nature genetics

SUPPLEMENTARY INFORMATION

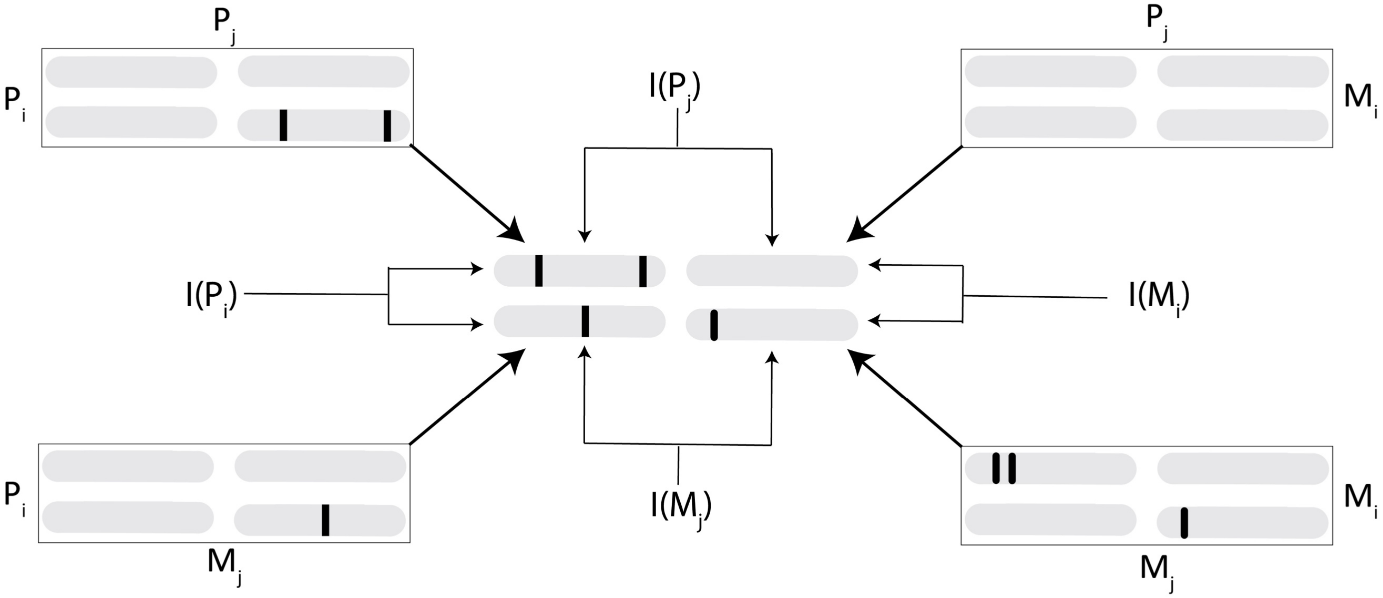
https://doi.org/10.1038/s41588-018-0178-9

In the format provided by the authors and unedited.

Relatedness disequilibrium regression estimates heritability without environmental bias

Alexander I. Young[1,2,3*], Michael L. Frigge[1], Daniel F. Gudbjartsson[1,4], Gudmar Thorleifsson[1], Gyda Bjornsdottir[1], Patrick Sulem[1], Gisli Masson[1], Unnur Thorsteinsdottir[1,5], Kari Stefansson[1,5] and Augustine Kong[1,3,4*]

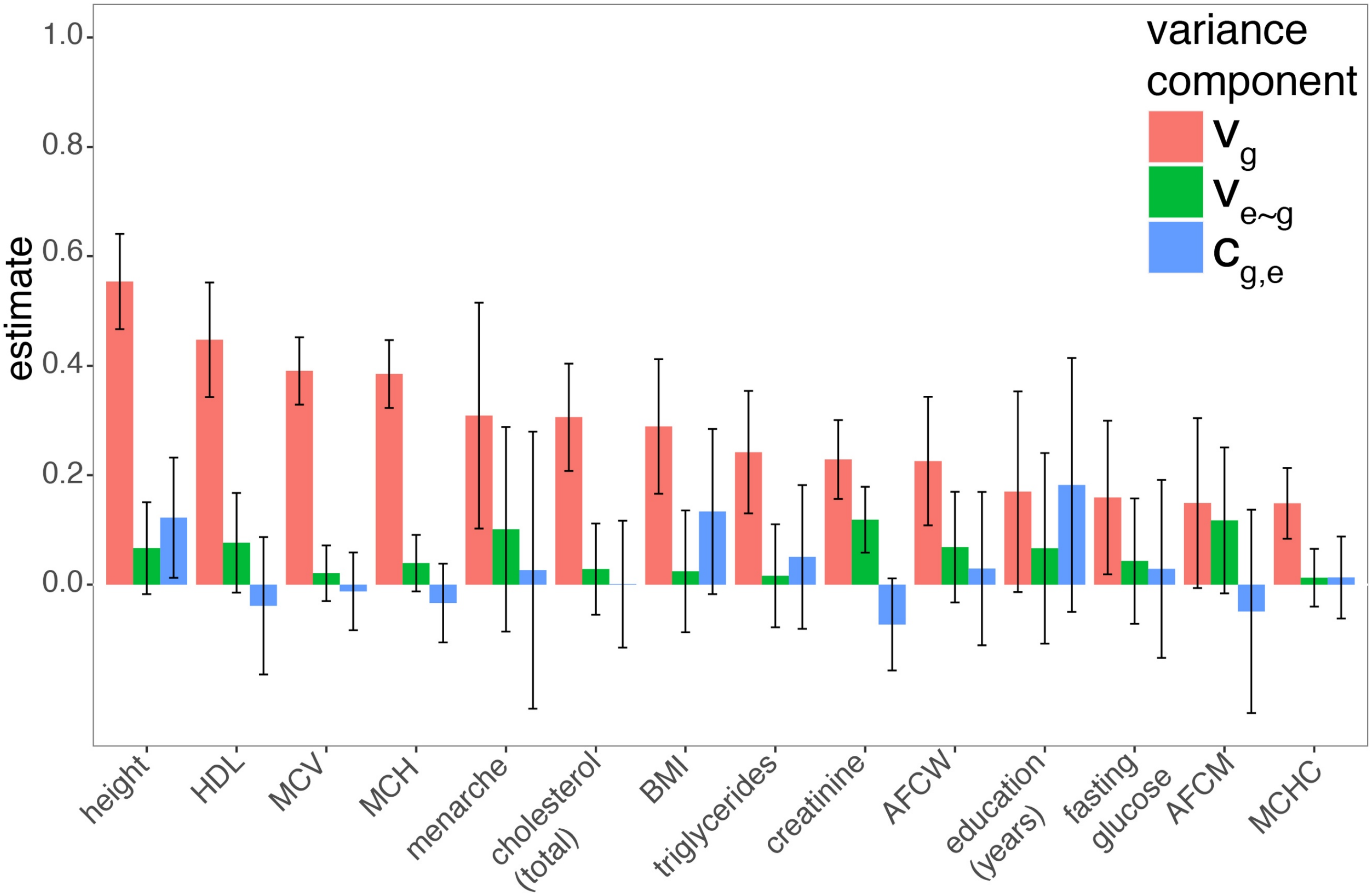
[1]deCODE genetics/Amgen Inc., Reykjavik, Iceland. [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [3]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK. [4]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. [5]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. *e-mail: alexander.young@bdi.ox.ac.uk; augustine.kong@bdi.ox.ac.uk



Supplementary Figure 1

Determination of offspring relatedness

The diagram shows how the identity-by-descent sharing states of two individuals ***i*** and ***j*** are determined by the identity-by-descent sharing states of their parents and the segregation events in the parents during meiosis. The identity-by-descent sharing states of ***i*** and ***j*** are represented by the four chromosomes in the centre, with black bands indicating regions shared identical-by-descent. The four chromosomes represent the four possible pairs of homologous chromosomes (maternal-maternal, paternal-maternal, maternal-paternal, and paternal-paternal): the identity-by-descent sharing between the chromosome inherited from ***i*** 's father, P_i, and ***j*** 's mother, M_j, etc. The identity-by-descent sharing states of the four possible pairs of parents, one from each individual, are shown in the corners (P_i and P_j, P_j and $M_{i,}$ P_j and M_i, and M_j and M_i). The segregation event in ***i*** 's father is represented by $I(P_i)$, the segregation event in ***j*** 's mother represented by $I(M_j)$, etc. Note that for simplicity we ignore recombination in this diagram. See the Relatedness Disequilibrium Lemma in the Supplementary Note for a mathematical description of this process and its consequences.



Supplementary Figure 2

RDR variance component estimates

Estimated variance components of the RDR covariance model for 14 quantitative traits in Iceland (Supplementary Table 4), expressed as a % of phenotypic variance, shown with intervals +/- 1.96 standard errors around the estimate. Trait abbreviations: BMI, body mass index; AFCW, age at first child in women; AFCM, age at first child in men; education (years), educational attainment (years); HDL, high density lipoprotein; MCH, mean cell haemoglobin; MCHC, mean cell heamoglobin concentration; MCV, mean cell volume.

	$\mathbb{E}[\hat{h}^2]$		$\mathrm{SD}[\hat{h}^2]$		Est. $\mathrm{SE}[\hat{h}^2]$
Trait	Max Lik.	L.Sq.	Max Lik.	L.Sq.	Max Lik.
additive	39.3 (0.62)	39.8 (0.65)	13.81	14.51	14.10
regional	38.3 (0.6)	40 (0.71)	13.50	15.60	13.16
maternal environment	38.9 (0.58)	37.6 (0.73)	13.03	15.80	12.42
genetic nurturing	39.4 (0.49)	39.8 (0.68)	10.99	15.28	10.58

Supplementary Table 1: **Comparison of maximum likelihood and least-squares approaches to fitting RDR variance components.** For each of four simulated traits based upon actual genetic data in Iceland, we give the mean estimate, along with its standard error in brackets, of the heritability, $\hat{h}^2$, from fitting the RDR covariance model by both (restricted) maximum likelihood (Max Lik.) and by least-squares regression (L.Sq.). The heritability estimates are expressed as a percentage of the phenotypic variance, and the true heritability was 40% for all the traits. We give the standard deviation of $\hat{h}^2$ from the two methods over the 500 independent replicates of each trait ($\mathrm{SD}[\hat{h}^2]$), and the average estimated standard error of $\hat{h}^2$ across 500 replicates, 'Est. $\mathrm{SE}[\hat{h}^2]$', for the maximum likelihood method. The regression method removes parent-offspring and grandparent-grandchild pairs from the analysis, whereas the maximum likelihood method considers all pairs.

	RDR			Kinship	Kinship F.E.	
	h^2	$v_{e\sim g}$	$c_{g,e}$	h^2	h^2	v_c
additive	39.3 (0.62)	-0.3 (0.49)	0.8 (0.71)	40.4 (0.15)	40.5 (0.18)	-0.1 (0.14)
genetic nurturing	39.4 (0.49)	9.4 (0.4)	29.2 (0.57)	92.7 (0.094)	82.8 (0.14)	9.9 (0.11)
maternal	38.9 (0.58)	47.5 (0.48)	-15.9 (0.67)	76.3 (0.17)	39.9 (0.18)	39.9 (0.11)
regional	38.3 (0.6)	2.1 (0.48)	12.2 (0.7)	59 (0.17)	58.3 (0.2)	0.7 (0.13)
rare SNPs	35.0 (0.64)	-0.8 (0.47)	3.4 (0.69)	39.5 (0.15)	39.4 (0.19)	0.1 (0.14)
epistatic	41.3 (0.6)	0.7 (0.5)	1.1 (0.71)	44.2 (0.16)	43.3 (0.19)	1.1 (0.13)
dominance	40.5 (0.63)	1.5 (0.52)	-0.1 (0.74)	42.7 (0.15)	41.1 (0.19)	2.0 (0.13)

Supplementary Table 2: **Variance components from simulations.** The mean variance component estimates, expressed as a % of the phenotypic variance, from the RDR, Kinship, and Kinship F.E. methods. For the Kinship F.E. method, the v_c column gives the estimate of the variance explained by shared family environment. We determined whether individuals shared a family environment by whether they shared a mother according to the deCODE Genealogy database. We simulated 500 replicates of each trait based on actual Icelandic genetic data for 10,000 individuals. Ten thousand SNPs with median frequency 23% were given additive effects for all the traits other than the rare SNPs trait, for which 2,200 SNPs with frequency between 0.1% and 1% (median 0.26%) were used. The true (narrow-sense) heritability of each trait was 40%. To this additive genetic component, only noise was added for the additive trait and the rare SNPs trait. For the epistatic trait, 10% of the phenotypic variance was due to pairwise interactions between SNPs. For the dominance trait, 10% of the phenotypic variance was due to dominance effects. For the other traits, effects representing different sources of environmental confounding were added in addition to noise and the additive genetic component. For the regional trait, each region of Iceland (sysla) was given an effect; for the 'maternal environment' trait, an environmental effect shared between those who share mothers was added. For the 'genetic nurturing' trait, the genotypes of the parents were also given effects to simulate 'genetic nurturing' effects[14], so that true $v_{e\sim g} = 10\%$, and true $c_{g,e} \approx 28.3\%$. For the 'regional' trait, the Kinship and Kinship F.E. methods also included adjustment for 20 genetic principal components.

	True	RDR			Kinship	Sib-Regression
	h^2	h^2	$v_{e\sim g}$	$c_{g,e}$	h^2	h^2
additive	0	0.7 (0.5)	0.1 (0.5)	0.7 (0.7)	0.7 (0.1)	0.1 (0.9)
additive	80	79.3 (0.37)	-2.1 (0.36)	2.6 (0.46)	80.4 (0.1)	83.1 (0.63)
genetic nurturing	15	14.9 (0.3)	14.2 (0.29)	22.1 (0.38)	72.7 (0.08)	15.4 (0.58)

Supplementary Table 3: **Additional simulations with different variance components.** The mean variance component estimates, expressed as a % of the phenotypic variance, from the RDR, Kinship, and Sib-Regression methods. The ‘additive’ trait was determined by additive, direct genetic effects and noise. For the genetic nurturing trait, in addition to additive, direct genetic effects explaining 15% of the variance, each genetic variant in the parents was also given an effect on the proband, a ‘parental genetic nurturing effect’[14]. The parental genetic nurturing effect of each variant differed only by a constant scale factor from the direct effect of the genetic variant in the offspring, so that $v_{e\sim g} = 15$, and $c_{g,e} \approx 21$.

		Kinship	Kinship F.E.		RDR			LR
	n	h^2	h^2	v_c	h^2	$v_{e\sim g}$	$c_{e,g}$	p
BMI	19589	54.8 (1.7)	46.7 (2.5)	6.6 (1.5)	28.9 (6.3)	2.4 (5.7)	13.4 (7.7)	6.6e-07
height	21802	83.6 (1.2)	78 (1.9)	4.9 (1.2)	55.4 (4.4)	6.7 (4.3)	12.2 (5.6)	1.6e-17
AFCW	22367	34.6 (1.6)	33.5 (2.1)	2.6 (1.3)	22.6 (6)	6.8 (5.2)	2.9 (7.1)	1.1e-04
AFCM	17117	21.6 (1.9)	16.3 (2.6)	5.4 (1.7)	14.9 (7.9)	11.8 (6.8)	-4.9 (9.5)	4.0e-03
menarche	11242	51.1 (2.6)	41.9 (4)	7.6 (2.4)	30.9 (10.5)	10.1 (9.5)	2.6 (12.9)	5.6e-03
education (years)	12035	54.6 (2.2)	52.4 (3.7)	3.3 (2.1)	17 (9.4)	6.6 (8.9)	18.2 (11.8)	7.8e-09
total chol.	27320	35 (1.4)	32.2 (1.8)	2.7 (1.1)	30.6 (5)	2.8 (4.3)	0.1 (5.9)	2.5e-01
HDL	24570	49.6 (1.5)	45.1 (2.1)	3.9 (1.2)	44.8 (5.3)	7.7 (4.6)	-3.9 (6.4)	2.0e-02
triglycerides	24099	34.7 (1.5)	29.8 (2)	4.7 (1.3)	24.2 (5.7)	1.6 (4.8)	5.1 (6.7)	3.0e-02
fasting glucose	19500	25.7 (1.8)	23.6 (2.3)	3 (1.6)	15.9 (7.2)	4.3 (5.8)	2.9 (8.3)	3.1e-02
creatinine	38929	27.7 (1.1)	22.2 (1.3)	7 (0.9)	22.9 (3.7)	11.9 (3.1)	-7.3 (4.3)	3.3e-08
MCH	43917	38.5 (1)	36.8 (1.2)	2 (0.7)	38.5 (3.2)	3.9 (2.6)	-3.4 (3.7)	2.0e-01
MCHC	43963	18.4 (0.9)	18.4 (1.1)	0.5 (0.8)	14.9 (3.3)	1.3 (2.7)	1.3 (3.8)	1.5e-01
MCV	43919	40.1 (1)	38.5 (1.2)	1.7 (0.7)	39.1 (3.1)	2.1 (2.6)	-1.2 (3.6)	4.9e-01

Supplementary Table 4: **Variance component estimates from RDR, Kinship and Kinship F.E. methods**. For each trait, the sample size used is given under ‘n’. Each variance component estimate is expressed as a percentage of the phenotypic variance and is followed by its standard error in brackets. For the Kinship F.E. method, the v_c column gives the estimate of the variance explained by shared family environment. We determined whether individuals shared a family environment by whether they shared a mother according to the deCODE Genealogy database. We give the p-value from the likelihood ratio test comparing the RDR covariance model to the model without the parental and parent-offspring relatedness matrices under the ‘LR’ column. Note that all of these estimates are from the exact same Icelandic samples with both parents genotyped, so any differences in heritability estimates are due to methodological differences. Samples were restricted to those born between 1951 and 1997 for BMI and traits measured from blood, and samples were restricted to those born between 1951 and 1995 for height. Trait abbreviations: BMI, body mass index; AFCW, age at first child in women; AFCM, age at first child in men; menarche, age at menarche (years); education, educational attainment (years); total chol., total cholesterol; HDL, high density lipoprotein; glucose, fasting glucose; MCH, mean cell haemoglobin; MCHC, mean cell heamoglobin concentration; MCV, mean cell volume.

		Kinship F.E.			RDR (Max. Lik)		RDR (L.Sq.)
	n	BPG	BPG-sysla	random	BPG	BPG-sysla	BPG
BMI	19589	46.7 (2.5)	46.3 (2.5)	48.9 (3.5)	28.9 (6.3)	28.7 (6.3)	33.2 (8.3)
height	21802	78.0 (1.9)	78.1 (1.9)	90.4 (2.4)	55.4 (4.4)	55.8 (4.5)	44.8 (9.0)
AFCW	22367	33.5 (2.1)	31.7 (2.1)	29.0 (1.7)	22.6 (6.0)	23.4 (6.0)	25.1 (7.2)
AFCM	17117	16.3 (2.6)	15.5 (2.6)	21.1 (2.1)	14.9 (7.9)	14.9 (7.9)	16.5 (8.8)
menarche	11242	41.9 (4.0)	41.2 (4.0)	40.1 (2.8)	30.9 (10.5)	31.0 (10.6)	33.4 (13.5)
education	12035	52.4 (3.7)	50.6 (3.8)	45.5 (2.2)	17.0 (9.4)	15.7 (9.5)	9.2 (13.5)
total chol.	27320	32.2 (1.8)	32.1 (1.8)	29.7 (2.4)	30.6 (5.0)	30.7 (5.0)	32.4 (5.7)
HDL	24570	45.1 (2.1)	45.0 (2.1)	46.3 (2.7)	44.8 (5.3)	44.1 (5.3)	46.5 (6.6)
triglycerides	24099	29.8 (2.0)	29.6 (2.0)	33.8 (2.7)	24.2 (5.7)	24.1 (5.7)	23.4 (6.5)
fasting glucose	19500	23.6 (2.3)	22.9 (2.3)	23.2 (3.0)	15.9 (7.2)	15.4 (7.2)	15.6 (7.8)
creatinine	38929	22.2 (1.3)	22.1 (1.3)	22.8 (1.7)	22.9 (3.7)	22.6 (3.7)	23.6 (4.1)
MCH	43917	36.8 (1.2)	36.7 (1.2)	40.6 (1.6)	38.5 (3.2)	38.5 (3.2)	38.7 (3.8)
MCHC	43963	18.4 (1.1)	17.9 (1.1)	19.8 (1.6)	14.9 (3.3)	14.9 (3.3)	14.3 (3.5)
MCV	43919	38.5 (1.2)	38.5 (1.2)	40.7 (1.6)	39.1 (3.1)	39.1 (3.1)	39.3 (3.8)

Supplementary Table 5: **Robustness of heritability esitmates.** Here we show RDR and Kinship F.E. heritability estimates from the sample with both parents genotyped (BPG), which is the sample used for the main results in the Main Text; heritability estimates from the sample with both parents genotyped after adjusting for mean differences between different regions (syslas) of Iceland (BPG-sysla); and Kinship F.E. heritability estimates from a random sample from all of the genotyped Icelandic individuals (random). The random sample was chosen to be of the same size as the sample with both parents genotyped, 54,888. We also show RDR estimates when estimated by least-squares regression of the sample phenotypic covariance matrix on the the elements of the three relatedness matrices, excluding parent-offspring and grandparent-grandchild pairs (Methods). We estimated the standard errors for the least-squares estimator using a procedure that takes into account dependence between pairs (Supplementary Note).

trait	study	twin pairs	ages	estimate	method
BMI	Carlsson[36]	16,732	58.7 (mean)	65 (3.8)	ACE
height	Silventoinen[37]	8,747	20-40	81	ACE
menarche	Baker[38]	756	16-17	75 (6.9)	ACE
education	Branigan[23]	32,814	NA	43.2 (3.6)	ACE
total chol.	Rahman[39]	9,066	66.2 (mean)	57 (3.8)	$2(r_{\mathrm{MZ}} - r_{\mathrm{DZ}})$
HDL	Rahman[39]	9,086	66.2 (mean)	69 (3.1)	$2(r_{\mathrm{MZ}} - r_{\mathrm{DZ}})$
triglycerides	Rahman[39]	9,072	66.2 (mean)	61(3.7)	$2(r_{\mathrm{MZ}} - r_{\mathrm{DZ}})$
fasting glucose	Rahman[39]	8,908	66.2 (mean)	59 (4.0)	$2(r_{\mathrm{MZ}} - r_{\mathrm{DZ}})$
creatinine	Arpegard[40]	12,313	64.9 (mean)	59 (1.5)	ADE

Supplementary Table 6: **Summary of twin studies estimates used.** Estimates are given as percentages of phenotypic variance along with standard errors in brackets. Estimates were taken from published studies on the Swedish Twin Registry[19], apart from education, where the estimate is from a meta-analysis of Scandinavian countries, including Sweden[23]. We took the published estimate from the ACE (additive-common-environment) model where available. If ACE estimates were not provided, we calculated a moment based esimtate of the heritability from the ACE model using published twin correlations (Materials and Methods). Method abbreviations: ACE, the additive-common-environment twin model, fit by maximum likelihood; $2(r_{\mathrm{MZ}} - r_{\mathrm{DZ}})$, the moment estimator of the heritability from the ACE model, where r_{MZ} is the correlation between monozygotic (MZ) twins, and r_{DZ} is the correlation between dizygotic (DZ) twins; ADE, the additive-dominance-environment model, fit by maximum likelihood.

	RDR-SNP		RELT-SNP		GREML-SNP	
	genome	causal	genome	causal	genome	causal
additive	47.1 (0.27)	40.1 (0.12)	48.0 (0.20)	39.8 (0.09)	43.7 (0.11)	40.0 (0.06)
genetic nurturing	45.2 (0.21)	40.1 (0.07)	93.0 (0.32)	74.1 (0.14)	89.2 (0.09)	73.1 (0.05)
maternal	47.2 (0.26)	40.0 (0.12)	47.9 (0.20)	39.9 (0.10)	63.1 (0.13)	44.8 (0.07)
regional	47.4 (0.26)	40.1 (0.11)	46.8 (0.19)	39.3 (0.10)	54.4 (0.13)	43.0 (0.06)
rare	12.0 (0.27)	40.3 (0.08)	13.5 (0.16)	75.1 (0.19)	24.4 (0.12)	40.3 (0.06)
epistatic	47.1 (0.27)	40.1 (0.11)	48.0 (0.21)	39.9 (0.10)	45.6 (0.12)	40.6 (0.07)
dominance	47.4 (0.27)	40.3 (0.12)	48.5 (0.21)	40.3 (0.10)	45.2 (0.11)	40.6 (0.07)

Supplementary Table 7: **Variance components from simulations for SNP based methods.** The mean variance component estimates, expressed as a % of the phenotypic variance, from the RDR-SNP, RELT-SNP, and GREML-SNP methods. Here, the GREML-SNP method, using restricted maximum likelihood, was applied to the full sample without pruning of relative pairs. We provide GREML-SNP estimates here purely as a point of comparison. We simulated 500 replicates of each trait based on Icelandic genetic data for 10,000 individuals. Ten thousand SNPs with median frequency 23% were given additive effects for all the traits other than the rare SNPs trait, for which 2,200 SNPs with frequency between 0.1% and 1% (median 0.26%) were used. The true (narrow-sense) heritability of each trait was 40%. We provide two different estimates for each method: estimates from using $\sim 600,000$ genome-wide SNPs typically found on Illumina genotyping arrays (genome), and estimates from using the causal SNPs for the simulated traits (causal). For the 'regional' trait, the RELT-SNP and GREML-SNP methods included adjustment for 20 genetic principal components.

		% var	RELT-SNP	RELT-SNP	RDR-SNP			LR
	n	20 PCs	h^2 (0 PCs)	h^2 (20 PCs)	h^2	$v_{e\sim g}$	$c_{e,g}$	p
BMI	19589	0.46	36.1 (3.4)	31.8 (3.2)	34.2 (2.9)	10.3 (2.8)	1.2 (3.6)	1.6e-14
height	21802	1.17	67.9 (4.7)	55.2 (4.4)	44.5 (2.3)	9.6 (2.2)	14.3 (2.7)	4.0e-59
AFCW	22367	1.29	27.7 (2.5)	20.1 (2.3)	11.7 (2.6)	11.1 (2.6)	2.9 (3.3)	2.3e-26
AFCM	17117	1.01	19.7 (2.5)	12.3 (2.2)	11.5 (3.4)	6.6 (3.2)	0.7 (4.2)	6.0e-06
menarche	11242	0.28	33.9 (4.2)	29.9 (4.1)	26.8 (5.0)	6.5 (4.7)	6.3 (6.2)	5.9e-06
education (years)	12035	2.89	46.2 (4.9)	29.2 (4.4)	17.3 (4.4)	14.8 (4.4)	8.8 (5.6)	6.1e-26
total chol.	27320	0.16	24.2 (2.2)	22.1 (2.1)	23.5 (2.3)	5.7 (2.1)	-0.5 (2.7)	1.0e-06
HDL	24570	0.52	29.7 (2.7)	24.2 (2.5)	32.0 (2.5)	8.3 (2.2)	0.5 (2.9)	3.3e-13
triglycerides	24099	0.42	25.8 (2.4)	22.1 (2.2)	23.8 (2.6)	4.7 (2.3)	1.0 (3.0)	1.6e-05
fasting glucose	19500	0.37	16.8 (2.3)	11.3 (2.1)	15.8 (3.1)	8.5 (2.9)	-2.6 (3.8)	3.0e-06
creatinine	38929	0.10	17.2 (1.6)	16.0 (1.5)	16.9 (1.6)	8.1 (1.5)	-1.9 (2.0)	1.7e-17
MCH	43917	0.10	28.7 (1.9)	27.5 (1.9)	29.3 (1.5)	3.7 (1.3)	0.5 (1.7)	1.4e-07
MCHC	43963	0.14	13.0 (1.2)	11.5 (1.2)	12.5 (1.5)	2.3 (1.3)	0.4 (1.7)	3.4e-04
MCV	43919	0.07	29.8 (2.0)	29.2 (2.0)	31.1 (1.5)	5.9 (1.3)	-1.9 (1.7)	1.1e-10

Supplementary Table 8: **Variance component estimates from RDR-SNP and RELT-SNP**. For each trait, the sample size used is given under 'n'. Each variance component estimate is expressed as a percentage of the phenotypic variance and is followed by its standard error in brackets. We give the RELT-SNP estimates without adjustment for PCs (0 PCs) and with adjustment for 20 PCs. We also give the percentage of trait variance explained by regression on the top 20 PCs under '% var 20 PCs'. We give the p-value from the likelihood ratio test comparing the RDR-SNP covariance model to the model without the parental and parent-offspring relatedness matrices under the 'LR' column. Samples were restricted to those born between 1951 and 1997 for BMI and traits measured from blood, and samples were restricted to those born between 1951 and 1995 for height. Trait abbreviations: BMI, body mass index; AFCW, age at first child in women; AFCM, age at first child in men; menarche, age at menarche (years); education, educational attainment (years); total chol., total cholesterol; HDL, high density lipoprotein; glucose, fasting glucose; MCH, mean cell haemoglobin; MCHC, mean cell heamoglobin concentration; MCV, mean cell volume.

	Sim. S.D.	S.E. Est	% error
additive	4.40	4.57	3.40
genetic nurturing	7.10	7.26	2.66
maternal	4.60	5.24	13.69
regional	4.30	4.96	14.12
rare	3.70	3.68	-0.22
epistatic	4.60	4.65	0.40
dominance	4.80	4.63	-3.91

Supplementary Table 9: **Accuracy of standard error estimates for RELT-SNP.** For the simulated traits, we compare the standard deviation of the simulation estimates (Sim S.D.) over 500 replicates to the mean estimated standard error (S.E. Est). The traits are as described in the Methods. The standard deviations and standard errors are expressed as a percentage of the phenotypic variance. Under '% error', we give the error as a percentage of the standard deviation of the simulation estimates. The mean error across the simulated traits was 4.3%. We used a procedure for estimating the standard error of the RELT-SNP estimates that takes into account the non-independence of different pairs of phenotype observations (Supplementary Note).

Supplementary Note for: ‘Estimating heritability without environmental bias’

Alexander I. Young

Contents

1 RDR Theory

1.1 Variance decomposition

Consider a phenotype, Y, variation in which is determined by the direct effect of the number of copies of an allele at a single locus, g, and an environmental effect, e. Assuming additivity, the phenotype of the i^{th} individual is

$$Y_i = \mu + \delta g_i + e_i, \tag{1}$$

where μ is some finite constant. The univariate regression estimate of the direct effect δ, $\hat{\delta}$, will be biased if g is correlated with e:

$$\mathbb{E}[\hat{\delta}] = \delta + \frac{\text{Cov}(g, e)}{\text{Var}(g)}. \tag{2}$$

This is the basic problem of genetic association testing: inherited genetic variants are often correlated with environmental influences on a trait, due to family effects and population stratification.

It is, however, possible to take advantage of the random nature of segregation to isolate the effects of genetic inheritance from environmental effects. This relies on the fact that the genotype of the offspring is determined by both the genotype of the parents and the segregation events that occurred during meiosis. One can assume that the segregation events in the parents of i that produced the genome of i are independent of the environmental effects on the phenotype of i. We can write g in a way that reflects segregation:

$$g_i = \text{I}_{\text{pp}}(i) g^{\text{p}}_{\text{p}(i)} + \text{I}_{\text{pm}}(i) g^{\text{m}}_{\text{p}(i)} + \text{I}_{\text{mm}}(i) g^{\text{m}}_{\text{m}(i)} + \text{I}_{\text{mp}}(i) g^{\text{p}}_{\text{m}(i)}, \tag{3}$$

where $g^{\text{p}}_{\text{p}(i)}$ is the paternally inherited binary genotype of the father of i, $g^{\text{m}}_{\text{p}(i)}$ is the maternally inherited binary genotype of the father of i, and $g^{\text{m}}_{\text{m}(i)}$ and $g^{\text{p}}_{\text{m}(i)}$ are the equivalents for the mother; $\text{I}_{\text{pp}}(i)$ is the indicator variable for whether the paternal variant of i was passed down the patrilineage, and $\text{I}_{\text{mp}}(i)$ is the indicator for whether the maternal variant of i was passed down from i's maternal grandfather. Note that $\text{I}_{\text{mp}}(i) = 1 - \text{I}_{\text{mm}}(i)$, since the maternal variant of i was either inherited from i's maternal grandfather or grandmother. One can assume that $\text{I}_{\text{pp}}(i)$ and $\text{I}_{\text{mp}}(i)$ are Bernoulli(0.5) variables, independent of e. Therefore,

$$g_i \perp e_i \mid g^{\text{p}}_{\text{p}(i)}, g^{\text{m}}_{\text{p}(i)}, g^{\text{m}}_{\text{m}(i)}, g^{\text{p}}_{\text{m}(i)}. \tag{4}$$

In other words, the genotype of the child is conditionally independent of environmental effects on the child given the genotypes of the parents of the child.

Any dependence between the inherited genetic variants and environmental effects flows through the parental genotypes. Furthermore, because the expectation of the offspring genotype conditional on the parental genotypes is a linear function of the parental genotypes,

$$\mathbb{E}[g_i|g^{\text{p}}_{\text{p}(i)}, g^{\text{m}}_{\text{p}(i)}, g^{\text{m}}_{\text{m}(i)}, g^{\text{p}}_{\text{m}(i)}] = \frac{1}{2}(g^{\text{p}}_{\text{p}(i)} + g^{\text{m}}_{\text{p}(i)} + g^{\text{m}}_{\text{m}(i)} + g^{\text{p}}_{\text{m}(i)}), \tag{5}$$

any linear dependence between g and and e flows through $g^{\text{par}}_i = g^{\text{p}}_{\text{p}(i)} + g^{\text{m}}_{\text{p}(i)} + g^{\text{m}}_{\text{m}(i)} + g^{\text{p}}_{\text{m}(i)}$. This notion is proven formally by the Conditional Independence Lemma (Appendix B), which implies that,

$$Y_i = \mu + \delta g_i + \eta g^{\text{par}}_i + \epsilon_i, \text{ for some } \epsilon_i \text{ such that } \text{Cov}(\epsilon_i, g_i) = \text{Cov}(\epsilon_i, g^{\text{par}}_i) = 0, \tag{6}$$

where η is the regression coefficient of e_i on g^{par}_i, and ϵ_i is the residual environmental effect on the phenotype of i after regression of e_i on g^{par}_i. We further assume, without loss of generality, that $\mathbb{E}[\epsilon_i] = 0$, as any non-zero mean can be incorporated into μ by reparameterisation. Note we have not made any assumption about the nature of the relationship between g^{par}_i and e_i, which could be non-linear. This leads to a simple decomposition of the phenotypic variance:

$$\text{Var}(Y_i) = \delta^2\text{Var}(g_i) + \eta^2\text{Var}(g^{\text{par}}_i) + 2\delta\eta\text{Cov}(g_i, g^{\text{par}}_i) + \text{Var}(\epsilon_i), \tag{7}$$

where $\delta^2\text{Var}(g_i)$ is the variance explained by the direct effect of genetic inheritance at the locus, $\eta^2\text{Var}(g^{\text{par}}_i)$ is the variance of the part of the environmental component of the phenotype that is correlated with parental genotype, $2\delta\eta\text{Cov}(g_i, g^{\text{par}}_i)$ is the covariance between the direct genetic effect and environmental effects, and $\text{Var}(\epsilon_i)$ is variance of the component of the phenotype that is uncorrelated with both parent and offspring genotype. We retain the subscript i to allow for heteroskedasticity between individuals, arising from different levels of inbreeding or different environmental variances. A variance decomposition for the whole population could then be found by applying the law of total variance:

$$\text{Var}(Y) = \mathbb{E}_i[\text{Var}(Y_i)] + \text{Var}_i(\mathbb{E}[Y_i]). \tag{8}$$

The decomposition is true for traits that satisfy the assumptions of additivity of direct genetic effects and no interaction between direct genetic effects and environmental effects. We note that $v_{e\sim g}$ will capture variance explained by (additive) parental genetic nurturing effects, among other sources of environmental variation.

Alternative variance decompositions are possible. We can write

$$Y_i = \delta\left(g_i - \frac{1}{2}g^{\text{par}}_i\right) + \left(\eta + \frac{\delta}{2}\right)g^{\text{par}}_i + \epsilon_i. \tag{9}$$

The first component is the variation in offspring genotype unexplained by parental genotype, so $\text{Cov}(g^{\text{par}}_i, (g_i - \frac{1}{2}g^{\text{par}}_i)) = 0$,

$$\Rightarrow \text{Var}(Y_i) = \delta^2\text{Var}\left(g_i - \frac{1}{2}g^{\text{par}}_i\right) + \left(\eta + \frac{\delta}{2}\right)^2\text{Var}(g^{\text{par}}_i) + \text{Var}(\epsilon). \tag{10}$$

This variance decomposition only has two genetic components, so may be preferred for some applications. However, $\delta^2\text{Var}\left(g_i - \frac{1}{2}g_i^{\text{par}}\right) < v_g$, with the difference depending on the level of inbreeding in the parents and relatedness between the parents in the population, so the interpretation of the variance components from fitting this decomposition may not be clear.

Although we cannot in general know ϵ, we can obtain a consistent estimator of δ by regressing Y_i jointly onto g_i and g_i^{par}. This follows from the fact that $\text{Cov}(g_i, \epsilon_i) = 0$ and $\mathbb{E}[\epsilon_i] = 0$ by standard regression theory.

The sum of the parental genotypes and the offspring genotype are linearly dependent in expectation, so the information to fit this regression model comes entirely from the deviation of the offspring genotype around its expectation. We will show that fitting the covariance model implied by this regression model gives an estimator of v_g that removes environmental bias, with the information coming from the deviation in relatedness between offspring around the expectation given by the relatedness of the parents.

1.2 Covariance between relatives

We now derive the covariance between pairs of individuals conditional on the identity-by-descent (IBD) sharing states between their inherited chromosomes and their parents' chromosomes. (Note we use the initials IBD to refer to both 'identity-by-descent' and 'identical-by-descent' depending on context.) We consider the causal variant to be located randomly with respect to identity-by-descent sharing. For a genome comprised of L locations, and for $k, l = \text{m,p}$, where 'm' indicates the maternal variant, and 'p' indicates the paternal variant:

$$\begin{aligned} \text{IBD}_{ij}^{kl} &= \mathbb{P}(\text{the } k\text{-variant of } i \text{ is IBD with the } l\text{-variant of } j) && (11) \\ &= \frac{1}{L}\sum_{s=1}^{L} \text{IBD}_{ij}^{kl}(s), && (12) \end{aligned}$$

where

$$\text{IBD}_{ij}^{kl}(s) = \begin{cases} 1 & \text{if } k\text{-chromosome of } i \text{ is IBD with } l\text{-chromosome of } j \text{ at position } s; \\ 0 & \text{otherwise.} \end{cases}$$

We now derive the covariance under population genetic assumptions of a random-mating population with finite ancestral size[33]. The covariance between two individuals becomes complicated when the residual environment of one individual is correlated with the genetic variant of the other. For the covariance between a pair i and j, we first assume that ϵ_i is uncorrelated with g_j and g_j^{par} and that ϵ_j is uncorrelated with g_i and g_i^{par}. If we assume this is true for all pairs, then, if the allele frequency of g is f, the covariance matrix for a vector of phenotype observations, $\mathbf{Y}$, is

$$\text{Cov}(\mathbf{Y}) = \delta^2 2f(1-f)\text{R} + \eta^2 4f(1-f)\text{R}_{\text{par}} + 2\delta\eta 2f(1-f)\text{R}_{\text{o,par}} + \text{Cov}(\epsilon), \quad (13)$$

where R is the additive relatedness matrix in a finite population, with i, j^{th} element equal to[33]

$$\mathrm{R}_{ij} = \frac{1}{2} \sum_{k,l=\mathrm{m,p}} \frac{\mathrm{IBD}_{ij}^{kl} - \mathrm{K}_0}{1 - \mathrm{K}_0}, \tag{14}$$

where K_0 is the mean kinship over all pairs in the population; $\mathrm{R}_{\mathrm{par}}$ is the additive relatedness matrix between the parents of individuals in sample, with i, j^{th} element equal to

$$[\mathrm{R}_{\mathrm{par}}]_{ij} = \frac{\mathrm{K}_{\mathrm{p}(i)\mathrm{p}(j)} + \mathrm{K}_{\mathrm{p}(i)\mathrm{m}(j)} + \mathrm{K}_{\mathrm{m}(i)\mathrm{p}(j)} + \mathrm{K}_{\mathrm{m}(i)\mathrm{m}(j)} - 4\mathrm{K}_0}{1 - \mathrm{K}_0}, \tag{15}$$

where $\mathrm{K}_{\mathrm{p}(i)\mathrm{m}(j)}$ is the kinship between the father of i and the mother of j; and $\mathrm{R}_{\mathrm{o,par}}$ is the additive relatedness between parents and offspring of individuals in the sample, with i, j^{th} element equal to

$$[\mathrm{R}_{\mathrm{o,par}}]_{ij} = \frac{\mathrm{K}_{i\mathrm{p}(j)} + \mathrm{K}_{i\mathrm{m}(j)} + \mathrm{K}_{\mathrm{p}(i)j} + \mathrm{K}_{\mathrm{m}(i)j} - 4\mathrm{K}_0}{1 - \mathrm{K}_0}, \tag{16}$$

where $\mathrm{K}_{i\mathrm{m}(j)}$ is the kinship between i and the mother of j, etc. Note that both $\mathrm{R}_{\mathrm{par}}$ and $\mathrm{R}_{\mathrm{o,par}}$ have diagonal elements equal to one in an infinite, outbred, random-mating population, where K_0 is zero[33].

The covariance matrix can be written in terms of the variance components defined in (7) in an infinite, outbred, random-mating population:

$$\mathrm{Cov}(\mathbf{Y}) = v_g \mathrm{R} + v_{e\sim g} \mathrm{R}_{\mathrm{par}} + c_{g,e} \mathrm{R}_{\mathrm{o,par}} + \mathrm{Cov}(\epsilon), \tag{17}$$

where $v_g = \delta^2 2f(1-f)$ is the variance explained by the direct effect of genetic inheritance at the locus, $v_{e\sim g} = \eta^2 4f(1-f)$ is the variance of the part of the environmental component of the phenotype that is correlated with parental genotype, $c_{g,e} = 2\delta\eta 2f(1-f)$ is the covariance between the direct genetic effect and environmental effects

1.3 Residual genetic correlations

We now relax the assumption that the residual environment of i is uncorrelated with g_j and vice-versa. We can write ϵ_i in terms of its regression on g_j and g_j^{par} and the residual of this regression:

$$\epsilon_i = a_{ji} g_j + b_{ji} g_j^{\mathrm{par}} + \epsilon_{ji}, \tag{18}$$

where $\mathrm{Cov}(\epsilon_{ji}, g_j) = \mathrm{Cov}(\epsilon_{ji}, g_j^{\mathrm{par}}) = 0$. The equivalent can be done for ϵ_j:

$$\epsilon_j = a_{ij} g_i + b_{ij} g_i^{\mathrm{par}} + \epsilon_{ij}, \tag{19}$$

where $\mathrm{Cov}(\epsilon_{ij}, g_i) = \mathrm{Cov}(\epsilon_{ij}, g_i^{\mathrm{par}}) = 0$. We note that if $a_{ij} \neq 0$ and $a_{ji} \neq 0$, this implies that the genotype of i affects the phenotype of j and vice-versa, which we term a 'reciprocal genetic effect'. If this is the case, it creates problems for estimating heritability for all methods based on genetic relatedness (Appendix A).

The covariance between i and j is therefore

$$\begin{aligned}\text{Cov}(Y_i, Y_j) = v_g \text{R}_{ij} + v_{e\sim g}[\text{R}_{\text{par}}]_{ij} + c_{g,e}[\text{R}_{\text{o,par}}]_{ij} + \delta[\text{Cov}(g_j, \epsilon_i) + \text{Cov}(g_i, \epsilon_j)] + \\ \eta[\text{Cov}(g_j^{\text{par}}, \epsilon_i) + \text{Cov}(g_i^{\text{par}}, \epsilon_j)] + \text{Cov}(\epsilon_i, \epsilon_j).\end{aligned} \tag{20}$$

We have that

$$\text{Cov}(\epsilon_i, g_j) = 2f(1-f)(a_{ji}\text{R}_{jj} + b_{ji}[\text{R}_{\text{o,par}}]_{jj}), \tag{21}$$

$$\text{Cov}(\epsilon_i, g_j^{\text{par}}) = 2f(1-f)(a_{ij}[\text{R}_{\text{o,par}}]_{jj} + 2b_{ij}[\text{R}_{\text{par}}]_{jj}), \tag{22}$$

with equivalent results for ϵ_j. This gives

$$\begin{aligned}&\text{Cov}(Y_i, Y_j) = v_g \text{R}_{ij} + v_{e\sim g}[\text{R}_{\text{par}}]_{ij} + c_{g,e}[\text{R}_{\text{o,par}}]_{ij} + \text{Cov}(\epsilon_i, \epsilon_j) + \\ &\delta 2f(1-f)[a_{ji}\text{R}_{jj} + a_{ij}\text{R}_{ii}] + \eta 4f(1-f)[b_{ji}[\text{R}_{\text{par}}]_{jj} + b_{ij}[\text{R}_{\text{par}}]_{ii}] + \\ &2f(1-f)[(\delta b_{ji} + \eta a_{ji})[\text{R}_{\text{o,par}}]_{jj} + (\delta b_{ij} + \eta a_{ij})[\text{R}_{\text{o,par}}]_{ii}].\end{aligned} \tag{23}$$

The covariance of the residual environmental effects is

$$\text{Cov}(\epsilon_i, \epsilon_j) = 2f(1-f)(a_{ji}a_{ij}\text{R}_{ij} + b_{ij}b_{ji}[\text{R}_{\text{par}}]_{ij} + (b_{ji}a_{ij} + b_{ij}a_{ji})[\text{R}_{\text{o,par}}]_{ij}) + c_{ij} \tag{24}$$

where c_{ij} is a sum of terms involving the covariances between g_i, g_i^{par} and ϵ_{ji}, and between g_j, g_j^{par} and ϵ_{ij}. Since $\text{Cov}(\epsilon_{ij}, g_i) = 0$ and $\text{Cov}(\epsilon_{ji}, g_j) = 0$, c_{ij} cannot contain any terms directly due to the covariance between g_i and g_j. We therefore have that the coefficient of R_{ij} in $\text{Cov}(Y_i, Y_j)$ is $v_g + 2f(1-f)a_{ji}a_{ij}$.

1.4 Consistent estimation of heritability

The covariances due to residual environmental effects are, in general, unknown. RDR fits a simplified model of the phenotypic covariance matrix:

$$\text{Cov}(\mathbf{Y}) = v_g\text{R} + v_{e\sim g}\text{R}_{\text{par}} + c_{g,e}\text{R}_{\text{o,par}} + \sigma^2\text{I}. \tag{25}$$

We prove that fitting the RDR covariance model by least-squares regression of the off-diagonal elements of the sample phenotypic covariance matrix jointly onto the off-diagonal elements of R, R_{par}, and $\text{R}_{\text{o,par}}$ gives a consistent estimator for v_g. To do this, however, we have to show that fitting R_{par} and R jointly removes any confounding between the elements of R and residual environmental effect sharing. This is necessary because, even though ϵ and g are uncorrelated, their associated covariance matrices could be related. This may be the case when environmental effects uncorrelated with parental genetics are more similar for more related people, either due to broad-scale environmental effects or to family level environmental effects.

We showed that when indirect genetic effects are present, specifically reciprocal genetic effects, this can generate covariance between pairs of individuals that is indistinguishable from the covariance due to the effects of directly inherited genetic variants (Appendix A and Subsection 1.3). This may seem an insurmountable problem for any method of estimating heritability using the change in phenotypic covariance with genetic relatedness, such as twin studies or Sib-Regression[5]. However, we prove that our estimator will still converge to the true heritability provided that the fraction of pairs exhibiting indirect genetic effects on each other tends to zero with sample size.

There is a further complication when pairs in the sample are related by direct descent. This is because the segregation events between pairs that are related by direct descent are not independent. For the purposes of this analysis, we consider monozygotic twins as related by direct descent.

We first prove consistency when there are no indirect genetic effects between pairs in the sample and no pairs in the sample are related by direct descent. To do this, we prove the following lemma for pairs of individuals that *do not* exhibit indirect genetic effects on each other's residual environments and are not related by direct descent.

1.4.1 Relatedness Disequilibrium Lemma

Lemma 1. Let

$$\mathrm{IBD}_{ij}^{\mathrm{par}} = \{\mathrm{IBD}_{k(i)l(j)}^{k'l'}(s)\}_{k,l,k',l'=\mathrm{m,p};\ s=1,\dots,L}, \tag{26}$$

and let $\mathrm{I}(i) = \{\mathrm{I}_{k\mathrm{m}}(i,s)\}_{k=\mathrm{m,p};s=1,\dots,L}$, then, for i and j not related by direct descent,

$$\mathrm{R}_{ij}, [\mathrm{R}_{\mathrm{o,par}}]_{ij} \perp \epsilon_i, \epsilon_j | \mathrm{IBD}_{ij}^{\mathrm{par}}, \text{ if } \epsilon_j \perp \mathrm{I}(i) \text{ and } \epsilon_i \perp \mathrm{I}(j). \tag{27}$$

Remark. This says that the additive relatedness between a pair i, j, R_{ij}, is independent of the residual environmental effects on i and j given the IBD sharing states between the parents of i and the parents of j, if the residual environment of i is independent of the segregation events leading to j's genotype and vice-versa. This is true because the IBD sharing between i and j is determined by the IBD sharing between the parents of i and the parents of j and the segregation events in those parents (Supplementary Figure 1).

Proof. We now express $\mathrm{IBD}_{ij}^{kl}(s)$ as a function of the IBD sharing between the parents of i and the parents of j at position s, and the segregation events in those parents. For $k, k' = \mathrm{m,p}$, we define the segregation variables in the parents of i:

$$\mathrm{I}_{kk'}(i,s) = \begin{cases} 1 & \text{if the genetic variant at position } s \text{ on the } k\text{-chromosome of } i \\ & \text{was inherited from the } k' \text{ parent of } k; \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

For example, $\mathrm{I}_{\mathrm{mp}}(i,2)$ is 1 if the maternal variant of i at position 2 was inherited from i's maternal grandfather. Note that $\mathrm{I}_{\mathrm{mp}}(i,s) = 1 - \mathrm{I}_{\mathrm{mm}}(i,s)$, as the maternal variant of

i was either inherited from the maternal grandmother or the maternal grandfather. Let $\mathrm{m}(i)$ be the mother of i and $\mathrm{p}(j)$ be the father of j, etc., then

$$\mathrm{IBD}^{kl}_{ij}(s) = \sum_{k',l'=\mathrm{m,p}} \mathrm{I}_{kk'}(i,s)\mathrm{I}_{ll'}(j,s)\mathrm{IBD}^{k'l'}_{k(i)l(j)}(s). \tag{29}$$

$$\Rightarrow \mathrm{IBD}^{kl}_{ij} = \sum_{k',l'=\mathrm{m,p}} \left[\frac{1}{L}\sum_{s=1}^{L} \mathrm{I}_{kk'}(i,s)\mathrm{I}_{ll'}(j,s)\mathrm{IBD}^{k'l'}_{k(i)l(j)}(s)\right]. \tag{30}$$

If there is no recombination and only one chromosome, then $\mathrm{I}_{kk'}(i,s) = \mathrm{I}_{kk'}(i)$ and $\mathrm{I}_{ll'}(j,s) = \mathrm{I}_{ll'}(j)$ for $s = 1,\ldots,L$, and

$$\mathrm{IBD}^{kl}_{ij} = \sum_{k',l'=\mathrm{m,p}} \left[\mathrm{I}_{kk'}(i)\mathrm{I}_{ll'}(j)\frac{1}{L}\sum_{s=1}^{L} \mathrm{IBD}^{k'l'}_{k(i)l(j)}(s)\right] = \sum_{k',l'=\mathrm{m,p}} \mathrm{I}_{kk'}(i)\mathrm{I}_{ll'}(j)\mathrm{IBD}^{k'l'}_{k(i)l(j)} \tag{31}$$

$$\Rightarrow \mathrm{IBD}^{kl}_{ij} \perp \epsilon_i | \{\mathrm{IBD}^{k'l'}_{k(i)l(j)}\}_{k',l'=\mathrm{m,p}} \text{ if } \epsilon_i \perp \mathrm{I}_{l\mathrm{m}}(j), \mathrm{I}_{k\mathrm{m}}(i). \tag{32}$$

For example, for $k = m$ and $l = p$, this means that the IBD sharing between the maternal chromosome of individual i and paternal chromosome of j is conditionally independent of ϵ_i, given the IBD sharing states between the mother of i and the father of j, if ϵ_i is independent of the relevant segregation events in j's father and i's mother. From the assumptions of the model, the environment of i, and therefore ϵ_i, is independent of $\mathrm{I}_{k\mathrm{m}}(i)$. From the assumptions of the Lemma, $\epsilon_i \perp \mathrm{I}_{l\mathrm{m}}(j)$, so

$$\mathrm{IBD}^{kl}_{ij} \perp \epsilon_i | \{\mathrm{IBD}^{k'l'}_{k(i)l(j)}\}_{k',l'=\mathrm{m,p}} \tag{33}$$

Also from the assumptions of the Lemma, $\epsilon_j \perp \mathrm{I}_{k\mathrm{m}}(i)$, implying

$$\mathrm{IBD}^{kl}_{ij} \perp \epsilon_i, \epsilon_j | \{\mathrm{IBD}^{k'l'}_{k(i)l(j)}\}_{k',l'=\mathrm{m,p}}. \tag{34}$$

This is true for for $k, l = \mathrm{m,p}$, so

$$\mathrm{R}_{ij} \perp \epsilon_i, \epsilon_j | \{\mathrm{IBD}^{k'l'}_{k(i)l(j)}\}_{k,l,k',l'=\mathrm{m,p}}. \tag{35}$$

This says that R_{ij} is conditionally independent of ϵ_i and ϵ_j, given the IBD sharing states between the parents of i and the parents of j.

While we have derived this under the assumption of no-recombination, we can relax this assumption to get the more general result. Let

$$\mathrm{IBD}^{\mathrm{par}}_{ij} = \{\mathrm{IBD}^{k'l'}_{k(i)l(j)}(s)\}_{k,l,k',l'=\mathrm{m,p};\, s=1,\ldots,L}, \tag{36}$$

and let $\mathrm{I}(i) = \{\mathrm{I}_{k\mathrm{m}}(i,s)\}_{k=\mathrm{m,p};s=1,\ldots,L}$,

$$\mathrm{R}_{ij} \perp \epsilon_i, \epsilon_j | \mathrm{IBD}^{\mathrm{par}}_{ij}, \text{ if } \epsilon_j \perp \mathrm{I}(i) \text{ and } \epsilon_i \perp \mathrm{I}(j). \tag{37}$$

A similar result applies to $\mathrm{R}_{\mathrm{o,par}}$. The elements of $\mathrm{R}_{\mathrm{o,par}}$ are constructed from the kinship coefficients between i and the parents of j, and vice-versa, such as

$$\mathrm{K}_{i\mathrm{p}(j)} = \frac{1}{4} \sum_{k=\mathrm{m,p}} \sum_{l'=\mathrm{m,p}} \mathrm{IBD}_{i\mathrm{p}(j)}^{kl'}. \tag{38}$$

$$\mathrm{IBD}_{i\mathrm{p}(j)}^{kl'} = \frac{1}{L} \sum_{s=1}^{L} \mathrm{IBD}_{i\mathrm{p}(j)}^{kl'}(s) = \frac{1}{L} \sum_{s=1}^{L} \sum_{k'=\mathrm{m,p}} \mathrm{I}_{kk'}(i,s) \mathrm{IBD}_{k(i)\mathrm{p}(j)}^{k'l'}(s) \tag{39}$$

$$\Rightarrow \mathrm{K}_{i\mathrm{p}(j)} = \frac{1}{4} \sum_{k=\mathrm{m,p}} \sum_{k',l'=\mathrm{m,p}} \frac{1}{L} \sum_{s=1}^{L} \mathrm{I}_{kk'}(i,s) \mathrm{IBD}_{k(i)\mathrm{p}(j)}^{k'l'}(s). \tag{40}$$

We therefore have that $\mathrm{K}_{i\mathrm{p}(j)} \perp \epsilon_i | \mathrm{IBD}_{ij}^{\mathrm{par}}$, and, if $\epsilon_j \perp \mathrm{I}(i)$, then $\mathrm{K}_{i\mathrm{p}(j)} \perp \epsilon_i, \epsilon_j | \mathrm{IBD}_{ij}^{\mathrm{par}}$. This also applies to $\mathrm{K}_{i\mathrm{m}(j)}$. Similarly, for $l = \mathrm{m,p}$, we have $\mathrm{K}_{l(i)j} \perp \epsilon_j | \mathrm{IBD}_{ij}^{\mathrm{par}}$, and, if $\epsilon_i \perp \mathrm{I}(j)$, then $\mathrm{K}_{l(i)j} \perp \epsilon_i, \epsilon_j | \mathrm{IBD}_{ij}^{\mathrm{par}}$. These results imply that $[\mathrm{R}_{\mathrm{o,par}}]_{ij} \perp \epsilon_i, \epsilon_j | \mathrm{IBD}_{ij}^{\mathrm{par}}$, given that $\epsilon_i \perp \mathrm{I}(j)$ and $\epsilon_j \perp \mathrm{I}(i)$. □

1.4.2 Proof of consistency

Let $L(A)$ be the column vector of lower-triangular elements of a matrix A, excluding the diagonal, in lower-triangular order. If $\mathbf{y}$ is the column vector of phenotype observations and $\bar{y}$ is the sample mean phenotype, then the sample covariance matrix is $S = (\mathbf{y} - \bar{y})(\mathbf{y} - \bar{y})^T$. We regress

$$L(S) \sim X = [L(\mathrm{R}) \; L(\mathrm{R}_{\mathrm{par}}) \; L(\mathrm{R}_{\mathrm{o,par}})]. \tag{41}$$

Let $\theta = [v_g, v_{e\sim g}, c_{g,e}]^T$, then our least-squares estimate of θ is

$$\hat{\theta} = [\hat{v}_g, \hat{v}_{e\sim g}, \hat{c}_{g,e}]^T = (X^T X)^{-1} X^T L(S). \tag{42}$$

We are regressing over all pairs of individuals in a sample. In this context, an unconditional expectation, for example $\mathbb{E}[(g_i - \bar{g})(g_j - \bar{g})]$, is the expectation over all pairs $i \neq j$, where $\bar{z}$ implies the sample mean of z. If, however, we condition on genetic relatedness, this is then the expectation over all pairs with genetic relatedness equal to R_{ij}, $\mathbb{E}[(g_i - \bar{g})(g_j - \bar{g}) | \mathrm{R}_{ij}] = 2f(1-f)\mathrm{R}_{ij}$.

The consistency of the estimator of v_g derives from the fact that the correlation over all pairs of the genetic relatedness, R_{ij}, with covariation of residual environmental effects, $\epsilon_i \epsilon_j$, is removed by regressing jointly with the relatedness between the parents of i and j, $[\mathrm{R}_{\mathrm{par}}]_{ij}$. However, this is not true for pairs of individuals who have indirect genetic effects upon each other and for pairs related by direct descent (Appendix A and Lemma 1). We therefore assume that no pairs are related by direct descent. We further assume that the genotype of i does not affect the residual environment of j and vice-versa for all pairs

i, j, and then we examine the conditions for consistency of the estimator of v_g when this assumption is not true.

We have shown that the phenotypes of individuals i and j can be written as (6)

$$\begin{aligned} y_i &= \delta g_i + \eta g_i^{\text{par}} + \epsilon_i, \\ y_j &= \delta g_j + \eta g_j^{\text{par}} + \epsilon_j, \end{aligned} \tag{43}$$

If we define $G_i = \delta g_i + \eta g_i^{\text{par}}$, then

$$\begin{aligned} S_{ij} = (y_i - \bar{y})(y_j - \bar{y}) = &(G_i - \bar{G})(G_j - \bar{G}) + (G_i - \bar{G})(\epsilon_j - \bar{\epsilon}) + \\ &(G_j - \bar{G})(\epsilon_i - \bar{\epsilon}) + (\epsilon_i - \bar{\epsilon})(\epsilon_j - \bar{\epsilon}). \end{aligned} \tag{44}$$

We can write this in terms of its expectation given X and the sampling error that is uncorrelated with X. The covariance of the genetic components given X is

$$\mathbb{E}[(G_i - \bar{G})(G_j - \bar{G})|X] = v_g \text{R}_{ij} + v_{e\sim g}[\text{R}_{\text{par}}]_{ij} + c_{g,e}[\text{R}_{\text{o,par}}]_{ij}. \tag{45}$$

For the components of the sample covariance involving the residual environments of i and j, we have that (18 and 19)

$$\epsilon_i = a_{ji} g_j + b_{ji} g_j^{\text{par}} + \epsilon_{ji}, \text{ and } \epsilon_j = a_{ij} g_i + b_{ij} g_i^{\text{par}} + \epsilon_{ij}, \tag{46}$$

for $\text{Cov}(\epsilon_{ij}, g_i) = \text{Cov}(\epsilon_{ij}, g_i^{\text{par}}) = \text{Cov}(\epsilon_{ji}, g_j) = \text{Cov}(\epsilon_{ji}, g_j^{\text{par}}) = 0$.

We first prove consistency without indirect genetic effects between pairs in the sample.

Theorem 2. If

1. for all pairs i, j, i and j are not related by direct descent;
2. for all pairs i, j, the residual environment of i is independent of $\text{I}(j)$ and the residual environment of j is independent of $\text{I}(i)$ (Definition E.9);
3. the sample is drawn from a random-mating population;
4. genetic variants with direct causal effects are located at random with respect to identity-by-descent sharing;
5. direct effects of inherited genetic variants are additive, including no parent-of-origin effects;
6. there is no gene-environment interaction;

then,

$$\lim_{n\to\infty} \hat{v}_g = v_g. \tag{47}$$

Proof. If $\epsilon_i \perp \mathrm{I}(j)$ and $\epsilon_j \perp \mathrm{I}(i)$, then $a_{ji} = a_{ij} = 0$ (19 and 18). We therefore have that (23)

$$\begin{aligned}\mathbb{E}[(G_i - \bar{G})(\epsilon_j - \bar{\epsilon}) + (G_j - \bar{G})(\epsilon_i - \bar{\epsilon})|X] = &\eta 4f(1-f)[b_{ji}[\mathrm{R}_{\mathrm{par}}]_{jj} + b_{ij}[\mathrm{R}_{\mathrm{par}}]_{ii}] + \\ &\delta 2f(1-f)[b_{ji}[\mathrm{R}_{\mathrm{o,par}}]_{jj} + b_{ij}[\mathrm{R}_{\mathrm{o,par}}]_{ii}].\end{aligned} \tag{48}$$

In addition to $a_{ji} = a_{ij} = 0$, $\epsilon_i \perp \mathrm{I}(j)$ and $\epsilon_j \perp \mathrm{I}(i)$ implies that b_{ij} and b_{ji} are independent of $\mathrm{I}(i)$ and $\mathrm{I}(j)$. This means that, although the coefficients b_{ij} and b_{ji} may rely upon the relatedness between the parents of i and the parents of j, $[\mathrm{R}_{\mathrm{par}}]_{ij}$, they cannot rely directly upon the relatedness between i and j, R_{ij}. This implies that $\mathrm{R}_{ij} \perp b_{ij}, b_{ji} | \mathrm{IBD}^{\mathrm{par}}_{ij}$. Furthermore, this allows us to apply the Conditional Independence Lemma (Appendix B) to elements of (48).

We can therefore express S_{ij} as:

$$\begin{aligned}S_{ij} = &v_g \mathrm{R}_{ij} + v_{e\sim g}[\mathrm{R}_{\mathrm{par}}]_{ij} + c_{g,e}[\mathrm{R}_{\mathrm{o,par}}]_{ij} + \eta 4f(1-f)[b_{ji}[\mathrm{R}_{\mathrm{par}}]_{jj} + b_{ij}[\mathrm{R}_{\mathrm{par}}]_{ii}] + \\ &\delta 2f(1-f)[b_{ji}[\mathrm{R}_{\mathrm{o,par}}]_{jj} + b_{ij}[\mathrm{R}_{\mathrm{o,par}}]_{ii}] + (\epsilon_i - \bar{\epsilon})(\epsilon_j - \bar{\epsilon}) + \xi_{Gij},\end{aligned} \tag{49}$$

where ξ_{Gij} represents sampling error uncorrelated with X, and $\mathbb{E}[\xi_{Gij}|X] = 0$.

From the Relatedness Disequilibrium Lemma (Lemma 1), we know that, when $\epsilon_i \perp \mathrm{I}(j)$ and $\epsilon_j \perp \mathrm{I}(i)$ and i and j are not related by direct descent,

$$\mathrm{R}_{ij}, [\mathrm{R}_{\mathrm{o,par}}]_{ij} \perp \epsilon_i, \epsilon_j | \mathrm{IBD}^{\mathrm{par}}_{ij}, \tag{50}$$

and from Lemma 4 that, for i, j not related by direct descent,

$$\mathbb{E}[\mathrm{R}_{ij}|\mathrm{IBD}^{\mathrm{par}}_{ij}] = \frac{1}{2}[\mathrm{R}_{\mathrm{par}}]_{ij}, \text{and } \mathbb{E}[[\mathrm{R}_{\mathrm{o,par}}]_{ij}|\mathrm{IBD}^{\mathrm{par}}_{ij}] = [\mathrm{R}_{\mathrm{par}}]_{ij}, \tag{51}$$

where $\mathrm{IBD}^{\mathrm{par}}_{ij}$ (Definition E.7) represents the genome-wide IBD sharing states between the parents of i and the parents of j. Therefore, by the Conditional Independence Lemma (Lemma 1), since $[\mathrm{R}_{\mathrm{par}}]_{ij}$ is a linear combination of the elements of $\mathrm{IBD}^{\mathrm{par}}_{ij}$, for some $\xi_{\epsilon ij}$ such that $\mathrm{Cov}(\mathrm{R}_{ij}, \xi_{\epsilon ij}) = \mathrm{Cov}([\mathrm{R}_{\mathrm{o,par}}]_{ij}, \xi_{\epsilon ij}) = \mathrm{Cov}([\mathrm{R}_{\mathrm{par}}]_{ij}, \xi_{\epsilon ij}) = 0$,

$$(\epsilon_i - \bar{\epsilon})(\epsilon_j - \bar{\epsilon}) = \zeta_{\mathrm{par}}[\mathrm{R}_{\mathrm{par}}]_{ij} + \xi_{\epsilon ij}, \tag{52}$$

where ζ_{par} is the regression coefficient from regression of $(\epsilon_i - \bar{\epsilon})(\epsilon_j - \bar{\epsilon})$ on $[\mathrm{R}_{\mathrm{par}}]_{ij}$. Note we have not made an assumption that the relationship between $[\mathrm{R}_{\mathrm{par}}]_{ij}$ and $(\epsilon_i - \bar{\epsilon})(\epsilon_j - \bar{\epsilon})$ is linear to derive this.

We now analyse the dependence between R_{ij} and the diagonal elements of $\mathrm{R}_{\mathrm{par}}$ and $\mathrm{R}_{\mathrm{o,par}}$. For

$$[\mathrm{R}_{\mathrm{par}}]_{ii} = \frac{\mathrm{K}_{\mathrm{p}(i)\mathrm{p}(i)} + \mathrm{K}_{\mathrm{p}(i)\mathrm{m}(i)} + \mathrm{K}_{\mathrm{m}(i)\mathrm{p}(i)} + \mathrm{K}_{\mathrm{m}(i)\mathrm{m}(i)} - 4\mathrm{K}_0}{1 - \mathrm{K}_0}, \tag{53}$$

we have that $[\mathrm{R}_{\mathrm{par}}]_{ii} \perp \mathrm{I}(i), \mathrm{I}(j)$. Therefore,

$$\mathrm{R}_{ij}, [\mathrm{R}_{\mathrm{o,par}}]_{ij} \perp [\mathrm{R}_{\mathrm{par}}]_{ii} | \mathrm{IBD}^{\mathrm{par}}_{ij}. \tag{54}$$

The same applies for $[\mathrm{R}_{\mathrm{par}}]_{jj}$, so

$$\mathrm{R}_{ij}, [\mathrm{R}_{\mathrm{o,par}}]_{ij} \perp [\mathrm{R}_{\mathrm{par}}]_{jj}, [\mathrm{R}_{\mathrm{par}}]_{ii} | \mathrm{IBD}_{ij}^{\mathrm{par}}. \tag{55}$$

By the Conditional Independence Lemma (Appendix B), since $\mathbb{E}[\mathrm{R}_{ij}|\mathrm{IBD}_{ij}^{\mathrm{par}}] = [\mathrm{R}_{\mathrm{par}}]_{ij}/2$, for some constant b and some ξ_{bij} such that

$$\mathrm{Cov}(\xi_{bij}, \mathrm{R}_{ij}) = \mathrm{Cov}(\xi_{bij}, [\mathrm{R}_{\mathrm{o,par}}]_{ij}) = \mathrm{Cov}(\xi_{bij}, [\mathrm{R}_{\mathrm{par}}]_{ij}) = 0, \tag{56}$$

$$\eta 4f(1-f)[b_{ji}[\mathrm{R}_{\mathrm{par}}]_{jj} + b_{ij}[\mathrm{R}_{\mathrm{par}}]_{ii}] = b[\mathrm{R}_{\mathrm{par}}]_{ij} + \xi_{bij}. \tag{57}$$

For

$$[\mathrm{R}_{\mathrm{o,par}}]_{ii} = \frac{2(\mathrm{K}_{i\mathrm{p}(i)} + \mathrm{K}_{i\mathrm{m}(i)}) - 4\mathrm{K}_0}{1 - \mathrm{K}_0}, \tag{58}$$

where (Lemma 5)

$$\mathrm{K}_{i\mathrm{p}(i)} + \mathrm{K}_{i\mathrm{m}(i)} = \frac{1}{4} \sum_{k,l=\mathrm{m,p}} \sum_{k',l'=\mathrm{m,p}} \frac{1}{L} \sum_{s=1}^{L} \mathrm{I}_{kk'}(i,s) \mathrm{IBD}_{k(i)l(i)}^{k'l'}(s), \tag{59}$$

we have that $\mathrm{R}_{ij} \perp [\mathrm{R}_{\mathrm{o,par}}]_{ii} | \mathrm{IBD}_{ij}^{\mathrm{par}}, \mathrm{I}(i)$. From Lemma 5, we have that

$$\mathbb{E}[\mathrm{R}_{ij}|\mathrm{IBD}_{ij}^{\mathrm{par}}, \mathrm{I}(i)] = \frac{\mathrm{K}_{i\mathrm{p}(j)} + \mathrm{K}_{i\mathrm{m}(j)} - 2\mathrm{K}_0}{1 - \mathrm{K}_0}. \tag{60}$$

Therefore, by the Conditional Independence Lemma (Appendix B), for some constant b' and some $\xi_{b'0ij}$ such that $\mathrm{Cov}(\xi_{b'0ij}, \mathrm{R}_{ij}) = 0$,

$$\delta 2f(1-f) b_{ij} [\mathrm{R}_{\mathrm{o,par}}]_{ii} = b' \left(\frac{\mathrm{K}_{i\mathrm{p}(j)} + \mathrm{K}_{i\mathrm{m}(j)} - 2\mathrm{K}_0}{1 - \mathrm{K}_0} \right) + \xi_{b'0ij}. \tag{61}$$

However, we may have that $\mathrm{Cov}(\xi_{b'0ij}, [\mathrm{R}_{\mathrm{par}}]_{ij}) \neq 0$. Although we do not prove it, we believe it to be an obvious extension of the Conditional Independence Lemma (Appendix B) that, for some constants b' and b'_{par} and some $\xi_{b'0ij}$ such that

$$\mathrm{Cov}(\xi_{b'0ij}, \mathrm{R}_{ij}) = \mathrm{Cov}(\xi_{b'0ij}, [\mathrm{R}_{\mathrm{par}}]_{ij}) = 0, \tag{62}$$

$$\delta 2f(1-f) b_{ij} [\mathrm{R}_{\mathrm{o,par}}]_{ii} = b' \left(\frac{\mathrm{K}_{i\mathrm{p}(j)} + \mathrm{K}_{i\mathrm{m}(j)} - 2\mathrm{K}_0}{1 - \mathrm{K}_0} \right) + \frac{b'_{\mathrm{par}}}{2} [\mathrm{R}_{\mathrm{par}}]_{ij} + \xi_{b'0ij}, \tag{63}$$

The coefficients b' and b'_{par} represent the joint regression:

$$\delta 2f(1-f) b_{ij} [\mathrm{R}_{\mathrm{o,par}}]_{ii} \sim \left[\left(\frac{\mathrm{K}_{i\mathrm{p}(j)} + \mathrm{K}_{i\mathrm{m}(j)} - 2\mathrm{K}_0}{1 - \mathrm{K}_0} \right), [\mathrm{R}_{\mathrm{par}}]_{ij} \right] \tag{64}$$

over all pairs i and j. By symmetry, we have

$$\delta 2f(1-f)b_{ji}[\mathrm{R}_{\mathrm{o,par}}]_{jj} = b'\left(\frac{\mathrm{K}_{j\mathrm{p}(i)} + \mathrm{K}_{j\mathrm{m}(i)} - 2\mathrm{K}_0}{1-\mathrm{K}_0}\right) + \frac{b'_{\mathrm{par}}}{2}[\mathrm{R}_{\mathrm{par}}]_{ij} + \xi_{b'1ij}, \tag{65}$$

for some $\xi_{b'1ij}$ such that $\mathrm{Cov}(\xi_{b'1ij}, R_{ij}) = \mathrm{Cov}(\xi_{b'1ij}, [\mathrm{R}_{\mathrm{par}}]_{ij}) = 0$. Therefore,

$$\delta 2f(1-f)[b_{ji}[\mathrm{R}_{\mathrm{o,par}}]_{jj} + b_{ij}[\mathrm{R}_{\mathrm{o,par}}]_{ii}] = b'[\mathrm{R}_{\mathrm{o,par}}]_{ij} + b'_{\mathrm{par}}[\mathrm{R}_{\mathrm{par}}]_{ij} + \xi_{b'ij}, \tag{66}$$

where $\mathrm{Cov}(\xi_{b'ij}, R_{ij}) = \mathrm{Cov}(\xi_{b'ij}, [\mathrm{R}_{\mathrm{o,par}}]_{ij}) = \mathrm{Cov}(\xi_{b'ij}, [\mathrm{R}_{\mathrm{par}}]_{ij}) = 0$.

We can therefore express S_{ij} as:

$$\begin{aligned} S_{ij} =& v_g R_{ij} + (v_{e\sim g} + b + b'_{\mathrm{par}} + \zeta_{\mathrm{par}})[\mathrm{R}_{\mathrm{par}}]_{ij} + (c_{g,e} + b')[\mathrm{R}_{\mathrm{o,par}}]_{ij} + \\ & \xi_{bij} + \xi_{b'ij} + \xi_{\epsilon ij} + \xi_{Gij}. \end{aligned} \tag{67}$$

We can therefore express the vector of lower triangular elements of S as

$$L(S) = X \begin{bmatrix} v_g \\ v_{e\sim g} + b + b'_{\mathrm{par}} + \zeta_{\mathrm{par}} \\ c_{g,e} + b' \end{bmatrix} + \xi_b + \xi_{b'} + \xi_\epsilon + \xi_G, \tag{68}$$

where ξ_b, $\xi_{b'}$, ξ_ϵ, and ξ_G are the vectors, in lower-triangular order for all pairs i, j, of ξ_{bij}, $\xi_{b'ij}$, $\xi_{\epsilon ij}$, and ξ_{Gij} respectively. Therefore, the least squares estimator of θ is

$$\hat{\theta} = \begin{bmatrix} v_g \\ v_{e\sim g} + b + b'_{\mathrm{par}} + \zeta_{\mathrm{par}} \\ c_{g,e} + b' \end{bmatrix} + (X^T X)^{-1}[X^T\xi_b + X^T\xi_{b'} + X^T\xi_\epsilon + X^T\xi_G]. \tag{69}$$

If we take the limit of this as the sample size tends to infinity,

$$\lim_{n\to\infty}\hat{\theta} = \begin{bmatrix} v_g \\ v_{e\sim g} + b + b'_{\mathrm{par}} + \zeta_{\mathrm{par}} \\ c_{g,e} + b' \end{bmatrix} + \mathrm{Cov}(X)^{-1}\lim_{n\to\infty}\frac{2}{n(n-1)}[X^T\xi_b + X^T\xi_{b'} + X^T\xi_\epsilon + X^T\xi_G] \tag{70}$$

In general, $\lim_{n\to\infty}(2/(n(n-1)))[X^T\xi] = [\mathbb{E}[R_{ij}\xi_{ij}], \mathbb{E}[[\mathrm{R}_{\mathrm{par}}]_{ij}\xi_{ij}], \mathbb{E}[[\mathrm{R}_{\mathrm{o,par}}]_{ij}\xi_{ij}]]^T$. Since $\mathbb{E}[R_{ij}] = \mathbb{E}[[\mathrm{R}_{\mathrm{par}}]_{ij}] = \mathbb{E}[[\mathrm{R}_{\mathrm{o,par}}]_{ij}] = 0$,

$$\lim_{n\to\infty}\frac{2}{n(n-1)}[X^T\xi] = [\mathrm{Cov}(R_{ij}, \xi_{ij}), \mathrm{Cov}([\mathrm{R}_{\mathrm{par}}]_{ij}, \xi_{ij}), \mathrm{Cov}([\mathrm{R}_{\mathrm{o,par}}]_{ij}, \xi_{ij})]^T. \tag{71}$$

Since ξ_{bij}, $\xi_{b'ij}$, and $\xi_{\epsilon ij}$ are not correlated with R_{ij}, $[\mathrm{R}_{\mathrm{par}}]_{ij}$ or $[\mathrm{R}_{\mathrm{o,par}}]_{ij}$ (see above), we have

$$\lim_{n\to\infty}\frac{2}{n(n-1)}[X^T\xi_b + X^T\xi_{b'} + X^T\xi_\epsilon] = 0. \tag{72}$$

Furthermore, we have that $\mathbb{E}[\xi_{Gij}|X] = 0$. Therefore, $\mathbb{E}[R_{ij}\xi_{Gij}] = \mathbb{E}[\mathbb{E}[R_{ij}\xi_{Gij}|X]] = \mathbb{E}[R_{ij}\mathbb{E}[\xi_{Gij}|X]] = 0$. Similarly, $\mathbb{E}[[\mathrm{R}_{\mathrm{par}}]_{ij}\xi_{Gij}] = \mathbb{E}[[\mathrm{R}_{\mathrm{o,par}}]_{ij}\xi_{Gij}] = 0$. Therefore,

$$\lim_{n\rightarrow\infty}(2/(n(n-1)))X^T\xi_G = 0. \tag{73}$$

We therefore have

$$\lim_{n\rightarrow\infty}\hat{\theta} = \begin{bmatrix} v_g \\ v_{e\sim g} + b + b'_{\mathrm{par}} + \zeta_{\mathrm{par}} \\ c_{g,e} + b' \end{bmatrix}. \tag{74}$$

□

Remark. We now comment on the assumptions of Theorem 2:

1. *for all pairs i, j, i and j are not related by direct descent.* This is necessary because when i and j are related by direct descent, the segregation events in i and j's parents become dependent upon each other. If parent-offspring (or grandparent-grandchild, etc.) pairs are included in the sample, then this could introduce bias, with the bias becoming larger in proportion to the the proportion of pairs related by direct descent.

2. *For all pairs where $i \neq j$, the residual environment of i is independent of $I(j)$ and the residual environment of j is independent of $I(i)$.* This means that, although the residual environment of i can be correlated with the genotype of j, it cannot be directly affected by it, and vice-versa. It also means that the correlation between the genotype of i and the residual environment of j, and vice-versa, cannot rely directly upon the genetic relatedness of i and j, although it can rely upon the relatedness between the parents of i and the parents of j, and through that be correlated with the relatedness between i and j.

3. *The sample is drawn from a random-mating population.* This is a standard assumption made to derive the relationship between kinship and genetic covariance, which is not sensitive to small deviations from random mating.

4. *genetic variants with direct causal effects are located at random with respect to identity-by-descent sharing.* This assumption may be violated for ascertained samples and when there is assortative mating with respect to the phenotype.

5. *direct effects of inherited genetic variants are additive, including no parent-of-origin effects.* Non-additive direct genetic effects could be incorporated but they may be correlated with the residual environment as only the linear correlation with genetic variation has been regressed out of the environment. Corresponding non-additive associations between the parental genotypes and the environment would have to be considered to guarantee the residual environment is uncorrelated with the non-additive genetic effects.

6. *No gene-environment interaction.* Interaction between genetic variants and environment can complicate heritability estimation by making the effects of genetic and environmental variation difficult to separate.

Although we have derived the theory under these assumptions, the consistency of the estimator of v_g may hold in a wider set of scenarios, at least approximately. For example, when there is population structure, the relationship between IBD sharing and genetic covariance does not hold exactly. However, the conditional independence of offspring genotype and environment given parental genotype still holds, and therefore so does the variance decomposition. The Relatedness Disequilibrium Lemma also still holds. Therefore, any lack of consistency of the estimator when there is population structure does not result from environmental confounding, but from improper calculation of the genetic covariance from IBD sharing, which is likely to be a small deviation except in cases of extremely strong population structure.

While we have shown the consistency of a least-squares regression based estimator of v_g, in practice it is often more statistically efficient to assume the phenotype Y follows a multivariate normal distribution and to fit the variance components by restricted maximum likelihood. The restricted maximum likelihood estimator for v_g and the least-squares estimator of v_g should converge to the same value, however, if $\mathbf{Y}$ has multivariate normal distribution, and so have the same asymptotic properties in terms of bias due to environmental effects.

We now prove consistency with indirect genetic effects when the fraction of pairs with indirect genetic effects on each other tends to zero.

Corollary 2.1. Let ρ_{n1} be the number of pairs i, j, out of a sample of size n, for which $(\epsilon_i \perp \mathrm{I}(j)$ and $\epsilon_j \perp \mathrm{I}(i))$ is not true. Then $\lim_{n\rightarrow\infty} \hat{v}_g = v_g$ if assumptions 1 and 3-6 of Theorem 2 hold and $\rho_{n1} \in o(n^2)$.

Proof. When the genotype of i affects the residual environment of j, or vice-versa, the Relatedness Disequilibrium Lemma (Lemma 1) no-longer holds. We partition X and $L(S)$ into the ρ_{n0} pairs i, j such that $\epsilon_i \perp \mathrm{I}(j)$ and $\epsilon_j \perp \mathrm{I}(i)$, which are represented by X_0 and S_0, and the ρ_{n1} remaining pairs, which are represented by X_1 and S_1: $X^T = [X_0^T \; X_1^T]$, and $L(S)^T = [S_0^T \; S_1^T]$. Applying Theorem 2 within pairs i, j such that $\epsilon_i \perp \mathrm{I}(j)$ and $\epsilon_j \perp \mathrm{I}(i)$, we have that

$$S_0 = X_0\theta_0 + \xi_0, \; \theta_0 = \begin{bmatrix} v_g \\ v_{e\sim g} + b + b'_{\text{par}} + \zeta_{\text{par}} \\ c_{g,e} + b' \end{bmatrix}, \tag{75}$$

for some ξ_0 such that $\lim_{n\rightarrow\infty} \rho_{n0}^{-1} X_0^T \xi_0 = 0$. For the remaining ρ_{n1} pairs, we have

$$S_1 = X_1(\theta_0 + b_1) + \xi_1, \tag{76}$$

for some ξ_1 such that $\lim_{n\to\infty} \rho_{n1}^{-1} X_1^T \xi_1 = 0$. This allows for some additional bias in the estimation of θ arising from the breaking of the Relatedness Disequilibrium Lemma (Lemma 1). Our overall estimate of θ therefore becomes

$$\hat{\theta} = \theta_0 + (X^T X)^{-1} X_1^T X_1 b_1 + (X^T X)^{-1}[X_0^T \xi_0 + X_1^T \xi_1]. \tag{77}$$

Let $S_n = 2(X^T X)/(n(n-1))$ and let $S_{n1} = X_1^T X_1 / \rho_{n1}$, then

$$\hat{\theta} = \theta_0 + \frac{2\rho_{n1}}{n(n-1)} S_n^{-1} S_{n1} b_1 + (X^T X)^{-1}[X_0^T \xi_0 + X_1^T \xi_1]. \tag{78}$$

There is therefore a bias term proportional to $2\rho_{n1}/(n(n-1))$, the fraction of pairs i, j for which ($\epsilon_i \perp \mathrm{I}(j)$ and $\epsilon_j \perp \mathrm{I}(i)$) is not true for a sample of size n. Taking the limit,

$$\lim_{n\to\infty} \hat{\theta} = \theta_0 + \left(\lim_{n\to\infty} \frac{2\rho_{n1}}{n(n-1)} \right) \mathrm{Cov}(X)^{-1} \mathrm{Cov}(X_1) b_1. \tag{79}$$

We therefore have that

$$\lim_{n\to\infty} \hat{v}_g = v_g \text{ if } \lim_{n\to\infty} \frac{2\rho_{n1}}{n(n-1)} = 0. \tag{80}$$

□

Remark. In this, we have assumed that

$$\lim_{n\to\infty} S_n^{-1} S_{n1} = \mathrm{Cov}(X)^{-1} \mathrm{Cov}(X_1) < \infty. \tag{81}$$

It is, however possible, that $S_n^{-1} S_{n1} \to \infty$, and then the bias may not disappear. This could be the case if there is only relatedness disequilibrium (variation in R_{ij} that is independent of $[\mathrm{R}_{\mathrm{par}}]_{ij}$ and $[\mathrm{R}_{\mathrm{o,par}}]_{ij}$) for the ρ_{n1} pairs with indirect effects on each other. Furthermore, if the relatedness disequilibrium is large for pairs with indirect genetic effects on each other, which would be the case if those pairs are siblings, then the magnitude of the bias will reflect this.

A case where the estimator could not be consistent would be when the genotype of each individual affects the residual environment of every other individual in the population. This may apply to certain traits that depend upon the entire network of individuals in a population.

2 GREML-SNP and RDR-SNP

In this section, we first construct the RDR-SNP covariance model, then we analyse the standard GREML-SNP method in the more general RDR-SNP covariance model.

In this section, we consider the causal SNPs as known and the genotypes at the causal SNPs as observed. This is to highlight the environmental bias properties of RDR-SNP and GREML-SNP. In reality, the causal SNPs are unknown and possibly unobserved. Using SNPs other than the causal SNPs to estimate relatedness matrices can introduce bias to heritability estimates[4].

2.1 RDR-SNP

We model a vector of phenotype observations for n probands for a trait with l bi-allelic SNPs with direct effects. Let X be the $[n \times l]$ matrix of standardised proband genotypes at the l causal SNPs, with

$$[X]_{ij} = \frac{g_{ij} - 2f_j}{\sqrt{2f_j(1-f_j)}}, \tag{82}$$

where $g_{ij} \in \{0, 1, 2\}$ is the count of one of the alleles of SNP j in proband i, and f_j is the frequency of the allele. In general, $\mathbb{E}[[X]_{ij}] = 0$. In an infinite, outbred, random-mating population, $\mathrm{Var}([X]_{ij}) = 1$.

Let X_p be the $[n \times l]$ matrix of standardised genotypes of the probands' fathers, with

$$[X_\mathrm{p}]_{ij} = \frac{g_{\mathrm{p}(i)j} - 2f_j}{\sqrt{2f_j(1-f_j)}}, \tag{83}$$

where $g_{\mathrm{p}(i)j} \in \{0, 1, 2\}$ is the count of one of the alleles of SNP j in the father of proband i, and f_j is the frequency of the allele. In general, $\mathbb{E}[[X_\mathrm{p}]_{ij}] = 0$. In an infinite, outbred, random-mating population, $\mathrm{Var}([X_\mathrm{p}]_{ij}) = 1$.

We define X_m to be the equivalent matrix of standardised genotypes of the probands' mothers. We then define the parental genotype matrix as $X_\mathrm{par} = X_\mathrm{p} + X_\mathrm{m}$. In general, $\mathbb{E}[[X_\mathrm{par}]_{ij}] = 0$. In an infinite, outbred, random-mating population, $\mathrm{Var}([X_\mathrm{par}]_{ij}) = 2$.

A vector of phenotype observations, $\mathbf{Y}$, can be written as the sum of a direct, genetic component and an environmental component:

$$\mathbf{Y} = X\delta + \mathbf{e}, \tag{84}$$

where δ is an l-vector of appropriately scaled direct effects of the l causal SNPs, and e is an n-vector of environmental effects, including noise. Let the genotypes of the parents of the probands be G_par. We have that $X \perp e | G_\mathrm{par}$ and $\mathbb{E}[X|G_\mathrm{par}] = X_\mathrm{par}/2$. Therefore, by the Conditional Independence Lemma (Appendix B):

$$\mathbf{Y} = \mu + X\delta + X_\mathrm{par}\eta + \epsilon, \tag{85}$$

for some constant μ, and for some ϵ that is uncorrelated with both proband and parental genotype and has expectation zero. If no other sources of correlation between proband genotype and $\mathbf{e}$ are present, then η is the vector of parental genetic nurturing effects of the l causal loci.

2.1.1 Random Effects Model

In RDR, we consider the effects as fixed and the genotypes at the causal loci as unknown. In GREML-SNP, genotypes at causal loci are treated as observed, and effects are treated as drawn from a normal distribution. We develop RDR-SNP to be analogous to typical GREML-SNP analysis. We therefore treat X and X_{par} as observed and model δ and η as drawn from a normal distribution:

$$\mathbf{Y} = X\delta + X_{\text{par}}\eta + \epsilon; \begin{bmatrix} \delta \\ \eta \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \frac{v_g}{l}\mathrm{I}_l & \frac{c_{\mathrm{g,e}}}{2l}\mathrm{I}_l \\ \frac{c_{\mathrm{g,e}}}{2l}\mathrm{I}_l & \frac{v_{e\sim g}}{2l}\mathrm{I}_l \end{bmatrix}\right). \tag{86}$$

Here, v_g is the variance explained by the direct effects of the l SNPs, $v_{e\sim g}$ is the variance explained by regression on parental genotypes at the l SNPs, and $c_{g,e}$ is the total covariance between the direct effects of the l SNPs and the environmental component of the trait. We make the further assumption that $\epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon \mathrm{I}_n)$ and that ϵ is independent of δ and η.

We then have that:

$$\mathbf{Y}|X, X_{\text{par}} \sim \mathcal{N}(\mu, v_g \mathrm{R}^{\text{snp}} + v_{e\sim g}\mathrm{R}^{\text{snp}}_{\text{par}} + c_{g,e}\mathrm{R}^{\text{snp}}_{\text{o,par}} + \sigma^2_\epsilon \mathrm{I}_n), \tag{87}$$

where

$$\mathrm{R}^{\text{snp}} = \frac{XX^T}{l}; \quad \mathrm{R}^{\text{snp}}_{\text{par}} = \frac{X_{\text{par}}X^T_{\text{par}}}{2l}; \quad \mathrm{R}^{\text{snp}}_{\text{o,par}} = \frac{XX^T_{\text{par}} + X_{\text{par}}X^T}{2l}. \tag{88}$$

It is straightforward to prove that $\mathbb{E}[[\mathrm{R}^{\text{snp}}]_{ii}] = 1$, $\mathbb{E}[[\mathrm{R}^{\text{snp}}_{\text{par}}]_{ii}] = 1$, and $\mathbb{E}[[\mathrm{R}^{\text{snp}}_{\text{o,par}}]_{ii}] = 1$ for $i = 1, \ldots, n$ in an infinite, outbred, random-mating population. Given that the assumptions of the random effects model are satisfied, the variance parameters v_g, $v_{e\sim g}$, and $c_{g,e}$ can be interpreted in the same way as in the RDR variance decomposition (1.1).

2.1.2 SNP relatedness disequilibrium

The matrices R^{snp}, $\mathrm{R}^{\text{snp}}_{\text{par}}$, and $\mathrm{R}^{\text{snp}}_{\text{o,par}}$ are the SNP equivalents of the matrices R, R_{par}, and $\mathrm{R}_{\text{o,par}}$, calculated from genome-wide IBD sharing. It is natural to ask whether the relationships between R^{snp}, $\mathrm{R}^{\text{snp}}_{\text{par}}$, and $\mathrm{R}^{\text{snp}}_{\text{o,par}}$ mirror the relationships between R, R_{par}, and $\mathrm{R}_{\text{o,par}}$.

It is straightforward to see that, conditional on parental genotypes, any variation in R^{snp} must be due to random Mendelian segregations in the parents of the probands. It is also fairly straightforward to prove that, for a pair i, j not related by direct descent, $\mathbb{E}[[\mathrm{R}^{\text{snp}}]_{ij}|G_{\text{par}}] = [\mathrm{R}^{\text{snp}}_{\text{par}}]_{ij}/2$. This implies that a version of the Relatedness Disequilibrium

Lemma (Lemma 1) applies for RDR-SNP, and that RDR-SNP should therefore have similar environmental bias properties to RDR.

While RDR and RDR-SNP may have similar environmental bias properties, RDR-SNP makes assumptions about the distribution of effect sizes that RDR does not. However, RDR-SNP does not make the assumptions about random-mating that RDR does in order to derive the phenotypic covariance matrix.

2.2 GREML-SNP

GREML-SNP only uses the halves of the parental genomes transmitted to probands. The parental genotypes can also be written as $X_{\text{par}} = X + X_{\text{NT}}$, where $X_{\text{NT}} = X_{\text{par}} - X$ contains the genotypes of the halves of the parental genomes not transmitted to the offspring. We can therefore write the phenotype as

$$\mathbf{Y} = X(\delta + \eta) + X_{\text{NT}}\eta + \epsilon. \tag{89}$$

The phenotypic covariance matrix is therefore

$$\text{Cov}(\mathbf{Y}) = (v_g + v_{e\sim g}/2 + c_{g,e})\text{R}^{\text{snp}} + (v_{e\sim g}/2)\text{R}^{\text{snp}}_{\text{NT}} + (c_{g,e} + v_{e\sim g})\text{R}^{\text{snp}}_{\text{TNT}} + \sigma^2_\epsilon \text{I}_n, \tag{90}$$

where

$$\text{R}^{\text{snp}}_{\text{NT}} = \frac{X_{\text{NT}} X_{\text{NT}}^T}{l}, \ \text{R}^{\text{snp}}_{\text{TNT}} = \frac{X X_{\text{NT}}^T + X_{\text{NT}} X^T}{2l}. \tag{91}$$

Standard GREML-SNP analysis fits the following model

$$\mathbf{Y} \sim \mathcal{N}(\mu, v_g \text{R}^{\text{snp}} + \sigma^2_\epsilon \text{I}_n). \tag{92}$$

In an infinite, outbred, random-mating population, the non-transmitted parts of the parental genomes are independent from the transmitted parts. Furthermore, for unrelated pairs, the elements of R^{snp} will be uncorrelated with the elements of $\text{R}^{\text{snp}}_{\text{NT}}$ and $\text{R}^{\text{snp}}_{\text{TNT}}$. Therefore, assuming that $\text{Cov}(\epsilon) = \sigma^2_\epsilon \text{I}_n$, GREML-SNP estimates of v_g will tend towards $v_g + v_{e\sim g}/2 + c_{g,e}$ in a sample of unrelated individuals from an infinite, outbred, random-mating population. When the sample includes close relative pairs, further bias could be introduced. If $\text{Cov}(\epsilon) \neq \sigma^2_\epsilon \text{I}_n$, this could introduce further bias to GREML-SNP estimates, especially when due to population stratification. Assortative mating and other forms of non-random mating could further introduce bias by inducing correlations between transmitted and non-transmitted parts of the parental genomes.

3 Variance of least-squares regression estimators

In this section and in Appendix D, we build on previous work expressing the Haseman-Elston Regression as a quadratic form[35] to derive simple formulae for variances and covariances of variance component estimates.

3.1 RELT-SNP as a modified Haseman-Elston Regression

The Haseman-Elston Regression is a simple way to estimate v_g by regressing elements of the sample phenotypic covariance matrix, $S = (\mathbf{y} - \bar{y})(\mathbf{y} - \bar{y})^T$, onto corresponding elements of the relatedness matrix. Here, we are interested in calculating the variance of the RELT-SNP estimator, so the relatedness matrix is $\mathrm{R}^{\mathrm{snp}}$. However, the results would apply to any relatedness matrix.

The Haseman-Elston estimator for v_g based on $\mathrm{R}^{\mathrm{snp}}$ is

$$\hat{v}_g = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} S_{ij} [\mathrm{R}^{\mathrm{snp}}]_{ij}}{\sum_{i=1}^{n} \sum_{j=i+1}^{n} [\mathrm{R}^{\mathrm{snp}}]_{ij}^2}. \tag{93}$$

For the following results, it is useful to define the Frobenius norm of a $[n \times m]$ matrix A:

$$||A||_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} [A]_{ij}^2}. \tag{94}$$

Let us define the $[n \times n]$ matrix R^-:

$$\mathrm{R}_{ij}^- = \frac{\mathrm{R}_{ij}^{\mathrm{snp}}}{\sqrt{||\mathrm{R}^{\mathrm{snp}}||_F^2 - \sum_{i=1}^{n} [\mathrm{R}^{\mathrm{snp}}]_{ii}^2}} \text{ for } i \neq j, \text{ and } \mathrm{R}_{ii}^- = 0 \, \forall \, i. \tag{95}$$

In other words, R^- is $\mathrm{R}^{\mathrm{snp}}$ with the diagonal elements set to zero, and normalised so that $||\mathrm{R}^-||_F^2 = 1$. Then we equivalently have

$$\hat{v}_g = (\mathbf{y} - \bar{y})^T \mathrm{R}^- (\mathbf{y} - \bar{y}). \tag{96}$$

This is a quadratic form. We derive a simple formula for the variance of a quadratic form in normal random variables in Appendix D. Assuming that $\mathbf{Y} \sim \mathcal{N}(0, \Sigma)$,

$$\mathrm{Var}(\hat{v}_g) = 2 \, \mathrm{Tr}(\mathrm{R}^- \Sigma \mathrm{R}^- \Sigma) = 2||\mathrm{R}^- \Sigma||_F^2 \tag{97}$$

For the RELT-SNP estimator, we simply modify $\mathrm{R}^{\mathrm{snp}}$ by setting all off-diagonal elements with entries greater than the threshold to zero, so that they do not contribute to the estimate of v_g.

3.2 Variance of heritability estimate

Let us define the sample variance of Y, $S_y^2 = (\mathbf{y} - \bar{y})^T(\mathbf{y} - \bar{y})/(n-1)$. Heritability can be estimated by

$$\hat{h}^2 = \frac{(\mathbf{y} - \bar{y})^T \mathrm{R}^- (\mathbf{y} - \bar{y})}{S_y^2}. \tag{98}$$

We can calculate the covariance between the numerator and denominator by realising that S_y^2 is also a quadratic form, and using a formula we derive for the covariance between two quadratic forms (Appendix D.1):

$$\mathrm{Cov}(\hat{v}_g, S_y^2) = \frac{2\,\mathrm{Tr}(\mathrm{R}^-\Sigma^2)}{n-1} = \frac{2\langle \mathrm{R}^-\Sigma, \Sigma\rangle_F}{n-1}, \tag{99}$$

where $\langle A, B\rangle_F$ is the Frobenius inner product between two matrices A and B.

To approximate the variance of $\hat{h}^2$, we also need the variance of S_y^2:

$$\mathrm{Var}(S_y^2) = 2\,\mathrm{Tr}(\Sigma^2)/(n-1)^2 = 2||\Sigma||_F^2/(n-1)^2. \tag{100}$$

We then approximate the variance of the heritability estimator using a first-order Taylor expansion:

$$\mathrm{Var}(\hat{h}^2) \approx \frac{\hat{v}_g^2}{(S_y^2)^2}\left(\frac{\mathrm{Var}(\hat{v}_g)}{\hat{v}_g^2} - 2\frac{\mathrm{Cov}(\hat{v}_g, S_y^2)}{\hat{v}_g S_y^2} + \frac{\mathrm{Var}(S_y^2)}{(S_y^2)^2}\right). \tag{101}$$

3.3 Variance of RDR least-squares estimator

The above procedure for computing the variance of the RELT-SNP estimator can be applied to compute the variance of the RDR least-squares estimator. The RDR least-squares estimator regresses elements of the sample phenotypic covariance matrix jointly onto elements of R, $\mathrm{R}_{\mathrm{par}}$ and $\mathrm{R}_{\mathrm{o,par}}$. The estimator for v_g can be expressed as

$$\hat{v}_g = (\mathbf{y} - \bar{y})^T \mathrm{R}_d (\mathbf{y} - \bar{y}), \tag{102}$$

where R_d is a linear combination of R, $\mathrm{R}_{\mathrm{par}}$ and $\mathrm{R}_{\mathrm{o,par}}$, with diagonal elements set to zero and parent-offspring pairs and grandparent-grandchild pairs set to zero. This enables estimation of the standard error of the RDR least-squares estimator based on the above results.

To calculate R_d, first set diagonal elements and parent-offspring and grandparent-grandchild pairs to zero for R, $\mathrm{R}_{\mathrm{par}}$ and $\mathrm{R}_{\mathrm{o,par}}$. As in Equation 41, let

$$X = [L(\mathrm{R}), L(\mathrm{R}_{\mathrm{par}}), L(\mathrm{R}_{\mathrm{o,par}})], \tag{103}$$

where $L(A)$ gives the vector of lower-triangular elements of a square matrix A. Further, let $C = (X^T X)^{-1}$, then

$$\mathrm{R}_d = C[1,1]\mathrm{R} + C[1,2]\mathrm{R}_{\mathrm{par}} + C[1,3]\mathrm{R}_{\mathrm{o,par}}. \tag{104}$$

3.4 Dealing with an unknown phenotypic covariance matrix

In real applications, the phenotypic covariance matrix Σ is unknown. We replace Σ with an estimate of Σ from the model: $\Sigma = v\mathrm{R} + \sigma^2\mathrm{I}$. We compute estimates of v, $\hat{v}$, by a standard Haseman-Elston regression without excluding any relative pairs, and we compute estimates of σ^2, $\hat{\sigma}^2$, by subtracting $\hat{v}$ from the estimated phenotypic variance. We then use $\hat{\Sigma} = \hat{v}\mathrm{R} + \hat{\sigma}^2\mathrm{I}$ in place of Σ in the formulae derived above.

4 Trait Simulations

4.1 Simulations using Icelandic data

For all traits other than the ‘rare SNPs’ trait, we used imputed genotypes at SNPs from the Illumina Framework SNP set (Online Methods). We filtered the SNPs so that the minimum imputation information was 0.9999, removing around half of the SNPs. Out of the remaining SNPs passing the filter, we randomly sampled 10,000 SNPs to use as the causal SNPs in our simulations. In the 10,000 selected SNPs, the median imputation information was 1.0000, the minimum minor allele frequency (MAF) was 0.52%, and the median MAF was 22.8%. For the ‘rare SNPs’ trait, we randomly sampled SNPs from all imputed SNPs with MAF between 1% and 0.1% and with imputation information at least 0.9999 and p-value for Hardy-Weinberg deviation greater than 0.05. We sampled 100 such SNPs from each chromosome, giving 2,200 SNPs in total. For each type of trait, we simulated 500 independent replicates.

Each trait had a direct, additive genetic component that explained 40% of the phenotypic variance, which we describe the simulation of here. Apart from for the ‘rare SNPs’ trait, we standardised genotypes so that each SNPs genotype vector had sample mean zero and sample variance one. Let G represent the matrix of standardised genotypes at the 10,000 causal SNPs. We sampled additive effects of SNPs from a normal distribution. Let β represent the vector of SNP effects. The additive genetic component, A, was then calculated as $G\beta$. The noise component was simulated as $\epsilon \sim \mathcal{N}(0, \mathrm{I})$. The additive genetic component was scaled to have sample variance 1. The additive phenotype was then simulated as:

$$\mathbf{Y} = \sqrt{0.4}A + \sqrt{0.6}\epsilon. \tag{105}$$

The same process was used for the ‘rare SNPs’ trait, except genotypes were not standardised because the standardisation becomes highly sensitive to fluctuations in estimated allele frequencies for rare SNPs.

For the ‘epistatic’ trait, we simulated a genetic component due to pairwise interactions between SNPs. To do this, we sampled 100 SNPs from the 10,000 SNPs given additive effects. We formed pairwise interaction variables for all pairs of SNPs from the 100 selected SNPs by multiplying the standardised genotypes of each pair of SNPs together. Let G_{epi} be the resulting matrix of SNP-SNP interaction variables. We standardised the columns of G_{epi} so that each column had sample mean zero and sample variance one. We simulated interaction effects from a normal distribution, $\beta_{\text{epi}} \sim \mathcal{N}(0, \mathrm{I})$, and we formed the pairwise interaction genetic component as $A_{\text{epi}} = G_{\text{epi}}\beta_{\text{epi}}$, and standardised A_{epi} so that it had sample mean zero and sample variance one. The epistatic trait was then formed as:

$$\mathbf{Y} = \sqrt{0.4}A + \sqrt{0.1}A_{\text{epi}} + \sqrt{0.5}\epsilon. \tag{106}$$

For the regional trait, we gave each of the 22 regions of Iceland (called syslas) a different normally distributed effect, and we scaled the overall variance explained by variation in sysla to be 20% of the phenotypic variance. The phenotype was thus simulated as:

$$\mathbf{Y} = \sqrt{0.4}A + \text{sysla} + \sqrt{0.5}\epsilon, \tag{107}$$

where 'sysla' represents the vector of sysla effects.

For the 'maternal environment' trait, we added an environmental effect that was shared between individuals who shared mothers according to the deCODE genealogy database. The effect due to each mother was drawn from a normal distribution, and resulting vector of effects due to maternal environment, M, was scaled to have variance 0.4. The phenotype was simulated as:

$$\mathbf{Y} = \sqrt{0.4}A + \mathrm{M} + \sqrt{0.5}\epsilon. \tag{108}$$

Note, if two individuals had the same mother in the deCODE genealogy database, then their maternal environment variables were the same.

For the 'genetic nurturing' trait, we simulated a component reflecting parental genetic nurturing effects: each genetic variant in the parents was also given an additive effect on the trait of the proband. The additive genetic component was $A = G\beta$, scaled to have variance 1. Let G_{par} be the matrix of standardised parental genotypes, where the parental genotype is defined as the sum of the mother's genotype and the father's genotype. Then the genetic nurturing component was simulated as $A_{\text{par}} = G_{\text{par}}\beta$, scaled to have sample variance 1. This implies that the parental genetic nurturing effects differ only by a constant scale factor from the direct effect of the genetic variant in the offspring. The phenotype was then simulated as

$$\mathbf{Y} = \sqrt{0.4}A + \sqrt{0.1}A_{\text{par}} + \sqrt{0.5 - \sqrt{0.08}}\epsilon, \tag{109}$$

where the scale factor for the residual variance was calculated so that the total phenotypic variance, which includes the covariance between and A and A_{par}, was 1. For this trait, $v_{e \sim g} = 0.1$ and $c_{g,e} = \sqrt{0.08} \approx 0.283$.

4.2 Simulations in the UK Biobank

To simulate GREML-SNP inference on a sample of distantly individuals, we identified a subset of the genotyped individuals in the UK Biobank who also had both of their parents genotyped. To identify parent-offspring pairs, we used the kinship table provided by UK Biobank that identifies all pairs with third degree relationship or closer[20]. As in the UK Biobank documentation[20], we determined parent-offspring pairs as those pairs that were in the kinship table but had IBS0 < 0.0012 and kinship less than $2^{-\frac{3}{2}}$. Of the remaining pairs, we checked whether the recorded ages at recruitment were consistent with a parent-offspring relationship. We removed pairs that had a recorded age difference of less than 11, and we determined which of the pair was the parent by whichever was older. We then filtered based on sample quality control metrics provided by UK Biobank[20]: removing those with a putative sex chromosome aneuploidy, excess relatives, and those with outlying heterozygosity. To remove the influence of population structure, we restricted the sample to those identified by UK Biobank as having been born in Britain and having predominantly British ancestry[20]. From the remaining parent-offspring pairs, we identified 973 individuals with both parents genotyped.

To estimate relatedness in the set of 973 individuals with both parents genotyped, we used all genotyped SNPs[20] with minor allele frequency (MAF) greater than 5% and missingness less than 1%. We identified 35 pairs with estimated relatedness greater than 0.05. We removed one of each pair at random, leaving 938 individuals. We removed one further individual after discovering evidence that there may have been a sample duplication of one of the individuals parents, leading us to spuriously infer both parents were genotyped. This gave a final sample of 937.

To select causal SNPs for phenotype simulation, for each chromosome we randomly sampled 1,500 SNPs then removed those with MAF less than 5% or more than 0.5% missing genotypes. This gave a set of 11,771 SNPs. We mean imputed missing genotypes for both parents and offspring.

We simulated 10,000 independent traits determined by additive genetic effects and noise. Let $l = 11,771$. We standardised offspring genotypes so that the genotypes at each SNP had mean zero and variance 1. Let G be the matrix of standardised offspring genotypes. For each trait, we simulated a normally distributed vector of effects for the l SNPs: $\beta \sim \mathcal{N}(0, 0.2l^{-1}\mathrm{I})$. The additive genetic component of the trait, A, was then calculated as $A = G\beta$. The noise component was simulated as $\epsilon \sim \mathcal{N}(0, 0.8\mathrm{I})$. The 'additive' trait was simulated as $Y = A + \epsilon$.

We simulated 10,000 independent traits with both additive and genetic nurturing effects. In addition to an additive genetic component simulated as above, we also simulated a genetic nurture component. We formed parental genotypes by summing the unnormalised genotype matrices for the mothers and fathers, and then we standardised parental genotypes to have mean zero and variance 2. In an outbred population, the variance for the parental genotypes is naturally twice that of the offspring genotype as it is the sum of maternal and paternal genotypes. Let G_{par} represent the matrix of standardised parental genotypes. The genetic nurturing component of the trait, A_{par}, was then calculated as $A_{\mathrm{par}} = G_{\mathrm{par}}\beta/3$, where β is the same vector of effects as for the direct, additive component, $A = G\beta$. To make the phenotypic variance approximately one, the noise component was simulated as $\epsilon \sim \mathcal{N}(0, 0.6222\mathrm{I})$. The trait with genetic nurturing effects was then simulated as

$$Y = A + A_{\text{par}} + \epsilon. \tag{110}$$

The relatedness matrices R^{snp}, $\mathrm{R}^{\text{snp}}_{\text{par}}$, $\mathrm{R}^{\text{snp}}_{\text{o,par}}$ were computed from the 11,771 causal SNPs as outlined in Section 2 and the Online Methods. GREML-SNP and RDR-SNP analysis was performed on the traits using unconstrained restricted maximum likelihood in GCTA[34].

4.2.1 Comparing RELT-SNP and GREML-SNP

We computed both GREML-SNP and RELT-SNP estimates for the simulated traits in the UK Biobank. For the trait determined only by additive, direct effects and noise, the mean RELT-SNP heritability estimate was 19.78% (0.16% S.E.), close to the true heritability, 20%, and the mean GREML-SNP estimate, 19.76% (0.15% S.E.). For the trait determined by both direct genetic effects and parental genetic nurturing effects, the mean RELT-SNP heritability estimate was 35.16% (0.16% S.E.), almost exactly the same as the mean GREML-SNP estimate, 35.15% (0.16% S.E.). These results show that GREML-SNP and RELT-SNP estimates exhibit the same bias from genetic nurturing effects.

5 Sensitivity analysis of heritability results

We performed analyses to test whether our results were driven by atypical properties of the sample with both parents genotyped or by differences between the regions of Iceland.

5.1 Comparison to full genotyped sample

To test whether the subsample with both parents genotyped was atypical, we also computed Kinship F.E. heritability estimates in a random subsample from the set of all genotyped Icelanders in our data (Supplementary Table 5). Kinship F.E. estimates were slightly higher on average (mean difference 1.2%) in the random subsample than in the subsample with both parents genotyped. The difference was small for most traits. However, for height, the Kinship F.E. estimate was 12% higher in the random subsample, and for educational attainment, the Kinship F.E. estimate was 6.9% lower in the random subsample. Even though the Kinship F.E. estimate for educational attainment was 6.9% lower in the random sample, it was still significantly different from the RDR estimate in the sample with both parents genotyped (difference=28.5%, $p < 1.5 \times 10^{-3}$). These results argue that overestimation of heritability by the Kinship F.E. method is not a consequence of atypical properties of the sample with both parents genotyped.

5.2 Controlling for regional differences

Controlling for mean differences between regions can reduce confounding in Kinship type methods. We checked whether RDR and Kinship F.E. estimates changed much when we adjusted traits for mean differences between regions of Iceland (called syslas) (Supplementary Table 5). Kinship F.E. estimates reduced by 0.5% of the phenotypic variance on average, with a max reduction of 1.8% for educational attainment. RDR estimates reduced by 0.1% on average, with a max reduction of 1.3% for educational attainment. These results show our analysis is robust to controlling for mean differences between regions. However, we chose not to control for region (sysla) in our main analysis because controlling for regional differences removes part of the phenotypic variation, which could bias heritability estimates for the overall population.

6 References

4 Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of SNP-based heritability. Nat. Genet. 49, (2017).

5 Visscher, P. M. et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2, e41 (2006).

14 Kong, A. et al. The nature of nurture: Effects of parental genotypes. Science 359, 424428 (2018).

19 Pedersen, N. L., Lichtenstein, P. & Svedberg, P. The Swedish Twin Registry in the Third Millennium. Twin Res. 5, 427432 (2002).

20 Bycroft, C. et al. Genome-wide genetic data on 500,000 UK Biobank participants. bioRxiv (2017). doi:10.1101/166298

23 Branigan, A. R., Mccallum, K. J. & Freese, J. Variation in the heritability of educational attainment: An international meta-analysis. Soc. Forces 92, 109140 (2013).

33 Young, A. I. & Durbin, R. Estimation of Epistatic Variance Components and Heritability in Founder Populations and Crosses. Genetics 198, 14051416 (2014).

34 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 7682 (2011).

35 Tamar, S. Confidence intervals for heritability via Haseman-Elston regression. Statistical Applications in Genetics and Molecular Biology 16, 259 (2017).

36 Carlsson, S., Ahlbom, A., Lichtenstein, P. & Andersson, T. Shared genetic influence of BMI, physical activity and type 2 diabetes: A twin study. Diabetologia 56, 10311035 (2013).

37 Silventoinen, K. et al. Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res. 6, 399408 (2003).

38 Baker, J. H., Thornton, L. M., Bulik, C. M., Kendler, K. S. & Lichtenstein, P. Shared genetic effects between age at menarche and disordered eating. J. Adolesc. Heal. 51, 491496 (2012).

39 Rahman, I. et al. Genetic dominance influences blood biomarker levels in a sample of 12,000 Swedish elderly twins. Twin Res. Hum. Genet. 12, 286294 (2009).

40 Arpegrd, J. et al. Comparison of heritability of Cystatin C- and creatinine-based estimates of kidney function and their relation to heritability of cardiovascular disease. J. Am. Heart Assoc. 4, e001467 (2015).

A Indirect genetic effects

When the genotype of one individual affects the environment of another, i.e. when there is an 'indirect genetic effect', this complicates separation of genetic and environmental components of the covariance between the pair of individuals. In particular, for a pair i, j, where the genotype of i affects the environment of j and vice-versa, what we term a 'reciprocal genetic effect', this implies a component of the covariance between the environments of i and j is proportional to the genetic relatedness between i and j, making it impossible to distinguish from the covariance due to directly inherited genetic variants. These reciprocal genetic effects are expected to be strongest for close relatives of the same generation and other pairs of individuals that have many opportunities for social interaction during development.

When there are reciprocal genetic effects between siblings, this can create large biases for heritability estimation methods based on sibling pairs, such as Sib-Regression[5] and twin studies.

To illustrate the problems introduced by reciprocal genetic effects between relatives, we consider a simple model where the genotype of one sibling, i, directly affects the phenotype of the other, Y_i^{sib}:

$$Y_i = \delta g_i + \alpha g_i^{\text{sib}} + \xi_i, \; Y_i^{\text{sib}} = \delta g_i^{\text{sib}} + \alpha g_i + \xi_i^{\text{sib}}; \tag{111}$$

where ξ_i^{sib} and ξ_i are independent of g_i and g_i^{sib}.

A.1 Covariance between siblings

The covariance between the siblings' phenotypes is:

$$\text{Cov}(Y_i, Y_i^{\text{sib}}) = (v_g + \alpha^2 2f(1-f))\text{R}_{i\text{sib}(i)} + \delta\alpha 2f(1-f)[\text{R}_{ii} + \text{R}_{\text{sib}(i)\text{sib}(i)}] + \text{Cov}(\xi_i, \xi_i^{\text{sib}}), \tag{112}$$

where $\text{R}_{i\text{sib}(i)}$ is the relatedness between the siblings, and $\text{R}_{\text{sib}(i)\text{sib}(i)}$ is the relatedness of the sibling of i with itself. Let $v_r = \alpha^2 2f(1-f)$.

While this model is too simple to be realistic, it demonstrates that a method that looks only at the relatedness between siblings, such as Sib-Regression, will be expected to overestimate v_g in proportion to v_r.

A.2 Effect on RDR estimator of heritability

We give a proof (Theorem 2 and Corollary 2.1) that, assuming the proportion of pairs in the sample exhibiting indirect genetic effects tends to zero as sample size grows, the RDR estimator is consistent. In this section, we complement this by giving intuition on the effect of a specific model of reciprocal genetic effects on the RDR estimator of heritability.

The covariance between an arbitrary pair of individuals in this model would be

$$\mathrm{Cov}(Y_i, Y_j) = v_g \mathrm{R}_{ij} + \delta\alpha 2f(1-f)[\mathrm{R}_{i\mathrm{sib(j)}} + \mathrm{R}_{j\mathrm{sib(i)}}] + v_r \mathrm{R}_{\mathrm{sib}(i)\mathrm{sib}(j)} + \mathrm{Cov}(\xi_i, \xi_j), \quad (113)$$

where $\mathrm{R}_{\mathrm{sib}(i)\mathrm{sib}(j)}$ is the additive relatedness between the sibling of i and the sibling of j.

Over all pairs that are not siblings, R_{ij} would be correlated with $\mathrm{R}_{\mathrm{sib}(i)\mathrm{sib}(j)}$. However, this correlation would be entirely mediated through the relatedness between the parents of i and the parents of j (Relatedness Disequilibrium Lemma (1)). A similar argument can be made that, although R_{ij} and $\mathrm{R}_{i\mathrm{sib}(j)} + \mathrm{R}_{j\mathrm{sib}(i)}$ would be correlated over all pairs, this would be mediated through the relatedness between i and the parents of j and j and the parents of i, which is captured by $[\mathrm{R}_{\mathrm{o,par}}]_{ij}$ (there is a similar but more formal argument in the proof of consistency (Theorem 2)). Therefore, fitting the relatedness disequilibrium covariance model, which jointly fits R_{ij}, $[\mathrm{R}_{\mathrm{o,par}}]_{ij}$, and $[\mathrm{R}_{\mathrm{par}}]_{ij}$, would not result in any bias in the estimation of v_g over non-sibling pairs.

It would be expected that any bias in the RDR estimator of v_g from sibling reciprocal genetic effects would be in proportion to the proportion of sibling pairs in the sample.

B Conditional Independence Lemma

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

Lemma 3. 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,

$$\mathbb{E}[x|z] = \alpha + ba^T z, \tag{114}$$

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

$$\text{Cov}(x, r) = 0, \text{ where } r = y - \frac{\text{Cov}(y, a^T z)}{\text{Var}(a^T z)} a^T z \tag{115}$$

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,

$$b = \frac{\text{Cov}(x, a^T z)}{\text{Var}(a^T z)}, \tag{116}$$

so

$$\text{Cov}(x, r) = \text{Cov}(y, x) - \text{Cov}(y, a^T z) b. \tag{117}$$

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

By the Law of Total Covariance

$$\text{Cov}(y, x) = \mathbb{E}_z[\text{Cov}(y, x|z)] + \text{Cov}_z(\mathbb{E}[x|z], \mathbb{E}[y|z]) = \text{Cov}_z(\mathbb{E}[x|z], \mathbb{E}[y|z]), \tag{118}$$

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

$$\text{Cov}(y, x) = \text{Cov}_z \left(\alpha + ba^T z, \mathbb{E}[y|z] \right) = \text{Cov}(\mathbb{E}[y|z], a^T z) b \tag{119}$$

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 $\mathbb{E}[\epsilon|z] = 0$,

$$y = \mathbb{E}[y|z] + \epsilon. \tag{120}$$

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

$$\text{Cov}(y, x) = \text{Cov}(y, a^T z) b \;\Rightarrow \text{Cov}(x, r) = 0. \tag{121}$$

□

C Conditional expectation results for relatedness matrices

Lemma 4. Let

$$\mathrm{IBD}_{ij}^{\mathrm{par}} = \{\mathrm{IBD}_{k(i)l(j)}^{k'l'}(s)\}_{k,l,k',l'=\mathrm{m,p};\ s=1,\dots,L}, \tag{122}$$

then, for i, j such that i and j are not related by direct descent and are not monozygotic twins,

$$\mathbb{E}[R_{ij}|\mathrm{IBD}_{ij}^{\mathrm{par}}] = \frac{1}{2}[\mathrm{R}_{\mathrm{par}}]_{ij} \text{ and } \mathbb{E}[[\mathrm{R}_{\mathrm{o,par}}]_{ij}|\mathrm{IBD}_{ij}^{\mathrm{par}}] = [\mathrm{R}_{\mathrm{par}}]_{ij}. \tag{123}$$

Proof. First consider, because the segregation events in the parents of i and j are independent Bernoulli(0.5) variables,

$$\mathbb{E}[\mathrm{IBD}_{ij}^{kl}|\mathrm{IBD}_{ij}^{\mathrm{par}}] = \frac{1}{4}\sum_{k',l'=\mathrm{m,p}} \mathrm{IBD}_{k(i)l(j)}^{k'l'} = \mathrm{K}_{k(i)l(j)}. \tag{124}$$

Therefore,

$$\mathbb{E}[R_{ij}|\mathrm{IBD}_{ij}^{\mathrm{par}}] = \frac{1}{2}\frac{\sum_{k,l=\mathrm{m,p}}\mathbb{E}[\mathrm{IBD}_{ij}^{kl}|\mathrm{IBD}_{ij}^{\mathrm{par}}] - 4\mathrm{K}_0}{1-\mathrm{K}_0} \tag{125}$$

$$\mathbb{E}[R_{ij}|\mathrm{IBD}_{ij}^{\mathrm{par}}] = \frac{1}{2}\frac{\mathrm{K}_{\mathrm{p}(i)\mathrm{p}(j)} + \mathrm{K}_{\mathrm{p}(i)\mathrm{m}(j)} + \mathrm{K}_{\mathrm{m}(i)\mathrm{p}(j)} + \mathrm{K}_{\mathrm{m}(i)\mathrm{m}(j)} - 4\mathrm{K}_0}{1-\mathrm{K}_0} = \frac{1}{2}[\mathrm{R}_{\mathrm{par}}]_{ij}. \tag{126}$$

From the Relatedness Disequilibrium Lemma (Lemma 1),

$$\mathrm{K}_{i\mathrm{p}(j)} = \frac{1}{4}\sum_{k=\mathrm{m,p}}\sum_{k',l'=\mathrm{m,p}}\frac{1}{L}\sum_{s=1}^{L}\mathrm{I}_{kk'}(i,s)\mathrm{IBD}_{k(i)\mathrm{p}(j)}^{k'l'}(s). \tag{127}$$

Therefore,

$$\mathbb{E}[\mathrm{K}_{i\mathrm{p}(j)}|\mathrm{IBD}_{ij}^{\mathrm{par}}] = \frac{1}{2}\sum_{k=\mathrm{m,p}}\frac{1}{4}\sum_{k',l'=\mathrm{m,p}}\mathrm{IBD}_{k(i)\mathrm{p}(j)}^{k'l'} = \frac{1}{2}(\mathrm{K}_{\mathrm{m}(i)\mathrm{p}(j)} + \mathrm{K}_{\mathrm{p}(i)\mathrm{p}(j)}) \tag{128}$$

$$\Rightarrow \mathbb{E}[\mathrm{K}_{i\mathrm{p}(j)} + \mathrm{K}_{i\mathrm{m}(j)} + \mathrm{K}_{\mathrm{m}(i)j} + \mathrm{K}_{\mathrm{m}(i)j}|\mathrm{IBD}_{ij}^{\mathrm{par}}] = \mathrm{K}_{\mathrm{p}(i)\mathrm{p}(j)} + \mathrm{K}_{\mathrm{p}(i)\mathrm{m}(j)} + \mathrm{K}_{\mathrm{m}(i)\mathrm{p}(j)} + \mathrm{K}_{\mathrm{m}(i)\mathrm{m}(j)}. \tag{129}$$

$$\Rightarrow \mathbb{E}[[\mathrm{R}_{\mathrm{o,par}}]_{ij}|\mathrm{IBD}_{ij}^{\mathrm{par}}] = [\mathrm{R}_{\mathrm{par}}]_{ij}. \tag{130}$$

□

Remark. When i is a direct ancestor of j, or vice-versa, the segregation events that determine which ancestral material j inherits and which i inherits are not independent, so these relationships do not hold.

It is easy to show that, in an infinite, outbred population, when i is a parent of j or vice-versa, $R_{ij} = 0.5$, $[\mathrm{R}_{\mathrm{par}}]_{ij} = 0.5$, and $\mathbb{E}[[\mathrm{R}_{\mathrm{o,par}}]_{ij}] = 3/4$.

Lemma 5. Let

$$\mathrm{I}(i) = \{\mathrm{I}_{k\mathrm{m}}(i,s)\}_{k=\mathrm{m,p};s=1,\dots,L} \tag{131}$$

represent the segregation variables relevant for the genotype of i, then, for i, j such that i and j are not related by direct descent and are not monozygotic twins,

$$\mathbb{E}[R_{ij}|\mathrm{IBD}^{\mathrm{par}}_{ij}, \mathrm{I}(i)] = \frac{\mathrm{K}_{i\mathrm{p}(j)} + \mathrm{K}_{i\mathrm{m}(j)} - 2\mathrm{K}_0}{1-\mathrm{K}_0}. \tag{132}$$

Proof. From Lemma 4, we have by addition

$$\mathrm{K}_{i\mathrm{p}(j)} + \mathrm{K}_{i\mathrm{m}(j)} = \frac{1}{4}\sum_{k,l=\mathrm{m,p}}\sum_{k',l'=\mathrm{m,p}}\frac{1}{L}\sum_{s=1}^{L}\mathrm{I}_{kk'}(i,s)\mathrm{IBD}^{k'l'}_{k(i)l(j)}(s). \tag{133}$$

From the Relatedness Disequilibrium Lemma, we have

$$\mathrm{IBD}^{kl}_{ij} = \sum_{k',l'=\mathrm{m,p}}\frac{1}{L}\sum_{s=1}^{L}\mathrm{I}_{kk'}(i,s)\mathrm{I}_{ll'}(j,s)\mathrm{IBD}^{k'l'}_{k(i)l(j)}(s). \tag{134}$$

$$\mathbb{E}[\mathrm{IBD}^{kl}_{ij}|\mathrm{IBD}^{\mathrm{par}}_{ij}, \mathrm{I}(i)] = \frac{1}{2}\sum_{k',l'=\mathrm{m,p}}\frac{1}{L}\sum_{s=1}^{L}\mathrm{I}_{kk'}(i,s)\mathrm{IBD}^{k'l'}_{k(i)l(j)}(s). \tag{135}$$

Therefore,

$$\begin{aligned}\sum_{k,l=\mathrm{m,p}}\mathbb{E}\left[\mathrm{IBD}^{kl}_{ij}|\mathrm{IBD}^{\mathrm{par}}_{ij}, \mathrm{I}(i)\right] &= \frac{1}{2}\sum_{k,l=\mathrm{m,p}}\sum_{k',l'=\mathrm{m,p}}\frac{1}{L}\sum_{s=1}^{L}\mathrm{I}_{kk'}(i,s)\mathrm{IBD}^{k'l'}_{k(i)l(j)}(s) \\ &= 2(\mathrm{K}_{i\mathrm{p}(j)} + \mathrm{K}_{i\mathrm{m}(j)}).\end{aligned} \tag{136}$$

$$\Rightarrow \mathbb{E}[R_{ij}|\mathrm{IBD}^{\mathrm{par}}_{ij}, \mathrm{I}(i)] = \frac{\mathrm{K}_{i\mathrm{p}(j)} + \mathrm{K}_{i\mathrm{m}(j)} - 2\mathrm{K}_0}{1-\mathrm{K}_0}. \tag{137}$$

□

Corollary 5.1. By symmetry, we have,

$$\mathbb{E}[R_{ij}|\mathrm{IBD}^{\mathrm{par}}_{ij}, \mathrm{I}(j)] = \frac{\mathrm{K}_{j\mathrm{p}(i)} + \mathrm{K}_{j\mathrm{m}(i)} - 2\mathrm{K}_0}{1-\mathrm{K}_0}. \tag{138}$$

Therefore,

$$\mathbb{E}[R_{ij}|\mathrm{IBD}^{\mathrm{par}}_{ij}, \mathrm{I}(j)] + \mathbb{E}[R_{ij}|\mathrm{IBD}^{\mathrm{par}}_{ij}, \mathrm{I}(i)] = [\mathrm{R}_{\mathrm{o,par}}]_{ij}. \tag{139}$$

D Variances of quadratic forms in normal random variables

Any real, $[n \times n]$ symmetric matrix $\mathrm{R} = \mathrm{R}^T$ defines a quadratic form in n variables, $\mathbf{y} = [y_1, y_2, ..., y_n]^T$:

$$\mathbf{y}^T \mathrm{R} \mathbf{y} = \sum_{i=1}^{n} \sum_{j=1}^{n} y_i \mathrm{R}_{ij} y_j. \tag{140}$$

If the n variables have a multivariate normal distribution,

$$\mathbf{y} \sim \mathcal{N}(0, \Sigma), \tag{141}$$

then the distribution of the quadratic form can be easily derived. First, $\mathbf{y}$ has the same distribution as $\Sigma^{\frac{1}{2}}\mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(0, I)$. Therefore,

$$\mathbf{y}^T \mathrm{R} \mathbf{y} \overset{d}{=} \mathbf{z}^T \Sigma^{\frac{1}{2}} \mathrm{R} \Sigma^{\frac{1}{2}} \mathbf{z}. \tag{142}$$

Since R is symmetric and Σ is a covariance matrix, so therefore also symmetric, $\Sigma^{\frac{1}{2}} \mathrm{R} \Sigma^{\frac{1}{2}}$ is symmetric and has eigendecomposition

$$\Sigma^{\frac{1}{2}} \mathrm{R} \Sigma^{\frac{1}{2}} = U \lambda U^T, \text{ where } UU^T = U^T U = I, \tag{143}$$

and where λ is a diagonal matrix containing the eigenvalues of $\Sigma^{\frac{1}{2}} \mathrm{R} \Sigma^{\frac{1}{2}}$. This implies that

$$\mathbf{y}^T \mathrm{R} \mathbf{y} \overset{d}{=} \mathbf{z}^T U \lambda U^T \mathbf{z} = \tilde{\mathbf{z}}^T \lambda \tilde{\mathbf{z}} = \sum_{i=1}^{n} \lambda_i \tilde{z}_i^2, \tag{144}$$

where $\tilde{\mathbf{z}} = U^T \mathbf{z} \sim \mathcal{N}(0, I)$. The $\tilde{z}_i^2$ are distributed as independent χ_1^2 random variables. The distribution of the quadratic form is thus a linear combination of independent χ_1^2 random variables, with coefficients given by the eigenvalues of $\Sigma^{\frac{1}{2}} \mathrm{R} \Sigma^{\frac{1}{2}}$.

This gives the mean and variance of the quadratic form as

$$\mathbb{E}[\mathbf{y}^T \mathrm{R} \mathbf{y}] = \sum_{i=1}^{n} \lambda_i; \ \mathrm{Var}(\mathbf{y}^T \mathrm{R} \mathbf{y}) = 2 \sum_{i=1}^{n} \lambda_i^2. \tag{145}$$

If the mean and variance converge to finite values with n, then, by the Central Limit Theorem,

$$\mathbf{y}^T \mathrm{R} \mathbf{y} \xrightarrow{d} \mathcal{N}\left(\sum_{i=1}^{n} \lambda_i, 2 \sum_{i=1}^{n} \lambda_i^2 \right). \tag{146}$$

We provide a simpler expression for the variance of a quadratic form that is easier to compute than directly computing eigenvalues of $\Sigma^{\frac{1}{2}} \mathrm{R} \Sigma^{\frac{1}{2}}$:

$$\mathrm{Var}(\mathbf{y}^T \mathrm{R} \mathbf{y}) = 2 \sum_{i=1}^{n} \lambda_i^2 = 2 \,\mathrm{Tr}(\Sigma^{\frac{1}{2}} \mathrm{R} \Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}} \mathrm{R} \Sigma^{\frac{1}{2}}) = 2 \,\mathrm{Tr}(\mathrm{R} \Sigma \mathrm{R} \Sigma) = 2 ||\mathrm{R}\Sigma||_F^2. \tag{147}$$

D.1 Covariances

We use the above results to derive the covariance between two real, symmetric quadratic forms in normal variables:

$$\mathrm{Cov}(\mathbf{y}_1^T \mathrm{R}_1 \mathbf{y}_1, \mathbf{y}_2^T \mathrm{R}_2 \mathbf{y}_2) = 2\,\mathrm{Tr}(\mathrm{R}_1 \Sigma_{12} \mathrm{R}_2 \Sigma_{12}^T) = 2\langle \mathrm{R}_1 \Sigma_{12}, \Sigma_{12} \mathrm{R}_2 \rangle_F, \tag{148}$$

for

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}\right). \tag{149}$$

This is derived by considering $\mathrm{Var}(\mathbf{y}_1^T \mathrm{R}_1 \mathbf{y}_1 + \mathbf{y}_2^T \mathrm{R}_2 \mathbf{y}_2)$, which can be computed from the previous results. First, $\mathbf{y}_1^T \mathrm{R}_1 \mathbf{y}_1 + \mathbf{y}_2^T \mathrm{R}_2 \mathbf{y}_2$ can be expressed as a real, symmetric quadratic form:

$$\mathbf{y}_1^T \mathrm{R}_1 \mathbf{y}_1 + \mathbf{y}_2^T \mathrm{R}_2 \mathbf{y}_2 = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}^T \begin{bmatrix} \mathrm{R}_1 & 0 \\ 0 & \mathrm{R}_2 \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}. \tag{150}$$

Using the previous results, this gives

$$\mathrm{Var}(\mathbf{y}_1^T \mathrm{R}_1 \mathbf{y}_1 + \mathbf{y}_2^T \mathrm{R}_2 \mathbf{y}_2) = \mathrm{Var}(\mathbf{y}_1^T \mathrm{R}_1 \mathbf{y}_1) + \mathrm{Var}(\mathbf{y}_2^T \mathrm{R}_2 \mathbf{y}_2) + 4\,\mathrm{Tr}(\mathrm{R}_1 \Sigma_{12} \mathrm{R}_2 \Sigma_{12}^T), \tag{151}$$

and therefore

$$\mathrm{Cov}(\mathbf{y}_1^T \mathrm{R}_1 \mathbf{y}_1, \mathbf{y}_2^T \mathrm{R}_2 \mathbf{y}_2) = 2\,\mathrm{Tr}(\mathrm{R}_1 \Sigma_{12} \mathrm{R}_2 \Sigma_{12}^T) = 2\langle \mathrm{R}_1 \Sigma_{12}, \Sigma_{12} \mathrm{R}_2 \rangle_F, \tag{152}$$

where $\langle A, B \rangle_F$ is the Frobenius inner product between two matrices A and B.

It is then trivial to derive the correlation between two quadratic forms:

$$\mathrm{Corr}(\mathbf{y}_1^T \mathrm{R}_1 \mathbf{y}_1, \mathbf{y}_2^T \mathrm{R}_2 \mathbf{y}_2) = \frac{\langle \mathrm{R}_1 \Sigma_{12}, \Sigma_{12} \mathrm{R}_2 \rangle_F}{\sqrt{||\mathrm{R}_1 \Sigma_{11}||_F^2 ||R_2 \Sigma_{22}||_F^2}}. \tag{153}$$

E Mathematical definitions

Definition E.1. The father of i is denoted as $\mathrm{p}(i)$.

Definition E.2. The mother of i is denoted as $\mathrm{m}(i)$.

Definition E.3. The genotype of individual i is the number of copies of an allele at a locus in individual i's genome: $g_i = g_i^{\mathrm{p}} + g_i^{\mathrm{m}}$, where g_i^{p} is an indicator variable for the presence of the allele on the paternally inherited chromosome, and g_i^{m} is an indicator variable for the presence of the allele on the maternally inherited chromosome.

Definition E.4. The parental genotype of individual i is the number of copies of an allele at a locus in the genomes of the mother of i and the father of i: $g_i^{\mathrm{par}} = g_{\mathrm{m}(i)} + g_{\mathrm{p}(j)} = g_{\mathrm{m}(i)}^{\mathrm{p}} + g_{\mathrm{m}(i)}^{\mathrm{m}} + g_{\mathrm{p}(i)}^{\mathrm{p}} + g_{\mathrm{p}(i)}^{\mathrm{m}}$, where $g_{\mathrm{m}(i)}$ is the genotype of the mother of i, and $g_{\mathrm{p}(j)}$ is the genotype of the father of j.

Definition E.5. For $k, l = \mathrm{m,p}$, and for $s = 1, \ldots, L$,

$$\mathrm{IBD}_{ij}^{kl}(s) = \begin{cases} 1 & \text{if } k\text{-chromosome of } i \text{ is IBD with } l\text{-chromosome of } j \text{ at position } s; \\ 0 & \text{otherwise.} \end{cases}$$

Definition E.6. For $k, l = \mathrm{m,p}$,

$$\begin{aligned} \mathrm{IBD}_{ij}^{kl} &= \mathbb{P}(\text{the } k\text{-variant of } i \text{ is IBD with the } l\text{-variant of } j) \\ &= \frac{1}{L}\sum_{s=1}^{L} \mathrm{IBD}_{ij}^{kl}(s) \end{aligned} \tag{154}$$

Definition E.7.

$$\mathrm{IBD}_{ij}^{\mathrm{par}} = \{\mathrm{IBD}_{k(i)l(j)}^{k'l'}(s)\}_{k,l,k',l'=\mathrm{m,p};\ s=1,\ldots,L} \tag{155}$$

Definition E.8. For $k, k' = \mathrm{m,p}$, where 'm' indicates 'maternal' and 'p' indicates 'paternal', we define the segregation variables in the parents of i:

$$\mathrm{I}_{kk'}(i,s) = \begin{cases} 1 & \text{if the genetic variant at position } s \text{ on the } k\text{-chromosome of } i \\ & \text{was inherited from the } k' \text{ parent of } k; \\ 0 & \text{otherwise.} \end{cases} \tag{156}$$

Definition E.9.

$$\mathrm{I}(i) = \{\mathrm{I}_{k\mathrm{m}}(i,s)\}_{k=\mathrm{m,p};s=1,\ldots,L} \tag{157}$$

Definition E.10. Kinship between i and j:

$$\mathrm{K}_{ij} = \frac{1}{4}\sum_{k,l=\mathrm{m,p}} \mathrm{IBD}_{ij}^{kl}. \tag{158}$$

Definition E.11. The mean kinship in the population is K_0.

Definition E.12. Additive relatedness between i and j:

$$\mathrm{R}_{ij} = \frac{1}{2} \sum_{k,l=\mathrm{m,p}} \frac{\mathrm{IBD}_{ij}^{kl} - \mathrm{K}_0}{1 - \mathrm{K}_0}. \tag{159}$$

Definition E.13. Additive relatedness between the parents of i and the parents of j:

$$[\mathrm{R}_{\mathrm{par}}]_{ij} = \frac{\mathrm{K}_{\mathrm{p}(i)\mathrm{p}(j)} + \mathrm{K}_{\mathrm{p}(i)\mathrm{m}(j)} + \mathrm{K}_{\mathrm{m}(i)\mathrm{p}(j)} + \mathrm{K}_{\mathrm{m}(i)\mathrm{m}(j)} - 4\mathrm{K}_0}{1 - \mathrm{K}_0}, \tag{160}$$

where $\mathrm{K}_{\mathrm{p}(i)\mathrm{m}(j)}$ is the kinship between the father of i and the mother of j, etc.

Definition E.14.

$$[\mathrm{R}_{\mathrm{o,par}}]_{ij} = \frac{\mathrm{K}_{i\mathrm{p}(j)} + \mathrm{K}_{i\mathrm{m}(j)} + \mathrm{K}_{\mathrm{p}(i)j} + \mathrm{K}_{\mathrm{m}(i)j} - 4\mathrm{K}_0}{1 - \mathrm{K}_0}, \tag{161}$$

where $\mathrm{K}_{i\mathrm{m}(j)}$ is the kinship between i and the mother of j, and $\mathrm{K}_{\mathrm{p}(i)j}$ is the kinship between j and the father of i, etc.

Glossary

identity-by-descent A segment of a chromosome on two or more haplotypes is said to be identical-by-descent if that segment was inherited from a common ancestor without recombination. The identity-by-descent sharing states between two chromosomes reflect which segments of the chromosomes are identical-by-descent. 5, 11, 15

indirect genetic effect An effect of genetic material in one individual on another individual through the environment. 8, 29

parental genetic nurturing effect A specific type of indirect genetic effect where the genotype of the parent affects the trait of the offspring through its environment. 4

reciprocal genetic effect A specific type of indirect genetic effect between a pair of individuals: the genotype of one individual affects the environment of the other and vice-versa. 6, 8, 29, 30