A unified linear mixed model for familial relatedness and population structure in genetic association studies

Nicholas Devogel, Paul Auer, Regina Manansala, Andrea Rau, Tao Wang

A unified linear mixed model for familial relatedness and population structure in genetic association studies

Nicholas Devogel[1], Paul L. Auer[2], Regina Manansala[2], Andrea Rau[2,3], Tao Wang[1]

July 13, 2020

[1]Division of Biostatistics, Institute for Health and Equity, Medical College of Wisconsin, Milwaukee, WI

[2]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI

[3]Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350, Jouy-en-Josas, France

Running title: A LMM for familial relatedness and population structure

Correspondence

Tao Wang, Division of Biostatistics, Institute for Health and Equity, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226-0509.

Email: taowang@mcw.edu

Funding information

UK Biobank Resource, Project number 19746. EU in the framework of the Marie-Curie FP7 COFUND People Programme, through the award of an AgreenSkills/AgreenSkills+ fellowship under grant agreement number 609398.

Abstract

Familial relatedness (FR) and population structure (PS) are two major sources for genetic correlation. In human population, both FR and PS can further breakdown into additive and dominance components to account for potential additive and dominance genetic effects. In this study, besides the classical additive genomic relationship matrix, a dominance genomic relationship matrix is introduced. A link between the additive/dominance genomic relationship matrices and the cocancestry (or kinship)/double coancestry coefficients is also established. In addition, a way to separate the FR and PS correlations based on the estimates of coancestry and double coancestry coefficients from the genomic relationship matrices is developed. A unified linear mixed model is also proposed, which can account for both the additive and dominance effects of FR and PS correlations as well as their possible random interactions. Finally, this unified linear mixed model is applied to analyze a real data set from UK Biobank.

KEYWORDS

Linear mixed model, familial relatedness, population structure, coancestry coefficient, variance components, random interactions

1 Introduction

There are two broad sources for genetic correlation: the familial relatedness (FR) and population structure (PS). Appropriate modeling and adjustment of FR and PS correlation is crucial in family based genetic association studies. The FR correlation comes from the fact that two family members may share certain alleles or genomic regions in identity-by-descent (IBD); i.e., which are inherited from a founder (e.g., a parent or grand parent) of the family. The PS correlation may arise when two individuals share certain alleles in IBD from some common ancestors. Similar to FR, PS can be treated as FR under a much larger space and time scale. When a study sample consists of family data with heterogeneous racial or ethnic background, both FR and PS correlations need to be accounted for in order to appropriately control for the false positive rate at a nominal level.

In human population, both FR and PS can further breakdown into additive and dominance components to account for potential additive and dominance genetic effects. For a continuous disease phenotype, the linear mixed model (LMM) has been proposed to adjust for the additive FR and PS correlations simultaneously (Yu et al., 2006; Kang et al., 2008, 2010; Hoffman, 2013). But the possible FR and PS correlations from the dominance effects are often ignored. There has been a lack of knowledge on how to separate the FR and PS correlations from the observed genomic relationship matrices. There is also no discussion on how to model possible random interactions between FR and PS.

In this study, besides the classical additive genomic relationship matrix, a dominance genomic relationship matrix is introduced. A link between the additive/dominance genomic relationship matrices and the coancestry (or kinship)/double coancestry coefficients is established. A separation of the FR and PS correlations based on the estimates of coancestry and double coancestry coefficients from the genomic relationship matrices is also developed. In addition, a unified linear mixed model is proposed which can account for both the additive and dominance effects of FR and PS correlations as well as their possible random interactions. Strategies on fitting this type of LMM is discussed. This unified linear mixed model is further applied to analyze a real data set from UK Biobank in Britain.

2 Interpretation of the genomic relationship matrices

In population genetics, the kinship (or coancestry) and double coancestry coefficients are well known parameters for describing the genetic relatedness among individuals (Falconer and Mackay, 1996; Weir, 1996; Lynch and Walsh, 1998). For two relatives, their kinship coefficient is defined as the probability that a paternal or maternal allele at a putative locus from one individual is IBD with a paternal or maternal allele at the same locus from the other individual. Their coancestry coefficient is 2 times the kinship coefficient. Their double coancestry coefficient is the probability that both the paternal and maternal alleles at a putative locus from one individual are IBD with the paternal and maternal alleles at the same locus from the other individual. Consider a collected sample of family data with n individuals. When the family structures are known, the expected kinship and double coancestry coefficients from FR can be calculated using the classical Melecot or Wright methods (Falconer and Mackay, 1996). The kinship and double coancestry coefficients from the joint FR and PS can also be estimated from the genome-wide single nucleotide polymorphisms (SNPs) (Sun et al., 2016; Dou et al., 2017). Assume that there are m biallelic SNPs. Let A_j (and a_j) denote the minor (common) allele at locus j with minor allele frequencies (MAF) $p_j = P(A_j)$ and $q_j = P(a_j) = 1 - p_j$ for $j = 1, \cdots, m$. Following the Fisherian modeling strategy (see Zeng et al., 2005; Wang and Zeng, 2009), the following indicator variables can be introduced to describe the inheritance of the two parental alleles for each individual i at the j-th locus

$$z_{1ij} = \begin{cases} 1, & \text{the inherited paternal allele is } A_j \\ 0, & \text{the inherited paternal allele is } a_j \end{cases}$$

$$z_{2ij} = \begin{cases} 1, & \text{the inherited maternal allele is } A_j \\ 0, & \text{the inherited maternal allele is } a_j \end{cases}$$

Then the following centered (or mean-corrected) index variables can be defined

$$\tilde{z}_{1ij} = \begin{cases} 1 - p_j, & \text{the inherited paternal allele is } A_j \\ -p_j, & \text{the inherited paternal allele is } a_j \end{cases}$$

$$\tilde{z}_{2ij} = \begin{cases} 1 - p_j, & \text{the inherited maternal allele is } A_j \\ -p_j, & \text{the inherited maternal allele is } a_j \end{cases}$$

As the parental origins (i.e., phases) of the alleles are usually unknown in the observed genotypes of SNPs, the above indicator and index variables are not identifiable. However, the following mean-corrected genotype coding variables are well defined (Wang and Zeng, 2009)

$$w_{ij} = \tilde{z}_{1ij} + \tilde{z}_{2ij} = \begin{cases} 2(1-p_j), & \text{if } g_{ij} = A_jA_j \\ 1-2p_j, & \text{if } g_{ij} = A_ja_j \\ -2p_j, & \text{if } g_{ij} = a_ja_j \end{cases}$$

$$v_{ij} = \tilde{z}_{1ij}\tilde{z}_{2ij} = \begin{cases} (1-p_j)^2, & \text{if } g_{ij} = A_jA_j \\ -p_j(1-p_j), & \text{if } g_{ij} = A_ja_j \\ p_j^2, & \text{if } g_{ij} = a_ja_j \end{cases}$$

where g_{ij} denotes the phase-unknown genotype of individual i at the j-th SNP for $j = 1, \cdots, m$.

Consider a random sample from a study population. Then the indicator (or index) and genotype coding variables defined above can be treated as random variables. Let f_i^j denote the probability that the paternal and maternal alleles of an individual i at locus j are IBD (i.e., the inbreeding coefficient at locus j for individual i). By assuming that non-IBD alleles are inherited independently, it can be shown that

$$\text{Var}(w_{ij}) = 2(1+f_i^j)p_jq_j \quad (1)$$

$$\text{Var}(v_{ij}) = [1-4f_i^j-(f_i^j)^2](p_jq_j)^2 + f_i^jp_jq_j \quad (2)$$

When $f_i^j = 0$, then $\text{Var}(w_{ij}) = 2p_jq_j$ and $\text{Var}(v_{ij}) = (p_jq_j)^2$. An approximated normalization on the mean-corrected genotype coding variables can be made through the following

$$w_{ij}^* = w_{ij}/\sqrt{2p_jq_j}, \quad v_{ij}^* = v_{ij}/p_jq_j \quad (3)$$

for $i = 1, \cdots, n$ and $j = 1, \ldots, m$.

The variables w_{ij}^* and v_{ij}^* are referred as the mean-corrected additive and dominance genotype coding variables, respectively. Based on these variables, the additive and dominance design (or standardized genotype coding) matrices of the genotypes can be defined as $W = (w_{ij}^*)_{n\times m}$ and $V = (v_{ij}^*)_{n\times m}$, respectively. The additive and dominance genomic relationship matrices are defined as $\Sigma = WW^T/m$ and $\Delta = VV^T/m$, respectively. The same additive genomic relationship matrix Σ has been proposed previously to model the

subject-by-subject genetic correlation (Kang et al., 2008, 2010; Hoffman, 2013). This additive genomic matrix Σ is also the same as the genomic relationship matrix used in GCTA (see Yang et al., 2011). However, the dominance genomic matrix introduced in this study is new and has not been used before.

In the following, we construct a link between the additive and dominance genomic relationship matrices and the coancestry (or kinship) and double coancestry coefficients. Consider two individuals i, i' $(i \neq i')$ within a family. Let $r^{f,j}_{i_1 i'_1}$ be the probability that two paternal alleles A_j carried by individuals i, i' at locus j are IBD due to the FR (i.e., the two alleles come from the same founder allele within a family), and $r^{p,j}_{i_1 i'_1}$ be the probability that two paternal alleles A_j carried by individuals i, i' at locus j are IBD due to PS (i.e., the two alleles come from two different founder alleles within the family but they share IBD owing to the PS). Note that the IBD from PS refers to the origin of alleles, while the IBD from FR comes from different familial relationships. Given a familial relationship, the IBD probabilities from FR are mainly driven by Mendel's law of segregation in cell meiosis, which should not depend on the origin of the alleles (i.e., PS) in most of the genomic regions. Assuming that the IBD probabilities from FR are independent of PS and non-IBD alleles are inherited independently at locus j, then

$$\begin{aligned} E(z_{1ij} z_{1i'j}) &= r^{f,j}_{i_1 i'_1} \cdot p_j + (1 - r^{f,j}_{i_1 i'_1})[r^{p,j}_{i_1 i'_1} \cdot p_j + (1 - r^{p,j}_{i_1 i'_1}) p_j^2] \\ &= (r^{f,j}_{i_1 i'_1} + r^{p,j}_{i_1 i'_1} - r^{f,j}_{i_1 i'_1} r^{p,j}_{i_1 i'_1}) p_j q_j + p_j^2 \end{aligned}$$

and

$$\mathrm{Corr}(z_{1ij}, z_{1i'j}) = r^{f,j}_{i_1 i'_1} + r^{p,j}_{i_1 i'_1} - r^{f,j}_{i_1 i'_1} r^{p,j}_{i_1 i'_1} \triangleq r^{j}_{i_1 i'_1} \tag{4}$$

where $r^{j}_{i_1 i'_1}$ is the probability that two paternal alleles A_j carried by individuals i, i' at locus j are IBD due to either FR or PS; i.e., the kinship coefficient between the two paternal alleles for individuals i, i' at locus j. Therefore, the kinship coefficient $r^{j}_{i_1 i'_1}$ can be interpreted as the correlation coefficient between z_{1ij} and $z_{1i'j}$. From the above equation, it is also interesting to see that the IBD probabilities $r^{j}_{ii'}$, $r^{f,j}_{ii'}$ and $r^{p,j}_{ii'}$ have the relationship $(1 - r^{j}_{ii'}) = (1 - r^{f,j}_{ii'})(1 - r^{p,j}_{ii'})$.

Suppose that the kinship coefficients are the same for both paternal and maternal alleles carried by individuals i and i'; i.e., $r^{j}_{i_1 i'_1} = r^{j}_{i_1 i'_2} = r^{j}_{i_2 i'_1} = r^{j}_{i_2 i'_2} \triangleq r^{j}_{ii'}$. Then $\mathrm{Cov}(w_{ij}, w_{i'j}) = 4 r^{j}_{ii'} p_j q_j$, where $r^{j}_{ii'}$ is the kinship coefficient between individuals i, i' at locus j. The nor-

malization (3) on w_{ij} leads to $2r^j_{ii'} = \text{Cov}(w^*_{ij}, w^*_{i'j}) = \text{Corr}(w_{ij}, w_{i'j})$. By further assuming that the kinship coefficients are the same across all the marker loci, then the (i, i') element of matrix Σ can be interpreted as an estimator of the coancestry coefficient $2r_{ii'}$ (or 2 times the kinship coefficient $r_{ii'}$) between two individuals i and i'; i.e., $2\hat{r}_{ii'} = \sum_{j=1}^{m} w^*_{ij}w^*_{i'j}/m$. When $i = i'$, from equation (1), the diagonal element (i, i) of the matrix Σ is an estimator of $(1 + f_i)$. There has been inconsistency on whether normalization on the genotype coding variables w_{ij} is needed. Patterson et al. (2006) showed that the normalization can improve simulation results and make population structure clearer in some real data. Here a link between the normalized w_{ij} and the coancestry coefficient is established, which further rationalizes the normalization method.

For the additive correlation between two individuals, the genetic correlation between two parental alleles within each individual (i.e., inbreeding) is not involved. However, the inbreeding can affect the correlation between the additive and dominance or among the dominance genotype coding variables. For example, the probability (or inbreeding coefficient) that the parental and maternal alleles at locus j carried by individual i are IBD is $f^j_i = E(v^*_{ij})$. Thus, to examine the dominance genomic relationship matrix, a more detailed description of the IBD status is needed. Based on the extended 15 IBD states and probabilities (see Cockerham, 1971; Weir et al., 2006), then

$$\begin{aligned}
\text{Cov}(v_{ij}, v_{i'j}) &= E(\tilde{z}_{1ij}\tilde{z}_{2ij}\tilde{z}_{1i'j}\tilde{z}_{2i'j}) - E(\tilde{z}_{1ij}\tilde{z}_{2ij})E(\tilde{z}_{1i'j}\tilde{z}_{2i'j}) \\
&= \delta^j_{i_1i_2i'_1i'_2}[E(\tilde{z}^4_{1ij}) - E(\tilde{z}^2_{1ij})E(\tilde{z}^2_{1i'j})] + (\delta^j_{i_1i'_1,i_2i'_2} + \delta^j_{i_1i'_2,i_2i'_1})E(\tilde{z}^2_{1ij})^2 \\
&= \delta^j_{i_1i_2i'_1i'_2}p_jq_j(p_j - q_j)^2 + (\delta^j_{i_1i'_1,i_2i'_2} + \delta^j_{i_1i'_2,i_2i'_1})p^2_jq^2_j
\end{aligned}$$

where $\delta^j_{i_1i_2i'_1i'_2}$ is the probability that all four alleles carried by individuals i, i' at locus j are IBD; $\delta^j_{i_1i'_1,i_2i'_2}$ is the probability that the parental and maternal alleles carried by individual i are IBD with the parental and maternal alleles carried by individual i' at locus j, respectively; and $\delta^j_{i_1i'_2,i_2i'_1}$ is the probability that the parental and maternal alleles carried by individual i are IBD with the maternal and paternal alleles carried by individual i' at locus j, respectively. Note that $\delta^j_{ii'} = (\delta^j_{i_1i'_1,i_2i'_2} + \delta^j_{i_1i'_2,i_2i'_1})$ is the probability that the two parental alleles carried by individuals i' are double IBD with the two parental alleles carried by individuals i' (i.e., double coancestry coefficient) from either FR or PS. Thus,

$$\text{Cov}(v^*_{ij}, v^*_{i'j}) = \delta^j_{ii'} + \delta^j_{i_1i_2i'_1i'_2}(p_j - q_j)^2/(p_jq_j) \tag{5}$$

As there is no inbreeding between close relatives in human population, the IBD probability $\delta^j_{i_1 i_2 i'_1 i'_2}$ mainly originates from PS, which is usually much weaker than FR. By ignoring this higher order IBD probability, then $\text{Cov}(v^*_{ij}, v^*_{i'j}) \approx \delta^j_{ii'}$. If we further assume that the IBD probabilities $\delta^j_{ii'}$ are the same across all the marker loci, then the (i, i') element of matrix Δ can be treated as an estimator of the double coancestry coefficient $\delta_{ii'}$ between individuals i and i'; i.e., $\hat{\delta}_{ii'} = \sum_{j=1}^{m} v^*_{ij} v^*_{i'j}/m$. When $i = i'$, from equation (2) and (3), $\text{Cov}(v^*_{ij}, v^*_{ij}) = [1 - 4f^j_i - (f^j_i)^2] + f^j_i/(p_j q_j)$, which depends on the allele frequencies. Therefore, the diagonal elements of the matrix Δ have no simple interpretation except that these diagonal elements should be close to 1 when inbreeding level is low across all loci.

In the above, we clarified that each off-diagonal element (i, i') of the additive genomic matrix Σ can provide an estimator of the coancestry coefficient $2r_{ii'}$. Also, each off-diagonal element (i, i') of the dominance genomic matrix Δ can be treated as an estimate of the double coancestry coefficient $\delta_{ii'}$ between two individuals i and i' under certain assumptions. Besides, the coancestry or double coancestry coefficients can be interpreted as correlations of alleles. As pointed out in Weir and Goudet (2017), by taking into account the correlation across different populations, the correlations of alleles could also be negative. However, the additive and dominance genomic matrices cannot be directly interpreted as correlation matrices because their diagonal elements could exceed or below one.

In the above derivation, it was assumed that the IBD probabilities from FR are independent of PS. If the inheritance of SNP alleles at certain genomic region really depends on the origin of alleles (e.g., the ancestral informative SNPs), these ancestral related SNP may need special care and should be excluded from this analysis. In practice, another potential problem in using the additive and dominance genomic matrices is that the allele frequencies in a study population are often estimated using sample allele frequencies. The deviation of sample allele frequencies from the true allele probabilities could bias the estimates of the coancestry and double coancestry coefficients. To reduce this bias, SNPs with rare alleles should be excluded. The allele frequencies should also be estimated using unrelated individuals from the sample.

3 A separation of FR and PS correlations

The additive and dominance genomic matrices provide estimates of the kinship and double coancestry coefficients from the combined FR and PS. In order to assess FR and PS

separately, one need to consider the correlation from FR and PS individually. Let $r^f_{ii'}$ and $\delta^f_{ii'}$ be the expected kinship and double coancestry coefficients from FR, and $r^p_{ii'}$ and $\delta^p_{ii'}$ be the kinship and double coancestry coefficients from PS. By assuming that the IBD probabilities from FR are independent of PS and the kinship coefficients are the same across all the marker loci, then from equation (4) the kinship coefficients have the relationship $r_{ii'} = r^f_{ii'} + r^p_{ii'} - r^f_{ii'}r^p_{ii'}$. Or, equivalently, $(1 - r_{ii'}) = (1 - r^f_{ii'})(1 - r^p_{ii'})$. For the double coancestry coefficients, assume that the IBD probabilities from FR are independent of PS and the double coancestry coefficients are the same across all the marker loci. It can be shown similarly that $(1 - \delta_{ii'}) = (1 - \delta^f_{ii'})(1 - \delta^p_{ii'})$. It should be pointed out that a similar relationship was previously established for Wright's F-statistics (see Wright, 1950; Holsinger and Weir, 2009). Here the relationship is extended to FR and PS correlations based on the IBD probabilities.

The above relationship between FR and PS correlations provides a way to construct separate genetic correlation matrices for PS and FR. From using the genome-wide genetic markers such as single nucleotide polymorphisms (i.e., SNPs), one can first estimate the joint kinship and double coancestry coefficients as $\hat{r}_{ii'}$ and $\hat{\delta}_{ii'}$ from the combined FR and PS. For family members with known family structures, by assuming that no genetic correlation or inbreeding among founders, their expected kinship and double coancestry coefficients $r^f_{ii'}$ and $\delta^f_{ii'}$ from FR can usually be derived from the classical Malecot or Wright methods (Falconer and Mackay, 1996). Note that the $r^f_{ii'}$ and $\delta^f_{ii'}$ are defined as the IBD probabilities raised by FR only, while the inbreeding and genetic correlation among founders come from PS. So the assumption of no genetic correlation and inbreeding among founders for Malecot or Wright methods holds well.

When the family structures are unknown, the expected familiar correlations can also be extracted from the $r_{ii'}$ and $\delta_{ii'}$ estimates. As the PS correlation is usually much weaker than FR correlation, one way to distinguish the FR from PS is to choose a cut-off threshold of the coancestry coefficients for identification of the family members and determination of the FR correlation. For some common familial relationships, the expected coancestry usually follow certain patterns (see "`https://en.wikipedia.org/wiki/Coefficient_of_relationship`"). Typically, the expected coancestry from FR can take values 1 for monozygotic twins, 1/2 for parent-child or full sibs, 1/4 for half-sibs or grand parents and grand children, 1/8 for great grand parents and children, 1/16 for half-grandaunt/uncle or grandniece/nephew, etc. Similarly, the expected double coancestry from FR can take values 1 for monozygotic twins,

1/2 for full sibs, 0 for parent-child, etc. (Falconer and Mackay, 1996). If one use, say, a threshold of $1/32 = 0.03125$ to define family members. It can probably account for most of the common FR from the recent $1 \sim 3$ generations. The expected coancestry and double coancestry coefficients from FR can then be extracted from the $r_{ii'}$ and $\delta_{ii'}$ estimates by equating them being 0, 1/32, 1/16, 1/8, 1/4, 1/2 or 1, whichever is the closest. After that, the kinship and double coancestry coefficients $r_{ii'}^{PS}$ and $\delta_{ii'}^{PS}$ from PS can be calculated as

$$\begin{cases} r_{ii'}^{p} &= (\hat{r}_{ii'} - r_{ii'}^{f})/(1 - r_{ii'}^{f}) \\ \delta_{ii'}^{p} &= (\hat{\delta}_{ii'} - \delta_{ii'}^{f})/(1 - \delta_{ii'}^{f}) \end{cases}$$

If $r_{ii'}^{f} = 1$ (or $\delta_{ii'}^{f} = 1$), one can set $2r_{ii'}^{p} = 0$ (or $\delta_{ii'}^{p} = 0$). When $i = i'$, note that the coancestry and double coancestry coefficents $2r_{ii}^{f}$, δ_{ii}^{f} can be treated as correlation coefficients. Therefore, one can set $2r_{ii}^{f} = 1$, $\delta_{ii}^{f} = 1$, $2r_{ii}^{p} = 2\hat{r}_{ii}$ and $\delta_{ii}^{p} = \hat{\delta}_{ii}$.

In the above, it was implicitly assumed that the genetic correlation comes from either FR or PS. One critical issue to separate FR from PS is how to choose a cut-off threshold of the coancestry coefficients for identification of the family members. From the definition of PS and FR, the distinction between FR and PS really depends on the time scale. If we refer PS as the genetic correlation from different ancestral populations or races and the rest as FR, then FR would include familiar correlation from tens or even hundreds of generations. On the other hand, if we define FR as the familial correlation from several recent generations and the rest as PS, then the PS correlation would include the familial correlation before the recent generations. As the PS correlation is supposedly to be much weaker than FR correlation, one may intend to choose a small cut-off threshold of the coancestry coefficients for identification of the FR correlation. In practice, however, a small cut-off threshold will lead to a weak PS correlation. To have the effect of PS correlation detectable, the cut-off threshold should not be too small.

4 A unified linear mixed model

Consider a random sample of s families from a study population with n_i individuals in the i-th family for $i = 1, \cdots, s$ and $n = \sum_{i=1}^{s} n_i$. Let y_i, $i = 1, \cdots, n$, denote the observed quantitative values for a disease phenotype. The disease phenotype can usually be modeled as $y_i = x_i^T\beta + a_i + d_i + \epsilon_i$, where x_i is a vector of the fixed covariates, a_i (or d_i) denote the random additive (or dominance) genetic effect, and ϵ_i is the residual error. Let $\tilde{a} = (a_1, \cdots, a_n)$ and

$\tilde{d} = (d_1, \cdots, d_n)$. From quantitative genetics (see Falconer and Mackay, 1996; Lynch and Walsh, 1998), when PS is ignored, it has been well known that the genetic covariances can be expressed as $\text{Cov}(\tilde{a}) = 2\Phi_f \sigma^2_{a,f}$ and $\text{Cov}(\tilde{d}) = \Delta_f \sigma^2_{d,f}$, where $\sigma^2_{a,f}$ and $\sigma^2_{d,f}$ denote the additive and dominance genetic variance components of FR, respectively; and $\Phi_f = (r^f_{ii'})_{n\times n}$ and $\Delta_f = (\delta^f_{ii'})_{n\times n}$ with $r^f_{ii'}$ and $\delta^f_{ii'}$ being the kinship and double coancestry coefficients from FR between individuals i and i', respectively. Similarly, when FR is ignored, one would have $\text{Cov}(\tilde{a}) = 2\Phi_p \sigma^2_{a,p}$ and $\text{Cov}(\tilde{d}) = \Delta_p \sigma^2_{d,p}$, where $\sigma^2_{a,p}$ and $\sigma^2_{d,p}$ denote the additive and dominance genetic variance components of PS, respectively; and $\Phi_p = (r^p_{ii'})_{n\times n}$ and $\Delta_p = (\delta^p_{ii'})_{n\times n}$ with $r^p_{ii'}$ and $\delta^p_{ii'}$ being the kinship and double coancestry coefficients from PS between individuals i and i', respectively.

In general, when both FR and PS are present, the FR and PS correlations need to be accounted for simultaneously. One naive way is to estimate $r_{ii'}$ and $\delta_{ii'}$ from the genetic markers, and then assume that $\text{Cov}(\tilde{a}) = (2r_{ii'})_{n\times n}\sigma^2_a$ and $\text{Cov}(\tilde{d}) = (\delta_{ii'})_{n\times n}\sigma^2_d$. However, this modeling strategy basically assumes that the FR and PS correlations contribute similar effects on the phenotypic covariance. It also cannot assess the effects of FR and PS correlations separately. Alternatively, the FR and PS can be treated as two independent random sources. Separate additive and dominance effects for FR and PS can be introduced in the model. In addition, noting that the FR and PS as two independent factors could be crossed in complicated ways in human population, the interactions between FR and PS can also be included in the joint model if needed.

Let $\tilde{a}_f = (a_{1f}, \cdots, a_{nf})$ and $\tilde{a}_p = (a_{1p}, \cdots, a_{np})$ represent the additive effects from FR and PS, respectively. Following the classical strategy in generating interaction terms, the additive random effects $\tilde{a}$ can be partitioned into three components: $\tilde{a} = \tilde{a}_f + \tilde{a}_p + \alpha\tilde{a}_f \odot \tilde{a}_p$, where $\odot$ denotes the element-by-element Hadamard product of vectors or matrices, $\tilde{a}_f \odot \tilde{a}_p$ represents the interaction variables between FR and PS, and α is a scalar which quantifies the effect of the interaction variables. The covariance matrices of $\tilde{a}_f$ and $\tilde{a}_p$ are given by $\text{Cov}(\tilde{a}_f) = (2r^f_{ii'})_{n\times n}\sigma^2_{a,f}$ and $\text{Cov}(\tilde{a}_p) = (2r^p_{ii'})_{n\times n}\sigma^2_{a,p}$, respectively. To derive the covariance matrix of $\tilde{a}_f \odot \tilde{a}_p$, consider two individuals i and i' with $\text{Cov}(a_{if}, a_{i'f}) = 2r^f_{ii'}\sigma^2_{a,f}$ and $\text{Cov}(a_{ip}, a_{i'p}) = 2r^p_{ii'}\sigma^2_{a,p}$. Assuming that the additive effects $\{a_{if}, a_{i'f}\}$ of FR are independent of the additive effects $\{a_{ip}, a_{i'p}\}$ of PS, it can be shown that the covariance $\text{Cov}(a_{if}a_{ip}, a_{i'f}a_{i'p}) = 4r^f_{ii'}r^p_{ii'}\sigma^2_{a,f}\sigma^2_{a,p}$. Therefore, the correlation $\text{Corr}(a_{if}a_{ip}, a_{i'f}a_{i'p}) = 4r^f_{ii'}r^p_{ii'}$ and the correlation matrix $\text{Corr}(\tilde{a}_f \odot \tilde{a}_p) = 4\Phi_f \odot \Phi_p$. The additive random effect $\tilde{a}$ can be re-expressed as $\tilde{a} = \tilde{a}_f + \tilde{a}_p + \tilde{a}_f \odot \tilde{a}_p$, where $\tilde{a}_f \odot \tilde{a}_p \sim N(0, 4\Phi_f \odot \Phi_p\sigma^2_{aa,fp})$ and $\sigma^2_{aa,fp} = \alpha^2\sigma^2_{a,f}\sigma^2_{a,p}$. It can

also be shown that the covariance matrices $\text{Cov}(\tilde{a}_f, \tilde{a}_f \odot \tilde{a}_p) = \text{Cov}(\tilde{a}_p, \tilde{a}_f \odot \tilde{a}_p) = 0$. Thus, there is an orthogonal partition $\text{Cov}(\tilde{a}) = \text{Cov}(\tilde{a}_f) + \text{Cov}(\tilde{a}_p) + \text{Cov}(\tilde{a}_f \odot \tilde{a}_p)$.

Similarly, the dominance effects $\tilde{d}$ can be partitioned into three components: $\tilde{d} = \tilde{d}_f + \tilde{d}_p + \tilde{d}_f \odot \tilde{d}_p$, where $\tilde{d}_f \odot \tilde{d}_p$ represents the dominance by dominance interactions between FR and PS with $\text{Corr}(\tilde{d}_f \odot \tilde{d}_p) = \Delta_f \odot \Delta_p$. In addition, the FR and PS interactions may also include the additive by dominance interaction $\tilde{a}_f \odot \tilde{d}_p$ and the dominance by additive interaction $\tilde{a}_p \odot \tilde{d}_f$ with $\text{Corr}(\tilde{a}_f \odot \tilde{d}_p) = 2\Phi_f \odot \Delta_p$ and $\text{Corr}(\tilde{a}_p \odot \tilde{d}_f) = 2\Phi_p \odot \Delta_f$. A unified LMM to account for both the additive and dominance effects of FR and PS as well as their possible interactions is then given by

$$Y = X\beta + \tilde{a}_f + \tilde{a}_p + \tilde{d}_f + \tilde{d}_p + \tilde{a}_f \odot \tilde{a}_p + \tilde{a}_f \odot \tilde{d}_p + \tilde{a}_p \odot \tilde{d}_f + \tilde{d}_f \odot \tilde{d}_p + Je_f + \epsilon \qquad (6)$$

where $Y = (y_1, \cdots, y_n)^T$, $X = (x_1, \cdots, x_n)^T$ is a $n \times p$ design matrix for the fixed effects β including an intercept, $\tilde{a}_f \sim N(0, 2\Phi_f \sigma^2_{a,f})$, $\tilde{a}_p \sim N(0, 2\Phi_p \sigma^2_{a,p})$, $\tilde{d}_f \sim N(0, \Delta_f \sigma^2_{d,f})$, $\tilde{d}_p \sim N(0, \Delta_p \sigma^2_{d,p})$, $\tilde{a}_f \odot \tilde{a}_p \sim N(0, 4\Phi_f \odot \Phi_p \sigma^2_{aa,fp})$, $\tilde{a}_f \odot \tilde{d}_p \sim N(0, 2\Phi_f \odot \Delta_p \sigma^2_{ad,fp})$, $\tilde{a}_p \odot \tilde{d}_f \sim N(0, 2\Phi_p \odot \Delta_f \sigma^2_{da,fp})$, $\tilde{d}_f \odot \tilde{d}_p \sim N(0, \Delta_f \odot \Delta_p \sigma^2_{dd,fp})$, and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T \sim N(0, \sigma^2 \text{I}_n)$ is a vector of model residuals. Note that here $e_f = (e_1, \cdots, e_s)^T \sim N(0, \sigma^2_s \text{I}_s)$ denotes the family shared random effects with each element e_i being shared by all the members in the i-th family, while J is a $n \times s$ design matrix for e_f with elements of 0's or 1's. In a prospective or retrospective cohort with family data, without genetic marker information, this random vector is usually added to account for some unobserved genetic and environmental factors shared by the family members. With the genetic marker information, the family shared genetic effects can be separated from the family shared environmental effects. Besides, the genetic correlation among family members can vary depending on the coancestry or double coancestry coefficients. By including the FR correlations, the vector e_f in the above model mainly represents the unobserved environmental factors shared by the family members.

In model (6), we assumed that the additive and dominance random effects of FR and PS are independent. Note that this assumption is different from that the IBD probabilities from FR are independent of PS in the previous Sections. Here the assumption that the additive and dominance random effects of FR and PS are independent refers to a situation where the additive and dominance FR correlation structures stay similar under different PS, which may not always hold in practice. But this assumption is not necessary for modeling the additive and dominance effects of FR and PS and their interactions. When effects of

FR and PS are dependent, however, we may not have the nice covariance structures such as $\text{Corr}(\tilde{a}_f \odot \tilde{a}_p) = 4\Phi_f \odot \Phi_p$, $\text{Corr}(\tilde{a}_f \odot \tilde{d}_p) = 2\Phi_f \odot \Delta_p$, $\text{Corr}(\tilde{a}_p \odot \tilde{d}_f) = 2\Phi_p \odot \Delta_f$ or $\text{Corr}(\tilde{d}_f \odot \tilde{d}_p) = \Delta_f \odot \Delta_p$. When the random effects of FR and PS are indeed independent, then an orthogonal partition of the covariance $\text{Cov}(Y|X)$ is given by

$$\begin{aligned}\text{Cov}(Y|X) = {} & 2\Phi_f\sigma^2_{a,f} + 2\Phi_p\sigma^2_{a,p} + \Delta_f\sigma^2_{d,f} + \Delta_p\sigma^2_{d,p} + 4\Phi_f \odot \Phi_p\sigma^2_{aa,fp} \\ & + 2\Phi_f \odot \Delta_p\sigma^2_{ad,fp} + 2\Phi_p \odot \Delta_f\sigma^2_{da,fp} + \Delta_f \odot \Delta_p\sigma^2_{dd,fp} + JJ'\sigma^2_s + \text{I}_n\sigma^2\end{aligned}$$

Model (6) can assess both the additive and dominance effects of FR and PS as well as their interactions. The interaction presents when certain family members are genetically related via not only FR but also PS. Meanwhile, these family members preserve a stronger or weaker phenotypic covariance than just an addition of the separate covariances from FR and PS. In practice, however, it is usually difficult to detect the interactions due to limited information in the study samples. By ignoring the interaction terms, model (6) reduces to a LMM including only the additive and dominance effects of FR and PS. When no family data are involved, model (6) can also be used to fit a sample with only the additive and dominance effects of PS. This unified LMM (6) provides a general framework which includes all these scenarios as special cases.

In order to fit a LMM (6), it is more convenient to re-formulate it into a standard LMM in which all the random vectors have their elements being independent and identically distributed (i.i.d.). This can be achieved by applying a spectral decomposition on the covariance matrices (see Hoffman, 2013; Wang et al., 2015). Suppose that $A_f, A_p, D_f, D_p, L_{aa}, L_{ad}, L_{da}$ and L_{dd} are the square roots of matrices $2\Phi_f, 2\Phi_p, \Delta_f, \Delta_p, 4\Phi_f \odot \Phi_p, 2\Phi_f \odot \Delta_p, 2\Phi_p \odot \Delta_f$ and $\Delta_f \odot \Delta_p$, respectively. Then model (6) can be re-written as the following

$$\begin{aligned}Y \quad = {} & X\beta + A_f a_f + A_p a_p + D_f d_f + D_p d_p + L_{aa}(aa)_{fp} \\ & + L_{ad}(ad)_{fp} + L_{da}(da)_{fp} + L_{dd}(dd)_{fp} + Je_f + \epsilon \end{aligned} \tag{7}$$

where $a_f \sim N(0, \sigma^2_{a,f}\text{I}_n)$, $a_p \sim N(0, \sigma^2_{a,p}\text{I}_n)$, $d_f \sim N(0, \sigma^2_{d,f}\text{I}_n)$, $d_p \sim N(0, \sigma^2_{d,p}\text{I}_n)$, $(aa)_{fp} \sim N(0, \sigma^2_{aa,fp}\text{I}_n)$, $(ad)_{fp} \sim N(0, \sigma^2_{ad,fp}\text{I}_n)$, $(da)_{fp} \sim N(0, \sigma^2_{da,fp}\text{I}_n)$, and $(dd)_{fp} \sim N(0, \sigma^2_{dd,fp}\text{I}_n)$. It is easy to see that model (7) can provide the same covariance structure for the phenotype Y as model (6).

Comparing to the classical LMM, one challenge in fitting model (7) is that the dimensions

of some random vectors could be as large as n. One way to simplify the model fitting is to treat PS and its interactions with FR as fixed effects and apply the classical principal component approach. Typically, the principle components (PCs) can be constructed by applying the singular value decomposition (SVD) on the standardized genotype coding matrices W and V (Patterson et al., 2006). By applying SVD on W and V, then

$$W = U_1 S_1 C_1^T, \; V = U_2 S_2 C_2^T \tag{8}$$

where U_j are $n \times n$ orthogonal matrices (i.e., $U_j^T U_j = I_n$), C_j are $m \times m$ orthogonal matrices (i.e., $C_j^T C_j = I_m$) for $j = 1, 2$, S_1 (or S_2) is a $n \times m$ matrix with singular values $\lambda_{11} \geq \cdots \geq \lambda_{1r} > 0$ (or $\lambda_{21} \geq \cdots \geq \lambda_{2s} > 0$) plus zeros as diagonal elements and zeros as off-diagonal elements, and $rank(S_1) = rank(W) = r$ and $rank(S_2) = rank(V) = s$. From (8), one can see that each column vector of matrix U_1 (or U_2) is a linear combination of the column vectors of the standardized genotype coding matrices W (or V). So the column vectors of U_1 and U_2 can be referred as the PCs of the matrix W and V, respectively. Note that $2\Phi = WW^T = U_1 S_1^2 U_1^T$ and $\Delta = VV^T = U_2 S_2^2 U_2^T$. Thus, the PCs can also be constructed via direct spectral decompositions on 2Φ and Δ.

Following the same strategy, one can construct PCs for PS by applying the spectral decomposition on $2\Phi_p$ and Δ_p, respectively. Suppose that $2\Phi_p = U_1 S_1^2 U_1^T$ and $\Delta_p = U_2 S_2^2 U_2^T$. Then, the square root matrices $A_p = U_1 S_1 U_1^T$ and $D_p = U_2 S_2 U_2^T$. Note that one could also take $A_p = U_1 S_1$ and $D_p = U_2 S_2$ as the design matrices in model (7). Unlike the previous square root matrices, the latter ones are no longer symmetric matrices. But they can still provide the same covariance matrices $\mathrm{Cov}(A_p a_p) = A_p A_p^T \sigma_{a,p}^2 = 2\Phi_p \sigma_{a,p}^2$ and $\mathrm{Cov}(D_p d_p) = D_p D_p^T \sigma_{d,p}^2 = \Delta_p \sigma_{d,p}^2$. When a_p, d_p in model (7) are treated as fixed effects, it is simpler to just take $A_p = U_1$ and $D_p = U_2$. This is equivalent to treat the PCs as fixed covariates. As pointed out in Hoffman (2013), the eigen-spectrum of PS correlation decays quickly. Therefore, only a few major PCs corresponding to the leading eigenvalues of $\{\lambda_{1r}\}$ and $\{\lambda_{2s}\}$ are usually needed to adjust for the PS effects. The correlation matrices for FR and PS interactions can be handled similarly.

Unlike PS, the correlation matrices $2\Phi_f$ or Δ_f of FR can have their eigenvalues persist above certain positive level. The re-formulation method proposed in Wang et al. (2015) can be used here to deal with the FR correlction. After excluding PS, the correlation matrices $2\Phi_f$ and Δ_f for FR are usually block diagonal matrices with the diagonal sub-matrices

$2\Phi_f^i$ and Δ_f^i for FR correlation within the i-th family ($i = 1, \cdots, s$). Separate Cholesky decompositions can be applied to obtain $2\Phi_f^i = A_i A_i^T$ and $\Delta_f^i = D_i D_i^T$ for each family i. The matrices A_i and D_i can be expanded to have their number of columns all equal the maximum family size $r = \max_{1 \le i \le s} \{n_i\}$ by adding extra columns of 0's if needed. After that, the matrices A_i and D_i can be concatenated vertically to construct the $n \times r$ design matrices A_f and D_f in model (7). Then model (7) can be fitted using PROC NLMIXED or PROC GLIMMIX procedures in SAS software (SAS Institute Inc, Cary, NC) or Bayesian approach by treating different families as independent clusters (see details in Wang et al., 2015).

5 Example

We applied the proposed LMM (7) to a real data set from UK Biobank (Sudlow et al., 2015). The data set consists of n=5,820 Caucasian from European population who were reported to be related to each other. The counts of white blood CD4+ T cells were considered as an outcome. Three covariates include: age, gender and BMI. For a simple interpretation, the 'age', which ranges from 40 to 70, is discretized into 6 groups: 40-45, 46-50, 51-55, 56-60, 60-65, and above 65. Similarly, the BMI is categorized into 5 groups: ≤ 20; (20-25], (25,30], (30,35] and greater than 35.

To construct the additive and dominance genomic matrices, the following criteria were applied to filtering the SNPs: 1) exclude SNPs with minor allele frequencies (MAF) $< 5\%$; 2) exclude SNPs with missingness $> 1.5\%$; 3) LD pruning to exclude SNPs which have $r^2 > 0.1$ with another tagged SNP; 4) remove C/G and A/T SNPs; and 5) exclude SNPs in regions with long-range LD. These were the same criteria used in Astle et al. (2016), which left us with approximately 270k SNPs.

The additive and dominance genomic relationship matrices W and V are constructed based on these SNPs. Then a threshold of $1/2^5 = 0.03125$ is used to determine the family members based on the additive genomic matrix Σ. That is if two individuals have their coancestry coefficient estimate greater or equal than 0.03125, then they are classified as familial relatives. After this clustering, only familial relatives can share a coancestry coefficient of 0.03125 or above. Members from different families have their coancestry coefficient estimates less than 0.03125. From the coancestry estimates, we identified 179 independent individuals, 2732 paired family members, 5 families with 3 individuals each, 35 families with

4 individuals each, 2 families with 6 individuals each, and one family with 10 individuals. The total number of estimated families is $s = 2,955$.

Within each family, we extract the expected FR correlations from $\Sigma = WW^T/m$ and $\Delta = VV^T/m$ by equating the expected coancestry and double coancestry coefficients being 0, 1/32, 1/16, 1/8, 1/4, 1/2 or 1, whichever is the closest. Then the PS correlation matrices are calculated using the formulae in the previous Section 3. Based on the FR correlation pattern from the coancestry and double coancestry coefficients, we can further identify some common familiar relationships. For examples, among the 2732 paired family members, there were 12 monozygotic twin pairs ($r^f_{ii'} = 0.5, \delta^f_{ii'} = 1$), 422 parent-child pairs ($r^f_{ii'} = 0.25$ and $\delta^f_{ii'} = 0$), 1452 sib-pairs ($r^f_{ii'} = 0.25$ and $\delta^f_{ii'} = 0.25$), 657 grandparent and child pairs ($r^f_{ii'} = 0.125$ and $\delta^f_{ii'} = 0$), 129 pairs with $r^f_{ii'} = 0.25$ and $\delta^f_{ii'} = 0.125$, and the rest 60 pairs with other types of familial relatedness. Among the 5 families with 3 individuals, it appears that 4 families include 2 siblings. Among the 35 families with 4 individuals, 23 families include at least two siblings.

First, a LMM without adjustment for FR and PS is fitted. It shows that both age (overall P=0.0024) and BMI (overall P$<$ 0.0001) are significantly associated with the outcome, while gender is not (P=0.57). Individuals of age greater than 65 have higher CD4 counts than individuals of $50 <$ age ≤ 65. Among the 5 BMI groups, the two groups of patients with BMI$\geq$ 25 have bigger means of CD4 cell counts than that of group BMI$\leq$ 20. All the 4 BMI groups with BMI$>$ 20 have their means of CD4 cell counts differ from each other.

For PS effects, we extract 10 leading PCs from the PS correlation matrices $2\Phi_p$ and Δ_p separately. For FR and PS interactions, we also extract 10 leading eigenvectors from $4\Phi_f \odot \Phi_p$, $2\Phi_f \odot \Delta_p$, $2\Phi_p \odot \Delta_f$ and $\Delta_f \odot \Delta_p$ each. We treat all these eigenvectors as fixed covariates and apply a stepwise forward selection procedure with a threshold of P$<$ 0.05 for both entry and stay in the model. Two PCs from PS (the 2nd leading one from $2\Phi_p$ and the 5th leading one from Δ_p) are identified to be associated with the outcome but no significant interactions of FR and PS are detected. This is probably expected as the interactions are often difficult to uncover and this UK Biobank data set mainly consists of Caucasian from European population.

To test for the FR correlation, we fit several LMM to the data set using PROC GLIMMIX and PROC NLMIXED procedures in SAS. All these LMM include age, gender, BMI and the two PCs from PS as fixed covariates. The results show that there is a strong additive FR correlation ($\hat{\sigma}^2_{a,f} = 0.025$, P$<$ 0.0001) but the dominance FR correlation is not significant.

Without adjusting for FR correlations, the family shared correlation is significant ($\hat{\sigma}_s^2$=0.011, P< 0.0001). After adjusting for the additive FR correlation, the family shared correlation is no longer significant ($\hat{\sigma}_s^2$=0.0048, P=0.010). This indicates that the family shared correlation is mainly reflected by the familial genetic correlation from the additive FR, while the family shared environmental correlation is weaker. After adjusting for the additive FR effect, the overall P=0.050 for age, the overall P< .0001 for BMI, and P=0.57 for gender. Meanwhile, the PC from $2\Phi_p$ has P=0.017 and the PC from Δ_p has P=0.013, which indicate that both the additive and dominance PS correlations may play a role on affecting the outcome after adjusting for the additive FR correlation.

6 Discussion

In human genomics, with two parental alleles at each locus, possible allelic interactions or so-called dominance effects between the paternal and maternal alleles may present within a gene locus in addition to the additive allelic effects. This feature can also lead to various allelic interactions between different loci when more than one gene or genomic locus are involved. Similarly, the FR or PS correlation can breakdown into additive and dominance components in order to account for potential additive and dominance genetic effects. Besides, possible random interactions between FR and PS could also arise. This makes the genetic modeling for FR and PS correlation more complicated than it appears.

In this study, a unified LMM is proposed which can assess both the additive and dominance effects of FR and PS correlations as well as their possible random interactions. Unlike the modeling for fixed genetic effects, the random effects of FR, PS and their random interactions could presumably affect the covariance structure of the phenotypes rather than the phenotypic means. In practice, it is usually difficult to really detect the interactions due to limited information in the study samples. This unified LMM provides a general framework under which one can at least test for the additive and dominance effects of FR and PS or their possible interactions.

The extension of the unified LMM to categorical or survival outcomes is plausible. For example, under the generalized linear mixed model (GLMM) framework, both the additive and dominance effects of FR and PS as well as their interaction can be incorporated into the GLMM as discussed in Wang et al. (2015). When the link function g is non-linear, the variance components represent the unspecified additive allelic effects or allelic interactions

from FR or PS that contribute to the covariance structure of the g-transformed phenotypic means.

The unified LMM (6) or (7) include multiple variance components and high dimensional random vectors, which makes the model fitting difficult. One way to simplify the model fitting is to treat PS and its interactions with FR as fixed effects. The PS effects can then be assessed via an overall effect of its PCs. Its interactions with FR can also be assessed via the PCs from each random component. The method proposed in Wang et al. (2015) is suitable for family data with moderate family sizes. But when the data set includes large pedigrees, how to fit the LMM to this type of family data using existing software is still a challenge and need further exploration.

7 Acknowledgements

This research was conducted with the UK Biobank Resource, Project number 19746. A.R. received the support of the EU in the framework of the Marie-Curie FP7 COFUND People Programme, through the award of an AgreenSkills/AgreenSkills+ fellowship under grant agreement number 609398.

References

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M. A., Lambourne, J. J., Sivapalaratnam, S., Downes, K., Kundu, K., Bomba, L., Berentsen, K., Bradley, J. R., Daugherty, L. C., Delaneau, O., Freson, K., Garner, S. F., Grassi, L., Guerrero, J., Haimel, M., Janssen-Megens, E. M., Kaan, A., Kamat, M., Kim, B., Mandoli, A., Marchini, J., Martens, J. H. A., Meacham, S., Megy, K., O'Connell, J., Petersen, R., Sharifi, N., Sheard, S. M., Staley, J. R., Tuna, S., van der Ent, M., Walter, K., Wang, S.-Y., Wheeler, E., Wilder, S. P., Iotchkova, V., Moore, C., Sambrook, J., Stunnenberg, H. G., Di Angelantonio, E., Kaptoge, S., Kuijpers, T. W., Carrillo-de Santa-Pau, E., Juan, D., Rico, D., Valencia, A., Chen, L., Ge, B., Vasquez, L., Kwan, T., Garrido-Martn, D., Watt, S., Yang, Y., Guigo, R., Beck, S., Paul, D. S., Pastinen, T., Bujold, D., Bourque, G., Frontini, M., Danesh, J., Roberts, D. J., Ouwehand, W. H., Butterworth, A. S., and Soranzo, N. (2016). The allelic land-

scape of human blood cell trait variation and links to common complex disease. *Cell*, 167:1415–1429.e19.

Cockerham, C. C. (1971). Higher order probability functions of identity of alleles by descent. *Genetics*, 69:235–246.

Dou, J., Sun, B., Sim, X., Hughes, J. D., Reilly, D. F., Tai, E. S., Liu, J., and Wang, C. (2017). Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. *PLoS genetics*, 13:e1007021.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to quantitative genetics.* Pearson Education, 4th edition.

Hoffman, G. E. (2013). Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PloS One*, 8:e75707.

Holsinger, K. E. and Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting f(st). *Nature reviews. Genetics*, 10:639–650.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42:348–354.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178:1709–1723.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.

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

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12:e1001779.

Sun, M., Jobling, M. A., Taliun, D., Pramstaller, P. P., Egeland, T., and Sheehan, N. A. (2016). On the use of dense snp marker data for the identification of distant relative pairs. *Theoretical population biology*, 107:14–25.

Wang, T., He, P., Ahn, K. W., Wang, X., Ghosh, S., and Laud., P. (2015). A reformulation of generalized linear mixed models to fit family data in genetic association studies. *Front Genet.*, 6:120.

Wang, T. and Zeng, Z. B. (2009). Contribution of genetic effects to genetic variance components with epistasis and linkage disequilibrium. *BMC Genetics*, 10:Article 52.

Weir, B. S. (1996). *Genetic data analysis II: methods for discrete population genetic data*. Sinauer, Massachusetts.

Weir, B. S., Anderson, A. D., and Hepler, A. B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nature Reviews. Genetics*, 7:771–780.

Weir, B. S. and Goudet, J. (2017). A unified characterization of population structure and relatedness. *Genetics*, 206:2085–2103.

Wright, S. (1950). Genetical structure of populations. *Nature*, 166(4215):247–249.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88:76–82.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38:203–208.

Zeng, Z.-B., Wang, T., and Zou, W. (2005). Modeling quantitative trait loci and interpretation of models. *Genetics*, 169(3):1711–1725.