Title
Methods and Models for the Analysis of Human Genetic Data

Author
Brown, Robert Paul

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Methods and Models for the Analysis of Human Genetic Data

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

by

Robert Brown

2017

ABSTRACT OF THE DISSERTATION

Methods and Models
for the Analysis of Human Genetic Data

by

Robert Brown
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2017
Professor Bogdan Pasaniuc, Chair

The advent of time- and cost-effective technologies for genotyping and sequencing human DNA has massively increased both the type and amount of genetic data available for study. In order to best utilize this data, new methods must be developed to better assess how human history affects genetics and how genetics affects human phenotypes such as height, eye color and disease risk.

This work presents five new methods that build upon each other to address this challenge. The first method leverages geographic information contained in rare genetic variation to infer the genetic ancestry of individuals at each location in the genome. It increases ancestry inference accuracy when applied to cohorts of continentally admixed individuals. This method also allows inference of local ancestry when studying cohorts containing subcontinentally admixed individuals.

The second method applies the idea of highly structured geographic information in rare variation to create a better variant filtering approach for finding the causal variation in monogenic disorders. By finding better estimates of allele frequencies both within and across populations, it reduces the number of variants that must be considered as potentially disease causing. This results in decreased time and cost expenditures in necessary follow-up analyses.

Due to multiple testing issues, compound heterozygous architectures and haplotype affects are difficult to detect as contributing to complex diseases or gene regulation. The next

two methods present ways to detect these complex features. Compared to standard marginal association approaches, these two methods show that compound heterozygous architectures and haplotype effect models often better capture the genetic contributions to traits. The results demonstrate the need for future fine-mapping approaches that seek complex causal architectures.

The final method in this work searches for causal relationships in gene expression networks. These networks are formed by genes with highly correlated expression levels. However, the correlation may be due to unobserved confounding variables. By utilizing genetic variants as instrumental variables, this method finds causal gene-on-gene effects. Knowing the direction and magnitude of gene-on-gene effects is vital to better understanding regulatory networks in disease pathways and for the identification of drug targets.

The dissertation of Robert Brown is approved.

Rita Cantor

Eleazar Eskin

Kirk Lohmueller

Janet Sinsheimer

Bogdan Pasaniuc, Committee Chair

University of California, Los Angeles

2017

To my family

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost I would like to acknowledge my advisor, Bogdan Pasaniuc, who joined his lab after I did. His encouragement and guidance were invaluable as I carried out my research, learned to be an effective communicator and developed the skills necessary to succeed in an academic environment. I sincerely appreciate the patience and thought he applied to all of our interactions. He provided valuable structure and advice while allowing me freedom to explore the ideas relating to my work.

Secondly I would like to acknowledge Eleazar Eskin. Prior to being co-advised by him in my last year, I had many opportunities to work with him on non-research endeavors. With his help, Ashley Cass, myself and others were able to found the Bioinformatics Student Retreat, participate in planning the RECOMB-Genetics satellite meeting in Santa Monica and carry out other initiatives. As an advisor, Eleazar has also been an outstanding source of guidance and helped me become a more efficient researcher.

For many years I was fortunate to have Kirk Lohmueller and his group in joint lab meetings. His feedback and ideas on projects have shaped many of my research questions and strongly contributed to the population genetics lean of much of my work.

I am grateful to all the members of my thesis committee: Rita Cantor, Eleazar Eskin, Kirk Lohmueller, Janet Sinsheimer and Bogdan Pasaniuc. Both before and after forming my committee I have enjoyed formal classwork and advising with them as well as informal hallway and seminar conversations. They have always taken a keen interested in my work and ideas and their encouragement has meant a great deal to me.

I have had the privilege of being surrounded by outstanding labmates in Bogdan's group. While Gleb Kichaev, Huwenbo Shi, Nicholas Mancuso and James Bookcock have been co-authors on my papers, that does not mean that I do not thoroughly appreciate the hours of lab meetings, white wall working, lunch conversations and lab socials I have had with my other labmates that have contributed to my time at UCLA. So a special shout out to all my

other labmates in Bogdan's group: Wen-un Yang, Kathy Burch, Claudia Giambartolomei, Pagé Goddard, Malika Kumar, Ruthie Johnson, Megan Roytman and Valarie Arboleda.

While my time in Eleazar's lab has been short, I would also like to give a special thanks to Robert Smith who has helped me settle in to the lab and strongly supported all of the projects I have worked on. Lana Martin who has always been there to provide great conversation in addition to her outstanding editorial comments on my manuscripts. Michael Bilow who has designed an impressive project interface for a class that we both TA and works with me on my current project. They and others in the lab have helped make my time there enjoyable and productive and I am excited for our future collaborations.

I also want to give a special thanks to Petko Fiziev and Larry Lam and Chelsea Ju. While we don't see each other as often as we did our first and second years, I always enjoy when we happen to all be together.

One of my favorite parts about Bioinformatics at UCLA is the active participation in the program, especially form Ashley Cass who was a co-organizer and founder with me of the Bioinformatics Student Retreat. Many others have contributed to making the UCLA program strong through their leadership and I have greatly benefited from their efforts. I apologize if I miss any of you, but thank you for your help Artur Jarosziwicz, Rebecca Walker, Kimberly Ensigne, Larry Lam, Adriana Sperlea, Christopher Hartl, Alden Huang, Kikuye Koyano and Ivette Zelaya.

A special thanks to Pamela Hurley, who's knowledge of the inner-UC workings is next to none. I am grateful for her help and support both for me personally and in helping the students succeed in our many initiatives.

Half-way through my graduate studies I met my wonderful wife Rebecca! I have loved being able to tell her about my projects and work. I am grateful for the interest that she shows, the support she provides, the delicious dinners she cooks me and the fun we have together. Having her by my side has helped me through the tough parts and made the good parts so much better.

Finally, I would like to thank my mother, father, brother and sister for their patience,

love and consistent support. It has helped make this journey enjoyable and unforgettable. I am forever grateful to them all.

This work was partially supported through the Systems and Integrative Biology Training Program (NIH T32-GM008185) and the Genome Analysis Training Program (NIH T32-HG002536). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the dissertation or published works below.

Chapter 2 was published in PLoS computational biology, 10(4):e1003555, April 2014. Robert Brown and Bogdan Pasaniuc. "Enhanced methods for local ancestry assignment in sequenced admixed individuals." The dissertation author was the primary investigator and author of this paper.

Chapter 3 was published in European Journal of Human Genetics, 24(1):113119, 2016. Robert Brown, Hane Lee, Ascia Eskin, Gleb Kichaev, Kirk E Lohmueller, Bruno Reversade, Stanley F Nelson, and Bogdan Pasaniuc. "Leveraging ancestry to improve causal variant identification in exome sequencing for monogenic disorders." The dissertation author was the primary investigator and author of this paper.

Chapter 4 was published in Bioinformatics (Oxford, England), 2017. Robert Brown, Gleb Kichaev, Nicholas Mancuso, James Boocock, and Bogdan Pasaniuc. "Enhanced methods to detect haplotypic effects on gene expression." The dissertation author was the primary investigator and author of this paper.

Chapter 5 is currently prepared for submission. Robert Brown, Eleazar Eskin and Bogdan Pasaniuc, "Haplotype-based eQTL Mapping Increases Power to Idenitfy eGenes" The dissertation author was the primary investigator and author of this paper.

Chapter 6 is currently prepared for submission. Robert Brown, Jong Wha J. Joo, Robert Smith, Bogdan Pasaniuc and Eleazar Eskin. "GGmend: A Mendelian randomization method for finding gene-on-gene regulatory effects in the presence of unobserved confounders." The dissertation author is joint-first author with Jong Wha J. Joo. Joo formulated the question and causal model while the dissertation author developed the statistical method to estimate effect size, assess significance and ran all simulations and real data analyses.

Vita

2009	B.A. (Astrophysics), Columbia University, New York, New York.
2009-2010	Middle School Math Teacher, Soldier Hollow Charter School, Midway, UT.
2010-2011	Researcher, University of Michigan, Ann Arbor, MI.
2011–Present	Graduate Student, Bioinformatics IDP, University of California, Los Angeles, CA.
2012-2013	Systems and Integrative Biology Training Grant, University of California, Los Angeles, CA.
2014-2016	Genomics Analysis Training Program, University of California, Los Angeles, CA.
Winter 2016	Teaching Assistant, Life Sciences Department, University of California, Los Angeles, CA.
Winter 2017	Teaching Assistant, Computer Science Department, University of California, Los Angeles, CA.
Spring 2017	Teaching Assistant, Computer Science Department, University of California, Los Angeles, CA.

Publications

Robert Brown and Bogdan Pasaniuc. "Enhanced methods for local ancestry assignment in sequenced admixed individuals." PLoS computational biology, 10(4):e1003555, April 2014.

Robert Brown, Hane Lee, Ascia Eskin, Gleb Kichaev, Kirk E Lohmueller, Bruno Reversade, Stanley F Nelson, and Bogdan Pasaniuc. "Leveraging ancestry to improve causal variant identification in exome sequencing for monogenic disorders." European Journal of Human Genetics, 24(1):113119, 2016.

Robert Brown, Gleb Kichaev, Nicholas Mancuso, James Boocock, and Bogdan Pasaniuc. "Enhanced methods to detect haplotypic effects on gene expression." Bioinformatics (Oxford, England), 2017.

Robert Brown, Eleazar Eskin and Bogdan Pasaniuc, "Haplotype-based eQTL Mapping Increases Power to Idenitfy eGenes." In preparation.

Robert Brown, Jong Wha J. Joo, Robert Smith, Bogdan Pasaniuc and Eleazar Eskin. "GGmend: A Mendelian randomization method for finding gene-on-gene regulatory effects in the presence of unobserved confounders." In preparation.

CHAPTER 1

Introduction

Essentially inaccessible a few decades ago at large scale, the three billion base pair long string of human deoxyribonucleic acid (DNA) can now be read in both a time-efficient and cost-efficient manner. Composed of four possible nucleotide base pairs represented by 'A', 'C','G' and 'T', each individual's DNA is unique and responsible for a proportion of phenotypes (traits such as height, eye color or disease risk) that are inherited from parents. With the advent of high throughput technologies, researchers can collect and study the DNA from many individuals in order to gain insights and understanding into how populations grow and evolve, the risk factors for diseases and the inner regulatory networks that control growth and development.

In humans, each individual carries two copies of the human genome. The copies are broken into chromosomes. There are 22 non-sex chromosomes, implying that each individual carries 44: two copies of chromosome 1, two of chromosome 2 and so forth. The 3 billion base pair long string of DNA that makes up the human genome is divided unevenly among the 22 chromosomes. Researchers are most interested in where and why the genome differs between individuals and how these differences influence human traits.

In order to quantify these differences, researchers use a 'reference genome.' The reference genome is simply a universally accessible genome that allows comparison with the genome of any individual; locations with differences are annotated as genetic variation. After surveying many individuals, it is possible to quantify in a population the frequency of differences from the reference genome at each nucleotide. Locations where there are individuals with differences from the reference genome are referred to as single nucleotide polymorphisms (SNPs). The nucleotides (which we will refer to as alleles) that match the allele on the

reference genome are referred to as 'reference alleles,' and those that do not match are called 'alternate alleles'. At any given SNP location, the vast majority of individuals carry either the reference allele or the same alternate allele.

Since each individual has two copies of their DNA, at each SNP position they can either carry two reference alleles, one reference and one alternate allele, or two alternate alleles. The genotype of an individual at a SNP position is normally defined as the number of alternate alleles an individual carries at that position. Interestingly, analysis of SNPs show that, relative to the size of the genome, there are very few genetic differences between any two randomly chosen individuals. In other words, we are more genetically similar than we are different.

While genotypes at SNPs are the most basic unit of genetic data, it is important to understand their arrangement in the human genome. For example, when looking at two adjacent SNP positions on chromosome 1, if an individual has one reference and one alternate allele at each position, there are four possible arrangements of the alleles: 1) both alternate alleles can be on the first copy of chromosome 1, 2)both alternate alleles are on the second copy of chromosome 1, 3) the first copy of the chromosome has the first alternate allele and the second copy carries the second alternate allele or 4) the second copy of the chromosome carries the first alternate allele and the first copy carries the second. These possible arrangements are referred to as haplotypes. While the example is of a 2-SNP haplotype, haplotypes can be defined by any number of SNPs occurring adjacently on the same chromosome.

When humans reproduce, they pass on chunks of DNA from each chromosome. In other words, haplotypes, the specific arrangement of reference and alternate alleles in the parent, are partially passed on to children. Because the arrangements of SNPs are also inherited, this induces correlations between observed genotype values at nearby SNPs. This correlation is referred to as linkage disequilibrium.

SNPs, haplotypes and linkage disequilibrium together make up the fundamental building blocks of information that scientists use to understand the relationship between genetics, phenotypes and human history. The following chapters will leverage these three sources of

information in order build better models and methods for understanding population structure, finding alleles that cause monogenic disorders, identifying the genetic variation that regulates gene expression and for understanding how the expression from one gene can affect the expression of another.

Chapter 2 investigates the differences in SNP frequency distributions between populations. There are a large number of SNPs where alternate alleles are observed in one population, such as Europeans, but not in any others, such as Asians and Africans (these SNPs can be referred to as continental specific variants (CSVs). CSVs tend to have either the alternate or the reference allele being rare in the population where it is observed. This is due to population structure, where there has not been enough time for the CSV to move to other populations through migration and drift. The method developed in this chapter recognizes that, while rare, observed CSVs contain a significant amount of geographic information. The method uses this potentially useful information in order to infer the continental (or subcontinental) location from which each location in an individual's genome originates. Identifying the local ancestry of genomic regions is especially important when considering admixed individuals such as African Americans, who have genetic ancestry form multiple continental populations. This information is important for finding risk locations for diseases, understanding genetic recombination and inferring human demographic events. The method is published in *PloS Computational Biology* [18].

Building off insights from Chapter 2, Chapter 3 presents an improved method for identifying genetic variants that cause monogenic disorders such as sickle cell anemia, hemophilia or color blindness. While these are common examples, the majority of monogenic disorders are extremely rare affecting at most a few individuals each year across the globe. Since they rarely occur in the general population, the disease causing alleles must also be rare in all populations. Standard approaches that try to narrow down the list of potential causal variants therefore filter out any SNP from consideration that is observed above a defined frequency such as 0.1%. However, standard approaches calculate this frequency across continental populations, such as across all Europeans. Because rare variation is highly structured even within a population, averaging across continental populations downwardly biases frequency

estimates resulting in higher false discovery rates. The method presented in this chapter corrects for this bias by more accurately defining populations for estimating the allele frequencies. The improved frequency estimates result in more efficient variant filtering. This method is published in the *European Journal of Human Genetics* [17].

One of the causal architectures in monogenic disorders is called a compound heterozygote. In such an architecture, the function of a gene in both copies of a chromosome is disrupted, but the disruption occurs at different SNP locations. While this architecture is common in monogenic disorders, it is difficult to find in complex disease studies and gene expression studies. There is no way to know *a priori* which SNPs together form a compound heterozygote. Chapter 4 presents a method to search for compound heterozygotes in gene expression data. The results indicate that compound heterozygous architectures, in many cases, better tag the underlying and unknown true regulatory architecture. The method and results of this study are published in *Bioinformatics* [16].

The compound heterozygous architecture is a special case of a more general haplotype effect architecture model. Chapter 5 lays out the method for the more general haplotype-based approach that seeks genetic regions regulating gene expression. By looking for associations between gene expression and all possible sets of haplotypes in 10,000 base pair regions of the genome, the method is able to find more and stronger associations, even though there is a much higher significance threshold due to the multiple testing correction. The results indicate that many genes are regulated by architectures that are not significantly correlated with any marginal SNP, rather, these genes have significant associations with specific arrangements of SNPs. This result is very important; it demonstrates that future fine-mapping methods need to incorporate complex haplotype-based architectures into their models when trying to identify the true causal genetic variation. This work is prepared for submission to the RECOMB-Genetics 2017 Conference in Los Angeles.

The final chapter (6), is unique. In all previous chapters, SNPs are used specifically because they contribute to disease risk, gene regulation or contain geographic information. In this Chapter 6, SNPs are only important as a tool. The genotype of an individual is both random and fixed. While an individual's genotypes can influence their phenotypes,

the phenotypes cannot influence their genotypes. This characteristic allows the SNPs to solve the question of correlation versus causation when looking at gene expression networks. Many genes have highly correlated expression levels due to unobserved confounding variables. Here, the central question is whether the expression from one gene causally affects gene expression levels from another gene. This chapter shows that leveraging SNPs as instrumental variables can determine if there is causation, in addition to correlation, within gene expression networks. This work is prepared for submission to the RECOMB-Genetics 2017 Conference in Los Angeles.

CHAPTER 2

Enhanced Methods for Local Ancestry Assignment in Sequenced Admixed Individuals

2.1 Introduction

Advances in high-throughput genotyping technologies have enabled large-scale studies of genetic variation, from genome-wide association studies (GWAS) [78] to inference of population history from genetic data[139]. The most notable use of high-throughput genotyping has been in GWAS where researchers have reproducibly identified thousands of genetic variants associated with many diseases[26]. Although initial studies have focused on homogenous populations[27], the development of accurate methods for discerning population structure has enabled studies across individuals of different ethnicities such as admixed populations (i.e. populations with genetic ancestry from more than one continent)[157, 144, 86, 167]. Owing to their recent demographic history, admixed individuals have genomes that are a mosaic of segments originating from different continents. A key component of genetic studies in recently admixed populations is the inference of ancestry at each locus in the genome (i.e. the continental origin of each variant, local ancestry). Although local ancestry has been traditionally used to map genes to diseases through admixture mapping[31, 170, 167], the past few years have seen the use of local ancestry analyses in a wide range of genetic applications. Recent work has shown that admixture mapping can be used to localize missing

The work appearing in this chapter is published: Robert Brown and Bogdan Pasaniuc. "Enhanced methods for local ancestry assignment in sequenced admixed individuals." PLoS computational biology, 10(4):e1003555, April 2014.

sequences from the human reference genome[57], while other analyses of local ancestry in large samples of African American individuals have yielded novel insights into the dynamics of recombination rates across the genome[77, 194]. Local ancestry also can be leveraged to make demographic inferences from genetic data of admixed populations[83, 22, 90, 64] as well as in finding signals of natural selection in African Americans[82]. Finally, local ancestry is also important for disease genetics in correcting for spurious associations in fine-mapping studies[156] as well as in finding new disease risk loci through a combination of association and admixture mapping[144, 171, 205, 51, 198].

Many methods have been developed to infer local ancestry in admixed individuals. Early methods[122, 126, 146] relied on ancestry informative markers within hidden Markov models to achieve high accuracy. With decreasing genotyping costs, newer methods[13, 142, 163, 182] were designed to use the increasing amount of data from genome-wide genotyping arrays while accounting for linkage disequilibrium (LD) among variants. The currently established methods[4, 154, 181] model LD in the form of haplotypes to achieve superior accuracy over non-haplotype aware approaches. Recent work in parallel to ours[119] explored the use of conditional random forests in performing local ancestry analysis. Although extremely accurate for African Americans, these methods have not achieved the same level of high accuracy in Latino Americans, partially due to the lack of good proxies for the Native American component[143] and more recent divergence among ancestral populations. Rapid cost decreases in sequencing technologies coupled with the increased power for assessing genetic variation has made sequencing the approach of choice for many of the coming genetic studies[73, 3, 8, 38, 43, 62, 72, 125, 130, 145, 158, 178, 186]. The amount of variants identified by sequencing makes local ancestry inference in large cohorts of sequenced individuals prohibitively time consuming (e.g. existing HMM-based approaches will take 5 CPU years to infer local ancestry in 15,000 sequenced African Americans, or 18 days per core on a 100-core cluster). This is particularly important as sample sizes continue to increase to hundreds of thousands of individuals. For example, a recent study of obesity included over 15,000 African Americans[30] and another study included 30,000 African Americans for recombination mapping[77].

Here we present improved methods for local ancestry inference for fully sequenced admixed genomes. Sequencing, as opposed to genotyping, is able to catalogue much larger sets of variants with a large component of such variants being continent-specific (i.e. variants that are observed only in individuals from one continental group such as Europeans or Africans). For example, the 1000 Genomes Project[39] has found that 17% of variants with frequencies between 0.5-5% and 53% of variants with frequencies <0.5% are continent-specific when comparing European, African, East Asian and American populations. We hypothesized that these variants can be used for ultra-fast assignment of ancestry at every locus in the genome. We term these variants as continent-specific variants (CSVs) and model them within standard hidden Markov models of local ancestry to achieve an accurate and computationally efficient method for local ancestry inference (Lanc-CSV). Our model accounts for potential errors induced by low-coverage sequencing as well as by the finite sample size of the reference panels used for local ancestry inference. As opposed to most previous local ancestry methods that require phased reference panels, our approach only requires allele frequency information for each continental group.

Our approach is significantly faster than existing standard haplotype-based approaches making it the approach of choice for large-scale sequencing studies (e.g. our approach is able to infer local ancestry in under 42 CPU days in 15,000 sequenced genomes, or 0.42 days per core if a 100-core cluster is available). The very-fast computational speed of our approach allows it to be sample aware by iteratively improving the quality of the CSV calls using the admixed individuals themselves to further boost accuracy by eliminating spuriously identified CSVs. We use simulations of recently admixed individuals starting from 1000 Genomes data to show that Lanc-CSV achieves comparable accuracy to existing methods (e.g. mean $r^2 = 0.92$ across simulations of African Americans, Mexicans, and Puerto Ricans as compared to 0.93 for LAMP-LD[4], 0.84 for RFMix[119] and 0.80 for MULTIMIX[36]).

We investigate the effect of low coverage sequencing on our method in simulations and show that at 5x coverage our approach achieves an $r^2 = 0.86$ in African Americans, 0.70 in Mexicans and 0.78 in Puerto Ricans. More importantly, we investigate whether similar results can be obtained in real data. We infer local ancestry using our approach in the real

African American, Mexican, and Puerto Rican individuals from 1000 Genomes and find that Lanc-CSV agrees with the published consensus local ancestry calls (mean $r^2 = 0.79$ across the three sets of comparisons as compared to a mean $r^2 = 0.81$ for a haplotype-based method, see Results). While our current method achieves comparable results to existing methods with the given data sets, we demonstrate that the iterative sample aware CSV updating continues to increase the overall accuracy as the sample size increases. With large studies this may give Lanc-CSV a further accuracy advantage over existing methods. Finally, we extend the concept of CSVs to sub-continental population-specific variants (sCSVs) and show that they can be used to perform ancestry assignment with individuals admixed from two ancestries from the same continent.

As the costs of sequencing rapidly decreases and genetic studies sequence more samples, the tradeoff between computational runtime and accuracy becomes critical for local ancestry inference. Using our proposed approaches we can reliably infer local ancestry in very large sequenced cohorts at a fraction of the computational cost of existing approaches.

2.2 Methods

2.2.1 Data and simulations

The 1000 Genomes Project[39] has produced a public catalog of human genetic variation through sequencing in individuals from populations across the world. In this work we use the 88 Yoruba (YRI) and 97 Luhya (LWK) individuals as proxy for the African haplotypes; the 85 Utah residents with northern and western European ancestry (CEU) and 97 Tuscans in Italy (TSI) individuals were used as proxy for the European haplotypes; the 88 Japanese in Tokyo (JPT), 97 Han Chinese in Beijing (CHB) and 100 Southern Han Chinese (CHS) individuals were used as a proxy for the Native American haplotypes. The 14 Iberian populations in Spain (IBS), 93 Finnish in Finland (FIN) and 89 British in England and Scotland (GBR) individuals are also used for determining sCSVs. We used the 1000 Genomes phased haplotypes from each individual. We restricted our analysis to chromosome 10. The TSI,

JPT and LWK haplotypes were used as training haplotypes for CSVs and all of the CEU, CHB+CHS and YRI haplotypes were used as simulation haplotypes so that the training and simulation haplotypes would be disjoint and unmatched. Following previous works we filtered A/T and C/G variants from the analysis[143] leaving 1,581,313 (50,000) SNPs used for sequencing (array) simulations.

Similar to previous works[154], we simulate admixed chromosomes as a random walk over the 1000 Genomes haplotypes. Distance to the next crossover is sampled from an exponential distribution with parameter $(G\lambda)^{-1}$ where $\lambda = 10^{-8}$ base pairs per generation and G is the number of generations since admixture[4, 154]. At a crossover event, an ancestry (i.e. continental group) is chosen according to admixture-specific proportions and a random haplotype is drawn uniformly from that continental group. We simulate 2000 haplotypes this way and paired them to form 1000 genotypes with no simulated haplotype used more than once. We used the following admixture proportions (θ) for the European, Native American and African ancestry: 0.45:0.5:0.05 for Mexicans and 0.67:0.13:0.2 for Puerto Ricans and 0.2:0.0:0.8 for African-Americans[118, 24, 153, 185]. For African Americans we simulated data assuming 6 generations since admixture ($G = 6$) and for Mexicans and Puerto Ricans we assumed 15 generations ($G = 15$).

2.2.2 Continent-specific variants in the 1000 Genomes data

Comparing sequenced samples from different continental groups identifies continent-specific variants. A CSV is identified if the reference or alternate allele is observed in only one of the continental groups being compared. A CSV is only informative of an individuals ancestry if it is observed in that individual. We used the reference panels to estimate CSVs and then identified how many European, Native American, and African CSVs per megabase per haplotype are present in the haplotypes used for simulations. That is, we count the total CSVs from each group observed in the simulation haplotypes of given group and normalize by sample size and chromosome length. The expected number of informative CSVs per megabase per haplotype gives an indication of how well local ancestry can be inferred using

only CSVs.

2.2.3 Local ancestry inference using CSVs

Following previous works[142], we consider admixed populations arising from K ancestral populations $A_1, \ldots, A_K$ that have been mixing for G generations. For a given admixed genotype from the admixed population, we describe each individual genotype as a vector g, where $g_i \in (0, 1, 2)$ is the number of alternative alleles of that individual at SNP i. At position i, the individuals two alleles have either both descended from the same ancestries (i.e. continental group) or from two different ancestries. We are interested in determining the ancestry origin of the two alleles at each position i in the genome. Our model is based on an HMM described by a triple $H = (Q, \delta, \epsilon)$, where Q is the set of states, δ is the transition probability function and ϵ is the emission probability function. A different HMM is estimated for each individual at each iteration with parameters estimated from the locations of informative CSVs in each individual.

We denote by Q each possible combination (including the same ancestry) of ancestries in a diploid genome. The transition function δ changes at each step j as a function of the genetic distance between informative CSVs. The emission probabilities ϵ are constant for each state in Q. For any number of ancestral groups K, there are nine transition types that are typical of all possible transitions (not all are needed if $K < 4$)(see Table 2.1). The transition functions described can describe the transition from any state q at step $j-1$ to any state q' at step j (see Figure 2.1). Here r is the probability of one or more recombinations occurring between the $j-1^{th}$ informative CSV and the j^{th} informative CSV and is a function of the genetic distance between the two of them. This is modeled as a Poisson process with parameter $dG\lambda$ as the probability of one or more recombinations occurring between two SNPs separated by distance d, having recombined G generations ago and with a rate parameter λ. There is significant linkage between many of the CSVs so we set $\lambda = 10^{-15}$ in order to minimize the effect of close highly linked CSVs.

It is impossible to perfectly determine which CSVs are spurious from the reference sets,

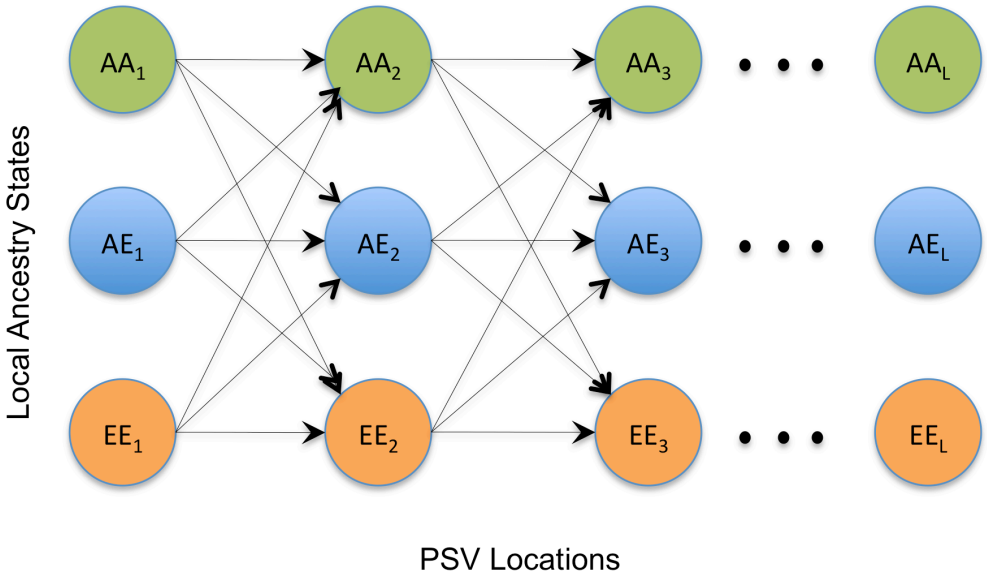

Figure 2.1: The hidden Markov model for a 2-way admixed individual (e.g. African American). The three types of states represent the three types of possible ancestry combinations: homozygous for African ancestry, homozygous for European ancestry or heterozygous for African and European ancestry. The probability of transitioning between the previous state q_{l-1} and q_l is a function of the genetic distance between the previous CSV_{l-1} and CSV_l.

so emission probabilities must reflect the possibility of errors in determining which variants are continent-specific (see Table 2.2). In a section of heterogeneous ancestry, emissions from the two ancestries are expected to occur proportional to the expected number of informative continent-specific variants seen in the two ancestries. In a section of homogenous ancestries, emissions are expected from only the one ancestry. We assume a low spurious CSV rate of $\epsilon_{CSV} = 10^{-5}$ and allow for the iterations to correct for errors by removing spurious CSVs identified in confidently called homozygous ancestry sections. We assume that the first state (q_0) of the HMM is silent. With the HMM defined for each individual, the probability of the individuals continent-specific variants is computed by summing over all paths π of length L (the number of CSVs showing alternate alleles in that respective individual):

$$P(CSV|HMM) = \sum_{\pi} \prod_{j=1}^{L} \delta_j(\pi_{j-1}, \pi_{j-1}) \epsilon(CSV_j|\pi_j) \tag{2.1}$$

The HMM is posterior decoded and local ancestry is called by assigning each CSV location the ancestry pair that had the highest posterior probability. Ancestry was called at all

variants by calling a variants ancestry as the same as the proceeding informative CSV.

Once ancestry calls have been assigned, reference panels are updated to reflect newly identified CSVs and to remove spurious CSVs. For each reference continental group k and for each allele i in the genome, the sample allele frequency p_{ki} is found by summing the alternate allele count across all individuals at allele i with homozygous ancestry for group k at that allele and then dividing by twice the number of homozygous calls at that locus for group k. This is performed for all homozygous ancestry SNP locations in individual i except for at SNPs that are within 10 SNPs of an ancestry transition since these are likely to be less confidently called. The minor allele frequency $\tilde{p}_{ki}$, first calculated from the reference haplotypes is then updated.

$$\tilde{p}_{ki} = \max(p_{ki}, \tilde{p}_{ki}) \tag{2.2}$$

The maximum value is used because allele frequencies are used as indicators of the presence of a CSV in a population; that the frequency is equal to zero or greater than zeros is what is important for training CSVs. Another iteration of posterior decoding is performed using this new $\tilde{p}_{ki}$ in order to determine CSV locations. We generally observed negligible improvement in accuracy between the 3^{rd} and 4^{th} iteration.

	(A_1, A_1)	(A_1, A_2)
(A_1, A_1)	$((1-r_j)+r_j\theta_1)^2$	$((1-r_j)+r_j\theta_1)\theta_1$
(A_1, A_2)	$2((1-r_j)+r_j\theta_1)r_j\theta_2$	$((1-r_j)+r_j\theta_1)((1-r_j)+r_j\theta_2)+r_j^2\theta_1\theta_2$
(A_3, A_3)	$r_j^2\theta_3^2$	$r_j^2\theta_3^2$
(A_2, A_3)	$r_j^2\theta_2\theta_3$	$((1-r_j)+r_j\theta_2)r_j\theta_3+r_j^2\theta_2\theta_3$
(A_3, A_4)	$r_j^2\theta_3\theta_4$	$r_j^2\theta_3\theta_4$

Table 2.1: The transition probabilities between ancestry pairs. If A_k represent a specific ancestry and θ_k represents the admixture proportion of that ancestry in the admixed population, then these equations are the transition probabilities for all possible types of transitions given a probability r_j of one or more recombinations occurring between the $(j-1)^{th}$ informative CSV and the j^{th} informative CSV. The columns represent the ancestry state at the $(j-1)^{th}$ CSV and the rows the ancestry state being transitioned into at the j^{th} CSV.

	$\pi = (A_1, A_1)$	$\pi = (A_1, A_2)$
$CSV = A_1$	$1 - \epsilon_{CSV} K$	$\frac{N_1}{N_1+N_2} - \epsilon_{CSV} \frac{K}{2}$
$CSV \neq A_1$	$\epsilon_{CSV} K$	NA
$CSV = A_2$	NA	$\frac{N_2}{N_1+N_2} - \epsilon_{CSV} \frac{K}{2}$
$CSV = A_3$	NA	$\epsilon_{CSV} K$

Table 2.2: Probability of emitting an informative CSV from an ancestry state. The probability of seeing a CSV from a different ancestry in a homozygous ancestry state is ϵ_{CSV}. In heterozygous states, CSVs are expected to be observed proportional to the ratio of the expected number of informative CSVs per haplotype per megabase per individual (see Table 2.4) in the two populations. N_k represents the expected number of informative CSVs per haplotype per megabase per individual in population k.

2.2.4 Comparison to array-based methods

We compared Lanc-CSV to LAMP-LD (v1.0) and MULTIMIX (v1.1.0), two widely used state-of-the art methods for local ancestry inference and a concurrently published method, RFMix (v1.0.2). We used unphased genotype data as input for all methods except RFMix, which requires phased haplotypes. We ran LAMP-LD using the same parameter settings used by 1000 Genomes[39] (number of states 25 and window size 100). We ran RFMix using the default settings with no EM iterations because of the large reference panel sizes. RFMix must be used with phased haplotypes that we computed with Beagle[20] using 30 haplotypes each from the African, European and Native American (Asian) reference panels as haplotype references for phasing. We ran MULTIMIX using the MULTIMIX_MCMCgeno method (which cannot be run with the resolve step). We ran it using the suggested misfit rates for two-way admixture [0.95 0.05; 0.05 0.95] and [0.95 0.025 0.025; 0.025 0.95 0.025; 0.025 0.025 0.95] for three-way admixture. For sequencing results we passed the fully sequenced reference haplotypes and the 200 simulated admixed individuals data to the programs. For array-based results we passed only the data at variants present on the Illumina 1M genotyping array (down sampled randomly to 50,000 variants in order to run on LAMP-LD). We parallelized LAMP-LD, RFMix and MULTIMIX for the fully sequenced data by splitting the data into small segments ($\sim$50,000 SNPs per segment) across the chromosome. We

computed accuracies by correlating the true and inferred local ancestry at each SNP across individuals only at the Illumina 1M chip variants.

2.2.5 Low-coverage sequencing

Using the same 200 genotypes for each simulated admixed population as above, we simulate read data for each individual. We assume that the number of reads covering each variant in each individual is drawn from a Poisson distribution with the rate parameter set to the average read coverage across the genome. We simulate reads for 0.1x, 1x, 2x, 5x, 10x, 20x, and 30x average coverage across the genome.

We adapted the inference method above to function with input read count data instead of genotype data. Given a set of read data for an individual at SNP i, $r_i = (ref_i, alt_i)$, where ref and alt are the counts of reads of the reference allele and the alternate allele. We first compute genotype dosage (d_i) at SNP location i using the admixture proportion weighted mean frequency in admixed individuals ($\hat{p}_i$) of the alternate allele. Let ancestral population k have admixture proportion θ_k on average in the admixed individuals.

$$\bar{p}_i = \sum_{k=1}^{K} \theta_k \tilde{p}_{ki} \tag{2.3}$$

Then $P(g_i)$, where g_i is the genotype, is assumed to follow Hardy-Weinberg Equilibrium with alternate allele frequency $\bar{p}_i$. $P(r_i|g_i)$ follows a binomial distribution modeling the number of alternate alleles seen given the number of trials equal to the total number of reads and the probability of an alternate allele equal to $1 - \epsilon_s$, 0.5 and ϵ_s for $g = 0$, 1, or 2. We assume a sequencing error rate of $\epsilon_s = 0.01$. We then calculate the genotype dosage:

$$d_i = \frac{\sum_{g_i=0}^{2} g_i P(r_i|g_i) P(g_i)}{P(r_i|g_i) P(g_i)} \tag{2.4}$$

When $d_i > 0.6$, we assume that the alternate allele is present at position i and can then run Lanc-CSV as previously described. We choose this threshold value so that the false positive rate is below 0.0025 and the false discovery rate of observed CSVs is below 0.2 for all coverage levels at or above 1x.

The Wahlund Effect[188] decreases heterozygosity and breaks Hardy-Weinberg Equilibrium when individuals from multiple populations are sampled and have different allele frequencies. CSVs are very rare and 98% of CSVs have an allele frequency <5% in the population in which they are observed. In order to ensure that the Wahlund Effect does not significantly affect our method, we calculate the probability of each genotype for a CSV with frequency 5% in one population admixing with another population with CSV frequency 0%, where the admixture proportion of the population with the observed CSV is 10%, 50% and 80% (see Table 2.3). The magnitude of the effect is decreases as CSV frequency decreases so these are the most extreme expected values. The effect also assumes that sampled populations have not mixed, so each generation since admixture will further decrease the effect size.

2.2.6 Effect of sample aware inference

In order to determine if the accuracy of Lanc-CSV increases with increasing numbers of admixed individuals, we used an additional 800 African Americans, Puerto Ricans, and Mexicans each giving a total of 1000 simulated admixed individuals for each population. We then run Lanc-CSV for 50, 100, 200, 400, 800 and 1000 individuals in each sample and compute r^2 after 4 iterations. Each set of individuals contains the individuals from smaller data sets. Figure 2.8 shows the increasing accuracy with increasing sample size.

2.2.7 Analysis of real admixed individuals from 1000 Genomes

In order to assess the performance of our approach in real data, we used the Americans of African Ancestry in South Western USA (ASW), Mexican Ancestry from Los Angeles (MXL), and Puerto Ricans from Puerto Rico (PUR) genotypes from real individuals contained in 1000 Genomes. Since the true ancestry is not known, we evaluate the accuracy of Lanc-CSV by comparing to the local ancestry calls provided by 1000 Genomes. The 1000 Genomes calls are the consensus calls of four established local ancestry methods (including LAMP-LD). Calls were made at a locus when 3 of the 4 methods agreed on the local ances-

	P(G=0)	P(G=1)	P(G=2)
Hardy-Weinberg equilibrium with 10% admixture proportion	0.99	0.01	0.00
Unmixed 10% admixture proportion	0.99	0.05	0.00
Hardy-Weinberg equilibrium with 50% admixture proportion	0.95	0.05	0.00
Unmixed 50% admixture proportion	0.95	0.05	0.00
Hardy-Weinberg equilibrium with 80% admixture proportion	0.92	0.08	0.00
Unmixed 80% admixture proportion	0.92	0.08	0.00

Table 2.3: Wahlund Effect on genotype probabilities. When an allele has different frequencies in different populations and the populations are looked at as a single population, the Wahlund Effect predicts a decrease in heterozygosity. The magnitude of the effect decreases with the difference in the allele frequencies and with mixing between the populations. 98% of CSVs have an allele frequency ¡5%. Here we report the genotype probabilities assuming the admixed populations have established Hardy-Weinberg Equilibrium, and assuming they are completely unmixed (the most extreme version of the Wahlund Effect). We report these values for 10%, 50% and 80% admixture proportion of the CSV containing population. This demonstrates that the Wahlund Effect will have negligible effect on our methods performance.

try at that locus. We used the 1000 Genomes consensus ancestry calls in place of the true ancestry and r^2 was calculated the same way as previously described. This is not a measure of accuracy since the true ancestry is not known, but a measure of calling consensus between our approach and other ancestry inference methods.

To check possible causes of poor correlation with the consensus calls, we selected a subset

of 20 individuals from the real Mexican data, and determined CSVs using both the reference haplotypes as well as the regions of the remaining (not from the subset of 20) Mexican individuals genotypes that are homozygous for a local ancestry. We then reran our method on the subset of Mexicans, first using both reference haplotypes and the held out Mexicans for training CSVs and second using only the reference haplotypes. We trained on the held out admixed genotypes by using the homozygous ancestry regions from 1000 Genomes local ancestry calls to identify new and spurious CSVs after training on the reference haplotypes. We see significant increases in accuracy when we do this, demonstrating poor reference panels as a major driver of the poor correlation.

2.2.8 Ancestry calling for closely related populations

CSVs, as demonstrated in Table 1, contain sufficient information to distinguish continental groups from each other. However, it is possible to distinguish sub-continental populations from each other as well, such as distinguishing a JPT haplotype from a CHS haplotype, both of which are in the Asian continental group.

In order to distinguish sub-continental haplotypes we define sub-continental population-specific variants (sCSVs) as variants seen in one of the 1000 Genomes populations (e.g. GBR) but not in any of the other populations of all continental groups including its own. We perform a leave one out analysis where we remove one of the haplotypes from one of the populations, then train sCSVs on all remaining haplotypes. We then determine how many sCSVs from each of the populations we see on the held out haplotype.

We repeated this analysis, but instead of using the full haplotype to ask how many sCSVs are seen on each haplotype, we randomly choose sections from each haplotype between 0.05 and 30 megabases long and call the ancestral population of each haplotype segment as the population of which the most sCSVs were seen on the segment. With ten populations, random guessing results in an accuracy of 10%. We also calculate the accuracy of correctly calling the continental group from which each haplotype segment was drawn against the haplotype length.

In order to address the accuracy of sub-continental population calling in real data, we look at ASW individuals. In regions where the continental group ancestry (using the 1000 Genomes consensus calls) was called as African-African or European-European, we counted the number of sCSVs seen in each population. We normalized the counts to the number of observed sCSVs per megabase per haplotype. We compared these counts for the African-African ancestry regions to the expected number of observed sCSVs per megabase for a YRI haplotype and a LWK haplotype (calculated from the expected counts from the haplotypes used for Figure 2.9 which were then normalized by the length of the chromosome in megabases).

2.3 Results

2.3.1 Continent-specific variants in the 1000 Genomes data

Using data from the 1000 Genomes Project, we investigate whether CSVs can be used to perform accurate local ancestry inference. We define CSVs as single nucleotide variants in which one of the alleles is only observed in one of the continental groups (e.g. European) and absent from other continental groups. Determining CSVs can be quickly achieved using reference panels such as data generated by the 1000 Genomes Project[39]. Although extremely useful, 1000 Genomes was sequenced at low coverage (4x) with potentially many rare variants (likely to be CSVs[65]) being left uncalled. In addition, some variants are spuriously called as CSVs due to the finite sample of the reference panels; for example, variants that would be observed in larger samples from more than one continental group are mislabeled as CSVs due to the small size of the reference panels. We call these variants spurious CSVs.

We assess the presence of informative and spurious CSVs for the purposes of local ancestry inference in the real 1000 Genomes data. To mimic local ancestry inference, we used data from the TSI(97), JPT(88) and LWK(97) populations (as proxies for the European, Native American and African continental groups, numbers represent the number of individuals from each population) to infer CSVs and used a different set of haplotypes (CEU(85),

	European CSVs (TSI)	African CSVs (LWK)	Asian CSVs (JPT)
European Haplotypes (CEU)	15.70 (93%, 2.25)	1.00 (6%, 0.60)	0.21 (1%, 0.12)
Afican Haplotypes (YRI)	0.57 (<1%, 0.33)	123.48 (99%, 5.22)	0.33 (<1%, 0.14)
Asian Haplotypes (CHS+CHB)	0.40 (3%, 0.26)	0.64 (5%, 0.32)	10.75 (91%, 1.45)

Table 2.4: The average number of observed CSVs per haplotype per megabase from each ancestry. Parentheses are the percentages of CSVs on each haplotype and the standard deviations. To estimate CSVs we used TSI, LWK, and JPT individuals as proxies for the European, African and Native American ancestries. We calculated the number of European, African and Asian CSVs seen on CEU, YRI, and CHS+CHB haplotypes. The values in parentheses represent the percentages of each ancestry type of CSV seen on a haplotype from a specific population.

CHB+CHS(197) and YRI(88)) to determine the number of observed CSVs from each continental group on a haplotype of a given group. We observe that only a fraction of called CSVs using the reference panel are spurious in the target panel; e.g. an average of 15.70 per mega-base per chromosome of European CSVs in the reference are also observed on a target European haplotype as compared to 1.20 per mega-base per chromosome that are spuriously called (i.e. was a Native American or African CSV seen on the European haplotype) (see Table 2.4). The spacing between observed CSVs on a haplotype ranges on the average from 10 kb for African chromosomes to 100 kb for Asian chromosomes. Since we used data from different populations within the same continental groups (e.g. TSI and CEU), some of the European CSVs are missed as they are specific to only one population within the same continent. Therefore the numbers in Table 1 represent a lower bound on the total amount of CSVs informative for continental local ancestry inference. As previously reported, we ob-

serve a much larger number of African CSVs owing to the larger genetic diversity observed within Africa[39]. We also observe that the percentage of spurious African CSVs is much lower than that of European and Asian CSVs that are falsely identified (0.7% vs. 7.2% and 8.8%).

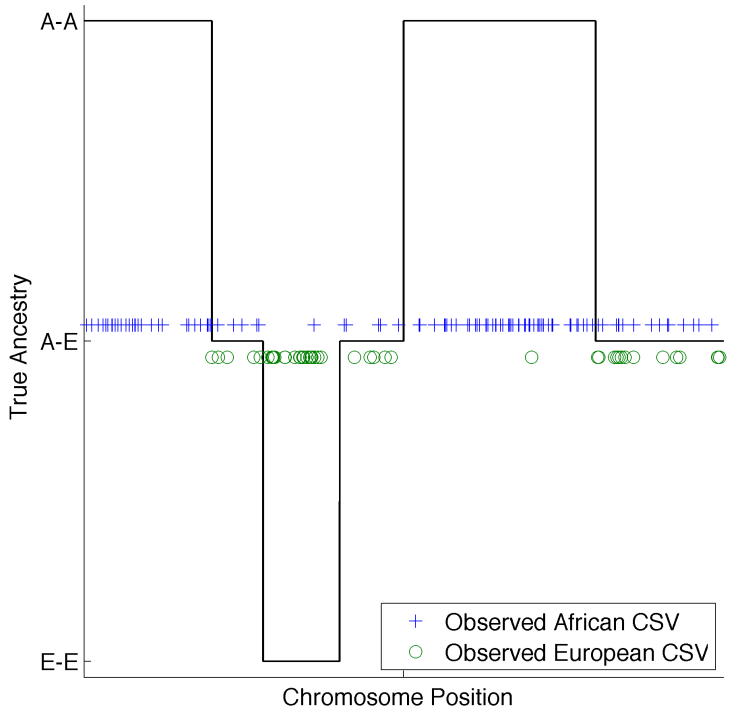

Figure 2.2: Example of CSVs in a 2-way admixed individual (e.g. African American). Lines denote the true local ancestry while the dots denote CSVs. Different dot types denote the continental ancestry of each CSV. From visual inspection it is relatively easy to discern the true ancestry from the three observed patterns. Spurious CSVs are denoted by CSVs mislabeling the true ancestry state.

2.3.2 Accurate local ancestry inference using CSVs

The admixture process creates chromosomal segments of different ancestry in recently admixed individuals[167]. Therefore, if we visualize CSVs along the genome of a recently admixed individual, we expect to observe continuous segments with only CSVs from one continent (at loci where both alleles have the same ancestry) or a mixture of CSVs from two continents (at loci where one allele comes from one ancestry and another allele from a different ancestry) (see Figure 2.2). In practice we do not know the true local ancestry and

we observe CSVs along the genome in an admixed individual (with potential errors) and we seek to infer the underlying local ancestry status. We extend standard hidden Markov models (HMM) for local ancestry to model CSVs as emissions and local ancestry as the underlying hidden state. We use this model to calculate the probability of the ancestral state at each locus in the genome conditional on the observed sequence of CSVs.

	African American	Mexican	Puerto Rican
LAMP-lD (array data)	0.98 (1.00, 0.99)	0.89 (0.97, 0.93)	0.91 (0.98, 0.96)
MULTIMIX (array data)	0.93 (0.99, 0.98)	0.73 (0.94, 0.80)	0.74 (0.93, 0.86)
RFMix (array data)	0.90 (0.98, 0.97)	0.79 (0.93, 0.87)	0.82 (0.95, 0.91)
LAMP-LD (full genome)	0.85 (0.97, 0.95)	0.80 (0.94, 0.89)	0.79 (0.95, 0.90)
MULTIMIX (full genome)	0.46 (0.84, 0.72)	0.44 (0.73, 0.49)	0.40 (0.74, 0.56)
RFMix (full genome)	0.92 (0.99, 0.97)	0.83 (0.95, 0.89)	0.85 (0.96, 0.92)
Lanc-CSV	0.96 (0.99, 0.99)	0.87 (0.96, 0.92)	0.92 (0.98, 0.96)

Table 2.5: Local ancestry accuracy in simulations of African Americans, Mexicans and Puerto Ricans. Accuracy is reported as mean r^2 (haploid accuracy, diploid accuracy). "Array data" denotes that a method was run only on the variants present on the Illumina 1M genotyping array."Full genome" denotes methods were run using all the variants. RFMix requires phased haplotype input that was phased using Beagle; all other methods received unphased genotype data as input. Correlation values are the mean squared correlation across SNPs of the true vs. inferred ancestry across individuals. Accuracy is reported as mean r^2 (haploid accuracy, diploid accuracy). LAMP-LD and MULTIMIX were optimized to run with genotyping array data, possibly explaining the steep drop in accuracy when they are run using full sequencing data.

We used simulations of African Americans, Mexicans, and Puerto Ricans to quantify the performance of our approach. As a baseline for comparison, we used LAMP-LD and MULTIMIX, two of the fastest and most accurate methods for local ancestry inference[4, 36]. LAMP-LD models haplotypes within HMMs of haplotype diversity for ancestry assignment and has been recently shown to attain similar accuracy as another HMM-based approach (HAPMIX[154]) for African Americans and superior accuracy in Latino Americans. MULTIMIX models correlations among SNPs using a multivariate Gaussian approach; all methods utilize a window-based framework to integrate results across the genome. As a metric of

accuracy, we use the squared correlation coefficient (r^2) between the true simulated ancestry and the inferred one; the correlation coefficient measures the loss in association power for admixture mapping from errors in the local ancestry estimates[4]. We also report the percent of correctly inferred ancestry calls. Lanc-CSV attains similar results as best performing methods for ancestry inference across simulations of African Americans, Mexicans, and Puerto Ricans (e.g. mean r^2 of 0.92 with Lanc-CSV across the considered populations)(see Figure 2.4 and Table 2.5). Interestingly, we observe that the accuracy of both LAMP-LD and MULTIMIX deteriorates when sequencing data is used; e.g. mean r^2 of 0.93 when only SNPs on the Illumina-1M array are used, as compared to 0.71 when all sequencing data are used with LAMP-LD. Similar results are seen with MULTIMIX (see Figure 2.4 and Table 2.5). This is likely due to the fact that both LAMP-LD and MULTIMIX have been optimized for GWAS genotyping array data and not for the significant number of rare variants identified through sequencing. Recent work in parallel to ours has proposed the use of conditional random forests in local ancestry inference (RFMix[119]). We assessed RFMix accuracy on our simulations and we observe comparable accuracy as other methods for array data (see Figure 2.4 and Table 2.5). In addition, RFMix accuracy slightly increases when sequencing data is available from an $r^2 = 0.84$ to 0.87. We also observe a lower performance of MULTIMIX as compared to LAMP-LD and RFMix in our simulations. The average distance between a true switch point and the inferred switch point for Lanc-CSV is 76 kb and for LAMP-LD is 91 kb, both have a standard deviation greater than 100 kb (see Figure 2.3). Importantly, our approach requires significantly less computational runtime than both LAMP-LD and MULTIMIX run on genotyping array data (Lanc-CSV is 3-5x faster) or sequencing data (Lanc-CSV is 40-150x faster). Lanc-CSV is slightly faster than RFMix when run on sequencing data (1.4x reduced runtime) (see Table 2.6 and Figures 2.5 and 2.1). However, RFMix requires phased haplotype data, which can take significant time to calculate with unrelated individuals. If phasing time is included Lanc-CSV is 12.5x faster than RFMix on sequencing data (see Table 2.6).

	Runtime for 200 individuals on chromosome 10	Estimated runtime for 200 individuals on all chromosomes	Estimated runtime for 15000 individuals on all chromosomes
LAMP-LD (array data)	0.11	2.79	69.83
MULTIMIX (array data)	0.06	1.50	113.25
RFMix (array data)	0.004 (+0.01 for phasing)	0.11 (+0.18 for phasing))	8.06 (+13.12 for phasing)
LAMP-LD (full genome)	4.39	109.79	1,790.20
MULTIMIX (full genome)	3.30	82.58	6,193.80
RFMix (full genome)	0.03 (+0.22 for phasing)	0.79 (+5.5 for phasing)	59.46 (+412.5 for phasing)
Lanc-CSV (full genome)	0.02	0.54	41.41

Table 2.6: Runtime in CPU days for LAMP-LD, MULTIMIX, RFMix and Lanc-CSV. Runtimes were estimated by running each method on chromosome 10 in 200 individuals and extrapolated to full genome. Results are in total CPU days. All methods can be parallelized for proportional decreases in computing time. RFMix requires phased haplotype data and phasing time is reported in the parentheses.

2.3.3 Extension to low-coverage sequencing

Recent works have shown that low-coverage sequencing yields superior association power per unit of cost as compared to genotyping arrays in GWAS[141]. The accuracy of genotype calling from sequencing data is directly related to the read coverage. High read coverage increases the likelihood of observing true CSVs, while low read coverage increases the likelihood of both not observing a CSV and spurious CSVs due to errors in the genotype calling from read data. We extend our method to low-coverage sequencing data by means of a preprocessing step where a CSV is called present at a locus if the genotype dosage (i.e. the expected count of alternate alleles given the observed reads) is above a set threshold level

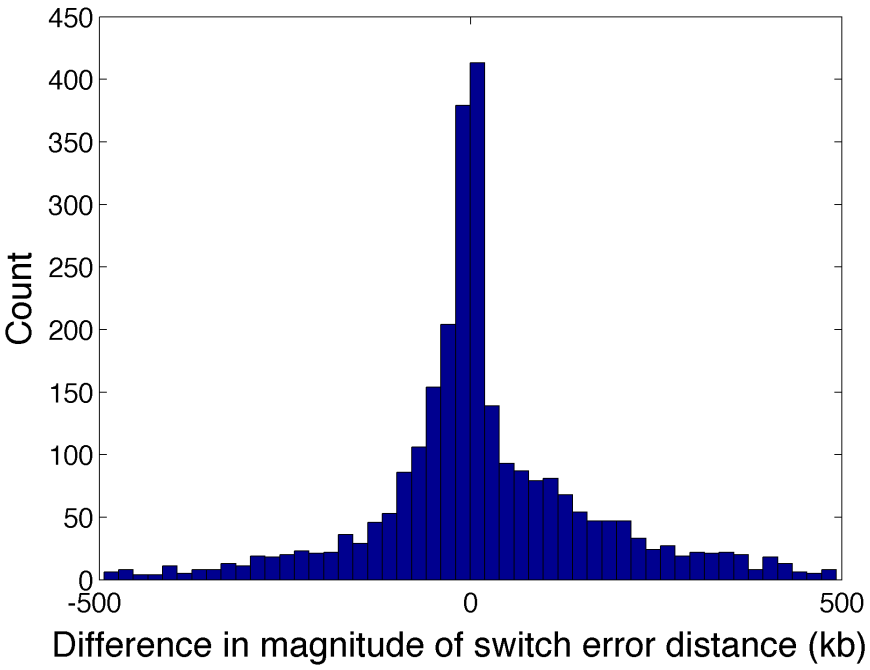

Figure 2.3: Resolution in determining ancestry switch locations in LAMP-LD and Lanc-CSV. For each true ancestry switch location in the simulated Puerto Rican data we calculated the distance in base pairs to the nearest inferred ancestry switch point for both LAMP-LD and Lanc-CSV from the true ancestry switch point. We only considered true switches where the inferred switches from both LAMP-LD and Lanc-CSV were less than 500 kb from the true switch point. The mean distance to the switch point for LAMP-LD was 91,145 bp and 75,644 bp for Lanc-CSV. For each true switch, we take the difference between the LAMP-LD error distance and Lanc-CSVs error distance and plot a histogram of these values. Positive values imply that at a true switch location LAMP-LD had greater error, negative values that our method had greater error; a zero value indicates that both methods are equally accurate.

at a CSV location. We estimate the genotype dosage from reads using standard techniques. Through simulations, we determine that the Wahlund Effect[188] is likely not going to impact our assumptions of Hardy-Weinberg Equilibrium at the allele frequencies of most CSVs (see Table 2.3). Starting from the previous simulations, we simulated sequencing data at various coverages using standard parameters for sequencing. At 5x coverage we observe an accuracy of 0.86 for African-Americans, 0.70 in Mexicans and 0.78 in Puerto Rican simulations. As expected accuracy increases as coverage increases with little gains in accuracy coming above 10x (e.g. an accuracy of 0.91 at 10x in Puerto Rican simulations) (see Figure 2.7).

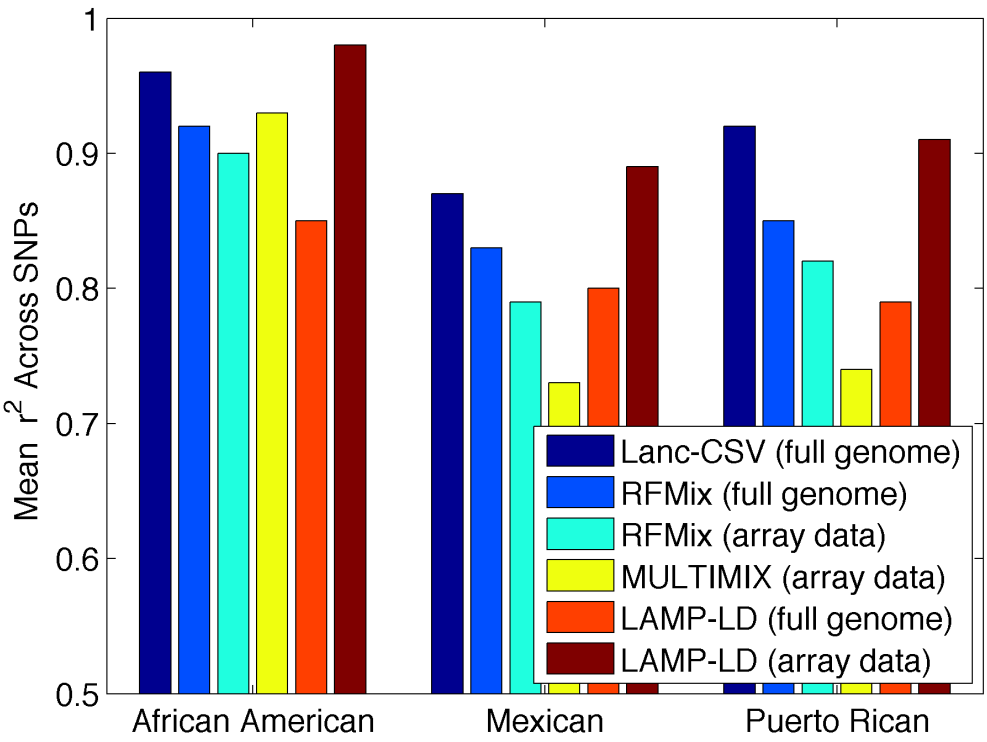

Figure 2.4: Local ancestry inference accuracy in three simulated populations. Array data denotes that a method was run only on the variants present on the Illumina 1M genotyping array. Full genome denotes methods were run using all the variants. RFMix requires phased haplotype input, which was infered using Beagle; all other methods received unphased genotype data as input. Correlation values are the mean squared correlation across SNPs of the true vs. inferred ancestry across individuals. LAMP-LD and MULTIMIX were optimized to run with genotyping array data, possibly explaining the steep drop in accuracy when they are run using full sequencing data. MULTIMIX is not plotted when run on full sequencing data because it performed very poorly, possibly due to inaccurate parameters for sequencing data. Haploid and diploid errors are reported in Table 2.5.

2.3.4 Sample-aware inference of local ancestry improves accuracy

Accurate methods for local ancestry inference leverage reference panels of haplotypes to use as proxies for the missing ancestral individuals that mixed to form current admixed populations. Recent works have shown that local ancestry inference can be improved when using the admixed samples themselves to rebuild the reference panels of haplotypes[143]. A major advantage of our approach for local ancestry inference is that we can iteratively re-estimate CSVs by incorporating information from the inferred ancestry regions in the admixed samples themselves. In particular, we first estimate CSVs using external reference panels of

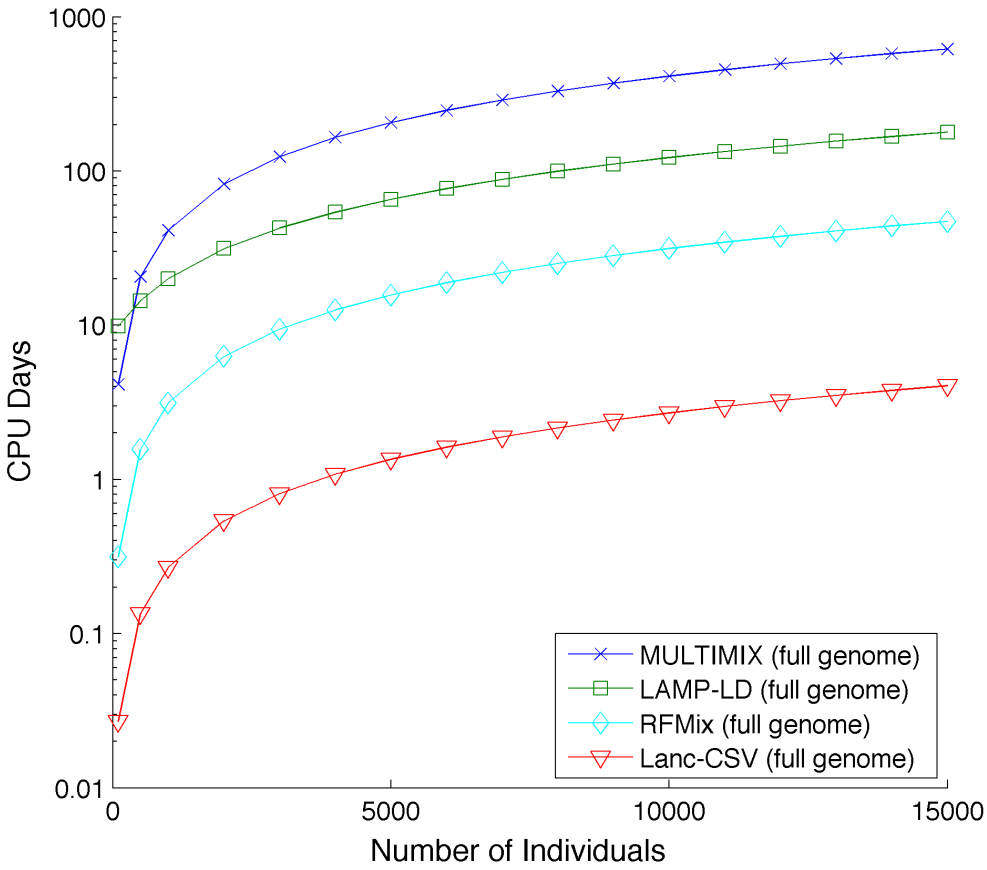

Figure 2.5: Runtime (in CPU days) as a function of the number of individuals in a study with sequencing data. Lanc-CSV is always faster than LAMP-LD and MULTIMIX when run on either full genome sequencing data or genotyping array data (see Figure 2.6 and Table 2.6). The full sequencing data contained ∼30 times more alleles than the genotyping array data. Only RFMix has comparable speed for full sequenced data and is faster for genotype array data. We show the runtime for RFMix with phasing time included.

haplotypes (e.g. 1000 Genomes), then call local ancestry and in an iterative fashion, re-call CSVs using confidently called ancestry segments from the sample itself. This procedure reduces the number of spuriously called CSVs while determining new CSVs and increasing the overall accuracy of the method. In addition, this allows for sample-aware reference panels that are better proxies for the true ancestral population of current day admixed individuals. For example we observe an increase in accuracy from $r^2 = 0.87$ to 0.92 in 200 simulated admixed Puerto Ricans after four iterations. The greatest increase in accuracy is after the first iteration and very little increase in accuracy comes with the fourth iteration. The main source of errors in Mexicans and Puerto Ricans is in distinguishing European and Native

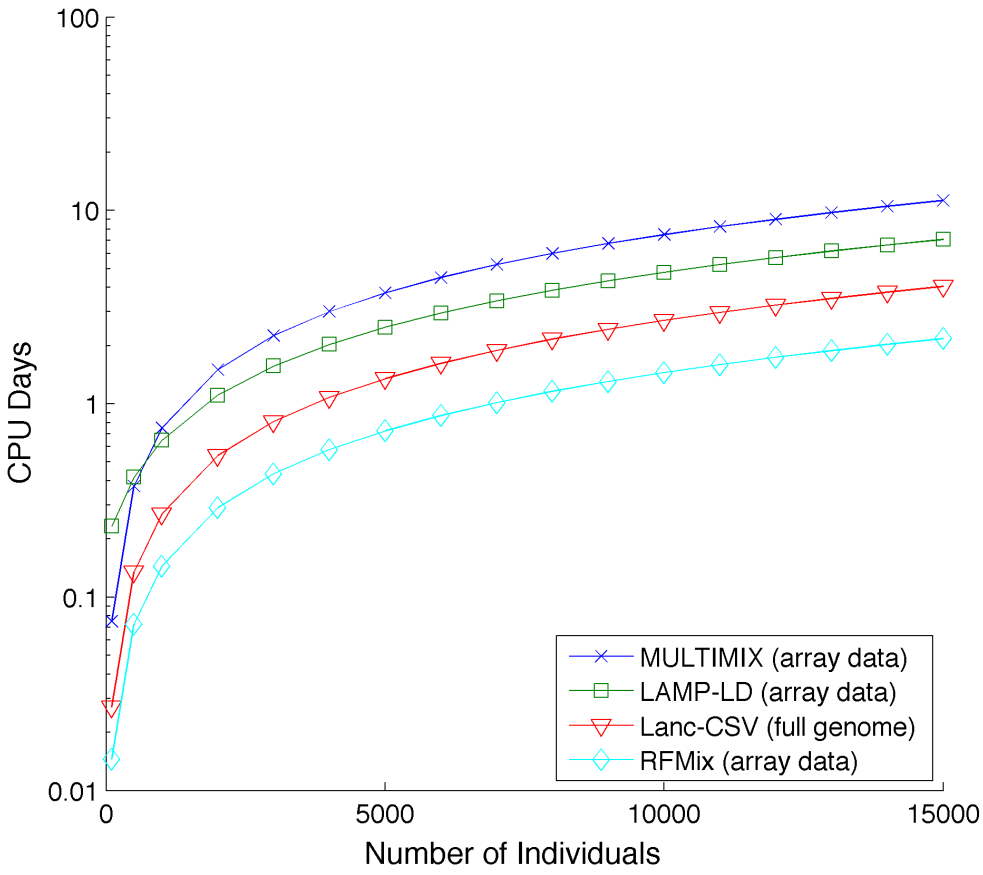

Figure 2.6: Runtime (in CPU days) as a function of the number of individuals in a study with genotyping array data (and sequencing data for Lanc-CSV). Lanc-CSV is always faster than LAMP-LD and MULTIMIX when run on either full genome sequencing data (see Figure 2.5 and Table 2.6) or genotyping array data. The full sequencing data contained ∼30 times more alleles than the genotyping array data. Only RFMix has comparable speed for full sequenced data and is faster for genotype array data. We show the runtime for RFMix with phasing time included.

American regions that have a much lower signal to noise ratio than in African and European or African and Native American regions (see Table 2.4). African American inference is highly accurate at all sample sizes because even without any iterations, accuracy is high due to the strong signal to noise ratio allowing African and European segments to be easily distinguished.

As compared to previous methods that do not use information from the other admixed individuals when calling local ancestry, Lanc-CSV will continue to increase in accuracy as the admixed sample size increases. Figure 2.8 plots the accuracy as a function of the number

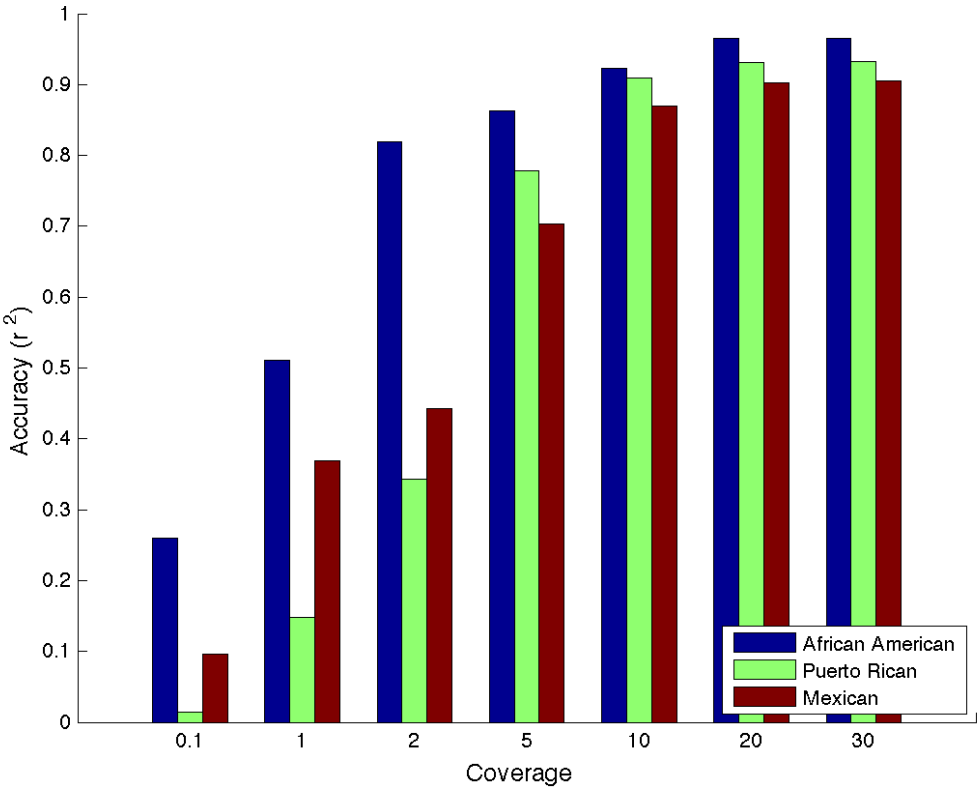

Figure 2.7: Accuracy as a function of sequencing coverage. African-Americans with only two distinct ancestral populations increases fastest in accuracy.

of admixed individuals. As expected we observe that the accuracy increases as the number of individuals increases with 200 samples being sufficient for high accuracy comparable to LAMP-LD in these simulations. However, as more simulated samples are added in, accuracy exceeds that of LAMP-LD in both the Mexican and Puerto Rican ancestries.

2.3.5 Analysis of real admixed individuals from 1000 Genomes

We investigated whether similar results can be achieved in real admixed genomes. We used our approach and LAMP-LD to call ancestry in the real data from the Americans of African Ancestry in South Western USA (ASW), Mexican Ancestry from Los Angeles (MXL), and Puerto Ricans from Puerto Rico (PUR) individuals from the 1000 Genomes Project. 1000 Genomes provided local ancestry calls for these individuals based on the consensus of four current local ancestry inference methods[4, 154, 119, 36]. Since the true ancestry is not known for these individuals, we measured the correlation between the local ancestry calls of

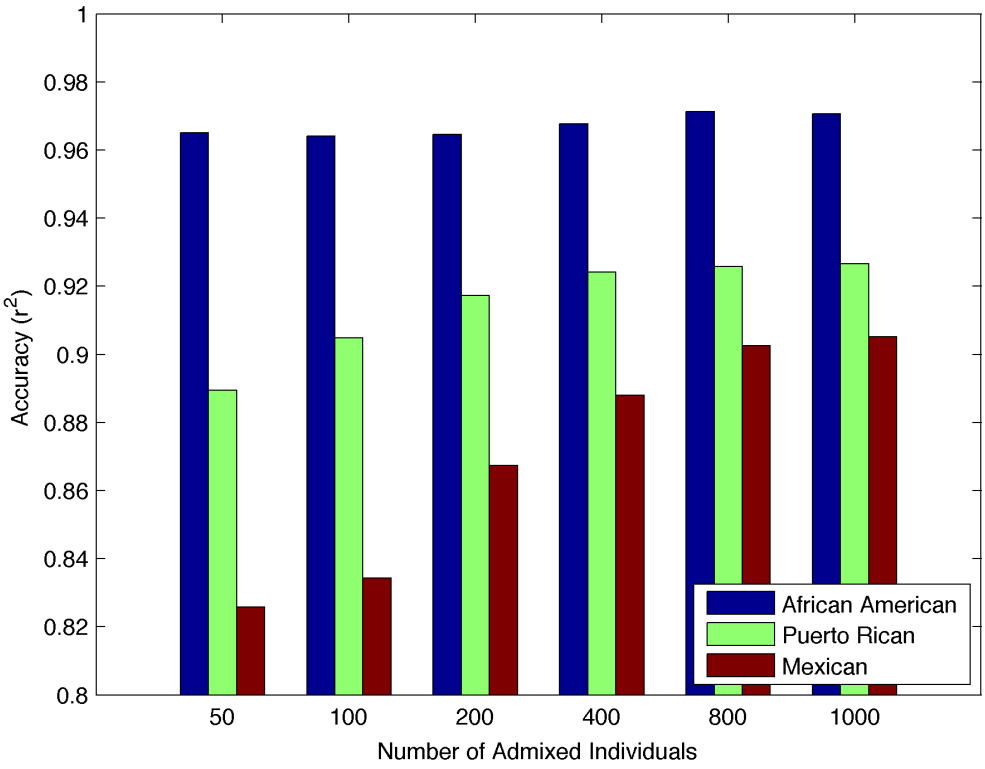

Figure 2.8: Accuracy as a function of sample size. While accuracy increases with increasing numbers of admixed individuals, the most significant increase is seen in Mexican individuals. We report accuracy for Lanc-CSV using 200 admixed individuals, but accuracy exceeds this as the number of admixed individuals increases. This is due to the method being better able to correct for spurious CSVs and to add in new CSVs when there are more individuals.

our approach with the ancestry calls provided by 1000 Genomes. We observed an average correlation rate (r^2) on chromosome 10 of 0.94, 0.63, and 0.81 for Lanc-CSV and 0.99, 0.66 and 0.79 for LAMP-LD (which was used as part of inferring the consensus calls) in African Americas, Mexicans and Puerto Ricans respectively (haploid and diploid errors reported in Table 2.7).

These low r^2 values are likely a result of poor reference panels in our inference since we are using the Asian haplotypes as proxy for Native American panels (1000 Genomes project used a specially designed panel for Native American[118]). To further investigate this hypothesis, we used our method to infer local ancestry in 20 of the Mexican individuals using the rest of the Mexicans as reference panel (that is, we used the consensus ancestry calls provided by 1000 Genomes to call CSVs). We observe a large increase in accuracy when incorporating

the other Mexicans (and their local ancestry consensus calls) in the reference panel (mean r^2 of 0.80 versus 0.66 if only Asian samples are used as reference).

This demonstrates that the low accuracy of both LAMP-LD and Lanc-CSV in real data is likely due to poor reference panels. It also demonstrates that a sample aware method could overcome this obstacle if sufficient admixed individuals are available. Therefore, we use the consensus calls of the Mexicans and Puerto Ricans of the 1000 Genome data to build improved CSV reference panels and provide them as a free resource to be used with Lanc-CSV for new sequenced admixed individuals.

	African American (ASW)	Mexican (MXL)	Puerto Rican (PUR)
LANC-CSV	0.94 (0.994 0.988)	0.63 (0.84, 0.68)	0.81 (0.96, 0.92)
LAMP-LD	99.12 (1, 1)	0.66 (0.89, 0.77)	0.79 (0.94, 0.79)

Table 2.7: Correlation of ancestry calls between our approach and the 1000 Genomes calls in real admixed individuals from 1000 Genomes. Accuracy reported as r^2 (haploid accuracy, diploid accuracy). The 1000 Genomes consensus local ancestry calls were made using LAMP-LD as one of the four methods. This demonstrates that poor accuracy is likely a result of poor reference panels.

2.3.6 Sub-continental ancestry calling

We extend continent-specific variants to sub-continental population-specific variants (sCSVs). We define sCSVs as variants that are observed in only one of the 1000 Genomes populations and not in any other (e.g. a variant observed only in the individuals from Great Britain (GBR) and never in any of the other populations). Using a leave one out analysis we demonstrate in Figure 2.9 that the chromosomes from 9 out of 10 populations have more observed sCSVs from the population from which it was observed than from any other. Due to limited reference panel size and the closeness of the sub-continental populations, there are considerable numbers of spurious sCSVs, but not enough to make sub-continental ancestry calling impossible in some scenarios. The two exceptions are the IBS population that only has 28 reference haplotypes (not enough to accurately determine sCSVs) and the CHB and

CHS that are genetically very similar.

	CEU	TSI	GBR	FIN	IBS	CHB	CHS	JPT	YRI	LWK
CEU	1.00	0.92	0.92	0.61	0.11	0.04	0.06	0.04	0.13	0.18
TSI	0.28	1.00	0.28	0.18	0.05	0.02	0.02	0.02	0.08	0.13
GBR	0.53	0.52	1.00	0.36	0.07	0.03	0.03	0.03	0.09	0.12
FIN	0.10	0.09	0.10	1.00	0.01	0.01	0.01	0.01	0.02	0.02
IBS	0.71	1.00	0.81	0.48	0.46	0.07	0.06	0.06	0.32	0.39
CHB	0.02	0.03	0.02	0.04	0.00	0.98	1.00	0.56	0.05	0.07
CHS	0.01	0.01	0.01	0.02	0.00	0.50	1.00	0.25	0.03	0.05
JPT	0.00	0.01	0.00	0.01	0.00	0.11	0.10	1.00	0.01	0.02
YRI	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.38
LWK	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.19	1.00

Figure 2.9: Proportions of sCSVs from each population observed on a held out haplotype. Each row represents the ancestry of the haplotype that was held out and each column represents the average number of sCSVs observed on the held out haplotype from the given population. Each row is normalized by the maximum value of the row so that the population with the most sCSVs observed has a value of 1. In each row, higher values are associated with populations in the same continental group as would be expected. The IBS have only fourteen individuals, which makes determining IBS sCSVs extremely difficult.

We assess the ability of correctly calling the population through a leave-one-out procedure starting from the real 1000 Genomes haplotype data. For each held out haplotype we randomly select short segments between 0.1 and 30 megabases and assign them to the population that has the maximum sCSV count across this segment. We plot the accuracy of this nave calling as a function of segment size in Figure 2.10. We also calculate the accuracy of assigning each haplotype to the correct continental group based on sCSVs in Figure 2.11. Correlating the accuracy of assigning the correct population to the haplotype segment (length 10 megabases) with the size of the reference panel of the called segment achieves a correlation coefficient of r=0.65 (p-value=0.042) showing that larger reference panel sizes

are associated with more accurate sub-continental population ancestry inference. The two African populations (YRI and LWK) as well as the Finnish (FIN) and (JPT) have very accurate ancestry calls possibly due to a higher degree of genetic differentiation as compared to the other sub-continental populations. The CEU individuals are Utah residents with northern and western European ancestry and may already be sub-continentally admixed which could potentially explain the low accuracy seen with the CEU. The IBS do not have enough reference panels to be able to call IBS sCSVs. As expected, when errors are being made, most of the errors resulted in another population from the same continental group being called (Figure 2.11). We also simulated diploid admixed individuals from pairs of sub-continental European populations with moderate accuracy in Lanc-CSV (Table 2.8).

	FIN	GBR	TSI	CEU
FIN	na	0.76	0.77	0.77
GBR		na	0.74	0.75
TSI			na	0.77
CEU				na

Table 2.8: Accuracy of Inference on 100 simulated admixed individuals among pairs of countries in Europe. We used admixture proportions of (0.5,0.5) and 6 generations of admixture. Accuracy is reported as haploid error. We observe a high proportion of heterozygous ancestry calls (over 90%), consistent with increased ambiguity in the calling using sCSVs for closely related populations.

In order to determine the effectiveness of the sCSV approach in real data, we counted the sCSVs observed per megabase in African-African and European-European continental called ancestry regions of the ASW individuals on chromosome 10 (using the 1000 Genomes consensus local ancestry calls). Figure 2.12 shows that in the African-African regions there is strong enrichment for YRI sCSVs. We additionally plot the expected number of observed sCSVs on a YRI haplotype (red diamonds) and the expected number of observed sCSVs on an LWK haplotype (green squares). The observed counts more closely resemble the count profile expected from the YRI haplotypes. This supports the established hypothesis that the African component of the ASW is likely from western Africa. When looking at the European-

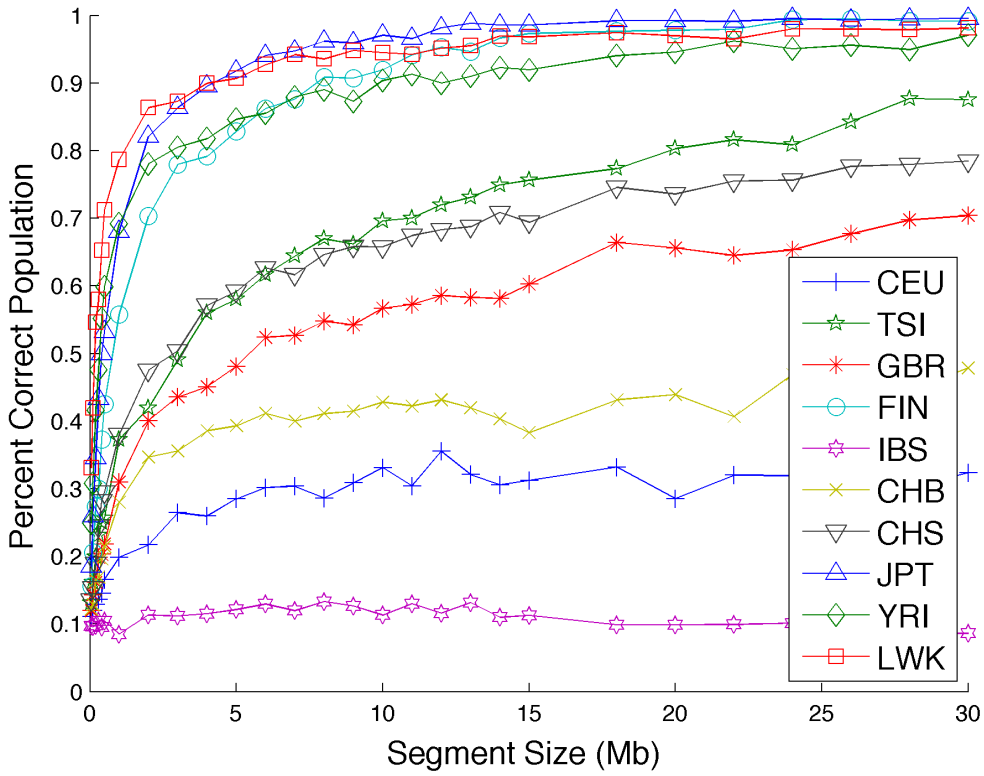

Figure 2.10: sCSVs allow for calling the sub-continental population of a haplotype. Randomly drawn segments of haplotypes from known populations can be accurately assigned to the population of origin. Accuracy for each population is significantly correlated with the number of reference haplotypes for that population (r=0.65, p-value=0.042). The highest accuracies are seen in populations that are more isolated from other populations in their continents.

European segments of the ASW (Figure 2.13), the most sCSVs are CEU followed by GBR supporting the hypothesis that the European ancestors are more related to northwestern Europeans. However given the small admixture proportion of European ancestry in African Americans, there are only a few small regions of European-European ancestry resulting in the very low sCSV counts for the ASW in these regions as compared to the African-African ancestry regions.

2.4 Discussion

We have presented here an approach for local ancestry inference in fully sequenced recently admixed individuals. Our approach makes use of alleles that are found to be present only

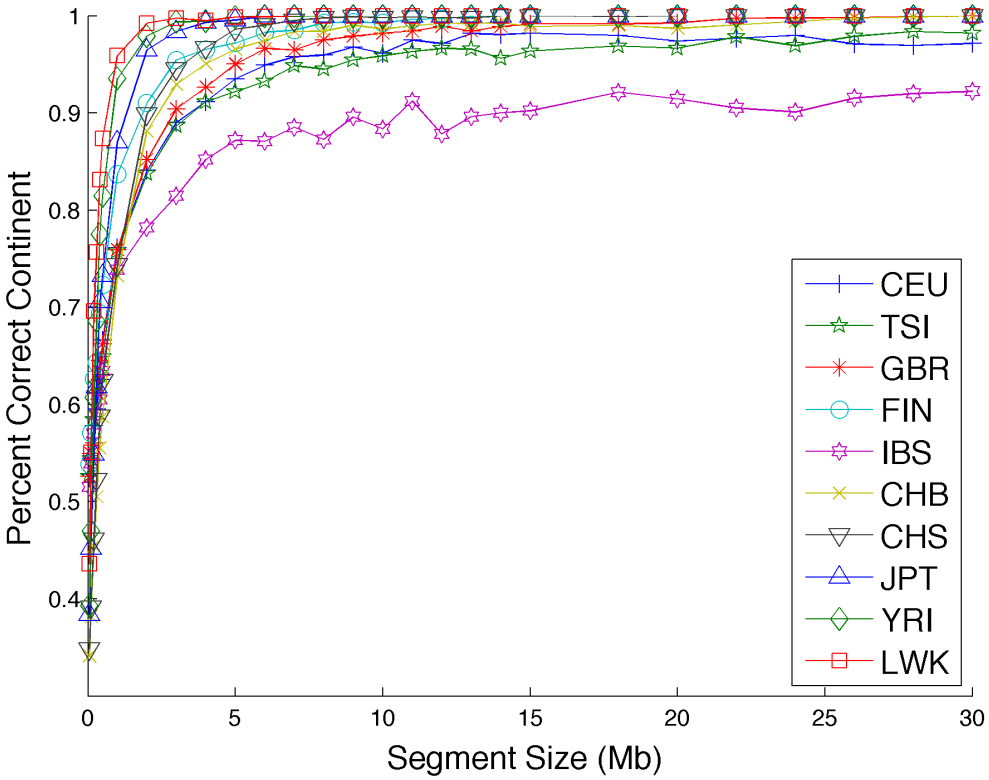

Figure 2.11: sCSVs are able to assign the correct continental group to small haplotype segments with high accuracy. This shows most of the incorrectly called accuracies still call to the correct continental group.

in individuals from a given continental group (continent-specific variants, CSVs). Through the use of real data from 1000 Genomes we have shown that the density of such CSVs is high enough across the genome to allow for fast and accurate inference of local ancestry. It should be noted that the 1000 Genomes haplotypes are based on 4x sequencing data. Not only does the low coverage make this data noisier, but it also misses many CSVs that are in the individuals but not called due to the low coverage. As more high coverage reference panels are constructed our method will become increasingly accurate as more CSVs are identified and spurious CSVs removed. Having no pre-compute time and fast runtime per individual allows for our approach to be sample-aware in an iterative fashion. As opposed to previous approaches Lanc-CSV shows increased accuracy as more admixed samples are being analyzed. We show that as the method is run on increasing numbers of simulated individuals, it exceeds the accuracy that is obtained by LAMP-LD on the 200 Mexican and

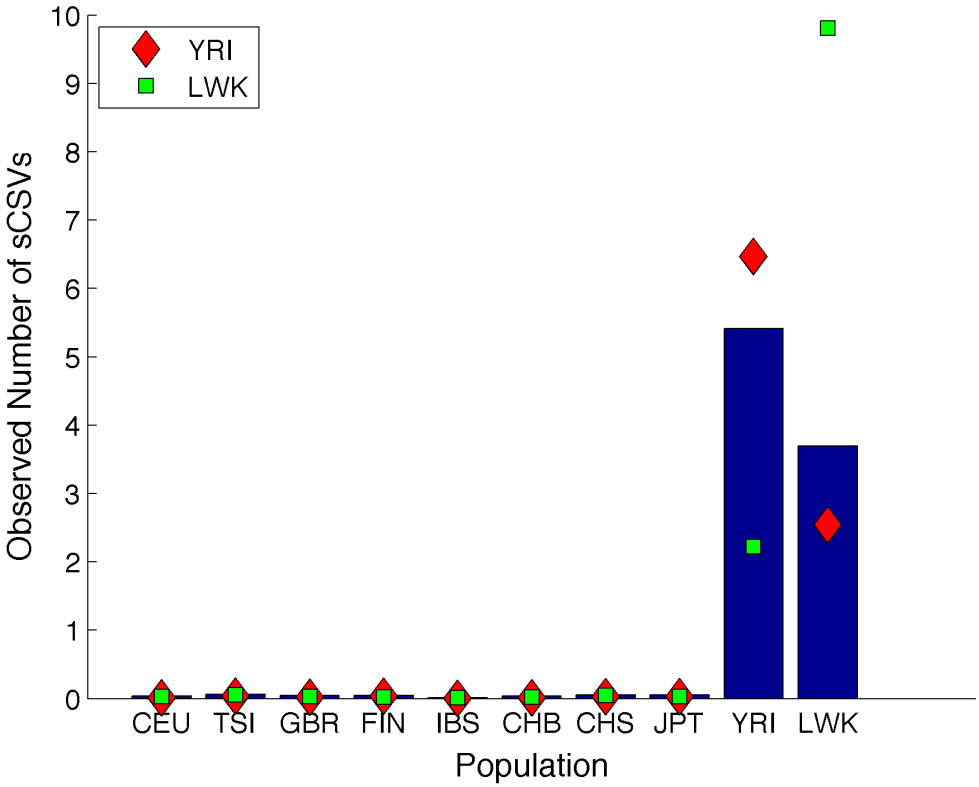

Figure 2.12: The average number of sCSVs from each 1000 Genomes population observed per megabase on the African-African called local ancestry regions of the real ASW individuals on chromosome 10. The large number of YRI sCSVs seen in these regions supports the hypothesis that the African admixture component in African Americans comes from western Africa. We plot the expected number of observed sCSVs per megabase on a YRI haplotype (red diamonds) and the expected number of observed sCSVs on an LWK haplotype (green squares). The observed counts more closely resemble the count profile expected from the YRI haplotypes.

Puerto Rican samples. We expect this feature to become more important as larger sample sizes are being analyzed since the accuracy should continue to increase.

The real data analysis demonstrates the necessity of having reference panels well matched to the admixed population or having a sample aware method that can correct for poorly matched reference panels. Lanc-CSV achieves comparable results to LAMP-LD in these few real African-American, Puerto Rican and Mexican individuals. Unlike LAMP-LD, we expect our approach to continue improving in accuracy as more sequenced individuals from each continental population become available. We extended the concept of continent-specific variants to sub-continental population-specific variants and showed that under some scenarios it

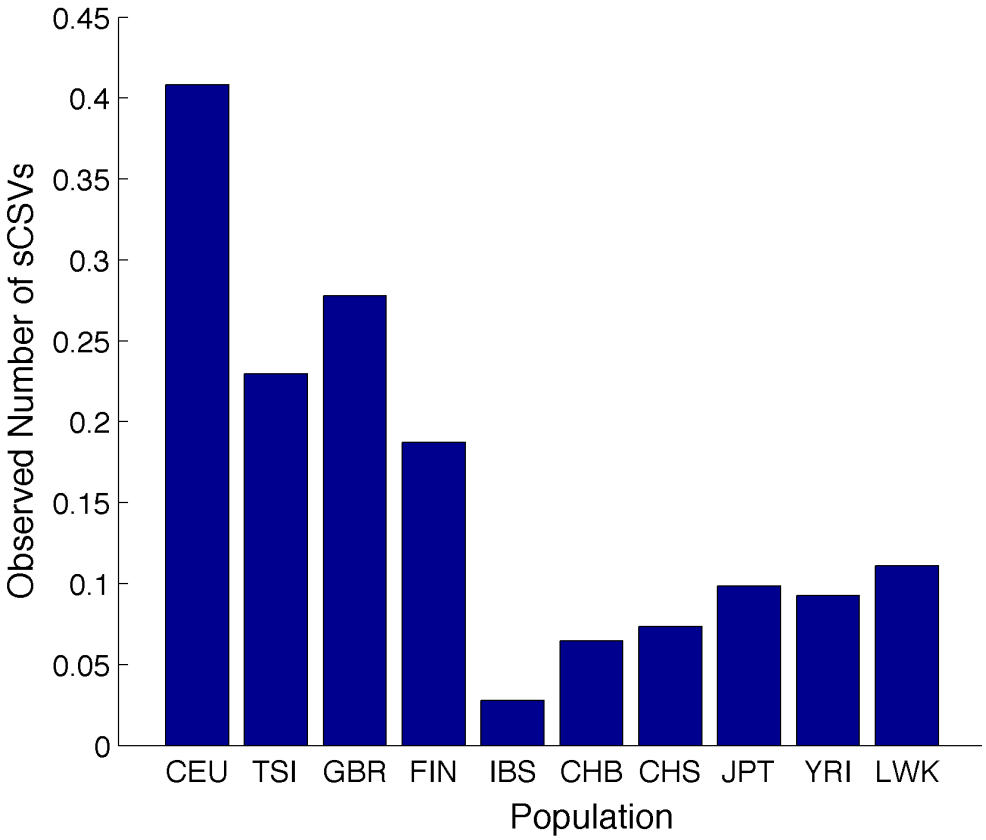

Figure 2.13: The average number of sCSVs from each 1000 Genomes population observed on the European-European called local ancestry regions of the real ASW individuals.

is possible to determine the sub-continental ancestry. We confirmed that in real ASW individuals, admixture was most likely between individuals from western Africa (near or related to the YRI); as more reference panels become available for these and other populations, we expect sCSVs to be increasingly informative of the sub-continental population ancestries. Although sCSVs show potential for sub-continental ancestry calling in haploid data, more sophisticated methods may prove fruitful for diploid calling.

A future direction for research that may prove fruitful is to relax assumptions used in our approach and by finding better ways to parameterize the method. Linkage disequilibrium among the CSVs is a main contributor to errors and further work explicitly modeling the LD structure between CSVs may provide increased accuracy. The current method has a uniform error rate for spurious CSVs across all ancestries. However the number of spurious European CSVs is much higher than the number of spurious African CSVs and the number of reference

haplotypes is not controlled for in determining the error rate; therefore a non-uniform error model may further increase accuracy.

CHAPTER 3

Leveraging ancestry to improve causal variant identification in exome sequencing for monogenic disorders

3.1 Introduction

Vast decreases in the cost of exome sequencing have allowed for major advancements in the identification of causal variants for rare monogenic traits and disorders[59, 3, 135, 99].Although each individual carries 20,000-24,000 single nucleotide variants, most are common in the population and are unlikely to explain a rare monogenic trait. Variants that are too common to be consistent with the prevalence of a rare disorder are removed from consideration[3] and the remaining variants are prioritized based on functional, structural and conservation properties[134, 138, 81]. Recent prioritization approaches use cross-species comparisons[160] or a combination of scores from several stand-alone methods for increased performance[63, 112, 108, 95]. Although such techniques are very powerful when family data is available[136, 174, 202, 159, 9, 133, 3, 99], hundreds of variants often remain for follow-up validation when only a single case individual is sequenced[108, 114, 115].

Variant filtering in exome sequencing studies is usually performed using frequencies that are estimated across large databases of human variation either by ignoring ancestry, or by

The work appearing in this chapter is published: Robert Brown, Hane Lee, Ascia Eskin, Gleb Kichaev, Kirk E Lohmueller, Bruno Reversade, Stanley F Nelson, and Bogdan Pasaniuc. "Leveraging ancestry to improve causal variant identification in exome sequencing for monogenic disorders." European Journal of Human Genetics, 24(1):113119, 2016.

matching at the level of continental ancestry (e.g. the Exome Variant Server[166])[116] thus ignoring sub-continental ancestry. Although F_{ST} values calculated within continental populations are usually low (mostly due to the dependency of F_{ST} on allele frequency)[80, 7], detectable population structure still exists[139, 200]. Population genetic models predict that rare variants show greater clustering within continental populations than more common variants[121] and empirical studies have supported this prediction[53, 39, 65, 184, 129, 207, 131]. Therefore, a rare variant might appear rare (<1%) when its frequency is estimated across many populations, when in reality it is only rare in most populations and less rare or even common (>1%) in a one or more clustered sets of populations (see Figure 3.1). For example, variant rs17046386 is common in Africans (therefore unlikely to be pathogenic) and generally rare or absent in non-Africans[107, 162] (see Figure 3.2). However, this variant would not be discarded in the filtering step based on frequency estimates from European reference panels thus increasing the validation burden in the subsequent steps. In addition, the limited size of existing reference panels, especially when defining ancestry at the level of a country (often <100 individuals), induces significant statistical variance in allele frequency estimates that needs to be accounted for (e.g. a variant with true frequency of 0.5% has 9.0% probability of being observed with a frequency >1% in a sample of 100 individuals and thus erroneously discarded).

In this work we investigate the use of sub-continental allele frequencies (typically estimated at the level of a country[127, 191]) for a discrete frequency-based filtering step in exome sequencing studies but such ideas can also be applied to general statistical methodologies aimed at finding causal genes in exome scans. We propose approaches that leverage the frequency estimates across all sub-continental populations in filtering while accounting for the statistical noise introduced by the smaller number of individuals used to estimate frequencies. We use simulations starting from the 1000 Genomes[39], the NHLBI Exome Sequencing Project (ESP) Exome Variant Server (EVS)[166] and the ClinVar[100] data, to show that our approach improves efficacy of filtering (e.g. a reduction of 16% in the number of variants to be followed up in case-only simulations). Importantly, we show that the standard approach that ignores statistical noise in the allele frequency estimation is miscal-

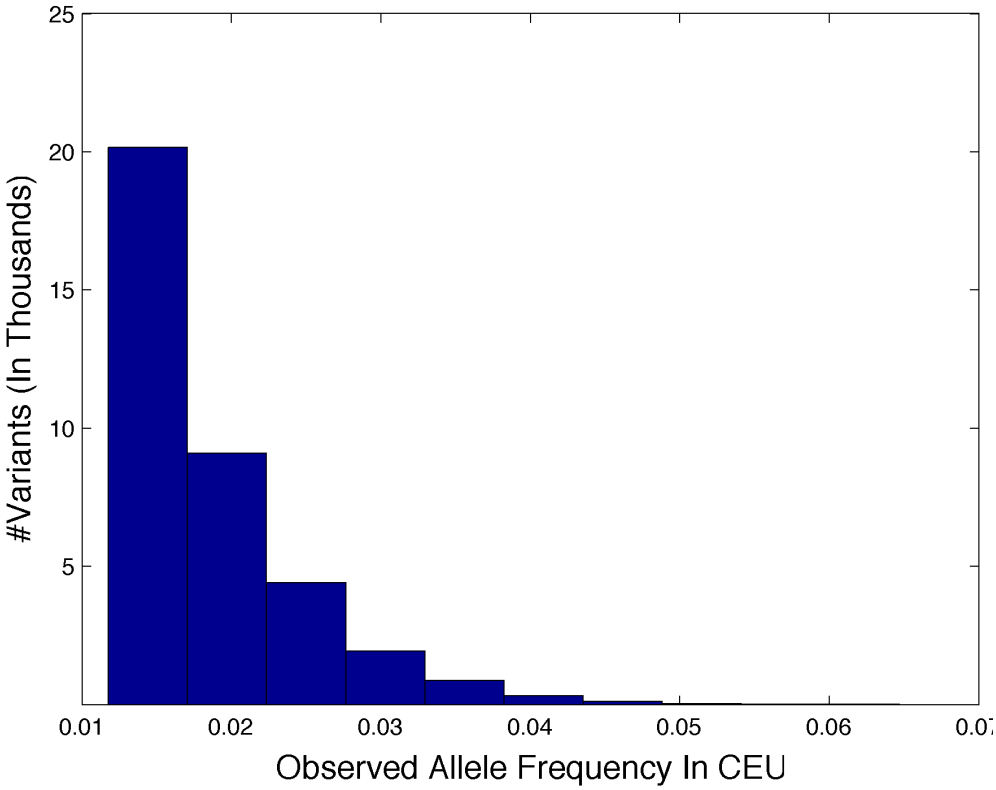

Figure 3.1: Histogram of variants with allele frequencies <1% in 1000 Genomes but >1% in the CEU. It shows that allele frequencies can be highly structured for rare variants and that averaging across too many genetically dissimilar populations can have a downward-bias effect on frequency estimates of alleles present in a given population.

ibrated with respect to the false negative rate (i.e. the probability of filtering out a true causal variant). Finally, we validate our approach using exome-sequencing data from 20 real individuals with monogenic disorders for which the true causal variants are known. Here we successfully reduce the number of variants to be functionally tested (a 38% reduction from 750 to 468 in the heterozygous case), while never discarding the known causal variants. Our results show that existing filtering pipelines for studies of monogenic traits can be significantly improved by incorporating ancestry while accounting for statistical noise in the filtering step. Interestingly, utilizing sub-continental population reference panels overcomes the reduction in performance due to higher statistical noise from the smaller panels.

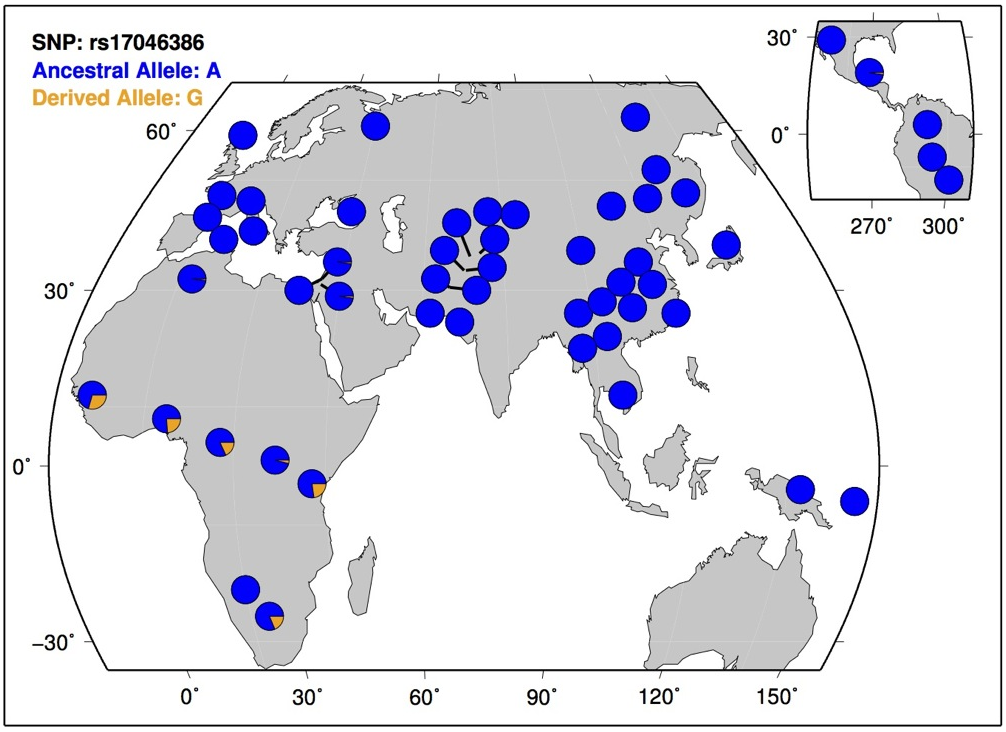

Figure 3.2: Geographic distribution of rs17046386 across the Human Genome Diversity Panel CEPH data. The minor allele is rare in non-African populations, but not rare in African populations.

3.2 Methods

3.2.1 Datasets

The 1000 Genomes Project[39] has produced a public catalog of human genetic variation through sequencing from several populations: Han Chinese in Beijing (CHB), Japanese in Tokyo (JPT), Southern Han Chinese (CHS), Utah Residents with Northern and Western European ancestry (CEU), Toscani in Italia (TSI), Finnish in Finland (FIN), British in England and Scotland (GBR), Iberian population in Spain (IBS), Yoruba in Ibadan (YRI), Luhya in Webuye (LWK), Americans of African Ancestry in SW USA (ASW), Mexican ancestry from Los Angeles (MXL), Puerto Ricans from Puerto Rico (PUR) and Colombians from Medellin (CLM). We use the 1000 Genomes data (with the exception of IBS individuals, only 14 in total) to evaluate the effectiveness of various filtering approaches. Since the vast majority of causal variants for monogenic traits are located in the exome[35], we restrict our

analysis to coding regions of autosomal chromosomes. For admixed individuals we downloaded and used the 1000 Genomes Project local ancestry calls (the consensus calls from four inference methods[4, 36, 119, 154]). Damaging scores for each single nucleotide variant were estimated using KGGSeq with default parameters[108] that combines the functional scores from dbNSFP[111] v2.0.

The NHLBI Exome Sequencing Project (ESP) Exome Variant Server (EVS) has released allele counts from 4,300 European-Americans and 2,203 African Americans[166] along with PolyPhen2 scores for missense variants and we used those in our analyses. A set of 1395 pathogenic variants (as reported by multiple submitters) was obtained from the ClinVar database[100] (accessed Dec 4, 2014).

To compare simulations to real data, we used exomes of 101 individuals with self- reported countries of origin including Turkey, Jordan, Tunisia, Egypt, Israel, Iran, Syria and Palestine. We grouped these individuals into a single supplemental population for estimating best matching allele frequencies. Of the 101 individuals, nine were known to harbor heterozygous variants in genes causing autosomal dominant disorders, ten had homozygous variants, and one had compound heterozygous variants in a gene causing an autosomal recessive disorder (see Table 3.6 and 3.7 and 3.8).

Exome sequencing was performed using llumina TruSeq Exome Enrichment Kit or Agilent SureSelectXT Human All Exon 50Mb kit. Illumina HiSeq2000 was used for sequencing as 100bp paired-end runs at the UCLA Clinical Genomics Center or the UCLA Broad Stem Cell Research Center. Sequence reads were aligned to the human reference genome (Human GRCh37 (hg19) build) using Novoalign (v2.07, http://www.novocraft.com/main/index.php). PCR duplicates were identified by Picard (v1.42, http://picard.sourceforge.net/) and GATK (Genome Analysis Toolkit) (v1.1, http://www.broadinstitute.org/gatk/)[123] was used to realign indels, recalibrate the quality scores, call, filter, recalibrate and evaluate the variants. SNVs and INDELs across the sequenced protein-coding regions and flanking junctions were annotated using Variant Annotator X (VAX), a customized Ensembl Variant Effect Predictor[203].

3.2.2 False negative rate estimation

We estimate the probability of filtering out a true causal variant (false negative rate) at a given frequency threshold as a function of a reference panel and the maximum true allele frequency of the causal variant. The filtering threshold can be adjusted in order to provide a desired FNR. Let t be the nominal frequency threshold that is used for filtering. We define the corresponding FNR at this threshold as:

$$FNR(t) = \frac{\int_0^{max(f_c)} fP(f_{ref,N} > t|f)P(f)df}{\int_0^{max(f_c)} fP(f)df} \tag{3.1}$$

where f is the frequency of the variant in the population, $max(f_c)$ is the maximum assumed frequency of the causal variant in the population, $P(f)$ is the proportion of variants with frequency f in the population and $P(f_{ref,N} > t|f)$ is the probability that a variant with frequency f is observed at a frequency greater than t in the reference panel of N individuals randomly drawn from the population.

The FNR computation requires knowledge about the distribution of variants across all frequencies in the population; this can be estimated from population genetic theory under various demographic assumptions[65, 94, 93, 150, 120, 10] or empirically from the data. In this work, we estimate the distribution $P(f)$ from reference panel allele counts and perform the above integration across the observed site frequency spectrum as follows:

$$FNR(t) = \frac{\sum_{f_i \leq max(f_c)} f_i P(f_{ref,N} > t|f_i)P'(f_i)}{\sum_{f_i \leq max(f_c)} f_i P'(f_i)} \tag{3.2}$$

Here f_i represents each of the unique allele frequencies observed in the reference panel of N individuals and $P'(f_i)$ represents the proportion of variants in the reference panel that have estimated frequency f_i. $P(f_{ref,N} > t|f_i)$ is modeled as a binomial draw with the frequency of success equal to f_i and the number of draws equal to the number of allele counts ($2N$). Since the integration is over a discrete space we calculate the probability that the number of success is greater than the threshold times $2N$. We propose to filter variants using the minimum frequency threshold t such that $FNR(t) < 0.05$. If multiple populations are used in filtering (i.e. removing variants that are common in any population, see below), we employ

a Bonferroni correction for the threshold; that is, we require $FNR(t) < {}^{0.05}/_{K}$ in each of the K considered populations.

3.2.3 Leveraging population structure for improved filtering

We simulate individuals with monogenic disorder by drawing two individuals from a specific 1000 Genomes population and then simulating an offspring assuming Mendelian inheritance and independence between SNPs. We compare three possible disease scenarios (Case-Only, Trio-Dominant and Trio-Recessive) using 40 simulated individuals per scenario and population. The Case-Only scenario assumes there is no information on parental genotypes. The Trio-Dominant scenario assumes that both parental exomes are sequenced and the offspring and one of the parents has the disorder. The Trio-Recessive scenario assumes that both parents are exome sequenced and heterozygous for the causal allele and that the offspring has two copies of the causal allele. Prior to frequency filtering, we remove all variants that do not result in an amino acid change or do not create or remove a stop codon. In addition we remove variants inconsistent with the disease scenario.

We consider multiple frequency-filtering approaches. The $NoAncestry$, f >1% and $NoAncestry$, FNR < 5% approaches estimate allele frequencies and FNRs across all 1000 Genomes individuals. The key intuition here is that statistical noise is decreased with large reference panels but at the cost of ignoring population structure. $NoAncestry$, f > 1% filters out variants with allele frequency > 1% without regard for the FNR; $NoAncestry$, FNR < 5% filters out variants above a threshold determined to ensure a 5% FNR. The $PopMatched$, FNR<5% approach uses only the reference individuals from the sub-continental population (country-level, see 1000 Genomes[39]) of the simulated individual. The $AllPop$, FNR < 5% approach filters out variants observed in any population above a conservative Bonferroni-corrected FNR < 5% threshold. We assume that populations are independent and set the desired FNR for each population to 0.05 divided by the number of populations used in filtering (e.g. 14 for simulation results). $MaxPopFreq$ filters variants if observed above 1% allele frequency in any 1000 Genomes continental population

and is similar to a strategy implemented by ANNOVAR[190] that filters variants if observed above 1% in any 1000 Genomes continental population or the EVS European or African American populations.

For admixed populations we only simulated the Case-Only scenario by using the genotypes of real admixed individuals from 1000 Genomes as case individuals. In addition to methods above, we considered a method that utilizes local ancestry calls (*PopMatched-LA*, FNR<5%). In each individual at loci that are homozygous for African, European or Native American ancestry, we used the corresponding continental allele frequency estimates obtained by averaging across all 1000 Genomes individuals from a given continent. In local ancestry heterozygous regions we used a 50-50 weighting of the matching continental frequencies. We use the maximum continental FNR-based frequency threshold from the African, European and Asian continents as the filtering threshold.

3.3 Results

3.3.1 Modeling statistical uncertainty increases filtering efficacy

We use simulations from the European-American Exome Variant Server (EVS)[166] dataset to assess filtering based on a false negative rate (FNR) as compared to the standard approach of ignoring statistical noise in the allele frequency estimates . We use simulations of various reference panel sizes created with binomial sampling from the frequencies estimated across all the European (or African American) EVS data. As expected, the frequency threshold that maintains a 5% FNR increases as reference panel size decreases (see Figure 3.3). As the maximum frequency of the true causal variant ($max(f_c)$) decreases the number of variants for follow-up per individual also decreases thus increasing filtering performance (see Figure 3.3). Overall, we find a diminishing return in performance for reference panels larger than 500 individuals.

Next, we investigated the FNR attained by the standard approach that ignores statistical noise and filters based on the mean frequency estimate. At small reference panels the

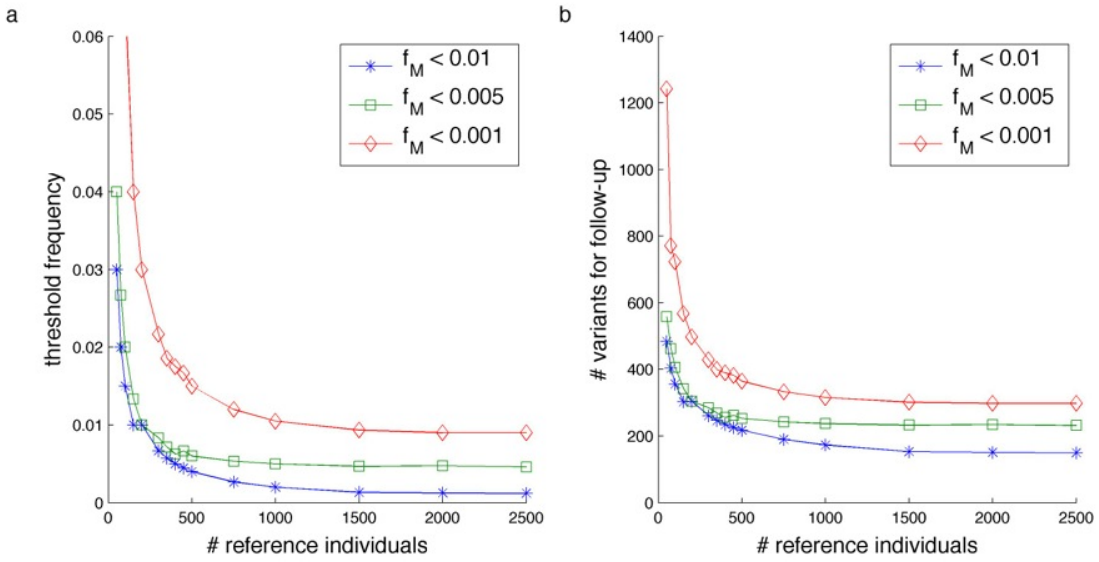

Figure 3.3: Reference panel size impacts the efficacy of filtering in exome sequencing in European simulations from the EVS data. We simulated reference panels at various sizes using a Binomial sampling from the EVS frequencies. Figure 3.3a shows the threshold on the variant frequency needed to achieve a 5% FNR for various assumptions about the maximum frequency of the causal variant in the population (from 0.001 to 0.01). Figure 3.3b displays the number of variants that remain to be followed up post-filtering at a 5% FNR rate. As expected with larger reference panel sizes, the estimated frequency from the reference panel becomes more accurate making the 5% FNR threshold converge to the maximum assumed frequency of the causal variant (f_M) which in turn increases the efficacy of filtering. We observe limited gains in accuracy for reference panels over 500 individuals. Similar results are obtained for simulations of African Americans (see Figure 3.4)

standard approach is mis-calibrated attaining an FNR close to 25% thus removing causal variants from consideration (see Table 3.1). In contrast, the approach that maintains an $FNR < 5\%$ significantly increases the number of variants for follow-up from 298 to 724 on average; this is necessary as it reduces the FNR to the desired 5% (see Table 3.1). With large reference panels the frequency-based approach is conservative (FNR ~0%) yielding twice as many variants for follow-up than the FNR-based approach if the maximum causal frequency is 0.1%. Qualitatively similar results were observed for simulations from the EVS African-American data (see Figure 3.4).

Max True Frequency	Method	100 Reference Individuals			2500 Reference Individuals		
		Threshold	Number of Variants for Follow-up	Probability of FIltering True Causal	Threshold	Number of Variants for Follow-p	Probability of Filtering True Causal
1.00%	(f >1%)	1.0%	298.0	25.4%	1.0%	310.9	2.2%
	FNR <5%	6.5%	724.1	4.6%	0.9%	298.3	4.8%
0.10%	(f >1%)	1.0%	298.0	12.1%	1.0%	310.9	0.0%
	FNR <5%	1.5%	356.0	4.3%	0.1%	149.8	3.6%
0.05%	(f >1%)	1.0%	298.0	8.0%	1.0%	310.9	0.0%
	FNR <5%	1.5%	356.0	1.9%	0.1%	141.2	2.3%

Table 3.1: Method comparisons for different reference panel sizes and maximum causal allele frequencies. We compare two methods. The first is a method (f >1%) that filters out any variants at an observed frequency >1% ignoring the statistical noise on the frequency estimates (and thus the FNR). The second is a method (FNR <5%) that filters out variants if observed above a threshold frequency guaranteed to provide less than a 5% chance of filtering out the true causal variant. At small reference panel sizes it is critical to incorporate statistical noise from the reference panel to not over-filter the true causal variants. Conversely, with large reference panels, a hard 1% frequency filter is too conservative and significantly increases the number of variants remaining for follow-up analysis.

	Variants for Follow-up			Variants for Follow-up with KGGSeq Variants		
	Case-only	Trio-Dominant	Trio-Recessive	Case-only	Trio-Dominant	Trio-Recessive
NoAncestry, f >1%	679.3	330.5	5.2	410.5	200.4	2.8
NoAncestry, FNR <5%	702.1	346.4	5.9	422.5	208.8	3.1
MaxPopFreq	358.3	176.6	1.0	235.4	115.9	1.0
PopMatched, FNR <5%	675.4	332.2	4.2	400.4	196.8	2.2
AllPop, FNR <5%	570.1	279.2	3.1	353.7	173.6	1.7

Table 3.2: Average number of variants that remain for follow-up post-filtering in simulations of non-admixed individuals. All FNR approaches assume the maximal causal variant frequency of 1%. *NoAncestry*, f >1% and *MaxPopFreq* have increased FNRs of 6% and 50% respectively. The *AllPop*, FNR <5% approach outperforms all other FNR-based approaches. The *PopMatched*, FNR <5% approach is the second best performing FNR-based approach demonstrating that the improvements from better population matching outweigh the effects of increased statistical noise from smaller reference panels.

3.3.2 Leveraging ancestry to increase filtering performance

Next, we assessed the performance of filtering with or without accounting for the highly structured nature of rare variants[28, 53, 39, 65]. Using simulated exome data we investi-

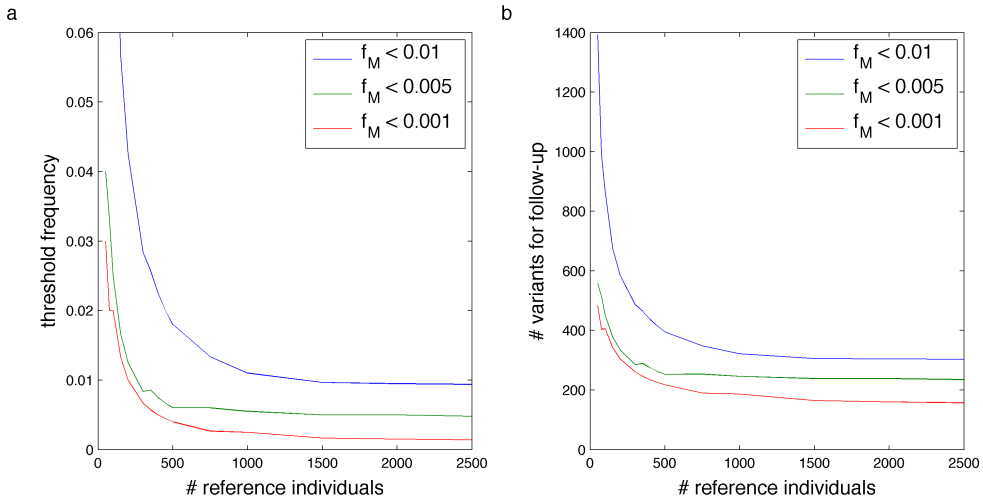

Figure 3.4: Reference panel size impacts the efficacy of filtering in exome sequencing in African-American simulations from the EVS data.(See Figure 3.3 for European simulations.)

gated the efficacy of filtering across a wide range of methods and sequencing studies. When comparing the methods that do not leverage ancestry, we observe that the *NoAncestry*, FNR <5% approach leads to a slightly increased number of variants that need to be functionally followed up over the *NoAncestry*, f >1% approach (Table 3.2). The increased number of variants is necessary to attain a correct 5% FNR rate (*NoAncestry*, f >1% attains an FNR of 6%). The MaxPopFreq approach yields the fewest number of variants by filtering at 1% frequency in any continental population, but has a 50% probability of filtering out the true causal variant (FNR =50%) (see Table 3.2).

Among all methods that maintain an FNR <5% and use ancestry information, we observe that the method that incorporates data across all populations (*AllPop*, FNR <5%) attains the best performance across all simulated scenarios (an average 16% reduction across all populations from the *NoAncestry*, FNR <5% method in the Case-Only scenario, see Table 3.2). The improvement is likely due to variants common in at least one population that are filtered out as unlikely to be pathogenic. This benefit from assaying variants across many populations comes even at the expense of multiple testing correction (a Bonferroni adjustment is made to the FNR required in each individual population resulting in an average filtering threshold of 3.2%). This demonstrates that the benefit of better population

matching outweighs the cost of higher statistical noise from the small reference panels. The greatest improvement from population matching comes with the African populations where there is a 26% decrease in the number of variants remaining for follow-up (see Table 3.3). We also investigated other types of clinical scenarios. As expected, the Trio-Dominant scenario has approximately half as many variants for follow-up as the Case-Only scenario (see Table 3.2). The Trio-Recessive scenario, simulated without inbreeding, shows <6 variants remaining for all scenarios and methods (see Table 3.2). Finally, we observe a similar pattern of improved performance when also filtering non-damaging variants as predicted by KGGSeq (Table 3.2). Therefore, improvements of ancestry-aware filtering do not come preferentially from variants with non-damaging predictions.

3.3.3 Ancestry-aware filtering in admixed individuals

We extend our approach to admixed individuals (e.g. African Americans) with genetic ancestry from multiple continents. We incorporate the local ancestry structure in the filtering step with the $PopMatched - LA$, FNR <5% approach that matches reference panels according to the ancestry at each site in an individuals genome. This significantly lowers the number of variants for follow-up in the admixed populations as compared to the local ancestry naive method ($PopMatched$, FNR <5%) (see Figure 3.5). For example, in African-American individuals we observe a reduction from 664 to 487 variants from just matching the local ancestry to continental populations as compared to using all 1000 Genomes data with a FNR <5%. When using information from all populations in the 1000 Genomes dataset, there is improvement for all admixed populations over the method that ignores ancestry ($NoAncestry$, FNR <5%)(see Figure 3.5).

3.3.4 Ancestry-aware filtering in ClinVar data

In the simulations above, we have made the assumption that the frequency distribution of causal variants is well approximated using the rare variants in our data which may not hold in practice. To investigate deviations from this assumption, we filter the set of Clin-

1000 Genomes Population (Number of Individuals)	*NoAncestry*, FNR <5%	*PopMatched*, FNR <5%	*AllPop*, FNR <5%
ASW* (61)	6645 (78)	487 (33)	514 (53)
CEU (85)	311 (38)	393 (43)	302 (40)
CHB (97)	321 (33)	323 (35)	282 (32)
CHS (100)	322 (16)	317 (17)	282 (16)
CLM* (60)	335 (44)	377 (32)	309 (29)
FIN (93)	289 (19)	312 (28)	264 (18)
GBR (89)	293 (29)	355 (40)	286 (30)
JPT (89)	344 (25)	341 (34)	295 (26)
LWK (97)	833 (32)	610 (31)	605 (29)
MXL* (66)	312 (27)	392 (34)	308 (24)
PUR* (55)	353 (52)	369 (37)	321 (39)
TSI (98)	326 (27)	386 (31)	321 (27)
YRI (88)	765 (26)	566 (23)	547 (23)

Table 3.3: Different levels of genetic diversity across populations induce a variation in the average number of variants remaining for follow-up in an individual. The highest number of variants remaining for follow-up is seen in African populations (YRI and LWK) as well as African-Americans (ASW); this is consistent with these populations have the greatest amount of genetic diversity. These populations also show the greatest benefit from better population matching and from applying the *AllPop*, FNR <5% approach. * denotes admixed populations where results from the $PopMatched - LA$, FNR <5% approach are reported. Standard deviations given in parentheses.

Var pathogenic SNPs according to our methods. We find that the *AllPop*, FNR <5%, *PopMatched*, FNR <5% and *NoAncestry*, f >1% approaches filtered out 42 (3.0% FNR), 18 (1.3% FNR), and 38 (2.7% FNR) of the 1395 variants respectively. This shows that all approaches are conservative with respect to FNR and suggests that by approximating the distribution of frequencies at causal variants from real data we do not artificially increase

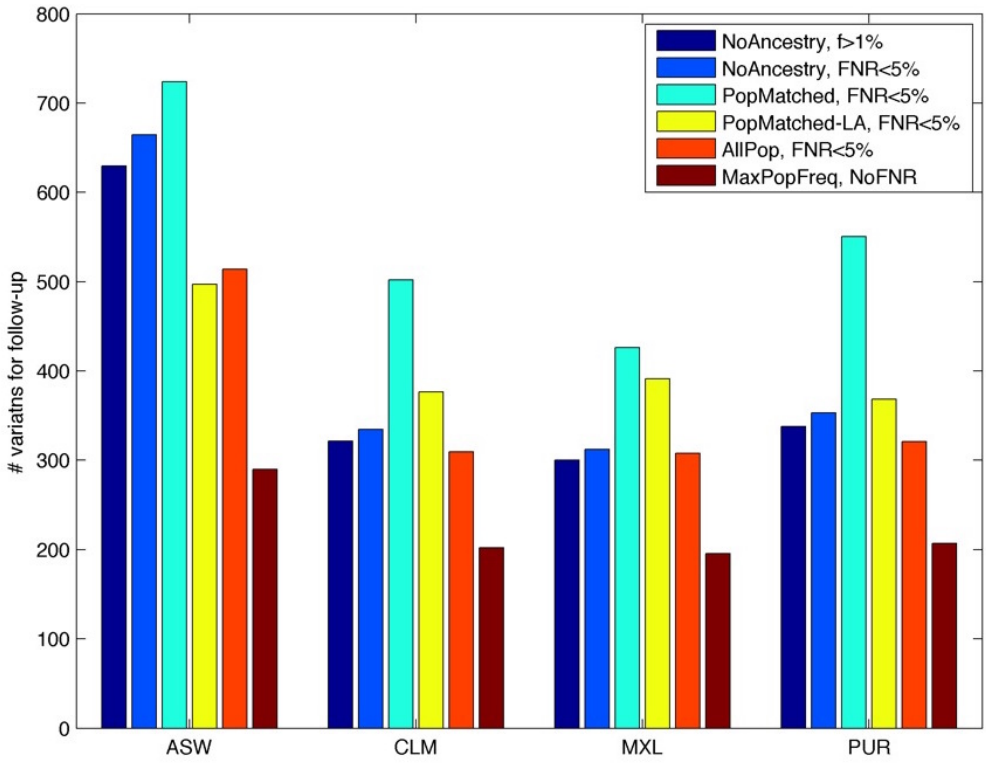

Figure 3.5: Population matching using local ancestry information improves performance over local ancestry nave population matching in admixed populations. The *PopMatched*, FNR <5% approach performs poorly because the admixed reference panel sizes are much smaller than non-admixed reference panels leading to increased filtering thresholds. The *AllPop*, FNR <5% outperforms all other FNR-based approaches.

the FNR in empirical data.

3.3.5 Analysis of 20 exomes of individuals with monogenic traits

To examine the performance of the different filtering strategies when applied to actual data, we used the data from 20 of the 101 real exome sequenced individuals who had monogenic disorders where the causal variants have been previously identified. We assumed a maximum causal allele frequency of 1% for all cases because there was no prevalence data[3]. For all modes of inheritance, the number of variants in an individual for follow-up after filtering was lower when filtering with the *PopMatched*, FNR <5% and *AllPop*, FNR <5% approaches as opposed to the *NoAncestry*, f >1% approach that does not account for the FNR (See Table 3.6). We filtered out all variants except those with damaging annotations:

splice acceptors, stop gains, frame shifts, stop losses, initiator codon changes, inframe insertions, inframe deletions, missense variants, splice region variants and KGGSeq predicted damaging variants. The 101 individuals form a supplemental population with the test individual held out. We included these individuals when estimating average frequencies across all populations in the 1000 Genomes project for the real data. Using our *AllPop*, FNR <5% approach only 468 variants need to be followed up for dominant disorders as compared to 750 for the *NoAncestry*, f >1% approach (see Table 3.4). The true causal variant identified in these individuals was never filtered out. This demonstrates that using multiple population frequency estimates significantly reduces the number of variants remaining for follow-up analysis, while still maintaining an appropriate false negative rate. In Table 3.6 we report the variants remaining in each individual along with country data, presumed inheritance pattern, the zygosity of the causal variant and disease. Removing outliers based on PCA from the 101 self-reported Middle Eastern individuals makes no significant difference in the number of variants remaining for follow-up (see Figure 3.6 and Table 3.5).

Method	Recessive (cases=10)	Dominant (cases=9)	Compound Heterozygous (cases=1)
NoAncestry, f >1%	57.7 (34.8)	749.7 (91.0)	604
PopMatched, FNR <5%	40.1 (32.5)	604.8 (107.1)	426
AllPop, FNR <5%	29.2 (21.4)	467.7 (61.5)	370

Table 3.4: Average number of variants that remain for follow-up post-filtering in real exome studies of 20 individuals with monogenic disorders. None of the filtering approaches removed the true casual variants from consideration. Across all disorder architectures, we observe a significant decrease in the number of variants that need to be followed up if ancestry is incorporated in the filtering step. Parentheses denote standard deviations. Variants were eliminated from consideration as potentially true causal variants if they are not annotated as damaging (see Methods 3.2.1) and if they are not observed twice if the disorder is assumed to be autosomal recessive or at least once if it is assumed to be dominant (heterozygous) or compound heterozygous.

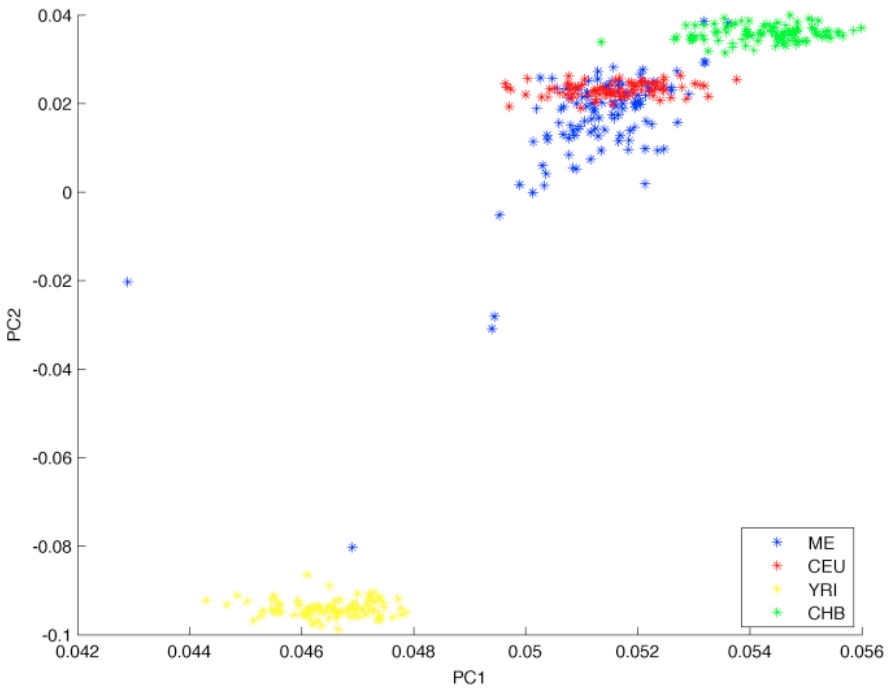

Figure 3.6: Principle component analysis of CEU, YRI, CHB and 101 self-reported Middle Eastern (ME) individuals. Analysis is based on 24,791 variants where there is no missing data for any of the Middle Eastern exomes and where the variants were called in the 1000 Genomes data. The YRI, CEU and CHB populations are well separated and the Middle Eastern population clusters with the CEU and trails towards the YRI. This is similar to what is observed in previous work[201]. Removing the two most extreme ME individuals (far left and far bottom blue dots) results in a marginal increase in the number of variants remaining in the Middle Eastern individuals (see Table 3.5)

3.4 Discussion

In this work, we introduce approaches that account for the finite sample size of the existing reference panels used in filtering while leveraging sub-continental ancestry to improve the filtering step in monogenic disease studies. Both the real data analysis of 20 exomes of individuals with known monogenic disorders and the simulations show that our approaches reduce the number of variants that need to be further investigated, thus increasing the effectiveness of identifying causal variants using exome sequencing of unrelated individuals. This work demonstrates that in a clinical setting, even a small reference panel of 100 individuals from a well-matched population can have significant impact on the filtering efficacy.

Method	Recessive (cases=10)	Dominant (cases=9)	Compound Heterozygous (cases=1)
NoAncestry, f >1%	58.6 (35.4)	758.0 (92.8)	614
PopMatched, FNR <5%	40.8 (33.2)	614.4 (110.8)	431
AllPop, FNR <5%	29.4 (21.5)	471.3 (62.8)	374

Table 3.5: Average number of variants that remain for follow-up post-filtering in real exome studies of 20 individuals with monogenic disorders after controlling the Middle Eastern reference panel by removing the two most extreme PCA outliers.

Our methods are limited in that they do not account for the cases where a second genetic or environmental factor is required for the phenotype to appear and this increases the risk of filtering the true causal if the second factor is rare in some populations. Errors in variant calling in reference populations may also falsely elevate the frequency of a true causal variant and so using multiple technologies for estimating allele frequencies would be a best practice. While our work is presented for use with exome sequencing studies, its central idea will be extendable to whole genome sequencing since rare variation both in and out of the exome will show population clustering.

The current bottleneck in using population structure to help identify rare variants is the limited size of the reference panels for narrowly defined sub-continental populations. Large databases such as the Exome Variant Server could increase their impact if they could report sub-continental allele frequencies in addition to just European and African-American allele frequencies. Recent projects such as the UK10K[128] will be extremely valuable as it is a large reference panel of a specific population. The ALFRED database[32] will also be a very valuable resource for cross-population work with monogenic diseases when it can provide sequencing level data. Founder populations (e.g. Amish or Iceland) where some non-causal variants are pulled to high frequency may be powerful in eliminating non-causal variants if the disease is rare or not present in the founder population[169, 110]. Tools such as Kaviar[61] will allow researchers to quickly search these emerging sources of population

frequency data. Finally, a Bayesian approach to integrate cross-population prevalence, allele frequencies, annotation and functional data in a filter-free probabilistic manner is possible but left to explore in future work.

Ind	Country	Presumed Inheritance	Causal Variant zygosity	Disease	Variants Remaining Pre-Frequency Filtering	$No\ Ancestry$, f >1%	$Pop\ Matched$, FNR <5%	$AllPop$, FNR <5%
1	Jordan	Recessive	Homo	Brachydactyly, type A2	5016	65	43	33
2	Turkey	Recessive	Homo	LADD Syndrome	5049	50	20	20
3	Turkey	Recessive	Homo	Oral-facial-digital syndrome type VI	4752	25	23	20
4	Iran	Recessive	Homo	Spastic paraplegia 11, autosomal recessive	4674	39	15	13
5	Syria	Recessive	Homo	Mitchell-Riley syndrome	4751	26	18	16
6	Tunisia	None Given	Homo	Steroid-11 beta-hydroxylase deficiency	4834	37	19	11
7	Jordan	Recessive	Homo	Microcephaly	4722	22	11	5
8	Jordan	Recessive	Homo	Rickets vitamin D-dependent type 1A	4973	104	102	56
9	Iran	None Given	Homo	Oral-facial-digital syndrome type VI	5653	112	67	52
10	Jordan	None Given	Homo	Cranioectodermal dysplasia 3	5938	97	83	66
11	Jordan	None Given	Het	Microphthalmia, syndromic 5	15694	872	750	532
12	Turkey	Recessive	Het	Split-hand/foot malformation 1 with sensorineural hearing loss	12171	619	445	361
13	Tunisia	Dominant	Het	Palmoplantar MSSE	14111	813	747	505
14	Iran	Dominant	Het	Parkinson's Disease	13727	726	535	446
15	Iran	Dominant	Het	Ataxia, sensory, 1, autosomal dominant	13417	661	512	410
16	Jordan	None Given	Het	Cerebral cavernous malformations 3	14713	871	690	545
17	Jordan7	None Given	Het	Brachydactyly, type B1	13472	668	543	421
18	Jordan	None Given	Het	Anophthamia	13789	748	623	494
19	Iran	None Given	Het	Spinocerebellar ataxia 6	15128	769	598	495
20	Turkey	Recessive	Comp-Het	Oral-Facial-Digital type VI	12919	604	426	370

Table 3.6: Analysis of real data by individual exome. Variants remaining pre-frequency filtering reflects the number of variants remaining when all variants without damaging annotations are filtered out and when variants inconsistent with the disorder architecture are removed but before frequency-based filtering has been preformed.

Ind	Genomic Position (hg19)	Gene	HGVSc	HGVSp	Causal Variant Zygosity
1	4:96035914	BMPR1B	c.188_207delGGTT GCCTGTGG TCACTTCT	p.Val66ArgfsX22	Hom
2	10:123256236	FGFR2	c.1400G>A	p.Gly467Glu	Hom
3	5:37205557	C5orf42	c.3150-1G>T	-	Hom
4	15:44941193	SPG11	c.169_170delCT	p.Leu57AspfsX66	Hom
5	6:117240406	RFX6	c.1129C>T	p.Arg377X	Hom
6	8:143956714	CYP11B1	c.1136G>T	p.Gly379Val	Hom
7	8:6266892	MCPH1	c.114+1G>T	-	Hom
8	12:58157481	CYP27B1	c.1319_1325dup CCCACCC	p.Phe443GlyfsX24	Hom
9	5:37121819	C5orf42	c.5616delA	p.Leu1872Phefs Ter44	Hom
10	14:76455258	IFT43	c.85G>T	p.Glu29X	Hom
11	14:57270998	OTX2	c.181C>G	p.Leu61Val	Het
12	7:96650164	DLX5	c.754T>C	p.Tyr252His	Het
13	-	-	-	-	Het
14	12:40653350	LRRK2	c.1487C>T	p.Thr496lle	Het
15	8:42729096	RNF170	c.191G>A	p.Arg64Gln	Het
16	3:167414800	PDCD10	c.264dupA	p.Glu89ArgfsX6	Het
17	9:94486233	ROR2	c.2543C>T	p.Pro848Leu	Het
18	14:57269015	OTX2	c.382_331delAATG	p.Asn110GlufsX6	Het
19	19:13411385	CACNA1A	c.2261C>A	p.Ala754Glu	Het
20	5:37183490	C5orf42	c.4793C>A	p.Thr1598Lys	Com Het
	5:37226725		c.1972G>C	p.Gly658Arg	

Table 3.7: Identification information for causal variants identified in the real individuals and their zygosity.

Ind	Variant reported in Literature	ExAC	EVS	HGMD variant types	SIFT	Polyphen	Prediction*
1	No[105]	NP	NP	Missense/ nonsense	NA	NA	Lkly pathogenic
2	No[161]	1 het	NP	Missense	Delet	Prb damage	Lkly pathogenic
3	Yes[113]	NP	NP	Missense/ nonsense	NA	NA	Pathogenic
4	No[177]	1 het	NP	Mostly nonsense	NA	NA	Lkly pathogenic
5	No[173]	NP	NP	Missense/ nonsense	NA	NA	Lkly pathogenic
6	Yes[88]	NP	NP	Missense/ nonsense	Delet	Benign	Pathogenic
7	No[45]	NP	NP	Missense/ nonsense	NA	NA	Lkly pathogenic
8	Yes[189]	NP	NP	Missense/ nonsense	NA	NA	Pathogenic
9	No[113]	NP	NP	Missense/ nonsense	NA	NA	Lkly pathogenic
10	No[2]	NP	NP	Missense	NA	NA	Lkly pathogenic
11	No[165]	NP	NP	Missense/ nonsense	Delet	Prb damage	Lkly pathogenic
12	No[168]	NP	NP	Missense/ nonsense	Delet	Poss damage	Lkly pathogenic
13	No[117]	-	-	-	-	-	Lkly pathogenic
14	No[210]	NP	NP	Missense/ nonsense	Toler	Benign	Lkly pathogenic
15	No[2]	NP	NP	Missense	Delet	Prb damage	Lkly pathogenic
16	Np[5]	NP	NP	Missense/ nonsense	NA	NA	Lkly pathogenic
17	No[140]	15 het†	2 het	Nonsense	Delet	Poss damage	VUS
18	No[165]	NP	NP	Missense	NA	NA	Lkly pathogenic
19	No[204]	NP	NP	Missense	Delet	Prb damage	Lkly pathogenic
20	No[113]	3 hom†	3 het	Missense/nonsense	Delet	Prb damage	VUS
		NP	NP		Toler	Prb damage	VUS

Table 3.8: Summary of evidence for identifying causal variants in real exome sequencing data. Variants not found in a database are signified as not present (NP). Variants that have exact matches to published variants are signified as reported in literature with the citation given. Variants without exact matches cite the literature of the gene in which the variant falls that is associated to the phenotype. The HGMD variant types list frame shifts and splice site variants as nonsense variants. *Variants reported before in the literature were predicted to be pathogenic. Variants not reported in literature were evaluated in the context of 1) how well the phenotypes match, 2) is the variant absent or extremely rare in the population, 3) does the variant type match the known or predicted mechanism of how the gene can be disrupted, 4) is the in silico prediction concordant (for missense variants) and predicted to be likely pathogenic if all four were in agreement. †Phenotypic data is not publically available from ExAC database. Patient #17's phenotype is relatively mild and it is likely that 15 individuals in ExAC with the same variant are affected or carriers. One variant in Patient #20 is observed as homozygous in 3 individuals in ExAC but phenotypic data on these 3 individuals are not available. A follow-up functional study is warranted to call the potential compound heterozygous variants likely pathogenic. Information on Individual #13 is withheld because it is a novel finding and in preparation for publication; the citation corresponding to it clinically defines the disease and maps it to one locus.

CHAPTER 4

Enhanced methods to detect haplotypic effects on gene expression

4.1 Introduction

Expression quantitative trait loci (eQTLs) are genetic variants, typically single nucleotide polymorphisms (SNPs), associated with gene expression levels. eQTLs are found through association scans that test for an additive effect of SNPs on expression[149, 179]. In addition to additive effects, effects from interacting SNPs can moderate gene expression[44, 74, 106, 151]. Some types of *cis*-interactions can only be captured by phase-aware methods[23, 46]. However, many estimated interaction effects are explained by un-typed variants or confounders that cast doubt on the importance and prevalence of interactions in humans[52, 74, 196, 75]. Despite this, marginal SNP tests and SNP interaction tests cannot represent the full range of possible genetic architectures that can influence gene expression.

Studies of monogenic disorders (i.e. diseases caused by damaging mutations in a single gene) have been particularly successful in determining the genetic mechanisms responsible for disease. Recessive monogenic disorders can have an underlying compound heterozygous architecture of causal mutations that are usually loss-of-function (LOF) [59, 60, 136]. These architectures arise when a gene is heterozygous at two different positions for LOF variants on different haplotypes. LOF compound heterozygous architectures are known to be important

The work appearing in this chapter is published: Robert Brown, Gleb Kichaev, Nicholas Mancuso, James Boocock, and Bogdan Pasaniuc. "Enhanced methods to detect haplotypic effects on gene expression." Bioinformatics (Oxford, England), 2017.

in complex traits as well [109], but are challenging to detect due to multiple testing issues[58]. Intuitively, for fully penetrant recessive disorders, additional LOF mutations on the same haplotype have no additional effect since the gene function has already been disrupted. With widespread evidence of compound heterozygote architectures in monogenic disorders, in this work we extend such ideas to finding their effects on gene expression.

Transcriptional processes are controlled through multiple layers of genome organization[66, 96, 193, 91]. We hypothesize that specific sets of SNP alleles have *cis*-acting effects[102] on transcriptional processes. Specifically, in this work we assume the effect of having one of the alleles on a haplotype is the same as having multiple. For example, having either of two alleles on a haplotype may have the same effect on an epigenetic state affecting expression from that haplotype as having both alleles on that same haplotype[47, 48, 69, 124, 183]. As an alternative example under this model, if alleles of two SNPs can each alone disrupt the function of an enhancer, then having both alleles on a haplotype will have the same effect as just having one of either allele on that same haplotype. To test this hypothesis, we define compound regulatory predictors (CRPs) that encode the number of haplotypes in each individual carrying at least one alternate allele from a predefined set of SNPs and test for association between the CRPs and gene expression levels. This does not preclude SNPs not in the set from having other independent effects for the same gene. We restrict our analysis to looking at CRPs composed from pairs of two SNPs.

Using simulations of multiple causal architectures, we demonstrate our method is better able to capture the signal from underlying CRP architectures leading to an increased number of eGenes discovered after controlling the false discovery rate (FDR). Importantly, the combined SNP and CRP method has no loss of power relative to the marginal SNP test to detect single causal SNPs.

To investigate the extent of CRPs in real data, we apply our method to data from the GEUVADIS eQTL study[101]. We find that 2,222 of the 3,529 identified eGenes (genes with at least one association) contain both a SNP eQTL and a CRP eQTL. Of these genes, 822 have more of the expression variance captured with a CRP eQTL than a SNP eQTL. Of all eGenes, 974 (27.6%) have a CRP as the top association. There are 153 genes with a CRP

eQTL but no SNP eQTL. Our combined SNP and CRP test finds 29 (0.8%) more eGenes than the marginal SNP test despite a larger multiple testing burden. While this is only a small increase in overall power, the results as a whole demonstrate that some underlying genetic architectures affecting expression are better captured using a compound regulatory predictor model.

4.2 Methods

We start with an overview of our proposed approach. We first regress gene expression on marginal SNP genotypes (the SNP test) in a 1Mb window centered on a transcription start site. We then re-encode genotypes so that the alternate allele is positively associated with expression levels. This way alternate alleles forming CRPs will have the same effect direction on expression levels. We then encode CRPs as the number of haplotypes in an individual with at least one alternate allele at either of two SNP positions ($g_{CRP} \in (0, 1, 2)$). Lastly we regress gene expression on g_{CRP}. To avoid a large increase in the number of tests, we limit multiple testing through a SNP pair selection process.

We illustrate the importance of the CRP model with a toy example in Figure 4.1. Alternate alleles, encoded as g_1 and g_2, can each affect a transcriptional process in such a way as to completely prevent gene expression from the haplotype(s) carrying the alternate allele(s). Since most eQTLs have small to modest effect sizes[1, 68, 101], full loss of expression due to a SNP allele is an extreme example for illustrative purposes and not assumed by our model. Both the SNP test and the SNP x SNP interaction test[44, 106] have reduced power since neither g_1, g_2 nor $g_1 g_2$ are perfectly correlated with g_{CRP}. Since the alternate alleles have a *cis*-acting effect on the transcriptional process, gene expression is dependent both on the genotypes and the phase of the alleles in the special case of (g_1,g_2)=(1,1).

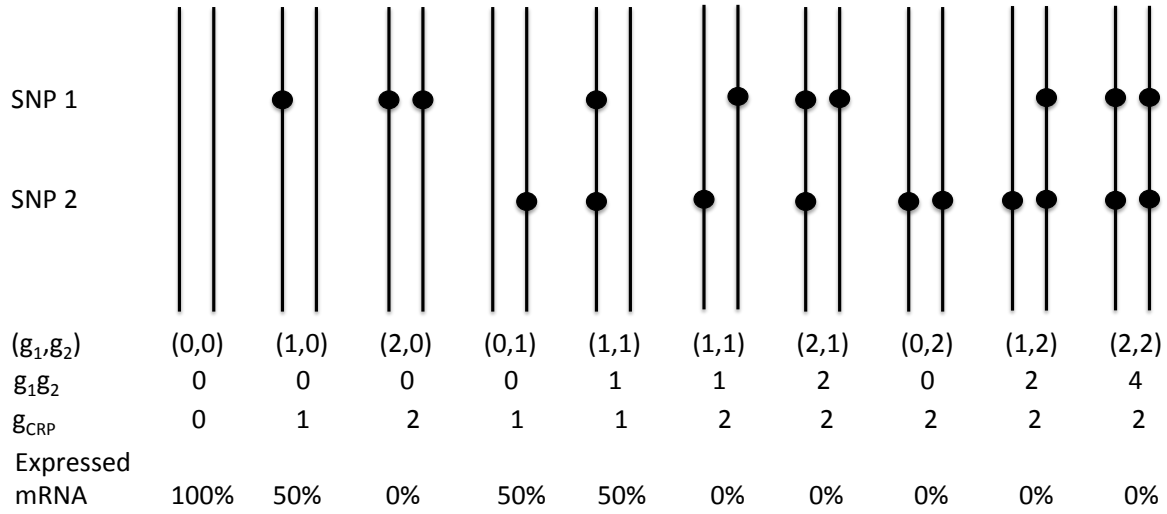

Figure 4.1: Example of a causal CRP architecture. Each pair of vertical bars represents a maternal and paternal haplotype (unordered). A dot represents an alternate allele with g_1 and g_2 denoting the genotypes of the SNPs for an individual. The number of haplotypes carrying at least one alternate allele is given by g_{CRP}. The example phenotype, expressed mRNA, represents the percentage of the maximum amount of mRNA that can be produced and is linearly dependent on g_{CRP}. Full loss of expression due to the alleles is an extreme example for illustrative purposes. The term g_1g_2 represents the product of the two genotypes. The example shows two instances of (g_1,g_2)=(1,1) where the phase will lead to different values for g_{CRP} and expression.

4.2.1 The CRP model

A general additive 2-SNP haplotype model in which each possible haplotype has an effect on the phenotype y is

$$y = \beta_{00}h_{00} + \beta_{10}h_{10} + \beta_{01}h_{01} + \beta_{11}h_{11} + \epsilon \tag{4.1}$$

Here h indicates the number (0, 1 or 2) of each of the four possible haplotypes carried by an individual, β is the effect size of each haplotype, and $\epsilon \sim \mathcal{N}(0, \sigma_e^2)$. The subscripts specify the allele combinations for each haplotype. We focus on the model in which alternate alleles form a CRP ($\beta_{CRP} = \beta_{10} + \beta_{01} + \beta_{11}$).

We introduce a new variable ($g_{CRP} = h_{01}+h_{10}+h_{11}$) to indicate the number of haplotypes containing at least one alternate allele. We rewrite the model in terms of g_{CRP} as

$$y = \beta_{CRP}g_{CRP} + \epsilon \tag{4.2}$$

Given genotype data g_i for SNP_i (or g_{CRP}) and phenotype data y for n individuals, a standard measure of association is the Wald statistic:

$$z_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} = \frac{Cov(g_i, y)\sqrt{n}}{\sqrt{Var(g_i)\sigma_e^2}} \tag{4.3}$$

which asymptotically follows a normal distribution with variance 1 and a non-centrality parameter (NCP) given by

$$\lambda_i\sqrt{n} = \frac{\beta_i\sqrt{Var(g_i)}}{\sigma_e}\sqrt{n} \tag{4.4}$$

The NCP governs the power of rejecting the null hypothesis that there is no association between g_i and the phenotype at a specified family-wise error rate (FWER). Non-causal SNPs ($\beta = 0$) have an induced-NCP if they are in linkage disequilibrium with a causal SNP[79, 89, 97, 155, 192, 206].

Similarly, an induced-NCP can exist for SNPs comprising or tagging a causal CRP. We let x and y* represent mean 0 and variance 1 transformed genotypes and phenotypes and β^* represent the β for the transformed data. We obtain an estimate for each β^* in a linear additive model.

$$\begin{bmatrix} \hat{\beta}_i^* \\ \hat{\beta}_j^* \\ \hat{\beta}_{CRP}^* \end{bmatrix} = \frac{1}{n}\begin{bmatrix} x_i^T & x_j^T & x_{CRP}^T \end{bmatrix} y^* \tag{4.5}$$

$$= \frac{1}{n}\begin{bmatrix} x_i^T & x_j^T & x_{CRP}^T \end{bmatrix} \left(\begin{bmatrix} x_i \\ x_j \\ x_{CRP} \end{bmatrix} \beta^* + \epsilon \right) \tag{4.6}$$

$$= \begin{bmatrix} 1 & r_{i,j} & r_{i,CRP} \\ r_{i,j} & 1 & r_{j,CRP} \\ r_{i,CRP} & r_{j,CRP} & 1 \end{bmatrix} \beta^* + \frac{1}{n}\begin{bmatrix} x_i^T & x_j^T & x_{CRP}^T \end{bmatrix} \epsilon \tag{4.7}$$

$$= V\beta^* + \frac{1}{n}\begin{bmatrix} x_i^T & x_j^T & x_{CRP}^T \end{bmatrix} \epsilon \tag{4.8}$$

We rewrite the $\hat{\beta}^*$ estimates as random variables drawn from a multivariate normal distribution with means given by $V\beta^*$ and variance $\sigma_e^2 V n^{-1}$, in which V is the correlation matrix

of the standardized genotypes.

$$\hat{\beta}^* \sim MVN\left(V\beta^*, \frac{\sigma_e^2}{n}V\right) \tag{4.9}$$

For a causal CRP architecture in which $\beta^* = \begin{bmatrix}0 & 0 & \beta^*_{CRP}\end{bmatrix}^T$, $\lambda = V\beta^*$ is the mean values of $\hat{\beta}^*$,

$$\begin{bmatrix}\lambda_i \\ \lambda_j \\ \lambda_{CRP}\end{bmatrix} = \begin{bmatrix}1 & r_{i,j} & r_{i,CRP} \\ r_{i,j} & 1 & r_{j,CRP} \\ r_{i,CRP} & r_{j,CRP} & 1\end{bmatrix}\begin{bmatrix}0 \\ 0 \\ \beta^*_{CRP}\end{bmatrix} = \begin{bmatrix}r_{i,CRP}\lambda_{CRP} \\ r_{j,CRP}\lambda_{CRP} \\ \lambda_{CRP}\end{bmatrix} \tag{4.10}$$

Here a SNP i that comprises or tags the CRP will appear to have a mean effect size $\lambda_i = r_{i,CRP}\lambda_{CRP}$. The mean effect size λ gives the NCP when testing a SNP or CRP for association with a phenotype for a given sample size.

4.2.2 Correlation between SNPs and CRPs

Haplotype	h_1	h_2	h_{CRP}	Haplotype Probability (p)
h_{00}	0	0	0	$(1-f_1)(1-f_2)+D$
h_{01}	0	1	1	$f_1(1-f_2)-D$
h_{10}	1	0	1	$(1-f_1)f_2-D$
h_{11}	1	1	1	f_1f_2+D

Table 4.1: Two-SNP haplotype characterization. Each 2-SNP haplotype is characterized by the presence or absence of an alternate allele at the first and second SNP position (h_1 and h_2). The variable h_{CRP} indicates if a haplotype carries either of the two alternate alleles. The allele frequencies (f_1 and f_2) and the linkage between the SNPs (D) govern the haplotype probability in a sample.

We calculate the correlation $r_{i,CRP}$ for SNPs from a hypothetical sample where the probability of each 2-SNP haplotype is a function of the allele frequencies and linkage (D) (see Table 4.1). Here two SNPs define each haplotype, with h_1 and h_2 representing the presence of an alternate allele at the first and second SNP positions. The maximum linkage (D_{max}) between SNPs is a function of their allele frequencies (f_1 and f_2) and puts an upper bound on their correlation (r)[76].

$$D = r\sqrt{(1-f_1)(1-f_2)f_1f_2} \tag{4.11}$$

$$D_{max} = \begin{cases} \min\{f_1f_2,\, (1-f_1)(1-f_2)\} & \text{when } D < 0 \\ \min\{f_1(1-f_2),\, (1-f_1)f_2\} & \text{when } D > 0 \end{cases} \tag{4.12}$$

Assuming that haplotypes are inherited independently, there are 16 possible maternal and paternal 2-SNP haplotype combinations for each individual. The haplotype probability (p) is the probability of drawing a specific haplotype with replacement. For each pair of haplotypes (indexed with superscripts k and l) we can compute the probability of the haplotype pair as p^kp^l. The equations $g_i^{k,l} = h_i^k + h_i^l$ and $g_j^{k,l} = h_j^k + h_j^l$ give the genotypes of an individual at SNPs i and j who has one k^{th} and one l^{th} haplotype. The $g_{CRP}^{k,l}$ term, given by $g_{CRP}^{k,l} = h_{CRP}^k + h_{CRP}^l$ is the number of haplotypes in an individual with one k^{th} and one l^{th} haplotype that contain either alternate allele. From these values, the correlation between g_i and g_{CRP} is computed using the following relationships:

$$r_{i,CRP} = \frac{\sum_{k=1}^4 \sum_{l=1}^4 p^kp^l(g_{CRP}^{k,l} - \mu_{CRP})(g_i^{k,l} - \mu_i)}{\sigma_{CRP}\sigma_i} \tag{4.13}$$

$$\sigma_i^2 = \sum_{k=1}^4 \sum_{l=1}^4 p^kp^l(g_i^{k,l} - \mu_i) \quad where\, \mu_i = \sum_{k=1}^4 \sum_{l=1}^4 p^kp^l g_i^{k,l} \tag{4.14}$$

$$\sigma_{CRP}^2 = \sum_{k=1}^4 \sum_{l=1}^4 p^kp^l(g_{CRP}^{k,l} - \mu_{CRP}) \quad where\, \mu_{CRP} = \sum_{k=1}^4 \sum_{l=1}^4 p^kp^l g_{CRP}^{k,l} \tag{4.15}$$

4.2.3 Power analysis to detect CRP effects

Using the model given in Equation (4.2) with the phenotype standardized to have mean 0 and variance 1 ($\sigma_Y^2 = 1$) we compute the power to reject the null hypothesis with a 0.05 significance threshold for a given sample and effect size. Let f_{CRP} be the frequency of risk haplotypes ($f_{CRP} = \mathbb{E}[g_{CRP}]/2$).

$$\sigma_Y^2 = 2f_{CRP}(1-f_{CRP})\beta_{CRP}^2 + \sigma_e^2 \tag{4.16}$$

Let $\sigma_{CRP}^2 = 2f_{CRP}(1-f_{CRP})\beta_{CRP}^2$ such that $\sigma_Y^2 = \sigma_{CRP}^2 + \sigma_e^2 = 1$ where σ_{CRP}^2 is the variance of the phenotype explained by the CRP. We can then estimate the variance of $\hat{\beta}_{CRP}$ and approximate the NCP for the Wald statistic.

$$Var(\hat{\beta}_{CRP}) = \frac{\sigma_e^2}{nVar(g_{CRP})} \approx \frac{\sigma_e^2}{2nf_{CRP}(1-f_{CRP})} \tag{4.17}$$

$$\lambda_{CRP}\sqrt{n} = \frac{\beta_{CRP}}{\sqrt{Var(\hat{\beta}_{CRP})}} = \sqrt{n\frac{\sigma_{CRP}^2}{1-\sigma_{CRP}^2}} \tag{4.18}$$

This result is identical to that of testing a single SNP for association with a phenotype, but uses g_{CRP} as the predictor as opposed to a SNP genotype. Assuming the true causal architecture is a CRP, SNPs will have induced-NCPs given by Equation (4.10). We calculate the power of a test to have a significant association given that the true causal architecture is a CRP:

$$POWER = \Phi\left(\Phi^{-1}(\alpha/2) + \lambda\sqrt{n}\right) + 1 - \Phi\left(-\Phi^{-1}(\alpha/2) + \lambda\sqrt{n}\right) \tag{4.19}$$

Here λ can be either the NCP of the CRP (λ_{CRP}) or the induced-NCP at SNP i ($r_{i,CRP}\lambda_{CRP}$). $\alpha = 0.05/M$ is the desired family-wise error rate (FWER), where M is the number of tests performed for each gene.

4.2.4 CRPs in gene expression data

In order to search for SNP and CRP eQTLs in both real and simulated gene expression, we begin with the SNP test that regresses expression levels on centered and standardized SNP genotypes in a 1 Mb window around the genes transcription start site. Following the GEUVADIS analysis, we included the top three genotype-based principal components (PCs) as covariates as well as a binary variable denoting whether individuals were originally obtained from the 1000 Genomes[39] Phase 1 or imputed. We only use SNPs with estimated maf > 0.05. We re-encode the genotype data so that alternate alleles are positively associated with expression levels.

We limit the number of tests by only performing the CRP test on selected SNP pairs. To select SNP pairs, we look at all possible pairs of SNPs in the window being tested; when both SNPs in the pair pass a suggestive 0.4 significance threshold (Bonferroni corrected based on the number of SNP tests performed) and when each SNP in the pair has $|r_{i,CRP}| < 0.8$, we

test the CRP formed by the SNP pair for association with expression. This process looks for CRPs primarily in genes that already have a significant or near significant marginal association, so it is not expected that this test will greatly increase power.

For real data analysis we determine an empirical p-value using an adaptive permutation procedure following the GTEx approach[68]. We perform at least 1,000 permutations and at most 10,000 permutations. After the first 1,000 permutations, an exit criteria is reached if 15 permutations have a stronger association than the observed association. Therefore, all p-values are estimated with at least 1,000 permutations. For each gene, we permute the expression levels and then rerun the entire SNP test as well as the entire SNP and CRP test including the SNP selection. We then control for a 0.05 false discovery rate (FDR) across genes using the Benjamini-Hochberg procedure. In the real data analysis the largest significant p-value after FDR control was 0.0095, which indicates that all significant genes required more than 1,000 permutations before reaching the exit criteria.

For simulated data, we permute each gene on chromosome 22 10,000 times and use the resulting null distribution of association statistics for each gene to determine the p-values for simulated genes.

4.2.5 Simulations for multiple causal architectures

We base our simulations on the chromosome 22 genotypes of Europeans (n=373) from the GEUVADIS study[101]. We ran Beagle 4.1[21, 19] to impute and phase missing or unphased genotypes for both the simulations and for the real expression analysis. Simulations draw either 0, 1 or 2 SNPs to be causal from a 1 Mb window centered on a randomly drawn transcription start site. After simulating a phenotype (see below), we run the tests as explained in Section 4.2.4. We also run an interaction test where we use an f-test to compare the model containing just the top marginal association to the model the contains the top marginal association as well as the product of the SNP genotypes for the same SNPs that are being evaluated as a CRP. For each causal architecture, we simulate 200 sets of 18,000 genes and report the mean number of genes with a significant association after controlling

for the FDR.

Phenotypes are simulated using an additive model so that the causal genetic architecture explains a fixed $\sigma_g^2 = 0.08$ proportion of the variance in expression. We simulate five underlying causal architectures using either common SNPs with maf>0.05 or rare SNPs with 0.01<maf< 0.05: (1) We randomly choose a single common or rare SNP to be causal. (2) Two causal common SNPs are randomly chosen and each explains half of σ_g^2 after accounting for linkage disequilibrium. (3) A causal CRP formed by two randomly chosen SNPs with either both common or both rare. We also simulate CRPs with two common SNPs but require that the SNPs are correlated either with $r^2 > 0.8$ or < 0.2. The high LD simulation replicates conditions likely seen in a regulatory element where SNPs are often strongly linked. (4) The genotypes of two randomly chosen common SNPs are multiplied to form a causal interaction effect. (5) A null model where the phenotype is simply a draw from a normal distribution. We run the simulations using either masked or unmasked causal SNPs to determine how the methods will perform with un-typed variation and confounders.

4.2.6 Real data analysis

We re-analyzed data from the GEVUDADIS project[101]. Following the original work[101], we filter out non-autosomal genes and genes that did not have >0 quantification in >90% of individuals resulting in 18,621 genes. We standardize the RPKM and PEER normalized gene expression levels sampled from human lymphoblastoid cells after removing non-European data. Lastly, we run the tests as described in Section 4.2.4. We compute a centered and standardized g_{CRP} from the phased genotype data for SNP pairs selected for the CRP test.

To determine if the top CRP eQTL is confounded due to correlation with the top SNP eQTL, we perform conditional regression that removes the effect of the top SNP eQTL if it is has an empirical p-value <0.05. We then re-run the CRP analysis. Significance of the CRP is determined using the permutation method described above.

4.3 Results

4.3.1 Underlying SNPs poorly tag CRPs

The test statistic (z_i) is drawn from a normal distribution with a mean given by the NCP or induced-NCP. Under certain frequency and linkage conditions, the induced-NCPs at SNPs that comprise a causal CRP can be significantly lower than the CRPs NCP (see Figure 4.2 and 4.3). For example, two SNP genotypes (g_1 and g_2) each with maf=0.5 and under no LD ($r_{1,2=0}$) each have a correlation ($r_{1,CRP}$ and $r_{2,CRP}$) of 0.58 with the CPR (g_{CRP}). In this case, if the CRP were causal, the SNP tests induced-NCP is only 58% the NCP of the CRP. This could result in a significant loss of power.

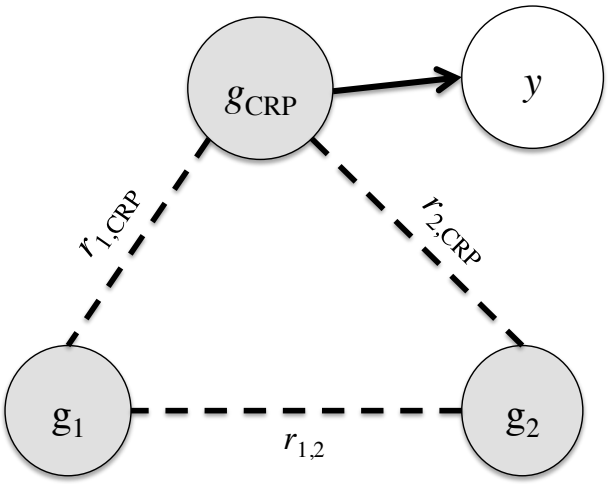

Figure 4.2: Correlation structure between the SNP genotypes (g_1 and g_2) and the CRP (g_{CRP}). The phenotype y is dependent on the CRP.

4.3.2 Power to detect CPRs

We computed the power at a 0.05 Bonferroni corrected significance level of the SNP test and the combined SNP and CRP test to have an association with a trait having an underlying CRP architecture. Because this does not take SNP selection into account for testing CRPs, the combined test results represent the power achievable if testing the causal CRP directly. We correct using the number of SNP (2,266) or SNP and CRP (3,064) tests performed on average per gene in the real gene expression data. As the variance explained due to a causal

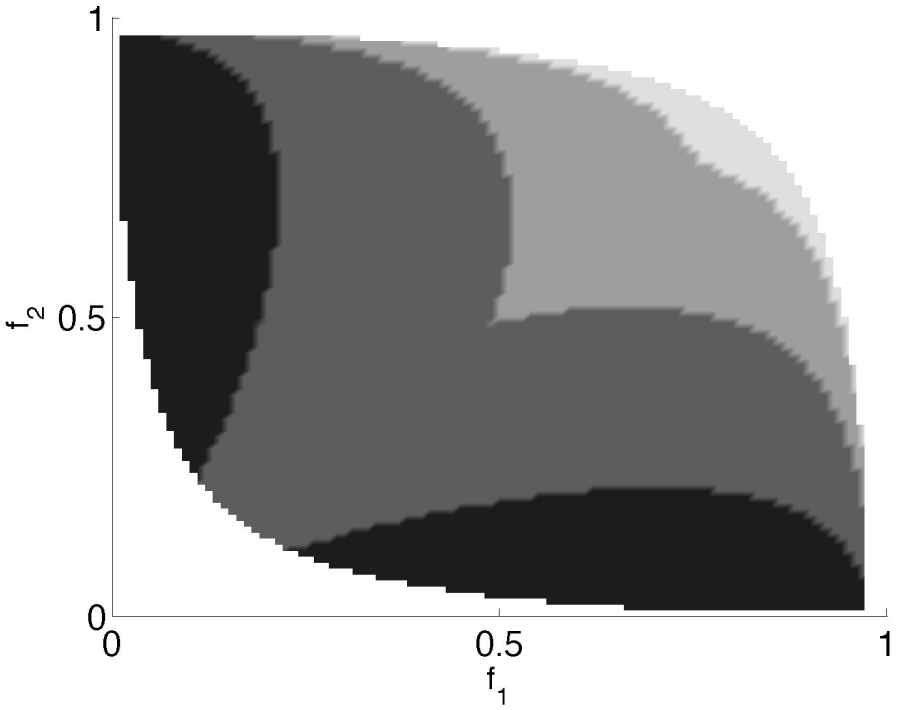

Figure 4.3: The correlation between SNPs and CRPs. The greyscale represents the absolute maximum of $r_{1,CRP}$ and $r_{2,CRP}$ given the SNP frequencies indicated by the x and y-axis and a correlation ($r_{1,2}$) of -0.2 between the SNPs. From darkest to lightest, the greyscale represents absolute maximum $r_{i,CRP}$ from [1,0.75), [0.75,0.5), [0.5,0.25) and [0.25,0].

CRP decreases, the combined test outperforms the SNP test (see Figure 4.4). The combined test has 92% maximum possible power to detect a CRP with $\sigma^2_{CRP} = 0.08$, assuming the CRP is directly tested, as opposed to 62% power with the SNP test.

We simulate gene expression under different causal architectures to evaluate the effect of confounders and to determine how the SNP selection process affects the power of the combined SNP and CRP test versus the marginal SNP test (see Table 4.2). In our simulations, we fix the percentage of phenotypic variance due to the underlying architecture at $\sigma^2_g = 0.08$ and simulate 200 sets of 18,000 genes under each architecture.

Using the null simulations, we observe that the SNP test, combined SNP and CRP test, and the combined SNP and interaction tests have mean false positive rates of 4.97%, 4.96% and 4.97% respectively, each having a standard error of 0.01%. This indicates that the three tests are well calibrated under the null.

We evaluate the power of each test by comparing the average number of genes that have a significant association after controlling FDR at 0.05 in each 18,000 gene set using Benjamini-Hochberg (see Table 4.2). No genes from the Null simulations are significant after controlling the FDR. When the underlying causal model is a single causal SNP, the three tests find approximately the same number of eGenes: 17,404 (SNP test), 17,404 (SNP and interaction test), and 17,405 (SNP and CRP test). Even though the tests find only one difference in the total number of eGenes, the sets of eGenes found by each test are not subsets of the most powerful test. The SNP and CRP test on average finds 7 unique eGenes not included in the set of eGenes found by the SNP test. Interestingly, the rare single SNP causal model is the only simulated architecture where the combined SNP and CRP model is outperformed by the other models, indicting that CRPs poorly tag SNPs with maf < 0.05.

For simulated CRP architectures, the combined SNP and CRP test significantly outperforms the SNP test (and the SNP and interaction test) by finding 93 (and 92) more eGenes. However it found 117 unique eGenes not found by the SNP test indicating that in those genes marginal SNPs were much poorer tags of the underlying CRP. The most extreme example of this is found when looking at CRPs formed by two rare SNPs where the SNP and CRP test finds 211 eGenes not found by the SNP test even though it only finds 46 total more eGenes.

The combined SNP and CRP test has increased power over the SNP test when SNPs forming a CRP are in low LD. In this case the combined test finds 100 more eGenes than the SNP test, 124 being unique to the SNP and CRP test. Conversely, for the high LD CRP simulations, the combined test finds the same number of eGenes with 9 being unique. There is no increase in the total number of eGenes discovered since CRPs formed by high LD SNPs are very well tagged by single SNPs (see Figures 4.2 and 4.3). This also explains why the number of unique eGenes found by the SNP and CRP test is similar to what was found with the single causal SNP architecture.

For the two SNP and interaction architectures, the combined SNP and CRP test significantly outperforms the SNP and combined SNP and interaction tests. This is likely due to the CRP test being able to tag combinations of haplotypes poorly tagged by single SNPs and the fact that the SNP selection method used for both the CRP and the interaction tests,

is optimized for finding CRPs.

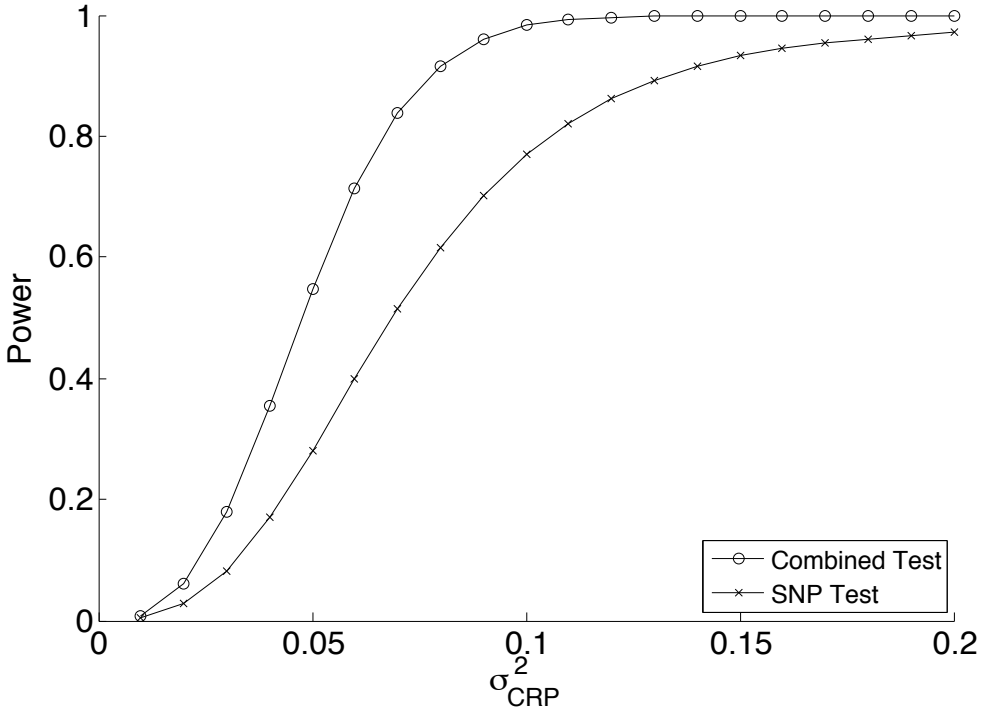

Figure 4.4: Power to detect a causal CRP with 373 individuals and a 0.05 Bonferroni corrected significance level.

4.3.3 CRPs in real gene expression data

Looking at all 18,621 genes that passed the filtering criteria, our combined SNP and CRP test identifies 3,529 eGenes while the marginal SNP test only finds 3,500. Of the 3,529 eGenes, 1,154 have a SNP eQTL but no CRP eQTL and 2,222 have both a SNP and a CRP eQTL. In 37.0% of the 2,222, the CRP eQTL has a larger effect size than the SNP eQTL. For these eGenes, the top CRP eQTL from the combined test on average captures 12.6% of the variance in expression, as opposed to 10.8% with the top SNP eQTL. Finally, 153 identified eGenes have a CRP eQTL but no SNP eQTL. In these eGenes the top CRP eQTL captures 7.1% of the expression variance while the top (not significant) SNP captures 4.9%. These results demonstrate that the combined test is both more powerful than the marginal SNP test and in many eGenes better captures the signal from the genetic effect on expression.

The CRP model makes two predictions. The first is that the mean expression levels of individuals who are heterozygous at the two SNPs (g_1, g_2)=(1,1) that form a CRP will depend on the phase of the alleles. The individuals will have g_{CRP}=1 if the alleles are in phase or g_{CRP}=2 if they are out of phase (see Figure 4.1). The second prediction is that there should be no difference in mean expression levels between individuals with (g_1, g_2)=(2,0) and the individuals with (g_1, g_2)=(0,2). Both of these groups of individuals will have g_{CRP}=2.

There are 887 eGenes with a CRP eQTL where at least four individuals fall into each of the groups. Using a Hochberg-Benjamini FDR control (α=0.05 applied to a t-test, we find 11 CRPs where there is a significant difference in mean expression levels between individuals with (g_1, g_2)=(1,1) and (g_{CRP}=1) and individuals with (g_1, g_2)=(1,1) and (g_{CRP}=2) but found no significant difference between individuals with (g_1, g_2)=(0,2) and those with (g_1, g_2)=(2,0).

In order to determine if the top CRP eQTL tags the top SNP eQTL, we condition gene expression on the top SNP eQTL with an empirical p-value < 0.05 and then re-run the CRP analysis and permutations. This results in 3,218 eGenes where there is only a SNP eQTL, 158 eGenes that contain both a SNP and a CRP eQTL, and 38 eGenes that contain only a CRP eQTL. This analysis shows that while the top CRP eQTLs are highly correlated with the top SNP eQTLs for most genes, in some cases the CRP eQTLs are capturing signal not included with the top SNP eQTL.

After running the SNP test and the combined SNP and CRP test, there are three SNPs of interest: g_m is the SNP that has the strongest marginal association with gene expression, $g_{CRP,1}$ and $g_{CRP,2}$ are the two SNPs that form the top CRP (g_{CRP}). We use Akaike information criterion (AIC) to compare eight models that use the following predictors: (1) g_m, with k=3 (2) $g_{CRP,1}$ with k=3, (3) $g_{CRP,2}$ with k=3, (4) g_{CRP} with k=3, (5) $g_{CRP,1}$*$g_{CRP,2}$ with k=3, (6) g_m and $g_{CRP,1}$*$g_{CRP,2}$ with k=4 (7) g_m and g_{CRP} with k=4 (8) $g_{CRP,1}$ and $g_{CRP,2}$ and $g_{=CRP,1}$*$g_{CRP,2}$. with k=5.

Using AIC we determine if a CRP (g_{CRP}) effect is more likely than a SNP interaction ($g_{CRP,1}$*$g_{CRP,2}$) by comparing models 6 and 7 that each also include the main marginal

effect(g_m). For the 2,222 eGenes with both a SNP eQTL and a CRP eQTL, the model with the CRP is 100 times more likely than the model with the interaction effect for 263 of the eGenes. When looking at the 153 eGenes that only have a CPR eQTL, the model with the CRP is 100 times more likely than the model with the interaction effect in 21 of the eGenes.

We then compare the model that only includes the CRP effect (model 4) to all other models. For the 2,222 eGenes, the CRP only model is most likely compared to all other models in 154 of the eGenes. When looking at the 153 eGenes that only have a CPR eQTL, the CRP only model is most likely in 31 of the eGenes.

4.4 Discussion

In this work we introduce a new method to detect haplotype effects on gene expression. Motivated by monogenic disorders, we extend ideas behind compound heterozygotes to gene expression through a compound regulatory predictor. Our method performs almost identically to the standard marginal SNP methods when the underlying architecture is a single causal, but outperforms it when there are more complex underlying architectures.

The main limitation of our combined test is that it only allows for CRPs composed of two SNPs. It is possible that any number of SNPs affect a transcriptional process or tag haplotypes with similar effect size. Due to the conservative SNP selection process, our method is best able to find CRP associations in genes that already have significant or close to significant associations. It is underpowered to find CRPs when the CRPs are poorly tagged by all marginal SNPs.

Though the gain in power of the combined test over the marginal SNP test is small in real data, the eGenes identified using the CRP model suggest that marginal tests of common SNPs do not fully tag the genetic architectures that influence gene expression. Without comprehensive functional analysis, it is impossible to know if a CRP eQTL causally changes expression levels, or if it is simply tags an un-typed causal variant, interaction, or a more complex causal mechanism.

Given the stronger CRP eQTL signals seen in many genes, our model may be useful for imputing gene expression that can then be leveraged in transcript-wide association studies[55, 70] Finally, the CRP eQTLs motivate two future directions. First, the method can be adapted to increase the power of genome-wide association studies to find novel associated loci. Second, fine-mapping methods used to prioritize potentially causal variants may become more accurate by explicitly modeling CRP architectures.

Causal Architecture	SNP test	SNP and Interaction Test		SNP and CRP Test	
Null	0	0	(0)	0	(0)
		Unmasked			
SNP (c)	17,404	17,404	(0)	17,405	(7)
CRP (c)	14,421	14,422	(1)	14,514*	(117)
CRP (c) low LD	14,402	14,403	(1)	14,502*	(124)
CRP (c) high LD	16,135	16,136	(0)	16,135	(9)
2 SNPs (c)	14,529	14,529	(0)	14,640*	(134)
G_1G_2 (c)	10,477	10,485	(9)	10,538*	(124)
		Masked			
SNP (c)	16,395	16,395	(0)	16,400	(18)
SNP (r)	3,298	3,303	(6)	3,278	(150)
CRP (c)	13,565	13,565	(0)	13,670*	(132)
CRP (r)	2,863	2,864	(2)	2,909†	(122)
2 SNPs (c)	13,701	13,701	(0)	13,824*	(153)

Table 4.2: Average number of eGenes identified after controlling the FDR for different underlying causal genetic architectures. The table reports the mean number of simulated genes with at least one significant association for a given test and simulated causal architecture after controlling the FDR at 0.05. The * represents a significant difference in the number of eGenes discovered between the SNP and CRP test and both other tests and the † represents a significant difference between the SNP and interaction test and the SNP and CRP test (using a t-test with a significance threshold of 0.05/22). The SNP and interaction test was never significantly different form the SNP test. The values in parentheses represent the number of genes found by the specified combined test but included in the set of eGenes found by the SNP test. The (c) and (r) represent architectures using common or rare SNPs.

CHAPTER 5

Haplotype-based eQTL Mapping Increases Power

5.1 Introduction

Expression quantitative trait loci (eQTLs) are genetic variants that regulate gene expression. Association scans to find eQTLs have been successfully applied to multiple datasets to find many eGenes, genes with at least one eQTL regulating their expression. These studies have shown that much of the variation in gene expression is heritable[33, 197, 199, 152] and that the genetic architectures that regulate expression are often found near the gene they regulate[34, 187]. The GEUVADIS[101] and GTEx[68] projects have publically provided genotype and expression data for hundreds of individuals across multiple tissues. These data have been successfully used to probe how genetic variation influences complex diseases through gene expression regulation[70, 55, 42, 209].

The standard test of association assumes an underlying additive model[149, 179] that relates genotype to expression level using a marginal SNP test. eQTLs found through studies using this model may have independent additive effects, but could also have more complicated interactions that cannot be captured by a marginal SNP test. The marginal SNP test does not account for the presence of multiple eQTLs existing for the same eGene[15] or for interactions between eQTLs for the same gene[208]. Recent work[44, 151, 106, 74] has searched for evidence of eQTLs arising from SNPxSNP interactions where one SNP moderates the effect of another. Such interactions are known to exist in yeast[12] but their relevance to gene expression in humans has been difficult to ascertain[196, 75, 14]. Other work has found evidence of haplotype effects when using a likelihood model for short 10 SNP haplotypes[56].

In this work we present a new approach (HapSet) for investigating haplotype effects on gene expression. We hypothesize that each haplotype in a 10 kb region can have a specific effect on gene expression. The HapSet approach divides haplotypes from a region into all possible two set combinations and looks for a difference in the average effect size between the sets. In order to use the marginal SNP approach as a subset of our approach, we force our approach to include haplotype sets defined by the genotypes at single SNP locations. Our method is not biased by filtering or test selection methods that utilize marginal test statistics, since it determines haplotype sets independent of the expression data. We compute significance thresholds for the haplotype set tests in order to control the family-wise error rate (FWER) at desired levels.

We simulate gene expression assuming five underlying architectures: single causal SNPs, multiple causal SNPs, SNPxSNP interactions, haplotypes with non-zero effects and a null model. Our simulations show that both approaches maintain a 0.05 FWER under the null model. With the common SNP-based simulated architectures, the marginal SNP approach has slightly higher power due to a less stringent significance threshold. We expect this result since the SNP approach is a subset of the HapSet approach. For example, when a single common SNP accounts for 5% of the variance in phenotype, the SNP approach has 78% power while the HapSet approach has 71%. However, when the underlying model is based on a random set of haplotypes assigned the same non-zero effect size, the HapSet approach has 71% power compared to the SNP approach that only has 56%. This demonstrates that there is insufficient SNP density to tag many of the possible haplotype combinations.

We apply our method to find eGenes with the GEUVADIS data using expression and genotype data from 373 Europeans individuals. Of the 18,621 genes in the data, both the marginal SNP approach and the HapSet approach identify an overlapping 4,495 genes as eGenes. The marginal SNP approach also finds 606 eGenes not identified by the HapSet approach, and the HapSet approach finds 707 eGenes not identified by marginal SNP approach. Since the SNP tests are a subset of the HapSet tests, the 707 eGenes only identified by the HapSet test are those that the SNP test still could not identify even with a lower significance threshold. This indicates that the single causal model poorly captures the genetic

architecture regulating expression of those genes. Overall, the HapSet approach identifies 101 more eGenes than the marginal SNP approach. This highlights the importance of exploring haplotype-based models both for association studies and for fine-mapping approaches.

5.2 Methods

A SNP represents variation at one specific location in the genome. Haplotypes represent variation across a set of successive SNPs on a single chromosome. Generally, a small number of haplotypes are representative of the majority of the haplotypes in a sample when looking at a small region. Each of these haplotypes can have different effect sizes depending on the genetic architectures occurring on the haplotypes. Our proposed approach seeks to maximize the difference between the frequency-weighted mean effect size on gene expression of haplotypes within a defined set (H) and those not in the set (H^c). Once a set is defined, haplotypes from individuals are identified as either in or out of the set, and a pseudo-genotype can encode for the number of haplotypes an individual carries that are within the set. Gene expression can then be regressed on this pseudo-genotype to estimate the mean difference in effect size of haplotypes in and out of the set.

5.2.1 Haplotype effect model

We divide the genome up into S equal length sections. We assume that the i^{th} haplotype in the s^{th} section has its own specific and independent effect β_{si} on gene expression y

$$y = \sum_{s=1}^{S} \sum_{i} \beta_{si} h_{si} + \epsilon \tag{5.1}$$

Here y is the gene expression for an individual, h_{si} indicates the number of the i^{th} haplotype in the s^{th} section that an individual carries (either 0, 1 or 2) and $\epsilon \sim N(0, \sigma_e^2)$. When only interested in finding associations between y and any of the haplotypes in a region s, we can rewrite our model (Equation 5.2). Here ϵ now includes the variance due to all of the

haplotypes not in region s as well as environmental noise.

$$y = \sum_i \beta_i h_i + \epsilon \tag{5.2}$$

5.2.2 Marginal SNP approach

The marginal SNP approach for identifying an association between a SNP and expression fits the following additive model

$$y = \beta_{gj} g_j + \epsilon \tag{5.3}$$

where g_j is the number of alternate alleles an individual carries at the j^{th} SNP. Assuming the model in Equation 5.2, β_{g_j} will be the difference between the frequency weighted average effect of haplotypes containing an alternate allele at SNP j and the average effect of haplotypes that carry the reference allele:

$$\beta_{g_j} = \sum_i \frac{\beta_i f(h_i) I\left(h_i(g_j) = 1\right)}{f(h_i) I\left(h_i(g_j) = 1\right)} - \sum_i \frac{\beta_i f(h_i) I\left(h_i(g_j) = 0\right)}{f(h_i) I\left(h_i(g_j) = 0\right)} \tag{5.4}$$

Here $I\left(h_i(g_j) = 1\right)$ is equal to 1 if h_i contains an alternate allele at the j^{th} SNP (g_j) otherwise it is 0. Similarly, $I\left(h_i(g_j) = 0\right)$ is equal to 1 if h_i contains a reference allele at the j^{th} SNP (g_j) otherwise it is 0. The frequency of haplotype h_i is given by $f(h_i)$.

5.2.3 Haplotype set approach

The haplotype set (HapSet) approach fits an additive model similar to Equation 5.3,

$$y = \beta_H g_H + \epsilon \tag{5.5}$$

where g_H is a pseudo-genotype for the number of haplotypes an individual carries that are members of a set H of haplotypes (see Section 5.2.5). β_H is the difference between the average effect size of haplotypes in set H and the average effect size of haplotypes not in set H (i.e. in set H^c)

$$\beta_H = \sum_i \frac{\beta_i f(h_i) I\left(h_i(H)\right)}{f(h_i) I\left(h_i(H)\right)} - \sum_i \frac{\beta_i f(h_i) I\left(h_i(H^c)\right)}{f(h_i) I\left(h_i(H^c)\right)} \tag{5.6}$$

Here $I\left(h_i(H)\right)$ and $I\left(h_i(H^c)\right)$ are equal to 1 if haplotype h_i is a member of set H (or H^c) and 0 otherwise.

5.2.4 Testing for significant associations

A standard measure of association is the Wald statistic for an expression level y and genotype (or pseudo-genotype) of SNP j (g_j) for n individuals

$$z_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{Cov(g_j, y)\sqrt{n}}{\sqrt{Var(g_j)\sigma_e^2}} \tag{5.7}$$

which asymptotically follows a normal distribution with variance 1 and a non-centrality parameter given by

$$\lambda_j\sqrt{n} = \frac{\beta_j\sqrt{Var(g_j)}}{\sigma_e}\sqrt{n}. \tag{5.8}$$

Since the $\mathbb{E}(\hat{\beta}_j) = \beta_j$, under the marginal SNP approach, testing the SNP with the largest β_{gj} after standardizing g_j will result in the largest association statistic for the marginal SNP test. Likewise, for the HapSet approach, a set H that maximizes β_H for a standardized g_H will maximize the association statistic for the HapSet approach. This is because for each gene, only the maximum of all the observed test statistics is kept (a Z_{max} like test).

5.2.5 Determining haplotype sets

Since there is no way to know *a priori* the H that maximizes the difference, we test all possible Hs from all 10 kb windows within 500 kb of a genes transcript start site. This results in $\binom{N_H}{2}$ haplotype sets (and statistical tests) where N_H is the number of haplotypes observes in a 10kb window. To determine the Hs, we remove all SNPs with minor allele frequency <0.05 and find all the haplotypes in a 10 kb region. We treat all haplotypes with frequency <0.05 as a single haplotype. If there are more than 8 haplotypes in the group, we use the 7 most frequently occurring haplotypes and treat all the remaining haplotypes as the 8^{th} haplotype. We then compute all possible ways to form two sets (H and H^c). Using the phased genotype data, for each H, we encode for each individual the number of haplotypes they carry that are contained in H as g_H. We compute g_H for each H of every 10 kb region in the genome.

We force the marginal SNP approach to be a subset of the HapSet approach. For each SNP with minor allele frequency >0.05, we define all haplotypes with an alternate allele at

the SNP location to be in set H and all haplotypes with the reference allele at the SNP location to be in set H^c.

5.2.6 Power simulations

We simulate four causal architectures using genotype data from chromosome 22 of the 373 Europeans in the GEUVADIS project. We use the non-transformed genotype data. We simulate using the following models where the proportion of the variance due to the underlying architecture is h^2: null, single causal SNP, two causal SNPs, SNPxSNP interaction and a haplotype set model (Equations 5.9, 5.10, 5.11, 5.12 and 5.13 respectively). Causal SNPs (g_{C1} and g_{C2}) are randomly chosen from SNPs within 500 kb of a transcription start site. All SNPs have minor allele frequency >0.05 except when using a rare single causal SNP with frequency between 0.01 and 0.05. We fix all β values to be 1 and add noise to achieve the desired h^2. The haplotype set model assumes that a random set of haplotypes H all have the same non-zero effect size. We draw the causal H from one of the Hs determined in Section 5.2.5 that is within 500 kb of a transcription start site. When masking the causal haplotypes, we do not test any SNPs or haplotypes within 10kb of the simulated causal haplotypes. We then regress the simulated expression on all genotypes and g_Hs within 500 kb of the transcription start site. We simulate each trait 65,000 times for each h^2 and the null model 650,000 times.

$$y \sim N(0,1) \tag{5.9}$$

$$y = \beta_{C1} g_{C1} + \epsilon \quad where \quad \epsilon \sim N\left(0, Var(\beta_{C1} g_{C1}) \frac{1-h^2}{h^2}\right) \tag{5.10}$$

$$y = \beta_{C1} g_{C1} + \beta_{C2} g_{C2} + \epsilon \quad where \quad \epsilon \sim N\left(0, Var(\beta_{C1} g_{C1} + \beta_{C2} g_{C2}) \frac{1-h^2}{h^2}\right) \tag{5.11}$$

$$y = \beta_{C1C2} g_{C1} g_{C2} + \epsilon \quad where \quad \epsilon \sim N\left(0, Var(\beta_{C1C2} g_{C1} g_{C2}) \frac{1-h^2}{h^2}\right) \tag{5.12}$$

$$y = \beta_H g_H + \epsilon \quad where \quad \epsilon \sim N\left(0, Var(\beta_H g_H)\frac{1-h^2}{h^2}\right) \tag{5.13}$$

5.3 Results

5.3.1 Data for simulations and analysis

Our analyses are based on publically available genotype and lymphoblastoid expression data of 373 European individuals provided by the GEUVADIS[101] project. Following the GEUVADIS project, we only use expression from genes that had >0 quantifications in >90% of individuals. We center and standardize the RPKM and PEER normalized gene expression levels.

5.3.2 Controlling the family-wise error rate

Using all of the g_H values calculated from the chromosome 22 genotypes of the 373 European individuals in the GEUVADIS data and the genotype data for SNPs with minor allele frequency above 0.05, we compute significance thresholds using SLIDE [71]. Since eQTL testing only looks at predictors within 500 kb of each gene, we ensure a FWER level for each gene by scaling the threshold for a 1 Mb region. For a 0.05 desired FWER, this results in a $2.0\text{x}10^{-5}$ significance threshold for our HapSet approach. Running the same procedure using the genotype data for the SNP approach, we estimate a per megabase significance threshold of $5.8\text{x}10^{-5}$. The results from the different FWER levels (see Table 5.1) suggest that the HapSet approach needs to correct for approximately three times as many independent tests in comparison to the marginal SNP approach.

5.3.3 Power analysis

Our simulations show that the marginal SNP approach and HapSet approach are both well calibrated under the null with 0.043 and 0.046 discovery rates when controlling for a 0.05 FWER. The power of both approaches increases as the variance due to the underlying genetic

	SNP	HapSet
0.01 FWER	$9.8\text{x}10^{-6}$	$3.3\text{x}10^{-6}$
0.05 FWER	$5.8\text{x}10^{-5}$	$2.0\text{x}10^{-5}$
0.10 FWER	$1.2\text{x}10^{-4}$	$4.0\text{x}10^{-5}$

Table 5.1: Per megabase significance thresholds estimated with SLIDE. For the 0.01, 0.05 and 0.1 FWERs analyzed, the HapSet approach performs approximately three times as many independent tests as the SNP approach.

architecture increases. The discovery rate of the SNP approach is slightly superior to that of the HapSet approach when the underlying model is based on common SNPs (see Figure 5.1 and Table 5.2). However, if the underlying model is a single rare SNP with minor allele frequency between 0.01 and 0.05, the HapSet approach outperforms the marginal SNP approach. When the underlying model is a random set of haplotypes H, the HapSet approach shows a substantial increase in power over the marginal SNP approach (see Figure 5.2 and Table 5.2). For example, for a h^2=0.05 and a haplotype set architecture, the SNP approach has 56% power compared to 71% for the HapSet approach (see Table 5.2). When the set of causal haplotypes is masked, the HapSet approach still outperforms the marginal SNP approach.

	Null (h^2=0)	Common SNP	Two SNPs	SNPxSNP Interaction	Rare SNP	Haplotype Set	Haplotype Set (Masked)
SNP	0.043	0.78	0.35	0.75	0.20	0.56	0.43
HapSet	0.046	0.71	0.30	0.73	0.22	0.71	0.47

Table 5.2: Discovery rate of the marginal SNP and HapSet approaches when h^2=0.05. The marginal SNP approach outperforms the HapSet approach when the underlying genetic architecture is based on SNPs. The exception is when the underlying architecture is based on a rare SNP with allele frequency between 0.01 and 0.05. When the architecture is based on haplotype sets, the HapSet approach strongly outperforms the SNP approach with 71% discovery rate compared to a 56% rate for the SNP approach. When the haplotype sets and SNPs within 10 kb of the simulated casual haplotype set are masked, the HapSet method still outperforms the SNP approach. This indicates that there is not enough SNP density to tag some haplotype combinations.

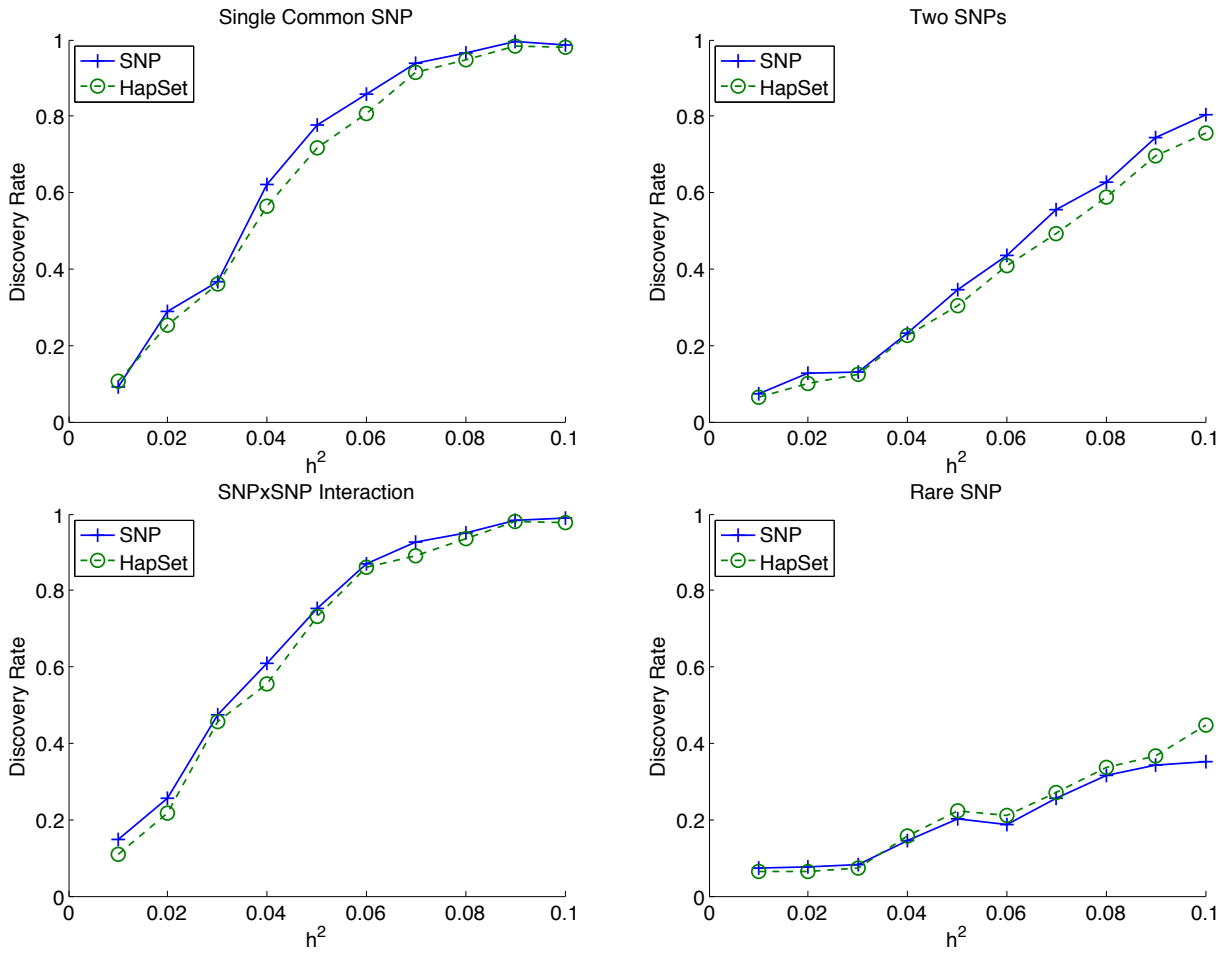

Figure 5.1: Discovery rate of the marginal SNP and HapSet approaches with SNP-based simulated architectures. The proportion of variance due to the underlying genetic architecture is given by h^2. The HapSet approach slightly outperforms the SNP approach for causal SNPs with minor allele frequencies between 0.01 and 0.05. Since the HapSet approach has a more stringent significance threshold to control the FWER at 0.05, this indicates that the HapSet approach is better tagging the causal SNP than the SNP approach. For all other SNP-based causal architectures, the HapSet method performs slightly below the SNP approach. Since the SNP approach is a subset of the HapSet approach, this is due only to the difference in significance thresholds.

5.3.4 Analysis of GEUVADIS data

We apply the two approaches to the GEUVADIS gene expression data. Following the GEUVADIS analysis, we regress expression on the genotypes or HapSets (both standardized to mean 0 and variance 1) while controlling for the top three genotype-based PCs[101]. We use the 18,621 genes that passed the filtering criteria. We report in Table 5.3 the number

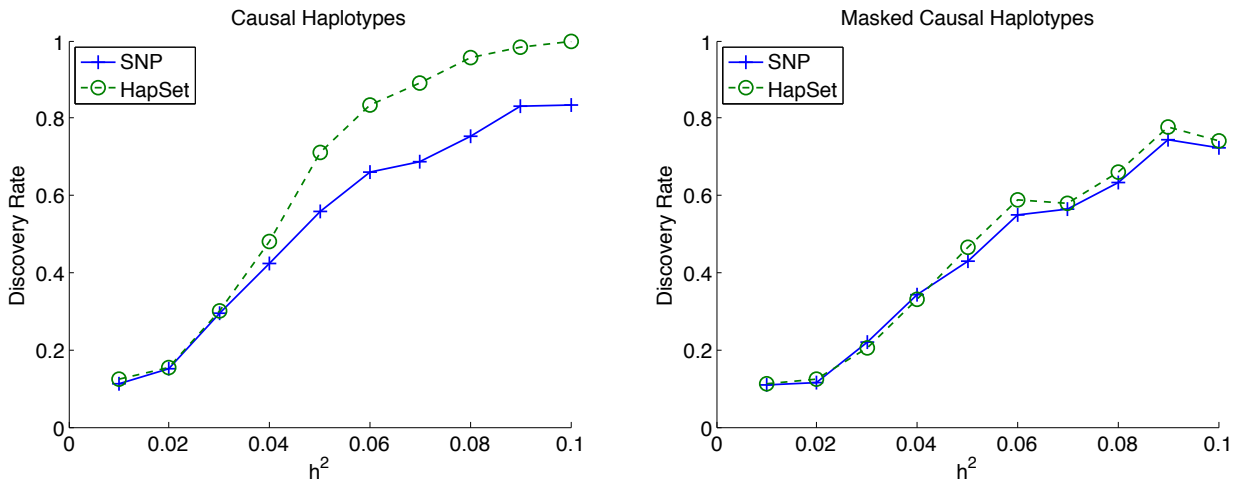

Figure 5.2: Discovery rate of the marginal SNP and HapSet approaches with haplotype-based simulated architectures. When a random set of haplotypes has a non-zero effect size, the HapSet approach has a large power advantage over the marginal SNP approach. When the set of causal haplotypes is masked, the HapSet approach slightly outperforms the SNP approach.

of eGenes found by the SNP approach and the Haplotype approach while controlling for different FWER levels.

When controlling for a 0.05 FWER, the two approaches found 4,495 overlapping eGenes (see Table 5.3). The HapSet approach also found 707 eGenes that were not detectable using genotype data. Since the SNP approach is a subset of the HapSet approach, yet has a less stringent significance threshold, this indicates that marginal SNPs poorly tag the regulatory architecture for these genes. Conversely, the 606 eGenes found only by the SNP approach represent eGenes that had p-values smaller than $5.8\text{x}10^{-5}$ (the SNP approach threshold for a 0.05 FWER) but larger than $2.0\text{x}10^{-5}$ (the HapSet threshold for a 0.05 FWER). In total, there are 101 more eGenes identified by the HapSet approach than by the marginal SNP approach. Using a more stringent 0.01 FWER shows that the marginal SNP test identifies 20 more eGenes than the HapSet approach. However, by allowing for a looser FWER of 0.1, the HapSet approach again outperforms the marginal SNP approach and identifies 112 more eGenes.

We evaluate the average squared effect sizes for the top SNP or HapSet for each eGene when controlling for a 0.05 FWER. The 4,495 eGenes found by both approaches had mean

	Only SNP	Both	HapSet
0.01 FWER	348	3,408	328
0.05 FWER	606	4,495	707
0.10 FWER	874	5,157	986

Table 5.3: Number of eGenes identified by the marginal SNP and HapSet approaches while controlling for a given FWER. The marginal SNP approach identifies 20 more eGenes than the HapSet approach when using a 0.01 FWER, but with the 0.05 and 0.1 FWERs, the HapSet approach identifies 101 and 112 more eGenes respectively.

squared effect sizes of 0.097 and 0.099 for the SNP and HapSet approach, respectively. The 606 eGenes only found by the SNP approach had effects sizes of 0.045 (SNP) and 0.045 (HapSet). The SNP approach is a subset of the HapSet approach but has a less stringent significance threshold, which explains why these eGenes were not found by the HapSet approach but have the same mean squared effect size. The mean squared effect sizes for the eGenes found only by the HapSet approach have the largest difference of 0.034 (SNP) and 0.055 (HapSet). This difference may indicate that there are underlying causal architectures that the HapSet approach is capable of effectively tagging but the marginal SNP approach cannot identify. We regress out the effect of the top associated HapSet for the 707 eGenes found only by the HapSet approach and compute the mean squared effect of the same top (though not significant) SNP associations with the residuals. We observe a mean squared effect of 0.013. This indicating the top SNPs were largely picking up on signal better captured by the HapSet approach.

5.4 Discussion

In this work we present a new framework for tagging additive haplotype effects that are insufficiently captured using the standard SNP-based regression approach. We implement our method and demonstrate that it is well calibrated under the null hypothesis. We also show that it has little loss of power due to an increase in the multiple testing burden for common SNP architectures, but it has increased power for SNPs with <0.05 minor allele

frequency and haplotype-based architectures. When we apply our approach to real gene expression data, we see increased power to detect eGenes as compared to the standard marginal SNP approach. Most importantly, we observe a large number of eGenes that are detected by the HapSet approach yet undetected by the SNP approach. Since the SNP approach is a subset of the HapSet approach, but with a less stringent significance threshold, this indicates that in many genes there are effects that are not well tagged by common SNPs.

Our method gives no indication of the specific causal architectures underlying each of the eGenes found with the HapSet approach. Further analyses must be performed in order to identify if there are multiple independent causal SNPs, a SNPxSNP interaction or a more complicated haplotype based effect. It is possible that our method may be identifying rare SNPs or un-typed genetic variation. Confounding from such variables has been observed in tests associating SNPxSNP interactions with gene expression[74, 196, 75, 52].

Future work can implement a version of our approach for identification of associated loci in complex traits. Many complex traits have GWAS loci that overlap known eQTLs[137], and applying the HapSet approach to complex trait data is a natural extension. The second direction is to apply the HapSet approach to fine mapping. Modeling more than 2 or 3 causal SNPs in a region for fine mapping can be computationally challenging[89, 175, 195, 50, 79, 148, 54]. However, knowing which combinations of haplotypes are associated, and to what degree, may improve efficient selection of SNPs and interactions between SNPs for testing.

CHAPTER 6

Detecting causal gene-on-gene regulatory effects acting through expression level

6.1 Introduction

Over the past decade, expression quantitative trait loci (eQTL) studies have identified many single nucleotide polymorphisms (SNPs) that regulate gene expression levels [67, 180, 40]. These studies have shown that a substantial amount of gene expression variation is accounted for by SNPs close to the gene, referred to as *cis*-eQTLs. Furthermore, a central interest of these studies is to identify SNPs affecting expression of genes that are distant in the genome, referred to as *trans*-eQTLs. One possible mechanism of these *trans*-eQTLs is that they are *cis*-eQTLs for one gene; in other words, the expression of the *cis*-gene interacts directly with a distant gene. We refer to this as a gene-on-gene effect. Identifying gene-on-gene effects is a step forward in constructing gene regulatory networks. A major problem in identifying these effects is the presences of noise created by unobserved confounding factors. In many cases, the noise can create correlation between gene expressions not due to the genetic variation. [37, 49, 11].

If the confounder influences expression from two genes in the data that have no gene-on-gene effect, the genes may appear to be correlated to each other even though this correlation arises only from the confounder. It has been well documented in the literature that confounding factors greatly affect gene expression levels and induce observed correlation between genes [104, 87, 176, 84]. Common confounders include gender, shared *trans*-regulators or environmental effects and population structure. Traditional statistical inference methodologies correct for the known confounding effects such as gender and population structure by in-

cluding them (or their proxies) as covariates in the analysis. However, unknown confounding effects cannot be corrected for in this way. Even after correcting for known confounders, an observed correlation between gene expression levels cannot be interpreted as a gene-on-gene effect since the correlation may be due to an unobserved confounder.

Several existing methods use a causal inference framework to infer the relationships between genes[164, 29, 85, 132]. The overarching difficulty of these approaches is that a confounding factor, either biological or technical, may affect the expression of both genes and obscure the causal relationships. For example, the Chen et al. [29] method considers a"triple" of a SNP and two genes in yeast, and it determines regulatory networks based on the probability that one gene regulates another. Similarly Neto et al. [132] use quantitative trait loci (QTLs) to infer the causal relationships in correlated phenotype networks. While different, these approaches attempt to intuitively identify if there is a causal relationship between the two genes by examining the joint distributions and conditional distributions between the genes and the SNP. These approaches are related to Mendelian randomization, which is rooted in classical randomized control trial theory and causal inference [147], where the SNP is treated as an instrumental variable.

We leverage insights from recent eQTL studies in our approach to estimate the direction and the magnitude of causal relationships between genes. Because wide-spread *cis*-eQTLs are present throughout the genome, we focus on a quartet of two genes and their corresponding two *cis*-eQTLs. Our quartet idea is related to the use of multiple instrumental variables in Mendelian randomization (referred to as bi-directional Mendelian randomization). These types of structures have been of interest in the community (e.g. see Kreimer and Pe'er 2014 [98]). To examine our method, we utilized both simulated and real datasets. When applied to several simulated datasets, our method is robust to different levels of confounding effects. From the simulations it is clear that when we apply our method to the Genotype-Tissue Expression (GTEx) data where at mosts there are only 338 individuals per tissue sample, the method will only have power to identify very strong gene-on-gene effects. Applied to four tissues in the GTEx data [40], we find a single gene pair that is significant after controlling the false discovery rate (FDR) at $\alpha = 0.05$.

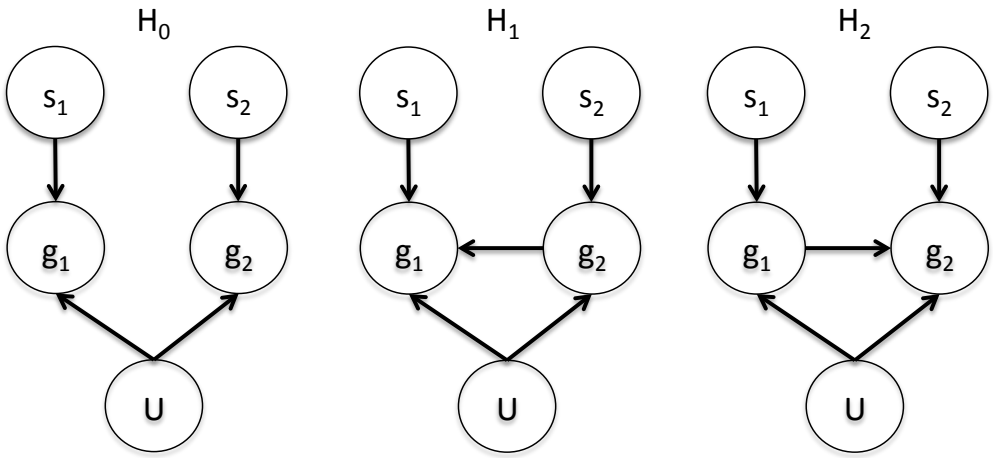

Figure 6.1: Possible causal graphs relating two eGenes, g_1 and g_2. s_1 is the *cis*-eQTL of g_1, s_2 is the *cis*-eQTL of g_2. We use U to represent all unobserved factors u_1 and u_2. u_1 and u_2 may have unknown correlation ρ that may create correlation between g_1 and g_2 in all models, even when there is no gene-on-gene effect.

6.2 Methods

6.2.1 Overview

eQTL studies have verified several genes with *cis*-acting SNP effects, referred to as eGenes [40]. We model the relationship in a quartet of two eGenes and their two *cis*-acting SNPs, including unknown confounding factors that affect both of the eGenes. Figure 6.1 shows the possible causal graphs for a quartet. U represents u_1 and u_2 that are the unobserved factors and the noise that may or may not be correlated. This noise creates correlation in g_1 and g_2 that is not due to a gene-on-gene effect.

H_0 in Figure 6.1 shows the case when there is no gene-on-gene effect between the two eGenes, however, the unknown confounding effects may induce indirect correlations between the eGenes. H_1 and H_2 in Figure 6.1 show causal graphs when the two eGenes contain direct gene-on-gene effects from one to another. Regressing g_2 on g_1, or the reverse, to determine a gene-on-gene effect would be confounded and result in a false positive. However, regressing g_2 on the *cis*-component of g_1 due to s_1 will not be confounded because there is no back door path connecting s_1 to U. This is because s_1 is subject to Mendelian randomization. Using this approach and the known *cis*-effects, we can estimate the gene-on-gene effects and

perform permutations to establish significance.

6.2.2 Generative model

Our method is based on a linear model framework with the following generative model:

$$g_1 = \beta_{s_1 g_1} s_1 + \beta_{g_2 g_1} g_2 + u_1 \tag{6.1}$$

$$g_2 = \beta_{s_2 g_2} s_2 + \beta_{g_1 g_2} g_1 + u_2 \tag{6.2}$$

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix}\right) \tag{6.3}$$

Let n be the number of samples, then g_i is a vector of length n, which contains the expression levels of gene i, where each element in g_i represents the expression of gene i on individual. s_i is a vector of length n, which contains the genotype values of a SNP that *cis*-regulates gene i and has mean 0 and variance 1. $\beta_{s_1 g_2}$ is the *cis*-effects of s_1 on g_2, $\beta_{s_2 g_1}$ is similar. $\beta_{g_1 g_2}$ is the gene-on-gene causal effect of g_1 on g_2 and $\beta_{g_2 g_1}$ is similar. u_1 and u_2 are the noise of g_1 and g_2 and may be correlated due to unobserved confounders. u_1 and u_2 follow a multivariate normal distribution with means 0 and a variance covariance structure such that u_1 and u_2 have a correlation of ρ and variances of σ_1^2 and σ_2^2.

We assume that the gene-on-gene effect can only be in one direction (no feedback and therefore require that either $\beta_{g_1 g_2}$ or $\beta_{g_2 g_1}$, or both, be zero. Assuming that $\beta_{g_2 g_1} = 0$, we can rewrite Equation 6.1 and substitute it into Equation 6.2 for the g_1 term as follows:

$$g_1 = \beta_{s_1 g_1} s_1 + u_1 \tag{6.4}$$

$$g_2 = \beta_{s_2 g_2} s_2 + \beta_{g_1 g_2} \beta_{s_1 g_1} s_1 + \beta_{g_1 g_2} u_1 + u_2 \tag{6.5}$$

From these equations, the effect of g_1 on g_2 is broken into the two components $\beta_{g_1 g_2}\beta_{s_1 g_1} s_1$ and $\beta_{g_1 g_2} u_1$. We assume that s_1 and s_2 are independent of each other and of u_1 and u_2, therefore regressing g_2 on s_1 will have an expected regression coefficient of $\beta_{g_1 g_2}\beta_{s_1 g_1}$.

6.2.3 Gene-on-gene effect size estimation

In all three models shown in Figure 6.1 there is no variable that d-separates g_1 and g_2. Since u_1 and u_2 are unobserved with unknown correlation and cannot be conditioned on, g_1 and g_2 can be dependent on each other due to the confounding effects even though there is no gene-on-gene effect between g_1 and g_2. Since s_1 is d-connected to g_1 and d-separated from g_2, we use it as an instrumental variable to determine the effect of g_1 on g_2.

We are able to use s_1 and s_2 as instrumental variables because they are subject to Mendelian randomization [103, 172, 25], which can be assumed to be independent of the traits of interest. We can thus assume that the observed genotype values are independent of any observed or unobserved confounders that would create correlation between g_1 and g_2. One assumption of this method is that the only way the instrumental variable s_1 can influence g_2 is through the modeled causal pathway. In order to assure this in real data, we require that gene pairs be on different haplotypes. Thus any observed effect cannot be a *cis*-effect.

The following procedure is used in our method to determine if there is a direct effect of g_2 on g_1 such as in H_2 in Figure 6.1. We separately regress g_1 and g_2 on the instrumental variable s_1 to fit the following regression equations:

$$g_1 = r_1 s_1 + \varepsilon_1 \tag{6.6}$$

$$g_2 = r_2 s_1 + \varepsilon_2 \tag{6.7}$$

where r_1 is the ordinary least squares estimate of $\beta_{s_1 g_1}$, r_2 is the estimate of $\beta_{g_1 g_2}\beta_{s_1 g_1}$. ε_1 and ε_2 represent the residuals. To reduce the noise in the estimates, first we regress out the *cis*-effect of the opposite gene. For example, we regress s_2 out of g_2, then regress the residual on s_1. This removes a portion of the variance in g_2 due to the $\beta_{s_2 g_2} s_2$ term in Equation 6.2 and Equation 6.5, resulting in a less noisy estimate of r_2. The correlation between u_1 and u_2 will also affect the variance in the estimate of r_2 but will not affect the expected estimate.

Since s_1 and s_2 are both independent of u_1 and u_2 (see Equations 6.4, 6.5, 6.8 and 6.9), $E(r_1) = \beta_{s_1 g_1}$ and $E(r_2) = \beta_{g_1 g_2}\beta_{s_1 g_1}$, the ratio $E(r_2/r_1) = \beta_{g_1 g_2}\beta_{s_1 g_1}/\beta_{s_1 g_1} = \beta_{g_1 g_2}$ provides an

estimate of the effect of g_1 on g_2. This allows us to circumvent the effect of confounding due to the correlation of u_1 and u_2. We use the same approach to estimate $\beta_{g_2g_1}$. We assume that there is no feedback, therefore only one of the gene-on-gene effects can be non-zero. We take the absolute maximum of $\beta_{g_1g_2}$ and $\beta_{g_2g_1}$ to have a non-zero value.

6.2.4 Estimate p-values through permutations

In order to determine if $\beta_{g_1g_2}$ or $\beta_{g_2g_1}$ are significant, we use a permutation approach. We permute the labels of one expression and a set of genotypes from one of the genes. While this maintains the *cis*-effects of SNPs on gene expression levels and the observed distributions of genotypes and expressions, it breaks any gene-on-gene effect and the covariance in the expression levels (see Section 6.3.1 for further discussion). We then estimate $\beta_{g_1g_2}$ and $\beta_{g_2g_1}$ for each permuted dataset and record the absolute maximum of those. We repeat this procedure 2,000 times for the power analysis and 50,000 times for the GTEx data analysis. We estimate an empirical p-value as the percentage of permutations with larger gene-on-gene effect than in the unpermuted data.

6.2.5 Simulated Data

We simulate data using the generative model described in Section 6.2.2. We draw s_1 and s_2 from standard normal distributions. We fix the *cis*-effects to be $\beta_{s_1g_1} = 0.2$ and $\beta_{s_2g_2} = 0.2$ Since our estimate of $\beta_{g_1g_2}$ is based on the ratio r_2/r_1, we want to avoid very small estimates of r_1 as the ratio would approach dividing by 0. We then choose values for $\beta_{g_1g_2}$ of 0, 0.15, 0.2 and 0.9. The correlation between u_1 and u_2, ρ, is set to be -0.9, -0.5, -0.2, 0, 0.2, 0.5 or 0.9. Finally, we solve for σ_1 and σ_2 from the following equations based on taking the variance of Equations 6.4 and 6.5 and recognizing that the variance of the all s_i and g_i are set to 1:

$$1 = \beta_{s_1g_1}^2 + \sigma_1^2 \tag{6.8}$$

$$1 = \beta_{s_2g_2}^2 + \beta_{g_1g_2}^2\beta_{s_1g_1}^2 + \beta_{g_1g_2}^2\sigma_1^2 + \sigma_2^2 + 2\beta_{g_1g_2}\rho\sigma_1\sigma_2 \tag{6.9}$$

We then have all the parameters for simulating under the generative model laid out in Equations 6.1, 6.2 and 6.3. For each set of simulation parameters, we estimate the family-

wise error rate (FWER) by setting $\beta_{g_1 g_2} = 0$) and the power by setting $\beta_{g_1 g_2} \neq 0$). We use 24,000 simulations for each set of parameters.

6.2.6 GTEx Data: Tissues and Filtering

We apply our method to four of tissues from the GTEx project: Whole Blood from 338 samples, Muscle-Skeletal from 361 samples, Artery-Tibial from 285 samples and Adipose-Subcutaneous from 298 samples. We transform the genotype data and the expression data for each gene to have mean 0 and variance 1 within each tissue. We regress out the top three principle components. We do not assume that they correct for all of the confounding. We then filter out all genes where the eQTL has an absolute effect size < 0.2. Since genes that have highly correlated expression are likely to be in the same network, it is probable that a gene in a network can have a regulatory effect on other genes in that same network. Therefore we only look for gene-on-gene effects between genes that have absolute expression correlations of at least 0.8. In order to ensure that a SNP eQTL does not have a direct *cis*-effect on both genes (which would lead to a false positive), we require that the genes in each pair be located on different chromosomes. In the future, a Hi-C study could also confirm that there is no interacting chromatin region near the two genes. We assume that this also eliminates the correlation between s_1 and s_2. We summarize the effects of this filtering scheme in Table 6.3. We then apply our method as described in Sections 6.2.3 and 6.2.4 to estimate the magnitude, direction and significance of the gene-on-gene effects for each gene pair. We use the Benjamini-Hochberg procedure to control for a 0.05 FDR of gene-on-gene effects in each tissue.

6.3 Results

6.3.1 Simulated Study

Using the model described in the Section 6.2.2, we simulated gene-on-gene effects in 24,000 simulations for each combination of parameters. We fix $\beta_{s_1 g_1} = \beta_{s_2 g_2} = 0.2$ and draw

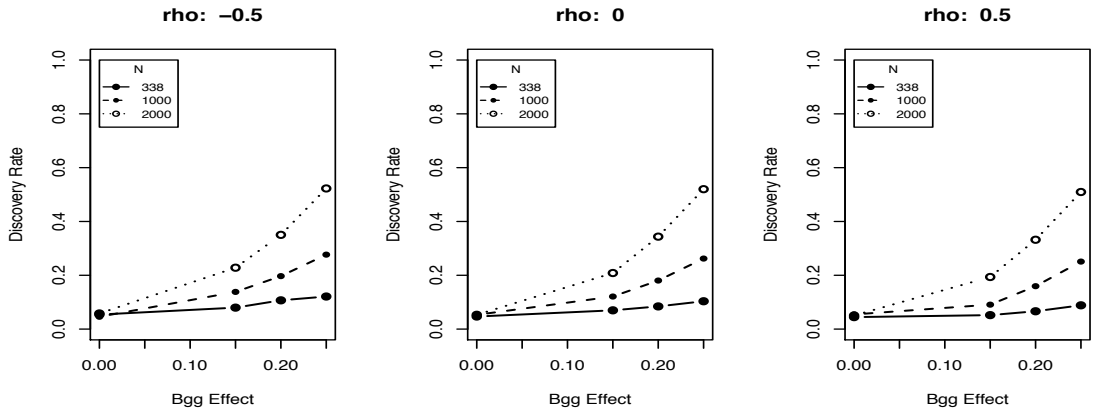

Figure 6.2: Discovery rate as a function of sample size n, gene-on-gene effect $\beta_{g_1g_2}$ and correlation between u_1 and u_2 (ρ, rho). The discovery rates in the $n = 2,000$ simulations are robust to changes in correlation between u_1 and u_2. Simulations with smaller sample sizes have a decrease in discovery rate with increasing correlation.

genotype values from a standard normal distribution for simulations under both H_0 and H_1. We run our simulation set up with $\beta_{g_1g_2} = 0$ in order to estimate the FWER. Table 6.1 shows that for ρ values ranging from -0.9 to 0.9, our method is well calibrated under the null regardless of sample sizes. According to our results, permutations that break the correlations in the confounding factors do not have an effect on the FWER.

As expected, simulations demonstrate that the power of our method increases as the sample size n and the magnitude of $\beta_{g_1g_2}$ increase. This is observed regardless of the correlation between u_1 and u_2 (see Figure 6.2 and Table 6.2). While the discovery rate for a sample size of 2,000 and $\beta_{g_1g_2} = 0.2$ remains near constant around 0.52, the discovery rate for the smaller samples sizes of n=1,000 and n=338, is largely influenced by ρ. We expect this, because a large and negative correlation reduces noise in the estimate of $\beta_{g_1g_2}$ relative to when the correlation is large and positive. For larger sample sizes, the effect of confounding is less pronounced since the sample size makes the estimate more accurate. Unfortunately, the largest GTEx sample size is $n = 338$, which means power will be quite low to observe anything but extremely large $\beta_{g_1g_2}$ effects in the real data.

n	-0.9	-0.5	-0.2	0.0	0.2	0.5	0.9
338	0.050	0.054	0.052	0.050	0.051	0.053	0.048
1000	0.051	0.050	0.048	0.050	0.051	0.051	0.049
2000	0.054	0.054	0.051	0.051	0.052	0.051	0.048

Table 6.1: Family-wise error rate as a function of sample size and correlation of u_1 and u_2 ranging from -0.9 to 0.9 when simulating under H_0.

n	-0.9	-0.5	-0.2	0.0	0.2	0.5	0.9
338	0.151	0.123	0.113	0.106	0.093	0.081	0.065
1000	0.304	0.279	0.273	0.264	0.256	0.246	0.232
2000	0.521	0.520	0.526	0.520	0.518	0.516	0.520

Table 6.2: Power to detect a $\beta_{g_1 g_2} = 0.2$ effect as a function of sample size and correlation (ρ) between u_1 and u_2.

6.3.2 GTEx data

We apply our method to four tissues from the GTEx project that have the largest number of individuals with both expression data and genotype data. We summarize effects of filtering and the top result from each tissue in Tables 6.3 and 6.4. When controlling the FDR at $\alpha = 0.05$ using the Benjamini-Hochberg procedure, the Artery-Tibial data contains a significant gene-on-gene effect for with the genes ENSG00000130300.4 and ENSG00000184497.8. With only 5 gene pairs passing the filtering requirements, the p-value for the gene pair (0.005) would still be significant when controlling FDR at $\alpha = 0.02$. This gene pair is also reported in a co-expression network in BRCA tumors as one of the 20 most highly correlated gene pairs[92].

The Whole Blood analysis had the second largest sample size. 49 gene pairs passed the filtering criteria. This tissue contains the most marginally significant $\beta_{g_1 g_2}$ estimate (p-value = 0.003) for the genes ENSG00000075303.8 and ENSG00000173890.12, however this gene pair would only be significant if controlling the FDR at $\alpha = 0.14$. In the Muscle-Skeletal

Tissue	Samples	eGenes	$\|cis$-effect$\|$ >0.2	$\|\rho\| > 0.8$
Artery-Tibial	285	8056	4523	5
Muscle-Skeletal	361	7082	3572	5
Whole Blood	338	6784	3182	49
Adipose-Subcontaneous	298	8500	4908	3

Table 6.3: Effect of filtering criteria on genes in the four GTEx tissues. eGenes represents the total number of eGenes identified in the tissue. $|cis$-effect$|$ >0.2 is the number of eGenes remaining after filtering out eGenes with small *cis*-effects. $|\rho| > 0.8$ represents the number of eGene pairs evaluated after the *cis*-effect filtering and requiring that eGenes been on different chromosomes and have an absolute correlation of at least 0.8.

tissue, the ENSG00000128928.4 and ENSG00000114054.9 gene pair as a marginal p-value of 0.012 but would only pass FDR filtering if controlling at an $\alpha = 0.06$. There were no marginally significant gene pairs observed in the Adipose-Subcutaneous tissue.

In order to see if the gene-on-gene effects are maintained across tissues, we take the four gene pairs reported in Table 6.4 and examine the observed effect in the tissues where it was not found (see Table 6.5). We do not observe any significant effect sizes. When looking across tissues, *cis*-effects can significantly differ across tissues and this results in division by very small numbers when estimating $E(r_2/r_1) = \beta_{g_1g_2}\beta_{s_1g_1}/\beta_{s_1g_1} = \beta_{g_1g_2}$. Finally, due to the complex regulatory and epigenetic changes that occur across tissues[41, 6], it is very possible that a gene-on-gene effect in one tissue will not exist in another.

6.4 Discussion

In this work we present a method to use SNPs as instrumental variables for identification of causal relationships between the expression levels of genes. By drawing on concepts from Mendelian randomization we are able to estimate both the effect size and the direction of these gene-on-gene effects. Gene-gene correlation networks have been used extensively, but are limited because identified correlations may be either causal relationships in the network or due to confounders. Our new method will allows estimation of effect direction and magnitude

Tissue	Gene Pair	P-value	FDR threshold	$\lvert\beta_{g_1g_2}\rvert$
Artery-Tibial	ENSG00000130300.4 ENSG00000184497.8	0.005	0.02	0.75
Muscle-Skeletal	ENSG00000128928.4 ENSG00000114054.9	0.012	0.06	0.54
Whole Blood	ENSG00000075303.8 ENSG00000173890.12	0.003	0.14	0.77
Adipose-Subcontaneous	ENSG00000170889.9 ENSG00000149273.10	0.086	0.259	0.44

Table 6.4: Top results form GTEx analysis in four tissues. the Gene Pair column is the pair of genes with the most significant gene-on-gene effect, P-value is the p-value for this gene pair. FDR threshold can be interpreted as the smallest FDR that can be controlled for such that the gene-on-gene effect passes FDR control.

and this forms a valuable additional component for gene network analysis studies.

While our method is robust to most forms of confounding, a false positive will occur if the confounder is correlated to the SNP genotype that is used as the instrumental variable. Such a false positive would likely still involve a gene-on-gene effect as the means for the SNP to affect the confounder, but the true model would not match either H_1 or H_2 in Figure 6.1. It is also possible that a SNP coud be a *cis* regulator for genes g_1 and g_2 and that g_2 could have a *tran* effect on a third gene g_3. This would result in a false detection of a gene-on-gene effect of g_1 on g_3, but the effect of g_2 on g_3 would still be a true gene-on-gene effect if it were detected.

The main limitation of applying our method to the GTEx data is the small sample sizes. Especially after controlling the FDR, due to sample sizes, our method is limited to finding genes with the largest effect size. This can be seen with the large effect size estimates for the top gene pairs in the GTEx data. Future work could increase the power of analysis by leveraging cross-tissue gene-on-gene effect estimates.

Gene Pair	Artery-Tibial	Muscle-Skeletal	Whole Blood	Adipose-Subcontaneous
ENSG00000130300.4 ENSG00000184497.8	0.75*	49.92	0.16	0.66
ENSG00000128928.4 ENSG00000114054.9	0.33	0.54*	.27	-0.29
ENSG00000075303.8 ENSG00000173890.12	-0.55	-2.47	0.77*	-5.98
ENSG00000170889.9 ENSG00000149273.10	-1.12	-0.13	0.40	0.44*

Table 6.5: The estimated gene-on-gene effects for the top gene-pairs from each tissue, estimated in each tissue. The * indicates the tissue where the gene pair had the top gene-on-gene effect (see Table 6.4). None of the estimated gene-on-gene effects in other tissues were statistically significant based on permutations.

References

[1] Francois Aguet, Andrew A Brown, Stephane Castel, Joe R Davis, Pejman Mohammadi, Ayellet V Segre, Zachary Zappala, Nathan S Abell, Laure Fresard, Eric R Gamazon, Ellen Gelfand, Machael J Gloudemans, Yuan He, Farhad Hormozdiari, Xiao Li, Xin Li, Boxiang Liu, Diego Garrido-Martin, Halit Ongen, John J Palowitch, YoSon Park, Christine B Peterson, Gerald Quon, Stephan Ripke, Andrey A Shabalin, Tyler C Shimko, Benjamin J Strober, Timothy J Sullivan, Nicole A Teran, Emily K Tsang, Hailei Zhang, Yi-Hui Zhou, Alexis Battle, Carlos D Bustamonte, Nancy J Cox, Barbara E Engelhardt, Eleazar Eskin, Gad Getz, Manolis Kellis, Gen Li, Daniel G MacArthur, Andrew B Nobel, Chiara Sabbati, Xiaoquan Wen, Fred A Wright, GTEx Consortium, Tuuli Lappalainen, Kristin G Ardlie, Emmanouil T Dermitzakis, Christopher D Brown, and Stephen B Montgomery. Local genetic effects on gene expression across 44 human tissues. Technical report, September 2016.

[2] Heleen H Arts, Ernie MHF Bongers, Dorus A Mans, Sylvia EC van Beersum, Machteld M Oud, Emine Bolat, Liesbeth Spruijt, Elisabeth AM Cornelissen, Janneke HM Schuurs-Hoeijmakers, Nicole De Leeuw, et al. C14orf179 encoding ift43 is mutated in sensenbrenner syndrome. *Journal of medical genetics*, 48(6):390–395, 2011.

[3] Michael J Bamshad, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah A Nickerson, and Jay Shendure. Exome sequencing as a tool for mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755, 2011.

[4] Yael Baran, Bogdan Pasaniuc, Sriram Sankararaman, Dara G Torgerson, Christopher Gignoux, Celeste Eng, William Rodriguez-Cintron, Rocio Chapela, Jean G Ford, Pedro C Avila, et al. Fast and accurate inference of local ancestry in latino populations. *Bioinformatics*, 28(10):1359–1367, 2012.

[5] F Bergametti, C Denier, P Labauge, M Arnoult, S Boetto, M Clanet, P Coubes, B Echenne, R Ibrahim, B Irthum, et al. Mutations within the programmed cell death 10 gene cause cerebral cavernous malformations. *The American Journal of Human Genetics*, 76(1):42–51, 2005.

[6] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, et al. The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, 28(10):1045–1048, 2010.

[7] Gaurav Bhatia, Nick Patterson, Sriram Sankararaman, and Alkes L Price. Estimating and interpreting fst: the impact of rare variants. *Genome research*, 23(9):1514–1521, 2013.

[8] Leslie G Biesecker, Wylie Burke, Isaac Kohane, Sharon E Plon, and Ron Zimmern. Next-generation sequencing in the clinic: are we ready? *Nature Reviews Genetics*, 13(11):818–824, 2012.

[9] Kaya Bilgüvar, Ali Kemal Öztürk, Angeliki Louvi, Kenneth Y Kwan, Murim Choi, Burak Tatlı, Dilek Yalnızoğlu, Beyhan Tüysüz, Ahmet Okay Çağlayan, Sarenur Gökben, et al. Whole-exome sequencing identifies recessive wdr62 mutations in severe brain malformations. *Nature*, 467(7312):207–210, 2010.

[10] Adam R Boyko, Scott H Williamson, Amit R Indap, Jeremiah D Degenhardt, Ryan D Hernandez, Kirk E Lohmueller, Mark D Adams, Steffen Schmidt, John J Sninsky, Shamil R Sunyaev, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*, 4(5):e1000083, 2008.

[11] William S. Branham, Cathy D. Melvin, Tao Han, Varsha G. Desai, Carrie L. Moland, Adam T. Scully, and James C. Fuscoe. Elimination of laboratory ozone leads to a dramatic improvement in the reproducibility of microarray gene expression measurements. *BMC Biotechnol*, 7:8, 2007.

[12] Rachel B Brem, John D Storey, Jacqueline Whittle, and Leonid Kruglyak. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*, 436(7051):701–703, August 2005.

[13] Abra Brisbin, Katarzyna Bryc, Jake Byrnes, Fouad Zakharia, Larsson Omberg, Jeremiah Degenhardt, Andrew Reynolds, Harry Ostrer, Jason G Mezey, and Carlos D Bustamante. Pcadmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human biology*, 84(4):343–364, 2012.

[14] Andrew Anand Brown, Alfonso Buil, Ana Viñuela, Tuuli Lappalainen, Hou-Feng Zheng, J Brent Richards, Kerrin S Small, Timothy D Spector, Emmanouil T Dermitzakis, and Richard Durbin. Genetic interactions affecting human gene expression identified by variance association mapping. *eLife*, 3:e01381, April 2014.

[15] Christopher D Brown, Lara M Mangravite, and Barbara E Engelhardt. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS genetics*, 9(8):e1003649, 2013.

[16] Robert Brown, Gleb Kichaev, Nicholas Mancuso, James Boocock, and Bogdan Pasaniuc. Enhanced methods to detect haplotypic effects on gene expression. *Bioinformatics (Oxford, England)*, 2017.

[17] Robert Brown, Hane Lee, Ascia Eskin, Gleb Kichaev, Kirk E Lohmueller, Bruno Reversade, Stanley F Nelson, and Bogdan Pasaniuc. Leveraging ancestry to improve causal variant identification in exome sequencing for monogenic disorders. *European Journal of Human Genetics*, 24(1):113–119, 2016.

[18] Robert Brown and Bogdan Pasaniuc. Enhanced methods for local ancestry assignment in sequenced admixed individuals. *PLoS computational biology*, 10(4):e1003555, April 2014.

[19] Brian L Browning and Sharon R Browning. Genotype Imputation with Millions of Reference Samples. *American journal of human genetics*, 98(1):116–126, January 2016.

[20] Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 2007.

[21] Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics*, 81(5):1084–1097, November 2007.

[22] Katarzyna Bryc, Adam Auton, Matthew R Nelson, Jorge R Oksenberg, Stephen L Hauser, Scott Williams, Alain Froment, Jean-Marie Bodo, Charles Wambebe, Sarah A Tishkoff, et al. Genome-wide patterns of population structure and admixture in west africans and african americans. *Proceedings of the National Academy of Sciences*, 107(2):786–791, 2010.

[23] Alfonso Buil, Andrew Anand Brown, Tuuli Lappalainen, Ana Viñuela, Matthew N Davies, Hou-Feng Zheng, J Brent Richards, Daniel Glass, Kerrin S Small, Richard Durbin, Timothy D Spector, and Emmanouil T Dermitzakis. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nature genetics*, 47(1):88–91, January 2015.

[24] EG Burchard, LN Borrell, S Choudhry, M Naqvi, H Tsai, JR Rodriguez-Santana, et al. Race, genetics, and health disparities. latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research. *American Journal of Public Health*, 95(12):2161–2168, 2005.

[25] Stephen Burgess, Nicholas J Timpson, Shah Ebrahim, and George Davey Smith. Mendelian randomization: where are we now and where are we going?, 2015.

[26] Paul R Burton, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[27] Carlos D Bustamante, M Francisco, and Esteban G Burchard. Genomics for the world. *Nature*, 475(7355):163–165, 2011.

[28] Ferran Casals, Alan Hodgkinson, Julie Hussin, Youssef Idaghdour, Vanessa Bruat, Thibault de Maillard, Jean-Cristophe Grenier, Elias Gbeha, Fadi F Hamdan, Simon Girard, et al. Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet*, 9(9):e1003815, 2013.

[29] Lin S. Chen, Frank Emmert-Streib, and John D. Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol*, 8(10):R219, 2007.

[30] Ching-Yu Cheng, WH Linda Kao, Nick Patterson, Arti Tandon, Christopher A Haiman, Tamara B Harris, Chao Xing, Esther M John, Christine B Ambrosone, Frederick L Brancati, et al. Admixture mapping of 15,280 african americans identifies obesity susceptibility loci on chromosomes 5 and x. *PLoS Genet*, 5(5):e1000490, 2009.

[31] Ching-Yu Cheng, David Reich, Christopher A Haiman, Arti Tandon, Nick Patterson, Selvin Elizabeth, Ermeg L Akylbekova, Frederick L Brancati, Josef Coresh, Eric Boerwinkle, et al. African ancestry and its correlation to type 2 diabetes in african americans: a genetic admixture analysis in three us population cohorts. *PLoS One*, 7(3):e32840, 2012.

[32] KH Cheung, PERRY L Miller, Judith R Kidd, Kenneth K Kidd, Michael V Osier, and ANDREW J Pakstis. Alfred: a web-accessible allele frequency database. In *Pac Symp Biocomput*, volume 2000, pages 639–50, 2000.

[33] Vivian G Cheung, Laura K Conlin, Teresa M Weber, Melissa Arcaro, Kuang-Yu Jen, Michael Morley, and Richard S Spielman. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature genetics*, 33(3):422–425, March 2003.

[34] Vivian G Cheung, Richard S Spielman, Kathryn G Ewens, Teresa M Weber, Michael Morley, and Joshua T Burdick. Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, 437(7063):1365–1369, October 2005.

[35] Murim Choi, Ute I Scholl, Weizhen Ji, Tiewen Liu, Irina R Tikhonova, Paul Zumbo, Ahmet Nayir, Ayin Bakkaloğlu, Seza Özen, Sami Sanjad, et al. Genetic diagnosis by whole exome capture and massively parallel dna sequencing. *Proceedings of the National Academy of Sciences*, 106(45):19096–19101, 2009.

[36] Claire Churchhouse and Jonathan Marchini. Multiway admixture deconvolution using phased or unphased ancestral panels. *Genetic epidemiology*, 37(1):1–12, 2013.

[37] Gary A. Churchill. Fundamentals of experimental design for cdna microarrays. *Nat Genet*, 32 Suppl:490–5, 12 2002.

[38] Elizabeth T Cirulli and David B Goldstein. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415–425, 2010.

[39] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[40] GTEx Consortium. The genotype-tissue expression (gtex) project. *Nat Genet*, 45(6):580–5, 6 2013.

[41] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[42] William Cookson, Liming Liang, Gonçalo Abecasis, Miriam Moffatt, and Mark Lathrop. Mapping complex disease traits with global gene expression. *Nature reviews. Genetics*, 10(3):184–194, March 2009.

[43] Gregory M Cooper and Jay Shendure. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, 12(9):628–640, 2011.

[44] Heather J Cordell. Detecting gene-gene interactions that underlie human diseases. *Nature reviews. Genetics*, 10(6):392–404, June 2009.

[45] H Darvish, S Esmaeeli-Nieh, GB Monajemi, M Mohseni, S Ghasemi-Firouzabadi, SS Abedini, I Bahman, P Jamali, S Azimi, F Mojahedi, et al. A clinical and molecular genetic study of 112 iranian families with primary microcephaly. *Journal of medical genetics*, 47(12):823–828, 2010.

[46] Antigone S Dimas, Barbara E Stranger, Claude Beazley, Robert D Finn, Catherine E Ingle, Matthew S Forrest, Matthew E Ritchie, Panos Deloukas, Simon Tavaré, and Emmanouil T Dermitzakis. Modifier effects between regulatory and protein-coding variation. *PLoS genetics*, 4(10):e1000244, October 2008.

[47] ENCODE Project Consortium, Ewan Birney, John A Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R Gingeras, Elliott H Margulies, Zhiping Weng, Michael Snyder, Emmanouil T Dermitzakis, Robert E Thurman, Michael S Kuehn, Christopher M Taylor, Shane Neph, Christoph M Koch, Saurabh Asthana, Ankit Malhotra, Ivan Adzhubei, Jason A Greenbaum, Robert M Andrews, Paul Flicek, Patrick J Boyle, Hua Cao, Nigel P Carter, Gayle K Clelland, Sean Davis, Nathan Day, Pawandeep Dhami, Shane C Dillon, Michael O Dorschner, Heike Fiegler, Paul G Giresi, Jeff Goldy, Michael Hawrylycz, Andrew Haydock, Richard Humbert, Keith D James, Brett E Johnson, Ericka M Johnson, Tristan T Frum, Elizabeth R Rosenzweig, Neerja Karnani, Kirsten Lee, Gregory C Lefebvre, Patrick A Navas, Fidencio Neri, Stephen C J Parker, Peter J Sabo, Richard Sandstrom, Anthony Shafer, David Vetrie, Molly Weaver, Sarah Wilcox, Man Yu, Francis S Collins, Job Dekker, Jason D Lieb, Thomas D Tullius, Gregory E Crawford, Shamil Sunyaev, William S Noble, Ian Dunham, France Denoeud, Alexandre Reymond, Philipp Kapranov, Joel Rozowsky, Deyou Zheng, Robert Castelo, Adam Frankish, Jennifer Harrow, Srinka Ghosh, Albin Sandelin, Ivo L Hofacker, Robert Baertsch, Damian Keefe, Sujit Dike, Jill Cheng, Heather A Hirsch, Edward A Sekinger, Julien Lagarde, Josep F Abril, Christoph Flamm, Claudia Fried, Jörg Hackermüller, Jana Hertel, Manja Lindemeyer, Kristin Missal, Andrea Tanzer, Stefan Washietl, Jan Korbel, Olof Emanuelsson, Jakob S Pedersen, Nancy Holroyd, Ruth Taylor, David Swarbreck, Nicholas Matthews, Mark C Dickson, Daryl J Thomas, Matthew T Weirauch, James Gilbert, Jorg Drenkow, Ian Bell, XiaoDong Zhao, K G Srinivasan, Wing-Kin Sung, Hong Sain Ooi, Kuo Ping Chiu, Sylvain Foissac, Tyler Alioto, Michael Brent, Lior Pachter, Michael L Tress, Alfonso Valencia, Siew Woh Choo, Chiou Yu Choo, Catherine Ucla, Caroline Manzano, Carine Wyss, Evelyn Cheung, Taane G Clark, James B Brown, Madhavan Ganesh, Sandeep Patel, Hari Tammana, Jacqueline Chrast, Charlotte N Henrichsen, Chikatoshi Kai, Jun Kawai, Ugrappa Nagalakshmi, Jiaqian Wu, Zheng Lian, Jin Lian, Peter Newburger, Xueqing Zhang, Peter Bickel, John S Mattick, Piero Carninci, Yoshihide Hayashizaki, Sherman Weissman, Tim Hubbard, Richard M Myers, Jane

Rogers, Peter F Stadler, Todd M Lowe, Chia-Lin Wei, Kevin Struhl, Mark Gerstein, Stylianos E Antonarakis, Yutao Fu, Eric D Green, Ulaş Karaöz, Adam Siepel, James Taylor, Laura A Liefer, Kris A Wetterstrand, Peter J Good, Elise A Feingold, Mark S Guyer, Gregory M Cooper, George Asimenos, Colin N Dewey, Minmei Hou, Sergey Nikolaev, Juan I Montoya-Burgos, Ari Löytynoja, Simon Whelan, Fabio Pardi, Tim Massingham, Haiyan Huang, Nancy R Zhang, Ian Holmes, James C Mullikin, Abel Ureta-Vidal, Benedict Paten, Michael Seringhaus, Deanna Church, Kate Rosenbloom, W James Kent, Eric A Stone, NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Serafim Batzoglou, Nick Goldman, Ross C Hardison, David Haussler, Webb Miller, Arend Sidow, Nathan D Trinklein, Zhengdong D Zhang, Leah Barrera, Rhona Stuart, David C King, Adam Ameur, Stefan Enroth, Mark C Bieda, Jonghwan Kim, Akshay A Bhinge, Nan Jiang, Jun Liu, Fei Yao, Vinsensius B Vega, Charlie W H Lee, Patrick Ng, Atif Shahab, Annie Yang, Zarmik Moqtaderi, Zhou Zhu, Xiaoqin Xu, Sharon Squazzo, Matthew J Oberley, David Inman, Michael A Singer, Todd A Richmond, Kyle J Munn, Alvaro Rada-Iglesias, Ola Wallerman, Jan Komorowski, Joanna C Fowler, Phillippe Couttet, Alexander W Bruce, Oliver M Dovey, Peter D Ellis, Cordelia F Langford, David A Nix, Ghia Euskirchen, Stephen Hartman, Alexander E Urban, Peter Kraus, and Va... Identification and analysis of functional elements in 1human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, June 2007.

[48] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis, and Bradley E Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, May 2011.

[49] Thomas L. Fare, Ernest M. Coffey, Hongyue Dai, Yudong D. He, Deborah A. Kessler, Kristopher A. Kilian, John E. Koch, Eric LeProust, Matthew J. Marton, Michael R. Meyer, Roland B. Stoughton, George Y. Tokiwa, and Yanqun Wang. Effects of atmospheric ozone on microarray data quality. *Anal Chem*, 75(17):4672–5, 9 2003.

[50] Laura L Faye, Mitchell J Machiela, Peter Kraft, Shelley B Bull, and Lei Sun. Reranking sequencing variants in the post-GWAS era for accurate causal variant identification. *PLoS genetics*, 9(8):e1003609, 2013.

[51] Laura Fejerman, Gary K Chen, Celeste Eng, Scott Huntsman, Donglei Hu, Amy Williams, Bogdan Pasaniuc, Esther M John, Marc Via, Christopher Gignoux, et al. Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in us latinas. *Human molecular genetics*, 21(8):1907–1917, 2012.

[52] Alexandra E Fish, John A Capra, and William S Bush. Are Interactions between cis-Regulatory Variants Evidence for Biological Epistasis or Statistical Artifacts? *American journal of human genetics*, September 2016.

[53] Wenqing Fu, Timothy D OConnor, Goo Jun, Hyun Min Kang, Goncalo Abecasis, Suzanne M Leal, Stacey Gabriel, Mark J Rieder, David Altshuler, Jay Shendure, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220, 2013.

[54] Daniel J Gaffney, Jean-Baptiste Veyrieras, Jacob F Degner, Roger Pique-Regi, Athma A Pai, Gregory E Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Dissecting the regulatory architecture of gene expression QTLs. *Genome biology*, 13(1):R7, January 2012.

[55] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, Nancy J Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091–1098, September 2015.

[56] Sophie Garnier, Vinh Truong, Jessy Brocheton, Tanja Zeller, Maxime Rovital, Philipp S Wild, Andreas Ziegler, Cardiogenics Consortium, Thomas Münzel, Laurence Tiret, Stefan Blankenberg, Panos Deloukas, Jeannette Erdmann, Christian Hengstenberg, Nilesh J Samani, Heribert Schunkert, Willem H Ouwehand, Alison H Goodall, François Cambien, and David-Alexandre Trégouët. Genome-wide haplotype analysis of cis expression quantitative trait loci in monocytes. *PLoS genetics*, 9(1):e1003240, 2013.

[57] Giulio Genovese, Robert E Handsaker, Heng Li, Nicolas Altemose, Amelia M Lindgren, Kimberly Chambert, Bogdan Pasaniuc, Alkes L Price, David Reich, Cynthia C Morton, et al. Using population admixture to help complete maps of the human genome. *Nature genetics*, 45(4):406–414, 2013.

[58] Greg Gibson. Rare and common variants: twenty arguments. *Nature reviews. Genetics*, 13(2):135–145, February 2011.

[59] Christian Gilissen, Alexander Hoischen, Han G Brunner, and Joris A Veltman. Unlocking Mendelian disease using exome sequencing. *Genome biology*, 12(9):228, 2011.

[60] Christian Gilissen, Alexander Hoischen, Han G Brunner, and Joris A Veltman. Disease gene identification strategies for exome sequencing. *European journal of human genetics : EJHG*, 20(5):490–497, May 2012.

[61] Gustavo Glusman, Juan Caballero, Denise E Mauldin, Leroy Hood, and Jared C Roach. Kaviar: an accessible system for testing snv novelty. *Bioinformatics*, 27(22):3216–3217, 2011.

[62] David B Goldstein, Andrew Allen, Jonathan Keebler, Elliott H Margulies, Steven Petrou, Slavé Petrovski, and Shamil Sunyaev. Sequencing studies in human genetics: design and interpretation. *Nature Reviews Genetics*, 14(7):460–470, 2013.

[63] Abel González-Pérez and Nuria López-Bigas. Improving the assessment of the outcome of nonsynonymous snvs with a consensus deleteriousness score, condel. *The American Journal of Human Genetics*, 88(4):440–449, 2011.

[64] Simon Gravel. Population genetics models of local ancestry. *Genetics*, 191(2):607–619, 2012.

[65] Simon Gravel, Brenna M Henn, Ryan N Gutenkunst, Amit R Indap, Gabor T Marth, Andrew G Clark, Fuli Yu, Richard A Gibbs, Carlos D Bustamante, David L Altshuler, et al. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, 2011.

[66] Fabian Grubert, Judith B Zaugg, Maya Kasowski, Oana Ursu, Damek V Spacek, Alicia R Martin, Peyton Greenside, Rohith Srivas, Doug H Phanstiel, Aleksandra Pekowska, Nastaran Heidari, Ghia Euskirchen, Wolfgang Huber, Jonathan K Pritchard, Carlos D Bustamante, Lars M Steinmetz, Anshul Kundaje, and Michael Snyder. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, 162(5):1051–1065, August 2015.

[67] Elin Grundberg, Kerrin S. Small, Åsa K. Hedman, Alexandra C. Nica, Alfonso Buil, Sarah Keildson, Jordana T. Bell, Tsun-Po P. Yang, Eshwar Meduri, Amy Barrett, James Nisbett, Magdalena Sekowska, Alicja Wilk, So-Youn Y. Shin, Daniel Glass, Mary Travers, Josine L. Min, Sue Ring, Karen Ho, Gudmar Thorleifsson, Augustine Kong, Unnur Thorsteindottir, Chrysanthi Ainali, Antigone S. Dimas, Neelam Hassanali, Catherine Ingle, David Knowles, Maria Krestyaninova, Christopher E. Lowe, Paola Di Meglio, Stephen B. Montgomery, Leopold Parts, Simon Potter, Gabriela Surdulescu, Loukia Tsaprouni, Sophia Tsoka, Veronique Bataille, Richard Durbin, Frank O. Nestle, Stephen O'Rahilly, Nicole Soranzo, Cecilia M. Lindgren, Krina T. Zondervan, Kourosh R. Ahmadi, Eric E. Schadt, Kari Stefansson, George Davey Smith, Mark I. McCarthy, Panos Deloukas, Emmanouil T. Dermitzakis, Tim D. Spector, and Multiple Tissue Human Expression Resource (MuTHER) Consortium. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet*, 44(10):1084–9, 10 2012.

[68] GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, May 2015.

[69] Matthew G Guenther, Stuart S Levine, Laurie A Boyer, Rudolf Jaenisch, and Richard A Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88, July 2007.

[70] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W J H Penninx, Rick Jansen, Eco J C de Geus, Dorret I Boomsma, Fred A Wright, Patrick F Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J Lusis, Terho Lehtimäki, Emma Raitoharju, Mika Kähönen, Ilkka Seppälä, Olli T Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L Price, Päivi Pajukanta, and Bogdan Pasaniuc.

Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–252, March 2016.

[71] Buhm Han, Hyun Min Kang, and Eleazar Eskin. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS genetics*, 5(4):e1000456, April 2009.

[72] R David Hawkins, Gary C Hon, and Bing Ren. Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, 11(7):476–486, 2010.

[73] Hannes Helgason, Patrick Sulem, Maheswara R Duvvari, Hongrong Luo, Gudmar Thorleifsson, Hreinn Stefansson, Ingileif Jonsdottir, Gisli Masson, Daniel F Gudbjartsson, G Bragi Walters, et al. A rare nonsynonymous sequence variant in c3 is associated with high risk of age-related macular degeneration. *Nature genetics*, 45(11):1371–1374, 2013.

[74] Gibran Hemani, Konstantin Shakhbazov, Harm-Jan Westra, Tõnu Esko, Anjali K Henders, Allan F McRae, Jian Yang, Greg Gibson, Nicholas G Martin, Andres Metspalu, Lude Franke, Grant W Montgomery, Peter M Visscher, and Joseph E Powell. Detection and replication of epistasis influencing transcription in humans. *Nature*, 508(7495):249–253, April 2014.

[75] Gibran Hemani, Konstantin Shakhbazov, Harm-Jan Westra, Tõnu Esko, Anjali K Henders, Allan F McRae, Jian Yang, Greg Gibson, Nicholas G Martin, Andres Metspalu, Lude Franke, Grant W Montgomery, Peter M Visscher, and Joseph E Powell. Hemani et al. reply. *Nature*, 514(7520):E5–6, October 2014.

[76] W G Hill and A Robertson. Linkage disequilibrium in finite populations. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 38(6):226–231, June 1968.

[77] Anjali G Hinch, Arti Tandon, Nick Patterson, Yunli Song, Nadin Rohland, Cameron D Palmer, Gary K Chen, Kai Wang, Sarah G Buxbaum, Ermeg L Akylbekova, et al. The landscape of recombination in african americans. *Nature*, 476(7359):170–175, 2011.

[78] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

[79] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, October 2014.

[80] Mattias Jakobsson, Michael D Edge, and Noah A Rosenberg. The relationship between fst and the frequency of the most frequent allele. *Genetics*, 193(2):515–528, 2013.

[81] Asif Javed, Saloni Agrawal, and Pauline C Ng. Phen-gen: combining phenotype and genotype to analyze rare disorders. *Nature methods*, 11(9):935–937, 2014.

[82] Wenfei Jin, Shuhua Xu, Haifeng Wang, Yongguo Yu, Yiping Shen, Bailin Wu, and Li Jin. Genome-wide detection of natural selection in african americans pre-and post-admixture. *Genome research*, 22(3):519–527, 2012.

[83] Nicholas A Johnson, Marc A Coram, Mark D Shriver, Isabelle Romieu, Gregory S Barsh, Stephanie J London, and Hua Tang. Ancestral components of admixed genomes in a mexican cohort. *PLoS Genet*, 7(12):e1002410, 2011.

[84] Jong Wha J. Joo, Jae Hoon Sul, Buhm Han, Chun Ye, and Eleazar Eskin. Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies. *Genome Biol*, 15(4):r61, 2014.

[85] Eun Yong Kang, Chun Ye, Ilya Shpitser, and Eleazar Eskin. Detecting the presence and absence of causal relationships between expression of yeast genes with very few samples. *J Comput Biol*, 17(3):533–46, 3 2010.

[86] Guolian Kang, Guimin Gao, Sanjay Shete, David T Redden, Bao-Li Chang, Timothy R Rebbeck, Jill S Barnholtz-Sloan, Nicholas M Pajewski, and David B Allison. Capitalizing on admixture in genome-wide association studies: a two-stage testing procedure and application to height in african-americans. *Frontiers in genetics*, 2, 2011.

[87] Hyun Min Kang, Chun Ye, and Eleazar Eskin. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4):1909–25, 12 2008.

[88] M Kharrat, S Trabelsi, M Chaabouni, F Maazoul, L Kraoua, L Ben Jemaa, N Gandoura, S Barsaoui, Y Morel, R Mrad, et al. Only two mutations detected in 15 tunisian patients with 11β-hydroxylase deficiency: the p. q356x and the novel p. g379v. *Clinical genetics*, 78(4):398–401, 2010.

[89] Gleb Kichaev, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L Price, Peter Kraft, and Bogdan Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS genetics*, 10(10):e1004722, October 2014.

[90] Jeffrey M Kidd, Simon Gravel, Jake Byrnes, Andres Moreno-Estrada, Shaila Musharoff, Katarzyna Bryc, Jeremiah D Degenhardt, Abra Brisbin, Vrunda Sheth, Rong Chen, et al. Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. *The American Journal of Human Genetics*, 91(4):660–671, 2012.

[91] Helena Kilpinen, Sebastian M Waszak, Andreas R Gschwind, Sunil K Raghav, Robert M Witwicki, Andrea Orioli, Eugenia Migliavacca, Michaël Wiederkehr, Maria Gutierrez-Arcelus, Nikolaos I Panousis, et al. Coordinated effects of sequence variation on dna binding, chromatin structure, and transcription. *Science*, 342(6159):744–747, 2013.

[92] Pora Kim, Feixiong Cheng, Junfei Zhao, and Zhongming Zhao. ccmgdb: a database for cancer cell metabolism genes. *Nucleic acids research*, 44(D1):D959–D968, 2016.

[93] M Kimura. The neutral theory of molecular evolution. 1983, cambridge [cambridgeshire].

[94] Motoo Kimura. The neutral theory of molecular evolution. *Scientific American*, 241:98–126, 1979.

[95] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–315, 2014.

[96] Linda Koch. Genomics: Adding another dimension to gene regulation. *Nature reviews. Genetics*, 16(10):563–563, October 2015.

[97] Emrah Kostem, Jose A Lozano, and Eleazar Eskin. Increasing power of genome-wide association studies by collecting additional single-nucleotide polymorphisms. *Genetics*, 188(2):449–460, June 2011.

[98] Anat Kreimer and Itsik Pe'er. Co-regulated transcripts associated to cooperating esnps define bi-fan motifs in human gene networks. *PLoS Genet*, 10(9):e1004587, 9 2014.

[99] Chee-Seng Ku, Nasheen Naidoo, and Yudi Pawitan. Revisiting mendelian disorders through exome sequencing. *Human genetics*, 129(4):351–370, 2011.

[100] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, 42(D1):D980–D985, 2014.

[101] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter A C 't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G MacArthur, Monkol Lek, Esther Lizano, Henk P J Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B Montgomery, Peter Donnelly, Mark I McCarthy, Paul Flicek, Tim M Strom, Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G Gut, Xavier Estivill, and Emmanouil T Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, September 2013.

[102] Nicholas B Larson, Shannon McDonnell, Amy J French, Zach Fogarty, John Cheville, Sumit Middha, Shaun Riska, Saurabh Baheti, Asha A Nair, Liang Wang, Daniel J Schaid, and Stephen N Thibodeau. Comprehensively Evaluating cis-Regulatory Variation in the Human Prostate Transcriptome by Using Gene-Level Allele-Specific Expression. *American journal of human genetics*, 96(6):869–882, 2015.

[103] Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

[104] Jeffrey T. Leek and John D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):1724–35, 9 2007.

[105] Katarina Lehmann, Petra Seemann, Sigmar Stricker, Marai Sammar, Birgit Meyer, Katrin Süring, Frank Majewski, Sigrid Tinschert, Karl-Heinz Grzeschik, Dietmar Müller, et al. Mutations in bone morphogenetic protein receptor 1b cause brachydactyly type a2. *Proceedings of the National Academy of Sciences*, 100(21):12277–12282, 2003.

[106] Juan Pablo Lewinger, John L Morrison, Duncan C Thomas, Cassandra E Murcray, David V Conti, Dalin Li, and W James Gauderman. Efficient two-step testing of gene-gene interactions in genome-wide association studies. *Genetic epidemiology*, 37(5):440–451, July 2013.

[107] Jun Z Li, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, Gregory S Barsh, Marcus Feldman, Luigi L Cavalli-Sforza, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *science*, 319(5866):1100–1104, 2008.

[108] Miao-Xin Li, Johnny SH Kwan, Su-Ying Bao, Wanling Yang, Shu-Leong Ho, Yong-Qiang Song, and Pak C Sham. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet*, 9(1):e1003143, 2013.

[109] Elaine T Lim, Soumya Raychaudhuri, Stephan J Sanders, Christine Stevens, Aniko Sabo, Daniel G MacArthur, Benjamin M Neale, Andrew Kirby, Douglas M Ruderfer, Menachem Fromer, Monkol Lek, Li Liu, Jason Flannick, Stephan Ripke, Uma Nagaswamy, Donna Muzny, Jeffrey G Reid, Alicia Hawes, Irene Newsham, Yuanqing Wu, Lora Lewis, Huyen Dinh, Shannon Gross, Li-San Wang, Chiao-Feng Lin, Otto Valladares, Stacey B Gabriel, Mark dePristo, David M Altshuler, Shaun M Purcell, NHLBI Exome Sequencing Project, Matthew W State, Eric Boerwinkle, Joseph D Buxbaum, Edwin H Cook, Richard A Gibbs, Gerard D Schellenberg, James S Sutcliffe, Bernie Devlin, Kathryn Roeder, and Mark J Daly. Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron*, 77(2):235–242, January 2013.

[110] Elaine T Lim, Peter Würtz, Aki S Havulinna, Priit Palta, Taru Tukiainen, Karola Rehnström, Tõnu Esko, Reedik Mägi, Michael Inouye, Tuuli Lappalainen, et al. Distribution and medical impact of loss-of-function variants in the finnish founder population. *PLoS Genet*, 10(7):e1004494, 2014.

[111] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. dbnsfp v2. 0: a database of human non-synonymous snvs and their functional predictions and annotations. *Human mutation*, 34(9):E2393–E2402, 2013.

[112] Margarida C Lopes, Chris Joyce, Graham RS Ritchie, Sally L John, Fiona Cunningham, Jennifer Asimit, and Eleftheria Zeggini. A combined functional annotation score for non-synonymous variants. *Human heredity*, 73(1):47–51, 2012.

[113] Estelle Lopez, Christel Thauvin-Robinet, Bruno Reversade, Nadia El Khartoufi, Louise Devisme, Muriel Holder, Hélène Ansart-Franquet, Magali Avila, Didier Lacombe, Pascale Kleinfinger, et al. C5orf42 is the major gene responsible for ofd syndrome type vi. *Human genetics*, 133(3):367–377, 2014.

[114] Daniel G MacArthur, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, Lukas Habegger, Joseph K Pickrell, Stephen B Montgomery, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, 2012.

[115] Daniel G MacArthur and Chris Tyler-Smith. Loss-of-function variants in the genomes of healthy humans. *Human molecular genetics*, 19(R2):R125–R130, 2010.

[116] DG MacArthur, TA Manolio, DP Dimmock, HL Rehm, J Shendure, GR Abecasis, DR Adams, RB Altman, SE Antonarakis, EA Ashley, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469–476, 2014.

[117] Ons Mamaï, Lobna Boussofara, Mohamed Denguezli, Nathalie Escande-Beillard, Wahiba Kraeim, Barry Merriman, Ilhem Ben Charfeddine, Giovanni Stevanin, Sana Bouraoui, Abdelbasset Amara, et al. Multiple self-healing palmoplantar carcinoma: a familial predisposition to skin cancer with primary palmoplantar and conjunctival lesions. *The Journal of investigative dermatology*, 135(1):304, 2015.

[118] Xianyun Mao, Abigail W Bigham, Rui Mei, Gerardo Gutierrez, Ken M Weiss, Tom D Brutsaert, Fabiola Leon-Velarde, Lorna G Moore, Enrique Vargas, Paul M McKeigue, et al. A genomewide admixture mapping panel for hispanic/latino populations. *The American Journal of Human Genetics*, 80(6):1171–1178, 2007.

[119] Brian K Maples, Simon Gravel, Eimear E Kenny, and Carlos D Bustamante. Rfmix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2):278–288, 2013.

[120] Gabor T Marth, Eva Czabarka, Janos Murvai, and Stephen T Sherry. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166(1):351–372, 2004.

[121] Iain Mathieson and Gil McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, 44(3):243–246, 2012.

[122] PM McKeigue, JR Carpenter, EJ Parra, and MD Shriver. Estimation of admixture and detection of linkage in admixed populations by a bayesian approach: application to african-american populations. *Annals of human genetics*, 64(2):171–186, 2000.

[123] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[124] Graham McVicker, Bryce van de Geijn, Jacob F Degner, Carolyn E Cain, Nicholas E Banovich, Anil Raj, Noah Lewellen, Marsha Myrthil, Yoav Gilad, and Jonathan K Pritchard. Identification of genetic variants that affect histone modifications in human cells. *Science*, 342(6159):747–749, November 2013.

[125] Michael L Metzker. Sequencing technologiesthe next generation. *Nature reviews genetics*, 11(1):31–46, 2010.

[126] Giovanni Montana and Jonathan K Pritchard. Statistical tests for admixture mapping with case-control and cases-only data. *The American Journal of Human Genetics*, 75(5):771–789, 2004.

[127] Carrie B Moore, John R Wallace, Daniel J Wolfe, Alex T Frase, Sarah A Pendergrass, Kenneth M Weiss, and Marylyn D Ritchie. Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. *PLoS Genet*, 9(12):e1003959, 2013.

[128] Dawn Muddyman, Carol Smee, Heather Griffin, and Jane Kaye. Implementing a successful data-management framework: the uk10k managed access model. *Genome medicine*, 5(11):100, 2013.

[129] Sean Myles, Dan Davison, Jeffrey Barrett, Mark Stoneking, and Nic Timpson. Worldwide population differentiation at disease-associated snps. *BMC medical genomics*, 1(1):22, 2008.

[130] Anton Nekrutenko and James Taylor. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, 13(9):667–672, 2012.

[131] Matthew R Nelson, Daniel Wegmann, Margaret G Ehm, Darren Kessner, Pamela St Jean, Claudio Verzilli, Judong Shen, Zhengzheng Tang, Silviu-Alin Bacanu, Dana Fraser, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–104, 2012.

[132] Elias Chaibub Neto, Mark P. Keller, Alan D. Attie, and Brian S. Yandell. Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Ann Appl Stat*, 4(1):320–339, 3 2010.

[133] Maggie C Y Ng, Jessica M Hester, Maria R Wing, Jiang Li, Jianzhao Xu, Pamela J Hicks, Bong H Roh, Lingyi Lu, Jasmin Divers, Carl D Langefeld, Barry I Freedman, Nichole D Palmer, and Donald W Bowden. Genome-wide association of BMI in African Americans. *Obesity (Silver Spring, Md.)*, 20(3):622–627, March 2012.

[134] Pauline C Ng and Steven Henikoff. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.*, 7:61–80, 2006.

[135] Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, 42(1):30–35, 2010.

[136] Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, Jay Shendure, and Michael J Bamshad. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, 42(1):30–35, January 2010.

[137] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS genetics*, 6(4):e1000888, April 2010.

[138] Jo Nishino and Shuhei Mano. The number of candidate variants in exome sequencing for mendelian disease under no genetic heterogeneity. *Computational and mathematical methods in medicine*, 2013, 2013.

[139] John Novembre and Anna Di Rienzo. Spatial patterns of variation due to natural selection in humans. *Nature Reviews Genetics*, 10(11):745–755, 2009.

[140] Michael Oldridge, M Ana, Monika Maringa, Peter Propping, Sahar Mansour, Christine Pollitt, Thomas M DeChiara, Robert B Kimble, David M Valenzuela, George D Yancopoulos, et al. Dominant mutations in ror2, encoding an orphan receptor tyrosine kinase, cause brachydactyly type b. *Nature genetics*, 24(3):275–278, 2000.

[141] Bogdan Pasaniuc, Nadin Rohland, Paul J McLaren, Kiran Garimella, Noah Zaitlen, Heng Li, Namrata Gupta, Benjamin M Neale, Mark J Daly, Pamela Sklar, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature genetics*, 44(6):631–635, 2012.

[142] Bogdan Pasaniuc, Sriram Sankararaman, Gad Kimmel, and Eran Halperin. Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, 25(12):i213–i221, 2009.

[143] Bogdan Pasaniuc, Sriram Sankararaman, Dara G Torgerson, Christopher Gignoux, Noah Zaitlen, Celeste Eng, William Rodriguez-Cintron, Rocio Chapela, Jean G Ford, Pedro C Avila, et al. Analysis of latino populations from gala and mec studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics*, page btt166, 2013.

[144] Bogdan Pasaniuc, Noah Zaitlen, Guillaume Lettre, Gary K Chen, Arti Tandon, WH Linda Kao, Ingo Ruczinski, Myriam Fornage, David S Siscovick, Xiaofeng Zhu, et al. Enhanced statistical tests for gwas in admixed populations: assessment using african americans from care and a breast cancer consortium. *PLoS Genet*, 7(4):e1001371, 2011.

[145] Tomi Pastinen. Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics*, 11(8):533–538, 2010.

[146] Nick Patterson, Neil Hattangadi, Barton Lane, Kirk E Lohmueller, David A Hafler, Jorge R Oksenberg, Stephen L Hauser, Michael W Smith, Stephen J OBrien, David Altshuler, et al. Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics*, 74(5):979–1000, 2004.

[147] J. Pearl. Causal inference in statistics: An overview. statistics surveys 3 96–146, 2009.

[148] Joseph K Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American journal of human genetics*, 94(4):559–573, April 2014.

[149] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, April 2010.

[150] Andrzej Polanski and Marek Kimmel. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, 165(1):427–436, 2003.

[151] Snehit Prabhu and Itsik Pe'er. Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome research*, 22(11):2230–2240, November 2012.

[152] Alkes L Price, Agnar Helgason, Gudmar Thorleifsson, Steven A McCarroll, Augustine Kong, and Kari Stefansson. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS genetics*, 7(2):e1001317, February 2011.

[153] Alkes L Price, Nick Patterson, Fuli Yu, David R Cox, Alicja Waliszewska, Gavin J McDonald, Arti Tandon, Christine Schirmer, Julie Neubauer, Gabriel Bedoya, et al. A genomewide admixture map for latino populations. *The American Journal of Human Genetics*, 80(6):1024–1036, 2007.

[154] Alkes L Price, Arti Tandon, Nick Patterson, Kathleen C Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, 5(6):e1000519, 2009.

[155] J K Pritchard and M Przeworski. Linkage disequilibrium in humans: models and data. *American journal of human genetics*, 69(1):1–14, July 2001.

[156] Huaizhen Qin, Nathan Morris, Sun J Kang, Mingyao Li, Bamidele Tayo, Helen Lyon, Joel Hirschhorn, Richard S Cooper, and Xiaofeng Zhu. Interrogating local population structure for fine mapping in genome-wide association studies. *Bioinformatics*, 26(23):2961–2968, 2010.

[157] Huaizhen Qin and Xiaofeng Zhu. Power comparison of admixture mapping and direct association analysis in genome-wide association studies. *Genetic epidemiology*, 36(3):235–243, 2012.

[158] Heidi L Rehm. Disease-targeted sequencing: a cornerstone in the clinic. *Nature Reviews Genetics*, 14(4):295–300, 2013.

[159] Jared C Roach, Gustavo Glusman, Arian FA Smit, Chad D Huff, Robert Hubley, Paul T Shannon, Lee Rowen, Krishna P Pant, Nathan Goodman, Michael Bamshad, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978):636–639, 2010.

[160] Peter N Robinson, Sebastian Köhler, Anika Oellrich, Kai Wang, Christopher J Mungall, Suzanna E Lewis, Nicole Washington, Sebastian Bauer, Dominik Seelow, Peter Krawitz, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome research*, 24(2):340–348, 2014.

[161] Edyta Rohmann, Han G Brunner, Hülya Kayserili, Oya Uyguner, Gudrun Nürnberg, Erin D Lew, Angus Dobbie, Veraragavan P Eswarakumar, Abdullah Uzumcu, Melike Ulubil-Emeroglu, et al. Mutations in different components of fgf signaling in ladd syndrome. *Nature genetics*, 38(4):414–417, 2006.

[162] Noah A Rosenberg, Jonathan K Pritchard, James L Weber, Howard M Cann, Kenneth K Kidd, Lev A Zhivotovsky, and Marcus W Feldman. Genetic structure of human populations. *science*, 298(5602):2381–2385, 2002.

[163] Sriram Sankararaman, Srinath Sridhar, Gad Kimmel, and Eran Halperin. Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, 82(2):290–303, 2008.

[164] Eric E. Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj Guhathakurta, Solveig K. Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, Pek Yee Lum, Amy Leonardson, Rolf Thieringer, Joseph M. Metzger, Liming Yang, John Castle, Haoyuan Zhu, Shera F. Kash, Thomas A. Drake, Alan Sachs, and Aldons J. Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*, 37(7):710–7, 7 2005.

[165] KF Schilter, Adele Schneider, Tanya Bardakjian, J-F Soucy, Rebecca C Tyler, Linda M Reis, and Elena V Semina. Otx2 microphthalmia syndrome: four novel mutations and delineation of a phenotype. *Clinical genetics*, 79(2):158–168, 2011.

[166] Exome Variant Server NHLBI GO Exome Sequencing Project (ESP) Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) (Accessed August 2013).

[167] Michael F Seldin, Bogdan Pasaniuc, and Alkes L Price. New approaches to disease mapping in admixed populations. *Nature Reviews Genetics*, 12(8):523–528, 2011.

[168] Hanan E Shamseldin, Maha A Faden, Walid Alashram, and Fowzan S Alkuraya. Identification of a novel dlx5 mutation in a family with autosomal recessive split hand and foot malformation. *Journal of medical genetics*, pages jmedgenet–2011, 2011.

[169] Sagiv Shifman and Ariel Darvasi. The value of isolated populations. *Nature genetics*, 28(4):309–310, 2001.

[170] Daniel Shriner, Adebowale Adeyemo, Edward Ramos, Guanjie Chen, and Charles N Rotimi. Mapping of disease-associated variants in admixed populations. *Genome biology*, 12(5):223, 2011.

[171] Daniel Shriner, Adebowale Adeyemo, and Charles N Rotimi. Joint ancestry and association testing in admixed individuals. *PLoS Comput Biol*, 7(12):e1002325, 2011.

[172] George Davey Smith and Shah Ebrahim. Mendelian randomization: prospects, potentials, and limitations. *International journal of epidemiology*, 33(1):30–42, 2004.

[173] Stuart B Smith, Hui-Qi Qu, Nadine Taleb, Nina Y Kishimoto, David W Scheel, Yang Lu, Ann-Marie Patch, Rosemary Grabs, Juehu Wang, Francis C Lynn, et al. Rfx6 directs islet formation and insulin production in mice and humans. *Nature*, 463(7282):775–780, 2010.

[174] Nara LM Sobreira, Elizabeth T Cirulli, Dimitrios Avramopoulos, Elizabeth Wohler, Gretchen L Oswald, Eric L Stevens, Dongliang Ge, Kevin V Shianna, Jason P Smith, Jessica M Maia, et al. Whole-genome sequencing of a single proband together with linkage analysis identifies a mendelian disease gene. *PLoS Genet*, 6(6):e1000991, 2010.

[175] Sarah L Spain and Jeffrey C Barrett. Strategies for fine-mapping complex traits. *Human Molecular Genetics*, 24(R1):R111–9, October 2015.

[176] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc*, 7(3):500–7, 3 2012.

[177] G Stevanin, C Paternotte, P Coutinho, S Klebe, N Elleuch, JL Loureiro, E Denis, VT Cruz, A Dürr, J-F Prudhomme, et al. A new locus for autosomal recessive spastic paraplegia (spg32) on chromosome 14q12-q21. *Neurology*, 68(21):1837–1840, 2007.

[178] Mark Stoneking and Johannes Krause. Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics*, 12(9):603–614, 2011.

[179] Barbara E Stranger, Alexandra C Nica, Matthew S Forrest, Antigone Dimas, Christine P Bird, Claude Beazley, Catherine E Ingle, Mark Dunning, Paul Flicek, Daphne Koller, Stephen Montgomery, Simon Tavaré, Panos Deloukas, and Emmanouil T Dermitzakis. Population genomics of human gene expression. *Nature genetics*, 39(10):1217–1224, September 2007.

[180] Jae Hoon Sul, Buhm Han, Chun Ye, Ted Choi, and Eleazar Eskin. Effectively identifying eqtls from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genet*, 9(6):e1003491, 6 2013.

[181] Andreas Sundquist, Eugene Fratkin, Chuong B Do, and Serafim Batzoglou. Effect of genetic divergence in identifying ancestral origin using hapaa. *Genome research*, 18(4):676–682, 2008.

[182] Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, 79(1):1–12, 2006.

[183] Aaron Taudt, Maria Colomé-Tatché, and Frank Johannes. Genetic sources of population epigenomic variation. *Nature reviews. Genetics*, 17(6):319–332, June 2016.

[184] Jacob A Tennessen, Abigail W Bigham, Timothy D OConnor, Wenqing Fu, Eimear E Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *science*, 337(6090):64–69, 2012.

[185] Chao Tian, David A Hinds, Russell Shigeta, Sharon G Adler, Annette Lee, Madeleine V Pahl, Gabriel Silva, John W Belmont, Robert L Hanson, William C Knowler, et al. A genomewide single-nucleotide–polymorphism panel for mexican american admixture mapping. *The American Journal of Human Genetics*, 80(6):1014–1023, 2007.

[186] Joris A Veltman and Han G Brunner. De novo mutations in human genetic disease. *Nature reviews. Genetics*, 13(8):565, 2012.

[187] Jean-Baptiste Veyrieras, Sridhar Kudaravalli, Su Yeon Kim, Emmanouil T Dermitzakis, Yoav Gilad, Matthew Stephens, and Jonathan K Pritchard. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS genetics*, 4(10):e1000214, October 2008.

[188] Sten Wahlund. Zusammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas*, 11(1):65–106, 1928.

[189] Jonathan T Wang, Chin-Jia Lin, Sandra M Burridge, Glenn K Fu, Malgorzata Labuda, Anthony A Portale, and Walter L Miller. Genetics of vitamin d 1α-hydroxylase deficiency in 17 families. *The American Journal of Human Genetics*, 63(6):1694–1702, 1998.

[190] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[191] Xuexia Wang, Xiaofeng Zhu, Huaizhen Qin, Richard S Cooper, Warren J Ewens, Chun Li, and Mingyao Li. Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics*, 27(5):670–677, 2011.

[192] Zhanyong Wang, Jae Hoon Sul, Sagi Snir, Jose A Lozano, and Eleazar Eskin. Gene-Gene Interactions Detection Using a Two-stage Model. *Journal of computational biology : a journal of computational molecular cell biology*, 22(6):563–576, June 2015.

[193] Sebastian M Waszak, Olivier Delaneau, Andreas R Gschwind, Helena Kilpinen, Sunil K Raghav, Robert M Witwicki, Andrea Orioli, Michael Wiederkehr, Nikolaos I Panousis, Alisa Yurovsky, Luciana Romano-Palumbo, Alexandra Planchon, Deborah Bielser, Ismael Padioleau, Gilles Udin, Sarah Thurnheer, David Hacker, Nouria Hernandez, Alexandre Reymond, Bart Deplancke, and Emmanouil T Dermitzakis. Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell*, 162(5):1039–1050, August 2015.

[194] Daniel Wegmann, Darren E Kessner, Krishna R Veeramah, Rasika A Mathias, Dan L Nicolae, Lisa R Yanek, Yan V Sun, Dara G Torgerson, Nicholas Rafaels, Thomas Mosley, et al. Recombination rates in admixed individuals identified by ancestry-based inference. *Nature genetics*, 43(9):847–853, 2011.

[195] Wellcome Trust Case Control Consortium, Julian B Maller, Gilean McVean, Jake Byrnes, Damjan Vukcevic, Kimmo Palin, Zhan Su, Joanna M M Howson, Adam Auton, Simon Myers, Andrew Morris, Matti Pirinen, Matthew A Brown, Paul R Burton, Mark J Caulfield, Alastair Compston, Martin Farrall, Alistair S Hall, Andrew T Hattersley, Adrian V S Hill, Christopher G Mathew, Marcus Pembrey, Jack Satsangi, Michael R Stratton, Jane Worthington, Nick Craddock, Matthew Hurles, Willem Ouwehand, Miles Parkes, Nazneen Rahman, Audrey Duncanson, John A Todd, Dominic P Kwiatkowski, Nilesh J Samani, Stephen C L Gough, Mark I McCarthy, Panagiotis Deloukas, and Peter Donnelly. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics*, 44(12):1294–1301, December 2012.

[196] Andrew R Wood, Marcus A Tuke, Mike A Nalls, Dena G Hernandez, Stefania Bandinelli, Andrew B Singleton, David Melzer, Luigi Ferrucci, Timothy M Frayling, and Michael N Weedon. Another explanation for apparent epistasis. *Nature*, 514(7520):E3–5, October 2014.

[197] Fred A Wright, Patrick F Sullivan, Andrew I Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, Rick Jansen, Wonil Chung, Yi-Hui Zhou, Abdel Abdellaoui, Sandra Batista, Casey Butler, Guanhua Chen, Ting-Huei Chen, David D'Ambrosio, Paul Gallins, Min Jin Ha, Jouke Jan Hottenga, Shunping Huang, Mathijs Kattenberg, Jaspreet Kochar, Christel M Middeldorp, Ani Qu, Andrey Shabalin, Jay Tischfield, Laura Todd, Jung-Ying Tzeng, Gerard van Grootheest, Jacqueline M Vink, Qi Wang, Wei Wang, Weibo Wang, Gonneke Willemsen, Johannes H Smit, Eco J de Geus, Zhaoyu Yin, Brenda W J H Penninx, and Dorret I Boomsma. Heritability and genomics of gene expression in peripheral blood. *Nature genetics*, 46(5):430–437, May 2014.

[198] Jun J Yang, Cheng Cheng, Meenakshi Devidas, Xueyuan Cao, Yiping Fan, Dario Campana, Wenjian Yang, Geoff Neale, Nancy J Cox, Paul Scheet, et al. Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. *Nature genetics*, 43(3):237–241, 2011.

[199] Shengjie Yang, Yiyuan Liu, Ning Jiang, Jing Chen, Lindsey Leach, Zewei Luo, and Minghui Wang. Genome-wide eQTLs and heritability for gene expression traits in unrelated individuals. *BMC genomics*, 15(1):13, January 2014.

[200] Wen-Yun Yang, John Novembre, Eleazar Eskin, and Eran Halperin. A model-based approach for analysis of spatial structure in genetic data. *Nature genetics*, 44(6):725–731, 2012.

[201] Xiong Yang, Suzanne Al-Bustan, Qidi Feng, Wei Guo, Zhiming Ma, Makia Marafie, Sindhu Jacob, Fahd Al-Mulla, and Shuhua Xu. The influence of admixture and consanguinity on population genetic diversity in middle east. *Journal of human genetics*, 59(11):615–622, 2014.

[202] Yaping Yang, Donna M Muzny, Jeffrey G Reid, Matthew N Bainbridge, Alecia Willis, Patricia A Ward, Alicia Braxton, Joke Beuten, Fan Xia, Zhiyv Niu, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *New England Journal of Medicine*, 369(16):1502–1511, 2013.

[203] Michael Yourshaw, S Paige Taylor, Aliz R Rao, Martín G Martín, and Stanley F Nelson. Rich annotation of dna sequencing variants by leveraging the ensembl variant effect predictor with plugins. *Briefings in bioinformatics*, page bbu008, 2014.

[204] Qing Yue, Joanna C Jen, Stanley F Nelson, and Robert W Baloh. Progressive ataxia due to a missense mutation in a calcium-channel gene. *The American Journal of Human Genetics*, 61(5):1078–1087, 1997.

[205] Noah Zaitlen, Sara Lindström, Bogdan Pasaniuc, Marilyn Cornelis, Giulio Genovese, Samuela Pollack, Anne Barton, Heike Bickeböller, Donald W Bowden, Steve Eyre, et al. Informed conditioning on clinical covariates increases power in case-control association studies. *PLoS Genet*, 8(11):e1003032, 2012.

[206] Noah Zaitlen, Bogdan Pasaniuc, Tom Gur, Elad Ziv, and Eran Halperin. Leveraging genetic variability across populations for the identification of causal variants. *American journal of human genetics*, 86(1):23–33, January 2010.

[207] Matthew Zawistowski, Mark Reppell, Daniel Wegmann, Pamela L St Jean, Margaret G Ehm, Matthew R Nelson, John Novembre, and Sebastian Zöllner. Analysis of rare variant population structure in europeans explains differential stratification of gene-based tests. *European Journal of Human Genetics*, 22(9):1137–1144, 2014.

[208] Tanja Zeller, Philipp Wild, Silke Szymczak, Maxime Rotival, Arne Schillert, Raphaele Castagne, Seraya Maouche, Marine Germain, Karl Lackner, Heidi Rossmann, Medea Eleftheriadis, Christoph R Sinning, Renate B Schnabel, Edith Lubos, Detlev Mennerich, Werner Rust, Claire Perret, Carole Proust, Viviane Nicaud, Joseph Loscalzo, Norbert Hübner, David Tregouet, Thomas Münzel, Andreas Ziegler, Laurence Tiret, Stefan Blankenberg, and François Cambien. Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PloS one*, 5(5):e10693, May 2010.

[209] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, and Jian Yang. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics*, 48(5):481–487, May 2016.

[210] Alexander Zimprich, Saskia Biskup, Petra Leitner, Peter Lichtner, Matthew Farrer, Sarah Lincoln, Jennifer Kachergus, Mary Hulihan, Ryan J Uitti, Donald B Calne, et al. Mutations in lrrk2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron*, 44(4):601–607, 2004.