Genetic Epidemiology 14:629–634 (1997)

Association of Posterior p-Values of S.A.G.E. SIBPAL Proportion-IBD and Haseman-Elston Statistics for ACTHR112

Derek Gordon, Stephen J. Finch, Adam L. Jacobs, Nancy R. Mendell, Richard M. Single, and Thomas G. Marr

Cold Spring Harbor Laboratory (D.G., A.L.J., T.G.M.), Cold Spring Harbor, New York; SUNY at Stony Brook (S.J.F., N.R.M.), Stony Brook, New York; St. Olaf College (R.M.S.), Northfield, Minnesota

A common practice among researchers performing linkage studies is the use of equal allele frequencies as input when reporting p-values from computer linkage programs such as S.A.G.E. SIBPAL. Our results, using 5,000 sets from a uniform-prior distribution of allele frequencies, showed that such input may be problematic. Further, we found that the S.A.G.E. SIBPAL test for proportion of alleles shared identical by descent among concordantly affected sib pairs showed a greater percentage of significant p-values with decreasing parental genotype information (Table III), while the S.A.G.E. SIBPAL Haseman-Elston test produced significant p-values comparatively less frequently (Table IV).
© 1997 Wiley-Liss, Inc.

Key words: affected sib-pair methods, allele frequency, bipolar illness, Haseman-Elston, identity-by-descent

INTRODUCTION

In this paper we present a method using a Bayesian approach and produce a probability measure that can be used to assess the stability of performance of statistical tests in S.A.G.E. SIBPAL [Tran et al., 1994] to specified values of allele frequencies. We are assuming a uniform prior distribution of allele frequencies, and estimating the posterior distribution of the p-values. Our two goals for this paper were: 1) to mathematically show the effects of having to rely on unknown allele frequencies, and 2) take these mathematical results and apply them to a specific data set with commonly used programs.

Address for reprint requests to Dr. Derek Gordon, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724.

© 1997 Wiley-Liss, Inc.

The accurate estimation of population allele frequencies has been a long standing problem in population genetics [see Wilson and Bossert, 1971]. Estimation depends almost entirely on the assumptions being made (selection, genetic drift, etc.) and the resultant model being used. The question has been raised, to what degree does uncertainty in allele frequency specification affect output from methods that require allele frequencies as input? This question has been addressed in applying identity-by-state (IBS) methods. Perhaps the most well known of the IBS methods is the affected-relative-pair (APM) method, by D. Weeks and K. Lange [1988, 1992]. Babron et al. [1993] and Weeks and Harby [1995] demonstrated that misspecification of allele frequencies leads APM to produce false-positive results, especially when the single-locus model is used.

We focus on two well known linkage tests, namely the test for the proportion of alleles shared identical by descent (IBD) at a marker locus for affected sib pairs (hereafter referred to as Proportion-IBD test) and the regression of the squared-trait difference on the estimated proportion of alleles shared IBD at a marker (hereafter referred to as the H-E test [Haseman and Elston, 1972]). Both tests appear in S.A.G.E. SIBPAL. We consider the IBD test because it is considered one of the most powerful for linkage in diseases, such as bipolar illness where the population prevalence is low [Risch, 1990]. The H-E test, on the other hand, allows us to incorporate the information provided by the distribution of alleles IBD in discordant sib pairs and unaffected sib pairs as well.

Given that the allele frequencies are specified by their random prior distribution, the p-values of the Proportion-IBD test and the H-E test for a given data set are random variables. Therefore, we can discuss their distribution functions and compute the probability that each random variable is less than specified critical values [Hogg and Craig, 1978]. These computed probabilities estimate how likely it is that the p-value observed from a test is less than the values most commomly used as significance thresholds (e.g., 0.05, 0.01, 0.001 and 0.0001) and hence give us one quantitative measure of the question of stability of performance of each test for a given data set.

METHODS

We defined state-known sib-pair percentage (SKSP) of a family data set as the percentage of affected sib pairs in that data set for whom complete marker-genotype information is available on both parents. We used the SKSP parameter because the mathematical formulae in S.A.G.E. SIBPAL for estimating the proportion of alleles shared by a sib pair use the marker-allele frequencies when full genotype information on the parents is not available [Tran et al., 1994]. We then examined the relationship between the p-values reported in the Proportion-IBD test using a given specification of allele frequencies with the p-values reported in the H-E test using the same specification. Additionally, we examined the association of these results with differing SKSPs.

We chose ACTHR112, from the GAW10-NIMH1 data set, as our example marker for this analysis, because Berrettini et al. [1994] have listed this marker as a candidate gene for linkage to bipolar illness. In the GAW10-NIMH1 family data provided by Gershon, Goldin, and Berrettini, ACTHR112 has three alleles. The affection model that we chose is referred to as Model 2 in the GAW booklet; this model includes all schizo-affective codes, bipolar I, mania, bipolar II with major depression, and recurrent unipolar as affected.

Entries with SKSP equal to 49.5% resulted from the analysis of the GAW data set as given. We randomly deleted about half of the parental genotype information to generate results with SKSP equal to 21.9%; we deleted all parental genotype information to generate results with SKSP equal to 8.3% (0% could not be achieved since S.A.G.E. SIBPAL infers marker-genotype information on parents when it can be unequivocally inferred from siblings' marker-genotype information).

Since the marker locus we focused upon (ACTHR112) has three alleles, we created a 5,000 x 3 matrix, *M*, of random numbers between 0 and 1, using the random number generator drand48 [Kernighan and Ritchie, 1988]. Each row of *M* was normalized so that the row sums to 1; we used these rows of this new matrix (M_1) as the input frequencies to our analysis. We then ran S.A.G.E SIBPAL five thousand times, using a different row of M_1 each time as the set of input allele frequencies. The p-values from each run were collected into a file, for further analysis. In calculating the summary statistics probabilities and correlations for the different SKSPs (Tables I-V), we used the same matrix M_1.

RESULTS

In Tables I and II, we gave the p-values observed when we ran S.A.G.E SIBPAL using allele frequency of 1/3 for each of the three alleles. In Table II, we presented the effective degrees of freedom (Effective D.F.) calculated for use with the S.A.G.E SIBPAL H-E test [Tran et al., 1994]. Numbers of sib pairs and effective degrees of freedom changed because we considered various degrees of missing data. We defined as (suggested) linkage any observed p-value below a given significance-level threshold.

Depending upon the SKSP, the statistical test chosen (Proportion-IBD or H-E), and the significance threshold chosen, random assignment of allele frequencies can render differing conclusions about linkage. For example, when we considered the 49.5% SKSP (the complete family data), our calculations for the Proportion-IBD test at the 0.01 significance-level threshold indicated that there was linkage approximately one quarter (24%) of the time with random assignments of allele frequencies (Table III). Similarly, for the Proportion-IBD test with a 21.9% SKSP, and a 0.01 significance-level threshold,

TABLE I. Summary Statistics for Posterior p-Values of S.A.G.E. SIBPAL Proportion-IBD Test (NIMH1 Family Data)

SKSP	Min. p-value	Max. p-value	Median	Standard deviation	Eq. freq. p-value	Number of sib pairs
49.5%	0.000	0.196	0.015	0.008	0.017	99
21.9%	0.000	0.134	0.011	0.007	0.013	73
8.3%	0.000	0.057	0.001	0.005	0.013	72

TABLE II. Summary Statistics for Posterior p-Values of S.A.G.E. SIBPAL H-E Test (NIMH1 Family Data)

SKSP	Min. p-value	Max. p-value	Median	Standard deviation	Eq. freq. p-value	Effective D. F.
49.5%	0.005	0.318	0.028	0.011	0.027	176
21.9%	0.019	1.000	0.054	0.055	0.050	78
8.3%	0.019	0.793	0.031	0.031	0.023	100

our computations showed that linkage was established for approximately 47% of the random assignments of allele frequencies. Using the Proportion-IBD test with an 8.3% SKSP, at the 0.01 significance-level threshold, our computations indicated that linkage was determined 54% of the time when random assignments of allele frequencies were used (Table III). Using the H-E test with a 21.9% SKSP at the 0.05 significance-level threshold led us to conclude linkage for approximately one out of every three random assignments of allele frequencies (32.8% -- Table IV). Similarly, using the H-E test with a 8.3% SKSP at the 0.05 significance-level threshold led to our conclusion of linkage approximately three times as often as not (76.9% -- Table IV).

The choice of equal frequencies for each of the alleles at a marker is not necessarily the most conservative choice when using these statistical tests. In fact, when we looked at the Proportion-IBD test, the probability of observing a p-value greater than the equal-frequencies p-value was more than 0.19 for each SKSP (Table III). The results for the H-E test were more striking: all SKSPs showed probabilities of greater than 0.5 for observing p-values greater than the equal-frequencies p-value (Table IV).

Finally, we noted the considerable difference among most of the corresponding probabilities that appear in the Proportion-IBD test and the H-E test (Tables III and IV). The differences were greatest for the significance-level threshold of 0.001 and 0.01, but even at the 0.05 level, for 8.3% and 21.9% SKSPs, the Proportion-IBD test had a probability of 0.99 for each, and the H-E test had corresponding probabilities of 0.77 and 0.33, respectively. To investigate these differences further, we performed a correlation analysis among the Proportion-IBD test p-values and the H-E test p-values for all of the SKSPs (Table V). We used the Spearman correlation coefficients because, in our tests of distribution-fitting, none of the data sets was well-fit by a normal distribution or even an approximately normal distribution (results not shown). Also, the Spearman correlation coefficients are more robust to outliers than the Pearson correlation coefficients, and in each of our data sets we had outliers (Tables I and II -- maximum observation versus median).

In Table V, we found further empirical evidence that the Proportion-IBD test and the H-E test have p-values that respond differentially with respect to specification of the allele frequencies. The correlation for the 8.3% SKSP was -0.91. This result suggested that random assignment in allele frequencies caused the p-values from the respective tests

TABLE III. Probability of Observing Certain Posterior p-values for S.A.G.E. SIBPAL Proportion-IBD Test (NIMH1 Family Data)

SKSP	≤ 0.0001	≤ 0.001	≤ 0.01	≤ 0.05	≥ Equal freq. p-value
49.5%	0.004	0.014	0.243	0.999	0.353
21.9%	0.031	0.076	0.466	0.999	0.368
8.3%	0.036	0.092	0.541	0.999	0.197

TABLE IV. Probability of Observing Certain Posterior p-values for S.A.G.E. SIBPAL H-E Test (NIMH1 Family Data)

SKSP	≤ 0.0001	≤ 0.001	≤ 0.01	≤ 0.05	≥ Equal freq. p-value
49.5%	0.000	0.000	0.031	0.999	0.550
21.9%	0.000	0.000	0.000	0.328	0.643
8.3%	0.000	0.000	0.000	0.769	0.777

TABLE V. Spearman Correlation Between p-values Observed with the H-E Test and Proportion-IBD Test

SKSP	Spearman correlation
49.5%	0.27
21.9%	0.26
8.9%	-0.91

for the 8.3% SKSP to move in opposite directions. Considering the summary statistics from Tables I and II, as well as the probabilities from Tables III and IV, we observed that the Proportion-IBD test appeared to result in lower p-values, while the H-E test appeared to result in larger p-values. The largest correlation was 0.27, occurring for a SKSP of 49.5%.

We performed this analysis on another marker from the GAW10 data sets (namely, D18S37 from the Hopkins study) and observed similar results regarding differing responses for or against suggested linkage with the Proportion-IBD and the H-E tests, when using a uniform-prior distribution of allele frequencies. Also, we observed low positive correlation among the two tests, for varying SKSPs (results not shown). We note that, for the Hopkins data, there were five alleles at marker 37, and the unmodified family data had an SKSP of 76.1%, with 46 sib pairs (results not shown).

DISCUSSION

Based on the data set as given (SKSP = 49.5%), our technique would lead us to conclude both the Proportion-IBD and H-E test are significant at the 0.05 level, but not at the 0.01 level for either test. Using our technique in an analysis of this data set without parental genotype information, we would conclude a linkage significant at the 0.05 level with the Proportion-IBD test but not with the H-E test.

Also, for the Proportion-IBD test, we observe a consistent increase in the probability of observing posterior p-values below a given significance threshold as the percentage-genotype information on parents decreases, independent of the significance threshold (with the exception of 0.05, which was always 0.999 -- Table III).

While we recognize that some of the rows in our matrix M_1 (see methods) may not be plausible as marker-allele frequency estimates, we note that a uniform prior distribution is a good preliminary distribution for computing our posterior probabilities, especially if our alternative is equal allele frequencies as our input for the computer programs.

ACKNOWLEDGMENTS

We thank Dr. Michael Zhang for his suggestions about creating sets of random allele frequencies, Rebecca Koskela for her assistance with the preliminary computer programming, Dr. Scott Sutherland for his suggestions regarding random number generators, Carol Marcincuk for her typesetting, and Leigh Johnson for her editorial comments. Most of the results for this paper were obtained using the program package S.A.G.E., which is supported by U.S. Public Health Service resource grant RR03655 from the Division of Research Resources.

REFERENCES

Babron MC, Martinez M, Bonaiti-Pellie C, Clerget-Darpoux F (1993): Linkage detection by the affected-pedigree-member method: What is really tested? Genet Epidemiol 10:389-394.

Berrettini W, Ferraro TN, Goldin LR, Weeks DE, Detera-Wadleigh S, Nurnberger JI, Gershon ES (1994): Chromosome 18 DNA markers and manic depression: evidence for a susceptibility gene. Proc Natl Acad Sci USA 91:5918-5921.

Haseman JK, Elston RC (1972): The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 1:3-19.

Hogg RV, Craig AT (1978):"Introduction to Mathematical Statistics." Fourth Edition, New York: Macmillan Publishing Company.

Kernighan BW, Ritchie DM (1988): "The C Programming Language." Second Edition, Englewood, NJ: Prentice Hall.

Risch N (1990): Linkage strategies for genetically complex traits II - The power of affected relative pairs. Am J Hum Genet 46:229-241.

Risch N, Botstein D (1996): A manic depressive history. Nat Genet 12: 351-353.

Tran LD, Elston RC, Keats BJB, Wilson AF (1994): Sib-Pair Linkage Program Version 2.6. Part of S.A.G.E. Release 2.2 documentation. Computer package available from Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, OH.

Weeks D, Harby LD (1995): The affected-pedigree-method: Power to detect linkage. Hum Hered 45: 13-24.

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

Weeks D, Lange K (1992): A multilocus extension of the affected-pedigree-method of linkage analysis. Am J Hum Genet 50:859-868.

Wilson EO, Bossert WH (1971): "A Primer of Population Biology." Fourth Printing, Stamford, CT: Sinauer Associates.