Biological Insights From Population Differentiation

Biological Insights from Population Differentiation

a dissertation presented
by
Kevin Joseph Galinsky
to
The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts
January 2017

Thesis advisor: Professor Alkes L. Price Kevin Joseph Galinsky

Biological Insights from Population Differentiation

Abstract

Population genetics studies the genetic variation within and between populations to gain understanding of human history and insight into underlying biological processes. My dissertation introduces three distinct methods: a linear-time principal components analysis (PCA) algorithm, a scan for natural selection along continuous principal components (PCs), and a relationship between the cross-population correlation of genetic effects at all single nucleotide polymorphisms (SNPs) and the correlation of genetic effects at typed SNPs. These methods are all related to the statistical concept of the correlation matrix. The first two build off the genetic correlation matrix across individuals (also known as a genetic relationship matrix, or GRM), and the last on the correlations between SNPs in a population (also known as the linkage disequilibrium matrix, or LD matrix).

With the PCA algorithm, we are now able to study the population structure of European American and British populations with finer resolution using very large population samples (55k and 113k samples, respectively). These PCs were fed into the scan for natural selection that detected signals of selection at known loci as well as several novel loci, including a gene protective against alcoholism in Europeans and several genes associated with blood pressure in the British. Lastly, using the SNP LD patterns in several populations we computed a factor that can be used in cross-population heritability scans to correct for the differential tagging efficiency within those populations.

Contents

Listing of figures

Listing of tables

Acknowledgments

I WOULD LIKE TO THANK my advisor, Alkes Price. Without his guidance there is a good chance that this doctorate would not have been completed. I thank him for keeping me on track and making sure that I published in a timely manner.

I WOULD ALSO LIKE TO THANK my wife, Nicole. She has been my rock during the majority of my studies, and without her there is a good chance that this doctorate would not have been completed. I thank her for encouraging me during the many times I have wanted to quit.

LASTLY, I WOULD LIKE TO THANK everyone associated with the Price lab and the HSPH Biostatistics department. I'm a firm believer in peer effects when it comes to education. The lab is full of rock stars and I know they pushed me to do better.

0

Introduction

The comparative study of human populations has led to numerous advances in the field of genetics. Aside from allowing us to study the history of human expansions and migrations[24,118,90,97,4,62], population genetics has allowed us to correct for confounding due to population stratification in genome-wide association (GWA) studies[112,113], scan for signals of natural selection which may be associated with disease[78,124], and to better understand the genetic architecture of complex traits[31,18]. Here we present three methods that advance the field of population genetics in these respective areas.

Population stratification induces confounding in genetic association studies when an-

cestry is associated with both the phenotype of interest and the genotypes at a given locus[113]. In doing so, ancestry induces a spurious correlation between genotype and phenotype which must be corrected. One way of doing this is using principal components analysis (PCA)[112]. PCA is a data transformation technique which produces an orthogonal set of principal components (PCs) where each PC contains the maximum amount of variation unaccounted for by previous principal components. When applied to genomic data, the top PCs represent population structure in the sample[97,112,103]. The drawback to PCA is that its computational efficiency isn't linear[121,55]. Its resource requirements are between quadratic and cubic with the number of samples. An implementation of PCA which may run in seconds on only a few hundred samples can be computationally intractable with modern datasets containing over a hundred thousand samples[131]. The first two chapters of this work will describe our implementation of a linear-time FastPCA algorithm[43,44], which computes a highly accurate approximation of the first few PCs, and its application to a European-American dataset[6] and a British dataset[131].

Our second method extends earlier population differentiation approaches[78,12] to detect positive natural selection[124]. Given a pair of populations differentiated by some F_{ST}[148,13], one can detect allele frequencies that are more differentiated than what one would expect if only genetic drift were occurring. SNPs that exhibit unusual levels of population differentiation are assumed to be under selective pressure in one population of the pair. Examples of such SNPs that have been subsequently linked to phenotypes include lactase persistence[11], hypoxia response[15], and malaria resistence[5]. Previous work to detect natural selection from population differentiation required either discrete subpopulations or were unable to produce a p-value[157]. We have developed a method that converts the SNP load-

ings that can be computed as part of PCA to chi-square (1 d.o.f.) statistics. Combined with the FastPCA algorithm, this approach allows us to rapidly detect natural selection in large sample cohorts. Additionally, population differentiation approaches benefit from large sample sizes and closely-related populations. By examining a large cohort composed of closely-related subpopulations, we have the power to detect many new signals of selection.

Our last project is the study of cross-population heritability. Mixed model methods[153] have been used to examine the genetic architecture of complex traits, and bivariate methods[31,154,74,76,85] have been able to detect the correlation of joint-fit SNP effects between either two complex traits or the same effect in two populations. Additionally, summary-statistics based methods have been developed to do the same thing[18,21,20]. The cross-population correlation of joint-fit SNP effects is affected by differential linkage disequilibrium (LD) in the two populations. We have developed a theoretical correction factor which computes the ratio of the cross-population correlation of joint-fit SNP effects and all causal SNP effects. By applying this correction factor to the estimated cross-population correlation of joint-fit SNP effects, we can estimate the cross-population correlation of causal SNP effects.

1

Population structure and natural selection in European Americans

1.1 Preface

The following work was published in the March 2016 issue of *The American Journal of Human Genetics*[43], titled *Fast principal components analysis reveals convergent evolution of ADH1B in Europe and East Asia* with co-authors Gaurav Bhatia, Po-Ru

Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J. Patterson and Alkes L. Price. It introduces our implementation of a linear-time PCA algorithm[121,55] and develops a PC-based natural selection statistic based upon population differentiation[78,12]. We applied these two methods to a European-American dataset[6] containing 54k samples and detected a novel signal of selection in Europeans in the alcohol dehydrogenase gene.

1.2 Abstract

Searching for genetic variants with unusual differentiation between subpopulations is an established approach for identifying signals of natural selection. However, existing methods generally require discrete subpopulations. We introduce a method that infers selection using principal components (PCs) by identifying variants whose differentiation along top PCs is significantly greater than the null distribution of genetic drift. To enable the application of this method to large data sets, we developed the FastPCA software, which employs recent advances in random matrix theory to accurately approximate top PCs while reducing time and memory cost from quadratic to linear in the number of individuals, a computational improvement of many orders of magnitude. We apply FastPCA to a cohort of 54,734 European Americans, identifying 5 distinct subpopulations spanning the top 4 PCs. Using the PC-based test for natural selection, we replicate previously known selected loci and identify three new genome-wide significant signals of selection, including selection in Europeans at *ADH1B*. The coding variant rs1229984*T has previously been associated to a decreased risk of alcoholism and shown to be under selection in East Asians; we show that it is a rare example of independent evolution on two continents. We also detect selection signals at *IGFBP3* and *IGH*, which have also previously been associ-

ated to human disease.

1.3 Introduction

Searching for genetic variants with unusual differentiation between populations is an established approach for identifying signals of natural selection[124,95,96,126,66,130]. We and others have employed this approach to identify signals of selection in a wide range of settings, informing our understanding of genes under evolutionary adaptation. Examples includes genes linked to lactase persistence[11,143], starch hydrolysis[104], fatty acid decomposition[41], red blood cell abundance[158], hypoxia response[15], alcoholism[57], kidney disease[69], malaria[56,5,12,51], HIV/AIDS[152], autoimmune disease[58], cancer[12] [OMIM 602470], cystic fibrosis[2] and hypertension[51] . However, the signals of selection identified thus far may represent "only the tip of the iceberg[70]", implying that further research on selection will provide additional insights about human disease. Unlike extended haplotype homozygosity (EHH) or allele frequency spectrum based tests for selection, the population differentiation approach is able to detect older selection events and selection on standing variation[124,96]. In addition, signals of selection detected using population differentiation can flag stratified genetic variants that are susceptible to false-positive associations in genome-wide association studies[111].

Recent work on detecting selection using population differentiation has focused on methods that evaluate deviations from genome-wide patterns of genetic drift between discrete populations, such as Locus-Specific Branch Length (LSBL)[130], Population Branch Statistic (PBS)[158] and TreeSelect[12]. These ideas are derived from the Lewontin and Krakauer test[78] and its extensions to the multinomial-Dirichlet model (F-model)[9] (later

incorporating a Bayesian framework[38], hierarchical population structure[35] and complex demography[39]) and to population trees[17] (see also Nicholson *et al.*[94] for a similar method that uses population trees and Günther and Coop[47] which uses population kinships). The population differentiation approach has greatest power when comparing very closely related populations with very large sample size[12]. The increasing availability of very large population cohorts for genetic analysis provides strong prospects for analyzing subtle differences in ancestry in large sample sizes, but raises the challenge of how to select subpopulations to compare; a population cohort with a single continental ancestry may be better represented by continuous clines rather than discrete clusters[112,110,98], and/or may contain a large number of discrete subpopulations corresponding to a large number of possible population comparisons[142,77]. Principal components analysis (PCA)[112,103] offers an appealing alternative to model-based clustering methods[116,3] for modeling human genetic diversity, and has been applied to infer population structure in many settings[110,98,103,141,129,81,65,128,73,91]. One advantage of PCA is that results for top PCs are not sensitive to the number of PCs analyzed, whereas results of model-based clustering methods often vary with the number of clusters. Another advantage of PCA is its low computational cost, as top PCs can be inferred in time only linear in the number of samples by drawing upon recent advances in random matrix theory[121,54,55], implemented in the FastPCA software that we introduce here. We thus developed a test for selection that uses the SNP weights from PCA to calculate the differentiation of each locus along top PCs; our approach is similar in spirit to a recently proposed test for selection based on Bayesian factor analysis[33] but has much lower computational cost.

Specifically, the squared correlation of each SNP to a PC, rescaled to account for ge-

netic drift, follows a chi-square (1 d.o.f.) distribution under the null hypothesis of no selection. Our PC-based test produces a p-value at each locus and is able to detect signals at genome-wide significance, a key consideration in genome scans for selection[12].

We ran FastPCA on 54,734 individuals of European descent from the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort; FastPCA required only 57 minutes of compute time and 2.6GB of RAM for this analysis, orders of magnitude better than any other publicly available software. We detected evidence of population structure along the top 4 PCs, which separated samples into several subpopulations. Using our PC-based test for selection, we replicate previously known selected loci *LCT*, *HLA*, *OCA2* and*IRF4* and identify three additional signals of selection at *IGH*, *IGFBP3* and *ADH1B* . The signal in *ADH1B* at coding variant rs1229984 has previously been associated to alcoholism[34,149,79,46] and shown to be under selection in East Asians[57,79,100,105]; we show that it is a rare example of independent evolution on two continents[143,104].

1.4 Methods

1.4.1 Overview of methods

We first describe the FastPCA algorithm, which is an implementation of the *blanczos* method from Rokhlin *et al.*[121,54,55]. As with our previous work on PCA[112,103], FastPCA makes use of existing computational literature and does not contain any new computational ideas; nonetheless, we anticipate that the software will be widely used, since to our knowledge it is the only publicly available software for computing top PCs on genetic data in linear time. The algorithm generalizes the method of power iteration[48], a tech-

nique to estimate the largest eigenvalue and corresponding eigenvector of a matrix. Multiplying a random vector by a square matrix projects that vector onto the eigenvectors of that matrix and then scales it according the respective eigenvalues of that matrix. After repeating, the projection along the eigenvector with the largest eigenvalue grows fasters than the rest and the repeated matrix by vector product converges to this eigenvector. Additional eigenvectors can be found by repeating this process and orthogonalizing to previously-found PCs. The *blanczos* method improves on this method by initially estimating more PCs than ultimately desired. The original estimates are perturbed from the true PCs, but this missing variation is captured by estimating the extra PCs. The genotype matrix is then projected onto this set of eigenvectors, reducing its dimension while preserving the variation along the top PCs. Traditional PCA methods are applied to this reduced matrix to find accurate estimates of the top PCs of the original matrix.

We next describe our PC-based selection statistic, which generalizes a previous selection statistic developed for discrete populations[12]. We detect unusual allele frequency differences along inferred PCs by making use of the fact that the squared correlation of each SNP to a PC, rescaled to account for genetic drift, follows a chi-square (1 d.o.f.) distribution under the null hypothesis of no selection. We have released open-source software implementing the FastPCA algorithm and PC-based selection statistic.

1.4.2 FastPCA algorithm

We are given an input $M \times N$ genotype matrix $\boldsymbol{X}$, where M is the number of SNPs and N is the number of individuals (e.g. each row is a SNP, each column is a sample). Each entry in this matrix takes its values from $\{0, 1, 2\}$ indicating the count of variant alleles

for a sample at a SNP. From this matrix we can generate the normalized genomic matrix $\boldsymbol{Y}_{M\times N} = \left(\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, \ldots, \boldsymbol{y}_M^T\right)^T$ where each row $\boldsymbol{y}_i$ has approximately mean 0 and variance 1 for SNPs in Hardy-Weinberg equilibrium.

$$\begin{aligned} \hat{p}_i &= \frac{\sum_{j=1}^{N} x_{ij}}{2N_i} = \frac{\boldsymbol{x}_i \mathbf{1}}{2\mathbf{1}^T\mathbf{1}} \\ y_{ij} &= \frac{x_{ij} - 2\hat{p}_i}{\sqrt{2\hat{p}_i\left(1-\hat{p}_i\right)}} \\ \boldsymbol{y}_i &= (y_{i1}, y_{i2}, \ldots, y_{iN}) = \frac{x_{ij} - 2\hat{p}_i}{\sqrt{2\hat{p}_i\left(1-\hat{p}_i\right)}} \end{aligned} \tag{1.1}$$

Here, $\boldsymbol{x}_i$ is the row vector of genotypes for SNP i and $\boldsymbol{y}_i$ is the normalized row vector. x_{ij} and y_{ij} are the genotype/normalized genotype at SNP i for sample j. N_i is the number of valid genotypes at SNP i. All this is used to calculate $\hat{p}_i$, the sample allele frequency for SNP i, which is used to normalize the genotypes. In practice, the genotype matrix is normalized through the use of a lookup table mapping from genotypes (stored as 0, 1 or 2 copies of the alternate allele, or missing data) to normalized genotypes (using the above formula, with missing data having a normalized value of 0).

We are seeking the top K PCs for the normalized genomic matrix $\boldsymbol{Y}$. Traditional PCA algorithms compute the PCs by performing the eigendecomposition of the genetic relationship matrix ($GRM = \boldsymbol{Y}^T\boldsymbol{Y}/M$), a costly procedure which returns all the principal components. FastPCA, which makes use of recent advances in random matrix theory[121,54,55], speeds this process up by only approximating the top K PCs.

FastPCA is seeded with a random $N \times L$ matrix $\boldsymbol{G}_0$ composed of values drawn from a standard Gaussian distribution. L affects the accuracy of the result and L should be greater than K. For $K = 10$, $L = 20$ is a good choice. Then, for I iterations, we calculate

$\boldsymbol{H}_i = \boldsymbol{Y} \times \boldsymbol{G}_i$ and $\boldsymbol{G}_{i+1} = \boldsymbol{Y}^T \times \boldsymbol{H}_i / M$, where the $\boldsymbol{H}_i$s are $M \times L$ matrices and $\boldsymbol{G}_i$s are $N \times L$ matrices like $\boldsymbol{G}_0$. In simulated samples with discrete populations, $I = 3$ was sufficient, but in real datasets, $I = 10$ was found to provide accurate results.

After the iterative step completes, we stack the $\boldsymbol{H}_i$ matrices to produce the matrix $\boldsymbol{H}_{M \times (I+1)L} = (\boldsymbol{H}_0, \boldsymbol{H}_1, \ldots, \boldsymbol{H}_I)$, and the singular value decomposition of matrix $\boldsymbol{H}$ is taken: $\boldsymbol{H} = \boldsymbol{U}_H \boldsymbol{\Sigma}_H \boldsymbol{V}_H^T$. $\boldsymbol{U}_H$ is a low-rank approximation to the column-space of $\boldsymbol{Y}$ with dimension $M \times (I+1)\, L$, where $\boldsymbol{Y} \approx \boldsymbol{U}_H \boldsymbol{U}_H^T \boldsymbol{Y}$. $\boldsymbol{Y}$ is then projected onto $\boldsymbol{U}_H$ to produce $\boldsymbol{T}_{(I+1)L \times N} = \boldsymbol{U}_H^{\boldsymbol{T}} \boldsymbol{Y}$. The SVD of $\boldsymbol{T} = \boldsymbol{U}_T \boldsymbol{\Sigma}_T \boldsymbol{V}_T^T$ can be computed efficiently and approximates the SVD of $\boldsymbol{Y}$ since $\boldsymbol{Y} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^T \approx \boldsymbol{U}_H \boldsymbol{T} = \boldsymbol{U}_H \boldsymbol{U}_T \boldsymbol{\Sigma}_T \boldsymbol{V}_T^T$. For the PCA, we are only interested in the left K columns of $\boldsymbol{V}_T$ and the first K entries along the diagonal of $\boldsymbol{\Sigma}_T$.

FastPCA runs in linear time and memory relative to M and N. There are $O(I)$ matrix multiplications where each multiplication takes $O(MNL)$ time. Then, the SVD of $\boldsymbol{H}$ takes $O\left(MI^2L^2\right)$ and the SVD of $\boldsymbol{T}$ takes $O\left(NI^2L^2\right)$ time. Taking I and L to be constants, the overall running time simplifies to $O(MN)$. This is much faster than traditional $O\left(MN^2 + N^3\right)$ PCA methods as well as the $O\left(MN^2\right)$ of flashpca.

1.4.3 Selection statistic

We first consider the simple case of an ancestral population that split into two extant populations with genetic distance F_{ST}. We consider the allele frequencies at SNP i for the ancestral population (p_i) and the two extant populations (p_{i1} and p_{i2}). If there is no selection and SNPs are randomly ascertained, $p_{i1} - p_{i2}$ has expectation 0 (because allele frequencies can drift either up or down in each population) and variance $2p_i(1-p_i)F_{ST}$[94].

In the case where p_i is not close to 0 or 1 and F_{ST} is small, the distribution of this difference approximately follows a normal distribution:

$$\begin{aligned} E\left[p_{i1} - p_{i2}\right] &= 0 \\ Var\left[p_{i1} - p_{i2}\right] &= 2p_i\left(1 - p_i\right)F_{ST} \\ p_{i1} - p_{i2} &\sim N\left[0, 2p_i\left(1 - p_i\right)F_{ST}\right], \quad F_{ST} \ll 1,\ 0 \ll p_i \ll 1 \end{aligned} \tag{1.2}$$

In practice, we do not have access to either the ancestral allele frequency or the extant population allele frequencies. Instead, we have sample allele frequencies for the two extant populations, $\hat{p}_{1i}$ and $\hat{p}_{i2}$. Assuming a large enough sample size from each population (N_1 and N_2) and that the true population allele frequency is not close to 0 or 1, these sample allele frequency estimates approximately follow a normal distribution with respect to the true allele frequencies. If we additionally assume that the ancestral allele frequency can be approximated by averaging the sample allele frequencies and that the true population allele frequencies are not that different, the sample allele frequency difference also follows a normal distribution[5,111,12]:

$$\begin{aligned} &\hat{p}_{i1} \sim N\left[p_{i1}, \frac{p_{i1}\left(1 - p_{i1}\right)}{2N_1}\right],\ \hat{p}_{i2} \sim N\left[p_{i2}, \frac{p_{i2}\left(1 - p_{i2}\right)}{2N_2}\right],\ N_1, N_2 \gg 0, 0 \ll p_{i1}, p_{i2} \ll 1 \\ &D_i = \hat{p}_{i1} - \hat{p}_{i2} \sim N\left[0, \sigma_D^2\right] = N\left[0, \hat{p}_i\left(1 - \hat{p}_i\right)\left(2F_{ST} + \frac{1}{2N_1} + \frac{1}{2N_2}\right)\right], \\ &p_i \approx \hat{p}_i = \frac{\hat{p}_{i1} + \hat{p}_{i2}}{2},\ p_{i1} \approx p_{i2} \end{aligned} \tag{1.3}$$

Below, we build the intuition behind our PC-based statistic by rewriting the discrete-population statistic using vector notation, then extending this statistic to individuals with

fractional ancestries, and then to continuous-valued PCs.

In the case with two discrete populations, we define a vector $\boldsymbol{\alpha}$ where α_j indicates the ancestry in population 1 (e.g. $\alpha_j = 1$ if sample j is in population 1 and 0 if sample j is in population 2). D_i can be rewritten as:

$$\hat{p}_1 = \frac{\boldsymbol{x}_i\boldsymbol{\alpha}}{2\mathbf{1}^T\boldsymbol{\alpha}}, \quad \hat{p}_2 = \frac{\boldsymbol{x}_i\left(\mathbf{1}-\boldsymbol{\alpha}\right)}{2\mathbf{1}^T\left(\mathbf{1}-\boldsymbol{\alpha}\right)}, \quad D_i = \frac{\boldsymbol{x}_i\boldsymbol{\alpha}}{2\mathbf{1}^T\boldsymbol{\alpha}} - \frac{\boldsymbol{x}_i\left(\mathbf{1}-\boldsymbol{\alpha}\right)}{2\mathbf{1}^T\left(\mathbf{1}-\boldsymbol{\alpha}\right)} \tag{1.4}$$

If we run PCA on the normalized genotype matrix $\boldsymbol{Y}$ from a sample with two discrete populations, we would ideally get an eigenvector $\boldsymbol{v}$ that has value v_1 for individuals in population 1 and $-v_2$ for individuals in population 2, where (since $\boldsymbol{v}^T\mathbf{1} = 0$, $\boldsymbol{v}^T\boldsymbol{v} = 1$)

$$v_q = \frac{1}{N_q}\sqrt{\frac{N_1N_2}{N}} \tag{1.5}$$

In this case, D_i can be rewritten as:

$$D_i = \frac{1}{2}\sqrt{\frac{N_1N_2}{N}}\boldsymbol{x}_i\boldsymbol{v} \tag{1.6}$$

In the limiting case where F_{ST} approaches 0, the statistic becomes:

$$\frac{D_i^2}{\sigma_D^2} = \frac{\frac{1}{4}\frac{N_1N_2}{N}\left(\boldsymbol{x}_i\boldsymbol{v}\right)^2}{\hat{p}_i\left(1-\hat{p}_i\right)\left(\frac{1}{2N_1}+\frac{1}{2N_2}\right)} = \left[\left(\frac{\boldsymbol{x}_i - 2\hat{p}_i\mathbf{1}^T}{2\hat{p}_i\left(1-\hat{p}_i\right)}\right)\boldsymbol{v}\right]^2 = \left[\boldsymbol{y}_i\boldsymbol{v}\right]^2 \tag{1.7}$$

Thus, the square of the SNP weight follows a chi-square 1-d.o.f. distribution in the case where $F_{ST} \to 0$. In the case where $F_{ST} \neq 0$, then the scaling parameter has to be changed, but D_i still follows a normal distribution.

In the case with fractional ancestry ($\alpha_j \in [0,1]$), $\hat{p}_1$, $\hat{p}_2$ and D_i can still be estimated

using equation (1.4). The individual $\hat{p}_q$s will still asymptotically follow a normal distribution (because of the Lyapunov central limit theorem[16]), but will be correlated due to individuals with fractional ancestry contributing to both estimates. Thus, D_i will still follow a normal distribution, but the variance of equation (1.3) will not hold.

Now consider the case where we do not have fractional ancestries, but rather an eigenvector that separates individuals along some axis of variation. (We assume that extreme outlier individuals detected by PCA have been removed[112], as PCs dominated by such outliers may violate normality assumptions.) We can treat the eigenvector as a linear transformation of the ancestry vector:

$$\boldsymbol{\alpha} = \beta_0 + \beta_1 \boldsymbol{v} \tag{1.8}$$

Substituting these values into (1.4), we find:

$$D_i = \frac{\beta_1}{2N\beta_0\left(1-\beta_0\right)} \boldsymbol{x}_i \boldsymbol{v} \propto \boldsymbol{y}_i \boldsymbol{v} \tag{1.9}$$

Thus, our new selection statistic D_i is based on the dot product of the normalized genotypes and the eigenvector. Since the variance of D_i is not known, it will need to be rescaled in order to follow a $N\left(0, 1^2\right)$ distribution.

If we are operating on the same set of SNPs that we used for PCA, then the rescaling of $\boldsymbol{y}_i \boldsymbol{v}$ is straightforward. Because PCA is the same as SVD, we see that:

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \\ \boldsymbol{U} &= \boldsymbol{Y}\boldsymbol{V}\boldsymbol{\Sigma}^{-1} \end{aligned} \tag{1.10}$$

Here, $\boldsymbol{V}$ contains the right singular vectors which are equivalent to the PCs, $\boldsymbol{U}$ contains the left singular vectors which are rescaled SNP weights and $\boldsymbol{\Sigma}$ contains the singular values which are the square roots of the eigenvalues of the GRM. $\boldsymbol{V}$ and $\boldsymbol{U}$ are unitary, so the columns of $\boldsymbol{U}$ are guaranteed to have a norm of 1. Multiplying $\boldsymbol{U}$ by $\sqrt{M}$ will then produce a properly normalized vector of differences $\boldsymbol{D} = (D_1, D_2, \ldots, D_M)^T$. In other words:

$$\begin{aligned} \frac{\sqrt{M}}{\Sigma_k}\boldsymbol{y}_i\boldsymbol{v}_k &\sim N(0,1) \\ \frac{M}{\Sigma_k^2}(\boldsymbol{y}_i\boldsymbol{v}_k)^2 &\sim \chi_1^2 \end{aligned} \tag{1.11}$$

In the case of non-random SNP ascertainment and non-random choice of reference and variant allele, the expectation of D_i may be non-zero. However, if we randomly flip the reference and variant alleles in such a situation, the resulting principal components and values of D_i remain unchanged up to a factor of -1 and the expectation of D_i becomes 0. As a result, even if there are systematically positive or negative SNP loadings, D_i^2 still follows a chi-square 1-d.o.f distribution.

In the case where we are computing selection statistics on a different set of SNPs than the one for which we computed PCs, then the above property is not guaranteed to hold. Specifically, inflation can occur if SNPs with higher differentiation tend to have higher LD, which can occur as a consequence of true selection signals[123].

One assumption underlying the statistic is that the true minor allele frequency is not extremely small, otherwise the assumption of normality will not hold[12]. For this reason, the selection statistic was only computed for those SNPs containing minor allele frequency greater than 1% in our sample.

1.4.4 Simulation framework

Genotypes were simulated at M independent SNPs and N independent individuals in four steps:

1. The ancestral allele frequency (p_i) for a given SNP i was sampled from a $Uniform(0.05, 0.95)$ distribution.
2. Allele frequencies for Q populations ($\boldsymbol{P}_i = (p_{i1}, p_{i2}, \ldots, p_{iQ})^T$) were generated by simulating random drift (see below).
3. Admixture ($\boldsymbol{\alpha}_j$) for individual j was sampled from a $Dirichlet(\boldsymbol{a})$ distribution.
4. Genotype g_{ij} was sampled from a $Binomial(2, \boldsymbol{\alpha}_j^T \boldsymbol{P}_i)$ distribution.

Population allele frequencies were generated by simulating random drift in Q populations of fixed size N_e for τ generations and stored in $Q \times 1$ vector $\boldsymbol{P}_i = (p_{i1}, p_{i2}, \ldots, p_{iQ})^T$. The number of alternate alleles z_{iqt} at SNP i in population q at generation t were sampled from a $Binomial\,(2N_e, p_{i,q,t-1})$ distribution, where p_{iq0} is the ancestral allele frequency p_i. The population allele frequency at this generation was then calculated as $p_{iqt} = \frac{z_{iqt}}{2N_e}$. For most simulations, population allele frequency simulations were run for $\tau = 200$ total generations and population size N_e was calculated for a target F_{ST} by using the formula $F_{ST} = -\log\left(1 - \frac{\tau}{2N_e}\right)$ [12]. For $F_{ST} \approx 0.1$, 0.01 and 0.001, $N_e = 1k$, $10k$ and $100k$ respectively. To detect the effect of population bottlenecks at the same level of F_{ST}, simulations were also run for $\tau = 20$ and $N_e = 100$, $1k$ and $10k$, again producing populations with genetic distance $F_{ST} \approx 0.1$, 0.01 and 0.001. Most simulations were run with two populations, but we also simulated 5 populations with a phylogenetic structure as follows. We set $N_e = 10k$ and $\tau = 200$ for populations 1 and 2, and $\tau = 180$ for an intermediary ancestral population of populations 3, 4 and 5, yielding allele frequency

p_i^*. This was then fed back into the random drift model for an additional 20 generations for populations 3, 4 and 5. The pairwise genetic distance between populations 3, 4 and 5 is $F_{ST} \approx 0.001$ while the genetic distance between any other pair of populations is $F_{ST} \approx 0.01$.

We also considered simulations with admixed samples. In these simulations, the $Q \times 1$ population membership vector $\boldsymbol{\alpha}_j$ for individual j was sampled from a $Dirichlet\,(\boldsymbol{a})$ distribution, where $\boldsymbol{a}$ is a vector containing ancestry weightings. In the most simple case $\boldsymbol{a} = a\mathbf{1}$, where a is the admixture coefficient. For $a = 0$, this does not form a proper distribution and instead ancestry was selected by alternating individual ancestry between each of the populations. Increasing this coefficient increases admixture. When $a = 1$, this is effectively a uniform distribution and when $a > 1$, the mode of the distribution is one containing even admixture between all the populations.

The individual ancestries $\boldsymbol{\alpha}_j$ make up the rows of ancestry matrix $\boldsymbol{A}$, which has dimension $N \times Q$. Multiplying this ancestry matrix by the population allele frequency vector ($\boldsymbol{P}_i$), which (for a given SNP i) has length Q, generated an $N \times 1$ vector of allele frequencies for each individual ($\boldsymbol{P}'_i = \boldsymbol{A}\boldsymbol{P}_i$). Individual genotypes g_{ij} were generated from a $Binomial(2, P'_{ij})$ distribution.

To assess running time, the simulated datasets had $F_{ST} = 0.01$, $M = 100k$ SNPs and $N \approx \{1k, 1.5k, 2k, 3k, 5k,\ 7k, 10k, 15k, 20k, 30k, 50k, 70k, 100k\}$ individuals (since we used 6 populations of equal sample size, we rounded N to multiples of 6). Throughout this paper we report CPU time, but due to multi-threading present in the GSL[42] and OpenBLAS libraries, run time was about 60% of CPU time. FastPCA accuracy was assessed using $M = 50k$ SNPs and $N \approx 10k$ individuals at $F_{ST} = \{0.001, 0.002, \ldots, 0.010\}$. Cal-

ibration and power of the selection statistic was assessed using 2 populations at $F_{ST} = \{0.1, 0.05, 0.02, 0.01, 0.005, 0.002, 0.001, 0.0005\}$, and also using 5 populations withthe tree structure described above. We set $M = 60k$, the effective number of independent SNPs in genotype array data[156]. When testing the power of the statistic, we wished to control the absolute difference in allele frequencies (D) between pairs of populations. For this purpose, SNPs under selection were generated in a similar manner as the above, except population allele frequencies were fixed at $p_{iq*} = 0.5 + \frac{D}{2}$ for one population and $p_{iq} = 0.5 - \frac{D}{2}$ for the remaining population(s); this approximates allele frequency differences under a population genetic selection model with strong selection in one population, because the magnitude of allele frequency differences caused by strong selection is much larger than the magnitude of allele frequency differences caused by genetic drift.

1.4.5 Assessing PC accuracy

Accuracy was assessed via the Mean of Explained Variances (MEV) of eigenvectors. Two different sets of K N-dimensional principal components each produce a K-dimensional column space. A metric for the performance of a PCA algorithm against some baseline is to see how much the column spaces overlap. This is done by projecting the eigenvectors of one subspace onto the other and finding the mean lengths of the projected eigenvectors. If we have a reference set of PCs $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K)$ against which we wish evaluate the performance a set of computed PCs $(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_K)$, then the performance calculation becomes:

$$MEV = K^{-1} \sum_{j=1}^{K} \sqrt{\sum_{j=1}^{K} (\boldsymbol{v}_k \cdot \boldsymbol{u}_j)^2} = K^{-1} \sum_{j=1}^{K} \left\| \boldsymbol{U}^T \boldsymbol{v}_k \right\| \tag{1.12}$$

Here, $\boldsymbol{U}$ is a matrix whose column vectors are the PCs which we are testing. The test matrix can either be the result of another computation or the truth for a simulated sample. K eigenvectors can describe the population structure in a dataset with $K + 1$ populations. They can be constructed by first creating a vector $\boldsymbol{v}_k^* = \left(v_{k,1}^*, v_{k,2}^*, \ldots, v_{k,N}^*\right)$ where $v_{k,j}^* = 1$ if individual j is in population k and 0 otherwise. The set of eigenvectors $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K\}$ are constructed by taking K of these vectors, normalizing them to have mean 0, and scaling/orthogonalizing them via the Gram-Schmidt process.

1.4.6 GERA data set

The GERA dataset includes 62,318 individuals from Northern California typed on a European-specific 670,176-SNP array[6]. This dataset underwent two levels of filtration: a quality control step to produce the QC set of SNPs used to detect natural selection, and a second step used to produce the LD-pruned set of SNPs for PCA.

For the QC step, individuals were filtered to remove those with missing sex information, individuals related according to the provided pedigree data or with observed genomic relatedness greater than 0.05 in the GRM[117] and individuals with less than 90% European ancestry as predicted by SNPweights[26] using a worldwide dataset containing European, African, and Asian ancestry. After filtering, 54,734 individuals remained. Additionally, SNPs were initially filtered to remove non-autosomal SNPs, SNPs with minor allele frequency less than 1%, and SNPs with >1% missing data, leaving 608,981 SNPs.

The second stage of filtering removed SNPs that failed PLINK's Hardy-Weinberg Equilibrium test[117] with $p < 10^{-6}$, and performed LD-pruning using PLINK. Due to regions of long-range LD, LD persisted even after one filtering run. Multiple rounds of LD filtering were performed using an r^2 cutoff of 0.2 until additional rounds of LD filtering did not remove additional SNPs, leaving 162,335 SNPs.

FastPCA was run on the pruned set of 162,335 SNPs, while selection statistics were computed on the full set of 608,981 SNPs, prior to H-W filtering and LD-pruning. We note that many of the SNPs producing signals of selection generated significant H-W p-values (see Results - e.g. H-W $p = 1.37 \times 10^{-79}$ for LCT SNP rs6754311), which is an expected consequence of unusual population differentiation.

SNPweights[26] was used to predict fractional Northwest European, Southeast European, and Ashkenazi Jewish ancestry for each individual. For plotting purposes, percentage ancestry in each of these three populations was mapped to an integer in $[0, 255]$, which was then used for the RGB color value for that sample, so a NW sample would appear red, SE would appear green and AJ would appear blue.

1.4.7 PC Projection

POPRES[92] individuals were projected onto these PCs. The left singular vectors ($\boldsymbol{U}$) were generated by multiplying normalized genotypes for all SNPs in GERA ($\boldsymbol{Y}_{GERA}$) by the PCs ($\boldsymbol{V}$) and scaling by the singular values ($\boldsymbol{\Sigma}$), the number of SNPs used to calculate the PCs (M) and the number of SNPs used for projection (M_{GERA}): $\boldsymbol{U} = \boldsymbol{Y}_{GERA}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\sqrt{M/M_{GERA}}$. Projected PCs were then calculated by multiplying the corresponding set of SNPs in POPRES by these singular vectors and scaling again by the

singular values: $\boldsymbol{V}_{POPRES} = \boldsymbol{Y}_{POPRES}^{T}\boldsymbol{U}\boldsymbol{\Sigma}^{-1}$. The projected individuals were overlaid on the PCA plot of GERA individuals and colored according to population membership and consistently with population assignment from SNPweights[26].

1.5 Results

1.5.1 FastPCA Simulations

We used simulated data to compare the running time and memory usage of FastPCA to three previous algorithms: smartpca[112,103], PLINK2-pca[117], and flashpca[1]. We simulated genotype data from six populations with a star-shaped phylogeny using 100k SNPs (typical for real data after LD-pruning) and up to 100k individuals (see Methods). For each run, running time was capped at 100 hours and memory usage was capped at 40GB. The running time and memory usage of FastPCA scaled linearly with simulated dataset size (i.e. $O(MN)$ cost) (Figure 1.1), compared with quadratically or cubically for other methods. The computation became intractable at 50k-70k individuals for smartpca, PLINK2-pca and flashpca. The largest dataset, with 100k SNPs and 100k individuals, required only 56 minutes and 3.2GB of memory with FastPCA (Table A.1). (We also note that shellfish, a parallel PCA implementation, requires $O\left(MN^2 + N^3\right)$ and is not computationally tractable on large data sets, as previously demonstrated[1]). Thus, FastPCA - unlike other publicly available software packages for analyzing genetic data - enables rapid principal components analysis without specialized computing facilities.

We next assessed the accuracy of FastPCA, using PLINK2-pca[117] as a benchmark. We used the same simulation framework as before, with 10k individuals (1,667k individuals

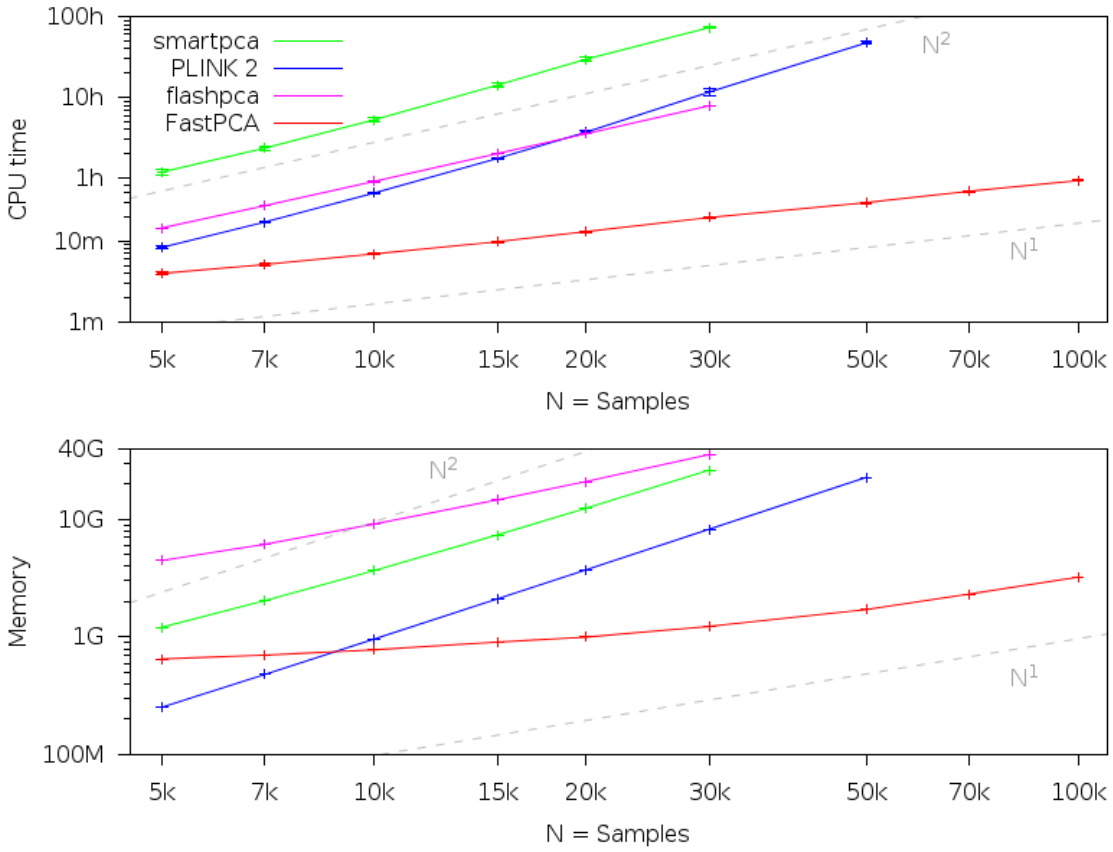

Figure 1.1: Running time and memory requirements of FastPCA and other algorithms. The CPU time and memory usage of FastPCA scale linearly with the number of individuals. On the other hand, smartpca and PLINK2-pca scale between quadratically and cubically, depending on whether computing the GRM (quadratic) or the eigendecomposition (cubic) is the rate-limiting step. The running time of flashpca scales quadratically (because it computes the GRM), but its memory usage scales linearly because it stores the normalized genotype matrix in memory. With 50k individuals, smartpca exceeded the time constraint (100 hours) and flashpca exceeded the memory constraint (40GB). With 70k individuals, PLINK2-pca exceeded the memory constraint (40GB). Run times are based on one core of a 2.26-GHz Intel Xeon L5640 processor; we caution that run time comparisons may vary by a small constant factor as a function of the computing environment. Numerical data is provided in Table A.1.

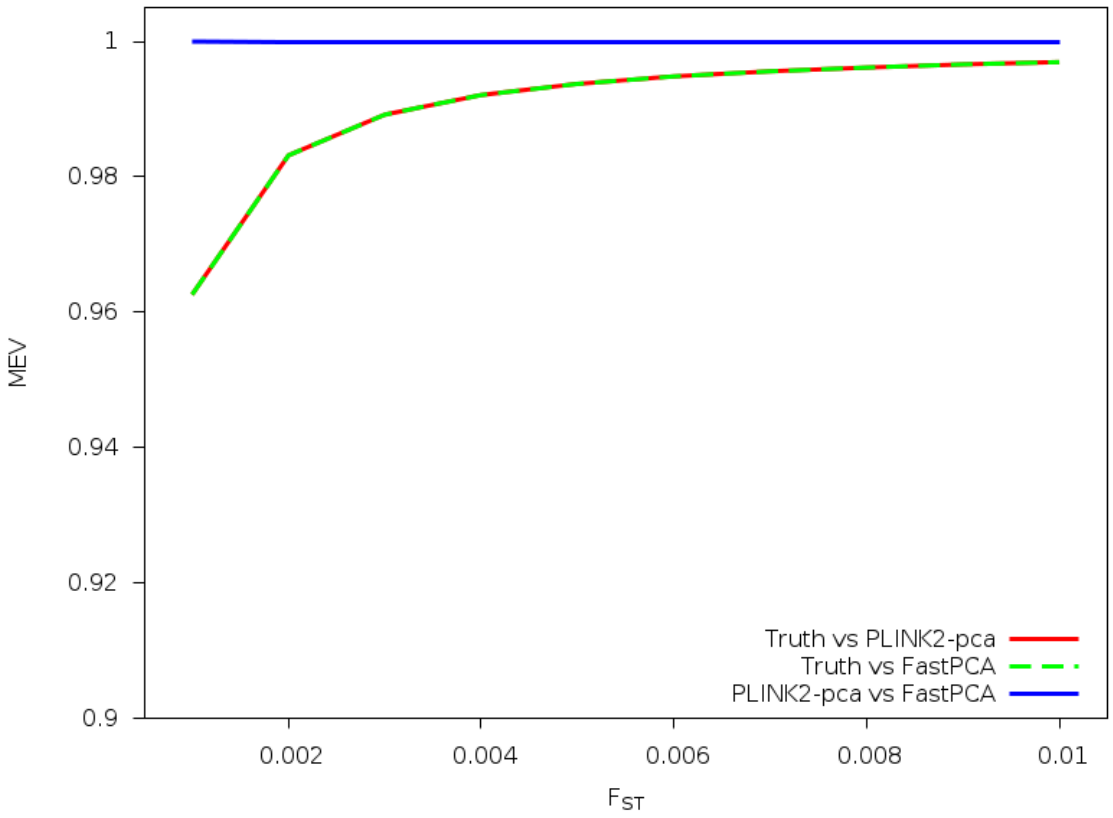

Figure 1.2: Accuracy of FastPCA and PLINK2-pca. FastPCA and PLINK2-pca were run on simulated populations of varying divergence. The simulated data comprised 50k SNPs and 10k total individuals from six subpopulations derived from a single ancestral population. PCs computed by PLINK2-pca and FastPCA were compared to the true population PCs and to each other using the Mean of Explained Variances (MEV) metric (see text). FastPCA explained the same amount of true population variance as PLINK2-pca in all experiments, and the methods output nearly identical PCs (MEV>0.999).

per population) and 50k SNPs. We varied the divergence between populations, as quantified by F_{ST}[13]. We assessed accuracy using the Mean of Explained Variances (MEV) of the 5 population structure PCs (see Methods). We determined that the results of FastPCA and PLINK-pca were virtually identical (Figure 1.2). This indicates that FastPCA performs comparably to standard PCA algorithms while running much faster.

1.5.2 PC-based Selection Statistic Simulations

We evaluated the calibration and power of the PC-based selection statistic. To evaluate calibration, we simulated 60k SNPs undergoing random drift with up to $N = 50$k individuals from two populations differentiated by $F_{ST} = \{0.1, 0.01, 0.001\}$. At all values of N and F_{ST}, the proportion of truly null SNPs reported as significant was well-calibrated at p-value thresholds ranging from 10^{-1} to10^{-5}. Similar results indicating appropriate calibration were obtained for simulations with admixture (Table A.2), as expected since the drift model still applies in the case of admixture[110]. The median of the selection statistic was slightly inflated at $F_{ST} = 0.1$ due to a deficiency in the tail (Figure A.1, Table A.2 and Table A.3), but well-calibrated at the small values of F_{ST} that correspond to our analyses of real data. The selection statistic in the presence of a population bottleneck performed identically to populations differentiated by the same F_{ST} level (Table A.3). We also simulated five populations with a phylogenetic structure (see Methods) that mimics the population structure found in the GERA data (see below) and found that the statistic remained well-calibrated here as well (Figure A.1 and Table A.4).

We evaluated power using the same number of SNPs and samples but at $F_{ST} = \{0.1, 0.05, 0.02, 0.01, 0.005, 0.002, 0.001, 0.0005\}$ and using a separate set of

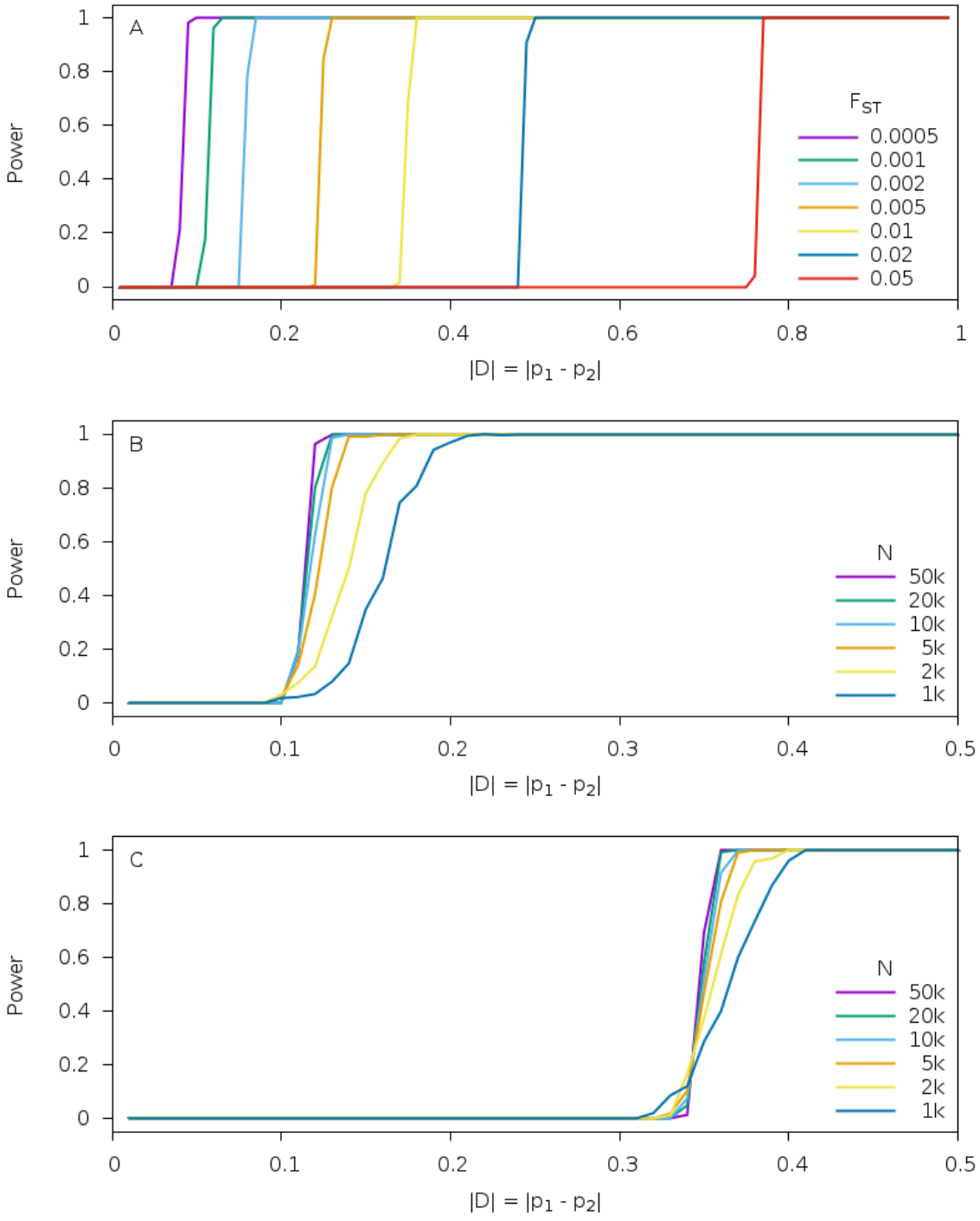

Figure 1.3: Power of PC-based selection statistic. The allele frequency difference at selected SNPs was varied between two populations separated by varying F_{ST}. The significance threshold was set to 8.3×10^{-7} based on 60K SNPs tested. (a) With 50k samples, the power curves for $F_{ST} = \{0.05, 0.02, 0.01, 0.005, 0.002, 0.001, 0.0005\}$ showed a phase change. (b) Varying the number of samples for $F_{ST} = 0.001$ demonstrated that this phase change was more gradual at smaller sample sizes. (c) Varying the number of samples at $F_{ST} = 0.01$ showed that the impact of sample size was less pronounced than at $F_{ST} = 0.001$.

SNPs under selection where the allele frequency between the two populations was varied ($|D| = |p_1 - p_2|$). The significance threshold was set to 8.3×10^{-7} based on 60K SNPs tested. There was no power to detect selection at $F_{ST} = 0.1$. We observed a phase-change in the power simulations that was sharper for smaller F_{ST}, where there was no power to detect selection below a specified allele frequency difference threshold, but there was complete power to detect selection at a slightly higher threshold (Figure 1.3a). We examined this effect in more depth using a range of samples sizes, and determined that the transition from no-power to complete-power was more sample size dependent at $F_{ST} = 0.001$ (Figure 1.3b) than at $F_{ST} = 0.01$ (Figure 1.3c), indicating that sample size is more important when analyzing more closely related populations. The PC-based selection statistic performed very similarly to the discrete-population test of selection[12] in the case of data from discrete subpopulations (Figure A.2). We also assessed effect of admixture on power by sampling ancestry for individuals between the two populations using a $Beta\,(a, a)$ distribution. We determined that increasing the admixture parameter a (which reduces the variation in ancestry across samples) had a similar effect to reducing sample size (Figure A.3).

1.5.3 Application of FastPCA to a European American Cohort

We ran FastPCA on the GERA cohort, a large European American dataset containing 54,734 individuals and 162,335 SNPs after QC filtering and LD-pruning (see Methods). This computation took 57 minutes and 2.6GB of RAM. PC1 and PC2 separated individuals along the canonical Northwest European (NW), Southeast European (SE) and Ashkenazi Jewish (AJ) axes[111], as indicated by labeling the individuals by predicted fractional

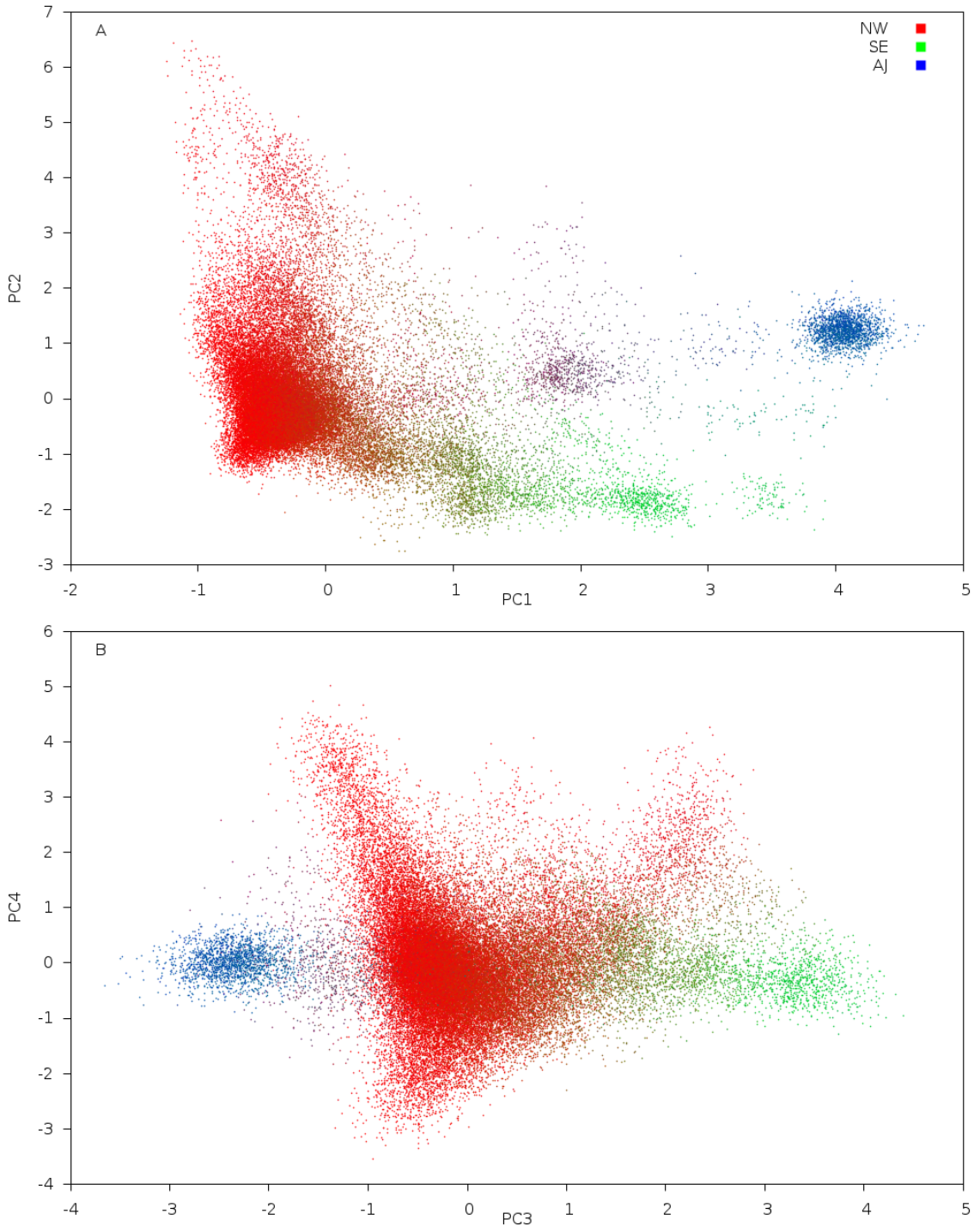

Figure 1.4: FastPCA results on GERA data set. FastPCA and SNPweights [26] were run on the GERA cohort and the principal components from FastPCA were plotted. Individuals were colored by mapping Northwest European (NW), Southeast European (SE) and Ashkenazi Jewish (AJ) ancestry estimated by SNPweights to the red/green/blue color axes (see Methods). PC1 and PC2 separate the GERA cohort into northwest (NW), southeast (SE) and Ashkenazi Jewish (AJ) subpopulations. PC3 separates the AJ and SE individuals, while PC3 and PC4 further separates the NW European individuals.

ancestry from SNPweights[26] (Figure 1.4). These results are consistent with Banda *et al.* 2015[6] which also examined this dataset. PC3 and PC4 detected additional population structure within the NW population.

To further investigate this subtle structure, we projected POPRES individuals from throughout Europe[92] onto these PCs[103] (see Methods). This analysis recapitulated the position of SE populations via the placement of the Italian individuals, and determined that PC3 and PC4 separate the NW individuals into Irish (IR), Eastern European (EE) and Northern European (NE) populations (Figure 1.5). This visual subpopulation clustering was confirmed via k-means clustering on the top 4 PCs, which consistently grouped the AJ, SE, NE, IR and EE populations separately (Figure A.4). We note that, in general, K PCs can cluster samples into $K + 1$ subpopulations.

1.5.4 Application of PC-based Selection Statistic to a European American Cohort

For each of the top PCs, we computed our PC-based selection statistic for 608,981 non-LD-pruned SNPs (see Methods). The resulting Manhattan plots for PCs 1-4 are displayed in Figure 1.6 (QQ plots are displayed in Figure A.5). Analyses of PCs 5-10 indicated that these PCs do not represent true population structure (Figure A.6), but are either dominated by a small number of long-range LD loci[141,36,160] or correlated with the missing data rate across individuals. Selection statistics for PCs 1-4 exhibited little or no inflation, particularly after removing Table 1.1 regions (Table A.5).

Genome-wide significant signals (listed in Table 1.1) included several known selection regions[11,30,147,22,106] and signals at *ADH1B*, *IGFBP3* and *IGH* (see below). Suggestive

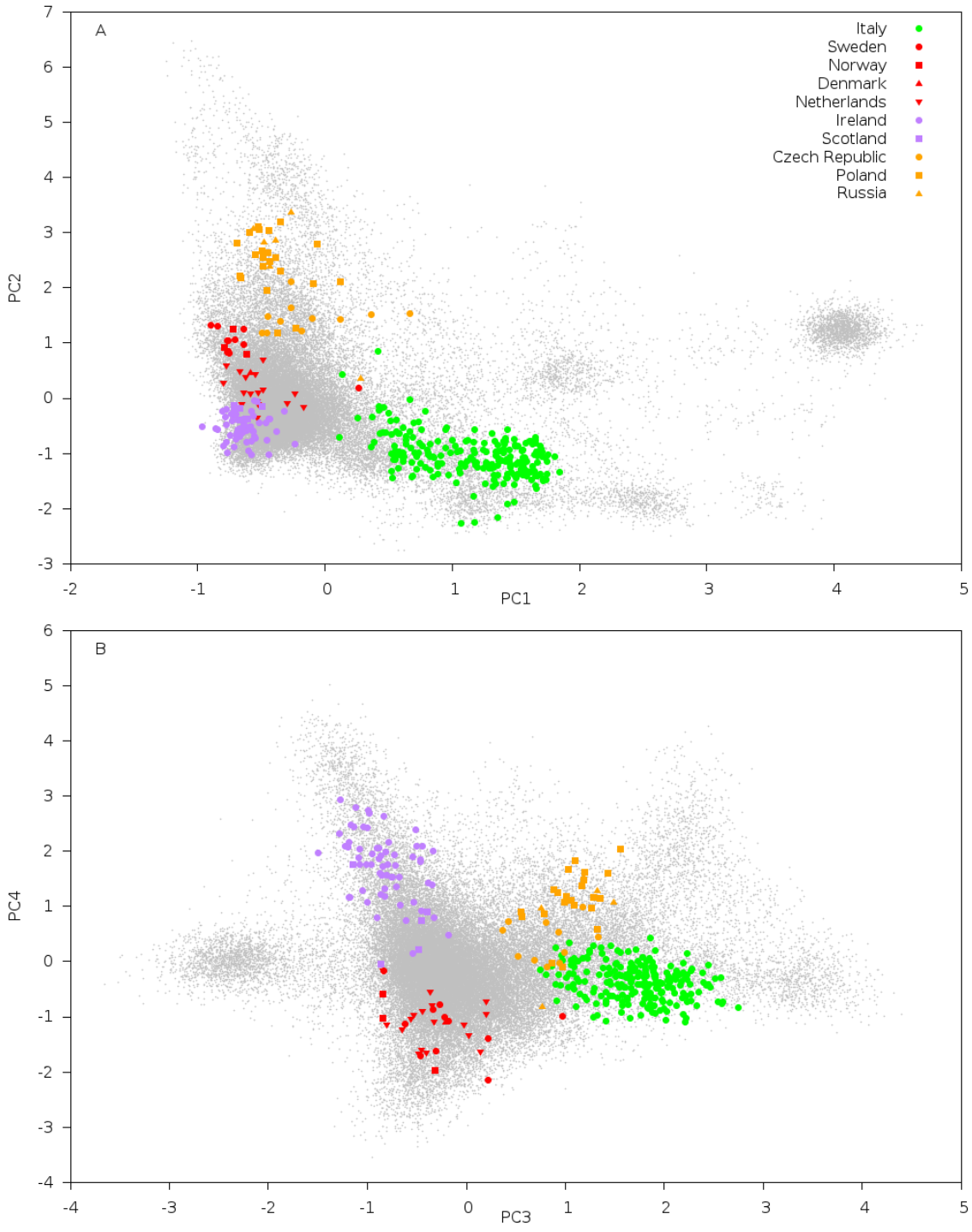

Figure 1.5: Separation of Irish, Eastern European and Northern European individuals in GERA data. We report results of projecting POPRES [92] individuals onto top PCs. The plot of PC3 vs PC4 shows that the Northwest European (NW) individuals are further separated into Irish and Eastern European and Northern European populations. Projected populations were colored based on correspondence to the ancestry assignment from SNPweights [26], except that Irish and Eastern European individuals were colored purple and orange, respectively, to indicate additional population structure.

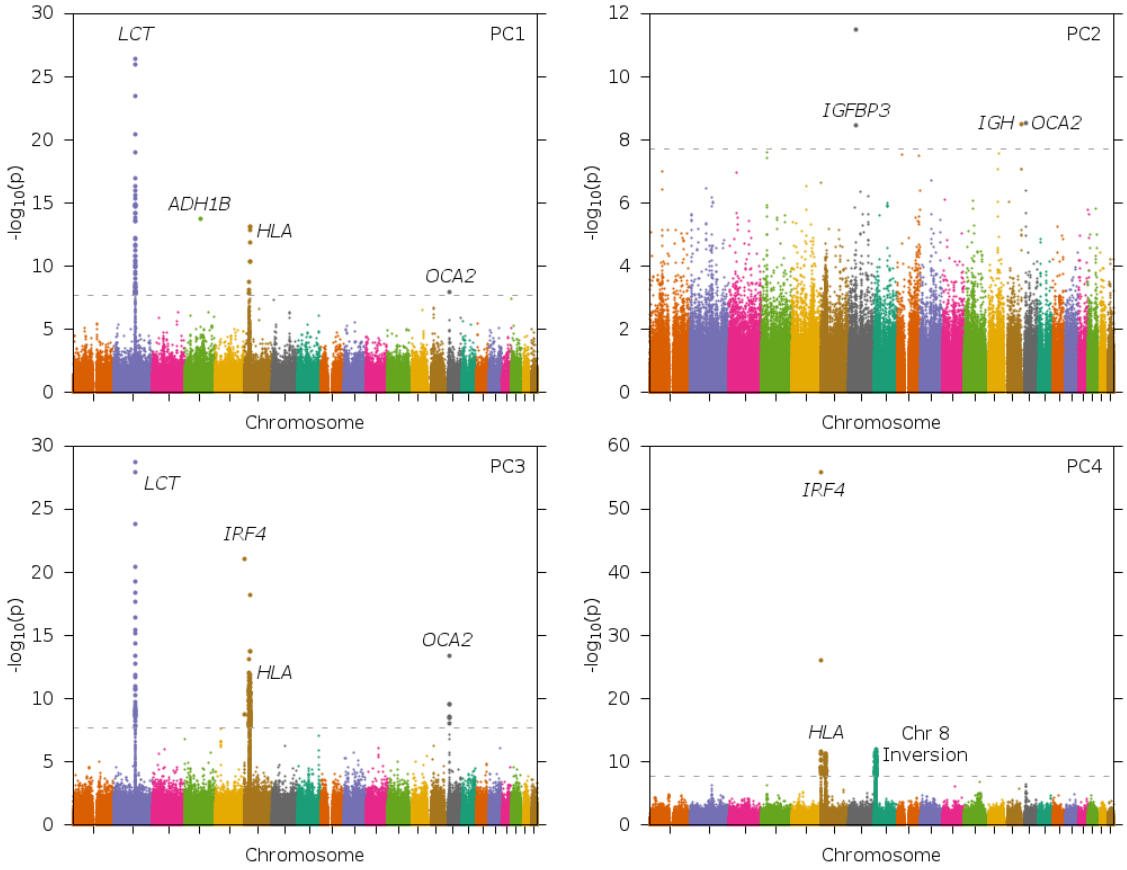

Figure 1.6: [Signals of selection in the top PCs of GERA data. We display Manhattan plots for selection statistics computed using each of the top 4 PCs. The grey line indicates the genome-wide significance threshold of 2.05×10^{-8} based on 2,435,924 hypotheses tested ($\alpha = 0.05$, 608,981 SNPs $\times$ 4 PCs).

signals were observed at additional known selection regions[22,125] (Table A.6). After removing the regions in Table 1.1, rerunning FastPCA and recalculating selection statistics, all of these regions remained significant except for a region on chromosome 8 with a known chromosomal inversion[141,36] (Figure A.7 and Table A.7). Thus, the remaining regions are not due to PC artifacts caused by SNPs inside these regions. We also found that a significantly greater proportion of SNPs under selection failed Hardy-Weinberg equilibrium, although the converse is not true, indicating that signals of selection are not a result of H-W artifacts (Figure A.8). Detecting subtle signals of selection benefited from the large sample size, as subsampling the GERA data set at smaller sample sizes and recomputing PCs and selection statistics generally led to less significant signals (Table 1.2). We note that several suggestive selection signals, including signals at the known selected loci *TLR1*[22] and *SLC45A2*[125], are on the cusp of being significant and further increases to sample size may increase power to detect selection at suggestive loci.

We identified a genome-wide significant signal of selection at rs1229984, a coding SNP (Arg47His) in the alcohol dehydrogenase gene (*ADH1B*) (Table 1.1). The allele rs1229984*T has been shown to have a protective effect on alcoholism risk[34,149,79,46] and to produce an REHH signal in East Asians[57,79,100,105], but was not previously known to be under selection in Europeans. (Previous studies noted the higher frequency of the rs1229984*T allele in western Asia compared to Europe, but indicated that selection or random drift were both plausible explanations[80,145].) We examined the allele frequency of the rs1229984*T allele in the five subpopulations: AJ, SE, NE, IR and EE (Table A.8). We observed allele frequencies of 0.21 in AJ, 0.10 in SE, and 0.05 or lower in other subpopulations, consistent with the higher frequency of the rs1229984*T in western Asia. A comparison of NE

Locus	Chr	Region (Mb)	PC	Best Hit	p-value
LCT[11]	2	134.8 - 137.6	1	rs6754311	4.15×10^{-27}
			3	rs4988235	1.83×10^{-29}
ADH1B	**4**	**100.5**	**1**	**rs1229984**	$\mathbf{1.67 \times 10^{-14}}$
IRF4[22,106]	6	0.3 - 0.5	3	rs12203592	8.69×10^{-22}
			4		1.83×10^{-56}
HLA[30]	6	30.8 - 33.3	1	rs382259	7.95×10^{-14}
			3	rs9268628	6.52×10^{-19}
			4	rs34707463	4.76×10^{-12}
IGFBP3	**7**	**45.3-45.9**	**2**	**rs150353309**	$\mathbf{3.14 \times 10^{-12}}$
Chr8 Inversion[141]	8	8.2 - 11.9	4	rs6984496	9.21×10^{-13}
IGH	**14**	**106.0-106.1**	**2**	**rs34614900**	$\mathbf{3.34 \times 10^{-9}}$
OCA2[147,106]	15	25.9 - 26.2	1	rs12916300	1.12×10^{-8}
			2		3.07×10^{-9}
			3		4.29×10^{-14}

Table 1.1: Genome-wide significant signals of selection in GERA data. We list regions with genome-wide significant ($\alpha = 0.05$, Bonferroni correction with 608,981 SNPs $\times$ 4 PCs = 2,435,924 hypotheses tested, $p < 2.05 \times 10^{-8}$) evidence of selection in the top 4 PCs. We provide previous reference(s) where available; remaining loci are indicated in bold font. The chromosome 8 inversion signal is due to a PC artifact (see main text). Regions with suggestive evidence of selection ($10^{-6} < p < 2.05 \times 10^{-8}$) are listed in Table A.5.

to the remaining subpopulations using the discrete subpopulation selection statistic[12] also produced a genome-wide significant signal after correcting for all hypotheses tested (Table A.9); this is not an independent experiment, but indicates that this finding is not due to assay artifacts affecting PCs.

To further understand the selection at this locus, we examined the allele frequency of rs1229984*T in 1000 Genomes project[136] populations, along with the allele frequency of the regulatory SNP rs3811801 that may also have been a target of selection in Asian populations[79]. The haplotype carrying rs3811801*A (and corresponding haplotype H7) was absent in populations outside of East Asia (Table A.10). This indicates that if natural selection acted on this SNP in Asian populations, selection acted independently at this locus in Europeans. One possible explanation for these findings is that rs1229984 is an

		Sample Size						
Locus	**SNP**	**Full**	**1k**	**2k**	**5k**	**10k**	**20k**	**50k**
LCT	rs6754311	2.15e-25	4.91e-17	2.97e-20	1.53e-23	1.17e-24	2.63e-25	1.02e-26
	rs4988235	1.15e-27	7.44e-17	9.80e-20	4.64e-23	3.11e-24	2.69e-25	1.62e-27
	rs17346504	8.41e-7	2.86e-2	1.25e-2	9.49e-4	6.03e-5	8.12e-6	9.80e-7
ADH1B	rs1229984	1.26e-13	3.91e-9	3.51e-11	1.97e-12	5.54e-13	1.50e-13	1.31e-13
IRF4	rs12203592	5.52e-55	3.15e-6	9.18e-12	7.47e-25	7.21e-36	7.02e-45	2.19e-54
HLA	rs382259	5.38e-13	8.68e-9	1.23e-10	7.07e-12	1.85e-12	7.51e-13	5.77e-13
	rs9268628	8.66e-18	3.62e-5	3.41e-7	5.97e-12	2.10e-14	2.68e-16	1.00e-17
	rs4394275	9.36e-12	8.40e-2	1.94e-3	1.44e-5	4.00e-8	7.86e-10	1.24e-11
IGFBP3	rs150353309	5.82e-12	5.90e-4	1.49e-5	2.72e-8	3.61e-10	3.34e-11	6.61e-12
IGH	rs34614900	5.23e-9	6.33e-3	2.24e-4	2.26e-6	2.01e-7	3.32e-8	5.32e-9
OCA2	rs12916300	2.80e-13	6.29e-6	1.07e-7	3.67e-9	1.94e-11	5.29e-12	3.11e-13
	rs2703951	5.11e-7	1.12e-1	2.45e-2	7.96e-4	7.17e-5	4.52e-6	5.74e-7
TLR1	rs5743611	5.42e-8	8.05e-3	4.27e-4	9.41e-6	1.19e-6	2.17e-7	5.60e-8
	rs4833095	6.52e-7	6.07e-4	3.37e-4	7.35e-5	3.64e-5	6.03e-6	7.10e-7
SLC45A2	rs16891982	6.89e-8	8.25e-4	2.17e-4	1.93e-5	4.55e-6	2.46e-7	7.31e-8

Table 1.2: Performance of natural selection statistic in subsampled data. The selection statistic was computed in random subsets of individuals of specified size for each SNP in Table 1.1 (except for the chromosome 8 inversion region) and the known selection regions *TLR1* [22] and *SLC45A2* [125] in Table A.6. We report the median selection statistic P-value across 100 random subsets.

older SNP under selection in Europeans, while rs3811801 is a newer SNP under strong selection in Asian populations leading to the common haplotype found in those populations.

The insulin-like growth factor-binding protein gene (*IGFP3*) had two SNPs reaching genome-wide significance. Genetic variation in *IGFBP3* has been associated with breast cancer[49], height, blood pressure[45] and hypertension[159], although the published associated SNPs are not in LD with the two SNPs we detected. The immunoglobulin heavy locus (*IGH*) had one genome-wide-significant SNP and two suggestive SNPs with p-value $< 10^{-6}$ (Table 1.1). Genetic variation in *IGH* has been associated with multiple sclerosis[19], although the published associated SNPs are not in LD with the three SNPs we detected. The *IGFBP3* and *IGH* SNPs each had substantially higher minor allele frequencies in Eastern Europeans, but were not genome-wide significant under the discrete subpopulation selection statistic[12] (Table A.11 and Table A.12). The existence of multiple SNPs at each of these loci with $p < 10^{-6}$ for the PC-based selection statistic suggests that these findings are not the result of assay artifacts.

1.6 Discussion

We have detected new, genome-wide significant signals of selection by applying a PC-based selection statistic to top PCs computed using FastPCA, a computationally efficient (linear-time and linear-memory) algorithm. Although mixed model association methods are increasingly appealing for conducting genetic association studies[156,83], we anticipate that PCA will continue to prove useful in population genetic studies, in characterizing population stratification when present in association studies, in supplementing mixed

model association methods by including PCs as fixed effects in studies with extreme stratification, and in correcting for stratification in analyses of components of heritability[153,155]. Our PC-based selection statistic extends previous statistics developed for discrete populations[12]. In contrast to previous work on detecting selection using PCs[160,132] or using the spatial ancestry analysis (SPA) method[157], our statistic is able to detect signals at genome-wide significance, a key consideration in genome scans for selection[14]. Our work demonstrates the advantages of comparing closely related populations in very large sample sizes to detect subtle signals of selection, whereas very recent studies applying related methods to smaller sample sizes detected genome-wide significant signals only at previously known loci[60,27]. In particular, we detected genome-wide significant evidence of selection in Europeans at *ADH1B*, which was previously reported to be under selection in East Asian populations[57,79,100,105] using REHH[123] (which can only detect relatively recent signals and does not work on standing variation[96]). We also detected genome-wide significant evidence of selection at the disease-associated *IGFBP3* and *IGH*. While the SNPs under selection at these loci are not in LD with the disease-associated SNPs identified in previous association studies, these genes are biologically important and there may be other phenotypes associated with the selected SNPs. Although we emphasize the importance of genome-wide significance, loci with suggestive signals of selection that do not reach genome-wide significance could potentially be used to increase the power of disease mapping[68].

We note that our work has several limitations. First, top PCs do not always reflect population structure, but may instead reflect assay artifacts[28] or regions of long-range LD[141]; however, PCs 1-4 in GERA data reflect true population structure and not assay

artifacts, because the PCs (and the signals of selection they detect) remained nearly unchanged after removing regions with significant signals of selection (Table 1.1) and rerunning PCA. Second, common variation may not provide a complete description of population structure, which may be different for rare variants[88]; we note that based on analysis of real sequencing data with known structure, we recommend that LD-pruning and removal of singletons (but not all rare variants) be applied in data sets with pervasive LD and large numbers of rare variants. Third, our selection statistic is only capable of detecting that selection occurred, but not when or where it occurred; indeed, top PCs may not perfectly represent the geographic regions in which selection occurred, underscoring that interpretation of results can be a fundamental limitation of model-free methods. Fourth, our selection statistic performs best when allele frequencies vary linearly along a PC; the SPA method[157] (see above) models allele frequency as a logistic function and is not constrained by this limitation. Despite these limitations, we anticipate that FastPCA and our PC-based selection statistic will prove valuable in analyzing the very large data sets of the future.

2

Population structure and natural selecton in the United Kingdom

2.1 Preface

The following work was published in the November 2016 issue of *The American Journal of Human Genetics*[44], titled *Population structure of UK Biobank and ancient Eurasians reveals adaptation at genes influencing blood pressure* with co-authors Po-Ru

Loh, Swapan Mallick, Nick J. Patterson and Alkes L. Price. We apply the FastPCA algorithm and PC-based natural selection statistic to 113k samples in the UK Biobank dataset[131]. We detected five axes of variation in individuals of British ancestry in the UK, onto which we projected ancient DNA samples[73,52,87] to detect how ancient population migrations affected the genetic make-up of the UK. We detected natural selection in fucosyltransferase 2 and combined our natural selection statistic with one that compares modern populations to ancestral ones[87] to determine additional signals of selection at genes that affect blood pressure.

2.2 Abstract

Analyzing genetic differences between closely related populations can be a powerful way to detect recent adaptation. The very large sample size of the UK Biobank is ideal for detecting selection using population differentiation, and enables an analysis of UK population structure at fine resolution. In analyses of 113,851 UK Biobank samples, population structure in the UK is dominated by 5 principal components (PCs) spanning 6 clusters: Northern Ireland, Scotland, northern England, southern England, and two Welsh clusters. Analyses with ancient Eurasians show that populations in the northern UK have higher levels of Steppe ancestry, and that UK population structure cannot be explained as a simple mixture of Celts and Saxons. A scan for unusual population differentiation along top PCs identified a genome-wide significant signal of selection at the coding variant rs601338 in *FUT2* ($p = 9.16 \times 10^{-9}$). In addition, by combining evidence of unusual differentiation within the UK with evidence from ancient Eurasians, we identified genome-wide significant ($p < 5 \times 10^{-8}$) signals of recent selection at two additional loci: *CYP1A2/CSK* and

F12. We detected strong associations to diastolic blood pressure in the UK Biobank for the variants with selection signals at *CYP1A2/CSK* ($p = 1.10 \times 10^{-19}$) and for variants with ancient Eurasian selection signals in the *ATXN2/SH2B3* locus ($p = 8.00 \times 10^{-33}$), implicating recent adaptation related to blood pressure.

2.3 Introduction

Detecting signals of selection can provide biological insights into adaptations that have shaped human history[123,95,96,126]. Searching for genetic variants that are unusually differentiated between populations is a powerful way to detect recent selection on standing variation[130]; this approach has been applied to detect signals of selection linked to lactase persistance[11,143], fatty acid decomposition[41], hypoxia response[158,15,84], malaria resistance[56,5,51], and other traits and diseases[71,104,58,69].

Leveraging population differentiation to detect selection is particularly powerful when analyzing closely related subpopulations with large sample sizes[12]. Here, we analyze 113,851 samples of UK ancestry from the UK Biobank in conjunction with recently published People of the British Isles (PoBI)[77] and ancient DNA[73,52,87,127] data sets to draw inferences about population structure and recent selection. We employ a recently developed selection statistic that detects unusual population differentiation along continuous principal components (PCs) instead of between discrete subpopulations[43], and combine our results with independent results from ancient Eurasians[87]. We detect three interesting signals of selection, and show that genetic variants at these and previously reported[87] signals of selection are strongly associated to diastolic blood pressure in UK Biobank samples.

2.4 Methods

2.4.1 UK Biobank data set

The UK Biobank phase 1 data release contains 847,131 SNPs and 152,729 samples. We removed SNPs that were multi-allelic, had a genotyping rate less than 99%, or had minor allele frequency (MAF) less than 1%. We also removed samples with non-British ancestry (including admixed samples) as well as samples with a genotyping rate less than 98%. This left 510,665 SNPs and 118,650 samples, a data set that we call "QC*." Using PLINK2[25], we removed SNPs not in Hardy-Weinberg equilibrium ($p < 10^{-6}$), and we LD-pruned SNPs to have $r^2 < 0.2$. We then generated a genetic relationship matrix (GRM) and removed one of each any pair of samples with relatedness greater than 0.05. This data set, which we call "LD," contained 210,113 SNPs and 113,851 samples. Taking the full set of SNPs from the QC* data set and the set of unrelated samples from the LD data set produces the final "QC" dataset.

2.4.2 PoBI and POPRES data sets

The 2,039 UK PoBI samples were a subset of the 4,371 samples collected as part of the PoBI project[77]. The 2,039 samples were a subset of the 2,886 samples genotyped on the Illumina Human 1.2M-Duo genotyping chip, with 2,510 samples passing QC procedures and 2,039 samples with all four grandparents born within 80km of each other. This dataset allows us to examine the population genetics of the UK prior to the migrations of the late 19^{th} and early 20^{th} centuries. We also examined 2,988 European POPRES samples from the LOLIPOP and CoLaus collections[92]. These samples were genotyped on the

Affymetrix GeneChip 500K Array. The POPRES dataset allows us to compare the UK Biobank population structure with that of continental Europe.

2.4.3 Ancient DNA data sets

Ancient DNA was gathered from several regions. 9 Steppe samples were collected from the Yamna oblast in Russia[52], 7 west-European hunter-gatherers from Loschbour[73], 26 Neolithic farmer samples from the Anatolian region[52], and 10 Saxon samples from three sites in the UK[127]. DNA was extracted from bone tissue, PCR amplified and then purified using a hybrid capture approach[52,87,127]. The resulting DNA was sequenced on Illumina MiSeq, HiSeq or NextSeq platforms. Sequenced reads were aligned to the human genome using BWA and called SNPs were intersected with the SNPs found on the Human Origins Array[102].

2.4.4 PCA

We ran PCA on the UK Biobank LD dataset using the FastPCA software in EIGENSOFT[43]. We identified several artifactual PCs that were dominated by regions of long-range LD (Figure B.1). Removing loci with significant or suggestive selection signals (Table B.1) along with their flanking 1Mb regions from the LD data set and rerunning PCA eliminated these artifactual PCs (Figure B.2). We refer to the resulting data set with 202,486 SNPs and 113,851 samples as the "PC" dataset. This dataset allows us to better capture axes of variation that correspond to population structure rather than artifacts due to LD.

2.4.5 PC Projection

We projected PoBI[77] (642,288 SNPs, 2,039 samples from 30 populations), POPRES[92] (453,442 SNPs, 4,079 samples from 60 populations) and ancient DNA[52,87] (159,588 SNPs, 52 samples from 4 populations) samples onto the UK Biobank PCs via PC projection[103]. The SNPs in the UK Biobank QC data set were intersected with those in the projected data set and A/T and C/G SNPs were removed due to strand ambiguity (75,254, 37,593 and 24,467 SNPs for PoBI, POPRES and ancient DNA, respectively). The intersected set of SNPs was stringently LD-pruned for $r^2 < 0.05$ using PLINK2[25] (leaving 27,769, 20,914 and 15,722 SNPs respectively). SNP weights were computed for the intersected set of SNPs and these weights were then used to project the new samples onto the UK Biobank PCs[103].

2.4.6 Population cluster analysis

After running PCA, we clustered individuals with k-means clustering using 6 clusters on 5 PCs. We labeled clusters by comparing the centroids of each cluster with the centroids of the projected PoBI populations, as well as by visual inspection. These clusters were then analyzed by running TreeMix[108,107] with default settings, in order to assess the hierarchical population structure between the clusters.

2.4.7 Pairwise discrete subpopulation-based selection statistic

Although we focus primarily on the PCA-based selection statistic from ref.[43] (see below), we also applied the discrete subpopulation-based selection statistic from ref.[12],

which we briefly review. Suppose we are given two populations with genetic distance F_{ST} that are descended from a single ancestral population. The allele frequencies at a particular SNP in these two populations (p_1 and p_2) follows a normal distribution with $p_1, p_2 \sim N\left(p_a, F_{ST} p_a \left(1 - p_a\right)\right)$, where p_a is the ancestral allele frequency of this SNP. As a result, the allele frequency difference ($D = p_1 - p_2$) also follows a normal distribution with mean 0 and variance $2F_{ST} p_a \left(1 - p_a\right)$. Thus, the sample allele frequency difference $\hat{D} = \hat{p}_1 - \hat{p}_2 \sim N(0, p_a \left(1 - p_a\right) \left(2F_{ST} + N_1^{-1} + N_2^{-1}\right)$, where N_1 and N_2 are the number of observed haplotypes from each population. By estimating $\widehat{F_{ST}}$ and $\hat{p}_a = (\hat{p}_1 + \hat{p}_2)/2$, we can assess the statistical significance of unusually large values of $\hat{D}$. We applied this statistic to pairs of population clusters.

2.4.8 PCA-based selection statistic

We applied the PCA-based selection statistic from ref.[43], which we briefly review. PCA is equivalent to the singular value decomposition ($\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$) where $\boldsymbol{X}$ is the normalized genomic matrix, $\boldsymbol{U}$ is the matrix of left singular vectors, $\boldsymbol{V}$ is the matrix of right singular vectors, and $\boldsymbol{\Sigma}$ is a diagonal matrix of singular values. The singular values are related to the eigenvalues of the genetic relationship matrix (GRM) by the relationship $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^2/M$, where M is the number of SNPs used to compute the GRM $\boldsymbol{X}^T\boldsymbol{X}/M$. The matrix $\boldsymbol{U}$ has the properties $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{U} = \boldsymbol{X}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}$. By the central limit theorem, the elements of $\boldsymbol{U}$ follow a normal distribution and after rescaling by M they follow a chi-square (1 d.o.f.) distribution. In other words, the statistic $M(\boldsymbol{X}_i\boldsymbol{V}_k)^2/\Sigma_k^2 = (\boldsymbol{X}_i\boldsymbol{V}_k)/\boldsymbol{\Lambda}_k$ for the i^{th} SNP at the k^{th} PC follows a chi-square (1 d.o.f.) distribution[43]. One benefit of this statistic is that the PCs can be generated on one set of SNPs (here we used the PC

dataset described earlier in order to capture axes of variation related to true population structure) and the selection statistic can be calculated on another set of SNPs (we used the QC dataset in order to maximize the set of SNPs evaluated for signals of selection).

Signals of selection were clustered by considering all SNPs for which the p-value with respect to at least one PC was less than an initial threshold (which we set at 10^{-6}) and clustering together SNPs within 1Mb. SNPs with signals on different PCs but in close proximity were clustered together because loci often have signals of selection on multiple PCs. We defined genome-wide significant loci based on clusters that contained at least one SNP with a p-value smaller than the genome-wide significance threshold. Since we analyzed 5 PCs and 510,665 SNPs, the genome-wide significance threshold was $0.05/\left(5 \times 510,665\right) = 1.96 \times 10^{-8}$. We defined suggestive loci based on clusters with at least two SNPs crossing the initial threshold (but none crossing the genome-wide significance threshold).

2.4.9 Combined selection statistic

We intersected the chi-square (4 d.o.f.) ancient Eurasian selection statistics for 1,004,613 SNPs from Mathieson *et al.*[87] with the PC-based chi-square (1 d.o.f.) UK Biobank selection statistics for 510,665 QC SNPs, producing a list of 115,066 SNPs. For each SNP and each PC, we added the ancient Eurasian selection statistics to the UK Biobank selection statistics for that PC, producing chi-square (5 d.o.f.) statistics which we corrected using genomic control.

2.4.10 Association tests

Association analyses were performed using PLINK2[25] with the top 5 PC as covariates using the "–linear" or "–logistic" flags.

2.5 Results

2.5.1 Population Structure in the UK Biobank

We restricted our analyses of population structure to 113,851 UK Biobank samples of UK ancestry and 202,486 SNPs after quality control (QC) filtering and linkage disequilibrium (LD) pruning (see Methods). We ran principal components analysis (PCA) on this data, using our FastPCA implementation[43]. We determined that the top 5 PCs represent geographic population structure (Figure 2.1) by visually examining plots of the top 10 PCs (Figure B.2), observing that the eigenvalues for the top 5 PCs were above background levels. PC1 through PC4 were also strongly correlated with birth coordinate (Table B.2). The eigenvalue for PC1 was 20.99, which corresponds to the eigenvalue that would be expected at this sample size for two discrete subpopulations of equal size with an F_{ST} of 1.76×10^{-4}(Table B.2).

We ran k-means clustering on these 5 PCs to partition the samples into 6 clusters, since K PCs can differentiate $K + 1$ populations (Figure 2.1, Table 2.1, Figure B.3). To identify the populations underlying the 6 clusters, we projected the PoBI dataset[77], comprising 2,039 samples from 30 regions of the UK, onto the UK Biobank PCs (Figure 2.2, Figure B.4). The individuals in the PoBI study were from rural areas of the UK and had all four grandparents born within 80 km of each other, allowing a glimpse into

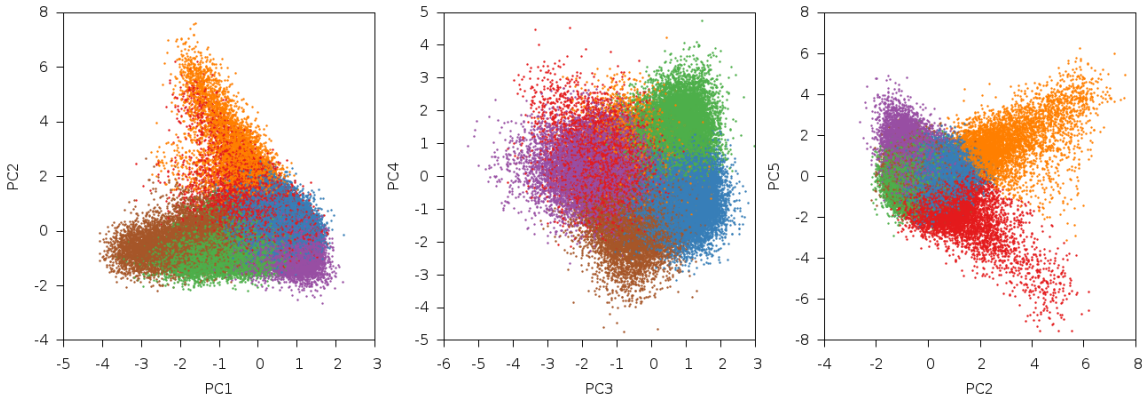

Figure 2.1: Results of PCA with k-means clustering. The top 5 PCs in UK Biobank data are displayed. Samples were clustered using these PCs into 6 clusters with k-means clustering (see Table 2.1). PC5 is plotted against PC2, because PC5 primarily separated the orange and red clusters, which were separated from the other clusters by PC2.

Color	Count	Cluster Name	PoBI Populations
Purple	19,452	Northern England	Yorkshire, Lancashire
Blue	41,494	Southern England	Hampshire, Devon, Norfolk
Brown	12,895	Northern Ireland	Northern Ireland
Green	21,215	Scotland	Argyll and Bute, Banff and Buchan, Orkney
Red	14,190	North Wales	North Wales
Orange	4,605	South Wales / Pembrokeshire	North Pembrokeshire, South Pembrokeshire

Table 2.1: Correspondence between UK Biobank clusters and PoBI populations. We report the PoBI population that most closely corresponds to each UK Biobank cluster (see main text).

the genetics of the UK before the increase in mobility of the 20^{th} century. We selected representative PoBI sample regions that best aligned with the 6 UK Biobank clusters by comparing centroids of each projected population region with those from the UK Biobank clusters via visual inspection (see Methods, Table 2.1). The largest cluster represented southern and eastern England, three clusters represented different regions in the northern UK (northern England, Northern Ireland and Scotland) and two clusters represented north and south Wales. The PCs separated the six UK clusters along two general geographical axes: a north-south axis and a Welsh-specific axis. PC1 and PC3 both separated individuals on north-south axes of variation, with southern England on one end and one of the northern UK clusters on the other. PC2 separated the Welsh clusters from the rest of the UK. PC4 separated the Scotland cluster from the Northern Ireland cluster. PC5 separated the north Wales and south Wales (also known as Pembrokeshire) clusters from each other. To confirm these clusterings, we ran TreeMix[108,107] on our UK Biobank clusters (Figure B.5) as well as the UK Biobank clusters and PoBI populations (Figure B.6), and found that the Celtic subpopulations were grouped separately from the Saxon-related subpopulations; surprisingly, the north and south Wales clusters were separated by TreeMix, which we attribute to the north Wales cluster potentially containing Saxon-related samples (we note the low F_{ST} values between north Wales and southern England; see F_{ST} values in Table B.6). Overall, our results were generally similar to those from the PoBI study[77]; for example, both analyses identified a Welsh axis of differentiation (PC2) and split northern and southern Wales (PC5). We also observed some differences due to the different sampling schemes; in particular, the UK Biobank data set contained many more Irish and Scottish samples (driving variation along PC1) and fewer Orkney samples,

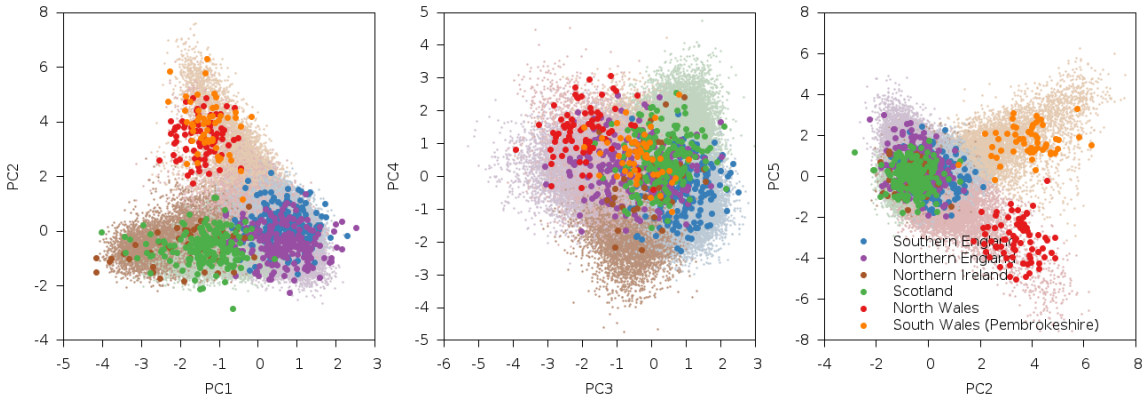

Figure 2.2: Results of PCA with projection of PoBI samples. The top 5 PCs in UK Biobank data are displayed with PoBI samples projected onto these PCs. PoBI populations which visually best matched the clusters from k-means clustering were used to assign names to the six clusters (Table 2.1).

which impacted the clustering of PoBI samples.

We next analyzed UK Biobank population structure in conjunction with ancient DNA samples. Modern European populations are currently thought to have descended from three ancestral populations: Steppe, Mesolithic Europeans and Neolithic farmers[73,52]. We projected ancient samples from these three populations as well as ancient Saxon samples[127] onto the UK Biobank PCs (Figure 2.3, Figure B.7, see Methods). These populations were primarily differentiated along PC1 and PC3, indicating higher levels of Steppe ancestry in northern UK populations. Additionally, the lack of any ancient sample correlation with PC2 suggests that Welsh populations are not differentially admixed with any ancient population in our data set, and likely underwent Welsh-specific genetic drift. We confirmed these findings by projecting pan-European POPRES[92] samples onto the UK Biobank PCs (see Methods, Figure B.8). We note that the Irish and Scottish POPRES populations are projected on top of their corresponding UK Biobank population clusters. Of the continental European populations, Russians (who have the most Steppe ancestry) are projected farther in the Steppe direction along PC1 and PC3 compared with Spanish

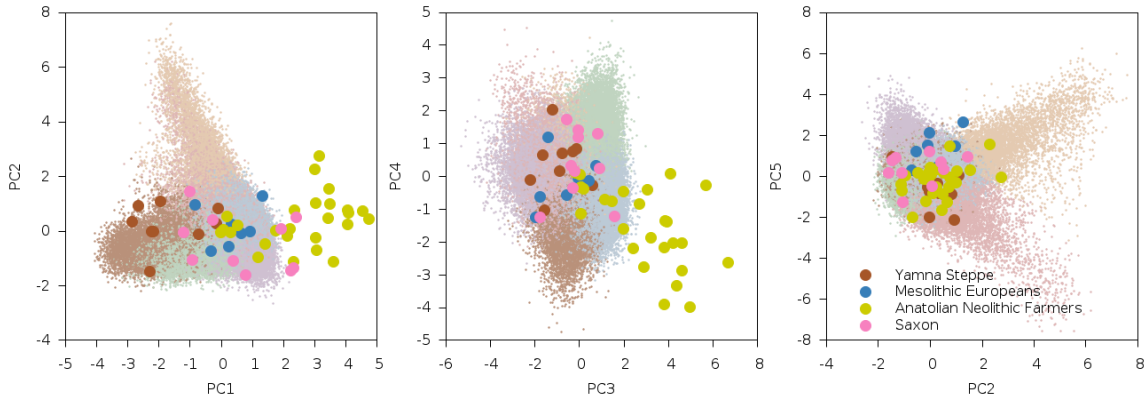

Figure 2.3: Results of PCA with projection of ancient samples. The top 5 PCs in UK Biobank data are displayed with ancient samples projected onto these PCs.

and Italians (who have least Steppe ancestry[52]). Additionally, none of the continental European populations projected onto the same regions as the Welsh on PC2 and PC5. We ran TreeMix on the UK Biobank clusters and ancient populations (Figure B.9) as well as the UK Biobank clusters and POPRES populations (Figure B.10); while population structure inferences using the ancient populations were challenging, the UK Biobank samples were most closely grouped with the Scottish and Irish POPRES samples.

In addition to the impact of ancient Eurasian populations, we know that the genetics of the UK has been strongly impacted by Anglo-Saxon migrations since the Iron Age[127], with the Angles arriving in eastern England and the Saxons in southern England. The Anglo-Saxons interbred with the native Celts, which explains much of the genetic landscape in the UK. We analyzed a variety of samples from predominantly Celtic (Scotland and Wales) and Anglo-Saxon (southern and eastern England) populations from modern Britain in conjunction with the PoBI samples[77] and 10 ancient Saxon samples from eastern England[127] in order to assess the relative amounts of Steppe ancestry. We computed f_4 statistics[102] of the form $f_4\,(Steppe, Neolithic\ Farmer; Pop1, Pop2)$ where *Steppe* and

		Pop2		
Grouping	***Pop1***	Hampshire	Devon	Norfolk
Ancient	Saxon	2.543	3.732	5.118
Scotland	Argyll and Bute	3.323	6.223	9.560
North Wales	North Wales	1.918	5.239	8.490
South Wales	North Pembrokeshire	1.759	4.430	7.124

Table 2.2: Results of f_4 statistics in ancient and modern British samples. We report f_4 statistics of of the form $f_4\,(Steppe,\ Neolithic\,Farmer;\ Pop1,\ Pop2)$, representing a z-score with positive values indicating more Steppe ancestry in *Pop1* than *Pop2* . Samples for *Pop1* were either modern Celtic (Scotland and Wales) or ancient Saxon. Samples for *Pop2* were modern Anglo-Saxon (southern and eastern England).

Neolithic Farmer populations are from ref.[73,52], *Pop1* is either a modern Celtic (Scotland or Wales) or ancient Saxon population and *Pop2* is a modern Anglo-Saxon (southern and eastern England) population (Table 2.2, Table B.3). This statistic is sensitive to Steppe ancestry with positive values indicating more Steppe ancestry in *Pop1* than *Pop2*. We consistently obtained significantly positive f_4 statistics, implying that both the modern Celtic samples and the ancient Saxon samples have more Steppe ancestry than the modern Anglo-Saxon samples from southern and eastern England. This indicates that southern and eastern England is not exclusively a genetic mix of Celts and Saxons, which each have more Steppe ancestry. There are a variety of possible explanations, but one is that the present genetic structure of Britain, while subtle, is quite old, and that southern England in Roman times already had less Steppe ancestry than Wales and Scotland.

2.5.2 Signals of Natural Selection

We searched for signals of selection using a recently developed selection statistic that detects unusual population differentiation along continuous PCs[43]. Notably, this statistic is able to detect selection signals at genome-wide significance. We analyzed the top 5 UK Biobank PCs (which were computed using LD-pruned SNPs), and computed selection

statistics at 510,665 SNPs, reflecting the set of SNPs after QC but before LD-pruning (see Methods). The Manhattan plot for PC1 is reported in Figure 2.4, with additional plots in Figure B.11. We detected genome-wide significant signals of selection at *FUT2* and at several loci with widely known signals of selection (Table 2.3). Loci with suggestive signals of selection ($p < 10^{-6}$) are reported in Table B.4. *FUT2* has also previously been reported as a target of natural selection[37,40]; those results focused on frequency differences between highly diverged continental populations, whereas our results implicate much more recent selection because UK Biobank populations diverged much more recently than continental populations. *FUT2* encodes fucosyltransferase 2, an enzyme that affects the Lewis blood group. The SNP with the most significant p-value, rs601338, is a coding variant where the variant rs601338*G encodes the secretor allele and the rs601338*A variant encodes the nonsecretor allele, which protects against the Norwalk norovirus[140,23]. This SNP also affects the progression of HIV infection[67] , and is associated with vitamin B_{12} levels[59], Crohn's disease[89] , celiac disease and inflammatory bowel disease[101], possibly due to changes in gut microbiome energy metabolism[144]. rs601338*A is more common in northern UK samples (Table B.5). The GERA[6] and PoBI[77] data sets do not include rs601338, but exhibited similar allele frequency patterns at rs492602 and rs676388 (Table B.5), two linked SNPs in *FUT2* whose allele frequencies vary on a north-south axis in UK Biobank data. All three SNPs had genome-wide significant signals of selection in UK Biobank, and rs601338 and rs492602 were also genome-wide significant when analyzing the 6 UK Biobank clusters described above using a test for selection based on unusual differentiation between pairs of discrete subpopulations (Table B.6). On the other hand, rs492602 and rs676388 were only suggestively significant ($p < 1.00 \times 10^{-6}$)

Locus	Chromosome	Position (Mb)	PC	Top SNP	p-value
LCT[11]	2	134.9 - 137.2	1	rs7570971	3.96×10^{-15}
TLR1[61]	4	38.8 - 38.9	1	rs4833095	7.96×10^{-15}
			2		1.27×10^{-8}
			3		7.89×10^{-9}
			4		1.54×10^{-11}
IRF4[22,106]	6	0.4 - 0.5	1	rs62389423	2.31×10^{-43}
HLA[30]	6	31.1 - 32.9	1	rs9366778	8.45×10^{-9}
FUT2	19	49.2 - 49.2	1	rs601338	9.16×10^{-9}

Table 2.3: Top signals of selection for UK Biobank along PC1-PC5. We report the top signal of natural selection for each locus reaching genome-wide significance ($p < 1.96 \times 10^{-8}$) along any of the top five PCs. Neighboring SNPs <1Mb apart with genome-wide significant signals were grouped together into a single locus.

in tests for selection using the GERA data set (Table B.7), emphasizing the advantage of analyzing more closely related subpopulations in very large sample sizes in the UK Biobank data set.

To detect additional signals of selection, we combined our PC-based selection statistics from the UK Biobank data with a previously described selection statistic that detects unusual allele frequency differences after the admixture of ancient Eurasian populations by identifying SNPs whose allele frequencies are inconsistent with admixture proportions inferred from genome-wide data[87]. For each of PC1-PC5 in UK Biobank, we summed our chi-square (1 d.o.f.) selection statistics for that PC with the chi-square (4 d.o.f.) selection statistics from ref. 23 to produce chi-square (5 d.o.f.) statistics that combine these independent signals (see Methods). We confirmed the independence of the two selection statistics by examining the correlations between the two selection statistics and checking that the combined statistics were not substantially inflated, obtaining λ_{GC} values of 1.04-1.06 (Table B.8; see Figure B.12 for P-P plot). In order to produce maximally conservative statistics, we corrected our combined statistics by these λ_{GC} values. We looked for

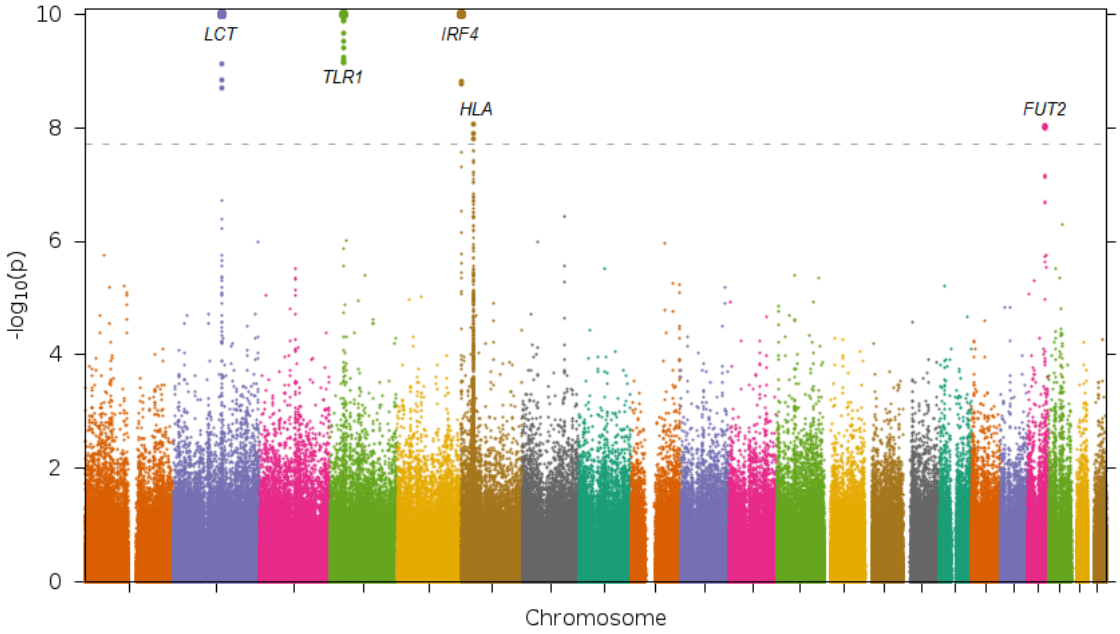

Figure 2.4: Selection statistics for UK Biobank along PC1. A Manhattan plot with $-\log_{10}(p)$ values is displayed. Values above the significance threshold (dotted line, $p = 1.96 \times 10^{-8}, \alpha = 0.05$ after correcting for 5 PCs and 510,665 SNPs) are displayed as larger points and are labeled with the locus they correspond to (see Table 2.3). $-\log_{10}(p)$ values larger than 10 are truncated at 10 for easier visualization and are displayed as even larger points.

signals that were genome-wide significant in the combined selection statistic but not in either of the constituent UK Biobank or ancient Eurasian selection statistics. Results are reported in Table 2.4.

We detected genome-wide significant signals of selection at the *F12* and *CYP1A2 /CSK* loci. We are not currently aware of previous evidence of selection at *F12*. *F12* codes for coagulation factor XII, a protein involved in blood clotting[119]. The SNP at the *F12* locus, rs2545801 was suggestively significant in the ancient Eurasian analysis ($p = 5.35 \times 10^{-8}$), and combining it with the UK Biobank selection statistic on PC2 produced a genome-wide significant signal. This SNP has been associated with activated partial thromboplastin time, a measure of blood clotting speed where shorter time is a risk factor for strokes[135]. An additional significant SNP at *F12*, rs2731672, affects expression

Locus	Chr	Position (Mb)	PC	Top SNP	Combined p-value	UK Biobank p-value	Ancient Eurasian p-value
F12	5	33.9 - 34.0	4	rs2545801	2.05×10^{-10}	3.99×10^{-5}	5.35×10^{-8}
CYP1A2 / CSK	15	75.0 - 75.1	2	rs1378942	4.65×10^{-8}	1.05×10^{-2}	1.08×10^{-7}

Table 2.4: Top signals of selection for combined selection statistics. We report the top selection statistic for each locus reaching genome-wide significance, restricting to loci that were not genome-wide significant in either the UK Biobank selection statistic or the ancient Eurasian selection statistics. Neighboring SNPs $<$1Mb apart with genome-wide significant signals were grouped together into a single locus.

of *F12* in liver[64] and is associated with plasma levels of factor XII[50]. The *CYP1A2/CSK* locus has previously been reported as a target of natural selection when comparing intercontinental allele and haplotype frequencies[150,32], but our results implicate much more recent selection. The two detected SNPs at this locus are in strong LD ($r^2 = 0.858$). The top SNP, rs1378942, is in an intron in the *CSK* gene. This SNP has greatly varying allele frequency across continents[32], is associated with blood pressure[93,134] and systemic sclerosis (an autoimmune disease affecting connective tissue)[86] . The second SNP, rs2472304 in *CYP1A2*, is associated with esophageal cancer[151] , caffeine consumption[29] and may mediate the protective effect of caffeine on Parkinson's disease[109] .

We tested SNPs with genome-wide significant signals of selection in the constituent UK Biobank or ancient Eurasian scans or the combined scan for association with 15 phenotypes in the UK Biobank data set, using the top 5 PCs as covariates (Table B.9, see Methods). The top SNP at *F12* (rs2545801) was associated with height ($p = 4.8 \times 10^{-11}$), and the top SNP at *CYP1A2/CSK* (rs1378942) was associated with diastolic blood pressure (DBP) ($p = 3.6 \times 10^{-19}$) and hypertension ($p = 4.8 \times 10^{-9}$), consistent with previous findings[63]. We detected additional associations with DBP ($p = 8.00 \times 10^{-33}$) and hypertension ($p = 1.30 \times 10^{-1}$) at the *ATXN2/SH2B3* locus which was reported as under selection in the ancient Eurasian scan. The top SNP in *ATXN2/SH2B3*, rs3184504, is known to be associated with blood pressure[138]. We note that PC1 and PC3 were strongly associated with height in the UK Biobank data set, and PC3 and PC4 were associated with DBP (Table B.10). *GRK4*[124], *AGT*[124] and *ATP1A1*[51] have also been reported to be under selection and to be associated with DBP or hypertension. None of the SNPs in *GRK4* or *ATP1A1* were found to be under selection or associated with DBP or hyperten-

sion in our analyses. The *AGT* SNP rs699 was associated with DBP ($p = 7.2 \times 10^{-10}$) and nominally associated to hypertension ($p = 4.8 \times 10^{-4}$), although it did not produce a significant signal of selection in our analyses.

2.6 Discussion

In this study, we used PCA to analyze the population structure of a large UK cohort ($N = 113,851$). We detected 5 PCs representing geographic population structure that partitioned this cohort into six subpopulation clusters. Projecting ancient samples onto these PCs revealed greater Steppe ancestry in northern UK samples. No ancient samples were found to vary along the Welsh-specific axis, suggesting that the Welsh populations differ from the rest of the UK due to drift and not different levels of admixture. We also determined that UK population structure cannot be explained as a simple mixture of Celts and Saxons.

We leveraged the subtle population structure and large sample size of the UK Biobank data set to detect signals of natural selection. We determined that the rs601338*A allele of *FUT2* was more common in northern UK samples, suggesting that pathogens may have exerted selective pressure in those populations. Combining a selection statistic that detects selection via population differentiation within the UK with a separate statistic that detects selection since ancient population admixture in Europe, we were able to detect selection at two additional loci, *F12* and *CYP1A2/CSK*. We additionally found associations to diastolic blood pressure at *CYP1A2/CSK* and at the *ATXN2/SH2B3* locus implicated in a previous selection scan.

We conclude by noting three limitations in our work. First, we employed PCA, a widely

used method for analyzing population structure[43,103,97], but haplotype-based methods such as fineSTRUCTURE may be more powerful[77,72,139]; recent advances in computationally efficient phasing[82,99] increase the prospects for applying such methods to biobank scale data. Second, we employed methods designed to detect selection at individual loci, but did not employ methods to detect polygenic selection[115,114,146,10,120]; our observation that top PCs were correlated with height and DBP in the UK Biobank data set, which could potentially be consistent with the action of polygenic selection on these traits, motivates further analyses of possible polygenic selection. Finally, the PC-based test for selection that we employed assumes that allele frequencies vary linearly along a PC. The spatial ancestry analysis (SPA) method[157,8,7] allows for a logistic relationship between allele frequency and ancestry, and is not constrained by this limitation. However, the advantage of the PC-based test for selection over SPA is that it provides an assessment of statistical significance (p-values), allowing for the detection of genome-wide significant signals, a key consideration in genome scans for selection.

3

Estimating cross-population genetic correlations of causal effect sizes

3.1 Preface

The following work is being prepared with co-authors Po-Ru Loh, Hilary Finucane, Yakir Reshef, Nick J. Patterson and Alkes L. Price.

3.2 Abstract

Recent studies have examined genetic correlations of SNP effect sizes across pairs of populations to better understand the genetic architectures of complex traits. These studies have estimated ρ_g, the cross-population correlation of joint-fit effect sizes at typed SNPs. However, the value of ρ_g depends both on the cross-population correlation of true causal effect sizes (ρ_b) and on the similarity in linkage disequilibrium (LD) patterns in the two populations, which drive tagging effects. Here, we derive the value of the ratio ρ_g/ρ_b as a function of LD in each population. By applying existing methods[74,76,75] to obtain estimates of ρ_g, we can use this ratio to estimate ρ_b. Our estimates of ρ_b were equal to 0.55 (s.e. 0.14) between Europeans and East Asians averaged across 9 traits in the GERA data set, 0.54 (s.e. 0.18) between Europeans and South Asians averaged across 13 traits in the UK Biobank data set, and 0.48 (s.e. 0.06) and 0.65 (s.e. 0.09) between Europeans and East Asians in summary statistic data sets for type 2 diabetes (T2D) and rheumatoid arthritis (RA), respectively. These results implicate substantially different causal genetic architectures across continental populations.

3.3 Introduction

Differing patterns of linkage between SNPs (LD) across populations has been a blessing and a curse in modern genetic studies. Because genome-wide association studies (GWAS) analyze a subset of SNPs in the genome, differential LD between untyped causal SNPs and nearby typed SNPs has led to advances in fine-mapping[53] while also causing issues in transferring the results of GWAS across populations[18,122] . In this paper, we seek to

address how much of the transferability of GWAS is due to different LD patterns versus differences in underlying causal effect sizes.

Previous work[18,31,85] on this topic has extended methods used to estimate the correlation of effect sizes jointly fit at typed SNPs[76,154,20] to cross-population effect size correlations. For a pair of phenotypes, it is intuitive that each SNP may have a similar effect on the phenotypes, and measuring the correlation of SNP effect sizes is a measure of how related these two phenotypes are. By applying the same approach to multiple populations, we get a measure of how similarly typed SNPs affect different populations. These methods provide a measurement that is affected by several factors. Aside from the potential of the causal effect sizes being different across populations (possibly due to gene-by-gene or gene-by-environment interactions), differential LD and allele frequencies also affect correlation measurements. To illustrate the effects of differential LD, consider the situation in which there is an untyped SNP with the same non-zero effect size in a pair of populations with two nearby typed SNPs. If the causal SNP is in LD with one of the SNPs while being independent of the other in the first population, and the situation is reversed in the second population, the resulting measured effect size correlation will be 0 when in fact it should be 1. To illustrate the effects of differential allele frequencies, consider the situation in which a causal SNP is typed in two populations with the same per-allele effect size in both. If the SNP is rare in one population and common in the other, the per-normalized-allele effect size in the population where the SNP is rare will be much lower than the per-normalized-allele effect size in the other population.

Here, we propose a method to remove the effects of differential LD and allele frequencies on cross-population SNP effect correlations. We have developed a theoretical predic-

tion of the ratio of the joint effect size correlation and the causal effect size correlation. Additionally, we only center, but do not scale the genotypes so that we are measuring per-allele rather than per-normalized allele effect sizes, although it has been shown that the per-allele and per-normalized-allele effect size ratios are very similar[18]. We first estimate the per-allele joint fit effect size correlation, and then divide this quantity by the τ-ratio to arrive at an estimate of the per-allele causal effect size correlation.

3.4 Methods

3.4.1 Overview of methods

Our methods focus on relating per-allele genetic effects at all SNPs (b) and the joint genetic effects at typed SNPs (g). We note that g is a population level quantity, but can be thought of as the linear estimate of the effect sizes in the limit the sample size approaching infinity. The genetic effects at typed SNPs are dependent upon the genetic effects at all SNPs through the LD as well as which set of SNPs is typed on a microarray. In relating the per-allele effect sizes, we will consider SNP covariance as a measure of LD (which we call S) and centered genotypes X. We will also consider the parallel derivations for per-normalized-allele genetic effects (β and γ for causal and joint effect sizes), using SNP correlations of LD (which we call Σ) and the normalized genotypes W.

There are two ultimate goals from this work. The first is to test if the correlation of per-allele genetic effects at all SNPs is in fact equal to one ($\rho_b = 1$); the correlation of per-normalized-allele genetic effects (ρ_β) may not equal one simply due to allele frequency differences between a pair of populations. We approach this by measuring the correlation

of genetic effects at typed SNPs (ρ_g) using existing methods[18,154], and then deriving a relation between ρ_g and ρ_b which we use to estimate $\hat{\rho}_b$.

3.4.2 Estimating cross-population correlations of joint-fit effect sizes at typed SNPs

We begin by considering the joint effect sizes at typed SNPs, which is the quantity being estimated by mixed model methods[153]. We used bivariate REML[74,76,75] on the raw genotypes, as implemented in GCTA[154], to estimate $\hat{\rho}_g$ and $\hat{\rho}_\gamma$. REML models SNP effect sizes as random variables through the following equations:$Y_k = X_{k,T}^T g_k + e_{g,k}$In this formulation, Y_k is a mean-centered $N_k \times 1$ vector of phenotypes in population k consisting of N_k samples, $X_{k,T}$ is a mean-centered $M_T \times N_k$ matrix of genotypes at M_T typed SNPs in population k, g_k is a $M_T \times 1$ vector of per-allele joint fit effect sizes, and $e_{g,k}$ and $e_{\gamma,k}$ are $N \times 1$ random error vectors. In this formulation, the effect sizes for two populations are i.i.d. across SNPs follow a normal distribution for SNP i.

$$\begin{pmatrix} g_{1i} \\ g_{2i} \end{pmatrix} \sim N \left[\mathbf{0}, \begin{pmatrix} \sigma_{g_1}^2 & \sigma_{g_1}\sigma_{g_2}\rho_g \\ \sigma_{g_1}\sigma_{g_2}\rho_g & \sigma_{g_2}^2 \end{pmatrix} \right] \tag{3.1}$$

The per-normalized-allele effect sizes (γ_k) follow a similar derivation, except using W_k, the normalized genotypes, in place of X_k. It is clear that the two formulations are not equivalent, because rare alleles do not affect the variance of the phenotypes as much for the per-allele effect size formulation. When computing the genetic relationship matrix (GRM) for REML, it is important to compute it using X or W when attempting to compute the ρ_g or ρ_γ respectively. The GRM can be computed as

$$A_g = \begin{pmatrix} X_{1,T}^T X_{1,T}^T & X_{1,T}^T X_{2,T}^T \\ X_{2,T}^T X_{1,T}^T & X_{2,T}^T X_{2,T}^T \end{pmatrix} \text{or} A_\gamma = \begin{pmatrix} W_{1,T}^T W_{1,T}^T & W_{1,T}^T W_{2,T}^T \\ W_{2,T}^T W_{1,T}^T & W_{2,T}^T W_{2,T}^T \end{pmatrix} \tag{3.2}$$

An alternate approach is Popcorn[18], a maximum-likelihood based method can compute per-allele and per-normalized-allele (genetic effect and genetic impact) correlations using reference LD panels and summary statistics. We adjusted previously-published Popcorn estimates of cross-population heritability between European and East Asian type-2 diabetes and rheumatoid arthritis[18].

3.4.3 Relationship between cross-population genetic correlations of joint-fit effect sizes and causal effect sizes

As noted above, the joint-fit effect sizes at typed SNPs can be thought of as linear estimates of the effect sizes in the limit of the sample size approaching infinity. We derive the relationship for per-allele affect sizes, although we note that the math holds for per-normalized-allele effect sizes as well.

$$\begin{aligned} g_k = \lim_{N_k \to \infty} \hat{g}_k &= \lim_{N_k \to \infty} \left(X_{k,T} X_{k,T}^T \right) X_{k,T} Y_k \\ &= \lim_{N_k \to \infty} \left(X_{k,T} X_{k,T}^T \right) X_{k,T} \left[X_k^T b_k + e_{b,k} \right] \\ &= \lim_{N_k \to \infty} \left(X_{k,T} X_{k,T}^T \right) X_{k,T} X_{k,A}^T b_k \\ &= \left[S^{(k)} \right]_{TT}^{-1} S_{TA}^{(k)} b_k \end{aligned} \tag{3.3}$$

$$S_{TT}^{(k)} g_k = S_{TA}^{(k)} b_k$$

$X_{k,A}$ is the $M \times N_k$ matrix of all mean-centered SNPs in population k, while $S_{TT}^{(k)}$ and $S_{TA}^{(k)}$ are $M_T \times M_T$ and $M_T \times M$ covariance matrices between just typed SNPs and between typed and all SNPs, respectively, in population k. The next step is to note how the right hand side of this equation relates to ρ_b:

$$\begin{aligned} E\left[\left(S_{TA}^{(1)}b_1\right)^T\left(S_{TA}^{(2)}b_2\right)\right] &= E\left[tr\left(b_1^T S_{AT}^{(1)} S_{TA}^{(2)} b_2\right)\right] \\ &= tr\left(S_{TA}^{(1)} E\left[b_1^T b_2\right] S_{AT}^{(2)}\right) \\ &= \rho_b \sigma_{b_1}\sigma_{b_2} tr\left(S_{TA}^{(1)} S_{AT}^{(2)}\right) \\ \frac{E\left[b_1^T S_{AT}^{(1)} S_{TA}^{(2)} b_2\right]}{\sqrt{E\left[b_1^T S_{AT}^{(1)} S_{TA}^{(1)} b_1\right] E\left[b_2^T S_{AT}^{(2)} S_{TA}^{(2)} b_2\right]}} &= \rho_b \frac{tr\left(S_{TA}^{(1)} S_{AT}^{(2)}\right)}{\sqrt{tr\left(S_{TA}^{(1)} S_{AT}^{(1)}\right) tr\left(S_{TA}^{(2)} S_{AT}^{(2)}\right)}} \\ &= \rho_b \tau\left(S_{TA}^{(1)}, S_{TA}^{(2)}\right) \end{aligned} \tag{3.4}$$

The key to this formula is noting that pairs of causal effect sizes ((b_{1i}, b_{2i})) are iid, which is what allows us to convert the expectation into a scalar. However, this same assumption is made when estimating pairs of joint effect sizes ((g_{1i}, g_{2i})). Thus, we show:

$$\begin{aligned} \rho_g \tau\left(S_{TT}^{(1)}, S_{TT}^{(2)}\right) &= \rho_b \tau\left(S_{TA}^{(1)}, S_{TA}^{(2)}\right) \\ \frac{\rho_g}{\rho_b} &= \frac{\tau\left(S_{TT}^{(1)}, S_{TT}^{(2)}\right)}{\tau\left(S_{TA}^{(1)}, S_{TT}^{(2)}\right)} \\ \frac{\rho_\gamma}{\rho_\beta} &= \frac{\tau\left(\Sigma_{TT}^{(1)}, \Sigma_{TT}^{(2)}\right)}{\tau\left(\Sigma_{TA}^{(1)}, \Sigma_{TT}^{(2)}\right)} \end{aligned} \tag{3.5}$$

Here, Σ refers to the SNP correlation matrices. The ratio of τ functions that corresponds to $\frac{\rho_g}{\rho_b}$ or $\frac{\rho_\gamma}{\rho_\beta}$ will be referred to as the τ-ratio throughout this paper. We note that the trace of the product of the LD matrices corresponds to the sum of the entries of the

Hadamard product of the two matrices.

$$\begin{aligned} tr\left(\left[S^{(1)}\right]^T S^{(2)}\right) &= \sum_{i,j} S_{ij}^{(1)} S_{ij}^{(2)} \\ tr\left(\left[S^{(k)}\right]^T S^{(k)}\right) &= \sum_{i,j} \left[S_{ij}^{(k)}\right]^2 \end{aligned} \tag{3.6}$$

The denominator of the τ function contains the sums of LD scores, while the numerator contains the sum of a cross-population analog of LD scores. Since squared correlations are upwardly biased[21], we can remove this bias by adjusting our squared correlation estimates as in LD score regression[21] and propagating that adjustment to squared covariance estimates.

$$\begin{aligned} \widetilde{\Sigma^2}_{ij}^{(k)} &= \left[\widehat{\Sigma}_{ij}^{(k)}\right]^2 - \frac{1 - \left[\widehat{\Sigma}_{ij}^{(k)}\right]^2}{N_k - 2} \\ \widetilde{S^2}_{ij}^{(k)} &= \widehat{S}_{ii}^{(k)} \widehat{S}_{jj}^{(k)} \widetilde{\Sigma^2}_{ij}^{(k)} \\ \hat{\tau}_\Sigma\left(\widehat{\Sigma}^{(1)}, \widehat{\Sigma}^{(2)}\right) &= \frac{\sum_{i,j} \widehat{\Sigma}_{ij}^{(1)} \widehat{\Sigma}_{ij}^{(2)}}{\sqrt{\left(\sum_{i,j} \widetilde{\Sigma^2}_{ij}^{(1)}\right)\left(\sum_{i,j} \widetilde{\Sigma^2}_{ij}^{(2)}\right)}} \\ \hat{\tau}_S\left(\widehat{S}^{(1)}, \widehat{S}^{(2)}\right) &= \frac{\sum_{i,j} \widehat{S}_{ij}^{(1)} \widehat{S}_{ij}^{(2)}}{\sqrt{\left(\sum_{i,j} \widetilde{S^2}_{ij}^{(1)}\right)\left(\sum_{i,j} \widetilde{S^2}_{ij}^{(2)}\right)}} \end{aligned} \tag{3.7}$$

We also performed windowing on our LD calculations by setting $\widehat{S}_{ij}^{(k)}$ and $\widehat{\Sigma}_{ij}^{(k)}$ to 0 if the distance between SNPs i and j was greater than 1MB as in LD score regression[21].

3.4.4 Simulations with real genotypes and simulated phenotypes

For realistic simulations, we sampled real genotypes from GERA (see below), generated phenotypes and set aside a set of typed SNPs. We sampled N_1 GERA-EUR and N_2 GERA-EAS samples. We sampled M_T typed SNPs from the available SNPs on one chromosome, and M_C causal SNPs again sampled from either the set of all SNPs or the set of untyped SNPs. We simulated M_C pairs of causal effect sizes by sampling them from a $N(0, \mathrm{P}_b)$ distribution, where P_b is the 2×2 equicorrelation matrix with ρ_b, the causal effect correlation, on the off-diagonal. We multiplied the simulated genotypes by the vectors of population effect sizes within their corresponding population to create phenotypes. The resulting phenotypes were scaled to have mean 0 and variance h^2. We added random error sampled from $N\left(0, 1-h^2\right)$ to the phenotypes to create the final phenotypes.

3.4.5 1000 Genomes dataset

The 1000 Genomes[137] dataset contains 503 individuals of European ancestry (EUR), 504 individuals of East Asian ancestry (EAS) and 489 individuals of South Asian ancestry (SAS). We performed QC in each population separately, retaining only bi-allelic SNPs in Hardy-Weinberg equilibrium (p>0.001) with MAF>0.1% and excluded SNPs with duplicate IDs, leaving 13,258,254 EUR SNPs, 12,285,372 EAS SNPs and 24,463,301 SAS SNPs. We then performed additional MAF thresholding (1% for the main analysis) and retained SNPs that were above the MAF threshold in each pair of populations as was done in other work[18,31,85]. This left 1,352,543 EUR-EAS SNPS and 2,115,911 EUR-SAS SNPs.

3.4.6 GERA data set

The Genetic Epidemiology Research on Adult Health and Aging (GERA)[6] data set contains 62,318 individuals of European ancestry (GERA-EUR) and 5,188 individuals of East Asian ancestry (GERA-EAS) genotyped on population-specific microarrays containing 657,184 and 694,877 SNPs, respectively. We performed QC in each population separately, retaining only bi-allelic SNPs with MAF greater than 1% (as was done in other work[18,31,85]) and missing genotype rate less than 2%. Only SNPs that passed filters for both populations were retained for simulations, leaving 351,421 SNPs. This set was additionally intersected with the 1000 Genomes SNPs, leaving 315,434 SNPs. Related individuals and individuals with a greater than 2% missing data rate were also excluded, leaving 45,725 GERA-EUR and 3,357 GERA-EAS samples.

3.4.7 UK Biobank data set

The UK Biobank[131] data set contains 120,286 individuals of British ancestry (UKB-EUR) and 1,784 individuals of South Asian ancestry (UKB-SAS) genotyped at 847,131 SNPs. We performed QC as with the GERA data set, retaining only SNPs common to both populations and 1000 Genomes, leaving 392,598 SNPs, 116,478 UKB-EUR samples and 1,706 UKB-SAS samples.

3.4.8 RA and T2D summary statistics

Previous work[18] has computed ρ_g on 2,539,629 typed or imputed SNPs in rheumatoid arthritis (RA) and 1,054,079 typed or imputed SNPs for type-2 diabetes (T2D). These

estimates of ρ_g were derived from summary statistics that were computed from 58,284 individuals of European descent and 22,515 individuals of East Asian descent for RA and 69,033 individuals of European descent and 18,817 individuals of East Asian descent for T2D.

3.5 Results

3.5.1 Simulations

To evaluate the performance of the ρ_g/ρ_b ratio in a situation where ρ_b is known, we sampled real genotypes from GERA chromosome 11 and simulated phenotypes (see Methods). We then computed $\hat{\rho}_g$ using REML, and compared $\hat{\rho}_g/\rho_b$ to our theoretical ρ_g/ρ_b ratio. We first noted an issue where the theoretical ρ_g/ρ_b was lower than $\hat{\rho}_g/\rho_b$. We noted that truncating LD beyond 1MB (see Methods) did corrected this effect, leading us to the conclusion that the theoretical ρ_g/ρ_b ratio is sensitive to noise in long-range estimates of SNP correlation (Figure C.1). This effect was not as prevalent when examining the τ-ratio in 1000 Genomes SNPs (Figure C.2), though adjusting for inflation in $\widehat{r^2}$ was crucial. After removing LD beyond 1MB, the τ-ratio converted estimates of $\hat{\rho}_g$ to $\hat{\rho}_b$ which were accurate under a range of values for ρ_b (Figure 3.1). The key finding from our simulations was that estimates of ρ_g were not accurate when heritability was small (Figure 3.2). Otherwise, our correction factor worked for different numbers of typed SNPs, causal SNPs, samples in each population and an imbalanced number of samples (Figure C.3).

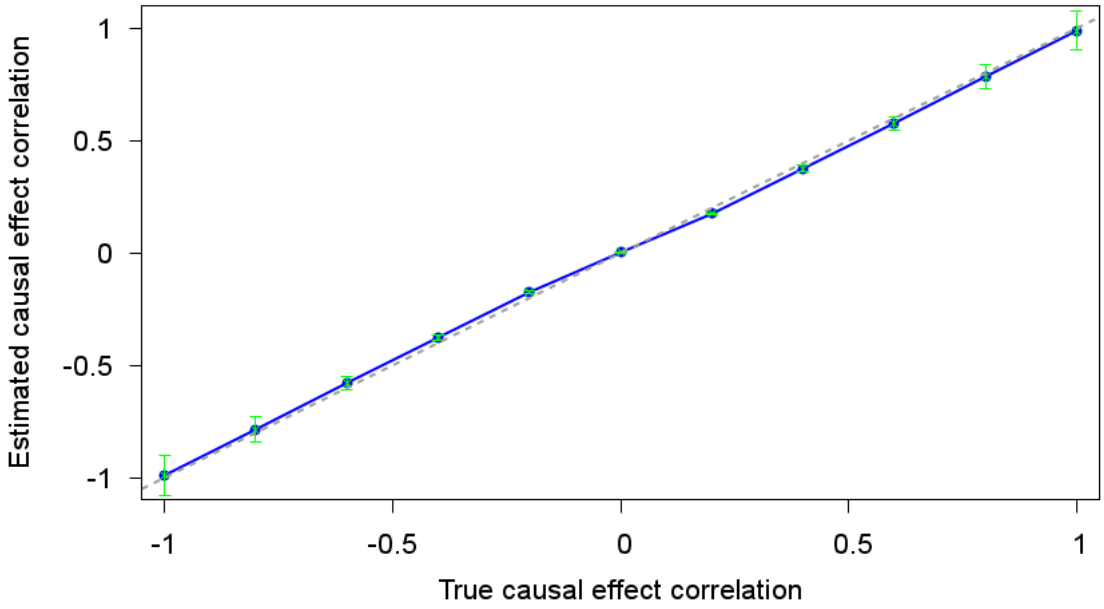

Figure 3.1: Estimates of cross-population heritability in chromosome 11 simulations are accurate. We performed simulations with 2k European and 2k East Asian samples on chromosome 11 with 5k typed SNPs and 100 causal SNPs (see Methods). ρ_b was varied and h^2 was fixed at 0.8. After estimating ρ_g with GCTA [154], we applied our correction factor to estimate ρ_b. We find that this method accurately estimated ρ_b.

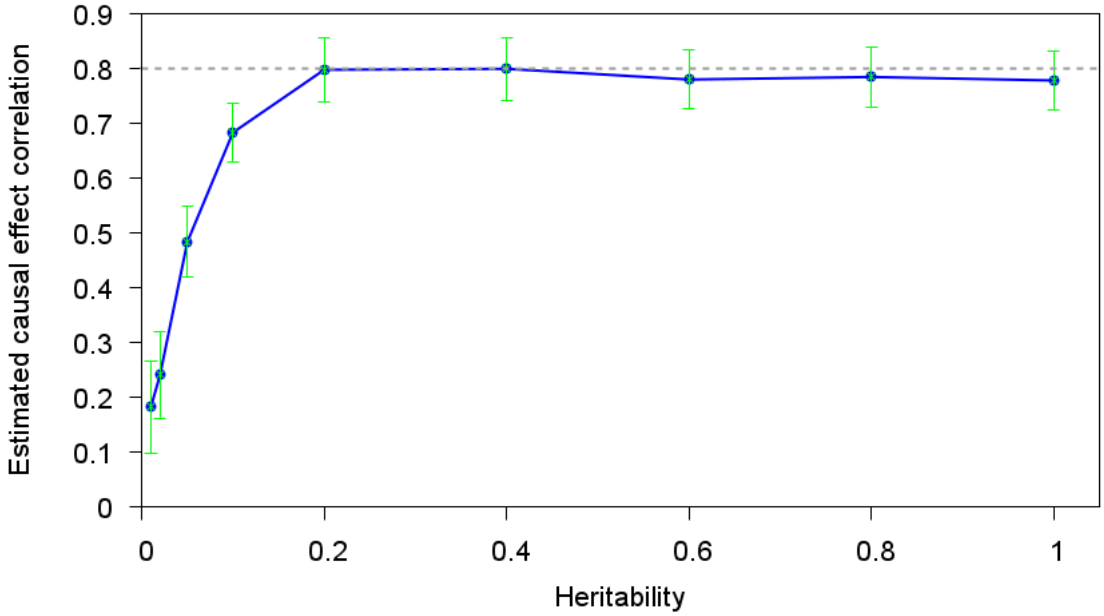

Figure 3.2: Estimates of cross-population heritability are inacurate when heritability is low. We performed simulations with 2k European and 2k East Asian samples on chromosome 11 with 5k typed SNPs and 100 causal SNPs (see Methods). ρ_b was fixed at 0.8 and h^2 was varied. After estimating ρ_g with GCTA [154], we applied our correction factor to estimate ρ_b. We find that for $h^2 < 0.2$, the estimate of ρ_b is innacurate.

Phenotype	$\hat{\rho}_g$		$\hat{\rho}_\gamma$		$\hat{\rho}_b$		$\hat{\rho}_\beta$	
Allergic rhinitis	1.00	(1.06)	1.00	(1.00)	1.08	(1.14)	1.08	(1.08)
Asthma	-0.04	(0.40)	-0.34	(0.33)	-0.04	(0.43)	-0.37	(0.35)
Cardiovascular Disease	0.48	(0.35)	0.30	(0.32)	0.52	(0.38)	0.32	(0.34)
Type 2 Diabetes	0.35	(0.27)	0.29	(0.27)	0.38	(0.29)	0.31	(0.29)
Dyslipidemia	0.52	(0.21)	0.47	(0.21)	0.56	(0.23)	0.51	(0.23)
Hypertension	0.27	(0.19)	0.24	(0.19)	0.29	(0.20)	0.26	(0.20)
Macular Degeneration	1.00	(2.09)	1.00	(2.08)	1.08	(2.25)	1.08	(2.24)
Osteoarthritis	0.53	(0.35)	0.42	(0.32)	0.57	(0.38)	0.45	(0.34)
Osteoporosis	-0.07	(0.47)	-0.12	(0.53)	-0.08	(0.51)	-0.13	(0.57)

Table 3.1: Cross-population heritability for GERA phenotypes. We estimated $\hat{\rho}_g$, $\hat{\rho}_\gamma$, $\hat{\rho}_b$ and $\hat{\rho}_\beta$ for nine phenotypes in GERA. Reported are the estimates and standard errors. The inverse-variance weighted estimates for average $\hat{\rho}_g$, $\hat{\rho}_\gamma$, $\hat{\rho}_b$, and $\hat{\rho}_\beta$ are 0.51 (0.13), 0.41 (0.13), 0.55 (0.14) and 0.44 (0.14).

3.5.2 Application to 9 traits from GERA data set

We applied our method for estimating ρ_b to 9 traits from the GERA data set (Table 3.1), analyzing data from Europeans (EUR) and East Asians (EAS). We first computed the τ-ratio for GERA target SNPs relative to 1000 Genomes reference SNP and found it to be 0. Then we calculated the GRM using all GERA-EAS samples and 10k GERA-EUR samples. We then ran bivariate REML using GCTA for 9 GERA phenotypes. We report the resulting estimates of $\hat{\rho}_g$, $\hat{\rho}_\gamma$, $\hat{\rho}_b$ and $\hat{\rho}_\beta$ (Table 3.1). The inverse-variance weighted average of $\hat{\rho}_g$ is 0.51 with standard error 0.13, translating to an estimate of $\hat{\rho}_b$ of 0.55 and standard error 0.14. The estimates of $\hat{\rho}_\gamma$ and $\hat{\rho}_\beta$ were lower, with estimates of 0.41 (s.e. 0.13) and 0.44 (s.e. 0.14) respectively.

3.5.3 Application to 13 traits from UK Biobank data set

We next applied our method for estimating ρ_b to 13 traits from the UK Biobank data set (Table 3.2), analyzing data from Europeans (EUR) and South Asians (SAS). We

Phenotype	$\hat{\rho}_g$		$\hat{\rho}_\gamma$		$\hat{\rho}_b$		$\hat{\rho}_\beta$	
Bone-densitometry of heel	0.60	(0.18)	0.49	(0.19)	0.62	(0.18)	0.50	(0.20)
Height	0.77	(0.26)	0.63	(0.24)	0.78	(0.26)	0.64	(0.24)
Weight-height ratio	1.00	(2.19)	1.00	(2.64)	1.02	(2.24)	1.02	(2.69)
Diastolic blood pressure	1.00	(0.56)	0.73	(0.35)	1.02	(0.57)	0.74	(0.36)
Systolic blood pressure	1.00	(0.91)	0.76	(0.59)	1.02	(0.93)	0.77	(0.60)
College education	0.36	(0.22)	0.38	(0.22)	0.37	(0.23)	0.39	(0.22)
Smoking status	0.37	(0.39)	0.22	(0.37)	0.38	(0.40)	0.22	(0.38)
Eczema	-0.19	(0.62)	0.14	(0.59)	-0.19	(0.63)	0.15	(0.60)
Asthma	0.92	(1.49)	0.65	(0.72)	0.94	(1.52)	0.66	(0.73)
Hypertension	0.32	(0.32)	0.38	(0.32)	0.32	(0.33)	0.38	(0.33)
FEV1	0.57	(0.27)	0.50	(0.29)	0.58	(0.28)	0.51	(0.30)
FEV1-FCV ratio	0.39	(0.29)	0.58	(0.41)	0.40	(0.30)	0.59	(0.42)
Age at menarche	0.70	(1.07)	0.59	(1.00)	0.71	(1.09)	0.60	(1.02)

Table 3.2: Cross-population heritability for UK Biobank phenotypes. We estimated $\hat{\rho}_g$, $\hat{\rho}_\gamma$, $\hat{\rho}_b$ and $\hat{\rho}_\beta$ for thirteen phenotypes in the UK Biobank. Reported are the estimates and standard errors. The inverse-variance weighted estimates for average $\hat{\rho}_g$, $\hat{\rho}_\gamma$, $\hat{\rho}_b$, and $\hat{\rho}_\beta$ are 0.53 (0.17), 0.50 (0.17), 0.54 (0.18) and 0.51 (0.17).

first computed the τ-ratio for UK Biobank target SNPs relative to the 1000 Genomes reference SNPs and found it to be 0.98. This is larger than the τ-ratio between Europeans and East Asians, despite the similar number of target SNPs because Europeans and South Asians are more closely related than Europeans and East Asians[133]. We then performed REML on 13 UK Biobank phenotypes at all UKB-SAS samples and 10k UKB-EUR samples. We report estimates of $\hat{\rho}_g$, $\hat{\rho}_\gamma$, $\hat{\rho}_b$ and $\hat{\rho}_\beta$ (Table 3.2). In inverse-variance weighted averages of these correlations is 0.53 (s.e. 0.17), 0.50 (s.e. 0.17), 0.54 (s.e. 0.18) and 0.51 (s.e. 0.17), respectively.

3.5.4 Application to RA and T2D summary statistics

We next applied our method for estimating ρ_b to RA and T2D summary statistics in EUR and EAS previously analyzed by Brown *et al.*[18], who reported estimates of ρ_g of 0.463 (s.e. 0.058) and 0.621 (s.e. 0.088) for RA and T2D, respectively. The τ-ratios for the RA target SNPs and T2D target SNPs relative to 1000 Genomes reference SNPs were 0.96 and 0.97, respectively. This is substantially larger than the τ-ratio for the GERA with the same population due to the summary statistics being computed on many more SNPs. The RA and T2D datasets contained 2,539,629 and 1,054,079 typed or imputed SNPs (respectively) compared to the 315,434 typed SNPs found in GERA. The resulting estimates of $\hat{\rho}_b$ are 0.48 (s.e. 0.06) and 0.65 (s.e. 0.09), which are still significantly less than 1.

3.6 Discussion

We have presented a τ-ratio which performs well in simulations and have used it to measure $\hat{\rho}_b$ for multiple phenotypes in GERA, UK Biobank and from summary statistics. We have found that the mean estimates of $\hat{\rho}_b$ in were 0.55 (s.e. 0.14) in GERA and 0.54 (s.e. 0.18) in UK Biobank, which are significantly different from 0 and 1. We have additionally computed theoretical ρ_g/ρ_b ratios which can assist in future studies.

These findings suggest that even after correcting for the effects of differential LD and allele frequency, that the underlying causal effect sizes are not uniform across populations. Our estimates of ρ_g/ρ_b for GERA and UK Biobank were close to one, indicating that existing techniques can recover most of the causal effect size correlations by examining just

typed SNPs.

Our work is limited in a few ways. To begin, LD-based methods do not work well in admixed populations[21], and thus the current approach is not applicable to cross-population genetic correlation estimates in admixed populations[31,85]. Next, if estimates of $\hat{\rho}_g$ and $\hat{\rho}_\gamma$ are noisy, the resulting $\hat{\rho}_b$ and $\hat{\rho}_\beta$ estimates will also be noisy. Additionally, we have not derived the standard error of the τ-ratio metric, which we will compute in future work.

4

Conclusion

We have presented three methodologies that make use of differences between populations to study human genetics. With our FastPCA algorithm, we rapidly detected fine-scale population structure in extremely large cohorts, accurately scaling to over a hundred thousand individuals with modest computational requirements. Using this tool, we detected two novel axes of variation in individuals of European ancestry and detected very subtle population structure in individuals of British descent.

Next, we introduced a natural selection statistic which operates on principal components. Unlike similar tests of natural selection, this statistic only needs one cohort to

operate and it can produce p-values which allow us to detect signals of selection that have genome-wide significance. We have run this tool on the GERA and UK Biobank datasets and detected novel signals of selection at the alcohol dehydrogenase and fucosyltransferase genes. We have also combined this signal of selection with an external one to detect signals of selection at genes influencing blood pressure.

Lastly, we estimated the cross-population correlation of genetic effects at causal SNPs. Using LD from sequenced reference samples, we derived the ratio of cross-population correlation at joint effects at typed SNPs to the cross-population at causal SNPS, and applied this ratio to the estimated joint effect correlations. After applying this methodology to European-East Asian and European-South Asian correlations, we found that aggregate cross-population correlation at typed SNPs was less than one, indicating that gene-by-gene or gene-by-environment correlations may be causing SNP effects to not be consistent across populations. This result has significance for the transferability of the results of genome-wide association studies

A

Supplementary Materials for Chapter 1

A.1 Supplementary Figures

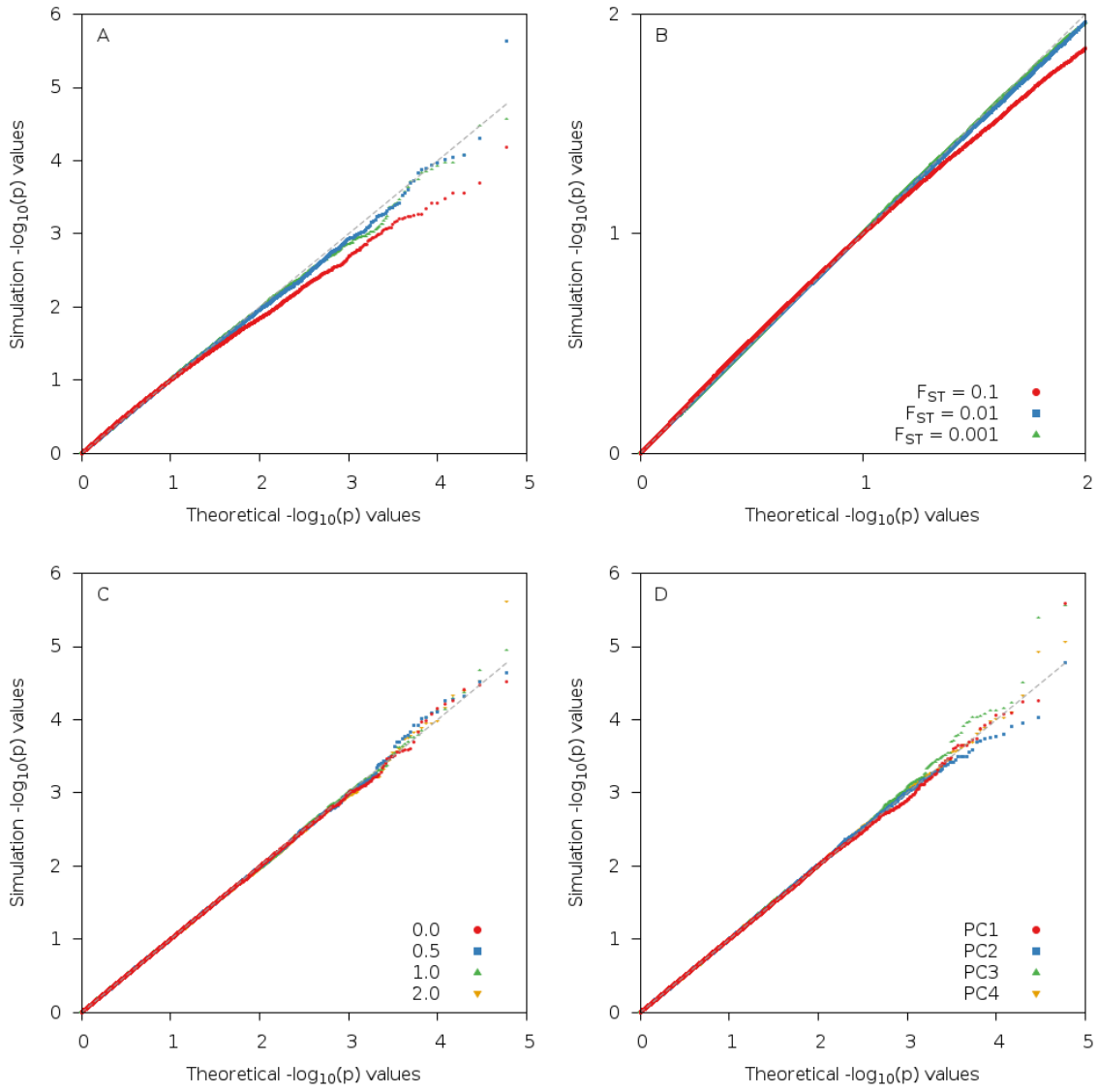

Figure A.1: QQ-plot of the selection statistic in null simulations. The selection statistic was generated for the first PC in null simulations containing 2 populations and differing by $F_{ST} = 0.001, 0.01$ and 0.1. (a) Examining all the p-values, the selection statistic was well calibrated for $F_{ST} = 0.001$ and 0.01, with deflation in the tails for $F_{ST} = 0.1$. (b) Looking only at p-values greater than 0.01, the selection statistic was well calibrated for $F_{ST} = 0.001$ and 0.01, but slightly inflated for p-values greater than 0.1 for $F_{ST} = 0.1$. This explains the results in Table A.2 and Table A.3. (c) In the case with 2 populations differing by $F_{ST} = 0.001$, admixed individuals were generated with admixture proportion drawn from a $Beta\,(a, a)$ distribution, where increasing a means more admixture. (d) Five subpopulations were generated from a phylogenetic structure (see Methods), where the F_{ST} between populations 3, 4 and 5 was 0.001 and the F_{ST} between any other pair of populations was 0.01. In this case with five subpopulations, four principal components are sufficient to describe the population structure. For both examples with more complicated population structure, (c) and (d), the selection statistic remains well calibrated.

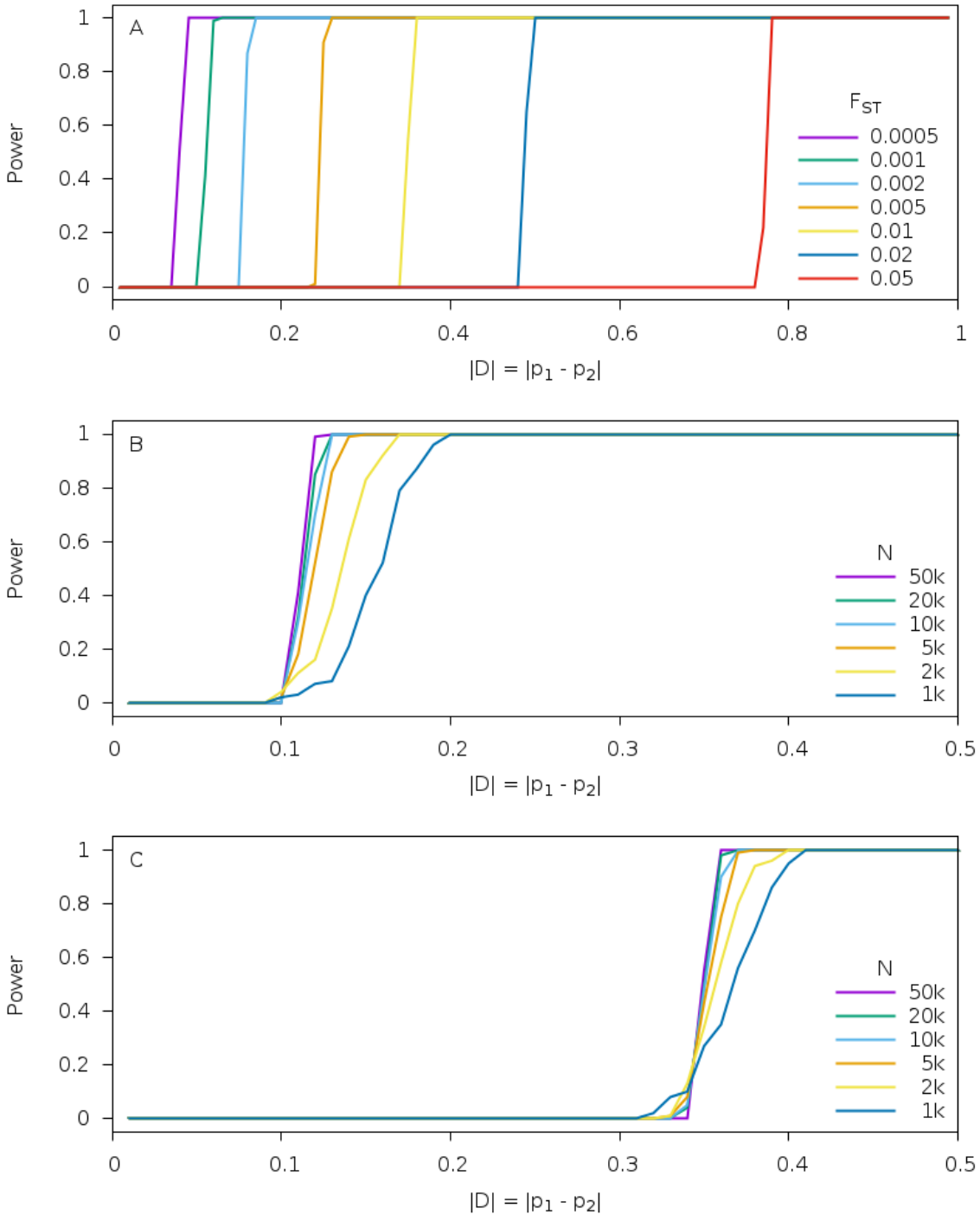

Figure A.2: Power of the discrete-population selection statistic. We ran the discrete-population selection statistic on the same simulations as in Figure 1.3 and found that the discrete-population and the PC-based selection statistics performed nearly identically in these regimes.

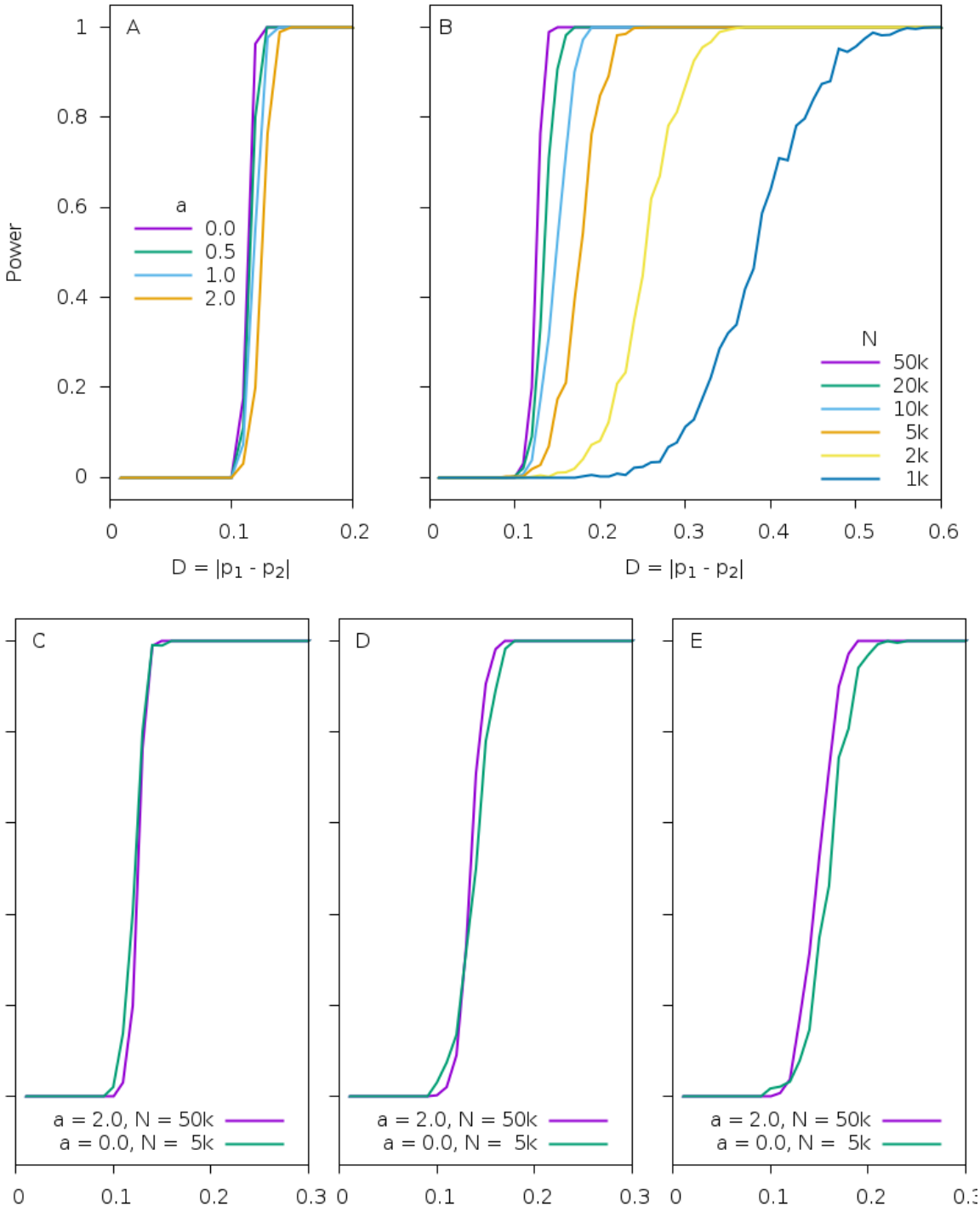

Figure A.3: Power of the PC-based selection statistic in the presence of admixture. Admixture or clinal variation in allele frequencies was simulated by samplings ancestry fraction between two ancestral populations from a $Beta\,(a, a)$ distribution. The two populations were differentiated by $F_{ST} = 0.001$. (a) Increasing a has a similar effect to reducing sample size (Figure 1.3). (b) Varying the number of samples when $a = 2.0$ had a dramatic effect, indicating that sample size is quite important in real data which will have small F_{ST} and non-discrete populations. (c-e) Setting $a = 2$ is roughly the same as having 10% of the data in a dataset with discrete populations.

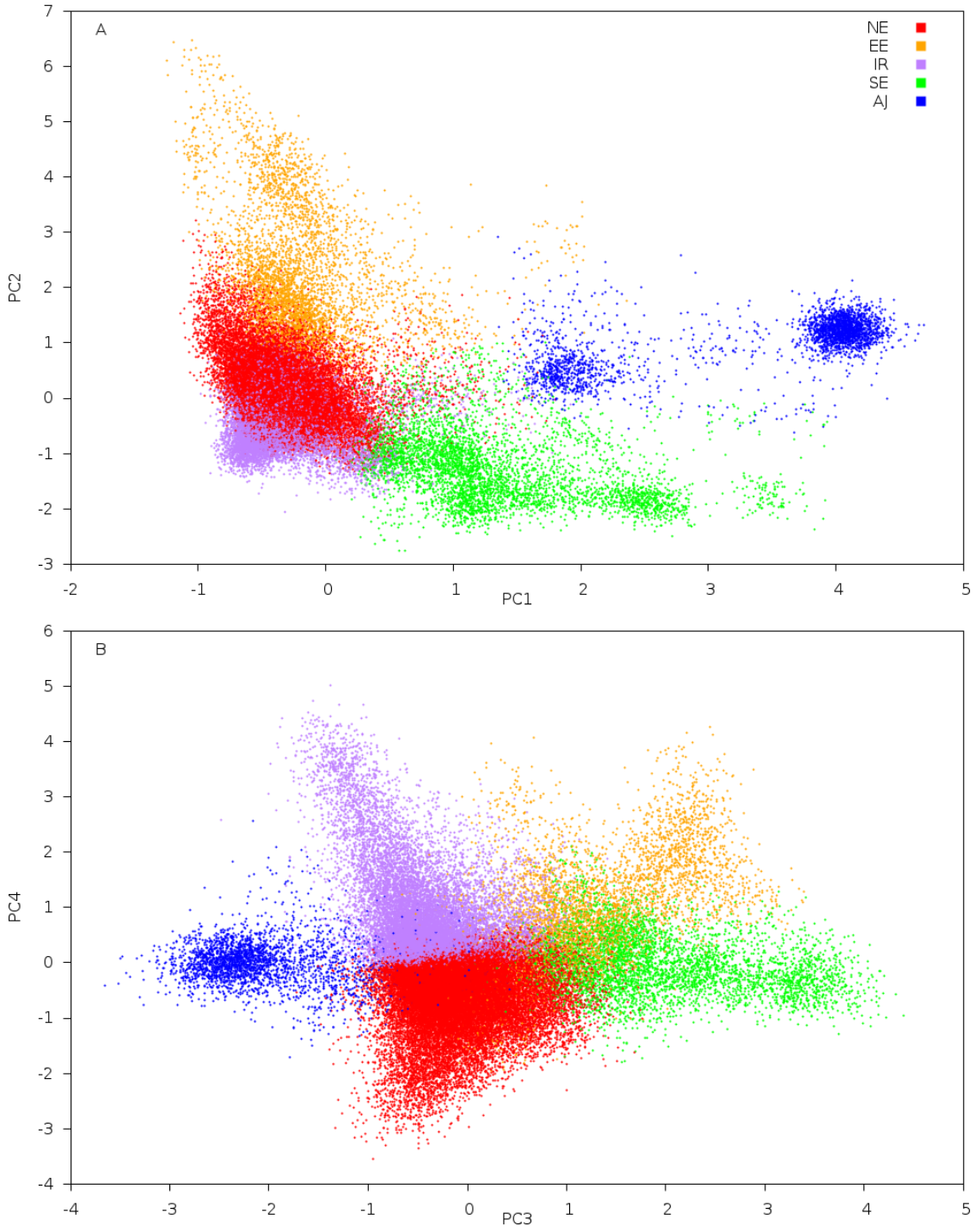

Figure A.4: k-Means clustering confirms visually-observed subpopulations. Individuals were clustered using k-means clustering with $k = 5$ on the top 4 PCs. 5 clusters were the minimum number of clusters that produced results consistent between runs. Clusters were labeled and assigned colors based upon where they fell relative to predicted fractional ancestry and where projected populations lay.

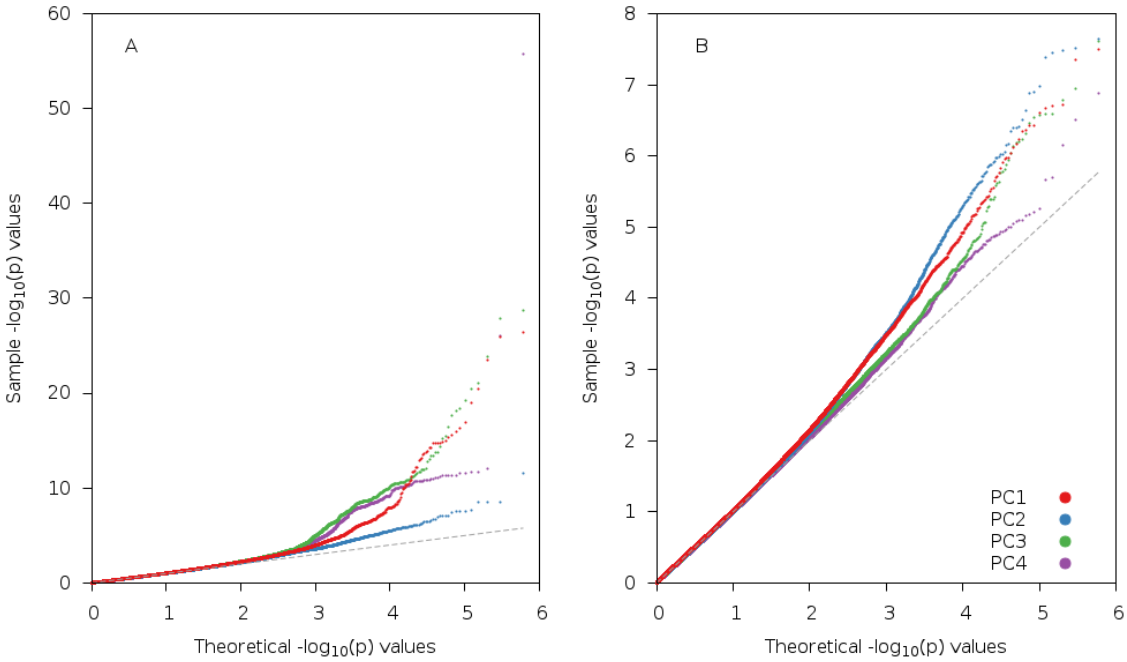

Figure A.5: QQ-plot of the selection statistic for PCs 1-4 in GERA data. QQ-plots of actual vs. theoretical p-values are provided for (A) selection statistics for 608,981 SNPs in the GERA sample that passed the first stage of QC, and (B) selection statistics for 599,992 SNPs excluding the genome-wide significant loci listed in Table 1.1. Despite clear evidence of signal at the extreme tails, the overall distribution of test statistic was not inflated in the original set of SNPs ($0.96 \leq \lambda_{GC} \leq 1.06$) nor in the filtered set ($0.94 \leq \lambda_{GC} \leq 1.05$).

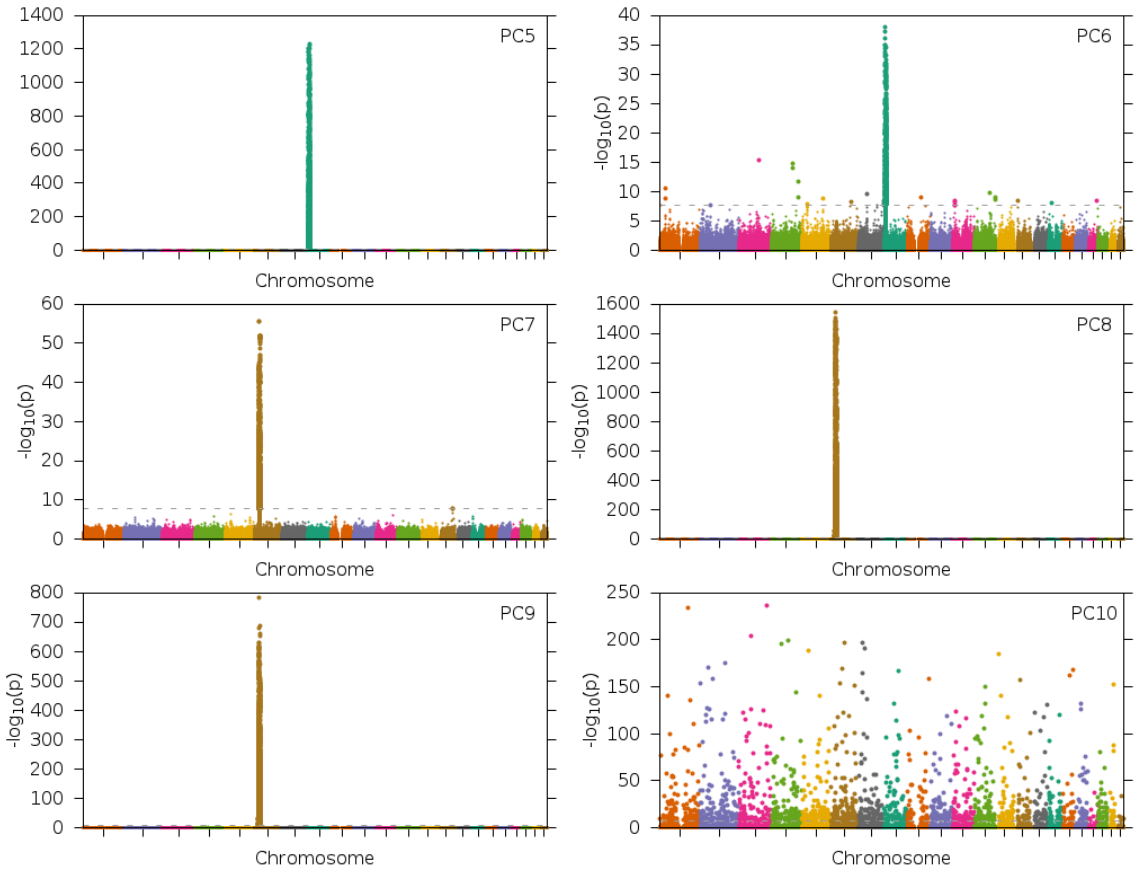

Figure A.6: Selection statistics for PCs 5-10 in GERA data. The selection statistics for PCs 5-10 were dominated by exceedingly large signals at one locus (PCs 5-9) or substantial correlation with missing data rate per individual (PC10 $\rho = 0.07, p < 2.2 \times 10^{-16}$), suggesting that these PCs are caused by PC artifacts and do not represent true population structure. PCs 1-4 were not significantly correlated with missing data.

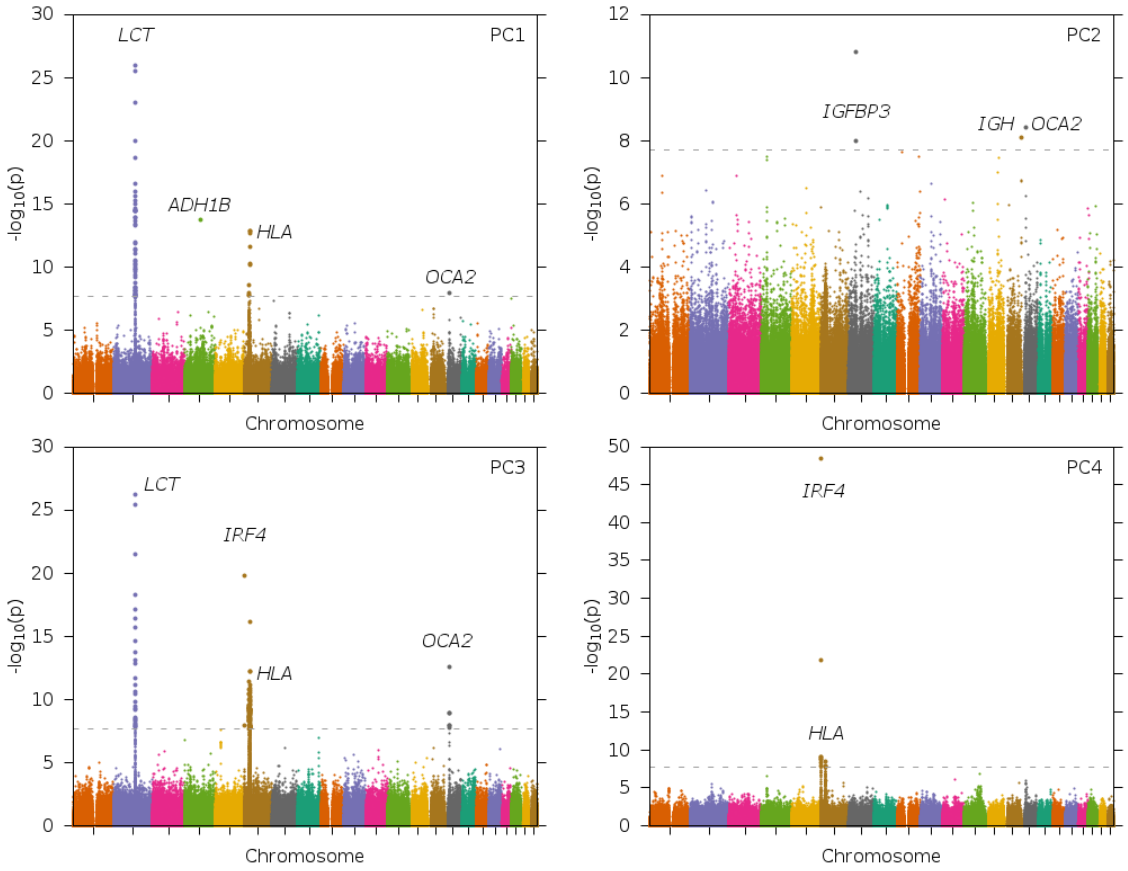

Figure A.7: Selection statistics for PCs 1-4 in GERA data after removing significant regions. We removed the genome-wide significant regions listed in Table 1.1, reran FastPCA and calculated the selection statistic across the genome. The significant hits in PCs 1-4 remain largely unchanged (Figure 1.6). The notable exception is the removal of the inversion on chromosome 8 spanning from 8-12 Mb. This indicates that the signal in that region was artifactual.

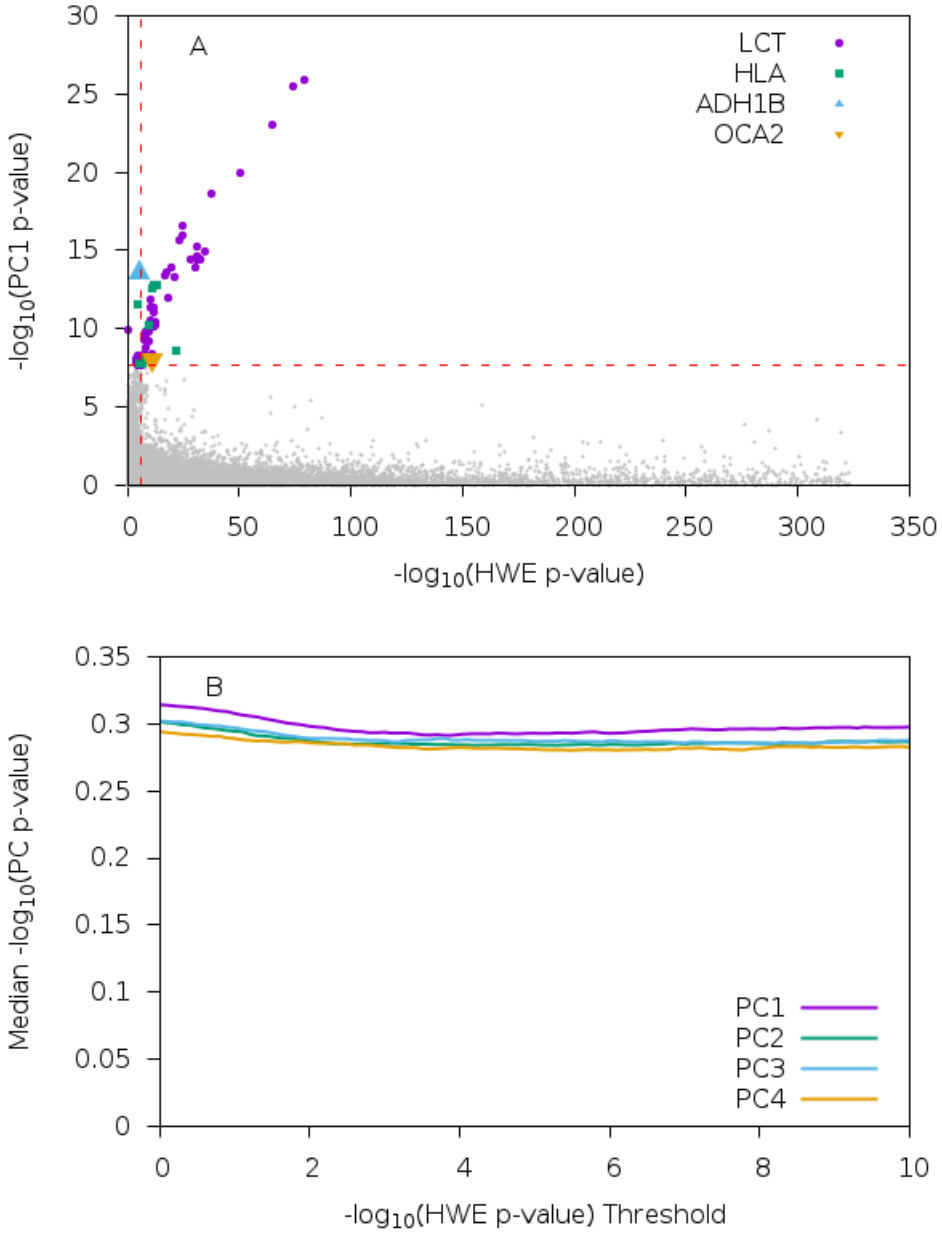

Figure A.8: Comparison of selection statistic and Hardy-Weinberg disequilibrium p-values Removing SNPs with a Hardy-Weinberg p-value less than 10^{-6} (those to the right of the vertical red line) removes many significant signals of selection. (a) For PC1, 51/63 significant SNPs have low Hardy-Weinberg p-values (for PCs 2-4 those numbers are 1/4, 39/116 and 2/12), compared with 3.9% of overall QC SNPs having HW p-value less than 10^{-6}. (b) We found no evidence of more significant selection statistics across PCs 1-4 for SNPs with strongly significant Hardy-Weinberg p-values.

A.2 Supplementary Tables

F_{ST}	N	α	Inflation	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
0.001	50k	0.0	1.01	9.98e-2	9.80e-3	1.02e-3	9.83e-5	1.33e-5
		0.5	1.00	9.98e-2	9.79e-3	9.55e-4	9.83e-5	1.00e-5
		1.0	1.00	9.99e-2	9.78e-3	9.97e-4	9.33e-5	1.17e-5
		2.0	1.01	9.97e-2	9.88e-3	1.01e-3	1.12e-4	1.17e-5
	5k	0.0	1.01	9.97e-2	9.83e-3	1.01e-3	1.02e-4	1.00e-5
		0.5	1.00	9.99e-2	9.88e-3	1.07e-3	9.33e-5	5.00e-6
		1.0	1.00	1.00e-1	9.85e-3	9.47e-4	9.50e-5	1.00e-5
		2.0	1.00	1.00e-1	9.98e-3	1.06e-3	1.18e-4	5.00e-6
	500	0.0	1.01	9.99e-2	9.58e-3	9.03e-4	9.67e-5	6.67e-6
		0.5	1.00	1.00e-1	9.92e-3	9.75e-4	8.50e-5	3.33e-6
		1.0	1.01	1.00e-1	9.82e-3	9.73e-4	7.33e-5	8.33e-6
		2.0	1.00	1.00e-1	1.00e-2	9.63e-4	1.00e-4	1.17e-5
0.01	50k	0.0	1.02	9.95e-2	8.95e-3	8.30e-4	5.83e-5	3.33e-6
		0.5	1.02	9.95e-2	9.00e-3	8.43e-4	6.00e-5	3.33e-6
		1.0	1.02	9.97e-2	8.92e-3	8.37e-4	5.83e-5	3.33e-6
		2.0	1.02	9.96e-2	9.07e-3	8.52e-4	5.67e-5	5.00e-6
	5k	0.0	1.02	9.96e-2	8.87e-3	8.28e-4	6.17e-5	0
		0.5	1.02	9.96e-2	8.99e-3	8.13e-4	5.67e-5	3.33e-6
		1.0	1.02	9.96e-2	9.10e-3	7.78e-4	7.67e-5	3.33e-6
		2.0	1.02	9.99e-2	9.07e-3	8.43e-4	5.50e-5	3.33e-6
	500	0.0	1.03	9.94e-2	8.76e-3	7.72e-4	6.17e-5	1.67e-6
		0.5	1.02	9.94e-2	9.28e-3	8.42e-4	7.00e-5	8.33e-6
		1.0	1.02	1.00e-1	9.24e-3	8.27e-4	8.17e-5	3.33e-6
		2.0	1.01	1.00e-1	9.45e-3	9.55e-4	8.67e-5	1.00e-5
0.1	50k	0.0	1.18	9.32e-2	5.65e-3	2.62e-4	8.33e-6	0
		0.5	1.18	9.32e-2	5.66e-3	2.65e-4	6.67e-6	0
		1.0	1.18	9.32e-2	5.63e-3	2.58e-4	6.67e-6	0
		2.0	1.18	9.32e-2	5.64e-3	2.67e-4	6.67e-6	0
	5k	0.0	1.18	9.32e-2	5.64e-3	2.52e-4	8.33e-6	0
		0.5	1.18	9.34e-2	5.65e-3	2.55e-4	8.33e-6	0
		1.0	1.18	9.33e-2	5.64e-3	2.50e-4	6.67e-6	0
		2.0	1.18	9.34e-2	5.69e-3	2.53e-4	8.33e-6	0
	500	0.0	1.18	9.35e-2	5.61e-3	2.62e-4	1.67e-6	0
		0.5	1.18	9.39e-2	5.78e-3	2.53e-4	8.33e-6	0
		1.0	1.16	9.46e-2	5.87e-3	2.72e-4	3.33e-6	0
		2.0	1.15	9.47e-2	6.23e-3	2.77e-4	5.00e-6	0

Table A.2: Inflation of the selection statistic in simulated data with admixture. We ran 10 simulations containing 60k SNPs and various numbers of simulated individuals (N) in two populations under different levels of admixture and calculated the selection statistic under the null. Admixture was sampled from a $Beta(a, a)$ where an increase in the admixture parameter (a) represents a greater probability of fractional ancestry. When $a = 0$ there is no admixture and when $a = 1$, fractional ancestry follows a $Uniform(0, 1)$ distribution. We report the inflation of the median selection statistic (median divided by the theoretical value of 0.455 under the null) and the proportion of SNPs that attain significance at different thresholds.

Samples	FastPCA			flashpca			PLINK2-PCA			smartpca		
(x1000)	CPU		Mem	CPU		Mem	CPU		Mem	CPU		Mem
1	0:01:42	(0:06)	0.54	0:00:55	(0:01)	1.25	0:00:19	(0:01)	0.02	0:02:10	(0:10)	0.17
1.5	0:02:00	(0:04)	0.55	0:01:41	(0:01)	1.64	0:00:42	(0:01)	0.03	0:05:39	(0:33)	0.25
2	0:02:18	(0:06)	0.57	0:02:44	(0:01)	2.03	0:01:15	(0:01)	0.05	0:10:11	(0:48)	0.35
3	0:02:53	(0:07)	0.59	0:05:38	(0:02)	2.82	0:02:53	(0:04)	0.09	0:23:38	(1:18)	0.58
5	0:03:58	(0:08)	0.64	0:14:31	(0:06)	4.44	0:08:20	(0:17)	0.25	1:11:21	(7:09)	1.19
7	0:05:08	(0:07)	0.69	0:27:24	(0:04)	6.13	0:17:13	(0:19)	0.47	2:21:24	(8:13)	2.02
10	0:06:56	(0:05)	0.77	0:54:37	(0:16)	9.11	0:39:15	(1:08)	0.94	5:15:58	(16:59)	3.64
15	0:09:50	(0:08)	0.89	2:01:16	(0:42)	14.71	1:45:43	(3:51)	2.10	14:13:13	(38:46)	7.39
20	0:13:05	(0:09)	0.98	3:32:55	(0:55)	21.04	3:41:55	(10:06)	3.70	29:34:22	(41:27)	12.44
30	0:19:36	(0:10)	1.22	7:53:56	(2:00)	35.96	11:41:39	(12:20)	8.27			
50	0:29:57	(0:36)	1.69				47:16:16	(50:39)	0.02			
70	0:41:18	(1:16)	2.30									
100	0:56:00	(1:25)	3.20									

Table A.1: CPU time and memory requirements of FastPCA and other methods. We report the running time (in CPU seconds) and memory usage (GB) of PCA implementations, with standard deviation in parentheses. The standard deviation of memory usage was 0.00 GB for all runs. Runs in which smartpca, PLINK2-pca and flashpca exceeded the time constraint (100 hours) or memory constraint (40GB) are denoted as blank entries. When there are few individuals, PLINK2-pca ran faster and consumed less memory than FastPCA. However, FastPCA was able to run on 100k individuals and 100k SNPs in 56 minutes using 3.2GB of memory.

F_{ST}	N_e	τ	N	Inflation	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
0.001	100k	200	50k	1.01	9.98e-2	9.80e-3	1.02e-3	9.83e-5	1.33e-5
			5k	1.01	9.97e-2	9.83e-3	1.01e-3	1.02e-4	1.00e-5
			500	1.01	9.99e-2	9.58e-3	9.03e-4	9.67e-5	6.67e-6
	10k	20	50k	1.00	1.00e-1	9.90e-3	9.75e-4	1.00e-4	8.33e-6
			5k	1.00	1.00e-1	1.01e-2	1.09e-3	1.00e-4	8.33e-6
			500	1.01	1.00e-1	9.61e-3	8.88e-4	1.04e-4	1.25e-5
0.01	10k	200	50k	1.02	9.95e-2	8.95e-3	8.30e-4	5.83e-5	3.33e-6
			5k	1.02	9.96e-2	8.87e-3	8.28e-4	6.17e-5	0
			500	1.03	9.94e-2	8.76e-3	7.72e-4	6.17e-5	1.67e-6
	1k	20	50k	1.02	1.00e-1	9.06e-3	8.22e-4	7.78e-5	1.67e-5
			5k	1.02	1.01e-1	9.17e-3	7.33e-4	6.11e-5	1.11e-5
			500	1.02	1.00e-1	9.07e-3	7.78e-4	7.78e-5	5.56e-6
0.1	1k	200	50k	1.18	9.32e-2	5.65e-3	2.62e-4	8.33e-6	0
			5k	1.18	9.32e-2	5.64e-3	2.52e-4	8.33e-6	0
			500	1.18	9.35e-2	5.61e-3	2.62e-4	1.67e-6	0
	100	20	50k	1.18	9.33e-2	5.76e-3	2.37e-4	0	0
			5k	1.18	9.34e-2	5.75e-3	2.30e-4	0	0
			500	1.18	9.33e-2	5.88e-3	2.07e-4	3.33e-6	0

Table A.3: Inflation of the selection statistic in simulated data with population bottlenecks. We investigated the effect of population bottlenecks on the selection statistic. For a fixed F_{ST}, we would generate two simulated datasets differing in the effective population size (N_e) and number of generations (τ). The statistic remained well calibrated under tighter population bottlenecks. As with Table A.2, the median selection statistic was inflated for simulations with large F_{ST} (at large F_{ST} it is impossible for the selection statistic to be extremely significant, and this deficiency in the tail implies a higher ratio of median to average; see Figure A.1), but well behaved at the small values of F_{ST} that correspond to our analyses of real data. The proportion of SNPs that attain significance was well-calibrated in all experiments.

N_e	PC	Inflation	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
500k	1	1.02	9.96e-2	9.53e-3	8.53e-4	7.5e-5	5.00e-6
	2	1.00	9.94e-2	1.04e-2	1.12e-3	1.02e-4	6.67e-6
	3	0.99	9.95e-2	1.04e-2	1.12e-3	1.32e-4	1.83e-5
	4	0.99	1.00e-2	1.03e-2	1.04e-3	1.10e-4	1.83e-5
50k	1	1.02	9.94e-2	9.55e-3	8.80e-4	7.33e-5	3.33e-6
	2	1.00	9.93e-2	1.06e-2	1.11e-3	9.33e-5	1.17e-5
	3	0.99	1.00e-1	1.05e-2	1.17e-3	1.38e-4	1.50e-5
	4	0.99	1.00e-1	1.03e-2	1.11e-3	1.27e-4	2.33e-5
500	1	1.02	9.95e-2	9.33e-3	8.42e-4	8.00e-5	3.33e-6
	2	0.99	1.00e-1	1.02e-2	1.03e-3	8.50e-5	8.33e-6
	3	1.00	1.00e-1	1.01e-2	1.00e-3	9.17e-5	1.00e-5
	4	1.00	9.98e-2	1.02e-2	1.06e-3	1.20e-4	1.33e-5

Table A.4: Inflation of the selection statistic in simulated data with two levels of F_{ST}. We considered the effect of a more complicated population structure on the selection statistic. We simulated five populations with a phylogenetic structure where three of the populations are more closely related than the other two (see Methods). We again did not see inflation in the median selection statistic nor the proportion of SNPs that attain different significance thresholds.

PCA	LD-pruned				LD-pruned, Table 1.1 Removed			
Selection					Table 1.1 removed			
	LD-pruned		Full		LD-pruned		Full	
	Mean	Med	Mean	Med	Mean	Med	Mean	Med
PC1	1.00	1.02	1.07	1.06	1.00	1.03	1.05	1.06
PC2	1.00	0.98	1.03	1.00	1.00	0.98	1.01	1.00
PC3	1.00	0.95	1.07	0.99	1.00	0.96	1.02	0.99
PC4	1.00	0.96	1.03	0.96	1.00	0.97	0.99	0.96
PC5	1.00	0.12	2.81	0.21	1.00	0.90	0.97	0.89
PC6	1.00	0.89	1.02	0.88	1.00	0.96	0.99	0.96
PC7	1.00	0.92	1.26	0.94	1.00	0.50	0.86	0.47
PC8	1.00	0.34	8.12	0.33	1.00	0.86	0.93	0.81
PC9	1.00	0.40	5.56	0.39	1.00	0.95	0.95	0.89
PC10	1.00	0.49	0.94	0.46	1.00	0.78	0.97	0.70

Table A.5: Inflation of the selection statistic in GERA data. This table indicates the average value of the selection statistic as well as the median selection statistic divided by the theoretical median (0.455) in GERA data. PCA was run on the set of 162,335 LD-pruned SNPs, and the selection statistic was applied to either the set of 162,335 LD-pruned SNPs or the full set of 608,981 SNPs passing QC. Additional analyses were performed with the significant regions from Table 1.1 removed from all SNP sets. When computing selection statistics using the full set of SNPs passing QC, inflation can occur if SNPs with higher differentiation tend to have higher LD, which can occur as a consequence of true selection. PCs 2-4 show moderate inflation when examining the means, but no inflation when looking at the median chi-squared (1 d.o.f) statistic, indicating that inflation is driven by outliers in the distribution. Removing Table 1.1 regions decreased the mean for these PCs, without affecting the median value. For PC1, a qualitatively similar reduction was observed, although a slight inflation in the mean remained. However, after conservatively correcting selection statistics for inflation in the mean and/or median, all SNPs in Table 1.1 remained genome-wide significant except for the OCA2 locus (a known signal of selection) on PC1. For PCs 5-10, the unusual mean and/or median values are consistent with the fact that these PCs are caused by PC artifacts and do not represent true population structure (Figure A.5).

Locus	Chromosome	Region (Mb)	PC	Best Hit	p-value
	1	79.3 - 79.4	2	rs17590370	1.47×10^{-7}
INPP4A	2	98.5 - 98.5	2	rs78108890	5.00×10^{-7}
ANO10	3	43.7 - 43.7	2	rs116086673	1.57×10^{-7}
	4	4.8 - 4.8	3	rs12186237	3.90×10^{-7}
ARAP2	4	35.9 - 35.9	2	rs116105213	3.78×10^{-8}
TLR1[22]	4	38.5 - 38.5	2	rs5743611	5.42×10^{-8}
			4	rs4833095	6.52×10^{-7}
SLC45A2[125]	5	34.0 - 34.0	3	rs16891982	6.89×10^{-8}
	5	89.5 - 89.5	2	rs72779178	4.22×10^{-7}
	6	93.7 - 93.7	1	rs1538270	5.80×10^{-7}
DGKB	7	14.2 - 14.2	1	rs59706690	1.43×10^{-7}
CCDC146	7	76.8 - 76.8	2	rs17151162	5.96×10^{-7}
CADPS2	7	121.8 - 121.8	2	rs6947805	8.58×10^{-7}
PVT1	8	129.1 - 129.1	3	rs12676558	2.26×10^{-7}
EQTN	9	27.3 - 27.3	2	rs41305329	4.25×10^{-8}
RALGPS1	9	128.8 - 128.8	2	rs76798990	4.88×10^{-8}
	9	135.4 - 135.4	2	rs79784812	5.65×10^{-7}
TET1	10	70.1 - 70.1	2	rs7896856	2.71×10^{-7}
	12	94.5 - 94.5	4	rs79822723	2.64×10^{-7}
	13	77.2 - 77.2	2	rs75892602	1.30×10^{-7}
	13	80.4 - 80.4	2	rs117888143	4.13×10^{-8}
	13	83.0 - 83.0	1	rs73234476	7.14×10^{-7}
	14	40.2 - 40.2	1	rs8021234	5.55×10^{-7}
	20	1.8 - 1.8	1	rs6045087	1.05×10^{-7}

Table A.6: Suggestive signals of selection in GERA data. We report the regions with suggestive ($10^{-6} < p < 2.05 \times 10^{-8}$) evidence of selection (analogous to Table 1.1).

Locus	Chromosome	Region (Mb)	PC	Best Hit	p-value
LCT	2	135.0 – 137.1	1	rs6754311	1.23×10^{-26}
			3	rs4988235	5.65×10^{-27}
ADH1B	4	100.5	1	rs1229984	1.76×10^{-14}
IRF4	6	0.3 – 0.5	3	rs12203592	1.61×10^{-20}
			4		3.29×10^{-49}
HLA	6	31.1 – 32.8	1	rs382259	1.47×10^{-13}
			3	rs9268628	7.15×10^{-17}
			4	rs1265103	2.84×10^{-9}
IGFBP3	7	45.3-45.9	2	rs150353309	1.53×10^{-11}
IGH	14	106.0	2	rs34614900	7.86×10^{-9}
OCA2	15	25.9 – 26.2	1	rs12916300	1.26×10^{-8}
			2		3.76×10^{-9}
			3		2.67×10^{-13}

Table A.7: Top signals of selection in GERA data using PCs computed from SNPs in other regions. We report the regions with suggestive ($10^{-6} < p < 2.05 \times 10^{-8}$) evidence of selection (analogous to Table 1.1).

		AJ	**EE**	**IR**	**NE**	**SE**
Count		2,750	4,196	14,771	28,439	4,578
ADH1B	rs1229984	21.37%	4.99%	2.66%	2.96%	9.58%
IGFBP3	rs150353309	1.66%	4.38%	0.76%	1.10%	0.79%
	rs35751739	2.47%	7.71%	2.68%	3.06%	2.19%
IGH	rs34614900	13.63%	26.78%	17.29%	18.92%	12.73%

	AJ	**EE**	**IR**	**NE**
EE	0.00684			
IR	0.00671	0.00095		
NE	0.00655	0.00073	0.00013	
SE	0.00345	0.00239	0.00193	0.00182

Table A.8: Allele frequencies for highlighted loci in GERA subpopulations. The GERA sample was clustered into 5 discrete subpopulations using k-means clustering run on the top 4 PCs. Individual clusters were labelled to coincide with SNPweights and projected POPRES individuals. These were Ashkenazi Jewish (AJ), Eastern European (EE), Irish (IR), Northern European (NE) and South-east European (SE). Results are reported only for genome-wide significant SNPs at highlighted loci. We also report F_{ST} between each pair of subpopulations.

rs1229984	AJ	EE	IR	NE	SE
AJ	1.47e-6				
EE	4.15e-5	0.556			
IR	8.31e-7	0.00731	1.83e-8		
NE	1.04e-6	0.00932	0.293	2.61e-10	
SE	0.000121	0.0126	4.98e-6	8.84e-6	0.00012

Table A.9: Natural selection at ADH1B between discrete subpopulations. The discrete-population selection statistic [12] (see Methods) for each pair of populations was calculated (below the diagonal) as well as the statistic comparing the frequency of rs1229984 in that population with the set of remaining individuals (diagonal). Genome-wide significant comparisons are those with $p < 5.47 \times 10^{-9}$ (608,981 SNPs $\times$ 15 subpopulation comparisons = 9,134,715 tests with $\alpha = 0.05$).

Haplotype	rs1693439	rs3811801	rs1159918	rs1229984	rs4147536	rs2075633	rs2066701	rs17033	rs1042026	Asian (CHB, CHS, JPT)	European (CEU, FIN, GBR, IBS, TSI)	African (ASW, LWK, YRI)
H1b	G	G	C	C	C	T	G	T	T	1.96%	40.11%	14.97%
H1c	G	G	C	C	A	T	G	T	T	0%	0.14%	5.21%
H2	G	G	A	C	C	T	G	T	T	0%	0.84%	18.66%
H2b	G	G	A	C	C	T	G	C	T	9.46%	10.10%	9.33%
H3	G	G	C	C	C	C	A	T	C	8.04%	27.21%	4.34%
H3c	G	G	C	C	C	C	G	T	T	0%	0%	0.43%
H4	G	G	A	C	A	T	G	T	T	6.96%	17.67%	46.42%
H4b	A	G	A	C	A	T	G	T	T	0%	1.96%	0.65%
H5	G	G	C	T	C	T	G	T	T	0.36%	1.12%	0%
H5b	A	G	A	T	A	T	G	T	T	0.18%	0.56%	0%
H6	G	G	C	T	C	C	A	T	C	12.14%	0.28%	0%
H7	G	A	C	T	C	C	A	T	C	60.89%	0%	0%

Table A.10: We computed frequencies of known haplotypes in 1000 genomes Asian, European and African populations. 9 SNPs were used to determine haplotype and haplotypes not described in Li *et al.*[79] were excluded from the analysis. 98% of the European haplotypes did not contain rs1229984*T (above line) compared to 20.8% of Asian haplotypes. The "A" allele of regulatory SNP rs3811801 was not found at all in European populations, while haplotype H7 which contains this allele is the most common haplotype in Asian populations.

rs150353309	**AJ**	**EE**	**IR**	**NE**	**SE**
AJ	0.755				
EE	0.178	4.07e-7			
IR	0.48	4.38e-7	0.00441		
NE	0.678	4.62e-7	0.0429	0.217	
SE	0.351	0.0014	0.955	0.6	0.374
rs35751739					
AJ	0.675				
EE	0.0438	1.24e-7			
IR	0.909	5.99e-7	0.0703		
NE	0.757	2.33e-7	0.207	0.451	
SE	0.827	0.000332	0.614	0.379	0.233

Table A.11: Natural selection at IGFBP3 between discrete subpopulations. As in Table A.9, but for SNPs rs150353309 and rs150353309 in *IGFBP3* which were under selection. Genome-wide significant comparisons are those with $p < 5.47 \times 10^{-9}$ (608,981 SNPs $\times$ 15 subpopulation comparisons = 9,134,715 tests with $\alpha = 0.05$).

rs34614900	**AJ**	**EE**	**IR**	**NE**	**SE**
AJ	0.23				
EE	0.00557	8.17e-8			
IR	0.386	4.43e-7	0.12		
NE	0.214	2.65e-6	0.0165	0.173	
SE	0.754	6.35e-7	0.0437	0.00577	0.00347
rs35237072					
AJ	0.378				
EE	0.0151	2.76e-7			
IR	0.554	1.37e-6	0.151		
NE	0.373	3.21e-6	0.0569	0.432	
SE	0.771	1.13e-5	0.139	0.0384	0.0245
rs34479337					
AJ	0.616				
EE	0.0472	1.52e-6			
IR	0.745	1.39e-5	0.371		
NE	0.613	9.15e-6	0.247	0.655	
SE	0.305	6.72e-6	0.0489	0.0183	0.0079

Table A.12: Natural selection at IGH between discrete subpopulations. As in Table A.9, but for SNP rs34614900 in *IGH* which was under selection and SNPs rs35237072 and rs34479337 were suggestive with p-value $< 10^{-6}$. Genome-wide significant comparisons are those with $p < 5.47 \times 10^{-9}$ (608,981 SNPs $\times$ 15 subpopulation comparisons = 9,134,715 tests with $\alpha = 0.05$).

B

Supplementary Materials for Chapter 2

B.1 Supplementary Figures

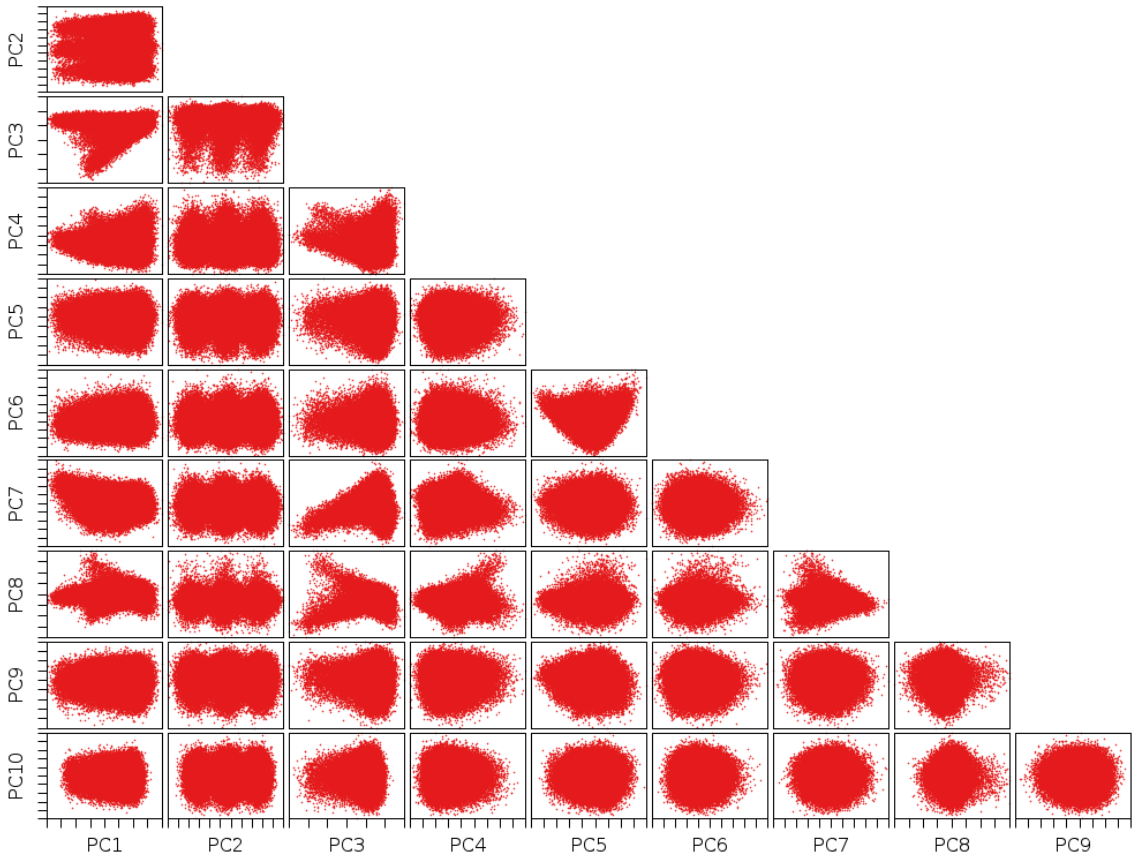

Figure B.1: Results of initial PCA run. Shown are the PCA plots for PC1 to PC10 after the initial PCA run. Several of these PCs are dominated by regions of long-range LD. In particular, the three clusters along PC2 indicate 0, 1 or 2 copies of a chromosome 8 inversion variant.

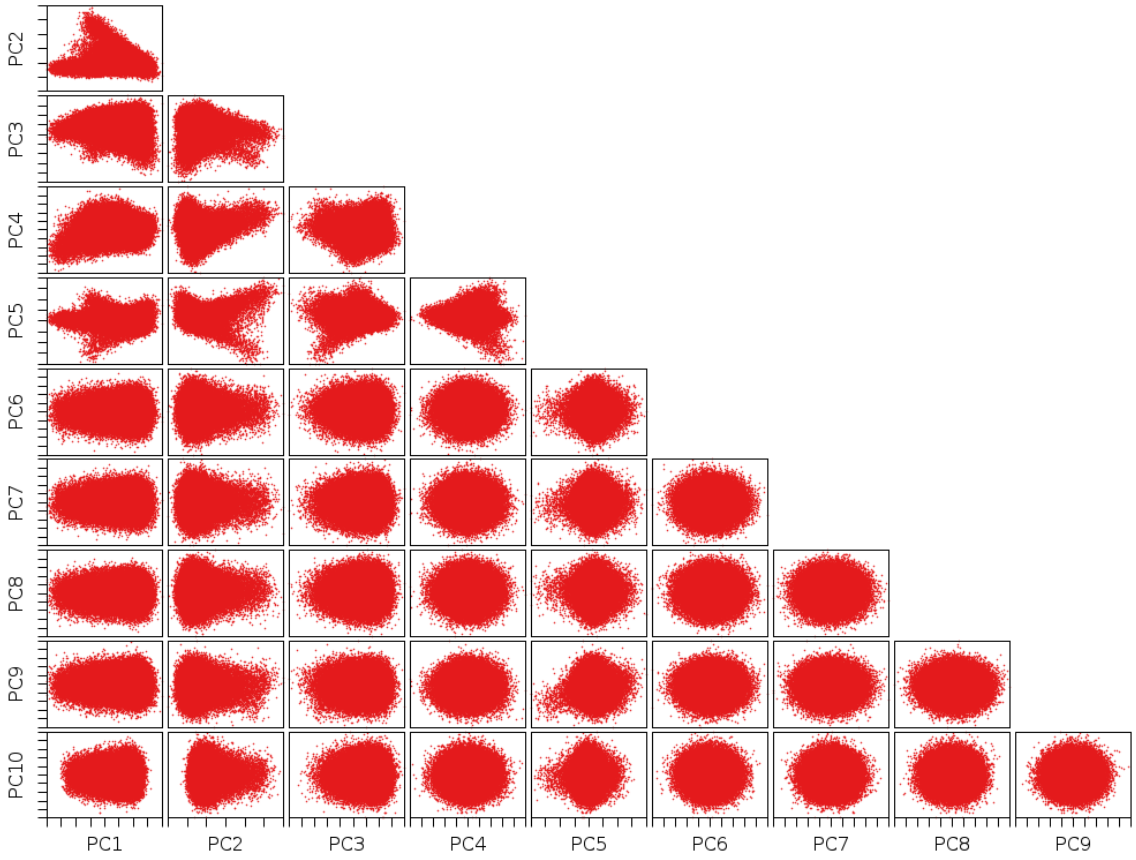

Figure B.2: Results of PCA after removing long-range LD regions. Regions with high SNP weights from the first PCA run were removed and PCA was run on the remainder of the genome (see Methods). The resulting PCs are no longer influenced by long-range LD regions. A visual inspecion suggests that PC1-PC5 have interesting population structure while PC6-PC10 do not.

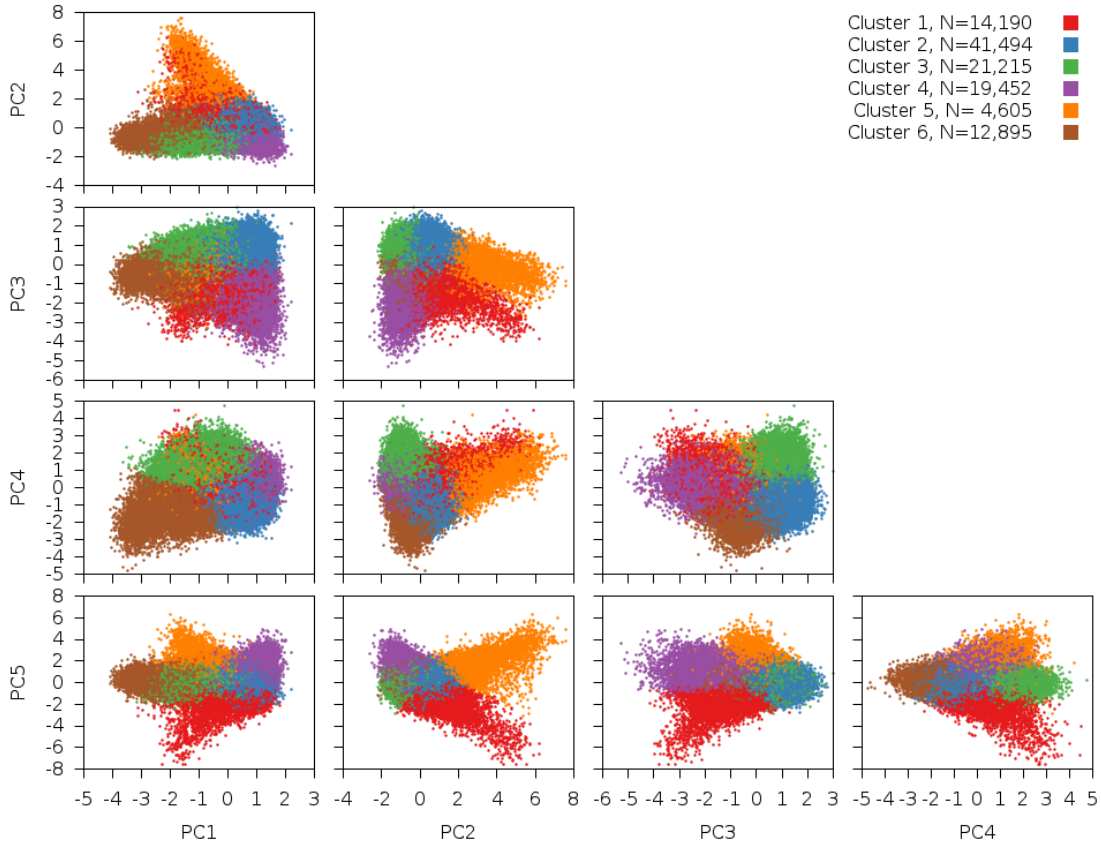

Figure B.3: Results of PCA with k-means clustering for all PCs. This is an expanded set of plots similar to Figure 2.1, except that plots of all pairs of top PCs are displayed.

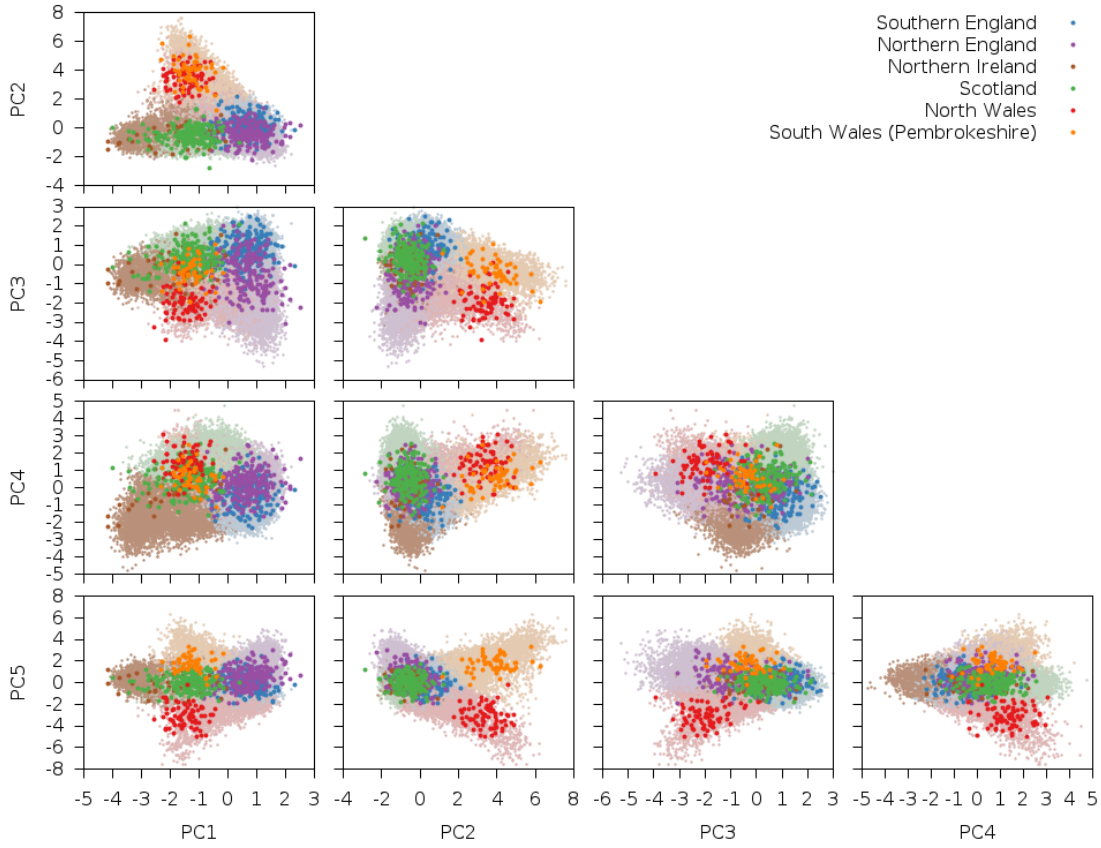

Figure B.4: Results of PCA with projection of PoBI samples for all PCs. This is an expanded set of plots similar to Figure 2.2, except that plots of all pairs of top PCs are displayed.

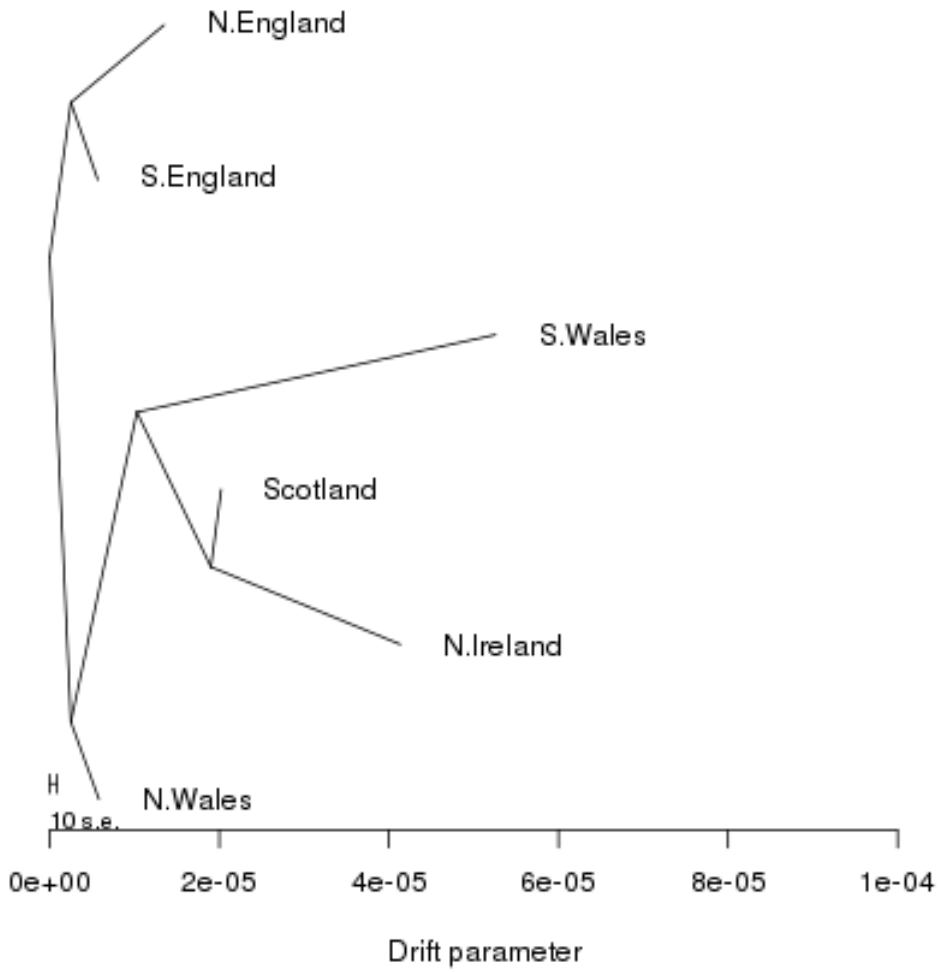

Figure B.5: Tree-based clustering of UK Biobank subpopulation clusters. We clustered UK Biobank subpopulation clusters with TreeMix. We found that Northern and Southern England clusters were grouped together while the Celtic-related clusters formed a separate branch to the tree. The disparity between the north and south Welsh clusters is due to the north Wales cluster containing Saxon samples (see F_{ST} values in Table B.6).

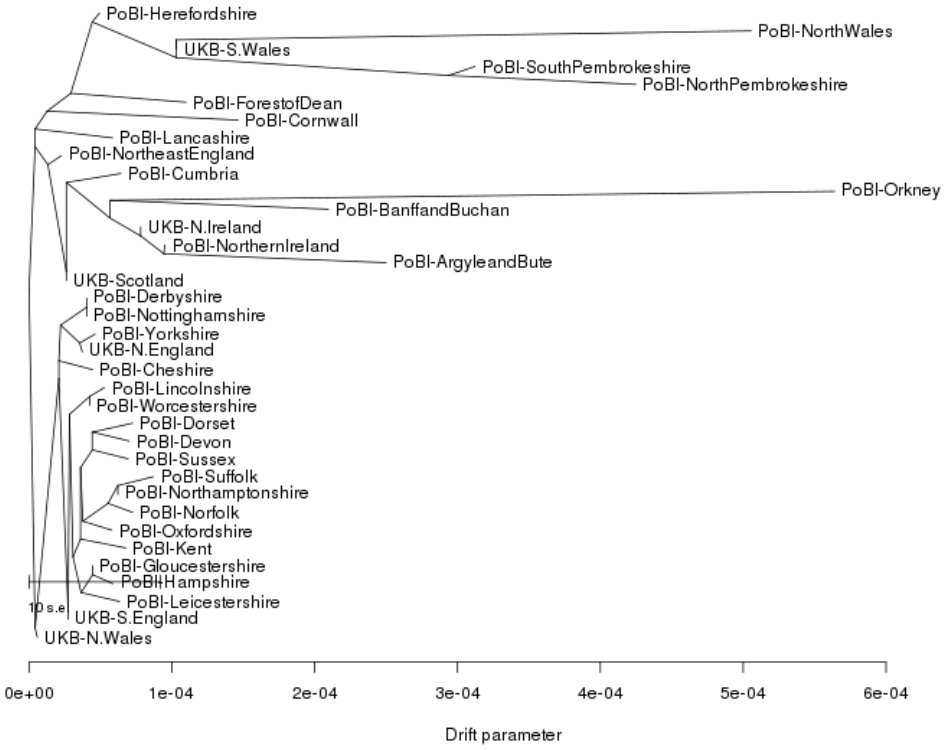

Figure B.6: Tree-based clustering of UK Biobank clusters and PoBI populations. Here we see how the UK Biobank subpopulations cluster with the PoBI ones. As in Figure B.5, the tree is split roughly into the "Celtic" (top) and "Saxon" (bottom) groups. The south Wales (UKB-S.Wales) cluster is grouped with the Welsh populations from PoBI while the north Wales (UKB-N.Wales) cluster is in the bottom group with the north and south England clusters. As in Figure B.5, the disparity between the north and south Welsh clusters is due to the north Wales containing Saxon samples (see F_{ST} values in Table B.6).

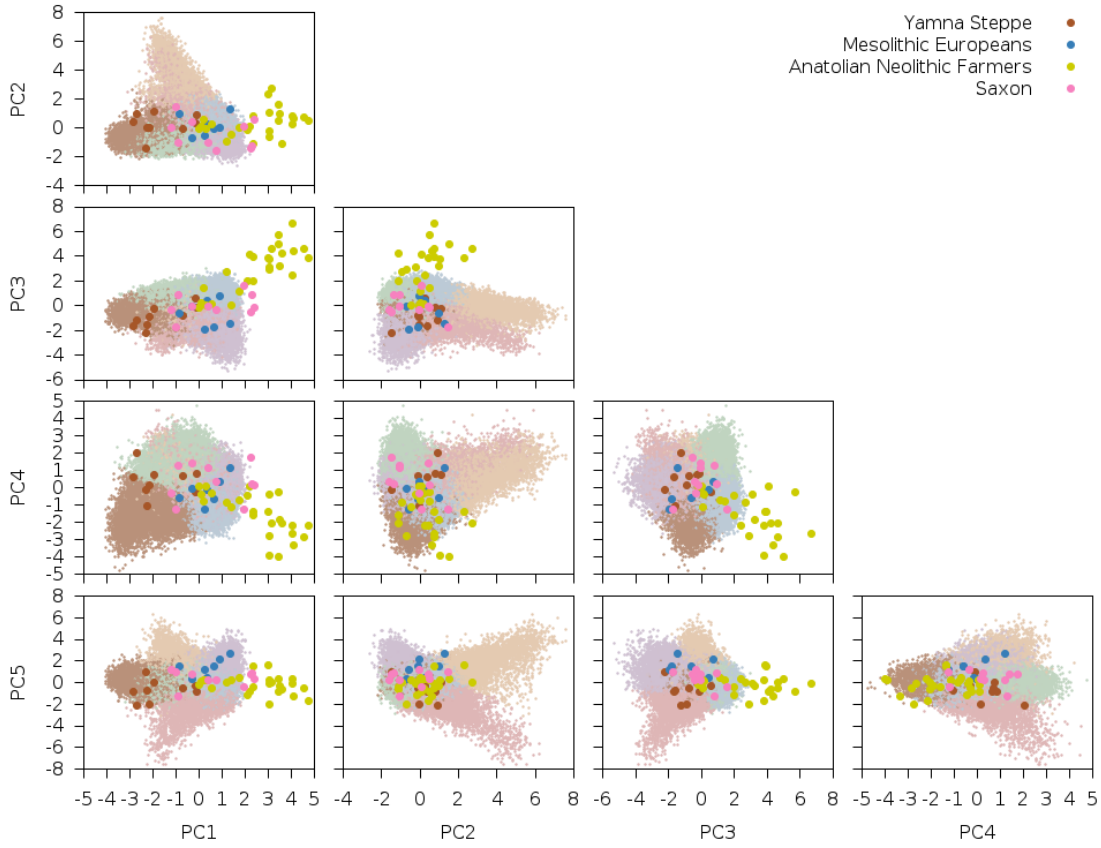

Figure B.7: Results of PCA with projection of ancient samples for all PCs. This is an expanded set of plots similar to Figure 2.3, except that plots of all pairs of top PCs are displayed.

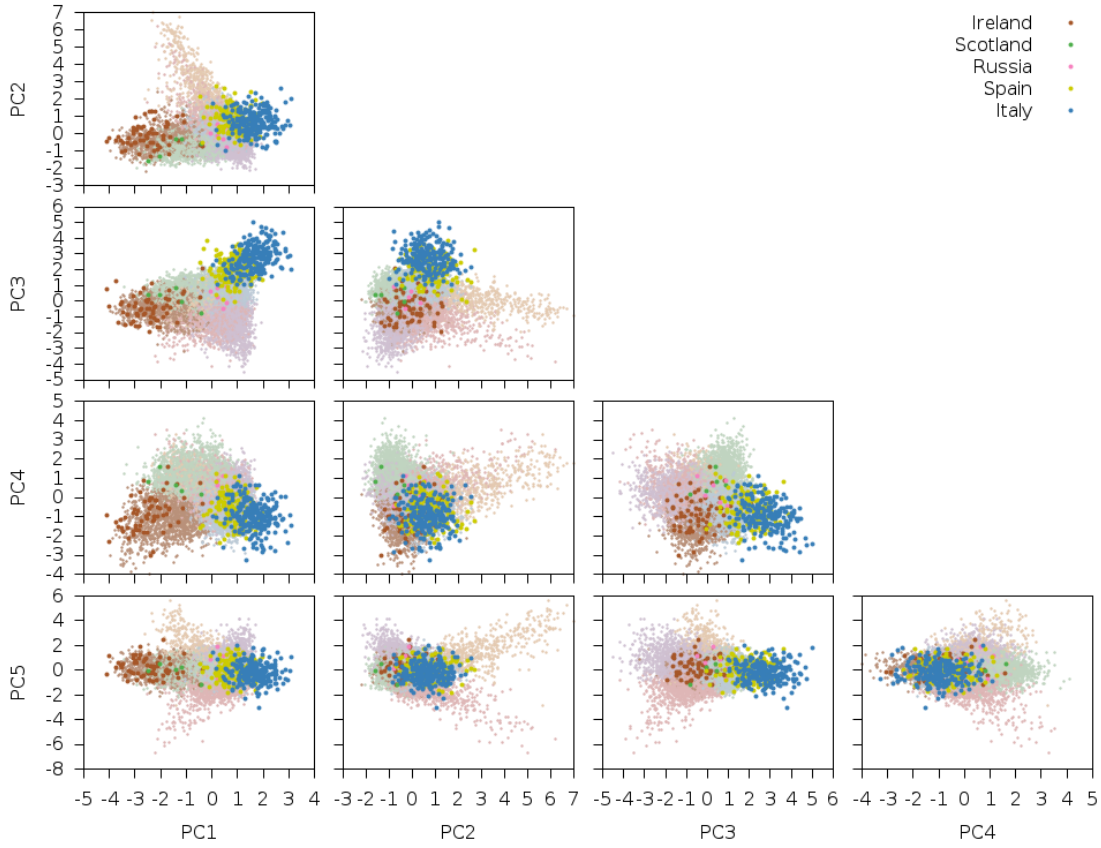

Figure B.8: Results of PCA with projection of POPRES samples for all PCs. This set of plots is similar to Figure B.3, except that POPRES samples are projected on top.

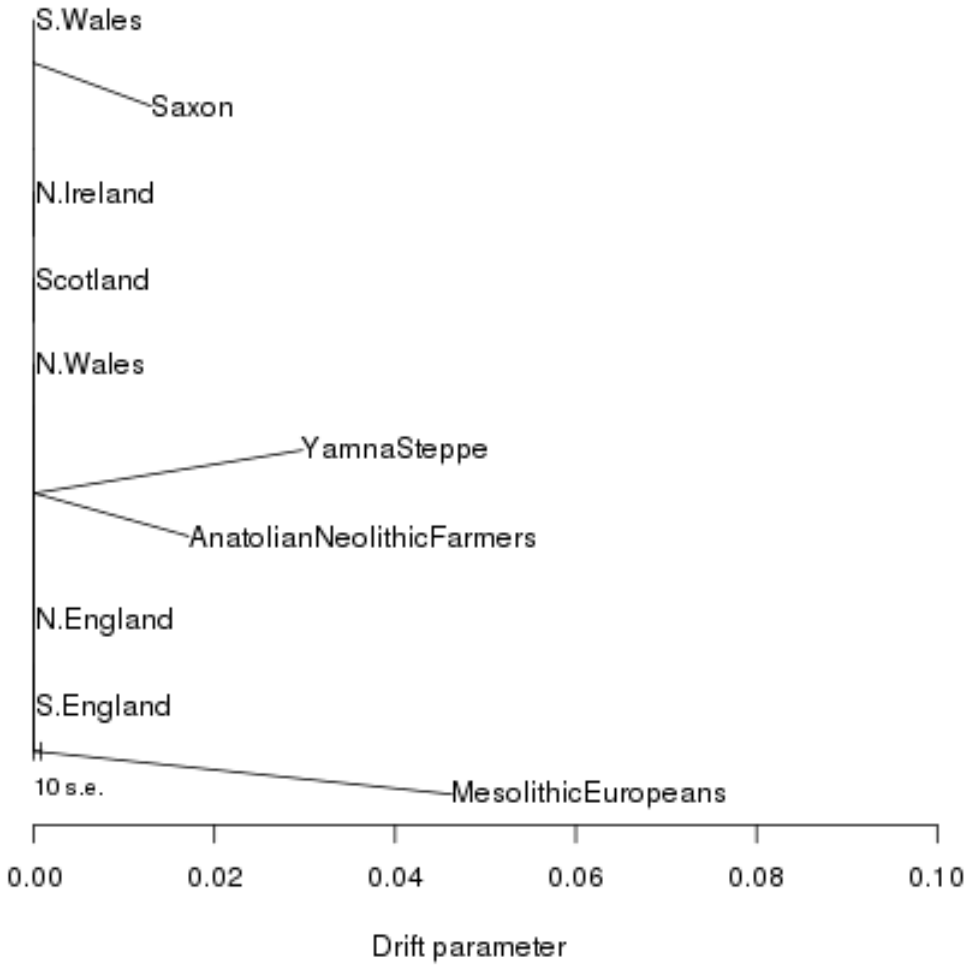

Figure B.9: Tree-based clustering of UK Biobank clusters and ancient populations. We found that the ancient samples were too diverged from both the UK Biobank samples and from each other to form meaningful trees. Of note, the Mesolithich Europeans formed an outgroup and the Saxons were grouped with the south Wales cluster.

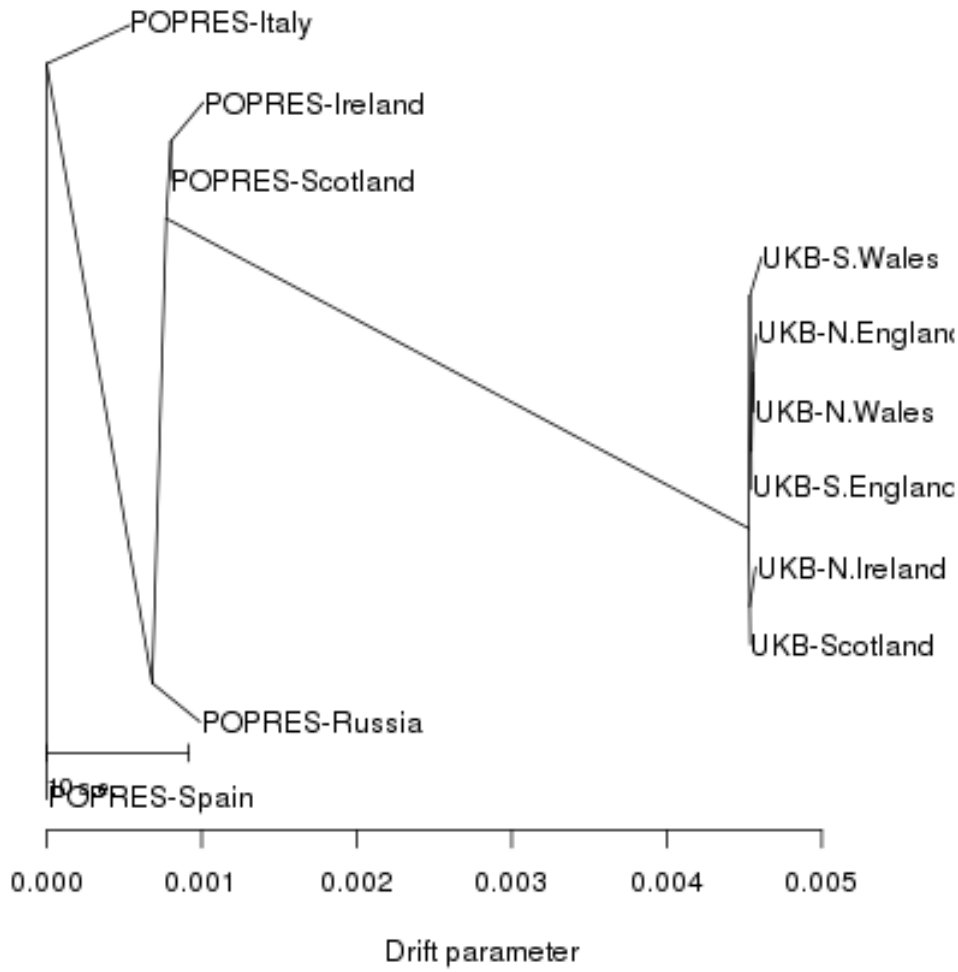

Figure B.10: Tree-based clustering of UK Biobank clusters and POPRES populations. We see a bit of a batch effect differentiating the UK Biobank clusters from the POPRES populations. However, the UK Biobank samples do lie near the POPRES Irish and Italian samples. The Russian samples (which contain more Steppe ancestry) also cluster more closely with the UK Biobank samples than the Italian and Spanish samples.

Figure B.11: Selection statistic for UK Biobank along PC1-PC5. This is an expanded set of plots similar to Figure 2.4, except that plots for each of the top 5 PCs are displayed.

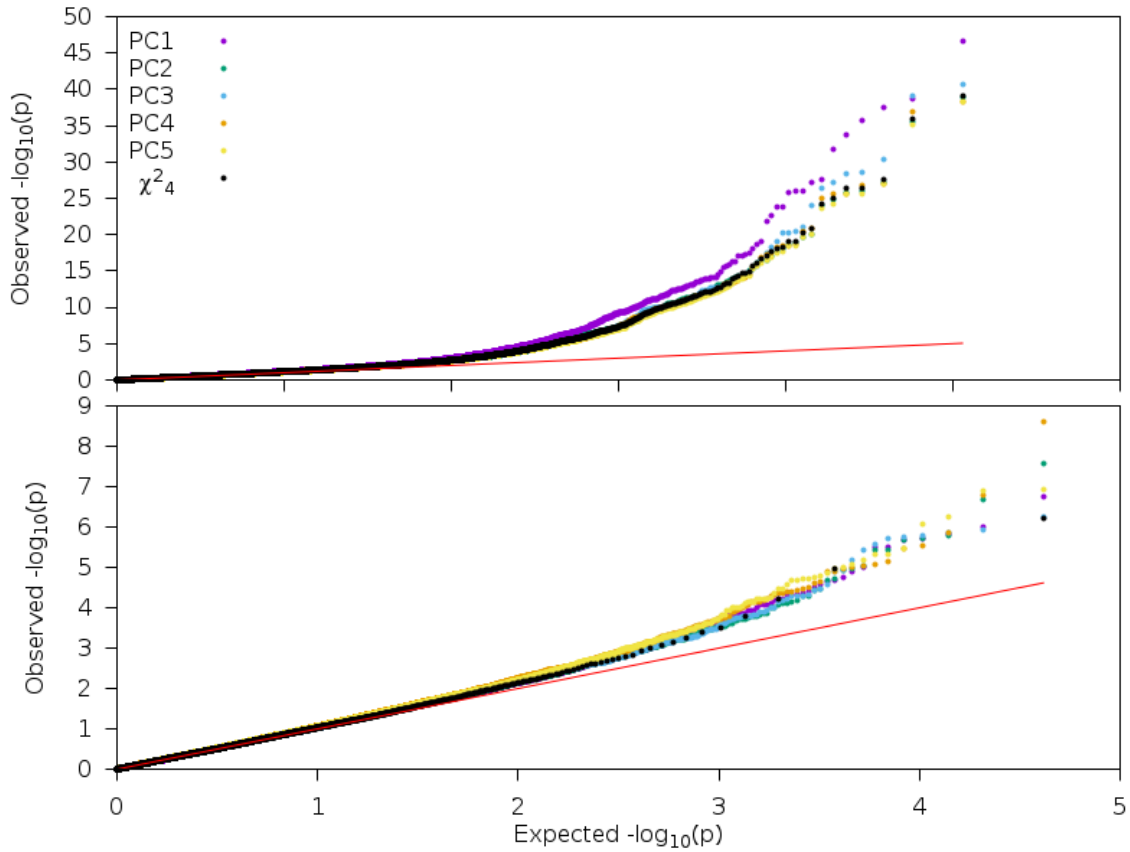

Figure B.12: P-P plot of the combined selection statistic. We plot the observed $-\log_{10}(p)$ values compared to the expected distribution for the combined selection statistics as well as for the χ^2_4 statistic from Mathieson et al. The plot of all overlapping SNPs (top) suggests some inflation in the tails of the combined selection statistics, although λ_{GC} values were only slightly inflated (1.04-1.06; see Table B.8) and this is largely due to the inflation of the χ^2_4 statistic from Mathieson et al. After LD-pruning by intersecting the SNPs with the PC dataset and removing two SNPs with $p < 5 \times 10^{-8}$ for the χ^2_4 statistic from Mathieson et al., we see much less inflation in the tails (λ_{GC} values were reduced to 1.01-1.03), indicating that the inflation in the tails was largely due to SNPs in LD with top hits from Mathieson et al. In order to produce a maximally conservative statistic, we corrected our combined statistics by their λ_{GC} values (Table B.8).

B.2 Supplementary Tables

Chrom	Locus (Mb)	PC	Best hit	*p*-value
1	2.2 - 2.2	3	rs79907870	4.68e-7
1	54.8 - 54.8	1	rs17390412	9.13e-7
1	56.0 - 56.1	8	rs1875068	1.86e-8
2	88.7 - 88.7	4	rs1713939	1.02e-7
2	133.3 - 144.0	1	rs7570971	7.21e-18
		4	rs1446585	3.02e-14
		7		3.09e-13
		10	rs72847650	<1e-50
2	159.8 - 160.0	7	rs1522699	3.89e-7
2	223.9 - 223.9	7	rs1900725	2.71e-7
3	46.3 - 46.4	7	rs9990343	2.80e-8
4	23.3 - 23.3	3	rs114557362	2.95e-7
4	38.7 - 38.9	1	rs4833095	9.29e-16
		3		8.54e-11
		4		1.21e-10
		7		2.18e-16
5	60.6 - 60.6	3	rs10471511	6.04e-7
5	101.5 - 101.6	8	rs411954	1.20e-8
5	114.8 - 114.8	8	rs895291	5.87e-7
5	164.8 - 164.9	3	rs77635680	6.70e-10
6	0.4 - 0.7	1	rs62389423	1.29e-47
		7		1.57e-9
6	23.9 - 36.7	1	rs151341075	3.76e-11
		3	rs2253908	5.43e-7
		4	rs151341075	3.35e-17
		5	rs3131618	<1e-50
		6	rs204999	<1e-50
		7	rs2596573	3.27e-54
		8	rs41268932	2.58e-23
		9	rs2596573	<1e-50
		10	rs9266258	4.14e-7

Table B.1: Significant or suggestive signals of selection in initial PCA run. We report significant or suggestive signals of selection in the initial PCA run. Neighboring SNPs <1Mb apart with genome-wide significant signals were grouped together into a single locus. The significant signals may represent either signals of selection or regions of long-range LD. All of these regions were removed from the main PCA run (see Methods).

Chrom	Locus (Mb)	PC	Best hit	*p*-value
6	46.8 - 46.8	7	rs9395218	9.92e-7
6	86.0 - 87.0	10	rs2816583	2.23e-8
7	41.4 - 41.4	1	rs76920365	3.66e-7
7	64.4 - 66.4	3	rs79415723	1.17e-8
7	118.2 - 118.3	1	rs187417794	1.13e-7
8	7.2 - 12.7	2	rs11250099	<1e-50
9	14.0 - 14.0	3	rs12380860	4.45e-7
10	133.2 - 133.2	4	rs57105422	7.45e-7
15	28.4 - 28.4	3	rs12913832	9.17e-10
15	50.8 - 50.8	5	rs148783236	3.18e-194
		6		<1e-50
		8		2.18e-13
		9		<1e-50
16	9.5 - 9.5	4	rs12149526	6.26e-7
16	26.5 - 26.5	3	rs73528772	4.01e-7
16	53.7 - 53.7	3	rs61747071	1.36e-7
16	89.7 - 89.8	4	rs449882	5.37e-7
17	29.6 - 29.6	3	rs11655238	5.98e-7
19	33.8 - 33.8	3	rs41355649	3.65e-8
19	49.2 - 50.2	1	rs601338	1.05e-9
20	39.1 - 39.1	1	rs2143877	8.34e-8
22	32.9 - 32.9	5	rs115815765	7.40e-31
		6		<1e-50

Table B.1: (Continued)

	Eigenvalue	F_{ST}	East-West Correlation	p-value	North-South Correlation	p-value
PC1	20.99	1.76e-4	0.4154	<1e-50	-0.3981	<1e-50
PC2	9.35	7.33e-5	-0.0865	<1e-50	-0.4322	<1e-50
PC3	7.76	5.94e-5	0.1894	<1e-50	-0.1262	<1e-50
PC4	5.18	3.68e-5	-0.1418	<1e-50	0.3409	<1e-50
PC5	5.13	3.63e-5	0.0019	5.27e-1	-0.0124	4.14e-5
PC6	4.62	3.18e-5	-0.0163	6.84e-8	0.0150	6.93e-7
PC7	4.61	3.17e-5	-0.0025	4.01e-1	0.0049	1.04e-1
PC8	4.59	3.15e-5	0.0216	7.75e-13	0.0047	1.16e-1
PC9	4.59	3.15e-5	-0.0522	<1e-50	-0.0119	7.92e-5
PC10	4.57	3.14e-5	-0.0143	2.23e-6	0.0121	6.47e-5

Table B.2: PC eigenvalues and geographical correlations. PC1-PC5 all had elevated eigenvalues, while PC6-PC10 had eigenvalues which were close to background levels. In the case where there are two equal-sized sample sets from distinct populations, the F_{ST} between the two populations can be estimated from the top eigenvalue (λ) via the following formula: $F_{ST} = (\lambda - 1)/N$, where N is the total number of samples. The top eigenvalue reflects an F_{ST} of 1.76×10^{-4}, indicating very subtle population structure within the UK. PC1 was most strongly correlated with east-west birth coordinate and PC2 was most strongly correlated with north-south birth coordinate.

		Pop2				
Grouping	***Pop1***	Norfolk	Suffolk	Hampshire	Kent	Devon
Saxon	Saxon	5.118	5.268	2.543	3.953	3.32
Scotland	Argyll and Bute	9.560	9.370	3.323	6.411	6.223
	Banff and Buchan	7.609	77.545	1.234	4.440	4.379
	Orkney	11.229	10.583	3.620	7.310	7.259
N. Wales	North Wales	8.490	8.393	1.918	5.163	5.239
S. Wales	North Pembrokeshire	7.124	7.287	1.759	4.542	4.430
	South Pembrokeshire	6.301	6.189	2.315	4.336	4.171

Table B.3: Expanded results of f_4 statistics in ancient and modern British samples. We report f_4 statistics of the form $f_4\,(Steppe, Neolithic\,Farmer; Pop1, Pop2)$, representing a z-score with positive values indicating more Steppe ancestry in *Pop1* than *Pop2* . Samples for *Pop1* were either modern Celtic (Scotland and Wales) or ancient Saxon. Samples for *Pop2* were modern Anglo-Saxon (southern and eastern England).

Annotation	Chromosome	Locus (Mb)	PC	Best hit	p-value
-	1	208.8 - 208.8	2	rs75602597	9.71e-7
ABCD3	1	226.4 - 226.8	4	rs72759068	8.62e-8
-	4	45.2 - 45.2	1	rs77147311	9.78e-7
-	5	164.8 - 164.9	2	rs77635680	2.13e-8
ZDHHC14	6	158.1 - 158.1	4	rs73584091	5.46e-7
-	7	64.9 - 64.9	2	rs79415723	8.81e-7
-	7	118.2 - 118.2	1	rs187417794	3.66e-7
-	13	66.2 - 66.2	3	rs1417218	8.38e-7
OCA2[30,22]	15	28.4 - 28.4	2	rs12913832	5.55e-8
RPGRIP1L	16	53.7 - 53.7	2	rs61747071	7.81e-7
BPIFB9P	20	31.8 - 32.0	4	rs293709	3.00e-7
-	20	39.1 - 39.1	1	rs2143877	5.03e-7

Table B.4: Suggestive signals of selection in UK Biobank. We report the top signal of natural selection for each locus not reaching genome-wide significance ($p > 1.96 \times 10^{-8}$) but yielding a suggestive signal ($p < 1.00 \times 10^{-6}$) along any of the top five PCs. Neighboring SNPs <1Mb apart with suggestives significant signals were grouped together into a single locus.

Dataset	Cluster	rs601338 (G/A)	rs492602 (G/A)	rs676388 (T/C)
UK Biobank	Northern Ireland	0.4406	0.4413	0.4207
	Northern England	0.4633	0.4638	0.4407
	Pembrokeshire	0.4864	0.4871	0.4580
	North Wales	0.5006	0.501	0.4781
	Yorkshire	0.5025	0.503	0.4763
	Southern England	0.5109	0.5111	0.4847
GERA	Irish	-	0.4754	0.4522
	Northern European	-	0.5215	0.4962
	Southern European	-	0.5248	0.5021
	Ashkenazi Jewish	-	0.5530	0.5200
	Eastern European	-	0.5840	0.5586

Table B.5: Allele frequency of FUT2 alleles. We report the allele frequency of the most significant hit, rs601338, along with two other linked SNPs in GERA and the PoBI datasets.

Dataset	Cluster	rs601338 (G/A)	rs492602 (G/A)	rs676388 (T/C)
PoBI	Argyll and Bute	-	-	0.2949
	North Pembrokeshire	-	-	0.3171
	Banff and Buchan	-	-	0.3365
	Northern Ireland	-	-	0.3667
	Cumbria	-	-	0.4039
	Derbyshire	-	-	0.4091
	Dorset	-	-	0.4125
	Herefordshire	-	-	0.4259
	Worcestershire	-	-	0.4265
	Lancashire	-	-	0.4306
	Devon	-	-	0.4416
	Yorkshire	-	-	0.4517
	Orkney	-	-	0.4531
	Lincolnshire	-	-	0.4619
	Kent	-	-	0.4661
	Suffolk	-	-	0.4699
	Cornwall	-	-	0.4716
	Leicestershire	-	-	0.4726
	North Wales	-	-	0.4737
	Northeast England	-	-	0.4844
	Cheshire	-	-	0.4875
	Forest of Dean	-	-	0.4881
	Norfolk	-	-	0.4951
	Nottinghamshire	-	-	0.5000
	Northamptonshire	-	-	0.5000
	Oxfordshire	-	-	0.5054
	Sussex	-	-	0.5167
	Gloucestershire	-	-	0.5417
	South Pembrokeshire	-	-	0.5417
	Hampshire	-	-	0.5833

Table B.5: (Continued)

Population 1	Population 2	F_{ST}	rs601338	rs492602	rs676388
N. England	S. England	7.36e-5	2.22e-1	2.29e-1	2.13e-1
	N. Ireland	2.96e-4	1.28e-6	1.42e-6	1.31e-5
	Scotland	1.48e-4	2.38e-5	2.50e-5	1.25e-4
	N. Wales	8.53e-5	7.94e-1	7.99e-1	8.15e-1
	S. Wales	3.75e-4	2.77e-1	2.85e-1	2.17e-1
S. England	N. Ireland	2.67e-4	***6.04e-9***	7.35e-9	1.07e-7
	Scotland	1.10e-4	***2.61e-9***	***3.11e-9***	3.38e-8
	N. Wales	7.46e-5	1.42e-1	6.86e-2	3.41e-1
	S. Wales	2.90e-4	6.45e-2	6.86e-2	4.27e-2
N. Ireland	Scotland	1.22e-4	9.31e-3	9.86e-3	2.12e-2
	N. Wales	2.23e-4	1.42e-7	1.58e-7	4.45e-7
	S. Wales	3.43e-4	1.47e-3	1.48e-3	9.38e-3
Scotland	N. Wales	1.23e-4	1.95e-5	2.00e-5	1.81e-5
	S. Wales	3.10e-4	9.12e-2	8.93e-2	2.06e-1
N. Wales	S. Wales	2.60e-4	2.72e-1	2.78e-1	1.18e-1

Table B.6: Discrete test for natural selection at FUT2 in UK Biobank. We report results of tests for selection using discrete subpopulations for *FUT2* in the UK Biobank data set, using the UK Biobank subpopulations derived from k-means clustering. With 15 comparisons per SNP and 510,665 SNPs (p-value threshold of 6.53×10^{-9}), we are still able to find genome-wide-significant results when comparing southern England with Scotland as well as Northern Ireland (emphasized in bold italic).

Population 1	Population 2	F_{ST}	rs492602	rs676388
Ashkenazi Jewish	Eastern European	6.84e-3	5.96e-1	5.13e-1
	Irish	6.71e-3	1.84e-1	2.45e-1
	Northern European	6.54e-3	5.83e-1	6.79e-1
	Southeast European	3.45e-3	5.05e-1	6.73e-1
Eastern European	Irish	9.44e-4	***1.48e-6***	***2.49e-6***
	Northern European	7.23e-4	1.58e-3	1.69e-3
	Southeast European	2.39e-3	9.25e-2	1.10e-1
Irish	Northern European	1.26e-4	***1.34e-7***	***4.53e-7***
	Southeast European	1.91e-3	1.16e-1	1.12e-1
Northern European	Southeast European	1.80e-3	9.12e-1	8.46e-1

Table B.7: Discrete test for natural selection at FUT2 in GERA. We report results of tests for selection using discrete subpopulations for *FUT2* in the GERA data set. *FUT2* does not reach genome-wide significance in the GERA dataset, however there are several suggestive signals when comparing the "Irish" subgroup with the Northern European and Eastern European subpopulation (emphasized in bold italic).

	Inflation Genome-wide	Overlap	Combined	Correlation
Ancient	1.00	1.07	-	-
UKB PC1	1.02	1.08	1.06	18.8%
UKB PC2	0.95	1.00	1.05	2.8%
UKB PC3	0.95	1.00	1.05	6.5%
UKB PC4	0.88	0.94	1.04	2.5%
UKB PC5	0.86	0.91	1.04	0.0%

Table B.8: Independence of UK Biobank and ancient Eurasian scans for selection. The UK Biobank and ancient Eurasian selection statistics were not substantially inflated genome-wide, nor at the overlapping SNPs in both datasets. Similarly, the combined selection statistics were not substantially inflated. The correlation between the two statistics is also small, with the UK Biobank PC1 and ancient Eurasian statistics being most correlated with $r = 0.188$.

Locus	Chr	Locus (Mb)	Phenotype	Top SNP	p-value
LCT	2	134.9 - 137.0	lung_FVCzSMOKE	rs6716536	4.90e-10
TLR1	4	38.8 - 38.9	disease_ALLERGY_ECZEMA_DIAGNOSED	rs5743614	5.80e-26
SLC22A4	5	131.4 - 131.8	body_HEIGHTz	rs1050152	3.70e-18
			disease_ASTHMA_DIAGNOSED	rs2188962	6.00e-9
F12	5	176.8 - 176.8	body_HEIGHTz	rs2545801	4.80e-11
HLA	6	28.3 - 33.0	body_HEIGHTz	rs2256183	4.10e-23
			body_WHRadjBMIz	rs521977	1.80e-10
			bp_DIASTOLICadjMEDz		2.00e-10
			disease_ALLERGY_ECZEMA_DIAGNOSED	rs3135377	1.60e-11
			disease_ASTHMA_DIAGNOSED	rs204993	5.10e-9
			lung_FEV1FVCzSMOKE	rs3891175	3.50e-21
			lung_FVCzSMOKE	rs6456834	3.70e-9
FADS1	11	61.5 - 61.6	bmd_HEEL_TSCOREz	rs174548	1.20e-8
ATXN2/ SH2B3	12	111.9 - 112.9	bp_DIASTOLICadjMEDz	rs3184504	8.00e-33
			bp_SYSTOLICadjMEDz		1.90e-13
			disease_HYPERTENSION_DIAGNOSED		1.30e-9
CYP1A2/ CSK	15	75.0 - 75.1	bp_DIASTOLICadjMEDz	rs2472304	1.10e-19
			bp_SYSTOLICadjMEDz		4.20e-10
			disease_HYPERTENSION_DIAGNOSED		2.60e-9

Table B.9: Phenotype associations at SNPs with signals of selection. We tested SNPs with genome-wide significant signals of selection in the constituent UK Biobank or ancient Eurasian scans or the combined scan for association with 15 phenotypes in the UK Biobank data set, using the top 5 PCs as covariates.

Phenotype	PC1	PC2	PC3	PC4	PC5
bmd_HEEL_TSCOREz	2.33e-1	4.75e-1	2.39e-23	3.54e-7	2.97e-1
body_BMIz	2.36e-27	7.98e-1	2.59e-11	2.21e-3	1.14e-1
body_HEIGHTz	<1e-50	4.66e-1	<1e-50	8.38e-3	2.41e-3
body_WHRadjBMIz	2.75e-6	3.35e-1	3.71e-3	2.94e-24	1.06e-1
bp_DIASTOLICadjMEDz	7.34e-1	5.42e-1	2.70e-8	6.16e-13	2.37e-1
bp_SYSTOLICadjMEDz	1.75e-3	7.67e-2	6.31e-1	6.15e-7	3.09e-2
cov_EDU_COLLEGE	6.73e-1	9.75e-25	2.01e-29	8.59e-36	7.96e-7
cov_SMOKING_STATUS	7.13e-1	5.67e-1	3.61e-3	9.63e-7	8.08e-1
disease_ALLERGY_ECZEMA_DIAGNOSED	1.76e-16	1.50e-3	1.07e-8	7.15e-3	7.87e-1
disease_ASTHMA_DIAGNOSED	7.04e-1	2.50e-6	6.12e-1	2.89e-5	3.84e-1
disease_HYPERTENSION_DIAGNOSED	7.84e-1	2.45e-1	1.09e-3	5.92e-1	3.32e-1
lung_FEV1FVCzSMOKE	9.85e-1	3.31e-7	1.89e-13	5.58e-10	3.43e-7
lung_FVCzSMOKE	8.69e-2	1.93e-45	2.21e-3	3.48e-40	3.99e-4
repro_MENARCHE_AGE	1.14e-4	1.11e-5	1.93e-3	3.11e-1	3.75e-2
repro_MENOPAUSE_AGE	1.46e-15	7.51e-4	9.54e-7	1.27e-3	1.93e-1

Table B.10: PC-phenotype associations in UK Biobank. We report the results of tests of associations (p-value) between top PCs and 15 phenotypes in UK Biobank. This analysis does not distinguish between environmental and genetic effects

C

Supplementary Materials for Chapter 3

C.1 Supplementary Figures

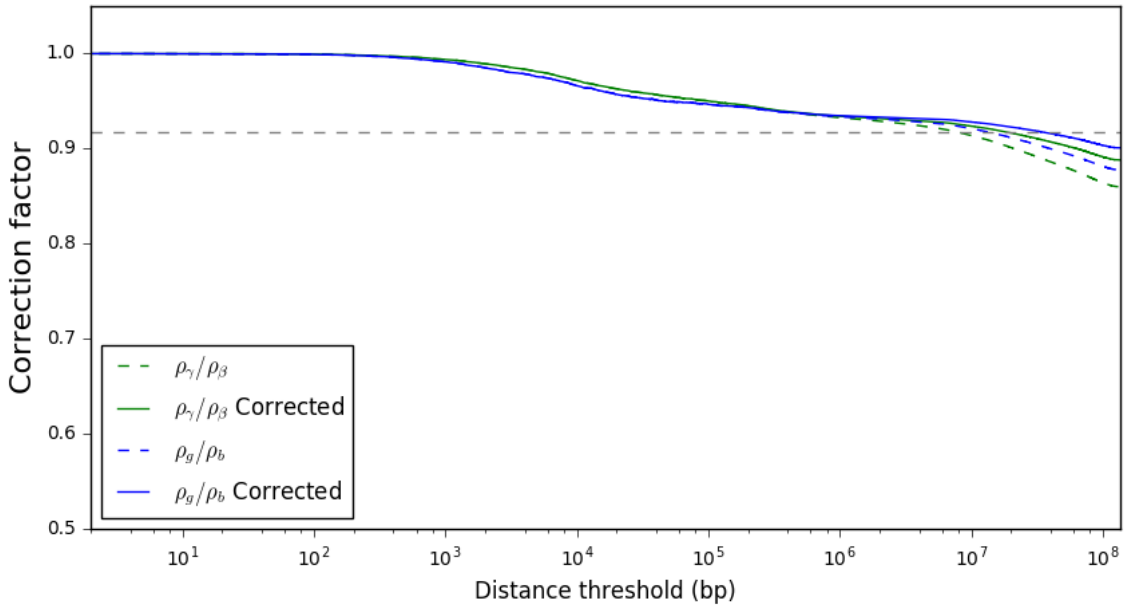

Figure C.1: Calculating LD in tighter windows alters the τ-ratio on GERA chromosome 11. The τ-ratio is sensitive to long-range noise. The dashed line is $\hat{p}_{g,REML}/\rho_b$ in simulations where the true ρ_b is known. Constraining the τ-ratio to 1MB LD windows as in LD-score regression results in a τ-ratio estimate that is close to $\hat{p}_{g,REML}/\rho_b$.

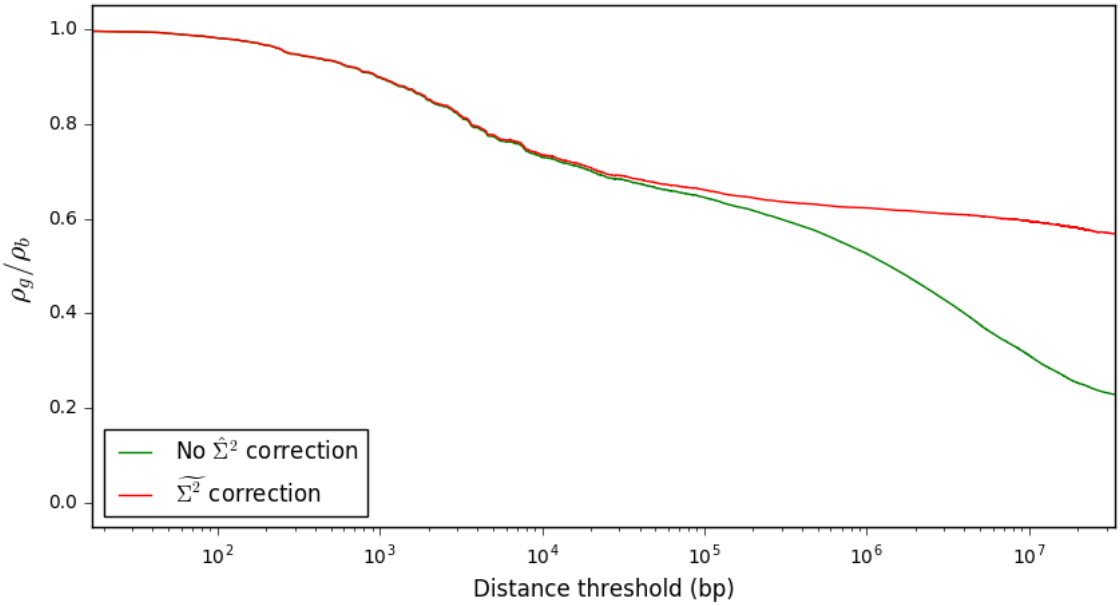

Figure C.2: Correcting the r^2 bias fixes the τ-ratio in 1000 Genomes chromosome 22. When calculating the τ-ratio using EUR and EAS populations from the 1000 Genomes project on chromosome 22, we found that correcting for bias in estimates of $\widehat{r^2}$ resulted in a plateau in estimates of the τ-ratio beyond a window size of 100kb.

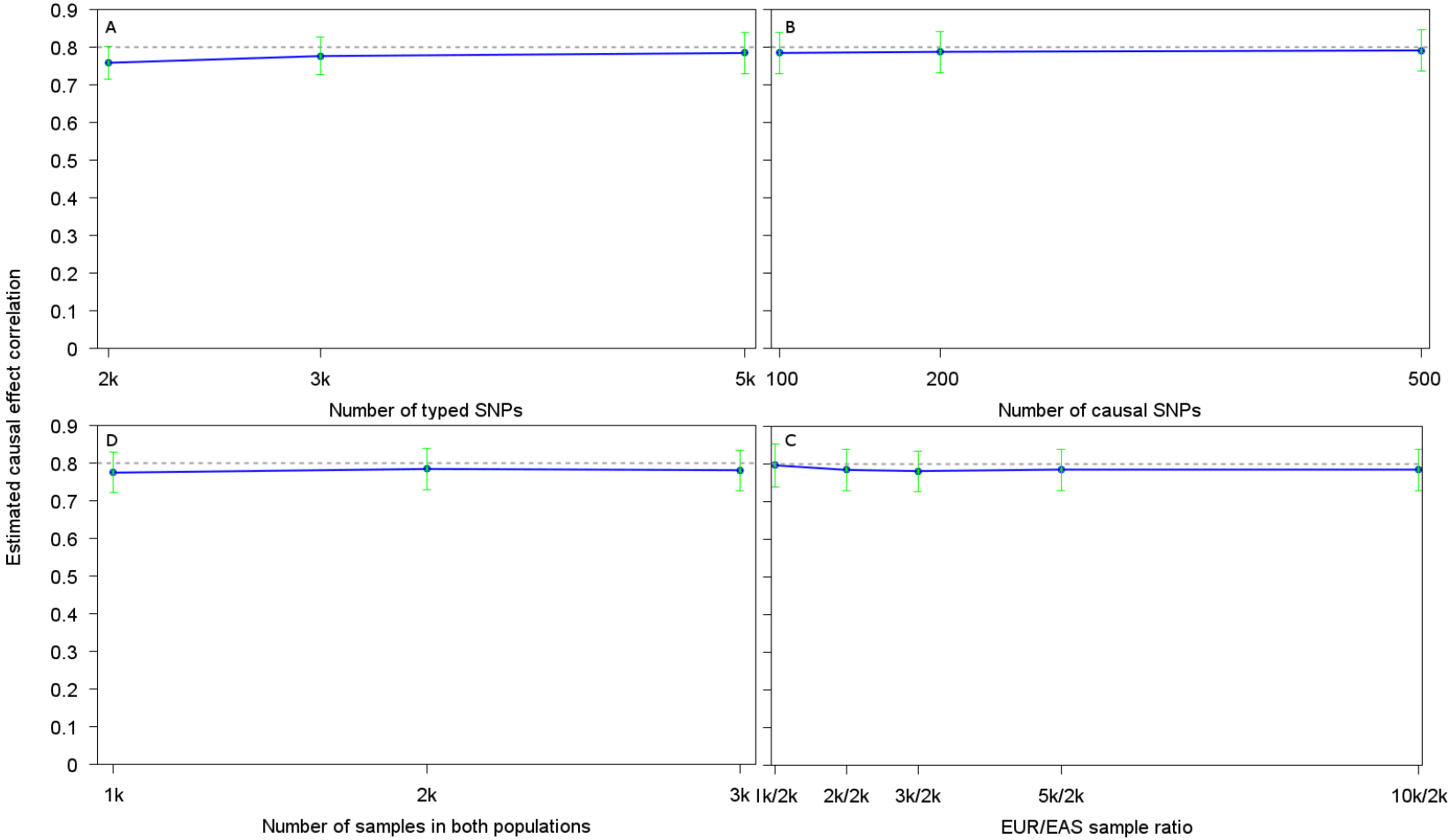

Figure C.3: Estimates of cross-population heritability are accurate when varying several parameters. As with Figure 1 and Figure 2, we performed simulations with European and East Asian samples on chromosome 11. We fixed ρ_b at 0.8 and h^2 at 0.8. Unless the quantity was being varied, the number of typed SNPs was 5k, the number of causal SNPs was 100, and we had 2k European and East Asian samples. After estimating ρ_g with GCTA [154], we applied our correction factor to estimate ρ_b. We find that this method accurately estimated ρ_b when varying the number of typed SNPs, causal SNPs and the number of samples in each population.

References

[1] Abraham, G. & Inouye, M. (2014). Fast Principal Component Analysis of Large-Scale Genome-Wide Data. PLoS ONE, 9(4), e93766.

[2] Akey, J. M., Zhang, G., Zhang, K., Jin, L., & Shriver, M. D. (2002). Interrogating a High-Density SNP Map for Signatures of Natural Selection. Genome Research, 12(12), 1805–1814.

[3] Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19(9), 1655–1664.

[4] Armitage, S. J., Jasim, S. A., Marks, A. E., Parker, A. G., Usik, V. I., & Uerpmann, H.-P. (2011). The Southern Route "Out of Africa": Evidence for an Early Expansion of Modern Humans into Arabia. Science, 331(6016), 453–456.

[5] Ayodo, G., Price, A. L., Keinan, A., Ajwang, A., Otieno, M. F., Orago, A. S. S., Patterson, N., & Reich, D. (2007). Combining Evidence of Natural Selection with Association Analysis Increases Power to Detect Malaria-Resistance Variants. The American Journal of Human Genetics, 81(2), 234–242.

[6] Banda, Y., Kvale, M. N., Hoffmann, T. J., Hesselson, S. E., Ranatunga, D., Tang, H., Sabatti, C., Croen, L. A., Dispensa, B. P., Henderson, M., Iribarren, C., Jorgenson, E., Kushi, L. H., Ludwig, D., Olberg, D., Quesenberry, C. P., Rowell, S., Sadler, M., Sakoda, L. C., Sciortino, S., Shen, L., Smethurst, D., Somkin, C. P., Eeden, S. K. V. D., Walter, L., Whitmer, R. A., Kwok, P.-Y., Schaefer, C., & Risch, N. (2015). Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics, 200(4), 1285–1295.

[7] Baran, Y. & Halperin, E. (2015). A Note on the Relations Between Spatio-Genetic Models. Journal of Computational Biology, 22(10), 905–917.

[8] Baran, Y., Quintela, I., Carracedo, n., Pasaniuc, B., & Halperin, E. (2013). Enhanced Localization of Genetic Samples through Linkage-Disequilibrium Correction. The American Journal of Human Genetics, 92(6), 882–894.

[9] Beaumont, M. A. & Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. Molecular Ecology, 13(4), 969–980.

[10] Berg, J. J. & Coop, G. (2014). A Population Genetic Signal of Polygenic Adaptation. PLOS Genet, 10(8), e1004412.

[11] Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., & Hirschhorn, J. N. (2004). Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. The American Journal of Human Genetics, 74(6), 1111–1120.

[12] Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., Palmer, C. D., Adeyemo, A. A., Akylbekova, E. L., Cupples, L. A., Divers, J., Fornage, M., Kao, W. H. L., Lange, L., Li, M., Musani, S., Mychaleckyj, J. C., Ogunniyi, A., Papanicolaou, G., Rotimi, C. N., Rotter, J. I., Ruczinski, I., Salako, B., Siscovick, D. S., Tayo, B. O., Yang, Q., McCarroll, S., Sabeti, P., Lettre, G., De Jager, P., Hirschhorn, J., Zhu, X., Cooper, R., Reich, D., Wilson, J. G., & Price, A. L. (2011). Genome-wide Comparison of African-Ancestry Populations from CARe and Other Cohorts Reveals Signals of Natural Selection. The American Journal of Human Genetics, 89(3), 368–381.

[13] Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. Genome Research, 23(9), 1514–1521.

[14] Bhatia, G., Tandon, A., Patterson, N., Aldrich, M. C., Ambrosone, C. B., Amos, C., Bandera, E. V., Berndt, S. I., Bernstein, L., Blot, W. J., Bock, C. H., Caporaso, N., Casey, G., Deming, S. L., Diver, W. R., Gapstur, S. M., Gillanders, E. M., Harris, C. C., Henderson, B. E., Ingles, S. A., Isaacs, W., De Jager, P. L., John, E. M., Kittles, R. A., Larkin, E., McNeill, L. H., Millikan, R. C., Murphy, A., Neslund-Dudas, C., Nyante, S., Press, M. F., Rodriguez-Gil, J. L., Rybicki, B. A., Schwartz, A. G., Signorello, L. B., Spitz, M., Strom, S. S., Tucker, M. A., Wiencke, J. K., Witte, J. S., Wu, X., Yamamura, Y., Zanetti, K. A., Zheng, W., Ziegler, R. G., Chanock, S. J., Haiman, C. A., Reich, D., & Price, A. L. (2014). Genome-wide Scan of 29,141 African Americans Finds No Evidence of Directional Selection since Admixture. The American Journal of Human Genetics, 95(4), 437–444.

[15] Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J. M., Mei, R., Scherer, S. W., Julian, C. G., Wilson, M. J., López Herráez, D., Brutsaert, T., Parra, E. J., Moore, L. G., & Shriver, M. D. (2010). Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data. PLoS Genet, 6(9).

[16] Billingsley, P. (1995). Probability and Measure. New York: Wiley-Interscience, 3 edition edition.

[17] Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., & SanCristobal, M. (2010). Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended. Genetics, 186(1), 241–262.

[18] Brown, B. C., Ye, C. J., Price, A. L., & Zaitlen, N. (2016). Transethnic Genetic-Correlation Estimates from Summary Statistics. The American Journal of Human Genetics, 99(1), 76–88.

[19] Buck, D., Albrecht, E., Aslam, M., Goris, A., Hauenstein, N., Jochim, A., International Multiple Sclerosis Genetics Consortium, W. T. C. C. C., Cepok, S., Grummel, V., Dubois, B., Berthele, A., Lichtner, P., Gieger, C., Winkelmann, J., & Hemmer, B. (2013). Genetic variants in the immunoglobulin heavy chain locus are associated with the IgG index in multiple sclerosis. Annals of Neurology, 73(1), 86–94.

[20] Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R., Patterson, N., Robinson, E. B., Daly, M. J., Price, A. L., & Neale, B. M. (2015a). An Atlas of Genetic Correlations across Human Diseases and Traits. Nature genetics, 47(11), 1236–1241.

[21] Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M. J., Price, A. L., & Neale, B. M. (2015b). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics, 47(3), 291–295.

[22] Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., Todd, J. A., Donnelly, P., Barrett, J. C., Davison, D., Easton, D., Evans, D., Leung, H.-T., Marchini, J. L., Morris, A. P., Spencer, C. C. A., Tobin, M. D., Attwood, A. P., Boorman, J. P., Cant, B., Everson, U., Hussey, J. M., Jolley, J. D., Knight, A. S., Koch, K., Meech, E., Nutland, S., Prowse, C. V., Stevens, H. E., Taylor, N. C., Walters, G. R., Walker, N. M., Watkins, N. A., Winzer, T., Jones, R. W., McArdle, W. L., Ring, S. M., Strachan, D. P., Pembrey, M., Breen, G., Clair, D. S., Caesar, S., Gordon-Smith, K., Jones, L., Fraser, C., Green, E. K., Grozeva, D., Hamshere, M. L., Holmans, P. A., Jones, I. R., Kirov, G., Moskvina, V., Nikolov, I., O'Donovan, M. C., Owen, M. J., Collier, D. A., Elkin, A., Farmer, A., Williamson, R., McGuffin, P., Young, A. H., Ferrier, I. N., Ball, S. G., Balmforth, A. J., Barrett, J. H., Bishop, D. T., Iles, M. M., Maqbool, A., Yuldasheva, N., Hall, A. S.,

Braund, P. S., Dixon, R. J., Mangino, M., Stevens, S., Thompson, J. R., Bredin, F., Tremelling, M., Parkes, M., Drummond, H., Lees, C. W., Nimmo, E. R., Satsangi, J., Fisher, S. A., Forbes, A., Lewis, C. M., Onnie, C. M., Prescott, N. J., Sanderson, J., Mathew, C. G., Barbour, J., Mohiuddin, M. K., Todhunter, C. E., Mansfield, J. C., Ahmad, T., Cummings, F. R., Jewell, D. P., Webster, J., Brown, M. J., Lathrop, G. M., Connell, J., Dominiczak, A., Marcano, C. A. B., Burke, B., Dobson, R., Gungadoo, J., Lee, K. L., Munroe, P. B., Newhouse, S. J., Onipinla, A., Wallace, C., Xue, M., Caulfield, M., Farrall, M., Barton, A., (braggs), T. B. i. R. G. a. G., Bruce, I. N., Donovan, H., Eyre, S., Gilbert, P. D., Hider, S. L., Hinks, A. M., John, S. L., Potter, C., Silman, A. J., Symmons, D. P. M., Thomson, W., Worthington, J., Dunger, D. B., Widmer, B., Frayling, T. M., Freathy, R. M., Lango, H., Perry, J. R. B., Shields, B. M., Weedon, M. N., Hattersley, A. T., Hitman, G. A., Walker, M., Elliott, K. S., Groves, C. J., Lindgren, C. M., Rayner, N. W., Timpson, N. J., Zeggini, E., Newport, M., Sirugo, G., Lyons, E., Vannberg, F., Hill, A. V. S., Bradbury, L. A., Farrar, C., Pointon, J. J., Wordsworth, P., Brown, M. A., Franklyn, J. A., Heward, J. M., Simmonds, M. J., Gough, S. C. L., Seal, S., (uk), B. C. S. C., Stratton, M. R., Rahman, N., Ban, M., Goris, A., Sawcer, S. J., Compston, A., Conway, D., Jallow, M., Rockett, K. A., Bumpstead, S. J., Chaney, A., Downes, K., Ghori, M. J. R., Gwilliam, R., Hunt, S. E., Inouye, M., Keniry, A., King, E., McGinnis, R., Potter, S., Ravindrarajah, R., Whittaker, P., Widden, C., Withers, D., Cardin, N. J., Ferreira, T., Pereira-Gale, J., Hallgrimsdóttir, I. B., Howie, B. N., Su, Z., Teo, Y. Y., Vukcevic, D., Bentley, D., & Compston, A. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447(7145), 661–678.

[23] Carlsson, B., Kindberg, E., Buesa, J., Rydell, G. E., Lidón, M. F., Montava, R., Mallouh, R. A., Grahn, A., Rodríguez-Díaz, J., Bellido, J., Arnedo, A., Larson, G., & Svensson, L. (2009). The G428a Nonsense Mutation in FUT2 Provides Strong but Not Absolute Protection against Symptomatic GII.4 Norovirus Infection. PLOS ONE, 4(5), e5593.

[24] Cavalli-Sforza, L. L. & Feldman, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. Nature Genetics, 33, 266–275.

[25] Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4, 7.

[26] Chen, C.-Y., Pollack, S., Hunter, D. J., Hirschhorn, J. N., Kraft, P., & Price, A. L. (2013). Improved ancestry inference using weights from external reference panels. Bioinformatics, 29(11), 1399–1406.

[27] Chen, G.-B., Lee, S. H., Zhu, Z.-X., Benyamin, B., & Robinson, M. R. (2015). EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. bioRxiv, (pp. 023457).

[28] Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E., Nutland, S., Howson, J. M. M., Faham, M., Moorhead, M., Jones, H. B., Falkowski, M., Hardenbol, P., Willis, T. D., & Todd, J. A. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. Nature Genetics, 37(11), 1243–1246.

[29] Cornelis, M. C., Monda, K. L., Yu, K., Paynter, N., Azzato, E. M., Bennett, S. N., Berndt, S. I., Boerwinkle, E., Chanock, S., Chatterjee, N., Couper, D., Curhan, G., Heiss, G., Hu, F. B., Hunter, D. J., Jacobs, K., Jensen, M. K., Kraft, P., Landi, M. T., Nettleton, J. A., Purdue, M. P., Rajaraman, P., Rimm, E. B., Rose, L. M., Rothman, N., Silverman, D., Stolzenberg-Solomon, R., Subar, A., Yeager, M., Chasman, D. I., van Dam, R. M., & Caporaso, N. E. (2011). Genome-Wide Meta-Analysis Identifies Regions on 7p21 (AHR) and 15q24 (CYP1a2) As Determinants of Habitual Caffeine Consumption. PLoS Genetics, 7(4).

[30] de Bakker, P. I. W., McVean, G., Sabeti, P. C., Miretti, M. M., Green, T., Marchini, J., Ke, X., Monsuur, A. J., Whittaker, P., Delgado, M., Morrison, J., Richardson, A., Walsh, E. C., Gao, X., Galver, L., Hart, J., Hafler, D. A., Pericak-Vance, M., Todd, J. A., Daly, M. J., Trowsdale, J., Wijmenga, C., Vyse, T. J., Beck, S., Murray, S. S., Carrington, M., Gregory, S., Deloukas, P., & Rioux, J. D. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nature Genetics, 38(10), 1166–1172.

[31] de Candia, T. R., Lee, S. H., Yang, J., Browning, B. L., Gejman, P. V., Levinson, D. F., Mowry, B. J., Hewitt, J. K., Goddard, M. E., O'Donovan, M. C., Purcell, S. M., Posthuma, D., Visscher, P. M., Wray, N. R., & Keller, M. C. (2013). Additive Genetic Variation in Schizophrenia Risk Is Shared by Populations of African and European Descent. The American Journal of Human Genetics, 93(3), 463–470.

[32] Ding, K. & Kullo, I. J. (2011). Geographic differences in allele frequencies of susceptibility SNPs for cardiovascular disease. BMC Medical Genetics, 12, 55.

[33] Duforet-Frebourg, N., Bazin, E., & Blum, M. G. B. (2014). Genome scans for detecting footprints of local adaptation using a Bayesian factor model. Molecular Biology and Evolution, (pp. msu182).

[34] Edenberg, H. J. & Foroud, T. (2013). Genetics and alcoholism. Nature Reviews Gastroenterology and Hepatology, 10(8), 487–494.

[35] Excoffier, L., Hofer, T., & Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. Heredity, 103(4), 285–298.

[36] Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., Cozzi-Lepri, A., Luca, A. D., Easterbrook, P., Francioli, P., Mallal, S., Martinez-Picado, J., Miro, J. M., Obel, N., Smith, J. P., Wyniger, J., Descombes, P., Antonarakis, S. E., Letvin, N. L., McMichael, A. J., Haynes, B. F., Telenti, A., & Goldstein, D. B. (2007). A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. Science, 317(5840), 944–947.

[37] Ferrer-Admetlla, A., Sikora, M., Laayouni, H., Esteve, A., Roubinet, F., Blancher, A., Calafell, F., Bertranpetit, J., & Casals, F. (2009). A Natural History of FUT2 Polymorphism in Humans. Molecular Biology and Evolution, 26(9), 1993–2003.

[38] Foll, M. & Gaggiotti, O. (2008). A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. Genetics, 180(2), 977–993.

[39] Foll, M., Gaggiotti, O. E., Daub, J. T., Vatsiou, A., & Excoffier, L. (2014). Widespread Signals of Convergent Adaptation to High Altitude in Asia and America. The American Journal of Human Genetics, 95(4), 394–407.

[40] Fumagalli, M., Cagliani, R., Pozzoli, U., Riva, S., Comi, G. P., Menozzi, G., Bresolin, N., & Sironi, M. (2009). Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Research, 19(2), 199–212.

[41] Fumagalli, M., Moltke, I., Grarup, N., Racimo, F., Bjerregaard, P., Jørgensen, M. E., Korneliussen, T. S., Gerbault, P., Skotte, L., Linneberg, A., Christensen, C., Brandslund, I., Jørgensen, T., Huerta-Sánchez, E., Schmidt, E. B., Pedersen, O., Hansen, T., Albrechtsen, A., & Nielsen, R. (2015). Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science, 349(6254), 1343–1347.

[42] Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., & Fabrice, R. (2009). GNU Scientific Library Reference Manual. Network Theory Limited, 3rd edition.

[43] Galinsky, K. J., Bhatia, G., Loh, P.-R., Georgiev, S., Mukherjee, S., Patterson, N. J., & Price, A. L. (2016a). Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1b in Europe and East Asia. The American Journal of Human Genetics, 98(3), 456–472.

[44] Galinsky, K. J., Loh, P.-R., Mallick, S., Patterson, N. J., & Price, A. L. (2016b). Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure. The American Journal of Human Genetics, 99(5), 1130–1139.

[45] Ganesh, S. K., Chasman, D. I., Larson, M. G., Guo, X., Verwoert, G., Bis, J. C., Gu, X., Smith, A. V., Yang, M.-L., Zhang, Y., Ehret, G., Rose, L. M., Hwang, S.-J., Papanicolau, G. J., Sijbrands, E. J., Rice, K., Eiriksdottir, G., Pihur, V., Ridker, P. M., Vasan, R. S., Newton-Cheh, C., Raffel, L. J., Amin, N., Rotter, J. I., Liu, K., Launer, L. J., Xu, M., Caulfield, M., Morrison, A. C., Johnson, A. D., Vaidya, D., Dehghan, A., Li, G., Bouchard, C., Harris, T. B., Zhang, H., Boerwinkle, E., Siscovick, D. S., Gao, W., Uitterlinden, A. G., Rivadeneira, F., Hofman, A., Willer, C. J., Franco, O. H., Huo, Y., Witteman, J. C. M., Munroe, P. B., Gudnason, V., Palmas, W., van Duijn, C., Fornage, M., Levy, D., Psaty, B. M., & Chakravarti, A. (2014). Effects of Long-Term Averaging of Quantitative Blood Pressure Traits on the Detection of Genetic Associations. The American Journal of Human Genetics, 95(1), 49–65.

[46] Gelernter, J., Kranzler, H. R., Sherva, R., Almasy, L., Koesterer, R., Smith, A. H., Anton, R., Preuss, U. W., Ridinger, M., Rujescu, D., Wodarz, N., Zill, P., Zhao, H., & Farrer, L. A. (2014). Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci. Molecular Psychiatry, 19(1), 41–49.

[47] Günther, T. & Coop, G. (2013). Robust Identification of Local Adaptation from Allele Frequencies. Genetics, 195(1), 205–220.

[48] Golub, G. H. & Van Loan, C. F. (1996). Matrix Computations. Baltimore: Johns Hopkins University Press, 3rd edition edition.

[49] Group, T. E. H. a. B. C. C. (2010). Insulin-like growth factor 1 (IGF1), IGF binding protein 3 (IGFBP3), and breast cancer risk: pooled individual data analysis of 17 prospective studies. The Lancet Oncology, 11(6), 530–542.

[50] Guerrero, J. A., Rivera, J., Quiroga, T., Martinez-Perez, A., Antón, A. I., Martínez, C., Panes, O., Vicente, V., Mezzano, D., Soria, J.-M., & Corral, J. (2011). Novel

loci involved in platelet function and platelet count identified by a genome-wide study performed in children. Haematologica, 96(9), 1335–1343.

[51] Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R. S., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., Doumatey, A. P., Asiki, G., Seeley, J., Sisay-Joof, F., Jallow, M., Tollman, S., Mekonnen, E., Ekong, R., Oljira, T., Bradman, N., Bojang, K., Ramsay, M., Adeyemo, A., Bekele, E., Motala, A., Norris, S. A., Pirie, F., Kaleebu, P., Kwiatkowski, D., Tyler-Smith, C., Rotimi, C., Zeggini, E., & Sandhu, M. S. (2015). The African Genome Variation Project shapes medical genetics in Africa. Nature, 517(7534), 327–332.

[52] Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R. G., Hallgren, F., Khartanovich, V., Khokhlov, A., Kunst, M., Kuznetsov, P., Meller, H., Mochalov, O., Moiseyev, V., Nicklisch, N., Pichler, S. L., Risch, R., Rojo Guerra, M. A., Roth, C., Szécsényi-Nagy, A., Wahl, J., Meyer, M., Krause, J., Brown, D., Anthony, D., Cooper, A., Alt, K. W., & Reich, D. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature, 522(7555), 207–211.

[53] Haiman, C. A., Chen, G. K., Blot, W. J., Strom, S. S., Berndt, S. I., Kittles, R. A., Rybicki, B. A., Isaacs, W. B., Ingles, S. A., Stanford, J. L., Diver, W. R., Witte, J. S., Chanock, S. J., Kolb, S., Signorello, L. B., Yamamura, Y., Neslund-Dudas, C., Thun, M. J., Murphy, A., Casey, G., Sheng, X., Wan, P., Pooler, L. C., Monroe, K. R., Waters, K. M., Marchand, L. L., Kolonel, L. N., Stram, D. O., & Henderson, B. E. (2011). Characterizing Genetic Risk at Known Prostate Cancer Susceptibility Loci in African Americans. PLOS Genetics, 7(5), e1001387.

[54] Halko, N., Martinsson, P., Shkolnisky, Y., & Tygert, M. (2011a). An Algorithm for the Principal Component Analysis of Large Data Sets. SIAM Journal on Scientific Computing, 33(5), 2580–2594.

[55] Halko, N., Martinsson, P., & Tropp, J. (2011b). Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. SIAM Review, 53(2), 217–288.

[56] Hamblin, M. T. & Di Rienzo, A. (2000). Detection of the Signature of Natural Selection in Humans: Evidence from the Duffy Blood Group Locus. The American Journal of Human Genetics, 66(5), 1669–1679.

[57] Han, Y., Gu, S., Oota, H., Osier, M. V., Pakstis, A. J., Speed, W. C., Kidd, J. R., & Kidd, K. K. (2007). Evidence of positive selection on a class I ADH locus. American Journal of Human Genetics, 80(3), 441–456.

[58] Hancock, A. M., Witonsky, D. B., Alkorta-Aranburu, G., Beall, C. M., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J. K., Coop, G., & Di Rienzo, A. (2011). Adaptations to Climate-Mediated Selective Pressures in Humans. PLoS Genet, 7(4), e1001375.

[59] Hazra, A., Kraft, P., Selhub, J., Giovannucci, E. L., Thomas, G., Hoover, R. N., Chanock, S. J., & Hunter, D. J. (2008). Common variants of FUT2 are associated with plasma vitamin B12 levels. Nature Genetics, 40(10), 1160–1162.

[60] He, Y., Wang, M., Huang, X., Li, R., Xu, H., Xu, S., & Jin, L. (2015). A probabilistic method for testing and estimating selection differences between populations. Genome Research, (pp. gr.192336.115).

[61] Heffelfinger, C., Pakstis, A. J., Speed, W. C., Clark, A. P., Haigh, E., Fang, R., Furtado, M. R., Kidd, K. K., & Snyder, M. P. (2014). Haplotype structure and positive selection at TLR1. European Journal of Human Genetics, 22(4), 551–557.

[62] Henn, B. M., Cavalli-Sforza, L. L., & Feldman, M. W. (2012). The great human expansion. Proceedings of the National Academy of Sciences, 109(44), 17758–17764.

[63] Hong, K.-W., Jin, H.-S., Lim, J.-E., Kim, S., Go, M. J., & Oh, B. (2010). Recapitulation of two genomewide association studies on blood pressure and essential hypertension in the Korean population. Journal of Human Genetics, 55(6), 336–341.

[64] Innocenti, F., Cooper, G. M., Stanaway, I. B., Gamazon, E. R., Smith, J. D., Mirkov, S., Ramirez, J., Liu, W., Lin, Y. S., Moloney, C., Aldred, S. F., Trinklein, N. D., Schuetz, E., Nickerson, D. A., Thummel, K. E., Rieder, M. J., Rettie, A. E., Ratain, M. J., Cox, N. J., & Brown, C. D. (2011). Identification, Replication, and Functional Fine-Mapping of Expression Quantitative Trait Loci in Primary Human Liver Tissue. PLoS Genetics, 7(5).

[65] Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H.-C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A., & Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature, 451(7181), 998–1003.

[66] Karlsson, E. K., Kwiatkowski, D. P., & Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. Nature Reviews Genetics, 15(6), 379–393.

[67] Kindberg, E., Hejdeman, B., Bratt, G., Wahren, B., Lindblom, B., Hinkula, J., & Svensson, L. (2006). A nonsense mutation (428g→A) in the fucosyltransferase FUT2 gene affects the progression of HIV-1 infection:. AIDS, 20(5), 685–689.

[68] Ko, A., Cantor, R. M., Weissglas-Volkov, D., Nikkola, E., Reddy, P. M. V. L., Sinsheimer, J. S., Pasaniuc, B., Brown, R., Alvarez, M., Rodriguez, A., Rodriguez-Guillen, R., Bautista, I. C., Arellano-Campos, O., Muñoz-Hernández, L. L., Salomaa, V., Kaprio, J., Jula, A., Jauhiainen, M., Heliövaara, M., Raitakari, O., Lehtimäki, T., Eriksson, J. G., Perola, M., Lohmueller, K. E., Matikainen, N., Taskinen, M.-R., Rodriguez-Torres, M., Riba, L., Tusie-Luna, T., Aguilar-Salinas, C. A., & Pajukanta, P. (2014). Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. Nature Communications, 5.

[69] Ko, W.-Y., Rajan, P., Gomez, F., Scheinfeldt, L., An, P., Winkler, C. A., Froment, A., Nyambo, T. B., Omar, S. A., Wambebe, C., Ranciaro, A., Hirbo, J. B., & Tishkoff, S. A. (2013). Identifying Darwinian Selection Acting on Different Human APOL1 Variants among Diverse African Populations. The American Journal of Human Genetics, 93(1), 54–66.

[70] Kwiatkowski, D. P. (2005). How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. The American Journal of Human Genetics, 77(2), 171–192.

[71] Lamason, R. L., Mohideen, M.-A. P. K., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., Jurynec, M. J., Mao, X., Humphreville, V. R., Humbert, J. E., Sinha, S., Moore, J. L., Jagadeeswaran, P., Zhao, W., Ning, G., Makalowska, I., McKeigue, P. M., O'Donnell, D., Kittles, R., Parra, E. J., Mangini, N. J., Grunwald, D. J., Shriver, M. D., Canfield, V. A., & Cheng, K. C. (2005). SLC24a5, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans. Science, 310(5755), 1782–1786.

[72] Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of Population Structure using Dense Haplotype Data. PLoS Genet, 8(1), e1002453.

[73] Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P. H., Schraiber, J. G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Fu, Q., Bos, K. I., Nordenfelt, S., Li, H., de Filippo, C., Prüfer,

K., Sawyer, S., Posth, C., Haak, W., Hallgren, F., Fornander, E., Rohland, N., Delsate, D., Francken, M., Guinet, J.-M., Wahl, J., Ayodo, G., Babiker, H. A., Bailliet, G., Balanovska, E., Balanovsky, O., Barrantes, R., Bedoya, G., Ben-Ami, H., Bene, J., Berrada, F., Bravi, C. M., Brisighelli, F., Busby, G. B. J., Cali, F., Churnosov, M., Cole, D. E. C., Corach, D., Damba, L., van Driem, G., Dryomov, S., Dugoujon, J.-M., Fedorova, S. A., Gallego Romero, I., Gubina, M., Hammer, M., Henn, B. M., Hervig, T., Hodoglugil, U., Jha, A. R., Karachanak-Yankova, S., Khusainova, R., Khusnutdinova, E., Kittles, R., Kivisild, T., Klitz, W., Kučinskas, V., Kushniarevich, A., Laredj, L., Litvinov, S., Loukidis, T., Mahley, R. W., Melegh, B., Metspalu, E., Molina, J., Mountain, J., Näkkäläjärvi, K., Nesheva, D., Nyambo, T., Osipova, L., Parik, J., Platonov, F., Posukh, O., Romano, V., Rothhammer, F., Rudan, I., Ruizbakiev, R., Sahakyan, H., Sajantila, A., Salas, A., Starikovskaya, E. B., Tarekegn, A., Toncheva, D., Turdikulova, S., Uktveryte, I., Utevska, O., Vasquez, R., Villena, M., Voevoda, M., Winkler, C. A., Yepiskoposyan, L., Zalloua, P., Zemunik, T., Cooper, A., Capelli, C., Thomas, M. G., Ruiz-Linares, A., Tishkoff, S. A., Singh, L., Thangaraj, K., Villems, R., Comas, D., Sukernik, R., Metspalu, M., Meyer, M., Eichler, E. E., Burger, J., Slatkin, M., Pääbo, S., Kelso, J., Reich, D., & Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature, 513(7518), 409–413.

[74] Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., (pgc Scz), T. S. P. G.-W. A. S. C., (isc), T. I. S. C., (mgs), T. M. G. o. S. C., Sullivan, P. F., Goddard, M. E., Keller, M. C., Visscher, P. M., & Wray, N. R. (2012a). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nature Genetics, 44(3), 247–250.

[75] Lee, S. H., Yang, J., Chen, G.-B., Ripke, S., Stahl, E. A., Hultman, C. M., Sklar, P., Visscher, P. M., Sullivan, P. F., Goddard, M. E., & Wray, N. R. (2013). Estimation of SNP Heritability from Dense Genotype Data. The American Journal of Human Genetics, 93(6), 1151–1155.

[76] Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., & Wray, N. R. (2012b). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics, 28(19), 2540–2542.

[77] Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E. C., Cunliffe, B., Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, Lawson, D. J., Falush, D., Freeman, C., Pirinen, M., Myers, S., Robinson, M., Donnelly, P., & Bodmer, W.

(2015). The fine-scale genetic structure of the British population. Nature, 519(7543), 309–314.

[78] Lewontin, R. C. & Krakauer, J. (1973). Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms. Genetics, 74(1), 175–195.

[79] Li, H., Gu, S., Han, Y., Xu, Z., Pakstis, A. J., Jin, L., Kidd, J. R., & Kidd, K. K. (2011). Diversification of the ADH1b Gene during Expansion of Modern Humans. Annals of Human Genetics, 75(4), 497–507.

[80] Li, H., Mukherjee, N., Soundararajan, U., Tárnok, Z., Barta, C., Khaliq, S., Mohyuddin, A., Kajuna, S. L. B., Mehdi, S. Q., Kidd, J. R., & Kidd, K. K. (2007). Geographically Separate Increases in the Frequency of the Derived ADH1b*47his Allele in Eastern and Western Asia. The American Journal of Human Genetics, 81(4), 842–846.

[81] Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., & Myers, R. M. (2008). Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. Science, 319(5866), 1100–1104.

[82] Loh, P., Palamara, P. F., & Price, A. L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nature Genetics, (pp. in press. http://biorxiv.org/content/early/2015/10/04/028282).

[83] Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N., & Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nature Genetics, 47(3), 284–290.

[84] Lorenzo, F. R., Huff, C., Myllymäki, M., Olenchock, B., Swierczek, S., Tashi, T., Gordeuk, V., Wuren, T., Ri-Li, G., McClain, D. A., Khan, T. M., Koul, P. A., Guchhait, P., Salama, M. E., Xing, J., Semenza, G. L., Liberzon, E., Wilson, A., Simonson, T. S., Jorde, L. B., Kaelin Jr, W. G., Koivunen, P., & Prchal, J. T. (2014). A genetic mechanism for Tibetan high-altitude adaptation. Nature Genetics, 46(9), 951–956.

[85] Mancuso, N., Rohland, N., Rand, K. A., Tandon, A., Allen, A., Quinque, D., Mallick, S., Li, H., Stram, A., Sheng, X., Kote-Jarai, Z., Easton, D. F., Eeles, R. A., the PRACTICAL Consortium, Le Marchand, L., Lubwama, A., Stram, D., Watya, S., Conti, D. V., Henderson, B., Haiman, C. A., Pasaniuc, B., & Reich, D. (2016).

The contribution of rare variation to prostate cancer heritability. Nature Genetics, 48(1), 30–35.

[86] Martin, J.-E., Broen, J. C., Carmona, F. D., Teruel, M., Simeon, C. P., Vonk, M. C., Slot, R. v. t., Rodriguez-Rodriguez, L., Vicente, E., Fonollosa, V., Ortego-Centeno, N., González-Gay, M. A., García-Hernández, F. J., Peña, P. G. d. l., Carreira, P., Voskuyl, A. E., Schuerwegh, A. J., Riel, P. L. C. M. v., Kreuter, A., Witte, T., Riemekasten, G., Airo, P., Scorza, R., Lunardi, C., Hunzelmann, N., Distler, J. H. W., Beretta, L., Laar, J. v., Chee, M. M., Worthington, J., Herrick, A., Denton, C., Tan, F. K., Arnett, F. C., Assassi, S., Fonseca, C., Mayes, M. D., Radstake, T. R. D. J., Koeleman, B. P. C., & Martin, J. (2012). Identification of CSK as a systemic sclerosis genetic risk factor through Genome Wide Association Study follow-up. Human Molecular Genetics, 21(12), 2825–2835.

[87] Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E. R., Llamas, B., Dryomov, S., Pickrell, J., Arsuaga, J. L., de Castro, J. M. B., Carbonell, E., Gerritsen, F., Khokhlov, A., Kuznetsov, P., Lozano, M., Meller, H., Mochalov, O., Moiseyev, V., Guerra, M. A. R., Roodenberg, J., Vergès, J. M., Krause, J., Cooper, A., Alt, K. W., Brown, D., Anthony, D., Lalueza-Fox, C., Haak, W., Pinhasi, R., & Reich, D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature, 528(7583), 499–503.

[88] Mathieson, I. & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nature Genetics, 44(3), 243–246.

[89] McGovern, D. P. B., Jones, M. R., Taylor, K. D., Marciante, K., Yan, X., Dubinsky, M., Ippoliti, A., Vasiliauskas, E., Berel, D., Derkowski, C., Dutridge, D., Consortium, I. I. G., Fleshner, P., Shih, D. Q., Melmed, G., Mengesha, E., King, L., Pressman, S., Haritunians, T., Guo, X., Targan, S. R., & Rotter, J. I. (2010). Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Human Molecular Genetics, 19(17), 3468–3476.

[90] Mellars, P. (2006). Going East: New Genetic and Archaeological Perspectives on the Modern Human Colonization of Eurasia. Science, 313(5788), 796–800.

[91] Moreno-Estrada, A., Gignoux, C. R., Fernández-López, J. C., Zakharia, F., Sikora, M., Contreras, A. V., Acuña-Alonzo, V., Sandoval, K., Eng, C., Romero-Hidalgo, S., Ortiz-Tello, P., Robles, V., Kenny, E. E., Nuño-Arana, I., Barquera-Lozano, R., Macín-Pérez, G., Granados-Arriola, J., Huntsman, S., Galanter, J. M., Via,

M., Ford, J. G., Chapela, R., Rodriguez-Cintron, W., Rodríguez-Santana, J. R., Romieu, I., Sienra-Monge, J. J., Navarro, B. d. R., London, S. J., Ruiz-Linares, A., Garcia-Herrera, R., Estrada, K., Hidalgo-Miranda, A., Jimenez-Sanchez, G., Carnevale, A., Soberón, X., Canizales-Quinteros, S., Rangel-Villalobos, H., Silva-Zolezzi, I., Burchard, E. G., & Bustamante, C. D. (2014). The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science, 344(6189), 1280–1285.

[92] Nelson, M. R., Bryc, K., King, K. S., Indap, A., Boyko, A. R., Novembre, J., Briley, L. P., Maruyama, Y., Waterworth, D. M., Waeber, G., Vollenweider, P., Oksenberg, J. R., Hauser, S. L., Stirnadel, H. A., Kooner, J. S., Chambers, J. C., Jones, B., Mooser, V., Bustamante, C. D., Roses, A. D., Burns, D. K., Ehm, M. G., & Lai, E. H. (2008). The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research. The American Journal of Human Genetics, 83(3), 347–358.

[93] Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M. D., Bochud, M., Coin, L., Najjar, S. S., Zhao, J. H., Heath, S. C., Eyheramendy, S., Papadakis, K., Voight, B. F., Scott, L. J., Zhang, F., Farrall, M., Tanaka, T., Wallace, C., Chambers, J. C., Khaw, K.-T., Nilsson, P., van der Harst, P., Polidoro, S., Grobbee, D. E., Onland-Moret, N. C., Bots, M. L., Wain, L. V., Elliott, K. S., Teumer, A., Luan, J., Lucas, G., Kuusisto, J., Burton, P. R., Hadley, D., McArdle, W. L., Wellcome Trust Case Control Consortium, Brown, M., Dominiczak, A., Newhouse, S. J., Samani, N. J., Webster, J., Zeggini, E., Beckmann, J. S., Bergmann, S., Lim, N., Song, K., Vollenweider, P., Waeber, G., Waterworth, D. M., Yuan, X., Groop, L., Orho-Melander, M., Allione, A., Di Gregorio, A., Guarrera, S., Panico, S., Ricceri, F., Romanazzi, V., Sacerdote, C., Vineis, P., Barroso, I., Sandhu, M. S., Luben, R. N., Crawford, G. J., Jousilahti, P., Perola, M., Boehnke, M., Bonnycastle, L. L., Collins, F. S., Jackson, A. U., Mohlke, K. L., Stringham, H. M., Valle, T. T., Willer, C. J., Bergman, R. N., Morken, M. A., Döring, A., Gieger, C., Illig, T., Meitinger, T., Org, E., Pfeufer, A., Wichmann, H. E., Kathiresan, S., Marrugat, J., O'Donnell, C. J., Schwartz, S. M., Siscovick, D. S., Subirana, I., Freimer, N. B., Hartikainen, A.-L., McCarthy, M. I., O'Reilly, P. F., Peltonen, L., Pouta, A., de Jong, P. E., Snieder, H., van Gilst, W. H., Clarke, R., Goel, A., Hamsten, A., Peden, J. F., Seedorf, U., Syvänen, A.-C., Tognoni, G., Lakatta, E. G., Sanna, S., Scheet, P., Schlessinger, D., Scuteri, A., Dörr, M., Ernst, F., Felix, S. B., Homuth, G., Lorbeer, R., Reffelmann, T., Rettig, R., Völker, U., Galan, P., Gut, I. G., Hercberg, S., Lathrop, G. M., Zelenika, D., Deloukas, P., Soranzo, N., Williams, F. M., Zhai, G., Salomaa, V., Laakso, M., Elosua, R., Forouhi, N. G., Völzke, H., Uiterwaal, C. S., van der

Schouw, Y. T., Numans, M. E., Matullo, G., Navis, G., Berglund, G., Bingham, S. A., Kooner, J. S., Connell, J. M., Bandinelli, S., Ferrucci, L., Watkins, H., Spector, T. D., Tuomilehto, J., Altshuler, D., Strachan, D. P., Laan, M., Meneton, P., Wareham, N. J., Uda, M., Jarvelin, M.-R., Mooser, V., Melander, O., Loos, R. J., Elliott, P., Abecasis, G. R., Caulfield, M., & Munroe, P. B. (2009). Genome-wide association study identifies eight loci associated with blood pressure. Nature Genetics, 41(6), 666–676.

[94] Nicholson, G., Smith, A. V., Jónsson, F., Gústafsson, m., Stefánsson, K., & Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(4), 695–715.

[95] Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., & Clark, A. G. (2007). Recent and ongoing selection in the human genome. Nature Reviews Genetics, 8(11), 857–868.

[96] Novembre, J. & Di Rienzo, A. (2009). Spatial patterns of variation due to natural selection in humans. Nature Reviews Genetics, 10(11), 745–755.

[97] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., & Bustamante, C. D. (2008). Genes mirror geography within Europe. Nature, 456(7218), 98–101.

[98] Novembre, J. & Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. Nature Genetics, 40(5), 646–649.

[99] O'Connell, J. R., Sharp, K., Delaneau, O., & Marchini, J. (2016). Haplotype estimation for biobank scale datasets. Nature Genetics, (pp. accepted in principle.).

[100] Osier, M. V., Pakstis, A. J., Soodyall, H., Comas, D., Goldman, D., Odunsi, A., Okonofua, F., Parnas, J., Schulz, L. O., Bertranpetit, J., Bonne-Tamir, B., Lu, R.-B., Kidd, J. R., & Kidd, K. K. (2002). A Global Perspective on Genetic Variation at the ADH Genes Reveals Unusual Patterns of Linkage Disequilibrium and Diversity. The American Journal of Human Genetics, 71(1), 84–99.

[101] Parmar, A. S., Alakulppi, N., Paavola-Sakki, P., Kurppa, K., Halme, L., Färkkilä, M., Turunen, U., Lappalainen, M., Kontula, K., Kaukinen, K., Mäki, M., Lindfors, K., Partanen, J., Sistonen, P., Mättö, J., Wacklin, P., Saavalainen, P., & Einarsdottir, E. (2012). Association study of FUT2 (rs601338) with celiac disease and inflammatory bowel disease in the Finnish population. Tissue Antigens, 80(6), 488–493.

[102] Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient Admixture in Human History. Genetics, 192(3), 1065–1093.

[103] Patterson, N., Price, A. L., & Reich, D. (2006). Population Structure and Eigenanalysis. PLoS Genet, 2(12), e190.

[104] Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C., & Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. Nature Genetics, 39(10), 1256–1260.

[105] Peter, B. M., Huerta-Sanchez, E., & Nielsen, R. (2012). Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation. PLoS Genet, 8(10), e1003011.

[106] Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., & Pritchard, J. K. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Research, 19(5), 826–837.

[107] Pickrell, J. K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., Kure, B., Mpoloka, S. W., Nakagawa, H., Naumann, C., Lipson, M., Loh, P.-R., Lachance, J., Mountain, J., Bustamante, C. D., Berger, B., Tishkoff, S. A., Henn, B. M., Stoneking, M., Reich, D., & Pakendorf, B. (2012). The genetic prehistory of southern Africa. Nature Communications, 3, 1143.

[108] Pickrell, J. K. & Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLOS Genet, 8(11), e1002967.

[109] Popat, R., Van Den Eeden, S., Tanner, C., Kamel, F., Umbach, D. M., Marder, K., Mayeux, R., Ritz, B., Ross, G. W., Petrovitch, H., Topol, B., McGuire, V., Costello, S., Manthripragada, A., Southwick, A., Myers, R., & Nelson, L. M. (2011). Coffee, ADORA2a, and CYP1a2: the caffeine connection in Parkinson's disease. European journal of neurology : the official journal of the European Federation of Neurological Societies, 18(5), 756–765.

[110] Price, A. L., Butler, J., Patterson, N., Capelli, C., Pascali, V. L., Scarnicci, F., Ruiz-Linares, A., Groop, L., Saetta, A. A., Korkolopoulou, P., Seligsohn, U., Waliszewska, A., Schirmer, C., Ardlie, K., Ramos, A., Nemesh, J., Arbeitman, L., Goldstein, D. B., Reich, D., & Hirschhorn, J. N. (2008). Discerning the Ancestry of European Americans in Genetic Association Studies. PLoS Genet, 4(1), e236.

[111] Price, A. L., Helgason, A., Palsson, S., Stefansson, H., St. Clair, D., Andreassen, O. A., Reich, D., Kong, A., & Stefansson, K. (2009). The Impact of Divergence Time on the Nature of Population Structure: An Example from Iceland. PLoS Genet, 5(6), e1000505.

[112] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38(8), 904–909.

[113] Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics, 11(7), 459–463.

[114] Pritchard, J. K. & Di Rienzo, A. (2010). Adaptation – not by sweeps alone. Nature Reviews Genetics, 11(10), 665–667.

[115] Pritchard, J. K., Pickrell, J. K., & Coop, G. (2010). The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Current Biology, 20(4), R208–R215.

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

[117] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics, 81(3), 559–575.

[118] Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America, 102(44), 15942–15947.

[119] Renné, T., Schmaier, A. H., Nickel, K. F., Blombäck, M., & Maas, C. (2012). In vivo roles of factor XII. Blood, 120(22), 4296–4303.

[120] Robinson, M. R., Hemani, G., Medina-Gomez, C., Mezzavilla, M., Esko, T., Shakhbazov, K., Powell, J. E., Vinkhuyzen, A., Berndt, S. I., Gustafsson, S., Justice, A. E., Kahali, B., Locke, A. E., Pers, T. H., Vedantam, S., Wood, A. R., van Rheenen, W., Andreassen, O. A., Gasparini, P., Metspalu, A., Berg, L. H. v. d., Veldink, J. H., Rivadeneira, F., Werge, T. M., Abecasis, G. R., Boomsma, D. I.,

Chasman, D. I., de Geus, E. J. C., Frayling, T. M., Hirschhorn, J. N., Hottenga, J. J., Ingelsson, E., Loos, R. J. F., Magnusson, P. K. E., Martin, N. G., Montgomery, G. W., North, K. E., Pedersen, N. L., Spector, T. D., Speliotes, E. K., Goddard, M. E., Yang, J., & Visscher, P. M. (2015). Population genetic differentiation of height and body mass index across Europe. Nature Genetics, 47(11), 1357–1362.

[121] Rokhlin, V., Szlam, A., & Tygert, M. (2009). A Randomized Algorithm for Principal Component Analysis. SIAM Journal on Matrix Analysis and Applications, 31(3), 1100–1124.

[122] Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., & Boehnke, M. (2010). Genome-wide association studies in diverse populations. Nature reviews. Genetics, 11(5), 356–366.

[123] Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature, 419(6909), 832–837.

[124] Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., & Lander, E. S. (2006). Positive Natural Selection in the Human Lineage. Science, 312(5780), 1614–1620.

[125] Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Xue, H., Wong, J. T.-F., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallée, C.,

Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Song, Y. Q., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., Bakker, P. I. W. d., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Sham, P. C., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Johnson, T. A., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Yakub, I., Birren, B. W., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., & Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature, 449(7164), 913–918.

[126] Scheinfeldt, L. B. & Tishkoff, S. A. (2013). Recent human adaptation: genomic approaches, interpretation and insights. Nature Reviews Genetics, 14(10), 692–702.

[127] Schiffels, S., Haak, W., Paajanen, P., Llamas, B., Popescu, E., Loe, L., Clarke, R., Lyons, A., Mortimer, R., Sayer, D., Tyler-Smith, C., Cooper, A., & Durbin, R. (2016). Iron Age and Anglo-Saxon genomes from East England reveal British migration history. Nature Communications, 7, 10408.

[128] Schlebusch, C. M., Skoglund, P., Sjödin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., Jongh, M. D., Singleton, A., Blum, M. G. B., Soodyall, H., & Jakobsson,

M. (2012). Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. Science, 338(6105), 374–379.

[129] Seldin, M. F. & Price, A. L. (2008). Application of Ancestry Informative Markers to Association Studies in European Americans. PLoS Genet, 4(1), e5.

[130] Shriver, M. D., Kennedy, G. C., Parra, E. J., Lawson, H. A., Sonpar, V., Huang, J., Akey, J. M., & Jones, K. W. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Human Genomics, 1(4), 274.

[131] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (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.

[132] Suo, C., Xu, H., Khor, C.-C., Ong, R. T., Sim, X., Chen, J., Tay, W.-T., Sim, K.-S., Zeng, Y.-X., Zhang, X., Liu, J., Tai, E.-S., Wong, T.-Y., Chia, K.-S., & Teo, Y.-Y. (2012). Natural positive selection and north–south genetic diversity in East Asia. European Journal of Human Genetics, 20(1), 102–110.

[133] Sved, J. A., McRae, A. F., & Visscher, P. M. (2008). Divergence between Human Populations Estimated from Linkage Disequilibrium. The American Journal of Human Genetics, 83(6), 737–743.

[134] Tabara, Y., Kohara, K., Kita, Y., Hirawa, N., Katsuya, T., Ohkubo, T., Hiura, Y., Tajima, A., Morisaki, T., Miyata, T., Nakayama, T., Takashima, N., Nakura, J., Kawamoto, R., Takahashi, N., Hata, A., Soma, M., Imai, Y., Kokubo, Y., Okamura, T., Tomoike, H., Iwai, N., Ogihara, T., Inoue, I., Tokunaga, K., Johnson, T., Caulfield, M., Consortium, P. M. o. b. o. t. G. B. P. G., Umemura, S., Ueshima, H., & Miki, T. (2010). Common Variants in the ATP2b1 Gene Are Associated With Susceptibility to Hypertension The Japanese Millennium Genome Project. Hypertension, 56(5), 973–980.

[135] Tang, W., Schwienbacher, C., Lopez, L. M., Ben-Shlomo, Y., Oudot-Mellakh, T., Johnson, A. D., Samani, N. J., Basu, S., Gögele, M., Davies, G., Lowe, G. D., Tregouet, D.-A., Tan, A., Pankow, J. S., Tenesa, A., Levy, D., Volpato, C. B., Rumley, A., Gow, A. J., Minelli, C., Yarnell, J. W., Porteous, D. J., Starr, J. M., Gallacher, J., Boerwinkle, E., Visscher, P. M., Pramstaller, P. P., Cushman, M., Emilsson, V., Plump, A. S., Matijevic, N., Morange, P.-E., Deary, I. J., Hicks, A. A., & Folsom,

A. R. (2012). Genetic Associations for Activated Partial Thromboplastin Time and Prothrombin Time, their Gene Expression Profiles, and Risk of Coronary Artery Disease. American Journal of Human Genetics, 91(1), 152–162.

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

[137] The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74.

[138] The International Consortium for Blood Pressure Genome-Wide Association Studies (2011). Genetic Variants in Novel Pathways Influence Blood Pressure and Cardiovascular Disease Risk. Nature, 478(7367), 103–109.

[139] The UK10K Consortium (2015). The UK10k project identifies rare variants in health and disease. Nature, 526(7571), 82–90.

[140] Thorven, M., Grahn, A., Hedlund, K.-O., Johansson, H., Wahlfrid, C., Larson, G., & Svensson, L. (2005). A Homozygous Nonsense Mutation (428g→A) in the Human Secretor (FUT2) Gene Provides Resistance to Symptomatic Norovirus (GGII) Infections. Journal of Virology, 79(24), 15351–15355.

[141] Tian, C., Plenge, R. M., Ransom, M., Lee, A., Villoslada, P., Selmi, C., Klareskog, L., Pulver, A. E., Qi, L., Gregersen, P. K., & Seldin, M. F. (2008). Analysis and Application of European Genetic Substructure Using 300 K SNP Information. PLoS Genet, 4(1), e4.

[142] Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., Smith, M. W., Thera, M. A., Wambebe, C., Weber, J. L., & Williams, S. M. (2009). The Genetic Structure and History of Africans and African Americans. Science, 324(5930), 1035–1044.

[143] Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., & Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nature Genetics, 39(1), 31–40.

[144] Tong, M., McHardy, I., Ruegger, P., Goudarzi, M., Kashyap, P. C., Haritunians, T., Li, X., Graeber, T. G., Schwager, E., Huttenhower, C., Fornace, A. J., Sonnenburg,

J. L., McGovern, D. P., Borneman, J., & Braun, J. (2014). Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism. The ISME Journal, 8(11), 2193–2206.

[145] Treutlein, J., Frank, J., Kiefer, F., & Rietschel, M. (2014). ADH1b Arg48his Allele Frequency Map: Filling in the Gap for Central Europe. Biological Psychiatry, 75(10), e15.

[146] Turchin, M. C., Chiang, C. W., Palmer, C. D., Sankararaman, S., Reich, D., & Hirschhorn, J. N. (2012). Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nature Genetics, 44(9), 1015–1019.

[147] Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A Map of Recent Positive Selection in the Human Genome. PLoS Biol, 4(3), e72.

[148] Weir, B. S. & Hill, W. G. (2002). Estimating F-statistics. Annual Review of Genetics, 36, 721–750. WOS:000180365100024.

[149] Whitfield, J. B. (2002). Alcohol Dehydrogenase and Alcohol Dependence: Variation in Genotype-Associated Risk between Populations. The American Journal of Human Genetics, 71(5), 1247–1250.

[150] Wooding, S. P., Watkins, W. S., Bamshad, M. J., Dunn, D. M., Weiss, R. B., & Jorde, L. B. (2002). DNA Sequence Variation in a 3.7-kb Noncoding Sequence 5 of the CYP1a2 Gene: Implications for Human Population History and Natural Selection. The American Journal of Human Genetics, 71(3), 528–542.

[151] Xie, Q., Ratnasinghe, L. D., Hong, H., Perkins, R., Tang, Z.-Z., Hu, N., Taylor, P. R., & Tong, W. (2005). Decision Forest Analysis of 61 Single Nucleotide Polymorphisms in a Case-Control Study of Esophageal Cancer; a novel method. BMC Bioinformatics, 6(2), 1–9.

[152] Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., Gong, X., Wang, H., Shen, Y., Pan, X., He, Y., Yang, Y., Wang, Y., Fu, W., An, Y., Wang, J., Tan, J., Qian, J., Chen, X., Zhang, X., Sun, Y., Zhang, X., Wu, B., & Jin, L. (2009). Genomic Dissection of Population Substructure of Han Chinese and Its Implication in Association Studies. The American Journal of Human Genetics, 85(6), 762–774.

[153] Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. Nature Genetics, 42(7), 565–569.

[154] Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011a). GCTA: A Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics, 88(1), 76–82.

[155] Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G., Hill, W. G., Landi, M. T., Alonso, A., Lettre, G., Lin, P., Ling, H., Lowe, W., Mathias, R. A., Melbye, M., Pugh, E., Cornelis, M. C., Weir, B. S., Goddard, M. E., & Visscher, P. M. (2011b). Genome partitioning of genetic variation for complex traits using common SNPs. Nature Genetics, 43(6), 519–525.

[156] Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., & Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nature Genetics, 46(2), 100–106.

[157] Yang, W.-Y., Novembre, J., Eskin, E., & Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. Nature Genetics, 44(6), 725–731.

[158] Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., Zou, J., Shan, Y., Li, S., Yang, Q., Asan, Ni, P., Tian, G., Xu, J., Liu, X., Jiang, T., Wu, R., Zhou, G., Tang, M., Qin, J., Wang, T., Feng, S., Li, G., Huasang, Luosang, J., Wang, W., Chen, F., Wang, Y., Zheng, X., Li, Z., Bianba, Z., Yang, G., Wang, X., Tang, S., Gao, G., Chen, Y., Luo, Z., Gusang, L., Cao, Z., Zhang, Q., Ouyang, W., Ren, X., Liang, H., Zheng, H., Huang, Y., Li, J., Bolund, L., Kristiansen, K., Li, Y., Zhang, Y., Zhang, X., Li, R., Li, S., Yang, H., Nielsen, R., Wang, J., & Wang, J. (2010). Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science, 329(5987), 75–78.

[159] Zhu, X., Feng, T., Tayo, B. O., Liang, J., Young, J. H., Franceschini, N., Smith, J. A., Yanek, L. R., Sun, Y. V., Edwards, T. L., Chen, W., Nalls, M., Fox, E., Sale, M., Bottinger, E., Rotimi, C., Liu, Y., McKnight, B., Liu, K., Arnett, D. K., Chakravati, A., Cooper, R. S., & Redline, S. (2015). Meta-analysis of Correlated Traits via Summary Statistics from GWASs with an Application in Hypertension. The American Journal of Human Genetics, 96(1), 21–36.

[160] Zou, F., Lee, S., Knowles, M. R., & Wright, F. A. (2010). Quantification of Population Structure Using Correlated SNPs by Shrinkage Principal Components. Human Heredity, 70(1), 9–22.