Am J Hum Genet 36:177–186, 1984

Estimation of the Marker Gene Frequency and Linkage Disequilibrium from Conditional Marker Data

ARAVINDA CHAKRAVARTI,[1] CHING CHUN LI, AND KENNETH H. BUETOW

SUMMARY

A method is proposed to calculate the maximum likelihood estimate of gene frequency and linkage disequilibrium from disease-codominant marker conditional data. The method is illustrated using data on sickle-cell anemia and Duchenne muscular dystrophy and linked polymorphic restriction endonuclease cleavage sites.

INTRODUCTION

With the advent of recombinant DNA technology, it is now possible to study variation in human DNA sequences using restriction endonucleases. These variations are detected as restriction fragment length polymorphisms (RFLPs) and a general method for discovering RFLPs near any structural gene currently exists [1]. There has recently been a concerted effort for discovering those polymorphisms that are closely linked to a disease gene for purposes of prenatal diagnosis [2]. It is therefore important to estimate the gene frequencies at the marker locus and the linkage disequilibrium between the disease and marker loci, because these quantities determine the usefulness of the marker locus in prenatal diagnosis [3, 4].

Several maximum likelihood methods for estimating gene frequency and linkage disequilibrium from a random sample of individuals are known [5]. However, a random sample will often fail to include all the genotypes at a disease locus since the frequency of the disease gene will generally be quite small. Therefore, several investigators [6–9] have studied the distribution of marker genotypes among individuals chosen randomly from given disease genotypes. The maximum likelihood estimation procedure that we suggest is based on such conditional marker

Received April 19, 1983; revised July 5, 1983.

Part of this research was supported by grant 2R01AM13983-14 from the National Institutes of Health and grant Y-77 from the Health Research and Services Foundation.

[1] All authors: Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261.

© 1984 by the American Society of Human Genetics. All rights reserved. 0002-9297/84/3601-0016$02.00

distributions. In fact, most studies of linkage disequilibrium between a disease locus and RFLPs are of this type [9].

AUTOSOMAL RECESSIVE DISEASE

Consider a disease locus with a normal allele D and a disease allele d, with known frequencies p and q, respectively ($p + q = 1$). Furthermore, let M_1 and M_2 be codominant marker alleles with frequencies m_1 and m_2, respectively ($m_1 + m_2 = 1$). We suppose that a particular study has investigated the numbers of M_1M_1, M_1M_2, and M_2M_2 genotypes among given DD, Dd, and dd genotypes. This situation can be presented as follows:

Given disease genotype	Marker genotypes			Total no.	Conditional frequency of M_2
	M_1M_1	M_1M_2	M_2M_2		
DD	n_{11}	n_{12}	n_{13}	N_1	$\hat{a}_1$
Dd	n_{21}	n_{22}	n_{23}	N_2	$\hat{a}_2$
dd	n_{31}	n_{32}	n_{33}	N_3	$\hat{a}_3$

In this table, the n_{ij}'s are the numbers of individuals with the jth marker genotype ($j = 1, 2, 3$) given that their disease status is i ($i = 1, 2, 3$). The numbers N_1, N_2, and N_3 are the sizes of the *separate* and *independent* samples taken from DD, Dd, and dd individuals, respectively, and they do not reflect the relative frequency of the DD, Dd, and dd genotypes in the population.

Recently, Kan and Dozy [6] discovered a polymorphism in the recognition site for the endonuclease Hpa I in people of African ancestry, this site being 5-kilobase (kb) pairs distal to the 3′ end of the β-globin structural gene and closely linked to it. Several studies [6–9] also demonstrated that in American blacks with sickle-cell anemia, a high proportion of β^S chromosomes are associated with a 13-kb fragment at the restriction marker locus. Two of these data sets are shown in table 1 with the simplification that two other common fragments, 7 kb and 7.6 kb, are pooled into a single 7-kb class.

TABLE 1

DISTRIBUTION OF Hpa I MARKER GENOTYPES AMONG NORMAL (AA), SICKLE-CELL TRAIT (AS), AND SICKLE-CELL ANEMIA (SS) INDIVIDUALS

Given disease genotype	Marker genotypes (Hpa I)			Total no.	Conditional frequency of "13"
	7,7	7,13	13,13		
A. Kan and Dozy (1980) [7]: San Francisco					
AA	39	4	0	43	4/86 = .047 ± .023
AS	32	53	1	86	55/172 = .320 ± .027
SS	4	24	30	58	84/116 = .724 ± .042
B. Panny et al. (1981) [9]: Baltimore					
AA	14	3	0	17	3/34 = .088 ± .049
AS	4	11	1	16	13/32 = .406 ± .066
SS	5	7	8	20	23/40 = .575 ± .078

For the two-locus model mentioned above, there are four segregating gametes, namely, DM_1, DM_2, dM_1, and dM_2. Their frequencies may be written as:

$$\begin{aligned} (DM_1)\, y_1 &= pm_1 + \epsilon\ , \qquad (DM_2)\, y_2 = pm_2 - \epsilon\ , \qquad y_1 + y_2 = p \\ (dM_1)\, y_3 &= qm_1 - \epsilon\ , \qquad (dM_2)\, y_4 = qm_2 + \epsilon\ , \qquad y_3 + y_4 = q\ , \end{aligned} \tag{1}$$

where $\epsilon = y_1y_4 - y_2y_3$ is an index of linkage disequilibrium, and p and q are fixed known values.

The conditional frequency of allele M_2 in the classes DD, dd, and Dd are, respectively,

$$a_1 = \text{Prob}\,(M_2|DD) = (y_1y_2 + y_2^2)/p^2 = y_2/p\ , \tag{2a}$$

$$a_3 = \text{Prob}\,(M_2|dd) = (y_3y_4 + y_4^2)/q^2 = y_4/q\ , \tag{2b}$$

and,

$$\begin{aligned} a_2 &= \text{Prob}\,(M_2|Dd) = (y_1y_4 + y_2y_3 + 2y_2y_4)\,/2pq \\ &= (y_2/p + y_4/q)/2 = \tfrac{1}{2}(a_1 + a_3)\ . \end{aligned} \tag{2c}$$

In terms of m_2, these may be written as, using equation (1),

$$\begin{aligned} a_1 &= m_2 - \epsilon/p\ , \qquad a_3 = m_2 + \epsilon/q \\ a_2 &= m_2 + \epsilon\,(p - q)/2pq = \tfrac{1}{2}(a_1 + a_3)\ . \end{aligned} \tag{3}$$

Since these are conditional frequencies of M_2 for DD, Dd, and dd genotypes, the frequency of M_2 in the population is:

$$m_2 = p^2a_1 + 2pqa_2 + q^2a_3\ , \tag{4}$$

by the well-known theorem $\text{P(A)} = \Sigma_i\, \text{P(A}|\text{H}_i) \cdot \text{P(H}_i)$, where H_i are the conditions.

There are three solutions for ϵ from equation (3):

$$\epsilon = pq(a_3 - a_1)\ ,\ \epsilon = 2pq(a_3 - a_2)\ ,\ \epsilon = 2pq(a_2 - a_1)\ . \tag{5}$$

The average value of ϵ from solutions (5) is

$$\epsilon = pq(a_3 - a_1)\ . \tag{6}$$

The last formula has also been obtained by Nei and Li [10] by a different procedure.

For homozygous genotypes DD or dd, the distribution of the marker genotypes is in Hardy-Weinberg proportions. Hence, the maximum likelihood estimates of a_1 and a_3 are

$$\hat{a}_1 = \frac{n_{12} + 2n_{13}}{2N_1}, \qquad \hat{a}_3 = \frac{n_{32} + 2n_{33}}{2N_3}, \tag{7}$$

with variance

$$V(\hat{a}_1) = \frac{a_1(1 - a_1)}{2N_1}, \qquad V(\hat{a}_3) = \frac{a_3(1 - a_3)}{2N_3}. \tag{7V}$$

But, given heterozygous genotype *Dd*, the distribution of the marker genotypes is not in the Hardy-Weinberg form. The APPENDIX shows that the maximum likelihood method yields the estimate:

$$\hat{a}_2 = \frac{n_{22} + 2n_{23}}{2N_2}, \qquad V(\hat{a}_2) = \frac{a_2(1 - a_2)}{2N_2} - \frac{(a_1 - a_3)^2}{8N_2}. \tag{8}$$

What $\hat{a}_2$ estimates is actually the average conditional frequencies of M_2 for *DD* and *dd* genotypes; that is, $(a_1 + a_3)/2$.

Therefore, one may estimate the M_2 gene frequency and the linkage disequilibrium by

$$\hat{m}_2 = p^2\hat{a}_1 + 2pq\hat{a}_2 + q^2\hat{a}_3 \tag{9}$$

and

$$\hat{\epsilon} = pq(\hat{a}_3 - \hat{a}_1), \tag{10}$$

with variances given by

$$V(\hat{m}_2) = p^4V(\hat{a}_1) + 4p^2q^2V(\hat{a}_2) + q^4V(\hat{a}_3) \tag{11}$$

and

$$V(\hat{\epsilon}) = p^2q^2[V(\hat{a}_1) + V(\hat{a}_3)]. \tag{12}$$

The estimates $\hat{m}_2$ and $\hat{\epsilon}$ are maximum likelihood estimates (MLE) because they are linear combinations of MLEs. Our estimation procedure requires knowledge of the gene frequency q, which is usually obtained from large surveys. Thus, for sickle-cell anemia in the United States, $q \approx .05$ [11].

AUTOSOMAL DOMINANT DISEASE

For autosomal dominant diseases, affected homozygotes (*dd*) are extremely rare ($q^2 \approx 0$), and so only affected *Dd* individuals will usually be sampled. Thus, in all probability, $N_3 = 0$. In this case, our estimation procedure can still be used with the following changes. Observe that $p^2 \approx 1 - 2q$ and $2pq \approx 2q$. Therefore, the estimates are given by:

$$\hat{m}_2 = (1 - 2q)\,\hat{a}_1 + 2q\hat{a}_2 \,, \tag{13}$$

$$\hat{\epsilon} = 2pq\,(\hat{a}_2 - \hat{a}_1) \,, \tag{14}$$

with variances

$$V(\hat{m}_2) = (1 - 2q)^2\,V(\hat{a}_1) + 4q^2\,V(\hat{a}_2) \,, \tag{15}$$

$$V(\hat{\epsilon}) = 4p^2q^2[V(\hat{a}_1) + V(\hat{a}_2)] \,. \tag{16}$$

X-LINKED RECESSIVE DISEASE

The situation for the sex-linked recessive diseases is identical with autosomal dominants since affected females are usually not observed. Therefore, equations (13–16) can be directly used. However, there is the additional possibility that marker genotypes may be studied among a sample of normal and affected males. Thus, from normal and affected males, one may estimate b_1 and b_3, respectively, the conditional gene frequency of the M_2 allele. If the male and female gamete frequencies [equation (1)] are equal, then $b_1 = a_1$ and $b_3 = a_3$. The male data can be used to test this possibility.

NUMERICAL EXAMPLES

Sickle-cell Anemia (SCA)

For SCA, there are two major studies in different geographical regions of the United States that may be used to illustrate our estimation procedure. The data that we have used are presented in table 1 along with the references to the studies. These data allow computations of $\hat{m}_2$ and $\hat{\epsilon}$ from equations (9–12) with the assumption that $q = .05$. The vaues we obtained are as follows:

Area	$\hat{m}_2$	$\hat{\epsilon}$
San Francisco	.0742 ± .0207	.0322 ± .0022
Baltimore	.1197 ± .0443	.0231 ± .0044

The standard errors associated with each estimate of the gene frequency m_2 are high and show that none of the values are significantly different from each other. However, it is more important to test if $\hat{\epsilon} = 0$. For this purpose, the statistic $\hat{\epsilon}^2/V(\hat{\epsilon})$ based on a χ^2 distribution with 1 df may be used. The two samples give $\chi^2_1 = 205.2$ (San Francisco) and $\chi^2_1 = 28.0$ (Baltimore), both of which are highly significant ($P < .00001$ in both cases). This demonstrates that we can reject the hypothesis of no association between β-globin and Hpa I.

Duchenne Muscular Dystrophy (DMD)

Recently, a polymorphic Taq I endonuclease site has been discovered 10 centimorgans away from the DMD locus [12]. This RFLP has two alleles, $M_1 = 3.2$ kb and $M_2 = 5.3$ kb. The data collected are presented in table 2, and for this

TABLE 2

DISTRIBUTION OF Taq I MARKER GENOTYPES AMONG NORMAL MALES (D), NORMAL FEMALES (DD), AND CARRIER FEMALES (Dd) AT THE DUCHENNE MUSCULAR DYSTROPHY LOCUS

GIVEN DISEASE GENOTYPE	MARKER GENOTYPE (Taq 1)			TOTAL NO.	CONDITIONAL FREQUENCY OF "5.3"
	MALE				
	3.2		5.3		
D	20		3	23	3/23 = .130 ± .070
	FEMALE				
	3.2, 3.2	3.2, 5.3	5.3, 5.3		
DD	19	7	0	26	7/52 = .135 ± .047
Dd	36	9	0	45	9/90 = .100 ± .032

data, $q = 1/5{,}000 = .0002$ has been assumed [12]. Since $\hat{a}_1 = .135 \pm .047$ and $\hat{a}_2 = .100 \pm .030$ for the females, we obtain from equations (13–16): $\hat{m}_2$(5.8-kb allele) = .1346 ± .0473 and $\hat{\epsilon} = -.0000138 \pm .0000224$. It is clear from the estimate of $\hat{\epsilon}$ that it is not significantly different from zero ($\chi^2_1 = 0.383$, $.52 < P < .58$). It is, in general, expected that the disequilibrium between two loci that are 10 centimorgans apart will be small and negligible.

The value of $\hat{b}_1$ is .130 ± .070 from the males and is not significantly different from $\hat{a}_1$($\chi^2_1 = 0.049$, $.82 < P < .84$), so that we may conclude that the male and female gamete frequencies are the same, albeit, from the limited sampling conducted.

DISCUSSION

We have provided a simple method to estimate the marker gene frequencies and linkage disequilibrium coefficients from conditional data and used simple statistical tests of hypotheses. One should note that the χ^2 test used may not be applicable unless the sample sizes are large. In fact, the distribution of ϵ may be quite skewed so that the P values reported may be incorrect. Our analyses also depend on the assumption that the population is in equilibrium so that the gamete frequencies are constant. Furthermore, we assume that the effects of selection and recombination are small and negligible, at least in a single generation.

One should also notice that in the two samples the gene frequency in San Francisco is less than that in Baltimore but the linkage disequilibrium in San Francisco is greater than that in Baltimore. It is known that the 13-kb allele is primarily of African origin and that Caucasians in this country are generally monomorphic for the 7-kb allele [7–9]. Therefore, if the blacks in the United States were all derived from the same ancestral population, the different values of m_2 could be accounted for by differential migration of the 7-kb allele. We would therefore expect more Caucasian gene admixture in San Francisco than in Baltimore. In general, since migration will contribute A7 chromosomes, we will also expect an increase in the linkage disequilibrium as observed.

The estimates of gene frequency and linkage disequilibrium obtained by our procedure may be used to evaluate the usefulness of a marker locus for prenatal diagnosis [3]. This is important because these parameters crucially affect the proportion of informative pregnancies, that is, those pregnancies where the marker locus can be used to diagnose the disease status accurately. In reality, more than two alleles may exist at the marker locus, such as for HLA. For these situations, the estimation procedure may be easily generalized. Thus, if $M_1, M_2, \ldots, M_k$ are marker alleles with frequencies $m_1, m_2, \ldots, m_k$ ($m_1 + m_2 + \ldots + m_k = 1$), the gamete frequencies may be written as: $DM_j : y_j = pm_j + \epsilon_j$, $dM_j : y_{k+j} = qm_j - \epsilon_j$, where $j = 1, 2, \ldots, k$ and

$$\sum_{j=1}^{k} \epsilon_j = 0 .$$

Let us also suppose that for the ith disease phenotype class ($i = 1, 2, 3$), the conditional frequency marker M_j is a_{ij}. Then, $a_{1j} = y_j/p$, $a_{3j} = y_{k+j}/q$, and $a_{2j} = (a_{1j} + a_{3j})/2$. Therefore, as before, $\hat{m}_j = p^2\hat{a}_{1j} + 2pq\hat{a}_{2j} + q^2\hat{a}_{3j}$ and $\hat{\epsilon}_j = -pq(\hat{a}_{3j} - \hat{a}_{1j})$ for $j = 1, 2, \ldots, k$. Note that because

$$\sum_{j=1}^{k} \epsilon_j = 0 ,$$

only $k - 1$ of the ϵ_j's are independent.

Finally, we have to emphasize the fact that our procedure is unique to the type of data analyzed. Thus, not only should the individuals chosen in any particular disease locus phenotype class (*DD*, *Dd*, *dd*) be independent, but individuals in any one class (say, *Dd*) should not be biologically related to individuals in any other class (say, *dd*). If related individuals are included in a sample, then a more complex maximum likelihood pedigree analysis method needs to be used.

ACKNOWLEDGMENTS

We thank S. Bale, P. P. Majumder, and an anonymous reviewer for comments on this manuscript.

REFERENCES

1. Botstein D, White RL, Skolnick M, Davis RW: Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314–331, 1980
2. Orkin SH: Genetic diagnosis of the fetus. *Nature* 296:202–203, 1982
3. Chakravarti A: Utility and efficiency of linked marker genes in genetic counseling. III. Proportion of informative families under linkage disequilibrium. *Am J Hum Genet* 35:592–610, 1983
4. Nei M: Proportion of informative families for genetic counseling with linked marker genes. *Jpn J Hum Genet* 24:131–142, 1979
5. Hill WG: Estimation of linkage disequilibrium in randomly mating populations. *Heredity* 33:229–239, 1974
6. Kan YW, Dozy AM: Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. *Proc Natl Acad Sci USA* 75:5631–5635, 1978

7. KAN YW, DOZY AM: Evolution of the hemoglobin S and C genes in world populations. *Science* 209:388–391, 1980
8. FELDENZER J, MEARS JG, BURNS AL, NATTA C, BANK A: Heterogeneity of DNA fragments associated with the sickle-globin gene. *J Clin Invest* 64:751–755, 1979
9. PANNY SR, SCOTT AF, SMITH KD, ET AL.: Population heterogeneity of the *Hpa I* restriction site associated with the β-globin gene: implications for prenatal diagnosis. *Am J Hum Genet* 33:25–35, 1981
10. NEI M, LI W-H: Non-random association between electromorphs and inversion chromosomes in finite populations. *Genet Res (Camb)* 35:65–83, 1980
11. LIVINGSTONE FB: *Abnormal Hemoglobins in Human Populations*. Chicago, Aldine, 1967
12. MURRAY JM, DAVIES KE, HARPER PS, MEREDITH L, MUELLER CR, WILLIAMSON R: Linkage relationship of cloned DNA sequence on the short arm of the X chromosome to Duchenne muscular dystrophy. *Nature* 300:69–71, 1982

APPENDIX
MAXIMUM LIKELIHOOD ESTIMATION OF a_2

From the gamete frequencies defined in equation (1), we obtain the following population composition and conditional frequencies of the marker genotypes for given disease genotypes, where $a_1 = y_2/p$ and $a_3 = y_4/q$.

	Population composition				Conditional distribution		
	M_1M_1	M_1M_2	M_2M_2		M_1M_1	M_1M_2	M_2M_2
DD	y_1^2	$2y_1y_2$	y_2^2	p^2	$(1 - a_1)^2$	$2a_1(1 - a_1)$	a_1^2
Dd	$2y_1y_3$	$2y_1y_4 + 2y_2y_3$	$2y_2y_4$	$2pq$	$(1 - a_1)(1 - a_3)$	$a_1(1 - a_3) + (1 - a_1)a_3$	a_1a_3
dd	y_3^2	$2y_3y_4$	y_4^2	q^2	$(1 - a_3)^2$	$2a_3(1 - a_3)$	a_3^2

The conditional distributions of the marker genotypes for *DD* and *dd* are in Hardy-Weinberg proportions; the estimates of a_1 and a_3 and their variances are given in equations (7) and (7V) and need no comment. To estimate the frequency of M_2 from the conditional distribution given the disease genotype *Dd*, we find the following reparameterization convenient: sum, $s = a_1 + a_3$; product, $t = a_1a_3$. In terms of these new parameters, the method of maximum likelihood estimation may be readily applied. Consider the following sample:

	M_1M_1	M_1M_2	M_2M_2	Total
Probability:	$u = 1 - s + t$	$v = s - 2t$	t	1.00
Observed:	n_1	n_2	n_3	N

Then, the log likelihood is given by $L = n_1 \log u + n_2 \log v + n_3 \log t + \text{constant}$. To obtain the MLEs of s and t, we solve the following equations:

$$\frac{\partial L}{\partial s} = \frac{-n_1}{u} + \frac{n_2}{v} = 0$$

$$\frac{\partial L}{\partial t} = \frac{n_1}{u} - \frac{2n_2}{v} + \frac{n_3}{t} = 0 .$$

Adding these two equations, we obtain:

$$\frac{n_2}{s - 2t} = \frac{n_3}{t}, \qquad t = \frac{n_3 s}{n_2 + 2n_3},$$

so that substituting for t in $\partial L/\partial s = 0$, we obtain: $n_1 (s - 2t) = n_2 (1 - s + t)$, which on solving yields:

$$\hat{s} = \frac{n_2 + 2n_3}{N}, \qquad \hat{t} = \frac{n_3}{N}.$$

Hence,

$$\hat{a}_2 = \widehat{½s} = ½(a_1 + a_3) = \frac{n_2 + 2n_3}{2N},$$

as we calculated in table 1. To obtain the variances of s and t, we calculate the second derivatives as:

$$\frac{\partial^2 L}{\partial s^2} = -\frac{n_1}{u^2} - \frac{n_2}{v^2}; \qquad \frac{\partial^2 L}{\partial s \partial t} = \frac{n_1}{u^2} + \frac{2n_2}{v^2};$$

$$\frac{\partial^2 L}{\partial t^2} = -\frac{n_1}{u^2} - \frac{4n_2}{v^2} - \frac{n_3}{t^2}.$$

Since the expected values of n_1, n_2, and n_3 are Nu, Nv, and Nt, respectively, the information with respect to s and t are

$$-E\left(\frac{\partial^2 L}{\partial s^2}\right) = N \cdot \left(\frac{1}{u} + \frac{1}{v}\right) = I_{ss};$$

$$-E\left(\frac{\partial^2 L}{\partial s \partial t}\right) = -N \cdot \left(\frac{1}{u} + \frac{2}{v}\right) = I_{st};$$

$$-E\left(\frac{\partial^2 L}{\partial t^2}\right) = N \cdot \left(\frac{1}{u} + \frac{4}{v} + \frac{1}{t}\right) = I_{tt}.$$

The information matrix is

$$I = \begin{bmatrix} I_{ss} & I_{st} \\ I_{st} & I_{tt} \end{bmatrix}, \text{ so that } I^{-1} = \frac{1}{\Delta}\begin{bmatrix} I_{tt} & -I_{st} \\ -I_{st} & I_{ss} \end{bmatrix},$$

where the determinant $\Delta = I_{ss}I_{tt} - I^2_{st} = N^2/uvt$. Then the variance and covariance of the estimates are

$$V(\hat{t}) = \frac{I_{ss}}{\Delta} = \frac{t(1 - t)}{N},$$

$$\operatorname{cov}(\hat{s}, \hat{t}) = \frac{-I_{st}}{\Delta} = \frac{t(2 - s)}{N},$$

$$V(\hat{s}) = \frac{I_{tt}}{\Delta} = \frac{vt + 4ut + uv}{N} = \frac{s(1 - s) + 2t}{N},$$

and

$$V(\hat{a}_2) = \tfrac{1}{4}V(\hat{s}) = \frac{a_2(1 - 2a_2) + a_1a_3}{2N}$$

$$= \frac{a_2(1 - a_2) - \tfrac{1}{4}(a_1 - a_3)^2}{2N}.$$

If there is no disequilibrium, $a_1 = a_2 = a_3$, then $V(\hat{a}_2) = a_2(1 - a_2)/2N$, correctly.