Confounding of indirect genetic and epigenetic effects

Ilse Krätschmer[1*], Matthew R. Robinson[1*]

[1]Institute of Science and Technology Austria, Klosterneuburg, Austria

[*]To whom correspondence should be addressed:
ilse.kraetschmer@ist.ac.at, matthew.robinson@ist.ac.at

An individual's phenotype reflects a complex interplay of the direct effects of their DNA, epigenetic modifications of their DNA induced by their parents, and indirect effects of their parents' DNA. Here, we present theory that shows there is a greater degree of potential confounding between direct, epigenetic imprinting, maternal and paternal genetic effects than has been previously considered, especially under assortative mating. We show how covariances among these effects contribute substantially to the population-level variance. Assortative mating alters the variance by inducing increased homozygosity and correlations within and across loci, which may lead to covariance of parental genetic effects, a measure of the similarity of the indirect genetic effects among parents on their children. We propose that this assortment for parental characteristics, where biological parents create similar environments for their children, can create shared parental effects across traits and the appearance of cross-trait assortative mating. Our results demonstrate that single-locus, marginal estimates commonly made in genome-wide association studies are impossible to interpret causally, even from within-family studies. Understanding the genetic basis of complex traits requires controlling for the covariances among loci across the genome (both cis- and trans-correlations) when estimating genetic effects, which could be achieved by fitting all variants and all forms of genetic effect (direct, maternal, paternal, parent-of-origin) jointly.

Keywords: epigenetic, parent-of-origin, genetic nurture, confounding, GWAS

Introduction

In humans, parental influence persists long beyond childhood, shaping adult health and disease[1]. A basic principle of genetics is that observable characteristics result from two sources: the expression of an individual's DNA and the environment that they experience[2]. Parental characteristics shape a child's environment, giving rise to indirect genetic effects, where the genotypes of both the mother and father influence the traits of their children, beyond simply the alleles that are inherited[2], as displayed in Figure 1. Many studies have now shown that failing to separate the effects of an individual's DNA from the indirect effects of their parental genotypes results in confounding within genetic association studies[3,4,5]. Additional sources of confounding in population-level studies are: (i) correlations between loci and the environment (environmental confounding); (ii) correlations between non-causal loci and other causal loci, driven by population structure, assortative mating, and selection (genetic confounding); and (iii) an inability to separate genetic from environmental effects if an allele's effect is dependent on the environment (gene-environment interactions). With the increasing availability of genotyped family members, family-based studies have been proposed as a method to overcome confounding, as Mendelian segregation provides a natural randomization experiment[6]. These experimental designs are assumed to produce unbiased estimates of direct genetic effects[7,8,9], defined as effects of the DNA of an individual on its phenotype, and to accurately determine the extent to which parental genotypes (or other family genotypes) causally affect phenotype through development (parental genetic effects, or 'genetic nurture'), by studying associations between untransmitted parental alleles and the child's phenotype. Recent work[10] has however shown that even family-based genome-wide association studies (GWAS) may be confounded and affected by bias.

Here, without limiting ourselves to GWAS assumptions, we show the confounded nature of direct and indirect genetic effects as well as parent-of-origin effects, which are typically not considered in many studies. Epigenetic effects are increasingly recognized as an important potential source of variation in complex traits and have emerged as the focus of a rapidly expanding area of research. Principle among these effects is genomic imprint-

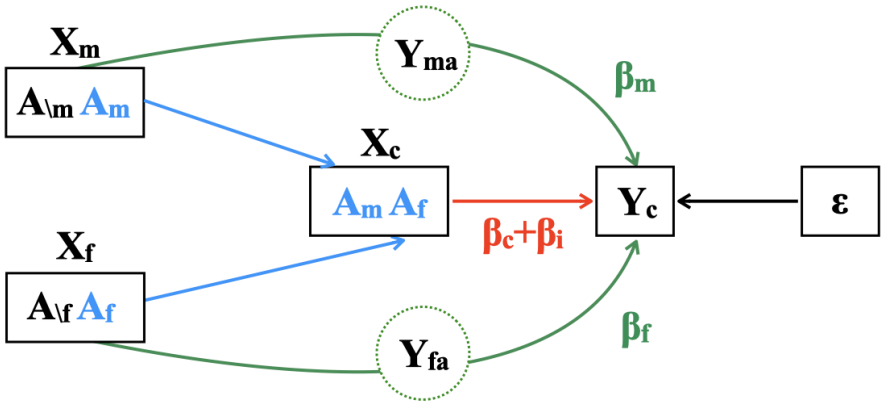

Figure 1: Direct ($\boldsymbol{\beta_c}$), parent-of-origin ($\boldsymbol{\beta_i}$) and parental ($\boldsymbol{\beta_m}$, $\boldsymbol{\beta_f}$) genetic effects on the trait of children ($\boldsymbol{Y_c}$). The parental genotypes (X_m, X_f) are split into transmitted alleles (A_m, A_f), which constitute the genotype of the child (X_c), and untransmitted ones ($A_{\backslash m}$, $A_{\backslash f}$). Parental effects can be mediated through any parental phenotype (Y_{ma}, Y_{fa}). The variable $\boldsymbol{\epsilon}$ represents variations created by other environmental effects experienced by the children.

ing, which has generally been examined by testing for parent-of-origin-dependent effects of alleles[11]. We use the experimental designs of child-mother-father trios and sibling pairs, and extend previous theory models[12,13,14,15] to show that studies of both parental genetic and parent-of-origin effects are confounded because untransmitted allelic effects can produce the same patterns of phenotypic variation expected for various forms of imprinting and vice versa. We then demonstrate how population-level variance is shaped by potentially complex relationships between direct, indirect and imprinting genetic effects that render their separation and quantification almost impossible, especially when assortative mating occurs within the population.

Child-mother-father trios

We first consider a simple single autosomal locus with two alleles (A_1 and A_2) with frequencies q_1 and q_2 that can have the following four effects on the child's phenotype: direct (β_c), indirect maternal (β_m), indirect paternal (β_f) and parent-of-origin (imprinting, β_i). The alleles of the child at the locus are ordered, referring first to the maternally

Table 1: A simple model for the genetic basis of human bi-parental nurture with epigenetic effects: Maternal (X_m), paternal (X_f) and child genotypes (X_c), phenotypic values of the child (Y_c), mating frequencies under random mating (z) and under assortative mating (z_{AM}) taken from Ref.[16] where $S = 2q_1q_2(1+\rho)$. Parental heterozygous genotypes are not ordered and are used interchangeably. Refer to text for more details.

X_m	X_f	X_c (A_mA_f)	Y_c				z	z_{AM}
A_1A_1	A_1A_1	A_1A_1					q_1^4	$[q_1^4 + 2\rho q_1^3 q_2 + \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S} 4q_2^2]$
	A_1A_2	A_1A_1			$+\beta_f$		$q_1^3 q_2$	$[q_1^3 q_2 + \rho q_1^2 q_2(q_2 - q_1) - \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S} 2q_2(q_2 - q_1)]$
		A_1A_2	$+\beta_c$		$+\beta_f$	$-\beta_i$	$q_1^3 q_2$	$[q_1^3 q_2 + \rho q_1^2 q_2(q_2 - q_1) - \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S} 2q_2(q_2 - q_1)]$
	A_2A_2	A_1A_2	$+\beta_c$		$+2\beta_f$	$-\beta_i$	$q_1^2 q_2^2$	$[q_1^2 q_2^2 + \rho q_1 q_2(q_1^2 + q_2^2) + \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S}(-4q_1 q_2)]$
A_1A_2	A_1A_1	A_1A_1		$+\beta_m$			$q_1^3 q_2$	$[q_1^3 q_2 + \rho q_1^2 q_2(q_2 - q_1) - \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S} 2q_2(q_2 - q_1)]$
		A_2A_1	$+\beta_c$	$+\beta_m$		$+\beta_i$	$q_1^3 q_2$	$[q_1^3 q_2 + \rho q_1^2 q_2(q_2 - q_1) - \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S} 2q_2(q_2 - q_1)]$
	A_1A_2	A_1A_1		$+\beta_m$	$+\beta_f$		$q_1^2 q_2^2$	$[q_1^2 q_2^2 - 2\rho q_1^2 q_2^2 + \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S}(q_1 - q_2)^2]$
		A_2A_1	$+\beta_c$	$+\beta_m$	$+\beta_f$	$+\beta_i$	$q_1^2 q_2^2$	$[q_1^2 q_2^2 - 2\rho q_1^2 q_2^2 + \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S}(q_1 - q_2)^2]$
		A_1A_2	$+\beta_c$	$+\beta_m$	$+\beta_f$	$-\beta_i$	$q_1^2 q_2^2$	$[q_1^2 q_2^2 - 2\rho q_1^2 q_2^2 + \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S}(q_1 - q_2)^2]$
		A_2A_2	$+2\beta_c$	$+\beta_m$	$+\beta_f$		$q_1^2 q_2^2$	$[q_1^2 q_2^2 - 2\rho q_1^2 q_2^2 + \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S}(q_1 - q_2)^2]$
	A_2A_2	A_1A_2	$+\beta_c$	$+\beta_m$	$+2\beta_f$	$-\beta_i$	$q_1 q_2^3$	$[q_1 q_2^3 + \rho q_1 q_2^2(q_1 - q_2) - \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S} 2q_1(q_1 - q_2)]$
		A_2A_2	$+2\beta_c$	$+\beta_m$	$+2\beta_f$		$q_1 q_2^3$	$[q_1 q_2^3 + \rho q_1 q_2^2(q_1 - q_2) - \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S} 2q_1(q_1 - q_2)]$
A_2A_2	A_1A_1	A_2A_1	$+\beta_c$	$+2\beta_m$		$+\beta_i$	$q_1^2 q_2^2$	$[q_1^2 q_2^2 + \rho q_1 q_2(q_1^2 + q_2^2) + \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S}(-4q_1 q_2)]$
	A_1A_2	A_2A_1	$+\beta_c$	$+2\beta_m$	$+\beta_f$	$+\beta_i$	$q_1 q_2^3$	$[q_1 q_2^3 + \rho q_1 q_2^2(q_1 - q_2) - \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S} 2q_1(q_1 - q_2)]$
		A_2A_2	$+2\beta_c$	$+2\beta_m$	$+\beta_f$		$q_1 q_2^3$	$[q_1 q_2^3 + \rho q_1 q_2^2(q_1 - q_2) - \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S} 2q_1(q_1 - q_2)]$
	A_2A_2	A_2A_2	$+2\beta_c$	$+2\beta_m$	$+2\beta_f$		q_2^4	$[q_2^4 + 2\rho q_1 q_2^3 + \rho^2 q_1^2 q_2^2][1 + \frac{\rho}{S} 4q_1^2]$

inherited allele and second to the paternally inherited one, i.e. A_mA_f. The genotypes, X_c, X_m and X_f, are coded as 0, 1 and 2 for A_1A_1 homozygotes, A_1A_2 or A_2A_1 heterozygotes and A_2A_2 homozygotes, respectively. For heterozygous children, the parent-of-origin effect is positive for maternally inherited A_2 and negative for paternally inherited A_2. From these definitions, the possible genotypic, X_c, and phenotypic values, Y_c, of the child given the genotype of the parents and imprinting are shown in Table 1. The phenotypic values represent the deviations from the case where both parents and child have A_1A_1 genotypes.

Table 1 shows that indirect parental genetic effects and epigenetic effects are confounded due to the fact that not every child genotype can be produced by all parental genotypes. For example, parents who are both the same type of homozygote (both either A_1A_1 or A_2A_2) will only have A_1A_1 and A_2A_2 children, respectively. Each of the two types of homozygous parents (A_1A_1 and A_2A_2) can produce only one of the two types

of reciprocal heterozygote (A_1A_2 and A_2A_1). The children of A_1A_1 fathers all have $+\beta_i$, with the exception of A_1A_1 children whose parents are both A_1A_1. The children of A_2A_2 fathers all have $-\beta_i$, with the exception of A_2A_2 children whose parents are both A_2A_2. The reverse is true for mothers, where the children of A_1A_1 mothers all have $-\beta_i$, with the exception of A_1A_1 children whose parents are both A_1A_1, and the children of A_2A_2 mothers all have $+\beta_i$, with the exception of A_2A_2 children whose parents are both A_2A_2. This clearly shows that when there are indirect parental genetic effects, the phenotypes of the reciprocal heterozygotes will differ, creating the appearance of parent-of-origin effects. Likewise, differences between reciprocal heterozygotes due to parent-of-origin effects will appear as parental genetic effects.

Table 1 also provides the results of the expected probability of the different parent-child combinations within the population (mating frequency), z, under random mating and Hardy-Weinberg equilibrium, where $(q_1 + q_2) = 1 = (q_1 + q_2)^2$, as well as the probabilities under assortative mating, z_{AM}, which will be discussed in the next section. The phenotypic values, Y_c, and expected probabilities, z, in Table 1 can be used to calculate the mean and variance contributed by direct, indirect genetic and epigenetic effects within the human population under random mating. The population mean is given by

$$\mu_{RM} = \sum_j Y_{cj} z_j = 2q_2(\beta_c + \beta_m + \beta_f), \tag{1}$$

where j loops over all possible genotypes of the child and is not influenced by the imprinting effects. We find that the variance contributed by this single locus reflects the following components:

$$V_{RM} = \sum_j Y_{cj}^2 z_j - \mu_{RM}^2 = 2q_1q_2(\beta_c^2 + \beta_m^2 + \beta_f^2 + \beta_i^2 + \beta_c\beta_m + \beta_c\beta_f + \beta_m\beta_i - \beta_f\beta_i). \tag{2}$$

More details on the calculations are provided in the Supplementary Material.

Equation 2 shows that direct genetic effects covary with both maternal and paternal genetic effects, which in turn covary with imprinting effects, creating a greater degree of potential confounding between all four components than has been previously considered. The additive genetic variance among offspring genotypes, apart from direct effects (β_c^2), also contains components contributed by parental genetic effects (β_m^2, β_f^2), direct-parental

genetic covariance ($\beta_c\beta_m$, $\beta_c\beta_f$), parent-of-origin effects (β_i^2), and parental-parent-of-origin genetic covariance ($\beta_i\beta_m$, $\beta_i\beta_f$). Parent-of-origin effects contribute not only directly, but also dependent on their covariance with parental genetic effects, i.e. $\beta_i^2 + \beta_i\beta_m - \beta_i\beta_f$, which represents the degree to which loci with parental genetic effects are also imprinted loci in the child. Note the negative sign of $\beta_i\beta_f$, which is due to the definition of parent-of-origin effects (positive for the maternally inherited allele and negative for the paternal one).

The implications of Equation 2 are that phenotypic variation can be attributable to the single nucleotide polymorphism (SNP) markers of an individual (i.e., SNP heritability $h^2_{SNP} > 0$) and marker-phenotype associations can be discovered in population studies, but the underlying pattern of causality can be anything from: (i) the locus does not directly affect the expression of the trait ($\beta_c^2 = 0$), but instead only has an effect in one or both of the parents; to (ii) both direct and parental genetic effects may be present at a locus, but the parental effects may cancel out if they covary negatively; or (iii) any combination of terms and their covariance.

The Supplementary Material describes a simulation study that we conducted to confirm our theoretical calculations. We compare the expected theoretical variance given in Equation 2 to the estimated variance from a simple simulation where a single causal locus within a segment of DNA with realistic linkage disequilibrium (LD) is generated. The left panel of Figure 2 shows that the expected theoretical variance matches the estimated variance for a single causal locus for a wide range of scenarios: (i) only direct effects ($\beta_c = 0.5$), (ii) direct and imprinting effects ($\beta_c = 0.5, \beta_i = 0.1$), (iii) direct, maternal and imprinting effects ($\beta_c = 0.5, \beta_i = 0.1, \beta_m = 0.1$), and (iv) direct, maternal, paternal and imprinting effects ($\beta_c = 0.5, \beta_i = 0.1, \beta_m = 0.1, \beta_f = 0.1$). As already demonstrated numerically in Equation 2, the population level variance increases when indirect genetic and imprinting effects are present. If there are no parental influences apart from parent-of-origin effects, then the variance simply increases by β_i^2, with no covariance between imprinting and direct effects. If parental genetic effects are also present, covariances then emerge and contribute substantially to the population-level variance. These covariances will be dependent upon the direction and magnitude of the maternal, paternal and imprinting

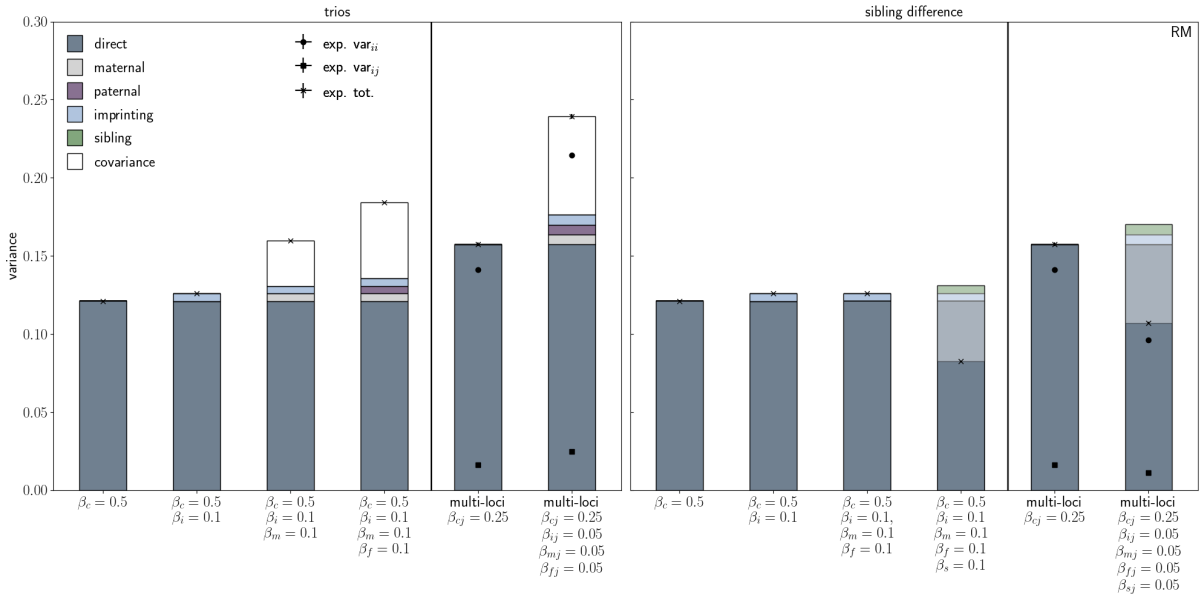

Figure 2: Expected theoretical total variance (exp. tot.), sum of variances in case of multiple loci (exp. var_{ii}), covariance between loci (exp. var_{ij}) and estimated variance (bars) for trios (left) and sibling differences (right) under random mating when considering various effects (given as labels in x-axis) for the single-locus case and for five causal loci (multi-loci) with each having the effects $\beta_c = 0.25, \beta_m = 0.05, \beta_f = 0.05, \beta_i = 0.05$, and in case of the sibling difference, $\beta_s = 0.05$. Note that the direct-sibling covariance is negative and is thus shown opaquely, covering parts of the variance.

effects of a given locus. The results given in Figure 2 clearly show that components of variance cannot be simply separated into independent direct and indirect sources that then simply sum and highlights again the confounded nature of direct, indirect genetic and parent-of-origin effects.

Extension of a single locus to a polygenic additive genetic model, summing over all l causal variants and taking into account correlations, k_{ij}, between loci, gives the total genetic variance under random mating as

$$V_{tot_{RM}} = \sum_{j=1}^{l} V_{RM_j} + \sum_{i \neq j} k_{ij} \sqrt{V_{RM_i} V_{RM_j}}. \tag{3}$$

The covariances between loci, given in the second term of Equation 3, are dependent upon the correlation among the causal variants and the concordance in their effects size. In Figure 2, the estimated variance for five causal loci, each having an effect of

$\beta_c = 0.25, \beta_m = 0.05, \beta_f = 0.05, \beta_i = 0.05$ on the phenotypic value, agree with the expected total variance from Equation 3 for the situation where causal variants are randomly selected and thus approximately unlinked.

Equation 3 highlights that single-locus, marginal estimates common to genome-wide association studies (GWAS) are impossible to interpret causally, even from within-family studies. Controlling for the covariances among loci across the genome (both cis- and trans-correlations) when estimating genetic effects would require fitting all variants and all forms of genetic effect (direct, maternal, paternal, parent-of-origin) jointly. Bayesian penalized regression models may be suitable for this, but even then large sample sizes will be required to differentiate between these highly correlated components.

Differences in siblings

The model presented in the previous section can be extended to also include sibling effects, as is shown in the Supplementary Material. However, it is more interesting to look at the relationship of the phenotypic and genotypic differences in siblings as the difference in siblings is assumed to provide an unbiased estimate of direct genetics effects[7,8,9]. But typical analyses do not consider parent-of-origin effects, nor the indirect genetic effects from siblings. Using the phenotypic differences, $Y_c - Y_s$, which include also sibling effects, and the expected probabilities of the genotypes for the different sibling combinations given in Table 2, we can calculate the expected mean and variance at population level. The total mean is 0, while the variance is

$$V_{s_{RM}} = \sum_j (Y_{cj} - Y_{sj})^2 z_j = 2q_1q_2(\beta_c^2 + \beta_i^2 + \beta_s^2 - 2\beta_c\beta_s), \tag{4}$$

which shows that indirect parental effects do indeed cancel in the difference, but indirect sibling and imprinting effects remain.

Importantly, if there is an indirect effect from the sibling on an individual's phenotype, the variance increases due to the squared sibling's effects, but decreases because of the covariance between the individual and its sibling. As shown in Table 2, this results in a total confounding of direct and indirect sibling effects, which always oppose each other in sign. This behaviour is also observed with the simulation study, as shown in the right

Table 2: Extension of Table 1 to include siblings and their differences. Mating frequencies are given in Table 1 which have to be multiplied with the factor f_z to account for the different sibling combinations. Parental heterozygous genotypes are not ordered and are used interchangeably.

X_m	X_f	X_c (A_mA_f)	X_s (A_mA_f)	Y_c					Y_s					$Y_c - Y_s$			f_z
A_1A_1	A_1A_1	A_1A_1	A_1A_1														1
	A_1A_2	A_1A_1	A_1A_1			$+\beta_f$					$+\beta_f$						$\frac{1}{2}$
			A_1A_2			$+\beta_f$		$+\beta_s$	$+\beta_c$		$+\beta_f$	$-\beta_i$		$-\beta_c$	$+\beta_i$	$+\beta_s$	$\frac{1}{2}$
		A_1A_2	A_1A_1	$+\beta_c$		$+\beta_f$	$-\beta_i$				$+\beta_f$		$+\beta_s$	$+\beta_c$	$-\beta_i$	$-\beta_s$	$\frac{1}{2}$
			A_1A_2	$+\beta_c$		$+\beta_f$	$-\beta_i$	$+\beta_s$	$+\beta_c$		$+\beta_f$	$-\beta_i$	$+\beta_s$				$\frac{1}{2}$
	A_2A_2	A_1A_2	A_1A_2	$+\beta_c$		$+2\beta_f$	$-\beta_i$	$+\beta_s$	$+\beta_c$		$+2\beta_f$	$-\beta_i$	$+\beta_s$				1
A_1A_2	A_1A_1	A_1A_1	A_1A_1		$+\beta_m$					$+\beta_m$							$\frac{1}{2}$
			A_2A_1		$+\beta_m$			$+\beta_s$	$+\beta_c$	$+\beta_m$		$+\beta_i$		$-\beta_c$	$-\beta_i$	$+\beta_s$	$\frac{1}{2}$
		A_2A_1	A_1A_1	$+\beta_c$	$+\beta_m$		$+\beta_i$			$+\beta_m$			$+\beta_s$	$+\beta_c$	$+\beta_i$	$-\beta_s$	$\frac{1}{2}$
			A_2A_1	$+\beta_c$	$+\beta_m$		$+\beta_i$	$+\beta_s$	$+\beta_c$	$+\beta_m$		$+\beta_i$	$+\beta_s$				$\frac{1}{2}$
	A_1A_2	A_1A_1	A_1A_1		$+\beta_m$	$+\beta_f$				$+\beta_m$	$+\beta_f$						$\frac{1}{4}$
			A_1A_2		$+\beta_m$	$+\beta_f$		$+\beta_s$	$+\beta_c$	$+\beta_m$	$+\beta_f$	$-\beta_i$		$-\beta_c$	$+\beta_i$	$+\beta_s$	$\frac{1}{4}$
			A_2A_1		$+\beta_m$	$+\beta_f$		$+\beta_s$	$+\beta_c$	$+\beta_m$	$+\beta_f$	$+\beta_i$		$-\beta_c$	$-\beta_i$	$+\beta_s$	$\frac{1}{4}$
			A_2A_2		$+\beta_m$	$+\beta_f$		$+2\beta_s$	$+2\beta_c$	$+\beta_m$	$+\beta_f$			$-2\beta_c$		$+2\beta_s$	$\frac{1}{4}$
		A_1A_2	A_1A_1	$+\beta_c$	$+\beta_m$	$+\beta_f$	$-\beta_i$			$+\beta_m$	$+\beta_f$		$+\beta_s$	$+\beta_c$	$-\beta_i$	$-\beta_s$	$\frac{1}{4}$
			A_1A_2	$+\beta_c$	$+\beta_m$	$+\beta_f$	$-\beta_i$	$+\beta_s$	$+\beta_c$	$+\beta_m$	$+\beta_f$	$-\beta_i$	$+\beta_s$				$\frac{1}{4}$
			A_2A_1	$+\beta_c$	$+\beta_m$	$+\beta_f$	$-\beta_i$	$+\beta_s$	$+\beta_c$	$+\beta_m$	$+\beta_f$	$+\beta_i$	$+\beta_s$		$-2\beta_i$		$\frac{1}{4}$
			A_2A_2	$+\beta_c$	$+\beta_m$	$+\beta_f$	$-\beta_i$	$+2\beta_s$	$+2\beta_c$	$+\beta_m$	$+\beta_f$		$+\beta_s$	$-\beta_c$	$-\beta_i$	$+\beta_s$	$\frac{1}{4}$
		A_2A_1	A_1A_1	$+\beta_c$	$+\beta_m$	$+\beta_f$	$+\beta_i$			$+\beta_m$	$+\beta_f$		$+\beta_s$	$+\beta_c$	$+\beta_i$	$-\beta_s$	$\frac{1}{4}$
			A_1A_2	$+\beta_c$	$+\beta_m$	$+\beta_f$	$+\beta_i$	$+\beta_s$	$+\beta_c$	$+\beta_m$	$+\beta_f$	$-\beta_i$	$+\beta_s$		$+2\beta_i$		$\frac{1}{4}$
			A_2A_1	$+\beta_c$	$+\beta_m$	$+\beta_f$	$+\beta_i$	$+\beta_s$	$+\beta_c$	$+\beta_m$	$+\beta_f$	$+\beta_i$	$+\beta_s$				$\frac{1}{4}$
			A_2A_2	$+\beta_c$	$+\beta_m$	$+\beta_f$	$+\beta_i$	$+2\beta_s$	$+2\beta_c$	$+\beta_m$	$+\beta_f$		$+\beta_s$	$-\beta_c$	$+\beta_i$	$+\beta_s$	$\frac{1}{4}$
		A_2A_2	A_1A_1	$+2\beta_c$	$+\beta_m$	$+\beta_f$				$+\beta_m$	$+\beta_f$		$+2\beta_s$	$+2\beta_c$		$-2\beta_s$	$\frac{1}{4}$
			A_1A_2	$+2\beta_c$	$+\beta_m$	$+\beta_f$		$+\beta_s$	$+\beta_c$	$+\beta_m$	$+\beta_f$	$-\beta_i$	$+2\beta_s$	$+\beta_c$	$+\beta_i$	$-\beta_s$	$\frac{1}{4}$
			A_2A_1	$+2\beta_c$	$+\beta_m$	$+\beta_f$		$+\beta_s$	$+\beta_c$	$+\beta_m$	$+\beta_f$	$+\beta_i$	$+2\beta_s$	$+\beta_c$	$-\beta_i$	$-\beta_s$	$\frac{1}{4}$
			A_2A_2	$+2\beta_c$	$+\beta_m$	$+\beta_f$		$+2\beta_s$	$+2\beta_c$	$+\beta_m$	$+\beta_f$		$+2\beta_s$				$\frac{1}{4}$
	A_2A_2	A_1A_2	A_1A_2	$+\beta_c$	$+\beta_m$	$+2\beta_f$	$-\beta_i$	$+\beta_s$	$+\beta_c$	$+\beta_m$	$+2\beta_f$	$-\beta_i$	$+\beta_s$				$\frac{1}{2}$
			A_2A_2	$+\beta_c$	$+\beta_m$	$+2\beta_f$	$-\beta_i$	$+2\beta_s$	$+2\beta_c$	$+\beta_m$	$+2\beta_f$		$+\beta_s$	$-\beta_c$	$-\beta_i$	$+\beta_s$	$\frac{1}{2}$
		A_2A_2	A_1A_2	$+2\beta_c$	$+\beta_m$	$+2\beta_f$		$+\beta_s$	$+\beta_c$	$+\beta_m$	$+2\beta_f$	$-\beta_i$	$+2\beta_s$	$+\beta_c$	$+\beta_i$	$-\beta_s$	$\frac{1}{2}$
			A_2A_2	$+2\beta_c$	$+\beta_m$	$+2\beta_f$		$+2\beta_s$	$+2\beta_c$	$+\beta_m$	$+2\beta_f$		$+2\beta_s$				$\frac{1}{2}$
A_2A_2	A_1A_1	A_2A_1	A_2A_1	$+\beta_c$	$+2\beta_m$		$+\beta_i$	$+\beta_s$	$+\beta_c$	$+2\beta_m$		$+\beta_i$	$+\beta_s$				1
	A_1A_2	A_2A_1	A_2A_1	$+\beta_c$	$+2\beta_m$	$+\beta_f$	$+\beta_i$	$+\beta_s$	$+\beta_c$	$+2\beta_m$	$+\beta_f$	$+\beta_i$	$+\beta_s$				$\frac{1}{2}$
			A_2A_2	$+\beta_c$	$+2\beta_m$	$+\beta_f$	$+\beta_i$	$+2\beta_s$	$+2\beta_c$	$+2\beta_m$	$+\beta_f$		$+\beta_s$	$-\beta_c$	$+\beta_i$	$+\beta_s$	$\frac{1}{2}$
		A_2A_2	A_2A_1	$+2\beta_c$	$+2\beta_m$	$+\beta_f$		$+\beta_s$	$+\beta_c$	$+2\beta_m$	$+\beta_f$	$+\beta_i$	$+2\beta_s$	$+\beta_c$	$-\beta_i$	$-\beta_s$	$\frac{1}{2}$
			A_2A_2	$+2\beta_c$	$+2\beta_m$	$+\beta_f$		$+2\beta_s$	$+2\beta_c$	$+2\beta_m$	$+\beta_f$		$+2\beta_s$				$\frac{1}{2}$
	A_2A_2	A_2A_2	A_2A_2	$+2\beta_c$	$+2\beta_m$	$+2\beta_f$		$+2\beta_s$	$+2\beta_c$	$+2\beta_m$	$+2\beta_f$		$+2\beta_s$				1

panel of Figure 2, where the expected and estimated within-siblings variance are shown to match under random mating for different scenarios at a single-locus and five causal loci. Thus, any indirect sibling genetic effect will act to alter the variance of the direct genetic effects. But it does appear possible to separate parent-of-origin effects within this experimental design under random mating.

Extending the single-locus model for differences in siblings to multiple loci can be done through Equation 3, substituting V_{RM} with Equation 4. For the situation where causal variants are randomly selected and unlinked, the theoretical and empirical estimates are again identical, as shown in the right panel of Figure 2. Again, controlling for the covariances among loci across the genome (both cis- and trans-correlations) when estimating genetic effects would also require fitting all variants and all forms of differences in genetic effects jointly, likely within a form of Bayesian penalized regression model.

Extension to assortative mating

Assortative mating (AM) acts to alter the mating frequencies in the population as it increases parental similarity at trait-associated loci, which is reflected in the correlation, ρ, among the parents in their alleles. This changes the genotype frequencies, and thus the probabilities of parents producing offspring of a certain genotype, given as z_{AM} in Table 1, but does not affect the minor allele frequencies in the population. The altered offspring frequencies, z_{AM}, are taken from Ref.[16]. This modification leaves the population mean unchanged under AM, but it alters the variance.

Under assortative mating, we find that a single locus with direct, indirect parental and imprinting effects makes the following contribution to the phenotypic variance:

$$\begin{aligned} V_{AM} =& 2q_1q_2(\beta_c^2 + \beta_i^2 + \beta_m^2 + \beta_f^2 + \beta_c\beta_m + \beta_c\beta_f + \beta_m\beta_i - \beta_f\beta_i) \\ &+ q_1q_2\rho(\beta_c^2 + 2\beta_m^2 + 2\beta_f^2 - \beta_i^2 + 4\beta_c\beta_m + 4\beta_c\beta_f + 4\beta_m\beta_f) \\ &+ q_1q_2\rho^2(\beta_c^2 - \beta_i^2 + 2\beta_c\beta_m + 2\beta_c\beta_f - 2\beta_m\beta_i + 2\beta_f\beta_i + 4\beta_m\beta_f). \end{aligned} \tag{5}$$

The ρ term reflects a correlation among mates at causal SNP variants (or those correlated with an underlying causal variant). Equation 5 shows that the increase in the variance

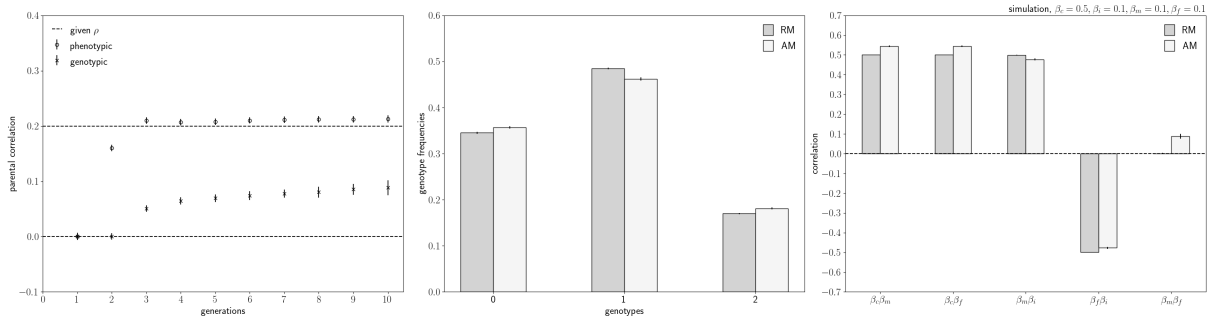

Figure 3: Simulation study for a single locus with effects $\beta_c = 0.5, \beta_i = 0.1, \beta_m = 0.1, \beta_f = 0.1$: (Left) Parental phenotypic and genotypic correlation as function of generations compared to the given correlation between parental phenotypes of 0.2, fixed in the simulation parameters. (Middle) Genotype frequencies under random mating (RM) and assortative mating (AM). (Right) Correlations between effects under RM and AM.

contributed by a locus under AM is dependent upon the relationships among the direct, parental genetic and imprinting effects.

To confirm the increase in variance due to AM, we extended our forward-in-time simulation study presented above, details of which are provided in the Supplementary Material. An example of the change in genotype frequencies for simulated data for a single locus with minor allele frequency of 0.41 is shown in the middle panel of Figure 3. A correlation between parental phenotypes is introduced in ten generations, as displayed on the left side of Figure 3. This also generates a genotypic correlation between the parents which in turn creates the covariance term, $\beta_m\beta_f$, as can be seen in the right panel of Figure 3. This represents the covariance of the parental indirect genetic effects, which is an important component of variance and can be interpreted as a measure of the similarity of the indirect genetic effects among parents on their children. This suggests that assortative mating on a trait may also create assortment for parental characteristics, where biological parents create similar environments for their children, creating covariance in the indirect genetic effects that may have a large role to play in shaping the population-level variation. Until now this covariance has been generally overlooked. We propose that shared parental effects across traits can also create the cross-trait assortative mating patterns that are observed in the human population[17].

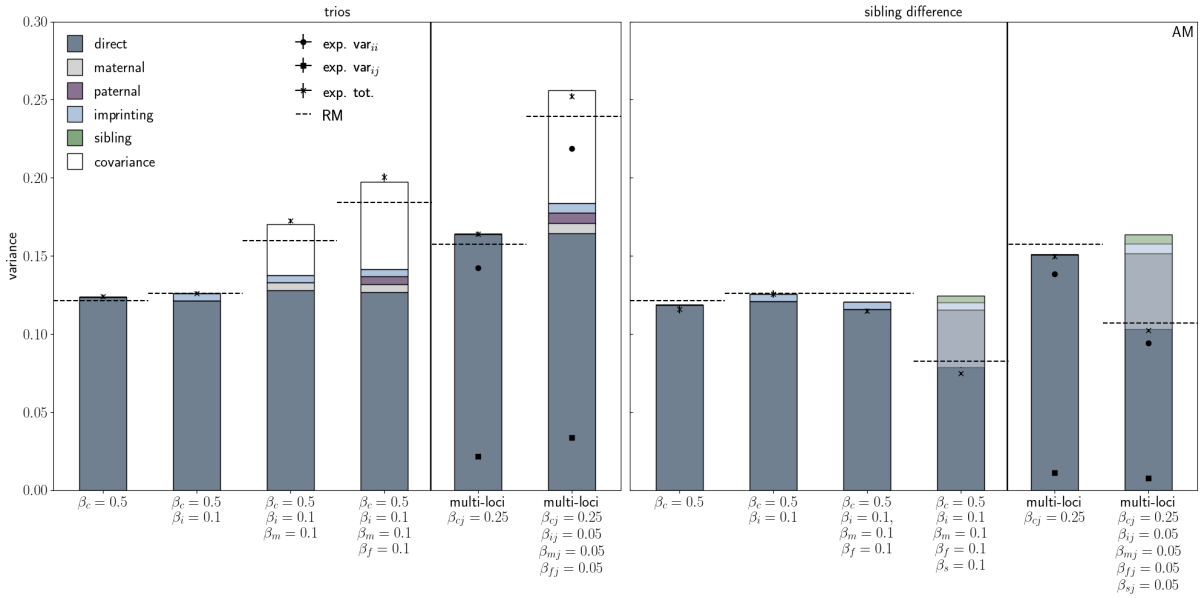

Figure 4: Expected theoretical total variance (exp. tot.), sum of variances in case of multiple loci (exp. var_{ii}), covariance between loci (exp. var_{ij}) and estimated variance (bars) for trios (left) and sibling differences (right) under assortative mating ($\rho_Y = 0.2$) when considering various effects (given as labels in x-axis) for the single-locus case and for 5 causal loci (multi-loci) with each having the effects $\beta_c = 0.25, \beta_m = 0.05, \beta_f = 0.05, \beta_i = 0.05$, and in case of the sibling difference, $\beta_s = 0.05$. Note that the direct-sibling covariance is negative and is thus shown opaquely, covering parts of the variance. The corresponding values under RM are shown as dashed lines.

For our simulation study of assortative mating we are not seeking to reach an equilibrium in either the phenotypic or genetic correlations. We appreciate that we create mate assortment by ordering the phenotypes according to a specified phenotypic correlation, which is a simplified and not realistic process. However, our aim was to simply confirm our theory and to demonstrate the directions of the (co)variances and the degree of confounding under AM. The left panel of Figure 4 shows the expected theoretical variance and the estimated variance for trios under assortative mating for a correlation between the parental phenotypes, $\rho_Y = 0.2$, for a single causal locus, as well as the multiple loci case where the expected values are calculated using Equation 3 substituting V_{RM} with V_{AM} which already considers changes in variance due to AM within the locus. The estimated variances match the expected ones, supporting our theoretical model.

The left panel of Figure 4 highlights our findings from Equation 5: if for simplicity we consider a single locus with only direct effects, the variance is enlarged only due to the increased homozygosity of causal loci that is expected under AM. Our theory also captures the correlations both within and across loci in the multi-locus examples of Figure 4. Even with only weak parental genetic and parent-of-origin effects, the covariances across terms contributes strongly to the variance in both the single- and the multi-locus example. This highlights the important role of the covariances among components, especially those that are typically overlooked, such as the parental indirect genetic covariance, the covariance of the parent-of-origin effects with both parental genetic effects, and the covariance of the direct and parental genetic effects. Thus, in the trio case (and in population-level studies), assortative mating acts to inflate the variances and covariances, to increase the confounding among the terms, and leads to greater difficulty in separating the contributions of each term. Thus, it is very difficult, maybe impossible, to control for both the covariances among loci across the genome (both cis- and trans-correlations) and the inflation of (co)variance under AM at loci when estimating genetic effects with current GWAS methods.

The change in variance due to AM induced correlations between alleles of mates at a locus and also between loci has been shown by previous theory[18] and empirical research[19]. An equilibrium is reached under AM, where there is inflated genetic variance, a larger correlation between relatives for traits driving AM and increased homozygosity at causal loci. However, previous work does not take into account the increased homozygosity of causal loci that is expected under AM when calculating the expected increase in genetic variance. Therefore, Equation 5 extends theory for the correlation induced by AM between alleles of mates at a locus under direct, indirect, and imprinting effects.

To put our work in the context of previous theory[18], we use a similar approach to understand the effect of AM on the variance taking into account multiple loci and parental influences. We follow the approach outlined in Chapter 4.7 of Ref.[18], where all alleles of l causal loci are assumed to have the same frequencies, q_1 and $q_2 = 1 - q_1$, and the same effect α on the trait under study. In our work, we consider the parental genetic influence as well as direct effects, and thus α needs to be replaced, depending upon whether the

allele is inherited from the mother or the father. The variance of a single allele from the mother is determined to be $q_1 q_2 (\beta_c + \beta_m)^2$, assuming that the effect of the allele is either $(\beta_c + \beta_m)$ or 0, with frequency q_1 and q_2. We temporarily assume that parent-of-origin effects are absent, as the procedure of Ref.[18] is based on the variance of single alleles, but we will introduce these effects again later. The mean of the maternal allele is then given by $q_1(\beta_c + \beta_m) + q_2 \cdot 0 = q_1(\beta_c + \beta_m)$. The variance is calculated as

$$\begin{aligned} V(A_m) =& q_1[(\beta_c + \beta_m) - q_1(\beta_c + \beta_m)]^2 + q_2[0 - q_1(\beta_c + \beta_m)]^2 \\ =& q_1 q_2 (\beta_c + \beta_m)^2. \end{aligned} \tag{6}$$

Similarly, the variance for one paternal allele is

$$V(A_f) = q_1 q_2 (\beta_c + \beta_f)^2 \tag{7}$$

Without any correlations between loci or alleles, the sum of the total variance of an individual, $V(X)$, is just the sum of the variances over all loci, where one allele at each locus is inherited from the mother and one from the father

$$\begin{aligned} V(X) =& \sum_{i=1}^{l} (V(A_{m_i}) + V(A_{f_i})) \\ =& l q_1 q_2 [(\beta_c + \beta_m)^2 + (\beta_c + \beta_f)^2] \\ =& 2 l q_1 q_2 (\beta_c^2 + \frac{1}{2}\beta_m^2 + \frac{1}{2}\beta_f^2 + \beta_c \beta_m + \beta_c \beta_f), \end{aligned} \tag{8}$$

which is the expectation for the additive genetic variance in a randomly mating population in the presence of direct and indirect parental effects.

Through AM, correlations between alleles of mates and between loci are introduced, as shown in Chapter 4.7 of Ref.[18]. Figure 5 only displays the relevant correlations in the child. The total variance then changes due to the correlations to be

$$V(X) = \sum_{i=1}^{l} (V(A_{m_i}) + V(A_{f_i})) + \sum_{i \neq j} cov_{ij}, \tag{9}$$

where the covariance term cov_{ij} consists of the following three parts, adapted from Ref.[18]:

1. The covariance of a pair of A_m and A_f alleles at the same locus for all l pairs is given by $l f q_1 q_2 (\beta_c + \beta_m)(\beta_c + \beta_f)$, where f represents the correlation between A_m and A_f.

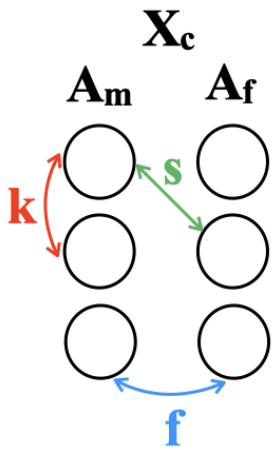

Figure 5: Correlations within and between loci introduced through AM. The child's DNA (X_c) consists of maternally inherited alleles (A_m) and paternally inherited ones (A_f). The correlation between alleles at a locus is given by f, the one between alleles at different loci within the same gamete by k and the one between pairs of A_m and A_f alleles at different loci by s.

2. The covariance between alleles at different loci, but within the same gamete, is $l(l-1)/2 \cdot kq_1q_2(\beta_c+\beta_{m/f})$, where the correlation is denoted with k. The covariance differs for maternal and paternal alleles in the parental effect $\beta_{m/f}$.

3. There are $l(l-1)$ pairs of A_m and A_f alleles at different loci with correlation s and covariance $l(l-1)sq_1q_2(\beta_c+\beta_m)(\beta_c+\beta_f)$.

These three covariance terms all act to inflate the variability, as is expected from AM. However, as AM induces a change in the mating frequencies, the first covariance term is not correct because it assumes that the correlation f acts in the same way on all genotypes, which is akin to assuming that the mating frequencies are not changed. Thus, while previous theory has established that an equilibrium is reached under AM, where there is inflated genetic variance, a larger correlation between relatives for traits driving AM and increased homozygosity at causal loci[18], the increased homozygosity is not accounted for when calculating the expected increase in genetic variance.

Our single locus model under AM presented above incorporates changes in the mating frequency under AM, which accounts for the increase in variability at one locus that is induced by f (which we denote as ρ above). To extend our model to multiple loci, we therefore only need to consider the covariances across loci. Instead of studying the correlations between a pair of alleles, we look at the correlation k between a pair of genotypes at different loci, i.e. $(A_mA_f)_i$ and $(A_mA_f)_j$. This also enables us to include

imprinting effects.

For $k = corr((A_m A_f)_i, (A_m A_f)_j)$, the covariance for $l(l-1)$ possible pairs is given by $l(l-1)k\sqrt{V_{AMi}V_{AMj}} = l(l-1)kV_{AM}$, using the variance under AM defined in either Equation 5 for trios or Equation 10 for differences in siblings. Correlations induced by AM between loci will thus inflate the variance by the factor $l(l-1)k$. Extending this to the case where the correlations between the loci are not equal and effects and allele frequencies are not the same for each locus, the covariance changes to $\sum_{i \neq j} k_{ij}\sqrt{V_{Ai}V_{Aj}}$. As a result, all statistical models which analyse the data one DNA variant at a time will yield estimates that are inflated by this "trans-LD". Only estimation of all marker effects jointly controlling for LD will result in estimates of the direct genetic effects that are unbiased of assortative mating and other forms of indirect effects.

For the sibling difference experimental design, we find the single-locus variance under assortative mating is reduced by the parental correlation, ρ:

$$V_{sAM} = 2q_1q_2(1-\rho)(\beta_c^2 + \beta_i^2 + \beta_s^2 - 2\beta_c\beta_s) \tag{10}$$

We then use Equation 3 to extend the model to multiple loci, substituting V_{RM} with V_{sAM}. The right panel of Figure 4 shows that the estimated variance for sibling differences under assortative mating using $\rho_Y = 0.2$ agrees with the expected theoretical variance for a single causal locus, as well as the multiple locus case under a range of different effect size scenarios. The indirect parental genetic effects do not directly influence the variance, although we note that with parental genetic effects the correlation within and among loci under AM alters, which changes the variance as compared to the case when parental genetic effects are absent. Also, when modelling differences in siblings, AM leads to greater difficulty in separating the contributions of each term. However, as the variances in the trio study design increase and the variances in the sibling design decrease, contrasting the estimates across within-family studies may give the upper and lower bounds of the degree to which the components influence the phenotypic variance within the human population.

Finally, we show our simulations and conclusions still hold under a more extreme assortative mating of $\rho_Y = 0.4$, as shown in Supplementary Figure S1 and S2.

Transmitted and untransmitted parental alleles

We sought to place our work in the wider context of previous research, which has focussed on transmitted and untransmitted alleles[3]. The phenotype of a child, Y_c, that is influenced by direct, indirect maternal, indirect paternal and imprinting effects across l loci, can be written in terms of transmitted (A_m, A_f) and untransmitted maternal and paternal alleles ($A_{\backslash m}$, $A_{\backslash f}$):

$$\begin{aligned} Y_c &= \sum_{j=1}^{l}(X_{c_j}\beta_{c_j} + X_{m_j}\beta_{m_j} + X_{f_j}\beta_{f_j} + X_{i_j}\beta_{i_j}) + \epsilon \\ &= \sum_{j=1}^{l}[(A_{m_j} + A_{f_j})\beta_{c_j} + (A_{m_j} + A_{\backslash m_j})\beta_{m_j} + (A_{f_j} + A_{\backslash f_j})\beta_{f_j} + (A_{m_j} - A_{f_j})\beta_{i_j}] + \epsilon \\ &= \sum_{j=1}^{l}[A_{m_j}(\beta_{c_j} + \beta_{m_j} + \beta_{i_j}) + A_{f_j}(\beta_{c_j} + \beta_{f_j} - \beta_{i_j}) + A_{\backslash m_j}\beta_{m_j} + A_{\backslash f_j}\beta_{f_j}] + \epsilon, \end{aligned} \tag{11}$$

where ϵ represents the residual error due to other environmental influences. Note that the imprinting effect will have a positive sign if only A_m is transmitted, a negative one if only A_f is transmitted, and will be equal to 0 for all the other cases. Thus, a transmitted allele has direct, indirect and imprinting effects, while an untransmitted allele only has indirect effects on a child's phenotype.

Assuming Hardy-Weinberg equilibrium and random mating, thus that the four alleles are independent, the total variance that each transmitted and untransmitted allele contributes to the trait of a child per locus j is

$$\begin{aligned} V_j &= q_{1_j}q_{2_j}[(\beta_{c_j} + \beta_{m_j} + \beta_{i_j})^2 + (\beta_{c_j} + \beta_{f_j} - \beta_{i_j})^2 + \beta_{m_j}^2 + \beta_{f_j}^2] \\ &= 2q_{1_j}q_{2_j}(\beta_{c_j}^2 + \beta_{i_j}^2 + \beta_{m_j}^2 + \beta_{f_j}^2 + \beta_{c_j}\beta_{m_j} + \beta_{c_j}\beta_{f_j} + \beta_{m_j}\beta_{i_j} - \beta_{f_j}\beta_{i_j}), \end{aligned} \tag{12}$$

where q_{1_j} and q_{2_j} are the allele frequencies. Equation 12 is equivalent to Equation 2.

When effect sizes are estimated in GWAS, a linear relationship between an individual's phenotype and their genotype, i.e. the two transmitted alleles, is assumed, ignoring imprinting effects. Thus, the SNP heritability, defined as the proportion of phenotypic variance among children attributable to the SNP genotypes of the children, would corre-

spond to

$$h^2_{SNP} = \sum_{j=1}^{l} 2q_{1_j}q_{2_j}(\beta^2_{c_j} + 0.5\beta^2_{m_j} + 0.5\beta^2_{f_j} + \beta_{c_j}\beta_{m_j} + \beta_{c_j}\beta_{f_j}), \tag{13}$$

assuming that all loci are independent. Our results suggest that in reality, it is not known how much the indirect effects contribute to the GWAS estimate unless all the direct and indirect terms are taken into account, as we do not know the size and magnitude of the covariances between direct and paternal effects, nor do we know if indirect components are actually totally estimated or estimated as part of the residual variance, nor is it clear how the abundant covariances across loci will influence the effect estimates (both cis- and trans-LD). It is also important to note that the variance due to the untransmitted alleles does not equal the total variance of the indirect effects, but only half (at most). Thus, the effects attributed to the untransmitted alleles are in fact not a good estimate of indirect parental effects, as has already been noted by Ref.[10].

Previous work[3] used polygenic risk scores computed from transmitted alleles, $\hat{\boldsymbol{g}}_T = \mathbf{A}_m\hat{\boldsymbol{\beta}}_{GWAS} + \mathbf{A}_f\hat{\boldsymbol{\beta}}_{GWAS}$, and untransmitted ones $\hat{\boldsymbol{g}}_{NT} = \mathbf{A}_{\backslash m}\hat{\boldsymbol{\beta}}_{GWAS} + \mathbf{A}_{\backslash f}\hat{\boldsymbol{\beta}}_{GWAS}$, to separate direct from indirect effects. The effects of transmitted and untransmitted alleles, θ_T and θ_{NT}, are estimated from the linear regression,

$$\mathbf{y} = \theta_T\hat{\boldsymbol{g}}_T + \theta_{NT}\hat{\boldsymbol{g}}_{NT} + \boldsymbol{\epsilon}, \tag{14}$$

where $\mathbf{y}$ is a vector of phenotypic observations and $\mathbf{y}$, $\hat{\boldsymbol{g}}_T$ and $\hat{\boldsymbol{g}}_{NT}$ are all standardized to have mean zero and variance 1. The direct effects of the transmitted polygenic risk score, δ, are then determined by calculating the difference between transmitted and untransmitted effects, $\delta = \theta_T - \theta_{NT}$. The estimated variance accounted for by the direct effect alone is then given by

$$\begin{aligned} R^2_\delta &= (\theta_T - \theta_{NT})^2/\theta^2_T \\ &= \sigma^2_C/(\sigma^2_C + 0.5 * \sigma^2_M + 0.5 * \sigma^2_F), \end{aligned} \tag{15}$$

where σ^2_C is the variance contributed by the direct effects $\sigma^2_C = \sum_{j=1}^{l} 2q_{1_j}q_{2_j}\beta^2_{c_j}$, σ^2_M is the variance contributed by the indirect maternal effects $\sigma^2_M = \sum_{j=1}^{l} 2q_{1_j}q_{2_j}\beta^2_{m_j}$, and σ^2_F is the variance contributed by the direct effects $\sigma^2_F = \sum_{j=1}^{l} 2q_{1_j}q_{2_j}\beta^2_{f_j}$.

Equation 15 makes the following assumptions: (i) $\theta_T = (\sigma^2_C + 0.5 * \sigma^2_M + 0.5 * \sigma^2_F)$ and $\theta_{NT} = (0.5 * \sigma^2_M + 0.5 * \sigma^2_F)$; (ii) loci are independent; (iii) there are no covariances among direct and indirect effects; (iv) imprinting effects are absent; (v) transmitted and untransmitted alleles have the same indirect effects; and (vi) estimation error can be ignored. Note that the estimated variance for direct effects, R^2_δ, in Equation 15 is defined differently than the direct variance, β^2_c, in Equations 2-10. R^2_δ is a relative estimate of the direct effects in the transmitted predictor, while β^2_c is an absolute value and can only be transformed to something similar to R^2_δ by making an assumption on the heritability of the transmitted alleles.

Discussion

Here, we show that there is a greater degree of potential confounding between direct genetic, indirect genetic and epigenetic effects than has been previously considered, especially under assortative mating. While previous theory has established that an equilibrium is reached under AM, where there is inflated genetic variance, a larger correlation between relatives for traits driving AM and increased homozygosity at causal loci[18,6], imprinting effects are not considered and the increased homozygosity is not accounted for when calculating the expected increase in genetic variance. Our results show that single-locus, marginal estimates common to GWAS are impossible to interpret causally, even from within-family studies. Controlling for the covariances among loci across the genome (both cis- and trans-correlations) when estimating genetic effects would require fitting all variants and all forms of genetic effect (direct, maternal, paternal, parent-of-origin) jointly. We show how covariances depend upon the direction and magnitude of the direct, maternal, paternal and imprinting effects within and across loci. These covariances contribute substantially to the population-level phenotypic variance, if parental genetic effects are present. In particular, covariance of the parental indirect genetic effects, a measure of the similarity of the indirect genetic effects among parents on their children, may be an important component of variance. We propose that assortative mating on a trait may also create assortment for parental characteristics, where biological parents cre-

ate similar environments for their children, creating shared parental effects across traits and cross-trait assortative mating[17].

The separation of imprinting and indirect genetic effects becomes more complex when one considers the underlying potential causal mechanism. Genomic imprinting has been considered as a form of parental effect in the sense that there is parental influence on the genotype-phenotype relationship in the offspring; while others suggest that because the offspring genotype accounts for all of the genetic variance in the offspring trait, imprinting is a form of direct effect as there is no variation explained by the parental genotype. Differentiating the genes in the parental genome that cause imprinting from parental genetic effects and from the genes in the offspring genome that become imprinted is difficult - if not impossible - especially if there are parental loci that control the imprinting state of other loci in the offspring. This complex phenomenon creates variation in offspring traits that is dependent on the combination of parental and offspring genotypes and it is important to recognize that: (i) patterns interpreted as indirect genetic effects can result from genomic imprinting (and vice versa), and (ii) it is not straightforward to distinguish parental effects on imprinting from imprinted genes showing parental expression; and (iii) under assortative mating, which is common for many human phenotypes, obtaining accurate estimates will be an even more difficult task.

Modelling sibling differences instead of trios removes indirect parental effects from the variance, but alters the variance of direct genetic effects in the presence of indirect sibling genetic effects. But this experimental design seems to be able to separate parent-of-origin effects under random mating. Under assortative mating, the variances in the trio study design increase and the variances in the sibling design decrease. Therefore, we suggest that contrasting the estimates across within-family studies may give the upper and lower bounds of the degree to which the components influence the phenotypic variance within the human population.

We have used the terms parent-of-origin and the epigenetic phenomenon of imprinting interchangeably, assuming that parent-of-origin effects are underlain by epigenetic mechanisms, where a parent influences the genotype-phenotype relationship of their children. However, we appreciate that other factors may create parent-of-origin effects in the popu-

lation[11]. Additionally, the ρ term used throughout this work reflects a correlation among mates at causal SNP variants (or those correlated with an underlying causal variant). A non-zero ρ term can be induced through actively choosing mates similar (or dissimilar) to oneself at specific phenotypic characters, termed direct assortative (disassortative) mating, which creates a phenotypic correlation among parents that translates to a correlation in underlying causal variants. Correlations can also appear through cultural, environmental, or geographical stratification in mating patterns that correlate with trait and trait-associated allele frequency stratification in the population. We do not discriminate between these here as they are likely inseparable in population-level data. Rather, we focus on demonstrating how correlations between mates at causal loci influence the phenotypic variance attributable to the DNA.

In summary, population-level variance in human phenotypes is shaped by potentially complex relationships between direct, indirect genetic and imprinting effects that render their separation and quantification almost impossible, especially when assortative mating occurs within the population. Currently, it is invalid to interpret either associations of untransmitted alleles causally in terms of indirect genetic effects, or parent-of-origin effects causally in terms of imprinting. Statistical modelling developments are required if we are to understand the genetic basis of human complex traits.

Acknowledgements

We thank members of the Medical Genomics group at ISTA for their comments, which improved this manuscript. This work was funded by an SNSF Eccellenza Grant to MRR (PCEGP3-181181), and by core funding from the Institute of Science and Technology Austria.

Author contributions

IK and MRR conceived and designed the study, derived the theory, conducted the analyses and wrote the paper.

Author competing interests

MRR receives research funding from Boehringer Ingelheim for work unrelated to that presented here. IK declares no competing interests.

Code availability

The code used to simulate data can be found at https://github.com/medical-genomics-group/familyMC

References

[1] Gluckman, P. D., Hanson, M. A., Cooper, C. & Thornburg, K. L. Effect of in utero and early-life conditions on adult health and disease. *New England Journal of Medicine* **359**, 61–73 (2008). PMID: 18596274.

[2] Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits.* (Sinauer, Sunderland, Mass, 1998).

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

[4] Shen, H. & Feldman, M. W. Genetic nurturing, missing heritability, and causal analysis in genetic statistics. *Proceedings of the National Academy of Sciences* **117**, 25646–25654 (2020).

[5] Eilertsen, E. M. *et al.* Direct and indirect effects of maternal, paternal, and offspring genotypes: Trio-gcta. *Behavior Genetics* **51**, 154–161 (2021).

[6] Fisher, R. A. Statistical methods in genetics. *Heredity* **6**, 1–12 (1952).

[7] Morris, T. T., Davies, N. M., Hemani, G. & Smith, G. D. Population phenomena inflate genetic associations of complex social traits. *Science Advances* **6** (2020).

[8] Brumpton, B. *et al.* Avoiding dynastic, assortative mating, and population stratification biases in mendelian randomization through within-family analyses. *Nature Communications* **11** (2020).

[9] Young, A. I. *et al.* Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nature Genetics* **54**, 897–905 (2022).

[10] Veller, C. & Coop, G. M. Interpreting population- and family-based genome-wide association studies in the presence of confounding. *PLOS Biology* **22**, 1–35 (2024).

[11] Lawson, H. A., Cheverud, J. M. & Wolf, J. B. Genomic imprinting and parent-of-origin effects on complex traits. *Nature Reviews Genetics* **14**, 609–617 (2013).

[12] Wolf, J. B. & Hager, R. A maternal–offspring coadaptation theory for the evolution of genomic imprinting. *PLOS Biology* **4**, 1–6 (2006).

[13] Hager, R., Cheverud, J. M. & Wolf, J. B. Maternal Effects as the Cause of Parent-of-Origin Effects That Mimic Genomic Imprinting. *Genetics* **178**, 1755–1762 (2008).

[14] Santure, A. W. & Spencer, H. G. Influence of Mom and Dad: Quantitative Genetic Models for Maternal Effects and Genomic Imprinting. *Genetics* **173**, 2297–2316 (2006).

[15] Wolf, J. B. & Wade, M. J. Evolutionary genetics of maternal effects. *Evolution* **70**, 827–839 (2016).

[16] Stark, A. E. A model of assortative mating. *Annals of Human Genetics* **39**, 455–464 (1976).

[17] Border, R. *et al.* Cross-trait assortative mating is widespread and inflates genetic correlation estimates. *Science* **378**, 754–761 (2022). URL `https://www.science.org/doi/abs/10.1126/science.abo2059`. `https://www.science.org/doi/pdf/10.1126/science.abo2059`.

[18] Crow, J. & Kimura, M. *An Introduction to Population Genetics Theory* (Blackburn Press, 1970).

[19] Yengo, L. *et al.* Imprint of assortative mating on the human genome. *Nature Human Behaviour* **2**, 948–954 (2018).

[20] Dimitromanolakis, A., Xu, J., Krol, A. & Briollais, L. sim1000g: a user-friendly genetic variant simulator in r for unrelated individuals and family-based designs. *BMC Bioinformatics* **20** (2019).

Supplementary Material

Confounding of indirect genetic and epigenetic effects

Ilse Krätschmer, Matthew R. Robinson

Direct, indirect and imprinting effects in trios

One locus model

We first consider a single autosomal locus with two alleles (A_1 and A_2) with frequencies q_1 and q_2 that can have the following four effects on the child's phenotype: direct (β_c), indirect maternal (β_m), indirect paternal (β_f) and parent-of-origin (imprinting, β_i). The alleles at the locus are ordered, referring first to the maternally inherited allele and second to the paternally inherited one, i.e. $A_m A_f$. The genotypes, X_c, X_m and X_f, are coded as 0, 1 and 2 for A_1A_1 homozygotes, A_1A_2 or A_2A_1 heterozygotes and A_2A_2 homozygotes, respectively. For heterozygous children, the parent-of-origin effect is positive for maternally inherited A_2 and negative for paternally inherited A_2. The possible genotypic and phenotypic values of the child, Y_c, given the genotype of the parents and imprinting are shown in Table 1, where the phenotypic values represent the deviations from the case where both parents and child have A_1A_1 genotypes.

We begin by assuming random mating and Hardy-Weinberg equilibrium, where $(q_1 + q_2) = 1 = (q_1 + q_2)^2$. The mean phenotypic value over the population is calculated as

$$\mu = \sum_j Y_{cj} z_j, \tag{16}$$

where j loops over all possible genotypes of the child, Y_c is the expected phenotype produced by a given genotype and z represents the corresponding mating frequency. The means for the different genotype classes of the child and the total mean are determined to be

$$\mu(A_1A_1) = q_1^2 q_2(\beta_m + \beta_f)\,, \tag{17}$$

$$\mu(A_1A_2) = q_1 q_2(\beta_c - \beta_i + q_2\beta_m + (1 + q_2)\beta_f)\,, \tag{18}$$

$$\mu(A_2A_1) = q_1 q_2(\beta_c + \beta_i + (1 + q_2)\beta_m + q_2\beta_f)\,, \tag{19}$$

$$\mu(A_2A_2) = q_2^2[2\beta_c + (1 + q_2)(\beta_m + \beta_f)]\,, \tag{20}$$

$$\mu = 2q_2(\beta_c + \beta_m + \beta_f)\,. \tag{21}$$

The variance is calculated as

$$V_{RM} = \sum_j Y_{cj}^2 z_j - \mu^2, \tag{22}$$

where

$$\sum_j Y_{cj}^2 z_j = 2q_2(q_2+1)(\beta_c^2+\beta_m^2+\beta_f^2) + 2q_1q_2(\beta_i^2+\beta_i\beta_m-\beta_i\beta_f) + \\ + 2q_2(3q_2+1)(\beta_c\beta_m+\beta_c\beta_f) + 8q_2^2\beta_m\beta_f, \tag{23}$$

and results in

$$V_{RM} = 2q_1q_2(\beta_c^2+\beta_m^2+\beta_f^2+\beta_i^2+\beta_c\beta_m+\beta_c\beta_f+\beta_m\beta_i-\beta_f\beta_i). \tag{24}$$

Extension to consider assortative mating

The expected phenotype of an offspring of a given genotype depends on the probability that it was produced by each of the different possible parental genotypes. Assortative mating (AM) acts to alter the mating frequencies in the population as it increases parental similarity at trait-associated loci which is reflected in the correlation, ρ, among the parents in their alleles. This changes the genotype frequencies, but does not affect the minor allele frequencies in the population. The genotype frequencies under AM are given in Ref.[16]. The means for the different genotype classes of the child and the total mean are determined to be

$$\mu(A_1A_1) = q_1q_2(1-\rho)(\beta_m+\beta_f)[q_1+\rho(\frac{1}{2}-q_1)]\,, \tag{25}$$

$$\mu(A_1A_2) = q_1q_2[\beta_c-\beta_i+q_2(1-\rho)\beta_m+(1+q_2+\rho q_1)\beta_f] \\ + \frac{1}{2}\rho q_1q_2[(1+\rho)(-\beta_c+\beta_i)+(q_1-q_2)(1-\rho)\beta_m-(3+\rho-2q_1+2\rho q_1))\beta_f], \tag{26}$$

$$\mu(A_2A_1) = q_1q_2[\beta_c+\beta_i+(1+q_2+\rho q_1)\beta_m+q_2(1-\rho)\beta_f] \\ + \frac{1}{2}\rho q_1q_2[(1+\rho)(-\beta_c-\beta_i)-(3+\rho-2q_1+2\rho q_1))\beta_m+(q_1-q_2)(1-\rho)\beta_f], \tag{27}$$

$$\mu(A_2A_2) = q_2^2[2\beta_c+(1+q_2+\rho q_1)(\beta_m+\beta_f)] \\ + \rho q_1q_2[(1+\rho)\beta_c+\frac{1}{2}(3+\rho-2q_1+2\rho q_1)(\beta_m+\beta_f)]\,, \tag{28}$$

$$\mu = 2q_2(\beta_c+\beta_m+\beta_f)\,. \tag{29}$$

The genotypic mean is the same as the one assuming that there are no correlations between the parental genotypes, which is expected as the allele frequencies do not change.

The variance is then

$$\begin{aligned} V_{AM} =& 2q_1q_2(\beta_c^2+\beta_i^2+\beta_m^2+\beta_f^2+\beta_c\beta_m+\beta_c\beta_f+\beta_m\beta_i-\beta_f\beta_i) \\ &+ q_1q_2\rho(\beta_c^2+2\beta_m^2+2\beta_f^2-\beta_i^2+4\beta_c\beta_m+4\beta_c\beta_f+4\beta_m\beta_f) \\ &+ q_1q_2\rho^2(\beta_c^2-\beta_i^2+2\beta_c\beta_m+2\beta_c\beta_f-2\beta_m\beta_i+2\beta_f\beta_i+4\beta_m\beta_f), \end{aligned} \tag{30}$$

using

$$\begin{aligned} \sum_j Y_j^2 z_j =& 2q_2(q_2+1)(\beta_c^2+\beta_m^2+\beta_f^2)+2q_1q_2(\beta_i^2+\beta_m\beta_i-\beta_f\beta_i)+ \\ &+ 2q_2(3q_2+1)(\beta_c\beta_m+\beta_c\beta_f)+8q_2^2\beta_m\beta_f+ \\ &+ q_1q_2\rho(\beta_c^2-\beta_i^2+2\beta_m^2+2\beta_f^2+4\beta_c\beta_m+4\beta_c\beta_f+4\beta_m\beta_f) \\ &+ q_1q_2\rho^2(\beta_c^2-\beta_i^2+2\beta_c\beta_m+2\beta_c\beta_f-2\beta_m\beta_i+2\beta_f\beta_i+4\beta_m\beta_f). \end{aligned} \tag{31}$$

All terms are increased under assortative mating with the exception of imprinting which is reduced. Note that there is now an additional correlation term between the parental genetic effects, $\beta_m\beta_f$, that depends on ρ.

Direct, indirect and imprinting effects in siblings

One locus model

We first extend our single autosomal locus model to the case of siblings where there are the following five effects on the child's phenotype: direct (β_c), indirect maternal (β_m), indirect paternal (β_f), parent-of-origin (imprinting, β_i) and sibling (β_s). The possible genotypic and phenotypic values of the child, Y_c, given the genotype of the parents, sibling and imprinting are shown in Table 2, where the phenotypic values represent the deviations from the case where all parents and children have A_1A_1 genotypes.

We begin by assuming random mating and Hardy-Weinberg equilibrium, where the means for the different genotype classes of the child and the total mean are determined to be

$$\mu(A_1A_1) = q_1^2q_2(\beta_m+\beta_f+\beta_s)\,, \tag{32}$$

$$\mu(A_1A_2) = q_1q_2[\beta_c-\beta_i+q_2\beta_m+(1+q_2)\beta_f+\frac{1}{2}(1+2q_2)\beta_s]\,, \tag{33}$$

$$\mu(A_2A_1) = q_1q_2[\beta_c + \beta_i + (1+q_2)\beta_m + q_2\beta_f + \frac{1}{2}(1+2q_2)\beta_s]\,, \tag{34}$$

$$\mu(A_2A_2) = q_2^2[2\beta_c + (1+q_2)(\beta_m + \beta_f + \beta_s)]\,, \tag{35}$$

$$\mu = 2q_2(\beta_c + \beta_m + \beta_f + \beta_s)\,. \tag{36}$$

The variance is calculated as

$$V_{RM} = \sum_j Y_{cj}^2 z_j - \mu^2, \tag{37}$$

where

$$\begin{aligned}\sum_j Y_{cj}^2 z_j =& 2q_2(q_2+1)(\beta_c^2 + \beta_m^2 + \beta_f^2 + \beta_s^2) + 2q_1q_2(\beta_i^2 + \beta_m\beta_i - \beta_f\beta_i) + 8q_2^2\beta_m\beta_f + \\ &+ 2q_2(3q_2+1)(\beta_c\beta_m + \beta_c\beta_f + \beta_c\beta_s + \beta_m\beta_s + \beta_f\beta_s),\end{aligned} \tag{38}$$

and results in

$$V_{RM} = 2q_1q_2(\beta_c^2+\beta_i^2+\beta_m^2+\beta_f^2+\beta_s^2+\beta_c\beta_m+\beta_c\beta_f+\beta_m\beta_i-\beta_f\beta_i+\beta_c\beta_s+\beta_m\beta_s+\beta_f\beta_s). \tag{39}$$

The mean and population-level variance depend on the mating frequency, z, of the parents. We again introduce a correlation ρ between the maternal and paternal alleles which impacts the mating frequencies, z_{AM}, as can be seen in Table 1. The means for the different genotype classes of the child and the total mean are now determined to be

$$\mu(A_1A_1) = q_1q_2(1-\rho)(\beta_m + \beta_f + \beta_s)[q_1 + \rho(\frac{1}{2} - q_1)]\,, \tag{40}$$

$$\begin{aligned}\mu(A_1A_2) =& q_1q_2[\beta_c - \beta_i + q_2(1-\rho)\beta_m + (1+q_2+\rho q_1)\beta_f + \frac{1}{2}(1+2q_2+\rho q_1-\rho q_2)\beta_s] \\ &+\frac{1}{2}\rho q_1q_2[(1+\rho)(-\beta_c+\beta_i) + (q_1-q_2)(1-\rho)\beta_m - (3+\rho-2q_1+2\rho q_1))\beta_f - 2(q_2+\rho q_1)\beta_s],\end{aligned} \tag{41}$$

$$\begin{aligned}\mu(A_2A_1) =& q_1q_2[\beta_c + \beta_i + (1+q_2+\rho q_1)\beta_m + q_2(1-\rho)\beta_f + \frac{1}{2}(1+2q_2+\rho q_1-\rho q_2)\beta_s] \\ &+\frac{1}{2}\rho q_1q_2[(1+\rho)(-\beta_c-\beta_i) - (3+\rho-2q_1+2\rho q_1))\beta_m + (q_1-q_2)(1-\rho)\beta_f - 2(q_2+\rho q_1)\beta_s],\end{aligned} \tag{42}$$

$$\begin{aligned}\mu(A_2A_2) =& q_2^2[2\beta_c + (1+q_2+\rho q_1)(\beta_m + \beta_f + \beta_s)] \\ &+ \rho q_1q_2[(1+\rho)\beta_c + \frac{1}{2}(3+\rho-2q_1+2\rho q_1)(\beta_m + \beta_f + \beta_s)]\,,\end{aligned} \tag{43}$$

$$\mu = 2q_2(\beta_c + \beta_m + \beta_f + \beta_s)\,. \tag{44}$$

The genotypic mean is the same as the one assuming that there are no correlations between the parental genotypes, which is expected as the allele frequencies do not change. The variance is then

$$
\begin{aligned}
V_{AM} &= 2q_1q_2(\beta_c^2+\beta_i^2+\beta_s^2+\beta_m^2+\beta_f^2+\beta_c\beta_m+\beta_c\beta_f+\beta_m\beta_i-\beta_f\beta_i+\beta_c\beta_s+\beta_m\beta_s+\beta_f\beta_s)\\
&+q_1q_2\rho(\beta_c^2-\beta_i^2+\beta_s^2+2\beta_m^2+2\beta_f^2+4\beta_c\beta_m+4\beta_c\beta_f+4\beta_c\beta_s+4\beta_m\beta_s+4\beta_f\beta_s+4\beta_m\beta_f)\\
&+q_1q_2\rho^2(\beta_c^2-\beta_i^2+\beta_s^2+2\beta_c\beta_m+2\beta_c\beta_f+2\beta_c\beta_s+2\beta_m\beta_s+2\beta_f\beta_s-2\beta_m\beta_i+2\beta_f\beta_i+4\beta_m\beta_f),
\end{aligned}
\tag{45}
$$

using

$$
\begin{aligned}
&\sum_j Y_j^2 z_j =\\
&= 2q_2(q_2+1)(\beta_c^2+\beta_m^2+\beta_f^2+\beta_s^2)+2q_1q_2(\beta_i^2+\beta_m\beta_i-\beta_f\beta_i)\\
&+8q_2^2\beta_m\beta_f+2q_2(3q_2+1)(\beta_c\beta_m+\beta_c\beta_f+\beta_c\beta_s+\beta_m\beta_s+\beta_f\beta_s)\\
&+q_1q_2\rho(\beta_c^2-\beta_i^2+\beta_s^2+2\beta_m^2+2\beta_f^2+4\beta_c\beta_m+4\beta_c\beta_f+4\beta_c\beta_s+4\beta_m\beta_s+4\beta_f\beta_s+4\beta_m\beta_f)\\
&+q_1q_2\rho^2(\beta_c^2-\beta_i^2+\beta_s^2+2\beta_c\beta_m+2\beta_c\beta_f+2\beta_c\beta_s+2\beta_m\beta_s+2\beta_f\beta_s-2\beta_m\beta_i+2\beta_f\beta_i+4\beta_m\beta_f).
\end{aligned}
\tag{46}
$$

This shows that indirect effects of siblings are also confounded with the direct effects, the parental genetic effects and the imprinting effects. Assortative mating acts to inflate the covariance of the indirect sibling genetic effects and all other components, making a partitioning of variance components intractable.

Modelling the phenotypic difference of siblings

Previous studies[7] have proposed to control for indirect effects by looking at the phenotypic and genetic difference in siblings. Table 2 shows the possible genotypic and phenotypic values of the child, Y_c, and sibling, Y_s, given the genotype of the parents, the respective sibling and imprinting, as well as the phenotypic difference between the offspring, $Y_c - Y_s$.

The means in phenotypic difference for the various genotype classes of the child and the total mean are determined to be

$$\mu(A_1A_1) = q_1^2q_2(-\beta_c+\beta_s)\,, \tag{47}$$

$$\mu(A_1A_2) = \frac{1}{2}q_1q_2[(1-2q_2)(\beta_c-\beta_s)-\beta_i]\,, \tag{48}$$

$$\mu(A_2A_1) = \frac{1}{2}q_1q_2[(1-2q_2)(\beta_c-\beta_s)+\beta_i]\,, \tag{49}$$

$$\mu(A_2A_2) = q_1q_2^2(\beta_c-\beta_s)\,, \tag{50}$$

$$\mu = 0\,. \tag{51}$$

The total mean is 0 as expected when calculating the difference between siblings. Thus, the variance is

$$V_{sRM} = \sum_j (Y_{cj}-Y_{sj})^2 z_j = 2q_1q_2(\beta_c^2+\beta_i^2+\beta_s^2-2\beta_c\beta_s), \tag{52}$$

which shows that indirect parental effects do indeed cancel in the difference, but indirect sibling and imprinting effects remain.

Extending the model to also include the change in genotype frequencies due to assortative mating, we find

$$\mu(A_1A_1) = q_1q_2(1-\rho)(-\beta_c+\beta_s)[q_1+\rho(\frac{1}{2}-q_1)]\,, \tag{53}$$

$$\mu(A_1A_2) = \frac{1}{2}q_1q_2(1-\rho)[(1-2q_2+\rho-2\rho q_1)(\beta_c-\beta_s)-\beta_i]\,, \tag{54}$$

$$\mu(A_2A_1) = \frac{1}{2}q_1q_2(1-\rho)[(1-2q_2+\rho-2\rho q_1)(\beta_c-\beta_s)+\beta_i]\,, \tag{55}$$

$$\mu(A_2A_2) = q_1q_2(1-\rho)(\beta_c-\beta_s)[q_2-\rho(\frac{1}{2}-q_1)]\,, \tag{56}$$

$$\mu = 0\,. \tag{57}$$

The variance is reduced by the parental correlation, ρ,

$$V_{sAM} = \sum_j (Y_{cj}-Y_{sj})^2 z_{AMj} = 2q_1q_2(1-\rho)(\beta_c^2+\beta_i^2+\beta_s^2-2\beta_c\beta_s) \tag{58}$$

Forward in time simulation study

A simple, but realistic simulation (based loosely on Ref.[20]) for parents and children is set up and moved forward in time for a number of generations, *ngen*, to simulate assortative mating. Genotypes of unrelated individuals in generation 0 are created with a realistic LD based on randomly selected number of SNPs of chromosome 4. The data of chromosome 4 is taken from the 1000 genomes project (`http://ftp.1000genomes.`

ebi.ac.uk/vol1/ftp/release/20130502/) using vcfrandomsample (https://github.com/vcflib/vcflib#vcflib) to downsample the data to make it more manageable. In the end, 100 variants are selected for the simulation, out of which only a fixed number, l, are set to have an effect. In order to avoid inbreeding when moving forward in time, a high number of individuals, n, is required to be generated. The number of individuals is kept constant across generations. Individuals in generation 0 are separated into g groups and mated randomly within a group. They produce two offspring by recombining the parents' haplotypes based on a genetic map (release b37, for example from https://github.com/odelaneau/shapeit5/tree/main/resources/maps/b37). The offspring of generation 0 become the parents for the next generation and mate with parents from a different group. These parents are then merged into one group whose offspring is then used in the next generation to mate with yet another group. This way, inbreeding is avoided.

To simulate assortative mating, starting from generation 1, the parents are no longer mated randomly within each group, but mate based on their ordered phenotypes. The phenotypes are calculated as:

$$\boldsymbol{Y} = \boldsymbol{X_c}\boldsymbol{\beta_c} + \boldsymbol{X_m}\boldsymbol{\beta_m} + \boldsymbol{X_f}\boldsymbol{\beta_f} + \boldsymbol{X_i}\boldsymbol{\beta_i} + \boldsymbol{X_s}\boldsymbol{\beta_s} + \boldsymbol{\epsilon} \tag{59}$$

where ϵ is a value drawn from a Normal distribution with mean 0 and standard deviation of 0.1 to facilitate ordering. The phenotypes are ordered to introduce a correlation, ρ, between the parents to simulate AM. This procedure is a simplified version of the unknown realistic process and has some limitations in particular for the single locus case, where the parental phenotpyes can only have a limited number of values. For the implementation of the AM procedure, see Code availability.

For the simulation results presented here the simulation is moved forward in time for $ngen = 10$ generations. This gave, $g = 2^{ngen-1} = 512$ groups with 2000 individuals within each group. Effect sizes for either $l = 1$ or 5 are set depending on simulation the simulation scenario presented in each figure. The correlation between the parents' phenotypes is either set to $\rho = 0.4$ or 0.2. Ten simulations are generated, where the causal loci are fixed. The average and standard deviations of the ten simulations are calculated

and shown as estimated variances in Figures 2, 4 and S2. The standard deviations are negligible as the simulation process is not stochastic and the same causal loci with the same effect size are used for each simulation.

Supplementary Figures

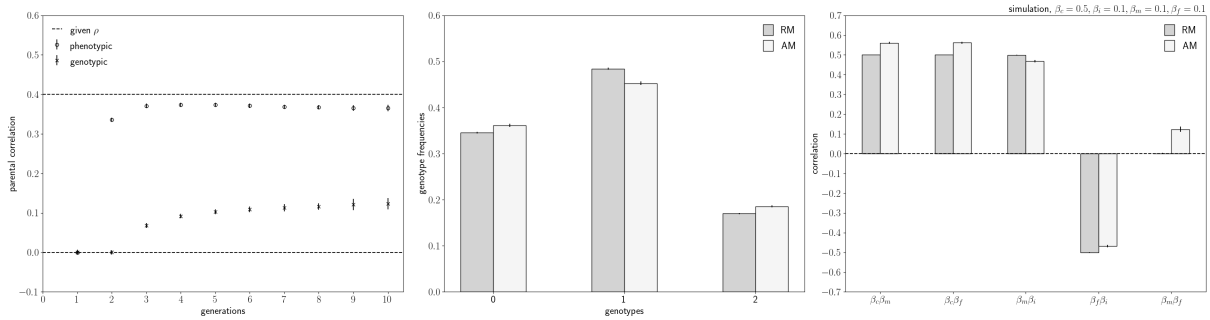

Figure S1: Simulation study for a single locus with effects $\beta_c = 0.5, \beta_i = 0.1, \beta_m = 0.1, \beta_f = 0.1$: (Left) Parental phenotypic and genotypic correlation as function of generations compared to the given correlation between parental phenotypes, $\rho_Y = 0.4$, fixed in the simulation parameters. (Middle) Genotype frequencies under random mating (RM) and assortative mating (AM). (Right) Correlations between effects under RM and AM.

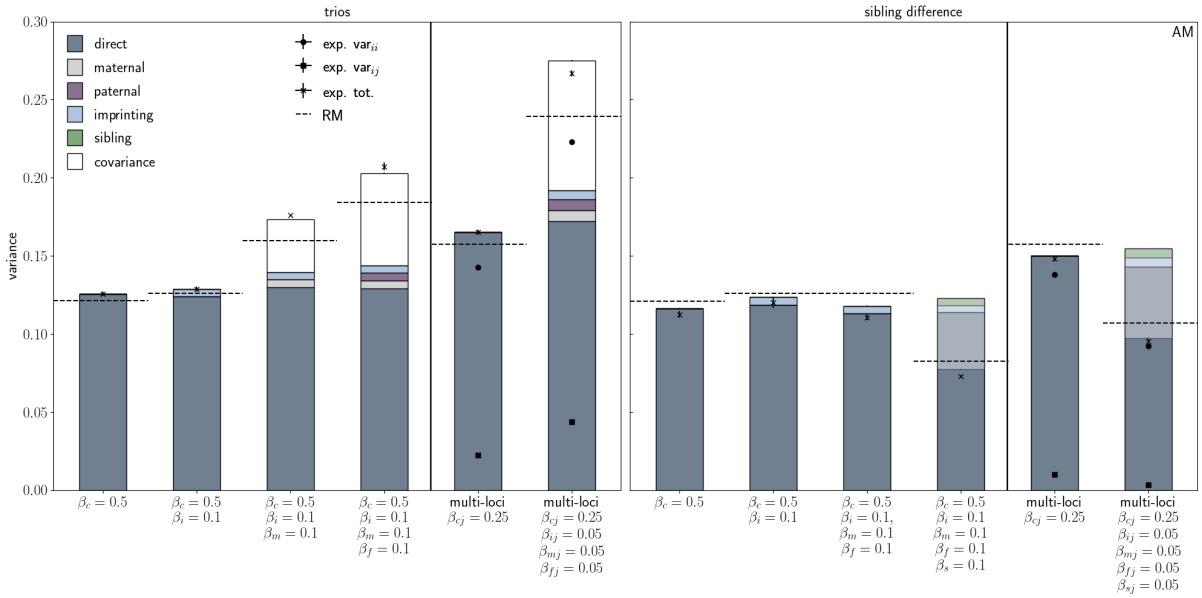

Figure S2: Expected theoretical total variance (exp. tot.), sum of variances in case of multiple loci (exp. var_{ii}), covariance between loci (exp. var_{ij}) and estimated variance (bars) for trios (left) and sibling differences (right) under assortative mating ($\rho_Y = 0.4$) when considering various effects (given as labels in x-axis) for the single-locus case and for 5 causal loci (multi-loci) with each having the effects $\beta_c = 0.25, \beta_m = 0.05, \beta_f = 0.05, \beta_i = 0.05$, and in case of the sibling difference, $\beta_s = 0.05$. Note that the direct-sibling covariance is negative and is thus shown opaque, covering parts of the variance. The corresponding values under RM are shown as dashed lines.