nificantly between cases and controls, the correct approach is to perform a heterogeneity test, in which one calculates whether the overall likelihood is significantly higher if different frequencies are allowed than if the same frequencies apply to both groups. An incorrect approach is to estimate haplotype counts by multiplying the frequencies by twice the sample size and then to treat these counts as if they were actually observed. The counts may be compared using a Pearson χ^2 test on a contingency table, by a permutation test as implemented in the CLUMP program (Sham and Curtis 1995) or by the newly described entropy method (Zhao et al. 2005). In every case, the test based on estimated counts will be anticonservative.

To illustrate that this is the case, we randomly generated case-control samples genotyped for two markers, assuming that the population frequencies of the haplotypes were the same for all subjects, under the assumption of random mating. For each data set, we applied a Pearson χ^2 test and the entropy test to the counts of the simulated haplotypes. We then combined pairs of haplotypes into two-locus genotypes, and we used the GENECOUNTING program (Zhao et al. 2002) to obtain estimated haplotype frequencies in the cases, controls, and combined sample, along with the associated likelihoods. We applied a heterogeneity test to these likelihoods and again applied the Pearson χ^2 and entropy tests, this time to the estimated counts. Illustrative results are given in table 1, for which the population frequencies of the four haplotypes were set at 0.5, 0.2, 0.2, and 0.1, and a sample size of 500 cases and 500 controls was used. The Pearson χ^2 and entropy tests perform appropriately when applied to the actual haplotype counts, as does the heterogeneity test using likelihoods based on estimated frequencies. However, both of the tests that use estimated counts are markedly anticonservative.

It is not appropriate to treat estimated haplotypes as if they were observed, and tests that do so will produce unacceptably high type I error rates. As we have said, this will apply even if a permutation test is performed on the estimated haplotypes—for example, by inputting them into the CLUMP program (Sham and Curtis 1995). However, a valid test can be devised if, instead, the original data are repeatedly permuted and then, for each permuted data set, haplotypes are estimated and a test statistic is derived. The rank of the test statistic obtained from the original data set can then be used to obtain an empirical significance level (North et al. 2003), and such an approach could be used for the entropy-based statistic. Without such a permutation procedure, we do not see how the entropy test can be applied to case-control data.

Table 1

Number of Times Each Statistic Reaches a Given *P* Value in 100,000 Simulations

	Real Counts			Estimated Counts	
P	χ^2	Entropy Test	Heterogeneity Test	χ^2	Entropy Test
.05	5,013	4,971	5,072	11,115	11,089
.01	970	968	1,034	3,678	3,678
.001	103	107	114	797	813
.0001	10	7	11	206	209
.00001	0	0	1	51	54
.000001	0	0	1	16	17

David Curtis[1] and Pak C. Sham[2]
[1]*Academic Department of Psychiatry, Queen Mary's School of Medicine and Dentistry, and* [2]*Social, Genetic and Developmental Psychiatry Research Centre, Institute of Psychiatry, London*

References

North BV, Curtis D, Sham PC (2003) A note on the calculation of empirical *P* values from Monte Carlo procedures. Am J Hum Genet 72:498–499

Sham PC, Curtis D (1995) Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. Ann Hum Genet 59: 97–105

Zhao J, Boerwinkle E, Xiong M (2005) An entropy-based statistic for genomewide association studies. Am J Hum Genet 77:27–40

Zhao JH, Sham PC (2002) Faster haplotype frequency estimation using unrelated subjects. Hum Hered 53:36–41

Address for correspondence and reprints: Dr. David Curtis, Academic Department of Psychiatry, Royal London Hospital, Whitechapel, London E1 1BB, United Kingdom. E-mail: david.curtis@qmul.ac.uk

0002-9297/2006/7804-0021$15.00

Am. J. Hum. Genet. 78:730–731, 2006

Reply to Wirtenberger et al.

To the Editor:

Wirtenberger et al. (2006) analyzed the SNP content of 82 large (median length 157 kb) common copy-number polymorphisms (CNPs), selected from the Database of Genomic Variations, and determined the number of SNPs included in the GeneChip Mapping 100K arrays (Affymetrix). The data they presented showed that the density of these SNPs within the CNPs is lower than would be expected, with 52.4% of CNPs having no SNP coverage (median length 120 kb) and only 8.5% having

a SNP density equal to or higher than the overall mean intermarker density for all SNPs on the array.

As suggested by Wirtenberger et al. (2006), the underlying reason for this low density of Mapping 100K SNPs in their selected CNPs is the selection criteria used for SNPs on the array. The SNP selection criteria for the Mapping 100K arrays select strongly but not completely against SNPs in segmental duplications. The selection is based on genotyping accuracy, Mendelian inheritance, Hardy-Weinberg equilibrium, robustness, and reproducibility—all of which are characteristics likely to give poor results in genotyping SNPs that are located in CNPs. Despite a selective bias against SNPs in CNPs, some SNPs on the Mapping 100K arrays are able to provide CNP information. For example, as Wirtenberger et al. (2006) indicate, 14.6% of the CNP regions contained more than four of the SNPs on the array.

Even with modifications in SNP selection, the current algorithm implemented in CNAT (Affymetrix) would still need to be modified, because it compares copy-number data from the test sample with data from a large pool of normal reference individuals, thereby decreasing the likelihood of detecting CNPs. Future advances in SNP selection, algorithm development, and density will be required to identify frequent CNPs by use of SNP arrays.

For investigation of CNPs, the advice of Wirtenberger et al. (2006) to be aware of the limitation of Mapping 100K microarrays is sound. However, it is worth remembering that we (Slater et al. [2005]) describe their use for detection of clinically significant chromosome abnormalities. Exclusion of SNPs within common CNPs is arguably an advantage in the diagnostic scenario when virtually nothing is currently known of the clinical significance of these CNPs.

HOWARD R. SLATER,[1,2] DIONE K. BAILEY,[5] HUA REN,[3] MANQIU CAO,[5] KATRINA BELL,[3] STEVEN NASIOULAS,[1] ROBERT HENKE,[4] K. H. ANDY CHOO,[2,3] AND GIULIA C. KENNEDY[5]
[1]*Genetic Health Cytogenetics Laboratory,* [2]*University of Melbourne Department of Paediatrics, and* [3]*Murdoch Children's Research Institute, Royal Children's Hospital, Melbourne;* [4]*Millennium Biosciences, Box Hill, Australia; and* [5]*Affymetrix, Santa Clara, CA*

References

Slater HR, Bailey DK, Ren H, Cao M, Bell K, Henke R, Choo KHA, Kennedy GC (2005) High-resolution identification of chromosomal abnormalities using oligonucleotide arrays containing 116,204 SNPs. Am J Hum Genet 77:709–716

Wirtenberger M, Hemminki K, Burwinkel B (2006) Identification of frequent chromosome copy-number polymorphisms by use of high-resolution single-nucleotide–polymorphism arrays. Am J Hum Genet 78:520–522

Address for correspondence and reprints: Dr. Howard R. Slater, Cytogenetics Laboratory, Genetic Health Services Victoria, Royal Children's Hospital, Parkville, Victoria 3052, Australia. E-mail: howard.slater@ghsv.org.au

0002-9297/2006/7804-0022$15.00