New Computational Approaches for Analyzing Admixed Populations

by

Danny Park

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

New Computational Approaches for Analyzing Admixed Populations

by

Danny Park

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

6/8/17 Chair

Elad Ziv 6/5/17

John Witte 6/5/17

Committee in Charge

Dedication and Acknowledgments

I am sure I could write an entire thesis on the people I am grateful for and please accept my apologies for any oversight.

My time at UCSF was an unexpected one since I had never really planned on doing a PhD. I was convinced by a good friend to give genomics/statistical genetics a try and it turned out to be a fascinating and fulfilling field to be a part of. I owe much of my growth as a scientist to the culture of my home program at UCSF, PSPG. The program gave me the flexibility to explore new areas, given my background in math and computer science. It also provided support for me when I felt as if I had failed (e.g. Cell Biology). Naturally, the culture of the program is due to the direction of its leadership. Deanna Kroetz, Nadav Ahitiuv, and Debbie Acoba have been instrumental to whatever successes I have achieved and my completion of the program owes so much to their guidance and support. I hope their leadership of the program continues to foster a culture of innovation and collaboration. My classmates have also been with me since the beginning of this journey. Luckily, my class ended up being a tight knight group that has kept in touch regularly. It has been an honor knowing you guys and watching everybody's personal growth. I hope we stay great friends in the future.

Noah Zaitlen, my advisor, has <u>obviously</u> been a tremendous part of my career at UCSF. I was Noah's first rotation student, and subsequently his first graduate student. I joined Noah's lab just as he was joining UCSF and although joining a junior PI's lab is often seen as a bit of a gamble, it never felt like one. With Noah I was able to work on intellectually challenging and fulfilling projects and was pushed to grow as a scientist. I am sure that I pushed Noah to be a better PI given the Korean drama my PhD career ended up becoming. His generosity, compassion, and support have made my graduate school career infinitely better and I am immensely grateful.

All the members of the Zaitlen, Hoffman, Witte, and Ziv labs also deserve tremendous credit for being great listeners. I appreciate them letting me continuously bounce ideas off them and for their constructive feedback. I also want to thank Chris Gignoux and Eimear Kenny for their friendship and their intellectual help. I look forward to working with all you guys well into the future.

Of course I would not be here without all my close friends. Josh Wellman and Jarreau Gilbert, thanks for hanging out with me since first grade! I can't imagine how bland elementary/middle/high school would have been without you guys. Eric Gamazon, I can not thank you enough for making sure I went to college and for your continued guidance. You truly are my brother from another mother. Who knows what I would be doing right now had you not encouraged me to go to the University of Chicago. Lance Leong, I am grateful that we ended up being friends in college. I appreciate your insight into touchy issues and willingness to have an open mind. Will Song, Jaehoon Choe, Jordan Weil, and Danny Zadkovich, thank you for the great friendship and all the great times from college until now. Phil Cheng and Kristina Estrada, thanks for always accommodating me and being gracious hosts whenever I am in San Diego. Rich and Julie Song, thank you for always opening your home to me and always welcoming me into your busy lives. John and Jeannie Bang, thanks for keeping me in your lives and visiting me regularly in SF. Simon Wong and Calvin Ku, thanks for competing with me and keeping me on my toes on the racetrack. Dustin Eichler, Janet Choi, and Hanyul Lee, thanks for taking me in while I was in New York and making me feel like part of your family. Again, thank you to all my friends. Your friendships have been an integral part of my life and I hope my friendship has brought some amount of happiness to yours.

My family unfortunately has had to endure some very trying times during my time at UCSF. However, during that time they never wavered in their support for me. My parents have set the tone for how hard I should work ever since I saw them working three jobs as a child. Thank you so much for showing how far hard work and dedication goes. I will never forget and never stop to appreciate the effort you have put into building our lives here in America.

Last but not least, I want to thank Andrea, my fiancée, for your belief in me and for always being more sure of this than I was. I am lucky to have met someone who was willing to put up with my ignorance and who has opened my eyes to the world. Thanks for giving me the courage to continue and try new things. I can't wait to see what the future holds for us and I am excited to share more memories with you.

*The text of this thesis/dissertation/manuscript is a reprint of the material as it appears in BMC Bioinformatics (PIGS) and Bioinformatics (Adapt-Mix). The co-authors listed in this publication helped with the research that forms the basis for the dissertation/thesis.

Abstract

This work revolves around developing, applying, and evaluating computational methods for the analysis of genomic datasets of recently admixed populations. Individuals from recently admixed populations, such as African Americans and Hispanic Latinos, derive their genomes from multiple genetically distinct ancestral populations. For example, African Americans have locus-specific ancestry from African and European genomes, which reflects demographic history and influences disease predisposition. Genomic studies of admixed populations therefore provide an enormous opportunity to investigate the influence of genetic variation on human phenotypic diversity. Furthermore, such studies offer a framework to test the generalizability of findings on genotype-phenotype relationships originally obtained in more homogeneous populations (i.e. Europeans), potentially yielding insights into underlying mechanisms. Here we present four novel statistical/computational approaches that leverage the unique genetic makeup of admixed populations to aid in deepening our understanding of the effect of human genetic variation on the phenome.

Table of Contents

List of Tables

Chapter 1: Building Upon Existing Methods in Statistical Genetics

Chapter 2: Leveraging Admixed Populations to Understand Genes and Environment

List of Tables (cont'd)

List of Figures

Chapter 1: Building Upon Existing Methods in Statistical Genetics

Chapter 2: Leveraging Admixed Populations to Understand Genes and Environment

List of Figures (cont'd)

Introduction

Recent major advances in technologies have led to explosive growth in the volume of biological data available for research. Whole-genome, transcriptome, epigenetic, and other omics data for thousands of individuals are now available for scientific research(T. 1. G. P. Consortium, 2012; Gibbs et al., 2003; The ENCODE Project Consortium, 2004). Many of these datasets are linked to health records or other medically relevant phenotypes and covariates that have been collected through carefully planned scientific studies(Burchard, Avila, Nazario, & Casal, 2004; T. C. A. D. C. G. Consortium, 2011; Sudlow et al., 2015). However, analyzing these datasets come with numerous well-known challenges, including population structure, data size, and environmental noise(Aschard et al., 2012; Bryc et al., 2010; Seldin, Pasaniuc, & Price, 2011; Sul et al., 2016). To address these issues, computational and statistical methods have been developed to analyze these datasets in a computationally efficient and statistically powerful manner. These methods have often been fruitful and resulted in a large number of discoveries to date(Visscher, Brown, McCarthy, & Yang, 2012). Of note is the large number of Genome Wide Association Studies (GWAS) loci that have been found to be associated with a wide range of phenotypes(Welter et al., 2014). Although we have had successes in identifying genetic loci that contribute to disease risk, the majority of GWAS studies have been and continue to be focused on individuals of European descent(Bustamante, La Vega, & Burchard, 2011; Popejoy & Fullerton, 2016). As Bustamante *et al.* have warned, however, failing to incorporate diverse populations in genomic studies may result in a biased understanding of the impact of genetic variation on the human phenome, and genomic medicine may end up benefiting "largely a privileged few"(Bustamante et al., 2011).

This work revolves around developing, applying and evaluating computational methods for the analysis of datasets of recently admixed populations. Recently admixed populations, such as African Americans and Hispanic Latinos derive their genomes from multiple genetically distinct ancestral populations. For example, African Americans derive parts of their genome from African and European ancestors, reflecting recent demographic history and influencing predisposition to disease and other traits. Consequently, genomic studies of admixed populations provide a unique opportunity to evaluate the generalizability of findings initially derived from more homogeneous populations (i.e. Europeans).

First, we present improvements to existing methods that are used to analyze human genetic data and apply them to admixed populations. We start by presenting an approach for improved detection of segments of the genome that are inherited identical-by-descent (IBD). IBD has been used for such applications as detecting cryptic relatedness between individuals, estimating components of heritability, inferring evolutionary and demographic history, and mapping disease loci(B. L. Browning & Browning, 2011; Gusev et al., 2011; He, 2013; Palamara, Lencz, Darvasi, & Pe'er, 2012; Zaitlen et al., 2013). Therefore, the identification of IBD segments from genotyping and sequencing data has important implications for studies of complex human phenotypes. We then consider how to improve summary statistics based approaches by learning local genetic correlations. Summary statistics of association tests, such as effect size estimates and their standard errors, are becoming the datatype of choice for many analyses because they are orders of magnitude faster than their genotype-based counterparts and because they are often necessary due to privacy and legal issues(Schork et al., 2013). Moving forward, integration of summary statistics into genomic analyses will be vital for advancing our knowledge of various complex diseases and phenotypes.

Next, we present methods that leverage the unique genetic makeup of admixed populations to detect the existence of gene-environment interactions and quantify the effects of the genetic component versus environmental factors on quantitative traits. Admixed genomes can be thought of as a mosaics of different ancestral genomic segments(Seldin et al., 2011). For each individual, the proportion of ancestry derived from population p (θ_p) can be accurately estimated (e.g. the fraction of European/African ancestry in African Americans). θ_p, or global ancestry, has been shown to be correlated with an array of environmental and biomedical covariates as well as many genetic markers(Burchard et al., 2004; Cheng et al., 2012; Florez et al., 2011; Ziv et al., 2006). θ_p can therefore be used as a proxy for environmental and genetic covariates and tested for interaction with genetic loci. We present a method for testing for gene-ancestry interactions and investigate its suitability for studies of the transcriptome and the methylome. We also develop a method that partitions the correlation between phenotype and genetic ancestry θ_p into genetic and environmental components. Suppose a disease has differential prevalence between Africans and Europeans. The extent to which genetic component drives the difference, will induce a correlation between disease status and genetic ancestry in admixed populations that derive their genomes from the two continental populations. However, genetic ancestry has also been shown to be correlated with environmental factors such as smoking and socio-economic status, and thus it is unclear how much of a given trait-ancestry correlation is due to genetic variation(Fejerman et al., 2008; Florez et al., 2011). Partitioning the correlation into genetic and environmental components has important implications for precision medicine and global health since knowing their relative contribution to a disease may provide crucial insights into pathophysiological mechanisms and therapeutic strategies.

Finally, we will end by discussing the future implications of the methods developed here. Previously isolated populations are now significantly less isolated from each other as a result of world-historical demographic events such as colonization, slave trade, and mass migration. These events are reflected in the existence of recently admixed populations, such as Native Hawaiians and Hispanic Latinos(Seldin et al., 2011). In the future, further admixtures between populations are likely to occur, given the enormous advances that have brought populations into closer contact; analyzing admixed genomes will thus require powerful approaches that leverage and account for their unique ancestral genetic makeups. As we will discuss, this will allow for a greater understanding of the relative contributions of genetic and environmental components to disease risk and other traits. This understanding promises to tailor treatment options as well as to create therapies for previously untreatable diseases. As discussed previously, most studies to date have focused on individuals of European descent and there is an urgent need to broaden our understanding of human genetic variation. This is especially the case since knowledge of genetic variation in one population is not broadly applicable to other populations, as Martin *et al.* has previously suggested(Martin et al., 2017). This work hopes to contribute to the broadening of understanding of the impact of genetic variation on human phenotypic diversity.

Chapter 1: Building Upon Existing Methods in Statistical Genetics

1. Improving Identity-by-Descent Segment Discovery by Probabilistic IBD Graph Sampling

Introduction

Identity-by-descent (IBD) is a fundamental genetics concept with broad applications to both medical and population genetics(Visscher, Hill, & Wray, 2008). Two haplotypes are identical-by-state (IBS) if they share the same sequence. Two haplotypes are IBD if they are both IBS and were inherited from a common ancestor(Whittemore & Halpern, 1994). IBD therefore contains information both about sequence similarity and about the historical relationship of individuals. IBD has been used for such applications as detecting cryptic relatedness between individuals(Weir, Anderson, & Hepler, 2006), estimating components of heritability(Zaitlen et al., 2013), inferring evolutionary and demographic history(Albrechtsen, Moltke, & Nielsen, 2010; Hochreiter, 2013; Palamara et al., 2012), and mapping disease loci(Albrechtsen et al., 2009; B. L. Browning & Browning, 2011; S. R. Browning & Thompson, 2012; Gusev et al., 2009; Han, Kang, Raychaudhuri, & de Bakker, 2013; Purcell et al., 2007). Therefore, the identification of IBD segments from genome-wide genotyping studies, and more recently sequencing studies, has important implications for studies of complex human phenotypes.

The identification of IBD segments is challenging for both statistical and computational reasons. IBD segments may be missed due to genotyping or sequencing errors. Since IBD occurs at the level of haplotypes, the data are typically phased and phasing errors can induce false negatives. Small segments of IBD are especially challenging because their haplotype frequency must be accurately modeled often resulting in both false positive and false negative IBD calls(B. L. Browning & Browning, 2013). Finally, there are computational challenges because the number of potential IBD relationships at a locus is $O(2^{h(h-1)/2})$, where h is the number of haplotypes.

Two classes of methods for computing the probability of IBD between haplotypes have been developed. Multiway methods such as Moltke *et al.* 2011(Moltke, Albrechtsen, Hansen, Nielsen, & Nielsen, 2011), simultaneously estimate the probability of IBD over the haplotypes of all individuals in a study. While powerful, generally multiway approaches are not computationally efficient enough to examine whole genome data sets over a large number of individuals(Letouzé et al., 2012). Recently an efficient mulitway method, HapFABIA(Hochreiter, 2013), has been proposed but focuses on detecting very ancient IBD segments (i.e. <<1cM) and relies on the existence of rare variation in the data. In practice, pairwise methods such as Germline(Gusev et al., 2009) and Refined IBD(B. L. Browning & Browning, 2013) are used to detect segments of IBD between pairs of haplotypes independently. Germline uses a sliding-window dictionary approach to find putative IBD segments and relies on fragment length to estimate IBD probability. Refined IBD utilizes the Germline approach to identify putative IBD segments and then applies a hidden markov model (HMM) to compute haplotype frequencies and estimate IBD probabilities. Since these methods consider pairs of individuals independently, they are computationally efficient at the genome-wide scale. However, they do not exploit the clique structure of true IBD segments(Gusev et al., 2011; He, 2013), and lack power relative to multiway approaches for smaller IBD segments(He, 2013; Hochreiter, 2013; Moltke et al., 2011).

Here we introduce a novel method PIGS, which combines the computational efficiency of pairwise methods with the power of multiway methods. PIGS takes as input the IBD probabilities output by pairwise approaches. Then, to update the probability that a pair of haplotypes are IBD, it computes the probability of IBD conditional on the IBD probabilities of all other haplotypes pairs at the locus. Consider a pair of haplotypes with a low probability of being IBD according to a pairwise method. If both haplotypes are IBD with high probability to a shared set of many other haplotypes, then that pair has a higher probability of being IBD conditioned on the shared set. By leveraging the graph structure of the complete set of IBD segments we are able to produce more accurate estimate of IBD probabilities and thereby produce more powerful identification of IBD segments. We first present an exact algorithm for computing conditional IBD probabilities. However, because of the large number of potential IBD graphs we cannot compute exact probabilities in all cases. Instead, we propose an efficient sampling algorithm to approximate these probabilities in practice.

Method

We begin by defining an IBD graph that is constructed over a set of N haplotypes at a genomic locus as follows. Each haplotype is represented by a node and there exists an edge between nodes if the two haplotypes to which the nodes correspond are IBD at the locus. Valid IBD graphs obey a transitivity property such that if individuals 1 and 2 are IBD and individuals 2 and 3 are IBD, then individuals 1 and 3 are IBD(Braak et al., 2010). An IBD graph is *transitive* if the edges obey the transitivity property; otherwise the graph is *intransitive* and cannot represent the true state of IBD at the locus. Due to the transitivity property, all connected components of a valid IBD graph at a locus are cliques. We leverage the clique property of IBD graphs to improve the pairwise probabilities of IBD output by existing software packages such as Refined IBD(B. L. Browning & Browning, 2013).

Given the probability of IBD between all pairs of N haplotypes at a genomic locus, we construct a probabilistic IBD graph G_p=(*N, P*) as follows. Each haplotype i is represented by a node $n_i \in N$. For every pair of haplotypes there is an edge assigned probability $p_{ij} \in P$ where p_{ij} is the probability of IBD between haplotypes i and j at that locus.

Given a probabilistic IBD graph G_p we consider a proposed IBD graph g=(*N,E*) over the nodes of G_p. Any proposed IBD graph g represents a different scenario of how individuals in G_p. can be IBD to each other at the given genomic locus. The probability of g conditional on the probabilistic IBD graph $G_p$is computed as follows. We define *I* as the set of all IBD graphs derived from the nodes of G_p. For each $g \in I$ the conditional probability of g on G_p, *P*($g|G_p$) is the product of induced edge probabilities. An edge e_{ij} = 1 if it is present in g and it has induced probability p_{ij}. An edge e_{ij} = 0 if it is not in g and has induced probability (*1*-p_{ij}).

Eq. 1

$$P(g|G_p) = \prod_{\forall i,j;i \neq j} p_{ij}^{e^{ij}}(1 - p_{ij})^{(1-e^{ij})}$$

Our objective is to update the probability of each pair of individuals being IBD by conditioning on the probabilities of all pairs in the graph. The intuition is best understood with an example. Consider a probabilistic IBD graph with three nodes ($i \in \{1, 2, 3\}$) and suppose the initial pairwise probabilities have been assigned by a pairwise IBD-calling algorithm such that p_{12} = 0.9, p_{13} = 0.9, and p_{23} = 0.1. The edges e_{12} and e_{13} have a higher probability of IBD than e_{23}, but given that true IBD graphs obey transitivity, the probability of e_{12} and e_{13} conditioned on edge e_{23} will be lower. Similarly, the probability of e_{23} will be higher when conditioned on e_{12} and e_{13} as shown in Figure 1b. By the transitive rule when 2 of the 3 edges in a triangle have a high probability of IBD we expect the third edge to have a high probability of IBD as well.

The conditional probability of an edge given the probabilities of the graph, $\dot{p}_{ij} =$

$P(e_{ij} = 1 \mid G_p)$, is the sum of the probabilities of all transitive graphs in which the edge is present, divided by the sum of the probability of all transitive graphs. We compute the conditional probability using only transitive graphs since we know that an intransitive graph is a biologically implausible scenario. We define V as the set of transitive IBD graphs derived from the nodes of G_p.

Eq. 2

$$\dot{p}_{ij} = P(e_{ij} = 1 \mid G_p) = \frac{\sum_{g \in V; e_{ij} \in g} P(g \mid G_p)}{\sum_{g \in V} P(g \mid G_p)}$$

All such transitive graphs and their respective probabilities are shown in Figure 1a. For illustrative purposes, we include an intransitive graph in the bottom right of Figure 1a. We update each edge by using Equation 2 and the resulting graph with updated probabilities is shown in Figure 1b. To further clarify how we compute a conditional edge probability, we compute the conditional probability of edge e_{23}:

$$\dot{p}_{23} = P(e_{23} = 1 \mid G_p) = \frac{1(0.081) + 1(0.001) + 0(0.081) + 0(0.081) + 0(0.009)}{(0.081) + (0.001) + (0.081) + (0.081) + (0.009)} = 0.324$$

Computing exact conditional probabilities requires computing the probability of every transitive IBD graph, which has a sample space of size $O(2^{h(h-1)/2})$, where h is the number of haplotypes. Unfortunately, this is computationally infeasible to enumerate and so we develop a sampling method that can be used to efficiently approximate conditional edge probabilities.

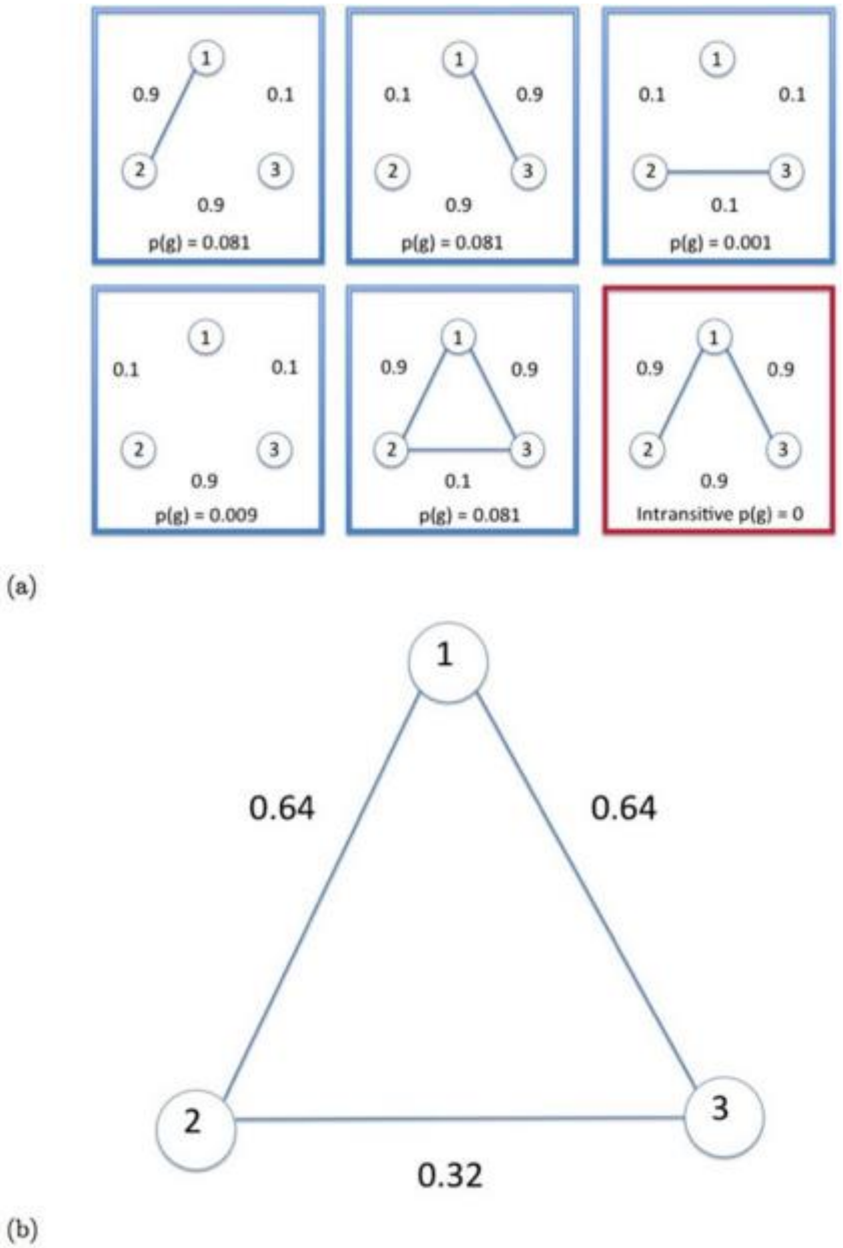


Figure 1 Example IBD Graphs.(a) Sample space of graphs and their respective probabilities. (b) Probabilistic IBD Graph with updated edge probabilities conditioned on the initial graph.

Efficient Computation of Conditional IBD Probabilities

Our sampling method starts by generating the probabilistic IBD graph for a given genomic location. We only consider the unique positions along the genome where the IBD graph changes, or more specifically, the points where the initial IBD segments begin or end. Analyzing other positions would be redundant because the positions provide no information about how the

IBD graph changes. An initial graph is generated by adding in all edges output by Refined IBD that pass a LOD score threshold. Alternative pairwise IBD probability methods may be used if desired. We identify the connected components of this graph because edges that are part of disjoint components have no effect on each other when computing updated probabilities. For each connected component *c*, we construct G_p by translating the RefinedIBD LOD scores to probabilities. A connected component *c* may have edges that were never called by Refined IBD and thus have a probability of 0. We assign uncalled edges the probability ε = 0.0046 in order to ensure that $P(g|G_p) > 0$ and that the edge can be sampled. Then instead of enumerating the set of all possible valid graphs V inducible by G_p we sample from V. We define N_g as the current sum of probabilities of all sampled graphs so far and N_{ij} as the current sum of probabilities of all sampled graphs containing edge e_{ij}. At any stage in the sampling process, the estimate of the conditional probability that individuals i and j are IBD is $\hat{p}_{ij} = \frac{N_{ij}}{N_g}$. If all valid graphs are sampled with equal N_g probability, this converges to the exact conditional probability shown in equation 2. The sampling procedure is given in Algorithm 1.

Algorithm 1 Graph sampling

```
Input: $G_p$
Output: $N_g$, $N_{ij}$ ($\forall i,j$; $i \neq j$)

Set $N_{ij}$= 0, $N_g$= 0
for all i, j do
        if $p_{ij} \geq 0.99$ then $e_{ij}$=1
        else $e_{ij}$= 0
Add edges to make all connected components of $G_p$cliques
Compute $P(g|G_p)$
$N_g$+= $P(g|G_p)$
for all i, j where $e_{ij}$=1 do $N_{ij}$+= $P(g|G_p)$
while $\hat{p}_{ij}$ has not converged $\forall i, j$ do
        Sample a random $e_{ij}$and set $e_{ij}$= 1 with probability $p_{ij}$
        Ensure graph transitivity
        Compute $P(g|G_p)$
        $N_g$+= $P(g|G_p)$
        for all i,j where $e_{ij}$=1 do $N_{ij}$+= $P(g|G_p)$
```

Edges are sampled according to a weighted distribution where weight w_{ij} is based on p_{ij} and is defined as:

$$w_{ij} = \begin{cases} Normal\ CDF(p_{ij};\ \mu = 0.5, \sigma = 0.234), if\ (p_{ij}) \leq 0.5 \\ 1 - Normal\ CDF(p_{ij};\ \mu = 0.5, \sigma = 0.234), otherwise \end{cases}$$

If σ ≈ 0, then edges with pij ≈ 1 or 0 will almost never be sampled since the selection weights of such edges will be infinitesimally small. Similarly if σ is too large, then all edges will be assigned similar selection weights and as a result graphs will be sampled uniformly instead of in proportion to their probability. We selected σ = 0.234 because it allows for efficient convergence times.

This weighted sampling scheme assures that edges with p_{ij}≈ 1 or 0 are sampled less often than edges with p_{ij}≈ 0.5. Intuitively this makes sense because we sample proposed IBD graphs in proportion to their respective $P(g|G_p)$. Edges with p_{ij}≈ 1 induce a proposed graph with a greater $P(g|G_p)$ when they are present. Similarly edges with p_{ij}≈ 0 induce a proposed graph with a greater $P(g|G_p)$ when they are missing. Thus, to sample the most probable graphs more often, edges with high values of p_{ij} should typically have e_{ij}= 1 and edges with low values of p_{ij}should typically have e_{ij}= 0. Changing the state of an edge can cause the proposed graph g to be intransitive. Therefore we add or remove edges from g to ensure transitivity. At each iteration if an edge has $p \geq 0.99$ then p_{ij} is set to 1 so that we never sample very high probability edges.

Algorithm 2 Ensure graph transitivity

Input: G_p, e_{ij}that was just added or removed
Output: Transitive G_p

```
if $e_{ij}$= 1 then
    Add edges to make all connected components of $G_p$cliques
else
    $S_i$ = nodes connected to $i$, $S_j$= nodes connected to $j$
    for all $p_{mk}$<0.99 do $e_{mk}$=0
    for each connected component $X$, where $|X| > 1$ do
        $\bar{p}_{iX} = \frac{p_{iX}}{|X|}, \bar{p}_{jX} = \frac{p_{jX}}{|X|}$ ∀ nodes $x \in X$
        if $\bar{p}_{iX} > \bar{p}_{jX}$ then add all nodes of $X$ to $S_i$
        else if $\bar{p}_{iX} < \bar{p}_{jX}$ add all nodes of $X$ to $S_j$
        else flip a fair coin
    for randomly selected $k \in (S_i \; U \; S_j)$ do
        $\bar{p}_{kS_i} = \frac{p_{ky}}{|S_i|}, \forall y \in S_i$
        $\bar{p}_{kS_j} = \frac{p_{kz}}{|S_j|}, \forall z \in S_j$
        if $\bar{p}_{kS_i} > \bar{p}_{kS_j}$ then add $k$ to $S_i$
        else if $\bar{p}_{kS_i} < \bar{p}_{kS_j}$ then add $k$ to $S_j$
        else flip a fair coin
    Make $S_i$ and $S_j$ cliques
```

We continue selecting edges until we reach a convergence criterion or a user-set time limit has been reached. As a convergence criterion we check if all edge probabilities change less than 1x10^{-11} for 5000 sequential iterations. With larger graphs that may never reach the convergence criteria, we allow the user to set a runtime limit. After convergence or hitting the user runtime limit, we output all edges and their respective updated probabilities. We tried a variety of sampling schemes to explore the space of graphs and selected this one due to its performance over simulated datasets.

IBD segments can span multiple regions and our method analyzes IBD at a single region. The probability of IBD between two individuals can therefore be output at multiple adjacent regions by our method. Furthermore, the IBD probability may be assigned a different value in each region due to the inexact nature of the sampling method. If the same IBD segment is assigned different probabilities across multiple loci we use the maximum value across all regions.

Once an IBD graph is analyzed using the sampling procedure, edges that were previously missing (i.e. those that were not called by Refined IBD) are output with a start and stop site that is equal to the intersection of all IBD segment boundaries in the graph. Since we do not look in the region for sequence identity between haplotypes we can only output the probability that IBD exists somewhere within the region. These new segments may also overlap with other called IBD segments. In order to reconcile overlapping IBD segments, we merge them provided that they pass a probability threshold set by the user and that they lie on the same haplotype. As the final probability, we use the maximum $\hat{p}_{ij}$ of the merged segments. For all analyses presented here, we only merged segments that had a probability of 0.99 or greater.

Converting LOD Scores to Probabilities

To find the relationship between LOD scores and the true positive rate of IBD segments we ran Refined IBD on simulated data using a LOD score cutoff of 0.1 and a length cutoff of 0.1

centimorgans. A true positive segment is defined as a predicted segment that is at least 50% true IBD. We fit a curve to the observed relationship between LOD score and true positive rate of IBD segments (see Figure 2). The equation of our curve is of the form p = (2*o*+*af*)/(*o*+*f*) where *o* = posterior odds, *f* = (*prior**(10^3)/.997)−(*prior**(10^3)), *a* = $(1 - LOD)^3/7$ if $LOD \leq 1$ 1, and a = −0.15 otherwise. The values for *f* and *a* were chosen to maximize the fit of the curve. Since the Refined IBD LOD score is the negative base 10 log likelihood of one shared haplotype divided by the likelihood of no shared haplotypes, we use Bayes rule for odds to convert a LOD score into a posterior odds: $O(A|B) = O(A) * \frac{L(A|B)}{L(A^c|B)}$, where $O(A|B)$ is the posterior odds, $O(A)$ is the prior odds, and $\frac{L(A|B)}{L(A^c|B)}$ is the likelihood ratio.

For the prior, we use the probability of any two individuals in the sample being IBD at any point in the genome, $\varepsilon = 0.0046$. This is the average proportion of the genome shared IBD between all pairs of individuals estimated using results from Refined IBD over simulated data. For edges that have a probability of 0 (i.e. an edge with no pairwise call), we assign a probability equal to the prior because otherwise these edges would never be sampled and graphs would have a probability of 0.

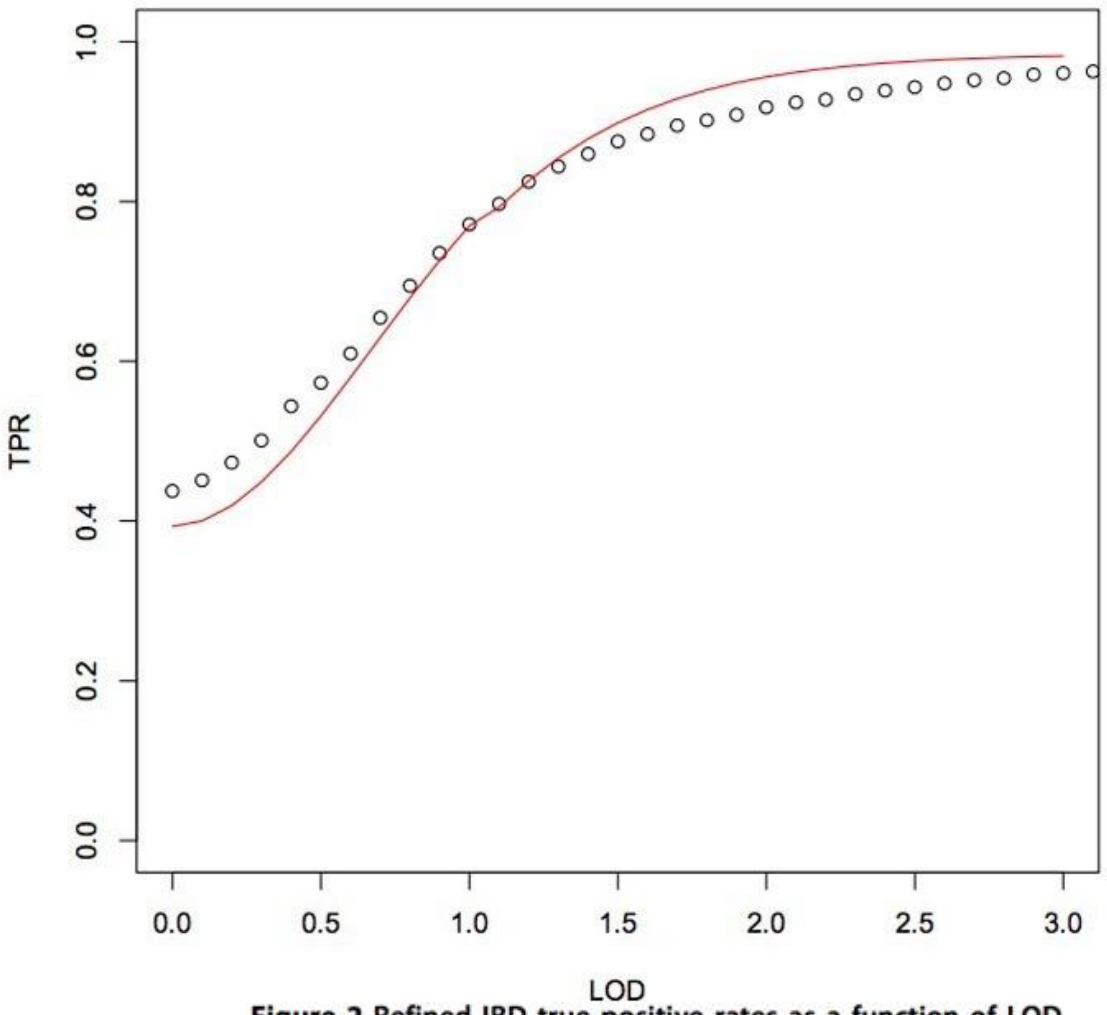


Figure 2 Refined IBD true positive rates as a function of LOD score. Refined IBD true positive rates as a function of LOD score are shown as black dots and our converted probabilities as a function of LOD score is shown as a red line.

Ideally, the relationship between LOD score and true positive rate is given by *p = odds/ (1 + odds)*. However, the relationship between LOD score and true positive rate in our sample of simulated individuals deviates from this theoretical relationship. Our function and *p = odds/ (1 + odds)* are of the same form (i.e. *g (x) = (cx +d)/ (mx +n)*). This served as our motivation in defining the conversion function. Lastly, given that the simulated data we generated is reflective of European population growth, the relationship between LOD score and true positive rate may differ in other populations.

Merging Results Across Graphs and Inferring New Segments

IBD segments can span multiple regions and our method analyzes IBD at a single region. The probability of IBD between two individuals can therefore be output at multiple adjacent regions by our method. Furthermore, the IBD probability may be assigned a different value in each region due to the inexact nature of the sampling method. If the same IBD segment is assigned different probabilities across multiple loci we use the maximum value across all regions.

Once an IBD graph is analyzed using the sampling procedure, edges that were previously missing (i.e. those that were not called by Refined IBD) are output with a start and stop site that is equal to the intersection of all IBD segment boundaries in the graph. Since we do not look in the region for sequence identity between haplotypes we can only output the probability that IBD exists somewhere within the region. These new segments may also overlap with other called IBD segments. In order to reconcile overlapping IBD segments, we merge them provided that they pass a probability threshold set by the user and that they lie on the same haplotype. As the final probability, we use the maximum $\hat{p}_{ij}$ of the merged segments. For all analyses presented here, we only merged segments that had a probability of 0.99 or greater.

Creating Simulated IBD Data

We generated simulated genotype data as previously described by Browning *et al.*(B. L. Browning & Browning, 2013). To start, we use Fastsimcoal(Excoffier & Foll, 2011) to generate phase known DNA sequence data of 2000 diploid individuals. A single individual is represented as one chromosome consisting of ten independent 30 MB regions, each with a mutation rate of 2.

5×10^{-8} and a recombination rate of 10^{-8}. The population simulated begins with an effective population size of 3000 diploid individuals with a growth rate of 1.8% at time t=300 (where t is the number of generations ago from the present). Moving forward in time, the growth rate was changed to 5% and to 25% at times t=50 and t=10 respectively, resulting in a final effective populations size of 24,000,000 at t=0. The simulation is reflective of European population sizes estimated from the linkage disequilibrium of common variants(Tenesa et al., 2007).

Using the DNA sequence data we create genotype data by first filtering single nucleotide polymorphisms (SNPs) that were not bi-allelic with a minor allele frequency (MAF) less than 2%. Next, we choose 10,000 variants uniformly by MAF (where $2\% \leq \text{MAF} \leq 50\%$) per 30 MB region. This SNP density is in line with that of a 1,000,000 SNP genotyping array. Finally, we remove all phase information and apply a genotyping error at a rate of .05% by turning heterozygous genotypes into homozygous genotypes and vice versa. Using the simulated genotype data, we use Refined IBD to phase the data and call pairwise IBD. We define true IBD segments as those segments longer than or equal to 0.1 centimorgan. A potential consequence of this approach to creating simulated data is that the resulting IBD graph may not completely obey transitivity.

Results

Convergence Properties and Runtime

We first verify that the conditional probabilities estimated from our sampling ap- proach, $\hat{p}_{ij}$, converge to the true edge conditional probabilities, $\dot{p}_{ij}$. We randomly create three to eight node probabilistic IBD graphs with edge probabilities drawn uniformly from the open interval (0, 0.99). For each graph, we enumerated every transitive IBD graph to compute the exact conditional edge probability $\dot{p}_{ij}$. It is computational infeasible to compute exact probabilities for graphs larger than 8 nodes since all transitive graphs must be enumerated. We then ran our sampling approach over each graph and at each iteration l, we calculated the average percent difference between $\hat{p}_{ij}$ and $\dot{p}_{ij}$, which we callδ_l.

$$\delta_l = \sum_{\forall i \neq j} \frac{|p_{ij} - \hat{p}_{ij}^l|}{p_{ij}}, \qquad \text{where } \hat{p}_{ij}^l = \text{conditional edge probablities at iteration } l$$

We ran PIGS 25 times and calculated δ_l^{25}. which is δ_l averaged over all 25 runs. From Figure 3 we see that for graphs with 3 to 7 nodes, edges are within 1% of true conditional probability after 5000 iterations. For 8 node graphs, the probabilities are within 15% of the true $\dot{p}_{ij}$ after 5000 iterations and within 5% within 7500 iterations. We recorded the average runtime of the 25 runs and show the results in (Table 1). While it is computationally feasible to sample until convergence for small graphs, this approach will not scale to genome-wide IBD studies of a large number of individuals. Instead PIGS takes as input a user specified time limit for sampling each region.

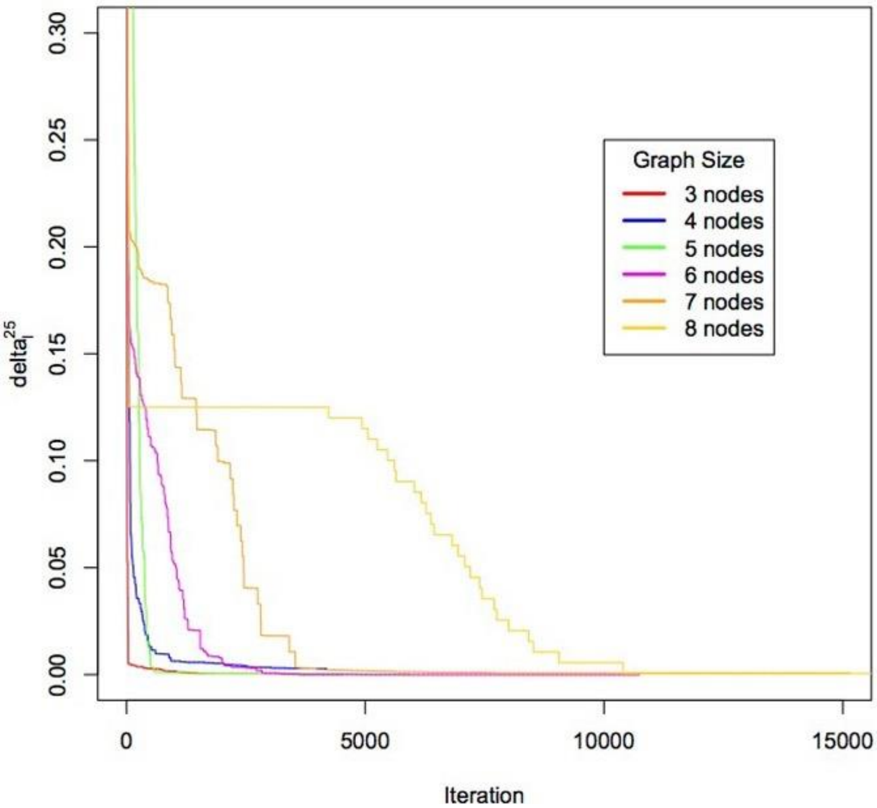


Figure 3 Iterations needed for convergence. On the x-Axis is the number of iterations and on the y-axis is the value of δ_l^{25} which is the average percentage edge delta over 25 runs.

Table 1 Average Runtime of different sized graphs over 25 iterations

Nodes	Time (CPU seconds)
3	1.34
4	2.79
5	6.31
6	17.33
7	46.99
8	80.19

Application to Simulated Data

Ultimately, the metrics of merit are the IBD calls themselves, not IBD probabilities. IBD calls can be made from IBD probabilities using a thresholding approach in which all probabilities exceeding a threshold are output as IBD. Alternatively, methods such as DASH(Gusev et al., 2011), EMI(Qian, Browning, & Browning, 2013), and IBD-Groupon(He, 2013) leverage the clique nature of IBD graphs to output cliques over a region as opposed to IBD pairs. The choice of IBD calling method is a function of the objective of the study. For example, DASH was designed specifically for association testing in which individuals in a clique are given a psuedo-genotype of 1 and all others are given a pseudo-genotype of 0. Other testing methods examine the distribution of IBD between cases and controls(S. R. Browning & Thompson, 2012; Han et al., 2013; Purcell et al., 2007) and rely on IBD calls that powerfully and accurately cover true IBD segments. For population genetics purposes such as inferring demographic history(Palamara et al., 2012), the distribution of IBD segments sizes is the figure of merit.

This diversity of uses of IBD precludes any single metric as being the gold standard for assessing the quality of IBD calls. Therefore, we compare several different methods of computing IBD probabilities and calling IBD over a range of metrics. We compare a thresholding approach to calling IBD applied to PIGS probabilities as well as Refined IBD LOD scores. We also examine the behavior of the clique-calling approaches DASH and EMI when applied to Refined IBD output and PIGS output. We attempted to include IBD-Groupon but in its current implementation some hard-coded parameters make it unsuitable for the sample sizes we examined here. This will be addressed in a future release (personal communication with Dan He). We created simulated genotype data on ten 30MB regions for 2000 individuals (see Creating simulated IBD data).

We generated IBD calls from Refined IBD by using a LOD threshold of 3. For PIGS, we first generated pairwise graphs from Refined IBD by using a LOD threshold of 0.1 and a segment length cutoff of 0.1 centimorgans. PIGS was then run over the pairwise graphs for a maximum of 2 minutes and IBD calls were made using a probability threshold of 0.99. IBD calls for Germline were generated using their suggested parameters "-haploid -bin out -min m 1 -bits 32 -err hom 1 -err het 1"(Gusev et al., 2009) after phasing genotype data using fastIBD(B. L. Browning & Browning, 2011). DASH was run over the Refined IBD calls passing a LOD threshold of 3. All results were filtered to have a minimum segment length of 0.5 centimorgans.

Identification of IBD Segments

For a given genomic locus, the power of tests comparing the distribution of IBD in cases or between cases and controls(S. R. Browning & Thompson, 2012; Han et al., 2013; Purcell et al., 2007), is a function of the number of true IBD segments intersected by predicted segments. We therefore performed an analysis of the total number of true IBD segments intersected by IBD calls from Refined IBD, Germline, and PIGS. The results shown in Figure 4a show that PIGS substantially outperforms Refined IBD for small IBD segments. DASH was not included in this analysis because it was not designed for this purpose and the resulting error rates were 10 fold higher than PIGS and Refined IBD even at 1 centimorgan segments. For predicted segments of size 0.5, 0.6, 0.7, 0.8, 0.9, and 1 centimorgans, there was an increase of 95%, 43%, 27%, 17%, 12%, and 9% in the number of predicted segments intersecting a true segment over Refined IBD. For predicted segments of size 1.1, 1.2, and 1.3 centimorgans, Germline was able to detect 60%, 27%, and 12% more segments than PIGS but the calls were less accurate.

In order to assess the error rate we examined the fraction of segments that did not intersect any true IBD segment. Note that this is error rate may be inflated due to the fact that true segments are required to be at least 0.1 centimorgans. The results shown in Figure 4b demonstrate that PIGS has nearly identical error rates to Refined IBD at small segments. However, at 0.5 centimorgans the error rate increases from 0.3% to 0.7%; this is a modest increase relative to the 95% increase in the number of segments identified. Germline, in general, was similar to the other methods in terms of error rates for segments between 1 and 2 centimorgans.

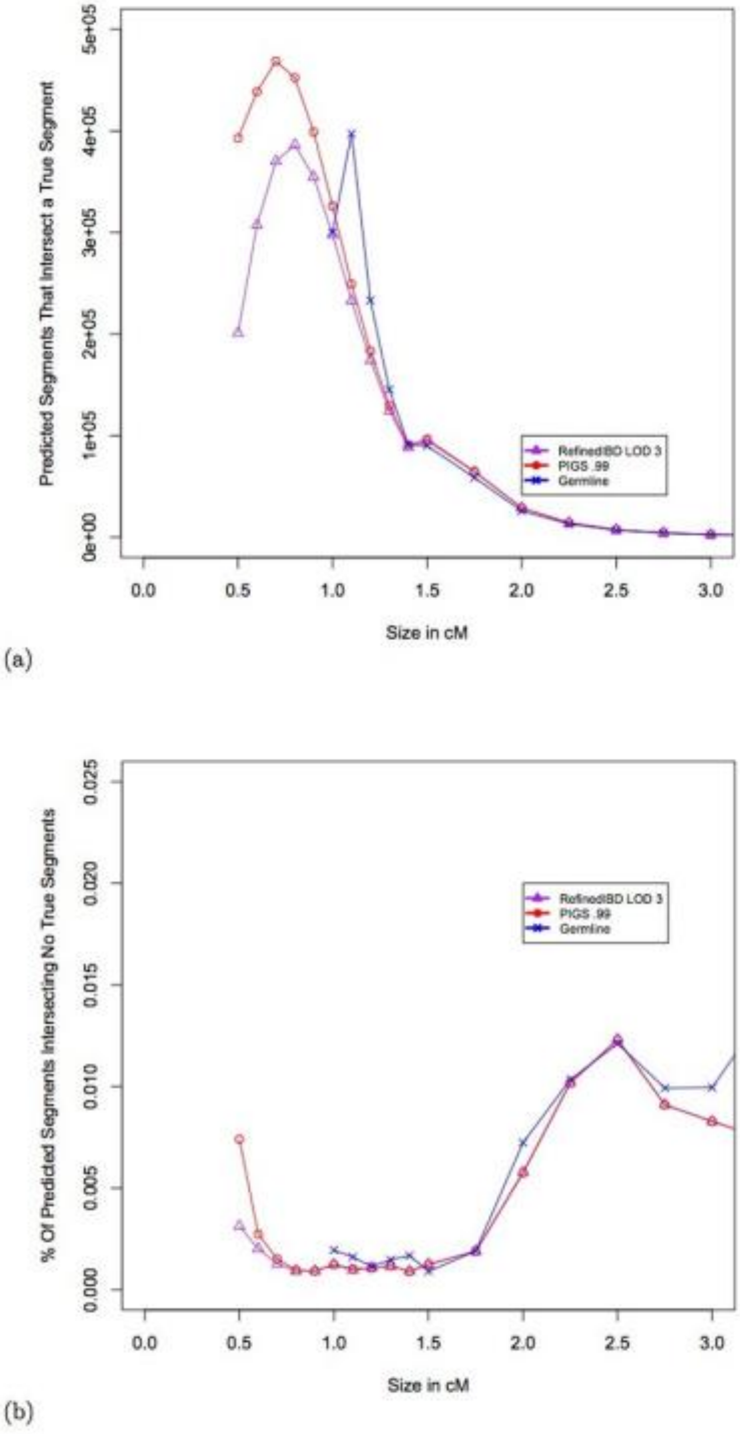


Figure 4 Power and error rates as a function of IBD segment counts. (a)Number of predicted segments overlapping a true IBD segment is shown on the y-axis. The x-axis shows the size of the predicted segment in centimorgans. (b)The percentage of predicted segments that have no overlap with a true segment is shown on the y-axis. The x-axis shows the size of the predicted segment in centimorgans.

Accuracy of IBD Segments

In population genetics settings, such as inferring demography(Palamara et al., 2012; Tenesa et al., 2007), methods often rely on the distribution of IBD segment lengths. The figure of merit here is related to the accuracy of predicted segments recovered. We first examined power, the average proportion of true IBD segments that were overlapped by predicted segments. For true segments between 0.5 and 2.5 centimorgans our method had modestly greater power (Figure 5a). This came at the expense of a slight increase in false discovery rate (FDR) as shown in Figure 5b. The false discovery rate is defined as the average proportion of predicted segments that does not overlap a true IBD segment. This is somewhat expected since new segments from PIGS use existing IBD segment boundaries to approximate the new start and stop sites. The greatest decrease comes at 0.5 centimorgans where PIGS predicts 95% more segments than Refined IBD. However, the difference in FDR between Refined IBD and PIGS is still less than 5% (10% versus 14%). On the other hand, for segments of size between 0.6 and 1.5 centimorgans, PIGS predicts 23% more segments while keeping the FDR within 1% of Refined IBD.

We also examined the true positive rate, defined as the percentage of predicted segments with at least 50% overlap with a true segment. Compared to Refined IBD the true positive rate for PIGS drops slightly for segments that are smaller than 1 centimorgan but the difference is less than 1% for all sizes except for at 0.5 centimorgans where it is 3% (Figure 6). The reason for this drop in performance is at least partly due to the fact that we add new segments according to the IBD graph without specifically examining the sequence. Given the results of the previous section, the most likely explanation is that the PIGS predicted segments intersect true IBD segments, but not at the 50% threshold required by definition of a true positive. Based on these results PIGS could be used for population genetics purposes, but users should take into account the slight increase in error rates for smaller segment sizes.

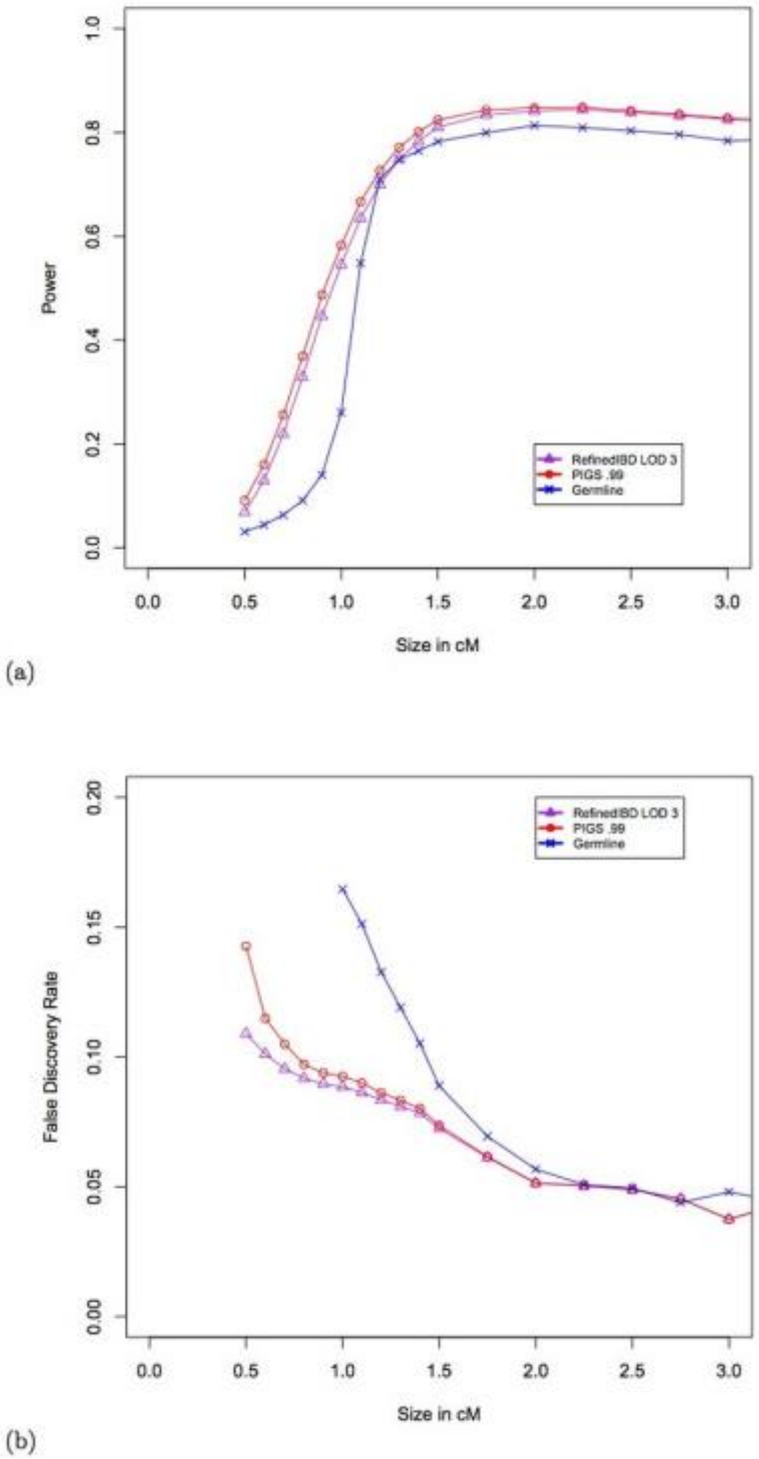


Figure 5 Power and error rates as a function of IBD segment proportions. (a)The average proportion of a true segment detected is shown on the y-axis. The x-axis shows the size of the true segment in centimorgans. (b)The average proportion of predicted segments that do not intersect any true IBD segment is shown on the y-axis. The x-axis shows the size of the predicted segment in centimorgans.

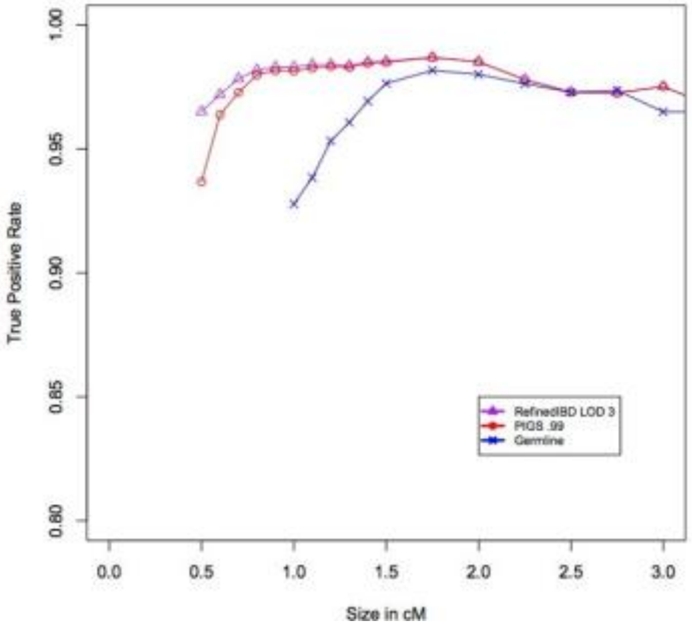


Figure 6 True positive rate. The percentage of predicted segments that overlaps at least 50% with a true IBD segment is shown on the y-axis. The x-axis shows the size of the predicted IBD segment in centimorgans.

Identification of Cliques

In a genome wide association study (GWAS) association tests are typically per- formed on individual single nucleotide polymorphisms. Cliques of IBD segments can also be leveraged to increase power in association tests(S. R. Browning & Thompson, 2012; Gusev et al., 2011). In this setting IBD serves as a representation of SNPs not contained on the genotyping platform, and the figure of merit is how well the true IBD cliques are captured by predicted IBD segments. For 10000 random positions along the genome, we created predicted IBD graphs for Refined IBD, PIGS, P-DASH, P-EMI (DASH and EMI using PIGS as input), R-DASH, and R-EMI (DASH and EMI using Refined IBD as input). For Refined IBD and PIGS, all segments of size 0.5 centimorgans or greater were used if they passed a LOD threshold of 3 and probability threshold of 0.99 respectively. DASH and EMI are both algorithms that create cliques in a given window. DASH starts with the biggest connected component and creates dense subgraphs by cutting out false-positive edges. EMI on the other hand starts with seed subgraphs and adds edges that it believes to be true IBD. For DASH we used default parameters "- win 500000 -density 0.6 -r2 0.85 -min 4". For EMI we used the parameters "-win bp 200000 -den 0.6 -min 3 -wgt bp 100000 1000000". However with these EMI parameters, R-EMI had an error rate 3 to 8 times greater than the other methods depending on the size of the clique. Instead we used the weight parameter "-wgt 7th 3 40" for Refined IBD input to EMI which uses the LOD score instead of the length to weight the edges and improved performance.

At each position we examined cliques in the true graph that overlapped with a connected component in the predicted graph for any method. The true graph was generated with all true IBD calls regardless of size and all connected components were converted to cliques. Table 2 shows the power of each method to detect an edge of a true clique of a given size. The power here is defined as the average proportion of edges in a true clique that are called correctly by a given method. This is not the power to recover an entire clique, but an estimate of the number of edges in a clique that are recovered. At any clique size, PIGS detects a higher proportion of edges than Refined IBD. For P-DASH and R-DASH the power of both methods are very similar with P-DASH only showing a very modest increase in power depending on the clique size. However, when comparing P-EMI and R- EMI we see 2-5% increases in power for P-EMI. To verify that the gain in power for P-DASH and P-EMI was due to PIGS and not due to leveraging clique information twice, we used EMI and DASH output as input into a second round of EMI and DASH. We observed virtually no change in power or error rate showing that PIGS is providing the increase in performance. All methods lacked power when considering very large cliques and this is most likely due to the fact that very large cliques are generated from small segments of IBD (i.e. < 0.5 centimorgan).

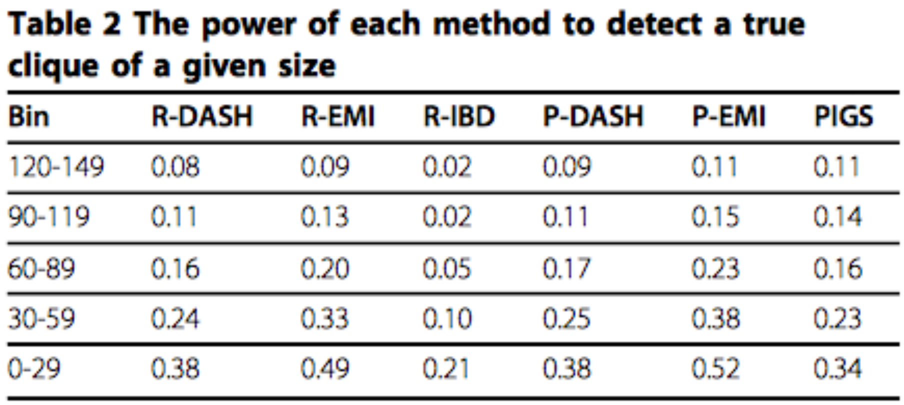


Table 2 The power of each method to detect a true clique of a given size

Bin	R-DASH	R-EMI	R-IBD	P-DASH	P-EMI	PIGS
120-149	0.08	0.09	0.02	0.09	0.11	0.11
90-119	0.11	0.13	0.02	0.11	0.15	0.14
60-89	0.16	0.20	0.05	0.17	0.23	0.16
30-59	0.24	0.33	0.10	0.25	0.38	0.23
0-29	0.38	0.49	0.21	0.38	0.52	0.34

We also assessed the false positive rate of each clique-based method. The false positive rate was defined as the average proportion of predicted edges that are not part of a true clique. Table 3 shows the false positive rate of each method for a given size of a predicted clique. The false positive rate of PIGS is slightly higher than Refined IBD for most clique sizes, but the increase in false positive rate is modest (within 2%) for all clique sizes. As was the case with power, the error rates of P-DASH and R-DASH are nearly identical. The biggest difference is for cliques with 60-89 nodes, where P-DASH has a 4% higher false positive rate. We see similar behavior for R-EMI and P-EMI, where for cliques of size 90-119 nodes, the error rate goes from 5% to 9%. Based on these results we recommend using EMI to perform clique calling on PIGS output as it provides lower error rates and higher power than DASH.

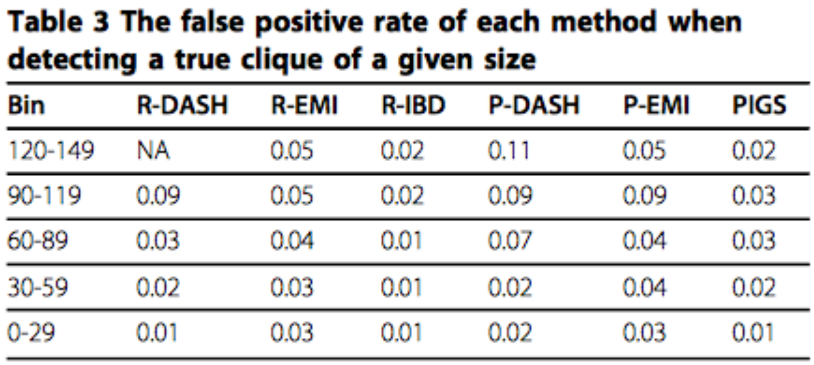


Table 3 The false positive rate of each method when detecting a true clique of a given size

Bin	R-DASH	R-EMI	R-IBD	P-DASH	P-EMI	PIGS
120-149	NA	0.05	0.02	0.11	0.05	0.02
90-119	0.09	0.05	0.02	0.09	0.09	0.03
60-89	0.03	0.04	0.01	0.07	0.04	0.03
30-59	0.02	0.03	0.01	0.02	0.04	0.02
0-29	0.01	0.03	0.01	0.02	0.03	0.01

Application To Real Data

Identification of IBD Segments

We applied PIGS, RefinedIBD, DASH, and EMI to 489 Latino trios from the Genetics of Asthma in Latino Americans (GALA) cohort(Burchard et al., 2004). The availability of trio genotype data allows us to phase the genotypes with high accuracy by taking into account the rules of Mendelian segregation. The increased phasing accuracy in turn boosts the power to detect IBD segments because phasing errors are a major source of difficulty in calling IBD(B. L. Browning & Browning, 2013). To evaluate how well a given method is able to identify segments of IBD in real data we used IBD segment calls made using Refined IBD in a trio-aware mode (trio-IBD segments). Trio-IBD segments were thresholded at a length of 0.1 centimorgans and at a LOD score of 3. We then asked how many IBD calls made by a given method without access to the near-perfect trio phasing overlapped with trio-IBD segments. DASH and EMI were run in the same way as we described for identifying cliques (using PIGS and RefinedIBD as input), how- ever the resulting clique edges were converted into IBD calls and merged with the original input. We include the clique calls here as IBD calls because we do not know the structure of the real IBD graph. All IBD calls were thresholded at a segment size of 0.5 centimorgans.

As shown in Figure 7, when considering PIGS and Refined IBD calls at 0.5 centimorgans, there is an increase of 10% in the number of segments identified by PIGS over Refined IBD. After applying DASH and EMI to the input of both methods we see an increase of 8% and 7%, respectively, for PIGS input. It is clear that both DASH and EMI improve the power of both main approaches to detect IBD for use in association studies regardless of the segment size. DASH and EMI seem to per- form similarly in terms of boosting power when called segments are bigger than 0.8 centimorgans, but EMI appears to have the upper hand for anything smaller. For example, at 0.5 centimorgans the difference between EMI and DASH for PIGS input is 8% but at 0.8 centimorgans the difference is only 0.8%. Across all segment sizes, we see increases of 4%, 3%, and 2.5% for PIGS, P-DASH, and P-EMI over their Refined IBD counterparts. The increases are more modest than in the simulated data, most likely due to the fact that without sequencing data we are underpowered to detect small segments of IBD even when trio phased genotypes are available.

PIGS and Refined IBD called 3134591 and 2968480 segments that overlapped with at least one of trio-IBD segment, respectively, which equates to a 6% increase. Similar increases are also seen when using PIGS input to DASH and EMI, with a 4% increase (3177234 versus 3047734) for DASH and a 3% increase (3263594 versus 3158818) for EMI. 1330207 PIGS and 803527 Refined IBD calls did not overlap with any trio-IBD segments. Because we only have access to true positives in the real data, there is no perfect way to determine the false positive rate of any of these methods, and it could be argued that PIGS increases the power to detect IBD at the expense of a higher false positive rate. To determine if this was the case, we made random IBD calls along the genome. As an example consider the calls of size 0.5 centimorgans, where PIGS made 349221 calls and Refined IBD made 262064 calls, a difference of 87157 calls. Of these, 207122 PIGS and 187582 Refined IBD segments overlap a trio-IBD segment, which is an increase of 19540 segments. After making 87157 random calls (of length 0.5 centimorgans) we only identified 453 segments compared to the 19540 we observed originally. This means that even if Refined IBD were to make an additional 87157 random guesses along the genome we would not expect Refined IBD to have the same power as PIGS, showing the increase in performance is not entirely due to false positives. Furthermore, if we assumed all non-overlapping segments are false positives both Refined IBD and PIGS have an error rate over

20%, which is not reflective of the simulations where the error rate for both methods was below 1% (see Figure 4b)(B. L. Browning & Browning, 2013).

Given these results, we conclude that the increased performance in PIGS was not driven by the extra IBD calls and that the majority of the non-overlapping segments are indeed true as suggested by the simulation results. Assuming that the true false positive rate in real data is similar to simulation data, the large increase in predicted segments that overlap trio-IBD segments when using PIGS (with or without a clique calling method) shows the potential for substantial power increases when using PIGS for IBD mapping studies.

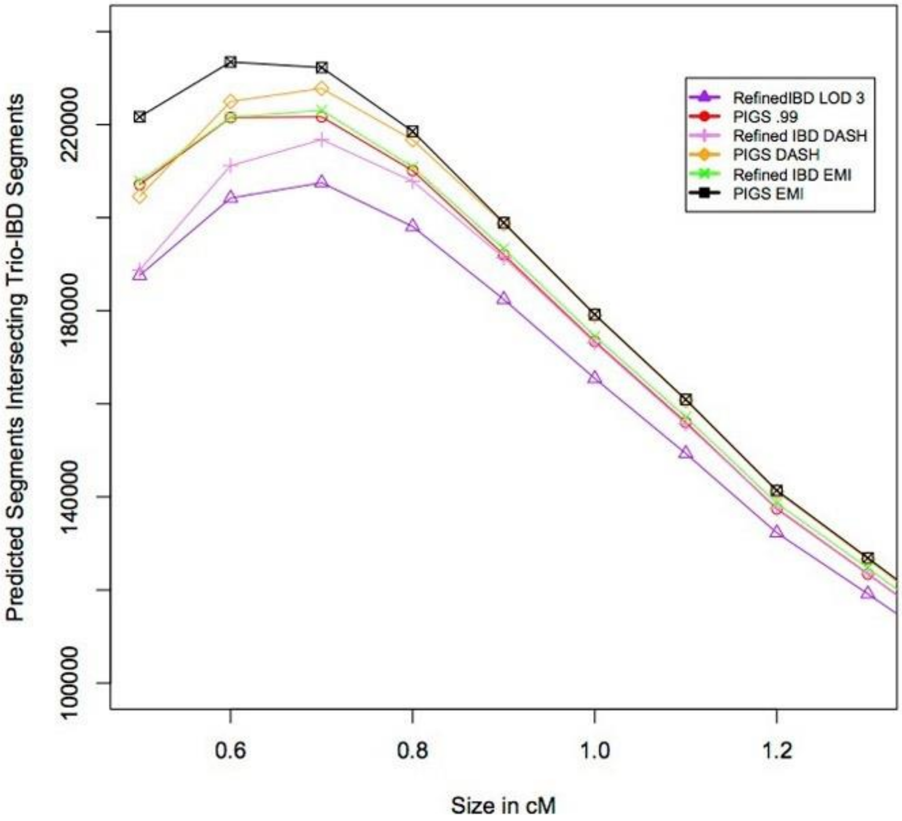


Figure 7 Number of trio-IBD segments identified. The number of predicted segments that overlaps a trio-IBD segment is shown on the y-axis. The x-axis shows the size of the predicted IBD segment in centimorgans.

Conclusion

We have developed a new efficient method (PIGS) for simultaneously computing the probability of IBD between multiple haplotypes at a genomic region. PIGS combines the computational efficiency of pairwise methods with the power advantages of multiway methods. We demonstrated that PIGS converges to the correct probabilities of conditional IBD probabilities for small IBD graphs. For IBD graphs with both small and large numbers of individuals we showed that the approximate probabilities from PIGS produce a substantial improvement in the power to identify small IBD segments and recover IBD edges from cliques relative to previous approaches.

PIGS relies on accurate pairwise probabilities in order to compute conditional probabilities. In this work we scaled the probabilities according to the results of simulated segments of IBD. This has been the approach of previous methods, as there is currently no mechanism for assessing true probabilities in real data. This approach is not guaranteed to be accurate for all populations(B. L. Browning & Browning, 2013; Gusev et al., 2011; He, 2013; Moltke et al., 2011; Purcell et al., 2007). If the demographic history of the population of interest is substantially different from the one simulated here, additional simulations should be done to assess the relationship between LOD scores and probability of IBD.

In some scenarios, such as the inference of demographic history(Palamara et al., 2012), the metric of merit is not the power to identify segments, but the accuracy of the distribution of IBD segment lengths. Because PIGS does not currently utilize genotype or sequence data to refine newly identified IBD segments it is not as accurate as Refined IBD for small segments. One possible future approach is to use powerful, but computationally expensive multiway IBD calling methods such as the MCMC proposed by Moltke *et al.*(Moltke et al., 2011) to examine the new regions identified from PIGS.

In our analysis here we restricted our analysis to segments that were at least 0.5 centimorgans in size. There may be IBD segments that are much smaller in size (<< 1 centimorgan) and methods such as HapFABIA(Hochreiter, 2013) are able to identify these small segments. HapFABIA uses an efficient bi-clustering approach but relies on the existence of rare variation in the data. Given that we did not have sequencing data available, we did not compare our method to HapFABIA. However in the future, sequencing data will be more readily available and we hope to see how if PIGS can also be leveraged to improve the power existing methods.

Clique-calling methods such as DASH, EMI, and IBD-Groupon use IBD probabilities such as those output from PIGS and Refined IBD to identify cliques of IBD segments. Clique-calling methods are typically used to increase the power of IBD mapping studies. We showed that these methods can substantially increase the power to detect the edges of IBD graphs. The exact relationship between the power of a given IBD mapping approach and the number of edges discovered remains to be shown. Going forward, having a better grasp of how power and false positive rates of predicted graphs affect IBD mapping methods will be important to maximize the utility of clique based mapping approaches.

The current sampling scheme for PIGS was selected for its performance in identifying IBD segments. There are many different methods of exploring the space of transitive graphs. Our focus in this work was medical genetics, but alternative sampling schemes could be explored to optimize segment accuracy instead of power to detect segments. Given the substantial improvement in the number of identified IBD segments of our method, we expect that PIGS will facilitate improvements in IBD based disease association studies and provide new inroads into identifying small segments of IBD.

2. Using Machine Learning to Create Combined Reference for Summary Statistic Based Methods in Admixed Populations

Introduction

Summary statistics of association tests, such as effect size estimates and their standard errors, are becoming the data type of choice in many genetic analyses due to two significant advantages. First, summary statistics based methods are generally orders of magnitude faster than their genotype based counterparts. The rapidly increasing size of existing and planned cohorts is causing computational bottlenecks for some standard analyses. Second, analyses of summary statistics are often a necessity since access to individual-level data is complicated by privacy and other issues(Gymrek, McGuire, Golan, Halperin, & Erlich, 2013). Publication of summary statistics is now required for all Nature Genetics genome wide association study (GWAS) papers, and these statistics have already been released for a large number of traits. For these reasons a growing number of summary statistics based methods, including imputation of z-scores, joint-testing, fine mapping of causal variants, quality control of GWAS results, and gene based tests, have recently been published(Bulik-Sullivan et al., 2014; Han, Hackel, & Eskin, 2011; Hormozdiari, Kostem, Kang, Pasaniuc, & Eskin, 2014; Kichaev et al., 2014; Liu et al., 2010; Pasaniuc, Zaitlen, Shi, Bhatia, Gusev, Pickrell, Hirschhorn, Strachan, Patterson, & Price, 2014; Yang et al., 2012). Moving forward, the integration of summary statistics will be vital for increasing our knowledge of various complex diseases and phenotypes(Schork et al., 2013).

Summary statistics based methods typically require estimates of linkage-disequilibrium (LD) between markers as input. Existing tools use “best guess” reference panels to estimate LD. For example, Yang *et al.*(Yang et al., 2012) used European ancestry individuals from the Queensland Institute of Medical Research reference panel to estimate LD for an analysis of statistics produced from the European ancestry GIANT consortium(Speliotes et al., 2010). This approach is not optimal and has the potential to produce misleading results in the case of admixed populations. Admixed individuals' genomes can be viewed as mosaics, where different segments of the genome are derived from various ancestral groups. Previous work has shown that the proportions of ancestry for individuals from admixed populations are highly variable(Bryc et al., 2010; Silva-Zolezzi et al., 2009; Wang et al., 2008). Given this high variability in admixed populations, “best guess" panels are more likely to have LD estimates that are not in concordance with original datasets, and which vary in their local structure. This will be especially true if the population of interest has no reference panel available. Furthermore, several genotype based methods have shown that learning local structure from multi-population reference panels improves performance even in the case of homogenous study populations(Howie, Donnelly, & Marchini, 2009; Pasaniuc et al., 2013).

In this work, we develop a method, Adapt-Mix, to accurately estimate the local SNP correlation matrix for each region of the genome from summary statistics of an arbitrary population study. We compute the correlation matrix using a mixture of existing reference panels, such as the 1000 genomes(T. 1. G. P. Consortium, 2012), where the mixture proportion for each reference population is learned from summary statistics. Unlike previous approaches, our method incorporates data from multiple reference panels when computing the correlation matrix and allows for adaptation to local structure. We first provide a closed form solution for the expected correlation structure from a mixture of populations in a genomic locus. Then, using this derivation, we efficiently search for the mixture of populations in each genomic locus that maximizes/minimizes an objective function most relevant to the problem in question. For example, in this work we consider the problems of imputation and joint-testing from summary statistics, using imputation error and joint-test accuracy as the objective function, respectively. In practice, arbitrary objective functions can be used provided they can be computed efficiently.

We apply our method to summary statistics from simulated phenotypes over real genotypes from the Genes-environments & Admixture in Latino Americans (GALA II, (Borrell et al., 2013)) cohort that is composed of Mexican and Puerto Rican individuals. We also apply our method to real coronary artery disease summary statistics from the CARDIoGRAMplusC4D consortium(T. C. A. D. C. G. Consortium, 2011; Schunkert et al., 2011). In the simulated data sets we show significant improvements in the mean-squared error of our mixture correlation coefficients compared to the most relevant reference panels. We also demonstrate the direct impact of the improved correlation estimates for imputation and joint-testing methods, which take correlation matrices as input. For both the simulated summary statistics over the GALA II study as well as the meta-analysis results we show significant improvement in both summary statistics based imputation and joint-testing(Pasaniuc, Zaitlen, Shi, Bhatia, Gusev, Pickrell, Hirschhorn, Strachan, Patterson, & Price, 2014; Yang et al., 2012).

Methods

First, we describe the situation where Adapt-Mix may be applied. We then derive a formula for the genotype correlation matrix as a mixture of several reference populations, and describe our procedure for optimizing the mixture frequencies for various objective functions. We end the section by discussing the simulation framework in which we evaluate our method.

GWAS summary statistics typically consist of an effect size β_i and standard error σ_i for each SNP *i* examined in a study. For simplicity, β_i and σ_i can be converted to a Wald test statistic (*Z*-score) z_i. When dealing with case-control phenotypes $z_i = \sqrt{N}\,\frac{p_i^+ - p_i^-}{\sqrt{2p_i(1-p_i)}}$, where *N* is the sample size, p_i^+ (p_i^-) is the frequency of the reference allele in cases (controls), and p_i is the overall frequency. For quantitative phenotypes $z_i = \sqrt{N} * cor(\vec{g}_i, \vec{q})$, where $\vec{g}_i$ are the genotypes of the individuals and $\vec{q}$ are the phenotypes. Here, $\vec{g}_i = \{g_{i1} \dots g_{iN}\}$ for $g_{id} \in \{0,1,2\}$, g_{id} being the count of the reference allele for individual d.

As input, most summary statistics based methods take Z-scores and a correlation matrix Σ(Bulik-Sullivan et al., 2014; Han et al., 2011; Hormozdiari et al., 2014; Kichaev et al., 2014; Liu et al., 2010; Pasaniuc, Zaitlen, Shi, Bhatia, Gusev, Pickrell, Hirschhorn, Strachan, Patterson, & Price, 2014; Yang et al., 2012). For each pair of SNPs *i, j* the correlation matrix has the value $\Sigma_{ij} = r_{ij}$, where r_{ij} is the Pearson correlation coefficient between the SNPs in the study. If individual level genotypes are available, the correlation can be computed by $r_{ij} = cor(\vec{g}_i, \vec{g}_j)$. When individual level genotypes are unavailable, r_{ij} is typically estimated using a reference panel of genotypes from a population similar to the source population of the data being analyzed. In this work we develop a method to provide a better estimate of r_{ij} using a combination of reference panels from different populations. Given a set of *K* reference populations, we generate a correlation matrix for each genomic locus using a new mixture population, where the frequency of population $k \in K$ in the mixture population is f_k. The objective of our work is to select the frequencies, f_k, that optimizes the performance of the summary statistics method of interest.

Estimating the Mixture Correlation Matrix

Given a set of mixture frequencies, $\vec{f} = \{f_1, \dots, f_K\}$, where $f_k \in \vec{f}$ is the frequency for population $k \in K$. We wish to compute the expected correlation between each pair of SNPs in the mixture population. For simplicity, we begin by deriving the mixture variance of the allele frequencies (σ_i^2) at SNP i, in a mixture population composed of 2 reference populations. At SNP i, the 2 reference populations will have separate variances (σ_{1i}^2, σ_{2i}^2), sample sizes (n_1, n_2), and allele frequencies (p_{1i}, p_{2i}).

Additionally, assume that each reference population has a mixture frequency equal to their proportion of sample size, i.e. $f_1 = \frac{n_1}{\square_1 + n_2}$ and $f_2 = \frac{n_2}{n_1 + n_2}$. We can then express the mixture variance as

$$\sigma_i^2 = \frac{\sum_{z=1}^{n_1}(g_{1z} - 2p_i)^2 + \sum_{q=1}^{n_2}\left(g_{2q} - 2p_i\right)^2}{(n_1 + n_2)}$$

where g_{kd} is the genotype of individual *d* in population *k*, and $2p_i = f_1 2p_{1i} + f_2 2p_{2i}$ is the genotype frequency in the mixture population. Let us now consider only $\sum_{z=1}^{n_1}(g_{1z} - 2p_i)^2$. This term is equal to

$$\sum_{z=1}^{n_1}(g_{1z} - 2p_i)^2 = \sum_{z=1}^{n_1}[(g_{1z} - 2p_{1i}) + (2p_{1i} - 2p_i)]^2 = \sum_{z=1}^{n_1}(g_{1z} - 2p_{1i})^2 + n_1(2p_{1i} - 2p_i)^2$$

$$= n_1\sigma_{1i}^2 + n_1(2p_{1i} - 2p_i)^2$$

Applying the same logic to $\sum_{q=1}^{n_2}\left(g_{2q} - 2p_i\right)^2$we arrive at the formula for the variance for the mixture population.

$$\sigma_i^2 = \frac{n_1\sigma_{1i}^2 + n_2\sigma_{2i}^2}{n_1 + n_2} + \frac{n_1(2p_{1i} - 2p_i)^2 + n_2(2p_{2i} - 2p_i)^2}{n_1 + n_2}$$

$$= f_1\sigma_{1i}^2 + f_1(2p_{1i} - 2p_i)^2 + f_2\sigma_{2i}^2 + f_2(2p_{2i} - 2p_i)^2$$

We now extend from 2 to *K* populations. Suppose we have a set of reference panels representing *K* populations and their corresponding mixture frequencies, $\vec{f}$. Then for SNP *i* in population $k \in K$, let σ_{ki}^2 be the variance and $2p_{ki}$ be the frequency. The frequency in the mixture population is then $2p_i = \sum_{k=1}^{K} f_k 2p_{ki}$, and the combined variance at SNP *i* is

$$\sigma_i^2 = f_1 \sum_{z=1}^{n_1} (g_{1z} - 2p_i)^2 + \cdots + f_K \sum_{l=1}^{n_K} (g_{Kl} - 2p_i)^2$$

$$= f_1\sigma_{1i}^2 + f_1(2p_{1i} - 2p_i)^2 + \cdots + f_K\sigma_{Ki}^2 + f_K(2p_{Ki} - 2p_i)^2$$

$$= \sum_{k=1}^{K} f_k[\sigma_{ki}^2 + 4(p_{ki} - p_i)^2]$$

Next, we derive the covariance between SNPs *i* and *j* in the mixture population. If *x* and *y* are random variables, $\sigma_{x+y}^2 = cov(x+y, x+y) = cov(x,x) + cov(y,y) + cov(x,y) = \sigma_x^2 + \sigma_y^2 + 2cov(x,y)$, and thus $cov(x,y) = \frac{\sigma_{x+y}^2 - \sigma_x^2 - \sigma_y^2}{2}$.

Let $cov_k(i,j)$ be the covariance of SNPs *i* and *j* in population *k*. Then the covariance in the mixture population is:

$$2cov(i,j)_{i\neq j} = \left[\sigma_{i+j}^2 - \sigma_i^2 - \sigma_j^2\right]$$

$$= \sum_{k=1}^{K} f_k \left\{\left[\sigma_{ki}^2 + \sigma_{kj}^2 + 2cov_k(i,j) + 4\left(p_{k(i+j)} - p_{(i+j)}\right)^2\right]\right.$$

$$\left. - [\sigma_{ki}^2 - 4(p_{ki} - p_i)^2] - \left[\sigma_{kj}^2 - 4\left(p_{kj} - p_j\right)^2\right]\right\}$$

$$= \sum_{k=1}^{K} f_k \left\{\left[\sigma_{ki}^2 + \sigma_{kj}^2 + 2cov_k(i,j) + 4\left((p_{ki} - p_i) + \left(p_{kj} - p_j\right)\right)^2\right]\right.$$

$$\left. - [\sigma_{ki}^2 - 4(p_{ki} - p_i)^2] - \left[\sigma_{kj}^2 - 4\left(p_{kj} - p_j\right)^2\right]\right\}$$

$$\Rightarrow cov(i,j)_{i\neq j} = \sum_{k=1}^{K} f_k\left[cov_k(i,j) - 4(p_{ki} - p_i)\left(p_{kj} - p_j\right)\right]$$

By definition, the mixture correlation matrix is

$$\sum\nolimits_{ij} = \frac{cov(i,j)}{\sqrt{\sigma_i^2\sigma_j^2}}$$

Algorithm 1 details our procedure for computing the mixture correlation matrix over a set of SNPs. Given K populations and M SNPs, it takes as input the mixture frequencies ($\vec{f}$), a matrix of SNP variances ($\mathbf{V}_{K\mathrm{x}M} = \{\sigma_{ki}^2\}$), a matrix of the pairwise SNP covariances ($\mathbf{C}_{K\mathrm{x}M\mathrm{x}M} = \{cov_k(i,j)\}$), and a matrix of the genotype frequencies ($\mathbf{P}_{K\mathrm{x}M} = \{2p_{ki}\}$), and outputs the mixture correlation matrix.

```
Algorithm 1 Graph sampling
Input: $\vec{f}$, $\mathbf{V}$, $\mathbf{C}$, $\mathbf{P}$
Output: $\Sigma$

# Normalize mixture frequencies so they sum to 1
$\vec{f} = \vec{f}/sum(\vec{f})$

# Compute weights for mixture variances
$WeightedGT = \mathbf{P}\,\vec{f}^T$
$NegWeightedGT = \mathbf{P}\,\left[\left(\vec{f}-1\right)^T\right]$
$D$ = empty $K$ x $M$ matrix
for all $k$ in $\{1\ldots K\}$ do
    $D_k = NegWeightedGT_k + sum(WeightedGT_l), \forall l \neq k$

# Compute mixture variances
$MixVar = (D^2 + \mathbf{V})$

# Compute mixture covariances
$MixCov$ = empty $K$ x $M$ x $M$ matrix
for all $k$ in $\{1\ldots K\}$ do
    $tmp = f_k(\mathbf{C}_k + [D_k \otimes D_k])$
    $MixCov = \mathbf{C} + tmp$

# Compute mixture correlations
$denominators = \sqrt{MixVar \otimes MixVar}$, square-root applied element-wise
$\Sigma = MixCov/denominators$, element-wise division
```

Optimization of Mixture Frequencies

Given this algorithm for computing the correlation matrix Σ of the mixture population over a set of SNPs, we turn to the problem of selecting the mixture frequencies $\vec{f}$. We formulate this as a constrained optimization problem: minimizing (or maximizing) the value of a given objective function subject to the constraint that $\sum f = 1$ using the L-BFGS algorithm(Byrd, Lu, Nocedal, & Zhu, 2006). In this context, the "best guess" approach corresponds to setting $f_k = 1$ for the guessed population and $f_j = 0, \forall j \neq k$. In this work we consider the problems of imputation and joint-testing from summary statistics, and therefore selected the mean squared error (MSE) of imputed z-scores at observed SNPs and MSE of computed joint-test statistics as our objective functions, respectively (see the Imputation and Joint-testing sections below). However, other objective functions may be more appropriate depending on the purpose of the summary statistics based method. For example, one could chose to maximize the likelihood of the observed z-scores $\vec{Z}$ under a multivariate normal distribution.

In order to allow for variation in local correlation structure, the genome is separated into W equally sized non-overlapping windows. For each window, $w \in \{1 \ldots W\}$, we compute the correlation matrix using only SNPs in w, Σ_w. Using Σ_w, z-scores are imputed for all SNPs in w and the imputed values are used to compute the mean squared error from the true z-scores. We exclude SNPs from Σ_w with a minor allele frequency (MAF) less than 0.01 in any of the k populations, missing z-scores, $r^2 \leq 0.003$, or an undefined r with the SNP we are imputing. These SNPs are excluded because they only add noise to the imputation process. To ensure that Σ is invertible, λ is added to the diagonal of the matrix. The final correlation matrix is then $\Sigma = \Sigma^{unadj} + \lambda I$. Σ^{unadj} is the original correlation matrix prior to adding λ. The exact algorithm to compute the imputation mean squared error for a set of SNPs in a window is described in Algorithm 2.

Algorithm 2 Mean squared error objective function

```
Input: $\vec{f}$, $\mathbf{V}$, $\mathbf{C}$, $\mathbf{P}$, windowSize, $\lambda, \vec{Z}$
Output: meanSquaredError

# Normalize mixture frequencies so they sum to 1
$\vec{f} = \vec{f}/sum(\vec{f})$

# Compute number of windows
windows = length($\vec{Z}$)/windowSize

# Initialize numerator and denominator of MSE
numerator = 0
denominator = 0
for all $q \in \{1 \dots windows\}$ do
        # Compute Sigma using SNPs in window q
        $\Sigma^{(q)} = Create\ \Sigma(\vec{f}, \mathbf{V}^{(q)}, \mathbf{C}^{(q)}, \mathbf{P}^{(q)})$, see Algorithm 1
        $\Sigma^{(q)} = \Sigma^{(q)} + \lambda I$

        # Impute SNPs in window
        for all $s \in \{1 \dots windowSize\}$ do
                $z_s = \Sigma_{st}^{(q)} \left[\Sigma_{tt}^{(q)}\right]^{-1} \vec{Z}_t^{(q)}, \forall t \neq s$
                numerator = numerator + $\left(z_s - \vec{Z}_s^{(q)}\right)^2$
                denominatoy = denominator + 1

meanSquaredError = numerator/denominator
```

The procedure we have described is easily extendable from a window to any region, be it a whole genome, chromosome, or single locus. In this case, $\vec{f}$ is optimized by minimizing/maximizing the objective function over the sum of the non-overlapping windows. If there are a large number of SNPs in the region of interest, the convergence time of the algorithm will increase. To minimize the computation time when optimizing over the entire genome, we selected regions of the genome that have the largest absolute z-scores. Specifically, for every set of five adjacent windows we optimized using the two windows with the largest number of z-scores > 1.5.

Imputation

The z-score at a SNP i can be imputed from summary statistics and the correlation matrix, Σ, using the ImpG approach(Pasaniuc, Zaitlen, Shi, Bhatia, Gusev, Pickrell, Hirschhorn, Strachan, Patterson, & Price, 2014). Pasaniuc *et al.* used a gaussian approximation combined with a windowing approach to impute the z-score at i. The windowing aims to decrease runtime and reduce statistical noise that might be caused by distant SNPs with random non-zero correlation but no true LD. Define $\vec{Z}_t$ as the set of observed z-scores within a given window size around *i*. The imputed z-score is then $z_i = \Sigma_{it}^{(q)} \left[\Sigma_{tt}^{(q)}\right]^{-1} \vec{Z}_t$, for all SNPs *t* in the window.

Joint Testing

At genomic loci where at two SNPs are negatively correlated, using a marginal test often underestimates effect sizes(Galarneau et al., 2010; Sanna et al., 2011; Yang et al., 2012). A joint analysis is more powerful than a marginal test when analyzing such SNPs. Given two z-scores computed at SNPs i and j using a marginal test, a χ^2 test-statistic with 2 degrees of freedom, J_{ij} can be calculated as shown in Equation 3.

$$J_{ij} = \frac{1}{1-\Sigma_{ij}^2}\left(z_i^2 + z_j^2 - 2\Sigma_{ij}\mathrm{z}_i\mathrm{z}_j\right)$$

In our tests, the calculation of J_{ij} is restricted to SNPs that have a pairwise correlation $|r| < 0.8$ because small changes in *r* can cause large fluctuations in J_{ij} as $|r|$ approaches 1.

Simulation Framework

We simulated data using individuals from the Genes-environments & Admixture in Latino Americans (GALA II) cohort(Borrell et al., 2013), which is composed of 1245 Mexican and 1785 Puerto Rican individuals. The Mexican individuals have predominantly European and

Native American ancestry whereas their Puerto Rican counterparts tend to have mostly European and African ancestry. We conducted separate simulations for each group due to the differences in ancestry. We generated quantitative phenotypes and z-scores for every non-overlapping window of 1000 SNPs. For each window, a binomial trial ($p = 0.01$) was used to determine if the phenotype should be drawn from the null or alternate. Under the null, individuals' phenotypes were drawn from a $\mathcal{N}(0,1)$. Under the alternate, we assumed an effect size of 0.2, and drew individuals' phenotypes from $\mathcal{N}(0.2g_{id}, 1)$, where g_{id} is the genotype of individual d at SNP i. The phenotypes were generated using the SNP in the middle of each window, and z-scores were computed at all SNPs as described in the introduction of the Methods section.

Reference Panels

Reference panels were generated using the 1000 Genomes (1KG) Phase 3 data from the following 11 populations: CEU, IBS, FIN, GBR, TSI, YRI, MXL, PUR, CHB, JPT, GIH. For each dataset we analyzed (i.e. GALA II, CARDIoGRAMplusC4D) we removed any A/T and G/C SNPs to avoid strand issues. We then took an intersection of rsids between our data and the 1KG data to determine which SNPs to include in our reference panels. All SNPs for the reference panels were coded as the number of reference alleles an individual had (i.e. 0, 1, 2).

Results

We applied Adapt-Mix to summary statistics from simulated and real data to estimate the pairwise SNP correlation matrix (Σ). In this work, we use z-score imputation and joint-testing. For both datasets we used several approaches to estimate Σ and impute z-scores. All imputation was done using a window size of 200 SNPs and $\lambda = 0.1$. The values for window size and λ were chosen based on the recommended settings used in Pasaniuc *et al*(Pasaniuc, Zaitlen, Shi, Bhatia, Gusev, Pickrell, Hirschhorn, Strachan, Patterson, & Price, 2014). We measured the impact of using different methods to estimate Σ on z-score imputation by computing the mean-squared error (MSE) and Pearson correlation coefficient (*r*) between the imputed z-scores and true z-scores. In addition to imputation, we also performed joint-testing in the simulated data because we had access to the individual genotypes and thus the could compute the true SNP correlation matrix. Again, we measured the effect of several Σ estimation methods on joint-testing by computing the MSE and *r* between the true joint statistics and the estimated joint statistics.

Simulated Data

Simulated z-scores from the GALA II genotypes (see Simulation Framework) were used to determine whether our method gave more accurate results for a) imputing z-scores and b) computing joint-test statistics. Since there are multiple ways to optimize mixture frequencies using Adapt-Mix, we compared the use of several optimization strategies against the "best guess" approach. Using Adapt-Mix, we estimated Σ using 1KG reference panels by optimizing over each chromosome (1KG-Chrom), over the whole genome (1KG-Genome), and per window (1KG-Window). We note that any SNP used to measure imputation quality was excluded during optimization. Additionally, to evaluate how our method affects imputation and joint-testing when a "best guess" panel is unavailable, we removed both MXL and PUR panels and optimized frequencies over the chromosomes (1KG-No-PUR-MXL).

Population Frequencies

We applied our method to simulated data over Mexican and Puerto Rican individuals from the GALA II cohort(Borrell et al., 2013). Figure 1 shows the average frequency assigned to each population when frequencies were optimized per chromosome. When matching reference populations are included in the optimization (MXL for the Mexicans and PUR for the Puerto Ricans), nearly one third of the mixture is assigned to the matching reference panel. The rest of the frequencies are distributed to populations in a similar manner to the admixture proportions of each group(Baran, Pasaniuc, Sankararaman, Torgerson, Gignoux, Eng, Rodríguez-Cintrón, Chapela, Ford, Avila, Rodriguez-Santana, Burchard, & Halperin, 2012). Having predominantly Native American and European ancestry, Mexicans have frequencies distributed amongst European and East Asian panels in addition to MXL. However, when MXL and PUR are not included, we see an increase in frequency assigned to the East Asian panels. Puerto Ricans have more African ancestry than Native American ancestry, and we observe a correspondingly larger frequency of the YRI (African) panel and lower frequencies of East Asian panels.

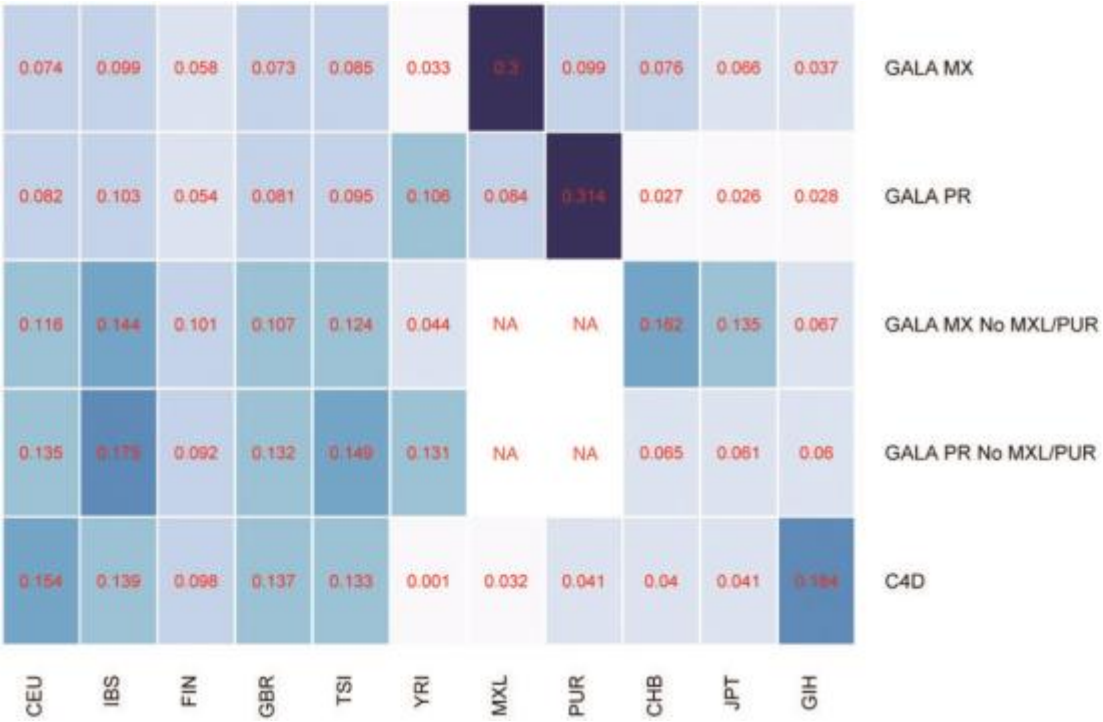


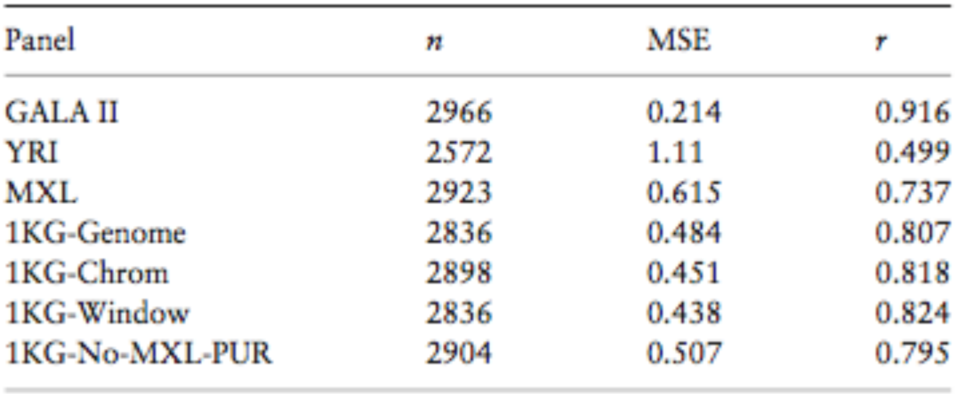
Fig. 1. This heatmap shows the average mixture frequency assigned to each reference population when optimizing over independent chromosomes for various datasets

Imputation

We next evaluated the imputation performance of the different approaches to estimating Σ. We measured each method's impact on imputation by computing the mean-squared error (MSE) and Pearson correlation coefficient (r) between the imputed z-scores and true z-scores. We imputed the z-score of the 100th SNP in every window. We restricted our analysis to SNPs with a MAF ≥ 0.01 in the reference panel since imputation quality tends to be poor for rare SNPs. We also removed from Σ SNPs that had a $r^2 \leq 0.003$ with the SNP we were imputing. When using a mixture reference panel we filtered SNPs using a mixture MAF. The mixture MAF for SNP i is $\sum_{k=1}^{K} f_k \mathrm{MAF}_{ki}$, where f_k is the mixture frequency assigned to population k and MAF_{ki} is the MAF of SNP i in k.

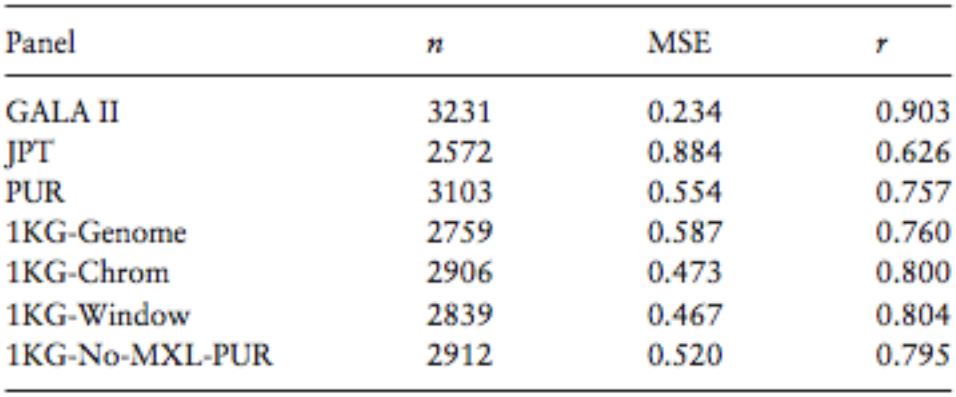
As the gold standard, the original GALA II genotypes were used to estimate Σ. It is clear from Tables 1 and 2 that using the original genotypes results in very high imputation quality. To demonstrate that using the wrong reference panel can cause a huge decrease in performance, we imputed z-scores using YRI and JPT as reference panels for the Mexicans and Puerto Ricans, respectively. Using the wrong reference panel resulted in MSE increasing over 400% in the Mexicans and over 250% in the Puerto Ricans.

Table 1. Performance of each reference panel when imputing *z*-scores for GALA II Mexicans



Panel	*n*	MSE	*r*
GALA II	2966	0.214	0.916
YRI	2572	1.11	0.499
MXL	2923	0.615	0.737
1KG-Genome	2836	0.484	0.807
1KG-Chrom	2898	0.451	0.818
1KG-Window	2836	0.438	0.824
1KG-No-MXL-PUR	2904	0.507	0.795

Table 2. Performance of each reference panel when imputing *z*-scores for GALA II Puerto Ricans



Panel	*n*	MSE	*r*
GALA II	3231	0.234	0.903
JPT	2572	0.884	0.626
PUR	3103	0.554	0.757
1KG-Genome	2759	0.587	0.760
1KG-Chrom	2906	0.473	0.800
1KG-Window	2839	0.467	0.804
1KG-No-MXL-PUR	2912	0.520	0.795

Next, z-scores were imputed using Adapt-Mix to estimate LD. We found that for imputation in admixed individuals, locally optimizing mixture frequencies over each window performs the best. For z-scores imputed over the whole genome there is a 28.8% decrease in MSE for the Mexicans and a decrease of 15.7% for the Puerto Ricans (See Tables 1 and 2). Similar decreases in MSE are seen when optimizing frequencies over the chromosome and the entire genome. Even when MXL and PUR were removed, we see that our method approach to estimating Σ outperforms the "best guess" panel. We also see increases in the *r* of imputed and true z-scores in the Mexicans and the Puerto Ricans when using Adapt-Mix. The increase in *r* is equivalent to an increase of 25.0% and 12.8% in effective sample size for the Mexicans and Puerto Ricans, respectively. Interestingly, the local optimization approach does not necessarily find mixture frequencies that are closest to the study's overall mixture of ancestry. The results here indicate that using such a mixture may not be the best for imputation accuracy and highlights the benefits of using the correct objective function when optimizing mixture frequencies for the selected summary statistics based method.

Joint-Test

Joint-testing of pairs of SNPs from summary statistics also relies on estimates of the pairwise correlation between SNPs(Yang et al., 2012). Using SNPs on chromosome 22 we computed true joint statistics using Σ computed from the genotypes of the GALA II individuals. The estimated joint statistics were computed using Σ estimated using Adapt-Mix. The mixture frequency optimization strategies were the same as those used in z-score imputation. We computed Joint statistics for SNPS that had a MAF or mixture MAF ≥ 0.05 in all of the Σ estimation approaches. Tables 3 and 4 show that using a Σ estimated from a mixture reference panel results in increased performance over using a "best guess" reference panel.

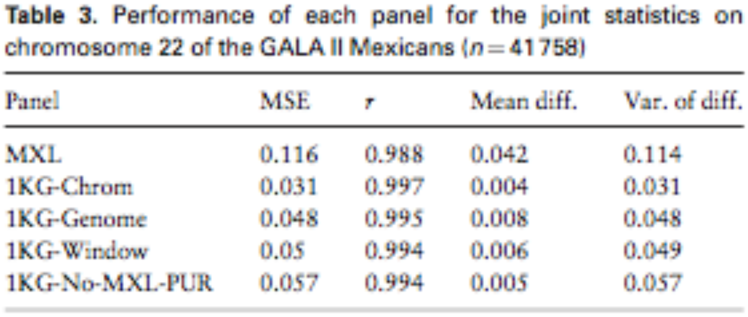


Table 3. Performance of each panel for the joint statistics on chromosome 22 of the GALA II Mexicans ($n = 41\,758$)

Panel	MSE	r	Mean diff.	Var. of diff.
MXL	0.116	0.988	0.042	0.114
1KG-Chrom	0.031	0.997	0.004	0.031
1KG-Genome	0.048	0.995	0.008	0.048
1KG-Window	0.05	0.994	0.006	0.049
1KG-No-MXL-PUR	0.057	0.994	0.005	0.057

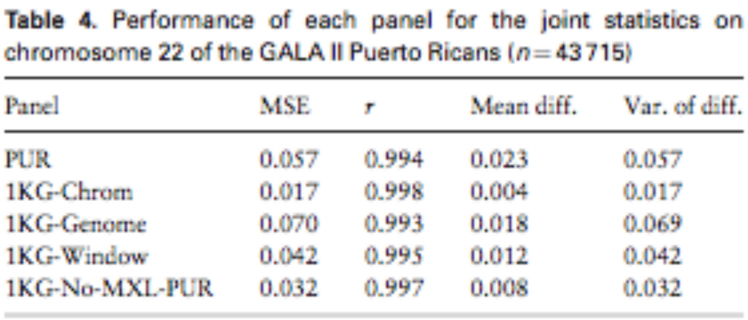


Table 4. Performance of each panel for the joint statistics on chromosome 22 of the GALA II Puerto Ricans ($n = 43\,715$)

Panel	MSE	*r*	Mean diff.	Var. of diff.
PUR	0.057	0.994	0.023	0.057
1KG-Chrom	0.017	0.998	0.004	0.017
1KG-Genome	0.070	0.993	0.018	0.069
1KG-Window	0.042	0.995	0.012	0.042
1KG-No-MXL-PUR	0.032	0.997	0.008	0.032

In both populations, the frequencies optimized per chromosome (1KG-Chrom) performed the best. Compared to using a “best guess” panel, we observed a 73.7% decrease in MSE for the Mexicans and a 70.2% decrease in MSE for the Puerto Ricans. We plotted the estimated joint statistics versus the true joint statistics for Mexicans and Puerto Ricans for different choices of Σ (Figure 2). The results show that joint statistics computed using the combined reference panels are in higher concordance with the truth than the “best guess" panel. Remarkably, even when MXL and PUR are removed from the mixture, estimates of Σ improvements can be clearly seen (Figure 2c and 2d).

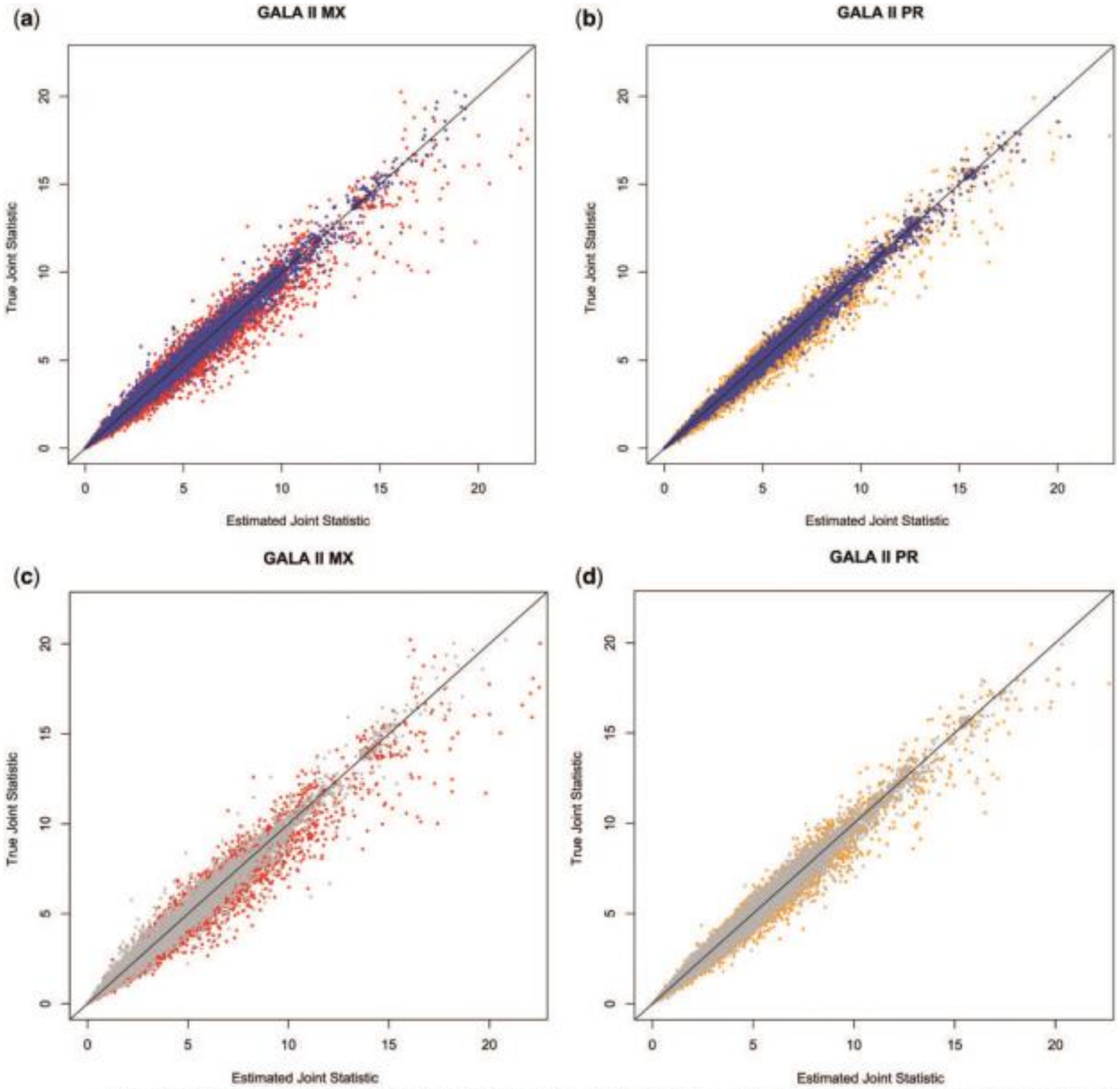


Fig. 2. Estimated joint statistic (*x* axis) versus the true joint statistic (*y* axis) in the GALA II individuals using Σ estimated from a 'best guess' reference panel and Adapt-Mix. (**a**) Joint statistics for the GALA II Mexicans using MXL (red) and 1KG-Chrom (blue). (**b**) Joint statistics for the GALA II Puerto Ricans using PUR (orange) and 1KG-Chrom (blue). (**c**) Joint statistics for the GALA II Mexicans using MXL (red) and 1KG-No-MXL-PUR (gray). (**d**) Joint statistics for the GALA II Puerto Ricans using PUR (orange) and 1KG-No-MXL-PUR (gray)

To show that the joint statistics produced by using our method for estimating correlations are unbiased (i.e. $E[J_{ij} - \hat{J}_{ij}] = 0$), we looked at the mean difference between the true statistics and estimated statistics. Tables 3 and 4 show that the mean difference is closer to 0 when our approach is used in both the Mexicans and Puerto Ricans. The 1KG-Chrom based correlation estimates generated differences in true versus estimated that were the closest to zero amongst all approaches. We can see from Tables 3 and 4 that 1KG-Chrom has the smallest variance for the differences in true versus estimated joint statistics. The “best guess" panels had the highest variance of all approaches except for 1KG-Genome in the Puerto Ricans. Additionally, we examined all estimated joint statistics that were more than 2 chi-squared units from the truth. In Mexicans, we saw 122 such statistics for the MXL and 22 for 1KG-Chrom (Figure 3a). A similar trend is seen in Puerto Ricans as well, with 53 large deviations for the PUR and 3 for 1KG-Chrom (Figure 3b). The decrease in frequency and magnitude of large differences demonstrates that using Adapt-Mix can help reduce the number of false positives in a joint analysis using reference panels. However, high deviations seen in both methods indicate that regardless of approach there is potential to misestimate the pairwise correlation coefficients of SNPs.

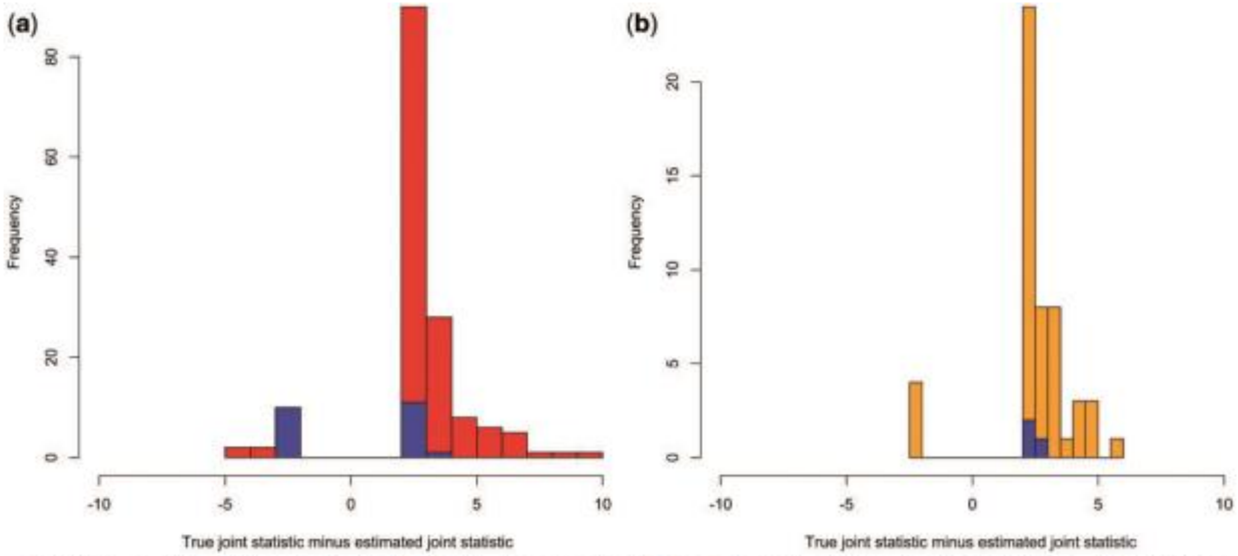


Fig. 3. Histogram of the deviations from the true joint statistic when using a 'best guess' panel and Adapt-Mix to estimate Σ for joint-testing. **(a)** Joint testing for GALA II Mexicans. MXL deviations are shown in red and 1KG-Chrom is shown in blue. **(b)** Joint testing for GALA II Puerto Ricans. PUR deviations are shown in orange and 1KG-Chrom is shown in blue

Real Data

We applied our method to the C4D coronary artery disease dataset from the CARDIoGRAMplusC4D consortium(T. C. A. D. C. G. Consortium, 2011; Schunkert et al., 2011). In the C4D study, the discovery cohort consisted of 14,790 South Asians and 15,692 Europeans. South Asians are known to have undergone admixture between two ancestral populations, with one of the ancestral populations being genetically similar to Europeans(Moorjani et al., 2013; Reich, Thangaraj, Patterson, Price, & Singh, 2009). Consistent with the admixture seen in South Asians, we see mixture frequencies for C4D that are assigned primarily to the European and the South Asian panels (see Figure 1).

Imputation

The C4D data provided us with an opportunity to assess how our method affects the performance of z-score imputation in the context of a dataset with different population structure than that used in the simulations. Unlike our simulations, where everybody was admixed, the summary statistics in C4D were generated using a mixture of individuals with homogenous ancestries (Europeans) and heterogeneous ancestries (South Asians). As we did for the simulated data, we used MSE and r of the imputed z-scores as our performance metrics. Here, we estimated Σ using a "best guess" reference panel, 1KG-Chrom, and 1KG-Window. We chose to optimize frequencies for the 1KG reference panels over each chromosome and each window because these two approaches performed the best in our simulations. We imputed the 100th SNP in each window and we restricted our analyses here to SNPs that had (mixture) MAF ≥ 0.01.

As the "best guess" reference panel for C4D we used GIH and CEU because the C4D discovery cohort was composed of roughly an equal number of individuals with a European or South Asian ancestry. When imputing we saw similar results to our simulations. Compared to using CEU or GIH there was a decrease of 30.1% or 36% in MSE, respectively (Table 5). In terms of r we saw increases of about 7% over CEU and about 9% over GIH for both 1KG-Window and 1KG-Chrom. The increase in correlation is equivalent to an increase of 15% in effective sample size compared to CEU.

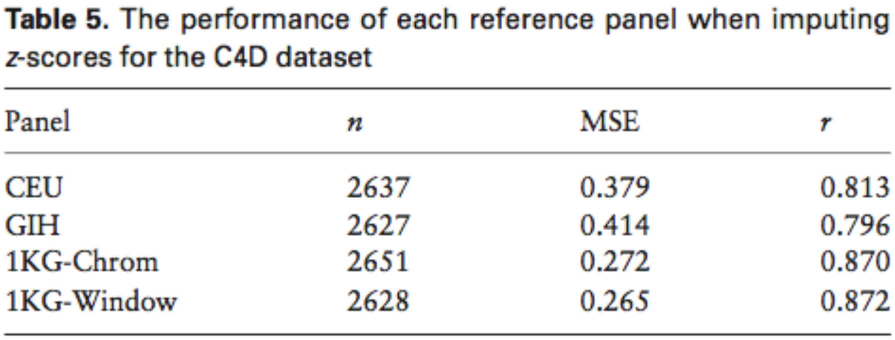


Table 5. The performance of each reference panel when imputing *z*-scores for the C4D dataset

Panel	n	MSE	r
CEU	2637	0.379	0.813
GIH	2627	0.414	0.796
1KG-Chrom	2651	0.272	0.870
1KG-Window	2628	0.265	0.872

Discussion

Summary statistics based methods requiring an estimate of the genetic correlation matrix are becoming increasingly popular, however, very few GWAS include LD information in their released data. In prior work, this information has been approximated by using LD information from "best guess" reference panels, but here we show that this can lead to high error rates even when a population closely matching the study population is available(Zaitlen, Kang, & Eskin, 2009). Our method can be used to improve the accuracy of any summary statistics based method that requires LD information by more accurately estimating the local genetic correlation structure using information available across several reference populations.

Our simulations have demonstrated the importance of accurately estimating the genetic correlation matrix. Using Adapt-Mix to estimate LD for summary statistics methods can increase their power and decrease their false positive rates. For example for z-score imputation, Pasaniuc *et al.*(Pasaniuc, Zaitlen, Shi, Bhatia, Gusev, Pickrell, Hirschhorn, Strachan, Patterson, & Price, 2014) showed that as long as there is a "best guess" reference panel available, there is no increase in false positive rate when imputing summary statistics. However, in the case that there is no "best guess" panel available, we have shown that there is a potential for increased false positives by using the wrong reference panel.

One of the biggest benefits of our method is allowing the analysis of arbitrary populations when a matching reference panel is not available. We were able impute z-scores and compute joint statistics with better precision "best guess" panels alone even after leaving out the relevant "best guess" panels from our computation of Σ. For datasets with admixed individuals, the high variability of ancestry proportions may make it harder to consistently model LD in an accurate manner with a single reference panel. For example, in in the Native American component Latinos, there is a high level of population substructure(Wang et al., 2008). In the 1000 Genomes reference panels, there are currently no Native American reference panels available. Although proxy populations such as CHB and JPT are often used, they are unlikely to capture the full resolution of each underlying sub-population. Accounting for all the fine scale differences seen in admixed individuals will improve with the collection of additional reference panels.

In this work we aimed to minimize the mean squared error of imputed summary statistics in our objective function because imputation was one of our main focuses. For other purposes it may be more appropriate to use a different objective depending on how the pairwise correlation estimates will ultimately be used. For example, Hormozdiari *et al.*(Hormozdiari et al., 2014) use summary statistics to fine map causal variants by finding the set of variants that maximize the likelihood of a multivariate normal distribution. In this case, optimizing frequencies for reference panels by using the multivariate normal likelihood may improve performance.

Improvements to Adapt-Mix may be made by using an out-of-sample approach to learning the mixture frequencies due to the potential of overfitting. Typically, overfitting will cause high prediction error variances. We have shown though, with the example of joint-testing, that overfitting should not be a major concern as the error variances are smaller when using Adapt-Mix compared to a "best guess" panel. Another enhancement could be made to Adapt-Mix by using partial correlations. Often covariates such as principal components are included in GWAS, which alter the genetic correlation structure of the individuals being studied. Partial correlations, which account for these covariates may provide even more accurate estimates of the Σ for use in summary statistics methods.

Chapter 2: Leveraging Admixed Populations to Understand Genes and Environment

1. An Ancestry Based Approach for Detecting Interactions

Introduction

Genetic association studies in humans have focused primarily on the identification of additive single nucleotide polymorphism (SNP) effects through marginal tests of association. There is growing evidence that both epistatic and gene-environment $(G \times E)$ interactions contribute significantly to phenotypic variation in humans and model organisms(Hemani et al., 2014; Jemal et al., 2011; E. Y. Kang et al., 2014; M. Lee, Raj, & Castillo, 2012; Rouhani et al., 2014). In addition to explaining additional components of missing heritability, interactions lend insights into biological pathways that regulate phenotypes and improve our understanding of their genetic architectures. However, identification of interactions in human studies has been complicated by the computational and multiple testing burden in the case of epistatic interactions, and the lack of consistently measured environmental covariates in the case of $G \times E$ interactions(Eichler, Flint, Gibson, Kong, & Leal, 2010; Manolio et al., 2009).

To overcome these challenges, we leverage the unique nature of genomes from recently admixed populations such as African Americans, Latinos, and Pacific Islanders. Admixed genomes are mosaics of different ancestral segments(Seldin et al., 2011) and for each admixed individual it is possible to accurately estimate θ, the proportion of ancestry derived from each ancestral population (e.g. the fraction of European/African ancestry in African Americans)(Alexander, Novembre, & Lange, 2009). Ancestry has been previously leveraged to demonstrate that an array of environmental and biomedical covariates are correlated with θ (Burchard, Ziv, Coyle, & Gomez, 2003

; Cheng et al., 2012; Choudhry et al., 2006; Florez et al., 2011; Kumar et al., 2013; 2010; Price et al., 2008; Reiner et al., 2007; Sanchez et al., 2010; Shaffer et al., 2007; Ziv et al., 2006) and we therefore consider its use as a surrogate for unmeasured and unknown environmental exposures. θ is also correlated with the genotypes of SNPs that are differentiated between the ancestral populations, suggesting that θ may be effectively used as a proxy for detecting multi-way epistatic interactions. Therefore, we propose a new SNP by θ test of interaction in order to detect evidence of interaction in admixed populations.

We first investigate the properties of our method through simulated genotypes and phenotypes of admixed populations. In our simulations we demonstrate that differential linkage-disequilibrium (LD) between ancestral populations can produce false positive SNP by θ interactions when local ancestry is ignored. To accommodate differential LD, we include local ancestry in our statistical model and demonstrate that this properly controls this confounding factor. We also show that our approach, the Ancestry Test of Interaction with Local Ancestry (AITL), is well-powered to detect $G \times E$ interactions when θ is correlated with the environmental covariates of interest and multi-way epistatic interactions. The power for detecting pairwise $G \times G$ interactions at highly differentiated SNPs is lower than direct interaction tests even after accounting for the additional multiple testing burden. However, the results of our simulations show that AITL is well powered to detect multi-way epistasis involving tens or hundreds of SNPs of small effects, not detectable by pairwise tests.

We first examined molecular phenotypes by applying our method to gene expression data from African Americans, as well as DNA methylation data from Latinos. Gene expression traits have previously been shown to have large-scale differences as a function of genetic ancestry(Price et al., 2008). Other molecular phenotypes, such as LDL levels, have also been shown to be associated with genetic ancestry (Fraser, Lam, Neumann, & Kobor, 2012; Galanter et al., 2016; Peralta et al., 2009; Price et al., 2008; Reiner et al., 2007; Spielman et al., 2007). For gene expression in particular, Price *et al.* showed that the effects of ancestry on expression are widespread and not restricted to a handful of genes. Additionally, molecular phenotypes are often used in deep phenotyping and Mendelian randomization studies and are thus directly relevant to elucidating disease biology(Delude, 2015; Vimaleswaran et al., 2013).

We identified one genome-wide significant interaction $(p < 5\times10^{-8})$ associated with gene expression in the African Americans and eight significant interactions $(p < 5\times10^{-8})$ associated with methylation in the Latinos. Two of the eight interactions associated with DNA methylation in the Latinos also replicated and the remaining six were enriched for low p-values ($p < 1.8\times10^{-6}$). To demonstrate that our approach works in larger data sets we also applied AITL to asthma case-control data from Latinos and observed well-calibrated test statistics. Together, these results provide evidence for the existence of interactions regulating expression and methylation and show that our approach is statistically sound.

Methods

Our approach is best illustrated with an example. First consider testing a SNP for interaction with an environmental covariate E. θ can serve as a proxy for E if the two are correlated, even if E is unknown or unmeasured (see Figure 1a). Now consider testing a SNP s for interaction with a SNP $j \neq s$ that is highly differentiated in terms of ancestral allele frequencies. For example, a SNP that has a high allele frequency in one ancestral population and a low allele frequency in the other ancestral population. θ can be used as a proxy for j because θ and the genotypes of SNP j will be correlated. Consider the case where j has a frequency of 0.9

in population 1 and frequency of 0.1 in population 2. Individuals with large values of θ (percentage of ancestry from population 1) are more likely to have derived j from population 1 and on average have greater genotype values at j. Similarly, individuals with small values of θ are more likely to have derived j from population 2 and on average have smaller genotype values. Thus, θ will be correlated with the genotypes of the individuals for highly differentiated SNPs and can serve as a proxy for detecting interactions (see Figure 1b).

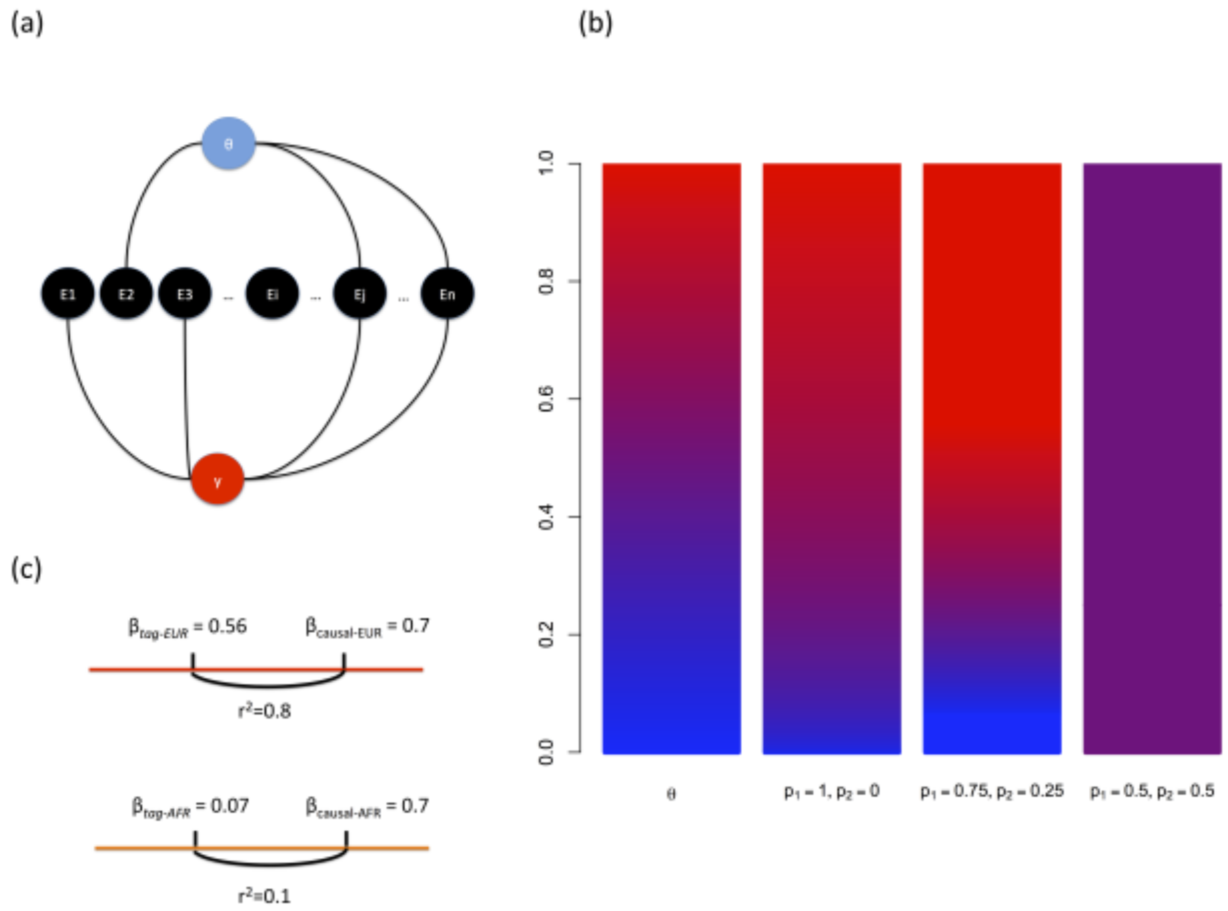


Figure 1. Examples of How Genetic Ancestry Can Be A Proxy for Interacting Covariates. (a) Model of how genetic ancestry θ can be correlated with various environmental exposures, some of which affect a phenotype. (b) Example of how the correlation between the probability of an AA genotype (bars 2-4) and values of θ (bar 1) increase with higher levels of SNP allele frequency differentiation. In this plot p_1 and p_2 denote the allele frequency of allele A in ancestral populations 1 and 2 respectively. (c) Example of how effect sizes at a tag-SNP may differ due to differential LD on distinct ancestral backgrounds (here, EUR and AFR).

Consider an admixed individual i who derives his or her genome from k ancestral populations. We denote individual i's global ancestry proportion as $\langle\theta_{i1}, \theta_{i2}, \dots, \theta_{ik}\rangle$, where $\sum_{j=1}^{k} \theta_{ij} = 1$.. The local ancestry of individual i at a SNP is denoted as $\gamma_{ia} \in \{0, 1, 2\}$ and is equal to the number of alleles from ancestry $a \in \{1 \dots k\}$ inherited at that SNP. Current methods allow us to estimate ancestry directly from genotype data both globally and at specific SNPs(Alexander et al., 2009; Baran, Pasaniuc, Sankararaman, Torgerson, Gignoux, Eng, Rodríguez-Cintrón, Chapela, Ford, Avila, Rodriguez-Santana, Burchard, & Halperin, 2012; Sankararaman, Sridhar, & Kimmel, 2008). We denote the genotype of an individual i at a given SNP as $g_i \in \{0, 1, 2\}$ and the corresponding phenotype as y_i.

In this work, we model continuous phenotypes in an additive linear regression framework. Assuming n (unrelated) individuals, define $\vec{y}$ to be the vector of all individuals' phenotypes. The model for the phenotype is then

$$\vec{y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon}$$

where $\vec{\varepsilon} \sim \mathcal{N}(0, \sigma)$ is a n×1 vector of error terms, $\boldsymbol{X}$ is a $n \times v$ matrix of v covariates, and $\vec{\beta}$ is a $v \times 1$ vector of the covariate effect sizes. We note that in our notation $\vec{v}^2 = \vec{v}^T \vec{v}$ for a vector $\vec{v}$. Assuming independence, the likelihood under this model is:

$$L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\left(-\frac{1}{2\sigma^2}\left(\vec{y} - \boldsymbol{X}\vec{\beta}\right)^2\right)$$

Then the log likelihood is given by the following expression:

$$\log(L) = -n\log(\sqrt{2\pi}) - n\log(\sigma) - \frac{\left(\vec{y} - \boldsymbol{X}\vec{\beta}\right)^2}{2\sigma^2}$$

Let L_1 denote the likelihood under the alternative and L_0 the likelihood under the null. We can compute the log-likelihood ratio statistic (D) using a maximum likelihood approach:

$$D = -2\ (\log L_1 - \log L_0)$$

$$= -2\left(n\log\left(\hat{\sigma}_{L_1}\right) + \frac{\left(\vec{y} - \boldsymbol{X}\hat{\vec{\beta}}_{L_1}\right)^2}{2\hat{\sigma}_{L_1}^2} \right) + 2\left(n\log\left(\hat{\sigma}_{L_0}\right) + \frac{\left(\vec{y} - \boldsymbol{X}\hat{\vec{\beta}}_{L_0}\right)^2}{2\hat{\sigma}_{L_0}^2} \right)$$

where $\hat{\vec{\beta}}_{L_1}$ and $\hat{\sigma}_{L_1}$are the maximum likelihood estimators of $\vec{\beta}$ and σ under the alternative, respectively and $\hat{\vec{\beta}}_{L_0}$ and $\hat{\sigma}_{L_0}$are the maximum likelihood estimators (MLEs) of $\vec{\beta}$ and σ under the null, respectively.

We note that for a case-control phenotype we would use the following likelihood and log-likelihood ratio statistic, assuming a logistic regression model:

$$L = \prod_{i=1}^{n}\left[\frac{1}{1+e^{-X_i\vec{\beta}}}\right]^{y_i}\left[1 - \frac{1}{1+e^{-X_i\vec{\beta}}}\right]^{1-y_i}$$

$$D = -2\ (\log L_1 - \log L_0)$$

$$= -2\left(\sum_{i=1}^{n} - \log\left(1 + e^{-X_i\hat{\vec{\beta}}_{L1}}\right) + \sum_{i=1}^{n} y_i\left(X_i\hat{\vec{\beta}}_{L_1}\right)\right)$$

$$+ 2\left(\sum_{i=1}^{n} - \log\left(1 + e^{-X_i\hat{\vec{\beta}}_{L0}}\right) + \sum_{i=1}^{n} y_i\left(X_i\hat{\vec{\beta}}_{L_0}\right)\right)$$

where $X_i$is the *i*-th row of the matrix $\boldsymbol{X}$, which correspond to the covariates of individual *i* and y_i ∈ {0,1} is the phenotype of individual *i.*

For linear regression, the MLE of the effect sizes is $\hat{\vec{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{y}$, and the MLE of the error variance is $\hat{\sigma}^2 = \frac{1}{n}\left(\vec{y} - \boldsymbol{X}\hat{\vec{\beta}}\right)^2$. $(\hat{\vec{\beta}}_{L_1}, \hat{\sigma}_{L_1}^2)$ and $(\hat{\vec{\beta}}_{L_0}, \hat{\sigma}_{L_0}^2)$ are the effect sizes and error variance estimates that maximize the respective likelihoods. *D* is distributed as χ^2 with *k* degrees of freedom (*df*), where *k* is the number of parameters constrained under the null.

1-df Ancestry Interaction Test (AIT)

The first test we present is the standard direct test of interaction. We test for a SNP's interaction with θ instead of an environmental covariate or another genotype. Let $\vec{g} = \langle g_1, \dots, g_n \rangle$ be the vector of the individuals' genotypes at a given SNP, $\vec{\theta}_a = \langle \theta_{1a}, \dots, \theta_{na} \rangle$ be the vector of their global ancestries for ancestry a, and $\vec{g} \times \vec{\theta}_a$ be the vector of interaction terms which result from the component-wise multiplication of the genotype and global ancestry vectors. We test the alternative hypothesis $\left(\hat{\beta}_{G\times\theta} \neq 0\right)$ against the null hypothesis $\left(\hat{\beta}_{G\times\theta} = 0\right)$.

$$H_1 : \vec{y} = \hat{\beta}_g \vec{g} + \hat{\beta}_{G\times\theta}(\vec{g} \times \vec{\theta}_a) + \hat{\beta}_\theta \vec{\theta}_a$$

$$H_0 : \vec{y} = \hat{\beta}_g \vec{g} + \hat{\beta}_\theta \vec{\theta}_a$$

In this test of interaction, we test a single ancestry versus the other ancestries that may be present in the population of interest. One parameter is constrained under the null which results in a statistic with k=1 *df.* Let $\hat{\beta}_{L_{\{0,1\}}(g)}$, $\hat{\beta}_{L_{\{0,1\}}(G\times\theta)}$, and $\hat{\beta}_{L_{\{0,1\}}(\theta)}$ denote the effect sizes of genotype, interaction, and global ancestry under a given hypothesis respectively. The statistic is given below.

$$D = -2\left(n\log\left(\hat{\sigma}_{L_1}\right) + \frac{\left[\vec{y} - \boldsymbol{X}\langle \hat{\beta}_{L_1(g)}, \hat{\beta}_{L_1(G\times\theta)}, \hat{\beta}_{L_1(\theta)} \rangle\right]^2}{2\hat{\sigma}^2_{L_1}} \right)$$

$$+ 2\left(n\log\left(\hat{\sigma}_{L_0}\right) + \frac{\left[\vec{y} - \boldsymbol{X}\langle \hat{\beta}_{L_0(g)}, 0, \hat{\beta}_{L_0(\theta)} \rangle\right]^2}{2\hat{\sigma}^2_{L_0}} \right)$$

where $\boldsymbol{X}$ is an n×3 matrix composed of $\vec{g}$, $\vec{\theta}_a$, and $\vec{g} \times \vec{\theta}_a$ as columns.

1-df Ancestry Interaction Test with Local Ancestry (AITL)

Given that the individuals we analyze in this work are assumed to be admixed, there is potential for confounding due to differential LD. An interaction that is not driven by biology could occur due to the possibility that a causal variant may be better tagged by a SNP being tested on one ancestral background versus another (See Figure 1c). We account for the different LD patterns on varying ancestral backgrounds by including local ancestry as an additional covariate in AITL. By including local ancestry, we assume that the SNP being tested is on the same local ancestry block as the causal SNP that it may be tagging. Such an assumption is reasonable because admixture in populations such as Latinos and African Americans are relatively recent events and their genomes have not undergone many recombination events. As a result, local ancestry blocks on average stretch for several hundred kilobases(Price et al., 2007; M. W. Smith et al., 2004).

Let $\vec{\gamma}_a = \langle \gamma_{1a}, \dots, \gamma_{na} \rangle$ be the vector of local ancestry calls for all individuals for ancestry a and let $\vec{g} \times \vec{\gamma}_a$ be the interaction terms from piecewise multiplication of the two vectors. We use the following alternative and null hypotheses:

$$H_1 \colon \vec{y} = \hat{\beta}_g \vec{g} + \hat{\beta}_{G\times\theta}(\vec{g} \times \vec{\theta}_a) + \hat{\beta}_\theta \vec{\theta}_a + \hat{\beta}_\gamma \vec{\gamma}_a + \hat{\beta}_{G\times\gamma} \vec{g} \times \vec{\gamma}_a$$

$$H_0 \colon \vec{y} = \hat{\beta}_g \vec{g} + \hat{\beta}_\theta \vec{\theta}_a + \hat{\beta}_\gamma \vec{\gamma}_a + \hat{\beta}_{G\times\gamma} \vec{g} \times \vec{\gamma}_a$$

Here we are testing for an interaction effect, i.e. $\hat{\beta}_{G\times\theta} \neq 0$, and constrain one parameter under the null resulting in a statistic with k=1 *df*. Let $\hat{\beta}_{L_{\{0,1\}}(G\times\gamma)}$ and $\hat{\beta}_{L_{\{0,1\}}(\gamma)}$ denote the effect sizes of the interaction between genotype and local ancestry and just local ancestry, respectively. The log likelihood ratio statistic is given by

$$D = -2\left(n\log\left(\hat{\sigma}_{L_1}\right) + \frac{\left[\vec{y} - \boldsymbol{X}\langle\hat{\beta}_{L_1(g)}, \hat{\beta}_{L_1(G\times\theta)}, \hat{\beta}_{L_1(\theta)}, \hat{\beta}_{L_1(\gamma)}, \hat{\beta}_{L_1(G\times\gamma)}\rangle\right]^2}{2\hat{\sigma}_{L_1}^2}\right)$$

$$+ 2\left(n\log\left(\hat{\sigma}_{L_0}\right) + \frac{\left[\vec{y} - \boldsymbol{X}\langle\hat{\beta}_{L_0(g)}, 0, \hat{\beta}_{L_0(\theta)}, \hat{\beta}_{L_0(\gamma)}, \hat{\beta}_{L_0(G\times\gamma)}\rangle\right]^2}{2\hat{\sigma}_{L_0}^2}\right)$$

where $\boldsymbol{X}$ is an $n\times 5$ matrix composed of $\vec{g}$, $\vec{\theta}_a$, $\vec{g}\times\vec{\theta}_a$, $\vec{\gamma}_{as}$, and $\vec{g}_s\times\vec{\gamma}_{as}$ as columns. All of these test-statistics are straightforwardly modified to jointly incorporate several ancestries in the case of multi-way admixed populations.

Standard Pairwise Test of Interaction and Controlling Confounding in Admixed Populations

Here we present the standard approach for testing for interaction between two SNPs *s* and *j*. We use the following alternative and null hypotheses.

$$H_1: \vec{y} = \hat{\beta}_s\vec{g}_s + \hat{\beta}_j\vec{g}_j + \hat{\beta}_{s\times j}\vec{g}_s\times\vec{g}_j + \hat{\beta}_\theta\vec{\theta}_a$$

$$H_0: \vec{y} = \hat{\beta}_s\vec{g}_s + \hat{\beta}_j\vec{g}_j + \hat{\beta}_\theta\vec{\theta}_a$$

If AITL is significant for a given SNP *s,* then any SNP *j* tested for interaction with *s* may be biased if *j* is correlated with covariates that are also correlated with θ. Furthermore, if the effects of the covariates correlated with θ are non-linear then controlling for the main effects of the SNPs and ancestry will account for the non-linear effects. We thus, propose the following alternative and null hypotheses:

$$H_1: \vec{y} = \hat{\beta}_s\vec{g}_s + \hat{\beta}_j\vec{g}_j + \hat{\beta}_{s\times j}\vec{g}_s\times\vec{g}_j + \hat{\beta}_\theta\vec{\theta}_a + \hat{\beta}_{s\times\theta}\vec{g}_s\times\vec{\theta}_a$$

$$H_0: \vec{y} = \hat{\beta}_s\vec{g}_s + \hat{\beta}_j\vec{g}_j + \hat{\beta}_\theta\vec{\theta}_a + \hat{\beta}_{s\times\theta}\vec{g}_s\times\vec{\theta}_a$$

We note that the utility of this test will require further investigation (see Discussion).

Simulation Framework

For all our simulations, we simulated 2-way admixed individuals. Global ancestry for ancestral population 1 (θ_1) was drawn from a normal distribution with $\mu = 0.7$ and $\sigma = 0.2$. Individuals i with $\theta_{i1} > 1$ or $\theta_{i1} < 0$ were assigned a value of 1 or 0, respectively. We simulated phenotypes of individuals to investigate our method in four different scenarios: $G\times E$ interactions, pairwise epistatic interactions, multi-way epistatic interactions, and false positive interactions due to local differential tagging.

To simulate phenotypes under the situation of a $G\times E$ interaction, we simulated a single SNP. For each individual i, we assigned the local ancestry or the number of alleles derived from population 1 (γ_{i1}) for each haplotype by performing two binomial trials with the probability of success equal to θ_{i1}. We then drew ancestry specific allele frequencies following the Balding-Nichols model by assuming a $F_{ST} = 0.16$ and drawing two population frequencies, p_1 and p_2, from the following beta distribution(Balding & Nichols, 1995).

$$p_1, p_2 \sim Beta\left(\frac{p(1-F_{ST})}{F_{ST}}, \frac{(1-p)(1-F_{ST})}{F_{ST}}\right)$$

where p is the ancestral population allele frequency and is set to 0.2. Genotypes were drawn using a binomial trial for each local ancestry haplotype with the probability of success equal to p_1 or p_2 for values of $\gamma_{i1} = 0, 1$, or 2. Environmental covariates correlated with the proportion of ancestry from population 1, E_i, were generated for each individual i by drawing from a normal distribution $\mathcal{N}(\mu = \theta_{i1}, \sigma_E)$, where σ_E is the standard deviation of the environmental covariates. σ_E was varied from 0 to 5 in increments of 0.005 to create E_i's that were correlated with individuals' global ancestries in varying degrees. We generated phenotypes for individuals assuming only an interaction effect by drawing from a normal distribution, $\mathcal{N}(\mu = \beta_{G\times E}\times g_{i1}\times E_i, \sigma = 1)$ for a given interaction effect size ($\beta_{G\times E}$).

To simulate phenotypes based on pairwise epistatic interactions, we simulated two SNPs. At both SNPs, we assigned the local ancestry values as described for the $G\times E$ case. We assigned genotypes for individuals at the first SNP assuming an allele frequency of 0.5 for both populations and drawing from two binomial trials. We assigned genotypes at the second SNP over a wide range of ancestry specific allele frequencies to simulate different levels of SNP differentiation. Ancestry specific allele frequencies were initially $p_1 = p_2 = 0.5$ and iteratively increasing p_1 by 0.005 while simultaneously decreasing p_2 by 0.005 until $p_1 = 0.95$ and $p_2 = 0.05$. Genotypes at the second SNP were drawn using the same approach described for $G\times E$. Using the simulated genotypes, phenotypes were drawn from a normal distribution, $\mathcal{N}(\mu = \beta_{G\times G}\times g_{i1}\times g_{i2}, \sigma = 1)$, where g_{is} is the genotype for individual *i* at the simulated SNP.

To simulate phenotypes based on multi-way epistatic interactions, we simulated a SNP *s* and *m* (independent) SNPs with pairwise interactions with *s*. Genotypes for individuals at SNP *s* were assigned assuming an allele frequency of 0.5 for both populations and drawing from two binomial trials. Genotypes at the *m* interacting SNPs were assigned in the same manner as the 2nd SNP in the pairwise interaction simulations. Using the simulated genotypes, phenotypes were drawn from a normal distribution, $\mathcal{N}\left(\mu = \sum_{j=1}^{m} \beta_{s\times j}\ g_{is}\times g_{ij}, \sigma = 1\right)$ where g_{ix} is the genotype for individual *i* at the simulated SNP *x*.

To simulate the scenario of differential LD on different ancestral backgrounds leading to false positives, we simulated phenotypes based on a single causal SNP that was tagged by another SNP. At both SNPs, local ancestries were assigned as described previously and genotypes were drawn using ancestry specific allele frequencies. Ancestral allele frequencies were assigned such that the average r^2 between the causal and tag SNP was 0.272 on the background of ancestral population 1 and 0.024 on the background of ancestral population 2. Thus, the tag SNP was only a tag on the population1 background and not on the population 2 background. Phenotypes were drawn from a normal distribution, $\mathcal{N}(\mu = \beta_{Causal} \times g_{ic}, \sigma = 1)$, assuming no interaction and $\beta_{Causal} = 0.7$, where g_{ic} is the genotype of individual i at the causal variant c.

We implemented our approach in an R package (GxTheta), which is available for download at http://www.scandb.org/newinterface/GxTheta.html

Ancestry Inference

Global ancestry inference was done using ADMIXTURE (Alexander et al., 2009) and local ancestry inference was done using LAMP-LD (Baran, Pasaniuc, Sankararaman, Torgerson, Gignoux, Eng, Rodríguez-Cintrón, Chapela, Ford, Avila, Rodriguez-Santana, Burchard, & Halperin, 2012). CEU and YRI from 1000 Genomes Phase 3 (T. 1. G. P. Consortium, 2012) were used as the European and African reference panels. For the Native American reference panels, 95 Native Americans genotyped on the Axiom LAT1 array were used(Drake et al., 2014).

Filtering for Related Individuals

All analyses in real data were filtered for related individuals due to the possibility of cryptic relatedness causing false positives. To filter for related individuals, we estimated kinship coefficients between all pairs of individuals using REAP (Thornton et al., 2012). We defined two individuals as related if they had a kinship coefficient greater than 0.025. For a pair of related individuals, we removed the one with a greater number of other individuals to whom he or she was related. In the case of a tie, we removed one of the pair at random.

Data Normalization

Gene Expression Normalization

Gene expression data (see Results) were first standardized for each gene such that mean expression was 0 and variance was 1. We then computed a covariance matrix of individual's expression values and performed PCA on the covariance matrix. Residuals were computed for all expression values by adjusting for the top 10 principal components and the mean for each gene was added back to the residuals. Due to the high dynamic range of gene expression compared to methylation we conservatively chose to additionally perform quantile normalization. We then sorted the gene expression residuals and used the quantiles of their rank order to draw new expression values from a normal distribution, $\mathcal{N}(\mu = 0, \sigma = 1)$, by using the inverse cumulative density function[24,25].

Methylation Data Normalization

Raw methylation values (see Results) were first normalized using Illumina's control probe scaling procedures. All probes with median methylation less than 1% or greater than 99% were removed and the remaining probes were logit-transformed as previously described(Du et al., 2010). To control for extreme outliers, we truncated the distribution of methylation values. For a given probe, we first computed the mean and standard deviation of the methylation values.

We then set any methylation values deviating more than 2.58 standard deviations from the mean to the methylation value corresponding to the 99.5th quantile.

Results

<u>*Simulated Data*</u>

To determine the utility of using θ as a proxy for unmeasured and unknown environmental covariates, we applied the AITL to simulated 2-way admixed individuals. We tested θ_1, the proportion of ancestry from ancestral population 1, for interaction with simulated SNPs (see Simulation Framework). Power was computed over 1,000 simulations, assuming 10,000 SNPS being tested, and using a Bonferroni correction p-value cutoff of 5×10^{-6}. We calculated the power using assumed interaction effect sizes (either $\beta_{G\times G}$ or $\beta_{G\times E}$) of 0.1, 0.2, 0.3, and 0.4 (see Simulation Framework). Although the few interactions reported for human traits and diseases have smaller effects in terms of the phenotypic variance they explain, we simulated large effects because genetic and environmental effect sizes in omics data, such as the expression and methylation data considered here, are known to be of larger magnitude. For example, some cis-eQTL SNPs explain up to 50% of the variance of gene expression(Grundberg et al., 2012). However for most phenotypes, known interactions will explain a very small proportion of the phenotypic variance, mainly due to the fact that so few interactions have been identified and replicated(Aschard et al., 2012).

Power When Using θ as a Proxy for Highly Differentiated SNPs

To determine whether using θ as a proxy for highly differentiated SNPs is more powerful than testing all pairs of potentially interacting SNPs directly, we simulated two interacting SNPS in 1000 admixed individuals (see Simulation Framework). We then tested for an interaction using AITL by replacing the genotypes at the highly differentiated SNP with $\vec{\theta}_1$. We observed that even with moderate effect sizes, using θ in place of the actual genotypes does not provide any increase in power even after accounting for multiple corrections (see Figure 2a). This is in agreement with recent work showing the limited utility of local ancestry by local ancestry interaction test to identify underlying SNP by SNP interaction when genotype data are available(Aschard, Gusev, Brown, & Pasaniuc, 2015). For the larger effect sizes we simulated, we do see power increasing as the delta between ancestral frequencies increases. The plots show that AITL has little power unless the effect was very strong. Figure 2b reveals that even with the multiple correction penalty, testing all pairwise SNPS directly is always more powerful. We note that when testing the interacting SNPs directly, we used a cutoff p-value of 1×10^{-9} since in theory we were testing all unique pairs of 10,000 SNPs. Based on these results, we would recommend testing for pairs of interacting SNPs directly if pairwise $G\times G$ interactions are a subject of interest in the study.

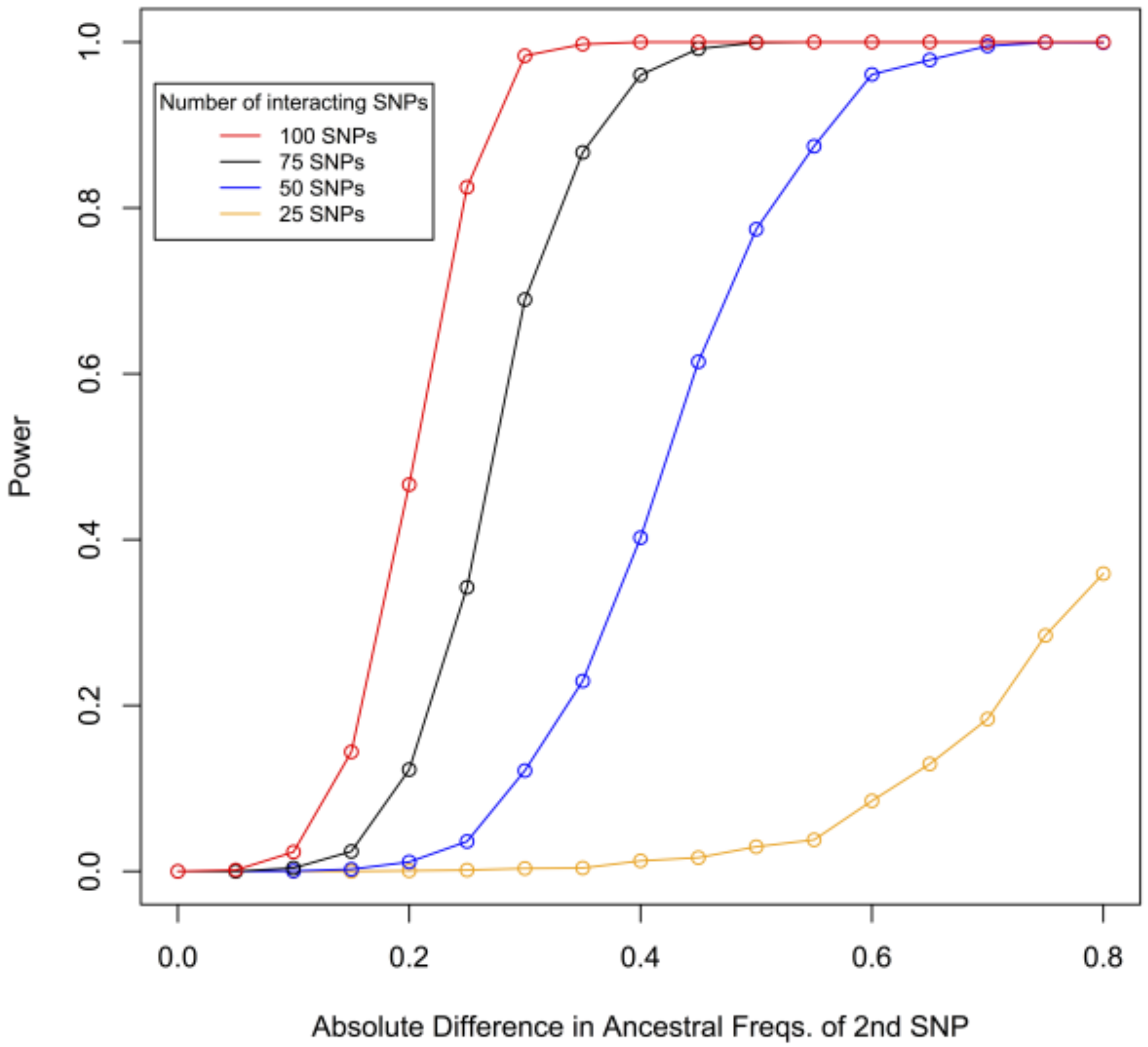


Figure 2. Power Plots for Pairwise Interaction Simulations.
Power of testing $G \times \theta$ (a) versus testing pairwise SNPs directly (b) as a function of the difference in the ancestral allele frequencies at a differentiated SNP.

However, when multi-way interactions are considered, AITL may become more powerful since differentiated SNPs across the genome will be correlated with genetic ancestry. These simulations are important as other studies have suggested that higher order interactions may be important for some traits(De, Hu, Moore, & Gilbert-Diamond, 2015; Hemani et al., 2014; Ritchie et al., 2001). To evaluate the ability of θ to serve as a proxy for multiple (independent) differentiated SNPs, we simulated a scenario where a candidate SNP *z* had interactions with m SNPs (see Simulation Framework). For each interaction, we assumed a small interaction effect size ($\beta_{G\times G} = 0.025$), which would not be detectable using a pairwise approach, as we demonstrated in the pairwise simulation. Figure 3 shows that AITL is better powered to detect the existence of interactions than a pairwise approach in the presence of multiple interacting SNPs with a candidate SNP.

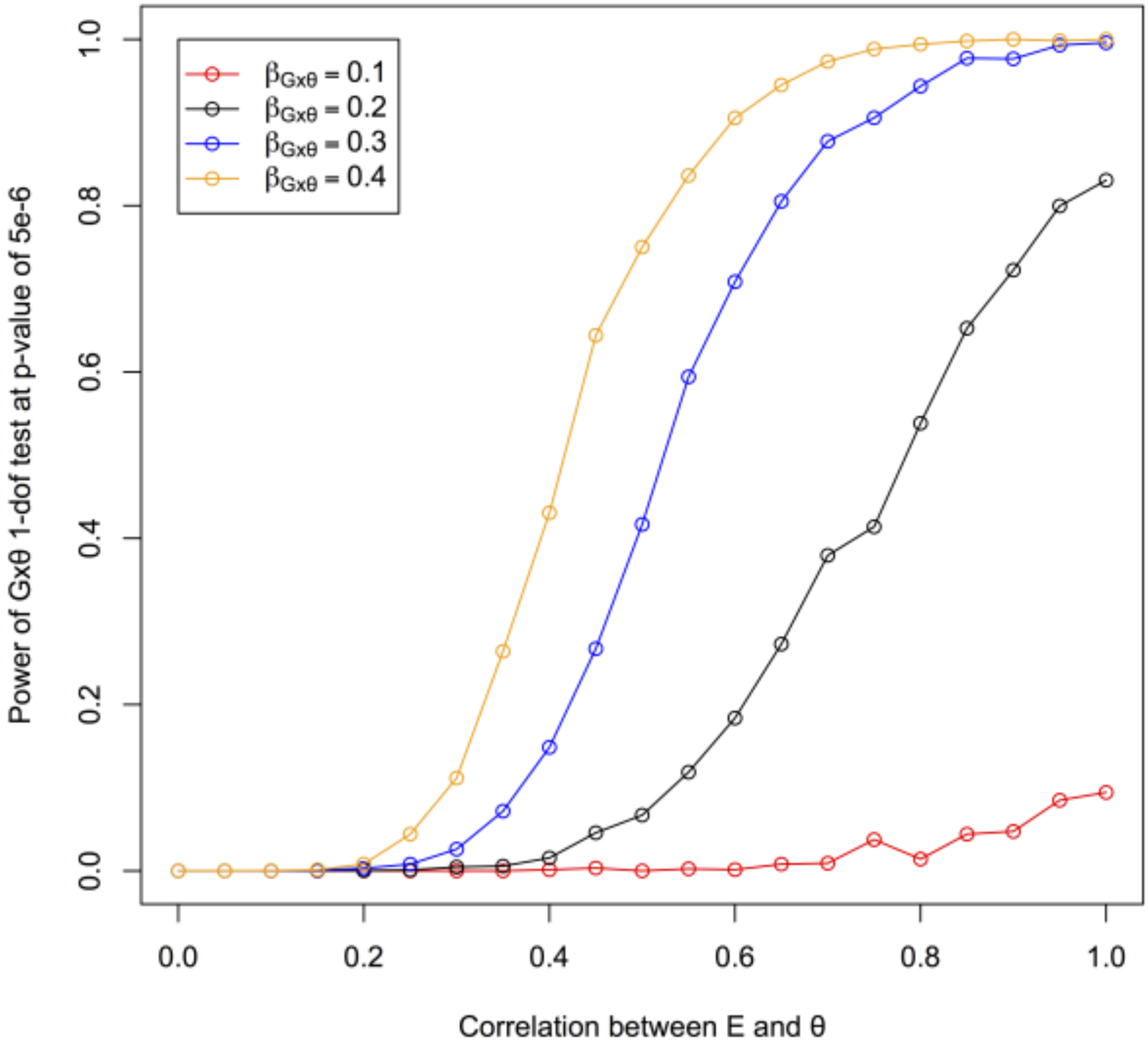


Figure 3. Power Plots for Multi-way Pairwise Interaction Simulations.
Power of testing $G\times\theta$ as a function of the difference in the ancestral allele frequencies for multiple interacting SNPs.

Power When Using θ as a Proxy Environmental Covariate

When assessing the utility of θ as a proxy for an environmental covariate E, we simulated 3000 individuals. E was simulated such that it was correlated with the global ancestries in varying degrees (see Simulation Framework). Figure 4 shows the power of the AITL as a function of the Pearson correlation between $\vec{\theta}_1$ and E. The power of testing E directly is exactly the power of the AITL when the correlation is equal to 1. As expected, as the correlation increases, the power increases as well. When the effect size is 0.1, the power to detect a $G\times E$ interaction is low whether one uses θ_1 or E. However, both tests are much better powered for effect sizes greater or equal to 0.2, with the AITL's power being dependent on the level of correlation. Note that using θ as a proxy for E is equivalent to testing G$\text{x}E$ in the presence of measurement error. Under the assumption of non-differential error with regard to the outcome (e.g. the correlation between θ and E is equal among cases and control) such a test is underpowered but has a controlled type I error rate under the null(Wong, Day, Luan, Chan, & Wareham, 2003).

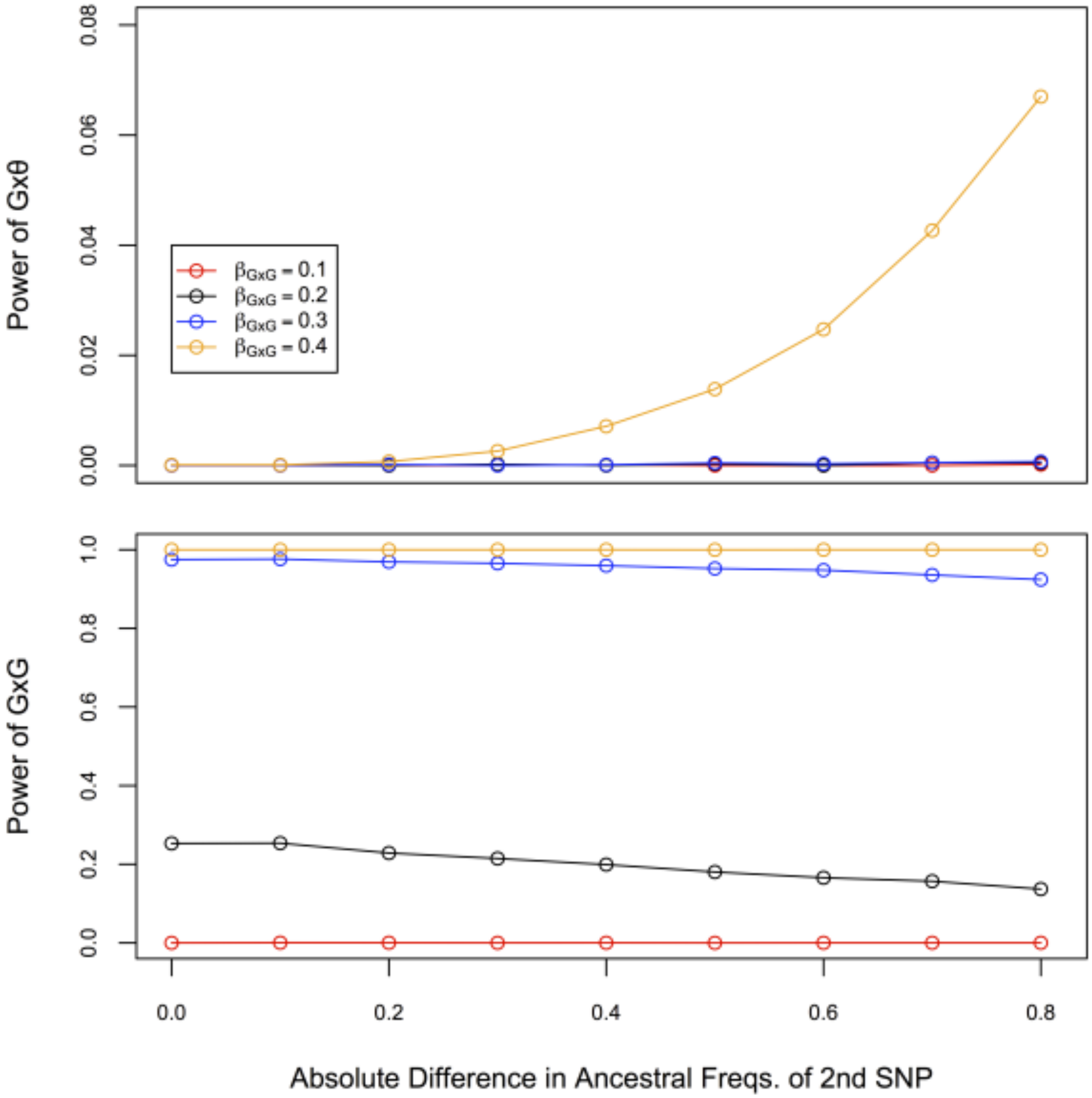


Figure 4. Power Plots for $G \times E$ Interaction Simulations.
Power of testing $G \times \theta$ as a function of the correlation between an environmental covariate and genetic ancestry.

Differential LD

To demonstrate that differential LD has the potential to cause inflated test-statistics, we ran 10,000 simulations of 1000 admixed individuals. For each individual we simulated 2 SNPs, a causal SNP and a tag SNP. The LD between the tag SNP and causal SNP was different based on the ancestral background the SNPs were on (see Simulation Framework). Over 10,000 simulations, we computed the mean χ_1^2 test-statistic for the AIT and the AITL. We note that the phenotypes for these simulations were generated under a model that assumed no interaction. We observed a mean $\chi_1^2 = 0.996$ with a standard deviation of 1.53 for AITL. AIT, which does not condition on local ancestry, had a mean $\chi_1^2 = 3.59$ with a standard deviation of 3.60. We also looked at genomic control λ_{GC}, the ratio of the observed median χ^2 over the expected median χ^2 under the null(Devlin & Roeder, 2004). λ_{GC} compares the median observed χ^2 test-statistic versus the true median under the null. In our simulations, we observed $\lambda_{GC} = 5.81$ for AIT and $\lambda_{GC} = 0.980$ for AITL (see Supplementary Figure S1). Last, we computed the proportion of test-statistics that passed a p-value threshold of .05 and .01 in our simulations. The AIT had 3687 statistics passing a p-value of .05 and 1687 at a threshold of .01, whereas AITL had 464 and 96 at the same p-value thresholds. The results for AITL are as expected under a true null. The results from our simulations show that not accounting for local ancestry can result in inflated test-statistics and can potentially lead to false positive findings.

Real Data

Coriell Gene Expression Results

We first applied our method to the Coriell gene expression dataset(Simon-Sanchez et al., 2007). The Coriell cohort is composed of 94 African-American individuals and the gene expression values of ~8800 genes in lymphoblastoid cell lines (LCLs). Since African Americans derive their genomes from African and European ancestral backgrounds, we tested for interaction between a given SNP and the proportion of European ancestry, θ_{EUR}. Each SNP by θ_{EUR} term was tested once for association with the expression of the gene closest to the SNP. We observed well-calibrated statistics with a λ_{GC} equal to 1.04 (see Supplementary Figure S2). In the LCLs, we found that interaction of rs7585465 with θ_{EUR} was associated with ERBB4 expression (AITL $p = 2.95\times10^{-8},$ marginal $p = 0.404$) at a genome-wide significant threshold ($p \leq 5\times10^{-8}$). rs7585465 has a 'C' allele frequency of 0.218 in the Corriell data and appears to be differentiated between CEU and YRI with allele frequencies of 0.619 and 0.097 in the respective populations.

Given that the gene expression values come from LCLs (all cultured according to the same standards), the SNPs may be interacting with epigenetic alterations due to environmental exposures that have persisted since transformation into LCLs. This scenario is unlikely, and we believe that signals are driven by multi-way epistatic interactions. In our simulations, we showed that using θ as a proxy for a single highly differentiated SNP is underpowered compared to testing all pairs of potentially interacting SNPs directly. However, there are many SNPs that are highly differentiated across the genome with which θ will be correlated. It is therefore possible that θ is capturing the interaction between the aggregate of many differentiated trans-SNPs (i.e. global genetic background) and the candidate SNP. This is consistent with a recently reported finding, conducted in human iPS cell lines, that genetic background accounts for much of the transcriptional variation(Martin et al., 2014; Rouhani et al., 2014).

Although we believe the ERBB4 result to be representative of multi-way epistasis, we performed a standard pairwise interaction test (see Methods) to check for interaction between rs7585465 and other SNPs genome-wide. Interestingly, we found that the standard interaction test (see Methods) showed substantial departure from the null with a λ_{GC} equal to 1.8 (see Supplementary Figure S3). Since the interaction of rs7585465 by θ was significant, the pairwise interaction test-statistics of rs7585465 by any SNP j can be inflated if j is correlated with θ. We found that including the original significant SNP by θ term in the null (see Methods) brought the λ_{GC} down to 1.05, and controlled for such scenarios in this dataset (See Supplementary Figure S3). As we had previously anticipated, identifying the exact interactions driving the SNP by θ interaction proved to be difficult. We found one borderline significant SNP (rs4839709, $p = 3.08\times10^{-7}$) but no interactions that passed genome-wide significance. These results are consistent with what we have observed in simulations, in which even though a standard pairwise interaction test is underpowered to detect interactions, AITL is able to identify the main locus involved in a multi-way interaction.

GALA II Case-Control

To determine if our method is biased in large structured GWAS data, we applied AITL to case-control data from a study of asthmatic Latino individuals called the Genes-environments and Admixture in Latino Americans (GALA II)(Borrell et al., 2013). The dataset includes 1158 Mexicans and 1605 Puerto Ricans, which were analyzed separately. Case status was assigned to individuals if they were between the ages of 8 and 40 years with a physician-diagnosed mild to moderate-to-severe asthma. Additionally, they had to have experienced 2 or more asthma related symptoms in the previous 2 years at the time of recruitment(Torgerson et al., 2012). In the Mexicans and Puerto Ricans there were 548 and 797 cases, respectively. In our analysis, we also included BMI, age, and sex as additional covariates. We observed well-calibrated statistics with a λ_{GC} equal to 1.00 and 0.98 in the Mexicans and Puerto Ricans, respectively (see Supplementary Figure S5). In contrast to the molecular phenotype data, searches for interactions in these phenotypes did not yield any findings passing genome-wide significance. This is consistent with previous disease studies that have failed to find many replicable interactions in disease studies(Aschard et al., 2012). In the data here, the lack of any findings may be due to the relatively small sample size or because the effects of the interactions are extremely small (if they exist for covariates correlated with θ_{EUR}).

GALA II Methylation Results

We searched for interactions in methylation data derived from a study of GALA II asthmatic Latino individuals(Borrell et al., 2013). The methylation data is composed of 141 Mexicans and 184 Puerto Ricans. As the phenotype, we used DNA methylation measurements on ~300,000 markers from peripheral blood. As we had done with gene expression, we tested for interaction between a given SNP and θ_{EUR} using AITL. All SNPs within a 1 MB window centered around the methylation probe were tested. We used the European component of ancestry because it is the component shared most between Mexicans and Puerto Ricans (see Table 1). We observed well-calibrated test-statistics with λ_{GC} equal to 1.06 in the Mexicans and 0.96 in the Puerto Ricans (see Supplementary Figure S6). We tested 128,794,325 methylation-SNP pairs, which result in a Bonferroni corrected p-value cutoff of 3.88×10^{-10}. However, this cutoff is extremely conservative given the tests are not independent. We therefore report all results that are significant at 5×10^{-8} in either set as an initial filter. We found 5 interactions in the Mexicans and 3 in the Puerto Ricans that are significant at this threshold (see Table 2).

Table 1. Distribution of Ancestry in Coriell and GALA II.

Dataset	θ_{EUR}	θ_{AFR}	θ_{NAM}
Coriell	μ=0.212, σ=0.021	μ=0.788, σ=0.021	NA
GALA II MX	μ=0.396, σ=0.149	μ=0.043, σ=0.025	μ=0.561, σ=0.159
GALA II PR	μ=0.641, σ=0.094	μ=0.246, σ=0.101	μ=0.113 σ=0.024

Mean and variance of the global ancestry distributions for each dataset.

Table 2. GALA II DNA Methylation Analysis Results.

GALA II Population	Probe Gene	Probe ID	rsid	Distance of SNP to Probe	Marginal p-value	AITL p-value	AITL Replication p-value
MX	CNFN	cg14327995	rs16975986	280795	2.49E-09	5.69E-09	9.27E-03
MX	C11orf95	cg16678159	rs7106153	249768	2.58E-01	2.52E-08	9.39E-02
MX	NA	cg05697734	rs1560919	13711	1.14E-01	2.21E-08	8.18E-03
MX	TNK2	cg01792640	rs67217828	278866	4.49E-01	6.38E-09	1.43E-02
MX	HDAC4	cg06533788	rs925736	9548	4.51E-01	3.09E-09	2.80E-02
PR	NA	cg07436864*	rs8117083	31813	7.46E-02	1.34E-09	5.34E-03
PR	NA	cg16803083*	rs4312379	63847	3.69E-01	2.29E-08	2.31E-04
PR	SERPINA6	cg10025865	rs17091085	247796	6.83E-01	2.97E-08	8.05E-03

P-values for AITL applied to the methylation data in the GALA II Latinos. MX and PR denote Mexicans and Puerto Ricans respectively in the GALA II population columns. The probe gene column shows the gene that the methylation probe lies in. The marginal column is the p-value for standard linear regression of methylation on genotype while controlling for population structure. * indicates results that replicated between the Mexicans and Puerto Ricans.

Unlike the Coriell individuals, who are 2-way admixed, the GALA II Latinos are 3-way admixed and derive their ancestries from European, African, and Native American ancestral groups. Consequently, to confirm that incomplete modeling or better tagging on one of the non-European ancestries was not driving the results, we retested all significant interactions including a second component of ancestry for AITL. In the case of the Mexicans, we included African and European ancestry, and in the case of the Puerto Ricans, we included European and Native American ancestry. Even after adjusting for the second ancestry the interactions between SNP and θ_{EUR} remained highly significant (see Supplementary Table 1).

As we did for the gene expression data, we attempted to identify pairwise interactions involved in the methylation data results. For each genome-wide significant result, we performed a standard pairwise interaction test of all SNPs with the original SNP found to be significant with AITL. We were unable to identify any significant interactions after applying genomic control to the results. For all tests, we included the significant SNP by θ term (see Methods) in the null. For this dataset, unlike the gene expression data, we observed substantial remaining departure from the null (see Supplementary Table S2) even after including the original significant SNP by θ term, suggesting there may be other factors that need to be accounted for when testing for interactions in admixed populations. The results from our pairwise scan are what we would anticipate, given that in simulations only AITL (not the standard pairwise interaction test) was able to identify the main locus involved in the multi-way interaction.

We then performed a replication study of the significant Puerto Rican associations in the Mexican cohort and vice versa. To account for the fact that we are replicating eight total results across both populations, we used a Bonferroni corrected p-value threshold equal to $.05/8 = 6.25\times10^{-3}$. The interaction of rs4312379 and rs4312379 with ancestry in the Puerto Ricans replicated in the Mexicans. Furthermore, there was a highly significant enrichment of low p-values in the replication study among the discovery results (permutation $p < 1{\times}10^{-4}$). Furthermore, 5 out of the 6 non-replicating results have a p-value less than 0.05 (binomial test $p < 1.8{\times}10^{-6}$). The results of the permutation and binomial test suggests that the interactions that did not replicate are likely to do so with bigger sample sizes. It is important to note that replicated interactions and the enrichment for low p-values do not necessarily indicate that the same genetic or environmental covariates are interacting with the genetic locus in both populations. The covariates correlated with θ_{EUR} in one population are not necessarily those correlated with θ_{EUR} in the other population. There may be correlations which exist in both populations but θ_{EUR} serves as a proxy for all such correlated covariates and therefore should not be necessarily viewed as a proxy for any specific one. Overall, our results from the GALA II (methylation) cohort suggest there are both genetic and environmental variables contributing to epistasis that have yet to be discovered in admixed individuals.

Discussion

For many disease architectures, interactions are believed to be a major component of missing heritability(Eichler et al., 2010). Finding new interactions has proven to be difficult for logistical, statistical, biological, and computational reasons. In this study, we have demonstrated that in admixed populations, testing for $G{\times}\theta$ interactions can be leveraged to overcome some of the difficulties typically encountered when searching for interactions. The computational cost is minimal and has the same order as running a standard GWAS.

One drawback of our method is that it does not identify which covariate is interacting with a genetic locus. Nevertheless, the approach can show whether an interaction effect exists in a given dataset and if it does exist, our method ensures that an underlying genetic or environmental covariate(s) is correlated with ancestry. Additionally, in the case where there is no marginal effect, our approach identifies new loci and shows that the genetic locus influences the phenotype and exerts its effects through interactions, which has important implications for the genetic architecture of the phenotype. The relative contribution of additive and non-additive genetic effects to variability in molecular phenotypes and disease risk is an important area of investigation, and our approach provides a direct test for detecting non-additive contributions(Powell et al., 2013).

Environmental covariates are often not consistently measured across cohorts whereas genetic ancestry is nearly perfectly replicable. Testing for the presence of interaction using a nearly perfectly reproducible covariate may enhance our understanding of the genetic basis of disease and other traits. Our method also provides the additional benefit of not being confounded by interactions between unaccounted-for covariates(Keller, 2014).

Association testing for interaction effects involving continuous environmental exposures in the context of mixed-models remains an open problem. For binary environmental exposures, it has been shown that mixed-models control for population structure nominally better than including genetic ancestry (or principal components) as a covariate(Sul et al., 2016). Because it is unclear how mixed-models perform with continuous environmental exposures, especially those correlated with ancestry, in our analyses we took the standard approach of filtering related individuals and including ancestry as a covariate.

It has been shown that 2-step analyses may be more powerful for detecting interactions when exposures are binary (Hsu et al., 2012; Kooperberg & LeBlanc, 2008; Murcray, Lewinger, & Gauderman, 2009). However, these studies have primarily been done in a single homogeneous population, and the correct null distribution for the interaction effect must assume that the 2^{nd} stage procedure is independent of the marginal effect test-statistic. In real data, using a 2-step approach in conjunction with AITL to test for interactions may be problematic because the interaction effect size will not necessarily be independent of the marginal effect size, as the allele frequency at any SNP will be a function of ancestry in an admixed population. Additionally, only 1 of the interaction results that we report here had a marginal effect ($p < 0.05$) and thus would have been missed by a 2-step approach. Thus, our approach can serve to complement or extend the frequently used 2-step procedure for detecting interaction effects.

Results from our multi-way epistasis simulation analyses and empirical data in cell lines suggest that genetic ancestry is a good proxy for genetic background, since all highly differentiated SNPs across the genome will be correlated with genetic ancestry. Our simulations also demonstrated that genetic ancestry can be a good proxy for an environmental covariate depending on the correlation between the two. However, it may be the case that there are multiple environmental factors interacting with a genetic locus, all of which are correlated with θ in differing degrees and effect sizes. Such a situation would mirror what we saw in our multi-way $G\times G$ simulations where a single interaction may not be detectable by using a traditional $G\times E$ test, but because θ aggregates the effects of all interacting covariates, AITL would be able to detect it. There are also other contexts in which modeling SNP by θ may be useful, such as using variance components. For example, SNP by θ interaction terms can be used in a mixed-model framework to test for interaction effects because genetic ancestry is correlated with many genetic markers and environmental covariates(Yang, Yang, et al., 2010).

For some traits, there may be systematic differences between ancestral populations in the genetic effects on the trait. In admixed individuals with these ancestral populations, the effect of genetic variation on phenotype will be reflected in the correlation between phenotype and θ, thereby affecting epistatic and $G\times E$ interactions. It will be interesting to see how much of the phenotype-ancestry correlations are due to epistatic and $G\times E$ interactions.

In our analysis of real data, we discovered gene by θ interactions associated with genes that have known interactions. In the GALA II Mexicans, the interaction of rs925736 with ancestry was associated with the methylation of HDAC4, a known histone deaceytlase (HDAC). In concert with DNA methylases, HDACs function to regulate gene expression by altering chromatin state(Z. D. Smith & Meissner, 2013). In Europeans, HDACs have been shown to be associated with lung function through direct genetic effects and through environmental interactions(Artigas et al., 2011; Liao, Lin, & Christiani, 2013). For the GALA II Puerto Ricans, rs17091085 showed an interaction associated with the methylation state of SERPINA6. Of note, interaction between birth weight and SERPINA6 has been previously associated with Hypothalamic-Pituitary-Adrenal axis function(L. N. Anderson et al., 2014). Further investigations of our interaction findings are thus warranted.

In the GALA II (methylation) dataset, two of the eight significant associations replicated and, in general, the results had an enrichment of low p-values in the replication dataset. However, we note that if the interactions detected by AITL are multi-way epistasis it is more likely that the results will replicate. This is because most SNPs differentiated in the Mexicans will still be differentiated in the Puerto Ricans, and thus still be correlated with θ. If the interactions detected by AITL are $G\times E$ interactions, then the interactions are less likely to replicate because the same environmental covariate(s) will need to be correlated with ancestry in both groups.

Another caveat is that the Mexicans and Puerto Ricans, though independent, are part of the same study and occasionally technical artifacts, such as issues with genotyping or measuring methylation, can affect downstream analyses of both populations. For our analyses, we have taken careful quality-control steps to ensure that this is not the case and there is no apparent inflation of test-statistics as demonstrated by our values for genomic control. Future research of interactions using AITL should keep such caveats in mind.

We investigated in detail the potential of single SNP-SNP interactions driving the results that were found both in the gene expression and methylation datasets. As demonstrated by the wide range of λ_{GC} values, we observed that non-linear effects can cause substantial departure from the null when testing for pairwise SNP-SNP interactions. This is especially true when testing for interaction between SNPs s and j, where s has a significant interaction with θ and j is correlated with covariates that are also correlated with θ. As we saw in the gene expression data, including the significant SNP by θ term can properly control for such situations, but its use in standard pairwise interaction tests warrants further investigation.

Our analysis revealed the existence of interactions but does not provide a direct way to determine the covariate that is interacting with a SNP. Further methodological work is required to uncover the exact environmental exposures or genetic loci with which SNPs are interacting. The existence of gene by θ interactions in GALA II underscores why modeling interactions should be considered for future association studies and for heritability estimation in admixed populations.

2. Partitioning Phenotype-Ancestry Correlations

Background

The prevalence of some common complex diseases differs between populations. For example, hypertension prevalence in men ranges from 3.4% in parts of rural North India to 50.1% in Japan(Kearney et al., 2005). Whether these differences are a consequence of genetic or environmental factors is an important question, potentially providing crucial insights into underlying disease mechanisms. Under the simplifying assumption that environmental exposures are constant, genetic risk factors would be responsible for the differences between populations. However, as studies have shown, both genetic and environmental factors contribute to disease(Lichtenstein et al., 2000; Willett, 2002). Understanding their relative contribution has important implications for global health and precision medicine.

Studies of admixed populations provide a unique opportunity to deepen our knowledge about the relative contributions of genetic and environmental risk factors to differences in disease prevalence between populations. The genomes of admixed individuals are derived from multiple ancestral populations. For example, African Americans have ancestry from European and African populations while Hispanic Latinos have ancestry from Europeans, Africans, and Native Americans. Recent advances in technology have allowed us to infer the ancestry of an individual's genome with extremely high accuracy and resolution(Alexander et al., 2009; Baran, Pasaniuc, Sankararaman, Torgerson, Gignoux, Eng, Rodríguez-Cintrón, Chapela, Ford, Avila, Rodriguez-Santana, Burchard, & Halperin, 2012; Maples, Gravel, & Kenny, 2013; Price et al., 2006; Pritchard, Stephens, & Donnelly, 2000). As a result, admixed populations provide an opportunity to analyze the influence of population differences in genetic variation on phenotype, but confined to a more consistent environment. Specifically, data from admixed populations in a fixed geographical locale provides genetic variation derived from continental populations, but with environmental differences limited to relatively small variation. In admixed populations, the extent to which genetic variation drives phenotypic differences between ancestral populations will induce a correlation between phenotype and the proportion of genetic ancestry from a given ancestral group (θ). As an example, consider African Americans who have ancestry from African and European ancestors. If a single nucleotide polymorphism (SNP) increases disease risk and occurs at an 80% frequency in Africans and a 20% frequency in Europeans, then African Americans with higher proportions of African ancestry will be at higher risk and will have a higher disease prevalence than those with lower proportions of African ancestry. This will result in a correlation between θ and disease status. However, θ has also been shown to be correlated with various environmental covariates. In the context of breast cancer, θ has been shown to be correlated with education, alcohol intake, and daily caloric intake(Fejerman et al., 2008; Ziv et al., 2006). Other environmental and biomedical covariates, such as socio-economic status and lung function, have also been shown to be correlated with θ(Burchard, Ziv, Coyle, & Gomez, 2003; Cheng et al., 2012; Choudhry et al., 2006; Florez et al., 2011; Kumar et al., 2010; 2013; Price et al., 2008; Reiner et al., 2007; Sanchez et al., 2010; Shaffer et al., 2007). Thus, it is unclear how much of a given phenotype-ancestry correlation is actually driven by the genetic component.

In this work, we present a novel statistical method to partition phenotype-ancestry correlations (2PAC) into genetic and environmental components. We show analytically and via extensive simulations that our approach provides unbiased estimates of the genetic and environmental contributions to the correlation between ancestry and phenotype. We further show that existing methods for estimating heritability in admixed populations are biased when ancestry is correlated with phenotype and our approach corrects this bias. We apply our method to admixed populations and phenotypes in the Bio*Me* Biobank at the Icahn School of Medicine at Mount Sinai to investigate the relationship between genetic ancestry and disease risk. We analyzed self-reported African American (AA; n=3,705) and Hispanic/Latino (HL; n=5,104) participants in Bio*Me* using phenotype data from electronic health records (EHRs) linked to genotype data. EHR data is comprised of over 14,000 medical billing codes (ICD-9), which classify diseases, injuries, and health encounters. Analyzing the full medical phenome, we found 18, 3 known and 15 novel, significant correlations between ancestry and ICD-9 based phenotypes including asthma ($p<3.0x10^{-3}$; HL), hypertension ($p<1.0x10^{-4}$; HL), cardiac dysrhythmias ($p<0.02$; HL), and anemia ($p<1.0x10^{-4}$; AA). For each of these associations we estimate the contribution of genetic and environmental factors to the phenotype-ancestry correlation and show that the contributions vary between phenotypes. We discuss the implications of these findings for medical research and clinical practice.

Materials and Method

We begin with an example to demonstrate the intuition behind 2PAC. For simplicity sake, we consider a 2-way admixed population, such as African Americans, with ancestral populations P_1 and P_2, and θ the proportion of ancestry from P_1. Suppose that there is a disease-causing SNP that occurs at 80% frequency in P_1 and 5% frequency in P_2. Then individuals with higher values of θ will have a higher disease prevalence than those with lower values of θ. This is because individuals with higher values of θ are more likely to have the disease causing SNP from P_1 and as a result the disease. This will result in a correlation between disease status and θ

as well as disease risk and θ. Now consider the case where the disease is also influenced by an environmental factor such as diet. If diet is correlated with θ this will also induce a correlation between disease status and θ.

In summary, the genetic portion of the correlation will be a function of the genetic effects and the differences in genetics between the ancestral populations. The environmental portion will be a function of the correlation between environmental and θ. To estimate the genetic component of the correlation, we need to be able to estimate accurately, in aggregate, the genetic effects and the differences between allele frequencies (i.e. the level of differentiation) between the ancestral populations. For the environmental component of the correlation, we need to be able to estimate the effect of environment, correlated with θ, on the disease without also capturing the genetic effect simultaneously. We describe our approach to partitioning phenotype-ancestry correlations (2PAC) in detail below.

Correlation as a Function of Genetics and Environment

Consider an admixed individual *i* who derives his or her genome from *k* ancestral populations. We denote individual *i*'s global ancestry proportion as $\theta_i = \langle \theta_{i1}, \theta_{i2}, \dots, \theta_{ik} \rangle$, where $\sum_{j=1}^{k} \theta_{ij} = 1$ and the genotype at a SNP *s* as $g_{is} \in \{0, 1, 2\}$. We model the phenotype of individual *i* as:

$$y_i = \sum_{s} \beta_s g_{is} + \beta_\theta \theta_{ij} + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, β_s is the effect size at SNP *s*, β_θ is the effect size of environmental factors correlated with the *j*-th component of global ancestry.

Let $\boldsymbol{y}$ be the vector of all individuals' phenotypes, $\boldsymbol{\theta_j}$be the vector of the corresponding *j*-th component of global ancestry, and $\boldsymbol{g_s}$ be the vector of corresponding genotypes at SNP *s*. Then by definition

$$cor(\boldsymbol{y},\boldsymbol{\theta}_j) = \frac{cov(\boldsymbol{y},\boldsymbol{\theta}_j)}{\sigma_y\sigma_{\theta_j}}$$

We begin by considering the case where there is no genetic effect, i.e. $\beta_s = 0\ \forall s$. If the effects at all SNPs are equal to zero, it then follows that

$$cov(\boldsymbol{y},\boldsymbol{\theta}_j) = cov(\beta_\theta\boldsymbol{\theta}_j + \boldsymbol{\varepsilon},\boldsymbol{\theta}_j) = \beta_\theta cov(\boldsymbol{\theta}_j,\boldsymbol{\theta}_j) + cov(\boldsymbol{\varepsilon},\boldsymbol{\theta}_j) = \beta_\theta\sigma_{\theta_j}^2$$

$$\sigma_y^2 = var(\beta_\theta\boldsymbol{\theta}_j + \boldsymbol{\varepsilon}) = var(\beta_\theta\boldsymbol{\theta}_j) + var(\boldsymbol{\varepsilon}) = \beta_\theta^2\sigma_{\theta_j}^2 + \sigma_\varepsilon^2$$

$$\Rightarrow\ cor(\boldsymbol{y},\boldsymbol{\theta}_j) = \frac{\beta_\theta\sigma_{\theta_j}^2}{\sigma_{\theta_j}\sqrt{\beta_\theta^2\sigma_{\theta_j}^2 + \sigma_\varepsilon^2}} = \frac{\beta_\theta^2\sigma_{\theta_j}^2}{\beta_\theta^2\sigma_{\theta_j}^2 + \sigma_\varepsilon^2}$$

We now consider the case where there are genetic effects at all or some of the SNPs. This then implies that

$$\sigma_y^2 = var\left(\sum_s \beta_s\boldsymbol{g}_s\right) + var(\beta_\theta\boldsymbol{\theta}_j) + var(\boldsymbol{\varepsilon}) + 2cov\left(\sum_s \beta_s\boldsymbol{g}_s\,,\beta_\theta\boldsymbol{\theta}_j\right)$$

$$+\ 2cov\sum_{s\neq t}(\beta_s\boldsymbol{g}_s,\beta_t\boldsymbol{g}_t)$$

$$= \sum_s \beta_s^2\sigma_{g_s}^2 + \beta_\theta^2\sigma_{\theta_j}^2 + \sigma_\varepsilon^2 + 2\beta_\theta\sum_s \beta_s cov(\boldsymbol{g}_s,\boldsymbol{\theta}_j) + 2cov\sum_{s\neq t}(\beta_s\boldsymbol{g}_s,\beta_t\boldsymbol{g}_t)$$

We note that if effect sizes are independent, then $2\beta_\theta\sum_s \beta_s cov(\boldsymbol{g}_s,\boldsymbol{\theta}_j) + 2cov\sum_{s\neq t}(\beta_s\boldsymbol{g}_s,\beta_t\boldsymbol{g}_t)$ is equal to zero in expectation and we are left with the standard assumptions of a linear mixed model.

$$cov(\boldsymbol{y},\boldsymbol{\theta}_j) = cov\left(\sum_s \beta_s\boldsymbol{g}_s,\boldsymbol{\theta}_j\right) + cov(\beta_\theta\boldsymbol{\theta}_j,\boldsymbol{\theta}_j) + cov(\boldsymbol{\varepsilon},\boldsymbol{\theta}_j) = \sum_s \beta_s cov(\boldsymbol{g}_s,\boldsymbol{\theta}_j) + \beta_\theta\sigma_{\theta_j}^2$$

Note that at a SNP s

$$cov(\boldsymbol{g}_s, \boldsymbol{\theta}_j) = E[\boldsymbol{g}_s\boldsymbol{\theta}_j] - E[\boldsymbol{g}_s]E[\boldsymbol{\theta}_j]$$

$$E[\boldsymbol{g}_s\boldsymbol{\theta}_j] = \int_{\theta_j=0}^{1} \sum_{g_s=0}^{2} P(g_s, \theta_j) g_s \theta_j \, d\theta_j = \int_{\theta_j=0}^{1} \sum_{g_s=0}^{2} P(g_s|\theta_j) P(\theta_j) g_s \theta_j \, d\theta_j$$

where $P(g_s|\theta_j) = \binom{2}{g_s} p^{g_s}(1-p)^{(2-g_s)}$ and p is the allele frequency at s.

$$\Rightarrow cov(\boldsymbol{g}_s, \boldsymbol{\theta}_j) = \left(\int_{\theta_j=0}^{1} \sum_{g_s=0}^{2} P(g_s|\theta_j) P(\theta_j) g_s \theta_j \, d\theta_j \right) - \mu_{g_s}\mu_{\theta_j}$$

This provides a closed form expression for the correlation between phenotype and genetic ancestry that is partitioned between genetic effects and environmental effects.

$$\rho_{y\theta} = cor(\boldsymbol{y}, \boldsymbol{\theta}_j) = \ cor_{genetics} + cor_{environment} = \rho_{g\theta} + \rho_{e\theta} = \frac{\sum_s \beta_s cov(\boldsymbol{g}_s, \boldsymbol{\theta}_j)}{\sigma_y \sigma_{\theta_j}} + \frac{\beta_\theta \sigma_{\theta_j}^2}{\sigma_y \sigma_{\theta_j}}$$

$$= \frac{\sum_s \beta_s \left[\left(\int_{\theta_j=0}^{1} \sum_{g_s=0}^{2} P(g_s|\theta_j) P(\theta_j) g_s \theta_j \, d\theta_j \right) - \mu_{g_s}\mu_{\theta_j} \right] + \beta_\theta \sigma_{\theta_j}^2}{\sigma_{\theta_j} \sqrt{\sum_s \beta_s^2 \sigma_{g_s}^2 + \beta_\theta^2 \sigma_{\theta_j}^2 + \sigma_\varepsilon^2 + 2\beta_\theta \sum_s \beta_s cov(\boldsymbol{g}_s, \boldsymbol{\theta}_j) + 2cov \sum_{s \neq t}(\beta_s \boldsymbol{g}_s, \beta_t \boldsymbol{g}_t)}}$$

Estimating Effect Sizes using Linear Mixed Models

Assuming independence of SNPs, we could estimate the genetic effect at each SNP using linear regression and use the marginal effect size estimates for the SNPs to estimate the total contribution to trait variance. However, due to linkage disequilibrium (LD), we use a linear mixed model (LMM) framework to estimate the effects of all SNPs jointly. Assuming that our data consists of n individuals and m SNPs, we first estimate a kinship matrix, $\boldsymbol{K} = \boldsymbol{Z}\boldsymbol{Z}^T/m$, where $\boldsymbol{Z}$ is a $n \times m$ matrix of genotypes normalized to have mean 0 and variance 1 for each SNP. Using restricted maximum likelihood we find the parameters, σ_g^2 and σ_ε^2, that maximize the model $\boldsymbol{y} \sim \mathcal{N}\left(0, \boldsymbol{K}\sigma_g^2 + \boldsymbol{I}\sigma_\varepsilon^2\right)$(Yang, Benyamin, et al., 2010). The solution to the fixed effects are given by $\widehat{\boldsymbol{\beta}}_{fixed} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y})$, where $\boldsymbol{X}$ is a $n \times c$ matrix of fixed covariates and $\boldsymbol{V} = \boldsymbol{K}\hat{\sigma}_g^2 + \boldsymbol{I}\hat{\sigma}_\varepsilon^2$. The best linear unbiased predictor (BLUP) of SNP effect sizes is given by $\widehat{\boldsymbol{\beta}}_{BLUP} = \frac{\hat{\sigma}_g^2}{m}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{fixed}\right)$. Since the genotypes are normalized to have mean 0 and variance 1, the effect sizes are given on the same scale. By dividing through by the vector of the observed genotype standard deviations, $\widehat{\boldsymbol{SD}}_g$, we can get an estimate of the effect sizes on the observed scale:

$$\widehat{\boldsymbol{\beta}}_{BLUP,OBS} = \frac{\widehat{\boldsymbol{\beta}}_{BLUP}}{\widehat{\boldsymbol{SD}}_g}$$

BLUP SNP effect estimates are shrunk towards zero and therefore we need to multiply it by $\frac{1}{\widehat{shrink}}$ to obtain our final effect size estimates. We describe how to estimate $\widehat{shrink}$ in the following section.

Estimating the Shrink

In order to fit all SNPs jointly, the LMM framework shrinks the SNP effect sizes towards 0 relative to the marginal effect sizes. Consider the case of independent SNPs, where the marginal effect sizes provide unbiased estimates of the SNP effect sizes. The shrink applied to the BLUP SNP effects is, in this case, proportional to the estimated narrow sense heritability, $\hat{h}^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_\varepsilon^2}$, and is given by $\left(\frac{\hat{h}^2}{\hat{h}^2 + \frac{m}{n}}\right)$ (Vilhjálmsson et al., 2015). LMM assumes that $\sigma_g^2 = \sum_s \beta_s^2$, given that the genotypes are scaled and centered, and thus $\hat{\sigma}_g^2$ is an estimate of the sum of the squared true effect sizes. We propose the ratio, $\frac{\sum_s \hat{\beta}_{s,BLUP}^2}{\hat{\sigma}_g^2}$, to estimate the shrink. We show the intuition for why it works below, and provide simulations to show that it works in practice as well in the presence of LD (see Results).

For simplicity, assume that we have independent SNPs, $\boldsymbol{\beta} \sim \mathcal{N}\left(0, \frac{\sigma_g^2}{m}\right)$, and that the phenotype is given by $\boldsymbol{y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{Z}$ is the matrix of normalized genotypes. We can estimate the marginal effect at a SNP *s* by $\hat{\beta}_{s,OLS} = (\boldsymbol{Z}_s^T \boldsymbol{Z}_{\boldsymbol{s}})^{-1}(\boldsymbol{Z}_s^T \boldsymbol{y})$. The marginal estimates are unbiased estimates because SNPs are independent. The mean and variance of the estimate is then given by

$$E\left[\hat{\beta}_{s,OLS}\right] = E\left[\frac{1}{n}\boldsymbol{Z}_s^T \boldsymbol{Z}_s \beta_s\right] = \beta_s$$

$$var\left(\hat{\beta}_{s,OLS}\right) = E\left[\frac{1}{n}\boldsymbol{Z}_s^T \boldsymbol{Z}_s \beta_s + \frac{1}{n}\sum_{j \neq s} \boldsymbol{Z}_s^T \boldsymbol{Z}_j \beta_j + \frac{1}{n}\boldsymbol{Z}_s^T\ \boldsymbol{\varepsilon} - E[\beta_s]\right]$$

$$= E\left[\frac{1}{n}\sum_{j \neq s} \boldsymbol{Z}_s^T \boldsymbol{Z}_j \beta_j + \frac{1}{n}\boldsymbol{Z}_s^T\ \boldsymbol{\varepsilon}\right]^2 = E\left[\frac{1}{n^2}\sum_{j \leq s} \left(\boldsymbol{Z}_s^T \boldsymbol{Z}_j \beta_j\right)\left(\boldsymbol{Z}_s^T \boldsymbol{Z}_j \beta_j\right)^T\right] + E\left[\frac{1}{n^2}(\boldsymbol{Z}_s^T\ \boldsymbol{\varepsilon})(\boldsymbol{Z}_s^T\ \boldsymbol{\varepsilon})^T\right]$$

$$= E\left[\frac{1}{n^2}\left(\boldsymbol{Z}_s^T\left(\sum_{j \leq s} \beta_j^2 \boldsymbol{Z}_j \boldsymbol{Z}_j^T\right)\boldsymbol{X}_s\right)\right] + E\left[\frac{1}{n^2}(\boldsymbol{Z}_s^T\ \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \boldsymbol{Z}_s)\right]$$

$$= E\left[\frac{1}{n^2}\left(\boldsymbol{Z}_s^T\left((m-1)\boldsymbol{I}_m\right)\boldsymbol{Z}_s E[\beta_s^2]\right)\right] + \ E\left[\frac{1}{n^2}(\boldsymbol{Z}_s^T\ (1-h^2)\boldsymbol{I}_n \boldsymbol{Z}_s)\right]$$

$$= E\left[\frac{1}{n^2}(n(m-1)\boldsymbol{I}_m E[\beta_s^2])\right] + E\left[\frac{1}{n^2}(m(1-h^2)\boldsymbol{I}_n)\right] = \frac{h^2 - \frac{h^2}{m} + 1 - h^2}{n}$$

$$= \frac{1 - \frac{h^2}{m}}{n} \approx \frac{1}{n}$$

Since $\hat{\beta}_{s,OLS}$ given $\boldsymbol{\beta}$ is normally distributed with the mean and variance we derived above, we can derive the posterior distribution of $\boldsymbol{\beta}$ given the estimate $\hat{\beta}_{s,OLS}$ by using a normal-normal conjugate prior update. We have the prior $\boldsymbol{\beta} \sim \mathcal{N}\left(0, \frac{\sigma_g^2}{m}\right) = \mathcal{N}(\mu_0, \sigma_0^2)$ and the likelihood $(\widehat{\boldsymbol{\beta}}_{s,OLS} | \boldsymbol{\beta}) \sim \mathcal{N}\left(\boldsymbol{\beta}, \frac{1-\frac{h^2}{m}}{n}\right)$ which gives us the following posterior mean and variance

$$E\left[\widehat{\boldsymbol{\beta}}_{s,BLUP}\right] = E\left[\boldsymbol{\beta}_s | \widehat{\boldsymbol{\beta}}_{s,OLS}\right] = \widehat{\boldsymbol{\beta}}_{s,OLS}\ (shrink)$$

$$var\left(\widehat{\boldsymbol{\beta}}_{s,BLUP}\right) = var\left(\boldsymbol{\beta}_s | \widehat{\boldsymbol{\beta}}_{s,OLS}\right) = \frac{1}{n}(shrink)$$

where $shrink = \left(\frac{\widehat{h}^2}{\widehat{h}^2 + \frac{m}{n}}\right)$. We then have

$$\sigma_g^2 = \sum_s \beta_s^2 \approx \hat{\sigma}_g^2 = \sum_s \left(\frac{1}{\widehat{shrink}} * \hat{\beta}_{s,BLUP}\right)^2 = \left(\frac{1}{\widehat{shrink}}\right)^2 \sum_s \hat{\beta}_{s,BLUP}^2$$

$$\Rightarrow \left(\widehat{shrink}\right)^2 = \frac{\sum_s \hat{\beta}_{s,BLUP}^2}{\hat{\sigma}_g^2}$$

Plugging in the variance of the BLUP betas shows us why this ratio is approximately the shrink.

$$\left(\widehat{shrink}\right)^2 = \frac{\sum_s \hat{\beta}_{s,BLUP}^2}{\hat{\sigma}_g^2} = \frac{var\left(\widehat{\boldsymbol{\beta}}_{BLUP}\right)}{\hat{\sigma}_g^2} = \frac{\frac{m}{n}\left(\widehat{shrink}\right)}{\hat{\sigma}_g^2}$$

$$\Rightarrow \frac{\sum_s \hat{\beta}_{s,BLUP}^2}{\hat{\sigma}_g^2} = \widehat{shrink}$$

Simulation Framework

For all of our non-real data simulations, we simulated 500 2-way admixed individuals and 750 SNPs. Global ancestry for ancestral population 1 (θ_1) was drawn from a truncated normal distribution with $\mu = 0.7$ and $\sigma = 0.2$. The boundaries of the truncated distribution were set at a 0 and 1. For each individual i and every SNP s, we assigned the local ancestry or the number of alleles derived from population 1 (γ_{ais}) for each haplotype by performing two binomial trials with the probability of success equal to θ_{i1}. We then drew ancestry specific allele frequencies following the Balding-Nichols model by assuming a $F_{ST} = 0.2$ and drawing two ancestral frequencies, p_{1s} and p_{2s}, from the following beta distribution.

$$p_{1s}, p_{2s} \sim Beta\left(\frac{p_s(1 - F_{ST})}{F_{ST}}, \frac{(1 - p_s)(1 - F_{ST})}{F_{ST}}\right)$$

where p_s is the underlying MAF in the entire population and is drawn from a uniform distribution with bounds 0.01 and 0.5. Genotypes were drawn using a binomial trial for each local ancestry haplotype with the probability of success equal to p_{1s} or p_{2s} for values of $\gamma_{ais} = 0$ or 1, respectively.

For both non-real and real data simulations, we drew effect sizes in 3 ways to generate phenotypes that were correlated with ancestry in various degrees as a function of genetics and environment. In the first approach ("low") SNP effect sizes were drawn from a normal distribution with $\mu = 0$ and $\sigma = 0.2$. Next, we used a probabilistic approach ("mid") that ties the direction of the effect size to the ancestral minor allele frequency. For a SNP s, we performed a binomial trial where the odds of success is $\frac{p_{1s}}{p_{2s}}$. If the trial was a success, the effect size is positive and otherwise the effect size is negative. Last, we used a deterministic approach ("high") where if $p_{1s} > p_{2s}$ then the effect size is positive and otherwise negative. These scenarios induce low, mid, and high correlations between phenotype and ancestry, respectively. In the scenario that there was an environmental effect for covariates correlated with ancestry, we used $\beta_\theta = 25$. To set the heritability of our simulated phenotypes at a specified level, we added a noise term, ε, drawn from a normal distribution with $\mu = 0$ and $\sigma_\varepsilon = \sqrt{\sigma_{g_s}^2\left(\frac{1}{h^2} - 1\right)}$. This way $h^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_\varepsilon^2}$ where $\sigma_g^2 = var(\sum_s \beta_s \boldsymbol{g_s})$.

Filtering for Related Individuals

All analyses in real data were filtered for related. To filter for related individuals, we estimated kinship coefficients between all pairs of individuals using REAP. We defined two individuals as related if they had a kinship coefficient greater than 0.025. For a pair of related individuals, we removed the one with a greater number of other individuals to whom he or she was related. In the case of a tie, we removed one of the pair at random.

BioMe Data

We analyzed data from the *BioMe* biobank in Mount Sinai Medical Center in New York City. The *BioMe* Biobank at Mt. Sinai is composed of >32,000 multi-ethnic patients with phenotype data from electronic health records (EHRs) linked to genotype data from the Illumina OmniExpress platform. After filtering for genotypes with a minor allele frequency ≥ 0.05 and a genotyping rate ≥ 0.01, we had about 610,000 SNPs left in our analysis. The individuals analyzed here were admixed patients consisting of 3,705 African Americans and 5,104 Hispanic/Latinos. Global ancestry inference was done using ADMIXTURE(Alexander et al., 2009). The health records contained information about the ICD9 billing codes that individuals had, measurements on various continuous traits (i.e. white blood cell count, creatinine, cholesterol), and information on age, sex, and body mass index (BMI). For continuous phenotypes, we limited our analyses to those that had at least 1000 samples with measurements for a given trait, which resulted in 48 phenotypes for African Americans and 49 for Hispanic/Latinos. For binary phenotypes, we restricted our analyses to ICD9 codes that had at least 20 individuals with a given ICD9 code. We treated each ICD9 code as a case/control phenotypes where individuals with the ICD9 code were assigned a case status and they were assigned a control status otherwise. We also ignored ICD9 codes that had to do with injury caused by an external event. In total we looked at 1182 ICD9 code based phenotypes for African Americans and 1458 for Hispanic/Latinos.

BioMe Permutations

We explain the general framework here for our permutation approach. For simplicity, we leave out extra covariates, but we note that covariates may be included when testing the association between phenotype and ancestry. Let $\boldsymbol{Y}$ be an $n \times p$ matrix of phenotypes for n individuals and p phenotypes. Then we can test the association between $\boldsymbol{Y}_p,$ for $p \in \{1...P\}$ and $\boldsymbol{\theta}_j$ by using linear regression for continuous phenotypes and logistic regression for binary phenotypes. However, if correlations exist within $\boldsymbol{Y}$ between different phenotypes, then we must preserve these correlations when we permute the data to generate our empirical null distribution. For each permutation u, we generate $\boldsymbol{Y_u}$ by permuting the rows of $\boldsymbol{Y}$ and then perform the regression $\boldsymbol{Y}_{u,p} \sim \hat{\beta}_{u,p,\theta_j} \boldsymbol{\theta}_j \; \forall p$. We then include the maximum Wald test statistic $\left(\frac{\hat{\beta}_{u,p,\theta_j}}{SE(\hat{\beta}_{u,p,\theta_j})}\right)^2 \forall p$ in our empirical null distribution. We repeat this process 10,000 times to generate a null distribution composed of 10,000 chi-squared test statistics. We then compare the original observed test statistic $\left(\frac{\hat{\beta}_{p,\theta_j}}{SE(\hat{\beta}_{p,\theta_j})}\right)^2$ to the null distribution and compute the p-value as

$\frac{r+1}{q+1}$, where r is the number of permutation test statistics that are greater than or equal to the original statistics, and q is the number of permutations.

Results

Simulated Data

To determine how 2PAC behaves under the simple scenario of independent SNPs, we performed 1000 simulations for the "low", "mid", and "high" scenarios that we describe in *Simulation Framework.* For each scenario we simulated phenotypes for h^2 values of 0.35 and 0.65. Furthermore, for every scenario and heritability combination we also tested 2PAC when there is and is not an environmental effect, for covariates correlated with ancestry, by setting β_θ equal to 25 and 0, respectively. We show the results for $h^2 = 0.65$ for brevity, but the conclusions we drew from the results when $h^2 = 0.35$ were the same.

We first wanted to see if we could obtain unbiased estimates for the genetic component of the correlation of phenotype and ancestry ($\rho_{g\theta}$). As shown in Table 1, we were able to obtain accurate and unbiased estimates of $\rho_{g\theta}$ in the case of independent SNPs regardless of whether there was an environmental effect or not.

Table 1. Mean Estimates of $\rho_{g\theta}$ for $h^2 = 0.65$ with and without an environmental effect

	$\beta_\theta = 25$		$\beta_\theta = 0$	
	Truth (SE)	2PAC (SE)	Truth (SE)	2PAC (SE)
Low	-0.006 (0.004)	-0.010 (0.008)	0.003 (0.006)	-0.0004 (.011)
Mid	0.448 (0.001)	0.446 (0.010)	0.669 (0.001)	0.663 (0.006)
High	0.572 (0.004)	0.567 (0.004)	0.756 (0.001)	0.759 (0.006)

In these simulations, we also saw that β_θ was overestimated as a function of the correlation between phenotype and genetic ancestry ($\rho_{y\theta}$), whether it be with linear mixed models or with ordinary least squares. As shown in Table 2, by obtaining an unbiased estimate of $\rho_{g\theta}$ we can recover an unbiased estimate of β_θ by rewriting the equation for $\rho_{y\theta}$ as

$$\beta_\theta = \left(\frac{\rho_{y\theta} - \rho_{g\theta}}{\sigma^2_{\theta_j}}\right)\sigma_y\sigma_{\theta_j}$$

Table 2. Mean Estimates of β_θ for $h^2 = 0.65$ with and without an environmental effect

	$\beta_\theta = 25$		$\beta_\theta = 0$	
	LMM (SE)	2PAC (SE)	LMM (SE)	2PAC (SE)
Low	25.0 (0.106)	25.2 (0.201)	0.073 (0.107)	0.115 (.210)
Mid	46.5 (0.150)	23.7 (0.275)	21.4 (0.146)	-1.07 (0.259)
High	67.9 (0.177)	25.3 (0.386)	42.8 (0.165)	-0.286 (0.379)

Furthermore when $\rho_{y\theta} > 0$, there is a downward bias in heritability when estimating h^2 using a LMM approach such as GCTA(Yang, Yang, et al., 2010). In a standard LMM framework, it is generally assumed that $\beta_s \sim \mathcal{N}(\mu = 0, \sigma^2)$. Recall that

$$\sigma_g^2 = \sum_s \beta_s^2 \sigma_{g_s}^2 + 2cov \sum_{s \neq t} (\beta_s \boldsymbol{g_s}, \beta_t \boldsymbol{g_t})$$

but if $\beta_s \sim \mathcal{N}(\mu = 0, \sigma^2)$ then $E[2cov \sum_{s \neq t}(\beta_s \boldsymbol{g_s}, \beta_t \boldsymbol{g_t})] = 0$ and thus $\sigma_g^2 = \sum_s \beta_s^2 \sigma_{g_s}^2$, which is exactly what is modeled by LMM. However, when there is selection on a trait in a population, or $\rho_{y\theta} > 0$, then it is unlikely that $E[2cov \sum_{s \neq t}(\beta_s \boldsymbol{g_s}, \beta_t \boldsymbol{g_t})] = 0$ because causal loci will have correlated effect sizes(Edge & Rosenberg, 2015). This is analogous to what happens in the "mid" and "high" scenarios. We see that in the scenarios that induce $\rho_{y\theta} > 0$, a standard LMM fails to account for the covariance between the effects of different genetic loci. As a result, the estimates of σ_g^2 and h^2 are downwardly biased. It is also straightforward to see that because we can obtain unbiased estimates of β_θ and $\rho_{g\theta}$, we can also recover unbiased estimates of heritability by rewriting σ_y^2

$$\sigma_g^2 = \sigma_y^2 - \beta_\theta^2 \sigma_{\theta_j}^2 - \sigma_\varepsilon^2 - 2\beta_\theta \sum_s \beta_s cov(\boldsymbol{g_s}, \boldsymbol{\theta_j}) = \sigma_y^2 - \beta_\theta^2 \sigma_{\theta_j}^2 - \sigma_\varepsilon^2 - 2\beta_\theta \rho_{g\theta} \sigma_\theta \sigma_y$$

We show how the bias in heritability estimates increase as the correlation between ancestry and phenotype increase in Figure 1.

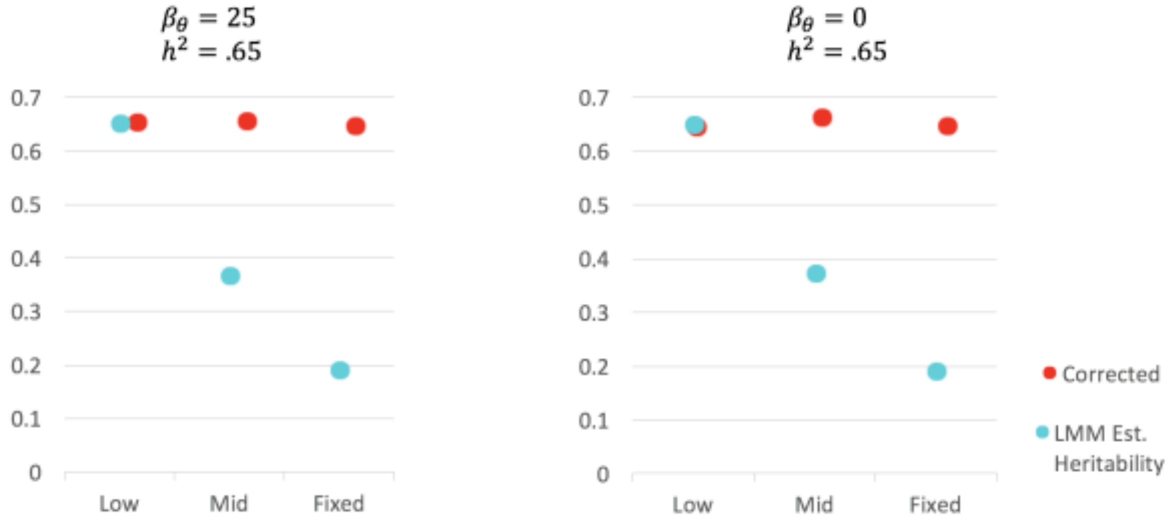


Figure 1. Corected heritability estimates after using 2PAC to correct the estimate of σ_g^2. The average phenotype-ancestry correlation due to genetics, $\rho_{g\theta}$, was -0.006 for "low", 0.448 for "mid", and 0.572 for "high".

Simulations using Real Data

We also wanted to determine how 2PAC behaves when there is LD between SNPs. We simulated phenotypes as described in *Simulation Framework* but for genotypes we used chromosomes 1-3 from the African Americans in *BioMe.* As we did with simulated genotypes, we performed 200 simulations under the "low", "mid", and "high" scenarios at a heritability of 0.35. Again, we tested 2PAC when there is an environmental effect, for covariates correlated with ancestry, by setting β_θ equal to 225. We chose these values of β_θ to allow the environmental effect to account for at least 10% of the heritability on average during the simulations.

Admixture will induce long range LD between SNPs due to allele frequency differences in the mixing populations(X. Zhu, Tang, & Risch, 2008). We account for ancestry induced LD (ALD) by adjusting the genotypes for ancestry using the linear regression $\boldsymbol{g}_s \sim \beta_{s,\theta}\boldsymbol{\theta}_{AFR}$. We used the residuals of this regression to generate the kinship matrix $\boldsymbol{K}_{res}$ and the residualized genotype matrix $\boldsymbol{Z}_{res}$ (scaled and centered after taking the residuals). We used these matrices to compute the BLUPs for estimating the shrink. Not adjusting for ALD results in overestimating the shrink by a very large margin as shown in Table 3.

Table 3. Mean and Median Estimates of the shrink for $h^2 = 0.35$ using Regular BLUP betas

	$\beta_\theta = 225$		$\beta_\theta = 0$	
	True (SE)	Estimated (SE)	True (SE)	Estimated (SE)

Low	Mean = 1.89 (0.668)
	Median = 0.342
	n=176
		Mean = 120.8 (1.65)
		Median = 116.5
			Mean = 1.17 (0.559)
			Median = 0.687
			n=183
				Mean = 123.2 (1.59)
				Median = 13.2

Mid	Mean = 23.3 (1.41)
	Median = 17.5
	n=169
		Mean = 159.6 (3.58)
		Median = 148.8
			Mean = 18.9 (0.146)
			Median = 16.0
			n=155
				Mean = 153.6 (3.39)
				Median = 144.9

High	Mean = 93.3 (20.3)
	Median = 62.3
	n=177
		Mean = 615.9 (156)
		Median = 529
			Mean = 116.3 (16.0)
			Median = 72.1
			n=175
				Mean = 961.8 (135)
				Median = 587.5

Using the residualized genotypes, we could recover accurate estimates of the shrink, although they were not unbiased. We note that GCTA did not converge when estimating σ_g^2 for all simulations. In Table 4 we provide the mean and median estimates of the shrink. We define here the true shrink as $\frac{\rho_{g\theta}}{\hat{\rho}_{g\theta,shrunk}}$, where $\hat{\rho}_{g\theta,shrunk}$ is computed using the regular non-residualized BLUP betas. We include the median shrink over the simulations because there were outliers which influenced the mean. In general, the median values are generally closer to each other than the mean values. However, even with the outliers, the mean true and estimated values are very close after factoring in standard error in the case when there is a correlation between phenotype and ancestry (i.e. "mid" and "high"). In the case of the "low" scenario, although the estimated shrink is very large compared to the truth, it makes very little difference in the estimation of $\rho_{g\theta}$ because $\hat{\rho}_{g\theta,shrunk}$ is already very small number (see Table 5).

Table 4. Mean and Median Estimates of the shrink for $h^2 = 0.35$ using Residualized BLUP betas

	$\beta_\theta = 225$		$\beta_\theta = 0$	
	True (SE)	Estimated (SE)	True (SE)	Estimated (SE)

Low	Mean = 1.89
	(0.668)
	Median = 0.342
	n=176
		Mean = 13.3
		(0.179)
		Median = 12.8
			Mean = 1.17
			(0.559)
			Median = 0.687
			n=183
				Mean = 13.6
				(0.174)
				Median = 13.2

Mid	Mean = 23.3
	(1.41)
	Median = 17.5
	n=169
		Mean = 17.5
		(0.391)
		Median = 16.3
			Mean = 18.9
			(0.146)
			Median = 16.0
			n=155
				Mean = 16.8
				(0.370)
				Median = 15.98

High	Mean = 93.3
	(20.3)
	Median = 62.3
	n=177
		Mean = 67.4
		(17.1)
		Median = 57.9
			Mean = 116.3
			(16.0)
			Median = 72.1
			n=175
				Mean = 105.1
				(14.8)
				Median = 64.3

As shown in Table 5, we were able to get unbiased estimates of $\rho_{g\theta}$ when there was LD between SNPs using our approach to estimate the shrink (see *Estimating the Shrink*). This was both in the case when there was an effect of environment and when there was not. This also implies that we would be able to recover unbiased estimates of β_θ, σ_g^2, and h^2.

Table 5. Mean Estimates of $\rho_{g\theta}$

	$\beta_\theta = 225$		$\beta_\theta = 0$	
	True (SE)	Estimated (SE)	True (SE)	Estimated (SE)
Low	0.003 (0.004)	0.011 (0.013)	0.002 (0.005)	0.011 (0.021)
Mid	-0.207 (0.002)	-0.206 (0.009)	-0.311 (0.004)	-0.327 (0.015)
High	-0.461 (0.001)	-0.442 (0.013)	-0.561 (0.005)	-0.561 (0.016)

Analysis of BioME ICD9 Codes

Since we were unsure where we would find correlations between the human phenome and genetic ancestry, we used the biobank, *BioMe.* Using a biobank allowed us to analyze a large number of human phenotypes for significant associations with genetic ancestry. We performed our analyses on both continuous and ICD9 code based binary phenotypes for a cohort of African American and Hispanic/Latinos (see *BioMe Data*). We started by inferring proportions of European, African, and Native American ancestry (θ_{EUR}, θ_{AFR}, and θ_{NAM}) for all individuals. For each phenotype, *p*, we the performed regression $\boldsymbol{y}_p \sim \hat{\beta}_{p,\theta_j}\boldsymbol{\theta}_j + \hat{\beta}_{p,BMI}\boldsymbol{BMI} + \hat{\beta}_{p,AGE}\boldsymbol{AGE} + \hat{\beta}_{p,SEX}\boldsymbol{SEX},\ \forall j \in \{EUR, AFR, NAM\}$, using linear regression for continuous traits and logistic regression for binary traits. Since biobank phenotypes may be correlated due to billing practices, we used permutations to create a null distribution of test statistics to assess the significance of $\hat{\beta}_{p,\theta_j}$ (see *BioMe Permutations*). In summary, we found 4 significant phenotype-ancestry correlations in the African American data and 18 in the Hispanic/Latino data (Tables 6 and 7). With the significant associations, we also applied 2PAC to partition the phenotype-ancestry correlation into estimates of the genetic and environmental components $\left(\hat{\rho}_{g\theta}, \hat{\rho}_{\varepsilon\theta}\right)$. We used

GCTA to compute the estimate $\hat{\sigma}_g^2$ and compute the BLUP betas. However, due to sample sizes we were unable to partition the correlations with any meaningful resolution due to large standard errors, which were obtained via jackknifing by leaving one individual out and re-computing the $\hat{\rho}_{g\theta}$ n times. Instead, we tested for evidence that the correlation between phenotype and ancestry is driven by genetics using the chi-squared (*df* = 1) Wald test $\left(\frac{\hat{\rho}_{g\theta}}{SE(\hat{\rho}_{g\theta})}\right)^2$. We found that in Hispanic/Latinos, essential hypertension (ICD9=401.9) had evidence for a genetically driven phenotype-ancestry correlation with a p-value of 0.014. Essential hypertension is defined as high blood pressure without any known secondary causes such as renovascular disease or renal failure. Essential hypertension also accounts for about 95% of all hypertension cases and is thought to have both genetic and environmental factors which cause it(Carretero & Oparil, 2000).

Table 6. Phenotype-Ancestry Correlations for *BioMe* African Americans

Phenotype (ICD9)	Ancestry	Partial Correlation
Neutrophil Count	African	-0.161 (p=1e-4)
Neutrophil %	African	-0.119 (p=0.011)
White Blood Cell #	African	-0.142 (p=1e-4)
Anemia (285.9)	African	0.098 (p=2e-4)

Table 7. Phenotype-Ancestry Correlations for *BioMe* Hispanic/Latinos

Phenotype (ICD9)	Ancestry	Partial Correlation
Hematocrit	African	-0.076 (p=0.005)
Hemoglobin	African	-0.099 (p=1e-4)
Lymphocyte %	African	0.135 (p=1e-4)
Mean Hemoglobin (HGB)	African	-0.113 (p=1e-4)
Mean HGB Conc.	African	-0.147 (p=1e-4)
Mean Corpuscular Volume	African	-0.085 (p=4e-4)
Neutrophil %	African	-0.135 (p=1e-4)
RBC Distribution Width	African	0.102 (p=1e-4)
Triglycerides	African	-0.097 (p=4e-4)
Iron deficiency anemia (280.9)	African	0.069 (p=0.002)
Cardiac dysrhythmias (427.89)	African	0.063 (p=0.017)
Reaction to tuberculin skin test (795.51)	African	0.060 (p=0.033)
Anemia (285.9)	African	0.065 (p=0.008)
Uterine leiomyoma (218.9)	African	0.063 (p=0.014)
Hypertension (401.9)	African	0.065 (p=1e-4)
Asthma (493.9)	Native American	-0.071 (p=0.003)
Tobacco use disorder (305.1)	Native American	-0.092 (p=2e-4)
Hypertensive Renal Disease (403.91)	European	-0.062 (p=0.015)

Discussion

In previous studies, it has been unclear whether correlations between phenotype and genetic ancestry are due to effects from genetic or environmental risk factors, or a combination of both. In this study, we have developed a method, 2PAC, which partitions the correlation into genetic and environmental components. Being able to determine whether the correlation is driven by genetics or environment provides potential insight into underlying mechanisms for differences in prevalence for some complex traits. This is especially important in the context of precision medicine because it deepens our understanding of disease etiologies and proposes hypotheses on therapeutic strategies. Diseases that have a correlation with genetic ancestry driven mostly by the genetic component are likely to be targeted differently from those that have a correlation driven predominantly by environmental factors.

In this work, we used the LMM framework as implemented by GCTA and rely on its theoretical assumptions (Yang, Lee, Goddard, & Visscher, 2013). Others have shown that the model assumptions in LMM are critical, and deviations from these assumptions can affect convergence or stability of the method(Gamazon & Park, 2016; Speed, Cai, Johnson, Nejentsev, & Balding, 2017). If appropriate, 2PAC can be used together with any other LMM framework, such as LDAK(Speed, Hemani, Johnson, & Balding, 2012), as long as the BLUP betas and the shrink can be estimated accurately. Intuitively, using a LMM framework that is less sensitive to confounders, such as the LD structure between causal variants, would increase 2PAC's accuracy as well, although it remains to be shown. We have shown here that performing heritability estimates in admixed populations can, in certain scenarios, result in biased results that can be addressed using our approach. We have also shown that ALD has undesirable effects when estimating the BLUP betas and the shrink applied to them. It will be interesting to see if there is a way to directly account for potential heritability biases in admixed populations directly in the LMM framework, and if addressing LD issues in the theoretical framework can address long range LD induced by ancestry.

There are future extensions to 2PAC that will be interesting to implement in the future. One way 2PAC may be extended is to apply the method to identify regions of the genome that account for the largest proportion of $\rho_{g\theta}$. Identifying such regions of the genome may provide identify loci responsible for observed population level differences in phenotype. 2PAC may also be extended to interrogate phenotype-ancestry correlations for non-admixed populations, such as Europeans. It is unclear though whether population structure in Europeans will provide as much information as the genetic ancestry of admixed populations. This is because the amount of genetic differentiation in admixed populations is much higher than that of just Europeans. Our approach to estimate the shrink is also fundamentally interesting. It could be incorporated into all GCTA like analyses, of any population, to learn about the genetic architecture of disease because it indirectly estimates the percentage of causal variants in the genome for that phenotype.

In our analysis of phenotypes from *BioMe*, we were able to identify multiple significant correlations between phenotype and genetic ancestry in both African Americans and Hispanic/Latinos. We also found evidence that the correlation between essential hypertension and Native American ancestry is driven partially be genetics. The data analyzed here are from genotyping chips and the estimates we provide of $\rho_{g\theta}$ are, as a result, estimates of the contribution of genotyped SNPs to $\rho_{g\theta}$. There may be rare variation or untyped genetic variation that also contributes to the correlations we examined, and should be investigated in the future. As more hospitals collect genetic data and more precise records of their patients, we will be able to learn more about the causes of various diseases and phenotypes. The *BioMe* data we analyzed was mostly composed of case/control data based on ICD9 codes, which contain a wealth of information, but do not provide a refined picture of disease biology. In the future, 2PAC together with deep phenotyping could paint a more detailed picture of what is driving the genetic component of phenotype-disease correlations, and help us to understand at a deeper level underlying disease mechanisms (Delude, 2015; Robinson, 2012).

Conclusion

World-historical events such as the slave trade, colonization of the Americas, and mass migrations have resulted in the emergence of populations with complex patterns of population admixture. Individuals are now able to travel around the world with unprecedented ease thanks in part to advances in manufacturing, worldwide infrastructure, and the general increase in individuals' economic power. As a result, the emergence of new patterns of admixture is certain to occur and there is a need to address the gap in our understanding of how genetic variation influences trait variation, including disease risk, in human populations. This includes understanding whether the genetic architectures of complex traits are similar across populations, determining whether the contributions of genetic and environmental components differ between populations, and testing whether the results of genome studies that have been obtained for Europeans are generalizable to other populations. Without aiming for a more comprehensive account of human genetic variation, we risk falling short of the goal of precision medicine and providing the promise of genomic medicine (e.g., improved healthcare) to all(Bustamante et al., 2011).

In this work, our primary aim was to develop methods that improve the genomic analysis of diverse populations, specifically admixed populations. We presented methods that improve upon existing methods for calling IBD (PIGS) and using reference panels for summary statistics based approaches (Adapt-mix). Calling IBD has a multitude of uses including detecting cryptic relatedness, inferring demographic history, and association mapping. With PIGS we provided a computationally efficient way to leverage existing IBD calls; our method finds many more IBD segments than other approaches. The method contributes to facilitating future IBD based disease association studies. Furthermore, PIGS may provide new inroads into identifying small segments of IBD and therefore add to our understanding of human demographic history. With Adapt-mix we aimed at addressing the growing importance of summary statistics based methods for genomic studies of admixed populations and of populations without available reference panels. Due to computational and privacy issues, summary statistics will be used more widely in the future(Schork et al., 2013). Adapt-mix enables the accurate use of summary statistic based methods as genetic data from new populations are collected and high quality reference panels are not readily available. Adapt-mix may enable improved summary statistic based analyses of admixed populations (that have sub-optimal reference panels), and allow researchers to analyze populations that have not been previously studied in depth.

In addition to improving upon existing methodology, we developed two methods (AITL and 2PAC) that evaluate the extent to which genetic variation, environment, and their interactions, contribute to various traits. With AITL we leveraged the fact that environmental and medical covariates are associated with genetic ancestry to detect gene-environment and multiway gene-gene interactions. Interactions have been difficult to detect for logistical, statistical, biological, and computational reasons. However, they are believed to be responsible for a significant portion of missing heritability for several phenotypes. The relative contribution of additive and non-additive genetic effects to variability in molecular phenotypes and disease risk is an important area of investigation, and AITL provides a way to find such contributions(Powell et al., 2013). With 2PAC, we introduced a way to partition the relative contributions of genetic and environmental components to the correlation between ancestry and phenotype. Gaining insight into whether genetic component or environment drives differences in population phenotypic distributions has direct implications for precision medicine. Treatment of a disease should differ if the disease is caused primarily by genetic variation as opposed to an environmental factor and vice versa. We also proposed ways that 2PAC might be extended to identify regions of the genome that contribute substantially to the correlation between ancestry and phenotype. This extension could be used to find potential drug targets in the future.

It will be interesting to witness how genomics shapes the future landscape of healthcare and quality of life. More and more, there is a great push to "personalize" not just medicine but general lifestyle (i.e., exercise, beauty) based on an individual's genetic profile. The success of these efforts will hinge on a broad understanding of genetic variation and its impact on the full spectrum of human phenotypic variability.

References

1. Albrechtsen, A., Moltke, I., & Nielsen, R. (2010). Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, *186*(1), 295–308. http://doi.org/10.1534/genetics.110.113977
2. Albrechtsen, A., Sand Korneliussen, T., Moltke, I., van Overseem Hansen, T., Nielsen, F. C., & Nielsen, R. (2009). Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic Epidemiology*, *33*(3), 266–274. http://doi.org/10.1002/gepi.20378
3. Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. http://doi.org/10.1101/gr.094052.109
4. Anderson, L. N., Briollais, L., Atkinson, H. C., Marsh, J. A., Xu, J., Connor, K. L., et al. (2014). Investigation of Genetic Variants, Birthweight and Hypothalamic-Pituitary-Adrenal Axis Function Suggests a Genetic Variant in the SERPINA6 Gene Is Associated with Corticosteroid Binding Globulin in the Western Australia Pregnancy Cohort (Raine) Study. *PloS One*, *9*(4), e92957. http://doi.org/10.1371/journal.pone.0092957
5. Artigas, M. S., Loth, D. W., Wain, L. V., Gharib, S. A., Obeidat, M., Tang, W., et al. (2011). Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nature Genetics*, *43*(11), 1082–1090. http://doi.org/10.1038/ng.941
6. Aschard, H., Gusev, A., Brown, R., & Pasaniuc, B. (2015). Leveraging local ancestry to detect gene-gene interactions in genome-wide data. *BMC Genetics*, *16*(1), 124. http://doi.org/10.1186/s12863-015-0283-z
7. Aschard, H., Lutz, S., Maus, B., Duell, E. J., Fingerlin, T. E., Chatterjee, N., et al. (2012). Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. *Human Genetics*, *131*(10), 1591–1613. http://doi.org/10.1007/s00439-012-1192-0
8. Balding, D. J., & Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Human Identification: the Use of DNA Markers*.
9. Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., Rodríguez-Cintrón, W., Chapela, R., Ford, J. G., Avila, P. C., Rodriguez-Santana, J., Burchard, E. G., & Halperin, E. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, *28*(10), 1359–1367. http://doi.org/10.1093/bioinformatics/bts144
10. Borrell, L. N., Nguyen, E. A., Roth, L. A., Oh, S. S., Tcheurekdjian, H., Sen, S., et al. (2013). Childhood Obesity and Asthma Control in the GALA II and SAGE II Studies. *dx.doi.org* (Vol. 187, pp. 697–702). American Thoracic Society. http://doi.org/10.1164/rccm.201211-2116OC
11. Braak, ter, C. J. F., Boer, M. P., Totir, L. R., Winkler, C. R., Smith, O. S., & Bink, M. C. A. M. (2010). Identity-by-descent matrix decomposition using latent ancestral allele models. *Genetics*, *185*(3), 1045–1057. http://doi.org/10.1534/genetics.110.117390
12. Browning, B. L., & Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *American Journal of Human Genetics*, *88*(2), 173–182.

http://doi.org/10.1016/j.ajhg.2011.01.010

13. Browning, B. L., & Browning, S. R. (2013). Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data. *The American Journal of Human Genetics*, *93*(5), 840–851. http://doi.org/10.1016/j.ajhg.2013.09.014
14. Browning, B. L., & Browning, S. R. (2013). Improving the Accuracy and Efficiency of Identity by Descent Detection in Population Data. *Genetics*. http://doi.org/10.1534/genetics.113.150029
15. Browning, S. R., & Thompson, E. A. (2012). Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*, *190*(4), 1521–1531. http://doi.org/10.1534/genetics.111.136937
16. Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., et al. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans, *107*(2), 786–791. http://doi.org/10.1073/pnas.0909559107
17. Bulik-Sullivan, B., Loh, P. R., Finucane, H., Ripke, S., Yang, J., Psychiatric Genomics Consortium, S. W. G., et al. (2014). LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. *Nature Genetics, 47(3), 291-., http://doi.org/10.1038/ng.3211*
18. Burchard, E. G., Avila, P. C., Nazario, S., Casal, J., Torres, A., Rodríguez-Santana, J. R., et al. (2004). Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. *American Journal of Respiratory and Critical Care Medicine*, *169*(3), 386–392. http://doi.org/10.1164/rccm.200309-1293OC
19. Burchard, E. G., Ziv, E., Coyle, N., & Gomez, S. L. (2003). The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine, 348(12), 1170-5.*
20. Bustamante, C. D., La Vega, De, F. M., & Burchard, E. G. (2011). Genomics for the world. *Nature*, *475*(7355), 163–165. http://doi.org/10.1038/475163a
21. Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (2006). A Limited Memory Algorithm for Bound Constrained Optimization. *Dx.Doi.org*, *16*(5), 1190–1208. http://doi.org/10.1137/0916069
22. Carretero, O. A., & Oparil, S. (2000). Essential hypertension. Part I: definition and etiology. *Circulation*, *101*(3), 329–335. http://doi.org/10.1161/01.CIR.101.3.329
23. Cheng, C.-Y., Reich, D., Haiman, C. A., Tandon, A., Patterson, N., Elizabeth, S., et al. (2012). African Ancestry and Its Correlation to Type 2 Diabetes in African Americans: A Genetic Admixture Analysis in Three U.S. Population Cohorts. *PloS One*, *7*(3), e32840. http://doi.org/10.1371/journal.pone.0032840
24. Choudhry, S., Burchard, E. G., Borrell, L. N., Tang, H., Gomez, I., Naqvi, M., et al. (2006). Ancestry–Environment Interactions and Asthma Risk among Puerto Ricans. *American Journal of Respiratory and Critical Care Medicine*, *174*(10), 1088–1093. http://doi.org/10.1164/rccm.200605-596OC
25. Consortium, T. 1. G. P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. http://doi.org/10.1038/nature11632
26. Consortium, T. C. A. D. C. G. (2011). A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nature Genetics*, *43*(4), 339–344. http://doi.org/10.1038/ng.782
27. De, R., Hu, T., Moore, J. H., & Gilbert-Diamond, D. (2015). Characterizing gene-gene interactions in a statistical epistasis network of twelve candidate genes for obesity.

BioData Mining, *8*(1), 1–16. http://doi.org/10.1186/s13040-015-0077-x

28. Delude, C. M. (2015). Deep phenotyping: The details of disease. *Nature*, *527*(7576), S14–S15. http://doi.org/10.1038/527S14a
29. Devlin, B., & Roeder, K. (2004). Genomic Control for Association Studies. *Biometrics*, *55*(4), 997–1004. http://doi.org/10.1111/j.0006-341X.1999.00997.x
30. Drake, K. A., Torgerson, D. G., Gignoux, C. R., Galanter, J. M., Roth, L. A., Huntsman, S., et al. (2014). A genome-wide association study of bronchodilator response in Latinos implicates rare variants. *Journal of Allergy and Clinical Immunology*, *133*(2), 370–378.e15. http://doi.org/10.1016/j.jaci.2013.06.043
31. Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, *11*(1), 587. http://doi.org/10.1186/1471-2105-11-587
32. Edge, M. D., & Rosenberg, N. A. (2015). Implications of the apportionment of human genetic diversity for the apportionment of human phenotypic diversity. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *52*, 32–45. http://doi.org/10.1016/j.shpsc.2014.12.005
33. Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, *11*(6), 446–450. http://doi.org/10.1038/nrg2809
34. Excoffier, L., & Foll, M. (2011). fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, *27*(9), 1332–1334. http://doi.org/10.1093/bioinformatics/btr124
35. Fejerman, L., John, E. M., Huntsman, S., Beckman, K., Choudhry, S., Perez-Stable, E., et al. (2008). Genetic ancestry and risk of breast cancer among U.S. Latinas. *Cancer Research*, *68*(23), 9723–9728. http://doi.org/10.1158/0008-5472.CAN-08-2039
36. Florez, J. C., Price, A. L., Campbell, D., Riba, L., Parra, M. V., Yu, F., et al. (2011). Strong Association of Socioeconomic Status and Genetic Ancestry in Latinos: Implications for Admixture Studies of Type 2 Diabetes. In *Racial Identities, Genetic Ancestry, and Health in South America* (pp. 137–153). Palgrave Macmillan US. http://doi.org/10.1007/978-1-137-00170-2_7
37. Fraser, H. B., Lam, L. L., Neumann, S. M., & Kobor, M. S. (2012). Population-specificity of human DNA methylation, *13*(2), R8. http://doi.org/10.1186/gb-2012-13-2-r8
38. Galanter, J. M., Gignoux, C. R., Oh, S. S., Torgerson, D., Pino-Yanes, M., Thakur, N., et al. (2016). Methylation Analysis Reveals Fundamental Differences Between Ethnicity and Genetic Ancestry. *bioRxiv*, 036822. http://doi.org/10.1101/036822
39. Galarneau, G., Palmer, C. D., Sankaran, V. G., Orkin, S. H., Hirschhorn, J. N., & Lettre, G. (2010). Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nature Genetics*, *42*(12), 1049–1051. http://doi.org/10.1038/ng.707
40. Gamazon, E. R., & Park, D. S. (2016). SNP-based heritability estimation: measurement noise, population stratification and stability. *bioRxiv*, 040055. http://doi.org/10.1101/040055
41. Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., et al. (2003). The International HapMap Project. *Nature*, *426*(6968), 789–796.

http://doi.org/doi:10.1038/nature02168

42. Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, *44*(10), 1084–1089. http://doi.org/10.1038/ng.2394
43. Gusev, A., Kenny, E. E., Lowe, J. K., Salit, J., Saxena, R., Kathiresan, S., et al. (2011). DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. *American Journal of Human Genetics*, *88*(6), 706–717. http://doi.org/10.1016/j.ajhg.2011.04.023
44. Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., et al. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, *19*(2), 318–326. http://doi.org/10.1101/gr.081398.108
45. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, *339*(6117), 321–324. http://doi.org/10.1126/science.1229566
46. Han, B., Hackel, B. M., & Eskin, E. (2011). Postassociation cleaning using linkage disequilibrium information. *Genetic Epidemiology*, *35*(1), 1–10. http://doi.org/10.1002/gepi.20544
47. Han, B., Kang, E. Y., Raychaudhuri, S., & de Bakker, P. (2013). Fast pairwise IBD association testing in genome-wide association studies. *Bioinformatics*, *30*(2), 206-13. http://doi.org/10.1093/bioinformatics/btt609
48. He, D. (2013). IBD-Groupon: an efficient method for detecting group-wise identity-by-descent regions simultaneously in multiple individuals based on pairwise IBD relationships. *Bioinformatics*, *29*(13), i162–70. http://doi.org/10.1093/bioinformatics/btt237
49. Hemani, G., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A. K., Mcrae, A. F., et al. (2014). Detection and replication of epistasis influencing transcription in humans. *Nature*, *508*(7495), 249–253. http://doi.org/10.1038/nature13005
50. Hochreiter, S. (2013). HapFABIA: identification of very short segments of identity by descent characterized by rare variants in large sequencing data. *Nucleic Acids Research*, *41*(22), e202. http://doi.org/10.1093/nar/gkt1013
51. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., & Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics*, *198*(2), 497–508. http://doi.org/10.1534/genetics.114.167908
52. Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, *5*(6), e1000529. http://doi.org/10.1371/journal.pgen.1000529
53. Hsu, L., Jiao, S., Dai, J. Y., Hutter, C., Peters, U., & Kooperberg, C. (2012). Powerful Cocktail Methods for Detecting Genome-Wide Gene-Environment Interaction. *Genetic Epidemiology*, *36*(3), 183–194. http://doi.org/10.1002/gepi.21610
54. Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: a Cancer Journal for Clinicians*, *61*(2), 69–90. http://doi.org/10.3322/caac.20107
55. Kang, E. Y., Han, B., Furlotte, N., Joo, J. W. J., Shih, D., Davis, R. C., et al. (2014). Meta-Analysis Identifies Gene-by-Environment Interactions as Demonstrated in a Study of 4,965 Mice. *PLoS Genetics*, *10*(1), e1004022. http://doi.org/10.1371/journal.pgen.1004022

56. Kearney, P. M., Whelton, M., Reynolds, K., Muntner, P., Whelton, P. K., & He, J. (2005). Global burden of hypertension: analysis of worldwide data. *The Lancet*, *365*(9455), 217–223. http://doi.org/10.1016/S0140-6736(05)17741-1

57. Keller, M. C. (2014). Gene × Environment Interaction Studies Have Not Properly Controlled for Potential Confounders: The Problem and the (Simple) Solution. *Biological Psychiatry*, *75*(1), 18–24.

58. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., et al. (2014). Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genetics*, *10*(10), e1004722. http://doi.org/10.1371/journal.pgen.1004722

59. Kooperberg, C., & LeBlanc, M. (2008). Increasing the power of identifying gene × gene interactions in genome-wide association studies. *Genetic Epidemiology*, *32*(3), 255–263. http://doi.org/10.1002/gepi.20300

60. Kumar, R., Nguyen, E. A., Roth, L. A., Oh, S. S., Gignoux, C. R., Huntsman, S., et al. (2013). Factors associated with degree of atopy in Latino children in a nationwide pediatric sample: The Genes-environments and Admixture in Latino Asthmatics (GALA II) study. *The Journal of Allergy and Clinical Immunology*, *132*(4), 896–905.e1. http://doi.org/10.1016/j.jaci.2013.02.046

61. Kumar, R., Seibold, M. A., Aldrich, M. C., Williams, L. K., Reiner, A. P., Colangelo, L., et al. (2010). Genetic Ancestry in Lung-Function Predictions. *New England Journal of Medicine*, *363*(4), 321–330. http://doi.org/10.1056/NEJMoa0907897

62. Lee, M., Raj, T., & Castillo, I. W. (2012). ImmVar Project: Genetic architecture of leukocyte gene expression in healthy humans. JOURNAL OF ….

63. Letouzé, E., Sow, A., Petel, F., Rosati, R., Figueiredo, B. C., Burnichon, N., et al. (2012). Identity by descent mapping of founder mutations in cancer using high-resolution tumor SNP data. *PloS One*, *7*(5), e35897. http://doi.org/10.1371/journal.pone.0035897

64. Liao, S. Y., Lin, X., & Christiani, D. C. (2013). Gene-environment interaction effects on lung function-a genome-wide association study within the Framingham heart study. *Environ Health*.

65. Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., et al. (2000). Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal of Medicine*, *343*(2), 78–85. http://doi.org/10.1056/NEJM200007133430201

66. Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., et al. (2010). A Versatile Gene-Based Test for Genome-wide Association Studies. *The American Journal of Human Genetics*, *87*(1), 139–145. http://doi.org/10.1016/j.ajhg.2010.06.009

67. Manolio, T. A., Manolio, T. A., Collins, F. S., Collins, F. S., Cox, N. J., Cox, N. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. http://doi.org/10.1038/nature08494

68. Maples, B. K., Gravel, S., & Kenny, E. E. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics, 93(2), 278-88. http://doi.org/10.1016/j.ajhg.2013.06.020*

69. Martin, A. R., Costa, H. A., Lappalainen, T., Henn, B. M., Kidd, J. M., Yee, M.-C., et al. (2014). Transcriptome Sequencing from Diverse Human Populations Reveals Differentiated Regulatory Architecture. *PLoS Genetics*, *10*(8), e1004549.

http://doi.org/10.1371/journal.pgen.1004549

70. Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., et al. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *American Journal of Human Genetics*, *100*(4), 635–649. http://doi.org/10.1016/j.ajhg.2017.03.004

71. Moltke, I., Albrechtsen, A., Hansen, T. V. O., Nielsen, F. C., & Nielsen, R. (2011). A method for detecting IBD regions simultaneously in multiple individuals--with applications to disease genetics. *Genome Research*, *21*(7), 1168–1180. http://doi.org/10.1101/gr.115360.110

72. Moorjani, P., Thangaraj, K., Patterson, N., Lipson, M., Loh, P.-R., Govindaraj, P., et al. (2013). Genetic Evidence for Recent Population Mixture in India. *The American Journal of Human Genetics*, *93*(3), 422–438. http://doi.org/10.1016/j.ajhg.2013.07.006

73. Murcray, C. E., Lewinger, J. P., & Gauderman, W. J. (2009). Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology*, *169*(2), 219–226. http://doi.org/10.1093/aje/kwn353

74. Palamara, P. F., Lencz, T., Darvasi, A., & Pe'er, I. (2012). Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. *The American Journal of Human Genetics*, *91*(5), 809–822. http://doi.org/10.1016/j.ajhg.2012.08.030

75. Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Zaitlen, N., Eng, C., et al. (2013). Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics*, *29*(11), btt166–1415. http://doi.org/10.1093/bioinformatics/btt166

76. Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D. P., Patterson, N., & Price, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, *30*(20), btu416–2914. http://doi.org/10.1093/bioinformatics/btu416

77. Peralta, C. A., Risch, N., Lin, F., Shlipak, M. G., Reiner, A., Ziv, E., et al. (2009). The Association of African Ancestry and Elevated Creatinine in the Coronary Artery Risk Development in Young Adults (CARDIA) Study. *American Journal of Nephrology*, *31*(3), 202–208. http://doi.org/10.1159/000268955

78. Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164. http://doi.org/10.1038/538161a

79. Powell, J. E., Henders, A. K., Mcrae, A. F., Kim, J., Hemani, G., Martin, N. G., et al. (2013). Congruence of Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and SNP Data. *PLoS Genetics*, *9*(5), e1003502. http://doi.org/10.1371/journal.pgen.1003502

80. Price, A. L., Patterson, N., Hancks, D. C., Myers, S., Reich, D., Cheung, V. G., & Spielman, R. S. (2008). Effects of cis and trans Genetic Ancestry on Gene Expression in African Americans. *PLoS Genetics*, *4*(12), e1000294. http://doi.org/10.1371/journal.pgen.1000294

81. Price, A. L., Patterson, N., Yu, F., Cox, D. R., Waliszewska, A., McDonald, G. J., et al. (2007). A Genomewide Admixture Map for Latino Populations. *The American Journal of Human Genetics*, *80*(6), 1024–1036. http://doi.org/10.1086/518313

82. Price, A. L., Patterson, N. J., Patterson, N. J., Plenge, R. M., Plenge, R. M., et al. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. http://doi.org/10.1038/ng1847

83. Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(2), 945–959.

84. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. http://doi.org/10.1086/519795

85. Qian, Y., Browning, B. L., & Browning, S. R. (2013). Efficient clustering of identity-by-descent between multiple individuals. *Bioinformatics*, *30*(7), btt734–922. http://doi.org/10.1093/bioinformatics/btt734

86. Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, *461*(7263), 489–494. http://doi.org/10.1038/nature08365

87. Reiner, A. P., Carlson, C. S., Ziv, E., Iribarren, C., Jaquish, C. E., & Nickerson, D. A. (2007). Genetic ancestry, population sub-structure, and cardiovascular disease-related traits among African-American participants in the CARDIA Study. *Human Genetics*, *121*(5), 565–575. http://doi.org/10.1007/s00439-007-0350-2

88. Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., & Moore, J. H. (2001). Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *The American Journal of Human Genetics*, *69*(1), 138–147. http://doi.org/10.1086/321276

89. Robinson, P. N. (2012). Deep phenotyping for precision medicine. *Human Mutation*, *33*(5), 777–780. http://doi.org/10.1002/humu.22080

90. Rouhani, F., Kumasaka, N., de Brito, M. C., Bradley, A., Vallier, L., & Gaffney, D. (2014). Genetic Background Drives Transcriptional Variation in Human Induced Pluripotent Stem Cells. *PLoS Genetics*, *10*(6), e1004432. http://doi.org/10.1371/journal.pgen.1004432

91. Sanchez, E., Webb, R. D., Rasmussen, A., Kelly, J. A., Riba, L., Kaufman, K. M., et al. (2010). Genetically determined Amerindian ancestry correlates with increased frequency of risk alleles for systemic lupus erythematosus. *Arthritis & Rheumatism*, *62*(12), 3722–3729. http://doi.org/10.1002/art.27753

92. Sankararaman, S., Sridhar, S., & Kimmel, G. (2008). Estimating local ancestry in admixed populations. *The American Journal of Human Genetics. 82(2), 290-303. http://doi.org/10.1016/j.ajhg.2007.09.022*

93. Sanna, S., Li, B., Mulas, A., Sidore, C., Kang, H. M., Jackson, A. U., et al. (2011). Fine Mapping of Five Loci Associated with Low-Density Lipoprotein Cholesterol Detects Variants That Double the Explained Heritability. *PLoS Genetics*, *7*(7), e1002198. http://doi.org/10.1371/journal.pgen.1002198

94. Schork, A. J., Thompson, W. K., Pham, P., Torkamani, A., Roddey, J. C., Sullivan, P. F., et al. (2013). All SNPs Are Not Created Equal: Genome-Wide Association Studies Reveal a Consistent Pattern of Enrichment among Functionally Annotated SNPs. *PLoS Genetics*, *9*(4), e1003449. http://doi.org/10.1371/journal.pgen.1003449

95. Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, *43*(4), 333–338. http://doi.org/10.1038/ng.784

96. Seldin, M. F., Pasaniuc, B., & Price, A. L. (2011). New approaches to disease mapping in

admixed populations. *Nature Reviews Genetics*, *12*(8), 523–528. http://doi.org/10.1038/nrg3002

97. Shaffer, J. R., Kammerer, C. M., Reich, D., McDonald, G., Patterson, N., Goodpaster, B., et al. (2007). Genetic markers for ancestry are correlated with body composition traits in older African Americans. *Osteoporosis International*, *18*(6), 733–741. http://doi.org/10.1007/s00198-006-0316-6
98. Silva-Zolezzi, I., Hidalgo-Miranda, A., Estrada-Gil, J., Fernandez-Lopez, J. C., Uribe-Figueroa, L., Contreras, A., et al. (2009). Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(21), 8611–8616. http://doi.org/10.1073/pnas.0903045106
99. Simon-Sanchez, J., Scholz, S., Fung, H.-C., Matarin, M., Hernandez, D., Gibbs, J. R., et al. (2007). Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. *Human Molecular Genetics*, *16*(1), 1–14. http://doi.org/10.1093/hmg/ddl436
100. Smith, M. W., Patterson, N., Lautenberger, J. A., Truelove, A. L., McDonald, G. J., Waliszewska, A., et al. (2004). A High-Density Admixture Map for Disease Gene Discovery in African Americans. *The American Journal of Human Genetics*, *74*(5), 1001–1013. http://doi.org/10.1086/420856
101. Smith, Z. D., & Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, *14*(3), 204–220. http://doi.org/10.1038/nrg3354
102. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S., & Balding, D. J. (2017). Reevaluation of SNP heritability in complex human traits. *Nature Genetics*, *35*, 212. http://doi.org/10.1038/ng.3865
103. Speed, D., Hemani, G., Johnson, M. R., & Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics*, *91*(6), 1011–1021. http://doi.org/10.1016/j.ajhg.2012.10.010
104. Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, *42*(11), 937–948. http://doi.org/10.1038/ng.686
105. Spielman, R. S., Bastone, L. A., Burdick, J. T., Morley, M., Ewens, W. J., & Cheung, V. G. (2007). Common genetic variants account for differences in gene expression among ethnic groups. *Nature Genetics*, *39*(2), 226–231. http://doi.org/10.1038/ng1955
106. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*, *12*(3), e1001779. http://doi.org/10.1371/journal.pmed.1001779
107. Sul, J. H., Bilow, M., Yang, W.-Y., Kostem, E., Furlotte, N., He, D., & Eskin, E. (2016). Accounting for Population Structure in Gene-by-Environment Interactions in Genome-Wide Association Studies Using Mixed Models. *PLoS Genetics*, *12*(3), e1005849. http://doi.org/10.1371/journal.pgen.1005849
108. Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., & Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, *17*(4), 520–526. http://doi.org/10.1101/gr.6023607

109. The ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, *306*(5696), 636–640. http://doi.org/10.1126/science.1105136

110. Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., & Risch, N. (2012). Estimating kinship in admixed populations. *American Journal of Human Genetics*, *91*(1), 122–138. http://doi.org/10.1016/j.ajhg.2012.05.024

111. Torgerson, D. G., Gignoux, C. R., Galanter, J. M., Drake, K. A., Roth, L. A., Eng, C., et al. (2012). Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. *The Journal of Allergy and Clinical Immunology*, *130*(1), 76–82.e12. http://doi.org/10.1016/j.jaci.2012.02.040

112. Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindstrom, S., Genovese, G., et al. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, *97*(4), 576–592. http://doi.org/10.1016/j.ajhg.2015.09.001

113. Vimaleswaran, K. S., Berry, D. J., Lu, C., Tikkanen, E., Pilz, S., Hiraki, L. T., et al. (2013). Causal Relationship between Obesity and Vitamin D Status: Bi-Directional Mendelian Randomization Analysis of Multiple Cohorts. *PLoS Med*, *10*(2), e1001383. http://doi.org/10.1371/journal.pmed.1001383

114. Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24. http://doi.org/10.1016/j.ajhg.2011.11.029

115. Visscher, P. M., Hill, W. G., & Wray, N. R. (2008). Heritability in the genomics era--concepts and misconceptions. *Nature Reviews Genetics*, *9*(4), 255–266. http://doi.org/10.1038/nrg2322

116. Wang, S., Ray, N., Rojas, W., Parra, M. V., Bedoya, G., Gallo, C., et al. (2008). Geographic Patterns of Genome Admixture in Latin American Mestizos. *PLoS Genetics*, *4*(3), e1000037. http://doi.org/10.1371/journal.pgen.1000037

117. Weir, B. S., Anderson, A. D., & Hepler, A. B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, *7*(10), 771–780. http://doi.org/10.1038/nrg1960

118. Welter, D., Macarthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(D1), D1001–D1006. http://doi.org/10.1093/nar/gkt1229

119. Whittemore, A. S., & Halpern, J. (1994). Probability of gene identity by descent: computation and applications. *Biometrics*, *50*(1), 109–117.

120. Willett, W. C. (2002). Balancing life-style and genomics research for disease prevention. *Science*, *296*(5568), 695–698. http://doi.org/10.1126/science.1071055

121. Wong, M. Y., Day, N. E., Luan, J. A., Chan, K. P., & Wareham, N. J. (2003). The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *International Journal of Epidemiology*, *32*(1), 51–57.

122. Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. http://doi.org/10.1038/ng.608

123. Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Consortium, G. I. O. A. T. G., Replication, D. G., et al. (2012). Conditional and joint multiple-SNP analysis of

GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, *44*(4), 369–375. http://doi.org/10.1038/ng.2213

124. Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2013). Genome-Wide Complex Trait Analysis (GCTA): Methods, Data Analyses, and Interpretations. In *Genome-Wide Association Studies and Genomic Prediction* (Vol. 1019, pp. 215–236). Totowa, NJ: Humana Press. http://doi.org/10.1007/978-1-62703-447-0_9
125. Yang, J., Yang, J., Benyamin, B., Benyamin, B., McEvoy, B. P., McEvoy, B. P., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. http://doi.org/10.1038/ng.608
126. Zaitlen, N., Kang, H. M., & Eskin, E. (2009). Linkage effects and analysis of finite sample errors in the HapMap. Human heredity. http://doi.org/10.1111/j.1469-1809.2006.00383.x/full
127. Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., & Price, A. L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genetics*, *9*(5), e1003520. http://doi.org/10.1371/journal.pgen.1003520
128. Zhu, X., Tang, H., & Risch, N. (2008). Admixture Mapping and the Role of Population Structure for Localizing Disease Genes. In *Genetic Dissection of Complex Traits* (Vol. 60, pp. 547–569). Elsevier. http://doi.org/10.1016/S0065-2660(07)00419-1
129. Ziv, E., John, E. M., Choudhry, S., Kho, J., Lorizio, W., Pérez-Stable, E. J., & Burchard, E. G. (2006). Genetic Ancestry and Risk Factors for Breast Cancer among Latinas in the San Francisco Bay Area. *Cancer Epidemiology Biomarkers & Prevention*, *15*(10), 1878–1885. http://doi.org/10.1158/1055-9965.EPI-06-0092

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

______________________________ 5/1/17

Author Signature Date