General Epistatic Models of the Risk of Complex Diseases

Yun S. Song\textsuperscript{1}, Fulton Wang\textsuperscript{2}, and Montgomery Slatkin\textsuperscript{2}

\textsuperscript{1} Departments of EECS and Statistics, University of California, Berkeley CA, 94720-1776

\textsuperscript{2} Department of Integrative Biology, University of California, Berkeley CA, 94720-3140

September 17, 2010
Running head: General epistatic models

Keywords:
complex disease, gene interactions, missing heritability, recurrence risk, Markov chain Monte Carlo

Corresponding author:
Yun S. Song
University of California, Berkeley
Department of EECS
683 Soda Hall # 1776
Berkeley, CA 94720-1776
U.S.A.
Tel: +1 510 642 2211
Fax: +1 510 642 5775
Email: yss@eecs.berkeley.edu
ABSTRACT

The range of possible gene interactions in a multilocus model of a complex inherited disease is studied by exploring genotype-specific risks subject to the constraint that the allele frequencies and marginal risks are known. We quantify the effect of gene interactions by defining the interaction ratio, $C_R = K_R/K'_R$, where $K_R$ is the recurrence risk to relatives with relationship $R$ for the true model and $K'_R$ is the recurrence risk to relatives for a multiplicative model with the same marginal risks. We use a Markov chain Monte-Carlo procedure to sample from the space of possible models. We find that the average of $C_R$ increases with the number of loci for both low frequency ($p = 0.03$) and higher frequency ($p = 0.25$) causative alleles. Furthermore, the probability that $C_R > 1$ is nearly 1. Similar results are obtained when more weight is given to risk models that are closer to the comparable multiplicative model. These results imply that, in general, gene interactions will result in greater heritability of a complex inherited disease than is expected on the basis of a multiplicative model of interactions, and hence may provide a partial explanation for the problem of missing heritability of complex diseases.
INTRODUCTION

Although many genome-wide association studies (GWAS) have been performed and have found hundreds of SNPs associated with higher risk of complex inherited diseases, those SNPs so far account for only a small fraction of the inherited risk of those diseases (Altshuler et al., 2008). Several not-mutually-exclusive explanations have been proposed for the “missing heritability,” i.e., the heritability that is not yet accounted for by SNPs found in GWAS (Manolio et al., 2009): (i) common alleles of small effect that have not been found because GWAS done so far have been underpowered, (ii) low frequency alleles of moderate effect that are difficult to find using HapMap SNPs, (iii) rare copy-number variants that are not in strong linkage disequilibrium (LD) with HapMap SNPs, (iv) inherited epigenetic factors that are not in strong LD with HapMap SNPs, (v) interactions among causative alleles that conceal their true contribution to heritability. In this paper we investigate the last possibility and determine the extent to which interactions may account for missing heritability.

Our analysis is in the same spirit as that of Culverhouse et al. (2002). We assume that the risk of being affected by a complex disease is determined by an individual’s genotype at 2 or more loci, and that the frequencies of causative alleles and the average risks for each one-locus genotype (the marginal risks) are known. Culverhouse et al. assumed the marginal risks were the same for all genotypes and all loci. In that case, causative alleles have odds ratios of 1; they contribute to risk only through their interactions. Culverhouse et al. found the risk function which maximized the heritability and showed that the maximum possible heritability attributable to interactions increased with the number of loci. They concluded that it is quite possible that interactions among loci that have no main effect could contribute
substantially to the heritability of a complex disease and indeed could account for “virtually all the variation in affection status for diseases with any prevalence” (Culverhouse et al., 2002, page 468).

We generalize the analysis of Culverhouse et al. in three ways. First, we allow causative alleles to have odds ratios greater than 1. Second, we explore the entire space of models instead of focusing only on the risk model that maximizes heritability. Third, we examine how the importance of gene interactions depends on the “distance” between a risk model and a comparable multiplicative model. We will show that gene interactions can substantially increase the heritability of risk as measured by recurrence risk, $K_R$, and that the effect increases with the number of loci carrying causative alleles. Furthermore, we show that these results are true even if more weight is given to models that are closer to a comparable multiplicative model.

Geometrically, the space of feasible genotype-specific risks subject to the aforementioned constraints (i.e., that the allele frequencies and marginal risks are known) corresponds to a high-dimensional convex polytope, and the computational problem of interest involves integrating a quadratic function over the polytope. The dimension of the polytope grows exponentially with the number of loci, and, therefore, analytic computation is intractable for more than 2 loci. Hence, we devise a Monte Carlo approach to tackle the problem. Note that, because of high dimensionality, rejection algorithms are not appropriate for this kind of problem. We instead employ a Markov chain Monte Carlo (MCMC) algorithm based on a random walk that always stays inside the polytope. We present empirical results for up to 5 loci and obtain a closed-form formula for the minimum of $K_R$ over the polytope; the latter result applies to an arbitrary number of loci. Interestingly, the minimum of $K_R$ decreases as
the number $L$ of loci increases, but the average of $K_R$ over the polytope increases with $L$.

MULTILOCUS MODEL OF RISK

Model constraints: We assume that the risk of an individual being affected by a dichotomous complex disease depends on the genotype at $L$ diallelic loci. We use $f(k)$ to denote the probability that an individual with the $L$-locus genotype $k = (k_1, \ldots, k_L)$ is affected, where $k_i = 0, 1, 2$ indicates the number of copies of the higher risk allele (denoted by $A_i$) at locus $i$. Note that $0 \leq f(k) \leq 1$ for all genotypes $k$. We assume that the frequency $p_i$ of $A_i$ in a population is known. Further, we assume that the loci are unlinked, and are in Hardy-Weinberg and linkage equilibrium.

The average risk in the population is

$$K = \sum_k \Pr(k)f(k),$$

where $\Pr(k)$ is the probability of genotype $k$ in the population and the sum is over all genotypes. The marginal risks for each one-locus genotype are obtained by averaging over the other loci:

$$\bar{f}_i(k) = \sum_{k' = (k'_1, \ldots, k'_L): k'_i = k} \Pr(k' | k'_i = k)f(k').$$

In this paper, we assume that these one-locus marginal risks $\bar{f}_i(k)$ are known for each locus $i$ and genotype $k$. Necessarily,

$$(1 - p_i)^2 \bar{f}_i(0) + 2p_i(1 - p_i)\bar{f}_i(1) + p_i^2 \bar{f}_i(2) = K,$$

for all loci $i$.

Recurrence risk and the induced multiplicative model: To characterize a risk model, we begin with the recurrence risk, $K_R$, which is the risk of the disease to a relative
with relationship $R$ of an individual that has the disease. Risch’s (1990) recurrence risk ratio, $\lambda_R$, is defined as $K_R/K$. The increase in $K_R$ over $K$ indicates the effect of causative alleles shared by the two relatives because of identity by descent. In our notation, the recurrence risk corresponding to a risk model $f = (f(k))_{k \in \{0,1,2\}}^L$ is given by
\[
K_R(f) = \frac{1}{K} \left[ \sum_{k \in \{0,1,2\}^L} \sum_{k' \in \{0,1,2\}^L} f(k) f(k') \Pr(k,k' \mid R) \right],
\] (4)
where $\Pr(k,k' \mid R)$ denotes the joint probability of the $L$-locus genotypes of two relatives with relationship $R$. In this paper, we will be concerned with the recurrence risk in full siblings, parents and offspring, half siblings, and first full cousins. Since we assume that all $L$ loci are unlinked, the joint probability can be decomposed as
\[
\Pr(k,k' \mid R) = \prod_{i=1}^L \Pr(k_i,k'_i \mid R),
\]
where $\Pr(k_i,k'_i \mid R)$ is the marginal joint probability for locus $i$. See Liu and Weir (2005) for details on computing $\Pr(k_i,k'_i \mid R)$ for various relationships.

We define for any risk model the multiplicative model with the same average risk $K$ and marginal risks $\bar{f}_i(k)$. We call this model the induced multiplicative model. It satisfies
\[
f_I(k_1, \ldots, k_L) = \frac{1}{KL-1} \prod_{i=1}^L \bar{f}_i(k_i).
\] (5)
The recurrence risk for the induced model $f_I = (f_I(k))_{k \in \{0,1,2\}}^L$ is denoted by
\[
K^I_R := K_R(f_I) = \frac{1}{KL-1} \prod_{i=1}^L K^{(i)}_R,
\] (6)
where the one-locus recurrence risk $K^{(i)}_R$ for locus $i$ is defined as
\[
K^{(i)}_R = \frac{1}{K} \left[ \sum_{k=0}^2 \sum_{k'=0}^2 \bar{f}_i(k) \bar{f}_i(k') \Pr(k,k' \mid R) \right].
\] (7)
The quantity we use to characterize the deviation of a risk model $f$ from the induced multiplicative model is the interaction ratio

$$C_R(f) = \frac{K_R(f)}{K^I_R}. $$

For a multiplicative model, $C_R = 1$. If $C_R < 1$, then $K_R$ is smaller than expected under a comparable multiplicative model. In that case, assuming a multiplicative model would overestimate the actual heritability of risk as measured by $K_R$. If $C_R > 1$, assuming a multiplicative model would underestimate the actual heritability of risk, in which case, gene interactions are concealing some of the heritability.

The space of disease risk models: In what follows, we provide a geometric description of the space of disease risk models that are consistent with given constrains. Since $f(k) \in [0, 1]$ for all $k$, the $3^L$-tuple $f = (f(k))_{k \in \{0, 1, 2\}^L}$ takes value in the $3^L$-dimensional unit hypercube, denoted $Y$. Note that different points in $Y$ correspond to different disease risk models.

For each locus $i \in \{1, \ldots, L\}$ and genotype $k \in \{0, 1, 2\}$, (2) relates the disease risk function $f$ to the marginal risk $\bar{f}_i(k)$. More explicitly, we have

$$\bar{f}_i(k) = \sum_{k=(k_1, \ldots, k_L): k_i=k} \left[ f(k) \prod_{j \neq i} q_{j, k_j} \right], \quad (8)$$

where $q_{j, k_j}$ denotes the equilibrium frequency of genotype $k_j$ at locus $j$, given by

$$q_{j, k_j} = \begin{cases} 
(1 - p_j)^2, & \text{if } k_j = 0, \\
2p_j(1 - p_j), & \text{if } k_j = 1, \\
p_j^2, & \text{if } k_j = 2.
\end{cases}$$
Note that the $3L$ equations in (8) are not all independent, since, for each $i \in \{1, \ldots, L\},$

$$K = \sum_{k=0}^{2} q_{i,k} f_i(k) = \sum_{k=(k_1, \ldots, k_L) \in \{0,1,2\}^L} f(k) \prod_{j=1}^{L} q_{j,k_j}.$$  \hspace{1cm} (9)

Now, suppose that $K$, $p_i$, and $\bar{f}_i(k)$ are given, for each locus $i$ and genotype $k$. Then, (8) and (9) define $2L+1$ linearly independent affine hyperplanes $H_1, \ldots, H_{2L+1}$ in the unit hypercube $Y$, and the intersection of those hyperplanes restricted to $Y$ defines the space $\Delta$ of disease risk models consistent with the given information; i.e.,

$$\Delta = Y \cap H_1 \cap \cdots \cap H_{2L+1}.$$  

Note that $\Delta$ is a convex polytope of dimension $3^L - 2L - 1$ embedded in the $3^L$-dimensional $Y$. The induced model with the risk function $f_I(k)$ shown in (5) is consistent with the given constraints (8) and (9), and it corresponds to a particular point in the polytope $\Delta$.

In our study, instead of specifying the marginal risks $\bar{f}_i(k)$ directly, we assume a multiplicative model within each locus and specify the genotype relative risk $GRR_i$ for each locus $i$. Under this assumption, the marginal risks are related by $\bar{f}_i(1) = GRR_i \bar{f}_i(0)$ and $\bar{f}_i(2) = GRR_i^2 \bar{f}_i(0)$, and the first equality in (9) implies

$$\bar{f}_i(0) = \frac{K}{(1 - p_i)^2 + 2p_i(1 - p_i) GRR_i + p_i^2 GRR_i^2}.$$  \hspace{1cm} (10)

Since $\bar{f}_i(0), \bar{f}_i(1), \bar{f}_i(2)$ are linear in $K$, note that $K'_R$ defined in (6) is proportional to $K$.

**Sampling from the polytope:** The main computational problem involved in our work concerns computing the expectation $E[C_R(f)]$ with respect to the uniform distribution over the polytope $\Delta$. The dimension of the polytope grows exponentially with the number of loci, rendering analytic integration intractable for more than 2 loci. Hence, we employ a Markov chain Monte Carlo algorithm to estimate the expectation.
Over the past 20 years, there has been a series of theoretical developments (e.g., see Dyer et al. (1991); Kannan et al. (1997); Lovász and Vempala (2006)) on fully polynomial-time randomized approximation schemes for computing the volume of convex bodies in $\mathbb{R}^n$. The key component of these algorithms concerns the problem of sampling points uniformly at random from convex bodies. The convexity property implies that one can devise an MCMC algorithm with acceptance rate 1. In our work, we employed the hit-and-run sampling algorithm (Smith, 1984; Lovász, 1999), which goes as follows: Suppose that the current state in the Markov chain is $x_t \in \Delta$. To sample the next state, choose a direction uniformly at random and move uniformly along that direction, restricted to $\Delta$. The point after the move is the next state $x_{t+1} \in \Delta$.

The mixing time of the hit-and-run algorithm has been shown to be $O^*(n^3)$, after some appropriate preprocessing (Lovász, 1999), where $n$ denotes the dimension of the ambient space. We remark that there are at least two sources of complication in applying this complexity result in practice. First, it is an asymptotic result and the $O^*$ notation actually hides a large constant, as well as the dependence on error parameters. Second, in the application we are considering, the dimension $n$ grows exponentially with the number $L$ of loci. Hence, to check convergence, we took an empirical approach, by keeping track of the running average of $C_R(f)$.

For $L = 2, 3, 4$, we ran our sampler for 20 million iterations, taking samples every 20 iterations to compute the expectation $E[C_R(f)]$. For $L = 5$, we used 30 million iterations, again taking samples every 20 iterations to compute $E[C_R(f)]$. For a few parameter settings, we tried using 100 million iterations and obtained results very close to that for 30 million iterations. In all runs, we started the chain at the induced multiplicative model $f_I$. The
computation was done using a single core of a Mac Pro with two 3.0 GHz Quad-Core Intel Xeon processors, and the running time for a given parameter setting was 88 seconds for \( L = 2 \); 822 seconds for \( L = 3 \); 2.2 hours for \( L = 4 \); and 28.3 hours for \( L = 5 \).

RESULTS

Random sampling of risk models: With \( L \) diallelic loci, there are \( 3^L \) genotypes and hence \( 3^L \) degrees of freedom for an arbitrary risk function. If we fix the marginal risks at each locus, the risk function is subject to \( 2L + 1 \) constraints (Culverhouse et al., 2002). That can be seen by noting first that there are \( 3L \) marginal risks but that (3) implies that only two of those marginal risks are independent. The average risk, \( K \), must be specified also, adding one more constraint. Even for small \( L \), there is a large range of risk functions subject to these constraints. There are 4 degrees of freedom in the risk function for \( L = 2 \); 20 for \( L = 3 \); 72 for \( L = 4 \); and 232 for \( L = 5 \). To explore efficiently the space of feasible risk functions subject to these constraints, we implemented an MCMC algorithm described in the previous section. That algorithm allowed us to randomly sample risk functions from the space of feasible risk functions subject to the given constraints. In our empirical study, instead of specifying the marginal risks directly, we assumed a multiplicative model within each locus and specified the genotype relative risk \( GRR_i \) for each locus \( i \). Recall that specifying \( K \), \( p_i \), and \( GRR_i \) completely fixes the marginal risks for locus \( i \); see (10).

Figure 1 summarizes the results for full and half siblings. The results for parents and offspring are similar to those for full siblings. For full first cousins, the overall effect of gene interactions is weaker than for half siblings but the average \( C_R \) still increases with \( L \). The
qualitative pattern for \( K = 0.005 \) is the same for both low frequency (\( p = 0.03 \)) and higher frequency (\( p = 0.25 \)) causative alleles but the average of \( C_R \) is larger for rare alleles. Similar results were found for \( K = 0.02 \). The average \( C_R \) increases with \( L \) and the probability that \( C_R > 1 \) is nearly 1 even for \( L = 2 \). That is, gene interactions will tend to make the true similarity of relatives larger than expected if the multiplicative model is assumed to be correct, and the effect is more pronounced when causative alleles are rarer.

The results in Figure 1 were based on the assumption that all loci have the same genotype relative risks and allele frequencies. When there are both rare and common alleles, the increase in average \( C_R \) increases with the proportion of rare alleles, except when \( L = 2 \). See Figure 2.

**Non-random sampling of risk models:** The above results were obtained by sampling all feasible risk functions with equal probability. On intuitive grounds, however, it is reasonable to assume that risk functions which are closer in some sense to the induced multiplicative model are more likely alternatives than risk functions which are more unlike the induced multiplicative model. That assumption embodies the intuition that, although genes are not completely independent in their effect on risk, they are more likely to be partially independent than they are to interact in a completely arbitrary way. To test whether taking account of the similarity to a multiplicative model affects the above conclusions, we define a distance between a risk function \( f \) and the induced multiplicative risk function \( f_I \) to be

\[
d = \sqrt{\sum_{k_1,\ldots,k_L=0}^{2} [f(k_1,\ldots,k_L) - f_I(k_1,\ldots,k_L)]^2}. \tag{11}
\]

This is a Euclidean distance in the ambient space with \( 3^L \) dimensions.

In exploring the space of feasible models, we computed the \( d \) for each model. The first
part of Figure 3 shows the cumulative distribution of \( d \) for one of the sets of parameter values. Other results are similar. The second part of Figure 3 shows the conditional average of the interaction ratio \( C_R \) for full siblings, given that the distance \( d \) from the induced multiplicative model is within the \( y \)th percentile. The graph shows that, except for risk models that are very close to the multiplicative model, the average of \( C_R \) is nearly the average obtained by randomly sampling risk models. In other words, even models of risk that are similar to the induced multiplicative model conceal on average substantial amounts of heritability.

The minimum of \( C_R \): The extrema of \( C_R \) over the space of feasible disease models can be computed numerically using quadratic programming methods. Surprisingly, it turns out that we can actually obtain a closed-form formula for the minimum of \( C_R \) by considering the following interpretation of the recurrence risk (4): Given an individual \( a \) in a population, let \( X_a \) denote an indicator random variable such that \( X_a = 0 \) (respectively, \( X_a = 1 \)) if the individual is unaffected (respectively, affected) by a given complex disease. Then, for two relatives \( a \) and \( b \) with relationship \( R \), the recurrence risk (4) can be written as

\[
K_R = K + \frac{\text{Cov}(X_a, X_b| R)}{K}.
\]

Since the recurrence risk \( K^I_R \) for the induced multiplicative model is completely determined by the given constraints, minimizing \( C_R = K_R/K^I_R \) is equivalent to minimizing \( K_R \), which in turn translates to minimizing the covariance term \( \text{Cov}(X_a, X_b| R) \). Now, since \( \text{Cov}(X_a, X_b| R) \) can be partitioned into additive, dominance, and interaction variance components (James, 1971)—which are all non-negative—we conclude that the minimum \( \text{Cov}(X_a, X_b| R) \) is attained by a model with vanishing interaction variances. In such a
model, \( \text{Cov}(X_a, X_b \mid R) \) is given by a sum of one-locus variance components over the \( L \) loci:

\[
\frac{\text{Cov}(X_a, X_b \mid R)}{K} = \sum_{i=1}^{L} \left[ K^{(i)}_R - K \right].
\]

In summary,

\[
\min C_R = \frac{K + \sum_{i=1}^{L} \left( K^{(i)}_R - K \right)}{\sum_{i=1}^{L} K^{(i)}_R} = K^{L-1} \left[ \frac{K + \sum_{i=1}^{L} \left( K^{(i)}_R - K \right)}{\prod_{i=1}^{L} K^{(i)}_R} \right].
\]  \( \text{(12)} \)

In the symmetric case with \( p_i = p \) and \( \bar{f}_i(k) = \bar{f}(k) \), for all loci \( i \), the risk function \( f(k) \) that minimizes \( C_R \) is given by

\[
f(k_1, \ldots, k_L) = n_0a_0 + n_1a_1 + n_2a_2,
\]

where \( n_k \) is the number of loci with genotype \( k \) and \( a_k = \bar{f}(k) - \frac{K^{(L-1)}}{L} \).

It can be checked that numerical optimization based on the method of Lagrange multipliers produces results which coincide with (12). In the cases we considered in this paper, the marginal risks \( \bar{f}_i(0), \bar{f}_i(1), \bar{f}_i(2) \) are linear in \( K \). Therefore, \( K^{(i)}_R \) are proportional to \( K \), and hence the minimum of \( C_R \) is independent of \( K \). Shown in Figure 4 are plots of the minimum of \( C_R \) for \( p_i = p \) and \( \text{GRR}_i = 1.25 \) for all loci \( i \). Note that the minimum of \( C_R \) decreases as the number of loci increases, and that the rate of decrease is more rapid for larger values of \( p \). It is interesting that the average of \( C_R \) increases with the number \( L \) of loci, despite the fact that the minimum of \( C_R \) decreases as \( L \) increases. Computing the maximum of \( C_R \) is a more difficult problem, both numerically and analytically, and we do not address it in this paper. For \( L \leq 4 \) and special values of \( K \) and \( p \), Culverhouse et al. (2002) estimated the maximum of \( C_R \) for purely epistatic models, in which case \( \bar{f}_i(k) = K \) for all \( i = 1, \ldots, L \) and \( k = 0, 1, 2 \).
DISCUSSION

In human genetics, gene interactions are regarded with ambivalence. On one hand, what has been learned about metabolic pathways and gene-regulatory networks makes clear that genes and gene products may interact in so many ways that independent effects of individual genes are probably the exception rather than the norm (Cordell, 2002; Carlberg and Haley, 2004; Moore and Williams, 2005; Phillips, 2008). On the other hand, tradition, convenience, and a large body of evidence from quantitative genetic studies encourage the use of models of disease risk that assume causative loci are independent in their effects or are additive on an underlying scale of disease liability (Risch, 1990; Hill et al., 2008; Wray and Goddard, 2010). Although many statistical methods have been developed to detect interacting genes in GWAS data sets (Marchini et al., 2005; Zhao et al., 2006; Zhang and Liu, 2007; Gayán et al., 2008), those and similar methods are intended to test for specific pairs of loci that interact and not the overall importance of gene interactions to the average risk and heritability of complex diseases.

In this paper, we have quantified the potential effects of gene interactions. We have shown that, whether or not there are detectable marginal effects of individual loci, gene interactions will tend to increase the heritability of disease risk as measured by the recurrence risk, $K_R$, from what is expected based on the assumption of no gene interactions. Furthermore, a greater increase in heritability is expected if there are more causative loci and when causative alleles are rare rather than common, and the pattern is found even with models of disease risk that are closer than average to a model that assumes no gene interactions. Therefore, relatively weak interactions among multiple causative loci could conceal substantial
The magnitude of the effect of interactions depends on the frequencies of causative alleles. To illustrate what our results mean, consider the case with low frequency causative alleles ($p = 0.03$), each of which increases disease risk by a factor of 1.25. If there are 5 such loci, the interaction ratio, $C_R$ for full siblings is about 1.57 (see Figure 2) for $K = 0.005$. If these 5 loci were identified and the genotypic relative risks (GRRs) estimated and a multiplicative model of interactions is assumed, we could conclude that these loci together would cause the recurrence risk $K_R$ to exceed $K$ by a factor of 1.01. As a consequence, we would conclude that these 5 loci contribute very little to the increase risk to full siblings of an affected individual. However, our results indicate that, if the interaction model is sampled randomly from the space of models with the same GRRs, these loci would cause $K_R$ to exceed $K$ by an average of 1.59. Furthermore, that would be true even if the risk model were chosen to be close to a multiplicative model in the sense we defined above. In this example, then, most of the increase in the risk to close relatives would be concealed if the true model were not multiplicative. The effect is weaker for higher frequency risk alleles. If $p = 0.25$, $C_R$ for full siblings is about 1.14. The increase in $K_R$ is 1.05 if a multiplicative model is assumed. The expected increase for a randomly chosen model of interactions is about 1.20.

Untyped rare causative alleles may enhance the potential role of epistasis in explaining missing heritability. As our results show, the effect of epistasis on heritability is more pronounced for rare causative alleles than for common ones. Suppose $c$ common causative alleles were typed in a GWAS, while $r$ rare causative alleles were not typed. Our results (see Figure 2) suggest that the recurrence risk $K_R$ for the $c+r$ loci can be substantially greater than that for the $c$ common causative alleles only, and that this difference in general grows with
the number $r$ of untyped rare causative alleles. In contrast, the recurrence risk $K'_{R_i}$ in the
induced multiplicative model will be similar in the two cases.

Our analysis cannot tell us whether gene interactions actually contribute to disease risk
and the heritability of complex disease. Only empirical studies can do that. But our results
indicate that there is ample room for gene interactions to have a strong effect on disease
heritability. It is not necessary to assume extreme or bizarre types of gene interaction. Most
deviations from independent gene action have an important effect on heritability, so the
possibility of extensive interactions should be kept in mind when interpreting the results of
GWAS and family studies of complex inherited diseases.

ACKNOWLEDGMENTS

The research was supported in part by an NIH grant R00-GM080099, an Alfred P. Sloan
Research Fellowship, and a Packard Fellowship for Science and Engineering to YSS, and by
an NIH grant R01-GM40282 to MS.


Figure 1: Expected increase in the interaction ratio \( (C_R) \) and the fraction of models for which \( C_R > 1 \) as functions of the number of causative loci, for low and high frequency alleles. The average \( C_R \) increases with the number of loci and the probability that \( C_R > 1 \) is nearly 1 for all loci. The effects of gene interactions are more pronounced when causative alleles are rarer. We used \( K = 0.005 \) and \( p_i = p \) for all loci \( i \). Within each locus, we assumed a multiplicative model and set the genotype relative risk \( (GRR) \) to 1.25. For all loci \( i \), the corresponding marginal risks are \( (\bar{f}_i(0), \bar{f}_i(1), \bar{f}_i(2)) = (0.0049, 0.0062, 0.0077) \) for \( p = 0.03 \) and \( (\bar{f}_i(0), \bar{f}_i(1), \bar{f}_i(2)) = (0.0044, 0.0055, 0.0069) \) for \( p = 0.25 \). See (10).
Figure 2: Expected increase in the interaction ratio as a function of the number of loci with rare causative alleles, given the total number of loci is $L$. Except for the case of $L = 2$, when there are both rare and common alleles, the increase in average $C_R$ increases with the proportion of rare alleles. The same parameter values as in Figure 1 are assumed here. In particular, rare risk alleles are assumed to have frequency $p = 0.03$, while common risk alleles are assumed to have frequency $p = 0.25$. 
Figure 3: Effects of non-random sampling of risk models for $L = 4$. The left plot shows the cumulative probability that a randomly sampled risk model has a Euclidean distance $d$ (measured as a fraction of the empirical maximum possible distance $d_{\text{max}}$) from the induced multiplicative model for low and high frequency causative alleles. The right plot shows the conditional expectation of the interaction ratio $C_R$ for full siblings, given that the distance from the induced multiplicative model is within the $y$th percentile. This result implies that even models of risk that are similar to the induced multiplicative model conceal on average substantial amounts of heritability. The same parameter values as in Figure 1 are assumed here.
Figure 4: A graph of the minimum interaction ratio $C_R$ for full siblings plotted against the number of causative loci. The minimum of $C_R$ does not depend on $K$. Also, it decreases as the number of loci increases, and the rate of decrease is more rapid for larger values of $p$. It is interesting that, although the minimum of $C_R$ decreases as the number $L$ of loci increases, the average of $C_R$ over the polytope increases with $L$. 