

Genome-wide strategies for detecting multiple loci that influence complex diseases

Jonathan Marchini¹, Peter Donnelly¹ & Lon R Cardon²

After nearly 10 years of intense academic and commercial research effort, large genome-wide association studies for common complex diseases are now imminent. Although these conditions involve a complex relationship between genotype and phenotype, including interactions between unlinked loci¹, the prevailing strategies for analysis of such studies focus on the locus-by-locus paradigm. Here we consider analytical methods that explicitly look for statistical interactions between loci. We show first that they are computationally feasible, even for studies of hundreds of thousands of loci, and second that even with a conservative correction for multiple testing, they can be more powerful than traditional analyses under a range of models for interlocus interactions. We also show that

plausible variations across populations in allele frequencies among interacting loci can markedly affect the power to detect their marginal effects, which may account in part for the well-known difficulties in replicating association results. These results suggest that searching for interactions among genetic loci can be fruitfully incorporated into analysis strategies for genome-wide association studies.

Since the completion of the human genome project, genome-wide association studies have been considered to hold promise for unravelling the genetic etiology of complex traits². It is now possible to assess this promise, as the emergence of large marker panels, large collections of well-phenotyped human samples and high-throughput genotyping

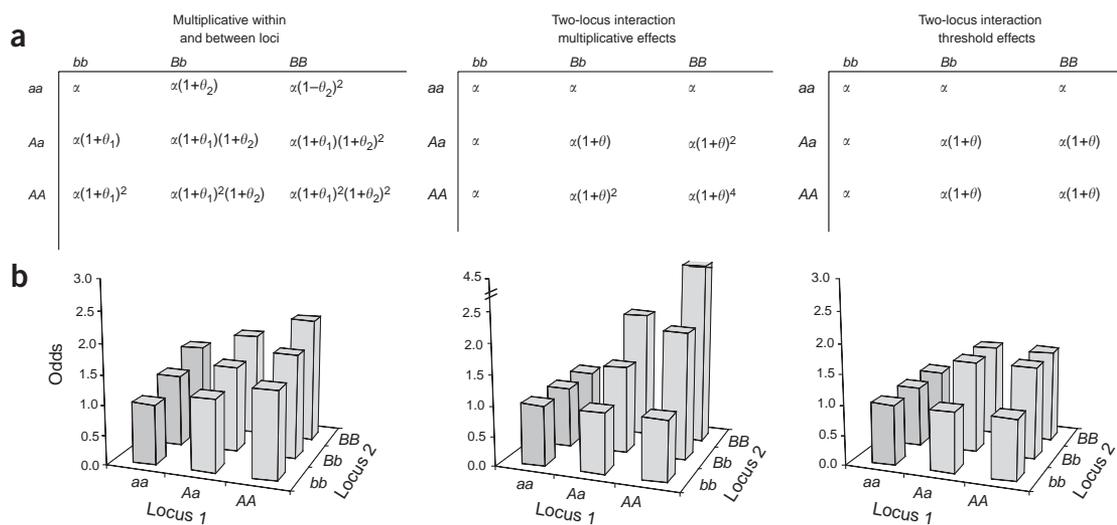


Figure 1 Multilocus models of disease. **(a)** The odds of disease for two loci under the epistatic scenarios considered. In model 1, the odds increase multiplicatively with genotype both within and between loci. In model 2, the odds have a baseline value (α) unless both loci have at least one disease-associated allele. After that, the odds increase multiplicatively within and between genotypes. Model 3 is similar to model 2 but specifies a threshold of disease effects rather than multiplicative gene action. Both loci have the same effect size. As models 2 and 3 include no explicit marginal effects, they are expected to be harder to detect without an interaction-based search strategy. **(b)** Examples of the genotypic risks under illustrative parameters. In these examples, $\pi_A = \pi_B = 0.25$ and $\lambda = 0.20$, which permits derivation of the genotypic effects, θ , as 0.20, 0.45 and 0.53 for the examples shown (left to right); $\alpha = 1.0$ for illustration purposes.

¹Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. ²Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. Correspondence should be addressed to L.R.C. (lon.cardon@well.ox.ac.uk).

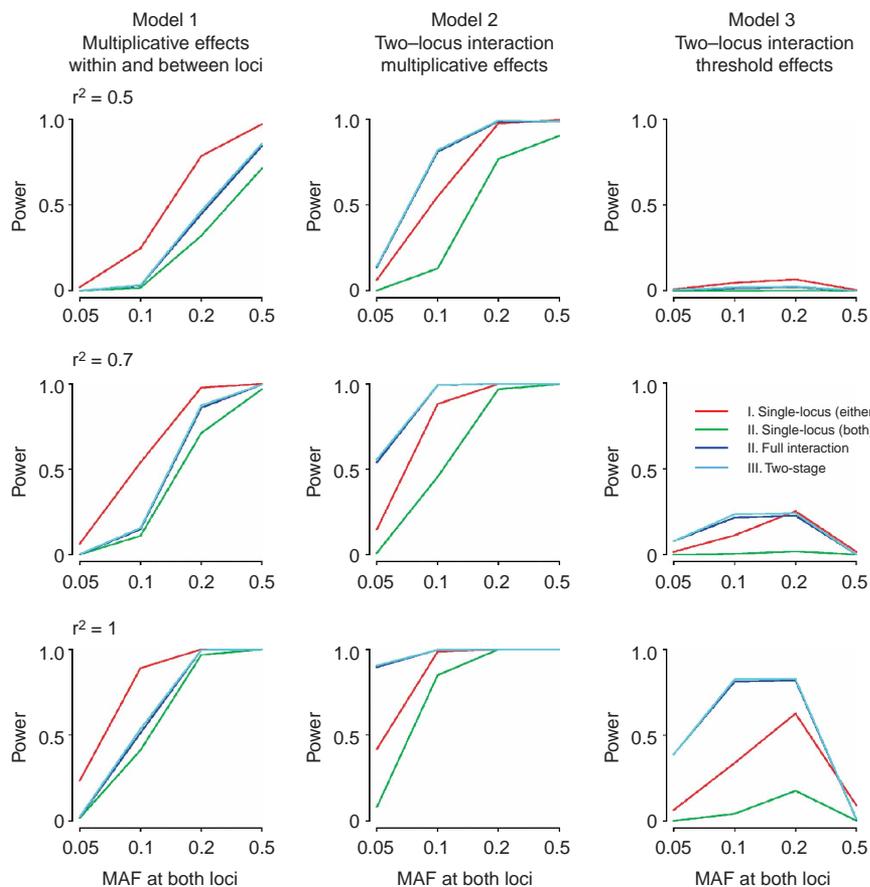


Figure 2 Power to detect genetic association using different search strategies. Individual panels show the power to detect association using each of the search strategies considered. Red lines, strategy Ia (requiring that at least one of the two loci meets the significance threshold); green lines, strategy Ib (requiring significance at both loci); dark blue lines, full interaction approach; light blue lines, two-stage strategy. The columns show the three different 'true' disease gene models. The rows show different levels of LD between the unobserved disease-associated loci and measured markers using constraint 2. λ is the size of the marginal effect at the causative loci. The y axis shows the statistical power for each of the strategies. The x axes show the minor allele frequency (MAF) of the disease-causing alleles (assumed equal for both loci). For this subset of results, $n = 2,000$ cases and 2,000 controls and $\lambda = 0.50$. The nominal significance threshold for all simulations was set at $\alpha = 0.05$, with the initial screening threshold for two-stage tests set at $\alpha_1 = 0.10$. A more comprehensive set of outcomes is provided in **Supplementary Note** online.

are enabling genome-wide assessment. Initial reports of such studies are appearing in the literature^{3,4}.

Despite the achievements that render genome-wide studies feasible, it is not obvious how to analyze the resulting data productively. Most analytical methods proceed by considering each genetic marker or haplotype individually⁵, but increasing empirical evidence from model organisms^{6–9} and human studies^{10,11} suggests that interactions among loci contribute broadly to complex traits. Although there are many possible biological configurations by which just two loci can interact, much recent statistical work has focused on interaction models that have little or no marginal effects at each locus^{12–14}. Here we address the following question: given the plausibility of interactions between genetic loci with non-negligible marginal effects, how might we design and analyze genome-wide association studies?

We consider three different statistical models for interlocus interactions that attempt to mimic simple biological mechanisms (**Fig. 1**). Model 1 involves multiplicative effects within and between loci: on the appropriate scale, this model is additive and has marginal effects that should be 'detectable' independent of other loci. Models 2 and 3 explicitly include interactions, in two different ways that are consistent with plausible models¹⁵ in humans. We set the parameters of each model so that the marginal effect (*i.e.*, the effect at one locus considered individually) is in the range suggested by empirical studies in humans, namely relative risks of 1.2–2.0 (refs. 16,17).

We examined three strategies for analyzing genome-wide association studies: strategy I, locus-by-locus search; strategy II, search over all pairs of loci; and strategy III, a two-stage strategy in which all loci meeting some low threshold in a single-locus search are subsequently examined for a significant full model fit. This approach differs from

those that require that single loci meet strict statistical significance in the first stage¹⁸; such approaches will miss loci with modest marginal effects and large interactions¹⁹.

For power considerations, because the interaction strategies (strategies II and III) consider two disease-associated loci simultaneously, they are directly comparable to a single-locus approach that defines success on the basis of detecting both loci individually (which we call strategy Ib). In initial genome-wide screens, however, the primary objective may be to detect any locus irrespective of others involved. Therefore, we also compared the interaction strategies with one in which either of the interacting loci is detected (strategy Ia).

To assess the power of each strategy, we simulated genotypes at two interacting loci in n cases and n controls. The power calculations consider $L = 300,000$ genotyped markers with only a single pair of (unobserved) causative loci, each of which is in linkage disequilibrium (LD) with one of the genotyped markers. We used Bonferroni corrections to account for the large number of tests done in each strategy. In the two strategies that look explicitly for interactions, we fit a full model, not the interaction model under which the data are simulated.

An illustrative selection of the simulation results is presented in **Figure 2** for 2,000 cases and 2,000 controls in which the marginal heterozygote odds ratio at both loci is equal to 1.5. The most notable outcome is that there are many configurations in which the interaction-based search strategies are more powerful than searching locus-by-locus for all three models considered. These results are unexpected because the interaction-based searches involve statistical correction for as many as 10^5 times the number of tests of the single-locus searches. Marginal effects still exist in these models, but the multilocus information is so great as to negate the multiple-testing cost. As

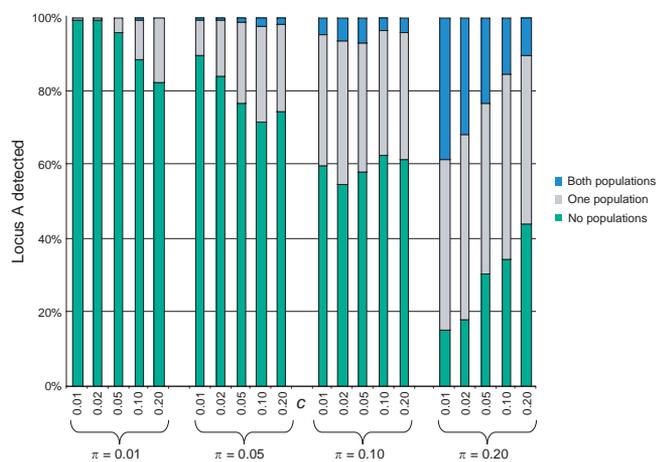


Figure 3 Replication of marginal association effects among interacting loci. Two interacting loci (A and B) were simulated in two populations using the model of population structure validated in ref. 29. The mean allele frequencies at the two loci were 0.1 and π , respectively. The variability of the actual population allele frequencies was controlled through the parameter c , which is equal to the traditional F_{ST} measure under these scenarios. A single-locus logistic regression with 1 degree of freedom was fitted to test for association at locus A. Each bar shows the three possible study outcomes for detecting locus A: not detected in either population (green), detected in only one population (*i.e.*, detected in one sample but not replicated in the second; gray) or detected in both populations (*i.e.*, detected and replicated; blue). As the disease-associated allele frequency increases, the power to detect locus A increases, but as the allele frequencies become more distinct between samples, the frequency of replication decreases sharply and the proportion of times locus A is seen in only one sample increases.

expected²⁰, power is strongly correlated with allele frequency at the disease-associated loci and increases with LD between the marker and disease-associated loci (**Fig. 2** and **Supplementary Note** online). These relationships hold rather generally across the disease models and search strategies that we examined. An exception occurs for model 1, such that single-locus searches for either of the two interacting loci (strategy Ia) often yield more power than interaction-based searches, but at the expense of uncovering only one of the loci. Our results thus extend the role of interaction-based searches beyond situations where they are of obvious importance (*e.g.*, large effects in which all individual loci are essential for detection^{12,13}) to some that are less obvious (*e.g.*, modest single-locus effects that may not meet statistical significance individually but that are detectable when considered jointly¹⁹).

The computational burden of searching for interactions can prevent full assessment¹⁴. A notable finding of this study is that all three strategies considered are computationally feasible for large sample sizes and genome-wide settings, with the most demanding strategy (strategy II) taking ~ 33 hours to analyze on a ten-node cluster for 1,000 cases and 1,000 controls (**Supplementary Note** online). The fact that this strategy is possible is especially relevant for models involving no marginal effects^{12–14}, as it is the only one of the three we considered that would uncover the loci involved.

Our results bear on the reported failure to replicate association studies upon follow-up^{17,21}. In addition to the oft-cited factors of statistical overinterpretation, small sample sizes, genetic and phenotypic heterogeneity, and population structure, these results highlight the possibility that for interacting loci, differences in allele frequencies between initial and replicate populations affect the power of single-locus strategies and thus hinder reproducibility^{7,19,22}. (This is different

from the problem of false positives caused by cryptic structure in a study population.) To explore this possibility further, we simulated two unlinked loci (denoted A and B) in a study of two separate populations and examined how often one of the loci (locus A) would be detected in none, only one or both of the populations. The studies can often differ in their detection of a disease-associated locus (**Fig. 3**), especially as the two populations become more genetically differentiated. This effect is most pronounced when the interacting disease-associated allele (at locus B) is common in the initial population ($\pi \geq 0.10$ in **Fig. 3**), where replication was generally not achieved in $>30\%$ of the simulations. In practice, such nonreplication will be exacerbated by differences in the frequencies of causative environmental factors.

There are several ways in which our analyses understate the potential utility of analytical strategies that explicitly look for interactions. First, we applied the simplest correction for multiple testing (Bonferroni), which is conservative. The multiple-testing cost of fitting interaction models is much greater than that for the single-locus analyses. Therefore, with a less conservative penalty, the relative power gain would be greater for the interaction strategies. Permutation-based strategies, though computationally expensive, may help to reduce the multiple testing burden. Second, obtaining the correct error probabilities in sequential tests is not straightforward, and our simple implementation of the two-stage strategy (strategy III) is conservative (which explains why this approach does not greatly outperform the full interaction approach (strategy II) in our comparisons). A more sophisticated sequential test would increase power and, hence, increase the utility of explicitly considering interactions. Third, all our models assume some level of marginal effects. In cases where trait variation arises exclusively from interactions, interaction-based searches will always perform better than single-locus tests.

The determination of a single best strategy for the detection of loci in a general multilocus model is complicated because both the number of interacting loci and the form of the interaction can vary, yielding many possible models with different properties. Here we began with a simple system of two loci. To gain a preliminary view of higher-order statistical interactions, we extended our assessments to an analogous class of three-locus models. We asked how well the one- and two-locus search strategies perform when there are three interacting loci and, more generally, whether there are better strategies for uncovering all three loci under these models. Our conclusions regarding the first point are similar to those for two-locus models: loci with large marginal effects relative to their interaction effects are detected well using single-locus searches, but loci with explicit three-way interactions are more likely to be detected by searching for two-locus marginal effects than by single-locus screening. Regarding the second question, we also found that searching explicitly (using a two-stage strategy) for all three loci together could be more powerful than both single-locus and two-locus searches (**Supplementary Note** online).

There are several ways in which these analyses may be extended. For simplicity, we considered models and analyses in which causative alleles are single SNPs rather than haplotypes. Examination of haplotype-based models would require many more assumptions, but we would expect the same general conclusions to hold. In addition, we focused on gene-gene interactions, but gene-environment interactions could be handled by similar models, effectively by treating the environmental variable as a locus. There is considerable interest in study designs that pool DNA from sampled individuals to reduce genotyping costs²³, but pooling precludes the possibility of fitting interaction models, which is a potential disadvantage of such designs.

We conclude that in analyzing genome-wide association studies, fitting models that explicitly allow for interactions between loci can add

substantially to single-locus searches. Perhaps unexpectedly, not only are interaction-based searches computationally feasible for genome-wide studies, but they can also be more powerful than single-locus approaches, even when accounting for the multiple-testing cost. This will not be true, however, when the single-locus effects are large relative to the interaction effects, particularly if they are sufficient to identify at least one of the loci. Although the power of any search strategy depends on the underlying model, a useful compromise between exhaustive searching and locus-by-locus tests may be obtained using a two-stage approach that first identifies a set of single loci under liberal statistical criteria and then evaluates all possible two-way interactions among them under rigorous criteria, corrected for multiple testing.

METHODS

Two-locus models. There is a broad spectrum of scenarios for interactions among genetic loci, ranging from situations in which no effects would be detected by searching one locus at a time (reviewed in ref. 12) to those in which the results of genetic interaction would be reflected in the marginal effects of the two individual loci involved. The most general two-locus model for diallelic loci has nine parameters in the 3×3 table of genotypes. We selected three submodels of the general two-locus case for our comparisons of search strategies. **Figure 1a** shows these models in terms of the odds of disease for each combination of genotypes at two loci (A and B), parameterized as baseline effects, α , and genotypic effects, θ .

Model 1 specifies that the odds of disease increase in a multiplicative fashion both within and between two loci. In this model, an individual who is heterozygous at locus A has increased odds of $1 + \theta_1$ relative to those of an individual who is homozygous *aa*; the *AA* homozygote has further multiplicative odds of $(1 + \theta_1)^2$. Similar effects for locus B are reflected in θ_2 , and the odds of disease for each combination of genotypes at loci A and B is the product of the two within-locus effects.

Model 2 is a statistical interaction model that has explicit marginal effects. In this model, at least one disease-associated allele must be present at each locus for the odds to increase beyond the baseline level. Beyond that, each additional copy of the disease-associated allele at loci A or B further increases the odds by the multiplicative factor $1 + \theta$. Both loci have the same effect size (i.e., $\theta = \theta_1 = \theta_2$).

Model 3 takes the same form as Model 2 in requiring at least one copy of the disease-associated alleles at both loci A and B, but additional copies of the disease-associated alleles do not increase the risk further. This model reflects disease threshold effects, in which a single copy of the disease-associated allele at each locus is required to increase odds of disease, but having both copies of the disease-associated allele at either locus has no additional influence as the disease threshold has already been met. In classical terms¹⁵, model 1 is multiplicative and models 2 and 3 are variants of complementary gene models.

Marginalizing multilocus models. Most models for interaction between loci still have an effect (the marginal effect) at each of the loci separately. The magnitude of the marginal effect at a particular locus will depend on the model parameters, θ and α , and the allele frequencies at the other locus²⁴. There is relatively little data to indicate realistic interaction effect sizes for complex traits. In contrast, there is increasing empirical information about the magnitude of the marginal effect sizes¹⁷. To make use of the empirical information, we first fixed the marginal effect sizes under our three models and then worked backwards to determine the magnitude of the interaction effects. For this approach, we defined a marginal parameter, λ , and a disease prevalence, p (here $p = 0.01$), set the heterozygote odds ratio to a value of $1 + \lambda$ and then numerically derived the values of the model parameters θ and α under a range of allele frequencies (details provided in **Supplementary Note** online). As an example, for Model 2,

$$\lambda = \frac{\frac{2\theta\pi_B(1 - \pi_B)}{1 + \alpha(1 + \theta)} + \frac{\theta(2 + \theta)\pi_B^2}{1 + \alpha(1 + \theta)^2}}{\frac{(1 - \pi_B)^2}{1 + \alpha} + \frac{2\pi_B(1 - \pi_B)}{1 + \alpha(1 + \theta)} + \frac{\pi_B^2}{1 + \alpha(1 + \theta)^2}}$$

This shows the size of effect we can expect to see marginally (at locus A) for an interaction parameterized by θ that involves an unobserved locus (locus B) with allele frequency π_B .

Given the parameters of the two-locus model, we also considered the slightly more complicated situation of LD between the disease-associated loci and otherwise anonymous markers. By specifying the level of LD (using the pairwise parameter r^2) between a marker, X, in LD with disease-associated locus A and, similarly, the level of LD between an unlinked marker, Y, in LD with disease-associated locus B, we extended our approach to the situation in which the disease-predisposing loci are not observed but two correlated markers are genotyped instead. The derivation of this extension is provided in **Supplementary Note** online.

Strategies for searching for interactions. We present the disease models in terms of the odds of disease. For statistical assessment and comparisons of search strategies, it is somewhat more natural to work with the logarithm of the odds, because the multiplicative relationships become additive on the log-odds scale. This is the natural setting for logistic regression, for which there is well-developed theory for case-control studies²⁵. We used this framework to compare search strategies, taking advantage of the composition of genotype data for computational efficiency (**Supplementary Note** online).

For the three models in **Figure 1**, we simulated genotypes at loci X and Y for a range of parameter settings ($n = 1,000, 2,000$ or $4,000$; $\pi_A = \pi_B = 0.05, 0.1, 0.2$ or 0.5 ; $r^2 = 0.5, 0.7$ or 1.0 ; and $\lambda = 0.2, 0.5$ or 1.0). By selecting these settings, we focused on effects and sample sizes for which choice of search strategies could matter. In other settings, where all approaches have either very low or very high power, the comparisons are less interesting. For each combination of these parameters, we carried out 1,000 simulations and assessed the power of the following three strategies to detect the interacting loci.

Strategy I: single locus. For any single locus there are three possible genotypes, and we fitted a full logistic model with a parameter for each observed genotype. In quantitative genetics terms, this parameterization is the full single-locus model involving an intercept plus additive and dominance terms²⁶. To ensure an overall type I error of at most α , we used a Bonferroni correction to set the significance level of the test at each locus to α/L . For comparisons with interaction search strategies, we evaluated this strategy by two criteria: (i) requiring that at least one of the two loci meet the significance threshold, irrespective of the other locus, or (ii) requiring that both loci are significant. The former criterion is appropriate when the main aim is to find any genetic locus, whereas the latter is more appropriate for comparing different strategies to detect interactions. As these situations relate to different scientific questions, we assessed them both and refer to the 'either locus' and 'both loci' scenarios as strategies Ia and Ib, respectively.

Strategy II: full interaction. We fitted the full logistic regression model (with at most nine parameters) to the 3×3 table of observed genotypes at the pair of loci. The parameters comprise an intercept, additive and dominance terms for each locus, and four interaction terms. We used a Bonferroni correction to set the significance level of each test to α/L^2 . We defined 'success' on the basis of a significant model fit, which is different from testing the interaction terms over and above the main effects.

Strategy III: two-stage. In the first stage, we identified all loci that were significant in single-locus tests (as above) at a liberal level α_1 . We called this set of loci $I_1 \subseteq \{1, \dots, L\}$. We let d_1 be the degrees of freedom of the single-locus model fitted at stage one for locus l (maximum 2 degrees of freedom if all three genotypes are present) and defined k_l such that $P(\chi_{d_1}^2 > k_l) = \alpha_1$ for $l \in I_1$.

In the second stage, for each pair of loci l and m identified in stage one ($l, m \in I_1, l \neq m$), we calculated the log likelihood ratio statistic $R_{(l,m)}$ for the full interaction model. Because of the way in which loci l and m were identified, $R_{(l,m)} \geq k_l + k_m$. Therefore, we defined a new statistic $R'_{(l,m)} = R_{(l,m)} - (k_l + k_m)$ and assessed the significance of this statistic against a $\chi_{d'}^2$ distribution in which d' is the degrees of freedom of the full model fitted at the two loci. We set the level of significance using a Bonferroni correction based on the expected number of tests to be done ($\alpha/(\alpha_1^2 L^2)$). Through simulation, we found this procedure to provide a conservative test of interaction between two loci (data not shown).

In the above three strategies, we used log likelihood ratio tests for the full logistic regression model²⁷. Given the nine parameters (at most) in each model fitted and the reasonably large sample sizes that we assumed, we avoided the estimation bias of logistic regression in the presence of sparse data^{12,14}. For all simulations, we set the nominal significance threshold at $\alpha = 0.05$. Our two-stage approach is similar in principle to that of ref. 28, but we set a liberal first-stage screening level (here $\alpha_1 = 0.10$) in an attempt to detect loci with large interactions but small marginal effects¹⁹. We deliberately chose the sample sizes to be large to correspond to expected requirements for complex trait studies, but even for these large samples, there exist many models in which the power of detection will be low for some or all search strategies considered (e.g., for rare alleles).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the Wellcome Trust, the US National Institutes of Health and the SNP Consortium for support.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 1 December 2004; accepted 23 February 2005

Published online at <http://www.nature.com/naturegenetics/>

- Phillips, P.C. The language of gene interaction. *Genetics* **149**, 1167–1171 (1998).
- Risch, N.J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
- Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32**, 650–654 (2002).
- Roses, A. The genome era begins. *Nat. Genet.* **33** (Supp 2), 217 (2003).
- Thomas, D.C. *Statistical Methods in Genetic Epidemiology* (Oxford University Press, Oxford, 2004).
- Mackay, T.F. Quantitative trait loci in *Drosophila*. *Nat. Rev. Genet.* **2**, 11–20 (2001).
- Williams, S.M., Haines, J.L. & Moore, J.H. The use of animal models in the study of complex disease: all else is never equal or why do so many human studies fail to replicate animal findings? *Bioessays* **26**, 170–179 (2004).
- Routman, E.J. & Cheverud, J.M. Gene effects on a quantitative trait: Two-locus epistatic effects measured at microsatellite markers and at estimated QTL. *Evolution* **51**, 1654–1662 (1997).
- Segre, D., Deluna, A., Church, G.M. & Kishony, R. Modular epistasis in yeast metabolism. *Nat. Genet.* **37**, 77–83 (2005).
- Sing, C.F. & Davignon, J. Role of the apolipoprotein E polymorphism in determining normal plasma lipid and lipoprotein variation. *Am. J. Hum. Genet.* **37**, 268–285 (1985).
- Zerba, K.E., Ferrell, R.E. & Sing, C.F. Complex adaptive systems and human health: the influence of common genotypes of the apolipoprotein E (ApoE) gene polymorphism and age on the relational order within a field of lipid metabolism traits. *Hum. Genet.* **107**, 466–475 (2000).
- Hoh, J. & Ott, J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Rev. Genet.* **4**, 701–709 (2003).
- Culverhouse, R., Suarez, B.K., Lin, J. & Reich, T. A perspective on epistasis: limits of models displaying no main effect. *Am. J. Hum. Genet.* **70**, 461–471 (2002).
- Moore, J.H. & Ritchie, M.D. The challenges of whole-genome approaches to common diseases. *JAMA* **291**, 1642–1643 (2004).
- Kempthorne, O. *An Introduction to Genetic Statistics* (John Wiley & Sons, New York, 1957).
- Carlson, C.S., Newman, T.L. & Nickerson, D.A. SNPing in the human genome. *Curr. Opin. Chem. Biol.* **5**, 78–85 (2001).
- Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. & Hirschhorn, J.N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* **33**, 177–182 (2003).
- Carlson, C.S., Eberle, M.A., Kruglyak, L. & Nickerson, D.A. Mapping complex disease loci in whole-genome association studies. *Nature* **429**, 446–452 (2004).
- Templeton, A.R. Epistasis and complex traits. in *Epistasis and the Evolutionary Process* (eds. Wolf, J.B., Brodie, E.D.I. & Wade, M.J.) 41–57 (Oxford University Press, New York, 2000).
- Zondervan, K.T. & Cardon, L.R. The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* **5**, 89–100 (2004).
- Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A. & Contopoulos-Ioannidis, D.G. Replication validity of genetic association studies. *Nat. Genet.* **29**, 306–309 (2001).
- Moore, J.H. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.* **56**, 73–82 (2003).
- Sham, P., Bader, J.S., Craig, I., O'Donovan, M. & Owen, M. DNA pooling: A tool for large-scale association studies. *Nat. Rev. Genet.* **3**, 862–871 (2002).
- Tiwari, H.K. Deriving components of genetic variance for multilocus models. *Genet. Epidemiol.* **14**, 1131–1136 (1997).
- Schlesselman, J.J. *Case-Control Studies: Design, Conduct, Analysis* (Oxford University Press, Oxford, 1982).
- Sham, P. *Statistics in Human Genetics* (Hodder Arnold, London, 1997).
- Clayton, D. Population association. in *Handbook of Statistical Genetics* (eds. Balding, D.J., Bishop, M. & Cannings, C.) 519–540 (John Wiley & Sons, New York, 2001).
- Hoh, J. *et al.* Selecting SNPs in two-stage analysis of disease association data: a model-free approach. *Ann. Hum. Genet.* **64**, 413–417 (2000).
- Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36**, 512–517 (2004).