PLOS | GENETICS

Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples

Peter M. Visscher[1,2]*, Gibran Hemani[1,2], Anna A. E. Vinkhuyzen[1], Guo-Bo Chen[1], Sang Hong Lee[1], Naomi R. Wray[1], Michael E. Goddard[3,4], Jian Yang[1,2]*

[1] The University of Queensland, Queensland Brain Institute, Brisbane, Queensland, Australia, [2] The University of Queensland Diamantina Institute, The Translational Research Institute, Brisbane, Queensland, Australia, [3] University of Melbourne, Department of Food and Agricultural Systems, Parkville, Victoria, Australia, [4] Biosciences Research Division, Department of Primary Industries, Bundoora, Victoria, Australia

Abstract

We have recently developed analysis methods (GREML) to estimate the genetic variance of a complex trait/disease and the genetic correlation between two complex traits/diseases using genome-wide single nucleotide polymorphism (SNP) data in unrelated individuals. Here we use analytical derivations and simulations to quantify the sampling variance of the estimate of the proportion of phenotypic variance captured by all SNPs for quantitative traits and case-control studies. We also derive the approximate sampling variance of the estimate of a genetic correlation in a bivariate analysis, when two complex traits are either measured on the same or different individuals. We show that the sampling variance is inversely proportional to the number of pairwise contrasts in the analysis and to the variance in SNP-derived genetic relationships. For bivariate analysis, the sampling variance of the genetic correlation additionally depends on the harmonic mean of the proportion of variance explained by the SNPs for the two traits and the genetic correlation between the traits, and depends on the phenotypic correlation when the traits are measured on the same individuals. We provide an online tool for calculating the power of detecting genetic (co)variation using genome-wide SNP data. The new theory and online tool will be helpful to plan experimental designs to estimate the missing heritability that has not yet been fully revealed through genome-wide association studies, and to estimate the genetic overlap between complex traits (diseases) in particular when the traits (diseases) are not measured on the same samples.

Citation: Visscher PM, Hemani G, Vinkhuyzen AAE, Chen G-B, Lee SH, et al. (2014) Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples. PLoS Genet 10(4): e1004269. doi:10.1371/journal.pgen.1004269

Editor: Gregory S. Barsh, Stanford University School of Medicine, United States of America

Received September 25, 2013; **Accepted** February 3, 2014; **Published** April 10, 2014

Funding: This research was supported by the Australian Research Council (FT0991360, DP1093502 and DP130102666), the Australian National Health and Medical Research Council (APP1011506, APP1047956, APP1048853 and APP1052684), and the National Institutes of Health (GM099568, GM075091 and MH100141). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: peter.visscher@uq.edu.au (PMV); jian.yang@uq.edu.au (JY)

Introduction

Genome-wide association studies (GWAS) have been extremely successfully in identifying genetic variants associated with complex traits and diseases in humans [1]. In GWAS, hundreds of thousands or millions of SNPs are tested one by one for statistical evidence of association with a trait, and to avoid false positive discoveries due to the very large number of statistical tests being conducted, usually a very stringent p-value threshold, e.g. 5×10^{-8}, is used to report a significant finding. Therefore, if there are many genes each with a small effect affecting the trait, most of these genetic variants will fail to pass the stringent threshold and remain undetected. This is one of the explanations of the 'missing heritability' question, that genetic variants identified from GWAS so far explain a fraction of the heritability for complex traits [2]. We proposed a method, which is able to estimate the total amount of variance explained by all SNPs together without testing the SNPs individually for a quantitative trait [3], and subsequently extended it to the estimation of missing heritability for binary disease data from ascertained case-control studies [4]. The analyses until recently only included common SNPs (e.g. minor allele frequency >0.01). The estimate quantifies the overall contribution from the additive effects of all SNPs, which is the upper limit of the proportion of variance that is captured by the additive effects of the set of SNPs used in the estimation, and is also the lower limit of the narrow-sense heritability of the trait. We also extended the method to estimate the genetic correlation between two traits using SNP data [5,6]. In contrast to the traditional (co)variance estimation methods that rely on pedigree information (family/twin studies), our method uses unrelated samples from a general population and the genetic (co)variance is estimated using a genetic relationship matrix (GRM) estimated from SNPs. The estimate of genetic variance using SNP data in unrelated individuals is free of confounding from common environment effects shared between close relatives that are difficult to model in family-based analyses, and is directly comparable to results from GWAS, because both are based on the same experimental design. For multiple trait analysis, the SNP-based approach allows the estimation of the genetic correlation between complex traits measured on different samples [6,7]_ENREF_8. This is important in particular for estimating the genetic correlation between diseases because multiple diseases are unlikely to co-segregate in sufficiently large pedigrees to allow estimation using traditional pedigree design. The SNP-based method has the flexibility of

Author Summary

Genome-wide association studies (GWAS) have identified thousands of genetic variants for hundreds of traits and diseases. However, the genetic variants discovered from GWAS only explained a small fraction of the heritability, resulting in the question of "missing heritability". We have recently developed approaches (called GREML) to estimate the overall contribution of all SNPs to the phenotypic variance of a trait (disease) and the proportion of genetic overlap between traits (diseases). A frequently asked question is that how many samples are required to estimate the proportion of variance attributable to all SNPs and the proportion of genetic overlap with useful precision. In this study, we derive the standard errors of the estimated parameters from theory and find that they are highly consistent with those observed values from published results and those obtained from simulation. The theory together with an online application tool will be helpful to plan experimental design to quantify the missing heritability, and to estimate the genetic overlap between traits (diseases) especially when it is unfeasible to have the traits (diseases) measured on the same individuals.

estimating the genetic correlation between any two diseases using completely independent case-control data. Other methods to estimate genetic parameters from individual-level or summary GWAS data have also been reported [8–10].

We previously named the SNP-based method mentioned GREML [11], as a complement to GBLUP [12] where variance components are assumed known, and have been implemented them in the software tool GCTA [13]. One outstanding question is the statistical power of detecting genetic variation using the population-based estimation method, for example how many samples are required to achieve estimates that are sufficiently accurate to detect genetic (co)variance of complex traits. In this paper, we derive the sampling variance of the estimate of genetic (co)variance by analytical derivations and verify our derivations by simulations under a range of scenarios. We also provide an online tool for power calculation.

Methods and Results

Univariate analysis

The methods of using SNP data to estimate genetic variance in unrelated individuals have been detailed elsewhere [3,13]. In brief, given GWAS data, we can model the phenotype as

$$\mathbf{y} = \mathbf{g} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is an $N \times 1$ vector of phenotypes with N being the sample size, $\mathbf{g}$ is an $N \times 1$ vector with each of its elements being the total genetic effect of an individual captured by all SNPs, and $\mathbf{e}$ is an $N \times 1$ vector of residuals. We have $\mathbf{g} \sim N(0, \sigma_G^2 \mathbf{A})$ and $\mathbf{e} \sim N(0, \sigma_e^2 \mathbf{I})$, where σ_G^2 is the genetic variance captured by all SNPs, $\mathbf{A}$ is the genetic relationship matrix (GRM) estimated from SNPs [3], σ_e^2 is the residuals variance and $\mathbf{I}$ is an identity matrix. The genetic relationships, also known as 'genomic relationships' or 'genetic similarity relationships', are referenced to the current population, and so can be negative as they are distributed about a mean of zero. Equation (1) is a typical mixed linear model with $\text{var}(\mathbf{y}) = \sigma_G^2 \mathbf{A} + \sigma_e^2 \mathbf{I}$, in which the variance components can be estimated using a restricted maximum likelihood (REML) approach [13,14]. The proportion of variance explained by all SNPs (SNP heritability) is defined as $h_G^2 = \sigma_G^2 / (\sigma_G^2 + \sigma_e^2)$.

For power calculation, we need to know the sampling variance of the estimate of σ_G^2, i.e. $\text{var}(\hat{\sigma}_G^2)$. In practice, the asymptotic sampling variance (standard error squared) of a variance component is calculated from a diagonal element of the inverse of the information matrix in maximum likelihood analysis [15–18]. Each element of the information matrix, however, comprises complex forms of matrix algebra including a matrix inverse. It is therefore unfeasible to derive $\text{var}(\hat{\sigma}_G^2)$ directly from the inverse of the information matrix. We show below an equivalent approach to obtain $\text{var}(\hat{\sigma}_G^2)$ under the simple regression framework.

For unrelated individuals, where the phenotypic correlation between individuals is small, mixed linear model analysis using the REML approach is asymptotically equivalent to simple regression analysis of pairwise phenotypic similarity/difference on pairwise genetic similarity, as measured by identity-by-descent (IBD) or identity-by-state (IBS) at genome-wide markers [17–20]. Under such circumstance, a regression of the cross-product of the phenotypes is equivalent to using both the squared difference and squared sum of the pairwise phenotypes, and using the cross-product is equivalent to using maximum likelihood [19]. The model for the regression-based analysis can be written as

$$z_{ij} = \mu + bA_{ij} + \varepsilon_{ij} \tag{2}$$

where $z_{ij} = y_i y_j$ with y_i and y_j being the phenotypes of individuals i and j $(i > j)$, A_{ij} is the ij-th element of the GRM $\mathbf{A}$, and ε_{ij} is the residual of this regression. There are $n = N(N-1)/2 \approx N^2/2$ observations (contrasts) in the regression. The regression coefficient b is equivalent to σ_G^2 because

$$\begin{aligned} b &= \text{cov}(A_{ij}, y_i y_j)/\text{var}(A_{ij}) = \text{cov}(A_{ij}, g_i g_j)/\text{var}(A_{ij}) \\ &= \text{E}(A_{ij} g_i g_j)/\text{var}(A_{ij}) = \sigma_G^2 \text{E}(A_{ij}^2)/\text{var}(A_{ij}) \\ &= \sigma_G^2 \end{aligned}$$

In such a simple regression, the sampling variance of the estimate of the regression coefficient is

$$\text{var}(\hat{\sigma}_G^2) = \text{var}(\hat{b}) = \text{var}(\varepsilon_{ij}) / [n\, \text{var}(A_{ij})] \tag{3}$$

If the samples are unrelated and the phenotypes have been standardized with mean of 0 and variance of 1, then $\text{E}(z_{ij}) = 0$ and $\text{var}(z_{ij}) = 1$. Since $\text{var}(A_{ij})$ is small, there is hardly any variance in z_{ij} that can be explained by A_{ij} so that $\text{var}(\varepsilon_{ij}) \approx \text{var}(z_{ij}) = 1$. We therefore have

$$\text{var}(\hat{\sigma}_G^2) \approx 2/[N^2\, \text{var}(A_{ij})] \tag{4}$$

Under circumstances when $\text{var}(A_{ij})$ is large, for example when the GRM is calculated from pedigree data, a substantial proportion of variance in z_{ij} could be explained by A_{ij}, so that $\text{var}(\varepsilon_{ij})$ will be smaller than $\text{var}(z_{ij})$ and the sampling variance of estimate of genetic variance will be reduced accordingly. In general, $\text{var}(A_{ij})$ and the residual variance in equation (2) depend on the number of SNP that are used to calculate the GRM and their correlation structure. Although $\text{var}(A_{ij})$ can be calculated empirically from the data, theoretical work suggest it is approximately 2×10^{-5} for genome-wide coverage of common SNPs in human populations

[21]. Since the phenotypic variance is usually estimated with very high precision,

$$\mathrm{var}(\hat{h}^2_{\mathrm{G}}) \approx \mathrm{var}(\hat{\sigma}^2_{\mathrm{G}}) \approx 2/[N^2\,\mathrm{var}(A_{ij})] = 10^5/N^2 \qquad (5)$$

This suggests that the standard error (SE) of $\hat{h}^2_{\mathrm{G}}$ depends only on sample size, and is approximately $316/N$. We show by simulations based on real genotype data (Text S1) that this approximation is very accurate (Figure 1 and Table S1). The SEs calculated from the approximation theory are also highly consistent with those reported from our previous studies for human height and body mass index (BMI). For example, the reported SE of $\hat{h}^2_{\mathrm{G}}$ for height was 0.083 using 3925 unrelated samples [3] and 0.029 for both height and BMI, irrespective to $\hat{h}^2_{\mathrm{G}}$, using 11586 unrelated samples [22], and the SE calculated from the approximation theory is 0.081 for $N = 3925$ and 0.027 for $N = 11586$.

Bivariate analysis (traits measured on the same individuals)

For a bivariate analysis where the two traits are measured on the same individuals, the mixed linear model can be written as [6]

$$\mathbf{y}_1 = \mathbf{g}_1 + \mathbf{e}_1 \text{ for trait \#1 and } \mathbf{y}_2 = \mathbf{g}_2 + \mathbf{e}_2 \text{ for trait \#2} \qquad (6)$$

where $\mathbf{y}_1$ and $\mathbf{y}_2$ are $N \times 1$ vector of phenotypes, $\mathbf{g}_1$ and $\mathbf{g}_2$ are $N \times 1$ vectors of genetic effects with $\mathbf{g}_1 \sim N(0, \sigma^2_{\mathrm{G1}}\mathbf{A})$ and $\mathrm{g}_2 \sim N(0, \sigma^2_{\mathrm{G2}}\mathbf{A})$, $\mathbf{e}_1$ and $\mathbf{e}_2$ are $N \times 1$ vectors of residuals with $\mathbf{e}_1 \sim N(0, \sigma^2_{\mathrm{e1}}\mathbf{I})$ and $\mathrm{e}_2 \sim N(0, \sigma^2_{\mathrm{e2}}\mathbf{I})$, and N is the sample size. The variance covariance matrix is

$$\mathrm{var}\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \sigma^2_{\mathrm{G1}}\mathbf{A} + \sigma^2_{\mathrm{e1}}\mathbf{I} & \sigma_{\mathrm{G1G2}}\mathbf{A} + \sigma_{\mathrm{e1e2}}\mathbf{I} \\ \sigma_{\mathrm{G1G2}}\mathbf{A} + \sigma_{\mathrm{e1e2}}\mathbf{I} & \sigma^2_{\mathrm{G2}}\mathbf{A} + \sigma^2_{\mathrm{e2}}\mathbf{I} \end{bmatrix}$$

where σ_{G1G2} is the genetic covariance between the two traits and σ_{e1e2} is the residual covariance. The genetic variance and covariance components can also be estimated using REML [6]. The genetic correlation is estimated as $\hat{r}_{\mathrm{G}} = \hat{\sigma}_{\mathrm{G1G2}}/\sqrt{\hat{\sigma}^2_{\mathrm{G1}}\hat{\sigma}^2_{\mathrm{G2}}}$. Since $\hat{r}_{\mathrm{G}}$ is a non-linear function of $\hat{\sigma}_{\mathrm{G1G2}}$, $\hat{\sigma}^2_{\mathrm{G1}}$ and $\hat{\sigma}^2_{\mathrm{G2}}$, there is no explicit derivation for $\mathrm{var}(\hat{\sigma}_{\mathrm{G1G2}}/\sqrt{\hat{\sigma}^2_{\mathrm{G1}}\hat{\sigma}^2_{\mathrm{G2}}})$. Reeve (1955) and Robertson (1959) provided an approximation of $\mathrm{var}(\hat{r}_{\mathrm{G}})$ in the context of balanced pedigree design as $\dfrac{(1-r^2_{\mathrm{G}})^2\sqrt{\mathrm{var}(\hat{h}^2_{\mathrm{G1}})\mathrm{var}(\hat{h}^2_{\mathrm{G2}})}}{2h^2_{\mathrm{G1}}h^2_{\mathrm{G2}}}$ [23,24] and Koots and Gibson (1996) proposed a modified version as $\dfrac{(1-r^2_{\mathrm{P}})^2\sqrt{\mathrm{var}(\hat{h}^2_{\mathrm{G1}})\mathrm{var}(\hat{h}^2_{\mathrm{G2}})}}{2h^2_{\mathrm{G1}}h^2_{\mathrm{G2}}}$ [25], where r_{P} is the phenotypic correlation between the traits. However, both approximations have an unsatisfying property that $\mathrm{var}(\hat{r}_{\mathrm{G}})$ will approach 0 if r_{G} or r_{P} is close to 1. We derived an approximation, which does not have this problem (Text S2), i.e.

$$\mathrm{var}(\hat{r}_{\mathrm{G}}) \approx \frac{(1-r_{\mathrm{G}}r_{\mathrm{P}})^2 + (r_{\mathrm{G}} - r_{\mathrm{P}})^2}{h^2_{\mathrm{G1}}h^2_{\mathrm{G2}}N^2\,\mathrm{var}(A_{ij})} \qquad (7)$$

As described above, $\mathrm{var}(A_{ij}) \approx 2 \times 10^{-5}$ for a GRM estimated from common SNPs in unrelated individuals in human populations,

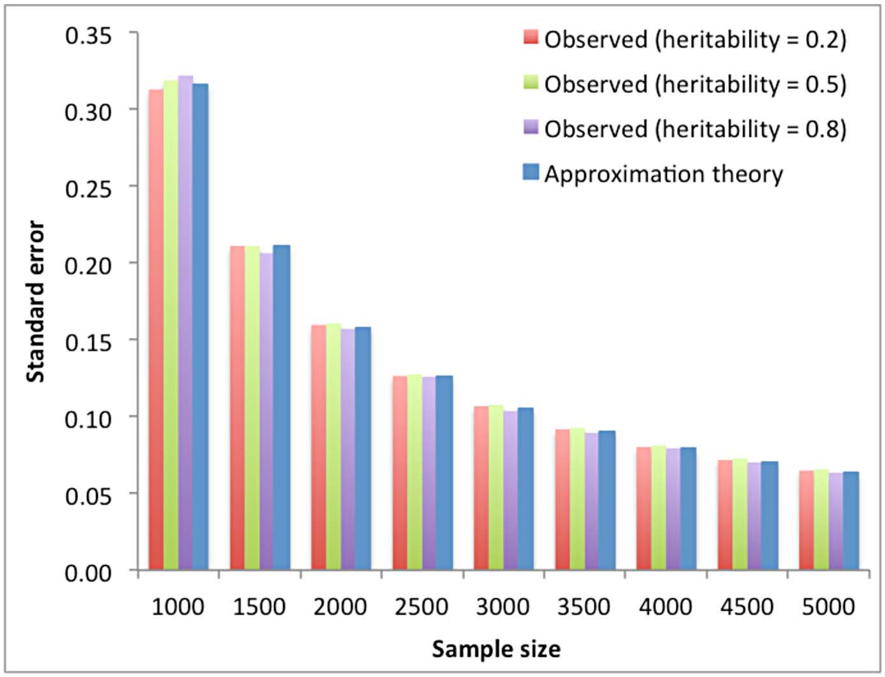

Figure 1. Standard error of the estimate of variance explained by all SNPs vs. sample size. The first three columns are the averaged standard error observed from 100 simulations under three heritability levels. The last column is the predicted standard error from our approximation theory. The plotted data can be found in Table S1.

doi:10.1371/journal.pgen.1004269.g001

Table 1. Standard error of the estimate of genetic correlation from a bivariate analysis of two traits measured on the same or different samples using genome-wide SNP data.

	Same sample					Different samples					
r_G	*N*	Est.	SE (Obs.)	s.e.m.	SE (Approx.)	N_1	N_2	Est.	SE (Obs.)	s.e.m.	SE (Approx.)
0.0	4000	0.00	0.128	0.0024	0.114	1000	3000	0.04	0.288	0.0126	0.264
0.0	6000	0.00	0.085	0.0009	0.076	2000	4000	0.00	0.191	0.0062	0.161
0.0	8000	−0.01	0.065	0.0006	0.057	3000	5000	−0.01	0.129	0.0021	0.118
0.0	10000	0.00	0.053	0.0004	0.046	4000	6000	−0.01	0.103	0.0014	0.093
0.4	4000	0.42	0.112	0.0019	0.108	1000	3000	0.32	0.295	0.0124	0.309
0.4	6000	0.39	0.076	0.0009	0.072	2000	4000	0.39	0.230	0.0141	0.182
0.4	8000	0.38	0.057	0.0006	0.054	3000	5000	0.37	0.136	0.0036	0.131
0.4	10000	0.39	0.046	0.0005	0.043	4000	6000	0.38	0.107	0.0015	0.103
0.8	4000	0.80	0.081	0.0024	0.091	1000	3000	0.62	0.417	0.0274	0.418
0.8	6000	0.80	0.056	0.0017	0.061	2000	4000	0.83	0.248	0.0127	0.232
0.8	8000	0.80	0.042	0.0012	0.046	3000	5000	0.86	0.198	0.0069	0.164
0.8	10000	0.79	0.036	0.0010	0.036	4000	6000	0.83	0.133	0.0034	0.127

Same sample: two traits are measured on the same set of samples. Different sample: two traits are measured on the different sets of samples. r_G: parameter of genetic correlation (i.e. proportion of simulated causal variants shared between the two traits). Est.: estimate of genetic correlation from 100 simulations. SE(Obs.): mean of the observed standard errors from 100 simulations. s.e.m.: standard error of the mean (i.e. SE(Obs.)). SE(Approx.): standard error calculated from our approximation theory.

doi:10.1371/journal.pgen.1004269.t001

therefore $\mathrm{SE}(\hat{r}_G) \approx \frac{224}{N}\sqrt{\frac{(1-r_G r_P)^2+(r_G-r_P)^2}{h^2_{G1}h^2_{G2}}}$. When $r_G = r_P = 0$, i.e. the traits are completely independent, $\mathrm{SE}(\hat{r}_G) \approx \frac{224}{N\sqrt{h^2_{G1}h^2_{G2}}}$. We tested equation (7) by simulations based on real genotype data (Text S1). The simulation results suggest that the approximation is reasonably accurate (Table 1). For real data analysis, we previously estimated the genetic correlation between intelligence at age 11 years and in old age of 0.62 with a SE of 0.23 using 1729 samples [5], consistent with the predicted SE of 0.22 from the approximation theory.

Bivariate analysis (traits measured on different sets of individuals)

For a bivariate analysis where the two traits are measured on different sets of individuals, e.g. height in males and blood pressure in females, the variance-covariance matrix is [6]

$$\mathrm{var}\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\sigma^2_{G1}\mathbf{A}_1+\sigma^2_{e1}\mathbf{I}_1 & \sigma_{G1G2}\mathbf{A}_{12}\\ \sigma_{G1G2}\mathbf{A}_{12} & \sigma^2_{G2}\mathbf{A}_2+\sigma^2_{e2}\mathbf{I}_2\end{bmatrix}$$

where $\mathbf{y}_1$ is an $N_1 \times 1$ vector of phenotypes in sample set #1 (e.g. males), and $\mathbf{y}_2$ is an $N_2 \times 1$ vector of phenotypes for in sample set #2 (e.g. females), with N_1 and N_2 being the sample sizes of the two sets. $\mathbf{A}_1$ is an $N_1 \times N_1$ GRM for individuals in sample set #1, $\mathbf{A}_2$ is an $N_2 \times N_2$ GRM in sample set #2 and $\mathbf{A}_{12}$ is an $N_1 \times N_2$ GRM between the two sets of samples. σ^2_{G1} and σ^2_{G2} are the genetic variance for the two traits. σ^2_{e1} and σ^2_{e2} are the residual variances with the corresponding identify matrix $\mathbf{I}_1$ and $\mathbf{I}_2$. σ_{G1G2} is the genetic covariance between traits. Since the two traits are measured on different sets of samples, the residual covariance is ignored because it is assumed that there is no covariance between the unrelated individuals apart from that caused by genetic factors. The genetic correlation is also estimated as $\hat{r}_G = \hat{\sigma}_{G1G2}/\sqrt{\hat{\sigma}^2_{G1}\hat{\sigma}^2_{G2}}$, however, the sampling variance of $\hat{r}_G$ is different from that described above. Since the traits are measured in different sets of samples, $\mathrm{cov}(\hat{\sigma}^2_{G1},\hat{\sigma}^2_{G2}) = \mathrm{cov}(\hat{\sigma}^2_{G1},\hat{\sigma}_{G1G2}) = \mathrm{cov}(\hat{\sigma}^2_{G2},\hat{\sigma}_{G1G2}) = 0$. Therefore, from a second order Taylor series approximation [15]

$$\mathrm{var}(\hat{r}_G) \approx r^2_G\left[\frac{\mathrm{var}(\hat{\sigma}^2_{G1})}{4\sigma^4_{G1}} + \frac{\mathrm{var}(\hat{\sigma}^2_{G2})}{4\sigma^4_{G2}} + \frac{\mathrm{var}(\hat{\sigma}_{G1G2})}{\sigma^2_{G1G2}}\right] \quad (8)$$

This approximation involves the sampling variance of $\hat{\sigma}_{G1G2}$. We show below an equivalent approach to obtain $\mathrm{var}(\hat{\sigma}_{G1G2})$.

Analogous to the univariate analysis, estimation of genetic covariance by a bivariate mixed linear model analysis is asymptotically equivalent to the following linear regression model

$$z_{12(ij)} = \mu + bA_{12(ij)} + \varepsilon_{ij} \quad (9)$$

where $z_{12(ij)} = y_{1(i)}y_{2(j)}$ i.e. the product of phenotypes between the i-th individual in set #1 and the j-th individual in set #2, and $A_{12(ij)}$ is the ij-th element of the GRM $\mathbf{A}_{12}$, i.e. the genetic relationship between the i-th individual in set 1 and the j-th individual in sample set #2. The regression coefficient is equivalent to genetic covariance between the two traits because

$$\begin{aligned}b &= \mathrm{cov}(A_{12(ij)},y_{1(i)}y_{2(j)})/\mathrm{var}(A_{12(ij)}) = \mathrm{cov}(A_{12(ij)},g_{1(i)}g_{2(j)})/\mathrm{var}(A_{12(ij)})\\ &= \mathrm{E}(A_{12(ij)}g_{1(i)}g_{2(j)})/\mathrm{var}(A_{12(ij)}) = \sigma_{G1G2}\mathrm{E}(A^2_{12(ij)})/\mathrm{var}(A_{12(ij)})\\ &= \sigma_{G1G2}\end{aligned}$$

If the two sample sets are independent and phenotypes for both traits have been standardized with mean of 0 and variance of 1, then $\mathrm{E}(z_{12(ij)}) = 0$ and $\mathrm{var}(z_{12(ij)}) = 1$. Since $\mathrm{var}(A_{12(ij)})$ is small, $\mathrm{var}(\varepsilon_{ij}) \approx \mathrm{var}(z_{12(ij)}) = 1$. We then have $\mathrm{var}(\hat{\sigma}_{G1G2}) \approx \mathrm{var}(\varepsilon_{ij})/[N_1N_2\,\mathrm{var}(A_{12(ij)})] \approx 1/[N_1N_2\,\mathrm{var}(A_{12(ij)})]$.

We know from the derivations above that $\mathrm{var}(\hat{\sigma}^2_{G1}) \approx 2/[N_1^2\,\mathrm{var}(A_{1(ij)})]$ and $\mathrm{var}(\hat{\sigma}^2_{G2}) \approx 2/[N_2^2\,\mathrm{var}(A_{2(ij)})]$. For unrelated individuals sampled from the same population, $\mathrm{var}(A_{1(ij)}) = \mathrm{var}(A_{2(ij)}) = \mathrm{var}(A_{12(ij)}) = \mathrm{var}(A_{ij})$, we therefore get

$$\mathrm{var}(\hat{r}_G) \approx \frac{r^2_G(N_1^2h^4_{G1}+N_2^2h^4_{G2}) + 2h^2_{G1}h^2_{G2}N_1N_2}{2h^4_{G1}h^4_{G2}N_1^2N_2^2\,\mathrm{var}(A_{ij})} \quad (10)$$

This was also tested by simulations (Text S1) and the approximated standard errors were highly consistent with those observed from simulations, especially when sample size was large (Table 1). When $r_G = r_P = 0$, i.e. two traits are completely independent, $\mathrm{var}(\hat{r}_G) \approx 1/[h^2_{G1}h^2_{G2}N^2\,\mathrm{var}(A_{ij})]$ for traits measured on the same sample, and $\mathrm{var}(\hat{r}_G) \approx 1/[h^2_{G1}h^2_{G2}N_1N_2\,\mathrm{var}(A_{ij})]$ for traits measured on different samples. Therefore, for independent traits, the ratio of sampling variance of genetic correlation between the two traits measured on the same sample to that on different samples is simply N_1N_2/N^2.

Case-control studies

For case-control studies, the proportion of variance in case-control status (0 or 1) that is explained by all SNPs on the observed scale (h^2_O) can be estimated using a linear model [4]. Therefore, the same approximations to the sampling variance of genetic variance and genetic correlation for quantitative traits can be applied directly to case-control studies. As shown in equation (5), in a univariate analysis, the sampling variance of SNP-based heritability depends only on sample size and variance in genetic relatedness, independent of the properties of the phenotype, so that $\mathrm{var}(\hat{h}^2_O)$ is also approximately $2/[N^2\,\mathrm{var}(A_{ij})]$ in a case-control study with N being the total number of cases and controls. We show in Table 2 that the observed standard errors of the estimates of h^2_O from published studies are highly consistent with those predicted from our approximation theory.

To calculate power, however, we would need to specify h^2_O, which is a parameter with non-intuitive properties, and depends on the prevalence of the disease in the population (K), the proportion of variance in disease liability that is captured by the SNPs at population level, and the proportional of cases in the sample (v). For this reason we define h^2_L as the variance explained by all SNPs at the population level on the unobserved underlying scale of disease liability, and use a linear transformation to transform h^2_O to h^2_L on the liability scale [4], i.e. $h^2_L = ch^2_O$ and $\mathrm{var}(\hat{h}^2_L) = c^2\,\mathrm{var}(\hat{h}^2_O)$ with $c = (1-K)^2/[v(1-v)i^2]$. We then get

$$\mathrm{var}(\hat{h}^2_L) = \frac{2(1-K)^4(N_{\mathrm{case}}+N_{\mathrm{control}})^2}{N^2_{\mathrm{case}}N^2_{\mathrm{control}}i^4\,\mathrm{var}(A_{ij})} \quad (11)$$

where N_{case} is the number of cases, N_{control} is the number of

Table 2. Standard errors of the estimates of variance explained by all SNPs on the observed scale (h^2_O) from published analyses of case-control studies for a number of diseases vs. those predicted from the approximation theory.

Disease	N_{cases}	$N_{controls}$	*Prevalence*	h^2_O	**SE(Obs.)**	**SE(Approx.)**
Multiple sclerosis [34]	1604	1953	0.001	0.851	0.088	0.089
Alzheimer's disease [34]	3290	3849	0.020	0.364	0.049	0.044
Endometriosis [34]	3154	6981	0.080	0.231	0.036	0.031
Schizophrenia [7]	9087	12171	0.010	0.410	0.015	0.015
Bipolar disorder [7]	6704	9031	0.010	0.441	0.021	0.020
MDD [7]	9041	9381	0.150	0.177	0.017	0.017
ASD [7]	3303	3428	0.010	0.310	0.046	0.047
ADHD [7]	4163	12040	0.050	0.253	0.020	0.020

N_{cases}: number of cases. $N_{controls}$: number of controls. SE(Obs.): reported standard error of the estimate of h^2_O from real data analysis. SE(Approx.): standard error of h^2_O calculated from our approximation theory. MDD: major depression disorder. ASD: autism spectrum disorders. ADHD: attention-deficit/hyperactivity disorder.
doi:10.1371/journal.pgen.1004269.t002

controls, and i is the selection intensity which is a function of K [4]. We illustrate in Figure 2 the dependency of the SE of h^2_L on disease prevalence (K) and proportion of cases in the sample (v) due to the transformation.

As shown in equation (10), in a bivariate analysis where traits are measured on different sets of samples, the sampling variance of genetic correlation depends on sample sizes, trait heritabilities and the genetic correlation parameter, which is also independent of the properties of the phenotypes. Therefore, in a bivariate analysis of two independent case-control disease studies,

$$\mathrm{var}(\hat{r}_G) \approx \frac{r^2_G(N^2_1 h^4_{O1} + N^2_2 h^4_{O2}) + 2h^2_{O1} h^2_{O2} N_1 N_2}{2h^4_{O1} h^4_{O2} N^2_1 N^2_2 \, \mathrm{var}(A_{ij})} \qquad (12)$$

where N_1 and N_2 are the total numbers of cases and controls of the two case-control studies, respectively. This also applies to a bivariate analysis of a quantitative trait and a cases-control disease study on different sets of samples, i.e.

$$\mathrm{var}(\hat{r}_G) \approx \frac{r^2_G(N^2_1 h^4_O + N^2_2 h^4_G) + 2h^2_O h^2_G N_1 N_2}{2h^4_O h^4_G N^2_1 N^2_2 \, \mathrm{var}(A_{ij})} \qquad (13)$$

These two equations can also be expressed with respect to h^2_L, given $h^2_O = h^2_L / c$ (see above). We show in Table 3 that the reported SEs of r_G from bivariate analyses of psychiatric diseases are also highly in line with the predicted SEs from the approximation theory.

Statistical power

Statistical power is calculated from the population value of the parameter and its sampling variance, which was derived above. If the parameter is θ, where θ is either the proportion of phenotypic variance captured by SNPs (h^2_G) in the univariate case or the

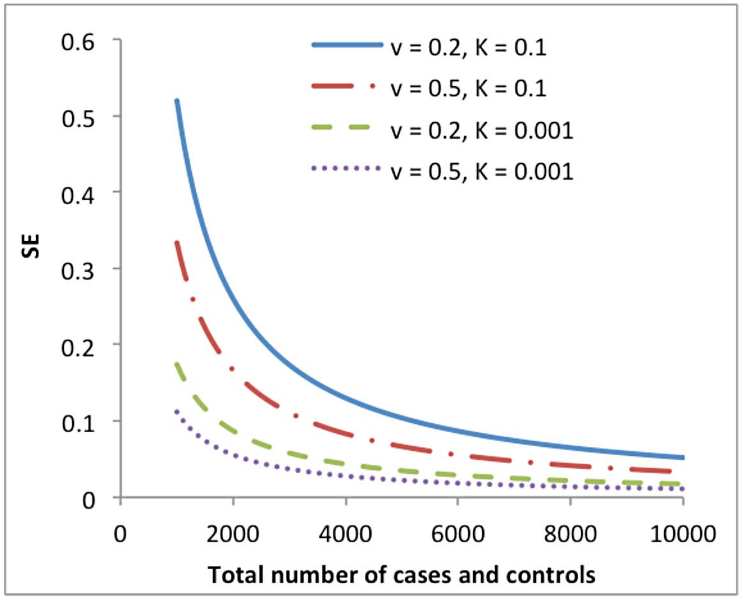

Figure 2. Standard error (SE) of the estimate of variance explained by all SNPs on the underlying scale (h^2_L) from a univariate analysis of a case-control study vs. total number of cases and controls (sample size). The SE is predicted from the approximation theory given different levels of disease prevalence (K) and proportion of cases in the sample (v).
doi:10.1371/journal.pgen.1004269.g002

genetic correlation (r_G) in the bivariate case, then $\hat{\theta}^2/\mathrm{var}(\hat{\theta}^2)$ is asymptotically distributed as a non-central χ^2 with 1 degree of freedom and non-centrality parameter (NCP) of $\lambda = \theta^2/\mathrm{var}(\hat{\theta}^2)$. Given λ and the type-I error rate of α, statistical power is the probability that a non-central χ^2 variable is larger than the central χ^2 threshold that is determined by α. We show in Figure 3 the statistical power based on the sampling variance from our approximation theories to detect h_G^2 in a univariate case and r_G in a bivariate case under a range of scenarios. For example, for a quantitative trait, approximately 8900, 4500, 3000 and 2300 independent individuals are required to detect h_G^2 of 0.1, 0.2, 0.3 and 0.4 with >80% power at a type-I error rate of 0.05, respectively. For two quantitative traits measured on the same sample, approximately 7000, 4700, 2500 and 1600 independent individuals are required to detect r_G of 0.2, 0.4, 0.6 and 0.8 with > 80% at a type-I error rate of 0.05, respectively.

Online tool

We have also developed an online calculator (GCTA Power Calculator, http://spark.rstudio.com/ctgg/gctaPower), as part of the GCTA [13] software package (http://ctgg.qbi.uq.edu.au/software/gcta), using R-Shiny (http://shiny.rstudio.org) to calculate the SE of genetic variance or genetic correlation and statistical power given user-defined parameters.

Discussion

We have derived the approximate sampling variance of the estimate of variance explained by all common SNPs (h_G^2) for a quantitative trait or case-control study of a disease, and genetic correlation (r_G) between two quantitative traits, between two diseases, or between a quantitative trait and a disease, using genome-wide SNP data in unrelated individuals. We believe that the derivations and the online tool will be helpful for researchers to determine how many samples are required to detect h_G^2 (or r_G) and to estimate h_G^2 (or r_G) with adequate precision before collecting the genotype data.

The sampling variance of $\hat{h}_G^2$ for a complex trait is inversely proportional to sample size (N) and the variance in SNP-based genetic relatedness ($\mathrm{var}(A_{ij})$), and independent of h_G^2. The sampling variance of $\hat{r}_G$ between two complex traits is a function of r_G, N of the two samples, h_G^2 of the two traits and $\mathrm{var}(A_{ij})$ when the traits are measured on different samples, and further depends on the phenotypic correlation (r_P) when traits are measured on the same samples. All the approximation theories apply to case-control studies of diseases since the case-control data can be analysed using a linear model on the observed 0–1 scale. The sampling variance for the estimate on the observed scale can then be transformed to that on the underlying liability scale using well-established theory. The standard errors (square root of sampling variance) of either $\hat{h}_G^2$ or $\hat{r}_G$ observed in published studies were all highly consistently with those predicted from our approximation theories, which were also confirmed by simulations based on real genotype data.

Analytical expressions for the sampling variance of the estimates of genetic (co)variance from pedigree analyses have been around for over 50 years [17,26], and statistical power can be derived from these by using the sampling variance and population value of the parameter. However, these expressions are typically for specific structured pedigrees, such as fullsib or halfsib families or twin pairs. There are to our knowledge no simple approximations for general pedigrees, because the inverse of the variance-covariance matrix

Table 3. Standard errors of the estimates of genetic correlations from published bivariate analyses of case-control studies for psychiatric diseases [7] vs. those predicted from the approximation theory.

Disease	K	N_{cases}	$N_{controls}$	h_O^2	Disease	K	N_{cases}	$N_{controls}$	h_O^2	r_G	SE (Obs.)	SE (Approx.)
SCZ	0.01	9032	7980	0.40	**BPD**	0.01	6664	5258	0.39	0.68	0.044	0.049
SCZ	0.01	9051	10385	0.38	**MDD**	0.15	8998	7823	0.16	0.43	0.055	0.057
SCZ	0.01	9111	12146	0.41	**ASD**	0.01	3226	3308	0.29	0.16	0.059	0.057
SCZ	0.01	9013	10115	0.42	**ADHD**	0.05	4108	9936	0.22	0.08	0.046	0.045
BPD	0.01	6665	7408	0.42	**MDD**	0.15	8997	7680	0.17	0.47	0.061	0.063
BPD	0.01	6704	9030	0.43	**ASD**	0.01	3207	3294	0.31	0.04	0.065	0.061
BPD	0.01	6656	7041	0.38	**ADHD**	0.05	4099	9873	0.25	0.05	0.053	0.052
MDD	0.15	9031	9370	0.17	**ASD**	0.01	3239	3331	0.31	0.05	0.089	0.090
MDD	0.15	8936	8668	0.16	**ADHD**	0.05	4098	11233	0.24	0.32	0.071	0.073
ASD	0.01	3156	3254	0.27	**ADHD**	0.05	4181	12022	0.23	−0.13	0.087	0.090

SCZ: schizophrenia. BPD: bipolar disorder. MDD: major depression disorder. ASD: autism spectrum disorders. ADHD: attention-deficit/hyperactivity disorder. N_{cases}: number of cases. $N_{controls}$: number of controls. K: disease prevalence. h_O^2: estimate of variance explained by all SNPs on the observed scale, which was calculated from the reported h_L^2 and disease prevalence in Supplementary Table 1 of Lee et al. [7]. r_G: genetic correlation. SE(Obs.): reported standard error of the estimate of r_G from real data analysis. SE(Approx.): standard error of r_G calculated from our approximation theory.

doi:10.1371/journal.pgen.1004269.t003

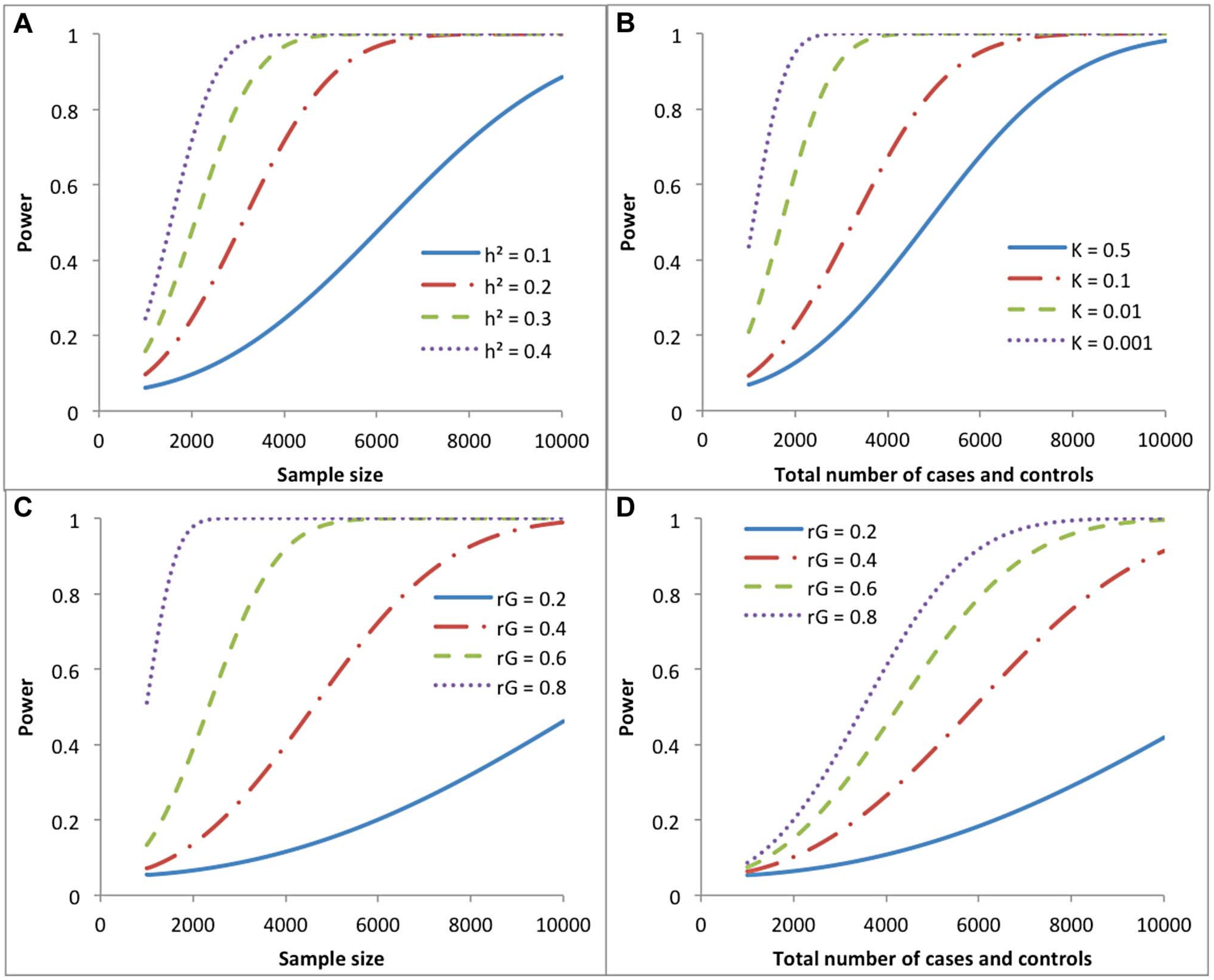

Figure 3. Statistical power of detecting genetic variance (correlation) under different study designs. a) Univarite analysis of a quantitative trait. **b**) Univariate analysis of a case-control study assuming equal number of cases and controls ($v=0.5$) and heritability of liability (h^2_L) of 0.2. **c**) Bivariate analysis of two quantitative traits measured on the same set of individuals, assuming heritability of 0.2 for both traits. **d**) Bivariate analysis of two case-control studies on independent sets of samples, assuming equal numbers of cases and controls for each disease, and equal sample size (total number of cases and controls), equal heritability of liability ($h^2_L=0.2$) and equal prevalence ($K=0.01$) for both diseases.
doi:10.1371/journal.pgen.1004269.g003

is required and this is conditional on the actual pedigree structure. The sampling variance of the estimated parameters in a general complex pedigree is usually derived post hoc after the analysis has been performed.

Methods for calculating the power of detecting quantitative trait loci (QTL) in family-based linkage studies have been investigated extensively in the past two decades [16–18,27]. These methods were developed to calculate the power of detecting a QTL but can be generalized for variance components estimation, e.g. estimating the genetic variance using pedigree information. The non-centrality parameter of the test-statistic from a maximum likelihood analysis of variance components is $\mathrm{NCP}=\mathrm{E}(2\ln L_1)-\mathrm{E}(2\ln L_0)=-\mathrm{E}(\ln|\mathbf{V}_1|)+\ln|\mathbf{V}_0|$, where L is the likelihood function, and $\mathbf{V}_0$ and $\mathbf{V}_1$ are the variance covariance matrix under the null and alternative hypotheses respectively [17,18]. For a specific balanced pedigree design, e.g. fullsibs or nuclear families, the determinant (or inverse) of the $\mathbf{V}$ matrix can be computed explicitly, so that the NCP can be calculated without making approximation [16,17]. For an arbitrary pedigree, $\ln|\mathbf{V}|$ can be calculated approximately using Taylor expansions given the variance in family relatedness [18,27]. Therefore, all these methods explicitly or implicitly require a known pedigree. When the correlations between relatives are small, the first order approximation of the NCP in Rijsdijk et al [18] can be written in our notations as $\mathrm{NCP}\approx\sum_{i>j}\mathrm{var}(A_{ij})h^4_{\mathrm{G}}\approx N^2h^4_{\mathrm{G}}\,\mathrm{var}(A_{ij})/2$, which is the same as we derived (i.e. $h^4_{\mathrm{G}}/\mathrm{var}(\hat{h}^2_{\mathrm{G}})$, see Equation (4) for $\mathrm{var}(\hat{h}^2_{\mathrm{G}})$), even though our deviations were based on least squares regression analysis in unrelated samples whereas the derivations in Rijsdijk et al [18] were based on maximum likelihood approach in family data. This approximation is reasonably accurate when correlations between relatives are small for a pedigree-based design, which is not an issue for a population-based design where the genetic relationships between unrelated samples are very small as demonstrated in Yang et al [3]. We show by simulations (Text S1) that for a univariate analysis the LRT statistics calculated based on REML are highly consistent with the chi-squared test-statistics calculated by the Wald test using the sampling variance either observed from the simulations or predicted from our approximation theory (Figure S1).

For a given population, a set of common SNPs and the method of calculating the genetic relationship matrix that we have used here, $\text{var}(A_{ij})$ is a fixed quantity because it depends only on effective population size of the human populations [28]. We used $\text{var}(A_{ij}) = 2 \times 10^{-5}$, which was calculated from theory based upon an effective population size of 10,000. Variance in genetic relatedness (and therefore power of detection) can decrease by including many rare SNPs in calculating the GRM because adding more rare SNPs increases the effective population size reflecting recent population expansion. The variance in relatedness can also increase by sampling closer relatives (see below for more discussion) or, for example, by creating a relationship matrix based upon haplotype information. Modifying the GRM can also affect the variance of the off-diagonal elements. For example by applying a weighting of SNPs depending on linkage disequilibrium the variance in the estimates of genetic relationships will decrease so that the sampling variance of the estimate of SNP-based heritability will be increased [29]. Although we derive the theory and show the results based on the SNPs on the whole genome, our approximation theories are also applicable in analyses using a subset of SNPs, e.g. SNPs from a single chromosome. In that case, $\text{var}(A_{ij})$ used in the approximation equations should be either observed empirically from data or derived from theory [28] based on the subset of SNPs.

If there are unknown related samples in the data (cryptic relatedness), $\hat{h}^2_{\text{G}}$ will possibly be inflated due to shared environment between close relatives and/or the effects of causal variants in LD with the SNPs but captured by family relatedness, and $\text{var}(\hat{h}^2_{\text{G}})$ will be deflated due to the increase of $\text{var}(A_{ij})$. In fact, the interpretation of $\hat{h}^2_{\text{G}}$ changes if there is a substantial proportion of close relatives in the data [30,31]. This, however, affects GWAS result in a similar way, where the SE of the estimate of a SNP effect from a single SNP analysis (e.g. linear regression for a quantitative trait and logistic regression for a case-control study) will be deflated, causing an inflation of the test-statistics GWAS (often called "genomic inflation" [32]). For the estimation of h^2_{G} using all common SNPs, to avoid possible confounding from shared environments and uncaptured causal variants, we suggested in Yang et al. (2010) a stringent threshold, i.e. 0.025, to remove cryptic relatedness from the data so that the estimate of h^2_{G} can be compared directly to the results from GWAS in response to the "missing heritability" problem [2]. In practice, observing a much smaller SE of $\hat{h}^2_{\text{G}}$ using all common SNPs than that predicted from theory is a caveat suggesting substantial cryptic relatedness remaining in the data.

Using the same experimental design of a sample of conventionally unrelated individuals, the experimenter can increase power by increasing sample size. Fortunately, power increases quadratically with sample size because every new sample is contrasted with all existing samples. The sampling variance of the estimate of the genetic correlation is generally much larger than that of the proportion of variance explained from a univariate analysis, consistent with the theory of the sampling variance of genetic correlations in pedigree designs [33].

Supporting Information

Figure S1 Likelihood ratio test (LRT) statistic vs. Chi-squared test-statistic in a univariate analysis.
(PDF)

Table S1 Standard error of the estimate of h^2_{G} (variance explained by all SNPs) observed from 100 simulations vs. that calculated from our approximation theory.
(PDF)

Text S1 Simulations.
(PDF)

Text S2 Sampling variance of genetic correlation.
(PDF)

Text S3 Acknowledgments to dbGaP data.
(PDF)

Acknowledgments

This study uses data obtained from dbGaP through accession numbers [phs000090] and [phs000091] (a full list of acknowledgments to the dbGaP data can be found in Text S3).

Author Contributions

Conceived and designed the experiments: PMV JY. Performed the experiments: JY PMV MEG. Analyzed the data: JY PMV. Contributed reagents/materials/analysis tools: NRW SHL AAEV GBC. Wrote the paper: PMV JY. Developed the online tool: GH JY.

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106: 9362–9367.
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.
3. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565–569.
4. Lee SH, Wray NR, Goddard ME, Visscher PM (2011) Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88: 294–305.
5. Deary IJ, Yang J, Davies G, Harris SE, Tenesa A, et al. (2012) Genetic contributions to stability and change in intelligence from childhood to old age. Nature 482: 212–215.
6. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics 28: 2540–2542.
7. Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, et al. (2013) Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet 45: 984–994.
8. So HC, Li M, Sham PC (2011) Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. Genet Epidemiol 35: 447–456.
9. Dudbridge F (2013) Power and predictive accuracy of polygenic risk scores. PLoS Genet 9: e1003348.
10. Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, et al. (2012) Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet 44: 483–489.
11. Benjamin DJ, Cesarini D, van der Loos MJ, Dawes CT, Koellinger PD, et al. (2012) The genetic architecture of economic and political preferences. Proc Natl Acad Sci U S A 109: 8026–8031.
12. Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.
13. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88: 76–82.
14. Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545–554.
15. Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits: Sunderland, MA: Sinauer Associates.
16. Williams JT, Blangero J (1999) Power of variance component linkage analysis to detect quantitative trait loci. Ann Hum Genet 63: 545–563.

17. Sham PC, Cherny SS, Purcell S, Hewitt JK (2000) Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. Am J Hum Genet 66: 1616–1630.
18. Rijsdijk FV, Hewitt JK, Sham PC (2001) Analytic power calculation for QTL linkage analysis of small pedigrees. Eur J Hum Genet 9: 335–340.
19. Sham PC, Purcell S (2001) Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. Am J Hum Genet 68: 1527–1532.
20. Visscher PM, Hopper JL (2001) Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data. Ann Hum Genet 65: 583–601.
21. Vinkhuyzen AA, Wray NR, Yang J, Goddard ME, Visscher PM (2013) Estimation and Partitioning of Heritability in Human Populations Using Whole Genome Analysis Methods. Annual Review of Genetics 47.
22. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, et al. (2011) Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet 43: 519–525.
23. Reeve ECR (1955) The variance of the genetic correlation coeffi- cient. Biometrics 11: 357–374.
24. Robertson A (1959) The sampling variance of the genetic correlation coefficient. Biometrics 15: 469–485.
25. Koots KR, Gibson JP (1996) Realized sampling variances of estimates of genetic parameters and the difference between genetic and phenotypic correlations. Genetics 143: 1409–1416.
26. Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics: England: Longman.
27. Chen WM, Abecasis GR (2006) Estimating the power of variance component linkage analysis in large pedigrees. Genet Epidemiol 30: 471–484.
28. Goddard ME (2009) Genomic selection: prediction of accuracy and maximisation of long term. Genetica 136: 245–257.
29. Speed D, Hemani G, Johnson MR, Balding DJ (2012) Improved heritability estimation from genome-wide SNPs. Am J Hum Genet 91: 1011–1021.
30. Visscher PM, Yang J, Goddard ME (2010) A Commentary on 'Common SNPs Explain a Large Proportion of the Heritability for Human Height' by Yang et al. (2010). Twin Res Hum Genet 13: 517–524.
31. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, et al. (2013) Author reply to A commentary on Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14: 894.
32. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
33. Visscher PM (1998) On the sampling variance of intraclass correlations and genetic correlations. Genetics 149: 1605–1614.
34. Lee SH, Harold D, Nyholt DR, Goddard ME, Zondervan KT, et al. (2013) Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. Hum Mol Genet 22: 832–841.