Letters to the Editor

Am. J. Hum. Genet. 75:716–718, 2004

No "Bias" Toward the Null Hypothesis in Most Conventional Multipoint Nonparametric Linkage Analyses

To the Editor:

We would like to comment on the Schork and Greenwood (2004) article dealing with the inherent "bias" toward the null hypothesis in the context of nonparametric linkage analysis. The authors point out that, in certain situations, a loss of evidence for linkage can result from the practice of assigning expected allele-sharing values to affected relative pairs that are uninformative for their identity-by-descent (IBD) status. They explained this by setting up a likelihood function and studying its properties by simulation, clearly illustrating the negative impact of using expected IBD values for uninformative pairs. However, we would like to point out that their likelihood does not reflect how the majority of nonparametric linkage analysis programs compute statistics in practice. Indeed, the "problem" has been known and well discussed for years. Some of the concerns we discuss here have also been raised by Cordell (2004).

Schork and Greenwood (2004) set up the likelihood formulation as follows. Let n_i be the number of sib pairs sharing i alleles IBD ($i = 0$, 1, or 2). If all families had unambiguous IBD sharing, then the LOD score evaluated at the sharing vector (p_0, p_1, p_2) is calculated as

$$\begin{aligned}\text{LOD} &= \log_{10}\left\{\frac{p_0^{n_0}\,p_1^{n_1}\,p_2^{n_2}}{0.25^{n_0}\,0.50^{n_1}\,0.25^{n_2}}\right\}\\ &= n_0\log_{10}(4p_0)+n_1\log_{10}(2p_1)+n_2\log_{10}(4p_2)\ . \quad (1)\end{aligned}$$

In their model, Schork and Greenwood (2004) said that fully uninformative sibling pairs contribute 0.25, 0.50, and 0.25, respectively, to the counts n_0, n_1, and n_2 used in equation (1). If so, then the presence of uninformative sib pairs can lower the LOD score. However, in most software implementations, expected allele-sharing values are *not* used to compute nonparametric LOD scores. For example, consider the maximum LOD score (MLS) statistic proposed by Risch (1990). Let w_i be the probability of the observed marker phenotypes of the pair, given that they share i alleles IBD ($i = 0$, 1, or 2). Then, the likelihood of the observed marker data for the pair is given by

$$L = \sum_{i=0}^{2} w_i\,p_i\ ,$$

where p_i is the posterior probability that the pair shares i alleles IBD, given that both members of the pair are affected. Suppose, in addition, that we know that $n_{2,1}$ pairs share either 2 or 1 alleles, $n_{2,0}$ pairs share either 2 or 0 alleles, $n_{1,0}$ pairs share either 1 or 0 alleles, and n_{un} is the number of pairs that are fully uninformative. According to Risch (1990), the LOD score can be written as

$$\begin{aligned}\text{LOD} = {} & n_0\log_{10}(4p_0)+n_1\log_{10}(2p_1)+n_2\log_{10}(4p_2)\\ & +n_{2,1}\log_{10}(2p_2+p_1)+n_{2,0}\log_{10}[2(p_2+p_0)]\\ & +n_{1,0}\log_{10}(p_1+2p_0)+n_{\text{un}}\log_{10}(p_0+p_1+p_2)\ .\end{aligned}$$

Maximizing this likelihood gives consistent and asymptotically unbiased estimates of the IBD-sharing probabilities. Cordell (2004) confirms this by simulation.

To verify that most implementations of nonparametric linkage statistics are not altered by uninformative families, we used FastSLINK (Ott 1989; Weeks et al. 1990; Cottingham et al. 1993) to simulate 200 fully genotyped affected–sib-pair families under disease model 1 of Schork and Greenwood (2004). The disease locus was completely linked to a two-allele marker with equally frequent alleles. We then used a variety of programs to compute linkage statistics on two data sets: (1) all 200 families and (2) the 147 families that remained after removal of the fully uninformative families. As shown in table 1, the majority of the linkage statistics, as implemented in widely used software, are exactly the same for the two data sets.

There are two statistics in table 1 that are less significant when all 200 families are used than when the uninformative families are removed. These two statistics are the GeneHunter NPL S_{all} Z score and the SIBPAL mean test Z value. In both of these cases, the reduction in evidence for linkage is caused by the use of the "perfect data approximation" to compute the variance of the

Table 1

Comparison of Linkage Statistics Analyses Using All 200 Families and Using Only the 147 Informative Families

Statistic and Software	Result for All 200 Families	Result for 147 Informative Families	Reference
Mean test Z value:			
SIBPAL	14.07	17.56	Haseman and Elston 1972
MLS LOD score (2 df):			
SPLINK	36.34	36.34	Holmans 1993
MLS LOD score (1 df):			
GeneHunter	22.20	22.20	Kruglyak and Lander 1995
ASPEX sib_phase	22.20	22.20	Hinds and Risch 1996
NPL S_{all} Z score:			
GeneHunter	6.70	7.82	Kruglyak et al. 1996
Allegro	7.82	7.82	Gudbjartsson et al. 2000
Merlin	7.82	7.82	Abecasis et al. 2002
GeneHunter-Plus S_{all} LOD score:			
GeneHunter-Plus	22.20	22.20	Kong and Cox 1997
Allegro	22.20	22.20	Gudbjartsson et al. 2000
Merlin	22.20	22.20	Abecasis et al. 2002

statistics. The "perfect data approximation" performs well if most of the families are informative for IBD sharing, but, as the proportion of uninformative families increases, it becomes increasingly conservative, leading to a loss of power (Kruglyak et al. 1996). In fact, the loss of power due to "bias" that Schork and Greenwood (2004) identify is, mathematically, exactly the same thing as the loss of power due to the "perfect data approximation."

The negative effects of the "perfect data approximation" can be illustrated by a simple example. Consider the sib-pair IBD-sharing statistic

$$\frac{\sum_i (\pi_i - 1/2)}{\sqrt{\mathrm{var}\left(\sum_i \pi_i\right)}},$$

where π_i is the estimated proportion of alleles shared IBD for the ith affected sib pair. Suppose we have two data sets: (1) 50 fully informative affected–sib-pair families and (2) 50 fully informative and 50 uninformative families. Suppose π_i in our fully informative families takes on the values 0, 1/2, and 1, with probabilities 1/8, 1/2, and 3/8, respectively, whereas π_i is 1/2 in our uninformative families. The numerator of the statistic is identical for both data sets. However, different approaches to computing the variance in the denominator can lead to different statistic values for the two data sets. Under the "perfect data approximation," the value of the statistic is 2.50 for the first data set and 1.77 for the second data set—an undesirable reduction in the evidence for linkage. Use of the correct variance (given that the number of uninformative families remains constant) leads to statistic values of 2.50 for both data sets. Another option is to use the empirical variance, which reflects the alternative hypothesis rather than the null hypothesis and can be quite powerful; the empirical variance gives an expected IBD-sharing statistic of 2.50 for both example data sets. A score test using empirical variances was one of the best statistics in a recent evaluation of methods for QTL mapping using selected sibling pairs (T.Cuenco et al. 2003).

To avoid the negative consequences of using the "perfect data approximation," Kong and Cox (1997) proposed a nonparametric statistic that performs much better in the presence of uninformative families. This statistic has been implemented in GeneHunter-Plus (Kong and Cox 1997), Allegro (Gudbjartsson et al. 2000), and Merlin (Abecasis et al. 2002) and, as illustrated by our simple simulation experiment in table 1, is insensitive to the presence of fully uninformative families. Similarly, in the context of the Haseman-Elston (HE) test (Haseman and Elston 1972), in which trait values are regressed on IBD sharing, the problem of using estimated IBD sharing has long been recognized. For example, Kruglyak and Lander (1995) developed a missing-value regression approach to compute a modified HE test that has much better behavior in the presence of uninformative families than the original test.

Whereas it is always useful to remind the scientific community that proper statistical analyses of linkage data requires deep insight into the potential weaknesses of the chosen methodology and software implementation, we feel that Schork and Greenwood's concerns are overstated. Indeed, as we have shown, not only has this potential problem been known since at least the mid-

1990s, but, in addition, the majority of implementations of linkage statistics in commonly used software do not suffer from this "bias" toward the null hypothesis in the presence of uninformative families. Furthermore, the use of highly informative markers in a multipoint analysis will result in very few families being fully uninformative for IBD sharing.

Acknowledgments

This work was supported by the University of Pittsburgh and National Institutes of Health grants 5D43TW006180-02 and 5R01MH064205-06. Some of the results of this paper were obtained using the S.A.G.E. package of genetic epidemiology software, which is supported by U.S. Public Health Service Resource grant RR03655 from the National Center for Research Resources.

INDRANIL MUKHOPADHYAY,[1] ELEANOR FEINGOLD,[1,2] AND DANIEL E. WEEKS[1,2]

Departments of [1]Human Genetics and [2]Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101

Cordell HJ (2004) Bias toward the null hypothesis in model-free linkage analysis is highly dependent on the test statistic used. Am J Hum Genet 74:1294–1302

Cottingham RW, Idury RM, Schäffer AA (1993) Faster sequential genetic linkage computations. Am J Hum Genet 53: 252–263

Gudbjartsson DF, Jonasson K, Frigge ML, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. Nat Genet 25:12–13

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

Hinds D, Risch N (1996) The ASPEX package: affected sib-pair exclusion mapping. Available at: http://aspex.sourceforge.net/. Accessed August 2, 2004

Holmans P (1993) Asymptotic properties of affected–sib-pair linkage analysis. Am J Hum Genet 52:362–374

Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–1188

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

Ott J (1989) Computer-simulation methods in human linkage analysis. Proc Natl Acad Sci USA 86:4175–4178

Risch N (1990) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242–253

Schork NJ, Greenwood TA (2004) Inherent bias toward the null hypothesis in conventional multipoint nonparametric linkage analysis. Am J Hum Genet 74:306–316

T.Cuenco K, Szatkiewicz JP, Feingold E (2003) Recent advances in human quantitative-trait–locus mapping: comparison of methods for selected sibling pairs. Am J Hum Genet 73:863–873

Weeks DE, Ott J, Lathrop GM (1990) SLINK: a general simulation program for linkage analysis. Am J Hum Genet 47: A204

Address for correspondence and reprints: Dr. Daniel E. Weeks, Department of Human Genetics, University of Pittsburgh, Crabtree Hall, Room A302A, 130 DeSoto Street, Pittsburgh, PA 15261. E-mail: dweeks@watson.hgen.pitt.edu

0002-9297/2004/7504-0020$15.00

Am. J. Hum. Genet. 75:718–720, 2004

Conventional Multipoint Nonparametric Linkage Analysis Is Not Necessarily Inherently Biased

To the Editor:

Schork and Greenwood (2004) recently reported that there is an inherent bias toward the null hypothesis in conventional multipoint linkage analysis in which expected values are used for allele sharing between relatives when, in fact, there is no information on their identity-by-descent (IBD) sharing status. The implications of Schork and Greenwood's results are serious, because they suggest that the power of detection of disease genes or QTLs is compromised. Here, we show that their results are based on a comparison of test statistics that have different variance (and, therefore, have different distribution) and so should not be compared directly and that the usual way in which inference is made from multipoint nonparametric linkage is, in fact, correct. In addition, we demonstrate that, for linkage analysis of quantitative traits, the effect of mixing informative and uninformative sib pairs on the test statistic is very small and very unlikely to be of practical importance.

Schork and Greenwood (2004) use the analogy of a coin-tossing experiment to make their main point, and we use the same experiment to contest their conclusion. Suppose a coin is tossed 100 times to test the hypothesis that it is fair (i.e., that it gives a 1:1 ratio of heads to tails). The outcome of the experiment is observed in only 50 tosses, and, of those 50 tosses, 40 are heads. The estimate of the probability of heads ($\hat{p}$) from the observation that 40 of 50 observed tosses are heads is thus 0.80. If we assign the expected values (under the null