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

hypothesis) for the 50 unobserved tosses (25 tails and 25 heads), the estimate of the probability of heads is 0.65. Schork and Greenwood use the fact that the estimate of 0.65 is <0.80 to make their point that there is an inherent bias toward the null hypothesis when unobserved outcomes are assigned an expected value under the null hypothesis. However, to draw statistical inference from this experiment, we need to compare the observed statistic (in this case, the estimate of the probability of heads) with the variance of that statistic under the null hypothesis (H_0), to create a test statistic. In the case of $\hat{p} = 0.80$, the variance under the null hypothesis is $0.5(1 - 0.5)/50 = 0.005$. If, without loss of generality, we create a test statistic (T) that is the deviation of the estimate from its expectation, divided by the SE of the estimate—that is, $T = (\hat{p} - 0.50)/\sigma(\hat{p}|H_0)$—we obtain the T value $(0.80 - 0.50)/\sqrt{0.005} = 3\sqrt{2}\,(= 4.24)$. The variance of the estimate of 0.65 is $[(50)(0.50)(1 - 0.50) + 0]/100^2 = 0.00125$, and, in this case, the test statistic is $(0.65 - 0.50)/\sqrt{0.00125} = 3\sqrt{2}$. Hence, the two test statistics are identical, and the inference from both experiments is the same, if the correct variance of the sufficient statistic is used. Despite the lower estimate of the probability of heads for the case in which unobserved outcomes were assigned the expected value, the test statistic is the same, because the variance of the estimate is lower. This should not be surprising, because all we have done in the second case is add a constant to a random variable and scale it by another constant.

With linkage analysis, the same analogy holds. Schork and Greenwood (2004) base their conclusions on presented statistics (likelihood-ratio scores) and implicitly assume that the distribution of these test statistics under the null hypothesis is the same for all comparisons, when it is not. For collections of small families, computer programs such as GeneHunter (Kruglyak et al. 1996) and Merlin (Abecasis et al. 2002) calculate the correct variance of the sharing statistic, conditional on all observed marker information; therefore, the correct test statistic and P value are computed (see also Cordell [2004]). For large complex pedigrees, the exact variance of the sharing statistic cannot be computed, and it has been pointed out elsewhere that to assume fully informative markers when there is missing information can reduce power of detection (Kong and Cox 1997). Kong and Cox (1997) present a modification of the test statistic, taking into account that the precision of the estimation of IBD allele sharing varies between pairs of relatives.

Linkage analysis of quantitative traits to map QTL is typically a two-stage procedure with several well-known approximations. In the first stage, IBD probabilities are calculated (or IBD proportions are estimated) and, in the second stage, a regression or variance analysis is performed using the phenotypes and IBD proportions. One implicit assumption of these methods is that the proportion of alleles shared IBD between a pair of individuals is known without error. This is most easily seen in those least-squares regression methods in which the estimated proportions of alleles shared IBD ($\hat{\pi}$) are the x variables, because, in regression analysis, the x values are taken as "fixed." If marker informativeness varies between families (or between pairs of individuals within a family), then this variation is not taken into account in these analyses, and one would not expect these QTL-mapping methods to be invariant with respect to uninformative pairs. The approximation in the use of the expected proportion of alleles shared IBD, instead of the full distribution, has been tested (e.g., by Gessler and Xu [1996]). Gessler and Xu (1996) explicitly make the distinction between the "distribution approach" and the "expectation approach" and conclude that there is little difference between them, in terms of power. Cordell (2004) performed simulations to investigate the "bias" in the test statistic for a number of regression and variance-components QTL-mapping methods. As acknowledged by the author, the simulation parameters used were rather extreme, because there was no sibling resemblance other than that due to a single diallelic QTL, and this QTL explained >90% of the phenotypic variance. Cordell (2004) showed the mean difference in the test statistic when uninformative pairs were left out or were kept in the analysis and showed the SD of that difference for a range of test statistics. However, the scale of the test statistic varies between methods, and the mean and SD of the difference in test statistics do not necessarily show how important these results are in practice. We have performed additional simulations, using both Cordell's parameters and a less extreme set of parameters, and have expressed the mean and SD of the difference in test statistics when uninformative pairs are left out or are kept in, as a function of the average test statistic and the SD of the test statistic. Results are shown in table 1 for the Haseman-Elston LOD (HE-LOD) and variance-components LOD (VC-LOD) methods (see Cordell [2004] for details). Clearly, when put in perspective, the effect on the test statistic either of keeping uninformative pairs in the analysis or of removing them is very small. For example, even in the extreme case of a QTL heritability of 98% and 50% uninformative pairs, the average difference in test statistics is only 4% (HE-LOD) and <1% (VC-LOD) of the average test statistic, and the SD of the difference in test statistics when uninformative pairs are kept in or left out is only 6% (HE-LOD) and 2% (VC-LOD) of the SD of the test statistic. As Cordell (2004) pointed out, the slight increase in the HE-LOD test statistic when uninformative pairs are removed is the result of a decrease in the residual variance in the regression analysis. The decrease in the VC-LOD when uninformative pairs are removed (too small to show in table 1 but reported by Cordell

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-

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.

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-

Table 1

SOLAR Variance-Components Analysis of a Quantitative Trait in which Some Families Are Forced to Be Uninformative

	LOD Scores				
Family ID	Original Data[a]	Linked[b]	Data Removed[c]	Homozygous[c]	Heterozygous[c]
1334	−.2215	...	−.0949	−.0949	−.0949
1345	−.2390	...	−.1149	−.1149	−.1149
1346	2.0463	2.1224	1.7108	1.7108	1.7108
1349	.1005	−.1837	.2588	.2588	.2588
1350	−.1488	...	−.2615	−.2615	−.2615
1358	.6113	...	.3778	.3778	.3778
1362	−.0085	...	−.4964	−.4964	−.4964
1377	.8464	1.4597	.6918	.6918	.6918
1408	.7518	1.1983	1.1969	1.1969	1.1969
1418	−.4073	...	−.4245	−.4245	−.4245
1421	.7240	.9921	1.1242	1.1242	1.1242
1424	.3491	.1458	.3342	.3342	.3342
Total LOD	4.4043	5.7346	4.3023	4.3023	4.3023

[a] Per-family scores produced by SOLAR with the original data.

[b] Scores computed on the basis of a reanalysis of only the families showing linkage in the original data.

[c] Per-family scores computed from an analysis in which the families contributing negative evidence for linkage in the original analysis were made uninformative at the marker locus by removing their genotype data (column 4), by making them homozygous for the same allele (column 5), or by making them heterozygous for the same alleles (column 6).

ric and nonparametric linkage–based gene-discovery strategies.

First, our commentators generally take offense to the use of the word "bias" in our description of what happens in a nonparametric linkage analysis when uninformative or partially informative observations (e.g., affected sibling pairs with non–completely informative marker-genotype data) are treated on equal footing with those that are fully informative. We are in no way wedded to the term "bias" and do not actually care how one refers to the issue we raised in our paper, whether as a "conservative handling" of partially informative observations or as a "power loss" due to the treatment of partially informative observations as though they were fully informative. We do want to emphasize that, as shown by Cordell (2004), the treatment of non–completely informative observations as though they were informative does, on average, deflate the test statistic toward values more consistent with the null hypothesis—as we showed in our simple and contrived example involving affected sibling pairs and the coin-flip example—and thus suggests that this phenomenon induces a tendency or "bias" (in a general sense) toward test-statistic values closer to the null hypothesis.

Second, our commentators dwell on the elegant work of Kong and Cox (1997), which considers the issue we describe in the context of affected–sibling-pair analyses. Kong and Cox (1997) provide a test statistic that appropriately combines uninformative and informative observations into a test statistic based on marker information. However, not all statistics currently in use exploit the principles described by Kong and Cox (1997). For example, a very recent survey by Cordell (2004) suggests that, indeed, statistics do exist that inappropriately treat non–completely informative observations as though they were fully informative, although the degree to which this phenomenon affects various linkage test statistics is context dependent. Thus, for example, of the six statistics for quantitative-trait analysis that Cordell examined, only one—the statistic implemented in the Merlin "Regress" software module—did not show the effects of this phenomenon. In this context, it could be said that perhaps the message of Kong and Cox (1997) simply has not reached the broader genetics community in the way our commentators would like. It should also be noted that Cordell's study was not exhaustive, suggesting that more research investigating other statistics is needed.

Since the SOLAR analysis program (Almasy and Blangero 1998) was not considered by Cordell (2004), we explored its handling of uninformative families in a simple study meant to showcase the issue of concern in a practical example. We want to emphasize that we believe SOLAR provides an excellent suite of genetic analysis tools despite the issue we expose (which is a result of the potential complexity of its handling in the setting of variance-components models). We used a subset of the data investigating the genetic determinants of a molecular phenotype in genotyped three-generation CEPH pedigrees (Greenwood et al. 2004; data available on re-

quest). We analyzed 12 CEPH pedigrees together and then forced 5 of them to be completely uninformative by three different methods. We then compared the results, which are presented in table 1, as per-family and overall LOD scores. Families with blank data in the second column not only contributed negative evidence for linkage to the overall original linkage signal (column 1 of table 1) but were forced to be completely uninformative in subsequent analyses—a phenomenon which, if accounted for properly (i.e., by not considering the contribution of the uninformative families to the linkage signal), should increase evidence for linkage, via the LOD score.

From table 1, it can be seen that the LOD score actually decreases (from 4.4 to 4.3) when the families providing negative evidence for linkage are made completely uninformative, which suggests that informativeness is not accounted for in this analysis. In addition, family 1358 was uninformative in the original data set yet contributed substantial positive evidence for linkage, which, again, is consistent with the potential for the inclusion of uninformative families to increase the value of the linkage statistic because of stochastic effects, as discussed in our article (Schork and Greenwood 2004) and Cordell's (2004). (Indeed, the individual informative and uninformative family LODs are simply summed without weighting, to give the total LOD, thus allowing the uninformative families to contribute to the LOD score for the sample.) We also found that the variance-components statistic implemented in the Merlin software package provided exactly the same overall LOD scores for these families as SOLAR did in each context, suggesting that Merlin is computing statistics in the same way as SOLAR.

Third, although simulation-based tests could be of value in helping determine the impact of the use of statistics that inappropriately treat non–completely informative observations as though they were informative in actual linkage studies (i.e., by simulating the process of including non–completely informative families in data sets and then estimating *P* values for observed statistics from these simulations), such practices can be problematic for a number of reasons:

1. Resorting to simulation studies merely reinforces the need to accommodate inappropriate handling of non–completely informative observations in the construction of a test statistic.
2. One would have to simulate in accordance with the exact mechanism creating the lack of informativeness (partial missing genotype data, marker informativeness, etc.), although the use of permutation tests of allele-sharing information in certain settings may ease this problem (note that not all computer programs provide, by default, *P* values for statistics based on simulation studies)—in addition, this would have to be pursued on a locus-by-locus basis to accommodate the marker information (and/or lack thereof) at each locus.
3. Point estimates of relevant parameters (sibling risk, variance explained, etc.) would not be as reliable as those obtained in a comparable sample of informative observations (as described in our analogy to flipping a coin).
4. Analyses that require simulation studies would produce actual test statistics that are highly context dependent (e.g., a low LOD score on one chromosome may have a low *P* value as a result of the reductions in the test statistic that arise from the inclusion of non–completely informative families as though they were completely informative, whereas a high LOD score on a different chromosome may have a high *P* value for the same reason), which would undermine conventional "guidelines" for assessment of linkage evidence based on test-statistic values—for example, to convey the value of a LOD score as an indication of linkage strength (Lander and Kruglyak 1995)
5. Because of the nonmonotonic relationships between test-statistic values that require simulations to assess significance, total sample size (i.e., a sample that is not adjusted for informativeness), and *P* values (from the simulations), one would have to be conscious not only of test statistics conveying linkage with artificially low values through these simulation studies but also of test statistics with artificially high values for the same reason—especially for statistics, such as variance-components statistics, that show wide variation in values when constructed without appropriate weighting for marker informativeness (Cordell 2004).

It is thus arguably better to use statistics that are designed to account for marker informativeness. In this context, however, studies that have not used, for example, locus-by-locus simulation studies to investigate the effect of the inclusion of non–completely informative observations on test-statistic values obtained throughout the genome might benefit from such studies, since interpretation of the statistical significance of their results is in doubt (see, e.g., the otherwise comprehensive and excellent studies by Panhuysen et al. [2003] and Arya et al. [2004])—a practice entirely consistent with the advice given in our article (Schork and Greenwood 2004).

Fourth, the problem of the inappropriate handling of non–completely informative observations is unique to nonparametric, as opposed to parametric, linkage analysis, since many conventional nonparametric linkage test statistics make use of assigned or imputed allele-sharing

values in their construction from available marker information. Thus, the inappropriate treatment of allele-sharing values assigned to observations that do not have informative marker data creates problems. This simply is not the case in conventional parametric linkage analysis, where, for example, uninformative observations simply do not contribute to a linkage statistic (i.e., they do not contribute positively or negatively to the signal but contribute a value of 0.0 to the overall LOD score, as though they were simply removed from the analysis).

To combat the issue we exposed, we suggest the following actions, all of which are consistent with our commentators' considerations: (1) software documentation should inform the user about (appropriate) potential problems in interpreting test statistics implemented in that software at face value (e.g., on the basis of the guidelines published by Lander and Kruglyak [1995] that focus on actual test-statistic values, such as LOD scores or *t* statistics); (2) simulation-based *P* values should be provided by default for problematic test statistics; and (3) greater emphasis should be placed on the derivation and use of statistics that, like the statistic in Kong and Cox (1997), are based on sound statistical principles for the treatment of non–completely informative observations.

The problems plaguing the reconciliation of multiple nonparametric linkage analysis results—in, for example, the combination of evidence to guide a positional cloning effort—are both numerous and vexing. Consider a recent example in which a LOD score of 11.68 implicating a susceptibility locus for myocardial infarction was reported (Wang et al. 2004*a*; see also the correspondence of Newton-Cheh et al. [2004] and Wang et al. [2004*b*]). On the basis of conventional guidelines, this LOD score should have (and was reported to have) an associated nominal *P* value of ~.00000000001, making it one of the (if not the single) most significant linkages ever reported for a complex trait. However, after simulation studies, this LOD score was found to have a *P* value of .0001 (still impressive but much less so). Although it is unclear if the statistic used to produce the LOD score of 11.68 was plagued by the stochastic effects of treating of non–fully informative observations as though they were informative, our article (Schork and Greenwood 2004) (and Cordell's [2004]) suggests that some statistics could (and, in fact, do) treat them this way and hence could lead to interpretive difficulties and discrepancies of this type. It is in this context that we provided the conclusion in our article, which we restate here with minor parenthetical qualifications (in brackets): "…researchers who have actually conducted relevant linkage studies (without completely informative data) in the past and ignored, or were not aware of, [the allele-sharing information] problem [i.e., by, e.g., knowingly or unknowingly using an available, though problematic, statistic without adjustment via, e.g., extensive locus-by-locus simulation studies] should go back and revisit their analyses" (Schork and Greenwood 2004, p. 316).

Acknowledgments

This work was supported by the following large-scale human genetics research programs: the National Heart, Lung, and Blood Institute (NHLBI) Family Blood Pressure Program (HL64777-01), the NHLBI hypertension SCOR program (HL54998), the National Institutes of Health Pharmacogenetics Network (HL69758-01), and the National Institute of Medical Health Consortium on the Genetics of Schizophrenia (1 R01 MH06557-01A1). The authors would like to thank Dr. Heather Cordell, for critical discussions and the opportunity to review her work in progress.

NICHOLAS J. SCHORK AND TIFFANY A. GREENWOOD
Polymorphism Research Laboratory,
University of California–San Diego,
La Jolla

References

Abecasis G, Cox N, Daly MJ, Kruglyak L, Laird N, Markianos K, Patterson N (2004) No bias in linkage analysis. Am J Hum Genet 75:722–723 (in this issue)

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

Arya R, Duggirala R, Jenkinson CP, Almasy L, Blangero J, O'Connell P, Stern MP (2004) Evidence of a novel quantitative-trait locus for obesity on chromosome 4p in Mexican Americans. Am J Hum Genet 74:272–283

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

Greenwood TA, Cadman PE, Stridsberg M, Nguyen S, Taupenot L, Schork NJ, O'Connor DT (2004) Genome-wide linkage analysis of chromogranin B expression in the CEPH pedigrees: implications for exocytotic sympathochromaffin secretion in humans. Physiol Genomics 18:119–127

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

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

Mukhopadhyay I, Feingold E, Weeks DE (2004) No "bias" toward the null hypothesis in most conventional multipoint nonparametric linkage analysis. Am J Hum Genet 75:716–718 (in this issue)

Newton-Cheh C, Larson M, Kathiresan S, O'Donnell C (2004) On the significance of linkage studies of complex traits. Am J Hum Genet 75:151–152

Panhuysen CIM, Cupples LA, Wilson PWF, Herbert AG, Myers RH, Meigs JB (2003) A genome scan for loci linked to quantitative insulin traits in persons without diabetes: the Framingham offspring study. Diabetologia 46:579–587

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

Visscher PM, Wray NR (2004) Conventional multipoint nonparametric linkage analysis is not necessarily inherently biased. Am J Hum Genet 75:718–720 (in this issue)

Wang Q, Rao S, Shen GQ, Li L, Moliterno DJ, Newby LK, Rogers WJ, Cannata R, Zirzow E, Elston RC, Topol EJ (2004*a*) Premature myocardial infarction novel susceptibility locus on chromosome 1P34–36 identified by genomewide linkage analysis. Am J Hum Genet 74:262–271 (erratum 74:1080)

Wang Q, Rao S, Topol EJ (2004*b*) Reply to Newton-Cheh et al. Am J Hum Genet 75:152–154

Address for correspondence and reprints: Dr. Nicholas J. Schork, Polymorphism Research Laboratory, University of California–San Diego, Department of Psychiatry 0603, 9500 Gilman Drive, La Jolla, CA 92093-0603. E-mail: nschork@ucsd.edu

0002-9297/2004/7504-0024$15.00

Am. J. Hum. Genet. 75:727–730, 2004

Germline *PHOX2B* Mutation in Hereditary Neuroblastoma

To the Editor:

We read with interest the study by Trochet and colleagues (2004), published in the April 2004 issue of *The American Journal of Human Genetics,* that described germline mutations of the paired-like homeobox 2B gene (*PHOX2B* [MIM 603851]) in neuroblastoma (MIM 256700). We have also considered *PHOX2B* as a candidate gene for predisposition to neuroblastoma, and we now report on a germline *PHOX2B* mutation in a pedigree with neuroblastoma. However, we also show that there is no evidence for mutation of this gene in eight other pedigrees with neuroblastoma screened to date. We think these data establish *PHOX2B* as the first bona fide gene that can predispose to neuroblastoma when mutated in the germline, and the findings further emphasize the complex genetics of this important pediatric malignancy.

We previously demonstrated linkage of hereditary neuroblastoma to 16p12-13 by use of a genomewide screening strategy (Maris et al. 2002). Positional cloning of a putative 16p12-13 hereditary neuroblastoma-predisposition gene (*HNB1*) is ongoing, but the critical genomic region for this gene remains large. We had previously considered and excluded other genes known to be mutated in Hirschsprung disease (MIM 142623) and/or in congenital central hypoventilation syndrome (CCHS [MIM 209880]) as candidates for *HNB1,* because these disorders can occur coincident with both sporadic and hereditary neuroblastoma (Maris et al. 2002). Because of the recent reports that the vast majority of patients with CCHS harbor *PHOX2B* mutations, including two patients also affected with neuroblastoma (Amiel et al. 2003; Weese-Mayer et al. 2003), we initiated a screen for germline mutations in this gene in our series of pedigrees with neuroblastoma.

Oligonucleotide primer pairs flanking the coding regions of exons 1, 2, and 3 of *PHOX2B* were designed by use of the program Primer 3.0; these primer pairs were used for PCR amplification and bidirectional sequencing of purified PCR products (primer sequences available on request). We screened germline DNA from the proband and an unaffected family member for each of the seven families that showed cosegregation of a 16p haplotype with disease, as well as for two pedigrees that consisted of cousins with neuroblastoma with no cosegregation of 16p marker haplotypes (see Maris et al. [2002] for details of pedigrees). We also sequenced 109 control DNA samples from the Coriell SNP500 Cancer Panel (Coriell Cell Repositories). All sequence aberrations were confirmed by repeat sequencing after cloning of purified PCR products (TOPO TA Cloning Kit [Invitrogen]), and DNA samples from the remaining available members of the pedigree were also screened for the variant. The Children's Hospital of Philadelphia institutional review board approved this work.

A heterozygous single-base deletion (676delG) was discovered in a complex pedigree with neuroblastoma (fig. 1) (see dbSNP Home Page). This family has seven members in three generations affected with neuroblastoma, and two of these individuals were also shown to have Hirschsprung disease. The proband was affected with neuroblastoma, Hirschsprung disease, and neurofibromatosis type 1 (MIM 162200). The putative nonsense mutation 676delG segregated with neuroblastoma through all three generations, and the frameshift was predicted to produce a slightly truncated protein that would no longer code for the second polyalanine tract. This family had previously been shown to cosegregate a 16p12-13 haplotype with neuroblastoma, and the proband was also shown to have an inactivating mutation in *NF1* (3775delT) that was not present in either of her parents (Maris et al. 2002). Tumor material was available only for patient 1-001, and the tumor exon 3 sequence remained heterozygous for the 676delG mutation. In addition, loss-of-heterozygosity studies using microsatellite markers (D4S2912, D4S1587, D4S405, D4S2971, and D4S428) that are closely linked to the *PHOX2B* locus showed no evidence for allelic deletion. The only other sequence variant discovered in the remaining eight pedigrees was a putative SNP (C552T) in pedigree 12 that is not predicted to affect the resultant protein sequence (S184S) (see dbSNP Home Page). This

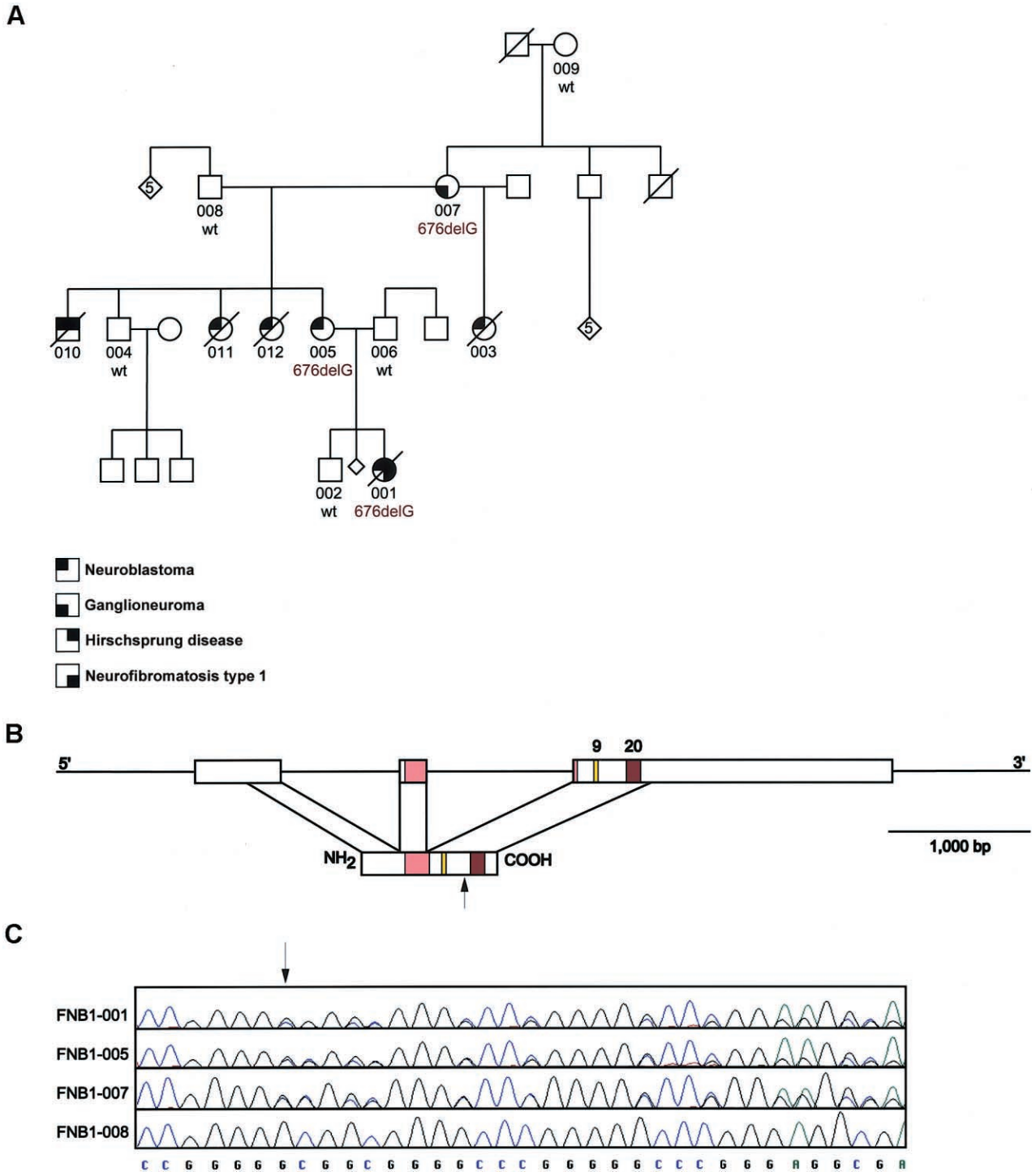

Figure 1 Germline *PHOX2B* mutation in a pedigree segregating neuroblastoma and Hirschsprung disease. *A*, Family 1 pedigree structure. DNA samples from this family with neuroblastoma were available only for patients with a *PHOX2B* result shown (wt = wild type; 676delG = heterozygous mutation segregating through three generations). Patients 1-001 and 1-010 also had Hirschsprung disease, and patient 1-001 was affected with neurofibromatosis type 1. *B*, Genomic organization of *PHOX2B* and schematic representation of the PHOX2B protein. The homeobox domain and the 9– and 20–amino-acid polyalanine repeats are shown for both schemas (*pink*, *yellow*, and *dark red boxes*, respectively). The location of the 676delG mutation is shown (*arrow*). *C*, DNA sequence electropherograms from exon 3 of the *PHOX2B* gene, showing a heterozygous single-base deletion at G676 (*arrow*) that segregates through all three generations but does not appear in a representative unaffected family member.

sequence variant was present in the proband but was not detected in the patient's affected father (no maternal DNA sample was available). It is important to note that neither sequence variant was identified in the bidirectional sequencing of 218 alleles from the control sample set. This strongly suggests that the 676delG sequence variant that segregates with the disease phenotype is a true mutation. The C552T sequence variant, which does not segregate with the disease, is more likely a very rare polymorphism, but we cannot formally exclude the possibility that there might be a functional effect of this presumably neutral polymorphism.

Accumulated data strongly implicate *PHOX2B* as an essential regulator of normal autonomic nervous system development (Pattyn et al. 1999; Brunet and Pattyn 2002). The discovery of polyalanine-expansion mutations in the majority of patients with CCHS clearly defines a role for this gene in human disease (Amiel et al. 2003; Weese-Mayer et al. 2003), and there appears to be a correlation between the severity of the respiratory symptoms and the length of polyalanine expansion (Weese-Mayer et al. 2003; Matera et al. 2004). Neuroblastoma represents perhaps the most aberrant phenotype that results from abnormal adrenergic tissue development. The rare but well-described synchronous appearance of neuroblastoma with other disorders of the autonomic nervous system has suggested a common genetic etiology often referred to as a "neurocristopathy" (Gaisie et al. 1979; Nemecek et al. 2003). Although other genes implicated in Hirschsprung disease and/or CCHS have not been excluded as hereditary neuroblastoma-predisposition genes (Maris et al. 2002; Perri et al. 2002), our data further establish *PHOX2B* as an important gene involved in the initiation of neuroblastoma tumorigenesis. However, the fact that the majority of pedigrees studied here do not show *PHOX2B* mutations clearly implicates locus heterogeneity for hereditary predisposition to neuroblastoma. Assuming that our inferences of linkage to 16p are correct, and in light of the observation of two germline mutations in the proband of the family presented here, we suggest that an oligogenic mechanism for neuroblastoma initiation should be considered, as has been shown for other diseases of neural crest–derived tissues (Gabriel et al. 2002).

It is not yet clear if the *PHOX2B* mutations discovered in patients with hereditary or sporadic neuroblastoma result in gain or loss of protein function. The hypothesis that *PHOX2B* functions as a tumor suppressor is supported by the potential predicted consequence of the five mutations described, to date, in patients with neuroblastoma. Weese-Mayer and colleagues discovered a nonsense mutation that predicts a significantly truncated protein that would miss most of exon 3, including all of the 20-alanine repeat motif (Weese-Mayer et al. 2003). The frameshift mutation described here is similar to that reported by Amiel and colleagues (2003) in a patient who also had CCHS, Hirschsprung disease, and neuroblastoma, and, in both cases, the changes in reading frame are predicted to abolish the polyalanine tract. Trochet and colleagues (2004) discovered two missense mutations, both of which map to a conserved portion of the homeodomain and thus may interfere with DNA binding. On the other hand, 4p12 is not a known site of frequent allelic deletion in neuroblastoma (Maris and Matthay 1999), and, to date, biallelic inactivation of *PHOX2B* has not been demonstrated, although far too few cases have been examined to assert this with confidence.

Taken together, these data suggest that *PHOX2B* mutations are involved in the initiation of neuroblastoma tumorigenesis, especially in patients with associated disorders of the autonomic nervous system. Our data also indicate that germline mutational events in this gene are not involved in the majority of hereditary neuroblastoma cases and that alternative genetic events may predispose to tumorigenesis. Examination of additional patients will facilitate the definition of *PHOX2B* mutation frequency in the genetic and (apparently) sporadic forms of neuroblastoma and will help to clarify the role of *PHOX2B* mutations in tumor initiation and progression.

Acknowledgments

This work was supported in part by grants from the National Institutes of Health (R01-CA78545 [to J.M.M.]), the Children's Hospital of Philadelphia General Clinical Research Center (M01-RR00240), and the Abramson Family Cancer Research Institute.

Yael P. Mosse,[1] Marci Laudenslager,[1] Deepa Khazi,[1] Alex J. Carlisle,[1] Cynthia L. Winter,[1] Eric Rappaport,[1] and John M. Maris[1,2,3]

[1]*The Children's Hospital of Philadelphia,* [2]*Department of Pediatrics, University of Pennsylvania School of Medicine, and* [3]*Abramson Family Cancer Research Institute, Abramson Cancer Center, University of Pennsylvania, Philadelphia*

Electronic-Database Information

The URLs for data presented herein are as follows:

dbSNP Home Page, http://www.ncbi.nlm.nih.gov/SNP/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for *PHOX2B,* neuroblastoma, Hirschsprung disease, CCHS, and neurofibromatosis type 1)

References

Amiel J, Laudier B, Attie-Bitach T, Trang H, de Pontual L, Gener B, Trochet D, Etchevers H, Ray P, Simonneau M, Vekemans M, Munnich A, Gaultier C, Lyonnet S (2003)

Polyalanine expansion and frameshift mutations of the paired-like homeobox gene *PHOX2B* in congenital central hypoventilation syndrome. Nat Genet 33:459–461

Brunet JF, Pattyn A (2002) *Phox2* genes: from patterning to connectivity. Curr Opin Genet Dev 12:435–440

Gabriel SB, Salomon R, Pelet A, Angrist M, Amiel J, Fornage M, Attie-Bitach T, Olson JM, Hofstra R, Buys C, Steffann J, Munnich A, Lyonnet S, Chakravarti A (2002) Segregation at three loci explains familial and population risk in Hirschsprung disease. Nat Genet 31:89–93

Gaisie G, Oh KS, Young LW (1979) Coexistent neuroblastoma and Hirschsprung's disease: another manifestation of the neurocristopathy? Pediatr Radiol 8:161–163

Maris JM, Matthay KK (1999) Molecular biology of neuroblastoma. J Clin Oncol 17:2264–2279

Maris JM, Weiss MJ, Mosse Y, Hii G, Guo C, White PS, Hogarty MD, Mirensky T, Brodeur GM, Rebbeck TR, Urbanek M, Shusterman S (2002) Evidence for a hereditary neuroblastoma predisposition locus at chromosome 16p12-13. Cancer Res 62:6651–6658

Matera I, Bachetti T, Puppo F, Di Duca M, Morandi F, Casiraghi GM, Cilio MR, Hennekam R, Hofstra R, Schober JG, Ravazzolo R, Ottonello G, Ceccherini I (2004) *PHOX2B* mutations and polyalanine expansions correlate with the severity of the respiratory phenotype and associated symptoms in both congenital and late onset central hypoventilation syndrome. J Med Genet 41:373–380

Nemecek ER, Sawin RW, Park J (2003) Treatment of neuroblastoma in patients with neurocristopathy syndromes. J Pediatr Hematol Oncol 25:159–162

Pattyn A, Morin X, Cremer H, Goridis C, Brunet JF (1999) The homeobox gene *Phox2b* is essential for the development of autonomic neural crest derivatives. Nature 399:366–370

Perri P, Longo L, McConville C, Cusano R, Rees SA, Seri M, Conte M, Romeo G, Devoto M, Tonini GP (2002) Linkage analysis in families with recurrent neuroblastoma. Ann NY Acad Sci 963:74–84

Trochet D, Bourdeaut F, Janoueix-Lerosey I, Deville A, De Pontual L, Schleiermacher G, Coze C, Philip N, Frebourg T, Munnich A, Lyonnet S, Delattre O, Amiel J (2004) Germline mutations of the paired-like homeobox 2B (*PHOX2B*) gene in neuroblastoma. Am J Hum Genet 74:761–764

Weese-Mayer DE, Berry-Kravis EM, Zhou L, Maher BS, Silvestri JM, Curran ME, Marazita ML (2003) Idiopathic congenital central hypoventilation syndrome: analysis of genes pertinent to early autonomic nervous system embryologic development and identification of mutations in *PHOX2B*. Am J Med Genet 123A:267–278

Address for correspondence and reprints: Dr. John M. Maris, Division of Oncology, The Children's Hospital of Philadelphia, Abramson Pediatric Research Center 902A, 3516 Civic Center Boulevard, Philadelphia, PA 19104-4318. E-mail: maris@email.chop.edu

0002-9297/2004/7504-0025$15.00

Am. J. Hum. Genet. 75:730–731, 2004

Comparative Frequency of Fragile-X (*FMR1*) and Creatine Transporter (*SLC6A8*) Mutations in X-Linked Mental Retardation

To the Editor:

The study by Rosenberg et al. (2004), in the July 2004 issue of *The American Journal of Human Genetics,* and the previous work by Dr. Salomons's laboratory on the implications of the creatine transporter gene, *SLC6A8,* for X-linked mental retardation (XLMR) are very important contributions to the field (Salomons et al. 2001; Rosenberg et al. 2004). I wish, however, to qualify the concluding sentences in the abstract and the discussion section of the study by Rosenberg et al. (2004), which may lead readers to overestimate the incidence of mutations in the creatine transporter gene in mental retardation (MR). The authors write in the abstract that the "frequency of *SLC6A8* mutations in XLMR is close to that of CGG expansions in *FMR1*" (Rosenberg et al. 2004, p. 97). This is certainly incorrect. Rosenberg et al. (2004) found a 2.2% prevalence of *SLC6A8* mutations in families with proven or possible XLMR (the latter are families with at least two males affected by MR and compatible with X-linked inheritance). On the other hand, the *FMR1* expansion mutation associated with fragile-X syndrome is found in ~2%–3% of males with MR who were not selected for family history (these figures are based on cohorts with little clinical preselection apart from the exclusion of clearly chromosomal or syndromic forms of MR) (see de Vries et al. 1997; Hecimovic et al. 2002; Pandey et al. 2002; Grønskov et al. 2004; Biancalana et al., in press). In fact, when selection is based on possible X-linked inheritance, the proportion of individuals with fragile-X syndrome is much higher. For instance, in the study by Fishburn et al. (1983), fragile-X syndrome accounted for MR in 12 of 45 male sib pairs with "nonspecific" MR, a proportion (27%) that is thus >10 times higher than the reported incidence of *SLC6A8* mutations in a cohort containing sib pairs such as these as well as families with even more obvious XLMR. In fact, we have proposed recently that, unless there are clear hotspots of mutations and/or a very large mutation target size (such as for Duchenne muscular dystrophy, Rett syndrome, and hemophilia A), the population incidence of X-linked diseases implicating genes of average size that lead to highly decreased reproductive fitness is 10–20 times lower than the incidence of fragile-X syndrome (1/50,000–1/100,000 for most X-linked diseases, compared with 1/~5,000 males for the fragile-X syndrome) (Chelly and Mandel 2001). Thus, one expects that the contribution to XLMR of an average gene that does not present mutation hotspots would be 10–20 times lower than that of *FMR1*.

SLC6A8 is such a gene (with 13 exons, a 635-aa coding sequence, and no indication of highly recurring mutations), and thus we predict that its incidence in "nonsyndromic" MR will be in the range of 0.1%–0.3%. Indeed, even for the *ARX* (*X-linked Aristaless*) gene, which has a clear mutation hotspot that accounts for ~6.6% of families with X-linked "nonsyndromic" MR, the incidence of this *ARX* recurring mutation in cohorts of patients with MR is much lower (~0.15%) than that of *FMR1* mutations (Grønskov et al. 2004; Mandel and Chelly 2004).

I also suggest that, in reporting prevalence estimates that are based on relatively small numbers of positive cases, it would be useful to give confidence intervals (CIs). Thus, the observed prevalence, in the study by Rosenberg et al. (2004), of 2.2% may indeed be an underestimate, since some mutations may have been missed and some variants of uncertain significance at present may prove pathogenic, or it may be an overestimate of the true prevalence, since, for the reported data, the CI for the prevalence of proven mutations is 1.0%–4.4%.

J. L. MANDEL
Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC)
CNRS/INSERM/Université Louis Pasteur/Collège de France
Illkirch, France

References

Biancalana V, Beldjord C, Taillandier A, Szpiro-Tapia S, Cusin V, Gerson F, Philippe C, Mandel JL, French National Working Group on Fragile X Syndrome. Five years of molecular diagnosis of fragile X syndrome (1997–2001): a collaborative study reporting 95% of the activity in France. Am J Med Genet (in press)

Chelly J, Mandel JL (2001) Monogenic causes of X-linked mental retardation. Nat Rev Genet 2:669–680

de Vries BB, van den Ouweland AM, Mohkamsing S, Duivenvoorden HJ, Mol E, Gelsema K, van Rijn M, Halley DJ, Sandkuijl LA, Oostra BA, Tibben A, Niermeijer MF (1997) Screening and diagnosis for the fragile X syndrome among the mentally retarded: an epidemiological and psychological survey. Collaborative Fragile X Study Group. Am J Hum Genet 61:660–667

Fishburn J, Turner G, Daniel A, Brookwell R (1983) The diagnosis and frequency of X-linked conditions in a cohort of moderately retarded males with affected brothers. Am J Med Genet 14:713–724

Grønskov K, Hjalgrim H, Nielsen IM, Brøndum-Nielsen K (2004) Screening of the *ARX* gene in 682 retarded males. Eur J Hum Genet. http://www.nature.com/cgi-taf/dynapage.taf?file=/ejhg/journal/vaop/ncurrent/full/5201222a.html (electronically published June 16, 2004; accessed August 16, 2004)

Hecimovic S, Tarnik IP, Baric I, Cakarun Z, Pavelic K (2002) Screening for fragile X syndrome: results from a school for mentally retarded children. Acta Paediatr 91:535–539

Mandel JL, Chelly J (2004) Monogenic X-linked mental retardation: is it as frequent as currently estimated? the paradox of the ARX (Aristaless X) mutations. Eur J Hum Genet (in press)

Pandey UB, Phadke S, Mittal B (2002) Molecular screening of FRAXA and FRAXE in Indian patients with unexplained mental retardation. Genet Test 6:335–339

Rosenberg EH, Almeida LS, Kleefstra T, deGrauw RS, Yntema HG, Bahi N, Moraine C, Ropers HH, Fryns JP, deGrauw TJ, Jakobs C, Salomons GS (2004) High prevalence of SLC6A8 deficiency in X-linked mental retardation. Am J Hum Genet 75:97–105

Salomons GS, van Dooren SJ, Verhoeven NM, Cecil KM, Ball WS, Degrauw TJ, Jakobs C (2001) X-linked creatine-transporter gene (*SLC6A8*) defect: a new creatine-deficiency syndrome. Am J Hum Genet 68:1497–1500

Address for correspondence and reprints: Dr. Jean-Louis Mandel, IGBMC, BP10142, 67404 Illkirch, CU Strasbourg, France. E-mail: mandeljl@igbmc.u-strasbg.fr

0002-9297/2004/7504-0026$15.00

Am. J. Hum. Genet. 75:731–732, 2004

Reply to Mandel

To the Editor:

On behalf of all the authors of our recent study (Rosenberg et al. 2004), we thank Dr. Mandel for his comments (Mandel 2004 [in this issue]), to which we fully subscribe, and we apologize for the misleading statement in our article. Indeed, *SLC6A8* mutations, although probably more common than mutations in other known nonsyndromic X-linked mental retardation (MRX) genes except *ARX,* must be much less frequent than pathogenic CGG expansions in the *FMR1* gene. As pointed out by Mandel (2004 [in this issue]), this is convincingly illustrated by the relative paucity of *ARX* mutations in nonselected cohorts of males with mental retardation (MR) (Grønskov et al. 2004). Mandel's second argument, which implies that mutation rates in X-linked genes can be inferred from their lengths and must be intrinsically much lower than the rate of CGG expansions in *FMR1,* is less compelling in view of the evidence for mutational hotspots in many disease genes, including *ARX* and *PQBP1,* a recent addition to the growing list of genes involved in MRX (Kalscheuer et al. 2003). Therefore, the existence of another common but hitherto-unknown cause of nonsyndromic MR cannot be ruled out yet, even though ongoing large-scale mutation screening in regions known to carry many

mutations (Ropers et al. 2003) has so far failed to identify such a gene.

Reliable estimation of the relative importance of *SLC6A8* and other MRX genes in the etiology of MR will have to await systematic screening of large, unselected cohorts of patients with MR. So far, comprehensive studies of this kind have only been reported for a few genes, including *FMR1* and *ARX*.

GAJJA S. SALOMONS[1] AND HANS-HILGER ROPERS[2,3]
[1]*Department of Clinical Chemistry, Metabolic Unit (1WBi2), VU University Medical Center, Amsterdam;* [2]*Department of Human Molecular Genetics, Max Planck Institute for Molecular Genetics, Berlin; and the* [3]*European X-Linked Mental Retardation (XLMR) Consortium*

References

Grønskov K, Hjalgrim H, Nielsen IM, Brøndum-Nielsen K (2004) Screening of the *ARX* gene in 682 retarded males. Eur J Hum Genet. http://www.nature.com/cgi-taf/dynapage.taf?file=/ejhg/journal/vaop/ncurrent/full/5201222a.html (electronically published June 16, 2004; accessed August 16, 2004)

Kalscheuer VM, Freude K, Musante L, Jensen LR, Yntema HG, Gecz J, Sefiani A, et al (2003) Mutations in the polyglutamine binding protein 1 gene cause X-linked mental retardation. Nat Genet 35:313–315

Mandel JL (2004) Comparative frequency of fragile-X (*FMR1*) and creatine transporter (*SLC6A8*) mutations in X-linked mental retardation. Am J Hum Genet 75:730–731 (in this issue)

Ropers HH, Hoeltzenbein M, Kalscheuer V, Yntema H, Hamel B, Fryns JP, Chelly J, Partington M, Gecz J, Moraine C (2003) Nonsyndromic X-linked mental retardation: where are the missing mutations? Trends Genet 19:316–320

Rosenberg EH, Almeida LS, Kleefstra T, deGrauw RS, Yntema HG, Bahi N, Moraine C, Ropers HH, Fryns JP, deGrauw TJ, Jakobs C, Salomons GS (2004) High prevalence of SLC6A8 deficiency in X-linked mental retardation. Am J Hum Genet 75:97–105

Address for correspondence and reprints: Dr. Gajja S. Salomons, VU University Medical Center, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands. E-mail: g.salomons@vumc.nl

0002-9297/2004/7504-0027$15.00