Family Analysis with Mendelian Imputations

Augustine Kong[1]*, Stefania Benonisdottir[1], and Alexander I. Young[1,2]*

1 Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, UK
2 Center for Economic and Social Research, University of Southern California, Los Angeles, CA, USA
*corresponding authors: contact augustine.kong@bdi.ox.ac.uk or alextisyoung@gmail.com

Abstract

Genotype-phenotype associations can be results of direct effects, genetic nurturing effects and population stratification confounding. Genotypes from parents and siblings of the proband can be used to statistically disentangle these effects. To maximize power, a comprehensive framework for utilizing various combinations of parents' and siblings' genotypes is introduced. Central to the approach is *mendelian imputation*, a method that utilizes identity by descent (IBD) information to non-linearly impute genotypes into untyped relatives using genotypes of typed individuals. Applying the method to UK Biobank probands with at least one parent or sibling genotyped, for an educational attainment (EA) polygenic score that has an R^2 of 5.7% with EA, its predictive power based on direct genetic effect alone is demonstrated to be only about 1.4%. For women, the EA polygenic score has a bigger estimated direct effect on age-at-first-birth than EA itself.

Introduction

Standard genotype-phenotype association analyses, such as those typically performed for genome-wide association studies (GWAS), involve only the phenotypes and the genotypes of the proband. However, to separate the direct genetic effects from the indirect genetic effects and other confounding factors, in addition to the proband's genotypes, genotypes of family members such as parents and siblings are often necessary[1]. Even though much more family data can be expected in the future, either through deliberate ascertainment or as a consequence of a substantial fraction of the population being genotyped, family data are currently somewhat limited. Thus, for now and for the future, it is important to develop methods that can get the most out of the data available. Here we consider a model where the proband's phenotype depends on the genotypes of four people --- the proband, the parents, and one sibling. When genotypes of one or more family members are unavailable, they are treated as missing-data, and imputed in an appropriate manner. This setup serves two purposes: (a) it allows different data types to be treated under one analytic framework, increasing flexibility and power, (b) statistical efficiency is increased through non-linear imputation of the missing genotypes using the observed genotypes. Genotyped sib-pairs with untyped parents, a common data type, benefit the most from this approach. Compare to standard analyses[2–5], our method of imputing parental genotypes, which incorporates the identical-by-decent (IBD) information between sibs, adds information and allows for the estimation of sibling genetic nurturing effect. Moreover, it includes the modelling of asymmetric sib-pairs, *e.g.* siblings of different gender, and highlights the utility of the genotypes of a sibling whose phenotype is either missing or is not directly comparable with that of the proband.

This paper is organized as follows. (i) The basic model and parameters are introduced. (ii) The fifteen possible genotype data patterns, one complete plus fourteen incomplete, are presented and (iii) the various forms of imputations, linear and IBD-based non-linear, are described and illustrated by examples. (iv) Provide conditions for the estimates obtained using data with imputations to be unbiased, *i.e.* the estimates, while having different standard errors, have the same interpretations as those obtained using complete data. We call this property *estimate consistency*. (v) Illustrate the extension from handling one phenotyped sibling (sib) to two phenotyped sibs and introduce a model that incorporates phenotypic asymmetry between sibs. A note on extension to genotyped sib-ships of size bigger than two. (vi) When the data or

estimates from different missing data patterns are combined, it is a form of multivariate (parameter) meta-analysis[6]. Someimtes the results are not intuitive, *e.g.* estimates often have smaller standard errors than one expects from applying univariate principles. (vii) Extension of analyses of individual variants to that of polygenic scores and discuss the consequences of assortative mating. (viii) Empirical study based on UK Biobank data. (ix) Discussion.

Models and Setup

The proband is defined as the person with known phenotype Y, the response variable. We start with a single-locus model where Y, conditional on the genotypes of the proband and the parents, has expectation

$$E(Y) = constant + \delta G + \alpha_P G_P + \alpha_M G_M \qquad (1.1)$$

where G is the genotype of the proband, G_P is the genotype of the father, G_M is the genotype of the mother, and δ is the direct effect. Y is treated as a quantitative variable, but it does not have to be normally distributed and indeed can be binary. As long as the variance explained by the G's is small relative to the variance of Y, results given will apply exactly or approximately. Note that (1.1) is in effect the same as a model previous used[7] where the explanatory variables are the transmitted and non-transmitted alleles, as the explanatory variables this model are a one-to-one linear transformation of those in the other. The parameters α_P and α_M can be written as

$$\eta_P + \omega \text{ and } \eta_M + \omega. \qquad (1.2)$$

where η_P and η_M denote parent-of-origin (PO) specific genetic nurturing effect, and ω captures all confounding effects that have not been adjusted out, including assortative mating induced confounding. Note that, following our previous work[7], the genetic nurturing effects of the parental alleles, η_P and η_M, are meant to incorporate not only the genetic nurturing effects of the parents, but also include the contributions from older ancestors and siblings. In particular, when the proband has a sibling, the model is extended to

$$E(Y) = constant + \delta G + \eta_S G_S + \beta_P G_P + \beta_M G_M \qquad (1.3)$$

where η_S denote the genetic nurturing effect of the sibling's genotype. Because a parental allele has ½ chance of being passed onto a sibling,

$$\beta_P = \alpha_P - \frac{\eta_S}{2} \text{ and } \beta_M = \alpha_M - \frac{\eta_S}{2}. \qquad (1.4)$$

The α's are more natural parameters than the β's, as the former are well defined regardless of whether the proband has any siblings. Nonetheless, the introduction of η_S and the β's is necessary when family analysis, in the absence of parental genotypes, is performed using the

genotypes of a sibling. While $(\delta, \beta_P, \beta_M, \eta_S)$ is the parameter vector being directed estimated through fitting the model (1.3), its estimate could easily be transformed into an estimate of other parameters vectors such as $(\delta, \alpha, \eta_M - \eta_P, \eta_S)$ where

$$\alpha = \frac{\alpha_P + \alpha_M}{2}. \qquad (1.5)$$

Notice that because the estimates of the four parameters are correlated, estimates of δ and α will change and usually have reduced standard errors if the assumptions that $\eta_P = \eta_M$ and $\eta_S = 0$ are made, a question of bias-variance trade-off. Finally, it is noted that the sum $(\delta + \alpha)$, referred to as the *population effect* here, corresponds to what is being estimated in a proband only genotype-phenotype analysis performed by most GWAS studies.

Treatment of Missing Data

Genotypes (explanatory variables) in the complete-data model (1.3) that are unobserved are treated as missing-data. Including the complete-data case, there are $2^4 - 1 = 15$ complete-missing data patterns (Fig. 1, genotyped individuals shaded). For example, Fig. 1a, 1b, 1e, and 1h, correspond respectively to complete data, parents-proband trios, genotyped sib-pairs, and the standard GWAS *singletons* setup with proband genotyped only. Missing genotypes are imputed either linearly (○ or blank) or non-linearly (+ or ⊕). Linear imputation is predicting an unobserved genotype using a linear combination of the observed genotypes with coefficients that minimize mean squared error (MSE). With alleles coded 0 or 1 and population frequency of allele 1 denoted by p, the four genotypes (G, G_S, G_P, G_M), assuming random mating, have known variance-covariance matrix (Table 1a). Even though the G's are not normally distributed, the formulas for best linear predictions are the same as those established for multivariate normal variables. However, as illustrated below, based on the known non-normal joint distribution of the G's, sometimes a missing genotype can be imputed as a non-linear function of the observed genotypes with higher correlation with the actual genotype than any linear imputation.

There are seven cases in Fig. 1 where IBD-based non-linear imputations are possible. However, with respect to the joint distribution of the genotypes, up to symmetry, there are only 3 distinct equivalent classes --- (1c,1d,1j,1k), (1e), and (1f,1g). One case from each of the equivalent class is covered below.

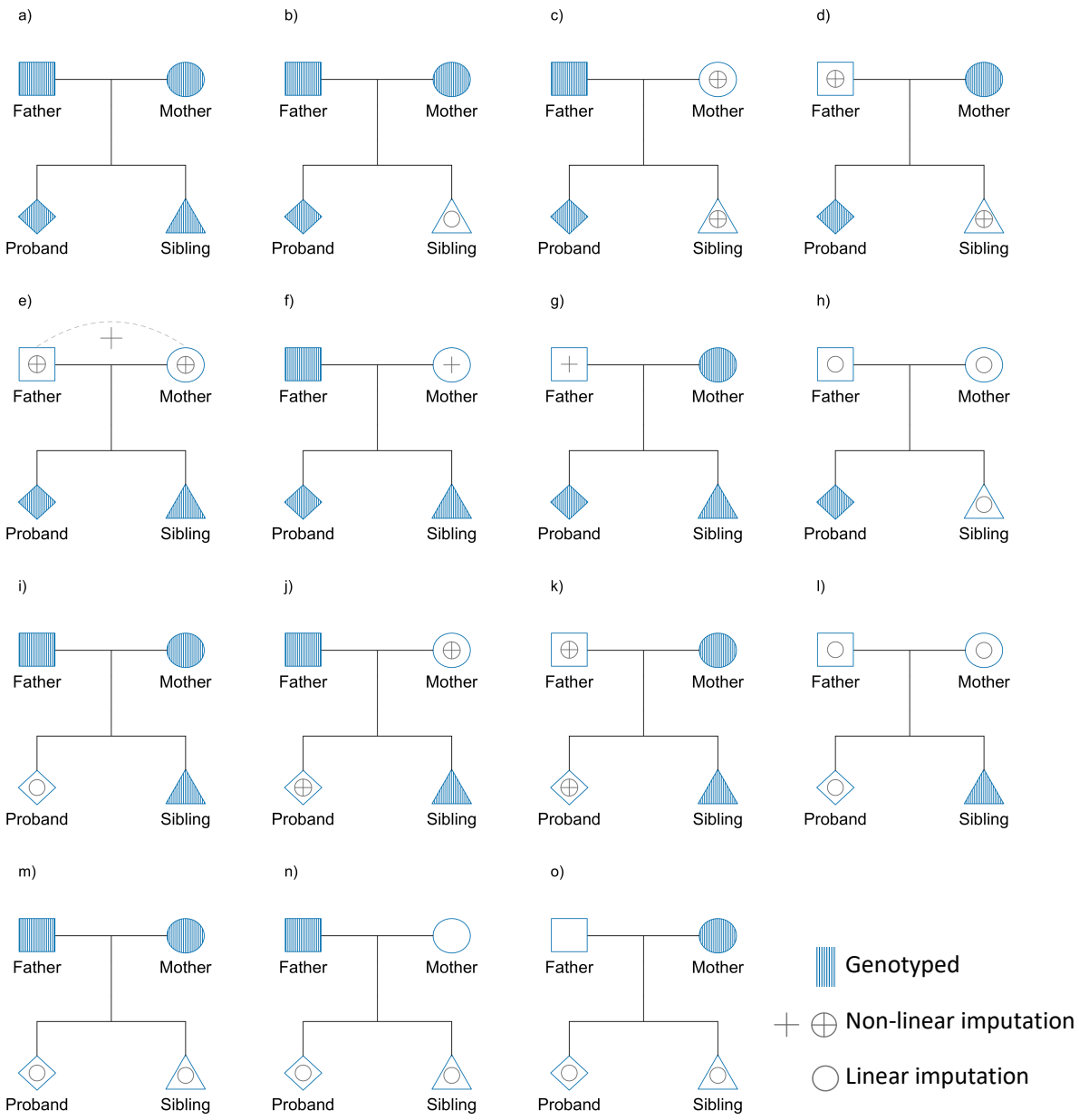

Figure 1: Mendelian imputation for 15 different missing data patterns for nuclear families. Proband refers to an individual with phenotype information. Shaded individuals are directly genotyped. ○ denotes linear imputation from observed genotypes, + denotes non-linear imputation from observed genotypes and IBD information, ⊕ denotes non-linear imputation in the case where the resulting covariance matrix of the four genotypes (observed and imputed) is not of full rank (see Table 1). The genotypes of the blanked mother in 1n and blanked father in 1o are imputed by a constant, the population frequency.

a)
$$\begin{array}{cccc} G & G_S & G_P & G_M \end{array}$$
$$\begin{pmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 0 \\ 1/2 & 1/2 & 0 & 1 \end{pmatrix}$$

b)
$$\begin{array}{cccc} G & G_S^* & G_P & G_M \end{array}$$
$$\begin{pmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 0 \\ 1/2 & 1/2 & 0 & 1 \end{pmatrix}$$

c)
$$\begin{array}{cccc} G & G_S^* & G_P & G_M^* \end{array}$$
$$\begin{pmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 3/8 & 1/2 & 1/4 \\ 1/2 & 1/2 & 1 & 0 \\ 1/2 & 1/4 & 0 & 1/2 \end{pmatrix}$$

d)
$$\begin{array}{cccc} G & G_S^* & G_P^* & G_M \end{array}$$
$$\begin{pmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 3/8 & 1/4 & 1/2 \\ 1/2 & 1/4 & 1/2 & 0 \\ 1/2 & 1/2 & 0 & 1 \end{pmatrix}$$

e)
$$\begin{array}{cccc} G & G_S & G_P^* & G_M^* \end{array}$$
$$\begin{pmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 3/8 & 3/8 \\ 1/2 & 1/2 & 3/8 & 3/8 \end{pmatrix}$$

f)
$$\begin{array}{cccc} G & G_S & G_P & G_M^* \end{array}$$
$$\begin{pmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 0 \\ 1/2 & 1/2 & 0 & 3/4 \end{pmatrix}$$

g)
$$\begin{array}{cccc} G & G_S & G_P^* & G_M \end{array}$$
$$\begin{pmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 3/4 & 0 \\ 1/2 & 1/2 & 0 & 1 \end{pmatrix}$$

h)
$$\begin{array}{cccc} G & G_S^* & G_P^* & G_M^* \end{array}$$
$$\begin{pmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/4 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

i)
$$\begin{array}{cccc} G^* & G_S & G_P & G_M \end{array}$$
$$\begin{pmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 0 \\ 1/2 & 1/2 & 0 & 1 \end{pmatrix}$$

j)
$$\begin{array}{cccc} G^* & G_S & G_P & G_M^* \end{array}$$
$$\begin{pmatrix} 3/8 & 1/2 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 0 \\ 1/4 & 1/2 & 0 & 1/2 \end{pmatrix}$$

k)
$$\begin{array}{cccc} G^* & G_S & G_P^* & G_M \end{array}$$
$$\begin{pmatrix} 3/8 & 1/2 & 1/4 & 1/2 \\ 1/2 & 1 & 1/2 & 1/2 \\ 1/4 & 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 & 1 \end{pmatrix}$$

l)
$$\begin{array}{cccc} G^* & G_S & G_P^* & G_M^* \end{array}$$
$$\begin{pmatrix} 1/4 & 1/2 & 1/4 & 1/4 \\ 1/2 & 1 & 1/2 & 1/2 \\ 1/4 & 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 & 1/4 \end{pmatrix}$$

m)
$$\begin{array}{cccc} G^* & G_S^* & G_P & G_M \end{array}$$
$$\begin{pmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 0 \\ 1/2 & 1/2 & 0 & 1 \end{pmatrix}$$

n)
$$\begin{array}{cccc} G^* & G_S^* & G_P & G_M^* \end{array}$$
$$\begin{pmatrix} 1/4 & 1/4 & 1/2 & 0 \\ 1/4 & 1/4 & 1/2 & 0 \\ 1/2 & 1/2 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

o)
$$\begin{array}{cccc} G^* & G_S^* & G_P^* & G_M \end{array}$$
$$\begin{pmatrix} 1/4 & 1/4 & 0 & 1/2 \\ 1/4 & 1/4 & 0 & 1/2 \\ 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 1 \end{pmatrix}$$

Table 1. Mendelian imputation and resulting scaled variance-covariance matrices. Each matrix shows the variance-covariance structure within the nuclear family given observed and imputed genotypes. The labels a) to o) correspond to the those in Fig. 1. G denotes proband's genotype, G_s denotes sibling's genotype, G_P denotes father's genotype and G_M denotes mother's genotype. * indicates an unobserved genotype that is imputed (either linearly or non-linearly) using observed genotypes. Displayed are $\Sigma_X^S = \Sigma_X/(2p(1-p))$. It is scaled so that the diagonal entry corresponding to an observed genotype is 1.

Mother and proband genotyped. In Fig. 1d, G and G_M are observed, but not G_P and G_S. Based on the known variance-covariance matrix of the genotypes,

$$G_P^{*L} = (2/3)\,G - (1/3)G_M + (4/3)p, \qquad (2.1)$$

is the best linear predictor of G_P (formula for computing linear imputations in Supplementary Information), with $E(G_P^{*L}) = 2p$, and $cor^2(G_P, G_P^{*L}) = 1/3$. G_P can be decomposed as T_P and NT_P, denoting respectively the allele transmitted to the proband and the allele not transmitted. If T_P is known, the best and natural prediction of G_P is

$$G_P^* = T_P + p. \qquad (2.2)$$

satisfying $E(G_P^*) = 2p$, and $cor^2(G_P, G_P^*) = 1/2$. T_P is equal to $G - T_M$, where T_M is the allele transmitted from mother to proband. Given G and G_M, T_M is known unless G and G_M are both heterozygotes. In that case, unless the target SNP is very close to a recombination event in the mother-proband meiosis, T_M can be deduced through a phased neighbouring SNP which is homozygous for one member of the mother-proband pair and heterozygous for the other (Fig. 2).

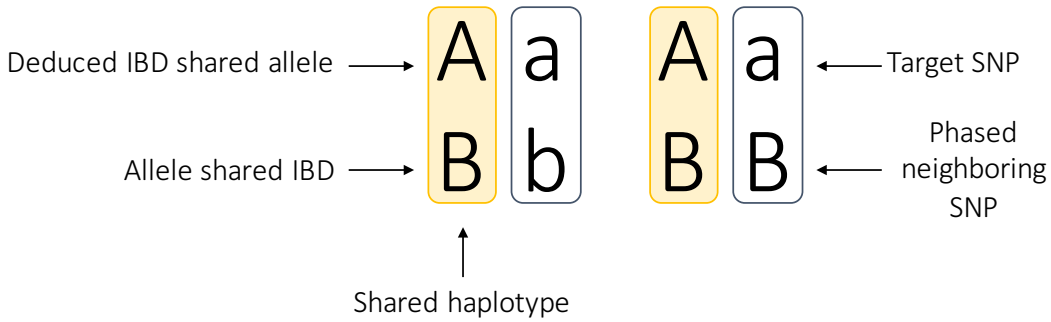

Figure 2: How the allele shared IBD between two individuals can be determined when both are heterozygous at the target SNP. This applies to a parent-offspring pair, who always shared one allele IBD, and a sibling pair at locations where they are determined to be sharing one allele IBD. A neighbouring SNP which has been phased with the target SNP, and is homozygous for one sib and heterozygous for the other is employed to resolve the uncertainty. For the individual on the left above, the B allele must be the allele shared with the individual on the right. Thus through the phased haplotype A-B, it is determined that allele A, as opposed to a, is the shared IBD allele.

To avoid the phasing step, with some loss of information, one could impute G_P by $2p$ when G and G_M are both heterozygotes[8], an event occurring with probability $p(1-p)$. Relative to linear imputation, the increased correlation between G_P and G_P^* leads to increased information for parameter estimation. If data with this pattern are analysed on its own, as opposed to mixing together with other patterns, then linear imputation corresponds to no imputation at all. In particular, if Y is regressed on G and G_M only, it can be shown that the fitted coefficients have expectations $[\delta + (2/3)\alpha_P]$ and $[\alpha_M - (1/3)\alpha_P]$ respectively. Without the assumption that $\alpha_M = \alpha_P$, it is not possible to obtain an unbiased estimate of δ. By contrast, by regressing Y on G, G_M and G_P^*, in which case the fitted coefficients would have expectations δ, α_M and α_P respectively, the same as if G_P is observed and included in the regression. However, replacing G_P by G_P^* would not just change the variances of individual parameter estimates, it would have an impact on the whole variance-covariance structure of the correlated estimates. Also, if the assumption $\alpha_M = \alpha_P$ is made, then δ can be estimated with or without non-linear imputations, but the variance of the estimate obtained with imputation is only $3/4$ of that of the estimate obtained without imputation. If data with this missing data pattern are analysed jointly with other data in a single regression with all four explanatory variables included, then $G_S^* = (G_M + G_P^*)/2$. However, even though G_S^* is a non-linear function of G and G_M, it is a linear function of G_M and G_P^*. Thus, if data of this pattern are analysed by themselves, with G_P^* already

included in the regression, adding G_S^* would only introduce collinearity (indicated by $\oplus$ in Fig. 1).

Siblings genotyped only. In Fig. 1e, G and G_S are observed, but not G_P and G_M. Without PO information, the best linear prediction of $G_{PM} = G_P + G_M$ is

$$G_{PM}^{*L} = (2/3)G + (2/3)G_S + (4/3)p, \qquad (2.3)$$

with $G_P^{*L} = G_M^{*L} = G_{PM}^{*L}/2$. At an autosomal locus, two siblings share 0, 1, or 2, alleles IBD with probability $1/4$, $1/2$, and $1/4$ respectively. With the large number of SNPs included in any of the recent genome-wide genotyping arrays, given their genotypes, the number of alleles shared IBD between a specific sib-pair at a specific locus can usually be determined quite accurately[9] unless the locus is close to one of the paternal or maternal recombination events for the siblings, which happens infrequently. With IBD number known (Fig. 3), G_{PM} can be non-linear imputed as

$$G_{PM}^* = [sum\ of\ observed\ alleles] + p \times [number\ of\ unobserved\ alleles] \quad (2.4)$$

with $G_P^* = G_M^* = G_{PM}^*/2$. The number of unobserved parental alleles is equal to four minus the IBD number, and observed parental alleles are summed without double counting. Similar to the *parent-proband* case (Fig. 2), when IBD = 1 and both siblings are heterozygous at the target SNP, there is uncertainty as to which is the IBD shared allele. This can again be resolved through a neighbouring phased SNP which is homozygous for one sib and heterozygous for the other. If phasing is not performed, with some loss of information, G_{PM} can be imputed as $1 + 2p$ in this situation[8]. Assuming the double-heterozygous situation is resolved through phasing, it can be shown that $cor^2(G_{PM}, G_{PM}^*) = (3/4)$, an increase over $cor^2(G_{PM}, G_{PM}^{*L}) = 2/3$. If sib-pair data are analysed by themselves, since $G_P^* = G_M^*$, it is natural to regress Y on G, G_S and G_{PM}^*. The respective fitted coefficients are then unbiased estimates of δ, η_S, and $\beta = (\beta_P + \beta_M)/2$. By contrast, if Y is regressed on G and G_S only, the respective fitted coefficients have expectations $[\delta + (2/3)\beta]$ and $[\eta_S + (2/3)\beta]$ respectively. By taking the difference of these two fitted coefficients, one can obtain an unbiased estimate for $(\delta - \eta_s)$. Without imputations, δ cannot be estimated without bias unless the assumption $\eta_S = 0$ is made. Even then this estimate is suboptimal: when only one sib is a proband, *i.e.* phenotyped, it has a variance $4/3$ times that of the estimate of δ obtained with imputations and conditioning on $\eta_S = 0$ (see below for the double-proband case).

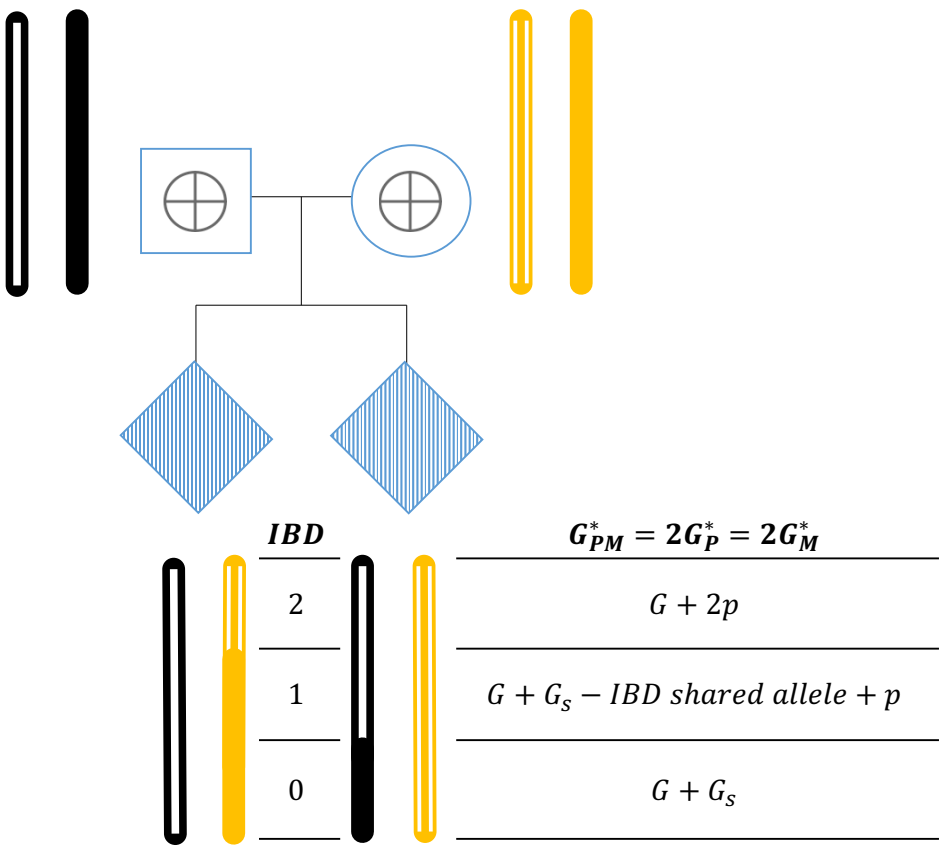

Figure 3: Mendelian imputation of parental alleles given the IBD status of genotyped sibling pairs. Parent-offspring pairs share one allele IBD at each locus. Siblings share 0, 1 or 2 alleles IBD with probabilities $1/4, 1/2.$, and $1/4$ respectively. Given array genotypes, the number of IBD alleles shared in realization can often be determined with little uncertainty. Illustrated is how $(G_P + G_M)$ is imputed given IBD number

Two siblings and mother genotyped. In Fig. 1g, G, G_S, and G_M are observed. Similar to the previous two examples, if the paternal alleles transmitted to the proband and sibling are estimated to be non-IBD, then G_P^* is the sum of those two alleles. Otherwise, G_P^* is the common paternal allele plus p. By regressing Y on G, G_S, G_M and G_P^*, the fitted coefficients are unbiased estimates of δ, η_S, β_M, and β_P, the same as with complete data.

Estimate Consistency and Multivariate Meta-Analysis

Let $\vec{G} = (G, G_S, G_P, G_M)$ and let $\vec{X} = (X_1, \dots, X_t), t \leq 4$, be explanatory variables used in an analysis satisfying the condition that the phenotype Y is correlated with $\vec{X}$ only through $\vec{G}$, *i.e.* Y is conditional independent of $\vec{X}$ given $\vec{G}$. Let matrices Σ, Σ_X, and Σ_{cov}, be respectively $var(\vec{G}), var(\vec{X})$, and $cov(\vec{X}, \vec{G})$. The model (1.3) can be rewritten as

$$Y = constant + \vec{G}'\vec{\theta} + \mathcal{E}. \qquad (3.1)$$

where $\vec{\theta} = (\delta, \eta_S, \beta_P, \beta_M)$, and $\mathcal{E}$ and $\vec{G}$ are uncorrelated. If Y is regressed on $\vec{X}$ linearly, the corresponding model is

$$Y = constant + \vec{X}'\vec{\gamma} + \epsilon, \qquad (3.2)$$

with $\vec{\gamma}$ satisfying

$$\Sigma_X \vec{\gamma} = \Sigma_{cov}\vec{\theta}. \qquad (3.3)$$

If Σ_X is of full rank, then

$$\vec{\gamma} = \Sigma_X^{-1}\Sigma_{cov}\vec{\theta}. \qquad (3.4)$$

In addition to being used in earlier examples to calculate the expectations of the fitted coefficients when $\vec{X}$ consists of a subset of the G's, this formula is key to understanding the regressions performed with imputed genotypes. For the proposed imputations, Table 1 gives

$$\Sigma_X^S = \frac{\Sigma_X}{2p(1-p)} \qquad (3.5)$$

for the fifteen data patterns in Fig. 1. Σ_X^S is a scaled version of Σ_X with the property that the diagonal entry of an observed genotype is one, *e.g.* instead of 0 or 1, alleles are coded as 0 or $1/\sqrt{2p(1-p)}$. (Σ_X^S is given for linear imputations in Supplementary Table 1.) Consider the Fig. 1g example studied earlier. Here $\vec{X} = (G, G_S, G_P^*, G_M)$. Regardless of how G_P^* is computed, because of the overlapping variables in $\vec{G}$ and $\vec{X}$, matrices Σ, Σ_X, and Σ_{cov}, are by definition the same apart from entries in row 3 and column 3, and Σ_X and Σ_{cov} are the same except for column 3. The imputation G_P^* we proposed further ensures that $\Sigma_X = \Sigma$ for all entries except entry [3, 3], where $\Sigma_X[3,3]/\Sigma[3,3] = 3/4$ (equivalent relationships reflected by Σ_X^S in Table 1a and Table 1g), and $\Sigma_X = \Sigma_{cov}$ for all entries (Supplementary Information). It follows that $\Sigma_X^{-1}\Sigma_{cov} = I$ and, most importantly

$$\vec{\gamma} = \vec{\theta}, \qquad (3.6)$$

the property we call *estimate consistency*. With our proposed imputations, the relationships between the matrices extend to all fifteen patterns in Fig. 1 (Supplementary Information). In particular, entries of Σ_X and Σ, and equivalently their scaled versions, are equal except for those entries where both indexes tag imputed genotypes, and

$$\Sigma_X = \Sigma_{cov}. \qquad (3.7)$$

The latter also implies that for an imputed genotype, the corresponding diagonal element in Σ_X^S (Table 1) is the correlation-squared between actual and imputed genotype. For example, in the 1g case,

$$cor^2(G_P^*, G_P) = \frac{cov^2(G_P^*, G_P)}{var(G_P^*)var(G_P)} = \frac{var^2(G_P^*)}{var(G_P^*)var(G_P)} = \frac{var(G_P^*)}{2p(1-p)} = \frac{3}{4}. \quad (3.8)$$

Among the fifteen patterns, seven involve non-linear imputations (+ or ⊕) and $rank(\Sigma_X) =$ number of genotyped (shaded) family members plus 1. For the others, $rank(\Sigma_X) =$ number of genotyped members. Thus Σ_X is of full rank for 1a, 1f and 1g. For the other twelve cases, if the data with any of these patterns are analysed by themselves, the number of explanatory variables would have to be reduced to eliminate collinearity, as demonstrated above for 1d and 1e. However, with data from more than one pattern, by imputing all the missing genotypes, with both linear and nonlinear imputations, a single regression can be performed with all the data mixed together. Specifically, consider data from k different patterns indexed by i. For $i = 1, \ldots, k$, let n_i be the sample sizes, $n = \sum_{i=1}^{k} n_i$, $w_i = n_i/n$, and Σ_{Xi} and Σ_{covi} be the Σ_X and Σ_{cov} of pattern i. The combined data have sample size n and variance-covariance matrix

$$\Sigma_{Xcomb} = \sum_{i=1}^{k}[w_i\Sigma_{Xi}] = \sum_{i=1}^{k}[w_i\Sigma_{covi}] = \Sigma_{covcomb}. \quad (3.9)$$

This is because $\Sigma_{Xi} = \Sigma_{covi}$ for each i, and our imputations satisfy $E(\vec{X}) = E(\vec{G})$ for all data patterns. If Σ_{Xcomb} is of full rank, then the combined data can be analysed by one regression based on the complete-data model (1.3). Notice that Σ_{Xcomb} would be full rank as long as we have some data from patterns 1a, 1f or 1g. It would also be of full rank if the data include pattern 1e and cases from either 1b, 1c or 1d. If Σ_{Xcomb} is of rank 3, and the data do not include any genotyped sibling, then model (1.1) can be considered as the complete-data model. Similarly, if Σ_{Xcomb} is of rank 3 and the parents' genotypes are always missing, then fitting a model with the parental genotypes combined is appropriate. In general, individually, the different data patterns have different variance-covariance structures for the explanatory variables and through their inverses impact the variance-covariance structures of the parameter estimates.

Both siblings are phenotyped

Here we consider the case where both siblings are phenotyped and thus both are probands. Let Y_1 and Y_2 denote respectively the phenotypes of sib1 and sib2, and let G and G_S denote their respective genotypes. Let $\vec{\theta}_1 = (\delta_1, \eta_{S1}, \beta_{P1}, \beta_{M1})$, and $\vec{\vartheta}_2 = (\eta_{S2}, \delta_2, \beta_{P2}, \beta_{M2})$. The complete genotype-data model is

$$Y_1 = constant + \vec{G}'\vec{\theta}_1 + \mathcal{E}_1, \quad (4.1a)$$

and

$$Y_2 = constant + \vec{G}'\vec{\vartheta}_2 + \mathcal{E}_2. \qquad (4.1b)$$

The subscript of 1 or 2 for the parameters allows for asymmetry between the sibs. For example, if sib1 is male and sib2 is female, or sib1 is the younger sib, the parameters could take on different values. Assuming Y_1 and Y_2 to be each standardized to have variance one, under asymmetry, $var(\mathcal{E}_1)$ and $var(\mathcal{E}_2)$ are not necessarily equal. For simplicity, we assume the difference is negligible and denote their average value as σ^2. The parameters (and their corresponding estimates) can be reparametrize as averages and differences:

$$\vec{\theta} \stackrel{\text{def}}{=} \begin{pmatrix} \delta \\ \eta_S \\ \beta_P \\ \beta_M \end{pmatrix} \stackrel{\text{def}}{=} \frac{1}{2} \times \begin{pmatrix} \delta_1 + \delta_2 \\ \eta_{S1} + \eta_{S2} \\ \beta_{P1} + \beta_{P2} \\ \beta_{M1} + \beta_{M2} \end{pmatrix}, \quad \vec{\theta}_- \stackrel{\text{def}}{=} \begin{pmatrix} \delta_- \\ \eta_{S-} \\ \beta_{P-} \\ \beta_{M-} \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} \delta_1 - \delta_2 \\ \eta_{S1} - \eta_{s2} \\ \beta_{P1} - \beta_{P2} \\ \beta_{M1} - \beta_{M2} \end{pmatrix}. \qquad (4.2)$$

From (4.2), related parameters are similarly defined, *e.g.* $\beta = (\beta_P + \beta_M)/2$, $\alpha = \beta + \eta_S/2$, and $\alpha_- = (\alpha_1 - \alpha_2)$. Assuming symmetry corresponds to conditioning on $\vec{\theta}_- = 0$. Estimates can be obtained by performing regressions that correspond to (4.1a) and (4.1b). However, because $\mathcal{E}_1$ and $\mathcal{E}_2$ are correlated, determining the variance-covariance matrix of the parameters is more complicated. In Supplementary Information, we show how to do that by reparameterizing the responses to $Y_- \stackrel{\text{def}}{=} Y_1 - Y_2$ and $Y_+ \stackrel{\text{def}}{=} Y_1 + Y_2$. Here we focus on the case where only the siblings are genotyped, one of the most common data-type. The variance-covariance of the average parameters (Supplementary Information) are

$$var\begin{pmatrix} \hat{\delta} \\ \hat{\eta}_S \\ \hat{\alpha} \end{pmatrix} \cong \begin{pmatrix} 2+r & 1+2r & -\left(\frac{3}{2}+r\right) \\ 1+2r & 2+r & -\left(1+\frac{3}{2}r\right) \\ -\left(\frac{3}{2}+r\right) & -\left(1+\frac{3}{2}r\right) & \frac{3}{2}+\frac{5}{4}r \end{pmatrix} \times \frac{\sigma^2}{n \times \pi}, \qquad (4.3)$$

where $\pi = 2p(1-p)$, $r = cor(\mathcal{E}_1, \mathcal{E}_2)$, and n is the number of sib-pairs. The variance-covariance matrix of the estimates of the difference parameters, which are uncorrelated with the estimates of the average parameters, are given in the Supplementary Information. Here, if we condition on $\eta_S = 0$, then estimate of (δ, α) is

$$\begin{pmatrix} \ddot{\delta} \\ \ddot{\alpha} \end{pmatrix} = \begin{pmatrix} \hat{\delta} \\ \hat{\alpha} \end{pmatrix} + \begin{pmatrix} \frac{-(1+2r)}{2+r} \\ \frac{1+3r/2}{2+r} \end{pmatrix} \hat{\eta}_S \qquad (4.4)$$

with variance

$$\begin{pmatrix} 3(1-r) & -2(1-r) \\ -2(1-r) & (2-r) \end{pmatrix} \times \frac{1+r}{2+r} \times \frac{\sigma^2}{n \times \pi}. \qquad (4.5)$$

Note that $var(\hat{\delta})$ is an increasing function of r, but $var(\ddot{\delta})$ is a decreasing function of r. Thus, for estimating δ, one positively correlated double-proband sib-pair is less informative than two single proband sib-pairs without assuming η_S equal to zero, but the opposite if η_S is assumed to be zero. By contrast, without imputations, by regressing Y_- on $(G - G_S)$ and $(G + G_S)$, the fitted coefficient for $(G - G_S)$, denoted by $\tilde{\delta}$, has expectation $(\delta - \eta_S)$. When Y_+ is regressed on $(G - G_S)$ and $(G + G_S)$, the fitted coefficient of $(G + G_S)$, denoted by $\tilde{\varphi}$, has expectation $\delta + \eta_S + (4/3)\beta$. It follows that $\tilde{\delta}$ is an unbiased estimate of $(\delta - \eta_S)$, and $(3/4)(\tilde{\varphi} - \tilde{\delta})$is an unbiased estimate of $\beta + (3/2)\eta_S = \alpha + \eta_S$. The variance-covariance of $(\tilde{\delta}, (3/4)(\tilde{\varphi} - \tilde{\delta})$ is approximately

$$\begin{pmatrix} 2(1-r) & -(3/2)(1-r) \\ -(3/2)(1-r) & (3/2) - (3/4)r \end{pmatrix} \times \frac{\sigma^2}{n \times \pi}. \qquad (4.6)$$

Thus, if $\eta_S \neq 0$, using $\tilde{\delta}$ to estimate δ has a bias of $-\eta_S$. By comparison, using $\ddot{\delta}$ to estimate δ, the bias shrinks to $-[(1 + 2r)/(2 + r)]\eta_S$. Even if η_S is actually zero, $\ddot{\delta}$ is more efficient, as measured by the inverse of the variance, than $\tilde{\delta}$. Assuming $\eta_S = 0$, Fig. 4 shows the efficiency of estimating δ, as a function of r, using n double-proband sibpairs. Solid black line is without imputation of parental genotypes and solid red line is with imputations. The efficiency presented is relative to $2n$ single proband genotyped sibpairs without imputations. Notably, if $r = 0$, the efficiency of n double-proband sibpairs is statistically equivalent to $2n$ single probands. Comparing the solid red line with the solid black line shows that, for $r = (0, 0.1, 0.2, 0.3)$, $\ddot{\delta}$ is $(33, 27, 22, 18)\%$ more efficient than $\tilde{\delta}$.

While data from *singletons* (Fig. 1h), by themselves, cannot provide an unbiased estimate of δ, as an augmentation to the sibpair data, they can increase the efficiency of estimating δ and other parameters, a characteristic of multivariate meta-analysis. Without imputations, for $r = (0, 0.1, 0.2, 0.3)$, adding $16n$ singletons to n genotyped-phenotyped sibpairs increases the efficiency respectively from $(1.0, 1.11, 1.25, 1.43)$ to $(1.29, 1.37, 1.49, 1.66)$ (broken black line in Fig. 4), a percentage increase of $(29, 24, 20, 16)$. With imputations, efficiency increases from $(1.33, 1.41, 1.53, 1.68)$ to $(1.89, 1.93, 2.01, 2.13)$ (broken red line in Fig. 4), a percentage increase of $(42, 37, 31, 27)$. As a consequence, when augmented by the singletons, for $r = (0, 0.1, 0.2, 0.3)$, imputing parental genotypes increases efficiency by a percentage of $(47, 31, 34, 29)$ respectively (by comparing the broken red line with the broken black line). The 16-fold *singletons* versus double-proband sibpairs chosen for demonstration here is approximately the ratio seen in the UK Biobank samples. In Fig. 4, the broken black line is just

below the solid red line. Indeed, if the sibpairs are augmented by an infinite number of singletons, then the two lines would coincide, *i.e.* the efficiency gain for estimating δ through IBD-based imputation of parental genotypes is the same as augmenting by practically an infinite number of singletons.

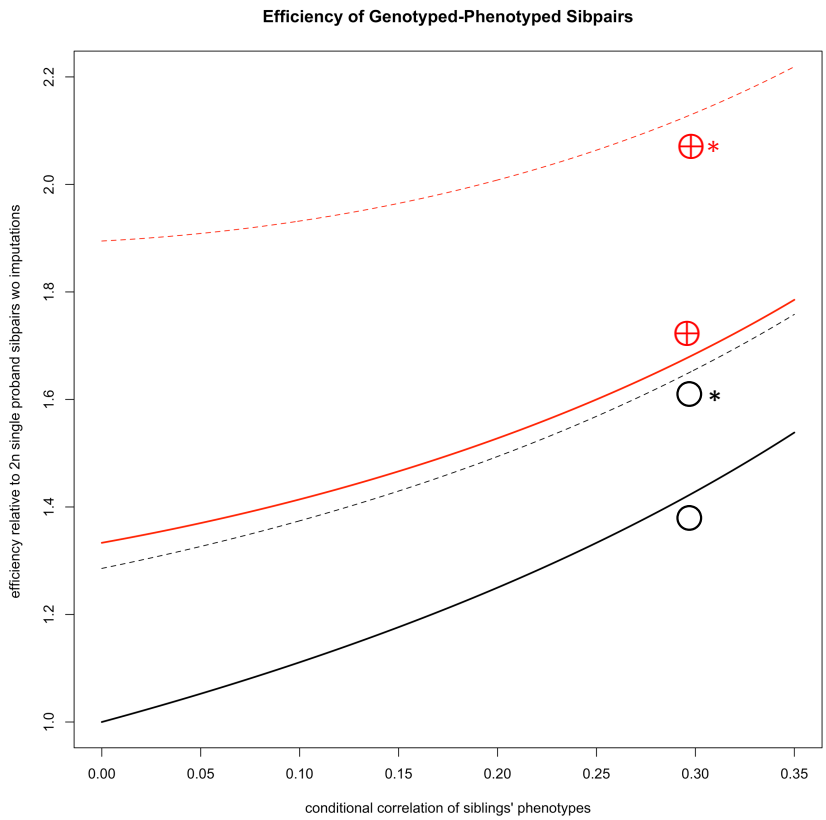

Figure 4: Efficiency of n double-proband genotyped sibpairs in estimating δ, under the assumption of $\eta_S = 0$. The parental genotypes are assumed missing, *i.e.* the case in Fig 1e. Efficiency is displayed relative to $2n$ single-proband genotyped sibpairs and as a function of $r = cor(\mathcal{E}_1, \mathcal{E}_2)$. Black solid line (○) represents efficiency without imputation of parental genotypes. Red solid line (⊕) is efficiency with imputations. Black broken line (○∗) is without imputation and augmented by $16n$ singletons. Red broken line (⊕∗) is with imputations and augmented by $16n$ singletons.

It is noted that our approach to sib-pairs can be extended naturally to incorporate sib-ships with three or more genotyped sibs[8]. In particular, with k genotyped sibs, on average $4(1 - 2^{-k})$ of the parental alleles can be deduced.

Polygenic Scores and Assortative Mating

Consider a polygenic score of T SNPs:

$$PGS = \sum_{t=1}^{T} w_t G_t \,, \qquad (5.1)$$

where G_t and w_t denote respectively genotype and weight of SNP t. When a person is not genotyped, the imputed polygenic score is

$$PGS^* = \sum_{t=1}^{T} w_t G_t^* \,, \qquad (5.2)$$

where each G_t^* is imputed as before. Assuming the T SNPs are in linkage equilibrium with each other, then essentially all the previous results apply. For example, if the observed and imputed PGSs are jointly scaled so that $var(PGS) = 1$, then Table 1 gives the variance-covariance matrix of the observed and imputed PGS for various missing data patterns. Most importantly, if (3.7) holds for individual SNP genotypes, then it also holds for the polygenic scores since the variance-covariance matrix of the PGSs is just a weighted average of the variance-covariance matrixes of the individual SNP genotypes. If model (1.3) is generalized as

$$E(Y) = constant + \delta PGS + \eta_S PGS_S + \beta_P PGS_P + \beta_M PGS_M, \qquad (5.3)$$

then estimate consistency will continue to apply with analysis performed with imputed polygenic scores. In practice, correlations, *i.e.* linkage disequilibrium (LD), between some of the SNPs are expected. However, as long as many SNPs contribute to the polygenic score and only a small fraction of the SNP pairs have non-negligible correlation, the effect on the imputations and estimates would be negligible. The phenomenon that requires consideration is assortative mating (ASM). For a trait with substantial assortative mating, contributing SNPs can become correlated regardless of their relative physical positions, an effect that is reduced, but usually not eliminated, by principal component (PC) adjustments[1]. Effects of ASM on genotype-phenotype associations are in general subtle and complicated, and have to be treated on a case by case basis. With imputations of parental genotypes based on genotypes of sib-pairs, ASM would lead to deviations between Σ_X and Σ_{cov}, *i.e.* violation of (3.7) and estimate consistency. However, as long as the trait is highly polygenic, *i.e.* the genetic component not dominated by a few variants, the estimates of δ and η_S remain essentially unbiased. The estimate of β or α will be magnified by a multiplicative factor, but the degree of magnification is small unless the case is extreme. For example, if couples' trait correlation is 0.30 and trait-PGS R^2 is 0.35, the multiplicative factor is around 1.05. By contrast, if imputations are not

performed and Y is regressed on PGS and PGS_S only, the fitted coefficient of PGS_S has expectation $\eta_S + (2/3)\beta$ under random mating. With ASM as described, the fitted coefficient has expectation around $\eta_S + .71 \times \beta = \eta_S + (1.065) \times (2/3) \times \beta$. Thus IBD-based imputations actually reduce bias relative to no imputations. Most importantly, if deemed necessary, observed correlation of PGS and PGS_S, which would go substantially above 0.5 with strong ASM, can be used to adjust the parameter estimates. It is noted that the bias referred to here is about using imputed data relative to having complete data. Even with complete data, as described previously[7], as a consequence of ASM, estimates of α or β would capture, in addition to indirect effects, some confounding effects due to the parental *PGS*s being correlated with the part of the genetic component of the traits that is not captured by the *PGS* studied.

Empirical Study

Using the developed methods, the effects of a EA polygenic score on EA, age-at-first-birth (AAFB)[10], height (HT) and body-mass-index (BMI) are examined. The weights of 510,290 SNPs underlying this polygenic score are calculated (Supplementary Information) based on a GWAS meta-analysis that includes 608,402 samples, a subset of a larger set[2], and includes 350,000 samples from UKB. A set of more than 39,000 probands from UKB with at least one sibling/parent genotyped is used to estimate the various effects of the polygenic score. The 350,000 individuals do not include any of the 39,000 probands or any third degree or higher relatives of the probands. Phenotypes are standardized for males and females separately to have variance one after adjusting for year-of-birth and 40 principal components. The number of probands for AAFB is about 40% of that for the other traits because AAFB information is only available for women from UKB and only applies to those who have children. This reduces sample size, but, in contrast to other sib-pair analyses that required both siblings to be phenotyped, our AAFB analysis includes 5216 probands, one-third of the total, with only genotyped male siblings.

After standardizing the polygenic scores so that those computed from observed genotypes have variance one, estimates of $\delta, \eta_S, \beta_M,$ and β_P are obtained. After reparametrization, values of $\hat{\delta}$, $\hat{\eta}_S$, $\hat{\alpha} \stackrel{\text{def}}{=} (\hat{\alpha}_M + \hat{\alpha}_P)/2$, and $(\hat{\alpha}_M - \hat{\alpha}_P)$ are displayed in Supplementary Table 2. For all four traits, $\hat{\eta}_S$ and $(\hat{\alpha}_M - \hat{\alpha}_P)$ are not significantly different from zero. This does not mean η_S is zero or that there are no parent-of-origin effects, only that they are not large enough to detect at our sample size. To simplify and to reduce variance, accepting the possibility of introducing

some small bias, estimates $\ddot{\delta}$, $\ddot{\alpha} \stackrel{\text{def}}{=} (\ddot{\alpha}_M + \ddot{\alpha}_P)/2$, and $(\ddot{\alpha}_M - \ddot{\alpha}_P)$ are computed conditioning on $\eta_S = 0$ (Table 2). For all four traits, $\ddot{\delta}$ and $\ddot{\alpha}$ are highly significant (absolute effect size at least 5 times standard error (SE)). For EA, AAFB, BMI and HT, the estimated direct-population effect ratio, $\ddot{\delta}/(\ddot{\delta} + \ddot{\alpha})$, is 0.49, 0.66, 0.63 and 0.44 respectively. By comparison, for a different but related EA polygenic score applied to Icelandic data[7], the corresponding ratio estimates are 0.70, 0.64, 0.72, and 0.42. The estimates are broadly consistent with the exception of EA. As R^2, or variance explained is proportional to effect[2], $\ddot{\delta}/(\ddot{\delta} + \ddot{\alpha}) = 0.49$ means the variance explained by the direct effect alone is only $0.49^2 = 0.24$ of the variance explained by the direct effect. In absolute terms, the population effect is estimated to explain $(0.118 + 0.121)^2 = 5.7\%$ of the variance of EA, while it is only $0.118^2 = 1.4\%$ for the direct effect alone. Another striking result with the EA polygenic score is that its estimated direct effect for AAFB, $\ddot{\delta} = 0.135$, is higher than that for EA itself, $\ddot{\delta} = 0.118$. To make a more direct comparison, analysis for EA is recalculated using the same AAFB probands, and the estimated direct effect is 0.114. For these probands, correlation of EA and AAFB is 0.24, significant but not extremely high, indicating the polygenic score influences EA and AAFB mainly through separate causal paths, not purely affecting one trait through the other.

trait	$\ddot{\delta}$	*(SE)*	$\ddot{\alpha}$	*(SE)*	$\ddot{\alpha}_M - \ddot{\alpha}_P$	*(SE)*
EA	0.118	(0.008)	0.121	(0.007)	-0.017	(0.026)
AAFB	0.135	(0.014)	0.069	(0.011)	0.033	(0.044)
BMI	-0.068	(0.009)	-0.041	(0.007)	0.027	(0.027)
HT	0.04	(0.007)	0.052	(0.007)	-0.017	(0.027)

Table 2. Estimated effects of an EA polygenic scores. Estimates of direct effect ($\ddot{\delta}$), average parental effect ($\ddot{\alpha}$), and the difference of the parental effects ($\ddot{\alpha}_M - \ddot{\alpha}_P$), for the traits educational attainment (EA), age at first birth (AAFB), body mass index (BMI), and height (HT). Sibling effect η_S is assumed to be zero. SE denotes standard error. Descriptions of the polygenic score and the UKB samples used are in the Supplementary Information.

It is noted that the singleton probands in the UKB data set cannot be used to augment the family analysis here because these singletons are part of the GWAS sample used to obtian the weights of the polygenic score.

Discussion

We introduce Mendelian imputations as a tool to perform family-based association analysis. Conceptually, this is similar to multipoint linkage analysis performed with pedigrees that include deceased members. It is also related to familial imputations (also called in silico genealogy-based genotyping)[11,12] and association by proxy[13,14] where genotypes of relatives are used to associate with phenotypes of un-genotyped probands. Even though our general framework can also incorporate association by proxy, e.g. Fig. 1i to 1o, the main focus here is to disentangle various effects that contribute to the associations between a proband's genotypes and phenotypes. As such, Mendelian imputations should also be applicable to family-based Mendelian randomization studies using genotyped sib-pairs[15].

Mendelian imputations allow us to combine data with different missing data patterns in a single analysis, maximizing power. Moreover, even with one data type, Mendelian imputation increases flexibility and power. Specifically, for genotyped sib-pairs, it is shown that a genotyped sibling of the proband, even without phenotype, can be used. When both sibs are phenotyped, we can examine whether the genotypes have different effects on the sibs with respect to birth order or gender through direct or indirect effects. Non-linear imputations of parents not only increase power, but allow for the estimation of sibling nurturing effect, and enable unbiased estimation of the direct effect when the sibling nurturing effect is present. A recent manuscript[16] proposed an imputation method that imputes each SNP without using IBD information inferred from neighbouring SNPs, *e.g.* when the two siblings are discordant homozygotes, then all four parental alleles can be inferred. While this method improves on not imputing at all, the gain is small relative to our method[8], analogous to the difference between single-point and multipoint linkage analyses.

By applying the proposed method to UKB data, in addition to replicating observations previously reported based on Icelandic data, there are two thought provoking results. The low direct-associate effect ratio of the EA polygenic score on EA might be because many UK samples are included in the GWAS being used to construct the PGS, and thus part of the PGS's predictive power with EA in other UKB samples could be population stratification effects that have not been eliminated by PC adjustments. It is known that the genetic components underling EA and AAFB have substantial overlap[17], and the genetic component to EA was shown to have a stronger effect on AAFB of women than that of men in Iceland[18]. Despite that, the fact that the EA polygenic has a higher estimated direct effect on AAFB than EA is surprising. This highlights the complexity of the nature of the genetic variants influencing EA and reproductive

traits. The issues raised here have important implications for how to interpret the population effects of a polygenic score, its portability and policy use. Family based analysis with more data and improved methodology could play a major role in the future research in this area.

Acknowledgements This work was supported by the Li Ka Shing Foundation. We thank the UK Biobank (application ID 11867). We thank Aysu Okbay for providing educational attainment summary statistics.

References

1. Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science.* **365**, 1396–1400 (2019).

2. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

3. Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* **8**, 1–47 (2019).

4. Mostafavi, H. *et al.* Variable prediction accuracy of polygenic scores within an ancestry group. *Elife* **9**, (2020).

5. Fulker, D. W., Cherny, S. S., Sham, P. C. & Hewitt, J. K. Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* **64**, 259–267 (1999).

6. Jackson, D., Riley, R. & White, I. R. Multivariate meta-analysis: potential and promise. *Stat. Med.* **30**, 2481–2498 (2011).

7. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science.* **359**, 424–428 (2018).

8. Young, A. I. *et al.* Mendelian imputation of parental genotypes for genome-wide estimation of direct and indirect genetic effects. *BioRxiv* (2020) doi:10.1101/185199.

9. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

10. Barban, N. *et al.* Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat. Genet.* **48**, 1462–1472 (2016).

11. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–75 (2008).

12. Gudbjartsson, D. F. *et al.* A frameshift deletion in the sarcomere gene MYL4 causes early-onset familial atrial fibrillation. *Eur. Heart J.* ehw379 (2016).

13. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case--control association mapping by proxy using family history of disease. *Nat. Genet.* **49**, 325 (2017).

14. Hujoel, M. L. A., Gazal, S., Loh, P.-R., Patterson, N. & Price, A. L. Liability threshold modeling of case--control status and family history of disease increases association power. *Nat. Genet.* **52**, 541–547 (2020).

15. Brumpton, B. *et al.* Within-family studies for Mendelian randomization: avoiding dynastic, assortative mating, and population stratification biases. *bioRxiv* (2019) doi:10.1101/602516.

16. Hwang, L.-D. *et al.* Estimating indirect parental genetic effects on offspring phenotypes using virtual parental genotypes derived from sibling and half sibling pairs. *BioRxiv* (2020).

17. Mills, M. C., Tropf, F. C., Brazel, D. M., Zuydam, N. Van & Vaez, A. Identification of 370 loci for age at onset of sexual and reproductive behaviour, highlighting common aetiology with reproductive biology, externalizing behaviour and longevity. *BioRxiv* (2020).

18. Kong, A. *et al.* Selection against variants in the genome associated with educational attainment. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E727–E732 (2017).