Two-Locus Heterogeneity Cannot Be Distinguished from Two-Locus Epistasis on the Basis of Affected-Sib-Pair Data

Veronica J. Vieland[1,2] and Jian Huang[1,3]

[1]Division of Statistical Genetics, Department of Biostatistics, College of Public Health, [2]Department of Psychiatry, Roy J. and Lucille A. Carver College of Medicine, and [3]Department of Statistics and Actuarial Science, University of Iowa, Iowa City

The observation of multiple linkage signals in the course of conducting genome screens for complex disorders raises the question of whether distinct genes represent independent causes of disease (heterogeneity) or whether they interact to produce the phenotype of interest (epistasis); and there has been a corresponding interest in statistical methods for detecting and/or exploiting the distinction between these two possibilities. At the same time, researchers are increasingly relying on affected-sib-pair (ASP) data. Here, we demonstrate an apparently unrecognized fact about two-locus (2L) models and ASP data, namely, 2L heterogeneity and 2L epistasis cannot, in general, be distinguished from one another on the basis of ASP marker data, as a matter of mathematical principle and therefore regardless of sample size. By the same token, correlations across ASPs in single-locus LOD scores or other measures also cannot be used to distinguish 2L heterogeneity from 2L epistasis. This raises questions about the measurement of gene-gene interactions in terms of patterns of correlation in marker data. Portions of our results carry over to larger pedigree structures as well, as long as only affected individuals are included in analyses; the extent to which our overall findings apply to general pedigrees (including unaffected individuals) remains to be investigated.

Introduction

The observation of multiple linkage signals in the course of conducting genome screens for complex disorders raises the question of whether distinct genes represent independent causes of disease (heterogeneity) or whether they interact to produce the phenotype of interest (epistasis); and there has been a corresponding interest in statistical methods for detecting and/or exploiting the distinction between these two possibilities. However, the subject is complicated, for two reasons. The first is that, although the terms "locus heterogeneity" and "epistasis" have straightforward definitions in classical genetics, it is nevertheless not obvious how these terms should be defined at the mathematical level for general (human) applications. For example, many authors (e.g., Durner et al. 1992; MacLean et al. 1993) consider two-locus (2L) models in which the penetrances are the same for disease-genotype carriers at each of two loci, as well as for disease-genotype carriers at both loci, as heterogeneity models. If, however, the vernacular meaning of "heterogeneity" involves independent gene effects at each locus, then surely the penetrance for carriers at both loci should be higher than the penetrance for carriers at one but not the other (see below for mathematical details). The definition of heterogeneity used by Risch (1990) precludes these equal-penetrance models, but it was also unclear to us whether his definition corresponded to the usual genetic concept of heterogeneity (again, see below for mathematical details).

It has also become commonplace to use the terms "epistasis" and "gene-gene interaction" interchangeably, and statistical measures of gene-gene interaction are typically based on some aspects of the correlational structure of the genotypic data across loci (e.g., Cox et al. 1999; Holmans 2002). But it was not at all clear to us initially how such correlations in genotypes related to the classical genetic concept of epistasis. Other definitions of epistasis entail "multiplicativity" in the penetrance matrices (see Hodge 1981; Risch 1990); but this class of models has mathematical properties that conflict with methods for detecting gene-gene interaction on the basis of positive correlations, since these models predict no correlation in the relevant (marginal) quantities (see Hodge 1981; MacLean et al. 1993).

Thus, the first difficulty in considering statistical methods for distinguishing heterogeneity from epistasis is simply to derive mathematical expressions that correspond appropriately to the ordinary genetic concepts. At the outset, it was unclear to us which, if any, of the mathematical models in the literature did, in fact, correspond to the genetic concepts of interest to us.

The second complication for any attempt to distinguish

Received February 6, 2003; accepted for publication April 21, 2003; electronically published June 26, 2003.

Address for correspondence and reprints: Dr. Veronica Vieland, Center for Statistical Genetics Research, 2190 Westlawn Building, University of Iowa Colleges of Public Health and Medicine, Iowa City, IA 52242-1008. E-mail: veronica-vieland@uiowa.edu

0002-9297/2003/7302-0002$15.00

classes of multilocus models from one another in human genetics is simply the number of parameters involved. For instance, the full 2L model (for a dichotomous trait) involves 11 parameters (two disease-allele frequencies and nine penetrances; see below for details). The values of these parameters are seldom known in advance, and it is not clear whether heterogeneity and epistasis can be truly distinguished by any method that involves fixing these parameters at arbitrary (incorrect) values. At the same time, it is not clear how many of these parameters can, in fact, be estimated from typical data.

In addition to these inherent difficulties, researchers are increasingly relying on affected-sib-pair (ASP) data rather than on the larger pedigrees (including full nuclear families) favored by an earlier generation of gene mappers. This trend exacerbates the difficulties of parameter estimation, since the number of parameters that can be estimated from such data is far fewer than the total parameters in a full 2L model. It might, however, still be the case that some summary measures, such as correlations, could be used to detect epistasis; indeed, the current literature on gene-gene interactions tends to focus on methods for ASP data (e.g., Holmans 2002). Similarly, MacLean et al. (1993) demonstrated the feasibility of estimating a "degree of epistasis" parameter from sibship data, and this method works in ASP data as well. But how is it possible to distinguish complex classes of 2L models when estimation of the underlying parameters is moot? Are these techniques really accomplishing what they set out to do?

In the present article, we rigorously investigate the feasibility of distinguishing, in typical human data sets, 2L heterogeneity (2L HET) from 2L epistasis (2L EPI) for dichotomous traits. We focus on ASP data, partly because of the current popularity of ASP designs and the proliferation of methods for detecting gene-gene interaction on the basis of ASP data and partly for purposes of mathematical clarity; we comment on extensions to general pedigree data as we go. We begin with the genetic concept of locus heterogeneity and derive the corresponding mathematical expression. We then prove that heterogeneity can never be established on the basis of ASP (or any affecteds-only) data, for reasons having to do with the underlying mathematical structure of the models themselves; and we show why both the approach of MacLean et al. (1993) and the methods based on correlations (e.g., Holmans 2002) are not addressing epistasis in the usual genetic sense. Our results also show, however, that there is a class of epistatic models that could, in principle, be distinguished from heterogeneity, provided that it were possible to estimate certain aspects of the penetrance structure of the model from the data. We then demonstrate that, as a general rule, this cannot be done on the basis of ASP data, as a matter of mathematical principle and therefore regardless of sample size.

Methods

Assumptions and Notation

Throughout the present article, we restrict our attention to dichotomous traits. For clarity of exposition, we assume that there are exactly two trait loci (the A locus and the B locus), each with 2 alleles (A and a, B and b, respectively); and we assume that these alleles are not directly observed. Define the allele frequencies $p_A = \mathrm{P(A)}, q_\mathrm{A} = 1 - p_\mathrm{A} = \mathrm{P(a)}$; and similarly for $p_\mathrm{B}, q_\mathrm{B}$. We assume that there is no clinical basis for differentiating phenotypic effects of the A locus and the B locus; and we assume that the trait loci are unlinked to one another. We assume that parental genotypes are known at each of two marker loci, $\mathrm{M_A}$ (linked to the A locus) and $\mathrm{M_B}$ (linked to the B locus) and that the two markers are also unlinked to one another; and we assume throughout that all matings are fully informative at the marker loci, so that identity-by-descent (IBD) sharing can be directly scored for each ASP. We assume linkage equilibrium between each marker and the trait, as well as between the two markers. Finally, we restrict our attention to a subclass of 2L models, in which two of the three marginal penetrances at each locus are equal to one another. This restriction enables us to describe each of the two loci as either dominant (D) or recessive (R), which greatly simplifies the exposition; however, it can be relaxed without altering our fundamental findings.

For ASP data, under these assumptions, all of the genotypic information conveyed by two markers can be captured by the 3 × 3 matrix of joint (two marker) observed IBD sharing; and a 2L model can accordingly be represented by the 3 × 3 matrix of joint IBD-sharing probabilities at the two marker loci. In general applications, for a 2L model, the nine entries in this IBD-sharing matrix are, in turn, functions of 13 underlying parameters: the recombination fractions between the A locus and $\mathrm{M_A}$ and between the B locus and $\mathrm{M_B}$, respectively; the disease-allele frequencies at each of the two loci; and nine penetrances, corresponding to each possible 2L genotype.

However, under our simplifying restriction on the penetrance structure, there are at most four distinct penetrances in the model. If we adopt, for the moment, the short-hand "carrier" to denote individuals who carry the disease genotype (aa in the case of a R model and AA or Aa in the case of a D model at the A locus; and similarly for the B locus), then these four penetrances become f_A (the penetrance for carriers at the A locus but not the B locus), f_B (the penetrance for carriers at the B locus but not the A locus), f_AB (the penetrance for individuals who are carriers at both loci), and f_P (the penetrance for noncarriers, which is sometimes called the "phenocopy rate"). Note that, as defined here, f_A and f_B apply only to carriers at one locus or the other but not both. Table 1 illustrates this parameterization in the case of a RR model.

Table 1

Penetrances for the Joint A Locus/B Locus Genotypes for the Class of Two-Locus Models Considered in the Present Article, Illustrated for an RR Model

Genotype	AA	Aa	aa
BB	f_P	f_P	f_A
Bb	f_P	f_P	f_A
bb	f_B	f_B	f_{AB}

NOTE.—We consider a model to represent 2L HET if and only if $f_{AB} = f_A + f_B - f_A f_B$; otherwise we consider the model to represent 2L EPI.

Definitions of Heterogeneity and Epistasis

We define "2L HET" as any model in which the genotype at each locus influences the phenotype independently of the genotype at the other locus. This *genetic* definition motivates our derivation of a mathematical definition, as follows: Let K be the total prevalence of disease, K_A be the prevalence due to the action of genotype at the A locus, and K_B be the prevalence due to the B locus. (We rigorously define these terms below.) Then, in the absence of any other causes of disease in the population, we seek a definition of 2L HET that produces a particular structure in the prevalence model, namely, one in which K is defined by the elementary probabilistic relationship

$$K = K_A + K_B - K_A K_B \ , \tag{1}$$

which is the general expression for the probability of a union of two independent events.

This expression in terms of prevalence determines a corresponding relationship among the penetrances. Let q_{ij} be the probability of the ith trait genotype at the A locus and the jth trait genotype at the B locus, and let f_{ij} be the corresponding penetrances. Then $K = \sum_i \sum_j q_{ij} \times f_{ij}$, where the sums are taken over all possible genotypes i and j at the A and B loci respectively. On the basis of this expression, it is readily confirmed by simple algebra that, to achieve the structure expressed by equation (1), a specific relationship must obtain among the penetrances as defined above. For example, when we consider an RR model as an illustration, under Hardy-Weinberg equilibrium, we have

$$\begin{aligned} K &= q_A^2(1-q_B^2)f_A + (1-q_A^2)q_B^2 f_B + q_A^2 q_B^2 f_{AB} \\ &= q_A^2 f_A + q_B^2 f_B - q_B^2 q_B^2 (f_A + f_B - f_{AB}) \ . \end{aligned}$$

Defining $q_A^2 f_A = K_A$ and $q_B^2 f_B = K_B$ and substituting in $(f_A + f_B - f_A f_B)$ for f_{AB} yields the desired structure $K = K_A + K_B - K_A K_B$.

The previous sentence contains definitions of K_A and K_B that are more rigorous than the informal ones we started with ("prevalence due to the A or B locus"), but it may, on first blush, appear somewhat odd, since f_A and f_B are not true marginal penetrances (which would each be a function of allele frequencies at the other locus). We started with a biological definition of "heterogeneity" in terms of independent gene action (the effects of a mutation at one gene do not depend on mutation status at the other gene); but the probabilistic definition of independence dictates that the penetrance for carriers at one locus will, in fact, be a function of carrier status at the other. Thus, equation (1) dictates that, to satisfy the biological definition of "heterogeneity," f_A and f_B—which are by definition the true penetrances for carriers at one but not the other locus—must also be the portion of the penetrance for carriers at both loci that can be attributed to the action of the A (or B) gene, or, more succinctly, the attributable penetrances for their respective loci.

Another way of putting this is to note that, from a genetic point of view, a heterogeneity model should be one in which the trait model at each individual locus can be written as an ordinary single-locus model, without needing to incorporate parameters of the model at the other locus. In this representation of the single-locus component of the 2L model, then, penetrance can no longer have the usual interpretation as simply the probability of being affected, because this quantity will involve parameters of the other locus. Rather, our results show that 2L HET entails writing these single-locus components in terms of the somewhat more abstract concept of "attributable penetrances."

More generally, for any simple dominance model (RR, RD, or DD), our genetic definition of "heterogeneity" will require that the penetrances observe the relationship

$$f_{AB} = f_A + f_B - (f_A \times f_B) \tag{2}$$

We therefore define "2L HET" as any 2L model that satisfies equation (2); and we refer to equation (2) as the "fundamental heterogeneity equation" (FHE).

Although our definition of 2L HET is designed to capture what we believe is commonly meant in the human genetics literature by the term "heterogeneity," we note that it differs from some definitions used elsewhere. For instance, except when $f_A = f_B = 1$, under our definition a penetrance matrix in which $f_{AB} = f_A = f_B$ represents (negative) epistasis rather than heterogeneity; it corresponds to a deficit in prevalence due to carriers of both disease genotypes, relative to what would be expected were the two genes to act in a probabilistically independent manner, and it does not satisfy the FHE. However, other authors do consider such cases as heterogeneity models (see Durner et al. 1992; MacLean et al. 1993).

Also, although our definition of 2L HET superficially resembles that of Risch (1990), his is derived from penetrance quantities that coincide neither with our terms f_A and f_B nor with the true marginal penetrances (which would be functions of allele frequencies); and his definition precludes, for instance, models with no phenocopies ($f_P = 0$). Our definition of 2L HET also differs from his definition of an additive model.

We define "2L EPI," by contrast with "2L HET," as any relationship among the penetrances that does not satisfy the FHE (eq. [2]). This definition allows for either a deficit or an excess penetrance for carriers of both disease genotypes, relative to what is predicted by the FHE. We note that this definition of 2L EPI may differ from the classical concept of epistasis, in which the only positive penetrance in the model applies to individuals who are carriers at both trait loci. By contrast, our definition allows for varying degrees of epistasis. Alternatively, we can think of this as allowance for degrees of interlocus dominance, by analogy with allowance for varying degrees of intralocus dominance. Thus, we can have "negative 2L EPI," in which $f_{AB} < f_A + f_B - (f_A \times f_B)$; "positive 2L EPI," in which $f_{AB} > f_A + f_B - (f_A \times f_B)$; or "complete 2L EPI" (classical epistasis), in which all penetrances except f_{AB} are 0. Of particular interest is the relationship between our definition of 2L EPI and definitions based on positive correlations in marginal IBD sharing (Cox et al. 1999; Holmans 2002). We return to this topic below.

Results

We derive the results in stages, beginning with a very simple model and progressing to a fairly general model. Accordingly, we divide the "Results" section into four subsections, as follows: in the section "Estimation of Degree of Epistasis," we consider the procedure of MacLean et al. (1993) for estimating degree of epistasis in a special elementary subclass of 2L models, and we show that it cannot be used to establish 2L HET; in the section "2L HET Can Never Be Established on the Basis of ASP Data," we show, using a slight generalization of the MacLean model, that this difficulty is inherent in the properties of the 2L HET model whenever the data comprise affected individuals only and, therefore, that 2L HET can never be established on the basis of ASP data; in the section "But Can We Sometimes Establish 2L EPI?," we investigate whether it is ever possible to establish 2L EPI based on ASP data, for the general class of 2L models as defined above, and show that, in general, aspects of the penetrance structure of the model necessary to do so cannot be estimated from ASP data; finally, in the section "Correlations and Gene-Gene Interaction," we consider the implications of these results for the interpretation of gene-gene interactions in ASP data defined in terms of marginal correlations in 2L IBD data, again, for fully general models.

Estimation of Degree of Epistasis

For the purposes of this section and the following section ("2L HET Can Never Be Established on the Basis of ASP Data"), we make certain additional simplifying assumptions. First, we assume that the recombination fraction between each locus and its corresponding (linked) marker is 0 and that this is known to the investigator (i.e., that these parameters do not need to be estimated). We also make a further, unrealistic, simplifying assumption that all parents are doubly heterozygous at each of the two trait loci as well as at the marker loci; thus, trait-locus allele frequencies also cancel out of all likelihoods. These additional assumptions serve only to simplify the algebraic derivation of IBD-sharing probabilities from the generating penetrance matrix and the resulting IBD formulas; they have no bearing on our fundamental results.

We begin with the elegant parameterization of a 2L model of MacLean et al. (1993). (We have adapted their notation to conform to our own.) When $f_A = f_B$, the penetrance structure for the class of 2L models we are considering can be parameterized in terms of just two quantities: f_{AB} and a parameter λ (their original notation) representing degree of epistasis, such that $f_A = f_B = \lambda f_{AB}$. Table 2 shows the resulting penetrance table for a simple 2L RR model, which we use to illustrate the results of this section.

Note that, when $\lambda = 0$, only individuals who are both aa and bb have a positive probability of becoming affected. Thus, $\lambda = 0$ represents complete, or classical, epistasis; and, as λ increases, the degree of epistasis decreases. When $\lambda = 1$, the penetrance is the same whether an individual has the disease genotype at either or both loci. MacLean et al. (1993) called this case "heterogeneity." As noted above, however, models in which $\lambda = 1$ will not satisfy the FHE unless $f_{AB} = 1$. Thus, under our def-

Table 2

Penetrances for the Joint A-Locus/B-Locus Genotypes

Genotype	AA	Aa	aa
BB	0	0	$\lambda \times f_{AB}$
Bb	0	0	$\lambda \times f_{AB}$
bb	$\lambda \times f_{AB}$	$\lambda \times f_{AB}$	f_{AB}

NOTE.—Penetrances are for the special model considered in the "Estimation of Degree of Epistasis" section; for purposes of illustration, an RR model is assumed.

Table 3

IBD-Sharing Probabilities, P(Sharing *i* Marker Alleles IBD at the A Locus, *j* Marker Alleles IBD at the B Locus | ASP), for the RR Model Considered in the "Estimation of Degree of Epistasis" Section

No. of IBD Alleles	$i = 0$	$i = 1$	$i = 2$
$j = 0$	$\frac{\lambda^2}{\Delta}$	$\frac{2\lambda^2}{\Delta}$	$\frac{\lambda(1+\lambda)}{\Delta}$
$j = 1$	$\frac{2\lambda^2}{\Delta}$	$\frac{4\lambda^2}{\Delta}$	$\frac{2\lambda(1+\lambda)}{\Delta}$
$j = 2$	$\frac{\lambda(1+\lambda)}{\Delta}$	$\frac{2\lambda(1+\lambda)}{\Delta}$	$\frac{(1/2)(1+6\lambda^2)}{\Delta}$

NOTE.—Δ = sum of the numerators over all cells. See table 2 for the penetrance model; see the "Estimation of Degree of Epistasis" section for additional modeling assumptions.

inition of 2L HET, it is not immediately apparent what value of λ corresponds to complete absence of epistasis.

Table 3 shows the IBD-sharing probabilities among ASPs for this generating model, again illustrated by an RR model. We note that the parameter f_{AB} has canceled out of the constituent probabilities and does not appear in the table; it therefore also does not appear in the likelihood written as a function of these probabilities. The parameter λ, on the other hand, does appear in the table. Thus, although f_{AB} cannot be estimated from ASP data, the observed IBD sharing matrix does contain information relevant to estimation of the penetrance structure of the model via the parameter λ.

The question then becomes whether this information is sufficient to enable us to distinguish heterogeneity from epistasis. And here we encounter a difficulty, because the value of λ that satisfies the FHE depends upon the (unknown) true, underlying penetrance f_{AB}, through the formula

$$\text{FHE} \Rightarrow \lambda = \frac{1-\sqrt{1-f_{AB}}}{f_{AB}} \; .$$

Since $0 < f_{AB} \leqslant 1$, some algebra shows that this formula implies $0.5 < \lambda \leqslant 1$; however, for $\lambda > 0.5$, whether any particular value of λ satisfies the FHE depends upon f_{AB}. Thus, if we estimate $\lambda > 0.5$, we cannot know whether the model represents 2L HET or 2L EPI without knowing the value of f_{AB}, and f_{AB} itself cannot be estimated from the IBD data. However, if an accurate estimate of $\lambda \leqslant 0.5$ were obtained, we would know that the model was 2L EPI rather than 2L HET. We note that there is nothing special about the RR model in this respect; the same results are obtained for DD and DR models.

Thus, we have the peculiar result that, although it is indeed possible to estimate the degree of epistasis for this simple class of 2L models, when $\lambda > 0.5$, it is nevertheless not possible to establish that there is *no* epistasis. Moreover, these findings pertain to the true, underlying (or generating) IBD-sharing probabilities. Alternatively, thinking in terms of estimation of sharing probabilities based on a data set of ASPs, our reasoning is equivalent to assuming that the observed IBD data follow their exact, expected proportions, as would occur in very large samples. Thus, all conclusions drawn here apply to asymptotic statistical inference.

This simple illustration illuminates the fundamental underlying problem: the distinction between 2L HET and 2L EPI is a matter of the penetrance structure of the model, and ASPs contain limited information on penetrances. As a result, the type of information needed to differentiate the two classes of models is not necessarily present in the IBD data.

2L HET Can Never Be Established on the Basis of ASP Data

We now show that the basic result of the previous section is not merely an artifact of the extreme simplicity of the 2L model considered thus far. To do this, we need only generalize the model somewhat to allow for the case $f_A \neq f_B$. Thus, we return to the parameterization shown in table 1, but still fixing $f_P = 0$. For clarity of exposition, all other simplifying assumptions remain as in the preceding section, "Estimation of Degree of Epistasis."

We again show that there is a class of 2L models

Table 4

IBD-Sharing Probabilities, P(Sharing *i* Marker Alleles IBD at the A Locus, *j* Marker Alleles IBD at the B Locus | ASP), for the RR Model Considered in the "2L HET Can Never Be Established on the Basis of ASP Data" Section

No. of IBD Alleles	$i = 0$	$i = 1$	$i = 2$
$j = 0$	$\frac{f_A/f_{AB} \times f_B/f_{AB}}{\Delta}$	$\frac{2(f_A/f_{AB} \times f_B/f_{AB})}{\Delta}$	$\frac{f_A/f_{AB}(1+f_A/f_{AB})}{\Delta}$
$j = 1$	$\frac{2(f_A/f_{AB} \times f_B/f_{AB})}{\Delta}$	$\frac{4(f_A/f_{AB} \times f_B/f_{AB})}{\Delta}$	$\frac{2[f_A/f_{AB}(1+f_A/f_{AB})]}{\Delta}$
$j = 2$	$\frac{f_B/f_{AB}(1+f_B/f_{AB})}{\Delta}$	$\frac{2[f_B/f_{AB}(1+f_B/f_{AB})]}{\Delta}$	$\frac{(1/2)[3(f_A/f_{AB})^2+1+3(f_B/f_{AB})^2]}{\Delta}$

NOTE.—Δ = sum of the numerators over all cells.

such that no amount of ASP data will suffice to establish the model as 2L HET rather than 2L EPI. What distinguishes this case from what was considered in the preceding section, "Estimation of Degree of Epistasis," is that we now have three penetrance parameters (f_A, f_B, and f_{AB}) appearing in the formulae for the IBD-sharing probabilities among ASPs, rather than just the one (λ) that appeared in the previous parameterization, as shown in table 4. However, we note that only the two penetrance ratios (f_A/f_{AB}, f_B/f_{AB}) appear in table 4; f_{AB} appears only in the context of these ratios. This again illustrates, as above, that it is possible to estimate at most all but one of the constituent penetrances in the model based on ASP data.

We now prove that for any such model satisfying the FHE and producing IBD-sharing matrix π, some other model exists that produces the same π but that does not satisfy the FHE. That is, for every possible $f = (f_A, f_B, f_{AB})$ satisfying the FHE, there also exists a set of penetrances $f^* = (f_A{}^*, f_B{}^*, f_{AB}{}^*)$ such that (*i*) f^* does not satisfy the FHE; and (*ii*) f^* gives rise to an IBD-sharing table π^* that is identical to π.

To show this result, we note that the FHE requires (dividing both sides through by f_{AB})

$$1 = \frac{f_A}{f_{AB}} + \frac{f_B}{f_{AB}} - \frac{f_A}{f_{AB}} \times \frac{f_B}{f_{AB}} \times f_{AB} \; . \qquad (3)$$

We now find our second penetrance vector f^* by imposing the following two equalities:

$$f_A^*/f_{AB}^* = f_A/f_{AB} \qquad (4)$$

and

$$f_B^*/f_{AB}^* = f_B/f_{AB} \qquad (5)$$

There are infinitely many solutions to these two equalities, only some of which will satisfy the FHE. For the sake of illustration, we select one pair of solutions: $f = (f_A, f_B, f_{AB}) = (0.400, 0.200, 0.520)$ and $f^* = (f_A^*, f_B^*, f_{AB}^*) = (0.577, 0.288, 0.769)$. It is readily confirmed that the equalities in equations (4) and (5) are both satisfied. Furthermore, the vector f satisfies the FHE in the form of equation (3), but the vector f^* does not. Thus, f represents 2L HET, whereas f^* represents positive 2L EPI. Finally, plugging either f or f^* into table 4 will produce numerically identical results for the IBD-sharing probabilities, since the penetrance ratios are identical for the two models by stipulation (equations [4] and [5]).

Thus, we see that when the true model is 2L HET, then we cannot establish, on the basis of the observed IBD-sharing table, that the underlying model satisfies the FHE, no matter how many ASPs we collect. Any IBD-sharing table that is compatible with 2L HET can also be produced by a 2L EPI model. Therefore any observed data compatible with 2L HET will also be equally compatible with 2L EPI. The algebra of this example makes clear that this result can be readily generalized to other, more complicated 2L HET models. Equation [3] shows that, in general, if we are able to estimate only ratios of penetrances, then for every set of ratios satisfying the FHE, we will always be able to find 2L EPI models that generate the same IBD probability distributions. Thus, based on all the information contained in a data set of ASPs, no matter how large the sample size, we can never establish that the true model represents heterogeneity rather than epistasis.

This argument runs in one direction only: some 2L EPI models do generate IBD tables that are incompatible with 2L HET. For example, any model in which $f_A/f_{AB} + f_B/f_{AB} < 1$ can never satisfy the FHE. The simplest example of this type of model is the special case in which $f_A/f_{AB} = f_B/f_{AB} = 0$ (and $f_{AB} > 0$), or "complete" epistasis. For example, under a very rare RR completely epistatic model, in the absence of recombination, genotyping errors, or other extraneous sources of variability, the IBD sharing matrix will have a 1 for the "2,2" cell, and a 0 in all other cells. This pattern is incompatible with any 2L HET model.

We note that, although we have restricted our attention to ASPs, in fact, the results from this section extend to any studies based on analysis of affected individuals only. Such data do not provide sufficient information to estimate the absolute values of the penetrances based on marker data but only (at most) the relative values or penetrance ratios; and, as we have seen, these ratios are never sufficient to establish 2L HET.

But Can We Sometimes Establish 2L EPI?

The existence of 2L EPI models whose penetrance ratios are incompatible with 2L HET raises the possibility that we could sometimes establish 2L EPI on the basis of ASP data. To do so, however, we would need to be able to estimate the penetrance ratios. But can the penetrance ratios be estimated from ASP data in a way that makes differentiation of 2L EPI from 2L HET possible? Because the parameters of the model have complicated relationships with one another and because of the complex nature of the 2L HET and 2L EPI constraints on the parameter space, asymptotic theory does not provide an answer to this question. Therefore, we have written a computer program for numerical maximum-likelihood estimation, as described below.

To produce results relevant to real applications, we again consider the general four-penetrance, 2L model shown (for an RR model) in table 1. But we now consider the full set of eight unknown parameters, including the following: the allele frequencies p_A, p_B; the four penetrances as shown; and two recombination fractions (one corresponding to each marker).

We have programmed an algorithm for generating the IBD-sharing probabilities from any given input vector of these genetic parameters. The algorithm is similar to that of Hodge (1981); see also Li and Sacks (1954); see appendix A for details. Let π_{ij} be the probability that an ASP shares i marker alleles IBD at the A locus and j marker alleles IBD at the B locus (i,j = 0,1,2). These π_{ij}s are functions of the underlying genetic parameters of the model (disease-allele frequencies, recombination fractions, and penetrances). We use the notation π^0_{ij} to denote the IBD-sharing probabilities evaluated at the true (generating) values of the underlying parameters. Let n_{ij} be the corresponding observed number of pairs sharing i,j marker alleles IBD. Because we are interested in asymptotic properties of the likelihood, we assume perfect data (i.e., we let $n_{ij} = N \times \pi^0_{ij}$, where N is the total number of ASPs in the data set).

The support function (log likelihood), defined up to an arbitrary additive constant, can then be written as

$$S(\pi) \propto \sum_{i=0}^{2}\sum_{j=0}^{2} \pi^0_{ij} \log \pi_{ij} \ . \qquad (6)$$

To examine the asymptotic behavior of this support function, for each generating model considered, equation (6) was evaluated at each point in a grid of values for each parameter in the model, as described in appendix A. From these calculations, all sets of parameters that produced the maximum value of S(π) were found; we refer to these as the "solutions" to the maximum support function. (In every case, the generating set was among the solutions, as expected.)

The set of solutions was further characterized in terms of the number of unique solutions in the penetrance ratio pairs (f_A/f_{AB}, f_B/f_{AB}). (As for any affecteds-only data, the absolute values of all penetrances in the model cannot be estimated.) Finally, for each generating model, each solution was evaluated to determine whether it satisfied the FHE (eq. [2]). Because f_{AB} was specified to three decimal places in the generating model, we considered a solution to represent a 2L HET model if the FHE was satisfied to three decimal places.

We have considered 29 generating models, covering RR, RD, and DD penetrance tables, over a range of generating values for each parameter; see appendix A for details. In virtually every case, we find multiple solutions in (f_A / f_{AB}, f_B / f_{AB}), that is, multiple sets of penetrance ratios leading to the same maximum value of S(π). Thus, the penetrance ratios are not identifiable from ASP data. Moreover, for almost every generating model we have examined so far, every maximizing vector (f_A/f_{AB},f_B/f_{AB}) occurs in the solution set with multiple values of f_{AB}, such that some of these solutions satisfy the FHE while others do not. Thus, it appears that in real applications it is not possible to estimate sufficient penetrance structure to allow establishment of either 2L HET or 2L EPI based on ASP data, even when the true underlying penetrance ratios are incompatible with 2L HET.

The only exception to these results that we have identified so far occurs when $f_P = 0$ (no phenocopies) and when the model represents "complete" epistasis ($f_A = f_B = 0$). In this case only, the joint IBD-sharing distribution gives rise to a single solution ($f_A/f_{AB} = f_B/f_{AB} = 0$), and this solution is sufficient to establish that the model is 2L EPI rather than 2L HET. Allowing for recombination or high disease-allele frequencies has no effect on this result.

The same is not true, however, if we allow for phenocopies (e.g., $f_P = f_A = f_B = 0.05$, $f_{AB} = 1$); in which case we again obtain multiple solutions, some compatible with the FHE and others not compatible with it. Since it is hard to imagine an application in which the baseline disease probability (phenocopy rate) is zero, at least for the common complex disorders, these few exceptions seem unlikely to be relevant to data-analytic practice.

Correlations and Gene-Gene Interaction

Intuitively, it might seem that 2L EPI should produce positive correlations in the marginal IBD sharing across markers, assuming each one is linked to one of the two genes in the 2L system, and 2L HET would produce negative correlations. But, on their face, our findings show that this intuition cannot be correct. For example, we showed in the section "2L HET Can Never Be Established on the Basis of ASP Data" that for every 2L HET model there was also a 2L EPI model that would produce the same IBD-sharing probabilities. This implies directly that whatever the intermarker correlation may be under the 2L HET model, an identical correlation can be obtained under some 2L EPI model. More generally, the result in section "But Can We Sometimes Establish 2L EPI?" that observed joint (two-marker) IBD-sharing matrices cannot distinguish 2L HET from 2L EPI for any but a small group of exceptional generating models implies directly that correlations based on marginal IBD sharing distributions also cannot distinguish 2L HET from 2L EPI. (However, there are three-locus or other types of multilocus models for which correlations may provide relevant information regarding 2L HET vs. 2L EPI; see MacLean et al. 1993; Cox et al. 1999.)

There has been considerable recent interest in detecting or exploiting "gene-gene interactions" in ASP data.

For example, Holmans (2002) investigated the power to detect gene-gene interaction or epistasis (he uses the terms interchangeably, as do many investigators), which he defined as positive correlation in the marginal IBD sharing (and this too seems to be a common practice in the contemporary literature). There is absolutely nothing wrong with this from a statistical point of view, and, insofar as marginal correlations are of interest, his findings provide a useful assessment of power. However, as we have shown, when we base our definitions of 2L HET and 2L EPI on the usual genetic concepts, then epistasis turns out to be something fundamentally different from positive correlation in the marginal IBD sharing. (We note, however, that the term "gene-gene interaction," as defined in the quantitative-trait literature, necessarily differs from our definition of "2L EPI" in terms of the penetrance model, a model that is only relevant for dichotomous traits.)

This point applies as well to any statistics that are functions of the marginal IBD sharing, such as single-locus LOD scores or the various model-free statistics. And, the strength of our intuitions notwithstanding, it applies to any method based on subsetting or "conditioning" at one marker based on the magnitude of a linkage statistic at the other locus: as it turns out, 2L EPI does *not* imply that ASPs with positive linkage evidence at one locus will tend to also have positive linkage evidence at the other; and 2L HET does *not* imply that they will tend to have negative linkage evidence at the other. This finding, however, may not carry over directly to statistical methods based on correlational structure in more complex pedigrees (including unaffected individuals); and the implications for inference in the context of more complicated multilocus models are also unclear (see, e.g., MacLean et al. 1993; Cox et al. 2001).

Although these fundamental mathematical results hold regardless of the value of the correlation, it is worth noting, perhaps surprisingly, that this implies the existence of epistasis models with fairly strong negative correlation in the marginal IBD sharing among ASPs. (As has been noted elsewhere [MacLean et al. 1993], 2L epistatic models tend to predict correlations very close to 0, rather than positive correlations; and indeed almost all models we considered in the section "But Can We Sometimes Establish 2L EPI?" yielded correlations close to 0 under both 2L HET and 2L EPI generating conditions.) For example, we generated a simple RR 2L HET model (both disease allele frequencies = 0.001; both recombination fractions = 0; and penetrances $f_P = 0.00$, $f_A = 0.20$, $f_B = 0.40$, and $f_{AB} = 0.520$) that yielded a correlation of −0.417. To choose just one example of a corresponding 2L EPI model, increasing f_{AB} to 0.90, which produces an epistatic model with penetrance ratios incompatible with the FHE, yields identical IBD-sharing probabilities to four decimal places. Thus, this particular model, representing strong positive epistasis, also yields a correlation of −0.417.

Discussion

Our starting point for this work was a rigorous definition of "heterogeneity" that was intended to capture the biological meaning of the term (independent gene action at each locus) through a mathematical representation in terms of probabilistic independence and to provide a definition of "epistasis" in terms of violations of probabilistic independence. Using these new definitions, we were able to show that 2L HET cannot be distinguished from 2L EPI on the basis of ASP data. This result follows from properties of the models themselves and from properties of ASP data. It is a matter of mathematical principle, which no quantity of data can overcome.

Even in retrospect, this is perhaps a surprising result. With data on a single marker, under linkage equilibrium, ASPs provide just two pieces of information: the proportion of pairs sharing zero and one marker alleles IBD, respectively (the proportion sharing two must be such that the three numbers sum to one). But the joint two-marker IBD matrix contains nine cells (the contents of which must sum to one), so that, in general, there are eight independent pieces of information from which to work in estimating parameters. (Under some circumstances there will be fewer than eight; for instance, the number will be smaller if the penetrances and allele frequencies are set up in such a way that the resulting IBD matrix is symmetric so that there are fewer independent cells.) In view of the limited information on penetrance conveyed by ASPs, together with the limitation on the number of estimable parameters, it seems reasonable that very little specific information on the penetrance structure of the underlying model could be captured from data on a single marker. But shouldn't we expect to be able to distinguish 2L HET from 2L EPI if we can use two markers to estimate as many as eight independent parameters?

When we began this project, we predicted that it would not be possible to distinguish 2L HET from 2L EPI in ASPs on the basis of a single marker but that simultaneous consideration of two markers might ameliorate the problem. In fact, however, one marker versus two makes a quantitative difference (through the standard errors on estimates of estimable parameters) but not a qualitative one: for the general class of models considered in the "But Can We Sometimes Establish 2L EPI?" section, except for complete-epistasis models with no phenocopies, 2L HET could not be distinguished from 2L EPI either on the basis of one marker nor on the basis of simultaneous consideration of IBD sharing at both markers. Thus, the inclusion of the additional, informative (linked) marker does increase the number of estimable (IBD) parameters, but it does not improve estimation of the penetrance structure necessary to distinguish 2L HET from 2L EPI.

Similarly, it might be supposed that, given the limited amount of genetic information in a single ASP, compared with a larger multiplex family, that many more ASPs would be required to differentiate 2L HET from 2L EPI relative to the number of larger families required. But again, the limitation in the ASP structure turns out to be not only quantitative but also qualitative: the information on the penetrance structure required for distinguishing 2L HET from 2L EPI simply is not present in ASP data; therefore, no number of ASPs will be sufficient to allow the distinction to be made. The problem is inherent in the data structure itself and cannot be overcome by increasing the number of these structures in the data set. This is perhaps ironic: one of the attractions of ASPs for many investigators is that their analysis does not require parameterization in terms of penetrances; as it turns out, the limited penetrance information they convey undercuts their utility for moving from simple linkage findings to more complex questions of etiology.

These results may serve as a cautionary note for any work in statistical genetics. Depending upon the genetic question one wishes to address, ensuring the appropriateness of the statistical model and of the available data can be tricky: mathematical features of the data that intuitively seem relevant, such as intermarker correlations, may bear quite counterintuitive relationships to underlying genetic concepts; alternative family structures exhibit qualitative, not just quantitative, differences in the types of genetic information they convey, so that adding more families of a given type to a data set will not necessarily overcome problems of inadequate information; and simultaneous consideration of additional markers or genomic locations will not necessarily be useful in all contexts, even under multilocus models.

Acknowledgments

This work was supported, in part, by National Institute of Mental Health grants R01-52841 (to V.J.V.) and K01-01541 (to J.H.). The authors wish to thank Yungui Huang, for valuable programming assistance, and Mark Logue and M. Anne Spence, for helpful discussions.

Appendix A

Computational Details for Results of the Section "But Can We Sometimes Establish 2L EPI?"

All assumptions are as stated in the text. Let F be the 3 × 3 penetrance matrix. Let p_1 be the frequency of allele A, $q_1 = 1 - p_1 = \mathrm{P(a)}$; and let p_2, q_2 be the corresponding allele frequencies at the B locus. For $i = 0$, 1, and 2, let $p_{1,i}$ be the column vector of nine conditional probabilities of the genotype at the first trait locus, given that the number of shared alleles at the first trait locus is *i;* these probabilities are enumerated in table 1 of the article by Haseman and Elston (1972). Let $p_{2,i}$ be the corresponding column vector for the second trait locus. Let $\mathrm{T_A}$, $\mathrm{T_B}$ be the number of shared alleles at the A and B loci, respectively. For $i,j = 0$, 1, and 2, the conditional probability

$$p_{ij} = \mathrm{P(ASP|T_A} = i \text{ and } \mathrm{T_B} = j) = \sum\nolimits_{\mathrm{(GA,GB)}} \mathrm{P(ASP|G_A,G_B)P(G_A|T_A} = i)\mathrm{P(G_B|T_B} = j) = \mathrm{p}'_{1i}(\mathrm{F} \otimes \mathrm{F})p_{2j}\ ,$$

where $\Sigma_{\mathrm{(GA,GB)}}$ denotes the sum over all possible two-locus genotypes, and F ⊗ F is the Kronecker product of the penetrance matrix F with itself (a 9 × 9 matrix). Note that p_{ij} depends on the penetrances and the allele frequencies at the two trait loci.

Let $\mathrm{M_A}$ and $\mathrm{M_B}$ be the number of shared alleles at the two marker loci respectively. (Note that this notation differs from the notation in the main body of the text, where $\mathrm{M_A}$ and $\mathrm{M_B}$ referred to the markers themselves.) Then

$$\begin{aligned}\mathrm{P(ASP)} &= \sum\nolimits_{(i=0\ldots2)}\sum\nolimits_{(j=0\ldots2)} \mathrm{P(ASP|T_A} = i,\ \mathrm{T_B} = j)\mathrm{P(T_A} = i,\ \mathrm{T_B} = j) \\ &= \sum\nolimits_{(i=0\ldots2)}\sum\nolimits_{(j=0\ldots2)} \mathrm{p}_{ij}\ (\mathrm{T_A} = i)\ \mathrm{P(T_B} = j)\end{aligned}$$

and

$$\begin{aligned}\mathrm{P}(\mathrm{ASP},\mathrm{M_A}=i,\mathrm{M_B}=j) &= \sum\nolimits_{(\mathrm{k}=0\ldots2)}\sum\nolimits_{(\mathrm{l}=0\ldots2)}\mathrm{P}(\mathrm{ASP},\mathrm{M_A}=i,\mathrm{M_B}=j,\mathrm{T_A}=k,\mathrm{T_B}=l)\\ &= \sum\nolimits_{(k=0\ldots2)}\sum\nolimits_{(l=0\ldots2)}\mathrm{P}(\mathrm{ASP}|\mathrm{T_A}=k,\mathrm{T_B}=l)\mathrm{P}(\mathrm{M_A}=i,\mathrm{T_A}=k)\mathrm{P}(\mathrm{M_B}=j,\mathrm{T_B}=l)\\ &= \sum\nolimits_{(k=0\ldots2)}\sum\nolimits_{(l=0\ldots2)}p_{kl}\mathrm{P}(\mathrm{M_A}=i,\mathrm{T_A}=k)\mathrm{P}(\mathrm{M_B}=j,\mathrm{T_B}=l)\ .\end{aligned}$$

As in the text, we have assumed that the A and B loci are unlinked to one another. The probabilities $\mathrm{P}(\mathrm{T_A}=i)=\mathrm{P}(\mathrm{T_B}=i)$ are 0.25, 0.50, and 0.25, for $i=0$, 1, and 2, respectively. The joint probabilities $\mathrm{P}(\mathrm{M_A}=i,\ \mathrm{T_A}=k)$ and $\mathrm{P}(\mathrm{M_B}=j,\ \mathrm{T_B}=l)$ are functions of the recombination fractions between each trait locus and its corresponding marker locus; they can be computed on the basis of formulas given in table 4 of the article by Haseman and Elston (1972).

Finally, we can calculate the joint IBD-sharing probabilities at the two marker loci by use of the equation

$$\mathrm{P}(\mathrm{M_A}=i,\mathrm{M_B}=j|\mathrm{ASP})=\frac{\mathrm{P}(\mathrm{M_A}=i,\mathrm{M_B}=j,\mathrm{ASP})}{\mathrm{P}(\mathrm{ASP})}\ ,$$

where P(ASP) can also be calculated by summing the numerator over $i,j=0$, 1, and 2, respectively.

These equations were used to calculate joint two-marker–locus IBD-sharing probabilities from an input set of genetic parameters (recombination fractions, disease allele frequencies, and two-locus penetrances). We then implemented a grid-search to compute exact asymptotic ln likelihoods (support) over a discretized grid of the entire parameter space. All parameters varied from 0 to 1, with the restrictions $f_\mathrm{P}<f_\mathrm{AB},f_\mathrm{P}\leq f_\mathrm{A},f_\mathrm{P}\leq f_\mathrm{B}$. For f_AB, the range (0,1) was covered in steps of 0.001; the remaining parameters were varied at step sizes of 0.01. The support for each parameter vector was compared with the maximum support (which was always obtained under the generating values) to six decimal places of precision in most cases and in some cases (for verification purposes) to 10. A parameter vector yielding support equal to the maximum support to the specified precision was considered as satisfying the FHE if and only if it met the condition $f_\mathrm{AB}=f_\mathrm{A}+f_\mathrm{B}-f_\mathrm{A}\times f_\mathrm{B}$ to three decimal places, the generating precision on f_AB.

In all, 17 DR, 9 RR, and 3 DD models were run. Disease allele frequencies were set as both = 0.001, both = 0.10, or one at each value; recombination fractions were set as both = 0, both = 0.05 or both = 0.10, or one at each of two different values (0, 0.05, or 0.10). Both 2L HET (for verification purposes) and 2L EPI penetrance models were tried; included were both simple complete 2L EPI models ($f_\mathrm{A}=f_\mathrm{B}=f_\mathrm{P}=0,f_\mathrm{AB}>0$) and also 2L EPI models whose penetrance ratios are incompatible with the FHE (e.g., $f_\mathrm{A}=0.40$, $f_\mathrm{B}=0.20$, and $f_\mathrm{AB}=0.90$). Although these generating models are obviously in no way exhaustive of the set of all possible generating models, they are fully adequate to provide counter examples to the hypothesis that the penetrance ratios are, in general, sufficiently identifiable from ASP data to allow differentiation of 2L HET from 2L EPI.

References

Cox NJ, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI, Kong A (1999) Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. Nat Genet 21:213–215

Durner M, Greenberg DA, Hodge SE (1992) Inter- and intrafamilial heterogeneity: effective sampling strategies and comparison of analysis methods. Am J Hum Genet 51:859–870

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19

Hodge SE (1981) Some epistatic two-locus models of disease. I. Relative risks and identity-by-descent distributions in affected sib pairs. Am J Hum Genet 33:391–395

Holmans P (2002) Detecting gene-gene interactions using affected sib pair analysis with covariates. Hum Hered 53: 92–102

Li CC, Sacks L (1954) The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. Biometrics 10:347–360

MacLean CJ, Sham PC, Kendler KS (1993) Joint linkage of multiple loci for a complex disorder. Am J Hum Genet 53: 353–366

Risch N (1990) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228