Am. J. Hum. Genet. 41:356–373, 1987

A Deductive Method of Haplotype Analysis in Pedigrees

ELLEN M. WIJSMAN

Department of Genetics, Stanford University, Stanford, CA

Derivation of haplotypes from pedigree data by means of likelihood techniques requires large computational resources and is thus highly limited in terms of the complexity of problems that can be analyzed. The present paper presents 20 rules of logic that are both necessary and sufficient for deriving haplotypes by means of nonstatistical techniques. As a result, automated haplotype analysis that uses these rules is fast and efficient, requiring computer memory that increases only linearly (rather than exponentially) with family size and the number of factors under analysis. Some error analysis is also possible. The rules are completely general with regard to any system of completely linked, discrete genetic markers that are autosomally inherited. There are no limitations on pedigree structure or the amount of missing data, although the existence of incomplete data usually reduces the fraction of haplotypes that can be completely determined.

INTRODUCTION

Highly polymorphic genetic markers are a necessity for efficient linkage analysis of the human genome. The markers that are most polymorphic tend to be those for which haplotypes have been defined (Willard et al. 1985). The haplotypes may consist of (1) members of a clustered gene family detected as protein markers and/or restriction-fragment-length polymorphisms (RFLPs) (Migone et al. 1983; Cohen-Haguenauer et al. 1985; Le Gall et al. 1985; Ohlsson et al. 1985; Johnson et al. 1986), (2) RFLPs for several different regions detected with probes defined by a cosmid insert (Litt et al. 1985; Litt and White 1985; Bufton et al. 1986; Buroker et al. 1986), (3) RFLPs for multiple enzymes

Received September 23, 1986; revision received March 4, 1987.

Address for correspondence and reprints: Dr. Ellen M. Wijsman, Department of Genetics, Stanford University, Stanford, CA 94305.

 0002-9297/87/4103-0003$2.00

detected with a single-copy probe (Chakravarti et al. 1984, 1986; Eng and Strom 1985; Julier et al. 1985; Lidsky et al. 1985; Elbein et al. 1986; Higgs et al. 1986), or (4) other possibilities. Extended haplotypes, such as those that combine RFLPs with classical markers, are particularly useful for linkage studies. The increase in polymorphism information content (PIC) (Botstein et al. 1980) can be substantial: in Caucasians, the PIC of Gm increases from 0.38 for classical haplotypes (estimated from the data of Steinberg and Cook [1981, pp. 21–57]) to 0.95 if one considers the haplotypes as including RFLPs (estimated from the data of Johnson [1984]).

The construction of extended haplotypes serves two additional purposes. First, an association between classical alleles and RFLPs may become clear. Second, when extended haplotypes that include a new RFLP are being determined, more fragments can be assigned to haplotypes than when haplotypes are determined for the new RFLP without the other markers.

Maximum-likelihood estimation of autosomal haplotypes from pedigrees is possible and has been developed for two or three linked loci (Larsen 1979), both as a part of a package of programs, FAP (Baur et al. 1984, 1985), and as a special case of multilocus linkage analysis in the programs LINKAGE (Lathrop and Lalouel 1984) and PAP (Hasstedt and Cartwright 1979; Hasstedt 1982). However, the substantial memory and time requirements limit the complexity of the systems that can be analyzed, and most of the programs are not readily implementable on small computers. For large problems it often remains necessary to initially find haplotypes in pedigrees by hand; for smaller problems this approach is also of benefit in reducing subsequent computations.

For a small number of markers, geneticists have little difficulty in determining haplotypes from pedigree data. As the number and complexity of the systems increases, however, so does the number of errors, incomplete deductions, and difficulty of convincing oneself and others of the validity and thoroughness of the analysis. In addition, the time necessary to do the analysis by hand becomes very substantial. Although the difficulty that accompanies the analysis of large data sets is partially a function of the amount of data, it is also the result of an incomplete understanding of how the geneticist determines haplotypes from pedigree data. The underlying rules used in haplotype analysis are neither extraordinarily numerous nor complex and can be stated in terms that do not require substantial training in genetics to be understood and applied.

The purpose of the present paper is to present these rules, and the main body of the paper is devoted to their description. The rules were formalized for two reasons: (1) to develop a computer program for automated analysis and (2) to provide a set of algorithms that can be used either for manual haplotype analysis or for understanding the logic behind haplotype analysis as performed by a computer or a geneticist. In addition, it seemed useful to present the rules in a fashion that strips them of much of their genetic terminology—and of the accompanying requirement that they be applied by a person with good knowledge of genetics. By doing this, the underlying logic of haplotype analyses of diverse genetic systems can be removed from the peculiarities of the individual systems. However, although a completely formal description would present the

rules with the symbols of logic, such a presentation would suffer from lack of clarity and would not increase brevity. So the rules are presented here in primarily verbal terms.

The program developed for the analysis is called PATCH (pedigree analysis to construct haplotypes). The time necessary to do the analysis on a microcomputer, although dependent on the computer in use, is usually <1 min; a very complicated problem may take a few minutes. Approximately 13.5 K of memory for data are required to handle 100-member pedigrees segregating for 20 factors; these numbers can easily be increased. Some error analysis is also possible, but not in as precise a fashion as the haplotype analysis. The program is written in C; copies with documentation and supporting data-management programs are available on request.

DEFINITIONS

The rules used to deduce haplotypes require the use of a few terms, which are defined in the Appendix. The importance of using these terms lies in their ability to transform all autosomal markers into a series of locus-like elements, each of which behaves as a locus having one dominant and one recessive allele. We need only to describe the rules of inheritance of such two-allele systems (subject to certain constraints) to be able to describe inheritance of any system.

Symbols used in the figures are given in figure 1. As an example to clarify the terms, consider the Rh blood group (table 1). If we ignore rare factors such as C^w, there are five factors: C, c, D, E, and e. (These factors are commonly considered to define three linked loci with alleles C vs. c, D vs. d, and E vs. e.) In figure 2(*a*), the phenotype of the father consists of the presence of all five factors; the phenotype of the mother consists of the presence of the c, D, and e factors and the absence of the C and E factors. Haplotypes that can be deduced

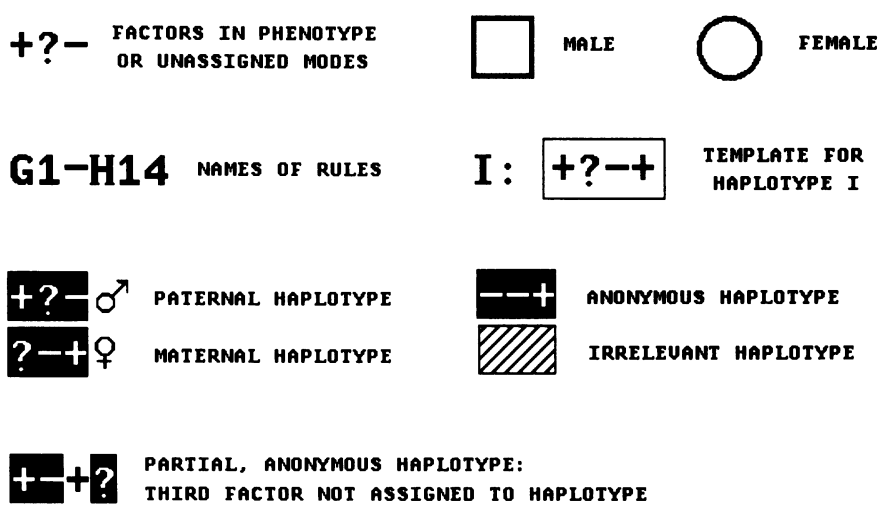


FIG. 1.—Symbols used in figures

TABLE 1

HAPLOTYPES FOR Rh DEFINED BY FIVE FACTORS

HAPLOTYPES	FACTORS				
	C	c	D	E	e
CDE	+	−	+	+	−
CDe	+	−	+	−	+
CdE	+	−	−	+	−
cDE	−	+	+	+	−
Cde	+	−	−	−	+
cDe	−	+	+	−	+
cdE	−	+	−	+	−
cde	−	+	−	−	+

can be represented as in figure 2(*b*) or 2(*c*). In 2(*c*), the haplotypes are given in the conventional manner; 2(*b*) shows the same haplotypes in the representation used in the present paper.

In figure 2(*b*), the father's two modes for factor C are a dominant mode on haplotype I and a recessive mode on haplotype II, i.e., the C factor is present on I and missing on II (each individual has two modes per factor—see Assumptions below). Both of these modes are known, and they are assigned to anonymous haplotypes. In the son, the modes for factor C are assigned to parental haplotypes, i.e., their origins are the respective parents. The genotype of the father is heterozygous for factor C; the mother and daughter are homozygous for factor C. For factor D, one of the modes in the father is unknown, so the father is neither homozygous nor heterozygous (see the Appendix for definition of these terms). In addition, the two modes for factor D are unassigned, i.e., their respective haplotypes are unknown. Haplotypes I and II in the father are partial. Haplotype I* in the mother is complete and full; haplotype II* is full but not complete.

An individual mode corresponds to the usual definition of an allele only when a single factor defines the locus, e.g., the D locus for Rh. If multiple factors define a locus, an individual allele becomes a haplotype or subhaplotype. For example, the C allele becomes the + − subhaplotype of Rh. Also, the notion of codominance loses meaning in this context. Codominance is a function of the

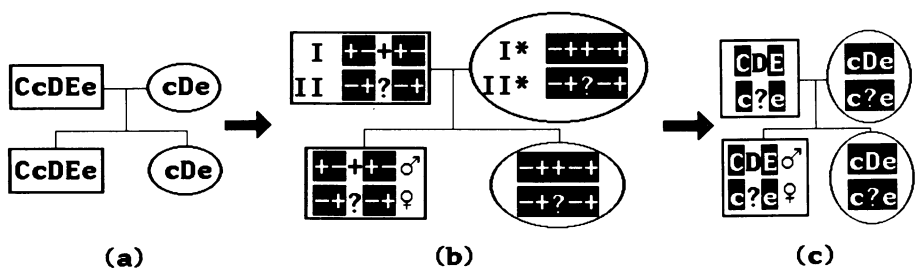


FIG. 2.—Family segregating for Rh. Factors are as in table 2. Symbols are as in fig. 1. (*a*), Phenotypes; (*b*), haplotypes coded as binary factors; and (*c*), traditional haplotypes.

absence of one or more possible combinations of modes on haplotypes (e.g., the $--$ and $++$ combinations for the Rh C gene do not exist) and is treated by maintaining a table of all possible haplotypes that are known to exist in a population. In table 1, we see codominance for the C and c alleles of Rh, since haplotypes containing neither or both factors do not exist. Not all systems will have such a table, particularly new systems. Such tables contain all information needed to take into account the peculiarities of inheritance of particular systems.

ASSUMPTIONS

The rules makes use of some assumptions. Some can be relaxed later, and others are less strict than they first seem. These rules are as follows:

1. Factors are completely linked: all haplotypes are inherited as complete units and no recombination occurs between factors.
2. No mutation occurs in observed pedigrees.
3. All individuals in the pedigrees are related as stated: no incorrect paternity, adoptions, etc. exist.
4. Phenotypes are accurately given: mistypings of factors do not exist.
5. If a set of subhaplotypes (e.g., alleles for a locus on the haplotypes) is specified for a section of the total haplotypes, this specification includes all such subhaplotypes that may occur in the population.
6. Factors show complete dominance: In a genotype the presence of one factor is indistinguishable from the presence of two factors.
7. Factors are on autosomes.
8. Individuals are diploid: an individual receives one haplotype from each parent for the genetic systems under analysis.

Haplotype analysis can sometimes identify violations in one or more of the above assumptions. It is not generally possible to identify with absolute certainty which assumption has been violated, although the circumstances may make one type of assumption error more likely than another. For example, incorrect paternity may be more likely than mutation. A thorough discussion of how to identify which rule is most likely violated is beyond the scope of the present paper. However, where violations can be detected, the rules below note how to do this.

RULES

In the description of the rules, the terms "parent," "child," and "individual" will be used to describe individuals who are under consideration. If the term "parent" is used, it is assumed that he or she has at least one child in the pedigree; if the term "child" is used, it is assumed that he or she has at least one parent in the pedigree. An individual with no parents has no parents *in the pedigree*. In rules that make use of the parent-child relationship, if either parent or child is missing from the pedigree, the rule does not apply.

Each rule makes some inference about one of two types of information: (1) *what* is inherited and (2) *how* it is inherited, i.e., from which parent a mode is derived or on which haplotype it is found. To make these inferences, certain

facts about genetic inheritance are used. All rules assume that each individual has received from each parent one haplotype containing only one mode for each factor. Application of this assumption comes into play when modes are assigned to haplotypes or when their origins are assigned to particular relatives: in the following description of the rules, whenever a mode is assigned to one haplotype in an individual, the other mode of the same factor is automatically assigned to the second haplotype in the individual.

The rules are allowed to make certain kinds of alterations: rules may force unknown modes to become known or to assign modes to parental or anonymous haplotypes. Certain alterations are forbidden: rules may not change known modes to unknown modes or unassign assigned origins. If a rule encounters a situation that demands a forbidden change, an error in the data and/or assumptions must have occurred.

The rules can conveniently be divided into two types: (1) the rules of genotyping and (2) the rules of haplotyping. For efficient analysis, certain rules should be applied before others; most important, the rules of genotyping should be applied before the rules of haplotyping. Within each category of rules, there is considerable flexibility in the order in which the rules can be applied, although certain orderings are more efficient than others. Figure 3 shows a flow diagram indicating which rules should come before others, and it can be followed either along with descriptions of the rules or as an aid in manual analysis. There are many points in the analysis that allow choices to be made concerning the next rule to be applied. All rules must be considered at some point during any given analysis, and rules must be applied until no more useful changes can be made. A useful change is one that eventually results in an assignment of a mode to a haplotype, the assignment of a haplotype to a parent, or the change of a mode from unknown to known. The logic flow looks complicated at first glance, but in reality it is not difficult to follow. There are three basic groups of rules of haplotyping, those based on (1) whether knowledge is available on all possible haplotypes that exist in a population (H3), (2) whether origins of haplotypes in a child are its parents (H4), and (3) whether such origins are anonymous (H5). These groups are linked through four simple rules that assign modes to haplotypes (H1 and H2) or change the origins of haplotypes from anonymous to parental and vice versa (H13 and H14). Each group also uses some rules that are common to the other combinations (H9 and H10).

Rules of Genotyping

Figure 4 gives an example of the rules of genotyping as they are applied to a small family and can be consulted concurrently with the description of the rules.

The first rule of genotyping, rule G1, deduces genotypes from phenotypes. G1 must be applied before any other rule and exactly once for each factor in each individual. There are three mutually exclusive parts to the rule.

> *Rule G1:* (a) If a factor is absent in the phenotype of an individual, both modes in the genotype must be recessive. (b) If a phenotype is positive, one mode must be

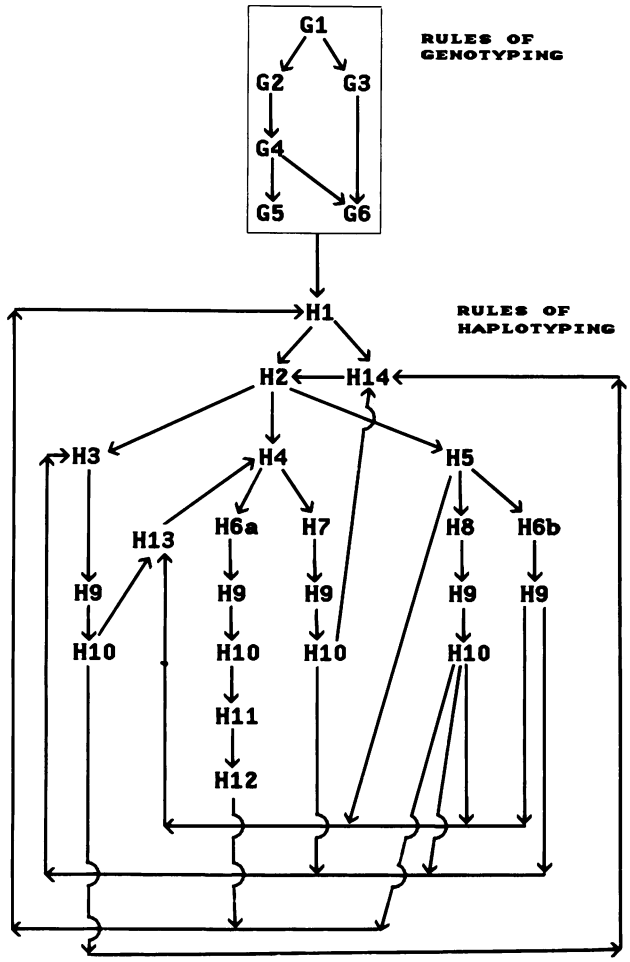


FIG. 3.—Logic flow of rules used in haplotype analysis. Rules are as described in text

dominant, but the other mode is unknown. (c) If a phenotype is unknown for a factor, both modes are also unknown.

The second rule of genotyping changes unknown modes to known modes on the basis of homozygosity of parent or child.

Rule G2: Suppose A and B are a parent and child (either one may be the parent). If, for a given factor, A is homozygous, its mode must exist in the genotype of B. So, if neither mode in B's genotype is known to be the same as that of A's genotype *and* there are one or more unknown modes, change one unknown mode in B's genotype to A's mode.
Error check: If the mode of A's homozygous genotype does not currently exist in B's genotype *and* both modes in B's genotype are known, an error exists.

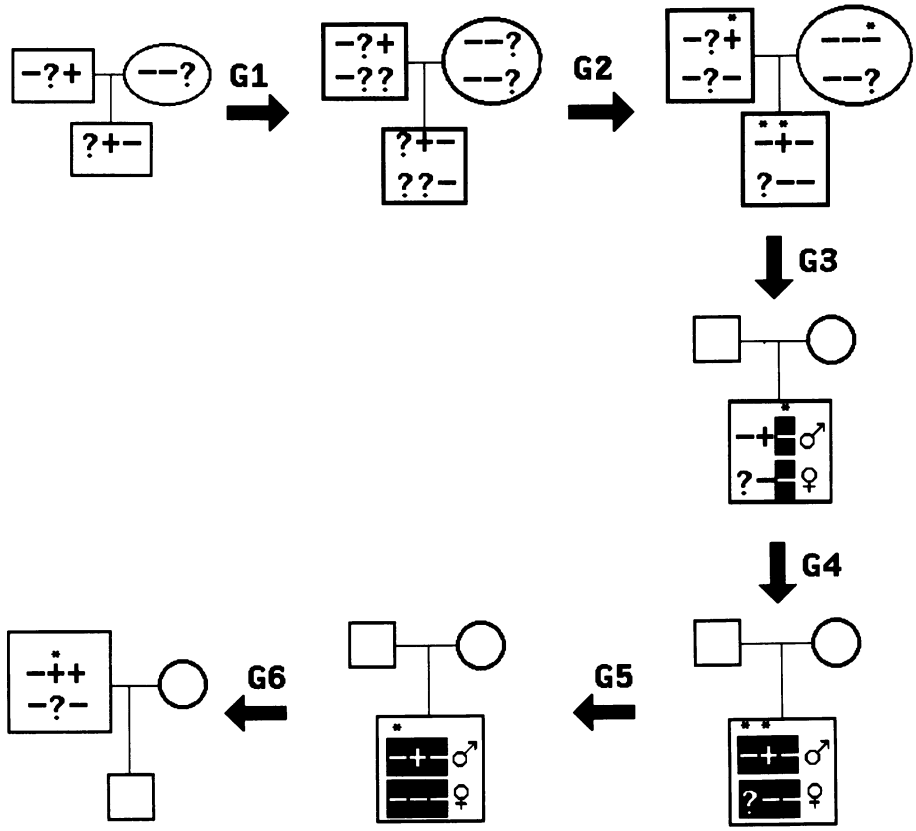


FIG. 4.—Example of one of seven possible orderings of rules for genotyping in a small family. Empty circles and squares are place holders, indicating no changes in those individuals. Other symbols are as in fig. 1. Asterisks denote factors for which a rule changed some information.

The function of rules G3 and G4 is to determine the origins of a pair of modes in a child.

Rule G3: If the genotype of a child is homozygous, assign the origins of the modes in the child to the parents, one assignment to each parent.

Rule G4: If a parent is homozygous and the origins of the child's modes have not yet been assigned to the parents, assign to the parent the origin of a mode in the child that is identical to the parental modes.

Once some modes have been assigned, it is possible to use the assignments to deduce more about the modes. The final two rules of genotyping do this.

Rule G5: If an unknown mode in a child is assigned to a homozygous parent, change the unknown mode to that of the parent's mode.

Rule G6: If the origin of a known mode in a child is a particular parent and the child's mode is not already known to exist in the parent, if there is an unknown mode in the parent, change it to the state of the child's mode.
Error check: It there is no unknown mode in the parent, there is an error.

Rules of Haplotyping

Once all possible changes have been made with the rules of genotyping, partial or complete haplotypes will be known in some individuals since all modes in an individual that are assigned to a given parent are on the same haplotype (by assumptions 1, 7, and 8 above). Further improvement of the haplotypes can only be made by using information about the joint inheritance of multiple modes. Specific haplotypes will be referred to by roman numerals: an individual has two unordered haplotypes that will be referred to by means of labels I and II. Examples of the rules are in figure 5 and can be used to clarify the rules given below.

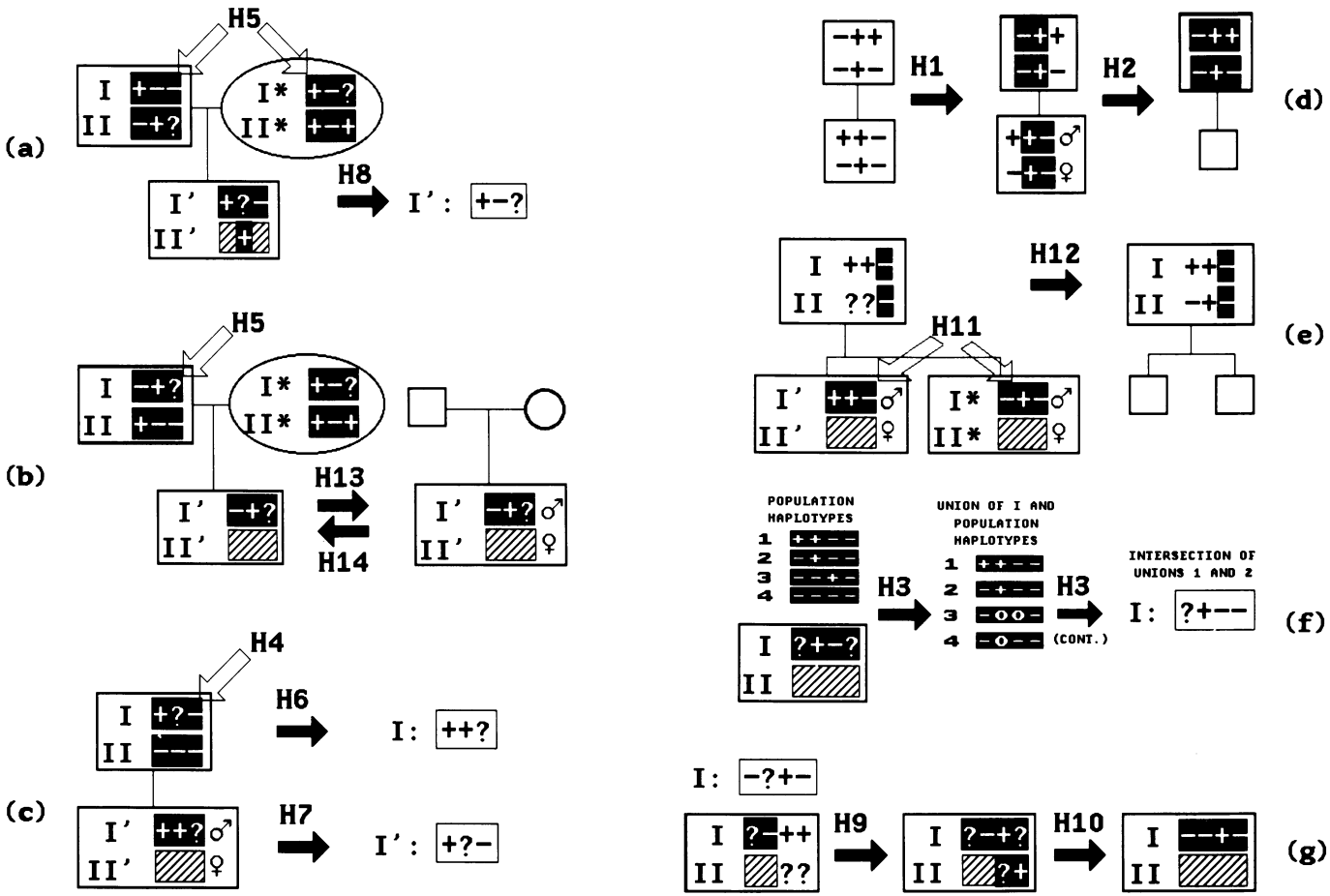
In all individuals, modes for at least one factor can be assigned to haplotypes, although such haplotypes may be anonymous. The first two rules of haplotyping (fig. 5[*d*]) ensure the existence of partial haplotypes in all individuals.

Rule H1: For all homozygous factors in an individual: (a) if the individual has no parents, anonymously assign all unassigned homozygous factors; (b) if the individual has at least one parent, assign all unassigned homozygous factors to the parents.

Rule H2: If no nonhomozygous factors in an individual have been assigned to haplotypes and either an individual has no parents or the haplotypes in the individual are anonymous, make one of the two following changes: (a) if a heterozygous factor exists, anonymously assign the modes of one heterozygous factor; (b) if no heterozygous factor exists, anonymously assign the modes of one factor with one known mode, if such a factor exists.

After H2 is applied, one of three rules must next be applied—H3, H4, or H5 (see fig. 3). These three rules embark on the general procedure used to increase information on haplotypes—i.e., create a template that describes as much as possible about a haplotype and then use the template to modify the haplotype. The three rules differ primarily in the source of information used to find the template. However, although H3 (fig. 5[*f*]) actually finds the template by identifying those modes that are in common to all alleles that are consistent with the haplotype in question, H4 (fig. 5[*c*]) and H5 (figs. 5[*a*], 5[*b*]) only identify

FIG. 5.—Examples of rules of haplotyping. Symbols are as in fig. 1. Black arrows represent rules that change modes or assignments to haplotypes or that produce templates. White arrows represent rules used to identify haplotypes. (*a*), H5 identifies I and I* as both possibly giving rise to I′. H8 uses I and I* to find a template for I′. (*b*), H5 identifies I as possibly giving rise to I′ but fails to identify I* or II* as giving rise to I′. H13 assigns the origin of I′ to the father. H14 makes the haplotypes in the child anonymous. (*c*), H4 identifies I as giving rise to I′. H6 and H7 then find templates for I and I′. (*d*), H1 assigns homozygous factors to haplotypes, and H2 adds a single heterozygous factor to the father's anonymous haplotypes. (*e*), H11 identifies I′ and I* as being different paternal haplotypes. H12 then changes unknown modes in the father to known modes. (*f*), H3 takes a complete set of known population haplotypes and an individual with a partially known haplotype to find a template for I. A zero on union haplotypes represents the null mode, indicating an impossible union. (*g*), H9 uses a template for I to assign modes to I and II, and H10 uses the template to change unknown modes to known modes.


(a)
H5
I +−−
II −+?
I* +−?
II* +−+
I' +?−
II' +
H8
I': +−?
(b)
H5
I −+?
II +−−
I* +−?
II* +−+
I' −+?
II'
H13
H14
I' −+? ♂
II' ♀
(c)
H4
I +?−
II −−−
H6
I: ++?
I' ++? ♂
II' ♀
H7
I': +?−
(d)
−++
−+−
H1
++−
−+−
−++
−+−
H2
−++
−+−
(e)
I ++
II ??
H11
I' ++ ♂
II' ♀
I* −+ ♂
II* ♀
H12
I ++
II −+
(f)
POPULATION HAPLOTYPES
1 + + − −
2 − + − −
3 − − + −
4 − − − −
I ?+−?
II
H3
UNION OF I AND POPULATION HAPLOTYPES
1 + + − −
2 − + − −
3 − o o −
4 − o − −
(CONT.)
H3
INTERSECTION OF UNIONS 1 AND 2
I: ?+−−
(g)
I: −?+−
I ?−++
II ??
H9
I ?−+?
II ?+
H10
I −−+−
II

haplotypes that are passed from parent to child and are subsequently used to find templates.

Rule H3: Suppose that (1) a group of n factors defines a set of known haplotypes or subhaplotypes in a population, (2) a set of haplotypes or subhaplotypes is known in a population, and (3) a partial, full, or complete haplotype, I, exists in an individual. Let P_i be the ith haplotype in the set of population haplotypes, and let U_i be the union of P_i and the observed haplotype. The jth mode on U_i is the jth mode on P_i if (a) the jth mode in the individual is assigned to I and is either unknown or is the same as the jth mode on P_i or (b) the jth mode is unassigned in the individual. If the mode is known and assigned to I and is different than the mode on P_i, the jth mode on U_i is null (see mode 2 for haplotype 4; fig. 5[*f*]). The template for I consists of the intersection of all the U_i that contain no nulls for any of the n factors: if the jth mode of all these U_i is known and identical, the jth mode of the template is this mode. Otherwise, the jth mode of the template is unknown.
Error check: If all U_i contain at least one null, no known haplotypes are consistent with the observed haplotype I; so an error must exist.

Rules H4 and H5 find haplotypes in relatives who will be used to make templates. Rule H4 finds a haplotype in a parent that *must* have been passed to a child, and rule H5 finds a haplotype that *may* have been passed to a child. These are the two most difficult rules, requiring knowledge of modes on at least one child's haplotype and on both the parents' haplotypes.

Rule H4: I in the parent must be identical to I′ in the child if (1) there is a factor such that the factor has been assigned to parental haplotypes in the child and to haplotypes in the parent, (2) the origin of I′ is this parent, (3) the factor is not homozygous in the parent, (4) II in the parent contains a known mode for this factor, and (5) the mode on I′ is known and different from the mode on II.

For example, in figure 5(*c*), the origin of I′ is the father, the first two modes on I′ are +, and the first two modes on II are −. Therefore, II could not have given rise to I′; so I′ came from I.

Rule H5: A factor identifies I in the parent as possibly giving rise to anonymous I′ in the child if the parent is not homozygous for the factor and either (a) the mode on I′ is known and is either identical to a known mode on I or different from a known mode on II or (b) the mode on I′ is unknown and the mode on II′ is known and is different from the known mode on I. Then I may have given rise to I′ if either all factors that satisfy (a) or (b) identify I as the parental haplotype or if the parent is homozygous for all assigned factors and all modes that are known and assigned in both parent and child are the same on I′ as they are on I (see figs. 5[*a*], 5[*b*]).

The next three rules describe how to use rules H4 and H5 in combination with rules H6–H8 to derive templates from relatives' haplotypes. Rules H6 and H7 require a single parent and child; rule H8 requires two parents. Examples of these rules are given in figures 5(*a*) and 5(*c*).

Rule H6: If (a) rule H4 identifies haplotype I in a parent as being I′ in a child or (b) rule H5 identifies I in a parent as being I′ in a child and H5 does not identify II in a

parent as being I in a child, then a template for I can be found such that the mode on the template is the mode on I′ for all assigned, known, modes on I′ (the mode on the template otherwise remaining unknown).

Rule H7: If rule H4 identifies I in a parent as being passed to I′ in a child or if the origins of the haplotypes in the child are the parents and all assigned factors in the parent are homozygous, a template for I′ can be made from I such that the mode on the template is the mode on I for all factors that are assigned to I (the mode for all other factors on the template remaining unknown).

Rule H8: If rule H5 identifies I and I* in the father and mother, respectively, as each *possibly* giving rise to I′ in the child and also identifies II and II* as not possibly giving rise to I′, then a template can be found for I′. For a given factor, the mode on the template is the mode on I if the factor is assigned on both I and I* and if the mode is known and identical on I and I*; otherwise, the mode on the template is unknown.

Once a template has been identified for a particular haplotype with rule H3 or rules H6–H8, rules H9 and H10 use the information on the template to change the information about the haplotype. An example of the application of these two rules is shown in figure 5(*g*).

Rule H9: If a template exists for I in an individual, make the following changes for unassigned factors that have known modes on the template: (1) If there is a known, unassigned mode that is identical to that on the template, assign the mode to I. (2) If there is a known, unassigned mode that differs from that on the template, assign the mode to II.
Error check: If a mode is known and assigned to I and the mode on the template is known and different from that on I, an error exists.

Rule H10: If a template exists for I in an individual and the template was not found with rule H6(b), for each mode that is unknown but assigned to I change the mode to that on the template.

There remain four additional rules of haplotyping. The first pair of rules is concerned with using the haplotypes in a pair of children to increase information about parental genotypes. Unlike the previous several rules, this does not involve making a template, but it does require determining whether two children received different haplotypes from a given parent (rule H11) and then changing unassigned, unknown modes (rule H12) (see fig. 5[*e*]).

Rule H11: If the origin of I′ in child 1 is a parent and the origin of II′ in child 2 is the same parent, then I′ and II′ are known to be different haplotypes if there exists a factor that is assigned on both I′ and II′ and that has different known modes on I′ and II′.

Rule H12: If the origin of I′ in child 1 is a parent and the origin of II′ in child 2 is the same parent and if I′ and II′ are different haplotypes, then, for any factor that has known and assigned modes on both I′ and II,′ make the following changes to the modes for the factor on the parental haplotypes: (a) If the mode on I′ is identical to the mode on II,′ change any unknown modes for the factor in the parent to the mode on I′. (b) If the known modes on I′ and II′ differ and exactly one mode in the parent is unknown, change the unknown mode in the parent to the mode that is different from the known mode. (c) If the known modes on I′ and II′ differ and both

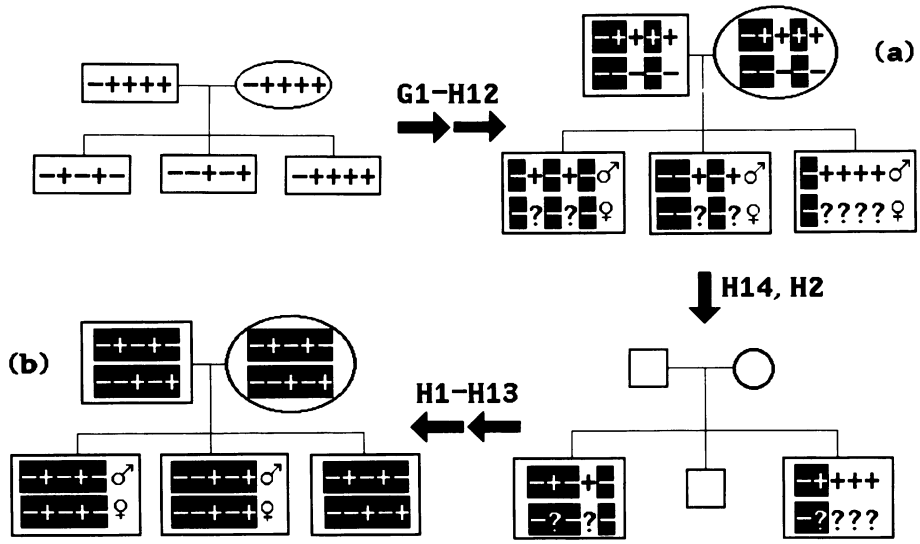


FIG. 6.—Example of the importance of rule H14. Symbols are as in fig. 1. (*a*), Endpoint of analysis without rule H14; (*b*), endpoint of analysis with rule H14.

modes are unknown in the parent, change one unknown mode to be dominant and the other to be recessive.
Error check: If the modes on I′ and II′ are identical and one mode in the parent is known and differs from that on I′ and II′, an error exists.

Finally, the last two rules, rules H13 and H14, can be used to change the assignment of haplotypes (fig. 5[*b*]).

Rule H13: If I′ in a child is anonymously assigned, it can be assigned to a particular parent either if (a) rule H5 fails to find a haplotype in the other parent that could have given rise to I′ or (b) all assigned factors in the child are homozygous.

Rule H14: If the assignment of haplotypes in an individual is to the parents, the assignments may be made anonymous.

Rule H14 appears to be innocuous, but if it is not included in the set of rules, haplotype analysis of a situation such as that depicted in figure 6 only reaches point (a), whereas with H14 the analysis reaches point (b).

DISCUSSION

The rules presented in the present paper derive haplotypes through logical rules rather than through statistical procedures. As a result, the computer program PATCH, which implements the rules, requires very little time and space, enabling the approach to be used for finding haplotypes, determining linkage disequilibria, and determining modes of inheritance such as codominance relationships among alleles. For example, codominance relationships are identified

by noting that modes for some subset of the total set of factors are not found in all possible arrangements on the haplotypes. The approach described here also may produce incomplete haplotypes, since ambiguities are not resolved through these methods. These partial data are still useful: individuals for whom no phenotypic data have been collected may be assigned partial haplotypes; the partial haplotypes can be subjected to a likelihood fit to produce complete haplotypes more efficiently than current implementations of the likelihood approach; and the results may suggest an efficient strategy for collecting additional data. It may sometimes also be possible to complete haplotypes if the assumption of complete dominance is relaxed. For example, for some RFLPs heterozygotes may be distinguishable from homozygotes on the basis of density scans.

A second important result of the minimal necessary computer resources is the possibility of using the rules to determine the most likely violation of one of the first five assumptions when application of the rules exposes an error in the analysis. Because the rules can be used for error analysis, although the assumptions behind haplotype analysis appear to be very strict, inclusion of the full set of assumptions actually introduces more flexibility and accuracy into the analysis than if, say, the assumption of no recombination were removed but the other assumptions (those necessary for standard linkage analysis, as have been described, e.g., by Ott [1985]) were left the same. This is true for five reasons: (1) The stimulus for determining haplotypes is most often the expectation that the factors under analysis are closely linked, e.g., are members of a gene family or RFLPs plus protein markers for the same gene. Under such circumstances, recombination may be less likely than either mistyping of data or some other source of error. (2) When haplotypes are desired for determining linkage disequilibrium, such disequilibrium is only expected to be significant when there is little recombination in the system. (3) Appropriate use of the procedure allows determination of individual linkage groups and their associated haplotypes when recombination does occur. (4) It is useful to be able to identify the most probable violations of assumptions, so that true errors can be corrected. (5) Identification of previously unknown haplotypes is possible.

Techniques for using the rules for error analysis consist of two parts: (1) identification of the most probable cause of the error and (2) analysis of the error. For the case of recombination, the procedure of analysis will be described below. Both parts of the error analysis consist of removing information from the data (pedigree links and/or typings of individual factors) followed by reanalysis to determine whether the error disappears.

When an error in analysis appears, it is reasonable to first test the possibility that there has been a mistyping. To do this, as well as to investigate the possibility of either incorrect identification of paternity or mutation, it is necessary to determine the individual or individuals in the pedigree who are the source of the problem. Such individuals can be isolated either by cutting some or all links between nuclear families, or by sequentially removing individuals from the family, or by both, followed by reanalysis. A single error in a pedigree

will be isolated to a single pair of parents and their children (possibly a specific subset of children) by noting when the modified data no longer give errors in analysis. If the error is due to mistyping of a factor, this can then be determined by sequentially changing single typings in the appropriate individual(s) to unknown typings, again followed by reanalysis.

If recombination is the likely source of error—e.g., because errors are common in a large number of families—individual linkage groups can be determined. Haplotype analysis of all pairs of factors in all families will produce a matrix, the elements of which indicate whether an error in analysis has been detected in one or more families for each pair of factors. If there are any trios of factors that in pairwise analysis fail to produce errors in construction of haplotypes, these trios can be analyzed together to determine whether they are compatible in haplotype analysis, etc. Eventually, a family of sets will exist in which each set contains a group of factors that are mutually compatible with inheritance as a single gene or group of tightly linked genes; the number of sets is a minimum estimate of the number of linkage groups, and the individual combinations that constitute alleles or haplotypes is obtained from the results of haplotype analysis of each set. It therefore may be possible, for complex systems such as that described by Jeffries et al. (1985), to delineate sets of fragments that belong to individual dispersed genes.

The situations in which the approach described in the present paper is most useful are the situations that are the worst for application of a likelihood approach: large pedigrees segregating for many factors, especially those pedigrees in which the components of the haplotypes may be in linkage disequilibrium. In such situations, likelihood techniques require intolerable amounts of time and computer memory. To reach a solution for a large problem, the current approach requires that the computer resources required be increased only a small amount over those required by a smaller problem. As a result, this approach is amenable to use with microcomputers. In many cases, further analyses will be unnecessary, but, when desired, a likelihood or Bayesian approach may be able to fill in the unknowns to produce the most likely haplotypes segregating in the pedigrees (although such haplotypes are subject to change if new data become available). Investigation of the use of statistical methods to extend the results that have been obtained with the methods presented in the present paper is in progress. Since the aim of linkage studies is to find systems that have many alleles or haplotypes, the approach to haplotype analysis discussed here may provide the tools necessary for rapid haplotype construction and error analysis.

ACKNOWLEDGMENTS

I wish to thank Drs. M. J. Johnson and L. L. Cavalli-Sforza for discussions that initially defined the problem and Drs. L. L. Cavalli-Sforza and M.-C. King for suggestions on the manuscript. This work was partially supported by National Institutes of Health grant GM28428.

APPENDIX

DEFINITIONS OF TERMS AND SYMBOLS USED IN DESCRIPTION OF RULES

Factor: a single, observable, discrete trait
- *Mode:* presence or absence of factor received from parent
  - *Dominant mode:* presence of factor
  - *Recessive mode:* absence of factor
  - *Known mode:* known to be dominant or recessive
  - *Unknown mode:* not known to be dominant or recessive
  - *Assigned mode:* known to be on a particular haplotype
  - *Unassigned mode:* not known to be on a particular haplotype
  - *Origin of mode:* origin of haplotype containing assigned mode
- *Phenotype:* observed state of factor (presence/absence/unknown)
- *Genotype:* set of two modes for a factor
  - *Homozygous:* both modes known and identical
  - *Heterozygous:* both modes known and different
- +: dominant mode or presence of factor in phenotype
- −: recessive mode or absence of factor in phenotype
- ? : unknown mode or state of factor in phenotype

Haplotype: a set of coinherited modes
- *Maternal haplotype:* haplotype received from mother
- *Paternal haplotype:* haplotype received from father
- *Parental haplotype:* haplotype known to be paternal or maternal
- *Anonymous haplotype:* haplotype not known to be paternal or maternal
- *Origin of haplotype:* source of haplotype—father, mother, or anonymous
- *Full haplotype:* haplotype to which known or unknown modes are assigned for all factors
- *Complete haplotype:* haplotype to which known modes are assigned for all factors
- *Partial haplotype:* haplotype to which one or more factors do not have a mode assigned
- I, II, I*, II*, I′, II′: labels for haplotypes; I and II are within an individual and * and ′ distinguish individuals.

REFERENCES

Baur, M. P., N. Neugebauer, and M. Sigmund. 1985. Genetic analysis workshop III: multipoint linkage analysis using FAP. Genet. Epidemiol. **2**:201–202.

Baur, M. P., M. Neugebauer, M. Sigmund, and J. Willems. 1984. FAP—family analysis program. Cytogenet. Cell. Genet. **37**:416.

Botstein, D., R. White, M. Skolnick, and R. W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. **32**:314–331.

Bufton, L., G. A. P. Bruns, R. E. Magenis, D. Tomar, D. Shaw, D. Brook, and M. Litt. 1986. Four restriction fragment length polymorphisms revealed by probes from a single cosmid map to chromosome 19. Am. J. Hum. Genet. **38**:447–460.

Buroker, N. E., R. E. Magenis, K. Weliky, G. Bruns, and M. Litt. 1986. Four restriction fragment length polymorphisms revealed by probes from a single cosmid map to human chromosome 12q. Hum. Genet. **72**:86–94.

Chakravarti, A., S. C. Elbein, and M. A. Permutt. 1986. Evidence for increased recombination near the human insulin gene: implication for disease association studies. Proc. Natl. Acad. Sci. USA **83**:1045–1049.

Chakravarti, A., J. A. Phillips III, K. H. Mellits, K. H. Buetow, and P. H. Seeburg. 1984. Patterns of polymorphism and linkage disequilibrium suggest independent origins of the human growth hormone gene cluster. Proc. Natl. Acad. Sci. USA **81**:6085–6089.

Cohen-Haguenauer, O., E. Robbins, C. Massart, M. Busson, I. Deschamps, J. Hors, J.-M. Lalouel, J. Dausset, and D. Cohen. 1985. A systematic study of HLA class II-β DNA restriction fragments in insulin-dependent diabetes mellitus. Proc. Natl. Acad. Sci. USA **82**:3335–3339.

Elbein, S. C., L. Corsetti, A. Ullrich, and M. A. Permutt. 1986. Multiple restriction fragment length polymorphisms at the insulin receptor locus: a highly informative marker for linkage analysis. Proc. Natl. Acad. Sci. USA **83**:5223–5227.

Eng, C. E. L., and C. M. Strom. 1985. Analysis of three restriction fragment length polymorphisms in the human type II procollagen gene. Am. J. Hum. Genet. **37**:719–732.

Hasstedt, S. J. 1982. A mixed-model likelihood approximation on large pedigrees. Comp. Biomed. Res. **15**:295–307.

Hasstedt, S. J., and P. E. Cartwright. 1979. PAP: pedigree analysis package. Tech. rep. no. 13. Department of Biophysics and Computing, University of Utah, Salt Lake City.

Higgs, D. R., J. S. Wainscoat, J. Flint, A. V. S. Hill, S. L. Thein, R. D. Nicholls, H. Teal, H. Ayyub, T. E. A. Peto, A. G. Falusi, A. P. Jarman, J. B. Clegg, and D. J. Weatherall. 1986. Analysis of the human α-globin gene cluster reveals a highly informative genetic locus. Proc. Natl. Acad. Sci. USA **83**:5165–5169.

Jeffries, A. J., V. Wilson, and S. L. Thein. 1985. Hypervariable 'minisatellite' regions in human DNA. Nature **314**:67–73.

Johnson, M. J. 1984. Molecular genetics of the human immunoglobulin heavy chain genes: analysis of DNA polymorphisms. Ph.D. diss., Stanford University, Stanford, CA.

Johnson, M. J., G. de Lange, and L. L. Cavalli-Sforza. 1986. Ig gamma restriction fragment length polymorphisms indicate an ancient separation of Caucasian haplotypes. Am. J. Hum. Genet. **38**:617–640.

Julier, C., M. Lathrop, J. M. Lalouel, A. Reghis, M. F. Szajnert, and J. C. Kaplan. 1985. New restriction fragment length polymorphisms on human chromosome 22 at loci SIS, MB and IGLV. Cytogenet. Cell Genet. **40**:664.

Larsen, S. O. 1979. A general program for estimation of haplotype frequencies from population diploid data. Comp. Prog. Biomed. **10**:48–54.

Lathrop, G. M., and J. M. Lalouel. 1984. Easy calculations of lod scores and genetic risks on small computers. Am. J. Hum. Genet. **36**:460–465.

Le Gall, I., A. Marcadet, M.-P. Font, C. Auffray, J. Strominger, J.-M. Lalouel, J. Dausset, and D. Cohen. 1985. Exuberant restriction fragment length polymorphism associated with the DQ α-chain gene and the DX α-chain gene. Proc. Natl. Acad. Sci. USA **82**:5433–5436.

Lidsky, A. S., F. D. Ledley, A. G. DiLella, S. C. M. Kwok, S. P. Daiger, K. J. H. Robson, and S. L. C. Woo. 1985. Extensive restriction site polymorphism at the human phenylalanine hydroxylase locus and application in prenatal diagnosis of phenylketonuria. Am. J. Hum. Genet. **37**:619–634.

Litt, M., N. E. Buroker, L. Bufton, and C. Maslen-McClure. 1985. DNA polymorphisms via cosmid probes. Cytogenet. Cell Genet. **40**:682.

Litt, M. and R. L. White. 1985. A highly polymorphic locus in human DNA revealed by cosmid-derived probes. Proc. Natl. Acad. Sci. USA **82**:6206–6210.

Migone, N., J. Feder, H. Cann, B. van West, J. Hwang, N. Takahashi, T. Honjo, A. Piazza, and L. L. Cavalli-Sforza. 1983. Multiple DNA fragment polymorphisms associated with immunoglobulin μ chain switch-like regions in man. Proc. Natl. Acad. Sci. USA **80**:467–471.

Ohlsson, M., J. Feder, L. L. Cavalli-Sforza, and A. von Gabain. 1985. Close linkage of α and β interferons and infrequent duplication of β interferon in humans. Proc. Natl. Acad. Sci. USA **82**:4473–4476.

Ott, J. 1985. Analysis of human genetic linkage. The Johns Hopkins University Press, Baltimore.

Steinberg, A. G., and C. E. Cook. 1981. The distribution of the human immunoglobulin allotypes. Oxford University Press, New York.

Willard, H. F., M. H. Skolnick, P. L. Pearson, and J.-L. Mandel. 1985. Report of the committee on human gene mapping by recombinant DNA techniques. Cytogenet. Cell Genet. **40**:360–489.