

Cladistic Analysis: Its Applications in Association Studies of Complex Diseases

C K Heng,*^{BSc (Hons), PhD}, P S Low,**^{FAMS, M Med (Paed), FRCP}

Abstract

Introduction: With the increase in genotype data generated by high throughput typing technologies, there is currently a lack of complexity-oriented analytical methods that can maximise the information obtained from these raw data for the study of complex diseases. We introduce the cladistic analysis that is traditionally applied in evolution studies and taxonomy, to specify relevant comparisons of traits associated with each haplotype/genotype in a population sample. **Methods:** Haplotypes are determined from the genotype data and linked to each other by their evolutionary relationships to form a cladogram. This is then used to specify relevant statistical comparisons. The central assumption is that any functionally important genetic variation causing a phenotypic effect at any point in the course of evolution will be embedded in the framework of haplotypes represented by the cladogram. **Applications:** There are various applications of cladistic analysis in the study of complex diseases. Basically, it helps in the identification of haplotypes that are associated with a disease state or a significantly altered level of quantitative trait. However, its limitations are that only polymorphic sites on the same DNA strand can be analysed and that recombination events must be relatively rarer than mutational events. **Conclusions:** In the absence of methods that can recognise the complexity of the genotype organization, and given its ability to exploit evolutionary information for optimising the analytical strategy, cladistic analysis would be a method of choice for studying multi-loci effects on a quantitative trait or disease outcome.

Ann Acad Med Singapore 2000; 29:313-21

Key words: Cladogram, Evolution, Genotypes, Haplotypes, Phylogenetic tree

INTRODUCTION

An increasing number of diseases are now found to be associated with one or more gene loci. Recent advances in genome-wide scanning techniques have resulted in the identification of susceptibility genes that are associated with complex diseases such as diabetes mellitus,^{1,2} coronary artery disease³⁻⁷ and various forms of cancers.⁸⁻¹¹ Complex diseases have multifactorial aetiology that include both environmental and genetic factors. Once susceptibility genes are identified, there is a wide range of mutation detection strategies to screen for genetic variations that contribute to the risk of the disease. Genotyping of a large number of individuals can now be efficiently carried out with high throughput typing technologies. The analysis of genetic data gets correspondingly complicated with increasing number of genetic markers studied. Investigators of complex diseases are confronted with very different research questions compared to those studying monogenic disorders that conform to classical Mendelian mode of inheritance. Besides being polygenic in aetiology, clinical outcomes of complex diseases are also determined by environmental factors and gene-environment interactions.

Recently, it is believed that common variants may contribute significantly to genetic risk for common diseases.¹²⁻¹⁴ This is called the common disease-common variant (CD-CV) hypothesis.¹⁵

Studying the Effect of a Genetic Variant

At this point, it would be helpful to technically define terms that are often confused. A genetic variant can either be a mutation or a polymorphism. A *mutation* is a change in the genetic material, either of a single gene or in the number or structure of the chromosomes.¹⁶ Unlike mutation, a *genetic polymorphism* is the occurrence in a population of two or more genetically determined forms in such frequencies that the rarest of them could not be maintained by mutation alone.¹⁷ The cutoff point of 1% to 2% for the rare allele is arbitrary and has no formal scientific basis. In a large population, the existence of a genetic condition that is not polygenic in nature at a frequency of more than, say 2%, must be the consequence of a selective advantage in the heterozygous form. In such a situation therefore, it is not realistic to attribute the high gene frequency of the condition to maintenance by high mutation rates, but by

* Research Fellow

** Professor and Head

Department of Paediatrics

National University of Singapore

Address for Reprints: Dr Heng Chew Kiat, Department of Paediatrics, National University of Singapore, 5 Lower Kent Ridge Road, Singapore 119074.

selective advantages.

“What is the effect of a genetic variation on its gene product?” This is one of the fundamental research questions that often arises when a genetic variant is detected. There are many possible effects depending on whether the genetic variation occurs at the regulatory sequences or the coding sequences. Variation at the regulatory region may alter the transcription/translation rate, which determines the quantity of product synthesised. A variation at the structural sequence may result in a change of the protein properties such as catalytic rate if it is an enzyme, or binding affinity, if it is a ligand or receptor. Recently, Shen et al¹⁸ demonstrated that single nucleotide polymorphisms (SNPs) can cause significant structural variations in allelic forms of mRNA and therefore allele-specific biological consequences would result from mRNA-structure-dependent processes. It is also possible that the variation is silent, regardless of whether it occurs at coding or non-coding region.

Genetic variations could be in the form of SNPs, insertion/deletion polymorphism and variable number of repetitive sequences (microsatellite, minisatellite, Alu repeats and L1 etc). For a known genetic variation, one of the simpler approaches to address the research question regarding its genetic effect is to carry out a cross-sectional association study in the population and determine if any of the genotypes present are associated with a significantly altered quantitative trait level. This approach works fine when there is only one genetic variation on the gene. However, this is usually not the case as the frequency of SNPs are expected to be between 0.3 to 5 per 1000 bp.¹⁹ As such, a large gene would theoretically have more polymorphic sites. In such a situation, the results obtained from the association study would not reveal a true effect of the polymorphism in question. This is because polymorphic sites elsewhere on the gene may have confounded its effect if one or more of them are functional variations.

Limitations of Traditional Approaches—The ApolipoproteinB Gene Example

We would like to draw from the example of the apolipoprotein B (apoB) gene to illustrate the study of a gene with many polymorphic sites. ApoB is a molecule found mainly on low-density lipoprotein (LDL). It serves as the ligand for LDL-receptor binding. Many polymorphic sites are found on the 43 kb long gene. This gene is known to influence lipid metabolism and therefore its genetic variations and quantitative plasma level of the gene product, the apoB molecule, are considered as risk factors for coronary artery disease (CAD).

The effects of individual apoB DNA polymorphisms on various plasma lipid traits in different ethnic groups of the world have been extensively studied. Unfortunately, results

from the majority of single-site genotype-phenotype association studies on apoB are inconsistent and therefore no conclusive effect can be established.²⁰ This may be due to the variable frequencies of the rare alleles in different ethnic groups.²¹ Other possible factors to account for the inconsistent findings include non-representative sampling criterion and small sampling size which inflates type I and II errors, respectively.²² Even if sample sizes could be large enough and representative of the population of interest, the study populations may be genetically heterogeneous. The situation may arise such that the functional DNA variation may be present in some but not in other populations, or linkage disequilibrium between polymorphic sites and DNA variations may differ among populations.^{23,24}

Although single-site genotype-phenotype studies have proven helpful in studying the role of candidate genes in the determination of lipid risk factors for CAD, conclusions about genetic architecture from these studies cannot be drawn precisely.²⁵ When the individual sites of apoB gene are studied in isolation, the overall effect of the gene cannot be determined as each of these sites may have effects of varying magnitude on plasma lipid and lipoprotein levels. Furthermore, the alleles in one site might have interaction with alleles at other sites, enhancing or limiting their effect. Some of these polymorphic sites may not be functional or could even be silent, but in linkage disequilibrium with another functional polymorphism elsewhere in the gene or chromosome. Linkage disequilibrium are prevalent between sites in short DNA regions. As such, phenotypic associations with polymorphic sites are not independent between sites. The complications arising as a result of dependence between sites can be avoided by studying simultaneously, associations of quantitative traits with the haplotypes as defined by all the available polymorphisms.

Currently, there is no known method that can be adequately applied to analyse the complex phenomenon of interactions between all susceptibility gene loci. Studying the effects of multiple loci of one gene like the apoB is already complicated, albeit more manageable than genome-wide variations. Ideally, one could type all known polymorphisms of the apoB gene and then work out an algorithm to determine their combined effects on plasma lipid traits and outcome of CAD. This is a more informative approach than carrying out a series of single locus association studies as had been done in earlier studies. Moving in this direction, investigators have recently started using multiple marker loci so that haplotypes can be constructed for genotype-phenotype association studies. However, as the number of markers increases, the problem of decreasing size for each combined genotype class arises, leading to a diminished statistical power to detect significant phenotypic differences between genotypes and haplotypes. Undesirably low level of statistical power becomes critical

in the study of quantitative lipid variables because it is expected that the majority of the effects contributed by genetic changes on the lipid variability will be small.²⁶ As the marker loci number increases, each individual will correspondingly gain a greater potential to have their own unique haplotypes and combined genotype. The number of possible combinations could exceed the number of individuals in the sample. Many possible genotypes will then likely to be absent from even a large sample. Consequently, even with modest numbers of marker loci, the number of exhaustive pair-wise comparisons between haplotypes or combined genotypes required to detect any association with quantitative changes in phenotypic lipid traits will increase the degrees of freedom for the statistical test used. In statistical tests, a larger difference must be present before a statistically significant level is attained when the degree of freedom is large.

An effective and efficient approach would have to reduce the number of irrelevant pair-wise comparisons but yet, avoid the biases that may be introduced when investigators preview the data and choose those contrasts that seem most likely to give significant results.²⁷ Such an approach can be satisfactorily carried out through cladistic analysis. This was developed by Templeton et al,²⁷⁻³¹ who specified the procedures for investigating the association between haplotype variations and its effect on phenotypic traits in a series of 5 papers.

Making Sense Out of the Myriad Genotypes, Haplotypes and their Quantitative Effects on Risk Factor Traits—The Cladistic Approach

Cladistics had its origin with evolution biologists in tracing the phylogenetic path of an organism of interest. It also contributes in the field of taxonomy. The three basic assumptions in cladistics are: 1) any group of organisms are related by descent from a common ancestor, 2) there is a bifurcating pattern of cladogenesis, and 3) change in characteristics occurs in lineages over time. These assumptions could be similarly applied to a DNA sequence instead of organisms. The use of cladistic analysis in studying complex diseases is therefore not a novel method but a novel application. It is an example of using an old tool for a new task. In this review, we introduce cladistic analysis as a tool for studying the effect of multiple genetic variations on quantitative traits, as well as for tracing the evolutionary path of a gene to shed some light on the genetic susceptibility of some haplotypes to diseases.

In cladistic analysis, haplotypes that are defined by multiple polymorphisms within a gene are used in constructing a cladogram. A cladogram represents the mutational steps that interrelate the observed haplotypes to one another. The cladogram can be rooted or unrooted.

Rooting requires identification of the ancestral haplotype. Therefore a rooted cladogram is basically a phylogenetic tree of the gene in question. The root of the cladogram constitutes the 0-step haplotype. There can only be one root or ancestral haplotype. Haplotypes that branch from the root represent the 1-step haplotypes. Those arising from the 1-step haplotypes form the 2-step haplotypes and so on. The highest step number occurs at the terminal tip and indicates the likelihood that the terminal haplotype is the most recently evolved.

The cladogram aids in analysis by defining a series of statistical tests that can be used to identify haplotypes which give rise to a marked phenotypic effect. The central assumption in cladistic analysis is that any functionally important genetic variation causing a phenotypic effect at any point in the course of evolution will be embedded in the framework of haplotypes represented by the cladogram.²⁷ Thus, a functional variation which occurs in a given haplotype will also be present in the succeeding haplotype.

Statistical analysis based on the framework defined by a cladogram increases both efficiency and power because evolutionary information is incorporated into the statistical design and contrast between haplotypes/combined genotypes could be made *a priori*.²⁷ In some cladistic studies, haplotypes were used as units of analysis to reduce the number of potential comparisons that would be required if genotypes were used.^{28,32} In general, n haplotypes define $n(n+1)/2$ genotypes. Therefore dividing data set into $n(n+1)/2$ treatment groups can seriously erode statistical power simply by making the sample size associated with each group small.

METHODOLOGY

Three major steps are involved in carrying out cladistic analysis. Firstly, haplotypes of each subject have to be determined. In the second step, evolutionary relationships among haplotypes have to be worked out using polymorphisms at multiple sites of the gene. Thirdly, the evolutionary relationships are used to specify the choice of haplotypes to make quantitative trait comparisons. We will use the example of apoB once again to illustrate the procedures of haplotype determination and cladogram construction.

Coding Procedures of Combined Genotypes and Haplotypes of the ApoB Polymorphisms

Numeric Coding of Combined Genotypes

In one of our studies, we used genotype data on six polymorphic sites of the apoB gene to construct the evolutionary pathways of the gene in Singaporean Chinese and Indians (Heng et al 1999).³³ These include

polymorphisms at the signal peptide (*sp*) region (insertion of Leu-Ala-Leu at nucleotide -51), four antigenic (*Ag*) sites: *Ag*(c/g) at nucleotide 421 (Thr to Ile), *Ag*(a₁/d) at nucleotide 1981 (Val to Ala), *Ag*(h/i) at nucleotide 11041 (Arg to Gln), *Ag*(t/z) at nucleotide 12669 (Glu to Lys) and a *Xba*I restriction site at nucleotide 7673 which is a silent mutation. In order to code for the combined genotypes of the six polymorphisms, a six-digit numeric label is assigned to each of the combined genotypes according to their relative positions on the apoB gene, that is, *sp*24/27, *Ag*(c/g), *Ag*(a₁/d), *Xba*I(X+/X-), *Ag*(h/i) and *Ag*(t/z). The numbers “0”, “1” and “2” were used to represent genotypes of homozygotes for the common alleles, heterozygotes, and homozygotes for the rare alleles respectively. For example, an individual with the following combined genotype (*sp*24/24, a₁d, cg, X-X-, ii, tt) would have a genotype code of “211000”.

Binary Coding of Unambiguous Haplotypes

From the list of combined genotypes, those that were homozygous at all loci or those not having more than one heterozygous site were manually selected as they were made up of unambiguous haplotypes. A binary coding system comprising the digit “0” to represent the common alleles and “1” to represent the rare alleles was used, retaining the relative position of the loci on the apoB gene as in the combined genotype codes. Estimation of the haplotype was not carried out for individuals who were heterozygous at two or more loci as haplotypes could not be determined unambiguously from these in the absence of family data.

Determining Ambiguous Haplotype and Their Frequencies by the Expectation-Maximisation (EM) Algorithm

In the presence of ambiguous haplotypes, which will usually be the case, a simple gene counting method is not possible because several haplotypes are possible for subjects who are heterozygous at more than one site. In such a situation, the maximum-likelihood haplotype frequencies can be computed using an EM algorithm.³⁴⁻³⁷ The EM algorithm is an iterative process that is very computationally intensive. It begins with an arbitrary estimate of the frequency of each haplotype using a random seed generator. These are then used to compute the expected genotype proportions, assuming Hardy-Weinberg equilibrium. This is known as the “E-step”. The relative genotype frequencies are used as weights for their two constituent haplotypes similar to the gene counting procedure. This leads to new estimates of haplotype frequencies. This is known as the “M-step”. The “E-” and the “M-” steps are repeated in a looped manner until the haplotype frequencies attain equilibrium with their corresponding genotype proportions. The limit of change permitted for each iterative step (the

epsilon value) is predefined by the investigator.

Construction of a Rooted Cladogram (Phylogenetic tree)

Establishment of the Ancestral Haplotype

The ancestral haplotype is usually inferred based on some biological indications. In the example of apoB, the ancestral haplotype is assumed to be *sp*27gdX-it. There are two supporting evidences for this assumption. The first is based on the findings of Rapacz et al,³⁸ who studied a group of unrelated chimpanzees as well as gorillas, and found only one apoB antigen group haplotype *Ag*(g,d,y,i,t) in them. From this, it was inferred that this is the ancestral haplotype from which all the others have evolved. Rapacz et al³⁸ also suggested that *Homo sapiens* had an African origin and subsequently migrated to Europe and the Pacific. The second basis for the assumption of this ancestral haplotype was the findings of Breguet et al²¹ that *Ag*(g), *Ag*(i) and *Ag*(t) were most likely the earliest alleles as they were found to be the most frequent in all the 13 human populations they studied. *Ag*(g, d, y, i, t) also represented the most common haplotype among the Bantu (frequency = 0.695) and the Senegalese (frequency = 0.577) who are both African populations. Subsequent studies have also used the same ancestral haplotype.^{39,40}

In cases where no ancestral haplotype could be inferred, the cladogram would be one with bi-directional arrows linking haplotypes (Fig. 1), unlike phylogenetic trees which have unidirectional arrows.

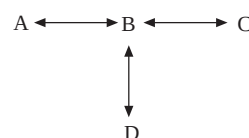


Fig. 1. A hypothetical unrooted cladogram.

Elucidating the Evolutionary Pathways

Regardless of whether the ancestral haplotype could be established, the next step involves using the marker information from the haplotypes to construct a cladogram. The haplotypes are linked according to their inferred evolutionary relationship. One of the methods for determining the evolutionary pathways is the maximum parsimony method.⁴¹⁻⁴³ The cladograms are constructed such that the fewest number of mutations are predicted to have occurred during the evolution of the observed haplotype array. This is the parsimony criterion of minimum net evolution. The assumptions underlying the construction of a cladogram are given in Sing et al.⁴⁴ In addition, the coalescent theory as proposed by Kingman^{45,46} could be

applied. Several points as discussed by Crandall et al⁴⁷ are very useful for rooting and resolving ambiguities in a cladogram. We provide here a summary of the assumptions and considerations from the aforementioned references as well as from other authors. These are:

1. Parsimony reflects evolutionary history and that all haplotypes present in the sample may be traced to a common ancestor.⁴⁴
2. The age of a haplotype is proportional to its frequency.⁴⁸ The older a mutation is, the greater is the chance that a second mutation had occurred. As such, those haplotypes differing by many mutations diverged from their common ancestor earlier than those differing by only a few.
3. Younger haplotypes that are recently evolved occur preferentially as terminal haplotypes.⁴⁹
4. Common haplotypes usually occur in the interior of the cladogram whereas rare ones usually occur at the terminal ends.⁴⁹
5. Recombination and gene conversion are rare events.²⁷ This assumption was reasonable in the case of the apoB gene because there were strong linkage disequilibrium among the apoB loci.
6. The rate of generation of functional genetic variations is lower than that of the variations in the marker loci.²⁷
7. Functional genetic variations can be traced through the same evolutionary history that is defined by the cladogram.²⁷

The above process of constructing the cladogram can be aided by available softwares such as Felsenstein's PHYLIP or Swofford's PAUP (Phylogenetic Analysis Using Parsimony). A web-based version of PHYLIP called

WebPHYLIP is also available (<http://sdmc.krdl.org.sg:8080/~lxzhang/phytip>). An example of a cladogram for apoB using genotype information from Singaporean Chinese is given in Figure 2. This approach has led us to identify 3 haplotypes that were exclusively found in the Chinese coronary artery disease population and that these 3 haplotypes were most recently evolved relative to the rest of the haplotypes.³³

Using Cladogram to Define Quantitative Contrast of Phenotypic Differences

Given n number of group, the number of pair-wise comparisons is given by the equation $(n^2-n)/2$. Consider a simple example consisting of 5 haplotypes named A to E. In a conventional statistical approach, one would have to make $(5^2-5)/2$ or 10 independent pair-wise multiple comparisons of their associated quantitative trait levels in order to determine if there is any significant difference between them. A cladogram could be constructed based on the evolutionary relationship of the 5 haplotypes as depicted in Figure 3.

With the cladistic strategy, the only pair-wise comparisons one would require to make are A-B, A-C, A-D and A-E. As

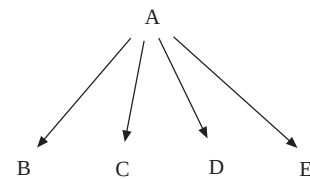


Fig. 3. A cladogram linking haplotypes A, B, C, D and E based on the evolutionary relationship with maximum parsimony.

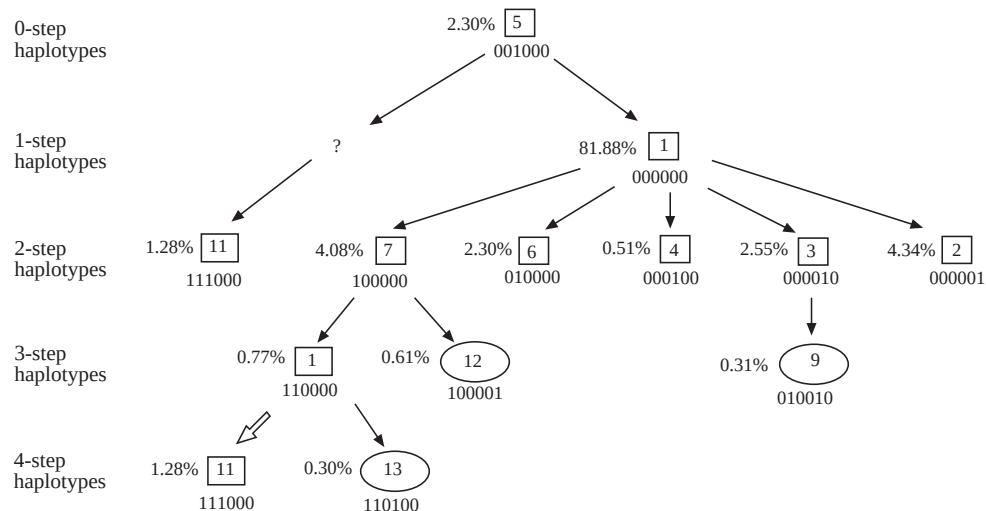


Fig. 2. Cladogram for Chinese (Squares = haplotypes present in healthy and CAD, Ellipse = Unique haplotypes in CAD patients, clear arrow indicates an unlikely path for derivation of H11). Frequencies of the haplotypes in are given as percentages. Frequency percentages of CAD haplotypes refer to their occurrence within the CAD sub-population. 0 = common alleles (sp27ga,X-it), 1 = rare alleles (sp24cdX+hz) (Reproduced with permission from Ann Hum Genet³³).

apparent, the 10 comparisons (9 degrees of freedom or *df*) involved in the conventional strategy would greatly corrode the statistical power to detect a true difference in associated quantitative trait levels between the haplotypes. The cladistic approach, however, could drastically reduce the number of comparisons to 4 (*df* = 3) and thus increases the likelihood of detecting a small difference that would otherwise be missed by the conventional method.

In a more complex cladogram, the strategy could be extended. This gives rise to a nested algorithm for making comparisons. For explanation, we will use a simple example of a cladogram that requires nesting as shown in Figure 4.

Comparison of quantitative traits begins with the terminal haplotype E and the internal haplotype B. If any significant differences could be detected between them, the mutational step associated with such a change in the quantitative level is already identified. Similar comparisons are subsequently carried out between F and B, H and G, as well as D and A. If no significant differences in trait levels are found, the terminal 0-step clades can be combined by moving back one mutational step. This union would form three 1-step clades consisting of BEF, GH and AD. Pair-wise testing for significant differences of quantitative traits could then be carried out for BEF with AD, and GH with C. If there is still no difference, the next union of one more mutational step backwards would leave only one comparison between ADBEF with CGH.

Details of the algorithm and various applications of the cladistic analysis can be obtained from Templeton et al's series of 5 papers,²⁷⁻³¹ which we will briefly review here.

In their first paper, Templeton et al²⁷ described the fundamental principles underlying the construction of the cladogram and its application in analysing the alcohol dehydrogenase gene of an inbred population of drosophila. However, the method is only applicable for a homozygous or haploid population. In their second paper in the series, Templeton et al²⁸ applied the cladistic analysis on a natural diploid human population with heterozygotes. Since the human population is diploid, the concept of average excess that Templeton introduced earlier had to be used.⁵⁰ Next, they evaluated the limits of parsimony in estimating cladogram and addressed the issue of confidence that one could place on them.²⁹ In paper IV,³⁰ they extended the analytical strategy to handle categorical phenotypes in addition to quantitative phenotypes. Further application of cladistic analysis for a case/control design was discussed in Templeton's final paper in the series.³¹ Following that, Templeton had another review paper on cladistics that put forth the merits of procedures which incorporate evolutionary history to identify determinants of variability in multifactorial phenotypes over procedures that do not.⁵¹

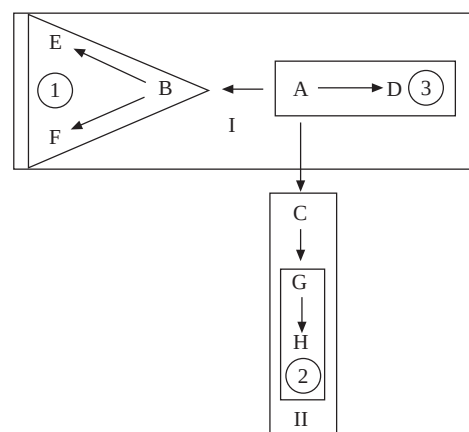


Fig. 4. A hypothetical cladogram that requires nesting.

Haplotypes as Units of Analysis

The use of haplotypes necessitates the calculation of average excess or average effect of each haplotype for statistical comparisons. Both average excess and average effect are ways of assigning a phenotypic effect to a haploid genotype although the phenotype is expressed only in diploid individuals. The equations for calculating these two quantities can be found in reference 50.

In instances where all genotypes are homozygous, as in experiments with drosophila in which homozygotes can be inbred,⁵² the problem of assigning haplotype effects is eliminated and the average excesses can be analysed with standard nested analysis of variance (ANOVA) procedures.²⁷ For a genetic epidemiological study, the subjects are outbred populations in which a single haplotype could exist in a variety of different genotypes without a one-to-one correspondence between haplotypes and the genotypes expressing the phenotype. When confounded by heterozygosity, haplotype effects cut across the sample of individual phenotypic measurements in a complex manner, disabling the use of conventional statistical tests like ANOVA and nested ANOVA.²⁸ One of the alternatives to circumvent this problem is to use a sample reuse procedure called the random permutation test developed by Edgington.⁵³ There are potential problems inherent in this procedure though. It is computation-intensive and its power is not adequately known. Another alternative used by Haviland et al⁵⁴ could be adopted. It involves only a subset of unambiguous haplotypes of which the one which is homozygous and *n*-1 step away is used to contrast with all the haplotypes of the *n*-step clades. All the individuals involved have one common copy of the haplotype, leaving the other variable haplotype for contrast. Diploid phenotypic variations can therefore be used directly without having to calculate average excesses of the haplotypes and more importantly, the data can be made amenable to relatively

less complex parametric analyses like the ANOVA. However, this approach only allows a subset of genotypes that has one haplotypes in common with others to be used.

EXAMPLES OF CLADISTICS APPLICATION

Despite its usefulness, cladistic analysis has not been popularly applied in the study of complex diseases. One of the earlier works was that of Dunning et al.³⁹ They postulated a phylogenetic tree for the human apoB gene in both healthy and myocardial infarction patients. Two of the haplotypes were found to be associated with elevated plasma apoB levels. Working also with the apoB gene, Hallman et al.³² found association between some apoB haplotypes with HDL-cholesterol, triglycerides and total cholesterol levels. With the cladistic analysis, Friedlander et al.⁵⁵ found apoB haplotypes that respond significantly to dietary challenge of total cholesterol and LDL-C in young people. Using similar approach, Haviland et al.⁵⁴ identified haplotypes of the apoAI-CIII-AIV gene cluster and the LDL receptor gene⁵⁶ that were associated with small pleiotropic effects on plasma lipid traits. In a study of 10 polymorphisms spanning 26 kb of the ACE gene, Keavney et al.⁵⁷ used cladograms to localise the polymorphic sites that significantly influence angiotensin converting enzyme levels. From another group, Lobos et al.⁵⁸ found significant association of the tyrosine hydroxylase gene polymorphisms with alcoholism but not in the case of the dopamine D2 receptor gene⁵⁹ by using the nested statistical analysis strategy derived from cladograms. The results of Kittles et al.⁶⁰ also demonstrated the influence of eight loci in the Y chromosomes on the risk of alcoholism in Finnish males using cladistic analysis.

EVALUATION OF CLADISTIC ANALYSIS—CAVEATS

Like all methodology, cladistic analysis is not without its limitations. Firstly, cladistic analysis can only analyse genetic information from sites that are on the same strand of DNA. Although this may span many genes, complex diseases usually have candidate genes that are located on various chromosomes. As such, it is only useful when the candidate gene or genes on the same cluster have multiple polymorphic sites. When candidate genes are spread over different chromosomes, cladistic analysis would not be applicable.

Secondly, haplotypes in the cladogram are linked to each other based on the mutational steps over time. However, some of the haplotypes may have been derived from recombination events and hence are not legitimate constituents of the cladogram. When recombination events are more frequent in the region of DNA being investigated but less frequent than mutational changes, the frequency of recombination events can be reduced to very rare or non-

existent by subdividing the region of study to smaller segments. Separate cladograms for each segment could be constructed to identify functional variations. If recombination frequency is common relative to mutational events, then cladistic analysis would not be appropriate.

Thirdly, the task of constructing a cladogram is relatively easier with genetic variations being single nucleotide substitution or simple insertion/deletion polymorphisms. When micro- or minisatellite repetitive sequences are present, the large number of alleles would result in the absence of many theoretically possible haplotypes. Even those that are present would only be carried by very few individuals. As such, the problem of large number of comparisons would still persist despite the cladogram-specified strategy.

CONCLUSIONS

In the absence of complexity-oriented analytical strategies, cladistic analysis is still by far a useful tool for identifying haplotypes that either contribute to a significantly altered quantitative trait or associated with clinical manifestations of a complex disease. Currently, no method is yet able to recognise the complexity of the genotype organization and exploit the coherency provided by evolutionary pathways to optimise an analytical strategy. As such, cladistic analysis would be a method of choice for studying multi-loci effects despite its many limitations.

Acknowledgement

The authors would like to thank Ms Jenny Tan for her critical comments on this article.

REFERENCES

1. Hanis C L, Boerwinkle E, Chakraborty R, Ellsworth D L, Concannon P, Stirling B, et al. A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nat Genet* 1996; 13:61-6.
2. Todd J A, Farrall M. Panning for gold: genome-wide scanning for linkage in type 1 diabetes. *Hum Mol Genet* 1996; 5:1443-8.
3. Ciccarese M, Pacifico A, Tonolo G, Pintus P, Nikoshkov A, Zuliani G, et al. A new locus for autosomal recessive hypercholesterolemia maps to human chromosome 15q25-q26. *Am J Hum Genet* 1999; 66:453-60.
4. Aouizerat B E, Allayee H, Cantor R M, Davis R C, Lanning C D, Wen P Z, et al. A genome scan for familial combined hyperlipidemia reveals evidence of linkage with a locus on chromosome 11. *Am J Hum Genet* 1999; 65:397-412.
5. Almasy L, Hixson J E, Rainwater D L, Cole S, Williams J T, Mahaney M C, et al. Human pedigree-based quantitative-trait-locus mapping: localization of two genes influencing HDL-cholesterol metabolism. *Am J Hum Genet* 1999; 64:1686-93.
6. Pajukanta P, Terwilliger J D, Perola M, Hiekkalinna T, Nuotio I, Ellonen P, et al. Genomewide scan for familial combined hyperlipidemia genes in finnish families, suggesting multiple susceptibility loci influencing triglyceride, cholesterol, and apolipoprotein B levels. *Am J Hum Genet* 1999; 64:1453-63.

7. Varret M, Rabè J P, Saint-Jore B, Cenarro A, Marinoni J C, Civeira F, et al. A third major locus for autosomal dominant hyper-cholesterolemia maps to 1p34.1-p32. *Am J Hum Genet* 1999; 64:1378-87.
8. Xu J, Meyers D, Frijie D, Isaacs S, Wiley K, Nusskern D, et al. *Nat Genet* 1998; 20:175-9.
9. Smith J R, Freije D, Carpten J D, Gronberg H, Xu J, Isaacs S D, et al. Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search. *Science* 1996; 274:1371-4.
10. Narod S A, Feunteun J, Lynch H T, Watson P, Conway T, Lynch J, et al. Familial breast-ovarian cancer locus on chromosome 17q12-q23. *Lancet* 1991; 338:82-3.
11. Hall J M, Lee M K, Newman B, Morrow J E, Anderson L A, Huey B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 1990; 250:1684-9.
12. Risch N, Merikangas K. The future of genetic studies of complex diseases. *Science* 1996; 273:1516-7.
13. Lander E S. The new genomics: global views of biology. *Science* 1996; 274:536-9.
14. Collins F S, Guyer M S, Chakravarti A. Variations on a theme: cataloging human DNA sequence variation. *Science* 1997; 278:1580-1.
15. Cargill M, Alshuler D, Ireland J, Sklar P, Ardlie K, Patil N, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999; 22:231-8.
16. Ford EB. *Mendelism and evolution*. London: Methuen, 1960.
17. Vogel F, Motulsky A G. *Human genetics : problems and approaches*. Berlin; New York: Springer, 1997:497.
18. Shen L X, Basilion J P, Stanton Jr V P. Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc Natl Acad Sci USA* 1999; 96:7871-6.
19. Chakravarti A. Population genetics-making sense out of sequence. *Nat Genet* 1999; 21:56-60.
20. Young S G. Recent progress in understanding apolipoprotein B. *Circulation* 1990; 82:1574-94.
21. Breguet G, Büttler R, Büttler-Brunner E, Sanchez-Mazas A. A worldwide population study of the Ag-system haplotypes, a genetic polymorphism of human low-density lipoprotein. *Am J Hum Genet* 1990; 46:502-17.
22. Humphries S E, Talmud P, Monsalve V, McKeigue P. RFLP studies in different ethnic groups. *Atherosclerosis* 1989; 75:249-50.
23. Haviland M B, Kessling A M, Davignon J, Sing C F. Estimation of Hardy-Weinberg and pairwise disequilibrium in the apolipoprotein AI-CIII-AIV gene cluster. *Am J Hum Genet* 1991; 49:350-65.
24. Cooper D N, Clayton J F. DNA polymorphism and the study of disease associations. *Hum Genet* 1988; 78:299-312.
25. Kessling A, Ouellette S, Bouffard O, Chamberland A, Betard C, Selinger E, et al. Patterns of association between genetic variability in apolipoprotein (apo) B, apoAI-CIII-AIV, and cholesterol ester transfer protein gene regions and quantitative variation in lipid and lipoprotein traits: influence of gender and exogenous hormones. *Am J Hum Genet* 1992; 50:92-106.
26. Tanksley S D. Mapping polygenes. *Ann Rev Genet* 1993; 27:205-33.
27. Templeton A R, Boerwinkle E, Sing C F. 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* 1987; 117:343-51.
28. Templeton A R, Sing C F, Kessling A, Humphries S. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping II. The analysis of natural populations. *Genetics* 1988; 120:1145-54.
29. Templeton A R, Crandall K A, Sing C F. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. III. Cladogram estimation. *Genetics* 1992; 132:619-33.
30. Templeton A R, Sing C F. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics* 1993; 134:659-69.
31. Templeton A R. 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. *Genetics* 1995; 140:403-9.
32. Hallman D M, Visvikis S, Steinmetz J, Boerwinkle E. The effect of variation in the apolipoprotein B gene on plasma lipid and apolipoprotein B levels. I. A likelihood-based approach to cladistic analysis. *Ann Hum Genet* 1994; 58:35-64.
33. Heng C K, Saha N, Low P S. Evolution of the apolipoprotein B gene and coronary artery disease: a study in low and high risk Asians. *Ann Hum Genet* 1999; 63:45-62.
34. Dempster A, Laird N, Rubin D. Maximum likelihood estimation from incomplete data via the EM algorithm. *J R Stat Soc* 1977; 39:1-38.
35. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995; 12:921-7.
36. Lange K. *Mathematical and statistical methods for genetic analysis*. New York: Springer, 1997.
37. Weir B S. *Genetic data analysis II: Methods for discrete population genetic data*. Sunderland, MA, USA: Sinauer Assoc Inc, 1996.
38. Rapacz J, Chen L, Büttler-Brunner E, Wu M J, Hasler-Rapacz J O, Büttler R, et al. Identification of the ancestral haplotype for apolipoprotein B suggests an African origin of *Homo sapiens sapiens* and traces their subsequent migration to Europe and the Pacific. *Proc Natl Acad Sci USA* 1991; 88:1403-6.
39. Dunning A M, Renges H H, Hamsten A, Talmud P, Humphries S. A postulated phylogenetic tree for the human apolipoprotein B gene: unpredicted haplotypes are associated with elevated apoB levels. *Biochim Biophys Acta* 1993; 1165:271-8.
40. Renges H H, Peacock R, Dunning A M, Talmud P, Humphries S E. Genetic relationships between the 3'-VNTR and diallelic apolipoprotein B polymorphisms: Haplotype analysis in individuals of European and South Asian origin. *Ann Hum Genet* 1992; 56:11-33.
41. Felsenstein J. Phylogenies from molecular sequences: inference and reliability. *Ann Rev Genet* 1988; 22:521-65.
42. Sober E. Parsimony in systematics: philosophical issues. *Ann Rev Ecol Syst* 1983; 14:335-57.
43. Stewart C B. Mapping polygenes. *Ann Rev Genet* 1993; 27:205-33.
44. Sing C F, Haviland M B, Templeton A R, Zerba K E, Reilly S L. Biological complexity and strategies for finding DNA variations responsible for inter-individual variation in risk of a common chronic disease, coronary artery disease. *Ann Med* 1992; 24:539-47.
45. Kingman J F C. The coalescent. *Stochastic Processes Appl* 1982; 13:235-48.
46. Kingman J F C. On the genealogy of large populations. *J Appl Probab* 1982; 19A:27-43.
47. Crandall K A, Templeton A R. Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics* 1993; 134:959-69.
48. Donnelly P, Tavaré S. The ages of alleles and coalescent. *Adv Appl Probab* 1986; 18:1- 19.
49. Golding G B. The detection of deleterious selection using ancestors inferred from a phylogenetic history. *Gene Res* 1987; 49:71-82.
50. Templeton A R. The general relationship between average effect and average excess. *Genet Res Camb* 1987; 49:69-70.

51. Templeton A R. Cladistic approaches to identifying determinants of variability in multifactorial phenotypes and the evolutionary significance of variation in the human genome. In: Chadwick D, Cardew G, editors. *Variation in the Human Genome*. Chichester: John Wiley & Sons, 1996:259-77.
 52. Aquadro C F, Deese S F, Bland M M, Langley C H, Laurie-Ahlberg C C. Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster*. *Genetics* 1986; 114:1165-90.
 53. Edgington E W. *Randomization tests*. 2nd ed. New York: Marcel Dekker, 1987.
 54. Haviland M B, Kessling A M, Davignon J, Sing C F. Cladistic analysis of the apolipoprotein AI-CIII-AIV gene cluster using a healthy French Canadian sample. I. Haploid analysis. *Hum Genet* 1995; 59:211-31.
 55. Friedlander Y, Berry E M, Eisenberg S, Stein Y, Leitersdorf E. Plasma lipids and lipoproteins response to a dietary challenge: analysis of four candidate genes. *Clin Genet* 1995; 47:1-12.
 56. Haviland M B, Ferrell R E, Sing C F. Association between common alleles of the low-density lipoprotein receptor gene region and interindividual variation in plasma lipid and apolipoprotein levels in a population-based sample from Rochester, Minnesota. *Hum Genet* 1997; 99:108-14.
 57. Keavney B, McKenzie C A, Connel J M C, Julier C, Ratcliffe P J, Sobel E, et al. Measured haplotype analysis of the angiotensin-I converting enzyme gene. *Hum Mol Genet* 1998; 7:1745-51.
 58. Lobos E A, Todd R D. Cladistic analysis of disease association with tyrosine hydroxylase: application to manic-depressive disease and alcoholism. *Am J Med Genet* 1997; 74:289-95.
 59. Lobos E A, Todd R D. Association analysis in an evolutionary context: Cladistic analysis of the DRD2 locus to test for association with alcoholism. *Am J Med Genet* 1998; 81:411-9.
 60. Kittles R A, Long J C, Bergen A W, Eggert M, Virkkunen M, Linnoila M, et al. Cladistic association analysis of Y chromosome effects on alcohol dependence and related personality traits. *Proc Natl Acad Sci USA* 1999; 96:4204-9.
-