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

© 2004 by The American Society of Human Genetics. All rights reserved.
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-

titative-trait analysis was explicitly addressed nearly a decade ago for sib pairs (Kruglyak and Lander 1995) and, more recently, for larger pedigrees (e.g., Sham et al. 2002), although several methods still in use today have not fully accounted for this issue, and users should be cognizant of this fact (Cordell 2004). Also, although nonparametric linkage (NPL) analysis has always been recognized to be conservative when the data is not fully informative (Kruglyak et al. 1996), this problem has long been resolved either by calculating LOD scores (Kong and Cox 1997) or by estimating significance empirically through simulation (e.g., Kruglyak and Daly 1998), an approach that is becoming increasingly practical even for whole-genome scans. Other methods are examined in detail by Cordell (2004), who comes to similar conclusions. Of course, it is well appreciated that all linkage methods (and all statistical tests, in general) have lower power when faced with less informative data, but this broadly recognized effect is distinct from the "bias" claimed by Schork and Greenwood.

Schork and Greenwood (2004) also make a problematic statement about parametric linkage analysis. They correctly note that the contribution to the LOD score of completely uninformative families is zero—exactly the same as when such families are simply excluded from analysis—but then inexplicably conclude that "uninformative families detract from a linkage signal in parametric settings as well" (Schork and Greenwood 2004, p. 312). Since the final statistic in parametric analysis is simply the sum of individual family LOD scores, uninformative families, obviously, have absolutely no effect on the overall results.

In conclusion, the "bias" in linkage analysis claimed by Schork and Greenwood does not affect most modern nonparametric (and parametric) linkage analysis methods. The handling of incomplete information remains an active area of research in some specialized linkage settings.

GONCALO ABECASIS,[1] NANCY COX,[2] MARK J. DALY,[3] LEONID KRUGLYAK,[5,6] NAN LAIRD,[4] KYRIACOS MARKIANOS,[5] AND NICK PATTERSON[3]
[1]*University of Michigan, Ann Arbor;* [2]*University of Chicago, Chicago;* [3]*Whitehead Institute, Cambridge, MA;* [4]*Harvard School of Public Health, Boston;* [5]*Fred Hutchinson Cancer Research Center, Seattle; and* [6]*Howard Hughes Medical Institute, Chevy Chase, MD*

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

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm [with discussion]. J Roy Stat Soc B 39:1–38

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 (1998) Linkage thresholds for two-stage genome scans. Am J Hum Genet 62:994–996

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

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 N, Greenwood T (2004) Inherent bias toward the null hypothesis in conventional multipoint nonparametric linkage analysis. Am J Hum Genet 74:197–207

Sham PC, Purcell S, Cherny SS, Abecasis GR (2002) Powerful regression-based quantitative-trait linkage analysis of general pedigrees. Am J Hum Genet 71:238–253

Address for correspondence and reprints: Dr. Leonid Kruglyak, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, D4-100, Seattle, WA 98109. E-mail: leonid@fhcrc.org

The authors are listed alphabetically.

© 2004 by The American Society of Human Genetics. All rights reserved.
0002-9297/2004/7504-0023$15.00

Am. J. Hum. Genet. 75:723–727, 2004

Got Bias? The Authors Reply

To the Editor:

We are happy to see that our colleagues have taken seriously the issue we raised in our article (Schork and Greenwood 2004), and, in essence, we do not disagree with much of the factual content of their letters (Abecasis et al. 2004; Mukhopadhyay et al. 2004; Visscher and Wray 2004 [all in this issue]). However, we strongly disagree with aspects of their commentaries and will concentrate on four related issues in our response: (1) the use of the word "bias" to characterize the effects of the treatment of non–fully informative observations as though they were fully informative, in a nonparametric linkage analysis setting; (2) the prevalence and pervasiveness of the inappropriate treatment of non–completely informative observations, in nonparametric linkage analyses; (3) the use of both simulation studies and published "guidelines" for the interpretation of linkage test statistics in the face of inappropriate treatment of non–fully informative observations; and (4) the difference between, and need for refinements in, paramet-