Biometrics 74, 155–164
March 2018

DOI: 10.1111/biom.12708

Computation of Ancestry Scores with Mixed Families and Unrelated Individuals

Yi-Hui Zhou,[1,*] James S. Marron,[2] and Fred A. Wright[3]

[1]Department of Biological Sciences, Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, U.S.A.
[2]Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, U.S.A.
[3]Department of Biological Sciences and Statistics, Bioinformatics Research Center, North Carolina State University, Raleigh, U.S.A.
[*]*email:* yihui_zhou@ncsu.edu

Summary. The issue of robustness to family relationships in computing genotype ancestry scores such as eigenvector projections has received increased attention in genetic association, and is particularly challenging when sets of both unrelated individuals and closely related family members are included. The current standard is to compute loadings (left singular vectors) using unrelated individuals and to compute projected scores for remaining family members. However, projected ancestry scores from this approach suffer from shrinkage toward zero. We consider two main novel strategies: (i) matrix substitution based on decomposition of a target family-orthogonalized covariance matrix, and (ii) using family-averaged data to obtain loadings. We illustrate the performance via simulations, including resampling from 1000 Genomes Project data, and analysis of a cystic fibrosis dataset. The matrix substitution approach has similar performance to the current standard, but is simple and uses only a genotype covariance matrix, while the family-average method shows superior performance. Our approaches are accompanied by novel ancillary approaches that provide considerable insight, including individual-specific eigenvalue scree plots.

Key words: Genetic association; Population stratification; Principal components.

1. Introduction

Differing ancestries of human subpopulations create systematic differences in genetic allele frequencies across the genome, a phenomenon known as population stratification or substructure. If a phenotypic trait such as disease is associated with subpopulation membership, a genetic association study can identify spurious relationships with genetic markers. Singular value decomposition (SVD) of genotype data or eigen decomposition of covariance matrices can be used to identify population stratification. The eigenvectors (essentially principal component scores) that correspond to large eigenvalues can be used as covariates in association analysis (Levine et al., 2013). The combined analysis of unrelated and related individuals is a common feature of genetic association studies (Zhu et al., 2008). However, the presence of close-degree relatives in a genetic dataset presents difficulties, as the family structure can greatly influence the eigenvalues and eigenvectors.

Cystic fibrosis (CF) is a recessive genetic lung disorder, caused by a mutation in the single gene *CFTR*. However, considerable genetic variation remains in the severity of disease, and evidence indicates this variation is complex and influenced by numerous genes (Wright et al., 2011). Genotypes gathered by the North American CF Consortium are typical of a large-scale genomewide association study (GWAS), with thousands of individuals and over 1 million genetic markers (Corvol et al., 2015). For covariate control, the eigenvectors are computed for a submatrix of the genotypes, after a "thinning" process in which only an ancestry-informative subset of markers which have low marker–marker correlation is retained (Patterson et al., 2006). We illustrate the proposed methods using the dataset from the CF patients described as "GWAS1" in Corvol et al. (2015), with 21,205 thinned ancestry markers and 3444 individuals. The dataset includes 2546 singletons (unrelated to others) and 438 small families of siblings (417 sets of 2 individuals, 20 sets of 3, and 1 set of 4). Figure 1 is a scatter plot of the fifth versus the first "ancestry scores" (right singular vectors for this example) from a naive analysis of all 3444 individuals (see Section 2).

Here, the PC5 scores are driven largely by membership in the family of size 4, rather than the ancestry substructure of interest. Several additional top-ranked eigenvectors are also driven by family membership. Accordingly, matrix projection methods have been proposed (Zhu et al., 2008), in which singular value decomposition is performed on singletons, followed by projections for the remaining families. However, this approach has been shown to produce shrunken projected scores for the family members (Lee et al., 2010). In Conomos et al. (2015), the PCAiR method was proposed to expand the set of individuals included in the SVD to include a single individual from each family, resulting in improved performance. However, the question remains as to whether scores for the remaining projected individuals will exhibit shrinkage, or if the methods can be further improved.

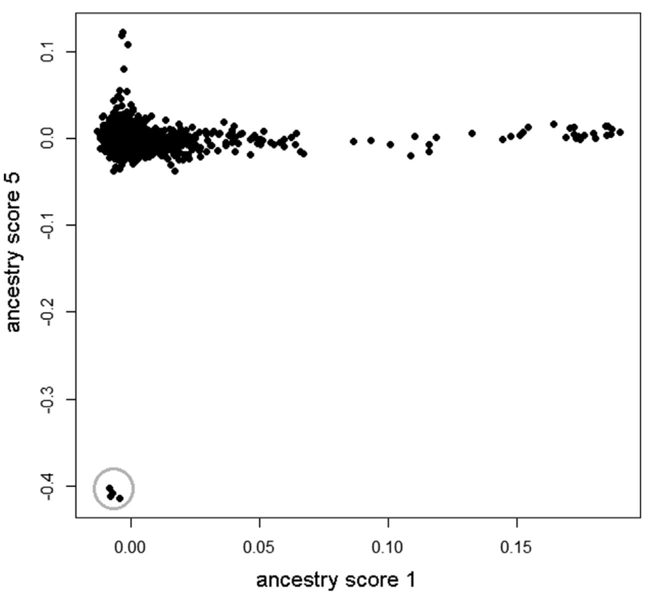


Figure 1. Ancestry score (right singular vector) 5 versus ancestry score 1 in a naive decomposition of the covariance matrix using all CF individuals. Membership in a family of size 4 (highlighted with a circle) is responsible for most of the variation in ancestry score 5.

In contrast to previous efforts, in this article we directly address the family covariance structures that complicate ancestry score calculation. We introduce several novel approaches to account for the family-specific correlation structures in a single analysis, avoiding difficulties posed by standard projection methods. Comparison via simulation and analysis of real data indicate that our approaches offer substantial improvements over existing methods in either simplicity or performance, and are straightforward to implement. The methods follow from an understanding of high-dimensional geometry and use the new device of smoothed individual scree plots for further exploration. The article is organized as follows. In Section 2, we introduce the existing and proposed approaches. Section 3 describes performance criteria. Section 4 described the simulation methods. Section 5 contains performance results. An accompanying Web Supplement contains details of the algorithms and numerous additional illustrations.

2. Methods

This article discusses a number of competing methods, and considerable notation is unavoidable. To reduce confusion, we adopt uniform notation where possible. We use $i = 1, \ldots, p$ to denote genetic markers (single nucleotide polymorphisms, SNPs), $j = 1, \ldots, n$ to denote individuals (including families), and typically $p \gg n$. The individuals can be partitioned into singletons ($\mathcal{S}$, unrelated to anyone else in the dataset), and family members ($\mathcal{F}$, related to at least one other individual), with respective sample sizes $n_{\mathcal{S}}$ and $n_{\mathcal{F}}$, so $n = n_{\mathcal{S}} + n_{\mathcal{F}}$. The set $\mathcal{F}$ is partitioned into distinct families $\{\mathcal{F}_f\}$ of size n_f, $f = 1, \ldots, F$. Let G be the original $p \times n$ genotype matrix, with elements taking on the values 0, 1, or 2, typically coded as the number of minor alleles, and $\bar{g}_{i.} = \sum_{j=1}^{n} g_{ij}/n$, the mean for SNP i. The scaled $p \times n$ genotype matrix X consists of elements $x_{ij} = (g_{ij} - \bar{g}_{i.})/\sqrt{\sum_{j'}(g_{ij'} - \bar{g}_{i.})^2/(n-1)}$, so that $\sum_j x_{ij} = 0$, $\sum_j x_{ij}^2 = n - 1$, for all $i = 1, \ldots, p$.

Family membership could be inferred by KING (Manichaikul et al., 2010), which is used for analyses for the PCAiR method described below. We find that a simple screening method for first and second-degree relationships is also effective, identifying pairs of individuals j_1, j_2 such that $\mathrm{corr}(x_{.j_1}, x_{.j_2}) > \eta$, after X has been provisionally residualized for gross ancestry structure. Using $\eta = 0.1$ identifies paired family members up to second-degree with high sensitivity and specificity (see Supplement Section S1).

2.1. *SVD and Eigen Decomposition*

The "naive" approach to handling the full dataset is to simply compute the singular value decomposition $X = UDV^T$, using the columns of V as informative scores for ancestry, in decreasing order of the singular values contained in the diagonal of D. However, as Figure 1 showed, this approach can be highly influenced by family structure. Other methods work with the matrix of sample covariances of the individuals, which for the full matrix X is the $n \times n$ matrix $M = \overline{X}^T\overline{X}/(p-1)$, where $\overline{X}$ is the column-centered version of X. Eigen decomposition of M provides eigenvectors that are nearly identical to the columns of V. Alternatively, a principal component (PC) decomposition provides PC scores that are identical or nearly identical (depending on column-centering) to V. For ease of discussion, we refer to the column output from the various methods simply as "ancestry scores," except when further specificity is required.

2.2. *The Singleton Projection (SP) Method*

Singleton projection (Zhu et al., 2008) first computes the SVD $X_{\mathcal{S}} = U_{\mathcal{S}}D_{\mathcal{S}}V_{\mathcal{S}}^T$. Ancestry scores for the complete data are given as the columns of the $n \times n_{\mathcal{S}}$ matrix $\widetilde{V}_{SP} = X^T U_{\mathcal{S}} D_{\mathcal{S}}^{-1}$, as in practice no more than $n_{\mathcal{S}}$ ancestry scores (PCs) will be used as covariates. Here and subsequently a tilde ("∼") will signify a matrix or vector that has been made robust to the effects of family relationships, and $\widetilde{V}$ with a corresponding subscript will be used to denote the matrix of ancestry scores for each method. The singleton projection approach is easily implemented in popular software such as EIGENSTRAT (Price et al., 2006). By ignoring families in the initial step, singleton projection loses accuracy, with the family ancestry scores suffering from the shrinkage phenomenon described in Lee et al. (2010), who also prescribed a bias-correction procedure to correct the shrinkage. However, the bias-correction is a multi-step procedure whose performance has not been established for a range of eigenvalues, and is not convenient for collections of families of various sizes.

2.3. *PCAiR*

To incorporate more information from the family data, PCAiR (Conomos et al., 2015) works with a set of unrelated individuals $\mathcal{U}$, where $\mathcal{U}$ includes the singletons plus a single member from each family. Thus $\mathcal{U}$ does not contain any related pairs, and we will use $\mathcal{R}$ to denote the complementary set of related individuals not in $\mathcal{U}$. The set $\mathcal{U}$ is not unique, and

PCAiR attempts to identify and use a maximally informative set. The full approach of Conomos et al. (2015) involves genotype normalization differing slightly from our scaling, identification of family members using KING (Manichaikul et al., 2010), and numerous matrix operations. However, a careful reading and simulations below show that the essence of the approach is similar to singleton projection, using columns of $\widetilde{V}_{PCAiR} = X^T U_{\mathcal{U}} D_{\mathcal{U}}^{-1}$ as scores, where $U_{\mathcal{U}}, D_{\mathcal{U}}$ are obtained from the SVD $X_{\mathcal{U}} = U_{\mathcal{U}} D_{\mathcal{U}} V_{\mathcal{U}}^T$. Although numerous ancestry estimation procedures have been proposed (Sankararaman et al., 2008), for the calculation of ancestry scores using eigenvectors or principal components, the results in Conomos et al. (2015) indicate that PCAiR represents the current state of the art. Here, we use the KING software and PCAiR code from Conomos et al. (2015) as recommended, except for Gaussian simulations. For comparison, we also use the straightforward algorithm above coded in *R*, with the family member included in $\mathcal{U}$ randomly chosen.

2.4. *Matrix Substitution (MS)*

As noted, ancestry scores can be obtained directly from a covariance matrix (Frudakis et al., 2003). We propose simply modifying the sample covariance matrix $M = \overline{X}^T\overline{X}/(p-1)$ so that family members do not have outsized influence. We construct a matrix $\widetilde{M}$ with entries $\widetilde{m}_{j_1 j_2}$ =*median entry in* M if $j_1 \neq j_2$ and j_1 and j_2 belong to the same family, and $\widetilde{m}_{j_1 j_2} = m_{j_1 j_2}$ otherwise. Co-family members are typically a small fraction of the pairs of individuals, and so M and $\widetilde{M}$ differ in only a small fraction of elements. Following matrix substitution, we compute $\widetilde{V}_{MS}$ as the eigenvectors of $\widetilde{M}$. One appealing quality of the approach is that it treats all family members symmetrically, and no individual need be chosen to "represent" the family.

Although the matrix substitution approach is simple and appealing, it does not provide "whitened" actual genotype data, which might be useful for other purposes, such as analyses of subsets of individuals or for careful investigation of marker–marker correlation (Lake et al., 2000). Interestingly, whitened genotype data consistent with matrix substitution can be computed using a series of matrix operations, in such a manner that singleton data remains unchanged. The approach is derived in the Supplement Section S2, and termed "covariance-preserving whitening" (CPW).

2.5. *Family Average (FA) Projection*

A potential concern with the PCAiR projection method of Section 2.3 is that only a single member is used from each family. We consider the approach of using the mean vector for each family, instead of a single representative member, to obtain loadings. Specifically, for family f indexed by $\mathcal{F}_f$, we compute a new data vector $\widehat{x}_{.f} = \bar{z}_{\mathcal{F}_f}(\sum_{j\in\mathcal{F}_f} ||x_{.j}||/n_f)$, where $\bar{z}_{\mathcal{F}_f}$ is the unit-length family mean vector from Web Supplement Section S3. Multiplication by the family average length ensures that $\widehat{x}_{.f}$ has a "typical" length—otherwise the variance contribution from the family mean vector would be much smaller than for an individual, distorting the approach. We construct a new matrix of singletons combined columnwise with the F rescaled family averages, $\underset{p\times(n_{\mathcal{S}}+F)}{X_{\mathcal{A}}} = [\underset{p\times n_{\mathcal{S}}}{X_{\mathcal{S}}}, \underset{p\times F}{\widehat{X}}]$, and compute the SVD $X_{\mathcal{A}} = U_{\mathcal{A}} D_{\mathcal{A}} V_{\mathcal{A}}^T$. Finally, the projected ancestry scores are computed for all individuals, as the columns of $\widetilde{V}_{FA} = X^T U_{\mathcal{A}} D_{\mathcal{A}}^{-1}$.

2.6. *Geometric Rotation/Family Whitening (FW)*

Yet another approach to computing ancestry scores would be to include all of the stratification data, but to first modify genotypes within families only to reduce the family-specific impact on SVD analysis. The derivation is provided in Web Supplement Section S3, presented using both geometric motivation and a more standard matrix whitening. Such family whitening is entirely for the purpose of stratification analysis—the modified genotypes are not intended to be used for trait association. This approach is quite different from PCAiR or FA, as it modifies all family members and does not use a single representative or averaged individual from within a family. We describe the steps of family whitening in the Web Supplement, for completeness and because the approach has intellectual appeal. However, later simulation results show that family whitening has poorer performance than the other proposed methods, and so is not emphasized after initial evaluations.

3. Criteria for Evaluation

Here, we describe criteria to evaluate the performance of ancestry score calculations. The first two criteria reflect the ability to discriminate among known (by simulation) subpopulations, while providing family ancestry scores that are comparable to those from singletons. The third criterion, which can be assessed with real data even though true ancestry is unknown, measures the tendency for ancestry scores to remain stable for an individual who belongs to a family, depending on whether the individual's family members are also included in the analysis. Finally, we end this section by introducing the "individual scree plot," a novel visualization tool to provide insight into the behavior of ancestry scores.

3.1. R^2 *Criteria for Prediction of Ancestry*

We assume the population consists of K ancestry subgroups, and a true ancestry value $a_{jk} \in [0,1]$ for individual j is the proportion of the autosomal genome derived from the kth subpopulation, $\sum_k a_{jk} = 1$. If $a_{jk} = 1$, then the individual is entirely derived from the kth ancestry subgroup, and values in $(0,1)$ correspond to admixture. We performed linear regression of true ancestry on ancestry scores as follows. The ancestry scores are columns of a matrix $\widetilde{V}$, where each entry v_{jl} is the lth ancestry score for individual j. Using multiple linear regression to predict a from the set of ancestry scores, we have $(1-R_k^2) = \frac{\sum_j (a_{jk}-\hat{a}_{jk})^2}{\sum_j (a_{jk}-\bar{\bar{a}}_{.k})^2}$, where $\hat{a}_{jk}$ is the prediction and $\bar{\bar{a}}_{.k}$ is the grand mean for the kth ancestry, and an overall score $(1-R^2) = (1-\sum_k R_k^2/K)$. We note that in the special case of binary a (distinct subpopulations), the same overall R^2 can be obtained by averaging over analyses of variance using each column of $\tilde{V}$ as the response, with membership in each of the K strata as predictors (Cabanski et al., 2010).

Using PCAiR as a baseline, for each method we also compute a proportional reduction in prediction error $e_k = (R^2_{method,k} - R^2_{PCAiR,k})/(1-R^2_{PCAiR,k})$, expressed as a percentage. A final single error reduction index can be computed as the average $\bar{e} = \sum_k e_k/K$. The rationale for using prediction accuracy for true ancestry is intuitive, but we are not aware of a clear description in the literature. Thus for completeness, we describe the rationale in the Web Supplement Section S4, in the context of using predicted ancestry for covariate control.

3.2. *The Relateds Square Error (RSE) Criterion*

Most of the methods described in this article use a partition into family members $\mathcal{F}$ versus singletons $\mathcal{S}$. An important performance aspect that is not fully captured by $(1 - R^2)$ is the tendency for the family members to exhibit reduced variation in the ancestry scores. For the initial simulations, we introduce a measure of the tendency for ancestry scores of family members to overlap their singleton counterparts, calculated within each stratum before summarizing.

For each stratum k, we further partition Ω_k into $\Omega_{k,\mathcal{F}}$ and $\Omega_{k,\mathcal{S}}$, corresponding to family members and singletons within the stratum, of sizes $n_{k,\mathcal{F}}$ and $n_{k,\mathcal{S}}$. Let $\overline{v}_{\Omega_{k,\mathcal{S}}l}$ denote the average of the lth ancestry scores for individuals in $\Omega_{k,\mathcal{S}}$. For the lth ancestry score, we compute the Relateds Squared Error (RSE),

$$RSE_l = \sqrt{\frac{\sum_{k=1}^{K} \sum_{j\in\Omega_{k,\mathcal{F}}} (v_{jl} - \overline{v}_{\Omega_{k,\mathcal{S}}l})^2/(n_{k,\mathcal{F}} - 1)}{\sum_{k=1}^{K} \sum_{j\in\Omega_{k,\mathcal{S}}} (v_{jl} - \overline{v}_{\Omega_{k,\mathcal{S}}l})^2/(n_{k,\mathcal{S}} - 1)}}.$$

In other words, for both family members and singletons, we compute the average squared deviation from the mean of singletons. For a method that performs well, projected family members will behave similarly to singletons, and RSE_l will be near 1.0. We average the first 5 RSE_l values to obtain an overall RSE. For PCAiR, we compute the RSE using $\mathcal{U}$ and $\mathcal{R}$ instead of $\mathcal{S}$ and $\mathcal{F}$, respectively.

3.3. *An Instability Index*

The criteria above require knowledge of the true population strata. Here, we describe a performance criterion based on *stability* of the eigenvector values for family members, as compared to an internally computed standard. It can be performed for real data, without knowing true ancestry. We will let $\underset{n\times n}{W}$ denote a “gold standard” ancestry score matrix to be used subsequently, and $\underset{n\times n}{Q}$ a comparison matrix, and for both matrices the columns are arranged in the same order as X.

Suppose we wish to compute ancestry scores for an individual j who belongs to a family. One approach, robust to family structure, is to combine j with the singletons, computing $\underset{p\times(n_{\mathcal{S}+1})}{X_{\mathcal{S}\cup j}} = UD \underset{(n_{\mathcal{S}+1})\times(n_{\mathcal{S}+1})}{V^T}$. As j is unrelated to $\mathcal{S}$, we will use the last column of V as the jth column of W, that is, $w_{.j} = v_{.(n_{\mathcal{S}}+1)}$. We perform this procedure in succession for all $j \in \mathcal{F}$ to populate the family ($\mathcal{F}$) columns of W. Alternately, we populate the $\mathcal{F}$ columns of Q by performing, for each $f \in \mathcal{F}$, the family-robust methods described in this article, applied for each f using the genotype data for $S \cup \mathcal{F}_f$. In other words, W is computed by combining each family member with $\mathcal{S}$ one at a time, while Q is computed by combining each *family* with $\mathcal{S}$. We consider W as the gold standard, because it is computed using only unrelated individuals in each step. For an ancestry method that is robust to family structure, we expect Q to be similar to W. The instability index for the lth ancestry score is $instability_l = \sum_{j\in\mathcal{F}}(q_{jl} - w_{jl})^2/\sum_{j\in\mathcal{F}} q_{jl}^2$, with an ideal value of zero.

3.4. *Individual Scree Plots*

Scree plots (Cattell, 1966) are a useful method to visualize the relative importance of eigenvectors and PCs. Here, we take the scree plot in a new direction, by studying the corresponding plot for each individual, that is, studying the squares of the projections of each individual. For the SVD $X = UDV^T$, these projections are $(X^T U)^2 = (VD)^2$, and the column sums of $(VD)^2$ are the squared singular values of X. These values essentially correspond to principal component variance values, which are also used in overall scree plots. Accordingly, for the robust ancestry methods described in this article, we use rows of $(\widetilde{V}\widetilde{D})^2$ as *individual* scree values, reflecting the contribution of each individual to the overall influence of each ancestry score. The individual scree values are noisy (see Web Supplement Section S5) and cover several orders of magnitude, so we plot them on the $\log_{10}$ scale and perform loess smoothing to discern important patterns.

4. Genotype Simulation Methods and Settings

Much of the behavior of the various methods can be understood largely in terms of covariance patterns, and are not unique to discrete genotype data. This is seen and motivated by idealized Gaussian simulations provided in Web Supplement Section S6, which illustrate that the MS and FA methods have promising performance. Another informative set of simulations more directly reflects the special properties of genotype data, studied next.

4.1. *Idealized Simulation of Genotypes and Family Sibships*

Web Supplement Section S7.1 describes our procedure for realistic simulation of founder genotype data for K population strata, following the Balding–Nichols model with $F_{ST} = 0.01$. The model uses modest serial correlation of successive markers of approximately 0.2 in blocks of 20 markers, 20,000 markers in total, and matches the allele frequencies in the CF data. To simulate a family sibship of size n_f, we followed a realistic autosomal recombination model. First, we generated enough singletons within each subpopulation so that parents could be simulated and then discarded. For each family, from the singletons we randomly selected two parents at random from a stratum (subpopulation) without replacement. Artificial grandparental haplotype genomes were generated for each parent by randomly dividing the alleles. Children were then simulated using an artificial recombination process, with recombinations in each parent simulated as a geometric random variable for successive SNPs, at a rate such that on average 30 recombinations occurred per meiosis. For each family, the n_f children were simulated independently from the same parental pair.

For the *balanced* simulations, we generated $K = 5$ subpopulations using the approach above. Sibships of $n_f = 3$ were simulated such that the proportion of individuals belonging to families *prop* was the same in each subpopulation. The total sample sizes used were $n = \{500, 1000, 2000\}$, with family proportions $prop = \{0.2, 0.5, 0.8\}$, respectively, and the total number of families was $n(prop)/3$.

For the *unbalanced* simulations, again 5 subpopulations were simulated with total $n = \{500, 1000, 2000\}$. However, all of the families, again with $n_f = 3$, were simulated from a single subpopulation, such that 20% of the total sample size belonged to these sibships. This scenario was intentionally extreme, to determine the robustness of various methods for handling families.

4.2. *Resampled Data from the 1000 Genomes Project*

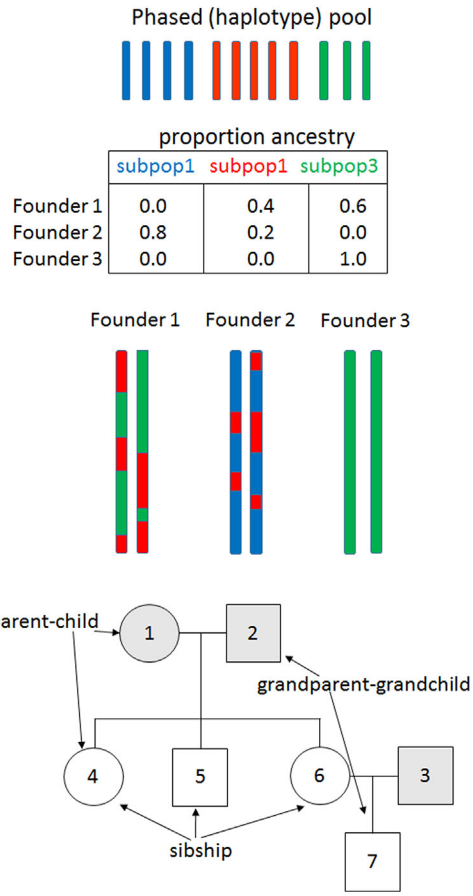
In order to represent realistic linkage disequilibrium, population structure, and additional family types, we developed novel simulation code, drawing from haploid genomes sequenced for the 1000 Genomes Project Phase 3 v5 (1000 Genomes Project Consortium, 2015). The approach is similar to that used in HAP-SAMPLE (Wright et al., 2007), but with refined data and special attention to ancestral subpopulations. A subset of 19,681 markers were chosen as representative of those that might arise from a whole genome association study (details in Web Supplement Section S7.2). After removing several subpopulations with extensive admixture, the collection of the remaining 20 subpopulations show clear evidence of continental-level stratification, as well as a range of more modest stratification among more closely related subpopulations. The simulation scheme is shown in Figure 2, and detailed in the Web Supplement. Briefly, unrelated individuals and family founders have pre-specified target proportions of their genome from each of K subpopulations, and segments are chosen from each subpopulation using an artificial meiotic recombination process. Simulation of a family of size 7 proceeds by choosing three founders, with the remaining individuals determined by the recombination process. From each family, first-degree relatives (parent-child pairs or siblings) or second-degree relatives (grandparent-grandchild pairs) can be selected.

For a pure subpopulation scenario, 1000 unrelated individuals were chosen, and 500 families. For each relative-type scenario, the parent-child pairs, sibships, or grandparent-grandchild pairs were combined with the unrelated individuals, for a total of 2000 individuals for unrelateds + parent-child, 2500 individuals for unrelateds + sibships, and 2000 individuals for unrelateds + grandparent-grandchild. The entire process was performed 10 times, and reported results averaged over the simulations.

An admixed subpopulation scenario, the process was the same as above, but with each individual showing a random admixture between two randomly selected subpopulations as described in the Web Supplement Section S7.2 (average admixture proportion about 0.09).

5. Results

The Web Supplement Figure S3 shows results for Gaussian simulations for $p = 10{,}00$ and varying proportions with unrelated individuals and “family pairs” that have correlation 0.5. The number of strata was $K = 3$, so two ancestry scores are sufficient to capture the relationships, and visual impressions can be formed. The figure illustrates that singleton projection results in extreme shrinkage of projected family members $\mathcal{F}$, while the PCAiR algorithm results in modest shrinkage of the individuals in $\mathcal{R}$. Matrix whitening shows modest shrinkage for $\mathcal{F}$, while the remaining novel methods all show good and similar performance. For highly unbalanced data with all families coming from a single subpopulation (Web Supplement Figure S4), the conclusions are similar, although shrinkage is less extreme due to a higher overall proportion of singletons. The findings are sensible, reflecting the simple fact that inclusion of an individual when computing loadings results in better performance. For family whitening, the change in cross correlations with individuals outside the family results in family shrinkage.



	subpop1	subpop1	subpop3
Founder 1	0.0	0.4	0.6
Founder 2	0.8	0.2	0.0
Founder 3	0.0	0.0	1.0

Figure 2. Simulation scheme for 1000 Genomes data. For each family of size 7, the founder genomes are simulated according to specified ancestry proportions, with artificial meioses and random haplotypes drawn from the appropriate subpopulations. The remaining family members are simulated from the founder haploid genomes, and the desired relative pairs or triplets are selected from the family. Unrelated individuals are simulated in the same manner as founders.

5.1. *Results for Idealized Simulated Genotypes*

Web Supplement Figure S5 depicts results in heatmap form for our idealized genotype simulations in which the proportion of families is balanced across the 5 strata. For the $(1 - R^2)$ criterion, matrix substitution, CPW, and family averaging appear to perform similarly and somewhat better than PCAiR. Although CPW is theoretically identical to MS, we investigated empirically, due to some reduced-rank issues that affect computation (Web Supplement Section S2).

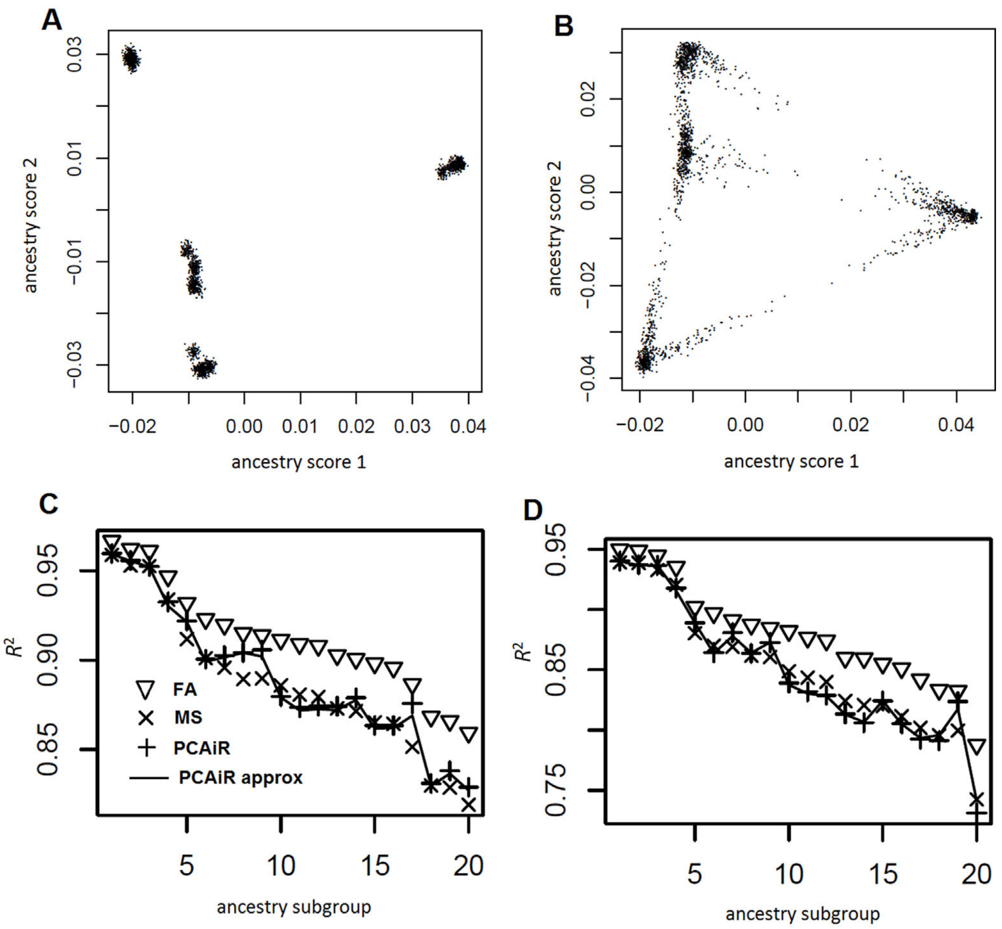


Figure 3. Results for the 1000 Genomes simulations. (A) Illustrative ancestry scores for unrelated individuals belonging to $K = 20$ subpopulations. (A) Ancestry scores where each individual has varying degrees of admixture between a randomly chosen pair of subpopulations. (C) (1-R^2) for linear predictions of each $a_{.k}$ vector (ancestry subgroup) from the first 20 columns of $\tilde{V}$ from each method, for sibships simulated under the pure subpopulation scenario. (D) (1-R^2) for linear predictions of each $a_{.k}$ vector from the first 20 columns of $\tilde{V}$ from each method, for sibships simulated under the admixed subpopulation scenario.

For the RSE criterion, differences are more noticeable, and again matrix substitution/CPW, and family averaging perform best. For large samples ($n = 2000$) and a modest proportion of family members (20%), family averaging performs best. Family whitening performs poorly.

Web Supplement Figure S6 shows the performance of the methods under the unbalanced genotype simulation with 20% of individuals belonging to families, from a single stratum. The left panel shows the $(1 - R^2)$ performance, for which family averaging offers a slight improvement over MS/CPW, followed by PCAiR. For the RSE criterion, ranking of methods is similar, with family averaging performing especially well for larger sample sizes. As expected, performance generally improves with increasing sample size.

Due to the poor performance of singleton projection and family whitening, and the fact that CPW is essentially the same as MS, for the remaining simulations we analyzed only PCAiR, Matrix Substitution, and Family Averaging.

5.2. *Results for 1000 Genomes Simulations*

For the 1000 Genomes simulations, both KING and our family-identification method performed well, correctly identifying family members or with at most two errors in all the simulations (in each instance, a family member was considered as a singleton). Thus family identification was a trivial source of variation among the methods, and we used KING with PCAiR as recommended, and our family-identification code otherwise, including the PCAiR approximation. The top panels of Figure 3 show the results for the first two eigenvectors among unrelated individuals, for pure subpopulations (panel A) and admixed subpopulations (panel B). The lower panels show the prediction R^2 results across $K = 20$ ancestry subgroups for unrelateds + sibships, using the top 20 columns of $\tilde{V}$ as predictors, although in principle only 19 should be necessary. For both the pure and admixed scenarios, the Family Average approach dominates the others. Matrix Substitution is similar to PCAiR, and our PCAiR approximation, in which a random member of each family is used and following our matrix operations, is nearly identical to PCAiR.

Table 1

Percentage reduction in ancestry-prediction $(1-R^2)$ versus PCAiR

		Distinct subpops			With admixture		
Family type	Data portion	PCAiR approx	MS	FA	PCAiR approx	MS	FA
Parent-child	Overall	−1.3	1	13.5	−1.2	1.3	14.9
	Among singletons	−1.3	3.5	4.4	−1.5	3.9	4.1
	Among families	−1.2	−1.5	22.3	−1.3	−1.5	25.5
Sibship	Overall	−1.4	−3.7	19.7	0	1.1	20.1
	Among singletons	−1.4	5.5	9.3	0.2	8.1	10.5
	Among families	−1.4	−10.1	26.9	−0.1	−3.9	27
Grandparent-grandchild	Overall	−1.3	9.8	26.4	0.1	12.2	26
	Among singletons	−1.4	8.5	8.7	0.2	10.9	11.5
	Among families	−1.2	10.9	41.1	0.1	13.3	40.3

Analysis of these profiles across the ancestry subgroups k showed a roughly constant ratio of $(1-R_k^2)$ for each method in comparison to PCAiR, justifying the average ratio $\bar{e}$ as an overall performance index. We summarize the results in Table 1, expressed as percentage reductions compared to PCAiR.

To best understand the behavior of the methods, we report reductions in the overall error $\bar{e}$ relative to PCAiR, as well as the portions attributable to each data portion of singletons and family members. The 'family member' set includes, for PCAiR, the individual used as part of set $\mathcal{U}$, and thus may understate any improvements relative to the PCAiR-projected family member(s). As expected, the major differences arise in the families. Matrix substitution performs slightly better than PCAiR overall, but performance varies, and in some instances seems worse (e.g., in sibships). Family Averaging dominates the other methods, with performance improvements among family members ranging from 22 to 41%. The greatest improvements were among the grandparent-grandchild pairs, which might be expected, because letting one family member "stand" for the other may introduce additional error as the degree of relationship becomes more distant. Perhaps surprisingly, family averaging produced an improvement of a few percent even among singletons. We believe this improvement reflects the inclusion of additional family members, which improve the loadings for all individuals.

The introduction of admixture had little effect on the overall results. As expected, our approximate PCAiR approach was nearly identical to PCAiR, with an average of about 1% higher error.

5.3. *Results for the CF Dataset*

Finally, we applied the methods to the CF dataset, using the instability index approach described earlier. To do so, we first performed 898 separate analyses of $\mathcal{S} \cup j$ for each $j \in \mathcal{F}$. We then performed 438 analyses of $\mathcal{S} \cup \mathcal{F}_f$ for each $f = 1, \ldots, 438$, and compared the two sets of analyses using the instability index, for each of the first 6 ancestry scores, chosen based on Tracy–Widom testing (Patterson et al., 2006).

The three scatterplots in Figure 4 show the results for the first and second ancestry scores using covariance matrix eigendecomposition and a single family with two siblings. The A and B panels show the position of ancestry scores when the two siblings are analyzed separately. Panel C shows the results for the entire family after matrix substitution, overplotted with the values from earlier panels, showing that they have changed little. The D panel shows the stability index values for ancestry scores 1–6 (which clearly meet significance thresholds, Corvol et al., 2015) for the various methods. Singleton projection and family whitening performed much more poorly, and are not shown. For the first four ancestry scores, matrix substitution and PCAiR performed similarly. However, for ancestry scores 5–6, PCAiR showed much higher values of the instability index. Family averaging showed considerably lower instability than PCAiR throughout, consistent with the simulation results.

5.4. *Individual Scree Plot Results*

Overall, the simulations and real data showed that the novel methods (except family whitening) dominate PCAiR and singleton projection. To gain further insight into the properties of the various methods, we examined the individual scree plots for the full CF dataset (Figure 5), with curves colored according to the size of the family (all sibships) that each individual belongs to. Panel A of Figure 5 shows the individual scree curves for the naive analysis, which simply applies SVD to the full dataset without regard to the presence of families. The colored curves (red, green, blue) show these curves for family members from families of various sizes (2, 3, 4, respectively). Although the individual scores are highly variable (see Web Supplement Figure S2), after smoothing the patterns are broadly consistent. Family members have higher values for the first components, because they tend to drive the highest-ranked ancestry scores in a naive analysis. Family members tend to have lower curves for the middle scores, because these ancestry directions are driven by the non-family members (as expected). Family members again have larger values for the last ancestry components, because these directions are driven by family component direction vectors that are *orthogonal* to the dominant family direction.

To carefully check these interpretations, we performed a simulation study using Gaussian data, with the approach described in the Web Supplement, and the numbers of each family type ($n_f = 2, 3, 4$) matching the real CF data (panel B

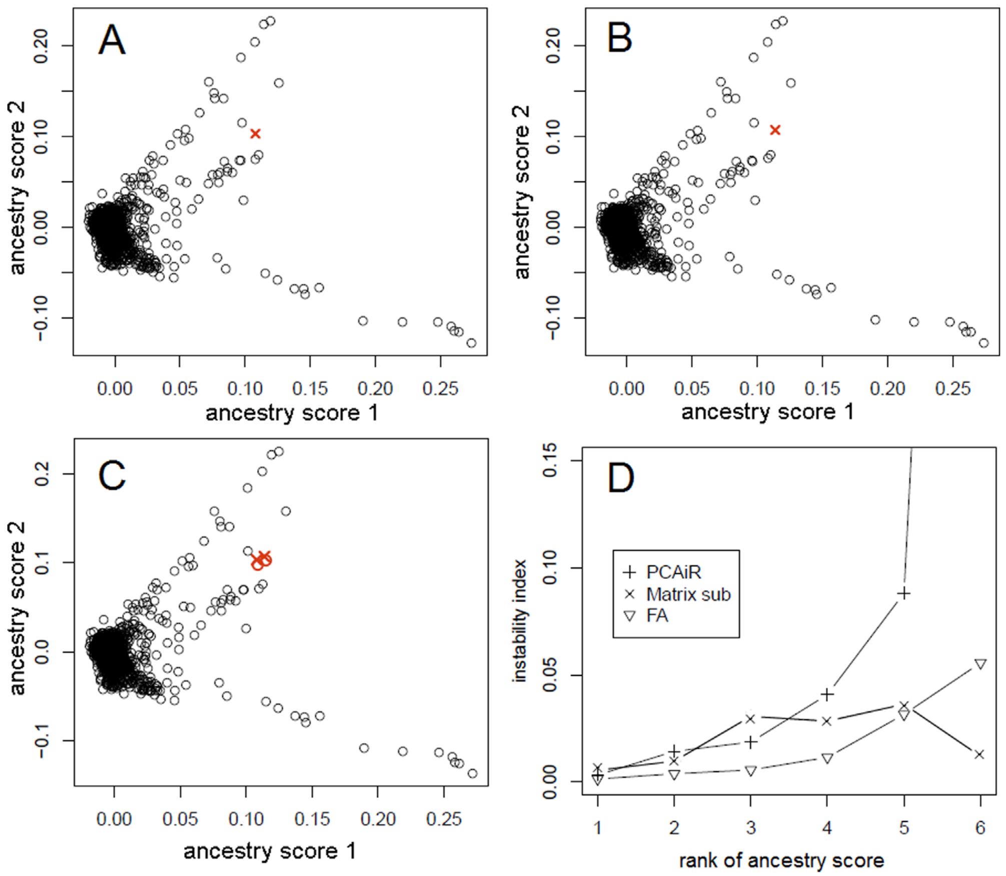


Figure 4. Illustration of the instability index for the CF dataset. A) Ancestry scores (eigenvectors of the covariance matrix) for all singletons plus the first sib in a family, marked as a red "X." B) Ancestry scores for all singletons plus the *second* sib in the family. C) Ancestry scores for matrix substitution, with the two individuals shown as circles, overplotted with values computed in A and B. D) The instability index for each method, providing a summary for each ancestry score across the 898 family members.

of Figure 5). The family patterns are very similar, although with somewhat less scatter, indicating that the geometric interpretations of these patterns are correct. Panel C shows the individual scree curves for matrix substitution, for which the curves of family members more closely overlap those of singletons. However, the curves for families of size 3 and 4 remain distinctive, as matrix substitution does not fully eliminate the effect of high correlation between family members. Panel D shows that the family average method achieves more general overlap of scree curves among the individuals.

6. Summary and Conclusions

With the CF dataset as a motivating example, we have introduced several new methods to obtain family-robust informative ancestry scores in genetic stratification analysis. Several of the methods offer improvements over the current standard, and yet are quite simple to perform using standard matrix operations in our package PCFAM. Our careful genotype simulations and analysis of the CF data support the general motivating discussion in the Web Supplement. In particular, both singleton projection and (to a lesser extent) PCAiR suffer due to the exclusion of individuals when computing loadings.

Among the new methods, family average projection has the best performance. The matrix substitution method has a potential advantage in that it relies only on the $n \times n$ covariance matrix, which is typically much smaller than the original genotype dataset. In addition, matrix substitution can be easily re-computed for different assumptions or thresholds in identifying family members or perhaps for cryptic relatedness. However, we have here examined only first- and second-degree relatives. Covariance-preserving whitening may be appealing if the resulting whitened matrix is to be used in further investigations of linkage disequilibrium structure, or perhaps in substructure analysis of individual chromosomes.

Alternative stratification control methods have included case-control modeling based on stratification scores (Epstein et al., 2007) or in the method of Song et al. (2015), which rely importantly on high-dimensional data summaries as part of the modeling procedure. Thus we foresee the methods described herein as providing useful ancestry scores for subsequent careful modeling of disease risk in combined sets of related and unrelated individuals.

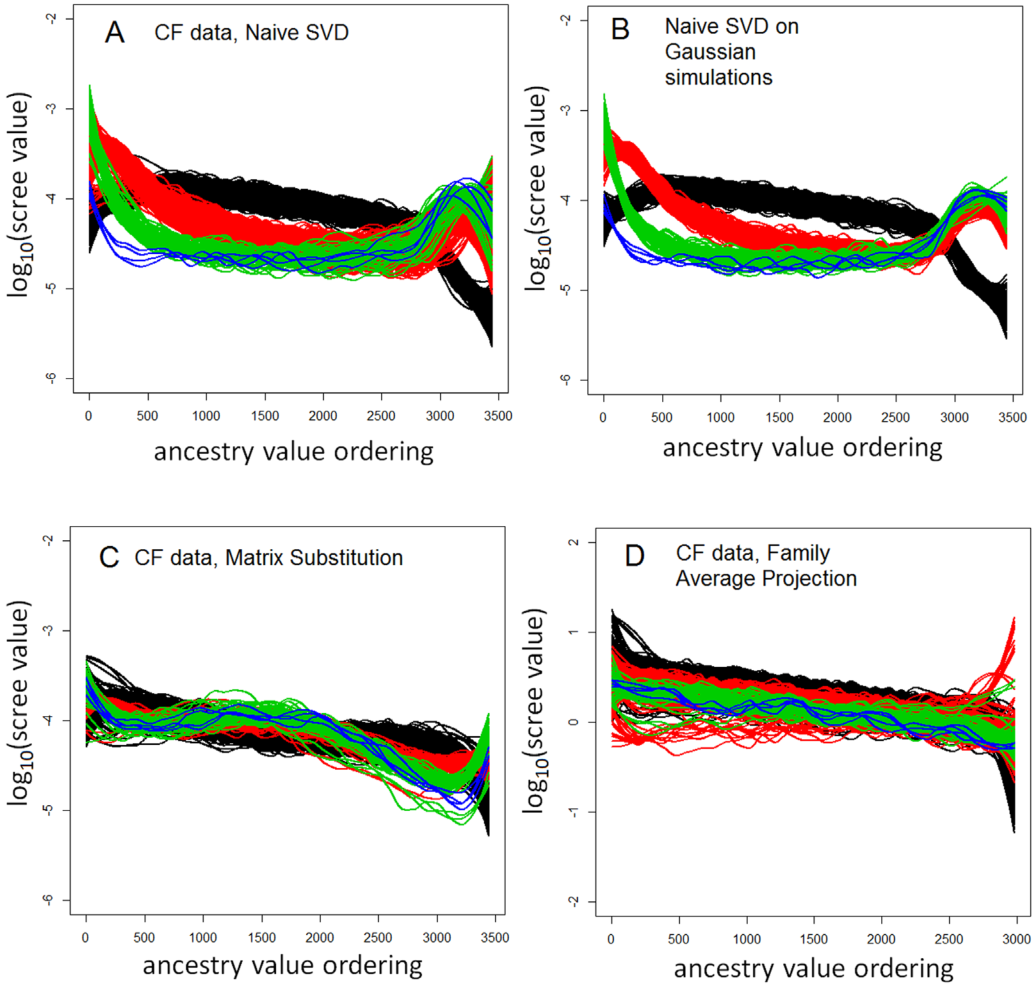


Figure 5. Individual scree plots for several methods. Black curves are for the singletons; red curves show members of families of size 2; green curves are for families of size 3 and blue curves are for the family of size 4. A) Individual scree curves for the full CF data using naive ancestry analysis, with all individuals included; B) The plot for simulated Gaussian data with the same family structure as the CF data; C) The plot for the full CF data using matrix substitution, showing that the "removal" of family effects persists through most of the ancestry values; D) The plot using the family average approach suggests further improved removal of family effects.

7. Supplementary Materials

Web Appendices and Figures referenced in Sections 2–5 are available with this article at the *Biometrics* website on Wiley Online Library. Our methods are implemented in the R package PCFAM (https://cran.r-project.org/web/packages/PCFAM/index.html).

Acknowledgements

This work supported in part by R21HG007840, EPA STAR RD83574701, and the Game-Changing Research Incentive Program at NC State University.

References

1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature* **526**, 68–74.

Cabanski, C. R., Qi, Y., Yin, X., Bair, E., Hayward, M. C., Fan, C., et al. (2010). Swiss made: Standardized within class sum of squares to evaluate methodologies and dataset elements. *PloS ONE* **5**, e9905.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research* **1**, 245–276.

Conomos, M. P., Miller, M. B., and Thornton, T. A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology* **39**, 276–293.

Corvol, H., Blackman, S. M., Boëlle, P.-Y., Gallins, P. J., Pace, R. G., Stonebraker, J. R., et al. (2015). Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nature Communications* **6**, 8382.

Epstein, M. P., Allen, A. S., and Satten, G. A. (2007). A simple and improved correction for population stratification in case-control studies. *The American Journal of Human Genetics* **80**, 921–930.

Frudakis, T., Venkateswarlu, K., Thomas, M., Gaskin, Z., Ginjupalli, S., Gunturi, S., et al. (2003). A classifier for the snp-based inference of ancestry. *Journal of Forensic Sciences* **48**, 771–782.

Lake, S. L., Blacker, D., and Laird, N. M. 2000. Family-based tests of association in the presence of linkage. *The American Journal of Human Genetics* **67**, 1515–1525.

Lee, S., Zou, F., and Wright, F. A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *Annals of Statistics* **38**, 3605.

Levine, D. M., Ek, W. E., Zhang, R., Liu, X., Onstad, L., Sather, C., et al. (2013). A genome-wide association study identifies new susceptibility loci for esophageal adenocarcinoma and barrett's esophagus. *Nature Genetics* **45**, 1487–1493.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics* **2**, e190.

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. *Nature Genetics* **38**, 904–909.

Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. 2008. Estimating local ancestry in admixed populations. *The American Journal of Human Genetics* **82**, 290–303.

Song, M., Hao, W., and Storey, J. D. (2015). Testing for genetic associations in arbitrarily structured populations. *Nature Genetics* **47**, 550–554.

Wright, F. A., Huang, H., Guan, X., Gamiel, K., Jeffries, C., Barry, W. T., et al. (2007). Simulating association studies: A data-based resampling method for candidate regions or whole genome scans. *Bioinformatics* **23**, 2581–2588.

Wright, F. A., Strug, L. J., Doshi, V. K., Commander, C. W., Blackman, S. M., Sun, L., et al. (2011). Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13. 2. *Nature Genetics* **43**, 539–546.

Zhu, X., Li, S., Cooper, R. S., and Elston, R. C. (2008). A unified association analysis approach for family and unrelated samples correcting for stratification. *The American Journal of Human Genetics* **82**, 352–365.

Received June 2016. *Revised March* 2017. *Accepted March* 2017.