Hum Genet (2007) 121:549–557
DOI 10.1007/s00439-007-0345-z

ORIGINAL INVESTIGATION

Testing for association based on excess allele sharing in a sample of related cases and controls

Lambertus Klei · Kathyrn Roeder

Received: 21 November 2006 / Accepted: 12 February 2007 / Published online: 7 March 2007

Abstract Samples consisting of a mix of unrelated cases and controls, small pedigrees, and much larger pedigrees present a unique challenge for association studies. Few methods are available for efficient analysis of such a broad spectrum of data structures. In this paper we introduce a new matching statistic that is well suited to complex data structures and compare it with frequency-based methods available in the literature. To investigate and compare the power of these methods we simulate datasets based on complex pedigrees. We examine the influence of various levels of linkage disequilibrium (LD) of the disease allele with a marker allele (or equivalently a haplotype). For low frequency marker alleles/haplotypes, frequency-based statistics are more powerful in detecting association. In contrast, for high frequency marker alleles, the matching statistic has greater power. The highest power for frequency-based statistics occurs when the disease allele frequency closely matches the frequency of the linked marker allele. In contrast maximum power of the matching statistic always occurs for intermediate marker allele frequency regardless of the disease allele frequency. Moreover, the matching and frequency-based statistics exhibit little correlation. We conclude that these two approaches can be viewed as complementary in finding possible association between a disease and a marker for many different situations.

L. Klei
Western Psychiatric Institute and Clinic,
University of Pittsburgh Medical Center,
Pittsburgh, PA 15213, USA

K. Roeder (✉)
Department of Statistics, Carnegie Mellon University,
Pittsburgh 15213, USA
e-mail: roeder@stat.cmu.edu

Introduction

Subjects originally recruited for linkage analysis of a binary trait are often available for follow up association studies. Frequently these samples are supplemented with unrelated cases and controls drawn from the same population to increase the power of the study. Although convenient to collect, few analysis methods are available for samples of complex structure, potentially including a mix of small pedigrees, much larger pedigrees, and unrelated subjects (Göring and Terwilliger 2000, Gordon et al. 2004). When too few closely-related individuals are sampled, the sample structure inhibits effective application of a transmission disequilibrium tests. In particular the unrelated subjects have little utility in this framework. On the other hand, most population-based tests for association expect the subjects to be unrelated. Ignoring family structure can lead to spurious associations (Newman et al. 2001; Abney et al. 2002). Of the few methods of analysis that can analyze disparate samples (e.g., Risch and Teng 1998; Slager and Schaid 2001; Bourgain et al. 2003), it remains an open question how well they work under a variety of conditions. In addition, the field of methods becomes even narrower when the pedigrees are composed of numerous distantly related subjects and pedigrees with inbreeding loops.

For unrelated samples, many tests for association can be grouped under the label ''frequency-based'' because they are based on assessing the difference in allele frequencies between affected and unaffected individuals. For convenience we call the affected individuals 'cases' and the unaffected individuals 'internal controls' or simply 'controls'. Building on the structure of these tests, Bourgain et al. (2003) developed two frequency-based methods, the corrected chi-square method, and the case-control quasi-likelihood (CC-QLS) method, that account for arbitrarily complex, but known, relatedness among subjects. We investigate a method similar to the corrected chi-square that is based on the Hellinger distance between case and control allele frequencies. Our method adjusts for complex relatedness among subjects through allele dropping simulations and takes into account pedigree structure when estimating allele frequencies.

Other statistical procedures have been developed that are based on detecting an excess of matching alleles or haplotypes among cases. Unlike the frequency-based methods, these statistics are based on pairwise comparisons of subjects. Like frequency-based methods, these tests have been developed either for unrelated subjects (te Meerman and van der Meulen 1997; Tzeng et al. 2003) or for pedigrees suitable for a transmission disequilibrium tests (Bourgain et al. 2000, 2001). Here we extend the matching approach to permit samples drawn from complex pedigrees and/or unrelated subjects. When restricted to a single large pedigree, the proposed test is similar to the linkage test known as the affected-pedigree-member method by Weeks and Lange (1988).

In what follows we describe this new matching method for association analysis and conduct a simulation study to assess its performance when the data consists of cases and controls drawn from complex pedigrees. The methods of Bourgain et al. (2003) are also evaluated in the simulations. The structure of the simulations is motivated by a study of the genetics of schizophrenia in the Oceanic population of Palau (Klei et al. 2005; Devlin et al. 2007), but the results should be relevant for many other complex sample structures and diseases.

Methods

Matching statistic

The matching statistic is based on the observation that when a disease allele is in linkage disequilibrium with a marker allele, one expects to observe individuals with matching marker alleles more often than predicted under the null hypothesis of no linkage. To detect this effect we propose to use the difference between the observed and expected identity by state (IBS) for each pair of cases as the test statistic. This matching test statistic, M_k for locus k, can be written as:

$$\mathrm{M}_k = \frac{1}{\binom{n_{\text{case}}}{2}} \sum_{i<j} \left(\mathrm{O}\left(\mathrm{IBS}_{k,ij}\right) - \mathrm{E}\left(\mathrm{IBS}_{k,ij}\right)\right) \tag{1}$$

in which n_{case} is the number of cases, $O(\mathrm{IBS}_{k,ij})$ is the observed IBS of locus k for case pair i,j, and $E(\mathrm{IBS}_{k,ij})$ is the expected IBS of locus k for case pair i,j.

Observed IBS

The observed IBS is a count of the number of alleles that are identical, either by decent or by chance. For an autosomal locus using an additive genetic model this variable takes on the values of 0, 1, or 2. For a sex-linked locus this concept can be extended naturally. Moreover, other measures of scoring that are more sensitive to dominant and recessive genetic models can be accommodated by adjusting the definition of the scoring function used for determining O(IBS) and E(IBS) accordingly.

Expected IBS

Consider an autosomal locus with m alleles. When a pair of individuals is unrelated the expected number of alleles IBS can be determined easily. For each possible allele pair, multiply the observed IBS by the probability of this allele combination occurring then sum this quantity over all $(m(m+1)/2)^2$ possible allele pairs. Formally this is:

$$E\left(\mathrm{IBS}_{k,ij}\right) = \sum_{\substack{a_1 \le a_2 \\ a_3 \le a_4}}^{m} P(a_1a_2, a_3a_4) O(\mathrm{IBS}(a_1a_2, a_3a_4)) \tag{2}$$

in which a_1a_2, is the index pair of alleles for person i, a_3a_4 , is the index pair of alleles for person j, $P(a_1a_2, a_3a_4)$ is the probability that combination a_1a_2, a_3a_4 occurs by chance, and $\mathrm{O}(\mathrm{IBS}(\mathrm{a}_1a_2, a_3a_4))$ is the observed number of IBS alleles in the combination a_1a_2, a_3a_4.

This calculation extends naturally to situations in which individuals are related. Related individuals are expected to have larger numbers of alleles in common because they share common ancestors. To account for this we use the identical by descent (IBD) probabilities defined by Li (1955). For a pair of individuals i and j, let $\Pi_{ij} = \begin{pmatrix} \pi_{0,ij} & \pi_{1,ij} & \pi_{2,ij} \end{pmatrix}$ be the vector indicating the probability of individuals i and j being IBD at 0, 1 or 2 alleles. For many common relationships such as full-sibs, half-sibs, and parent-offspring these probabilities are well known. However, for more complex pedigrees, including those with inbreeding loops, these quantities are hard to calculate. We therefore opted to approximate the IBD probabilities using an allele dropping simulation. In this experiment each founder was assigned two distinct alleles such that the total number of distinct alleles was twice the number of founders. After dropping the alleles through the pedigree, assuming Mendelian inheritance to assign alleles to children based on their parents' genotypes, we determined the IBD status between all pairs of individuals. By repeating this process 100,000 times we were able to obtain an accurate estimate of Π_{ij}.

The probability of a certain allele combination a_1a_2 and a_3a_4 for i and j depends upon the allele frequencies and the IBD status. It is computed conditionally on the relatedness, Π_{ij}, of individuals i and j.

$$\begin{aligned} P\left(a_1a_2, a_3a_4|\Pi_{ij}\right) &= \sum_{q=0}^{2} P(a_1a_2, a_3a_4|\mathrm{IBD}=q)P(\mathrm{IBD}=q) \\ &= \sum_{q=0}^{2} P(a_1a_2, a_3a_4|\mathrm{IBD}=q)\pi_{q,ij} \\ &= P(a_3a_4|a_1a_2,\mathrm{IBD}=q)P(a_1a_2)\pi_{q,ij}. \end{aligned} \tag{3}$$

Derivation of the quantities used in this equation can be found in Appendix I.

The expected IBS for an arbitrary pair of individuals i and j for locus k can then be determined using the expression for the conditional probability in (3) in the expression for the probability in (2) to obtain:

$$E\left(\mathrm{IBS}_{k,ij}\right) = \sum_{\substack{a_1 \leq a_2 \\ a_3 \leq a_4}}^{m} \left(\left(\sum_{q=0}^{2} P(a_1a_2, a_3a_4|\mathrm{IBD}=q)\,\pi_{q,ij}\right) O(\mathrm{IBS}(a_1a_2, a_3a_4))\right) \tag{4}$$

Notice, unrelated individuals are just a special case within the general framework of this statistic. Simulations were used to determine the significance of the resulting test statistic.

Sex-linked loci

Sex-linked loci affect the IBD probabilities, and the effect is not easily expressed in closed form for a general pedigree. Indeed, Π_{ij} is even dependent on the distribution of males and females in the pedigree. Again allele dropping simulations facilitate the estimation of Π_{ij}. In this case we take into account the specific inheritance patterns associated with sex-linked loci. Though we do not use sex-linked loci in the upcoming simulation study, this feature is implemented in our software.

Frequency based statistics

In contrast to the matching method, which looks for unusual levels of matching among pairs of cases, frequency-based statistics look for differences in allele frequencies between cases and controls. The corrected chi-square method and the CC-QLS method of Bourgain et al. (2003) both adjust the test statistics for relatedness among subjects. We describe a related method that is based on the Hellinger distance between case and control allele frequencies.

Hellinger Distance

The Hellinger distance test statistic for locus k (H_k) uses the difference in the square root of allele frequencies between cases and controls to determine association between a marker and a disease. The test statistic is defined as:

$$H_k = \sqrt{\sum_{\ell=1}^{m} \left(\sqrt{f_\ell(\mathrm{case})} - \sqrt{f_\ell(\mathrm{control})}\right)^2} \tag{5}$$

in which $f_\ell(\mathrm{case})$ is the frequency of the ℓth allele in the cases, and $f_\ell(\mathrm{control})$ is the frequency of the ℓth allele in the controls.

A large positive value for H_k, determined by simulation, indicates that a marker is associated with the disease. By its design the Hellinger distance puts more weight on differences in rare allele frequencies than the chi-square statistic. In the scenarios investigated with our simulations, however, the Hellinger distance statistic exhibited a near perfect correlation with the corrected chi-square method (results not show).

Estimating allele frequencies

Each of the three methods we investigate requires estimates of allele frequencies. These were determined using BLUE as described in McPeek et al. (2004). This method utilizes the relationships among genotyped individuals to determine an unbiased estimate for the allele frequencies in the founders. When many genotyped subjects are related, BLUE frequencies can differ substantially from estimates based on counting, the naïve estimates. The frequency-based methods require such estimates of the allele frequencies for both cases and controls. The matching method requires an estimate of the population allele frequencies to compute E(IBS). When cases are over-sampled, we suggest using only the controls for this estimate.

Determining significance

Asymptotic P values are typically valid only if rare alleles are combined: a common rule-of-thumb is to ensure that minor allele counts are at least five. The advantage of using simulations to determine P values, rather than asymptotic distribution theory, is that rare alleles can be retained in the analysis. This is appealing because in studies based on complex pedigrees, the sample size is often small and hence many alleles would be classified as rare. To determine the significance of our simulation test statistics we used allele dropping simulations for all three methods studied. In these simulations, we assumed a null model of no linkage between marker allele and disease allele. Founders were randomly assigned STR alleles based on the estimated allele frequencies, and Mendelian inheritance rules were used to assign alleles to children based on parental genotypes. In the simulations to determine significance we kept the disease status of the individuals the same as they were in the population of interest. In other words, the disease status and population structure were considered fixed while the genotypes of the founders were treated as random. We used at most 1,000 repetitions to determine the empirical P value of each test statistic. (For efficiency, we stopped the allele dropping as soon as we observed 100 simulated test statistics greater than the observed test statistic, ensuring that a P value of less than 0.1 could no longer be achieved. Thus our P values were more precise if they were small). For data analysis we suggest using 10,000 repetitions.

Simulations

Although our method is applicable to any type of genetic marker, and in special cases for haplotypes, for our simulations we decided to use short tandem repeats (STRs). While STRs are useful markers for association studies in their own right, they also serve as a convenient proxy for haplotypes. The distribution of STR alleles in a population is typically quite similar to that of haplotypes composed of single nucleotide polymorphisms (SNPs) in substantial LD. Both have one or two common alleles and several rare ones. Indeed, assuming little or no recombination in designated regions, haplotypes can be used interchangeably with STRs in a simulation setting. Allele frequencies were based on a set of STRs genotyped as part of a schizophrenia linkage study (Klei et al. 2005). Specifically we took the 50 STRs on chromosomes 1 and 2. Among these 50 STR the number of alleles per STR varied between 3 and 11, while allele frequencies varied between 0.01 and 0.80. A disease with a low prevalence ($K = 0.02$) was simulated. Three disease allele frequencies were selected ($P = 0.10, 0.30, 0.50$). The odds ratio (OR) of the disease was chosen to be 2.0, 3.0 or 5.0 for an additive genetic model producing the penetrances reported in Table 1. An STR allele was chosen such that it was in linkage disequilibrium with the disease allele using three different values for Lewontin's D' (Lewontin 1964) ($D' = 0.500$; 0.707; 1.000). We chose 0.500 and 0.707 since they match r^2 of 0.25 and 0.50, respectively, in situations in which the allele frequency of the disease allele equals the allele frequency of the linked STR allele.

To simulate realistic situations the linked STR allele was randomly chosen based on STR allele frequencies. Using the formulae for D' we then created haplotypes for the disease alleles and STRs alleles that were randomly assigned to the founders in our sample pedigree in Fig. 1 based on the haplotype frequencies. Haplotypes were subsequently assigned to children based on Mendelian inheritance rules ignoring possible recombination. Finally, case-control status was assigned to children in the last generation using the penetrance model described previously. With rejection sampling we generated a total of 70 families with at least two affected individuals in the last generation (individuals labeled 16–25 in Fig. 1). This type of data structure presented us with related (within family) as well as unrelated (across families) cases and controls. A further assumption was made that individuals in earlier generations were not diagnosed or genotyped. We also assumed that all affected individuals were genotyped and that of the unaffected individuals in the last generation, 65% were genotyped to serve as controls. This resulted in datasets with ~150 genotyped cases and ~350 genotyped controls. These values matched the number of cases and controls in our schizophrenia study (Devlin et al. 2007; Klei et al. 2005). This process was repeated 20 times for each of the 50 STRs, resulting in a total of 1,000 repetitions for each combination of p, OR, and D'.

Results

Type I error rate

To obtain the Type I error rate, we simulated an odds ratio of one and Lewontin's D' of one. The results show that, for the frequency-based statistics, the nominal Type I error rate is within the confidence range of the observed Type I errors (Table 2). For the matching statistic Type I error rates are slightly inflated, especially at $\alpha = 0.050$. To determine if this bias was associated with the relatively small size of the simulated population, we doubled the number of families in the simulation. Results, not shown here, indicate that for larger populations the observed Type I error rate for the matching statistic is at the nominal rate.

Power

Table 3 shows the power to detect the disease alleles for different levels of linkage disequilibrium when $\alpha = 0.01$. As expected, this Table shows that the power increases for all three detection methods when either the level of D' or the odds ratio (OR) increases, regardless of disease allele frequency (P) (Table 3). In all cases studied, except for $P = 0.10$ and OR = 2.0, Hellinger distance outperforms CC-QLS, whereas certain combinations of P, OR and D' show more power for Matching than Hellinger distance and vice versa. The matching statistic shows its strongest relative performance when the disease and STR allele are not in tight linkage disequilibrium ($D' < 1$) and the disease allele frequency is intermediate to high (0.30 and 0.50). This advantage appears to be more pronounced for lower values of OR. For all three methods and for OR up to 3, the maximum power occurs for a disease allele with a frequency of 0.30.

Correlations

Table 4 illustrates the correlation between the P values for the three detection methods for the different scenarios we explored. The high correlation between the two frequency-based methods, Hellinger distance and CC-QLS, indicates that in many cases these two methods are nearly interchangeable. On the other hand, the correlation between the matching and frequency-based methods is always less than 0.50, often substantially smaller. This suggests that the matching statistic is sensitive to different characteristics of the association than the frequency-based methods.

Linked STR allele frequency

We investigated the effect of varying odds ratio by examining the power of the three methods with linkage disequilibrium (D') set at 0.707, a disease allele with frequency 0.10, and a linked allele frequency varying between 0.01and 0.80 (Fig. 2a). Clearly frequency-based methods are more powerful when the disease allele is linked to a low frequency STR allele (0.01–0.30) while the matching statistic shows more power for the other situations. This result is consistent across all parameter combinations that we investigated.

We also investigated the performance patterns across varying disease allele frequencies (Fig. 2b). As seen in Fig. 2a, the frequency-based methods are more powerful at low allele frequencies while the matching statistic is more powerful at high frequencies. However, the more interesting result is revealed by looking at peak power for each of the methods (Fig. 2b). Whereas the matching statistic achieves its peak consistently in the 0.4–0.6 allele frequency range, the frequency-based methods achieve their peak at a frequency depending on the allele frequency of the disease allele. In particular, the peak power for the frequency-based methods occurs when the correlation between the disease allele and the linked STR allele is maximized, and this occurs when the STR allele frequency matches the disease allele frequency. Again, results shown here are indicative of patterns observed in the other scenarios we investigated.

Discussion

Few analysis methods are available when studying association in samples of complex structure, including large families and unrelated subjects. Samples of this structure are not suitable for traditional tests of transmission disequilibrium or population-based tests of association. Moreover, of the methods of analysis available for complex samples, little is known about which work well under various conditions. We explore the performance of existing methods and compare them to a new ‘‘matching’’ method that is similar in spirit to both the association test proposed by Tzeng et al. (2003) and the linkage test proposed by Weeks and Lange (1988).

If the matching method were applied to a single large pedigree, then it could detect a signal based purely on linkage information. Indeed, it would perform similarly to the affected-pedigree-member method of Weeks and Lange. Data of the structure we envision, however, possesses information about both linkage and association. Linkage is apparent when pairs of related individuals match more than expected. Association is detected when clusters of individuals (related or not) match, leading to more pair-wise matching than expected. The proposed test capitalizes on both forms of information to good effect.

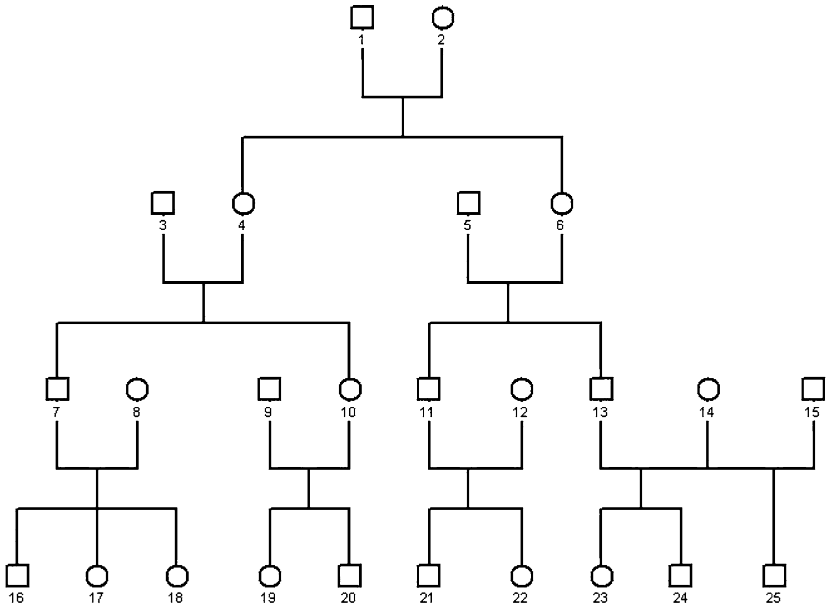
Frequently in association studies, concern exists about false positives due to population stratification. The effect of population substructure and unrecorded relatedness can be modeled using genomic control (Devlin and Roeder 1999). The effect of relatedness can be more effectively modeled

Table 1 Penetrances for the different values of the disease allele frequency (P) and odds ratio (OR) assuming a constant disease prevalence of $K = 0.02$ and an additive genetic model

P	OR	$f2$	$f1$	$f0$
0.10	2.0	0.0517	0.0341	0.0165
	3.0	0.0775	0.0456	0.0136
	5.0	0.1177	0.0634	0.0091
0.30	2.0	0.0410	0.0260	0.0110
	3.0	0.0542	0.0298	0.0054
	5.0	–[a]	–	–
0.50	2.0	0.0331	0.0200	0.0069
	3.0	0.0397	0.0200	0.0003
	5.0	–	–	–

[a] The combination of $P = 0.30$ or 0.50 and OR = 5.0 resulted in invalid penetrance values for $f2$

Fig. 1 Base pedigree used for the simulations



by using methods that overtly incorporate the known pedigree structure. In the methods we explore here we do not adjust for population structure. It is assumed that the sample is obtained from a homogeneous population such as the one motivating our study (Devlin et al. 2007). Indeed, in Devlin et al. (2007) we sample from a defined, island population, and the relationships among the affected and unaffected individuals can be quite complex.

Association tests are most often performed on three major types of genetic data: SNPs, STRs and haplotypes. We simulated STR markers rather than SNPs for two reasons. STR markers mimic haplotype blocks very effectively, providing a glimpse into two types of data structures in one simulation study. Moreover, the simulations correspond to a motivating study (Devlin et al. 2007). In an attempt to model a complex disease realistically we varied a number of factors in our simulations, including the frequency of the disease allele, the frequency of the linked allele under investigation, the linkage disequilibrium between the disease allele and the linked allele, and the penetrance model.

In our simulations of the frequency-based methods we found the Hellinger method performs almost identically to the corrected chi-square method of Bourgain et al. (2003), especially when we combined alleles with low frequency. Therefore we treat the two as interchangeable in our discussion. A more surprising finding is that the Hellinger method outperformed CC-QLS, the other method described by Bourgain et al. (2003), in every scenario in our study. This result is contrary to the findings of Bourgain et al. (2003). Presumably the different results are due to exploring different genetic marker types, relatedness, and classes of alternatives. In particular, Bourgain et al. (2003) simulated SNP markers rather than STRs and used a sta-

Table 2 Type I error rates and standard errors for the three methods for different disease models based on 1,000 replicates

α	0.050			0.010			0.001		
Method	*M*[a]	*H*	*C*	*M*	*H*	*C*	*M*	*H*	*C*
P[b] = 0.10	0.050	0.057	0.057	0.020	0.011	0.017	0.007	0.003	0.005
SE[c]	0.007	0.007	0.007	0.004	0.003	0.004	0.003	0.002	0.002
P = 0.30	0.075	0.059	0.062	0.020	0.013	0.007	0.005	0.003	0.002
SE	0.008	0.007	0.008	0.004	0.004	0.003	0.002	0.002	0.001
P = 0.50	0.086	0.031	0.039	0.016	0.006	0.005	0.006	0.003	0.000
SE	0.008	0.005	0.006	0.004	0.002	0.002	0.002	0.002	0.001

[a] *M* Matching, *H* Hellinger distance, *C* Case-control quasi likelihood

[b] Disease allele frequency

[c] Standard error of the Type I error rate

Table 3 Power at α = 0.01 for different linkage disequilibrium (D') values and disease models

P[a]	OR[b]	D' = 0.500			D' = 0.707			D' = 1.00		
		M[c]	H	C	M	H	C	M	H	C
0.10	2.0	0.019	0.028	0.029	0.036	0.038	0.045	0.103	0.112	0.195
	3.0	0.065	0.101	0.070	0.163	0.215	0.197	0.339	0.492	0.432
	5.0	0.211	0.316	0.308	0.456	0.620	0.585	0.794	0.856	0.814
0.30	2.0	0.081	0.069	0.061	0.165	0.145	0.131	0.404	0.395	0.354
	3.0	0.231	0.212	0.189	0.501	0.479	0.420	0.665	0.734	0.709
0.50	2.0	0.086	0.053	0.038	0.180	0.098	0.085	0.325	0.288	0.258
	3.0	0.201	0.122	0.106	0.350	0.315	0.270	0.494	0.534	0.508

[a] Disease allele frequency

[b] Odds ratio

[c] *M* Matching, *H* Hellinger distance, *C* Case-control quasi likelihood

tistical parameterization rather than a penetrance model to calibrate their alternatives. In addition, they assumed the causal SNP was the one under investigation and a sample large enough to warrant asymptotic distribution theory. Thus our simulation studies differed markedly in implementation. The two frequency-based methods were highly correlated in our simulations, so perhaps the difference in performance is of minor practical importance.

Of greater interest is the difference between the frequency-based methods and the matching method. Previous research using a similar form of matching-based tests designed for a sample of unrelated cases and controls, suggested that frequency-based statistics show relatively more power for low frequency alleles and matching-based statistics work well when the disease allele frequencies are relatively common (Tzeng et al. 2003). In that study, the marker allele was assumed to be the causal allele. Our simulations explore a much larger space of genetic alternatives and confirm that their findings extend to the methods adapted for association studies with relatives. Specifically, we find that frequency-based methods are more powerful when the disease allele is linked to a low frequency STR allele and matching statistic shows more power when the disease allele is linked to a common allele. The matching and frequency-based approaches have low correlation suggesting that it is often worthwhile to analyze a given data set with both approaches.

A related question concerns the similarity between the marker and causal allele frequencies. Surprisingly, the matching statistic achieves its peak power when the marker allele frequency is relatively large, regardless of the disease allele frequency, whereas the frequency-based methods achieve their peak power when the allele frequencies of the disease and marker alleles match closely. These results are obtained for a fixed level of D'.

Although results are not displayed, we found that the matching method has low power relative to the frequency-based methods when the analysis is based on SNPs, unless

Table 4 Correlations between detection methods for different disease model

P[b]	OR[c]	D' = 0.500[a]			D' = 0.707			D' = 1.000		
		MH[d]	MC	HC	MH	MC	HC	MH	MC	HC
0.10	2	0.03	0.00	0.86	0.09	0.03	0.87	0.02	–0.03	0.87
	3	0.05	0.02	0.88	0.04	0.00	0.89	0.05	–0.00	0.90
	5	0.02	-0.01	0.89	0.09	0.05	0.89	0.10	0.03	0.84
0.30	2	0.21	0.16	0.86	0.26	0.22	0.88	0.31	0.25	0.90
	3	0.31	0.27	0.88	0.31	0.27	0.91	0.37	0.33	0.86
0.50	2	0.19	0.14	0.84	0.30	0.25	0.87	0.40	0.34	0.90
	3	0.37	0.29	0.88	0.40	0.35	0.88	0.46	0.43	0.90

[a] Lewontin's linkage disequilibrium

[b] Disease allele frequency

[c] Odds ratio

[d] *MH* correlation between matching and hellinger distance, *MC* correlation between matching and case-control quasi likelihood, *HC* correlation between hellinger distance and case-control quasi likelihood

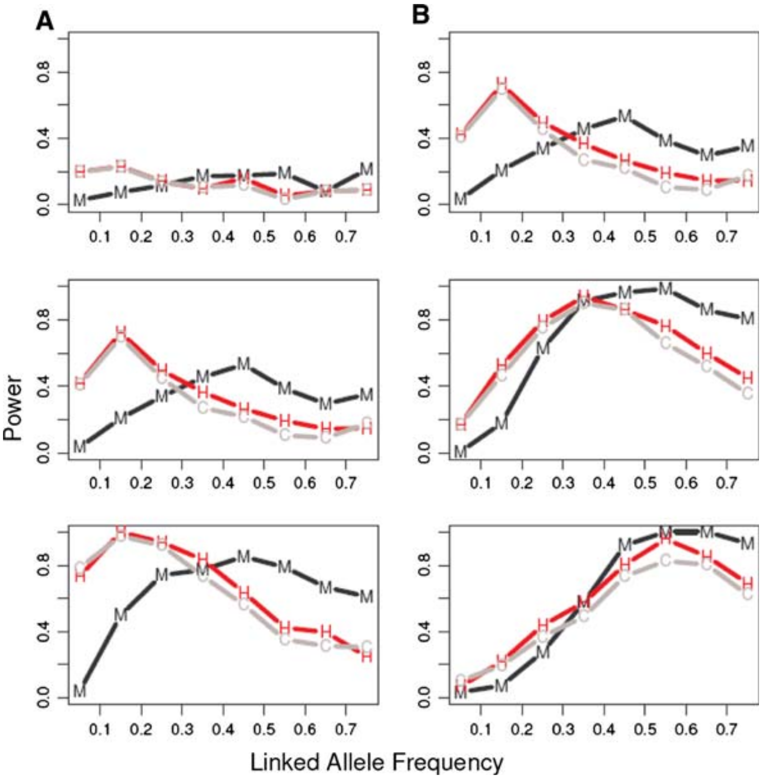


Fig. 2 Power of three approaches (*M* Matching statistic, *H* Hellinger distance, *C* Case-control quasi-likelihood) plotted as a function of linked allele frequency, for $D' = .707$ and $\alpha = 0.05$. For panel **a**, disease allele frequency = 0.1 and odds ratio = 2 (*top*), 3 (*middle*) and 5 (*bottom*). For panel **b**, odds ratio = 3 and disease allele frequency = 0.1 (*top*), 0.3 (*middle*) and 0.5 (*bottom*)

the disease is linked to the common allele and the penetrance is very low. This finding makes intuitive sense. The matching method automatically reduces the degrees of freedom to one, regardless of the number of alleles present. With SNP markers the degrees of freedom are already one, and so there is no benefit to be gleaned in terms of parsimony. When SNP markers are used to form haplotypes, however, the results should be similar to those displayed for STRs.

In our simulations we used allele dropping experiments to obtain *P* values for all methods explored. We find this to be a simple and powerful approach that is superior to using asymptotic *P* values. With this approach it is not necessary to pool rare alleles. Of course, one may want to combine rare alleles when using the frequency-based approaches to reduce the degrees of freedom and potentially increase power. Another advantage of allele dropping is that the software allows for missing data, which is almost certain to be present in practice. Finally, allele dropping allows for an easy extension to markers on the X chromosome. Although ignoring the family structure in the sample can lead to spurious associations (Newman et al. 2001; Abney et al. 2002), we found that all of the tests we investigated that obtain their *P* values through allele dropping were properly calibrated and obtained the desired Type I error rate.

In summary, we introduce a new matching statistic and compare it with frequency-based methods that are applicable to samples consisting of a mixture of unrelated cases and controls, and pedigrees. We find that, for high frequency marker alleles, matching statistics are more powerful in detecting association. For low frequency marker alleles the frequency based statistics are more powerful. Moreover, the relative performance of frequency-based statistics and matching statistics depends on how closely the disease allele frequency matches the frequency of the linked marker allele, a feature of the data that cannot be examined in practice. Overall, the matching and frequency-based statistics exhibit surprisingly little correlation. We conclude that these two approaches can profitably be used in tandem to optimize the chance of finding association between a disease and a marker allele.

Acknowledgments Research supported by NIH grants MH057881 and MH063295. The authors would like to thank Bernie Devlin for many valuable suggestions to improve this manuscript. Fortran95 source code for software used in this study is available on request from Lambertus 'Bert' Klei (kleil@upmc.edu).

Appendix I. Derivation of probabilities needed for calculating expected IBS

For this derivation assume that the loci are an ordered pair (locus a_1 (a_2 and a_3 (a_4) and that one is dealing with an autosomal locus. Furthermore let $f(a_i)$ be the frequency of allele *i*.

$$P(a_1a_2) = \begin{cases} f(a_1)^2 & \text{when } a_1 = a_2 \\ 2f(a_1)f(a_2) & \text{when } a_1 \neq a_2 \end{cases}$$

$$P(a_3a_4|a_1a_2, \text{IBD} = 0) = \begin{cases} f(a_3)^2 \text{ when } a_3 = a_4 \\ 2f(a_3)f(a_4) \text{ when } a_3 \neq a_4 \end{cases}$$

$$P(a_3a_4|a_1a_2, \text{IBD} = 1)$$
$$= \begin{cases} f(a_3) \text{ when } a_1 = a_2 \text{ and } a_1 = a_4 \\ f(a_4) \text{ when } a_1 = a_2 \text{ and } a_1 = a_3 \\ f(a_3)/2 \text{ when } a_1 \neq a_2 \text{ and } a_1 = a_4 \\ f(a_4)/2 \text{ when } a_1 \neq a_2 \text{ and } a_1 = a_3 \\ f(a_3)/2 \text{ when } a_1 \neq a_2 \text{ and } a_2 = a_4 \\ f(a_4)/2 \text{ when } a_1 \neq a_2 \text{ and } a_2 = a_3 \\ (f(a_1) + f(a_2))/2 \text{ when } a_1 \neq a_2 \text{ and } a_1 = a_3 \text{ and } a_2 = a_4 \\ 0 \text{ otherwise} \end{cases}$$

$$P(a_3a_4|a_1a_2, \text{IBD} = 2) = \begin{cases} 1 \text{ when } a_1 = a_3 \text{ and } a_2 = a_4 \\ 0 \text{ otherwise} \end{cases}$$

Specifics on the probabilities needed for calculating expected IBS for sex-linked loci can be obtained by request from the authors.

References

Abney M, Ober C, McPeek MS (2002) Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting seruminsulin level in the Hutterites. Am J Hum Genet 70:920–934

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

Bourgain C, Genin E, Holopainen P, Mustalahti K, Maki M, Partanen J, Clerget-Darpoux F (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, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C, McPeek MS (2003) Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. Am J Hum Genet 73:612–626

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

Devlin B, Klei L, Myles-Worsley M, Tiobech J, Otto C, Byerley W, Roeder K (2007) Genetic liability to schizophrenia in Oceanic Palau: a search in the affected and maternal generation. Human Genetics (under review)

Gordon D, Haynes C, Johnnidis C, Patel SB, Bowcock AM, Ott J (2004) A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping error and nay number of untyped parents. Eur J Hum Genet 12:752–761

Göring HHH, Terwilliger JD (2000) Linkage analysis in the presence of errors IV: Joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. Am J Hum Genet 66:1310–1327

Klei L, Bacanu SA, Myles-Worsley M, Galke B, Xie W, Tiobech J, Otto C, Roeder K, Devlin B, Byerley W (2005) Linkage analysis of a completely ascertained sample of familial schizophrenics and bipolars from Palau, Micronesia. Hum Genet 117:349–356

Lewontin RC (1964) The interaction of selection and linkage. I. General Considerations; Heterotic Models. Genetics 49:49–67

Li CC (1955) Population genetics. University of Chicago Press, Chicago

McPeek MS, Wu X, Ober C (2004) Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics 60:359–367

Newman DL, Abney M, McPeek MS, Ober C, Cox NJ (2001) The importance of genealogy in determining genetic association in complex inheritance. Am J Hum Genet 69:1146–1148

Risch N, Teng J (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling. Genome Res 8:1273–1288

Slager SL, Schaid DJ (2001) Evaluation of candidate genes in case-control studies: a statistical method to account for related subjects. Am J Hum Genet 68:1457–1462

te Meerman GJ, Van der Meulen MA (1997) Genomic sharing surrounding alleles identical by descent: effects of genetic drift and population growth. Genet Epidemiol 14:1125–1130

Tzeng JY, Devlin B, Wasserman L, Roeder K (2003) On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet 72:891–902

Weeks DE, Lange K (1988) The affected-pedigree-member method of linkage analysis. Am J Hum Genet 42:315–326