Analysis and Exploration of the Use of Rule-Based Algorithms and Consensus Methods for the Inferral of Haplotypes

Steven Hecht Orzack,*[,1] Daniel Gusfield,[†] Jeffrey Olson,[‡] Steven Nesbitt,[‡] Lakshman Subrahmanyan[‡] and Vincent P. Stanton, Jr.[‡]

**Fresh Pond Research Institute, Cambridge, Massachusetts 02140, [†]Department of Computer Science, University of California, Davis, California 95616 and [‡]Variagenics, Cambridge, Massachusetts 02139*

Manuscript received January 13, 2003
Accepted for publication June 20, 2003

ABSTRACT

The difficulty of experimental determination of haplotypes from phase-unknown genotypes has stimulated the development of nonexperimental inferral methods. One well-known approach for a group of unrelated individuals involves using the trivially deducible haplotypes (those found in individuals with zero or one heterozygous sites) and a set of rules to infer the haplotypes underlying ambiguous genotypes (those with two or more heterozygous sites). Neither the manner in which this "rule-based" approach should be implemented nor the accuracy of this approach has been adequately assessed. We implemented eight variations of this approach that differed in how a reference list of haplotypes was derived and in the rules for the analysis of ambiguous genotypes. We assessed the accuracy of these variations by comparing predicted and experimentally determined haplotypes involving nine polymorphic sites in the human apolipoprotein E (APOE) locus. The eight variations resulted in substantial differences in the average number of correctly inferred haplotype pairs. More than one set of inferred haplotype pairs was found for each of the variations we analyzed, implying that the rule-based approach is not sufficient by itself for haplotype inferral, despite its appealing simplicity. Accordingly, we explored consensus methods in which multiple inferrals for a given ambiguous genotype are combined to generate a single inferral; we show that the set of these "consensus" inferrals for all ambiguous genotypes is more accurate than the typical single set of inferrals chosen at random. We also use a consensus prediction to divide ambiguous genotypes into those whose algorithmic inferral is certain or almost certain and those whose less certain inferral makes molecular inferral preferable.

THEORETICAL and practical considerations suggest that a more complete causal understanding of many complex traits, such as human diseases, will often be gained by haplotype analysis, since such traits may often be the partial result of many genetic determinants (Chakravarti 1999; Judson *et al.* 2000; Hartman *et al.* 2001; Reich *et al.* 2001).

Since molecular inferral of haplotypes is difficult (*e.g.*, see Michalatos-Beloin *et al.* 1996), nonmolecular methods for haplotype inferral have been developed. One can distinguish three different approaches for a group of unrelated individuals. First, Clark (1990) presented a "rule-based" approach in which a list of haplotypes and a set of rules are used to infer the haplotypes underlying each of the ambiguous genotypes. Second, one can determine the maximum-likelihood solution, that is, the set of haplotype frequencies that generates the largest likelihood given the observed genotype frequencies. Present implementations of this approach rely upon the expectation-maximization algorithm (Templeton *et al.* 1988; Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long *et al.* 1995; Fallin and Schork 2000). This approach uses the assumption that the genotypes come from a population at or near Hardy-Weinberg equilibrium to infer the expected frequencies of possible haplotype pairs that underlie ambiguous genotypes. This assumption also underlies the related Bayesian approach of Niu *et al.* (2002). Finally, several approaches rely upon more substantive population genetics to infer haplotypes. For example, Stephens *et al.* (2001a) use the neutral coalescent model of sequence evolution as a basis for their method (see also Lin *et al.* 2002); this model implies that the population is at or near Hardy-Weinberg equilibrium.

Here, we analyze the accuracy of the rule-based approach when applied to a sample of genotypes with known haplotypes at the human apolipoprotein E (APOE) locus (see Fullerton *et al.* 2000 for detailed information on this locus). An important motivation for our work is that Clark's method is described as a plausible inferral method in many of the nearly 100 articles in which his article has been cited (Science Citation Index 2003). In some articles, Clark's method is described as "meritorious" (*e.g.*, Niu *et al.* 2002) or "effective" (*e.g.*, Zhu *et al.* 2001) in regard to haplotype inferral. Despite this attention, the accuracy of this method has

[1]*Corresponding author:* Fresh Pond Research Institute, 173 Harvey St., Cambridge, MA 02140. E-mail: orzack@freshpond.org

never been systematically assessed, especially with a data set containing real haplotype pairs.

MATERIALS AND METHODS

Subjects studied: Genomic DNA from each of 31 cell lines from the Coriell Cell Respository was sequenced to identify polymorphisms in the APOE locus. Using the data on race and ethnicity at the Coriell web site, we classified 8 cell lines as derived from Asian individuals, 8 from Blacks, and 15 from Caucasians. Some of the cell lines were from individuals with rare inborn errors of metabolism, although none suffered from a condition known to be associated with their APOE genotype. Of the 80 individuals whose haplotypes were phased, we classified 18 as Asians, 19 as Blacks, and 43 as Caucasians. All individuals were unrelated.

DNA sequencing and genotyping: Polymorphisms were discovered in the 31 cell lines by DNA sequencing using an ABI 3700 DNA sequencer and Polyphred software for identification of candidate polymorphic sites, substantially as described by Nickerson *et al.* (1997, 2000). Subsequently, primer extension (minisequencing) genotyping assays (as described by Dracopoli *et al.* 2001) were developed for each polymorphic site. The products of minisequencing reactions were analyzed by gel electrophoresis on an ABI 377 DNA sequencer.

Experimental haplotype determination: Haplotypes were determined first by using the genotype data to identify the 5′ flanking heterozygous site for each phase-unknown individual. (Sequence orientation is determined relative to the APOE locus.) Second, each of the two alleles of such individuals was amplified in a PCR reaction primed with an allele-specific primer at the 5′ flanking heterozygous site and a constant (non-allele-specific) primer located just 3′ of exon 4. Third, the haplotypes of the two allele-specific amplicons from each individual were determined by genotyping all internal heterozygous sites. A pair of allele-specific primers was designed and tested for all sites except the most 3′ site, 21388, which is next to the constant primer and therefore cannot circumtend two heterozygous sites. Primer design and PCR conditions were optimized for maximal allele selectivity at each site using known pairs of haplotypes. The sequence of both alleles and the genotyping results provided independent sources of information for all haplotype inferrals. Primer sequences are available upon request.

The rule-based approach—Clark (1990): We first define some terms. Any inferral method "resolves" an ambiguous genotype when it chooses a "resolution," that is, a pair of underlying haplotypes for an ambiguous genotype. A genotype left unresolved is an "orphan."

Clark (1990) described an inferral method that is appealing because it relies on unambiguous haplotypes and has weak assumptions about population structure and size. Once such haplotypes are identified, one proceeds as follows (Clark 1990, p. 112):

> For each known haplotype, we then look at all the remaining unresolved [genotypes] and ask whether the known haplotype can be made from some combination of the ambiguous sites. Each time such a haplotype is found, we immediately recover the complement of the haplotype as another potential haplotype. This chain of inference continues until all haplotypes have been recovered, or until no more haplotypes can be found.

Clark also proposed that this procedure be run more than once with the data being randomly reordered at the start each time so as to determine whether distinct resolutions would be generated for any given ambiguous genotype. Following Clark, we refer to the ensemble of pairs of inferred haplotypes associated with a given reordering as a "solution" for the genotypic data. On the basis of such iterated calculations, Clark (pp. 117–118) went on to describe an important claim: ". . . the solution with the fewest orphans is the valid solution and suggests that when a solution resolves all haplotypes it is likely to be unique." We interpret this claim to be the prediction that the set of inferred haplotypes that resolves the most individuals is the most accurate solution; that is, a higher percentage of its inferrals is correct than for any other solution. This is an important claim because it implies that the single, most accurate solution can be identified using this approach. However, despite this suggestion by Clark to generate and compare multiple solutions, it appears that no subsequent analysis is so structured.

The rule-based approach—a general overview: We define the list of "real" haplotypes as those obtained by serially inspecting the ordered set of input genotypes and adding the haplotypes of unambiguous genotypes to the end of a list. How duplicates are handled is discussed below.

During haplotype inferral, haplotypes derived from ambiguous genotypes might be added to the list of real haplotypes; this larger list is the "reference list" of haplotypes. Each iteration of a given variation began with a random reordering of the list of genotypes and the creation of the list of real haplotypes. Haplotype inferral proceeded in either of two ways. In the first, proceeding downward from the genotype at the top of the list, we scanned down the reference list of haplotypes to find one that could resolve the genotype in question. In the second, proceeding downward from the haplotype at the top of the reference list, we scanned down the list of ambiguous genotypes to find one that could be resolved by the haplotype in question. In either case, if a resolution was found, we then searched for additional resolutions. We either chose the first haplotype found as the basis for the resolution of the genotype in question or chose among all of the haplotypes that could resolve it (see below). Finally, we decided whether the chosen haplotype and/or its complement haplotype was added to the reference list.

Thus, each iteration of our implementation of the rule-based approach consisted of either a haplotype selection loop, which in turn contained a genotype selection loop, or a genotype selection loop, which in turn contained a haplotype selection loop. In a given iteration, one of these pairs of loops was executed until no further genotypes could be resolved. We found no systematic differences between the results stemming from the two methods; we present results based upon the second.

We developed eight variations of the rule-based approach that differ in the way that the initial list of real haplotypes and the reference haplotype list are formulated and in how the list of genotypes is analyzed. The development of these variations was motivated by questions that a user of the rule-based approach must answer before using it:

1. Does the reference list include just real haplotypes or inferred ones as well?
2. Are duplicate haplotypes removed or retained in the reference list at the beginning of each iteration? If they are removed, the initial list is like a phone book, with a unique "name" for each haplotype. If they are retained, they are represented as distinct copies.
3. Are duplicate ambiguous genotypes "consolidated" at the beginning of each iteration or left separate? If they are consolidated, identical genotypes will be resolved identically. If they are left separate, identical genotypes may be resolved differently, except when the reference list is not randomized (see below).

TABLE 1

The characteristics of the eight rule-based algorithms for haplotype inferral

Variation	Duplicate haplotypes: Real	Duplicate haplotypes: Inferred	Haplotype list randomized?	Ambiguous genotypes consolidated?	Preference for real haplotypes?	Frequency preference	Identical genotypes
1	Retained	Retained	Yes	No	No	Population	May be resolved differently
2	Retained	Retained	Yes	Yes	No	Population	Resolved identically
2a	Removed	Retained	Yes	Yes	No	Weak	Resolved identically
2b	Retained	Retained	No	No	Yes	Population	Resolved identically
2c	Removed	Retained	No	Yes	Yes	No	Resolved identically
3	Retained	Retained	No	No	No	Strong	May be resolved differently
4a	Removed	Removed	Yes	Yes	No	No	Resolved identically
4b	Removed	Removed	No	Yes	Yes	No	Resolved identically

4. Before trying to resolve an ambiguous genotype, it is important to ask is the current reference list of haplotypes randomized or not? Since the list is scanned from the top, reordering or not may determine whether a particular haplotype and its complement are used in the resolution.
5. If multiple haplotypes can resolve a given ambiguous genotype, is the haplotype chosen as the "resolving" haplotype the first one found, the one with the currently highest frequency, or one chosen randomly?
6. After a genotype is resolved by a given haplotype, how is the reference list of haplotypes updated? We used two alternatives. In the first, the resolving haplotype's complement haplotype is added to the list only if it is new, thereby maintaining the list as a phone book of names. In the second, a resolving haplotype and its complement are added to the reference list, thereby allowing any subsequent choice among competing resolving haplotypes for an ambiguous genotype to be influenced by their population frequencies.

In all of our variations, we assumed that inferred haplotypes are included on the reference list. We focused on how the answers to questions 2, 3, 4, 5, and 6 affect inferral accuracy. None of these questions has one right answer; for example, the decision as to whether identical genotypes should be resolved with the same haplotypes comes down to a judgment as to the prevalence of homoplasy.

The choices made for each variation in regard to the treatment of duplicate haplotypes, haplotype list randomization, and consolidation of ambiguous genotypes are shown in Table 1. These choices have consequences in regard to the choice of real *vs.* inferred haplotypes, the choice of more frequent haplotypes, and the resolution of identical ambiguous genotypes. For example, variation 1 is perhaps the simplest inferral algorithm: duplicate haplotypes were retained on the reference list, which therefore included an entry for each haplotype derived from unambiguous genotypes and resolved ambiguous genotypes. The order of the haplotypes on the reference list was randomized before each iteration. Duplicate ambiguous genotypes were not consolidated. The resolving haplotype was selected randomly from among all possible resolving haplotypes. These choices generated a "population" frequency preference: a more common resolving haplotype was likely chosen more often than a less common resolving haplotype, although identical genotypes might not be resolved identically.

Comments on the other variations are:

Variation 2: the same as variation 1 except that duplicate ambiguous genotypes were consolidated, which caused them to be resolved identically by the first resolving haplotype found.

Variation 2a: removal of duplicate real haplotypes from the reference list at the beginning of each iteration generated a "weak" frequency preference. Its magnitude was typically less than that of the population frequency preference because the reference list had no initial frequency differences (any duplicate inferred haplotypes were retained).

Variation 2b: lack of randomization of the haplotype reference list caused real haplotypes to have a preference as resolving haplotypes since they are nearer the top of the list than are inferred ones. Identical ambiguous genotypes were resolved identically by the first haplotype that could do so because haplotype order was fixed.

Variation 2c: identical ambiguous genotypes were resolved identically by the first haplotype that could do so because haplotype order was never changed. Consequently, there was a "no" frequency preference.

Variation 3: there was a "strong" frequency preference in that any given ambiguous genotype was solved by the currently most frequent haplotype that could do so. This is the only variation in which frequency preference was an explicit choice.

Variation 4a: the complement of a solution haplotype was added to the reference list only if it was new. Identical genotypes were solved identically because of consolidation.

Variation 4b: the same as variation 4a except that the order of the haplotypes in the reference list was not randomized. There was a preference for real over inferred haplotypes (since they were nearer the top of the reference list). These choices generated a "no" frequency preference. This variant appears to be the approach described by Clark (1990) although he did not provide an explicit algorithm.

One can devise additional rule-based algorithms. For example, one might assess the real or inferred status of a complement haplotype when deciding among resolving haplotypes. However, our present choices allowed us to explore the substantial biological heterogeneity among the eight variations of the rule-based approach to haplotype inferral. So, for example, the variations differed in having no frequency preference (variations 2c, 4a, and 4b), a weak frequency preference (2a), a population frequency preference (variations 1, 2, and 2b), or a strong frequency preference (variation 3). In addition, within either the first or the third group, the variations differed substantially in how a given frequency preference was manifested. For example, variation 1 had a population preference with no priority for real-over-inferred haplotypes. In contrast, variation 2b had a population preference that favored real

TABLE 2

Frequencies of APOE haplotypes in 80 subjects

Haplotype	e2, e3, e4 type[a]	Haplotype frequency[b]	Polymorphic site (base no.)[c]								
			17874	17937	18145	18476	19311	20334	21250	21349	21388
1	e3	53 (33.0)	A	T	G	G	G	G	T	C	C
2	e3	42 (26.0)	A	T	T	C	G	G	T	C	C
3	e3	12 (7.5)	T	T	G	G	G	G	T	C	C
4	e4	10 (6.3)	A	T	T	G	A	G	C	C	C
5	e2	9 (5.6)	A	C	G	G	G	G	T	C	T
6	e3	8 (5.0)	T	T	T	C	G	G	T	C	C
7	e4	5 (3.1)	T	T	T	G	G	G	C	C	C
8	e3	5 (3.1)	A	C	T	C	G	G	T	C	C
9	e2	3 (1.9)	T	T	G	G	G	G	T	C	T
10	e3	3 (1.9)	A	T	G	G	G	G	T	T	C
11	e3	2 (1.25)	A	T	T	G	G	G	T	C	C
12	e3	2 (1.25)	A	T	G	C	G	G	T	C	C
13	e4	2 (1.25)	A	T	G	G	G	G	C	C	C
14	e4	1 (0.62)	T	T	G	G	G	G	C	C	C
15	e4	1 (0.62)	A	T	T	G	G	G	C	C	C
16	e3	1 (0.62)	T	T	G	C	G	G	T	C	C
17	e3	1 (0.62)	A	T	T	C	G	A	T	C	C
		Variant nucleotides (major/minor):[d]	A/T	T/C	G/T	G/C	G/A	G/A	T/C	C/T	C/T
		Minor allele frequency (%):	18.75	8.75	46.3	36.9	6.3	0.63	11.9	1.9	7.5

[a] The e2, e3, e4 classification is determined by genotype at positions 21250 and 21388. e2 is T, T; e3 is T, C; e4 is C, C. There are 2 e2, 10 e3, and 5 e4 haplotype subgroups.
[b] Absolute and relative frequencies of the 17 observed haplotypes in 160 chromosomes. Numbers in parentheses are percentages.
[c] Position of variant nucleotides in GenBank accession no. AF050154.
[d] The more common nucleotide in the sample is listed first and the less common second.

haplotypes. We ran at least two independent sample paths of 10,000 iterations for each variation.

RESULTS

Identification of polymorphisms in the apolipoprotein E locus: A 415-bp segment of the APOE locus extending from 570 bp upstream of the transcription start site to the end of exon 4 (nucleotides 17800–21958 of GenBank accession no. AF050154) was sequenced and polymorphisms were identified at nucleotides 17874, 17937, 18145, 18476, 19311, 19753, 20334, 21250, 21349, and 21388. These polymorphisms have been described previously (Artiga *et al.* 1998; Bullido *et al.* 1998; Nickerson *et al.* 2000) except for the polymorphism at nucleotide 20334, which was identified in a single chromosome. It is a missense polymorphism that changes amino acid 18 of the primary translation product from alanine to threonine. Amino acids 1–18 comprise a leader sequence that is cleaved off to form the mature protein. The cleavage properties of the threonine allele are unknown. Nucleotides 21250 and 21388 correspond to amino acids 112 and 158 of the mature protein and account for the classical e2, e3, and e4 allele classification (see Table 2).

Experimental identification of haplotypes: We were unable to develop reliable allele-specific PCR reactions for one of the 10 polymorphic sites; this site (19753) was omitted from subsequent analyses. As shown in Table 2, a total of 17 haplotypes involving 9 sites were experimentally identified; these range in frequency from 0.62% (1 of 160) to 33.1% (53 of 160). Eleven of the haplotypes can be inferred from unambiguous genotypes.

Of the 80 individuals, 33 have unambiguous pairs of haplotypes and 47 have ambiguous pairs. There are 17 genotypes with two polymorphic sites, 20 genotypes with three such sites, 6 genotypes with four such sites, and 4 genotypes with five such sites. Haplotype pairs for all genotypes are available at http://www.genetics.org.

Genotypic proportions at all nine sites do not deviate significantly ($\alpha = 0.05$) from Hardy-Weinberg proportions: 17874 ($\chi^2 = 1.769$, $P = 0.183$), 17937 (3.775, 0.052), 18145 (0.250, 0.617), 18476 (0.290, 0.590), 19311 (0.356, 0.551), 20334 (0.003, 0.955), 21250 (0.868, 0.352), 21349 (0.029, 0.864), and 21388 (0.786, 0.375).

Computational results: Our first goal was to compare the performances of the variations. The second goal was to determine whether the important inferential principle enunciated by Clark was true, that is, whether the solution with the fewest orphans is the most accurate solution. Finally, we wished to determine whether multi-

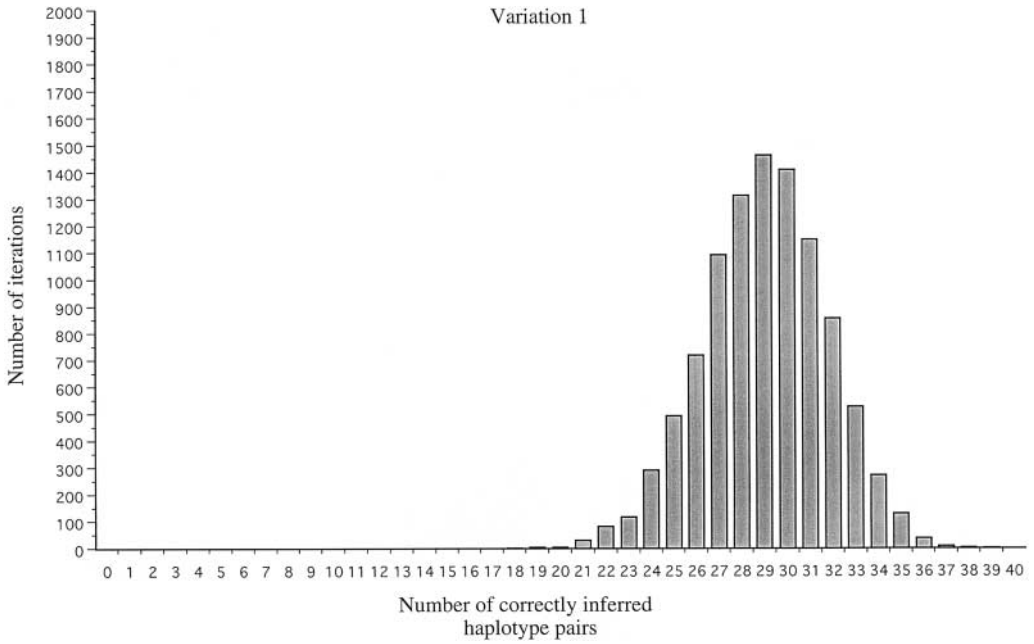

FIGURE 1.—The distribution of the number of correctly inferred pairs of haplotypes for a sample path of variation 1. Only the 47 ambiguous genotypes are considered.

ple complete solutions exist and, if so, to understand how they could be of use.

The frequency distribution of accuracy for a 10,000-iteration sample path of variation 1 is shown in Figure 1. The first important point is that there are "different" solutions, that is, those that differ in the resolution of at least one ambiguous genotype. In fact, each of the 10,000 iterations resulted in a different complete solution of the 47 ambiguous genotypes. None of these solutions had an orphan. It is clear for this locus, then, that the principle enunciated by Clark concerning the accuracy of the solution with the fewest orphans cannot apply since there is no one solution with the fewest orphans. Analyses of two other sets of phase-known genotypes and of simulated genotypic data also resulted in multiple solutions with no orphans (S. ORZACK and D. GUSFIELD, unpublished results). This underscores the potential for mistaken inferrals particularly when an investigator runs a method only once on a given data set. We expect the occurrence of multiple complete solutions to be common when using the rule-based approach.

Multiple distinct solutions were manifested in different ways, as shown in Table 3. In one variation there are only two solutions, in one there are <500, in some there are ~ 2000, and several had 10,000 or close to it. Thus, while the algorithmic differences among the variations at first may seem minor, they led to substantially different outcomes.

The normal-like distribution of accuracy shown in Figure 1 is not typical, as can be seen in Figure 2, which contains the distribution of accuracy for variation 2, and in Figure 3, which contains the distribution of accuracy for variation 3; here, there are only two solutions. This last result reflects the never-decreasing influence of frequency on the process of inferral; this restricts resolutions to a small part of the solution space.

The average accuracy score and its 95% confidence interval for each of the variations are shown in Table 4. It is clear that the variations differed substantially in their predictive accuracy, with variations 2a, 2c, 4a, and 4b performing poorly (36–40% accuracy) and variations 1, 2, 2b, and 3 performing better (62–68% accuracy).

Variations 2c, 4a, and 4b, with no frequency preference, performed poorly. Average accuracy tends to increase with increasing reliance on frequency, although most of this increase is achieved by a population frequency preference as a strong frequency preference added little to average accuracy.

Dealing with multiple solutions: For this data set, all of the rule-based variations generate multiple solutions of quite variable quality. Without an understanding of what sense, if any, can be made of multiple solutions, it is difficult to understand what insights any rule-based algorithm could provide an investigator. However, almost all of the variations generate some solutions in which $>80\%$ of the ambiguous genotypes are correctly inferred. If one could identify such solutions, it would be extremely advantageous.

Consensus methods: These considerations led us to develop consensus methods that are based on the multiple solutions that are present in any given simulation. We evaluated several ways to generate a consensus prediction.

The first way was to tally the different inferrals for any given genotype across all 10,000 iterations of the genotypic data. The inferral appearing in the highest number of iterations was taken to be the "full" consensus prediction for the genotype.

Our knowledge of the real haplotype pairs for each genotype allowed us to determine whether particular

TABLE 3

Number of complete solutions and number of distinct solutions for all variations

Sample path	No. of complete solutions	No. of distinct solutions
Variation 1		
1	10,000	10,000
2	10,000	10,000
Variation 2		
1	10,000	9,792
2	10,000	9,787
3	10,000	9,805
4	10,000	9,822
Variation 2a		
1	10,000	10,000
2	10,000	10,000
Variation 2b		
1	10,000	436
2	10,000	437
Variation 2c		
1	10,000	2,005
2	10,000	1,996
3	10,000	2,049
Variation 3		
1	10,000	2
2	10,000	2
Variation 4a		
1	10,000	10,000
2	10,000	10,000
3	10,000	10,000
4	10,000	10,000
Variation 4b		
1	10,000	2,015
2	10,000	2,022

types of solutions are more likely to have a higher number of correct inferrals. To generate a second kind of consensus prediction we used information on the number of haplotypes needed to resolve all of the ambiguous genotypes or to resolve all of the genotypes.

A plot of the relationship between the number of different haplotypes used in a solution and the average number of correct inferrals is shown in Figure 4. There is a strong negative relationship between the two, implying that smaller lists perform better than larger lists (Spearman rank correlation corrected for ties, ambiguous genotypes, -0.996, 13 d.f., $P < 0.001$; all genotypes, -1.0, 13 d.f., $P < 0.001$). This result occurred for all of our variations for this data set and in the analysis of two other phase-known data sets (S. Orzack and D. Gusfield, unpublished results). We surmise that the underlying reason for the negative relationship between list size and accuracy is that typical mutation and recombination rates do not substantially increase the expected number of haplotypes for genotypes with k ambiguous sites above $k + 1$, the number possible with no recombination and no recurrent mutation. This implies that a population will tend to contain few haplotypes (relative to k) and therefore that solutions generated by fewer haplotypes will be more accurate (S. Orzack, D. Gusfield and C. Wiuf, unpublished results).

We believe haplotype number will prove generally useful as the basis for making more accurate consensus predictions. Accordingly, the second way we generated a consensus prediction for each genotype was to use either just the solutions that were based on the fewest number of haplotypes ("minimum") or those solutions plus those that were based on one more than the fewest number of haplotypes ("minimum + 1"). As for full consensus, the most common inferral was taken to be the consensus prediction for a genotype.

The results shown in Figure 4 are relevant to claims that the rule-based approach satisfies a parsimony criterion in regard to the number of haplotypes used. For example, Stephens *et al.* (2001a, p. 979) wrote that Clark's approach "can be viewed as an attempt to minimize the number of haplotypes observed in the sample and, hence, as a sort of parsimony approach" and Niu *et al.* (2002, p. 158) wrote that "Clark's parsimony approach attempts to assign the smallest number of haplotypes for the observed genotype data." The incorrectness of these claims is indicated by the fact that variation 4b (the variation that appears to be that of Clark 1990) and all the other variations generated complete solutions that differ in the numbers of haplotypes used (see Figure 4). The genesis of these claims about parsimony is unclear; the only parsimony claim in Clark (1990, pp. 117–118) concerns the number of orphaned genotypes.

The results of consensus: In Table 5, we show the results of the full consensus and the minimum + 1 consensus calculations for our data. The results of consensus calculations are very consistent across sample paths of a given variation and we present results from only a single sample path.

A consensus solution can be used in two ways. First, such a solution provides a single prediction for any genotype; without such a synthetic prediction the investigator is left to pick a solution at random from among multiple complete solutions that are often of highly variable quality (see Figure 1). The results in Table 6 indicate that the average accuracy of complete solutions was always less than the accuracy of the consensus solution (although their difference can be small or substantial). This difference underscores the advantage of using a consensus prediction, even apart from its necessity when dealing with multiple solutions.

The distinction between the average number of correct inferrals and the consensus number of correct inferrals is reflected in the fact that the correlation be-

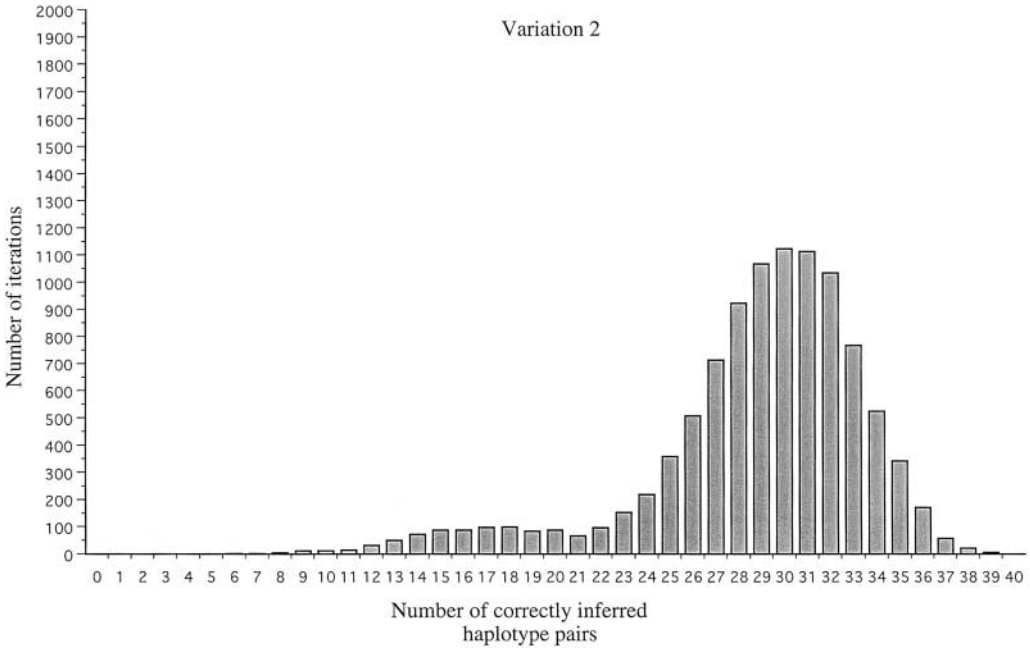

FIGURE 2.—The distribution of the number of correctly inferred pairs of haplotypes for a sample path of variation 2. Only the 47 ambiguous genotypes are considered.

tween the two was negative for all five consensus methods (Spearman rank correlation corrected for ties: full, −0.195; minimum, −0.708; minimum + 1, −0.878; minimum (all), −0.192; minimum + 1 (all), −0.169) and the average correlation is significantly negative (−0.509, 95% confidence interval: −0.747 to −0.158). These negative relationships indicate that the variations differed in the extent to which they explore the solution space. For example, variation 1 achieved a higher average percentage of correct inferrals (∼62%) as compared to, say, variation 4b (∼37%), but variation 1 had a poorer minimum consensus accuracy (36) as compared to variation 4b (40). This reveals that any given iteration of variation 4b tended to correctly predict only a subset of the ambiguous genotypes but not the same subset as another iteration; yet the ensemble prediction of such subsets had good accuracy.

The second way in which a consensus prediction can be used involves the assessment of the consensus values themselves so as to distinguish between more and less reliable inferrals. As shown in Table 5, as the consensus threshold or number of identical inferrals increases, there is a threshold value for which all inferrals are correct. One can use this approach to divide the data into inferrals that are certain or nearly certain as opposed to those that are less certain and for which experi-

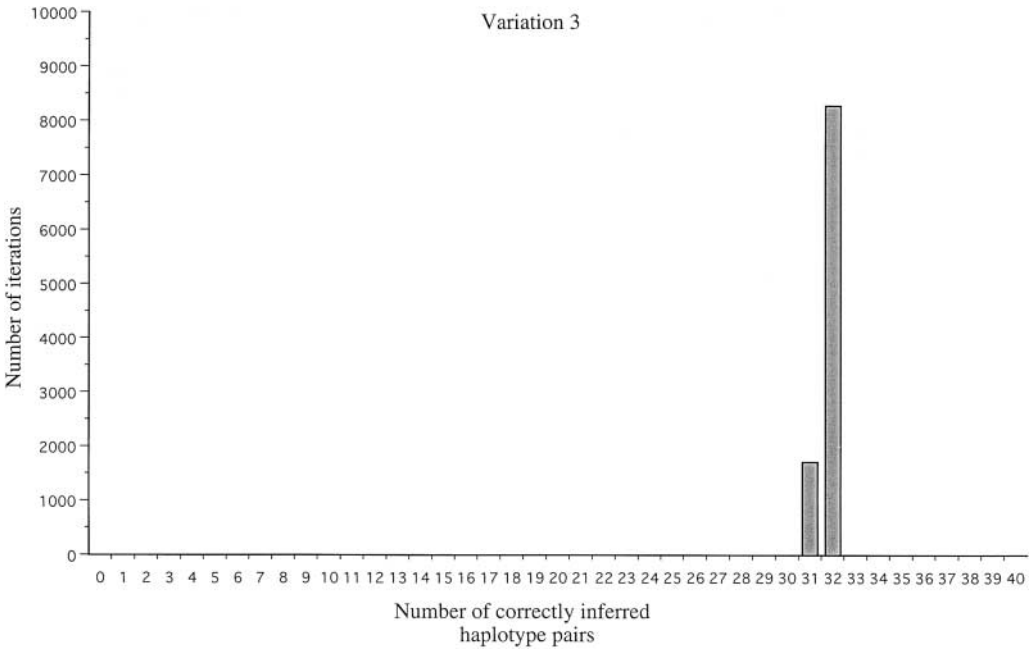

FIGURE 3.—The distribution of the number of correctly inferred pairs of haplotypes for a sample path of variation 3. Only the 47 ambiguous genotypes are considered.

TABLE 4

Average number of correctly inferred ambiguous genotypes, ~95% confidence interval, and the average percentage of correctly inferred ambiguous genotypes (47 possible)

Variation	Sample path	Average no. correct	95% confidence interval	Average % correct
1	1	29.0	28.99–29.10	61.7
	2	29.0	28.97–29.08	61.7
2	1	28.7	28.65–28.84	61.1
	2	28.7	28.58–28.77	61.1
	3	28.8	28.71–28.90	61.3
	4	28.8	28.70–28.89	61.3
2a	1	19.1	18.98–19.31	40.7
	2	19.0	18.81–19.14	40.4
2b	1	28.4	28.26–28.53	60.4
	2	28.6	28.43–28.70	60.8
2c	1	17.2	16.93–17.38	36.5
	2	17.3	17.04–17.49	36.7
	3	17.4	17.16–17.61	37.0
3	1	31.8	31.79–31.81	67.7
	2	31.8	31.79–31.81	67.7
4a	1	18.3	18.18–18.48	39.0
	2	17.5	17.31–17.61	37.1
	3	17.6	17.45–17.75	37.4
	4	17.6	17.48–17.78	37.5
4b	1	17.2	17.01–17.46	36.7
	2	17.5	17.23–17.68	37.1

mental inferral is mandated. For example, a full consensus threshold of 8000 (80%) would reduce the number of required experimental inferrals by ~50% (22/47). Similarly, a minimum + 1 consensus threshold of 80% would reduce the required experimental inferrals by ~>50% (25/47). This is a very substantial saving of experimental labor. We note that such consensus calculations did not perform well for variation 3 in that even high consensus thresholds are associated with a substantial proportion of incorrect inferrals (not shown). The

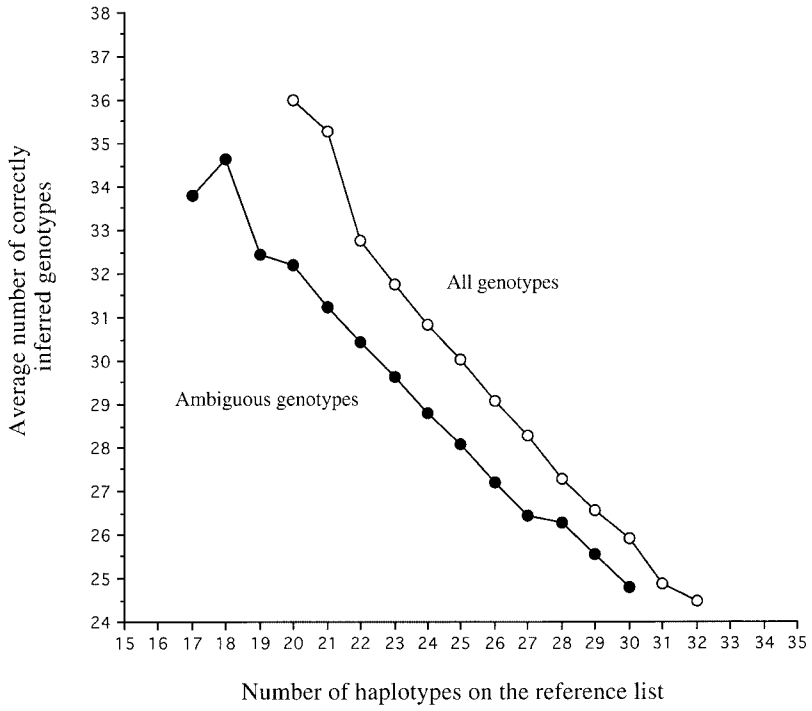

FIGURE 4.—The relationship between the average number of correctly inferred haplotype pairs and the number of different haplotypes on the reference list for a sample path of variation 1. We distinguish between the list used to resolve only ambiguous genotypes and the list used to resolve all genotypes. The latter will be larger if some haplotypes derived from unambiguous genotypes are not used to resolve ambiguous genotypes.

TABLE 5

Consensus calculations for a sample path of variation 1

	Full consensus											
	>1000	>2000	>3000	>4000	>5000	>6000	>7000	>8000	>9000	>9500		
C	31	31	31	31	29	26	24	22	14	1		
IC	16	16	16	14	7	5	3	0	0	0		
	Minimum + 1 consensus											
	>1	>2	>3	>4	>5	>6	>7	>8	>9	>10	>11	>12
C	34	34	34	34	34	34	34	34	34	34	34	34
IC	13	13	13	13	13	13	13	13	12	11	9	8
	>13	>14	>15	>16	>17	>18	>19	>20	>21	>22	>23	
C	34	34	33	30	28	25	25	25	24	23	22	
IC	7	6	7	2	2	2	0	0	0	0	0	

The placement of each consensus inferral that surpasses a given threshhold (*e.g.*, 1000 or 16) is determined by whether the consensus inferral is correct (C) or incorrect (IC). There are 10,000 iterations in the full calculations and 24 iterations in the minimum + 1 calculations.

strong frequency preference appears to amplify the "signal" of both incorrect and correct inferrals.

DISCUSSION

We believe that consensus calculations hold great promise for allowing the investigator to ultimately achieve correct inferrals for all of the genotypes contained in a sample. We are guided here by a belief that some measure of the reliability for any given inferral is an essential element of any inferral method (see also Stephens *et al.* 2001b) and that methods that just have an expected error rate for an ensemble of inferrals will not prove useful as the basis for the investigation of possibly weak genotype-phenotype associations. One needs as much certainty for inferrals as one can get, especially if phenotypes tend to be associated with rare genetic variants (see Pritchard 2001) for which the inferral process may tend to be more error prone.

We are developing methods to refine and extend these consensus results. Consensus calculations for another more complex locus do not perform as well in

TABLE 6

The average number of correctly inferred ambiguous genotypes and the number of ambiguous genotypes correctly predicted by various consensus methods

Variation	Average no. correct	Consensus method				
		Full	Minimum	Minimum + 1	Minimum (all)	Minimum + 1 (all)
1	29.0	31	36	34	36	39
2	28.7	31	35	35	38	35
2a	19.1	37	33	40	27	41
2b	28.4	31	40	38	40	35
2c	17.2	33	40	40	39	38
3	31.8	32	32	32	32	32
4a	17.5	38	38	42	25	26
4b	17.2	27	40	40	38	38
	Average:	32.5	36.8	37.6	34.4	35.5

Full, a consensus prediction based on all iterations. Minimum, a consensus prediction based on those iterations with the smallest number of haplotypes used in the resolution of ambiguous genotypes. Minimum + 1, a consensus prediction based on those iterations with the smallest or next-smallest number of haplotypes used in the resolution of ambiguous genotypes. Minimum (all), a consensus prediction based on those iterations with the smallest number of haplotypes used in the resolution of all genotypes (unambiguous and ambiguous). Minimum + 1 (all), a consensus prediction based on those iterations with the smallest or next-smallest number of haplotypes used in the resolution of all genotypes.

that even some high consensus scores for variation 1 are associated with some incorrect inferrals, although partitioning the locus so that separate inferrals occur for regions with different recombination rates improves consensus performance (S. Orzack and D. Gusfield, unpublished results). We hope to develop metrics so that an estimate of linkage disequilibrium can be used to select a consensus threshold that will be reliable at providing correct inferrals with certainty or near certainty. Our motivation is the belief that inferral methods should not be viewed as statistical machines competing against one another. Instead, they compete against experimental inferral, which can be viewed in the present context as being error free (or at least as having a much smaller error rate than that of algorithmic inferral). To this extent, the role of algorithmic inferral should be to guide the expenditure of experimental effort to achieve a correct solution.

The nature of mistaken inferrals and a guide to the investigator: The variations of the rule-based approach described here differ substantially in their algorithmic structure and population-genetic assumptions. It is essential to understand the basis for their differences in performance so that the potential user can compare the inferral capabilities of particular variations.

Table 7 contains a breakdown of the inferrals for a single sample path for each of the eight variations. Consider the results concerning full consensus calculations. The distribution of correct and incorrect inferrals does not differ across the variations for either two-ambiguous-site genotypes (uncorrected $\chi^2 = 10.46$, 7 d.f., $P > 0.05$) or three-ambiguous-site genotypes (uncorrected $\chi^2 = 4.12$, 7 d.f., $P > 0.05$). These results suggest that consensus inferral of haplotypes with two and three ambiguous sites can be performed with any of the variations. However, the variations differ significantly ($\alpha = 0.05$) in regard to their inferral success for more complex genotypes as indicated by the significant χ^2 values for four- and five-ambiguous-site genotypes (see Table 7). Inspection of the results indicates that variations 2a and 4a perform best in regard to the correct prediction of these more complex genotypes.

The breakdowns of correct inferrals for restricted consensus predictions are shown in Table 7. The χ^2 values for minimum consensus predictions indicate that the variations do not differ in their ability to correctly infer haplotypes involving up to five ambiguous sites. In contrast, the χ^2 values for minimum + 1 consensus predictions indicate that the variations do differ in their predictive ability for four- and five-site genotypes, with variations 2a and 4a again performing best.

The results shown in Tables 6 and 7 also indicate that restricted consensus calculations based on the number of haplotypes used to solve only ambiguous genotypes result in higher average numbers of correct inferrals than do restricted consensus calculations based on the number needed to solve all genotypes.

The rule-based approach and the approach of Stephens *et al.* (2001a): Both the consensus approach described here and the inferral method of Stephens *et al.* (2001a) allow one to distinguish between more and less certain inferrals and to thereby more efficiently allocate experimental effort. We regard such a discrimination to be essential for any haplotype inferral method.

These two methods differ in important ways. The rule-based approach is simpler to understand, to program oneself, and to modify. In addition, it is quicker since multiple iterations and associated consensus predictions can be obtained in a few minutes using typical microprocessors currently available. The approach of Stephens *et al.* (2001a) may often require much more CPU time. One reason is that the calculation of the posterior distribution needed for the identification of more and less certain inferrals in any given run is computationally intensive. Another reason is the need for multiple independent runs so that the consistency of haplotype inferrals can be assessed (see Stephens *et al.* 2001a, p. 985, and Stephens *et al.* 2002, p. 7).

We stress that we regard Stephens *et al.*'s approach to be an important inferral tool. In the present analysis, the performance of their program matches the best consensus performance of the rule-based approach (42 correctly inferred genotypes out of 47 possible); these best solutions differ (only 2 of the 5 incorrectly inferred genotypes are the same).

In our opinion, the performance of these methods is surprisingly good. After all, the vagaries of natural selection, genetic drift, and sampling biases, and the absence of assumptions about the genetic details of the APOE locus all combine to make haplotype inferral a formidably complex problem in an *a priori* sense. However, this performance does not mean that these methods can now be applied to most loci with the confidence that they will work well. Such a conclusion must await additional studies like the present one in which real phase-known genotypic data are analyzed.

Tests of haplotype inferral accuracy: Our results and those of Xu *et al.* (2002) and Niu *et al.* (2002) are the only studies in which the accuracy of a haplotype inferral method has been assessed by comparing real and inferred haplotype pairs. While praiseworthy in this regard, the Xu *et al.* (2002) study of *N*-acetyltransferase 2 genotypes is lacking because their results are based on one iteration of Clark's algorithm (C. Xu, personal communication); this iteration resolved 64 individuals and left 17 unresolved. Our results indicate that the results of a single iteration of a rule-based algorithm can be very misleading. Niu *et al.* (2002) also appear to have used a single iteration in their analysis of the 121 β_2-adrenergic genotypes shown in Table 2 of Drysdale *et al.* (2000) since they wrote (p. 161) that "Clark's algorithm made two mistakes (*i.e.*, predicted two individuals' phases incorrectly)." We believe that both of these studies should be regarded as giving an incomplete assess-

TABLE 7

The number of correctly inferred genotypes arranged by number of SNPs

No. of SNPs	No. of individuals	Variation:	Full consensus prediction 1	2	2a	2b	2c	3	4a	4b	χ^2
2	17		17	17	16	17	16	17	16	14	10.46
3	20		14	14	15	14	13	15	14	10	4.12
4	6		0	0	3	0	4	0	5	3	23.95
5	4		0	0	3	0	0	0	3	0	22.15
			Minimum consensus prediction								
No. of SNPs	No. of individuals	Variations: No. of solutions:	1 5	2 5	2a 1	2b 5	2c 4	3 8290	4a 5	4b 4	χ^2
2	17		17	17	17	17	17	17	16	17	7.05
3	20		14	14	14	16	16	15	15	16	1.60
4	6		3	3	1	4	4	0	4	4	11.27
5	4		2	1	1	3	3	0	3	3	10.00
			Minimum + 1 consensus prediction								
No. of SNPs	No. of individuals	Variations: No. of solutions:	1 24	2 109	2a 30	2b 15	2c 33	3 10000	4a 55	4b 16	χ^2
2	17		14	17	16	17	17	17	16	17	13.08
3	20		12	14	14	15	16	15	16	16	3.49
4	6		3	3	6	3	4	0	6	4	18.03
5	4		0	1	4	3	3	0	4	3	19.81
			Minimum consensus prediction (all)								
No. of SNPs	No. of individuals	Variations: No. of solutions:	1 1	2 6	2a 6	2b 5	2c 28	3 10000	4a 5	4b 17	χ^2
2	17		16	16	3	17	17	17	3	16	91.88
3	20		16	15	16	16	15	15	15	15	0.53
4	6		2	5	6	4	4	0	6	4	21.04
5	4		2	2	2	3	3	0	1	3	8.00
			Minimum + 1 consensus prediction (all)								
No. of SNPs	No. of individuals	Variations: No. of solutions:	1 16	2 112	2a 35	2b 75	2c 166	3 10000	4a 37	4b 137	χ^2
2	17		16	16	16	15	15	17	3	15	60.65
3	20		18	15	17	13	15	15	13	15	5.66
4	6		4	3	6	4	5	0	6	5	20.85
5	4		1	1	2	3	3	0	4	3	12.93

Critical χ^2 values for 7 d.f. are $P = 0.10$ (12.02), $P = 0.05$ (14.07), $P = 0.01$ (18.48), and $P = 0.005$ (20.28).

ment of the usefulness of the rule-based approach to haplotype inferral.

We have analyzed the Drysdale *et al.* (2000) data using the rule-based variations presented here. For example, the distribution of numbers of haplotype pairs correctly inferred by variation 1 is shown in Figure 5. All 10,000 iterations resolve all 79 ambiguous genotypes; there are 8480 different solutions. A total of 170 iterations make all of the inferrals correctly. The results of consensus calculations are shown in Table 8. A threshold for full consensus of, say, 50%, results in a complete and fully accurate solution while a threshhold of, say, 80%, results in 75 correct inferrals and no incorrect ones. Similarly, the minimum + 1 consensus calculations shown indicate that a consensus threshold of 50% results in a complete and fully accurate solution while a threshhold of 80% results in 78 correct inferrals and no incorrect ones. There is also a negative relationship between the number of haplotypes used in a solution and the inferral accuracy (not shown). These results suggest that the success of our APOE analyses is not somehow a consequence of features peculiar to that locus.

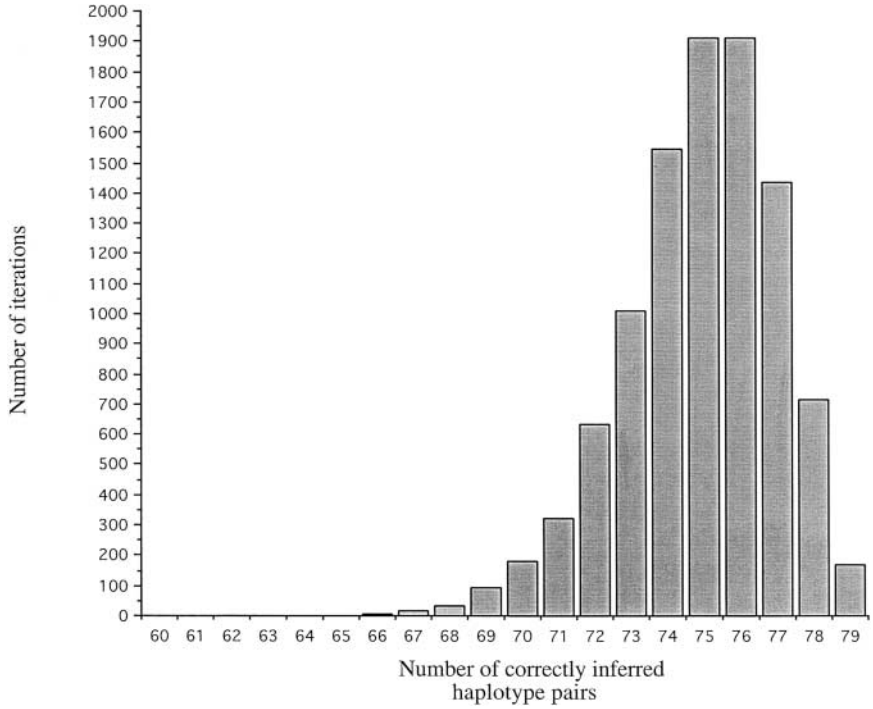

FIGURE 5.—The distribution of the number of correctly inferred pairs of haplotypes for a sample path of variation 1 applied to the data of DRYSDALE *et al.* (2000). Only the 79 ambiguous genotypes are considered.

Finally, we note that DRYSDALE *et al.* (2000, p. 10,485) commented in regard to a data set that includes the 121 aforementioned β_2-adrenergic genotypes that "assigning haplotypes from unphased genotype data from 200 individuals by using an extension [of Clark's algorithm] gave the same results as molecular haplotyping, except in a single subject because of a discrepancy at one SNP position." Since the authors did not describe this "extension," the significance of their claim cannot be assessed.

The general utility of consensus calculations: Consider the application of the consensus approach to the inferrals generated by the program of NIU *et al.* (2002). Their stochastic approach can also generate multiple solutions, as illustrated in Figure 6, which depicts the distribution of the number of correct inferrals for the APOE data from 1000 independent iterations of their program (Htyperv2). Two iterations of the program can differ at least twofold in the number of incorrect inferrals; multiple solutions also occur for another phase-unknown data set that we analyzed with their program (S. ORZACK and D. GUSFIELD, unpublished results). One can construct, say, a full consensus solution to make sense of these multiple solutions; it has 42 correct in-

TABLE 8

Consensus calculations for a sample path of variation 1 applied to the data of DRYSDALE *et al.* (2000)

	Full consensus									
	>1000	>2000	>3000	>4000	>5000	>6000	>7000	>8000	>9000	>9500
C	79	79	79	79	79	77	76	75	75	65
IC	0	0	0	0	0	0	0	0	0	0

	Minimum + 1 consensus							
	>20	>40	>60	>80	>100	>120	>140	>160
C	79	79	79	79	79	79	79	79
IC	0	0	0	0	0	0	0	0
	>180	>200	>220	>240	>260	>280	>300	>320
C	79	79	78	78	78	78	78	73
IC	0	0	0	0	0	0	0	0

The placement of each consensus inferral that surpasses a given threshhold (*e.g.*, 10,000 or 240) is determined by whether the consensus inferral is correct (C) or incorrect (IC). There are 10,000 iterations in the full calculations and 328 iterations in the minimum + 1 calculations.

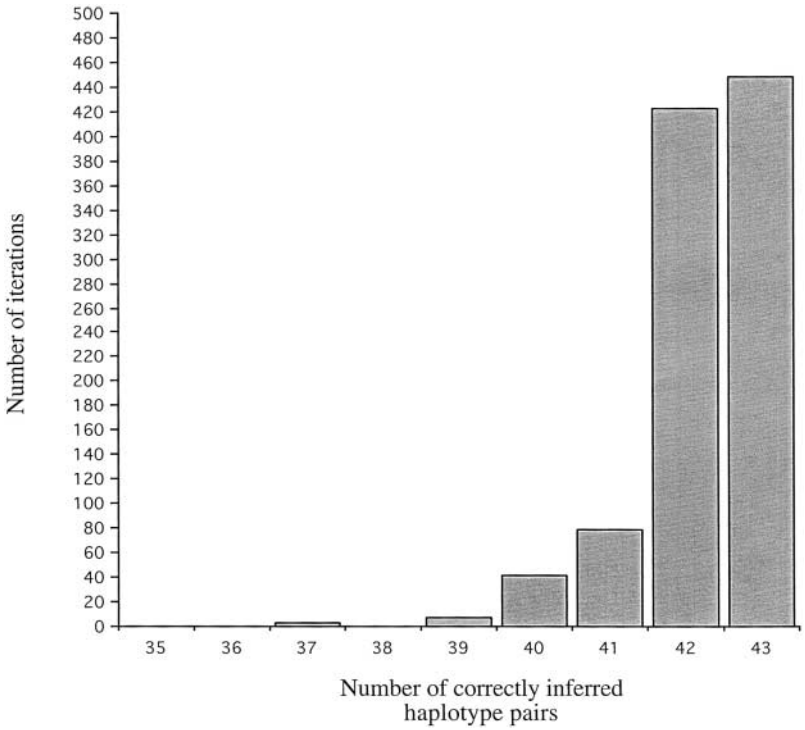

FIGURE 6.—The distribution of the number of correctly inferred pairs of haplotypes for 1000 independent iterations of the program Htyperv2 created by NIU *et al.* (2002). Only the 47 ambiguous genotypes are considered.

ferrals, making it slightly less accurate than the best solution found (43), but it allows one to avoid choosing randomly from among competing solutions. We note in passing that Htyperv2 also infers haplotype pairs that do not generate the observed genotype; in 1000 independent iterations, 290 iterations had one such "bad" genotype, 159 iterations had two, and three iterations had three.

Consensus methods can also be applied to the results generated by the method of STEPHENS *et al.* (2001a). We analyzed the APOE data with several thousand independent runs ($n = 2303$) of their program (each with 10,000 steps in the Markov chain, a thinning interval of 100, and a burn-in of 10,000). In this case, we found only one solution (as described above). Of course, this monomorphism could not be expected *a priori*. Multiple runs with their program of other data sets with real or artificial haplotype pairs reveal multiple solutions for which consensus calculations allow one to construct a single solution (S. ORZACK, L. SUBRAHMANYAN and D. GUSFIELD, unpublished results; S. M. FULLERTON, personal communication).

Any algorithmic inferral method has the potential to generate different solutions. This is as it should be, given that many genotypic configurations can support different haplotype configurations. Indeed, such a general consideration suggests that no inferral program be viewed as allowing "one-time-use" only. To this extent, techniques for reconciling competing solutions are needed. One approach is consensus; another is the use of goodness-of-fit statistics, as described by STEPHENS *et al.* (2002, p. 18). In either case, one must perform multiple analyses of the same data set.

The future development of rule-based algorithms and consensus calculations: If one has genotypes with known haplotype pairs, one advantage of the rule-based approach is that one can assess, using an integer programming analysis, whether a given variation could result in perfect prediction (see GUSFIELD 2001). Such knowledge has an obvious relevance when deciding whether to attempt to improve the variation. At present, if other methods fail to produce a perfect solution, there is no such framework for determining whether they could do so.

Such an analysis shows that such a perfect solution exists for variation 1. However, the best result we obtained for a single iteration was 40 correct resolutions out of a possible 47. Hence, even with a large number of iterations, this variation failed to achieve optimal performance. This failure motivates ongoing research on how to best implement particular variations to achieve optimal performance and on techniques that force these algorithms to do so (S. ORZACK and D. GUSFIELD, unpublished results).

We thank Anne Ferentz, Alan Templeton, David Posada, Carsten Wiuf, and three reviewers for critical comments. This work was partially supported by National Science Foundation (NSF) awards SES-9906997, DBI-9723346, and EIA-0220154 and by Variagenics.

LITERATURE CITED

ARTIGA, M. J., M. J. BULLIDO, I. SASTRE, M. RECUERO, M. A. GARCIA *et al.*, 1998 Allelic polymorphisms in the transcriptional regulatory region of apolipoprotein E gene. FEBS Lett. **421:** 105–108.

BULLIDO, M. J., M. J. ARTIGA, M. RECUERO, I. SASTRE, M. A. GARCIA *et al.*, 1998 A polymorphism in the regulatory region of APOE associated with risk for Alzheimer's dementia. Nat. Genet. **18:** 69–71.

Chakravarti, A., 1999 Population genetics—making sense out of sequence. Nat. Genet. **21** (Suppl.): 56–60.

Clark, A. G., 1990 Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol. **7:** 111–122.

Dracopoli, N. C., J. L. Haines, B. R. Korf, C. M. Morton, C. E. Seidman *et al.*, 2001 *Current Protocols in Human Genetics*. John Wiley & Sons, New York.

Drysdale, C. M., D. W. McGraw, C. B. Stack, J. C. Stephens, R. S. Judson *et al.*, 2000 Complex promotor and coding region β_2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. Proc. Natl. Acad. Sci. USA **97:** 10483–10488.

Excoffier, L., and M. Slatkin, 1995 Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. **12:** 921–927.

Fallin, D., and N. J. Schork, 2000 Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am. J. Hum. Genet. **67:** 947–959.

Fullerton, S. M., A. G. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor *et al.*, 2000 Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am. J. Hum. Genet. **67:** 881–900.

Gusfield, D., 2001 Inference of haplotypes from samples of diploid populations: complexity and algorithms. J. Comput. Biol. **8:** 305–323.

Hartman, J. L., IV, B. Garvik and L. Hartwell, 2001 Principles for the buffering of genetic variation. Science **291:** 1001–1004.

Hawley, M. E., and K. K. Kidd, 1995 HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J. Hered. **86:** 409–411.

Judson, R., J. C. Stephens and A. Windemuth, 2000 The predictive power of haplotypes in clinical response. Pharmacogenomics **1:** 15–26.

Lin, S., D. J. Cutler, M. E. Zwick and A. Chakravarti, 2002 Haplotype inference in random population samples. Am. J. Hum. Genet. **71:** 1129–1137.

Long, J. C., R. C. Williams and M. Urbanek, 1995 An E-M algorithm and testing strategy for multiple-locus haplotypes. Am. J. Hum. Genet. **56:** 799–810.

Michalatos-Beloin, S., S. A. Tishkoff, K. L. Bentley, K. K. Kidd and G. Ruano, 1996 Molecular haplotyping of genetic markers 10 kb apart by allele-specific long- range PCR. Nucleic Acids Res. **24:** 4841–4843.

Nickerson, D. A., V. O. Tobe and S. L. Taylor, 1997 Polyphred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res. **25:** 2745–2751.

Nickerson, D. A., S. L. Taylor, S. M. Fullerton, K. M. Weiss, A. G. Clark *et al.*, 2000 Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. Genome Res. **10:** 1532–1545.

Niu, T., Z. S. Qin, X. Xu and J. S. Liu, 2002 Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am. J. Hum. Genet. **70:** 157–169.

Pritchard, J. K., 2001 Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. **69:** 124–137.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199–204.

Science Citation Index, 2003 ISI Web of Knowledge, September 2003 (http://isiknowledge.com).

Stephens, M., N. J. Smith and P. Donnelly, 2001a A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. **68:** 978–989.

Stephens, M., N. J. Smith and P. Donnelly, 2001b Reply to Zhang *et al.* Am. J. Hum. Genet. **69:** 912–914.

Stephens, M., N. J. Smith and P. Donnelly, 2002 Documentation for PHASE, version 1.0 (http://www.stat.washington.edu/stephens/phase.html).

Templeton, A. R., C. F. Sing, A. Kessling and S. Humphries, 1988 A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. Genetics **120:** 1145–1154.

Xu, C., K. Lewis, K. L. Cantone, P. Khan, C. Donnelly *et al.*, 2002 Effectiveness of computational methods in haplotype prediction. Hum. Genet. **110:** 148–156.

Zhu, X., C. A. McKenzi, T. Forrester, D. A. Nickerson, U. Broeckel *et al.*, 2001 Localization of a small genomic region associated with elevated ACE. Am. J. Hum. Genet. **67:** 1144–1153.

Communicating editor: M. Uyenoyama