A Cladistic Analysis of Phenotypic Associations with Haplotypes Inferred from Restriction Endonuclease Mapping or DNA Sequencing. V. Analysis of Case/Control Sampling Designs: Alzheimer's Disease and the Apoprotein E Locus

Alan R. Templeton

Department of Biology, Washington University, St. Louis, Missouri 63130

Manuscript received September 7, 1994
Accepted for publication February 4, 1995

ABSTRACT

Present-day associations between haplotypes at a candidate locus and phenotypes exist when phenotypically important mutations occurred at some point during the evolution of the current array of genetic variation. A cladistic statistical design can be defined that focuses power by using the evolutionary history of the candidate DNA region. This paper shows how cladistic methodology is used for the analysis of case/control data, a common sampling design in genetic/disease association studies. A worked example is presented of the associations for sporadic early and late-onset forms of Alzheimer's disease with the *19q13.2* chromosomal region that includes the loci for apoproteins E, CI, and CII. This analysis confirms earlier reports of a strong association of the *ApoE* $\epsilon 4$ allele with Alzheimer's disease but indicates that it is premature to condsider this association causal, particularly for early onset cases. Associations were also found with the $\epsilon 2$ allele, as previously reported, and with the *1* allele at the *ApoCI* locus. However, this analysis indicates that it is inappropriate both statistically and medically to use single markers as risk predictors when haplotype data are available, even when the mutation leading to the marker is identified as having a strong phenotypic association.

GENETIC studies of quantitative traits have traditionally utilized phenotypic correlations between related and unrelated individuals to estimate the fraction of the interindividual variance in the population that is attributable to unmeasured genotypic differences. Recent advances in molecular genetics are making it possible to locate and characterize the loci that determine this genetic component of variance. If the trait of interest is determined by biochemical or physiological functions under the control of identified genes (candidate genes), then the population can be screened for restriction fragment length polymorphism (RFLP) or sequence variability in and/or near a candidate gene to define haplotype variation. One then analyzes the associations between haplotype variation at the candidate gene with phenotypic variation in the trait. Analyzing haplotypes instead of performing multiple individual analyses on each variable site in the candidate region avoids the problems of linkage disequilibrium and statistical nonindependence across variable sites. However, there is often much haplotype variation, so an efficient statistical design is required to implement this haplotype approach. A mutation with significant phenotypic effects arises at some point in evolution, and consequently this mutation should be imbedded in the evolutionary tree of genetic variation at this locus (*i.e.*, a gene tree) as long as recombination has been sufficiently rare. Accordingly, some branch or "clade" of this gene tree (cladogram) should display a similar phenotypic association. Therefore, by reconstructing the evolutionary tree of the genetic variation at the candidate region, an optimal statistical strategy is to look for associations between the phenotypes of interest with larger and larger branches of the gene tree.

The basic implementation of this cladistic approach was described in TEMPLETON *et al.* (1987) and was restricted to the analysis of quantitative phenotypes in homozygous or haploid stocks. Since then, the cladistic procedure has been extended and amplified to include diploid populations (TEMPLETON *et al.* 1988), to deal with the complexities of estimating the cladograms and quantifying their ambiguity (TEMPLETON *et al.* 1992; TEMPLETON and SING 1993), to deal with and utilize recombination events to obtain a rough physical localization of the mutations causing significant phenotypic effects (TEMPLETON *et al.* 1992; TEMPLETON and SING 1993), and to analyze categorical data (TEMPLETON and SING 1993). The purpose of this paper is to extend the cladistic approach to the analysis of data sampled with a case/control design.

The case/control design is common in many clinical and disease association studies. With this design, a population of individuals with the phenotype of interest is first identified (the "case" group). A nondiseased control group is then sampled, with an attempt being made to match the control sample with the case sample

Author e-mail: templeton@wustlb.wustl.edu

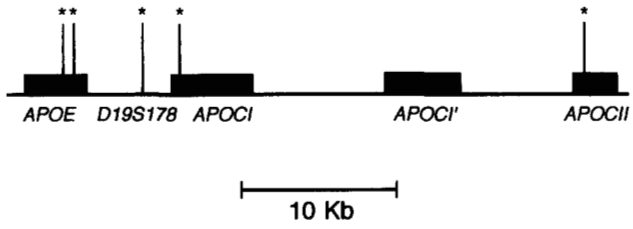

FIGURE 1.—Chromosome *19q13.2* genomic map. Lines with an asterisk show the approximate location of the markers used in this study. The locations of the following loci are also indicated on this map: *APOE*, apoprotein E locus; *APOCI*, apoprotein CI locus; *APOCI'*, apoprotein CI pseudogene; and *APOCII*, apoprotein CII locus. The position of the *D19S178* marker is not accurately known, so the indicated position is only a rough approximation.

for one or more variables that are also associated with the phenotype (*e.g.*, age, sex, smoking status, etc.). Associations are detected by looking for differences between the cases and the controls. The application of cladistics to this sampling design will be given through a worked example, using the data of CHARTIER-HARLIN *et al.* (1994) on haplotype variation in the *19q13.2* chromosomal region of the human genome with early and late-onset forms of Alzheimer's disease (AD).

MATERIALS AND METHODS

Populations: There are two independent AD populations. The first is a group of 36, late-onset sporadic cases from France with a mean age of onset of 78 ± 9 years (mean ± SD). The corresponding controls were 38 individuals with a matched mean age (80 ± 8 years) and shared environmental variables. The second case group was 34 individuals that were diagnosed as early-onset sporadic AD cases at St. Mary's Hospital in London, U.K. The mean age of onset was 57 ± 5 years, and the corresponding control group was a group of 36 individuals from the U.K. with a mean age of 55 ± 7.5 years. For further details, see CHARTIER-HARLIN *et al.* (1994).

Haplotypes: CHARTIER-HARLIN *et al.* (1994) scored genomic DNA at six marker locations in the *19q13.2* chromosomal region, but they used only five for haplotype reconstruction: an anonymous (CA)n repeat DNA marker (*D19S78*) that is most likely between the *APOE* and *APOCI* loci; three alleles at the *APOE* locus, which in turn are determined by nucleotide variation at two sites in the coding region of this locus (amino acid positions 112 and 158); a *Hpa*I RFLP in the 5′ end of the *APOCI* locus; and a (CA)n repeat polymorphism in the *APOCII* locus. The genetic variability for the (CA)n repeats was scored categorically into *S* (short) and *L* (long) alleles. Figure 1 shows a map indicating the positions of these markers in the candidate DNA region. Fifteen haplotypes were defined by these five markers, as indicated in Table 3 of CHARTIER-HARLIN *et al.* (1994). The frequencies of these haplotypes in both case and both control populations is also indicated in Table 3 of CHARTIER-HARLIN *et al.* (1994). These frequencies were multiplied by the number of chromosomes scored in each sample and rounded to the nearest integer to convert them into haplotype numbers.

Cladogram and nested design: The haplotypes were used to estimate a 95% plausible set of cladograms using the algorithm of TEMPLETON *et al.* (1992). The cladogram was then converted into a nested statistical design using the rules given in TEMPLETON *et al.* (1987) and TEMPLETON and SING (1993). The nested design groups together evolutionarily closely related haplotypes ("0-step" clades) into "one-step" clades, groups of evolutionarily closely related one-step clades into two-step clades, *etc.*

Nested analysis: Associations between haplotypes and AD were investigated by performing a series of nested two (cases and controls) × $n(i)$ contingency analyses, where $n(i)$ is the number of clades in nesting category i. The contingency table consists of the number of times a particular clade is found in the control and case populations. A series of nested contingency tests are asymptotically independent of one another (PRUM *et al.* 1990), and different contingency tests at the same clade step level are independent because they utilize nonoverlapping subsets of the data. Because of small sample size, an exact permutational chi-square test was performed using the algorithm of ROFF and BENTZEN (1989) with 1000 random permutations to achieve accurate 5% statistical inference (EDGINGTON 1986).

When a significant association is detected within a nesting category that has multiple evolutionary transitions within it, a series of 2 × 2 contingency tests are performed to localize the evolutionary position of the significant phenotypic change. The 2 × 2 comparisons are chosen to reflect contrasts only between the pairs of clades that are most closely related evolutionarily within the nesting category. These 2 × 2 tests correspond to the multiple comparisons procedure used in standard nested analyses of variance (TEMPLETON *et al.* 1987), and like these standard multiple comparisons, the significance levels of these contingency tests are adjusted by the Bonferroni procedure where the number of comparisons is known *a priori* (TEMPLETON *et al.* 1987).

Three different nested contingency analyses are performed: one on the late-onset French samples, a second on the early-onset British samples, and the third on all-onsets (*i.e.*, the cases from the United Kingdom and France were pooled and so were the controls).

RESULTS

Cladogram and nested design: In applying the algorithm of TEMPLETON *et al.* (1992), it was discovered that there was little to no association between the *S* and *L* alleles at the *APOCII* locus with the other four variable sites. Of the 15 haplotypes described by CHARTIER-HARLIN et al. (1994), 14 of them define seven haplotype pairs that are identical in their genetic state for the other five markers but with one of the pair having the *APOCII S* allele and the other the *L* allele. There are two likely explanations for this state. First, recombination could be common in the region separating the *APOCII* locus from the other four markers. As seen in Figure 1, the other four markers are in close proximity to one another (all within a piece of DNA 10 kb) whereas the *APOCII* locus is ~27 kb from the closest of these other four markers. Hence, the recombinational hypothesis is highly plausible. The second explanation is homoplasy; that is, multiple independent evolutionary events leading to the same allelic state (in this case at least six such events). This explanation is also highly plausible because the variation in the (CA)n repeats at this locus were scored only coarsely into short and long categories. Hence, different *S* and *L* alleles may not truly be the same allele by descent. Moreover, repeat regions

TABLE 1

Haplotypes and their numbers of occurrences in the French, British, and total case and control samples

Haplotype[a]	Marker genotypes			Late-onset cases	Late-onset controls	Early-onset cases	Early-onset controls	All-onset cases	All-onset controls
	APOE	D19S178	APOCI						
1 & 3	4	S	2	6	0	7	2	13	2
2 & 4	4	L	2	12	4	13	8	25	12
5 & 6	3	S	1	23	21	21	20	44	41
7 & 8	3	S	2	1	1	9	0	10	11
9 & 10	3	L	1	28	39	17	32	45	71
11	3	L	2	1	0	0	0	1	0
12 & 13	2	L	2	0	4	0	3	0	7
14 & 15	2	S	2	1	7	1	7	2	14

[a] The haplotype numbers from Table 3 of CHARTIER-HARLIN *et al.* (1994) are retained, but as the *APOCII* marker is not used in the present analysis, seven of the eight haplotypes are indicated by a pair of the original haplotype numbers.

tend to be highly polymorphic precisely because they are subject to frequent changes in repeat number (QUELLER *et al.* 1993), thereby greatly increasing the chances for homoplasy.

Regardless of which explanation is true, the algorithm of TEMPLETON *et al.* (1992) results in the recommendation of splitting this DNA region into two subsets; one consisting solely of the *APOCII* marker, and the other the 10-kb region straddling the *APOE* and *APOCI* loci. As CHARTIER-HARLIN *et al.* (1994) have already presented a single marker analysis of the *APOCII* marker, no further use of this marker will be made in this paper. The cladistic analysis will be applied only to the remaining four markers. These four markers define eight haplotypes. These haplotypes, as well as the number of observations of each haplotype in the sampled populations, are given in Table 1.

Figure 2 shows the 95% plausible set of cladograms for these eight haplotypes as estimated by the algorithm of TEMPLETON *et al.* (1992). Note that there are three adjoining loops of ambiguity in this plausible set. There are 486 different ways of breaking this triple loop, so there are 486 cladograms in the plausible set. These loops are created by potential homoplasy either for the evolution of the $\epsilon 4$ allele at the *APOE* locus, homoplasy for the *Hpa*I restriction site at the *APOCI* locus, or homoplasy for the *S* and *L* alleles at the *D19S178* marker. As mentioned above, homoplasy is highly likely for (CA)n repeat variants that are only coarsely scored. Accordingly, the homoplasy will be ascribed to the (CA)n repeats, with the resulting cladogram shown in Figure 3. Figure 3 also gives the nested design using the nesting rules given in TEMPLETON *et al.* (1987).

Nested contingency analysis: Using the nested design given in Figure 3, nested contingency analyses were performed upon the data given in Table 1. The results of the late-onset AD analysis are given in Table 2, the early-onset analysis in Table 3, and the all-onsets analysis in Table 4. In all three analyses, the only nesting category with a significant permutational chi-square is the entire cladogram with four one-step clades nested within it. As can be seen from Figure 3, three evolutionary transitions occur in this nesting category; the three mutational changes that interconnect the evolutionarily central one-step clade 1–2 to the three peripheral one-step clades, clades 1–1, 1–3 and 1–4. Hence, to localize the significant associations with AD observed in this nesting category, three additional contingency tests were performed; cases *vs.* controls crossed with 1–2 *vs.* 1–1; 1–2 *vs.* 1–3 and 1–2 *vs.* 1–4. Table 5 shows the results of these additional permutational chi-square analyses. The evolutionary transitions that are associated with these significant phenotypic changes as identified by the comparisons given in Table 5 are indicated by asterisks in Figure 3.

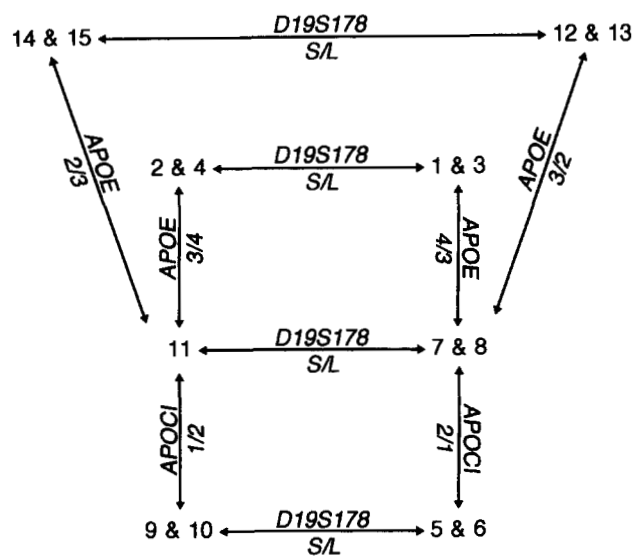

FIGURE 2.—The cladogram set estimated by the TEMPLETON *et al.* (1992) algorithm as derived from the eight haplotypes given in Table 1. Each arrow indicates one mutational event. The description of the event is indicated by the arrow, using the allelic notation given in CHARTIER-HARLIN *et al.* (1994). Because the network is unrooted, each arrow is double headed, and the type of change as a function of evolutionary direction is indicated by the symbol closest to the arrowhead that defines the evolutionary direction.

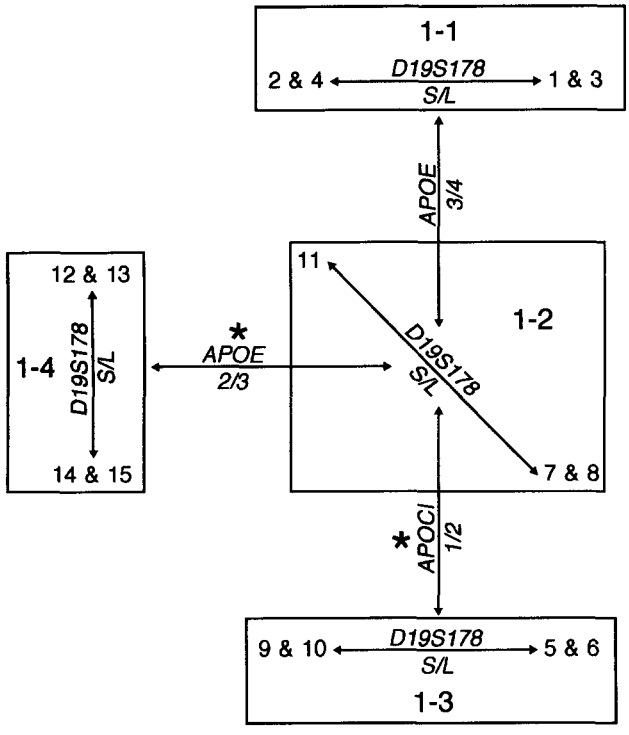

FIGURE 3.—The cladogram set estimated by the TEMPLETON *et al.* (1992) algorithm as derived from the eight haplotypes given in Table 1 when homoplasy is ascribed to the length variations of the *D19S178* (CA)n repeat. Arrows indicating mutations at other markers are connected to the middle of the arrows indicating *S/L* transitions because homoplasy prevents a more accurate localization. Boxes enclose haplotypes that are nested together to form one-step clades, as designated by the notation *1-#*. The next level of nesting would be the entire cladogram, so there is only one two-step clade that includes these four one-step clades within it. The significant phenotypic localizations for both early-onset and all-onsets from Table 5 are indicated by asterisks.

DISCUSSION

As with previous analyses of AD, the cladistic analysis has detected significant associations between genetic variation at the *APOE* locus and nearby markers with Alzheimer's disease. Significant genetic heterogeneity among one-step clades of haplotypes is observed for late, early and all-onsets. There was no statistically significant localization of this association for the late-onset data set (Table 5). This lack of statistical resolution can be attributed to the small sample size of clade 1–2, the pivotal central clade in the cladogram that is involved in all the localization contrasts. Only three observations fall into this clade for the late-onset data set, thereby precluding any statistically significant localization. Hence, for late-onset cases, the most appropriate conclusion is that significant genetic heterogeneity for AD exists, but the current sample sizes are too small to pinpoint the genetic source of this association.

For the early and all-onset data sets, localization of the genetic effect within the cladogram was possible (Table 5). In both cases, two significant phenotypic transitions were detected, subdividing the cladogram into three phenotypically distinct subsections. The first subsection consists of all haplotypes bearing the $\epsilon 4$ allele at the *APOE* locus combined with those haplotypes with the $\epsilon 3$ allele and the simultaneous presence of the *Hpa*I restriction site at the *APOCI* locus (haplotypes 1–4, 7, 8, and 11). This subset of haplotypes is associated with a high incidence of AD (*e.g.*, for early onset AD, 29 cases *vs.* 10 controls fall into this haplotype set). The second subset consists of haplotypes with the $\epsilon 3$ allele and the absence of the *Hpa*I restriction site at the *APOCI* locus (haplotypes 5, 6, 9, and 10). This subset of haplotypes is associated with a medium incidence of AD (*e.g*, for early onset AD, 38 cases and 52 controls fall into this haplotype set). The final subset consists of those haplotypes bearing the $\epsilon 2$ allele (haplotypes 12–15), and this subset is associated with a low incidence of AD (*e.g.*, for early onset AD, one case and 10 controls fall into this haplotype set).

TABLE 2

Nested exact contingency analysis of the late onset case/control data described in CHARTIER-HARLIN *et al.* (1994)

Source	Chi-square statistic	Probability
Zero-step clades		
Within 1-1	1.8333	0.30
Within 1-2	0.7500	1.00
Within 1-3	1.1749	0.33
Within 1-4	0.5455	1.00
One-step clades		
Within 2 (entire cladogram)	18.2071	0.00

The nested design is given in Figure 3. A standard contingency chi-square statistic is calculated, and its exact significance is determined by 1000 random permutations that preserve the marginal values. The probability column refers to the frequency with which these randomly generated chi-square statistics were equal to or greater than the observed chi-square.

CHARTIER-HARLIN *et al.* (1994) performed multiple single marker marginal analyses and found significant associations with the *D19S78*, *APOE*, and *APOCI* markers. However, multiple marginal analyses of single markers found within a small DNA region characterized by much linkage disequilibrium are statistically inappropriate because of a lack of marginal independence. The cladistic analysis, which circumvents this problem, does not support the inference of CHARTIER-HARLIN *et al.* (1994) of a significant association between AD and the *D19S78* marker (which characterizes the contrasts among haplotypes nested within one-step clades, all of which were nonsignificant as can be seen in Tables 2–4). Hence, this marginal association is most likely due to disequilibrium with the other markers in this DNA gene region, markers which are stronger and better predictors of AD risk.

The cladistic analysis does support the conclusion of

TABLE 3

Nested exact contingency analysis of the early onset case/control data described in CHARTIER-HARLIN *et al.* (1994)

Source	Chi-square statistic	Probability
Zero-step clades		
Within 1-1	0.7143	0.66
Within 1-3	2.4989	0.12
Within 1-4	0.4125	1.00
One-step clades		
Within 2 (entire cladogram)	21.7782	0.00

The nested design is given in Figure 3. A standard contingency chi-square statistic is calculated, and its exact significance is determined by 1000 random permutations that preserve the marginal values. The probability column refers to the frequency with which these randomly generated chi-square statistics were equal to or greater than the observed chi-square.

the *APOCI* restriction site marker being associated with AD risk. From Figure 3, it can be seen the evolutionary transition between presence and absence of the *Hpa*I restriction site is associated with a phenotypic transition from high to medium AD risk. However, the cladistic analysis indicates that it would still be inappropriate to use the presence and absence of the *Hpa*I restriction site as an indicator of AD risk. Note from Figure 3 that individuals that have the *Hpa*I restriction site are heterogeneous with respect to AD risk. In particular, this genetic class contains those individuals with the highest and with the lowest AD risk. Using this site as a marginal risk predictor of AD would lump these individuals together into a single risk category. Hence, the cladistic analysis clearly demonstrates that even when a particular marker displays a strong phenotypic association, it is still statistically and medically inappropriate to use that marker by itself as a disease risk predictor when haplotype data are available.

TABLE 4

Nested exact contingency analysis of all onset case/control data described in CHARTIER-HARLIN *et al.* (1994)

Source	Chi-square statistic	Probability
Zero-step clades		
Within 1-1	1.9788	0.19
Within 1-2	0.0992	1.00
Within 1-3	3.3454	0.08
Within 1-4	0.9583	0.42
One-step clades		
Within 2 (entire cladogram)	37.5445	0.00

The nested design is given in Figure 3. A standard contingency chi-square statistic is calculated, and its exact significance is determined by 1000 random permutations that preserve the marginal values. The probability column refers to the frequency with which these randomly generated chi-square statistics were equal to or greater than the observed chi-square.

TABLE 5

Localization of the significant phenotypic changes in the cladogram shown in Figure 3 via the evolutionarily relevant pairwise comparisons among one-step clades

Comparison	Chi-square statistic	Probability
Late onset data		
1-2 *vs.* 1-1	0.3788	1.000
1-2 *vs.* 1-3	0.5041	0.570
1-2 *vs.* 1-4	5.1047	0.074
Early onset data		
1-2 *vs.* 1-1	4.0345	0.124
1-2 *vs.* 1-3	10.9532	0.001*
1-2 *vs.* 1-4	16.3636	0.000*
All onset data		
1-2 *vs.* 1-1	1.8777	0.260
1-2 *vs.* 1-3	10.2098	0.003*
1-2 *vs.* 1-4	23.2522	0.000*

* Significant at the 5% level after Bonferroni adjustment. Test results that have a probability less than 0.017 are significant at the 5% level after adjustment for multiple comparisons by the Bonferroni procedure.

The inappropriateness of marginal associations is further indicated by the significant phenotypic associations found withe the *APOE* alleles. The cladistic analysis confirms the inference of CHARTIER-HARLIN *et al.* (1994) that the $\epsilon 2$ allele is associated with a low incidence of AD, an association found in other studies as well for late-onset cases (CORDER *et al.* 1994; WEST *et al.* 1994). However, the cladistic analysis does not indicate any significant phenotypic change associated with the evolutionary transition leading to the $\epsilon 4$ allele. This is particularly interesting because it was reports of a strong marginal association between the $\epsilon 4$ allele and AD (CORDER *et al.* 1993; VANDUIJN *et al.* 1994; YU *et al.* 1994) that motivated the study of CHARTIER-HARLIN et al. (1994), and one of their principal inferences was to confirm this strong marginal association and to conclude that $\epsilon 4$ is a major risk factor for AD. However, the cladistic analysis is not incompatible with this inference of a strong marginal association with the $\epsilon 4$ allele. The $\epsilon 4$ allele may indeed be associated with increased risk for AD, but the current analysis lacked sufficient statistical power to detect this association. Although the test of association was an exact test, and therefore has maximum power for tests of association in such cases, no test can circumvent the low power imposed by small sample sizes, as is the case here. Nevertheless, it is interesting to note that significant associations were detected with the other two contrasts involving the central clade 1–2, including the contrast of clade 1–2 with clade 1–4, the clade consisting of haplotypes bearing the $\epsilon 2$ allele. The sample sizes for clade 1–4 are 11 for the early onset AD sample (cases and controls) and 23 for the all onsets sample, which are considerably smaller than the respective sample sizes of 30 and 52 associated with clade 1–1 (the clade defined by the $\epsilon 4$ allele).

Hence, if the $\epsilon 4$ allele does have an effect on early onset or all onsets AD risk, it is most likely a minor one and much smaller than that associated with the $\epsilon 2$ allele. However, even if the $\epsilon 4$ allele has only a weak or no effect at all upon AD risk, the cladistic analysis still predicts that the $\epsilon 4$ allele will have a strong marginal association with AD risk. Note from Figure 3 that the high risk phenotypic category consists of two one-step clades, 1–1 (which has the $\epsilon 4$ allele) and 1–2 (which has the $\epsilon 3$ allele). However, clade 1–2 is a rare clade, so almost all members of the high risk group bear the $\epsilon 4$ allele, and almost all individuals with the $\epsilon 3$ allele have medium risk (clade 1–3). Therefore, a strong marginal association of AD with the $\epsilon 4$ allele is expected even though the cladistic analysis indicates that the evolutionary transition between the $\epsilon 3$ and $\epsilon 4$ alleles is *not* associated with any significant alteration of AD risk in early onset and all onsets cases. Once again, this illustrates the dangers and inappropriateness of using single marker marginal associations as risk predictors when haplotype data are available.

The failure of the evolutionary transition between the $\epsilon 3$ and $\epsilon 4$ alleles to be associated with any alteration of AD risk is patently relevant to speculations about the causal nature of the marginal association between $\epsilon 4$ and AD. CHARTIER-HARLIN *et al.* (1994) appropriately point out that this marginal association could be due either to linkage disequilibrium with a causal mutation in this DNA region, or the $\epsilon 4$ allele itself is the causal mutation. However, when a mutation is the cause of a phenotypic transition and when that mutation itself is used as a marker in the construction of the cladogram, the cladistic analysis should localize the phenotypic transition to that marker. If the phenotypic transition is localized to a second marker, then the first marker cannot be causative. For example, in the first application of cladistics to phenotypic/candidate gene associations, activity of the enzyme Alcohol dehydrogenase (Adh) in *Drosophila melanogaster* was examined for associations with haplotype variation at the *Adh* locus (TEMPLETON *et al.* 1987). A major difference in Adh activity had long been associated with animo acid coding changes leading to the *S* ("slow") and *F* ("fast") isozyme alleles at this locus. This *S/F* transition was one of the markers used in constructing the cladogram, but the cladistic analysis indicated that the major Adh activity differences were not associated with this transition, but rather with an adjacent evolutionary transition. Once again, this would lead to a strong marginal association of Adh activity with the *S* and *F* alleles, but the cladistic analysis lead to the prediction that the amino acid coding changes were not the cause of the activity difference (TEMPLETON *et al.* 1987)—a prediction that was later confirmed by more extensive molecular analyses (LAURIE *et al.* 1991). In the present case, the cladistic analysis leads to the prediction that the $\epsilon 4$ allele is *not* the cause of increased AD risk, at least for the early onset form. Recall that the sample sizes were inadequate to localize the effects to specific evolutionary transitions for the late-onset form, so the associations of $\epsilon 4$ with late-onset AD risk have yet to be determined. Thus, the present analysis does not eliminate as a possibility the proposed causal mechanisms relating the $\epsilon 4$ allele to late-onset AD (*e.g.*, STRITTMATTER *et al.* 1994). However, the lack of statistical resolution of the evolutionary step associated with increased risk to late-onset AD and the fact that the $\epsilon 4$-AD association is not universally found in all populations (LANNFELT *et al.* 1994) does caution against a premature conclusion of cause and effect.

It is important to point out that although the cladistic analysis can be used to eliminate markers as putative causal agents, the evolutionary localization of a phenotypic effect to a particular mutational transition does not mean that that transition is necessarily causative (TEMPLETON *et al.* 1987). For example, the low risk for AD that the cladistic analysis revealed for the *e2* allele should not be interpreted as evidence that the $\epsilon 2$ allele causes lowered AD risk. Rather, the cladistic analysis only indicates that the $\epsilon 2$ allele remains as a potential candidate for the causal mutation, but by itself does not discriminate between cause and association through disequilibrium with yet another, unscored mutation in this DNA region. This situation can be clarified by increasing the genetic resolution of this DNA region to increase the cladogram resolution. Thus, this initial cladistic analysis indicates that the *19q13.2* chromosomal region contains mutations that both increase and decrease the risk of AD, but further progress on understanding the nature of these disease associations will require larger sample sizes and increased genetic resolution within this DNA region.

I thank Drs. CHARLES F. SING and MARTHA HAVILAND and two anonymous reviewers for their comments on an earlier draft of this manuscript. This work was supported by National Institute of Heart, Lung, and Blood I grant 1 R01 HL-39107.

LITERATURE CITED

CHARTIER-HARLIN, M., M. PARFITT, S. LEGRAIN, J. PÉREZ-TUR, T. BROUSSEAU *et al.*, 1994 Apolipoprotein E, $\epsilon 4$ allele as a major risk factor for sporadic early and late-onset forms of Alzheimer's disease: analysis of the 19q13.2 chromosomal region. Hum. Mol. Genet. **3:** 569–574.

CORDER, E. H., A. M. SAUNDERS, W. J. STRITTMATTER, D. E. SCHMECHEL, P. C. GASKELL *et al.*, 1993 Gene dose of Apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science **261:** 921–923.

CORDER, E. H., A. M. SAUNDERS, N. J. RISCH, W. J. STRITTMATTER, D. E. SCHMECHEL *et al.*, 1994 Protective effect of Apolipoprotein E type 2 allele for late onset Alzheimer disease. Nat. Genet. **7:** 180–184.

EDGINGTON, E. S., 1986 *Randomization Tests*, Ed. 2. Marcel Dekker, New York.

LANNFELT, L., L. LILIUS, M. NASTASE, M. VIITANEN, L. FRATIGLIONI *et al.*, 1994 Lack of association between apolipoprotein-E allele epsilon-4 and sporadic Alzheimers disease. Neurosci. Lett. **169:** 175–178.

LAURIE, C. C., J. T. BRIDGHAM, and M. CHOUDHARY, 1991 Associa-

tions between DNA sequence variation and variation in expression of the Adh gene in natural populations of *Drosophila melanogaster*. Genetics **129:** 489–499.

PRUM, B., M. GUILLOUD-BATAILLE, and F. CLERGET-DARPOUX, 1990 On the use of c^2 tests for nested categoried data. Ann. Hum. Genet. **54:** 315–320.

QUELLER, D. C., J. E. STRASSMANN, and C. R. HUGHES, 1993 Microsatellites and kinship. Trends Evol. Ecol. **8:** 285–288.

ROFF, D. A., and P. BENTZEN, 1989 The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. Mol. Biol. Evol. **6:** 539–545.

STRITTMATTER, W. J., K. H. WEISGRABER, M. GOEDERT, A. M. SAUNDERS, D. HUANG *et al.*, 1994 Hypothesis—microtubule instability and paired helical filament formation in the Alzheimer disease brain are related to apolipoprotein E genotype. Exp. Neurol. **125:** 163–171.

TEMPLETON, A. R., and C. F. SING, 1993 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. Genetics **134:** 659–669.

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.

TEMPLETON, A. R., K. A. CRANDALL, and C. F. SING, 1992 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics **132:** 619–633.

VANDUIJN, C. M., P. DEKNIJFF, M. CRUTS, A. WEHNERT, L. M. HAVEKES *et al.*, 1994 Apolipoprotein e4 allele in a population-based study of early-onset Alzheimer's disease. Nat. Genet. **7:** 74–78.

WEST, H. L., G. W. REBECK, and B. T. HYMAN, 1994 Frequency of the Apolipoprotein E epsilon-2 allele is diminished in sporadic Alzheimer disease. Neurosci. Lett. **175:** 46–48.

YU, C., H. PAYAMI, J. M. OLSON, M. BOEHNKE, E. M. WIJSMAN *et al.*, 1994 The apolipoprotein E/CI/CII gene cluster and late-onset Alzheimer disease. Am. J. Hum. Genet. **54:** 631–642.

Communicating editor: B. S. WEIR