Am. J. Hum. Genet. 76:152–156, 2005

Report

Mathematical Assumptions versus Biological Reality: Myths in Affected Sib Pair Linkage Analysis

Robert C. Elston, Danhong Song, and Sudha K. Iyengar

Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland

Affected sib pair (ASP) analysis has become common ever since it was shown that, under very specific assumptions, ASPs afford a powerful design for linkage analysis. In 2003, Vieland and Huang, on the basis of a "fundamental heterogeneity equation," proved that heterogeneity and epistasis are confounded in ASP linkage analysis. A much more serious limitation of ASP linkage analysis is the implicit assumption that randomly sampled sib pairs share half their alleles identical by descent at any locus, whereas a critical assumption underlying Vieland and Huang's proof is that of joint Hardy-Weinberg equilibrium proportions at two trait loci. These are considered as examples of mathematical assumptions that may not always reflect biological reality. More-robust sib-pair designs and appropriate methods for their analysis have long been available.

Blackwelder and Elston (1985) investigated the design and analysis of sibship data to detect linkage between a genetic marker and a dichotomous trait. By considering tests for samples of affected sib pairs (ASPs), they showed that, for most rare-disease models, the mean test is the most powerful. Similarly, in the case of a rare protective allele, for most models, the mean test based on unaffected sib pairs would be the most powerful. The necessary assumptions underlying the validity of such a design were clearly stated but—until very recently—largely ignored. Vieland and Huang (2003*b*) stated that "as a matter of mathematical principle" two-locus heterogeneity and two-locus epistasis cannot be distinguished on the basis of ASP marker data. In this report, we consider these cases as two examples of the same unfortunate phenomenon: the misuse of mathematical proofs by ignoring the underlying assumptions necessary for their validity. In the former case, potential users of the method were fully cautioned, whereas in the latter case a critical assumption was not even mentioned by the authors under their heading "Assumptions and Notation."

Received September 7, 2004; accepted for publication October 13, 2004; electronically published November 11, 2004.

Address for correspondence and reprints: Dr. Robert C. Elston, Department of Epidemiology and Biostatistics, Case Western Reserve University, 2103 Cornell Road, Cleveland, OH 44106-7281. E-mail: rce@darwin.case.edu

0002-9297/2005/7601-0015$15.00

Performing a linkage analysis with only ASPs is analogous to performing an epidemiological study without controls. It is now well established that unselected sib pairs tend to share more than half their alleles identical by descent, especially in an inbred population (Leutenegger et al. 2002). Zöllner et al. (2004) found that, in their sample of 148 nuclear families from the Hutterite population of South Dakota, "the signal of increased sharing is spread broadly across the genome, and no single chromosomal location reaches genomewide significance" (p. 65), with pointwise significance at the 5% level occurring on chromosomes 1, 5, 8, 10, and 15. In his classic work on the use of sib pairs to detect linkage, Penrose (1935) pointed out that "four classes will be formed according to whether the sibs are like in both characters, unlike in both, or alike in one and unlike in the other" (p. 133), with respect to the characters under investigation for linkage. It was suggested by Elston et al. (1996) that, in any ASP study, an unaffected sib should be recruited for half the ASPs, so that, for an increase of only 25% in sample size, the concordantly affected sib pairs would be properly controlled by an equal number of discordant sib pairs (DSPs). Most recently, Lemire et al. (2004) have again pointed out the need to have DSPs as controls and have devised a statistic that contrasts the two types of sib pairs to overcome the problem. For independent sib pairs, their statistic is essentially identical to that used by Elston et al. (1973). Although it was not explicitly identified as such, this

method of analysis is mathematically identical to use of the regression statistic proposed by Haseman and Elston in 1972 (HE). That method can be used for the linkage analysis of any quantitative trait; a quantitative trait subsumes a binary trait that takes on only two values, which can, without loss of generality, be taken to be 0 and 1. In the original HE method, the squared sib-pair difference (which is always zero for concordant pairs and nonzero for discordant pairs) is regressed on an estimate of the proportion of alleles the sibs share identical by descent at a marker locus and a test performed to determine whether the regression coefficient is significantly negative. This is mathematically identical to testing whether the mean proportion of marker alleles shared identical by descent is larger for concordant pairs than for DSPs, which is the method proposed by Lemire et al. (2004). As noted by Schaid et al. (2003), whether one regresses the squared sib-pair trait difference on the estimated proportion of alleles identical by descent or regresses that proportion on the squared sib-pair trait difference, the test statistic is the same. The validity of the original HE method for binary traits was again pointed out by Zeegers et al. (2003). If some sibships contain more than one unaffected sib, the method pools the concordantly affected with the concordantly unaffected sibs, comparing them with the DSPs. A recent example of this method of analysis, appropriately allowing for the dependence of sib pairs within sibships, was published by Wiesner et al. (2003). Implementation of the HE method in the computer program SIBPAL was first made generally available as part of the program package S.A.G.E. almost 20 years ago, but it is currently much extended in S.A.G.E. 5.0, in a model that allows for covariates, marker-covariate interactions, and epistasis.

Bhende et al. (1952) first demonstrated that lack of the H antigen, an intermediary point in the production of the corresponding A and B antigens at the ABO locus, results in the apparent masking of the A and B alleles. This null phenotype—discovered among the natives of Bombay, India—was called "the Bombay phenotype," to distinguish it from that of other individuals with the O blood group. Lack of the H antigen was subsequently shown to be a rare recessive trait that segregated in families (Levine et al. 1955; Aloysia et al. 1961). The advent of molecular technology has demonstrated that at least two genes—the H-gene (*FUT1*) and the Secretor gene (*FUT2*), encoding alpha (1,2) fucosyltransferases—control the complex epistatic effect (Koda et al. 1997). Mutations in the *FUT1* gene lead to lack of ABH antigen on red blood cells, whereas mutations in *FUT2* suppress the formation of ABH antigen in saliva and other body fluids. The classic Indian Bombay phenotype results from a T725G mutation of *FUT1* and the gene deletion of *FUT2*. Additional molecular variants of the classic Bombay phenotype that involve only *FUT1* or *FUT2* have been described.

Vieland and Huang (2003*b*) used a definition of epistasis that differs not only from the classical concept of epistasis, as they freely admit, but also from the modern definition of epistasis as used in quantitative genetics. In classical genetics, epistasis and hypostasis are the interlocus analogues of the intralocus concepts of dominance and recessiveness. In the case of one locus, the effect of one allele masks the effect of another allele; the masking allele is termed "dominant" and the masked allele "recessive." In the case of two (or more) loci, the two alleles at one locus mask the effect of the two alleles at another locus; the masking alleles are termed "epistatic" and the masked ones "hypostatic." In our example, alleles at the ABO locus are hypostatic to alleles at *FUT1* and *FUT2,* whereas the latter are epistatic to those at the ABO locus. Farral (2003) and Cordell (2003) have discussed in detail that, under the definitions of epistasis used in quantitative genetics, it is not true that heterogeneity and epistasis cannot be distinguished on the basis of ASP marker data.

In the appendix, we show that, specifically for ASPs and their definition of epistasis, Vieland and Huang's mathematical derivation results from the assumption of joint Hardy-Weinberg equilibrium proportions for two trait loci—which we believe is an unrealistic mathematical assumption—rather than from any mathematical principle. We do this using the same notation, and making all the same assumptions, as in the section of their paper titled "Assumptions and Notation."

Although the assumption of joint Hardy-Weinberg equilibrium proportions—just like the assumption that unselected sib pairs, in the absence of inbreeding, share half their alleles identical by descent—may be a reasonable assumption at the time of conception, we believe it to be an unreasonably arbitrary assumption for any later time point. Most conceptions are never born (Croteau et al. 2002; Edwards 2003), so, by as early as birth, there is ample opportunity for selection or meiotic drive to disrupt both chance allele sharing at a single locus and joint equilibrium proportions at two loci. Selection may disrupt linkage equilibrium between a marker and a trait locus, or even between two marker loci—but we believe that disruption of joint equilibrium proportions for two trait loci is even more likely, especially if the trait is one for which there may be any form of selection, including meiotic drive.

In conclusion, the necessary assumption that randomly sampled sib pairs always share half their alleles identical by descent is a serious limitation of ASP linkage analysis, and we believe that epistasis is best defined as a statistical concept of dependent gene action rather than as a departure from a "fundamental heterogeneity equation" based on mathematical assumptions that do not

follow from well-established biological facts. However, it is clear that any ASP design that does not include some DSPs may be unable to validly detect epistasis, regardless of the validity of the proof offered by Vieland and Huang (2003*b*). A limitation of using an ASP design that is much more serious than the possible confounding of heterogeneity and epistasis (which is more a question of assumptions and definitions than biological reality) is that linkage results based on ASPs alone can easily lead to false-positive results. Similarly, any test that purports to gain power by pooling the "information" available from ASPs alone with that available from comparing ASPs and DSPs (Forrest and Feingold 2000) should be used with caution.

Acknowledgments

This work was supported in part by grants from the U.S. Public Health Service: research grant GM-28356, from the National Institute of General Medical Sciences; research grants DK-57292 and DK-54644, from the National Institute of Diabetes, Digestive and Kidney Diseases; and resource grant RR03655, from the National Center for Research Resources.

Appendix

We show here that, without the critical assumption of joint Hardy-Weinberg equilibrium proportions, equation (1) below does not necessarily follow from a probability truism that Vieland and Huang (2003*b*) used to capture the biological meaning of independent gene action at each of two loci.

Assume two trait loci, each with two alleles (*A* and *a*; *B* and *b*). The allele frequencies are $p_A = P(A)$, $q_A = 1 - p_A = P(a)$, and similarly for p_B, q_B. Assume that the trait loci are unlinked to each other, that each is linked to a marker, and that the two markers are unlinked to each other. There is linkage equilibrium between each marker and the trait, as well as between the two markers. We take as our example a double recessive trait phenotype with no phenocopies. The essence of Vieland and Huang's mathematical derivation then hinges on two further assumptions:

1. The two trait loci have a genotypic distribution that follows joint Hardy-Weinberg equilibrium proportions, with the result that

$$f_{AB} = f_A + f_B - (f_A \times f_B) \ , \tag{1}$$

where f_{AB} is the penetrance of the double homozygote *aabb*, f_A is the penetrance of *aa* whether the genotype at the *B* locus is *BB* or *Bb*, and f_B is the penetrance of *bb* whether the genotype at the *A* locus is *AA* or *Aa*.

2. Two-locus epistasis is defined to be any relationship among the penetrances that does not satisfy equation (1), "on the grounds that either the genes act independently or not" (Vieland and Huang 2003*a,* p. 1471).

Let *K* be the total prevalence of a disease, K_A the prevalence due to the action of genotypes at the *A* locus alone, and K_B the prevalence due to genotypes at the *B* locus alone. It then follows, from elementary probability, that

$$K = K_A + K_B - (K_A \times K_B) \ . \tag{2}$$

We now show that, in the absence of phenocopies, for a double recessive trait phenotype, equation (1) only follows from equation (2) on the assumption of joint Hardy-Weinberg equilibrium proportions. (An analogous argument can be made for a double dominant or a dominant-recessive trait phenotype.)

Our specific claim is that, if $f_A > 0$, $f_B > 0$, and $P(aa,bb) \neq q_A^2 \times q_B^2$, where $P(aa,bb)$ is the joint genotypic frequency of the two-locus genotype *aabb,* then, for a double recessive trait phenotype with no phenocopies, equation (1) does not follow from equation (2). We prove this by contradiction: Let $P(aa,bb) = q_A^2 \times q_B^2 + \varepsilon$, $\varepsilon \neq 0$. In the case of a double recessive trait phenotype, when there are no phenocopies and we assume Hardy-Weinberg equilibrium

at each locus, K_A is defined to be $q_A^2 \times f_A$, and K_B is defined to be $q_B^2 \times f_B$. Therefore, we have the following result:

$$\begin{aligned} K &= [P(aa,BB) + P(aa,Bb)] \times f_A + [P(AA,bb) + P(Aa,bb)] \times f_B + P(aa,bb) \times f_{AB} \\ &= [P(aa) - P(aa,bb)] \times f_A + [P(bb) - P(aa,bb)] \times f_B + P(aa,bb) \times f_{AB} \\ &= P(aa) \times f_A + P(bb) \times f_B - P(aa,bb) \times (f_A + f_B - f_{AB}) \\ &= q_A^2 \times f_A + q_B^2 \times f_B - P(aa,bb) \times (f_A + f_B - f_{AB}) \\ &= K_A + K_B - P(aa,bb) \times (f_A + f_B - f_{AB}) \ . \end{aligned}$$

This, along with the assumption that equation (2) is true, implies that

$$P(aa,bb)(f_A + f_B - f_{AB}) = K_A \times K_B$$

or, equivalently, that

$$(q_A^2 \times q_B^2 + \varepsilon)(f_A + f_B - f_{AB}) = q_A^2 \times q_B^2 \times f_A \times f_B \ . \tag{3}$$

But, if equation (1) follows from equation (2), then equation (1) is true—that is,

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

and so equation (3) becomes

$$(q_A^2 \times q_B^2 + \varepsilon)(f_A \times f_B) = q_A^2 \times q_B^2 \times f_A \times f_B \ .$$

This leads to

$$\varepsilon \times f_A \times f_B = 0 \ ,$$

which is a contradiction to $\varepsilon \neq 0$, $f_A > 0$, and $f_B > 0$.

Electronic-Database Information

The URL for data presented herein is as follows:

S.A.G.E. 5.0, http://darwin.cwru.edu/sage/

References

Aloysia M, Gelb AG, Fudenberg H, Hamper J, Tippett P, Race RR (1961) The expected "Bombay" group O(H-A1) and O(H-A2). Transfusion 1:212–217

Bhende YM, Deshpande CK, Bhatia HM, Sanger R, Race RR, Morgan WTJ, Watkins WM (1952) A "new" blood group character related to the ABO system. Lancet 1:903–904

Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2:85–97

Cordell HJ (2003) Affected-sib-pair data can be used to distinguish two-locus heterogeneity from two-locus epistasis. Am J Hum Genet 73:1468–1471

Croteau S, Andrade MF, Huang F, Greenwood CM, Morgan K, Naumova AK (2002) Inheritance patterns of maternal alleles in imprinted regions of the mouse genome at different stages of development. Mamm Genome 13:24–29

Edwards JH (2003) Sib-pairs in multifactorial disorders: the sib-similarity problem. Clin Genet 63:1–9

Elston RC, Guo X, Williams LV (1996) Two-stage global search designs for linkage analysis using pairs of affected relatives. Genet Epidemiol 13:535–558

Elston RC, Kringlen E, Namboodiri KK (1973) Possible linkage relationships between certain blood groups and schizophrenia or other psychoses. Behav Genet 3:101–106

Farrall M (2003) Reports of the death of the epistasis model are greatly exaggerated. Am J Hum Genet 73:1467–1468

Forrest WF, Feingold E (2000) Composite statistics for QTL mapping with moderately discordant sibling pairs. Am J Hum Genet 66:1642–1660

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

Koda Y, Soejima M, Johnson PH, Smart E, Kimura H (1997) Missense mutation of *FUT1* and deletion of *FUT2* are re-

sponsible for Indian Bombay phenotype of ABO blood group system. Biochem Biophys Res Commun 238:21–25

Lemire M, Roslin NM, Laprise C, Hudson TJ, Morgan K (2004) Transmission-ratio distortion and allele sharing in affected sib pairs: a new linkage statistic with reduced bias, with application to chromosome 6q25.3. Am J Hum Genet 75:571–586

Leutenegger AL, Genin E, Thompson EA, Clerget-Darpoux F (2002) Impact of parental relationships in maximum lod score affected sib-pair method. Genet Epidemiol 23:413–425

Levine P, Robinson E, Celano M, Briggs O, Falkinburg L (1955) Gene interaction resulting in suppression of blood group substance B. Blood 10:1100–1108

Penrose LS (1935) The detection of autosomal linkage in data which consists of pairs of brothers and sisters of unspecified parentage. Ann Eugen 6:133–138

Schaid DJ, Olson JM, Gauderman WJ, Elston RC (2003) Regression models for linkage: issues of traits, covariates, heterogeneity, and interaction. Hum Hered 55:86–96

Vieland VJ, Huang J (2003*a*) Reply to Cordell and Farrall. Am J Hum Genet 73:1471–1473

——— (2003*b*) Two-locus heterogeneity cannot be distinguished from two-locus epistasis on the basis of affected-sib-pair data. Am J Hum Genet 73:223–232

Wiesner GL, Daley D, Lewis S, Ticknor C, Platzer P, Lutterbaugh J, MacMillen M, Baliner B, Willis J, Elston RC, Markowitz SD (2003) A subset of familial colorectal neoplasia kindreds linked to chromosome 9q22.2-31.2. Proc Natl Acad Sci USA 22:12961–12965

Zeegers MPA, Rice JP, Rijsdijk FV, Abecasis GR, Sham PC (2003) Regression-based sib pair linkage analysis for binary traits. Hum Hered 55:125–131

Zöllner S, Wen X, Hanchard NA, Herbert MA, Ober C, Pritchard JK (2004) Evidence for extensive transmission distortion in the human genome. Am J Hum Genet 74:62–72