

Power of genome-wide association studies in the presence of interacting loci

Journal:	<i>Genetic Epidemiology</i>
Manuscript ID:	GenEpi-06-0164.R2
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	02-Apr-2007
Complete List of Authors:	Pickrell, Joseph; University of Chicago, Human Genetics; INSERM, U535 and Univ Paris-Sud Clerget-Darpoux, Françoise; INSERM, U535 and Univ Paris-Sud Bourgain, Catherine; INSERM, U535 and Univ Paris-Sud
Key Words:	linkage disequilibrium, genome-wide association, tagSNPs



Review

Power of genome-wide association studies in the presence of interacting loci

Joseph Pickrell^{1,2}, Françoise Clerget-Darpoux¹ and Catherine Bourgain¹

1: INSERM, U535 and Univ Paris-Sud, F-94817 Villejuif.

2: Present affiliation : Department of Human Genetics, University of Chicago

For Peer Review

Summary

Though multiple interacting loci are likely involved in the etiology of complex diseases, early genome-wide association studies (GWAS) have depended on the detection of the marginal effects of each locus. Here, we evaluate the power of GWAS in the presence of two linked and potentially associated causal loci for several models of interaction between them and find that interacting loci may give rise to marginal relative risks that are not generally considered in a one-locus model. To derive power under realistic situations, we use empirical data generated by the HapMap ENCODE project for both allele frequencies and LD structure. The power is also evaluated in situations where the causal SNPs may not be genotyped, but rather detected by proxy using a SNP in linkage disequilibrium (LD). A common simplification for such power computations assumes that the sample size necessary to detect the effect at the tSNP is the sample size necessary to detect the causal locus directly divided by the LD measure r^2 between the two. This assumption, which we call the “proportionality assumption”, is a simplification of the many factors that contribute to the strength of association at a marker, and has recently been criticized as unreasonable [Terwilliger and Hiekkalinna 2006], in particular in the presence of interacting and associated loci. We find that this assumption does not introduce much error in single locus models of disease, but may do so in certain two-locus models.

Introduction

Genetic association studies are currently an area of great interest as a tool for studying the genetic basis of human diseases. The association approach, in comparison to genetic linkage studies, has better resolution for the detection of a disease locus, a fact which first made association studies in candidate genes, chosen because of their function or because of their location under linkage peaks, a popular tool for identifying susceptibility alleles. The recent technological progress in SNP genotyping and the availability of resources like the HapMap [2005] have made genome-wide association studies (GWAS) possible. However, aside from the technological feasibility, the power of such a strategy for detecting new susceptibility loci in complex diseases is a key issue. There are many considerations that separate GWAS from association studies in candidate genes: the many thousands SNPs required to efficiently sample the genetic variability of the genome necessarily require a strong correction for multiple testing, large samples are thus necessary to have a chance of detecting the relatively weak signals expected in complex diseases, and risks of subtle hidden population structure are important [Clayton, et al. 2005; Wang, et al. 2005].

Further, though multiple loci are very likely involved in the etiology of complex diseases, the initial applications of the genome-wide association approach have depended on the detection of the marginal effects of each locus. However, complex interactions between loci are likely to be the norm, rather than the exception, in the study of complex diseases [Moore 2003]. Interacting loci need not be different genes or chromosomal regions; both linked and associated loci may contribute to susceptibility to diseases. Ultimately, if we consider any causal polymorphism (rather than an entire gene) as a locus, this situation, known as allelic heterogeneity, has been described in many cases [Cohen, et al. 2004; Hugot, et al. 2001; Li, et al. 2006; Weng, et al. 2005]. For some of them, the effects of particular variants are strong enough to be detected by their marginal effects [Duerr, et al. 2006; Klein, et al. 2005]. However, the variety of models is such that a more systematic power study of GWAS for these models is of interest.

The initial power studies of GWAS considered single locus models, which is equivalent to assuming an independent effect of all causal loci [de Bakker, et al. 2005;

Wang, et al. 2005]. Studies of more complex two-locus models have either limited themselves to particular allele frequencies [Howson, et al. 2005] or to unlinked loci [Evans, et al. 2006; Marchini, et al. 2005]. Here, we evaluate the power of GWAS in the presence of linked and potentially associated loci, considering several models of interaction between two causal SNPs. This complements the previous work by adding another layer of biologically relevant complexity to the disease model.

For a given case-control sample size and type I error, the power to detect a given disease model further depends on the allele frequencies of the two causal SNPs, the LD between them and the correction for multiple testing. To derive power under realistic situations, we use empirical data generated by the HapMap ENCODE project [2005] for both allele frequencies and LD structure. For each 2-locus genetic model, we determine an average power to detect the region, considering each pair of alleles in the ENCODE region ENM010 as causal.

In the course of this study, we evaluated the power in situations where the causal SNPs may not be genotyped, but rather its effect detected by proxy using a SNP in linkage disequilibrium (LD) with the causal SNP, referred to as a tagSNP (tSNP). In many previous power and sample size studies, taking this into account has been simple: the sample size necessary to detect an effect at a marker is assumed to be the sample size necessary to detect the effect directly at the causal locus divided by the LD measure r^2 between the two [Kruglyak 1999; Pritchard and Przeworski 2001; Zondervan and Cardon 2004]. This assumption, which we call the “proportionality assumption”, has made r^2 the statistic of choice in defining tSNPs and judging the coverage of a set of SNPs. Kruglyak [Kruglyak 1999] proposed a cutoff for a “useful” level of LD between a tSNP and a causal SNP at $r^2=0.1$ based on the proportionality assumption, which would suggest a 10X increase in the sample size needed to detect an effect. Currently, it seems that r^2 values superior to 0.8 have become the cutoff for “useful” LD [de Bakker, et al. 2005; Pe'er, et al. 2006]. However, the “proportionality assumption” is a simplification of the many factors that contribute to the strength of association at a marker, and has recently been criticized as unreasonable [Terwilliger and Hiekkalinna 2006], in particular in the presence of interacting and associated loci. While evaluating the power of the association

using tSNPs, we examine the error introduced by this simplification in models with both one and two associated causal loci.

In what follows, we start by presenting the two-locus power study and examine the error in the power computation introduced by the proportionality assumption afterwards.

For Peer Review

Methods and Results

Features of a two-locus disease model

A two-locus disease model can be described with the following genotype relative risk (GRR) matrix:

	<i>A/A</i>	<i>A/a</i>	<i>a/a</i>
<i>B/B</i>	GRR_{AABB}	GRR_{AaBB}	GRR_{aaBB}
<i>B/b</i>	GRR_{AABb}	GRR_{AaBb}	GRR_{aaBb}
<i>b/b</i>	GRR_{AAbb}	GRR_{Aabb}	1

where SNP 1 has alleles *A* and *a* and SNP 2 has alleles *B* and *b* and GRR_{g_i} is the GRR of genotype g_i , as compared to genotype *aabb*.

For our study, we define four GRR matrices to describe two-locus interactions. These are described in Table 1. We label the models in the following way: in the heterogeneity model, the risk for a two locus genotype corresponds to the highest marginal risk associated with the two corresponding one-locus genotypes. Matrix 1 is an example of this type of model with the same recessive effect of alleles within the two locus. In the epistasis model, the effect of a given SNP is only seen in the presence of a certain genotype or genotypes at the other SNP. Matrix 2 is an example of an epistatic model with a recessive effect of alleles within the two loci. In the additive model, the risk increase due to alleles at one locus is the same for all genotypes at the other locus. Matrix 3 is an example of such an additive model with an additive effect of alleles within the two loci and the same effect of the two loci. Finally, in the multiplicative model, the effect of the two SNPs are independent. Risks at the two loci simply multiply. Matrix 4 is an example of a multiplicative model with an additive effect of alleles within the two loci.

The equations for deriving the marginal GRRs at each locus and the sample sizes needed to detect them by single locus association testing are presented in Appendix A. As can be seen from equation (6), the GRR at a given SNP (SNP A, for example) not only depends on the two-locus GRR model, but also on the allele frequencies at both SNP A and SNP B and on the linkage disequilibrium between them. Consequently, a variety of marginal single locus GRR models are associated with each two-locus GRR model.

To illustrate the variation of single-locus GRR models, we present in Table 2 different possible GRRs at a causal locus when the underlying disease model is the two-locus heterogeneity model 1 from Table 1. In this example, the frequency of the causal allele at the first SNP (allele A) is set to 0.05 and the frequency of the SNP allele B and the LD between the two SNPs are allowed to vary, with the constraint that the two causal alleles (A and B) are out of phase (the LD between them is negative, $D < 0$). Under this constraint, the observed single-locus GRR model may be unrelated to the actual underlying disease model. Looking at the extreme case where the frequency of SNP B is 0.7 and $D' = 1$, we see that $GRR_{AA} = 1.3$, while $GRR_{Aa} = 0.65$, which implies an apparent protective effect of the heterozygote. In this case, the apparent protective effect is a consequence of the two causal loci being in strong LD with each other.

Thus, a simple two-locus model of disease can give rise to marginal relative risks outside of those that would be considered in a single-locus model.

Power to detect a two-locus disease model

We sought to estimate the power of a genome-wide association study in the presence of two potentially associated causal SNPs. The factors that influence the power are: the frequencies of the causal SNPs, the LD between them, the two-locus GRR matrix, and the correction for multiple testing. Further, if tSNPs are used, the power will also depend on their allele frequencies and on the LD between the tSNPs and the causal SNPs.

Because genome-wide association scans do not necessarily seek to directly identify the causal SNP, but rather a region or gene worthy of further inspection, we take, for each “full” model (where a “full” model includes a two-locus GRR model and the allele frequencies and LD of the causal SNPs) the maximum power from the two loci (or from the two tSNPs tagging the two causal loci) as the power to detect the region by single locus testing.

Power of allelic and genotypic association tests was calculated using the formulas given in Appendix A, for the different two-locus GRR models presented in Table 1. To consider a large variety of marginal single locus models for each two-locus GRR model, the data on allele frequencies and LD for the two interacting loci were taken from the

empirical data generated by the HapMap ENCODE project. This project provides an exhaustive look at the SNP polymorphisms in ten 500-kb regions in the genome. We use data from the region ENM010 in the European-origin population (CEU), which contains 716 polymorphic SNPs. The ENCODE genotype data was downloaded on 15/2/2006 from the HapMap website (www.hapmap.org). Only independent haplotypes (i.e. from the 60 parents) were used in the analysis. Calculation of linkage disequilibrium was done in Haploview v3.11 [Barrett, et al. 2005]. The distribution of allele frequencies and the LD plot for the region are presented in Figure 1 and Figure 2, respectively. The 1,023,880 possible pairs of SNP alleles in ENM010 (four allele combinations for each pair of SNPs) were considered, in turn, as being causal. Consequently, to each two-locus GRR model we attached a distribution of single locus GRRs and a distribution of risk genotype frequencies reflecting both the frequency and LD distributions of all SNP allele pairs. Median and ranges of single locus GRR distributions arising from the four two-locus GRR models are presented in Table 3.

The ENCODE data were also used to compare the power of different tagging strategies. First, for each tagging strategy (defined by a value of r^2 and a minor allele frequency threshold for the SNPs to be tagged), a tSNP was attached as a proxy to each of the 716 SNPs of the ENM010 region, using Tagger [de Bakker, et al. 2005] in Haploview. For each 2-locus “full” genetic model, the power was computed for the two tSNP proxies for the SNP pair that was causal in this “full” genetic model, using the formulas provided in the Appendix A. Second, to compare the power of different tagging strategy (including no tagging at all when all SNPs are tested), specific multiple testing corrections were derived for each strategy, including no tagging at all. To do so, the number of independent tSNPs (or “effective” number of tSNPs) for a whole genome scan were extrapolated from computation based on the 10 ENCODE region data. As the ENCODE data only represent a discrete sampling of 500kb regions on the genome, simply multiplying the mean number of tSNPs necessary in these 10 regions by the number of 500kb blocks on the genome would not properly account for the redundancy of adjacent regions. To do so, we proceeded as follows.

For any tagging strategy, let the number of tSNPs in a given 500-kb region be N_0 . Let the number of tSNPs selected using only the first 250-kb block and only the second

250-kb block be N_1 and N_2 , respectively. N_0 will always be less than or equal to $N_1 + N_2$ because of the redundancy removed when the region is divided in half, though the precise difference will depend on the LD structure in the region. We assume that the average amount of redundancy lost by splitting each of the 10 ENCODE regions will be a rough estimate of the redundancy lost by taking any discrete sample of the genome. Among the N_0 , N_1 and N_2 SNPs, let the “effective” number of SNPs calculated using the method of Li and Ji [Li and Ji 2005] in each region be M_0 , M_1 , and M_2 . Once again, M_0 will always be less than or equal to $M_1 + M_2$. Let the difference between the M_0 and $(M_1 + M_2)$ be C . If the size of the genome in kilobases is $Gsize$ (which we let be 3,059,200), then an estimation of the effective number of SNPs on the whole genome, T , is

$$T = \frac{Gsize}{500} * \bar{M}_0 - \left(\frac{Gsize}{500} - 1 \right) \bar{C}$$

where the bars represent the mean of the given value over the ten 500-kb regions available. The effective number of SNPs corresponds to the number of independent tests and can then be used in a Bonferroni correction. The effective number of tests for a whole genome search corresponding to different tagging strategies are presented in Table 4.

Figure 3 shows for the four two-locus GRR models, the percentage of “full” genetic models for which association is detectable as a function of sample sizes, with 80% power and a 5% corrected type I error. The calculations were done using both the one degree of freedom allelic test for association and the two degree of freedom genotypic test for association. Curves are shown for two strategies: no tagging and tagging with parameters $r^2 = 0.8$ and $MAF = 0.1$. In all four 2-locus GRR models considered in this study, the difference between the two tagging strategies and the two tests (genotypic and allelic) is minimal. Overall, the proportion of detectable models is remarkably low. With 5000 individuals, a maximum proportion of 60% is observed and it is for the multiplicative two-locus model. For both epistasis and heterogeneity models, if about 45% of the models are detectable with 4000 individuals, the proportion only weakly increases with increasing sample size. Even with 10,000 individuals, only slightly more than 50% of the models, at best, are detectable. For the additive model, the

proportion is smaller than 50% unless samples of 9000 individuals are considered. These three two-locus GRR models generate a high proportion of “full” genetic models that are not detected with single locus association test in conceivable sample sizes. The multiplicative two-locus GRR model is the least sensitive to the SNP frequency and LD distributions, provided that samples large enough are used. With samples of 10,000 individuals, more than 80% of the models are detectable. This multiplicative model also corresponds to the highest mean marginal ORs over all corresponding “full” 2-locus models (see Table 3)

To illustrate the impact of the marginal frequency distribution of the causal SNPs on the power to detect the different two-locus GRR models, we fixed the sample size at 5000 (2500 cases and 2500 controls) and asked which alleles are detected. In Figure 4, we show the percentage of “full” models that are detectable as a function of the allele frequency of the first causal allele. Once again, the difference between not tagging and tagging is minimal. For the multiplicative two-locus model, most of the models involving alleles with a frequency between 0.15 and 0.75, are detectable. The severe drops at both sides of the frequency scale reflect situations where the risk alleles are either too rare or too frequent for reasonable power of marginal association tests. The pattern is very similar for the additive model though the range of frequencies with a reasonable proportion (around 50%) of detectable models is only between 0.2 and 0.7. The pattern is slightly different for the epistasis and heterogeneity models that are clearly asymmetrical. Only models involving alleles with a minimum frequency of at least 0.5 and 0.3 for heterogeneity and epistasis, respectively, have a reasonable detection proportion (around 50%). Alleles with frequency as high as 0.90 still have the same detection proportion. This reflects the fact that these two models have a recessive effect. Unless risk alleles have a moderate frequency, homozygous at-risk genotypes are too rare to allow detection of the association.

Error in the proportionality assumption

In the course of doing the above power study, we examined the best way to determine the power to detect an association using a marker associated with a causal locus, rather than a directly genotyped causal locus. The power of a test for association at

a genotyped tSNP in linkage disequilibrium with an ungenotyped causal allele is a function of the disease model, the allele frequencies of both the tSNP and the causal SNP, and the LD between the two, as can be seen from Equation (5) of Appendix A. However, the relationship is commonly simplified as $N_2 = N_1/r^2$, where N_2 is the sample size (the total combined number of cases and controls) needed to detect the effect by proxy at the tSNP, N_1 is the sample size needed to detect the effect directly at the causal SNP, and r^2 is the measure of LD between the two loci in the general population. We call this the proportionality assumption, which was made explicit by Pritchard and Przeworski [Pritchard and Przeworski 2001].

In Appendix B, we highlight one of the assumptions underlying this demonstration: the allele frequencies of both the tSNP and the causal SNP are the same in the general population as in the combined sample of cases and controls. While this is generally true in a prospective study design or a QTL study, in a case-control study the cases are enriched for the causal allele, and thus, the proportionality assumption will only hold if there is no association between the allele tested and the disease. The difference in allele frequencies due to an association may also imply a difference in LD between the combined sample and the general population, as r^2 is sensitive to allele frequency. If the r^2 in the sample is greater than r^2 in the general population, the simplification will overestimate the sample size necessary to detect the association. In the opposite case, the proportionality assumption will underestimate the sample size necessary to detect the association.

The factors that determine the precise error in the proportionality assumption are the allele frequency of the causal allele, the allele frequency of the tSNP, the LD (as measured by D') between them, and the disease model. We define the error as $\varepsilon = (N_{\text{EXP}} - N_{\text{OBS}})/N_{\text{EXP}}$, where N_{EXP} is the estimated sample size needed to detect the effect of the causal allele with a tSNP using the proportionality assumption and N_{OBS} is the estimated sample size without this assumption, using Equation (7).

We started by examining the impact of different one-locus disease models on the error in the proportionality assumption. Figure 5-Figure 8 display for 4 genetic models (dominant, recessive, additive and multiplicative), the upper and lower bounds on the error of the allelic association test, for all possible values of LD and all possible allele

frequencies at both the tSNP and the causal SNP, summarized by r^2 . Thus, in this figure, a given r^2 value may correspond to a variety of different LD and allele frequency settings. For GRR smaller than 1.5 and in the absence of other causal alleles, the error in the proportionality assumption is around -20% to 20% . This error, however, is smaller for higher values of r^2 , with error values around -5% to 5% for $r^2 > 0.8$.

The presence of an additional causal locus in the region may lead to the violation of another hypothesis underlying the proportionality assumption, which states that the frequencies at the tSNP are independent from the disease status conditional on the causal allele. Indeed, if the tSNP is associated to two causal loci – a plausible situation in our two-locus model- this assumption is violated. This has been highlighted by Terwilliger and Hiekkalinna [Terwilliger and Hiekkalinna 2006], who provide an example in which this non independence leads to the extreme case where either of the two causal alleles would be detected if assayed directly, but neither is detected by proxy, even though r^2 between one causal SNP and the tSNP is 0.33 (a situation where the detection should be possible at the tSNP providing a three-fold increase of the sample size, according to the proportionality assumption). Because their example presents only a single case, in which the LD between the causal SNP and the tSNP is a fairly weak correlation, we take a more extensive look at this issue.

Following Terwilliger and Hiekkalinna [Terwilliger and Hiekkalinna 2006], assume we have two causal SNPs, A and B, and a tSNP C that tags A. First, if there is no LD between A and B, then the power to detect an association at C is entirely dependant on the relationship between A and C. Second, if the disease alleles A and B are in phase, the interaction will either have no effect or increase power; that is, the relative risks at C may be increased due to the additional effect of allele B. Third, if A and B are out of phase ($D < 0$), it is possible that the interaction may confound the ability to detect an effect. To examine this effect, we consider a situation where there is no ancestral recombination in the region (that is, at most three haplotypes are observed for each two locus combination). We let the mode of interaction between A and B be heterogeneity (see Table 1) and we examine the error in the proportionality assumption as a function of r^2 between loci A and C. In this framework, an error of $-\infty$ means that the allele is undetectable using a tSNP. Results are presented in Figure 9, in two different frequency

settings for the two out of phase causal alleles : the two causal alleles have different frequencies in Figure 9a, while in Figure 9b, the two causal alleles have the same frequencies. If the frequency of C is less than that of A, the proportionality assumption always holds. However, if the frequency of allele C is greater than that of allele A (which it tags), the sample size necessary to detect an effect may be greatly misestimated, even for high values of r^2 .

To evaluate more exhaustively the impact of two-locus models on the deviation from the proportionality assumption, we computed the error for all “full” 2 locus-models based on the ENM010 data and tSNPs, considering the set of tSNP selected with $r^2 = 0.8$ and MAF= 0.1. By construction, this set of tSNP has a $r^2=0.8$ only for the SNPs with a MAF>0.1. For all other SNPs in the region, no constraint is set on r^2 . We could thus assess the impact for a relatively wide range of r^2 . This was done for each of the four two-locus models. Figure 10 presents, by bins of r^2 , the median errors and the 90% ranges calculated in each bin. The median error ranges around -20 /+20 % for all disease models and r^2 values. However, even with $r^2=0.8$, situations of strong errors are observed, in particular for the heterogeneity and epistasis models.

Discussion

The resources and technology necessary for GWAS are now available. The effectiveness of the approach in the presence of multiple interacting susceptibility loci, as are expected in complex diseases, is contingent on the detection of the marginal effect at each locus. For this reason, many previously published GWAS power studies focused on single gene models, arguing that multiple loci models would only result in weakened marginal effects.

However, even simple parameters can give rise to marginal relative risks that are not normally considered in a one-locus model. Further, as outlined by Terwilliger and Hiekkalinna [Terwilliger and Hiekkalinna 2006], the presence of two associated causal loci may impact the power of tSNP based strategies.

In the present study, we evaluated the impact of two locus models on the power of GWAS, and we considered the possibility that the two causal loci are associated. By basing our power study on the HapMap ENCODE data, we measured the impact of a large spectrum of two-locus frequency and linkage disequilibrium patterns on power, including the power of realistic tagging strategies.

There has been much debate over the likely frequencies of causal alleles in complex diseases [Lohmueller, et al. 2003; Pritchard 2001; Pritchard and Cox 2002; Reich and Lander 2001]. Because the debate is not resolved, we chose to base our study on the largest set of two-locus allele frequencies and LD patterns. Therefore, we considered each SNP in the ENM010 region to have an equal likelihood of being causal, thus using the distribution of alleles in the genome as the distribution for causal alleles. Our evaluation of the proportion of detectable models should thus be taken as it is: a proportion of models for which association is possibly detected and not a proportion of models for which association will be detected.

For realistic sample sizes of 5000, the proportion of models for which association cannot be detected with 80% power is large, and this is particularly true for the models with interacting loci (all three non-multiplicative models) that we considered, where less than 50% of the models are detectable. The specificity of the no-interaction (multiplicative) two-locus model highlights the importance of specifically considering interaction models and not only considering that extrapolation from one-locus to multiple

loci is straightforward – as it is classically assumed in power studies where only marginal effects of loci are modeled [de Bakker, et al. 2005; Wang, et al. 2005].

By further examining the example provided by Terwilliger and Hiekkalinna [Terwilliger and Hiekkalinna 2006], we confirmed that, in some cases, tagging is ineffective, especially for tSNPs with relatively low LD with the causal variants ($r^2 < 0.6$). This is to say that unless very dense maps (over 250,000 SNPs) are chosen, this inefficacy of tagging is a potential problem. Multipoint methods for the analysis of tagSNPs may be a step towards remedying this [Chapman, et al. 2003]. On the other hand, based on the frequency and LD patterns of the ENM010 region, we found that for tSNPs chosen with high tagging parameters ($r^2 = 0.8$, $MAF = 0.1$), two-locus models are consistent with previous reports based on one-locus models in stating that the use of tSNPs is an efficient way to conserve genotyping costs while maintaining power to detect an effect [de Bakker, et al. 2005]. In all models examined, the loss of power due to the reduced effect size at the tSNP was almost completely countered by the gain in power due to a reduced multiple testing burden. However, this conclusion must be presented with two caveats: first, we used tSNPs from the ENCODE project, which were identified in extensive resequencing data, and are thus the best possible tSNPs in the region. Other regions of the genome are less well-covered by the HapMap. Second, we considered only models with weak relative risks, and therefore the loss of power due to tagging in some models may be minimal because the power even without tagging is close to zero. Many of our models are detected with low power when the causal allele frequency is low; this is also the region of the frequency spectrum where tagging is known to be less effective [de Bakker, et al. 2005].

Our results suggest that approximations due to the proportionality assumption, labelled the “Fundamental Theorem of the HapMap” by Terwilliger and Hiekkalinna [Terwilliger and Hiekkalinna 2006], do not have a major impact in the prediction of power studies in models with a single causative locus. Nevertheless, we would strongly encourage the use of exact power computation in lieu of the proportionality assumption simplification for the study of two-locus models.

Another conclusion of this study is that, the more complex models that we considered in our computation argue in favor of larger sample sizes for GWAS, and

suggest that prohibitive sample sizes might indeed be required in some cases. Our two causal locus model is only a very first step of higher complexity. Multi-locus models, even with non-associated susceptibility loci, may result in peculiar marginal models and very strong multiple locus GRR may result in flat marginal one-locus GRRs [Lesage, et al. 2002]. Further, we have only considered the power to detect SNP variation. Other variants, such as deletions, inversions, and copy number polymorphisms, are also likely to play a role in diseases. Results have suggested that deletions in the genome are correlated with SNPs, and thus are effectively detected by proxy [Hinds, et al. 2006]. However, this may not be true for copy number polymorphisms [Locke, et al. 2006]. But this should not prevent us from making our models more complex despite the fact that they might end up not being the most favorable for currently available methods. .

Appendix A: Relative risk and sample size calculations

Direct test for association. Let us consider a SNP with relative risks GRR_{AA} and GRR_{Aa} for genotypes AA and Aa compared to genotype aa and population genotype frequencies $P(AA)$, $P(Aa)$ and $P(aa)$. The expected frequencies of each genotype in the cases under the alternative hypothesis of association are the following:

$$\begin{aligned} P(AA)_{H1,Cases} &= GRR_{AA} * P(AA) / [GRR_{AA} * P(AA) + GRR_{Aa} * P(Aa) + P(aa)] \\ P(Aa)_{H1,Cases} &= GRR_{Aa} * P(Aa) / [GRR_{AA} * P(AA) + GRR_{Aa} * P(Aa) + P(aa)] \\ P(aa)_{H1,Cases} &= P(aa) / [GRR_{AA} * P(AA) + GRR_{Aa} * P(Aa) + P(aa)] \end{aligned} \quad (1)$$

Assuming the sample size is equally divided between cases and controls and controls are representative of the general population, the expected frequencies in both the cases and the controls under the null hypothesis of no association are:

$$\begin{aligned} P(AA)_{H0} &= [P(AA)_{H1,Cases} + P(AA)]/2 \\ P(Aa)_{H0} &= [P(Aa)_{H1,Cases} + P(Aa)]/2 \\ P(aa)_{H0} &= [P(aa)_{H1,Cases} + P(aa)]/2 \end{aligned} \quad (2)$$

The genotypic test for association follows a chi-squared distribution with 2 df under the null hypothesis, where $GRR_{AA} = GRR_{Aa} = GRR_{aa} = 1$. We use the classical assumption that the test follows a non-central chi-squared distribution under the alternative hypothesis (which holds asymptotically for local alternatives) with non-centrality parameter

$$\lambda = \sum_{g_i} N * \frac{(P(g_i)_{H1,Cases} - P(g_i)_{H0})^2}{P(g_i)_{H0}} \quad (3)$$

where $g_i \in \{AA, Aa, aa\}$, and N is the sample size (cases and controls). For fixed values defining an alternative (GRR_{AA} , GRR_{Aa} , $P(AA)$, $P(Aa)$), the power of the test for a nominal type I error of α is $\beta = 1 - \mathfrak{R}_{\lambda,2}(\kappa_{\alpha,2})$, where $\kappa_{\alpha,2}$ is the upper α^{th} percentile of a chi-squared distribution with 2 df and $\mathfrak{R}_{\lambda,2}(x)$ is the cumulative distribution function of a non-central χ^2 distribution with 2 df with non-centrality parameter λ .

Alternatively, for the 1 df allelic association test,

$$\begin{aligned}
 P(A)_{H1,Cases} &= 2 * P(AA)_{H1,Cases} + P(Aa)_{H1,Cases} \\
 P(a)_{H1,Cases} &= 2 * P(aa)_{H1,Cases} + P(Aa)_{H1,Cases} \\
 P(A)_{H0} &= [P(A)_{H1,Cases} + P(A)] / 2 \\
 P(a)_{H0} &= [P(a)_{H1,Cases} + P(a)] / 2
 \end{aligned} \tag{4}$$

$$\lambda = \sum_{a_i} 2N * \frac{(P(a_i)_{H1,Cases} - P(a_i)_{H0})^2}{P(a_i)_{H0}}$$

where $a_i \in \{A, a\}$.

Proxy test for association. If the association is not performed at the causal SNP A but at a tSNP C, the GRRs at locus C can be derived. Let $P(CC)$, $P(Cc)$ and $P(cc)$ be the population frequencies of genotypes CC , Cc , and cc , respectively, and let $P(g_i, CC)$, $P(g_i, Cc)$ and $P(g_i, cc)$ be the frequencies of genotypes g_iCC , g_iCc and $g_i cc$, where $g_i \in \{AA, Aa, aa\}$. The marginal GRRs of locus C are then:

$$GRR_{CC} = \frac{\left(\sum_{g_i} P(g_i, CC) GRR_{g_i} \right) P(cc)}{\left(\sum_{g_i} P(g_i, cc) GRR_{g_i} \right) P(CC)} \quad (5)$$

$$GRR_{Cc} = \frac{\left(\sum_{g_i} P(g_i, Cc) GRR_{g_i} \right) P(cc)}{\left(\sum_{g_i} P(g_i, cc) GRR_{g_i} \right) P(Cc)}$$

where GRR_{g_i} is the GRR of genotype g_i at the causal locus A. For fixed values defining an alternative (GRR_{AA} , GRR_{Aa} , $P(g_i)$, $P(g_i, CC)$, $P(g_i, Cc)$), the power of the tests can be computed using GRR_{CC} , GRR_{Cc} , $P(CC)$ and $P(Cc)$ in place of GRR_{AA} , GRR_{Aa} , $P(AA)$ and $P(Aa)$ in equations (1)-(4).

Direct test for association under a two-locus model of disease.

To determine the marginal genotype relative risks that result from a two-locus disease model, let the nine 2 locus-genotype frequencies be $P(AABB)$, $P(AABb)$, etc. The marginal genotype relative risks for SNP A can be calculated as follow :

$$GRR_{AA} = \frac{P(aa)[P(AABB) * GRR_{AABB} + P(AABb) * GRR_{AABb} + P(AAbb) * GRR_{AAbb}]}{P(AA)[P(aaBB) * GRR_{aaBB} + P(aaBb) * GRR_{aaBb} + P(aabb) * GRR_{aabb}]} \quad (6)$$

$$GRR_{Aa} = \frac{P(aa)[P(AaBB) * GRR_{AaBB} + P(AaBb) * GRR_{AaBb} + P(Aabb) * GRR_{Aabb}]}{P(Aa)[P(aaBB) * GRR_{aaBB} + P(aaBb) * GRR_{aaBb} + P(aabb) * GRR_{aabb}]}$$

where GRR_{g_i} is the GRR of the two-locus genotype g_i . The relative risks for BB and Bb follow similarly. From these relative risks, the power to detect an alternative with fixed genotype frequencies and relative penetrances can be calculated using equations (1)-(4).

Proxy test for association in a two-locus disease model. The calculations for the relative risks at a tSNP C presented for a single locus must be extended to take into account that the tSNP may be in LD with both causal loci A and B. Let $P(g_i, g_j, CC)$, $P(g_i, g_j, Cc)$ and $P(g_i, g_j, cc)$ be the probabilities of genotypes $g_i g_j CC$, $g_i g_j Cc$ and $g_i g_j cc$, where $g_i \in \{AA, Aa, aa\}$ and $g_j \in \{BB, Bb, bb\}$. The marginal genotype relative risks of locus C are then:

$$GRR_{CC} = \frac{\left(\sum_{i,j} P(g_i, g_j, CC) GRR_{ij} \right) P(cc)}{\left(\sum_{i,j} P(g_i, g_j, cc) GRR_{ij} \right) P(CC)} \quad (7)$$

$$GRR_{Cc} = \frac{\left(\sum_{i,j} P(g_i, g_j, Cc) GRR_{ij} \right) P(cc)}{\left(\sum_{i,j} P(g_i, g_j, cc) GRR_{ij} \right) P(Cc)}$$

where GRR_{ij} is the GRR of genotype $g_i g_j$. The power to detect an alternative for fixed genotype frequencies and relative risks can be calculated as in the one-locus model.

Appendix B

To reproduce the demonstration of Pritchard and Pzreworski (2001), assume we genotype N_1 individuals at locus A (with causal allele A), which we consider a disease locus, and N_2 individuals at locus C, which we consider as a marker. We wish to compare the power to detect an association at these two loci.

We define two classes of individuals: individuals with the disease (D) and unaffected individuals (U). Let $\pi_{DA} = P(A | D)$, $\pi_{UA} = P(A | U)$, $\pi_{DC} = P(C | D)$ and $\pi_{UC} = P(C | U)$. Let $q_{AC} = P(C|A)$ and $q_{aC} = P(C|a)$.

If we let $\varphi =$ the fraction of the sample that are cases (i.e. in the disease class), the allelic (one degree of freedom) χ^2 test of association at locus A is:

$$\chi_A^2 = \frac{(\hat{\pi}_{DA} - \hat{\pi}_{UA})^2 2N_1\varphi(1-\varphi)}{\hat{\pi}_A(1-\hat{\pi}_A)}$$

where the \wedge represents the sample frequency of the given quantity. The test statistic at locus C is similar, except with C in place of A and N_2 in place of N_1 . The distributions of the two test statistics are approximately the squares of random normal variables, with expectations

$$E[\chi_A^2] = (\pi_{DA} - \pi_{UA}) \sqrt{\frac{2N_1\varphi(1-\varphi)}{\bar{\pi}_A(1-\bar{\pi}_A)}}$$

and

$$E[\chi_C^2] = (\pi_{DC} - \pi_{UC}) \sqrt{\frac{2N_2\varphi(1-\varphi)}{\bar{\pi}_C(1-\bar{\pi}_C)}}$$

respectively, where

$$\bar{\pi}_A = \varphi\pi_{DA} + (1 - \varphi)\pi_{UA}$$

$$\bar{\pi}_C = \varphi\pi_{DC} + (1 - \varphi)\pi_{UC}$$

and variances of about 1 if the difference between cases and controls is small, in terms of frequency of A or C.

If we suppose that, conditional on A, C is independent from the disease, then

$$\pi_{DC} = q_{AC}\pi_{DA} + q_{aC}(1 - \pi_{DA}) \text{ and } \pi_{DC} - \pi_{UC} = (\pi_{DA} - \pi_{UA})(q_{AC} - q_{aC}).$$

By replacing $\pi_{DC} - \pi_{UC}$ with $(\pi_{DA} - \pi_{UA})(q_{AC} - q_{aC})$ in the equation for the mean of the χ^2 distribution for the test at locus C:

$$E[\chi_C^2] = (\pi_{DA} - \pi_{UA})(q_{AC} - q_{aC}) \sqrt{\frac{2N_2\varphi(1 - \varphi)}{\bar{\pi}_C(1 - \bar{\pi}_C)}}$$

Thus, the distributions at locus A and locus C are approximately the same if

$$N_2 = \frac{N_1\bar{\pi}_C(1 - \bar{\pi}_C)}{(q_{AC} - q_{aC})^2\bar{\pi}_A(1 - \bar{\pi}_A)}$$

If we define r^2 as:

$$r^2 = (q_{AC} - q_{aC})^2\pi_A(1 - \pi_A)\pi_C^{-1}(1 - \pi_C)^{-1}$$

we see that $N_2 = N_1/r^2$ if

$$\pi_A = \bar{\pi}_A, \pi_C = \bar{\pi}_C$$

References

2005. A haplotype map of the human genome. *Nature* 437(7063):1299-320.
- Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21(2):263-5.
- Chapman JM, Cooper JD, Todd JA, Clayton DG. 2003. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* 56(1-3):18-31.
- Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE and others. 2005. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* 37(11):1243-6.
- Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 305(5685):869-72.
- de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat Genet* 37(11):1217-23.
- Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A and others. 2006. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 314(5804):1461-3.
- Evans DM, Marchini J, Morris AP, Cardon LR. 2006. Two-stage two-locus models in genome-wide association. *PLoS Genet* 2(9):e157.
- Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* 38(1):82-5.
- Howson JM, Barratt BJ, Todd JA, Cordell HJ. 2005. Comparison of population- and family-based methods for genetic association analysis in the presence of interacting loci. *Genet Epidemiol* 29(1):51-67.
- Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M and others. 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411(6837):599-603.
- Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST and others. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* 308(5720):385-9.
- Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 22(2):139-44.
- Lesage S, Zouali H, Cezard JP, Colombel JF, Belaiche J, Almer S, Tysk C, O'Morain C, Gassull M, Binder V and others. 2002. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* 70(4):845-57.
- Li J, Ji L. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 95(3):221-7.
- Li M, Atmaca-Sonmez P, Othman M, Branham KE, Khanna R, Wade MS, Li Y, Liang L, Zarepari S, Swaroop A and others. 2006. CFH haplotypes without the Y402H

- coding variant show strong association with susceptibility to age-related macular degeneration. *Nat Genet* 38(9):1049-54.
- Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z, Schwartz S, Albertson DG, Pinkel D, Altshuler DM and others. 2006. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am J Hum Genet* 79(2):275-90.
- Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. 2003. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 33(2):177-82.
- Marchini J, Donnelly P, Cardon LR. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 37(4):413-7.
- Moore JH. 2003. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 56(1-3):73-82.
- Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ. 2006. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet* 38(6):663-7.
- Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 69(1):124-37.
- Pritchard JK, Cox NJ. 2002. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet* 11(20):2417-23.
- Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 69(1):1-14.
- Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. *Trends Genet* 17(9):502-10.
- Terwilliger JD, Hiekkalinna T. 2006. An utter refutation of the "Fundamental Theorem of the HapMap". *Eur J Hum Genet* 14(4):426-37.
- Wang WY, Barratt BJ, Clayton DG, Todd JA. 2005. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 6(2):109-18.
- Weng L, Kavaslar N, Ustaszewska A, Doelle H, Schackwitz W, Hebert S, Cohen JC, McPherson R, Pennacchio LA. 2005. Lack of MEF2A mutations in coronary artery disease. *J Clin Invest* 115(4):1016-20.
- Zondervan KT, Cardon LR. 2004. The complex interplay among factors that influence allelic association. *Nat Rev Genet* 5(2):89-100.

		<i>Genotypes</i>	<i>AA</i>	<i>Aa</i>	<i>aa</i>
1.	Heterogeneity	<i>BB</i>	2	2	2
		<i>Bb</i>	2	1	1
		<i>bb</i>	2	1	1
2.	Epistasis	<i>BB</i>	4	1	1
		<i>Bb</i>	1	1	1
		<i>bb</i>	1	1	1
3.	Additive	<i>BB</i>	3	2.5	2
		<i>Bb</i>	2.5	2	1.5
		<i>bb</i>	2	1.5	1
4.	Multiplicative	<i>BB</i>	4	3	2
		<i>Bb</i>	3	2.25	1.5
		<i>bb</i>	2	1.5	1

Table 1: Two-locus Genotype Relative Risk (GRR) matrices used in power calculations

Table 1 : Marginal GRR at a causal locus A with frequency 0.05, under the 2-locus heterogeneity model 1 of Error! Reference source not found., when the two causal alleles are out of phase. Marginal GRRs are given as a function of both the allele frequency at the other locus (in columns) and of D' between A and B (in lines). Risks for the locus A are presented in the format $GRR_{AA}/GRR_{Aa}/GRR_{aa}$.

		Allele Frequency at locus B				
		0.1	0.3	0.5	0.7	0.9
D'between loci A and B	0	1.98/1.00/1	1.83/1.00/1	1.60/1.00/1	1.34/1.00/1	1.10/1.00/1
	0.2	1.98/1.00/1	1.83/0.98/1	1.59/0.96/1	1.33/0.93/1	1.09/0.91/1
	0.4	1.98/1.00/1	1.83/0.96/1	1.59/0.91/1	1.32/0.86/1	1.08/0.81/1
	0.6	1.98/0.99/1	1.83/0.95/1	1.58/0.87/1	1.31/0.79/1	1.07/0.72/1
	0.8	1.98/0.99/1	1.82/0.93/1	1.57/0.83/1	1.31/0.72/1	1.06/0.62/1
	1	1.98/0.99/1	1.82/0.91/1	1.57/0.78/1	1.30/0.65/1	1.05/0.53/1

Penetrance model	GRR_{AA}	GRR_{Aa}
1 (heterogeneity)	1.6 [1, 2]	1 [0.5, 1.2]
2 (epistasis)	1.75 [1, 4]	1 [1, 1]
3 (multiplicative)	2 [1, 4]	1.5 [1.25, 2.25]
4 (additive)	1.66 [1, 3]	1.33 [1, 2]

Table 3: Median and range of single locus GRRs computed from the ENM010 data for different two-locus models. The first number is the median in all two-way models, and the 95% range is presented in brackets. GRR_{aa} is always one.

For Peer Review

Tagging Parameters	Estimated number of	
	tSNPs	Independent tests
No tagging	6,056,604	992,296
$r^2=0.8$, MAF=0.1	597,770	328,805
$r^2=0.8$, MAF=0.2	385,461	242,475
$r^2=0.6$, MAF=0.1	394,648	217,572
$r^2=0.6$, MAF=0.2	257,586	160,181

Table 4: Estimated numbers of tSNPs to cover the entire genome using different tagging parameters and estimated numbers of independent tests (“effective” numbers of SNPs) that they entail.

For Peer Review

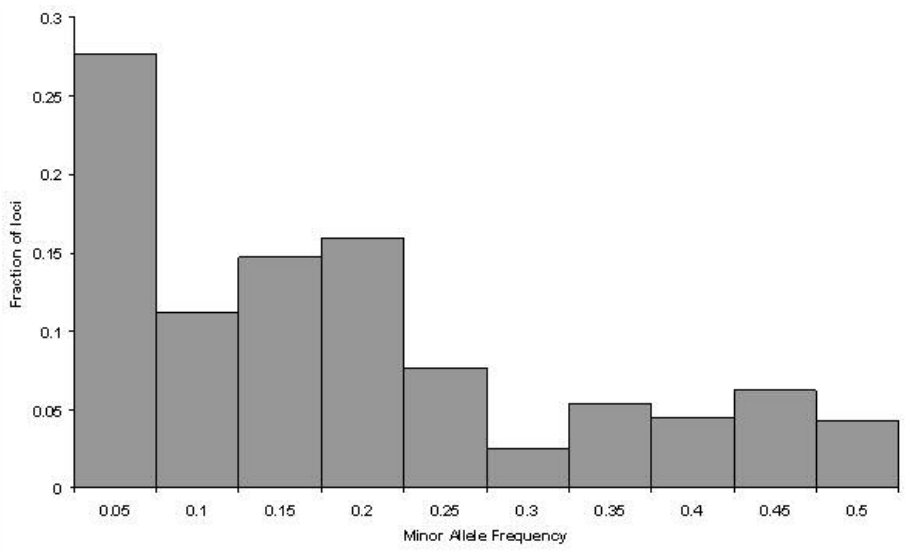


Figure 1: Distribution of minor allele frequencies in ENM010. There are 716 polymorphic SNPs in this region in the sample.

Review

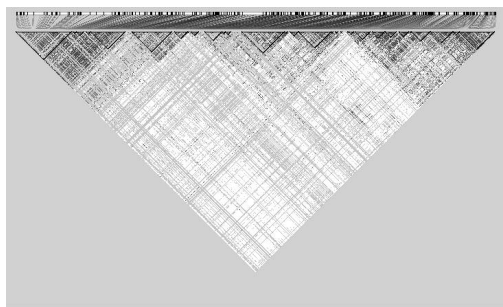


Figure 2: LD plot for the ENCODE region ENM010 generated in Haploview. Black squares represent SNPs in strong LD.

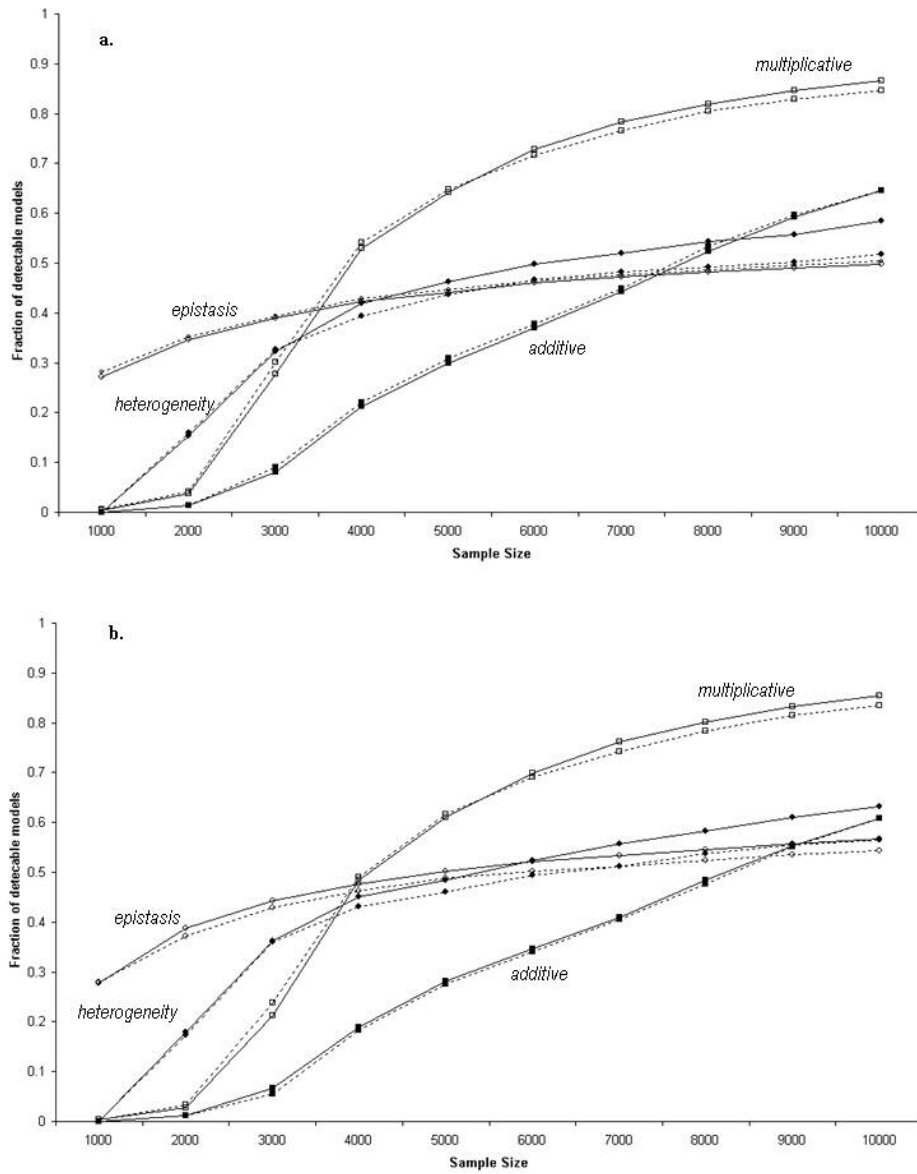


Figure 3: Fraction of detectable models in the ENM010 region under different two-locus models. The solid and dotted lines represent the curves for direct testing and the use of tSNPs, respectively. The filled circles, open circles, filled squares, and open squares represent models 1 (heterogeneity), 2 (epistasis), 3 (additive), and 4 (multiplicative) respectively. a. using a 1-dof allelic test for association and b. using a 2-dof genotypic test for association.

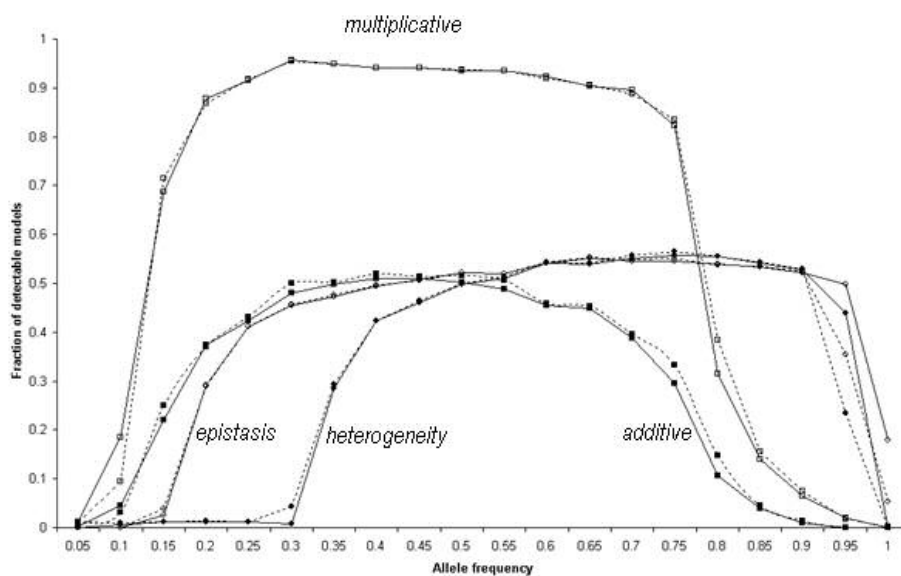


Figure 4: Fraction of detectable models using the allelic test for association in the ENM010 region as a function of the allele frequency of SNP1 under 4 two-locus GRR models. The solid and dotted lines represent the curves for direct testing and the use of tSNPs, respectively. The filled circles, open circles, filled squares, and open squares represent models 1 (heterogeneity), 2 (epistasis), 3 (additive), and 4 (multiplicative) respectively.

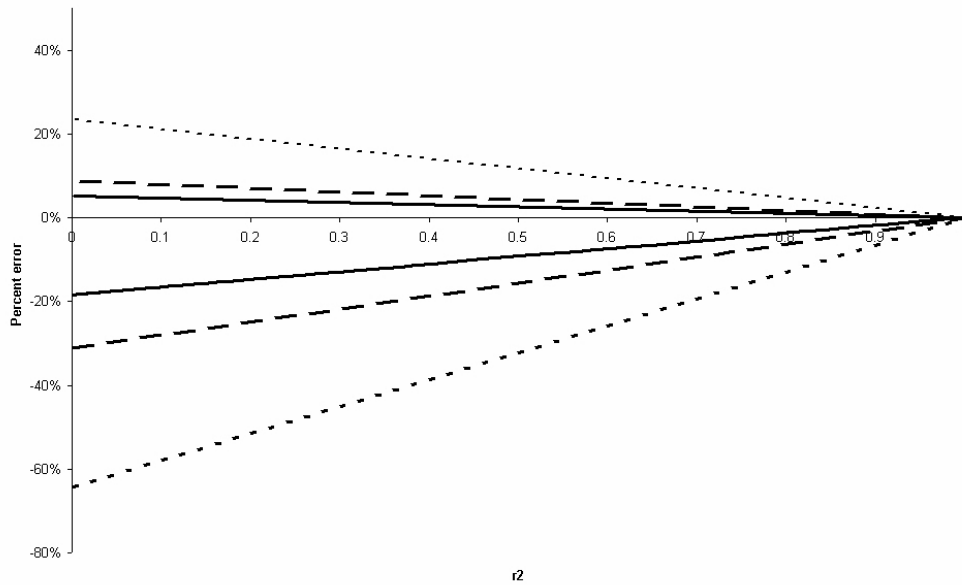


Figure 5: Upper and lower bounds on the error in the proportionality assumption for the 1df association test under a one-locus recessive model, where $GRR(AA)=x$ and $GRR(Aa)=GRR(aa)=1$. Errors are plotted as a function of r^2 between the tSNP and the causal SNP. Dotted, dashed and bold lines represent the bounds for $x=5, 2,$ and $1.5,$ respectively.

Review

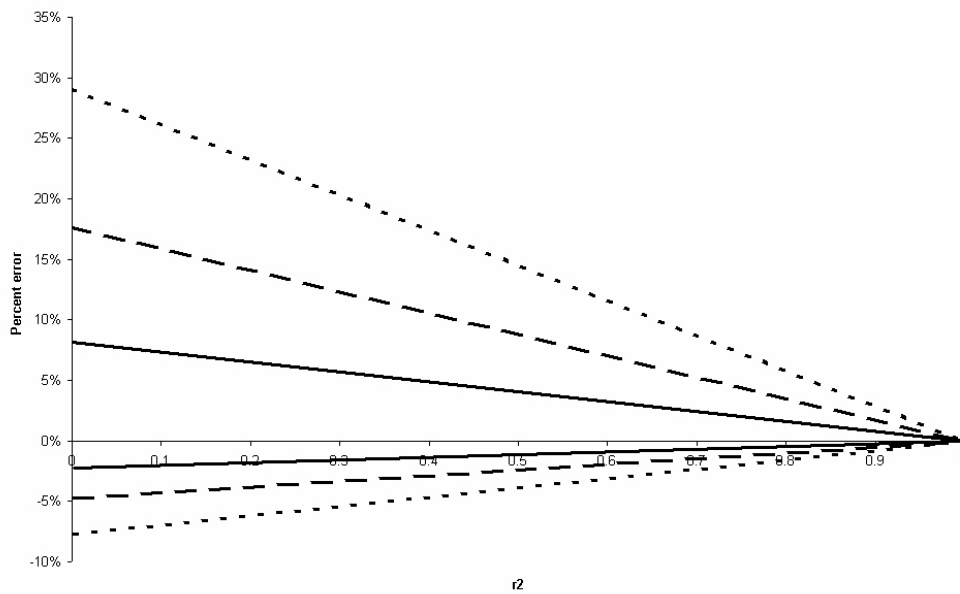


Figure 6: Upper and lower bounds on the error in the proportionality assumption for the 1df association test under a one-locus dominant model, where $GRR(AA)=GRR(Aa)=x$ and $GRR(aa)=1$. Errors are plotted as a function of r^2 between the tSNP and the causal SNP. Dotted, dashed, and bold lines represent the bounds for $x=2$, 1.5, and 1.2, respectively

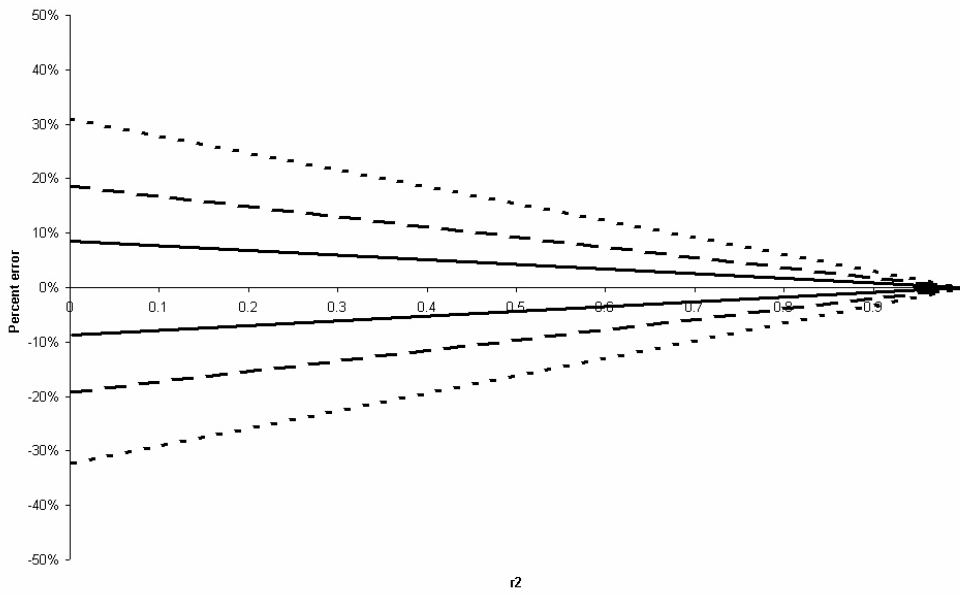


Figure 8: Upper and lower bounds on the proportionality assumption under for the 1df association test a one-locus multiplicative model, where $GRR(AA)= x^2$, $GRR(Aa)=x$, and $GRR(aa)=1$. Errors are plotted as a function of r^2 between the tSNP and the causal SNP. Dotted, dashed and bold lines represent the bounds for $x=2,1.5$, and 1.2 , respectively.

Review

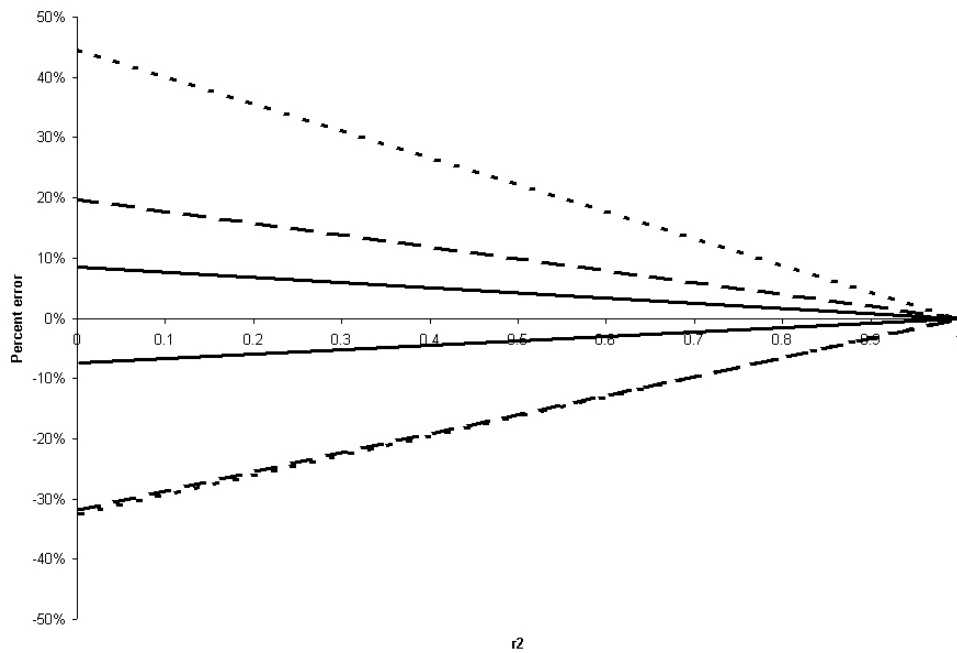


Figure 7: Upper and lower bound on the error in the proportionality assumption for the 1df association test under a one-locus additive model, where $GRR(AA)=2x$, $GRR(Aa)=x$, and $GRR(aa)=1$. Errors are plotted as a function of r^2 between the tSNP and the causal SNP. Dotted, dashed and bold lines represent the bounds for $x=3$, 1.5 and 1.2, respectively.

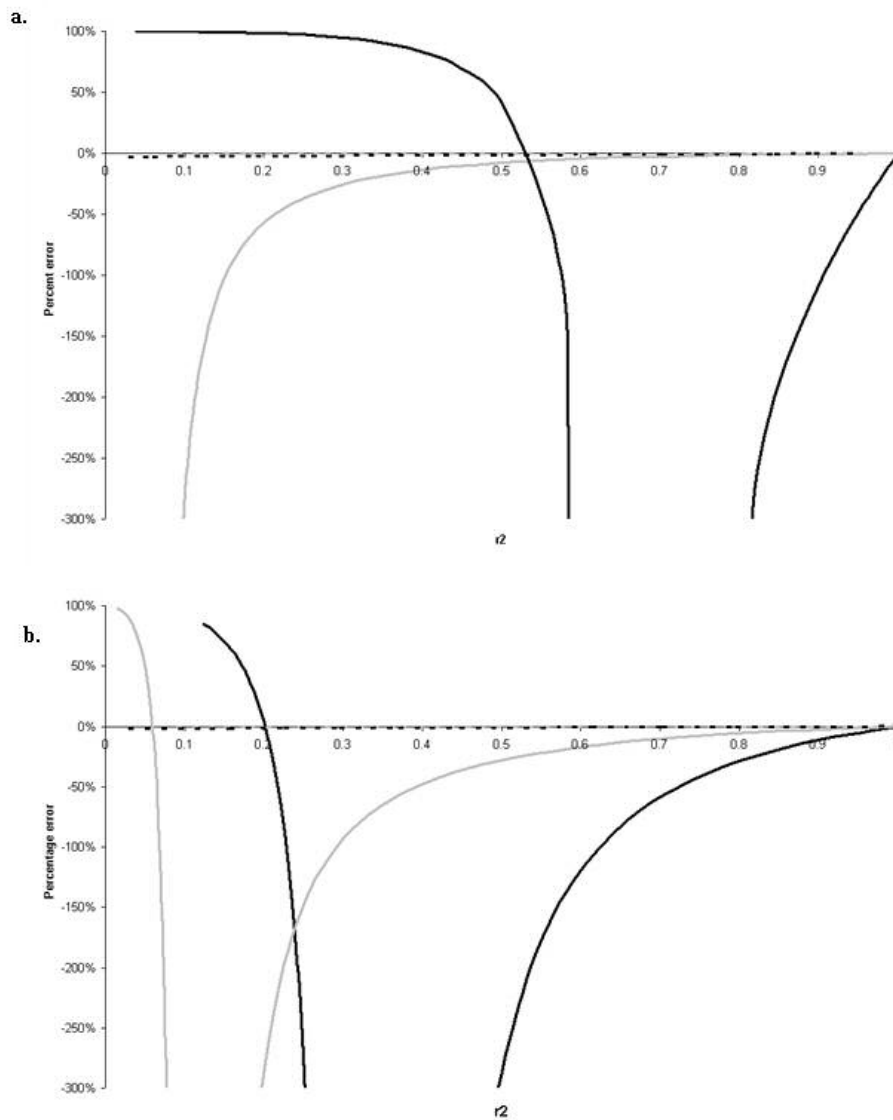


Figure 9: Error in the proportionality assumption for the 1df association test for a two-locus heterogeneity model where the causal alleles are out of phase, as a function of r^2 between the tSNP and the causal allele it is intended to tag. The bold and dotted portions represent the cases where the frequency of the tag allele is greater than or less than, respectively, the frequency of the causal allele it is intended to tag. a. For the black line, the tSNP tags an allele of frequency 0.1 in the presence of another causal allele of frequency 0.25. The reverse is true for the grey line. b. Both causal alleles have the same frequency. The grey and black lines represent allele frequencies of 0.1 and 0.25, respectively.

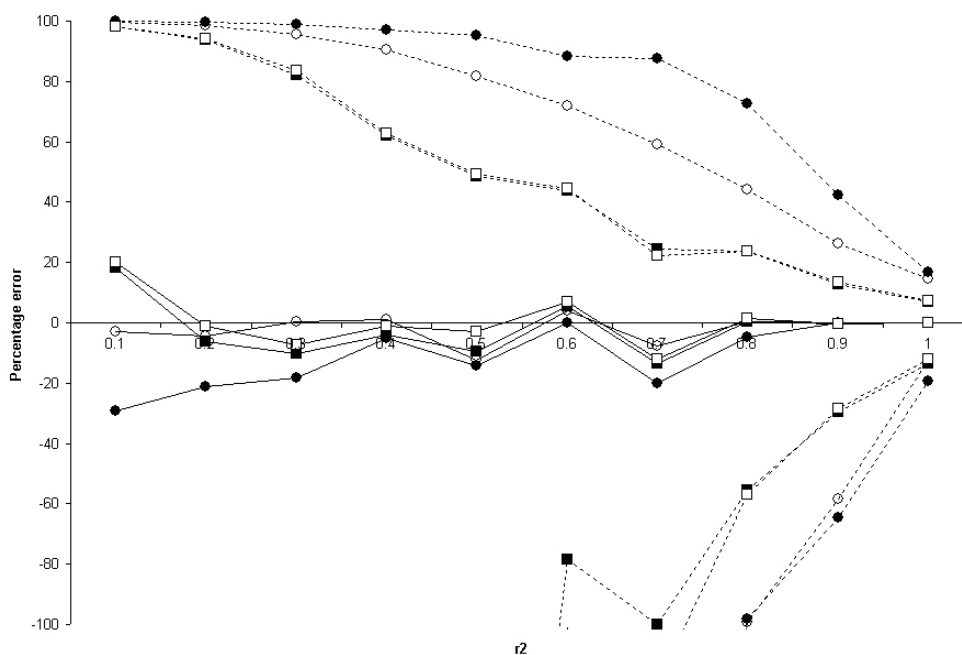


Figure 10: Median (bold) and 90% range (dotted) errors in the proportionality assumption for the two-locus GRR models based on the ENM010 data. The error in each model is binned according to the r^2 between the tSNP and one of the causal SNPs. The filled circles, open circles, filled squares, and open squares represent models 1 (heterogeneity), 2 (epistasis), 3 (additive), and 4 (multiplicative) respectively. The lines do not all converge at $r^2=1$ because the bin labelled 1 contains all the r^2 values from 0.9 to 1.

View