Evaluating the Heritability Explained by Known Susceptibility Variants: A Survey of Ten Complex Diseases

Hon-Cheong So,[1] Allen H.S. Gui,[1] Stacey S. Cherny,[1–3] and Pak C. Sham[1–3]*

[1] *Department of Psychiatry, University of Hong Kong, Hong Kong SAR, China*
[2] *Genome Research Centre, University of Hong Kong, Hong Kong SAR, China*
[3] *State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Hong Kong SAR, China*

Recently, an increasing number of susceptibility variants have been identified for complex diseases. At the same time, the concern of "missing heritability" has also emerged. There is however no unified way to assess the heritability explained by individual genetic variants for binary outcomes. A systemic and *quantitative* assessment of the degree of "missing heritability" for complex diseases is lacking. In this study, we measure the variance in liability explained by individual variants, which can be directly interpreted as the locus-specific heritability. The method is extended to deal with haplotypes, multi-allelic markers, multi-locus genotypes, and markers in linkage disequilibrium. Methods to estimate the standard error and confidence interval are proposed. To assess our current level of understanding of the genetic basis of complex diseases, we conducted a survey of 10 diseases, evaluating the total variance explained by the known variants. The diseases under evaluation included Alzheimer's disease, bipolar disorder, breast cancer, coronary artery disease, Crohn's disease, prostate cancer, schizophrenia, systemic lupus erythematosus (SLE), type 1 diabetes and type 2 diabetes. The median *total* variance explained across the 10 diseases was 9.81%, while the median variance explained per associated SNP was around 0.25%. Our results suggest that a substantial proportion of heritability remains unexplained for the diseases under study. Programs to implement the methodologies described in this paper are available at http://sites.google.com/site/honcheongso/software/varexp. *Genet. Epidemiol.* 35:310–317, 2011. © 2011 Wiley-Liss, Inc.

Key words: association study; variance explained; genetic architecture; liability threshold model

Additional Supporting Information may be found in the online version of this article.
Contract grant sponsor: Hong Kong Research Grants Council General Research Fund; Contract grant numbers: HKU 766906M; HKU 774707M; Contract grant sponsors: University of Hong Kong Strategic Research Theme of Genomics; Croucher Foundation Scholarship.

*Correspondence to: Pak C. Sham, Department of Psychiatry, 10/F Laboratory Block, LKS Faculty of Medicine, University of Hong Kong, Pokfulam, Hong Kong SAR, China. E-mail: pcsham@hku.hk

Received 19 July 2010; Revised 7 January 2011; Accepted 31 January 2011
Published online 2 March 2011 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.20579

INTRODUCTION

The last few years have seen a rapid rise in the number and scale of association studies. Particularly, genome-wide association studies (GWAS) have become increasingly popular [Manolio et al., 2008]. With increasing sample size and wider coverage of the genome, we have been able to identify many susceptibility loci for complex diseases. Since the heritabilities of many complex diseases have been estimated from twin and family studies, this raises the question of how much of the total genetic variance of a given complex disease the genes identified to date explain. For example, although large-scale GWAS have been used to study the genetic basis for height and at least forty loci have been identified, they only account for ~5% of the variance [Visscher, 2008]. GWAS's failure to explain a substantial portion of the genetic variance of complex disease has led to the question of where the "missing heritability" is located in the genome [Manolio et al., 2009]. However, although the case of "missing heritability" has been noted in the community, a systemic *quantitative* assessment of how much heritability is actually "missing" for complex diseases is lacking. Furthermore, there is no unified way to assess the heritability explained by individual genetic variants for binary outcomes.

In this study, we suggest the use of a multifactorial liability threshold model [Falconer, 1965] in assessing the contribution of individual variants to the total heritability. This model proposes a latent continuous liability, which is assumed to follow a normal distribution with mean 0 and variance 1. The model is theoretically justified by the central limit theorem, as we assume the overall liability is made up by many genetic variants and other risk factors with modest effects. The disease is assumed to be present in individuals whose liability exceeds a certain threshold (*T*), and is absent in all other individuals. As shown in Figure 1, the mean liability is assumed to differ among genotypes while the threshold is fixed. Genotypes carrying higher risk will have higher liabilities. We measure the *variance in liability explained* (Vg) by individual genetic variants. As heritabilities of diseases estimated from twin or family data are also based on the liability threshold model, variance in liability explained by a

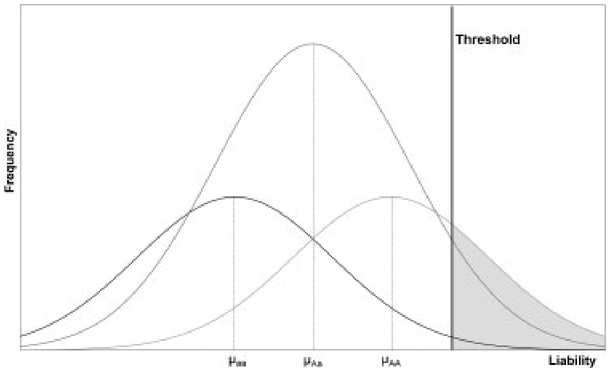

Fig. 1. Liability threshold model with different mean liability for each genotype. The shaded area represents the affected individuals.

variant can be directly interpreted as the locus-specific heritability.

While the measure of Vg has been used before to quantify the contribution of identified variants to the susceptibility of Crohn's disease in a recent GWAS [Barrett et al., 2008], the details of calculation were not described. The present paper is organized as follows. We first describe how to calculate the heritability explained by a single variant. We then extend the calculation of Vg to haplotypes, multi-allelic markers and multilocus genotypes, when interactions are present. We also devise a method for combining variances that takes into account the covariance structure of SNPs in LD. We provide four approaches for estimating the confidence interval (CI) and standard error (SE) of Vg. Finally we present a survey of 10 complex diseases, evaluating the total variance explained by the known common susceptibility variants for each trait.

METHODS

CALCULATION OF VARIANCE IN LIABILITY EXPLAINED

Consider the simple case with a single bi-allelic locus with risk allele A and protective allele a. The three genotypes have different mean liabilities for disease, while the threshold is assumed the same for the entire population (Fig. 1). We devise a method for calculating the variance in liability explained by the locus (Vg) from input information including the overall disease probability (K), the frequency of the high-risk allele (P_A), and the risk ratios (or relative risks) of the genotypes Aa and AA (RR_1 and RR_2), relative to genotype aa.

First, we calculate the frequencies of the three genotypes assuming Hardy-Weinberg disequilibrium. Then, the overall probability of disease (K) is related to risk ratios of genotypes Aa and AA by

$$K = P_{aa}f_{\mathrm{aa}} + P_{Aa}(RR_1 f_{aa}) + P_{AA}(RR_2 f_{aa})$$

where P are genotype frequencies assumed to be determined by allele frequencies via Hardy-Weinberg disequilibrium, and f_{aa} is the penetrance of genotype aa (i.e. probability of disease given genotype aa), which is given by

$$f_{aa} = \frac{K}{P_{aa} + P_{Aa}RR_1 + P_{AA}RR_2} \tag{1}$$

In practice, often only the odds ratios (OR) are available. The OR approximates the RR when the disease is rare. RR may be estimated from OR obtained in case-control studies, depending on the actual sampling scheme [So and Sham, 2010]. For example, if cases and controls are sampled from a cohort and the controls are disease-free at the end of the cohort, then the risk ratio (RR) is related to the OR as follows:

$$OR = \left(\frac{RRf_{aa}}{1 - RRf_{aa}}\right) \Big/ \left(\frac{f_{aa}}{1 - f_{aa}}\right)$$

which can be re-expressed as

$$RR = \frac{OR}{1 + f_{aa}(OR - 1)} \tag{2}$$

A simple procedure for calculating RR_1 and RR_2 from OR_1, OR_2 and K is then to set an initial value for f_{aa}, then iteratively use Equations (2) and (1) to alternately update (RR_1 and RR_2) and f_{aa} until convergence.

Under a liability-threshold model, the three genotypic classes have distinct liability distributions with different means but the same residual variance. Denoting the overall mean liability by μ_{all} and genotype-specific mean liabilities by μ_{AA}, μ_{Aa} and μ_{aa}, the variance explained by the locus may be written as

$$\text{Variance explained} = P_A^2(\mu_{AA} - \mu_{all})^2 + 2PAPa(\mu_{Aa} - \mu_{all})^2 + P_a^2(\mu_{aa} - \mu_{all})^2 \tag{3}$$

If we assume initially that the residual variance of each genotype is 1, then the genotype-specific mean liabilities are given by

$$\Phi^{-1}(1 - f_{aa}) = T - \mu_{aa}$$

$$\Phi^{-1}(1 - f_{Aa}) = T - \mu_{Aa}$$

$$\Phi^{-1}(1 - f_{AA}) = T - \mu_{AA}$$

where T is the liability threshold and Φ^{-1} is a function that returns the quantile of a normal distribution. We can further set one of the genotype-specific mean liabilities, say μ_{aa}, to zero. This will not affect the variance estimate. The equations can then be solved easily. The variance explained due to the locus is simply the variance of these three mean liabilities (weighted by the respective genotype frequencies), as in formula (3).

We denote the variance explained calculated from the above method by V^*. It is useful to evaluate a *standardized* estimate of V^*, denoted Vg, to allow for comparisons with heritability estimates. Vg is given by $V^*/(1+V^*)$. Note that while liability-threshold models with binary phenotypes typically model the liability distribution as a standard normal, the approach we adopted here ensures identification of Vg in an equivalent manner.

DEALING WITH IMPUTED GENOTYPES

GWAS and meta-analyses often involve imputed genotypes and the allelic counts may be fractional. Our methodology to evaluate Vg can still be applied in this scenario. Logistic regression is usually employed to analyze the association of imputed genotypes with the disease outcome. The OR of having one additional risk allele (i.e. AA versus Aa or Aa versus aa) can be obtained by exp(β) (β being the regression coefficient). The allele frequencies can also be easily derived from the fractional allelic counts. For example, if the sum of the fractional allelic counts of A equals 1,500 in 1,000 individuals, the frequency of A is 0.75. Vg can then be calculated given the allele frequencies and the corresponding OR.

EXTENSION TO MULTI-ALLELIC MARKERS, HAPLOTYPES AND MULTI-LOCUS GENOTYPES

Dealing with multiple alleles and/or markers involves a straightforward extension of the method above. Instead of considering only three distributions for a bi-allelic locus, we construct n distributions (each with a different mean) for a single marker with n genotypes or multiple markers with n multi-marker genotypic combinations. Similarly, in the case of gene-gene or gene-environment interactions, we may establish a distribution for each combination of different genotypes and/or environmental factors. For instance, for two bi-allelic markers there are nine genotypic combinations, each with a corresponding OR. In brief, we find the penetrance associated with each combination and set up equations in the form of

$$\Phi^{-1}(1-f) = T - \mu$$

We then find the variance of the mean liabilities (V^*) similar to above and apply the transformation $V^*/(1+V^*)$ to obtain the proportion of variance in liability explained by the set of genetic and/or environmental factors.

ACCOUNTING FOR LD BETWEEN MARKERS

For tightly linked loci, a marker may be associated with disease entirely or partially due to high LD with a causal variant. Simply adding up the explained variances of correlated SNPs would result in an inflated estimate of the overall variance. We have devised a method for combining variances that takes into account of the covariance structure of the SNPs. The method is based on regression theory. Consider the simplest case when there are only two SNPs and liability is assumed to be additive. The underlying liability (Y) may be written as a standardized regression equation

$$Y = \beta_1 x_1 + \beta_2 x_2$$

where x_1 and x_2 represent the two correlated SNPs, each coded additively (e.g. 0, 1 or 2 assuming they are bi-allelic) but transformed to have unit variance. The variance in liability explained by the SNPs then may be expressed as $\beta' \mathbf{r}_{XX}\beta$, where $\mathbf{r}_{XX}$ is a correlation matrix between the two (or more) SNPs, and β is a vector of regression coefficients given by $\beta = \mathbf{r}_{XX}^{-1}\mathbf{r}_{YX}$, where $\mathbf{r}_{XX}$ is the correlation matrix of the SNPs (which may be computed as the square root of the LD measure r^2, but retaining the same sign as D'), obtained from genotype data or from HapMap. The correlation between the liability Y and a particular marker x_i is simply the square root of the variance explained by x_i. In notation, $r_{YX} = \sqrt{\mathbf{Vg}}$, where **Vg** is a vector of the variance explained. After simplification, the variance explained by the correlated SNPs can be expressed as

$$Var(Y) = \mathbf{r}'_{YX}\mathbf{r}_{XX}^{-1}\mathbf{r}_{YX}$$

Note that the assumption that the liability is additive across risk alleles may be relaxed. Considering two loci (A, B), we may use two dummy variables to code for each locus. Hence

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

where the presence or absence of genotypes Aa, AA, Bb, BB are initially coded as 1 or 0, and subsequently transformed to have variance 1. The correlation between two genotypes (entries for $\mathbf{r}_{XX}$) is given by

$$r = \frac{\Pr(G1 \text{ and } G2) - P_{G1} \bullet P_{G2}}{\sqrt{P_{G1}(1-P_{G1})P_{G2}(1-P_{G2})}}$$

where $\Pr(G_1$ and $G_2)$ denotes the probability that both genotypes would appear together and P_{Gi} represents the frequency of the particular genotype. $\Pr(G_1$ and $G_2)$ can be derived from haplotype frequencies, which in turn depend on the LD between the markers. For instance, Pr(Aa and BB) $= 2 \times H_{AB} \times H_{aB,}$ where H denotes haplotype frequencies.

CI AND SE OF Vg

Simulation approach. The CI and SE of Vg can be calculated using a simulation approach. The population frequency of an allele a is assumed to follow a normal distribution with mean P_a and variance $P_a(1-P_a)/N$, where N is the number of individuals from which the allele frequency was estimated. Similarly, we assume a normal distribution for ln OR. The variance of ln OR can be obtained from meta-analyses or single studies. Random normal variables representing the allele frequency and ln OR are generated according to the above distributions and the corresponding Vg is calculated. This is repeated a large number of times to obtain a frequency distribution of Vg. The SE and 95% CI of Vg are estimated from the SD and the (0.025, 0.975) percentiles of this simulated distribution of Vg, respectively.

Non-parametric bootstrap. We also developed a non-parametric bootstrap approach to calculate the SE or CI of Vg. Such an approach does not assume particular distributions of the data and may be useful in cases when the sample sizes are small, minor allele frequencies are low or when normality assumptions are doubtful. If Vg is to be determined for a single study, a standard non-parametric bootstrap may be applied by re-sampling the observations with replacement and repeating the above method for Vg calculation on each bootstrapped sample. The inputs required are the counts of each genotype in cases and controls. Non-parametric bootstraps were carried out by the *R* package "boot." Besides the standard percentile approach, the option of constructing a bias-corrected and accelerated (BCa) CI is also available. The BCa interval adjusts for the median bias and rate of change of the SE of the estimated parameter value with respect to the true parameter value. Details are given in Efron [1987] and Efron and Tibshirani [1993].

If Vg is to be determined from meta-analyses of several (say N) studies, we may produce a bootstrapped sample for each study in each run and calculate the overall OR as obtained from a standard meta-analysis. The Vg can then be calculated for these N bootstrapped study samples (denoted Vg*). This process is repeated B times. The final SE and CI can be deduced from the SD and the (0.025, 0975) percentiles of the B Vg* values.

Delta method. The variance of the Vg estimate may be estimated analytically by the delta method. We regard risk allele frequency (P_A) and lnOR as independent variables and as the only variables with uncertainty. Since each of them is asymptotically normal, their joint distribution is asymptotically bivariate normal, and a differentiable function (such that partial derivatives exist) of these two variables should converge to a normal distribution [Lehmann, 1999].

Let g be a function that converts risk allele frequency and OR to Vg,

$$Vg = g(P_A, OR)$$

Variance of the allele frequency is equal to $P_A(1–P_A)/N$, where N is the sample size from which the allele frequency is derived. The variance of ln OR can be obtained from standard formulae for 2×2 tables (or from outputs of a logistic regression)

$$\mathrm{var}(\ln OR) \approx \mathrm{var}(OR)\left(\frac{d\ln OR}{dOR}\right)^2 = \mathrm{var}(OR)\left(\frac{1}{OR}\right)^2$$

$$\mathrm{var}(OR) \approx (OR)^2 \mathrm{var}(\ln OR)$$

$$\mathrm{var}(V_g) \approx \mathrm{var}(P_A)\left(\frac{\partial g}{\partial P_A}\right)^2 + \mathrm{var}(OR)\left(\frac{\partial g}{\partial (OR)}\right)^2$$

assuming no covariance between the allele frequency and OR. The partial derivatives are calculated with the *R* package "numDeriv."

Simple approximation method using the CI of OR. In addition to the above three methods, we may also approximate the CI of Vg by calculating the Vg using the upper and lower CI of the OR, respectively, assuming a constant allele frequency. The variance of allele frequency is usually small as compared to the variance of OR and may be considered fixed in the approximation.

TESTING PROPOSED METHODS FOR CI AND VARIANCE ESTIMATION

The four proposed methods for evaluating CI and variance were tested on three examples. The first two were extracted from the Alzgene database [Bertram et al., 2007] on SNPs rs1044925 and rs600879 in genes *SOAT1* and *SORCS1*. The original studies were described in Grupe et al. [2006] and Wollmer et al. [2003]. Since multiple population samples were used for each study, we only extracted the results from the (Greece+Italy) sample for rs1044925 and the Washington University patient sample for rs600879. The third example aims at examining the effect of a larger sample size on estimates of variance and CI. The genotype counts were multiplied by ten from the *SORCS1* example.

APPLICATIONS: A SURVEY ON TEN COMPLEX DISEASES

We conducted a survey on the variance explained by known susceptibility variants for 10 complex diseases: Alzheimer's disease, bipolar disorder, breast cancer, coronary artery disease, Crohn's disease, prostate cancer, schizophrenia, systemic lupus erythematosus (SLE), type 1 diabetes and type 2 diabetes. We only included susceptibility variants with compelling evidence of association, mostly having *P*-values passing the genome-wide significance threshold of 7.2×10^{-8} [Dudbridge and Gusnanto, 2008].

For type 1 diabetes, as a database (T1Dbase) [Hulbert et al., 2007] of known disease loci was available, we extracted relevant SNP information from it. The SNPs listed in the T1Dbase are based on the results of the most recent GWAS meta-analysis on type 1 diabetes [Barrett et al., 2009] which reported association of more than 40 loci. For type 1 DM, the MHC locus was not included due to the very complex LD pattern in this region. For Crohn's disease, we included loci confirmed in the most updated GWAS meta-analysis by Barrett et al. [2008]. The meta-analysis included three previous genome-wide scans [Libioulle et al., 2007; Rioux et al., 2007; Wellcome Trust Case Control Consortium, 2007]. Thirty-two loci were confirmed to be associated with the disease, including those identified previously. At the time of writing, there are no further GWAS on Crohn's disease that confirm further associated variants.

For other diseases, information on variants was extracted from the catalog of published GWAS established by the National Human Genome Research Institute (NHGRI) (http://www.genome.gov/gwastudies, accessed January 31, 2010). Details about catalog curation is given in Hindorff et al. [2009]. Briefly, the catalog is updated weekly and includes association studies capturing at least 100,000 SNPs at the initial stage. GWAS were identified by PubMed searches with the terms "genome-wide" OR "genome AND identification" OR "genome AND association." SNPs reaching *P*-values $<10^{-5}$ in the overall (initial+replication) sample were reported. If replication sample is not available, then SNPs from the discovery sample were reported. The risk allele frequencies, ORs and their corresponding CIs were calculated from the largest sample size, typically a joint analysis of the discovery and replication samples. Since the PubMed search strategy by the catalog does not necessarily guarantee that all confirmed loci are reported, we also searched for recent review articles and publications concerning the relevant diseases in Nature Genetics. We notice that these are still "crude" approaches to complement the NHGRI catalog and there may still be SNPs that are genome-wide significant but escape our detection. For the overall probability of disease (*K*), we considered the lifetime risk of disease as mentioned in the literature. Please refer to Supplementary Tables 1 and 2 for the sources of heritability and lifetime risk estimates for the diseases.

The majority of the studies were based on Caucasian samples. If the same SNP was reported both in Caucasians and another ethnic group(s), information about the SNP was extracted from the Caucasian sample, for the sake of consistency. If an identical SNP was reported in two or more studies or if two SNPs have $r^2>0.8$, we considered the study with the largest sample size. LD information was extracted from the HapMap CEU sample unless given in the original paper. For SNPs with $0.05<r^2<0.8$, the combined Vg is evaluated by taking into account of the covariance structure as previously described. In view of the large number of genetic variants surveyed, CIs were simply estimated from the reported upper and lower CI of ORs, and we made the approximation that the OR from case-control studies is close to the lifetime relative risk estimate. An additive model (on ln OR scale) was assumed unless otherwise specified.

RESULTS

RELATIONSHIP OF Vg WITH DISEASE PREVALENCES, RELATIVE RISKS AND ALLELE FREQUENCIES

Higher disease prevalences, bigger relative risks, and having allele frequencies closer to 0.5 all lead to a larger explained variance. Their relationships are shown in Supplementary Figures 1–6. It is noteworthy that the explained variance is rather small for common effect sizes (allele-wise ORs of 1.1 to 1.5).

EVALUATION OF THE CI AND SE OF Vg

The application of various approaches to estimate the SE and CI of Vg in three example data sets is shown in Tables I and II. The examples were derived from association data in AlzGene. For the estimation of SE, the delta method, simulation from normal distribution and non-parametric bootstrap gave similar estimates, though the discrepancy was larger in the second example when the minor allele frequency was lower.

Constructing CI by considering the lower and upper CI of the OR gave estimates very close to those obtained from the simulation approach. The CI obtained from non-parametric bootstrap were close to those obtained from the other two methods, but again in the second example the differences became larger, probably due to the low MAF. When the sample size increased by 10-fold (the third example), all methods considered gave similar results as the normal assumptions are more reliable in a larger sample size. The standard percentile and the BCa methods gave very close results. The sampling distribution of Vg, as estimated from the bootstrapped values, is non-normal in a modest sample size (Supplementary Figures 7–8). Hence CIs constructed from the SE are not reliable. From theory, Vg should approach normality as sample size increases and this is reflected in Supplementary Figure 9.

RESULTS OF SURVEY ON TEN COMPLEX DISEASES

The summary results of the survey on 10 complex diseases are listed in Table III. Supplementary Table 3 shows more detailed results of the survey, including the Vg explained by each variant for all 10 diseases.

TABLE I. Three example data sets to test the SD and confidence interval (CI) of Vg

	Case			Control		
Example	A/A(risk)	A/a	a/a	A/A(risk)	A/a	a/a
1	35	112	70	21	58	68
2	6	87	325	4	54	319
3	60	870	3,250	40	540	3,190

A is the risk allele in all examples.

The number of loci confirmed for each disease ranged from 4 to over 40. The total variance in liability explained was generally around 6–12%, with a median of 9.81%. For bipolar disorder and schizophrenia, the number of identified common variants was small (<5) and the Vg were only 2.14% and 0.32%, respectively. On the other hand, the Vg for Alzheimer's disease was high, mainly due to the large effect size of the *APOE* loci. Compared to the total heritability, the proportion explained ranged from ~0.4% to around 25%. Averaging over all diseases under study, the median Vg per associated SNP was ~0.25%, reflecting that in general individual common variants explain a very small proportion of the variance of liability. We also calculated the total heritability divided by the median Vg for each disease. The figure may serve as a crude and conservative estimate of the number of variants underlying each trait. Note that the variants we have already found are likely to confer larger effects than the undiscovered ones, so the actual median Vg is probably much lower. The value of heritability divided by median Vg ranges from approximately a hundred to more than a thousand (for schizophrenia). The results suggest some heterogeneity in the genetic architecture of the studied traits, though this may also be due to some diseases being studied more thoroughly than others. Though most established associations included here are common variants with small effects, rare variants that are likely to have large effects also tend to have low Vg (low frequencies lead to low Vg). Therefore, we still expect the median Vg to be low even if rare variants are considered. Rare variants with very large effects probably would have been detected by other means such as linkage studies.

DISCUSSION

In this study we estimated the variance explained by all genetic variants reported for 10 complex diseases. We demonstrated that knowing the variance explained by detected genetic loci enables us to assess the completeness of our knowledge about the genetic and/or environmental factors involved in complex diseases. There are a number of advantages associated with the use of Vg. First, it is based on the classical model of variance partitioning and would appear familiar to geneticists. In addition, Vg

TABLE II. SD and confidence intervals of Vg using different methods

	Method	Example 1	Example 2	Example 3
SD	Delta method	1.471E−02	7.376E−03	2.333E−03
	Sim from normal	1.506E−02	8.000E−03	2.340E−03
	Non-parametric bootstrap	1.500E−02	6.798E−03	2.116E−03
Lower CI	From CI of OR	1.233E−04	2.395E−04	4.457E−03
	Sim from normal	3.240E−04	2.697E−04	4.436E−03
	Non-parametric bootstrap (percentile)	2.894E−04	3.230E−04	4.650E−03
	Non-parametric bootstrap (BCa)	2.506E−04	2.920E−04	4.623E−03
Upper CI	From CI of OR	5.635E−02	2.939E−02	1.349E−02
	Sim from normal	5.593E−02	3.011E−02	1.358E−02
	Non-parametric bootstrap (percentile)	5.597E−02	2.578E−02	1.291E−02
	Non-parametric bootstrap (BCa)	5.562E−02	2.551E−02	1.286E−02

Sim, simulation; BCa, bias-corrected and accelerated bootstrap.

may be interpreted relative to all etiological influences (i.e. total variance in liability) as well as genetic influences (heritability) alone. It can also be readily extended to deal with complications such as multi-allelic loci, haplotypes and correlated markers.

Another similar measure is the locus-specific sibling recurrence risk ratio (λs) [Risch, 1990]. Locus-specific λs itself is difficult to interpret unless compared to the overall λs. On the other hand, variance explained can be either interpreted on its own or compared to the total heritability. In addition, the overall λs represents the combined effect of shared genes and environment while Vg may be compared to the heritability which separates shared genetic from shared environmental effects. The numerator of the overall λs (risk to sibling of a case) is also bound by the disease prevalence. Currently there are no methods to extend λs to deal with complications such as markers in LD or multilocus genotypes.

Population attributable risk (PAR) is another commonly employed measure. It is defined as the proportion of disease cases in the population that could be prevented if a causal risk factor were eliminated. As it is yet not possible to alter our genes, whether PAR is an appropriate concept in genetics is debatable. In addition, PAR is not additive and does not reflect the contribution of a genetic variant to the total liability of disease.

We conducted a survey to assess the variance explained by known common variants for 10 complex diseases. We note that there are limitations to this survey. For example, the effect size estimates are based on combining the discovery and replication samples, or just the discovery sample. It is well-known that effect size estimates are often biased upward for the selected markers passing a significance threshold, a phenomenon called "winner's curse" [Garner, 2007; Zollner and Pritchard, 2007]. The ORs reported here may therefore be subject to this bias, although the bias should not be large if the replication sample is large enough. As described in the methods section, we mainly relied on the NHGRI catalog to extract information on significant variants. However, some variants reaching genome-wide association may still be missed unless a comprehensive search is performed, preferably by researchers who are familiar with the genetics of the respective diseases. We believe that the proportion of variants missed is low and is unlikely to change the results much. Another point to note is that we focused on common variants in the survey and rare mutations (e.g. BRCA mutations) or structural variants (e.g. deletions in schizophrenic cases [St Clair, 2009]) are not considered.

In this study we assume the heritability refers to the variance in liability to a lifetime diagnosis of the disease and hence the liability threshold was determined from lifetime risk. We have also made the approximation that the OR from case-control studies is close to the lifetime relative risk estimate. We note that the approximation does not work very well when the disease is common. The actual lifetime relative risk, with consideration of competing risks, should be *smaller* than the observed OR. We have also computed the variance explained for breast and prostate cancers by using lifetime relative risks (RR) estimates as inputs. The details of converting ORs (from prevalent case-control studies) to lifetime RRs are presented elsewhere [So and Sham, 2010]. The resulting total Vg are smaller than when using the approximation.

TABLE III. Variance in liability explained for ten complex diseases

		Variance explained (%)								
	No. of loci	Sum	95% CI			Average	Median	h^2	Proportion of h^2 explained	h^2/median Vg
Alzheimer's disease	4	18.34	NA		NA	4.586	0.343	0.79	23.22	230
Bipolar disorder	5	2.14	1.18	–	3.44	0.427	0.239	0.77	2.77	323
Breast cancer[a]	13	6.63	3.44	–	11.17	0.510	0.392	0.53	12.52	135
Coronary artery disease (including myocardial infarction)	12	12.32	8.34	–	18.30	1.027	0.580	0.49	25.15	84
Crohn's disease	32	7.39	NA	–	NA	0.231	0.158	0.55	13.43	347
Prostate cancer[a]	27	15.43	8.61	–	26.00	0.417	0.344	0.50	31.16	144
Schizophrenia	4	0.32	0.14	–	0.58	0.079	0.065	0.81	0.39	1,238
Systemic lupus erythematosus	23	8.71	5.33	–	12.96	0.379	0.296	0.66	13.20	223
Type 1 diabetes[b]	45	10.90	6.69	–	16.19	0.242	0.104	0.80	13.63	770
Type 2 diabetes	25	11.84	6.36	–	19.04	0.474	0.305	0.42	27.93	139

h^2, heritability; NA, not available. The total heritability divided by the median Vg serves as a conservative and crude estimate of the number of variants underlying the disease.

[a] Corrected sum of Vg using lifetime relative risk as input: 5.70% for breast cancer and 12.50% for prostate cancer.

[b] MHC region not included.

In addition, the heritability may not refer to the variance in liability to a *lifetime* diagnosis of the disease, especially for late-onset diseases. It will depend on the ascertainment scheme and analysis methods of the twin or family samples. However, to maintain consistency, we have used lifetime risk throughout. It is very difficult to thoroughly address this problem as the sample collection schemes and analysis approaches for each heritability study may be different. For instance, some may follow the twins longer and some may be recruiting relatively young subjects. For late-onset diseases, young subjects who may be free of disease at the time of study may develop the condition if followed longer. This problem is usually known as right-censoring in survival analysis. The effects of censoring may be taken into account by employing more complex statistical techniques such as frailty models [Locatelli et al., 2007; Yashin and Iachine, 1995]. Considering these factors, the proportion of heritability explained for late-onset diseases may be biased upward. For the two diseases with the highest lifetime risk (type 2 DM and CAD), repeating the calculations with their prevalence estimates instead of lifetime risks give Vg that are roughly half of the original (5.24 and 5.33%, respectively).

From the survey we found that the known common susceptibility variants explain only a portion of the total genetic component of most complex diseases. The proportion of heritability explained differs across traits, and is rather low for the psychiatric diseases (excluding Alzheimer's disease). The differences may be due to the fact that some diseases are studied more extensively than others or may reflect true differences in the underlying genetic architecture. For example, three large-scale GWAS on schizophrenia have been published [Purcell et al., 2009; Shi et al., 2009; Stefansson et al., 2009] and a "partial meta-analysis" (SNPs with $P<10^{-3}$ were combined across samples) was done. Nevertheless, the number of loci revealed was small and total variance explained is small. It is yet unknown if this is due to greater heterogeneity in patients' phenotypes, inadequate sample sizes, different genetic architecture from other traits or some other reasons.

The confirmed common variants account for at most a quarter of the entire heritability. This confirms that a considerable proportion of heritability is still "missing." Manolio et al. [2009] gave a detailed review on the possible sources of "missing" heritability and strategies to uncover the genetic basis of complex diseases. For example, there may be rare variants, variants with very small effect sizes or structural variations (such as copy number variants, inversions and translocations) that cannot be readily detected with current technologies or sample sizes. Also of note is that the contribution of gene-gene and gene-environment interactions has not been fully explored. Epistasis may account for part of the heritability that remains undiscovered.

It is clear that we are still a long way from understanding the complete genetic basis of most complex diseases. Strategies like further increasing sample sizes, exploration of gene-gene and gene-environmental interactions, better coverage of rare variants and structural variations, may help to discover the hidden portion of heritability. It might also be interesting to apply the methods described here to a larger variety of diseases and repeat the analysis after a period of time when more susceptibility variants are found.

ACKNOWLEDGMENTS

H.C.S. was supported by a Croucher Foundation Scholarship.

WEB RESOURCES

Programs (mainly written in R) to implement the methodologies described in the paper are available at http://sites.google.com/site/honcheongso/software/varexp.

REFERENCES

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40:955–962.

Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, Plagnol V, Pociot F, Schuilenburg H, Smyth DJ, Stevens H, Todd JA, Walker NM, Rich SS, Type 1 Diabetes Genetics C. 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41:703–707.

Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. 2007. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet 39:17–23.

Dudbridge F, Gusnanto A. 2008. Estimation of significance thresholds for genomewide association scans. Genet Epidemiol 32:227–234.

Efron B. 1987. Better Bootstrap Confidence-Intervals. Journal of the American Statistical Association 82:171–185.

Efron B, Tibshirani R. 1993. An introduction to the bootstrap. New York: Chapman & Hall.

Falconer D. 1965. The inheritance of liability to certain diseases, estimated from the incidence among relatives. Annals of Human Genetics 29:51–76.

Garner C. 2007. Upward bias in odds ratio estimates from genome-wide association studies. Genet Epidemiol 31:288–295.

Grupe A, Li Y, Rowland C, Nowotny P, Hinrichs AL, Smemo S, Kauwe JS, Maxwell TJ, Cherny S, Doil L, Tacey K, van Luchene R, Myers A, Wavrant-De Vrieze F, Kaleem M, Hollingworth P, Jehu L, Foy C, Archer N, Hamilton G, Holmans P, Morris CM, Catanese J, Sninsky J, White TJ, Powell J, Hardy J, O'Donovan M, Lovestone S, Jones L, Morris JC, Thal L, Owen M, Williams J, Goate A. 2006. A scan of chromosome 10 identifies a novel locus showing strong association with late-onset Alzheimer disease. Am J Hum Genet 78:78–88.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367.

Hulbert EM, Smink LJ, Adlem EC, Allen JE, Burdick DB, Burren OS, Cassen VM, Cavnor CC, Dolman GE, Flamez D, Friery KF, Healy BC, Killcoyne SA, Kutlu B, Schuilenburg H, Walker NM, Mychaleckyj J, Eizirik DL, Wicker LS, Todd JA, Goodman N. 2007. T1DBase: integration and presentation of complex data for type 1 diabetes research. Nucleic Acids Res 35:D742–D746.

Lehmann EL. 1999. Elements of large sample theory. New York: Springer.

Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, Vermeire S, Dewit O, de Vos M, Dixon A, Demarche B, Gut I, Heath S, Foglio M, Liang L, Laukens D, Mni M, Zelenika D, Van Gossum A, Rutgeerts P, Belaiche J, Lathrop M, Georges M. 2007. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet 3:e58.

Locatelli I, Rosina A, Lichtenstein P, Yashin AI. 2007. A correlated frailty model with long-term survivors for estimating the heritability of breast cancer. Stat Med 26:3722–3734.

Manolio TA, Brooks LD, Collins FS. 2008. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118: 1590–1605.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. Nature 461:747–753.

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460:748–752.

Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, Green T, Kuballa P, Barmada MM, Datta LW, Shugart YY, Griffiths AM, Targan SR, Ippoliti AF, Bernard EJ, Mei L, Nicolae DL, Regueiro M, Schumm LP, Steinhart AH, Rotter JI, Duerr RH, Cho JH, Daly MJ, Brant SR. 2007. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 39:596–604.

Risch N. 1990. Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228.

Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, Dudbridge F, Holmans PA, Whittemore AS, Mowry BJ, Olincy A, Amin F, Cloninger CR, Silverman JM, Buccola NG, Byerley WF, Black DW, Crowe RR, Oksenberg JR, Mirel DB, Kendler KS, Freedman R, Gejman PV. 2009. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 460:753–757.

So HC, Sham PC. 2010. Effect size measures in genetic association studies and age-conditional risk prediction. Hum Hered 70: 205–218.

St Clair D. 2009. Copy number variation and schizophrenia. Schizophr Bull 35:9–12.

Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, Werge T, Pietilainen OP, Mors O, Mortensen PB, Sigurdsson E, Gustafsson O, Nyegaard M, Tuulio-Henriksson A, Ingason A, Hansen T, Suvisaari J, Lonnqvist J, Paunio T, Borglum AD, Hartmann A, Fink-Jensen A, Nordentoft M, Hougaard D, Norgaard-Pedersen B, Bottcher Y, Olesen J, Breuer R, Moller HJ, Giegling I, Rasmussen HB, Timm S, Mattheisen M, Bitter I, Rethelyi JM, Magnusdottir BB, Sigmundsson T, Olason P, Masson G, Gulcher JR, Haraldsson M, Fossdal R, Thorgeirsson TE, Thorsteinsdottir U, Ruggeri M, Tosato S, Franke B, Strengman E, Kiemeney LA, Genetic R, Outcome in P, Melle I, Djurovic S, Abramova L, Kaleda V, Sanjuan J, de Frutos R, Bramon E, Vassos E, Fraser G, Ettinger U, Picchioni M, Walker N, Toulopoulou T, Need AC, Ge D, Yoon JL, Shianna KV, Freimer NB, Cantor RM, Murray R, Kong A, Golimbet V, Carracedo A, Arango C, Costas J, Jonsson EG, Terenius L, Agartz I, Petursson H, Nothen MM, Rietschel M, Matthews PM, Muglia P, Peltonen L, St Clair D, Goldstein DB, Stefansson K, Collier DA. 2009. Common variants conferring risk of schizophrenia. Nature 460:744–747.

Visscher PM. 2008. Sizing up human height variation. Nat Genet 40:489–490.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Wollmer MA, Streffer JR, Tsolaki M, Grimaldi LM, Lutjohann D, Thal D, von Bergmann K, Nitsch RM, Hock C, Papassotiropoulos A. 2003. Genetic association of acyl-coenzyme A: cholesterol acyltransferase with cerebrospinal fluid cholesterol levels, brain amyloid load, and risk for Alzheimer's disease. Mol Psychiatry 8:635–638.

Yashin AI, Iachine IA. 1995. Genetic analysis of durations: correlated frailty model applied to survival of Danish twins. Genet Epidemiol 12:529–538.

Zollner S, Pritchard JK. 2007. Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am J Hum Genet 80:605–615.