Genetic Structure and the Search for Genotype-Phenotype Relationships: An Example from Disequilibrium in the *Apo B* Gene Region

Kim E. Zerba,* Anna M. Kessling,† Jean Davignon‡ and Charles F. Sing*

**Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109, †Department of Biochemistry and Molecular Genetics, St. Mary's Hospital Medical School, London W2 1PG, United Kingdom, and ‡Hyperlipidemia and Atherosclerosis Research Group, Clinical Research Institute of Montreal, Montreal, Quebec H2W 1R7, Canada*

Manuscript received January 11, 1991
Accepted for publication June 8, 1991

ABSTRACT

We analyzed allelic associations (disequilibria) for four restriction fragment length polymorphisms (RFLPs) in the region of the 43-kb *Apo B* gene in a sample of 233 unrelated individuals from Montreal, Canada, sampled for health. This total sample (T) included 160 individuals of known French Canadian (FC) ancestry. We present a rigorous application of current methodology to these samples, including estimation of type II error probabilities and correlations between markers for estimates of disequilibria. We then consider the utility of these estimates of allelic disequilibria for the interpretation of genotype-phenotype relations. Significant deviations from Hardy-Weinberg equilibrium were not predicted by proximity to other markers in disequilibrium. We found significant quadri-allelic disequilibrium for two marker pairs despite absence of significant deviations from Hardy-Weinberg equilibrium for either marker or tri-allelic disequilibrium, respectively. Altogether these results underscore the complexity of the genotypic structure of the data. A combination of nonevolutionary factors, including sampling for health, small sample size and data exclusion due to methodological constraints of not successfully typing all members of the sample for every RFLP, is a likely explanation for this complexity. These types of factors are common to many RFLP studies. Patterns of composite di-allelic disequilibrium indicated that some RFLP allele pairs may have a longer shared evolutionary history than others and that disequilibrium is not predicted by distance between RFLPs. Type II error probabilities were generally much higher than those for type I errors. Correlations between marker pairs for disequilibria were generally not high. We show from a review of 14 published studies of association between the *Xba*I RFLP and variation in a total of 15 lipid traits that deviations from Hardy-Weinberg equilibrium can cause substantial differences in the estimation of variability associated with phenotypic differences among marker genotypes relative to Hardy-Weinberg conditions.

MARKERS such as restriction fragment length polymorphisms (RFLPs) are in widespread use to identify DNA sequences containing mutations responsible for quantitative phenotypic variation in a population. When multiple markers of a gene region are available, they are commonly used individually in separate analyses of variance (ANOVAs) to detect associations of marker genotypes with phenotypic variation. Inferences are then made about an unknown functional mutation with a phenotypic effect, which is nonrandomly associated with marker alleles, considering all single-marker analyses simultaneously. This approach implicitly assumes the marker allele frequencies are independent. Inferences from this analytical strategy ignore the evolutionary history shared by the markers.

Evolutionary history events, including admixture, finite population size, selection, migration and mutation, determine the genetic structure defined by frequency associations among marker alleles on the same (gametic allele associations) and different homologous chromosomes (nongametic allele associations). Sampling design factors, including nonrandom sampling, small sample size and data exclusion due to methodological constraints, represent an additional, very important and largely unrecognized class of determinants of such allele frequency associations. Whichever the cause(s), nonrandom marker allele frequency associations (allelic disequilibria) result in two major problems for ANOVA studies of marker-phenotype associations.

First, the presence of gametic allele associations means that inferences from analyses of markers are interdependent (TEMPLETON, BOERWINKLE and SING 1987; TEMPLETON *et al.* 1988). In addition, since evolutionary history events are infrequent and unpredictable in time and in chromosomes they affect, a predictable relationship is not expected between the degree of gametic allele association and physical distance. This unpredictability applies to associations among marker alleles and between alleles of markers and the mutation(s) responsible for a phenotypic effect. Consequently, specificity of information from individual markers about the location of a mutation

with a phenotypic effect is not possible. Where there is gametic allelic disequilibrium, inferences from multiple single-marker analyses are limited to showing whether or not the region being marked is involved with the phenotype (KESSLING *et al.* 1991).

Second, nongametic allelic disequilibrium will affect inferences about association of markers with phenotypic variation. Phenotypic variation associated with marker allelic variation is dependent on three factors related to frequency and scale: (1) the frequency distribution of marker genotypes; (2) the association between a mutation with a phenotypic effect and marker alleles; and (3) the impact of the mutation on the phenotype (scale). Inferences about associations of single markers with phenotypic variation are often made assuming that genotypes are in Hardy-Weinberg proportions in the population. If it is assumed that the "truth" is represented by genotypes in Hardy-Weinberg proportions, then deviation of marker genotypes from Hardy-Weinberg expected values results in differences in the estimation of phenotypic variation associated with a single marker relative to these conditions.

The first step in dealing with these issues is to understand the genetic structure of marker variation. The strength of inferences about genetic structure becomes especially important because this information is critical in the proper design of statistical analyses to locate mutations with phenotypic effects. Power to detect disequilibrium given sample sizes used in most studies may be low (WARD and SING 1970; THOMPSON *et al.* 1988). Thus, the complement to power, *i.e.*, the type II error probability of not detecting disequilibrium when it is present, may be high. Type II errors may also play an important role in explaining discordant inferences about genetic structure of a region, estimated by different studies using the same set of markers. In addition, there is the concern regarding correlated type I and type II errors of inference resulting from estimating many allelic associations from the same set of data.

This report focuses on an analysis of the genetic structure of marker alleles in the *Apolipoprotein B* (*Apo B*) gene region in humans. The DNA sequence encoding the *Apo B* gene is on the short arm of chromosome 2, 43 kb in length and includes 29 exons (BLACKHART *et al.* 1986). *Apo B* plays a central role in assembly of triglyceride-rich lipoproteins (LEIGHTON *et al.* 1990), is a major component of low density lipoproteins (LDL) and the ligand for the LDL receptor (BROWN and GOLDSTEIN 1986). LDL is a major transporter of cholesterol in plasma. It is well established that elevated LDL cholesterol levels convey increased risk of coronary artery disease (CAD). Thus apolipoprotein B is a candidate for a role in the pathogenesis of CAD (SNIDERMAN and SILBERBERG 1990).

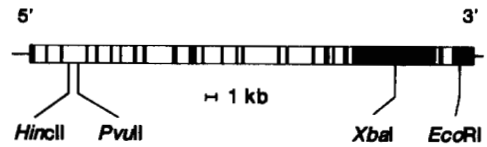

FIGURE 1.—The *Apo B* gene region and location of RFLPs. Filled bars represent exons; open bars represent introns.

We evaluate the genetic structure defined by allelic associations at each and between pairs of four RFLPs in the *Apo B* region. We present a rigorous application of current methodology to this data set including estimation of type II error probabilities and correlations between markers for estimates of allelic disequilibria. We then consider the utility of these estimates of genetic structure for the interpretation of marker genotype-phenotype relationships. We show from a review of 14 published studies of association between the *Apo B XbaI* RFLP and variation in a total of 15 lipid traits that deviations from Hardy-Weinberg equilibrium can cause substantial differences in the estimation of variability associated with phenotypic scale differences among marker genotypes relative to Hardy-Weinberg conditions.

MATERIALS AND METHODS

Sample: Unrelated white-collar workers, aged 20 to 59 years from the Hydro-Quebec power company of Montreal, Canada, were sampled according to health status by the Hyperlipidemia and Atherosclerosis Research Group at the Clinical Research Institute of Montreal to represent a population free of clinical manifestations of disease, including cardiovascular disease, hypertension, chronic disorders requiring medication, previously diagnosed thyroid dysfunction, diabetes, gout, renal dysfunction or hyperlipidemia (S. LUSSIER-CACAN, unpublished data). There were 233 individuals with typings for all four *Apo B* RFLPs; 146 male and 87 female. This was the total (T) sample used for this study. Of these, 160 individuals were of known French Canadian (FC) ancestry (KESSLING *et al.* 1991). These two samples were part of an overall sample of 347 individuals on which RFLP typings were attempted. The difference between this larger sample and the T sample used in this study is the result of data exclusion due to methodological constraints. Individuals were not included if they could not be typed for an RFLP because independent assessments by three investigators did not agree on the typing from reading the autoradiograms. In common with other RFLP studies, typings were repeated where DNA was sufficient but rejected where DNA was insufficient for a repeat of the assay.

Molecular genetic analyses: Restriction endonuclease analyses defining the four markers in this region for this sample have been described (KESSLING *et al.* 1991). The *Apo B* gene region and location of RFLPs is depicted in Figure 1. Two of the four polymorphisms are near the 5′ end with the others near the 3′ portion of the *Apo B* region. The HincII and *Pvu*II polymorphisms are in intron 4, 171 bp 3′ to exon 4 and within an Alu sequence 523 bp 5′ to exon 5 (HUANG, RIPPS and BRESLOW 1990), respectively. Both are restriction site polymorphisms resulting from single base substitutions. The *Xba*I polymorphism is in exon 26 and changes the third base of codon 2488, but does not alter an amino acid (BERG *et al.* 1986). The *Eco*RI polymorphism is

in exon 29. The presence of the *Eco*RI variable restriction site alters glutamic acid to lysine (SHOULDERS *et al.* 1985) and is associated with the Ag antisera type t/z (MA *et al.* 1987).

Statistical analyses: In the following expressions we employ the notation of WEIR and COCKERHAM (1989). The most frequent alleles of a pair of RFLPs are designated A and B, respectively. Genotype frequency estimates were the observed 12 single-marker and 54 two-marker genotype frequencies. For each pair of marker loci, nine two-marker genotypes were recognized; double heterozygotes were not distinguished. Allele frequencies were estimated by gene counting. Variances of allele frequencies were determined according to WEIR (1990). Using an approach and methodology outlined by WEIR and COCKERHAM (1989), for each RFLP, we estimated deviations of the three genotype frequencies from those expected at Hardy-Weinberg equilibrium (D_A). For pairs of RFLPs we estimated composite di-allelic (Δ_{AB}), two tri-allelic (D_{AAB}, D_{ABB}) and quadri-allelic (Δ_{AABB}) disequilibria. We also estimated variances for each of these disequilibria.

Bounds of all these disequilibrium estimates are affected by allele frequencies (HEDRICK 1987; LEWONTIN 1988). Disequilibrium estimates and their variances other than D_A are affected by other disequilibria estimates (WEIR and COCKERHAM 1989). We therefore determined the limits of each of the observed disequilibria and the percentage of each divided by its limit. Bounds for D_A were determined according to WEIR and BROOKS (1986). Since Δ_{AB} is a composite measure of two disequilibria having the same limits, we determined the bounds for Δ_{AB} as twice values for D_{AB} defined by WEIR and COCKERHAM (1989). Bounds for tri-allelic disequilibria were determined according to WEIR and COCKERHAM (1989). Bounds for quadri-allelic disequilibria have not been derived and hence were not estimated.

For each disequilibrium estimate, we tested the single degree of freedom null hypothesis that the disequilibrium parameter $\delta = 0$ by

$$\chi^2 = \delta^2/\text{var}(\delta)_0 \qquad (1)$$

where δ represents the disequilibrium estimated and $\text{var}(\delta)_0$ represents the variance of δ estimated under the null hypothesis with the δ set to zero and all other parameters at the respective point estimates. χ^2 in (1) is approximately distributed as chi-square with one degree of freedom. We use the conventional statistical significance level $\alpha = 0.05$ for single tests throughout.

We used computer simulation to determine the type II error probabilities of not detecting disequilibrium of the size determined by the observed allele and genotype frequencies and sample sizes. We constructed T and FC populations with the relative genotype and allele frequencies and hence the disequilibria observed in our samples. These observed disequilibria served as the alternatives to the null hypothesis of disequilibrium equal to zero for the type II error simulations. Then, for each disequilibrium, 10,000 replicate random samples were drawn from each of these populations with replacement. Samples were of size 233 or 160 depending on whether we were considering the T sample or FC subsample. For each replicate sample we calculated the disequilibrium coefficient of interest and tested the null hypothesis using the strategy described above. Type II error probability was determined as the proportion of the simulations in which the chi-square was less than or equal to the critical value 3.84 for a single degree of freedom test at $\alpha = 0.05$. Altogether, 28 type II error probability simulations were performed for the estimated disequilibria in our samples.

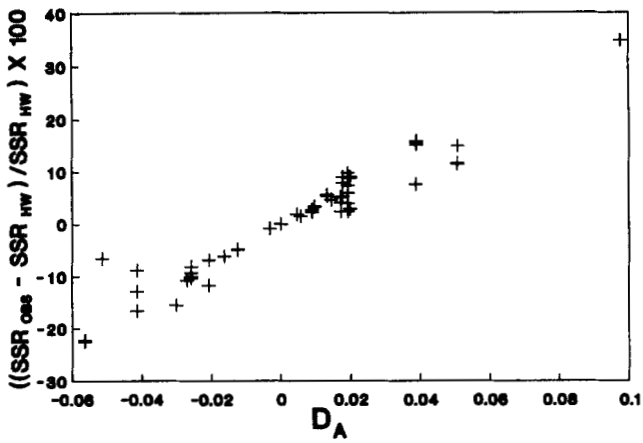

FIGURE 2.—Impact of deviations from Hardy-Weinberg equilibrium on estimation of sum of squares variability associated with marker genotypes. Relationship in figure based on review of 14 published association studies of a total of 15 lipid traits with the *Apo B Xba*I RFLP as marker (LAW *et al.* 1986; COCOZZA *et al.* 1987; TALMUD *et al.* 1987; AALTO-SETÄLÄ *et al.* 1988, 1989; DUNNING *et al.* 1988; LEREN *et al.* 1988; MONSALVE *et al.* 1988; DEMANT *et al.* 1988; HOULSTON *et al.* 1988; DARNFORS *et al.* 1989; MYANT *et al.* 1989; WIKLUND *et al.* 1989; PAULWEBER *et al.* 1990).

Since all of the tests for disequilibria were performed on the same data set, we were concerned with the potential for correlated type I (SOKAL and ROHLF 1981) and correlated type II errors of inference. We evaluated such potential correlated errors of inference using bootstrap techniques (EFRON 1982; WEIR and BROOKS 1986) to estimate the correlation between a pair of markers or two pairs of markers for the respective allelic disequilibria. A substantial positive correlation would support the potential for correlated errors of inference. We focused on the correlations between RFLPs for D_A in the T sample and FC subsample and between different pairs of RFLPs for composite measures of di-allelic disequilibrium in the FC subsample.

We estimated the effects of adjusting critical α levels for multiple tests on type II error rates. We estimated the overall probability of at least one type II error for families of tests of D_A in the T and FC subsamples and Δ_{AB} in the FC subsample. For each family of tests we simultaneously estimated all respective disequilibria for each of the 10,000 bootstrap replicate samples. Type II error probability was determined as the proportion of simulations in which the chi-square was less than or equal to the critical value for any of the disequilibria. For each family of tests the simulations were run using an individual test $\alpha = 0.05$ and an experimentwise $\alpha = 0.05$.

We reviewed the results of 14 published studies of association between variability in a total of 15 lipid traits and the *Apo B Xba*I RFLP (references under Figure 2). For each study we estimated the D_A. We then estimated the sum of squares variability associated with the *Xba*I RFLP using

$$\text{SSR} = \Sigma\Sigma f_{ij} (\bar{Y}_{ij} - \bar{Y}_{..})^2, \qquad (2)$$

where f_{ij} is the frequency of the $_{ij}$th genotype, $\bar{Y}_{ij}$ the $_{ij}$th genotype mean and $\bar{Y}_{..}$ the grand mean of the sample. We estimated SSR in two ways: (1) using the observed f_{ij} (SSR_{OBS}), and (2) using Hardy-Weinberg equilibrium expected f_{ij} (SSR_{HW}). We estimated the percent difference in estimating SSR relative to genotype proportions under Hardy-Weinberg conditions by

$$((\text{SSR}_{\text{OBS}} - \text{SSR}_{\text{HW}})/\text{SSR}_{\text{HW}}) \times 100. \qquad (3)$$

TABLE 1

Common allele frequencies and deviations from Hardy-Weinberg equilibrium in the total (T) sample (n = 233) and French Canadian (FC) subsample (n = 160)

	5′	*Hinc*II		*Pvu*II		*Xba*I		*Eco*RI	3′
Sample		T	FC	T	FC	T	FC	T	FC
p_A		0.70	0.70	0.90	0.90	0.52	0.51	0.83	0.83
$\text{Var}(p_A) \times 10^{-4}$		5.1	6.9	2.0	3.2	4.7	7.1	3.3	5.1
D_A		0.027	0.012	0.008	0.016	−0.032	−0.022	0.013	0.019
$\text{Var}(D_A) \times 10^{-4}$		2.0	2.8	0.5	1.1	2.6	3.9	1.1	1.8
$D_A/D_{A\ max} \times 100$		12.9	5.7	9.0	17.9	13.8	9.1	9.5	13.4
χ^2		3.85	0.52	1.88	5.10	3.73	1.24	2.08	2.88
Pr (χ^2)		0.0497	0.4721	0.1700	0.0240	0.0534	0.2663	0.1489	0.0896
Pr (type II error)		0.5024	0.8854	0.7065	0.4269	0.4901	0.7932	0.6892	0.6071

RESULTS

Common allele frequencies and variances for each RFLP for the T sample and FC subsample are in Table 1. RFLPs are listed in order 5′ to 3′. Relative frequencies of the common allele ranged from 0.5 for *Xba*I to 0.9 for *Pvu*II. They were not significantly different between the two samples. Variances of allele frequencies were larger in the FC subsample, reflecting its smaller sample size.

Estimates of D_A, corresponding variances, percent of maximum possible D_A, χ^2 tests and type II error probabilities are also presented in Table 1. Estimates of D_A were significant for the *Hinc*II and *Xba*I RFLPs in the T sample, but not in the FC subsample. In contrast, the estimate of D_A was not significant for *Pvu*II RFLP in the T sample, but was significant in the FC subsample. As expected, variances of all D_A estimates were larger in the smaller FC subsample. None of the estimates of D_A were at or near bounds set by allele frequencies. Type II error probability over all tests ranged from 0.43 to 0.89. For the three tests where significant D_A was detected in the actual samples, the type II error probability was 8.5 to 10.0 times higher (0.43–0.50) than the type I error probability given by the nominal level of significance α = 0.05 for a single test for the same three estimates.

Estimates of two-marker disequilibria, variances, percent of maximum possible disequilibrium, χ^2 tests and type II error probabilities are in Table 2. The inferences from the T sample and FC subsample were the same for 22 of the 24 tests. However, because of differences in significant D_A between the two samples, we present estimates only of two-marker disequilibria in the FC subsample. Composite measures of di-allelic disequilibria were significant for four of six pairs of markers; Δ_{AB} for the *Hinc*II RFLP paired with the two RFLPs at the other end of the region (*Xba*I and *Eco*RI) was not significant in either case. We note that Δ_{AB} for the *Pvu*II RFLP adjacent to the *Hinc*II RFLP was significant for all pairings with the other three markers. None of these composite estimates was at the bounds set by allele frequencies. None of the tri-allelic measures was significant and 3/12 were at the maximum possible values. *Hinc*II-*Xba*I and *Xba*I-*Eco*RI were the only two of the six marker pairs which gave evidence for significant quadri-allelic disequilibrium. Type II error probabilities over all tests ranged from 0.0 to 0.9989. We note that for the three non-significant estimates of tri-allelic disequilibria at bounds (100% maximum) set by allele frequencies, type II error probability was very high. In addition, two of the three bounded tri-allelic disequilibria are from the *Pvu*II-*Eco*RI RFLP pair. The estimate of quadri-allelic disequilibrium depends, in part, on these tri-allelic measures. Type II error probability for *Pvu*II-*Eco*RI quadri-allelic disequilibrium was the highest of all the tests (0.9989). Furthermore, for the six two-marker tests in which disequilibrium was detected, type II error probabilities ranged from 4.5 to 10.4 times higher (0.22–0.52; three tests) to over 100 times lower (0.0005–0; two tests) than the fixed α = 0.05 probability of type I error for each of the same six tests.

Bootstrap estimates of correlations between the RFLPs for estimates of D_A in the T sample and FC subsample are in Table 3. Estimated variances are on the diagonal of Table 3, with T above FC. These bootstrap variances were nearly the same as those estimated in Table 1 from the theoretical equations. Correlations for the T sample are in the upper diagonal and for the FC subsample in the lower diagonal of Table 3. All correlations were positive and low ($r \leq 0.27$). These results indicate the probability of correlated errors of inferences about D_A is low for this data set.

Bootstrap estimates of correlations between the estimates of two-marker composite measures of di-allelic disequilibria for the FC subsample are in Table 4. The bootstrap estimated variances were again nearly the same as those estimated from the theoretical equations in Table 2. The correlation structure among the Δ_{AB} was more complex than among the D_A with not

TABLE 2

Two-marker disequilibria in French Canadians

RFLP pair	Disequilibrium	Variance $\times 10^{-4}$	Percent maximum	χ^2	Pr (χ^2)	Pr (type II error)
*Hinc*II-*Pvu*II						
Δ_{AB}	0.05	1.9	38.1	19.54	<0.0001	0.0005
D_{AAB}	−0.006	0.2	50.2	1.80	0.1796	0.7279
D_{ABB}	0.006	0.3	50.3	1.37	0.2423	0.8449
Δ_{AABB}	0.002	<0.1		0.33	0.5641	0.9381
*Hinc*II-*Xba*I						
Δ_{AB}	−0.02	4.0	7.7	1.29	0.2570	0.7892
D_{AAB}	−0.001	0.4	2.4	0.05	0.8264	0.9376
D_{ABB}	−0.006	0.4	6.8	0.70	0.4042	0.8650
Δ_{AABB}	0.01	0.2		6.55	0.0105	0.2893
*Hinc*II-*Eco*RI						
Δ_{AB}	0.02	2.6	8.6	1.82	0.1778	0.7226
D_{AAB}	0.002	0.3	7.9	0.12	0.7345	0.9176
D_{ABB}	−0.006	0.3	20.6	1.79	0.1808	0.7416
Δ_{AABB}	0.002	0.1		0.24	0.6235	0.9384
*Pvu*II-*Xba*I						
Δ_{AB}	−0.03	1.6	34.3	6.4	0.0114	0.2238
D_{AAB}	0.006	0.2	65.7	1.9	0.1680	0.8359
D_{ABB}	−0.001	0.2	3.4	0.1	0.8118	0.9370
Δ_{AABB}	0.002	<0.1		1.1	0.2950	0.7798
*Pvu*II-*Eco*RI						
Δ_{AB}	−0.03	0.5	81.6	8.0	0.0047	0.0540
D_{AAB}	0.002	<0.1	100.0	0.5	0.4677	0.9915
D_{ABB}	<0.001	<0.1	100.0	<0.1	0.9977	0.9499
Δ_{AABB}	−0.0004	<0.1		0.2	0.6714	0.9989
*Xba*I-*Eco*RI						
Δ_{AB}	−0.07	2.5	43.6	21.8	<0.0001	0.0000
D_{AAB}	0.002	0.3	100.0	0.1	0.7625	0.9343
D_{ABB}	0.008	0.2	72.7	2.3	0.1285	0.7511
Δ_{AABB}	0.005	<0.1		4.07	0.0437	0.5196

TABLE 3

Bootstrap variances of and correlations between RFLPs for D_A in the total sample and French Canadian subsample

	*Hinc*II	*Pvu*II	*Xba*I	*Eco*RI
*Hinc*II	*2.1* *2.9*	0.12	0.09	0.02
*Pvu*II	0.15	*0.5* *1.1*	0.08	0.03
*Xba*I	0.21	0.10	*2.6* *3.8*	0.23
*Eco*RI	0.06	0.01	0.27	*1.1* *1.8*

Variances ($\times 10^{-4}$) are in italics on the diagonal with T above FC. Correlations for the T and FC samples are given in the upper and lower diagonals, respectively. Bootstrap samples were 3000 for each of the T and FC samples.

only larger, but also positive and negative correlations. The largest positive correlation was between *Hinc*II-*Xba*I and *Pvu*II-*Xba*I (0.42). The inferences about these two pairs of markers are not candidates for correlated type I or type II errors since Δ_{AB} was significant for the latter marker pair only. We conclude the probability of correlated errors of inferences about Δ_{AB} is low for this data set.

Estimates of at least one type II error in a family of tests were 0.9819, 0.9976 and 0.9248 for D_A in the T and FC subsamples and Δ_{AB} in the FC subsample, respectively. Adjusting critical α levels for multiple tests inflated these estimates of type II error rates to 0.9988, 0.9999 and 0.9942, respectively.

Assuming the "truth" is represented by genotypes in Hardy-Weinberg proportions, the relationship between percent difference in estimation of SSR associated with the *Apo B Xba*I RFLP relative to Hardy-Weinberg conditions for 14 published studies is shown in Figure 2 (references under Figure 2). The relationship is linear and predictable. Deviation of individual study points from linearity are due to differences among studies in the relative position of the heterozygote mean between the homozygotes and differences in allele frequencies. When the mean for the heterozygote falls outside of the range determined by the homozygotes the percent sum of squares difference relationship with D_A is still predictable, but more complex and nonlinear (K. E. ZERBA, unpublished results). We included results only from studies for traits where the heterozygote mean was estimated as between the homozygotes. As evident from Figure 2,

TABLE 4

Bootstrap variances of and correlations between RFLP pairs for Δ_{AB} in French Canadians

	*Hin*cII-*Pvu*II	*Hin*cII-*Xba*I	*Hin*cII-*Eco*RI	*Pvu*II-*Xba*I	*Pvu*II-*Eco*RI	*Xba*I-*Eco*RI
*Hin*cII-*Pvu*II	*1.9*	−0.40	−0.28	−0.66	−0.45	0.24
*Hin*cII-*Xba*I		*4.0*	−0.39	0.42	0.18	0.01
*Hin*cII-*Eco*RI			*2.6*	0.15	0.08	−0.23
*Pvu*II-*Xba*I				*1.6*	0.20	−0.21
*Pvu*II-*Eco*RI					*0.5*	0.05
*Xba*I-*Eco*RI						*2.7*

Variances ($\times 10^{-4}$) are in italics on the diagonal. Bootstrap sample was 3000.

there can be substantial difference in the estimation of trait variability associated with the *Apo B Xba*I RFLP relative to Hardy-Weinberg conditions, ranging from as much as 34% higher to 22% lower.

DISCUSSION

The pattern of significant D_A among the four RFLPs in each of the T and FC samples illustrates that D_A for a marker is not predicted by proximity to other markers in disequilibrium. This is a remarkable result given the high degree of composite di-allelic disequilibrium among most pairs of markers in this region. The differences between the T sample and FC subsample in significant estimates of D_A were also surprising given the RFLP allele frequencies were not different between the two samples. These differences in D_A may be partly explained by differences in genotypic structure between the samples of French Canadians and other populations of origin that comprise the T sample. Smaller sample size of the FC subsample may also explain some of the differences. Moreover, we found significant estimates of quadri-allelic disequilibrium for two marker pairs in the FC subsample, despite none of the single markers from these pairs and neither pair showing evidence for significant estimates of D_A or tri-allelic disequilibrium, respectively. We suggest that chance sampling associated with small sample size and data exclusion due to methodological constraints at the RFLP typing stage of analysis probably have very important roles in estimates of D_A and quadri-allelic disequilibrium. For these obvious reasons inferences about the genetic structure of the original sample of healthy individuals from Hydro-Quebec cannot be made from this study. Sampling design factors including samples consisting of multiple strata with respect to population of origin or race, nonrandom sampling for disease or health, small sample size and data exclusion due to methodological constraints at the RFLP typing level are features common to many RFLP studies. These results underscore the complex genotypic structure that may occur in any particular set of data and force reconsideration of the meaning of such concepts as Hardy-Weinberg equilibrium as conventionally considered for a gene region as a whole. The implications of such complex genotypic structure for construction of meaningful population genetic models offers an extraordinary challenge.

Many studies of genetic structure have considered only gametic di-allelic (linkage) disequilibrium (CHAKRAVARTI *et al.* 1984; but see CHAKRAVARTI *et al.* 1986; WEIR and HILL 1986), ignoring the potential for other types of nonrandom allelic associations (SINNOCK and SING 1972; WEIR 1979; WEIR and COCKERHAM 1989) that may influence the variance and hence significance of measures of linkage disequilibrium. Furthermore, haplotype frequencies are often estimated from nonfamily data assuming Hardy-Weinberg equilibrium, but not considering the possible presence of D_A. WEIR and BROOKS (1986) and HAVILAND *et al.* (1991) present the only other published applications of this more comprehensive approach of characterizing disequilibrium to human marker data sets.

Patterns of composite di-allelic disequilibria we observed in the FC subsample suggests that certain alleles for these RFLPs are often packaged together on individual chromosomes. The *Pvu*II RFLP was in significant disequilibrium with all three other RFLPs. However, even though *Pvu*II was in significant disequilibrium with the adjacent *Hin*cII RFLP, there was nonsignificant disequilibrium estimated when *Hin*cII was paired with the other two markers at the other end of the region. These results indicate *Hin*cII may have had a longer shared evolutionary history with *Xba*I and *Eco*RI than with *Pvu*II. However, a role for small sample size and data exclusion due to methodological constraints in determining these patterns is also a possibility.

JENNER *et al.* (1988) found no evidence for disequilibrium between *Xba*I and *Eco*RI. However, not detecting significant disequilibrium does not imply it is absent in the population (COX, BELL and XIANG 1988). There are at least four explanations for differences between our study and that of JENNER *et al.* (1988). First, the samples are from different populations (French Canadian *vs.* persons living in England) with possibly different evolutionary histories. This is supported by differences between the respective stud-

ies in common allele frequencies (*Xba*I: 0.51 *vs.* 0.69; *Eco*RI: 0.83 *vs.* 0.87). Second, the JENNER *et al.* study had very small sample sizes and most likely larger type II error probability with regard to detecting disequilibrium. Third, disequilibria can be at maximal possible values and not significant. The JENNER *et al.* study did not consider this possibility. Fourth, the JENNER *et al.* study eliminated double heterozygotes from consideration because of unknown allelic phase relationships, which could have affected not only allele frequencies, but also the estimate of disequilibrium.

Our results also indicate no direct relationship between physical distance between markers and degree of composite di-allelic disequilibrium. In theory, increasing distance between markers is accompanied by increased probability of recombination over time and subsequent decreases in gametic disequilibrium relationships with time. This relationship does not appear to hold in short DNA regions where the force of recombination on haplotype organization may be similar to or less than the force of mutation (LITT and JORDE 1986; THOMPSON *et al.* 1988; BØRRESEN, MØLLER and BERG 1988).

Most studies of genetic structure have focused on fixed α levels of 0.05 for individual tests. Other studies have recognized that many tests are being repeated on the same set of data with the potential for correlated type I errors and thus lower individual test a levels such that an experimentwise α level 0.05 is attained (HEGELE, PLAETKE and LALOUEL 1990). In contrast, WARD and SING (1970) showed huge sample sizes are needed to detect significant D_A at levels normally observed in most studies of human populations. THOMPSON *et al.* (1988) also showed, for linkage disequilibrium statistics, that large sample sizes are needed to detect disequilibrium, especially for negative linkage disequilibrium. We considered the very important complement to power, *i.e.*, the type II error probability, for each disequilibrium estimate from the samples considered. Our computer simulation results showed, for most individual tests that detected presence of significant D_A and two-marker disequilibrium (given the observed sample sizes and allele frequencies), the type II error probability can be much greater than the fixed type I error, α, probability of 0.05.

We can compare estimates of the overall probability of making at least one type I error *vs.* that of a type II error for a family of tests of disequilibrium. A standard probability argument is used to estimate an overall type I error probability α' for a family of k tests,

$$\alpha' = 1 - (1 - \alpha)^k. \quad (4)$$

For example, given the conventional $\alpha = 0.05$, the probability of making at least one type I error for the four tests of D_A in the T or FC samples is $\alpha' = 0.19$. For the six tests of Δ_{AB} in the FC subsample $\alpha' = 0.26$. A similar argument can be made regarding the probability of making at least one type II error for a family of k tests (β'). Our simulations indicated the type II error probabilities for these same three families of tests were quite high ($\beta' > 0.9$). These arguments assume the tests are independent, which of course they are not, given the bootstrap estimates of correlation structure among the tests presented in Tables 3 and 4. Furthermore, the type II error probabilities are *a posteriori* estimates, whereas type I error probability α is determined *a priori*. Our arguments do, however, provide crude estimates of the relative differences in probabilities of the two types of errors for sets of tests.

Extending this argument, to achieve an experimentwise $\alpha' = 0.05$, the individual test α is lowered. Our results showed, however, that using an experimentwise $\alpha' = 0.05$ for families of tests to lower the type I error rate simultaneously inflates already high type II error rates. Given these results and the bootstrap estimates of relatively low probabilities of both correlated type I and type II errors for this data set, we must conclude, for disequilibrium analyses of this data set and others like it, that using experimentwise type I error rates results in unacceptable experimentwise type II error rates. Using an experimentwise $\alpha' = 0.05$ for the family of tests of Δ_{AB} in Table 2 would mean that the value for *Pvu*II-*Xba*I is now not significant. It is true that this value may not represent real disequilibrium and could be significant by chance alone. It is also true, however, that any of the nonsignificant values for Δ_{AB} in Table 2 may represent real disequilibrium and could be not significant by chance alone. This point is especially important for inferences about measures of composite di-allelic disequilibrium in short DNA regions where we expect such disequilibrium (LITT and JORDE 1986). Also, given the potential for involvement of nonevolutionary sampling design factors in other types of disequilibrium for many studies, it is reasonable to consider a less conservative alternative to adjusting critical values for multiple tests. In this manner, any significant disequilibrium would require explanation. An argument could certainly be made for even increasing the α level above 0.05 for individual tests to achieve a better balance between the two types of potential errors.

Since physical distance between markers in this region does not predict the magnitude or significance of disequilibria among marker alleles, we conclude the physical distance relationships between markers and unknown functional mutations will be likewise unpredictable from the degree of associations inferred about individual markers and phenotypic variation. This unpredictability is a natural consequence of the unpredictable evolutionary history shared by alleles. Furthermore, nonindependence of RFLPs dictates

markers should be used as multisite haplotypes or genotypes in ANOVA studies to identify the chromosomes likely to carry the unknown functional mutations with significant phenotypic effects. Using multisite haplotypes or genotypes automatically incorporates the evolutionary history shared by marker and unknown functional mutation alleles. And lastly, it is clear from this study that genetic structure cannot be ignored. Our simple example of D_A from published results indicates nonrandom allele frequency associations will affect genotypic frequencies and hence inferences about genetic variance of quantitative traits if we assume in our inferences that such disequilibria do not exist. Work in progress examines the role of the more complex two-marker allelic associations on such inferences. Our argument assumes that completely random allelic associations are the "truth." If evolutionary factors are the sole explanation for nonrandom allelic associations then there is no bias in the observed genotype-phenotype relationships, If, however, nonevolutionary sampling factors are involved such as those we have described in this study, then the phenotypic variation associated with a marker is estimated with bias and will remain a problem in the search for genotype-phenotype relationships.

This work was supported in part by Health and Welfare, Canada grant 6605-2087-52, the Medical Research Council of Canada grant MA-5427, the Medical Research Council of Canada/CIBA-Ceigy Canada Ltd. grant U1-0029, grants from L'Institut de resherche en santé et en sécurité du travail (IRSST) du Québec, La Succession J. A. De Sève and Hydro-Québec, and National Institutes of Health grant HL39107. A. M. K. was supported as a postdoctoral fellow under a fellowship of the Medical Research Council of Canada. We are grateful to the administration at Hydro-Quebec whose participation was critical in completing this study. We appreciate the efforts of numerous individuals at the Clinical Research Institute of Montreal who likewise made this study possible, especially SUZANNE LUSSIER-CACAN, MARIANNE XHIGNESSE, CHRISTINE BÉTARD, ODETTE BOUFFARD, ANN CHAMBERLAND and SUZANNE OUELLETTE. We thank BRUCE WEIR, PAT MOLL, MARTHA HAVILAND, TIM REBBECK and SHARON REILLY for discussion and constructive suggestions during the preparation of this paper.

LITERATURE CITED

AALTO-SETÄLÄ, K., M. J. TIKAANEN, M.-R. TASKINEN, M. NIEMINEN, P. HOLMBERG and K. KONTULA, 1988 *Xba*I and c/g polymorphisms of the apolipoprotein B gene locus are associated with serum cholesterol and LDL-cholesterol levels in Finland. Atherosclerosis **74:** 47–54.

AALTO-SETÄLÄ, K., H. GYLLING, E. HELVE, P. KOVANEN, T. A. MIETTINEN, H. TURTOLA and K. KONTULA, 1989 Genetic polymorphism of the apolipoprotein B gene locus influences serum LDL cholesterol level in familial hypercholesterolemia. Hum. Genet. **82:** 305–307.

BERG, K., L. M. POWELL, S. C. WALLIS, R. PEASE, T. J. KNOTT and J. SCOTT, 1986 Genetic linkage between the antigenic group (Ag) variation and the apolipoprotein B gene: assignment of the Ag locus. Proc. Natl. Acad. Sci. USA **83:** 7367–7370.

BLACKHART, B. D., E. M. LUDWIG, V. R. PIEROTTI, L. CAIATI, M. A. ONASCH, S. C. WALLIS, L. POWELL, R. PEASE, T. J. KNOTT, M.-L. CHU, R. W. MAHLEY, J. SCOTT, B. J. MCCARTHY and B. LEVY-WILSON, 1986 Structure of the human apolipoprotein B gene. J. Biol. Chem. **25:** 15364–15367.

BØRRESEN, A. L., P. MØLLER and K. BERG, 1988 Linkage disequilibrium analyses and restriction mapping of four RFLPs at the proα2(I) collagen locus: lack of correlation between linkage disequilibrium and physical distance. Hum. Genet. **78:** 216–221.

BROWN, M. S., and J. L. GOLDSTEIN, 1986 A receptor-mediated pathway for cholesterol homeostasis. Science **232:** 34–47.

CHAKRAVARTI, A., K. H. BUETOW, S. E. Antonarakis, P. G. Waber, C. D. Boehm and H. H. KAZAZIAN, 1984 Nonuniform recombination within the human β-globin gene cluster. Am. J. Hum. Genet. **36:** 1239–1258.

CHAKRAVARTI, A., K. H. BUETOW, S. E. ANTONARAKIS, P. G. WABER, C. D. BOEHM and H. H. KAZAZIAN, 1986 Nonuniform recombination within the human β-globin gene cluster: a reply to B. S. Weir and W. G. Hill. Am. J. Hum. Genet. **38:** 779–781.

COCOZZA, S., A. MONTICELLI, GARAFALO, S., G. RICHARD, R. CORTESE, N. QUIRICO, A. RICCI, R. CICERONE, F. ROSSI and S. VARRONE, 1987 DNA polymorphisms as potential genetic risk markers for cardiovascular diseases. Boll. Soc. It. Biol. Sper. **58:** 771-777.

COX, N. J., G. I. BELL and K.-S. XIANG, 1988 Linkage disequilibrium in the human insulin/insulin-like growth factor II region of human chromosome 11. Am. J. Hum. Genet. **43:** 495–501.

DARNFORS, C., O. WIKLUND, J. NILSSON, B. GERARD, P. CARLSSON, S. JOHANSSON, C. BONDJERS and G. BJURSELL, 1989 Lack of correlation between the apolipoprotein B *Xba*I polymorphism and blood lipid levels in a Swedish population. Atherosclerosis **75:** 183–188.

DEMANT, T., R. S. HOULSTON, M. J. CASLAKE, J. J. SERIES, J. SHEPHERD, C. J. PACKARD and S. E. HUMPHRIES, 1988 Catabolic rate of low density lipoprotein is influenced by variation in the apolipoprotein B gene. J. Clin. Invest. **82:** 797–802.

DUNNING, A. M., P. DURIEZ, N. VU DAC, J. C. FRUCHART and S. E. HUMPHRIES, 1988 Association between epitopes detected by monoclonal antibody BIP-45 and the *Xba*I polymorphism of apolipoprotein B. Clin. Genet. **33:** 181–188.

EFRON, B., 1982 *The Jackknife, the Bootstrap, and Other Resampling Plans.* Society for Industrial and Applied Mathematics, Philadelphia.

HAVILAND, M. B., A. M. KESSLING, J. DAVIGNON and C. F. SING, 1991 Estimation of Hardy-Weinberg and pairwise disequilibrium in the apo AI-CIII-AIV gene cluster. Am. J. Hum. Genet. **49:** 350–365.

HEDRICK, P. W., 1987 Gametic disequilibrium measures: proceed with caution. Genetics **117:** 331–341.

HEGELE, R. A., R. PLAETKE and J.-M. LALOUEL, 1990 Linkage disequilibrium between DNA markers at the low-density lipoprotein receptor gene. Genet. Epidemiol. **7:** 69–81.

HOULSTON, R. S., P. R. TURNER, J. REVILL, B. LEWIS and S. E. HUMPHRIES, 1988 The fractional catabolic rate of low density lipoprotein in normal individuals is influenced by variation in the apolipoprotein B gene: a preliminary study. Atherosclerosis **71:** 81–85.

HUANG, L.-S., M. E. RIPPS and J. L. BRESLOW, 1990 Molecular basis of five apolipoprotein B gene polymorphisms in noncoding regions. J. Lipid Res. **31:** 71–77.

JENNER, K., A. SIDOLI, M. BALL, J. R. RODRIGUEZ, F. PAGANI, G. GIUDICI, C. VERGANI, J. MANN, F. E. BARALLE and C. C. SHOULDERS, 1988 Characterization of genetic markers in the 3′ end of the apo B gene and their use in family and population studies. Atherosclerosis **69:** 39–49.

KESSLING, A., S. OUELLETTE, O. BOUFFARD, A. CHAMBERLAND, C. BÉTARD, E. SELINGER, M. XHIGNESSE, S. LUSSIER-CACAN and J. DAVIGNON, 1991 Patterns of association between genetic

variability in apolipoprotein (apo) B, apo AI-CIII-AIV and cholsterol ester transfer protein gene regions and quantitative variation in lipid and lipoprotein traits: influence of gender and exogenous hormones. Am. J. Hum. Genet. (in press).

LAW, A., L. M. POWELL, H. BRUNT, T. J. KNOTT, D. G. ALTMAN, J. RAJPUT, S. C. WALLIS, R. J. PEASE, L. M. PRIESTLEY, J. SCOTT, G. J. MILLER and N. E. MILLER, 1986 Common DNA polymorphism within coding sequence of apolipoprotein B gene associated with altered lipid levels. Lancet **1:** 1301–1303.

LEIGHTON, J. K., J. JOYNER, J. ZAMARRIPA, M. DEINES and R. A. DAVIS, 1990 Fasting decreases apolipoprotein B mRNA editing and the secretion of small molecular weight apoB by rat hepatocytes: evidence that the total amount of apoB secreted is regulated post-transcriptionally. J. Lipid Res. **31:** 1663–1668.

LEREN, T. P., K. BERG, I. HJERMANN and P. LEREN, 1988 Further evidence for an association between the *Xba*I polymorphism at the apolipoprotein B locus and lipoprotein level. Clin. Genet. **34:** 347–351.

LEWONTIN, R. C., 1988 On measures of gametic disequilibrium. Genetics **120:** 849–852.

LITT, M., and L. B. JORDE, 1986 Linkage disequilibria between pairs of loci within a highly polymorphic region of chromosome 2q. Am. J. Hum. Genet. **39:** 166–178.

MA, Y., V. N. SCHUMAKER, R. BUTLER and R. S. SPARKES, 1987 Two DNA restriction fragment length polymorphisms associated with Ag(t/z) and Ag(g/c) antigenic sites of human apolipoprotein B. Arteriosclerosis **7:** 301–305.

MONSALVE, M. V., R. YOUNG, J. JOBSIS, S. A. WISEMAN, S. DHAMU, J. T. POWELL, R. M. GRENHALGH and S. E. HUMPHRIES, 1988 DNA polymorphisms of the gene for apolipoprotein B in patients with peripheral arterial disease. Atherosclerosis **70:** 123–129.

MYANT, N. B., J. GALLAGHER, M. BARBIR, G. R. THOMPSON, D. WILE and S. E. HUMPHRIES, 1989 Restriction fragment length polymorphisms in the apo B gene in relation to coronary artery disease. Atherosclerosis **77:** 193–201.

PAULWEBER, B., W. FRIEDL, F. KREMPLER, S. E. HUMPHRIES and F. SANDHOFER, 1990 Association of DNA polymorphism at the apolipoprotein B gene locus with coronary heart disease and serum very low density lipoprotein levels. Arteriosclerosis **10:** 17–24.

SHOULDERS, C. C., N. B. MYANT, A. SIDOLI, J. C. RODRIGUEZ, C. CORTESE, F. E. BARALLE and R. CORTESE, 1985 Molecular cloning of human LDL apolipoprotein B cDNA. Atherosclerosis **58:** 277–289.

SINNOCK, P., and C. F. SING, 1972 Analysis of multilocus genetic systems in Tecumseh, Michigan. II. Consideration of the correlation between non-alleles in gametes. Am. J. Hum. Genet. **24:** 393–415.

SNIDERMAN, A. D., and J. SILBERBERG, 1990 Is it time to measure apolipoprotein B? Arteriosclerosis **10:** 665–667.

SOKAL, R. R., and F. J. ROHLF, 1981 *Biometry. The Principles and Practice of Statistics in Biological Research*, Ed. 2. W. H. Freeman, New York.

TALMUD, P. J., N. BARNI, A. M. KESSLING, P. CARLSSON, C. DARNFORS, G. BJURSELL, D. GALTSON, V. WYNN, H. KIRK, M. R. HAYDEN and S. E. HUMPHRIES, 1987 Apolipoprotein B gene variants are involved in the determination of serum cholesterol levels: a study in normo- and hyperlipidaemic individuals. Atherosclerosis **67:** 81–89.

TEMPLETON, A. R., E. BOERWINKLE and C. F. SING, 1987 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. Genetics **117:** 343–351.

TEMPLETON, A. R., C. F. SING, A. KESSLING and S. HUMPHRIES, 1988 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. Genetics **120:** 1145–1154.

THOMPSON, E. A., S. DEEB, D. WALKER and A. G. MOTULSKY, 1988 The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes. Am. J. Hum. Genet. **42:** 113–124.

WARD, R. H., and C. F. SING, 1970 A consideration of the power of the χ^2 test to detect inbreeding effects in natural populations. Am. Nat. **104:** 355–365.

WEIR, B. S., 1979 Inferences about linkage disequilibrium. Biometrics **35:** 235–254.

WEIR, B. S., 1990 *Genetic Data Analysis. Methods for Discrete Population Genetic Data*. Sinauer Associates, Sunderland, Mass.

WEIR, B. S., and L. D. BROOKS, 1986 Disequilibrium on human chromosome 11p. Genet. Epidemiol. Suppl. **1:** 177–183.

WEIR, B. S., and C. C. COCKERHAM, 1989 Complete characterization of disequilibrium at two loci, pp. 86–110 in *Mathematical Evolutionary Theory*, edited by M. W. FELDMAN. Princeton University Press, Princeton, N.J.

WEIR, B. S., and W. G. HILL, 1986 Non-uniform recombination within the human β-globin gene cluster. Am. J. Hum. Genet. *38:* 776–778.

WIKLUND, O., C. DARNFORS, G. BJURSELL, J. NILSSON, T. LINDE'N, S. O. OLOFSSON, L. WILHELMSEN and G. BONDJERS, 1989 *Xba*I restriction fragment length polymorphism of apolipoprotein B in Swedish myocardial infarction patients. Eur. J. Clin. Invest. **19:** 255–258.

Communicating editor: B. S. WEIR