Statistical Science
1994, Vol. 9, No. 3, 320–333

Analysis of Genetic Data from the Polymerase Chain Reaction

W. Navidi and N. Arnheim

Abstract. The polymerase chain reaction (PCR) makes possible rapid generation of a very large number of copies of a specific region of DNA. This enables the typing of quantities of DNA as small as a single molecule. PCR has led to the development of laboratory experiments which provide new approaches to many classic problems in genetics, such as estimation of linkage, marker ordering and genetic disease diagnosis. We describe some of these experiments and the statistical techniques that have been used to design them and to analyze the data they produce.

Key words and phrases: Polymerase chain reaction, gene mapping, gene ordering, genetic disease diagnosis.

1. INTRODUCTION

1.1 Description of DNA

The macromolecule DNA is the basic storage vehicle for all hereditary information. A single strand of DNA consists of a string of bases attached to a sugar–phosphate backbone. There are four types of bases: adenine (A), guanine (G), cytosine (C) and thymine (T). The DNA molecule is composed of two such strands, with each base on one paired with a base on the other via hydrogen bonds. The base A is always paired with T, and G with C, so that the sequence of bases on one strand determines the sequence of its complementary strand. All hereditary differences among individuals result from differences in the sequence of base pairs in their DNA.

A fundamental property of DNA is its ability to replicate. During replication, the two strands separate, and a new complementary strand is synthesized on each. This process is catalyzed by enzymes known as DNA polymerases. Replication occurs in one direction only. New bases are added only to the end of the strand known as the 3′ end, which terminates with a hydroxyl group. The other end terminates with a phosphate group and is known as the 5′ end.

Human cells other than sex cells are diploid, which means that their nuclei contain two copies of each chromosome. Each complete chromosome set, or genome, contains 23 members in humans and is composed of about 3×10^9 base pairs of DNA. The DNA in any cell is for all practical purposes identical to the DNA in any other cell from the same individual, but differences can exist between the two members of a pair of chromosomes, since each contains the genetic information inherited from a different parent. Regions of the DNA in which the sequences of the two chromosomes may potentially differ are called polymorphic. Each distinct sequence which may appear at a particular polymorphic site is called an allele. An individual whose chromosomes carry different alleles at a locus is heterozygous at that locus; one whose chromosomes carry the same allele is homozygous. Within a population, the number of alleles to be found at a polymorphic site may be as few as two or as many as several thousand. Many situations arise in which it is desirable to identify the allele present at a given site on a given chromosome. In practice it often happens that only a few copies (or even a single copy) of the genome of interest are available. For example, the ability to type a single cell can be crucial to diagnosing genetic diseases before birth (see Section 3). Typing individual sperm cells can be of great importance in developing a genetic map which will assist in locating disease-causing and other important human genes (see Section 2). In forensic applications, sometimes only a very small amount of DNA can be recovered from a crime scene for use in helping to identify the criminal. Without a laboratory procedure to amplify the number of copies of a specific region of the genome, accurate typing in these situations would be impossible.

W. Navidi is Associate Professor of Biostatistics, Department of Preventive Medicine, University of Southern California, Los Angeles, California 90033. N. Arnheim is Professor, Department of Biological Sciences, University of Southern California, Los Angeles, California 90089.

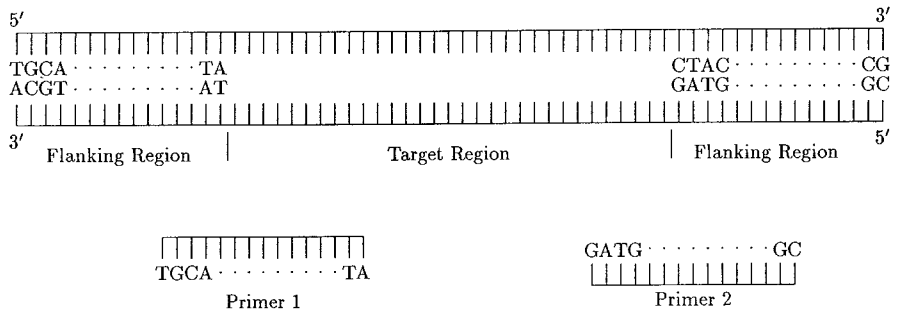

FIG. 1. *A schematic representation of a DNA fragment. Each primer sequence is designed to anneal to the 3′ end of one of the strands.*

1.2 The Polymerase Chain Reaction

The polymerase chain reaction (PCR) (Saiki et al., 1985, 1988; Mullis and Faloona, 1987) is a procedure by which DNA fragments are made to replicate many times in a test tube. Within a few hours, a single DNA fragment can be amplified to billions of copies. This PCR product contains enough DNA to be typed by any of several conventional techniques.

Figure 1 shows a DNA fragment. The target region, usually several hundred but potentially up to about 10,000 base pairs in length, is the region to be amplified. Short flanking regions, usually 20 base pairs in length, have already been sequenced. The flanking regions must not be polymorphic, since their sequences must be known from studies of other members of the population (human or other species). Typically the target region is polymorphic; the purpose of PCR amplification is to produce enough copies so the alleles can be identified. A large number (about 10^{13}) of primer sequences are chemically synthesized, each identical to a flanking region on one of the strands on one side of the target.

PCR is carried out by depositing the DNA fragment or fragments to be amplified into a reaction tube along with primer sequences, reagents and DNA polymerase. PCR is a cyclical process; the number of copies of the target increases at most by a factor of 2 at each cycle. Figure 2 is a schematic diagram of the first few cycles of PCR, applied to a single fragment. Figure 2a shows a double-stranded DNA molecule, with a box indicating the target region. To begin the cycle, the mixture is heated to about 95°C, which causes the two strands of DNA to separate (denature). Then the mixture is cooled to about 60°C, which provides favorable conditions for complementary strands to reanneal. It is extremely unlikely that the two strands of the original fragment will find each other to reanneal. Instead each will anneal with one of the much more numerous primer sequences, as shown in Figure 2b. The cycle is completed by reheating the mixture to about 72°C, whereupon the polymerase catalyzes the formation of new complementary strands (Figure 2c). Notice that the primers extend in one direction only (from the end with the asterisk). Figure 2d shows the product after two cycles. Each of the two original strands has been replicated as in Figure 2c. In addition, the new strands formed in the first cycle have been replicated. Because the primers extend in one direction only, these replicates terminate at the ends of the flanking regions. All the DNA strands in the final product except the two original strands and those strands replicated directly from them will terminate at the flanking regions. Thus, after many cycles, the final PCR product will consist almost entirely of DNA fragments as shown in Fig-

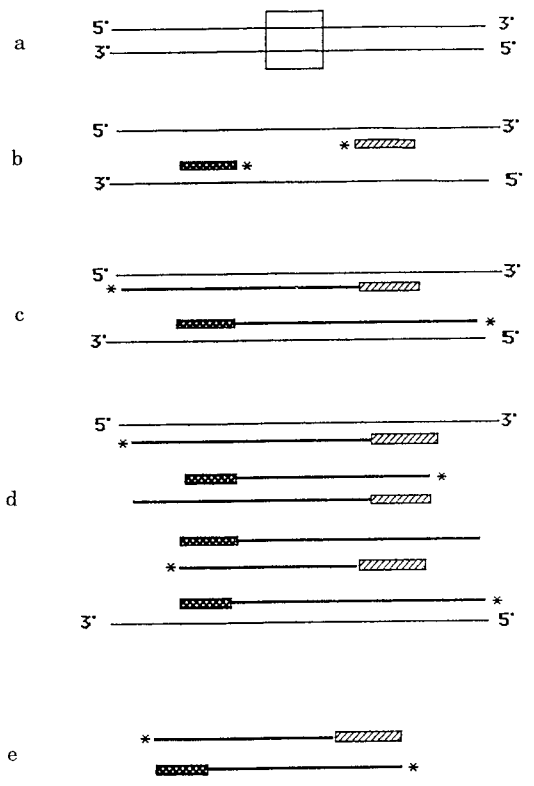

FIG. 2. (a–d) *The first two cycles of PCR*; (e) *a typical DNA fragment in the final PCR product.*

ure 2e. The maximum number of cycles feasible is about 50.

PCR became easy to carry out with the introduction of a DNA polymerase obtained from a thermophilic bacterium called *Thermus aquaticus* (Saiki et al., 1988). Known as *Taq* polymerase, it is unique among DNA polymerases in retaining its catalyzing ability after being subjected to the high temperature necessary to denature the DNA strands. Each cycle of PCR potentially doubles the number of copies of the target. In practice the increase is much less, because the process does not succeed with every fragment on every cycle. In one experiment, 50 cycles of PCR applied to a diploid amount of DNA (two molecules) resulted in a geometric mean increase per cycle of about 65% (Li et al., 1988).

Several conventional procedures exist for typing DNA. Methods involving allele-specific probes are appropriate when all commonly occurring alleles have been sequenced. Portions of the PCR product are brought into contact with synthesized DNA fragments (probes) corresponding to the various alleles. By determining to which probes the PCR product anneals, the alleles present in the product are identified. When the alleles differ from each other in length, the PCR product may be subjected to gel electrophoresis, which sorts the fragments by length. For more information on these and other typing procedures, see Saiki et al. (1986), Kogan, Doherty and Gitschier (1987), Ugozzoli and Wallace (1991) and Landegren et al. (1988).

The PCR process is subject to occasional errors which can result in mistyping the product. The errors fall into four categories: efficiency, contamination, deposit and mutation. Efficiency refers to the fact that sometimes a fragment subjected to PCR will fail to yield sufficient product to be detectable. Reasons for this are not known with certainty. In some cases, perhaps, the DNA is degraded or broken before the reaction begins. It might also happen that the fragment adheres to the side of the reaction tube and fails to denature. The efficiency of a PCR reaction is defined to be the probability that a single fragment produces a detectable level of product. In many PCR experiments, the efficiency is 95% or more, but efficiencies as low as 80% are not uncommon. In general, if several fragments are amplified simultaneously, it is assumed that they act independently in terms of efficiency; so if r represents the efficiency of the PCR reaction, then the probability that a reaction starting with n fragments will produce a detectable product is $1 - (1 - r)^n$. This also assumes that there is negligible probability that the n fragments together will produce enough product to be detected when none does individually. It is easy to see how efficiency errors can cause mistyping. Assume DNA from a single diploid cell from a heterozygous individual is amplified. Denoting the two alleles A and a, if the fragment containing the A allele is successfully amplified while the a allele is not, the individual will be incorrectly typed as homozygous AA.

Contamination is the most difficult error to deal with in terms of assessing impact on typing accuracy. Contamination occurs when exogenous DNA enters the reaction tube along with the target DNA. The contaminating fragments will be amplified along with the target, which may result in mistyping. The impact of contamination depends on the ratio of the number of contaminating molecules to the number of initial target molecules. If the number of contaminating molecules is much less, the typing signal from the contaminant is likely to be much weaker than that of the target, and thus indistinguishable from background noise. If the number of contaminating molecules is considerably greater, it is likely that the contaminant will be detected and the target not. If the number of molecules of both target and contamination are of the same order of magnitude, both are likely to be detected. Situations where the number of molecules of contaminant is too small to be detected can be ignored, since they do not cause typing errors. Further, it is thought that large numbers of contaminating molecules can be almost entirely avoided by carefully following appropriate laboratory procedures. Therefore it is usually assumed that, when contamination occurs, both target and contaminant occur in equal numbers. Certain types of contamination can be detected experimentally. For example, it is common practice to include several control tubes in a PCR experiment, into which no DNA is placed. If no signal is detected in the product from these tubes, then contamination events which would be likely to affect all tubes, such as contamination of reagents, may be ruled out. It is generally assumed for modeling purposes that when no contamination is found in the controls, the DNA-containing tubes are independent with respect to contamination. The frequency of contamination varies from experiment to experiment, depending on factors such as the laboratory setup and the skill of the technician conducting the reaction. Estimates of the contamination rate in well-run experiments range from 0 to 7% per tube (Cui et al., 1989; Goradia et al., 1991). In certain situations, mistyping due to contamination may be almost entirely avoided by using the so-called multiple-tubes approach (Navidi, Arnheim and Waterman, 1992), which involves apportioning the target DNA among several tubes. This is discussed in Section 3.

The third category of error is deposit error, which can occur in experiments in which individual cells such as sperm cells, oocytes or embryonic cells must

be micromanipulated into reaction tubes. The goal is usually to place a single cell in a tube, but there is positive probability that no cell or more than one cell will be deposited.

The fourth category of PCR error is mutation or misincorporation error. It has been estimated that during the extension phase of PCR by DNA polymerase the proportion of bases incorrectly specified per PCR cycle is between 10^{-5} and 10^{-4}, on average (Gelfand and White, 1990). Thus if the target region is 1,000 base pairs in length, then on average between 1 and 9.5% of fragments created during a given cycle will differ in at least one site from their immediate ancestors, assuming independence of mutation events across bases. As a result, many of the fragments in the final PCR product will not be identical to the original target. In general this does not lead to errors in typing PCR product when typing is done by allele-specific probes, because it is very unlikely that a fragment bearing a misincorporation error will have a sequence which happens to correspond to an allele. Misincorporation errors can pose a serious problem when the goal is to sequence the target region from one molecule of product which has been cloned, and in some other circumstances. To assess the impact of these errors, branching process models have been proposed for PCR. These are discussed in Section 5.

In Section 2, we describe some models for PCR in genetic mapping. In Section 3 we describe models for the use of PCR in genetic disease diagnosis and in forensics. In Section 4 we describe and analyze the impact of the PEP procedure, an elaboration of PCR which enables simultaneous amplification of a large portion of the human genome. In Section 5 we describe some branching process models for PCR.

2. GENETIC MAPPING

2.1 Recombination

Consider two polymorphic sites in a chromosome. Locus 1 has two alleles A and a, and locus 2 has two alleles B and b. Figure 3 shows the pedigree of a hypothetical family. The grandparents (1 and 2) are each doubly homozygous; their son (3) is doubly heterozygous. Individual 4, the wife of 3, is doubly homozygous. The mating between the doubly heterozygous 3 and the doubly homozygous 4 is called a double backcross. The set of alleles at different loci inherited from the same parent is called the haplotype. Knowing the genotypes of his parents, it is clear that individual 3 inherited haplotype AB from his father and haplotype ab from his mother. A doubly heterozygous individual from a different family might have haplotypes Ab and aB. These two possible

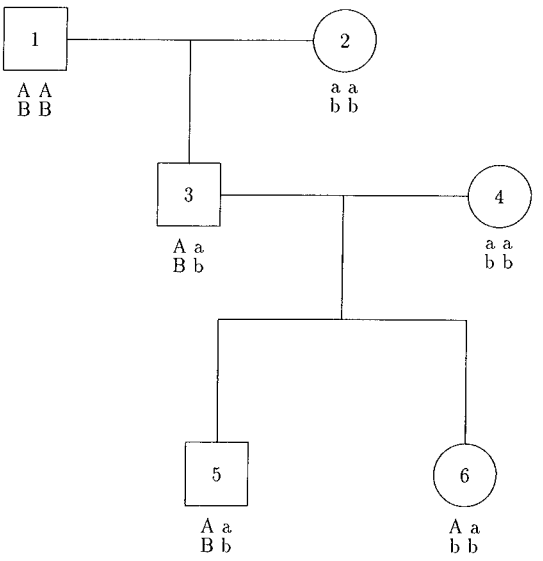

FIG. 3. *Individual 5 is a parental type; individual 6 is a recombinant.*

pairs of haplotypes for a doubly heterozygous individual are called the two phases. In general, the phase cannot be directly determined; it can only be inferred from sufficient knowledge of ancestral genotypes. Individual 3 has two children, 5 and 6. Individual 5 inherited haplotype AB from his father, while individual 6 inherited Ab. The haplotype AB is called a "parental" or "nonrecombinant" type, since it is the haplotype of the parent from whom it was inherited. The haplotype Ab is called a "recombinant" type. The "recombination fraction" between loci 1 and 2 is the probability that a parent will pass on a recombinant haplotype to an offspring. This probability differs somewhat between men and women, and it may also vary slightly among individuals of the same sex. The value of the recombination fraction is determined by the stochastic nature of random events which occur during the process of meiosis, the production of sex cells (sperm and eggs).

Sex cells are produced from so-called germ line cells, which are diploid. Before meiosis begins, the DNA replicates, so that at the beginning of meiosis, each chromosome consists of two identical fragments of DNA, called chromatids, joined at a point called the centromere. Each chromosome can be aligned with the other member of its chromosome pair. The two chromosomes may differ from each other, as one contains information inherited from one's father, and the other contains corresponding information inherited from one's mother. The chromosomes exchange genetic material, in a process known as crossing-over. Figure 4 is a schematic diagram of a crossover event. Crossovers always involve one chromatid from each chromosome. It is quite possible for several crossover

events to occur on a single chromosome during meiosis, involving different pairs of chromatids. Thus the chromatids after the completion of meiosis (now called chromosomes again) may consist of alternating stretches of genetic material from each of the original chromosomes. Each sex cell (sperm or egg) contains one chromatid chosen at random from each of the 23 sets of four.

If two loci are on different chromosomes, the alleles have probability $\frac{1}{2}$ of having originated from the same parental genome. Thus the recombination fraction is $\frac{1}{2}$. When two loci are on the same chromosome, the recombination fraction is determined by the stochastic nature of the crossing-over process. In general, the intensity of the crossover process varies considerably from region to region of the genome, so the recombination fraction between two loci cannot be reliably determined from the physical distance between them. The recombination fraction will be nearly zero for loci right next to each other, and for loci near opposite ends of a chromosome it will approach $\frac{1}{2}$. A pair of loci with a small recombination fraction is said to be tightly linked, since haplotypes are usually passed on intact. Loci with a recombination fraction of $\frac{1}{2}$ are said to be unlinked. Procedures to estimate recombination fractions are referred to as linkage analysis.

One purpose for estimating recombination fractions is to create a genetic map of the genome or a region of it. In a genetic map, the distance between pairs of loci is defined to be the expected number of crossovers per chromatid between the loci during meiosis. Since crossovers cannot be directly observed, this quantity must be estimated by transforming an estimate of the recombination fraction. When two loci are close together, the probability of more than one crossover between them is negligible, and the map distance is well approximated by the recombination fraction itself. For larger distances, the sites on the chromatid where crossovers occur must be modeled as a point process which specifies a functional relationship between map distance and recombination fraction. Such functions are called map functions and have been studied for nearly a century. Some recent references are Karlin and Liberman (1979) and Evans, McPeek and Speed (1993).

Another purpose is to study the variation in the crossover frequency along the length of a chromosome. Of particular interest is the location of so-called hot-spot regions, regions of unusually intense crossover activity. For example, on average the recombination fraction in humans between loci 10^6 base pairs apart is 0.01. Imagine that an otherwise average region 10^6 base pairs in length contains a subregion 10^3 base pairs long in which the crossover rate is 100 times the average. The recombination fraction between the loci at the ends of this region will be about 0.011. Thus, to determine the existence of this "hot spot," we must be able to estimate the recombination fraction to within an accuracy of 0.001. This illustrates the fact that locating hot spots requires extremely precise estimates of recombination fractions, which in turn require very large samples.

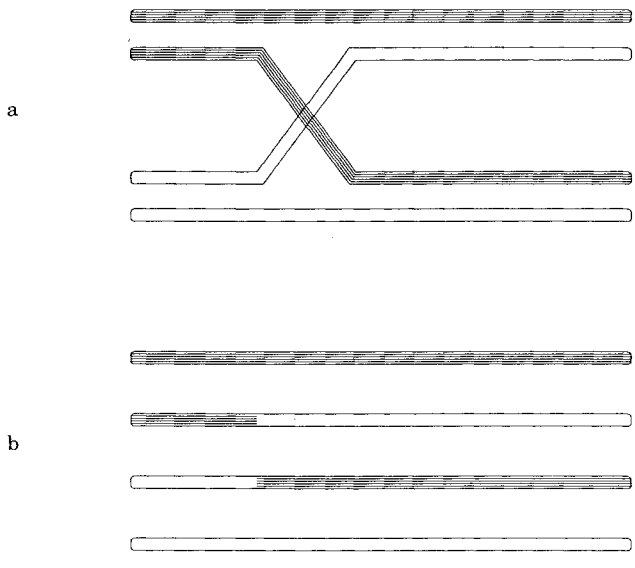

FIG. 4. *A schematic representation of a crossover event*: (a) *two chromatids exchange genetic material*; (b) *after the crossover, two of the chromatids contain genetic material from only one parent, while the other two contain material from both.*

2.2 The Family Method

The original and still indispensable method of linkage analysis is the family method, so called because it involves studying the inheritance of haplotypes within members of a family. Figure 5 shows an example of a two-generation family. Assume we know that the mother is doubly homozygous and the father is doubly heterozygous, with haplotypes AB and ab. If θ denotes the recombination fraction between the two loci and if n denotes the number of offspring, the number of recombinant offspring from this mating is binomial with parameters n and θ. The recombination fraction can be estimated by maximum likelihood. When the phase of the father is unknown, it is common to assign prior probabilities of $\frac{1}{2}$ to each phase and obtain a mixture of binomials for the likelihood. The assumption that each phase occurs equally often in the population is known as the hypothesis of linkage equilibrium. When the probability of observing a given set of offspring genotypes is a function of θ, the mating is said to be informative for linkage. If

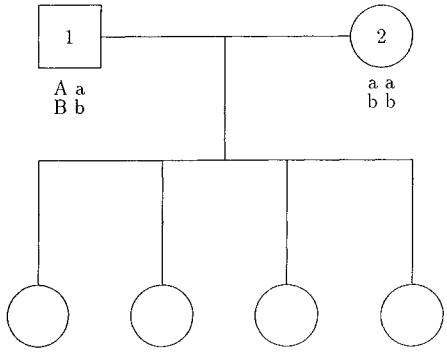

FIG. 5. *The number of recombinants among the four offspring is distributed* Bin(4, θ), *where* θ *is the recombination fraction.*

the father were not doubly heterozygous, the mating would be uninformative.

The family method of linkage analysis involves obtaining genetic data on a number of families. The pedigrees are often large, involving several generations. Many of the matings are likely not to be informative, and data are likely to be missing on some individuals. The idea is to construct a likelihood function for θ for each family. Usually it is appropriate to assume that families are independent, so that inferences about θ can be based on the product of the family-specific likelihoods. Techniques for obtaining likelihood functions can be quite involved. The primary limitation of the family method is lack of data. Existing data are sufficient for estimating recombination fractions with a resolution no greater than 0.01. For a detailed introduction to this subject see Ott (1991).

2.3 Sperm Typing

2.3.1 *Estimating recombination fractions*

In the example in Figure 5, offspring are classified as recombinant or not by comparing the haplotype inherited from their father with the father's haplotype. Determining the haplotype inherited from the father is equivalent to determining the haplotype of the sperm cell which was involved in the offspring's conception. Thus each sperm from a doubly heterozygous man contains as much linkage information as an offspring from an informative mating. PCR enables the DNA in a single sperm cell to be amplified to a level sufficient for typing. Since a man produces essentially unlimited numbers of sperm, the ability to type sperm should make it possible to estimate linkage with nearly unlimited accuracy. Of course, sperm typing can produce only a male meiotic map. This does not matter for gene ordering, since the order is the same in both sexes. Recombination fractions, however, differ between men and women.

TABLE 1
Results of typing 708 *sperm**

Observed sperm type	Total
1 ____	89
2 ___b	9
3 __B_	11
4 __Bb	1
5 _a__	18
6 _a_b	33
7 _aB_	227
8 _aBb	15
9 A___	21
10 A__b	196
11 A_B_	47
12 a_Bb	9
13 Aa__	0
14 Aa_b	6
15 AaB_	10
16 AaBb	16

*Types 7 and 10 are parental types; types 6 and 11 are recombinants; all other types result from PCR errors. Reproduced from Cui et al. (1989).

In a sperm typing study, individual sperm cells from a doubly heterozygous donor are placed into reaction tubes where the DNA is extracted from the cell, amplified by PCR and typed. Li et al. (1988) showed that it is possible to amplify two such loci simultaneously in a genome from a human sperm and to determine the haplotype. It is quite feasible to type more than a thousand sperm in a single experiment. Since PCR is not error free, some of the sperm are likely to be mistyped. It is best to estimate the recombination fraction by constructing a model to account for PCR error. Such a model was proposed by Cui et al. (1989), who estimated the recombination fraction between two loci on chromosome 11 by typing 708 sperm from two males. Let Aa and Bb denote the alleles at the loci, and let θ denote the recombination fraction. The two males were each doubly heterozygous with haplotypes Ab and aB. The four possible sperm types are parental types Ab and aB, each with probability $(1-\theta)/2$, and recombinants AB and ab, each with probability $\theta/2$. Because of the possibility of PCR errors, the PCR product in a given tube will produce one of 16 observations, corresponding to each of the four alleles A, a, B, b being present or absent. For example, the three alleles A, a and B would be detected if two sperm of types AB and aB entered the tube due to deposit error, or if a sperm of type aB entered the tube and the reaction was contaminated by a fragment containing the A allele, or if two sperm of types AB and ab entered the tube and the b allele failed to be amplified to a detectable level, among other possibilities. Cui et al. (1989) obtained the following results for 708 sperm (Table 1).

To account for the complexity of the data, Cui et al. (1989) postulated a 14-parameter model containing the recombination fraction θ as the parameter of interest, parameters $\gamma_n, n = 0, 1, 2, 3, 4$, representing the probability that n sperm are deposited in a tube, parameters $\alpha_A, \alpha_a, \alpha_B$ and α_b representing the probabilities that regions corresponding to alleles A, a, B and b will be amplified to a detectable level and four contamination probabilities $\beta_A, \beta_a, \beta_B$ and β_b representing the probabilities that a fragment carrying allele A, a, B or b, respectively, contaminates the reaction and is detected. Additionally, it is assumed that the outcome in each tube is statistically independent of the outcomes in the other tubes, so that the counts in Table 1 are multinomial. Calculation of the multinomial probabilities in terms of the 14 parameters is tedious but straightforward under the assumptions that, within a tube, distinct loci amplify independently, and independently of contamination and of the number of sperm in the tube. Cui et al. (1989) calculate the MLE's of the recombination fraction θ and of the nuisance parameters. Their results are reproduced in Table 2. It is interesting to note that a crude estimate of θ can be obtained by considering only the tubes in categories A_B_, A_b, _aB_ and _a_b, which ostensibly show one allele at each locus. If we assume that no errors were made, the MLE for θ based on these tubes is

$$\hat{\theta} = \frac{n_{AB} + n_{ab}}{n_{AB} + n_{aB} + n_{Ab} + n_{ab}} = \frac{80}{503} = 0.159,$$

with standard error $\sqrt{\hat{\theta}(1-\hat{\theta})/503} = 0.0163$. This is in very good agreement with the MLE and Fisher SE based on the full data set. Such good agreement is often but not always found in studies of this sort. More recent studies in which sperm typing has successfully been used to estimate linkage are reported in Hubert et al. (1992) and Lewin et al. (1992).

The practical applications of sperm typing differ from those of the family method. The primary application of the family method is to locate disease-causing genes. For example, assume a disease is dominant and let A be the disease allele, so that a person with genotype AA or Aa will have the disease, while a person with genotype aa will be disease-free. Let B and b be the alleles at a *marker* locus, a locus whose position on the genome is known and which it is thought might be close to the disease locus. Assume we can type people at the B locus by standard methods. For the disease locus, we know that disease-free people have type aa; for people with the disease, we distinguish between AA and Aa by obtaining disease information on their ancestors. The situation in Figure 5 would arise if the father was diseased, but had one nondiseased parent and the mother was not diseased. Offspring who are diseased must have type Aa at the disease locus; those not diseased have type aa. Given enough families with offspring from informative matings, we can accurately estimate the recombination fraction between the A and B loci. If the recombination fraction is small, the disease gene is known to lie in the vicinity of the marker locus.

TABLE 2
*Parameter MLE's based on 708 sperm**

Parameter	MLE (SE)
α_A	0.9731 (0.0187)
α_a	0.9679 (0.0181)
α_B	0.9483 (0.0205)
α_b	0.9012 (0.0243)
β_A	0.01403 (0.0151)
β_a	0.007239 (0.0131)
β_B	0.02839 (0.0169)
β_b	0.03698 (0.0170)
γ_0	0.1352 (0.0145)
γ_1	0.7878 (0.0218)
γ_2	0.07705 (0.0205)
γ_3	0.0
γ_4	0.0
θ	0.1618 (0.0168)

*α_A, α_a, α_B, α_b are probabilities that DNA fragments containing alleles A, a, B, b, respectively, are amplified to a detectable level. β_A, β_a, β_B, β_b are probabilities that a reaction is contaminated with a fragment containing allele A, a, B, b, respectively. γ_0, γ_1, γ_2, γ_3, γ_4 are probabilities that 0, 1, 2, 3, 4 sperm, respectively, are placed into a tube. θ is the recombination fraction between loci A and B. Reproduced from Cui et al. (1989).

Sperm typing, unlike family studies, cannot be used to locate disease genes directly, because the sperm cells do not manifest the disease symptoms as do members of a family. However, because of the potentially unlimited number of meioses which can be studied, sperm typing can produce a dense map of marker loci which will make it easier to localize disease genes with family studies. In addition, samples large enough to be useful in locating recombination "hot spots" can be obtained. The limited data available make the family method unsuitable for these purposes.

2.3.2 *Ordering markers*

Constructing a genetic map involves determining the linear order of a large number of loci. If a sufficiently accurate method for determining the order of three loci is available, then it is easy to see that arbitrary numbers of loci can be ordered by its repeated application. For a given map of $n-1$ ordered loci, an nth locus can be placed in its proper order among them by determining its position with respect to a sufficient number of pairs of the original $n-1$.

Boehnke et al. (1989) and Goradia and Lange (1990) discuss methods for ordering a new locus relative to a given set of loci. Since accurate ordering of n loci requires many three-locus orderings, it is important that a three-locus ordering procedure be nearly error free.

We discuss the problem of ordering three loci using sperm typing. We will assume that the three loci are rather tightly linked, since that is the case where the family method fails to provide enough data. Data can be collected from sperm derived from a male who is triply heterozygous with genotypes Aa, Bb and Cc at the three loci. Each sperm will be one of eight types: (1) ABC; (2) abc; (3) ABc; (4) abC; (5) AbC; (6) aBc; (7) Abc; (8) aBC. Two of these types are parental types, four are single recombinants and two are double recombinants. Assume the phase, or parental haplotype, is ABC/abc. In practice this would not be known, but since parental types are far more common than the others when the loci are tightly linked, it becomes obvious upon typing a few sperm. The identity of the double recombinant depends on the order of the loci. We do not distinguish right from left, so order ABC is the same as CBA. Thus, ordering three loci is equivalent to determining which locus is in the middle. If locus B is in the middle, then types 3, 4, 7 and 8 are single recombinants, representing recombination events between loci B and C or between loci A and B, and types 5 and 6 are double recombinants, obtainable only when recombination occurs between both A and B and between B and C. If locus A is in the middle, then types 7 and 8 are the double recombinants, and if locus C is in the middle, then types 3 and 4 are the double recombinants. Assume that locus B is in the middle. Let θ_1 be the recombination fraction between A and B, and let θ_2 be the recombination fraction between B and C. If we assume that crossovers occur independently in disjoint regions of the genome (no chiasma interference), and that, if more than one crossover occurs in a meiosis, the chromatids involved in each are selected independently (no chromatid interference), the expected proportion of sperm of each type is

$$
\begin{aligned}
P[\text{ABC or abc}] &= (1-\theta_1)(1-\theta_2),\\
P[\text{aBC or Abc}] &= \theta_1(1-\theta_2),\\
P[\text{ABc or abC}] &= (1-\theta_1)\theta_2,\\
P[\text{aBc or AbC}] &= \theta_1\theta_2.
\end{aligned}
$$

Corresponding probabilities for other orders are obtained by interchanging the appropriate letters. Since $\theta_1\theta_2$, the probability of a double recombinant, is the smallest of the four probabilities (this is true even without the assumption of no chiasma interference), the problem reduces to that of selecting the least common category from a multinomial distribution. Boehnke et al. (1989) were the first to propose ordering loci by sperm typing. Goradia and Lange (1990) discuss sequential strategies for inferring the least common type. They recommend typing sperm until the number of observations of the least frequently observed type is s fewer than the next least frequently observed. Reasonable values of s might be 2 or 3. Expected stopping times and error probabilities depend on the recombination fractions between the three pairs of loci AB, AC and BC. As these recombination fractions decrease, the stopping time increases and the error rate decreases. Goradia and Lange (1990) calculate close approximations to these quantities under the assumption of no PCR errors. As an example of their results, if the recombination fractions are both equal to 0.01, then, using $s = 3$, the expected number of sperm to type before a decision is reached is about 400 and the error probability is about 1.3×10^{-6}.

Goradia et al. (1991) describe a method for ordering three loci which takes PCR errors into account. It involves a model similar to the one proposed by Cui et al. (1989) and discussed above in the section on estimating recombination fractions. Since there are six alleles involved, there are $2^6 = 64$ possible typing results when PCR errors are taken into account. Given the phase and correct order of the three loci, the likelihood is a function of parameters reflecting the recombination fractions between pairs of loci, and the probabilities of efficiency, contamination and deposit errors. The likelihood is maximized for each phase–order combination. As before, the phase is usually obvious, and the task is to select among the three orders within a given phase by comparing the maximum values of the likelihoods under different orders. This presents the difficulty of interpreting likelihood ratios for nonnested models. Guerra et al. (1992) and Stephens and Smith (1992) suggest Bayesian approaches to this problem which could be adapted to the analysis of PCR data.

The model of Goradia et al. (1991) becomes computationally impractical when more than three loci are involved. Lazzeroni et al. (1993) propose a model in which multipoint calculations are feasible for a much larger number of loci. The model requires the assumptions of no chiasma or chromatid interference and that no more than two sperm are deposited into any tube.

Consider n loci, numbered $1, 2, \ldots, n$ in order. Let θ_i represent the recombination fraction between loci i and $i + 1$. Under the assumption of no interference, the probability of observing a given haplotype on a given sperm can be expressed in terms of the parameters θ_i and nuisance parameters that represent the efficiency, contamination and deposit error

rates. Maximum likelihood estimates of the parameters can be computed and maximum values of the likelihood can be compared for different orderings of the loci.

This model has at least two appealing features. First, it is not necessary that any one donor be heterozygous at all n loci. Under the assumption of no interference, the recombination fraction between any two of the n loci can be expressed in terms of the θ_i. Lazzeroni et al. use this fact to derive a likelihood function in the case where several donors are heterozygous at different subsets of the n loci. This is valuable because it may be difficult to find donors simultaneously heterozygous at several loci, and because it enables one to extend a map by combining data from a new donor, involving additional loci, with data from previous donors. A second appealing feature is that the likelihood function can be expressed in a form which allows its calculation by standard pedigree methods such as are found in the MENDEL package (Lange, Boehnke and Weeks, 1988).

Dear and Cook (1993) describe an ordering strategy not based on genetic recombination in which DNA is physically broken in vitro, rather than during meiosis. DNA is extracted from diploid cells (e.g., blood cells) and placed in a solution. Random breaks are introduced either by irradiation or by agitating the solution to shear the DNA into fragments. The solution is then diluted and aliquots are poured into reaction tubes. In each tube, the number of copies of any given locus is assumed to be Poisson with known mean, controllable through the dilution. Best results seem to occur when the mean is about 1. Each tube is subjected to PCR and typed for presence or absence of the loci of interest. The principles behind the procedure are, first, that the closer two markers are to each other, the more likely they are to be either both present or both absent on a tube and, second, that given three markers A, B and C, with B in the middle, then in tubes where B is not present, the presence of A will be independent of the presence of C.

Say that loci $L_1, \ldots, L_n$ are to be ordered. The data consist of typing results indicating the presence or absence of each marker in each tube. Given an ordering, the probability of any combination of markers present can be expressed as a function of the probabilities of breaks in between pairs of loci and the known mean number of copies of each locus per tube. In principle, probabilities of efficiency and contamination errors should be taken into account as well. The likelihood is maximized for each ordering, and the order with largest maximized likelihood is preferred. Problems of comparing likelihoods for nonnested models remain.

Ordering loci by using physically broken DNA has some potential advantages over ordering by genetic methods. Nonpolymorphic loci can be ordered, in contrast to genetic methods where heterozygous individuals must be found. However, statistically more rigorous methods for data analysis must be established.

2.4 Areas for Further Research

An advantage of sperm typing over the family method is that, because of the theoretically unlimited number of sperm available for typing, very small recombination fractions can be estimated. This enables ordering of closely spaced loci and accurate location of regions of unusually intense crossover activity. In practice it is likely that several experiments will be done, at different times and perhaps in different laboratories, each attempting to estimate the same recombination fraction. If the recombination fraction is small, it is likely that none of the MLE's will be close to normally distributed, and Fisher information will not give accurate estimates of their standard errors. Some experiments may yield MLE's of 0. It would be desirable to combine the results of such experiments to obtain a more efficient and approximately normally distributed estimate of the recombination fraction. The difficulty is that the nuisance parameters of efficiency, contamination and deposit error will vary from experiment to experiment, especially if experiments are done in different laboratories. Thus the combined counts of the observed types will not be multinomial. It is not clear how best to combine the results of such experiments.

The approach of Dear and Cook (1993) raises several issues. Optimal concentrations of fragments per aliquot need to be worked out. Likelihood functions incorporating the probability of PCR errors need to be constructed. The question of optimal length of broken fragments needs to be addressed as well. Too many breaks will disassociate all loci and destroy all order information. Too few breaks will fail to separate loci often enough. To some extent, the amount of breakage can be controlled by varying the breaking technique. In addition, it may be possible to sort fragments by length after breakage and to choose those within a narrow length range.

3. GENETIC DISEASE DIAGNOSIS AND FORENSICS

3.1 Preimplantation Genetic Disease Diagnosis

Some disease-causing alleles have been sequenced and thus can be detected through genotyping. Examples include the genes for cystic fibrosis and sickle-cell anemia. A couple at risk for bearing a child with a genetic disease may choose to reduce this risk by in vitro fertilization followed by implanting the em-

bryo in the mother's uterus after it has been determined that it is not diseased. This determination can be made by determining the genotype of the egg cell prior to fertilization or by typing the embryo itself. Both methods involve PCR to amplify DNA from one or a few cells for typing. We discuss the latter procedure, known as blastomere analysis. A more complete discussion of this and other procedures can be found in Winston and Handyside (1993). In blastomere analysis, the embryo is typed when it consists of about four to eight cells, called blastomeres. It is possible to remove at least one blastomere from the embryo without interfering with its development (Handyside et al., 1989). The DNA can be amplified by PCR and typed, and a decision can be made whether or not to implant the embryo into the uterus.

A simple calculation shows that blastomere analysis can considerably reduce the probability that a child is born with a genetic disease. For example, consider a recessive disease not located on the X or Y chromosome, and let a denote the disease allele, so that an individual has the disease if and only if his genotype is aa. Another way of saying this is that the aa genotype is fully penetrant. Assume both parents have genotype Aa. In the absence of intervention, a child born to this couple has probability 25% of having the disease. If blastomere analysis is undertaken, an embryo will be implanted if and only if it is typed AA or Aa. Let G_T be the true genotype of the embryo, and let G_O be the genotype inferred as the result of typing the PCR amplified product from a single blastomere. Consider the case where $G_O = \text{Aa}$. The probability that the embryo is diseased is

$$P[G_T = \text{aa} \mid G_O = \text{Aa}].$$

We can calculate this probability in terms of the PCR efficiency r, contamination rate c and deposit probability d as follows. (In Section 2.3, r, c and d were referred to as α, β and γ, respectively.) Let p_0, p_1 and p_2 be the probability that exactly 0, 1 or 2 of the alleles in the cell are amplified to a detectable level. In terms of the parameters r and d we have

$$(1) \qquad p_0 = d(1-r)^2 + 1 - d, \quad p_1 = 2dr(1-r), \quad p_2 = dr^2.$$

A cell with true genotype aa will be typed Aa if a DNA fragment carrying the A allele enters the reaction tube as a contaminant, and if at least one of the two a alleles in the cell is detected. If we assume that both alleles are equally likely to be contaminants, we have

$$(2) \qquad P[G_O = \text{Aa} \mid G_T = \text{aa}] = \frac{(p_1 + p_2)c}{2}.$$

Analogous considerations yield

$$(3) \qquad P[G_O = \text{Aa} \mid G_T = \text{Aa}] = p_2 + \frac{p_1 c}{2}$$

and

$$(4) \qquad P[G_O = \text{Aa} \mid G_T = \text{AA}] = \frac{(p_1 + p_2)c}{2}.$$

Since $P[G_T = \text{AA}] = P[G_T = \text{aa}] = 0.25$ and $P[G_T = \text{Aa}] = 0.5$, the value of $P[G_T = \text{aa} \mid G_O = \text{Aa}]$ can be computed using Bayes' rule. For example, if $r = 0.9$, $d = 0.9$ and $c = 0.05$, $P[G_T = \text{aa} \mid G_O = \text{Aa}] \approx 0.015$. If an embryo typed AA is implanted, the error probability is only about 0.005.

An interesting issue arises if two or three blastomeres could be made available for typing. Is the probability of implanting a diseased embryo lower when all cells are amplified together in one tube or when each is amplified separately? To make the issue clearer, assume the disease is recessive as before, so the embryo is disease free if and only if it carries the A allele. In the one-tube procedure, all the blastomeres will be amplified together, and the embryo will be implanted if a positive signal for the A allele is given. In the separate-tubes procedure, each blastomere will be deposited in a separate tube for amplification. We assume that the blastomeres are independent with respect to being deposited in their respective tubes and that the tubes are independent with respect to contamination. The embryo will be implanted if each tube gives a signal for the A allele. The probability that an implanted embryo will be diseased can be calculated in a way analogous to the single blastomere case. See Navidi and Arnheim (1991) for details. It turns out that the separate-tubes procedure is superior. For example, if two blastomeres are available and $r = 0.9$, $d = 0.9$ and $c = 0.05$, the probability of implanting a diseased embryo with the one-tube procedure is 0.009; for the separate-tubes procedure, the probability is 0.002.

One other issue deserves consideration. Since in practice a limited number of oocytes are available for fertilization, a procedure which rejects embryos with high probability is not useful, even if the probability of error given acceptance is low. The separate-tubes procedure rejects somewhat more often than the one-tube procedure, but the acceptance rate is high enough to be useful. See Navidi and Arnheim (1991) for details.

3.2 The Multiple-Tubes Approach

Dividing a small DNA sample among several tubes proves to be beneficial in other situations as well. Many biological samples gathered as criminal evidence contain very small amounts of DNA—

potentially as little as a single fragment. It is impossible to determine precisely how much DNA is contained in such a sample. When the number of fragments is very small, mistyping can occur as a result of sampling error. For example, if 10 fragments are sampled from an individual heterozygous at some locus of interest, there is about a 10% chance that 8 or more of the fragments will contain the same allele. In this event, if the 10 fragments are placed into a single tube, amplified by PCR and then the PCR product typed, it is likely that the signal from the less common allele will be comparatively quite weak and attributed to background noise. Thus an incorrect finding of homozygosity will be made.

The current protocol for amplifying very small quantities in a single tube for forensics purposes includes limiting the number of PCR cycles so that a sample of unreliably small size will not be amplified to a detectable level. Such a procedure is quite conservative and results in failure to determine a genotype in cases when the sample is large enough so sampling error is not an issue.

Navidi, Arnheim and Waterman (1992) proposed the "multiple-tubes approach." This involves dividing the DNA among several tubes, then amplifying and typing the contents of each tube separately. Each tube thus provides an estimated genotype for the individual who provided the sample DNA. These estimates provide the basis for a series of hypothesis tests. For each possible genotype, we test the null hypothesis that it is the genotype of the sample. Two levels α and α' are chosen, where $\alpha \geq \alpha'$. To conclude that a particular genotype is the genotype of the sample, we must fail to reject it at level α, while we reject each other genotype at level α'. On the basis of simulation studies, Navidi, Arnheim and Waterman (1992) suggest taking $\alpha = 0.05$ and $\alpha' = 0.01$.

The number of hypothesis tests to be performed is determined by the number of distinct alleles detected in all the tubes combined. If only one allele is detected, we test two hypotheses: (1) that the sample is homozygous; (2) that the sample is heterozygous with one allele escaping detection. If $n \geq 2$ alleles are detected, then n homozygous and $n(n-1)/2$ heterozygous genotypes can be formed from these alleles. This results in a total of $n(n+1)/2$ genotypes in all to be tested. The hypothesis tests are performed under the assumption that tubes are contaminated independently with known contamination probability c. In practice a rough estimate of c is adequate, and an overestimate is preferable to an underestimate. See Navidi, Arnheim and Waterman (1992) for a more complete discussion of contamination in the multiple-tubes approach.

The test of the hypothesis that the DNA comes from a homozygous individual (with genotype AA, say) is based on the fact that under the null hypothesis every tube containing an allele other than A is contaminated. If N is the number of tubes used, then the number of contaminated tubes is binomially distributed with known parameters N and c. We reject if the number of such tubes is large.

For a heterozygous genotype (Aa, say) we combine two hypothesis tests. First, under H_0, every tube containing an allele other than A or a is contaminated, so we perform the test described above. Second, let n_A be the number of tubes in which allele A is detected but a is not, and let n_a be the number of tubes in which allele a is detected but A is not. It can be shown (Navidi, Arnheim and Waterman, 1992) that, under the assumptions that both alleles are equally likely to be contaminants and that PCR is equally efficient for both alleles, the null distribution of n_A given $n_A + n_a$ is binomial with parameters $n_A + n_a$ and $\frac{1}{2}$. We perform a two-sided test, rejecting if the value of n_A is far from $(n_A + n_a)/2$. The overall P-value for the hypothesis can be taken to be the smaller of the P-values of these two tests. The Bonferroni correction of doubling the smaller P-value does not appear to be necessary.

As an example, imagine a DNA sample has been apportioned among 10 tubes and that, after PCR amplification and typing, one tube is negative for all alleles, three are positive for allele A only, two are positive for allele a only, three are positive for alleles A and a, and one is positive for alleles A and a* (a third allele). The contamination rate c is conservatively estimated to be 10%. The hypotheses of homozygosity for alleles A, a and a* have P-values 1.47×10^{-4}, 9.12×10^{-6} and 9.10×10^{-9}, respectively. The hypotheses of heterozygous Aa, Aa* and aa* have P-values 0.651, 9.12×10^{-6} and 9.12×10^{-6}, respectively. Thus the conclusion that the sample came from an individual heterozygous Aa is clearly indicated. As another example, if six tubes were positive for allele A only and four positive for both A and a, the hypothesis of homozygous A, homozygous a and heterozygous Aa would have P-values 0.0128, 1.0×10^{-10} and 0.313, respectively. Under the criteria suggested by Navidi, Arnheim and Waterman (1992), this result would be declared inconclusive.

Based on simulation studies, Navidi, Arnheim and Waterman (1992) suggested that 15 tubes be used. To reduce on average the number of PCR reactions necessary, they suggested a simple sequential approach. Four tubes are amplified and typed. If a conclusive finding of genotype can be made, no further amplification is necessary. Otherwise six more tubes are amplified and typed, and the results of all 10 are analyzed. If a conclusive finding results, the process stops. Otherwise the last five tubes are amplified. When this approach is adopted, the value of α' should

be divided by 3 to account for the multiple tests. Simulation results indicate that with this approach conclusive results are usually attained when the number of fragments is 30 or more. Incorrect findings occur less than 3 times in 1,000 if the assumption of independently contaminated tubes is valid. Under a certain dependence structure, the error rate can be about 1.5% when the sample is very small. In contrast, if the entire sample is amplified in a single tube, the error rate can be as high as 19% for small samples.

3.3 Areas for Further Research

The values of α and α', the number of tubes and the sequential algorithm were determined by Navidi, Arnheim and Waterman (1992) primarily through trial and error in simulation studies and may not be optimal. In particular, a more sophisticated sequential algorithm might reduce the average number of PCR reactions per experiment. Given the large number of typing procedures performed in commercial laboratories, a savings of a fraction of a tube per experiment could result in considerable savings for a commercial laboratory.

4. WHOLE GENOME AMPLIFICATION—THE PEP PROCEDURE

Conventional PCR requires that one or more loci be targeted for amplification. The recently developed primer–extension–preamplification (PEP) method (Zhang et al., 1992) is intended to amplify a large fraction of the genome in a single reaction. The PEP procedure is simple in concept. Primer sequences, 15 bases in length, are manufactured by a process which generates sequences at random, producing a collection which contains on average about 10^6 copies of each of the 4^{15} such sequences (DNA has four bases). The existence of many different primer sequences in the reaction allows annealing and extension to occur at many different places on the genome. As a result, a large portion of the genome is copied at each cycle.

The PEP procedure produces far fewer copies of any given locus than would conventional PCR targeted to that locus. By dividing PEP products for individual sperm into aliquots and testing each aliquot for the presence of a specific locus using conventional PCR, Zhang et al. (1992) estimated that PEP produces an average of about 60 copies per locus. A similar experiment in which aliquots from PEP product from a single sperm were tested for the same sequence resulted in a conservative estimate that at least 78% of the genome had been amplified to a level of 30 or more copies.

The PEP procedure has the potential to allow multilocus typing of very small DNA samples, as might be collected in a forensic study, or in an analysis of ancient DNA, or in a single cell. For example, if k loci are to be typed, the DNA would be amplified by PEP, then the PEP product would be divided into k aliquots. Each aliquot would be amplified by PCR and the product would be typed at one locus. The accuracy obtainable as a function of k has not been determined. If k is too large, it is likely that for some heterozygous locus one allele would have no copy in its tube, which would lead to a mistaken homozygous typing. On the other hand if k is sufficiently small, extra accuracy might be obtained by dividing the PEP product into mk aliquots and performing a version of the multiple-tubes procedure (see Section 3.2) on m tubes for each locus.

Dear and Cook (1993) have studied the use of PEP for multilocus typing in their physical mapping procedure (see Section 2.3). They subjected 180 DNA samples to PEP, then divided each of the 180 PEP products into 20 aliquots. They amplified one aliquot from each set of 20 by conventional PCR, and typed it at nine loci. Their results indicated that typing the PEP products from these aliquots yielded about the same accuracy as typing the samples themselves by conventional PCR. In principle, the 19 remaining sets of aliquots could be amplified at 19 different sets of markers, and the linkage information on all 20 sets of markers could be included in one map.

5. MUTATION ERRORS: BRANCHING PROCESS MODELS FOR PCR

5.1 A Simple Model

Mutation errors, discussed in Section 1.2, occur when a base is incorrectly specified during the extension phase of a PCR cycle. We distinguish between single- and double-stranded mutants. A DNA fragment, consisting of two complementary strands, is a double-stranded mutant if either strand differs from the corresponding strand in the original ancestral DNA. Each strand which so differs is a single-stranded mutant. If we assume that mutations are never repaired by reverse errors during later cycles, it follows that the proportion of fragments in the PCR product which are identical to the target will approach zero as the number of cycles increases. For some methods of typing, success depends on the absolute number of correct fragments, not on the fraction of the product they represent, since incorrect fragments may be assumed to be invisible to the typing process. In principle, mutation errors might prevent detection of an allele. Since this possibility is already accounted for by a parameter for efficiency

error, there is no need to consider mutation errors explicitly.

Mutation errors must be taken into account if the PCR product is to be used in conjunction with other typing procedures, such as restriction fragment length polymorphism (RFLP) analysis. A restriction site is a specific DNA sequence, four to six base pairs in length. A restriction enzyme cuts a DNA fragment at every location where its corresponding restriction sequence is found, breaking it into fragments of different lengths. These fragments can be sorted by length using gel electrophoresis, and the lengths themselves can be approximately measured. Thus RFLP analysis provides a method for characterizing an individual's DNA by specifying the distances between successive occurrences of a given restriction site.

If mutation errors occur at a restriction site during PCR, the correctly replicated fragments in the PCR product will be cut, while the mutant fragments will not be. Thus certain bands on the gel will be formed only by correct fragments while others are formed only by mutants. If the proportion of mutants is small, the mutant bands will be noticeably fainter than the correct ones and will thus not be misleading. Otherwise, incorrect conclusions may result.

Krawczak, Reiss and Rösler (1989) studied the distribution of the proportion of fragments after n PCR cycles in which a given restriction site found in the target DNA is correctly replicated. Let S_0 be the initial number of single-stranded fragments, and assume for the sake of simplicity that every fragment replicates every cycle, so that the number of fragments after n cycles is $2^n S_0$. Let $X_{j,n}$ be the number of correct fragments produced by the jth fragment during the nth cycle. Then $X_{j,n} = 2$ if the restriction site is correctly replicated, and $X_{j,n} = 1$ if a mutation occurs within the restriction site. Let S_n be the number of correct single-stranded fragments after n cycles, so $S_n = \sum_j^{S_{n-1}} X_{j,n}$. Let $p = P[X_{j,n} = 1]$. It is reasonable to take $p = m\rho$, where m is the number of base pairs in the restriction site and ρ is the misincorporation rate per base per cycle. Gelfand and White (1990) have estimated ρ to be between 10^{-5} and 10^{-4}.

The expectation and variance of S_n are given by Krawczak, Reiss and Rösler (1989) as

$$E(S_n) = S_0(2-p)^n, \tag{5}$$

$$\mathrm{Var}(S_n) = S_0 p(2-p)^{n-1}\left[(2-p)^n - 1\right]. \tag{6}$$

The number D_n of correct double-stranded fragments after the nth cycle can be studied by noting that D_n is equal to the number among the S_{n-1} correct single strands which correctly replicate. Thus

$$D_n = \sum_{j=1}^{S_{n-1}} (X_{j,n} - 1) = S_n - S_{n-1}. \tag{7}$$

It follows that

$$E(D_n) = (1-p)E(S_{n-1}), \tag{8}$$

$$\mathrm{Var}(D_n) = (1-p)^2\mathrm{Var}(S_{n-1}) + p(1-p)E(S_{n-1}). \tag{9}$$

Krawczak, Reiss and Rösler (1989) use Chebyshev's inequality to place an upper bound on the probability that the proportion of correct fragments falls below a given proportion of the total. They take 80% as the proportion of correct fragments needed for accurate results. For a six-base-pair restriction ($m = 6$), $S_0 = 100{,}000$ and $n = 40$, the probability that less than 80% of the double-stranded fragments will be correct is bounded above by 4.15×10^{-8}. Varying the number of cycles n does not change this value appreciably. For $m = 6$, $n = 40$ and $S_0 = 2$, the Chebyshev bound is about 0.0042, indicating that under this model a given restriction site will almost always be preserved by PCR amplification.

5.2 Areas for Further Research

The Chebyshev bounds for single-stranded fragments when S_0 is small are more crude than is desired. Further investigation of the probability distribution of the number of correct fragments would be useful. The model described here is simplistic in that it assumes perfect replication, while experimental evidence suggests that, during 50 cycles of PCR starting with a single cell, only about 65% of fragments replicate per cycle on average, although it is suspected that this percentage is much higher, perhaps approaching 90%, in later cycles. A two-type branching process might be considered, where on each cycle each mutant fragment produces zero nonmutants and either one or two mutants, and each nonmutant fragment produces either one nonmutant, two nonmutants or one nonmutant and one mutant.

ACKNOWLEDGMENTS

We would like to thank Terry Speed and Simon Tavaré for helpful discussions. This research was partially supported by NSF Grant DMS-90-05833.

REFERENCES

BOEHNKE, M., ARNHEIM, N., LI, H. and COLLINS, F. S. (1989). Fine-structure genetic mapping of human chromosomes using the polymerase chain reaction on single sperm: experimental design considerations. *American Journal of Human Genetics* **45** 21–32.

CUI, X., LI, H., GORADIA, T., LANGE, K., KAZAZIAN, H. H., JR., GALAS, D. and ARNHEIM, N. (1989). Single-sperm typing: determination of genetic distance between the $^{G}\gamma$-globin and parathyroid hormone loci by using the polymerase chain reaction and allele-specific oligomers. *Proc. Nat. Acad. Sci. U.S.A.* **86** 9389–9393.

DEAR, P. H. and COOK, P. R. (1993). Happy mapping: linkage mapping using a physical analogue of meiosis. *Nucleic Acids Research* **21** 13–20.

EVANS, S. N. MCPEEK, M. S. and SPEED, T. P. (1993). A characterisation of crossover models that possess map functions. *Theoret. Population Biol.* **43** 80–90.

GELFAND, D. H. and WHITE, T. J. (1990). Thermostable DNA polymerases. In *PCR Protocols* (M. A. Innis, D. H. Gelfand, J. J. Sninsky and T. J. White, eds.) 129–144. Academic, New York.

GORADIA, T. M. and LANGE, K. (1990). Multilocus ordering strategies based on sperm typing. *Annals of Human Genetics* **54** 49–77.

GORADIA, T. M., STANTON, V. P., JR., CUI, X., ABURATANI, H., LI, H., LANGE, K., HOUSMAN, D. E. and ARNHEIM, N. (1991). Ordering three DNA polymorphisms on chromosome 3 by sperm typing. *Genomics* **10** 748–755.

GUERRA, R., MCPEEK, M. S., SPEED, T. P. and STEWART, P. M. (1992). A Bayesian analysis of radiation hybrid data. In *Genetic Analysis Workshop 7* (MacCluer et al., eds.). *Cytogenet. Cell Genet.* **59** 104–106.

HANDYSIDE, A. H., PATTINSON, J. K., PENKETH, R. J., DELHANTY, J. D., WINSTON, R. M. and TUDDENHAM, E. G. (1989). Biopsy of human preimplantation embryos and sexing by DNA amplification. *The Lancet* **i** (8634) 347–349.

HUBERT, R., STANTON, V. P., JR., ABURATANI, H., WARREN, J., LI, H., HOUSMAN, D. E. and ARNHEIM, N. (1992). Sperm typing allows accurate measurement of the recombination fraction between D3S2 and D3S3 on the short arm of human chromosome 3. *Genomics* **12** 683–687.

KARLIN, S. and LIBERMAN, U. (1979). A natural class of multilocus recombination processes and related measures of crossover interference. *Adv. in Appl. Probab.* **11** 479–501.

KOGAN, S. C., DOHERTY, M. and GITSCHIER, J. (1987). An improved method for prenatal diagnosis of genetic diseases by analysis of amplified DNA sequences. *New England Journal of Medicine* **317** 985–990.

KRAWCZAK, M., REISS, J. and RÖSLER, U. (1989). Polymerase chain reaction: replication errors and reliability of gene diagnosis. *Nucleic Acids Research* **17** 2197–2201.

LANDEGREN, U., KAISER, R., SANDERS, J. and HOOD, L. (1988). A ligase-mediated gene detection technique. *Science* **241** 1077–1080.

LANGE, K., BOEHNKE, M. and WEEKS, D. E. (1988). Programs for pedigree analysis: MENDEL, FISHER and dGENE. *Genetic Epidemiology* **5** 471–472.

LAZZERONI, L., ARNHEIM, N., SCHMITT, K. and LANGE, K. (1993). Multipoint mapping calculations for sperm-typing. Unpublished manuscript.

LEWIN, H. A., SCHMITT, K., HUBERT, R., VAN EIJK, M. J. T. and ARNHEIM, N. (1992). Close linkage between bovine prolactin and BoLA-DRB3 genes: genetic mapping in cattle by single sperm typing. *Genomics* **13** 44–48.

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

MULLIS, K. B. and FALOONA, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase catalyzed chain reaction. *Methods in Enzymology* **155** 335–351.

NAVIDI, W. and ARNHEIM, N. (1991). Using PCR in preimplantation genetic disease diagnosis. *Human Reproduction* **6** 836–849.

NAVIDI, W., ARNHEIM, N. and WATERMAN, M. (1992). A multiple-tubes approach for accurate genotyping of very small DNA samples by using PCR: statistical considerations. *American Journal of Human Genetics* **50** 347–359.

OTT, J. (1991). *Analysis of Human Genetic Linkage*, rev. ed. Johns Hopkins Univ. Press.

SAIKI, R., BUGAWAN, T. L., HORN, G. T., MULLIS, K. B. and ERLICH, H. A. (1986). Analysis of enzymatically amplified β-globin and HLA-DQα DNA with allele-specific oligonucleotide probes. *Nature* **324** 163–166.

SAIKI, R., GELFAND, D. H., STOFFEL, S., SCHARF, S. J., HIGUCHI, R., HORN, G. T., MULLIS, K. B. and ERLICH, H. A. (1988). Primer directed amplification of DNA with a thermostable DNA polymerase. *Science* **239** 487–491.

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

STEPHENS, D. A. and SMITH, A. F. M. (1992). Bayesian inference in multipoint gene-mapping. Unpublished manuscript.

UGOZZOLI, L. and WALLACE, R. B. (1991). Allele-specific polymerase chain reaction. In *Methods—A Companion to "Methods in Enzymology* **2** (N. Arnheim, ed.)" (J. N. Ableson and M. I. Simon, eds.) 42–48. Academic, New York.

WINSTON, R. M. L. and HANDYSIDE, A. H. (1993). New challenges in human in vitro fertilization. *Science* **260** 932–936.

ZHANG, L., CUI, X., SCHMITT, K., HUBERT, R., NAVIDI, W. and ARNHEIM, N. (1992). Whole genome amplification from a single cell: implications for genetic analysis. *Proc. Nat. Acad. Sci. U.S.A.* **89** 5847–5851.