1 Nuclear family model

Direct genetic effects are effects of alleles in an individual on that individual's trait. Indirect genetic effects are the effects of alleles in one individual on another individual's traits. Indirect genetic effects from parents have been shown to be substantial for educational attainment[ref], and there is some evidence that indirect genetic effects from siblings also affect educational attainment[ref].

Consider a sample of n independent families with two siblings in each family. We consider a model for a single SNP with direct effects and indirect effects from parents and a single sibling:

$$Y_{i1} = \delta g_{i1} + \eta_s g_{i2} + \eta_p g_{p(i)} + \eta_m g_{m(i)} + e_{i1}, \tag{1}$$

$$Y_{i2} = \delta g_{i2} + \eta_s g_{i1} + \eta_p g_{p(i)} + \eta_m g_{m(i)} + e_{i2}, \tag{2}$$

where Y_{ij} is the observed phenotype value for sibling j in family i, g_{ij} is the allele count for sibling j in family i, $g_{p(i)}$ is the allele count for the father in family i, $g_{m(i)}$ is the allele count for the mother in family i, δ is the direct effect of the SNP, η_s is the indirect genetic effect from the sibling, η_p is the indirect genetic effect from the father, η_m is the indirect genetic effect from the mother, and e_i is the residual for individual i. For convenience, we assume that phenotypes and genotypes have been mean normalised.

1.1 Multiple siblings

Families have different numbers of siblings, with different distributions of ages and genders, among other potentially important variables. While the indirect genetic effect from a single sibling may depend upon the number of other siblings and other factors, for simplicity, we model the indirect effect from multiple siblings as an additive effect of the average of the siblings' genotypes:

$$Y_{ij} = \delta g_{ij} + \eta_s \bar{g}_{i,j} + \eta_p g_{p(i)} + \eta_m g_{m(i)} + e_i, \tag{3}$$

where $\bar{g}_{i,j}$ is the average genotype of the siblings of individual j in family i.

2 Association analysis with complete family data

The goal of the analysis is to estimate the parameter vector $\theta = [\delta, \eta_s, \eta_p, \eta_m]^T$. The residual e_i can be correlated with the genotypes due to population stratification. We show that this introduces bias to estimates of η_m and η_p but not δ and η_s. The unbiasedness of the estimates of δ and η_s follows from the fact that, given $g_{p(i)}$ and $g_{m(i)}$, g_{i1} and g_{i2} are conditionally independent of e_{i1} and e_{i2}. This follows from the fact that, given parental genotypes, offspring genotypes are determined by random Mendelian segregations in the parents, which are independent of environmental effects. Furthermore, the expectations of g_{i1} and g_{i2} given $g_{p(i)}$ and $g_{m(i)}$ are

a linear function of $g_{p(i)}$ and $g_{m(i)}$; namely, $(g_{p(i)} + g_{m(i)})/2$. Therefore, by the Conditional Independence Lemma A,

$$Y_{i1} = \delta g_{i1} + \eta_s g_{i2} + (\eta_p + b_p) g_{p(i)} + (\eta_m + b_m) g_{m(i)} + \epsilon_{i1}, \tag{4}$$

$$Y_{i2} = \delta g_{i2} + \eta_s g_{i1} + (\eta_p + b_p) g_{p(i)} + (\eta_m + b_m) g_{m(i)} + \epsilon_{i2}, \tag{5}$$

for some constants b_p and b_m and some ϵ_{i1} such that $\mathrm{Cov}(g_i, \epsilon_{i1}) = \mathrm{Cov}(g_{i2}, \epsilon_{i1}) = \mathrm{Cov}(g_{p(i)}, \epsilon_{i1}) = \mathrm{Cov}(g_{m(i)}, \epsilon_{i1}) = 0$.

Therefore, $\mathbf{Y}_i = X_i(\theta + \mathbf{b}) + \epsilon_i$, where

$$X_i = \begin{bmatrix} g_{i1} & g_{i2} & g_{p(i)} & g_{m(i)} \\ g_{i2} & g_{ii} & g_{p(i)} & g_{m(i)} \end{bmatrix} \tag{6}$$

, $\mathbf{Y}_i = [Y_{i1}, Y_{i2}]^T$, $\epsilon_i = [\epsilon_{i1}, \epsilon_{i2}]^T$,and $\mathbf{b} = [0, 0, b_p, b_m]^T$.

We model the correlations between residuals for siblings, giving the phenotypic covariance matrix for family i as

$$\Sigma_i = \mathrm{Cov}(Y_i) = \sigma_\epsilon^2 \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}, \tag{7}$$

where $r = \mathrm{Corr}(\epsilon_{i1}, \epsilon_{i2})$. The generalised least squares estimator for θ is

$$\hat{\theta} = \left(\sum_{i=1}^{n} X_i^T \Sigma_i^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{n} X_i^T \Sigma_i^{-1} Y_i \right), \tag{8}$$

which has expectation

$$\mathbb{E}[\hat{\theta}] = \theta + \mathbf{b}, \tag{9}$$

implying unbiased estimation of δ and η_s.

2.1 Sampling variance

We note that by a similar argument to the above, it can be shown that regressing Y onto only the proband and sum of maternal and paternal genotypes also gives an unbiased estimator of δ. We consider the sampling variance of this estimator to simplify computations and aid comparison with other results. The estimator is:

$$\hat{\theta} = \left(\sum_{i=1}^{n} X_i^T \Sigma_i^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{n} X_i^T \Sigma_i^{-1} Y_i \right), \tag{10}$$

where

$$X_i = \begin{bmatrix} g_{i1} & g_{\mathrm{par(i)}} \\ g_{i2} & g_{\mathrm{par(i)}} \end{bmatrix}; \text{ where } g_{\mathrm{par(i)}} = g_{m(i)} + g_{p(i)}. \tag{11}$$

We now compute $\mathrm{Var}(\hat{\theta})$:

$$\mathrm{Var}(\hat{\theta}) = \left(\sum_{i=1}^{n} X_i^T \Sigma_i^{-1} X_i \right)^{-1}. \tag{12}$$

As the sample size increases,

$$\sum_{i=1}^{n} X_i^T \Sigma_i^{-1} X_i \to \frac{2nf(1-f)}{\sigma_\epsilon^2(1-r^2)} \begin{bmatrix} 2-r & 2(1-r) \\ 2(1-r) & 4(1-r) \end{bmatrix}. \tag{13}$$

Therefore,

$$\left(\sum_{i=1}^{n} X_i^T \Sigma_i^{-1} X_i \right)^{-1} \to \frac{\sigma_\epsilon^2(1+r)}{8(2-r)nf(1-f)} \begin{bmatrix} 4(1-r) & -2(1-r) \\ -2(1-r) & (2-r) \end{bmatrix}. \tag{14}$$

Therefore,

$$\mathrm{Var}(\hat{\delta}) \to \frac{(1-r^2)\sigma_\epsilon^2}{2(2-r)nf(1-f)}. \tag{15}$$

3 One parent missing

3.1 Imputation

We impute the missing parental genotype as the expectation given the observed proband and parent genotypes. Assuming that the father's genotype is missing, this is

$$\hat{g}_{\mathrm{p(i)}} = \mathbb{E}[g_{\mathrm{p(i)}} | g_i, g_{\mathrm{m(i)}}] \tag{16}$$

$$= \frac{2[f(1-f)\mathbb{P}(g_i | g_{\mathrm{m(i)}}, g_{\mathrm{p(i)}} = 1) + f^2 \mathbb{P}(g_i | g_{\mathrm{m(i)}}, g_{\mathrm{p(i)}} = 2)]}{(1-f)^2 \mathbb{P}(g_i | g_{\mathrm{m(i)}}, g_{\mathrm{p(i)}} = 0) + 2f(1-f)\mathbb{P}(g_i | g_{\mathrm{m(i)}}, g_{\mathrm{p(i)}} = 1) + f^2 \mathbb{P}(g_i | g_{\mathrm{m(i)}}, g_{\mathrm{p(i)}} = 2)}, \tag{17}$$

which is derived from application of Bayes' Rule.

By applying the Laws of Mendelian Inheritance to compute the above probabilities, one can derive that:

		$g_{\mathrm{m(i)}}$		
		0	1	2
g_i	0	f	f	-
	1	$1+f$	$2f$	f
	2	-	$1+f$	$1+f$

Table 1: $\mathbb{E}[g_{\mathrm{p(i)}} | g_i, g_{\mathrm{m(i)}}]$

Note that it is impossible for a parent to have two copies of an allele and for the offspring to inherit zero copies, without mutation. (We ignore the possibility of genotyping error here.)

3.1.1 Multiple siblings

3.2 Association analysis

Consider a sample of n independent families with one parent genotyped. We assume the genotyped parent is the mother for all families for notational convenience.

The phenotype of the proband from family i can be expressed as

$$Y_i = \delta g_i + \eta_p^* g_{\mathrm{p(i)}} + \eta_m^* g_{\mathrm{m(i)}} + \epsilon_i, \tag{18}$$

for some mean-zero ϵ_i such that $\mathrm{Cov}(g_i, \epsilon_i) = \mathrm{Cov}(g_{\mathrm{p(i)}}, \epsilon_i) = \mathrm{Cov}(g_{\mathrm{m(i)}}, \epsilon_i) = 0$. (Note that any indirect effects from siblings will contribute one half of their value to η_p^* and η_m^*, so these parameters may differ from η_p and η_m defined above.)

Let $\hat{X}_p = [\mathbf{g}\, \mathbf{g_m}\, \hat{\mathbf{g}}_\mathbf{p}]$. We consider an estimator formed by regression of Y onto $\hat{X}_p$: $\hat{\theta}_p = (\hat{X}_p^T \hat{X})^{-1} \hat{X}_p^T \mathbf{Y}$. The imputed parental genotype is the conditional expectation given the proband and maternal genotype: $\hat{g}_{\mathrm{p(i)}} = \mathbb{E}[g_{\mathrm{p(i)}} | g_i, g_{\mathrm{m(i)}}]$. This means we can apply Theorem 3 to derive that $\lim_{n\to\infty} \hat{\theta}_p = (\delta, \eta_p^*, \eta_m^*)$.

We now derive the sampling variance of $\hat{\theta}_p$: given that η_p^* is small relative to the phenotypic variance,

$$\mathrm{Var}(\hat{\theta}_p) \approx \frac{\mathrm{Var}(\hat{X}_p)^{-1}}{n}. \tag{19}$$

To derive $\mathrm{Var}(\hat{g}_{\mathrm{p(i)}})$, we first derive the joint probabilities of the observed genotypes using Bayes' Rule and the Laws of Mendelian Inheritance:

		$g_{\mathrm{m(i)}}$		
		0	1	2
g_i	0	$(1-f)^3$	$f(1-f)^2$	0
	1	$f(1-f)^2$	$f(1-f)$	$f^2(1-f)$
	2	0	$f^2(1-f)$	f^3

Table 2: $\mathbb{P}(g_i, g_{\mathrm{m(i)}})$

From this, we can compute that $\mathrm{Var}(\hat{g}_{\mathrm{p(i)}}) = f(1-f)[1-f(1-f)]$.

By application of Lemma 2, we have that $\mathrm{Cov}(g_i, \hat{g}_{\mathrm{p(i)}}) = \mathrm{Cov}(g_i, g_{\mathrm{p(i)}}) = f(1-f)$, and that $\mathrm{Cov}(g_{\mathrm{m(i)}}, \hat{g}_{\mathrm{p(i)}}) = \mathrm{Cov}(g_{\mathrm{m(i)}}, g_{\mathrm{p(i)}}) = 0$. Therefore,

$$\mathrm{Var}(\hat{X}_p) = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 1-f(1-f) \end{bmatrix}; \tag{20}$$

and therefore

$$\mathrm{Var}(\hat{\theta}_p) \approx \frac{1}{f(1-f)[1-3f(1-f)]} \begin{bmatrix} 2-2f(1-f) & 1-f(1-f) & -2 \\ 1-f(1-f) & 1-2f(1-f) & 1 \\ -2 & 1 & 3 \end{bmatrix}. \tag{21}$$

Let $\hat{\delta}_p$ be the resulting estimator of δ, then

$$\mathrm{Var}(\hat{\delta}_p) \approx \frac{2 - 2f(1-f)}{[1 - 3f(1-f)]nf(1-f)} \tag{22}$$

This can be compared to the variance of the estimator of delta with both parental genotypes observed, $\hat{\delta}_{po}$: $\mathrm{Var}(\hat{\delta}_{po}) = (nf(1-f))^{-1}$; and

$$\frac{\mathrm{Var}(\hat{\delta}_{po})}{\mathrm{Var}(\hat{\delta}_p)} = \frac{1 - 3f(1-f)}{2 - 2f(1-f)} \tag{23}$$

Figure 1: Relative effective sample size for estimating direct genetic effects using imputed parental genotypes compared to observed parental genotypes as a function of allele frequency.

The penalty relative to using fully observed parental genotypes increases with the heterozygosity due to the fact that when both observed parent and child genotypes are heterozygous, the allele inherited by the child from the observed parent cannot be determined, so an average over the two possible inheritance patterns is taken as the imputation. This could be overcome by determining the segments of DNA that were inherited by the child from the parent, but that is not always computationally feasible to do for all genotyped/imputed SNPs.

4 Both parents missing

4.1 Using difference in sibling genotypes

Assuming that $\eta_s = 0$, sibling pairs can also be used to estimate δ and η, the average parental effect. The model for the two siblings' phenotypes is:

$$Y_{i1} = \delta g_{i1} + \eta g_{\mathrm{par(i)}} + \epsilon_{i1}; \tag{24}$$
$$Y_{i2} = \delta g_{i2} + \eta g_{\mathrm{par(i)}} + \epsilon_{i2}. \tag{25}$$

This can be transformed into two orthogonal variables:

$$Y_{i1} - Y_{i2} = \delta(g_{i1} - g_{i2}) + \epsilon_{i1} - \epsilon_{i2}; \tag{26}$$
$$Y_{i1} + Y_{i2} = \delta(g_{i1} + g_{i2}) + 2\eta g_{\mathrm{par(i)}} + \epsilon_{i1} + \epsilon_{i2}. \tag{27}$$

The first variable, $Y_{i1} - Y_{i2}$, gives information on δ. The second variable gives information on a linear combination of δ and η that, when combined with the information on δ from the difference in phenotypes, can give an estimate of *eta*.

By performing regression of differences between siblings' phenotypes onto differences in genotypes, one can estimate δ [ref]. Let $\hat{\delta}_\Delta$ be the resulting estimator. It can be shown that $\mathbb{E}[\hat{\delta}_\Delta] = \delta - \eta_s$ and

$$\mathrm{Var}(\hat{\delta}_\Delta) = \frac{(1-r)\sigma^2_\epsilon}{nf(1-f)}. \tag{28}$$

By performing the regression $(Y_{i1}+Y_{i2}) \sim (g_{i1}+g_{i2})$ one obtains an estimate of $\delta+(4/3)\eta$. Let this estimate be $\hat{\beta}$. It is trivial to show that $\mathrm{Var}(\hat{\beta}) = (1+r)\sigma^2_\epsilon/(3nf(1-f))$. We can obtain an estimate of η as $\hat{\eta}_\Delta = (3/4)(\hat{\beta} - \hat{\delta}_\Delta)$. From this, we have that $\mathrm{Var}(\hat{\eta}) = 3\sigma^2_\epsilon(2-r)/(8nf(1-f))$. We also have that $\mathrm{Cov}(\hat{\eta}_\Delta, \hat{\delta}_\Delta) = -3\mathrm{Var}(\hat{\delta}_\Delta)/4 = -3(1-r)\sigma^2_\epsilon/(4nf(1-f))$.

4.2 Imputation

When both parents are missing, we infer parental genotypes from sibling genotypes and the identity-by-descent (IBD) sharing state between the siblings. Consider a sibling pair. When neither alleles are shared IBD (IBD state 0), all four parental alleles have been observed; when one allele is shared IBD (IBD state 1), three parental alleles have been observed; when both alleles are shared IBD (IBD state 2), two parental alleles have been observed.

More formally, let $g_{par(i)} = g_{m(i)} + g_{p(i)}$ be the sum of the parental genotypes for individual i. This can also be written in terms of the parental alleles: $g_{par(i)} = g^p_{m(i)} + g^m_{m(i)} + g^p_{p(i)} + g^m_{p(i)}$, where $g^p_{m(i)}$ is the paternally inherited allele of the mother of i, and $g^m_{p(i)}$ is the maternally inherited allele of the father of i. Since parent-of-origin of alleles cannot be determined from sibling data alone, we can only estimate $g_{par(i)}$ rather than $g_{m(i)}$ and $g_{p(i)}$ separately. For now, we assume that we know which alleles are shared IBD in addition to the overall IBD state (0,1, or 2), which, for IBD state 1, requires phased data. We construct the estimate of $g_{par(i)}$, $\hat{g}_{par(i)}$, to be the expectation of $g_{par(i)}$ given the observed sibling genotypes and the IBD state of the siblings: $\hat{g}_{par(i)} = \mathbb{E}[g_{par(i)}|g_{i1}, g_{i2}, \mathrm{IBD}]$. This gives:

$$\hat{g}_{par(i)} = \begin{cases} g_{i1} + g_{i2} = g_{par(i)}, & \text{if IBD} = 0 \\ g_{i1} + g^k_{i2} + f, & \text{if IBD} = 1 \\ g_{i1} + 2f, & \text{if IBD} = 2, \end{cases} \tag{29}$$

where $k \in \{m, p\}$ is such that g^k_{i2} is not IBD with the alleles inherited by sibling 1 in family i.

If we do not have access to phased IBD data, then if both siblings are heterozygous and the IBD state is 1, then the shared allele cannot be determined. In this case, the shared allele is the allele with frequency f with probability $1-f$, and the shared allele is the allele with frequency $1-f$ with probability f. This can be derived from considering the relative frequencies of the three observed parental genotypes: when the allele with frequency f is shared, the probability of observing those three parental alleles is $f(1-f)^2$; and when the allele with frequency $(1-f)$

is observed, the probability of observing those three parental alleles is $f^2(1-f)$. Conditional on both siblings being heterozygous and being in IBD state 1, the probability of the allele with frequency f is shared is $f(1-f)^2/[f(1-f)^2 + f^2(1-f)] = f(1-f)^2/f(1-f) = 1-f$; this implies that the probability that the allele with frequency $1-f$ is shared is f. The imputed parental genotype is therefore the average over these two possibilities:

$$\mathbb{E}[g_{par(i)}|g_{i1}=1, g_{i2}=1, \text{IBD}=1] = f(2+f) + (1-f)(1+f) = 1+2f. \tag{30}$$

Let H_i be the event that both siblings are heterozygous, then the imputed parental genotype without phased IBD data is:

$$\hat{g}_{par(i)} = \begin{cases} g_{i1} + g_{i2} = g_{par(i)}, & \text{if IBD} = 0 \\ g_{i1} + g^k_{i2} + f, & \text{if IBD} = 1 \text{ and } \neg H_i \\ 1 + 2f, & \text{if IBD} = 1 \text{ and } H_i \\ g_{i1} + 2f, & \text{if IBD} = 2, \end{cases} \tag{31}$$

4.2.1 Multiple siblings

First we prove that it is impossible that, for more than three siblings, it is impossible for all pairs of siblings to be in an IBD 1 state.

Consider three siblings in family i. We write the genotype of a sibling in terms of parental alleles as $(g^j_{m(i)}, g^k_{p(i)})$ for $j,k \in \{m,p\}$. Without loss of generality, consider that the genotype of sibling 1 is $(g^p_{m(i)}, g^p_{p(i)})$ and that the genotype of sibling 2 is $(g^p_{m(i)}, g^m_{p(i)})$, so that sibling 1 and 2 are in an IBD 1 state. We now consider if it is possible for a third sibling to be in an IBD 1 state with both sibling 1 and sibling 2.

If sibling 3 inherits the same parental alleles as either sibling 1 or sibling 2, then sibling 3 is in an IBD 2 state with another sibling. There are two more possible inheritance patterns for siblings 3: $(g^m_{m(i)}, g^p_{p(i)})$, which implies sibling 3 is IBD 0 with sibling 2; and $(g^m_{m(i)}, g^m_{p(i)})$, which implies sibling 3 is IBD 0 with sibling 2. Therefore, it is impossible for sibling 3 to be IBD 1 with both sibling 1 and sibling 2.

This implies that, to impute parental genotypes with more than two siblings, the problem can be reduced to imputing exactly the parental alleles when at least one sibling pair is in an IBD 0 state; or imputing as if one has observed a single sibling pair in an IBD state 2 with each other if all siblings are in an IBD 2 state with each; or reduced to the problem of imputing from a sibling pair by reducing sets of siblings that are all IBD 2 with each other to a single sibling. (An additional sibling that is in an IBD 2 state with an existing sibling adds no further observed parental alleles.)

4.3 Association analysis

We reformulate the phenotype model (3) as

$$Y_{i1} = \delta g_{i1} + \eta_s g_{i2} + \eta g_{par(i)} + \epsilon_{i1}, \tag{32}$$

$$Y_{i2} = \delta g_{i2} + \eta_s g_{i1} + \eta g_{par(i)} + \epsilon_{i2}, \tag{33}$$

where $\mathrm{Cov}(g_{i1}, \epsilon_{i1}) = \mathrm{Cov}(g_{i2}, \epsilon_{i1}) = \mathrm{Cov}(g_{p(i)}, \epsilon_{i1}) = \mathrm{Cov}(g_{m(i)}, \epsilon_{i1}) = 0$ due to the Conditional Independent Lemma (A) [ref]. Here $\eta = (\eta_p + b_p + \eta_m + b_m)/2$, the average of the maternal and paternal indirect genetic effects and biases.

Let $\hat{\theta}_p$ be the estimator that results from regression of the phenotype onto the proband, sibling, and imputed parental genotypes. We show that $\lim_{n\to\infty} \hat{\theta}_p = [\delta, \eta_s, \eta]$:

Proof.

$$\lim_{n\to\infty} \hat{\theta}_p = \begin{bmatrix} \mathrm{Var}(g_{i1}) & \mathrm{Cov}(g_{i1}, g_{i2}) & \mathrm{Cov}(g_{i1}, \hat{g}_{\mathrm{par(i)}}) \\ \mathrm{Cov}(g_{i1}, g_{i2}) & \mathrm{Var}(g_{i2}) & \mathrm{Cov}(g_{i2}, \hat{g}_{\mathrm{par(i)}}) \\ \mathrm{Cov}(g_{i1}, \hat{g}_{\mathrm{par(i)}}) & \mathrm{Cov}(g_{i2}, \hat{g}_{\mathrm{par(i)}}) & \mathrm{Var}(\hat{g}_{\mathrm{par(i)}}) \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{Cov}(g_{i1}, Y_{i1}) \\ \mathrm{Cov}(g_{i2}, Y_{i1}) \\ \mathrm{Cov}(\hat{g}_{\mathrm{par(i)}}, Y_{i1}). \end{bmatrix} \tag{34}$$

We now compute $\mathrm{Var}(\hat{g}_{par(i)})$ and $\mathrm{Cov}(\hat{g}_{par(i)}, g_{i1})$, which is the same as $\mathrm{Cov}(\hat{g}_{par(i)}, g_{i2})$, by symmetry. The variance of the imputed parental genotype can be computed by the Law of Total Variance:

$$\mathrm{Var}(\hat{g}_{par(i)}) = \mathbb{E}_{\mathrm{IBD}}[\mathrm{Var}(\hat{g}_{par(i)}|\mathrm{IBD})] + \mathrm{Var}_{\mathrm{IBD}}(\mathbb{E}[\hat{g}_{par(i)}|\mathrm{IBD}]) = \mathbb{E}_{\mathrm{IBD}}[\mathrm{Var}(\hat{g}_{par(i)}|\mathrm{IBD})], \tag{35}$$

since the expectation of the imputed parental genotypes does not depend upon the IBD state of the siblings. The variance of the imputed parental genotype is directly proportional to the number of observed parental alleles:

$$\mathrm{Var}(\hat{g}_{par(i)}|\mathrm{IBD}) = \begin{cases} 4f(1-f), & \text{if IBD} = 0 \\ 3f(1-f), & \text{if IBD} = 1 \\ 2f(1-f), & \text{if IBD} = 2 \end{cases}. \tag{36}$$

Therefore, since $\mathbb{P}(\mathrm{IBD} = 0) = 0.25$, $\mathbb{P}(\mathrm{IBD} = 1) = 0.5$, and $\mathbb{P}(\mathrm{IBD} = 2) = 0.25$, $\mathrm{Var}(\hat{g}_{par(i)}) = 3f(1-f) = (3/4)\mathrm{Var}(g_{par(i)})$. The imputed parental genotype thus captures three quarters of the variance of the observed parental genotype.

Similarly, by applying the Law of Total Covariance, it can be shown that $\mathrm{Cov}(\hat{g}_{par(i)}, g_i) = 2f(1-f)$. Therefore,

$$\begin{bmatrix} \mathrm{Var}(g_{i1}) & \mathrm{Cov}(g_{i1}, g_{i2}) & \mathrm{Cov}(g_{i1}, \hat{g}_{\mathrm{par(i)}}) \\ \mathrm{Cov}(g_{i1}, g_{i2}) & \mathrm{Var}(g_{i2}) & \mathrm{Cov}(g_{i2}, \hat{g}_{\mathrm{par(i)}}) \\ \mathrm{Cov}(g_{i1}, \hat{g}_{\mathrm{par(i)}}) & \mathrm{Cov}(g_{i2}, \hat{g}_{\mathrm{par(i)}}) & \mathrm{Var}(\hat{g}_{\mathrm{par(i)}}) \end{bmatrix}^{-1} = \frac{1}{f(1-f)} \begin{bmatrix} 2 & 1 & -2 \\ 1 & 2 & -2 \\ -2 & -2 & 3 \end{bmatrix}. \tag{37}$$

Furthermore, $\mathrm{Cov}(\hat{g}_{par(i)}, g_{par(i)}) = 3f(1-f)$, and therefore

$$\begin{bmatrix} \mathrm{Cov}(g_{i1}, Y_{i1}) \\ \mathrm{Cov}(g_{i2}, Y_{i1}) \\ \mathrm{Cov}(\hat{g}_{\mathrm{par(i)}}, Y_{i1}). \end{bmatrix} = f(1-f) \begin{bmatrix} 2\delta + \eta_s + 2\eta \\ \delta + 2\eta_s + 2\eta \\ 2\delta + 2\eta_s + 3\eta \end{bmatrix}, \tag{38}$$

and therefore $\lim_{n\to\infty} \hat{\theta}_p = \theta_p$ can be verified. □

4.3.1 Dropping indirect effects from siblings

If the goal is to estimate direct genetic effects with precision, then dropping the sibling indirect effect may seem an attractive option. Let $\hat{\theta}_s$ be the estimator that results from regression of phenotype onto proband and imputed parental genotype. From the above results it is readily derived that

$$\lim_{n\to\infty} \hat{\theta}_s = \begin{bmatrix} \delta - \frac{\eta_s}{2} \\ \eta + \frac{\eta_s}{2} \end{bmatrix}. \tag{39}$$

We compute the sampling variance of $\hat{\theta}_s$ from n independent families with two siblings in each family, where

$$\hat{X}_i = \begin{bmatrix} g_{i1} & \hat{g}_{\text{par(i)}} \\ g_{i2} & \hat{g}_{\text{par(i)}} \end{bmatrix} \tag{40}$$

is the design matrix for family i, with g_{i1} the genotype of sibling 1 in family i, g_{i2} the genotype of sibling 2 in family i, and $\hat{g}_{\text{par(i)}}$ the imputed parental genotype for family i.

The generalised least-squares estimator is:

$$\hat{\theta}_s = \left(\sum_{i=1}^{n} \hat{X}_i^T \Sigma_i^{-1} \hat{X}_i \right)^{-1} \left(\sum_{i=1}^{n} \hat{X}_i^T \Sigma_i^{-1} Y_i \right). \tag{41}$$

Assuming that η^2 is negligible compared to the phenotypic variance,

$$\text{Var}(\hat{\theta}_s) \approx \left(\sum_{i=1}^{n} \hat{X}_i^T \Sigma_i^{-1} \hat{X}_i \right)^{-1}. \tag{42}$$

We have that:

$$X_i^T \Sigma_i^{-1} X_i = \frac{1}{\sigma_\epsilon^2 (1-r^2)} \begin{bmatrix} g_{i1} & g_{i2} \\ \hat{g}_{\text{par(i)}} & \hat{g}_{\text{par(i)}} \end{bmatrix} \begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix} \begin{bmatrix} g_{i1} & \hat{g}_{\text{par(i)}} \\ g_{i2} & \hat{g}_{\text{par(i)}} \end{bmatrix} \tag{43}$$

$$= \frac{1}{\sigma_\epsilon^2 (1-r^2)} \begin{bmatrix} g_{i1}^2 - 2r g_{i1} g_{i2} + g_{i2}^2 & \hat{g}_{\text{par(i)}} (g_{i1} - r(g_{i1} + g_{i2}) + g_{i2}) \\ \hat{g}_{\text{par(i)}} (g_{i1} - r(g_{i1} + g_{i2}) + g_{i2}) & 2(1-r)\hat{g}_{\text{par(i)}}^2 \end{bmatrix} \tag{44}$$

As the sample size increases,

$$\sum_{i=1}^{n} \hat{X}_i^T \Sigma_i^{-1} \hat{X}_i \to \frac{2nf(1-f)}{\sigma_\epsilon^2 (1-r^2)} \begin{bmatrix} 2-r & 2(1-r) \\ 2(1-r) & 3(1-r) \end{bmatrix}. \tag{45}$$

Therefore,

$$\left(\sum_{i=1}^{n} \hat{X}_i^T \Sigma_i^{-1} \hat{X}_i \right)^{-1} \to \frac{\sigma_\epsilon^2 (1+r)}{2(2+r)nf(1-f)} \begin{bmatrix} 3(1-r) & -2(1-r) \\ -2(1-r) & (2-r) \end{bmatrix}. \tag{46}$$

Therefore, for large samples, the estimator for δ, $\hat{\delta}_s$ has variance:

$$\mathrm{Var}(\hat{\delta}_s) \approx \frac{3(1-r^2)\sigma^2_\epsilon}{2n(2+r)f(1-f)}. \tag{47}$$

This can be compared to the variance of the estimator of δ using differences between siblings' genotypes only:

$$\mathrm{Var}(\hat{\delta}_\Delta) = \frac{(1-r)\sigma^2_\epsilon}{nf(1-f)}. \tag{48}$$

We therefore have that

$$\frac{\mathrm{Var}(\hat{\delta}_\Delta)}{\mathrm{Var}(\hat{\delta}_s)} \approx \frac{2(2+r)}{3(1+r)} \geq 1 \text{ for } r \in [-1, 1]. \tag{49}$$

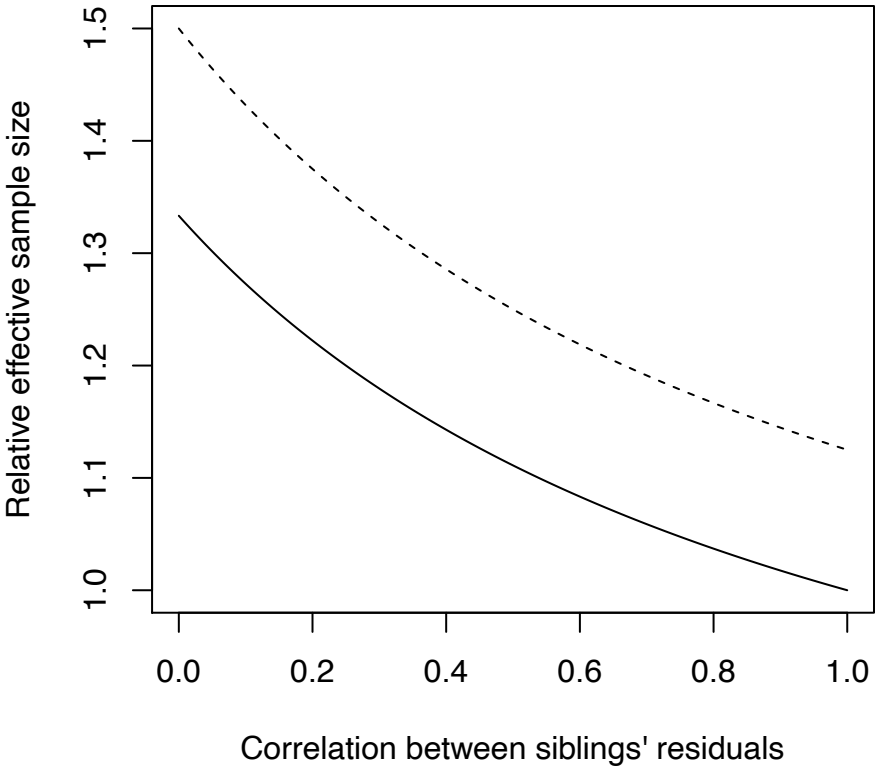

Figure 2: Relative effective sample size for estimating direct genetic effects using imputed parental genotypes compared to regressing phenotype differences between sibs onto genotype differences between sibs, as a function of the correlation between siblings' residuals.

Furthermore, let $\hat{\eta}$ be the estimator of η. From the above,

$$\mathrm{Var}(\hat{\eta}) = \frac{(1+r)(2-r)\sigma_\epsilon^2}{2(2+r)nf(1-f)}. \tag{50}$$

This can be compared to the variance of the estimator of η from using sibling genotypes alone without imputation, $\hat{\eta}_\Delta$. From the above,

$$\frac{\mathrm{Var}(\hat{\eta}_\Delta)}{\mathrm{Var}(\hat{\eta})} = \frac{3(2+r)}{4(1+r)}. \tag{51}$$

Let $\hat{\delta}$ be the estimator of δ using observed parental genotypes and proband genotypes. Then we have

$$\frac{\mathrm{Var}(\hat{\delta})}{\mathrm{Var}(\hat{\delta}_s)} \approx \frac{2+r}{3(2-r)} \leq 1 \text{ for } r \in [-1, 1]. \tag{52}$$

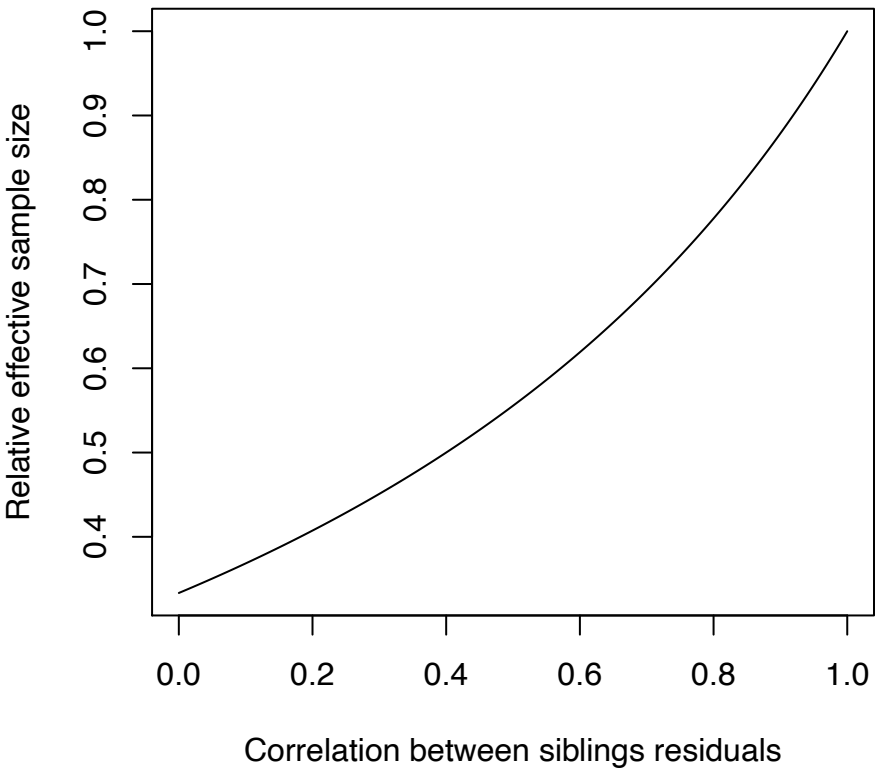

Figure 3: Relative effective sample size for estimating direct genetic effects using imputed parental genotypes compared to observed parental genotypes as a function of the correlation between the siblings' residuals.

4.3.2 Without phased IBD data

The variation in the parental genotype captured by the imputation is decreased without phased IBD data due to the inability to determine the shared allele when both siblings are heterozygous and in IBD state 1.

To compute the sampling distribution of the estimators that result from using the imputed parental genotype without phased IBD data, we first need to compute the variance of the imputed parental genotype.

The imputed parental genotype as a function of the observed sibling genotypes given IBD state 1 is:

		g_{i2}		
		0	1	2
	0	f	$1+f$	-
g_{i1}	1	$1+f$	$1+2f$	2+f
	2	-	$2+f$	$3+f$

Table 3: $\mathbb{E}[g_{\text{par(i)}}|g_{i1}, g_{i2}, \text{IBD} = 1]$

By considering the probability of observing the three observed parental alleles, one can derive the distribution of the observed sibling genotypes given that the IBD state is 1:

		g_{i2}		
		0	1	2
	0	$(1-f)^3$	$f(1-f)^2$	0
g_{i1}	1	$f(1-f)^2$	$f(1-f)$	$f^2(1-f)$
	2	0	$f^2(1-f)$	f^3.

Table 4: $\mathbb{P}(g_{i1}, g_{i2}|\text{IBD} = 1)$

From these two tables, one can compute that

$$\text{Var}(\hat{g}_{\text{par(i)}}|\text{IBD} = 1) = [3 - f(1-f)]f(1-f); \tag{53}$$

and therefore

$$\text{Var}(\hat{g}_{\text{par(i)}}) = [3 - f(1-f)/2]f(1-f). \tag{54}$$

Assuming that η^2 is negligible compared to the phenotypic variance,

$$\text{Var}(\hat{\theta}_s) \approx \left(\sum_{i=1}^{n} \hat{X}_i^T \Sigma_i^{-1} \hat{X}_i \right)^{-1}. \tag{55}$$

Following the same steps as above for the imputed parental genotypes with phased IBD data gives:

$$\lim_{n\to\infty} \sum_{i=1}^{n} \hat{X}_i^T \Sigma_i^{-1} \hat{X}_i = \frac{2nf(1-f)}{\sigma_\epsilon^2(1-r^2)} \begin{bmatrix} 2-r & 2(1-r) \\ 2(1-r) & (1-r)[3-f(1-f)/2]; \end{bmatrix} \tag{56}$$

and therefore

$$\lim_{n\to\infty} \sum_{i=1}^{n} (\hat{X}_i^T \Sigma_i^{-1} \hat{X}_i)^{-1} = \frac{\sigma_\epsilon^2(1+r)}{2nf(1-f)[2+r-(1-r/2)f(1-f)]} \begin{bmatrix} (1-r)[3-f(1-f)/2] & -2(1-r) \\ -2(1-r) & 2-r \end{bmatrix}. \tag{57}$$

This gives the variance of the estimate of the direct effect in large samples as:

$$\text{Var}(\hat{\delta}_s) \approx \frac{\sigma_\epsilon^2(1-r^2)[3-f(1-f)/2]}{[2+r-(1-r/2)f(1-f)]2nf(1-f)}. \tag{58}$$

Comparing this to the variance of the sib-difference estimator:

$$\frac{\mathrm{Var}(\hat{\delta}_{\Delta})}{\mathrm{Var}(\hat{\delta}_{s})} \approx \frac{2[2+r-(1-r/2)f(1-f)]}{[3-f(1-f)/2](1+r)}. \tag{59}$$

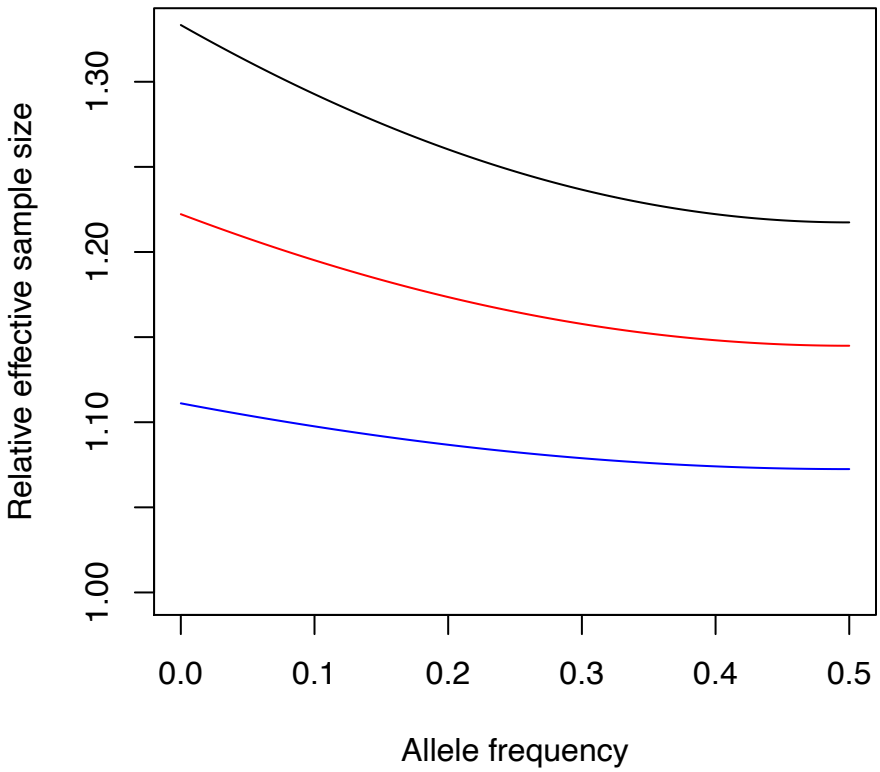

Figure 4: Relative effective sample size for estimating direct genetic effects using parental genotypes imputed from sibling genotypes with un-phased IBD data compared to using differences in sibling genotypes. We show the relative effective sample size as a function of the allele frequency for a correlation between sibling residulas of zero (black curve) and 0.2 (red curve).

5 Mixed model inference

The goal is to perform robust genome-wide association analysis using siblings. There are potentially strong correlations between siblings' phenotypes even after accounting for the genetic

component of a trait. To perform efficient estimation of direct effects of SNPs and obtain correct standard errors, it is therefore necessary to account for the correlation between phenotype observations. To do this, we model the expected phenotype within a family of siblings as a random effect.

The data are comprised of observations on N_F families. For family i, there are n_i observations, giving a total of $\sum_{i=1}^{N_F} n_i = N$ observations. We assume that the data has been ordered so that the observations from family 1 are indexed from 1 to n_1, the observations from individual 2 are indexed from $n_1 + 1$ to $n_1 + n_2$, etc.

The phenotype is an $[N \times 1]$ vector Y and the covariate matrix is an $[N \times c]$ matrix X. The X matrix can be constructed using the genotypes of the siblings and/or parents in different ways depending on the application. We introduce a $[N \times n]$ matrix Z such that $[Z]_{ij} = 1$ if observation i is from family j, and $[Z]_{ij}$ is zero otherwise. We assume that the within-family means are independently normally distributed, represented by an $[n \times 1]$ vector $u \sim \mathcal{N}(0, \sigma_F^2 \mathrm{I}_n)$. The model is

$$Y = X\alpha + Zu + \epsilon. \tag{60}$$

We assume that the residuals are I.I.D. Gaussians, $\epsilon \sim N(0, \sigma^2 I)$. The distribution of $Y|X$ is therefore,

$$Y|X \sim \mathcal{N}(X\alpha, \sigma_F^2 ZZ^T + \sigma_\epsilon^2 I). \tag{61}$$

It can readily be inferred that ZZ^T has a simple block-diagonal structure. For $N_F = 2$, the matrix has the following structure:

$$ZZ^T = \begin{bmatrix} 1_{n_1} 1_{n_1}^T & 0 \\ 0 & 1_{n_2} 1_{n_2}^T \end{bmatrix}, \tag{62}$$

where 1_k is the $[k \times 1]$ column vector of all 1s. The simple structure of ZZ^T allows for an efficient algorithm to compute the likelihood and gradients.

5.1 Loss function and gradients

Instead of the optimising the likelihood, we seek to minimise negative two times the log-likelihood as a loss function:

$$L = \log|\Sigma| + (y - X\alpha)^T \Sigma^{-1} (y - X\alpha), \tag{63}$$

where $\Sigma = \sigma_F^2 ZZ^T + \sigma_\epsilon^2 I$.

Naive computation of the loss function takes $\mathrm{O}(N^3)$ operations. However, the likelihood component of the loss function can be split into a sum over families. Let Σ_i be the diagonal block of Σ corresponding to observations on family i. Furthermore, let y_i be the $[n_i \times 1]$ vector of observations for individual i, and let X_i be the $[n_i \times c]$ matrix of covariate observations. Then,

$$L = \sum_{i=1}^{N_F} \log|\Sigma_i| + \sum_{i=1}^{N_F} (y_i - X_i\alpha)^T \Sigma_i^{-1} (y_i - X_i\alpha), \tag{64}$$

While this expression can be computed with fewer operations, naively it requires $\mathrm{O}(\sum_{i=1}^{n} n_i^3)$ operations to compute. If some of the n_i are large, this could become expensive.

We introduce $\tau = \sigma_\epsilon^2/\sigma_F^2$, and parameterise the model in terms of τ and σ_ϵ^2. Because the blocks of Σ are comprised of a diagonal plus a rank-one matrix, the determinant and inverse of each block can be computed analytically using the Sherman-Morrison-Woodbury identity and the Matrix Determinant Lemma. This gives

$$\Sigma_i^{-1} = \frac{1}{\sigma_\epsilon^2}\left(\mathrm{I}_{n_i} - \frac{1_{n_i} 1_{n_i}^T}{\tau + n_i}\right);\ \log|\Sigma_i| = n_i \log(\sigma_\epsilon^2) + \log\left(1 + \frac{n_i}{\tau}\right). \tag{65}$$

The loss function can thus be expressed as

$$L = N\log(\sigma_\epsilon^2) + \sum_{i=1}^{N_F} \log\left(1 + \frac{n_i}{\tau}\right) + \frac{(y - X\alpha)^T(y - X\alpha)}{\sigma_\epsilon^2} - \frac{1}{\sigma_\epsilon^2}\sum_{i=1}^{n} \frac{[1_{n_i}^T(y_i - X_i\alpha)]^2}{\tau + n_i}. \tag{66}$$

The loss function can be computed in $\mathrm{O}(N)$ operations, and does not require storage of the covariance matrix Σ or the design matrix Z. Gradients can be computed in the same time complexity class. The gradient with respect to α is,

$$\frac{\partial L}{\partial \alpha} = -\frac{2}{\sigma_\epsilon^2}(y - X\alpha)^T X + \frac{2}{\sigma_\epsilon^2}\sum_{i=1}^{N_F} \frac{[1_{n_i}^T(y_i - X_i\alpha)]1_{n_i}^T X_i}{\tau + n_i} \tag{67}$$

The gradients with respect to the variance parameters can be computed similarly easily.

It is straightforward to show that the MLE for α, $\hat{\alpha}$, given τ, must satisfy the linear system:

$$\left(X^T X - \sum_{i=1}^{N_F} \frac{n_i^2 \bar{X}_i^T \bar{X}_i}{\tau + n_i}\right)\hat{\alpha} = X^T y - \sum_{i=1}^{N_F} \frac{n_i^2 \bar{X}_i^T \bar{y}_i}{\tau + n_i}, \tag{68}$$

where $\bar{X}_i$ is the sample mean of the covariate vector within family i, and $\bar{y}_i$ is the sample mean phenotype within family i.

It is also straightforward to show that the asymptotic sampling variance of the MLE for α is:

$$\mathrm{Var}(\hat{\alpha}) = \sigma_\epsilon^2 \left(X^T X - \sum_{i=1}^{N_F} \frac{n_i^2 \bar{X}_i^T \bar{X}_i}{\tau + n_i}\right)^{-1}. \tag{69}$$

6 Optimisation

The parameters we are optimising over are $\theta = (\alpha, \sigma_\epsilon^2, \tau)$. However, since the MLE for α can be computed efficiently analytically given an estimate of τ, we instead optimise

$$L_{\mathrm{prof}}(\sigma_\epsilon^2, \tau) = L(\hat{\alpha}(\tau), \sigma_\epsilon^2, \tau), \tag{70}$$

where the optimisation takes place over $(\sigma^2_\epsilon, \tau)$ only, with the MLE for α for a given τ, $\hat{\alpha}(\tau)$, computed analytically.

We optimise the model with the L-BFGS-B algorithm, with $(\sigma^2_\epsilon, \tau)$ bounded below at $(10^{-5}, 10^{-5})$. We provide the functions for efficient computation of the likelihood, the MLE of α given τ, and gradients to the L-BFGS-B algorithm. Correct calculation of the loss function was checked by comparing to calculation from the expression in equation 64. Correct calculation of the gradient was checked by numerical differentiation of the loss function. (The checks are performed using unit testing in the *Python* package).

For application to a set of SNPs from a chromosome, a null model including no SNPs is first fit. By default, we initialise $(\sigma^2_\epsilon, \tau)$ to $(s^2_Y/2, 1)$, where s^2_Y is the sample estimate of the phenotypic variance The MLEs of τ and σ^2_ϵ from the null model are then fixed for all SNP specific models, allowing analytical computation of the (approximate) MLE for α for each SNP.

7 Inferring relations between effects

Consider estimating, for each SNP i, a vector of d effects, θ_i. This vector could include direct genetic effects, indirect genetic effects, and standard GWAS effects. Let $\hat{\theta}_i$ be the vector of effect estimates. When these effects are estimated from the same sample, we know the joint sampling distribution of these effects. Assuming that $\hat{\theta}_i$ is a consistent estimator of θ_i and that the sample is sufficiently large

$$\hat{\theta}_i | \theta_i \sim \mathcal{N}(\theta_i, S_i), \tag{71}$$

where S_i is the known variance-covariance matrix.

We assume that the true effect vector at each SNP, θ_i is drawn from a multivariate normal distribution:

$$\theta_i \sim \mathcal{N}(0, V). \tag{72}$$

From V, the phenotypic variance explained by the different effects and the correlations between the different effects can be derived.

To make inferences about V from the observed effect estimates and their sampling distributions, we derive the marginal dist ribution of $\hat{\theta}_i$ by integrating out the true effects:

$$f(\hat{\theta}_i) = \int_{\theta_i \in \mathcal{R}^d} f(\hat{\theta}_i | \theta_i) f(\theta_i) d\theta_i; \tag{73}$$

$$= \int_{\theta_i \in \mathcal{R}^d} (2\pi)^{-d} |S_i|^{-\frac{1}{2}} |V|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \left[(\hat{\theta}_i - \theta_i)^T S_i^{-1} (\hat{\theta}_i - \theta_i) + \theta_i^T V^{-1} \theta_i \right] \right) d\theta_i. \tag{74}$$

By using a multivariate analogue of the method of 'completing the square', the sum of quadratic forms in the exponential can be written as a single quadratic form in terms of θ_i and a residual that is independent of θ_i. Let $M_i = S_i^{-1} + V^{-1}$ and $b = S_i^{-1}\hat{\theta}_i$, then

$$(\hat{\theta}_i - \theta_i)^T S_i^{-1} (\hat{\theta}_i - \theta_i) + \theta_i^T V^{-1} \theta_i = (\theta_i - M_i^{-1} b)^T M_i (\theta_i - M_i^{-1} b)^T + \hat{\theta}_i^T S_i^{-1} [S_i - M_i^{-1}] S_i^{-1} \hat{\theta}_i, \tag{75}$$

which can be readily verified by expanding out the quadratics on both sides of the equation.

We therefore have that

$$f(\hat{\theta}_i) = (2\pi)^{-d}|S_i|^{-\frac{1}{2}}|V|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\hat{\theta}_i^T S_i^{-1}[S_i - M_i^{-1}]S_i^{-1}\hat{\theta}_i\right) \tag{76}$$

$$\int_{\theta_i \in \mathcal{R}^d} \exp\left(-\frac{1}{2}(\theta_i - M_i^{-1}b)^T M_i(\theta_i - M_i^{-1}b)^T\right) d\theta_i. \tag{77}$$

Since the integral is of the exponential component of a multivariate normal distribution with covariance matrix M_i^{-1}, we have that

$$\int_{\theta_i \in \mathcal{R}^d} \exp\left(-\frac{1}{2}(\theta_i - M_i^{-1}b)^T M_i(\theta_i - M_i^{-1}b)^T\right) d\theta_i = (2\pi)^{\frac{d}{2}}|M_i|^{-\frac{1}{2}}. \tag{78}$$

By application of the Matrix Determinant Lemma, we have that $|M_i| = |S_i + V||S_i|^{-1}|V|^{-1}$. Therefore,

$$f(\hat{\theta}_i) = (2\pi)^{-\frac{d}{2}}|S_i + V|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\hat{\theta}_i^T S_i^{-1}[S_i - M_i^{-1}]S_i^{-1}\hat{\theta}_i\right). \tag{79}$$

This implies that

$$\hat{\theta}_i \sim \mathcal{N}(0, S_i + V). \tag{80}$$

This is true since

$$\left(S_i^{-1}[S_i - M_i^{-1}]S_i^{-1}\right)^{-1} = S_i + V. \tag{81}$$

We show this using the Woodbury Matrix Identity:

$$\begin{aligned}
\left(S_i^{-1}[S_i - M_i^{-1}]S_i^{-1}\right)^{-1} &= \left(S_i^{-1} - S_i^{-1}M_i^{-1}S_i^{-1}\right)^{-1} && (82)\\
&= S_i + S_iS_i^{-1}(M_i - S_i^{-1}S_iS_i^{-1})^{-1}S_i^{-1}S_i && (83)\\
&= S_i + (V_i^{-1} + S_i^{-1} - S_i^{-1})^{-1} && (84)\\
&= S_i + V_i. && (85)
\end{aligned}$$

7.1 Likelihood and gradient

The contribution to the log-likelihood from SNP i is

$$l_i = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|S_i + V| - \frac{1}{2}\mathrm{tr}(\hat{\theta}_i\hat{\theta}_i^T(S_i + V)^{-1}). \tag{86}$$

The gradient is

$$\frac{dl_i}{dV} = -\frac{1}{2}(S_i + V)^{-1} + \frac{1}{2}(S_i + V)^{-1}\hat{\theta}_i\hat{\theta}_i^T(S_i + V)^{-1}. \tag{87}$$

7.2 Incorporating linkage disequilibrium

The LDSC model implies that the expected squared effect of a SNP increases in proportion to its LD-score. Let the LD-score of SNP i be r_i, then the our model can incorporate the LDSC model by modifying the distribution of true effects to be

$$\theta_i \sim \mathcal{N}(0, r_i V). \tag{88}$$

This implies that

$$\hat{\theta}_i \sim \mathcal{N}(0, S_i + r_i V). \tag{89}$$

If we consider a univariate version of our model with $S_i = \sigma_i^2$ and $V = h^2/L$, where L is the number of loci, we have

$$\hat{\theta}_i \sim \mathcal{N}\left(0, \sigma_i^2 + r_i \frac{h^2}{L}\right), \tag{90}$$

and

$$\hat{z}_i = \frac{\hat{\theta}_i}{\sigma_i} \sim \mathcal{N}\left(0, 1 + r_i \frac{h^2}{\sigma^2 L}\right). \tag{91}$$

In the derivation of LDSC, $\sigma_i^2 = N^{-1}$, due to normalisation of both genotype and phenotype. Therefore,

$$\hat{z}_i \sim \mathcal{N}\left(0, 1 + r_i \frac{N}{L} h^2\right), \tag{92}$$

and therefore

$$\mathbb{E}[\hat{z}_i^2] = 1 + r_i \frac{N}{L} h^2, \tag{93}$$

which is the basis of the simplest LDSC model.

This shows that our model with true effects distributed as $\mathcal{N}(0, r_i V)$ provides a generalisation of LDSC to the estimation of multiple effects of a SNP when the joint sampling distribution of those effects is known.

We can define a multivariate analogue of the z-score used in LDSC:

$$\hat{z}_i = S_i^{-\frac{1}{2}} \hat{\theta}_i \sim \mathcal{N}\left(0, \mathrm{I} + r_i S_i^{-\frac{1}{2}} V S_i^{-\frac{1}{2}}\right). \tag{94}$$

This implies that

$$\mathbb{E}[\hat{z}_i \hat{z}_i^T] = \mathrm{I} + r_i S_i^{-\frac{1}{2}} V S_i^{-\frac{1}{2}}, \tag{95}$$

which is a multivariate analogue of the key LDSC equation. While V could be estimated from such a model by multiple regressions of elements of $\hat{z}_i \hat{z}_i^T$ onto elements of $\mathrm{I} + r_i S_i^{-\frac{1}{2}} V S_i^{-\frac{1}{2}}$, this will be inefficient due to the fact that the regressions would not, in general, be independent. (If both S_i and V were diagonal, the regressions based on the diagonal elements of $\hat{z}_i \hat{z}_i^T$ would be independent, but this is unlikely to be an interesting scenario for application of this method.)

7.2.1 Approximate joint likelihood

There is the further question of how to deal with linkage disequilibrium between SNPs in forming a joint likelihood across all SNPs. We follow the weighting scheme proposed by the authors of LDSC: to weight the likelihood contribution at each locus by the inverse of the total squared correlation with other SNPs nearby in the regression. Let $u_i = \sum_{j \in \mathcal{R}_i} \text{Corr}(g_i, g_j)^2$, where $\mathcal{R}_i$ is the set of indices of SNPs in the model that are within the local region of i, usually defined to be within 1 cM of SNP i. Then the approximate joint log-likelihood, l, is

$$l = \sum_{i=1}^{L} \frac{1}{u_i} l_i = \sum_{i=1}^{L} -\frac{1}{2u_i} \left[d \log(2\pi) + \log |S_i + r_i V| + \text{tr}(\hat{\theta}_i \hat{\theta}_i^T (S_i + r_i V)^{-1}) \right] \tag{96}$$

The gradient of the approximate joint likelihood is therefore:

$$\frac{dl}{dV} = \sum_{i=1}^{L} -\frac{1}{2u_i} \left[(S_i + r_i V)^{-1} - (S_i + r_i V)^{-1} \hat{\theta}_i \hat{\theta}_i^T (S_i + r_i V)^{-1} \right]. \tag{97}$$

7.3 Incorporating allele frequency

To ensure that the V matrix corresponds to the phenotypic variance explained by the different effects, we first scale the estimated effects and corresponding sampling covariance:

$$\hat{\theta}_i \to \sqrt{2f_i(1-f_i)} \hat{\theta}_i; \; S_i \to 2f_i(1-f_i) S_i. \tag{98}$$

Given this, V multiplied by the number of loci, L, gives the variance-covariance matrix of the corresponding components of phenotypic variance. This makes an implicit assumption that average effect size increases in proportion to $[2f_i(1-f_i)]^{-\frac{1}{2}}$, and that the expected contribution of each SNP to the phenotypic variance is independent of allele frequency.

8 Application to UK Biobank

We apply the method to the subsample of the white British subsample of the UKB who also have a first degree relative genotyped in the UKB.

8.1 Identification of relatives

We used the UK Biobank sample in the White British cluster which had not been identified by UK Biobank to have excess relatives, excess heterozygosity, or sex chromosome aneuploidy. Then we used the kinship coefficients computed by UK Biobank to identify individuals with a first degree relative, where a first degree relation is defined as a kinship coefficient above 0.177 [ref].

We extracted the genotypes for that subsample of the UK Biobank, removing SNPs with missingness above 5%. We then used KING[ref] to infer the sibling and parent-offspring relations within that set of individuals. We do this using the '–related –degree 1' option in KING. We identified 157 duplicates/monozygotic twins and removed one from each pair from further analyses. We identified 17,296 families with at least two siblings, giving a total of 19,329 sibling pairs. The maximum number of siblings in a family was 6, and 913 families had more than two siblings. We identified 4,418 families with at least one parent and one child genotyped; 736 families had at least one child and the father genotyped but not the mother genotyped; 2,798 families had at least one child and the mother but not the father genotyped; 893 families had at least one child and both parents genotyped. We identified 31 families with at least two children and both parents genotyped, 'parent-sib quads'.

8.1.1 Inference of Identity-by-Descent

We inferred IBD segments between all first degree relatives using the KING –ibdsegs option. We confirmed the accuracy of the IBD segment inference by using the 31 white British families where two siblings and both of their parents have been genotyped. When both parents are heterozygous, IBS implies IBD except when both siblings are heterozygous. We computed the fraction of sites inferred to be IBD 0, 1, and 2 given the true IBD state, given in Table **??**. We 'smoothed' the true IBD inferred from the quads to account for genotyping errors: if the IBD state at a SNP differed from its two immediately adjacent neighbours, and both adjacent neighbours had the same IBD state, we changed the IBD state of the SNP to be the same as its neighbours.

		Inferred IBD		
		0	1	2
True IBD	0	0.997	0.002	0.000
	1	0.017	0.982	0.001
	2	0.002	0.023	0.975

Table 5: The estimated probability of each inferred IBD state given the true IBD state, rounded to 3 decimal places.

The overall probability of inferring the correct IBD state was estimated to be 98.4%. This was computed by averaging over the three possible IBD states 0, 1, and 2, which occur with probability 0.25, 0.5, and 0.25 respectively.

8.1.2 Imputation of missing parental genotypes

For the genotyped SNPs, we imputed the missing parental genotypes for the families without genotyped parents using the approach outlined in Section [ref]. We examined the bias in the

imputed parental genotypes by performing the imputation for the 31 families with two genotyped siblings and both parents genotyped (ignoring the parental genotypes), then comparing the imputed parental genotypes to observed parental genotypes. If the imputation has worked perfectly, then the regression coefficient of the imputed parental genotypes onto the observed parental genotypes should be 1. Using 22,128,451 SNPs from 31 families, we estimated the regression coefficient to be 0.993, indicating a very slight downward bias likely due to noise arising from genotyping errors and errors in IBD inference.

We imputed the missing parental genotypes for the families with one parent genotyped using the approach outlined in Section [ref].

We also imputed missing parental genotypes for the imputed bi-allelic SNPs with INFO> 0.99 and MAF> 1%. We used hard-called genotypes from the imputed data. We used a stringent INFO threshold so that any influence of genotype errors on the imputation procedure would be minimal. As for the genotyped SNPs, we regressed the parental genotypes imputed from the siblings onto the observed parental genotypes. Using 166,587,490 SNPs, we estimated the regression coefficient to be 0.996. This shows the imputation from the siblings based upon the imputed genotypes is less biased than the imputation from the siblings based upon the genotyped array SNPs, possibly due to a reduction in genotype error in the imputed data.

8.1.3 Simulated population

We simulated traits to examine the bias in estimates of direct and indirect effects, and to examine the bias in estimates of heritability and genetic correlation from LD-score-regression.

We simulated a trait using a simulated population. We simulated genotypes at 10,000 independent SNPs with minor allele frequency 0.5 for parent-sib quads for 30,000 families, and we set one of the parent's genotypes to be missing at random for 10,000 of those families, and both parents' genotypes to be missing for 10,000 of those families. The true IBD state for each SNP for each sibling pair was recorded and used for imputation of parental genotype for those families where we set both parental genotypes as missing. The code for simulating populations of this kind is available in the *sibreg* package and is used as part of the tutorial.

We simulated direct, indirect sibling effects, and indirect maternal and paternal effects from a multivariate normal distribution:

$$\begin{bmatrix} \delta \\ \eta_s \\ \eta_p \\ \eta_m \end{bmatrix} \sim \mathcal{N}\left(0, a \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{bmatrix}\right), \tag{99}$$

where a was chosen so that the total variance explained by the direct and indirect effects was 50% of the phenotypic variance, and the remaining 50% of the phenotypic variance due to random Gaussian noise.

We performed the imputation procedure outlined above for the families with missing parental genotypes, and performed the regression with proband, sibling, and (imputed) parental geno-

types separately for the subsets with no parental genotypes missing, one parental genotypes missing, and both parental genotypes missing.

We combined estimates using the multivariate meta-analysis approach derived in Appendix [ref]. We constructed Z-statistics for each effect for each SNP as $(\hat{\beta} - \beta)/\sqrt{\text{Var}(\hat{\beta})}$, where $\hat{\beta}$ is the estimated effect, and β is the true effect. We compared the sum of the squared Z-statistics for each effect to the Chi-Square distribution on 10,000 degrees of freedom, giving p-values of 0.11, 0.82, 0.52, and 0.78 for direct, sibling, paternal, and maternal effects respectively. These results show no evidence for bias in effect estimates or estimates of standard errors. We also regressed the effect estimates onto the true effects, obtaining regression coefficients of 1.008 (S.E. 0.0259), 1.0144 (S.E. 0.0252), 0.9985 (S.E. 0.0358), and 1.0179 (S.E. 0.0349) for direct, sibling, paternal, and maternal effects respectively. Again, these results show no evidence for bias in effect estimates.

8.1.4 Simulations with UK Biobank Data

We simulated traits using the real genetic data. We randomly sampled 10,000 SNPs from the imputed data to use as causal SNPs. We standardised SNPs to have mean zero and variance 1 before simulating the genetic components of the traits.

We simulated a trait with only direct, additive genetic effects, and with the remaining phenotypic variance due to random Gaussian noise. We simulated the additive direct genetic effects from a Normal distribution, and we scaled the effects so that the resulting heritability of the trait was 40%.

We simulated 40 independent replicates of a trait affected by direct genetic effects and indirect genetic effects from parents. As we do not have access to observed genotypes for both parents from many families, we used the imputed parental genotypes for the sample of families with at least two siblings and no parents genotyped. We normalised the imputed parental genotypes (which impute the sum of paternal and maternal genotypes) to have mean zero and variance two, as the variance of the sum of maternal and paternal genotypes should be twice that of the offspring genotype under random mating. We simulated direct effects, δ, and indirect parental effects (assumed to be the same for mothers and fathers), η, from a bivariate normal distribution with correlation 0.5 between the direct and indirect effects. The effects were scaled so that the total variance explained by the combined direct and indirect effects was 60% of the phenotypic variance, with the remaining phenotypic variance due to random Gaussian noise. The correlation between direct and combined direct and indirect effects (as estimated by standard GWAS methods) was 0.866.

For the simulated traits, we estimated direct and indirect effects for all the imputed SNPs using the sample of families with at least two genotyped siblings and no genotyped parents, where we used the imputed parental genotypes in place of the observed parental genotypes. We also obtained 'standard' GWAS effect estimates by regressing on the proband genotype alone.

We tested for bias in estimates of the direct genetic effects and indirect parental genetic effects by regressing the effect estimates onto the true effects. The results for both traits are in

Table [ref]:

trait	direct	parental
direct	0.988 (0.015)	-
direct and parental	0.994 (0.003)	1.001 (0.003)

Table 6: Bias in effect estimates for simulated traits.

We estimated heritability by applying LD-score regression to both the direct effect estimates and the standard GWAS effect estimates. To apply LD-score regression to direct, indirect, and GWAS effects, we adjusted the sample size input to LD-score regression to reflect the effective sample size for each effect at each SNP. Note that the effective sample size is considerably smaller for estimation of direct and indirect effects than for GWAS effects. Let $\hat{\beta}$ be the effect estimate for a SNP with allele frequency f and with sampling variance $\text{Var}(\hat{\beta})$. We estimated the effective sample size, N_{eff}, to be

$$N_{\text{eff}} = \frac{1}{2f(1-f)\text{Var}(\hat{\beta})}, \tag{100}$$

where we use the fact that the phenotypic variance for the simulated traits is 1, and that the effects are small, so the residual variance is approximately 1.

		direct			direct and parental	
parameter	True	Est.	S.E.	True	Est.	S.E.
h^2_δ	0.4	0.470	0.044	0.144	0.167	0.005
h^2_η	0	-	-	0.144	0.225	0.005
h^2_β	0.4	0.432	0.024	0.432	0.363	0.003
$r_{\delta\eta}$	-	-	-	0.5	0.497	0.028
$r_{\delta\beta}$	1	1.034	0.022	0.866	0.888	0.004

Table 7: LD score regression results for simulated traits.

8.2 Family GWAS in UK Biobank

8.2.1 Phenotypes

We performed family based GWAS on educational attainment, height, body mass index (BMI), neuroticism score, and 'ever smoked'. For educational attainment, we converted the answers to the qualifications question (Data Field [ref]) to years of education according to the method in [ref]. For all traits, we regressed out age, age^2, age^3 sex, and interactions between sex and age, age^2, and age^3, along with the 40 genetic principal components provided by UK Biobank. For quantitative traits measured on a continuous scale (height and BMI), we performed an inverse normal transformation on the residuals separately for males and females and then combined the male and female samples.

8.2.2 Estimation of effects

We estimated effects for all imputed variants with INFO>0.99 and MAF> 1%. To enable estimation under different models from one analysis of the data, we formed summary statistics for each SNP corresponding to the X^TX matrix and X^TY vector in standard multivariate linear regression. For the subsample of families with at least two genotyped siblings and no parents genotyped, the X matrix had columns corresponding to the proband's genotype, the mean genotype of the proband's siblings, and the imputed parental genotype. Let Σ be the phenotypic covariance matrix, then the estimate of the parameters under the full model with both sibling and parental effects is:

$$\hat{\theta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y. \tag{101}$$

The estimate under the model without sibling effects is obtained by dropping the rows and columns corresponding to the average genotype of the proband's siblings from $X^T\Sigma^{-1}X$ and $X^T\Sigma^{-1}Y$. Standard GWAS estimates are obtained by using only the rows and columns corresponding to the proband genotype from $X^T\Sigma^{-1}X$ and $X^T\Sigma^{-1}Y$.

For the subsample of families with one parent genotyped, the X matrix has columns corresponding to proband genotype, (imputed) paternal, and (imputed) maternal genotypes. For the subsample with both parents genotyped, the X matrix had columns corresponding to proband genotype, paternal genotype, and maternal genotype. We did not fit indirect genetic effects from siblings for these subsets of families because only a small fraction of these families had more than one genotyped sibling.

Direct effect estimates from the different subsamples were combined using fixed effects meta-analysis. Indirect sibling effects were estimated from the subsample of families with at least two siblings genotyped and no parents genotyped alone. For parental effects, we used the multivariate meta-analysis method outline in Appendix [ref] to get meta-analysis estimates of maternal and paternal effects separately, and we took the average of those estimates to give meta-analysis estimates of the average parental effect.

For the subset of families with at least two siblings genotyped but no parents genotyped, we also implemented the difference in sibling genotypes method. We computed the mean genotype of the siblings in each family i, $\bar{g}_i$, and regressed the phenotype of each proband jointly onto the deviation of the proband's genotype from $\bar{g}_i$ and $\bar{g}_i$ in the mixed model framework outline in Section [ref].

8.2.3 LD Score Regression Results

We estimated the genetic correlation between direct effects and GWAS effects using LDSC (Figure **??**).

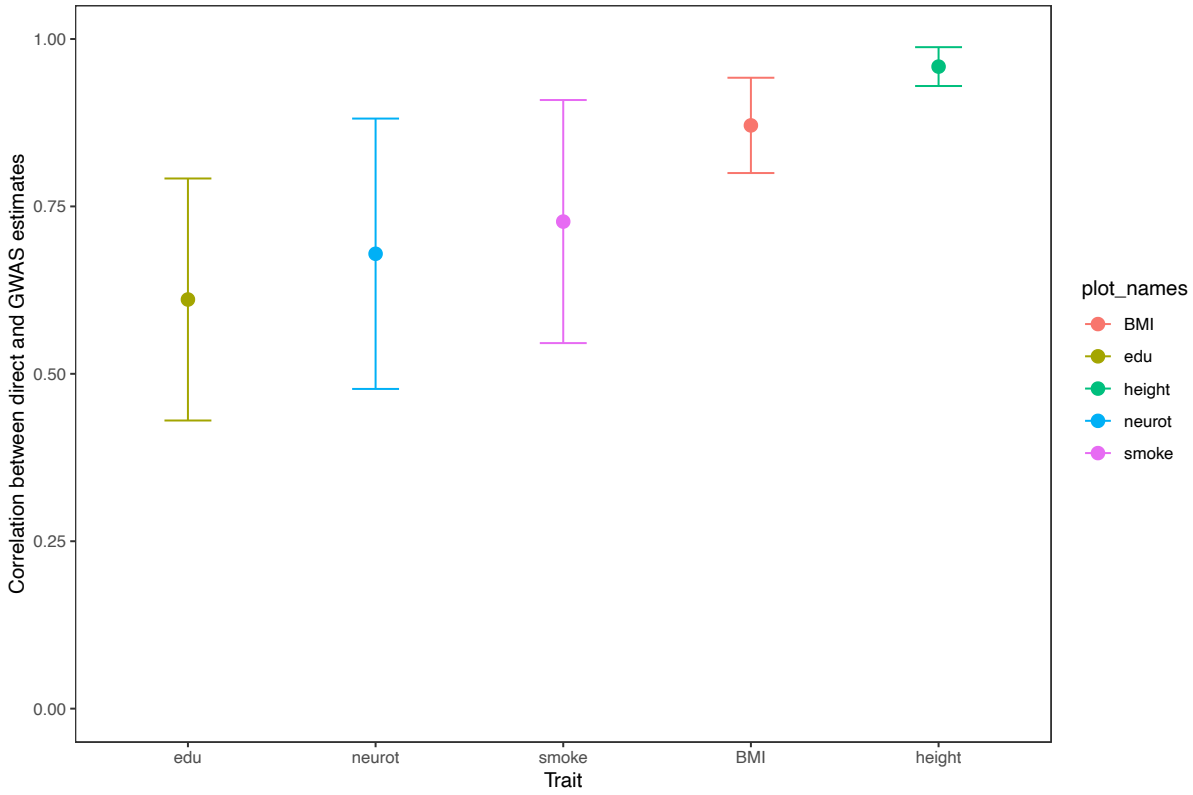

Figure 5: Genetic correlation between direct effects and GWAS effects.

We computed the relative gain in effective sample size for estimation of direct genetic effects that comes from using imputed parental genotypes over using the sib-difference method. We did this by comparing standard errors from our method with the sib-difference method for the subsample of families with at least two genotyped siblings but no genotyped parents. The relative gain in effective sample size was computed by taking the ratio of the squared standard error of the direct effect estimate from the sib-difference method to the squared standard error of the direct effect estimate from our method. The results for the five traits are in Figure **??**.

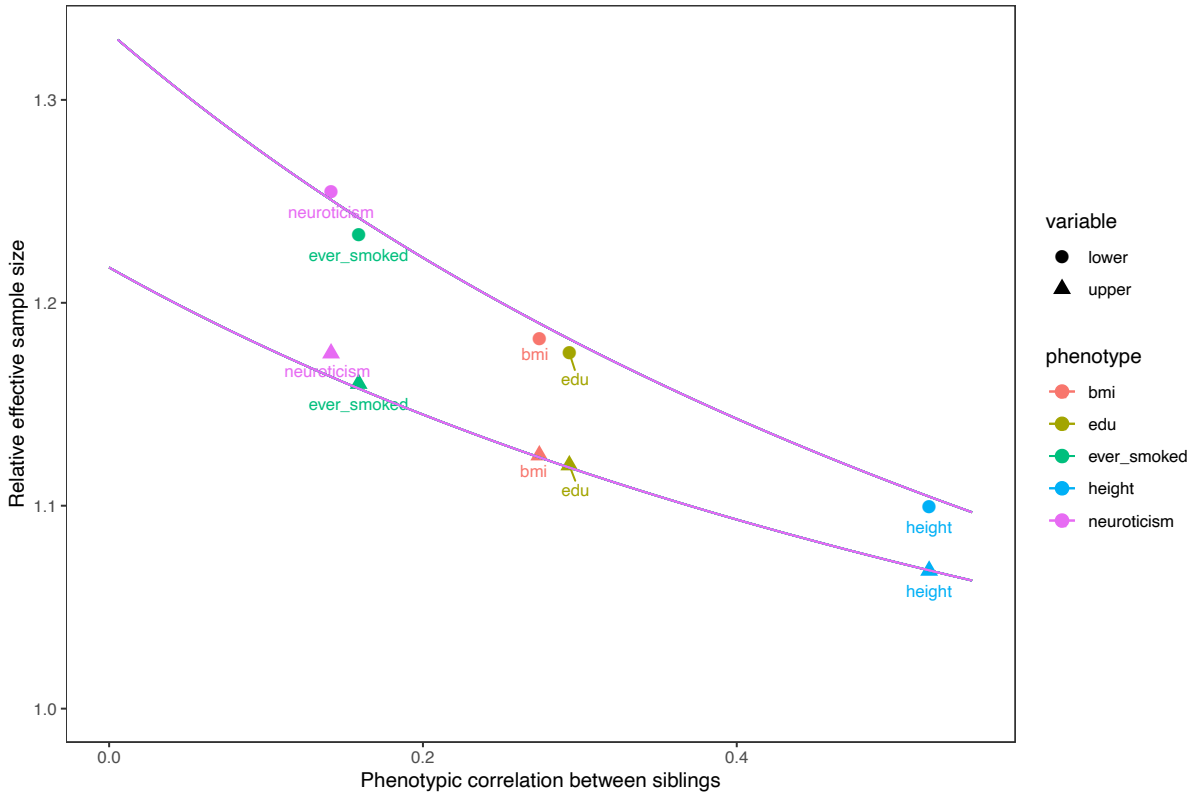

Figure 6: Increase in effective sample size from using imputed parental genotypes instead of the sib-difference method. Results are from the subset of families with at least two siblings genotyped and no parents genotyped. The relative effective sample size was calculated by taking the ratio of the squared standard errors for each SNP. For each trait, the median increase is given by the point, and the 1-99% quantile range is given by the line.

We also estimated genetic correlations between direct and GWAS estimates using direct genetic effect estimates from the sib-difference method applied to the subsample of families with at least two genotyped siblings but no parents genotyped. The results are in Table [ref]

trait	$r_{\delta\beta}$	S.E.	$\mathrm{P}(r_{\delta\beta} < 1)$
bmi	0.892	0.045	8.4×10^{-3}
EA	0.984	0.337	0.480
ever smoked	0.878	0.100	0.109
height	0.989	0.023	0.305
neuroticism	0.994	0.235	0.490

Table 8: LD score regression results for simulated traits.

While simulation results imply that, while LDSC estimates of genetic correlation are reliable,

the LDSC estimates of heritability are not reliable. Nevertheless, we present LDSC estimates of h^2_δ and h^2_β in Figure [ref] for completeness.

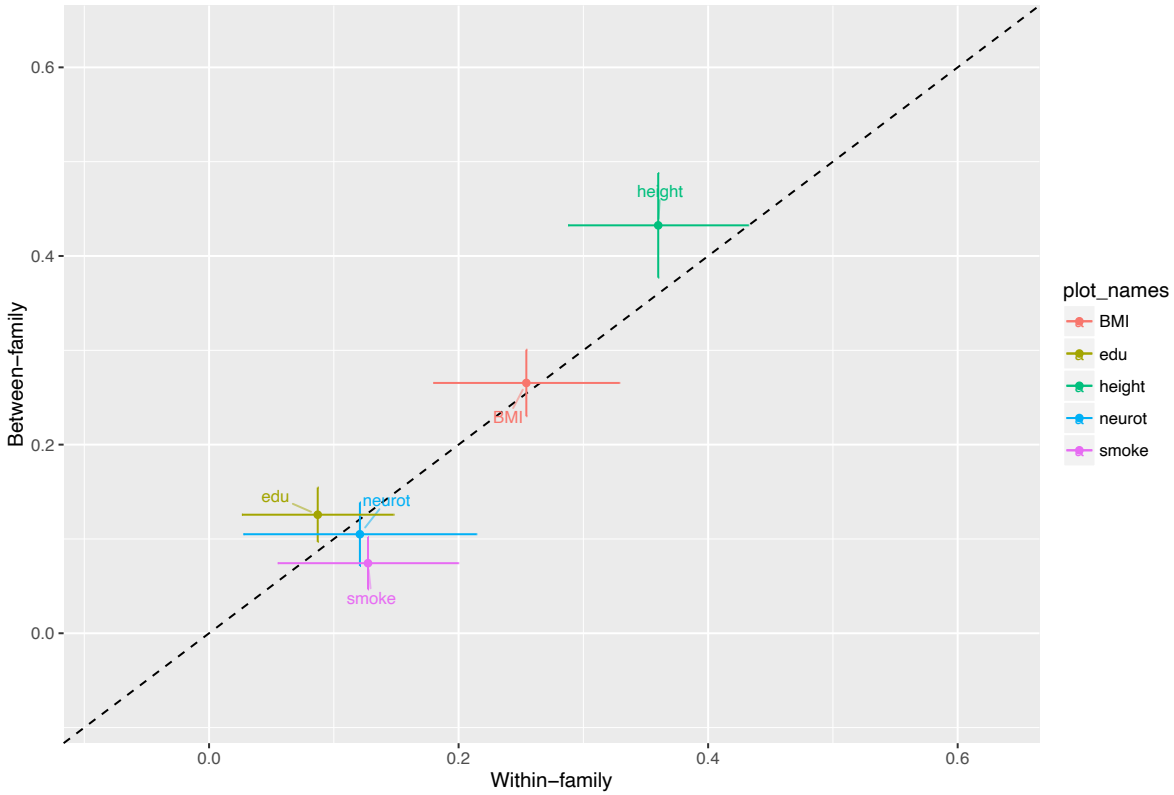

Figure 7: Increase in effective sample size from using imputed parental genotypes instead of the sib-difference method. Results are from the subset of families with at least two siblings genotyped and no parents genotyped. The relative effective sample size was calculated by taking the ratio of the squared standard errors for each SNP. For each trait, the median increase is given by the point, and the 1-99% quantile range is given by the line.

The simulation results imply that h^2_β may be underestimated whereas h^2_δ may be overestimated. This phenomenon may have suppressed actual differences in h^2_δ and h^2_β.

A Conditional Independence Lemma

While the following has probably been proven before, we prove it here for completeness.

Lemma 1. Consider random variables x, y, and random column vector z such that $x \perp y \,|\, z$, and, for constants α and b, and a constant vector of length equal to z, a,

$$\mathbb{E}[x|z] = \alpha + ba^T z, \tag{102}$$

i.e. the the conditional expectation of x given z is a linear function of some linear combination of the elements of z, then

$$\text{Cov}(x, r) = 0, \text{ where } r = y - \frac{\text{Cov}(y, a^T z)}{\text{Var}(a^T z)} a^T z \tag{103}$$

Remark. This means that if the expectation of x given z is a linear function of the elements of z, and if x is independent of y given z, then the residual of the regression of y on $a^T z$ is uncorrelated with x. Note that we only assume that there is a linear relationship between x and z, not y and z.

Proof. To prove it, first note that by standard regression theory,

$$b = \frac{\text{Cov}(x, a^T z)}{\text{Var}(a^T z)}, \tag{104}$$

so

$$\text{Cov}(x, r) = \text{Cov}(y, x) - \text{Cov}(y, a^T z) b. \tag{105}$$

It therefore suffices to show that $\text{Cov}(y, x) = \text{Cov}(y, a^T z) b$.

By the Law of Total Covariance

$$\text{Cov}(y, x) = \mathbb{E}_z[\text{Cov}(y, x|z)] + \text{Cov}_z(\mathbb{E}[x|z], \mathbb{E}[y|z]) = \text{Cov}_z(\mathbb{E}[x|z], \mathbb{E}[y|z]), \tag{106}$$

as $\text{Cov}(y, x|z) = 0$, because $x \perp y \,|\, z$. Therefore,

$$\text{Cov}(y, x) = \text{Cov}_z\left(\alpha + b a^T z, \mathbb{E}[y|z]\right) = \text{Cov}(\mathbb{E}[y|z], a^T z) b \tag{107}$$

It now suffices to show that $\text{Cov}(y, a^T z) = \text{Cov}(\mathbb{E}[y|z], a^T z)$. Without loss of generality, for some ϵ such that $\mathbb{E}[\epsilon|z] = 0$,

$$y = \mathbb{E}[y|z] + \epsilon. \tag{108}$$

Therefore, $\text{Cov}(y, a^T z) = \text{Cov}(\mathbb{E}[y|z], a^T z) + \text{Cov}(\epsilon, a^T z)$. $\text{Cov}(\epsilon, a^T z) = \mathbb{E}_z[a^T z \mathbb{E}[\epsilon|z]] - \mathbb{E}[a^T z]\mathbb{E}[\epsilon] = -\mathbb{E}[a^T z]\mathbb{E}[\epsilon]$, as $\mathbb{E}[\epsilon|z] = 0$. We also have that $\mathbb{E}[\epsilon] = \mathbb{E}_z[\mathbb{E}[\epsilon|z]] = 0$. Therefore, $\text{Cov}(\epsilon, a^T z) = 0$ and $\text{Cov}(y, a^T z) = \text{Cov}(\mathbb{E}[y|z], a^T z)$, implying

$$\text{Cov}(y, x) = \text{Cov}(y, a^T z) b \;\Rightarrow \text{Cov}(x, r) = 0. \tag{109}$$

□

B Imputation regression theory

Lemma 2. Consider two random column vectors X_0 and X_1. Let $\hat{X}_1 = \mathbb{E}[X_1|X_0]$, then $\text{Cov}(X_1, \hat{X}_1) = \text{Var}(\hat{X}_1)$ and $\text{Cov}(X_0, \hat{X}_1) = \text{Cov}(X_0, X_1)$.

Proof. First note that $\mathbb{E}[\hat{X}_1] = \mathbb{E}[\mathbb{E}[X_1|X_0]] = \mathbb{E}[X_1]$ by the Law of Iterated Expectations. We now compute

$$\begin{aligned} \mathrm{Cov}(X_1, \hat{X}_1) &= \mathbb{E}[X_1\hat{X}_1^T] - \mathbb{E}[X_1]\mathbb{E}[\hat{X}_1]^T && (110)\\ &= \mathbb{E}[\mathbb{E}[X_1\hat{X}_1^T|X_0]] - \mathbb{E}[\hat{X}_1]\mathbb{E}[\hat{X}_1]^T, && (111)\end{aligned}$$

by the Law of Iterated Expectations, and using the fact that $\mathbb{E}[\hat{X}_1] = \mathbb{E}[X_1]$. Since conditional on X_0, $\hat{X}_1$ is constant, we have that $\mathbb{E}[\mathbb{E}[X_1\hat{X}_1^T|X_0]] = \mathbb{E}[\mathbb{E}[X_1|X_0]\hat{X}_1^T] = \mathbb{E}[\hat{X}_1\hat{X}_1^T]$, and therefore

$$\mathrm{Cov}(X_1, \hat{X}_1) = \mathbb{E}[\hat{X}_1\hat{X}_1^T] - \mathbb{E}[\hat{X}_1]\mathbb{E}[\hat{X}_1]^T = \mathrm{Var}(\hat{X}_1). \quad (112)$$

We use the Law of Total Covariance to compute $\mathrm{Cov}(X_0, X_1)$:

$$\mathrm{Cov}(X_0, X_1) = \mathbb{E}[\mathrm{Cov}(X_0, X_1|X_0)] + \mathrm{Cov}(\mathbb{E}[X_0|X_0], \mathbb{E}[X_1|X_0]). \quad (113)$$

Since X_0 is a constant given X_0, $\mathrm{Cov}(X_0, X_1|X_0) = 0$. Therefore,

$$\mathrm{Cov}(X_0, X_1) = \mathrm{Cov}(\mathbb{E}[X_0|X_0], \mathbb{E}[X_1|X_0]) = \mathrm{Cov}(X_0, \hat{X}_1). \quad (114)$$

□

Theorem 3. Let $X = [X_0\ X_1]$; $\hat{X}_1 = \mathbb{E}[X_1|X_0]$; $\hat{X} = [X_0\ \hat{X}_1]$; and $Y = X\theta + \epsilon$, where $\epsilon \perp X$. Then $\hat{\theta} = (\hat{X}^T\hat{X})^{-1}\hat{X}^TY$ is a consistent estimator of θ provided that $\mathrm{Var}(\hat{X})$ is invertible.

Proof. Provided that $\mathrm{Var}(\hat{X})$ is invertible,

$$\lim_{n\to\infty} \hat{\theta} = \mathrm{Var}(\hat{X})^{-1}\mathrm{Cov}(\hat{X}, Y). \quad (115)$$

Using the above Lemma, we have that $\mathrm{Cov}(\hat{X}, Y) = \mathrm{Cov}(\hat{X}, X\theta) = \mathrm{Var}(\hat{X})\theta$. Therefore,

$$\lim_{n\to\infty} \hat{\theta} = \mathrm{Var}(\hat{X})^{-1}\mathrm{Var}(\hat{X})\theta = \theta. \quad (116)$$

□

C Meta-analysis

Consider a parameter vector α and independent observations $z_i \sim \mathcal{N}(A_i\alpha, \Sigma_i)$ for $i = 1, \ldots, k$, then it can be shown that the MLE for α is

$$\hat{\alpha} = \left(\sum_{i=1}^{k} A_i^T\Sigma_i^{-1}A_i\right)^{-1}\left(\sum_{i=1}^{k} A_i^T\Sigma_i^{-1}z_i\right), \quad (117)$$

with $\mathbb{E}[\hat{\alpha}] = \alpha$ and

$$\mathrm{Var}(\hat{\alpha}) = \left(\sum_{i=1}^{k} A_i^T \Sigma_i^{-1} A_i \right)^{-1}. \tag{118}$$

Note that this assumes that $\sum_{i=1}^{k} A_i^T \Sigma_i^{-1} A_i$ is invertible.

If the estimates are from generalised least squares such that $z_i = (X_i^T \Omega_i^{-1} X_i)^{-1} X_i^T \Omega_i^{-1} Y_i$, then we have that

$$\hat{\alpha} = \left(\sum_{i=1}^{k} A_i^T (X_i^T \Omega_i^{-1} X_i) A_i \right)^{-1} \left(\sum_{i=1}^{k} A_i^T (X_i^T \Omega_i^{-1} Y_i) \right), \tag{119}$$

and

$$\mathrm{Var}(\hat{\alpha}) = \left(\sum_{i=1}^{k} A_i^T (X_i^T \Omega_i^{-1} X_i) A_i \right)^{-1}. \tag{120}$$

Consider estimating parameters in the model

$$Y_i = \delta g_i + \eta_p g_{p(i)} + \eta_m g_{m(i)} + \epsilon_i, \tag{121}$$

where ϵ_i is uncorrelated with g_i, $g_{p(i)}$, and $g_{m(i)}$. We assume that indirect effects from siblings are zero, $\eta_s = 0$. We consider estimating $\theta = [\delta, \eta_p, \eta_m]^T$ using different samples with different observations of sibling and parental alleles. Let sample 1 be a sample where only proband genotypes are available and to which standard GWAS analysis has been applied. Let sample 2 be a sample of sibling pairs with observed genotype and phenotype, to which the sib-difference method has been applied. Let sample 3 be a sample where only proband and paternal genotype has been observed, and the estimate comes from regression of proband phenotype jointly onto proband and paternal genotype. Let sample 4 be a sample where only proband and maternal genotype has been observed, and the estimate comes from regression of proband phenotype jointly onto proband and maternal genotype. And let sample 5 be a sample where proband and both parents genotypes have been observed, and the estimate comes from regression of proband phenotype jointly onto proband, maternal, and paternal genotypes. Then we can combine the estimates from these regressions using the following matrices that give the linear transformation between θ and the expected estimates from the regressions in each subsample:

sample	observed genotypes	regression	$\mathbb{E}[z_i]$	A_i
1	proband	$Y_{ij} \sim g_i$	$\delta + (\eta_p + \eta_m)/2$	$\begin{bmatrix}1 & 0.5 & 0.5\end{bmatrix}$
2	sibling pair	$(Y_{i1} - Y_{i2}) \sim (g_{i1} - g_{i2})$	δ	$\begin{bmatrix}1 & 0 & 0\end{bmatrix}$
3	proband and paternal	$Y_i \sim g_i + g_{p(i)}$	$\begin{bmatrix}\delta + \frac{2}{3}\eta_m \\ \eta_p - \frac{1}{3}\eta_m\end{bmatrix}$	$\begin{bmatrix}1 & 0 & \frac{2}{3} \\ 0 & 1 & -\frac{1}{3}\end{bmatrix}$
4	proband and maternal	$Y_i \sim g_i + g_{m(i)}$	$\begin{bmatrix}\delta + \frac{2}{3}\eta_p \\ \eta_m - \frac{1}{3}\eta_p\end{bmatrix}$	$\begin{bmatrix}1 & \frac{2}{3} & 0 \\ 0 & -\frac{1}{3} & 1\end{bmatrix}$
5	proband, paternal, and maternal	$Y_i \sim g_i + g_{p(i)} + g_{m(i)}$	$\begin{bmatrix}\delta \\ \eta_p \\ \eta_m\end{bmatrix}$	I_3

Table 9: A matrices for meta-analysis combining different samples to estimate $\theta = [\delta,\, \eta_p,\, \eta_m]^T$.

We analyse theoretically a simple scenario where we combine results from a trio GWAS estimating direct and average maternal and paternal indirect effects, and a standard GWAS that estimates the sum of direct and indirect effects. Let $\alpha = [\delta,\, \eta]^T$ and let n_0 be the number of independent individuals with both parents genotyped used in the trio GWAS, and let n_1 be the number of independent individuals used in the standard GWAS. Then we have that

$$z_0 \sim \mathcal{N}\left(\alpha, \frac{\sigma^2_\epsilon}{n_0 2f(1-f)} \begin{bmatrix}1 & 1\\ 1 & 2\end{bmatrix}^{-1}\right); \tag{122}$$

and

$$z_1 \sim \mathcal{N}\left(\begin{bmatrix}1\\ 1\end{bmatrix}^T \alpha, \frac{\sigma^2_\epsilon}{n_1 2f(1-f)}\right). \tag{123}$$

We therefore have that

$$\mathrm{Var}(\hat{\alpha}) = \left(\frac{n_0 2f(1-f)}{\sigma^2_\epsilon} \begin{bmatrix}1 & 1\\ 1 & 2\end{bmatrix} + \frac{n_1 2f(1-f)}{\sigma^2_\epsilon} \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\right)^{-1} \tag{124}$$

$$= \frac{\sigma^2_\epsilon}{2f(1-f)} \begin{bmatrix}n_0 + n_1 & n_0 + n_1\\ n_0 + n_1 & 2n_0 + n_1\end{bmatrix}^{-1} \tag{125}$$

$$= \frac{\sigma^2_\epsilon}{2f(1-f)} \frac{1}{n_0(n_0 + n_1)} \begin{bmatrix}2n_0 + n_1 & -(n_0 + n_1)\\ -(n_0 + n_1) & n_0 + n_1\end{bmatrix}. \tag{126}$$

We therefore have that

$$\mathrm{Var}(\hat{\delta}) = \frac{\sigma^2_\epsilon}{2f(1-f)} \left(\frac{1}{n_0} + \frac{1}{n_0 + n_1}\right). \tag{127}$$

We can compare this to the variance of the estimator of direct effects using only the n_0 trios, $\hat{\delta}_0$:

$$\frac{\text{Var}(\hat{\delta}_0)}{\text{Var}(\hat{\delta})} = 1 + \frac{n_1}{2n_0 + n_1} \to 2 \text{ as } n_1 \to \infty. \tag{128}$$

Alternatively, if we have non-independent samples with known covariance between the estimates z_1, z_2, ..., z_k, then we can obtain an estimate by forming

$$z = \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_k \end{bmatrix}; \; A = \begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_k \end{bmatrix}; \text{ and } \Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{12} & \dots & \Sigma_{1k} \\ \Sigma_{12}^T & \Sigma_2 & \dots & \Sigma_{2k} \\ \vdots & \vdots & \dots & \vdots \\ \Sigma_{1k}^T & \Sigma_{2k}^T & \dots & \Sigma_k \end{bmatrix}, \tag{129}$$

where Σ_{ij} gives the sample covariance between z_i and z_j and can be estimated from LD-score regression. Then we have that

$$z \sim \mathcal{N}(A\alpha, \Sigma) \tag{130}$$

Then, from the above, the MLE for α is

$$\hat{\alpha} = (A^T \Sigma^{-1} A)^{-1} (A^T \Sigma^{-1} z), \tag{131}$$

with variance $\text{Var}(\hat{\alpha}) = (A^T \Sigma^{-1} A)^{-1}$.

D Joint distribution of sibling genotypes

The joint distribution of sibling genotypes can be derived by conditioning on the parental genotypes:

$$\mathbb{P}(g_{i1}, g_{i2}) = \sum_{g_{m(i)}, g_{p(i)}} \mathbb{P}(g_{i1}, g_{i2} | g_{m(i)}, g_{p(i)}) \mathbb{P}(g_{m(i)}, g_{p(i)}). \tag{132}$$

Since sibling genotypes are determined by independent random segregations in the parents, they are conditionally independent given parental genotype. Therefore,

$$\mathbb{P}(g_{i1}, g_{i2}) = \sum_{g_{m(i)}, g_{p(i)}} \mathbb{P}(g_{i1} | g_{m(i)}, g_{p(i)}) \mathbb{P}(g_{i2} | g_{m(i)}, g_{p(i)}) \mathbb{P}(g_{m(i)}, g_{p(i)}). \tag{133}$$

Under assumptions of random mating, the parental genotypes are independent. Therefore,

$$\mathbb{P}(g_{i1}, g_{i2}) = \sum_{g_{m(i)}, g_{p(i)}} \mathbb{P}(g_{i1} | g_{m(i)}, g_{p(i)}) \mathbb{P}(g_{i2} | g_{m(i)}, g_{p(i)}) \mathbb{P}(g_{m(i)}) \mathbb{P}(g_{p(i)}). \tag{134}$$

The above probabilities can be computed (laboriously) by application of Mendelian laws of inheritance and using parental genotype frequencies at Hardy-Weinberg equilibrium.

		g_{i2}		
		0	1	2
g_{i1}	0	$(1-f)^2(1-f/2)^2$	$f(1-f)^2(1-f/2)$	$f^2(1-f)^2/4$
	1	$f(1-f)^2(1-f/2)$	$f(1-f)[1+f(1-f)]$	$f^2(1-f)(1+f)/2$
	2	$f^2(1-f)^2/4$	$f^2(1-f)(1+f)/2$	$f^2(1+f)^2/4$

Table 10: $\mathbb{P}(g_{i1}, g_{i2})$

Variance-covariance matrix for [δ,η] using imputed parental genotype with variance equal to $2f(1-f)v$ is

$$\frac{\sigma_\epsilon^2(1+r)}{4nf(1-f)[(2-r)v-2(1-r)]}\begin{bmatrix}2(1-r)v & -2(1-r)\\ -2(1-r) & (2-r)\end{bmatrix} \tag{135}$$