Inference of Haplotypes from PCR-amplified Samples of Diploid Populations[1]

Andrew G. Clark
Department of Biology and Genetics Program, Pennsylvania State University

Direct sequencing of genomic DNA from diploid individuals leads to ambiguities on sequencing gels whenever there is more than one mismatching site in the sequences of the two orthologous copies of a gene. While these ambiguities cannot be resolved from a single sample without resorting to other experimental methods (such as cloning in the traditional way), population samples may be useful for inferring haplotypes. For each individual in the sample that is homozygous for the amplified sequence, there are no ambiguities in the identification of the allele's sequence. The sequences of other alleles can be inferred by taking the remaining sequence after "subtracting off" the sequencing ladder of each known site. Details of the algorithm for extracting allelic sequences from such data are presented here, along with some population-genetic considerations that influence the likelihood for success of the method. The algorithm also applies to the problem of inferring haplotype frequencies of closely linked restriction-site polymorphisms.

Introduction

Although the acquisition of sequences of multiple alleles from natural populations provides the ultimate description of genetic variation in a population, the time and labor involved in obtaining sequence data has limited the number of such studies. Any means of acquiring sequence data from population samples that decreases this effort could be of great utility. The advent of the polymerase chain reaction (PCR) (Saiki et al. 1985; Scharf et al. 1986) has greatly accelerated the process of going from genomic DNA to sequence data, by eliminating the cloning step. Direct sequencing of PCR products can work very well for mtDNA or for DNA from isogenic or otherwise homozygous or hemizygous regions, but heterozygosity in diploids results in amplification of both alleles. Using asymmetric amplification, with unequal concentrations of the two primers, one can obtain single-stranded DNA products that can be directly sequenced. In a heterozygote, asymmetric PCR results in amplification products of both homologues. The resulting superimposition of the two sequencing ladders for the two alleles produces a vast number of possible haplotypes for any heterozygous individual. If there are n such "ambiguous" sites in an individual, then there are 2^n possible haplotypes. The challenge is to devise a scheme whereby haplotypes can be inferred from a series of these ambiguous sequences constructed from samples of diploid natural populations.

The Algorithm

Suppose that we have a series of sequences from diploids with many ambiguous sites. Even with highly polymorphic sequences, a sample of sufficient size from a

Address for correspondence and reprints: Dr. Andrew G. Clark, Department of Biology and Genetics Program, 208 Mueller Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802.

1. Key words: haplotype, polymerase chain reaction, direct sequencing, population genetics.

Mol. Biol. Evol. 7(2):111–122. 1990.

0737-4038/90/0702-0001$02.00

population will have some homozygotes or individuals with just one heterozygous site. A homozygote is recognized by a lack of ambiguous sites on the sequencing gel, and, as soon as a homozygote is found, we have unambiguously identified a haplotype. If an individual has a single heterozygous site, then we have unambiguously identified two haplotypes. The algorithm begins by finding all homozygotes and single-site heterozygotes and tallying the resulting known haplotypes.

For each known haplotype, we then look at all the remaining unresolved sequences 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 new haplotypes can be found.

Suppose, for example, that one observes a sequencing gel with the sequence ATGGTAC. If this sequence has no ambiguous sites, then we infer that this is a true haplotype (the individual must have been homozygous at all seven sites). If one also observes the sequence $AT\frac{G}{C}G\frac{C}{T}AC$, then it is clear that the two ambiguous sites could result in any of four possible haplotypes. Because the original known haplotype is one of these, we assume that the genotype had this haplotype. The homologous allele that the genotype must have had in order to give the observed ambiguous phenotype is ATCGCAC.

Figure 1 is a diagram of the chain of inference used by the algorithm. Suppose that A_1A_1 is a homozygote for haplotype A_1, where A_1 represents a sequence of arbitrary

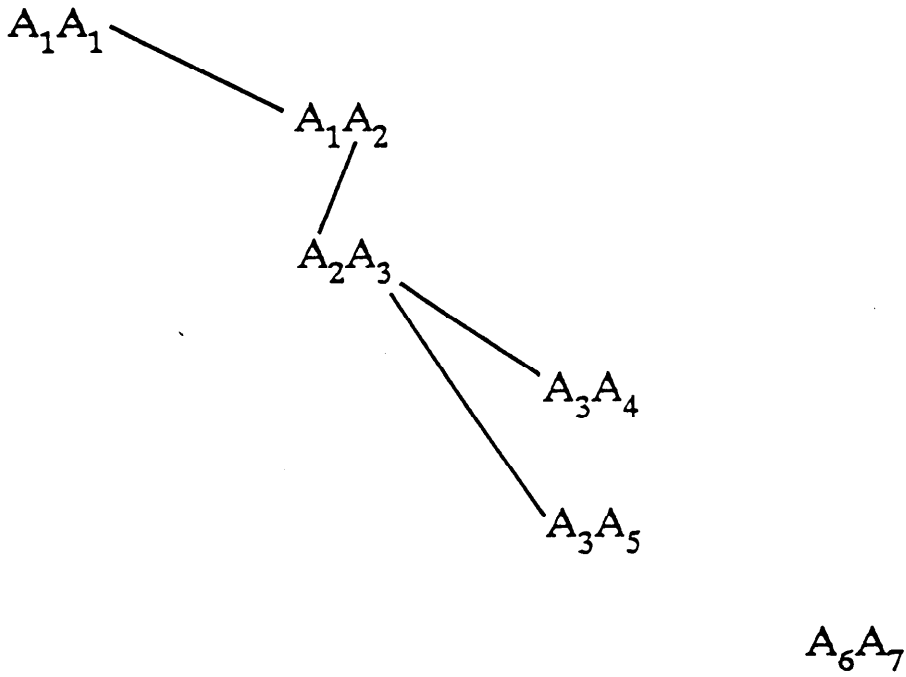

FIG. 1.—Diagram showing the cascade of inferences in the haplotype-inferring algorithm. First a homozygote is identified, yielding a definitive haplotype. If any other ambiguous sequence could have this haplotype as one of its two haplotypes, then the remaining bands determine the other haplotype. This chain is continued either until all haplotypes are resolved or until one identifies sequences that cannot be derived from any of the known haplotypes (as in genotype A_6A_7).

length. Let A_1A_2 be a heterozygote that would produce a superimposed sequencing ladder with potentially many ambiguous sites, but among the possible haplotypes lies the combination of A_1A_2. Observation of this fact would give us both haplotypes, A_1 and A_2, and we could then search for occurrence of either A_1 or A_2 from among the remaining ambiguous sequences. In summary, the algorithm is as follows: (1) Identify all homozygotes and single-site heterozygotes. Consider their haplotypes as "resolved." (2) Determine whether any of the resolved haplotypes could be one of the alleles in each of the remaining ambiguous sequencing ladders. If not, then stop; otherwise continue. (3) Each time a resolved haplotype is identified as one of the possible alleles in an ambiguous ladder, identify the homologue as the sequence with the other set of bands at each ambiguous site. Consider this newly identified homologue as resolved, and go to step 2. By performing these steps with different orderings of the data, the uniqueness of the solution can be determined. The solution that resolves the most haplotypes is almost always valid (see below). A simple implementation of the algorithm designed for inferring sequences is available from the author (please send formatted 5.25″ diskette).

The algorithm works on samples that are polymorphic for insertions and deletions, but implementing the algorithm on a computer can be cumbersome if it must be able to handle multiple nested insertions and deletions. In the simple case of presence/absence of one insertion, a heterozygote would produce a sequencing ladder with many ambiguous sites 3′ from the insertion. If either haplotype of this heterozygote had been identified in a homozygote, both haplotypes could be correctly inferred.

While the algorithm should work in principle, there are three problems that can arise: (1) One may fail to recover any homozygotes or single-site heterozygotes and may never get the cascade started. (2) There may be unresolved haplotypes left at the end (such as A_6 and A_7 in the example in fig. 1). (3) Haplotypes might be erroneously inferred if a crossover product of two actual haplotypes is identical to another true haplotype. The likelihoods of these problems can be estimated from population genetics theory and will clearly depend on such factors as the average heterozygosity per nucleotide site, the length of the DNA sequence observed, sample size, and rates of recombination among sites.

How Many Ambiguous Sites Will There Be?

To examine how many ambiguous sites there will be, note that the infinite-site model (Kimura 1969, 1971) allows one to estimate the number of mismatching sites between a pair of genes drawn at random from a natural population of diploids in steady state between mutation and drift. Under the infinite-site model, the expected number of mismatching sites for a DNA sequence of L nucleotides is given by $\theta = L\theta_{nt}$, where θ_{nt} is $4N\mu$ (Kimura 1969; Watterson 1975). Here N is the effective population size and μ is the neutral mutation rate per nucleotide site per generation.

Several studies of restriction-site variation in natural populations of *Drosophila melanogaster* have allowed estimation of the magnitude of heterozygosity per nucleotide. In the *Adh* region, θ_{nt} has been estimated at 0.006 (Kreitman 1983; Aquadro et al. 1986; Kreitman and Aquadé 1986*a*, 1986*b*; Simmons et al. 1989), and other chromosomal regions appear to exhibit similar levels of heterozygosity, including *Notch* at 0.007 (Schaeffer et al. 1988), *Amy* at 0.006 (Langley et al. 1988), the 87A heat-shock region at 0.002 (Leigh Brown 1983), *white* at 0.004–0.008 (Langley and Aquadro 1987; Miyashita and Langley 1988), *zeste-tko* at 0.004 (Aguadé et al. 1989*b*), and *rosy*

at 0.003 (Aquadro et al. 1988). One exceptional region spans the X-linked *yellow-achaete-scute* genes, with a per-nucleotide heterozygosity of 0.0003 (Aguadé et al. 1989*a*). Other species appear to exhibit higher levels of heterozygosity, with the *rosy* region of *D. simulans* having a heterozygosity of 0.019 (sixfold that of *melanogaster;* Aquadro et al. 1988) and *Adh* in *D. pseudoobscura* having a per-nucleotide heterozygosity of 0.021 (Schaeffer et al. 1987). Other estimates of heterozygosity at the nucleotide level include those of β-globin and growth hormone in man (0.002 each), factor IX in man (0.0002), and growth hormone in pig (0.007) (Nei and Hughes, accepted).

When $\theta_{nt} = 0.005$/nucleotide site is taken as an average for *D. melanogaster,* an upper bound for the value of θ can be estimated as $0.005L$. If the mutation rate is homogeneous across a gene, doubling the size of a gene should double the total mutation rate, but this yields an upper bound for the value of θ because the mutation-drift process generates correlations of heterozygosity across sites. Fragment lengths of 400, 1,000, and 2,000 nucleotides, taken from a population having $\theta_{nt} = 0.005$, would have θ's of 2, 5, and 10, respectively. The distribution of the number of mismatching sites expected when two genes are drawn from a population was calculated by Watterson (1975) as follows:

$$\Pr(2 \text{ sequences have } m \text{ mismatches}) = \left(\frac{1}{\theta+1}\right)\left(\frac{\theta}{\theta+1}\right)^m.$$

For a typical *Drosophila* gene of length 1,000 bp, we expect that the average direct-sequencing gel would reveal five ambiguous sites per individual (if we ignore for now the effects of insertion and deletion). Figure 2 gives the distribution of the number of mismatches for a part of the parameter space that might be reasonable for a typical population sample of *D. melanogaster.*

Can the Algorithm Get Started?

Without homozygotes or single-site heterozygotes, the algorithm cannot get started, so it is important to consider what population genetics theory says about the likelihood of recovering homozygotes. According to the infinite-allele model, in a population that is in steady state, with a balance between the rate of gain of alleles by mutation and loss due to drift, the probability that two genes will be identical is $F = 1/(1 + \theta)$, so the chance of having two different alleles is $1 - F$. The probability of drawing n diploids and getting no homozygotes is complicated by the fact that we are drawing without replacement, so that subsequent samples are not independent of one another. To calculate the probability of drawing n diploids and getting no homozygotes, we must exhaustively enumerate all possible configurations of alleles in a sample and determine their probabilities with the Ewens (1979, p. 95) sampling formula. The probability that we are after is the sum of the probabilities of the configurations weighted by the probability that each configuration lacks homozygotes. In the case of a sample of two diploids, the configurations may be $\{A_1/A_2, A_3/A_4\}$, $\{A_1/A_2, A_2/A_3\}$, $\{A_1/A_2, A_1/A_2\}$, $\{A_1/A_2, A_3/A_3\}$, $\{A_1/A_1, A_2/A_2\}$, $\{A_1/A_1, A_1/A_2\}$, and $\{A_1/A_1, A_1/A_1\}$. Only the first three of these configurations lack homozygotes, and their respective probabilities (from the Ewens sampling formula) are θ^3/S, $4\theta^2/S$, and $2\theta/S$, where $S = [(1+\theta)(2+\theta)(3+\theta)]$. The probability of obtaining no homozygotes in a sample of two diploids is the sum of these probabilities, or

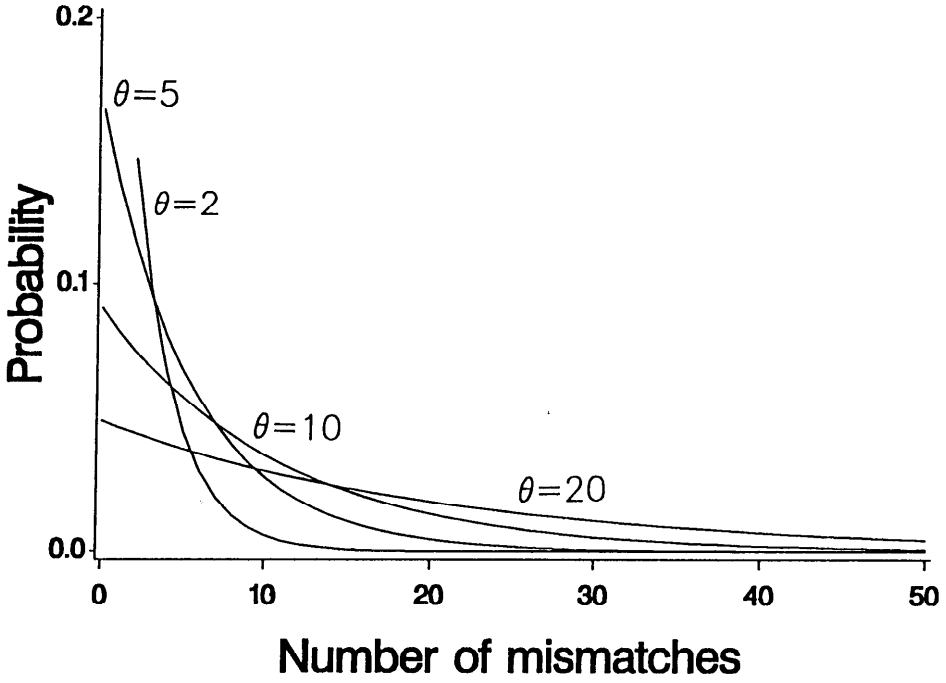

FIG. 2.—Probability density of the number of mismatches between two sequences drawn from a population at steady state under the infinite-site model. The curve for $\theta = 2$ is truncated and would meet the y-axis at 0.333 [the probability of no mismatches is $1/(1+\theta)$].

$$\Pr(\text{no homozygotes}) = \frac{\theta^3 + 4\theta^2 + 2\theta}{(1+\theta)(2+\theta)(3+\theta)}.$$

For larger sample sizes, one must determine the probability of all partitions of all configurations of alleles having no homozygotes, and this is a very cumbersome calculation. For large population sizes (large θ), successive samples from a population become nearly independent. The chance of drawing one homozygote is $1/(1+\theta)$, and Watterson's distribution gives the chance of drawing a single-site heterozygote as being $\theta/(1+\theta)^2$. If approximate independence is assumed, the chance of drawing n individuals and having none of them be either a homozygote or a single-site heterozygote (i.e., the probability of failing to get the algorithm started) is

$$\Pr(\text{failure}) \approx \left[1 - \frac{1}{1+\theta} - \frac{\theta}{(1+\theta)^2}\right]^n.$$

This approximation becomes very good when $\theta > 0.5$. As Figure 3 indicates, even with a highly diverse segment of DNA, homozygotes or single-site heterozygotes will be recovered if the sample is large enough. For example, even if θ is as large as 10, the chance of failing to get the algorithm started is <1% with a sample of 24 individuals. These calculations were based on the infinite-site model, and recombination would result in a greater allelic diversity and lower chance of obtaining no homozygotes. Simulations will address this issue. In the case of organisms that can be reared in the laboratory, one can use either isofemale lines or lines that are otherwise somewhat inbred to increase the likelihood of encountering homozygotes. Direct sequencing can

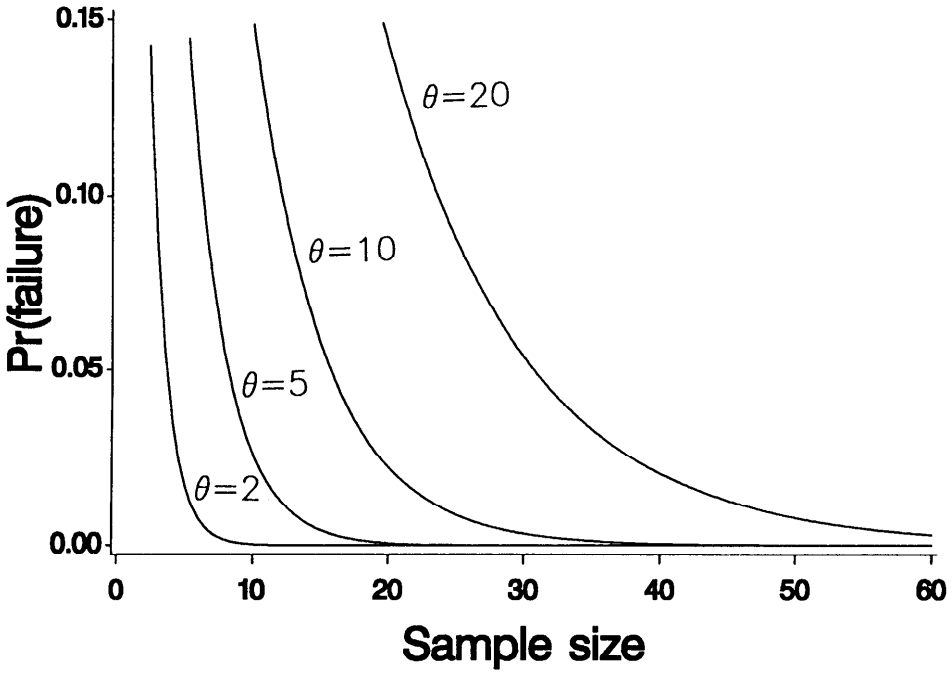

FIG. 3.—Probability of failing to start the algorithm (owing to a lack of homozygotes or single-site heterozygotes) as a function of θ and the sample size (the number of diploid genotypes). The curves indicate the probability of obtaining no homozygotes or single-site heterozygotes for given values of θ and sample size according to the infinite-site model.

detect homozygotes much more quickly than can the classical method of chromosome extraction.

Will There Be Orphaned Alleles?

If a genotype A_iA_j is found such that neither haplotype A_i nor A_j occurs in a homozygote or any other heterozygote, then these haplotypes cannot be resolved and will be referred to as "orphans." Simulations were performed to explore the fraction of times that orphans will be encountered under a range of values of θ and a range of sample sizes. Samples of $2n$ gametes were drawn from the frequency distribution expected under the infinite-allele model, constructed with the algorithm of Stewart (1977). These $2n$ gametes were combined to form n diploid genotypes. Homozygotes were identified, and paths connecting alleles were constructed, in an attempt to connect all alleles to a homozygote. If orphaned alleles remain, then this sample is scored as an orphaned sample. A total of 1,000 samples was drawn, each from a distinct realization of the Ewens distribution constructed for each combination of θ and n that is indicated in table 1. There are two noteworthy trends in table 1. First, larger values of θ result in a higher chance of obtaining orphans (for a given sample size) because larger θ results in greater allelic diversity at steady state. Second, as the sample size increases, orphans are less likely to remain, because large samples are more likely to contain paths connecting all alleles to homozygotes. With regard to the diagram shown in fig. 1, if the sample size were expanded, ambiguous genotype A_6A_7 would be resolved when genotypes such as A_1A_6 are found. The feature of being able to sample more individuals until all haplotypes are resolved has considerable practical importance.

Table 1
Fraction of Samples with Any Orphaned Alleles, and Average Frequency of Orphaned Alleles

	Sample Size (no. of individuals)			
θ	10	20	50	100
1.....	<0.001 (<0.001)	<0.001 (<0.001)	<0.001 (<0.001)	0.002 (<0.001)
2.....	0.008 (0.008)	<0.001 (<0.001)	<0.001 (<0.001)	0.003 (<0.001)
5.....	0.053 (0.205)	0.035 (0.008)	0.011 (0.001)	0.004 (<0.001)
10.....	0.204 (0.078)	0.102 (0.029)	0.037 (0.003)	0.011 (<0.001)
20.....	0.325 (0.118)	0.276 (0.092)	0.119 (0.018)	0.057 (0.001)

NOTE.—Figures represent the fraction of 1,000 samples, drawn from simulations of the infinite-allele model, that have any orphaned alleles after the method sketched in fig. 1 is applied. Figures in parentheses are the average frequencies of orphaned alleles in the samples. θ is the expected number of mismatching sites in a pair of alleles.

Will There Be Anomalous Matches?

An example will best illustrate the nature of this problem. Suppose one observes a homozygote for sequence ATTGCTGA. If one also observes a sequencing ladder with four ambiguous sites, composed of haplotypes ATCGCTAA and AGTGCGGA, the originally identified homozygous haplotype may be anomalously identified as part of this haplotype pair. If this happens, then the other haplotype inferred from this haplotype pair may be erroneous, and there will follow a cascade of errors.

To determine the frequency of such anomalous matches, the infinite-site model was simulated using a tree-based algorithm (Hudson 1983) provided by Dick Hudson. For given values of θ, n, and the recombination rate across the sequence, this algorithm gives a sample from a steady-state population obeying the assumptions of the infinite-site model. These samples were then presented to the computer algorithm that performs the haplotype-inference algorithm outlined above (table 2). There are no striking trends in the frequency of anomalous matches depending on either θ or sample size. Larger samples might encounter more genotypes that could anomalously match alleles, but homozygotes that resolve the ambiguity are also found in larger samples. There appears to be only a minor effect of recombination in this portion of the parameter space. The relatively minor effect of recombination is consistent with Hudson's (1983) simulations, which showed that, while recombination increases heterozygosity and the number of alleles, there is little effect of recombination on the expected heterozygosity conditioned on the number of alleles. The effectiveness of the algorithm appears to depend on the conditional heterozygosity, such that samples with more alleles can be resolved if there is higher heterozygosity. A higher heterozygosity assures that more of the common alleles are found in heterozygotes with rare alleles, so that, once the chain gets started (by finding one homozygote), it continues to resolve more alleles.

From the simulations described above, in each case when anomalous matches were found the computer routine tested whether the resulting false haplotype could be pulled from the sample of ambiguous sequences. In no case were any false complementary haplotypes found in the sample, and, in every case in which anomalous matches were obtained, that solution had orphaned alleles. In no case was a solution that obtained anomalous matches the solution with the fewest orphans. This empirically demonstrates a parsimony rule that the solution with the fewest orphans is the valid

Table 2
Fraction of Samples with Unresolved Anomalous Matches, and Frequency of Anomalous Haplotypes

θ	Sample Size (no. of individuals)			
	10	20	50	100
1.....	0.09 (0.07)	0.06 (0.04)	0.06 (0.06)	0.08 (0.09)
	0.07 (0.09)	0.06 (0.05)	0.07 (0.08)	0.08 (0.08)
2.....	0.15 (0.14)	0.14 (0.10)	0.12 (0.14)	0.15 (0.15)
	0.14 (0.11)	0.22 (0.14)	0.12 (0.13)	0.17 (0.17)
5.....	0.18 (0.22)	0.23 (0.00)	0.17 (0.23)	0.20 (0.34)
	0.10 (0.21)	0.23 (0.33)	0.20 (0.24)	0.26 (0.17)
10.....	0.00 (0.07)	0.17 (0.00)	0.13 (0.22)	0.20 (0.34)
	0.05 (0.13)	0.14 (0.31)	0.16 (0.28)	0.29 (0.30)
20.....	0.00 (0.20)	0.00 (0.00)	0.15 (0.24)	0.20 (0.34)
	0.02 (0.02)	0.02 (0.00)	0.15 (0.28)	0.35 (0.19)

NOTE.—Figures in each cell report the fraction of 100 samples, drawn from the infinite-site algorithm of Hudson (1983), that have anomalous matches and are not directly related to a homozygous allele. Figures in parentheses are the average frequency of anomalously matched alleles. Within each cell the upper figures are for no recombination, and the lower figures represent $4Nr = 0.5$.

solution and suggests that when a solution resolves all haplotypes it is likely to be unique.

The results of figure 3 and tables 1 and 2 can be summarized as follows: If no homozygotes or single-site heterozygotes are found, the method cannot even get started, and more sampling is needed. The probability of this difficulty is given in figure 3. Once the algorithm gets started (by finding homozygotes or single-site heterozygotes), then the chain of inference may be broken before all alleles are resolved. The probability of this problem is reported in table 1. While anomalous matches occur quite often (table 2), whenever they occur they result in anomalous complementary haplotypes that do not occur in the sample, and they thereby result in premature termination of the cascade of inference. In the simulations, the solution with the fewest orphans and fewest anomalous matches always resolved the greatest number of true haplotypes.

Tests of the Algorithm by Using *Adh* and *Est-6* Sequences

Kreitman (1983) reported 43 polymorphic sites in a sample of 11 alleles of *Adh* in *Drosophila melanogaster* and a total of nine different haplotypes. Cooke and Oakeshott (1989) reported 52 polymorphic sites in sequences of 13 alleles of *Est-6* in *D. melanogaster* and resolved 12 different haplotypes. While neither study represents a sample from a natural population, these data can be used to illustrate the algorithm with the structure of linkage disequilibrium among sites that is observed in nature. Samples of the haplotypes were drawn at random to form $n = 10$, 20, 50, and 100 diploid genotypes. The haplotype pairs of each genotype were randomized by swapping nucleotides at ambiguous sites, thus discarding all linkage phase information from the sample. These data were then presented to a computer routine that performs the haplotype-inferring algorithm described above.

In all cases, when the sample size was >20, all the haplotypes in the sample were resolved. Many of the samples of 20 did not contain all alleles, but the alleles in the

sample were successfully identified. Although all haplotypes were successfully identified, most samples could admit anomalous matches. Whenever an anomalous match was obtained, however, not all haplotypes were subsequently resolved. With samples of 10, fewer than half of the haplotypes were resolved, and in ~40% of these cases no homozygotes were recovered, so no haplotypes were resolved.

The tests that resampled 50 or 100 genotypes from Kreitman's sequences were not quite legitimate because with samples of this size one would expect to encounter more rare haplotypes. In addition, because the sequences were not obtained from one population, the alleles are likely to exhibit more differences than the infinite-site model would predict for a panmictic population (the sample is thus biased in favor of the algorithm working successfully). Nevertheless, the success at correctly resolving sequences with experimentally acceptable sample sizes is consistent with the theoretical conclusions and provides considerable encouragement for applying the method.

Other Approaches

The algorithm described here is intended to be able to resolve ambiguous sequences in the situation where there is only one unique pair of true haplotypes that correspond to a given ambiguous sequencing gel. If one were looking at only a few polymorphic sites, and if several of the linkage phases are present in the population, then a given sequence might be one of several haplotype pairs. In this situation, one needs to apply such methods as gene counting and the EM algorithm for frequency estimation (Dempster et al. 1977). For this purpose, large sample sizes are needed to obtain accurate frequency estimates. This problem would be more likely to arise if one were examining variation at a small number of loosely linked restriction sites, as might be done in studies of human polymorphism. In this situation, family data could increase the likelihood of being able to resolve linkage phases. With tightly linked, polymorphic restriction sites, where multisite linkage phases may still be unique, the algorithm might be quite effective.

There are also two direct experimental procedures for obtaining haplotype data by using PCR. The first is to isolate single sperm cells by micromanipulation and to amplify the DNA from these individual cells (Li et al. 1988). While this method has remarkable advantages for mapping (Boehnke et al. 1989), it may not always be practical to obtain sperm, and the data do not represent a population sample. Another experimental approach is to dilute genomic DNA down to the point where each reaction has an average of one DNA molecule (with replicate samples having a Poisson distribution about this mean). When these samples are amplified, some will reveal unambiguous haplotype information (Ruano et al., accepted). Each of these two experimental methods and the inferential method reported here will be optimal under different circumstances.

Discussion

While PCR has achieved remarkable popularity, its application to obtain direct sequences of multiple alleles has been largely limited to studies of haploid or otherwise homozygous or hemizygous genetic loci. Samples from diploids may produce ambiguous sequences, so direct sequencing has been applied only rarely to diploid material. Population genetic theory shows that for realistic levels of DNA sequence diversity an acceptable sample size will afford an excellent chance of resolving the sequences of alleles in the sample. The method is particularly well suited to studies of DNA

polymorphism in organisms in which chromosome extractions cannot be done and in which small quantities of DNA must be used. Because direct sequencing obviates the need for cloning, massive volumes of sequence data can be obtained, and, as sequencing itself becomes mechanized, the demand for automating basic analyses will grow.

Perhaps the most encouraging aspects of the present report are the observation of the high frequency with which simulation data are successfully resolved and the relatively small sample sizes needed to resolve known sequences of *Drosophila* genes. It is also useful to note that, if a given sample fails to resolve all haplotypes, increasing the sample size is likely to improve the chances of resolving all alleles.

While intragenic recombination does not seem to affect, to any great degree, the ability of the algorithm to resolve haplotypes, it should be recognized that samples of large fragments (large θ) are very likely to admit anomalous matches and that in these cases the confidence in the inferred haplotypes is weakened. Simulation results suggest that the principle of parsimony (minimum number of orphans) will yield the most accurate assignment of haplotypes. The algorithm works even with samples in which there is sufficient recombination to produce all four gametic types for most pairs of polymorphic sites. While this may seem counterintuitive, the algorithm depends more on high-order linkage disequilibrium to resolve haplotypes. What is more important than pairwise disequilibria is that only a small fraction of the possible haplotypes are actually found in a given sample. Despite the power of the method, the parsimony rule is only an observation based on simulations, and, if absolute confidence in all haplotypes is necessary, this algorithm should only be applied in situations in which anomalous matches are unlikely.

In the worst case—when there are orphan alleles and/or unresolvable anomalous matches—these problems are more likely to occur with rare alleles, since common alleles are more likely to be resolved through homozygotes or short paths to homozygotes. Many of the tests one might want to perform on population data might still be applicable when all but a few rare alleles are identified. Gene genealogies can be constructed for the known alleles, and orphans might be placed in the tree by a minimum-distance rule. Many evolutionary inferences can be made from the distribution of polymorphic sites alone.

Acknowledgments

I thank Richard Hudson for sharing his computer programs for generating samples from the infinite-site model and for carefully critiquing the analyses. Clay Stephens and Masatoshi Nei provided helpful suggestions for the manuscript. Ken Kidd, Steve Schaeffer, and Tom Whittam provided engaging and encouraging discussion. This work was supported by NIH grant HD21963.

LITERATURE CITED

AGUADÉ, M., N. MIYASHITA, and C. H. LANGLEY. 1989*a*. Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics **122**:607–615.

———. 1989*b*. Restriction-map variation at the *zeste-tko* region in natural populations of *Drosophila melanogaster*. Mol. Biol. Evol. **6**:123–130.

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

AQUADRO, C. F., K. M. LADO, and W. A. NOON. 1988. The *rosy* region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. Genetics **119**:875–888.

BOEHNKE, M., N. ARNHEIM, H. LI, and F. S. COLLINS. 1989. Fine-structure genetic mapping of human chromosomes using the polymerase chain reaction on single sperm: experimental design considerations. Am. J. Hum. Genet. **45**:21–32.

COOKE, P. W., and J. G. OAKESHOTT. 1989. Amino acid polymorphisms for esterase-6 in *D. melanogaster*. Proc. Natl. Acad. Sci. USA **86**:1426–1430.

DEMPSTER, A. P., N. M. LAIRD, and D. RUBIN. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. J. R. Stat. Soc. [B] **39**:1–38.

EWENS, W. J. 1979. Mathematical population genetics. Springer, Berlin.

HUDSON, R. R. 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23**:183–201.

KIMURA, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**:893–903.

———. 1971. Theoretical foundation of population genetics at the molecular level. Theor. Popul. Biol. **2**:174–208.

KREITMAN, M. 1983. Nucleotide polymorphism at the *alcohol dehydrogenase* locus of *Drosophila melanogaster*. Nature **304**:412–417.

KREITMAN, M. E., and M. AGUADÉ. 1986*a*. Excess polymorphism at the *Adh* locus in *Drosophila melanogaster*. Genetics **114**:93–110.

———. 1986*b*. Genetic uniformity in two populations of *Drosophila melanogaster* revealed by filter hybridization of four-nucleotide-recognizing restriction enzyme digests. Proc. Natl. Acad. Sci. USA **83**:3562–3566.

LANGLEY, C. H., and C. F. AQUADRO. 1987. Restriction-map variation in natural populations of *Drosophila melanogaster*: *white*-locus region. Mol. Biol. Evol. **4**:651–663.

LANGLEY, C. H., A. E. SHRIMPTON, T. YAMAZAKI, N. MIYASHITA, Y. MATSUO, and C. F. AQUADRO. 1988. Naturally-occurring variation in the restriction map of the *Amy* region of *Drosophila melanogaster*. Genetics **119**:619–629.

LEIGH BROWN, A. J. 1983. Variation at the 87A heat-shock locus in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **80**:5350–5354.

LI, H., U. B. GYLLENSTEN, X. CUI, R. K. SAIKI, H. A. ERLICH, and N. ARNHEIM. 1988. Amplification and analysis of DNA sequences in single human sperm and diploid cells. Nature **335**:414–417.

MIYASHITA, N., and C. H. LANGLEY. 1988. Molecular and phenotypic variation of the *white* locus region in *Drosophila melanogaster*. Genetics **120**:199–212.

NEI, M., and A. L. HUGHES. Polymorphism and evolution of the major histocompatibility complex loci in mammals. *In* R. K. SELANDER, A. G. CLARK, and T. S. WHITTAM, eds. Evolution at the molecular level. Sinauer, Sunderland, Mass. (accepted).

RUANO, G., K. K. KIDD, and J. C. STEPHENS. Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. Proc. Natl. Acad. Sci. USA (accepted).

SAIKI, R. K., S. SCHARF, F. FALOONA, K. B. MULLIS, G. T. HORN, H. A. ERLICH, and N. ARNHEIM. 1985. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science **230**:1350–1354.

SCHAEFFER, S. W., C. F. AQUADRO, and W. W. ANDERSON. 1987. Restriction-map variation in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Mol. Biol. Evol. **4**:254–265.

SCHAEFFER, S. W., C. F. AQUADRO, and C. H. LANGLEY. 1988. Restriction-map variation in the *Notch* region of *Drosophila melanogaster*. Mol. Biol. Evol. **5**:30–40.

SCHARF, S. J., G. T. HORN, and H. A. ERLICH. 1986. Direct cloning and sequence analysis of enzymatically amplified genomic sequences. Science **233**:1076–1078.

SIMMONS, G. M., M. E. KREITMAN, W. F. QUATTLEBAUM, and N. MIYASHITA. 1989. Molecular analysis of the alleles of *alcohol dehydrogenase* along a cline in *Drosophila melanogaster*. I. Maine, North Carolina and Florida. Evolution **43**:393–409.

STEWART, F. M. 1977. Computer algorithm for obtaining a random set of allele frequencies in an equilibrium population. Genetics **86**:482–483.

WATTERSON, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **6**:217–250.

MASATOSHI NEI, reviewing editor

Received June 27, 1989, revision received September 29, 1989

Accepted November 8, 1989