OPEN ACCESS Freely available online

PLoS one

Underestimated Effect Sizes in GWAS: Fundamental Limitations of Single SNP Analysis for Dichotomous Phenotypes

Sven Stringer[1]*, Naomi R. Wray[2,3], René S. Kahn[1], Eske M. Derks[1]

1 Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht, The Netherlands, **2** Psychiatric Genetics Laboratory, Queensland Institute of Medical Research, Brisbane, Australia, **3** Queensland Brain Institute, University of Queensland, Brisbane, Australia

Abstract

Complex diseases are often highly heritable. However, for many complex traits only a small proportion of the heritability can be explained by observed genetic variants in traditional genome-wide association (GWA) studies. Moreover, for some of those traits few significant SNPs have been identified. Single SNP association methods test for association at a single SNP, ignoring the effect of other SNPs. We show using a simple multi-locus odds model of complex disease that moderate to large effect sizes of causal variants may be estimated as relatively small effect sizes in single SNP association testing. This underestimation effect is most severe for diseases influenced by numerous risk variants. We relate the underestimation effect to the concept of non-collapsibility found in the statistics literature. As described, continuous phenotypes generated with linear genetic models are not affected by this underestimation effect. Since many GWA studies apply single SNP analysis to dichotomous phenotypes, previously reported results potentially underestimate true effect sizes, thereby impeding identification of true effect SNPs. Therefore, when a multi-locus model of disease risk is assumed, a multi SNP analysis may be more appropriate.

Citation: Stringer S, Wray NR, Kahn RS, Derks EM (2011) Underestimated Effect Sizes in GWAS: Fundamental Limitations of Single SNP Analysis for Dichotomous Phenotypes. PLoS ONE 6(11): e27964. doi:10.1371/journal.pone.0027964

Editor: Nicholas John Timpson, University of Bristol, United Kingdom

Received May 9, 2011; **Accepted** October 28, 2011; **Published** November 28, 2011

Funding: Eske Derks is supported by the Netherlands Scientific Organization (NWO; project number 451-08-010; www.nwo.nl). NWO 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: s.stringer@umcutrecht.nl

Introduction

Since the first GWA study in 2005[1], hundreds of GWA studies have been published, reporting more than 2000 associations[2]. However, despite large heritability estimates, relatively few associations have been reported for most complex traits. Moreover, associations found in GWA studies often explain only a small proportion of the phenotypic variation[3]. For example, although 71 independent loci have been identified as being associated with Crohn's Disease, they still account for only 23% of the estimated heritability[4]. GWA studies of psychiatric diseases show an even less favorable picture. For instance, schizophrenia has an estimated heritability of 80%[5,6], but observed genetic variants currently account for less than 1% of the variance[7].

One explanation of the missing heritability is that complex diseases are caused by a large number of causal variants with small effect sizes. Odds ratios (OR) reported in GWA studies are typically small (i.e., a median OR of 1.33[8]). The many associations that are tested require a very low significance threshold to prevent an inflated genome-wide type I error. This reduces the probability of identifying SNPs with small effect size, unless sample sizes are large enough to achieve sufficient power to identify such SNPs. Using large combined datasets within scientific consortia has significantly increased power in GWA studies. Despite this increase in power, still only a small number of associated variants have been identified[3]. A second explanation of the missing heritability is that risk SNPs are correlated with unobserved causal genetic variants, since they are unlikely to be causal themselves[9]. The lower the correlation between an observed risk SNP and the unobserved causal variant, the smaller the estimated effect size of the risk SNP, resulting in less explained variance and hence decreased power. This decrease in power is most dramatic for rare variants (i.e., SNPs with minor allele frequencies less than 5% or even 1%) and these variants are less likely to be tagged by the genotyped SNPs.

The present study addresses a fundamental limitation of traditional GWA analysis of dichotomous phenotypes which provides an additional explanation for the difficulty in identifying effect SNPs and the missing heritability. By definition complex diseases are caused by numerous risk variants. However, as single SNP analysis only considers a single SNP at a time, other SNPs associated with disease can be considered omitted covariates. Gail et al.[10] proved in the context of generalized linear models that omitting covariates can result in asymptotically underestimated effect sizes, even in the absence of confounders. Confounders are (possibly omitted) covariates that are associated with other covariates or variables of interest. Gail et al. showed that only the linear-link and log-link functions produce asymptotically unbiased effect sizes in generalized linear regression, although the log-link function can produce asymptotically biased intercepts[10]. In the context of logistic regression, this underestimation effect reduces the efficiency of effect size statistics[11]. Neuhauss

and Jewell[12] provided formulas to assess this bias for several common link functions, including the logit and probit link functions, which are most suitable for analyzing dichotomous phenotypes. In linear regression omitting covariates has no effect on the estimated effect size[11].

The underestimation effect of non-linear link functions can be best understood in terms of the statistical concept of collapsibility. Simpson[13] wrote a seminal paper on the surprising non-equivalence of conditional and marginal odds ratios, which has later been referred to as Simpsons's paradox[14,15]. Given three dichotomous variables X, Y, and Z, he showed that even if the odds ratios between X and Y conditional on the value of Z are equal ($OR_{XY|Z=0} = OR_{XY|Z=1} = OR_{XY|Z}$), this does not imply that the marginal odds ratios equal the conditional odds ratio ($OR_{XY} = OR_{XY|Z}$). In other words, the odds ratio is a non-collapsible effect measure, as the marginal effect measure (OR_{XY}) cannot generally be expressed as a weighted average of the conditional effect measures ($OR_{XY|Z=0}$ and $\mathrm{OR}_{\mathrm{XY|Z=1}}$). In the context of GWAS, Y is disease status, X is the genotype of an allele of interest, and Z is the number of risk variants in the genetic background. In this context Z is unlikely to be dichotomous. An effect size measure would be called collapsible if the marginal effect size of SNP X, averaged over all possible genetic backgrounds Z, can be expressed as a weighted average of all conditional effect sizes of SNP X (i.e., conditional on specific genetic background Z).[14,15]

Two conditions have been identified that do result in collapsible odds ratios[16]. The first condition is that disease status Y and background Z are independent given SNP X. This implies that ignoring SNPs which have no effect on disease will not result in underestimation. The second condition is that SNP X and genetic background Z are independent given disease status Y. This situation cannot arise if we (safely) assume that SNPs or the causal variants with which they are in linkage disequilibrium cause disease status and not vice versa (see Hernán et al.[14] for a discussion on the importance of causal assumptions when dealing with Simpson's paradox). In other words, conditional and marginal odds ratios are only equivalent if the SNP of interest or the genetic background is not associated with disease status.

Despite the use of the word 'bias' by earlier authors[10–12], Greenland et al. [15] note that non-collapsibility is technically not a bias. It reflects the mathematical fact that for some effect measures marginal and conditional effect sizes are non-equivalent. When choosing a non-collapsible effect size measure, one merely needs to decide whether the marginal, the conditional effect size or both are of interest[14]. We believe that in GWA studies the odds ratio conditional on a fixed genetic background reflects the relative importance of a single SNP better than the marginal odds ratio. A single SNP analysis would estimate the marginal odds ratio, whereas a multi SNP analysis would estimate the odds ratio conditional on a fixed genetic background. Risk difference and risk ratio are examples of collapsible effect measures[15]. However, as traditional GWA analyses are often based on odds ratios, we will focus here on the logistic or odds disease model.

Complex diseases in GWA studies can be characterized by numerous risk SNPs with small effect sizes. Although the average effect size is expected to be small, the variance in the genetic background increases with the number of true risk SNPs. In the present simulation study we investigate the potential implications of non-collapsibility for traditional GWA studies. We first study the relation between the marginal and the conditional odds ratio under a naive disease model. The simplicity of the naive model facilitates the simulation and mathematical analysis of the underestimation effect. We report how disease characteristics (e.g., prevalence, number of risk SNPs, minor allele frequency, and effect sizes) influence the underestimation effect. We also show how this underestimation affects the estimated explained variance. Subsequently, we illustrate the underestimation effect under a more realistic genetic architecture. Finally, we discuss the implications of underestimating effect size and suggest potential solutions.

Methods

Modeling a heritable disease requires a function relating genotype to disease risk. To simulate the implications of traditional GWA analysis using odds ratios, we constructed a disease generating model based on the odds model of disease risk. Before discussing this model in more detail, we illustrate the disease generating process of the odds model with an example. We assume that all risk alleles at different loci have equal frequency and equal effect size (these assumptions have been shown by others to have little impact on interpretation of results)[17–19]. For example, Figure 1 shows disease probability and the distribution of risk allele counts for a disease with a prevalence of 1%, assuming a total number of 200 effect alleles (i.e., 100 risk SNPs); the odds ratio of each risk allele is 1.6 and the risk allele frequencies are 0.25. Under this additive model on the log odds scale, people carry on average 50 risk alleles (binomial mean is 200×0.25) corresponding to a negligible disease risk. However, as the number of risk alleles exceeds a threshold, disease probability increases rapidly, demonstrating the highly non-linear relationship between genetic risk factors and disease risk. Those at highest risk of disease carry more risk alleles, >70 in this example, but each affected person could have a unique portfolio of risk alleles; the effect of a risk allele on disease depends on the genetic background (other risk alleles) carried by an individual. The (implicit) error variance in the odds model is $\pi^2/3$, the variance of the standard logistic distribution.

As mentioned before, the marginal odds ratio produced by single SNP analysis is averaged over all possible genetic

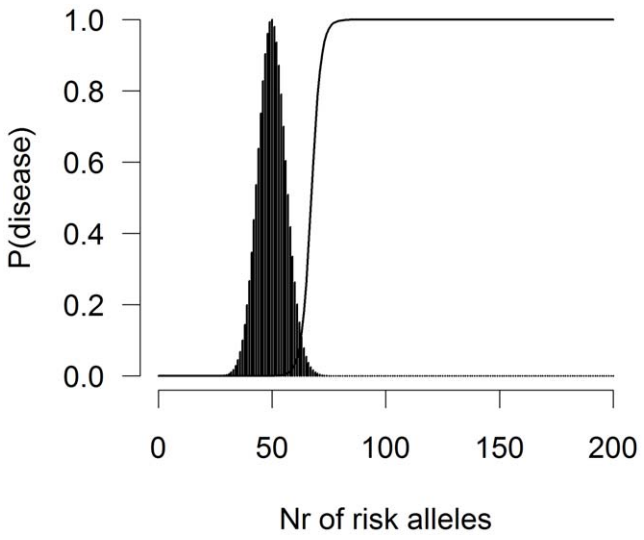

Figure 1. Disease model. Probability of disease as a function of the number of risk alleles (line) and the distribution of risk alleles in a large sample (n = 10,000) (histogram). Disease prevalence is 1%. The odds ratio of each risk SNP is 1.6 and the allele frequencies of risk alleles are 0.25. The maximum number of risk alleles is 200 (i.e., 100 SNPs). The (implicit) error variance of the odds model is $\pi^2/3$.
doi:10.1371/journal.pone.0027964.g001

backgrounds. However, the odds ratios from the odds model, which we assume generates the disease, are conditional on a fixed background odds of disease (see section A3.2 in Appendix S1). We will therefore refer to the conditional odds ratio as the true odds ratio OR_t and to the marginal odds ratio as the (possibly under-)estimated odds ratio OR_e. To relate OR_e to the prespecified OR_t, we performed the following four steps: (1) we specified a disease generating model based on disease characteristics including OR_t, (2) we mathematically derived the genotype distribution of a single SNP of interest given disease status and disease characteristics, (3) we repeatedly simulated a case-control sample of the SNP of interest based on this genotype distribution and computed the corresponding SNP-based odds ratio (OR_e), and (4) we reported the median of all estimated odds ratios OR_e, reflecting the asymptotic marginal odds ratio estimated by single SNP analysis. We now discuss the disease generating model.

We specified a disease generating model with four parameters: (1) disease prevalence p_D, (2) true (allelic) odds ratio OR_t, (3) minor allele frequency of risk alleles p_a, and (4) the total number of effect alleles n_a. Risk alleles can be either minor alleles or major alleles. We only consider minor risk alleles, as the analysis is analogous for major risk alleles. Let D be disease status and $z(x_a,\beta_0,\beta_1)=\beta_0+\beta_1 x_a$ a linear function of the number of risk alleles x_a with effect size β_1 and intercept β_0. Then the probability of disease conditional on the number of risk alleles is defined as (see also Equation S3 in Appendix S1)

$$P(D=1|X_a=x_a;\beta_0,\beta_1)=\frac{1}{1+\exp[-z(x_a,\beta_0,\beta_1)]} \qquad (1)$$

As effect size β_1 is defined on a log odds scale, $\exp(\beta_1)$ is the effect size on an odds scale. Therefore $OR_t=\exp(\beta_1)$ is the true odds ratio in the biological reality we aim to model.

So far we specified the probability of disease conditional on the number of risk alleles. To obtain a full probability model of disease, it is necessary to specify the distribution of risk alleles as well. Assuming Hardy-Weinberg equilibrium and linkage equilibrium for a total of n_a effect alleles (i.e., twice the number of risk SNPs) and risk allele frequency p_a, the number of risk alleles x_a in the population can be modeled with a binomial distribution

$$P(X_a|n_a,p_a)=\binom{n_a}{x_a}p_a^{x_a}(1-p_a)^{n_a-x_a} \qquad (2)$$

Combining distribution 1 and 2 results in a joint probability distribution of disease and number of risk alleles given four disease parameters: risk allele frequency p_a, total number of effect alleles n_a, effect size β_1 on a log odds scale, and intercept β_0.

$$P(D,X_a|\beta_0,\beta_1,n_a,p_a)=P(D|X_a,\beta_0,\beta_1)P(X_a|n_a,p_a) \qquad (3)$$

The probability of disease status $P(D|\beta_0,\beta_1,n_a,p_a)$ can be obtained by summing over all possible genetic liabilities X_a.

Although β_0 has an interpretation as the baseline (or background) log odds of disease, there is no strong prior information what this might be, as there is for the other three model parameters. However, as disease prevalence is an observed disease characteristic, it is possible to set β_0 such that the disease probability of the model $P(D=1|\beta_0,\beta_1,p_a,n_a)$ equals disease prevalence p_D. Although $P(D=1|\beta_0,\beta_1,p_a,n_a)=p_D$ cannot be solved analytically for β_0, an error function, such as the sum squared error can be defined (Equation S1 in Appendix S1). This error function can be minimized to obtain a numerical approximation of β_0 that satisfies the equality. Because $OR_t=\exp(\beta_1)$ and number of risk SNPs $n_s=\frac{1}{2}n_a$, the result is a model of disease with the four parameters: disease prevalence (p_D), true allelic odds ratio (OR_t), number of risk SNPs (n_s), and risk allele frequency in risk SNPs (p_a). As a fifth parameter, error variance on the liability trait could be included to model the proportion of variance explained by all SNPs (heritability), but as this was not required for the derivations in this paper, we left the error variance implicit and constant (see section A3.4 in Appendix S1). From the four-parameter disease model we derived the genotype distribution of SNP s given disease status and model parameters $P(X_s|D,p_D,OR_t,p_a,n_a)$ (see Equation S2 in Appendix S1). Based on this distribution we simulated 10,000 case-control samples and computed the median estimated SNP-based odds ratio OR_e. By relating the odds ratio OR_e obtained when performing a single SNP analysis to the true odds ratio OR_t, we could study the underestimation effect for different disease characteristics. Further details on simulation technicalities can be found in section A1 of Appendix S1.

Although the odds model is mathematically convenient, it assumes a constant effect size and minor allele frequency for all risk alleles. Therefore we performed a second simulation investigating the underestimation effect under a more realistic genetic architecture. In GWA studies absolute effect sizes on the log odds scale are roughly exponentially distributed[20,21]. Consequently, effect sizes were drawn from an exponential distribution with rate parameter 5. This corresponds with an expected OR_t of 1.25, but acknowledges that true effect sizes are frequently small and rarely large. To avoid rare variants, allele frequencies were assumed to be uniformly distributed between 0.05 and 0.95. Effect sizes and allele frequencies were drawn once and fixed in the rest of the simulation replicates. The odds disease model from the first simulation is easily extended to accommodate different fixed effect sizes by defining $z(x_a,\beta_0,\beta_1)=\beta_0+\sum_{i=1}^{n_s}\beta_i x_i$ in Equation 1, where n_s is the number of SNPs, β_i is the effect size of SNP i and $x_i\in\{0,1,2\}$ refers to the number of risk alleles at SNP i. The intercept $\beta_0$was chosen corresponding to a disease prevalence of 1%.

The asymptotic single SNP estimate was again assessed by generating 10,000 case-control samples and computing for each SNP the median odds ratio using a single SNP logistic regression. Case-control samples, 5000 subjects each, were generated by repeatedly drawing from the population distribution until 2500 cases and 2500 controls were sampled.

Results

If a disease is caused by a single risk SNP, the odds ratio estimated by single SNP analysis (OR_e) will, on average, reflect the true odds ratio (OR_t) (section A2 in Appendix S1). However, if a disease is caused by numerous risk SNPs, the median OR_e follows an asymptote. Figure 2 shows the relationship between median OR_e and OR_t for diseases caused by 100 risk SNPs with different prevalences (A) and different minor allele frequencies of the risk SNPs (B). A wide range of prevalences and minor allele frequencies results in upper limits for the median SNP-based odds ratio. This asymptotic effect is more dramatic in diseases with higher prevalences and/or higher minor allele frequencies. Depending on the model parameters, the upper bound is reached with true model odds ratios as low as 1.5. In that case traditional association testing cannot differentiate, for example, between a

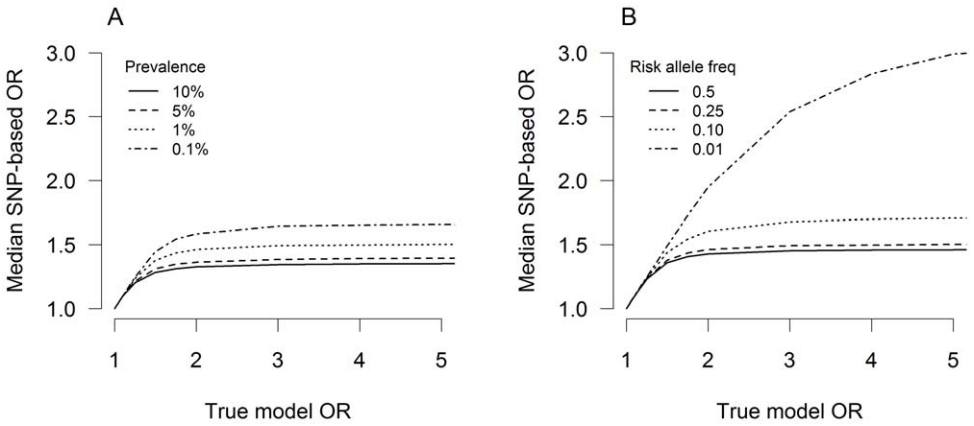

Figure 2. Numerous risk SNPs. Relationship between median estimated SNP-based odds ratio (OR_e) and true conditional model odds ratio (OR_t) for a disease with 100 effect SNPs. (A) Different prevalences with risk allele frequency 0.25. (B) Different risk allele frequencies with prevalence 1%. Simulations are based on a case-control study of 3500 subjects and a 1:1 case:control ratio. Medians are based on 10,000 case-control samples.
doi:10.1371/journal.pone.0027964.g002

true odds ratio of 1.5 and a true odds ratio of 3, as both will be estimated at 1.5, the maximum value that can be obtained. In other words, under this disease model large true effect sizes are not identified as such by single SNP association testing.

The asymptotic constraint on the estimated odds ratio is caused by two factors. First, single SNP odds ratios (OR_e) are estimated across an average over all possible background risks in cases and controls; this can be seen when computing the conditional probability of disease status given the genotype at a particular SNP (section A3.1 in Appendix S1). Only when the risk allele frequency (p_a) approaches zero, the background risk will approach zero, which is similar to a disease with a single risk SNP. This is why low risk allele frequencies (for example, $p_a = 0.01$) result in a delayed asymptotic effect compared to high risk allele frequencies ($p_a \geq 0.1$) (Figure 2B). If the odds ratio for an allele could be estimated in a subsample of the population that all carried the same background risk, then the SNP-based odds ratio (OR_e) would (almost) equal the true odds ratio (OR_t) (see section A3.2 in Appendix S1).

Although weighted averaging is part of the explanation of the constrained odds ratios, it is not a sufficient explanation, because for continuous phenotypes the asymptotic effect does not occur when computing SNP-based effect sizes (section A3.3 in Appendix S1). It is due to the non-collapsibility of the odds ratio that averaging over background risks results in a discrepancy between the estimated marginal odds ratio OR_e and the true conditional odds ratio OR_t.

A priori the total number of risk SNPs in a disease is unknown, but it is of course possible to simulate the results of traditional association testing for diseases with different numbers of risk SNPs. The asymptotic effect is stronger for diseases which are influenced by a large number of risk SNPs (Figure 3). In other words, an increase in the number of SNPs associated with disease results in increased underestimation. As complex diseases are assumed to be influenced by many risk SNPs, analyzing numerous large-effect SNPs with traditional association testing would result in considerable underestimation. This type of underestimation is not due to a lack of power as increasing sample size will decrease the variance of effect sizes obtained, but will not reduce underestimation due to the non-collapsibility of the odds ratio.

We will now show that underestimation of effect sizes can result in additional missing heritability. Narrow-sense heritability is the percentage of total phenotypic variance that is explained by additive genetic variance. Figure 4 compares the explained variance (on the log odds scale) of true odds models with the explained variance based on effect sizes obtained from single SNP association tests. Although many measures of explained variance exist for logistic regression, we adopted McKelvey-Zavoina's $\text{pseudo}-R^2$ [22], as it is defined on the log odds scale and closely mirrors the explained variance of continuous traits [23](see section A3.4 in Appendix S1 for more details on McKelvey-Zavoina's $\text{pseudo}-R^2$).

McKelvey-Zavoina's $\text{pseudo}-R^2$ strongly depends on the effect size of risk alleles and the genetic variance in risk SNPs. Therefore even true odds models show little explained variance in case of small effect sizes or low minor allele frequencies (e.g., $p_a = 0.01$). Except for diseases with rare causal variants, true models with moderate to large effect sizes explain more than 80% of total variance, approaching 100% for very large effect sizes, indicating large heritability. However, odds models based on

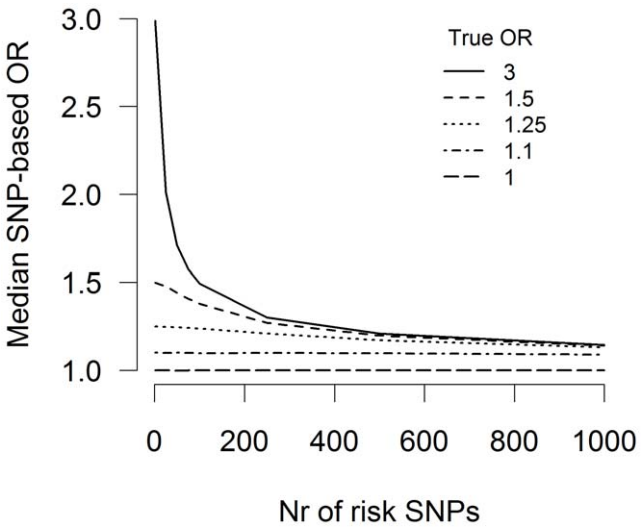

Figure 3. Number of risk SNPs. Effect of total number of risk SNPs on median SNP-based odds ratio (OR_e) for different true odds ratios (OR_t). An allele frequency of 0.25 for risk alleles and a prevalence of 1% is assumed. Simulation is based on a sample of 3500 subjects and a 1:1 case:control ratio. Median is based on 10,000 case-control samples.
doi:10.1371/journal.pone.0027964.g003

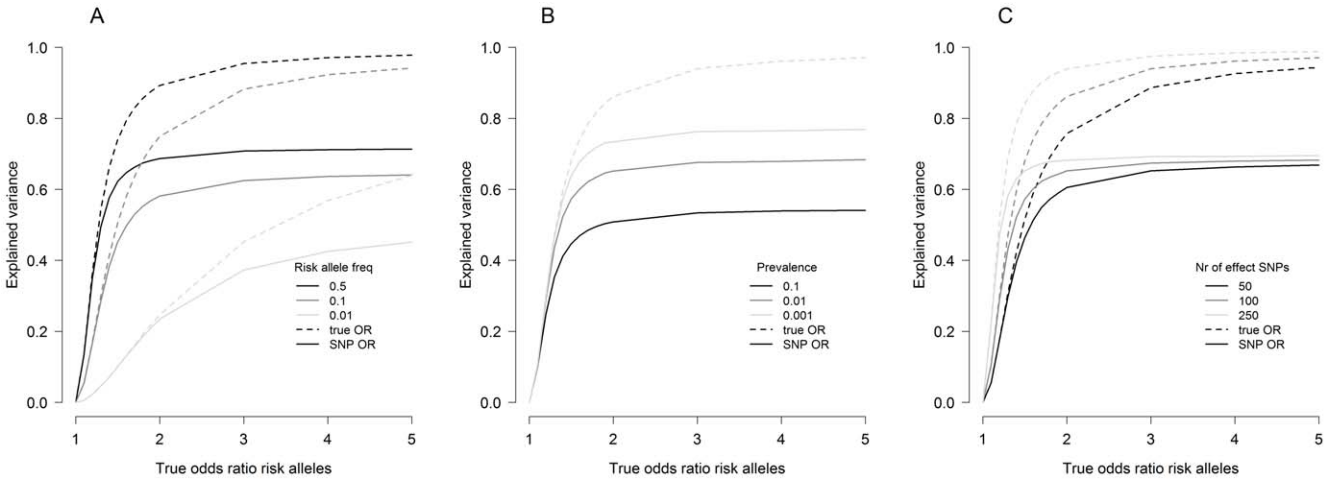

Figure 4. Explained Variance. McKelvey-Zavoina's pseudo $-R^2$ (on log odds scale) as a function of true effect size for an odds model with true odds ratio (dashed line) and an odds model with median odds ratio obtained by single SNP analyses (solid line) for (A) different risk allele frequencies, (B) different prevalences, and (C) different number of effect SNPs. Unless stated otherwise models are based on a disease prevalence of 1%, 100 effect SNPs with risk allele frequencies of 0.25, a case-control sample of 3500 subjects and a 1:1 case:control ratio.
doi:10.1371/journal.pone.0027964.g004

underestimated SNP-based odds ratios (OR_e) show a loss in explained variance compared to odds models based on true effect sizes (OR_t). In the unrealistic case of 100% heritability the typical loss of explained variance is around 20%. A more realistic disease with a heritability of 80%, prevalence of 1% and a minor allele frequency of 50%, still results in an expected loss of more than 10% in explained variance (see Figure 4A). Although prevalence does not affect the true heritability (dotted line), it does affect the heritability based on OR_e(solid line) (Figure 4B).

Truly associated SNPs are unknown a priori and effect sizes will be estimated with error. Nonetheless, this analysis shows that even if truly associated SNPs are known and effect sizes are estimated without error, traditional association testing on dichotomous phenotypes can result in a significant loss of explained variance.

The previous results were all based on the assumption of fixed effect size and allele frequency. Figure 5 shows odds ratios estimated with single SNP analysis, using a more realistically simulated data set in which absolute effect sizes are exponentially distributed and minor allele frequencies are uniformly distributed. Moderate and large odds ratios are underestimated and the underestimation effect increases with effect size. For example the highest risk SNP with a true (conditional) odds ratio of 4.74 has a marginal odds ratio of 4.36, resulting in underestimation of 9% on the odds scale. As expected, odds ratios close to one do not show underestimation. Similar to the naive disease model results, increasing the average true odds ratio, the number of effect SNPs, or the prevalence further increases the underestimation effect (data not shown).

Discussion

Summarizing, our analysis shows a fundamental limitation of applying single SNP association tests to dichotomous phenotypes. Single SNP tests can severely underestimate moderate and large effect sizes for diseases with numerous risk SNPs due to non-collapsibility of the odds ratio. Therefore the marginal odds ratios obtained by single SNP tests can be smaller than the true conditional odds ratios. This underestimation reduces the explained variance and hence contributes to the missing heritability. Underestimation is most pronounced in diseases with high-risk SNPs (i.e., mean $OR > 1.25$), common affect SNPs (i.e., $MAF > 0.1$), a large number of risk SNPs (i.e., 100 or more) and high prevalence ($>10\%$).

Our results are consistent with empirical findings in the GWAS literature. Odds ratios reported in GWA studies are generally small[8]. For example, a recent GWA study reported 57 regions outside the major histocompatibility complex associated with multiple sclerosis, none of which had an odds ratio much higher than 1.5 (see Figure 2 in[24]). Although occasionally large effect

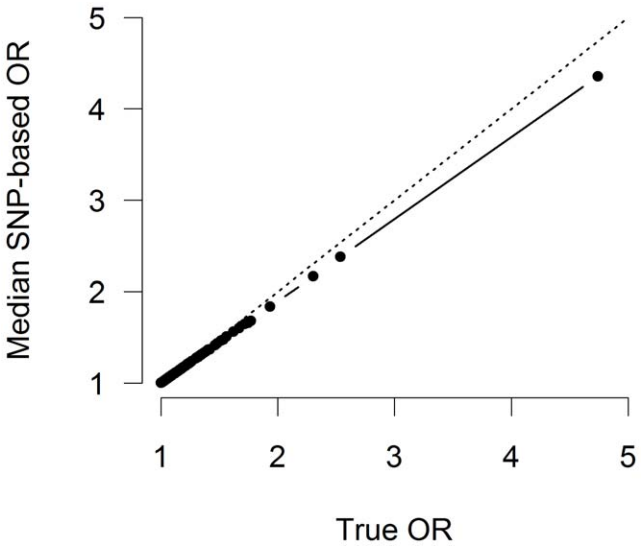

Figure 5. Varying effect sizes. Relationship between median estimated SNP-based odds ratio (OR_e) and true conditional model odds ratio (OR_t) for a disease with 100 effect SNPs and a disease prevalence of 1%. Effect sizes on log odds scale were drawn once for each SNP from an exponential distribution with rate parameter 5 and fixed for all 10,000 case-control simulations. Similarly, allele frequencies were drawn once for each SNP from a uniform distribution between 0.05 and 0.95 and fixed for all case-control simulations. Case-control simulation was based on a sample of 5000 subjects and a 1:1 case:control ratio.
doi:10.1371/journal.pone.0027964.g005

sizes have been reported, numerous common high-risk SNPs have not been identified for a single dichotomous trait. Searching the GWAS catalogue (http://www.genome.gov/gwastudies; accessed August 24, 2011) for SNPs with $OR>4$ and $p<10^{-8}$, shows that no single study reports a disease that is influenced by two or more common SNPs with $OR>4$. Diseases for which high odds ratios are reported for common SNPs (with minor allele frequency in controls >0.05) include auto-immune diseases such as type I diabetes ($OR=8.3$ and $OR=5.49$)[25,26] and ciliac disease ($OR=7.04$)[27]. These high-risk SNPs are part of the major histocompatibility complex.

Single SNP analysis cannot identify large effect sizes of numerous risk SNPs, even if many high risk SNPs would exist. This scenario is mostly of theoretical interest though, as research on quantitative traits, which are not affected by non-collapsibility, suggests that numerous high risk SNPs are not likely in practice. However, conditional odds ratios are likely to be larger than the marginal odds ratios commonly reported. The significance thresholds for marginal and conditional odds ratio are equal as both odds ratios are equivalent in case of no effect[15]. That is, under the null distribution underestimation is not an issue. Therefore, underestimation impedes the identification of SNPs above the significance threshold with underestimated values below the significance threshold.

GWA studies of diseases with high prevalence have reported less significantly associated genetic variants than similar studies of diseases with low prevalence. For example, GWA studies of major depression disorder, which has a life time prevalence of 15%, have reported no associations that reached genome-wide significance or have been solidly replicated[7,28]. On the other hand, studies of schizophrenia and bipolar disorder, which have life time prevalences of 1% or less, have reported several SNPs that did reach genome-wide significance and/or were replicated[7]. There are likely to be many factors contributing to the differential success of GWAS for psychiatric disorders. For example, a lower heritability for depression compared to schizophrenia could imply smaller effect sizes under an architecture of the same number of causal variants, hence requiring larger sample sizes to achieve the necessary power to detect variants that explain the same proportion of variance. Nonetheless, the empirical data are consistent with our result that the underestimation of effect size is larger and the explained variance in liability is lower for complex diseases with high prevalence compared to diseases with low prevalence.

The underestimation effect due to non-collapsibility has important implications for GWA studies of complex diseases. An important aim of GWA analyses is to select truly associated SNPs for use in subsequent analyses and to identify causal variants[19,29]. For selection purposes moderate underestimation of effect sizes need not be a problem, if sample sizes are large enough. However, underestimation of effect size requires larger sample sizes to identify both truly associated SNPs and causal variants. One solution to avoid underestimation of true effect sizes is to analyze continuous instead of dichotomous phenotypes, if available. Continuous phenotypes can usually be modeled with linear regression and under an additive genetic model SNPs are independent and single SNP association tests will not result in underestimation. The use of continuous phenotypes is consistent with the quest for endophenotypes for complex (psychiatric) diseases[30]. Another solution is to estimate effect sizes of all SNPs simultaneously rather than individually. It is for example feasible to estimate the effect sizes of more than 100,000 SNPs in a single analysis[31]. Based on the results of Robinson et al.[11], we expect that a multi SNP analysis is more powerful than a single SNP analysis in the context of a complex disease. Methods for estimating aggregate statistics such as explained variance, total number of risk SNPs, and average effect size of risk SNPs, which analyze all SNPs simultaneously, also exist[32–34]. Even in the context of continuous traits it might be beneficial to opt for multi SNP analysis, as adding covariates can reduce the standard error of the estimates, requiring a smaller sample size to achieve significance.

There are some limitations to our analysis. First of all, our conclusions are conditional on simple model assumptions. However, simpe assumptions do underscore the fundamental nature of the underestimation effect. A second limitation is that we have not proved that effect sizes reported in traditional GWA studies are indeed underestimated. Biases such as the winner's curse could also result in overestimation[35,36]. The winner's curse refers to the fact that due to stringent multiple testing correction it is likely that the first significant finding of a SNP will have a larger effect size than subsequent independent replications. It is therefore unclear whether in practice reported odds ratios are overestimated or underestimated. The major difference between the underestimation effect we discuss and the winner's curse bias, is that the latter will decrease as the sample size increases, whereas non-collapsibility results in a fundamental underestimation that is not affected by sample size. Finally, although we show that underestimation can partly explain missing heritability, this effect could be modest. Continuous traits such as human height are not affected by the underestimation effect, but also show missing heritability[37].

In conclusion, single SNP association testing on dichotomous phenotypes can be problematic. Our analysis implies that odds ratios typically reported in GWA studies [8] could be underestimates of the true conditional odds ratios. We argue that asymptotic underestimation is a serious draw-back, as it cannot be remedied by increasing sample size. We therefore recommend analyzing all SNPs simultaneously. As a variety of multi SNP methods have been proposed in the literature, we are currently comparing the performance of several of those on real GWAS data.

Supporting Information

Appendix S1 Supplemental Appendix.
(PDF)

Acknowledgments

We would like to thank Peter Visscher and Michael Goddard for providing comments on an earlier version of this paper and the anonymous reviewers for their suggestions.

Author Contributions

Wrote the paper: SS NRW RSK EMD.

References

1. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389.
2. Ku CS, Loy EY, Pawitan Y, Chia KS (2010) The pursuit of genome-wide association studies: where are we now? J Hum Genet 55: 195–206.
3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.
4. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 42: 1118–1125.
5. Cardno AG, Gottesman II (2000) Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. Am J Med Genet 97: 12–17.

6. Sullivan PF, Kendler KS, Neale MC (2003) Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch Gen Psychiat 60: 1187–1192.
7. Visscher PM, Goddard ME, Derks EM, Wray NR (2011) Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. Mol Psychiat (advance online publication). pp 1–12.
8. 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 106: 9362–9367.
9. Zondervan KT, Cardon LR (2004) The complex interplay among factors that inuence allelic association. Nat Rev Genet 5: 89–100.
10. Gail MH, Wieand S, Piantadosi S (1984) Biased Estimates of Treatment Effect in Randomized Experiments with Nonlinear Regressions and Omitted Covariates. Biometrika 71: 431–444.
11. Robinson LD, Jewell NP (1991) Some surprising results about covariate adjustment in logistic regression models. Int Stat Rev 58: 227–240.
12. Neuhaus JM, Jewell NP (1993) A geometric approach to assess bias due to omitted covariates in generalized linear models. Biometrika 80: 807–815.
13. Simpson EH (1951) The interpretation of interaction in contingency tables. J R Statist Soc B 13: 238–241.
14. Hernán M, Clayton D, Keiding N (2011) The Simpson's paradox unraveled. Int J Epidemiol. pp 1–6.
15. Greenland S, Robins JM, Pearl J (1999) Confounding and collapsibility in causal inference. Stat Sci 14: 29–46.
16. Guo J, Geng Z (1995) Collapsibility of logistic regression coefficients. J R Statist Soc B 57: 263–267.
17. Janssens ACJW, Aulchenko YS, Elefante S, Borsboom GJJM, Steyerberg EW, et al. (2006) Pre- dictive testing for complex diseases using multiple genes: fact or fiction? Genet Med 8: 395–400.
18. Wray NR, Goddard ME, Visscher PM (2007) Prediction of individual genetic risk to disease from genome-wide association studies. Genom Res 17: 1520–1528.
19. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748–752.
20. Flint J, Mackay TFC (2009) Genetic architecture of quantitative traits in mice, ies, and humans. Genom Res 19: 723–733.
21. Wray NR, Goddard ME, Visscher PM (2008) Prediction of individual genetic risk of complex disease. Curr Opin Genet Dev 18: 257–263.
22. McKelvey RD, Zavoina W (1975) A statistical model for the analysis of ordinal level dependent variables. J Math Sociol 4: 103–120.
23. DeMaris A (2002) Explained variance in logistic regression: A Monte Carlo study of proposed measures. Sociol Methods Res 31: 27–74.
24. Sawcer S, Hellenthal G, Pirinen M, Spencer CCA, Patsopoulos NA, et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476: 214–219.
25. Hakonarson H, Grant SFA, Bradfield JP, Marchand L, Kim CE, et al. (2007) A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature 448: 591–594.
26. Cardon L, Craddock N, Deloukas P, Duncanson A, Kwiatkowski D, et al. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
27. van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, et al. (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet 39: 827–829.
28. Wray NR, Pergadia ML, Blackwood DHR, Penninx BWJH, Gordon SD, et al. (2010) Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. Mol Psychiatry 37: 1–13.
29. Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet 86: 6–22.
30. Gottesman I, Gould T (2003) The endophenotype concept in psychiatry: etymology and strategic intentions. Am J Psyciat 160: 636–645.
31. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet 4: e1000130.
32. 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.
33. Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet 10: 681–690.
34. Wei YC, Wen SH, Chen PC, Wang CH, Hsiao CK (2010) A simple Bayesian mixture model with a hybrid procedure for genome-wide association studies. Eur J Hum Genet 18: 942–947.
35. Kraft P (2008) Curses-winner's and otherwise-in genetic epidemiology. Epidemiology 19: 649–651.
36. Garner C (2007) Upward bias in odds ratio estimates from genome-wide association studies. Genet Epidemiol 31: 288–295.
37. 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.