The American Journal of Human Genetics, Volume *103*

Supplemental Data

Estimating SNP-Based Heritability and Genetic Correlation in Case-Control Studies Directly and with Summary Statistics

Omer Weissbrod, Jonathan Flint, and Saharon Rosset

Contents

1 Supplemental Tables

Table S1: **Estimation correctness of the investigated methods**. LDSC behaves differently depending on whether covariates are present. The entries marked with * indicate that although the estimated quantity is empirically unbiased in simulations, it is given by the division of two biased estimates (the estimate in the second column divided by the product of the square roots of the estimates in the first column), suggesting that estimation errors cancel each other in the division. We are not currently aware of a theoretical justification for this behavior. The entry marked with ** is only empirically correct as long as covariates are excluded from the analysis.

		heritability	genetic covariance	genetic correlation
no covariates	PCGC	✓	✓	✓
	LDSC	✓	✓	✓
	REML	✗	✗	✓*
with covariates	PCGC	✓	✓	✓
	LDSC	✗	✗	✓**
	REML	✗	✗	✓*

Table S2: Please see Supplemental Excel file

Table S3: **Results of real data analysis of schizophrenia (SCZ) and bipolar disorder (BIP), when using the LDAK model assumptions [6] instead of the standard assumptions used in the results reported in the main text.**

Covariates		SCZ		BIP		
		$\hat{\sigma}_g^2$	$\hat{h}^2$	$\hat{\sigma}_g^2$	$\hat{h}^2$	Correlation
Omitted	PCGC-s	0.048 (0.051)	0.048 (0.051)	0.254 (0.056)	0.254 (0.056)	0.850 (0.439)
	PCGC-s-LD	0.052 (0.059)	0.052 (0.059)	0.288 (0.064)	0.288 (0.064)	0.862 (0.488)
Included	PCGC-s	0.407 (0.054)	0.400 (0.055)	0.471 (0.057)	0.461 (0.059)	0.438 (0.080)
	PCGC-s-LD	0.410 (0.060)	0.403 (0.062)	0.486 (0.063)	0.476 (0.066)	0.442 (0.086)

Table S4: **Results of real data analysis of type 1 diabetes (T1D) and coronary artery disease (CAD), using the ldsc software**[1]. Shown are the estimated genetic variance σ_g^2 and the genetic correlation, as obtained from the ldsc software. Marginal heritability estimates are not reported because they are not estimated in the ldsc software. Values marked with "-" could not be computed because of negative or illegal parameter estimates.

Covariates		T1D	CAD	
		$\hat{\sigma}_g^2$	$\hat{\sigma}_g^2$	Correlation
Omitted	LDSC-omit	0.385 (0.049)	0.644 (0.069)	0.298 (0.069)
	LDSC-omit+intercept	0.055 (0.066)	0.102 (0.102)	-1.07 (1.90)
Included	LDSC	-1.80 (0.040)	-0.338 (0.060)	-
	LDSC+intercept	-0.016 (0.037)	0.037 (0.090)	-

Table S5: **Results of real data analysis, using in-sample SNP normalization**. The table is similar to Table 2 in the main text, but both studies estimated the minor allele frequencies based on the (shared) controls rather than using HapMap 3 estimates.

Covariates		T1D		CAD		
		$\hat{\sigma}_g^2$	$\hat{h}^2$	$\hat{\sigma}_g^2$	$\hat{h}^2$	Correlation
Omitted	PCGC-s	0.295 (0.051)	0.295 (0.051)	0.469 (0.064)	0.469 (0.064)	0.231 (0.090)
	PCGC-s-LD	0.291 (0.050)	0.291 (0.050)	0.465 (0.064)	0.465 (0.064)	0.231 (0.090)
	LDSC-omit	0.284 (0.050)	0.284 (0.050)	0.451 (0.064)	0.451 (0.064)	0.215 (0.094)
	LDSC-omit + intercept	0.505 (0.552)	0.505 (0.552)	0.014 (0.131)	0.014 (0.131)	- -
Included	PCGC-s	0.277 (0.069)	0.210 (0.052)	0.498 (0.072)	0.457 (0.066)	0.239 (0.119)
	PCGC-s-LD	0.274 (0.068)	0.208 (0.052)	0.493 (0.064)	0.452 (0.059)	0.239 (0.119)
	LDSC	-1.80 (0.042)	- -	-0.45 (0.057)	- -	- -
	LDSC + intercept	0.040 (0.080)	0.030 (0.060)	-0.065 (0.099)	- -	- -

[1] `https://github.com/bulik/ldsc`

Table S6: **Results of real data analysis when regressing the top 10 principal components out of the genotypes and possibly using them as additional covariates.**

Covariates		T1D		CAD		
		$\hat{\sigma}_g^2$	$\hat{h}^2$	$\hat{\sigma}_g^2$	$\hat{h}^2$	Correlation
Omitted	PCGC-s	0.219 (0.043)	0.219 (0.043)	0.406 (0.063)	0.406 (0.063)	0.200 (0.117)
	PCGC-s-LD	0.226 (0.044)	0.226 (0.044)	0.419 (0.065)	0.419 (0.065)	0.198 (0.116)
	LDSC-omit	0.301 (0.045)	0.301 (0.045)	0.538 (0.066)	0.538 (0.066)	0.241 (0.085)
	LDSC-omit + intercept	-0.016 (0.097)	-0.016 (0.097)	-0.057 (0.112)	-0.057 (0.112)	- -
Included	PCGC-s	0.258 (0.066)	0.196 (0.050)	0.471 (0.069)	0.432 (0.063)	0.287 (0.123)
	PCGC-s-LD	0.266 (0.068)	0.202 (0.052)	0.487 (0.065)	0.446 (0.060)	0.284 (0.122)
	LDSC	-1.78 (0.039)	- -	-0.36 (0.057)	- -	- -
	LDSC + intercept	-0.060 (0.044)	- -	-0.11 (0.096)	- -	- -

Table S7: **Results of real data analysis when using in-sample SNP normalization and regressing the top 10 principal components out of the genotypes**. The table is similar to Supplemental Table 6, but both studies estimated the minor allele frequencies based on the (shared) controls rather than using HapMap 3 estimates.

Covariates		T1D		CAD		
		$\hat{\sigma}_g^2$	$\hat{h}^2$	$\hat{\sigma}_g^2$	$\hat{h}^2$	Correlation
Omitted	PCGC-s	0.237 (0.045)	0.237 (0.045)	0.456 (0.064)	0.456 (0.064)	0.210 (0.103)
	PCGC-s-LD	0.232 (0.044)	0.232 (0.044)	0.447 (0.063)	0.447 (0.063)	0.207 (0.102)
	LDSC-omit	0.195 (0.045)	0.195 (0.045)	0.383 (0.064)	0.383 (0.064)	0.091 (0.135)
	LDSC-omit + intercept	-0.081 (0.080)	- -	-0.096 (0.129)	- -	- -
Included	PCGC-s	0.278 (0.068)	0.211 (0.052)	0.534 (0.072)	0.489 (0.066)	0.306 (0.110)
	PCGC-s-LD	0.273 (0.066)	0.207 (0.050)	0.524 (0.065)	0.479 (0.059)	0.303 (0.108)
	LDSC	-1.84 (0.041)	- -	-0.50 (0.056)	- -	- -
	LDSC + intercept	-0.065 (0.046)	- -	-0.14 (0.106)	- -	- -

Table S8: **PCGC-s genetic correlation estimates between WTCCC1 phenotypes**. The traits and their assumed prevalences (following [1]) are Crohn's disease (CD, 0.1%), type 1 diabetes (T1D; 0.5%), bipolar disorder (BD, 0.5%), rheumatoid arthritis (RA; 0.75%), type 2 diabetes (T2D; 3%), coronary artery disease (CAD; 3.5%) and hypertension (HT; 5%). All analyses included sex as a covariate. T1D and RA analyses additionally excluded the MHC region from the analysis and used MHC SNPs as covariates (Supplemental Note).

	CD	T1D	BD	RA	T2D	CAD	HT
CD		0.067 (0.128)	0.217 (0.077)	0.047 (0.104)	0.155 (0.098)	0.114 (0.097)	0.259 (0.087)
T1D	0.067 (0.128)		0.090 (0.128)	0.387 (0.137)	-0.054 (0.161)	0.192 (0.139)	0.089 (0.155)
BD	0.217 (0.077)	0.090 (0.128)		-0.012 (0.117)	-0.145 (0.116)	0.011 (0.103)	0.136 (0.100)
RA	0.047 (0.104)	0.387 (0.137)	-0.012 (0.117)		0.228 (0.121)	0.314 (0.105)	0.226 (0.124)
T2D	0.155 (0.098)	-0.054 (0.161)	-0.145 (0.116)	0.228 (0.121)		0.343 (0.097)	0.371 (0.092)
CAD	0.114 (0.097)	0.192 (0.139)	0.011 (0.103)	0.314 (0.105)	0.343 (0.097)		0.280 (0.096)
HT	0.259 (0.087)	0.089 (0.155)	0.136 (0.100)	0.226 (0.124)	0.371 (0.092)	0.280 (0.096)	

2 Supplemental Figures

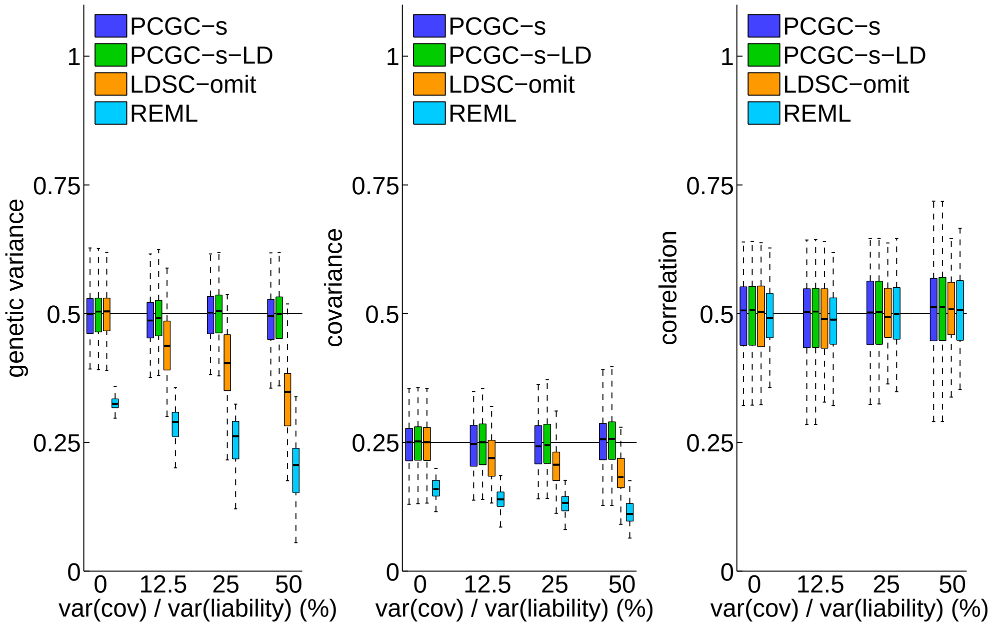

Figure S1: The performance of the evaluated methods when measuring the genetic variance $\sigma_g^{2^t}$ (also called the conditional heritability in the main text) instead of the marginal heritability $h^{2^t} = \frac{\sigma_g^{2^t}}{1+\text{var}\left(\boldsymbol{C}_i^t\boldsymbol{\beta}\right)}$ (see Methods in main text for further clarification regarding these terms).

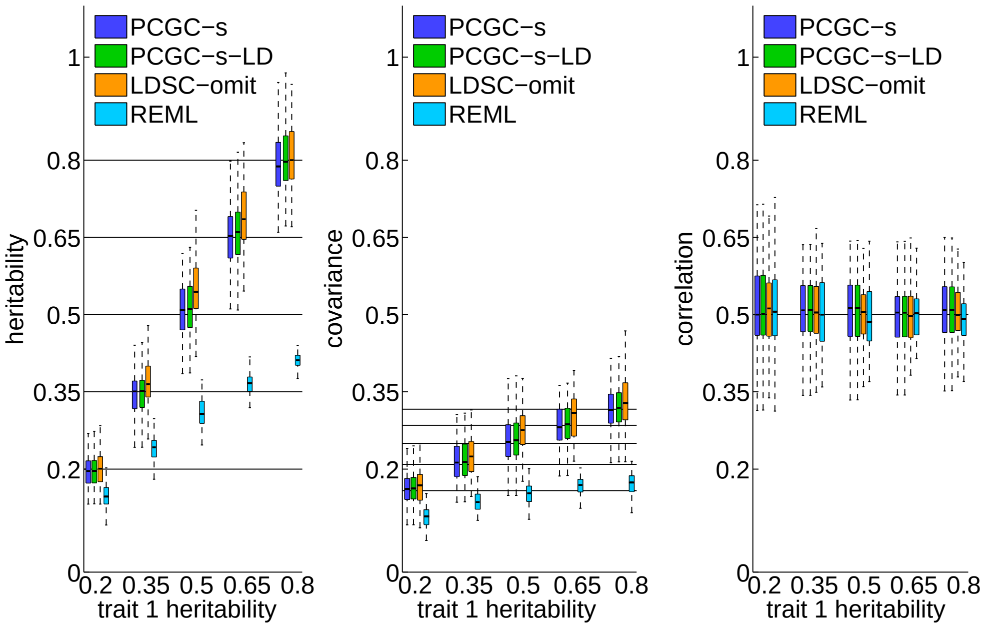

Figure S2: The performance of the evaluated methods under different heritability levels for trait 1. The black horizontal lines indicate the true parameter values. The genetic covariance values were set to obtain a genetic correlation of 50%.

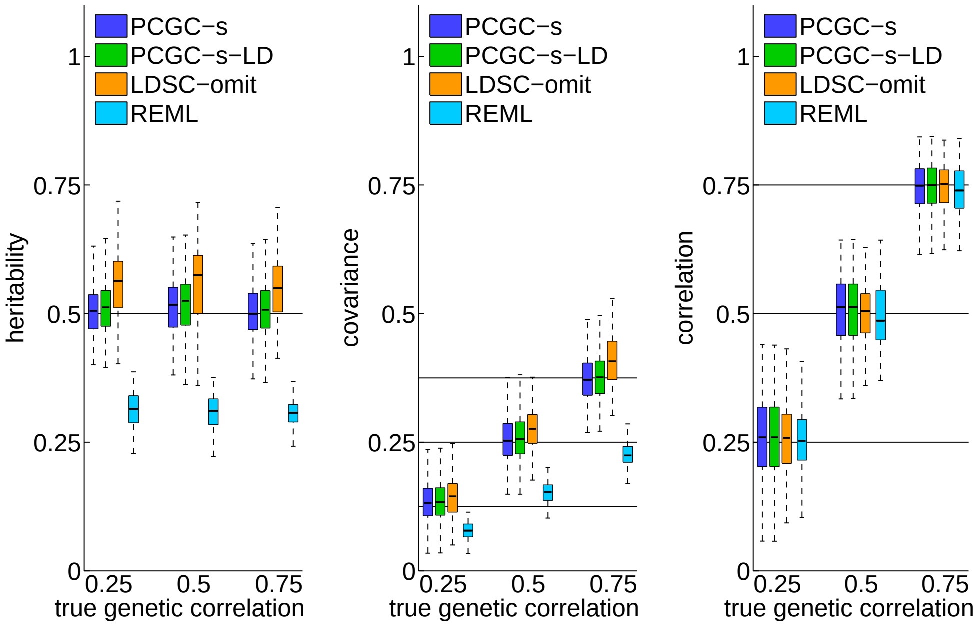

Figure S3: The performance of the evaluated methods under different genetic correlation levels. The black horizontal lines indicate the true parameter values.

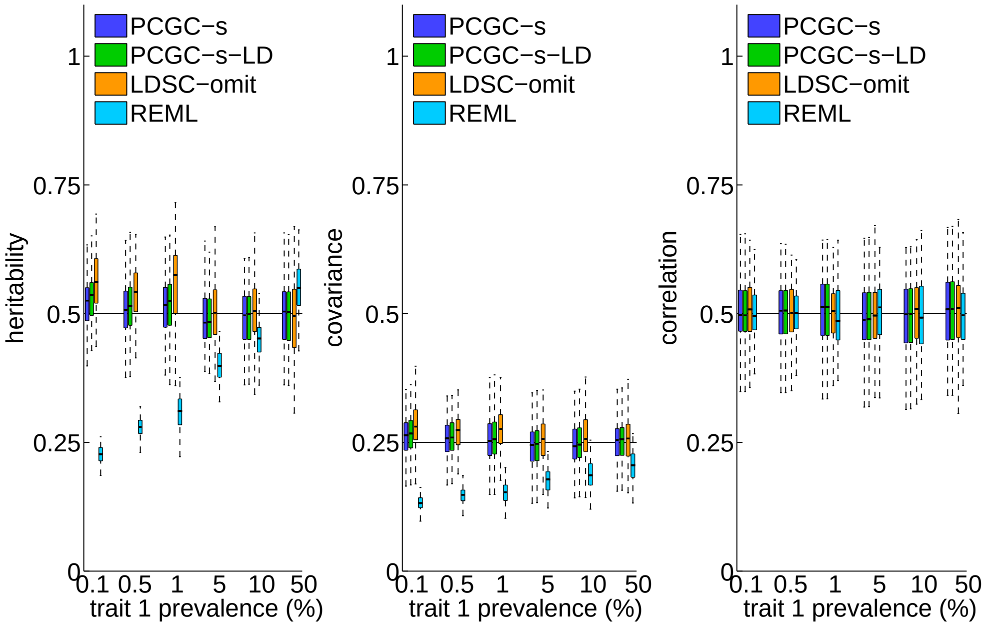

Figure S4: The performance of the evaluated methods under different prevalence levels. The in-sample case control ratio was 50% in all experiments.

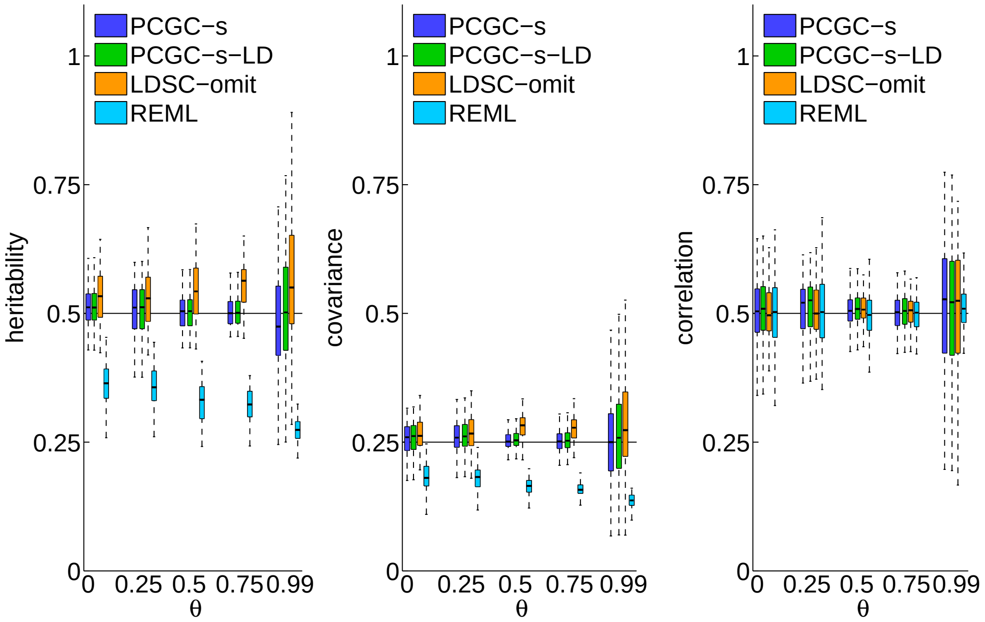

Figure S5: The effect of the LD parameter θ. Larger values of θ lead to a stronger correlation between adjacent SNPs. The standard error of all methods increases with the degree of LD.

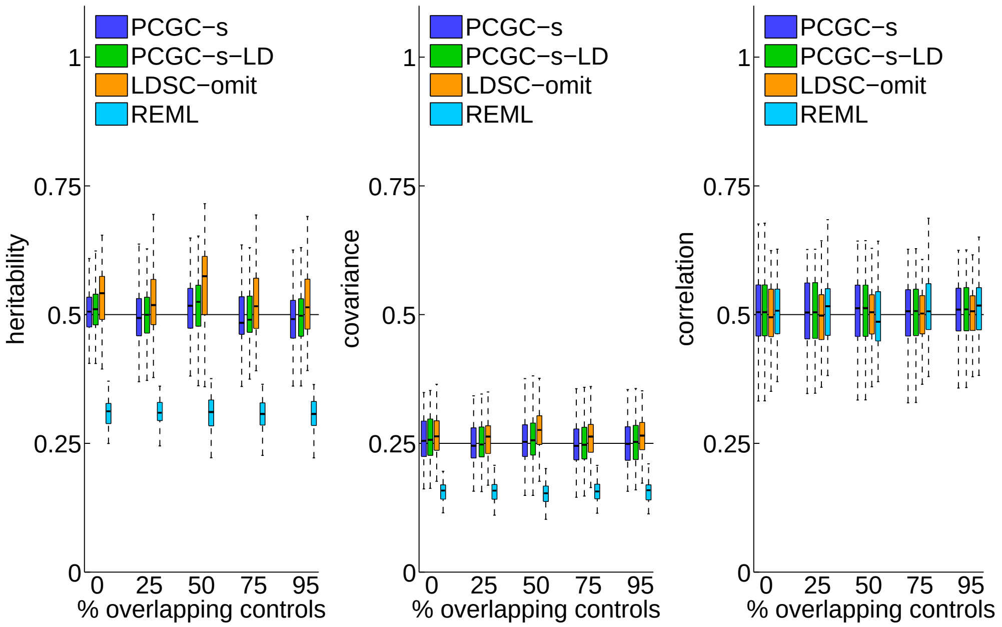

Figure S6: The performance of the evaluated methods under different levels of overlap between the control groups of the two studies.

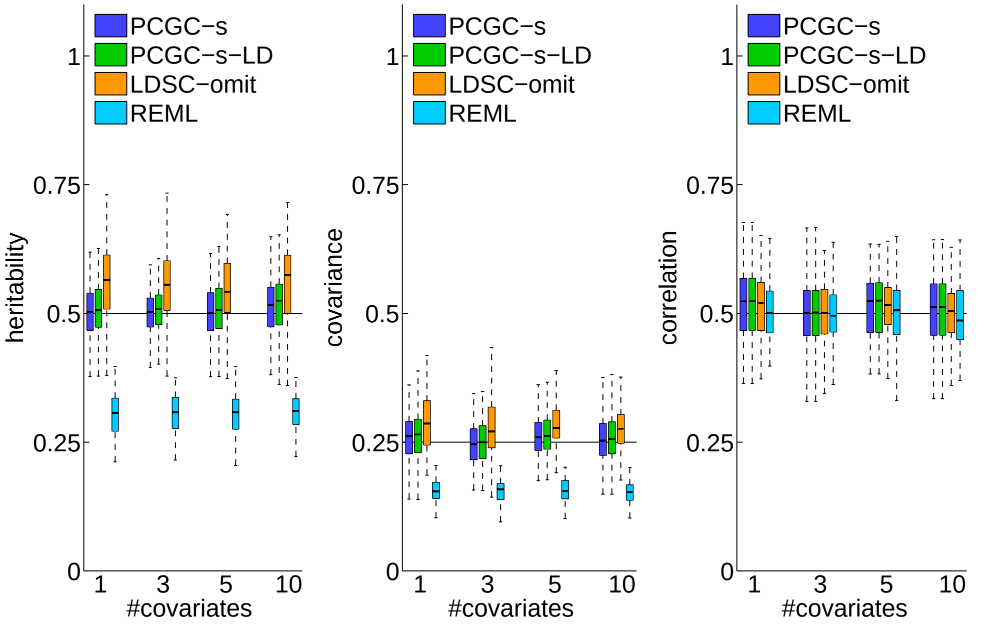

Figure S7: The performance of the evaluated methods under different numbers of measured covariates.

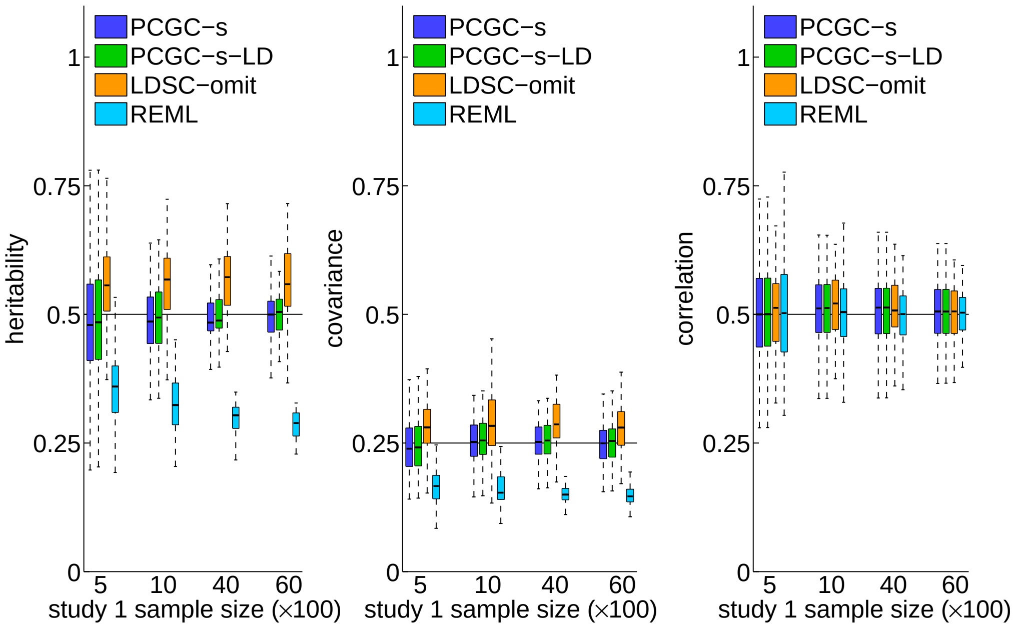

Figure S8: The performance of the evaluated methods under different sample sizes for study 1. PCGC and PCGC-s become increasingly more accurate as sample sizes increase.

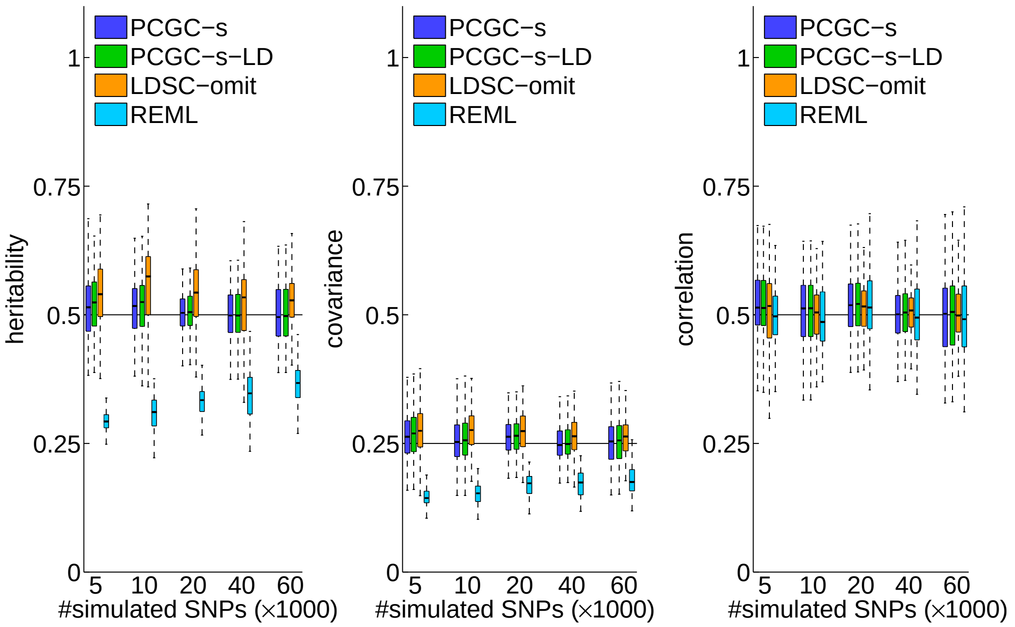

Figure S9: The performance of the evaluated methods under different numbers of simulated SNPs.

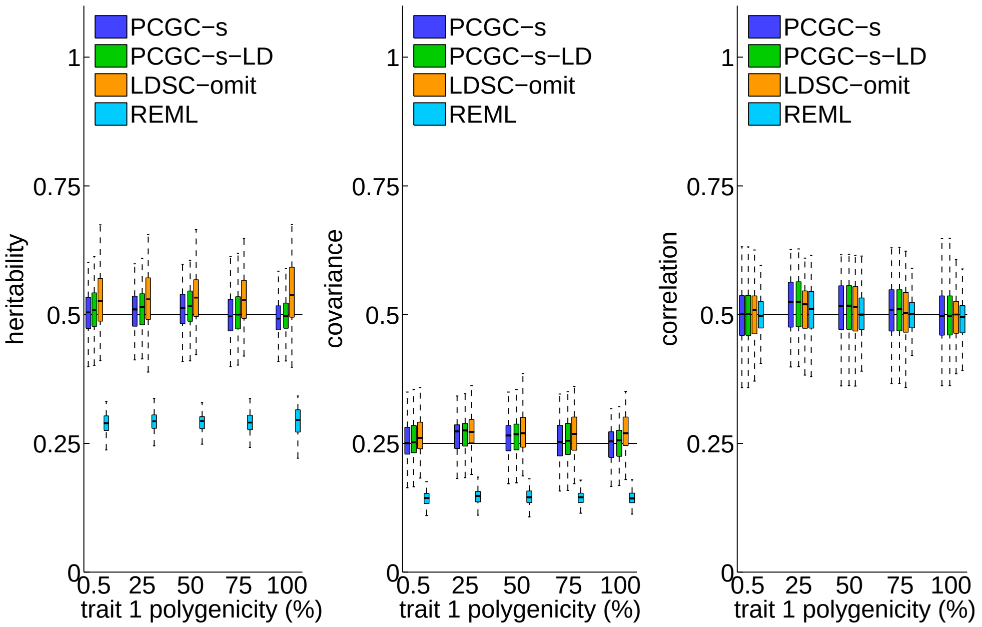

Figure S10: The performance of the evaluated methods under different polygenicity levels. The x axis is the fraction of SNPs in the genome that influence the trait of study 1.

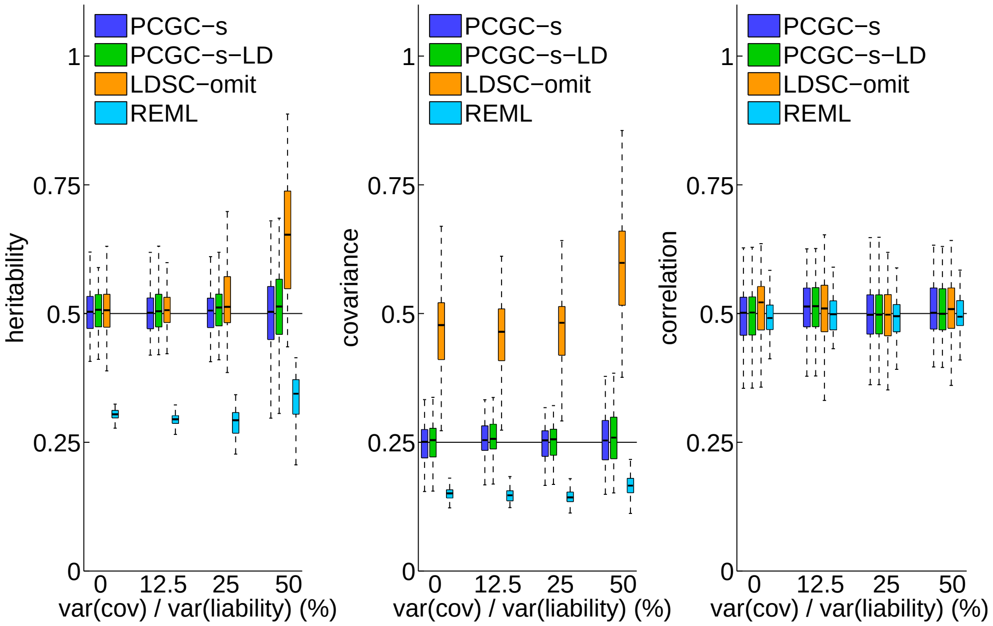

Figure S11: The performance of the evaluated methods when LDSC weights test statistics according to their postulated posterior variance (but still using a constrained intercept[2]), as implemented in the ldsc software[2].

[2] `https://github.com/bulik/ldsc`

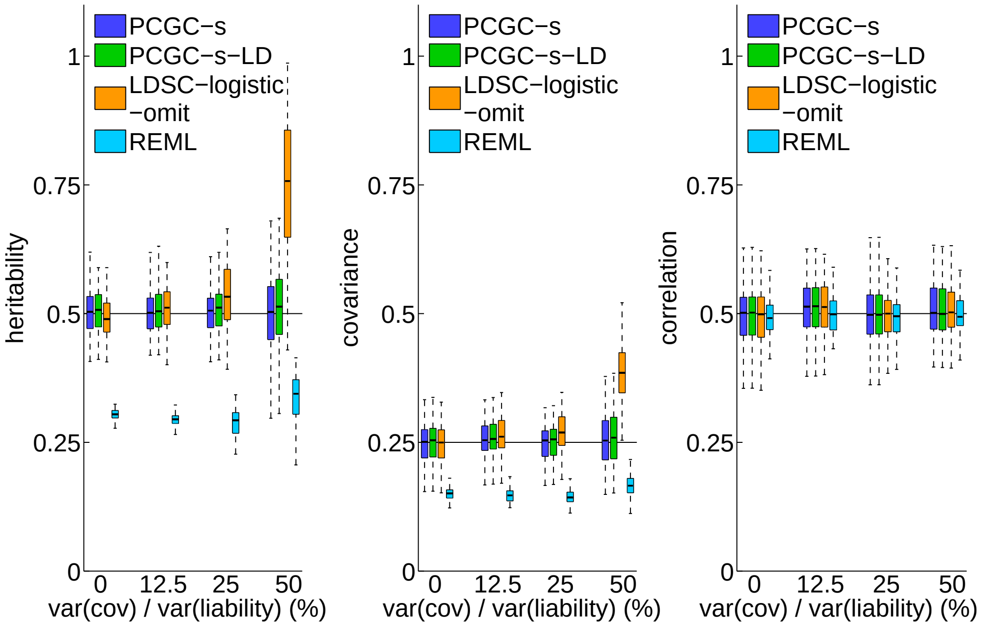

Figure S12: The performance of the evaluated methods when LDSC uses logistic regression rather than linear regression based summary statistics.

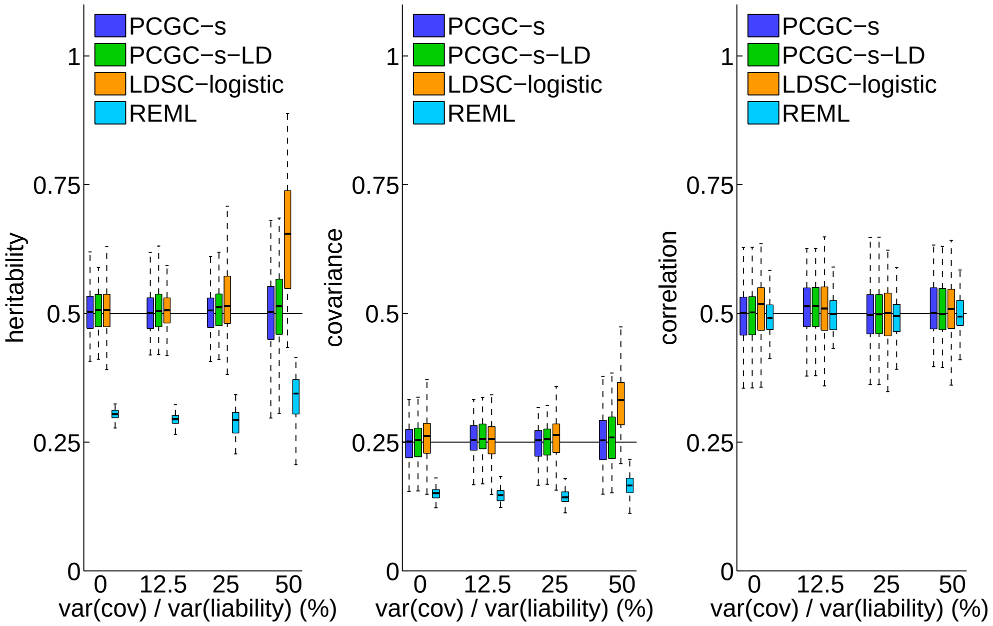

Figure S13: The performance of the evaluated methods when LDSC uses logistic regression rather than linear regression based summary statistics. Here, the logistic regression test statistics included the covariates instead of omitting them.

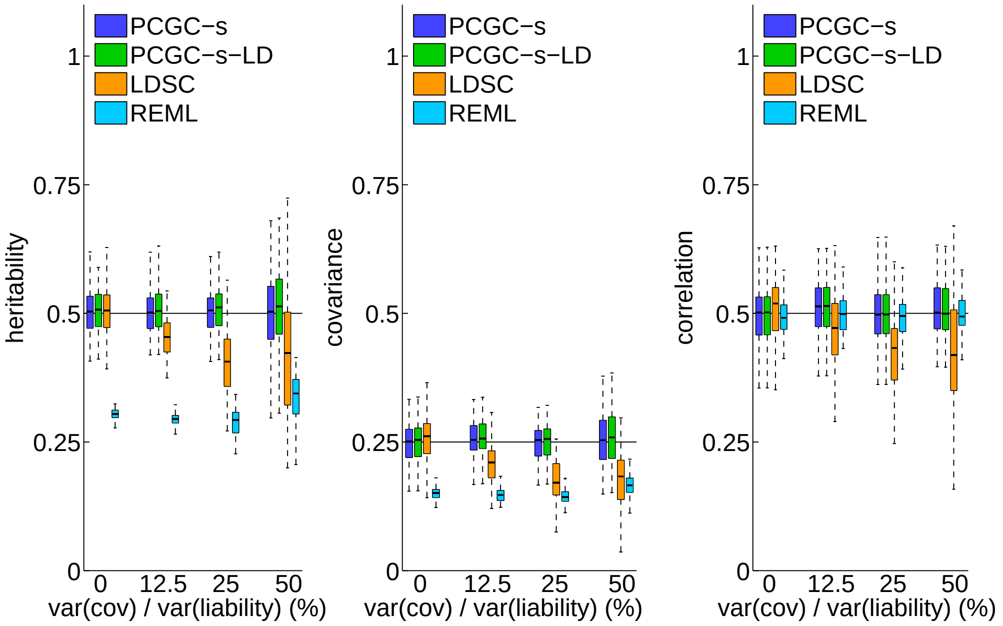

Figure S14: The performance of the evaluated methods under different levels of covariate strength, when LDSC regresses the covariates out of the phenotypes and genotypes.

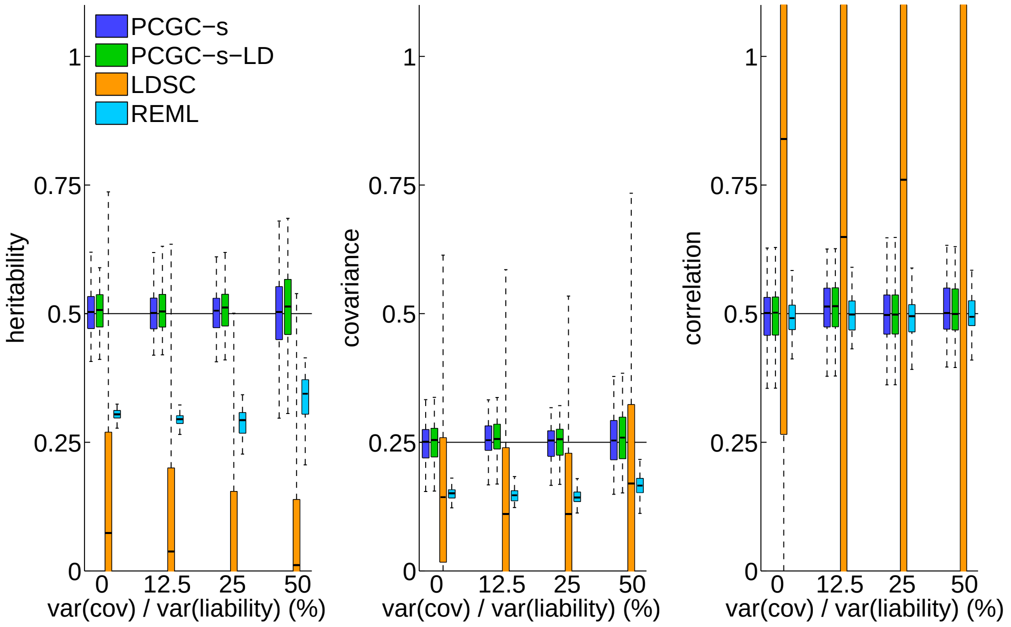

Figure S15: The performance of the evaluated methods under different levels of covariate strength, when LDSC regresses the covariates out of the phenotypes and genotypes and fits an intercept.

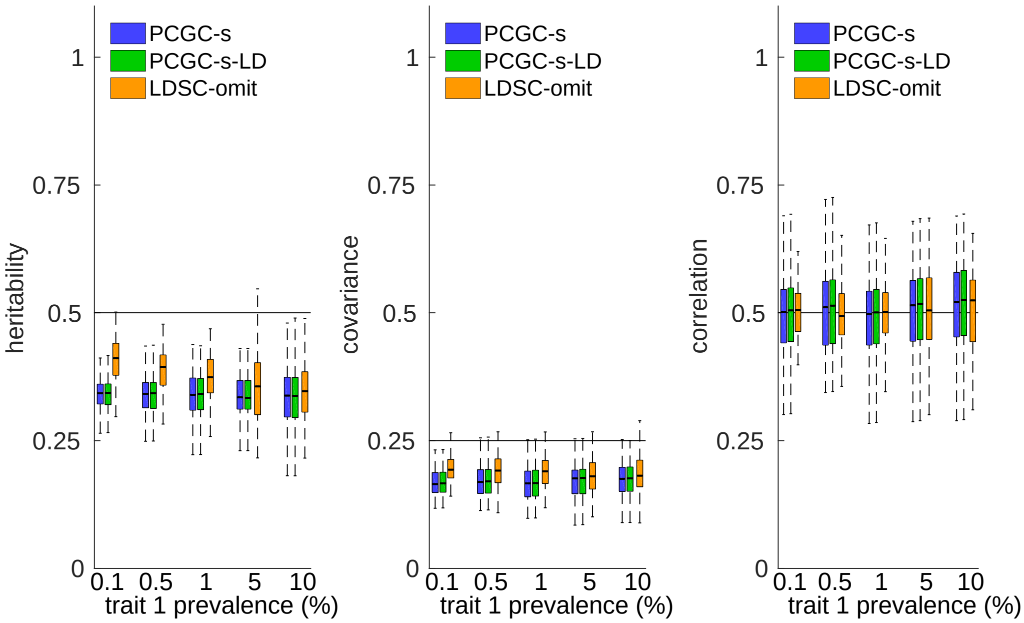

Figure S16: The performance of the evaluated methods when data is generated according to the LDAK model [6], under different prevalence levels for study 1. All methods yield biased estimates of heritability and of genetic covariance, because they use an incorrect model that assigns a uniform prior variance for the effect size of every SNP, regardless of its MAF and LD patterns. In contrast, genetic correlation estimates are unbiased, suggesting that the approximation errors of the heritabilities and of the genetic covariance are canceled when dividing the latter by the former. REML is not evaluated in this experiment because we are not aware of a REML-based method for estimation of genetic covariance under the LDAK model, which would be required for comparison with Figure S17.

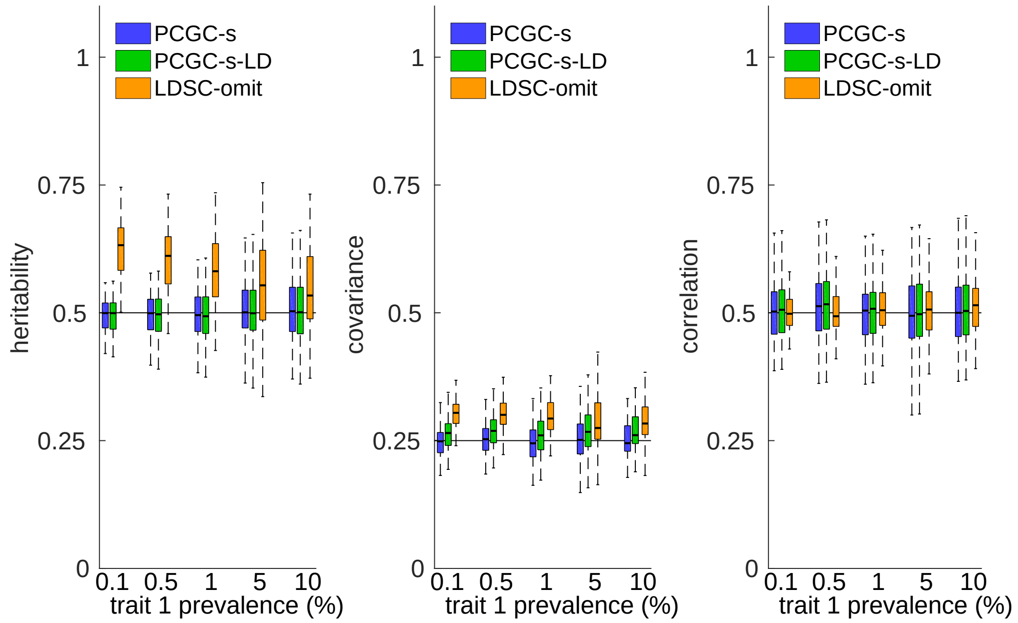

Figure S17: The performance of the evaluated methods when data is generated according to the LDAK model [6] (as in Figure S16), using modified versions of the evaluated methods which use the LDAK model for estimation. PCGC-s and PCGC-s-LD yield unbiased estimates because they use the correct underlying model, whereas LDSC-omit is biased because it ignores the effect of covariates. REML was not evaluated in this experiment because we are not aware of a REML-based method for estimation of genetic covariance under the LDAK model.

3 PCGC without Covariates

PCGC was described in [3] in the context of heritability estimation. Here we show the derivation for estimation of genetic covariance. This is a generalization of heritability, which in the absence of covariates can be seen as the genetic covariance of a trait with itself. We first present the derivation when there are no covariates. A derivation with covariates is presented in Section 4. The derivation here does not make use of summary statistics. A description of how PCGC can be reformulated to use summary statistics is presented in Section 5.

We first establish some notations. We assume the same mixed effects liability threshold model described in the main text. Namely, every individual is associated with a latent liability a_t^i for every studied trait t, where $a_t^i = g_t^i + e_t^i$, and g_t^i,e_t^i are genetic and environmental effects, respectively. We further assume $g_t^i \sim \mathcal{N}(0, \sigma_{g\,t}^2)$, $e_t^i \sim \mathcal{N}(0, 1-\sigma_{g\,t}^2)$. The environmental effects are assumed to be independent and identically distributed between individuals, and $\text{cov}(g_{t_1}^i, g_{t_2}^j) = \rho_{t_1,t_2} G_{t_1,t_2}^{i,j}$, where $G_{t_1,t_2}^{i,j}$ is the genetic similarity coefficient between individual i in study t_1 and individual j in study t_2. Every individual is also associated with an observed affection status indicator $y_t^i = \mathbb{1}\left[a_t^i > \tau_t\right]$, where $\tau_t = \Phi^{-1}(1 - K_t)$ is the affection cutoff for a trait with prevalence K_t, and where $\Phi^{-1}(\cdot)$ is the inverse standard normal cumulative distribution.
Note that when t_1 and t_2 refer to the same trait, the genetic covariance coincides with heritability, $\rho_{t_1,t_2} = \sigma_{g\,t}^2$. Our derivation therefore encapsulates heritability estimation as a special case.

We assume an ascertained case-control study where cases are overrepresented relative to the trait prevalence. Denote P_t as the case-control proportion in study t, and define $\tilde{y}_t^i \triangleq (y_t^i - P_t)/\sqrt{P_t(1-P_t)}$ as the standardized phenotype of individual i in study t. Further denote s_t^i as an observed selection indicator for individual i in study t, such that $s_t^i = 1$ for all individuals in the study. We assume that s_t^i is conditionally independent of all other variables given y_t^i. PCGC approximates the expected value of $\tilde{y}_{t_1}^i \tilde{y}_{t_2}^j$ conditional on the ascertainment scheme and on the genetic similarity coefficient of individuals i and j via a Taylor expansion around $G_{t_1,t_2}^{i,j} = 0$. Namely, the first order Taylor expansion when there are no covariates is given by:

$$E\left[\tilde{y}_{t_1}^i \tilde{y}_{t_2}^j \mid s_{t_1}^i, s_{t_2}^j, G_{t_1,t_2}^{i,j}\right] = G_{t_1,t_2}^{i,j} f\left(t_1, t_2\right) \rho_{t_1,t_2} + \mathcal{O}\left((G_{t_1,t_2}^{i,j})^2\right), \tag{1}$$

where s_t^i is a shorthand notation for $s_{t_1}^i = 1$, and where $f\left(t_1, t_2\right)$ is given by:

$$f\left(t_1, t_2\right) = \frac{\sqrt{P_{t_1}(1-P_{t_1})P_{t_2}(1-P_{t_2})}\phi(\tau_{t_1})\phi(\tau_{t_2})}{K_{t_1}(1-K_{t_1})K_{t_2}(1-K_{t_2})}. \tag{2}$$

Here, $\phi(\cdot)$ is the standard normal density. Therefore, ρ_{t_1,t_2} can be estimated by regressing $\tilde{y}_{t_1}^i \tilde{y}_{t_2}^j$ on $G_{t_1,t_2}^{i,j} f\left(t_1, t_2\right)$.

The derivation of Equation 1 is carried out as follows. We first write down the expected value of $\tilde{y}_{t_1}^i \tilde{y}_{t_2}^j$ conditional on the ascertainment scheme and on the genetic similarity coefficient of individuals i and j. By using Bayes rule and the assumption that s_t^i is conditionally independent of all other variables given y^i_t, we obtain:

$$
\begin{aligned}
E\left[\tilde{y}^i_{t_1}\tilde{y}^j_{t_2} \mid s^i_{t_1}, s^j_{t_2}, G^{i,j}_{t_1,t_2}\right] &= \sum_{y^i_{t_1}, y^j_{t_2}=0}^{1} \frac{y^i_{t_1} - P_{t_1}}{\sqrt{P_{t_1}(1-P_{t_1})}} \frac{y^j_{t_2} - P_{t_2}}{\sqrt{P_{t_2}(1-P_{t_2})}} P(y^i_{t_1}, y^j_{t_2} \mid s^i_{t_1}, s^j_{t_2}, G^{i,j}_{t_1,t_2}) \\
&= \frac{\sum_{y^i_{t_1}, y^j_{t_2}=0}^{1} \frac{y^i_{t_1} - P_{t_1}}{\sqrt{P_{t_1}(1-P_{t_1})}} \frac{y^j_{t_2} - P_{t_2}}{\sqrt{P_{t_2}(1-P_{t_2})}} P(y^i_{t_1}, y^j_{t_2} | G^{i,j}_{t_1,t_2}) P(s^i_{t_1} \mid y^i_{t_1}) P(s^j_{t_2} | y^j_{t_2})}{P(s^i_{t_1}, s^j_{t_2} \mid G^{i,j}_{t_1,t_2})}.
\end{aligned}
\tag{3}
$$

Next, we approximate Equation 3 via a Taylor expansion around $G^{i,j}_{t_1,t_2} = 0$. Denote the numerator as $A(G^{i,j}_{t_1,t_2})$ and the denominator as $B(G^{i,j}_{t_1,t_2})$. The Taylor expansion takes the form:

$$
E\left[\tilde{y}^i_{t_1}\tilde{y}^j_{t_2} \mid s^i_{t_1}, s^j_{t_2}, G^{i,j}_{t_1,t_2}\right] = \frac{A(0)}{B(0)} + \frac{A'(0)B(0) - B'(0)A(0)}{B(0)^2} G^{i,j}_{t_1,t_2} + \mathcal{O}\left((G^{i,j}_{t_1,t_2})^2\right). \tag{4}
$$

Equation 4 can be simplified because $A(0) = 0$. This can be verified by noting that setting $G^{i,j}_{t_1,t_2} = 0$ in Equation 4 yields $A(0)/B(0)$ on the one hand, but setting $G^{i,j}_{t_1,t_2} = 0$ also causes the random variables $\tilde{y}^i_{t_1}, \tilde{y}^j_{t_2}$ to become independent conditional on $s^i_{t_1}, s^j_{t_2}$, and therefore leads to the decomposition:

$$
E\left[\tilde{y}^i_{t_1}\tilde{y}^j_{t_2} \mid s^i_{t_1}, s^j_{t_2}, G^{i,j}_{t_1,t_2} = 0\right] = E\left[\tilde{y}^i_{t_1} \mid s^i_{t_1}\right] E\left[\tilde{y}^j_{t_2} \mid s^j_{t_2}\right] = 0, \tag{5}
$$

because $E\left[\tilde{y}^i_t \mid s^i_t\right] = 0$ by definition.

We conclude that the Taylor expansion takes the form:

$$
E\left[\tilde{y}^i_{t_1}\tilde{y}^j_{t_2} \mid s^i_{t_1}, s^j_{t_2}, G^{i,j}_{t_1,t_2}\right] = \frac{A'(0)}{B(0)} G^{i,j}_{t_1,t_2} + \mathcal{O}\left((G^{i,j}_{t_1,t_2})^2\right). \tag{6}
$$

To compute $B(0)$, we first compute the probability of cases and controls to participate in the study. Define $s_{t,0} = P(s^i_t = 1 | y^i_t = 0)$, $s_{t,1} = P(s^i_t = 1 | y^i_t = 1)$ as the selection probabilities of controls and cases, respectively. Using the definition of P_t and Bayes rule, we have:

$$
\begin{aligned}
P_t = P(y^i_t = 1 | s^i_t = 1) &= \frac{P(y^i_t = 1) P(s^i_t = 1 | y^i_t = 1)}{P(y^i_t = 0) P(s^i_t = 1 | y^i_t = 0) + P(y^i_t = 1) P(s^i_t = 1 | y^i_t = 1)} \\
&= \frac{K_t s_{t,1}}{(1 - K_t) s_{t,0} + K_t s_{t,1}}.
\end{aligned}
\tag{7}
$$

After rearrangement, we obtain:

$$
s_{t,0} = s_{t,1} \frac{K_t(1 - P_t)}{(1 - K_t) P_t}. \tag{8}
$$

We assume without loss of generalization that $s_{t,1} = 1$, but the results remain exactly the same regardless.

Next, we use the fact that the variables $s^i_{t_1}, s^j_{t_2}$ become independent given $G^{i,j}_{t_1,t_2} = 0$.

Therefore, by using Equation 8, $B(0)$ is given by:

$$\begin{aligned}
B(0) &= P(s^i_{t_1})P(s^j_{t_2}) \\
&= \left(P(y^i_{t_1}=0)s_{t_1,0} + P(y^i_{t_1}=1)s_{t_1,1}\right)\left(P(y^j_{t_2}=0)s_{t_2,0} + P(y^j_{t_2}=1)s_{t_2,1}\right) \\
&= \left((1-K_{t_1})\frac{K_{t_1}(1-P_{t_1})}{(1-K_{t_1})P_{t_1}} + K_{t_1}\right)\left((1-K_{t_2})\frac{K_{t_2}(1-P_{t_2})}{(1-K_{t_2})P_{t_2}} + K_{t_2}\right) \\
&= \frac{K_{t_1}}{P_{t_1}}\frac{K_{t_2}}{P_{t_2}}. \qquad (9)
\end{aligned}$$

It remains to derive $A'(0)$. We use the following lemma, derived in Section 2.2 of [3]. If the affection cutoffs of individuals i and j are τ_{t_1} and τ_{t_2}, respectively, then:

$$\begin{aligned}
\frac{d}{dG^{i,j}_{t_1,t_2}} P(y^i_{t_1} = y^j_{t_2} \mid G^{i,j}_{t_1,t_2}) \mid_{G^{i,j}_{t_1,t_2}=0} &= \phi(\tau_{t_1})\phi(\tau_{t_2})\rho_{t_1,t_2} \\
\frac{d}{dG^{i,j}_{t_1,t_2}} P(y^i_{t_1} \neq y^j_{t_2} \mid G^{i,j}_{t_1,t_2}) \mid_{G^{i,j}_{t_1,t_2}=0} &= -\phi(\tau_{t_1})\phi(\tau_{t_2})\rho_{t_1,t_2}. \qquad (10)
\end{aligned}$$

Therefore, A'(0) is explicitly given by:

$$\begin{aligned}
A'(0) = &\sqrt{\frac{P_{t_1}}{1-P_{t_1}}\frac{P_{t_2}}{1-P_{t_2}}} s_{t_1,0}s_{t_2,0}\phi(\tau_{t_1})\phi(\tau_{t_2})\rho_{t_1,t_2} \\
&+ \sqrt{\frac{P_{t_1}}{1-P_{t_1}}\frac{1-P_{t_2}}{P_{t_2}}} s_{t_1,0}s_{t_2,1}\phi(\tau_{t_1})\phi(\tau_{t_2})\rho_{t_1,t_2} \\
&+ \sqrt{\frac{1-P_{t_1}}{P_{t_1}}\frac{P_{t_2}}{1-P_{t_2}}} s_{t_1,1}s_{t_2,0}\phi(\tau_{t_1})\phi(\tau_{t_2})\rho_{t_1,t_2} \\
&+ \sqrt{\frac{1-P_{t_1}}{P_{t_1}}\frac{1-P_{t_2}}{P_{t_2}}} s_{t_1,1}s_{t_2,1}\phi(\tau_{t_1})\phi(\tau_{t_2})\rho_{t_1,t_2}. \qquad (11)
\end{aligned}$$

By incorporating the definition of $s_{t,0}$ in Equation 8 and assuming $s_{t,1}=1$, we obtain:

$$\begin{aligned}
A'(0) = &\frac{K_{t_1}}{1-K_{t_1}}\frac{K_{t_2}}{1-K_{t_2}}\sqrt{\frac{1-P_{t_1}}{P_{t_1}}\frac{1-P_{t_2}}{P_{t_2}}}\phi(\tau_{t_1})\phi(\tau_{t_2})\rho_{t_1,t_2} \\
&+ \frac{K_{t_1}}{1-K_{t_1}}\sqrt{\frac{1-P_{t_1}}{P_{t_1}}\frac{1-P_{t_2}}{P_{t_2}}}\phi(\tau_{t_1})\phi(\tau_{t_2})\rho_{t_1,t_2} \\
&+ \frac{K_{t_2}}{1-K_{t_2}}\sqrt{\frac{1-P_{t_1}}{P_{t_1}}\frac{1-P_{t_2}}{P_{t_2}}}\phi(\tau_{t_1})\phi(\tau_{t_2})\rho_{t_1,t_2} \\
&+ \sqrt{\frac{1-P_{t_1}}{P_{t_1}}\frac{1-P_{t_2}}{P_{t_2}}}\phi(\tau_{t_1})\phi(\tau_{t_2})\rho_{t_1,t_2} \\
= &\frac{\sqrt{\frac{(1-P_{t_1})(1-P_{t_2})}{P_{t_1}P_{t_2}}}\phi(\tau_{t_1})\phi(\tau_{t_2})\rho_{t_1,t_2}}{(1-K_{t_1})(1-K_{t_2})}. \qquad (12)
\end{aligned}$$

Finally, we combine Equations 9 and 12 into Equation 6 to obtain:

$$\begin{aligned}
E\left[\tilde{y}^i_{t_1}\tilde{y}^j_{t_2} \mid s^i_{t_1}, s^j_{t_2}, G^{i,j}_{t_1,t_2}\right] &= \frac{\frac{\sqrt{\frac{(1-P_{t_1})(1-P_{t_2})}{P_{t_1}P_{t_2}}}\phi(\tau_{t_1})\phi(\tau_{t_2})\rho_{t_1,t_2}}{(1-K_{t_1})(1-K_{t_2})}}{\frac{K_{t_1}}{P_{t_1}}\frac{K_{t_2}}{P_{t_2}}} G^{i,j}_{t_1,t_2} + \mathcal{O}\left((G^{i,j}_{t_1,t_2})^2\right) \\
&= \frac{\sqrt{P_{t_1}(1-P_{t_1})P_{t_2}(1-P_{t_2})}\phi(\tau_{t_1})\phi(\tau_{t_2})G^{i,j}_{t_1,t_2}}{K_{t_1}(1-K_{t_1})K_{t_2}(1-K_{t_2})}\rho_{t_1,t_2} + \mathcal{O}\left((G^{i,j}_{t_1,t_2})^2\right). \qquad (13)
\end{aligned}$$

This completes the derivation.

4 PCGC with Covariates

Here we derive the PCGC genetic covariance estimator in the presence of covariates. We extend the model presented in the previous section as follows. We assume that every individual in study t carries a vector of covariates $\boldsymbol{C}_t^i$, including an intercept. The liability a_t^i is now given by $a_t^i = g_t^i + e_t^i + (\boldsymbol{C}_t^i)^T \boldsymbol{\beta}_t$, where $\boldsymbol{\beta}_t$ is a vector of fixed effects. Denote P_t^i as the in-sample probability of individual i in study t being a case conditional on her covariates, $P_t^i = P(y_t^i = 1 \,|\, \boldsymbol{C}_t^i, s_t^i = 1\,;\, \boldsymbol{\beta}_t)$. We define the standardized phenotype of individual i as $\tilde{y}_t^i = (y_t^i - P_t^i)/\sqrt{P_t^i(1-P_t^i)}$.

We show below that the first order Taylor expansion of the conditional expectation of $\tilde{y}_{t_1}^i \tilde{y}_{t_2}^j$ is now given by:

$$\begin{aligned} E &\left[\tilde{y}_{t_1}^i \tilde{y}_{t_2}^j \,|\, \boldsymbol{C}_{t_1}^i, \boldsymbol{C}_{t_2}^j, s_{t_1}^i, s_{t_2}^j, G_{t_1,t_2}^{i,j}\right] \\ &= G_{t_1,t_2}^{i,j} f\left(\boldsymbol{C}_{t_1}^i, \boldsymbol{C}_{t_2}^j\,;\, \boldsymbol{\beta}_{t_1}, \boldsymbol{\beta}_{t_2}, t_1, t_2\right) \rho_{t_1,t_2} + \mathcal{O}\left(\left(G_{t_1,t_2}^{i,j}\right)^2\right), \end{aligned} \tag{14}$$

where $f\left(\boldsymbol{C}_{t_1}^i, \boldsymbol{C}_{t_2}^j\,;\, \boldsymbol{\beta}_{t_1}, \boldsymbol{\beta}_{t_2}, t_1, t_2\right)$ depends on the covariates of individuals i and j, on the fixed effects and on the case-control proportions and the prevalences of the two studied traits, and is explicitly given by:

$$\begin{aligned} f\left(\boldsymbol{C}_{t_1}^i, \boldsymbol{C}_{t_2}^j\,;\, \boldsymbol{\beta}_{t_1}, \boldsymbol{\beta}_{t_2}, t_1, t_2\right) \triangleq & \frac{\phi(\tau_{t_1}^i)}{\sqrt{P_{t_1}^i(1-P_{t_1}^i)}\left(K_{t_1}^i + (1-K_{t_1}^i)\frac{K_{t_1}(1-P_{t_1})}{P_{t_1}(1-K_{t_1})}\right)} \\ & \frac{\phi(\tau_{t_2}^j)}{\sqrt{P_{t_2}^j(1-P_{t_2}^j)}\left(K_{t_2}^j + (1-K_{t_2}^j)\frac{K_{t_2}(1-P_{t_2})}{P_{t_2}(1-K_{t_2})}\right)} \\ & \left[q_{t_1,t_2;0,0}^{i,j} + q_{t_1,t_2;0,1}^{i,j} + q_{t_1,t_2;1,0}^{i,j} + q_{t_1,t_2;1,1}^{i,j}\right], \end{aligned} \tag{15}$$

where $K_t^i = 1 - (1-P_t^i) \,/\, \left(1 + \frac{K_t(1-P_t)}{P_t(1-K_t)} P_t^i - P_t^i\right)$ is the population-level probability of being a case (derived in [3]), $\tau_t^i = \Phi^{-1}\left(1-K_t^i\right)$ is the individual-level affection cutoff, and the terms in the parentheses are given by:

$$\begin{aligned} q_{t_1,t_2;0,0}^{i,j} &= \frac{K_{t_1}(1-P_{t_1})}{P_{t_1}(1-K_{t_1})} \frac{K_{t_2}(1-P_{t_2})}{P_{t_2}(1-K_{t_2})} P_{t_1}^i P_{t_2}^j. \\ q_{t_1,t_2;0,1}^{i,j} &= \frac{K_{t_1}(1-P_{t_1})}{P_{t_1}(1-K_{t_1})} P_{t_1}^i (1-P_{t_2}^j) \\ q_{t_1,t_2;1,0}^{i,j} &= \frac{K_{t_2}(1-P_{t_2})}{P_{t_2}(1-K_{t_2})} (1-P_{t_1}^i) P_{t_2}^j \\ q_{t_1,t_2;1,1}^{i,j} &= (1-P_{t_1}^i)(1-P_{t_2}^j). \end{aligned} \tag{16}$$

Unlike the previous section, the term $f\left(\boldsymbol{C}_{t_1}^i, \boldsymbol{C}_{t_2}^j\,;\, \boldsymbol{\beta}_{t_1}, \boldsymbol{\beta}_{t_2}, t_1, t_2\right)$ is different for every pair of individuals. We first derive Equation 14 under the assumption that the fixed effects are known, and then describe estimation with unknown fixed effects.

The derivation of Equation 14 is carried out as follows. As before, we begin by writing down the conditional expectation of $\tilde{y}_{t_1}^i \tilde{y}_{t_2}^j$ and use Bayes rule and the conditional independence assumptions to obtain:

$$
\begin{aligned}
&E\left[\tilde{y}^i_{t_1}\tilde{y}^j_{t_2} \,|\, \boldsymbol{C}^i_{t_1}, \boldsymbol{C}^j_{t_2}, s^i_{t_1}, s^j_{t_2}, G^{i,j}_{t_1,t_2}\right] = \\
&\sum_{y^i_{t_1}, y^j_{t_2}=0}^{1} \frac{y^i_{t_1} - P^i_{t_1}}{\sqrt{P^i_{t_1}(1-P^i_{t_1})}} \frac{y^j_{t_2} - P^j_{t_2}}{\sqrt{P^j_{t_2}(1-P^j_{t_2})}} P(y^i_{t_1}, y^j_{t_2} \,|\, \boldsymbol{C}^i_{t_1}, \boldsymbol{C}^j_{t_2}, s^i_{t_1}, s^j_{t_2}, G^{i,j}_{t_1,t_2}) \\
&= \frac{\sum_{y^i_{t_1}, y^j_{t_2}=0}^{1} \frac{y^i_{t_1} - P^i_{t_1}}{\sqrt{P^i_{t_1}(1-P^i_{t_1})}} \frac{y^j_{t_2} - P^j_{t_2}}{\sqrt{P^j_{t_2}(1-P^j_{t_2})}} P(y^i_{t_1}, y^j_{t_2} | \boldsymbol{C}^i_{t_1}, \boldsymbol{C}^j_{t_2}, G^{i,j}_{t_1,t_2}) P(s^i_{t_1} \,|\, y^i_{t_1}) P(s^j_{t_2} | y^j_{t_2})}{P(s^i_{t_1}, s^j_{t_2} \,|\, \boldsymbol{C}^i_{t_1}, \boldsymbol{C}^j_{t_2}, G^{i,j}_{t_1,t_2})}.
\end{aligned}
\tag{17}
$$

Next, we approximate Equation 17 via a Taylor expansion around $G^{i,j}_{t_1,t_2} = 0$. As before, we denote the numerator and denominator as $A(G^{i,j}_{t_1,t_2})$ and $B(G^{i,j}_{t_1,t_2})$, respectively. The term $A(0)$ is once again equal to 0, as can be verified by seeing that setting $G^{i,j}_{t_1,t_2} = 0$ in the first order Taylor expansion of the expression $A(G^{i,j}_{t_1,t_2})/B(G^{i,j}_{t_1,t_2})$ leads to the expression $A(0)/B(0)$ on the one hand, but that the conditional expectation $E\left[\tilde{y}^i_{t_1}\tilde{y}^j_{t_2} \,|\, \boldsymbol{C}^i_{t_1}, \boldsymbol{C}^j_{t_2}, s^i_{t_1}, s^j_{t_2}, G^{i,j}_{t_1,t_2} = 0\right]$ decomposes into a product of conditional expectations of $\tilde{y}^i_{t_1}$ and of $\tilde{y}^j_{t_2}$, each of which is equal to 0 by definition. We therefore once again have a Taylor expansion of the form:

$$
E\left[\tilde{y}^i_{t_1}\tilde{y}^j_{t_2} \,|\, \boldsymbol{C}^i_{t_1}, \boldsymbol{C}^j_{t_2}, s^i_{t_1}, s^j_{t_2}, G^{i,j}_{t_1,t_2}\right] = \frac{A'(0)}{B(0)} G^{i,j}_{t_1,t_2} + \mathcal{O}\left((G^{i,j}_{t_1,t_2})^2\right). \tag{18}
$$

To compute $B(0)$, we use the fact that the variables $s^i_{t_1}, s^j_{t_2}$ become independent given $G^{i,j}_{t_1,t_2} = 0$ and the covariates. Therefore, by using Equation 8, $B(0)$ is given by:

$$
\begin{aligned}
B(0) &= P(s^i_{t_1} \,|\, \boldsymbol{C}^i_{t_1}) P(s^j_{t_2} \,|\, \boldsymbol{C}^j_{t_2}) \\
&= \left(K^i_{t_1} + (1-K^i_{t_1}) \frac{K_{t_1}(1-P_{t_1})}{(1-K_{t_1})P_{t_1}}\right) \left(K^j_{t_2} + (1-K^j_{t_2}) \frac{K_{t_2}(1-P_{t_2})}{(1-K_{t_2})P_{t_2}}\right).
\end{aligned}
\tag{19}
$$

To compute $A'(0)$, we rewrite the numerator of Equation 17 using the results in Equation 10, and additionally incorporate Equation 8 as follows:

$$
\begin{aligned}
A'(0) = &\frac{P^i_{t_1} P^j_{t_2}}{\sqrt{P^i_{t_1}(1-P^i_{t_1}) P^j_{t_2}(1-P^j_{t_2})}} \frac{K_{t_1}(1-P_{t_1})}{P_{t_1}(1-K_{t_1})} \frac{K_{t_2}(1-P_{t_2})}{P_{t_2}(1-K_{t_2})} \phi(\tau^i_{t_1})\phi(\tau^j_{t_2})\rho_{t_1,t_2} \\
&+ \frac{P^i_{t_1}(1-P^j_{t_2})}{\sqrt{P^i_{t_1}(1-P^i_{t_1}) P^j_{t_2}(1-P^j_{t_2})}} \frac{K_{t_1}(1-P_{t_1})}{P_{t_1}(1-K_{t_1})} \phi(\tau^i_{t_1})\phi(\tau^j_{t_2})\rho_{t_1,t_2} \\
&+ \frac{(1-P^i_{t_1})P^j_{t_2}}{\sqrt{P^i_{t_1}(1-P^i_{t_1}) P^j_{t_2}(1-P^j_{t_2})}} \frac{K_{t_2}(1-P_{t_2})}{P_{t_2}(1-K_{t_2})} \phi(\tau^i_{t_1})\phi(\tau^j_{t_2})\rho_{t_1,t_2}\phi(\tau^i_{t_1})\phi(\tau^j_{t_2})\rho_{t_1,t_2} \\
&+ \frac{(1-P^i_{t_1})(1-P^j_{t_2})}{\sqrt{P^i_{t_1}(1-P^i_{t_1}) P^j_{t_2}(1-P^j_{t_2})}} \phi(\tau^i_{t_1})\phi(\tau^j_{t_2})\rho_{t_1,t_2}.
\end{aligned}
\tag{20}
$$

Equation 14 is obtained by combining Equations 19 and 20 into Equation 18. This completes the derivation.

When the fixed effects are unknown we carry out a two steps procedure, as explained in [3]. In the first stage we estimate the fixed effects while ignoring the covariance structure via logistic regression. The theory of generalized estimating equations shows that such an estimation procedure tends to produce highly accurate estimates [4] (the formula for the variance of the estimators needs to be modified to account for the covariance structure, but this is out of the scope of our work). In the second stage we use the estimated fixed effects to compute a conditional in-sample affection probability $P_t^i = P(y_t^i = 1 \,|\, \boldsymbol{C}_t^i, s_{t_1}^i \,;\, \boldsymbol{\beta})$, which enables us to use the Taylor approximation described above.

5 Adapting PCGC to use Summary Statistics

Here we describe how PCGC can be modified to use summary statistics. Our derivation assumes the presence of covariates. Settings without covariates can be regarded as a special case with a single constant covariate carried by all individuals (a so-called intercept). To avoid dependency on the previous section, we first reestablish the relevant notations.

Denote P_t as the proportion of cases in study t and P_t^i as the in-sample probability of individual i in study t of being a case conditional on her covariates. Further denote $K_t^i = 1 - \left(1 - P_t^i\right) / \left(1 + \frac{K_t(1-P_t)}{P_t(1-K_t)} P_t^i - P_t^i\right)$ as the population-level probability of being a case, and define $\tau_t^i = \Phi^{-1}\left(1 - K_t^i\right)$. Note that in the absence of covariates $P_t^i = P_t$, $K_t^i = K_t$ and $\tau_t^i = \tau_t$ for all individuals.

The PCGC covariance estimator is given by regressing the conditionally-standardized phenotype products $\frac{y_{t_1}^i - P_{t_1}^i}{\sqrt{P_{t_1}^i(1-P_{t_1}^i)}} \frac{y_{t_2}^j - P_{t_2}^j}{\sqrt{P_{t_2}^j(1-P_{t_2}^j)}}$ on the conditionally-modified genotype products $G_{t_1,t_2}^{i,j} Q_{t_1,t_2}^{i,j}$, where $Q_{t_1,t_2}^{i,j}$ is given by:

$$Q_{t_1,t_2}^{i,j} \triangleq \sum_{a,b=0}^{1} u_{t_1,a}^i u_{t_2,b}^j, \tag{21}$$

where

$$u_{t,0}^i \triangleq \frac{\phi(\tau_t^i)}{\sqrt{P_t^i(1-P_t^i)}\left(K_t^i + (1-K_t^i)\frac{K_t(1-P_t)}{P_t(1-_t)}\right)} \frac{K_t(1-P_t)}{P_t(1-K_t)} P_t^i \tag{22}$$

$$u_{t,1}^i \triangleq \frac{\phi(\tau_t^i)}{\sqrt{P_t^i(1-P_t^i)}\left(K_t^i + (1-K_t^i)\frac{K_t(1-P_t)}{P_t(1-_t)}\right)} \left(1 - P_t^i\right). \tag{23}$$

Note that each term $u_{t,0}^i$ and $u_{t,1}^i$ depends only on information from study t.

A key ingredient in the adaptation of PCGC for summary statistics is the assumed form of the genetic similarity coefficients:

$$G_{t_1,t_2}^{i,j} \triangleq \frac{1}{m} \sum_{k=1}^{m} X_{t_1}^{k,i} X_{t_2}^{k,j}, \tag{24}$$

where $X_t^{k,i}$ is the k_{th} variant carried by individual i in study t, after normalization at the population level. Therefore, both $G_{t_1,t_2}^{i,j}$ and $Q_{t_1,t_2}^{i,j}$ are given by sums of products of terms, where each term depends only on an individual from one of the two studies. This is the underlying idea that enables to compute the PCGC estimator via summary statistics. However, we note that is is straightforward to extend our results to accommodate frequency or LD-dependent architectures or multiple variance components, as shown in Sections 6 and 7.

We now provide the full derivation of our results. The PCGC covariance estimator is explicitly given by:

$$\hat{\rho}_{t_1,t_2}^{\text{pcgc-covar}} = \frac{\sum_{i,j \notin S_{t_1,t_2}} \tilde{y}_{t_1}^i \tilde{y}_{t_2}^j G_{t_1,t_2}^{i,j} Q_{t_1,t_2}^{i,j}}{\sum_{i,j \notin S_{t_1,t_2}} \left(G_{t_1,t_2}^{i,j} Q_{t_1,t_2}^{i,j} \right)^2}, \tag{25}$$

where S_{t_1,t_2} is the set of all pairs of indices i,j that refer to the same individual who is shared between the two studies, and $\tilde{y}_t^i \triangleq \frac{y_{t_1}^i - P_{t_1}^i}{\sqrt{P_{t_1}^i (1-P_{t_1}^i)}}$.

To compute Equation 25 without having access to genetic and phenotypic data, we require the following summary statistics:

$$\begin{aligned} z_t^{k,\text{covar}} &\triangleq \sum_i \tilde{y}_t^i X_t^{k,i} \sum_{a=0}^{1} u_{t,a}^i \\ \hat{r}_t^{k,h,\text{covar}} &\triangleq \sum_i X_t^{k,i} X_t^{h,i} \sum_{a,b=0}^{1} u_{t,a}^i u_{t,b}^i. \end{aligned} \tag{26}$$

If the two studies include overlapping individuals, we also require the following summary statistics for each of the overlapping individuals:

$$w_t^i \triangleq \sqrt{G_{t,t}^{i,i}} \tilde{y}_t^i \left(\sum_{a=0}^{1} u_{t,a}^i \right). \tag{27}$$

The summary statistic w_t^i are not privacy preserving because they expose (a noisy version of) the phenotype of individual i, and some indirect information about her covariates. This is often not a problem, because overlapping individuals often consist of control cohorts, in which the phenotypes are already known. Nevertheless, we propose a privacy-preserving approximation in Section 5.2.

We now describe how Equation 25 can be rewritten to use only the above summary statistics. The numerator of Equation 25 can be rewritten to use only summary statistics as follows. We first decompose the numerator into two terms:

$$\sum_{i,j \notin S_{t_1,t_2}} \tilde{y}_{t_1}^i \tilde{y}_{t_2}^j G_{t_1,t_2}^{i,j} Q_{t_1,t_2}^{i,j} = \sum_{i,j} \tilde{y}_{t_1}^i \tilde{y}_{t_2}^j G_{t_1,t_2}^{i,j} Q_{t_1,t_2}^{i,j} - \sum_{i,j \in S_{t_1,t_2}} \tilde{y}_{t_1}^i \tilde{y}_{t_2}^j G_{t_1,t_2}^{i,j} Q_{t_1,t_2}^{i,j}. \tag{28}$$

We will handle each term separately. By using Equations 24 and 21, the first term on the right hand side of Equation 28 can be reformulated as follows:

$$\begin{aligned} \sum_{i,j} \tilde{y}_{t_1}^i \tilde{y}_{t_2}^j G_{t_1,t_2}^{i,j} Q_{t_1,t_2}^{i,j} &= \sum_{i,j} \tilde{y}_{t_1}^i \tilde{y}_{t_2}^j \left(\frac{1}{m} \sum_{k=1}^{m} X_{t_1}^{k,i} X_{t_2}^{k,j} \right) \left(\sum_{a,b=0}^{1} u_{t_1,a}^i u_{t_2,b}^j \right) \\ &= \frac{1}{m} \sum_{k=1}^{m} \left(\sum_i \tilde{y}_{t_1}^i X_{t_1}^{k,i} \sum_{a=0}^{1} u_{t_1,a}^i \right) \left(\sum_j \tilde{y}_{t_2}^j X_{t_2}^{k,j} \sum_{b=0}^{1} u_{t_2,b}^j \right) \\ &= \frac{1}{m} \sum_{k=1}^{m} z_{t_1}^{k,\text{covar}} z_{t_2}^{k,\text{covar}}. \end{aligned} \tag{29}$$

Using Equations 21 and 27, the second term on the right hand side of Equation 28 can be rewritten as follows:

$$\begin{aligned}
\sum_{i,j \in S_{t_1,t_2}} \tilde{y}^i_{t_1} \tilde{y}^j_{t_2} G^{i,j}_{t_1,t_2} Q^{i,j}_{t_1,t_2} &= \sum_{i,j \in S_{t_1,t_2}} \tilde{y}^i_{t_1} \tilde{y}^j_{t_2} \sqrt{G^{i,i}_{t_1,t_1} G^{j,j}_{t_2,t_2}} \sum_{a,b=0}^{1} u^i_{t_1,a} u^j_{t_2,b} \\
&= \sum_{i,j \in S_{t_1,t_2}} \left(\sqrt{G^{i,i}_{t_1,t_1}} \tilde{y}^i_{t_1} \sum_{a=0}^{1} u^i_{t_1,a} \right) \left(\sqrt{G^{j,j}_{t_2,t_2}} \tilde{y}^j_{t_2} \sum_{b=0}^{1} u^j_{t_2,b} \right) \\
&= \sum_{i,j \in S_{t_1,t_2}} w^i_{t_1} w^j_{t_2}. \qquad (30)
\end{aligned}$$

The derivation in Equation 30 requires having access to the genotypes and covariates of overlapping individuals. If there are no covariates and the number of overlapping individuals having each of the four possible combinations of phenotypes is known, Equation 30 can be simplified considerably by using the approximation $G^{i,j}_{t_1,t_2} \approx 1.0$ for overlapping individuals. However, we caution that this approximation is very sensitive to the data preprocessing, because $G^{i,j}_{t_1,t_2} \neq \sqrt{G^{i,i}_{t_1,t_1}} \sqrt{G^{j,j}_{t_2,t_2}}$ if studies t_1, t_2 were preprocessed separately (see Supplemental section on the effects of preprocessing the data for a discussion of this issue).

If a third party with no access to the overlapping individuals wishes to approximate the second term on the right hand side of Equation 28, she may do so using a sum of Taylor Expansions around the mean covariates vector for each of the four possible combinations of phenotypes of overlapping individuals. Typically the only overlapping individuals are controls, which simplifies this approximation. The derivation is provided in Section 5.2.

We now consider the denominator of Equation 25. As in the analysis of the numerator, we first decompose the denominator into two terms:

$$\sum_{i,j \notin S_{t_1,t_2}} \left(G^{i,j}_{t_1,t_2} Q^{i,j}_{t_1,t_2} \right)^2 = \sum_{i,j} \left(G^{i,j}_{t_1,t_2} Q^{i,j}_{t_1,t_2} \right)^2 - \sum_{i,j \in S_{t_1,t_2}} \left(G^{i,j}_{t_1,t_2} Q^{i,j}_{t_1,t_2} \right)^2. \qquad (31)$$

As before, we will handle each term separately. The first term on the right hand side of Equation 31 can be reformulated as follows:

$$\begin{aligned}
\sum_{i,j} \left(G^{i,j}_{t_1,t_2} Q^{i,j}_{t_1,t_2} \right)^2 &= \sum_{i,j} \left(\left(\frac{1}{m} \sum_{k=1}^{m} X^{k,i}_{t_1} X^{k,j}_{t_2} \right) \left(\sum_{a,b=0}^{1} u^i_{t_1,a} u^j_{t_2,b} \right) \right)^2 \\
&= \frac{1}{m_2} \sum_{i,j} \left(\sum_{k,h=1}^{m} X^{k,i}_{t_1} X^{k,j}_{t_2} X^{h,i}_{t_1} X^{h,j}_{t_2} \right) \left(\sum_{a,b,c,d=0}^{1} u^i_{t_1,a} u^j_{t_2,b} u^i_{t_1,c} u^j_{t_2,d} \right) \\
&= \frac{1}{m_2} \sum_{k,h=1}^{m} \left(\sum_{i} X^{k,i}_{t_1} X^{h,i}_{t_1} \sum_{a,c=0}^{1} u^i_{t_1,a} u^i_{t_1,c} \right) \left(\sum_{j} X^{k,j}_{t_2} X^{h,j}_{t_2} \sum_{b,d=0}^{1} u^j_{t_2,b} u^j_{t_2,d} \right) \\
&= \frac{1}{m_2} \sum_{k,h=1}^{m} \hat{r}^{k,h,\text{covar}}_{t_1} \hat{r}^{k,h,\text{covar}}_{t_2}. \qquad (32)
\end{aligned}$$

A possible drawback of the summary statistics $\hat{r}^{k,h,\text{covar}}_{t}$ is their large number. If the linkage disequilibrium (LD) patterns within both studies are approximately the same as in a reference population based on which LD patterns were computed, we can carry out an approximate analysis with only a single summary statistics, as described in Section 5.1.

Finally, the second term on the right hand side of Equation 31 can be easily computed by a research group with access to the genotypes and covariates of overlapping individuals (only covariate can suffice when using the approximation $G^{i,j}_{t_1,t_2} \approx 1$). Otherwise, we describe an approximation of this term in Section 5.2.

5.1 Approximate Summary Statistics without LD

Recall from Equation 32 that computation of the sum $\sum_{i,j}\left(G^{i,j}_{t_1,t_2}Q^{i,j}_{t_1,t_2}\right)^2$ via summary statistics requires summary statistics for every pair of variants. Here we describe an approximation that requires only a single summary statistics, $E\left[Q^{i,i}_{t,t}\right]$, and empirically yields very accurate approximations.

The approximation consists of assuming independence between covariates and genetic variants (technically, one needs to assume only that for every pair of individuals i,j and pair of variants k,h, the covariates of individuals i,j are independent of the product $X^{k,i}_{t_1}X^{h,i}_{t_1}X^{k,j}_{t_2}X^{h,j}_{t_2}$, which is a very mild assumption). Using this assumption and the law of large numbers, the denominator of Equation 25 can be approximated as:

$$\begin{aligned}
\sum_{i,j\notin S_{t_1,t_2}} \left(G^{i,j}_{t_1,t_2}Q^{i,j}_{t_1,t_2}\right)^2 &\approx \left|\{(i,j)\,|\,(i,j)\notin S_{t_1,t_2}\}\right| E_{i,j\notin S_{t_1,t_2}}\left[\left(G^{i,j}_{t_1,t_2}Q^{i,j}_{t_1,t_2}\right)^2\right] \\
&\approx \left|\{(i,j)\,|\,(i,j)\notin S_{t_1,t_2}\}\right| E_{i,j\notin S_{t_1,t_2}}\left[\left(G^{i,j}_{t_1,t_2}\right)^2\right] E_{i,j\notin S_{t_1,t_2}}\left[\left(Q^{i,j}_{t_1,t_2}\right)^2\right] \\
&= \left|\{(i,j)\,|\,(i,j)\notin S_{t_1,t_2}\}\right| E_{i,j\notin S_{t_1,t_2}}\left[\left(G^{i,j}_{t_1,t_2}\right)^2\right] E_{i,j\notin S_{t_1,t_2}}\left[\left(\sum_{a,b=0}^{1} u^i_{t1,a}u^j_{t2,b}\right)^2\right] \\
&= \left|\{(i,j)\,|\,(i,j)\notin S_{t_1,t_2}\}\right| E_{i,j\notin S_{t_1,t_2}}\left[\left(G^{i,j}_{t_1,t_2}\right)^2\right] E_{i,j\notin S_{t_1,t_2}}\left[\sum_{a,b,c,d=0}^{1} u^i_{t1,a}u^i_{t1,c}u^j_{t2,b}u^j_{t2,d}\right] \\
&= \left|\{(i,j)\,|\,(i,j)\notin S_{t_1,t_2}\}\right| E_{i,j\notin S_{t_1,t_2}}\left[\left(G^{i,j}_{t_1,t_2}\right)^2\right] E_{i,j\notin S_{t_1,t_2}}\left[Q^{i,i}_{t_1}Q^{j,j}_{t_2}\right]. \qquad (33)
\end{aligned}$$

To proceed, we first make use of the fact that when the in-sample LD patterns in both studies are the same, we have:

$$E_{i,j\notin S_{t_1,t_2}}\left[\left(G^{i,j}_{t_1,t_2}\right)^2\right] = \frac{n_{t_1}n_{t_2}}{m}E\left[\ell\right], \qquad (34)$$

where $E[\ell]$ is the unbiased estimate of the mean LD score among all genetic variants in the data (see [5] for an explanation). Second, we use the fact that $Q^{i,i}_{t_1}$, $Q^{j,j}_{t_2}$ are independent for $i,j\notin S_{t_1,t_2}$, as they depend only on the covariates of individuals i,j, which are sampled from their respective distributions. Using these facts, Equation 33 can be approximated as:

$$\sum_{i,j\notin S_{t_1,t_2}} \left(G^{i,j}_{t_1,t_2}Q^{i,j}_{t_1,t_2}\right)^2 \approx \frac{n_{t_1}n_{t_2}}{m}E\left[\ell\right]\,E\left[Q^{i,i}_{t_1}\right]E\left[Q^{j,j}_{t_2}\right]. \qquad (35)$$

We conclude that the denominator of the PCGC estimator (Equation 25) can be approximated as $\frac{n_{t_1}n_{t_2}}{m}E\left[\ell\right]\,E\left[Q^{i,i}_{t_1}\right]E\left[Q^{j,j}_{t_2}\right]$.

Finally, we note that if the covariate-genotypes independence assumption above does not exactly hold, one can obtain a better fit by assuming conditional independence given phenotypes, and then apply the approximation:

$$\sum_{i,j\notin S_{t_1,t_2}} \left(G^{i,j}_{t_1,t_2} Q^{i,j}_{t_1,t_2}\right)^2 \approx N^{\text{cas,cas}} E\left[\left(G^{\text{cas,cas}}_{t_1,t_2} Q^{\text{cas,cas}}_{t_1,t_2}\right)^2\right] + N^{\text{cas,con}} E\left[\left(G^{\text{cas,con}}_{t_1,t_2} Q^{\text{cas,con}}_{t_1,t_2}\right)^2\right]$$
$$+ N^{\text{con,cas}} E\left[\left(G^{\text{con,cas}}_{t_1,t_2} Q^{\text{con,cas}}_{t_1,t_2}\right)^2\right] + N^{\text{con,con}} E\left[\left(G^{\text{con,con}}_{t_1,t_2} Q^{\text{con,con}}_{t_1,t_2}\right)^2\right], \tag{36}$$

where $N^{\text{cas,cas}}$ is the number of non-overlapping individuals in the two studies who are cases for both traits, $E\left[\left(G^{\text{cas,cas}}_{t_1,t_2} Q^{\text{cas,cas}}_{t_1,t_2}\right)^2\right]$ is the mean value of of $(G^{i,j}_{t_1,t_2} Q^{i,j}_{t_1,t_2})^2$ for a pair of cases in the two studies, and the other quantities are defined analogously. One can then apply the approximation for each of the four summands separately. This approximation is typically not required because overlapping individuals consist mainly of shared controls.

5.2 Third Party Approximations

If two studies include overlapping individuals, the PCGC estimators cannot be computed exactly by a third party with no access to the covariates of these overlapping individuals. Here we propose a summary statistics based approximation. Recall that the denominator of the PCGC estimator (Equation 25) can be approximated without access to individual-level data using the approximation described in Section 5.1. We are therefore left with the task of approximating the second term in the numerator of Equation 25, given by $\sum_{i,j\in S_{t_1,t_2}} \tilde{y}^i_{t_1} \tilde{y}^j_{t_2} G^{i,j}_{t_1,t_2} Q^{i,j}_{t_1,t_2}.$.

As a first step, we can approximate $G^{i,j} \approx 1$ since i and j refer to the same individual, which simplifies this term to $\sum_{i,j\in S_{t_1,t_2}} \tilde{y}^i_{t_1} \tilde{y}^j_{t_2} Q^{i,j}_{t_1,t_2}$. We propose approximating this term by approximating the expectation for every combination of the two phenotypes:

$$\sum_{i,j\in S_{t_1,t_2}} \tilde{y}^i_{t_1} \tilde{y}^j_{t_2} Q^{i,j}_{t_1,t_2} \approx \sum_{y^i_{t_1}, y^j_{t_2} \in\{0,1\}} n^{y^i_{t_1}, y^j_{t_2}}_{t_1,t_2} E\left[\tilde{y}^i_{t_1} \tilde{y}^j_{t_2} Q^{i,j}_{t_1,t_2} \,\middle|\, y^i_{t_1}, y^j_{t_2}\right], \tag{37}$$

where $n^{y^i_{t_1}, y^j_{t_2}}_{t_1,t_2}$ is the number of overlapping individuals having phenotypes $y^i_{t_1}$ and $y^j_{t_2}$. However, unlike before we cannot make independence assumptions because the terms in the expectations refer to the same individuals, and therefore terms belonging to the two studies are likely to use the same covariates.

Instead, we propose to use summary statistics of the mean covariates vector for every combination of phenotypes, $E\left[\boldsymbol{C^i_{t_1}}; \boldsymbol{C^j_{t_2}} \mid y^i_{t_1}, y^j_{t_2}\right]$. Typically this is feasible because overlapping individuals consist almost exclusively of controls. Using this information, we can approximate the term $E\left[\tilde{y}^i_{t_1} \tilde{y}^j_{t_2} Q^{i,j}_{t_1,t_2} \mid y^i_{t_1}, y^j_{t_2}\right]$ via a first order Taylor expansion of $\tilde{y}^i_{t_1} \tilde{y}^j_{t_2} Q^{i,j}_{t_1,t_2}$ around the mean covariate values, respectively. Since the first-order Taylor expansion is linear in the covariates, the approximate expectation is linear in the mean covariates vector. Although the derivations are technically straightforward, they are extremely tedious to write down explicitly. Nevertheless, it is not difficult to code these computations algorithmically.

6 Allele-frequency and LD Dependent Genetic Architectures

The derivations in the sections above assume a genetic similarity matrix of the form $\mathbf{G} = \mathbf{X}\mathbf{X}^T/m$. It is possible to consider alternative architectures, which assign different weights to different SNPs based on their MAF levels, their LD-scores, or other properties [6]. In this case, the genetic similarity matrix can be written as $\mathbf{G} = \mathbf{X}\mathbf{W}\mathbf{X}^T/M$, where $\mathbf{W}$ is an $m \times m$ diagonal matrix of SNP weights, and $M = \sum_k W^{kk}$ is a normalization factor which guarantees that the mean entry in the diagonal of the resulting matrix is 1.0. Assuming that the weights are known, it is straightforward to adapt PCGC and PCGC-s for such architectures, as we now describe.

Adapting PCGC for alternative architectures is straightforward, by using the correct (weighted) form of the genetic similarity entries $G^{i,j}_{t_1,t_2}$ in Equation 25. To adapt PCGC-s, we need to adapt the summary statistics in Equation 26 by (a) multiplying each summary statistic $z_t^{k,\text{covar}}$ by $\sqrt{W^{kk}/M}$, and (b) multiplying each summary statistic $\hat{r}_t^{k,h,\text{covar}}$ by $\sqrt{W^{kk}W^{hh}}/M$.

Instead of using the summary statistics $\hat{r}_t^{k,h,\text{covar}}$, we can approximate the denominator of the modified form of Equation 25 (as described in Section 5.1) by replacing the average LD score $E[\ell] = \sum_{k,h} \left(\hat{r}^{kh}\right)^2/m$ in Equation 35 with the term $\sum_{k,h} W^{kk}W^{hh}\left(\hat{r}^{kh}\right)^2/M$.

Hence, we can approximate the denominator of the PCGC-s estimator via:

$$n_{t_1}n_{t_2}\frac{1}{M}\sum_{k,h} W^{kk}W^{hh}\left(\hat{r}^{kh}\right)^2 E\left[Q^{i,i}_{t_1}\right] E\left[Q^{j,j}_{t_2}\right]. \tag{38}$$

Importantly, the above derivation demonstrates that we can carry out estimation using the exact same summary statistics as in the unweighted version of PCGC-s. Hence, it is possible to evaluate heritability and genetic correlation under various sets of modeling assumptions given a single set of summary statistics.

7 Extension to Multiple Variance Components

The derivations in the sections above describe estimation of a single variance component. The extension to multiple variance components is straightforward [3]. In the presence of multiple variance components, the PCGC estimator is obtained via a multivariate Taylor expansion of the form:

$$\begin{aligned} &E\left[\tilde{y}^i_{t_1}\tilde{y}^j_{t_2} \mid \boldsymbol{C}^i_{t_1}, \boldsymbol{C}^j_{t_2}, s^i_{t_1}, s^j_{t_2}, G^{i,j}_{t_1,t_2}\right] \\ &\qquad = \sum_{v=1}^{V} G^{i,j}_{t_1,t_2;\,v} Q^{i,j}_{t_1,t_2}\rho_{t_1,t_2;\,v} + \sum_{v=1}^{V}\mathcal{O}\left(\left(G^{i,j}_{t_1,t_2;\,v}\right)^2\right), \end{aligned} \tag{39}$$

where V is the number of variance components, $G^{i,j}_{t_1,t_2;\,v}$ is the genetic similarity coefficient between individuals i and j according to variance component v and $\rho_{t_1,t_2;\,v}$ is the corresponding coefficient. The multivariate regression estimator is now given by:

$$\hat{\boldsymbol{\rho}}^{\text{pcgc-multi}}_{t_1,t_2} \triangleq \left(\left(\boldsymbol{Z}_{t_1,t_2}\right)^T \boldsymbol{Z}_{t_1,t_2}\right)^{-1}\left(\boldsymbol{Z}_{t_1,t_2}\right)^T \tilde{\boldsymbol{Y}}_{t_1,t_2}. \tag{40}$$

In Equation 40, $\tilde{\boldsymbol{Y}}_{t_1,t_2}$ is a $(n_{t_1}n_{t_2} - |S_{t_1,t_2}|) \times 1$ vector of $\tilde{y}^i_{t_1}\tilde{y}^j_{t_2}$ values for non-overlapping individuals and $\boldsymbol{Z}_{t_1,t_2}$ is a $(n_{t_1}n_{t_2} - |S_{t_1,t_2}|) \times V$ matrix where column v is a vector of $G^{i,j}_{t_1,t_2;\,v}Q^{i,j}_{t_1,t_2}$ values for non-overlapping individuals.

We now describe how Equation 40 can be computed via summary statistics. We consider the terms $\left((\boldsymbol{Z}_{t_1,t_2})^T \boldsymbol{Z}_{t_1,t_2}\right)^{-1}$ and $(\boldsymbol{Z}_{t_1,t_2})^T \tilde{\boldsymbol{Y}}_{t_1,t_2}$ separately.

We begin with the term $(\boldsymbol{Z}_{t_1,t_2})^T \tilde{\boldsymbol{Y}}_{t_1,t_2}$. This term is a vector with V elements, each of which corresponds to the summation $\sum_{i,j \notin S_{t_1,t_2}} G^{i,j}_{t_1,t_2;\,v}Q^{i,j}_{t_1,t_2}\tilde{y}^i_{t_1}\tilde{y}^j_{t_2}$. Following the derivation in Equations 28 and 29, each such term can be computed via:

$$\sum_{i,j \notin S_{t_1,t_2}} G^{i,j}_{t_1,t_2;\,v}Q^{i,j}_{t_1,t_2}\tilde{y}^i_{t_1}\tilde{y}^j_{t_2} = \frac{1}{m^v}\sum_{k \in v} z^{k,\text{covar}}_{t_1} z^{k,\text{covar}}_{t_2} - \sum_{i,j \in S_{t_1,t_2}} G^{i,j}_{t_1,t_2;\,v}Q^{i,j}_{t_1,t_2}\tilde{y}^i_{t_1}\tilde{y}^j_{t_2}, \quad (41)$$

where m^v is the number of variants used to compute genetic similarity coefficient v and k iterates over all variants participating in this genetic similarity coefficient. As before, the second term on the right hand side can be computed by a party with access to the covariates of overlapping individuals, using the summary statistics w^i_t in Equation 27 or using the approximations described in Section 5.2.

The term $\left((\boldsymbol{Z}_{t_1,t_2})^T \boldsymbol{Z}_{t_1,t_2}\right)$ is a $V \times V$ matrix, wherein each entry $\left((\boldsymbol{Z}_{t_1,t_2})^T \boldsymbol{Z}_{t_1,t_2}\right)^{v,w}$ corresponds to the summation $\sum_{i,j \notin S_{t_1,t_2}} G^{i,j}_{t_1,t_2;\,v}G^{i,j}_{t_1,t_2;\,w}\left(Q^{i,j}_{t_1,t_2}\right)^2$. Following the derivation in Equations 31 and 32, this term can be computed via summary statistics as follows:

$$\sum_{i,j \notin S_{t_1,t_2}} G_{t_1,t_2;\,v}G_{t_1,t_2;\,w}\left(Q^{i,j}_{t_1,t_2}\right)^2 =$$
$$\frac{1}{m^v m^w}\sum_{k \in v, h \in w} \hat{r}^{k,h,\text{covar}}_{t_1}\hat{r}^{k,h,\text{covar}}_{t_2} - \sum_{i,j \in S_{t_1,t_2}} G_{t_1,t_2;\,v}G_{t_1,t_2;\,w}\left(Q^{i,j}_{t_1,t_2}\right)^2. \quad (42)$$

The second term on the right hand side can be computed by a party with access to the covariates of overlapping individuals, or using approximations similar to those described in Section 5.1.

8 Estimating the Liability Variance Due to Covariates

Heritability estimation requires dividing the genetic variance estimator of study t, $\left(\hat{\sigma}^2_g\right)_t$, by an estimate of the liability variance $\text{var}(a^i_t) = 1 + \text{var}\left((\mathbf{C}^i_t)^T\boldsymbol{\beta}_t\right)$. However, since the data we have is ascertained, we cannot directly estimate $\boldsymbol{\beta}$. Instead we use the non-parametric variance estimator proposed in [3]. Namely, we employ the law of total variance to decompose $\text{var}(a^i_t)$ as follows:

$$\text{var}(a^i_t) = E\left[\text{var}\left(a^i_t \,\middle|\, (\mathbf{C}^i_t)^T\boldsymbol{\beta}_t\right)\right] + \text{var}\left(E\left[a^i_t \,\middle|\, (\mathbf{C}^i_t)^T\boldsymbol{\beta}_t\right]\right). \quad (43)$$

The first term on the right hand side of Equation 43 is equal to one by definition, so our task is estimating the second term, which is equal to $\text{var}\left((\mathbf{C}^i_t)^T)\boldsymbol{\beta}_t\right)$ by definition. Since $\tau_{t_i} \triangleq \tau_t + (\mathbf{C}^i_t)^T\boldsymbol{\beta}_t$, we can instead estimate $\text{var}(\tau_{t_i})$. We employ the law of total variance again to obtain:

$$\text{var}(\tau_{t_i}) = E\left[\text{var}\left(\tau_{t_i} \,\middle|\, y^i\right)\right] + \text{var}\left(E\left[\tau_{t_i} \,\middle|\, y^i\right]\right). \quad (44)$$

Following the derivation in [3] we can estimate the right hand side of Equation 44 as follows:

$$\begin{aligned} E\left[\mathrm{var}\left(\tau_{t_i}\,\middle|\,y_t^i\right)\right] &= K_t\mathrm{var}(\tau_t^i \mid y_t^i = 1) + (1-K_t)\mathrm{var}(\tau_t^i \mid y_t^i = 0). \\ \mathrm{var}\left(E\left[\tau_{t_i}\,\middle|\,y_t^i\right]\right) &= K_t(1-K_t)\left(E\left[\tau_t^i\,\middle|\,y_t^i = 1\right] - E\left[\tau_t^i\,\middle|\,y_t^i = 0\right]\right)^2. \end{aligned} \tag{45}$$

The affection cutoffs τ_{t_i} are conditionally independent of the selection variables s_t^i given the phenotypes y_t^i. We can therefore estimate the expectations and variances in Equation 45 by their sample estimates.

Consequently, heritability estimation via summary statistics (without having access to genotype or phenotype data) requires the summary statistic $\mathrm{var}(\tau_{t_i})$.

9 Logistic Regression based Summary Statistics

Case control studies often report logistic regression rather than linear regression based Z-scores. We demonstrate here that although the PCGC estimator should ideally be computed with linear regression Z-scores, logistic regression Z-scores are approximately the same as linear regression Z-scores under large sample sizes. Thus, the use of logistic regression based summary statistics is expected to yield accurate estimates as well, as verified in our simulations. Our derivation here assumes that variants are single nucleotide polymorphisms (SNPs) and does not consider covariates. The use of logistic regression based summary statistics with covariates leads to inaccurate estimates, as demonstrated in the main text. We note that a somewhat similar treatment was provided in [7], but this treatment only concerned the estimated effect sizes, whereas here we are concerned with the Z-scores of the estimates.

Recall that linear regression Z-scores are given by:

$$Z_t^{k,\text{linear}} \triangleq \frac{1}{\sqrt{n_t}} \sum_i \tilde{y}_t^i X_t^{k,i}. \tag{46}$$

Logistic regression Z-scores are given by:

$$Z_t^{k,\text{logistic}} \triangleq \frac{\hat{\beta}^k}{\sqrt{\mathrm{Var}(\hat{\beta}^k)}}, \tag{47}$$

where $\hat{\beta}^k$ is the estimate of the logistic regression coefficient of SNP k. We show that under several mild assumptions $Z_t^{k,\text{logistic}} \approx Z_t^{k,\text{linear}}$ under large sample sizes.

Our derivation is carried out in two stages. First, we apply a Taylor expansion to show that $\hat{\beta}^k \approx \frac{Z_t^{k,\text{linear}}}{\sqrt{n_t P_t(1-P_t)}}$, where P_t is the case-control proportion in study t. Then, we approximate $Z_t^{k,\text{logistic}}$ via a Taylor expansion around $\hat{\beta}^k = 0$ and incorporate the estimate of $\hat{\beta}^k$ from the first stage to complete the derivation.

First stage: We now demonstrate that $\hat{\beta}^k \approx \frac{Z_t^{k,\text{linear}}}{\sqrt{n_t P_t(1-P_t)}}$ under large sample sizes. We consider a logistic regression model with the covariates vector $\mathbf{X}_t^k$ and an intercept. Unfortunately, logistic regression does not admit a closed form solution. To circumvent this difficulty, we will approximate the solution by using a profile log likelihood instead of the true log likelihood. Namely, we will first find the maximum likelihood estimate (MLE) of the intercept coefficient β_0^k under the assumption $\beta^k = 0$ and then express the MLE of β^k as a function of $\hat{\beta}_0^k$ and of $Z_t^{k,\text{linear}}$. This approximation is likely to be accurate if $\beta^k \approx 0$, which is likely to hold for polygenic traits.

Finding the MLE of β_0^k under the assumption $\beta^k = 0$ is easy. The log likelihood is:

$$\ell(\beta_0^k) = -n_t P_t \log(1 + \exp(-\beta_0^k)) - n_t(1 - P_t)\log(1 + \exp(\beta_0^k)), \tag{48}$$

and after differentiating it and setting the derivative to zero, we obtain:

$$\hat{\beta}_0^k = \log(P_t/(1 - P_t). \tag{49}$$

To proceed we assume that under large sample sizes, logistic regression and linear regression estimate the same conditional mean function. The approximation is accurate when the first order approximation of the logistic function as a function of X is close to the actual function, which is the case when $\beta^k \approx 0$ (more generally, the approximation holds when both the logistic and linear approximation of the true function estimate a very small coefficient for the tested variant). Formally, denote $E_{\text{linear}}\left[\tilde{y}_{t_i} \mid X_t^{k,i}\right]$ as the linear regression estimated conditional mean of $\tilde{y}_{t_i}$ given $X_t^{k,i}$, and $E_{\text{logistic}}\left[y_{t_i} \mid X_t^{k,i}\right]$ as the logistic regression estimated conditional mean of y_t^i given $X_t^{k,i}$. Then we assume that for every value of $X_t^{k,i}$:

$$\begin{aligned} E_{\text{linear}}[\tilde{y}_t^i \mid X_t^{k,i}] &= E_{\text{linear}}\left[\left.\frac{y_t^i - P_t}{\sqrt{P_t(1-P_t)}}\,\right|\, X_t^{k,i}\right] \\ &= \frac{E_{\text{linear}}[y_t^i \mid X_t^{k,i}] - P_t}{\sqrt{P_t(1-P_t)}} \\ &\approx \frac{E_{\text{logistic}}[y_t^i \mid X_t^{k,i}] - P_t}{\sqrt{P_t(1-P_t)}}. \end{aligned} \tag{50}$$

This assumption enables us to express $\hat{\beta}^k$ as a function of $Z_t^{k,\text{linear}}$.

We begin by simplifying the logistic regression estimate. According to the definition of logistic regression, we have:

$$E_{\text{logistic}}[y_t^i | X_t^{k,i}] = P(y_t^i = 1 | X_t^{k,i}) = \frac{1}{1 + \exp(-X_t^{k,i}\hat{\beta}^k - \hat{\beta}_0^k)}. \tag{51}$$

We approximate this quantity via a Taylor expansion around $\hat{\beta}^k = 0$ and then plug in the approximated value of $\hat{\beta}_0^k$ from Equation 49 to obtain:

$$\begin{aligned} E_{\text{logistic}}[y_t^i | X_t^{k,i}] &\approx \frac{1}{1 + \exp(-\beta_0^k)} + \frac{X_t^{k,i}\exp(-\beta_0^k)}{(1 + \exp(-\beta_0^k))^2}\hat{\beta}^k \\ &\approx \frac{1}{1 + (1 - P_t)/P_t} + \frac{X_t^{k,i}(1 - P_t)/P_t}{(1 + (1 - P_t)/P_t)^2}\hat{\beta}^k \\ &= P_t + P_t(1 - P_t)X_t^{k,i}\hat{\beta}^k. \end{aligned} \tag{52}$$

We now consider a linear regression model. Denoting $\hat{\gamma}^k$ as the coefficient estimate of $X_t^{k,i}$ in the regression of $\tilde{y}_t^i$ on $X_t^{k,i}$, we have:

$$\hat{\gamma}^k = \sqrt{n_t}\frac{Z_t^{k,\text{linear}}}{\sum_i \left(X_t^{k,i}\right)^2} \approx \frac{Z_t^{k,\text{linear}}}{\sqrt{n_t}}, \tag{53}$$

where we made the approximation $\sum_i \left(X_t^{k,i}\right)^2 \approx n_t$. Although this it not guaranteed to hold in case-control studies where the allele frequencies are different from the population frequencies, the approximation remains accurate under high levels of polygenicity where each SNP exerts a very small effect on the phenotype. Therefore, the linear regression estimate of the conditional mean is closely approximated as follows:

$$E_{\text{linear}}[\tilde{y}_t^i \mid X_t^{k,i}] = X_t^{k,i}\hat{\gamma}^k \approx X_t^{k,i}\frac{Z_t^{k,\text{linear}}}{\sqrt{n_t}}. \tag{54}$$

Finally, we combine Equations 52 and 54 into Equation 50 and rearrange to obtain:

$$\hat{\beta}^k \approx \frac{Z_t^{k,\text{linear}}}{\sqrt{n_t P_t(1-P_t)}}. \tag{55}$$

This completes the derivation.

Second stage: In the second stage we demonstrate that $Z_t^{k,\text{logistic}} \approx Z_t^{k,\text{linear}}$ by approximating $Z_t^{k,\text{logistic}}$ via a Taylor expansion around $\hat{\beta}^k = 0$. We begin by deriving $\text{Var}(\hat{\beta}^k)$. Define $\tilde{\mathbf{X}}_t^k$ as a matrix with two columns, where the first column contains ones and the second column contains $\mathbf{X}_t^k$. The observed information matrix is given by $\left(\tilde{\mathbf{X}}_t^k\right)^T \mathbf{D}\ \tilde{\mathbf{X}}_t^k$, where $\mathbf{D}$ is given by:

$$\mathbf{D} = \text{diag}\left(\left(1+\exp\left(-X_t^{k,i}\hat{\beta}^k - \hat{\beta}_0^k\right)^{-1}\right)\left(1+\exp\left(X_t^{k,i}\hat{\beta}^k + \hat{\beta}_0^k\right)^{-1}\right)\right). \tag{56}$$

The covariance matrix $\left(\left(\tilde{\mathbf{X}}_t^k\right)^T \mathbf{D}\ \tilde{\mathbf{X}}_t^k\right)^{-1}$ can be computed analytically using the formula for inversion of a 2 × 2 matrix. Namely, $\text{Var}(\hat{\beta}^k)$ is given by:

$$\text{Var}(\hat{\beta}^k) = \frac{\sum_i D^{i,i}}{\left(\sum_i D^{i,i}\right)\left(\sum_i D^{i,i}(X_t^{k,i})^2\right) - \left(\sum_i D^{i,i}X_t^{k,i}\right)^2}. \tag{57}$$

To write $\text{Var}(\hat{\beta}^k)$ in an analytic form, denote n_t^0, n_t^1 and n_t^2 as the number of individuals with 0, 1 and 2 minor alleles in SNP k, respectively. Further denote D^0, D^1 and D^2 as the values of the diagonal entries of $\mathbf{D}$ corresponding to individuals with 0, 1 and 2 minor alleles, respectively (note that the i_{th} diagonal entry in $\mathbf{D}$ depends only on the number of minor alleles carried by individual i). Finally, denote $X_{t,0}^k$, $X_{t,1}^k$ and $X_{t,2}^k$ as the values of the genotypes carried by individuals with 0, 1 and 2 minor alleles, respectively, after normalization. Using these notations, we can rewrite Equation 57 as follows:

$$\text{Var}(\hat{\beta}^k) = \frac{\sum_{a=0}^2 n_t^a D^a}{\left(\sum_{a=0}^2 n_t^a D^a\right)\left(\sum_{a=0}^2 n_t^a D^a\left(X_{t,a}^k\right)^2\right) - \left(\sum_{a=0}^2 n_t^a D^a X_{t,a}^k\right)^2}. \tag{58}$$

We can now incorporate Equation 58 into Equation 47 and then compute a first order Talylor expansion of $Z_t^{k,\text{logistic}}$ around $\hat{\beta}^k = 0$. After also using the approximation of β_0^k from Equation 49 and applying some algebra, the first order Taylor approximation takes the form:

$$\begin{aligned} Z_t^{k,\text{logistic}} &\approx \frac{P_t\sqrt{\left(\sum_{a,b=0,a\neq b}^2 n_t^a n_t^b\left(X_{t,a}^k\right)^2\right) - 2\left(\sum_{a,b=0,a<b}^2 n_t^a n_t^b X_{t,a}^k X_{t,b}^k\right)}}{\sqrt{n_t P_t/(1-P_t)}}\hat{\beta}^k. \\ &= \frac{P_t\sqrt{\sum_{a,b=0,a<b}^2 n_t^a n_t^b\left(X_{t,a}^k - X_{t,b}^k\right)^2}}{\sqrt{n_t P_t/(1-P_t)}}\hat{\beta}^k. \end{aligned} \tag{59}$$

Next, we use the fact that the genotypes were initially coded as 0,1,2 and then standardized by substracting twice the minor allele frequency p^k and dividing by $\sqrt{2p^k(1-p^k)}$. We therefore have:

$$\frac{1}{\sqrt{2p^k(1-p^k)}} = X^k_{t,2} - X^k_{t,1} = X^k_{t,1} - X^k_{t,0} = \frac{X^k_{t,2} - X^k_{t,0}}{2}. \tag{60}$$

Using this fact, Equation 59 can be rewritten as:

$$Z_t^{k,\text{logistic}} \approx \frac{P_t\sqrt{n_t^0 n_t^1 + n_t^1 n_t^2 + 4n_t^0 n_t^2}}{\sqrt{n_t P_t/(1-P_t)}\sqrt{2p^k(1-p^k)}}\hat{\beta}^k. \tag{61}$$

To proceed, we note that under large sample sizes and assuming Hardy-Weinberg equilibrium (HWE), we have $n_t^0 \approx n_t(1-p^k)^2$, $n_t^1 \approx 2n_t p^k(1-p^k)$, $n_t^2 \approx n_t(p^k)^2$. In practice, HWE and these approximations do not hold exactly in case-control studies, but the deviation is very small for highly polygenic traits where each SNP has a small effect. By incorporating these approximations into Equation 61, we obtain:

$$\begin{aligned} Z_t^{k,\text{logistic}} &\approx \frac{P_t n_t\sqrt{2p^k(1-p^k)^3 + 2(p^k)^3(1-p^k) + 4(p^k)^2(1-p^k)^2}}{\sqrt{n_t P_t/(1-P_t)}\sqrt{2p^k(1-p^k)}}\hat{\beta}^k. \\ &= \frac{P_t n_t\sqrt{(1-p^k)^2 + (p^k)^2 + 2p^k(1-p^k)}}{\sqrt{n_t P_t/(1-P_t)}}\hat{\beta}^k \\ &= \sqrt{n_t P_t(1-P_t)}\hat{\beta}^k. \end{aligned} \tag{62}$$

Finally, we combine Equations 55 and 62 to obtain:

$$Z_t^{k,\text{logistic}} \approx \sqrt{n_t P_t(1-P_t)}\frac{Z_t^{k,\text{linear}}}{\sqrt{n_t P_t(1-P_t)}} = Z^{k,\text{linear}}. \tag{63}$$

This completes the derivation.

10 Additional Summary Statistics

In [2] it is proposed to use LD score regression in case control studies by treating each SNP as a pair of binary variables, which enables using summary statistics of the form:

$$Z_{t,m}^{k,\text{binary}} \triangleq \frac{\sqrt{n_t P_t(1-P_t)}\left(\hat{p}_{t,m}^{k,\text{cas}} - \hat{p}_{t,m}^{k,\text{con}}\right)}{\sqrt{\hat{p}_{t,m}^k\left(1-\hat{p}_{t,m}^k\right)}}, \tag{64}$$

where $Z_{t,m}^{k,\text{binary}}$ is the summary statistics of the maternal allele of SNP k in study t, $\hat{p}_{t,m}^k$ is the in-sample maternal allele frequency of SNP k in study t, and $\hat{p}_{t,m}^{k,\text{cas}}$, $\hat{p}_{t,m}^{k,\text{con}}$ are its in-sample maternal allele frequency among cases and controls, respectively. The paternal allele summary statistic $Z_{t,p}^{k,\text{binary}}$ is defined analogously. [2] argue that using LD score regression with these summary statistics and a constrained intercept yields approximately the correct genetic correlation estimate. Here we reach the same conclusion, by showing that these summary statistics are approximately proportional to linear regression-based summary statistics of maternal and paternal alleles, which can provably be used to estimate genetic correlation owing to the relation to PCGC discussed in the main text.

We begin by establishing some notations. Denote $X_{t,m}^{k,i}$, $X_{t,p}^{k,i}$ as the maternal and paternal alleles of SNP k of individual i in study t, respectively, where $X_{t,m}^{k,i}, X_{t,p}^{k,i} \in \{0,1\}$. Further denote $X_t^{k,i} = X_{t,m}^{k,i} + X_{t,p}^{k,i}$ as the un-standardized value of SNP k of individual i in study t, and denote $\tilde{X}_{t,m}^{k,i} = \frac{X_{t,m}^{k,i} - p^k}{\sqrt{2p^k(1-p^k)}}$ as the standardized value of the maternal allele, and denote $\tilde{X}_{t,p}^{k,i} = \frac{X_{t,p}^{k,i} - p^k}{\sqrt{2p^k(1-p^k)}}$ analogously for the paternal allele, where p^k is the minor allele frequency of SNP k. Note that the notations here are slightly different from those used in the rest of the paper and the Supplemental material, where we defined $X_t^{k,i}$ as the standardized value of SNP k. This modification facilitates the notations in this section.

Using these notations, we define the maternal, paternal and standard linear regression summary statistics of the form:

$$\begin{aligned} Z_{t,m}^{k,\text{linear}} &\triangleq \frac{1}{\sqrt{n_t}} \sum_i \tilde{X}_{t,m}^{k,i} \tilde{y}_t^i \\ Z_{t,p}^{k,\text{linear}} &\triangleq \frac{1}{\sqrt{n_t}} \sum_i \tilde{X}_{t,p}^{k,i} \tilde{y}_t^i. \\ Z_t^{k,\text{linear}} &\triangleq \frac{1}{\sqrt{n_t}} \sum_i \left(\tilde{X}_{t,m}^{k,i} + \tilde{X}_{t,p}^{k,i} \right) \tilde{y}_t^i. \end{aligned} \tag{65}$$

In the large sample limit we have:

$$\sum_k Z_{t_1,m}^{k,\text{linear}} Z_{t_2,m}^{k,\text{linear}} + \sum_k Z_{t_1,p}^{k,\text{linear}} Z_{t_2,p}^{k,\text{linear}} \approx \frac{1}{2} \sum_k Z_{t_1}^{k,\text{linear}} Z_{t_2}^{k,\text{linear}}. \tag{66}$$

The approximation is obtained by applying the approximation that for each $r \in \{m, p\}$ the sum $\sum_k Z_{t_1,r}^{k,\text{linear}} Z_{t_2,r}^{k,\text{linear}}$ is the same in the large sample limit. This sum is used in the numerator of the PCGC-s covariance estimator described in the main text. We conclude that the sum of the maternal and paternal linear regression-based summary statistics can be used to estimate genetic correlation, since genetic correlation is a ratio and is therefore invariant to scaling by $\frac{1}{2}$.

We now show that in the large sample limit we have:

$$\begin{aligned} Z_{t,m}^{k,\text{binary}} &\approx 2\sqrt{2} P_t(1-P_t) Z_{t,m}^{k,\text{linear}} + \frac{\sqrt{n_t p^k P_t(1-P_t)}}{\sqrt{(1-p^k)}} (2P_t - 1) \\ Z_{t,p}^{k,\text{binary}} &\approx 2\sqrt{2} P_t(1-P_t) Z_{t,p}^{k,\text{linear}} + \frac{\sqrt{n_t p^k P_t(1-P_t)}}{\sqrt{(1-p^k)}} (2P_t - 1). \end{aligned} \tag{67}$$

These approximations demonstrate that $Z_{t,m}^{k,\text{binary}}$ is approximately proportional to $Z_{t,m}^{k,\text{linear}}$ when the case-control ratio P_t is close to $\frac{1}{2}$, and can therefore be used to estimate genetic correlation as well. All the derivations henceforth refer to the maternal allele but are equally applicable to the paternal allele.

We first rewrite Equation 65 as follows:

$$\begin{aligned}
Z_{t,m}^{k,\text{linear}} &= \frac{1}{\sqrt{n_t}} \sum_i \tilde{y}_{t_i} \frac{X_{t,m}^{k,i} - p^k}{\sqrt{2p^k(1-p^k)}} \\
&= \frac{1}{\sqrt{n_t}} \sum_i \tilde{y}_{t_i} \frac{X_{t,m}^{k,i}}{\sqrt{2p^k(1-p^k)}} \\
&= \frac{1}{\sqrt{n_t}} \sum_i \frac{y_t^i - P_t}{\sqrt{P_t(1-P_t)}} \frac{X_{t,m}^{k,i}}{\sqrt{2p^k(1-p^k)}} \\
&= \frac{\sum_i y_t^i X_{t,m}^{k,i} - P_t \sum_i X_{t,m}^{k,i}}{\sqrt{2n_t P_t(1-P_t)p^k(1-p^k)}} \\
&\approx \frac{\sum_i y_t^i X_{t,m}^{k,i} - n_t P_t p^k}{\sqrt{2n_t P_t(1-P_t)p^k(1-p^k)}}. \qquad (68)
\end{aligned}$$

Here, the first equality uses the definition of $\tilde{X}_t^{k,i}$ as a standardized SNP, the second equality uses the fact that $\sum_i \tilde{y}_{t_i} = 0$ by definition, the third equality uses the definition of $\tilde{y}_{t_i}$, the fourth equality is a straightforward expansion and the final approximation uses a large sample approximation.

Next, we rewrite Equation 64 as follows:

$$\begin{aligned}
Z_{t,m}^{k,\text{binary}} &\triangleq \frac{\sqrt{n_t P_t(1-P_t)}\left(\hat{p}_{t,m}^{k,\text{cas}} - \hat{p}_{t,m}^{k,\text{con}}\right)}{\sqrt{\hat{p}_{t,m}^{k}\left(1-\hat{p}_{t,m}^{k}\right)}} \\
&\approx \frac{\sqrt{n_t P_t(1-P_t)}}{\sqrt{p^k\left(1-p^k\right)}} \frac{\sum_i X_{t,m}^{k,i} y^i - \sum_i X_{t,m}^{k,i}(1-y^i)}{n_t} \\
&= \frac{\sqrt{P_t(1-P_t)}}{\sqrt{n_t p^k\left(1-p^k\right)}} \left(2\sum_i X_{t,m}^{k,i} y^i - \sum_i X_{t,m}^{k,i}\right) \\
&\approx \frac{\sqrt{P_t(1-P_t)}}{\sqrt{n_t p^k\left(1-p^k\right)}} \left(2\sum_i X_{t,m}^{k,i} y^i - n_t p^k\right). \qquad (69)
\end{aligned}$$

Here, both approximations are based on large sample assumptions and the assumption that the in-sample maternal allele frequency $\hat{p}_{t,m}^{k}$ is approximately the same as the population-level allele frequency p^k for highly polygenic traits.

Finally, we combine Equations 68 and 69 to obtain:

$$\begin{aligned}
Z_{t,m}^{k,\text{binary}} &\approx \frac{\sqrt{P_t(1-P_t)}}{\sqrt{n_t p^k\left(1-p^k\right)}} \left(2\left(Z_{t,m}^{k,\text{linear}}\sqrt{2n_t P_t(1-P_t)p^k(1-p^k)} + n_t P_t p^k\right) - n_t p^k\right) \\
&= 2\sqrt{2}P_t(1-P_t)Z_{t,m}^{k,\text{linear}} + \frac{\sqrt{n_t p^k P_t(1-P_t)}}{\sqrt{(1-p^k)}}(2P_t - 1). \qquad (70)
\end{aligned}$$

We conclude that $Z_{t,m}^{k,\text{binary}}$ is approximately proportional to $Z_{t,m}^{k,\text{linear}}$ if the sample case-control ratio P_t is close to 0.5. Therefore, genetic correlation estimates are likely to be accurate when using the summary statistics $Z_{t,m}^{k,\text{binary}}$, $Z_{t,p}^{k,\text{binary}}$.

11 The Effect of Ignoring Covariates

Here we prove the result reported in the main text, that under certain conditions, omitting measured covariates does not bias heritability or genetic correlation estimates. Specifically, the estimates remain unbiased if the covariate effects are normally distributed and are uncorrelated with the genetic effect.

The proof proceeds as follows. Recall that the liability for individual i in study t is given by $a_t^i = g_t^i + e_t^i + \left(\mathbf{C}_t^i\right)^T \boldsymbol{\beta}_t$, where $\mathbf{C}_t^i$ is a vector of covariates and $\mathrm{var}(g_t^i) + \mathrm{var}(e_t^i) = 1$. If we treat the term $\left(\mathbf{C}_t^i\right)^T \boldsymbol{\beta}_t$ as an unobserved random variable and assume it is uncorrelated with the genetic effect, then the liability variance is given by $\mathrm{var}\left(g_{t_i}\right) + \mathrm{var}\left({e_{t_i}}^\dagger\right)$, where ${e_{t_i}}^\dagger = e_t^i + \left(\mathbf{C}_t^i\right)^T \boldsymbol{\beta}_t$. If we additionally assume that ${e_{t_i}}^\dagger$ is normally distributed, then all the model assumptions hold except for the constraint $\mathrm{var}\left(g_t^i\right) + \mathrm{var}\left({e_{t_i}}^\dagger\right) = 1$ (the variables ${e_{t_i}}^\dagger$ are by definition independently and identically distributed).

Since the liability is unobserved, it is unidentifiable up to multiplication. We can therefore define a new model in which covariates are omitted, and estimate the model parameters in this new model. Denote g_t^{i*} and ${e_{t_i}}^*$ as the genetic and environmental effects in this new model, respectively. We proceed by making the assumption $\mathrm{var}\left(g_t^{i*}\right) + \mathrm{var}\left({e_{t_i}}^*\right) = 1$. Define $\sigma_{g\,t}^{2*} = \mathrm{var}\left(g_t^{i*}\right)$, $\sigma_{2t}^{e*} = \mathrm{var}\left(e_t^{i*}\right)$, $\rho_{t_1,t_2}^* = \mathrm{cov}\left({g_{t_1}^i}^*, {g_{t_2}^i}^*\right)$ as the genetic variance, the environmental variance and the genetic covariance in this new model, respectively. Then we have:

$$\begin{aligned}
\sigma_{g\,t}^{2*} &= \frac{\mathrm{var}\left(g_t^i\right)}{\mathrm{var}\left(g_t^i\right) + \mathrm{var}\left({e_{t_i}}^\dagger\right)} \\
\sigma_{2t}^{e*} &= \frac{\mathrm{var}\left({e_{t_i}}^\dagger\right)}{\mathrm{var}\left(g_t^i\right) + \mathrm{var}\left({e_{t_i}}^\dagger\right)} \\
\rho_{t_1,t_2}^* &= \frac{\rho_{t_1,t_2}}{\sqrt{\mathrm{var}\left(g_{t_1}^i\right) + \mathrm{var}\left({e_{t_1}^i}^\dagger\right)}\sqrt{\mathrm{var}\left(g_{t_2}^i\right) + \mathrm{var}\left({e_{t_2}^i}^\dagger\right)}}.
\end{aligned} \tag{71}$$

Consequently, heritability and genetic correlation in this new model are given by:

$$\begin{aligned}
{h^{2*}}_t &\triangleq \frac{\sigma_{g\,t}^{2*}}{1} = \frac{\mathrm{var}\left(g_t^i\right)}{\mathrm{var}\left(g_t^i\right) + \mathrm{var}\left({e_{t_i}}^\dagger\right)} = \frac{\mathrm{var}\left(g_t^i\right)}{\mathrm{var}\left(a_t^i\right)} = {h^2}_t \\
{r_{t_1,t_2}^g}^* &\triangleq \frac{\rho_{t_1,t_2}^*}{\sqrt{\sigma_{g_{t_1}}^{2*}\sigma_{g_{t_2}}^{2*}}} = \frac{\rho_{t_1,t_2}}{\sqrt{\mathrm{var}\left(g_{t_1}^i\right)\mathrm{var}\left(g_{t_2}^i\right)}} = r_{t_1,t_2}^g.
\end{aligned} \tag{72}$$

We conclude that when the assumptions of covariate normality and lack of correlation with genetic effects hold, heritability and genetic correlation in the omitted-covariates model is the same as in the covariates model. Therefore, estimates of these quantities are unbiased when these assumptions hold.

12 Real Data Analysis

We performed stringent quality control preprocessing to avoid genotyping artifacts from biasing the results. SNPs were excluded if they had minor allele frequency $< 5\%$, missingness rates $> 1\%$, a significantly different missingness rate between cases and controls, or a significant deviation from Hardy Weinberg equilibrium among the controls group. In the Wellcome Trust Case Control Consortium (WTCCC) analysis, controls consisted of individuals from the national blood service control group. Individuals were excluded from the analysis if they were in the WTCCC exclusion lists or if they had missingness rates $> 1\%$. We further excluded individuals with a standardized similarity coefficient > 0.05 with at least one other individual, by greedily removing individuals according to the number of related individuals they had, until no related individuals remained. In Table 1 and in Supplemental Tables S3 and S5 we additionally projected all genotype vectors to the subspace that is orthogonal to the top 10 principal components to prevent spurious results due to population structure. However, we caution that the analysis is sensitive to this procedure (see next section).

The analysis of each pair of WTCCC traits included approximately 1,950 cases, 1,450 shared controls and 275,000 SNPs that passed the quality filtering in both data sets. Following [1], the assumed prevalence for the traits was Crohn's disease (CD, 0.1%), type 1 diabetes (T1D; 0.5%), bipolar disorder (BD, 0.5%), rheumatoid arthritis (RA; 0.75%), type 2 diabetes (T2D; 3%), coronary artery disease (CAD; 3.5%) and hypertension (HT; 5%).

The schizophrenia data set included 1745 cases and 2586 controls. The bipolar disorder data set included 1268 cases and 3707 controls. 2566 controls were shared between the two data sets. After quality control, the analysis included 635,339 SNPs shared between the two data sets. These SNPs were taken from genotyped and imputed data provided by the Psychiatric Genomics Consortium (PGC), and filtered to ensure that no two SNPs had $r_2 > 0.9$. The MHC region was excluded from all analyses of these disorders. The assumed population prevalence for both disorders was 1%.

When analyzing the PGC datasets according to the LDAK model assumptions [6], we first computed SNP LD-weightings using LDAK [6], and then multiplied the LD-weighting of SNP j by $(p^j(1-p^j))^{0.75}$ (where p_j is the MAF of SNP j) to obtain its final weight. We then used these weights to compute a weighted genetic similarity coefficient between every pair of individuals. These weighted genetic similarity coefficients were used in the PCGC-s estimator, as explained in Section 6. LDAK estimated the weight of 244,640 SNPs as zero, which decreased the number of SNPs used in the estimation to 390,699. The LDAK commands we used to compute weightings were:

```
ldak5.linux --bfile <plink_file_name>
        --cut-weights <file_name>
ldak5.linux --bfile <plink_file_name>
        --calc-weights-all <file_name>
```

Every SNP k was standardized by subtracting $2p^k$ and dividing by $\sqrt{2p^k(1-p^k)}$, where p^k is its minor allele frequency. The minor allele frequencies were computed using Hapmap 3 data rather than from the data itself, To ensure that the summary statistics in both data sets use the same normalization.

Sex was used as a covariate in all analyses. The top 10 principal components were used as additional covariates in Table 1 and in Supplemental Tables S3 and S5. To use SNPs from the MHC region as covariates for T1D and RA, we ranked all SNPs in chromosome 6 between loci 25963966 and 34013250 (hg18) according to their correlation with the phenotype and then selected the 24 top SNPs (for T1D) and the top 31 SNPs (for RA) by maximizing the area under the receiving operating characteristic curve (AUC) via a five-fold cross validation using a logistic regression model. All SNPs in the MHC region were excluded from the genetic similarity matrix computations of T1D and RA.

LD scores were computed in-sample using the overlapping controls with a 0.1 centiMorgan window via the ldsc software[3] and were used by both PCGC-s-LD and LDSC. In Table 2, LDSC used a predetermined intercept and weighted all summary statistics by the inverse of the LD scores as recommended in [2], but a second weighting was not performed (see the discussion in the main text for further elaboration on these issues). The results with two rounds of weighting were similar (Table S5).

In all analyses, confidence intervals were computed using a block jackknife procedure with 200 blocks of SNPs, as in LDSC [2].

13 The Effects of Preprocessing the Data

The main text compares the performance of LDSC with omitted covariates to PCGC with included covariates. Under ideal conditions, LDSC with omitted covariates is almost equivalent to PCGC with omitted covariates, and this indeed was the case under the simulation studies. However, this equivalence can break down due to preprocessing of the data. Here we provide a short discussion of these issues. We consider LDSC invoked with weighting of the test statistics by the inverse of the LD score (but not by their posterior variance) and with a predetermined intercept. Recall that the PCGC-s and LDSC estimators in the absence of covariates are given by:

$$\hat{\rho}_{t_1,t_2}^{\text{pcgc-s}} = \frac{\frac{\sqrt{n_{t_1}n_{t_2}}}{m}\sum_{k=1}^{m} z_{t_1}^{k} z_{t_2}^{k} - \sum_{i,j\in S_{t_1,t_2}} \tilde{y}_{t_1}^{i}\tilde{y}_{t_2}^{j} G_{t_1,t_2}^{i,j}}{\frac{n_{t_1}n_{t_2}}{m^2}\sum_{k,h=1}^{m} \hat{r}_{t_1}^{k,h}\hat{r}_{t_2}^{k,h} - \sum_{i,j\in S_{t_1,t_2}} \left(G_{t_1,t_2}^{i,j}\right)^2}. \tag{73}$$

$$\hat{\rho}_{t_1,t_2}^{\text{ldsc}} = \frac{\frac{\sqrt{n_{t_1}n_{t_2}}}{m}\sum_{k=1}^{m} z_{t_1}^{k} z_{t_2}^{k} - \sum_{i,j\in S_{t_1,t_2}} \tilde{y}_{t_1}^{i}\tilde{y}_{t_2}^{j}}{\frac{n_{t_1}n_{t_2}}{m} E\left[\ell\right]}. \tag{74}$$

Here, S_{t_1,t_2} is the set of pairs of indices i,j that refer to the same individual in the two studies, and $E\left[\ell\right]$ is the mean LD score among all variants in the study.

If both studies used the exact same preprocessing prior to computing the test statistics, Equations 73 and 74 are almost equivalent: The second term of the numerator of Equation 74 (the so-called LDSC intercept) very closely approximates the second term of the numerator of Equation 73, because the genetic similarity coefficient of an individual with herself is typically very close to 1.0 when using a large number of variants. Furthermore, the denominator of Equation 74 is an unbiased estimator of the denominator of Equation 73 when the LD patterns in the two studies are the same as in a reference population from which LD scores was computed [5]. The two aforementioned assumptions are unlikely to hold exactly under case-control sampling, but our simulation studies indicate that under a highly polygenic trait model, the deviation from these assumptions is very small.

Unfortunately, the two assumptions can be violated when the two studies differ in the preprocessing of the data. Namely, the genotype distribution in the two studies can differ when using different SNP normalizations or when regressing principal components out of the genotypes.

[3] `https://github.com/bulik/ldsc`

We first consider the effect of different normalization of SNPs. Under ideal conditions, every SNP k is standardized by substracting $2p^k$ and dividing by $\sqrt{2p^k(1-p^k)}$, where p^k is its minor allele frequency. However, p^k is an unknown parameter that needs to be estimated from data. Many studies estimate this value from the sample and thus use a a study-specific normalization. This can lead to situations where the genetic similarity coefficient of an individual shared between two studies, given by $\sum_k X_{t_1}^{k,i} X_{t_2}^{k,j} \,/\, m$ (where i and j refer to the same individual) can deviate from 1.0. Although the deviation is typically small (less than 0.03 on average in the analysis of the WTCCC data), the LDSC estimator is very sensitive to such small deviations.

Regression of principal components can also affect the approximate equivalence between Equations 73 and 74. Ref. [5] argues that regression of principal components is likely to have a minimal effect on short-range LD patterns in the data, and thus advocates estimating LD using a limited window size. However, regression of PCs can lead to small deviations in the genetic similarity coefficient estimates of an individual with herself from 1.0, which can severely affect the LDSC intercept.

We note that although genetic similarity coefficient estimates of an individual with herself that deviate from 1.0 are biased estimators, treating these genetic similarity coefficients as if they were 1.0, without also correcting the first part of the numerator accordingly, can increase rather than decrease the bias of $\hat{\rho}_{t_1,t_2}^{\text{ldsc}}$.

We conclude that estimating variance components via summary statistics is sensitive to preprocessing. We therefore recommend that researchers provide summary statistics that can minimize the bias due to preprocessing, by enabling other researchers to replicate the preprocessing procedure. Namely, we recommend that researchers publish the in-sample LD information of their samples after normalization and regression of principal components. We further recommend that researchers normalize SNPs according to published minor allele frequencies based on a reference population rather than in-sample estimates. Finally, we recommend that researchers publish the estimated principal components so that researchers with access to overlapping individuals can replicate the preprocessing exactly for these overlapping individuals. We carried out these steps in the results reported in the manuscript.

14 Simulations

The simulation procedure consisted of first generating SNPs with LD patterns and then generating phenotypes based on these SNPs. Here we describe these steps.

The simulation of LD is an active research topic, but existing simulation require an elaborate model of population history [8], which is beyond the scope of our study. The use of real genotypes is hardly an option, because simulation of case-control studies requires first obtaining a population sample with millions of individuals and then down-sampling cases and controls. Here we used a simple model based on a Gaussian field with a single parameter controlling the degree of LD, similarly to other simulations of case-control studies [9]. Briefly, we first sampled a minor allele frequency for each SNP k in the range [0.05,0.5]. Afterwards, two independent vectors $\boldsymbol{v}_m$,$\boldsymbol{v}_p$ corresponding to maternal and paternal chromosomes were sampled for each individual from a zero mean multivariate normal distribution with a covariance matrix $\boldsymbol{R}$ obeying $R_{kh} = \theta^{|k-h|}$, where $\theta \in [0,1]$ is a tunable parameter. Finally, the maternal and paternal alleles of each SNP k were set to the minor allele if v_m^k and v_p^k exceeded the normal distribution percentile corresponding to their respective MAF. The normal vectors were generated without explicitly computing the matrix $\boldsymbol{R}$, using the well known result that a first order normal autoregressive process with autocorrelation parameter θ has the covariance matrix $\boldsymbol{R}$ [10].

Each SNP was first encoded as a vector of {0,1,2} (corresponding to the number of minor alleles) and then standardized to have a zero mean and unit variance in the population. The two effects of each SNP were sampled from $\mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \sigma^2_{g_{t_1}}/m & \rho_{t_1,t_2}/m \\ \rho_{t_1,t_2}/m & \sigma^2_{g_{t_2}}/m \end{pmatrix}\right)$, where m is the number of SNPs.

Binary covariates were generated as vectors of {0,1} and then standardized. Covariate effects were generated in several stages. First, the effect of each covariate j in study t, β_t^j, was sampled from $\mathcal{N}(0,1)$. Afterwards, the effect of the first covariate was multiplied by a parameter $w \geq 1$. Finally, all effects β_t^j were scaled by a constant to ensure that the contribution of the covariates to the liability variance, $\sum_j \left(\beta_t^j\right)^2$, yields the desired heritability level. This procedure enables tuning the normality of the aggregated covariates effect via the parameter w; Values of w close to 1 yield an approximately normal distribution of the aggregate effect.

The liability of every individual was computed as the weighted sum of the SNPs and the covariates multiplied by their effects, and an environmental term sampled from $\mathcal{N}\left(0, 1 - \sigma^2_{g_t}\right)$. The affection cutoff was determined empirically from the data as the $1-K$ percentile of the liabilities, where K was the simulated prevalence level. Individuals with liability greater than the affection cutoff were marked as cases.

In each experiment we first generated a population of size n_t/K_t (where n_t is the desired sample size and K_t is the trait prevalence) and then sampled the desired number of cases and controls from this population.

To generate simulations with MAF and LD-dependent SNP weightings, we made several modifications to the simulations algorithm. First, we divided the SNPs in to 10 different LD blocks, where the correlation between the Gaussian fields in each LD block b is a different number θ_b, with $\theta_b \in (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95)$. This was done in order for some SNPs to have significantly different LD scores (and thus different weights) than the others. Second, the SNP effects were sampled from a MAF and LD-dependent distribution, using the LDAK model [6]. Specifically, the effect of every SNP k, β_k, was generated from a zero-mean normal distribution with variance $(p_k(1-p_k))^{0.75} w_k/M$, where the weights w_k were selected to minimize the average L_2 norm of the quantity $1 - \sum_{k=1}^m \left(r^{kj}\right)^2 w_k$ across all SNPs j.

We computed the LDAK weights using the LDAK software [6]. Specifically, after creating an (unascertained) population of individuals, we created a plink file for a randomly selected subset of 10,000 individuals, and then computed the SNP weights by invoking LDAK with the options:

```
ldak5.linux --bfile <plink_file_name>
--no-thin YES --cut-weights <file_name>
ldak5.linux --bfile <plink_file_name>
--calc-weights-all <file_name>  --quick-weights YES.
```

We used the quick-weights option to expedite the simulation studies (which encompasses hundreds of different simulations). We note that our results demonstrate that PCGC can be adapted for different genetic architectures, and so the specific weight values are not of central importance for this demonstration.

Unless otherwise stated, all simulated datasets consisted of two studies of two traits with 1% prevalence, 50% heritability and 50% genetic correlation, with each study having 2,000 cases, 1,000 unique and 1,000 overlapping controls, 10,000 single nucleotide polymorphisms (SNPs) with a correlation of between 25% and 90% between adjacent SNPs. In most simulations all SNPs influenced the phenotype, though we verified that relaxing this assumption does not affect the results (Figure S10). 100 simulations were conducted for each unique combination of settings.

15 Use of Alternative Methods

Here we describe how LDSC and REML were used in the results section.

REML estimates were computed via GCTA [11]. Specifically, heritability was estimated via the following two commands:

```
gcta64 --bfile <file_name> --make-grm --out <file_name>
gcta64 --grm <file_name> --reml-bivar --pheno <phenotypes_file>
        --reml-bivar-prevalence <trait 1 prevalence> <trait 2 prevalence>
        --qcovar <covariates file>
```

LDSC estimates in the real data analysis were computed via the command:

```
python ldsc.py --w-ld <file_name> --ref-ld <file_name>
        --rg <sumstats1>,<sumstats1>
        --samp_prev <sample_prevalence1>,<sample_prevalence2>
        --pop-prev <prevalence1>,<prevalence2> --M <#variants>
```

When using a predetermined intercept, we also added the arguments:

```
intercept-h2 1,1 --intercept-gencov 0, <intercept>
```

where the provided intercept was computed as described in [2].

References

[1] D. Golan and S. Rosset. Effective genetic-risk prediction using mixed models. *Am. J. Hum. Genet.* 95(4) (2014), 383–93.

[2] B. Bulik-Sullivan et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* 47(11) (2015), 1236–41.

[3] D. Golan, E.S. Lander, and S. Rosset. Measuring missing heritability: Inferring the contribution of common variants. *Proc. Natl. Acad. Sci. USA* 111(49) (2014), E5272–81.

[4] C.E. McCulloch, S.R. Searle, and J.M. Neuhaus. *Generalized, Linear, and Mixed Models.* 2nd. Wiley Series in Probability and Statistics, 2008.

[5] B. Bulik-Sullivan. Relationship between LD Score and Haseman-Elston Regression. *bioRxiv* (2015), 018283.

[6] D. Speed et al. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* 49(7) (2017), 986.

[7] M. Pirinen, P. Donnelly, C.C. Spencer, et al. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann. Appl. Stat.* 7(1) (2013), 369–390.

[8] S. Hoban, G. Bertorelle, and O.E. Gaggiotti. Computer simulations: tools for population and evolutionary genetics. *Nat. Rev. Genet.* 13(2) (2011), 110–22.

[9] B.K. Bulik-Sullivan et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47(3) (2015), 291–5.

[10] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning.* The MIT Press, 2006.

[11] J. Yang et al. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88(1) (2011), 76–82.