STATISTICAL METHODS
FOR
DNA FINGERPRINTING

by

VARSHA CHHATRE
University of Victoria, January 1995.

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

We accept this thesis as conforming
to the required standard

[redacted]

Dr. Mary Lesperance, Supervisor (Department of Mathematics & Statistics)

[redacted]

Dr. Roger Davidson, Department Member (Department of Mathematics & Statistics)

[redacted]

Dr. William Reed, Department Member (Department of Mathematics & Statistics)

[redacted]

Dr. Kenneth Stewart, External Examiner (Department of Economics)

Supervisor: Dr. Mary Lesperance (Department of Mathematics and Statistics)

Abstract

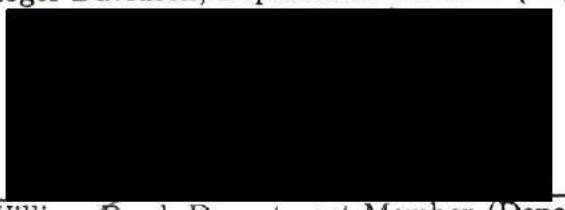
DNA fingerprinting plays an important role in forensic sciences. Hypervariable regions of the human genome provide valuable evidence which can be used for identifying an individual based on a DNA sample. Though each individual can't be uniquely determined from a DNA fingerprint, it is believed that the probability that the DNA patterns of two randomly chosen individuals match is highly unlikely. If two samples match, it can be concluded that they may have the same genotype by chance. It is important to know how likely it is that the two individual genotypes match by chance. This depends upon how frequent the observed genotype is in the general population. If the genotype is not commonly observed then it can be concluded that the samples came from the same individual. To reach any conclusion, it is necessary to get estimates of the distribution of allele sizes for the general population or a reference population.

Frequency estimates are obtained for five Variable Number of Tandem Repeats (VNTR) loci of Orange County's Sheriff-Coroner Department's(OCSD) reference population using Fixed Bin, the mixture model of Devlin et al. (1991) and a modification of their mixture model. Estimates obtained by our modification are compared with the estimates obtained by Devlin et al.(1991) for Life Codes (LC) VNTR loci D17S79. The fit of the suggested model is tested and variances of the estimates are obtained. There are some controversies about the statistical assumptions, computations and interpretation of the results. The thesis includes a discussion on the use of the estimated allele distribution in forensic cases and on the controversy surrounding its use.

Examiners:

Dr. Mary Lesperance, Supervisor (Department of Mathematics & Statistics)

Dr. Roger Davidson, Department Member (Department of Mathematics & Statistics)



Dr. William Reed, Department Member (Department of Mathematics & Statistics)

Dr. Kenneth Stewart, External Examiner (Department of Economics)

Contents

List of Tables

List of Figures

Acknowledgement

I am deeply indebted to Dr. Mary Lesperance for her guidance, great interest in my research work, and encouragement. This thesis would not have been completed without her help and incredible patience.

I would like to thank Dr. Roger Davidson, Dr. William Reed and Dr. Kenneth Stewart for their time to examine my thesis and provide me with suggestions and comments.

Thanks are due to Dr. Kathryn Roeder, Dr. Bernie Devlin and John Hartmann for providing me the necessary data, and contributing valuable information.

Deep appreciation is expressed to my friends and colleagues for their help and advice.

Finally I owe a special debt of thanks to my brother and his family for their love, support and encouragement. I would like to extend heartful thanks to other family members for their emotional support and prayers.

Varsha Chhatre

General Structure of the Report

Chapter 1 provides information on the basic terminology used in DNA analysis and the way DNA profiling is done. It also describes the problem of allele estimation, and the data used for the estimation. Chapter 2 provides the models used for solving the estimation problem. A modification of the Devlin et al. (1991) mixture model is derived and recommended for use with small samples. Method for testing the fit of the model and a method for obtaining the variances of the estimates are described. Chapter 3 provides the results of applying the methods described in Chapter 2 to the available data. Chapter 4 describes how inferences in criminal cases can be drawn with the estimated allele distribution and other available information, and it also discusses the interpretation of the results and the controversies over statistical issues about assumptions.

Chapter 1

Background and Data

DNA fingerprinting is considered to be a powerful technique in forensic as well as paternity cases. The culpability of a suspect is determined by the similarities between the DNA or genetic fingerprint of the suspect and the DNA sample found at the scene of the crime. If two samples are sufficiently similar, a match of the profile is declared and is presented with an estimate of the probability of obtaining a matching profile from a randomly selected individual from some appropriately selected reference population. Hence, it is important to determine how likely it is to observe a particular DNA fragment in a general population. The forensic labs have their own reference populations which consist of measurements on different locations of the DNA string for many individuals from different ethnic groups.

John Hartmann of the Orange County's Sheriff-Coroner Department (OCSD) approached Mary Lesperance with their entire 1992 database. The data consists of

measurements from 5 different locations(loci) of the human chromosomes, for each of approximately 1300 individuals from 7 ethnic groups. OCSD's main interest is to obtain the estimates of the frequency distribution of measurements for each locus and for each ethnic group separately. OCSD suggested that one method that is used to estimate the frequency distributions is given in Devlin et al. (1991) .

We review some of the basic terminology used in genetics and in the subject of DNA fingerprinting in Chapter 1. The background section includes discussion of the elements of genetics, DNA fingerprinting, a procedure for obtaining DNA fingerprints, determination of DNA fragment sizes and the sources of measurement error. The problem of allele estimation is expressed in mathematical language in the same section. The second section gives a description of the data used for estimation. It provides complete information on the loci used, files in which the data is stored and format of the files.

1.1 Background

1.1.1 Elements of DNA Fingerprinting

DNA fingerprinting is a powerful technique used in forensic sciences to differentiate individuals. The basic questions are what is DNA; what is DNA fingerprinting; how can it possess such an enormous quantity of information required for differentiation of an organism? This subsection on DNA fingerprinting helps to answer these questions.

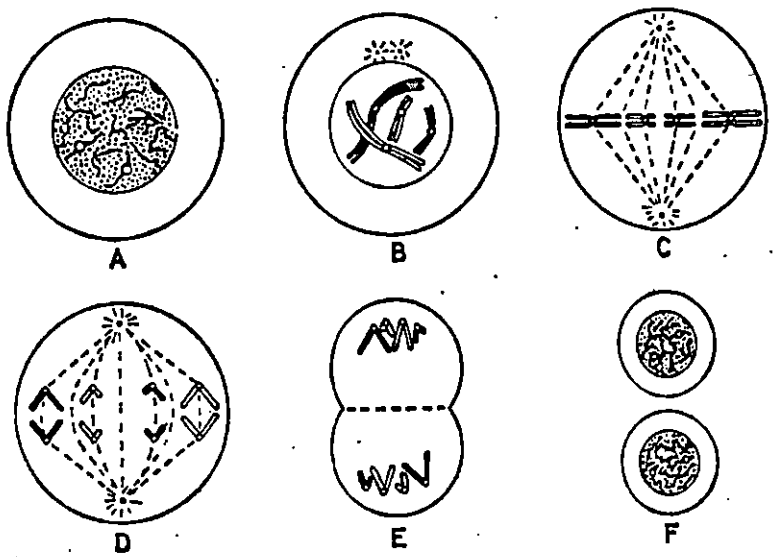


Figure 1.1: Various stages of cell division (mitosis): A=interphase; B= prophase; C=metaphase; D= anaphase; E=telophase; F= Daughter Cells.

Each cell in the body contains minute threadlike structures called *chromosomes.* In most cells, they are usually hidden unless the cell is in the process of division, as shown in Figure 1.1, from Koller (1967). Except for the reproductive cells(gametes), each cell in the human being contains 46 different chromosomes, 44 of which are paired, and the other 2 are sex chromosomes. One member from each pair is provided by the father, and the other from the mother. Each pair of chromosomes controls a set of characteristics displayed in the individual. Geneticists have done a considerable amount of work mapping out the functions of specific locations or 'loci' on the chromosome pairs.

The hereditary material inside a chromosome is called *Deoxyribonucleic Acid* or DNA. DNA molecules are long chains of sub-units called *nucleotides*, repeated some

number of times. Nucleotides are blocks composed of three parts: base, sugar (S) and phosphoric acid (P). Four bases are adenine(A), guanine(G), thymine(T), or cytosine(C). The combination of one base, one sugar and one phosphate molecule forms one nucleotide unit. Thousands or millions of nucleotides are joined together to form the polynucleotide chain. See Figure 1.2, from Koller (1967).

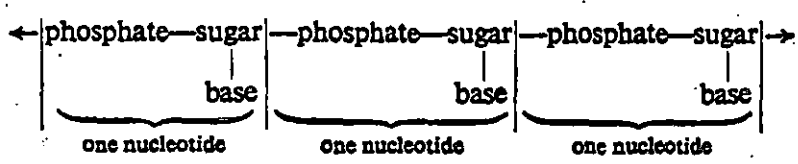


Figure 1.2: Structure of polynucleotide chain

The DNA molecule is composed of two polynucleotide chains in which bases are linked with hydrogen bonds. The molecular structure of the two polynucleotide chains in which the bases are linked together is shown in Figure 1.3, Koller (1967). Due to the different shape of the bases and the angle at which they are joined and linked with a sugar molecule, the two polynucleotide chains are twisted around each other. DNA is formed of a double helix like a spiral staircase with the base-pair bonds as the steps. We can see in Figure 1.4, from Koller (1967) that two sugar-phosphate-sugar (S-P-S) groups are held together by a hydrogen bond between the bases of two chains and this forms a double helix. Adenine is always paired with thymine {A,T}, and guanine with cytosine {G,C}. The sets {A,T}, {G,C} are called *base-pairs* or *bp.*

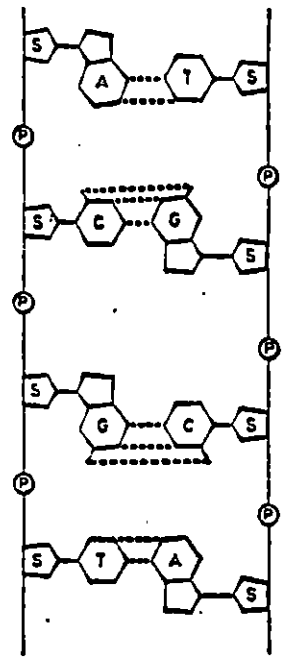


Figure 1.3: Molecular structure of two polynucleotide chains

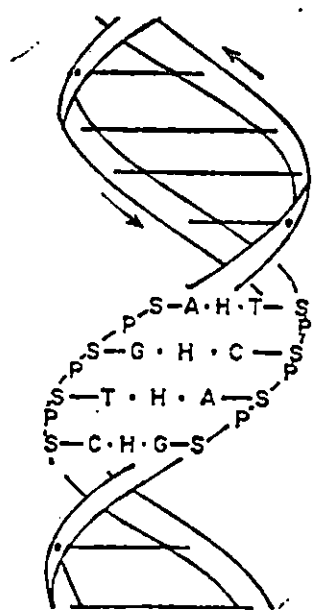


Figure 1.4: Model of the DNA molecule

DNA fingerprinting:

The theory of DNA fingerprinting, is very recent, dating from 1985. The DNA structure of human beings is very similar from individual to individual, however there are some loci on the human chromosome where the DNA structure shows a substantial variation among individuals. At these locations *core sequences* which are short sequences of base-pairs, are repeated in tandem a number of times. The repeated sequence of base-pairs is called a *Repeat.* Due to large variation in the *number* of tandem repeats, the length of the DNA molecule at these locations varies greatly between individuals. Hence Nakamura et al. (1987) introduced the term *Variable Number of Tandem Repeats* (VNTR) loci for these specific locations. The DNA pattern from a VNTR locus or set of VNTR loci is known as *DNA Profile* or *DNA Fingerprint.*

DNA identification depends on the information contained in the variable regions. The more variation between individuals present in the variable region, the more information they provide. It is easy to identify the sequences by laboratory techniques when substantial variation is observed among individuals.

For DNA identification it is necessary to cut out the variable regions from the DNA molecule. Restriction endonucleases or molecular scissors, are the enzymes which recognize the particular sequences and are used to separate the variable regions from the rest of the DNA molecule. These restriction enzymes are developed from various bacteria. Hundreds of endonucleases have been isolated from more than 200

different bacterial species. They are named by the bacterial species from which they are derived. The name consists of the first letters of the microorganism source and a roman number indicating the series number of enzymes derived from that organism. One of the restriction enzymes which is used quite often is *Hae III*. The microorganism source of *Hae III* is *Haemophilis aegyptius*.

As the number of repeated sequences as well as the number of base-pairs varies, fragments of various sizes are produced when the restriction enzyme chops the DNA. Alleles are defined as the DNA fragments of different lengths at any locus. Since chromosomes occur in pairs, alleles are also observed in pairs at each locus. The genotype at a particular locus is determined by the *true* lengths of the pair of alleles that an individual possesses at that locus. The phenotype is the *observed* lengths of the alleles at that locus. Individuals are called *heterozygous*, if they have distinct alleles at the locus; and *homozygous*, if they have same pair of alleles at the locus. Previous studies suggest the possibility of observing hundreds of different alleles at each VNTR locus while for loci other than VNTR the observed number of alleles are one to only a few.

Restriction enzymes cut the DNA molecule at specific sites such that each fragment is divided into two regions, namely the variable region and the flanking region. The important one is variable region, which consists of the base-pair sequence, repeated a number of times and linked in tandem. Each of these variable regions is flanked on both sides by base-pair sequences, so-called flanking-regions. These base-

depends upon where the restriction enzyme cuts the fragment. Hence, variation in the number of repeats, variation in the size of flanking-region or variation in both, can cause the difference in the actual size of two restriction fragments,(see Figure 1.5). This variation in length is a polymorphism and hence termed as *Restriction Fragment Length Polymorphism* (RFLP). The highly polymorphic nature of a VNTR loci helps to distinguish a fingerprint of any person from the rest of the population.

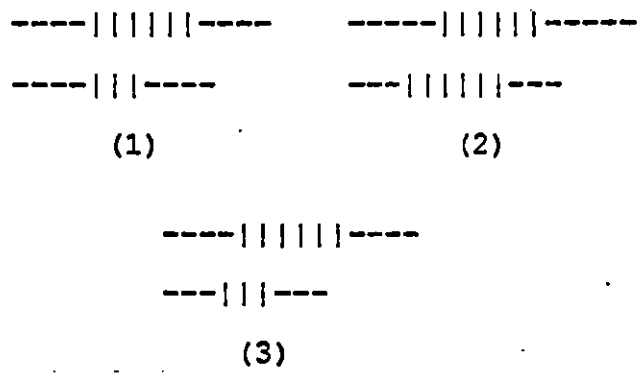


Figure 1.5: Sources of variation in alleles: - represents flanking region and | represents a repeat. (1) Variation in number of repeats (2) Variation in flanking region size (3) Variation in both

Due to substantial variation in the number of repeats, a large number of alleles is observed at any one VNTR locus, hence the probability that two different individuals have the same allele at a set of VNTR loci is very low. Therefore DNA sequences on VNTR loci are considered to be unique to individuals and it is thought that no two people can have the same fingerprint. This has made DNA fingerprinting an important tool in forensic sciences and paternity cases. Now that we know what DNA

fingerprinting is , the next task is to find out how to obtain the fingerprint?

1.1.2 DNA Identification

DNA profiling includes cleavage by restriction enzymes, electrophoresis, southern transfer, probe labeling, hybridization and print detection. The following description and the picture of the process, Figure 1.6. from Kirby (1990), describes how to get a DNA fingerprint.

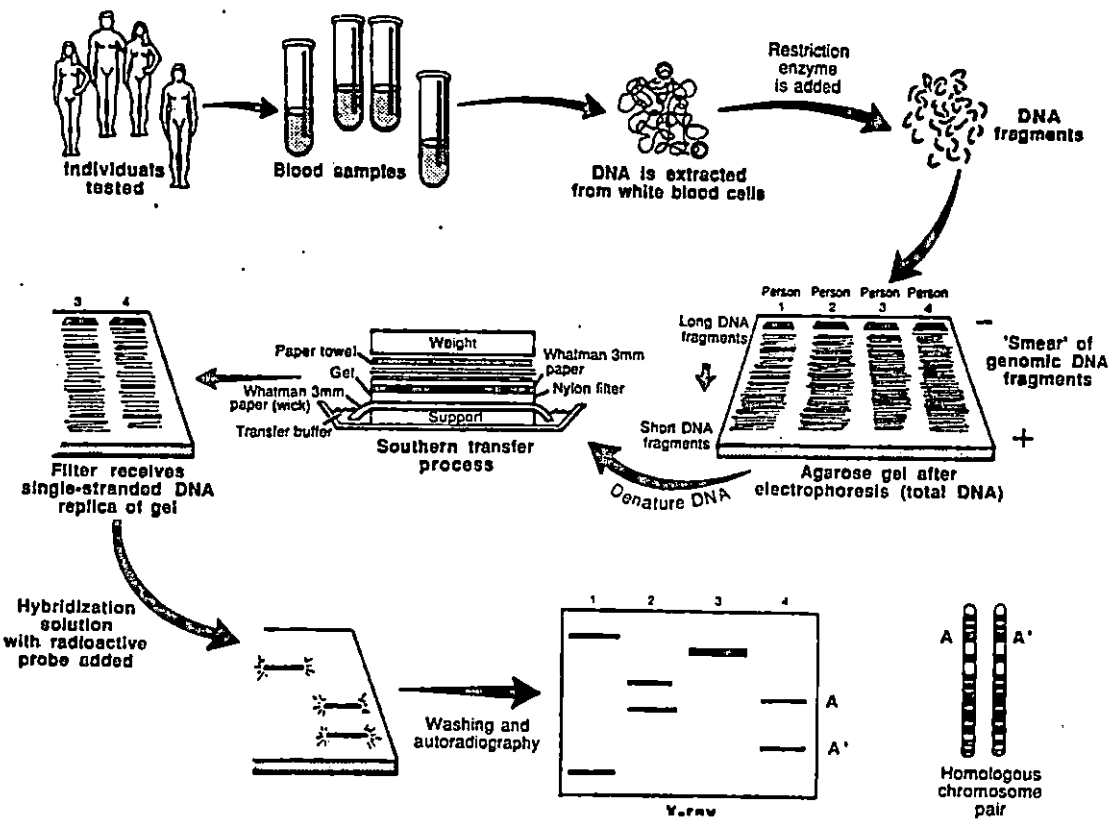


Figure 1.6: Single-locus multiallele DNA fingerprinting procedure

Following are the steps involved in obtaining the DNA fingerprint.

1. First the blood samples are collected from individuals. DNA is extracted from white blood cells and restriction enzymes are added to it. Restriction enzymes act as molecular scissors which chop the DNA at variable regions. If a restriction enzyme cuts in between the repeated sequence, it will result in the production of small fragments, which may be unresolvable and of no use for identification. Hence the choice of restriction enzymes is important.

2. Electrophoresis is the second step. Electrophoresis is the separation of large molecules in an electric field, hence it is desired to maintain a constant electric field for separation of the DNA. The electrophoresis setup is given in Figure 1.7. To maintain the balance of ions and to keep the current going, a chemical which is known as buffer is added to the agarose mixture as well as later on to the set gel. Agar is extracted from certain seaweeds and agarose is the neutral gelling fraction of agar commonly used in gels. The agarose mixture of temperature not greater than 50^0C is poured in a casting tray giving the gel thickness of 3-5 mm. The comb is then inserted in this mixture to create the wells for analyzing the individual DNA samples. When the gel has set, the surface is flooded with the same buffer used for agarose mixture and the comb is removed. The casting tray is then submerged in the electrophoresis tank. Size markers, controls and the individual DNA samples are added into separate wells. The tank cover is set in place, and the electrical controls are attached. The sample migrates from negative to positive. The electric field is set up such that, the rate at

which the fragments move is a function of their lengths.

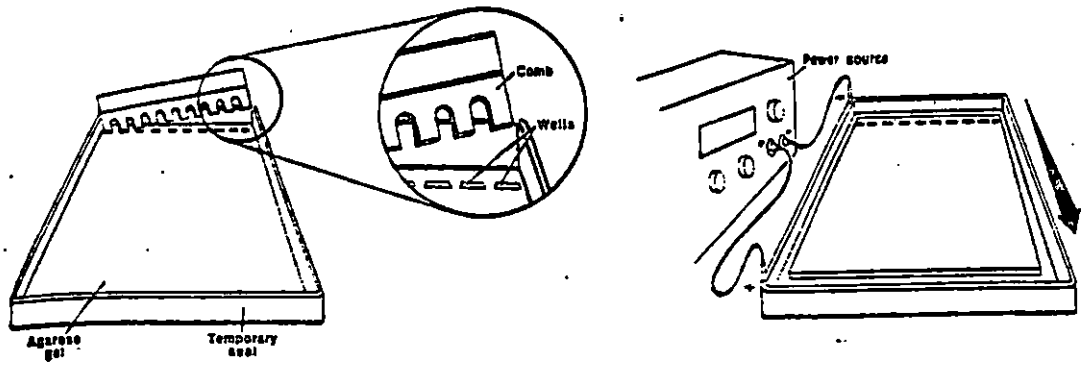


Figure 1.7: Gel electrophoresis setup

3. The third step is the transfer of DNA fragments from the gel to a nylon membrane and separation of the double helix into single strands. To carry out this procedure after electrophoresis, the gel is flooded with denaturing solution for at least 30 min. Denaturing solution is the solution which helps to break the hydrogen bonds between the complementary DNA strands. The denaturing solution is then replaced with neutralizer for 30 min or longer. Then the gel with DNA side down is placed on prewetted filter paper and by pressing, all air bubbles are removed. The nylon membrane is placed on the gel ensuring that all air bubbles are removed. Then the nylon membrane is covered with sheets of wetted filter paper. Absorbent papers such as paper towels and weights are added. Once the the process of transferring the fragments is completed in 5-6 hrs, the membrane is removed and rinsed to remove

agarose residuals, air dried and wrapped in saran wrap.

4. As the core sequences are important in DNA identification, it is necessary to identify them from remaining irrelevant fragments. There are special DNA probes, single stranded segments of DNA which can recognize the specific sequences and attach themselves to these sequences and help to locate them. If the core sequence is observed at only one locus, the probe is called a single locus probe. Normally forensic DNA profiling labs use several single-locus probes, each of which binds to a different site. After transferring the DNA fragments, the nylon membrane is placed in a medium containing radioactive probes.

5. After washing, the membrane is exposed to an X-ray film. The radioactive sites of these fragments appear as dark bands where the probes are attached to the DNA fragments on a X-ray film. This X-ray film is called autoradiogram. Comparing the band patterns, an individual can be identified. From the band patterns on the X-ray films, the lengths of the DNA fragments are determined with the help of size markers used.

Once the band pattern is obtained the next step is to determine the band sizes. The following section describes how the band sizes are estimated.

1.1.3 Fragment Size determination

The fragment length is measured in base-pairs(bp) or Kilo base-pairs (Kb), where 1 Kb = 1000 bp. A base-pair is pair of bases {A,T} or {C,G} linked by a hydrogen

bond. The mobility of the DNA fragment is the distance traveled by the fragment in the agarose gel. The fragment length is determined from the mobility of the fragment. Since mobility will always vary from membrane to membrane, the molecular length of the DNA fragments can be determined only by comparing the position of the bands to the position of standard markers on the same membrane. Standard markers or Size markers are the bands of known molecular lengths. Size markers used by OCSD and LC are given in Figure 1.10. On each membrane, size markers are added every 4 to 5 lanes (Kirby, 1990). To determine the unknown lengths from their mobilities, a relationship is established between the mobilities and lengths of standard fragments. Measurement error associated with the estimated lengths depends upon two factors, the measured mobility and the relationship used for determination. Factors which can cause error in measurement of mobilities are discussed in the next section. In this section we discuss the methods of measuring mobility and the relationships used for estimation.

All laboratories have their own method of measuring mobility. The simple one is measuring by hand using a metric ruler. The ruler is laid on the gel or gel photograph and the migration distance of the bands from the application slots are measured to the nearest millimeter.

Digitizing is another method which is often used. Galbriath, et al. (1991) discuss how to size the band using digitizing. In digitizing, first the autoradiograms are taped to an acetate film. A line is drawn through the most intense region of each band. If

the edges of the blots are visible they are also marked using fine-pointed markers. The autoradiogram of the standard markers is placed under this acetate overlay and their positions are marked in each lane on the overlay as shown in Figure 1.8. A digitizer pen is touched on this acetate overlay at each of the positions of the standard sizes and their mobility is determined in each lane. A log-linear model, log(length) = $c_0 + c_1$ (mobility), is fitted for the standard markers, for which the lengths are known. This fitted model is used to estimate the molecular lengths of the DNA fragments given mobility.

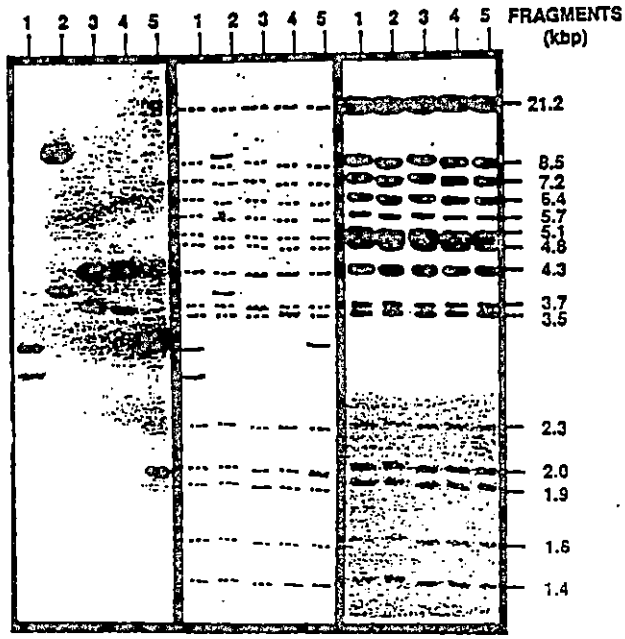


Figure 1.8: Acetate overlay

Holmlund et al. (1992) describe an image processing technique for measuring mobility where a video camera is attached to a microcomputer. The video camera scans the autoradiogram and an image of the total autoradiogram is produced on

the video monitor. The positions of the bands for marker lanes as well as for other sample lanes are determined from the monitor. The measurement is done in pixels and then converted to millimeters. For the image processing system used in Holmlund et al. (1992), four pixels represented one vertical mm and eleven pixels were scanned along the width of a lane. From the mobilities the fragment sizes are determined automatically by the computer and the results are displayed.

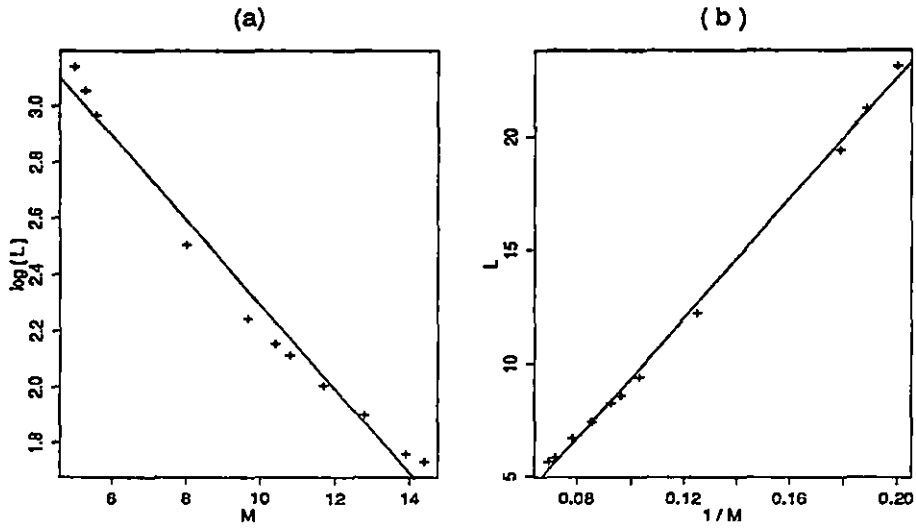


Figure 1.9: The plots of band mobility versus length for 11 standard fragments. (a) Plot of log(length) vs mobility. (b) Plot of length vs 1 / mobility. The standard lengths are 23.130, 21.226, 19.397, 12.220, 9.416, 8.614, 8.271, 7.421, 6.682, 5.804, 5.643. Mobilities observed are 5.0, 5.3, 5.6, 8.0, 9.7, 10.4, 10.8, 11.7, 12.8, 13.9, 14.4.

Elder and Southern (1983) suggest different relationships between length (L) and mobility (M). Fitting or choosing the wrong relationship would result in measurement error. Two commonly used models are the log model that is $log(L) = c_0 + c_1 M$ (Fisher and Dingman, 1971) and the inverse model $L = k/M$ (Southern 1979). Elder and Southern (1983) suggest that the inverse relationship gives the most accurate

results. It is possible that for high molecular weights the log-linear relationship shows departure from linearity.

In Figure 1.9, (a) shows departure from linearity for low mobility and large molecular weight while (b) gives a linear fit. Hence model $L = k/M$ can be used for estimation of unknown lengths.

There are several factors which can cause the variation in the mobilities of the fragments. Error in mobility gives error in the measurement of length. The following section includes a brief discussion of the problem of measurement error.

1.1.4 Sources of Measurement Error

To get reliable results, it is required to have accurate measurements of the fragment lengths. But it is observed that the fragments are measured with error. Besides the measurement errors caused by technicians, some other factors are also responsible.

1. **Mobility:** Factors responsible for variation in mobility are:

Gel concentration and Voltage: Factors like concentration of the gel and voltage in the electrophoresis can cause the mobility to differ from a predicted simple relationship. A gel concentration of 1% is usually used for human DNA. But care should be taken while reducing the concentration as the gel becomes fragile and there is a possibility of breaking it during handling. In case of high voltage and high agarose concentration, mobility which is a function of molecular length becomes nonlinear for larger fragments and hence estimation of fragments lengths becomes difficult. To

solve such problems, the relationships are modified using a correction factor. For example Southern (1979) observed curvature in the plot of the inverse model for high voltage gradients, hence he proposed a correcting factor (M_0) to give the best fit of the line $L = a_1 + a_2/(M - M_0)$

Band size and band width: When the individual DNA samples are placed in a slot at one end of the gel and forced through the gel using an electric field, fragments find their path of migration. These fragments appear in the gel as bands of varying width in a lane. Depending on the concentration of agarose gel, the larger fragments have difficulty in moving in the gel, and hence the resultant bands are narrow. Smaller fragments move faster than the larger fragments resulting in wider bands. In addition to this, if a set of fragments of similar size migrate, then each molecule follows a path of least resistance, producing wider bands.

Radioactive probes are used to locate the DNA fragments. The width of the bands increases with the exposure time. Long exposure can produce very wide bands.

Sometimes, for the larger fragments the bands are narrower than the pixel width of the camera, while for smaller fragments the bands are much wider than the pixel widths.

Wider bands can cause an increase in precision. The migration distance of the fragment is measured from the baseline to the position of the band or half of the width of the band. A shift of a band by one pixel changes the fragment size by some number of base-pairs and this difference is different for different ranges of fragment

size because of the nonlinear scale, (Holmlund, 1992).

Base sequence: In practice it is observed that the mobility not only depends upon the length but also depends upon the sequence of bases. According to information provided by John Hartmann, there is a difference in mobility depending upon the sequences of bases comprising a fragment.

Amount of DNA in a lane: Loading a larger amount of DNA in a lane yields more intense and wider bands. Also this may produce curved autoradiogram fragment bands, which introduces measurement error in mobility. If an old sample is used then the bands are of poor quality.

2. **Coalescence:** This is one of the serious sources of measurement error which should be taken into consideration in modeling allele frequency distributions. If the fragments are of same length, the two bands blur together, and instead of two, only one band is observed. This is termed coalescence. It is possible that both the alleles are the same, that is homozygotes, or they may be close heterozygotes termed pseudohomozygotes. Sometimes due to smaller length, one fragment is diffused and not identified. Larger fragments tend to coalesce more than the smaller fragments because they travel shorter distances.

3. **Gel effects:** Running the same sample in different lanes gives different results, which is called the within-gel effect. Samples near the sides of the gel also result in less precision. Running the same sample in two different gels gives more error than running in same gel. This is called between-gel effect. Material used for preparation

of the gel also makes a great deal of difference.

From the given size markers shown in Figure 1.10, it is observed that the difference between any two standard fragments is not consistent and the distribution of fragment length is nonlinear. Also note that the observed bands are of varying width. Thus the measurement error is not constant; it is a function of fragment length. For modeling the allele frequency distributions, it will be necessary to accommodate the probability of coalescence and the measurement error.

Next we discuss the objective of this project and present a mathematical model for allele lengths.

1.1.5 Problem

In forensic science, if the striped patterns from two samples appear to be the same, then a match is declared. To declare a match, it is important to find the probability of matching the DNA fingerprint of two randomly chosen individuals. This leads to the problem of obtaining allele lengths and then estimating the allele length frequency distribution.

If the lengths of the alleles are measured exactly, then they can be classified according to the number of repeats and estimation would be a simple multinomial problem. The allele length can be modeled as follows. If x is the observed length of a restriction fragment with r repeated sequences, each of length ρ and the total flanking region is u bp. Then given that there is no measurement error, the actual

length of the allele is discrete and it will be $u + r\rho$. Even though the number of repeats is not observed directly, all alleles can be classified according to the number of repeats. But the fact is that the alleles are measured with error, which is generally associated with the allele length. This measurement error is modeled continuously, hence the distribution of the observed fragments is no longer discrete. The resulting distribution is continuous and we estimate it using methods for mixture distributions.

Now that the problem is set up we can have look at the data available. The following section on Data gives the description of the databases to be analyzed.

1.2 Data

We analyze two sets of data here. One is from Lifecodes corporation(LC) laboratory and the other is from Forensic Science Service Orange County's Sheriff-Coroner Dept. (OCSD). The Lifecodes data consists of two VNTR loci, D2S44 and D17S79. The OCSD data consists of 5 VNTR loci, namely D2S44, D17S79, D1S7, D10S28 and D4S139.

Both the data sets contain band sizes for different individuals. These band sizes are the estimated sizes of DNA fragments that are obtained as a result of electrophoresis. Fragments from a particular locus along a particular chromosome are called by the locus name, for example in D4S139, '4' refers to the fourth chromosome. Large variation in allele lengths at this locus between individuals is observed.

The file 'hdna.dat' contains the band sizes for 5 VNTR loci of OCSD. The entire record is found on a single line. The first column gives sample identifier codes, which are given according to the ethnic/racial origin of an individual. The groups and their codes are:

ACHX : Asian, Chinese

AJPX : Asian, Japanese

AKOX : Asian, Korean

AVNX : Asian, Vietnamese

BXXX : Black

CXXX : Caucasian

HXXX : Hispanic

The data for each individual is found in pairs of columns in one row. Each pair consists of the band sizes in base-pairs for that sample at that locus. There are 5 loci, hence the data is in 5 pairs of columns. Band sizes for 1299 individuals are given. However, data for all five loci is available only for the Hispanic group. From repeated measurements of the same allele, the standard deviation(SD) of the measurement error is estimated to be 0.008 times the fragment length. Table-1.1 and Table- 1.2 below give a summary of the data according to range, repeat length (ρ), number of homozygotes and heterozygotes for OCSD and LC data.

The Lifecodes data is in file 'newdata.dat'. The first pair of columns contains band sizes for D2S44, and the second pair of columns contains band sizes for D17S79 locus.

Table 1.1: Database summary for OCSD data

Name	ρ (Kb)	Range (Kb)	No. of Apparent Homozygotes	No. of Heterozygotes
D2S44	0.031	0.649 - 9.680	77	1159
D1S7	0.009	7.650 - 23.186	33	1210
D4S139	0.031	2.068 - 23.118	53	1162
D10S28	0.033	0.651 - 11.580	53	1202
D17S79	0.038	0.974 - 3.277	28	167

Table 1.2: Database summary for LC data

Name	ρ (Kb)	Range (Kb)	No. of Apparent Homozygotes	No. of Heterozygotes
D2S44	0.031	6.87 - 20.94	390	2726
D17S79	0.038	2.06 - 05.95	502	2600

The DNA is cut with the restriction enzyme *PstI*. From repeated measurements of the same allele, the standard deviation(SD) of the measurement error is estimated to be 0.00575 times the fragment length (Devlin at al. 1991).

Note that the lengths of the fragments for the same locus of LC and OCSD are different. The main reason for this is that OCSD uses a different restriction enzymes than LC. For the given data, LC has used the restriction enzyme *PstI*, while OCSD has used *HaeIII*. As discussed earlier, when restriction enzymes cut the DNA fragments, the variable region is flanked by flanking-regions. Different sizes of flanking-regions give different sizes of fragments. Some of the restriction enzymes eliminate

most of the flanking-regions, resulting in smaller restriction fragments. Cutting the DNA with *PstI* results in larger fragments than cutting with *HaeIII*.

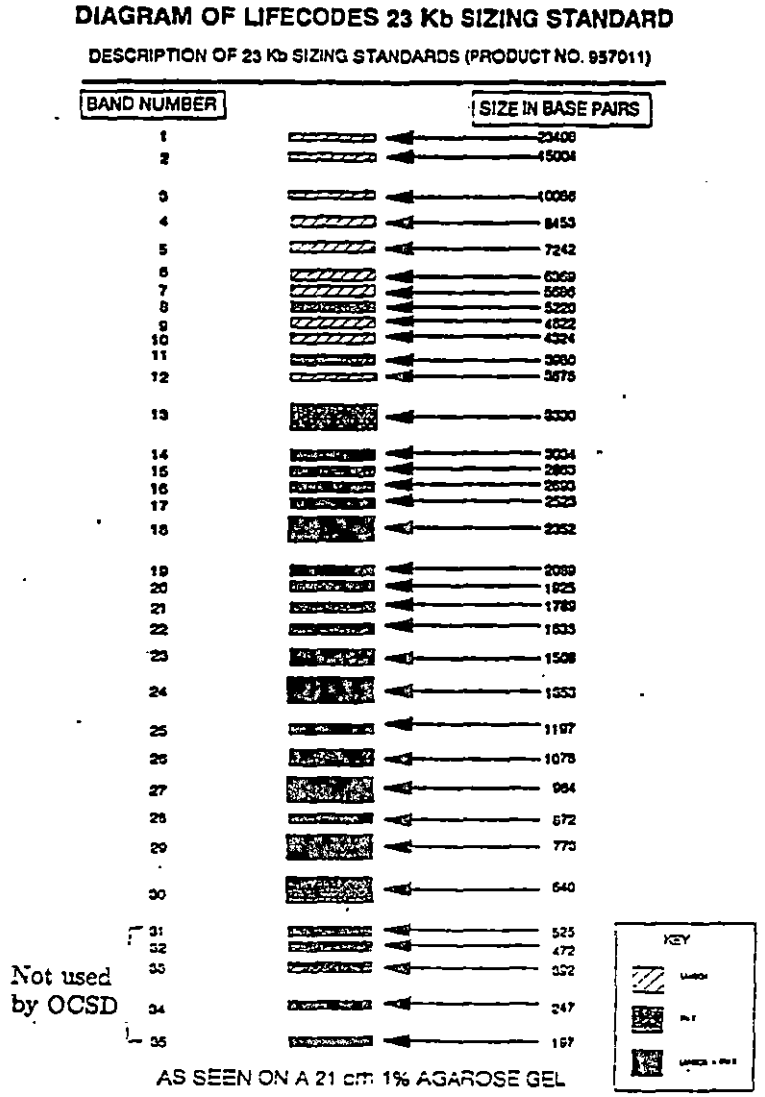


Figure 1.10: Size markers

Chapter 2

Methods of Allele Estimation

As discussed in Chapter 1, an estimate of the allele frequency distribution is required in forensic as well as paternity cases to reach any conclusions regarding the matching of DNA patterns. In Section 1 we review methods previously used to estimate allele frequency distributions, and Section 2 suggests a modification. The EM Algorithm and its application to the allele estimation problem is discussed in Section 3. Section 4 provides information on computations. The empirical Bayes method is explained in Section 5. Section 6 discusses how to test the goodness-of-fit and Section 7 shows how to obtain estimates of the variance of the frequency distribution estimates.

2.1 Estimating Allele Distributions

This section is divided into subsections on the method of binning, the theory of mixture models and its application to the allele estimation problem, the method of

maximum likelihood and the method of empirical Bayes.

2.1.1 Method of Binning

The method of binning is a conservative statistical method which permits classification of continuous data and provides simple and compact data analysis. Fixed-bin analysis for continuous distributions of alleles is discussed in Budowle et al. (1991) and Weir (1992). The fragments are classified or clustered within appropriate bins, and the frequency of occurrence is calculated using the number of sample fragments that reside in that bin. Bins are then treated as alleles and bin frequencies are used for further inference. The bin boundaries are determined by using a set of size-standard markers, Figure 1.10. The bins are sufficiently large so that more than one type of allele can reside within a bin. For each individual the two fragments at a locus are assigned to the appropriate bins and the pair of bins is then the estimated 'genotype' of the individual at that locus. Sometimes measurement error complicates the procedure of classification. If the observed length is near or on the boundary, it is possible that the true length belongs to the neighboring bin. The Federal Bureau of Investigation of the USA has adopted a window $[.975x, 1.025x]$ for fragment length x, along with the boundaries for assigning the fragments. If two bins have a common boundary then the bin with the highest frequency is chosen. For an example, suppose that locus D17S79 has two adjacent bins as 1.925- 2.088 Kb and 2.089- 3.329 Kb with frequency 23 and 8. If a fragment of size 2.089 is to be classified, its window is [2.036,

2.141] which overlaps the two bins. The maximum frequency is in bin 1.925- 2.088 Kb, hence the fragment is assigned to this class.

Another approach to estimation uses mixture models as suggested by Devlin et al. (1991). The following two sections briefly discuss the theory of mixture models and the method of maximum likelihood estimation.

2.1.2 Theory of mixture models

If a random variable or a random vector, $X \in \mathcal{X}$ is given and its probability density function, (or probability mass function, in the discrete case) can be written in the form

$$p(x) = \pi_1 f_1(x) + \ldots + \pi_k f_k(x)$$

where

$$\pi_i > 0, \;\; i = 1, \ldots, k \; ; \;\; \pi_1 + \ldots + \pi_k = 1$$

and

$$f_i(\cdot) \geq 0, \;\; \int_{\mathcal{X}} f_i(x)dx = 1, \;\; i = 1, \ldots, k,$$

then it is said that X has a *finite mixture distribution* and $p(\cdot)$, is a *finite mixture density function.* The quantities $\pi_1, \ldots, \pi_k$ are called the *mixing weights*, and $f_1(\cdot), \ldots, f_k(\cdot)$ are the *component densities* of the mixture. In many situations, $f_1(\cdot), \ldots, f_k(\cdot)$ have specified parametric forms like $f_1(x|\theta_1), \ldots, f_k(x|\theta_k)$, where θ_i are the parameters occurring in the component densities. It is not necessary that all

the component densities should belong to the same parametric family, but most of the time they come from the same family. The finite mixture density function of X in such cases is written as

$$p(x|\Psi) = \sum_{i=1}^{k} \pi_i f_i(x|\theta_i),$$

where $(\theta_1, \ldots, \theta_k)$ are all in same parameter space Θ and $\pi_i = P(\theta = \theta_i)$; $i = 1, \ldots, k$ and Ψ is the collection of all distinct parameters occurring in the model. If $G_\pi(\cdot)$ denotes the probability measure over Θ defined by π then the finite mixture density is also written as

$$p(x|\Psi) = \int_\Theta f(x|\theta) dG_\pi(\theta).$$

Once the form of the density is known, the next question is to estimate the parameters of the mixture distribution. Methods like Bayesian estimation, maximum likelihood, moment generating functions and the method of moments have been previously applied to mixtures. Devlin et al. (1991) suggested maximum likelihood(ML) and empirical Bayes(EB) methods for the estimation of the allele frequency distribution.

2.1.3 Method of Maximum Likelihood Estimation (ML)

The size of the flanking region depends upon where the restriction enzyme cuts the DNA. Two approaches are used to model flanking regions. Either the flanking region is considered to be the same for all observations or the flanking region takes on one of a fixed number of possible lengths. The latter is called a polymorphic flanking region

model. Models developed by Devlin et al. (1991) for both cases are discussed below.

1. Single flanking region Model

Let (x_j, y_j) denote the phenotype, the observed lengths of the 2 alleles, for the $j-$th two-band individual, where $j = 1, \ldots, n$. Let z_j be the phenotype of the j-th single band individual. Let n^* be the total number of two-band individuals, b be the total possible numbers of alleles associated with $r = 1, 2, \ldots, b$ tandem repeats. Let a_r be the length of an allele with r repeats, ρ be the true length of each repeat and u be the flanking region size. Then $a_r = u + r\rho$. If coalescence has not occurred, then the phenotype observed for allele r is $u + r\rho + \epsilon$, where ϵ is the measurement error. Prior studies of LC and OCSD (Communication with John Hartman) suggests that the measurement error is approximately normally distributed. Thus a measurement of an allele of length a_r is distributed $N(a_r, \sigma_r^2)$, where the standard deviation(SD), σ_r for allele a_r, is approximately ca_r and c is a constant depending on the restriction enzyme used. An estimate of c is generally obtained from duplicate measurements. The samples are processed in duplicate and the fragment sizes are determined on each autoradiogram by two independent operators. The standard deviation is obtained by calculating the difference between duplicate measurements as a function of the average size, that is SD is estimated as some % of the fragment size. This percentage is denoted by c. For LC, $c = 0.575\%$ (Baird et al. 1986) and for OC, $c = 0.8\%$ (personal communication with John Hartmann). Let $g_r(.)$ denote a normal density with mean

zero and SD σ_r. For a randomly chosen allele, the number of repeats, R is unknown. Let π_r denote the gene frequency of the r-th allele, and $\pi_r = P(R = r)$, $\sum_{r=1}^{b} \pi_r = 1$ and $\pi = (\pi_1, \ldots, \pi_b)$ denote the vector of gene relative frequencies.

The population is assumed to be in Hardy-Weinberg (H-W) equilibrium. The Hardy-Weinberg law states that in a large randomly mating population, in absence of excessive mutations, migration or selection, the relative frequencies of different genotypes remain constant over generations. Consider a two allele system, that is a pair of alleles is observed at each locus. If a pair of alleles A and B is observed with relative frequency p and q respectively then $p + q = 1$ and the relative frequency of observing genotype AA is p^2, BB is q^2 and AB is $2pq$. Besides H-W equilibrium, it is also assumed that if the observed pair of alleles is $u + r_1\rho + \epsilon_1$, and $u + r_2\rho + \epsilon_2$ then ϵ_1 and ϵ_2 are independent. If coalescence is ignored, then the probability density function for length $X = u + r\rho + \epsilon$ for a randomly selected fragment, is

$$f(x|\pi, u) = \sum_{r=1}^{b} \pi_r g_r(x - u - r\rho).$$

and the likelihood of phenotype (x_j, y_j) is proportional to the product of the densities, namely

$$L(\pi, u|x_j, y_j) \propto f(x_j|\pi, u) f(y_j|\pi, u).$$

If coalescence is considered then the above model changes. In Devlin et al. (1991) the probability of coalescence is modeled as a function of both the average pair length and the difference in allele lengths. Let $\delta(t, z)$ be the conditional probability of

coalescence given $t = \frac{1}{2}|x - y|$ and the average fragment size is $z = \frac{1}{2}(x + y)$. Given (x_j, y_j) the likelihood for a two-band observation (x_j, y_j) is:

$$L(\pi, u|x_j, y_j) \propto f(x_j|\pi, u) f(y_j|\pi, u)[1 - \delta(t_j, z_j)],$$

When the difference between the allele lengths is very large it is highly unlikely that they would coalesce. Hence the probability of coalescence can be considered to be negligible when $(x_j \neq y_j)$. Given a single band z_j, the likelihood is then

$$L(\pi, u|z_j) \propto \int_0^\infty f(z_j - t|\pi, u) f(z_j + t|\pi, u) \delta(t, z_j) dt.$$

The log likelihood for the sample is the sum over the n^* two-band individuals plus that over the $(n - n^*)$ single band individuals.

$$\begin{aligned} \mathcal{L}(\pi, u|data) &= \sum_{j=1}^{n^*} [\log f(x_j|\pi, u) + \log f(y_j|\pi, u)] \\ &+ \sum_{j=n^*+1}^{n} \log[\int_0^\infty f(z_j - t|\pi, u) f(z_j + t|\pi, u) \delta(t, z_j) dt]. \end{aligned} \quad (2.1)$$

Given prior estimates of b and the coalescence probability δ, estimates of u and π can be obtained. We obtained an estimate of b by dividing the difference of the largest and the smallest fragment by its repeat length. For example for D17S79 locus of LC, b is estimated as 93 by dividing the range 3.53 Kb by the repeat length 0.038 Kb. An estimate of an upper bound of u, can be obtained as follows. If a_{min} is the true minimum length with r_{min} repeats and flanking region u, then $a_{min} = u + r_{min}\rho$, where ρ is the repeat length. An estimate of the upper bound of the flanking region size is the smallest observed allele length minus $r_{min}\rho$. For example for D17S79 locus

of LC, the minimum length observed is 2.06 Kb, and $\rho = 0.038$ Kb. If the minimum length contains a minimum of one repeat, then an estimate of the maximum value of the flanking region would be 2.022 (2.06- 0.038). Note that the error term for the smallest observed allele length, $x_{min} = u + r_{min}\rho + \epsilon_{min}$ can be negative, and so, a more conservative estimate of the upper bound of u is simply x_{min}.

In the above model the errors of a pair of fragments (ϵ_1, ϵ_2) are assumed to be independent when in fact they may not be. The exact relationship is unknown, however Devlin et al. (1992) suggests that the correlation between measurement errors is a function of the absolute value of the difference in allele length t and the average length z. We do not incorporate this into our model.

2. Model for Polymorphic Flanking Regions

Assume that we have phenotypes for n individuals and that L different flanking-region sizes are possible. If $\ell = 1, \ldots, L$, indexes the flanking-region sizes, then the length of the allele with repeat r and with flanking region u_ℓ is given as $a_{r\ell} = u_\ell + r\rho$, where u_ℓ is the $\ell-$th flanking-region size and $u = (u_1, \ldots, u_L)$, $u_1 < \ldots < u_L$. Because the alleles are measured with error, the observed length of an allele is $u_\ell + r\rho + \epsilon$, where $\epsilon \sim N(0, \sigma^2_{r\ell})$ and $\sigma_{r\ell} = ca_{r\ell}$.

Let $\gamma_{r\ell} = P(U = u_\ell, R = r)$ denote the relative frequency of allele $a_{r\ell}$. Let $\phi_\ell = P(U = u_\ell)$ be the proportion of observations with flanking region size u_ℓ. Let $\xi_\ell = \phi_\ell^{-1}(\gamma_{1\ell}, \ldots, \gamma_{b\ell})$ denote the conditional distribution of allele size for the $\ell-$th

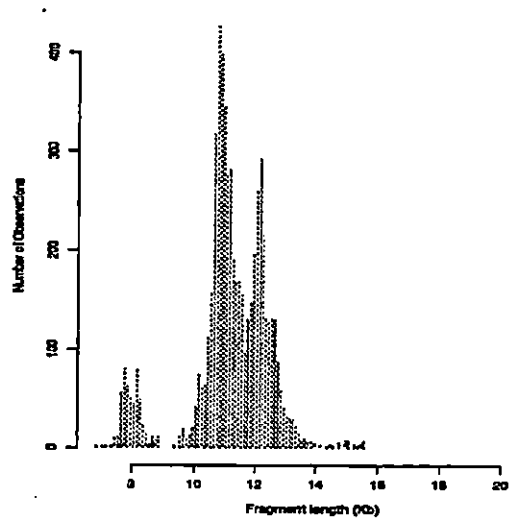


Figure 2.1: Frequency distribution of 6232 unpaired restriction fragments of D2S44 (LC).

flanking-region size. If flanking-region size U and repeat number R are independent then $\xi_1 = \ldots = \xi_L$. The error for allele $a_{r\ell}$ is assumed to distributed $N(0, \sigma^2_{r\ell})$. The probability density of a randomly selected fragment X is

$$f(x) = \sum_{\ell} \sum_{r} \gamma_{r\ell} g_{r\ell}(x - u_\ell - r\rho).$$

where $g_{r\ell}(.)$ is the normal density, with mean zero and SD $\sigma_{r\ell}$. This gives the general model.

The assumptions of the polymorphic flanking region model are similar to those for the single flanking region model. The errors are assumed independent when the population is in H-W equilibrium. If L is unknown then it can be difficult to estimate the parameters of the model because of non-identifiability. Note that patterns commonly occur in the empirical frequency distributions, an example of which is given in Figure

2.1. Because of this, when L is unknown, any repeating pattern in the repeat-unit mixing distribution can lead to an overestimate of L. In addition, if the distributions overlap, then unique estimates are not available because two alleles of the same length are not distinguishable according to their flanking-region sizes. Clearly $\gamma_{r\ell}$ and $\gamma_{r'\ell'}$ are not identifiable when $a_{r\ell} = u_\ell + r\rho = u_{\ell'} + r'\rho = a_{r'\ell'}$. Without proper information on the parameter space, estimating $\gamma_{r\ell}$'s can be difficult and the mixture model method may not be appropriate.

If the distributions are separated in space, then the estimation problem is similar to the single-flanking-region problem applied to L distinct subsets. It is recommended in Devlin et al. (1991) that in practice if the overlap of the distributions displayed by the empirical distribution is small, then the observations can be grouped separately and the single-flanking-region model can be used for estimation of the parameters for each group. While if the distributions overlap, but one of the $\phi_\ell{}'s$ appears to be small, then the presence of a rare distribution can be ignored since its effect on the estimates would be small. There is a case when the model is identifiable even if repeat-unit mixing distributions overlap. This special case arises when the flanking region size u and repeat size R are independent so that $\xi_1 = \ldots = \xi_L$ or $\gamma_{r\ell} = \phi_\ell \pi_r$.

The log likelihood has the same form as given for the single-flanking-region, that is the log likelihood for the sample is the sum over the $n^\star$ two-band individuals plus

that over the $(n - n^*)$ single band individuals.

$$\mathcal{L}(\pi, u|data) = \sum_{j=1}^{n^*}[\log f(x_j|\pi, u) + \log f(y_j|\pi, u)] + \sum_{j=n^*+1}^{n} \log[\int_0^\infty f(z_j - t|\pi, u) f(z_j + t|\pi, u)\delta(t, z_j)dt,] \quad (2.2)$$

where $f(x|u, \pi, \phi)$ the probability density of an randomly selected fragment $X = u_l + r\rho + \epsilon$ is given as

$$f(x|u, \pi, \phi) = \sum_{\ell=1}^{L} \phi_\ell \sum_{r=1}^{b} \pi_r g_{r\ell}(x - u_{r\ell} - r\rho).$$

Provided L is known, estimates of ϕ and u are obtained.

3. Probability of Coalescence

Both the models for single and polymorphic flanking regions incorporate $\delta(t, z)$, the probability of coalescence. In this subsection we review a method for estimating $\delta(t, z)$ proposed by Devlin et al. (1990). Sometimes for near heterozygotes the distance between the two bands is less than the minimum resolution possible for a particular device and the technician is not able to distinguish the bands with great precision. The bands are merged and only one band is observed. The length of this observed band is approximately equal to the average length of two bands. If phenotypes observed are affected by coalescence, it results in increasing the number of apparent homozygotes and decreasing the number of heterozygotes. An excess of apparent homozygotes indicates the possibility of observing close heterozygotes as homozygotes due to coalescence.

The presence of coalescence can be determined by plotting the length of larger segments(y) against the length of smaller segments(x) or by plotting the absolute value of the fragment pair difference against the average fragment pair length. In the plot of longer fragments against shorter fragments, points on the line $y = x$(equal fragment length) represent the apparent homozygotes. In the presence of coalescence, the number of points in intervals adjacent to the line $y = x$ is substantially less than the number of points on line $y = x$. This clearly indicates that some of the close heterozygotes, that is the heterozygotes with smaller differences, are measured as homozygotes. In the case of contamination of the data due to coalescence, observations for smaller differences are missing in the plot of absolute difference versus the average length. See for example Figure 3.8.

Devlin et al. (1990) estimate the probability of coalescence as a function of the absolute value of the difference of pairs of allele lengths, as follows. Let $t_j = |x_j - y_j|$, $j = 1, \ldots, n$ be the absolute differences for the n pairs. The range of t can be divided into m intervals. Let c_k be the midpoint of the k-th interval say B_k, and $2d$ be the length of each interval. The observed frequency O_k for interval $B_k, (k = 1, \ldots, m)$ can be obtained by counting the observations falling in that interval and the expected number E_k can be obtained by integrating over the distribution function,

$$E_k = n \int I[\ (c_k - d) < |x - y| \leq (c_k + d)\ ] dF(x) dF(y)$$

where I is an indicator function with value 1 if the condition is satisfied and 0 otherwise. As the distribution function $\mathcal{F}$ is unknown, E_k can be estimated using the

empirical distribution function. Hence

$$\hat{E}_k = \frac{1}{4n}\sum_{i=1}^{2n}\sum_{j=1}^{2n} I[(c_k - d) < |X_i - X_j| \leq (c_k + d)]$$

where X_i, $i = 1, \ldots, 2n$ denotes the observed unpaired heterozygote fragment lengths. Devlin et al. (1990) use $(1 - \frac{O_k}{\hat{E}_k})$ $k = 1, \ldots, m$, to estimate the probability of coalescence. Since, O_k is an estimate of

$$nP\{t \in B_k \; AND \; no \; coalescence\} = nP\{no \; coalescence \mid t \in B_k\}P\{t \in B_k\}$$

and $\hat{E}_k$ estimates $nP\{t \in B_k\}$, the ratio $\frac{O_k}{\hat{E}_k}$ is an estimate of $P\{no \; coalescence \mid t \in B_k\}$. Hence, a smooth curve $\delta(\cdot)$, fitted through the points $(1 - \frac{O_k}{\hat{E}_k})$ versus c_k, can be used to estimate the probability of coalescence as a function of t, the allele pair length difference.

Figure 3.2 shows an example of this procedure. The function $\delta(t)$ is used in likelihood equations (2.1) and (2.2) for further estimation.

2.2 Another Approach

In the case of small samples, it can be difficult to estimate the probability of coalescence by the method discussed above. If the sample size is small and the data is widely spread, then the observed frequencies can be higher than the expected for most of the class intervals and the estimates $[1 - (\frac{O_k}{\hat{E}_k})]$ for c_k are negative.

If the distance between two bands is not enough to distinguish them, then they are identified as a single band, see for example Figure 2.2. We are interested in the

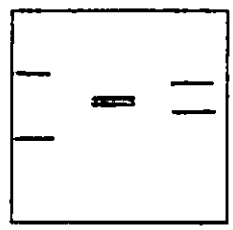


Figure 2.2: Audiogram showing that a single band is observed in lane 2 as the distance between the two bands is less than the minimum resolution distance. In other lanes two bands are observed.

likelihood of observing a single band. This can be written in a different way using the definition of coalescence. Coalescence comes into the picture when the distance between the observed band is less than the minimum resolution distance, say α. The molecular weight of a single observed band z, is approximately equal to the average of two bands. Most of the labs use resolution threshold, the minimum distance required for identification of two bands, as some percentage of the band weight of a single band observed. The resolution distance is generally considered to be $1-2\%$ of molecular weight of the single band observed (Berry, 1991; Morris, 1992). This resolution threshold can be estimated by plotting the percentage difference of lengths as a function of the average length (Morris, 1992).

Given any two alleles of length x and y, a single band will be observed if the difference between x and y is less than the resolution threshold. We have to determine the probability of observing a single band of length $z = \frac{1}{2}(x+y)$, $f(z)$, when $|x-y| \leq \alpha$. Using the transformation of variables techniques, let $Z = \frac{1}{2}(X+Y)$, $V = X-Y$

then

$$f(z) = \int_{-\alpha}^{\alpha} f(z,v)dv,$$

$$f(z,v) = f(x,y)\ |J|$$

If X and Y are independent, then $f(x,y) = f(x)\ f(y)$ where

$$X = Z + \frac{V}{2}\ , Y = Z - \frac{V}{2}$$

The likelihood of z can be written as

$$\begin{aligned} f(z) &= \int_{-\alpha}^{\alpha} f(z+\frac{v}{2}) f(z-\frac{v}{2}) dv \\ &= \int_{-\alpha}^{\alpha} \sum_{r_1=1}^{b} \sum_{r_2=1}^{b} \pi_{r_1}\pi_{r_2} g_{r_1}(z+\frac{v}{2}-u-r_1\rho) g_{r_2}(z-\frac{v}{2}-u-r_2\rho) dv \end{aligned}$$

Then the log likelihood for the sample is

$$\begin{aligned} \mathcal{L}(\pi, u|data) &= \sum_{j=1}^{n^*} [\log f(x_j|\pi,u) + \log f(y_j|\pi,u)] \\ &+ \sum_{j=n^*+1}^{n} \log[\int_{-\alpha}^{\alpha} f(z+\frac{v}{2}) f(z-\frac{v}{2}) dv]. \end{aligned} \tag{2.3}$$

This method of writing the likelihood is simple and clear to understand. We compare maximum likelihood estimates for allele distributions using the likelihood in (2.3) with (2.1) in the next chapter.

The use of the model (2.3) to estimate the allele distribution appears to be new. Berner and Morris (1990) use the resolution threshold to estimate the probability of a random match.

Henceforth, the likelihood given in equation (2.1) is referred to as the *coalescence likelihood* and the likelihood given in equation (2.3) is referred to as the *resolution likelihood.*

2.3 EM Algorithm

The EM Algorithm is a general approach for obtaining maximum likelihood estimates when the observations can be viewed as incomplete data (Dempster, et al. 1991). The term 'incomplete data' indicates the existence of two sample spaces $\mathcal{C}$ and $\mathcal{X}$ and a many-to-one mapping from $\mathcal{C}$ to $\mathcal{X}$. The observed data is $x \in \mathcal{X}$ and $c \in \mathcal{C}$ is indirectly observed through x; c is *complete* data and x is *incomplete* data. The corresponding family of sampling densities are derived from a family of sampling densities $f(c|\psi)$, depending on parameters ψ as

$$g(x|\psi) = \int_{\mathcal{C}(x)} f(c|\psi)dc.$$

The aim is to get a value of ψ which maximizes $g(x|\psi)$ making use of $f(c|\psi)$. Given the incomplete-data specification $g(x|\psi)$, there are many possible complete-data specifications $f(c|\psi)$ that will generate $g(x|\psi)$.

The EM algorithm is an iterative technique which alternatively estimates the expectation of a function of unknown parameters, given some prior estimates and then maximizes this expected function. Let $Q(\psi'|\psi) = E[\log f(c|\psi')|x, \psi]$, which exists for all pairs (ψ', ψ), where $f(c|\psi) > 0$ for $c \in \mathcal{C}$ for all $\psi \in \Omega$. Let $\psi^{(m)}$

denote the value of the parameter at the m-th iteration. Then the EM iteration $\psi^{(m)} \rightarrow \psi^{(m+1)}$ is

E-step : Compute $Q(\psi|\psi^{(m)})$.

M-step : choose $\psi^{(m+1)}$ which maximizes $Q(\psi|\psi^{(m)})$ over $\psi \in \Omega$.

Iterations are performed until some $\psi^* = \psi^{(m)} = \psi^{(m+1)}$ which maximizes $Q(\psi|\psi^{(m)})$ is obtained. In practice other stopping criteria are used. We used the difference between the value of objective function at two successive iterations. Iterations are performed until the absolute difference is smaller than the given tolerance limit 0.0001.

The EM algorithm can be applied to maximize the likelihood in the allele estimation problem. We outline the details for the coalescence likelihood. The derivation for the resolution likelihood is similar. Let us consider a single allele. Let X be the length of an allele measured with error. We have modeled this as $X = u + R\rho + \epsilon$, where the number of repeats, R is unknown, $1 \leq R \leq b$, and b is estimated from the data. Here, X is the incomplete data, and (X, R) can be called the complete data. It is convenient to use another representation of the complete data as $C = (X, H)$, where H is an indicator vector of length b, with entries zero everywhere except in the r-th location. The complete data likelihood for $\psi = (\pi, u)$ given the one observation $c = (x,\ H)$ is,

$$f(c|\psi) = \prod_{j=1}^{b} \pi_j^{h_j} g(x|R = j, u)^{h_j},$$

where $\pi_j = P(R = j)$. The complete data log-likelihood for one observation is,

$$f(c|\psi) = \sum_{j=1}^{b} h_j \log\ \pi_j\ +\ h_j \log\ g(x|R = j, u)$$

If we have n observations on X, then the log-likelihood for n observations is written as,

$$\begin{aligned} \log\ f(c|\psi) &= \sum_{i=1}^{n}\sum_{j=1}^{b} h_{ij}\ \log\ \pi_j + \sum_{i=1}^{n}\sum_{j=1}^{b} h_{ij}\ \log\ g(x_i|R = j, u) \\ &= \sum_{i=1}^{n} h_i^T \mathcal{V}(\pi) + \sum_{i=1}^{n} h_i^T \mathcal{U}_i(R, u), \end{aligned}$$

where $\mathcal{V}(\pi)$ and $\mathcal{U}_i(R, u)$ are vectors of length b with j-th element as $\log\ \pi_j$ and $\log\ g(x_i|R = j, u)$ respectively. In the above defined problem $Q(\psi|\psi^{(m)})$ is,

$$Q(\psi|\psi^{(m)}) = \sum_{i=1}^{n} E(h_i|x_i, \psi^{(m)})^T \mathcal{V}(\pi) + \sum_{i=1}^{n} E(h_i|x_i, \psi^{(m)})^T \mathcal{U}_i(R, u),$$

where $E(h_i|x_i, \psi^{(m)})$ is a vector of length b with j-th component as

$$\frac{\pi_j^{(m)} g(x_i|R = j, u^{(m)})}{\sum_{j=1}^{b} \pi_j^{(m)} g(x_i|R = j, u^{(m)})} = \frac{\pi_j^{(m)} g_j^{(m)}(x_i)}{\sum_{j=1}^{b} \pi_j^{(m)} g_j^{(m)}(x_i)}$$

The M-step of the EM iteration is to maximize $Q(\psi|\psi^{(m)})$ over $\psi \in \omega$ and obtain $\psi^{(m+1)}$.

In our estimation problem, we have three data vectors, x and y when two bands are observed for n^* individuals and z for $n - n^*$ individuals with a single band. Here the incomplete data vectors are x, y and z. We have b possible values of repeats for each measured length. Each of x_i, y_i or z_i is associated with one of the b repeats. Let $h = (h_1, h_2, \ldots, h_{n^*}), s = (s_1, s_2, \ldots, s_{n^*})$ are two indicator vectors associated with x,

y respectively. Since

$$P[z_j|\pi,u] = \int_0^\infty f(z_j - t|\pi,u) f(z_j + t|\pi,u)\delta(t,z_j)dt.$$

Let $v = (v_1, v_2, \ldots, v_{n-n^*})$, and $w = (w_1, w_2, \ldots, w_{n-n^*})$ be indicator vectors associated with $z_j - t$ and $z_j + t$ respectively.

Let $\psi = (\pi, u)$ then $\psi^{(m)} = (\pi^{(m)}, u^{(m)})$, where $\pi^{(m)}, u^{(m)}$ be the value of π and u at the m-th iteration. We know that the log-likelihood for the complete data is the sum over the n^* two-band individuals plus that over the $n - n^*$ single band individuals.

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{n^*}\left\{\sum_{r=1}^{b}[h_{ir}\log\ \pi_r + h_{ir}\log\ g_r(x_i) + s_{ir}\log\ \pi_r + s_{ir}\log\ g_r(y_i)]\right\} \\ &+ \sum_{i=1}^{n-n^*}\left\{\sum_{r=1}^{b}\sum_{r'=1}^{b} v_{ir}w_{ir'}\log\ \pi_r\pi_{r'}\int_0^\infty g_r(z_i - t)g_{r'}(z_i + t)\delta(t)dt\right\}.\end{aligned}$$

Let $A_{rr'}(z_i) = \int_0^\infty g_r(z_i - t)g_{r'}(z_i + t)\delta(t)dt$, and $H_r = \sum_{i=1}^{n^*} h_{ir}$, $S_r = \sum_{i=1}^{n^*} s_{ir}$, $V_r = \sum_{i=1}^{n-n^*} v_{ir}$, $W_r = \sum_{i=1}^{n-n^*} w_{ir}$.

Then complete log-likelihood is rewritten as

$$\begin{aligned}\mathcal{L} &= \sum_{r=1}^{b}\left([H_r + S_r + V_r + W_r]\log\pi_r + \sum_{i=1}^{n^*}[h_{ir}\log\ g_r(x_i) + s_{ir}\log g_r(y_i)]\right) \\ &+ \sum_{r=1}^{b}\sum_{r'=1}^{b}\sum_{i=1}^{n-n^*} v_{ir}w_{ir'}\log A_{rr'}(z_i).\end{aligned}$$

For the E-step, $Q(\psi|\psi^{(m)})$ is then written as

$$\begin{aligned}&= \sum_{r=1}^{b}\left(\ E[H_r|\mathbf{x},\psi^{(m)}] + E[S_r|\mathbf{y},\psi^{(m)}] + E[V_r|\mathbf{z},\psi^{(m)}] + E[W_r|\mathbf{z},\psi^{(m)}]\ \right)\log\pi_r \\ &+ \sum_{r=1}^{b}\sum_{i=1}^{n^*}\left(\ E[h_{ir}|x_i,\psi^{(m)}]\log\ g_r(x_i) + E[s_{ir}|y_i,\psi^{(m)}]\log\ g_r(y_i)\ \right)\end{aligned}$$

$$+ \sum_{r=1}^{b}\sum_{r'=1}^{b}\sum_{i=1}^{n-n^*} E[v_{ir}w_{ir'}|z_i,\psi^{(m)}]\ \log\ A_{rr'}(z_i),$$

because of the assumed independence of x, y, z. Here,

$$\begin{aligned}
E[h_{ir}|\mathbf{x},\psi^{(m)}] &= \frac{\pi_r^{(m)}g_r^{(m)}(x_i)}{\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(x_i)} = \frac{\pi_r^{(m)}g_r^{(m)}(x_i)}{f^{(m)}(x_i)},\\
E[s_{ir}|\mathbf{y},\psi^{(m)}] &= \frac{\pi_r^{(m)}g_r^{(m)}(y_i)}{\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(y_i)} = \frac{\pi_r^{(m)}g_r^{(m)}(y_i)}{f^{(m)}(y_i)},\\
E[v_{ir}w_{ir'}|\mathbf{z},\psi^{(m)}] &= P[v_{ir}=1,w_{ir'}=1|z_i,\psi^{(m)}]\\
&= \frac{\pi_r^{(m)}\pi_{r'}^{(m)}A_{rr'}^{(m)}(z_i)}{\sum_{r=1}^{b}\sum_{r'=1}^{b}\int_0^\infty \pi_r^{(m)}g_r^{(m)}(z_i-t)\pi_{r'}^{(m)}g_{r'}^{(m)}(z_i+t)\delta(t)dt}\\
&= \frac{\pi_r^{(m)}\pi_{r'}^{(m)}A_{rr'}^{(m)}(z_i)}{\int_0^\infty f^{(m)}(z_i-t)f^{(m)}(z_i+t)\ \delta(t)\ dt}\\
&= \frac{\pi_r^{(m)}\pi_{r'}^{(m)}A_{rr'}^{(m)}(z_i)}{K^{(m)}(z_i)}.
\end{aligned}$$

Maximizing $Q(\psi|\psi^{(m)})$ over π, such that $\sum_{r=1}^{b}\pi_r = 1$, then

$$\pi_r^{(m+1)} = \frac{E[H_r^{(m)}|\mathbf{x}] + E[S_r^{(m)}|\mathbf{y}] + E[V_r^{(m)}|\mathbf{z}] + E[W_r^{(m)}|\mathbf{z}]}{\sum_{r=1}^{b}\left\{E[H_r^{(m)}|\mathbf{x}] + E[S_r^{(m)}|\mathbf{y}] + E[V_r^{(m)}|\mathbf{z}] + E[W_r^{(m)}|\mathbf{z}]\right\}}.$$

or

$$\pi_r^{(m+1)} = \frac{\pi_r^{(m)}}{2n}\left(\sum_{i=1}^{n^*}\left[\frac{g_r^{(m)}(x_i)}{f^{(m)}(x_i)} + \frac{g_r^{(m)}(y_i)}{f^{(m)}(y_i)}\right] + \sum_{i=n^*+1}^{n}\frac{B^{(m)}(z_i)}{K^{(m)}(z_i)}\right), \tag{2.4}$$

where

$$B^{(m)}(z_i) = \int_0^\infty [\ f^{(m)}(z_i+t)g_r^{(m)}(z_i-t) + f^{(m)}(z_i-t)g_r^{(m)}(z_i+t)\ ]\ \delta(t)\ dt.$$

Maximizing $Q(\psi|\psi^{(m)})$ over u, we obtain the equation for estimating the flanking region size u. Here, $\frac{\partial}{\partial u}Q(\psi|\psi^{(m)})$, ignoring the dependence of $\hat{\sigma}_r = ca_r = c(u_r\rho)$ on

u, is

$$
\begin{aligned}
&= \sum_{i=1}^{n^*} \left\{ \sum_{r=1}^{b} \frac{\pi_r^{(m)} g_r^{(m)}(x_i)}{f^{(m)}(x_i)} \frac{(x_i - u - r\rho)}{\hat{\sigma}_r^2} + \sum_{r=1}^{b} \frac{\pi_r^{(m)} g_r^{(m)}(y_i)}{f^{(m)}(y_i)} \frac{(y_i - u - r\rho)}{\hat{\sigma}_r^2} \right\} \\
&+ \sum_{i=1}^{n-n^*} \sum_{r=1}^{b} \sum_{r'=1}^{b} \left\{ \frac{\pi_r^{(m)} \pi_{r'}^{(m)} A_{rr'}^{(m)}(z_i)}{K^{(m)}(z_i)} \left[\frac{\int_0^\infty g_r(z_i - t) \frac{(z_i - t - u - r\rho)}{\hat{\sigma}_r^2} g_{r'}(z_i + t) \delta(t)\, dt}{A_{rr'}(z_i)} \right] \right. \\
&+ \left. \left[\frac{\int_0^\infty g_r(z_i - t) g_{r'}(z_i + t) \frac{(z_i + t - u - r\rho)}{\hat{\sigma}_r^2} \delta(t)\, dt}{A_{rr'}(z_i)} \right] \right\} \qquad (2.5)
\end{aligned}
$$

which is,

$$
\begin{aligned}
&= \sum_{i=1}^{n^*} \left[\frac{D_1(x_i)}{f^{(m)}(x_i)} + \frac{D_1(y_i)}{f^{(m)}(y_i)} \right] \\
&+ \sum_{i=1}^{n-n^*} \sum_{r=1}^{b} \sum_{r'=1}^{b} \left\{ \frac{\pi_r^{(m)} \pi_{r'}^{(m)} A_{rr'}^{(m)}(z_i)}{K^{(m)}(z_i)} \left[\frac{\int_0^\infty g_r(z_i - t) \frac{(z_i - t - u - r\rho)}{\hat{\sigma}_r^2} g_{r'}(z_i + t) \delta(t)\, dt}{A_{rr'}(z_i)} \right] \right. \\
&+ \left. \left[\frac{\int_0^\infty g_r(z_i - t) g_{r'}(z_i + t) \frac{(z_i + t - u - r\rho)}{\hat{\sigma}_r^2} \delta(t)\, dt}{A_{rr'}(z_i)} \right] \right\}, \qquad (2.6)
\end{aligned}
$$

where

$$
D_1(x) = \sum_{r=1}^{b} \pi_r^{(m)} g_r^{(m)}(x) \frac{(x - u - r\rho)}{\hat{\sigma}_r^2}.
$$

An estimate of u at the (m+1)-st iteration, is the value $u^{(m+1)}$ such that $\frac{\partial Q}{\partial u} |_{u^{(m+1)}} = 0$. As $u^{(m)}$ and $u^{(m+1)}$ becomes close, we have that $\frac{A_{rr'}^{(m)}(z_i)}{A_{rr'}^{(m+1)}(z_i)} \cong 1$, and the estimating equation becomes

$$
\begin{aligned}
\frac{\partial Q}{\partial u} &= \sum_{i=1}^{n^*} \left[\frac{D_1(x_i)}{f^{(m)}(x_i)} + \frac{D_1(y_i)}{f^{(m)}(y_i)} \right] + \\
&\sum_{i=1}^{n-n^*} \frac{\int_0^\infty [\, f^{(m,*)}(z_i + t) D_1(z_i - t) + f^{(m,*)}(z_i - t) D_1(z_i + t)\,] \delta(t)\, dt}{K^{(m)}(z_i)}, \qquad (2.7)
\end{aligned}
$$

where $f^{(m,*)}(x) = \sum_{r=1}^{b} \pi_r{}^{(m)} g_r(x - u - r\rho)$. When we incorporate the derivative of $\hat{\sigma}_r$ with respect u, we obtain the following estimating equation

$$
\frac{\partial Q}{\partial u} = \sum_{i=1}^{n^*} \left[\frac{D_2(x_i)}{f^{(m)}(x_i)} + \frac{D_2(y_i)}{f^{(m)}(y_i)} \right] +
$$

$$\sum_{i=1}^{n-n^*} \frac{\int_0^\infty [\, f^{(m,*)}(z_i+t)D_2(z_i-t) + f^{(m,*)}(z_i-t)D_2(z_i+t)\,]\delta(t)\,dt}{K^{(m)}(z_i)}, \tag{2.8}$$

where

$$D_2(x) = \sum_{r=1}^{b} \pi_r^{(m)} g_r^{(m)}(x) \left[\frac{(x-u-r\rho)}{\hat{\sigma}_r^2} + c\frac{(x-u-r\rho)^2}{\hat{\sigma}_r^3} - \frac{c}{\hat{\sigma}_r} \right].$$

We equated $\frac{\partial Q}{\partial u}$ in equation (2.8) to zero, and solved for $u^{(m+1)}$, to obtain an estimate of u at the (m+1)-st iteration.

Our equation (2.4) for estimating π, is same as given in Devlin et al (1991) but the equation for estimating the flanking-region size, u in equation (2.8) is similar to that provided by them. Devlin et al (1991) disregard the dependence of $\hat{\sigma}_r$ on u, and obtain an equation for estimating flanking region size, by equating the derivative of the likelihood function in equation (2.1) evaluated at $\psi^{(m)} = (\pi^{(m)}, u^{(m)})$ to $u^{(m+1)}$. The equation given by Devlin et al (1991) is,

$$\begin{aligned} u^{(m+1)} &= \sum_{i=1}^{n^*} \left[\frac{D(x_i)}{f^{(m)}(x_i)} + \frac{D(y_i)}{f^{(m)}(y_i)} \right] \\ &+ \sum_{i=n^*+1}^{n} \frac{\int_0^\infty \left[\, f^{(m)}(z_i+t)D(z_i-t) + f^{(m)}(z_i-t)D(z_i+t)\,\right]\delta(t)\,dt}{K^{(m)}(z_i)} \end{aligned} \tag{2.9}$$

where

$$D(x) = \sum_{r=1}^{b} \pi_r^{(m)} g_r^{(m)}(x) \frac{(x-u^{(m)}-r\rho)}{\hat{\sigma}_r{}^2}.$$

When we used (2.9) to estimate u, we found that the routine would not converge despite the use of several different starting values.

The next section provides information on computations.

2.4 Computations

All the computations for this project were done on an HP-735 computer under HP-Unix. For obtaining allele frequency estimates, four Fortran programs were written in HP Fortran 77. Table 2.1 gives the names of the 4 routines according to the algorithm and the likelihood equation used.

Table 2.1: Method Names

	Algorithm	
Likelihood	E04VDF	EM
Coalescence (2.1)	SQPCOAL	EMCOAL
Resolution (2.3)	SQPRESL	EMRESL

Some routines which are taken from NAG Fortran Library, Mark 15 are listed below.

- E04VDF: This routine does constrained minimization of a nonlinear function of several variables using a Sequential Quadratic Programming(SQP) method. For maximizing a function, the negative of the function can be minimized. The function, first order derivatives of the function, linear constraints and initial values of u and $\bar{\pi}$ must be provided. For all the data sets we used starting values of π as $\frac{1}{b}$ and u close to the minimum observed fragment length. The objective function is said to converge if the relative change occurring between

two successive iterations is less than some prescribed tolerance.

- E04ZCF: This is used in conjunction with E04VDF to check the provided gradients of the likelihood function.

- D01BAF: Computes an estimate of a one dimensional definite integral using Gaussian-Legendre formula with a specified number of abscissae. (32 abscissae are used in each of the four programs).

- C05NCF: A routine to find a solution of a system of nonlinear equations by modification of the Powell Hybrid method.

SQPCOAL computes the allele distribution using the coalescence likelihood given in equation (2.1) and it includes NAG routines E04ZCF, E04VDF, and D01BAF. In the coalescence likelihood (2.1), the second term of product of densities and coalescence probability is integrated over the interval $(0, \infty)$, and is nonzero only for a finite range. Because of this, the final answers were different for different numerical approximation routines. Hence, the finite range was determined. After integrating the function over this finite range, all approximation routines gave the same results. For LC D17S79 locus, the product of densities and the probability of coalescence is nonzero only over the range (0, 0.5).

SQPRESL, the program which uses the resolution likelihood equation (2.3) also uses NAG routines E04ZCF, E04VDF, and D01BAF.

Programes EMCOAL and EMRESL use the EM algorithm discussed in Section

2.3. The function, estimating equations and initial values of the parameters must be provided. For all the data sets we used starting values of π as $\frac{1}{b}$ and u close to the minimum observed fragment length. From equation (2.8), it is necessary to solve the nonlinear equation in u to estimate the flanking region size u. C05NCF is used to estimate u, and then $\hat{u}$ is used for estimation of π. Such iterations are performed until the number of iterations exceed the maximum number of iterations or the absolute difference between the objective function at two successive iterations is less than the given tolerance limit. For example, we have used a tolerance limit 0.0001. EMCOAL and EMRESL are the EM algorithm codes with coalescence and resolution likelihood respectively.

All other computations are done in SPlus. The procedures for Fixed-bin analysis, drawing the samples, estimating correlation, testing the goodness-of-fit, obtaining the EB estimates and calculating the SD of the estimates, estimating resolution threshold and plotting the graphs are written in SPlus specially for this project.

2.5 Empirical Bayes (EB) estimates

The method of maximum likelihood estimation is well known and often used because of its intuitive appeal and asymptotic properties. In this application the measurement error is a function of the restriction fragment length and for large fragments, the standard deviation SD of the measurement error is large relative to the repeat

length. Hence there can be substantial overlap in the measurements of alleles which differ by a smaller number of repeats. Due to this overlapping, there is a possibility of misclassification of the alleles, which can cause over and underestimation in the frequency distribution. ML estimates may be unreliable in this case and they may require some local smoothing. It is suggested in Devlin et al. (1991) that EB estimates would be more reliable as compared to ML estimates.

In general, there is a large variation in the lengths of the alleles observed at the VNTR loci, which results in a large number of allele frequency parameters in the model. If the sample sizes are not large enough for the number of parameters to be estimated, the parameter estimates will have large variances. The error in one allele frequency estimate changes the neighboring frequency estimates. If one frequency is overestimated, neighboring frequencies are underestimated, and the error in estimation of a particular parameter tends to be negatively correlated with the error in estimation of the neighboring parameters. Hence more consistent, reliable estimates may be obtained by smoothing.

The empirical Bayes(EB) method suggested in Devlin et al. (1991) described below is based on Stein shrinkage estimators (James and Stein, 1961; Efron and Morris, 1973). Assume $Y_r \sim N(\theta_r, \sigma^2)$; $r = 1, \ldots, b$; $b \geq 3$. Instead of using $Y_1, \ldots, Y_b$ to estimate $\theta_1, \ldots, \theta_b$, estimates with smaller mean squared error (MSE) can be obtained by shrinking toward a common value ν. The expected MSE is improved no matter

what the true values of the parameters. The modified Stein estimator is of the form

$$\tilde{\theta}_r = (1 - B)Y_r + B\nu\,,$$

where the shrinkage factor is

$$B = min\left[\frac{(b-2)\sigma^2}{\sum_r (Y_r - \nu)^2}, 1\right].$$

The Bayesian interpretation of the Stein estimator, $\tilde{\theta}_r$ is as follows: Suppose that *a priori* $\theta_1, \ldots, \theta_b$ are assumed to be i. i. d. $N(\nu, \tau^2)$. Assume that $Y_r|\theta_r \sim N(\theta_r, \sigma^2)$. Then θ_r can be estimated by the mean of the posterior distribution of $\theta_r|Y_r$. This estimator is of the form as the Stein estimator $\tilde{\theta}_r$ with $B = \sigma^2/(\sigma^2 + \tau^2)$.

The amount of shrinkage toward the mean of the prior depends on the relative variances of the prior τ^2 and of the observations σ^2. The Stein shrinkage factor is an estimate of the Bayes shrinkage factor. If ν and τ are unknown, an EB estimator of the form $B = min\left[\frac{(b-3)\sigma^2}{\sum_r (Y_r - \bar{Y})^2}, 1\right]$ is often used.

If the allele frequency distribution is uniform, that is if we want alleles to appear equally likely, then an EB approach is to shrink toward the average allele frequency $(1/b)$. Since many of the allele frequencies are zero, shrinking toward the average frequency would give biased estimates. To remedy this, Devlin et al. (1991) suggest a commonly employed technique. Incorporate a covariate h into model and then shrink toward a function ν_h of the covariate (Berger, 1985). The covariate used here is the number of repeats r. The estimate of ν_r is a weighted average of $\hat{\pi}_r$ and the neighboring estimates :$\ldots \hat{\pi}_{r-2}, \hat{\pi}_{r-1}, \hat{\pi}_{r+1}, \ldots$ (Prakasa Rao, 1983). Consider a weight

function (Kernel) K_{ri}, where i is the index of $\hat{\pi}_i$, the neighbour that is being weighted to obtain the regression estimate at r. The sum of the kernel function over i must be one. The nonparametric regression estimate at r is

$$\nu_r^* = K_{rr}\hat{\pi}_r + \sum_{i \neq r} K_{ri}\hat{\pi}_i.$$

In this particular case, the sum of the probabilities has to be one, hence these estimates are normalized as $\nu_r = \nu_r^* / \sum_i \nu_i^*$.

Now what remains is to estimate the kernel function. It is clear that these weights should decrease as $|i-r|$ increases. Therefore the weights should depend upon on the distance at which misclassification would be likely to occur. A normal kernel with variance σ_i^2 and mean $i\rho$ evaluated at $r\rho$ is

$$K_{ri}^* = (\sqrt{2\pi}\sigma_i)^{-1} exp\{-1/2[(r\rho - i\rho)/\sigma_i]^2\}.$$

The sum of the kernel functions must be one, hence they can be standardized as $K_{ri} = K_{ri}^* / \sum_i K_{ri}^*, i = 1, \ldots, b$. To make the weights depend upon misclassification error the SD is taken as $\sigma_i = c\ (u+i\rho),$. If σ_i is small relative to $|i-r|\rho$, then for $i \neq r$, K_{rr} will be substantially larger than K_{ri} and very little smoothing will occur, because in this case little misclassification has occurred. If σ_i is large relative to $|i-r|\rho$, then K_{ri} will not be small and substantial smoothing will occur, because misclassification occurs within this larger neighborhood. The empirical Bayes estimates are obtained using the above described EB procedure. Let

$$\tilde{\pi}_r^* = (1 - B_r)\hat{\pi}_r + B_r\nu_r,$$

where

$$B_r = min\left[\frac{bs_r^2}{\sum_i(\hat{\pi}_i - \nu_i)^2}, 1\right],$$

where s_r^2 is the bootstrap estimate of the variance of $\hat{\pi}_r$. The amount of shrinkage toward ν_r depends on the variance of the maximum likelihood estimate. The greater the variance, the greater the shrinkage. Because the linear combination depends on r, these estimates are normalized, so that they sum to one. So EB estimates are $\tilde{\pi}_r = \tilde{\pi}_r^*/\sum_i \tilde{\pi}_i^*$. The estimates obtained by EB have a smaller mean squared error than ML estimates.

2.6 Fit of the model

The Pearson χ^2 Goodness-of-fit test should not be used to test the fit of the model in this case because such a large number of parameters are estimated in addition to the coalescence probabilities. Devlin, Risch and Roeder (1991), suggest a Monte Carlo test of Goodness-of-fit. The test statistic is obtained as follows.

The support of the data is divided into k intervals say $\Gamma_1, \ldots, \Gamma_k$. The observed frequency O_k for $k-$th class is calculated by counting the number of unpaired fragments x, y or z which fall in intervals Γ_k. The Expected frequency E_k is calculated from the estimated model as

$$E_k = \int_{\Gamma_k} f(.|\hat{u}, \hat{\pi}, \hat{\phi})dx.$$

The test statistic is

$$T = \sum_k \frac{(O_k - E_k)^2}{E_k}.$$

In Chapter 3 the test statistic is computed using S data sets generated from $f(.|\hat{u}, \hat{\pi}, \hat{\phi})$, each set having n randomly paired observations, where n is the number of observations in the data set. The data is coalesced using the estimated coalescence probability $\delta(t, z)$ and a test statistic T_i is obtained for each set. Goodness-of-fit is measured by a Monte Carlo P-value, which is the proportion of $T_i's$ larger than the observed test statistic T for the actual data.

2.7 Estimation of variance of the gene frequency estimates

Variance estimates of π can be obtained by estimating the asymptotic covariance for ML estimates as the inverse of the information matrix. In our case this is not feasible because π is very large, typically of size 61 to 2400. One way to get the variance estimates is by parametric bootstrap (Efron, 1982). First B data sets are generated from the estimated distribution $f(\cdot|\hat{\pi}, \hat{u})$. Each data set has $2n$ unpaired observations that are paired randomly and then coalesced based on $\delta(t, z)$. π is estimated for each data set and the variance of π is estimated by the sample variances of the B estimates of π. Ideally, for accuracy we desire that B is large however due to computational time, it is not possible to have a very large B. The suggested value of B is between

50 and 100. If the populations are not in Hardy-Weinberg equilibrium, ordinary bootstrap resampling on the paired data could not be used.

Chapter 3

Results

In this chapter we review the results of the analyses of the Lifecodes and the OCSD data. The chapter consists of 3 sections. The first section provides the results for locus D17S79 and D2S44 for LC data. Section 2 discusses the results for each locus in the OCSD database. Section 3 gives some information on independence of alleles within and between loci.

3.1 Allele Frequency Distribution Results : LC

For LC, data is available on two VNTR loci, D17S79 and D2S44 for Caucasions only. This section provides the results for these two loci. A descriptive summary of the data set for LC locus D17S79 follows.

1. D17S79

Descriptive summary

Number of individuals : 3102	Repeat Size : 0.038 Kb
Minimum fragment size : 2.06 Kb	Maximum fragment size : 5.59 Kb
Average length : 3.54 Kb	Estimated number of alleles : 93
SD = $ca_r = 0.00575a_r$	SD of average length : 0.0203 Kb

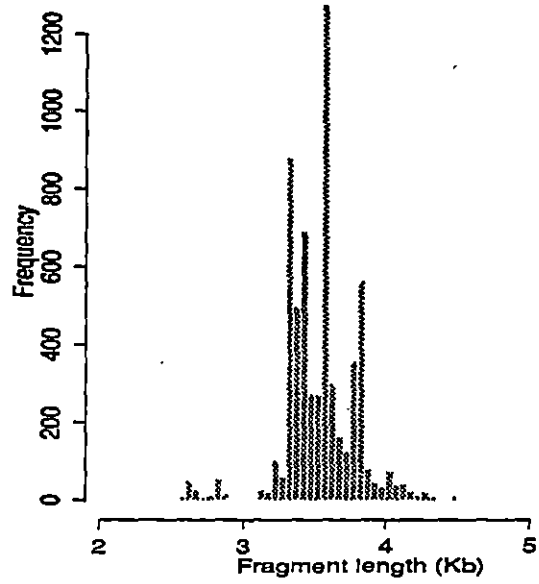


Figure 3.1: Frequency distribution of 6204 unpaired restriction fragments of D17S79.

Figure 3.1 is a frequency histogram of the fragment lengths for D17S79. The graph does not show a repeating pattern as for example displayed by Figure 2.1, therefore we conclude that the alleles are associated with a single flanking region. Most of the

fragments have lengths between 3.0 - 4.5 Kb and the maximum frequency is observed at 3.8 Kb. If coalescence is not considered, then 95% of the time the phenotype is expected to lie within $1.05 = \frac{1.96 \times 0.0203}{0.038}$ repeats on either side of the genotype for an allele of average length, 3.54 Kb.

Fixed-bin Results

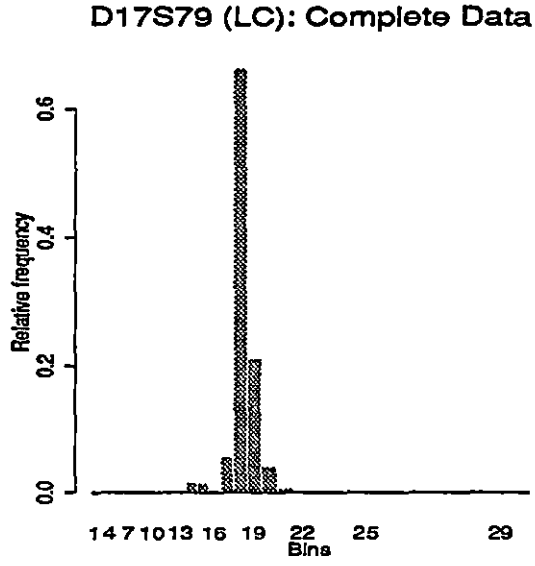


Figure 3.2: Plot of bin relative frequencies for D17S79 population.

In Figure 3.2, the numbers on the x-axis denote the bin numbers, where the bin boundaries are determined from standard markers given in Figure 1.9. Thirty bins are created from the given 30 bands sizes. All loci have minimum fragment length greater than 0.640 Kb, hence the first two bins are combined together, which reduces the number of bins to 29. Note that the bins are of unequal length. The relative frequencies obtained by the fixed-bin method gives us the shape of the distribution.

From Figure 3.2, note that only 8 bins have observable events. Most of the fragments are distributed in the range 2.351–4.323 Kb and the highest frequency is observed for bin 17, that is for the range 3.329–3.674 Kb.

Estimation of allele distribution

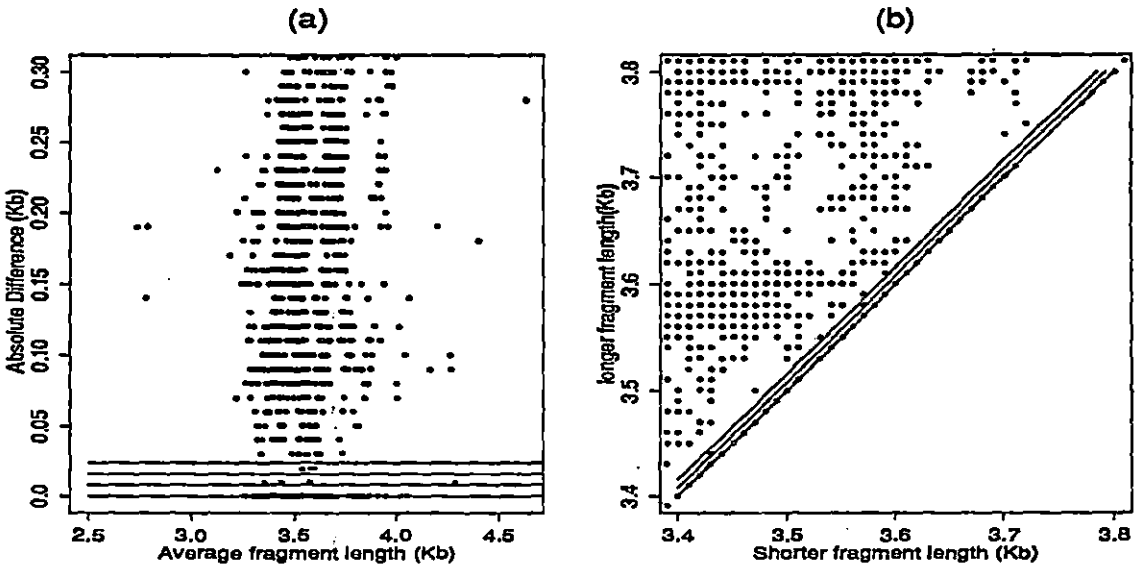


Figure 3.3: (a) Plot of the absolute difference of fragment lengths against the average length. (b) Plot of the shorter fragment length against the longer fragment length.

As suggested in section 2.1.2, we investigate the degree of coalescence through graphs of the length of larger segments against the length of smaller segments or the absolute value of the fragment pair difference against the average fragment pair length. From both the plots in Figure 3.3 coalescence is clearly observed. In Figure 3.3(a), observations with smaller differences are missing and in Figure 3.3(b) the number of points in the adjacent intervals of the line of equal length, that is $y = x$ is substantially less than the number of points on the line $y = x$. Due to the presence of

coalescence and the irregular patterns observed in Figure 3.1 it is difficult to comment upon the frequency of occurrence of any particular allele.

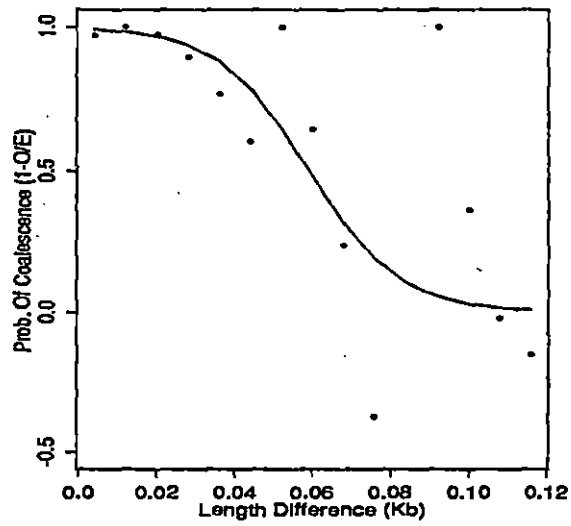


Figure 3.4: Plot of probability coalescence as a logistic function of the difference in fragment length

Using the method of Devlin et al. (1990), the probability of coalescence for a given difference is estimated for this locus as follows. The range of the difference t is divided into intervals of length 0.008 Kb. Observed(O_k) and Expected(E_k) frequencies of heterozygotes for each class interval k are obtained. When the points $[1 - \frac{O_k}{E_k}]$ versus the midpoint for each interval are plotted in Figure 3.4, the points roughly follow a decreasing sigmoidal curve. The logistic model with parameters β_0 and β_1 is fitted, that is for the k-th interval

$$\frac{O_k}{E_k} = f(c_k, \beta_0, \beta_1) = \frac{exp(\beta_0 + \beta_1 c_k)}{[1 + exp(\beta_0 + \beta_1 c_k)]}.$$

The probability of coalescence is estimated as $[1 - f(t, \beta_0, \beta_1)]$. Devlin et al. (1990)

computed the probability of coalescence for 1399 Caucasian individuals from LC. They estimated $\hat{\beta}_0 = -10.24$ and $\hat{\beta}_1 = 156.14$. Since we do not have the classification scheme of LC data, we used the entire data set of 3102 paired lengths for estimation. The logistic parameters for D17S79 are obtained as $\hat{\beta}_0 = -5.08$ and $\hat{\beta}_1 = 85.87$.

The results of estimating the allele distribution using different methods are summarized as follows.

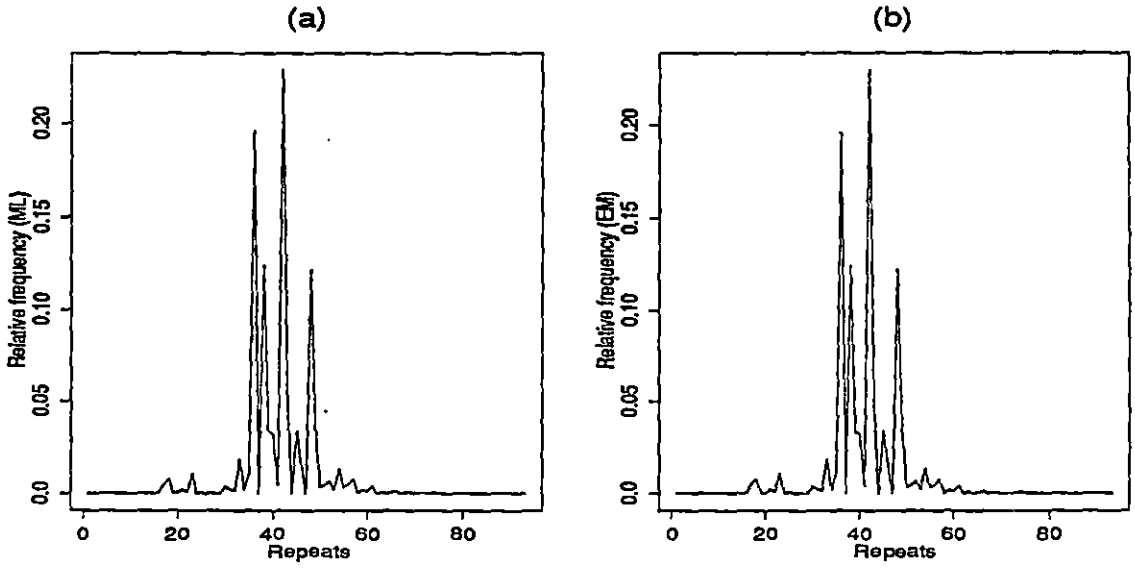


Figure 3.5: (a) Plot of the estimates of allele distribution against the number of repeats, for SQPCOAL(ML). (b) Plot of the estimates of allele distribution against the repeats for EMCOAL(EM).

The estimates of allele frequencies π and flanking-region size u are obtained using SQPCOAL under the restrictions, $\sum_{r=1}^{b} \pi_r = 1$, $0 \leq \pi_r \leq 1$ and $0 \leq u \leq 2.022$, where the upper limit of u is obtained as the minimum length observed minus the repeat length, that is $2.022 = 2.06 - 0.038$. The flanking-region size is estimated as 1.979 Kb and Figure 3.5(a) shows the graph of $\hat{\pi}$ versus the number of repeats, r.

Estimates of π and u were also obtained using EMCOAL. The estimated flanking-region size is 1.978 Kb and the estimate of π is the same as that obtained by SQP-COAL. The sensitivity of the algorithm towards the initial value of u, was checked using different starting values of u. Different values of u, resulted in the same final estimates of u and π. The plot of allele frequencies against the number of repeats show the same pattern for both sets of estimates, see Figures 3.5(a) and 3.5(b). The plot of the estimates of the allele frequencies follows the same pattern of the observed data frequency histogram given in Figure 3.1. Devlin et al. (1991) estimated the flanking-region size as 1.953 Kb and their plot of the allele frequencies against the number of repeats shows a similar pattern as that shown in Figures 3.5(a) and 3.5(b).

We did not check the fit of the model for this data because of the amount of computation required. Due to the large sample size, it took approximately 6 days to compute estimates of π and u, making a Monto-Carlo goodness-of-fit test impractical. Devlin et al. (1991) concluded that this model fits the data on D17S79(LC) on the basis of a Monto-Carlo test of goodness-of-fit (p-value=.09).

Morris et al. (1990) suggests the following as a method to obtain an estimate of the resolution threshold α. They graph the percentage difference of each pair of alleles, $s_i = 100\left(\frac{\max(x_i,y_i)-\min(x_i,y_i)}{z_i}\right)$ versus the average pair fragment length $z_i = \frac{x_i+y_i}{2}$. In their graphs, there is an interval of values, s, where few or no points are observed. They use a value for α which falls at the upper end if this interval. For example see Figure 3.12(a).

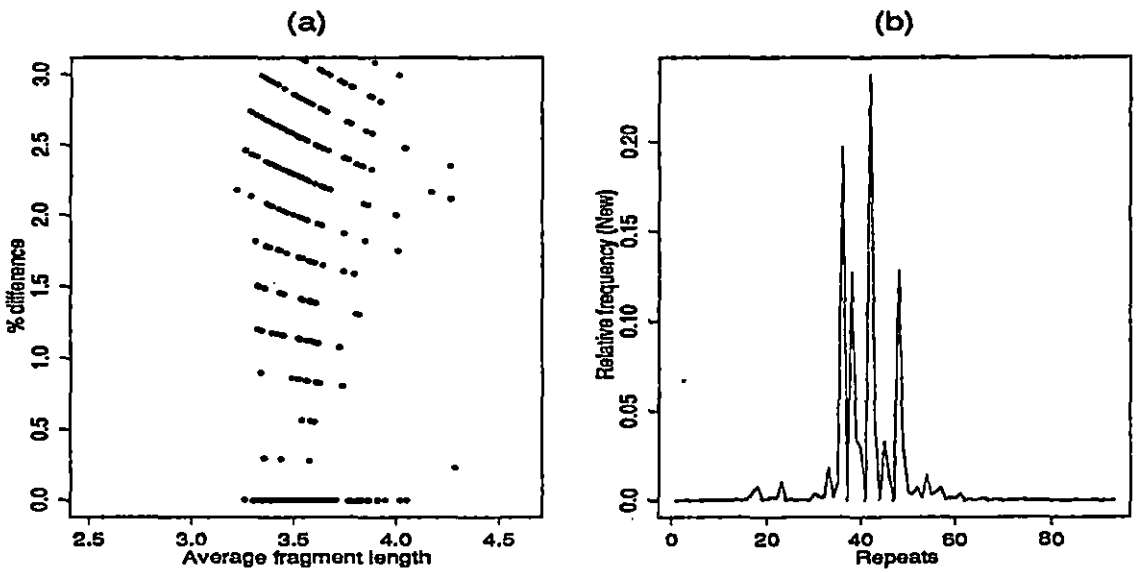


Figure 3.6: (a) Plot of the difference between fragments as a percentage of the average fragment length. (b) Plot of the estimates of allele distribution against the repeats for EMRESL.

The graph in Figure 3.6(a) is a plot of s versus the average fragment length. For the available data LC has used for LC D17S79. Since LC's fragment length measuring device, digipad does not measure the actual base-pairs. It rounds the length to 10 bp. Hence in Figure 3.6(a), band patterns separated by 10 bp are clearly observed. This graph is not consistent with those of Morris et al. (1990), and one cannot clearly pick off a value α from the graph. Alternatively, we use another technique also given in Morris et al. (1990). For a set of values $\alpha = 0.8\%, 0.9\%, 1.0\%$, we estimate the heterozygosity(h) as:

$$h_\alpha = \frac{number\ of\ pairs\ with\ s > \alpha}{total\ number\ of\ pairs}.$$

We use the largest value α such that $h_\alpha \cong h_0$, where h_0 is the observed percentage of

heterozygotes. For this set of data, we find that for $\alpha = 0.8\%$, $h_\alpha = 83.7 \cong h_0 = 83.8$, and h_α is smaller for the other values, $\alpha = 0.9\%$ and 1.0%. Therefore, we use 0.8% as our resolution threshold.

Using SQPRESL and EMRESL, the estimate of the flanking-region size and π are obtained. For both methods SQPRESL and EMRESL, the estimate of u is 1.979 Kb and the estimate of π is shown in Figure 3.6(b). The allele frequency estimates obtained by SQPCOAL and SQPRESL or EMRESL are the same.

2. D2S44

The following is the descriptive summary of the data set for LC D2S44 locus.

Descriptive summary

Number of individuals : 3116	Repeat Size : 0.031 Kb
Minimum fragment size : 6.87 Kb	Maximum fragment size : 20.94 Kb
Average length : 11.17 Kb	Estimated number of alleles : 454
SD $= ca_r = 0.00575a_r$	SD of average length : 0.0642 Kb

The frequency distribution in Figure 3.7, clearly displays a repeating pattern, which suggests that the measured fragments are associated with two flanking-regions. Two observed repeated patterns are nonoverlapping. Hence fragments could be divided into two groups and then the mixture model with single flanking region could be used to estimate the allele distribution. As observed in Figure 3.7, two groups

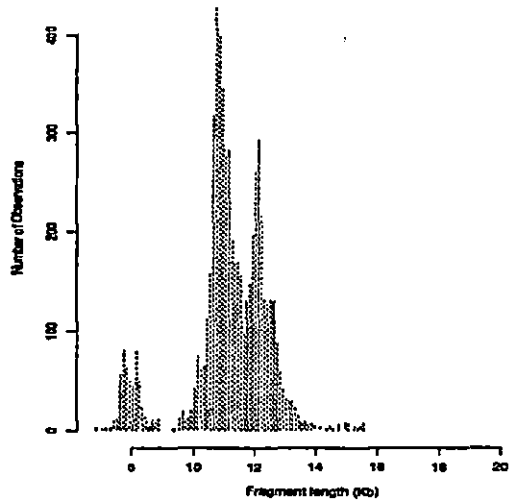


Figure 3.7: Frequency distribution of 6232 unpaired restriction fragments of D2S44 (LC).

can be made according to the fragment length less than 8.97 Kb or greater than 9.14 Kb. While grouping the data, it is observed that 526 individuals have one or both fragment sizes less than 8.97 Kb of which 495 individuals have one band greater than 8.97. This difficulty in classification makes the estimation of allele distribution more difficult, and further analysis is not performed on the locus D2S44 of LC.

Devlin et al. (1991) tried to estimate the frequency distribution using the EM algorithm for two flanking regions. They observed that the fit of the model was poor, and they concluded that the flanking and repeat distributions are not independent. They observed that D2S44 shows a polymorphism in the flanking region when it is excised with *PstI*. After inspecting supplementary data for sequences cut with *HaeIII*, they were able to split the data in two distinct groups corresponding to

two distinct flanking region sizes. They calculated the proportion of data in the two groups as 0.097 and 0.903 respectively. Using the EM algorithm they estimated π and u_1 and u_2 for the separated data sets. Their estimate for u_1, u_2 were 6.701 and 9.331 respectively. From the goodness-of-fit P-value=0.28, they concluded that the model fits the data.

3.2 Allele Frequency Distribution Results : OCSD

The following subsections provide the frequency estimates for all 5 VNTR loci. The results of fixed-bin analyses on OCSD data are also discussed.

1. D17S79

Note that the OCSD database for D17S79 contains data only for Hispanics whereas the LC database contains data for Caucasions only. The following is the descriptive summary of the data set for OCSD D17S79 locus.

Descriptive Summary

Number of individuals : 195	Repeat Size : 0.038 Kb
Minimum fragment size : 0.974 Kb	Maximum fragment size : 3.277 Kb
Average length : 1.595 Kb	Estimated number of alleles : 61
SD $= ca_r = 0.008a_r$	SD of average length : 0.0128 Kb

From the information given by OCSD, the measurement error is 0.8% of the ob-

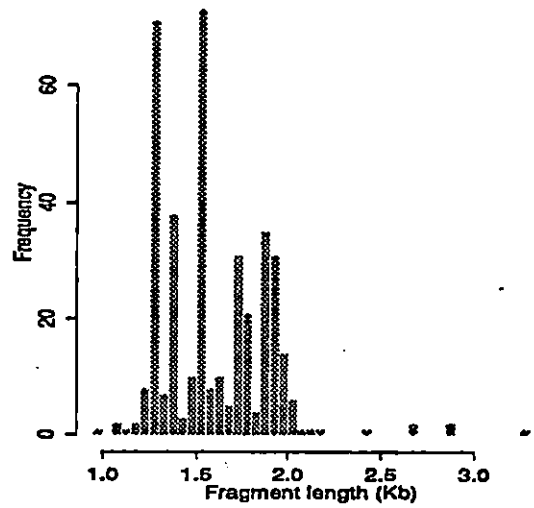


Figure 3.8: Histogram of 390 unpaired restriction fragments of D17S79.

served length. Hence 95% of the time, the phenotype is expected to lie within $0.65 = \frac{1.96 \times 0.0126}{0.038}$ repeats on either side of the genotype for an allele of average length, 1.595 Kb. The restriction enzyme used by OCSD is different than LC, hence the lengths of the fragments are small as compared to the LC D17S79 locus. The distribution in Figure 3.8, does not display a repeating pattern, which suggests that the measured fragments are associated with single regions. Most of the fragments are observed between 1.20 - 2.0 Kb and the maximum frequency is observed between 1.25 - 1.3 Kb and 1.5 - 1.6 Kb.

Fixed-bin Results

From Figure 3.9, note that only 14 bins have observable events. And out of these 14, six bins have counts below five. Most of the fragments are distributed in the range 1.197–2.088 Kb, that is, between the bins numbered 6 and 7.

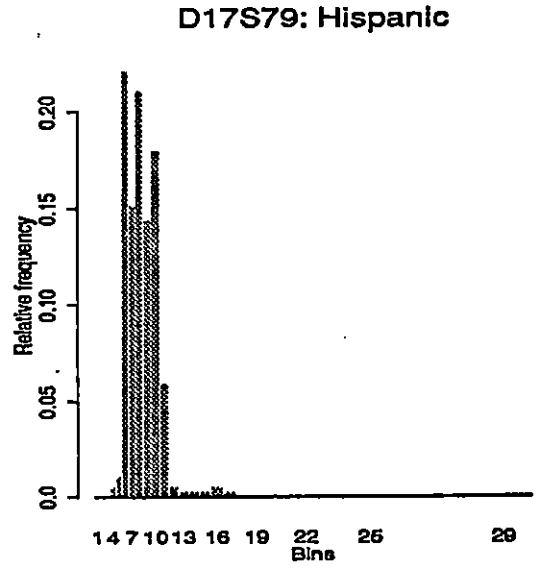


Figure 3.9: Plot of bin relative frequencies for D17S79 Hispanic population.

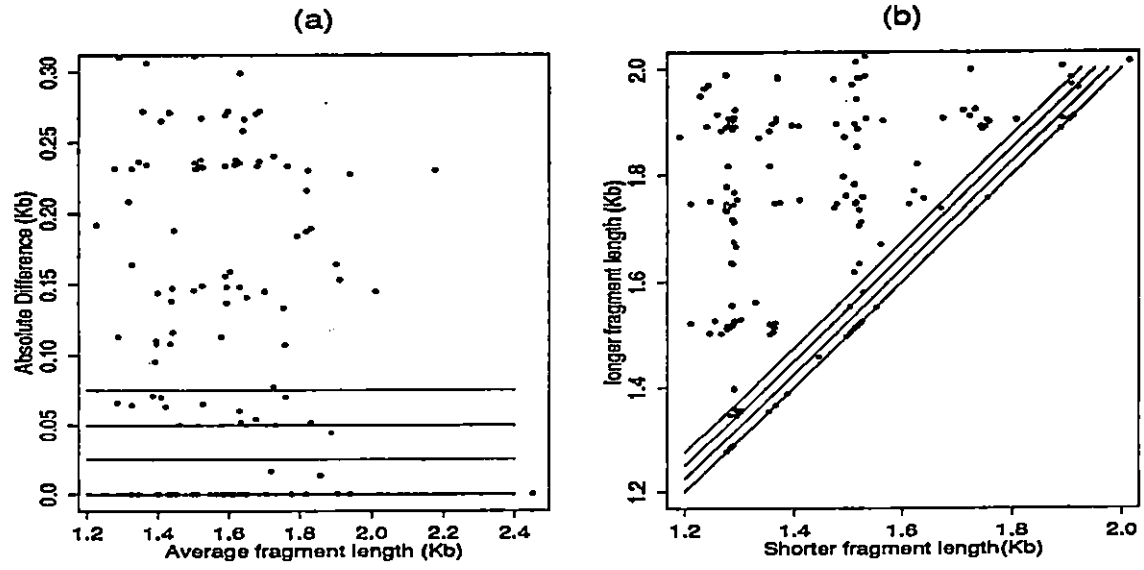


Figure 3.10: (a) Plot of the absolute difference of fragment lengths against the average length. (b) Plot of the shorter fragment length against the longer fragment length.

Estimation of allele distribution

Figures 3.10(a) and 3.10(b) suggest that some of the fragment pairs are coalesced. Note that in Figure 3.10(a) there are more observations with zero differences, on the zero-line as compared to adjacent intervals. This phenomenon is also present in Figure 3.10(b) where many points fall on the $y = x$ line, and fewer in the adjacent interval.

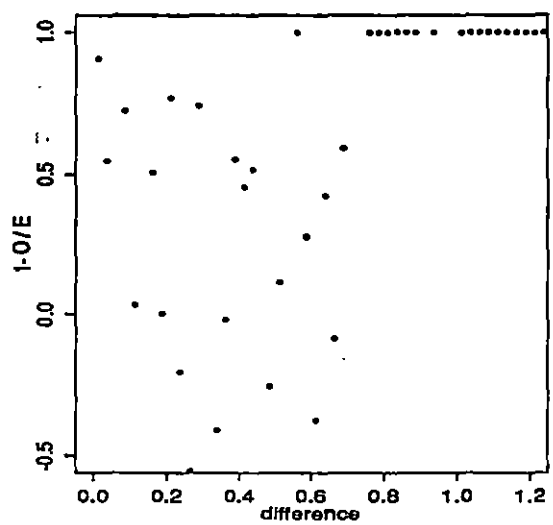


Figure 3.11: Plot of probability of coalescence as a function of the difference in fragment length

The method of estimating the probability of coalescence given in section 2.1.3.3 could not be used here. The range of pair differences is divided into intervals of length 0.025 Kb. As the sample size is small, many of the observed frequencies O_k are greater than the expected frequencies E_k, hence it is not possible to fit a reasonable function to the points on the plot $[1 - \frac{O_k}{E_k}]$ versus c_k in Figure 3.11. We tried the same

plots with different intervals of length 0.01 Kb and 0.04 Kb, but the plots were not much different than that shown in Figure 3.11. Therefore algorithms SQCOAL and EMCOAL are not used for estimating the allele distribution.

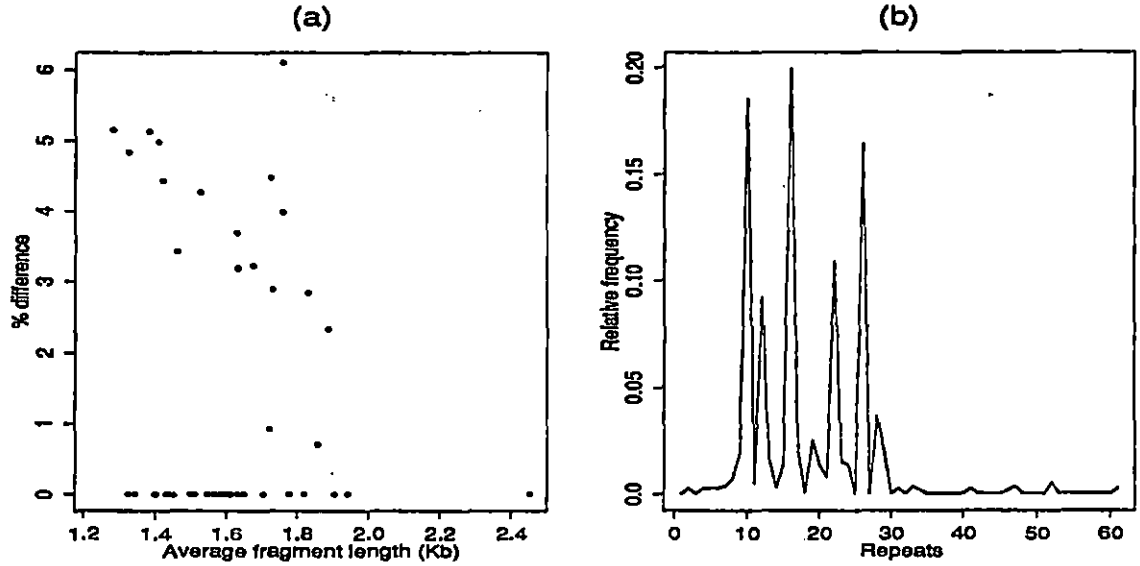


Figure 3.12: (a)Plot of the difference between fragments as a percentage of the average fragment length. (b) Plot of the allele frequency estimates against the number of repeats for EMRESL.

If the distance between two bands is not enough to distinguish them from each other, then they are identified as a single band. If a single band is observed, then the size of this band is almost the same as its average. Most of the labs set the resolution threshold, the minimum distance required for identification of two bands, as some percentage of the molecular weight of the band observed. This resolution threshold can be estimated, by plotting the percentage differences s as function of the average (Morris et al., 1992). Figure 3.12 gives the approximate idea of the minimum resolution distance α. For D17S79 this distance is observed to be around 2.7% of the

molecular weight.

The estimate of the flanking-region size is 0.9064 Kb by both algorithms SQPRESL and EMRESL. The plot of the allele distribution estimated using EMRESL is given in 3.12(b), and SQPRESL is given in Figure 3.13(a). The plots of the estimated allele distribution by both algorithms follow the same pattern as that of frequency histogram given in 3.8. Devlin et al. (1991) suggest that the estimated allele distribution can be smoothed by empirical Bayes(EB) method. Figure 3.14(a) shows both SQPRESL(ML) estimates together with the EB estimates. To get a better resolution of the graph, square-root transformed estimates of SQPRESL(ML) and EB are plotted in Figure 3.14(b). There is not much difference between the two estimates.

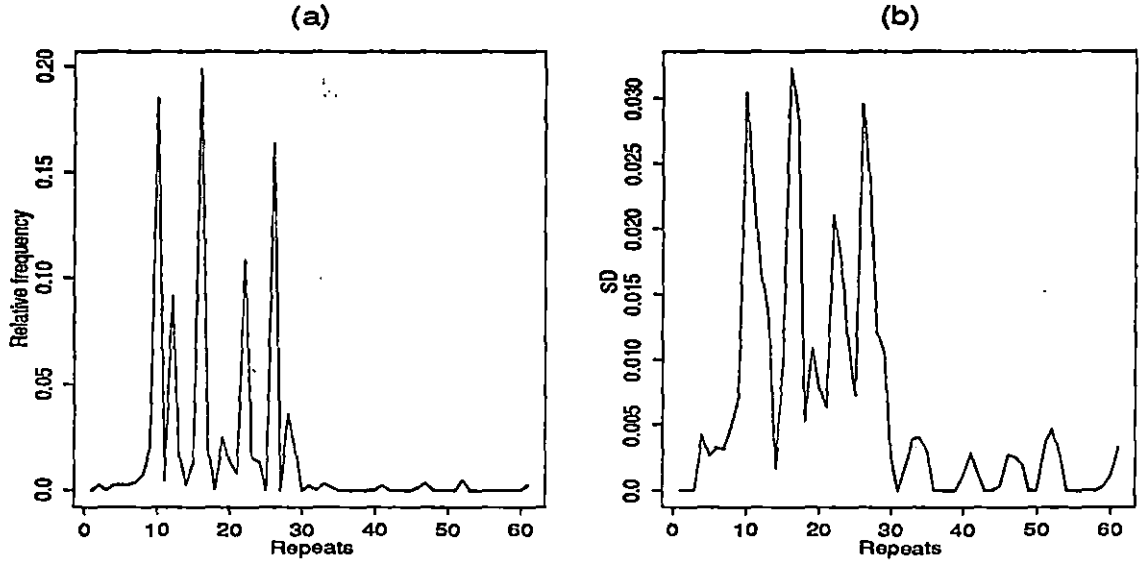


Figure 3.13: (a) Plot of the estimates of allele distribution against the repeats for SQPRESL. (b) Plot of Standard Deviation of estimates of allele distribution against repeats.

Parametric bootstrap is used to estimate the variance of the estimates. Fifty

data sets are generated from the estimated distribution $f(.|\hat{\pi}, \hat{u})$. For each data set bootstrap estimates of u and π are obtained and the variance of π is estimated from the 50 bootstrap estimates of π. Figure 3.13(b) shows the plot of the bootstrap standard deviation of π.

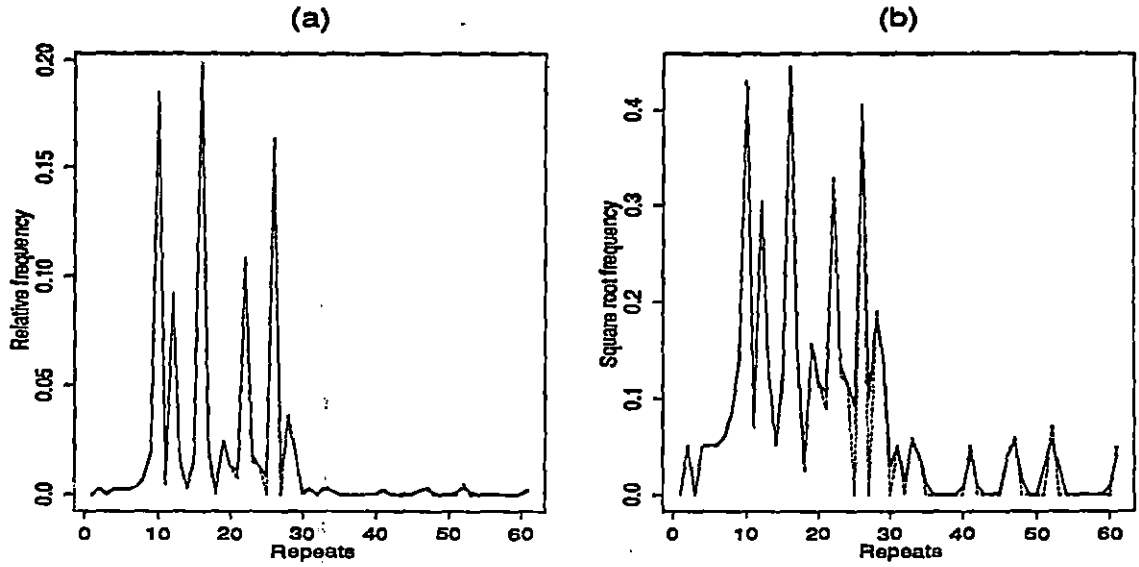


Figure 3.14: (a) Plot of the estimates of allele distribution against the repeats for empirical Bayes [solid line], and maximum likelihood [broken line]. (b) Square root transformed estimates of allele distribution for empirical Bayes [solid line], and maximum likelihood [broken line].

The fit of the model is checked using the Monto-Carlo goodness-of-fit test discussed earlier. 300 data sets are generated from $f(.|\hat{\pi}, \hat{u})$. This model fits the data very well since the P-value is 0.45. The shapes of the estimated allele distribution shown in Figure 3.13(a) and that given by Devlin et al. (1991) are almost the same.

In the fixed-bin analyses of D17S79 the maximum frequency of 0.2 is observed for the range 1.508–1.637 Kb and 1.197–1.352 Kb. By SQPCOAL, the maximum relative frequency 0.199 is estimated for 16 repeats that is for an allele of size 1.5144 Kb [.9064 + (16)(.038)]. The second largest relative frequency estimated is 0.185 for 10 repeats that is for an allele of size 1.2864 Kb. This shows that fixed-bin analyses provide the general idea of distribution but that the relative frequencies are overestimated.

3. D2S44

The descriptive summary of the data set for locus D2S44 is as follows.

Descriptive Summary

Number of individuals : 1236	Repeat Size : 0.031 Kb
Minimum fragment size : 0.649 Kb	Maximum fragment size : 6.693 Kb
Average length : 2.001 Kb	Estimated number of alleles : 195
SD = $ca_r = 0.008a_r$	SD of average length : 0.016 Kb

From the histogram of unpaired fragments in Figure 3.15(a), most of the fragments are observed between 0.64 - 4 Kb and the maximum frequency is observed between 1.4 - 1.6 Kb. No repeating pattern of the fragments is observed, which suggests that the measured fragments are associated with only one flanking-region. If coalescence is ignored, 95% of time the phenotypes are expected to lie within $1.011 = \frac{1.96 \times 0.016}{0.031}$ repeats on either side of the genotypes for an allele of average length, 2.001 Kb.

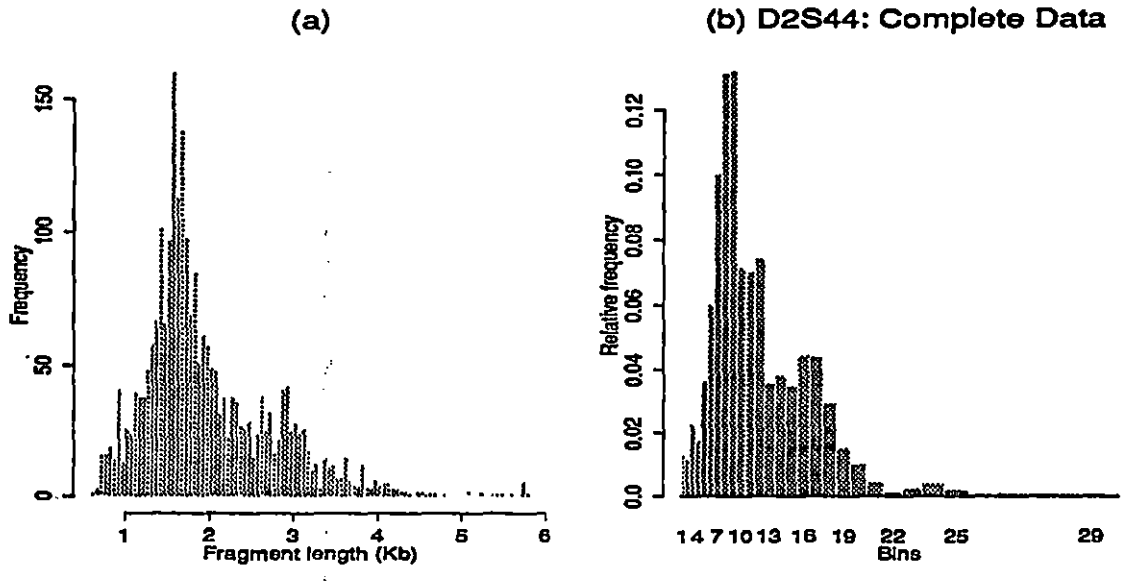


Figure 3.15: (a) Frequency distribution of 2472 unpaired restriction fragments of D2S44. (b) Plot of bin relative frequencies for D2S44 Complete data(n=1236)

Fixed-bin Results

From Figure 3.15(b), 26 bins contain observable events. The relative frequency of observing the fragments in the range 1.197–2.351 is high as compared to other ranges. The highest frequency of 0.13 is observed for bins 8 and 9, that is for the range, 1.508–1.637 Kb and 1.638–1.788 Kb. Fragments with length greater than 3.980 Kb have low chances of occurrence. The fixed-binning method is also applied to three ethnic groups, the Hispanic, Black and Caucasian populations. This analysis helps us to determine whether the distribution of alleles shows substantial variation among the three ethnic groups. Figure 3.16(a-d) gives the distribution of alleles in four subgroups. It is observed that the low frequency as well as high frequency bins are consistent Hispanic, Black and Caucasian populations. Since the Asian group consists

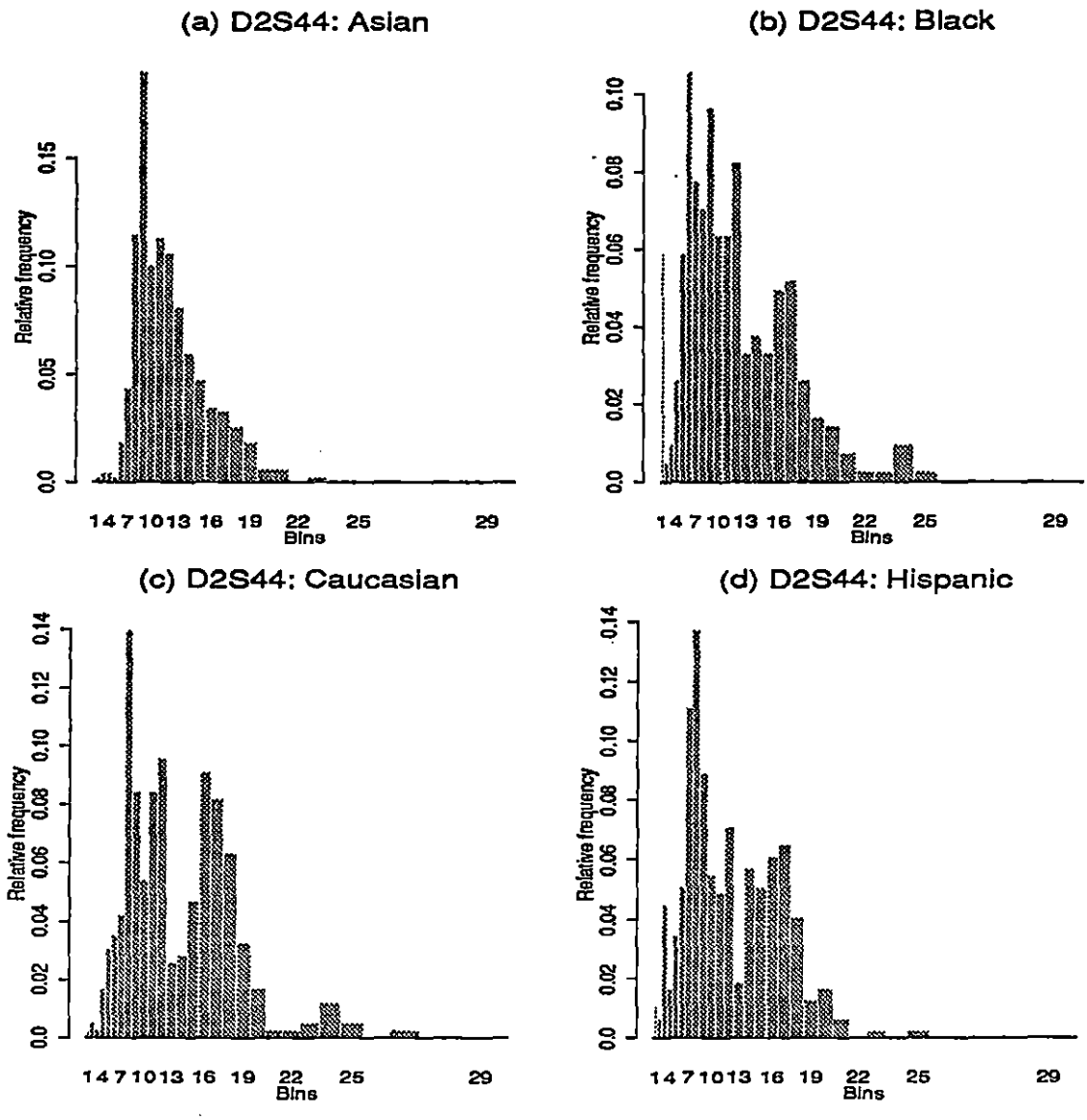


Figure 3.16: Plot of bin relative frequencies for Asian(n=560), Black(n=213), Caucasian(n=215), and Hispanic(n=248) individuals.

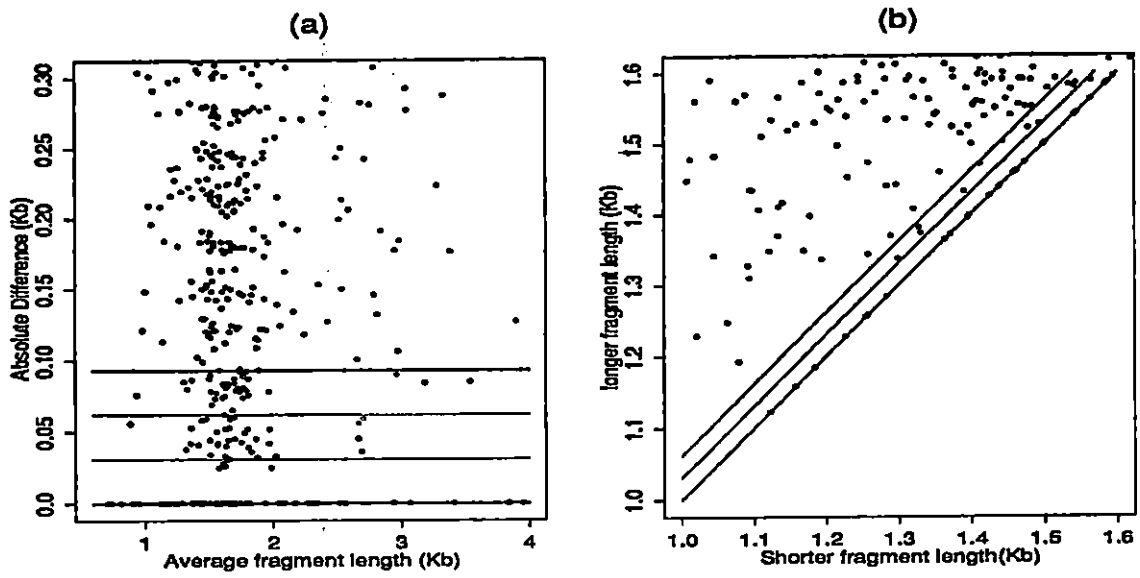


Figure 3.17: (a) Plot of the absolute difference of fragment lengths against the average length. (b) Plot of the shorter fragment length against the longer fragment length.

of several subpopulations, the distribution of alleles is different than the Hispanics, Blacks and Caucasians.

Estimation of allele distribution

Figure 3.17(a) gives the plot of the absolute difference of fragment lengths against the average length and Figure 3.17(b) is a plot of the shorter fragment length against the longer fragment length. Interval lines in Figure 3.17(a) and 3.17(b) are of width 0.031 Kb. In Figure 3.17(a), the number of observations with zero differences is larger than the number of points observed in adjacent intervals. In Figure 3.17(b) many points are observed on the line $y = x$, and fewer points are observed in adjacent intervals. This indicates that close heterozygotes may have been measured as homozygotes.

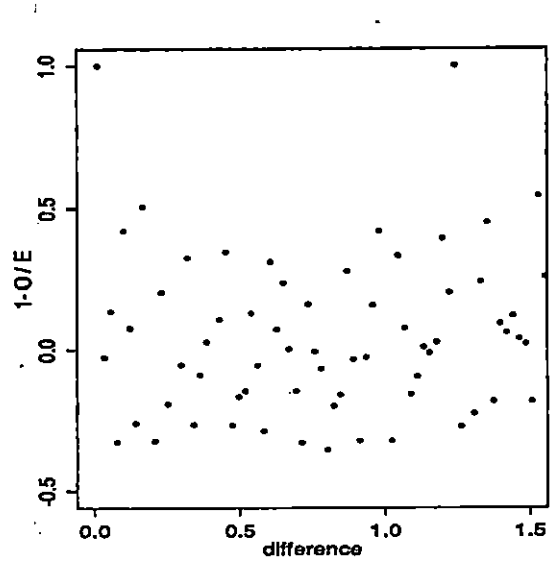


Figure 3.18: Plot of probability of coalescence as a function of the difference in fragment length

To estimate the probability of coalescence as suggested in Section 2.1.3.3, the range of the differences is divided into intervals of length 0.025 Kb. It is not possible to fit a reasonable function to the points on the plot $[1 - \frac{O_k}{E_k}]$ versus c_k in Figure 3.18. The pattern of points did not change much with intervals of length 0.016 Kb and 0.032 Kb. The probability of coalescence is not easily estimable for the given data, and we cannot use EMCOAL and SQPCOAL. Therefore we concentrated on SQPRESL and EMRESL for which we require the minimum resolution distance. From Figure 3.19(a), the minimum resolution distance is observed to be approximately 1.5% times the molecular weight. For the average length this distance is 0.04 Kb which is almost 1.2 repeats.

Using EMRESL, the flanking-region size is estimated as 0.6066 Kb, and the es-

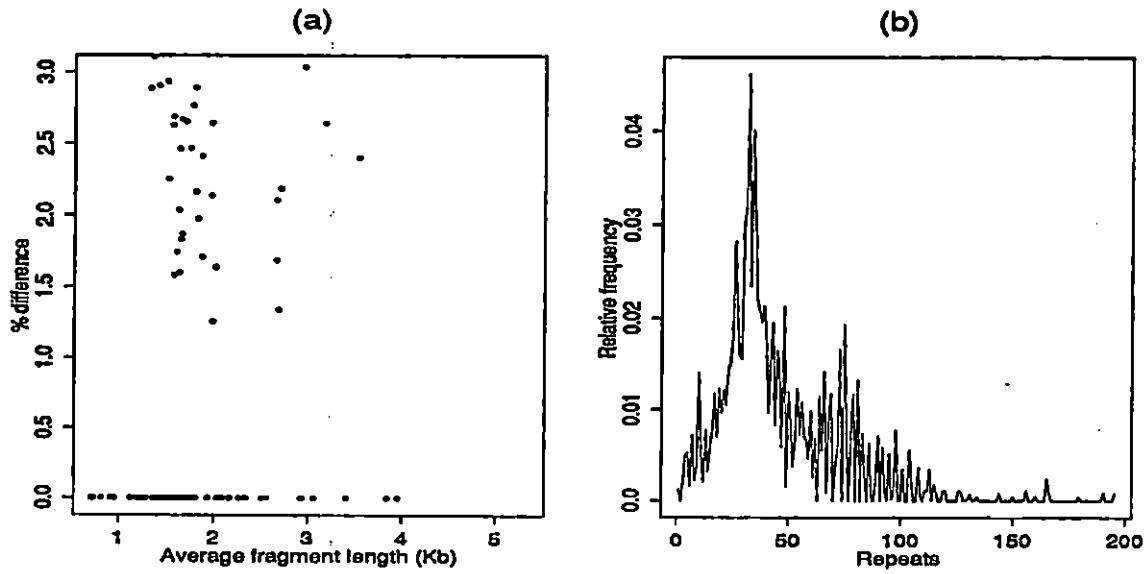


Figure 3.19: (a) Plot of the difference between fragments as a percentage of the average fragment length. (b) Plot of the allele frequency estimates against the number of repeats.

timates of π shown in Figure 3.19(b) follows the same pattern of the observed data frequency histogram in Figure 3.15. We tried to use SQPRESL, with initial starting values of ${\pi_r}^{(0)} = \frac{1}{195}, r = 1, 2, \ldots, 195$ and $u^{(0)} = 0.6Kb$. The routine could not move to improve upon the initial estimates ${\pi_r}^{(0)}, u^{(0)}$.

D4S139

The descriptive summary of the data set for locus D4S139 is as follows.

Descriptive Summary

Number of individuals : 1215	Repeat Size : 0.031 Kb
Minimum fragment size : 2.068 Kb	Maximum fragment size : 23.12Kb
Average length : 6.645 Kb	Estimated number of alleles : 679
SD $= ca_r = 0.008a_r$	SD of average length : 0.053 Kb

From the histogram of the unpaired fragment lengths shown in Figure 3.20(a) the distribution of lengths is more centered on the lower tail, and it is observed to be positively skewed. Most of the fragments are observed between 0.65 - 6 Kb. Without coalescence 95% of the time the phenotype is approximately $3.4 = \frac{1.96 \times 0.053}{0.031}$ repeats on either side of the genotype for an allele of average length 0.053 Kb.

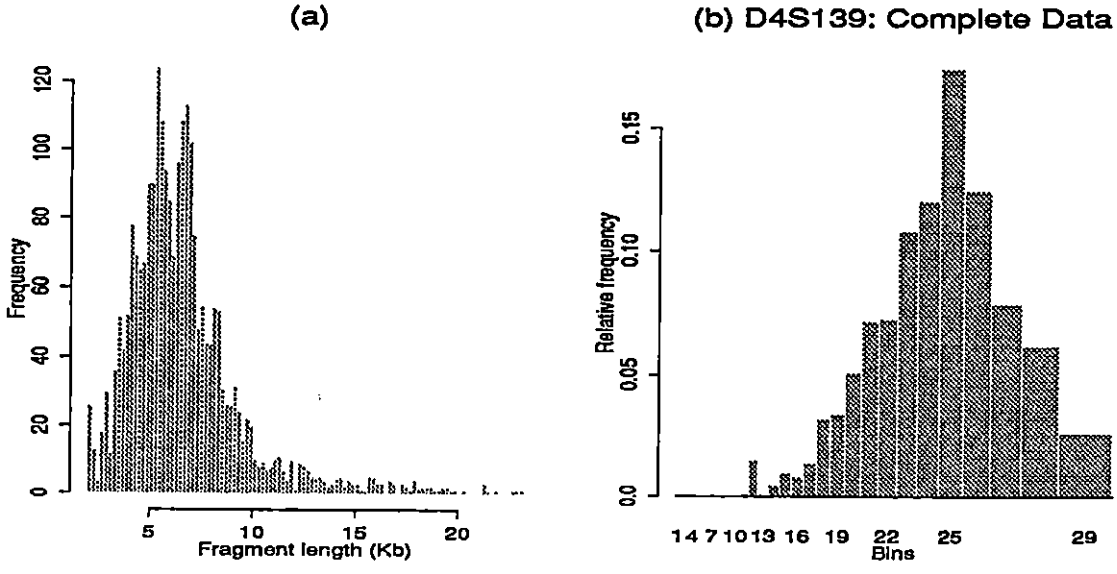


Figure 3.20: (a) Frequency distribution of 2430 unpaired restriction fragments of D4S139. (b) Plot of bin relative frequencies for D4S139 Complete data(n=1215)

Fixed-bin Results

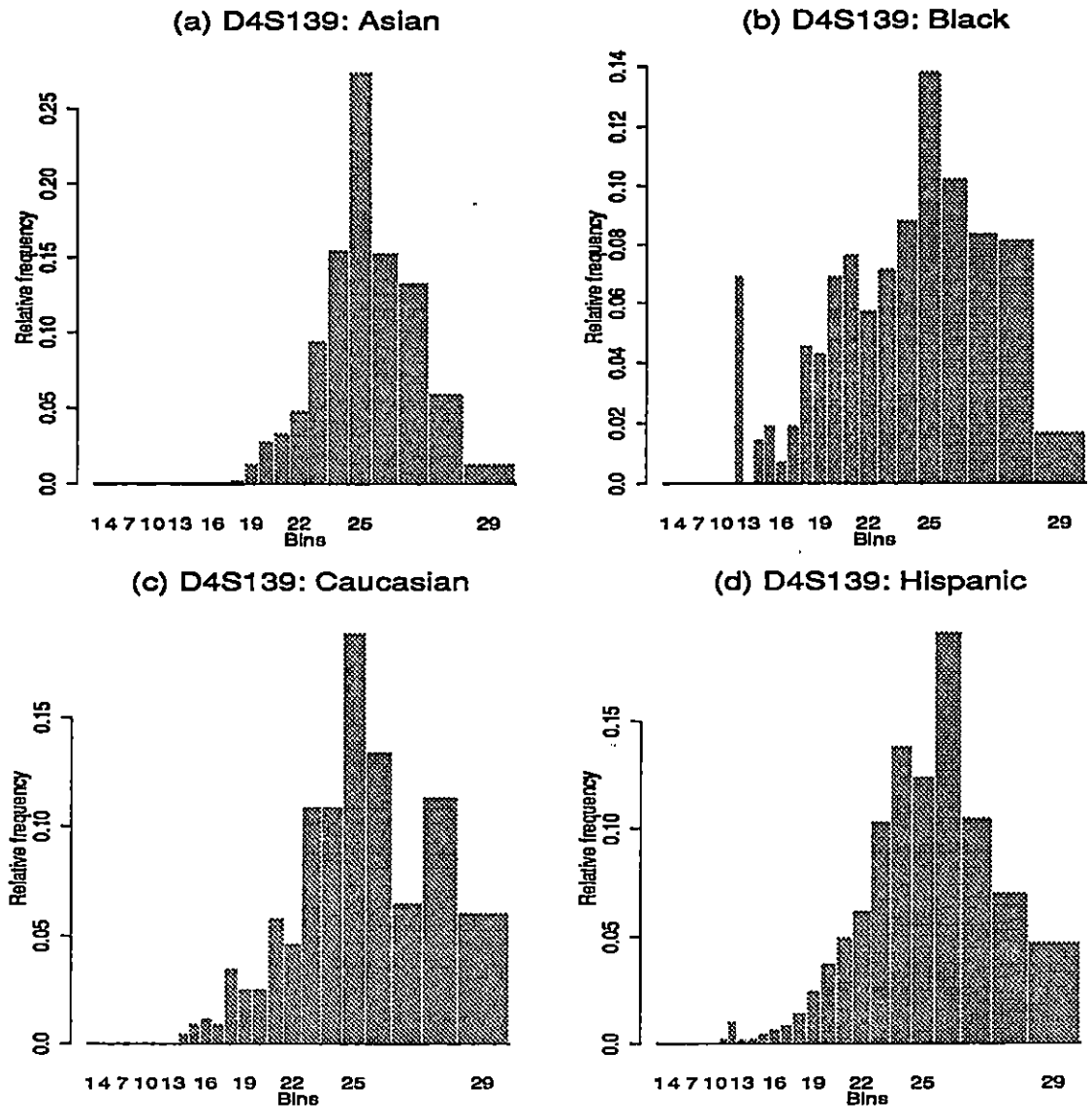


Figure 3.21: Plot of bin relative frequencies for Asian(n=1215) and Black(n=210), Caucasian(n=217), Hispanic(n=243) individuals.

Figure 3.20(b) provides the relative frequency distribution for locus D4S139. Out of 29, the first 10 bins have zero relative frequency. Like D1S7 locus, the relative frequency is observed to be high for large fragment lengths. Bins 24-27 have relative frequency greater than 0.11, and the highest frequency 0.174 is observed for the bin 26 that is for the fragment length range 6.369–7.241 Kb. The relative frequency of an allele in the range 0–3.033 Kb is quite small. Figure 3.21(a-d) gives the distribution of alleles in four subgroups. The low frequency as well as high frequency bins are consistent in all subgroups.

Estimation of allele distribution

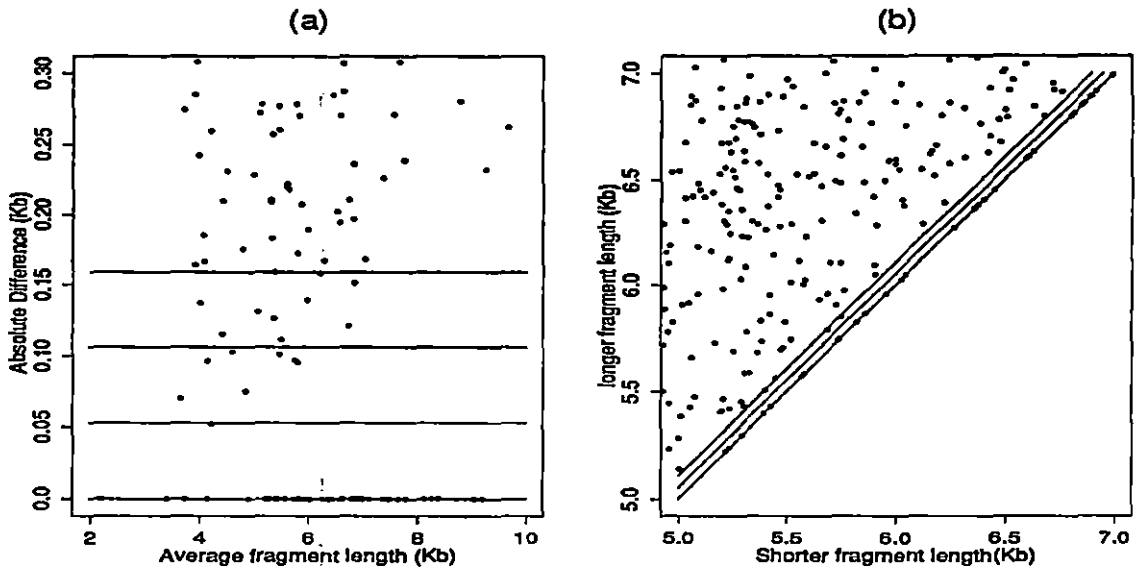


Figure 3.22: (a) Plot of the absolute difference of fragment lengths against the average length. (b) Plot of the shorter fragment length against the longer fragment length.

Both the plots in Figure 3.22 indicate the possibility of coalescence. Interval lines are of repeat length 0.031 Kb. There are more points with zero differences, on the

zero difference line as compared to the number of points observed in adjacent intervals in Figure 3.22(a). This phenomenon is also observed in Figure 3.22(b). More points are observed on the line $y = x$, and fewer points are observed in adjacent intervals. This suggests that some of the close heterozygotes are observed as homozygotes.

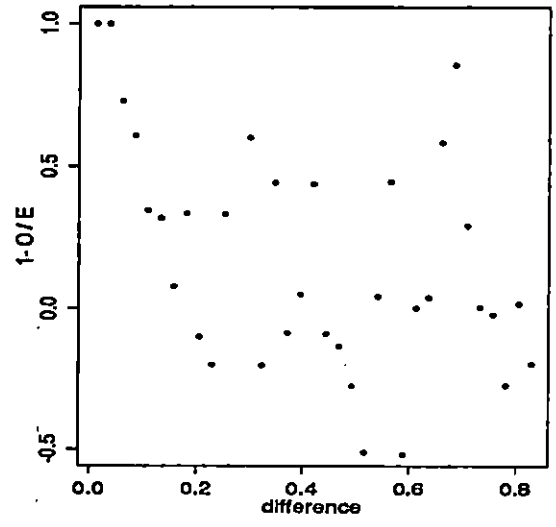


Figure 3.23: Plot of probability of coalescence as a function of the difference in fragment length

For the locus D4S139, the range of the fragment sizes and also the range of the pair differences is large. The range of the differences is divided into intervals of length 0.02 Kb. From Figure 3.23, it is observed that no reasonable function can be fitted to the points in the plot of $[1 - \frac{O_k}{E_k}]$ versus c_k. Similar plots were also tried with interval lengths 0.03 Kb and 0.05 Kb, but the plots were not much different from Figure 3.23. As it is not possible to estimate the probability of coalescence, the algorithms EMCOAL and SQPCOAL are not used for estimating the allele distribution. There-

fore we proceed with EMRESL and SQPRESL, for which we require an estimate of the minimum resolution distance. From the Figure 3.24(a) the minimum resolution distance is estimated to be 1.5% of the observed band.

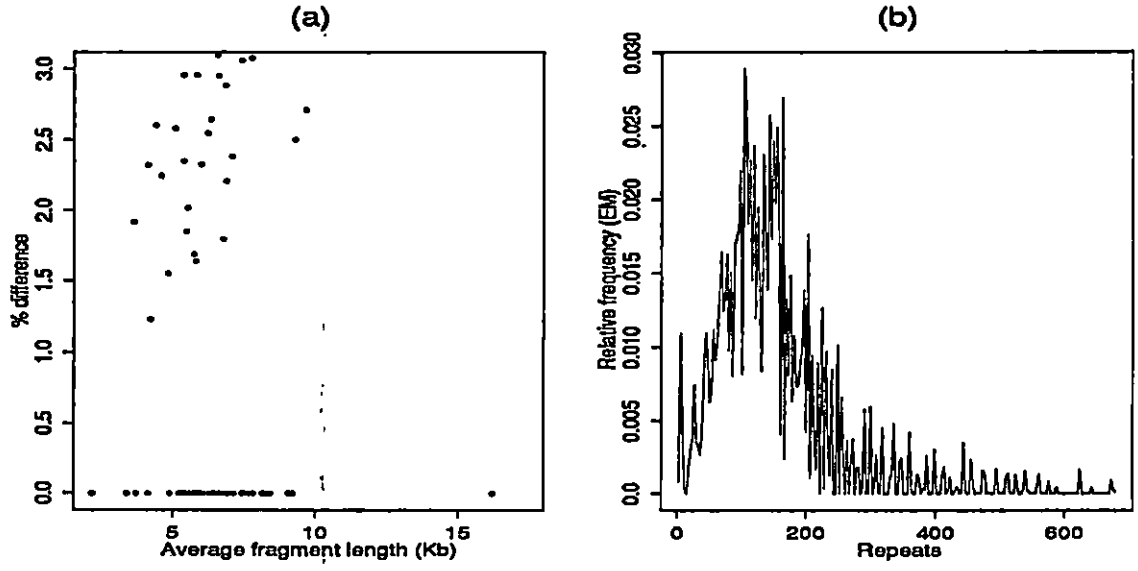


Figure 3.24: (a) Plot of the difference between fragments as a percentage of the average fragment length. (b) Plot of the allele frequency estimates against the number of repeats.

The estimated number of alleles is 679 and the sample size is 1215. Since the sample size is small as compared to the number of parameters to be estimated and most of the frequencies are zero, it may be difficult to obtain all of the 679 allele frequencies. Hence SQPRESL and EMRESL were tried with different step-sizes, $r = mj, j = 1, \ldots, \frac{679}{m}$. SQPRESL did not converge with step-size less than 5, for maximum of 400 iterations. While EMRESL gave an optimal solution of the frequency estimates with step-size three. Hence 226 allele frequency estimates, for $r = 3, 6, 9, \ldots, 678$ are obtained using EMRESL. The estimate of the flanking-region

is 2.05 Kb. Figure 3.24(b) gives the plot of the estimates of allele distribution against the number of repeats. The pattern of the estimates of frequencies is similar to that of observed in histogram of unpaired fragment lengths (Figure 3.20)

D10S28

The descriptive summary of the data set for locus D10S28 is as follows.

Descriptive Summary

Number of individuals : 1255	Repeat Size : 0.033 Kb
Minimum fragment size : 0.651 Kb	Maximum fragment size : 11.58 Kb
Average length : 2.372 Kb	Estimated number of alleles : 330
SD $= ca_r = 0.008a_r$	SD of average length : 0.0189 Kb

Figure 3.25(a) does not show any repeating pattern of the fragment lengths which indicates that the fragments are associated with a single flanking region. The distribution of fragment length seems to be positively skewed. Most of the fragments are observed between 0.64 - 5 Kb. The maximum frequency is observed between 1.4 - 1.6 Kb. Without coalescence, for an allele of average length, 2.372 Kb the phenotype is expected to overlap $1.13 = \frac{1.96 \times 0.0189}{0.033}$ repeats on either side of the genotype 95% of the time.

Fixed-bin Results

The allele frequency distribution for locus D10S28 is given in Figure 3.25(b). The relative frequency of observing a length greater than 15 Kb is zero. Fragments with

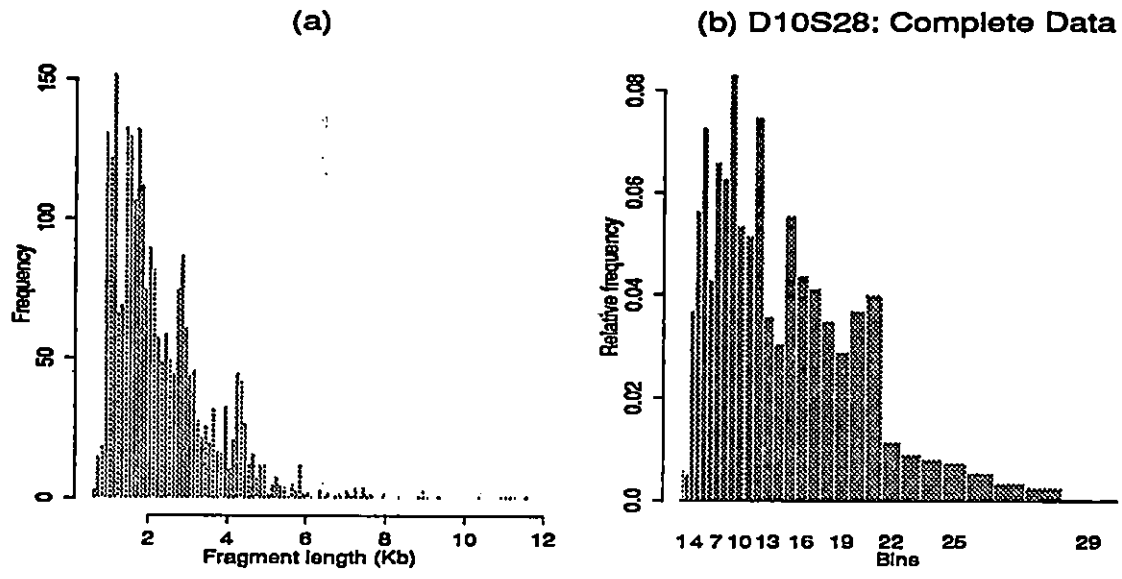


Figure 3.25: (a) Frequency distribution of 2510 unpaired restriction fragments of D10S28. (b) Plot of bin relative frequencies for D10S28 Complete data (n=1255).

length below 0.871 Kb or above 5.220 Kb have relative frequency less than 0.01. The relative frequency of the fragment lengths between 1.197–4.821 Kb is less than 0.035. The highest relative frequency 0.083 is observed for bin 9, that is for the range 1.638-1.788 Kb. Figure 3.26(a-d) gives the distribution of alleles in four subgroups. It is observed that the low frequency as well as high frequency bins are consistent in Asian, Hispanic, Black and Caucasian populations. Since the Asian group consists of several subgroups, distribution of alleles is different than the Hispanics, Blacks and the Caucasians.

Estimation of allele distribution

Figure 3.27(a) and 3.27(b) suggest that some of the fragment pairs are coalesced. Interval lines are of width 0.033 Kb. In Figure 3.27(a), note that there are more

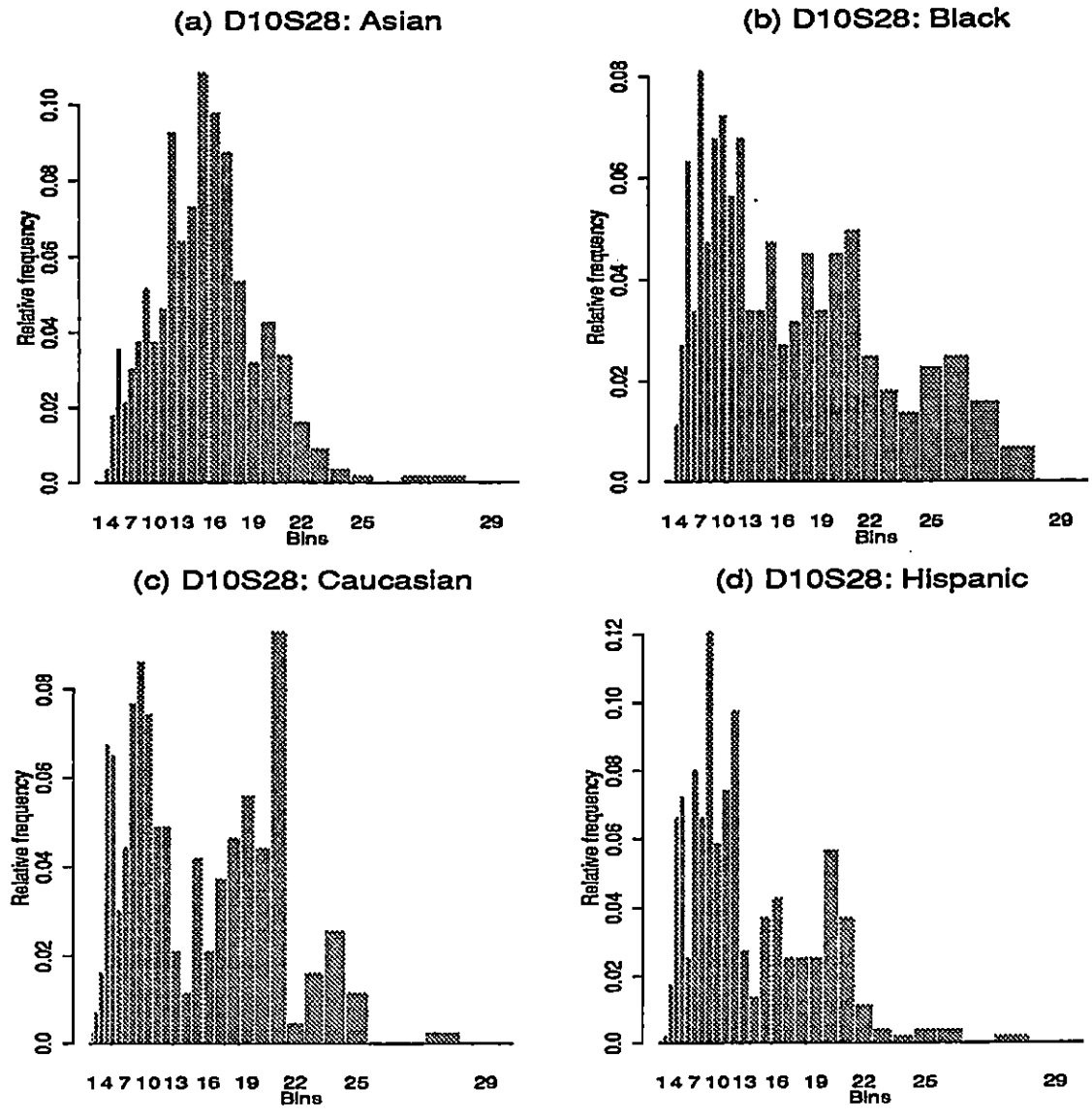


Figure 3.26: Plot of bin relative frequencies for Asian (n=572), Black(n=222), Caucasian(n=215), and Hispanic(n=256) population.

observations with zero differences and fewer observation in adjacent intervals. Many points fall on the line $y = x$, in graph 3.27(b), and fewer observations are observed in intervals adjacent to this line. Both the plots in Figure 3.27 indicate the probability of coalescence, that is observing the close heterozygotes as homozygotes.

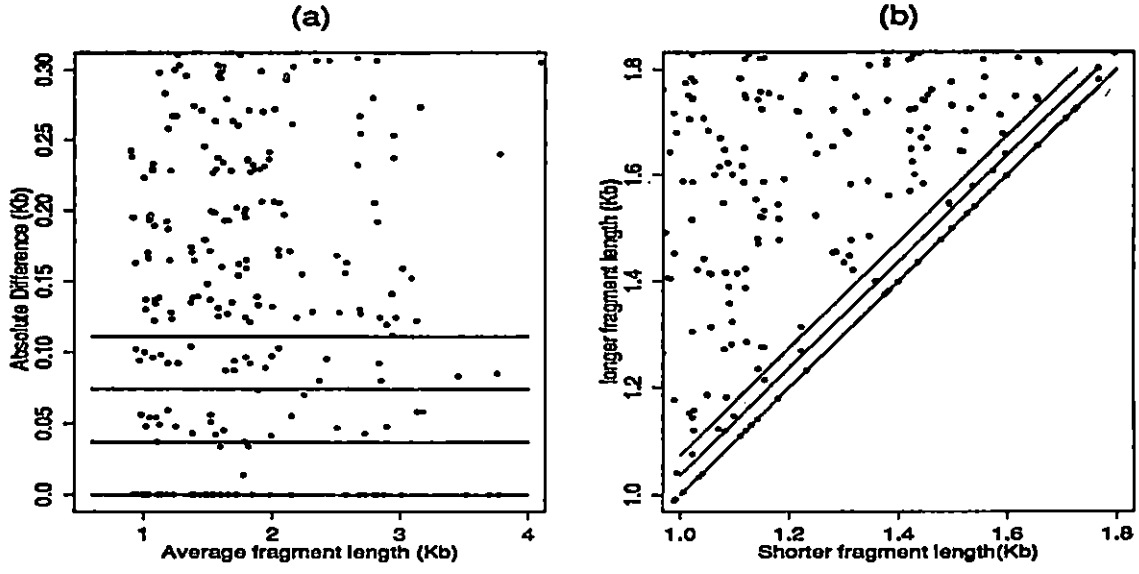


Figure 3.27: (a) Plot of the absolute difference of fragment lengths against the average length. (b) Plot of the shorter fragment length against the longer fragment length.

To compute the probability of coalescence, as suggested in section 2.1.3.3, we plotted the points $[1 - \frac{O_k}{E_k}]$ versus c_k for the intervals of the length 0.016 Kb, 0.028 Kb and 0.03 Kb. Figure 3.28, is the plot with the interval length 0.028 Kb. As it is not possible to fit any reasonable function to the points in this plot, the probability of coalescence cannot be easily estimated. SQPCOAL and EMCOAL cannot be used for obtaining the allele distribution. To compute the relative frequencies using EMRESL or SQPRESL, we need to estimate the resolution distance α. From Figure 3.29(a),

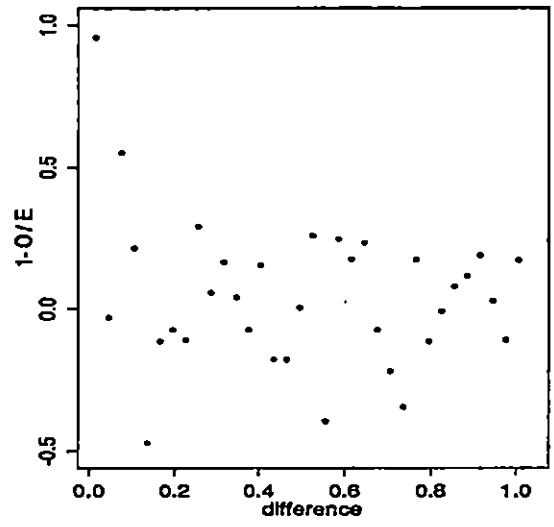


Figure 3.28: Plot of probability of coalescence as a function of the difference in fragment length

the resolution threshold is determined to be around 1.6% of the molecular weight of the single band observed.

The estimates of relative frequencies and the flanking-region size are obtained using EMRESL algorithm. The flanking region is estimated to be 0.6056 Kb. The pattern of the estimates of the allele distribution in Figure 3.29(b) shows a spiky nature and smoothed estimates could be obtained using empirical Bayes methods. Since the whole procedure takes a large amount of computational time, the EB estimates as well as estimates of the variance of π are not obtained. The pattern of the ML estimates is similar to that observed in the histogram of unpaired fragment lengths, Figure 3.25. The estimates of the flanking-region and the relative frequencies for 110 alleles, with step size 3 were also obtained using SQPRESL. Since the algorithm

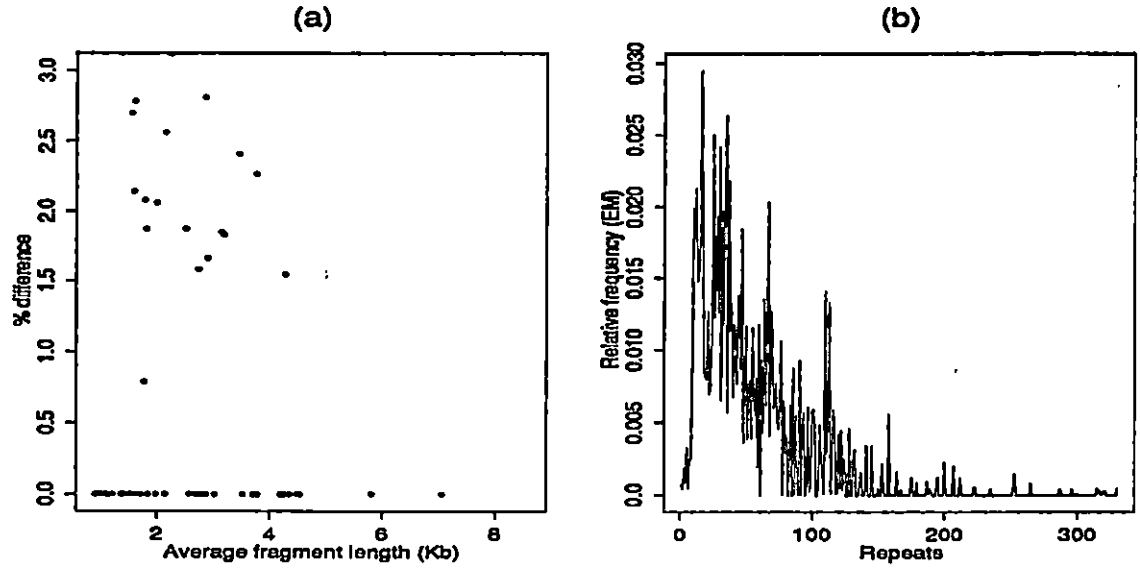


Figure 3.29: (a) Plot of the difference between fragments as a percentage of the average fragment length. (b) Plot of the estimate of allele distribution against the number of repeats.

SQPRESL could not converge with step-size less than 3, the results are not reported here.

D1S7

The descriptive summary of the data set for locus D1S7 is as follows.

Discriptive Summary

Number of individuals : 1243	Repeat Size : 0.009 Kb
Minimum fragment size : 0.756 Kb	Maximum fragment size : 23.186 Kb
Average length : 5.538 Kb	Estimated number of alleles : 2490
SD $= ca_r = 0.008a_r$	SD of average length : 0.043 Kb

Figure 3.30 does not show any repeating pattern of the fragment lengths which

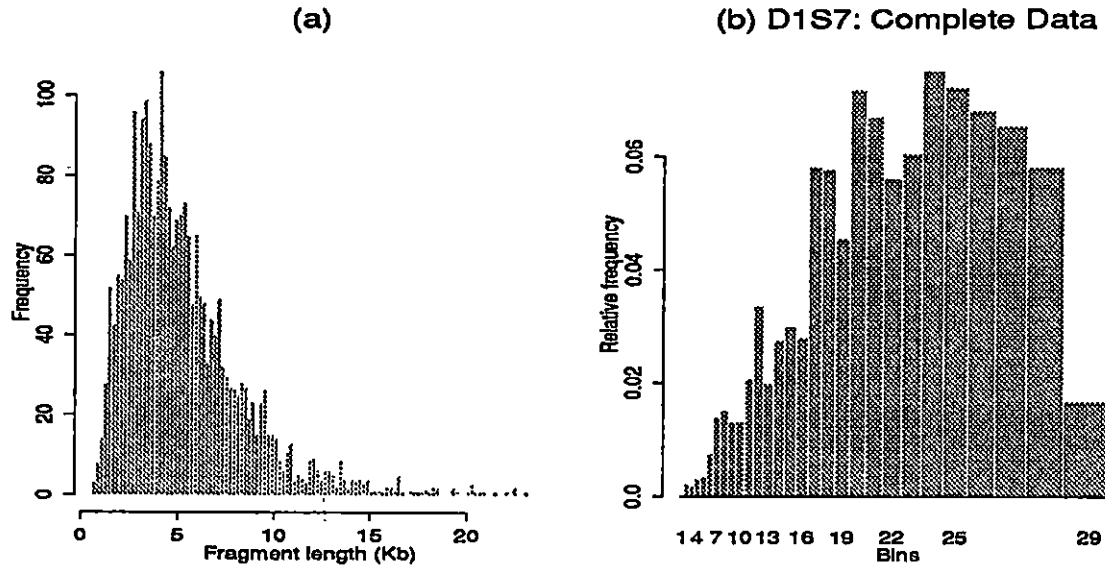


Figure 3.30: (a) Frequency distribution of 2486 unpaired restriction fragments of D1S7. (b) Plot of bin relative frequencies for D1S7 Complete data(n=1243).

indicates that the fragments are associated with a single flanking region. Most of the fragments are observed between 0 - 10 Kb. As the repeat length is very small, for an allele of average length, 5.538 Kb without coalescence, the phenotypes are expected to overlap $9.4 = \frac{1.96 \times 0.043}{0.009}$ repeats on either side of the genotypes 95% of the time.

Fixed-bin Results

For locus D1S7 the repeat length of 9 bp is very small and the observed fragment lengths are widely spread hence the number of alleles to be estimated is quite large. From Figure 3.30(b) it is observed that most of the fragments are between 0.75 - 6 Kb. Figure 3.31(b-d) gives the distribution of alleles in four subgroups. From Figure 3.31(a-d) of the distribution of alleles in four subgroups, it is observed that the low frequency as well as high frequency bins are consistent in all four groups. In addition,

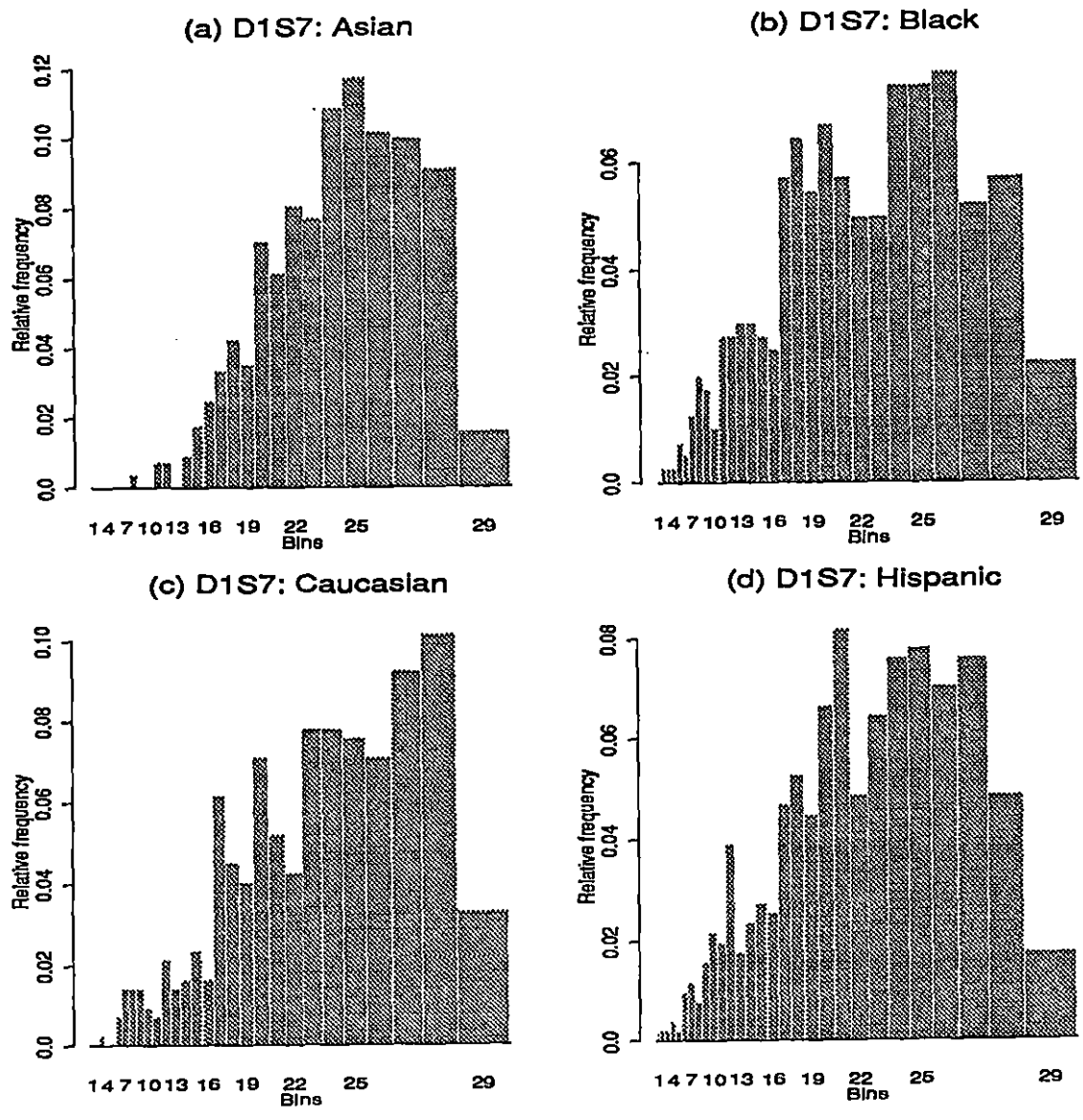


Figure 3.31: Plot of bin relative frequencies for Asian (n=1243), Black(n=202), Caucasian(n=212), and Hispanic(n=257) individuals.

it can be concluded that Asian, Hispanic, Black and Caucasian populations have similar allele distributions.

Estimation of allele distribution

Coalescence is observed from both the plots in Figure 3.32. In Figure 3.32(a) there are more points on the zero difference line in adjacent intervals. This phenomenon is also observed in Figure 3.32(b). More points are observed on line $y = x$, and fewer points are observed in adjacent intervals. This shows that some of the close heterozygotes are observed as homozygotes.

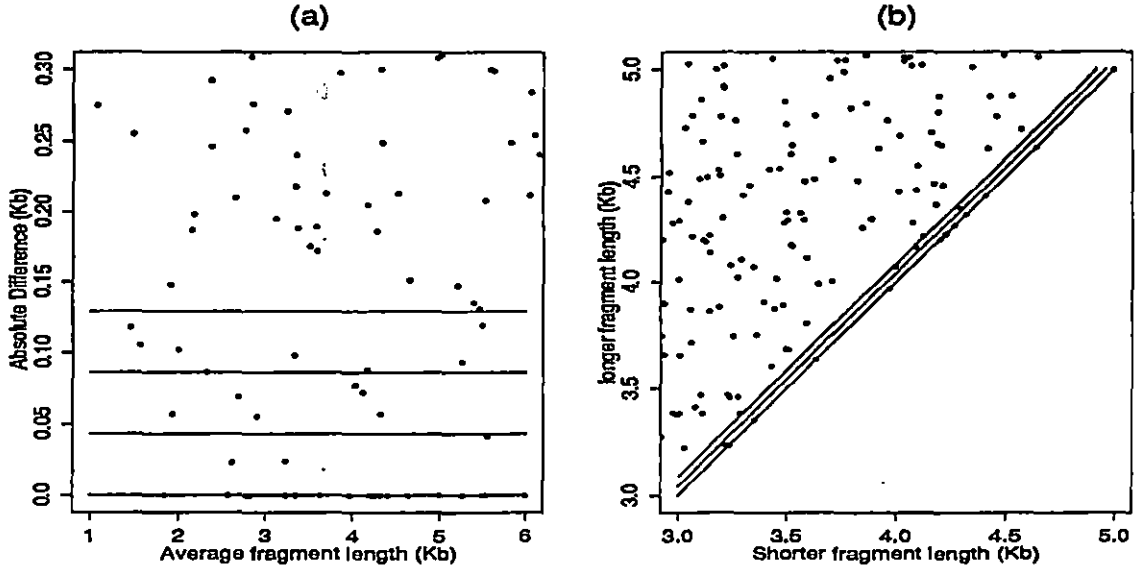


Figure 3.32: (a) Plot of the absolute difference of fragment lengths against the average length. (b) Plot of the shorter fragment length against the longer fragment length.

To compute the probability of coalescence, as suggested in section 2.1.3.3, we plotted the points $[1 - \frac{O_k}{E_k}]$ versus c_k for the intervals of the length 0.05 Kb, 0.025 and 0.09 Kb. The plot with the interval length 0.05 Kb is shown in Figure 3.33, and the

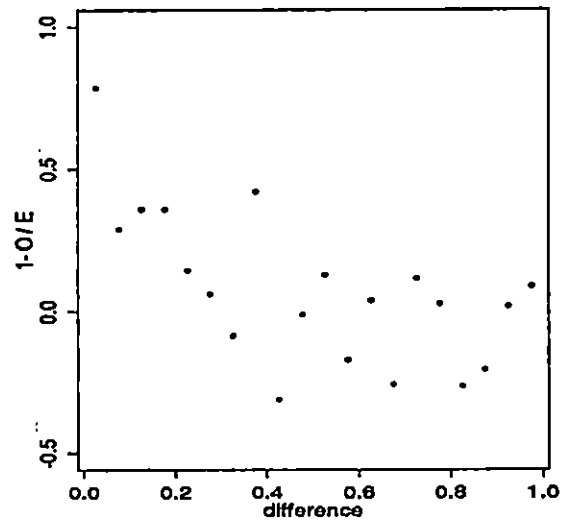


Figure 3.33: Plot of probability of coalescence as a function of the difference in fragment length

other two plots are also similar to this plot. It is not possible to fit any reasonable function to the points in Figure 3.33, hence the algorithm SQPCOAL and EMCOAL cannot easily be used for obtaining the allele distribution. To compute the relative frequencies using EMRESL or SQPRESL, we need to estimate the resolution distance α. From Figure 3.34(a), the minimum resolution distance is approximately 1.75% of the observed band.

The number 2490 of alleles to be estimated, is large as compared to the sample size of 1243. SQPRESL was tried with initial estimates $\pi_r^{(0)} = \frac{1}{249}$, $u^{(0)} = 0.6$ Kb for $r = 10j$, $j = 1, 2, \ldots, 249$, a step-size of 10. The routine could not improve upon initial estimates, therefore the allele distribution is estimated using EMRESL. Frequency estimates with $r = 10j$, $j = 1, 2, \ldots, 249$ are obtained. The flanking-region

is estimated as 0.7 Kb. The estimated distribution of allele frequency estimates in Figure 3.34(b) shows the same pattern as observed in histogram Figure 3.30.

Figure 3.31(a), gives the allele distribution of D1S7 locus for complete data as well as subpopulations. All 29 bins have observable events. The highest relative frequency is 0.0748. The relative frequency is observed to be high for large fragments. The allele distribution does not show substantial variation among the three subpopulations.

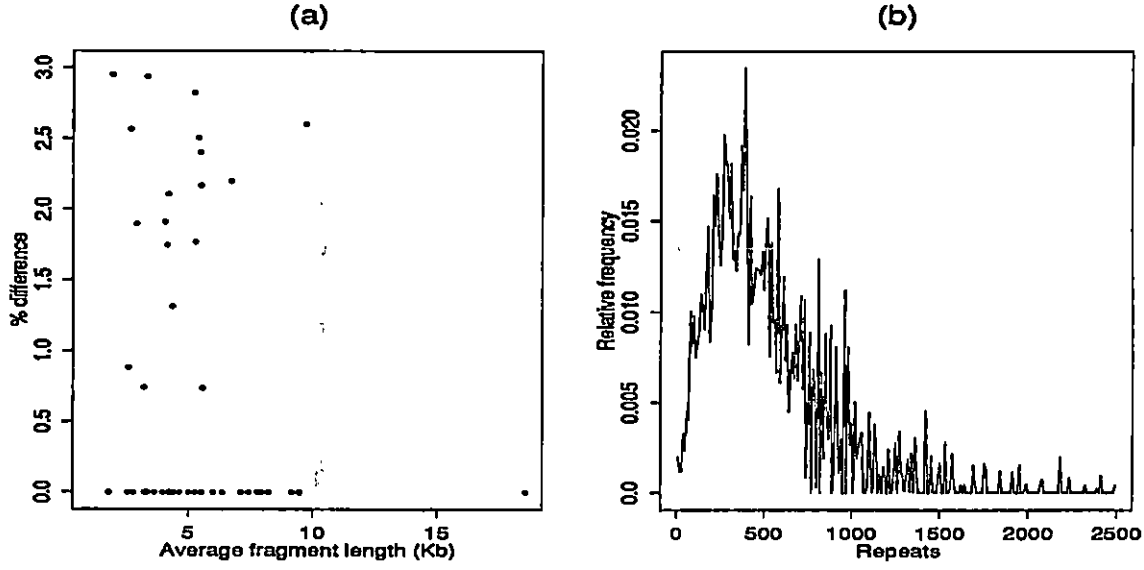


Figure 3.34: (a) Plot of the difference between fragments as a percentage of the average fragment length. (b) Plot of the estimate of allele distribution.

3.3 Independence of VNTR loci

Estimates of the frequency distribution of unpaired fragments were computed in Sections 3.1 and 3.2. Generally a pair of fragments is observed. Weir(1992) suggested that the two fragment lengths observed at each locus can be assigned to the bins they

Table 3.1: Frequency of observing genotypes in fixed-bin analyses

Bins	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1															
2	1	0														
3	1	0	0													
4	1	0	1	0												
5	2	0	0	0	0											
6	0	0	0	0	0	0										
7	2	0	1	0	0	2	2									
8	2	2	1	1	2	2	4	1								
9	2	0	0	2	2	2	2	0	1							
10	0	0	2	1	0	1	1	2	0	2						
11	3	0	0	0	2	1	5	2	0	2	0					
12	3	2	1	1	2	3	3	5	4	6	6	3				
13	0	1	1	1	3	1	3	3	5	4	7	4	2			
14	4	0	0	2	3	4	1	1	2	2	3	8	2	0		
15	6	0	0	0	3	2	1	2	1	3	5	4	3	2	1	
16	0	0	0	1	0	0	0	1	0	2	1	0	1	1	0	0

belong to and then this pair of bins is considered to be the genotype of the individual for that VNTR locus. For example, Table 3.1 gives the counts for each pair observed in the Black population for locus D4S139.

The summary statistics for binned data in the OCSD data base is given in Table 3.2. Here the column 'N' is the number of individuals, and 'No. of bins' gives the number of bins with nonzero frequency. The number of possible genotypes is the number of all pairs possible from the available number of nonzero frequency bins. For example, consider the Black population of locus D4S139. There are $f = 16$ available bins, and so the total number of pairs possible is $\frac{f\times(f+1)}{2}$. The number of observed

Table 3.2: Summary for binned data sets in OCSD database

Locus	Ethnic Group	N	No.of Bins	No. of Genotypes Possible	seen	Gene Diversity
D1S7	Black	202	22	253	119	0.9472
	Caucasian	212	23	276	131	0.9435
	Hispanic	257	24	300	146	0.9449
D2S44	Black	213	21	231	110	0.9360
	Caucasian	215	19	190	104	0.9258
	Hispanic	248	20	210	107	0.9158
D4S139	Black	210	16	136	89	0.9201
	Caucasian	217	14	105	69	0.8935
	Hispanic	243	14	105	72	0.8913
D10s28	Black	222	26	351	145	0.9512
	Caucasian	215	23	276	124	0.9431
	Hispanic	256	22	253	133	0.9344
D17S79	Hispanic	195	8	36	29	0.8273

genotypes is the sum of nonzero frequency cells in the two-way frequency table, as shown in Table 3.1. In Table 3.1 only 89 cells have nonzero counts, hence there are 89 observed genotypes for the Black population of locus D4S139. The *gene diversity* of any locus as defined in Weir(1992) is the proportion of heterozygotes observed at that locus. The gene diversity is calculated as one minus the sum of squared bin relative frequencies, that is $1 - \sum_i^f p_i^2$ where p_i is the relative frequency for bin i. Now $\sum_i^f p_i^2$ is the probability of observing only homozygotes, thus the gene diversity is also a measure of heterozygosity. The larger the heterozygosity, the more useful is the locus for DNA fingerprinting. All of these loci have a high percentage of heterozygotes.

When modeling the distribution of allele sizes, we assumed that the allele sizes

are independent within each loci. Independence between VNTR loci is also required when estimating the probability of observing a set of allele pairs at several VNTR loci. Weir (1992) gives a method of estimating correlations between fragment lengths within a locus and for different loci.

If individual i has two fragments at the same locus, say x_{ij}, $j = 1, 2$ for $i = 1, 2, \ldots, n$ where x_{ij} are distributed about the same mean μ_x with same variance σ_x^2. Then ρ_x, the correlation coefficient between the two fragment lengths (x_{i1}, x_{i2}) is called the *intraclass correlation coefficient*, and it can be estimated using analysis of variance techniques and a random effects model. The ANOVA table is given in Table 3.3 where $x_{i.} = \sum_j x_{ij}$, $x_{..} = \sum_i x_{i.}$.

Table 3.3: Analysis of variance for fragment lengths

Source	d.f	Mean square	Expected MS
Between individuals	$n-1$	$MSB_x = \frac{1}{2(n-1)}\left(\sum_i x_{i.}^2 - \frac{1}{n}x_{..}^2\right)$	$(1+\rho_x)\sigma_x^2$
Within individuals	n	$MSW_x = \frac{1}{n}\left(\sum_i \sum_j x_{ij}^2 - \frac{1}{2}\sum_i x_{i.}^2\right)$	$(1-\rho_x)\sigma_x^2$

The intraclass correlation coefficient is then estimated as

$$\hat{\rho_x} = \frac{MSB_x - MSW_x}{MSB_x + MSW_x}$$

Estimates of intraclass correlation coefficients for OCSD data are given in Table 3.4.

To check whether the correlation coefficient differs significantly from zero, 1000 bootstrap samples are generated from the available database with replacement. The

coefficient of correlation between N pairs is estimated for 1000 samples. The range of coefficients into which the central $(1-\alpha)$ of estimates fall is considered to be a confidence interval for the original correlation coefficient. For any database, if such interval does not include zero, then the correlation coefficient is considered to be significantly different from zero at the α level of significance. Except for the Black population of locus D1S7, 95% confidence intervals of all populations given in Table 3.4 include zero. We conclude that the correlation coefficients between pair of fragments observed at each locus are not significantly different from zero at the 5% level of significance.

Table 3.4: Estimates of intraclass correlations for fragment lengths in OCSD database

Locus	Ethnic Group	N	Correlation Coefficient
D2S44	Black	213	0.0777
	Caucasian	215	0.0029
	Hispanic	248	-0.0012
D1S7	Black	202	-0.1687
	Caucasian	212	0.0082
	Hispanic	257	0.0573
D4S139	Black	210	-0.0744
	Caucasian	217	0.0256
	Hispanic	243	-0.0018
D10S28	Black	222	-0.0109
	Caucasian	215	-0.0288
	Hispanic	256	0.0233
D17S79	Hispanic	195	0.0074

The correlation between fragment lengths for different loci can be also estimated on similar lines. Suppose individual i has lengths (x_{i1}, x_{i2}) at one loci and (y_{i1}, y_{i2})

at a second loci, for $i = 1, \ldots, n$. Let the mean and variance for x_{ij} be μ_x and σ_x^2, and for y_{ij} be μ_y and σ_y^2 respectively. Then there are two correlations ρ_{xy_1} and ρ_{xy_2} between two loci depending on whether or not the two fragments were received from the same parental sex cell.

$$E(x_{ij}y_{ij}) = \mu_x\mu_y + \rho_{xy_1}\sigma_x\sigma_y$$

$$E(x_{ij}y_{ij'}) = \mu_x\mu_y + \rho_{xy_2}\sigma_x\sigma_y,\ j' \neq j.$$

Since we do not know which fragment came from which parent at each loci, we estimate the average of these two correlations. Hence the correlation for a random pair of fragments, one at each locus can be estimated. For computations the average fragment lengths are,

$$X_i = \frac{x_{i1} + x_{i2}}{2},\ \ Y_i = \frac{y_{i1} + y_{i2}}{2}.$$

The mean product between two individuals for the two loci is calculated as

$$MPB_{xy} = \frac{1}{(n-1)}\left[\sum_{i=1}^{n} X_iY_i - \frac{1}{n}\sum_{i=1}^{n} X_i \sum_i Y_i\right],$$

then the average correlation between the loci is estimated as

$$\frac{1}{2}\widehat{(\rho_{xy_1} + \rho_{xy_2})} = \frac{2MPB_{xy}}{\sqrt{(MSB_x + MSW_x)(MSB_y + MSW_y)}}.$$

Table 3.5 gives estimates of correlation coefficients for OCSD data base.

To check whether the correlation coefficient differs significantly from zero, 1000 bootstrap samples are generated from the database of 2 VNTR loci with replacement. The correlation coefficients are estimated for 1000 samples. The range of coefficients

Table 3.5: Estimates of correlation coefficients between random pair of fragment lengths at different loci

Loci	Black		Caucasian		Hispanic	
	N	Correlation	N	Correlation	N	Correlation
D2S44, D1S7	194	-0.0525	211	-0.0544	246	-0.0377
D2S44, D4S139	200	0.1061	208	-0.0062	232	-0.0559
D2S44, D10S28	213	0.0432	214	-0.0203	245	-0.0709
D2S44, D17S79	–	–	–	–	195	-0.0057
D1S7, D4S139	189	0.0012	204	-0.0267	241	0.0941
D1S7, D10S28	202	-0.0322	212	0.0312	254	0.0121
D1S7, D17S79	–	–	–	–	193	-0.0630
D4S139, D10S28	209	-0.0183	207	0.0039	240	-0.0169
D4S139, D17S79	–	–	–	–	192	-0.0033
D10S28, D17S79	–	–	–	–	193	-0.0273

into which the central $(1-\alpha)$ of estimates fall is considered to be a confidence interval for the original correlation coefficient. For any database if such interval does not include zero, then the correlation coefficient is considered to be significantly different from zero at the α level of significance. For each ethnic group and for all possible pairs of VNTR loci 95% confidence intervals are estimated. Since all confidence intervals include zero, we concluded that the correlation coefficients between loci are not significantly different from zero at 5% level of significance.

Chapter 4

Inferences: DNA fingerprinting

DNA fingerprinting is a very powerful technique in forensic science. It is used to draw inferences given measurements on bodily fluids and other material found at crime scenes. The culpability of a suspect is determined from the similarities between the suspect's profile with that of material found at the crime scene. A *match* is declared if two profiles show sufficient similarities, and then with some measure of the weight, this evidence is presented to the jury. If the chance that a randomly chosen innocent individual shows the same degree of match as that between the suspect and the evidence sample is very small, then the suspect is almost certainly the criminal; this is called *inclusion.* If the evidence sample contains a band not present in the sample from the suspect then, assuming no contamination of this sample and no errors in the laboratory analysis, the suspect is clearly not the criminal; this is called *exclusion.* If exclusion is declared, then the case does not go for a trial. Different

methods of drawing inference from DNA fingerprints are given in Berry(1991)and Roeder(1993). There are some controversies about the statistical assumptions made, the computations of allele frequencies and the methods drawing the inferences. A summary of this controversy is given in Roeder(1993). This chapter consists of two sections. Section 1 includes a discussion of the methods for drawing inferences from the DNA fingerprint using the estimated allele distribution and Section 2 contains a small discussion on factors related to the controversy of DNA fingerprinting.

4.1 Methods of Inference

4.1.1 Match/Binning

Match/binning is a reasonable inferential method in scientific settings. This method is easily understood by the legal community. The name of the method follows from the two stages of the procedure. First a match or exclusion is determined and if there is a match then the *match proportion*, the probability of observing this match from a randomly selected population, is calculated using a binning technique. The match proportion is the proportion of the reference population that falls in the bin containing the crime sample.

Declaring a match visually is not always possible. As we have seen earlier, the fragments are measured with error, hence forensic laboratories use a K standard deviation criterion. That is, if the DNA of the suspect and evidentiary sample are

separated by less than K standard deviations of the average measurement, a match is declared. As discussed in Roeder(1993) and Berry (1991), Lifecodes use a $K = 3$ standard deviation criterion. For LC, the SD of measurement error is 0.006 times the fragment length, hence the match is declared if the absolute difference between two bands is less than 0.018 times the average length. For calculating the match proportion, the fixed-bin analysis described in Chapter 2 is used. The probability that a fragment from the locus falls into a particular bin is estimated by the proportion of observations from the database in that bin. Consider a case of observing only one fragment for each of the suspect's and the crime scene sample at a locus. Let b_1 be the suspect's band size and b_2 be the band size for the evidentiary sample, and a be the average of (b_1, b_2). A match is declared if $b_2 \in \mathcal{I} = [b_1 - 0.018a,\ b_1 + 0.018a]$, with 3 SD criterion as used by LC. For OCSD, a match is declared if $b_2 \in \mathcal{I} = [b_1 - 0.024a,\ b_1 + 0.024a]$ for 3 SD. If a match is observed and p is the relative frequency of interval $\mathcal{I}$, then assuming homozygosity, the match proportion is p^2. Berry (1991), also suggests that, there is a possibility of not observing another band due to a small amount of DNA available or degradation of the sample. In that case the match proportion is considered to be $2p$. When two fragments are observed for both samples, then two intervals say $\mathcal{I}$, $\mathcal{J}$ are determined, one for the two larger bands and one for the two smaller bands respectively. If a match is declared for both the larger and smaller bands, and the relative frequency of $\mathcal{I}$ is p_1 and that of $\mathcal{J}$ is p_2, then the total match proportion is $2p_1p_2$. If a match is not observed for both larger

and smaller bands, then exclusion is declared. An example similar to that discussed in Berry(1991) is given here to explain the above discussion.

Example:

In one murder case, a blood spot is observed on the suspect's watch (crime scene sample). The suspect is declared a culprit if the match is observed between the two DNA fingerprints, of the blood on the watch and the victim's blood sample. The DNA fingerprint is obtained for two VNTR loci, D2S44 and D17S79. Table 4.1 gives the band sizes for both samples, the averages, 3SD intervals and the relative frequencies of these intervals.

Table 4.1: DNA pattern summary: for calculating match proportion

	D2S44	D17S79	
Victim's sample	10.162	3.869	3.464
Watch sample	10.32	3.877	3.52
Average	10.241	3.873	3.492
3 Sd interval	[9.978, 10.346]	[3.799, 3.939]	[3.401, 3.527]
Relative freq.	0.049	0.111	0.155

For the locus D2S44, only one band is observed in both samples. From Table 4.1, it is observed that the band in crime sample falls in the match interval, hence a match is declared. Assuming homozygosity the match proportion is calculated as

$(0.049)^2 \approx 1/420$. If it is assumed that due to the contamination of the sample, one band is missing in each sample then this proportion is $2(0.049) \approx 1/10$.

For D17S79, the average of the two larger bands is 3.873 Kb, which gives the interval $3.869 \pm 0.07 = [3.799, 3.939]$. For smaller bands, the average is 3.492 Kb and the interval for match is $[3.401, 3.527]$. As the relative frequencies are 0.111 and 0.155 for two intervals, the combined match proportion is $2(0.111)(0.155) \approx 1/29$. The estimated match proportion for observed DNA profile of two loci is $(1/10)(1/29) \approx 1/290$ for heterozygosity, assuming there is only band at D2S44 .. Assuming homozygosity at D2S44, this proportion is $(1/420)(1/29) \approx 1/12200$.

There is another method of determining the match proportion that involves the calculation of the likelihood ratio.

4.1.2 Calculation of Likelihood Ratio

Devlin et al. (1992) suggest a method of computing the likelihood ratio which they call the *identity index.* This index helps to determine whether the DNA of the suspect and the evidentiary sample comes from the same person or not. The method is as follows:

Let

H_0: The two samples come from different persons.

H_1: The two samples are derived from the same person.

Let (x_1, x_2) and (y_1, y_2) be a pair of fragments observed from the evidentiary sam-

ple and from the suspect respectively. The conditional density of the fragments is $g_{ij}(x_1, x_2)$, which is a joint normal distribution of the pair X_1, X_2, given that X_1 is the measurement of allele a_i with i repeats and X_2 is the measurement of allele a_j with j repeats. Though the population is assumed to be in H-W equilibrium, the measurement errors are assumed to be correlated. Assuming no coalescence, the joint marginal distribution of a pair of measured fragments is approximated as:

$$f(x_1, x_2) = \begin{cases} \sum_i \pi_i^2 g_i(x_1), & \text{if } x_1 = x_2 \\ 2\sum_{i<j} \pi_i \pi_j g_{ij}(x_1, x_2), & \text{if } x_1 < x_2 \end{cases}$$

Then the likelihood of H_0 is given by

$$lik(H_0|x_1, x_2, y_1, y_2) = f(x_1, x_2) f(y_1, y_2) \tag{4.1}$$

Given the data, the likelihood of H_1 is

$$lik(H_1|x_1, x_2, y_1, y_2) = \begin{cases} \sum_i \pi_i^2 g_i(x_1) g_i(x_2), & x_1 = x_2, y_1 = y_2 \\ 2\sum_{i<j} \pi_i \pi_j g_{ij}(x_1, x_2) g_{ij}(y_1, y_2), & \text{otherwise} \end{cases} \tag{4.2}$$

The likelihood ratio $L = \frac{lik(H_1|x_1,x_2,y_1,y_2)}{lik(H_0|x_1,x_2,y_1,y_2)}$, which is termed the *identity index* is then computed. If samples are on multiple loci, then assuming independence over loci, the identity index can be obtained by multiplying the likelihood ratios for each locus.

Computations of likelihood change in case of coalescence. Let $x_1 = x_2 = z_1$ and $y_1 = y_2 = z_2$ Then the likelihood of H_0 is

$$lik(H_0|z_1, z_2) = f^\star(z_1, z_1) f^\star(z_2, z_2),$$

where

$$f^*(z,z) = \sum_i \pi_i^2 g_i(z) + 2\sum_{i<j} \pi_i \pi_j \int_0^\infty g_{ij}(z-t, z+t)\delta(t,z)dt.$$

If only one pair is coalesced say (x_1, x_2) then $f^*(z_2, z_2)$ is replaced by $f(y_1, y_2)$. If the evidentiary sample is truly homozygous then under H_1 the suspect sample is also homozygous. Thus the likelihood of H_1 when both pairs of fragments have coalesced is

$$lik(H_1|z_1, z_2) = \sum_i \pi_i^2 g_i(z_1)\pi_i^2 g_i(z_2) + 2\sum_{i<j} \pi_i \pi_j \;\; I_1 I_2$$

where $I_1 = \int_0^\infty g_{ij}(z_1 - t, z_1 + t)\delta(t, z_1)dt$ and $I_2 = \int_0^\infty g_{ij}(z_2 - t, z_2 + t)\delta(t, z_2)dt$. Once the likelihood under H_0 and H_1 is obtained, the identity index can be computed as discussed earlier. The decision rule given by Devlin et al. (1992) is

$$\begin{aligned} L \;&\in\; (0, \frac{1}{K}] \;\; \text{Conclude } H_0 \\ &\in\; (\frac{1}{K}, K] \;\; \text{inconclusive} \\ &\in\; [K, \infty) \;\; \text{conclude } H_1 \end{aligned}$$

K is chosen such that Type I error is very small. Devlin et al. (1992) suggest the value of K as 100.

The following is the example given in Devlin et al.(1992). Suppose the DNA fingerprint is obtained for two samples on loci D2S44 and D17S79. Band sizes measured for D17S79 are $x_1 = 3.66, x_2 = 4.01, y_1 = 3.56, y_2 = 3.84$ and for D2S44 are $x_1 = 12.06, x_2 = 12.99, y_1 = 11.64, y_2 = 3.84$. L is calculated as 4.7×10^{-11}. From the

above decision rule if K=100, it can be concluded that the two samples come from different individuals.

4.2 Summary of Controversy

Since the discovery of fingerprinting, scientific testimony is considered to be a deciding factor in most of the civil and criminal cases. However, because of inadequate scientific background, judges and lawyers, do not review the scientific techniques used, and judicial decisions are often made on the basis of experts testimony which is not argued or challenged.

In many techniques, the underlying theory is well established and results are independently verified. These methods are generally accepted for scientific research but may not be reliable for forensic applications. Failing to test the reliability of the techniques may cause the conviction of an innocent person. To use these methods in forensics, they must be tested thoroughly for their usefulness and limitations. Neufeld and Colman (1990) and Cohen (1990) give examples of forensic trials where the DNA profiling results and their interpretation seemed doubtful. The controversy is mainly associated with statistical issues. Roeder (1993) lists these issues and also explains how they could have been avoided by careful consideration of statistical theory. Some of these issues include the appropriate method of summarizing data subject to measurement error, the assumption of independence of events in DNA pro-

file, heterogeneity of populations, appropriate sampling methods to develop reference populations and probabilistic evaluation of evidence under uncertainty of appropriate reference database.

In any forensic investigation, a match of two samples is first determined and if a match is observed then the match proportion is computed. As discussed in Section 4.1.1, a match is declared if the evidentiary sample bands are within 3 standard deviations of the bands observed in the suspects sample. Sometimes environmental factors such as temperature and humidity may degrade the samples. In such cases there is a possibility that the larger fragments are destroyed and the probe may yield only one smaller band. Hence contamination of the samples and band shifting are two major problems.

Forensic scientists usually estimate the match proportion of the suspect and evidentiary profile in a reference population by assuming independence of alleles within and between loci. This assumption has stirred the most controversy. If different subgroups of a population have different allele probability distributions then such heterogeneity would cause dependencies of alleles within and between loci. Studies by Lander (1989) and Cohen (1990) suggested that the population heterogeneity could lead to an underestimation of matching probabilities. Some other geneticists and statisticians (Devlin et al., 1990) have countered that, even though theoretically heterogeneity in the population causes dependence in loci, human populations rarely exhibit enough heterogeneity to have substantial impact on probability calculations.

A consequence of population heterogeneity is an excess of homozygotes in the mixed population which causes deviation from H-W equilibrium. Without assumptions of H-W equilibrium there is no other reliable way to convert allele frequencies into overall genotype frequencies. Devlin et al. (1990) suggest that there is no excess of homozygotes observed at VNTR loci and that some of the close heterozygotes are measured as homozygotes. Tests are developed by Berry et al. (1992), Weir (1992), and Devlin et al. (1992) to check the H-W equilibrium and independence of the loci, but some of these tests can be affected by correlated measurement error. Evett et al. (1993) have suggested a new method for computing match proportions, pointing out that the classical tests may fail to reject the independence hypothesis because of lack of power.

Besides independence, the choice of reference population is another factor of concern. Generally the reference population is chosen according to ethnicity of the suspect. Since the suspect is not guilty until proven, the population consisting of individuals with the same ethnicity is not the proper choice of reference population. For more information on the controversy refer to Roeder(1993), which discusses the statistical issues and consequences of the National Research Council's 1992 report in detail.

Neufeld and Colman (1990) point out some nonstatistical issues related to the unreliability of DNA profiling. Sometimes forensic labs declare a match even if the bands do not show a visual match and they are also not within the 3 standard

deviation interval. Generally, forensic laboratories carry out the required calculations on data that they have collected themselves, hence most of the data is not published, previewed or verified. Sometimes prosecutors refuse to divulge the raw data for verification and even if they are verified, different laboratories get different results due to different RFLP systems, probes and enzymes used. Hence the defense is unable to verify the results. In addition, lack of time and resources often results in no challenge of doubtful results. Most of the testing laboratories are not regulated by government, they do not have any restrictions on submitting proficiency tests, and they are within police or prosecutor agencies, so that their results are biased because technicians are aware of the facts of the case.

Neufeld and Colman (1990) and Cohen (1990) suggests that if applied properly DNA fingerprinting is a very reliable and powerful technique, and if some standards or requirement for certifying and licensing forensic laboratories are enforced under the national standards and regulation, DNA profiling will serve an important and beneficial role in criminal justice.

Chapter 5

Summary and Future Work

In this project we modeled the allele distribution for five VNTR loci of OCSD and one locus of LC. Two methods were used to estimate the allele distribution. One is fixed binning and the other uses mixture models. In the fixed-bin analysis, introduced by Budowle et al. (1991), fixed intervals are used to cluster the fragments into the same group and then the bins are treated as the alleles. While with mixture models, the allele frequencies are directly estimated. Both approaches have some advantages and disadvantages. Fixed-bin analysis is simple to understand, easy to use, the allele frequency distribution can be estimated for any loci, but the method loses the available information. The method of mixtures uses the information available but it is complex and time consuming.

In general, the allele length is the length of the flanking-region plus the number of repeats times the repeat length. The number of repeats is not observed directly

because the alleles are measured with error, which is generally associated with the allele length. The binning method cannot give the distribution of repeats, but with mixtures, we can estimate the distribution of repeats and the flanking-region. We used the mixture model for estimating the allele distribution for loci with single flanking region given by Devlin et al. (1991). Since coalescence is related to observing a single band, it plays quite an important role in estimating the distribution. Without estimating the probability of coalescence it is not feasible to use the model suggested in Devlin et al. (1991). For LC data, we estimated the allele distribution only for the locus D17S79. For the locus D2S44 it was not possible to estimate the allele distribution for two flanking regions with the available dataset. For OCSD's data on all VNTR loci, we were unable to estimate the probability of coalescence, hence we could not use SQPCOAL or EMCOAL. We modified the model given by Devlin et al. (1991). A single band is observed when the distance between the two bands is less than the resolution threshold, the minimum distance required to distinguish the two bands. The likelihood of the data is the sum of the likelihood of observing two bands and the likelihood of observing single band. We modified the likelihood of observing a single band using the resolution threshold. To check the model, we estimated the distribution of LC's D17S79 locus using the modified model, that is SQPRESL and EMRESL. The estimate of the flanking-region size and π are the same as estimated by SQPCOAL or EMCOAL. The fit of the model was also checked for OCSD's locus D17S79. The P-value of the test was 0.45, indicating that the model fitted the data.

We used SQPRESL and EMRESL for estimation with all of the OCSD data. SQPRESL could not improve upon the initial distribution sometimes or could not converge at all. Though the EM algorithm was slow, it converged for all the data sets. For D2S44 and D10S28 loci, we estimated the complete allele distribution, but for D1S7 and D4S139 loci we estimated the allele distributions with step size 10 and 3 respectively. Due to the large amount of computation time, we could not test the model fit for all datasets.

Using the resolution approach suggested in this project one can save computation time as well as the time required for estimating the probability of coalescence. This modified likelihood is useful to estimate the allele distribution for small to medium sized samples. It is possible to estimate the distribution of any VNTR locus with some step size by EMRESL.

Future work :

- Estimate the allele distribution using other measurement error distributions, such as Log-normal and Weibull.
- Estimate the allele distribution using other algorithms as VEM, VDM
- Investigate methods to reduce the execution time.

Bibliography

[1] Aitken, C. G., Stoney D. A. (unknown). *The Use of Statistics in Forensic Science.* Chapt. 5, 146-150.

[2] Balazs, I., Baired, M. and Meade, E. (1989). Human population genetic studies of five hypervariable DNA Loci. *American Journal of Human Genetics*, Vol 44: 182-90.

[3] Baird, M., Balazs, I., Giusti, A., Miyazaki, l., Nicholas, L., Wexler, K., Kanter, E., Glassberg, J., Allen, F., Rubenstein, P., and Sussman, L. (1986). Allele frequency Distribution of Two Highly Polymorphic DNA sequences in Three Ethnic Groups and its Application to the Determination of Paternity. *American Journal of Human Genetics*, Vol 39: 489- 501

[4] Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis.* New York: Springer.

[5] Berry, D. A. (1991). Inferences using DNA profiling in forensic identification and paternity cases. *Statistical Science*, Vol 6: 175-205.

[6] Berry, D. A., Evett, I. W., Pinchin, R. (1992). Statistical Inference in Crime Investigations using DNA Profiling: Single locus probes. *Applied Statistics*, Vol 41: 499 - 531.

[7] Budowle, B., Giusti, A. M., Waye, J. S., Baechtel, F. S., Fourney, R. M., Dwight, E. A., Lawrence, A. P., Harold, A. D., Monson, K. L. (1991). Fixed-Bin Analysis for Statistical Evaluations of Continuous Distributions of Allelic Data from VNTR loci, for use in Forensic Comparisons. *American Journal of Human Genetics*, Vol 48: 841-855.

[8] Cohen, J. E. (1990). DNA Fingerprinting for forensic identification: potential effects on data interpretations of subpopulation heterogeneity and band number variability. *American Journal of Human Genetics*, Vol 46: 358 - 368.

[9] Curnow, R. N., Wheeler, S. (1991). Assessment of Deoxyribonucleic Acid Fingerprinting Evidence in Paternity and Immigration cases. *Journal of Royal Statistical Society.*, Vol 154A: 97-99.

[10] Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society,*[b] Vol 39: 1-38.

[11] Devlin, B., Lindsay, B., Roeder, K. (1989). Application of Maximum likelihood methods to population genetic data for the estimation of individual fertilities. *Biometrics*, June : 363-379.

[12] Devlin, B., Risch, N., Roeder, K. (1990). No excess of homozygosity at DNA fingerprint loci. *Science*, Vol 249: 1416-1420.

[13] Devlin, B., Risch, N., Roeder, K. (1991). Estimation for allele frequencies for VNTR loci. *American Journal of Human Genetics*, Vol 48: 662-676.

[14] Devlin, B., Risch,N., Roeder, K. (1991). Forensic inference from DNA fingerprints. *Journal of the American Statistical Association*, Vol 87: 337-349.

[15] Devlin, B., Risch, N.(1992). Allele Frequencies for VNTR loci. *American Journal of Human Genetics*, Vol 51: 534-547..

[16] Efron, B., Morris, C. (1973). Stein's Estimation Rule and its Competitors - An empirical Bayes Approach. *Journal of the American Statistical Association* , Vol 68, 141 : 117-130.

[17] Elder, J. K., Southern, E. M. (1983). Measurement of DNA length of gel electrophoresis II: Comparison of methods for relating mobility to fragment length. *Analytical Biochemistry*, Vol 128 : 227 - 231.

[18] Evett, I. W., Scrange, J., Pinchin, R. (1993). An illustration of the advantages of efficient statistical methods for RFLP analysis in forensic science. *American Journal of Human Genetics*, Vol 52: 498 - 505.

[19] Galbraith, D. A., Boag, P. T., Gibbs, H., White, B. N. (1991). Sizing bands on autoradiograms: A study of precision for scoring DNA fingerprinting. *Electrophoresis*, Vol 12: 210 - 220.

[20] Holmlund, G., Karlberg, K., Gustavsson, B., Lindbolm, B. (1992). Calculation of restriction fragment lengths by image processing. *Electrophoresis*, Vol 13: 407 - 410.

[21] Jeffreys, A., Wilson, V., Thein, S. L. (1985). Hypervariable 'Minisatellite' regions in human DNA. *Nature*, Vol 314: 67 - 72.

[22] Kirby, L. T. (1990). *DNA Fingerprinting an Introduction.*. USA: Stockton Press.

[23] Koller, P. C. (1967). *Chromosomes and Genes: The Biological Basis of Heredity.* Edinburgh: Oliver & Boyd.

[24] Lander, E. (1989). DNA fingerprinting on trial. *Nature*, Vol 339 : 501-505.

[25] Lesperance, M., Kalbfleisch, J. D. (1992). An algorithm for computing the Nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association*, Vol 87: 120-126.

[26] NAG Fortran Library Manual Mark 15, (1991). (NAG central office, 7 Banbury Road, Oxford OX26NN), U.K.

[27] Nakamura, Y., Leppert, M., O'Connell, P., Wolfe, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., White, R. (1987). *Science*, Vol 235: 1616 - 1622.

[28] Neufeld, P., Colman, N. (1990). When Science Takes the Witness Stand. *Scientific American*, Vol 262, Number 5 : 46 - 53.

[29] Rao, P. (1983). *Nonparametric Functional Estimation.* New York: Academic Press.

[30] Roeder, K. (1993). DNA fingerprinting: A review of the controversy. *Statistical Science* , Vol 9: 222-278.

[31] Southern, E. M. (1979). Measurement of DNA length by gel electrophoresis. *Analytical Biochemistry*, Vol 100: 319 - 323.

[32] Shapiro, R. (1991). *The Human Blueprint: The race to unlock the secrets of our Genetic Script.* New York: St. Martin's Press.

[33] Titterington, D. M., Smith, A. F., Makov U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

[34] Weir, B. S. (1992). Independence of VNTR alleles defined as Fixed bins. *Genetics*, Vol 130: 873 - 887.

[35] Wyman, A. R., White, R. (1980). A highly polymorphic locus in human DNA. *Proc. of National Academy of Science* Vol 77: 6754 - 6758.

VITA

Surname: Chhatre Given Names: Varsha

Place of Birth: India Date of Birth: July 24, 1965

Educational Institutions Attended:

University of Victoria:	1992 to 1995
University of Poona:	1982 to 1987

Degree Awarded:

B.Sc	University of Poona	1985
M.Sc	University of Poona	1987

Honours and Awards:

University of Victoria Fellowship	1992-1994

PARTIAL COPYRIGHT LICENSE

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the Library of any other university, or similar institution, on its behalf or for one of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the University designated by me. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Title of Thesis: STATISTICAL METHODS FOR DNA FINGERPRINTING

Author 

(Signature)

CHHATRE VARSHA

(Name in Block Letter)

08/03/95

(Date)