Letters to the Editor

Am. J. Hum. Genet. 71:1235–1236, 2002

Using All Alleles in the Multiallelic Versions of the SDT and Combined SDT/TDT

To the Editor:

Horvath and Laird's sibling disequilibrium test (SDT) provides a nonparametric approach to testing genetic markers for both linkage and association with a disease (1998). The advantage over its parametric alternatives is its validity as a test of association when using sibships containing more than one affected sibling and/or more than one unaffected sibling. Horvath and Laird introduced an SDT for multiallelic markers and a biallelic combined SDT/transmission/disequilibrium test (TDT) when some parental genotypic information is available. Curtis et al. (1999) later developed a multiallelic combined SDT/TDT. The multiallelic versions of these tests are designed for situations in which there is no a priori knowledge of which allele at a marker might have an effect on disease status; otherwise, a biallelic test can be performed on the allele of interest versus all other alleles collapsed into one. A problem with the multiallelic extensions is that the statistic varies depending on which allele is omitted from the analysis. We present an alternative multiallelic SDT (mSDT) that takes into account all the allelic information and is consistent with the biallelic approach. This method can also be applied to the combined SDT/TDT.

In calculating the multiallelic versions of both the SDT and combined SDT/TDT, the statistics d^j, $j = 1,\dots,m$ for a marker with m alleles are used. In the SDT, $d^j = \sum_i d_i^j$, where d_i^j represents the difference between the average number of times allele j occurs in an affected sibling and the average number of times it occurs in an unaffected sibling within sibship i (Horvath and Laird 1998); for the combined SDT/TDT, d^j is the difference between the number of times allele j is transmitted and the number of times it is not transmitted from a heterozygous parent to an affected child (Sham 1997). As discussed in Stuart (1955), a quadratic form of the d^j can be used to create a statistic with an asymptotic χ^2 distribution. It is noted that since $\sum_{j=1}^{m} d^j = 0$, the df for the distribution are $m - 1$. Furthermore, since using all m columns of the variance-covariance matrix creates a singularity, and, thus, the matrix is uninvertible, the natural solution is to eliminate one d^j and the corresponding row and column in the variance-covariance matrix to make it full rank. The invariance of the χ^2 statistic according to which variate (d^j) is omitted from the statistic is demonstrated by Stuart (1955).

To create a nonparametric test, $S_i^j = \operatorname{sgn}(d_i^j)$ is used in place of d_i^j, where $\operatorname{sgn}(d) = -1,0,1$ for $d <, =, > 0$, respectively. Though the sum of the quantities d_i^j, $j = 1,\dots,m$, is 0 for each sibship $i = 1,\dots,N$ and $S_i^1 = -S_i^2$ in the biallelic case, for more than two alleles, the sum over j of the S_i^j is not similarly linearly constrained within a sibship. In fact, the S_i^j can sum over j to either -1, 0, or 1. Despite this fact, multiallelic extensions to the SDT and combined SDT/TDT are formed by arbitrarily dropping one of the $S^j = \sum_{i=1}^{N} S_i^j$ from the analysis. The resulting $\chi^2_{(m-1)}$ test statistic is no longer invariant to which allele's information has been omitted, since there is no linear dependency among the values of S^j; information is being discarded unnecessarily. Furthermore, the variance-covariance matrix $\mathbf{W}$ for $\mathbf{S} = (S^1,\dots,S^m)$ is nonsingular (exceptions are discussed below) before any of the m alleles are omitted. Thus, when all m alleles are used, a valid test statistic can still be created as $\mathbf{S}'\mathbf{W}^{-1}\mathbf{S}$, which has an asymptotic $\chi^2_{(m)}$ distribution (Hettmansperger 1984; Randles 1989).

There are, as mentioned, situations in which $\mathbf{W}$ will not be full rank. Among these are:

1. the biallelic case, in which the S^j are constrained, since there is a perfect negative correlation between S_i^1 and S_i^2 for all i ($\sum_{j=1,2} S_i^j = 0$ for all i);
2. the existence of at least one allele j, such that $S_i^j = 0$ for all N sibships, so that this allele will have a row and column of 0s in $\mathbf{W}$, creating a singularity; and
3. $\sum_{j=1}^{m} S_i^j = C$, the same constant, for all N sibships.

For these situations, we recommend the use of the Moore-Penrose generalized inverse (g-inverse) of the variance-covariance matrix $\mathbf{W}$, $\mathbf{W}^{-}$. This is a unique generalized inverse of $\mathbf{W}$ that satisfies the following conditions (Rao and Mitra 1971; Searle 1971): $\mathbf{W}\mathbf{W}^{-}$ and $\mathbf{W}^{-}\mathbf{W}$ are symmetric; $\mathbf{W}^{-}\mathbf{W}\mathbf{W}^{-} = \mathbf{W}^{-}$; and $\mathbf{W}\mathbf{W}^{-}\mathbf{W} = \mathbf{W}$. It is worth noting that the last two scenarios listed for a singular variance-covariance matrix are possible with the original SDT statistic, even after one allele has been omitted from the analysis, in which case the statistic cannot be calculated, since $\mathbf{W}$ is uninvertible.

When using $\mathbf{W}^-$ in place of $\mathbf{W}^{-1}$ in the quadratic form, the test statistic $\mathbf{S}'\mathbf{W}^-\mathbf{S}$ still has an asymptotic χ^2 distribution, now with df equal to the rank of $\mathbf{W}$ (Rao and Mitra 1971). Note that, for the biallelic case, in Horvath and Laird's notation (1998), the mSDT gives $\mathbf{S} = (b - c, c - b)$, and the $\mathbf{W}$ matrix will be of the form

$$\begin{bmatrix} b+c & -(b+c) \\ -(b+c) & b+c \end{bmatrix} .$$

The g-inverse is then calculated as

$$\begin{bmatrix} 1/(4b+4c) & -1/(4b+4c) \\ -1/(4b+4c) & 1/(4b+4c) \end{bmatrix} ,$$

which yields a χ^2 statistic of $(b-c)^2/(b+c)$ with 1 df, the same as the usual biallelic statistic.

To summarize our approach, we suggest modifying Horvath and Laird's SDT statistic (1998) and the combined SDT/TDT of Curtis et al. (1999) in the following manner to calculate the statistic for the mSDT:

1. Use all m alleles in the $\mathbf{S}$ vector and $\mathbf{W}$ matrix.
2. Use $\mathbf{W}^-$ in place of $\mathbf{W}^{-1}$ to create the χ^2 statistic (note that these are identical when $\mathbf{W}$ is full rank).
3. Use rank($\mathbf{W}$) as the df for the χ^2 distribution.

We give an example here, using simulated data from GAW9 (Hodge 1995). As in Spielman and Ewens (1998) and Knapp (1999), we focus on multiallelic markers D1G31 and D5G23, which contain actual disease alleles, M8 and M7, respectively. Table 1 shows the results of analyzing the data using the original Horvath and Laird SDT method, in which each allele is dropped in turn. Also shown are the results from analyzing the data using our mSDT approach. Note that each marker has eight alleles, so P values from the SDT are based on a χ^2_7 distribution, whereas the mSDT P values are from a χ^2_8 distribution, since the variance-covariance matrices for both markers are full rank. This example is not intended as any sort of power comparison but merely to illustrate that there is not necessarily a loss of power by introducing an additional df. The other thing to note from this table is the variation of the SDT P values depending on which allele is dropped. Although all test statistics are highly significant for marker D5G23, we can see quite a discrepancy between the SDT statistic for marker D1G31 when dropping allele M8 and any of the other seven SDT statistics. The mSDT approach will always give a unique χ^2 statistic, regardless of whether $\mathbf{W}$ is full rank. This method will be available in a future release of SAS/Genetics™.

Wendy Czika[1,2] and Jack J. Berry[1]
[1]*SAS Institute, Cary, NC; and* [2]*Department of Statistics, North Carolina State University, Raleigh*

Table 1

SDT and mSDT Statistics for Two Markers Linked and Associated with Disease

	Statistic for Marker			
Allele	D1G31		D5G23	
Dropped	χ^2	P	χ^2	P^a
M1	23.115255	.001628	52.441075	.000048
M2	23.543802	.001370	52.365979	.000049
M3	23.239746	.001548	52.382481	.000049
M4	23.621073	.001328	51.086058	.000088
M5	23.661028	.001307	52.546616	.000046
M6	23.648748	.001313	53.238694	.000033
M7	23.417311	.001441	45.631132	.001031
M8	14.806102	.038567	51.811979	.000064
mSDT	23.667390	.002605	53.455015	.000088

[a] P values multiplied by 10^4.

References

Curtis D, Miller MB, Sham PC (1999) Combining the sibling disequilibrium test and transmission/disequilibrium test for multiallelic markers. Am J Hum Genet 64:1785–1786

Hettmansperger TP (1984) Statistical inference based on ranks. John Wiley and Sons, New York

Hodge SE (1995) An oligogenic disease displaying weak marker associations: a summary of contributions to problem 1 of GAW9. Genet Epidemiol 12:545–554

Horvath S and Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 63:1886–1897

Knapp M (1999) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. Am J Hum Genet 64: 861–870

Randles RH (1989) A distribution-free multivariate sign test based on interdirections. J Am Stat Assoc 84:1045–1050

Rao CR and Mitra SK (1971) Generalized inverse of matrices and its applications. John Wiley and Sons, New York

Searle SR (1971) Linear Models. John Wiley and Sons, New York

Sham P (1997) Transmission/disequilibrium tests for multiallelic loci. Am J Hum Genet 61:774–778

Spielman RS and Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458

Stuart A (1955) A test of homogeneity of the marginal distributions in a two-way classification. Biometrika 42:412–416

Address for correspondence and reprints: Wendy Czika, R3212 SAS Campus Drive, Cary, NC 27513. E-mail: wendy.czika@sas.com

0002-9297/2002/7105-0025$15.00

Am. J. Hum. Genet. 71:1237–1238, 2002

Griscelli Syndrome Types 1 and 2

To the Editor:

In a recent report, Anikster et al. (2002) identified a *RAB27A* (MIM 603868) deletion in a kindred with Griscelli syndrome (GS) (MIM #214450). Several patients from this kindred displayed neurological manifestations related to the hemophagocytic syndrome (HPS). On the basis of their study, the authors suggest "that the neurological involvement in these patients with GS occurs secondarily to the hemophagocytic syndrome and that patients with primary CNS complications and *MYO5A* (MIM 160777) mutations have a related disorder, namely, Elejalde syndrome." This assertion is certainly correct but surprisingly is presented as new and as an "alternative explanation." Several previously published reports have unequivocally established that neurological manifestations occurring in patients with GS and caused by *RAB27A* mutation are related to lymphocyte infiltration of the CNS (Ménasché et al. 2000; Pastural et al. 2000; de Saint Basile and Fischer 2001), whereas patient(s) with GS caused by *MYO5A* mutations exhibit a primary neurological disease, potentially described as Elejalde syndrome, and is unrelated to the hematopoietic lineage, as also observed in *Myo5a* mutant *dilute* mice (Pastural et al. 1997, 2000; Sanal et al. 2000; de Saint Basile and Fischer 2001; Ivanovich et al. 2001). The common finding of both conditions is albinism that results from the same mechanism—a defective release of melanosome content to neighboring cells, such as keratinocytes in the skin. MyoVA and Rab27A have been shown to interact in the same molecular pathway, resulting in melanosome transport on actin filament to dock at plasma membrane (Marks and Seabra 2001; Hume et al. 2002; Provance et al. 2002; Seabra et al. 2002). There should not be any confusion left, since patients with partial albinism and manifestations of HPS, with or without neurological involvement, should be screened for *RAB27A* mutations and treated accordingly, whereas those with partial albinism and a primary neurological disease without HPS should be screened for *MYO5A* mutations, as discussed elsewhere (Ménasché et al. 2000)

There are numerous examples of conditions grouped under the same umbrella name (such as "Gaucher disease type I to III") because of shared biological mechanisms, but that have different outcomes and treatments. Griscelli syndromes 1 and 2 are other examples.

Incidentally, in table 1 of the Anikster et al. report, the *dilute* and *ashen* murine models have been inverted, potentially causing some confusion.

Gaël Ménasché,[1] Alain Fischer,[1,2]
and Geneviève de Saint Basile[1]
[1]*Unité de Recherche sur le Développement Normal et Pathologique du Système Immunitaire INSERM U429 and* [2]*Unité dImmuno-Hématologie et Rhumatologie Pédiatriques, Hôpital Necker-Enfants Malades, Paris, France*

Electronic-Database Information

Accession numbers and URLs for data presented herein are as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for Griscelli syndrome [MIM #214450], RAB27A [MIM 603868], and MYO5A [MIM 160777])

References

Anikster Y, Huizing M, Anderson PD, Fitzpatrick DL, Klar A, Gross-Kieselstein E, Berkun Y, Shazberg G, Gahl WA, Hurvitz H (2002) Evidence that Griscelli syndrome with neurological involvement is caused by mutations in *RAB27A*, not *MYO5A*. Am J Hum Genet 71:401–414

de Saint Basile G, Fischer A (2001) The role of cytotoxicity in lymphocyte homeostasis. Curr Opin Immunol 13:549–554

Hume AN, Collinson LM, Hopkins CR, Strom M, Barral DC, Bossi G, Griffiths GM, Seabra MC (2002) The leaden gene product is required with Rab27a to recruit myosin Va to melanosomes in melanocytes. Traffic 3:193–202

Ivanovich J, Mallory S, Storer T, Ciske D, Hing A (2001) 12-year-old male with Elejalde syndrome (neuroectodermal melanolysosomal disease). Am J Med Genet 98:313–316

Marks MS, Seabra MC (2001) The melanosome: membrane dynamics in black and white. Nat Rev Mol Cell Biol 2:738–748

Ménasché G, Pastural, Elodie , Feldmann J, Certain S, Ersoy F, Dupuis S, Wulffraat N, Bianci D, Fischer A, Le Deist F, de Saint Basile G (2000) Mutations in RAB27A cause Griscelli syndrome associated with hemophagocytic syndrome. Nat Genet 25:173–176

Pastural E, Barrat FJ, Dufourcq-Lagelouse R, Certain S, Sanal O, Jabado N, Seger R, Griscelli C, Fischer A, de Saint Basile G (1997) Griscelli disease maps to chromosome 15q21 and is associated with mutations in the myosin-Va gene. Nat Genet 16:289–292

Pastural E, Ersoy F, Yalman N, Wulffraat N, Grillo E, Ozkinay F, Tezcan I, Gediköglu G, Philippe N, Fischer A, de Saint Basile G (2000) Two genes are responsible for Griscelli syndrome at the same 15q21 locus. Genomics 63:299–306

Provance DW, James TL, Mercer JA (2002) Melanophilin, the product of the leaden locus, is required for targeting of myosin-Va to melanosomes. Traffic 3:124–132

Sanal O, Yel L, Kucukali T, Gilbert-Barnes E, Tardieu M, Texcan I, Ersoy F, Metin A, de Saint Basile G (2000) An allelic variant of Griscelli disease: presentation with severe hypotonia, mental-motor retardation, and hypopigmentation consistent with Elejalde syndrome (neuroectodermal melanolysosomal disorder). J Neurol 247:570–572

Seabra MC, Mules EH, Hume AN (2002) Rab GTPases, intracellular traffic and disease. Trends Mol Med 8:23–30

Address for correspondence and reprints: Dr. Geneviève de Saint Basile, INSERM U429, Hôpital Necker-Enfants Malades, 169 rue de Sevres, 75015 Paris, France. E-mail: sbasile@necker.fr

0002-9297/2002/7105-0026$15.00

Am. J. Hum. Genet. 71:1238, 2002

Reply to Ménasché et al.

To the Editor:
It is gratifying to learn that Ménasché et al. (2002 [in this issue]) agree with our analysis of the phenotypic differences between patients with *RAB27A* mutations and those with *MYO5A* mutations. We leave it to *Journal* readers to decide if previous publications have "unequivocally established" these points. We do apologize for the error in table 1, which we recognized and have corrected in an erratum.

Perhaps we could make two additional points. First, Gaucher disease types I, II, and III represent examples of defects in a single gene resulting in different phenotypes, whereas Griscelli/Elejalde syndromes represent examples of defects in two different genes resulting in phenotypes with some similarities. Second, we wonder what nomenclature should be employed for these two disorders. Ménasché et al. continue to use Griscelli syndromes types 1 and 2. However, Griscelli's original cases exhibited immune deficiency (Griscelli et al. 1978), whereas Elejalde first recognized a distinct, neurologically based disorder (Elejalde et al. 1979). Perhaps Dr. Elejalde should be credited for the accuracy of his ascertainment.

Marjan Huizing,[1] Y. Anikster,[2] and W. A. Gahl[1]
[1]*Section on Human Biochemical Genetics, Medical Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; and* [2]*Metabolic Unit, Sheba Medical Center, Tel Hashomer, Israel*

References

Elejalde BR, Holguin J, Valencia A, Gilbert EF, Molina J, Marin G, Arango LA (1979) Mutations affecting pigmentation in man. I. Neuroectodermal melanolysosomal disease. Am J Med Genet 3:65–80

Griscelli C, Durandy A, Guy-Grand D, Daguillard F, Herzog C, Prunicras M. (1978) A syndrome associating partial albinism and immunodeficiency. Am J Med 65:691–702

Ménasché G, Fischer A, de Sainte Basile G (2002) Griscelli syndrome type 1 and 2. Am J Hum Genet 71:1237–1238 (in this issue)

Address for correspondence and reprints: Dr. Marjan Huizing, 10 Center Drive, MSC 1851, Building 10, Room 10C-103, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892-1851. E-mail: mhuizing@mail.nih.gov

0002-9297/2002/7105-0027$15.00

Am. J. Hum. Genet. 71:1238–1239, 2002

Family-Based Association Tests Incorporating Parental Genotypes

To the Editor:
The report "Parental Genotypes in the Risk of a Complex Disease" (Labuda et al. 2002) gives several interesting examples of how parental genotypes can contribute to children's disease risk—for example, through maternal effects during pregnancy or paternal effects during spermatogenesis. The authors note that if disease risk depends on parents' genotypes but not their child's genotype, then the distribution of genotypes in cases will not differ from the Mendelian expectation given their parents' genotypes. Hence, the traditional transmission disequilibrium test (TDT) using case-parent trio data will (correctly) not detect any association between individuals' genotypes and disease. The authors present an example in which the TDT provides no evidence of an association between a variant allele and disease (in fact, the point estimate for the odds ratio is 1.0), whereas a comparison of case subjects' genotypes to those of population control subjects does provide evidence of association (estimated odds ratio = 3.4). The authors then compare maternal and paternal genotypes to control subjects' genotypes and find evidence that the prevalence of the variant allele is higher in parents of case subjects than in population control subjects.

However, there are other analytic options in this case—namely, flexible statistical methods for case-parent trios which can test for parental-genotype effects. These have the advantage of being robust to population-stratification bias and, in some situations, are even more powerful for testing for parental-genotype effects than case-control studies (Starr et al. 2002).

The log-linear model developed by Weinberg et al. and Wilcox et al. can test for parental-genotype and parent-of-origin effects after adjusting for possible case-genotype effects (Weinberg et al. 1998; Wilcox et al. 1998). In principle, this model can also test parental-genotype × case-genotype interactions—which could be relevant;

for example, in the study of complications during pregnancy such as preeclampsia (Kilpatrick 1999).

The log-linear model is equivalent to a conditional logistic analysis comparing the case to appropriately defined "pseudo-sibling controls" (Kraft 2002). If the gene under study is assumed not to play a direct role in an individual's risk of disease (or to be linked to any other such gene), then to test for an indirect role of maternal genotype (say) each case subject should be compared to a pseudo-sibling control subject whose mother has the genotype of the case subject's father. That is, if the genotypes of the mother and father are G_m and G_f, respectively, then the conditional logistic likelihood for the family is

$$\frac{e^{\beta Z(G_m)}}{e^{\beta Z(G_m)}+e^{\beta Z(G_f)}}\ ,$$

where $Z(\cdot)$ is some dominance coding.

This approach (reasonably) assumes that, given the set of parental marker genotypes $\{G_1,G_2\}$, it is equally likely that $G_m = G_1$ or G_2. In other words, "the frequency of heterozygous mothers married to homozygous variant fathers is the same as the frequency of heterozygous fathers married to homozygous variant mothers, and so on" (Wilcox et al. 1998). Furthermore, since this likelihood permutes the genotypes of "the parent contributing to disease risk" and the "the parent not contributing to disease risk," it cannot estimate joint effects of both parents' genotypes. However, for many diseases, only the mother (father) will plausibly contribute to a child's disease risk.

Although the case-parent trio analysis conditions on the parents' genotypes and hence is robust to population stratification bias, the analysis comparing parental genotypes to population-based controls is not (although Labuda et al. [2002] argue that this may not be an issue for the particular data they analyze in their report). Furthermore, even when there is no population stratification, the latter analysis is something of an "apples and oranges" comparison, as the exposure of interest is not the control subject's genotype, but his or her parent's genotype. The control's genotype serves as a surrogate for his or her parent's. In a simulation study with 175 unmatched case and control subjects (1,000 replicates), we found that the odds ratio comparing case subjects' maternal genotypes to control genotypes underestimated the odds ratio associated with each variant maternal allele by 11% (variant allele frequency 0.25; baseline probability of disease 14%; odds ratio per variant maternal allele 2). Of course, the data Labuda et al. (2002) analyzed did not contain parental genotype information for the controls. But if one were to design a case-population control study to detect the effects of maternal (paternal) genotypes, then one should plan to collect information on controls' maternal (paternal) genotypes.

Finally, figure 1*a* is misleading in that case subjects' parents are not representative of population controls if individuals' genotypes are associated with disease or there is population stratification.

Peter Kraft and Melissa Wilson
Department of Preventive Medicine
Keck School of Medicine
University of Southern California
Los Angeles

References

Kilpatrick D (1999) Influence of human leukocyte antigen and tumor necrosis factor genes on the development of preeclampsia. Hum Reprod Update 5:94–102

Kraft P (2002) Statistical methods in family-based gene-association studies. PhD dissertation, University of Southern California, Los Angeles

Labuda D, Krajinovic M, Sabbagh A, Infante Rivard C, Sinnet D (2002) Parental genotypes in the risk of a complex disease. Am J Hum Genet 71:193–197

Starr J, Hsu L, Schwartz SM (2002) Maternal genetics as risk factors for disease in offspring: statistical power of the log-linear approach to case-parent triads vs. a case-control design. Am J Epidemiol 155:S50

Weinberg C, Wilcox A, Lie RT (1998) A log linear approach to case-parent data: assessing effects of disease genes that act directly or through maternal effects, and may be subject to parental imprinting. Am J Hum Genet 62:969–978

Wilcox A, Weinberg C, Lie RT (1998) Distinguishing the effects of maternal and offspring genes through studies of "case parent triads." Am J Epidemiol 148:893–901

Address for correspondence and reprints: Dr. Peter Kraft, University of Southern California, 1540 Alcazar Street, CHP 218 MC 9010, Los Angeles, CA 90089-9010. E-mail: pkraft@usc.edu

0002-9297/2002/7105-0028$15.00

Am. J. Hum. Genet. 71:1239–1240, 2002

Regarding "Parental Genotypes in the Risk of a Complex Disease"

To the Editor:

Labuda et al. (2002) have proposed that parental genotypes might play a role in the causation of complex diseases. They seem unaware that this idea has been considered by others (e.g., Lande et al. 1989) and that methods have been developed to test for parentally mediated genetic effects, both for a dichotomous phenotype (Mitchell 1997; Weinberg et al. 1998; Wilcox et al.

1998) and for a quantitative phenotype (van den Oord 2000).

Furthermore, some of the assessments made by Labuda et al. miss the mark. They assume (see their fig. 1) that under scenario A, where the offspring genotype is the one that "counts," the parents of affected children will resemble control parents with respect to the gene under study. This ignores the fact that the genotypes of parents and their children are correlated. Just as the parents of offspring with Huntington disease will differ from population controls in their prevalence of the allele for Huntington disease, parents of offspring who have a complex disease will tend to differ from population controls. Thus, the case-control analyses reported in table 1 of Labuda et al. (2002) are not specific to parentally mediated genetic effects.

There are other reasons, biologic and technical, to doubt the interpretation offered by Labuda et al., who suggest that their data support a parent-mediated effect of *CYP2E1*5* on risk of childhood acute lymphoblastic leukemia. First, the mechanisms by which the maternal and paternal genotypes would influence offspring phenotype are very different (i.e., in utero environment vs. DNA replication errors that produce genetically abnormal sperm). It thus seems unlikely that the etiology of a given condition would be related to both maternal and paternal effects of a single gene. Rather, similar "effects" of the maternal and paternal genotypes, on the basis of case-control parental data, seem more likely due to the selection of a biased control group or to offspring-mediated effects that have confounded the comparison of the (correlated) parental genotypes. Thus, the data offered by Labuda et al., which show very similar odds ratios for the mother and for the father, may be seen more plausibly as reflecting either a systematic bias in the control group or a chance finding.

The final issue is analytic. The odds ratio parameter estimated by the case-control analysis is not the same as that estimated by transmissions. Labuda et al. evidently used a standard method for paired data, calculating the ratio of counts for discordant transmission pairs based on heterozygous parents. This approach estimates the relative penetrance for carriers of a single copy of the variant allele under a gene dose model in which the relative penetrance for two copies is the square of that for one copy. By contrast, in their case-control analysis, Labuda et al. use carrier status, which presumes a dominant model. The paired estimator based on transmissions can be shown to be biased toward 1.0 under such a model. Even if the two analyses were estimating the same parameter, there is considerable overlap in the CIs for the two estimates. For these reasons, the results presented by Labuda et al. (2002) should be seen as providing only very weak evidence for a parent-mediated effect of *CYP2E1*5*.

Clarice R. Weinberg[1] and Laura Mitchell[2]
[1]*National Institute of Environmental Health Sciences, Research Triangle Park, NC; and* [2]*Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia*

References

Labuda D, Krajinovic M, Sabbagh A, Infante-Rivard C, Sinnett D (2002) Parental genotypes in the risk of a complex disease. Am J Hum Genet 71:193–197

Lande R, Price T (1989) Genetic correlations and maternal effect coefficients obtained from offspring-parent regression. Genetics 122:915–922

Mitchell L (1997). Differentiating between fetal and maternal genotypic effects, using the transmission test for linkage disequilibrium. Am J Hum Genet 60:1006–1007

van den Oord EJCG (2000) The use of mixture models to perform quantitative tests for linkage disequilibrium, maternal effects, and parent-of-origin effects with incomplete subject-parent triads. Behav Genet 30:335–343

Weinberg C, Wilcox A, Lie R (1998) A log-linear approach to case-parent triad data: assessing effects of disease genes that act directly or through maternal effects, and may be subject to parental imprinting. Am J Hum Genet 62:969–978

Wilcox AJ, Weinberg CR, Lie RT (1998) Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads." Am J Epidemiol 148:893–901

Address for correspondence and reprints: Dr. C. R. Weinberg, MD A3-03, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709. E-mail: Weinberg@niehs.nih.gov

0002-9297/2002/7105-0029$15.00

Am. J. Hum. Genet. 71:1240–1242, 2002

Reply to Comments by Kraft and Wilson and by Weinberg and Mitchell on "Parental Genotypes in the Risk of a Complex Disease"

To the Editor:

Kraft and Wilson (2002 [in this issue]) point out that there are other analytical options to a joint application of case-control and TDT analysis in our study of the effect of parental genetics in the risk of a complex disease. They propose a "pseudo-sibling controls" design as an alternative to the approach proposed earlier by Weinberg and colleagues (1998) to study parental effects in case-parent trios. However, these tests are directed to evaluate the effect within a presumed model and are not designed to estimate joint effects of both parents' genotypes, which appeared to be the case with our data. Our study, inspired by original experimental observations, led us to understand the underlying genetic effects that did not follow established paradigms. We concluded that a number of complementary strategies will need to be used simultaneously to dissect genetic predisposition to complex disorders (Labuda et al. 2002). In this regard, we are in agreement with Kraft and Wilson (2002 [in this issue]) that additional collecting of control-parent trios would extend possibilities of testing the observed effects under a greater variety of genetic and statistical models.

In the context of simple Mendelian disorders, fig. 1 could be troubling, but our paper was intended to divert the reader from this paradigm. Indeed, in a highly penetrant autosomal recessive condition, such as cystic fibrosis, in which two defective gene copies mean disease, collecting patients obviously identifies heterozygous parents, who otherwise would be difficult to find in such numbers in a control population. The example of Huntington chorea used by Weinberg and Mitchell (2002 [in this issue]) is rather unfortunate, since, at time of diagnosis, the case carrier-parent would already succumb to this disease. In a complex, multifactorial, and multilocus disease, in which the effect of a given allele is likely due to gene-gene (i.e., the presence of another variant at a different locus) and/or gene-environment interaction (e.g., exposure in only a fraction of carriers), one does not necessarily expect the enrichment in the at-risk alleles among patients' parents, expected in turn to transmit these alleles "preferentially" to their case-offspring (fig. 1*A*). In other words, we believe that figure 1 provides a good illustration of the experimental situation we faced.

We apologize for not giving satisfactory credit to earlier developments, which was pointed out by Weinberg and Mitchell (2002 [in this issue]). The fact that reference to Lande and Price (1989) is also absent among articles cited by Weinberg et al. (1998) is not an excuse. Rather, it reflects the fact that these excellent methodological contributions were reported in the absence of experimental data, in contrast, for example, to a recent paper by Infante-Rivard et al. (2002*b*) where both the sampling scenarios include control-parent trios as well as testing for maternally mediated effects.

Obviously, the mechanisms through which the maternal and paternal genotypes could influence child phenotype might be very different, but their net effect relevant to cancer risk, such as an increased mutation burden, need not be. In respect to this, different pathways controlling the metabolism of carcinogens, the level of oxidative stress, or the efficiency of DNA repair may have their unique contributions to the increase in the level of DNA lesions and, consequently, cancer risk. The observed effect with *CYP2E1*5* is, therefore, not at all unlikely. It is, however, possible that, for the effect to occur, the *CYP2E1*5* carriers would have to undergo a particular environmental exposure (Infante-Rivard et al. 2002*a*). However, as with all such results, this is the first report that will have to be confirmed by other studies that include different populations.

Here, our population of case and control subjects, both of French-Canadian origin, seems to be excellent for association studies because of the common genetics and lifestyle. Moreover, as presented in our report, we independently tested this population for a possibility of stratification that, in light of the recent results of Ardlie et al. (2002), appears to be less of a problem in appropriately designed studies. Kraft and Wilson (2002 [in this issue]) evaluated an underestimation of 11% in the odds ratio, related to the use of "surrogate" parental controls. This 11% arises from the elevated disease probability in the chosen numerical example and actually corresponds only to 1 SD in the variant frequency (0.250 ± 0.023) estimated in a sample of 350 chromosomes. Such extent of variation is expected under experimental conditions.

Weinberg and Mitchell (2002 [in this issue]) in their comments were also concerned by the effect of the *CYP2E1*5/*5* homozygotes. Because of the rarity (see table 1 in Labuda et al. 2002) of the variant in question, we did not need to consider the effect of its homozygotes. The dominant effect is, therefore, the one to be assumed to be consistent with presumed phenotypic outcome of this allele, leading to higher inducibility and, therefore, to higher activity of the enzyme (see references in Labuda et al. 2002).

For the reasons discussed above, we believe that our study provides solid evidence for the parental effects. It provides also an experimental illustration of genetic effects that, although escaping a simple Mendelian paradigm, were anticipated in earlier studies such as that of Weinberg and her colleagues (Weinberg et al. 1998). There is, therefore, no reason to believe that these effects should not be expected in other complex diseases.

Damian Labuda,[1,2] Maja Krajinovic,[1,2]
Audrey Sabbagh,[1,*] Claire Infante-Rivard,[1,3]
and Daniel Sinnett[1,2]
[1]*Centre de recherche, Hôpital Sainte Justine,*
[2]*Département de pédiatrie, Université de Montréal,*
and [3]*Department of Epidemiology and Biostatistics*
and Occupational Health, McGill University,
Montreal

References

Ardlie KG, Lunetta KL, Seielstad M (2002) Testing for population subdivision and association in four case-control studies. Am J Hum Genet 71:304–311

Infante-Rivard C, Krajinovic M, Labuda D, Sinnett D (2002*a*) Childhood acute lymphoblastic leukemia associated with parental alcohol consumption and polymorphisms of carcinogen-metabolizing genes. Epidemiology 13:277–281

Infante-Rivard C, Rivard GE, Yotov WV, Genin E, Guiguet M, Weinberg C, Gauthier R, Feoli-Fonseca JC (2002*b*) Absence of association of thrombophilia polymorphisms with intrauterine growth restriction. N Engl J Med 347:19–25

Kraft P, Wilson M (2002) Family-Based Association Tests Incorporating Parental Genotypes. Am J Hum Genet 71: 1238–1239 (in this issue)

Labuda D, Krajinovic M, Sabbagh A, Infante-Rivard C, Sinnett D (2002) Parental genotypes in the risk of a complex disease. Am J Hum Genet 71:193–197

Lande R, Price T (1989) Genetic correlations and maternal effect coefficients obtained from offspring-parent regression. Genetics 122:915–922

Weinberg CR, Wilcox AJ, Lie RT (1998) A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet 62: 969–978

Weinberg CR, Mitchell, L (2002) Regarding "Parental Genotypes in the Risk of a Complex Disease." Am J Hum Genet 71:1239–1240 (in this issue)

Address for correspondence and reprints: Dr. Damian Labuda, Centre de recherche, Hôpital Sainte-Justine, 3175 Côte-Sainte Catherine, Montréal, Québec H3T 1C5 Canada. E-mail: damian.labuda@umontreal.ca

* Present affiliation: CNRS-Laboratoire d'Anthropologie Biologique, Musée de l'Homme, Paris, France.

0002-9297/2002/7105-0030$15.00

Am. J. Hum. Genet. 71:1242–1247, 2002

Partition-Ligation–Expectation-Maximization Algorithm for Haplotype Inference with Single-Nucleotide Polymorphisms

To the Editor:

The mapping of SNPs in human genomes has generated a lot of interest from both the biomedical research community and industry. In conjunction with SNP mapping, researchers have shown that haplotypes possess considerably greater potential than the traditional single-SNP approach in disease-gene mapping and in our understanding of complex landscapes of linkage disequilibrium (LD) (Goldstein 2001). *In silico* methods for haplotype reconstruction have attracted much attention because of their cost-effectiveness and accuracy (Tishkoff et al. 2000) and have played an important role in the definition of human haplotype block structure and in candidate-gene studies of complex traits (Tabor et al. 2002). In a recent publication, Niu et al. (2002) proposed a partition-ligation (PL) strategy and implemented it together with Gibbs sampling, to estimate haplotype phases for a large number of SNPs. Although the resulting program, HAPLOTYPER, has been in high demand from many research groups, a significant portion of researchers are also strongly interested in using an expectation-maximization (EM)–based algorithm. In the present letter, we describe how to combine the PL strategy with the EM algorithm and how to handle the local-mode problem. We also present a fast and robust method of computing the variance of the estimated haplotype frequencies. Some related issues concern the handling of missing data and the multiple imputations of haplotype phases.

The EM algorithm is arguably the most popular statistical algorithm, because of its interpretability and stability. Compared to the Gibbs sampler, the EM approach is a deterministic procedure, requires less computing time, and is easier for convergence check. The output of the EM algorithm, if not trapped in a local mode, is the maximum-likelihood estimate (MLE), which possesses well-established statistical properties. However, the capability of most EM-based approaches is restricted to approximately one dozen loci, because of the memory constraint. A recently developed program, SNPHAP (see David Clayton's Web site [SNPHAP: A Program for Estimating Frequencies of Large Haplotypes of SNPs]), is an exception that, although different from the PL strategy, can handle many more linked loci by using a progressive-extension technique.

The essential steps of the PL strategy (Niu et al. 2002) are as follows: One first breaks down all of the marker loci into stretches of "atomistic" units and then uses either the EM algorithm or the Gibbs sampler to construct haplotypes for each unit and to rebuild the phase hierarchically, through a bottom-up approach. For example, an individual represented in the lipoprotein lipase (LPL) gene SNP data set (Nickerson et al. 1998) has the genotype (01200001000000000100010), where 0 stands for heterozygote and 1 and 2 stand for wild-type and mutant homozygotes, respectively. Since there are 18 heterozygous loci, the standard EM algorithm has to consider 2^{18} possible haplotypes, making it extremely costly for haplotype estimation. Using the PL strategy, we divide the linked loci into four "atomistic" units—(012000), (010000), (000001), and (00010)—and use the EM algorithm to estimate partial haplotypes within each unit. Afterward, two adjacent partial haplotypes are "ligated" by using the EM algorithm again, just like phasing two linked multiallelic markers. The ligation process is repeated until the complete phase is determined.

It is well known that the EM algorithm can be trapped in a local mode. This problem becomes a more serious issue for the PL-EM strategy, because every atomistic haplotype construction or ligation step involves a complete EM algorithm implementation. A naive implementation of the ligation step considers only the partial haplotypes that have nonzero estimated frequencies in the previous EM step. However, it appears that one phase configuration (and the corresponding haplotypes with nonzero estimated frequencies) is more likely when looking only at a partial set of loci, whereas a different configuration is more likely when all loci are taken into consideration. For example, consider the set of individuals with the following genotype data on four loci—(A/A A/A T/T T/T), (A/A A/A T/T T/T), (A/A G/G T/T T/T), (A/A G/G C/C C/C), (A/A G/G C/C C/C), and (A/G A/G T/T T/T). If just the first two loci are concerned, then the EM algorithm estimates the haplotype frequencies as 7/12, 4/12, and 1/12, for (AG), (AA), and (GA), respectively. When all four loci are considered together, however, the EM gives rise to four haplotypes—(AATT), (AGCC), (AGTT), and (GGTT), with frequencies 5/12, 4/12, 2/12, and 1/12, respectively. Thus, had we thrown away the (GG) haplotype prematurely when only the first two SNP markers were analyzed, we would have not been able to reach the MLE.

To overcome this difficulty, we devised a "backup-buffering" strategy during the ligation step. In brief, in addition to keeping in a buffer those partial haplotypes that have EM-algorithm–estimated frequencies greater than a threshold value (e.g., $\epsilon = 10^{-5}$), we also retain in the buffer some partial haplotypes whose estimated frequencies are below ϵ. The criterion for choosing such a backup partial haplotype is based on the rank of its average estimated frequency over all the EM iterations. The buffer size—that is, the total number of candidate partial haplotypes in a buffer—is kept as a constant in the PL process. Not surprisingly, our simulation study based on the cystic fibrosis data showed that, the larger the buffer size is, the more accurate the phasing results are (for details, see fig. A1 [online only, at J. S. Liu's Web site]).

Niu et al. (2002) observed a modest performance improvement when recombination hotspots were used as the partition sites. Recently, hotspot-detection algorithms, such as a greedy algorithm (Patil et al. 2001) and a dynamic programming approach (Zhang et al. 2002), have been developed. Our PL-EM program can incorporate the information revealed by such algorithms by allowing the user to specify desirable partition points (for details and download of the PL-EM program, see J. S. Liu's Web site). We also conducted an empirical study on the effects that different partition sizes, K, have when hotspot information is absent. Although little difference in phasing performance was observed when three different partition sizes were used—3–4, 5–8, or 9–16 (see fig. A2 [online only, at J. S. Liu's Web site])—we found that the computation time increased sharply when the coarsest partition was used. Overall, $K = 5$–8 appeared to be a good choice for the atomistic unit size.

Several EM-based algorithms—including HAPLO (Hawley and Kidd 1995), Arlequin (Schneider et al. 2000), and the Mx program (Neale et al. 1999)—provide the variance estimates for the estimated haplotype frequencies. However, since these methods handle no more than ~20 loci, their variance-estimation method cannot be directly used by the PL-EM program. Instead, we implemented with the PL-EM program a simple and robust approach, to estimate the variances or SEs of the frequencies of those haplotypes that were selected at the final ligation stage.

Let Y be the observed genotype data, Z be the missing phase information, and $\boldsymbol{\theta}$ be the vector of haplotype frequencies. As noted by Louis (1982), the Hessian matrix of $\boldsymbol{\theta}$ can be computed via an identity analogous to the variance-decomposition rule,

$$-\frac{\partial^2 \log p(\boldsymbol{\theta}\mid Y)}{\partial\boldsymbol{\theta}^2} = E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 \log[p(\boldsymbol{\theta}\mid Y,Z)]}{\partial\boldsymbol{\theta}^2}\Big| Y\right\} -\mathrm{var}_{\boldsymbol{\theta}}\left\{\frac{\partial \log[p(\boldsymbol{\theta}\mid Y,Z)]}{\partial\boldsymbol{\theta}}\Big| Y\right\}\ , \tag{1}$$

and the variance-covariance matrix of the MLE, $\hat{\boldsymbol{\theta}}$, is the inverse of this matrix evaluated at $\hat{\boldsymbol{\theta}}$. The first term on the right-hand side of equation (1) can be computed as

$$\left(E\left\{-\frac{\partial^2 \log[p(\boldsymbol{\theta}\mid Y,Z)]}{\partial\boldsymbol{\theta}^2}\Big| Y\right\}\right)_{i,j} = \begin{cases} \dfrac{E(n_i\mid Y)}{\theta_i^2} + \dfrac{E(n_m\mid Y)}{(1-\theta_1-\cdots-\theta_{m-1})^2} & \text{if } i=j<m \\ \dfrac{E(n_m\mid Y)}{(1-\theta_1-\cdots-\theta_{m-1})^2} & \text{if } i<j<m \end{cases}\ ,$$

where m is the number of all candidate haplotypes, n_i is the number of occurrences of haplotype i in Z, and the expectation is taken for the n_i (which is a function of Z) with $\boldsymbol{\theta}$ fixed at the MLE. The second term on the right-hand side needs the variance-covariance matrix of

$$\frac{\partial \log p(\boldsymbol{\theta}\mid Y,Z)}{\partial\boldsymbol{\theta}} = \left(\frac{n_1}{\theta_1}-\frac{n_m}{\theta_m},\frac{n_2}{\theta_2}-\frac{n_m}{\theta_m},\cdots,\frac{n_{m-1}}{\theta_{m-1}}-\frac{n_m}{\theta_m}\right)\ .$$

The calculation of $\mathrm{cov}(n_i,n_j)$, for example, can be achieved by observing in each individual the probability of the joint occurrence of haplotypes i and j.

In the presence of many heterozygous loci, some rare haplotypes with very low frequencies are likely to occur. Then, the inversion of the Hessian matrix becomes computationally burdensome and numerically unstable. Since scientists are mostly concerned with the variance of each $\hat{\theta}_i$ instead of covariances among the $\hat{\theta}_i$s, we introduce a new, robust method of computing these marginal variances. Take $\mathrm{var}(\hat{\theta}_1)$, for example: by applying equation (1) to a reparameterization of the model with $\theta' = (\theta_1, 1 - \theta_1)$ and $\theta'' = (\theta_2, \ldots, \theta_m)/(1 - \theta_1)$, we have

$$-\frac{\partial^2 \log p(\theta_1 | Y)}{\partial \theta_1^2}\Big|_{\theta_1 = \hat{\theta}_1}$$

$$= E_{\hat{\theta}}\left(\frac{n_1}{\hat{\theta}_1^2} + \frac{2n - n_1}{(1 - \hat{\theta}_1)^2}\right) - \mathrm{var}_{\hat{\theta}}\left(\frac{n_1}{\hat{\theta}_1} - \frac{2n - n_1}{1 - \hat{\theta}_1}\Big| Y\right)$$

$$= \frac{2n}{\hat{\theta}_1(1 - \hat{\theta}_1)} - \frac{\mathrm{var}_{\hat{\theta}}(n_1)}{\hat{\theta}_1^2(1 - \hat{\theta}_1)^2} \ . \qquad (2)$$

Thus, $\mathrm{var}(\hat{\theta}_1)$ is equal to the reciprocal of the above quantity. Note that the new method and Louis's method give identical variance estimates if the inversion of the Hessian matrix (eq. [1]) is accurate. Intuitively, the first term on the right-hand side of equation (2) is the standard variance estimate when there is no uncertainty in phasing, and the second term accounts for the loss of information because of unknown phases.

An example of the SE calculation for estimated haplotype frequencies is shown, in table 1, for the LPL data from Nickerson et al. (1998). This example also illustrates that haplotypes can shed new light on population migration and admixture. To better understand the properties of the estimated SEs, we conducted a simulation study using the 12 distinct haplotypes from the β_2-adrenergic receptor (β_2AR) data set. Assuming that the 12 haplotypes have equal frequencies (1/12), we simulated 100 data sets, each consisting of 90 hypothetical individuals. The PL-EM algorithm was applied to each of the data sets, and a 95% CI for each $\hat{\boldsymbol{\theta}}$ was constructed on the basis of the estimated frequencies and SEs (i.e., $\hat{\boldsymbol{\theta}} \pm 1.96\hat{\sigma}$). The number of times (in 100 trials) that the 95% CI covered the true frequency ($\boldsymbol{\theta} = 1/12$) for the 12 haplotypes was 92, 88, 93, 96, 97, 96, 88, 93, 94, 92, 95, and 94, which average to 93.2%. For the purpose of calibration, we note that the average coverage of the true $\boldsymbol{\theta}$ was only 93.1% when the haplotype phase information was given.

The presence of a significant portion of missing genotypes is a common problem when a great number of linked loci are under investigation. This missing-data problem poses a serious challenge to the existing EM haplotype-inference algorithms, even when the total number of SNP loci is moderate. In the case of missing two allele calls at one locus, for example, all three different genotype configurations—(AA), (Aa), and (aa)—have to be accounted for by the algorithm, which greatly inflates the space of candidate haplotypes. As a consequence, the standard EM algorithm not only needs a lot more memory but also converges much more slowly. The PL-EM algorithm resolves this difficulty seamlessly because of its adoption of the divide-conquer-combine strategy.

It often occurs that, for some individuals with a large number of heterozygous loci, numerous haplotype pairs (each with a nonzero probability) are compatible with their genotype data. In this case, generating all compatible haplotype phases with nontrivial probabilities is more desirable than outputting only the best phase. There is some evidence (X. Lu and J. S. Liu, unpublished data) showing

Table 1

Application of PL-EM on the LPL Data

		Results from		
Haplotype	ID	Jackson, MS ($N = 24$; $k = 28$)	North Karelia, Finland ($N = 24$; $k = 20$)	Rochester, MN ($N = 23$; $k = 22$)
01000001000000000100000	H1	.063 (.035)	.219 (.059)	.348 (.069)
01000001100000000100000	H2	…	.073 (.039)	.045 (.031)
00100110100000000000000	H3	.146 (.051)	…	.043 (.030)
00100110100000010001100	H4	.104 (.044)	…	…
00100110000000010000000	H5	.063 (.035)	…	…
10101000011111110011101	H6	.063 (.035)	…	…
00100001000000000100000	H7	…	.146 (.051)	.023 (.023)
01000110000000000000000	H8	.021 (.021)	.125 (.048)	…
00100000111111100011100	H9	…	.083 (.040)	…
10111100111111111011111	H10	…	…	.087 (.042)

Note.—The LPL data are based on a study by Nickerson et al. (1998). A total of 88 sites in the 7.9-kb region have been reported among the 71 individuals. Of these 88 biallelic markers, 23 met the following two criteria: (1) minor-allele frequency >20% and (2) marker missing data <2%. Both PL-EM and HAPLOTYPER were applied, to phase the entire 71 subjects by using only these 23 markers. N and k represent the sample size and the number of distinct haplotypes, respectively. Numbers shown in parentheses represent SEs of the frequency estimates. PL-EM appears to output almost the same number of haplotypes as does HAPLOTYPER ($k = 28$, $k = 20$, and $k = 22$ vs. $k = 28$, $k = 19$, and $k = 22$, for the Jackson, North Karelia, and Rochester samples, respectively). The number of distinct haplotypes is greatest in the Jackson sample (African Americans) and is smallest in the North Karelia sample (white Europeans). The Rochester sample shares H1 and H3 with the Jackson sample and shares H1, H2, and H7 with the North Karelia sample, indicating that this American-white population may be the result of admixture between black and European-white populations.

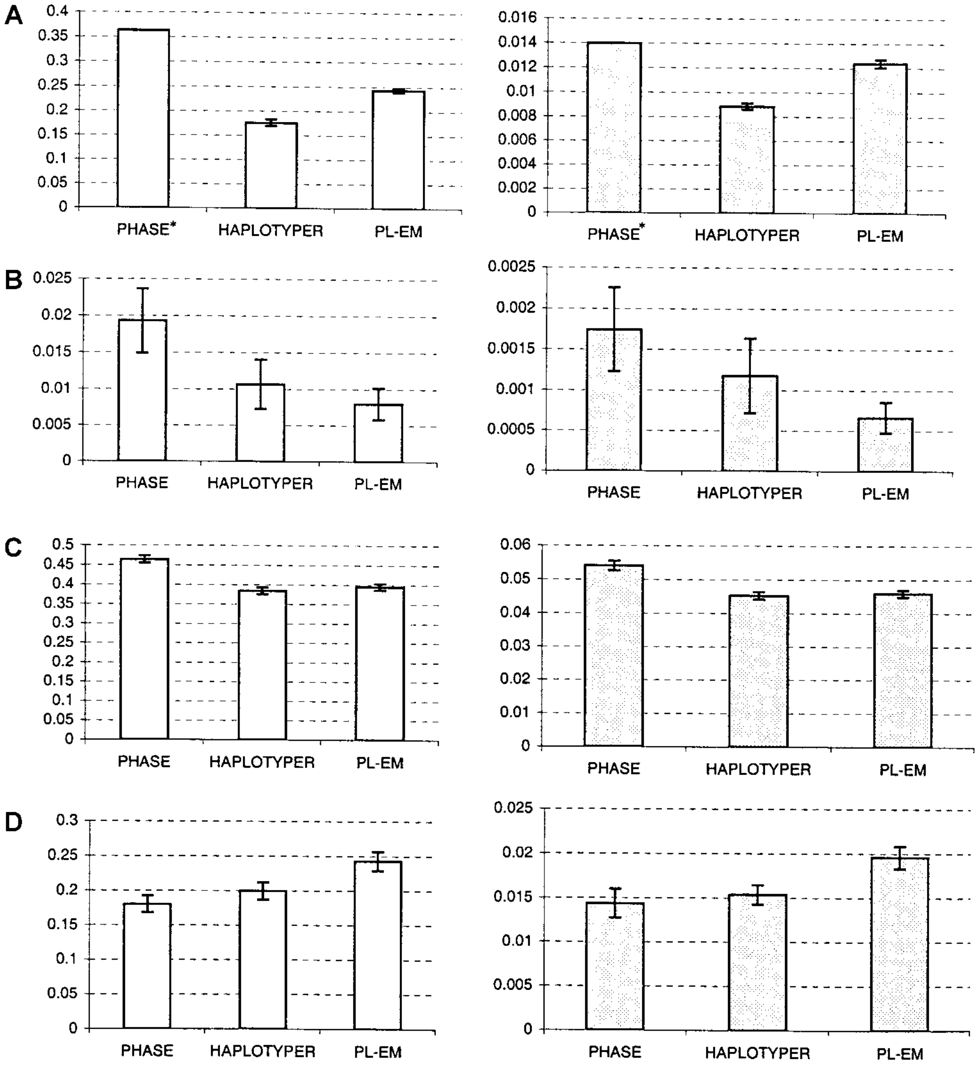

Figure 1 Histograms of the average error rates based on either individual phase calls (*open bars*) or the proportion of incorrectly inferred loci (*shaded bars*), for ACE (*A*), β_2AR (*B*), CFTR (*C*), and coalescence-simulation (*D*) data. For the ACE data, there are a total of 52 biallelic markers for 11 subjects (Rieder et al. 1999), and 100 independent runs for each algorithm were performed. For the β_2AR data, 15 haplotype pairs (each pair corresponding to one subject) were randomly drawn from a total of 10 distinct haplotypes according to their respective frequencies, as shown by Drysdale et al. (2000); this procedure was repeated to generate a total of 100 simulated data sets. For the CFTR data, the 100 data sets were generated by randomly pairing 56 of the 57 complete haplotypes of the 23 linked SNPs in a 1.8-Mb region near the CFTR gene provided by Kerem et al. (1989). The coalescence simulation was done using the Long Lab's algorithm. A total of 100 replications were performed for a regional size of 10 U of 4*Nc,* each of which consisted of 20 pairs of unphased chromosomes with 20 linked SNP loci. The error bars are shown as ±1 SE, for the new version of PHASE, HAPLOTYPER, and PL-EM. An asterisk (*) indicates that the old version of PHASE was used for this data set, because its performance is better than that of the new version.

that, by accounting for the phasing uncertainty, one can gain accuracy in LD mapping when using the algorithm BLADE (Liu et al. 2001; this algorithm employs a semi-hidden Markov model and a Markov-chain Monte Carlo method, for inference of the location of the disease mutation among a given set of linked markers with known genetic distances in a case-control setting). To accommodate this need of multiple-haplotype imputation, the PL-EM program can let the user choose to display either the top *f* most likely phases (if existing) for each individual or all phases with probabilities >0.1.

We evaluated the performances of PL-EM, HAPLOTYPER (Niu et al. 2002), and an enhanced version of PHASE (Stephens et al. 2001), using the angiotensin I–converting enzyme (ACE) data set, the β_2AR gene data set, the cystic fibrosis transmembrane conductance regulator (CFTR) gene data set, and data sets produced by coalescence model–based haplotype-simulation software (see the Long Lab's Web site [Tools: Statistical Analysis and Molecular Biology Tools]). All these data sets were constructed in the same way, as described by Niu et al. (2002). The results are summarized in the left panels of figure 1. The PL-EM program's error rate for individuals' phasing is comparable to HAPLOTYPER, but is lower than PHASE in the first three cases, which is consistent with the studies described by Niu et al. (2002). For the coalescence simulation, PL-EM and HAPLOTYPER respectively made 35% and 11% more errors than PHASE. Note that Stephens et al. (2001) reported that the EM algorithm made ~100% more errors than PHASE, indicating that PL-EM performed significantly better than the standard EM algorithm when the coalescence assumption is appropriate.

To investigate further how the inference errors were made by the three algorithms, we looked into the following two aspects: (1) how the incorrectly inferred haplotypes differ from the true ones and (2) whether different algorithms made errors on the same individuals. For the first three data sets, PL-EM appeared to produce the least amount of incorrectly inferred loci for those wrongly inferred haplotypes, whereas, for the coalescent-based simulated data, PL-EM and HAPLOTYPER respectively produced 36% and 7% more incorrectly inferred loci than did PHASE (fig. 1, *right panels*). In the first three cases, most of the errors made by HAPLOTYPER and PL-EM appeared to be a subset of the errors made by PHASE (see fig. A3 [online only, at J. S. Liu's Web site]).

In summary, the PL-EM algorithm can deal with a large number of linked loci that have moderate levels of LD. It is capable of variance estimation, multiple imputation, and the handling of incomplete genotype data. In addition, PL-EM was faster than HAPLOTYPER in these examples, even with the variance estimation. Hence, in practice, if a coalescence model for the population haplotypes is too strong to assume, then PL-EM can be an attractive alternative to HAPLOTYPER, further helping scientists in the haplotype-reconstruction endeavor.

Acknowledgments

We are grateful to Chi-Hse Teng and the two anonymous reviewers for insightful comments. This research was supported in part by the National Science Foundation grants DMS-0094613 and DMS-0104129 and National Institutes of Health grant R01 HG02518-01.

ZHAOHUI S. QIN,[1,*] TIANHUA NIU,[2,*]
AND JUN S. LIU[1]
[1]*Department of Statistics, Harvard University, Cambridge, MA; and* [2]*Program for Population Genetics, Harvard School of Public Health, Boston*

Electronic-Database Information

URLs for data presented herein are as follows:

J. S. Liu's Web Site, http://www.people.fas.harvard.edu/~junliu/plem/ (for supplemental figs. A1–A3 and detailed documentation and download instructions for the PL-EM algorithm)

SNPHAP: A Program for Estimating Frequencies of Large Haplotypes of SNPs, http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt

Tools: Statistical Analysis and Molecular Biology Tools, http://hjmuller.bio.uci.edu/~labhome/coalescent.html (for coalescence model–based haplotype-simulation software)

References

Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K, Arnold K, Ruano G, Liggett SB (2000) Complex promoter and coding region β^2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. Proc Natl Acad Sci USA 97:10483–10488

Goldstein DB (2001) Islands of linkage disequilibrium. Nat Genet 29:109–211

Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409–411

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, et al (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080

Liu JS, Sabatti C, Teng J, Keats BJ, Risch N (2001) Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res 11:1716–1724

Louis TA (1982) Finding the observed information matrix when using the EM algorithm. J R Stat Soc B 4:226–233

Neale MC, Boker S, Xie G, Maes H (1999) Mx: Statistical modeling. Department of Psychiatry, Medical College of Virginia, Richmond

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengård J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat Genet 19:233–240

Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157–169

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, et al (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723

Rieder MJ, Taylor SL, Clark AG, Nickerson DA (1999) Sequence variation in the human angiotensin converting enzyme. Nat Genet 22:59–62

Schneider S, Roessli D, Excoffier L (2000) Arlequin: a software for population genetics data analysis. Ver 2.000. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Geneva

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 3:391–397

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK (2000) The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 67:518–522

Zhang K, Deng M, Chen T, Waterman MS, Sun F (2002) A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci USA 99:7335–7339

Address for correspondence and reprints: Dr. Jun S. Liu, Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138. E-mail: jliu@stat.harvard.edu

* The first two authors contributed equally to this work.

0002-9297/2002/7105-0031$15.00

Am. J. Hum. Genet. 71:1247–1250, 2002

Alcohol Dehydrogenase and Alcohol Dependence: Variation in Genotype-Associated Risk between Populations

To the Editor:

Osier et al. (2002) report that haplotyping of the alcohol dehydrogenase (*ADH*) gene cluster at 4q21-23 showed unusually high values for F_{st}, an estimator of population differentiation. This was largely due to differences between populations in East Asia and those in other areas of the world. The finding was discussed in relation to the origin and maintenance of the distinct East Asian haplotype and in relation to possible association between genetic variation at this locus and the risk of alcohol dependence (MIM 103780). This letter draws attention to a potentially related difference between populations, in the magnitude of the alcohol dependence risk associated with the *ADH1B* (MIM 103720) Arg47His polymorphism (previously referred to as "*ADH2*2*"). One possible explanation for such a difference in risk is the presence of linkage disequilibrium between this marker and an undiscovered causative polymorphism, with the effect being stronger in East Asians and the relative risk associated with *ADH1B* Arg47His variation consequently being greater.

To update a previous meta-analysis of the effects of *ADH* polymorphisms (Whitfield 1997), articles reporting on *ADH1B* genotypes in control and alcohol-dependent subjects were identified by Medline search or from knowledge of data in conference proceedings, with elimination of articles in which subjects overlapped. Data from eight of the articles previously analyzed (all those listed in table 1 and published before 1997) and from nine new articles, were included. Information on *ADH1B* Arg47His genotypes in control and alcohol-dependent subjects was extracted. Data on alcohol-dependent subjects with known liver disease were excluded, because of the possibility that *ADH1B* variation may affect the risk of liver damage in alcoholics. Odds ratios were calculated from stratified 2 × 2 tables, using StatXact 5 (Cytel Software), with tests for heterogeneity across studies and estimation of common odds ratios. Whenever possible, two 2 × 2 tables were compiled from each article: one for the *ADH1B**47Arg/*47Arg versus *ADH1B**47Arg/*47His genotype comparison and the second for comparison of *ADH1B**47Arg/*47His against *ADH1B**47His/*47His.

Data from each article, exact odds ratios, and their 95% CIs are shown in table 1. For the *ADH1B**47Arg/*47Arg versus *ADH1B**47Arg/*47His (*ADH2*1/*1* versus *ADH2*1/*2*) comparison, there was significant heterogeneity of odds ratios across all the studies ($P < .0001$). Division of studies into those from Europe (including Russia and Australia) and those from Asia, with separate analyses for the two groups, showed no evidence of within-group heterogeneity among Europeans ($P = .397$), and the estimated common odds ratio was 2.11 (95% CI 1.32–3.44). However, there was still significant heterogeneity ($P < .0001$) among Asian studies. Inspection of the data suggested that results from Japanese and from Han Chinese groups were similar, whereas the minority ethnic groups within China, as well as Koreans, had lower odds ratios. As can be seen in table 1, the Han Chinese and the Japanese groups had very similar common odds ratios associated with *ADH1B**47Arg/*47Arg compared with *ADH1B**47Arg/*47His, which were substantially above those for Europeans and most of the other Asian groups.

The calculated odds ratios for *ADH1B**47Arg/*47His against ADH1B*47His/*47His (*ADH2*1/*2* versus **2/*2*) are also shown in table 1. There was no significant heterogeneity between studies ($P = .405$), and the estimated common odds ratio was 1.43 (95% CI 1.23–1.66). The difference in alcohol-dependence risk is therefore greater for *ADH1B**47Arg/*47Arg versus *ADH1B**47Arg/*47His than for *ADH1B**47Arg/*47His versus *ADH1B**47His/*47His, at least in the mainly East Asian populations in which the *ADH1B**47His allele frequency is high enough to allow a meaningful comparison.

Two conclusions may be drawn from this summary of published results. First, the *ADH1B**47His allelic effects on alcohol dependence risk are not additive. Heterozygotes are clearly more similar in risk to the *ADH1B**47His/*47His homozygotes than to the *ADH1B**47Arg/*47Arg homozygotes, and so the *ADH1B**47His allele shows quantitative (but not complete) dominance. Proposed mechanisms for the *ADH1B* Arg47His effect on dependence need to account for this

Table 1

Calculated Odds Ratios and Associated 95% CI for Alcohol Dependence by *ADH1B* Genotype

Population, Reference,	Control Subjects			Alcoholics			RR vs. RH		RH vs. HH	
and Sources of Subjects	RR	RH	HH	RR	RH	HH	OR	95% CI	OR	95% CI
Europeans:										
Gilder et al. 1993:										
England[a]	77	7	0	76	6	0	1.15	.32–4.35	NA	NA
Espinos et al. 1997:										
Spain	58	12	1	62	9	0	1.42	.51–4.13	NA	NA
Whitfield et al. 1998:										
Australia[b,c]	101	18	0	36	1	0	6.37	.94–274.60	NA	NA
Borras et al. 2000:										
France, Germany, Poland, Spain, Sweden[d]	214	10	0	226	5	0	2.11	.64–7.99	NA	NA
Ogurtsov et al. 2001:										
Russia (Moscow)[d]	15	29	6	24	12	1	3.80	1.39–10.94	2.44	.25–123.4
Frenzer et al. 2002:										
Australia [c,d]	184	14	2	54	3	0	1.37	.36–7.70	NA	NA
Common OR (all Europeans)							2.11	1.32–3.44	NA	NA
Asians:										
Thomasson et al. 1994:										
China (Atayal, Taiwan)	1	10	54	3	28	63	1.07	.08–61.93	2.39	1.01–6.03
Muramatsu et al. 1995:										
China (Han Chinese, Shanghai)	12	43	50	13	8	11	5.66	1.72–20.00	.85	.27–2.56
Chen et al. 1996:										
China (Han Chinese, Taipei)	0	19	44	14	15	17	NA	NA	2.03	.77–5.37
Shen et al. 1997:										
China (Han)	6	19	23	10	25	17	1.26	.34–5.03	1.77	.69–4.63
Korean	3	23	24	9	17	29	3.95	.82–26.11	.62	.25–1.51
Mongolian	6	14	15	11	15	5	1.69	.43–7.20	3.14	.80–14.10
Elunchun	12	22	3	13	15	3	1.58	.51–4.99	.69	.08–5.85
Osier et al. 1999:										
Taipei										
Han	6	56	73	40	39	49	9.42	3.52–29.86	1.04	.58–1.86
Ami	1	5	14	3	6	11	2.36	.13–156.6	1.51	.29–8.14
Atayal	0	7	13	0	6	15	NA	NA	.75	.16–3.38
Yin and Agarwal 2001:										
China (Han Chinese, Taipei)[e]	54	242	361	152	130	137	5.22	3.54–7.79	1.42	1.05–1.91
Higuchi 1994:										
Japan	31	152	247	204	224	227	4.46	2.86–7.10	1.60	1.21–2.13
Maezawa et al. 1995:										
Japan	2	22	36	30	28	38	11.48	2.45–109.90	1.20	.55–2.64
Nakamura et al. 1996:										
Japan	3	54	40	21	20	12	18.25	4.73–106.00	1.23	.50–3.11
Tanaka et al. 1996:										
Japan[d]	4	24	38	27	42	21	3.81	1.13–16.77	3.14	1.43–7.04
Lee et al. 2001:										
Seoul, Korea[d]	6	18	40	3	21	28	.44	.06–2.40	1.66	.70–3.98
Common OR:										
Han Chinese							5.19	3.74–7.26	1.36	1.07–1.72
Japanese							5.50	3.75–8.22	1.70	1.34–2.17

Note.—RR = *ADH1B**47Arg/*47Arg; RH = *ADH1B**47Arg/*47His; HH = *ADH1B**47His/*47His, OR = odds ratio, NA = Not applicable (odds ratio could not be calculated because of empty cells).

[a] U.K. or Irish descent.

[b] Men only.

[c] Australians of European descent.

[d] Control subjects versus alcoholics without alcoholic cirrhosis or pancreatitis.

[e] *ALDH2**11 and *12 subjects only.

feature. It is worth pointing out that a study that measured hepatic ADH activity and *ADH1B* genotype in human livers found that activity at pH 7.5 was approximately fivefold higher in *ADH1B**47Arg/*47His subjects and was only sixfold higher in *ADH1B**47His/*47His subjects than in those with the *ADH1B**47Arg/*47Arg genotype (Yao et al. 1997). It is not clear whether these two examples of nonadditive effects of this polymorphism are related.

Second, there was a notable difference between European and Chinese or Japanese risk estimates. At least two types of explanation for heterogeneity between populations in the relative risk conferred by, or associated with, a genetic polymorphism should be considered: genetic and social. If the polymorphism is not itself causative, then linkage disequilibrium with a causative locus will decrease with the passage of time after the original mutation event and may remain stronger in one group than in another. Alternatively, the same neutral polymorphism may have arisen independently in the two populations and may be in linkage disequilibrium with the causative polymorphism in only one. It will be seen from table 4 in the article by Osier et al. (2002) that the *ADH1B**47His allele occurs on a different haplotype background in East Asians (mainly 221221) and the European/Middle Eastern/European North American groups (mainly 221211, or 212211 in some Samaritans). Although this does not demonstrate independent mutations, it does suggest that the origin of *ADH1B* Arg47His is not recent and that changes have occurred in the nearby sequence.

It has generally been assumed that the *ADH1B* Arg47His polymorphism is causative and that the effect arises from the difference in $V_{\max}$ for ethanol (Bosron and Li 1986) between the enzymes produced. However, there are problems in extrapolating this in vitro activity difference to alcohol metabolism in vivo, and as Osier et al. (1999, 2002) discuss, another causative polymorphism within the *ADH* region cannot be excluded.

On the other hand, social factors or other unlinked genetic effects may modify the *ADH1B* Arg47His effect in the comparatively few Europeans who have the *ADH1B**47Arg/*47His or *ADH1B**47His/*47His genotypes, so the genotype-associated difference in risk is smaller. There is evidence (Higuchi et al. 1994) that the size of the protective effect associated with aldehyde dehydrogenase (ALDH2) deficiency has changed during the past 20 years in Japan—a period that, although it is far too short for genetic changes, has been a time of substantial alterations in the social environment. Lee et al. (2001) also comment on the social pressures to drink in Korea. Gene-environment interaction therefore presents an alternative explanation for the heterogeneity between populations.

We cannot yet determine whether social factors or variations in linkage disequilibrium are responsible for the difference in *ADH1B* Arg47His effects between Europeans and two major Asian groups. The question may be resolved by haplotype data across the *ADH* region in alcoholics and control subjects from different countries or regions, or by studies of alcoholics and control subjects of Asian descent living in European societies.

John B. Whitfield
Department of Clinical Biochemistry
Royal Prince Alfred Hospital
Sydney
Australia

Electronic-Database Information

Accession numbers and the URL for data presented herein are as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for ADH1B [MIM 103720], alcoholism [MIM 103780], and ALDH2 [MIM 100650]

References

Borras E, Coutelle C, Rosell A, Fernandez-Muixi F, Broch M, Crosas B, Hjelmqvist L, Lorenzo A, Gutierrez C (2000) Genetic polymorphism of alcohol dehydrogenase in Europeans: the ADH2*2 allele decreases the risk for alcoholism and is associated with ADH3*1. Hepatology 31:984–989

Bosron WF, Li TK (1986) Genetic polymorphism of human liver alcohol and aldehyde dehydrogenases, and their relationship to alcohol metabolism and alcoholism. Hepatology 6:502–510

Chen WJ, Loh EW, Hsu YP, Chen CC, Yu JM, Cheng AT (1996) Alcohol-metabolising genes and alcoholism among Taiwanese Han men: independent effect of ADH2, ADH3 and ALDH2. Br J Psychiatry 168:762–767

Espinos C, Sanchez F, Ramirez C, Juan F, Najera C (1997) Polymorphism of alcohol dehydrogenase genes in alcoholic and nonalcoholic individuals from Valencia (Spain). Hereditas 126:247–253

Frenzer A, Butler WJ, Norton ID, Wilson JS, Apte MV, Pirola RC, Ryan P, Roberts-Thomson IC (2002) Polymorphism in alcohol-metabolizing enzymes, glutathione S-transferases and apolipoprotein E and susceptibility to alcohol-induced cirrhosis and chronic pancreatitis. J Gastroenterol Hepatol 17:177–182

Gilder FJ, Hodgkinson S, Murray RM (1993) ADH and ALDH genotype profiles in Caucasians with alcohol-related problems and controls. Addiction 88:383–388

Higuchi S (1994) Polymorphisms of ethanol metabolizing enzyme genes and alcoholism. Alcohol Alcohol Suppl 2:29–34

Higuchi S, Matsushita S, Imazeki H, Kinoshita T, Takagi S, Kono H (1994) Aldehyde dehydrogenase genotypes in Japanese alcoholics. Lancet 343:741–742

Lee HC, Lee HS, Jung SH, Yi SY, Jung HK, Yoon JH, Kim CY (2001) Association between polymorphisms of ethanol-metabolizing enzymes and susceptibility to alcoholic cirrhosis in a Korean male population. J Korean Med Sci 16: 745–750

Maezawa Y, Yamauchi M, Toda G, Suzuki H, Sakurai S (1995) Alcohol-metabolizing enzyme polymorphisms and alcoholism in Japan. Alcohol Clin Exp Res 19:951–954

Muramatsu T, Wang ZC, Fang YR, Hu KB, Yan H, Yamada K, Higuchi S, Harada S, Kono H (1995) Alcohol and aldehyde dehydrogenase genotypes and drinking behavior of Chinese living in Shanghai. Hum Genet 96:151–154

Nakamura K, Iwahashi K, Matsuo Y, Miyatake R, Ichikawa Y, Suwaki H (1996) Characteristics of Japanese alcoholics with the atypical aldehyde dehydrogenase 2*2. I. A comparison of the genotypes of ALDH2, ADH2, ADH3, and cytochrome P-4502E1 between alcoholics and nonalcoholics. Alcohol Clin Exp Res 20:52–55

Ogurtsov PP, Garmash IV, Miandina GI, Guschin AE, Itkes AV, Moiseev VS (2001) Alcohol dehydrogenase ADH2-1 and ADH2-2 allelic isoforms in the Russian population correlate with type of alcoholic disease. Addict Biol 6:377–383

Osier M, Pakstis AJ, Kidd JR, Lee JF, Yin SJ, Ko HC, Edenberg

HJ, Lu RB, Kidd KK (1999) Linkage disequilibrium at the ADH2 and ADH3 loci and risk of alcoholism. Am J Hum Genet 64:1147–1157

Osier MV, Pakstis AJ, Soodyall H, Comas D, Goldman D, Odunsi A, Okonofua F, Parnas J, Schulz LO, Bertranpetit J, Bonne-Tamir B, Lu RB, Kidd JR, Kidd KK (2002) A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. Am J Hum Genet 71:84–99

Shen YC, Fan JH, Edenberg HJ, Li TK, Cui YH, Wang YF, Tian CH, Zhou CF, Zhou RL, Wang J, Zhao ZL, Xia GY (1997) Polymorphism of ADH and ALDH genes among four ethnic groups in China and effects upon the risk for alcoholism. Alcohol Clin Exp Res 21:1272–1277

Tanaka F, Shiratori Y, Yokosuka O, Imazeki F, Tsukada Y, Omata M (1996) High incidence of ADH2*1/ALDH2*1 genes among Japanese alcohol dependents and patients with alcoholic liver disease. Hepatology 23:234–239

Thomasson HR, Crabb DW, Edenberg HJ, Li TK, Hwu HG, Chen CC, Yeh EK, Yin SJ (1994) Low frequency of the ADH2*2 allele among Atayal natives of Taiwan with alcohol use disorders. Alcohol Clin Exp Res 18:640–643

Whitfield JB (1997) Meta-analysis of the effects of alcohol dehydrogenase genotype on alcohol dependence and alcoholic liver disease. Alcohol Alcohol 32:613–619

Whitfield JB, Nightingale BN, Bucholz KK, Madden PA, Heath AC, Martin NG (1998) ADH genotypes and alcohol use and dependence in Europeans. Alcohol Clin Exp Res 22: 1463–1469

Yao CT, Liao CS, Yin SJ (1997) Human hepatic alcohol and aldehyde dehydrogenases: genetic polymorphism and activities. Proc Natl Sci Counc Repub China B 21:106–111

Yin SJ, Agarwal DP (2001) Functional polymorphism of alcohol and aldehyde dehydrogenases. In: Agarwal DP, Seitz HK (eds) Alcohol in health and disease. Marcel Dekker, New York, pp 1–26

Address for correspondence and reprints: Dr. John B Whitfield, Department of Clinical Biochemistry, Royal Prince Alfred Hospital, Camperdown NSW 2050, Australia. E-mail: John.Whitfield@email.cs.nsw.gov.au

0002-9297/2002/7105-0032$15.00

Am. J. Hum. Genet. 71:1250–1251, 2002

Reply to Whitfield

To the Editor:

Dr. John B. Whitfield (Whitfield 2002 [in this issue]) writes to call attention to the variation in alcohol-dependence risk as a function of both the *ADH1B* Arg47His polymorphism *and* specific populations. We do not disagree with his result: our study (Osier et al. 2002) involved, primarily, normal individuals from multiple populations and showed considerable variation in allele frequencies at that site among the populations. In that report, we commented that the different risks of alcoholism associated with alleles at this site could be explained by other relevant variation on the specific haplotype at high frequency in eastern Asia. However, we did not show that haplotyping of the ADH Class I polymorphisms resulted in an unusually high F_{st} but that some individual sites in the gene cluster individually had unusually high F_{st} values. One particular haplotype does have a very large range of variation, but we do not have an appropriate empiric distribution for F_{st} values in multiallelic haplotype systems to show that it is unusually large. We showed that the "protective" allele, *ADH1B*47His,* occurs primarily on a specific haplotype in the Mediterranean and European populations studied but occurs on a different haplotype in eastern Asia. We concluded that the *ADH1B*47His* allele is likely to be old, a conclusion Whitfield reiterates.

We do disagree with some of the conclusions Whitfield reaches. He concludes that the effects of the *ADH1B*47His* allele are not additive. However, because of the strong linkage disequilibrium (LD) across the Class I gene cluster, Whitfield's analysis showing nonadditive allelic effects uses the *ADH1B*47His* allele as a surrogate for the entire haplotype. Those analyses do not provide sufficient evidence to limit the effect to just that allele; some other variant on that haplotype could be relevant. Taken together, the evidence that the risk difference associated with this polymorphism is not the same in Europe as it is in eastern Asia and our demonstration that the haplotype containing the *ADH1B*47His* allele is different in Europe from the one in eastern Asia require us to focus on the entire haplotype and not just on this one site. The nonadditive effects cannot be attributed to the *ADH1B*47His* allele exclusively, as Whitfield himself notes earlier in his introductory paragraph.

Whitfield repeats a common error when he says LD will decrease over time, without noting all of the assumptions involved in that deterministic result. In regions of high disequilibrium caused by very low levels of recombination, the effects of random genetic drift can easily outweigh the deterministic expectation, as we have demonstrated (Calafell et al. 2001). Modern humans have existed outside of Africa for a relatively short time and had small population sizes during much of that time. Thus, drift associated with the expansion out of Africa and diversification around the world can swamp any factors like recombination that tend to reduce LD. Since these ADH cluster genes are involved in alcohol metabolism, we of course have the additional complication of determining to what extent natural selection may have played a role in altering the frequency of particular haplotypes in different geographical regions.

Finally, Whitfield implies it is important to determine whether social factors or LD are responsible for the differences in effects of *ADH1B*47His* in eastern Asia and in Europe. Social factors can indeed be very important in modifying risk associated with different genotypes. For example, it will be difficult to determine the actual risk associated with ADH variation in northern Africa and the Middle East, since most populations in those regions are Muslim and consumption of alcohol is proscribed by their religion. However, we note that the relevant "genetic" component is not strictly differences in LD but differences in what haplotypes are present. A site in the ADH cluster but not in LD with the *ADH1B* Arg47His site could have an epistatic effect such that only those chromosomes with particular alleles in coupling account for the protective effect. Moreover, background genotype clearly differs at the population level between eastern Asian populations and European populations (e.g., Calafell et al. 1998) adding yet another level of confounding on the path to understanding the role of the *ADH1B* Arg47His polymorphism in risk of alcoholism.

Kenneth K. Kidd, Michael V. Osier,
Andrew J. Pakstis, and Judith R. Kidd
Department of Genetics
Yale University
New Haven

References

Calafell F, Grigorenko EL, Chikanian AA, Kidd KK (2001) Haplotype evolution and linkage disequilibrium: a simulation study. Hum Hered 51:85–96

Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK (1998) Short tandem repeat polymorphism evolution in humans. Eur J Hum Genet 6:38–49

Osier MV, Pakstis AJ, Soodyall H, Comas D, Goldman D, Odunsi K, Okonofua F, Parnas J, Schulz LO, Bertranpetit J, Bonne-Tamir B, Lu R-B, Kidd JR, Kidd KK (2002) A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. Am J Hum Genet 71:84–99

Whitfield JB (2002) Alcohol dehydrogenase and alcohol dependence: variation in genotype-associated risk between populations. 71:1247–1250 (in this issue)

Address for correspondence and reprints: Dr. Kenneth K. Kidd, Yale University School of Medicine, Department of Genetics, 333 Cedar Street, P.O. Box 208005, New Haven, CT 06520-8005. E-mail: kidd@biomed.med.yale.edu

0002-9297/2002/7105-0033$15.00

Am. J. Hum. Genet. 71:1251–1252, 2002

Detecting Polymorphisms and Mutations in Candidate Genes

To the Editor:
Currently, there is no consensus in the literature as to the number and the nature of controls that should be studied to distinguish between polymorphisms and disease-causing mutations (Bridge 1997). The quandary becomes particularly acute when we are trying to determine if a missense alteration in a candidate gene is important (disease associated). How many control samples should be tested, and what other considerations should go into the selection of controls? It is important to consider and report the status, race or ethnic background, and sex (if appropriate) of controls. Furthermore, how many patients should be studied when one is screening a gene for mutations? We have attempted to address these concerns in this letter.

First, one needs to consider if the control subjects could have the same disorder as the case subjects. Where did the control subjects originate, and how were they selected? The use of convenient control subjects (newborn samples, unused diagnostic samples, etc.) may inadvertently include individuals who are carriers or affected. If one uses control subjects selected for a particular study, they may or may not be appropriate for a different study. When studying psychiatric disorders, one needs to ensure that the controls do not have undiagnosed problems. When studying late-onset diseases, one needs to confirm that the control subjects are past the age of onset.

Marchuk (1998) suggested typing controls from similar racial, ethnic, and geographic backgrounds, since allele frequencies can differ between groups. In the past, ignoring this important tenet has caused some mutations to be misclassified. The peripheral myelin protein 22 Thr118Met substitution was believed to be a mutation in Charcot-Marie-Tooth disease, but was found to be a Swedish polymorphism (Nelis et al. 1997). The fibrillin-1 P1148A substitution was initially considered to be a Marfan syndrome mutation in a mixed population of patients, because it had not been found in white or African American control subjects. However, it was later found to be a polymorphism in Asians (Wang et al. 1997). The homeo box A1 A218G polymorphism was reported to increase susceptibility to autism; however, it was found to be more common in African Americans than in whites (Collins et al., in press). Thus, one could misinterpret a negative result if only a single racial or ethnic group is utilized as a control population.

The sex of the control subjects is of obvious importance in testing for polymorphisms in X-linked genes.

Table 1

Sample Sizes Needed to Detect Polymorphisms

N[a]	Polymorphism Frequency	Alpha	Power
40	.05	.05	.80
65	.05	.05	.95
210	.01	.05	.80
340	.01	.05	.95
2,400	.001	.05	.80
3,910	.001	.05	.95

[a] N signifies the number of chromosomes, and this applies to either X-linked or autosomal diseases.

Often, the sex of the control subjects used is not mentioned in the literature. If one looks at chromosomes from normal females in X-linked mental retardation (XLMR), the significance of finding an alteration is unclear, because females are not likely to be affected by a change that could be pathogenic in a male. Therefore, it is imperative, in studying XLMR, to examine chromosomes solely from males of normal intelligence in a reference population and to cite this in any publication.

How many normal controls should be analyzed to detect a 5%, 1%, or 0.1% polymorphism? We used power calculations performed by the Power and Precision program (Biostat) to determine the number of chromosomes required to detect a significant difference between the polymorphism frequency in the reference population and the expected frequency. The polymorphism proportion in the hypothetical control group was set to 0.001% (as close to 0% as possible), since 0% of the controls would be expected to carry a disease-causing mutation. The alpha, or significance level, was set to 5%. The power, or percent of studies expected to yield a significant effect, was set to both 80% and 95%. Power is commonly set at 80%; however, at that level, a polymorphism would be missed 20% of the time. If a power of 95% were used, there would be only a 5% possibility of missing a polymorphism.

Table 1 displays the number of chromosomes that should be examined to significantly determine if the polymorphism frequency in the reference population differs from the expected frequency. The examination of a minimum of 65 chromosomes is necessary to detect a 5% polymorphism with 95% power. Therefore, 95% of the time, the polymorphism will be detected if it is in the population. For a 1% polymorphism, a minimum of 340 chromosomes should be examined. This number is close to Marchuk's (1998) proposal of typing 300–400 or more chromosomes to detect a 1% polymorphism. Finally, 3,910 chromosomes would be required to detect a 0.1% polymorphism with 95% power.

These numbers of chromosomes can also be applied to the search for mutations in disease genes. If each disease gene has been found to cause a certain percentage of a disease, one can utilize that information to determine how many affected individuals should be screened for mutations. For example, it would appear that each known XLMR gene accounts for ~1% of XLMR (Chelly and Mandel 2001). Therefore, a minimum of 340 unrelated males with XLMR should be tested to detect a single alteration in any candidate gene with 95% power.

In summary, when a potential mutation is detected, the status, race, and number of control subjects used for polymorphism detection need to be carefully considered. For X-linked conditions, the sex of controls used should also be taken into consideration. These characteristics are crucial to formulating accurate results. The sample sizes in table 1 can also be applied to the identification of candidate gene mutations in affected individuals.

Acknowledgments

We are grateful for the editorial assistance provided by Holly Gilmore and Dr. Roger Stevenson. This work was supported in part by a grant from the National Institute of Mental Health (MH57840) to J.S.C. and a grant from the National Institute of Child Health and Human Development (HD26202) to C.E.S.

JULIANNE S. COLLINS AND CHARLES E. SCHWARTZ
JC Self Research Institute
Greenwood Genetic Center
Greenwood, SC

References

Bridge PJ (1997) The calculation of genetic risks: worked examples in DNA diagnostics. (2nd ed) Johns Hopkins University Press, Baltimore, p 137

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

Collins JS, Schroer RJ, Bird J, Michaelis RC. The HOXA1 A218G polymorphism and autism: lack of association in white and black patients from the South Carolina Autism Project. J Aut Devel Disord (in press)

Marchuk DA (1998) Laboratory approaches toward gene identification. In: Haines JL, Pericak-Vance MA (eds) Approaches to gene mapping in complex human diseases. Wiley-Liss, New York, pp. 371–372

Nelis E, Holmberg B, Adolfsson R, Holmgren G, Van Broeckhoven CV (1997) PMP22 Thr(118)Met: recessive CMT1 mutation or polymorphism? Nat Genet 15:13–14

Wang M, Mathews KR, Imaizumi K, Beiraghi S, Blumberg B, Scheuner M, Graham JM, Godfrey M (1997) P1148A in fibrillin-1 is not a mutation anymore. Nat Genet 15:12

Address for correspondence and reprints: Dr. Julianne S. Collins, JC Self Research Institute, Greenwood Genetic Center, 1 Gregor Mendel Circle, Greenwood, SC 29646. E-mail: julianne@ggc.org

0002-9297/2002/7105-0034$15.00