Original Paper

European Journal of Human Genetics

Eur J Hum Genet 1997;5:203–205

Received: November 22, 1996
Revision received: April 18, 1997
Accepted: May 10, 1997

Fatemeh Haghighi[a]
Jurg Ott[a,b]

[a] Department of Genetics and Development, Columbia University and
[b] Laboratory of Statistical Genetics, Rockefeller University, New York, N.Y., USA

Estimating Recessive Disease Allele Frequency Based on Genetic Maps

Key Words
Recessive disease
Dahlberg method
Allele frequency

Abstract
For a recessive disease whose gene has been localized on the human gene map, a new method is described for estimating the population frequency of the disease allele. The method focuses on affected individuals whose parents are first cousins, where parents and grandparents are genotyped for highly polymorphic markers at the disease gene. The primary statistic is the proportion of such probands who are autozygous (homozygous due to identity by descent of the two disease alleles), where this proportion is a function of the disease allele frequency. Our map-based method is compared to Dahlberg's method of estimating recessive disease allele frequencies, which is based on the proportion of affected individuals whose parents are first cousins; this proportion is also a function of the disease allele frequency. For small to moderate sample sizes and traits that are not too common, our new method is more efficient (for some parameter values dramatically more efficient) than Dahlberg's method.

Introduction

For a common trait, prevalence is easily estimated from a random sample of the population. However, this is prohibitively expensive for a rare disease, which is often ascertained through probands [1]. Population prevalence must then be estimated via the ascertainment probability.

Specialized indirect methods have been devised that do not rely on complete enumeration of all cases; they work with probands but use additional information that obviates the need for ascertainment corrections. In this paper, we are concerned with recessively inherited traits. For these, when q denotes the population frequency of the disease allele, q^2 is the disease incidence. Assuming an equal life expectancy for affected and unaffected individuals, q^2 is also the disease prevalence. A common method of estimating q is due to Dahlberg [2] and relies on the observation that recessive traits tend to occur more frequently among offspring of consanguineous matings than of unrelated parents. This method and extensions of it have been covered by Li [3] and applied, for example, to cystic fibrosis in Italy [4]. Below, we propose a new map-based method for estimating q and compare its efficiency relative to the Dahlberg method and that of simple population sampling.

For the two methods, probands are defined as follows. In our map-based method, a proband is an affected individual whose parents are first cousins, whereas in the Dahlberg method, a proband is any affected invididual. Thus, as will be outlined in the discussion, the cost of ascertaining probands varies among the two methods. For each method, N denotes the number of probands.

Methods

Map-Based Method

Consider a recessive trait with disease allele frequency, q. For a random individual, the probability of being affected (homozygous) is given by $Fq + (1 - F)q^2$, where F is the individual's inbreeding coeffi-

KARGER
E-Mail karger@karger.ch
Fax + 41 61 306 12 34
http://www.karger.ch

This article is also accessible online at:
http://BioMedNet.com/karger

Jurg Ott
Rockefeller University, Box 192
1230 York Avenue
New York, NY 10021-6399 (USA)
Tel. (212) 327 7991, fax (212) 327 7996, e-mail ott@linkage.rockefeller.edu

cient [3]. The first term indicates the probability of being autozygous, that is, homozygous due to having inherited the two disease alleles as copies of the same ancestral allele (identically by descent), while the second term refers to being allozygous. Therefore, for an affected individual whose parents are first cousins ($F = 1/16$), the conditional probability of being autozygous is given by

$$p = Fq/[Fq + (1 - F)q^2] = 1/(1 + 15q). \quad (1)$$

If the disease gene has a known genomic position and is located in a dense map of marker loci, marker typing of parents and grandparents will allow one to determine whether a proband is autozygous or allozygous. In other words, inheritance of alleles for markers tightly linked with the disease locus will show whether the two disease alleles in a proband are copies of one disease allele in one of the great-grandparents (autozygosity) or whether the two disease alleles have entered the pedigree separately (allozygosity). Consider N such probands of which an observed proportion, p, is autozygous. Based on (1), this estimate of p can be translated into a maximum likelihood estimate, $\hat{q}$, of the allele frequency. Because some values of p may lead to values of q exceeding 1, we define

$$\hat{q} = \begin{cases} (1 - \hat{p})/(15\hat{p}) & \text{if } \hat{p} > 1/16 \\ 1 & \text{if } \hat{p} \leq 1/16 \end{cases} \quad (2)$$

The variance of the estimate, $V(\hat{q})$, is computed numerically as follows. For a given sample size, N, and population allele frequency, q, each possible outcome (number of autozygous probands), i = 0,...N, occurs with binomial probability, $B(p, N)$, where p is given by (1). For each i, there is an associated $\hat{p} = i/N$ and corresponding $\hat{q}$ as given by (2). The variance of the allele frequency estimate is then obtained as $V(\hat{q}) = E(\hat{q}^2) - E^2(\hat{q})$, where E stands for expectation (mean).

Dahlberg's Method

To compare the variance of our estimate for q to that of the conventional (Dahlberg's) estimate, we applied Dahlberg's method as follows. Assume that among all matings in a population, a known proportion c is between first cousins while all other matings are between unrelated individuals (in practice, the latter category includes the rare matings between individuals of other relationships). Then, the proportion of recessive cases born to cousin marriages among all recessive cases in the population is

$$k = c(1 + 15q)/[c(1 - q) + 16q] \quad (3)$$

[2], which may be viewed as being analogous to (1). Consider a sample of N probands of which an observed proportion, $\hat{k}$, has parents who are first cousins. Based on (3), this estimate of k can be translated into a maximum likelihood estimate, $\hat{q}_D$, of the allele frequency. Because some values of k may lead to values of q exceeding 1, we define

$$\hat{q}_D = \begin{cases} c(1 - \hat{k})/[\hat{k}(16 - c) - 15c] & \text{if } \hat{k} > c \\ 1 & \text{if } \hat{k} \leq c. \end{cases} \quad (4)$$

The variance, $V(\hat{q}_D)$, is calculated in analogy to the calculation described above leading to $V(\hat{q})$.

Results

One of the best-studied recessive traits, phenylketonuria, shows a population frequency in the US of approximately 1 in 12,000 [5], that is, a disease allele frequency of just about 0.01 while other recessive traits appear to be more common. For a range of population allele frequencies, table 1 shows the standard error (square root of variance) of our allele frequency estimate, $\hat{q}$, depending on the sample size of N probands. Clearly, standard errors are quite high and it takes considerable sample sizes for reasonably accurate allele frequency estimates. For example, with $q = 0.05$ and $N = 100$, the 95% confidence interval approximately ranges from 0.03 to 0.07.

For Dahlberg's method, the population frequency of first cousin matings among all marriages must be known. In western societies, this rate is on the order of $c = 0.001$ [6], while in some eastern countries, it can be as high as $c = 0.200$ or higher [7]. For a range of these values, table 2 shows the efficiency of our map-based estimate versus the Dahlberg estimate, where efficiency is defined in the customary manner, that is, as the variance ratio, $V(\hat{q}_D)/V(\hat{q})$. This ratio expresses the relative accuracy of the two estimates but, as outlined in the discussion, does not take into account the costs associated with sampling probands. For rare diseases ($q \leq 0.02$) and small sample size ($N \leq 20$), map-based estimation is seen to be always more efficient than the Dahlberg method (i.e. for any of the c values considered). For cousin marriage rates of $c \leq 0.10$, the map-based estimate is generally more efficient except for small N and very large q. The differences in accuracy can be quite dramatic. For example, for a common allele frequency of 0.05 and $c = 0.001$, efficiency is 7.1 for $N = 10$, and 1,427 for $N = 100$. Thus, in most situations, the map-based estimate is clearly superior to Dahlberg's estimate.

Discussion

As shown above, for small to moderate sample sizes and traits that are not too common, our new method is more efficient than Dahlberg's method. In practice, grandparents are often not typed and the number of probands in linkage studies of recessive traits tends to be rather small. Thus it is fortuitous that this is the situation in which our method shines.

Clearly, obtaining probands suitable for the map-based method requires more resources than obtaining probands for Dahlberg's method. In principle, all relevant ancestors of the former probands must be genotyped but such data

Table 1. Standard error of map-based estimate

Number	Population frequency of disease allele				
	0.010	0.020	0.050	0.100	0.200
5	0.0211	0.0441	0.1320	0.2508	0.3590
10	0.0111	0.0192	0.0547	0.1349	0.2457
20	0.0072	0.0118	0.0276	0.0680	0.1747
50	0.0043	0.0070	0.0153	0.0327	0.0839
100	0.0030	0.0048	0.0104	0.0216	0.0516
200	0.0021	0.0034	0.0073	0.0148	0.0344
1,000	0.0009	0.0015	0.0032	0.0065	0.0148

are difficult or costly to obtain. With very highly polymorphic markers, identity by state is almost equivalent to identity by descent so that it may be possible to obtain approximate solutions based on homozygosity versus heterozygosity at closely linked marker loci. Such approaches are under investigation.

For Dahlberg's method, the population proportion of cousin marriages must be known. If it is unknown or inaccurate, this method presumably furnishes biased results. There is no such requirement for the map-based method.

Our approach focuses on one proband per family. In linkage studies of recessive traits, there is often more than one affected individual per shibship. We have not yet investigated how information from multiplex sibships could be used in the map-based method. This does not appear to be a simple problem.

The method introduced in this paper is tailored for a specific relationship of the parents. This relationship – first cousins – is the most common one among related parents in western civilizations. In other, for example, eastern populations, other relationships may be more common for which our method is not directly applicable.

Table 2. Relative efficiency (accuracy) of map-based versus Dahlberg's method of allele frequency estimation

Number	Population frequency of disease allele				
	0.01	0.02	0.05	0.10	0.20
c = 0.001					
5	76.41	10.11	0.62	0.12	0.05
10	520.63	103.39	7.06	0.84	0.20
20	2,243.40	515.83	53.55	6.44	0.79
50	11,182.55	3,055.78	393.23	64.61	8.04
100	27,212.25	9,375.35	1,427.38	260.91	38.40
200	39,788.14	20,979.89	4,220.17	866.02	141.30
1,000	6,999.18	33,658.73	20,687.90	5,394.65	956.15
c = 0.010					
5	466.37	76.21	5.30	1.10	0.44
10	2,002.71	593.63	52.10	6.73	1.70
20	3,536.72	1,737.43	291.98	41.41	5.51
50	1,521.23	2,253.64	877.15	212.19	32.16
100	804.59	3,382.57	1,956.02	486.01	78.15
200	27.67	1,148.25	2,637.20	945.00	178.81
1,000	2.69	10.97	491.49	1,687.81	686.56
c = 0.100					
5	110.08	64.22	12.00	3.65	1.79
10	195.74	298.74	74.42	12.62	3.48
20	22.58	190.69	203.91	46.37	7.05
50	0.85	12.06	206.14	137.38	27.00
100	0.70	2.08	78.48	174.18	58.28
200	0.65	1.65	14.41	135.06	96.42
1,000	0.62	1.44	5.33	18.00	105.49

c = Population frequency of first-cousin matings.

Acknowledgment

This work was supported by grant HG00008 from the National Human Genome Research Institute.

References

1 Morton NE: Outline of Genetic Epidemiology. Basel, Karger, 1982.
2 Dahlberg G: Mathematical Methods for Population Genetics. Basel, Karger, 1947.
3 Li CC: First Course in Population Genetics. Pacific Grove, Boxwood Press, 1978.
4 Romeo G, Bianco M, Devoto M, Menozzi P, Mastella G, Giunta AM, Micalizzi C, Antonelli M, Battistini A, Santamaria F, et al: Incidence in Italy, genetic heterogeneity, and segregation analysis of cystic fibrosis. Am J Hum Genet 1985;37:338–349.
5 Vogel F, Motulsky AG: Human Genetics. New York, Springer, 1986.
6 Lebel RR: Consanguinity studies in Wisconsin I: Secular trends in consanguineous marriage, 1843–1981. Am J Med Genet 1983;15:543–560.
7 Khlat M, Khoury M: Inbreeding and diseases: Demographic, genetic, and epidemiologic perspectives. Epidemiol Rev 1991;13:28–41.