Detection of Misspecified Relationships in Inbred and Outbred Pedigrees

Lei Sun[1], Mark Abney[1,2], Mary Sara McPeek[1,2]
[1]Department of Statistics, [2]Department of Human Genetics, University of Chicago, Chicago

Genome screen data collected for linkage analysis can be used to detect pedigree errors. We have developed methods applicable to a broad range of relationships. We discuss applications of our methods to data on asthma, in which we detect a number of likely misspecified relative pairs. We propose a graphical method for error detection in complex inbred pedigrees, with application to the Hutterites.

Key words: pedigree error, relationship estimation, software, PREST, ALTERTEST, likelihood ratio test, inbreeding

INTRODUCTION

The presence of pedigree errors in a data set may result in either reduced power or false positive evidence for linkage, so detection of pedigree errors can be useful prior to linkage analysis (Boehnke and Cox 1997). Genome screen data can provide considerable power to detect misspecified relationships. For detection of errors in general pedigrees, McPeek and Sun (2000) propose the expected identity by descent (EIBD), adjusted identity by state (AIBS), identity by state (IBS), and maximized log-likelihood ratio (MLLR) tests. They also propose a method for estimation of pairwise relationships. L. Sun, K. Wilder and M.S. McPeek (submitted) extend these methods to a broader range of relationships and implement them in the software programs PREST and ALTERTEST freely available on the web at *http://galton.uchicago.edu/~mcpeek/software/prest*. We apply the methods to the BUSS, GER and CSGA data. We extend the work of McPeek and Sun (2000) to include a graphical method for error detection in complex inbred pedigrees, which we apply to the Hutterite data.

Running Title: Detection of Pedigree Errors
Address reprint request to Mary Sara McPeek, Department of Statistics, University of Chicago, Chicago, IL 60637

METHODS

First consider pedigrees in which the majority of relative pairs fit into the following 11 relationship classes: MZ-twin, parent-offspring, full-sib, half-sib+first-cousin (a pair of individuals who have the same mother and different fathers who are brothers, or the same father and different mothers who are sisters), half-sib, grandparent-grandchild, avuncular, first-cousin, half-avuncular (the uncle/aunt is half-sib with the parent of the nephew/niece), half-first-cousin (a parent of one individual is half-sib with a parent of the other individual), and unrelated pairs. Later we will consider pedigrees, such as the Hutterites, for which these outbred relationships are not applicable. Leaving aside the MZ-twin pairs, which are not specified by the standard input format for pedigree data, we identify all pairs of the other 10 types within each pedigree. We then apply the two-stage screening procedure described in Sun, Wilder and McPeek (submitted). For each typed pair, in stage one, we perform the EIBD, AIBS and IBS tests, with the relationship indicated by the pedigree as the null hypothesis for the tests. We use a normal approximation to assess significance for each test. We also estimate $\boldsymbol{k} = (k_0, k_1, k_2)$, the probabilities of sharing 0, 1 and 2 alleles IBD, by the method of McPeek and Sun (2000). We then use the combined testing and estimation results to identify a set of pairs on whom the more powerful but more time-consuming MLLR test is performed in stage two. The MLLR statistic is maximized over a set of alternatives, A , which consists of the 11 relationships given above. To calculate the likelihood, in the presence of genotyping errors, for the cases of MZ-twin and parent-offspring pairs, we use the genotyping error model of Broman and Weber (1998) and Epstein et al. (2000). To assess significance for the MLLR test, for each pair, we simulate 10^5 or 10^6 realizations of the genotype data for that pair under the null relationship, with the same markers typed as in the data for that pair. If the null relationship indicated by the pedigree is rejected, it is useful to know what relationships are compatible with the data. When the MLLR test gives a small p-value, we use the estimate of $\boldsymbol{k}$ and the pattern of results among close relatives to select other likely relationships, which are then tested for fit to the data. Currently, PREST allows the 11 relationship classes given above as the null hypotheses for the tests.

For some pedigrees, such as the Hutterites, the simple outbred relationships considered above are not applicable; there are no relative pairs of exactly these types. For such pedigrees, we propose a graphical method for detection of pedigree errors. The first step is to calculate, for each pair, the probability distribution of the 9 condensed identity states [Jacquard, 1974] $\Delta_{1, \ldots, \Delta_9}$, which is obtained using the method of Abney, McPeek and Ober (2000). The second step is to calculate the EIBD, AIBS and IBS statistics. The last two are defined as in the outbred case, with kinship coefficient Φ calculated as $\Phi = \Delta_1 + (\Delta_3+\Delta_5+\Delta_7)/2+\Delta_8/4$. For the EIBD statistic, we assign states S_1, S_2, ..., S_9, as illustrated in McPeek and Sun (2000), to have 4, 0, 2, 0, 2, 0, 2, 1 and 0 alleles shared IBD by the pair. This definition ensures that the equation $4\Phi = E$ [EIBD] holds as in the case of non-inbred relative pairs. We do not calculate the variances of the statistics or perform the MLLR test because of the computational difficulties due to the complexity of the relationships. Instead, we plot the observed statistics for each pair vs. the kinship coefficient for that pair and look for apparent outliers in the graph. We also apply PREST to obtain estimates of pairwise relationships.

RESULTS

I. BUSS, GER and CSGA Data

No Mendelian errors are found through examination of every mother-father-child trio. Table 1 lists, for each data set, the number of typed pairs in each of the 9 relationship categories tested (no half-sib+first-cousin pairs in all the data sets), and the number of other relative pairs not tested. In the BUSS data, we observe that almost all the 80 unrelated pairs tested (the two parents in each pedigree) show significantly less sharing than expected, with *p*-values less than .00001. We suspect that the alleles in the BUSS data are family specific, i.e. allele numbers in the genotype data files refer to different alleles in different pedigrees. If so, the results of the tests are not meaningful, because the null means and null variances of the test statistics depend on the allele frequencies which are estimated using all the pedigrees. Table 2 lists the pairs in the GER data with *p*-value < .001 (uncorrected). Based on the results in Table 2, four pairs of putatively unrelated parents may actually be related approximately at the level of half-first-cousins. To apply the Bonferroni correction, we note that since all Mendelian errors have been cleaned, it would be impossible to reject any hypothesis test for a parent-offspring pair. Thus, we do not count the parent-offspring pairs in applying the Bonferroni correction, i.e., we multiply the uncorrected *p*-values by 252, instead of 694 (from Table 1). After this correction, only the last pair in Table 2 is significant. Note that the offspring genotypes provide no additional information on the relatedness of the parents, conditional on the parental genotype information.

TABLE 1. Summary of typed relative pairs within pedigrees for the BUSS, GER and CSGA data. p. o. (parent-offspring), f. sib (full-sib), h. sib (half-sib), g. p. c. (grandparent-grandchild), avun. (avuncular), f. cous. (first-cousin), h. avun. (half-avuncular), h. f. cous. (half-first-cousin), unrel. (unrelated), others (relationships that do not fit into the 11 classes given in the text).

Asthma Data Set	Number of Typed Relative Pairs									
	Tested									Not Tested
	p. o.	f. sib	h. sib	g. p. c	avun.	f. cous.	h. avun.	h. f. cous.	unrel.	others
BUSS	402	166	0	0	0	0	0	0	80	0
GER	442	155	0	0	0	0	0	0	97	0
CSGA	1365	754	113	226	345	193	21	3	706	47

TABLE 2. Results on possible misspecified relative pairs in the GER data. The results include the pedigree i.d., the i.d.s of the pair, the number of markers typed in both individuals, the null relationship given by the pedigree, the *p*-value of the test of the null, the estimated value of $\boldsymbol{k}$, a proposed relationship suggested by the estimate of $\boldsymbol{k}$ and the *p*-value of the test of the proposed relationship.

Ped. ID	ID1	ID2	No. of Mark.	Null Relationship	*p*-value	Estimated $\boldsymbol{k} = (k_0, k_1, k_2)$	Proposed Relationship	*p*-value
25	71478	57125	302	unrelated	.00026	(.884, .116, .000)	half-first-cousin	.447
51	74411	68580	317	unrelated	.00058	(.894, .096, .010)	half-first-cousin	.184
87	30259	95261	308	unrelated	.00068	(.871, .129, .000)	half-first-cousin	.907
90	63855	66532	312	unrelated	.00018	(.875, .107, .018)	half-first-cousin	.735

Table 3 gives the results for the pairs in the CSGA data with uncorrected p-value $< 2.1\times 10^{-5}$, which corresponds to a p-value of .05 after Bonferroni correction (again, not including the parent-offspring tests). Based on the results in Table 3, it is clear that the putative full sib pairs in pedigrees 1092 and 1202 are MZ twins or duplicated samples. There is strong evidence indicating that the half-sib pairs in pedigrees 1015, 1149, 1043 and 1097 are full-sib pairs, and that the full-sib pairs in pedigrees 1043, 1058, 1095, 1155 and 1199 are half-sib pairs. The evidence is also strong that the full-sib pairs in pedigrees 1097 and 1128 are half-sib pairs, and that some of the relevant avuncular pairs are half-avuncular pairs.

TABLE 3. Results on possible misspecified relative pairs in the CSGA data. (See legend of Table 2.)

Ped. ID	ID1	ID2	No. of Mark.	Null Relationship	p-value	Estimated $\boldsymbol{k} = (k_0, k_1, k_2)$	Proposed Relationship	p-value
1092	1	4	308	full-sib	0	(.000, .000, 1.00)	MZ-twin	.522
1202	5	6	298	full-sib	0	(.000, .000, 1.00)	MZ-twin	.481
1015	4	6	290	half-sib	0	(.238, .525, .237)	full-sib	.564
1149	1	4	288	half-sib	0	(.296, .479, .225)	full-sib	.261
1043	6	8	309	half-sib	0	(.176, .647, .117)	full-sib	.759
1043	3	8	309	full-sib	0	(.348, .636, .016)	half-sib	.457
1058	7	9	290	full-sib	0	(.463, .513, .025)	half-sib	.702
1095	5	8	310	full-sib	0	(.545, .454, .000)	half-sib	.507
1095	3	7	300	full-sib	0	(.482, .515, .004)	half-sib	.974
1097	5	8	301	half-sib	0	(.310, .450, .239)	full-sib	.609
1097	3	7	310	full-sib	0	(.449, .544, .007)	half-sib	.799
1097	3	8	306	avuncular	0	(.776, .224, .000)	half-avuncular	.591
1128	1	5	301	full-sib	0	(.542, .449, .010)	half-sib	.549
1128	5	6	309	avuncular	0	(.833, .136, .030)	half-avuncular	.215
1155	3	8	309	full-sib	0	(.557, .443, .000)	half-sib	.343
1199	3	8	291	full-sib	0	(.523, .477, .000)	half-sib	.221

II. HUTT Data

The Hutterite data consist of a single pedigree with 1544 individuals. Pedigree relationships between individuals are complicated; everyone is related and there are no relative pairs that fit into the 11 relationship classes considered. We identify 236,597 relative pairs with > 50 markers typed in common. No Mendelian errors are found. Figure 1 illustrates the observed EIBD statistic for each pair vs. the kinship coefficient for that pair. We find four obvious MZ twin pairs or duplicated samples (marked with diamonds in Figure 1), with all or nearly all the markers identical. They are **(**10075, 10076), (6863, 6864), (5206, 5205) and (9012, 9013). We also observe that individual 1768 has a number of relationship misfits (marked with x's in Figure 1). Figure 2 is a partial pedigree showing the position of 1768 relative to other individuals in the Hutterites. Based on the data, 1768 shows a large amount of over-sharing with the grandchildren of 1761 (7869, 10800, 10972), relative to what would be expected based on the pedigree. The estimates of $\boldsymbol{k}$ between 1768 and the grandchildren of 1761

are all about (.008, .992, .000). In fact, at almost every marker, 1768 shares at least 1 allele IBS with 7869, 10800 and 10972. This could be explained by the possibilities that 1768 and 3071 are either the same person or are MZ twins. There is also one inbred sib pair (marked with a triangle) that shows a large amount of over-sharing. This pair is from an inbred sibship of size 5, and none of the other 9 pairwise inbred sib pairs show over-sharing. The observed over-sharing could be due to chance.

DISCUSSION

We have developed a variety of statistical tools for detection of misspecified relationships. Our methods can be applied to a wide range of pedigree types, from sib pairs to complex inbred pedigrees. Analyses of the BUSS, GER, CSGA and HUTT data sets indicate a number of likely misspecified relative pairs and raise several issues. First, since allele frequencies are needed to use our methods, data in which allele definitions are family specific can be problematic. Second, the large number of hypothesis tests involved in checking a data set leads to a problem of multiple comparisons. We find that even using a conservative Bonferroni correction, we still have power to detect errors. Third, in a data set such as GER, with only 2 generations and all parents typed, nonpaternities/nonmaternities would be found by Mendelian errors. However, some unidentified relative marriages could be detected by our methods. Finally, there can be low power to detect small amounts of inbreeding in a sib pair. This suggests development of specially designed methods to detect inbreeding in a sibship with parents untyped.

ACKNOWLEDGMENTS

This work is supported by the National Institutes of Health grant HG01645 (to Mary Sara McPeek) and the NSF GIG postdoctoral fellowship (to Mark Abney). We thank Dr. Nancy Cox and Dr. Carole Ober for helpful discussions.

REFERENCES

Abney M, McPeek MS, Ober C (2000): Estimation of variance components of quantitative traits in inbred populations. Am J Hum Genet 66:629-650.

Boehnke M, Cox NJ (1997): Accurate inference of relationships in sib-pair linkage studies. Am J Hum Genet 61:423-429.

Broman KW, Weber JL (1998): Estimation of pairwise relationships in the presence of genotyping errors. Am J Hum Genet 63:1563-1564.

Epstein MP, Duren WL, Boehnke M (2000): Improved relationship inference for pairs of individuals. Am J Hum Genet 67:1219-1231.

Jacquard A (1974): "The genetic structure of populations." New York: Springer-Verlag.

McPeek MS, Sun L (2000): Statistical tests for detection of misspecified relationships by use of genome-screen data. Am J Hum Genet 66:1076-1094.

Sun L, Wilder K, McPeek MS (submitted): Enhanced pedigree error detection.

Legends for Figure 1 and Figure 2 (Figure 1 and Figure 2. appear before or after section II. HUTT Data).

Fig.1. Plot of EIBD statistic vs. kinships coefficient for the 236,597 relative pairs in the Hutterites, with at least 50 typed markers shared by each pair. Four possible MZ-twin pairs (or duplicated samples) are marked with diamonds, pairs with individual 1768 are marked with x's and the 10 pairs from the inbred sibship (9374, 9376, 9377, 9378, 9380) are marked with triangles.

Fig.2. A partial pedigree showing the position of individual 1768 relative to other individuals in the HUTT data set (but note that most of the founders of this partial pedigree are actually related). The starred individuals are not typed, and all the other individuals are typed for at least 330 markers.