Am. J. Hum. Genet. 48:22–25, 1991

Allele Frequency Estimation from Data on Relatives

Michael Boehnke

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

Summary

Given genetic marker data on unrelated individuals, maximum-likelihood allele-frequency estimates and their standard errors are easily calculated from sample proportions. When marker phenotypes are observed on relatives, this method cannot be used without either discarding a subset of the data or incorrectly assuming that all individuals are unrelated. Here, I describe a method for allele frequency estimation for data on relatives that is based on standard methods of pedigree analysis. This method makes use of all available marker information while correctly taking into account the dependence between relatives. I illustrate use of the method with family data for a VNTR polymorphism near the apolipoprotein B locus.

Introduction

For simple codominant markers and samples of unrelated individuals, allele-frequency estimation is easily carried out. Maximum-likelihood allele-frequency estimates (MLEs) may be calculated as sample proportions of the form $\hat{q} = y/n$, where y is the number of alleles of a particular type and n is the total number of alleles typed. Standard errors (SEs) for the allele-frequency estimates can be calculated as $[\hat{q}(1-\hat{q})/n]^{1/2}$.

However, at times it is desirable to estimate allele frequencies by using information on relatives. Families often are collected for some other purpose, such as segregation or linkage analysis, but allele-frequency estimates also may be needed. Alternatively, allele-frequency estimates may be required for complex genetic markers, such as VNTRs (Nakamura et al. 1987; Weber and May 1989), that are typed most readily and reliably within families. Finally, it may be desired to use family data to test for an association between a genetic marker and a genetic disease or some other familial trait (Bird et al. 1987; Schellenberg et al. 1987).

In each of these cases, allele frequencies can be estimated by selecting a subset of unrelated individuals from each family. This method wastes information, and the decision of whom to include in the analysis is often arbitrary and can affect the results. Alternatively, the dependence between relatives can be ignored. This method results in estimates that are no longer maximum likelihood, in an overestimate of the information present in the sample, and consequently in SEs that are too small.

In the current paper, I describe a third method of allele-frequency estimation for family data. This method estimates allele or haplotype frequencies by maximizing the likelihood of the family data in the framework of pedigree analysis (Elston and Stewart 1971; Lange and Boehnke 1983). The method enables use of the entire sample, and it results in MLEs. Such estimates have the desirable statistical properties of consistency, efficiency, and asymptotic unbiasedness (Rao 1973). SEs of the parameter estimates also are easily obtained by double differentiation of the logarithm of the likelihood function. I illustrate use of the method with family data for a VNTR polymorphism near the apolipoprotein B locus on human chromosome 2.

Material and Methods

Likelihood of the Pedigree

Given a pedigree of n individuals, let $\mathbf{x} = (x_1, \ldots, x_n)$, and let $\mathbf{g} = (g_1, \ldots, g_n)$ be the vectors of marker phenotypes and genotypes for the pedigree, respectively. For convenience I assume we wish to estimate allele frequencies for a single genetic locus. The same method

Received June 27, 1990; revision received September 4, 1990.

Address for correspondence and reprints: Michael Boehnke, Ph.D., Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109–2029.

0002-9297/91/4801-0003$02.00

applies to haplotype-frequency estimation for multilocus data. Let $A_1, \ldots, A_m$ be the possible alleles at the marker locus A, and let $\boldsymbol{q} = (q_1, \ldots, q_m)$ be the corresponding allele frequencies. Under the assumption of Hardy-Weinberg equilibrium (but see Discussion), the likelihood of the pedigree is

$$L(\boldsymbol{q}) = \sum_{\boldsymbol{g}} \prod_{i} P(x_i|g_i) \prod_{j} P(g_j) \prod_{k} P(g_k|g_{kf},g_{km}) \ , \quad (1)$$

where the sum ranges over all marker genotype vectors $\boldsymbol{g}$ and the products range over all pedigree members i of known marker phenotype, all pedigree originals j whose parents are not present in the pedigree, and all pedigree descendants k whose parents k_f and k_m are present in the pedigree (Elston and Stewart 1971; Lange and Boehnke 1983). Here, $P(x_i|g_i)$ is the penetrance of the marker phenotype x_i given marker genotype g_i, $P(g_j)$ is the prior probability of marker genotype g_j for an original who marries into the pedigree at random, and $P(g_k|g_{kf}, g_{km})$ is the transmission probability that an offspring k will be of marker genotype g_k, given parents of genotypes g_{kf} and g_{km}.

The allele frequencies $\boldsymbol{q}$ appear in the pedigree likelihood only in the prior probabilities $P(g_j)$. For an autosomal locus, the allele frequencies appear in terms of the form q_i^2 and $2q_iq_j$ for individuals of genotypes A_iA_i and A_iA_j ($1 \leqslant i, j \leqslant m$), respectively. For an X-linked locus, they also appear in terms of the form q_i for hemizygous genotype A_i ($1 \leqslant i \leqslant m$). The transmission and penetrance probabilities are required to insure that the relationships between pedigree members—and the correspondence between marker phenotypes and genotypes—are taken properly into account. Note that for the special case of a codominant marker typed on unrelated individuals, the likelihood (1) reduces to the product of the prior probabilities $P(g_j)$.

Given a sample of unrelated pedigrees, the joint likelihood for the entire sample is the product of the likelihoods (1) for each of the pedigrees.

Allele-Frequency Estimation

Given this pedigree likelihood framework, MLEs of the allele frequencies can be obtained by maximizing the likelihood as a function of $\boldsymbol{q}$. For (sets of) pedigrees of arbitrary structure, general closed-form solutions to this maximization problem are not available unless all pedigree originals are of known marker genotype; in that case, allele frequencies can be estimated by allele counting and sample proportions. If not all originals are of known marker genotypes, pedigree likelihood calculation programs may be used to estimate allele frequencies. I have written a version of the computer program MENDEL (Lange et al. 1988) that carries out this method of allele-frequency estimation. The program allows for pedigrees of any structure and for marker loci with alleles that demonstrate any pattern of dominance or codominance. MENDEL uses the variable metric routine SEARCH (Lange et al. 1988) for iterative maximization of the logarithm of the likelihood function. The covariance matrix of the parameter estimates is calculated as the negative of the inverse of the matrix of numerically calculated second partial derivatives of the logarithm of the likelihood function. SEs of the allele-frequency estimates are calculated as the square roots of the diagonal elements of the resulting matrix.

Equivalent Numbers of Alleles

For an allele-frequency estimate $\hat{q}$ based on a sample of n alleles observed in unrelated individuals, the SE of $\hat{q}$ is calculated as $\mathrm{SE}(\hat{q}) = [\hat{q}(1-\hat{q})/n]^{1/2}$. Given an allele-frequency estimate $\hat{q}$ and its SE based on a sample of related individuals, define the effective number of alleles as $n^* = \hat{q}(1-\hat{q})/\mathrm{SE}(\hat{q})^2$. n^* is thus the approximate number of alleles from unrelated individuals required to provide the same amount of information as is provided by the sample of families. I use the idea of the effective number of alleles to compare the three methods of allele-frequency estimation.

Example

Several investigators recently described a VNTR polymorphism near the apolipoprotein B (ApoB) locus on human chromosome 2 (Boerwinkle et al. 1989; Cuny et al. 1989; Ludwig et al. 1989). This polymorphism has at least 15 alleles that differ in size, as based on the number of copies of a tandem repeat of 15 bp. Genotypes on a sample of 233 Caucasian individuals from 52 sibships, nuclear families, and small extended pedigrees were provided to me by Dr. Scott Diehl of the Department of Human Genetics of the Medical College of Virginia. Twelve of the 15 known ApoB VNTR alleles were observed in this sample.

I calculated allele-frequency estimates and their SEs for these data by the three methods described in the Introduction (table 1). MLEs for the entire sample that take into account the dependence between relatives (table 1, column 1) may be compared with the estimates that ignore the dependence between relatives (table 1, column 2) and with the estimates based on a maximal subset of unrelated individuals (table 1, column 3).

Table 1

Allele-Frequency Estimates ± SE for the ApoB VNTR

	Entire Sample (n = 233)		Maximal Unrelated
Allele[a]	Relations Noted	Assumed Unrelated	Subset (n = 62)
1	.0533 ± .0156	.0579 ± .0108	.0403 ± .0177
2	.0720 ± .0183	.0644 ± .0114	.0806 ± .0245
3	.1995 ± .0283	.2167 ± .0191	.1935 ± .0355
4	.3589 ± .0342	.3562 ± .0222	.3629 ± .0432
5	.0679 ± .0175	.0601 ± .0110	.0887 ± .0255
6	.0096 ± .0068	.0043 ± .0030	.0000 ± .0000
7	.0048 ± .0048	.0043 ± .0030	.0081 ± .0080
8	.0000 ± .0000	.0000 ± .0000	.0000 ± .0000
9	.0965 ± .0208	.0901 ± .0133	.0968 ± .0266
10	.0989 ± .0210	.1052 ± .0142	.0968 ± .0266
11	.0290 ± .0117	.0215 ± .0067	.0242 ± .0138
12	.0048 ± .0048	.0150 ± .0056	.0000 ± .0000
13	.0048 ± .0048	.0043 ± .0030	.0081 ± .0080
14	.0000 ± .0000	.0000 ± .0000	.0000 ± .0000
15	.0000 ± .0000	.0000 ± .0000	.0000 ± .0000

[a] Designation is arbitrary.

Ignoring the dependence between relatives in this example gave allele-frequency estimates reasonably close to the MLEs. In relative terms, estimates usually differed from the MLEs by not much more than 10% of the values of the MLEs, although the estimates for the rare alleles 6, 11, and 12 differed from the MLEs by 55%, 26%, and 213%, respectively. Absolute errors in the estimates were in no case greater than .0172.

In contrast, the information content of the sample was substantially overestimated when dependence was ignored. Depending on the specific allele chosen, the effective number of alleles n^* provided by the sample ranged from 197 to 207. This corresponds to only 42%–44% of the 466 alleles that were actually typed. Resulting SEs when dependence was ignored (table 1, column 2) were all at least 31% less than those for the fully efficient MLEs, with the exception of the rare allele 12. Such a large overestimate in the information and corresponding underestimate in the SEs would result in a strong anticonservative bias in hypothesis tests on the allele frequencies.

Choosing a maximal subset of unrelated family members resulted in a sample size of 62. For each pedigree, the maximal subset was randomly selected if there were multiple subsets of the same size. Allele-frequency estimates based on the subset (table 1, column 3) differed more from the MLEs (table 1, column 1) than did those that simply ignored dependence (table 1, column 2). Indeed, the rare alleles 6 and 12 were no longer represented in the data. Still, absolute differences between these allele-frequency estimates and the MLEs were in no case greater than .0208.

Comparing the 124 alleles in the subset to the equivalent of 197–207 alleles provided by the entire sample suggests a loss of 37%–40% of the total information. SEs for these estimates were 13%–67% greater than those for the MLEs based on the entire sample. Thus, for the allele frequencies, hypothesis tests based on the subset would be substantially less powerful than tests based on the entire sample.

Discussion

The pedigree analysis method of allele-frequency estimation described here can be used for genetic markers with alleles that demonstrate any pattern of dominance or codominance and for samples of related or unrelated individuals. Given data on two or more linked loci, the method can also be used to estimate haplotype frequencies. Estimating both allele frequencies and haplotype frequencies permits a test of the assumption of linkage equilibrium by the likelihood-ratio criterion.

In principle, the method also provides a framework to estimate genotype frequencies. Instead of parameterizing the pedigree likelihood (1) in terms of allele frequencies under the assumption of Hardy-Weinberg equilibrium, the likelihood can be parameterized in terms of genotype frequencies $P(g_j)$, with likelihood maximization as a function of the genotype frequencies. Carrying out analyses for both allele frequencies and genotype frequencies permits a test of the assumption of Hardy-Weinberg equilibrium by the likelihood ratio criterion. The primary limitation to this approach is that as the number of alleles m becomes large, the number of genotypes $m(m+1)/2$ becomes much larger. From the $m = 15$ known alleles of the ApoB VNTR marker, 120 genotypes are possible.

Allele-frequency estimation for family data can also be used to test for an association between a familial disease and a genetic marker. We previously carried out such an analysis for familial Alzheimer disease (FAD) and polymorphisms at the apolipoprotein CII (ApoCII) (Schellenberg et al. 1987) and C4B loci (Bird et al. 1987). The strong association between an ApoCII allele and FAD might not have reached statistical significance if only an unrelated subset of the family data had been used, and it would have been overstated if the relationships between family members were ignored.

Family data are often collected for purposes other than allele-frequency estimation, e.g., for segregation or linkage analysis. Indeed, the families employed in the ApoB example described above were ascertained as part of a linkage study of schizophrenia. Such samples still can provide substantial information on allele frequencies, and it is only sensible to make efficient use of that information. The pedigree analysis method of allele-frequency estimation described here results in MLEs of the allele frequencies and in accurate estimates of their SEs. The alternative methods that ignore the dependence between relatives or choose a maximal subset of unrelated individuals provide less efficient allele frequency estimates and inaccurate estimates of their SEs.

The source code for the MENDEL version USERM13 that carries out allele- and haplotype-frequency estimation for data on relatives can be obtained free of charge from the author.

Acknowledgments

I thank Kenneth Lange and Daniel Weeks for developing and supporting MENDEL, and I thank Scott Diehl for providing the ApoB data. I also thank Drs. Lange and Diehl for reading and commenting on a previous version of the manuscript. This work was supported by grant R29-HG00376 from the National Institutes of Health.

References

Bird TD, Boehnke M, Anderson J, Lampe TH, Schellenberg G, Larson EB (1987) The frequency of C4B variants of complement in familial and sporadic Alzheimer disease. Alzheimer Dis Assoc Disord 1:251–255

Boerwinkle E, Xiong W, Fourest E, Chan L (1989) Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: application to the apolipoprotein B 3′ hypervariable region. Proc Natl Acad Sci USA 86: 212–216

Cuny G, Vigneron S, Senglat C, Roizes G (1989) Detection by PCR of the highly polymorphic 3′ Apo B gene minisatellite alleles. Cytogenet Cell Genet 51:982

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542

Lange K, Boehnke M (1983) Extensions to pedigree analysis. V. Optimal calculation of Mendelian likelihoods. Hum Hered 33:291–301

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

Ludwig EH, Friedl W, McCarthy BJ (1989) High-resolution analysis of a hypervariable region in the human apolipoprotein B gene. Am J Hum Genet 45:458–464

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, et al (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616–1622

Rao CR (1973) Linear statistical inference and its applications, 2d ed. John Wiley, New York, pp 364–366

Schellenberg GD, Deeb SS, Boehnke M, Bryant EM, Martin GM, Lampe TH, Bird TD (1987) Association of an apolipoprotein CII allele with familial dementia of the Alzheimer type. J Neurogenet 4:97–108

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396