Biometrics 000, 000–000 DOI: 000
000 0000

Web-based supplementary materials for "Computation of ancestry scores with mixed families and unrelated individuals"

Yi-Hui Zhou*

Bioinformatics Research Center

Department of Statistics and Biological Sciences

North Carolina State University, Raleigh, North Carolina, U.S.A.

J.S. Marron

Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, USA.

Fred A. Wright

Bioinformatics Research Center, Department of Statistics and Biological Sciences,

North Carolina State University, Raleigh, USA.

**email: yihui_zhou*@ncsu.edu

1. Family member identification

Standard results for shared genotype probabilities for related individuals are expressed in terms of kinship coefficients and identity-by-descent probabilities. Here we clarify, as is needed for this paper, the *correlation* of genotypes between (say first-degree) relatives. We focus on siblings, although a slight modification of the argument applies to parent-child relationships. Let q denote the minor allele frequency, and a pair of siblings have random genotypes g_1 and g_2, with means $2q$ and variances $2q(1-q)$. We have

$$corr(g_1, g_2) = \frac{E(g_1 g_2) - E(g_1)E(g_2)}{SD(g_1)SD(g_2)} = \frac{E(g_1 g_2) - (2q)^2}{2q(1-q)}.$$

The identity-by-descent (IBD) outcomes determine $E(g_1 g_2)$. For IBD=0, $E(g_1 g_2 | IBD = 0) = (2q)^2$. Also, $E(g_1 g_2 | IBD = 2) = E(g_1^2) = 2(\text{var}(g_1) + E(g_1)^2) = 2q(1-q) + (2q)^2$. If IBD=1, without loss of generality, we assume the shared allele comes from the mother. We use a_m to denote the allele from the mother and a_f from the father. Then $E(g_1 g_2) = E((a_{m1} + a_{f1})(a_{m1} + a_{f2})) = E(a_{m1}^2 + a_{f1}a_{m1} + a_{m1}a_{f2} + a_{f1}a_{f2}) = q + 3q^2$. Therefore

$$\begin{aligned} E(g_1 g_2) &= E(E(g_1 g_2 | IBD)) \\ &= \frac{1}{4}E(g_1 g_2 | IBD = 0) + \frac{1}{2}E(g_1 g_2 | IBD = 1) + \frac{1}{2}E(g_1 g_2 | IBD = 2) \\ &= \frac{1}{4}(2q)^2 + \frac{1}{2}(2q(1-q) + (2q)^2) + \frac{1}{4}(q + 3q^2) \end{aligned}$$

and plugging in to the correlation gives $1/2$, regardless of q. A similar approach can be used to show that the corrrelation between second-degree relatives is $1/4$.

All of our proposed methods can use families identified by external software such as KING (Manichaikul et al., 2010), but we found it convenient to utilize a method within our same R environment. For row-scaled genotype matrix X, we expect family members to have high correlation. However, population stratification distorts interpretation, so we first compute the $n \times n$ correlation matrix of X, then forcing the maximum off-diagonal value to be $1/8$. The spacing of successive eigenvalues from the resulting clipped correlation matrix are

examined on the log-scale, identifying the matrix P of the top k eigenvectors that exceed 4 standard deviations from the mean spacing. The matrix $X_{resid} = X - XPP^T$ is residualized for gross population stratification, and then X_{resid} is again row-scaled to obtain X', and families identified using entries for which the $n \times n$ correlation matrix of X' exceed η. We use $\eta = 0.1$ in this paper, effectively identifying first- and second-degree relatives. More careful investigation of higher-degree or cryptic relationships can use software for this purpose, beyond our intended scope.

2. Covariance-Preserving Whitening (CPW)

Here we describe an approach to modify the genotypes within families so that the final covariance matrix equals the modified covariance matrix $\widetilde{M}$ used for the matrix substitution approach, and families are orthogonalized while retaining their covariance with the remaining sample. The goal here is to find an $n \times n$ matrix B such that $Y = XB^T$ and $\frac{1}{p-1}Y^TY = \frac{1}{p-1}B\bar{X}^T\bar{X}B^T = \widetilde{M}$, where the entire sample, including all families, is handled at once. There are multiple possible solutions, but it is appealing to add the constraint that only family members be modified, as singletons do not contribute to the problem of "spurious" ancestry scores. We assume that the columns of X are arranged with singletons $\mathcal{S}$ followed by families $\mathcal{F}$. We then divide M (defined above) and B^T into submatrices as follows,

$$\underset{n\times n}{M} = \begin{bmatrix} \underset{(n-n_{\mathcal{F}})\times(n-n_{\mathcal{F}})}{M_{11}} & \underset{(n-n_{\mathcal{F}})\times n_{\mathcal{F}}}{M_{12}} \\ \underset{n_{\mathcal{F}}\times(n-n_{\mathcal{F}})}{M_{21}} & \underset{n_{\mathcal{F}}\times n_{\mathcal{F}}}{M_{22}} \end{bmatrix}, \quad \underset{n\times n}{B}^T = \begin{bmatrix} \underset{(n-n_{\mathcal{F}})\times(n-n_{\mathcal{F}})}{I_{n-n_{\mathcal{F}}}} & \underset{(n-n_{\mathcal{F}})\times n_{\mathcal{F}}}{C} \\ \underset{n_{\mathcal{F}}\times(n-n_{\mathcal{F}})}{0} & \underset{n_{\mathcal{F}}\times n_{\mathcal{F}}}{D} \end{bmatrix},$$

where $n_{\mathcal{F}}$ individuals belong to the all-families set $\mathcal{F}$, and $I_{n-n_{\mathcal{F}}}$ denotes an $(n-n_{\mathcal{F}})\times(n-n_{\mathcal{F}})$ identity matrix. Note that $\widetilde{M}$ differs from M in only the co-family pairs of the lower right submatrix, and we will use $\widetilde{M}_{22}$ to denote the corresponding $n_{\mathcal{F}} \times n_{\mathcal{F}}$ lower right submatrix of $\widetilde{M}$. The form of B^T, with the identity submatrix operating on the singletons in $Y = XB^T$, achieves the desired constraint that singletons be unchanged. C and D are unknown matrices,

to be solved for. We show below that the solution for full-rank X is

$$C = M_{11}^{-1}M_{12}(I_{n-n_{\mathcal{F}}} - D), \quad D = (M_{22} - S)^{-1/2}(\widetilde{M}_{22} - S)^{1/2},$$

where $S = M_{12}^T(M_{11}^{-1})^T M_{12} = M_{21}M_{11}^{-1}M_{12}$, with a slight modification to account for our situation that X has rank $n-1$. For $Y = XB^T$, ancestry scores are obtained as $\widetilde{V}_{CPW}$ in the SVD $Y = \widetilde{U}\widetilde{D}\widetilde{V}_{CPW}^T$. Figure S1 (right panel) shows the result of applying covariance-preserving whitening to the CF data. The plot shows that, for the new matrix Y, cross-correlations of families $\mathcal{F}$ to singletons $\mathcal{S}$ have indeed been preserved from the original X. In fact even the correlations between members of different families have been preserved (not shown).

2.1 *The Covariance-Preserving Whitening Solution*

We have

$$AMA^T = \begin{bmatrix} I_{n-n_{\mathcal{F}}} & 0^T \\ C^T & D^T \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} I_{n-n_{\mathcal{F}}} & C \\ 0 & D \end{bmatrix}$$

$$= \begin{bmatrix} M_{11} & M_{11}C + M_{12}D \\ C^TM_{11} + D^TM_{21} & \underbrace{C^TM_{11}C}_{a} + \underbrace{C^TM_{12}D}_{b} + \underbrace{D^TM_{21}C}_{c} + \underbrace{D^TM_{22}D}_{d} \end{bmatrix}$$

$$= \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & \tilde{M}_{22} \end{bmatrix}.$$

Comparing the last two expressions provides two equations in the two unknowns C and D. From the upper right, we have $M_{11}C + M_{12}D = M_{12}$, which implies $C = M_{11}^{-1}M_{12}(I_2 - D)$ (the lower left is the same equation written in transpose form). The lower right requires a bit more effort. We consider each of the four terms separately, plugging in the solution for C from above, giving

$$a = C^TM_{11}C = (M_{11}^{-1}M_{12}(I_2 - D))^T M_{11}M_{11}^{-1}M_{12}(I_2 - D)$$

$$= (I_2 - D)^T M_{12}^T(M_{11}^{-1})^T M_{12}(I_2 - D) = (I_2 - D)^T S(I_2 - D),$$

where $S = M_{12}^T (M_{11}^{-1})^T M_{12} = M_{21} M_{11}^{-1} M_{12}$.

$$b = C^T M_{12} D = (I_2 - D)^T M_{12}^T (M_{11}^{-1})^T M_{12} D = (I_2 - D)^T S D$$

$$c = D^T M_{21} C = D^T M_{21} M_{11}^{-1} M_{12} (I_2 - D) = D^T S (I_2 - D),$$

and d does not simplify. We have

$$a + b + c + d = (I_2 - D)^T S (I_2 - D) + (I_2 - D)^T S D + D^T S (I_2 - D) + D^T M_{22} D = \tilde{M}_{22},$$

and the expression reduces to $D^T (M_{22} - S) D = \tilde{M}_{22} - S$. Thus, finally, we have our solution

$$C = M_{11}^{-1} M_{12} (I_2 - D), \quad D = (M_{22} - S)^{-1/2} (\tilde{M}_{22} - S)^{1/2}.$$

The final expression is as desired, preserving singletons while rotating only the family members. The solution is unique if $X^T X$ is of full rank n. However, in our treatment, X has been row-centered, so no exact solution exists. To prove this by contradiction, suppose A exists such that $AMA^T = \tilde{M}$. When X is row-centered, M has rank $n - 1$, and the rank of the left-hand side cannot exceed $n - 1$. However, when matrix substitution is implemented in practice, the resulting $\tilde{M}$ typically has rank n, creating a contradiction. In practice, when X has been row-centered, we add a small value $\delta = 0.001$ to the diagonal of M before proceeding, which provides similar results to using a Moore-Penrose generalized inverse when solving C and D. Either approach results in $\frac{1}{p-1} Y^T Y$ as a close approximation to $\widetilde{M}$ in simulations and for the real CF data.

3. Geometric Rotation / Family Whitening (FW)

We first describe the problem in geometric terms, to gain an understanding of the nature of the modification, and follow with the simple matrix operation analogue. Our solution is to *rotate* the data to make individuals within a family orthogonal, performed within a plane such that the impact of the data rotation is otherwise minimal. The approach is easiest to explain for a family of size 2, and the data for each individual is the scaled genotype p-vector. Data vectors for first-degree relatives are expected to have a 60° angle, corresponding to a genotype

correlation of 0.5 (Appendix A). We first find the mean vector of the two members, and then rotate each member away from the mean vector to a target angle of 45°. This operation makes the new vectors orthogonal, which is approximately true for unrelated individuals.

In general, a family f consists of n_f individuals indexed by the set $\mathcal{F}_f$. The target rotation angle θ_f is the same as the angle in $\mathbb{R}^{n_f}$ between each coordinate unit vector and the direction vector $(\frac{1}{\sqrt{n_f}}, ..., \frac{1}{\sqrt{n_f}})^T$, which is $\theta_f = arccos(\frac{1}{\sqrt{n_f}})$. For example, when $n_f = 2$, $\theta_f = arccos\frac{1}{\sqrt{2}} = \pi/4$. Let $x_{.j}$ denote the data vector for individual j, with unit-length vector $z_j = \frac{x_{.j}}{||x_{.j}||}$, where $||x_{.j}||$ is the length $\sqrt{\sum_i x_{ij}^2}$. The mean vector $\bar{x}_{\mathcal{F}_f}$ is obtained by computing for each SNP i $\bar{x}_{i\mathcal{F}_f} = \sum_{j\in\mathcal{F}_f} x_{ij}/n_f$, with unit vector $\bar{z}_{\mathcal{F}_f} = \frac{\bar{x}_{\mathcal{F}_f}}{||\bar{x}_{\mathcal{F}_f}||}$. The unit length component of z_j which is orthogonal to $\bar{z}_{\mathcal{F}}$ is

$$\widetilde{z_j} = \frac{z_j - (z_j^T \bar{z}_{\mathcal{F}_f})\bar{z}_{\mathcal{F}_f}}{||z_j - (z_j^T \bar{z}_{\mathcal{F}_f})\bar{z}_{\mathcal{F}_f}||}.$$

In the plane determined by $\bar{z}_{\mathcal{F}}$ and $\widetilde{z_j}$, the unit vector with angle θ_f to $\widetilde{z_j}$ is $\widetilde{\mu_j} = \cos(\theta_f)\widetilde{z_j} + \sin(\theta_{n_f})\widetilde{z_j}$. The vector $\widetilde{x_{.j}} = \widetilde{\mu_j}||x_{.j}||$ is the natural rescaling of $\widetilde{\mu_j}$, and used as a replacement data vector for $x_{.j}$. Finally the data vector for each family member is centered and rescaled to match the mean and variance of the original data. This rotation operation is conducted in succession for each family $f = 1, ..., F$, and SVD is applied to the new whitened data matrix.

Geometric rotation has a matrix operation interpretation, de-correlating the members of a family f by an operation similar to classical multivariate sphering. Let $Z_{\mathcal{F}_f}$ be the $p \times n_f$ submatrix of scaled family genotype data, and $R_{\mathcal{F}_f}$ the corresponding (positive definite) $n_f \times n_f$ matrix of sample correlations. Then $\widetilde{Z}_{\mathcal{F}_f} = R_{\mathcal{F}_f}^{-1/2} Z_{\mathcal{F}_f}$ is a whitened matrix with identity correlation, and a final $\widetilde{X}_{\mathcal{F}_f}$ is obtained by recentering and scaling the columns of $\widetilde{Z}_{\mathcal{F}_f}$ to match the mean and variance of the original $X_{\mathcal{F}_f}$. Finally, the columns of singletons and newly whitened family data are combined into $\underset{p\times n}{\widetilde{X}} = [\underset{p\times n_{\mathcal{S}}}{X_{\mathcal{S}}}, \underset{p\times n_{\mathcal{F}}}{\widetilde{X}_{\mathcal{F}}}]$, and the ancestry scores are $\widetilde{V}_{FW}$ from the SVD $\widetilde{X} = \widetilde{U}\widetilde{D}\widetilde{V}_{FW}^T$.

In practice, geometric rotation and matrix whitening of the family are nearly identical,

with slight differences due to handling of column centering, and the matrix approach is used subsequently. Figure S1 (left panel) shows the result of family whitening in the CF dataset, in terms of the correlation of columns of $\widetilde{X}_{\mathcal{F}}$ compared to those of $X_{\mathcal{S}}$. This shows that the family whitening operation introduces some perturbation of the correlation structure. We will return to this issue below.

[Figure S1 about here.]

4. Rationale for predicting true ancestry from ancestry scores for association covariate control

We suppose Z is an $n \times p$ contains the genotype data at a SNP, as well as any important covariates, such as age, sex, etc., for a total of p covariates including the intercept. Let A be the (unobserved) true ancestry matrix for q strata, with each entry representing the proportion of each individual's genome from each of the q strata. In the special case of no admixture, then A has binary entries. A standard linear model would be

$$\underset{n\times 1}{Y} = \underset{n\times p}{Z}\ \underset{p\times 1}{\beta} + \underset{n\times q}{A}\ \underset{q\times 1}{\alpha} + \underset{n\times 1}{E}$$

where β and α are coefficients and E are random errors. Now, A is unobserved, but we believe is represented by the q-d subspace generated by the ancestry scores V, then

$$\underset{n\times q}{A} = \underset{n\times q}{V}\ \underset{q\times q}{C} \tag{5.1}$$

where C is a coefficient matrix representing "regression" of each of the q strata on the q ancestry scores. Following this assumption,

$$\begin{aligned}\underset{n\times 1}{Y} &= \underset{n\times p}{Z}\ \underset{p\times 1}{\beta} + \underset{n\times q}{V}\ \underset{q\times q}{C}\ \underset{q\times 1}{\alpha} + \underset{n\times 1}{E}\\ &= \underset{n\times p}{Z}\ \underset{p\times 1}{\beta} + \underset{n\times q}{V}\ \underset{q\times 1}{\delta} + \underset{n\times 1}{E},\end{aligned}$$

where $\delta = C\alpha$ is a coefficient vector. An investigator has access to observed Z and V, and

fits the final model. As long as (5.1) holds, then the regression is valid, even though we don't observe A directly.

The rationale above holds for generalized linear models, and thus includes logistic modeling, for example for case-control modeling. In practice, the mapping of V to A is not perfect, and even a relatively small error in prediction of A from V can degrade the covariate control effectiveness. However, (5.1) does serve as a rationale that the ability to linearly A from V is key, which can be summarized by multiple regression R^2 for each column of A from matrix V.

5. Individual scree plot illustration

The individual scree plots use loess smoothing of the individual scree values. Figure S2 shows the individual scree values for representative members of the singleton set (upper left panel), a member of a family of size 2 (upper right), a member of a family of size 3 (lower left), and a member of a family of size 4 (lower right). Although the individual scree values are noisy, after smoothing they produce curves that are characteristic of the family size as shown in Figure 6 of the main manuscript.

[Figure S2 about here.]

6. Gaussian simulations

We first illustrate the principles for our procedures using simple Gaussian simulations, so that the pure effect of individual-individual covariance structures can be demonstrated. Each simulated dataset consists of $p = 10,000$ markers and $n = 800$ individuals, including $n_{\mathcal{S}} = 200$ singletons, and $n_{\mathcal{F}} = 600$ individuals arising from 300 family pairs. An initial $p \times n$ error matrix E was first generated, reflecting family correlation structure but not population structure. For each singleton, the corresponding length-p column was generated as

independent $N(0,1)$ entries. For the (arbitrary) first member of a family, the corresponding column $e_{\cdot 1}$ was generated in the same manner, for each marker i as $e_{i1} \sim N(0,1)$. Then the second member of the same family was generated for each i as $e_{i2} = 0.5e_{i1} + \epsilon_i$, where $\epsilon_i \sim N(0, .75)$. Thus finally all entries were marginally $N(0,1)$ and family pairs had correlation 0.5, corresponding to "first-degree" relatives.

Population structure was generated by simulating p-vectors $P_k \sim N(0, .01)$, where $k = 1, 2, 3$, and randomly assigning each singleton and each family to a subpopulation with probability 1/3. A final matrix X was generated by adding, for each individual, the corresponding column of E to the appropriate subpopulation column vector from P, and X was again row-scaled.

We divided the family pairs into an arbitrary partition (set 1 vs. set 2), such that no members were related within a set. The various methods were represented as follows, and shown in Figure S3 (singletons in black, set 1 in gold, set 2 in cyan). Methods with good performance should show similar behavior for singletons and family members. (a) Singleton-based projection, with left singular vectors (loadings) based on the singletons only, shows extreme shrinkage of family members. (b) Using singletons plus family set 1 to obtain loadings shows modest shrinkage for set 2. For simple clustered stratification, there is no conceptual advantage to optimizing the choice of maximal unrelated set, and thus the results here are analogous to the PCAiR method. (c) Per-family whitening, which uses eigenvectors of the resulting covariance matrix, shows shrinkage of families. (d) The family-average approach performs SVD of the augmented data matrix consisting of singletons and the family averages, and shows good performance. (e) The covariance-preserving whitening approach also shows good performance, and is nearly identical to (f), the result from matrix substitution using zero as the replacement values.

Similar results can be observed for an unbalanced simulation, in which the total sample

size is fixed at 800, but 100 families all belong to a single population stratum (Figure S4). Here the shrinkage is not as strong, as the number of family members is a smaller proportion of the total sample size.

[Figure S3 about here.]

[Figure S4 about here.]

7. Simulation of genotypes

7.1 *Idealized simulations*

We simulated genotype data in a manner that respected local correlation structure, which is present but typically modest in SNPs used for stratification control, and reflected population ancestry. A SNP "block size" of 20 was chosen. An autoregressive normal model was used to simulate a set of modestly underlying correlated values, e.g. for one individual the value for the ith SNP is $Z_i = \rho Z_{i-1} + \epsilon$, where $\epsilon \sim N(0, 1-\rho^2)$, followed by reversal of sign of ρ with probability 0.5. Marginally, each $Z_i \sim N(0,1)$, and a modest $\rho = 0.2$ was used within each block and $\rho = 0$ at block boundaries, so that values across different blocks were uncorrelated. To convert the values to genotypes, we first generated random minor allele frequencies by drawing "ancestral" allele frequencies from the half-triangular distribution $f(x) = 2(x-a)/(a-b)^2$, where $a = 0.38, b = 0.50$, which corresponded closely to the observed minor allele frequency in the thinned CF dataset. For ancestral minor allele frequency q, the Balding-Nichols model was used for fixation index F_{ST} by drawing K subpopulation allele frequencies from the beta distribution with parameters $q(1-F_{ST})/F_{ST})$, and $(1-q)(1-F_{ST})/F_{ST}$. Conversion of the latent Z values to genotypes was performed by applying, for each SNP and individuals in subpopulation k with allele frequency q_k, an inverse quantile of ranked z-values such that the lowest z values were converted to genotype 0, the largest to

genotype 2, and genotypes 0, 1, and 2 occurred with frequencies $(1-q_k)^2$, $2q_k(1-q_k)$, and q_k^2 (i.e. Hardy-Weinberg equilibrium within each subpopulation k).

[Figure S5 about here.]

[Figure S6 about here.]

7.2 *Sampling from 1000 Genomes data*

Phased haplotype data were obtained from 1000 Genomes Phase3 v5, with 81.2 million markers (http://www.internationalgenome.org). In order to mimic a SNP array platform, we selected the subset of 2.1 million markers that overlapped with the HumanOmni2.5-8 v1.2 BeadChip. Using PLINK (Chang et al., 2015), we applied a filter of minor allele frequency (MAF)>0.25 and pairwise-LD $r^2<0.02$ (PLINK parameters: –maf 0.25 –indep-pairwise 50 5 0.02), which further narrowed the result to 19,681 ancestry-informative markers.

From the original 2504 individuals, inspection of the data revealed that subpopulations ACB, ASW, and AMR demonstrated high admixture. Thus we excluded these subpopulations, leaving 2000 individuals with 4000 phased autosomal haploid genomes as a haploid "pool," for 20 subpopulations: BEB (86 individuals), CDX (93), CEU (99), CHB (99), CHS (105), ESN (99), FIN (99), GBR (91), GIH (103), GWD (113), IBS (107), ITU (102), JPT (104), KHV (99), LWK (99), MSL (85), PJL (96), STU (102), TSI (107), and YRI (108). For each subpopulation the number of haploid genomes is twice the number if individuals.

To create realistic "pure" subpopulations, we followed the basic approach of HAP-SAMPLE Wright et al. (2007), in which phased autosomal haplotypes were selected from the pool of 4000 haploid genomes, and subjected to an artificial recombination process to create new haploid genomes. The new haploid genomes were then used as unrelated individuals, or as founder individuals in each 7-individual family as described in the main text.

The artificial recombination process followed a simple 1Mb=1cM assumption, for an approximately 30 Morgan genome. In any one "meiosis" and for any interval between succ-

cessive SNPs, crossover events are rare enough that this simple genetic map approach was effective and preserved the linkage disequilibrium structure, while adding novelty in terms of the genomes that could be generated. For each genome to be simulated, a target probability profile was selected, representing the proportion of the genome to be selected from each of the $K = 20$ subpopulations. For each simulated crossover event, the appropriate subpopulation for that portion of a simulated chromosome was chosen as a multinomial outcome, according to the target probability profile. Once the subpopulation was chosen, the particular haploid segment was chosen at random from among all haploids in the corresponding subpopulation pool.

For the pure subpopulation scenario, each unrelated individual and family founder was given a target probability of 1 for once of the $K = 20$ subpopulations, with a probability of $1/K$ for each subpopulation. The original individuals in the pool show some within-subpopulation variation, to varying degrees depending on each subpopulation. However, the simulation process produces individuals with no such variation, except for stochastic sampling variability due to the haploid segments.

For the admixture scenario, a single subpopulation was chosen for each individual as above, and then a secondary subpopulation at random from among the remaining $K - 1$ subpopulations. The proportion of the genome from the first and second subpopulations was then assigned as $1 - p$, p, respectively, where p was chosen from a beta(1,10) density (i.e. mean admixture =1/11, and about 10% of the individuals showed admixture greater than 20%).

The above approach was used to simulate all unrelated individuals, as well as the three founders in each 7-individual family. The remaining members of each family were simulated using the same artificial recombination process, but based on the haplotypes of their parents.

References

C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1): 7, 2015.

A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W.-M. Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

F. A. Wright, H. Huang, X. Guan, K. Gamiel, C. Jeffries, W. T. Barry, F. P.-M. de Villena, P. F. Sullivan, K. C. Wilhelmsen, and F. Zou. Simulating association studies: a data-based resampling method for candidate regions or whole genome scans. *Bioinformatics*, 23(19):2581–2588, 2007.

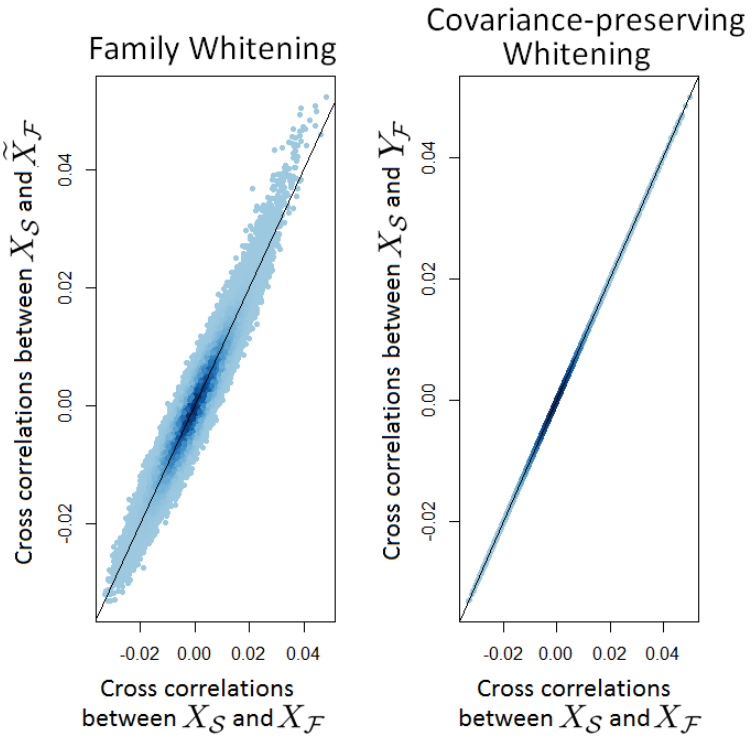

Figure S1: Cross-correlations between genotype vectors of set $\mathcal{S}$ vs. $\mathcal{F}$ for the cystic fibrosis data. On each axis, a point represents a correlation between an individual in $\mathcal{S}$ to an individual in $\mathcal{F}$, for a total of 2546× 898 points. Left panel: Cross correlations of $X_\mathcal{S} \times \widetilde{X}_\mathcal{F}$ vs. cross correlations of $X_\mathcal{S} \times X_\mathcal{F}$ show modest deviation. Right panel: Cross correlations of $X_\mathcal{S} \times Y_\mathcal{F}$ vs. cross correlations of $X_\mathcal{S} \times X_\mathcal{F}$ show that the goal of covariance-preserving whitening is achieved.

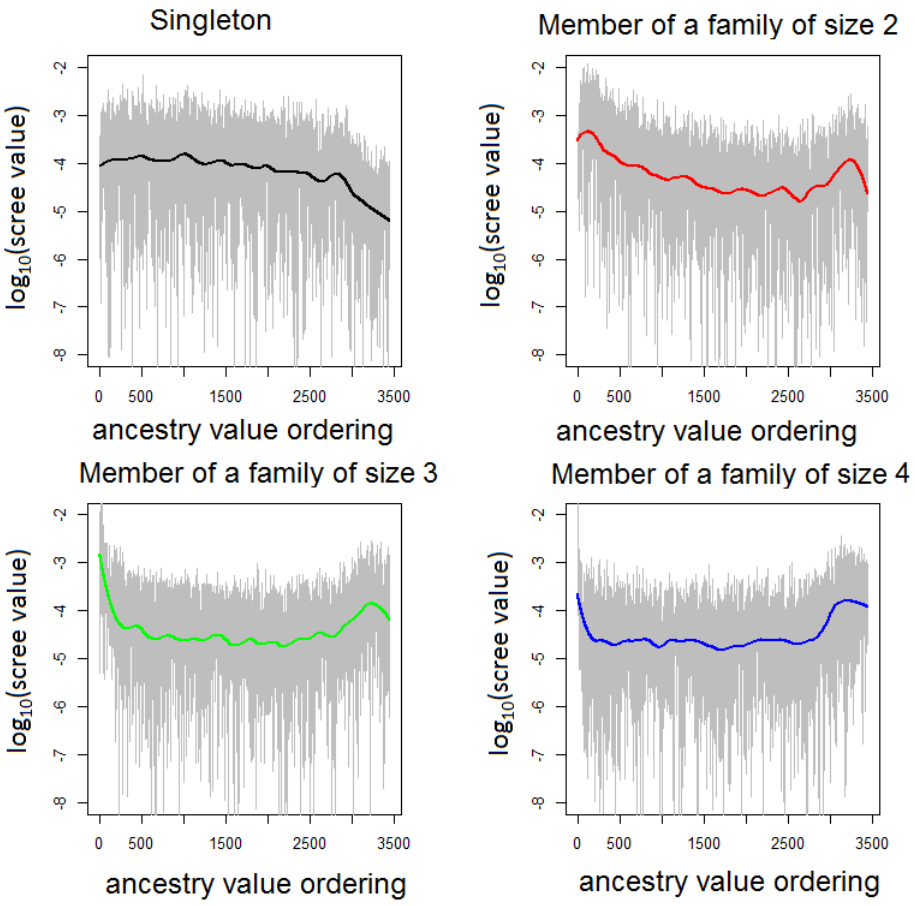

Figure S2: Several representative individuals, with noisy individual scree curves (grey), and corresponding loess fits with colors following the same scheme as in Figure 6 of the main manuscript.

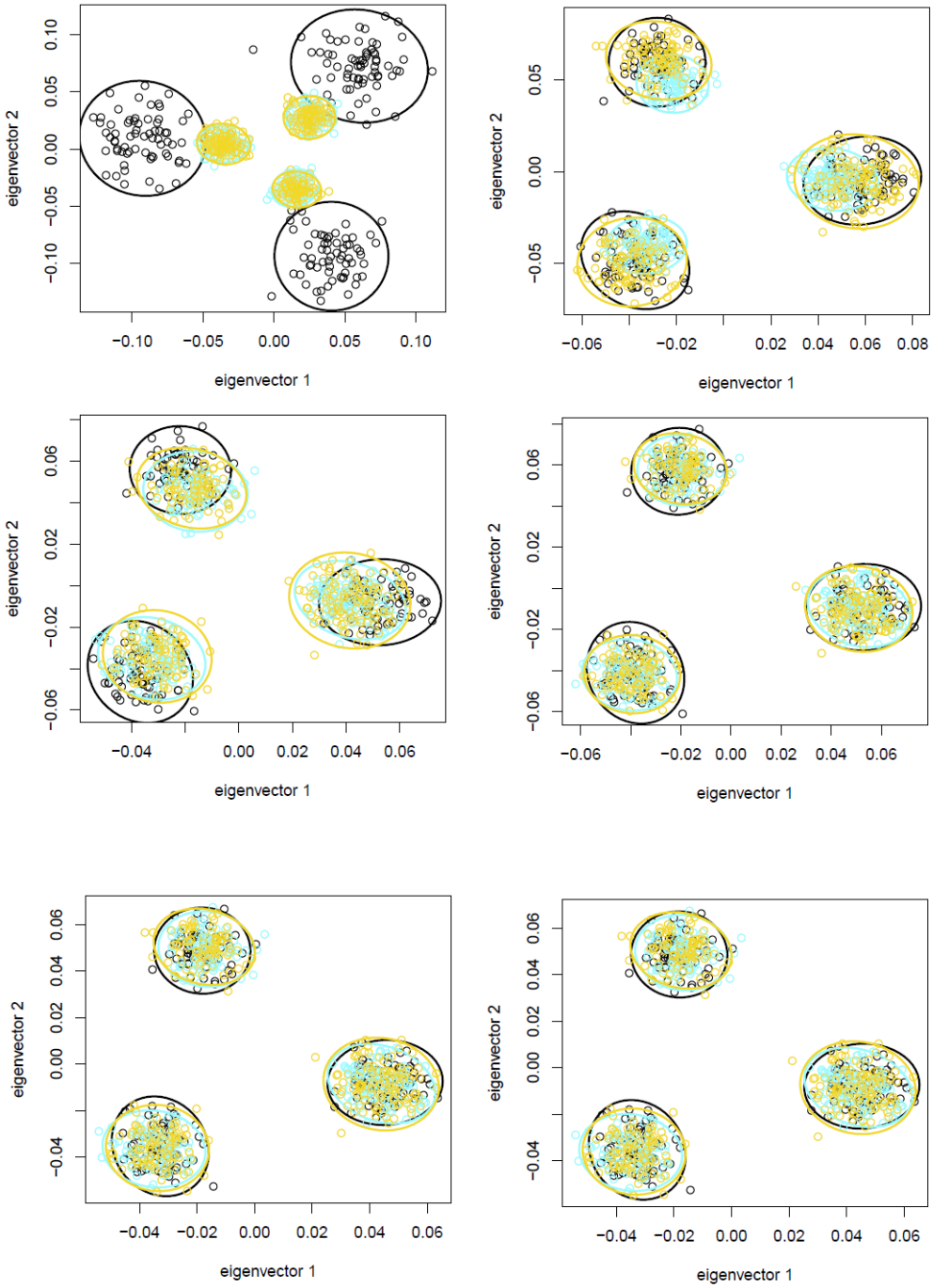

Figure S3: Illustration of family shrinkage for various methods, idealized normal data with 200 singletons (black) and 300 family pairs, divided into set 1 (gold) and set 2 (cyan). The first two eigenvectors are shown, with empirical 95% normal confidence ellipses for each subpopulation. (a) Singleton-based projection. (b) Using singletons plus family set 1 to obtain loadings. (c) Per-family whitening. (d) The family-average approach. (e) Covariance-preserving whitening. (e) Matrix substitution.

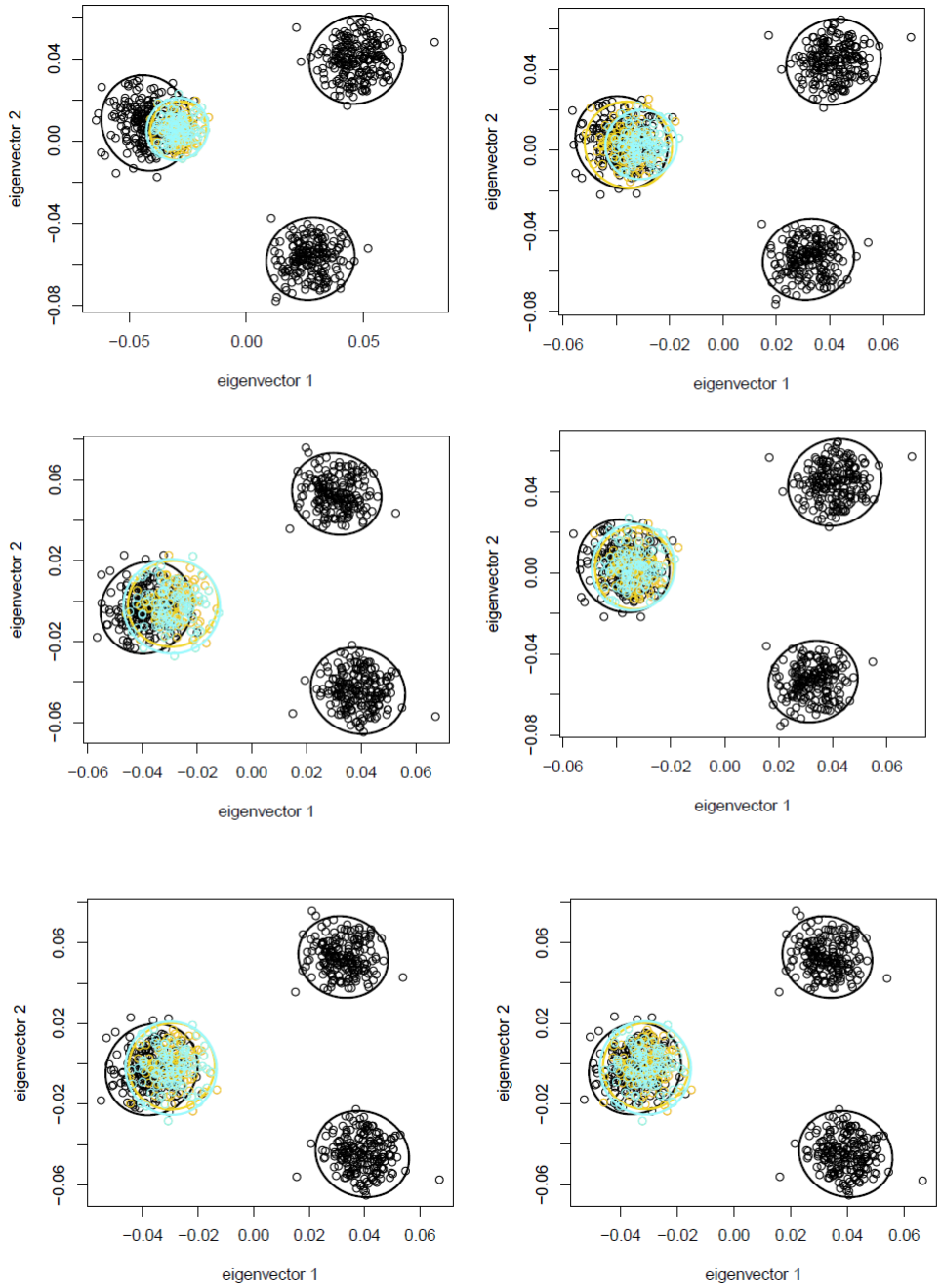

Figure S4: Illustration of family shrinkage for unbalanced data, with 600 singletons (black) and 100 family pairs residing in a single population stratum, divided into set 1 (gold) and set 2 (cyan). The first two eigenvectors are shown, with empirical 95% normal confidence ellipses for each subpopulation. (a) Singleton-based projection. (b) Using singletons plus family set 1 to obtain loadings. (c) Per-family whitening. (d) The family-average approach. (e) Covariance-preserving whitening. (e) Matrix substitution.

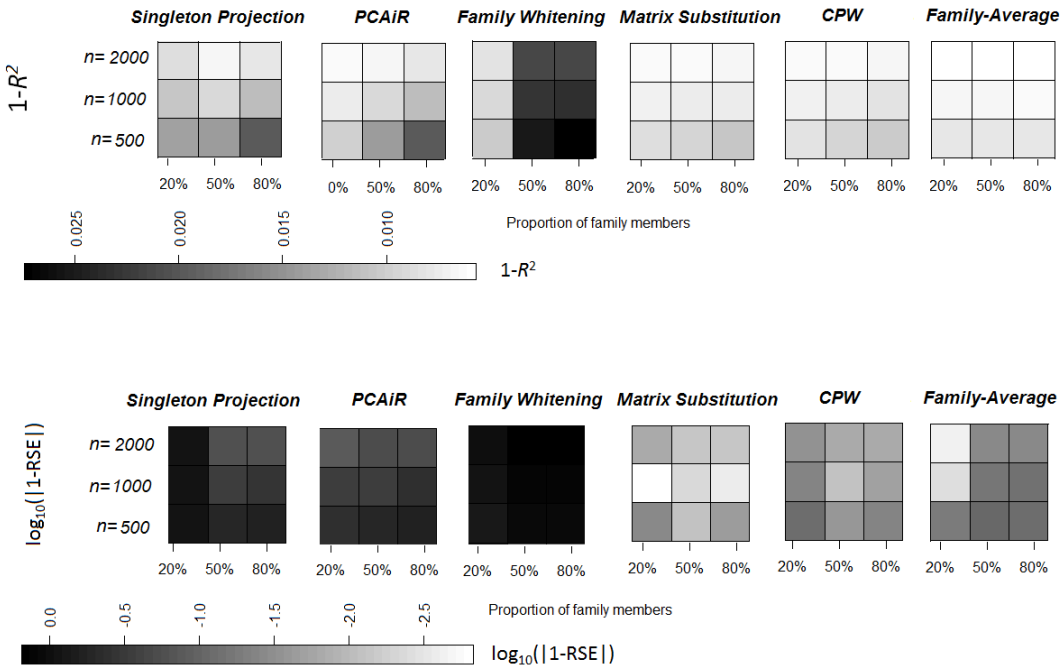

Figure S5: Heatmap for 1-R^2 and RSE performance, for the balanced simulations with the same proportion of family members in each of $K = 5$ subpopulation strata.

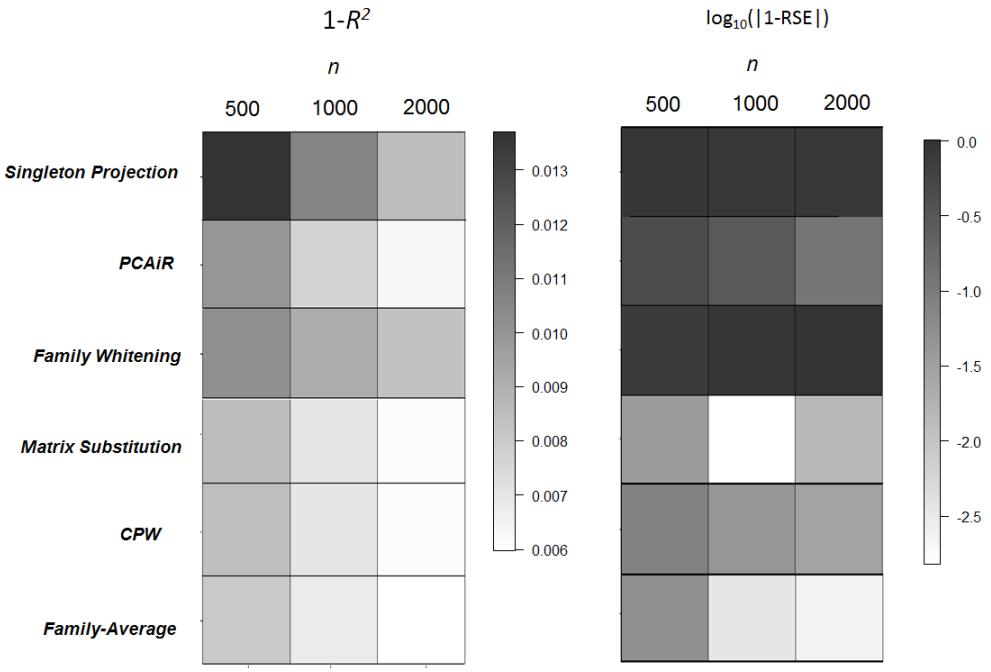

Figure S6: Heatmap for 1-R^2 and RSE performance, for the unbalanced simulations with 20% of the sample consisting of family members in a single stratum.