Am. J. Hum. Genet. 48:926–934, 1991

Testing for Association between Disease and Linked Marker Loci: A Log-linear-Model Analysis

Laurence Tiret,* Philippe Amouyel,† Roger Rakotovao,* Francois Cambien,* and Pierre Ducimetière*

*Institut National de la Santé et de la Recherche Médicale (INSERM) Unité 258, Hôpital Broussais, Paris; and †Institut Pasteur, Service de Santé Publique, Lille, France

Summary

One approach frequently used for identifying genetic factors involved in the process of a complex disease is the comparison of patients and controls for a number of genetic markers near a candidate gene. The analysis of such association studies raises some specific problems because of the fact that genotypic and not gametic data are generally available. We present a log-linear-model analysis providing a valid method for analyzing such studies. When studying the association of disease with one marker locus, the log-linear model allows one to test for the difference between allelic frequencies among affected and unaffected individuals, Hardy-Weinberg (H-W) equilibrium in both groups, and interaction between the association of alleles at the marker locus and disease. This interaction provides information about the dominance of the disease susceptibility locus, with dominance defined using the epidemiological notion of odds ratio. The degree of dominance measured at the marker locus depends on the strength of linkage disequilibrium between the marker locus and the disease locus. When studying the association of disease with several linked markers, the model becomes rapidly complex and uninterpretable unless it is assumed that affected and unaffected populations are in H-W equilibrium at each locus. This hypothesis must be tested before going ahead in the analysis. If it is not rejected, the log-linear model offers a stepwise method of identification of the parameters causing the difference between populations. This model can be extended to any number of loci, alleles, or populations.

Introduction

The existence of a genetic component in the etiology of multifactorial diseases, such as cardiovascular diseases, cancer, or psychiatric disorders, is now firmly established, and the identification of susceptibility loci for such diseases is of considerable interest for understanding their mechanisms and improving their prevention. One approach for examining the contribution of genes is to test for association between genetic markers and disease. The rationale underlying this approach is that if the marker locus and the unknown disease locus are in linkage disequilibrium, some specific alleles of the marker should be found more frequently among affected than among unaffected people. The mapping of a great number of new markers near candidate genes (genes possibly involved in the disease process) has contributed to the extension of such association studies.

These studies generally involve the comparison of groups of patients and controls who have been genotyped for a number of markers in the vicinity of a candidate gene. When several markers are investigated in the same study, they are generally selected on the basis of linkage disequilibrium existing between them, in order to increase the chance of detecting an association with the disease locus. However, the degree of disequilibrium must not be so high that the information content of haplotypes is decreased too much.

The statistical analysis of association studies raises some specific difficulties which are generally neglected or circumvented by authors. For example, the association of disease with one marker is treated by compar-

Received October 1, 1990; revision received December 21, 1990.

Address for correspondence and reprints: Laurence Tiret, INSERM U.258, Hôpital Broussais, 96 rue Didot, 75674 Paris, Cedex 14, France.

0002-9297/91/4805-0011$02.00

ing either allelic frequencies or genotypic frequencies between the two groups, ignoring the problem of interaction between alleles at the marker locus, although this problem has been treated in theoretical papers (Norwood and Hinkelmann 1978; Weir 1979; Khamis and Hinkelmann 1984).

When analyzing several markers simultaneously, a second problem arises in the determination of haplotypes for individuals heterozygous at several loci, since it is generally not possible to determine whether alleles are in coupling or repulsion phase. In the absence of available information from parental genotypes, these individuals are generally excluded from analysis, which may introduce a bias into the results. Methodological papers concerned with the problem of estimation of haplotypic frequencies and linkage disequilibrium from genotypic data have been published (Hill 1975; Weir 1979; Weir and Wilson 1986), but these papers have been confined to the case of one population, whereas it is the comparison of two populations which is the main goal of association studies.

In this paper, we propose an extension to the case of two populations for a log-linear analysis proposed by Weir and Wilson (1986) for one population. We will treat first the problem of association of disease with one marker, then the problem of association with two linked markers. For ease of presentation, we will consider diallelic markers, but an extension to multiallelic markers will be outlined at the end of the paper. Other extensions to more than two markers or more than two populations will also be discussed.

Method

One Locus

Consider one diallelic locus A, with alleles A and a. As the genotype Aa generally cannot be distinguished from the genotype aA, it will be assumed that allele proportions are the same in the maternal and the paternal populations, implying that the expected frequency of Aa individuals is the same as that of aA individuals. Using a multiplicative log-linear-model approach similar to that of Weir and Wilson (1986), we can write the three possible frequencies of individuals in terms of three quantities M, M_a, and M_{Aa}:

$$\begin{aligned} P_{AA} &= \mathrm{M}\ , \\ P_{Aa} &= 2\,\mathrm{M}\,M_a\,M_{Aa}\ , \\ P_{aa} &= \mathrm{M}\,M_a^2\ . \end{aligned}$$

M is a normalizing factor allowing us to work either with counts or with frequencies, M_a is a parameter relative to allelic frequency, and M_{Aa} is a term of interaction between alleles which measures the deviation from Hardy-Weinberg (H-W) equilibrium at locus A. There are three parameters to be fitted to three counts. Parameters and their standard errors are estimated by maximizing the likelihood of the sample, and the likelihood ratio test allows a test of each effect for significance. In a general way, if L_1 is the log likelihood of a general model with k_1 parameters and L_2 the log likelihood of a nested restricted model with k_2 parameters, under certain conditions, $2(L_1\text{-}L_2)$ is approximately distributed as a χ^2 with $k_1 - k_2$ df (Weir 1990). Thus, the departure from H-W equilibrium is tested by comparing the model estimating the three parameters to a submodel restricting M_{Aa} to 1 (χ^2 with 1 df).

Consider now two populations, one of affected and one of unaffected individuals. Again, because genotypes Aa and aA cannot be distinguished, we have to make a further assumption, that is, there are no maternal or paternal effects on disease incidence. The three frequencies of individuals in each population can be written as above in terms of the quantities M(1), $M_a(1)$, $M_{Aa}(1)$ and $M(0)$, $M_a(0)$, $M_{Aa}(0)$ (1 indexing affected and 0 unaffected individuals). Again, parameters are estimated by maximizing the log likelihood of the whole sample, which is the sum of the two log likelihoods of affected and unaffected samples. A global test for heterogeneity between the two populations is provided by comparing this log likelihood to that of a sample pooling the two populations, which is equivalent to setting restrictions of $M(1) = M(0)$, $M_a(1) = M_a(0)$, and $M_{Aa}(1) = M_{Aa}(0)$ (χ^2 with 3 df). Other tests can be performed by restricting only some of the parameters, for example, $M_{Aa}(1) = M_{Aa}(0)$ (χ^2 with 1 df). The nonsignificance of this test indicates either that the deviation from H-W equilibrium is the same in the two populations or that the two populations are in H-W equilibrium (this last hypothesis can be tested by setting the restriction $M_{Aa}(1) = M_{Aa}(0) = 1$ [χ^2 with 2 df]).

We can also give an epidemiological interpretation to the hypothesis $M_{Aa}(1) = M_{Aa}(0)$. Consider the odds ratios (ORs) for the disease associated with different genotypes:

$$\begin{aligned} \mathrm{OR(Aa/AA)} &= [P_{Aa}(1)\times P_{AA}(0)]/[P_{Aa}(0)\times P_{AA}(1)] \\ &= [M_a(1)/M_a(0)]\times[M_{Aa}(1)/M_{Aa}(0)]\ , \\ \mathrm{OR(aa/Aa)} &= [P_{aa}(1)\times P_{Aa}(0)]/[P_{aa}(0)\times P_{Aa}(1)] \\ &= [M_a(1)/M_a(0)]\times[M_{Aa}(0)/M_{Aa}(1)]\ , \\ \mathrm{OR(aa/AA)} &= [P_{aa}(1)\times P_{AA}(0)]/[P_{aa}(0)\times P_{AA}(1)] \\ &= [M_a(1)/M_a(0)]^2\ . \end{aligned}$$

We see that the condition $M_{Aa}(1) = M_{Aa}(0)$ implies that $OR(Aa/AA) = OR(aa/Aa) = \sqrt{OR(aa/AA)} = M_a(1)/M_a(0)$. With reference to genotype AA, it means that the OR of individuals carrying two doses of allele a is the square of the OR of individuals carrying one dose of a. Given the multiplicative model adopted, this can be interpreted as absence of dominance, in the population genetic sense of this term indicating that no allele dominates the other (Jacquard 1974). When $M_{Aa}(1)/M_{Aa}(0) = M_a(1)/M_a(0)$, it implies that OR(Aa/AA) = OR(aa/AA) and OR(aa/Aa) = 1, hence specifying complete dominance. Conversely, when $M_{Aa}(1)/M_{Aa}(0) = M_a(0)/M_a(1)$, it specifies recessivity.

Actually, the concept of dominance refers to the underlying disease-susceptibility locus S, whereas it is the marker locus A which is observed. The degree of dominance measured at the marker locus will depend on the strength of linkage disequilibrium between the two loci. If the two loci are in tight disequilibrium, the degree of dominance observed at the marker locus will reflect almost exactly that of the disease locus. For example, it can be shown that if the two loci are in complete disequilibrium—that is, haplotypes carrying the allele s at locus S always carry the allele a at locus A (s and a being conventionally the alleles leading to an increasing risk)—and if the disease locus is "recessive," then the marker locus will also exhibit recessivity (see App. A).

The degree of dominance measured at the marker locus will decrease with the decline of linkage disequilibrium between the two loci, so that two different situations could be observed:

1. No association between the marker locus and the disease-susceptibility locus: $M_a(1) = M_a(0)$ and $M_{Aa}(1) = M_{Aa}(0)$
2. Association between the marker locus and the disease-susceptibility locus: (*a*) absence of dominance of the disease-susceptibility locus, that is, heterozygotes have a risk strictly intermediate between low and high homozygotes (OR[ss/Ss] = OR[Ss/SS]): $M_a(1) > M_a(0)$ and $M_{Aa}(1) = M_{Aa}(0)$; (*b*) dominance of the disease-susceptibility locus, with heterozygotes having a risk closer to that of high homozygotes (OR[ss/Ss] < OR [Ss/SS]): $M_a(1)/M_a(0) \geq M_{Aa}(1)/M_{Aa}(0) > 1$; (*c*) dominance of the disease-susceptibility locus, with heterozygotes having a risk closer to that of low homozygotes (OR[ss/Ss] > OR[Ss/SS]): $M_a(0)/M_a(1) \leq M_{Aa}(1)/M_{Aa}(0) < 1$

In all cases, we should observe $M_{Aa}(1)/M_{Aa}(0) \leq M_a(1)/M_a(0)$. However, these relations should be verified in the case where there is no interaction between the disease locus and other genetic or environmental factors. In reality, the situation observed could be more complex.

Thus, the test on $M_{Aa}(1)$ and $M_{Aa}(0)$ provides information both on H-W equilibrium in the two populations and on the degree of dominance (as defined above) of the disease-susceptibility locus under study. An application will be shown in the next section.

Two Loci

Consider now one population and two diallelic loci A and B with alleles A, a and B, b, respectively. An extension of the previous model to two loci is the model proposed by Weir and Wilson (1986). The 10 possible frequencies of individuals are written in terms of 10 quantities M, M_a, M_b, M_{Aa}, M_{Bb}, $M_{abintra}$, $M_{abinter}$, M_{aab}, M_{abb}, and M_{abab}, where M, M_a, M_b, M_{Aa}, and M_{Bb} are defined as previously, $M_{abintra}$ and $M_{abinter}$ measure the intra- and intergametic linkage disequilibria, and M_{aab}, M_{abb}, and M_{abab} measure trigenic and quadrigenic disequilibria. In the usual situation in which double heterozygotes AB/ab and Ab/aB cannot be distinguished, there are only nine observable frequencies of individuals, and to reduce the number of parameters to nine, the absence of quadrigenic disequilibrium could be assumed. Moreover, in that case the parameters $M_{abintra}$ and $M_{abinter}$ cannot be estimated separately. Weir and Wilson (1986) proposed to use a composite measure of linkage disequilibrium based on the sum and the product of $M_{abintra}$ and $M_{abinter}$. Thus, their formulation allows a test for digenic and trigenic disequilibrium even when data on double heterozygotes are ambiguous. However, as Weir and Wilson themselves recognize, the practical interpretation of such effects is not easy. In fact, the existence of an intergametic measure of linkage disequilibrium could result from a departure from H-W equilibrium at one of the loci, as they showed in an example. The interpretation of these effects is even more difficult when the purpose is to compare two populations. Therefore, we shall assume subsequently that there is no intergametic disequilibrium, which is equivalent to assuming that the population is in H-W equilibrium at each locus. The parameters of the model are then M, M_a, M_b, and $M_{abintra}$ (the intra subscript will be omitted henceforth in the text). The nine expected frequencies are given in appendix B.

In the case in which the population under consideration is in H-W equilibrium, the model proposed by Weir and Wilson (1986) is mathematically equivalent to that proposed by Hill (1975) using haplotypic frequencies, and the frequencies f_{AB}, f_{Ab}, f_{aB}, and f_{ab} of the four possible haplotypes can be derived from the parameters of the model by the formulas $f_{AB} = \sqrt{M/N}$, $f_{Ab} = M_b\sqrt{M/N}$, $f_{aB} = M_a\sqrt{M/N}$, and $f_{ab} = M_aM_bM_{ab}\sqrt{M/N}$, where N is the total size of sample.

Consider now two populations, one of affected and one of unaffected individuals. Assuming that both populations are in H-W equilibrium, the log likelihood of the whole sample can be written in terms of eight parameters, $M(1)$, $M_a(1)$, $M_b(1)$, $M_{ab}(1)$, and $M(0)$, $M_a(0)$, $M_b(0)$, $M_{ab}(0)$. A global test for verifying the initial hypothesis that there is H-W equilibrium in both populations can be carried out by comparing the log likelihood of the model including these eight parameters to that of the saturated model, which is $\ln L = \Sigma_i n_{i1}\log(n_{i1}/N_1) + \Sigma_i n_{i0}\log(n_{i0}/N_0)$, for $i = 1, 2, \ldots 9$, $N_1 = \Sigma_i n_{i1}$, and $N_0 = \Sigma_i n_{i0}$.

If this first test is found to be significant, it indicates that the model resting on the assumption that there is no departure from H-W equilibrium in either population is not appropriate. The analysis should not be carried on using this model, since prior assumptions are not verified and the following tests would not be valid. However, the significance of this prior test already provides information. Actually, a significant test indicates that one or both populations are in H-W disequilibrium. Since the control population is supposed to be in H-W equilibrium (a condition which can be verified), it means that the affected population is in H-W disequilibrium at one or both of the two marker loci considered. If the two populations have been correctly sampled, this probably means, as shown in the case of one locus, that the marker loci are in linkage disequilibrium with the disease locus, and that the disease locus is dominant in the sense given above.

If the previous test is nonsignificant, the analysis can be carried on and a series of nested models can be tested. A first test can be performed to determine whether the two populations differ in the degree of linkage disequilibrium between the two loci A and B. This can be achieved by setting the restriction $M_{ab}(1) = M_{ab}(0)$ in the model including the eight parameters (χ^2 with 1 df). If this test is nonsignificant, the tests successively setting the restrictions $M_a(1) = M_a(0)$ and $M_b(1) = M_b(0)$ will indicate whether the two populations differ in the marginal frequencies of alleles a and b, respectively.

As shown above, haplotypic frequencies can be derived in each population from the parameters of the model. Hence, it is possible to estimate the OR for a disease associated with a given haplotype. For example, the OR associated with haplotype ab, with reference to haplotype AB, is

$$\begin{aligned}\mathrm{OR(ab/AB)} &= f_{ab}(1)f_{AB}(0)/f_{ab}(0)f_{AB}(1)\\ &= M_a(1)M_b(1)M_{ab}(1)/M_a(0)M_b(0)M_{ab}(0)\ .\end{aligned}$$

Extensions of the Model

Three Loci

The model can be easily extended to more than two loci, assuming H-W equilibrium at each locus. For example, in the case of three loci, there are 27 observable frequencies of individuals in the case in which double and triple heterozygotes are undistinguishable. In each population, the corresponding expected frequencies are written in terms of eight quantities, M, M_a, M_b, M_c, M_{ab}, M_{ac}, M_{bc}, and M_{abc}. The strategy of analysis is the same as for two loci. The OR associated with haplotype abc, with reference to haplotype ABC, will be

$$\begin{aligned}&\mathrm{OR(abc/ABC)}\\ &\quad= f_{abc}(1)f_{ABC}(0)/f_{abc}(0)f_{ABC}(1)\\ &\quad= \frac{M_a(1)M_b(1)M_c(1)M_{ab}(1)M_{ac}(1)M_{bc}(1)M_{abc}(1)}{M_a(0)M_b(0)M_c(0)M_{ab}(0)M_{ac}(0)M_{bc}(0)M_{abc}(0)}\ .\end{aligned}$$

Multiallelic Markers

The one-locus model written for a diallelic marker can be extended to involve a multiallelic marker. Consider one locus A, with alleles A_i, for $i = 1,2, \ldots m$. By conventionally setting to 1 the effect M_{A1}, the $m(m+1)/2$ possible frequencies of individuals can be written in terms of $m(m+1)/2$ quantities M, M_{Ai}, for $i = 2,3, \ldots m$, and M_{AiAj}, where $i < j$. A global test for H-W equilibrium can be performed by restricting all M_{AiAj} parameters to 1 (χ^2 with $m[m-1]/2$ df). The test for association of disease with a particular allele A_i will be performed by setting the restrictions $M_{Ai}(1) = M_{Ai}(0)$ and $M_{AiAj}(1) = M_{AiAj}(0)$, for $i \neq j$ (χ^2 with m df).

Several Populations

An obvious extension of this model is also to con-

Table 1

Contingency Tables of Fels Institute Data on Fingerprint Patterns and Ss Genotype, Part of MNSs blood group

	Population 1 (Whorl)	Population 0 (Loop)
Genotype SS.....	16	13
Genotype Ss.....	38	79
Genotype ss	59	78

NOTE.—Data are drawn from Khamis and Hinkelmann (1984).

Table 2

Genotypic Frequencies in Patients with Coronary Artery Disease and in Controls, for *MspI* RFLP of Apolipoprotein B Gene

	Patients (%)	Controls (%)
Genotype M1M1	160 (82)	109 (89)
Genotype M1M2	29 (15)	12 (10)
Genotype M2M2	6 (3)	1 (1)

NOTE.—Data are drawn from Genest et al. (1990). For patients, $n = 195$; for controls, $n = 122$.

sider more than two populations. Rather than proceed to pairwise comparisons, which may rapidly inflate the number of tests, the model generalized to k populations offers a global test of comparison.

Estimating Effects of the Model

All the models presented are based on the maximization of a likelihood function. Any maximization procedure can be employed. However, in the particular case of the one-locus model, it is possible to use a standard procedure of log-linear analysis. Data are entered in a 2 × 2 × 2 contingency table of disease status by maternal allele by paternal allele. Since it has been previously assumed in the model that there are no specific maternal or paternal effects, the observed count of heterozygotes in each disease group is divided into two equal counts corresponding to Aa and aA individuals. The symmetry constraint on maternal and paternal effects must be taken into account in the assessment of df, since there are six independent parameters to be estimated, instead of eight as in a usual 2 × 2 × 2 contingency table. For the two-locus model, it is not possible to employ a standard log-linear procedure, since double heterozygotes cannot be apportioned to AB/ab and Ab/aB individuals as in the case of one marker.

Applications

In all applications presented here, the likelihood function was maximized using the program GEMINI (Lalouel 1981).

A first application of the one-locus model was made for the example given by Khamis and Hinkelmann (1984), concerning the association between the fingerprint pattern coded in two categories (loop and whorl) and the Ss genotype, part of the MNSs blood group. Data are shown in table 1. The log likelihood of the saturated model, including $M(1)$, $M_a(1)$, $M_{Aa}(1)$ and $M(0)$, $M_a(0)$, $M_{Aa}(0)$ was -265.76, whereas that of the submodel with the restriction $M_{Aa}(1) = M_{Aa}(0)$ was -268.80. Therefore, the test for allele interaction is $\chi^2 = 6.08$ with 1 df ($P < .05$). In fact, $M_{Aa}(1)$ and $M_{Aa}(0)$ appear significantly different because population 0 (loop) is in H-W equilibrium, whereas population 1 (whorl) is in disequilibrium ($M_{Aa}[1] = 0.62 \pm 0.13$). The test with the restriction $M_{Aa}(1) = 1$ gives $\chi^2 = 5.00$ (1 df, $P < .05$), whereas the usual test for H-W equilibrium comparing observed to expected frequencies (Emigh 1980) gives $\chi^2 = 5.15$.

A second application of the one-locus model is given for data drawn from Genest et al. (1990). The data are genotypic frequencies in 195 patients with coronary

Table 3

Models Tested in Data from Genest et al. (1990)

Effects in Model	lnL	Alternative Model	df	χ^2
0. Saturated model.....	−152.72			
1. $M_{Aa}(1) = M_{Aa}(0)$.....	−152.77	0	1	.10
2. $M_a(1) = M_a(0)$, $M_{Aa}(1) = M_{Aa}(0)$.....	−154.72	1	1	3.90*

NOTE.—Estimates of effects in the most parsimonious model (model 1) are $M_a(1) = 0.189 \pm 0.038$, $M_a(0) = 0.108 \pm 0.033$, and $M_{Aa}(1) = M_{Aa}(0) = 0.488 \pm 0.121$.

* $P < .05$.

Table 4

Genotype Frequencies for the MNSs Blood Group in Three Populations from Austria, Nepal, and Tibet

	SS	Ss	ss
Austria (n = 509):			
MM.....	26	81	40
MN.....	19	111	124
NN.....	3	26	79
Nepal (n = 153):			
MM.....	9	30	33
MN.....	10	29	29
NN.....	3	4	6
Tibet (n = 126):			
MM.....	3	15	33
MN.....	2	19	35
NN.....	2	6	11

NOTE.—Data are from Mourant et al. (1976, pp. 300–311).

artery disease and 122 controls for the *Msp*I RFLP of the apolipoprotein B gene (table 2). The succession of models tested is shown in table 3. The comparison of model 1 to model 0 indicates that there is no allele interaction (χ^2 = 0.10 with 1 df; not significant). The comparison of model 2 to model 1 indicates that the frequency of alleles is significantly different between the two groups (χ^2 = 3.90 with 1 df; $P < .05$). In conclusion, the two populations differ by a higher frequency of the M2 allele among patients, and, in the absence of allele interaction, the estimated OR for disease of M1M2 individuals is the square root of the estimated OR of M2M2 individuals, with reference to M1M1 individuals. From the parameters of the model (table 3), these ORs are estimated as 1.75 and 3.06, respectively.

An application of the two-locus model is given using data of the MNSs blood system in three different populations from Austria, Nepal, and Tibet (Mourant et al. 1976, pp. 300–311). Data are presented in table 4, and models for comparing the populations are shown in table 5. Comparison of model 0, the saturated model, to model 1, which estimates all 12 parameters, indicates that the three populations are in H-W equilibrium (χ^2 = 7.20 with 15 df; $P > .90$). Comparison of models 2 and 1 indicates that the three populations differ in the degree of linkage disequilibrium between loci M and S (χ^2 = 24.42 with 2 df; $P < .001$). Actually, the two loci appear to be in equilibrium in the populations from Nepal and Tibet (χ^2 = 0.70 with 2 df), whereas they are in high linkage disequilibrium in the population from Austria (χ^2 = 57.06 with 1 df; $P < .001$), with a preferential association between alleles M and S. Since there is no association between the two loci either in Nepal or in Tibet, the marginal frequencies of alleles are compared between the two populations. The frequency of allele M is not different (χ^2 = 2.68 with 1 df), whereas allele S is significantly more frequent in Nepal than in Tibet (χ^2 = 12.54 with 1 df; $P < .001$). Estimates of haplotypic and allelic frequencies in each population were derived from parameters of the most parsimonious model and are shown in table 6. Because of the difference in linkage disequilibrium, the comparison of haplotypic frequencies rather than allelic frequencies is preferable when contrasting Austria with both other populations. On the other hand, comparing haplo-

Table 5

Models Tested for the Comparison of Populations from Austria (0), Nepal (1), and Tibet (2)

Effects in Model	ln L	Alternative Model	df	χ^2
0. Saturated model	−1,498.82			
1. M(2), M_a(2), M_b(2), M_{ab}(2), M(1), M_a(1), M_b(1), M_{ab}(1), M(0), M_a(0), M_b(0), M_{ab}(0)	−1,502.42	0	15	7.20 (NS)
2. M_{ab}(2) = M_{ab}(1) = M_{ab}(0)	−1,514.63	1	2	24.42***
3. M_{ab}(0) = 1	−1,530.95	1	1	57.06***
4. M_{ab}(2) = M_{ab}(1) = 1	−1,502.77	1	2	.70 (NS)
5. M_b(2) = M_b(1)	−1,509.04	4	1	12.54***
6. M_a(2) = M_a(1)	−1,504.11	4	1	2.68 (NS)

NOTE.—NS = not significant.
*** $P < .001$.

Table 6

Estimate of Haplotypic Frequencies for the MNSs Blood Group in Populations from Austria, Nepal, and Tibet

	Austria	Nepal	Tibet
Parameters of model 6:			
M	28.897	8.225	2.544
M_a	.295	.508	(.508)
M_b	1.259	1.860	3.667
M_{ab}	4.430	(1)	(1)
Haplotypic frequencies:			
f_{MS}	.238	.232	.142
f_{Ms}	.300	.431	.521
f_{NS}	.070	.118	.072
f_{Ns}	.392	.219	.265
Allelic frequencies:			
f_M	.538	.663	.663
f_N	.462	.337	.337
f_S	.308	.350	.214
f_s	.692	.650	.786

NOTE.—Values in parentheses are fixed in the model.

typic frequencies between Nepal and Tibet is not more informative than comparing allelic frequencies at both loci.

Discussion

For a few years, an increasing interest in the problem of association between disease and genetic markers has been developing. However, the analysis currently employed in association studies appears incomplete or even questionable when ambiguous data on multiple heterozygotes are excluded. The log-linear-model approach allows further insight into this problem and provides a valid method for analyzing the specific structure of these data.

Whereas the disease-marker association is usually characterized in reports of association studies by a higher frequency of a specific allele or a specific genotype among affected individuals, it is shown in this paper that taking into account the allele interaction may provide additional information about dominance of the disease-susceptibility locus. Dominance is defined here for a qualitative phenotype using the epidemiological notion of OR. It is also shown that dominance of the disease locus implies that the affected population is in H-W disequilibrium. The test of dominance is equivalent to the test of the coefficient ρ'_I of "association between genotype and disease due to gene interaction" introduced by Khamis and Hinkelmann (1984) and Norwood and Hinkelmann (1978), although those authors did not interpret it in terms of dominance.

The detection of an association between disease and a marker locus will depend both on the strength of linkage disequilibrium existing between the two loci and on the frequency of allele a compared to that of allele s (a and s being conventionally the alleles preferentially associated together and leading to an increasing risk). If the two loci are in tight disequilibrium and a is much more frequent than s in the population, then the marker A will have a high sensitivity but a low specificity, since there will be many haplotypes aS carrying the "allele of risk" at the marker locus although not carrying the disease allele itself. On the other hand, if s is much more frequent than a, the marker will have a high specificity but a low sensitivity, since there will be many haplotypes As. Both situations will be reflected by a lower OR at the marker locus than at the unobservable disease locus, and large sample sizes will be required to detect an association. In particular, this will be observed if the allele s determining the disease is associated with the more frequent allele of the marker locus, that is, if there is a negative linkage disequilibrium between the two loci. Thompson et al. (1988) showed that very large sample sizes would be required to detect an association between two loci in negative linkage disequilibrium.

In the case of several linked loci, the problem of genotype-disease association becomes rapidly complex unless it is assumed that both affected and unaffected populations are in H-W equilibrium at all marker loci. The condition of H-W equilibrium, which is generally fulfilled in the unaffected population, may appear very restrictive for the affected population, since in case of dominance of the disease locus, the affected population will be necessarily in H-W disequilibrium. However, a departure from H-W equilibrium is already a significant item of information by itself, since it would indicate that there is an association between the disease locus and one or several of the marker loci considered, and that the degree of dominance of the disease locus is not zero. Removing the assumption of H-W equilibrium in both populations would imply a more complex model including high-order interactions, as in the model proposed by Weir and Wilson (1986) for one population. This would be theoretically possible, but the interpretation of such effects would be difficult. In particular, it

would be impossible to determine whether populations differed in the inter- or the intragametic component of linkage disequilibrium.

In the case in which both populations are in H-W equilibrium, it is possible to go ahead in the analysis with this reduced model, and the log-linear approach offers a stepwise method of analysis which allows an identification of the parameters associated with the difference between populations. For example, if the two populations exhibit the same degree of linkage disequilibrium between two markers, analyzing haplotypic frequencies will not be more informative than analyzing the two markers separately. In other words, this will indicate that the OR for disease associated with the presence of a specific allele at one locus will be the same whatever the allele at the other locus.

This model can be extended to any number of loci, any number of alleles, or any number of populations, the only major practical limitation being the sample size necessary to estimate an increasing number of effects. When testing for the presence of associations within or between several loci, there is a need for conservatively adjusting the significance levels in order to account for the multiple comparisons performed. The Bonferroni procedure allows one to calculate the level of significance to be adopted for each individual test to achieve a given overall level of significance (Weir 1990).

Another extension of this model which seems particularly important would be to consider simultaneously the effects of environmental factors and their interaction with genotype, since a disease process mostly results from a complex interaction between genes and environment.

Appendix A

Demonstration of Recessivity at the Marker Locus A, When Recessivity Is Assumed at the Disease-Susceptibility Locus S and the Two Loci Are in Complete Disequilibrium

Let s denote the allele determining the disease at the locus S. The two loci are assumed to be in complete disequilibrium, allele s always being associated with allele a of the marker. This means that, conditional on allele s, the probability of a haplotype carrying allele A is 0 and the probability for a haplotype carrying allele a is 1. Recessivity of the disease locus implies, in terms of OR, that OR(Ss/SS) = 1, which is equivalent to $P_{Ss}(1)/P_{SS}(1) = P_{Ss}(0)/P_{SS}(0)$. The same relation now has to be demonstrated for locus A.

$$P_{Aa}(1)/P_{AA}(1) = \frac{P(Aa/SS)\,P_{SS}(1) + P(Aa/Ss)\,P_{Ss}(1) + P(Aa/ss)\,P_{ss}(1)}{P(AA/SS)\,P_{SS}(1) + P(AA/Ss)\,P_{Ss}(1) + P(AA/ss)\,P_{ss}(1)} .$$

If we assume that the population undergoes random mating, genotypic frequencies are the products of corresponding gametic frequencies, and the expression above becomes

$$P_{Aa}(1)/P_{AA}(1) = \frac{2\,P(A/S)\,P(a/S)\,P_{SS}(1) + P(A/S)\,P_{Ss}(1)}{P(A/S)^2\,P_{SS}(1)} ,$$

since $P(A/s) = 0$ and $P(a/s) = 1$. Canceling $P(A/S)$,

$$P_{Aa}(1)/P_{AA}(1) = \left(\frac{2\,P(a/S)\,P_{SS}(1) + P_{Ss}(1)}{P(A/S)\,P_{SS}(1)}\right) = \left(\frac{2\,P(a/S)\,P_{SS}(0) + P_{Ss}(0)}{P(A/S)\,P_{SS}(0)}\right) ,$$

since $P_{Ss}(1)/P_{SS}(1) = P_{Ss}(0)/P_{SS}(0)$.

$$\begin{aligned} P_{Aa}(1)/P_{AA}(1) &= \left(\frac{2\,Pa/S)\,[P_S(0)]^2 + 2\,P_S(0)P_s(0)}{P(A/S)\,[P_S(0)]^2}\right) \\ &= \left(\frac{2\,P(a/S)\,P_S(0) + 2\,P_s(0)}{P(A/S)\,P_S(0)}\right) \\ &= \left(\frac{2\,[P(a/S)\,P_S(0) + P(a/s)\,P_s(0)]}{P(A/S)\,P_S(0) + P(A/s)\,P_s(0)}\right) \\ &= 2P_a(0)/P_A(0) \\ &= [2\,P_A(0)\,P_a(0)]/[P_A(0)]^2 \\ &= P_{Aa}(0)/P_{AA}(0) . \end{aligned}$$

Appendix B

Two-Locus Model: Expected Values of the Nine Observable Frequencies of Individuals, Using the Multiplicative Log-linear Approach Proposed by Weir and Wilson (1986)

Denote by $P(AABB)$ the frequency of individuals having genotype AA at locus A and genotype BB at locus B. Assuming that there is no intergametic disequilibrium (the population is in H-W equilibrium at each locus), the nine expected frequencies of individuals are written as $P(AABB) = M$, $P(AABb) = 2MM_b$, $P(AAbb) = MM_b^2$, $P(AaBB) = 2MM_a$, $P(AaBb) = 2MM_aM_b(M_{ab} + 1)$, $P(Aabb) = 2MM_aM_b^2M_{ab}$,

$P(\text{aaBB}) = MM_a^2$, $P(\text{aaBb}) = 2MM_a^2M_bM_{ab}$, and $P(\text{aabb}) = MM_a^2M_b^2M_{ab}^2$.

References

Emigh TH (1980) A comparison of tests for Hardy-Weinberg equilibrium. Biometrics 36:627–642

Genest JJ, Ordovas JM, McNamara JR, Robbins AM, Meade T, Cohn SD, Salem DN, et al (1990) DNA polymorphisms of the apolipoprotein B gene in patients with premature coronary artery disease. Atherosclerosis 82:7–17

Hill WG (1975) Tests for association of gene frequencies at several loci in random mating diploid populations. Biometrics 31:881–888

Jacquard A (1974) The genetic structures of population. Springer, New York

Khamis HJ, Hinkelmann K (1984) Log-linear-model analysis of the association between disease and genotype. Biometrics 40:177–178

Lalouel JM (1981) GEMINI: a computer program for optimization of general nonlinear functions. Tech rep no. 14, Department of Biophysics and Computing, University of Utah, Salt Lake City

Mourant AE, Kopec AC, Domaniewska-Sobczak K (eds) (1976) The distribution of the human blood groups and other polymorphisms. Oxford University Press, London

Norwood PK, Hinkelmann K (1978) Measures of association between disease and genotype. Biometrics 34:593–602

Thompson EA, Deeb S, Walker D, Motulsky AG (1988) The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes. Am J Hum Genet 42:113–124

Weir BS (1979) Inferences about linkage disequilibrium. Biometrics 35:235–254

——— (1990) Genetic data analysis. Sinauer, Sunderland, MA

Weir BS, Wilson SR (1986) Log-linear models for linked loci. Biometrics 42:665–670