Genetic correlations of polygenic disease traits: from theory to practice

Wouter van Rheenen[1], Wouter J. Peyrot[2], Andrew J. Schork[3], S. Hong Lee[4] and Naomi R. Wray[5,6]**

Abstract | The genetic correlation describes the genetic relationship between two traits and can contribute to a better understanding of the shared biological pathways and/or the causality relationships between them. The rarity of large family cohorts with recorded instances of two traits, particularly disease traits, has made it difficult to estimate genetic correlations using traditional epidemiological approaches. However, advances in genomic methodologies, such as genome-wide association studies, and widespread sharing of data now allow genetic correlations to be estimated for virtually any trait pair. Here, we review the definition, estimation, interpretation and uses of genetic correlations, with a focus on applications to human disease.

Parameter
A numerical value that summarizes a characteristic of a population, such as the mean height of men, the lifetime risk of schizophrenia or the heritability of a specific trait.

[1]*Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, Netherlands.*

[2]*Department of Psychiatry, Amsterdam UMC, VU University Medical Center, Amsterdam, Netherlands.*

[3]*Institute for Biological Psychiatry, Mental Health Services Snct. Hans, Roskilde, Denmark.*

[4]*Australian Centre for Precision Health, University of South Australia Cancer Research Institute, University of South Australia, Adelaide, South Australia, Australia.*

[5]*Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia.*

[6]*Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia.*

**e-mail: w.vanrheenen-2@umcutrecht.nl; naomi.wray@uq.edu.au*

https://doi.org/10.1038/s41576-019-0137-z

The genetic correlation is a quantitative genetic parameter that describes the genetic relationship between two traits and has been expected to reflect pleiotropic action of genes or correlation between causal loci in two traits. Studies of genetically correlated traits improve our understanding of complex traits because they can reveal genetic variation that contributes to disease, improve genetic prediction and inform therapeutic interventions.

In humans, most lifestyle risk factors of disease, as well as the diseases themselves, are at least partially heritable[1]; genetic correlation estimates help to describe their complex relationships, which, particularly in the context of disease traits, may be unrecognized. For example, although genetic correlations had been hypothesized among psychiatric diseases[2], they long remained difficult to measure using traditional genetic epidemiological approaches, which require data from many families with two or more blood relatives recorded for each trait. Given that most diseases defined as common have lifetime risks of 0.5–5%, collating data sets that are informative for two diseases is difficult and subject to ascertainment biases. The Scandinavian registries[3], which comprise data on diagnosis codes from national hospital admissions and discharges for up to several million individuals, have been useful for calculating estimates of increased risk of a given disease in relatives of those with a different specified disease[4], which is the key observation for estimating genetic correlation between diseases from family data. However, these registries also have limitations for estimating genetic correlation between diseases; the data set is restricted to the size of the national population, and recording began quite recently for many disease traits[3], resulting in incomplete or censored observations for late-onset disease. Moreover, it remains challenging to disentangle genetic sharing from sharing of a common family environment, which in traditional epidemiology can only be separated by collecting large data sets of families that include different types of relatives (such as full siblings, cousins and parents–offspring) measured for both diseases.

Genome-wide association studies (GWAS) have provided a new paradigm for estimating genetic correlations among disease traits from data sets that have been independently collected for two diseases. Individual-trait data sets have much larger sample sizes and, because the data for the two traits are independently collected, the opportunity of confounding through shared common environmental factors is minimized. These new data, combined with new statistical methods to estimate genetic correlations from both individual-level genetic data or GWAS summary statistics and widespread data sharing, have greatly increased the potential to describe relationships between complex diseases and traits.

Here, we review the definition, estimation, interpretation and uses of genetic correlations for human complex traits, with a focus on disease. We discuss how genetic correlations between measured traits can be used to improve the power of association tests for identifying genetic variants contributing to disease risk, can improve genetic prediction and can aid inferences about causality to inform intervention strategies. We describe the interpretation of genetic correlation estimates and consider scenarios that can lead to misinterpretation. Supporting theory, together with simulations and code, are provided in the Supplementary note.

Defining genetic correlations

Genetic correlation measures pleiotropy. Pleiotropy is present when a genetic locus affects more than one trait. The term pleiotropy was introduced before the molecular characterization of DNA[5] and hence is attributed

to a genetic locus, which can imply pleiotropy at the level of either a DNA variant or a gene. In the context of a discussion about genetic correlation, our interest is in DNA variants. Pleiotropy between two traits can reflect different modes of action[5,6] (FIG. 1). The main distinction is whether the two phenotypes are part of a causal cascade (vertical or mediated pleiotropy) or not (horizontal or biological pleiotropy). Understanding horizontal pleiotropy may lead to a better understanding of biological processes that are common between traits. Vertical pleiotropy can inform on causality for intervention strategies for disease prevention. It is worth noting that in earlier studies[7,8] that laid the foundation for understanding pleiotropy, vertical pleiotropy was termed 'spurious pleiotropy', a term that we now use for spurious genetic correlation estimates due to bias, misclassification or linkage disequilibrium (FIG. 1). Methods to discriminate between vertical and horizontal pleiotropy indicate that often a combination of both contribute to the genetic correlation[9,10]. Genetic correlation describes the average effect of pleiotropy across all causal loci, but the underlying architecture of correlations at individual loci can vary (FIG. 2). Local genetic correlation can deviate from the genome-wide average and regions with strong positive or negative genetic correlation have been described for multiple traits even in the absence of genome-wide genetic correlation[11].

Defining genetic correlation mathematically. In a general quantitative genetic model, in which, for each individual, two traits (x and y) are each defined as the sum of a genetic value (g) and a residual value (e, with residual simply meaning the difference between the trait value and the genetic value):

$$x = g_x + e_x \quad (1)$$

$$y = g_y + e_y, \quad (2)$$

the genetic correlation (ρ_g) of the traits is:

$$\rho_g = \frac{\sigma_{g_x,g_y}}{\sqrt{\sigma^2_{g_x}\sigma^2_{g_y}}} \quad (3)$$

where σ_{g_x,g_y} is the covariance of the genetic values and $\sigma^2_{g_x}, \sigma^2_{g_y}$ is the genetic variance of the two traits in the population. As a result, ρ_g ranges from –1 to 1. Following convention, the Greek letters emphasize parameters (ρ_g), which are replaced by Roman letters for estimates (r_g), although we note that h^2 is commonly used to represent both the parameter and the estimate of heritability. Because the definition of genetic correlation depends on invoking a latent model, it is important to acknowledge that any estimates of genetic parameters may be biased if the assumed latent model is an imperfect representation of nature. Lack of clarity in distinguishing the conceptual parameters from the estimates made from the available data is a common problem[12].

If the traits are standardized (that is, phenotypic variance = 1) and the genetic values consider only the additive genetic effects, then the genetic variances are narrow-sense heritabilities and the numerator is the

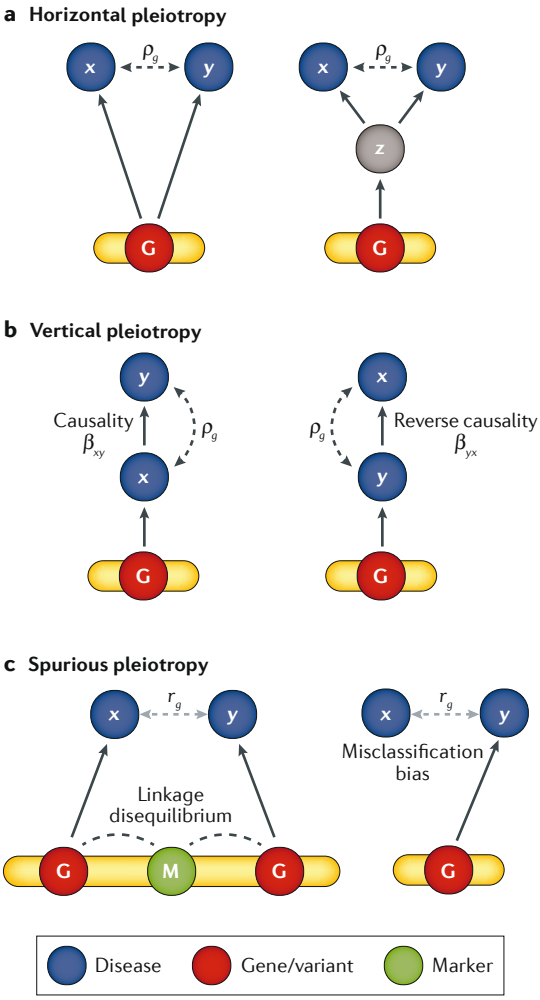


Fig. 1 | **Different mechanisms of pleiotropy between two diseases. a** | In horizontal pleiotropy, the genetic variant (G) contributes directly to risk of both diseases (*x* and *y*) or indirectly through an intermediate (endo)phenotype (z). **b** | In vertical pleiotropy, there is a causal relationship between disease *x* and *y* where disease *x* itself leads to an increased risk of disease *y* (left); in some circumstances, reverse causation may be observed (here we assume that there is a natural order in the traits to expect causation from trait *x* to trait *y*, with the reverse causation from *y* to *x* less expected) (right). In some circumstances, there may be causality, reverse causality and genetic correlation. **c** | Spurious pleiotropy at a locus can occur when a measured genetic marker (M) 'tags', via linkage disequilibrium, two distinct causal variants (left); however, this sort of spurious pleiotropy has to be consistent across loci to result in a non-zero estimate of genome-wide genetic correlation. Different sources of bias (such as disease misclassification) may also lead to the incorrect assumption of pleiotropy between disease *x* and *y* (right). In early work on pleiotropy, the term 'spurious pleiotropy' was used for what we term 'vertical pleiotropy'. β_{xy} is the expected change in trait y caused by each unit increase in trait x; β_{yx} is the expected change in trait x caused by each unit increase in trait y. ρ_g, genetic correlation; r_g, estimated genetic correlation.

Traits
Measurements or phenotypes that are usually studied as the outcome of statistical analyses. They can be quantitative (for example, height) or dichotomous (for example, schizophrenia).

Estimates
Approximations of a parameter based on a sample of observed data drawn from a population.

Ascertainment biases
Types of bias that occur when the studied trait or disease affects how data were ascertained. For example, patients with a family history of diabetes may have more frequent examinations for cardiovascular diseases.

Genome-wide association studies
Studies in which up to millions of mostly common single-nucleotide polymorphisms from across the genome are each tested for association with a trait.

GWAS summary statistics
The output of statistical tests of association of a trait with each single-nucleotide polymorphism generated by a genome-wide association study (GWAS), typically including the effect allele, signed effect estimate, standard error, test statistic (for example, a *z*-score) and/or *p*-value.

Power
The probability that a study correctly rejects the null hypothesis of no association or correlation, also described as 1 – type II error.

Bias
Phenomenon where statistical analyses produce estimates in observed data that systematically overestimate or underestimate the population parameter. Bias can arise from the ascertainment of the observed data or the statistical procedures used to generate the estimates.

Linkage disequilibrium
(LD). The non-random segregation of alleles at two distinct loci. LD induces a correlation between two single-nucleotide polymorphism (SNP) genotypes in the population and is caused by the fact that alleles of neighbouring SNPs are transmitted together until broken down by recombination events.

Genetic value
(g). The sum of the total effects of all genetic loci on the trait in an individual, that is $g = X\beta$ where X is a vector of genotypes for all loci and β is a vector with additive allelic effects on the trait. It is also called the genotypic value, true polygenic (risk) score or breeding value.

Covariance
($\sigma_{x,y}$). The expected product of the deviation of two random variables from their mean ($\sigma_{x,y} = E[(X-\mu_x)(Y-\mu_y)]$).

Genetic variance
(σ_G^2). The expected squared deviation of genetic values from the mean genetic value ($\sigma_G^2 = E[(G-\mu_g)^2]$), and can also be considered the covariance of a genetic value with itself.

Heritability
(h^2). The proportion of phenotypic variance (parameter σ_P^2, estimate V_P) attributable to variance in genetic factors. In the context of human traits, most often only additive genetic factors are considered for the genetic variance (parameter σ_A^2, estimate V_A) and the ratio of variances is the narrow-sense heritability.

Latent model
A collection of formalized assumptions to describe a data-generating process through which observed variables (such as disease occurrence) can be used to identify unobserved (latent) variables (for example, genetic parameters: heritability and genetic correlation).

Phenotypic variance
(σ_P^2). Variance of phenotypic values (for example, height or disease liability) after accounting for the variance attributable to fixed effects (for example, sex). When phenotypes are standardized, these phenotypic values are scaled such that $\mu_P = 0$ and $\sigma_P^2 = 1$.

Coheritability
(h_{xy}). The genetic covariance of standardized traits. This is a useful measure for comparisons of coheritabilities and heritabilities on the same scale.

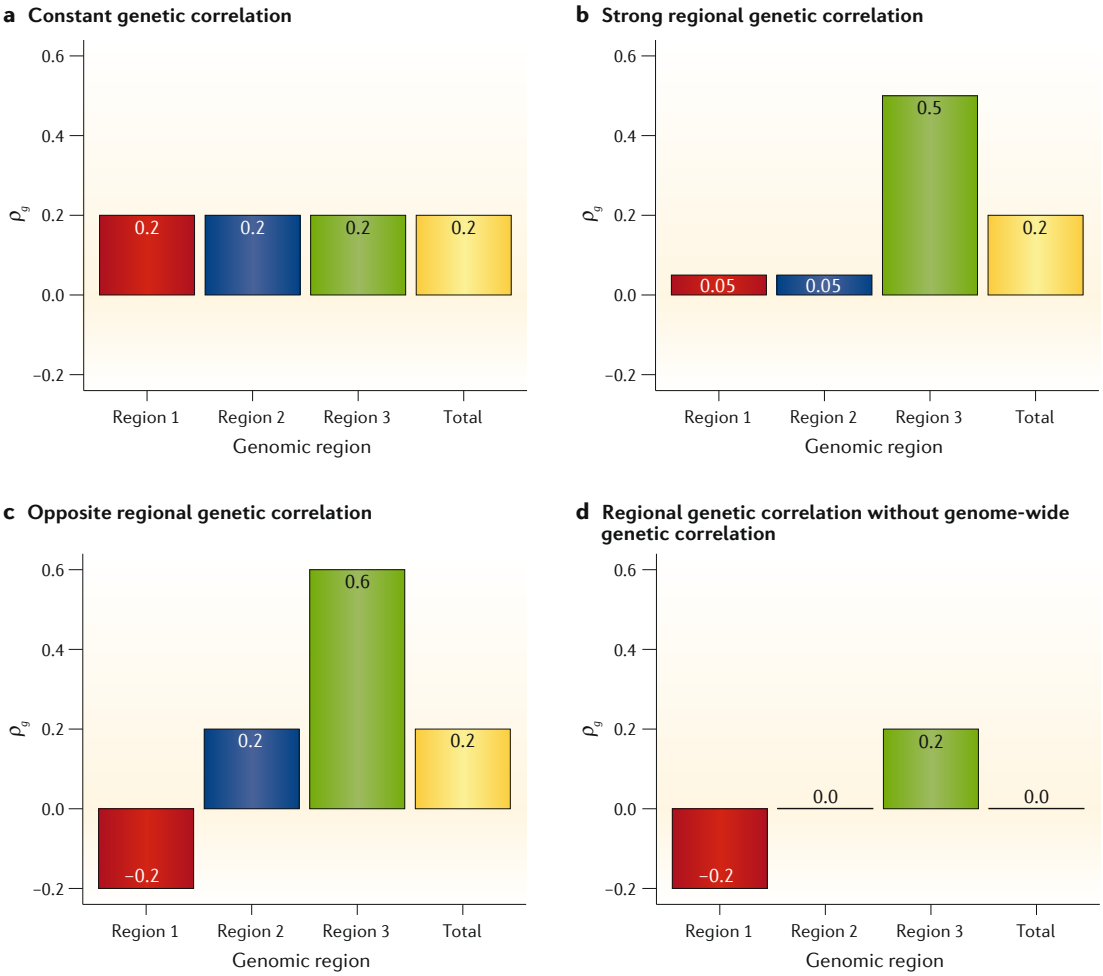


Fig. 2 | **Genome-wide genetic correlation versus regional genetic correlation.** The overall genetic correlation (as estimated from pedigree or genome-wide association study data) between two traits has been fixed at 0.2, but the underlying regional architecture of the genetic correlation can vary widely[11]. In part **a**, the genetic correlation is constant across the genome. Alternative scenarios include strong regional genetic correlation (part **b**) and a combination of both positive and negative regional correlations (part **c**); in both of these cases, the regional genetic correlation can far exceed the overall genome-wide genetic correlation. In part **d**, both positive and negative regional correlations occur in the absence of an overall genetic correlation. Regions can be interpreted as physical genomic loci, as allele-frequency bins or as functionally annotated categories (such as coding versus non-coding, biological pathways or tissue-specific expression). ρ_g, genetic correlation.

covariance of the standardized traits or the coheritability (h_{xy}). Equation 3 can therefore be rewritten as:

$$\rho_g = \frac{h_{xy}}{\sqrt{h_x^2 h_y^2}} \quad (4)$$

As the genetic covariance is scaled relative to the two genetic standard deviations when computing ρ_g, a high genetic correlation is possible even if there is only a small genetic contribution to the two traits. Hence, reporting estimates for both r_g and h^2 can provide a reference for the importance of the shared genetic contribution to the trait phenotypes by enabling coheritability to be estimated and benchmarked against the heritabilities.

Although the relationship between phenotypic correlation (ρ_p) and genetic correlation additionally includes the correlation between residual factors (ρ_e):

$$\rho_p = \sqrt{h_x^2 h_y^2}\,\rho_g + \sqrt{\left(1-h_x^2\right)\left(1-h_y^2\right)}\,\rho_e\,, \quad (5)$$

Cheverud's conjecture[13] proposes that estimates of ρ_p can be used to approximate estimates of ρ_g. The conjecture relies on the assumption that most environmental effects act in the same direction and through the same pathways as genetic effects, which leads to a similarity between phenotypic and genetic correlations. Importantly, unlike estimation of genetic correlation, the phenotypic correlation can be estimated from cohorts of unrelated individuals that have been measured for both traits. Although it is relatively easy to collect data to estimate phenotypic correlations between quantitative traits, it remains challenging to estimate phenotypic correlations for disease traits because of limited data availability, potential ascertainment biases and symptom overlap[14].

Linear mixed model
(LMM). A linear model that includes both fixed and random effects to describe phenotypic values and that allows a correlation structure between the random effect levels.

Restricted maximum likelihood
(REML). A method for maximum likelihood estimation of variance–covariance components of the parameters in linear mixed models.

Liability threshold model
A model that describes a dichotomous trait (disease) as a threshold partitioning of 'liability', which is a latent variable assumed to follow a standard normal distribution in the population. The liability threshold (T) defines lifetime risk (K) of disease as the proportion of individuals exceeding this threshold.

Risk ratio
Ratio between the risk of disease in a specific group (for example, relatives of affected individuals) and the risk of disease in the general population.

Tetrachoric correlation
The correlation between two latent normally distributed liability phenotypes assumed to underlie dichotomous population data and estimated from an observed 2 × 2 frequency table.

Genomic relationship matrix
(GRM). A matrix whose off-diagonal elements represent a coefficient of genetic sharing between individuals to describe the variance–covariance structure between their genetic values calculated from observed single-nucleotide polymorphism (SNP) data. GRM coefficients can be calculated based on different assumptions of the expected distribution of per-SNP heritability.

Independently-collected GWAS data sets for a range of important traits are now widely available and offer an alternative to family studies for estimating genetic correlations attributable to common genetic variants. However, the expectation that genetic correlation estimates from family phenotypic records are the same as those from GWAS data assumes that ρ_g is homogeneous across the allelic frequency spectrum of risk loci[15].

Methods to estimate genetic correlations

Methods to estimate genetic correlations depend on the data sets available, such as large cohorts of related individuals or GWAS. Here, we describe the methods that laid the foundation for studies of genetic correlation, including methods to study the distribution of genetic correlation across the genome (genome partitioning) and methods to study genetic correlations of the same trait in different environments. TABLE 1 presents a list of available methods and software.

Genetic epidemiological data for related individuals. A bivariate linear mixed model (LMM) can be used to estimate heritabilities and genetic correlations from large cohorts of families measured for two traits. In a bivariate LMM, each phenotype is modelled as a function of the latent genetic values of individuals, and they are assumed to be drawn from a bivariate normal (that is, polygenic) distribution, where in this case the variance–covariance structure of the genetic relationships is based on pedigree data. Best estimates of genetic and phenotypic variances and covariances can be obtained using restricted maximum likelihood (REML)[16]. For disease traits, a simpler approach is to estimate genetic correlation estimates using population disease risk and risks in pairs of related individuals (such as full siblings or parent–offspring), assuming, for example, a liability threshold model. Under this model, the heritabilities and genetic correlations can be estimated using normal distribution theory[17,18]. Although the equations look complex, they depend on only five measures: lifetime risk of disease x and y (K_x and K_y), lifetime risk of disease y or x in relatives of those with disease x or y ($K_{Ry,x}$, $K_{Rx,x}$, $K_{Ry,y}$) and the coefficient of relationship (α_R), which is 0.5 for full siblings or parent–offspring:

$$h_x^2 = \frac{T_x - T_{Rx,x}\sqrt{1-\left(1-T_x/i_x\right)\left(T_x^2 - T_{Rx,x}^2\right)}}{a_R\left[i_x + \left(i_x - T_x\right)T_{Rx,x}^2\right]} \quad (6)$$

$$r_g = \frac{T_y - T_{Ry,x}\sqrt{1-\left(1-T_x/i_x\right)\left(T_y^2 - T_{Ry,x}^2\right)}}{a_R\left[i_x + \left(i_x - T_x\right)T_{Ry,x}^2\right]h_x h_y} \quad (7)$$

Here, T_x and $T_{Ry,x}$ are the normal distribution thresholds that reflect the proportions K_x and $K_{Ry,x}$; and i_x is the mean phenotypic liability of those with disease x, which is calculated as z_x/K_x, where z_x is the height of the standard normal curve at T_x. The relationship between the increased risks to relatives ($K_{Ry,x}/K_y$; for example, cross-disorder risk ratio) and genetic correlation depends on the coefficient of kinship, heritability for both traits and disease prevalence (FIG. 3). Very similar estimates of h^2 and r_g are obtained for schizophrenia and bipolar disorder using a LMM approach on data from the large Swedish registry (h^2 of 0.64 (95% CI 0.62–0.68) and 0.59 (95% CI 0.56–0.62), respectively, and $r_g = 0.60$ (no 95% CI reported))[4] or meta-analysis of the estimates derived from the simple liability threshold equations (h^2 of 0.64 (95% CI 0.61–0.67) and 0.56 (95% CI 0.54–0.58), respectively, and $r_g = 0.47$ (95% CI 0.32–0.62))[19] described in equations 6 and 7, demonstrating that the simple approach is a good approximation of the complex analysis. The LMM approach also estimated a contribution for environment shared between relatives (c^2), which was 0.045 (95% CI 0.044–0.074) for schizophrenia and 0.034 (95% CI 0.023–0.062) for bipolar disorder.

Alternatively, the tetrachoric correlation (r_{tc}) of Pearson[20] can be used to estimate heritability[1] and genetic correlation from the 2 × 2 table of observations of disease[1,21,22] in related individuals:

$$h_x^2 = \frac{1}{a_R} r_{tc,x} \quad (8)$$

$$r_g = \frac{r_{tc,x,y}}{a_R\sqrt{h_x^2 h_y^2}} \quad (9)$$

Here, the main assumption is that health and disease are a result of dichotomizing an underlying bivariate normal distribution, which is consistent with the liability threshold model but requires that the proportion of cases in the study equals the population risk. Therefore, both methods yield similar estimates when applied to such data (Supplementary note).

Individual-level GWAS data for unrelated individuals. Estimation of the genetic correlation from individual-level GWAS data involves a bivariate extension[23] of the univariate genome-based REML (GREML) that uses a genomic relationship matrix (GRM) to estimate single-nucleotide polymorphism-based heritability (SNP-based heritability)[24,25]. As for traditional epidemiology data, this approach uses an LMM, in which the phenotype is modelled as a function of the genetic values of individuals, but the variance–covariance structure of genetic values is described by genetic relationships in the GRM constructed from observed genome-wide SNP data rather than from pedigree data. SNP-based heritability is expected to be lower than heritability estimated from epidemiological family records because it aims to capture only causal variation that is in linkage disequilibrium (LD) with the measured SNPs. Therefore, it provides insight into the relative importance of common SNP variation, which can differ among traits. Relatives closer than second or third cousins are excluded from the analysis to ensure that short haplotype segments are tracked by the shared genetic relationships between pairs of individuals. Compared to close relatives, distant relatives share negligible non-additive genetic variation[26] and are expected to have lower phenotypic correlation

Table 1 | **Summary of methods and software packages**

Name	Description	Input	Source	Refs
Estimate genetic correlation				
polycor	Estimate tetrachoric correlation through MLE	2 × 2 contingency tables from pedigrees	https://cran.r-project.org/web/packages/polycor/	22
GCTA (—reml-bivar)	Bivariate GREML; includes options for different model assumptions	Individual-level genotypes	http://cnsgenomics.com/software/gcta/	23,46
MTG2	Computationally efficient bivariate GREML	Individual-level genotypes	https://sites.google.com/site/honglee0707/mtg2	114
BOLT-REML	Computationally efficient approximate bivariate GREML with GRM with fully overlapping individuals for both traits	Individual-level genotypes	https://data.broadinstitute.org/alkesgroup/BOLT-LMM/	115
LDSC (—rg)	Weighted regression of product of GWAS summary statistics on LD scores	GWAS summary statistics	https://github.com/bulik/ldsc	28,29
LDAK	Calculate weighted kinship matrix to model distribution of causal variants across LD and/or MAF spectrum	Individual-level genotypes	http://dougspeed.com/ldak	41,43
SumHer (—sum-cors)	Analogue to LDSC, but adopts 'LDAK model'	GWAS summary statistics	http://dougspeed.com/sumher/	44
GCTA (—HEreg-bivar)	Bivariate Haseman–Elston regression	Individual-level genotypes	http://cnsgenomics.com/software/gcta/	46,51
S-PCGC	Phenotype correlation genotype correlation regression. Extension of Haseman–Elston regression, robust to ascertainment and covariates	GWAS summary statistics	• https://github.com/omerwe/S-PCGC • https://data.broadinstitute.org/alkesgroup/PCGC/	49,50,116
Popcorn	Bayesian estimate of transethnic genetic correlation	GWAS summary statistics	https://github.com/brielin/Popcorn	37
LD Hub	Server to estimate genetic correlation from published GWAS using LDSC	GWAS summary statistics	http://ldsc.broadinstitute.org/ldhub/	30
GNOVA	Annotation-partitioned genetic correlation	GWAS summary statistics	https://github.com/xtonyjiang/GNOVA	35
ρHESS	Local genetic correlation	GWAS summary statistics	https://github.com/huwenboshi/hess	11
Power calculation genetic correlation estimation				
GCTA power calculator	Calculate the power of bivariate GREML analysis in GCTA	User-defined parameters	https://cnsgenomics.shinyapps.io/gctaPower/	34
Multitrait association analysis				
MultiMeta	Inverse variance-weighted meta-analysis	SNP effects and standard errors	https://CRAN.R-project.org/package=MultiMeta	61
ASSET	Subset meta-analysis which can include correlated traits	SNP effects and standard errors	https://bioconductor.org/packages/release/bioc/html/ASSET.html	62
HIPO	Heritability-based weighting of correlated traits	SNP effects and standard errors	https://github.com/gqi/hipo	63
MTAG	Pleiotropy-informed SNP association analysis for single trait	SNP effects and standard errors	https://github.com/omeed-maghzian/mtag	64
CPMA	Test SNP pleiotropy through distribution of *p*-values	SNP *p*-values	http://coruscant.itmat.upenn.edu/software.html	117
LEP	Heterogeneous sharing of risk variants	SNP effects and standard errors	https://github.com/daviddaigithub/LEP	118
metaUSAT	Score-based association test	SNP effects and standard errors	https://github.com/RayDebashree/metaUSAT	65
TATES	Multivariate analysis of single SNPs	SNP *p*-values	https://ctg.cncr.nl/software/tates	73
metaCCA	Multivariate analysis of multiple SNPs using canonical correlation analysis	SNP effects and standard errors	https://bioconductor.org/packages/release/bioc/html/metaCCA.html	74

Table 1 (cont.) | **Summary of methods and software packages**

Name	Description	Input	Source	Refs
Multitrait association analysis (cont.)				
genomic SEM	Identify SNPs associated with general dimensions of cross-trait liability through SEM	SNP effects and standard errors	https://github.com/MichelNivard/GenomicSEM	72
cFDR	Bayesian conditional analysis	SNP *p*-values	https://github.com/jamesliley/cFDR-common-controls	75,76
CPBayes	Bayesian conditional analysis allowing >2 traits	SNP effects and standard errors	https://CRAN.R-project.org/package=CPBayes	77
GPA, GPA-MDS	Bayesian association analysis incorporating SNP annotation	SNP *p*-values	https://github.com/dongjunchung/GPA	78,79
EPS	Bayesian association analysis incorporating SNP annotation	SNP *p*-values	https://github.com/gordonliu810822/EPS	119
Multitrait prediction				
MTAG	Predictor based on β coefficients from pleiotropy-informed SNP associations	GWAS summary statistics + individual-level genotypes	https://github.com/omeed-maghzian/mtag	64
SMTpred	Combine SNP weights (β coefficient or BLUP) from multiple single-trait predictors	GWAS summary statistics + individual-level genotypes	https://github.com/uqrmaie1/smtpred	90
PleioPred	Bayesian framework for multitrait prediction potentially modelling genome annotation (PleioPred-anno)	GWAS summary statistics + individual-level genotypes	https://github.com/yiminghu/PleioPred	91
Inferences on causality				
MR-Egger	MR using Egger regression	GWAS summary statistics	https://cran.r-project.org/web/packages/MendelianRandomization/index.html	94
MRBase	Server for MR analysis using published GWAS	GWAS summary statistics	http://www.mrbase.org/	95
MR Steiger	Detect directionality in causal relationships	GWAS summary statistics	https://github.com/explodecomputer/causal-directions	97
GSMR	Generalized summary data-based MR, including HEIDI test to exclude SNPs with evidence for horizontal pleiotropy	GWAS summary statistics	http://cnsgenomics.com/software/gsmr/	9
MR-PRESSO	MR after detecting and correcting for horizontal pleiotropy	GWAS summary statistics	https://github.com/rondolab/MR-PRESSO	10
LCV	Latent causal variable model to infer causality, less biased by horizontal pleiotropy	GWAS summary statistics	https://github.com/lukejoconnor/LCV	103
Conditional analysis				
Multi-trait conditional GWAS analysis	Multitrait conditional analysis of GWAS in same individuals	GWAS summary statistics	https://github.com/yangq001/conditional	104
GCTA-mtCOJO	Multitrait conditional analysis of independent GWAS	GWAS summary statistics	http://cnsgenomics.com/software/gcta/	9
GWIS	Approximate conditioned GWAS summary statistics	GWAS summary statistics	https://sites.google.com/site/mgnivard/gwis	105
Fine-mapping causal variants				
RiVIERA	Bayesian framework to combine multitrait SNP associations and annotation for fine mapping	GWAS summary statistics	https://github.com/yueli-compbio/RiVIERA-beta	106
fastPAINTOR	Bayesian framework to combine multitrait SNP associations and annotation for fine mapping	GWAS summary statistics	https://github.com/gkichaev/PAINTOR_V3.0	107

BLUP, best linear unbiased predictor; GCTA, genome-wide complex trait analysis; GCTA-mtCOJO, genome-wide complex trait analysis–multitrait conditional and joint analysis; GREML, genetic restricted maximum likelihood; GRM, genomic relationship matrix; GWAS, genome-wide association study; HEIDI, heterogeneity in dependent instruments; LD, linkage disequilibrium; LDAK, linkage disequilibrium adjusted kinship; LDSC, linkage disequilibrium score regression; MAF, minor allele frequency; MLE, maximum likelihood estimation; MR, Mendelian randomization; SNP, single-nucleotide polymorphism; SEM, structural equation modelling.

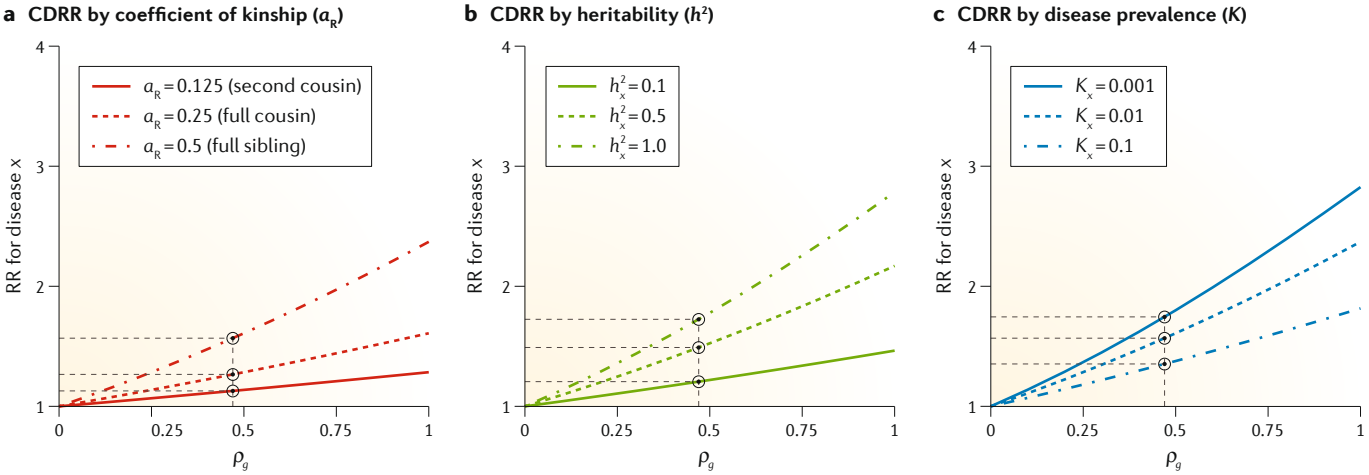


Fig. 3 | **Relation between cross-disorder relative risk (CDRR) and genetic correlation.** The graphs show the relationship between genetic correlation (ρ_g) and the risk ratio (RR) with varying coefficient of relationship (a_R) (part **a**), heritability of disease *x* (h^2_x) (part **b**) and lifetime risk of disease *x* (K_x) (part **c**). Reference parameters are $h^2_y = 0.4$, $K_y = 0.15$ (typical of major depression), $h^2_x = 0.65$, $K_x = 0.01$ (typical of schizophrenia) and $a_R = 0.5$ (that is, full sibling). In part **a**, the sibling ($a_R = 0.5$) of someone with major depressive disorder ($h^2_y = 0.4$, $K_y = 0.15$) has a 1.56-fold increased chance of having schizophrenia $h^2_x = 0.65$, $K_x = 0.01$, $\rho_g = 0.47$) compared to the general population. The risk ratio is lower for more distant relatives (1.26-fold and 1.13-fold increase for $a_R = 0.25$ and 0.125, respectively; dotted black lines). Using similar parameters, part **b** shows that the relative risk for the sibling increases with increasing heritability of disease *x* (1.21, 1.49 and 1.72 for $h^2_x = 0.1$, 0.5 and 1, respectively; dotted black lines). In part **c**, the relative risk for the sibling increases with decreasing lifetime risk of disease *x* (1.35, 1.57 and 1.75 for $K_x = 0.1$, 0.01 and 0.001, respectively; dotted black lines). The Supplementary note provides the theoretical background and code for this figure.

associated with the shared environment[24,25,27], so estimates are unlikely to be biased by these factors. Bivariate GREML analysis simultaneously estimates the genetic variances of the two traits and the genetic covariance between them that best fit the data given the model. It can be applied to data sets that have been collected independently because pairs of individuals across data sets are distantly related. Hence, bivariate models include a GRM that has two block diagonal matrices representing the two univariate GRMs, with the off-diagonal block representing the genetic relationships between pairs of individuals represented in the two data sets.

GWAS summary statistics for unrelated individuals. Linkage disequilibrium score regression (LDSC)[28] was the first method to propose estimation of genetic correlation from GWAS summary statistics. It is based on the observation, expected under polygenicity, that the greater the total amount of LD a SNP has with other genetic variants, the greater its chance of being correlated with causal variants, and the higher its expected association test statistic. Exploiting this relationship allows estimation of SNP-based heritability when using association test statistics for a single trait or estimation of SNP-based coheritability when combining association test statistics from two traits. Specifically, bivariate LDSC[29] uses a weighted regression framework to estimate the coheritability (h_{xy}) from GWAS association statistics for SNP *j* of both traits (z_{xj} and z_{yj}) and the SNP LD scores (l_j). The LD score of SNP *j* is the sum of r^2 LD of SNP *j* with other SNPs obtained from sequencing data and can thus be regarded as a measure of the genetic variation that is 'tagged' by SNP *j*. The regression relationship also depends on the sample size for the two traits (N_x, N_y) and the total number of SNPs (M). The intercept term estimates sample overlap (N_s) and hence reflects the proportion of shared individuals ($\frac{N_s}{\sqrt{N_x N_y}}$) and their phenotypic correlation (ρ_p):

$$E[z_{xj}\, z_{yj} | l_j] = \frac{\rho_p N_s}{\sqrt{N_x N_y}} + \frac{\sqrt{N_x N_y}\; h_{xy}}{M} l_j \qquad (10)$$

LDSC SNP-based heritabilities, which are needed to estimate the genetic correlation, are calculated similarly with $x = y$ (and with an additional intercept term to account for residual confounding, such as population stratification, within a data set). As LDSC is computationally very efficient, summary statistics from GWAS are widely shared and there is no bias introduced by sample overlap, genetic correlations between hundreds of traits can be studied[29–31]. LD Hub[30] provides a publicly available server that hosts LDSC calculations and a library of published GWAS summary statistics.

Genome partitioning. A question of key interest is whether causal variants for a trait are found randomly across the genome or are enriched based on genomic annotation. The univariate GREML approach can model multiple random effects and hence estimate multiple genetic variances using multiple GRMs, each built with SNPs selected on different annotations[32]. The REML approach optimizes the partitioning of the variance to these annotations. The computational efficiency of LDSC in estimating the enrichment of SNP-based heritability in sets of variants with particular genomic annotations has enabled study of genomic partitioning of genetic variance using a stratified LDSC[33] approach.

SNP-based heritability
An estimate of the proportion of the total phenotypic variance attributable to the additive effects of the class of variants (that is, common single-nucleotide polymorphisms (SNPs)) that are typically genotyped and imputed in pursuit of a genome-wide association study. It is often shortened to SNP heritability, but this should be avoided.

Extension of these methods to investigate differences in genetic correlations between traits based on genomic annotations is appealing, but they would generate estimates with high standard errors (which depend on the number of SNPs contributing to the estimates as well as sample sizes[34]). Heritability Estimation from Summary Statistics (ρHESS), which was developed to partition genetic correlations based on genomic regions, addresses this issue by reducing the noise in the LD matrix through principal component-based regularization (that is, block diagonalization)[11]. By contrast, the GeNetic cOVariance Analyzer (GNOVA), which partitions genetic correlations based on functional annotations, uses the method of moments as the underlying framework instead of the weighted regression in LDSC[35]. Both methods have shown that the genetic correlation is not constant across the genome for different trait pairs. For example, 11 regions of statistically significant local genetic correlation (four positive, seven negative) were found between LDL and HDL cholesterol in the absence of genome-wide genetic correlation[11].

Same trait measured in different environments. Bivariate methods can be used to analyse data for the same trait that have been intentionally recorded in two different environments (or populations); data from the two environments are treated as different traits in the analysis. The resulting genetic correlation estimates can reflect the sensitivity of the genetic effects to the chosen environments, and estimates less than one may be indicative of genotype by environment (G × E) interaction. An important caveat, especially for GWAS-derived estimates, is that these analyses should always be benchmarked against estimates from different cohorts of the same trait recorded in the same environment. In this case, the true genetic correlation parameter is one, but, in practice, small sample sizes, differences in participant ascertainment or unrecognized differences in phenotype definition can induce sample heterogeneity, which results in lower estimates[15,36]. Estimates of genetic correlation between samples of different ethnicities are additionally affected by differences in allele frequency and LD structure that may lead to r_g estimates <1 (REF.[37]). In LMM methods, the bivariate GRM can be constructed using allele frequencies estimated from the two different samples, which accounts for both allele frequency and LD differences between the populations[38]. For example, the estimated genetic correlation between European cohorts and East Asian cohorts was 0.76 (s.e. 0.04) for Crohn's disease and 0.79 (s.e. 0.04) for ulcerative colitis[39]. By contrast, the genetic correlation between these ethnicities for attention-deficit hyperactivity disorder (ADHD) was only 0.39 (s.e. 0.15)[40]; however, the genetic correlation estimated between two European ancestry ADHD cohorts was only 0.71 (s.e. 0.17)[15], which indicates sample heterogeneity in these ADHD GWAS. The Popcorn method[37] extends LDSC to allow estimation of genetic correlation between two traits from GWAS conducted in populations of different ethnicity using LD reference panels from both populations. In the absence of sample heterogeneity, an interesting question is whether $r_g < 1$ is because of population-specific allelic effects and/or population-specific allele frequencies. With this in mind, Popcorn estimates the correlation of SNP effect sizes (the genetic-effect correlation) and the correlation of per SNP heritability (the genetic-impact correlation); the genetic-impact correlation is dependent on differences between populations in terms of both effect sizes and allele frequencies. Both in simulations and in application to real data, the estimates were found to be similar[37].

Interpretation of SNP-based estimates

SNP-based genetic correlation estimates are robust to most model assumptions. There is an ongoing debate about model assumptions of GREML and LDSC and their impact on SNP-based heritability estimates[41–45]. By contrast, estimates of genetic correlations are generally shown to be robust to these assumptions[29,44]. To support this conclusion, we summarize the current discussions for SNP-based heritability estimation and then justify why the issues have little impact on estimates of genetic correlations. Briefly, the basic GREML model, one of the models implemented in the genome-wide complex trait analysis (GCTA) software[46], assumes an infinitesimal model and that causal effects are drawn from a normal distribution. In simulations[41], estimates of SNP-based heritability were found to be robust to three key underlying assumptions relating to the genetic architecture of the trait: extent of polygenicity, normality of genetic effects and the inverse relationship between minor allele frequency (MAF) and effect size. However, the estimates were found to be sensitive to the assumption that causal variants are found randomly with respect to the LD patterns of the genome. This led to the introduction of the LD adjusted kinship (LDAK) REML methods[41,43], in which contributions from SNPs in low LD regions are assigned higher weights when constructing the GRM. Instead of using a weighted GRM, LD and MAF stratified GREML (GREML-LDMS)[42] introduced multiple-component GREML models that estimate multiple genetic variances stratified by LD and MAF which sum to the SNP-based heritability. In turn, a multivariance-component LDAK model with LDAK GRM stratified by MAF was also introduced[43]. A comprehensive comparison of analyses that use single or multiple-component GRM constructed with different underlying assumptions on LD and/or MAF-dependent architectures, and based on simulations from genome-sequence data, showed that multicomponent models performed better than single-component models, but biases were observed in all methods depending on the underlying architecture[47]. It is therefore difficult to foresee which biases may occur in real data, as the true genetic architecture is unknown. In these comparisons, LDSC estimates were biased downwards by 5–10% when causal variants were common, and this bias increased as causal variants became less common. Recently, the summary statistics-based method SumHer[44] was introduced, which includes the assumption that low LD SNPs should have higher effect sizes.

Discussion about model assumptions has focused mostly on the estimates of SNP-based genetic variance and heritabilities, but the same concerns apply to the

Genotype by environment (G × E) interaction
Differences in size and/or direction of the effect of genotype on disease risk in two different environments.

Sample heterogeneity
Differences in the effects of genotype on disease risk in two different cohorts. Potential causes include differences in phenotype criteria, ascertainment methods and unknown environmental differences with genotype by environment interaction.

Infinitesimal model
This model assumes that a trait is shaped by a very large number of variants with small (infinitesimal) effects resulting in a normally distributed phenotype. A polygenic architecture of > ~ 10 causal variants is approximated well by normal distribution infinitesimal model theory.

Haseman–Elston regression
Regression of the product of the standardized phenotypes of pairs of individuals on their coefficient of genetic sharing as defined in the genomic relationship matrix.

estimates of SNP-based genetic covariances and coheritabilities between two traits[44]. The key point of discussion is whether the genetic architecture assumed in a model of analysis matches the true, but unknown, genetic architecture of the trait under analysis. A particular concern is the LD properties of causal variants, but for causal variants shared by two traits the same LD properties apply as they are a property of the genome not of the trait. Hence, differences in estimates of (co)variances are expected to approximately cancel out through their impact on both the numerator and the denominator. Therefore, estimates of genetic correlations are observed to be much more robust to underlying assumptions in simulations conducted across different methods[29,44,48,49].

Precision and required sample size for genetic correlation estimates. The standard error of GREML SNP-based heritability estimates depend only on sample size (N) and SNP density (accurately approximated as $316/N$ for GWAS data[34]), but the standard errors of genetic correlation estimates are several-fold larger because they reflect errors of three estimated parameters, that is, the heritabilities of the two traits and the genetic correlation parameter itself (FIG. 4a). Because standard errors can be estimated with good accuracy, power calculations can be undertaken before conducting a study[34]. Empirical standard errors of summary statistics-based methods such as LDSC are ~2× higher than those of GREML. Therefore, LDSC is less powerful to detect genetic correlations that are significantly different from zero or one, for a given sample size, compared to GREML[28,48] (FIG. 4b). The major advantage of summary statistics-based methods is that larger sample sizes can be achieved and they require only a small fraction of the computational expertise and resources required for the methods that use individual-level data.

Genetic correlation estimates are robust to scale transformations. For disease traits, SNP-based heritability estimates are made relative to the phenotypic variance in the sample, which is a function of the proportion of cases in the sample. The raw estimates are transformed based on normal distribution theory to account for sample ascertainment, so that they are interpretable and can be compared across different samples[21,47]. Likewise, raw coheritability estimates reflect the proportions of cases in the two samples. As the transformations apply to the numerator and the denominator, the estimated genetic correlation is scale independent[29].

Genetic correlation estimates are robust to ascertainment and strong environmental factors. Ascertainment of cases (resulting in oversampling of individuals with both high genetic and environmental values) or the presence of environmental factors with strong effects can violate GREML (and hence also LDAK) assumptions that environmental values are normally distributed and lead to a downward bias of SNP-based heritability estimates[49,50]. Haseman–Elston regression is robust to this assumption but gives ~1.5× higher standard errors than GREML-based approaches[51]. Phenotype correlation genotype correlation (PCGC) regression is an extension of Haseman–Elston regression that accounts for ascertainment and covariates in estimation of SNP-based

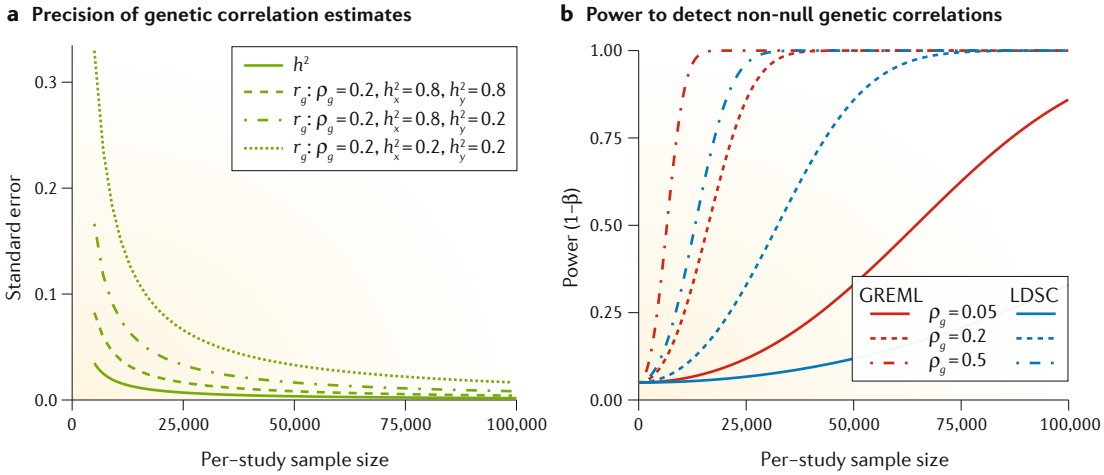
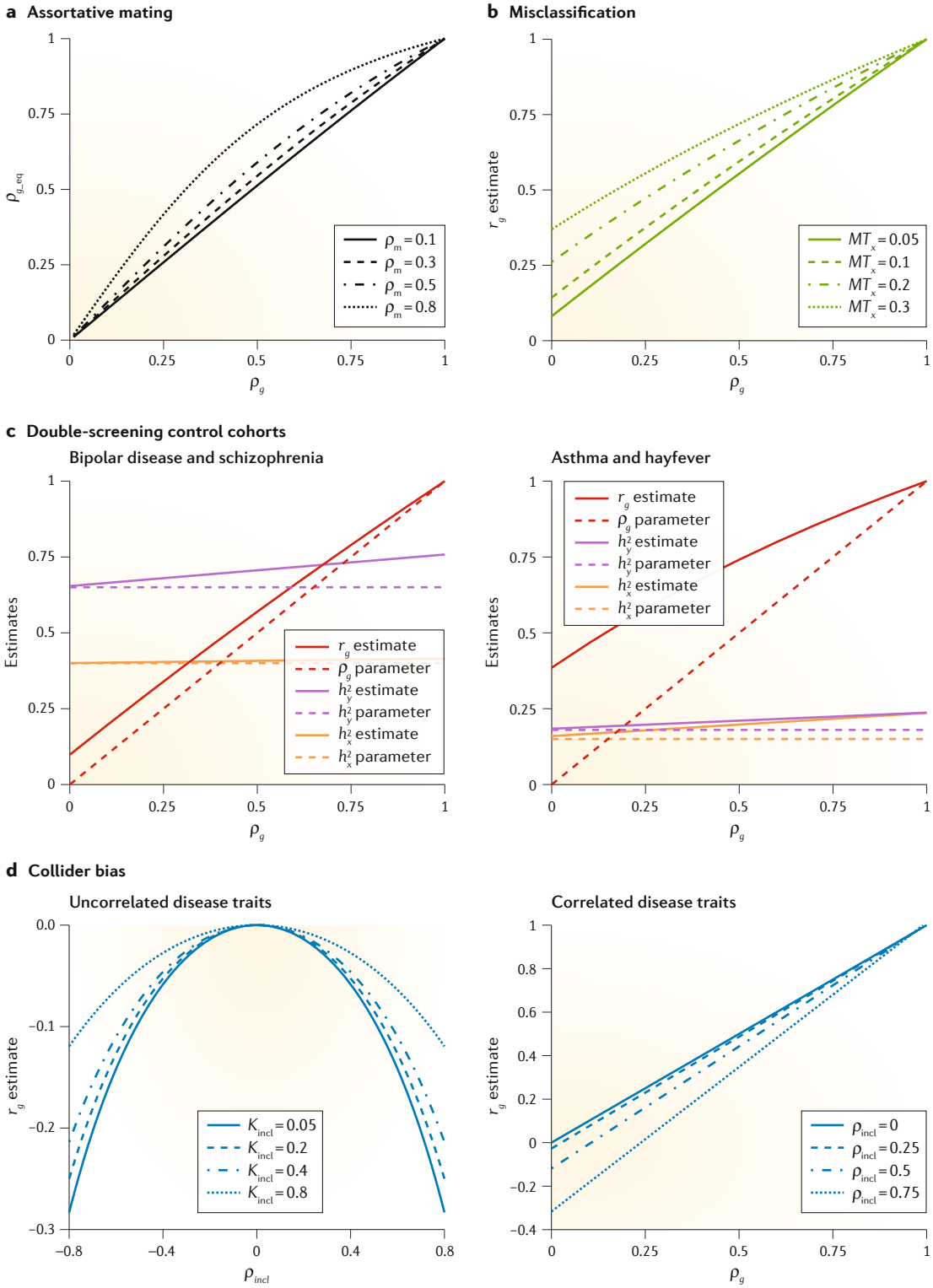


Fig. 4 | **Precision of genetic correlation estimates compared to heritability estimates and power for GREML and LDSC. a** | Standard errors for heritability and genetic correlation estimates obtained from genome-based restricted maximum likelihood (GREML). We use equal sample sizes for the two traits and show that the standard errors for genetic correlation estimates are substantially larger than for heritability estimates, meaning that larger sample sizes are needed to obtain equally accurate genetic correlation estimates compared to heritability estimates. **b** | Power calculations for GREML[34] and linkage disequilibrium score regression (LDSC). We assume that standard errors are approximately twice as large for LDSC regression compared to GREML based on observations from simulations and real data[28,48], in which the LDSC intercept was not constrained. For equal sample sizes, GREML is more powerful to detect genetic correlations than LDSC, but often the individual-level data sets needed for GREML are smaller than those contributing to the genome-wide association study summary statistics used in LDSC. Unless stated otherwise, the following parameters were fixed for both traits: heritability (h^2) = 0.2, lifetime risk (K) = 0.01, proportion of the study sample that are cases (P) = 0.5 and significance threshold (α) = 0.05. Power is described as 1 – type II error (that is, 1 – β). The Supplementary note provides the theoretical background and code for this figure. ρ_g, genetic correlation; r_g, estimated genetic correlation.



Confounding bias
A type of bias that emerges when a covariate, a 'confounder', causally influences the predictor variable and outcome variable. When the confounder is not accounted for, the relationship between predictor and outcome may be biased (confounded).

heritability. However, extensive simulations suggest that genetic correlation estimates obtained by GREML, LDSC and PCGC are not biased by strong ascertainment or strong environmental factors[49].

Impact of population stratification on genetic correlation estimates. Confounding bias due to population stratification can inflate (co)heritability estimates from GREML. Moreover, population and technical confounding is more likely to occur with a binary trait[24]. Hence, stringent SNP quality control is needed when applying GREML to disease data, as well as inclusion of ancestry-informative principal components as fixed-effect covariates. LDSC and SumHer attempt to model inflation of association statistics due to any residual population stratification when estimating SNP-based heritability, but bias may still remain[52]. The impact of population stratification on coheritability estimates and

Fig. 5 | **Bias in estimated genetic parameters. a** | Positive assortative mating increases genetic correlation estimates at equilibrium (ρ_{g_eq}). Here, ρ_m is the phenotypic correlation between mate pairs for trait *x*, heritability of trait *x* (h^2_x) = 0.65 and heritability of trait *y* (h^2_y) = 0.4. **b** | Misclassification inflates genetic correlation estimates (r_g). Here, MT_x represents the misclassification rate of trait *x* as trait *y*. There is no misclassification of trait *y*, h^2_x= 0.65 and h^2_y = 0.4. **c** | Extending the methodology that considered disease misclassification[56], double-screening controls can yield inflated estimates for heritability and genetic correlation. This bias is modest when at least one trait is relatively rare, for example (left) for major depressive disorder (h^2_x = 0.4, K_x = 0.15) and schizophrenia (h^2_y = 0.65, K_y = 0.01), but can be substantial for two common traits such as hayfever (h^2_x = 0.15, K_x = 0.15) and asthma (h^2_y = 0.18, K_y = 0.25)[113] (right). Dashed lines reflect true parameters. **d** | In the panel on the left, collider bias results in a downward bias of the genetic correlation estimates when both traits are associated with the probability of being included in the study, here modelled on the liability scale ($\rho_{incl,x} = \rho_{incl,y}$ as presented on *x* axis). This bias is most pronounced when a smaller proportion of samples is included in the study (K_{incl}). For this panel, $h^2_x = h^2_y = 0.4$ and $\rho_g = 0$. In the panel on the right, trait parameters are chosen to reflect major depressive disorder (h^2_x= 0.4, K_x = 0.15) and schizophrenia (h^2_y = 0.65, K_y = 0.01). Again, $\rho_{incl,x} = \rho_{incl,y}$ and we set $\rho_g = \rho_e$ as presented on the *x* axis. The bias in genetic correlation estimates due to collider bias is most pronounced when traits are uncorrelated. The Supplementary note provides the theoretical background and code for all panels of this figure. ρ_g, genetic correlation.

genetic correlation estimates for these methods has not been thoroughly investigated; however, for LDSC, theory predicts that it affects the bivariate LDSC intercept but not coheritability estimates[53].

Positive assortative mating increases genetic correlation. Compared to random mating, positive assortative mating on a trait (*x*) increases its genetic variance at equilibrium ($h^2_{x\_eq}$) and the genetic variance of any correlated trait (*y*)[54]. Their genetic correlation at equilibrium, ρ_{g_eq}, compared to the genetic correlation under random mating (ρ_g) can be expressed as[54]:

$$\rho_{g\_eq} = \rho_g \sqrt{\frac{1}{1-\rho_m h^2_{x\_eq}\left(1-\rho^2_g\right)}} \quad (11)$$

where ρ_m is the phenotypic correlation between mates for the assortatively-mated trait *x*. For disease traits, ρ_m would be the correlation between liability to disease. As a result, assortative mating increases genetic correlation estimates at equilibrium (FIG. 5a). For assortative mating typical of humans ($\rho_m < 0.3$)[55], ρ_{g_eq}/ρ_g has a maximum of ~1.2.

Assortative mating
Mating selection on a trait where the phenotypes of mates are positively correlated. Examples of assortative mating in humans include height or educational attainment.

Collider bias
A type of bias that emerges when estimates are conditioned on a covariate, a 'collider,' that is causally influenced by both the predictor variable and outcome variable.

Misclassification inflates genetic correlation estimates. Misclassification, or misdiagnosis, may occur when two traits share phenotypic characteristics, such as Crohn's disease and ulcerative colitis, and can lead to spurious estimates of genetic correlation[56]. The impact of misclassification on the estimated genetic correlation can be quantified from theory[56], and the bias is greatest when two diseases are genetically unrelated (FIG. 5b). However, if two diseases are truly genetically correlated then perhaps phenotypic overlap in clinical presentation leading to diagnostic ambiguity is expected. For example, in schizophrenia and bipolar disorder[57], changing diagnoses over time is likely a real reflection of the longitudinal symptom profile in some people. Hence, the likelihood of overestimation of genetic correlation should be guided by the context of realistic misclassification rates for each disease pair.

Individual-level genotype data can help detect potential misclassification. The Breaking Up Heterogeneous Mixture Based On cross(X)-locus correlations (BUHMBOX) algorithm leverages correlation patterns of disease risk loci to detect subgroup heterogeneity[58]. There have been few applications to real data but one example used rheumatoid arthritis data and indicated that the genetic correlation between seropositive and seronegative types was (partly) due to subgroup heterogeneity possibly introduced by false-negative serum rheumatoid blood factor tests (misclassification)[58]. Notably, subgroup heterogeneity can theoretically also result from (molecular) subtypes or vertical pleiotropy, which can, in contrast to misclassification, be of great interest.

Double-screening control cohorts can inflate genetic correlation estimates. In addition to screening controls to exclude the case trait, it is also common to exclude control subjects with potentially related diseases in what we term here as double-screening of control cohorts. Although this ascertainment bias may increase power for detection in GWAS, it can induce biased estimates of genetic parameters (FIG. 5c). When the true genetic correlation is zero, a non-zero estimate reflects the increased prevalence of risk alleles for the secondary trait in the cases relative to the doubly-screened controls. This bias increases with increased population risk (FIG. 5c) and increased differences in heritability between the primary and secondary trait.

Collider bias can affect genetic correlation estimates. Collider bias can introduce spurious correlations when two traits both influence a third 'collider' variable and their association is conditioned on this third variable. A special form of collider bias arises through selection bias in which both traits influence the probability that an individual is included in the study (FIG. 5d). Collider bias has been acknowledged as a potential pitfall in the use of large-scale biobanks[59], in which there may be a high degree of self-ascertainment. For example, in the UK Biobank study, only 5% of invitations to participate were accepted[60], with participants having higher educational status and lower prevalence of smoking[59]. Hence, the genetic correlation estimated between educational status and smoking obtained from UK Biobank data may be biased.

Uses of genetic correlations

If non-zero genetic correlations are estimated between two traits, analyses can be constructed that could improve power to detect new disease-associated variants, improve genetic risk prediction, make inferences on causality, perform conditional analyses and describe the biological aetiology of complex traits. Methods and software packages (TABLE 1) that can be used to achieve these aims are discussed below, with a focus on those that use GWAS summary statistics because these are broadly applicable.

Identification of new trait-associated variants. The combined association analyses of correlated traits can increase power to detect new SNP–trait associations. A wide variety of methods is available that combine GWAS summary statistics to identify trait-associated SNPs (TABLE 1). In general, methods encompass extensions of (inverse-variance weighted) meta-analysis[61–65], the score-based association test[65], linear combinations of GWAS test statistics[66–71], structural equation modelling[72], multivariate models[73,74] and Bayesian methods where the prior is informed by SNP associations in correlated traits[35,75–79]. A detailed description of these methods is beyond the scope of this review and is provided elsewhere[80,81]. The increase in power of the combined analyses compared to single-trait analyses can be interpreted as equivalent to the increase in sample size for the single-trait association analysis. For example, the multitrait analysis of the genetically correlated ($\rho_g \approx 0.7$) traits depressive symptoms ($N = 354{,}311$), neuroticism ($N = 168{,}105$) and subjective well-being ($N = 388{,}542$) with a large proportion of overlapping individuals led to an increase in power compared to the single-trait analysis, which is equivalent to a 27%, 55% and 55% increase in sample size, respectively[64].

Improved genetic risk prediction. Genetic risk prediction is of great interest for common complex traits because it can inform diagnostic decision-making and early intervention (prevention) strategies[82–84]. The accuracy of genetic risk prediction is dictated by disease characteristics (such as heritability and lifetime risk)[82], but also study design (including reference sample size, population and trait)[85]. Leveraging SNP effect estimates from GWAS of genetically correlated traits is equivalent to increasing the effective discovery sample size of the focal trait[86] (FIG. 6a). Predictors for case–control traits with relatively low heritability can particularly benefit from this multitrait approach[87]. The increase in predictive accuracy, measured as the area under the receiver operator curve, can be predicted from theory[85,88] (FIG. 6b). An increase in prediction accuracy has indeed been observed when combining individual-level genotypes for psychiatric traits using ridge regression[89] or for inflammatory bowel diseases using best linear unbiased predictors of SNP effects, such as MT-GBLUP[86]. Methods for multitrait prediction have also been extended to work with single-trait GWAS summary statistics (such as SMTPred[90] and Multi-Trait Analysis of GWAS (MTAG)[64]) and can include genome annotation in the prediction model (for example, PleioPred-anno)[91].

Inferences on causality. When traits *x* and *y* are found to be genetically correlated, it may be of interest to understand whether the correlation has been induced by a causal relationship, that is, trait *x* causes trait *y*

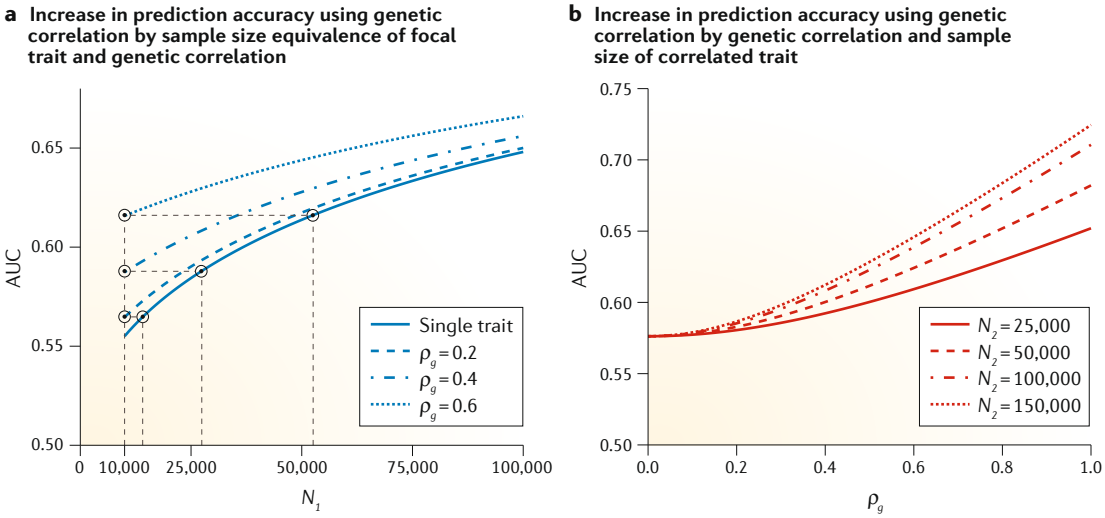


Fig. 6 | **Prediction accuracy increases when correlated traits are combined.** Based on the theory in quantitative trait genetics described in REF.[85], the prediction accuracy of a genetic predictor (parameterized as the area under receiver operating characteristic curve (AUC)) in the target population sample increases when two genetically correlated traits are combined to calculate single-nucleotide polymorphism (SNP) weights for 1,000,000 SNPs (*M*). The target trait in this example is major depressive disorder (heritability (h^2) = 0.40, lifetime risk (K) = 0.15, proportion of cases (P) = 0.5). The genetically correlated trait is schizophrenia (h^2 = 0.65, K = 0.01, P = 0.5). The effective number of chromosome segments is chosen to reflect an unrelated sample of the European population (M_{eff} = 50,000). Part **a** illustrates the increase in prediction accuracy that is achieved by including information from the correlated trait equivalent to the increase in the sample size of the first trait (N_1) in a single-trait analysis. Compared to the single-trait analysis in 10,000 individuals, adding a secondary trait with estimated genetic correlation (r_g) = 0.2, 0.4 or 0.6 and sample size N_2 = 50,000 results in an increase in prediction accuracy equivalent to an increase in sample size of focal trait genome-wide association studies (GWAS) (N_1) of 40%, 260% and 510%, respectively (dotted black lines). In part **b**, the prediction accuracy increases with increasing genetic correlation and increasing sample size of the discovery GWAS of the correlated trait (N_2), here N_1 = 20,000. The Supplementary note provides the theoretical background and code for this figure. ρ_g, genetic correlation.

(or vice versa). Identifying causality is particularly important if the putative causal trait is potentially modifiable, such as smoking or LDL cholesterol levels. However, only formal tests for causality justify these claims. The gold standard to prove causality is a randomized clinical trial, which can be costly and, in some instances, unethical or impossible to conduct. Genetic correlations can be leveraged to aid inferences on causality through Mendelian randomization (MR, reviewed elsewhere[92,93]) and is a cost-effective way to explore causality. Briefly, if trait *x* (for example, diabetes) increases the risk of trait *y* (for example, cardiovascular disease), then all risk factors (the instrumental variables), including genetic risk factors, for trait *x* will, to some consistent proportional extent, also increase the risk of trait *y*. A strong assumption in MR analyses is the absence of horizontal pleiotropy, that is, the instrumental variable does not affect trait *y* directly. Numerous methods for performing MR analyses[9,94,95] and detecting horizontal pleiotropy[9,10,96] are available (TABLE 1). MR Steiger[97] can help to detect directionality of causal relationships and to disentangle causal relationships between multiple related risk factors, and multivariate[98,99] or conditional[9] MR analyses can be applied. An illustrative example of conditional MR identified a causal effect of LDL cholesterol levels, but not HDL cholesterol levels, on cardiovascular disease[9], which is reflected in the results of randomized trials[100–102]. The latent causal variable (LCV) method[103] has been proposed to better differentiate causality and partial causality from horizontal pleiotropy, but a limitation is that it has not been extended to multitrait conditional analyses.

Conditional analysis. When it is known that two traits are genetically correlated and potentially causally related, it can be equally interesting to focus on which SNPs induce phenotypic heterogeneity and cause disease *x* to be different from disease *y*. Multitrait conditional analyses that condition a SNP–trait association on a second disease can provide this insight. When only summary statistics are available, conditional analyses can be performed for GWAS with fully overlapping individuals[104] and for completely independent GWAS (for example, using genome-wide complex trait analysis–multitrait conditional and joint analysis (GCTA-mtCOJO)[9]). If the genetic correlation between the traits reflects a purely causal relationship, then the SNP effects for trait *y* conditional on trait *x* are expected to be uncorrelated with SNP effects of trait *x* ($r_{gy|x,x} = 0$). However, if the genetic correlation between the traits includes horizontal pleiotropy where the genetic sharing may not be the same across the genome, then $r_{gy|x,x}$ may differ from zero. The Genome-Wide Inferred Statistics (GWIS)[105] method conditions trait *y* on trait *x*, forcing $r_{gy|x,x} = 0$. This approach was used to disentangle the genetic correlation between schizophrenia and educational attainment, attributing the observed genetic correlation to only those SNPs that are shared between schizophrenia and bipolar disorder[105]. Like GCTA-mtCOJO, GWIS assumes that the same adjustment applies across the genome. Therefore, both approaches may generate results that are difficult to interpret when the true sharing of genetic risk is variable across the genome.

Improved interpretation of GWAS results. The development of frameworks to include multiple correlated traits to translate GWAS results to functional biology is still in the early stages but is likely to become an area of active research. Similar to single-trait heritability enrichment analyses[33], the GNOVA framework aims to elucidate the biological processes underpinning genetic correlations by identifying functionally annotated sets of SNPs that contribute most to genetic correlations[35]. Furthermore, genetic correlations can be leveraged to help fine-map causal variants in GWAS loci, under an assumption of shared causal variants between traits, as illustrated by the Risk Variant Inference using Epigenomic Reference Annotation (RiVIERA)[106] method; this Bayesian framework approach identified more causal variants that regulated gene expression in disease-relevant tissue when multiple correlated traits were combined compared to single-trait analyses. Similarly, the Probabilistic Annotation INtegraTOR (fastPAINTOR)[107] algorithm combines multiple correlated traits and genome annotation to prioritize causal variants in GWAS loci, explicitly modelling multiple causal variants within a locus.

Conclusions and future perspectives

The past decade has seen great advances in our understanding of complex traits and common diseases and disorders. It is now well recognized that pleiotropy is ubiquitous[29,31,108] and that the number of uncorrelated traits is constrained[109,110]. The availability of large independently collected data sets for multiple traits and the sharing of GWAS summary statistics enable genetic correlations to be estimated and used on an unprecedented scale. Although ongoing discussion focuses on model assumptions and how they can bias estimates of genetic variances and covariance, simulation studies consistently conclude that genetic correlation estimates are robust to these assumptions. Large-scale genotype–phenotype resources show that a genetic contribution can be attributed to the vast majority of measured traits and lifestyle factors, further increasing the potential of studies on genetic correlation to describe disease biology[111]. In the short term, genetic correlation estimates will help to find disease-associated genetic variation and improve polygenic risk prediction as it makes its way into clinical practice[83,112]. They may also contribute to improved nosology and diagnosis, risk stratification to inform clinical trials and lifestyle interventions. Here, our focus is common, polygenic, disease traits, coded as a simple dichotomy of health and disease. However, the disease process and its estimated genetic contribution can be conceptualized as the complex product of multiple intermediate phenotypes at the level of gene expression, cumulative over time and cell types. The coming decades may generate the data that, through studies of genetically correlated traits, enable deconstruction of disease risk at the molecular level.

Published online: 06 June 2019

1. Polderman, T. J. C. et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).
2. Craddock, N. & Owen, M. J. The beginning of the end for the Kraepelinian dichotomy. *Br. J. Psychiatry* **186**, 364–366 (2005).
3. Maret-Ouda, J., Tao, W., Wahlin, K. & Lagergren, J. Nordic registry-based cohort studies: possibilities and pitfalls when combining Nordic registry data. *Scand. J. Public Health* **45** (Suppl. 17), 14–19 (2017).
4. Lichtenstein, P. et al. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**, 234–239 (2009).
This work reports a population-scale data set for estimation of genetic correlation between diseases based on family data.
5. Stearns, F. W. One hundred years of pleiotropy: a retrospective. *Genetics* **186**, 767–773 (2010).
6. Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. *Trends Genet.* **29**, 66–73 (2013).
7. Grüneberg, H. An analysis of the 'pleiotropic' effects of a new lethal mutation in the rat (*Mus norvegicus*). *Proc. R. Soc. Lond. B* **125**, 123–144 (1938).
8. Wagner, G. P. & Zhang, J. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet.* **12**, 204 (2011).
9. Zhu, Z. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
10. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
11. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* **101**, 737–751 (2017).
12. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl Acad. Sci. USA* **109**, 1193–1198 (2012).
13. Cheverud, J. M. A comparison of genetic and phenotypic correlations. *Evolution* **42**, 958–968 (1988).
This study describes phenotypic correlations as estimates of genetic correlations based on observation data.
14. Rzhetsky, A., Wajngurt, D., Park, N. & Zheng, T. Probing genetic overlap among complex human phenotypes. *Proc. Natl Acad. Sci. USA* **104**, 11694–11699 (2007).
15. Cross-Disorder Group of the Psychiatric Genomics Consortium et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
This study is among the first to estimate genetic correlation between diseases using independently collected GWAS samples.
16. Tenesa, A. & Haley, C. S. The heritability of human disease: estimation, uses and abuses. *Nat. Rev. Genet.* **14**, 139–149 (2013).
17. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **29**, 51–76 (1965).
18. Reich, T., James, J. W. & Morris, C. A. The use of multiple thresholds in determining the mode of transmission of semi-continuous traits. *Ann. Hum. Genet.* **36**, 163–184 (1972).
19. Wray, N. R. & Gottesman, I. I. Using summary data from the Danish national registers to estimate heritabilities for schizophrenia, bipolar disorder, and major depressive disorder. *Front. Genet.* **3**, 118 (2012).
20. Pearson, K. I. Mathematical contributions to the theory of evolution. — VII. On the correlation of characters not quantitatively measurable. *Philos. Trans. A Math. Phys. Eng. Sci.* **195**, 1–405 (1900).
21. Sham, P. *Statistics in Human Genetics* (Wiley, 1998).
22. Olsson, U. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44**, 443–460 (1979).
23. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).
This study introduces the bivariate GREML method to estimate genetic correlation from genome-wide SNP data.
24. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
25. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
26. Falconer, D. S. & Mackay, T. F. C. *Introduction to Quantitative Genetics* 4th edn (Pearson, 1996).
27. Zaitlen, N. et al. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLOS Genet.* **9**, e1003520 (2013).
28. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
29. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
This study introduces the LDSC method to estimate genetic correlation from GWAS summary data.
30. Zheng, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
This work introduces LD Hub, a server that hosts GWAS summary statistics and LDSC analyses to estimate genetic correlations.
31. Brainstorm Consortium et al. Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).
32. Yang, J. et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
33. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
34. Visscher, P. M. et al. Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLOS Genet.* **10**, e1004269 (2014).
35. Lu, Q. et al. A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. *Am. J. Hum. Genet.* **101**, 939–964 (2017).
36. Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
37. Brown, B. C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).
38. de Candia, T. R. et al. Additive genetic variation in schizophrenia risk is shared by populations of African and European descent. *Am. J. Hum. Genet.* **93**, 463–470 (2013).
39. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
40. Yang, L. et al. Polygenic transmission and complex neurodevelopmental network for attention deficit hyperactivity disorder: genome-wide association study of both common and rare variants. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **162B**, 419–430 (2013).
41. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
42. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
43. Speed, D. et al. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
44. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**, 277–284 (2019).
45. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling S-LDSC and LDAK models and functional enrichment estimates. Preprint at *bioRxiv* https://doi.org/10.1101/256412 (2018).
46. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
47. Evans, L. M. et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* **50**, 737–745 (2018).
48. Ni, G., Moser, G., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Wray, N. R. & Lee, S. H. Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. *Am. J. Hum. Genet.* **102**, 1185–1194 (2018).
49. Weissbrod, O., Flint, J. & Rosset, S. Estimating SNP-based heritability and genetic correlation in case–control studies directly and with summary statistics. *Am. J. Hum. Genet.* **103**, 89–99 (2018).
50. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl Acad. Sci. USA* **111**, E5272–E5281 (2014).
51. Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* **49**, 1304–1310 (2017).
52. Holmes, J. B., Speed, D. & Balding, D. J. Summary statistic analyses do not correct confounding bias. Preprint at *bioRxiv* https://doi.org/10.1101/532069 (2019).
53. Yengo, L., Yang, J. & Visscher, P. M. Expectation of the intercept from bivariate LD score regression in the presence of population stratification. Preprint at *bioRxiv* https://doi.org/10.1101/310565 (2018).
54. Gianola, D. Assortative mating and the genetic correlation. *Theor. Appl. Genet.* **62**, 225–231 (1982).
55. Peyrot, W. J., Robinson, M. R., Penninx, B. W. J. H. & Wray, N. R. Exploring boundaries for the genetic consequences of assortative mating for psychiatric traits. *JAMA Psychiatry* **73**, 1189–1195 (2016).
56. Wray, N. R., Lee, S. H. & Kendler, K. S. Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes. *Eur. J. Hum. Genet.* **20**, 668–674 (2012).
57. Bromet, E. J. et al. Diagnostic shifts during the decade following first admission for psychosis. *Am. J. Psychiatry* **168**, 1186–1194 (2011).
58. Han, B. et al. A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases. *Nat. Genet.* **48**, 803–810 (2016).
This work describes a method that tries to distinguish between genetic correlation driven by sample heterogeneity and that driven by trait pleiotropy.
59. Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider scope: when selection bias can substantially influence observed associations. *Int. J. Epidemiol.* **47**, 226–235 (2018).
60. Allen, N. et al. UK Biobank: current status and what it means for epidemiology. *Health Policy Technol.* **1**, 123–126 (2012).
61. Vuckovic, D., Gasparini, P., Soranzo, N. & Iotchkova, V. MultiMeta: an R package for meta-analyzing multi-phenotype genome-wide association studies. *Bioinformatics* **31**, 2754–2756 (2015).
62. Bhattacharjee, S. et al. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.* **90**, 821–835 (2012).
63. Qi, G. & Chatterjee, N. Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits. *PLOS Genet.* **14**, e1007549 (2018).
64. Turley, P. et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
65. Ray, D. & Boehnke, M. Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genet. Epidemiol.* **42**, 134–145 (2018).
66. O'Brien, P. C. Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087 (1984).
67. Xu, X., Tian, L. & Wei, L. J. Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics* **4**, 223–229 (2003).
68. Yang, Q., Wu, H., Guo, C.-Y. & Fox, C. S. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genet. Epidemiol.* **34**, 444–454 (2010).
69. Bolormaa, S. et al. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLOS Genet.* **10**, e1004198 (2014).
70. Zhu, X. et al. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet.* **96**, 21–36 (2015).

71. He, L. et al. Pleiotropic meta-analyses of longitudinal studies discover novel genetic variants associated with age-related diseases. *Front. Genet.* **7**, 179 (2016).
72. Grotzinger, A. D. et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* **3**, 513–525 (2019).
73. van der Sluis, S., Posthuma, D. & Dolan, C. V. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLOS Genet.* **9**, e1003235 (2013).
74. Cichonska, A. et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics* **32**, 1981–1989 (2016).
75. Andreassen, O. A. et al. Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLOS Genet.* **9**, e1003455 (2013).
76. Liley, J. & Wallace, C. A pleiotropy-informed Bayesian false discovery rate adapted to a shared control design finds new disease associations from GWAS summary statistics. *PLOS Genet.* **11**, e1004926 (2015).
77. Majumdar, A., Haldar, T., Bhattacharya, S. & Witte, J. S. An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations. *PLOS Genet.* **14**, e1007139 (2018).
78. Chung, D., Yang, C., Li, C., Gelernter, J. & Zhao, H. GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLOS Genet.* **10**, e1004787 (2014).
79. Wei, W. et al. GPA-MDS: a visualization approach to investigate genetic architecture among phenotypes using GWAS results. *Int. J. Genomics* **2016**, 6589843 (2016).
80. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).
81. Shriner, D. Moving toward system genetics through multiple trait analysis in genome-wide association studies. *Front. Genet.* **3**, 1 (2012).
82. Wray, N. R. et al. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).
83. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
84. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
85. Lee, S. H., Clark, S. & van der Werf, J. H. J. Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. *PLOS ONE* **12**, e0189775 (2017).
86. Maier, R. et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am. J. Hum. Genet.* **96**, 283–294 (2015).
87. Guo, G. et al. Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genet.* **15**, 30 (2014).
88. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLOS Genet.* **9**, e1003348 (2013).
89. Li, C., Yang, C., Gelernter, J. & Zhao, H. Improving genetic risk prediction by leveraging pleiotropy. *Hum. Genet.* **133**, 639–650 (2014).
90. Maier, R. M. et al. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat. Commun.* **9**, 989 (2018).
91. Hu, Y. et al. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLOS Genet.* **13**, e1006836 (2017).
92. Pingault, J.-B. et al. Using genetic data to strengthen causal inference in observational research. *Nat. Rev. Genet.* **19**, 566–580 (2018).
93. Smith, G. D., Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–R98 (2014).
94. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
95. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).
96. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
97. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLOS Genet.* **13**, e1007081 (2017).
98. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.* **181**, 251–260 (2015).
This study introduces MR, a method to determine whether genetic correlation results from a causal relationship.
99. Do, R. et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat. Genet.* **45**, 1345–1352 (2013).
100. Baigent, C. et al. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. *Lancet* **366**, 1267–1278 (2005).
101. Nissen, S. E. et al. Effect of torcetrapib on the progression of coronary atherosclerosis. *N. Engl. J. Med.* **356**, 1304–1316 (2007).
102. Barter, P. J. et al. Effects of torcetrapib in patients at high risk for coronary events. *N. Engl. J. Med.* **357**, 2109–2122 (2007).
103. O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat. Genet.* **50**, 1728–1734 (2018).
104. Deng, Y. & Pan, W. Conditional analysis of multiple quantitative traits based on marginal GWAS summary statistics. *Genet. Epidemiol.* **41**, 427–436 (2017).
105. Nieuwboer, H. A., Pool, R., Dolan, C. V., Boomsma, D. I. & Nivard, M. G. GWIS: genome-wide inferred statistics for functions of multiple phenotypes. *Am. J. Hum. Genet.* **99**, 917–927 (2016).
106. Li, Y. & Kellis, M. Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Res.* **44**, e144 (2016).
107. Kichaev, G. et al. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **33**, 248–255 (2017).
108. Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
109. Barton, N. H. Pleiotropic models of quantitative variation. *Genetics* **124**, 773–782 (1990).
110. Walsh, B. & Blows, M. W. Abundant genetic variation + strong selection = multivariate genetic constraints: a geometric view of adaptation. *Annu. Rev. Ecol. Evol. Syst.* **40**, 41–59 (2009).
This work puts forward arguments for multivariate genetic constraints and strong limits on the number of independent traits.
111. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
112. Inouye, M. et al. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).
113. Ferreira, M. A. et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* **49**, 1752–1757 (2017).
114. Lee, S. H. & van der Werf, J. H. J. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics* **32**, 1420–1422 (2016).
115. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
116. Loh, P.-R. et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
117. Cotsapas, C. et al. Pervasive sharing of genetic effects in autoimmune disease. *PLOS Genet.* **7**, e1002254 (2011).
118. Dai, M. et al. Joint analysis of individual-level and summary-level GWAS data by leveraging pleiotropy. *Bioinformatics* **35**, 1729–1736 (2018).
119. Liu, J., Wan, X., Ma, S. & Yang, C. EPS: an empirical Bayes approach to integrating pleiotropy and tissue-specific information for prioritizing risk genes. *Bioinformatics* **32**, 1856–1864 (2016).

Acknowledgements
W.v.R. was funded by the ALS Foundation Netherlands. W.J.P. was funded by an NWO Veni grant (91619152). S.H.L. is an ARC Future Fellow (FT160100229). N.R.W. acknowledges funding from the Australian National Health and Medical Research Council (1078901, 1087889 and 1113400). W.v.R. and N.R.W. acknowledge funding from the EU Joint Programme – Neurodegenerative Disease Research (JPND) project (Australia, NHMRC 1151854; The Netherlands, ZonMW project number 733051071). The authors thank K. Tilling, G. Davey Smith and the members of the University of Queensland Program in Complex Trait Genomics for their insightful discussions.

Author contributions
All authors researched data for the article, made substantial contributions to discussions of the content and reviewed and/or edited the manuscript before submission. W.v.R. and N.R.W. wrote the article.

Competing interests
The authors declare no competing interests.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reviewer information
Nature Reviews Genetics thanks D. Balding, B. Pasaniuc and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Supplementary information
Supplementary information is available for this paper at https://doi.org/10.1038/s41576-019-0137-z.

RELATED LINKS
BUHMBOX: http://software.broadinstitute.org/mpg/buhmbox/
fastPAINTOR: https://github.com/gkichaev/PAINTOR_V3.0
GCTA: http://cnsgenomics.com/software/gcta/
GNOVA: https://github.com/xtonyjiang/GNOVA
GWIS: https://sites.google.com/site/mgnivard/gwis
JPND: www.jpnd.eu
LCV: https://github.com/lukejoconnor/LCV
LDAK: http://dougspeed.com/ldak
LD Hub: http://ldsc.broadinstitute.org/ldhub/
LDSC: https://github.com/bulik/ldsc
MR Steiger: https://github.com/explodecomputer/causal-directions
MTAG: https://github.com/omeed-maghzian/mtag
PCGC: https://data.broadinstitute.org/alkesgroup/PCGC/
ρHESS: https://github.com/huwenboshi/hess
PleioPred: https://github.com/yiminghu/PleioPred
Popcorn: https://github.com/brielin/Popcorn
RiVIERA: https://github.com/yueli-compbio/RiVIERA-beta
SMTpred: https://github.com/uqrmaie1/smtpred
SumHer: http://dougspeed.com/sumher/