Quantifying the Underestimation of Relative Risks from Genome-Wide Association Studies

Chris Spencer[1]*, Eliana Hechter[2], Damjan Vukcevic[1], Peter Donnelly[1,2]*

1 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, **2** Department of Statistics, University of Oxford, Oxford, United Kingdom

Abstract

Genome-wide association studies (GWAS) have identified hundreds of associated loci across many common diseases. Most risk variants identified by GWAS will merely be tags for as-yet-unknown causal variants. It is therefore possible that identification of the causal variant, by fine mapping, will identify alleles with larger effects on genetic risk than those currently estimated from GWAS replication studies. We show that under plausible assumptions, whilst the majority of the per-allele relative risks (RR) estimated from GWAS data will be close to the true risk at the causal variant, some could be considerable underestimates. For example, for an estimated RR in the range 1.2–1.3, there is approximately a 38% chance that it exceeds 1.4 and a 10% chance that it is over 2. We show how these probabilities can vary depending on the true effects associated with low-frequency variants and on the minor allele frequency (MAF) of the most associated SNP. We investigate the consequences of the underestimation of effect sizes for predictions of an individual's disease risk and interpret our results for the design of fine mapping experiments. Although these effects mean that the amount of heritability explained by known GWAS loci is expected to be larger than current projections, this increase is likely to explain a relatively small amount of the so-called "missing" heritability.

Citation: Spencer C, Hechter E, Vukcevic D, Donnelly P (2011) Quantifying the Underestimation of Relative Risks from Genome-Wide Association Studies. PLoS Genet 7(3): e1001337. doi:10.1371/journal.pgen.1001337

Editor: Sebastian Zollner, University of Michigan, United States of America

Received June 20, 2010; **Accepted** February 15, 2011; **Published** March 17, 2011

Funding: This work was supported in part by the Wellcome Trust grants [085475/Z/08/Z], [075491/Z/04] (CS, DV, PD), the Rhodes Trust (EH), and Commonwealth Scholarship and Fellowship Plan (DV). PD is a Wolfson-Royal Society Merit Award holder and CS is a Scientific Leadership Fellow in the Nuffield Department of Medicine at the University of Oxford. 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: chris.spencer@well.ox.ac.uk (CS); peter.donnelly@well.ox.ac.uk (PD)

Introduction

Genome-wide association studies (GWAS) have been extremely successful across many diseases in identifying loci harbouring genetic variants that affect disease susceptibility. Virtually all associated variants identified from GWAS to date have relatively small effects: each additional copy of the risk allele typically increases disease risk by 10%–30% (see for example [1]). It has become clear that the variants discovered thus far account for only a small proportion of the genetic basis of each of the diseases, and there has been considerable speculation about where the "missing" heritability might lie [1].

One of several important factors in the success of the GWAS design has been the pattern of linkage disequilibrium in human populations. The strong correlations between nearby SNPs mean that commercially available genotyping chips, which assay 300,000–1,000,000 SNPs, can capture much of the common variation in the human genome, particularly in Caucasian populations [2]. Because genotypes at the causative loci will often be correlated with those at SNPs that are typed on the genotyping chip, it is typically not necessary to assay the true causative variant directly in order to detect a genetic association with disease.

While linkage disequilibrium is extremely helpful for GWAS discovery, the downside is that in most reported regions of association, the true causal variant or variants remain unknown. Therefore it is possible that many of the associated SNPs are only surrogates for the true causal variant(s). When it comes to quantifying the genetic effect, the genotype at the reported SNP acts as a noisy measurement of the genotype at the causal variant. This noise can dilute the apparent strength of the effect, and obscure the true relationship between genotype and phenotype. As we progress towards the identification of the causal variants, estimates of effect sizes for associated loci will thus tend to increase. In turn, the proportion of disease susceptibility explained by GWAS loci will also increase. Thus in addition to other plausible sources, such as secondary signals in GWAS loci, rare variants ($<1\%$ frequency), copy number polymorphisms, and epigenetic effects, some of the missing heritability is actually contained in loci already identified by GWAS, and is driven by common variation ($>1\%$ frequency).

In this paper we use an extensive simulation study to investigate, and quantify, this phenomenon. We show that estimates of the size of the genetic effect based on the best SNP from the GWAS genotyping chip can often closely approximate the effect size at the true causal SNP. In some cases the causal SNP has a large effect and is poorly tagged, leading to substantial underestimation of the true effect size. We investigate how much of the "missing" heritability could thus be hidden in reported GWAS loci, under several sets of assumptions about the nature of the effects at true causal SNPs. Our results also inform the design and value of fine mapping experiments in GWAS loci.

Results

Patterns of linkage disequilibrium (LD) in human populations are complicated, and preclude analytical results, so we adopted a

Author Summary

Genome-wide association studies (GWAS) exploit the correlation in genetic diversity along chromosomes in order to detect effects on disease risk without having to type causal loci directly. The inevitable downside of this approach is that, when the correlation between the marker and the causal variant is imperfect, the risk associated with carrying the predisposing allele is diluted and its effect is underestimated. Using simulations, where we know the true risk at the causal locus, we quantify the extent of this underestimation. We show that, for loci which have a modest effect on disease risk and are common in the population, the risk estimated from the most associated SNP is very close to the truth approximately two thirds of the time. Although the extent of the underestimation depends on assumptions about the frequency and strength of the risk allele, we predict that fine mapping of GWAS loci will, in rare cases, identify causal variants with considerably higher risk. Using three common diseases as examples, we investigate the expected cumulative effects of underestimation at multiple loci on our ability to stratify individuals by disease risk and to explain disease heritability.

simulation approach (see Methods for details). We describe the approach informally before describing our results. First, we chose each allele at each SNP in the HapMap ENCODE regions in turn, assuming it to be causative with a given effect size. We then used a previously reported simulation scheme (HAPGEN, [3]) to simulate a large population of chromosomes with European ancestry, whose patterns of LD match those in the CEU HapMap analysis panel. From this population a case-control sample is taken, with the controls sampled randomly from the population and the cases chosen by oversampling chromosomes carrying the causal allele in the appropriate way given its frequency and assumed effect size. To simulate a GWAS, we considered samples of 2000 cases and 2000 controls typed on the Affymetrix GeneChip Human Mapping 500K Array Set (see www.affymetrix.com). No single sample size can model all reported GWAS, but this size is typical of many. (Later, when considering associated loci from specific diseases that have been studied extensively, we simulated GWAS of larger size.) To simulate a GWAS on a particular commercial chip, we examined data at only those SNPs on the chip in question and checked to see whether any of these SNPs showed a p-value for association $<10^{-6}$. If this occurred we then modelled a replication study, using an additional 2000 cases and 2000 controls for definiteness. We took the best SNP from the simulated GWAS and examined it in the simulated replication sample to check whether it had $p<0.01$ in this replication sample. In what follows we only considered those simulations where the best SNP on the genotyping chip met both these criteria, as these model the ascertainment implicit in reported GWAS associations. For these simulations, we estimated the effect size at this associated SNP, which we call the *hit SNP*, in the replication datasets and compared it with the true effect size at the causal variant used for the simulation. The fact that we estimate the effect size from the replication data set is important, because it minimises the effect of "winner's curse", which would otherwise lead to the effect sizes being over estimated [4]. Simulated GWAS and replication samples were generated for a range of assumed true effect sizes.

Reported genome-wide association studies differ in many particular details, including the choice of genotyping chip used and the sizes of the discovery and replication samples. Specific assumptions are necessary for any simulation study, and ours aim to capture the general features of many reported studies. Investigation of different simulation scenarios, including different genotyping chips and sample sizes, did not change the broad conclusions that follow (data not shown).

Effect size estimates

To begin, we compare the estimated effect size at the replicated hit SNP with the true effect size at the causal SNP in the simulation. Figure 1 illustrates this comparison for three different values of the true effect size. For each we see a peak of estimates around the true effect size assumed at the causal SNP. But note also that there is often underestimation of the true effect size (mean estimated effect size 1.24, 1.86 and 3.32 for true relative risk of 1.25, 2 and 4 respectively), and that this underestimation can be substantial when the true effect is large. For example, when the true relative risk is 4, the estimated effect size was less than two in 12% of simulations of successful GWAS discovery of the effect.

In Figure 2 we plot the relative under- (or over-) estimation of the effect size (estimated effect size divided by true effect size) as a function of the correlation (as measured by the r^2 which is the

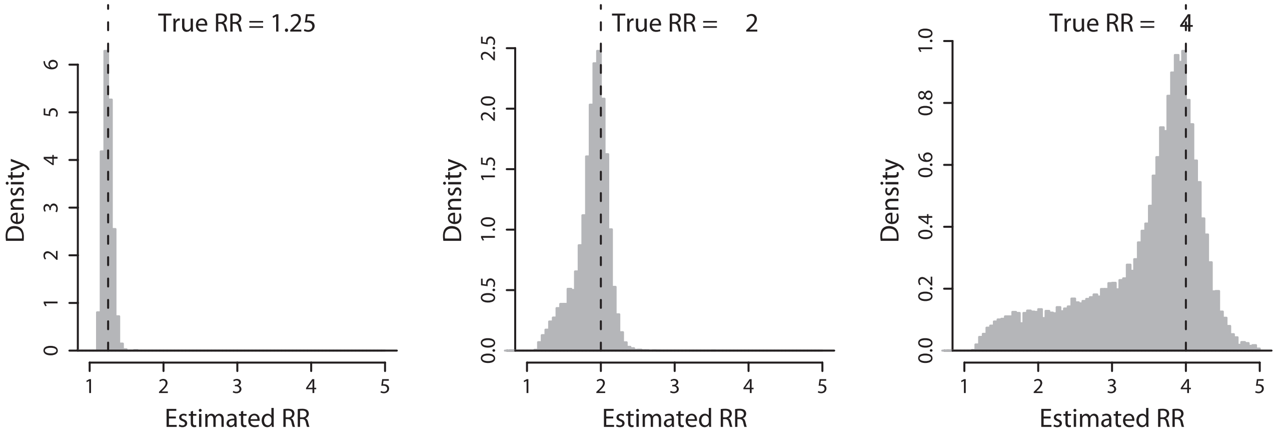

Figure 1. Distribution of estimated effect sizes. Histograms of estimated relative risks (RR), for three different true relative risks indicated by a vertical dashed line in each plot. Histograms include all simulations where the most associated (hit) SNP was significant in both the initial study and the replication study.
doi:10.1371/journal.pgen.1001337.g001

square of Pearson's correlation coefficient) between the hit SNP and the true causal variant. The underestimation is seen to be due to imperfect tagging: when the true causal variant is not well tagged by SNPs on the genotyping chip (the correlation is weak), the estimated effect at the hit SNP is often much lower than the true effect. Conversely, when the causal SNP is well tagged by a SNP on the chip, the estimated effects cluster around the true effect size. Note that while underestimation decreases as the correlation between the hit SNP and the causal SNP increases, there remains systematic underestimation even when the hit SNP has $r^2 \approx 0.8$ with the causative SNP. For example in one third of simulations when the true effect is two, the estimated effect will be under 1.8. Note also that when the true effect size is large, significant and replicable associations can be detected when the best tag SNP only has $r^2 \approx 0.2$ with the causal variant (Figure 2, relative risk = 4).

Imperfect tagging and an ascertainment effect also explain the feature of the plots whereby the underestimation is much less for smaller true effect sizes. If the true effect is small and the true causal variant is not well-tagged on the genotyping chip, there will not be enough power for the GWAS and subsequent replication to reach significance [5], with the result that the corresponding simulation will not contribute to the plot. But if the true effect is large there may still be power to see a significant result when the true variant is not well tagged, so the simulation contributes to the plot and shows the underestimation. Put another way, if the true effect is small, it will only be detected in an association study if the causal SNP is well tagged, and in this case the effect size will be estimated reasonably well. This second ascertainment effect explains the lack of underestimation at hit SNPs not strongly correlated to the causal SNP in the left panel of the Figure 2. Lastly, as low frequency SNPs are less well tagged by other SNPs [6], the extent of the underestimation also depends on the frequency of the risk allele (see Figure S1). Interestingly, the effect sizes at rare alleles are underestimated to a great extent, but only when the true effect size is large enough for the tag SNP of a rare allele to be detected and replicated in the simulated GWAS.

What true effects might underlie the effects estimated from GWAS?

The results above describe the distribution of estimated effect sizes as a function of known true effect sizes and the frequency of the risk allele. In practice we are actually interested in the reverse question, namely what true effect sizes are plausible in the light of the effect size actually estimated from a GWAS and follow-up study? We will see that this requires assumptions about the true distribution of effect sizes. Indeed, writing RR for relative risk, and RAF (risk allele frequency) for the allele frequency at the risk allele, application of Bayes' theorem gives

$$\begin{aligned}&\Pr(\mathit{true\ RR\ and\ RAF}|\mathit{observed\ RR\ and\ RAF}) \propto \\ &\Pr(\mathit{observed\ RR\ and\ RAF}|\mathit{true\ RR\ and\ RAF}) \times \Pr(\mathit{true\ RR\ and\ RAF}),\end{aligned} \tag{1}$$

where "true" refers to the value at the causal SNP and "observed" refers to the value at the hit SNP. Our simulation study allows us to estimate the first factor on the right hand side of (1), and we do so by discretising both the observed and true RR and RAF and creating a matrix of counts based on our simulations over the ENCODE regions. The second factor on the right hand side of (1) is the assumed joint distribution of true risk allele frequencies and effect sizes, which is of course unknown.

We proceed by making two different sets of assumptions about these unknowns. In each case we assume that the distribution of risk allele frequencies is given by the empirical distribution of allele frequencies in the ENCODE regions. In effect this assumes that any SNP variant is, *a priori*, equally likely to affect disease status. What differs between the sets of assumptions is the assumed effect size of a particular variant. Our first set of assumptions posits that the distribution of effect sizes is the same for all putative causal variants, regardless of their allele frequency, and that effect sizes are close to those observed in GWAS studies. The second set of assumptions explicitly assumes that there might be substantially larger effects at variants with smaller minor allele frequency. These priors are described in detail in the Methods section.

Different sets of assumptions about true effect sizes and risk allele frequencies necessarily lead to different conclusions, and it is impossible to study all possibilities. A number of theoretical analyses [7,8,9,10] have argued for a relationship between effect size, disease model, and minor allele frequency (MAF). As there is no consensus on the exact form and extent of the relationship we do not rely on them explicitly here, and instead our approach aims to capture two different perspectives on unknown effect sizes, with the subsequent analyses indicating a range of possibilities. The first perspective is that the range of true effect sizes will be close to those estimated from current GWAS. The second captures the

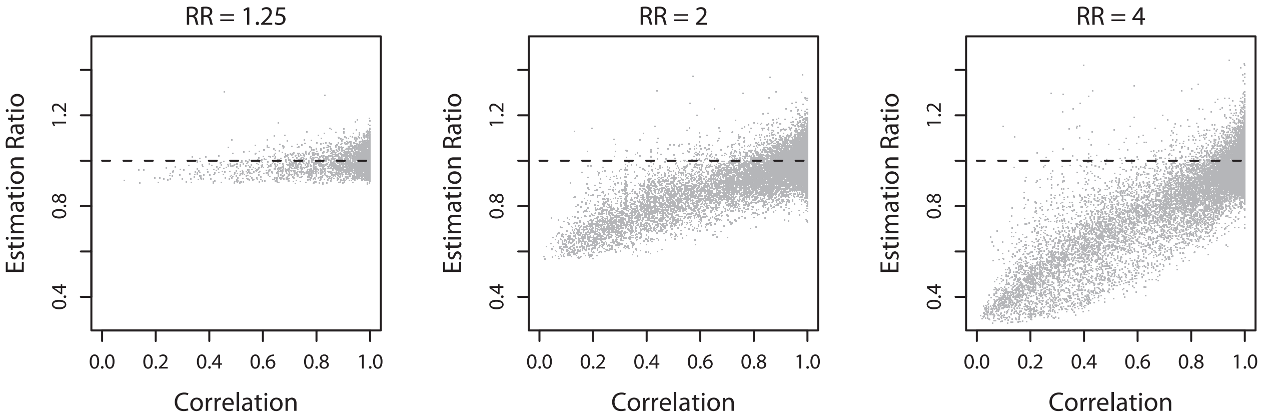

Figure 2. Relationship between underestimation and correlation. Scatter plots of the ratio of estimated to true relative risk against the correlation (r^2) between the disease SNP and the hit SNP, for three different true relative risks. The horizontal dashed line indicates a ratio of 1. Points below the line are under-estimated, and above are over-estimated. (Note that in the case where the hit SNP is also the causal SNP the correlation is 1.)
doi:10.1371/journal.pgen.1001337.g002

possibility that low-frequency variants may have considerably larger effect sizes.

Under either set of assumptions, we can use our simulation study, and Bayes' Theorem (1) to estimate the conditional distribution of true effect sizes and risk allele frequency (RAF) in the light of the observed data at the GWAS hit SNP. Figure 3 illustrates this, showing estimates of the posterior distribution of the true effect size conditional on observing a risk estimate between 1.2 and 1.3, for different observed risk allele frequencies, and under the two different prior assumptions on effect size distributions.

A common feature of the histograms in Figure 3 is that the mode of the posterior distribution on the true effect size is on, or very closes, to the observed estimate. That is, current estimates from GWAS studies of effect sizes from a common SNP, in the range 1.2–1.3 are most likely to be very close to truth. As expected, estimated effects within this range are more likely to be 1.3 than 1.2, because larger effects are more likely to generate a signal of association strong enough to pass the p-value thresholds commonly implemented in GWAS. This explains the left hand tail of the distributions represented in Figure 3.

Figure 3 also shows that there is some probability that the effect size at the causal variant is greater than estimated from the most associated SNP. Interestingly, the observed risk allele frequency impacts our posterior belief about the true effect size, under either set of prior assumptions, with underestimation be more marked when the risk allele at the hit SNP is rarer. Under the conservative prior, when the risk allele at the hit SNP has less than 20% frequency in the control population, the probability that the relative risk is above 1.325 is 55%, compared to 35% when the risk allele frequency is between 20–50%. The corresponding numbers for the *MAF-dependent* prior are 77% and 49%. There are several different phenomena at work here. If the hit SNP is the causal SNP then, assuming that the association is strong enough to be detected and replicated in the GWAS, there is no systematic under estimation (and very little over estimation as we assume the effect size is estimated from the replication sample). However, conditional on the hit SNP not being causal, the distribution of LD with true causal SNP, and therefore the propensity for under estimation, depends on its allele frequency. The posterior distribution on the true effect size given the observed frequency and effect of the hit SNP can be viewed as a mixture of these two scenarios, weighted by their conditional probability. Rarer SNPs are less likely to be tagged well by single markers, and as noted above, poor tagging leads to underestimation of effect sizes. In contrast, for a common SNP, the associated allele is more likely to be well correlated with the causal allele, so there is relatively less under estimation. Under the *MAF-dependent* prior, when the associated allele is low-frequency the causative allele will tend to be low-frequency as well, and so potentially of larger effect. In the scenario where we believe in larger effects at rare causal alleles and have observed a SNP with low RAF with estimated relative risk between 1.2 and 1.3 there is a 24% chance that the source of the signal is a variant which actually doubles or more than doubles risk with each copy of the risk allele.

Our observations are similar when the observed risk allele is the most common allele in the population (RAF>50%) and therefore

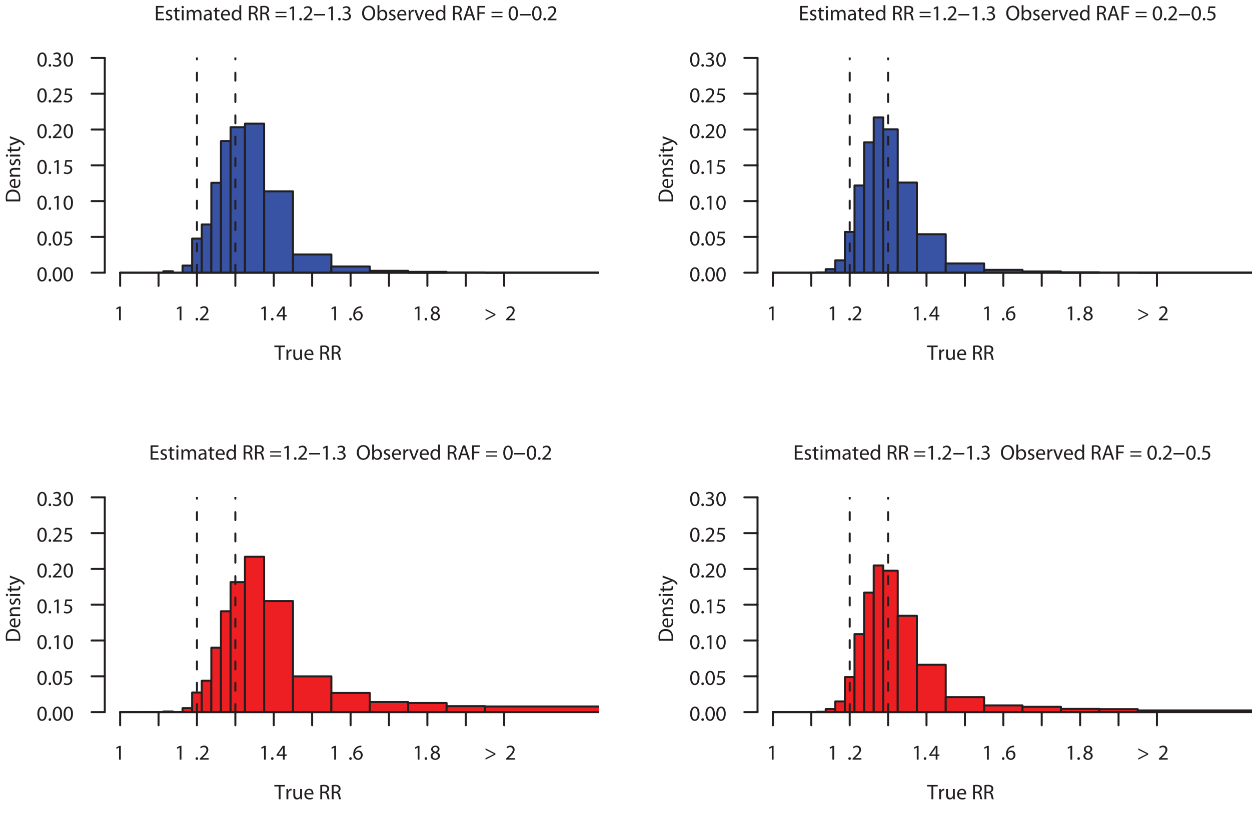

Figure 3. Posterior distribution on true relative risk. Histograms showing the posterior distribution on the true relative risk (RR) conditional on observing an estimated relative risk in the range 1.2–1.3 (vertical dashed lines). Left hand plots condition on the observed risk allele frequency (RAF) being between 20 and 50%, while the right hand plots condition on a RAF less than 20%. Results are shown using two different priors on RR and RAF: the blue histograms are the posterior distribution obtained using a *conservative* prior, and the red histograms are the posterior distribution obtained using the *MAF-dependent* prior.
doi:10.1371/journal.pgen.1001337.g003

the minor allele is protective (Figure S2). Qualitatively, the same conclusions also apply when the estimated effect size at the hit SNP is weaker, for example in the range 1.05 to 1.2 (Figure S3).

Consequences for individual disease risk

One consequence of the potential underestimation of effect sizes from GWAS findings is that as we move to better identification of the actual causal variants, through fine mapping and/or functional studies of associated regions, our estimates of their effect sizes might well increase. Assuming a multiplicative model of risk across loci, these small expected changes could combine to increase the relative risk of disease in those individuals with highest genetic risk of disease.

To investigate this, we simulated genotypes at known associated loci in a population of individuals (assuming Hardy Weinberg equilibrium and no linkage disequilibrium across loci) for each of breast cancer, type 2 diabetes and Crohn's disease, based on reported risk allele frequencies [11,12,13] (see Tables S3, S4, S5 for a list of loci). First treating the causal loci and relative risks for each disease as given by current GWAS estimates, we measured the average risk of individuals in the top x%, by risk, of the population (for differing values of x) and compared this to the mean risk in the population. We then repeated this simulation, allowing for the uncertainty in the estimation of true effect sizes by averaging over the uncertainty in both the RAF and effect size of the causal variant on the basis of the posterior distributions of these, given the GWAS findings, under the two priors described above. We assumed that risks combined multiplicatively across loci. For *NOD2* and *IL23R* in Crohn's disease where the causal variant is thought to be known, here and below, we used the effect sizes for the known variant, and did not average over uncertainty in these. Because all three diseases have been extensively studied, we approximated the GWAS discovery process as corresponding to a GWAS discovery sample of 5000 cases and 5000 controls, and a replication sample of 10,000 cases and controls. The actual discovery process for each of the diseases is complicated, often involving meta-analysis and/or multistage discovery, and not straightforward to model accurately, but the approach we use should capture the fact that GWAS-discovery were ascertained through study of large numbers of samples.

The results of the three simulations are given in Table 1.The unadjusted simulations give estimates of how much more at risk individuals with the greatest genetic propensity to disease are, based only on GWAS loci, relative to the average person in the population. As expected, the fold change in risk of individuals carrying a large fraction of risk variants is dependent on the number and magnitude of known loci. For example, individuals in the top 0.1% of risk for Crohn's disease are 20 times more likely

Table 1. Adjusted estimates of individual risks.

	Type II Diabetes						
x%	**50**	**25**	**10**	**5**	**1**	**0.5**	**0.1**
Unadjusted	1.31	1.58	1.92	2.17	2.73	2.96	3.49
	(1.3–1.32)	(1.56–1.6)	(1.89–1.95)	(2.12–2.21)	(2.64–2.82)	(2.85–3.1)	(3.25–3.8)
Conservative	1.42	1.83	2.36	2.78	3.8	4.26	5.38
	(1.36–1.52)	(1.69–2.05)	(2.12–2.77)	(2.44–3.36)	(3.19–4.9)	(3.52–5.62)	(4.25–7.58)
MAF-dependent	1.54	2.12	3.05	3.97	6.7	8.16	12.43
	(1.41–1.74)	(1.81–2.7)	(2.34–4.74)	(2.76–6.99)	(3.75–15.08)	(4.21–20.64)	(5.26–43.06)
	Crohn's Disease						
x%	**50**	**25**	**10**	**5**	**1**	**0.5**	**0.1**
Unadjusted	1.67	2.41	3.66	4.89	9.25	11.91	20.07
	(1.64–1.7)	(2.35–2.48)	(3.53–3.79)	(4.63–5.13)	(8.39–10.31)	(10.5–13.74)	(15.88–27.04)
Conservative	1.78	2.71	4.28	5.82	11.2	14.56	25.41
	(1.71–1.9)	(2.52–3)	(3.86–4.9)	(5.18–6.81)	(9.54–13.67)	(12.04–18.24)	(18.84–36.27)
MAF-dependent	2	3.36	6.01	8.87	19.77	27.19	54.11
	(1.82–2.27)	(2.83–4.26)	(4.56–8.69)	(6.29–14.13)	(12.32–38.88)	(15.91–58.84)	(26.59–150.26)
	Breast Cancer						
x%	**50**	**25**	**10**	**5**	**1**	**0.5**	**0.1**
Unadjusted	1.2	1.38	1.58	1.71	2.02	2.14	2.4
	(1.19–1.22)	(1.36–1.4)	(1.55–1.6)	(1.68–1.74)	(1.96–2.05)	(2.05–2.19)	(2.27–2.55)
Conservative	1.25	1.46	1.71	1.89	2.29	2.46	2.82
	(1.2–1.33)	(1.37–1.62)	(1.57–1.99)	(1.71–2.27)	(2.02–2.92)	(2.13–3.21)	(2.39–3.9)
MAF-dependent	1.29	1.59	2.14	2.74	4.06	4.69	6.41
	(1.2–1.5)	(1.4–2.25)	(1.64–4.25)	(1.8–6.09)	(2.14–11.12)	(2.28–16)	(2.58–36.14)

The median (and 95% confidence interval) of the increase in risk of individuals who are in the top x% of risk, relative to the average, for three common diseases. Values are estimated (using 100,000 simulations of a population of 10,000 individuals) from a set of replicated associations (see Tables S2, S3, S4) using published point estimates of the risk attributed to carrying a copy of the risk allele and its frequency in population controls. Unadjusted values use these numbers directly in the simulation of individual risks. For the adjusted values we simulate the relative risk and risk allele frequency from its posterior distribution using two different priors: conservative and *MAF-dependent* (see text).

doi:10.1371/journal.pgen.1001337.t001

than the average person to develop the condition, whereas for breast cancer, where the number of common loci and associated relative risks is typically smaller, the equivalent number is just over two-fold.

The second and third simulations attempt to average over the possible outcomes of our future efforts to map causal mutations, to reveal the likely gains in our ability to stratify individuals on the basis of risk. These use the methodology above, under both prior distributions, to average over the posterior distribution of the allele frequency and effect size at the causal SNPs underlying reported GWAS loci for the three diseases. These adjusted estimates are also shown in Table 1. Across diseases we see that there is a significant increase in the risk associated with carrying multiple risk variants. In particular we see that the biggest differences in risk are for those individuals in the extreme tail. It is these individuals who carry the stronger, likely rarer, risk alleles which are currently insufficiently characterised by the most significant signal of association in some regions identified to be important in disease. For example, the risk of an individual in the top 0.1% of the population for genetic risk typed at the *causal* loci underlying currently known GWAS loci will likely be increased by a factor of 3–6.5, 5–12, or 25–50, compared to an average individual, for breast cancer, type 2 diabetes and Crohn's disease. These are notably greater increases in risk than current prediction based in the hit SNPs from GWAS loci which would be 2.4, 3.5 and 20 respectively.

Missing heritability

We have shown above that as we move to identification of the true causal variants underlying GWAS associations, through fine mapping and functional studies, their effect sizes will tend to increase, in a minority of cases substantially, compared to current estimates from GWAS. This will, in turn, increase the amount of heritability explained by these diseases. We can use the approach developed here to try to quantify this effect.

We investigated this question in the context of the three diseases just described, namely breast cancer, type 2 diabetes, and Crohn's disease. For each disease we took the set of hit SNPs from published associated loci [11,12,13] (see Tables S3, S4, S5), and for our two prior distributions on effect sizes we estimated the posterior distribution of both the effect size and the allele frequency for the causal SNP at each locus, as described in the previous section. One commonly used measure of heritability is sibling recurrence risk ratio, often denoted by λ_S: the relative increase in risk to an individual if their sibling has the disease compared to the baseline risk in the population as a whole [14]. Assuming, as is usual for heritability calculations [15], that there is no interaction between loci, λ_S can be calculated as a function of the risk allele frequency and effect size for each causal variant. In order to allow for the uncertainty in the allele frequency and likely underestimation of the effect size at the causal variants underlying GWAS associations, we averaged this expression over the posterior distribution of these quantities, given the GWAS findings (see Methods for details).

The results are shown in Figure 4. For each disease they show that the heritability due to already identified GWAS loci will be higher than current estimates, under either set of assumptions about true effect sizes, but particularly under the *MAF-dependent* prior. Whereas at the time of writing the current estimates of the contribution to λ_S from GWAS loci are 1.03, 1.08, and 1.49 for breast cancer, type 2 diabetes, and Crohn's disease, these may well be 1.06, 1.14, and 1.61 (mean under the conservative prior) and they could plausibly be as high as 1.21, 1.39 and 2.46 (mean under the *MAF-dependent* prior). Whilst some of the "missing" heritability

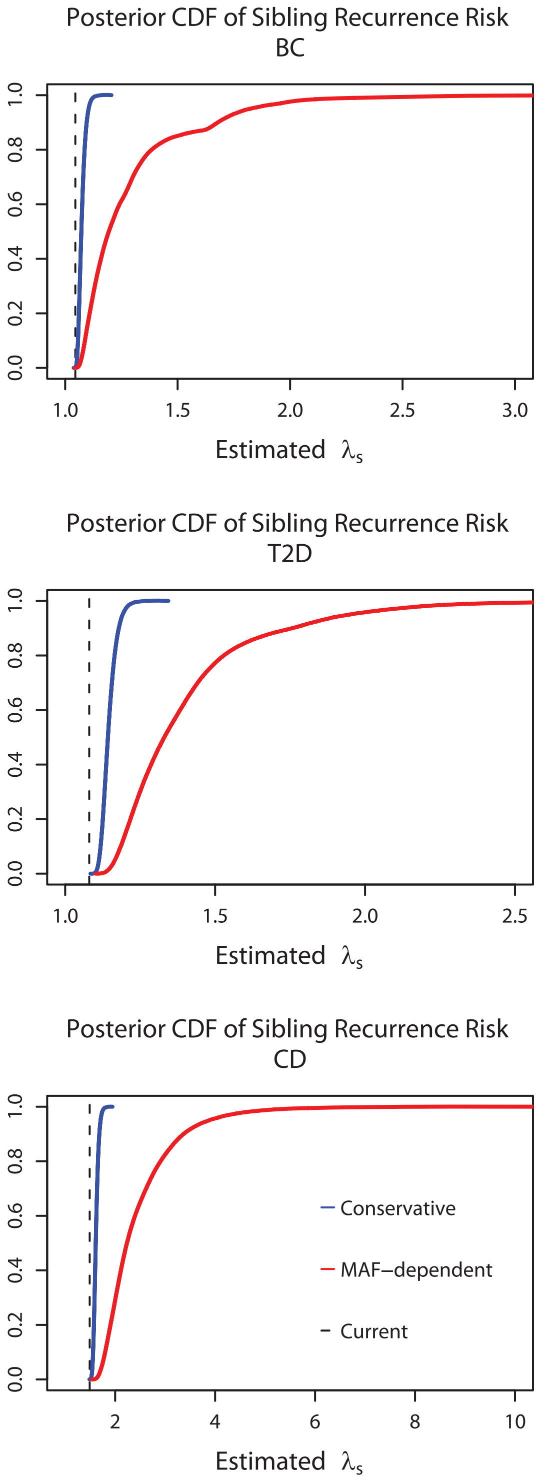

Figure 4. Adjusted estimates of explained heritability. Cumulative density functions of the posterior distribution of estimated sibling recurrence risk ratio (estimated λ_S) in breast cancer (BC), Type 2 diabetes (T2D), and Crohn's disease (CD) under the conservative and *MAF-dependent* priors. The dotted line indicates the λ_S based on replicated loci in each of the three diseases. The quoted values of sibling recurrence risk ratio (lower right hand corner) are from references given in main text. Posterior distributions were calculated assuming a GWAS of 5,000 cases and controls and a replication study of 10,000 cases and controls.
doi:10.1371/journal.pgen.1001337.g004

is thus disguised rather than missing, we note that this effect is unlikely to account for the extent of the gap between estimates of sibling relative risk (2, 1.8, and 10, respectively, from family studies [16,17,18]) and those explained by currently known loci. We return below to a discussion of the discrepancy.

Discussion

The correlation between alleles along the human genome has allowed GWAS to look for regions associated with disease without having to either genotype all known genetic variation or guess *a priori* which regions of the genome may be important. Although this approach has been a significant success, there is a predictable downside of using a subset of variation to tag, or predict, untyped diversity: for the vast majority of the SNPs identified as mediating disease risk, we are left uncertain as to whether they are causally involved in the pathway from genotype to phenotype, or, much more plausibly, are just a surrogate for the causal variation.

GWAS associations will thus typically relate to a noisy measurement of the causal variant. One consequence of this is that the size of the genetic effect associated with GWAS loci may be underestimated. We quantified this through an extensive simulation study designed to mimic patterns of linkage disequilibrium in European Caucasian populations. We draw two broad conclusions from these analyses. Firstly, a significant proportion of estimated relative risks will be biased downwards because the hit SNP is a powerful, but imperfect, tag for the true causal variation. In most cases this effect will be relatively minor, but in some instances, the best associated SNP may actually be a poor predictor of a, putatively rarer, SNP with a much larger effect, in which case the effect size estimated from the GWAS finding will substantially underestimate the true effect size.

The exact proportion of reported associations which fall into these two categories depends on properties of the design of the study from which the SNP was identified, and on one's belief about how likely low frequency (>1%) variants of large effect are to cause common diseases. The statistical power afforded by any particular association strategy sets a lower limit on the size of effect that can be under-estimated because an imperfect tag of an allele with a small effect size will simply fail to achieve genome-wide significance. Other properties of GWAS strategy, such as sample ancestry and the number of markers typed, also change our interpretation of observed effect sizes because they influence the distribution of linkage disequilibrium between putative hit SNPs and causal variants.

Our findings show that at any particular locus, especially if the associated SNP has a low MAF, the true effect could be quite large. But we would not expect this to be widespread. Were many true effects this large it would be extremely surprising for so few of them to have been observed: although any one such causal SNP may not be well tagged on the genotyping chips used for GWAS, some of them will happen to be at least moderately well tagged, and their detection would lead to much larger estimates than have been seen from current studies. In the context of this study these early observations suggest that, of the two prior distributions we investigated, it is the conservative prior that may better reflect the true distribution of effect sizes attributed to low and common frequency variants.

One way of viewing the posterior distribution on the true effects shown in Figure 3 is as a probability distribution on the outcome of efforts to fine map current regions of association. In this light, our results inform questions of the design and value of fine mapping experiments. First, simulations similar to those described above (assuming causal variation to be distributed like SNPs in ENCODE regions) suggest that less than 8% of the time will the hit SNP actually *be* the causal SNP. We note that there may be more reward in terms of gains in predictive ability and increases in effect size from fine mapping SNPs with lower minor allele frequency because they are, on average, more likely to be in poor LD with an unobserved causal variant. On the other hand, our simulations show that although they are unlikely to be causal, most common hit SNPs are likely to be very good surrogates markers for their causal variant. Indeed, in 25% of cases, the hit SNP will be a near-perfect surrogate (ie r^2>0.99) for the causal variant. Should this be the case, further genotyping will not reveal other SNPs with stronger associations, unless sample sizes are extremely large.

Here we have quantified the increased spread of genetic risk with genotypes just at known loci, and only considering a multiplicative disease model. But even in this restricted setting, there will be substantial differences in risk between high- and low-risk groups based on these genetic factors. For example the propensity of individuals in the top 0.1% of the population distribution of genetic risk of type 2 diabetes will be increased by a factor of 5–10, compared to the average. For breast cancer, in the analogous top-risk group this risk will be increased by a factor of 3–5 (on the basis of common variation). Importantly, with the growth of GWAS findings, both in terms of numbers of diseases and numbers of loci for particular diseases, more and more of the population will be in this most at risk category for at least one disease: assuming 100 independent diseases, nearly 10% of the population will be in the top 0.1% of risk of at least one disease. Knowing which individuals these are and what diseases they are most at risk of is therefore potentially useful information, both to the individual and at the population level. The issues involved in utilising such information in screening programmes (discussed for example in [13]) are complicated, but our results strengthen the arguments for consideration of this possibility.

We have shown that some of the "missing" heritability for common disease actually resides in known GWAS loci and have estimated this deficit for three particular diseases. While rather more heritability is likely to be explained by known GWAS loci than has been reported, this effect alone falls well short of explaining all the missing heritability. Note, however, that there are other reasons why existing loci may explain more heritability than currently thought. Current calculations (by others, and above) focus on a single causal variant in each associated region: more variants within regions will explain more heritability. They also ignore possible non-multiplicative disease effects, and also ignore interactions between variants at different loci. Power to detect either is low [19], so it is misleading to put much weight on the failure of existing designs to find such effects. As others have noted [20], parts of the missing heritability could be due to multiple rare variants of large effect, associations with other forms of genetic variation such as copy number polymorphisms, and epigenetic effects. Indeed it would be surprising if each did not play some role. Another possibility is that estimates of the "genetic" component of disease susceptibility, from epidemiological studies, confound shared environment with shared DNA, and so inflate heritability estimates [21,22].

Methods

Choice of genomic regions

In order to model the signal of association generated by disease-causing mutations, we chose to simulate data exploiting empirical surveys of human diversity. For this purpose we used data from the 10 ENCODE regions [23] within the CEU analysis panel of HapMap II [5], which have undergone SNP ascertainment by resequencing 48 individuals of diverse ancestry. These regions therefore show a fuller spectrum of SNPs than are represented in the HapMap data at large, and haplotypes are expected to be accurate due to the trio design of the CEU HapMap panel [24]. The regions over which we simulate data are centred on each of the 10 ENCODE regions (listed in Table S1) and include 500kb of flanking HapMap variation at the boundaries of each region.

Simulation of population data

As the typical sample size of most GWAS is much larger than the number of CEU HapMap individuals, we simulated 100,000 chromosomes using the HAPGEN software package. These 100,000 haplotypes we call the *reference* panel. GWAS case and control samples were then subsampled from the reference panel, as described below. HAPGEN uses a population genetic model that incorporates the processes of mutation and fine-scale recombination to generate individuals from an existing set of known haplotypes. We ran HAPGEN with an effective population size of 11418 (as recommended for the CEU population), a population scaled mutation rate of 1 per SNP, a population scaled recombination rate from estimates described in [25], with the known set of haplotypes taken from the CEU analysis panel of HapMap II as described above (see http://www.stats.ox.ac.uk/~marchini/software/gwas/hapgen.html).

Generating a case-control sample

For SNPs greater than 1% in frequency in the ENCODE regions we performed two hypothetical GWAS by letting each of the two alleles be causal in turn. We denote the causal allele by A and the protective allele by a. To generate the control sample we sampled the required number of haplotypes, without replacement, from the reference panel and combined these in pairs to form diploid individuals. This mimics the common use of population controls, rather than controls explicitly chosen for not having the disease under study. For the case sample, we sampled pairs of haplotypes from the reference panel according to the genotype frequencies at the causal SNP dictated by the assumed disease model: If δ is the risk of the AA genotype, and α is the risk of the Aa genotype, both relative to the aa genotype, then we sample case individuals (without replacement) on the basis of their genotypes at the SNP assumed to be causal with success probabilities proportional to:

$$p(AA)\propto\delta f^2 \quad p(Aa)\propto 2\alpha f(1-f) \quad p(aa)\propto(1-f)^2, \qquad (2)$$

where f is the frequency of the risk allele A in the reference panel. Throughout, for definiteness, we adopted a multiplicative model for disease risk (additive on the log scale) defined by $\delta=\alpha^2$. We refer to α as the relative risk (RR) or effect size associated with the causal variant. To approximate a GWAS, we thinned the generated data set to include only those SNPs present on the Affymetrix 500K array that had a minor allele frequency in sampled controls of greater than 1%. This set may or may not include the assumed causal SNP.

For analyses involving only simulated data, we sampled 2,000 cases and 2,000 controls from the reference panel to emulate a typical large GWAS. For the subsequent analyses of heritability and individual risk profiling for type 2 diabetes, breast cancer and Crohn's disease that studied particular reported associations, we simulated 5,000 cases and 5,000 controls to obtain results more comparable to the size of study from which the associations were ascertained. We simulated under a range of relative risks at 24 grid points from 1.05 to 6. In attempting to simulate the signal of disease at rare alleles (1% to 5%) in a GWAS of 5000 cases and controls there were a small number of simulations in which there were insufficient haplotypes in our reference panel to generate the required number of genotypes at the causal SNP for large effect sizes. These simulations were discarded, but as the numbers were small (3% when the RR = 4 and 11% when RR = 6) we do not believe this greatly affects the results presented below.

Testing for association

Following common practice, for each simulated case control sample, we tested for association between genotype and case control status using the Cochran Armitage trend test [26] at each SNP with frequency greater than 1% in the simulated panel of chromosomes. We calculated the p-value of this test statistic which is χ^2 distributed with 1 degree of freedom under the null hypothesis of no association. If any test across the region obtained a p-value$<10^{-6}$ the location of the most significant SNP (termed the *hit SNP*) was recorded and we simulated this SNP in an independent replication sample.

Simulating the replication processes

We simulated the replication experiment in three stages. First we simulated the frequency of the causal allele in cases and controls in the replication population. We then simulated the frequency of the hit SNP conditional on the frequency of the causal allele. Finally, we simulated the genotype counts for a sample of cases and controls in this replication population.

We motivated sampling of the frequency of the causal allele in controls in the replication population by thinking of the replication sample as an additional sample from the same population as the original GWAS sample. (Other assumptions are possible here, but seem unlikely to affect the main conclusions.) Specifically, we placed a uniform prior distribution on the unobserved population frequency and sampled a value, f', from the posterior distribution of this frequency given the data in the reference panel. (Given the large size of the reference panel, the frequency in the replication sample will be very close to that in the reference panel.) Conditional on f', the population replication frequency in cases was calculated from equation (2). To obtain the replication population frequencies at the hit SNP we estimated the conditional distribution in the reference sample of alleles at the hit SNP in each of cases and controls, given those at the causal SNP, and used these for the replication sample. This corresponds to assuming that the LD between the causal and hit SNP in the replication sample will be the same as that in the reference sample. Finally, conditional on the population replication frequencies in cases and controls, we take multinomial samples of the required size to mimic the replication case and control samples. A test of association using the trend test was performed at the hit SNP on the simulated replication samples and deemed a significant replication if the p-value was less than 10^{-2}.

Estimating effect sizes

We estimate the effect size, or relative risk, α, at the hit SNP by maximum likelihood under the model described above by equation (2). For studies with population controls this can be achieved in practice by fitting a logistic regression model for case status [27].

Priors on effect size

We implement two different sets of prior assumption on the effect size and its relationship with minor allele frequency. Our first set of assumptions is that if α is the effect size at a causal variant, then log(α) is normally distributed with mean 0 and standard deviation 0.2, independent of RAF. We refer to this as the *conservative* prior, since it places little weight on relative risks greater than 1.5. To get a sense for this distribution, it assumes that 81% of true effect sizes are less than 1.3 with 96% less than 1.5, and 99.9% less than 2. A further discussion of the choice of prior on effect sizes can be found in [19] and [28].

Our second set of assumptions, which we call the *MAF-dependent* prior, again assumes a normal distribution for log(α) with mean 0, but here the standard deviation, σ, is allowed to depend on the RAF. The dependence of the distribution of the effect size on allele frequency has no theoretical justification, but is chosen on pragmatic grounds to give a gradual increase in the average effect size as the alleles at causal SNP become rarer in the population. It is implemented by increasing σ by a weight defined by an exponential density with parameters chosen such that, when the RAF is near 0.5 (a common SNP), this prior is approximately the same as the conservative prior, with $\sigma = 0.2$. As the RAF approaches 0 or 1 (corresponding to rarer SNPs), then considerably more weight is put on larger RRs. See Figure S4 and Table S2 for details. For example, when the MAF is less than 5% the second prior gives an approximately 45% chance that the risk associated with each copy of the causal allele is larger than 2.5. Note that we used an empirical prior on the frequency of the risk allele (Figure S5) by choosing each allele, at each SNP, with in the ENCODE region to be causal in turn.

Estimating heritability

A commonly used measure of heritability is based on considering the risk of disease to an individual j conditional on them having an affected ($Y = 1$) sibling k relative to the unconditional probability (which is just the prevalence of the disease):

$$\lambda_s = \frac{\Pr(Y_j = 1 | Y_k = 1)}{\Pr(Y_j = 1)} = \frac{\Pr(Y_j = 1, Y_k = 1)}{\Pr(Y_k = 1)} \times \frac{1}{\Pr(Y_j = 1)}$$

We can calculate the above, using assuming that $\Pr(Y_k = 1) = \Pr(Y_j = 1)$, and by summing over the genotypes $g \in \{0,1,2\}$ of the siblings and of the mother M and father F (see [15]):

$$\frac{\sum_{g_M} \sum_{g_F} \sum_{g_j} \sum_{g_k} \Pr(Y_j = 1 | g_j) \Pr(Y_k = 1 | g_k) \Pr(g_j | g_M, g_F) \Pr(g_k | g_M, g_F) \Pr(g_M) \Pr(g_F)}{(\sum_g \Pr(Y = 1 | g) \Pr(g))^2}$$

If we divide through by the square of the risk associated with most protective genotype (which we can define to be $g = 0$) then we can write the above in terms of the per allele relative risk α, and assume the genotype probabilities follow Hardy-Weinberg equilibrium with risk allele frequency f as above:

$$\Pr(Y = 1 | g_0) / \Pr(Y = 1 | g_0) = 1,\ \Pr(Y = 1 | g_1) / \Pr(Y = 1 | g_0) = \alpha,$$

$$\Pr(Y = 1 | g_2) / \Pr(Y = 1 | g_0) = \alpha^2$$

$$\Pr(g_0) = (1 - f)^2, \quad \Pr(g_1) = 2f \times (1 - f), \quad \Pr(g_2) = f^2$$

By making the further assumption that loci are independent an estimate of the heritability explained by a set of hit SNPs can be obtained by multiplying together the λ_S values calculated at each individual locus. We calculated sibling relative risk in this manner using estimates of RR and RAF of replicated loci from studies of Type 2 diabetes, Crohn's disease and breast cancer (see Tables S3, S4, S5).

We then simulated 100,000 times from the posterior of true RR and RAF of each locus conditional upon the reported RR and RAF, using the simulation approach and the two different priors as described in the paper. For each set of simulations, for each disease, we recalculated λ_S at each locus and multiplied over loci, giving a sample from the posterior distribution of sibling risk that could be explained by the current set of report loci if the causal loci where typed directly.

Supporting Information

Figure S1 Average relative underestimation of effect size as a function of allele frequency and true effect size. Line plot shows the mean ratio of the effect size estimated from the most associated GWAS SNP to the true effect size at the causal locus. Lines are shown for 4 different risk allele frequency (RAF) bins.
Found at: doi:10.1371/journal.pgen.1001337.s001 (0.01 MB EPS)

Figure S2 Posterior distribution on true relative risk when the minor allele is protective. Histograms showing the posterior distribution on the true relative risk (RR) conditional on observing an estimated relative risk and risk allele frequency (RAF) at the hit SNP. Results are shown using two different priors on RR and MAF: the blue histograms are the posterior distribution obtained using a conservative prior, and the red histograms are the posterior distribution obtained using the MAF-dependent prior.
Found at: doi:10.1371/journal.pgen.1001337.s002 (0.92 MB EPS)

Figure S3 Posterior distribution on true relative risk for low estimated effect sizes. Histograms showing the posterior distribution on the true relative risk (RR) conditional on observing an estimated relative risk and risk allele frequency (RAF) at the hit SNP. Results are shown using two different priors on RR and MAF: the blue histograms are the posterior distribution obtained using a conservative prior, and the red histograms are the posterior distribution obtained using the MAF-dependent prior.
Found at: doi:10.1371/journal.pgen.1001337.s003 (0.92 MB EPS)

Figure S4 Priors on relative risks. Probability distributions of relative risk as a function of minor allele frequency for the MAF-dependent and conservative priors (the blue line). The MAF-dependent prior is pictured for five values of MAF
Found at: doi:10.1371/journal.pgen.1001337.s004 (3.14 MB EPS)

Figure S5 Empirical prior on risk allele frequency. Cumulative distribution of the frequency of SNPs within the ENCODE regions used for simulations. At each SNP, each allele is chosen is turn chosen to be the risk allele so the distribution is symmetric around a half.
Found at: doi:10.1371/journal.pgen.1001337.s005 (3.14 MB EPS)

Table S1 ENCODE regions used in simulations. The build 35 coordinates of the regions of HapMap CEU data used by HapGen to simulate genome-wide association study data. When simulating haplotypes and testing for association a 500kb buffer window was included either side of the listed regions.
Found at: doi:10.1371/journal.pgen.1001337.s006 (0.03 MB DOC)

Table S2 MAF dependent prior. Table of the standard deviation of the prior distribution of the log relative risk (RR) as a function of risk allele frequency (RAF).
Found at: doi:10.1371/journal.pgen.1001337.s007 (0.06 MB DOC)

Table S3 SNPs for Type 2 diabetes. (See main text for reference).
Found at: doi:10.1371/journal.pgen.1001337.s008 (0.04 MB DOC)

Table S4 Replicated SNPs for Crohn's disease. (See main text for reference).
Found at: doi:10.1371/journal.pgen.1001337.s009 (0.05 MB DOC)

Table S5 Replicated SNPs for breast cancer. (See main text for reference).
Found at: doi:10.1371/journal.pgen.1001337.s010 (0.03 MB DOC)

Acknowledgments

We thank Rory Bowden, Gil McVean, and Simon Myers for helpful discussion.

Author Contributions

Conceived and designed the experiments: CS DV PD. Performed the experiments: CS EH DV. Analyzed the data: CS EH DV PD. Contributed reagents/materials/analysis tools: CS. Wrote the paper: CS PD.

References

1. Manolio TA, Brooks LD, Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118: 1590–1605.
2. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.
3. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39: 906–913.
4. Zollner S, Pritchard JK (2007) Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am J Hum Genet 80: 605–615.
5. Iles MM (2008) What can genome-wide association studies tell us about the genetics of common disease? PLoS Genet 4: e33. doi:10.1371/journal.pgen.0040033.
6. (2005) A haplotype map of the human genome. Nature 437: 1299–1320.
7. Wakefield J (2008) Bayes factors for genome-wide association studies: comparison with P-values. Genet Epidemiol 33: 79–86.
8. Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6: 109–118.
9. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69: 124–137.
10. Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. Nat Rev Genet 5: 89–100.
11. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40: 955–962.
12. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40: 638–645.
13. Pharoah PD, Antoniou AC, Easton DF, Ponder BA (2008) Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med 358: 2796–2803.
14. Risch N (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46: 229–241.
15. Thomas DC (2004) Statistical methods in genetic epidemiology. Oxford: Oxford University Press. xxi, 435 p.
16. Pharoah PD, Day NE, Duffy S, Easton DF, Ponder BA (1997) Family history and the risk of breast cancer: a systematic review and meta-analysis. Int J Cancer 71: 800–809.
17. Weijnen CF, Rich SS, Meigs JB, Krolewski AS, Warram JH (2002) Risk of diabetes in siblings of index cases with Type 2 diabetes: implications for genetic studies. Diabet Med 19: 41–50.
18. Orholm M, Munkholm P, Langholz E, Nielsen OH, Sorensen TI, et al. (1991) Familial occurrence of inflammatory bowel disease. N Engl J Med 324: 84–88.
19. The 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.
20. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18–21.
21. Clayton DG (2009) Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS Genet 5: e1000540. doi:10.1371/journal.pgen.1000540.
22. Wallace C, Clayton D (2003) Estimating the relative recurrence risk ratio using a global cross-ratio model. Genet Epidemiol 25: 293–302.
23. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636–640.
24. (2003) The International HapMap Project. Nature 426: 789–796.
25. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321–324.
26. Armitage P (1955) Tests for linear trends in proportions and frequencies. Biometrics 11: 375–386.
27. Schouten EG, Dekker JM, Kok FJ, Le Cessie S, Van Houwelingen HC, et al. (1993) Risk ratio and rate ratio estimation in case-cohort designs: hypertension and cardiovascular mortality. Stat Med 12: 1733–1745.
28. Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet 10: 681–690.