Table 1

Results from Simulations to Show the Relative Impact of Uninformative Pairs on QTL Analysis for Different Values of Polygenic and QTL Heritability and Different Percentages of Uninformative Markers

Heritability		Uninformative Markers		HE-LOD[b]			VC-LOD[b]		
Polygenic	QTL	(%)	No. of Pairs	$\delta T/E(T)$	$\sigma(\delta)/E(T)$	$\sigma(\delta)/\sigma(T)$	$\delta T/E(T)$	$\sigma(\delta)/E(T)$	$\sigma(\delta)/\sigma(T)$
0	.90	50	1,000[a]	4	7	6	0	1	2
0	.98	50	1,000[a]	4	11	9	0	2	2
.5	.10	10	10,000	0	2	1	0	1	0
.5	.10	20	10,000	0	2	1	0	1	0
.5	.10	30	10,000	0	3	1	0	1	0
.5	.10	40	10,000	0	4	1	0	1	0
.5	.10	50	10,000	0	5	1	0	2	0
.5	.10	60	10,000	0	6	2	0	2	0
.5	.10	70	10,000	0	8	2	0	2	0
.5	.10	80	10,000	0	11	2	0	4	1
.5	.10	90	10,000	0	16	3	0	8	1

[a] The data in the first two rows correspond to the simulated scenario (2) of Cordell (2004) and are based on 10,000 replicates. For all other data, a normally distributed additive QTL was simulated, and results are averages from 1,000 replicates.

[b] δT is the average difference between the test statistic achieved when uninformative pairs are removed from the analysis and the one achieved when they are kept in the analysis. $\sigma(\delta)$ is the SD of the difference between the test statistic achieved when uninformative pairs are removed from the analysis and the one achieved when they are kept in the analysis. $E(T)$ is the average test statistic achieved when uninformative markers are removed from the analysis. $\sigma(T)$ is the SD of the test statistic when uninformative markers are removed from the analysis. All ratios are expressed as percentages.

[2004]) is very small because, presumably, the phenotypes of the uninformative pairs provide information on the estimation of the sibling variance and average covariance, and this information is used in the maximum-likelihood analysis. Hence, removal of uninformative pairs may indirectly decrease information on linkage.

We conclude that commonly used nonparametric allele-sharing methods, as implemented in major statistical-genetics computer programs, do not suffer from an inherent bias toward the null hypothesis when expected values of IBD sharing are used in the absence of observed IBD sharing and that QTL-mapping methods are not invariant but are robust to mixtures of informative and uninformative pairs.

Acknowledgments

The authors are supported by the U.K. Biotechnology and Biological Sciences Research Council, the U.K. Medical Research Council, and Organon NV. We thank Heather Cordell and Eleanor Feingold for helpful discussions.

Peter M. Visscher and Naomi R. Wray
School of Biological Sciences and Molecular Medicine Centre, University of Edinburgh, Edinburgh

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

Gessler DG, Xu S (1996) Using the expectation or the distribution of the identity by descent for mapping quantitative trait loci under the random model. Am J Hum Genet 59: 1382–1390

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

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

Address for correspondence and reprints: Dr. P. M. Visscher, University of Edinburgh, School of Biological Sciences, Ashworth Laboratories, West Mains Road, Edinburgh EH9 3JT, United Kingdom. E-mail: peter.visscher@ed.ac.uk

0002-9297/2004/7504-0021$15.00

Am. J. Hum. Genet. 75:720–722, 2004

"Bias toward the Null" Means Reduced Power

To the Editor:

In a recent article published in the *Journal*, Schork and Greenwood (2004) discuss the effects of uncertainty in inferred identity-by-decent (IBD) sharing on nonparametric linkage analysis. Tests based on inferred IBD

sharing that do not account for the ambiguity in data have long been known to have type I error smaller than the nominal level of the test (Kruglyak et al. 1996; Ekstrom 2001). This property, which is classically called *conservativeness* and leads to *loss of power,* is what the authors refer to as "bias toward the null hypothesis." The term "bias" is misleading and, in fact, incorrect here. For affected sib pairs, which are the focus of this letter, let **X** denote the data, at any given locus, and let R be the rejection region for testing H_0:$\{p_0,p_1,p_2\} = \{0.25,0.5,0.25\}$ versus H_A:$\{p_0,p_1,p_2\} \neq \{0.25,0.5,0.25\}$, which is the hypothesis test formulated by Schork and Greenwood (2004). A *biased* test would be one in which

$$P_{\{p_0,p_1,p_2\}}(\mathbf{X} \in R) < P_{\{0.25,0.5,0.25\}}(\mathbf{X} \in R)$$

for some $\{p_0,p_1,p_2\}$. In other words, a biased test is one in which there exists a set of parameters for which the probability of rejecting H_0 is smaller than the true level of the test (Casella and Berger 1990). This is clearly not the case here. In this case, a misspecification of the variance, arising from uncertainty in the IBD sharing, causes the test to be conservative, but it is still unbiased.

The conservativeness of allele-sharing tests has been addressed elsewhere. Teng and Siegmund (1998) used a score statistic and computed the appropriate critical value to attain the correct level. Kong and Cox (1997) used a likelihood model for the missing information. This likelihood model has been shown to result in tests that have the appropriate type I error rate (Badner et al. 1998) and is implemented in several common multipoint linkage packages, including MERLIN (Abecasis et al. 2002) and ALLEGRO (Gudbjartsson et al. 2000).

Schork and Greenwood (2004) propose five "methods" to deal with the conservativeness of traditional tests. Two of these "methods" are not methods so much as general truths about how to improve linkage analysis. These practices—namely, exploring measures of information to identify regions of low informativeness and increasing marker density in these regions to increase the multipoint information—should be done routinely. Two of the remaining methods involve weighting families according to the informativeness of the genotypes. We feel these methods are extremely dangerous, because of the asymmetrical nature of information about sharing. Consider the case of sib pairs with ungenotyped parents. When the pair have no common alleles, the IBD state of sharing no alleles can be inferred with complete certainty; however, when the pair share alleles in common, uncertainty about the IBD state always exists. The common alleles could be shared by descent or in state only. Thus, down-weighting (or removing completely) pairs with low information will systematically remove pairs with shared alleles and will result in a conservative test for linkage and an anticonservative test for exclusion mapping. The last method proposed by Schork and Greenwood (2004) involves using mixture models, but the authors themselves admit that this will work only in special cases and then will only partially alleviate the problem of conservativeness.

A further issue deserving mention is the actual calculation of IBD probabilities, conditional on multipoint marker genotype data. Under the assumption of no crossover interference and with reasonable estimates of marker allele frequencies and the genetic map, IBD probabilities are *calculated* rather than *estimated*. Some computer programs, such as MERLIN (Abecasis et al. 2002), ALLEGRO (Gudbjartsson et al. 2000), and GENEHUNTER (Kruglyak et al. 1996), compute these probabilities exactly; others, such as SOLAR (Almasy and Blangero 1998), use approximations that can give poor results (see Sobel et al. 2001). A discrepancy between the results of two such methods should not raise suspicion in both, since only one may be wrong.

Unfortunately, the authors' incomplete knowledge of the relevant literature may lead readers of Schork and Greenwood (2004) to believe that investigators' inability to identify genes contributing to complex diseases is the result of inadequacies in the statistical methods. However, these perceived inadequacies have been largely overcome through methods clearly superior to those proposed in the article. There can be little doubt that more and better data (e.g., improved phenotyping, additional families, and more complete genotype data) will provide improved results; however, the key challenge in identifying genes for complex diseases lies in the complex nature of the diseases.

Solveig K. Sieberts,[1] Karl W. Broman,[2] and Daniel F. Gudbjartsson[1]
[1]*deCODE Genetics, Reykjavík; and* [2]*Department of Biostatistics, Johns Hopkins University, Baltimore*

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

Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1211

Badner JA, Gershon ES, Goldin LR (1998) Optimal ascertainment strategies to detect linkage to common disease alleles. Am J Hum Genet 63:880–888

Casella G, Berger R (1990) Statistical inference. Duxbury Press, Belmont, CA

Ekstrom C (2001) Power of multipoint identity-by-descent methods to detect linkage using variance component models. Genet Epidemiol 21:285–298

Gudbjartsson D, Jonasson K, Frigge M, A K (2000) Allegro,

a new computer program for multipoint linkage analysis. Nat Genet 25:12–13
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
Schork NJ, Greenwood TA (2004) Inherent bias toward the null hypothesis in conventional multipont nonparametric linkage analysis. Am J Hum Genet 74:306–316
Sobel E, Sengul H, Weeks DE (2001) Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. Hum Hered 52: 121–131
Teng J, Siegmund D (1998). Multipoint linkage analysis using affected relative pairs and partially informative markers. Biometrics 54:1247–1265

Address for correspondence and reprints: Dr. Solveig K. Sieberts, deCODE Genetics, Sturlugata 8, 101 Reykjavik, Iceland. E-mail: solveig.sieberts@decode.is

0002-9297/2004/7504-0022$15.00

Am. J. Hum. Genet. 75:722–723, 2004

No Bias in Linkage Analysis

To the Editor:

In a recent article, Schork and Greenwood (2004) made the alarming claim that nonparametric linkage analysis methods have a previously unrecognized inherent bias against detection of linkage and proposed that linkage studies that have used these methods should be reexamined. It is fortunate for the genetics community that this claim is not well founded. The "bias" discussed by Schork and Greenwood is simply conservative handling of incomplete information. This issue is well appreciated by statistical geneticists, and most nonparametric linkage analysis methods—as implemented in commonly used programs such as GeneHunter (Kruglyak et al. 1996), Merlin (Abecasis et al. 2002), and many other software packages—already handle incomplete information correctly (see Cordell [2004]). The examples to the contrary provided by Schork and Greenwood (2004) derive from a contrived statistic explicitly implemented by these authors to handle incomplete information incorrectly.

This is best illustrated with Schork and Greenwood's (2004) example of testing whether a coin is fair. They write that if a coin is tossed 100 times, but the outcomes of only 50 tosses are observed, and 40 of these come up heads, then the estimate of the probability of heads is, of course, 0.80. They then write that if the 50 unobserved losses are assigned a 25-25 split expected of a fair coin, then the overall estimate of the probability of heads would be 0.65, which underestimates the true probability of heads and leads to a bias against detection of an unfair coin. This is, of course, true, and, for that very reason, no sound statistical procedure assigns a 25-25 split to the unobserved events. Rather, all correct missing-data–estimation procedures appropriately compute the probability of heads to be 0.80 in this example. Schork and Greenwood's statistic, unlike real-world linkage statistics, implements the equivalent of the former (incorrect) procedure when faced with incomplete data (i.e., uninformative markers or evaluation of linkage between marker locations).

The method directly examined by Schork and Greenwood (2004) is based on the popular maximum LOD score (MLS) approach introduced by Risch (1990). In this approach, the fraction of alleles that are shared identical by descent (IBD) by affected pairs of relatives (the quantity represented by the probability of heads in the coin-toss analogy) is estimated by maximum likelihood, and significance is evaluated via a likelihood-ratio test. The expectation-maximization (EM) algorithm (Dempster et al. 1977) is most commonly used to account for incomplete specification of IBD sharing by the data. The EM algorithm, as originally described (Dempster et al. 1977) and when correctly implemented (e.g., by Kruglyak and Lander [1995]), computes the IBD-sharing estimates iteratively, using standard missing-data techniques to update the "imputed values" at each iteration, and provides an accurate and unbiased estimate of the fraction of alleles shared IBD (and the LOD score) at the final iteration (see Cordell [2004]).

The statistic used by Schork and Greenwood (2004) is superficially similar, but, unlike any statistical analysis in the widely used linkage-analysis programs, does not use EM but rather simply assigns to uninformative pairs the sharing fraction expected under the null hypothesis of no linkage, making no attempt to properly estimate the sharing for uninformative data under the alternative hypothesis of linkage. Although the authors do not describe in detail how they implemented the method, their equation (1) (as well as their definition of maximum-likelihood estimates for the IBD-sharing parameters) applies only to the case of fully informative pairs and is inappropriate for other cases. The appropriate formulation is clearly stated in the article by Risch (1990) that originally described the method, as well as in Kruglyak and Lander (1995).

It is important to note that, although we have focused on the case of the MLS approach and the EM algorithm, appropriate handling of incomplete information has been a key consideration in the design and implementation of other nonparametric linkage methods. For example, the problem of incomplete information in quan-