Am. J. Hum. Genet. 59:946–950, 1996

Identifying Marker Typing Incompatibilities in Linkage Analysis

Heather M. Stringham and Michael Boehnke

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor

Summary

A common problem encountered in linkage analyses is that execution of the computer program is halted because of genotype(s) in the data that are inconsistent with Mendelian inheritance. Such inconsistencies may arise because of pedigree errors or errors in typing. In some cases, the source of the inconsistencies is easily identified by examining the pedigree. In others, the error is not obvious, and substantial time and effort are required to identify the responsible genotype(s). We have developed two methods for automatically identifying those individuals whose genotypes are most likely the cause of the inconsistencies. First, we calculate the posterior probability of genotyping error for each member of the pedigree, given the marker data on all pedigree members and allowing anyone in the pedigree to have an error. Second, we identify those individuals whose genotypes could be solely responsible for the inconsistency in the pedigree. We illustrate these methods with two examples: one a pedigree error, the second a genotyping error. These methods have been implemented as a module of the pedigree analysis program package MENDEL.

Introduction

When marker genotypes that are incompatible with Mendelian inheritance are present in a linkage data set, they are flagged by a linkage program through calculation of a zero likelihood. Such incompatibilities may be due to incorrect marker typing or data entry or to pedigree errors such as false paternity, unknown adoption, or sample switch. Frequently, visual inspection quickly identifies the error; but, particularly in moderate to large disease pedigrees and/or when many individuals are not genotyped, the error may not be obvious from visual inspection. In these situations, substantial time and effort may be required to trace the source of the inconsistency, and a method of automatic error detection would be useful.

The pedigree in figure 1, from the research of Wijsman and Guo (Ott 1993), is an example of a pedigree in which the error is not immediately obvious. This pedigree has 49 members, 37 of whom were typed for a marker with eight alleles. The incompatibility can be identified as follows: the genotype of person II:7 can be inferred from his spouse and children to be 3/7. Similarly, person II:3 can be inferred to be 1/7. Since person II:9 is 3/5, person II:1 must have two alleles from the set {1, 3, 5, 7}. However, since the son (III:3) of person II:1 is 2/8, person II:1 must also have either a 2 or an 8, resulting in a contradiction (Ott 1993). In this case, just determining the incompatibility was a chore. Even after doing so, however, it is still not clear who is the source of the error. In this paper, we introduce a method for automatically identifying those individuals who are most likely to have caused the inconsistency as well as a computationally simpler method that identifies individuals whose genotypes could be solely responsible for the incompatibility. Both of these methods were first described by Boehnke and Guo (1992).

Methods

Posterior Probability of Genotyping Error

To identify those pedigree members who are likely to have caused the incompatibility with Mendelian inheritance, we calculate the posterior probability $P(E_k|\mathbf{x})$ that individual k's observed marker phenotype represents a typing error (E_k), conditional on all the observed information for that marker in his/her pedigree ($\mathbf{x}$). This calculation requires a model for the way the errors occur in marker typing. The simplest such model is that all typing errors are independent and equally likely. For this model, given G possible genotypes at a codominant marker and a typing error rate $0 \leq e < 1$, the conditional probability of scoring someone's genotype as h, given that the true marker genotype is g is

$$P(h|g) = \begin{cases} 1 - e & \text{if } h = g \\ e/(G-1) & \text{if } h \neq g \end{cases} \quad (1)$$

(Lathrop et al. 1983*a*). Note that this method assumes

Received February 26, 1996; accepted for publication July 1, 1996.

Address for correspondence and reprints: Dr. Michael Boehnke, Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: boehnke@umich.edu

0002-9297/96/5904-0027$02.00

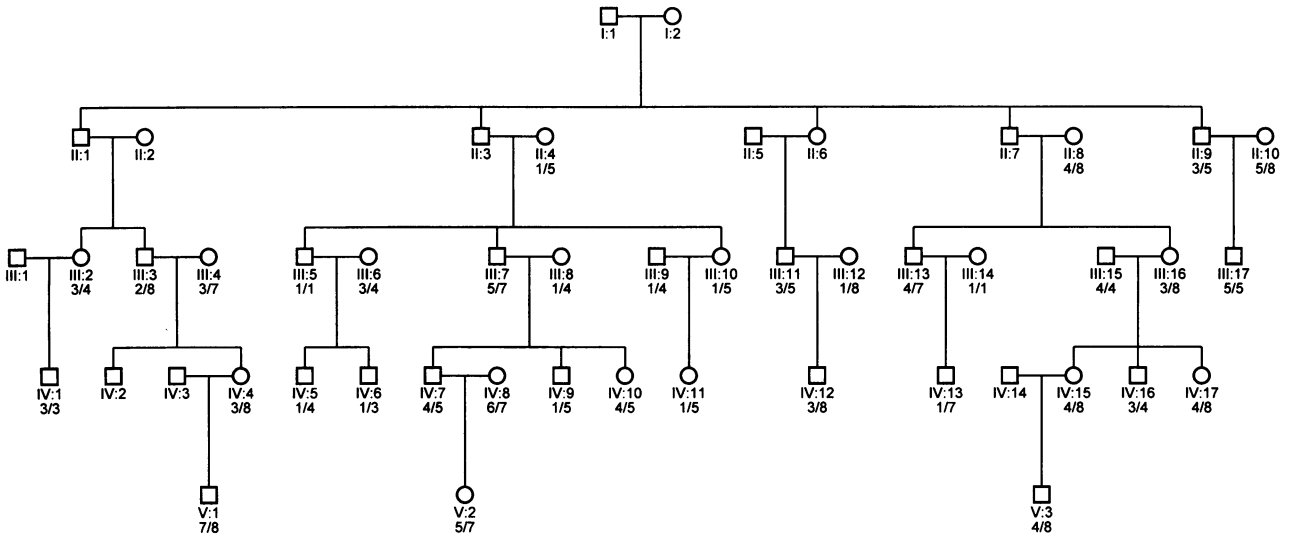

Figure 1 Pedigree from the research of Wijsman and Guo (Ott 1993), with marker typings inconsistent with Mendelian inheritance. Marker has eight alleles and $G = 36$ possible genotypes.

correct pedigree structure; that is, we are only directly allowing for typing error as an explanation for incompatibility. A person identified by this method to have a high probability of typing error could, however, have correct genotyping but represent a pedigree error (see Applications).

Calculating the conditional probabilities $P(E_k|\mathbf{x}; e)$ for all typed pedigree members k can be accomplished in the usual pedigree-analysis framework by calculating the ratio $P(E_k, \mathbf{x}; e)/P(\mathbf{x}; e)$. The denominator likelihood, $P(\mathbf{x}; e)$, is the probability of the pedigree data, allowing for error in all phenotypes. It can be calculated once using function (1) above in place of the standard 0/1 marker penetrances in which $e = 0$ and the correct phenotype has penetrance one with all other penetrances zero. Each numerator likelihood, $P(E_k, \mathbf{x}; e)$, the probability of the data *and* that k's phenotype is incorrect, can be calculated by using (1) for the penetrances for all pedigree members except with $1 - e$ replaced by zero for person k. The error rate e can either be set arbitrarily to some value such as $e = .01$ or 1/(number of typed people in the pedigree), or can be estimated directly from the data by maximizing the likelihood $P(\mathbf{x}; e)$.

A Computationally Simpler Method

While the above analysis is simple in principle, computational requirements can be nontrivial, since, given the penetrances (1), every genotype is possible for every pedigree member, typed or not. Since the number of genotypes $G = a(a + 1)/2$ for a locus with a alleles, $[a(a + 1)/2]^3$ genotypes must be cycled through for each father-mother-offspring trio in the likelihood calculation. Thus, for a marker with eight alleles, $36^3 = 46{,}656$ genotypes must be considered for each such trio. In linkage programs that use ordered genotypes, an even greater number of genotypes must be cycled through. To simplify computation, we can calculate the conditional probability $P(F_k|\mathbf{x})$ that individual k's phenotype at the marker locus is due to a typing error conditional on the phenotypic information $\mathbf{x}$ on his/her pedigree under the additional assumption that no other individual in the pedigree is incorrectly typed. This method allows all pedigree members besides person k to have standard 0/1 penetrances, resulting in rapid computation, since there is only one possible genotype for each typed individual except for person k. For person k, the penetrances remain as described in the previous section. The resulting conditional probabilities are either zero or one, with ones indicating which of the pedigree members could, by themselves, remove the inconsistency in the pedigree if only their genotype were changed.

To see why the conditional probabilities must be either zero or one, consider first the case where person k is not a possible source of the inconsistency. Then, no matter what genotype we assign to person k, the inconsistency will still be present in the pedigree, and a zero likelihood will be calculated. In contrast, suppose person k is a possible source of the inconsistency, and there is a genotype g_k, not equal to the observed genotype h_k, which would remove the inconsistency. Then $P(F_k, \mathbf{x}; e) = P(\mathbf{x}; e) > 0$, so that $P(F_k|\mathbf{x}; e) = 1$.

Applications

We analyzed the data for the pedigree given in figure 1, using both of the methods described above. Results

Table 1

Posterior Probabilities of Error, $P(E_k|\mathbf{x}; e)$, for Genotyped Members of Pedigree in Figure 1

		e	
PERSON	$\hat{e} = .035$	.01	.10
II:10	.007	.002	.023
III:2	.019	.005	.057
III:3	.995	.999	.986
III:4	.072	.022	.181
III:7	.021	.006	.066
III:9	.015	.004	.046
III:12	.007	.002	.023
III:14	.007	.002	.023
III:17	.007	.002	.021
IV:1	.016	.004	.050
IV:4	.040	.012	.102
IV:8	.007	.002	.023
V:1	.015	.004	.048
V:3	.007	.002	.022

NOTE.—Probabilities are rounded to the nearest .001. Persons with probabilities <.02 for each of the three values of the error rate e are not shown.

for the first method are given in table 1. Person III:3 is clearly pinpointed as the most likely source of the inconsistency, with a posterior probability of genotyping error of .995 when we use the maximum likelihood estimate of the error rate, $\hat{e} = .035$, which is slightly greater than 1/(number of typed people). Assuming a larger error rate of .10 or a smaller error rate of .01 had only a modest impact on the posterior probabilities. Person III:3 was also identified as the only pedigree member who could remove the inconsistency by changing only his/her genotype. Computation time for this simpler method was <1 min on a Sun SPARC 10 compared with ~40 min for the first method. Further study of the pedigree demonstrated that the error is due to false paternity (Ott 1993), suggesting that our methods, while directly addressing genotyping error, may also be useful for identifying pedigree error.

We also analyzed data for an eye-disease pedigree, shown in figure 2. This pedigree has 36 members, 28 of whom are typed for a marker with eight alleles. We discovered the inconsistency in this pedigree when we tried to estimate allele frequencies as the first step in a linkage analysis. Inspection of the pedigree revealed the inconsistency as follows: persons III:13, III:15, III:16, III:17, and III:18 are the offspring of persons II:9 and II:10. Among these offspring, alleles 115, 117, and 121 are present. Since person II:10 is 117/117, person II:9 must have a 115 and a 121 allele. However, since person III:18 is 117/117, person II:9 must also have a 117.

We computed the posterior probabilities of genotyping error for the typed members of this pedigree using allele frequencies modified from the Genome Data Base (GDB) so that all alleles present in the pedigree were included and so that allele frequencies sum to one. These error probabilities are given in table 2. Computation time for this analysis was ~25 min on a Sun SPARC 10. In this example, the posterior probabilities do not clearly identify a single person as the most likely source of the inconsistency but do suggest persons III:15, III:18, and IV:1 as the individuals most likely to be the source of the inconsistency, with posterior probabilities of genotyping error of .679, .315, and .219, respectively ($\hat{e} = .047$). Using the computationally simpler method, which again produced results in <1 min, we found that any of persons II:10, III:15, or III:18 could remove the inconsistency if only their genotype were changed. Thus, although person IV:1 had a moderately high probability of error compared to other members of the pedigree, his genotype is not the sole source of the inconsistency. In fact, person IV:1 is not a possible source of the inconsistency at all, as seen by inspection; only because he carries the rare 129 allele (frequency .01) is his genotype flagged. Subsequent retyping of the pedigree revealed a marker typing error in person III:15.

We also reran the analysis of the eye disease pedigree on the assumption of eight equally frequent alleles, resulting in posterior probabilities of genotyping error of .599 and .404 for persons III:15 and III:18, respectively. All other individuals had probabilities of error <.02 in this analysis. Thus, assuming equally frequent alleles may help to lower posterior probabilities that are moderately high solely because of the presence of a rare allele. However, it may also dilute the evidence for the true error, as was the case here. It also appears to be helpful to compare the results of the first method with that of the computationally simpler method, as in this case the simpler method established that person IV:1 alone could not explain the inconsistency. Using the simpler method first and then computing the posterior probabilities of error, $P(E_k|\mathbf{x}; e)$, for only those individuals flagged by the simpler method may be a useful technique for time-consuming problems.

Discussion

The methods described above for calculating the posterior probabilities of genotyping error for typed members of pedigrees with Mendelian inconsistencies are easy to interpret and may be computed using standard likelihood methods. They are useful in providing the automatic identification of a small number of individuals most likely to be the source(s) of the inconsistency in the pedigree. Such individuals may then be checked for errors in data entry or genotype scoring, thus allowing a quick return to the linkage analysis of the data.

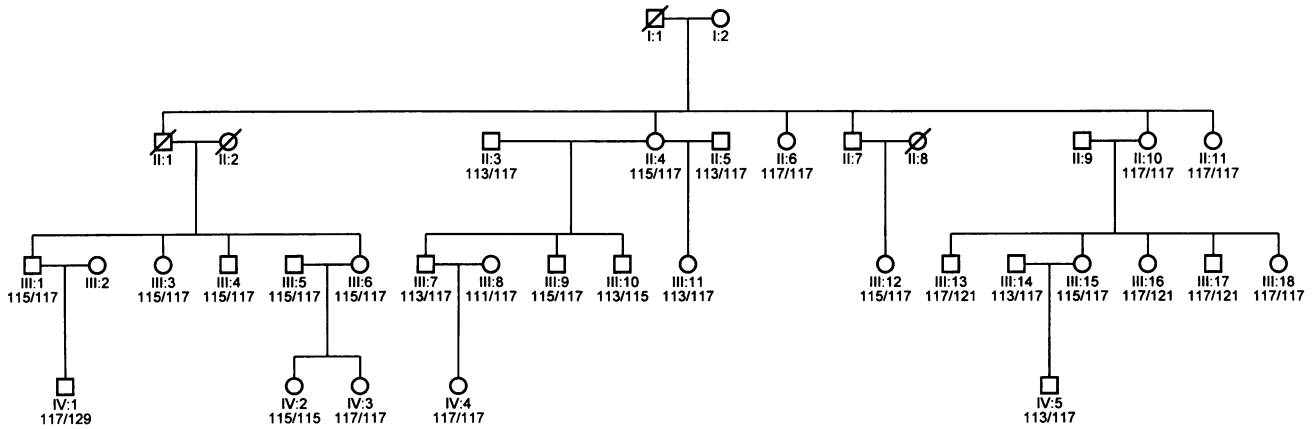

Figure 2 Eye-disease pedigree, with marker typings inconsistent with Mendelian inheritance. Marker has eight alleles and $G = 36$ possible genotypes.

Other Methods of Error Detection

The LINKAGE/FASTLINK programs (Cottingham et al. 1993) include the preprocessing program UNKNOWN, which detects marker genotype incompatibilities. UNKNOWN (from FASTLINK version 2.3P and beyond) determines the possible genotypes for each untyped person in the pedigree by using Boolean logic and tries to identify a child or a parent from nuclear families in which there is an error. The list of nuclear families can be quite long, even if there is only a single error in the pedigree, and there is no indication of which of the individuals are most likely to be the source of error. For example, a run of UNKNOWN version 5.20 on the pedigree in figure 1 identified nine nuclear families involved in the incompatibility.

Ott (1993) suggested comparing the conditional probability $P(g_k|x_k)$ of an individual's marker genotype (g_k), given his/her marker phenotype (x_k), with the conditional probability $P(g_k|\mathbf{x})$ of the marker genotype, given the marker phenotypes at that locus for all pedigree members ($\mathbf{x}$), using the same error model we employ. He measured the discrepancy between these two probabilities for each individual k by calculating the sum of squared deviations $[P(g_k|x_k) - P(g_k|\mathbf{x})]^2$ over all genotypes. His method identified the same individual as ours in figure 1 but lacks the direct interpretation of our two methods.

Lincoln and Lander (1992) used multilocus genotypes and, after calculating the maximum likelihood genetic map allowing for error, calculated the posterior probability distribution for each true multilocus genotype, given the complete typing data $\mathbf{x}$, the error rate e, and the recombination fraction estimate. They then calculated a LOD score for error, using the penetrance function given in equation (1) for the case of an F2 intercross ($G = 3$). For a single locus, the LOD score for error given by Lincoln and Lander is equivalent to

$$\log_{10} \frac{P(E_k|\mathbf{x};\, e)/(1 - P(E_k|\mathbf{x};\, e))}{e/(1 - e)}. \qquad (2)$$

They remarked that the extension from experimental data to human pedigrees, although theoretically straightforward, may be computationally challenging. They suggested several possible solutions for reducing the computational time. Among these is the idea of detecting errors separately in each offspring, assuming the genotypes of the other offspring are correct, which is analogous to our second method.

Other authors have looked at different aspects of error detection. Lathrop et al. (1983*b*) described a method for discriminating between pedigree error and typing error in families with inconsistencies. They evaluated the posterior probability of the true relationships, taking the possibility of mistyping into account. Ehm et al.

Table 2

Posterior Probabilities of Error, $P(E_k|\mathbf{x};\, e)$, for Genotyped Members of Eye-Disease Pedigree in Figure 2

Person	GDB	Equal
III:8	.044	.008
III:15	.679	.599
III:18	.315	.404
IV:1	.219	.017
$\hat{e}$	.047	.039

NOTE.—Probabilities were calculated for error rate $e = \hat{e}$ using GDB-based allele frequencies and assuming eight equally frequent alleles. Probabilities are rounded to the nearest .001. Persons with probabilities <.02 for both sets of allele frequencies are not shown.

(1995, 1996) have proposed a method for detecting errors which are consistent with Mendelian laws in the context of multipoint linkage.

Comments and Extensions

Our error-detection methods do not appear sensitive to the choice of the error rate *e*. This was clear in both of our applications using equal allele frequencies and was also noted by Lincoln and Lander (1992) and by Ehm et al. (1996).

We have found these methods to be successful at detecting errors not only in genotyping and pedigree errors, as illustrated in the Applications, but in other situations as well. These include handwritten genotypes misread when typed into the computer, two people with the same ID number appearing only once in the pedigree file with the family relationships of one person and the genotypes of the other, and a shift in the pedigree file in which a portion of the column of genotypes was offset from the column of person ID's by one line.

While we have found these methods to be quite useful for identifying incompatibilities in a variety of examples, there are some situations when they may fail or perform less satisfactorily. If there are multiple errors in a single pedigree, the second method will generally fail, because there will be no single person who is solely responsible for the inconsistencies in the pedigree. However, the first method still appears to perform well in such situations until the number of errors becomes very large. As mentioned earlier, presence of a rare allele can lead to moderately high probabilities of error in individuals not involved in the true inconsistency. On the other hand, if the incorrect allele is a very common one, the correct person may not be pointed to as strongly as if it had been a rare allele. For instance, in the eye-disease pedigree example, III:15's erroneous allele is 115, with allele frequency .27. If it had been the rare 129 allele (frequency .01) instead, the probability of error for III:15 would have been .983, much more clearly pinpointing the source of error (in addition, the probability of error for III:18 would have dropped to .021). Even so, the error could be identified in either case.

Other methods could be used to reduce computation time in a manner similar to our second method. For example, setting a person's phenotype to unknown and determining whether the resulting likelihood is positive would also identify those individuals who could be the sole source of the inconsistency.

Other, more complicated typing error models could be incorporated in the current framework. These could include frequency-dependent genotyping errors, allele-specific typing error rate, or allele-shift typing errors. False paternity or other pedigree error could also be modeled explicitly. The success of our relatively simple methods suggests that such elaborations are unnecessary.

We have written a module, USER*M14*, for MENDEL v3.31, that estimates the error rate *e* and calculates the conditional probabilities of error described above. This program is now part of the standard MENDEL package (Lange et al. 1988) available from Kenneth Lange (klange@umich.edu).

Acknowledgments

Support for this work was provided by grant HG00376 (to M.B.) from the National Institutes of Health and a University of Michigan Regents' Pre-Doctoral Fellowship (to H.M.S.). We thank Jürg Ott for bringing this problem to our attention, Julia E. Richards for allowing us to include data from the eye-disease pedigree, Deborah Wong for retyping that pedigree, and Kenneth Lange for providing comments on an earlier version of this manuscript.

References

Boehnke M, Guo S-W (1992) Statistical approaches to identify marker typing errors in linkage analysis. Am J Hum Genet Suppl 51:A183

Cottingham RW Jr, Idury RM, Schäffer AA (1993) Faster sequential genetic linkage computations. Am J Hum Genet 53:252–263

Ehm MG, Kimmel M, Cottingham RW (1995) Error detection in genetic linkage data for human pedigrees using likelihood ratio methods. J Biol Syst 3:13–25

——— (1996) Error detection for genetic data, using likelihood methods. Am J Hum Genet 58:225–234

Lange K, Weeks D, Boehnke M (1988) Programs for pedigree analysis. Genet Epidemiol 5:471–472

Lathrop GM, Hooper AB, Huntsman JW, Ward RH (1983*a*) Evaluating pedigree data. I. The estimation of pedigree error in the presence of marker mistyping. Am J Hum Genet 35: 241–262

Lathrop GM, Huntsman JW, Hooper AB, Ward RH (1983*b*) Evaluating pedigree data. II. Identifying the cause of error in families with inconsistencies. Hum Hered 33:377–389

Lincoln SE, Lander ES (1992) Systematic detection of errors in genetic linkage data. Genomics 14:604–610

Ott J (1993) Detecting marker inconsistencies in human gene mapping. Hum Hered 43:25–30