Genetic Epidemiology 17(Suppl 1): S79–S83 (1999)

Cleaning Genotype Data

Karl W. Broman

Marshfield Medical Research Foundation, Marshfield, Wisconsin

The identification of genes contributing to variation in complex phenotypes requires genetic data of high fidelity. Thus, the identification of pedigree and genotyping errors is a crucial prerequisite to the analysis of data from a genome scan for disease genes. The problem has been given little attention in most gene hunting papers; the focus has often been on eliminating mendelian inconsistencies in order that the analysis may proceed, rather than on achieving the best possible data. Though a number of computer programs are available to assist in the identification of genotyping and pedigree errors, the process is still not completely automated. While the Collaborative Study on the Genetics of Alcoholism (COGA) data set for GAW11 is completely compatible with Mendel's rules, there are still some errors present. We inspected the COGA data for the presence of additional errors, and identified five possible pedigree errors.

Key words: data cleaning, genotyping errors, nonpaternity, pedigree errors

INTRODUCTION

The identification of genes contributing to variation in complex phenotypes requires genetic data of high fidelity. Thus, the identification of pedigree and genotyping errors is a crucial prerequisite to the analysis of data from a genome scan for disease genes. Pedigree errors include problems such as nonpaternity and unreported adoption or twinning (where the reported relationships between individuals are incorrect), errors in the entry of pedigree information into a computer database and also sample mix-ups (where the sample genotyped does not correspond to the individual believed to have been genotyped). Genotyping errors occur when the observed genotype does not correspond to the true underlying genetic information, as a result of a mistake in data entry or a misinterpretation of the pattern on a gel. Mutations at a marker locus may mimic genotyping errors; it is just as important to identify and resolve such mutation events, as the later analysis requires the determination of the transmission of alleles through a pedigree.

Address reprint requests to Dr. K.W. Broman, Department of Biostatistics, John Hopkins University, School of Hygiene and Public Health, 615 N. Wolfe Street, Baltimore, MD 21205.

Although several computer programs are available to assist in the identification of errors, the process has not yet been completely automated. As a result, the cleaning of genotype data requires a careful analysis by an experienced individual. The process remains tedious, and is often viewed as an unfortunate burden that must be overcome in order that later analysis programs will run properly, rather than as an important step in ensuring that the later analysis is most valuable. We believe that the process of cleaning genotype data should not be relegated to the lab generating the data, but rather should be assigned to the analysts responsible for assessing the relationship between genotype and phenotype. Analysts should obtain data in the rawest form possible.

Regarding the identification of pedigree errors, in many cases pedigree problems may be easily seen when performing checks for mendelian inheritance: when a family displays an inordinate number of mendelian inconsistencies, the cause may often be traced to an error in the pedigree information. When parental data are missing, a more sophisticated approach may be necessary, because pedigree errors may not show up through a check for mendelian inheritance. Several computer programs are now available for verifying all pairwise relationships in a study [Boehnke and Cox, 1997; Göring and Ott, 1997; Ehm and Wagner, 1998]: for each pair of individuals in a study, their entire set of autosomal genetic data is used to infer their relationship, which may then be compared with the reported relationship. In most cases it is sufficient to consider the relationships monozygotic (MZ) twins, parent/offspring, full sibs, half sibs, and unrelated. One advantage of this type of approach is that the correct pedigree structure is often made clear, whereas when pedigree errors are observed by looking for large numbers of mendelian inconsistencies, it can be tricky to determine what change in the pedigree structure will eliminate the problem.

The approach of Boehnke and Cox [1997] is especially valuable because it takes account of the known linkage relationship between the genetic markers and yet is incredibly fast. Broman and Weber [1998] recently described a modification of this approach that accounts for possible genotyping errors in the genetic data, thus allowing the extension of the method to MZ twins and parent/offspring relationships. It is important that the verification of pedigree information be undertaken prior to the removal of genotypes that are inconsistent with mendelian inheritance, because that data provide important information about the correct relationships and because a proper modification of the relationship information may eliminate the need to remove those genotypes.

Regarding the identification of genotyping errors, one first identifies genotypes that do not conform to Mendel's rules and then determines which individual or individuals are responsible for the problem. Generally one seeks the most parsimonious explanation for the problem, finding the fewest genotypes that must be removed to eliminate the inconsistency. Two recent computer programs, PedCheck [O'Connell and Weeks, 1998] and a new module in the Mendel package [Stringham and Boehnke, 1996], make this process simpler, performing extensive checks of the data and identifying the individuals most likely to be in error. It is important that one not simply remove the genotypes flagged as inconsistent with Mendel's rules, because in many cases a single error (e.g., in a parent) will lead to a number of genotypes being flagged, and so by identifying the single individual responsible for the problem, one may retain much more of the genotype information.

Ideally, one would go beyond such one-marker-at-a-time checks for genotyping errors, looking further for unlikely multiple recombination events which may indicate the presence of genotyping errors. Unfortunately, the typical density at which most genome scans are performed makes this a largely useless effort, since a double recombinant within 30 or 40 cM cannot be immediately ascribed to a genotyping error.

To verify the relationships in the COGA data set, we applied a modified version of the method of Boehnke and Cox [1997], described in Broman and Weber [1998]. For each pair of individuals within a family, the genotype data for markers on the 22 autosomes were used to calculate the likelihood for the five relationships MZ twins, parent/offspring, full sibs, half sibs, and unrelated, assuming a genotyping error rate of 1%. The reported relationship for each pair was then compared with the inferred relationship, that giving the maximum likelihood.

RESULTS

We identified five possible pedigree errors in the COGA data (see Table I). Four of these involve a nonpaternity where the father was not genotyped; the other appears to be a sample mix-up, where the genotypes of a mother/daughter pair (family 81, individuals 921 and 925) are likely to be from the same sample. The sample for individual 921 is clearly that of the mother of the other siblings in the family, but the genotypes for individuals 921 and 925 share two alleles identical by state (IBS) at 203/215 markers for which they were both typed, far more than would be seen if they were truly mother/daughter.

Three of the identified nonpaternities are clear from the genetic data, but the situation in family 66 is not obvious. Individual 725 appears to be a half sib of 726 and 728 but a full sib of 727, and individuals 726, 727, and 728 appear to be a set of full sibs. A possible source for this discrepancy may be that many of the genotypes for this family are missing, most likely because the COGA data were cleaned to ensure a lack of mendelian inconsistencies, so that at any marker where these four reported siblings had more than four distinct alleles, at least one of the individuals' genotypes would have been deleted.

Among the 992 individuals typed in this study, the average proportion of missing genotype data is 9.2%. Individuals 725, 726, 727, and 728 in family 66 have 28%, 20%, 25% and 29% missing data, respectively, far above the average. The other three individuals indicated to be nonpaternities also have a great deal of missing genotype data: individuals 21-198, 42-435, and 74-822 have 16%, 50%, and 19% missing data, respectively. The data for individual 81-925 indicated to be a sample mix-up exhibited 23% missing data.

Many other individuals in this study also have a great deal of missing data: six individuals are missing more than half of their genotypes, and 22 are missing more than 25%. It may be that some of these samples show such high rates of missing data because their reported relationships were incorrect, but that we were unable to identify the problem due to the small amount of available data and due to the removal of the genotypes that would provide information about the relationship problem. An alternative explanation is that the DNA samples were of lower quality.

TABLE I. Pedigree Errors Identified in the COGA Data

Family-individual	Problem	Description
21-198	Nonpaternity	198 half sib of 196,197
42-435	Nonpaternity	435 half sib of 436, 437, 438
66-725	Possible nonpaternity	725 possibly half sib of 726, 728
74-822	Nonpaternity	822 half sib of 821, 823
81-925	Sample mix-up	925 same sample as 921

The genotypes for the mother/daughter pair 82-921 and 82-925 indicate that the genotyping error rate in this study may be relatively high. Among the 215 markers for which data were available on both individuals, the pair share two alleles IBS at 203 markers and one allele IBS at the other 12 markers. The genotypes thus appear to be from the same sample, so at the markers where IBS = 1, at least one of the genotypes was in error. This gives an estimated error rate of approximately 3%. If we allow that the individuals had matching genotypes at the 81 markers with missing data, the estimated error rate would be 2%.

Because the COGA data had been cleaned of mendelian inconsistencies prior to distribution, we were not able to use these data to illustrate the process of identifying and removing genotypes that do not follow Mendel's rules. The presence of additional genotyping errors, which did not result in mendelian inconsistencies, is indicated by the genetic maps estimated from the COGA data, which are much longer than the maps estimated using the CEPH families: the markers on the 22 autosomes span approximately 35 M, as estimated from these data. On the most recent maps estimated from the CEPH data [Broman et al., 1998], these markers span approximately 29 M. Such an inflation in map length is consistent with the presence of unidentified errors [Buetow, 1991; Lincoln and Lander, 1992].

We attempted to use the *chrompic* option of the CRI-MAP program [Green et al., 1990] to identify multiple recombination events that might indicate the presence of additional errors. Unfortunately, because of the marker density for this genome scan, approximately 14 cM, one could not assume that double recombinants were due to genotyping errors rather than true recombination events, and so this proved to be a useless enterprise. While it is clear that many of the observed double recombinants are due to unidentified genotyping errors, one cannot determine the origin of any particular double recombinant. For example, Figure 1 contains the *chrompic* output for the chromosome 2 data for two individuals in family 66. Both of these individuals exhibit three recombination events within a 48 cM region. While these events may each be due to a single recombination and a single genotyping error, one cannot rule out the possibility that they are truly triple recombinants in the region.

DISCUSSION

The cleaning of genotype data should be an integral part of the analysis of a genome scan for disease genes. It should not be viewed as a burden in the analysis of the relationship between phenotype and genotype but rather as a crucial step in ensuring that such analysis is made most valuable. Though the effect of genotyping errors on the

```
731   ooooo i-iii --i-iioio- oooo
      iiiiio-i-i i---iii-oo oooo

736   ooo-oo-o-o --oooioi- ii-i
      iii-ii-i-i i--iioo-oi ii-i
```

Fig. 1. *Chrompic* output from CRI-MAP for the chromosome 2 data for two individuals in family 66. Note: The two lines for each individual correspond to the maternal and paternal chromosomes. The inferred grandparental origin of each allele is denoted using " i " and " o. " The " – " indicates that the parent was homozygous or the individual's genotype was missing.

estimation of genetic maps is well known [Buetow, 1991; Lincoln and Lander, 1992], the effect of such errors on the power to map disease genes is not. A study assessing the influence of errors in genotype data on the power to map disease genes could be quite valuable; without such information it should be assumed that such errors may spell the difference between finding and not finding a gene. The routine use of the programs RELPAIR [Boehnke and Cox, 1997] and PedCheck [O'Connell and Weeks, 1998] will enable the identification of as many errors as possible while retaining as much genetic data as possible. The resulting, more refined data set will have maximal power to detect disease genes.

In the future, we expect that genome scans may be performed using dense arrays of diallelic markers. Though it is not yet clear what sort of error rate will accompany this new technology, it will no doubt be greater than 0. New tools will be needed in order to detect genotyping errors in diallelic markers, since with just three possible genotypes at each marker, most errors will conform to mendelian inheritance. Thus, we see a need to develop tools for identifying errors by observing multiple recombination events among closely linked markers.

While the COGA data set has been cleaned of mendelian inconsistencies, a number of errors are still present. We identified five possible pedigree errors and found evidence for the presence of additional genotyping errors. While analysts no doubt appreciated receiving "clean" data, we feel that, because the cleaning process has not yet been made completely automatic and routine, analysts should receive data in a more raw form and should take responsibility for identifying and resolving errors in genetic data. Still, it is important for laboratory scientists to continue to strive for the highest quality data, and it may be that an improved knowledge of methods for identifying errors will lead to new insight into how to better call genotypes.

REFERENCES

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): Estimating pairwise relationships in the presence of genotyping errors. Am J Hum Genet 63:1563-1564.

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998): Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet 63:861-869.

Buetow KH (1991): Influence of aberrant observations on high-resolution linkage analysis outcomes. Am J Hum Genet 49:985-994.

Ehm MG, Wagner M (1998): A test statistic to detect errors in sib-pair relationships. Am J Hum Genet 62:181-188.

Göring HHH, Ott J (1997): Relationship estimation in affected sib pair analysis of late-onset diseases. Eur J Hum Genet 5:69-77.

Green P, Falls K, Crooks S (1990): Documentation for CRI-MAP, version 2.4.

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

O'Connell JR, Weeks DE (1998): PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 6:259-266.

Stringham HM, Boehnke M (1996): Identifying marker typing incompatibilities in linkage analysis. Am J Hum Genet 59:946-950.