A Cladistic Analysis of Phenotype Associations with Haplotypes Inferred From Restriction Endonuclease Mapping. II. The Analysis of Natural Populations

Alan R. Templeton,* Charles F. Sing,† Anna Kessling‡ and Stephen Humphries§

**Department of Biology, Washington University, St. Louis, Missouri 63130, † Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109-0618, ‡Clinical Research Institute of Montreal, 110 Quest, Avenue des Pins, Montreal, Quebec H2W 1R7, Canada, and §Charing Cross Medical Research Centre, Lurgan Avenue, Hammersmith, London W6 8LW, Great Britain*

Manuscript received April 25, 1988
Revised copy accepted August 18, 1988

ABSTRACT

Genes that code for products involved in the physiology of a phenotype are logical candidates for explaining interindividual variation in that phenotype. We present a methodology for discovering associations between genetic variation at such candidate loci (assayed through restriction endonuclease mapping) with phenotypic variation at the population level. We confine our analyses to DNA regions in which recombination is very rare. In this case, the genetic variation at the candiate locus can be organized into a cladogram that represents the evolutionary relationships between the observed haplotypes. Any mutation causing a significant phenotypic effect should be imbedded within the same historical structure defined by the cladogram. We showed, in the first paper of this series, how to use the cladogram to define a nested analysis of variance (NANOVA) that was very efficient at detecting and localizing phenotypically important mutations. However, the NANOVA of haplotype effects could only be applied to populations of homozygous genotypes. In this paper, we apply the quantitative genetic concept of average excess to evaluate the phenotypic effect of a haplotype or group of haplotypes stratified and contrasted according to the nested design defined by the cladogram. We also show how a permutational procedure can be used to make statistical inferences about the nested average excess values in populations containing heterozygous as well as homozygous genotypes. We provide two worked examples that investigate associations between genetic variation at or near the *Alcohol dehydrogenase* (*Adh*) locus and Adh activity in *Drosophila melanogaster*, and associations between genetic variation at or near some apolipoprotein loci and various lipid phenotypes in a human population.

RECOMBINANT DNA technology makes it possible to survey populations for restriction site variability in small regions of the chromosome containing genes with known biochemical or physiological functions. By simultaneously studying traits that are related to the genes' known biochemical or physiological functions, it is possible to assign phenotypic effects to specific alleles or haplotypes at or near the genes of known function. This measured genotype approach allows one to investigate many of the problems traditionally addressed by quantitative genetic techniques, but with the advantage that the genes that are being considered are well defined, molecularly characterized units.

There are several problems that need to be addressed in order to implement this approach (Templeton, Boerwinkle and Sing 1987). They are (1) how to define a unit of genetic analysis when the data consist of multiple polymorphic sites with linkage disequilibrium between sites? (2) how to detect indirect phenotypic associations with the measured genetic unit that are caused by linkage disequilibrium? (3) which haplotype categories are likely to be associated with multiple phenotypic effects? (4) how to detect such phenotypic heterogeneity? (5) where are the phenotypically important genetic variants located in the cloned DNA region? (6) can a method be developed that deals with quantitative phenotypes as well as categorical phenotypes? and (7) how to estimate and test for the presence of phenotypic effects of specific alleles or haplotypes for a phenotype that is expressed only in diploid individuals?

An obvious solution to the first problem is to regard the unit of genetic analysis as the haplotypes determined by the states of all available polymorphic sites considered simultaneously (Templeton, Boerwinkle and Sing 1987). We retain this basic unit of analysis in this paper. However, the haplotypes present in heterozygous diploid individuals are often not unambiguously defined from the restriction fragment data. In this paper, we will present an algorithm for assigning the most probable haplotypes and for estimating haplotype frequencies in diploid populations containing heterozygotes.

Given that there is very little recombination in the DNA region of interest, a cladogram can be constructed that represents the evolutionary relationships between the present-day haplotypes. The cladogram is then used to define a nested statistical analysis that addresses problems two, three and six (TEMPLETON, BOERWINKLE and SING 1987). This cladistic approach does not directly address problem five, but it does identify haplotype categories that may warrant further molecular characterization in the attempt to understand the causes of the observed phenotypic associations.

The utility of this cladistic approach was first illustrated by a worked example on associations between restriction site variation observed in the *Alcohol dehydrogenase* (*Adh*) genetic region with Adh activity levels in *Drosophila melanogaster* (TEMPLETON, BOERWINKLE and SING 1987). The Adh data were obtained from homozygous strains with a common genetic background (AQUADRO *et al.* 1986). Thus, experimental genetic manipulations to create contrasts between homozygotes were used to solve problem seven. However, when performing surveys on outbred populations, a single haplotype can exist in a variety of different genotypes. Hence, there is no longer a 1 to 1 correspondence between haplotypes and the genotypes expressing the phenotype. In this paper we extend the statistical strategy given in TEMPLETON, BOERWINKLE and SING (1987) to the analysis of samples from outbred populations. This extension will be illustrated by two worked examples. First, we will reanalyze the Adh activity data disregarding the fact that only homozygous strains are involved. This will allow a direct comparison of the results of the techniques developed in this paper to the results obtained by the more traditional statistical analyses that are possible when experimentally derived homozygous genotypes are available. Second, we will analyze a sample from a human population of unrelated individuals who were scored for restriction site haplotypes in a region coding for three apolipoproteins and measured for various lipid variables. The cladistic analysis of these human data will suggest that certain haplotypes may be phenotypically heterogeneous in their effects. Hence, this example will be used to illustrate a solution to problem four—how does one statistically detect phenotypic heterogeneity within a haplotype class?

"PHENOTYPES" OF HAPLOTYPES IN A DIPLOID POPULATION

One of the oldest and most fundamental problems in quantitative genetics stems from the fact that although most phenotypes of interest are expressed only in diploid individuals, any genetic component that contributes to that phenotype has to be passed on to the next generation through a haploid gamete. Consequently, statistics that assign "phenotypic measures" to haploid genetic elements (alleles or haplotypes) are at the very core of quantitative genetic theory.

Two such measures are commonly used today, and both were invented by R. A. FISHER (1918). The first is the average excess. The average excess of a particular haplotype is the average phenotypic measurements made on bearers of that haplotype minus the overall population mean (TEMPLETON 1987). The second measure is the average effect, which is the least-squares regression coefficient that defines the linear relationship between the phenotype and the number of copies of each haplotype (zero, one or two) borne by an individual. The average effects can always be calculated from the average excesses and data on genotype frequencies, and the two measures are identical under Hardy-Weinberg genotype frequencies (TEMPLETON 1987). Because the average excess is simpler to calculate and has a more straightforward biological meaning, we will use the average excess as our means of assigning phenotypes to haplotypes.

We now introduce the notation we will use in the computation of the average excess measures. Let y_k be the phenotypic value of individual k. Let $\bar{y}$ be the average phenotypic value in a sample of n individuals. Let h_{ik} be the number of haplotypes of type i borne by individual k, which has the possible values of 0, 1 or 2. With these definitions, the average excess of haplotype i is

$$a_i = \sum_{k=1}^{n} y_k h_{ik} \bigg/ \sum_{k=1}^{n} h_{ik} - \bar{y}. \tag{1}$$

Using equation (1), we can now estimate the phenotypic effects associated with any particular haplotype. However, we still need to address the issue of testing the significance of these estimates.

A NESTED PERMUTATIONAL TEST OF AVERAGE EXCESSES

Direct statistical testing of average excesses is not generally possible because the phenotypic value of each heterozygote contributes to two different average excess measurements. Hence, the haplotype effects cut across the sample of individual phenotypic observations in a confounded fashion, thereby invalidating the use of many standard statistical techniques, such as an analysis of variance of the haplotype effects. As an alternative, investigators have often analyzed the phenotypic differences of genotypes to indirectly confirm the significance of the average excess measures. For example, BOERWINKLE *et al.* (1987) estimated the average effects (which are the same as the average excesses in this case because the population analyzed had Hardy-Weinberg genotype frequencies) of the three common alleles at the apolipoprotein E locus on the levels of several traits involved in lipid metabolism. However, the statistical confirmation of significant genetic effects was accomplished by performing analyses of variance (ANOVA) using the genotypes, not alleles, as the treatment stratifications. This is a useful approach as long as the number of alleles or haplotypes is relatively small. However, restriction enzyme mapping has the potential for distinguishing a large number of haplotypes in the region of the DNA where a gene is located. In general, N haplotypes define $N(N + 1)/2$ genotypes. Dividing a data set into $N(\mathrm{N} + 1)/2$ treatment effects can seriously erode statistical power simply by making the sample size associated with each treatment effect small. Therefore, for a given sample size, one has greater statistical power to detect differences among N haplotype effects than among $N(N + 1)/2$ genotypic effects unless there is much over- or underdominance or other nonadditive interaction effects.

When all genotypes are made homozygous, as with the Adh data of AQUADRO *et al.* (1986), the problem of assigning haplotype effects is eliminated and average excesses can be analyzed with standard nested ANOVA (NANOVA) procedures (TEMPLETON, BOERWINKLE and SING 1987). However, when the haplotype effects are confounded by heterozygosity, the ANOVA, or NANOVA, cannot be used. As an alternative, we propose a sample reuse procedure. Sample reuse procedures are commonly used for estimation and testing of statistics whose distributions are not known or easily derived (EFRON 1982). In particular, random or systematic permutations of the original data are appropriate for ANOVA-like situations (EDGINGTON 1987). We will use random permutations of the sample to generate the distribution of NANOVA-like statistics under the null hypothesis that the haplotypes and clades of haplotypes are not associated with the phenotypic variability measured by their corresponding average excess values.

As in TEMPLETON, BOERWINKLE and SING (1987), the first step of the analysis is to construct a cladogram of the haplotypes that reflects their evolutionary relationships. The basic rationale for this step is that any mutation causing phenotypically important alterations will be embedded somewhere in the same historical/evolutionary framework defined by the restriction sites. The consequence is a correlation between phenotypic effects and evolutionary relatedness. Figure 1 shows the cladogram for the *Adh* haplotypes from AQUADRO *et al.* (1986) that was used in the analysis of TEMPLETON, BOERWINKLE and SING (1987).

The second step is to use the cladogram to define a nested statistical analysis. As discussed in more detail in TEMPLETON, BOERWINKLE and SING (1987), we start with the average excess values assigned to haplotypes. Next, average excess values are estimated for larger and larger branches (clades) of the cladogram, defined in a nested fashion, until the next level of nesting would encompass the entire cladogram (hence, the largest clades are one step below the entire cladogram). We will call the units defined by the various nested branches of the cladogram as c-step clades, where c indicates the level of nesting and represents the maximum number of mutational differences that interconnect the haplotypes within a c-step clade. In this terminology, the haplotypes correspond to 0-step clades (TEMPLETON, BOERWINKLE and SING 1987). Evolutionarily related sets of 0-step clades are then pooled together to form a smaller number of 1-step clades, and evolutionarily related sets of 1-step clades are in turn pooled together to form an even smaller number of 2-step clades, etc., etc. The nesting of the *Adh* region haplotypes is indicated in Figure 1 (from TEMPLETON, BOERWINKLE and SING 1987).

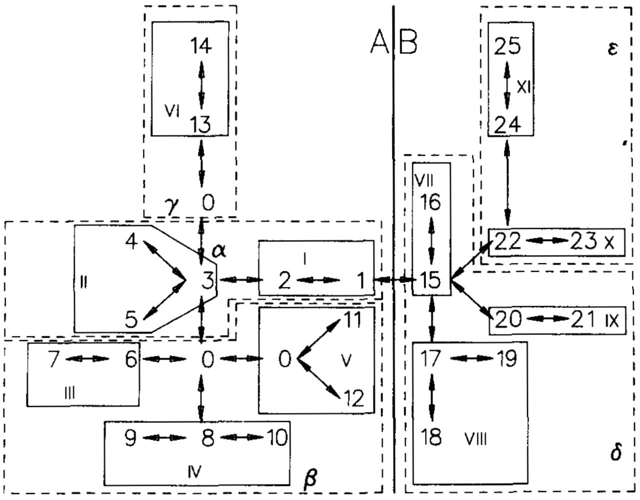

FIGURE 1.—Cladogram of the haplotypes found in the *Alcohol dehydrogenase* genetic region of *D. melanogaster* (from TEMPLETON, BOERWINKLE and SING 1987). Numbers *1* through *25* represent distinct haplotypes as detected by restriction site mapping (AQUADRO *et al.* 1986). The number "0" refers to inferred intermediate haplotypes that were not actually present in the sample but that are needed to interconnect the existing haplotypes. Each double-headed arrow represents a single mutational change detectable by restriction mapping. The haplotypes are then nested according to the algorithm of TEMPLETON, BOERWINKLE and SING (1987). The 0-step clades (haplotypes) are indicated by Arabic numerals, 1-step clades (enclosed by solid lines) are indicated by Greek letters, 2-step clades (dotted lines) are indicated by Roman numerals, and 3-step clades (the *A*,*B* partition of the cladogram) are indicated by capital letters.

Since average excesses are the phenotypic measures we are analyzing, we need to define a nested set of average excess measures in order to implement this nested statistical analysis. Let $a_{i(M)}$ be the average excess of the c-step clade, i, nested within c+1 step clade, M. This is the average excess assigned to the ith grouping of haplotypes at the c-step level measured as a deviation from the mean of all haplotypes associated with the $c+1$ step clade, M (the usual average excess measures deviations from the general population mean):

$$a_{i(M)} = \sum_{k=1}^{n} y_k h_{ik} \Big/ \sum_{k=1}^{n} h_{ik} - \sum_{k=1}^{n} y_k h_{Mk} \Big/ \sum_{k=1}^{n} h_{Mk} \qquad (2)$$

where h_{Mk} refers to the total number of clade M haplotypes carried by individual k.

In analogy to a NANOVA, sum of squares statistics can be defined from these nested average excesses. Let $p_{i(M)}$ be the relative frequency of c-step clade, i, *within* the $c+1$ step clade, M, category; that is,

$$p_{i(M)} = \sum_{k=1}^{n} h_{ik} \Big/ \sum_{k=1}^{n} h_{Mk}. \qquad (3)$$

Then, the weighted sum of squares of the c-step clades nested within $c+1$ step clade M, say $S_{c(M)}$, is

$$S_{c(M)} = \sum_{i} p_{i(M)} a_{i(M)}{}^2, \qquad (4)$$

where i is summed over all c-step clades contained within $c+1$ step clade M. The relative frequency of clade M of the $c+1$ step in the total population is given by

$$p_M = \sum_{k=1}^{n} h_{Mk}/(2n). \qquad (5)$$

The total sum of squares at level c nested within level $c+1$ is

$$S_c = \sum_{M} p_M S_{c(M)}, \qquad (6)$$

where the M are summed over all $c+1$ step clades.

The statistical significance of the sums of squares defined by Equations 4 and 6 are evaluated by a random permutation procedure. Because we are using a nested design, the permutations are carried out in a hierarchical, nested fashion. Starting at the highest clade level, say $c_{\max}$, the null hypothesis that there are no phenotypic associations at this level implies that the data are finitely exchangeable across the $c_{\max}$-level clades. This exchangeable distribution can be simulated by randomly permuting the entire vector of haplotype counts for individual k, $\{h_{i,k}\}$ over the individual phenotypic values. The null hypothesis is rejected if the probability of realizing the observed sums of squares statistics among the large set of values generated by random permutation falls below a preassigned significance level, say 0.05.

One could test for phenotypic associations at the lower levels in the hierarchy of clades by permuting observations that are nested within the next higher clade level. However, this procedure would result in a serious erosion of statistical power because the data set becomes increasingly subdivided into smaller and smaller nested groups of observations at the lower clade levels. Since the power of a permutational test depends critically upon the number of observations that are available for permutation, there would often be very little power at these lower levels of nesting even if the total sample size were large.

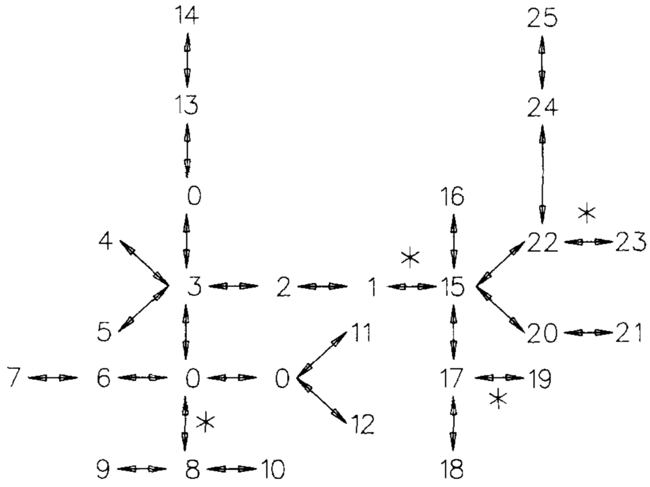

FIGURE 2.—Diagrammatic summary of the results of the NANOVA analysis given in TEMPLETON, BOERWINKLE and SING (1987). The *Adh* genetic region cladogram is given, with asterisks indicating the localization of phenotypically important changes within the cladogram.

We avoid this problem with a conditional, hierarchical permutation procedure. As one proceeds from the highest clade level to the lower levels, the only groupings that are retained as nesting categories for between individual permutations are those for which the hypothesis of exchangeability has been rejected. For example, suppose the $c_{\max}$-level analysis lead to the rejection of the null hypothesis of exchangeability. Then, to test the null hypothesis that there are no phenotypic associations with the ($c_{\max} - 1$)-level clades, it is important that the h values are exchanged at random only between individuals bearing haplotypes from the same $c_{\max}$-level clades. On the other hand, if the null hypothesis of exchangeability at the $c_{\max}$-level is accepted, then the null hypothesis of no ($c_{\max} - 1$)-level associations is tested by permuting the h values over all individuals.

COMPARISON OF THE PERMUTATION TEST AND THE NANOVA: AN APPLICATION TO THE *Drosophila* Adh DATA

We have applied this permutation procedure to the *Drosophila* Adh activity data previously analyzed by TEMPLETON, BOERWINKLE and SING (1987). Figure 1 gives the nested design defined by the cladogram, and Figure 2 summarizes the results inferred from the standard NANOVA given in TEMPLETON, BOERWINKLE and SING (1987). We performed a nested per-

TABLE 1

Nested permutational analysis of alcohol dehydrogenase activities in *D. melanogaster* using the hierarchical, conditional permutation procedure described in the text

Source	Sums of squares	Significance
3-Step clades	4.299	0.000***
2-Step clades	0.051	0.562
Within *A*	0.080	0.223
Within *B*	0.005	0.842
1-Step clades	0.288	0.083
(Observations permuted within *A*)		
Within α	0.078	0.281
Within β	0.598	0.007**
(Observations permuted within *B*)		
Within δ	0.672	0.207
Within ϵ	0.031	0.736
0-Step clades	0.467	0.188
(Observations permuted within *A* minus *IV*)		
Within *I*	0.059	0.490
Within *II*	0.179	0.133
Within *III*	0.005	0.828
Within *V*	0.073	0.511
Within *VI*	0.064	0.310
(Observations permuted within *IV*)		
Within *IV*	0.201	0.317
(Observations permuted within *B*)		
Within *VII*	0.000	0.987
Within *VIII*	3.498	0.048*
Within *IX*	0.828	0.276
Within *X*	1.035	0.104
Within *XI*	0.016	0.803

Asterisks highlight results that are significant at the 5% level (*), the 1% level (**), and 0.1% level (***).

TABLE 2

Localization of the significant phenotypic effect detected at the 1-step clade level for the Adh activity data in the permutational analysis given in Table 1

Contrast (i *vs.* j)	$a_{i(B)}$-$a_{j(B)}$	Significance
III vs. IV	−1.654	0.006**
III vs. V	−0.277	0.703
IV vs. V	1.378	0.039*

The significance of the three possible pairwise contrasts of the nested average excesses within 2-step clade are given as determined by 1000 random permutations of the observations nested within 3-step clade *A*.

mutational analysis of the same data by generating 1000 random sets of nested permutations, which is sufficient for accurate inference at the 5% level of significance (Edgington 1987). Significantly large values of statistic (6) indicate the clade level at which phenotypically significant associations are found, and these phenotypic associations can be further localized to the particular clade at that level of the cladogram by decomposing statistic (6) into its component statistic (4) values. If a phenotypic effect has been localized to the *c*-level clades within a particular *c*+1 step clade (say *M*) but there is still some ambiguity as to its precise location, further localization is possible by examining the statistics $a_{i(M)} - a_{j(M)}$, where *i* and *j* are a pair of *c*-step clades within clade *M* that are adjacent in the cladogram. The significances of these contrasts are also evaluated through permutational testing.

Table 1 gives the significance levels of the resulting statistics as determined by 1000 random permutations of the original data set. The permutational analysis detects a significant phenotypic effect at the 3-step level, which corresponds to the asterisk over the arrow connecting haplotypes *1* and *15* in Figure 2. Because the permutation analysis rejects the hypothesis of exchangeability between the two 3-step clades, all lower level permutations must be nested within the 3-step clades *A* and *B* (Figure 1). Going down to the 2-step level in Table 1, we see that the hypothesis of exchangeability is not rejected for the 2-step clades nested within 3-step clades *A* or *B*. Given that there are no 2-step effects, we now permute the data at the 1-step level nested within the two 3-step clades, *A* and *B*. A 1-step effect is found in 2-step clade beta that is significant at the 1% level.

The significant 1-step effect within clade beta can be further localized by contrasting the average excess values, as described earlier. The results are shown in Table 2. The significant phenotypic association is clearly localized to the mutational step leading to 1-step clade *IV*, the same conclusion reached by the NANOVA (Figure 2).

Moving on to the 0-step effects, our analysis up to this point indicates that all the data nested within clade *B* are exchangeable under the null hypothesis of no 0-step effects. Within clade *A*, our analysis indicates two exchangeable categories under the null hypothesis of no 0-step effects: the observations nested within 1-step clade *IV*, and the observations found in 1-step clades *I*, *II*, *III*, *V* and *VI* (*i.e.*, all the observations within clade *A* that are not in clade *IV*, which is designated at "*A* minus *IV*" in Table 1). The 0-step permutational analysis was then carried out nested within the three categories *B*, *A-IV*, and *IV*. As shown in Table 1, a significant 0-step effect is detected and localized within 1-step clade *VIII* (the asterisk under the *17–19* transition in Figure 2).

As can be seen by contrasting Table 1 with Figure 2, both the permutational analysis and the NANOVA detected and localized identically three phenotypically important mutations. In addition, the NANOVA detected a significant effect at the 0-step clade level associated with haplotype *23* that was not detected by the permutational analysis. This reduction in power

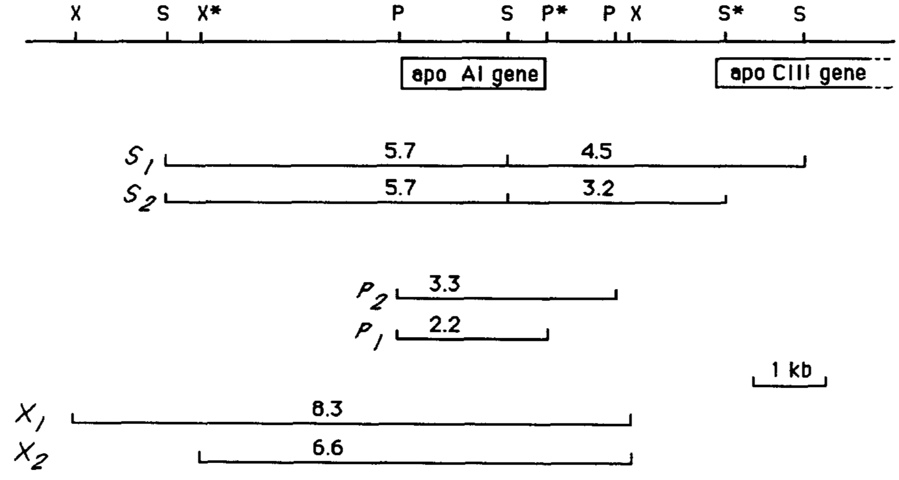

FIGURE 3.—Diagram of the *Apolipoprotein A-I, C-III, A-IV* gene complex on the long arm of chromosome *11,* showing the locations of restriction enzyme cleavage sites. The following abbreviations are used: X = *Xmn*I, S = *Sst*I, and P = *Pst*I. Asterisks indicate the locations of the polymorphic restriction sites. The resulting fragment sizes caused by the polymorphic and invariant sites are indicated below the chromosomal diagram. The locations of the *A-I* and *C-III* genes are indicated on the diagram. The *A-IV* gene is located to the right of the *C-III* gene.

of the permutational analysis with respect to the NANOVA is not surprising. The permutation test is a non-parametric procedure, and as expected it is less powerful than the parametric NANOVA. Although some power is lost, the permutational procedure does have some important advantages. First, being nonparametric, it is more robust to possible deviations from the underlying normality assumed in the NANOVA. Even more importantly, the permutational test can be applied to complex samples for which the NANOVA is not applicable. As we will now illustrate, the permutational analysis can be applied in a similar manner to a sample from a natural, outbreeding population.

AN ANALYSIS OF VARIATION IN TRIGLYCERIDE AND CHOLESTEROL LEVELS IN A HUMAN POPULATION

Our second example will utilize data collected by the St. Mary's Hospital Metabolic Unit group described in KESSLING, HORSTHEMKE and HUMPHRIES (1985). This group consists of 89 unrelated, adult individuals of both sexes who were biased in favor of being normo- or hyperlipidemic. Southern blots of digested DNA from all individuals were hybridized with a cloned 2.2-kb genomic *Pst*I fragment containing the *apoAI* gene, in the *apo AI-CIII-AIV* gene region on chromosome *11* (KESSLING, HORSTHEMKE and HUMPHRIES 1985). Figure 3 indicates the location of three polymorphic restriction sites detected on hybridizing the probe to DNA digested with the enzymes *Xmn*I, *Pst*I, and *Sst*I. All 89 individuals had this DNA region scored with all three of these restriction enzymes.

Our first task is to assign haplotypes to all the individuals. Complications arise because some individuals have restriction fragment length profiles that are compatible with more than one haplotype configuration. Table 3 summarizes the haplotype information obtained from the Southern blots. As can be seen, in 78 of the 89 individuals, there was only one haplotype configuration consistent with the observed fragment lengths. The remaining 11 individuals had two alternative possible haplotype configurations, and their haplotype numbers were estimated by maximum likelihood using an E-M algorithm on the total data set (HILL 1974).

The next step of the analysis is the construction of the cladogram. As can be seen from Table 3, nonzero frequencies are assigned to 7 of the 8 possible haplotypes defined by the three polymorphic restriction site makers. However, haplotypes *111* and *100* are not found in the unambiguous subset of the data. We feel it is best to confine the cladistic analysis to those haplotypes that are known to exist. Hence, the analysis will be confined to the first 5 haplotypes shown in Table 3.

Figure 4 shows the maximum parsimony cladogram constructed for these five haplotypes. As can be seen, three of the haplotypes can be derived by single mutational steps from haplotype *010,* the most common haplotype. Haplotype *001* (found only in one unambiguous heterozygote) requires either a recombinational event between haplotypes *011* and *000* or a convergent mutational step.

Ignoring for the moment the possibility of recombination, the position of haplotype *001* in the cladogram shown in Figure 4 is still ambiguous, as shown by the dashed lines which represent alternative evolutionary pathways. As shown by TEMPLETON (1983), convergence caused by a loss of a restriction site is much more likely than convergence caused by a gain of a site. In particular, for 6-base cutters like *Pst*I and *Sst*I (the two enzymes used to detect the sites of possible convergence associated with haplotype *001*), a convergent loss is about 18 times more likely than a gain over the relatively short periods of evolutionary time that are relevant for intraspecific polymorphisms (TEMPLETON 1983). However, to use this knowledge to discriminate between the two alternatives shown in Figure 4, we also need information about the root

TABLE 3

Estimated haplotype configurations in the St. Mary's population

		No. of each haplotype (h_{ik})						
No. of individuals	Possible genotypes	*010*	*011*	*000*	*110*	*001*	*111*	*100*
45	*010/010*	2	0	0	0	0	0	0
7	*010/011*	1	1	0	0	0	0	0
10	*010/000*	1	0	1	0	0	0	0
12	*010/110*	1	0	0	1	0	0	0
1	*011/001*	0	1	0	0	1	0	0
2	*000/000*	0	0	2	0	0	0	0
1	*110/110*	0	0	0	2	0	0	0
3	*010/001* or *011/000*	0.67	0.33	0.33	0	0.67	0	0
6	*010/100* or *000/110*	0.56	0	0.44	0.44	0	0	0.56
2	*111/010* or *110/011*	0.23	0.77	0	0.77	0	0.23	0
Estimated haplotype frequencies:		0.70	0.06	0.10	0.10	0.02	0.00	0.02

A "1" indicates the presence of a restriction site, and a "0" its absence. The order of polymorphic sites is always *Xmn*I, *Pst*I and *Sst*I, corresponding to their physical order on the DNA molecule (Figure 3).

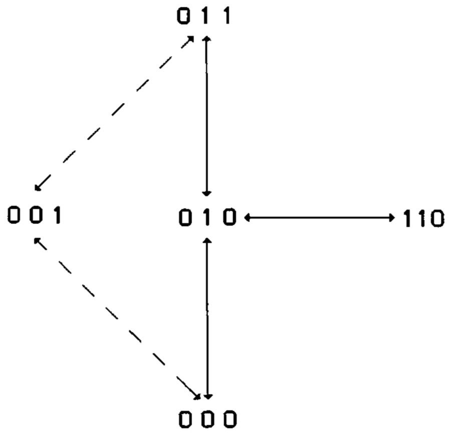

FIGURE 4.—Cladogram of the *ApoA-I, C-III, A-IV* region haplotypes known to exist in the St. Mary's population. The cladogram portrays the mutational steps (arrows) needed to interrelate the haplotypes to one another. The arrows are double-headed because the ancestral haplotype is not known. The dashed arrows leading from haplotype *011* and *000* towards haplotype *001* indicate that *001* is either a recombinant of *011* and *000* or has been derived by an additional mutation from either *011* or *000.*

(ancestral haplotype) of the cladogram. Because this information is lacking, no meaningful discrimination can be made between these alternatives. Moreover, it is not clear at this point exactly how common recombination is in this region. Since we only have three polymorphic sites, it is very difficult to distinguish between recombination *vs.* mutational convergence from the haplotype states. Since we cannot exclude the possibility of recombination, and because even if we assume there is no recombination, we still cannot resolve the position of haplotype *001* in the cladogram, it is best to exclude haplotype *001* from the cladogram. This results in a very simple cladogram in which haplotype *010* is related to all other haplotypes by a single mutational step.

Using the nesting algorithm of TEMPLETON, BOERWINKLE and SING (1987), there is no nesting above the haplotype level in this simplified cladogram. In other words, when haplotype *001* is excluded, the design implied by the resulting cladogram collapses into a single level unnested design. Hence, we are using the human data primarily to illustrate that this technique can be applied to natural, outbred populations, whereas the Adh example illustrates the strengths of a nested, cladistic approach. Although simple, the cladogram shown by the solid arrows in Figure 4 still defines the relevant contrasts needed to localize any phenotypic associations in the human data; namely, the contrasts between haplotype *010* (the central haplotype of the cladogram) versus haplotypes *011*, *000* and *110*. Hence, our statistical design is determined by cladistic criteria even in this unnested case.

We now turn our attention to the phenotypic measurements. The 89 individuals had their triglyceride, total cholesterol, and high density lipoprotein (HDL) cholesterol levels determined in serum samples taken after a 12-hr fast using standard methods (KESSLING, HORSTHEMKE and HUMPHRIES 1985). However, there are reasons for believing that certain transformations of these original phenotypic measurements might have greater biological meaning. These traits are measures of concentrations of certain lipid intermediates along an interconnected metabolic pathway. Genetic variation in the apolipoproteins would be expected to alter the dynamics of some or all of this pathway. As shown by SAVAGEAU (1976), the natural space to describe the dynamics of a metabolic pathway is a logarithmic one. Accordingly, we took the natural logarithms of the triglyceride, total cholesterol, and HDL cholesterol levels (hereafter abbreviated as ln(T), ln(C), and ln(H)).

Because age and sex differences are major deter-

TABLE 4

Average excesses for each of the six traits[a]

	Trait					
Haplotype	Trig.	Chol.	HDL	Ln(T)	Ln(C)	Ln(H)
010	−0.410	−0.073	0.023	−0.055	−0.007	0.016
011	1.603	−0.206	−0.067	0.449	−0.014	−0.024
000	0.300	0.371	−0.014	−0.070	0.039	0.009
110	0.527	0.093	−0.038	0.081	0.006	−0.045
001	0.331	−0.419	0.032	0.052	−0.033	−0.003

[a] Triglyceride level (Trig.), total cholesterol level (Chol.), HDL cholesterol level (HDL), the natural logarithm of triglyceride level (Ln(T)), the natural logarithm of cholesterol level (Ln(C)), and the natural logarithm of HDL cholesterol level (Ln(H)). All traits have been adjusted for age and sex.

TABLE 5

Frequencies with which the observed contrasts between the average excess of haplotype *010* with the average excesses of the other haplotypes exceeded in magnitude the random contrasts generated by 1000 permutations of the phenotype labels with respect to the haplotype numbers for all six traits

	Traits					
Contrast of *010 vs.*:	Trig.	Chol.	HDL	Ln(T)	Ln(C)	Ln(H)
011	0.128	0.814	0.465	0.034*	0.935	0.698
000	0.604	0.374	0.717	0.943	0.471	0.938
110	0.465	0.751	0.584	0.475	0.835	0.488
001	0.753	0.725	0.979	0.835	0.856	0.917

Asterisk indicates contrast that is significant at the 5% level.

minants of lipid levels in humans, we adjusted each trait for age and age-squared separately for males and females and then removed the mean difference between sexes before performing the genetic analysis (BOERWINKLE *et al.* 1987). Table 4 gives the average excesses for the haplotypes and phenotypes as estimated from Equation 1. Next, the significance of the contrasts between the average excess of haplotype *010 vs.* those of the other haplotypes is determined by generating 1000 random permutations of the data. Table 5 gives the significance levels of the contrasts as determined by computer simulation.

As can be seen from Table 5, a significant phenotypic contrast is detected for the phenotype of ln(triglyceride) and it is localized to the transition between haplotypes *010* and *011* (which is associated with the *Sst*I site). If the phenotypically important mutation is the same as the *Sst*I site mutation, this phenotypic association should be a clean one. However, if the two mutations are not identical, it is possible for either the haplotype *010* or the haplotype *011* category to be phenotypically heterogeneous; that is, some of the chromosomes in the haplotype *010* (or *011*) category could bear the phenotypically important mutation, and others not. Consequently, the next stage of the analysis is to look for heterogeneity with respect to ln(T) in the haplotype *010* and *011* classes. Unfortunately, the St. Mary's data set is inadequate to detect heterogeneity within the haplotype *011* category. Only 11 individuals unambiguously have this haplotype, and 10 of them are *010/011* heterozygotes. Phenotypic heterogeneity within this genotypic class could be due to either heterogeneity in the *010* or *011* haplotype categories, thereby making it impossible to conclude that any detected heterogeneity would be caused by heterogeneity within haplotype *011*. Fortunately, there are 45 individuals that are *010/010* homozygotes. Hence, we have a reasonable, albeit small, sample size for detecting heterogeneity within the haplotype *010* class.

Figure 5 gives the histogram of the ln(T) phenotypes of the *010/010* homozygotes. The figure also gives the maximum likelihood partitioning of the distribution into two heterogeneous phenotypic classes as estimated by the procedure of ENGLEMAN and HARTIGAN (1969). However, this partitioning is not significant at the 5% level, using the tables given in ENGLEMAN and HARTIGAN (1969). Hence, we fail to reject the null hypothesis that haplotype *010* has the same phenotypic effects in all individuals homozygous for that haplotype.

DISCUSSION

TEMPLETON, BOERWINKLE and SING (1987) have discussed the merits and limitations of the cladistic analysis of phenotypic associations with restriction site polymorphisms. Hence, this presentation will be limited to the features of the analytical method that are novel to applications to data collected from outbred populations.

We have shown how the quantitative genetic concept of average excess can be generalized to a nested average excess measure. We then use this nested average excess measure and associated statistics to extend the nested cladistic analysis of TEMPLETON, BOERWINKLE and SING (1987) to populations that are not genetically manipulated to consist solely of homozygotes. Consequently, this type of analysis can be applied to outbred as well as experimental populations. The Adh example supports the statistical validity of the permutational null distribution by yielding virtually identical statistical inferences in a case for which both NANOVA and permutation testing is possible. The validity of our analysis of the St. Mary's data is also supported by other reports of an association between the *Sst*I site and the categorical phenotype of hypertriglyceridaemia (REES *et al.* 1983, 1985).

The fact that we could detect an association between the *Sst*I site and elevated triglyceride levels in such a small sample illustrates another strength of our approach—enhanced statistical power when dealing with outbred populations. One motivation for devel-

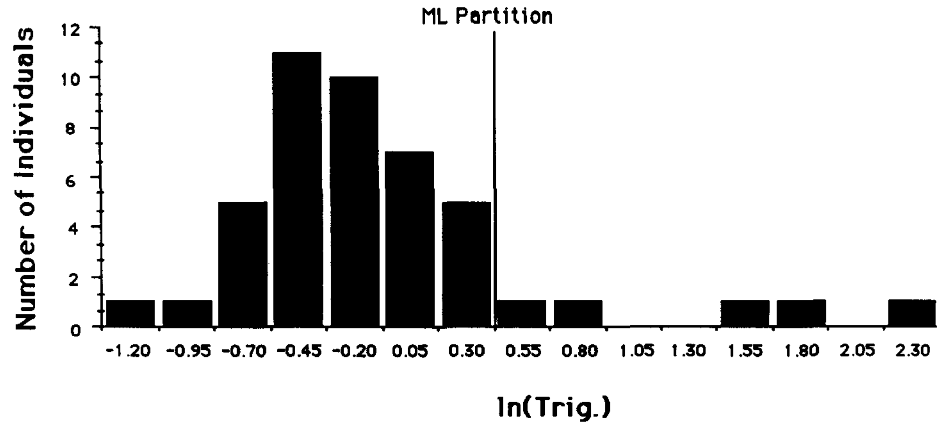

FIGURE 5.—Histogram of the logarithm of triglyceride levels (adjusted for age and sex) in the *010/010* homozygotes of the St. Mary's group. A line indicates the maximum likelihood partition of this population into two clusters, as determined by the procedure of ENGLEMAN and HARTIGAN (1969).

TABLE 6

Analyses of variances of the lipid traits using genotypes as the treatment effects

Source	Sum of squares	Degrees of freedom	Mean square	F-statistic	Significance
Trig.					
Treatment	341.89	9	37.99	1.38	0.21
Error	2174.3	79	27.52		
Total	2516.1	88			
Chol.					
Treatment	24.56	9	2.72	0.62	0.78
Error	347.15	79	4.39		
Total	371.61	88			
HDL					
Treatment	1.72	9	0.19	1.09	0.38
Error	13.28	76	0.17		
Total	15.00	85			
Ln(T)					
Treatment	9.15	9	1.02	1.49	0.17
Error	53.86	79	0.68		
Total	63.01	88			
Ln(C)					
Treatment	0.31	9	0.03	0.45	0.90
Error	6.02	79	0.07		
Total	6.33	88			
Ln(H)					
Treatment	1.27	9	0.14	1.17	0.33
Error	9.20	76	0.12		
Total	10.47	85			

Trait abbreviations are as given in Table 4. All traits have been adjusted for age and sex.

oping the permutational analysis was the belief that genetic effects on phenotypic variability can be more efficiently detected by looking directly at haplotype effects as opposed to genotype effects whenever there are several haplotypes. This enhanced statistical power can be illustrated by performing standard analyses of variances on the St. Mary's patient group for the six phenotypic measures. The 10 genotypic categories given in Table 3 will be the treatment effects. Table 6 gives the results of these ANOVAs. In contrast to the results given in Table 5, no significant effects of genotypic variability are detectable for any of the traits. Hence, the haplotype analysis of this population was more sensitive in detecting genetic effects than the traditional analysis which contrasts genotypic values.

However, it should be noted that the enhanced power of a haplotype analysis should exist only when the phenotypic values of the genotypic classes are close to their additive values. Interaction effects (such as recessiveness for rare haplotypes, epistasis, etc.) can reduce the predictability of a genotype's phenotypic value from its component haplotypes. Hence, our procedure is primarily limited to the analysis of a single DNA region with little internal recombination in which the genetic variance of the phenotypes of interest is mostly additive.

In summary, we have presented a strategy for performing a cladistic analysis of associations between haplotypes and phenotypic variation in samples from both genetically homozygous and outbred populations. The use of the average excess measure and permutational testing when there is heterozygosity greatly broadens the range of applications of this methodology. The worked examples presented in this paper illustrate that permutational testing may be less powerful than a standard nested analysis of variance when dealing with genetically homozygous populations, but it is more powerful than a standard analysis of variance using genotypes as treatments when dealing with an outbred population that is polymorphic for several haplotypes that are primarily additive in their phenotypic effects. Therefore, coupling permutation testing with a cladistic design can be a valuable tool in searching for associations between quantitative phenotypic variations and polymorphic restriction site markers in samples from natural populations.

This work was supported by National Institutes of Health Grants 1 R01 HL39107 (to C.F.S.) and R01 AG02246 (to A.R.T.). We wish to thank Drs. Bruce Weir and Francois Bonhomme and two anonymous reviewers for their excellent suggestions for improving an earlier version of this work.

LITERATURE CITED

AQUADRO, C. F., S. F. DEESE, M. M. BLAND, C. H. LANGLEY and C. LAURIE-AHLBERG, 1986 Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster.* Genetics **114:** 1165–1190.

Boerwinkle, E., S. Visvikis, D. Welsh, J. Steinmetz, S. M. Hanash and C. F. Sing, 1987 The use of measured genotype information in the analysis of quantitative phenotypes in man. II. The role of the apolipoprotein E polymorphism in determining levels, variability and covariability of cholesterol, betalipoprotein and triglycerides in a sample of unrelated individuals. Am. J. Med. Genet. **27:** 567–582.

Edgington, E. W., 1987 *Randomization Tests,* Ed. 2. Marcel Dekker, New York.

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

Engleman, L., and J. A. Hartigan, 1969 Percentage points of a test for clusters. J. Am. Stat. Assoc. **64:** 1647–1648.

Fisher, R. A., 1918 The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb. **52:** 399–433.

Hill, W. G. 1974 Estimation of linkage disequilibrium in randomly mating populations. Heredity **33:** 229–239.

Kessling, A. M., B. Horsthemke and S. E. Humphries, 1985 A study of DNA polymorphisms around the human apolipoprotein AI gene in hyperlipidaemic and normal individuals. Clin. Genet. **28:** 296–306.

Rees, A., J. Stocks, C. C. Shoulders, D. J. Galton and F. E. Baralle, 1983 DNA polymorphism adjacent to human apoprotein AI gene: relation to hypertriglyceridaemia. Lancet **1:** 444–446.

Rees, A., J. Stocks, C. R. Sharpe, M. A. Vella, C. C. Shoulders, J. Katz, N. I. Jowett, F. E. Baralle and D. J. Galton, 1985 Deoxyribonucleic acid polymorphism in the apolipoprotein A-I-C-III gene cluster. Association with hypertriglyceridemia. J. Clin. Invest. **76:** 1090–1095.

Savageau, M., 1976 *Biochemical Systems Analysis.* Addison-Wesley, Reading, Mass.

Templeton, A. R., 1983 Convergent evolution and nonparametric inferences from restriction data and DNA sequences. pp. 151–179. In: *Statistical Analysis of DNA Sequence Data,* Edited by B. S. Weir. Marcel Dekker, New York.

Templeton, A. R., 1987 The general relationship between average effect and average excess. Genet. Res. **49:** 69–70.

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.

Communicating editor: B. S. Weir