Am. J. Hum. Genet. 78:1094, 2006

A Note on Permutation Tests in Multistage Association Scans

To the Editor:

There is currently a great deal of interest in performing whole-genome scans for association between genetic markers—mainly SNPs—and biological or clinical end points.[1] Often, the most cost-effective strategy for these studies is a staged design in which a subset of the full sample is genotyped for all SNPs, and only those SNPs that show a trend of association are genotyped in the remainder of the sample.[2]

For calculating the significance of a genome scan, permutation tests have been suggested to adjust for multiple testing while preserving the correlation structure among linked markers.[3] In the staged design, however, permutation may result in a marker being selected for the second stage that had not been selected in the original analysis. Such a marker will not have been genotyped in the full sample, and data will not be available to complete the analysis of the permuted data. Recently, Lin[4] proposed a Monte Carlo method for assessing significance in two-stage association scans. The method is sound but is limited to analysis based on efficient score functions and does not use permutation. Other investigators have reported methods to address this problem.[5]

I wish to draw attention to a property of genome scans that permits a simple permutation procedure for staged designs, which is that the sample sizes are large enough for the null distributions to be asymptotically stable. Although this observation is trivial, its utility might have escaped some readers, because of the origins of permutation testing in small-sample inference. It means that any large subset of the data can be used to simulate the null distribution. In particular, we can simulate a staged design with just the first-stage subjects, by using a subset of the first stage as the simulated first stage, selecting markers on the basis of that subset, and using the remainder of the first stage as the simulated second stage. This ensures that full genotype data are always available and will generate approximately the same null distribution as exists for the full sample.

More precisely, consider a two-stage scan of a set of markers, M, in a set of subjects, S. In the first stage, all markers in M are genotyped in a subset of subjects, $S_1 \subset S$. An algorithm, $A(M; S_1)$, selects a subset of markers, M_1, on the basis of the data for S_1, which are then genotyped in the remaining subjects $S_2 = S \setminus S_1$. Next, perform a permutation test by using just the first-stage subjects as follows. Choose a simulated first-stage subsample, $S_1^* \subset S_1$, and a second-stage subsample, $S_2^* = S_1 \setminus S_1^*$. After each permutation, select markers $M_1^* = A(M; S_1^*)$. Compute statistics for markers M_1^* in subjects S_1, and compare them with the statistics of the original data for markers M_1 in subjects S. Assume that (i) there exists an asymptotic joint null distribution of test statistics on M and (ii) subjects are exchangeable between S_1 and S_2. Then, for sufficiently large $|S_1^*|$, $|S_2^*|$, and $|S_2|$, the permutation test will sample from the same null distribution (up to an arbitrary accuracy) as holds for the two-stage analysis of the full sample S.

For illustration and to confirm that the sample sizes proposed for genomewide scans are sufficiently large, a simulation was performed using 1,000 cases and 1,000 controls, which is a smaller sample than current estimates for well-powered scans.[6] Chromosomes were drawn from the phased CEU (CEPH subjects from Utah) data of chromosome 1, released in phase 1 of the International HapMap Project.[7] Parental chromosomes were drawn independently and grouped in pairs, and gametes were constructed using the supplied recombination maps, under the assumption of the Kosambi function with no interference between adjacent SNPs. Chromosomes of children were assigned from the constructed gametes according to Mendelian transmission and random union of gametes and were randomly assigned to the case or control group. In each replicate, 50% of subjects were used in the first stage, with the 10% most-significant markers considered in the second stage.[2] The significance of individual SNPs was calculated by the trend test,[8] and empirical distributions of the maximum trend statistic were generated from 1,000 replicates.

It is sufficient to show that the two-stage analysis of the first 500 cases and controls yields the same distribution as the analysis of all 1,000. The distributions were compared by the two-sample Kolmogorov-Smirnov test and also by the Kuiper test, which is more sensitive in the tail. No significant difference was found, implying that the null distribution is indeed stable at this sample size.

The main assumption of this approach is that subjects are exchangeable between stages, meaning that the null distribution is independent of the allocation of subjects to stages. This is true when the sample population is homogeneous but not when there are systematic differences between subpopulations. In particular, different patterns of linkage disequilibrium will invalidate this approach, as will population stratification in which differences in both allele frequency and trait distribution create a relationship between the null distribution and the specific subjects analyzed. When the sample consists of known proportions of different populations, the approach can be used if the proportions in the original data are preserved in the permutation test. Also, the large-sample assumption implies that only common variation is included; this is true for Hapmap SNPs, but, if rare variation is included, the permutation test will be less accurate. Nevertheless, for most well-designed scans of common variation, this approach is a practical and easily implemented solution for permutation testing in staged designs.

Acknowledgments

F.D. is supported by European Union contract LSHM-CT-2004-503485. Thanks to Doug Levinson and Peter Holmans for discussions.

Frank Dudbridge
Medical Research Council Biostatistics Unit
Cambridge
United Kingdom

References

1. Thomas DC, Haile RW, Duggan D (2005) Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet 77:337–345
2. Sagatopan JM, Venkatraman ES, Begg CB (2004) Two-stage designs for gene-disease association studies with sample size constraints. Biometrics 60:589–597
3. Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971
4. Lin DY (2006) Evaluating statistical significance in two-stage genomewide association studies. Am J Hum Genet 78:505–509
5. Lewinger JP, Thomas DC (2005) Controlling the family-wise error rate in multistage genome-wide association studies [abstract]. Genet Epidemiol 29:262
6. Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118
7. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
8. Sasieni P (1997) From genotypes to genes: doubling the sample size. Biometrics 53:1253–1261

Address for correspondence and reprints: Dr. Frank Dudbridge, MRC Biostatistics Unit, Robinson Way, Cambridge CB2 2SR, United Kingdom. E-mail: frank.dudbridge@mrc-bsu.cam.ac.uk

0002-9297/2006/7806-0025$15.00