Am. J. Hum. Genet. 72:891–902, 2003

On the Identification of Disease Mutations by the Analysis of Haplotype Similarity and Goodness of Fit

Jung-Ying Tzeng,[1] B. Devlin,[2] Larry Wasserman,[1] and Kathryn Roeder[1]

[1]Department of Statistics, Carnegie Mellon University, and [2]Department of Psychiatry, University of Pittsburgh, Pittsburgh

The observation that haplotypes from a particular region of the genome differ between affected and unaffected individuals or between chromosomes transmitted to affected individuals versus those not transmitted is sound evidence for a disease-liability mutation in the region. Tests for differentiation of haplotype distributions often take the form of either Pearson's χ^2 statistic or tests based on the similarity among haplotypes in the different populations. In this article, we show that many measures of haplotype similarity can be expressed in the same quadratic form, and we give the general form of the variance. As we describe, these methods can be applied to either phase-known or phase-unknown data. We investigate the performance of Pearson's χ^2 statistic and haplotype similarity tests through use of evolutionary simulations. We show that both approaches can be powerful, but under quite different conditions. Moreover, we show that the power of both approaches can be enhanced by clustering rare haplotypes from the distributions before performing a test.

Introduction

Mutations inducing simple genetic diseases are often found in tight linkage disequilibrium with alleles surrounding them (de la Chapelle and Wright 1998; Jorde 2000). When they are, various analyses have shown that disease alleles could be fine-mapped (McPeek and Strahs 1999) and even detected (Houwen et al. 1994) by statistical analysis of unusual haplotype sharing. For complex disorders, however, we expect much more etiologic heterogeneity, even within the same population. Nevertheless, on the basis of theory and practice (Puffenberger et al. 1994; Feder et al. 1996), we expect to find disproportionately large clusters of affected individuals sharing common haplotypes in the region flanking some diseases mutations, especially for mutations of recent origin (Fan and Lange 1998).

When the genealogical relationships among a sample are known, simulation (Houwen et al. 1994) or approximation (Durham and Feingold 1997; Service et al. 1999) methods can be used to measure the expected amount of haplotype sharing within a particular region. In many instances, however, the genealogy of the sample will not be known. Instead, one might look for a sample of unaffected individuals to contrast their degree of haplotype sharing with that from the sample of affected individuals.

Several statistical procedures have been proposed that are based on searching for excess similarity among haplotypes from affected individuals. Unlike the usual comparisons of allele or haplotype frequencies, these statistics are based on pairwise comparisons of all haplotypes in a sample of affected individuals. The initial concept was put forth by van der Meulen and te Meerman (1997), for analysis of excess matching of case haplotypes in a case-control and family-based samples. Bourgain et al. (2000, 2001) developed a more rigorous basis for the concept and extended these ideas to family-based studies, showing how such an approach would have proven useful for the study of celiac disease. In a parallel development, Grant et al. (1999) proposed a computationally intensive test statistic that searched for clusters of chromosomal segments exhibiting excess matching. This method was tailored to genome mismatch scanning (Cheung and Nelson 1998), a promising technique that is not yet feasible on a large scale. Devlin et al. (2000) developed a similar approach, relying on measurement of a dense set of genetic markers rather than genome mismatch scanning. Both of these methods have the advantage of not requiring a control sample but the disadvantage of requiring demanding laboratory procedures.

Recognizing that neither genome mismatch scanning nor a very dense set of genetic markers is likely to be feasible in the near future, Tzeng et al. (in press) extended the basic ideas of haplotype matching to a more practical level, which required a far cruder level of genotyping but reincorporated controls into the study de-

Received September 27, 2002; accepted for publication January 8, 2003; electronically published February 27, 2003.

Address for correspondence and reprints: Dr. Kathryn Roeder, Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. E-mail: roeder@stat.cmu.edu

0002-9297/2003/7204-0011$15.00

sign. They also showed how the matching statistic could be adjusted for population substructure, by using the concept of "genomic control" introduced by Devlin and Roeder (1999). Still, Tzeng et al.'s (in press) approach, as well as most others described previously, requires the input of haplotype data; in most case-control studies, however, haplotypes often can only be inferred with probability from data on multilocus genotypes.

In the present study, we develop a class of haplotype-sharing models applicable to multilocus genotype data. These models share many of the features developed by Tzeng et al. (in press) and other methods in the literature, but in a more general framework. The classic competitor to these haplotype-sharing models is Pearson's χ^2 statistic, a goodness-of-fit (GOF) test. Surprisingly, insofar as we are aware, the performance of these sharing statistics, relative to that of GOF statistics, is not well known. Therefore, we also evaluate the performance of these statistics, by simulation. We show that the power of both GOF and sharing statistics can be strikingly different, depending on evolutionary history, and that neither is uniformly most powerful.

Methods

We first develop several measures of haplotype similarity suitable for a case-control sample of haplotypes with known phase, and we then extend them to account for phase uncertainty.

Analysis of Haplotype Data

Consider a region of interest from which n haplotypes are samples from affected individuals (case haplotypes) and m haplotypes are sampled from unaffected individuals (control haplotypes). Let H_i denote the ith sampled haplotype, and assume that there are R distinct haplotypes in the population. Let $\pi_{al} = \Pr(H_i = l)$ for a case haplotype, $i = 1,\ldots,n$ and $l = 1,\ldots,R$, and $\pi_{ul} = \Pr(H_i = l)$ for a control haplotype. A GOF test with $R - 1$ df offers one possible test to determine whether these haplotype distributions, $\Pi_a = (\pi_{a1},\ldots,\pi_{aR})$ and $\Pi_u = (\pi_{u1},\ldots,\pi_{uR})$, differ.

Alternatively, statistics based on pairwise comparisons between haplotypes can reveal unusual clustering of similar haplotypes in the affected versus unaffected individuals, essentially using only 1 df. For instance, to measure the degree of matching within a sample, suppose we draw two case haplotypes at random. The chance that they will have the same version of the haplotype is $\sum_l \pi_{al}^2$ (analogous to the homozygosity). If a mutation leading to increased risk of disease occurred in the population, it is likely to be embedded within a relatively common haplotype. Furthermore, if a cluster of case haplotypes traces back to this common ancestor, then a sample of affected individuals will likely have a higher frequency of matching haplotypes, relative to the controls, and excess matching can be measured by the difference in the matching indexes (Van der Meulen and Te Meerman 1997). This measure, which Tzeng et al. (in press) call the "matching measure," tends to be large if substantial clusters of case haplotypes derive from one (or, at most, several) common ancestor(s), such as would be anticipated under the alternative hypothesis of association.

Statistics based on the sum of all pairwise comparisons between sampled haplotypes take the general form of a U statistic (Lee 1990):

$$U = \frac{2}{n(n-1)} \sum_{i<j}^{n} K(H_i,H_j) \ ,$$

where $K(H_i,H_j)$ is a symmetric kernel function of some feature in the comparison of the ith and jth haplotypes. By defining $K(H_i,H_j)$ to be 1 if the haplotypes match and 0 otherwise, we obtain the matching measure. It should be clear that the matching measure cannot be optimal for all evolutionary or practical scenarios. For example, because it treats each haplotype as a distinct category, as does GOF, information about clustering of similar but nonidentical haplotypes is not assimilated. There are alternative ways to measure similarity of haplotypes. Statistics that are sensitive to haplotypes that nearly match, as well as those that match at every measured marker, can be obtained by defining $K(H_i,H_j)$ to be the length spanned by the longest continuous interval of matching alleles or the number of alleles in common between haplotypes i and j in the region. We call the former "length measure" and the latter "counting measure." The length measure is quite similar to the maximum identity length contrast statistic (Bourgain et al. 2000) and forms the basis for other related statistics (MacLean et al. 2000; Qian and Thomas 2001; Thomas et al. 2001). All three haplotype similarity measures estimate quantities that can be expressed in a quadratic form, $\Pi' \mathbf{A} \Pi$, where $\mathbf{A}$ is a symmetric matrix containing the entries defined by the U-statistic kernel.

Each of these measures has notable strengths and weaknesses. With a dense grid of markers in a narrowly defined haplotype, the matching measure has the simple interpretation of comparing the level of excess sharing, ideally identity-by-descent (IBD) sharing, between cases and controls. It is not robust to genotyping errors, missing data, and recent marker mutations, however, and it can be sensitive to the length of the haplotypes under investigation, which is often an arbitrarily chosen value. The length measure captures partial sharing due to recombinations in the ancestral haplotype, and it is fairly robust to the length of the haplotype under investigation.

It is, however, nonrobust to genotyping errors, missing data, and recent mutations. In contrast to the other measures, the counting measure is robust to genotyping errors, missing data, and recent mutations. In addition, it can be thought of as a compromise of between the matching and length measures. The counting measure possesses an unexpected and convenient statistical feature: this statistic can be computed directly from multilocus genotype data (appendix A). Consequently, genotype data are essentially as informative as haplotype data for this measure.

If haplotype samples are available, $\hat{\pi}_{al} = 1/n \times$ (No. of case haplotypes of type l) is the maximum-likelihood estimator (MLE) for the haplotype frequencies in affected individuals, and $\hat{\pi}_{ul}$ can be defined analogously for the unaffected individuals. An estimator of $\Pi^t\mathbf{A}\Pi$ that is essentially equivalent to the U statistic estimator can be obtained by simply substituting the MLE $\hat{\Pi}$ for Π. The plug-in MLE estimator conveniently extends when genotypes, rather than haplotypes, are measured, or when there are missing data.

In summary, a test statistic for association can be based on the difference in haplotype similarities between cases and controls, properly normalized:

$$D = \hat{\Pi}_a^t \mathbf{A} \hat{\Pi}_a - \hat{\Pi}_u^t \mathbf{A} \hat{\Pi}_u \ , \ T = \frac{D}{\sigma(\hat{\Pi})} \ ,$$

where $\sigma^2(\Pi) = \mathrm{Var}\,(D)$, under the assumption of no population substructure or cryptic relatedness. A formula for the exact variance is derived by Tzeng et al. (in press); the expression for this variance is given in appendix B. In practice, $\hat{\Pi}$, obtained from the pooled sample of case and control haplotypes, is substituted for Π in $\sigma(\Pi)$. Provided Π is bounded away from singularities in the limiting distribution, T is approximately distributed as a standard normal. Each of these measures is likely to outperform the other and the GOF test, under certain evolutionary models. Which measure to use depends upon the nature of the genetic markers and the history of the population under investigation.

Analysis of Genotypic Data

From genotype data, haplotype frequencies Π can be estimated by maximum likelihood—for example, using algorithms such as the EM. For typical sample sizes, the measurement error in these estimates is primarily due to sampling error rather than to haplotype phase uncertainty (Fallin and Schork 2000; Gabriel et al. 2002). For this reason, we develop a test based on the estimated haplotype frequencies, with minor adjustments to account for haplotype uncertainty.

Although there is likely to be a limited set of haplotypes in the population (Gabriel et al. 2002), phase uncertainty can make it appear that there are many more, on the basis of an examination of multilocus genotypes. Many of the estimated haplotype frequencies are likely to be small, and the rare ones are likely to have arisen from genotyping errors, missing data, and recent marker mutations. Therefore, it seems good practice to prune the list of potential haplotypes back to an intermediate number of relatively common haplotypes before computing the test statistic. In addition, simulations (not presented) show that $\sigma(\Pi)$ is poorly estimated by $\sigma(\hat{\Pi})$ unless the rare haplotypes are removed from the field of potential types. Although pruning can be especially important when analyzing multilocus genotype data, it can also be important for the analysis of data on unambiguous haplotypes.

To prune the list, first order the haplotypes by their relative frequencies. On the basis of the haplotype frequencies, choose R^*, the number of common categories to be retained. We see two natural choices, which we will call the "pooling" and "clustering" methods. With the pooling method, define categories 1 through R^* to be the R^* most common types in the combined case/control sample, and then pool all other types into the $(R^* + 1)$ category. With the clustering method, again define categories 1 through R^* by the R^* most common types; among the $(R^* + 1)$ to R remaining types, however, add any that differ from the common types by a one-step mutation to that category; if more than one category differs by a one-step mutation, assign the rare type in question to the most common of those in the equivalence category. Finally, pool all other types into the $(R^* + 1)$ category. For both methods, add the relative frequencies of all those types assigned to a category to obtain the new vector of relative frequencies. Because the "other" category is meaningless when discussing haplotype similarity, we discard this category for all haplotype similarity tests.

When haplotype phase is unknown, $\mathrm{Var}\,(D) \geqslant \sigma^2(\Pi)$ and the inequality is typically strict. However, there is one surprising exception to this rule. For the counting measure, D can be computed directly from multilocus genotype data and, hence, $\mathrm{Var}\,(D) = \sigma^2(\Pi)$. Otherwise, the test statistic T must be adjusted to account for the extra variance due to phase uncertainty.

To compute the test statistic, the following steps are required:

1. Obtain haplotype frequencies $\hat{\Pi}$, $\hat{\Pi}_a$, and $\hat{\Pi}_u$, for the pooled and unpooled data, using any haplotype algorithm.
2. Prune the list of potential haplotypes and update $\hat{\Pi}$, $\hat{\Pi}_a$, and $\hat{\Pi}_u$ to reflect the relative haplotype frequencies of the pruned list for the pooled and unpooled samples.
3. Compute T_{obs} (using appendix B) and $p^* = 1 - \Phi(T_{\mathrm{obs}})$, the tail of the standard normal distribution.

For the counting measure, p^* is the actual P value, and for the other measures it is a lower bound on the P value. (If p^* is not small, there is no reason to proceed to the next step; the results are not significant.)

4. For the matching and length measures perform a bootstrap experiment, randomly drawing n "case" haplotypes and m "control" haplotypes with probability determined by the pooled $\hat{\Pi}$. Compute $\hat{\Pi}_b$ and T for each of B bootstrap samples. Obtain an empirical P value from the bootstrap experiment. Note that, for the matching and length measures, the empirical P value will be somewhat larger than p^*. This is to be expected, because of the variance inequality. It is also worth noting that, for small samples, one should obtain an empirical P value via the bootstrap for the counting measure.

Production of Simulated Data

Data were obtained using the methods and program described by Lam et al. (2000). The evolutionary program mimicked features of natural populations to the extent possible, by using direct simulation methods. In the simulations, diploid individuals paired at random in their generation, mated, and produced a random number of children. Recombination followed the Haldane model. Each population was founded by 1,000 individuals and remained at that size for 50 generations. This initialization, together with small population growth in early generations, generated random linkage disequilibrium among alleles on normal chromosomes. After 50 generations, a disease mutation was introduced on one chromosome, and the population grew exponentially for 200 generations to a final size of 50,000 individuals. Twelve STR markers were simulated, covering a 1.2-Mb critical region, with spacings of 0.11 Mb between markers. The disease mutation was located between markers 6 and 7. Alleles at each locus were simulated to mimic STRs with a mutation rate of 0.001. The average heterozygosity across STR markers was 0.80. To illustrate the effect of the number of markers, we analyzed haplotypes with the following subsets of markers included: 1, 3, 5, 7, 9, and 11; 3, 5, 7, and 9; and 3, 5, and 7, which yields 0.22 Mb between markers.

The simulation program produced populations from which samples of haplotypes with or without a disease mutation could be drawn. To mimic a complex disease of heterogeneous origin, we generated samples of case and control individuals by drawing haplotypes in the following way: for each affected individual, with probability 0.2, draw a haplotype bearing a mutation—otherwise, draw a "normal" chromosome; for each unaffected individual, draw two normal haplotypes. Consequently, in a sample of 100 affected individuals, only 40 of the 200 haplotypes will bear a mutation.

Results

We examine 18 populations that could be classified into two types: 9 populations in which the disease mutation occurred on a haplotype that is common in the population, and 9 for which the disease mutation occurs on a rare haplotype. This fundamental distinction—derivation from a rare or common haplotype—will be key to understanding the performance of the GOF versus sharing methods. From each population, 200 samples consisting of 100 affected and 100 unaffected individuals are generated, and these data are analyzed using GOF, matching, length, and counting statistics.

Each population has a large number of haplotypes (~26 types for the three-marker system and 43 for the six-marker system). Figure 1 shows the distribution of haplotypes among cases and controls for four populations. As is often observed in natural populations, a few haplotypes dominate. Because the sample size is moderate relative to the dimension of the problem, the GOF statistic does not achieve its asymptotic limiting distributions. We performed permutation tests to obtain accurate P values. When case and control haplotypes are drawn from the same population, the size of all four tests is close to the nominal 0.05 level for both kinds of populations (table 1).

Our initial power analyses assume that haplotype phases are known. For the data generated from populations in which the mutation occurred on a common haplotype, all three similarity statistics perform far better than the GOF test (fig. 2*A*). The maximum power differential is 63%. These nine populations roughly constitute three conditions, and the labels (1–9) are assigned to emphasize a gradient of performance (1–3 = excellent, 4–6 = good, and 7–9 = moderate). The six-marker system achieved somewhat more power than the three-marker system (fig. 2*A* vs. 2*B*). Results for the four-marker system followed the same pattern, with power intermediate to the six- and three-marker systems. For these simulations, which have no missing data or genotype errors and in which all populations are generated by identical evolutionary forces, the three haplotype similarity statistics performed equivalently. These results are not likely to hold in general. To evaluate power for the case of unknown haplotype phase, we obtain the maximum likelihood estimate of haplotype frequencies through use of the EM algorithm. From these simulations and by using the bootstrap to obtain exact P values, the same, general pattern of results is obtained; however, there is a slight drop in power when

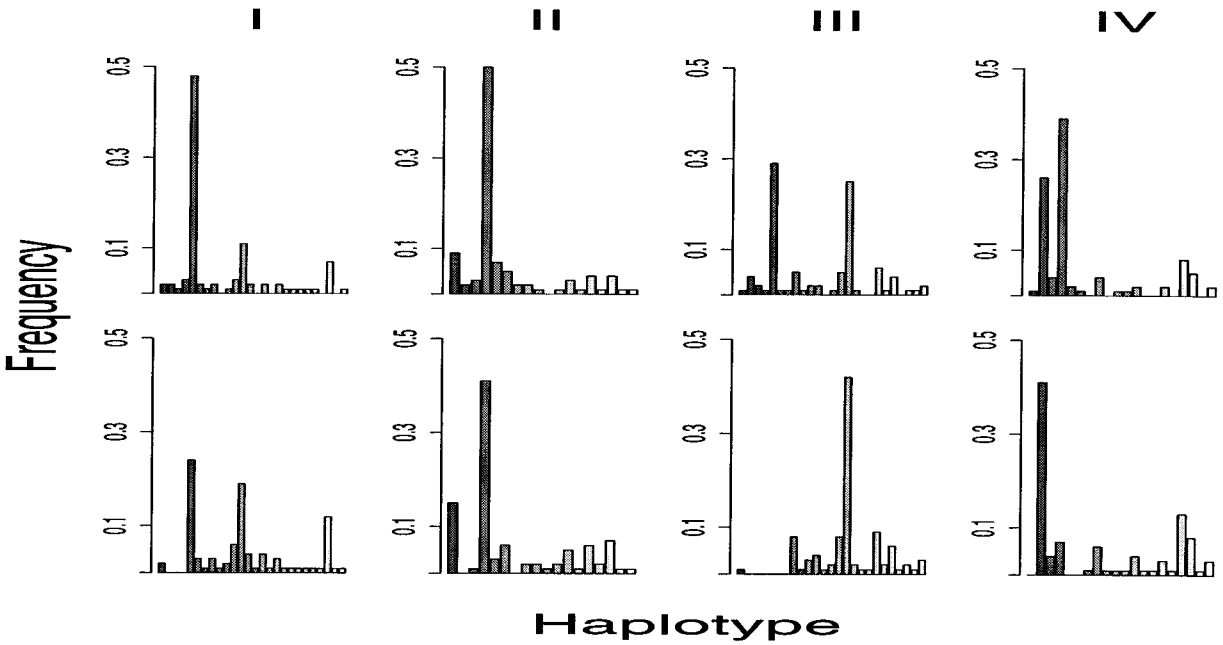

Figure 1 Distribution of haplotypes in affected (*top*) and unaffected (*bottom*) individuals for four populations. In populations I and II, the disease mutation occurred on a common background; in population III, it occurred on a haplotype so rare that it was not sampled in our simulation of the controls; and in population IV, it occurred on a haplotype that is rare in the normal population. In these bar plots, haplotypes that differ by a single mutation tend to appear adjacent to one another.

genotypes rather than haplotypes are analyzed (fig. 2*C* vs. 2*B*).

In summary, when the mutation occurs on a common haplotype, the similarity measures produce superior power compared with GOF, yet there is a clear gradient of performance (fig. 2). Examination of the haplotype distributions bearing mutated and normal alleles (e.g., fig. 1) suggests the gradient can be explained by two features: (1) although all samples from affected individuals have a dominant haplotype that bears the mutation, the relative frequency of this haplotype is much greater in the populations producing high power (1–3) than it is in the populations producing moderate power (7–9); and (2) the distribution of normal haplotypes also exhibits greater entropy in the populations producing high power. Populations 4–6, which produce intermediate power, are intermediate in both of these defining characteristics. The haplotype distributions for population 1 and 8 (fig. 1, *panels I* and *II*) are good examples for these contrasting features.

When the mutation occurs on a rare haplotype, the power differential switches. The haplotype similarity statistics have low power, whereas the GOF test has high power. Of the nine populations (10–18) investigated using the three-marker haplotypes, four exhibited no power for any of the haplotype similarity measures (fig. 2*D*). The haplotype distributions for populations 10 and 14 (fig. 1, *panels III* and *IV*) are good examples illustrating haplotype similarity measures that have no and low power, respectively. In population 10, for example, the case samples have greater entropy than the control samples; therefore, the haplotype similarity test statistics are negative, and one-sided tests cannot be powerful. In contrast, the greater power for population 14 results because the case sample has slightly less entropy than the control sample (fig. 1, *panel IV*).

To determine a rule for dimension reduction (DR) we used the same set of simulations and investigated the power for rules that chose R^* such that $\hat{\pi}_l \geq c\pi_{\max}$, $l = 1, \ldots, R^*$, for $c = 1/10$, $1/6$, $1/5$, $1/4$, or $1/3$. Using three-marker genotype data, we record which c produces the greatest power for each population. When this experiment is performed for both the rare (1–9) and the common (10–18) populations, the performance of DR seems to depend on the nature of the haplotype distribution, and no definitive rule for choosing R^* emerges. For some populations, the power of the haplotype similarity tests is essentially constant for a large range of values of R^*. Nevertheless, by ranking the performance of each rule across conditions, we determined that the cutoff with $c = 0.2$ tends to be more reliable in improving the power of both the GOF test and the

Table 1

Type I Error Rates

	Type I Error Rate			
	Haplotypes		Genotypes	
Tests	Six-Marker	Three-Marker	Three-Marker	Three-Marker (DR)
Common:				
Length	.051	.050	.048	.049
Counting	.049	.050	.049	.049
Matching	.049	.050	.048	.050
χ^2	.052	.050	.051	.050
Rare:				
Length	.049	.050	.051	.050
Counting	.049	.050	.051	.048
Matching	.047	.049	.050	.047
χ^2	.049	.050	.049	.048

Note.—Analyses were for four different tests, four different marker configurations (DR = dimension reduction), and two population scenarios—namely, disease mutation on common or rare haplotypes. Each entry is based on 1,000 experiments.

haplotype similarity tests, across both rare and common populations.

For the three-marker haplotypes, the clustering and pooling DR methods perform fairly similarly. For the six-marker haplotypes, however, the pooling method appears to break down (data not shown), because, in this situation, only a very few haplotypes are relatively common, and it is difficult to determine how to best pool the data. For this reason, we recommend using the clustering method of DR in practice.

Our investigations of DR, using populations 1–18 and the clustering DR rule with $c = 0.2$, suggest that clustering rare haplotypes from the distribution improves power in many instances and rarely decreases power by more than a negligible amount (table 2). Not surprisingly, the DR had the biggest impact on the GOF tests, because it reduces the df of the tests, and it had less of an effect on the power of haplotype similarity tests, because they are 1-df tests. For those populations in which the mutation occurs on a rare haplotype, even when DR induces a substantial increase in power for a haplotype similarity test (e.g., population 18, matching statistic), it is not sufficient to overcome the power advantage of GOF. DR improves the reliability of haplotype similarity tests, especially when the number of potential haplotypes is very large, by increasing the accuracy of the estimated denominator of the test, $\sigma(\hat{\Pi})$.

To explore the impact of incomplete penetrance, we also generated control samples that included 5% diseased haplotypes. Incomplete penetrance reduces the power of the tests, but the size of the tests is unchanged (data not shown).

Discussion

The search for disease mutations can sometimes be enhanced by the analysis of data on haplotypes. For example, in a sample of affected individuals and their parents, the fact that the distributions of transmitted and untransmitted haplotypes are distinctly different for a particular genomic region is strong evidence that a disease mutation lies in that region (Jorde 2000). A similar conclusion holds when the haplotype distribution differs between affected and unaffected individuals, under the assumption of population homogeneity. There are many ways to contrast haplotype distributions, especially for a sample of affected and unaffected individuals. In our reading of the literature, two common approaches emerge: GOF tests, such as a Pearson χ^2 statistic; and contrasts of various measures of haplotype similarity. Surprisingly, insofar as we are aware, there is little in the literature to differentiate between these two approaches.

In this article, we unify some of these measures of haplotype similarity, showing that they can be cast in the same quadratic form. Although such analyses are simplest when haplotype phase is known, unphased multilocus genotypes present no particular problem. Using this quadratic form, we investigate the power of haplotype-sharing measures and GOF for case-control sampling and two distinct settings: when the disease mutation arose on a haplotype that is common in the population (common case), and when the disease mutation arose on a rare haplotype (rare case).

We find that haplotype similarity measures have substantially higher power for data from the common case. In contrast, GOF tests have substantially higher power

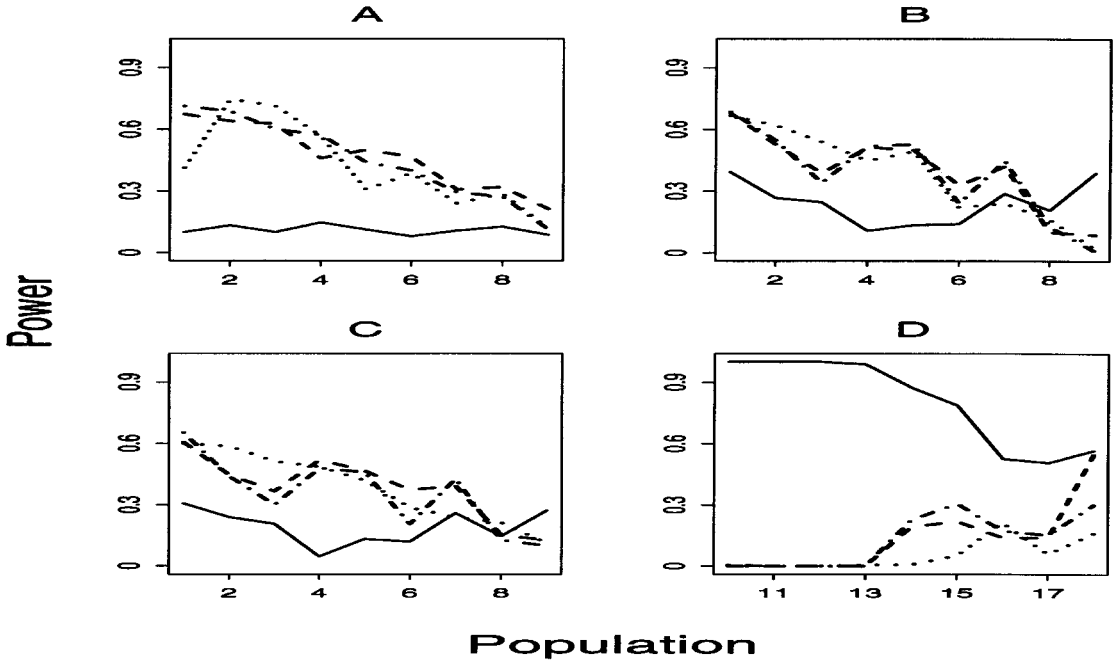

Figure 2 Performance of GOF and similarity statistics, in terms of power. The dotted, dashed, dot-dashed, and solid lines display the matching, count, length, and GOF statistics, respectively. The results represent nine separate populations (*X*-axis) ordered by power, which are connected to facilitate visualization. For panels *A*–*C*, the disease mutation occurred on a common background; for panel *D*, the disease mutation occurred on a rare background. Tests are performed using six-marker haplotypes (*A*), three-marker haplotypes (*B* and *D*) and three-marker genotypes (*C*).

for data from the rare case. This power advantage for GOF in the rare case occurs because there are very different haplotypes with elevated frequency in the affected versus unaffected individuals. Oversampling of affected individuals elevates the frequency of the rare haplotype at the expense of more common haplotypes not bearing disease alleles. Haplotype similarity measures, being 1-df tests, cannot detect this type of signal, because they do not keep track of which categories are relatively common and rare. It is implicitly presumed that the cases and controls share common haplotypes. To a first order of approximation, the haplotype similarity measures are comparing the entropy of the case and the control haplotype distributions. Under this scenario, the entropy can be quite similar for these distributions, or even greater, in the case distribution. This leads to a test with essentially no power.

According to evolutionary and statistical theory, a disease mutation is more likely to occur on a common haplotype, which in turn leads to a haplotype similarity statistic with good power. Such an event cannot be guaranteed a priori, however. For this reason, we recommend applying both GOF and one or more of the haplotype similarity tests. In addition, we recommend using the clustering DR method with cutoff value $0.2\pi_{max}$ or some other theoretically derived value, regardless of which test is used.

Tests for differentiation between haplotype distributions do not always fall into the neat categories of GOF and similarity tests; however, one can usually point to basic features of the tests that make them most like one group. For example, tests that contrast haplotype frequency distributions, by use of likelihood ratio, score, or permutation tests (e.g., Fallin and Schork 2000; Culverhouse et al. 2001), and tests that contrast transmitted and nontransmitted haplotype distributions (e.g., Clayton 1999) can be grouped into the GOF category. We expect those tests to perform well when a disease mutation falls on a rare haplotype, which is then elevated in frequency by the oversampling of affected individuals. When all possible pairwise comparisons of haplotypes are made, the goal is typically to assess and contrast haplotype similarity (e.g., tests based on the length of shared chromosomal segments [Van der Meulen and Te Meerman 1997; Bourgain et al. 2000; Qian and Thomas 2001; Thomas et al. 2001]). Our results suggest that these tests will perform well when the disease mutation falls on a common haplotype.

Table 2

Power of Tests with Full and Reduced Dimensions

	Power of Test							
	Length		Counting		Matching		χ^2	
Population[a]	Full	Reduced	Full	Reduced	Full	Reduced	Full	Reduced
Common:								
1	.65	.64	.61	.65	.60	.65	.31	.55
2	.44	.46	.45	.45	.59	.57	.24	.38
3	.30	.32	.37	.33	.51	.50	.21	.23
4	.48	.49	.52	.50	.49	.51	.05	.15
5	.46	.49	.47	.48	.42	.45	.13	.25
6	.21	.31	.37	.37	.29	.31	.12	.20
7	.43	.49	.39	.49	.24	.39	.26	.39
8	.13	.13	.15	.14	.21	.26	.15	.23
9	.10	.15	.13	.17	.11	.15	.27	.35
Rare:								
10	.01	.01	.00	.01	.00	.00	.99	1.00
11	.00	.00	.00	.00	.00	.00	.99	1.00
12	.01	.01	.00	.01	.00	.01	.99	1.00
13	.00	.02	.00	.02	.01	.05	.91	.99
14	.23	.27	.19	.23	.01	.03	.78	.91
15	.31	.23	.21	.19	.05	.09	.64	.95
16	.20	.17	.15	.17	.19	.19	.42	.57
17	.15	.15	.14	.15	.05	.15	.53	.76
18	.30	.51	.50	.49	.16	.53	.49	.77

Note.—The tests are applied to three-marker genotype data.

[a] The 18 populations are classified by origin of disease haplotype (common or rare).

Some tests, however, are more difficult to categorize. For example, Service et al. (1999) introduce a test that contrasts the distribution of haplotypes in affected individuals versus that expected on the basis of the genealogy of the sample descended from a common ancestral haplotype. If the identity of the ancestral haplotype were known, then the method would seem closest to GOF and would have optimal power if the ancestral haplotype were rare. Because the ancestral haplotype is unknown, however, Service et al. (1999) consider each kind of haplotype as a possible ancestral and obtain the likelihood of the data as a weighted sum over distinct haplotypes. In that case, the method resembles a test of haplotype similarity, in that the power may be optimized if the ancestral haplotype were common.

The distinction is more than academic. For example, using a population isolate from Costa Rica, Ophoff et al. (2002) present a genomewide linkage disequilibrium analysis to map alleles affecting liability to bipolar disorder. For analyses based on haplotypes, they choose the methods of Service et al. (1999). Their results point to several regions of the genome as potentially harboring alleles affecting liability, especially 8p. Still, it would be useful to know under what evolutionary scenarios the methods of Service et al. (1999) are optimal. If our conjecture about the properties of Service et al.'s (1999) methods is correct, a genomewide analysis based on GOF with the DR described herein might identify additional loci.

A common concern for case-control studies is confounding of disease status with genetic information because of population substructure (Devlin et al. 2001*a*, 2001*b*). Similarity and GOF tests using haplotype data from case-control studies are susceptible to confounding, just as single-locus studies are. When a large number of regions are haplotyped, however, there are ways of overcoming the confounding by use of genomic control (Devlin and Roeder 1999) or structured analysis (Pritchard et al. 2000). For example, by using arguments similar to those of Devlin and Roeder (1999), Tzeng et al. (in press) show that, when phase is assumed to be known, the variance inflation due to confounding is approximately constant across the genome; specifically $\mathrm{Var}(T_k) \approx \tau^2\sigma_k^2$, where τ is not a function of the region, k. This convenient form follows because, across the genome, an individual has a common evolutionary and demographic history (in expectation, and under the assumption of no evolutionary selection). As a result, τ can be estimated, provided many regions are sampled. Tzeng et al. (in press) also introduce a false-discovery–rate procedure to determine which regions show significant association.

If only multilocus genotypes are available, the variance of T_k is inflated over σ_k^2, because of phase uncer-

tainty as well as various evolutionary processes. Provided the sample size is not small, for the matching and length measures, a GC analysis can be performed under the assumption that $\mathrm{Var}(T_k) \approx \tilde{\tau}^2\sigma_k^2$, where $\tilde{\tau}^2$ incorporates all sources of the extra multinomial variance. Under this assumption, which appears to hold in our simulations (data not shown), there is no need to perform the bootstrap procedure. In the genomic control setting, the extra variability, formerly captured by the bootstrap experiment, is included in the estimate of the inflation factor $\tilde{\tau}^2$. In the same vein, we are currently considering how to compensate for population structure for GOF tests by using either GC or structured analysis.

Acknowledgments

This research was supported by National Institute of Health grants MH57881 and CA54852-07 and National Science Foundation grant DMS-9803433.

Appendix A

Counting Measure

We show that the numerator (D) of the test statistic for the counting measure can be computed directly from genotype data. This result does not apply to the other two measures.

Let $X_{\{b_1b_2\ldots b_J\}}$ be the haplotype counts of haplotype $\{b_1b_2\ldots b_J\}$, and $b_j \in \{1,2,\ldots,L_j\}$. Let $\hat{\Pi} = (\hat{\pi}_{\{11\ldots1\}},\ldots,\hat{\pi}_{\{b_1b_2\ldots b_J\}},\ldots,\hat{\pi}_{\{L_1L_2\ldots L_J\}})$, and $\hat{\pi}_{\{b_1b_2\ldots b_J\}} = X_{\{b_1b_2\ldots b_J\}}/n$. Let $Y_{\{b_1B_1\times b_2B_2\times\ldots\times b_JB_J\}}$ be the genotype counts of genotype $\{b_1B_1 \times b_2B_2 \times \ldots \times b_JB_J\}$.

Theorem 1. For **A** defined by the counting measure,

$$\hat{\Pi}^{\top}\mathbf{A}\hat{\Pi} = \frac{1}{n^2}\Big[\sum_{b_1}\Big(\sum_{b_2,\ldots,b_J} X_{\{b_1b_2\ldots b_J\}}\Big)^2 + \sum_{b_2}\Big(\sum_{b_1,b_3,\ldots,b_J} X_{\{b_1b_2\ldots b_J\}}\Big)^2 + \cdots + \sum_{b_J}\Big(\sum_{b_1,\ldots,b_{J-1}} X_{\{b_1b_2\ldots b_J\}}\Big)^2\Big]\ , \tag{A1}$$

$$\text{where } \sum_{b_2,\ldots,b_J} X_{\{b_1b_2\ldots b_J\}},\ \sum_{b_1,b_3,\ldots,b_J} X_{\{b_1b_2\ldots b_J\}},\ \cdots \text{ and } \sum_{b_1,\ldots,b_{J-1}} X_{\{b_1b_2\ldots b_J\}} \text{ depend only on genotype counts } Ys\ . \tag{A2}$$

Proof:

We need to show that equation (A1) and statement (A2) are true.

For equation (A1): The right hand side (RHS) of (A1) implies that, for an arbitrary haplotype pair $\{b_1b_2\ldots b_J\}$ and $\{B_1B_2\ldots B_J\}$, the number of alleles that these two haplotypes share in common determines the number of times that $X_{\{b_1b_2\ldots b_J\}}$ and $X_{\{B_1B_2\ldots B_J\}}$ show up in the same summation square $(\sum_{b_j s} X_{\{b_1b_2\ldots b_J\}})^2$. For example, haplotype {136} and haplotype {146} have two alleles in common. Hence, $X_{\{136\}}$ and $X_{\{146\}}$ meet twice, at $(\sum_{b_2,b_3} X_{1b_2b_3\}})^2$ and $(\sum_{b_1,b_2} X_{\{b_1b_26\}})^2$. Rewrite the RHS as sums of product $X_{\{b_1b_2\ldots b_J\}} \times X_{\{B_1B_2\ldots B_J\}}$; the coefficient of $X_{\{b_1b_2\ldots b_J\}} \times X_{\{B_1B_2\ldots B_J\}}$ is the number of times that $X_{\{b_1b_2\ldots b_J\}}$ and $X_{\{B_1B_2\ldots B_J\}}$ meet in the same summation square $(\sum_{b_j s} X_{\{b_1b_2\ldots b_J\}})^2$ multiplied by two. In other words, the coefficient is 2 times the number of alleles shared in common by haplotype $\{b_1b_2\ldots b_J\}$ and haplotype $\{B_1B_2\ldots B_J\}$. Consequently, RHS equals LHS by the design of the matrix **A**.

For statement (A2): Each haplotype count (X) can be calculated as the sum of the corresponding genotype counts (Y). These genotypes can be classified into three types: (i) homozygous genotypes, (ii) singly-heterozygous genotypes, and (iii) all other genotypes. Here we use $J = 3$ as an example to illustrate:

$$X_{\{b_1b_2b_3\}} = \underbrace{2Y_{\{b_1b_1\times b_2b_2\times b_3b_3\}}}_{\text{homozygous}} + \underbrace{Y_{\{b_1B_1\times b_2b_2\times b_3b_3\}} + Y_{\{b_1b_1\times b_2B_2\times b_3b_3\}} + Y_{\{b_1b_1\times b_2b_2\times b_3B_3\}}}_{\text{singly heterozygous}} + w_i \times \underbrace{Y_{\{b_1B_1\times b_2B_2\times b_3B_3\}}}_{\text{other}}\ ,$$

where w_i will be defined shortly. For homozygous and singly heterozygous genotypes, we can reconstruct the corresponding haplotype pairs without ambiguity and hence count haplotype directly. For the "other" part, more

haplotype pairs are possible. Assume in the data that each type of haplotype pair has frequency w_1, w_2, w_3, and w_4, respectively; then we have

$$X_{\{b_1b_2b_3/B_1B_2B_3\}} = w_1 \times Y_{\{b_1B_1 \times b_2B_2 \times b_3B_3\}},$$
$$X_{\{B_1b_2b_3/b_1B_2B_3\}} = w_2 \times Y_{\{b_1B_1 \times b_2B_2 \times b_3B_3\}},$$
$$X_{\{b_1B_2b_3/B_1b_2B_3\}} = w_3 \times Y_{\{b_1B_1 \times b_2B_2 \times b_3B_3\}}, \text{ and}$$
$$X_{\{b_1b_2B_3/B_1B_2b_3\}} = w_4 \times Y_{\{b_1B_1 \times b_2B_2 \times b_3B_3\}}.$$

From genotype data, although we do not know the values of w_i, we know $\sum_i w_i = 1$. On the other hand,

$$\begin{array}{llll} \text{for} & X_{\{b_1b_2b_3/B_1B_2B_3\}}, & X_{\{b_1b_2b_3\}} & = X_{\{B_1B_2B_3\}} \\ \text{for} & X_{\{B_1b_2b_3/b_1B_2B_3\}}, & X_{\{B_1b_2b_3\}} & = X_{\{b_1B_2B_3\}} \\ \text{for} & X_{\{b_1B_2b_3/B_1b_2B_3\}}, & X_{\{b_1B_2b_3\}} & = X_{\{B_1b_2B_3\}} \\ \text{for} & X_{\{b_1b_2B_3/B_1B_2b_3\}}, & X_{\{b_1b_2B_3\}} & = X_{\{B_1B_2b_3\}}. \end{array}$$

Thus, we have

$$\begin{aligned} \sum_{b_2,b_3} X_{\{b_1b_2b_3\}} &= \sum_{b_1,b_3} X_{\{b_1b_2b_3\}} = \sum_{b_1,b_2} X_{\{b_1b_2b_3\}} \\ &= X_{\{b_1b_2b_3/B_1B_2B_3\}} + X_{\{B_1b_2b_3/b_1B_2B_3\}} + X_{\{b_1B_2b_3/B_1b_2B_3\}} + X_{\{b_1b_2B_3/B_1B_2b_3\}} \\ &= Y_{\{b_1B_1 \times b_2B_2 \times b_3B_3\}}. \end{aligned}$$

By the same manner, we have statement (A2).

Appendix B

Variance of *D*

Assume $n\hat{\Pi}_a$ is distributed multinomial $(n;\Pi_a)$ and $m\hat{\Pi}_u$ is distributed multinomial $(m;\Pi_u)$. We compute $\sigma^2(\Pi)$ the variance of $D = \hat{\Pi}_a^{\top}\mathbf{A}\hat{\Pi}_a - \hat{\Pi}_u^{\top}\mathbf{A}\hat{\Pi}_u$ under the null hypothesis $\Pi_a = \Pi_u \equiv \Pi$.

Define $\Sigma_n = 1/n\,[\mathrm{Diag}(\Pi) - \Pi\Pi^{\top}]$, where $\mathrm{Diag}(\Pi)$ is the diagonal matrix of $(\pi_1, \ldots, \pi_R)$, $\mathrm{Diag}\,[\Pi^{(2)}]$ is the diagonal matrix of $(\pi_1^2, \ldots, \pi_R^2)$ and $\Pi^{(2)\top} = (\pi_1^2, \pi_2^2, \ldots, \pi_R^2)$.

Let $\mathbf{A} = [a_{ij}]$, $(d\mathbf{A})^{\top} = (a_{11}, a_{22}, \ldots, a_{RR})$, $\mathbf{A}^{(2)} = [a_{ij}^2]$, and $(d\mathbf{A}^{(2)})^{\top} = (a_{11}^2, a_{22}^2, \ldots, a_{RR}^2)$. Define $(d0\mathbf{A})$ as $\mathbf{A}$ with the diagonal elements zeroed out, $\mathbf{A}_{-l}$ = matrix $(d0\mathbf{A})$ with the lth-row and the lth-column deleted,

$$\mathbf{B} = \begin{bmatrix} a_{11} & a_{22} & \ldots & a_{RR} \\ a_{11} & a_{22} & \ldots & a_{RR} \\ \vdots & \vdots & & \vdots \\ a_{11} & a_{22} & \ldots & a_{RR} \end{bmatrix},$$

and $E^{\top} = (a_{11}\pi_1, a_{22}\pi_2, \ldots, a_{RR}\pi_R)$.

Now $\mathrm{Var}(\hat{\Pi}_a^{\top}\mathbf{A}\hat{\Pi}_a)$ equals

$$\begin{aligned}\mathrm{Var}\left(\hat{\Pi}_a^{\top}\mathbf{A}\hat{\Pi}_a\right) = {} & \frac{(n-1)(n-2)(n-3)}{n^3} \times [\Pi^{\top}\mathbf{A}\Pi]^2 + \frac{(n-1)(n-2)}{n^3} \times \left\{2 \times \text{sum of all elements of}\left[\sum_{l=1}^{R} \pi_l a_{ll} \times \mathbf{A}_{-1}\right]\right. \\ & +4 \times \text{sum of off} - \text{diagonal elements of } \left[\mathrm{Diag}(\Pi)\ \mathbf{A}\ \mathrm{Diag}(\Pi)\ \mathbf{A}\ \mathrm{Diag}(\Pi)\right] \\ & +4 \times \mathrm{tr}\left[(d0\mathbf{A})\ \mathrm{Diag}(\Pi)\ \mathbf{B}\ \mathrm{Diag}(\Pi^{(2)})\right] \\ & \left.+2 \times \Pi^{\top}(d\mathbf{A})(d\mathbf{A})^{\top}\Pi^{(2)} + 4 \times \Pi^{\top}\mathbf{A}^{(2)}\Pi^{(2)}\right\} \\ & +\frac{(n-1)}{n^3} \times \{4 \times E^{\top}\mathbf{A}\Pi + \Pi^{\top}(d\mathbf{A})(d\mathbf{A})^{\top}\Pi + 2 \times \Pi^{\top}\mathbf{A}^{(2)}\Pi\} \\ & +\frac{1}{n^3} \times (d\mathbf{A}^{(2)})^{\top}\Pi - [\mathrm{tr}(\mathbf{A}\Sigma_n) + \Pi^{\top}\mathbf{A}\Pi]^2 \ .\end{aligned}$$

Using the same general form, $\mathrm{Var}(\hat{\Pi}_u^{\top}\mathbf{A}\hat{\Pi}_u)$ is obtained; then $\sigma^2(\Pi) = \mathrm{Var}(\hat{\Pi}_a^{\top}\mathbf{A}\hat{\Pi}_a) + \mathrm{Var}(\hat{\Pi}_u^{\top}\mathbf{A}\hat{\Pi}_u)$.

References

Bourgain C, Génin E, Holopainen P, Mustalahti K, Mäki M, Partanen J (2001) Use of closely related affected individuals for the genetic study of complex diseases in founder populations. Am J Hum Genet 68:154–159

Bourgain C, Génin E, Quesneville H, Clerget-Darpoux F (2000) Search for multifactorial disease susceptibility genes in founder populations. Ann Hum Genet 64:255–265

Cheung VG, Nelson SF (1998) Genomic mismatch scanning identifies human genomic DNA shared identical by descent. Genomics 47:1–7

Clayton DG (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65:1170–1177

Culverhouse R, Lin J, Liu KY, Suarez BK (2001) Exploiting linkage disequilibrium in population isolates. Genet Epidemiol 21 Suppl 1:S429–S434

de la Chapelle A, Wright FA (1998) Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. Proc Natl Acad Sci USA 95:12416–12423

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Devlin B, Roeder K, Bacanu SA (2001*a*) Unbiased methods for population-based association studies. Genet Epidemiol 21:273–284

Devlin B, Roeder K, Wasserman L (2000) Genomic control for association studies: a semiparametric test to detect excess-haplotype sharing. Biostatistics 1:369–387

——— (2001*b*) Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60:155–166

Durham LK, Feingold E (1997) Genome scanning for segments shared identical by descent among distant relatives in isolated populations. Am J Hum Genet 61:830–842

Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947–959

Fan R, Lange K (1998) Models for haplotype evolution in a nonstationary population. Theor Popul Biol 53:184–198

Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, et al (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet 13:399–408

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296: 2225–2229

Grant GR, Manduchi E, Cheung VG, Ewens WJ (1999) Significance testing for direct identity-by-descent mapping. Ann Hum Genet 63:441–454

Houwen RHJ, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA, Freimer NB (1994) Genome screening by searching for shared segments: Mapping a gene for benign recurrent intrahepatic cholestasis. Nat Genet 8:380–386

Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes. Genome Res 10:1435–1444

Lam JC, Roeder K, Devlin B (2000) Haplotype fine mapping by evolutionary trees. Am J Hum Genet 66:659–673

Lee AJ (1990) U-statistics: theory and practice. Marcel Dekker, New York

MacLean CJ, Martin RB, Sham PC, Wang H, Straub RE, Kendler KS (2000) The trimmed-haplotype test for linkage disequilibrium. Am J Hum Genet 66:1062–1075

McPeek MS, Strahs A (1999) Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet 65:858–875

Ophoff RA, Escamilla MA, Service SK, Spesny M, Meshi DB, Poon W, Molina J, Fournier E, Gallegos A, Mathews C,

Neylan T, Batki SL, Roche E, Ramirez M, Silva S, De Mille MC, Dong P, Leon PE, Reus VI, Sandkuijl LA, Freimer NB (2002) Genomewide linkage disequilibrium mapping of severe bipolar disorder in a population isolate. Am J Hum Genet 71:565–574

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–181

Puffenberger EG, Kauffman ER, Bolk S, Matise TC, Washington SS, Angrist M, Weissenbach J, Garver KL, Mascari M, Ladda R (1994) Identity by descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. Hum Mol Genet 3:1217–1225

Qian D, Thomas DC (2001) Genome scan of complex traits by haplotype sharing correlation.Genet Epidemiol 21 Suppl 1:S582–S587

Service SK, Temple Lange DW, Freimer NB, Sandkuijl LA (1999) Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. Am J Hum Genet 64:1728–1738

Thomas DC, Morrison JL, Clayton DG (2001) Bayes estimates of haplotype effects. Genet Epidemiol 21 Suppl 1:S712–S717

Tzeng JY, Byerley W, Devlin B, Roeder K, Wasserman L. Outlier detection and false discovery rates for whole-genome DNA matching. J Am Stat Assoc (in press)

Van der Meulen MA, Te Meerman GJ (1997) Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. Genet Epidemiol 14: 915–920