Leveraging gene expression and local ancestry to investigate regulatory epistasis in humans

By

Alexandra Fish

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Human Genetics
May, 2017
Nashville, Tennessee

Approved:

Douglas P. Mortlock, Ph.D.

William Scott Bush, Ph.D.

John Anthony Capra, Ph.D.

Melinda C. Aldrich, Ph.D.

Joseph Lee Rodgers, Ph.D.

DEDICATION

I dedicate this work to both my family and Mike – I couldn't imagine ever having gotten here without you all. It certainly wouldn't have been as much fun.

Dad – You've had faith me in when I didn't. I'll always remember your sixth sense about Science and Math, and the infamous meeting with the high-school counselor. Mom – I have the courage to go anywhere and do anything, because I know you'll be at my back. We'll always have the fields of Sparta. Nick – Not many little brothers would willingly give up going to Prom to come to their sister's graduation. I'm so lucky to have one who is. Mike – You've been with me in this almost every step of the way – we've traveled the world together, and faced science-demons hand in hand…or, as close as is possible while still complying with fifth floor rules.

Now, enough of that sappy stuff. Let's get to science.

ACKNOWLEDGEMENTS

Science – unless perhaps you're a physicist – doesn't occur in a vacuum. The work presented here wouldn't have been possible without the guidance and support of my mentors, Drs. Will Bush and Tony Capra. Will's enthusiasm got me hooked on quantitative genetics, and his encouragement gave me the confidence to try my hand at something entirely new – coding. Tony took me in when times were a changing, and has been an unwavering source of support, encouragement, and reality checks ever since. Both have provided invaluable scientific guidance on the goals and design of this work, and have made phenomenal improvements on the (in some cases many) versions of the resulting manuscripts. It wouldn't have been possible without them.

My fellow graduate students have also provided vital support and advice as this project has developed. I would especially like to thank Corinne Simonti, for being a fantastic brainstorming-buddy, and R. Michael Sivley, for going through the harrowing experience of teaching me the fundamentals of Python. They both have enabled and advised me on the work presented here, for which I am very grateful.

Additionally, this work would not have been possible without the financial support of the Human Genetics training grant, T32GM080178. Money matters.

TABLE OF CONTENTS

Page

Chapter

Appendix

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
EPISTASIS IN HUMANS: EXPECTATIONS AND CHALLENGES

Epistasis is a phenomenon wherein the effect of a genetic variant on the phenotype is dependent on another genetic variant. The importance of epistasis – in terms of both prevalence, and effect sizes – to human health is controversial. Evidence from model organisms suggests that epistasis is widespread, and accounts for a notable proportion of phenotypic variance.[1,2] Similarly, additive genetic effects fail to fully account for the estimated heritability underlying the majority of human traits;[3,4] however, concrete evidence for epistasis influencing complex traits in humans remains elusive due to both experimental and computational limitations. Studies of epistasis in human have largely relied on observational studies. While the computational burden of performing genome-wide association studies for epistasis is no longer prohibitive for common variants, major statistical experimental limitations still complicate the field. In this dissertation, I take two approaches to address major concerns still facing the statistical study of epistasis in humans: a development of best practices, to ensure statistical interactions are indicative of biological epistasis, and then an investigation of epistasis between cis-regulatory variants influencing both a low-level phenotype directly tied to the nucleotide sequence – gene expression – and an array of complex phenotypes derived from electronic health records (EHR).

The different faces of epistasis – perspectives from statistics and molecular biology

The study of epistasis has taken several forms over the course of the last century, resulting in related but distinct definitions for the term in the fields of molecular and statistical genetics. William Bateson first coined the term epistasis in 1909[5], referring to a phenomenon wherein an allele at one locus is able to mask the effect of an allele at another locus. Bateson, who was expanding the work of Mendel, illustrated epistasis in sweet peas: he observed that peas with white flowers produced peas with purple flowers in non-Mendelian ratios (9:7) when crossed together.[6] From this, he concluded that the alleles interacted with one another to determine flower pigmentation. Further work has confirmed his hypothesis: these strains harbor mutations in two distinct genes within the enzymatic pathway responsible for processing anthocyanin, the pigment responsible for flower color.[7] Either mutation is therefore sufficient for the production of white flowers. Notably, the same principles underlie coat coloration in mammals: recessive, loss-of-function variants in genes responsible for synthesizing melanin, a pigment, result in albinos.[8] Thus, the molecular genetics definition of epistasis – in which an allele masks the effect of an allele at another locus – has been well illustrated, especially in the context of biological pathways.

The idea of epistasis was then adapted by R.A. Fisher in 1918 to describe a statistical phenomenon: interactions.[9] Mathematically, an interaction is defined as a deviation from additivity, which occurs when the combined effect of alleles at different loci is not equivalent to the sum of their individual effects.[10] This is conceptually analogous to the colloquial phrase “The whole is greater (or less) than the sum of the parts,” and an example is provided in Figure 1. This type of epistatic relationship can be represented by the inclusion of interaction terms in the mathematical models typically used in genetic association studies. However, this is a purely statistical construct, and as will be discussed in Chapter 2, the existence of a statistically significant interaction term does not necessarily indicate there is underlying biological epistasis.

For the remainder of this work, I will be referring to statistical epistasis as defined by R.A. Fisher.

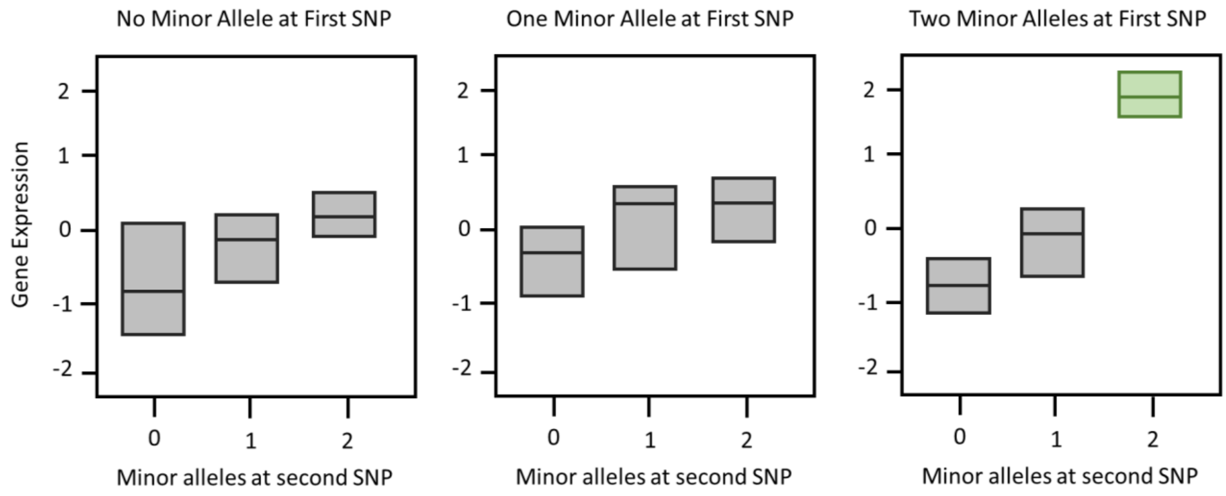


Figure 1. An example of a statistical interaction. In this visualization, individuals are stratified based on the number of minor alleles they have at two single nucleotide polymorphisms (SNPs). A summary of the gene expression for the individuals in each of the possible nine genotype combinations is provided as a boxplot. In this example, there is an increase in mean gene expression per minor allele of the second variant, indicating this variant is associated with gene expression levels independently. The first SNP has little influence on gene expression, except when there are two minor alleles at both loci. This group, highlighted in green, is a statistical interaction as the mean increase in gene expression anticipated for the second variant is far exceeded when it occurs on the background of two minor alleles at the first locus. Many other forms of statistical interactions are possible.

Epistasis is pervasive in model organisms

The genetic architecture underlying a variety of both fundamental biological processes and complex traits has been extensively studied in in an array of model organisms due to an inability to experimentally investigate the same questions, due to either ethical or methodological limitations, in humans. Below is a survey of the study of epistasis in three of the most commonly used model organisms: *Saccharomyces cerevisiae*, or yeast; *Drosphila melanogaster,* or fly; and *Mus musculus,* or mouse.

S. cerevisiae has long-served as model for discovering the genetic underpinnings of fundamental cellar processes, the basic principles of which are frequently shared with higher organisms. As a result, several strains of yeast have previously been genotyped,[11] efficient protocols exist to collect a variety of phenotypic data, and powerful molecular approaches, such as the two-hybrid cross, have already been developed to dissect the genetic underpinnings of traits.[12] Altogether, this makes yeast an ideal species to begin the exploration of epistasis.

Gene expression is an ideal phenotype to begin the study of epistasis, due to several methodological and biological considerations. Foremost, gene expression is a low-level phenotype in the sense that it is closely linked to the nucleotide composition; i.e., mutations in promoters or enhancers can directly disrupt the binding of transcriptional machinery, thereby influencing transcript levels. Moreover, two high-throughput methodologies exist to ascertain the expression of the majority of genes simultaneously: microarrays and RNA-sequencing. This enables the investigation of many phenotypes simultaneously, and thus results in a more comprehensive representation of the types of genetic architectures possible. Finally, and as will be discussed in more detail in Chapter 2, disruption of gene expression is believed to play a fundamental role in the development of complex disease. Consequently, understanding the genetic regulation of gene expression may translate into an increased knowledge of the genetic etiology of complex diseases pertinent to human health.

Brem and Kruglyak[13] conducted one of the first large-scale studies of epistasis, and capitalized on both the suitability of *S. cerevisiae* and the utility of gene expression. First, they crossed a common laboratory yeast strain and a wild isolate. In the recombinant offspring, they determined that roughly half of all genes had highly heritable expression levels, and moreover, almost six hundred (16%) had statistical evidence for epistasis. They determined this using a modified Lynch and Walsh test, which examines whether the mean expression level for a given transcript is the same for both the offspring and parents (as would be anticipated with additive genetic effects).[13] To identify the specific variants underlying the epistatic effects, they performed an expression quantitative trait loci analysis for epistasis. In this, a linear regression model was constructed for each pair of variants, in which an interaction term was included to represent epistasis. They identified over two hundred high-confidence statistical interactions between genetic variants.[14] As a proof of principle that the identified interactions were biological phenomenon, they further investigated an interaction between genetic variants that appeared to regulate the expression of multiple transcripts. They engineered isogenic yeast strains (i.e., strains genetically identical aside from the interacting variants) and found that the resulting changes in gene expression levels matched the previously observed patterns for 10 of the 15 transcripts. These studies illustrate that, in yeast, genetic variants interact with one another to have non-additive effects on the expression level of at least several hundred transcripts.

Next, Bloom et al.[1] addressed whether the genetic architecture of gene expression levels translated to that of more complex traits. Using the same yeast cross, Bloom et al.[1] investigated whether there was evidence for epistasis underlying twenty distinct growth-traits. They partitioned phenotypic variance into additive and epistatic genetic effects and found interactions accounted for between 2.2% and 21.2% of phenotypic variance, depending on the trait. They then performed a scan for pairwise interactions, and identified several hundred interactions between specific variants, which accounted for roughly half of the phenotypic variance attributed to interaction effects. The effect sizes of these interactions, however, were markedly reduced compared to those observed for additive effects. This illustrates that additive genetic effects likely dominate the genetic architecture for most traits; however, epistasis accounts for a notable proportion of phenotypic variance, largely due to large numbers of small-effect pairwise interactions.

While yeast serve a critical role in better understanding the genetic principles of higher organisms, it is always critical to illustrate that the principles discovered generalize between species. There are notable differences in the quantity and type of regulatory elements observed between yeast, fly, and higher level eukaryotes; namely, gene expression in yeast[15–17] is largely regulated by regulatory elements in close proximity to the target gene, whereas gene regulation in higher level eukaryotes such as fly[18,19], mouse[20–22] and human[20,21,23–25] is more frequently controlled by a combination of both proximal and distal-acting regulatory elements. Such striking differences in the regulatory landscape between species may critically influence the prevalence of epistasis, as perturbations to gene expression levels play a critical role in the development of complex traits pertinent to human health. Consequently, an exploration of epistasis in fly and mouse is a critical step in setting expectations for the phenomenon in human.

Epistasis has been extensively investigated in fly. This endeavor has been greatly aided by the *D. melanogaster* Genetics Reference Panel (DGRP), a repository of over 200 inbred lines with available whole-genome sequences and a variety of quantitative traits, including gene expression quantified by RNA-sequencing.[26] Huang et al. used linear regression to identify both expression quantitative trait loci (eQTL), or variants associated with the mean transcript level, and variance eQTL (veQTL), or variants associated with the variance in transcript levels in the DGRP.[27] Ultimately, they found that the vast majority of veQTL significantly interacted with a variant within the target gene's *cis*-regulatory region to regulate the variance in gene expression; however this could be accounted for by confounding factors addressed in the next Chapter.[27] Moreover, genetic variants associated with complex traits in the DGRP – specifically startle response, starvation resistance, and chill coma recovery time – demonstrated evidence for epistasis; the majority of variants associated with these traits engaged in at least one significant interaction.[2] Significant variants associated with aggressive behaviors also frequently engaged in epistatic relationships with other variants.[28] These variants were then used to create a gene-gene interaction network, wherein the nodes were the genes that variants mapped to (using proximity), and the edges were pairwise epistatic relationships. To validate the epistatic network, Shorter et al.[28] examined the phenotypic effects resulting from the knockout of either one, or both, gene partners in two pairs of directly interacting genes and two pairs of indirectly interacting genes. For three of the four gene pairs examined, the effect of the single-gene knockouts did not equal that of the double-gene knock out, indicating evidence of epistasis.

Consequently, an exploration of epistasis in mouse is a critical step in setting expectations for the phenomenon in human. Chromosome substitution strains (CSSs) have been an integral component of the study of epistasis in mouse; these strains of mouse have a single chromosome from a donor strain on the background of a host strain. This creates a homogeneous genetic background that reduces the 'noise' in phenotypic variation, which improves statistical power to detect effects.[29] Shao et al.[30] used a complete CSS panel between two strains, wherein a CSS had been constructed for each chromosome, to examine the genetic architecture of 41 blood, bone, and metabolic traits in mouse. The CSS construct enabled them to quantify the individual effect of each chromosome on each phenotype. Assuming additive genetic effects, the cumulative effect of each chromosome should equal the observed phenotypic difference between the two strains used to create the CSS panel. However, for 40 of 41 traits, the cumulative effect was significantly greater than the parental difference. Moreover, the median discrepancy was quite large, with the cumulative chromosome prediction anticipating a phenotypic difference eight times greater than that observed.[30] Thus, this study indicates that in mouse, epistasis between genetic variants is pervasive, with cumulatively large effect sizes for the vast majority of studied complex traits.

Altogether, evidence from model organisms suggests that epistasis is pervasive for both low-level molecular phenotypes such as gene expression, and high-level complex traits such as aggressive behavior. Studies of interactions between specific variants across multiple species have demonstrated that pairwise epistasis is not a rare phenomenon.[13,27,31] And while these pairwise interactions may have smaller effects than those observed for single variants,[1] in aggregate they account for a notable proportion of phenotypic variance depending on the trait of interest.[2,30] However, the approaches used to study epistasis in model organisms frequently use selective breeding strategies that would not be either ethical or practical to use for the study of epistasis in humans. Thus, while evidence from a diverse range of model organisms supports the idea that epistatic relationships between genetic variants influence complex phenotypes, different approaches must be taken to determine whether epistasis underlies complex traits in humans.

Insights from GWAS: The importance of the regulatory genome

Within the last decade, the genetic etiology of complex disease in humans has been interrogated primarily through genome-wide association studies (GWAS).[32–34] Previously, the interrogation of the genetic underpinnings of disease relied on two major approaches: linkage studies, and candidate studies.[35–37] Linkage studies identified genomic regions that segregated through families with the phenotype of interest; however, pinpointing the casual variation within these regions was often difficult due to limited resolution.[38–40] In contrast, candidate gene studies targeted a specific genomic region based on a priori knowledge of its function or on the basis of linkage studies, but were capable of identifying specific variants associated with the phenotype.[41,42] Due to recent technological advances, GWAS are able to combine many of the advantages of both of these approaches. In GWAS, millions of genetic variants genome-wide are rapidly interrogated for association with the phenotype, enabling both comprehensive coverage and the identification of comparatively narrow genomic regions.[35] Moreover, as this is a hypothesis-free approach, GWAS has the potential to identify novel genomic loci associated with disease.[34,35,41]

Close to three thousand GWAS have been performed for a multitude of complex traits, and a striking trend has emerged: the vast majority of genetic variants associated with complex disease are non-protein coding.[32,43] It was hypothesized that these variants were regulatory in nature, and influenced the expression level of gene products, rather than disrupting function of the gene product (i.e., protein) directly.[33,43–45] The development of both microarray and RNA-sequencing technologies in conjunction with genome-wide genetic data enabled the identification of variants associated with changes in gene expression, termed expression quantitative trait loci (eQTL).[46–54] Disease-associated variants identified in GWAS are enriched not only for eQTL status within relevant cell types,[55,56] but are additionally enriched for other signatures of regulatory function: DNase I hypersensitivity sites,[57] enhancers,[58] and transcription factor binding sites.[57] These results indicate that the majority of common variants associated with complex disease influence the phenotype via alterations in gene regulation.

For select examples, a causal link between perturbation of gene expression levels and complex diseases have been demonstrated. For example, Musunuru et al.[59] demonstrated that a variant in a non-coding region consistently associated with myocardial infarction and its causal risk factor LDL-cholesterol levels created a novel transcription factor binding site. This change resulted in a significant difference in the regulatory potential of this sequence in reporter assays, and was associated with the expression of the nearby gene *SORT1* in eQTL analyses. They demonstrated the change in *Sort1* levels causally influenced LDL-cholesterol levels by both knocking down and overexpressing *Sort1* in mice; the resulting changes in LDL-cholesterol levels were consistent with the direction of effect observed in humans. Thus, the disruption of gene expression levels by a single nucleotide polymorphism (SNP) can result in a complex phenotype. There are similar examples linking disease-associated variants to changes in gene expression and ultimately complex disease, indicating the disruption of regulatory elements is a general mechanism through which non-coding variation may influence complex traits.[60–64]

Given the preponderance of evidence highlighting the importance of gene regulation in the etiology of complex disease, I focus on regulatory epistasis in this work. However, epistasis between protein-coding variants influencing protein stability[65–67], protein-protein interactions[68], and protein-DNA interactions[69,70] has been documented and is reviewed within the literature.[71]

Evidence for regulatory epistasis in humans remains elusive

Evidence for regulatory epistasis within humans comes from three major sources: heritability studies, reporter assays of regulatory elements, and genetic association studies. Each approach provides unique insights into regulatory epistasis, and are reviewed below.

The first step in understanding the genetic architecture of a trait is often a heritability study. There are many different forms of heritability studies; however, they all provide a heritability estimate, which is a single score representing the proportion of phenotypic variance attributable to genetic variance across the entire genome. Thus, heritability estimates have traditionally been used to demonstrate the trait of interest has some genetic underpinning. Since the advent of GWAS, they have also frequently been used to estimate what proportion of the overall genetic effect is accounted for by individual genetic variants associated to the trait.[4] Strikingly, even when well-powered GWAS are performed, a notable proportion of the anticipated genetic effect is not accounted for by the individual variants.[4,72] Several potential explanations exist for this ‘missing heritability’ phenomenon, including epistasis as the vast majority of GWAS assume additive effects.[4,72–74] Ultimately, heritability studies demonstrate that additive genetic effects do not currenty account for all the genetic heritability of a trait; however, they provide only indirect evidence to support the existence of epistasis.

The principles of transcription factor binding also provide indirect evidence that cis-regulatory epistasis is plausible. Regulatory elements such as enhancers and promoters regulate gene expression levels via the binding of transcription factors. Transcription factors recognize specific sequence patterns, called motifs. For a recognized motif, the likelihood of the transcription factor binding is influenced by the multitude and relative location of additional motifs, highlighting the combinatorial nature of transcription factor binding.[75–77] Given the complex relationships between transcription factors and their critical role in regulatory activity, it is plausible that combinations of genetic variants influencing multiple transcription factor motifs could have epistatic relationships with one another.

This potential for epistasis within the cis-regulatory region has been directly investigated through the combination of massively parallel reporter assays and synthesized regulatory sequences. Kwasnieski et al. investigated the effect of genetic variation within *Rhodopsin* promoter on its regulatory function.[78] They engineered over a thousand single- and double-mutant sequences, and tested their regulatory activity in mammalian cells with *cis*-regulatory element sequencing (CRE-Seq). CRE-Seq is a reporter assay with methodological improvements, such that the regulatory function of thousands of sequences can be quantified simultaneously. For any given pair of variants, they were able to investigate epistasis by comparing the total effect of the two single-mutant sequence to that of the double-mutant sequence. They found that the majority of double mutants had regulatory activity levels significantly different than the anticipated combined effect of the single-mutants. Their work illustrates that epistasis within the cis-regulatory region is possible; however, it does not address how frequent epistasis is within natural populations.

Genetic association studies of epistasis may provide insight into how frequently epistatic relationships between variants occur in natural populations. Using Fisher's definition, such studies typically test pairs of variants for epistasis by including both variants' main effects and an interaction term between them.[79–82] Two studies have directly investigated regulatory epistasis influencing gene expression levels using naturally observed genetic variants. Hemani et al. performed a genome-wide association study of epistasis wherein all possible pairs of common genetic variants were tested for interactions associated with the expression of each gene expressed in whole blood.[80] They identified several hundred interactions that passed a strict Bonferroni multiple testing correction, and 30 interactions replicated with consistent directions of effect in independent datasets. Similarly, Brown et al. identified and replicated 57 interactions between pairs of variants that influenced gene expression levels in lymphoblastoid cell lines.[81] However, in a reply, Wood et al. demonstrated that these apparent interactions were in fact attributable to haplotype effects.[83] Essentially, the two putatively interacting variants were tagging a single, causal variant through linkage disequilibrium patterns. Ultimately, all example of epistasis identified by Hemani et al. were consistent with this phenomenon, calling into question whether epistasis occurs between regulatory variants.

Rather than investigation gene regulation directly, many other studies have investigated interactions between variants for a variety of complex traits in humans, including: bipolar disorder,[84] type 1 diabetes,[85] and Alzheimer's disease.[86] However, several factors complicate the interpretation of these results, which are discussed extensively in Chapter 2. Briefly, replication of interactions is notoriously difficult; many studies of epistasis either do not attempt replication,[85] or are unable to replicate the interactions.[84,86] Additionally, there are often alternative explanations for observed interactions such as scale effects[85] or haplotype effects.[83] While only a small subset of the studies of epistasis for complex traits are mentioned here, these issues are so pervasive within the field that a recent review concluded that "compelling statistical evidence is absent for the vast majority of reported epistatic interactions." [79] Thus, it is unclear how pervasive regulatory epistasis is in humans.

Ultimately, these results suggest that there is a great potential for regulatory epistasis: the transcriptional machinery acts in a combinatorial fashion, and additive genetic effects are insufficient to account for all of the heritability observed for complex traits. However, identifying the interacting elements via statistical association faces a major challenge: lack of clear best practices designed to ensure statistical interactions represent biological epistasis. In Chapter 2, I address this by performing a genome-wide investigation of epistasis between variants influencing gene expression using standard quality control procedures. I then systematically investigate confounding processes – both statistical and biological – that result in statistical interactions, but are not a product of biological epistasis. For each, I provide at least one example of an interaction identified in my analysis that was produced by the confounding process. I then provide guidelines and recommendations to correct for the confounding process as either a part of quality control or through post-hoc analyses. These are critical best practice guidelines for future studies of epistasis.

Regulatory epistasis within haplotypes: evidence, difficulties, and new approaches

Due to confounding issues discussed in Chapter 2, genetic association studies of epistasis typically focus on unlinked variants; however, there is evidence that epistasis occurs within haplotypes as well. Haplotypes are combinations of genetic variants that do not segregate independently from one another due to limited recombination. When a *de novo* genetic variant occurs, it therefore does so on a limited local background that is largely maintained; this could create the ideal opportunity for epistatic relationships between variants to be maintained. For example, deleterious coding variation might be masked by linked regulatory variation that reduces gene expression, and therefore maintained.

Three lines of evidence support the existence of regulatory epistasis within haplotypes. First, reporter assays illustrating regulatory epistasis investigate very small genomic regions – 50 base pairs in the case of Kwasnieski et al.[78] Due to their close physical proximity, it is likely that variants in these regions would be in high linkage disequilibrium (LD) with one another in natural populations. Secondly, Corradin et al. illustrated that multiple variants within haplotypes observed in natural populations can each influence the expression levels of the same gene.[87] While they did not demonstrate epistasis specifically, their findings support the hypothesis that nearby functional variants may occur on the same haplotype. Finally, Lappalainen et al.[88] demonstrated that rare, derived coding variants often arose on the background of common regulatory variants that decreased gene expression levels. This suggests that the formation of haplotypes may be influenced by functional relationships between variants. Thus, there is suggestive evidence for the existence of regulatory epistasis within haplotypes.

While limited recombination between variants creates a unique biological environment in which epistasis can arise, it also creates a missing data problem that complicates its study through statistical association methods. Without all combinations of genetic variants, regression-based approaches cannot accurately partition phenotypic variance into genetic components.[89] This difficulty has thus far largely prevented association-based studies of epistasis within haplotypes in natural populations. In Chapter 3, I leverage unique properties of admixed populations to investigate epistasis within haplotypes. Admixed populations arise when populations that have been historically reproductively isolated interbreed; for example, an admixture event between Europeans and West Africans beginning approximately eight generations ago has resulted in a two-way admixed population – African Americans.[90] Admixed populations represent a unique opportunity to investigate epistasis within haplotypes, as the recombination rate at a given genomic locus differs by continental ancestry.[91] Consequently, European haplotypes may be broken apart by African-specific recombination sites, and vice versa, in admixed populations. As ancestral haplotype boundaries are broken, I hypothesize that potentially novel combinations of genetic variants are formed that would enable the investigation of epistasis within the context of haplotypes.

This phenomenon may be detected by transitions in local ancestry, or the genetic ancestry at a specific genomic locus. There are many methods to estimate local ancestry, such as RFMix,[92] LAMP-LD,[93,94] and HAPMIX.[95] While the precise methodological approach differs between these methods, they follow a similar conceptual approach. First, the genetic information is phased, meaning that variants are placed within haplotypes, by comparing the observed genotypes to observed haplotypes within reference populations.[96] Next, the haplotype is assigned an ancestry, based on which of the continental populations it is more frequent in. Using these approaches, local ancestry – and transitions between ancestries – can be inferred genome-wide in admixed populations.

In Chapter 3, I use transitions in local ancestry to identify genomic regions in which ancestry haplotypes may frequently be broken apart by recombination events. To investigate epistasis in these contexts, I asked whether variants interact with these downstream ancestry transitions to influence an array of phenotypes derived from electronic medical records.

The use of electronic medical records for genetic research: opportunities and challenges

The vast majority of genetic studies are designed around a specific phenotype, or set of related phenotypes. However, the recent construction of biobanks linking electronic health record (EHR) data with genetic data has greatly expanded the depth of phenotypic information available on individuals.[97] This information is often encoded in the form of ICD-9 codes, which are diagnostic codes used for medical billing purposes. ICD-9 codes cover a wide array of phenotypes, which they are often used as proxies for. In addition to ICD-9 codes, the EHR also contains lab values for many tests, which can provide quantitative endophenotypes for many traits of interest. Thus, biobanks contain a wealth of both low-level, quantitative traits and abstractions of complex phenotypes on the same individuals, all linked to genetic information.

The unique properties of biobanks linking genetic information with EHR result in several advantages relevant to the study of epistasis. The first advantage is practical: samples are essentially already ascertained for many traits. This saves both the time and costs associated with sample ascertainment, and may enable research that would otherwise be cost-prohibitive, especially for younger investigators. Second, role of epistasis within the genetic architecture can be investigated for a broad array of phenotypes; consequently, the full spectrum of its prevalence can be estimated, rather than generalizing its frequency and importance within the genetic architecture based on a specific phenotype. Third, the phenotype of interest does not have to be predetermined; instead, variants of interest can be tested for association with any phenotype contained within the EHR. This has enabled a new type of genetic association test – phenome-wide association studies (PheWAS) – in which a genetic variant is systematically tested for association with all phenotypes.[98] As discussed in Chapter 3, this is a major benefit for investigating epistasis within haplotypes. Essentially, the genomic regions of interest are typically non-coding, and it is difficult to link these regions to a specific gene, much less a complex phenotype. Consequently, biobanks linking EHR and genetic data offer a unique opportunity to investigate the effects of epistasis across an array of phenotypes.

PheWAS has the potential to shed light on a variety of biological mechanisms and functions, in addition to epistasis. For example, both Verma et al. and Ye et al. performed a PheWAS for a subset of stop-gain variants, and identified and replicated both known and novel associations.[99,100] Their work helps to validate PheWAS as an approach, as they detected known associations, and suggests that PheWAS of stop-gain variation could be used to clinically characterize genes of unknown function. Simonti et al. performed a PheWAS for genetic variants derived from ancient admixture events with Neanderthals, linking evolutionarily intriguing variants of previously unknown function to phenotypes such as hypercoagulable state and protein-calorie malnutrition.[101] Others have used PheWAS to investigate pleiotropy, a phenomenon wherein a genetic variant influences multiple, sometimes seemingly distinct, phenotypes.[102,103] These studies could shed light on the shared genetic etiology of complex disease, and prioritize targets for drug development. Overall, PheWAS hold substantial promise for both better understanding genomic regions of unknown function, and better understanding cross-phenotype associations.

However, there are several issues and challenges that confront PheWAS. First, there is the concern that phenotypes derived from ICD-9 codes do not truly capture the phenotypes they purportedly represent. For example, codes may be assigned that reflect potential, rather than confirmed, diagnoses. This is especially true for autoimmune disorders, which frequently require diagnosis by a specialist.[104] To easily address this concern, some researches use the 'rule of two,' requiring at least two instances of the ICD-9 code to be considered a case; in the cases of type 2 diabetes, this has been shown to improve the positive predictive value.[105,106] Still, there is variability in the specificity and positive predictive values of each ICD-9 code's ability to represent the underlying phenotype, and performing a manual chart review for every ICD-9 code is not currently feasible.[106] To address this concern, Denny et al. performed a PheWAS for known disease-genotype associations derived from the GWAS Catalog, and were able to replicate approximately two thirds of associations they were well-powered to detect.[107] Other PheWAS have also replicated known associates derived from the GWAS-Catalog, illustrating that ICD-9 code defined phenotypes can recapitulate known associations to their analogous, more traditionally defined phenotypes.[99,108–110]

Multiple testing correction is an additional challenge for PheWAS. There are over 17,000 ICD-9 codes; if each is tested, a Bonferroni threshold of 2.9×10^{-6} would be required to adjust for a single variant.[105] This issue will only be further exacerbated with the transition ICD-10 codes, of which there are over 150,000 to improve diagnostic resolution.[105] However, many of these codes are not actually independent from one another; for example, there are several hundred codes for tuberculosis, each of which specifies a unique site of infection.[104] To address this concern, 'PheCodes' have been developed, which both combine redundant ICD-9 codes and provide exclusion criteria for putative controls.[98,111] There are 1,724 PheCodes, which substantially reduces the number of association tests performed.[104] However, PheWAS are often restricted a small number of variants, or tested for association with specific subset of phenotypes, to maintain sufficient power to detect effects.

Overall, PheWAS offer many advantages to better understand both the clinical consequences of genomic regions of unknown function, and to elucidate the shared genetic architecture between phenotypes. However, the underlying data was not designed for research purposes and likely contains more noise than traditionally ascertained cohorts. Consequently, replication and careful consideration of observed genotype-phenotype associations are essential for PheWAS.[104]

Challenges facing the investigation of regulatory epistasis

In this work, I address two major challenges within the field of regulatory epistasis: the development of statistical best practices, and the investigation of epistasis within haplotypes. In Chapter 2, I identify and replicate statistical interactions between cis-regulatory variants. I then review known sources of statistical confounding for the study of epistasis, and introduce novel forms of confounding. Ultimately, I develop a set of best statistical practices for the study of epistasis that address these confounders via additional quality control procedures or post-hoc analyses. In Chapter 3, I use unique properties of admixed populations to investigate epistasis within the context of haplotypes. Due to ancestry-specific recombination hotspots, haplotypes can be broken apart in admixed populations thereby enabling the detection of epistasis; this can be detected via transitions in local ancestry. I performed a PheWAS to better understand how regions harboring many local ancestry transitions influenced an array of clinical phenotypes. I identified several interactions between variants and nearby local ancestry transitions influencing red blood cell traits, which serves as a proof of principle for the utility of this approach.

CHAPTER 2
ARE STATISTICAL INTERACTIONS EVIDENCE FOR BIOLOGICAL EPISTASIS?[1]

Introduction

The importance of epistasis to the development of complex traits in humans has been highly contested. Despite evidence for wide-spread epistasis in model organisms,[1,2,112] evidence for epistasis influencing complex traits in human remains elusive. This may be attributable either an actual lack of epistasis in humans, the inherent inability to tightly control a variety of factors when studying phenotypes in humans, or the fact that most phenotypes studied are several steps removed from the underlying biological processes that influence them. These last two explanations are methodological limitations that make it unclear whether the lack of observed epistasis in humans is a true feature of the genetic architecture, or if epistasis is simply much more difficult to observe outside experimental systems.

Human-derived cell lines, while a proxy for primary tissue, provide a unique opportunity to investigate epistasis. Like model systems, the environment for cell lines can be tightly controlled, and moreover, comprehensive genetic and gene expression data can readily be collected by high-throughput methodologies. Gene expression is an ideal phenotype to study epistasis for a variety of reasons. First, the genetic architecture underlying thousands of genes' expression can be investigated simultaneously with either microarray or RNA-sequencing, meaning that the full spectrum of epistasis is likely to be captured. Secondly, the molecular mechanisms that drive gene expression are directly tied to the nucleotide sequence itself: transcription factors recognize and bind motif sequences to regulate gene expression, and disruption of these nucleotide sequences can alter expression levels.[59] Also, the regulation of gene expression is known to involve complex molecular interactions among transcription factors and regulatory sequences, and experimental maps of chromatin looping and transcription factor binding enable biological interpretations for observed statistical interactions.[25,113] Finally, the study of gene expression is directly relevant to complex disease: the vast majority of variants identified in genome-wide association studies are non-protein coding. Thus it is presumed that the disruption of gene regulation is causally involved in the development of many common diseases.[43,114] In several instances, it has been shown that single nucleotide variants regulate gene expression by altering the function of regulatory elements, and that these altered gene expression profiles result in clinical phenotypes.[59,61] By better understanding the genetic control of gene expression, I may therefore better understand the genetic architectures underlying complex disease.

Genetic variants associated with gene expression levels – termed expression quantitative trait loci (eQTL) – have been studied extensively in primary human tissue and in cell lines. In many eQTL analyses, a gene-based approached is taken wherein variants within the *cis*-regulatory region for a given gene are tested for association with its expression. Until recently, the number of association tests required to perform a similar genome-wide association test for interactions was not computationally feasible. However, advances in computational power are

[1] Adapted from Fish et al., Are Interactions between cis-Regulatory Variants Evidence for Biological Epistasis or Statistical Artifacts?, The American Journal of Human Genetics (2016), http://dx.doi.org/10.1016/j.ajhg.2016.07.022

continually diminishing this barrier and two genome-wide studies of epistasis have identified replicating interactions.[80,81] The validity of these interactions, however, was questioned when it was demonstrated that through complex linkage disequilibrium (LD) patterns, these putative interactions could tag single variant eQTL.[83] In such cases, the genotypes at the two putatively interacting loci together were highly informative of the genotype at single variant eQTL; consequently, they were identified as statistically interacting, although this relationship disappears when the effect of the single variant eQTL is conditioned on. Notably, all of the interactions identified in prior studies were either no longer significant or were strongly attenuated when the effects of additional *cis*-eQTL were considered. This illustrates that, compared to single-locus analyses, the statistical models used to detect epistasis are subject to novel confounding factors, which are rarely addressed in studies of epistasis.

In this study, I investigate whether evidence for epistasis within the *cis*-regulatory region in humans persists after systematically accounting for technical, statistical, and biological confounding factors. I performed a targeted investigation of interactions regulating gene expression levels in human lymphoblastoid cell lines (LCLs): the analysis was restricted to nominal eQTL within the target gene's *cis*-regulatory region ($p<0.05$) to drastically reduce the number of association tests performed[1,115] while retaining the genomic regions most likely to harbor pertinent regulatory elements. Few genes showed evidence of epistasis (165 of 11,465 genes tested), although multiple interactions were often detected for the same gene. A total of 1,119 interactions were identified, many of which replicated in an independent dataset (90 of 803 possible). I then investigated confounding factors – technical (variants within probe binding sites, ceiling/floor effect), statistical (missing genotype combinations, population stratification), and biological (haplotype effects, tagging *cis*-eQTL) – that provide alternative, more parsimonious explanations than biological epistasis. Ultimately, each of the interactions identified could be accounted for by an alternative mechanism, suggesting that the majority of statistical interactions identified without accounting for confounding factors are spurious associations. Many of these confounding factors are inherent to the statistical models used, and will therefore generalize to other phenotypes; consequently, the analytic framework of this study will be of use to many future studies of statistical epistasis.

Subjects and methods

Genotyping and gene expression data

The discovery dataset was comprised of individuals ascertained as part of the International HapMap Project, PhaseI+II,[116] which consisted of 210 unrelated individuals with genome-wide genotyping data (Phase I+II, release 24). For each of these individuals, Stranger et al. collected and normalized gene expression levels from immortalized LCLs using the Sentrix Human-6 Expression Bead Chip, v1.[53] All probes with a HapMap SNP underlying the expression probe were removed from analysis.[53] I applied a population normalization procedure, described by Veyrieras et al.,[117] to the gene expression values that such that the expression of each gene within each population followed a normal distribution. This removed population-level differences in gene expression, which enabled us to combine all ethnicities in our analysis. Our replication dataset consists of 232 unrelated individuals from the 1000 Genomes Project (1KG), for whom gene expression in LCLs was available. These individuals had been sequenced at low coverage as part of the 1KG Project;[118] I used genetic data from phase I, version 3. Stranger et al. also collected and normalized gene expression levels in LCLs for these individuals using Illumina Sentrix Human-6 Expression BeadChip, v2.[54] I applied the same population normalization procedure [117] to these data. Both the discovery and replication dataset are multiethnic; the sample composition by ethnicity is shown in Table 1.

Two additional replication datasets were used to investigate a promising interaction. The first consisted of 283 European-descent individuals from the Genotype-Tissue Expression (GTEx) Project, for whom gene expression in whole blood was assessed by RNA-sequencing.[119] Genotype data for these individuals was collected on both the HumanOmni5-Quad Array and the Infinium Exome Chip, and then imputed to 1KG.[119] The second dataset consisted of brain samples from autopsied European-descent individuals in the Mayo Late Onset Alzhemier's Disease Consortium.[120] These individuals were genotyped on the Illumina HumanHap300-Duo Genotyping Beadchip and gene expression was collected using the Illumina Whole-Genome DASL HT BeadChip.[120] 370 individuals had expression data available from cerebellum, and 385 had expression in the temporal cortex.

Analysis	Total	Ethnicity							
	Sample Size	CHB	CEU	GIH	JPT	LWK	MXL	MKK	YRI
Discovery	210	45	60	-	45	-	-	-	60
Replication	232	34	-	-	35	80	38	-	45

Table 1. Dataset Composition by Ethnicity. The number of individuals of each ethnicity (1KG abbreviations) in the discovery and replication analyses.

Generating SNP pairs for interaction testing

To generate SNP-pairs for each gene, I first identified all common SNPs within the gene's *cis*-regulatory region. To be considered common, variants had to have a MAF > 5% when all ethnicities were combined. Based on *cis*-eQTL analyses,[117] the *cis*-regulatory region was defined as starting 500 kb upstream of the gene's start and ending 500 kb downstream of the gene's stop (including the gene itself); gene boundaries were taken from ENSEMBL. Previously, these variants were individually tested for association with the gene's expression level in the discovery dataset by Veyrieras et al.[117] Based on this analysis, I filtered out SNPs whose marginal effects were not nominally associated with gene expression (excluded p > 0.05), under the hypothesis that nominally associated variants may represent weak marginal effects from a true underlying interaction. I then considered all possible SNP-pairs amongst the remaining variants. Once this was done for each gene, over 21 million SNP-pairs were generated for interaction testing.

Identifying significant interactions

Each SNP pair was tested for interactions significantly associated with the expression of the gene for which it was generated. The below interaction model (Equation 1)[121] was used, which contains additive and dominant effects for each variant and all four possible interaction terms in order to ensure that variance is properly partitioned across the genetic terms.

$$y = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2 + i_{aa}x_1x_2 + i_{ad}x_1z_2 + i_{da}z_1x_2 + i_{dd}z_1z_2 + PC_{1-3}$$

(Equation 1)

where y represents gene expression, x_1 and x_2 use additive encoding to represent the genotype at SNP A and SNP B respectively, z_1 and z_2 use Cordell's [121] dominant encoding to represent the genotype at SNP A and B respectively, a_1 and d_1 are estimated coefficients representing the additive and dominant effects of SNP A, a_2 and d_2 are estimated coefficients representing the additive and dominant effects of SNP B, and i_{aa}, i_{ad}, i_{da} and i_{dd} are estimated coefficients representing both additive and dominant interaction effects. The top three principal components were also included as covariates (PC_{1-3}). To determine the significance of interactions, this model was compared to a reduced model lacking the four interaction terms using a likelihood ratio test (LRT) (Equation 2).

$$y = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2 + PC_{1-3} \quad \text{(Equation 2)}$$

This test was implemented using the program INTERSNP.[122] I calculated an FDR of 5% using the qvalue package in R.[123]

Identification of representative interaction eQTL models for distinct pairs of interacting genomic loci

Some interaction eQTL (ieQTL) models identified in the discovery analysis were redundant due to LD. For two ieQTL models to be considered redundant, each SNP within one significant ieQTL model had to be in high LD ($r^2 \geq 0.9$) with a SNP within the second ieQTL model, and vice versa. By using this criterion, the pairs were effectively correlated at $r^2 \geq 0.8$, the threshold typically used for tag-SNP selection. Redundant ieQTL models were grouped together. The model with the most significant LRT p-value in the discovery analysis was used to represent the entire group in most analyses, so that each pair of interacting genomic loci was equally represented.

Statistical power estimation

We performed simulation analyses to determine the power to identify interactions. I first randomly sampled a set of 20,000 SNP-pairs having all nine genotype combinations present, and then used the observed genetic data to simulate gene expression values. I simulated gene expression values based on the observed genotypes, the actual additive and dominant main effects for each of the two interacting variants, an error term drawn from a standard normal distribution, and embedded interaction terms of varying strength.

To properly represent the main effects of the variants, I used βs for the additive and dominant terms for each variant reflecting the actual effects within our dataset. I used the following model (Equation 3):

$$y = \mu + a_1x_1 + d_1z_1 + PC_{1-3} \text{ (Equation 3)}$$

where y represents gene expression, uses x_1 additive encoding to represent the genotype for the variant, z_1 uses Cordell's [121] dominant encoding to represent the genotype, and the top three principal components were included as covariates ($PC_{1\text{-}3}$).

I then determined the effect size for the interaction terms. There are four interaction terms in the model: additive by additive (i_{aa}); additive by dominant; dominant by additive; and dominant by dominant. The i_{aa} term is significant in all significant interaction models identified in the actual discovery analysis, whereas the other terms are not – these terms are included so that phenotypic variance is appropriately partitioned between genetic components. Consequently, these three interaction terms were treated as nuisance variables when simulating gene expression values; their βs were drawn from a normal distribution (mean = 0, standard deviation = 0.03). I used the effect sizes of *cis*-eQTL ($p < 5.0 \times 10^{-8}$) in our analysis to establish a 'moderate' anticipated effect size (*cis*-eQTL median: β=0.771) and a 'high' anticipated effect size (*cis*-eQTL 75th percentile: β=0.908). These βs are well within the range of observed effect sizes for significant interactions (i_{aa} median: β = 0.65 and i_{aa} max: β = 2.57). I then simulated gene expression data for each of the two effect sizes for each pair of SNPs.

Next, I performed the same LRT used in the discovery analysis to identify significant interactions. All interactions with p-values below the FDR=5% threshold ($p \leq 1.328 \times 10^{-5}$) were considered significant. I then repeated this process 10 times using the same 20,000 pairs of variants. In each of these ten iterations, power was calculated as the total number of pairs found to have a significant interaction divided by the total number of simulated interactions tested.

Variants within the probe binding site

To determine if variants were within the probe binding locations, I first used BLAT to identify the probe binding location in hg19 coordinates. Some probes returned multiple hits; consequently, I filtered the binding sites (binding sites had to be on the same chromosome as the gene, have a length > 30 base pairs, and an identity score > 95%) to identify unique binding locations. I then exclusively looked within a subset of our discovery dataset with sequencing data in the 1KG Project (n=174) to determine if there were any variants within binding sites that might confound the interaction analysis.

Ceiling/floor effect

Microarrays have a limited dynamic range that is not able to capture the extremes of gene expression. If the combined additive effect of two variants exceeds the threshold of detection, their apparent combined effect will be less than the sum of their individual effects. Thus, they may be spuriously identified as interacting. If this occurs, there will be a characteristic pattern of βs: the main effects for variants will be in the same direction, and the interaction term β will be in the opposite direction. I looked for this characteristic pattern to determine an upper bound of the prevalence of the ceiling/floor effect within our results. First, I identified the significant variables (β±SE could not contain zero) in the model. All interactions were then categorized as having 0, 1, or 2 SNPs with a significant main effect - either additive or dominant main effects counted; if both additive and dominant main effects were significant for the same variant, the one with the largest effect size was used to represent the main effect. For interactions where both variants had at least one significant main effect, I determined whether or not they had a concordant direction of effect. For those pairs with concordant directions of effect, I compared the significant interaction term with the largest absolute effect size to determine if it was discordant with the main effects. If this was the case, the interaction had a pattern consistent with a ceiling/floor effect, and was not considered clear evidence for epistasis.

Population specific cis-eQTL

Population-specific *cis*-eQTL can confound the interaction analysis, even though gene expression values were population normalized and the top three PCs were included as covariates. To investigate this, I first stratified the discovery dataset by each of the three ethnicities (CEU, YRI, CHB+JPT), and tested each interaction for significance, using the same methodology. For interactions that were not significant ($p < 0.05$) in any of the populations, I determined if the interacting variants were population-specific *cis*-eQTL using Equation 3. Variants with nominally significant ($p < 0.05$) main effects were considered *cis*-eQTL. If a variant was identified as a *cis*-eQTL in only a subset of populations, it was considered population-specific.

Conditional cis-eQTL analysis

To determine if ieQTL pairs were tagging a *cis*-eQTL as suggested by Wood et al.,[83] I first identified all nominal *cis*-eQTL ($p < 0.05$) for genes with significant ieQTL. To identify all nominal *cis*-eQTL, I used a subset of the discovery analysis individuals (n=174) who were also sequenced as part of the 1KG Project.[118] I used the called genotypes from Phase III, v5. The same gene expression data previously described for the discovery set was used. Within this subset, I performed a single-marker *cis*-eQTL analysis for each common variant (MAF > 5%) within the *cis*-regulatory region using Equation 4:

$$y = \mu + a_1x_1 + PC_{1-3} \qquad \text{(Equation 4)}$$

where *y* represents gene expression, x_1 uses additive encoding to represent the genotype for the variant, and the top three principal components were included as covariates ($PC_{1\text{-}3}$). Variants with nominal significant ($p < 0.05$) main effects were considered *cis*-eQTL.

To determine if any of these *cis*-eQTL could account for the interaction, I created all pairs of *cis*-eQTL and ieQTL for the same gene. I incorporated each *cis*-eQTL into each interaction model (Equation 5) as shown below.

$$y = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2 + a_3x_3 + d_3z_3 + i_{aa}x_1x_2 + i_{ad}x_1z_2 + i_{da}z_1x_2 + i_{dd}z_1z_2 + PC_{1-3} \qquad \text{(Equation 5)}$$

where *y* represents gene expression, x_1 and x_2 use additive encoding to represent the genotype at interacting SNPs A and B respectively, z_1 and z_2 use Cordell's dominant encoding to represent the genotype at interacting SNPs A and B respectively, a_1 and d_1 are estimated coefficients representing the additive and dominant effects of SNP A, a_2 and d_2 are estimated coefficients representing the additive and dominant effects of SNP B, and i_{aa}, i_{ad}, i_{da} and i_{dd} are estimated coefficients representing both additive and dominant interaction effects. The main effect of the *cis*-eQTL is represented with additive encoding by x_3 and with dominant encoding by z_3; the estimated coefficients corresponding to the main effects are a_3 and d_3 respectively. The top three principal components were also included as covariates ($PC_{1\text{-}3}$). I then performed a LRT comparing this model to a reduced model lacking the interaction terms (Equation 6).

$$y = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2 + a_3x_3 + d_3z_3 + PC_{1-3} \qquad \text{(Equation 6)}$$

If the LRT p-value of an interaction was nominally significant ($p < 0.05$) for all conditional analyses, I considered this evidence that the interaction and *cis*-eQTL represented independent signals.

Results

Discovery and replication of genetic interactions that impact gene expression levels

We identified interactions between nominal *cis*-eQTL that were significantly associated with gene expression levels. Our analysis was conducted using 210 individuals from the HapMap Project, Phase I+II, on whom both genotyping[116] and gene expression data within LCLs[53] were available. A population normalization procedure was applied to the gene expression data, so that there were no systematic differences between populations.[117] The overall workflow for the analysis is shown in Figure 2. For each gene with expression data (n=11,465), I identified common SNPs (global MAF > 5%) within its *cis*-regulatory region, defined as 500 kb upstream to 500 kb downstream of the gene. To increase power, I only considered variants nominally associated with the gene's expression ($p < 0.05$) in a single-marker analysis.[117] I analyzed all pairwise combinations of these variants for each gene, resulting in over 21 million SNP pairs. I then performed a likelihood ratio test (LRT) comparing a full model, which contains the top three PCs, main effects, and interaction terms, to a reduced model, containing only the covariates and main effects, to determine which interactions significantly improved model fit.[121] Given the large number of correlated tests, I controlled the false discovery rate (FDR) at 5% ($p \leq 1.328 \times 10^{-5}$) across p-values from all LRT performed.[123]

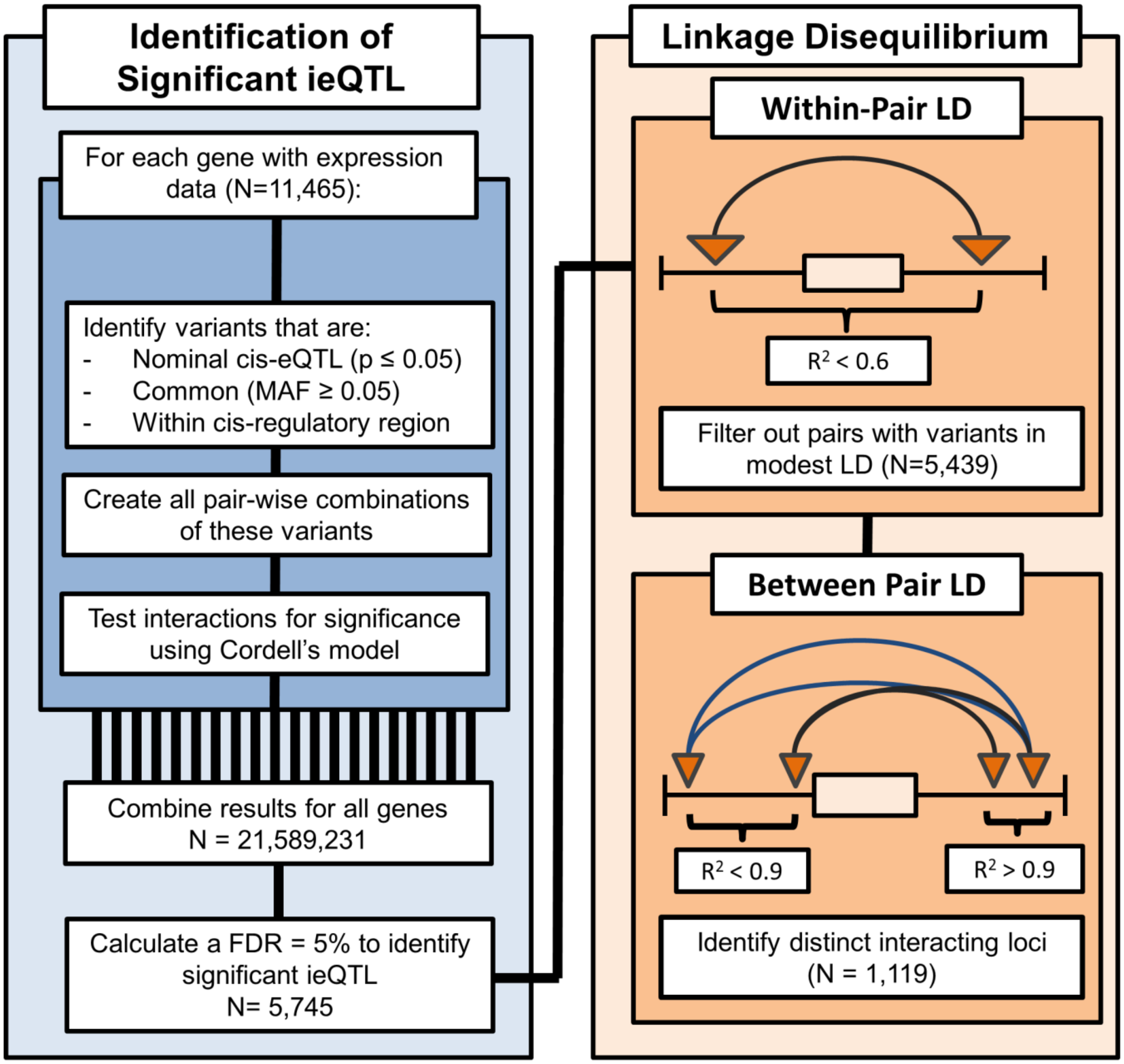

Identification of Significant ieQTL
For each gene with expression data (N=11,465):
Identify variants that are:
- Nominal cis-eQTL (p ≤ 0.05)
- Common (MAF ≥ 0.05)
- Within cis-regulatory region
Create all pair-wise combinations of these variants
Test interactions for significance using Cordell's model
Combine results for all genes N = 21,589,231
Calculate a FDR = 5% to identify significant ieQTL N= 5,745
Linkage Disequilibrium
Within-Pair LD
R2 < 0.6
Filter out pairs with variants in modest LD (N=5,439)
Between Pair LD
R2 < 0.9
R2 > 0.9
Identify distinct interacting loci (N = 1,119)

Figure 2. Workflow used to identify and group ieQTL. In the discovery analysis, nominally significant *cis*-eQTL (denoted by triangles) were paired together and tested for interactions significantly associated with gene expression levels (denoted by arcs). The within-pair LD was then calculated (Figure 3), and interactions composed of variants in modest LD ($r^2 > 0.6$) with one another were removed from the remainder of the analysis. Some of the remaining interactions represented the same pair of interacting genomic loci (Figure 4), and were partitioned into distinct groups (denoted by the arc color). For two interactions to be grouped together, each SNP within one significant ieQTL model had to be in high LD ($r^2 \geq 0.9$) with a SNP within the second ieQTL model, and vice versa.

An objective of this analysis is to characterize how frequently epistasis occurs; therefore, I next performed a power analysis to determine our ability to detect interactions in this dataset. First, I randomly sampled a set of 20,000 SNP-pairs tested in this analysis. Then, I simulated gene expression values using the observed genotypes and the observed main effect for each of the variants. I then imbedded an interaction effect into the simulated gene expression values, using both a moderate and large effect size, which were derived from the observed cis-eQTL effect sizes in this dataset. Assuming moderate and large effect sizes respectively, I had 21.6% - 55.3% and 44.3% – 78.9% power to detect interactions between high-frequency variants (MAF 0.2 – 0.5) in low LD with one another (Table 2). Thus, many potential examples of epistasis within the *cis*-regulatory region may not have been detected by this analysis, especially for low-frequency variants or those in high LD with one another.

Low LD ($r^2 < 0.05$)				
MAF Range		Percentage	Effect Size	
Variant 1	Variant 2		Moderate	Large
0.05 <= MAF < 0.1	0.05 <= MAF < 0.1	0.02	0.0 ± 0.0	3.3 ± 10.5
	0.1 <= MAF < 0.2	0.22	2.0 ± 2.1	5.5 ± 1.9
	0.2 <= MAF < 0.3	0.55	3.5 ± 1.1	9.5 ± 2.4
	0.3 <= MAF < 0.4	0.87	5.5 ± 1.5	12.4 ± 3.0
	0.4 <= MAF <= 0.5	1.11	4.3 ± 1.1	11.9 ± 1.6
0.1 <= MAF < 0.2	0.1 <= MAF < 0.2	1.04	5.7 ± 1.4	16.7 ± 3.2
	0.2 <= MAF < 0.3	5.30	10.8 ± 1.1	25.8 ± 0.6
	0.3 <= MAF < 0.4	6.79	14.9 ± 1.2	33.3 ± 1.1
	0.4 <= MAF <= 0.5	8.16	16.2 ± 1.0	36.2 ± 1.1
0.2 <= MAF < 0.3	0.2 <= MAF < 0.3	4.47	21.6 ± 1.4	44.3 ± 1.2
	0.3 <= MAF < 0.4	10.59	30.9 ± 0.6	57.6 ± 1.3
	0.4 <= MAF <= 0.5	11.01	35.8 ± 0.7	62.7 ± 1.0
0.3 <= MAF < 0.4	0.3 <= MAF < 0.4	5.23	44.3 ± 1.0	71.1 ± 0.9
	0.4 <= MAF <= 0.5	10.58	50.3 ± 0.6	75.5 ± 0.8
0.4 <= MAF <= 0.5	0.4 <= MAF <= 0.5	4.75	55.3 ± 1.9	78.9 ± 0.6

Moderate LD ($0.05 <= r^2 < 0.3$)				
MAF Range		Percentage	Effect Size	
Variant 1	Variant 2		Moderate	Large
0.05 <= MAF < 0.1	0.05 <= MAF < 0.1	0.04	0.0 ± 0.0	1.4 ± 4.5
	0.1 <= MAF < 0.2	0.32	2.4 ± 1.8	4.9 ± 2.7
	0.2 <= MAF < 0.3	0.21	2.7 ± 2.5	6.6 ± 2.8
	0.3 <= MAF < 0.4	0.04	1.4 ± 4.3	4.3 ± 9.6
	0.4 <= MAF <= 0.5	0.02	3.3 ± 10.0	0.0 ± 0.0
0.1 <= MAF < 0.2	0.1 <= MAF < 0.2	0.86	4.2 ± 1.8	10.4 ± 1.9
	0.2 <= MAF < 0.3	1.88	7.6 ± 1.5	19.8 ± 1.7
	0.3 <= MAF < 0.4	1.43	10.2 ± 1.2	24.6 ± 2.9
	0.4 <= MAF <= 0.5	0.86	11.9 ± 1.8	31.2 ± 2.9
0.2 <= MAF < 0.3	0.2 <= MAF < 0.3	1.62	14.6 ± 0.7	32.1 ± 2.0
	0.3 <= MAF < 0.4	3.40	20.4 ± 1.7	42.5 ± 1.7

	0.4 <= MAF <= 0.5	3.34	22.5 ± 1.7	46.2 ± 1.5
0.3 <= MAF < 0.4	0.3 <= MAF < 0.4	2.18	25.9 ± 1.5	52.5 ± 2.4
	0.4 <= MAF <= 0.5	5.24	31.0 ± 0.8	56.0 ± 1.1
0.4 <= MAF <= 0.5	0.4 <= MAF <= 0.5	2.85	35.2 ± 1.7	61.6 ± 2.2

High LD (0.3 <= r^2 < 0.6)				
MAF Range		Percentage	Effect Size	
Variant 1	Variant 2		Moderate	Large
0.05 <= MAF < 0.1	0.05 <= MAF < 0.1	0.01	0.0 ± 0.0	0.0 ± 0.0
	0.1 <= MAF < 0.2	0.03	2.0 ± 6.0	2.0 ± 6.3
	0.2 <= MAF < 0.3	0.00	-	-
	0.3 <= MAF < 0.4	0.00	-	-
	0.4 <= MAF <= 0.5	0.00	-	-
0.1 <= MAF < 0.2	0.1 <= MAF < 0.2	0.20	1.0 ± 1.2	2.5 ± 2.4
	0.2 <= MAF < 0.3	0.29	2.1 ± 1.5	7.1 ± 3.6
	0.3 <= MAF < 0.4	0.02	0.0 ± 0.0	0.0 ± 0.0
	0.4 <= MAF <= 0.5	0.00	-	-
0.2 <= MAF < 0.3	0.2 <= MAF < 0.3	0.65	3.1 ± 0.9	12.2 ± 2.4
	0.3 <= MAF < 0.4	0.67	6.0 ± 2.1	17.1 ± 2.1
	0.4 <= MAF <= 0.5	0.11	7.3 ± 4.2	18.6 ± 10.2
0.3 <= MAF < 0.4	0.3 <= MAF < 0.4	0.82	7.8 ± 2.1	17.4 ± 3.9
	0.4 <= MAF <= 0.5	1.11	10.1 ± 2.1	22.9 ± 3.5
0.4 <= MAF <= 0.5	0.4 <= MAF <= 0.5	1.18	9.8 ± 1.7	24.9 ± 3.1

Table 2. Power to Detect Interactions by MAF and LD. Power to detect interactions is contingent upon both the MAF of the two variants and the LD between the variants. To calculate power, I randomly selected 20,000 pairs of variants tested in this analysis and simulated gene expression values with interaction effects at a moderate (median β of *cis*-eQTLs; $\beta = 0.771$) and a large (75th percentile β of *cis*-eQTLs; $\beta = 0.908$) effect size (Methods). I then binned interactions according to their MAF and LD, and calculated power as the number of significant interactions divided by the total number of interactions within each bin. I repeated this process ten times, and computed the mean power and its standard deviation across all 10 runs for each bin, which is reported here. For each bin, I also report the percentage it accounted for of the 20,000 interactions.

LD between variants complicates the interpretation of the interaction models. I addressed two types of LD in significant interaction models: within-pair LD, defined as the LD between the variants in the same interaction model, and between-pair LD, defined as the LD between variants in different interaction models. Modest within-pair LD indicates the variants may be identifying a haplotype, which could carry a single variant that is actually driving the association with gene expression. Wood et al. have demonstrated that even very stringent LD-pruning thresholds ($r^2 > 0.1$ or D’ > 0.1) are insufficient to protect against confounding by *cis*-eQTL,[83] therefore I adopted a two-stage strategy to address this concern. First, I removed all pairs with variants in modest LD with one another ($r^2 > 0.6$) from the remainder of the analysis (median r^2 between remaining pairs of interacting variants was 0.06, Figure 3). I then directly tested for confounding by *cis*-eQTL in a later analysis. Ultimately, 5,439 interaction models were both significant and passed the within-pair LD filtering criteria; they were significantly associated with the expression of 165 unique genes, which are provided in Fish et al.'s[124] Supplemental Table 2. I then calculated between-pair LD, or the correlation of variants in different interaction models. Highly correlated interaction models were grouped together (Methods; Figure 2) because they likely represent the same pair of interacting genomic loci, as evidenced by their very similar statistical models (Figure 4). The 5,439 interaction models represented 1,119 pairs of interacting genomic loci (Fish et al.'s[124] Supplemental Table 2). The interaction model with the most significant p-value in the discovery analysis was selected to represent the entire group in all subsequent analyses, unless specifically stated otherwise, to ensure that each pair of interacting genomic loci was equally represented.

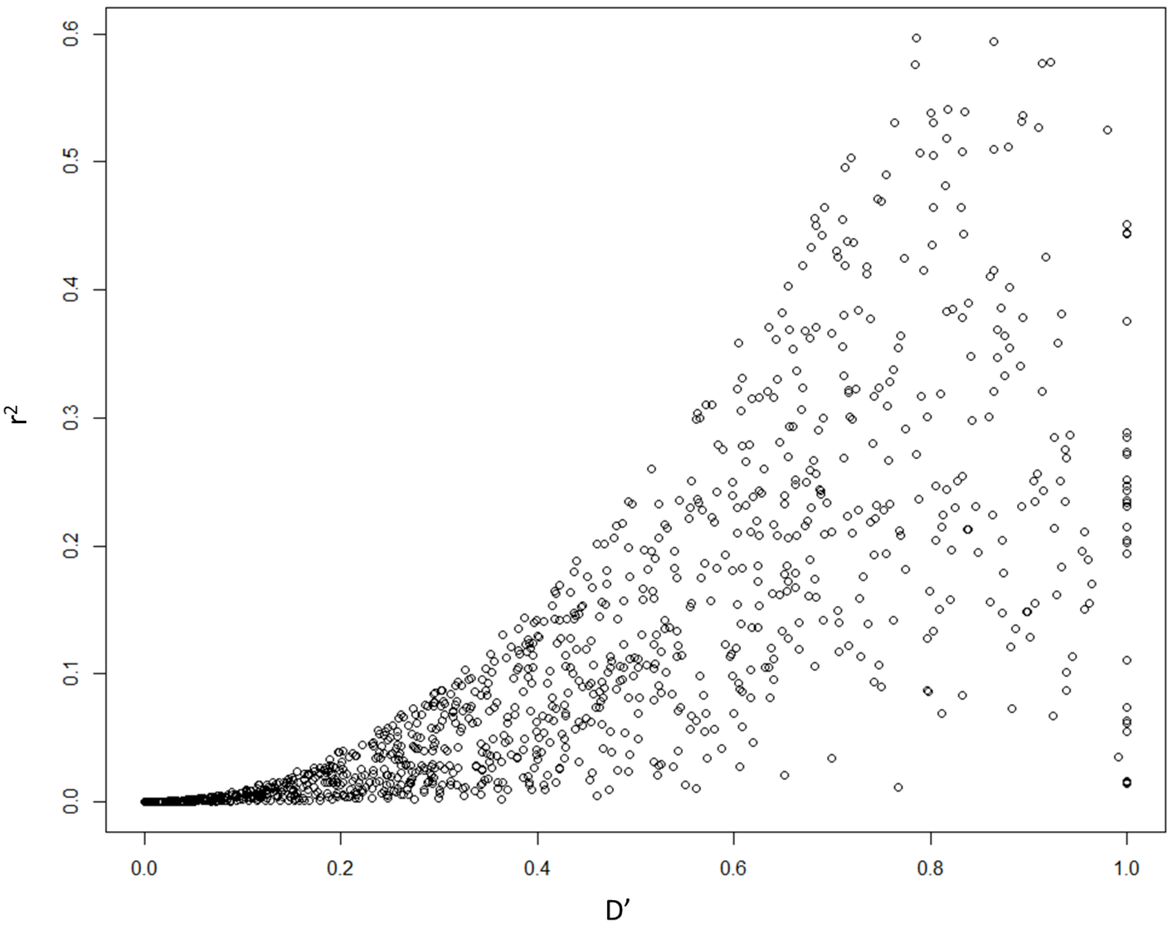


Figure 3. Linkage disequilibrium between interacting variants. I calculated LD between interacting variants using both r^2 and D' to determine if they were on the same haplotype. Interactions between variants in modest LD ($r^2 > 0.6$) had been removed from all stages of the analysis, and hence are not shown here.

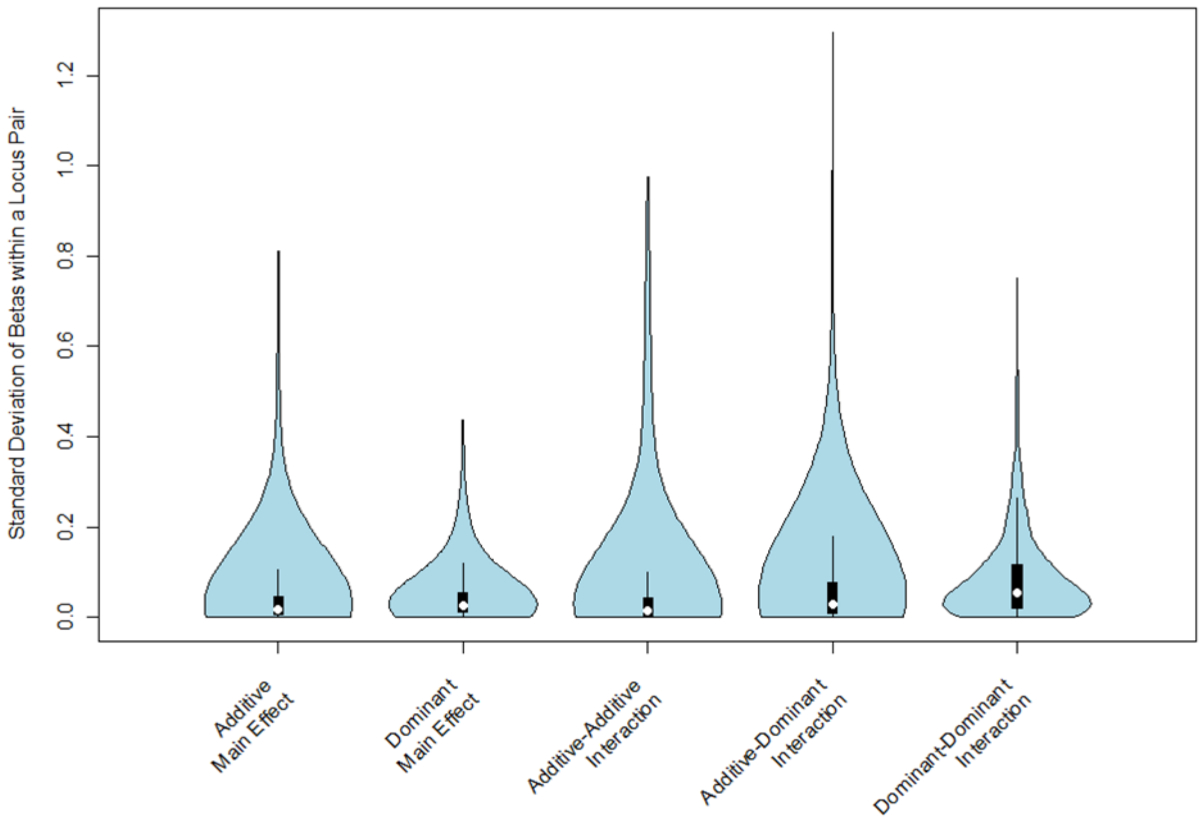


Figure 4. Redundant SNP-pairs have very similar parameter estimates. I grouped together all pairs of interacting SNPs (n=5,439) identified as being redundant through LD measures. For each group, I identified all terms that were significant in at least one of the associated interactions ($p < 0.05$). I extracted the βs for these significant terms from all interactions within the group. I then calculated the standard deviation of the βs for each significant term within each group to determine how similar the parameter estimates were across all interactions in the same group. The distribution of these standard deviations, categorized by type of variable, is shown above.

Next, I performed a replication analysis using an independent dataset of 232 unrelated individuals from the 1KG Project who had both whole-genome sequencing [118] data and gene expression levels in LCLs[54] available. All ieQTL composed of variants that were common (MAF > 5%) and had available genotyping data were tested for significant interactions with the same procedure used in the discovery analysis. Of the 803 ieQTL tested, 363 had p-values < 0.05 and 90 passed a Bonferroni multiple testing correction for all tests performed in the replication analysis.

Many factors confound interaction testing

Statistical interactions can be produced by a variety of factors other than biological epistasis, including technical artifacts, statistical artifacts, and LD artifacts driven by other biological processes. Technical artifacts are caused by the limitations of the data itself; for instance, limitations in the dynamic range of measureable gene expression can result in interactions being identified through the ceiling/floor effect. Statistical artifacts can result in an incorrect inference from a statistical model; for example, when there are population-level differences in the phenotype, analyzing multiple ethnicities together can produce spurious associations due to population stratification. Technical and statistical artifacts are especially troubling since they are unlikely to represent a real biological association between the loci and phenotype. Other biological phenomena, namely haplotype effects and *cis*-eQTL effects, can be captured by interaction analyses due to LD patterns. I investigated whether the observed 1,119 significant ieQTL models from the discovery analysis could be explained by each of these phenomena.

Some statistical interactions are consistent with confounding by technical limitations

The gene expression data used in this analysis was collected using microarrays. Microarray technology has a limited dynamic range, meaning that the upper and lower bound on the level of gene expression that microarrays can detect does not cover the full range observed in nature. If the combined effect of two variants behaving additively exceeds the detectable limit, their individual effects will not be fully captured as they hit the maximum (i.e., ceiling) or minimum (i.e., floor) value detectable by microarrays. This phenomenon, known as the ceiling/floor effect, may result in such pairs of variants being spuriously identified as epistasis.[125] Interactions caused by the ceiling/floor effect have a characteristic pattern of effects: the main effects of both variants have the same direction, and the interaction terms are in the opposite direction. For example, both main effects may increase gene expression, but the interactions will decrease gene expression. An example of an interaction putatively caused by the ceiling effect is shown in Figure 5. Of 1,119 locus pairs, 48 exhibited a pattern consistent with the ceiling/floor effect. It is possible that true genetic interactions could also produce this pattern; consequently, I consider this an upper bound of the influence of ceiling/floor artifacts within our analysis.

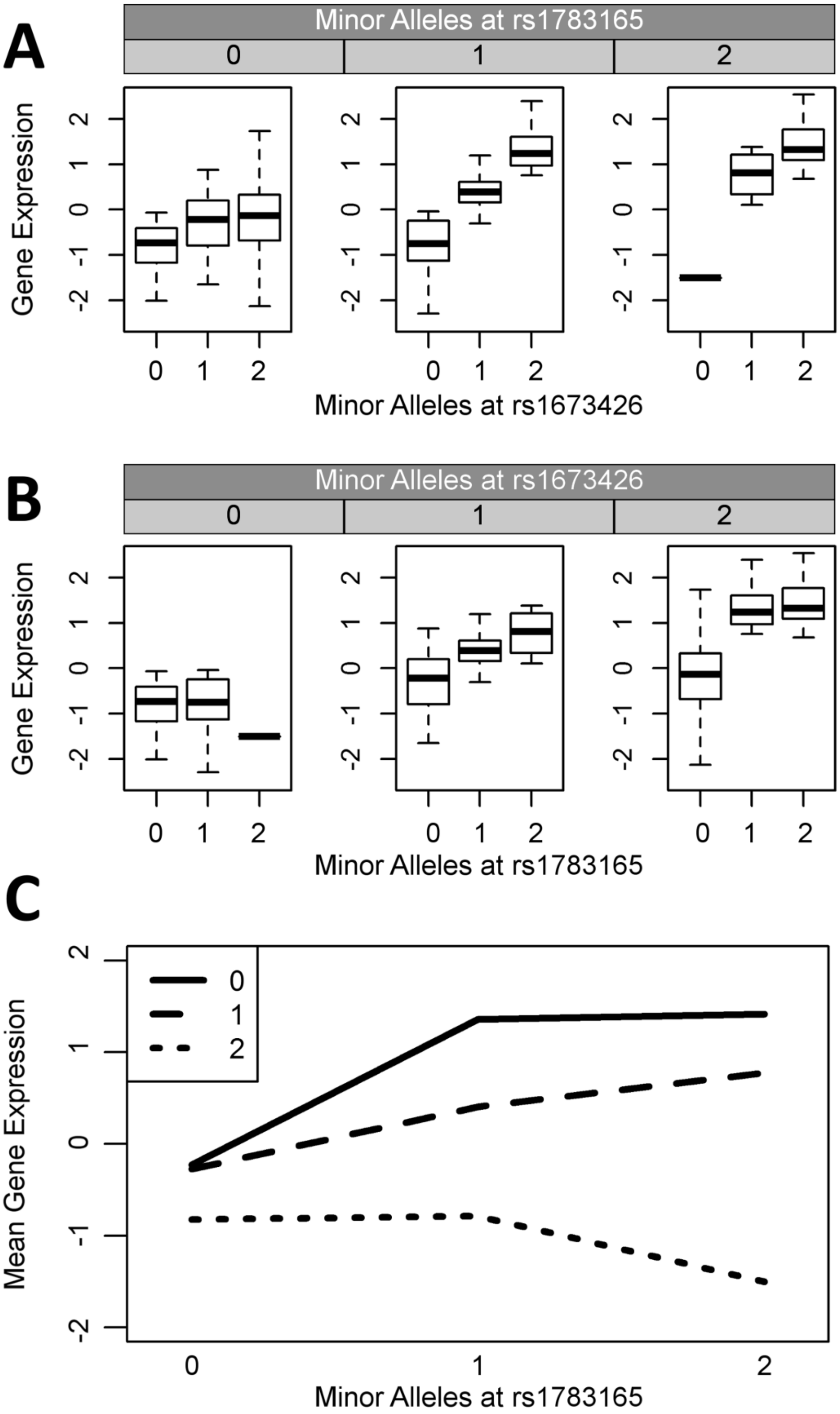


Figure 5. The interaction between rs1783165 and rs1673426 associated with the expression of *PKHD1L1* may be a ceiling effect. The ceiling effect, caused by limitations in the detectable range of gene expression, has a hallmark pattern – both variants have main effects with concordant direction of effect, and the interaction term has a discordant direction. (A) The minor

allele of rs1673426 increases the expression of *PKHD1L1*. (B) The minor allele of rs1783165 also increases the expression of *PKHD1L1*, meaning both variants have a concordant direction of effect. The interaction plot (C) depicts the mean gene expression for all individuals with the specified genotype combination, with each line representing the number of minor alleles at rs1673426. When there is only one minor allele at rs1673426, the mean gene expression increases for each minor allele at rs1783165; however, when there are two minor alleles at rs1673426, the increase in gene expression due to minor alleles at rs1783165 reaches a 'maximum' at one minor allele. There is no additional increase in expression for having two minor alleles at rs1783165. This is denoted by the flat line connecting the two genotype combinations. Given that each minor allele at rs1783165 increases gene expression on the background of one minor allele at rs1673426, and that the 'maximum' reached on the background of two minor alleles at rs1673426 is very close to the maximum gene expression levels possible to observe, I consider this an example of the ceiling effect.

The interpretation of microarray data is also complicated by genetic variants in the probe binding site, as different alleles may have different affinities for the probe. Probes containing any HapMap variant had previously been removed from the analysis;[53,117] however, HapMap does not provide comprehensive coverage of genetic variants. Consequently, I looked in a subset of individuals from the discovery analysis (n=174) with low-coverage sequencing data through the 1KG Project to see if genetic variants within the probe binding site may result in apparent interactions. The probes for 508 of 1,119 ieQTL contained a SNPs or indel in the 1KG Project. The probes for 255 ieQTL contained at least one common (MAF > 5%) variant. While the conditional analysis (Methods) performed later would likely account for the effect of these variants, I did not consider ieQTL with a common variant in the binding site evidence for biological epistasis. The probes for the remaining 253 ieQTL contained at least one rare variant, but no common variation. To determine if these rare variants could result in the interaction, I performed the interaction analysis using only the 1KG individuals who did not have a rare variant in the probe binding site. The interactions for 200 ieQTL remained nominally significant ($p < 0.05$) when all individuals with rare variants were removed. Consequently, the interactions for 811 ieQTL are not attributable to variants within the probe binding sites.

Missing genotype combinations may result in ieQTL

Linear regression models for epistasis may be unable to accurately decompose variance between genetic terms if there is either LD between the interacting variants or if there are missing genotype combinations. The issue of LD has previously been explored, and the Cordell model is robust to LD between variants when all genotype combinations are present.[89] Consequently, I examined all interactions within the discovery dataset to see if all of the nine possible two-locus genotype combinations were present. For 457 of the 1,119 ieQTL, at least one genotype combination was absent. While failure to see certain two-locus genotypes may be due to lethal combinations, and thus perhaps is evidence for epistasis, it may also simply be a result of certain combinations being uncommon due to allele frequencies and the proximity between variants. Either way, the statistical model used cannot provide robust estimates unless all genotype combinations are present, and therefore, I do not consider these interactions as evidence for biological epistasis.

Haplotype effects captured through complex LD patterns may produce ieQTL

In some LD architectures, a combination of two variants can identify haplotypes. While there is evidence to suggest haplotypes form in response to biological interactions between variants,[88,126] haplotypes may simply carry a single variant that additively regulates gene expression. Thus, interactions between two variants in LD with one another may simply be tagging a *cis*-eQTL. Wood et al. demonstrated that this could occur even when strict LD-pruning thresholds ($r^2 > 0.1$ or D' > 0.1) were used; therefore, I consider it unlikely that any LD-pruning threshold would be sufficient to eliminate confounding by *cis*-eQTL.[83] Consequently, I adopted a two-stage strategy to address haplotype effects, wherein I first use a lenient LD-threshold to filter out interactions and then directly tested whether the interaction can be accounted for by *cis*-eQTL.

In the first stage, I used LD-patterns to filter out variants in moderate LD with one another, as they likely represent a haplotype. I did this by first removing all interaction models composed of variants in modest LD with one another ($r^2 > 0.6$) from all portions of the study, as previously mentioned. I then investigated whether or not variants within the same interaction model were in modest LD with one another as assessed by D'; of the 1,119 interacting loci, 806 had D' values < 0.6. I did not consider any of the variants with D' thresholds exceeding this threshold as evidence for epistasis, as they likely carry a single variant driving the effect. An example of this phenomenon observed in our data is illustrated in Figure 6. The distribution of LD statistics, both r^2 and D', for interaction models is shown in Figure 3.

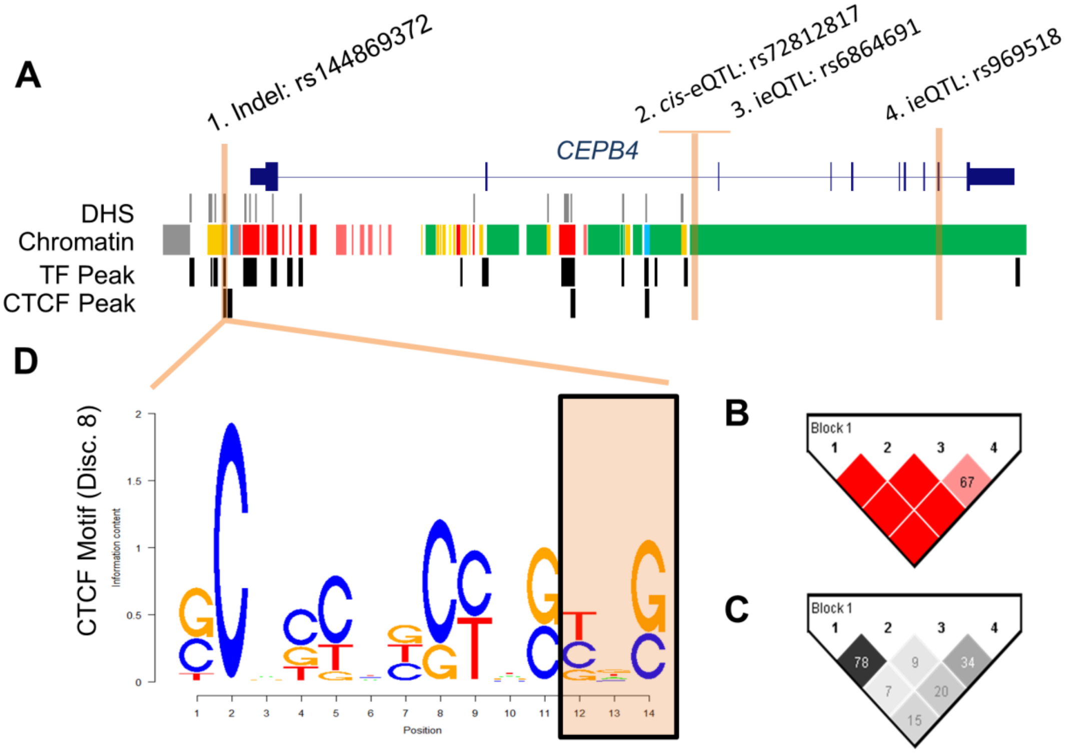


Figure 6. Interactions impacting the expression of *CPEB4* may represent haplotype effects. (A) A significant interaction between rs6864691 and rs969518 regulating the expression

of *CPEB4* was identified. The *cis*-eQTL rs72812817 mediated this interaction in the conditional analysis; however, none of these variants were within putative regulatory elements in GM12878 assayed by the ENCODE Project. However, a D' heatmap (B) of the region (the numbers correspond to SNP labels in A) illustrated that an indel, rs144869372, always occurred on the background of the *cis*-eQTL (D' = 1). This occurs despite modest r^2 values, as shown in the r^2 heatmap of the region (C). There is evidence from ENCODE (A) suggesting the indel may be functional, as it occurs within both a ChromHMM strong enhancer (yellow) and a CTCF binding peak in GM12878. (D) Notably, the indel is predicted to alter the binding of CTCF by HaploReg, by altering the last three nucleotides in the binding motif. Given the functional genomics evidence, the indel may be the causal variant and is detected by interactions that tag the haplotype carrying the indel.

In the second stage of the analysis, I directly tested whether or not the interaction could be accounted for by cis-eQTL by conditioning the interaction on each of the target gene's *cis*-eQTL in turn. I first identified all nominal, common *cis*-eQTL ($p < 0.05$) for the interaction's regulated gene using a subset of individuals from our discovery dataset (n=174) with sequencing data available through the 1KG Project so that I would have a comprehensive list of genetic variation. While the 1KG sequencing data is low coverage, it is extremely unlikely I would fail to detect the effect of a common *cis*-eQTL – 1KG estimates they had 99.3% power to detect variants of 1% frequency.[118] Even if a common *cis*-eQTL was missed, all variants that could tag it through LD would additionally have to be absent for its effect to not be captured in the conditional analysis. I then created all pairs of *cis*-eQTL and ieQTL for the same gene. For each of these combinations, I performed a conditional analysis in which the additive and dominant main effect for the *cis*-eQTL were incorporated into both the full and reduced model used in the LRT to determine the significance of the interaction. The majority of interactions appeared to be mediated by *cis*-eQTL (Figure 7); however, 139 of the 965 testable ieQTL remained significant ($p < 0.05$) in all conditional analyses performed, indicating that these interactions are not explained by *cis*-eQTL.

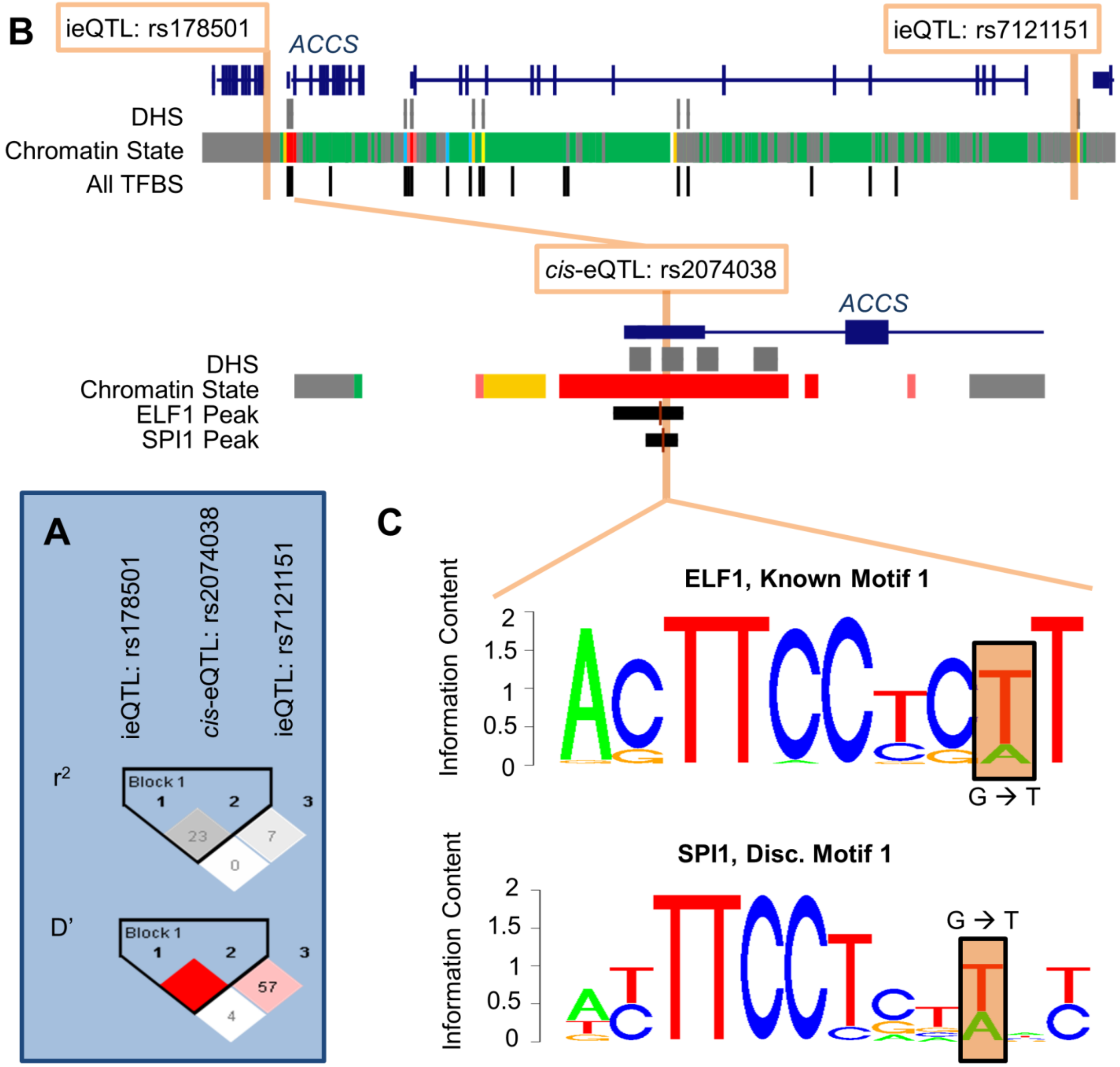


Figure 7. The interacting SNPs regulating *ACCS* are likely tagging a single-variant *cis*-eQTL through linkage disequilibrium. The interaction between rs178501 and rs7121151 is mediated by the *cis*-eQTL rs2074038 in the conditional analysis (interaction p-value > 0.05). (A) While the interacting variants are in low LD with the *cis*-eQTL based on r^2, their high D' indicates they often occur on the same haplotype. (B) The interacting variants are not located within DNase hypersensitivity sites, predicted chromatin states with a regulatory function (GM12878 Combined), or any of the uniform binding peaks identified for all transcription factors tested in GM12878 by ENCODE; however, the *cis*-eQTL is located within the canonical promoter for *ACCS*, a DNase hypersensitivity site, and numerous transcription factor binding peaks identified in GM12878 by ENCODE. (C) Notably, the *cis*-eQTL occurs within a binding peak for both ELF1 and SPI1 in GM12878, and also alters the binding motifs of these

transcription factors at the position highlighted in orange. Thus, the *cis*-eQTL rs2074038 is likely the causal variant, and the interaction is simply capturing its effect through LD.

Population specific eQTLs may produce statistical interactions

In our discovery and replication analyses I analyzed multiple ethnicities together. When there are population differences in both the distribution of genotypes and phenotypes, analyzing multiple populations together can lead to spurious results, due to a phenomenon known as population stratification. The population normalization procedure applied to the gene expression data removes systematic population differences in the phenotype, thereby enabling multiple ethnicities to be combined for analysis without risk of known complications from population stratification. While this approach has been used in other studies, I also controlled for the top three PCs in our analysis to adjust for residual ethnicity-dependent effects.[117,127] Furthermore, I performed a stratified analysis, wherein I tested each of the 1,119 ieQTL in each of the three discovery ethnicities (CEU, YRI, and CHB+JPT) separately. While the Cordell model was not robust in the stratified analysis in many cases (due to the reduced sample size, all nine possible two-locus genotype combinations were often not observed in all populations), 859 of 1,119 ieQTL were at least nominally significant ($p < 0.05$) in at least one population, suggesting that population stratification is unlikely to account for their significance.

However, the interaction for 260 ieQTL was completely attenuated in the stratified analysis. In some cases, this may be attributed to reduced power to detect effects as the sample size is smaller; however, it could also suggest that interaction testing was subject to a novel form of population stratification. Upon further investigation, I found that 234 of 260 ieQTL attenuated in the stratified analysis involved at least one population-specific *cis*-eQTL, meaning that a variant was only a significant *cis*-eQTL in a subset of populations. Population-specific *cis*-eQTL may be a product of reduced power to detect effects when allele frequencies are different between populations; however, there were also instances in which variants with very similar allele frequencies had different marginal effects across populations (Figure 8).[54] Such variants might be a product of population-dependent ability to tag causal *cis*-eQTL due to differential LD patterns. In relation to interaction testing, systematic differences in both the main effect of each variant and the frequency of two-locus genotype combinations between populations resulted in a spurious interaction signature; an example is provided in Figure 9. To investigate whether population-specific effects may impact the 859 ieQTL that were nominally significant in at least one population, I calculated the within-population LD between each pair of interacting variants. 689 of 859 ieQTL were significant in at least one population where the variants were not in LD with one another (r^2 and D' < 0.6) (Supplemental Table 3, provided by Fish et al.[124]). I did not consider the 170 ieQTL that were exclusively significant in populations with population-specific haplotypes as clear evidence for biological epistasis. Ultimately, 689 of the 1,119 ieQTL were inconsistent with population-specific effects.

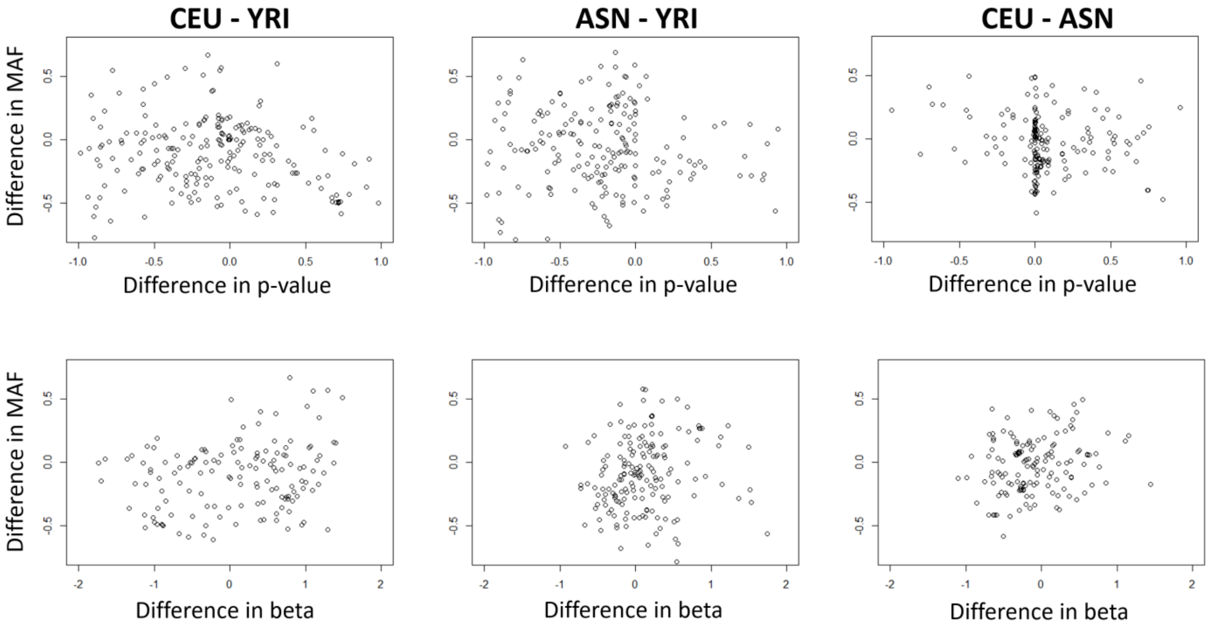


Figure 8. Investigation of population-specific *cis*-eQTL. To investigate whether or not population-specific *cis*-eQTL were caused by reduced power to detect significant marginal effects in the stratified analysis, or by different marginal effects for the same variant, I performed pairwise comparisons of MAF, additive β (marginal), and p-value (of the *cis*-eQTL) by ethnicity.

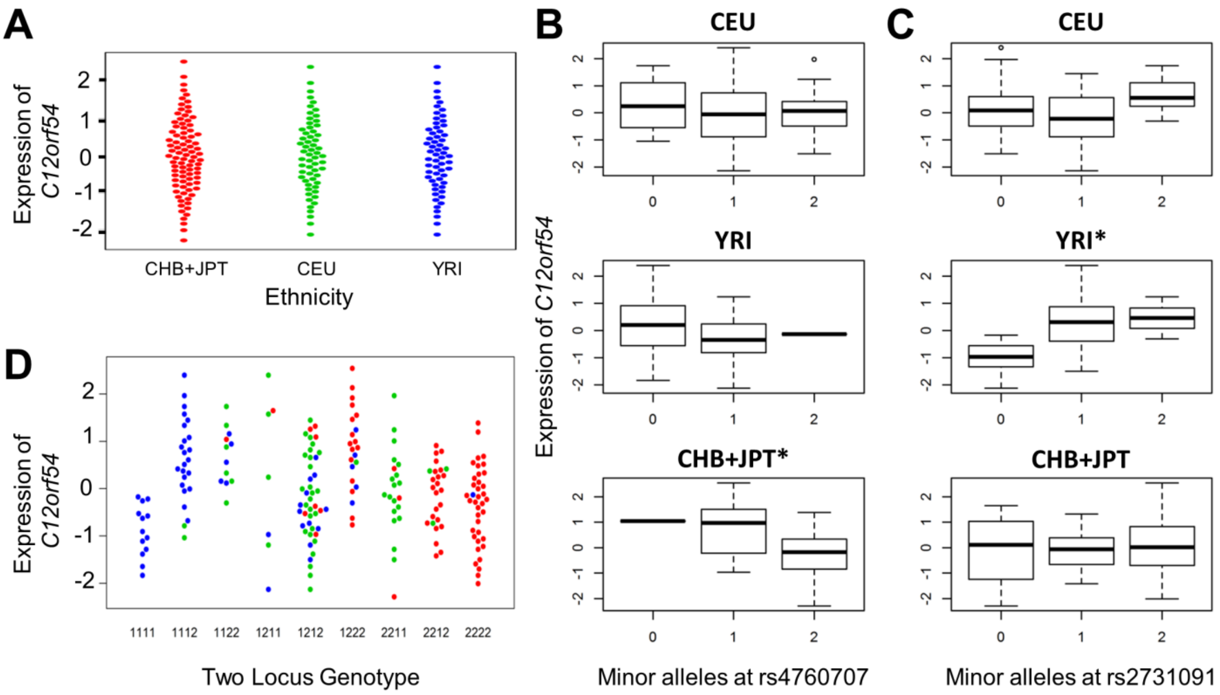


Figure 9. Population specific eQTLs may underlie ieQTL regulating *C12orf54*. The interaction between rs2731091 and rs4760707 regulating *C12orf54* replicated, but was not

nominally significant ($p < 0.05$) in any population in the stratified analysis. (A) Due to the population normalization procedure, there are not systematic differences in the expression of *C12orf54* between populations; however, I found that each variant was a population-specific *cis*-eQTL. (B) rs4760707 was a *cis*-eQTL in CHB+JPT (p=7.25x10^{-6}), but not in YRI (p=0.17) or CEU (p=0.96). (C) rs2731091 significantly regulated gene expression as a *cis*-eQTL in YRI (p = 7.28x10^{-6}), but not CEU (p = 0.14) or CHB+JPT (p=0.84). (D) There were clear population differences in the frequency of two-locus genotypes between populations; in combination, it appears the population differences in two-locus genotypes and population specific *cis*-eQTL produced a nuanced form of population stratification.

IeQTL can be entirely accounted for by alternative mechanisms

Ultimately, I investigated whether confounding factors could cumulatively account for all the interactions identified in this analysis (Supplemental Table 3[124]; Table 3). Of the 1,119 interacting genomic loci identified, 90 significantly replicated using a Bonferroni multiple testing correction threshold. Of these, 26 ieQTL could be explained by technical artifacts (i.e., the ceiling/floor effect and/or variants within the probe binding sites). 50 of the remaining 64 ieQTL could be explained by statistical artifacts (i.e., population stratification and/or missing genotypes). Biological explanations other than epistasis – namely haplotype effects or the tagging of *cis*-eQTL – could account for all remaining ieQTL that replicated at the most stringent Bonferroni level.

Confounder	**All Interactions (n=1,119)**		**Bonferroni Replicating Interactions (n=90)**	
	Total	(%)	Total	(%)
Ceiling/Floor Effect	48	(4.30)	11	(12.22)
Variants in Probe	308	(27.52)	15	(16.68)
Cis-eQTL	980	(87.58)	78	(86.68)
D’ Haplotype	313	(27.97)	43	(47.78)
Population-Specific Effects	430	(38.43)	58	(64.44)
Missing Genotypes	457	(40.84)	37	(41.11)

Table 3. Proportion of Interactions Consistent with Confounding Factors. I counted the number of interactions consistent with each alternative explanation; interactions can be consistent with multiple confounders. I considered two categories of interactions: all interactions identified (n=1,119), and the subset of those that replicated with p-values exceeding the Bonferroni multiple testing correction threshold for the entire replication analysis (n=90).

We additionally investigated the impact of filtering out interactions consistent with confounding prior to the replication analysis. Removing these interactions prior to replication testing had a considerable influence on the multiple testing correction threshold: only 86 of the

1,119 interactions identified in the discovery analysis were not consistent with the ceiling/floor effect, population stratification, variants within the probe binding site, missing genotype combinations, haplotype effects, or the tagging of *cis*-eQTL (Supplemental Table 4, provided in Fish et al.[124]). 37 of the 86 ieQTL had sufficient data to be tested in the replication analysis, and while none replicated at the adjusted Bonferroni multiple testing correction threshold, two interactions did replicate with nominal significance ($p < 0.05$). One of these, the interaction between rs1549791 and rs7115749 to regulate *APIP*, did not have a consistent direction of effect between the discovery and replication datasets (Figure 10), and thus was not considered evidence for epistasis. The remaining interaction, between rs1262808 and rs11615099 regulating the expression of *MYRFL*, had concordant effects in both the discovery and replication datasets (Figure 11). As it did not pass the multiple testing correction threshold in the initial replication analysis ($p=2.03\text{x}10^{-3}$) though, I further examined it additional datasets.

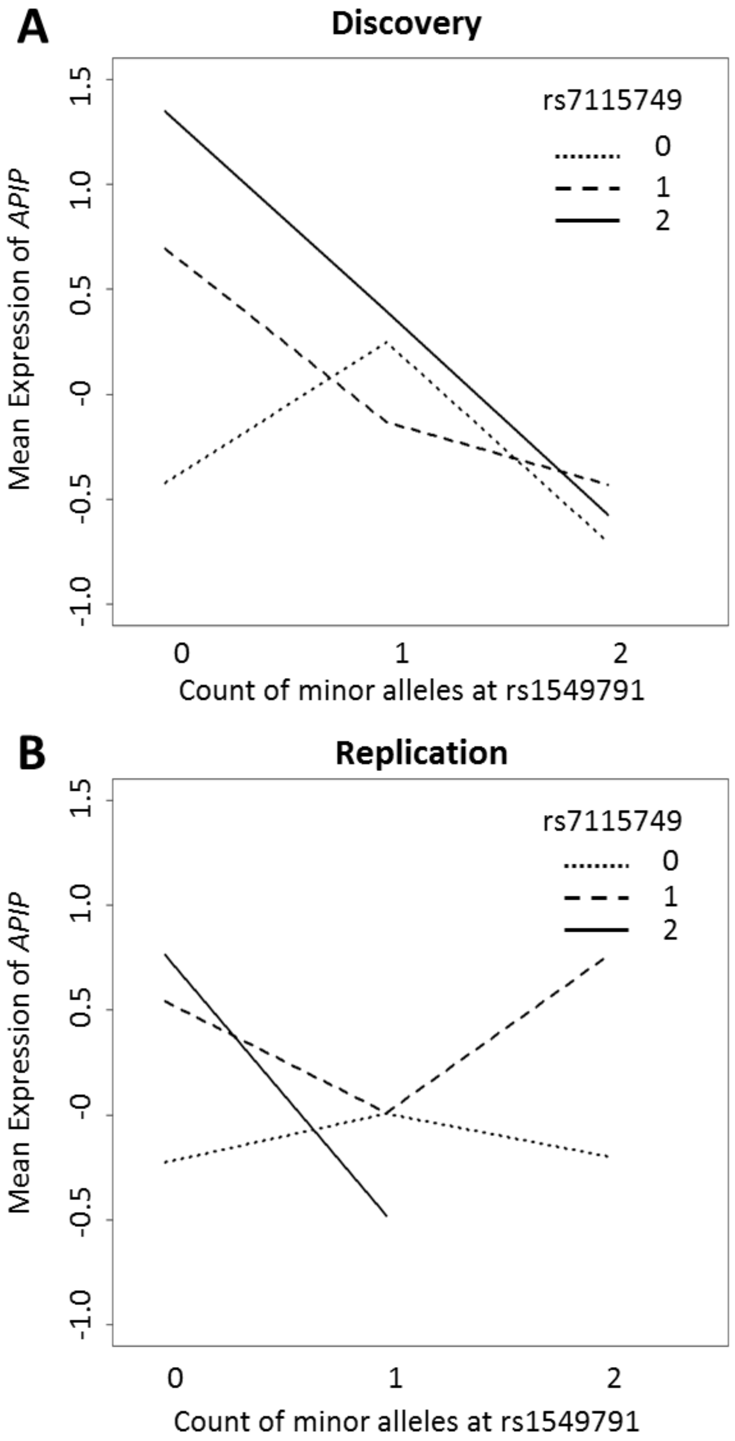


Figure 10. The interaction between rs1549791 and rs7115749 associated with the expression of *APIP* is not consistent between the discovery and replication datasets. In the interaction plot, each individual is categorized according to their two-locus genotype at rs1549791 and rs7115749. This results in nine possible genotype combinations, and the mean expression of *APIP* for each combination is shown here for the (A) discovery and (B) replication datasets.

There are markedly different patterns in gene expression by two-locus genotype between the two datasets, illustrating the putative interaction does not replicate with a consistent direction of effect.

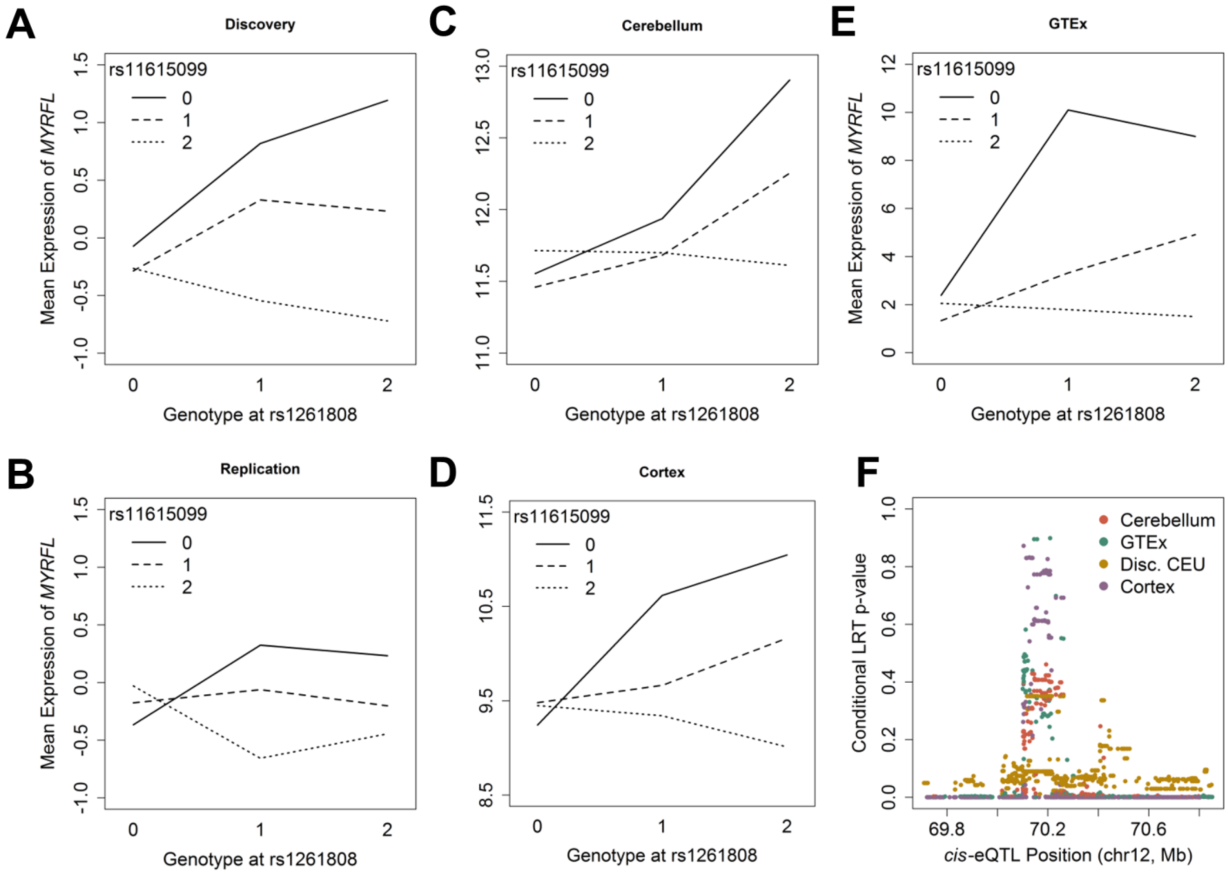


Figure 11. Despite consistent replication, the interaction regulating *MYRFL* is attributable to *cis*-eQTL. In each interaction plot, all individuals are categorized according to their two-locus genotype at rs1262808 and rs11615099. The mean expression of *MYRFL* for all individuals with each of the nine possible two-locus genotypes is shown here for the (A) discovery; (B) replication; (C) Mayo, cerebellum; (D) Mayo, cortex; (E) GTEx, whole blood datasets. The interaction plot illustrates a consistent trend across all datasets, this interaction is mediated by *cis*-eQTL. (F) Conditional *cis*-eQTL analyses were conducted in the discovery (CEU only, yellow); GTEx (purple); Mayo, cerebellum (teal); and Mayo, temporal cortex (orange). For each conditional analysis, the conditional LRT p-value is plotted by the genomic position of the *cis*-eQTL conditioned on. The p-value peak observed in this region illustrates that *cis*-eQTL completely attenuate the interaction when they are conditioned on.

Discussion

In this study, I analyzed more than 21 million pairs of *cis*-regulatory variants for epistatic interactions influencing gene expression, and found limited evidence for epistasis within the *cis*-regulatory region of genes. Fewer than 2% of genes tested (165 of 11,465) had significant interactions between regulatory genetic variants that appeared to influence their expression in the tightly controlled context of LCLs. Nonetheless, 90 of the 1,119 significant interactions replicated in independent datasets. I then performed a comprehensive investigation of known and novel potential confounding factors on the identified interactions (haplotype effects, ceiling/floor effect, single variant eQTL tagged through LD, missing genotype combinations, population stratification, and others), and found that all the interactions – even those that replicated – could be explained by at least one technical, statistical, or biological confounder. Thus, our findings do not support a major role for large effect interactions between common variants within the *cis*-regulatory region influencing the regulation of gene expression in LCLs.

Additionally, this study provides a trait-independent framework for protecting future interaction studies from confounding. Prior to performing any association testing, there are two levels of quality control required for statistical studies of epistasis: those adopted in GWAS best practices[128–131], which are aimed at ensuring individual genetic variants are called with high accuracy, and then those that check whether a given pair of genetic variants is appropriate for interaction testing (i.e. missing genotype and the within-pair LD filters). Even when these quality control measures are applied prior to the discovery analysis, significant interactions need to be further examined for evidence of confounding by single variants tagged through LD and for population-specific effects. I advise removing interactions consistent with these confounders prior to replication, as this reduced the number of putative interactions carried forward substantially, and consequently, the multiple testing penalty. The ceiling/floor effect is a more complicated confounder, as it is difficult to statistically disambiguate whether consistent interactions are caused by technical limitations or by biological epistasis. Consequently, I recommend interactions consistent with the ceiling/floor effect be flagged, rather than filtered out, and validated with an alternative technology if possible. It is still critical to replicate interactions to ensure they have robust, consistent effects, despite replication being insufficient to protect against confounding. Given how pervasive confounding factors are, it is critical to explicitly account for them through additional quality control procedures and post-hoc analyses in future studies to reduce spurious results.

To strike a balance between maximizing the power to detect effects and thoroughly investigating potentially interacting loci, I performed a focused analysis of common variants with significant marginal effects in the *cis*-regulatory region, which harbors the majority of known regulatory elements. I was moderately powered to detect interactions between common variants in low LD with one another with effects commensurate with the single-locus eQTL found in this dataset. While additional statistical interactions with either smaller effect sizes or between less frequent genotype combinations would likely be identified with increased power, every example of a significant interaction I did identify was consistent with at least one confounding factor. Thus, I did not find compelling evidence that *cis*-regulatory interactions contribute strongly to the genetic architecture of gene expression; however, there are several additional limitations to our study. First, cell lines are a model system, and thus are not perfectly representative of primary tissue. Second, I analyzed multiple ethnicities simultaneously in an effort to increase sample size; however, doing so also increased the heterogeneity of our sample, which may have obfuscated some interactions. Therefore, our findings do not preclude the existence of epistasis within the *cis*-regulatory region, and I recommend that future studies of regulatory epistasis consider potential interactions that: 1) occur within haplotypes (consistent with reports from Corradin et al.[126] and Lappalainen et al.[88]), 2) have smaller effect sizes than those detected in similarly powered single-locus eQTL studies, 3) occur among less frequent genotype combinations, including rare variants 4) involve variants without marginal eQTL effects (though evidence in model organisms suggests these are rare[1]), and/or 5) are context-dependent (e.g. inducible eQTL effects). Observing statistical interactions in these contexts could reconcile our findings with molecular studies, many of which use mutagenesis to generate genetic variation that would not be observed in population-based studies, that illustrate that transcription factors (TF) interact with each other to influence promoter and enhancer activity.[77,78,132]

Genetic interactions involving distant variants could also be a mechanism through which epistasis influences complex traits. However, I did not investigate interactions involving variants outside of the *cis*-regulatory region because evidence from eQTL studies in humans suggests that *trans*-eQTL effects are less robust, less common, and have smaller effect sizes.[133,134] This, coupled with the substantial increases in the number of association tests required to investigate *trans*- interactions, would have resulted in reduced power to detect such effects. Nonetheless, interactions between distant variants (i.e., gene by gene interactions) may still be important to the biology of disease in humans. Increases in the sample size of eQTL datasets and the corresponding increases in statistical power will enable future in-depth studies of *trans*-interactions that may help to illuminate the biological mechanisms through which genetic variants are associated with disease. However, *trans*- interactions are not protected from many of the confounders influencing the study of *cis*- interactions,[83] and thus studies of *trans*- interactions will need to explicitly account for these issues as well.

Our findings (along with prior reports)[83] illustrate that significant interaction effects can be due to a variety of confounding factors. This demonstrates that significant statistical interactions do not necessarily imply either a biological relationship with the phenotype, or between the variants themselves. To account for this, some confounders can be addressed as part of quality control procedures prior to performing any association tests (i.e., missing genotype check, removing variants in probe binding sites, and LD-filtering), while others – such as confounding by single variants with strong effects – will likely require specific post-hoc analyses after the initial association is identified. Furthermore, replication – long held as the gold standard for genetic association studies – does not safeguard against these confounders, as they can be due to artifacts that are consistent across multiple datasets. Given the pervasive nature of confounding, it must be considered in all future studies of epistasis. The analytic approach used in this study provides a trait-independent framework for explicitly examining confounding factors in interaction studies and avoiding reporting spurious results.

CHAPTER 3
EPISTASIS IN ADMIXED POPULATIONS

Introduction

Regulatory epistasis may occur between variants on the same haplotype, which are combinations of physically linked genetic variants that co-occur more often than anticipated. As reviewed in detail in Chapter 1, evidence from reporter assays clearly demonstrating epistasis between variants typically investigate a narrow genomic region, unlikely to be broken apart by recombination in natural populations.[78] Additionally, follow-up of GWAS-Catalog variants has demonstrated that multiple variants on the associated haplotype influence the expression of the same target gene in the rare instances when those variants are separated via recombination.[87] Finally, genetic variants that are associated with decreased gene expression levels are associated with an increased burden of recently-derived rare variants.[88,135] However, the study of epistasis within haplotypes is complicated by the same properties that make it biologically intriguing; the tight linkage between genetic variants on the same haplotype inherently means that all combinations of variants are either rarely observed, or absent. Consequently, linear regression models are unable to accurately partition phenotypic variance to genetic components, thereby complicating the study of epistasis within haplotypes.[89]

The structure of haplotypes is in large part dictated by the location of recombination hotspots. Indeed, the genome can be divided into blocks of variants in high LD with one another, separated by regions that frequently undergo recombination events.[136–138] These boundaries are highly correlated between ethnicities; however, there are some discrepancies.[136,139] First, LD-blocks are typically shorter in African-descent populations; this is consistent with the Out-of-Africa hypothesis, as both population bottlenecks increase the length of LD-blocks in European and Asian-descent populations, and older populations (i.e., African) have had more recombination events that reduce the length of LD-blocks.[139,140] Second, recombination rates also vary considerably by population. The location of recombination hotspots is regulated by *PRDM9*, a methyltransferase with a zinc finger domain that recognizes specific sequence motifs.[141–143] In European descent populations, there are two common alleles of *PRDM9*, A and B, which occur with a frequency of 90% and 5% respectively.[141] Individuals with rarer alleles that no longer recognize the canonical binding motif have drastically shifted landscapes of recombination hotspots.[141] In African descent populations, there is third allele, C, which occurs with a frequency of ~35%.[144,145] Individuals carrying the C allele of *PRDM9* do not appear to share any of the recombination hotspots recognized by the A allele, and vice versa.[144] This has resulted in drastically different recombination landscapes between European and African descent populations; there are more than two thousand recombination hotspots that are observed in populations of West African descent, but are absent in European descent populations.[91] Thus, there are population-level differences in haplotype structure that could potentially be exploited to investigate epistasis within haplotypes.

Due to population-level differences in recombination hotspots, admixed populations provide a unique opportunity to investigate epistasis within haplotypes. I hypothesize that African-specific recombination hotspots may disrupt European haplotypes, and vice versa. When ancestral haplotypes are broken apart by these recombination events, there is an increased likelihood of observing all possible genotype combinations, such that epistasis within the ancestral haplotype can be investigated using traditional linear regression approaches discussed in Chapter 2.

In this Chapter, I investigate this hypothesis in African Americans. African Americans are a population derived from a two-way admixture event between European-descent and African-descent individuals. Historical records indicate this admixture event began with the Trans-Atlantic slave trade, in which approximately 11 million individuals were forcibly brought form coastal regions of Africa to the Americas throughout the 15th to 19th centuries.[146,147] Current estimates predict that, on average, six to seven generations have passed since the initial admixture event.[148,149] On average, African Americans have approximately 20% ancestry from European descent populations with the remainder of African descent, although these proportions can vary substantially between individuals.[148–150] Various methods exist that predict local ancestry at specific genomic loci by comparing the observed, inferred haplotypes to those seen in reference populations of African and European descent.[92,93,95]

In this Chapter, I leverage transitions in local ancestry between European and African descent to identify genomic regions in which ancestral haplotypes might be disrupted. I investigated this in 9,559 African American adults with EHR linked to genetic data collected on the Metabochip, a custom genotyping array.[97,151,152] The Metabochip is designed to fine-map approximately two hundred genomic loci previously associated to with type 2 diabetes, obesity, and coronary artery disease, and corresponding endophenotypes. The wealth of phenotypic data within the EHR allowed me to explore whether these transitions in local ancestry interacted with nearby genetic variants to influence three categories of phenotypes: those the Metabochip was designed around; phenotypes with associations in the GWAS-Catalog in Metabochip regions; and finally, all possible phenotypes (i.e., a PheWAS). Thus, I was able to investigate a wide array of possible phenotypic consequences resulting from the disruption of ancestral haplotypes within Metabochip regions.

Subjects and methods

Subjects and genotyping

In this Chapter, I investigated whether local ancestry transitions interacted with genetic variants to influence a variety of EHR-derived phenotypes in 9,559 African-American adults. All samples used in this analysis were part of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study, which used Vanderbilt University's de-identified biorepository to link patient EHR data with their genetic data.[152] EAGLE selected individuals for inclusion based upon minority-status, rather than for specific health phenotypes. Consequently, this is a cross-sectional study design that minimizes ascertainment bias. This sample was comprised of 6,249 females, and 3,310 males. The mean age was 46.9 years, with a standard deviation of 16.4 years. The mean BMI was 28.9, with a standard deviation of 6.56. All individuals self-reported as African American. We inferred global ancestry from our local ancestry estimates, by determining the proportion of variants with European ancestry. In this sample, we observed a mean European ancestry of 21.5%, with a standard deviation of 14.5%, consistent with the average proportion of European ancestry reported within the literature.

The samples were genotyped on Illumina's Metabochip, a custom array of almost 200,000 SNPs that targets genomic regions previously associated with type 2 diabetes, obesity, and coronary artery disease for fine-mapping purposes.[151] As part of quality control, variants were removed that did not have at least a 95% genotyping efficiency rate, or that did not vary in this dataset, leaving a total of 192,093 variants for analysis.

Determining local ancestry

Determining local ancestry is a two-step process: first, individual chromosomes are phased, and then the local ancestry is assigned. I phased the data using the program SHAPEITv2[96] and the 1000 Genomes Phase 3 reference panel (available for download at https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference). 171,439 variants were successfully phased; when variants failed it was typically due to inconsistencies with the reference panel. I then used RFMixv1.5.4 to determine the local ancestry of the phased genetic data, using a window size of 0.1 cM, and a minimum node size of 5. For the phasing reference panel, I used all YRI and CEU individuals from 1000G, phase 3v5a. I developed a custom pipeline to perform all necessary data processing and file-type conversions between these programs, which will be made publicly available.

Phenotype processing and quality control

Typically, individuals have multiple measures for body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP), and low-density lipoprotein (LDL) levels. To assign a single score to each individual, I took distinct approaches depending on the phenotype. For BMI, I computed the median measurement for each year with data available in the EHR, and then computed the median of these scores. For SBP and DBP, I had three distinct scores: the first measurement, regardless of medication status; a pre-medication measurement, which was the median of yearly medians for measurements made prior to any references to blood pressure medications; and a post-medication measurement, which was the median of yearly medians for measurements made after a reference to blood pressure medications. Similarly for LDL, three distinct measurements were used for analysis: the median of yearly medians, regardless of medication status; a pre-medication measurement, which was the median of yearly medians for measurements made prior to any references to lipid medications; and a post-medication measurement, which was the median of yearly medians for measurements made after a reference to lipid medications. For all of the above phenotypes, I then performed the following quality controls: I removed clearly non-valid scores (i.e., scores of zero or one), and then removed outliers (those scores more than three standard deviations away from the mean). The distributions of each phenotype are provided in Figure 12.

Statistical modeling

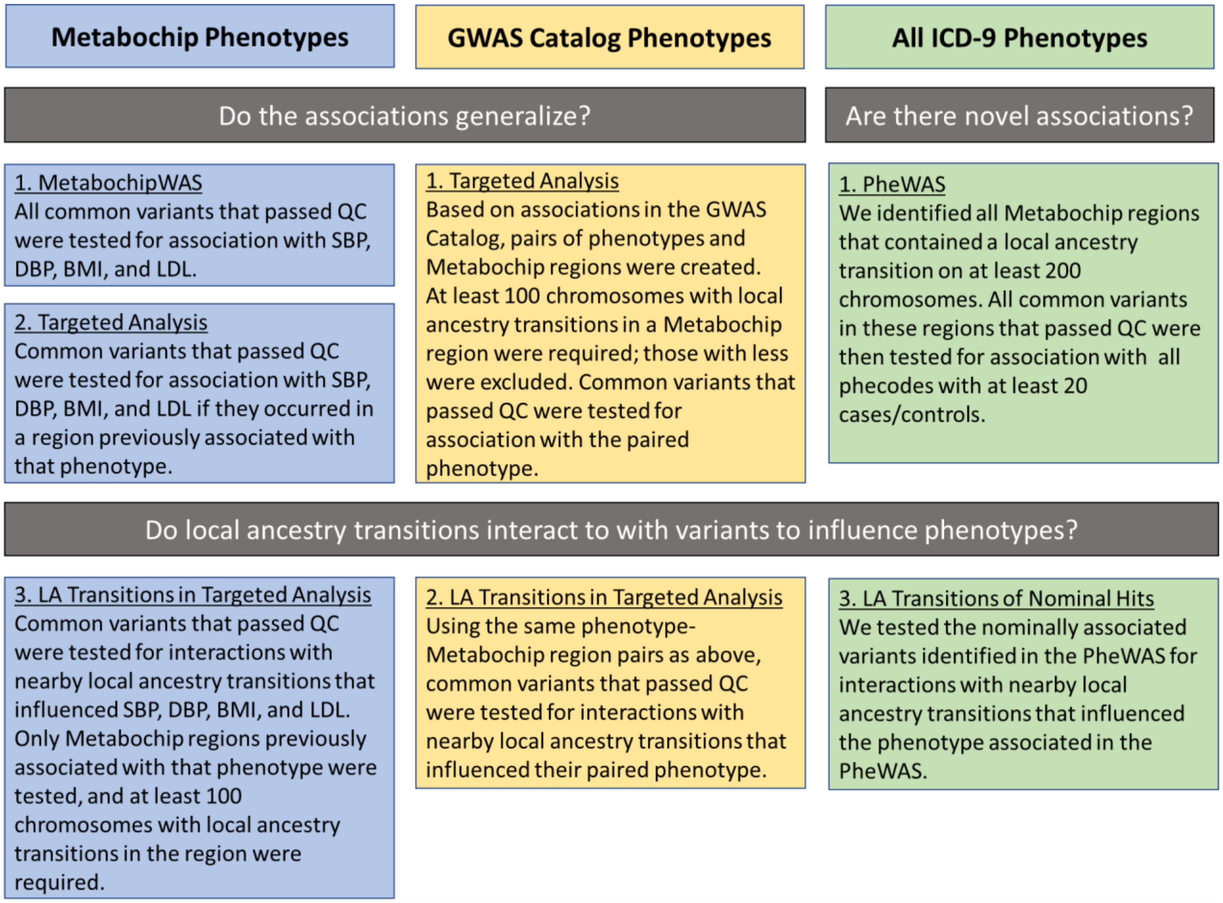
Given that local ancestry is specific to a given chromosome, I performed all analyses on the level of the chromosome, rather than the individual. I used linear regression to determine whether local ancestry transitions interacted with the allele to influence the phenotypes of interest, using the model:

$$y = A + LA + TRANS + A * TRANS + PC_{1-3} + AGE + GENDER + BMI$$

where y is the phenotype of interest; A corresponds to the allele status (binary variable: 0, absence of the allele; 1, presence of the allele); LA corresponds to the local ancestry at the variant (binary variable: 0, African ancestry; 1, European ancestry); TRANS indicates the presence of a local ancestry transition within the Metabochip region (binary variable: 0, no transition; 1, no transition); A*TRANS represents the interaction term between the allele and local ancestry transition (binary variable: 1 indicates presence of both the allele and a local ancestry transition; 0 encompasses all other possibilities). The covariates included varied by the analysis. AGE, GENDER, BMI, and the top three principal components (PC_{1-3}) were included in the investigation of phenotypes the Metabochip was designed to fine-map (BMI was not a covariate when it was the phenotype being investigated). AGE, GENDER, and the top three principal components (PC_{1-3}) were included in the investigation of phenotypes derived from the GWAS-catalog with associations in Metabochip regions. No covariates were included in the PheWAS. In the case of binary phenotypes, logistic regression was used.

Results

In this Chapter, I investigate whether transitions in local ancestry interacted with genetic variants to influence three distinct categories of EHR-derived phenotypes in an admixed population. All of these analyses require local ancestry, and the identification of genomic regions harboring transitions in local ancestry. To derive local ancestry, I first phased the genetic data using SHAPEIT2[96], then assigned local ancestry using RFMix[92]. I was able to pinpoint the location of local ancestry transitions, e.g. where changes in continental ancestry occurred along the chromosome, using the local ancestry calls. For each phenotype, I then analyzed the significance of single variants, regardless of local ancestry, to determine if there was genetic association within a region. Then, I performed an interaction analysis to determine if local ancestry transitions interacted with variants to influence the phenotype. I examine three separate categories of phenotypes: those that the Metabochip was designed to fine-map; those with associations in Metabochip regions in the GWAS Catalog for African-descent populations; and finally, all phenotypes encoded by ICD-9 with sufficient numbers of cases and controls. The first two categories are targeted analyses based on previous associations, whereas the final category – a PheWAS – is designed to discover novel associations. A table of the analyses performed is provided as a guide in Figure 12.



Metabochip Phenotypes

GWAS Catalog Phenotypes

All ICD-9 Phenotypes

Do the associations generalize?

Are there novel associations?

1. MetabochipWAS
All common variants that passed QC were tested for association with SBP, DBP, BMI, and LDL.

2. Targeted Analysis
Common variants that passed QC were tested for association with SBP, DBP, BMI, and LDL if they occurred in a region previously associated with that phenotype.

1. Targeted Analysis
Based on associations in the GWAS Catalog, pairs of phenotypes and Metabochip regions were created. At least 100 chromosomes with local ancestry transitions in a Metabochip region were required; those with less were excluded. Common variants that passed QC were tested for association with the paired phenotype.

1. PheWAS
We identified all Metabochip regions that contained a local ancestry transition on at least 200 chromosomes. All common variants in these regions that passed QC were then tested for association with all phecodes with at least 20 cases/controls.

Do local ancestry transitions interact to with variants to influence phenotypes?

3. LA Transitions in Targeted Analysis
Common variants that passed QC were tested for interactions with nearby local ancestry transitions that influenced SBP, DBP, BMI, and LDL. Only Metabochip regions previously associated with that phenotype were tested, and at least 100 chromosomes with local ancestry transitions in the region were required.

2. LA Transitions in Targeted Analysis
Using the same phenotype-Metabochip region pairs as above, common variants that passed QC were tested for interactions with nearby local ancestry transitions that influenced their paired phenotype.

3. LA Transitions of Nominal Hits
We tested the nominally associated variants identified in the PheWAS for interactions with nearby local ancestry transitions that influenced the phenotype associated in the PheWAS.

Figure 12. Outline of the analyses performed to investigate interactions between local ancestry and variants. In this Chapter, I investigated three categories of phenotypes: previous associations to phenotypes Metabochip was designed to fine-map; previous associations in African-descent populations in the GWAS Catalog; and all ICD-9 phenotypes contained in the EHR. For the first two phenotype categories, I first conducted analyses to investigate whether prior associates generalized to our sample. In contrast, investigation of all the ICD-9 phenotypes, which were translated to phecodes, was designed to identify novel associations. I then conducted a second set of analyses for each of these phenotypes to investigate whether variants interacted with local ancestry transitions to influence the specified phenotypes.

European associations for phenotypes targeted by Metabochip do not generalize to African American populations

The Metabochip genotyping array was designed to fine-map genetic associations for type 2 diabetes, coronary artery disease, myocardial infarction, and their associated quantitative traits in genome-wide association studies.[151] Based on prior associations, each of the over 200 densely-genotyped genomic regions is assigned to specific phenotypes. I analyzed quantitative traits associated with these diseases that are frequently collected as part of routine clinical visits, including: body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP), and low-density lipoprotein (LDL) levels (Chapter 3; Subjects and methods: Phenotype processing and quality control). Given the prevalence of drugs designed to alter both blood pressure and lipid levels, I investigated both a baseline, pre-, and post-medication value for SBP, DBP, and LDL levels (Methods). The distribution for these traits resembles non-clinical populations, and there is a median of 5446 individuals with data for these phenotypes (Figure 13).

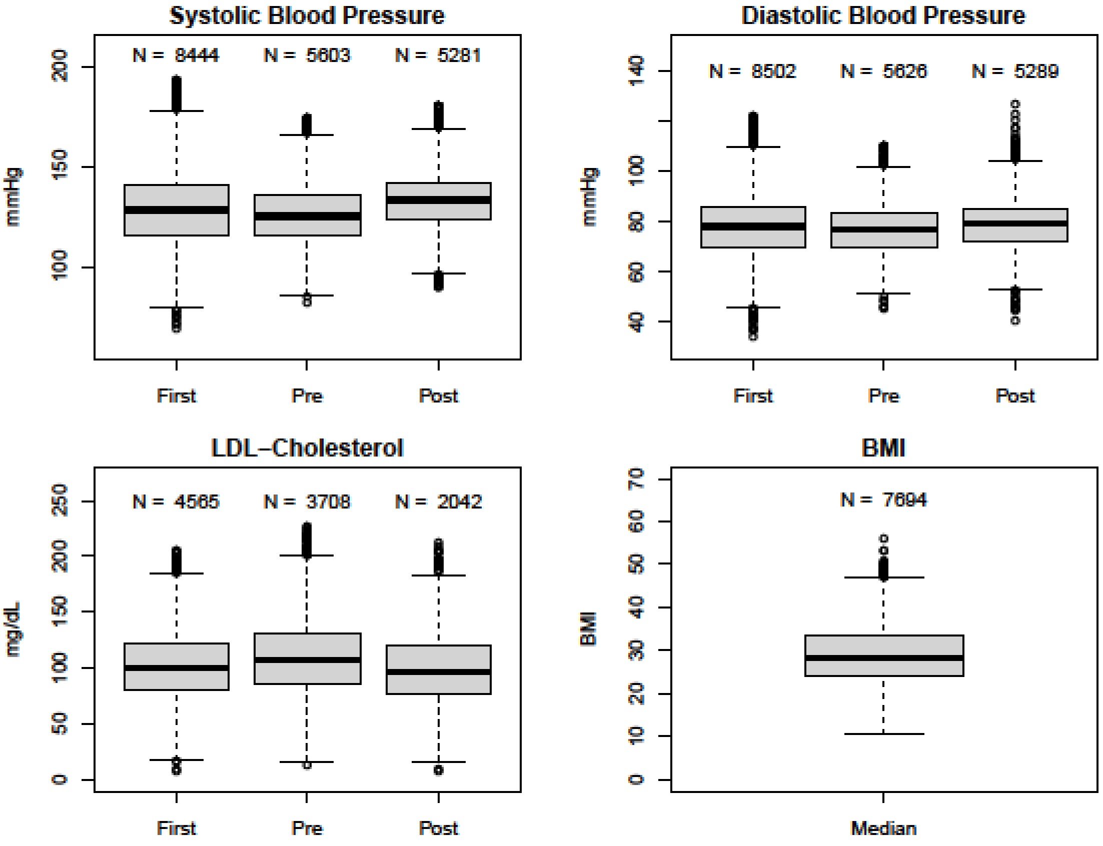


Figure 13. Distributions of EHR-derived SBP, DBP, LDL, and BMI measurements. For each of the 9,559 African American individuals genotyped on the Metabochip as part of the EAGLE Project, I derived their SBP, DBP, LDL, and BMI values from their medical record (Methods). For SBP, DBP, and LDL there are three values – the first measurement in the medical record, the median of yearly medians prior to mention of phenotype-altering medications, and the median of yearly medians after the mention of phenotype-altering medications.

To determine whether European-associations generalized to African Americans, I first performed a Metabochip-wide association test (Metabochip-WAS) for each of these traits. For each variant (MAF > 1%), I performed a linear regression analysis in which I assumed an additive genetic effect and included age, gender, the top three principal components (PCs), and BMI (except for when BMI was the trait of interest) as covariates. To adjust for multiple testing, a trait-specific Bonferroni significance threshold was set. No variants for any trait passed this significance threshold (Figures 14-16), suggesting that either European associations do not generalize to African Americans, or that there was insufficient power to detect them in this dataset.

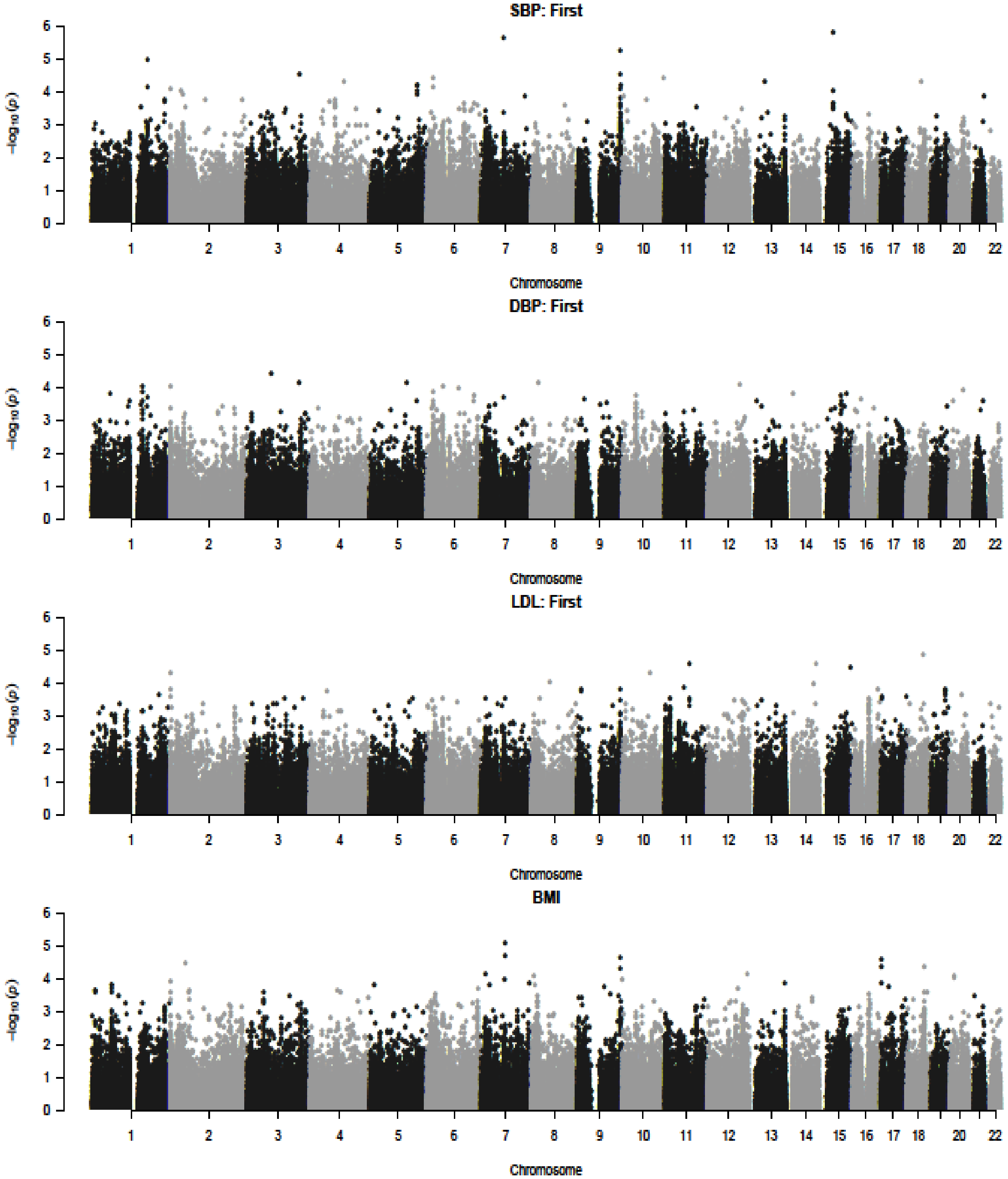


Figure 14. No variants significantly associated with SBP, DBP, LDL, or BMI in a Metabochip-WAS. Manhattan plots for a Metabochip-WAS of A) SBP, B) DBP, C) LDL, and D) BMI are provided. The Metabochip-WAS Bonferroni multiple testing correction for each trait separately is $3.2x10^{-7}$, which none of the variants passed.

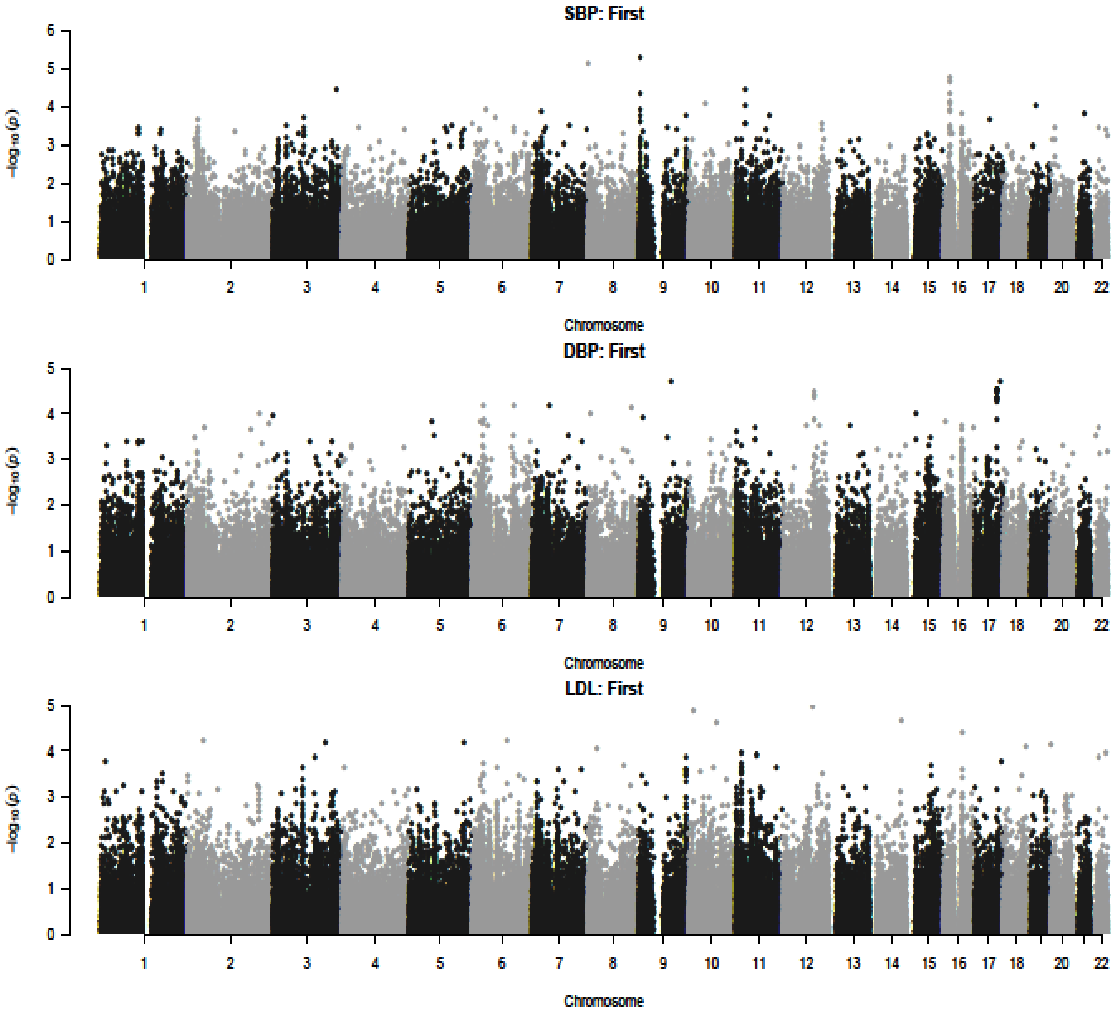


Figure 15. No variants significantly associated with pre-medication measurements for SBP, DBP, or LDL in a Metabochip-WAS. Manhattan plots for a Metabochip-WAS of pre-medication measurements for A) SBP, B) DBP, and C) LDL are provided. The Metabochip-WAS Bonferroni multiple testing correction for each trait separately is 3.2x10^{-7}, which none of the variants passed.

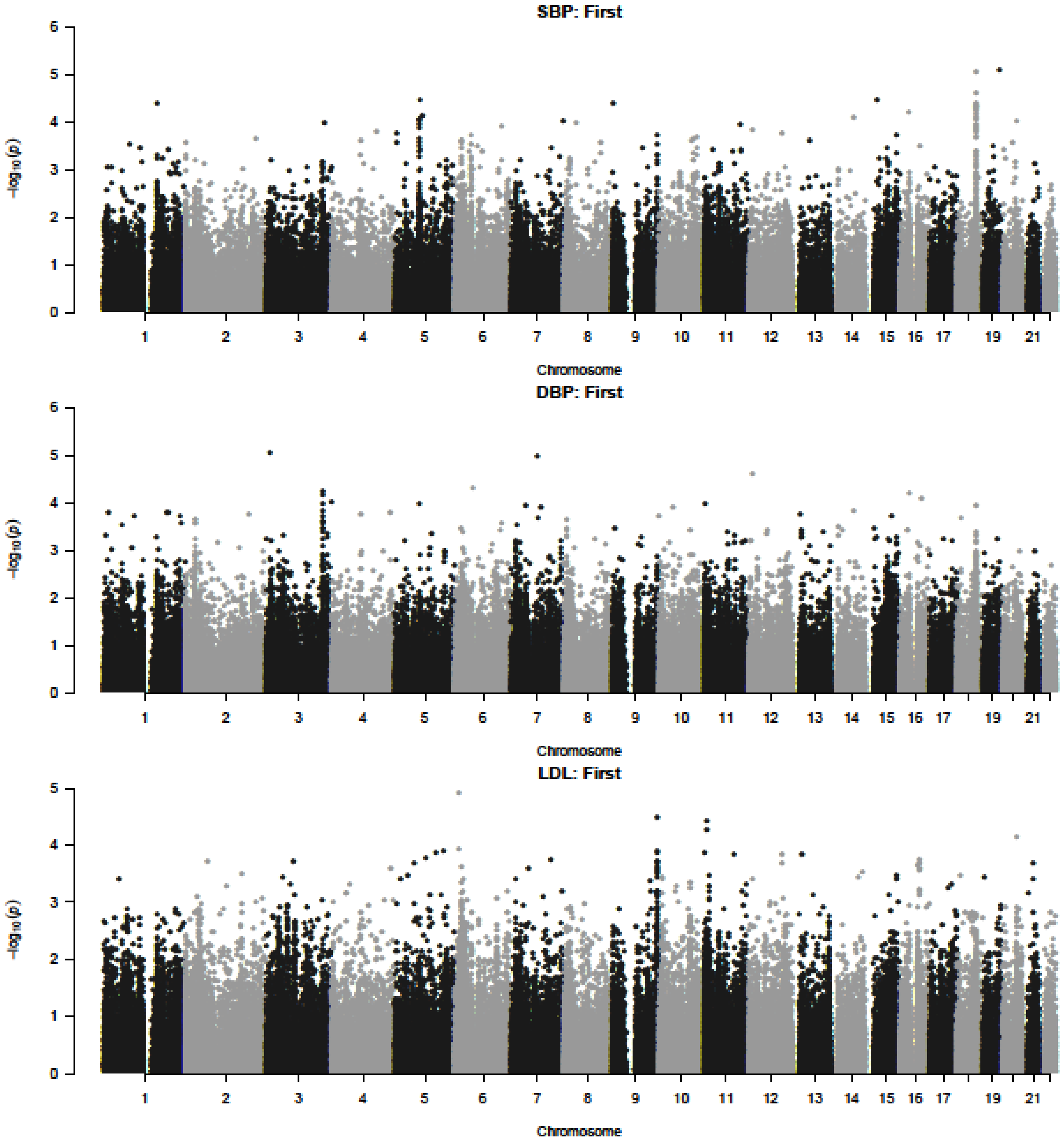
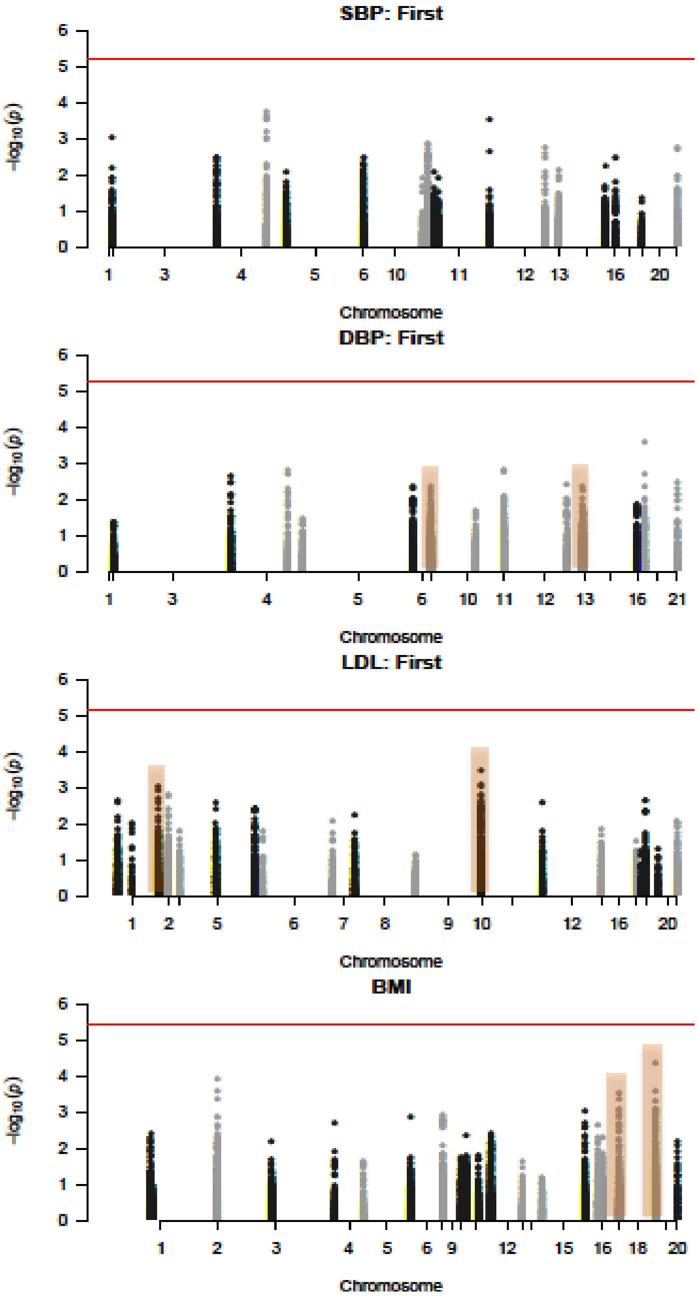


Figure 16. No variants significantly associated with post-medication measurements for SBP, DBP, or LDL in a Metabochip-WAS. Manhattan plots for a Metabochip-WAS of post-medication measurements for A) SBP, B) DBP, and C) LDL are provided. The Metabochip-WAS Bonferroni multiple testing correction for each trait separately is 3.2x10^{-7}, which none of the variants passed.

Relative to many GWAS, our study has reduced power to detect effects given the smaller sample size. To improve power, I next performed a targeted association analysis wherein I only tested variants for association with each for these traits if they occurred in a Metabochip region that had previously been associated with that trait (Table 4). Exclusively examining these regions dropped the number of association tests performed for each trait from a median of 156,004 to 8,477; however, no variants passed a trait-specific multiple testing correction threshold in the targeted analysis, despite its greater leniency (Figure 17-19). There were several suggestive associations though, which I hypothesized might be influenced by local ancestry transitions. To investigate this, I further filtered the Metabochip regions included in the targeted analysis: only previously associated regions with at least a hundred local ancestry transitions observed were considered. Two regions apiece met these criteria for DBP, LDL, and BMI – none did for SBP. Notably, these regions were those with the significant variants from the broader targeted analysis for both BMI and LDL (Figure 17). For common variants (MAF > 5%; LD-pruned at $r^2 > 0.9$) in these regions, I then performed a regression analysis to determine if the presence of a local ancestry transition within the region interacted with the variant to influence the previously associated phenotype (Methods). The significance of these interaction terms is provided in Table 5; however, no terms passed either a Bonferroni multiple-testing correction threshold, or a false-discovery rate (FDR) of 5%. Thus, there is little evidence local ancestry transitions interact with genetic variants to influence these phenotypes; however, there is also little evidence that there were any genetic associations within these regions.

Trait	*Regions*	*Trait*	*Regions*
LDL	chr1:25525907-25908241	***BMI***	chr1:72513687-72958905
	chr1:55498949-55513521		chr1:74961817-75078975
	chr1:109655637-110043693		chr1:177753776-177936525
	chr2:21226560-21451827		chr2:471136-719889
	chr2:44057030-44100849		chr3:85651797-86050826
	chr5:74568112-74956052		chr3:185747042-185862593
	chr5:156331094-156505233		chr4:45099376-45187658
	chr6:16104254-16134837		chr5:74562373-75123052
	chr6:160468278-160579527		chr6:50534485-51100751
	chr7:44374092-44676286		chr9:28403443-28499099
	chr8:126441650-126543928		chr11:8394189-8707147
	chr9:136042324-136482476		chr11:27452706-27749725
	chr11:126219429-126279347		chr11:47243424-48094879
	chr12:121304826-121488876		chr12:50168189-50290056
	chr16:71996291-72147683		chr14:30436558-30543794
	chr19:11183837-11211208		chr15:67649978-68215300
	chr19:19301232-19792250		chr16:19704224-20019432
	chr19:45396899-45444266		chr16:28306987-29001460
	chr20:39083142-39128578		chr16:53539509-54185787
	chr20:39613984-40010045		chr18:57727147-58094636
			chr19:34295278-34333501
SBP	chr1:11794676-11968356	**DBP**	chr1:11824260-11909736
	chr3:169087965-169195349		chr3:169087965-169195349
	chr4:81155937-81207963		chr4:81155937-81207963
	chr5:32689850-32867260		chr4:103121726-103218446
	chr5:157713315-157952955		chr5:157713315-157952955
	chr10:95869815-95949432		chr6:25235303-26141375
	chr10:104217441-104999266		chr10:63381832-63553849
	chr11:9886230-10370634		chr10:104217441-104999266
	chr11:16844924-16988268		chr12:89824040-90118890
	chr11:100497893-100698228		chr12:111505708-113105952
	chr12:89788633-90118890		chr12:115343492-115438209
	chr12:111681897-112225304		chr15:74864568-75449674
	chr15:74864568-75374591		chr20:10941849-10998754
	chr15:91390400-91441094		chr20:57660009-57790618
	chr17:43147554-43273187		
	chr20:57660009-57790618		

Table 4. Metabochip regions associated with LDL, BMI, SBP, and DBP. This chart indicates which Metabochip regions were previously associated with each trait, and therefore included in the targeted analysis.


SBP: First

DBP: First

LDL: First

BMI

Figure 17. No variants significantly associated with SBP, DBP, LDL, or BMI in a targeted analysis of previously associated regions. Manhattan plots for a Metabochip-WAS of A) baseline SBP, B) baseline DBP, C) baseline LDL, and D) BMI are provided. The Bonferroni multiple-testing correction threshold for each trait in the targeted analysis (SBP: $p = 6.57 \times 10^{-6}$, DBP: $p = 5.35 \times 10^{-6}$, LDL: $p = 7.23 \times 10^{-6}$, BMI: $p = 3.83 \times 10^{-6}$), which none of the variants passed, is indicated by the red line. Regions with at least 100 local ancestry transitions are highlighted in orange.

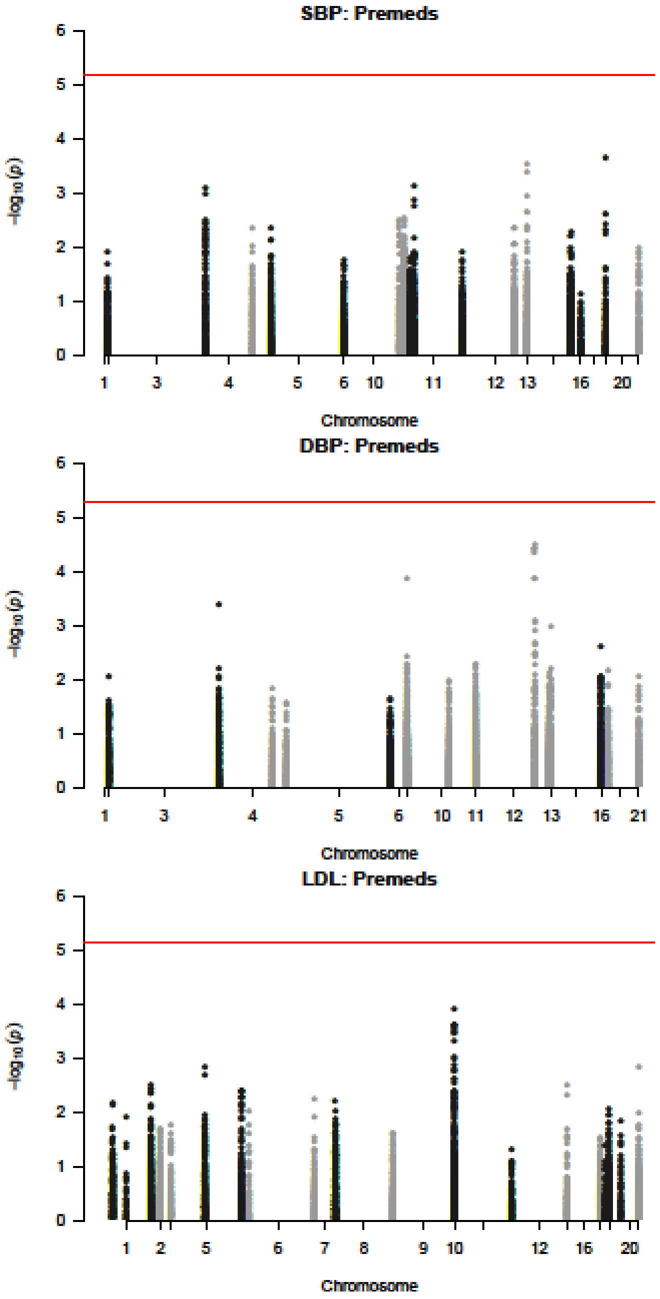


Figure 18. No variants significantly associated with pre-medication measurements for SBP, DBP, or LDL in a targeted analysis of previously associated regions. Manhattan plots for a Metabochip-WAS of pre-medication measurements for A) SBP, B) DBP, and C) LDL are provided. The Bonferroni multiple-testing correction threshold for each trait in the targeted analysis (SBP: $p = 6.58 \times 10^{-6}$, DBP: $p = 5.35 \times 10^{-6}$, LDL: $p = 7.25 \times 10^{-6}$), which none of the variants passed, is indicated by the red line.

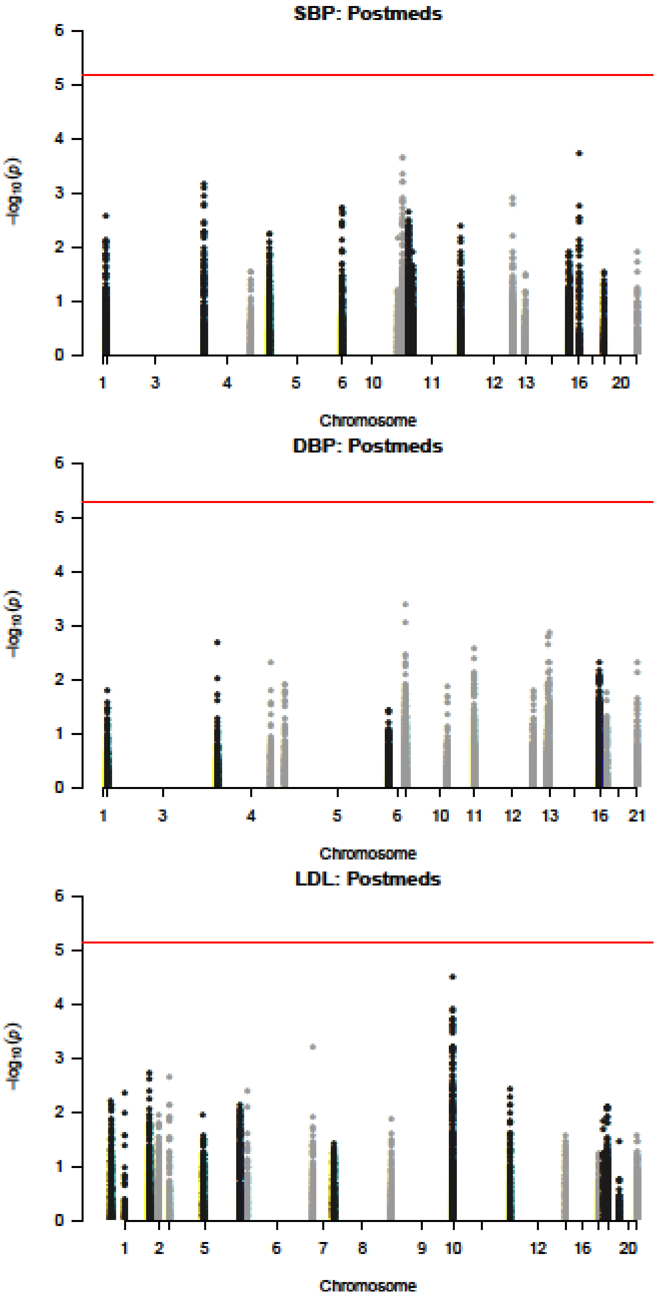


Figure 19. No variants significantly associated with post-medication measurements for SBP, DBP, or LDL in a targeted analysis of previously associated regions. Manhattan plots for a Metabochip-WAS of post-medication measurements for A) SBP, B) DBP, and C) LDL are provided. The Bonferroni multiple-testing correction threshold for each trait in the targeted analysis (SBP: p = 6.58x10^{-6}, DBP: p = 5.35x10^{-6}, LDL: p = 7.21x10^{-6}), which none of the variants passed, is indicated by the red line.

Trait	Metabochip Region	SNPs (N)	Nominal (p < 0.05)	Bonferroni	FDR = 5%
BMI	chr16:53539509-54185787	499	14	0	0
	chr18:57727147-58094636	277	1	0	0
DBP	chr6:25235303-26141375	674	25	0	0
	chr12:111505708-113105952	577	1	0	0
LDL	chr1:109655637-110043693	224	2	0	0
	chr9:136042324-136482476	301	25	0	0

Table 5. Local ancestry transitions do not significantly interact with variants to influence DBP, LDL, or BMI. I tested whether local ancestry transitions interact with variants in previously-associated regions with at least 100 local ancestry transitions to influence the indicated phenotype (Methods). Here, I report the number of SNPs tested within that region, and the number of interaction terms that passed various significance threshold cut-offs. No interactions were significant with multiple-testing corrections.

There are two potential explanations for the lack of genetic association within these regions: there is insufficient power to detect these effects, or that European-associations do not generalize to African American populations. To distinguish between these two possibilities, I performed power calculations across a range of allele frequencies and effect sizes. For each trait, I determined the minimum effect size that I had a power of 80% to detect using the actual sample size, mean, standard deviation, and number of association tests performed using Quanto.[153] I considered both a rare (MAF = 5%) and a common (MAF = 25%) allele frequency to capture the range in power across the allele frequency spectrum. Table 6 contains the effect sizes that I was well-powered to detect for each trait. In the GWAS Catalog, there are common variants reported for all four traits that have effect sizes greater than those I am well-powered to detect; these are additionally reported in Table 6. Thus, I was well-powered to detect effects commensurate with those observed in European-populations across a broad range of allele-frequencies, which suggests that European-associations did not generalize to this dataset. Consequently, the lack of genetic association within the region likely accounts for why no significant interactions between variants and local ancestry transitions were observed.

Trait	Effect size detectable at MAF: 0.05	Effect size detectable at MAF: 0.25	GWAS Catalog Effect Size MAF ≥ 0.05
SBP: First	3.75 mmHg	2.00 mmHg	5.43 mmHg
SBP: Pre	3.75 mmHg	1.75 mmHg	5.43 mmHg
SBP: Post	3.75 mmHg	2.00 mmHg	5.43 mmHg
DBP: First	2.50 mmHg	1.25 mmHg	3.20 mmHg
DBP: Pre	2.25 mmHg	1.25 mmHg	3.20 mmHg
DBP: Post	2.50 mmHg	1.25 mmHg	3.20 mmHg
LDL: First	8.25 mg/dL	4.25 mg/dL	12.30 mg/dL
LDL: Pre	10 mg/dL	5.00 mg/dL	12.30 mg/dL
LDL: Post	13 mg/dL	6.75 mg/dL	12.30 mg/dL
BMI	1.25 units	0.75 units	1.54 units

Table 6. Targeted analysis is well-powered to detect anticipated effects for SBP, DBP, LDL, and BMI. I calculated the effect size at which there was 80% power to detect effects for each trait, using the actual mean, standard deviation, sample size, and number of association tests performed. I considered both a rare allele frequency (MAF = 0.05) and a common allele frequency (MAF = 0.25). For each trait, I identified the strongest-effect variant associated in the GWAS Catalog for a common variant (MAF ≥ 0.05).

Local ancestry transitions interact with variants to influence GWAS Catalog traits

The GWAS Catalog[32] contained additional phenotypes that had been associated to regions fine-mapped by Metabochip. I exclusively analyzed associations for variants that occurred in Metabochip regions with at least 100 observed local ancestry transitions, as I ultimately wished to investigate the interaction of these transitions with variants. I additionally limited the associations to those that made reference to African ancestry in the study sample, were for phenotypes that could be readily derived from the EHR, and had at least 200 cases in the EHR. This resulted in 28 phenotype-Metabochip region pairs (Table 7). I did not restrict the analysis to the specific variant referenced in the GWAS Catalog, due to both differential LD structure between populations and representation of different variants on different genotyping platforms. Instead, I tested all variants in the Metabochip region for association with the phenotype. There was at least one nominal genetic association for each trait-Metabochip region pair, illustrating some level of genetic association within the region.

GWAS Catalog Trait	Corresponding EHR Trait	Metabochip Region
Urate levels	UricA*	chr6:25235303-26141375
Type 2 diabetes	PAGE T2D Algorithm[154]	chr11:2444094-2943115
Red blood cell traits	RBC*; RDW*	chr6:25235303-26141375
Iron status biomarkers	TIBC*	chr6:25235303-26141375
Weight	Weight**	chr16:53539509-54185787 chr18:57727147-58094636
Hematology traits	Alb*; AlkP*; AN-GAP*; BUN*; Ca*; Cl*; CO2*; Creat*; GluBed*; Gluc*; Hgb*; K*; MCHC*; MCH*; MCV*; NA*; RBC*; RDW*; SGOT*; SGPT*; TBil*; WBC; MPV*; Plt-Ct*; TIBC*	chr6:25235303-26141375
Mean platelet volume	MPV*	chr12:111290599-113206306 chr12:111505708-113105952 chr12:111681897-112225304 chr6:25235303-26141375
Height	Height**	chr7:27784039-28282062
Obesity-related traits	BMI**	chr16:53539509-54185787
Platelet count	Plt-Ct*	chr12:111290599-113206306
Coronary artery disease	Cases at least one ICD-9 Codes (410 – 414); all others were controls	chr12:111290599-113206306 chr12:111505708-113105952 chr12:111681897-112225304 chr13:110795080-111049623 chr18:57727147-58094636
LDL cholesterol	First LDL-C measurement	chr1:109655637-110043693 chr1:109789347-109826136
Body mass index	BMI**	chr12:111290599-113206306 chr12:111505708-113105952 chr12:111681897-112225304 chr16:53539509-54185787 chr18:57727147-58094636 chr3:122976919-123206919 chr3:123039584-123139034

Table 7. GWAS Catalog traits with genetic associations in Metabochip regions. I identified variants in the GWAS Catalog that were contained within Metabochip regions, and then filtered these down to a subset of associations with sufficient data to examine in this dataset. Here, I provide the trait as described in the GWAS Catalog, the EHR-implementation of that trait, and the Metabochip region to which it corresponded. In the case of lab values marked with an

asterisk, the median value was taken. In the case of values marked with a double asterisk, the median of yearly medians was taken.

I next investigated whether nearby transitions in local ancestry influenced the effects of these variants, as previously described. In Figure 20, the significance of all terms across Metabochip region chr6:25235303-26141375 with mean corpuscular hemoglobin (MCH) is provided as an example. I identified five significant interactions between local ancestry transitions and the allele across all traits, using a region-phenotype specific Bonferroni multiple testing correction. I first visually characterized these interactions, grouping chromosomes together based on their local ancestry, allele, and local ancestry transition status. From this, it is apparent that two interactions (one between the variant rs9467458 and creatinine levels, the other between rs4712930 and white blood cell (WBC) counts) are driven by chromosome combinations that are rarely observed (i.e., have low cell counts) (Figure 21). With so few observations, it is difficult to discern whether the chromosome category actually has an effect on the phenotype, or whether the individual with that chromosome category happens to fall on an extreme end of the normal phenotypic distribution for other reasons. This is evidenced by their lack of specific chromosome categories that are significant (Figure 21). The interaction between the variant rs1410438 and CO2 levels has a sufficient number of chromosomes in each category, and multiple chromosomes with a local ancestry transition seem to contribute to the significance of the interaction; however, it does not appear to be relevant whether the local ancestry transition is upstream or downstream of the variant (Figure 21). This is suggestive of the transition itself, rather than any interaction with local ancestry, driving the effect. The two remaining interactions do have low numbers of chromosomes that meet the criteria for some categories; however, categories with low cell counts closely resemble the sample median and are not driving the trend. Consequently, two interactions remain promising and are further investigated.

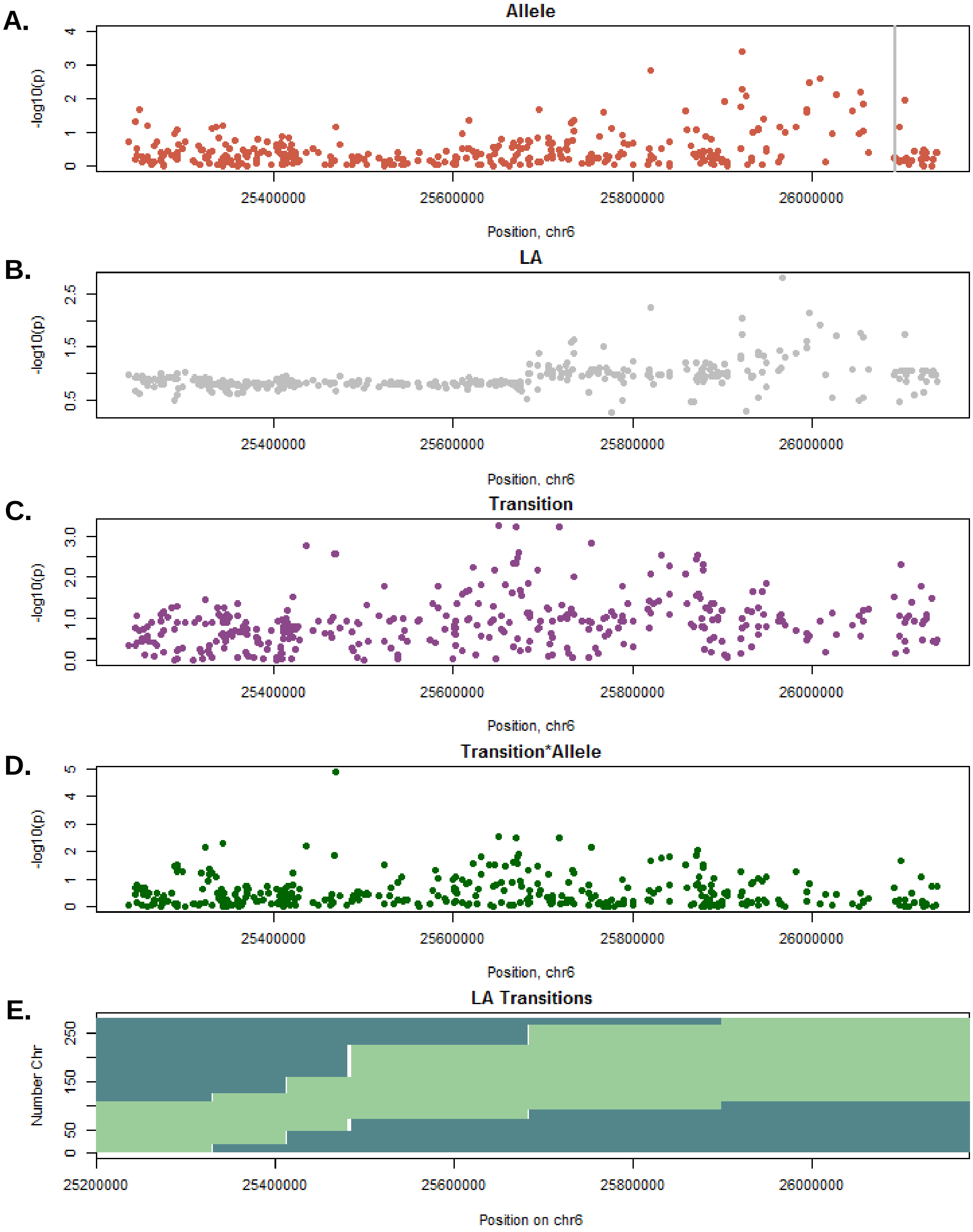


Figure 20. The association of genetic variants and local ancestry to MCH. The variant rs1800562 (location marked by gray line in A) has been associated with a variety of iron-related phenotypes, and is located on a region of chromosome six that was densely genotyped on the Metabochip platform. I tested all the variants within this region for association with MCH using a linear regression model, in which I included covariates (top three principal components, age at measurement, and gender) and terms for the allele, local ancestry, presence of a local ancestry

transition in the region, and an interaction between the allele and local ancestry transition (Methods). The Manhattan plots for these terms are provided in A-D (note the difference in scale), respectively. The specific local ancestry transitions observed in this region are shown in E. Dark green indicates European ancestry along the chromosome; light green indicates African ancestry.

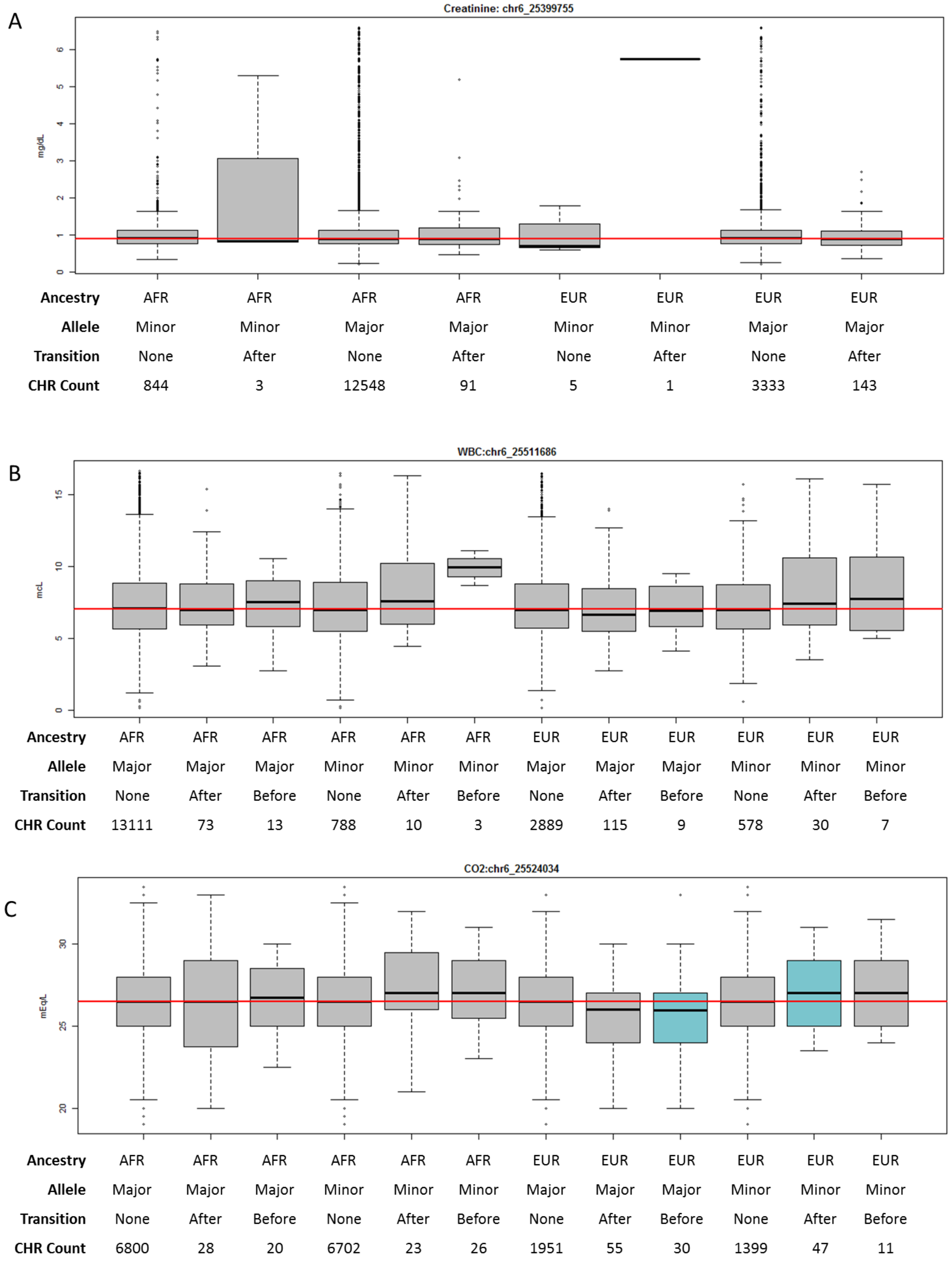


Figure 21. Interactions between local ancestry transitions and variants regulating creatinine, white blood cell counts, and CO$_2$ levels. For each significant interaction between

local ancestry transitions and a variant, I characterized the interaction by stratifying chromosomes based on: the local ancestry at the variant (EUR or AFR); whether they had the major or minor allele; whether there was a local ancestry transition on that chromosome within the broader Metabochip region, and if so, whether it occurred after (i.e., downstream) or before (i.e., upstream) of the variant. The number of chromosomes in each category is additionally provided. I conduced pairwise Mann Whitney U tests, comparing each category to the remainder of the sample, to determine significance. Categories significantly different ($p < 0.05$) are shown in blue. The overall median is shown in red. The interaction between variant rs9467458 and local ancestry transitions to influence creatinine levels (A) is driven by a single chromosome of European ancestry, with the minor allele, and a downstream local ancestry transition. This chromosome category, while the most significant, does not significantly differ from the rest of the sample ($p = 0.087$). The interaction between variant rs4712930 and local ancestry transitions to regulate white blood cell (WBC) counts is attributable to three chromosomes with African ancestry, the minor allele, and an upstream ancestry transition. Again, while the most significant chromosome category, it is not significant ($p = 0.051$). Given the small cell counts, these are not further investigated. The interactions between variant rs1410438 and local ancestry transitions associated with CO2 levels (C), while not due to low cell counts, does not have a clear biological interpretation. Two chromosome categories, highlighted in blue, are significantly different ($p < 0.05$) from the rest of the categories.

rs16890640 interacts with local ancestry transitions to influence red blood cell traits

We further investigated the two remaining interactions between local ancestry transitions and the allele to better understand the biological mechanisms underlying them. These two interactions identified the same variant, rs16890649, as interacting with a local ancestry transition to influence both MCH and mean corpuscular volume (MCV). These two phenotypes are highly correlated with one another, and consequently, the interactions strongly resemble one another. As shown in Figure 22, individuals with the variant on a European background and a downstream local ancestry transition have markedly lower MCH/MCV than any other combination. I further stratified these individuals on the basis of where their local ancestry transition occurred. There is a position-dependent effect wherein individuals with a local ancestry transition at the closest transition point (bp: 25481231) had lower MCH than did individuals with either of the two more distant transitions points (Figure 23). This suggests that the functional element that rs16890649 is putatively interacting with is located between rs16890649 and the second transition point.

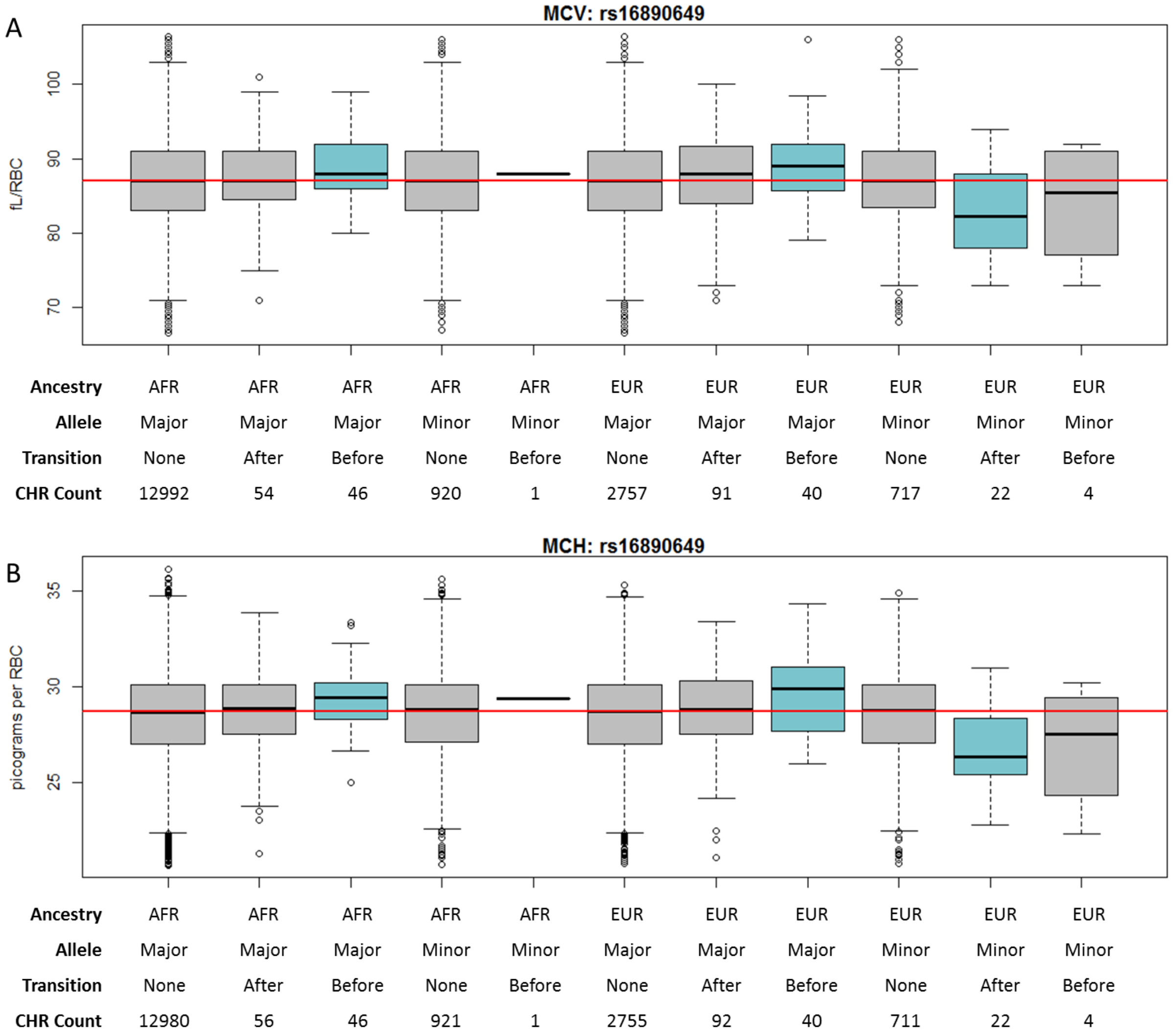


Figure 22. Chromosomes with the minor allele for rs16890649 on a European background and a downstream local ancestry transition are associated with lower MCH and MCV. To better understand the interactions for MCV (A) and MCH (B), I stratified chromosomes based on: the local ancestry at rs16890649 (EUR or AFR); whether they had the major or minor allele; whether there was a local ancestry transition on that chromosome within the broader Metabochip region, and if so, whether it occurred after (i.e., downstream) or before (i.e., upstream) of the variant. The count row provides the number of chromosomes observed in each category. Individuals with a chromosome that: has the minor allele of rs16890649 on a European background with a downstream local ancestry transition have lower MCH and MCV levels than the sample median (indicated by the red line). To determine which chromosome categories were driving the interaction, I performed a Mann Whitney U test comparing each chromosome category against the rest of the population; all significant interactions ($p < 0.05$) are in blue. Only one chromosome category was significant with multiple testing corrections for each pairwise test: the minor allele of rs16890649, on a European ancestry, with a downstream local ancestry transition for MCH ($p = 0.0024$).

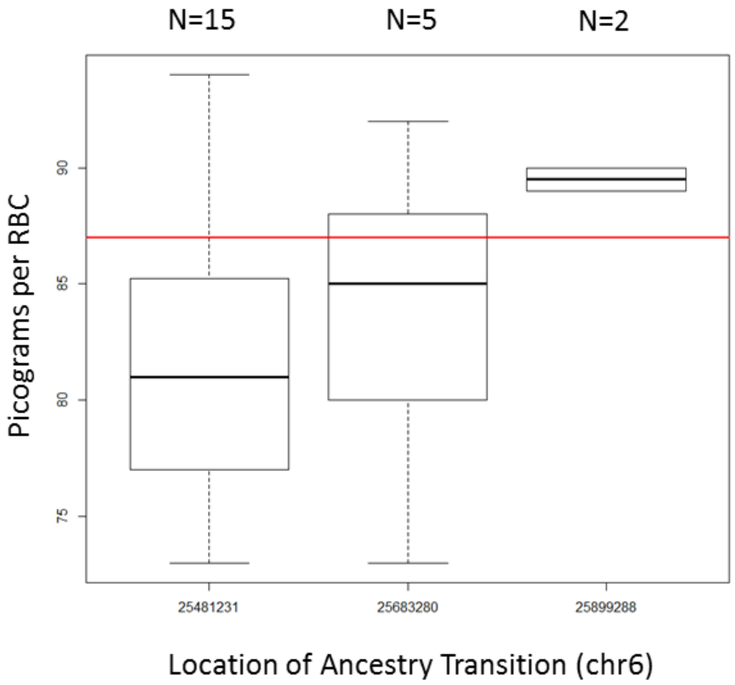


Figure 23. The effect of downstream local ancestry transitions on MCH is position-dependent. There were three possible points at which downstream local ancestry transitions occurred (Figure 20). Individuals with the transition point at chr6:25481231 had the lowest average MCH levels. Individuals with transitions at the two subsequent transition points began to approach the median MCH level, which is shown in red. The number of chromosomes with European ancestry, the variant, and a local ancestry transition at each of these locations is provided above each boxplot. The MCH levels were not significantly different ($p > 0.05$; Mann Whitney U test) from one another; however, this is likely due to the small number of chromosomes with transitions at the later points.

To better understand the biological mechanism mediating this interaction, I annotated this variant. First, it is roughly three times more common in European-descent populations (CEU = 23%) than it is in African-descent populations (YRI = 7%). It occurs within an intron of LRRC16A, which encodes a cytoskeleton-associated protein involved in regulation of actin polymerization; it is also associated with platelet development and production. rs16890640 is an eQTL for LRRC16A in whole blood, although it is not predicted to be an enhancer based on histone-modification patterns. However, the variant does fall within an observed binding site for a relevant transcription factor MAFK (Figure 24); when knocked out in mice, this transcription factor results in reduced MCV and MCH levels[155]. Additionally, it is less than 500 base pairs upstream of a predicted insulator element (Figure 24); however, chromatin looping patterns indicate that contacts occur on either side of this putative insulator (Figure 25). Thus, rs16890640 occurs within a plausibly relevant genomic-region, and is more frequent in Europeans.

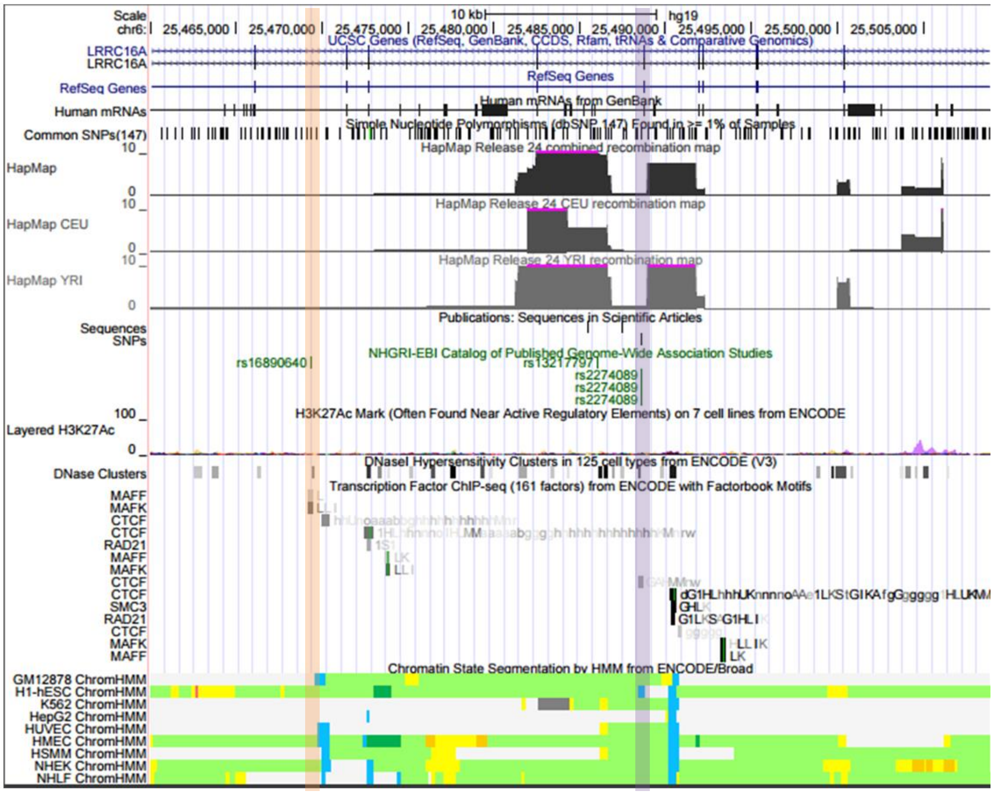


Figure 24. Ancestry-specific recombination hotspots may disrupt functional elements pertinent to MCH and MCV. rs16890640 (highlighted in orange) is located within binding sites for MAFF and MAFK in HEPG2. Additionally, it occurs approximately 500 base pairs upstream of a predicted insulator element. This variant interacts with a downstream local ancestry transition, which likely occurs at one of the two recombination hotspots shown here. The first is shared between populations, whereas the second is specific to YRI, an African population. This recombination peak is in close proximity to rs2274089 (highlighted in purple), a GWAS catalog variant for related traits, and overlaps the next insulator element in GM12878.

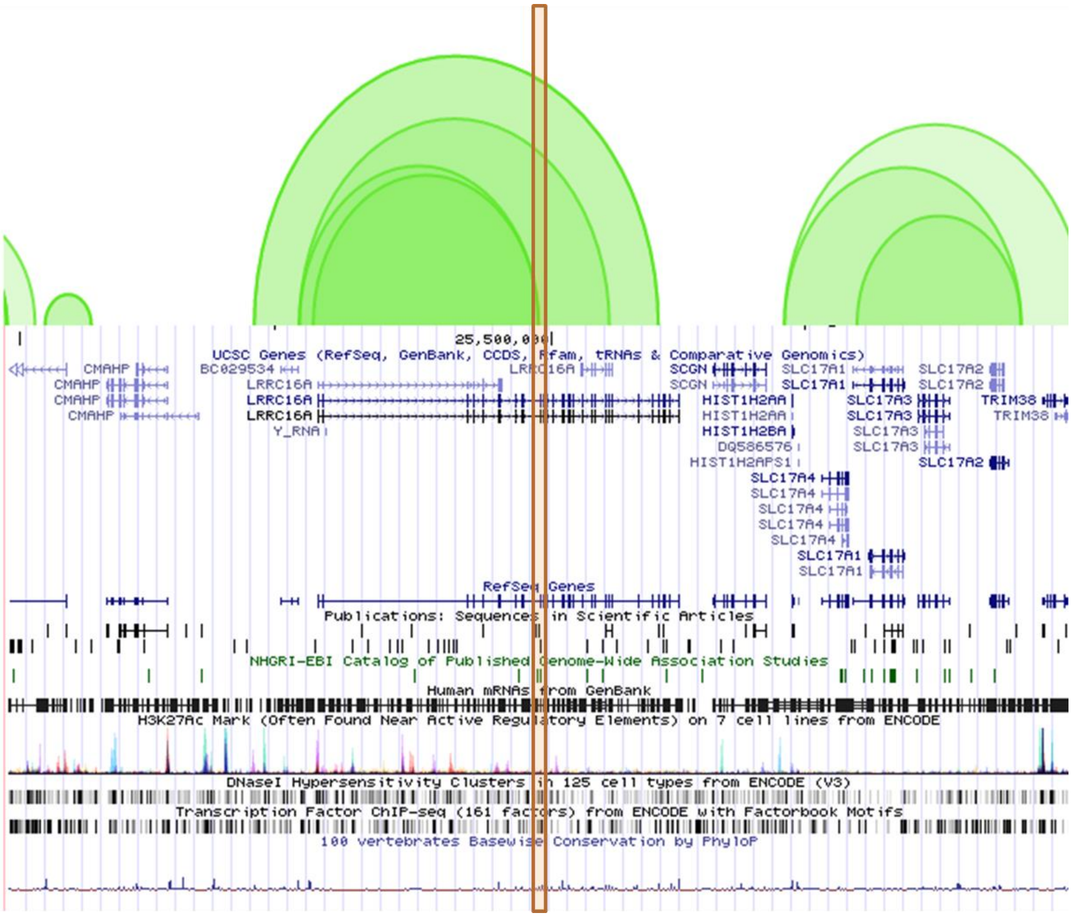

25,500,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
CMAHP
BC029534
LRRC16A
Y_RNA
SCGN
HIST1H2AA
HIST1H2BA
DQ586576
HIST1H2APS1
SLC17A1
SLC17A2
SLC17A3
SLC17A4
TRIM38
RefSeq Genes
Publications: Sequences in Scientific Articles
NHGRI-EBI Catalog of Published Genome-Wide Association Studies
Human mRNAs from GenBank
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
100 vertebrates Basewise Conservation by PhyloP

Figure 25. Local ancestry transitions may perturb chromatin looping patterns within the region. The genomic region containing the African-specific recombination peak physically interacts with the promoter of LRRC16A based on ChIA-PET data for RAD21 in GM12878. The GWAS variant rs2274089, associated with a relevant phenotype, is highlighted in orange.

I next investigated why this variant might interact with a downstream local ancestry transition to influence MCH and MCV levels. I identified a relatively close (within 20kb) GWAS-catalog variant associated with a related phenotype, serum transferrin levels (i.e., the amount of glycoproteins that bind free iron). This suggests that the genomic region may be functionally important for MCH and MCV as well. Notably, this variant, rs2274089, is flanked by the two recombination peaks that could result in a local ancestry transition in the area of interest (Figure 24). The first of these recombination peaks is observed in both European (CEU) and African (YRI) descent populations; however, the second recombination peak is African-specific. This African-specific recombination peak overlaps a predicted insulator element (Figure 24), although this ChromHMM prediction is based largely on the presence of CTCF binding. Chromatin looping data and gene expression data suggest this region may be an enhancer: the region contacts the LRRC16A promoter in GM12878 (Figure 25), and the variant rs2274089 is an eQTL for LRRC16A in whole blood. Regardless of whether the region is an enhancer or insulator, it is clearly engaged in regulatory chromatin looping and is pertinent to related phenotypes.

As elaborated on more fully in the Discussion, I hypothesize that the African-specific recombination site introduces novel genetic variants that disrupt the regulatory functions of this genomic region, that then permit the European variant rs16890640 to engage in ‘off-target’ effects.

PheWAS approach does not identify significant interactions between local ancestry transitions and EHR-derived phenotypes

The GWAS Catalog is an incomplete representation of phenotype-genotype associations: there are many health-relevant phenotypes for which GWAS have not been performed, but which are available in the EHR, and there are many phenotypes in the GWAS Catalog that have not been investigated in an African American population. I performed a phenome-wide association study, or PheWAS, to determine if Metabochip regions harbored any novel associations. Consequently, I initially performed a standard single-marker association analysis that did not incorporate local ancestry information. I did, however, restrict my analysis to Metabochip regions with at least 200 local ancestry transitions as my ultimate goal was to investigate interactions between local ancestry transitions and alleles. Following quality control steps (MAF > 5%; LD-pruned at $r^2 > 0.9$), this left 2,856 variants with local ancestry assignments for analysis. None of the associations identified in the PheWAS passed either a Bonferroni multiple testing correction or a false discovery rate threshold of 10%. However, 168 phenotype-genotype associations were nominally significant ($p < 5x10^{-5}$). Using the previously described methodology, I investigated whether local ancestry transitions interacted with the allele to influence the phenotype for these nominal associations (Table 8) (Full details are available in the Appendix: Nominal PheWAS associations do not interact with local ancestry transitions to influence phenotypes). Only one interaction passed a Bonferroni-multiple testing correction; however, further investigation revealed that it was attributable to low cell counts. Thus, only nominal genetic associations were identified, and these variants did not interact with local ancestry transitions significantly.

Phecode	*Description*	*Variant*	*Variant p-value*	*Variant*Local ancestry transition p-value*
480.5	Bronchopneumonia and lung abscess	rs17641977_G	2.16E-05	2.42E-06
707.1	Decubitus ulcer	chr16.52362593_G	2.34E-06	7.97E-04
250.13	Type 1 diabetes with ophthalmic manifestations	chr6.25733715_G	1.45E-05	1.51E-03
204.4	Multiple myeloma	chr1.109619786_G	8.62E-06	1.11E-02
614.1	Pelvic peritoneal adhesions, female (postoperative) (postinfection)	chr7.14901079_G	8.18E-06	1.61E-02
440	Atherosclerosis	rs2744238_G	3.86E-05	1.98E-02
276.11	Hyperosmolality and/or hypernatremia	chr1.160348446_A	2.95E-05	2.91E-02
270.3	Disorders of plasma protein metabolism	chr7.14760490_G	1.56E-05	3.20E-02
647.1	Infections of genitourinary tract during pregnancy	chr16.52588027_A	1.43E-06	3.87E-02
331.9	Cerebral degeneration, unspecified	chr13.109701508_G	3.24E-05	4.59E-02
628	Ovarian cyst	chr7.14566856_G	4.18E-05	5.13E-02

Table 8. Top associations from PheWAS for interactions between local ancestry transitions and alleles. I examined whether variants nominally associated ($p < 5x10^{-5}$) in the PheWAS interacted with local ancestry transitions to influence the associated phenotype. The top ten results from this analysis are reported here, the rest are in Appendix A. While one interaction passed the Bonferroni multiple testing correction for the interaction analysis, it was driven by a single chromosome (i.e., low cell counts) and was not considered further.

Discussion

In this Chapter, I hypothesized that ancestry-variable recombination events would disrupt the haplotype boundaries typically observed, thereby enabling the detection of epistasis within haplotypes. I investigated this in almost ten thousand African American adults, with both EHR-derived phenotypes and genetic data on the Metabochip. I first sought to identify significant genetic associations to phenotypes around which the Metabochip was designed; however, these European-based associations did not generalize to African American populations, despite our study being well-powered to detect reported effect sizes. To investigate epistasis within haplotypes, I then examined whether alleles interacted with nearby local ancestry transitions to influence these phenotypes, and found little support. However, when I performed the same analysis for associations that had been originally identified in African-descent population, I found an interaction between a variant and a nearby local ancestry transition. This suggests that combinations of genomic regions from differential continental ancestries may interact with one another to influence health traits in humans.

A variety of potential biological mechanisms exist through which this might occur. In the case of the interaction I identified, I hypothesize that events frequent in one continental population can each alter local regulatory events, and that when in combination with one another, influence a clinical phenotype. Specifically, the variant rs16890640 was associated with reduced MCV and MCH levels when it occurred on a European background and there was an immediate downstream local ancestry transition. The variant was roughly three times more frequent in Europeans, and it may have a regulatory effect as it both occurs within observed binding sites for trait-relevant transcription factors and is an eQTL. The transition could occur at either a hotspot shared between populations or at one specific to African-descent. These potential recombination sites flank a GWAS Catalog variant for a related trait, illustrating the phenotypic-relevance of the genomic region. The region also is densely annotated for regulatory function, and engages in chromatin looping to nearby promoters. I hypothesize that the African-specific recombination hotspot, which overlaps putative insulators, is introducing low-frequency genetic variants that alter its function.[156] When this occurs, the regulatory variant rs16890640 is then able to engage in ‘off-target’ effects, which ultimately reduce MCH and MCV levels. For example, the regulatory region could interact with *HFE*, a gene approximately 600 kb away that regulates iron uptake.[157,158] Mutations in this gene cause hereditary haemochromatosis, wherein excess iron is deposited within organs, ultimately leading to their failure.[159] To investigate this hypothesis, I am first examining whether there are such disruptive variants at the African-specific recombination peak. There are also other possible explanations: regardless of recombination-induced mutagenesis, an African haplotype may simply be carrying a variant that interacts with rs16890640 to influence MCH and MCV. However, it is less clear whether downstream elements directly relate to MCH or MCV levels. Ultimately, functional validation of any hypotheses – such as examining whether chromatin looping is altered – will be required to discern between possibilities.

The interaction I identified provides potential evidence for epistasis influencing health-related phenotypes in humans. The variant rs16890640 is not significantly associated with the phenotype on its own – it is only in combination with the downstream transition to African ancestry that an association to the phenotype is observed. While it is possible that this combination of events somehow is tagging a causal variant within this region, I consider this unlikely as nearby variants did not demonstrate a strong association. Instead, it highlights that admixed populations provide a unique opportunity to investigate epistasis, as novel combinations of variants are generated and unique population-specific recombination hotspots may disrupt functional haplotypes.

That I identified only a single interaction between a variant and local ancestry transitions should not be taken as evidence that these events are rare due to several limitations of our study. First, the Metabochip is a custom genotyping array specifically designed to capture variation within specific genomic regions; it is not a genome-wide platform. Thus, I only investigated a subset of the local ancestry transitions that occur within the whole genome. Additionally, it should be recognized that while EHRs provide a wealth of medical information, they are often incomplete representations – not all diagnoses may be contained within a patient’s EHR, especially when Vanderbilt Hospitals are not the patient’s source of primary care. In some cases, this may result in actual cases being considered controls. While this should not result in spurious associations, it would reduce our power to detect effects. Finally, all local ancestry transitions were treated the same in this study, regardless of where they occurred, or whether it was from European to African, or vice-versa. I grouped the transitions together due to their general infrequency; however, this may have diluted signal originating from specific combinations. Consequently, interactions between variants and local ancestry transitions may be a more frequent mechanism influencing health-related phenotypes.

An additional limitation was the lack of robust genetic associations within this region, regardless of local ancestry transitions. Given that these regions largely contained variants with significant associations in the GWAS Catalog, we were initially surprised by the lack of robust signal. However, studies generalizing associations between European and African descent populations have had mixed success thus far. Many find that while the direction of effect may be consistent between populations, the significance of this effect is influenced by differential LD structure with the causal variant, frequency differences between populations, and sample size.[154,160–162] As a consequence, a notable proportion of variants fail to generalize between populations when significance of the association is the primary metric, which we used in this analysis.[162,163] While this may account for the lack of generalization of signals between populations, the major issue for our study is the lack of robust associations, rather than the generalization of associations, as it obfuscates the interpretation of negative results. It is unclear in many cases whether local ancestry transitions do not interact with the variant to influence the phenotype, or whether the region is unrelated to the phenotype in African Americans.

There are several avenues I wish to explore in the future. First, it is critical to replicate the interaction we identified in an additional dataset. I am currently examining the feasibility of doing so in either Geisinger or the eMERGE network. Also, I wish to implement statistical approaches designed to pinpoint precisely what element the variant rs16890640 is interacting with to influence MCH and MCV, such as testing combinations of variants within the recombination region for epistasis with rs16890640. Our hypothesis that genetic variants of different continental ancestries may also be investigated more broadly; for instance, instead of focusing on local ancestry transitions on a chromosome, complete genotyping or sequencing data would enable the investigation of differential local ancestry combinations in biological pathways. Similarly, differences in ancestry between the mitochondria or Y-chromosome and autosomal regions may be relevant. Finally, investigation of this hypothesis between three-way admixture populations, such as Hispanics, may improve power to detect effects, as local ancestry transitions are likely to be more frequent.

Here, I propose a mechanism in which genetic variants from different continental ancestries, when combined in admixed populations, result in phenotypic associations not otherwise observed. I identified a specific interaction in which this appears to occur, however this mechanism is general: it could apply to many phenotypes, all admixed populations, and may encompass other sorts of continental ancestry combinations, such as mitochondrial/Y-chromosome with autosomal regions, or regions within biological pathways. It highlights the need to perform genetic ancestry studies within admixed populations (as the variants may not have an effect in either continental population) in order to address health disparities, and the potential role of epistasis in human health.

CHAPTER 4
CONCLUSIONS AND FUTURE DIRECTIONS

In this work, I developed a set of best practices for the study of statistical epistasis, and investigated whether there was evidence for regulatory epistasis in humans in two distinct biological contexts. In Chapter 2, I first developed a set of quality control procedures to identify statistical interactions likely attributable to biological epistasis, rather than confounding processes. I then investigated whether *cis*-regulatory variants interact to regulate gene expression levels, a quantitative low-level phenotype, in cell lines, where environmental factors are kept constant. Once confounding explanations were addressed, I found little evidence for epistasis between *cis*-regulatory variants influencing gene expression levels. In Chapter 3, I investigated epistasis in admixed populations, under the hypothesis that ancestry-specific recombination hotspots may break apart haplotypes, thereby enabling the detection of epistasis. In contrast to molecular phenotypes examined in Chapter 2, the majority of phenotypes I investigated here were complex, EHR-derived phenotypes. I identified a promising interaction between a variant, and a downstream local ancestry transition, that influenced red blood cell traits.

Overall, our results suggest that when rigorous statistical criteria are applied, interactions with a moderate effect size between common, unlinked variants are either uncommon at the population level across a broad range of phenotypes, or are not detectable by our approach. On the surface, this appears to contradict both studies in model organisms, where epistasis is pervasive, and findings from massively parallel reporter assays. Here, I reconcile these findings, and make recommendations for the future study of epistasis based on the consensus of these bodies of literature.

Reconciliation with findings from model organisms

Evidence from model organisms indicates that epistasis accounts for a notable component of phenotypic variance in yeast[1], Drosophila[2,27], and mouse[30] – yet I found little evidence for epistasis in humans. There are several differences in both the genetic architecture of model organisms and in the experimental approach taken by these studies that may both account for this discrepancy and shed light on future approaches for the study of epistasis in humans.

First, there is frequently a fundamental difference in the scientific question being addressed between studies of epistasis in model organisms versus those generally conducted in humans. Studies in model organisms often quantify the overall effect of epistasis across the genome, whereas I sought to identify specific pairs of interacting elements. This is analogous to comparing the amount of phenotypic variance attributable to additive effects in a heritability study to the amount of phenotypic variance explained by variants identified in a GWAS. To provide a more appropriate comparison to model organisms, future studies of epistasis might consider heritability analyses capable of partitioning phenotypic variance into non-additive genetic components. Studies of epistasis in model organisms that seek to identify specific interacting variants are more analogous to human-based studies; in yeast, they typically find that the interactions have smaller effect sizes than those observed for single variants.[1] Bloom et al. found that the largest epistatic effect was approximately a fifth of the size of the largest additive effect for variants influencing gene expression levels.[1] I was underpowered to detect smaller effect sizes in both Chapters 2 and 3 (Tables 2 and 6). Therefore, findings in both bodies of literature would be consistent with large numbers of small-effect interactions that account for the majority of phenotypic variance attributable to epistatic effects.

An alternative, and not necessarily exclusive, explanation is that I investigated epistasis on a different scale than that used in many model organism studies. I specifically investigated epistasis between variants proximal to one another, either within the same *cis*-regulatory region or in specific genomic regions densely genotyped by Metabochip. In contrast, studies of model organisms often quantify epistasis on the chromosomal level,[30] or between genes on different chromosomes.[7,8] Thus, epistasis may be occurring primarily between distant genomic elements, which are frequently referred to as gene by gene interactions. Comprehensive studies of pairwise interactions between such distal variants incur a steep multiple testing correction, and thus have limited power to detect effects in sample sizes typically used in genetic association studies.[80] Careful filtering of the pairs tested based on biological knowledge, such as pathways or protein-protein interactions, may preserve power by reducing the number of associations tested.

Finally, model organisms generally have a more homogenous genetic background than natural populations. This reduces the phenotypic 'noise' attributable to other genetic variants in the genome, and thereby improves power to detect associations. In contrast, the increased genetic diversity within natural populations such as humans results in greater phenotypic variance and reduced power to detect effects. There are population isolates – such as the Amish – that are more homogenous both genetically and culturally than are many other populations. Investigation of epistasis in population isolates may therefore improve power to detect effects.

Ultimately, the conclusions drawn from animal studies and those presented in this work are consistent with one another. And while minimal evidence for epistasis was found in Chapters 2 and 3, the approaches used in model organisms highlight potential ways to move the study of epistasis in humans forward – largely through statistical methods that are capable of capturing aggregate effects, investigation of gene by gene interactions, and/or the usage of population isolates with homogeneous genetic backgrounds.

Reconciliation with findings from massively parallel reporter assays

In addition to model organisms, massively-parallel reporter assays have been used to study epistasis for human regulatory sequences with engineered mutations. Kwasnieski et al. found that the majority of double mutants showed evidence of epistasis influencing the regulatory function of the *Rhodopsin* promoter.[78] This is most closely analogous to the interrogation of epistasis between *cis*-regulatory variants in Chapter 2, and draws markedly different conclusions. Several methodological differences may account for these discrepancies.

First, Kwanieski et al.[78] engineered genetic variation, rather than relying on observed genetic variation within the natural population. Additionally, they only investigated effects within a single *cis*-regulatory sequence. Thus, while their results neatly demonstrate the potential for epistasis within the *cis*-regulatory region, they cannot be taken as a measure of how common such interactions may be. Secondly, they engineered genetic variants within a 52 base pair window, meaning that most variant combinations tested were in very close proximity to one another. In contrast, common genetic variants in human populations are found approximately every 300 base pairs. Even if common variants were in such close proximity to one another, it is unlikely that a recombination event would occur in the limited space between these variants such that they are broken apart. Thus, scenarios described by Kwasnieski et al.[78] would most likely occur within haplotypes in natural populations. As discussed in Chapters 2 and 3, without all possible genotype combinations present regression-based techniques are unable to detect epistasis. Finally, the ability to detect effects within a reporter assay is in no way dependent on the frequency of the alleles within the population. In contrast, the power to detect effects in Chapters 2 and 3 was intimately tied to the frequency of the allele-combinations within the population.

Ultimately, Kwasnieski et al. found that engineered variants within a cis-regulatory region interact with one another the majority of the time; however, their results do not shed light on how frequently such combinations of variants actually occur within the natural population. Additionally, their findings indicate that epistasis within this region is likely to occur within haplotypes, or between rare variants. In these situations, reporter assays may be ideal methodological approaches to detect regulatory epistasis.

Where is epistasis?

We conclude that epistasis is a component of the genetic architecture in humans – the debate is over where it occurs, and how much phenotypic variation it accounts for. I demonstrate in Chapters 2 and 3 that interactions with large effects between common, unlinked variants in proximal regions are likely uncommon. This is consistent with analogous studies in model organisms and is not contradicted by results from massively parallel reporter assays. Based on both our results and these bodies of literature, I propose that epistasis is primarily based on small-effect interactions, or occurs in the following contexts: within haplotypes, as suggested by Kwasnieski et al.[78] and Corradin et al.[126]; between uncommon variant combinations; or between distant regions, as shown repeatedly in model organisms.[7,8,30] These explanations are not inherently exclusive; for example, distal interactions may have small effect sizes. I recommend that future studies of epistasis investigate it within these contexts.

Association-based methods may be ill-suited for future studies of epistasis

Standard association tests that use regression to detect relationships between genotypes and phenotypes rely on genetic diversity within natural populations to detect epistasis, and may be ill-suited to detect epistasis in the above situations. First, they require that all nine possible genotype combinations be represented within the sample in order to accurately partition phenotypic variance amongst genetic components. In the case of both haplotype effects and rare variants, this is unlikely to occur. Secondly, the power to detect effects in such studies is a function of five factors: allele frequencies, LD between variants, effect size, sample size, and the number of association tests performed. Researchers cannot alter either allele frequencies, LD patterns, or the effect size; thus, the only way to improve power is either to increase sample size or to reduce the number of association tests. To comprehensively investigate *trans*-interactions (i.e., gene by gene interactions), the number of association tests increases dramatically, as does the multiple testing correction; Hemani et al. conducted such a study, and faced a multiple testing correction threshold of $2.91\text{x}10^{-16}$.[80] Thus, investigating either small-effect interactions or *trans*- interactions comprehensively will require an increase in sample size above and beyond that required by standard GWAS. While many factors influence the power to detect epistasis, even the best of circumstances (i.e., common, unlinked variants) will require sample sizes of approaching 75,000 to be well-powered to detect modest effect size interactions (Table 9). Thus, the utility of association-based approaches to detect epistasis is limited by its reliance on naturally occurring genetic variation within a population, pervasive confounding influences, and a limited power to detect effects.

MAF Variant 1	MAF Variant 2	Cases	Controls	Power
0.05	0.05	300,000	350,000	0.76
	0.25	70,000	100,000	0.80
	0.5	40,000	110,000	0.81
0.25	0.25	15,000	50,000	0.75
	0.5	11,500	50,000	0.79
0.5	0.5	10,000	50,000	0.80

Table 9. Sample sizes required for adequate power to detect small effect interactions or trans-effects. Here, I estimate the sample sizes required to detect epistasis assuming simplified models (i.e., no linkage between variants) and a Bonferroni multiple testing correction of 1.0×10^{-14}, which is more lenient than the threshold used for a comprehensive analysis of epistasis genome-wide.[80] I assumed only an effect of the interaction (odds ratio = 1.2), without marginal effects of the variants. Approximations were made using the QIMR's Epistasis Power Calculator (https://gump.qimr.edu.au/general/manuelF/epistasis/epipower4i.html).

Alternative approaches to the study of epistasis

Alternative methods are required to identify epistasis in haplotypes, between rare variants, with small effect sizes, or between distant variants. Below, I make recommendations on how to approach the study of epistasis in these contexts.

Detection of epistasis on a haplotype requires that it be broken, such that the effect of each variant individually can be quantified and compared to the joint effect. This can be accomplished by either synthesizing sequences, as was done by Kwasnieski et al.[78], or through genome-editing approaches such as CRISPR. Once the required combinations of genetic variation have been generated, a functional assay is required to measure their effect. In the case of regulatory sequences, this can be readily accomplished through reporter assays. For coding variants within the same gene, epistasis has been quantified through comparing changes in Gibbs's free energy,[65–67] other thermodynamic properties, or the predicted 3D structure. For coding variants in different genes, epistasis between variants within protein-protein interfaces has been biochemically assayed to determine if they disrupt binding.[68] Ultimately, investigation of epistasis within haplotypes will be best accomplished through methods that can create all combinations of the observed variants, and then have a high-throughput assay to measure the phenotypic effect.

Detection of epistasis between rare variants can be best accomplished when power to detect effects is not contingent on the frequency within the population. This can be accomplished through two major methodological approaches: family-based studies and functional assays such as those just described. Family studies would be best suited for the study of epistasis between variants distant from one another, as recombination unlikely to break apart variants in close proximity to one another (i.e., haplotypes). Functional assays that rely on either synthesized or edited sequences, however, could be amenable to either *cis*- or *trans*-interactions.

Thus, the investigation of epistasis involving rare variants is likely not best accomplished within the general population, but rather through the usage of families or high throughput functional assays.

Detection of epistatic interactions with small-effect sizes is mostly impeded by lack of power. There are two possible ways to circumvent this issue – methods that look for the aggregate effect of epistasis genome-wide, rather than individual effects, and methods that improve power. First, methods such as genome-wide complex trait analysis (GCTA) quantify the phenotypic variance attributable to additive genetic effects; however, this method can be expanded to partition phenotypic variance to non-additive effects such as epistasis. This approach would be especially useful, as it would indicate the extent of epistatic effects that could be anticipated under ideal circumstances. Secondly, there are ways in which power can be improved – namely, reducing the number of association tests or changing attributes of the sample. Thus, small-effect epistasis could be investigated using current association-based approaches; however, this would be possible for only a limited number of pairs of variants. By investigating epistasis in population isolates, which have reduced phenotypic noise, small-effect epistasis could be examined more comprehensively. Ideally, I recommend a two-step approach to the investigation of small-effect epistasis: first, quantify the aggregate effect; secondly, perform a targeted analysis of specific pairs of variants.

Finally, detection of epistasis between variants on different chromosomes, i.e. gene by gene interactions, is limited by power to detect effects in the face of numerous association tests. Approaches that are able to use biological knowledge – such as pathways, or protein-protein interactions – to prune down the number of association tests may increase the likelihood of identifying epistasis; however, they are limited by a priori knowledge and cannot address the extent of epistasis. Alternatively, the number of association tests performed could be reduced by collapsing genetic variants into a single variable. For instance, a collapsing or burden score could indicate whether there were any variants within a gene predicted to have deleterious effects.[164–169] Then, all pairwise combinations of genes could be investigated for epistatic effects. This could be applied to a specific phenotype or, pending the availability of EHR data, to a PheWAS. Alternatively, data from chromatin conformation capture (e.g., 3C) approaches could be used to identify genomic regions in physical contact with one another, and epistasis could be investigated between these regions specifically. Overall, the interrogation of trans-effects is unlikely to be able to quantify the prevalence of epistasis genome-wide; however, careful selection of variants may enable the detection of epistasis in specific contexts.

Ultimately, the study of epistasis still has the potential to shed light on the biological mechanisms underlying complex disease in humans; however, it is critical that it be carefully investigated. In Chapter 2, I demonstrate that the quality control procedures sufficient for single-variant association analysis do not address rampant confounding influences. In Chapters 2 and 3, I demonstrate that common, cis-regulatory variants do not interact with one another with effect sizes anticipated based on single-variant analyses. Thus, both the biological expectations and statistical methodologies sufficient for single-marker analyses are not suited for the study of epistasis. Instead, the approaches which I outlined above are more likely to identify epistasis where it likely resides – within haplotypes, between rare variants, in interactions with small effects, or between distant variants.

APPENDIX

A. Nominal PheWAS associations do not interact with local ancestry transitions to influence phenotypes.

I performed a PheWAS to identify potential genetic associations within Metabochip regions not represented in the GWAS Catalog. While no associations passed a Bonferroni multiple-testing correction threshold, 168 variants were nominally significant ($p < 5x10^{-5}$). These associations are reported below. I then examined whether these variants interacted with local ancestry transitions (Methods) to influence the associated phenotype. The results are sorted based on the significance of the interaction p-value. While one interaction passed the Bonferroni multiple testing correction for the interaction analysis, it was driven by a single chromosome (i.e., low cell counts) and was not considered further.

Phecode	*Description*	*Variant*	*Variant p-value*	*Variant*Local ancestry transition p-value*
480.5	Bronchopneumonia and lung abscess	rs17641977_G	2.16E-05	2.42E-06
707.1	Decubitus ulcer	chr16.52362593_G	2.34E-06	7.97E-04
250.13	Type 1 diabetes with ophthalmic manifestations	chr6.25733715_G	1.45E-05	1.51E-03
204.4	Multiple myeloma	chr1.109619786_G	8.62E-06	1.11E-02
614.1	Pelvic peritoneal adhesions, female (postoperative) (postinfection)	chr7.14901079_G	8.18E-06	1.61E-02
440	Atherosclerosis	rs2744238_G	3.86E-05	1.98E-02
276.11	Hyperosmolality and/or hypernatremia	chr1.160348446_A	2.95E-05	2.91E-02
270.3	Disorders of plasma protein metabolism	chr7.14760490_G	1.56E-05	3.20E-02
647.1	Infections of genitourinary tract during pregnancy	chr16.52588027_A	1.43E-06	3.87E-02
331.9	Cerebral degeneration, unspecified	chr13.109701508_G	3.24E-05	4.59E-02
628	Ovarian cyst	chr7.14566856_G	4.18E-05	5.13E-02
578.2	Blood in stool	chr13.109644522_G	3.62E-06	5.39E-02
41.2	Streptococcus infection	chr7.14250501_A	3.65E-05	5.57E-02
669	Complications of labor and delivery NEC	chr6.25453823_G	1.57E-05	6.36E-02
781	Symptoms involving nervous and musculoskeletal systems	chr7.14586312_C	1.17E-05	7.28E-02
592.1	Cystitis	rs10800394_G	1.21E-05	7.44E-02

710.12	Chronic osteomyelitis	chr7.15045801_G	2.29E-05	7.55E-02
427.9	Palpitations	chr13.109782559_G	2.09E-05	8.43E-02
348.8	Encephalopathy, not elsewhere classified	chr11.2692602_A	2.37E-05	9.38E-02
427.9	Palpitations	chr13.109782978_G	5.51E-06	1.03E-01
608	Other disorders of male genital organs	chr7.14334642_G	4.39E-05	1.15E-01
250.14	Type 1 diabetes with neurological manifestations	chr16.52263655_A	4.53E-05	1.18E-01
577	Diseases of pancreas	chr1.160589579_A	4.81E-05	1.22E-01
592	Cystitis and urethritis	rs10800394_G	1.52E-05	1.26E-01
994	Sepsis and SIRS	chr6.25596244_A	2.63E-05	1.28E-01
117.1	Histoplasmosis	chr1.160477578_A	1.49E-05	1.38E-01
539	Bariatric surgery	chr7.14532220_C	7.97E-06	1.55E-01
614.3	Pelvic inflammatory disease (PID)	chr7.14585551_C	6.61E-06	1.55E-01
627.22	Need for Hormone replacement therapy (postmenopausal)	rs12340741_G	2.71E-05	1.57E-01
285.21	Anemia in chronic kidney disease	rs6456688_A	3.55E-05	1.58E-01
704.2	Hirsutism	chr13.109648374_A	3.95E-05	1.63E-01
290.1	Dementias	chr6.26230912_C	1.27E-05	1.66E-01
769	Nonallopathic lesions NEC	chr7.14163174_G	4.16E-05	1.67E-01
601.3	Orchitis and epididymitis	chr7.14548630_A	9.54E-06	1.70E-01
245.2	Chronic thyroiditis	chr13.109632419_A	6.93E-07	1.73E-01
245.21	Chronic lymphocytic thyroiditis	chr13.109632419_A	6.93E-07	1.73E-01
707.1	Decubitus ulcer	chr7.15035730_A	8.89E-06	1.73E-01
740.1	Osteoarthritis; localized	chr13.109679060_G	1.67E-05	1.75E-01
614.54	Abscess or ulceration of vulva	chr7.14337575_G	2.51E-06	1.75E-01
242	Thyrotoxicosis with or without goiter	chr16.52671149_A	1.46E-05	1.78E-01
245	Thyroiditis	chr13.109632419_A	4.70E-05	1.82E-01
513.3	Hypoventilation	chr7.14812267_A	3.79E-05	1.87E-01
251	Other disorders of pancreatic internal secretion	chr7.14820434_G	9.25E-06	1.89E-01
473	Diseases of the larynx and vocal cords	chr13.109805546_A	2.33E-05	2.02E-01
274	Gout and other crystal	chr13.109844573_	2.72E-05	2.03E-01

	arthropathies	C		
769	Nonallopathic lesions NEC	chr7.14162727_G	2.73E-05	2.04E-01
204	Leukemia	chr6.25517293_G	2.33E-05	2.10E-01
149.4	Cancer of larynx	chr13.109693401_C	2.67E-05	2.11E-01
473	Diseases of the larynx and vocal cords	chr13.109803476_A	2.68E-06	2.13E-01
623	Hypertrophy of female genital organs	chr11.2519713_A	3.23E-05	2.14E-01
457	Encounter for long-term (current) use of anticoagulants, antithrombotics, aspirin	chr16.52658019_G	3.69E-05	2.30E-01
614.52	Vaginitis and vulvovaginitis	chr16.52565323_C	1.40E-05	2.39E-01
772.3	Muscle weakness	chr13.109638381_A	1.18E-05	2.42E-01
415.2	Chronic pulmonary heart disease	chr16.52645735_A	8.46E-06	2.43E-01
614.52	Vaginitis and vulvovaginitis	chr11.2754881_G	2.85E-05	2.57E-01
204	Leukemia	chr16.52102494_G	2.54E-05	2.71E-01
526	Diseases of the jaws	chr16.52449474_A	2.39E-05	2.72E-01
696.41	Psoriasis vulgaris	chr13.109715932_A	2.22E-06	2.75E-01
614.5	Inflammatory disease of cervix, vagina, and vulva	chr16.52565323_C	4.73E-05	2.77E-01
245.2	Chronic thyroiditis	rs4721366_A	1.59E-05	2.82E-01
245.21	Chronic lymphocytic thyroiditis	rs4721366_A	1.59E-05	2.82E-01
592	Cystitis and urethritis	chr16.52109683_C	5.82E-06	2.83E-01
245	Thyroiditis	rs4721366_A	4.66E-06	2.89E-01
473.4	Voice disturbance	chr13.109803934_C	4.49E-05	2.92E-01
433.8	Late effects of cerebrovascular disease	chr13.109680753_A	3.94E-05	2.95E-01
613.1	Inflammatory disease of breast	chr7.15001772_A	3.02E-05	2.96E-01
715.1	Sacroiliitis NEC	chr13.109780358_G	1.92E-05	2.98E-01
245.2	Chronic thyroiditis	chr13.109621283_G	1.08E-05	3.00E-01
245.21	Chronic lymphocytic thyroiditis	chr13.109621283_G	1.08E-05	3.00E-01
696.4	Psoriasis	chr13.109715932_A	6.93E-06	3.03E-01
696	Psoriasis and related disorders	chr13.109715932_A	6.42E-06	3.12E-01
592	Cystitis and urethritis	chr1.160539069_A	4.09E-05	3.13E-01
946	Anaphylactic shock NOS	chr7.14390177_G	4.66E-05	3.16E-01

642.1	Preeclampsia and eclampsia	chr16.52560791_G	3.34E-05	3.20E-01
375.1	Dry eyes	chr6.25764640_C	2.38E-05	3.25E-01
334	Degenerative disease of the spinal cord	chr13.109654095_A	3.91E-06	3.57E-01
706.1	Acne	chr6.25374688_A	2.34E-05	3.58E-01
735.21	Hammer toe (acquired)	chr6.25653929_G	2.26E-05	3.59E-01
706.1	Acne	chr6.25381901_T	4.07E-05	3.64E-01
496.3	Bronchiectasis	chr7.14420590_G	7.06E-06	3.65E-01
225.1	Benign neoplasm of brain, cranial nerves, meninges	chr16.52658905_A	6.47E-06	3.83E-01
626.2	Dysmenorrhea	chr7.15031313_A	2.44E-05	3.86E-01
202.2	Non-Hodgkins lymphoma	chr16.52743030_A	2.38E-05	3.89E-01
585.33	Chronic Kidney Disease, Stage III	chr9.135184633_A	1.29E-05	3.93E-01
150	Cancer of esophagus	chr6.25373969_A	6.54E-06	3.97E-01
646	Other complications of pregnancy NEC	chr1.160257615_G	2.72E-05	3.98E-01
990	Effects radiation NOS	chr11.2747784_A	2.94E-06	3.99E-01
245	Thyroiditis	chr7.14820824_A	4.66E-05	3.99E-01
368	Visual disturbances	chr16.52685380_A	4.05E-05	4.02E-01
250.14	Type 1 diabetes with neurological manifestations	chr7.14683390_A	2.27E-05	4.07E-01
225.1	Benign neoplasm of brain, cranial nerves, meninges	chr16.52658156_C	3.68E-05	4.14E-01
614	Inflammatory diseases of female pelvic organs	chr16.52565323_C	3.54E-05	4.18E-01
225	Benign neoplasm of brain and other parts of nervous system	chr16.52658905_A	2.68E-05	4.23E-01
573.2	Liver replaced by transplant	chr1.109762407_G	5.87E-06	4.40E-01
642.1	Preeclampsia and eclampsia	chr16.52566594_G	1.37E-06	4.46E-01
635.2	Antepartum hemorrhage, abruptio placentae, and placenta previa	chr1.160331877_G	4.98E-07	4.51E-01
695.42	Systemic lupus erythematosus	chr9.135125327_A	1.41E-05	4.67E-01
578.2	Blood in stool	chr13.109662438_G	5.27E-06	4.74E-01
707.3	Chronic ulcer of unspecified site	rs9295676_A	1.90E-05	4.86E-01
377.3	Optic neuritis/neuropathy	rs4784323_A	3.39E-05	4.87E-01
745	Pain in joint	chr1.160405816_A	4.82E-05	4.87E-01
170.1	Bone cancer	chr1.160435746_G	4.65E-05	4.88E-01
281.9	Deficiency anemias	rs16835127_G	9.91E-06	4.89E-01
473.4	Voice disturbance	chr13.109803476_A	6.23E-06	4.97E-01
290	Delirium dementia and amnestic and other cognitive disorders	chr6.26230912_C	3.24E-06	5.01E-01

512.7	Shortness of breath	rs4395714_C	4.17E-06	5.04E-01
635	Hemorrhage during pregnancy; childbirth and postpartum	chr1.160331877_G	2.31E-05	5.19E-01
285	Other anemias	chr11.2891879_A	2.78E-05	5.22E-01
202.22	Reticulosarcoma	chr16.52743030_A	4.74E-06	5.23E-01
592.1	Cystitis	chr16.52109683_C	1.10E-05	5.32E-01
513.3	Hypoventilation	chr1.160272821_G	5.56E-07	5.40E-01
681.2	Cellulitis and abscess of face/neck	chr9.135280428_G	2.63E-05	5.45E-01
368.4	Visual field defects	chr6.25504869_A	8.24E-06	5.46E-01
270.32	Paraproteinemia	chr9.135300838_G	1.06E-05	5.48E-01
296.1	Bipolar	chr13.109604690_A	1.31E-05	5.50E-01
593.1	Gross hematuria	chr16.52331386_G	2.58E-05	5.50E-01
772.1	Muscular wasting and disuse atrophy	chr13.109638381_A	1.88E-05	5.52E-01
536	Disorders of function of stomach	chr7.14943393_A	1.70E-05	5.61E-01
379.5	Disorders of iris and ciliary body	chr1.160380841_A	2.88E-05	5.66E-01
446	Polyarteritis nodosa and allied conditions	chr11.2800624_G	1.17E-05	5.72E-01
628	Ovarian cyst	chr13.109650477_A	2.95E-05	5.77E-01
506	Empyema and pneumothorax	rs16835127_G	2.26E-07	5.82E-01
635.2	Antepartum hemorrhage, abruptio placentae, and placenta previa	chr1.160321867_C	2.55E-05	5.93E-01
946	Anaphylactic shock NOS	chr7.15032089_T	5.59E-06	5.93E-01
496.3	Bronchiectasis	chr6.25522825_A	6.37E-06	5.95E-01
285.21	Anemia in chronic kidney disease	chr7.14760490_G	2.36E-05	6.08E-01
697	Sarcoidosis	chr1.160272821_G	2.09E-05	6.12E-01
288.3	Eosinophilia	chr1.109538695_A	3.44E-05	6.32E-01
560.1	Paralytic ileus	chr1.160626500_A	4.25E-05	6.40E-01
473.3	Paralysis/spasm of vocal cords or larynx	chr7.14175664_G	9.28E-06	6.52E-01
735.2	Acquired toe deformities	chr6.25653929_G	4.82E-05	6.53E-01
946	Anaphylactic shock NOS	chr7.15022570_T	2.21E-05	6.73E-01
280.2	Iron deficiency anemia secondary to blood loss (chronic)	chr7.14294180_A	4.33E-05	6.80E-01
540.11	Acute appendicitis	chr16.52557265_G	2.29E-05	6.82E-01
871.2	Open wound of finger(s)	chr1.109509843_A	2.45E-05	7.07E-01
530.1	Esophagitis, GERD and related	chr7.14684017_A	4.20E-05	7.08E-01

	diseases			
352	Disorders of other cranial nerves	chr16.52611296_A	2.71E-05	7.08E-01
357	Inflammatory and toxic neuropathy	chr6.26206734_A	2.82E-05	7.15E-01
348	Other conditions of brain	chr13.109680753_A	4.01E-05	7.30E-01
695.7	Prurigo and Lichen	chr16.52713187_A	2.95E-05	7.34E-01
446	Polyarteritis nodosa and allied conditions	chr11.2797941_A	1.28E-05	7.52E-01
389.2	Conductive hearing loss	chr16.52713187_A	3.26E-06	7.55E-01
721	Spondylosis and allied disorders	chr11.2712577_A	3.19E-05	7.66E-01
819	Skull and face fracture and other intercranial injury	chr7.14351815_G	7.91E-06	7.73E-01
963.1	Antineoplastic and immunosuppressive drugs causing adverse effects	chr1.160491418_G	2.65E-05	7.77E-01
530	Diseases of esophagus	chr7.14684017_A	8.48E-06	7.79E-01
446	Polyarteritis nodosa and allied conditions	chr11.2803003_A	2.83E-05	7.80E-01
579	Other symptoms involving abdomen and pelvis	chr16.52681013_A	3.53E-05	7.83E-01
457	Encounter for long-term (current) use of anticoagulants, antithrombotics, aspirin	chr16.52657870_G	2.59E-05	7.90E-01
727.4	Ganglion and cyst of synovium, tendon, and bursa	chr7.14484502_A	4.15E-05	7.98E-01
367.9	Blindness and low vision	chr13.109829353_A	3.14E-05	8.00E-01
596	Other disorders of bladder	rs739677_A	4.38E-05	8.08E-01
353	Nerve root and plexus disorders	chr13.109783177_G	4.51E-06	8.18E-01
686	Other local infections of skin and subcutaneous tissue	chr9.135044095_G	3.69E-05	8.19E-01
345.3	Convulsions	chr11.2707861_A	2.12E-05	8.21E-01
578.2	Blood in stool	chr1.109570412_G	2.09E-05	8.36E-01
227.3	Benign neoplasm of pituitary gland and craniopharyngeal duct (pouch)	chr13.109840841_A	6.33E-06	8.48E-01
272.11	Hypercholesterolemia	chr6.25835361_A	3.42E-05	8.65E-01
942	Infusion and transfusion reaction	chr7.14266803_C	3.92E-05	8.85E-01
275.6	Hypercalcemia	chr11.2569137_A	3.17E-05	8.92E-01
369	Infection of the eye	chr6.26073531_G	1.16E-05	8.92E-01
942	Infusion and transfusion reaction	chr7.14271227_G	4.43E-05	8.94E-01

761	Cervicalgia	chr1.160257615_G	2.82E-05	9.02E-01
430.2	Intracerebral hemorrhage	chr7.14702689_A	1.96E-05	9.19E-01
743	Osteoporosis, osteopenia and pathological fracture	chr6.25517293_G	3.36E-05	9.19E-01
80	Postoperative infection	chr6.26006590_G	3.55E-05	9.66E-01
942	Infusion and transfusion reaction	chr7.14271721_G	8.60E-06	9.76E-01
362	Other retinal disorders	chr6.25985598_A	4.82E-05	1.00E+00

REFERENCES

1. Bloom, J.S., Kotenko, I., Sadhu, M.J., Treusch, S., Albert, F.W., and Kruglyak, L. (2015). Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nat. Commun. *6*, 8712.

2. Huang, W., Richards, S., Carbone, M.A., Zhu, D., Anholt, R.R.H., Ayroles, J.F., Duncan, L., Jordan, K.W., Lawrence, F., Magwire, M.M., et al. (2012). Epistasis dominates the genetic architecture of Drosophila quantitative traits. Proc. Natl. Acad. Sci. U. S. A. *109*, 15553–15559.

3. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. U. S. A. *109*, 1193–1198.

4. Manolio, T. a, Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L. a, Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

5. Bateson, W. (1909). Mendel’s Principles of Hereditary (Cambridge: Cambridge University Press).

6. Bateson, W., Saunders, E., Punnett, R., and Hurst, C. (1905). Reports to the Evolution Committee of the Royal Society, Report II (London: Harrison and Sons).

7. Phillips, P.C. (2008). Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet *9*, 855–867.

8. Griffiths, A., Miller, J., and Suzuki, D. (2000). An Introduction to Genetic Analysis. (New York: W.H. Freeman),.

9. Fisher, R. (1918). The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinburgh *31*, 151–162.

10. Cordell, H.J. (2002). Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans. Hum. Mol. Genet. *11*, 2463–2468.

11. Wilkening, S., Tekkedil, M.M., Lin, G., Fritsch, E.S., Wei, W., Gagneur, J., Lazinski, D.W., Camilli, A., and Steinmetz, L.M. (2013). Genotyping 1000 yeast strains by next-generation sequencing. BMC Genomics *14*, 90.

12. Botstein, D., and Fink, G.R. (2011). Yeast: An Experimental Organism for 21st Century Biology. Genetics *189*, 695–704.

13. Brem, R.B., and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci. U. S. A. *102*, 1572–1577.

14. Brem, R.B., Storey, J.D., Whittle, J., and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. Nature *436*, 701–703.

15. Borneman, A.R., Leigh-Bell, J. a., Yu, H., Bertone, P., Gerstein, M., and Snyder, M. (2006). Target hub proteins serve as master regulators of development in yeast. Genes Dev. *20*, 435–448.

16. Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. (2002). Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science (80-.). *298*, 799 LP – 804.

17. Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.-B., Reynolds, D.B., Yoo, J., et al. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99–104.

18. Kvon, E.Z., Kazmar, T., Stampfel, G., Yanez-Cuna, J.O., Pagani, M., Schernhuber, K., Dickson, B.J., and Stark, A. (2014). Genome-scale functional characterization of Drosophila

developmental enhancers in vivo. Nature *512*, 91–95.

19. Arnold, C.D., Gerlach, D., Spies, D., Matts, J. a., Sytnikova, Y. a., Pagani, M., Lau, N.C., and Stark, A. (2014). Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nat. Genet. *46*, 685–692.

20. Villar, D., Berthelot, C., Flicek, P., Odom, D.T., Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., and Pignatelli, M. (2015). Enhancer Evolution across 20 Mammalian Species. Cell *160*, 554–566.

21. Ravasi, T., Suzuki, H., Cannistraci, C. V, Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. Cell *140*, 744–752.

22. Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D.M., Groudine, M., Bender, M., Kaul, R., Canfield, T., et al. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. *13*, 418.

23. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

24. Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J. a., Birney, E., et al. (2013). Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. *41*, 827–841.

25. Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

26. Mackay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. Nature *482*, 173–178.

27. Huang, W., Carbone, M.A., Magwire, M.M., Peiffer, J. a., Lyman, R.F., Stone, E. a., Anholt, R.R.H., and Mackay, T.F.C. (2015). Genetic basis of transcriptome diversity in Drosophila melanogaster. Proc. Natl. Acad. Sci. U. S. A. *112*, E6010–E6019.

28. Shorter, J., Couch, C., Huang, W., Carbone, M.A., Peiffer, J., Anholt, R.R.H., and Mackay, T.F.C. (2015). Genetic architecture of natural variation in Drosophila melanogaster aggressive behavior. Proc. Natl. Acad. Sci. *112*, E3555–E3563.

29. Nadeau, J.H., Forejt, J., Takada, T., and Shiroishi, T. (2012). Chromosome substitution strains: gene discovery functional analysis and systems studies. Mamm. Genome *23*, 693–705.

30. Shao, H., Burrage, L.C., Sinasac, D.S., Hill, A.E., Ernest, S.R., O'Brien, W., Courtland, H.-W., Jepsen, K.J., Kirby, A., Kulbokas, E.J., et al. (2008). Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. Proc. Natl. Acad. Sci. *105* , 19910–19914.

31. Brem, R.B., Storey, J.D., Whittle, J., and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. Nature *436*, 701–703.

32. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, 1001–1006.

33. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five Years of GWAS Discovery. Am. J. Hum. Genet. *90*, 7–24.

34. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet *9*, 356–369.

35. Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat Rev Genet *6*, 95–108.
36. Wang, W.Y.S., Barratt, B.J., Clayton, D.G., and Todd, J.A. (2006). GENOME-WIDE ASSOCIATION STUDIES : THEORETICAL AND PRACTICAL CONCERNS.
37. Risch, N.J. (2000). Searching for genetic determinants in the new millennium. Nature *405*, 847–856.
38. Altmu, J., Palmer, L.J., Fischer, G., Scherb, H., and Wjst, M. (2001). Genomewide Scans of Complex Human Diseases : True Linkage Is Hard to Find. 936–950.
39. Bush, W.S., and Haines, J. (2001). Overview of Linkage Analysis in Complex Traits. In Current Protocols in Human Genetics, (John Wiley & Sons, Inc.),.
40. Lander, E.S., and Kruglyak, L. (1995). Genetic Dissection of Complex Traits: Guidelines for Interpreting and Reporting Linkage Analysis. Nat. Genet. *11*, 241–247.
41. Tabor, H.K., Risch, N.J., and Myers, R.M. (2002). practical considerations. *3*, 1–7.
42. Kwon, J.M., Goate, A.M., and Phil, D. (2000). The Candidate Gene Approach. Alcohol Res. Heal. *24*, 164–168.
43. Hindorff, L. a, Sethupathy, P., Junkins, H. a, Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T. a (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. *106*, 9362–9367.
44. Hu, X., and Daly, M. (2012). What have we learned from six years of GWAS in autoimmune diseases, and what is next? Curr. Opin. Immunol. *24*, 571–575.
45. Monteiro, A.N.A., and Freedman, M.L. (2013). Lessons from post-genome-wide association studies: functional analysis of cancer predisposition loci. J. Intern. Med. *274*, 414–424.
46. Gilad, Y., Rifkin, S. a, and Pritchard, J.K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. *24*, 408–415.
47. Michaelson, J.J., Loguercio, S., and Beyer, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). Methods *48*, 265–276.
48. Gaffney, D.J., Veyrieras, J.-B., Degner, J.F., Pique-Regi, R., Pai, A. a, Crawford, G.E., Stephens, M., Gilad, Y., and Pritchard, J.K. (2012). Dissecting the regulatory architecture of gene expression QTLs. Genome Biol. *13*, R7.
49. Powell, J.E., Henders, A.K., McRae, A.F., Wright, M.J., Martin, N.G., Dermitzakis, E.T., Montgomery, G.W., and Visscher, P.M. (2012). Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. Genome Res. *22*, 456–466.
50. Grundberg, E., Small, K.S., Hedman, Å.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.-P., Meduri, E., Barrett, A., et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat. Genet. *44*, 1084–1089.
51. Nica, A.C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., et al. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. *7*, e1002003.
52. Consortium, T.G. (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.
53. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. Nat. Genet. *39*, 1217–1224.
54. Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M., et al. (2012). Patterns of cis regulatory variation in diverse human populations. PLoS Genet. *8*, e1002639.

55. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. *6*, e1000888.
56. Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. PLOS Genet. *6*, e1000895.
57. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science (80-.). *337*, 1190–1195.
58. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.
59. Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K. V, Li, X., Li, H., Kuperwasser, N., Ruda, V.M., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature *466*, 714–719.
60. Krieger, E., Smemo, S., Campos, L.C., Moskowitz, I.P., Pereira, A.C., and Nobrega, M.A. (2012). Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *21*, 3255–3263.
61. Smemo, S., Tena, J.J., Kim, K.-H., Gamazon, E.R., Sakabe, N.J., Gómez-Marín, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature *507*, 371–375.
62. Hsu, A.P., Johnson, K.D., Falcone, E.L., Sanalkumar, R., Sanchez, L., Hickstein, D.D., Cuellar-rodriguez, J., Lemieux, J.E., Zerbe, C.S., Bresnick, E.H., et al. (2017). GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome. *121*, 3830–3838.
63. Lecerf, L., Kavo, A., Ruiz-ferrer, M., Baral, V., Watanabe, Y., Chaoui, A., Pingault, V., Borrego, S., and Bondurand, N. (2013). An Impairment of Long Distance SOX10 Regulatory Elements Underlies Isolated Hirschsprung Disease. 7–11.
64. Dodd, A.W., Syddall, C.M., and Loughlin, J. (2012). A rare variant in the osteoarthritis-associated locus GDF5 is functional and reveals a site that can be manipulated to modulate GDF5 expression. *21*, 517–521.
65. Olson, C.A., Wu, N.C., and Sun, R. (2014). Article A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. Curr. Biol. *24*, 2643–2651.
66. Pollock, D.D., Thiltgen, G., and Goldstein, R.A. (2012). Amino acid coevolution induces an evolutionary Stokes shift. Proc. Natl. Acad. Sci. *109* , E1352–E1359.
67. Risso, V.A., Manssour-Triedo, F., Delgado-Delgado, A., Arco, R., Barroso-delJesus, A., Ingles-Prieto, A., Godoy-Ruiz, R., Gavira, J.A., Gaucher, E.A., Ibarra-Molero, B., et al. (2015). Mutational Studies on Resurrected Ancestral Proteins Reveal Conservation of Site-Specific Amino Acid Preferences throughout Evolutionary History. Mol. Biol. Evol. *32*, 440–455.
68. Podgornaia, A.I., and Laub, M.T. (2015). Pervasive degeneracy and epistasis in a protein-protein interface. Science (80-.). *347*, 673 LP – 677.
69. Anderson, D.W., McKeown, A.N., and Thornton, J.W. (2015). Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. Elife *4*, e07864.
70. Haldane, A., Manhart, M., and Morozov, A. V (2014). Biophysical Fitness Landscapes for

Transcription Factor Binding Sites. PLOS Comput. Biol. *10*, e1003683.
71. Starr, T.N., and Thornton, J.W. (2016). Epistasis in protein evolution. *25*, 1204–1218.
72. Bloom, J.S., Ehrenreich, I.M., Loo, W.T., Lite, T.-L.V., and Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. Nature *494*, 234–237.
73. Young, A.I., and Durbin, R. (2014). Estimation of epistatic variance components and heritability in founder populations and crosses. Genetics *198*, 1405–1416.
74. Sadee, W., Hartmann, K., Seweryn, M., Pietrzak, M., Handelman, S.K., and Rempala, G. a. (2014). Missing heritability of common diseases and treatments outside the protein-coding exome. Hum. Genet. *133*, 1199–1215.
75. Ferraris, L., Stewart, A.P., Kang, J., DeSimone, A.M., Gemberling, M., Tantin, D., and Fairbrother, W.G. (2011). Combinatorial binding of transcription factors in the pluripotency control regions of the genome. Genome Res. *21*, 1055–1064.
76. White, M.A., Kwasnieski, J.C., Myers, C.A., Shen, S.Q., Corbo, J.C., and Cohen, B.A. (2016). A Simple Grammar Defines Activating and Repressing Cis-Regulatory Elements in Photoreceptors. Cell Rep. *17*, 1247–1254.
77. Fiore, C., and Cohen, B. (2016). Interactions between pluripotency factors specify cis-regulation in embryonic stem cells. Genome Res. gr.200733.115.
78. Kwasnieski, J.C., Mogno, I., Myers, C. a, Corbo, J.C., and Cohen, B. a (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc. Natl. Acad. Sci. *109*, 19498–19503.
79. Wei, W.-H., Hemani, G., and Haley, C.S. (2014). Detecting epistasis in human complex traits. Nat. Rev. Genet. *15*, 722–733.
80. Hemani, G., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A.K., McRae, A.F., Yang, J., Gibson, G., Martin, N.G., Metspalu, A., et al. (2014). Detection and replication of epistasis influencing transcription in humans. Nature *508*, 249–253.
81. Brown, A.A., Buil, A., Viñuela, A., Lappalainen, T., Zheng, H.F., Richards, J.B., Small, K.S., Spector, T.D., Dermitzakis, E.T., and Durbin, R. (2014). Genetic interactions affecting human gene expression identified by variance association mapping. Elife *2014*, 1–16.
82. Turner, S., and Bush, W.S. (2011). Multivariate analysis of regulatory snps: empowering personal genomics by considering cis-epistasis and heterogeneity. Pacific Symp. Biocomput. 276–287.
83. Wood, A.R., Tuke, M. a, Nalls, M., Hernandez, D., Singleton, A., Melzer, D., Ferrucci, L., Frayling, T.M., and Weedon, M.N. (2014). An alternative explanation for apparent epistasis. Nature *514*, 1–7.
84. Prabhu, S., and Pe’er, I. (2012). Ultrafast genome-wide scan for SNP–SNP interactions in common complex disease. Genome Res. *22*, 2230–2240.
85. Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N.L.S., and Yu, W. (2010). BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. Am. J. Hum. Genet. *87*, 325–340.
86. Kölsch, H., Lehmann, D.J., Ibrahim-Verbaas, C.A., Combarros, O., van Duijn, C.M., Hammond, N., Belbin, O., Cortina-Borja, M., Lehmann, M.G., Aulchenko, Y.S., et al. (2012). Interaction of insulin and PPAR-α genes in Alzheimer’s disease: the Epistasis Project. J. Neural Transm. *119*, 473–479.
87. Corradin, O., and Scacheri, P.C. (2014). Enhancer variants: evaluating functions in common disease. Genome Med. *6*, 85.
88. Lappalainen, T., Montgomery, S.B., Nica, A.C., and Dermitzakis, E.T. (2011). Epistatic

selection between coding and regulatory variation in human evolution and disease. Am. J. Hum. Genet. *89*, 459–463.

89. Zeng, Z.-B. (2004). Modeling Quantitative Trait Loci and Interpretation of Models. Genetics *169*, 1711–1725.

90. Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J., Wambebe, C., Tishkoff, S.A., et al. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. *107*,.

91. Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., et al. (2011). The landscape of recombination in African Americans. Nature *476*, 170–175.

92. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2017). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. Am. J. Hum. Genet. *93*, 278–288.

93. Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. Bioinformatics *28*, 1359–1367.

94. Paşaniuc, B., Sankararaman, S., Kimmel, G., and Halperin, E. (2009). Inference of locus-specific ancestry in closely related populations. Bioinformatics *25*, 213–221.

95. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLOS Genet. *5*, e1000519.

96. Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat Meth *10*, 5–6.

97. Roden, D.M., Pulley, J.M., Basford, M.A., Bernard, G.R., Clayton, E.W., Balser, J.R., and Masys, D.R. (2008). Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. Clin. Pharmacol. Ther. *84*, 362–369.

98. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics *26*, 1205–1210.

99. Verma, A., Verma, S.S., Pendergrass, S.A., Crawford, D.C., Crosslin, D.R., Kuivaniemi, H., Bush, W.S., Bradford, Y., Kullo, I., Bielinski, S.J., et al. (2016). eMERGE Phenome-Wide Association Study (PheWAS) identifies clinical associations and pleiotropy for stop-gain variants. BMC Med. Genomics *9*, 32.

100. Ye, Z., Mayer, J., Ivacic, L., Zhou, Z., He, M., Schrodi, S.J., Page, D., Brilliant, M.H., and Hebbring, S.J. (2015). Phenome-wide association studies (PheWASs) for functional variants. Eur. J. Hum. Genet. *23*, 523–529.

101. Simonti, C.N., Vernot, B., Bastarache, L., Bottinger, E., Carrell, D.S., Chisholm, R.L., Crosslin, D.R., Hebbring, S.J., Jarvik, G.P., Kullo, I.J., et al. (2016). The phenotypic legacy of admixture between modern humans and Neandertals. Science (80-.). *351*, 737–741.

102. Denny, J.C., Crawford, D.C., Ritchie, M.D., Bielinski, S.J., Basford, M.A., Bradford, Y., Chai, H.S., Bastarache, L., Zuvich, R., Peissig, P., et al. (2011). Variants Near FOXE1 Are Associated with Hypothyroidism and Other Thyroid Conditions: Using Electronic Medical Records for Genome- and Phenome-wide Studies. Am. J. Hum. Genet. *89*, 529–542.

103. Hall, M.A., Verma, A., Brown-Gentry, K.D., Goodloe, R., Boston, J., Wilson, S., McClellan, B., Sutcliffe, C., Dilks, H.H., Gillani, N.B., et al. (2014). Detection of Pleiotropy

through a Phenome-Wide Association Study (PheWAS) of Epidemiologic Data as Part of the Environmental Architecture for Genes Linked to Environment (EAGLE) Study. PLoS Genet. *10*, e1004678.
104. Bush, W.S., Oetjens, M.T., and Crawford, D.C. (2016). Unraveling the human genome-phenome relationship using phenome-wide association studies. Nat. Publ. Gr. *17*, 129–145.
105. Hebbring, S.J. (2014). The challenges, advantages and future of phenome-wide association studies. Immunology *141*, 157–165.
106. McCarty, C.A., Mukesh, B., Giampietro, P., and Wilke, R.A. (2010). Healthy People 2010 disease prevalence in the Marshfield Clinic Personalized Medicine Research Project cohort : opportunities for public health genomic research. *4*, 183–190.
107. Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotech *31*, 1102–1111.
108. Hebbring, S.J., Schrodi, S.J., Ye, Z., Zhou, Z., Page, D., and Brilliant, M.H. (2013). A PheWAS approach in studying HLA-DRB1[ast]1501. Genes Immun *14*, 187–191.
109. Pendergrass, S.A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E.S., Goodloe, R., Ambite, J.L., Avery, C.L., Buyske, S., Bůžková, P., et al. (2013). Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. PLOS Genet. *9*, e1003087.
110. Liao, K.P., Kurreeman, F., Li, G., Duclos, G., Murphy, S., Guzman, R., Cai, T., Gupta, N., Gainer, V., Schur, P., et al. (2013). Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid arthritis cases and non–rheumatoid arthritis controls. Arthritis Rheum. *65*, 571–581.
111. Carroll, R.J., Bastarache, L., and Denny, J.C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics *30*, 2375–2376.
112. Tyler, A.L., Donahue, L.R., Churchill, G. a., and Carter, G.W. (2016). Weak Epistasis Generally Stabilizes Phenotypes in a Mouse Intercross. PLOS Genet. *12*, e1005805.
113. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell *159*, 1665–1680.
114. Schaub, M. a, Boyle, A.P., Kundaje, A., and Frazer, K. a (2012). Linking disease associations with regulatory information in the human genome Toward mapping the biology of the genome. 1748–1759.
115. Kooperberg, C., and LeBlanc, M. (2008). Increasing the power of identifying gene × gene interactions in genome-wide association studies. Genet. Epidemiol. *32*, 255–263.
116. Frazer, K. a, Ballinger, D.G., Cox, D.R., Hinds, D. a, Stuve, L.L., Gibbs, R. a, Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
117. Veyrieras, J.-B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M., and Pritchard, J.K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. *4*, e1000214.
118. Consortium, T. 1000 G.P. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *135*, 0–9.
119. The GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-.). *348*, 648–660.

120. Zou, F., Chai, H.S., Younkin, C.S., Allen, M., Crook, J., Pankratz, V.S., Carrasquillo, M.M., Rowley, C.N., Nair, A. a., Middha, S., et al. (2012). Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. PLoS Genet. *8*,.
121. Cordell, H.J. (2002). Epistasis : what it means , what it doesn ' t mean , and statistical methods to detect it in humans. *11*, 2463–2468.
122. Herold, C., Steffens, M., Brockschmidt, F.F., Baur, M.P., and Becker, T. (2009). INTERSNP: Genome-wide interaction analysis guided by a priori information. Bioinformatics *25*, 3275–3281.
123. Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U. S. A. *100*, 9440–9445.
124. Fish, A.E., Capra, J.A., and Bush, W.S. (2016). Are Interactions between cis-Regulatory Variants Evidence for Biological Epistasis or Statistical Artifacts? Am. J. Hum. Genet. *99*, 817–830.
125. Lewis-Beck, M.S., Bryman, A., and Liao, T.F. The Sage encyclopedia of social science research methods.
126. Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal lari, R., Lupien, M., Markowitz, S., and Scacheri, P.C. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. *24*, 1–13.
127. Becker, J., Wendland, J.R., Haenisch, B., Nöthen, M.M., and Schumacher, J. (2012). A systematic eQTL study of cis-trans epistasis in 210 HapMap individuals. Eur. J. Hum. Genet. EJHG *20*, 97–101.
128. Pluzhnikov, A., Below, J.E., Konkashbaev, A., Tikhomirov, A., Kistner-Griffin, E., Roe, C.A., Nicolae, D.L., and Cox, N.J. (2010). Spoiling the Whole Bunch: Quality Control Aimed at Preserving the Integrity of High-Throughput Genotyping.
129. Laurie, C., Doheny, K., and Mirel, D. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. Genet. … *34*, 591–602.
130. Zuvich, R.L., Armstrong, L.L., Bielinski, S.J., Bradford, Y., Carlson, C.S., Crawford, D.C., Crenshaw, A.T., de Andrade, M., Doheny, K.F., Haines, J.L., et al. (2011). Pitfalls of merging GWAS data: Lessons learned in the eMERGE network and quality control procedures to maintain high data quality. Genet. Epidemiol. *35*, 887–898.
131. Turner, S., Armstrong, L.L., Bradford, Y., Carlson, C.S., Crawford, D.C., Crenshaw, A.T., de Andrade, M., Doheny, K.F., Haines, J.L., Hayes, G., et al. (2011). Quality control procedures for genome-wide association studies. Curr. Protoc. Hum. Genet. *Chapter 1*, 1–18.
132. Gertz, J., Siggia, E.D., and Cohen, B. a (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature *457*, 215–218.
133. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423–428.
134. Grundberg, E., Kwan, T., Ge, B., Lam, K.C.L., Koka, V., Kindmark, A., Mallmin, H., Dias, J., Verlaan, D.J., Ouimet, M., et al. (2009). Population genomics in a disease targeted primary cell model. Genome Res *19*, 1942–1952.
135. Montgomery, S.B., Lappalainen, T., Gutierrez-Arcelus, M., and Dermitzakis, E.T. (2011). Rare and common regulatory variation in population-scale sequenced human genomes. PLoS Genet. *7*, e1002144.
136. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins,

J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The Structure of Haplotype Blocks in the Human Genome. Science (80-.). *296*, 2225–2229.
137. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., and Lander, E.S. (2001). High-resolution haplotype structure in the human genome. *29*, 229–232.
138. Jeffreys, A.J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *29*, 217–222.
139. De La Vega, F.M., Isaac, H., Collins, A., Scafe, C.R., Halldórsson, B. V, Su, X., Lippert, R.A., Wang, Y., Laig-Webster, M., Koehler, R.T., et al. (2005). The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. Genome Res. *15*, 454–462.
140. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199–204.
141. Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). PRDM9 is a major determinant of Meiotic Recombination Hotspots in Human and Mice. Science (80-.). *327*,.
142. Parvanov, E.D., Petkov, P.M., and Paigen, K. (2010). Prdm9 Controls Activation of Mammalian Recombination Hotspots. *327*, 9–10.
143. Myers, S., Bowden, R., Tumain, A., Bontrop, R., Freeman, C., Macfie, T.S., Mcvean, G., and Donnelly, P. (2010). Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. 876–880.
144. Berg, I.L., Neumann, R., Sarbajna, S., Odenthal-hesse, L., Butler, N.J., and Jeffreys, A.J. (2011). Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations.
145. Berg, I.L., Neumann, R., Lam, K.G., Sarbajna, S., May, C.A., and Jeffreys, A.J. (2011). PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *42*, 859–863.
146. Reed, F.A., and Tishkoff, S.A. (2000). African human diversity , origins and migrations.
147. Salas, A., Richards, M., Lareu, M.-V., Scozzari, R., Coppa, A., Torroni, A., Macaulay, V., and Carracedo, Á. (2004). The African Diaspora: Mitochondrial DNA and the Atlantic Slave Trade. Am. J. Hum. Genet. *74*, 454–465.
148. Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., O'Brien, S.J., Altshuler, D., et al. (2004). Methods for High-Density Admixture Mapping of Disease Genes. Am. J. Hum. Genet. *74*, 979–1000.
149. Smith, M.W., Patterson, N., Lautenberger, J.A., Truelove, A.L., McDonald, G.J., Waliszewska, A., Kessing, B.D., Malasky, M.J., Scafe, C., Le, E., et al. (2004). A High-Density Admixture Map for Disease Gene Discovery in African Americans. Am. J. Hum. Genet. *74*, 1001–1013.
150. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J., Doumbo, O., et al. (2009). The Genetic Structure and History of.
151. Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (2012). The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. PLoS Genet. *8*, 1–12.
152. Matise, T.C., Ambite, J.L., Buyske, S., Carlson, C.S., Cole, S.A., Crawford, D.C., Haiman, C.A., Heiss, G., Kooperberg, C., Marchand, L. Le, et al. (2011). The Next PAGE in

Understanding Complex Traits: Design for the Analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. Am. J. Epidemiol. *174*, 849–859.
153. Gauderman, W.J. (2002). Sample size requirements for matched case-control studies of gene-environment interaction. Stat. Med. *21*, 35–50.
154. Haiman, C.A., Fesinmeyer, M.D., Spencer, K.L., Bůžková, P., Voruganti, V.S., Wan, P., Haessler, J., Franceschini, N., Monroe, K.R., Howard, B. V, et al. (2012). Consistent Directions of Effect for Established Type 2 Diabetes Risk Variants Across Populations: The Population Architecture using Genomics and Epidemiology (PAGE) Consortium. Diabetes *61*, 1642–1647.
155. Onodera, K., Shavit, J.A., Motohashi, H., Yamamoto, M., and Engel, J.D. (2000). Perinatal synthetic lethality and hematopoietic defects in compound mafG::mafK mutant mice. EMBO J. *19*, 1335–1345.
156. Arbeithuber, B., Betancourt, A.J., Ebner, T., and Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proc. Natl. Acad. Sci. *112* , 2109–2114.
157. Drakesmith, H., Sweetland, E., Schimanski, L., Edwards, J., Cowley, D., Ashraf, M., Bastin, J., and Townsend, A.R.M. (2002). The hemochromatosis protein HFE inhibits iron export from macrophages. Proc. Natl. Acad. Sci. *99* , 15602–15607.
158. Zhou, X.Y., Tomatsu, S., Fleming, R.E., Parkkila, S., Waheed, A., Jiang, J., Fei, Y., Brunt, E.M., Ruddy, D.A., Prass, C.E., et al. (1998). HFE gene knockout produces mouse model of hereditary hemochromatosis. Proc. Natl. Acad. Sci. *95* , 2492–2497.
159. Feder, J.N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D.A., Basava, A., Dormishian, F., Domingo, R., Ellis, M.C., Fullan, A., et al. (1996). A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet *13*, 399–408.
160. Franceschini, N., Fox, E., Zhang, Z., Edwards, T.L., Nalls, M.A., Sung, Y.J., Tayo, B.O., Sun, Y. V, Gottesman, O., Adeyemo, A., et al. (2013). Genome-wide Association Analysis of Blood-Pressure Traits in African-Ancestry Individuals Reveals Common Associated Genes in African and Non-African Populations. 545–554.
161. Carlson, C.S., Matise, T.C., North, K.E., Haiman, C.A., Fesinmeyer, M.D., Buyske, S., Schumacher, F.R., Peters, U., Franceschini, N., Ritchie, M.D., et al. (2013). Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry : The PAGE Study. *11*,.
162. Gong, J., Schumacher, F., Lim, U., Hindorff, L. a., Haessler, J., Buyske, S., Carlson, C.S., Rosse, S., Bůžková, P., Fornage, M., et al. (2013). Fine mapping and identification of BMI loci in African Americans. Am. J. Hum. Genet. *93*, 661–671.
163. Jeff, J.M., Ritchie, M.D., Denny, J.C., Kho, A.N., Ramirez, A.H., Crosslin, D., Armstrong, L., Basford, M.A., Wolf, W.A., Pacheco, J.A., et al. Generalization of Variants Identified by Genome-Wide Association Studies for Electrocardiographic Traits in African Americans. 321–332.
164. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.
165. Morgenthaler, S., and Thilly, W.G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). Mutat. Res. Mol. Mech. Mutagen. *615*, 28–56.
166. Morris, A.P., and Zeggini, E. (2010). An Evaluation of Statistical Approaches to Rare Variant Analysis in Genetic Association Studies. Genet. Epidemiol. *34*, 188–193.
167. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations

using a weighted sum statistic. PLoS Genet. *5*, e1000384.
168. Li, B., and Leal, S.M. (2017). Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. Am. J. Hum. Genet. *83*, 311–321.
169. Ionita-Laza, I., Buxbaum, J.D., Laird, N.M., and Lange, C. (2011). A New Testing Strategy to Identify Rare Variants with Either Risk or Protective Effect on Disease. PLoS Genet. *7*, e1001289.