The Effect of Marker Heterozygosity on the Power to Detect Linkage Disequilibrium

Jurg Ott* and Daniel Rabinowitz†

* *Laboratory of Statistical Genetics, Rockefeller University, New York, New York 10021 and*
† *Department of Statistics, Columbia University, New York, New York 10027*

Manuscript received March 14, 1997
Accepted for publication June 30, 1997

ABSTRACT

The relationship between marker heterozygosity and the power to detect linkage disequilibrium is examined through the analysis of an example and through a simulation study. The analysis suggests that, despite the penalties for multiple testing incurred with multiple alleles, greater heterozygosity results in greater power. The results of the simulation study are in accord with those of the analysis.

IN human families, joint inheritance of alleles at two marker loci depends on the recombination fraction between them. When a parent passes a gamete to an offspring, that gamete may reveal whether a recombination or nonrecombination has occurred in the parent only if the parent is heterozygous at each of the two loci. Thus, marker heterozygosity is an important ingredient for linkage informativeness. Some 10 years ago, human genetic maps consisted of markers with two to three alleles each. Since then, highly polymorphic markers with heterozygosities of 80–90% have been developed, which has greatly increased informativeness for linkage analysis. Current proposals call for the development of markers with two alleles each but for extremely large numbers of them so that intermarker distances will be much smaller than they are now. As has been shown previously, high marker density may be a good substitute for low marker heterozygosity (TERWILLIGER *et al.* 1992).

Association studies have been proposed as a means to identify candidate disease loci. For such studies, the role of marker heterozygosity is not immediately clear. On the one hand, detecting association (linkage disequilibrium) with a rare marker allele is easier than with a common allele. Yet, with a large number of marker alleles, trying each of the alleles for possible association brings with it the penalties of multiple testing.

Here, we investigate the effects of marker heterozygosity on the power to detect allelic association to a disease locus with two alleles. Our approach is to compute the power to detect disequilibrium in a hypothetical example. By varying the marker heterozygosity in the example we obtain a qualitative understanding of the effect of heterozygosity on the power. Simulation experiments are used to verify the accuracy of the analysis.

Attention is given to settings in which there is a single founder responsible for introducing a disease mutation into a population. In order to focus on the effect of marker heterozygosity rather than the effect of other factors, we deliberately restrict attention, for the most part, to one simple combination of a sampling plan, an inferential procedure and a population model. In order to focus on the practically important situation in which disequilibrium is not obvious from a small sample, we restrict attention to fairly small amounts of disequilibrium.

WEIR and COCKERHAM (1978) and ZOUROS *et al.* (1977) consider a model similar to the model examined here, but focus on the effect of pooling alleles. They find that pooling usually leads to a loss of power, though WEIR and COCKERHAM note that this need not always be the case. GOLDING (1984) and HUDSON (1983) examine the sampling distribution of linkage disequilibrium, but do not consider directly the question of power. THOMPSON *et al.* (1988) consider the question of power, but restrict attention to markers with two alleles. LEWONTIN (1995) examines questions of power but does not focus on human populations and disequilibrium due to a founder effect.

METHODS

First, the population model for the hypothetical example is described. Consider a situation in which there is a single disease locus and a single marker locus. Let k denote the number of alleles at the marker locus, and let $\pi_1, \pi_2, \ldots, \pi_k$ denote the prevalences in the population of the marker alleles. Consider the situation in which a single founder in a population is responsible for a mutation at a disease locus. Suppose that any member of the population is equally likely to be the

Corresponding author: Daniel Rabinowitz, Department of Statistics, Columbia University, New York, NY 10027.
E-mail: dan@stat.columbia.edu

founder, so that the probability that the mutation is originally in coupling with the ith allele is equal to π_i.

At the time of data collection, in the population of chromosomes with the disease mutation, some proportion will have undergone at least one recombination event between the marker locus and the disease locus. That proportion will be denoted by ρ. We focus attention on settings in which ρ is fairly large, that is, in which the amount of disequilibrium is fairly small. Let θ denote the recombination fraction between the disease locus and the marker locus, and let g denote the number of generations since the introduction of the original mutation (including the original mutation). Then, the expected proportion having undergone a recombination is given by $1 - (1 - \theta)^{g-1}$. For fixed values of ρ and g, then, the corresponding recombination fraction is given by $1 - (1 - \rho)^{1/(g-1)}$.

Assume that, among the portion of chromosomes with the disease mutation that have undergone a recombination, the distribution of marker alleles is not different from that of the general population. This may be thought of as assuming that there is random mixing and a sufficient number of recombinations that there is only only negligible genetic drift. Simulation experiments in which genetic drift does exist suggest that this assumption does not influence the qualitative results of the analysis.

Now, the sampling plan and the inferential approach are described. Suppose that a sample of chromosomes known to have the disease mutation is available. Let n denote the total sample size, and let O_i denote the number of chromosomes in the sample that have the ith allele. Let $E_i = n\pi_i$ denote the expectation, under the null hypothesis of no disequilibrium, of O_i. Assume, for simplicity, that the π_i, and thus the E_i, are known to the data analyst. Suppose that inference is to be based on the usual χ^2 goodness of fit statistic

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

Under the null hypothesis of no disequilibrium, the statistic has for large n, approximately a χ^2 distribution on $k - 1$ d.f.

Now, the distribution of the goodness of fit statistic under the alternative hypothesis of disequilibrium is examined. For local alternatives, that is, for little disequilibrium, for large n, the statistic has approximately a noncentral χ^2 distribution. See, for example, the discussion in WEIR and COCKERHAM (1978) or COX and HINKLEY (1974). It will be seen that the sample size, n, and the proportion of chromosomes with the disease gene that have undergone a recombination, ρ, influence the noncentrality parameter through the quantity $\Delta = n^{1/2}(1 - \rho)$. Suppose that the original mutation was in coupling with a fixed but arbitrary allele i^*. Then, the prevalence of the ith allele in the population of chromosomes containing the original mutation would be equal to $\rho\pi_i$ for i not equal to i^* and would be equal to $(1 - \rho) + \rho\pi_{i^*}$ for i equal to i^*. Here, we have used the assumption that drift does not exist: among the chromosomes with the disease mutation that have undergone a recombination, the prevalences of the marker alleles are taken to be the prevalences in the general population. The noncentrality parameter for the χ^2 statistic would therefore be

$$\sum_{i=1}^{k} \frac{(EO_i - E_i)^2}{E_i} = \frac{(n(1 - \rho + \rho\pi_{i^*}) - n\pi_{i^*})^2}{n\pi_{i^*}} + \sum_{i \neq i^*} \frac{(n\rho\pi_i - n\pi_i)^2}{n\pi_i} = \Delta^2(1/\pi_{i^*} - 1).$$

The noncentrality parameter depends not only on Δ, but also on π_i^*, the prevalence of the marker originally in coupling with the original mutation. The probability that the original mutation occurs in coupling with a particular marker is assumed to be the prevalence of the marker in the general population, so the distribution of the test statistic, when there is a mutation, is a mixture of noncentral χ^2 distributions, with mixing probabilities corresponding the prevalence of the markers. That is, the distribution of the goodness-of-fit statistic is a mixture of noncentral χ^2 statistics on $k - 1$ d.f. with mixing probabilities π_i and with noncentrality parameters $\Delta^2(1/\pi_i^* - 1)$. In the special case that all of the π_i are the same ($\pi_i = 1/k$), then the distribution is that of a noncentral χ^2 statistic on $k - 1$ d.f. and noncentrality parameter $\Delta^2(k - 1)$.

Now, the simulation experiments are described. In all of the simulations, there were 3600 replications. The sample size, n, was taken to be 100. The numbers of marker alleles, k, was taken to be 2, 3, 4, 5, 6, 7, 8, 9, and 10. The alleles were taken to be all equally likely, so the π_i were all $1/k$. The numbers of generations were taken to be 20, 50, and 80. It was assumed that every chromosome with a disease mutation was replicated twice in each succeeding generation. With each replication, there was the possibility of a recombination between the disease locus and the marker locus.

In the simulations, Δ^2 was 1. To obtain a fixed value of Δ^2 for different values of g and n, it is necessary to vary the recombination fraction, θ. Note that ρ may be expressed in terms of Δ and n as $1 - \Delta/\sqrt{n}$. Substituting into the expression for the recombination fraction in terms of ρ and g reveals the required recombination fraction is $\theta = 1 - (\Delta/\sqrt{n})^{1/(g-1)}$.

The simulations were carried out in Fortran, using the IMSL (1994) library of Fortran subroutines. In each of the simulation experiments, the empirical power at level $\alpha = 0.05$ was recorded and compared to the nominal probability from the theoretical analysis. The empirical power was computed as the proportion of replications in which the observed goodness of fit statistic ex-

ceeded the 0.05 level critical values for a χ^2 distribution on $k - 1$ d.f. The nominal power was computed as the exact probability that that a noncentral χ^2 variable exceeds the α-level critical value. Note that the penalty for multiple comparisons is exacted in terms of the critical value: as k increases, the critical value also increases.

RESULTS

Our analysis of the special case of k equally likely alleles indicates that the distribution of the test statistic is χ^2 on $k - 1$ d.f. with noncentrality parameter $\Delta^2(k - 1)$. This is the distribution of the sum of squares of $k - 1$ independent mean Δ variance 1 normal random variables. Thus, the test statistic behaves like an aggregate of k independent equally powerful pieces of information. Our examination shows, therefore, that despite the penalty that must be paid for multiple comparisons, greater marker heterozygosity should result in increased power to detect linkage disequilibrium. The results of the simulation experiments are in accord with this result.

The results of the simulation experiments and the nominal power calculations based on the theoretical analysis are detailed in Table 1. The columns of the table correspond to the number of alleles, k, the number of generations simulated, the recombination fraction between the marker and the disease locus, Δ, and the power derived from the simulation experiments and from the theoretical analysis. The nominal values indicate that for larger values of k (greater heterozygosity) the power to detect linkage disequilibrium increases. Although the simulation results suggest that the nominal values are somewhat lower than the true power for fewer generations, the qualitative result, that increased power results from increased heterozygosity, is evident from the simulation results as well.

The analysis and the simulations were restricted to the special case of equally prevalent alleles. The noncentrality parameter, given that a particular allele i^* is coupled with the original mutation is $\Delta^2(1/\pi_{i^*} - 1)$. This quantity is smaller for larger values of the prevalence of allele i^*, π_{i^*}. Since alleles with larger prevalence are more likely to be coupled with the original mutation, the effect of unequal prevalences would be that the smaller noncentrality parameters would be more likely. Thus, with unequal prevalences, there would be less power than what was calculated under the assumption of equally prevalent alleles.

DISCUSSION

High marker heterozygosity is correlated with increased mutation rate, which is deleterious for association studies. See, for example, HARTL and CLARK (1989). For a disease locus at a location with an increased mutation rate, it is more likely that there would be more than one founder and also more likely that the the coupling of the disease mutations with the founders' marker alleles could become attenuated through mutations at the marker locus. These possibilities are disregarded in our analysis.

Here, as in WEIR and COCKERHAM (1978), situations are considered in which disequilibrium is slight but sample sizes are large. The assumption plays two roles in the χ^2 approximation to the distribution, under disequilibrium, of the test statistic. The first is that slight disequilibrium corresponds to many recombinations between the the disease and marker loci. From this it follows that the frequencies of the marker alleles on chromosomes that have undergone recombinations will tend to be close to the frequencies in the general population. The second is that the variability in the observed frequencies will be close to that expected under the null hypothesis of no disequilibrium. From this it follows that dividing by the counts expected under the null hypothesis will properly normalize the statistic. Although the analysis ignored these effects, the effects did exist in the simulations studies. The results of the simulations studies indicate that for the settings considered, the effects were not substantial.

TABLE 1

Empirical and nominal power

k	Generations	Recombination fraction	Nominal	Empirical
2	20	0.1141	0.17	0.24
3	20	0.1141	0.23	0.29
4	20	0.1141	0.27	0.34
5	20	0.1141	0.32	0.40
6	20	0.1141	0.36	0.43
7	20	0.1141	0.40	0.48
8	20	0.1141	0.44	0.53
9	20	0.1141	0.48	0.54
10	20	0.1141	0.51	0.56
2	50	0.0459	0.17	0.19
3	50	0.0459	0.23	0.26
4	50	0.0459	0.27	0.30
5	50	0.0459	0.32	0.34
6	50	0.0459	0.36	0.36
7	50	0.0459	0.40	0.42
8	50	0.0459	0.44	0.45
9	50	0.0459	0.48	0.48
10	50	0.0459	0.51	0.50
2	80	0.0287	0.17	0.20
3	80	0.0287	0.23	0.25
4	80	0.0287	0.27	0.28
5	80	0.0287	0.32	0.31
6	80	0.0287	0.36	0.37
7	80	0.0287	0.40	0.38
8	80	0.0287	0.44	0.43
9	80	0.0287	0.48	0.47
10	80	0.0287	0.51	0.49

This work was supported by grant HG-00008 (J.O.) from the National Human Genome Research Institute and grant GM-55978 (D.R.) from the National Institute of General Medical Sciences.

LITERATURE CITED

COX, D. R., and D. V. HINKLEY, 1974 *Theoretical Statistics.* Halsted Press, New York.

GOLDING, G. B., 1984 The sampling distribution of linkage disequilibrium. Genetics **108:** 257–274.

HARTL, D. L., and A. G. CLARK, 1989 *Principles of Population Genetics.* Sinauer Associates, Sunderland, MA.

HUDSON, R. R., 1983 The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics **109:** 611–631.

LEWONTIN, R. C., 1995 The detection of linkage disequilibrium in molecular sequence data. Genetics **140:** 377–388.

TERWILLIGER, J. D., Y. DING and J. OTT, 1992 On the relative importance of marker heterozygosity and intermarker distance in gene mapping. Genomics **13:** 951–956.

THOMPSON, E. A., S. DEEB, D. WALKER and A. G. MOTULSKY, 1988 The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes. Am. J. Hum. Genet. **42:** 113–124.

Visual Numerics, Inc., 1994 IMSL Fortran subroutines. Houston, TX.

WEIR, B. S., and C. C. COCKERHAM, 1978 Testing hypotheses about linkage disequilibrium with multiple alleles. Genetics **88:** 633–642.

ZOUROS, E., G. B. GOLDING and T. F. C. MACKAY, 1977 The effect of combining alleles into electrophoretic classes on detecting linkage disequilibrium. Genetics **85:** 543–556.

Communicating editor: T. F. C. MACKAY