Report

Testing Linkage Disequilibrium in Sibships

Kimberly D. Siegmund, Bryan Langholz, Peter Kraft, Duncan C. Thomas

Department of Preventive Medicine, University of Southern California, Los Angeles

We describe the use of multivariate regression for testing allelic association in the presence of linkage, using marker genotype data from sibships. The test is valid, provided that the correct mean structure is modeled but does not require the correlation structure within families to be specified. The test can be implemented using standard statistical software such as the SAS programming language. In a simulation study, we evaluated this new test in comparison with one from a standard, matched–case-control analysis. First, we noted that the genetic effect needed to be quite extreme before residual familial correlation due to linkage led to false inference using the standard, matched-pair analysis. Second, we showed that under examples of extreme residual familial correlation, the new test had the correct test size. Third, we found that the test was more powerful than the sibship disequilibrium test of Horvath and Laird. Finally, we concluded that although the standard analysis may lead to correct inference for practical purposes, the new test is valid, even under extreme residual familial correlation and with no cost in power at the causal locus.

Recently, many articles have discussed methods for the detection of linkage disequilibrium in families for the fine mapping of disease genes. The approaches combine the information from linkage of a trait locus and marker locus in families and the association between marker alleles and trait alleles in the population. Since these methods depend on within-family associations between the marker locus and the disease locus, they are robust to confounding, because of population admixture.

Family designs proposed for studying linkage disequilibrium include case-parent trios and case–sib-control pairs. Case-parent trios are useful for the study of early-onset diseases for which the parents are still available for genotyping, but alternate designs are necessary for the study of late-onset diseases. This recognition led to recommendations for the case–sib-control design (Curtis 1997; Schaid and Rowland 1998; Spielman and Ewens 1998). Conditional logistic regression, the standard epidemiologic method for the analysis of matched case-control data, was proposed for the analysis of the marker genotypes of the case–sib-control pairs. The sib transmission/disequilibrium test (S-TDT) was proposed independently for testing a single diallelic locus. Schaid and Rowland (1998) noted that the S-TDT was equivalent to a score test from the conditional likelihood having log-additive effects of the marker alleles.

Although conditional logistic regression can be applied to sibships of arbitrary size for testing linkage in the presence of allelic disequilibrium, it is not always valid for testing disequilibrium in the presence of linkage. The conditional likelihood assumes that disease status in sibships is conditionally independent, given the sib marker data. This assumption is violated when there is linkage between a disease and a marker locus, since sibs with the same disease status will tend to share the same marker alleles. As a result, the variance for the score test from the usual conditional likelihood is underestimated, and the test for association is liberal.

Recently, several authors have proposed new methods of testing for association in the presence of linkage (Curtis 1997; Horvath and Laird 1998). Curtis (1997) suggested reducing large sibships to a single case-control pair and analyzing them using standard conditional logistic regression. He proposed selecting the sib control with the maximally different marker genotype to that of the case. When more than one sib was affected, he suggested selecting an affected sib at random. Although this method resulted in a valid test statistic, many data

Received March 27, 2000; accepted for publication May 11, 2000; electronically published May 30, 2000.

Address for correspondence and reprints: Dr. Kimberly D. Siegmund, Department of Preventive Medicine, University of Southern California, 1540 Alcazar Street, Suite 220, Los Angeles, CA 90089-9011. E-mail: kims@usc.edu

0002-9297/2000/6701-0029$02.00

were not used. In response, Horvath and Laird (1998) developed a method that would use the data from all siblings. Their approach reduced the marker genotype data in sibships using a sign test of whether the average number of a specific marker allele carried by the affected sibs differed from the average number carried by the unaffected sibs. Again, the data reduction method may have resulted in loss of information.

We propose to apply multivariate regression for correlated outcome data to the analysis of marker data in sibships. We will test for allelic association using a Wald test with a robust variance estimate that takes into account the correlation in outcome from the multivariate (clustered) data. This has the benefit of using the marker data on all sibs and does not require that the exact correlation structure be specified. The method is analogous to the use of generalized estimating equations for the conditional logistic likelihood and can be applied using any software package that will allow one to compute the leverage residuals under the discrete logistic model. We describe the likelihood and the procedure for data analysis using the SAS programming language (SAS Institute).

The sibships contributing to the likelihood are those containing at least one affected and one unaffected sibling. Let $i = 1,\ldots,I$ denote sibship, let $\boldsymbol{D}_\mathrm{i}$ denote the set of affected siblings in sibship *i,* and let n_i denote the number of affected sibs. We let $\boldsymbol{M}_\mathrm{i}$ denote the marker genotypes in the *i*th sibship and $\mathbf{Z}_\mathrm{i}$ their coding. For the purpose of illustration, suppose we have a single diallelic marker locus. For individuals carrying 0, 1, or 2 copies of the variant allele, Z takes on the values 0, $\triangle$, or 1, respectively. The parameter $\triangle$ allows us to model the dominance effect ($\triangle = 1$ for dominant, $\triangle = 0$ for recessive, and $\triangle = ½$ for [log-] additive). Then, the conditional likelihood is

$$L(\beta) = \prod_{i=1}^{I} \Pr(\mathrm{D}_i | n_i, \mathrm{Z}_i) = \prod_{i=1}^{I} \frac{\prod_{j \in \mathrm{D}_i} \exp(Z_{ij}'\beta)}{\sum\limits_{\mathrm{S} \in \mathrm{C}_i} \prod_{j \in \mathrm{S}} \exp(Z_{ij}'\beta)} \ ,$$

where $\mathbf{C}_\mathrm{i}$ denotes the set of all possible subsets for which n_i-affected sibs are sampled from the *i*th sibship and β the log-odds ratio of disease in subjects carrying two copies of the variant allele relative to subjects carrying zero copies. A marker locus with *m* marker alleles can be studied by creating *m-1* independent variables. In that case, β is a vector of *m-1* regression coefficients. More general models can be considered by creating dummy variables for each unique genotype.

We propose to fit the conditional logistic regression model in SAS using the procedure PHREG. Using *dfbeta* residuals from this fit, we compute the robust variance estimate described by Therneau and Hamilton (1997). This robust estimate approximates a grouped jackknife estimate of variance where the groups are defined by the independent sibships. It accounts for the correlation in disease status among sibs sharing the same marker alleles, which arises under linkage between the marker and the trait locus. Using this robust variance estimate, we compute a robust Wald test that is valid for testing association in the presence of linkage. The variable definitions and SAS code are given in the Appendix.

We performed a limited simulation study to evaluate the properties of our new test with regard to test size and power. A preferred test would account for residual correlation due to linkage when correlation is present and would show no cost in power when such correlation is not present. These two situations can be distinguished as follows. Suppose we have two distinct loci and we are testing whether alleles at a measured marker locus are in linkage disequilibrium with alleles at a nearby unmeasured disease susceptibility locus (scenario 1). In this case, we have a marker locus that has no direct effect on disease ($\beta = 0$) but that is linked to a trait locus (recombination fraction [θ] < ½). The trait locus has a disease susceptibility allele with population frequency between 0 and 1 and an effect on disease measured by β_g ($\beta_\mathrm{g} \neq 0$). We test the null hypothesis of linkage equilibrium between the marker and trait alleles (disequilibrium measure [δ] = 0) versus the alternative that the alleles are in disequilibrium. Using the conditional likelihood, this is equivalent to testing H_0: $\beta = 0$ ($\delta = 0$, $\beta_\mathrm{g} \neq 0$, $\theta < ½$) vs. H_A: $\beta \neq 0$ ($\delta \neq 0$, $\beta_\mathrm{g} \neq 0$, $\theta < ½$), where linkage disequilibrium will result in a nonzero effect at the linked marker locus. For this scenario, residual familial correlation, resulting from linkage of the marker locus and the trait locus, may necessitate the use of a robust variance estimator for making valid inference. On the other hand, suppose we have a single causal locus and the marker allele is the actual disease susceptibility allele (scenario 2). For this case, the associated marker allele has the same population frequency as the trait susceptibility allele and the two alleles are in maximum disequilibrium with each other. Now we test the null hypothesis that the *susceptibility* allele has no effect versus the alternative that it does, H_0: $\beta = \beta_\mathrm{g} = 0$ vs. H_A: $\beta = \beta_\mathrm{g} \neq 0$. In this situation, there is no residual correlation within families because of linkage, and the standard Wald test is valid. In general, we will not know in advance which is the true state of nature, so we would like a test that is valid regardless of the presence of residual correlation and that has no penalty in power under (conditional) independence.

In our simulation study, we address two questions of primary importance: (1) when does linkage of a trait and marker locus noticeably affect the test size for disequilibrium at the marker locus using the standard conditional logistic likelihood, and (2) does the use of the robust variance estimator fix it and at what cost? Since

Table 1

Estimated Significance Level (%) of the Standard Score Test, for the Nominal Rate of 5% (500 Sibships Having at Least One Case and One Control; 10,000 Replications)

GOR and Trait-Allele Frequency	PAF (%)	Significance Level at θ =				
		.00	.05	.10	.20	.50
Sibship size 3:						
2:						
.1	7	5.14	4.90	5.29	5.20	4.92
.3	19	5.39	5.08	4.73	5.15	5.22
.5	29	5.11	5.59**	4.89	5.54**	5.34
20:						
.1	36	5.67**	5.06	4.98	5.21	5.03
.3	70	5.70**	5.80**	5.35	5.30	4.73
.5	84	5.54**	5.67**	5.14	5.04	5.14
Sibship size 4:						
2:						
.1	7	5.13	4.49	4.72	4.68	5.09
.3	19	5.83*	5.06	5.24	5.23	4.68
.5	29	5.35	5.39	4.96	5.25	4.91
20:						
.1	36	5.60**	6.06*	5.38	5.20	5.07
.3	70	5.93*	5.37	5.82*	6.17*	4.72
.5	84	5.53**	5.49**	5.26	5.55**	5.14

NOTE.—Marker allele 1 frequency 50%; population disease prevalence = 10%

* $P < .001$.

** $.001 \leqslant P < .05$.

it is computationally infeasible to address the first question through an extensive evaluation in SAS, we investigate the standard Score test using the programming language C++. Then, for a subset of the situations in which this test does not have the correct size, the standard and robust Wald tests are computed using SAS. Finally, power is estimated at the actual trait locus for the two Wald tests and the sibship disequilibrium test (SDT) (Horvath and Laird 1998). As discussed above, the standard Wald test is valid for this special case.

To study the effect of linkage on the test for disequilibrium, we estimated the type I error rate for the test, H_0: $\beta = 0$ vs. H_A: $\beta \neq 0$, for a marker locus that is linked to a trait locus but in linkage equilibrium in the population (scenario 1). We suppose that both the trait and marker loci are diallelic and assume a log-additive model at the trait locus. The disequilibrium between the marker 1 allele (m_1) and the trait allele (g) is measured by the difference between the haplotype frequency and the product of the marker and trait allele frequencies [$\delta = \Pr(m_1 g) - \Pr(m_1)\Pr(g)$]. We considered relatively common trait alleles with frequencies of 10%, 30%, and 50%, two effect sizes (one small and one large), and a disease prevalence of 10%. The amount of disease in the population explained by the gene is given by the population-attributable fraction (PAF), $1 - \Pr(\text{affected}|\text{noncarrier})/\text{prevalence}$. The genetic odds ratio (GOR) of two (small effect) yielded moderate PAFs of 7%–29%; the GOR of 20 (large effect) yielded high attributable fractions of 36%–84%. We let θ take on the values 0, 0.05, 0.1, 0.2, and 0.5. Sibships of sizes three and four were sampled under two different ascertainment events. In the first, we sampled sibships having at least one case and one control; in the second, we sampled those having at least two cases and one control ($n = 500$ sibships for each design). For all simulations, we consider a nominal test size of 5%.

The estimated test size of the standard Score test was higher under a larger genetic effect, under tighter linkage between the marker and trait locus, and when sibships were ascertained on the basis of having at least two cases compared with having at least one (tables 1 and 2). In sibships that were sampled with at least one affected and one unaffected sibling, we found that the estimated significance level slightly exceeded the nominal 5% value for the GOR of 20 (maximum false-positive rate = 6.17%). However, for genes with small effects (GOR = 2), the estimated test size was generally within sampling error of the 5% nominal rate. The false-positive rate under linkage of a trait and marker locus was higher in a sample of sibships having at least two affected siblings. The type I error rate also appeared to increase with sibship size. Still, for a gene having a small GOR, the estimated false-positive rate in sibships of size four remained <6%.

We repeated our simulations in SAS for the larger

Table 2

Estimated Significance Level (%) of the Standard Score Test, for the Nominal Rate of 5% (500 Sibships Having at Least Two Cases and One Control; 10,000 Replications)

GOR and Trait-Allele Frequency	PAF (%)	Significance Level at θ =				
		.00	.05	.10	.20	.50
Sibship size 3:						
2:						
.1	7	5.92*	5.68**	5.23	5.18	4.80
.3	19	5.45**	5.15	5.00	5.51**	5.07
.5	29	5.12	5.56**	5.07	5.43	5.22
20:						
.1	36	5.73**	6.01*	5.87*	5.38	5.34
.3	70	6.94*	6.96*	5.72**	5.77**	5.05
.5	84	6.47*	6.40*	5.69**	5.76**	4.77
Sibship size 4:						
2:						
.1	7	4.99	4.80	5.24	5.48**	4.95
.3	19	5.08	4.99	4.82	4.72	5.56**
.5	29	5.51**	4.72	5.21	5.32	5.41
20:						
.1	36	7.09*	6.28*	6.20*	5.25	5.41
.3	70	7.15*	6.70*	5.99*	5.91**	5.39
.5	84	6.92*	6.57*	6.15*	5.43	5.60**

NOTE.—See footnotes to table 1.

Table 3

Estimated Significance Level (%) of the Wald Test, for the Nominal Rate of 5% (500 Replications of 500 Sibships of Size Four, Having at Least Two Affected and One Unaffected Sibs)

Trait-Allele Frequency	Standard	Robust
.10	7.6[a]	5.4
.30	6.2	4.0
.50	7.4[a]	5.2

NOTE.—Marker allele 1 frequency 50%; θ = 0; population disease prevalence = 10%; GOR = 20.

[a] Different from 5% (P = .02).

sibship size, at least two affected sibs per sibship, and a GOR of 20. The estimated false-positive rate for the Wald test, using the robust variance estimate, was within sampling error of the 5% nominal value for all three simulations (table 3). The estimated error rate using the naïve variance estimate from ordinary conditional logistic regression exceeded the nominal value for two of the three scenarios.

Table 4 presents the power of the standard and robust Wald tests and the SDT at an actual trait locus (scenario 2). We set the marker 1 allele frequency equal to the trait allele frequency (q), fix θ at zero, and the disequilibrium parameter at its maximum possible value [$q \times (1\text{-}q)$]. This is equivalent to testing for a causal effect of the trait allele, H_0: $\beta = \beta_g = 0$ vs. H_A: $\beta \neq 0$. For this scenario, the standard Wald test is appropriate. We let the GOR take on values from one (no genetic effect) to four (large effect). Results showed that all three tests had the correct size. The robust and standard Wald tests had similar power and, in general, were more powerful than the SDT.

In summary, the proposed analysis provides a robust variance estimate for testing linkage disequilibrium in sibships of arbitrary size, provided that the correct mean structure is modeled. Since the expected marker allele frequencies under the null hypothesis are computed within families before summing over all families to accumulate evidence for linkage disequilibrium, the statistic is unaffected by population stratification. The analysis can be easily implemented using standard statistical software. The model is general and can be extended to include environmental exposures, and exposure-exposure interactions in addition to exposure-marker interactions.

These results suggest that the robust variance estimator correctly accounts for familial correlation in sibships because of linkage, with no cost in power at the true trait locus. However, the genetic effect needs to be quite extreme before the standard Wald test leads to incorrect inference, so that, for practical purposes, the standard test may perform adequately.

Acknowledgments

This work was supported by National Cancer Institute grants CA52862 (to K.D.S., P.K., D.C.T.) and CA42949 (to B.L.).

Table 4

Estimated Power (%) at a "Trait" Locus (500 Replications of 200 Sibships of Size Four)

GOR AND TRAIT-ALLELE FREQUENCY	ONE AFFECTED PLUS ONE UNAFFECTED			TWO AFFECTED PLUS ONE UNAFFECTED		
	Standard Wald	Robust Wald	SDT	Standard Wald	Robust Wald	SDT
1.0:						
.10	4.6	5.0	3.8	5.2	5.0	4.8
.30	6.0	6.6	5.8	3.6	3.6	3.8
.50	5.6	5.6	6.4	4.6	4.0	5.0
2.0:						
.10	26.2	26.2	24.2	37.2	37.0	28.8
.30	52.0	51.4	44.8	63.2	62.6	53.6
.50	60.0	59.4	53.8	67.4	68.0	57.8
3.0:						
.10	60.6	59.2	58.6	78.2	78.2	71.6
.30	90.2	90.0	87.2	97.4	97.8	94.0
.50	92.8	93.0	89.4	95.8	95.4	89.6
4.0:						
.10	80.8	79.8	77.2	96.6	96.2	91.4
.30	98.0	98.4	97.6	99.6	99.6	98.2
.50	99.0	99.0	97.4	100.0	99.8	98.2

NOTE.—Marker allele 1 frequency equals trait-allele frequency; disequilibrium equals maximum possible disequilibrium; θ = 0; population disease prevalence 10%; significance level 5%.

Appendix

First, we create two outcome variables (CASE and TIME) to describe the disease status of the sibs. The variable CASE is an indicator variable, which denotes whether a subject is affected (CASE = 1 for diseased; CASE = 0 for control). We code the TIME variable so that all the cases have the same event time and the controls have later censored times. We arbitrarily code TIME = 1 for cases and TIME = 2 for controls. Second, we create the explanatory variables. We let MARKER denote the coding for the marker genotypes and SIBSHIP the sibship ($i = 1,\ldots,I$). In this illustration, we use the log-additive coding for MARKER described earlier (0, ½, or 1). Finally, using these variables and the procedure PHREG, we can run conditional logistic regression by stratifying on sibship and using the option *ties = discrete* in the model statement. The robust variance estimate for the regression coefficient is then computed using the score residuals and the SAS/IML software. The code is based on the example provided in the SAS/STAT User's Guide, under the example for multiple failure outcomes using the PHREG procedure, and is given as follows:

```
proc phreg data=family outest=est2;
   model time*case(0)= marker / ties=discrete;
   strata sibship;
   output out=resids dfbeta=db1 / order=data;
   id sibship;
proc means data=resids noprint;
   by sibship;
   var db1;
   output out=out2 sum=db1;
proc iml;
   use out2;
   read all var{db1} into x;
   MARKER=x` * x;
   reset noname;
   _TYPE_={"ROBVAR"};
   print,"Estimated Covariance Matrix",,
      MARKER[colname=_TYPE_
      rowname=_TYPE_ format=10.5];
   create est1 from MARKER[colname=_TYPE_
      rowname=_TYPE_];
   append from MARKER[rowname=_TYPE_];
   quit;
```

References

Curtis D (1997) Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319–333

Horvath S, Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 63:1886–1897

Schaid DJ, Rowland C (1998) Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. Am J Hum Genet 63:1492–1506

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458

Therneau TM, Hamilton SA (1997) rhDNase as an example of recurrent event analysis. Statist Med 16:2029–2047