Fast Principal-Component Analysis Reveals Convergent Evolution of *ADH1B* in Europe and East Asia

Kevin J. Galinsky,[1,2,*] Gaurav Bhatia,[2,3] Po-Ru Loh,[2,3] Stoyan Georgiev,[4] Sayan Mukherjee,[5] Nick J. Patterson,[2,6] and Alkes L. Price[1,2,3,6,*]

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 datasets, 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 associated to human disease.

Introduction

Searching for genetic variants with unusual differentiation between populations is an established approach for identifying signals of natural selection.[1–6] 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.[7–24] Examples includes genes linked to lactase persistence[9,11] (MIM: 223100), starch hydrolysis[12] (MIM: 104700), fatty acid decomposition,[24] red blood cell abundance[17] (MIM: 611783), hypoxia response[18] (MIM: 609070), alcoholism[14] (MIM: 103780), kidney disease[21] (MIM: 612551), malaria[7,13,19,23] (MIM: 611162), HIV/AIDS[16] (MIM: 609423), autoimmune disease,[20] cancer[19] (MIM: 602470), cystic fibrosis[8] (MIM: 219700), and hypertension[23] (MIM: 145500). However, the signals of selection identified thus far might represent "only the tip of the iceberg,"[25] 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.[1,3] In addition, signals of selection detected via population differentiation can flag stratified genetic variants that are susceptible to false-positive associations in genome-wide association studies.[15]

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),[6] population branch statistic (PBS),[17] and TreeSelect.[19] These ideas are derived from the Lewontin and Krakauer test[26] and its extensions to the multinomial-Dirichlet model (F-model)[27] (later incorporating a Bayesian framework,[28] hierarchical population structure,[29] and complex demography[30]) and to population trees[31] (see also Nicholson et al.[32] for a similar method that uses population trees and Günther and Coop[33] for one that uses population kinships). The population differentiation approach has greatest power when comparing very closely related populations with very large sample size.[19] 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 might be better represented by continuous clines rather than discrete clusters,[34–36] and/or might contain a large number of discrete subpopulations corresponding to a large number of possible population comparisons.[37,38] Principal-component analysis (PCA)[34,39] offers an appealing alternative to model-based clustering methods[40,41] for modeling human genetic diversity and has been applied to infer population structure in many settings.[35,36,39,42–48] 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.

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [3]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [4]Google, Palo Alto, CA 94043, USA; [5]Departments of Statistical Science, Computer Science, and Mathematics, Duke University, Durham, NC 27708, USA
[6]These authors contributed equally to this work
*Correspondence: galinsky@fas.harvard.edu (K.J.G.), aprice@hsph.harvard.edu (A.L.P.)
http://dx.doi.org/10.1016/j.ajhg.2015.12.022. ©2016 by The American Society of Human Genetics. All rights reserved.

Another advantage of PCA is its low computational cost; by drawing upon recent advances in random matrix theory,[49–51] the time to infer the top PCs is linearly proportional to the number of samples. This is 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[52] but has much lower computational cost.

Specifically, 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. 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.[19]

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 min of compute time and 2.6 GB of RAM for this analysis, orders of magnitude better than any other publicly available software. We detected evidence of population structure along the top four PCs, which separated samples into several subpopulations. Using our PC-based test for selection, we replicate previously known selected loci (*LCT* [MIM: 603202], *HLA* [MIM: 142800], *OCA2* [MIM: 611409], and *IRF4* [MIM: 601900]) and identify three additional signals of selection at *IGH* (MIM: 147100), *IGFBP3* (MIM: 146732), and *ADH1B* (MIM: 103720). The signal in *ADH1B* at coding variant rs1229984 has previously been associated to alcoholism[53–56] and shown to be under selection in East Asians;[14,55,57,58] we show that it is a rare example of independent evolution on two continents.[11,12]

Material and Methods

Overview of Methods

We first describe the FastPCA algorithm, which is an implementation of the *blanczos* method from Rokhlin et al.[49–51] As with our previous work on PCA,[34,39] 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, because 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,[59] a technique 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.[19] 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 (see Web Resources).

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 $\boldsymbol{Y}_{M\times N} = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, \ldots, \boldsymbol{y}_M^T)^T$ where each row $\boldsymbol{y}_i$ has approximately mean 0 and variance 1 for SNPs in Hardy-Weinberg equilibrium.

$$\widehat{p}_i = \frac{\sum_{j=1}^{N} x_{ij}}{2N_i} = \frac{\boldsymbol{x}_i 1}{2 1^T 1}$$

$$y_{ij} = \frac{x_{ij} - 2\widehat{p}_i}{\sqrt{2\widehat{p}_i(1-\widehat{p}_i)}} \qquad \text{(Equation 1)}$$

$$\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iN}) = \frac{x_{ij} - 2\widehat{p}_i}{\sqrt{2\widehat{p}_i(1-\widehat{p}_i)}}$$

Here, $\boldsymbol{x}_i$ is the row vector of genotypes for SNP $\boldsymbol{i}$ and $\boldsymbol{y}_i$ is the normalized row vector. x_{ij} and y_{ij} are the genotype/normalized genotype at SNP $\boldsymbol{i}$ for sample $\boldsymbol{j}$. N_i is the number of valid genotypes at SNP $\boldsymbol{i}$. All this is used to calculate $\widehat{p}_i$, the sample allele frequency for SNP $\boldsymbol{i}$, 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 that returns all the principal components. FastPCA, which makes use of recent advances in random matrix theory,[49–51] 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 $\boldsymbol{T}_{(I+1)L\times N} = \boldsymbol{U}_H^T\boldsymbol{Y}$. The SVD of $\boldsymbol{T} = \boldsymbol{U}_T\boldsymbol{\Sigma}_T\boldsymbol{V}_T^T$ can be computed efficiently and approximates the SVD of $\boldsymbol{Y}$ because $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \approx \boldsymbol{U}_H\boldsymbol{T} = \boldsymbol{U}_H\boldsymbol{U}_T\boldsymbol{\Sigma}_T\boldsymbol{V}_T^T$. For the PCA, we are interested only 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 $O(M\,N\,L)$ time. Then, the SVD of H takes $O(M\,I^2\,L^2)$ and the SVD of T takes $O(N\,I^2\,L^2)$ time. Taking I and L to be constants, the overall running time simplifies to $O(M\,N)$. This is much faster than traditional $O(M\,N^2 + N^3)$ PCA methods as well as the $O(M\,N^2)$ of flashpca.

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 $2p_i(1-p_i)F_{ST}$.[32] 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:

$$E\left[p_{i1} - p_{i2}\right] = 0$$
$$Var\left[p_{i1} - p_{i2}\right] = 2p_i(1-p_i)F_{ST}$$
$$p_{i1} - p_{i2} \sim N\left[0, 2p_i(1-p_i)F_{ST}\right], F_{ST} \ll 1, 0 \ll p_i \ll 1. \qquad \text{(Equation 2)}$$

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, $\widehat{p}_{1i}$ and $\widehat{p}_{i2}$. 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:[13,15,19]

$$\widehat{p}_{i1} \sim N\left[p_{i1}, \frac{p_{i1}(1-p_{i1})}{2N_1}\right], \widehat{p}_{i2} \sim N\left[p_{i2}, \frac{p_{i2}(1-p_{i2})}{2N_2}\right],$$
$$N_1, N_2 \gg 0, 0 \ll p_{i1}, p_{i2} \ll 1$$
$$D_i = \widehat{p}_{i1} - \widehat{p}_{i2} \sim N\left[0, \sigma_D^2\right] = N\left[0, \widehat{p}_i(1-\widehat{p}_i)\left(2F_{ST} + \frac{1}{2N_1} + \frac{1}{2N_2}\right)\right],$$
$$p_i \approx \widehat{p}_i = \frac{\widehat{p}_{i1} + \widehat{p}_{i2}}{2}, p_{i1} \approx p_{i2}. \qquad \text{(Equation 3)}$$

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

$$\widehat{p}_1 = \frac{\boldsymbol{x}_i\boldsymbol{\alpha}}{21^T\boldsymbol{\alpha}}, \widehat{p}_2 = \frac{\boldsymbol{x}_i(1-\boldsymbol{\alpha})}{21^T(1-\boldsymbol{\alpha})}, D_i = \frac{\boldsymbol{x}_i\boldsymbol{\alpha}}{21^T\boldsymbol{\alpha}} - \frac{\boldsymbol{x}_i(1-\boldsymbol{\alpha})}{21^T(1-\boldsymbol{\alpha})}. \qquad \text{(Equation 4)}$$

If we run PCA on the normalized genotype matrix 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 (because $\boldsymbol{v}^T\boldsymbol{1} = 0$, $\boldsymbol{v}^T\boldsymbol{v} = 1$)

$$v_q = \frac{1}{N_q}\sqrt{\frac{N_1N_2}{N}}. \qquad \text{(Equation 5)}$$

In this case, D_i can be rewritten as

$$D_i = \frac{1}{2}\sqrt{\frac{N_1N_2}{N}}\boldsymbol{x}_i\boldsymbol{v}. \qquad \text{(Equation 6)}$$

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

$$\frac{D_i^2}{\sigma_D^2} = \frac{\frac{1}{4}\frac{N_1N_2}{N}(\boldsymbol{x}_i\boldsymbol{v})^2}{\widehat{p}_i(1-\widehat{p}_i)\left(\frac{1}{2N_1} + \frac{1}{2N_2}\right)} = \left[\left(\frac{\boldsymbol{x}_i - 2\widehat{p}_i1^T}{2\widehat{p}_i(1-\widehat{p}_i)}\right)\boldsymbol{v}\right]^2 = \left[\boldsymbol{y}_i\boldsymbol{v}\right]^2. \qquad \text{(Equation 7)}$$

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 ($\alpha_j \in [0,1]$), $\widehat{p}_1$, $\widehat{p}_2$, and D_i can still be estimated by Equation 4. The individual $\widehat{p}_q$ s will still asymptotically follow a normal distribution (because of the Lyapunov central limit theorem[60]) 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 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,[34] because PCs dominated by such outliers might violate normality assumptions.) We can treat the eigenvector as a linear transformation of the ancestry vector:

$$\boldsymbol{\alpha} = \beta_0 + \beta_1\boldsymbol{v}. \qquad \text{(Equation 8)}$$

Substituting these values into Equation 4, we find

$$D_i = \frac{\beta_1}{2N\beta_0(1-\beta_0)}\boldsymbol{x}_i\boldsymbol{v} \propto \boldsymbol{y}_i\boldsymbol{v}. \qquad \text{(Equation 9)}$$

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

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

$$\boldsymbol{U} = \boldsymbol{Y}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}. \qquad \text{(Equation 10)}$$

Here, $\boldsymbol{V}$ contains the right singular vectors that are equivalent to the PCs, $\boldsymbol{U}$ contains the left singular vectors that are rescaled SNP weights, and $\boldsymbol{\Sigma}$ contains the singular values that 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 $\boldsymbol{D} = (D_1, D_2, ..., D_M)^T$. In other words,

$$\frac{\sqrt{M}}{\Sigma_k}\boldsymbol{\gamma}_i\boldsymbol{v}_k \sim N(0,1)$$

$$\frac{M}{\Sigma_k^2}\left(\boldsymbol{\gamma}_i\boldsymbol{v}_k\right)^2 \sim \chi_1^2. \quad \text{(Equation 11)}$$

In the case of non-random SNP ascertainment and non-random choice of reference and variant allele, the expectation of D_i might 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.[61]

One assumption underlying the statistic is that the true minor allele frequency is not extremely small, otherwise the assumption of normality will not hold.[19] For this reason, the selection statistic was computed only for those SNPs containing minor allele frequency greater than 1% in our sample.

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 ($\boldsymbol{P}_i = (p_{i1}, p_{i2},..., p_{iQ})^T$)) were generated by simulating random drift (see below).
3. Admixture (α_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 $\boldsymbol{P}_i = (p_{i1}, p_{i2},..., p_{iQ})^T$. 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 $p_{iqt} = (z_{iqt}/2N_e)$. 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 $F_{ST} = -\log(1-(\tau/2N_e))$.[19] For $F_{ST} \approx 0.1$, 0.01, and 0.001, $N_e = 1$k, 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 five populations with a phylogenetic structure as follows. We set $N_e = 10$k 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 $F_{ST} \approx 0.001$ and 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, $\boldsymbol{a} = a\boldsymbol{1}$, 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 $F_{ST} = 0.01$, $M = 100$k SNPs, and $N \approx$ {1k, 1.5k, 2k, 3k, 5k, 7k, 10k, 15k, 20k, 30k, 50k, 70k, 100k} individuals (because we used six populations of equal sample size, we rounded N to multiples of six). Throughout this paper we report CPU time, but due to multithreading present in the GSL[62] and OpenBLAS libraries, run time was about 60% of CPU time. FastPCA accuracy was assessed using $M = 50$k SNPs and $N \approx 10$k individuals at $F_{ST} =$ {0.001,0.002,...,0.010}. Calibration and power of the selection statistic was assessed using two populations at $F_{ST} =$ {0.1,0.05, 0.02,0.01,0.005,0.002,0.001,0.0005} and also using five populations with the tree structure described above. We set $M = 60$k, the effective number of independent SNPs in genotype array data.[63] 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 $p_{iq*} = 0.5 + (D/2)$ for one population and $p_{iq} = 0.5 - (D/2)$ 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.

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 ($\boldsymbol{v}_1, \boldsymbol{v}_2,..., \boldsymbol{v}_K$) against which we wish evaluate the performance a set of computed PCs ($\boldsymbol{u}_1, \boldsymbol{u}_2,..., \boldsymbol{u}_K$), then the performance calculation becomes

$$MEV = K^{-1}\sum_{j=1}^{K}\sqrt{\sum_{j=1}^{K}\left(\boldsymbol{v}_k \cdot \boldsymbol{u}_j\right)^2} = K^{-1}\sum_{j=1}^{K}\|\boldsymbol{U}^T\boldsymbol{v}_k\|. \quad \text{(Equation 12)}$$

Here, $\boldsymbol{U}$ is a matrix whose column vectors are the PCs which we are testing. The test matrix can be either 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 $\boldsymbol{v}_k^* = (v_{k,1}^*, v_{k,2}^*, \ldots v_{k,N}^*)$ 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.

GERA Dataset

The GERA dataset includes 62,318 individuals from Northern California typed on a European-specific 670,176-SNP array.[64] 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,[65] and individuals with less than 90% European ancestry as predicted by SNPweights[66] 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[65] 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, and 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[66] 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.

PC Projection

POPRES[67] 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.[66]

Results

FastPCA Simulations

We used simulated data to compare the running time and memory usage of FastPCA to three previous algorithms: smartpca,[34,39] PLINK2-pca,[65] and flashpca[68] (see Web Resources). 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 Material and Methods). For each run, running time was capped at 100 hr and memory usage was capped at 40 GB. The running time and memory usage of FastPCA scaled linearly with simulated dataset size (i.e., $O(M\,N)$ cost) (Figure 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 min and 3.2 GB of memory with FastPCA (Table S1). (We also note that shellfish [see Web Resources], a parallel PCA implementation, requires $O(M\,N^2 + N^3)$ and is not computationally tractable on large datasets, as previously demonstrated.[68]) Thus, FastPCA—unlike other publicly available software packages for analyzing genetic data—enables rapid PCA without specialized computing facilities.

We next assessed the accuracy of FastPCA, using PLINK2-pca[65] as a benchmark. We used the same simulation framework as before, with 10k individuals (1,667k individuals per population) and 50k SNPs. We varied the divergence between populations, as quantified by F_{ST}.[69] We assessed accuracy using the mean of explained variances (MEV) of the five population-structure PCs (see Material and Methods). We determined that the results of FastPCA and PLINK-pca were virtually identical (Figure 2). This indicates that FastPCA performs comparably to standard PCA algorithms while running much faster.

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 $N = 50k$ 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} to 10^{-5}. Similar results indicating appropriate calibration were obtained for simulations with admixture (Table S2), as expected because the drift model still applies in the case of admixture.[35] The median of the selection statistic was slightly inflated at $F_{ST} = 0.1$ due to a deficiency in the tail (Figure S1 and Table S2) 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 S2). We also simulated five populations with a phylogenetic structure (see Material and 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 S1 and Table S2).

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 SNPs under

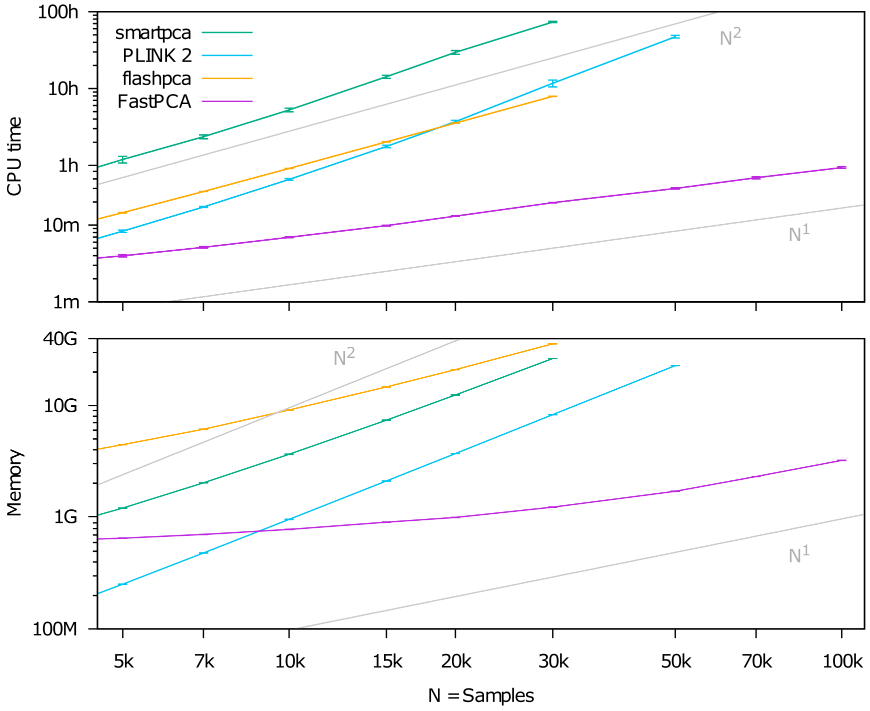

Figure 1. Running Time and Memory Requirements of FastPCA and Other Algorithms

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 hr) and flashpca exceeded the memory constraint (40 GB). With 70k individuals, PLINK2-pca exceeded the memory constraint (40 GB). Run times are based on one core of a 2.26-GHz Intel Xeon L5640 processor; we caution that run time comparisons might vary by a small constant factor as a function of the computing environment. Numerical data are provided in Table S1 and the error bars indicate ± 1 SD.

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 3A). We examined this effect in more depth by 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 3B) than at $F_{ST} = 0.01$ (Figure 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[19] in the case of data from discrete subpopulations (Figure S2). 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 S3).

Application of FastPCA to a European American Cohort

We ran FastPCA on the GERA cohort (see Web Resources), a large European American dataset containing 54,734 individuals and 162,335 SNPs after QC filtering and LD pruning (see Material and Methods). This computation took 57 min and 2.6 GB of RAM. PC1 and PC2 separated individuals along the canonical Northwest European (NW), Southeast European (SE), and Ashkenazi Jewish (AJ) axes,[15] as indicated by labeling the individuals by predicted fractional ancestry from SNPweights[66] (Figure 4). These results are consistent with Banda et al.,[64] 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[67] onto these PCs[39] (see Material and 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 5). This visual subpopulation clustering was confirmed via k-means clustering on the top four PCs, which consistently grouped the AJ, SE, NE, IR, and EE populations separately (Figure S4). We note that, in general, K PCs can cluster samples into $K + 1$ subpopulations.

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 Material and Methods). The resulting Manhattan plots for PCs 1–4 are displayed in Figure 6 (QQ plots are displayed in Figure S5). Analyses of PCs 5–10 indicated that these PCs do not represent true population structure (Figure S6), but rather are either dominated by a small number of long-range LD loci[42,70,71] 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 regions (Table S3).

Genome-wide significant signals (listed in Table 1) included several known selection regions[9,72–75] and signals at

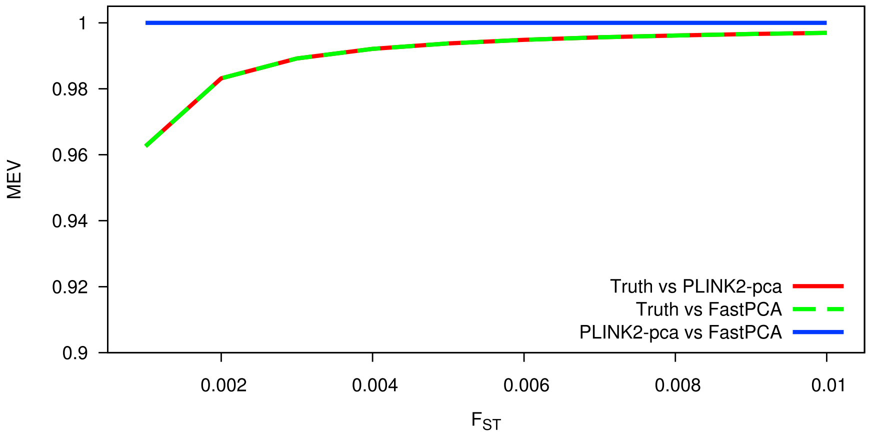

Figure 2. Accuracy of FastPCA and PLINK2-pca
FastPCA and PLINK2-pca were run on simulated populations of varying divergence. The simulated data comprised 50k SNPs and 10k total individuals from six subpopulations derived from a single ancestral population. PCs computed by PLINK2-pca and FastPCA were compared to the true population PCs and to each other using the mean of explained variances (MEV) metric (see Material and Methods). 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).

ADH1B, *IGFBP3*, and *IGH* (see below). Suggestive signals were observed at additional known selection regions[74,76] (Table S4). After removing the regions in Table 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[42,70] (Figure S7 and Table S5). 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 S8). Detecting subtle signals of selection benefited from the large sample size, as shown by the fact that subsampling the GERA dataset at smaller sample sizes and recomputing PCs and selection statistics generally led to less significant signals (Table 2). We note that several suggestive selection signals, including signals at the known selected loci *TLR1*[74] (MIM: 601194) and *SLC45A2*[76] (MIM: 606202), are on the cusp of being significant and further increases to sample size might 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). The allele rs1229984*T has been shown to have a protective effect on alcoholism risk[53–56] and to produce an REHH signal in East Asians,[14,55,57,58] 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.[77,78]) We examined the allele frequency of the rs1229984*T allele in the five subpopulations AJ, SE, NE, IR, and EE (Table S6). 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 to the remaining subpopulations using the discrete subpopulation selection statistic[19] also produced a genome-wide significant signal after correcting for all hypotheses tested (Table S7); 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[79] populations (see Web Resources), along with the allele frequency of the regulatory SNP rs3811801 that might also have been a target of selection in Asian populations.[55] The haplotype carrying rs3811801*A (and corresponding haplotype H7) was absent in populations outside of East Asia (Table S8). 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 older SNP under selection in Europeans, whereas 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[80] (MIM: 114480), height[81] (MIM: 606255), blood pressure,[82] and hypertension,[83] 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). Genetic variation in *IGH* has been associated with multiple sclerosis[84] (MIM: 126200), 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[19] (Tables S9 and S10). 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.

Discussion

We have detected new, genome-wide significant signals of selection by applying a PC-based selection statistic to top PCs computed via FastPCA, a computationally efficient (linear-time and linear-memory) algorithm. Although

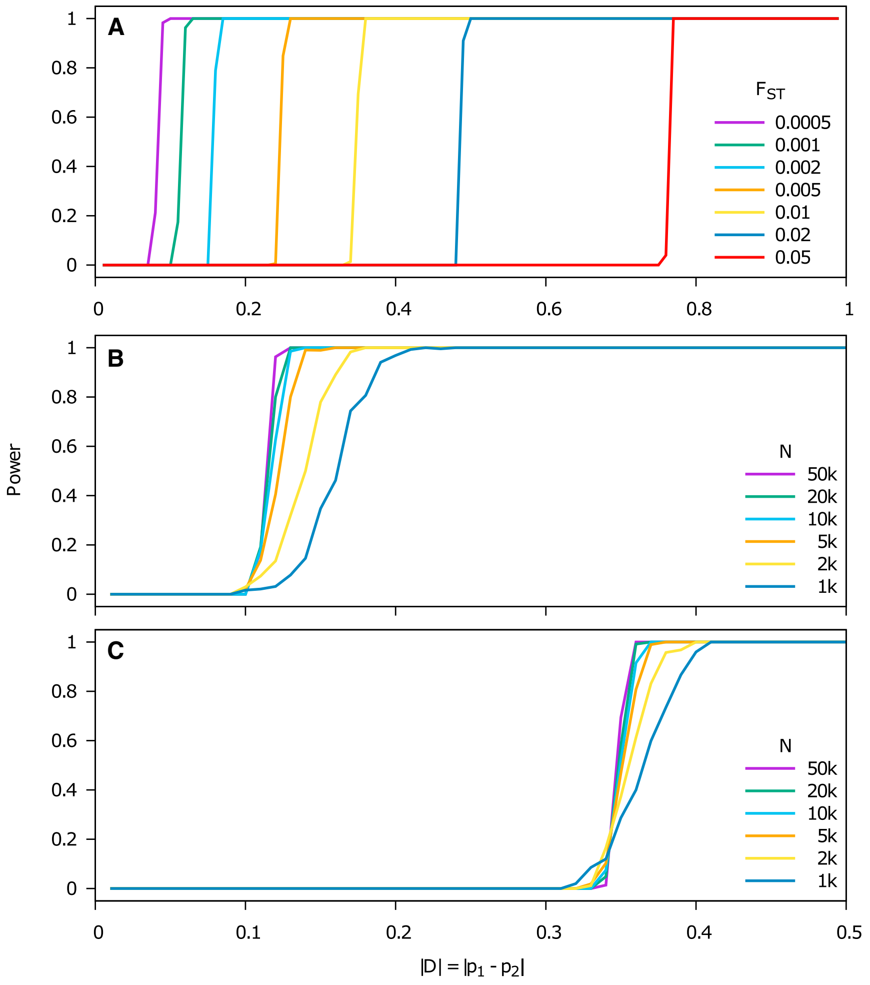

Figure 3. Power of PC-Based Selection Statistic

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 $F_{ST} = 0.001$.

mixed model association methods are increasingly appealing for conducting genetic association studies,[63,85] 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.[86,87] Our PC-based selection statistic extends previous statistics developed for discrete populations.[19] In contrast to previous work on detecting selection via PCs[71,88] or using the spatial ancestry analysis (SPA) method,[89] our statistic is able to detect signals at genome-wide significance, a key consideration in genome scans for selection.[90] 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.[91,92] 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[14,55,57,58] using REHH[61] (which can detect only relatively recent signals and does not work on standing variation[3]). We also detected genome-wide significant evidence of selection at the disease-associated *IGFBP3* and *IGH*. Although 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 might 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.[93]

We note that our work has several limitations. First, top PCs do not always reflect population structure, but can instead reflect assay artifacts[94] or regions of long-range LD;[42] 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) and rerunning PCA. Second, common variation might not provide a complete description of population structure, which might be different for rare variants;[95] 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 datasets with pervasive LD and large numbers of rare variants (see Appendix A). Third, our selection statistic is capable only of detecting that selection occurred, but not when or where it occurred; indeed, top PCs might 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

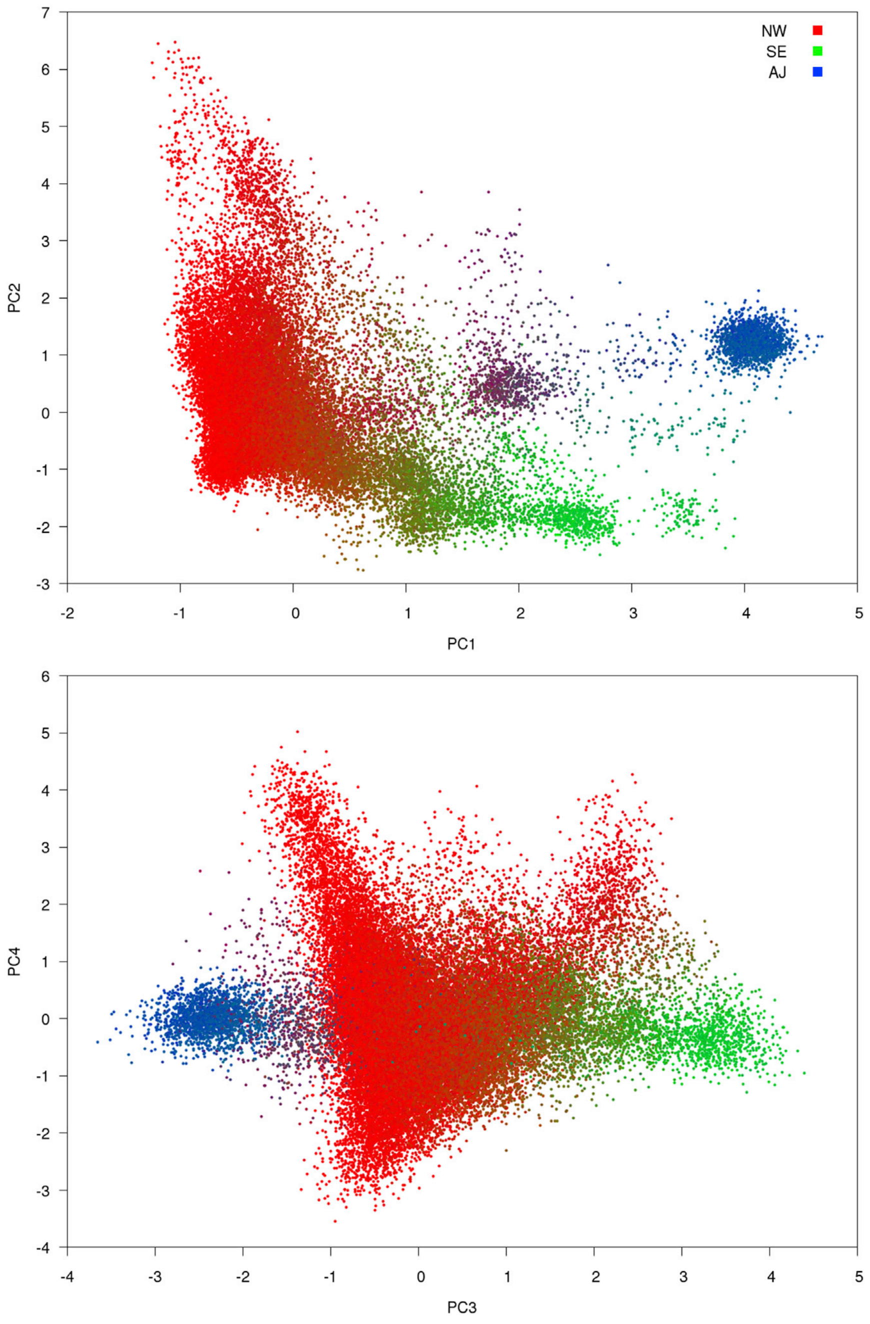

Figure 4. FastPCA Results on GERA Dataset
FastPCA and SNPweights[66] were run on the GERA cohort and the principal components from FastPCA were plotted. Individuals were colored by mapping Northwest European (NW), Southeast European (SE), and Ashkenazi Jewish (AJ) ancestry estimated by SNPweights to the red/green/blue color axes (see Material and Methods). PC1 and PC2 (top) separate the GERA cohort into northwest (NW), southeast (SE), and Ashkenazi Jewish (AJ) subpopulations. PC3 separates the AJ and SE individuals, and PC3 and PC4 (bottom) further separates the NW European individuals.

linearly along a PC; the SPA method[89] (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 datasets of the future.

Appendix A

Inferring ancestry from genetic data is a common problem in both population and medical genetic studies, and many methods exist to address it.[39,40,96] Principal-component analysis (PCA)[39] has been shown to be effective at elucidating geographic structure from genetic data[97] and correcting for confounding due to population stratification in association mapping.[34] These uses of PCA depend critically on its ability to separate genetically disparate subpopulations when analyzing data from commercial genotyping arrays. However, as high-throughput sequence data becomes more common, enabling ancestry inference from this new class of data is becoming increasingly relevant.

Because sequence data contain more variants and many more population-specific variants,[98] it might be reasonable to expect that PCA applied to high-throughput sequence data will be substantially more effective than the corresponding analysis on genotype data. However, our results suggest the opposite. Specifically, PCA makes assumptions about marker independence that are violated by the pervasive linkage disequilibrium in sequence data. In addition, assumptions about genetic drift that are reasonable for common SNPs on genotyping arrays are less so when applied to the numerous rare variants in sequence data.[95]

Methods

PCA is generally applied to a genetic relationship matrix (GRM) that is computed as

$$g_s = \frac{x_s - 2p_s}{\sqrt{2p_s(1-p_s)}}$$

$$G = \sum_{s \in SNPs} g_s g_s^T,$$

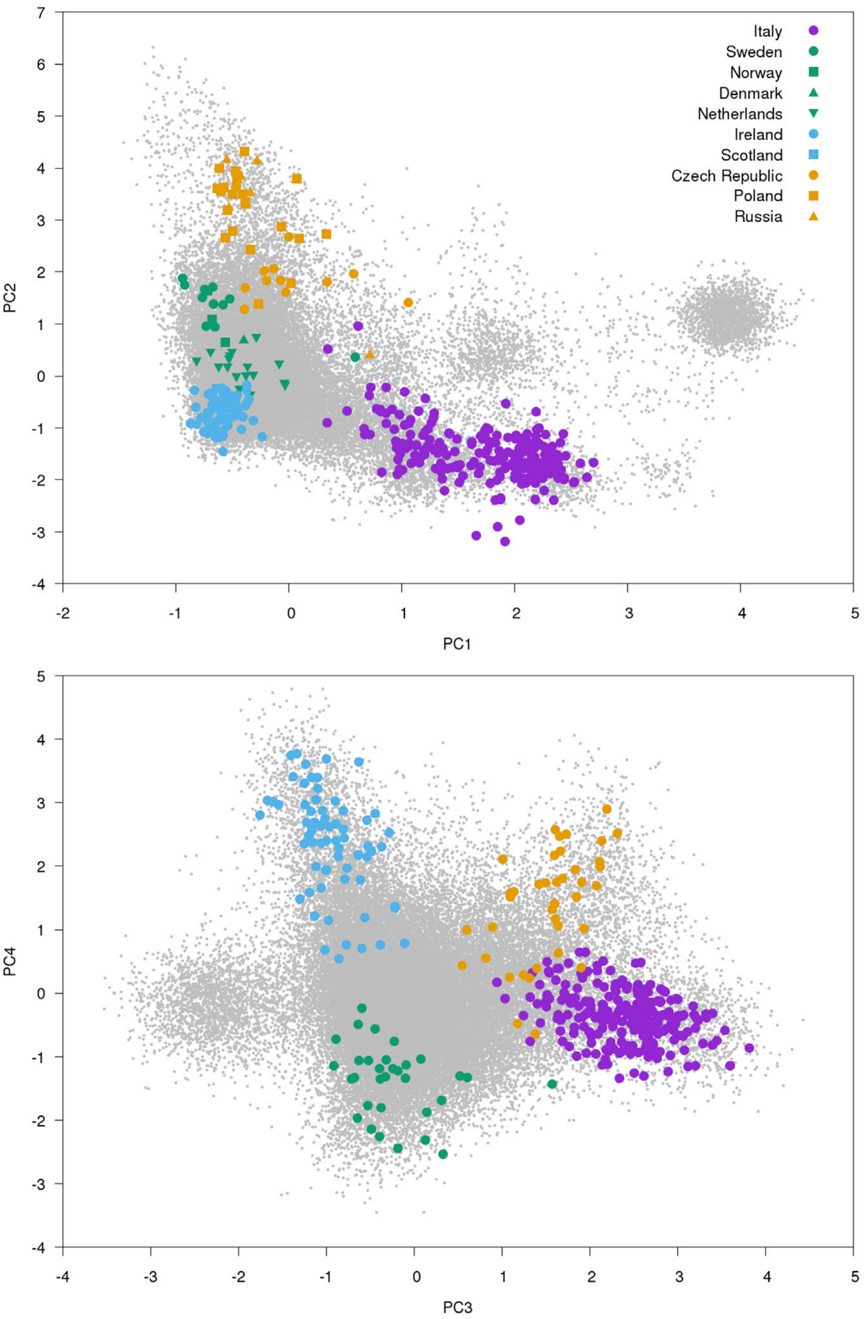

Figure 5. Separation of Irish, Eastern European, and Northern European Individuals in GERA Data

We report results of projecting POPRES[67] individuals onto top PCs. The plot of PC3 versus PC4 (bottom) 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,[66] except that Irish and Eastern European individuals were colored purple and orange, respectively, to indicate additional population structure.

where x_s is a vector of genotypes for SNP s and p_s is the minor allele frequency of SNP s. We propose modifications to standard PCA to deal with two challenges that are present in sequence data but absent from genotype data: pervasive linkage disequilibrium and rare variants. Specifically, we recommend that LD pruning be applied to sequence data and that singleton variants be removed. Although we evaluated more sophisticated approaches to handling these issues, they did not improve our results beyond these simpler approaches. Importantly, we recommend against a commonly used strategy of removing all low-frequency rare variants because these variants contain significant information for detecting population structure.

Linkage Disequilibrium

It is well known that application of PCA to regions of the genome containing long-range LD blocks can confound PCA's ability to separate disparate populations.[39,71] As a result, these LD blocks are often simply excluded from analysis. However, in sequence data, many regions of the genome outside of previously identified long-range LD blocks contain sufficient LD to bias results. As a result, we examine three methods to deal with LD: LD pruning, LD shrinkage,[71] and LD regression.[29,39,99]

LD pruning is a commonly applied approach to removing correlated SNPs from a dataset. To produce a dataset pruned for LD above a threshold T, one SNP of any pair of SNPs in LD ($r^2 > T$) is removed from the data.

LD shrinkage is a more sophisticated method of correcting for LD proposed by Zou et al.[71] In LD shrinkage, each SNP s is weighted by its LD to surrounding SNPs before inclusion in the genetic relationship matrix.

$$g_s = \frac{x_s - 2p_s}{\sqrt{2p_s(1-p_s)}}$$

$$w_s = \frac{1}{1 + \sum_{t \in window(s)} r^2_{s,t}}$$

$$G = \sum_{s \in SNPs} g_s g_s^T$$

We note that $t \in$ *window(s)* refers to SNPs t that are within some region of the genome surrounding SNP s. Intuitively,

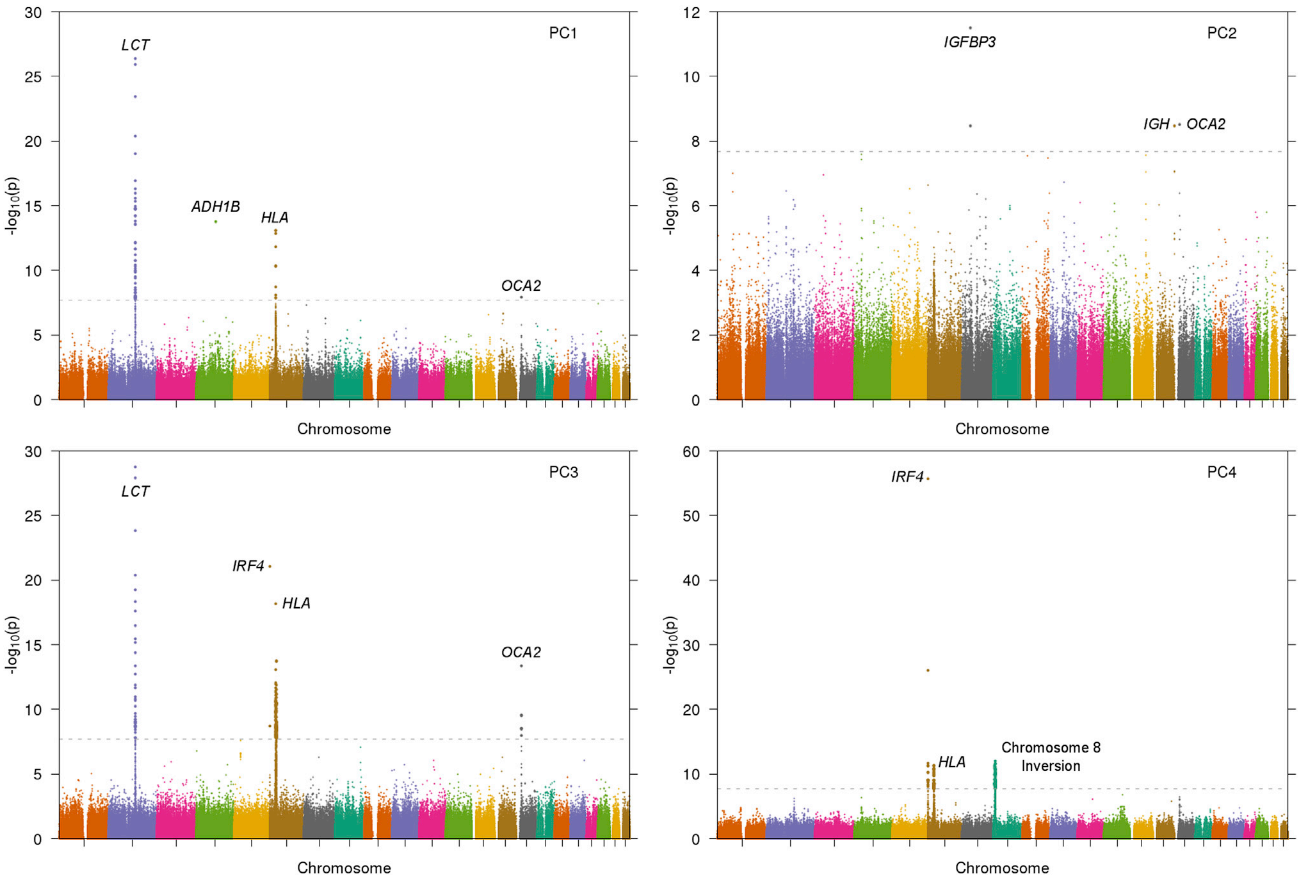

Figure 6. Signals of Selection in the Top PCs of GERA Data

We display Manhattan plots for selection statistics computed using each of the top four PCs. The gray 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 × 4 PCs).

this is a heuristic to correct for the over-representation in the GRM of some SNPs that are redundant with respect to nearby SNPs.

LD regression was originally proposed in Patterson et al.[39] and utilized extensively in Gusev et al.[99] Only the residual of a SNP—after regressing out other SNPs in LD—is included in the GRM:

$$g_s = \frac{x_s - 2p_s}{\sqrt{2p_s(1-p_s)}}$$

$$g_s \sim \sum_{t \in window(s)} g_t + \varepsilon_s$$

$$G = \sum_{s \in SNPs} \varepsilon_s \varepsilon_s^T$$

Rare Variants

In considering how to optimally include rare variants in the genome, we examined three strategies. The first strategy was to include all rare variants as described in the computations above without any modifications. The second strategy was to exclude all variants below a threshold, which is a standard strategy used in several recent papers. We compared these simple strategies to a strategy based on reweighting rare variants to optimize the separation between populations.

We considered a particular scenario to optimize. Specifically, we imagine that two populations that split from one another t generations ago are equally represented in our GRM. We would like to optimize the proportion of variance in our GRM that is explained by the true population labels. That is, our figure of merit is

$$\frac{\frac{1}{n(n-1)}\sum_i \sum_{j \in pop(i)} g_{i,j} - \frac{1}{n^2}\sum_i \sum_{j \in pop(i)} g_{i,j}}{\sqrt{\mathrm{Var}\left(g_{i,j}\right)}},$$

where $pop(i)$ refers to the subpopulation from which individual i came.

Now, considering the population split, our data contain two classes of variants: those variants that are result of mutations predating the population split (pre-split SNPs) and those variants arising after the population split (post-split SNPs). For pre-split SNPs, we invoke the normal approximation to genetic drift described. That is, the difference between allele frequencies p_1 and p_2 (for populations 1 and 2, respectively) is

$$(p_1 - p_2) \sim N(0, 2F_{ST}p(1-p)),$$

Table 1. **Genome-wide Significant Signals of Selection in GERA Data**

Locus	Chromosome	Region (Mb)	PC	Best Hit	p Value
LCT[9]	2	134.8–137.6	1	rs6754311	4.15×10^{-27}
			3	rs4988235	1.83×10^{-29}
ADH1B	4	100.5	1	rs1229984	1.67×10^{-14}
IRF4[74,75]	6	0.3–0.5	3	rs12203592	8.69×10^{-22}
			4		1.83×10^{-56}
HLA[72]	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	3.14×10^{-12}
Chr8 Inversion[42]	8	8.2–11.9	4	rs6984496	9.21×10^{-13}
IGH	14	106.0–106.1	2	rs34614900	3.34×10^{-9}
OCA2[73,75]	15	25.9–26.2	1	rs12916300	1.12×10^{-8}
			2		3.07×10^{-9}
			3		4.29×10^{-14}

We list regions with genome-wide significant ($\alpha = 0.05$, Bonferroni correction with 608,981 SNPs × 4 PCs = 2,435,924 hypotheses tested, $p < 2.05 \times 10^{-8}$) evidence of selection in the top four PCs. We provide previous reference(s) where available. The chromosome 8 inversion signal is due to a PC artifact (see Results). Regions with suggestive evidence of selection ($10^{-6} < p < 2.05 \times 10^{-8}$) are listed in Table S3.

where p is the allele frequency in the ancestral population prior to the split and F_{ST} quantifies the genetic drift that has occurred since the split. We note that this approximation is reasonable for common SNPs and for small values of F_{ST}. If we assume that our data contains only pre-split SNPs, then our figure of merit is optimized by the standard computation of the GRM given above. On the other hand, rare, post-split SNPs have the property that

$$|p_1 - p_2| = 2\hat{p}$$

Table 2. **Performance of Natural Selection Statistic in Subsampled Data**

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

The selection statistic was computed in random subsets of individuals of specified size for each SNP in Table 1 (except for the chromosome 8 inversion region) and the known selection regions *TLR1*[74] and *SLC45A2*[76] in Table S4. We report the median selection statistic p value across 100 random subsets.

where $\widehat{p}$ is the allele frequency estimated from the sample. This difference implies that the optimal weighting for pre-split SNPs is $(1/\sqrt{p_s(1-p_s)})$ identically

$$g_i^s = \frac{x_i - 2p_s}{\sqrt{p_s(1-p_s)}}, \text{for pre} - \text{split SNP } s$$

but the optimal weighting for post-spit SNPs is $\sqrt{(F_{ST}^2 + 2F_{ST} + 2)/(F_{ST}(1-2p_s))}$.

$$g_i^s = (x_i - 2p_s)\sqrt{\frac{F_{ST}^2 + 2F_{ST} + 2}{F_{ST}(1-2p_s)}}, \text{for post} - \text{split SNP } s$$

However, this modification requires knowledge of the F_{ST} between studied subpopulations and, more dauntingly, which SNPs are post-split. We believe it is reasonable to iterate over several values of F_{ST} (and find that in real data results are relatively robust to choice of F_{ST}). In order to deal with uncertainty over the set of post-split SNPs, we propose that a SNP be considered post-split if

$$\frac{1}{\sqrt{p_s(1-p_s)}} > \sqrt{\frac{F_{ST}^2 + 2F_{ST} + 2}{F_{ST}(1-2p_s)}}.$$

We examine the effect of both of these modifications on the effectiveness o f PCA to separate genetically disparate subpopulations.

Analysis of Northern versus Southern Europe in POPRES Targeted Sequencing Data

We analyzed 531 individuals from the UK referred to as Northern European and 146 Italian, 134 Portuguese, 100 Spaniards, and 7 Swiss Italian individuals collectively referred to as Southern European.[10] We excluded 25.9 kb of sequence data from genes on the X chromosome, focusing solely on the autosomes. In total, 8,469 SNPs were polymorphic in either of the Northern or Southern European samples. These variants were overwhelmingly rare, with 81.5% of variants having a MAF < 1% in the combined sample.

We tested various methods to correct for LD and better handle rare variants (see Material and Methods). The results are summarized in Table S11. These results indicate that handling of both rare variants and LD is critical to maximizing the performance of PCA on this class of data. Applying standard PCA, the top five PCs explained only 2.3% of the variance ($r^2 = 0.023$) of the true population labels. This was improved substantially by removing or reweighting rare variants with (r^2 = 0.287, 0.341, 0.352) for removing variants with MAF < 0.02, removing singletons and reweighting, respectively. This indicates that rare variants, particularly singletons, might be problematic when analyzed by PCA. However, the difference between removing variants with MAF < 0.02 and reweighting ($r^2 = 0.287$ versus 0.352) suggests that these variants do contain useful information for ancestry inference and should not be universally excluded.

Additionally, application of a method to correct for LD significantly improved performance of PCA when performed in conjunction with singleton exclusion or rare variant reweighting. With rare variant reweighting, LD shrinkage[8] ($r^2 = 0.563$) performed slightly better than LD regression (r^2 = 0.528)[2] and LD pruning (r^2 = 0.534). Although LD pruning performed well, this might be due to the fact that LD is broken up because the dataset contains sequence data from separated chunks of genome.

Recommendations

In datasets that do not include pervasive LD or large numbers of rare variants (i.e., genotyping data), standard techniques are likely to be successful in detecting population structure. However, in datasets that have pervasive LD and large numbers of rare variants, we recommend that LD pruning and singleton removal be applied. Although more sophisticated methods for dealing with these issues were assessed, we did not observe significant improvements above and beyond these simpler approaches. Importantly, we do not recommend that all low-frequency and rare variants (MAF < 0.02) be removed because these variants do significantly improve detection of population structure.

Supplemental Data

Supplemental Data include 8 figures and 11 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.12.022.

Acknowledgments

We are grateful to D. Reich for helpful discussions and S. Pollack for assistance with FastPCA software. This research was funded by NIH grant R01 HG006399. S.M. is funded by NSF grants DMS-1209155 and DMS-1418261.

Received: July 31, 2015
Accepted: December 31, 2015
Published: February 25, 2016

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes, http://browser.1000genomes.org
dbGaP, GERA, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000674.v1.p1
EIGENSOFT v.6.1, https://data.broadinstitute.org/alkesgroup/EIG6.1/
flashpca, https://github.com/gabraham/flashpca
OMIM, http://www.omim.org/
PLINK 1.9, https://www.cog-genomics.org/plink2/
Shellfish, http://www.stats.ox.ac.uk/~davison/software/shellfish/shellfish.php

References

1. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. Science *312*, 1614–1620.
2. Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A.G. (2007). Recent and ongoing selection in the human genome. Nat. Rev. Genet. *8*, 857–868.
3. Novembre, J., and Di Rienzo, A. (2009). Spatial patterns of variation due to natural selection in humans. Nat. Rev. Genet. *10*, 745–755.
4. Scheinfeldt, L.B., and Tishkoff, S.A. (2013). Recent human adaptation: genomic approaches, interpretation and insights. Nat. Rev. Genet. *14*, 692–702.
5. Karlsson, E.K., Kwiatkowski, D.P., and Sabeti, P.C. (2014). Natural selection and infectious disease in human populations. Nat. Rev. Genet. *15*, 379–393.
6. Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akey, J.M., and Jones, K.W. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum. Genomics *1*, 274–286.
7. Hamblin, M.T., and Di Rienzo, A. (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. *66*, 1669–1679.
8. Akey, J.M., Zhang, G., Zhang, K., Jin, L., and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome Res. *12*, 1805–1814.
9. Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. *74*, 1111–1120.
10. 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., et al. (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science *310*, 1782–1786.
11. 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., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. *39*, 31–40.
12. 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., et al. (2007). Diet and the evolution of human amylase gene copy number variation. Nat. Genet. *39*, 1256–1260.
13. Ayodo, G., Price, A.L., Keinan, A., Ajwang, A., Otieno, M.F., Orago, A.S.S., Patterson, N., and Reich, D. (2007). Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. Am. J. Hum. Genet. *81*, 234–242.
14. Han, Y., Gu, S., Oota, H., Osier, M.V., Pakstis, A.J., Speed, W.C., Kidd, J.R., and Kidd, K.K. (2007). Evidence of positive selection on a class I ADH locus. Am. J. Hum. Genet. *80*, 441–456.
15. Price, A.L., Helgason, A., Palsson, S., Stefansson, H., St Clair, D., Andreassen, O.A., Reich, D., Kong, A., and Stefansson, K. (2009). The impact of divergence time on the nature of population structure: an example from Iceland. PLoS Genet. *5*, e1000505.
16. Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., Gong, X., Wang, H., Shen, Y., Pan, X., et al. (2009). Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am. J. Hum. Genet. *85*, 762–774.
17. 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., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science *329*, 75–78.
18. 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., et al. (2010). Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet. *6*, e1001116.
19. Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., et al. (2011). Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. Am. J. Hum. Genet. *89*, 368–381.
20. Hancock, A.M., Witonsky, D.B., Alkorta-Aranburu, G., Beall, C.M., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J.K., Coop, G., and Di Rienzo, A. (2011). Adaptations to climate-mediated selective pressures in humans. PLoS Genet. *7*, e1001375.
21. Ko, W.-Y., Rajan, P., Gomez, F., Scheinfeldt, L., An, P., Winkler, C.A., Froment, A., Nyambo, T.B., Omar, S.A., Wambebe, C., et al. (2013). Identifying Darwinian selection acting on different human APOL1 variants among diverse African populations. Am. J. Hum. Genet. *93*, 54–66.
22. Engelken, J., Carnero-Montoro, E., Pybus, M., Andrews, G.K., Lalueza-Fox, C., Comas, D., Sekler, I., de la Rasilla, M., Rosas, A., Stoneking, M., et al. (2014). Extreme population differences in the human zinc transporter ZIP4 (SLC39A4) are explained by positive selection in Sub-Saharan Africa. PLoS Genet. *10*, e1004128.
23. Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. Nature *517*, 327–332.
24. Fumagalli, M., Moltke, I., Grarup, N., Racimo, F., Bjerregaard, P., Jørgensen, M.E., Korneliussen, T.S., Gerbault, P., Skotte, L., Linneberg, A., et al. (2015). Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science *349*, 1343–1347.
25. Kwiatkowski, D.P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. Am. J. Hum. Genet. *77*, 171–192.
26. Lewontin, R.C., and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics *74*, 175–195.
27. Beaumont, M.A., and Balding, D.J. (2004). Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. *13*, 969–980.
28. Foll, M., and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics *180*, 977–993.
29. Excoffier, L., Hofer, T., and Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. Heredity (Edinb) *103*, 285–298.
30. Foll, M., Gaggiotti, O.E., Daub, J.T., Vatsiou, A., and Excoffier, L. (2014). Widespread signals of convergent adaptation to high altitude in Asia and America. Am. J. Hum. Genet. *95*, 394–407.

31. Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., and Sancristobal, M. (2010). Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics *186*, 241–262.
32. Nicholson, G., Smith, A.V., Jónsson, F., Gústafsson, Ó., Stefánsson, K., and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. J. R. Stat. Soc. Ser. B. Stat. Methodol. *64*, 695–715.
33. Günther, T., and Coop, G. (2013). Robust identification of local adaptation from allele frequencies. Genetics *195*, 205–220.
34. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
35. Price, A.L., Butler, J., Patterson, N., Capelli, C., Pascali, V.L., Scarnicci, F., Ruiz-Linares, A., Groop, L., Saetta, A.A., Korkolopoulou, P., et al. (2008). Discerning the ancestry of European Americans in genetic association studies. PLoS Genet. *4*, e236.
36. Novembre, J., and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. Nat. Genet. *40*, 646–649.
37. 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., et al. (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035–1044.
38. Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E.C., Cunliffe, B., Lawson, D.J., et al.; Wellcome Trust Case Control Consortium 2; International Multiple Sclerosis Genetics Consortium (2015). The fine-scale genetic structure of the British population. Nature *519*, 309–314.
39. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.
40. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.
41. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
42. Tian, C., Plenge, R.M., Ransom, M., Lee, A., Villoslada, P., Selmi, C., Klareskog, L., Pulver, A.E., Qi, L., Gregersen, P.K., and Seldin, M.F. (2008). Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet. *4*, e4.
43. Seldin, M.F., and Price, A.L. (2008). Application of ancestry informative markers to association studies in European Americans. PLoS Genet. *4*, e5.
44. 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., and Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.
45. 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., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature *451*, 998–1003.
46. Schlebusch, C.M., Skoglund, P., Sjödin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M.G.B., et al. (2012). Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science *338*, 374–379.
47. Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409–413.
48. 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., et al. (2014). Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science *344*, 1280–1285.
49. Rokhlin, V., Szlam, A., and Tygert, M. (2009). A randomized algorithm for principal component analysis. SIAM J. Matrix Anal. Appl. *31*, 1100–1124.
50. Halko, N., Martinsson, P., Shkolnisky, Y., and Tygert, M. (2011). An algorithm for the principal component analysis of large data sets. SIAM J. Sci. Comput. *33*, 2580–2594.
51. Halko, N., Martinsson, P., and Tropp, J. (2011). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. *53*, 217–288.
52. Duforet-Frebourg, N., Bazin, E., and Blum, M.G.B. (2014). Genome scans for detecting footprints of local adaptation using a Bayesian factor model. Mol. Biol. Evol. *31*, 2483–2495.
53. Edenberg, H.J., and Foroud, T. (2013). Genetics and alcoholism. Nat. Rev. Gastroenterol. Hepatol. *10*, 487–494.
54. Whitfield, J.B. (2002). Alcohol dehydrogenase and alcohol dependence: variation in genotype-associated risk between populations. Am. J. Hum. Genet. *71*, 1247–1250, author reply 1250–1251.
55. Li, H., Gu, S., Han, Y., Xu, Z., Pakstis, A.J., Jin, L., Kidd, J.R., and Kidd, K.K. (2011). Diversification of the ADH1B gene during expansion of modern humans. Ann. Hum. Genet. *75*, 497–507.
56. Gelernter, J., Kranzler, H.R., Sherva, R., Almasy, L., Koesterer, R., Smith, A.H., Anton, R., Preuss, U.W., Ridinger, M., Rujescu, D., et al. (2014). Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci. Mol. Psychiatry *19*, 41–49.
57. Osier, M.V., Pakstis, A.J., Soodyall, H., Comas, D., Goldman, D., Odunsi, A., Okonofua, F., Parnas, J., Schulz, L.O., Bertranpetit, J., et al. (2002). A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. Am. J. Hum. Genet. *71*, 84–99.
58. Peter, B.M., Huerta-Sanchez, E., and Nielsen, R. (2012). Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genet. *8*, e1003011.
59. Golub, G.H., and Van Loan, C.F. (1996). Matrix Computations (Baltimore: Johns Hopkins University Press).
60. Billingsley, P. (1995). Probability and Measure (New York: Wiley-Interscience).
61. 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., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.
62. Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., and Fabrice, R. (2009). GNU Scientific Library Reference Manual (Network Theory Limited).
63. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. *46*, 100–106.

64. 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., et al. (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*, 1285–1295.

65. 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., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

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

67. 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., et al. (2008). The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am. J. Hum. Genet. *83*, 347–358.

68. Abraham, G., and Inouye, M. (2014). Fast principal component analysis of large-scale genome-wide data. PLoS ONE *9*, e93766.

69. Bhatia, G., Patterson, N., Sankararaman, S., and Price, A.L. (2013). Estimating and interpreting FST: the impact of rare variants. Genome Res. *23*, 1514–1521.

70. Fellay, J., Shianna, K.V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., et al. (2007). A whole-genome association study of major determinants for host control of HIV-1. Science *317*, 944–947.

71. Zou, F., Lee, S., Knowles, M.R., and Wright, F.A. (2010). Quantification of population structure using correlated SNPs by shrinkage principal components. Hum. Hered. *70*, 9–22.

72. 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., et al. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat. Genet. *38*, 1166–1172.

73. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol. *4*, e72.

74. 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., et al.; Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.

75. 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., and Pritchard, J.K. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.

76. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al.; International HapMap Consortium (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.

77. Li, H., Mukherjee, N., Soundararajan, U., Tárnok, Z., Barta, C., Khaliq, S., Mohyuddin, A., Kajuna, S.L.B., Mehdi, S.Q., Kidd, J.R., and Kidd, K.K. (2007). Geographically separate increases in the frequency of the derived ADH1B*47His allele in eastern and western Asia. Am. J. Hum. Genet. *81*, 842–846.

78. Treutlein, J., Frank, J., Kiefer, F., and Rietschel, M. (2014). ADH1B Arg48His allele frequency map: filling in the gap for Central Europe. Biol. Psychiatry *75*, e15.

79. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

80. Key, T.J., Appleby, P.N., Reeves, G.K., and Roddam, A.W.; Endogenous Hormones and Breast Cancer Collaborative Group (2010). Insulin-like growth factor 1 (IGF1), IGF binding protein 3 (IGFBP3), and breast cancer risk: pooled individual data analysis of 17 prospective studies. Lancet Oncol. *11*, 530–542.

81. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.

82. 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., et al.; Global Blood Pressure Genetics Consortium (2014). Effects of long-term averaging of quantitative blood pressure traits on the detection of genetic associations. Am. J. Hum. Genet. *95*, 49–65.

83. 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., et al.; COGENT BP Consortium (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. Am. J. Hum. Genet. *96*, 21–36.

84. Buck, D., Albrecht, E., Aslam, M., Goris, A., Hauenstein, N., Jochim, A., Cepok, S., Grummel, V., Dubois, B., Berthele, A., et al.; International Multiple Sclerosis Genetics Consortium; Wellcome Trust Case Control Consortium (2013). Genetic variants in the immunoglobulin heavy chain locus are associated with the IgG index in multiple sclerosis. Ann. Neurol. *73*, 86–94.

85. 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., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.

86. 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., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.

87. 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., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. *43*, 519–525.

88. 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., et al. (2012). Natural positive selection and north-south genetic diversity in East Asia. Eur. J. Hum. Genet. *20*, 102–110.

89. Yang, W.-Y., Novembre, J., Eskin, E., and Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. Nat. Genet. *44*, 725–731.

90. 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., et al. (2014). Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. Am. J. Hum. Genet. *95*, 437–444.

91. He, Y., Wang, M., Huang, X., Li, R., Xu, H., Xu, S., and Jin, L. (2015). A probabilistic method for testing and estimating selection differences between populations. Genome Res. *25*, 1903–1909.

92. Chen, G.-B., Lee, S.H., Zhu, Z.-X., Benyamin, B., and Robinson, M.R. (2015). EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. bioRxiv http://dx.doi.org/10.1101/023457.

93. 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., et al. (2014). Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. Nat. Commun. *5*, 3983.

94. 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., et al. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat. Genet. *37*, 1243–1246.

95. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. *44*, 243–246.

96. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS Genet. *8*, e1002453.

97. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature *456*, 98–101.

98. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A.; 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

99. Gusev, A., Bhatia, G., Zaitlen, N., Vilhjalmsson, B.J., Diogo, D., Stahl, E.A., Gregersen, P.K., Worthington, J., Klareskog, L., Raychaudhuri, S., et al. (2013). Quantifying missing heritability at known GWAS loci. PLoS Genet. *9*, e1003993.