

DNA FINGERPRINTING

An Interactive Qualifying Project Report

Submitted to the Faculty of the

WORCESTER POLYTECHNIC INSTITUTE

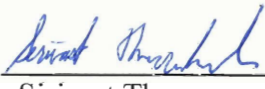
In partial fulfillment of the requirements for the

Degree of Bachelor of Science

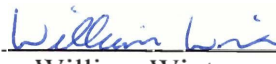
By



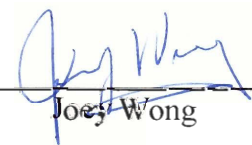
Yi Han



Sirinart Thanesvorakul



William Winter



Joey Wong

May 2, 2000

APPROVED:



Professor David S. Adams, BB
Project Advisor

ABSTRACT

With the increasing need in the accuracy of the forensic process, DNA fingerprinting becomes the most widely used technique in material-evidence analysis, allowing forensic scientists to accurately confirm or deny any forensic relationship in question. This project discusses the science behind DNA fingerprinting and its application in the court of law. Case studies show DNA fingerprinting to provide accurate evidence in many cases, while inadmissible due to improperly executed DNA analysis in others.

TABLE OF CONTENTS

Signature Page.....	1
Abstract.....	2
Table of Contents.....	3
Executive Summary.....	4
Project Objective.....	7
Introduction.....	8
Chapter 1: Genetics and Biology.....	9
Chapter 2: DNA Fingerprinting.....	18
Chapter 3: Applications of DNA Fingerprinting.....	33
Chapter 4: DNA in the Courtroom	38
Chapter 5: Probability and Statistics.....	59
Glossary.....	88
Bibliography.....	91

EXECUTIVE SUMMARY

In the midst of the greatest scientific achievements that occurred in the past century, biology was one of them. Biological information about humans, animals, and plants, from exoskeleton to cellular structure, has been accumulated over the years. With centuries of studies, it was only recently that breakthroughs beyond microbiology came in the discovery of genetic material.

With the identification of DNA as a genetic material and its characteristic as a biological blueprint, forensic scientists quickly sought DNA as a potential way to increase the accuracy of evidence analysis. With this growing interest, techniques of DNA analysis, or DNA fingerprinting, were soon adapted to prove or disprove the relationship between evidence and suspect in the court of law.

The purpose of this Interactive Qualifying Project is to analyze the use of DNA fingerprinting as a technique of forensic science and its success and failures in the court of law. We attempted to meet this goal by defining and exploring the molecular structure of DNA and how DNA fingerprinting is carried out. We also explored how it is used as evidence in court and the problems of finding and handling samples so that the evidence will remain effective and usable. Then, we studied some court cases that had been linked to the use of DNA fingerprinting in clearing the name of a person or in convicting a person of a crime. In addition, some statistics and probability are discussed in order to determine the match and estimate the population allele frequency distributions.

DNA is extracted from the evidence, mostly from blood, semen, tissue, or hair. By developing fragments using either Polymerase Chain Reaction (PCR) or Restriction Fragment Length Polymorphism (RFLP), a pattern can be detected and compared. This

pattern is in the form of bands on autoradiography or dots in reverse dot blot typing strips. The comparison of the patterns of evidence from the suspect and that from the crime scene will give a positive or negative connection of the suspect to the crime scene.

These comparisons are perhaps the most important evidence in determining a verdict. Because DNA is the blueprint of all life and it is unique to all individuals, one could consider DNA to be the most absolute fact of an individual. Both defense and prosecution use this to prove and disprove the accused. When DNA patterns match, it provides undeniable confirmation for the forensic relationship, whether it is positive or negative identification. However, since DNA fingerprinting is an analytical process performed by humans, its accuracy will no doubt be in question. In many cases when a positive result is found through DNA fingerprinting, the defense would challenge the validity of the results, stating that the analysis was improperly performed, and therefore inaccurate. This is a problem apparent to many DNA related cases, where the DNA evidence is either rejected by the court through the accuracy-challenge or ruled as inadmissible as evidence.

Determination of the probability of specimen match and estimation of population allele frequency distributions are two key areas of DNA profiling requiring probabilistic and statistical analyses. Therefore we described in great detail about Statistic and Probability in Chapter 5. For Statistical Methodology, we discussed *Random Sampling*, *Measure of Dispersion*, and *Statistical Inference*. For Probability, we discussed *Probability of Combined Events*, *Bayes' Theorem*, *Random Variables and Distributions*, and *Genetic Applications of Probability*.

DNA fingerprinting has proven to be one of the most accurate ways to analyze evidence. Due to its “potential” accuracy in distinguishing DNA samples, it has become more and more practical. Although this process does not provide absolutely perfect results, but with our constant technological advancements DNA fingerprinting will soon provide near perfect results.

PROJECT OBJECTIVE

The purpose of this Interactive Qualifying Project is to analyze the use of DNA fingerprinting as a technique of forensic science and its success and failures in the court of law. We attempted to meet this goal by defining and exploring the molecular structure of DNA and how DNA fingerprinting is carried out. We also explored how it is used as evidence in court and the problems of finding and handling samples so that the evidence will remain effective and usable. Then, we studied some court cases that had been linked to the use of DNA fingerprinting in clearing the name of a person or in convicting a person of a crime. In addition, some statistics and probability are discussed in order to determine the match and estimate the population allele frequency distributions.

INTRODUCTION

One of the greatest scientific advancements in the history of humans is the knowledge of biology. Over the last few centuries, especially the 20th, we have probed and mapped the structures of the body. We have developed techniques to interact with the constant physical changes of our being, and devices to counteract some of these changes. We have improved and increased the average life span of humans, and we will continue to do so with future improvements in biotechnology.

In the dawn of a new millennium, already with many breakthroughs in biology, it was not until late this century that these breakthroughs were widely adapted to use in other fields such as forensic science. Once adapted, the quality of forensic investigations was greatly improved. In legal cases where the identity of the suspect was unable to be determined through direct means, a technique of DNA analysis, called DNA fingerprinting, was used to aid positive identifications based on tissue samples obtained from the crime scene. The precision of such a technique has proven to be unmistakably the highest in all of forensic science. Hence the use of DNA fingerprints has been the number one evidence used in high landmark cases. Although DNA fingerprinting is a powerful new tool for forensic science, in many legal cases it has had trouble being accepted as evidence. This project will aim to explain why this is. In addition, many moral issues have arisen concerning the creation of DNA databases, thus the social impact of this new technology is complex and worth examining.

CHAPTER 1: GENETICS AND BIOLOGY

History of Genetics

The definition of Life is one of the most controversial terms in biological science. There are endless discussions as to the fundamental characteristics of life. Two frequently presented observations are the ability to reproduce and consume material. As they are all equally important in defining life, we will briefly expand on the property of reproduction and consumption of food, since it is appropriate to our discussion on **DNA fingerprinting**.

As we all know, on the molecular level all living things are unique to each other, with no two subjects sharing the same set of characteristics. However, at the organismal level, many living things superficially resemble each other in many ways perceivable to the eye. Some share the same basic properties such as appearance, physical structure, and behavioral habits. Scientists use the term *species* to classify these similarities.

How does an individual or a species maintain their basic similarities? Or what makes one living organism resemble another? Compatible reproduction. As we know all living things must reproduce. It is one of the defining features of life. In this complicated process, a new organism is created by either cell division (asexual reproduction) or the joining of two cells from different individuals of the same species (sexual reproduction). This new organism will retain most of the “parents” major traits. In early biology, this observed passing of physical information is called *inheritance*. In studying this phenomenon, scientists like Gregor Mendel used control or selective breeding to observe the inherited traits of pea plants. In 1865, Mendel proposed the

existence of a biological material responsible for this pass-over. This material was called **genes**. Early experiments proposed that genes carry information, instructions that specify the physical functions and properties of a living organism. In the process of sexual reproduction, each of the two parties contributes a copy of their genes that contains each of their traits. The resulting offspring exhibit traits from both parents.

Genetics was not an unfamiliar principal before Mendel. For centuries man had been improving the biological structures of plants and animals through selective breeding, so they could provide a better source of material consumption. Such techniques were also applied in our society as a tool to develop “better” offspring, hoping to improve the social and physical status of the family.

The biggest event in the evolution of biological science was the identification and characterization of this genetic material that is the basic mechanism of inheritance in all biology. For a long time, the principal of heredity was well known. However, the actual physical nature of the genetic material responsible for inheritance was still unknown. It was not until the mid-eighteen hundreds that genes were proposed as genetic material. In the early nineteen hundreds, however, a series of experiments surged the world of biology with the discovery of **DNA** as a genetic material.

In 1928, Frederick Griffith experimented (Figure 1.1) with the bacterium *Streptococcus pneumoniae*, which causes pneumonia in humans, but is fatal in mice. Griffith used two different strains of this bacterium, one virulent with a polysaccharide coating and one nonvirulent without the coating. He injected the mice with strains of dead virulent bacteria cells, and a mixture of dead virulent and nonvirulent cells. Mice injected with dead virulent cells survived, whereas mice given the mixture died. These

results showed that the dead virulent cells somehow converted the live nonvirulent cells to virulent, killing the mice. This process is called *transformation*.

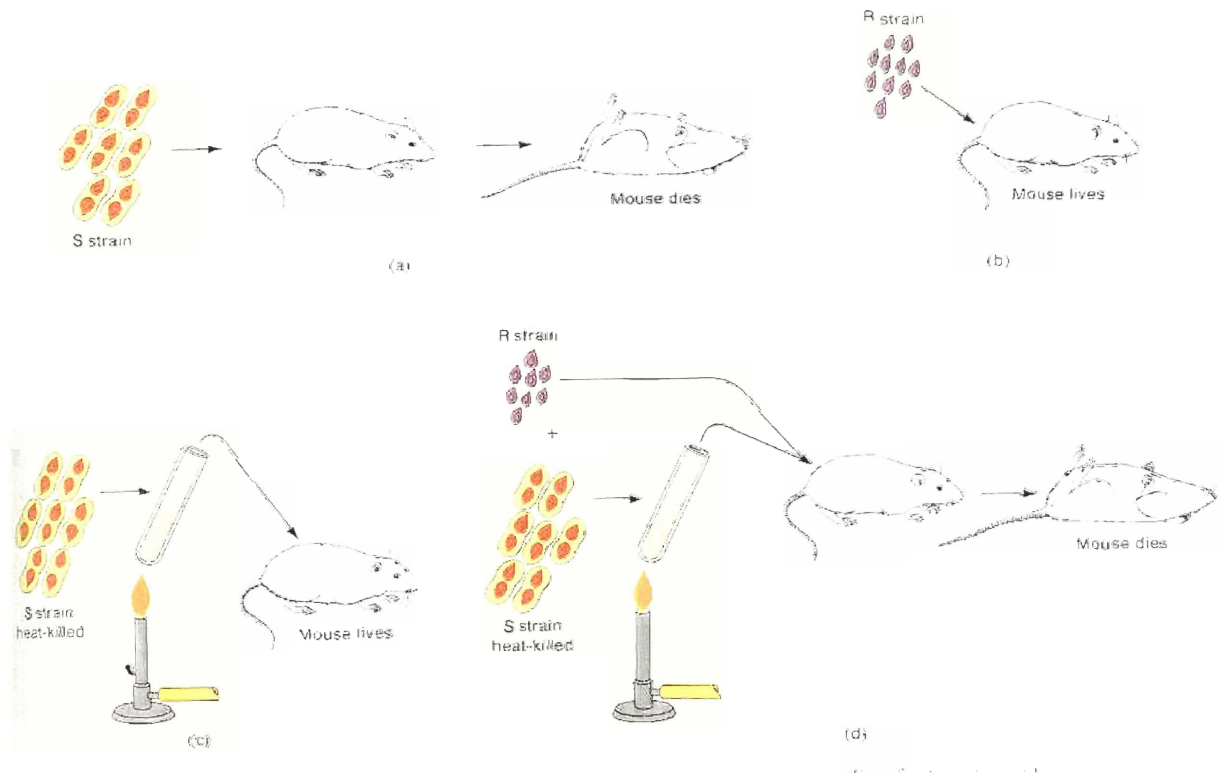


Figure 1.1: Griffith's *Streptococcus Pneumoniae* experiment was the first demonstration of DNA as a genetic material. (From A. J. F. Griffiths, J. H. Miller, D T. Suzuki, R. C. Lewontin, and W. M. Gelbart, *An Introduction to Genetic Analysis*, 6th ed. Copyright 1996 by W. H. Freeman and Company.)

In 1944, Oswald Avery, C. M. Macleod, and M. McCarty analyzed these dead cells. They found that the polysaccharide coating did not cause the transformation—the coating was basically a structural difference between the dead virulent cells and live nonvirulent, and this coating is what makes the cells virulent. At closer examination, Avery and his colleagues concluded that another molecule triggered this transformation. This type of molecule is called deoxyribonucleic acid, or DNA.

At first scientists were reluctant to accept DNA as the genetic material. Many questioned DNA's ability to store all the information pertinent to physical structures,

from organs to blood cells, based on the simplicity of the structure of the DNA molecule. It was in 1952 that scientists Alfred Hershey and Martha Chase solidified Avery's claims with a Phage virus experiment. They labeled the phage so that proteins could be distinguished from DNA. When the virus infected its target bacteria, it was found that DNA was passed onto the cell, not protein causing the infected cell to change. This observation provided undeniable evidence that DNA is the genetic material.

Components of DNA

DNA is composed of **nucleic acid**, a name that indicates that DNA is slightly

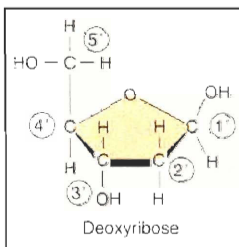


Figure 1.2: The sugar of DNA called deoxyribose is a five-carbon molecule. Each carbon is assigned by a number, 1'-5' starting at the carboxyl end (-COOH). (From Raven & Johnson, *Biology*, 4th ed. Copyright 1997 by The McGraw-Hill Companies, Inc.)

acidic. Each nucleic acid is a long chain of **nucleotides**. In the 1920s, a biochemist named P. A. Levine found that each nucleotide is made up of a five-carbon sugar called deoxyribose (Figure 1.2), attached to a phosphate (PO₄) group and a nitrogen-containing base.

The nitrogen base can be any of the four bases, **adenine**,

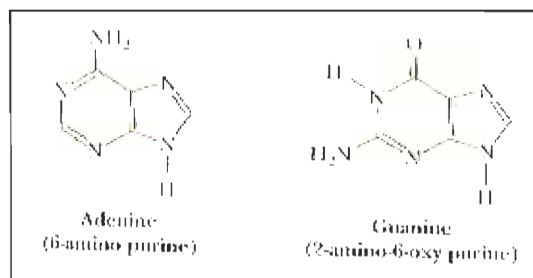


Figure 1.3: Adenine (A) and guanine (G) are the two purines. They possess two ring structures connected together. (From Garrett, R. G., and C. M. Grisham. *Biochemistry*, 2nd ed. Copyright 1999 by Saunders College Publishing.)

guanine, cytosine, and thymine. However, it is more convenient to abbreviate them as A, G, C, and T, respectively. Two of the bases, adenine and guanine, are double-ring

molecules, called **purines** (Figure 1.3) while the other two, cytosine and thymine, are **pyrimidines**, single-ring molecules (Figure 1.4).

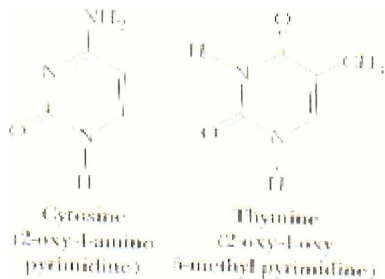


Figure 1.4: (left) Thymine (T) and Cytosine (C) are single-ring molecules called Pyrimidines. (From Garrett, R. G., and C. M. Grisham. *Biochemistry*, 2nd ed. Copyright 1999 by Saunders College Publishing.)

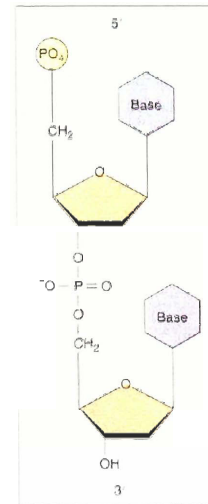


Figure 1.5: (right) A phosphodiester linkage. (From Raven & Johnson, *Biology*, 4th ed. Copyright 1997 by The McGraw-Hill Companies, Inc.)

In order to form a long chain of nucleotides, the phosphate group attached to the 5' position of the sugar in one subunit undergoes a dehydration reaction¹ with the hydroxyl group (-OH) at the 3' position of another, forming a linkage called a phosphodiester (P-O-C) bond (Figure 1.5). The two-unit polymer resulting from this reaction still has a free 5' phosphate group at one end and a free 3' hydroxyl group at the other, so it can be linked to other nucleotides on each end. In this way, millions of nucleotides can be joined together in long chains to form single strands of DNA.

Watson & Crick

Although the components of DNA had been discovered, several studies were carried out to find the exact structure of DNA. One of the studies was by Erwin

¹ A dehydration reaction is a chemical reaction involving the elimination of a water molecule and the formation of a covalent bond linking two chemical species. (From Raven & Johnson, *Biology*, 4th ed. Copyright 1997 by The McGraw-Hill Companies, Inc.)

Chargaff. He discovered that the amount of adenine present in DNA always equals the amount of thymine, and the amount of guanine always equals the amount of cytosine, and that there is always equal proportion of purines (A and G) and pyrimidines (C and T). These findings are commonly referred to as Chargaff's rules.

Using these as clues, Rosalind Franklin and Maurice Wilkins, who were studying DNA at King's College in London, began to work on DNA's structure using an experimental approach. In 1953, Franklin, who usually worked alone, had finally

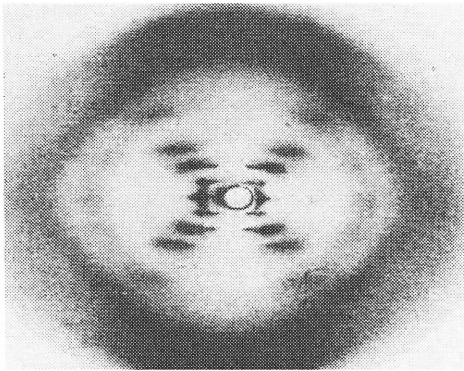


Figure 1.6: An X-ray diffraction photograph of DNA by Rosalind Franklin in 1953. (From Raven & Johnson, *Biology*, 4th ed. Copyright 1997 by The McGraw-Hill Companies, Inc.)

amassed X-ray diffraction data on DNA structure (Figure 1.6). In such experiments, a molecule is bombarded with a beam of X-rays. When individual rays encounter atoms, their path is bent or diffracted, and the diffraction pattern is recorded on photographic film. When carefully analyzed, they can yield information about the three-dimensional structure of a molecule. The X-ray

data showed the molecule to be helical (spiral-like).

At the same time, James Watson, a research fellow, and Francis Crick, a graduate student, from Cambridge University were also interested in the study of DNA. They wanted to figure out the accurate figure of DNA by building models. Knowing about Franklin's X-ray



Figure 1.7: In 1953, James Watson, a young American postdoctoral student (left), and the English Francis Crick (right) solved for the structure of DNA. (From Raven & Johnson, *Biology*, 4th ed. Copyright 1997 by The McGraw-Hill Companies, Inc.)

diffraction data before they were published, in 1953, they worked out a likely structure for the DNA molecule (Figure 1.7). They found the DNA structure to be a **double helix**,

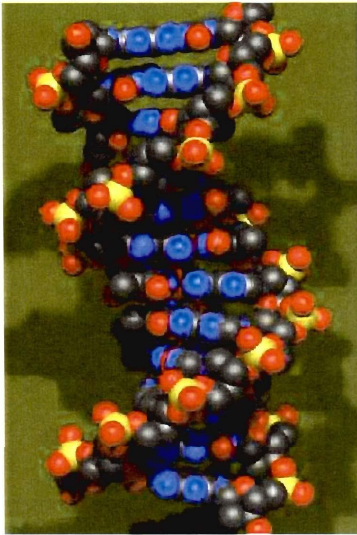


Figure 1.8: A three-dimensional view of DNA. (From Lundberg, Doug, 2000. <http://www.kaets.d20.co.edu/~lundberg/dnapic.html>).

which looks rather like two interlocked bedsprings (Figure 1.8). The backbone of each bedspring is composed of the sugars and phosphate groups, forming the outside of the molecule. In the inside, the two bedsprings (helices) are held together by hydrogen bonds, in which two electronegative atoms “share” a proton, between the bases. This forms what is called base pairs between two DNA strands. Each base-pair consists of one purine base and one pyrimidine base, paired according to the following rule: G pairs with C, and A pairs with T (figure 1.9). Because of

this specific pairing, the two strands of a DNA molecule are complementary: Each strand contains all the information required to specify the sequences of bases on the other.

Note also that the two backbones (strands) run in opposite directions. Thus, they are said to be **antiparallel** strands, defined by the 5' and 3' groups of deoxyribose. One strand runs in a 5' → 3' direction and the other in a 3' → 5' direction (Figure 1.10).

Double Helix

A double helix is the picture of DNA when viewed in three dimensions. Here, the bases actually form rather flat structures, and these flat bases partially stack on top of one

another in the twisted structure of the double helix. As seen in Figure 1.11, the stacking

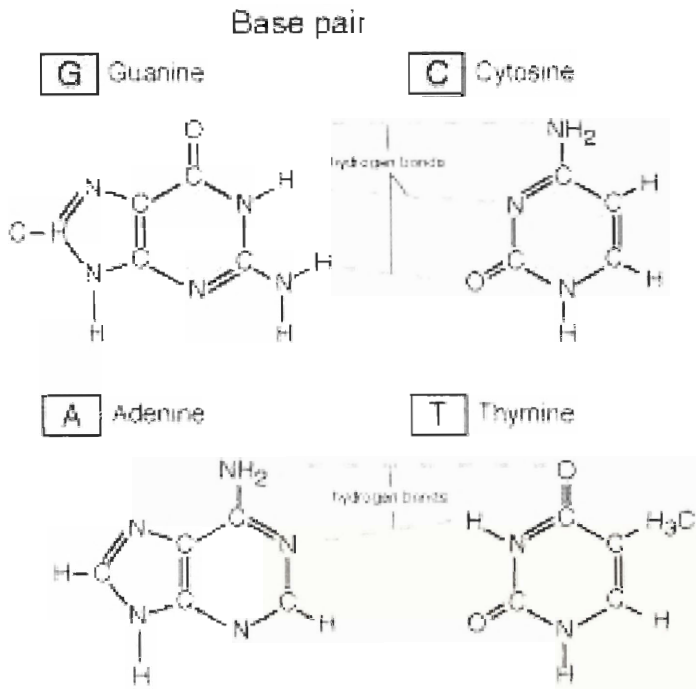


Figure 1.9: The base-pair rules indicate that there are only two pairs possible: G with C and A with T. (From *Graphics Gallery*, 2000. <http://www.accessexcellence.com/AB/GG/basePair1.html>.)

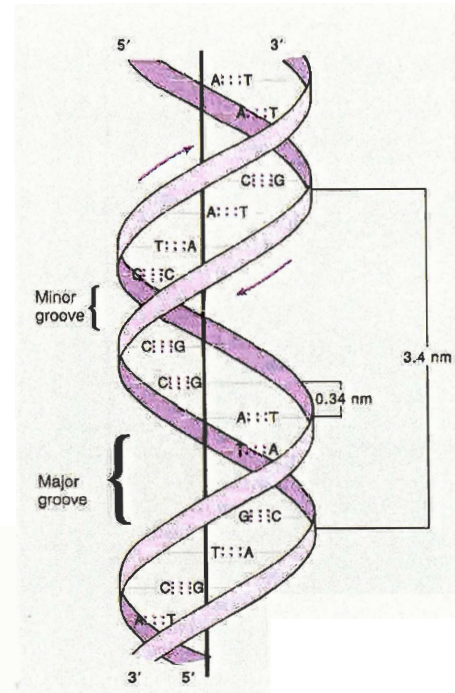


Figure 1.10: DNA strands are antiparallel. (From Raven & Johnson, *Biology*, 4th ed. Copyright 1997 by The McGraw-Hill Companies, Inc.)

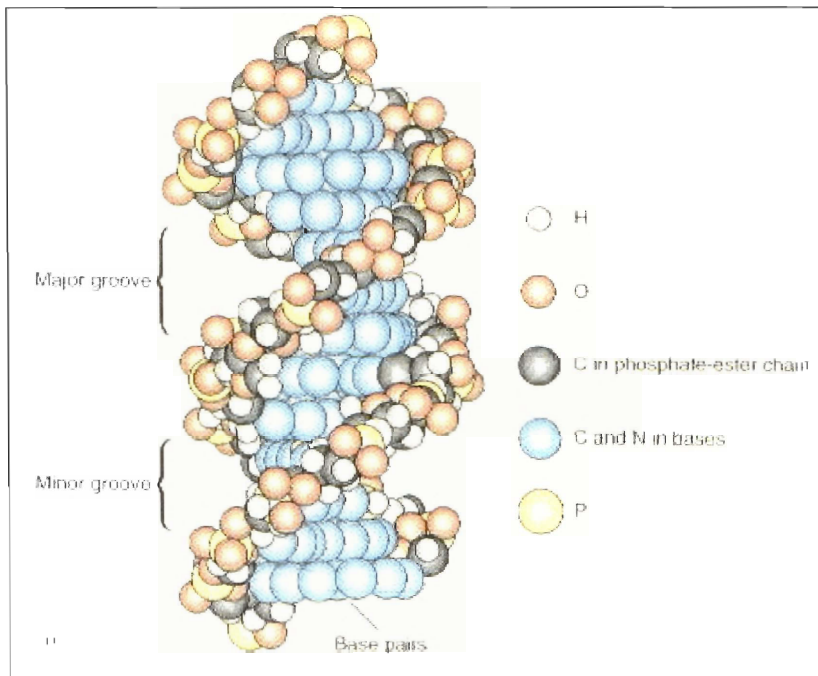


Figure 1.11: Stacking of base-pairs. (From Prescott, L. M., J. P. Harley, and D. A. Klein. *Microbiology*. 4th ed. Copyright 1999 by WCB/McGraw-Hill.)

of the base pairs results in two grooves in the sugar-phosphate backbones. These grooves are termed major and minor grooves.

The diameter of a helix is 2 nm (nanometer) and the distance between adjacent bases is 0.34 nm. The DNA molecule turns every 3.4 nm, thus, there are ten bases per turn of the helix.

However, it was subsequently realized that there were two forms of DNA called A form and B form. The A form of DNA is less hydrated, and the B form is more compact (Figure 1.12).

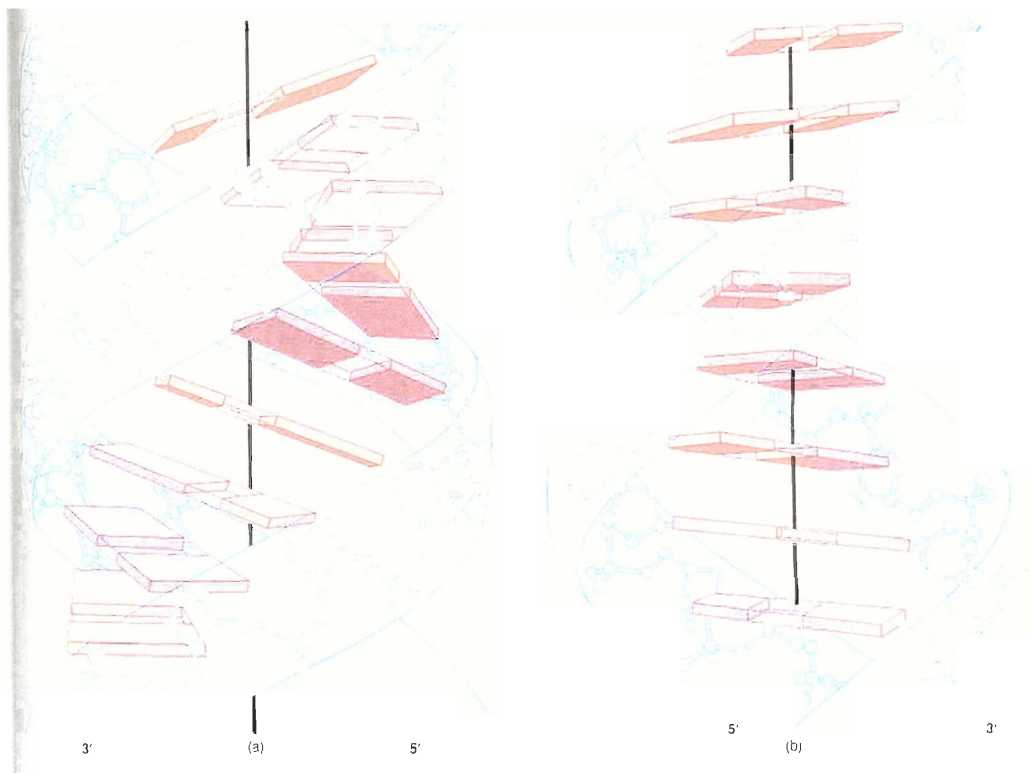


Figure 1.12: Schematic drawings of the structures of two forms of DNA. (a) A DNA. (b) B DNA. In A DNA, the base pairs are tilted and are pulled away from the axis of the double helix. In B DNA, on the other hand, the base pairs sit astride the helix axis and are perpendicular to it. (From A. J. F. Griffiths, J. H. Miller, D T. Suzuki, R. C. Lewontin, and W. M. Gelbart, *An Introduction to Genetic Analysis*, 6th ed. Copyright 1996 by W. H. Freeman and Company.)

CHAPTER 2: DNA FINGERPRINTING

Background to DNA Fingerprinting

Within the last century, detectives increasingly have turned to scientific evidence to help solve crimes. The distinctiveness of the human fingerprint was first described in 1892; soon to follow were the techniques for ABO blood typing¹ and leukocyte antigen tissue typing. DNA fingerprinting, first introduced in U.S. courts in 1988, was considered to be the greatest forensic advance since classical fingerprinting. So what exactly is DNA fingerprinting?

DNA fingerprinting is based on an observation made in 1984 by Alec Jeffreys. Jeffreys (Figure 2.1) found something rather odd in the non-coding region of the human genome: multiple copies of short nucleotide sequences, 3 to 30 base pairs long, repeated one after another 20 to 100 times. These groups of repeat sequences, called mini-satellites or VNTRs (variable number of tandem repeats), are now known to be widely scattered throughout the human genome. Everyone has these repeat units in their DNA, but the number of these regions at different **loci**² is different in each individual. Only identical twins end up with the same numbers and patterns of VNTRs. The DNA code for each individual is as unique as fingerprints, and

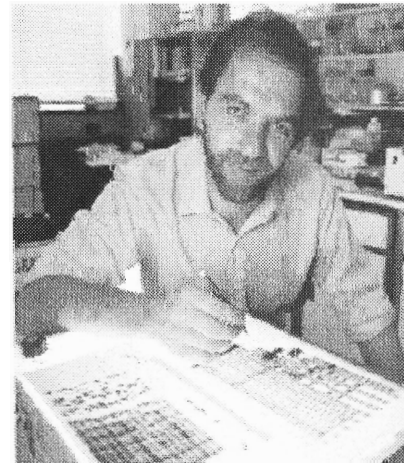


Figure 2.1: The “father” of genetic fingerprinting Alec Jeffreys. (1999. Alec Jeffreys – Genetic Evidence. <http://www.britcoun.org/science/science/personalities/text/ukperson/jeffreys.htm>)

¹ABO blood typing can give some preliminary indications about possible paternity. Its use is limited.

² Loci (plural for locus) are specific positions of a gene in chromosome.

DNA fingerprints do not change with age and are not affected by food, drugs or medicine. Chemical examination of blood, semen, and other body fluids at a crime scene can render positive identification when compared with the DNA molecules of a suspect. Evidence gathered from this technique had not been uniformly adopted for use in American court systems by the late 1980s.

DNA fingerprinting is more advanced than the simple tissue testing. Tissue testing is a process to determine either the blood type or tissue type from tissue samples collected. However, tissue testing requires a relatively large amount of fairly fresh tissue. Another problem with tissue testing is that there are many people in a population with the same blood type or tissue type, thus in court the result can only exclude a suspect but cannot prove guilt. On the other hand, DNA fingerprinting can theoretically identify the guilty individual with certainty because the DNA sequence of every person is unique. Although this does not mean that no 2 person can have exactly the same DNA fingerprint, but the probability of two people have the exact same DNA fingerprint is extremely low. In most legal cases, the probability of two people having identical DNA fingerprints is between one chance in 100,000 and one in a billion. The exact figure depends on how many **markers** are compared and on how common those markers are in populations.

History of DNA Fingerprinting

DNA Profiling is the newest and most powerful technique in forensic science, paternity testing, animal and plant sciences, and investigation of wildlife poaching. In 1980 Wyman and White established the foundation of such technique with the

observation of a polymorphic DNA locus characterized by a number of variable-length restriction fragments called *restriction fragment length polymorphisms* (RFLPs). With this finding came the idea that it was possible to analyze the unique portion of human DNA.

In the next ten years, several other approaches to DNA typing were published. In 1985, Alex Jeffreys' discovery of *hypervariable minisatellite* DNA developed the first real experimental approach for assaying unique DNA since Wyman and White. In his analysis, Jeffreys noticed a repeated region of DNA. These repeated regions are referred to as *minisatellites*, and other similar regions are *hypervariable* because the number of repeats is different among regions. Using these repeated patterns and studying different types of DNA, Jeffreys concluded that these profiles appear to be unique to each individual. Different types of repeats were later used to produce a number of different probes useful for fingerprinting.

In 1986, two more approaches to DNA typing were published. Both of these techniques were based on Jeffrey's hypervariable minisatellite concept. Tyler used the repeated sequence of Y-chromosomes as a probe, and a dot-blot technique to distinguish male and female blood. He and colleagues also used human **Alu** repeat sequences to distinguish dried animal blood from human. Although these two approaches are quick and simple, it proved to be as effective as Jeffreys'.

In later years more techniques of DNA typing were defined. Lifecodes investigators used the hybridization of human genomic DNA with specific probes to show the low probability of different fragments of DNA having the same **alleles** at a single specific locus. The experiments showed that the probability lowers with the

hybridization of more probes. Nakamura (1987) described the individual loci where the alleles are composed of tandem repeats as *variable number of tandem repeats* (VNTR). Based on this characteristic, scientists produced a number of probes for unique profiling.

Where do you find DNA?

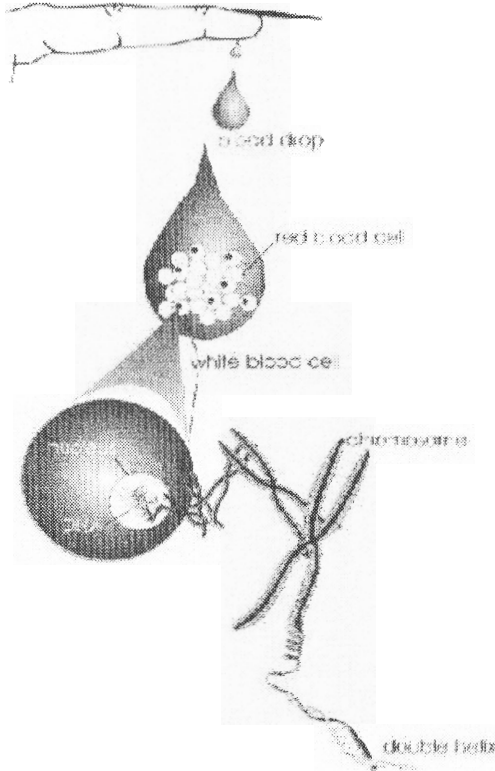


Figure 2.2: From blood to DNA. (From Inman Keith, and Norah Rudin. *An Introduction to Forensic DNA Analysis*. Copyright 1997 by CRC Press.)

The DNA fingerprint, of course is based on DNA. DNA can be found in **chromosomes** (Figure 2.2), which reside in the nucleus of the cells. It can be extracted from blood, semen, saliva, or hair of a person, using molecular chemical extraction, called Chelex extraction, or organic extraction before any type of testing can be carried out. The sample can be as much as four years old and still can be used in this fingerprinting method.

If a tiny amount of sample is to be extracted, Chelex extraction is often a preferred method. By boiling the sample with a solution

containing tiny chemical beads, Chelex, the cells are broken open and DNA is released. However, the Chelex must be removed along with other non-DNA components to yield only DNA. This technique will result in a single-stranded DNA since the two strands are broken apart in the process.

While Chelex extraction yields small pieces of single-stranded DNA, organic extraction results in larger and cleaner pieces of DNA. In this method, the sample (a

piece of cloth containing bloodstain, for instance) is cut into small pieces and soaked in a warm solution. From this step, the cells are released from any substance they are previously bound to. Then, another chemical mix is used (along with mild heat) to break open the cells, releasing DNA. DNA is later isolated by the use of several organic solvents, and is purified using special filtration method or precipitation, producing an extract of DNA (Figure 2.3).

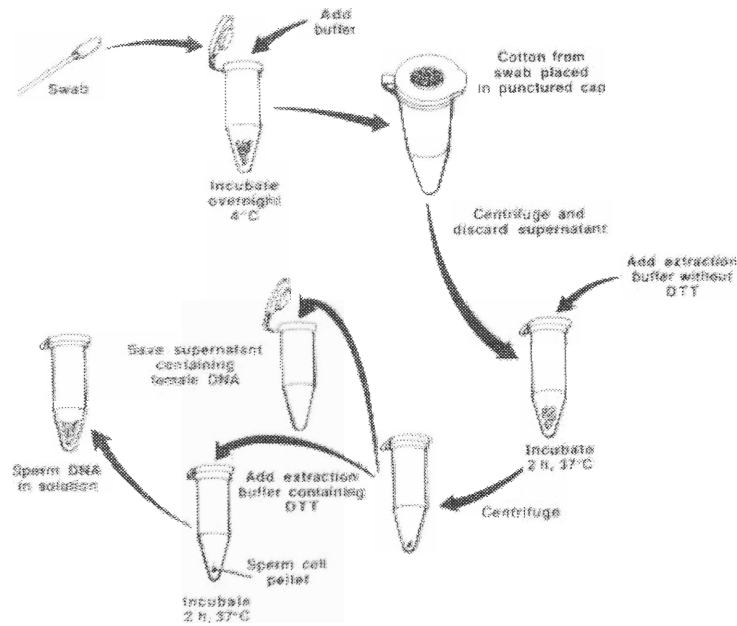


Figure 2.3: The procedure for isolation of DNA (From Kirby, Lorne T. *DNA Fingerprinting*. Copyright 1990 by Stockton Press.)

How to determine the quality and quantity of DNA obtained

Before any analysis can be carried out, it is very essential that the sample of DNA be observed to find how much DNA there is and how much it is degraded. If the sample consists mostly of large pieces of DNA, it is said to be of high molecular weight (HMW). If the Chelex extraction, which yields single-stranded DNA, has been used, then only the **slot blot** (Figure 2.4) method can be used to assess it. This method is a technique used to obtain information about the quantity of human DNA recovered from a sample. A small

portion of a sample is applied to a nylon membrane. A set of standard samples, for which

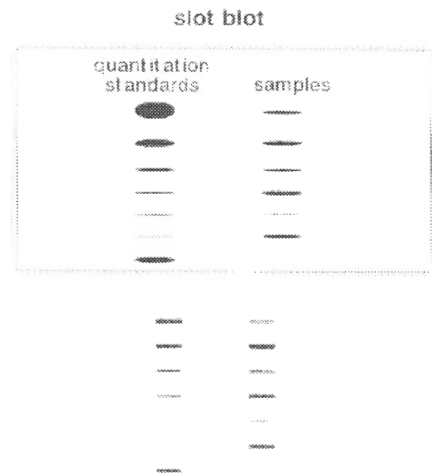


Figure 2.4: A picture of a slot blot. (From Inman Keith, and Norah Rudin. *An Introduction to Forensic DNA Analysis*. Copyright 1997 by CRC Press.)

the quantities are known are also applied for comparison. The samples are then probed with a small single-stranded fragment of DNA. This fragment is actually called **probe**. They are radioactively labeled and will hybridize with the samples. When exposed to X-ray film, a black band corresponding to the region detected by the probe is produced.

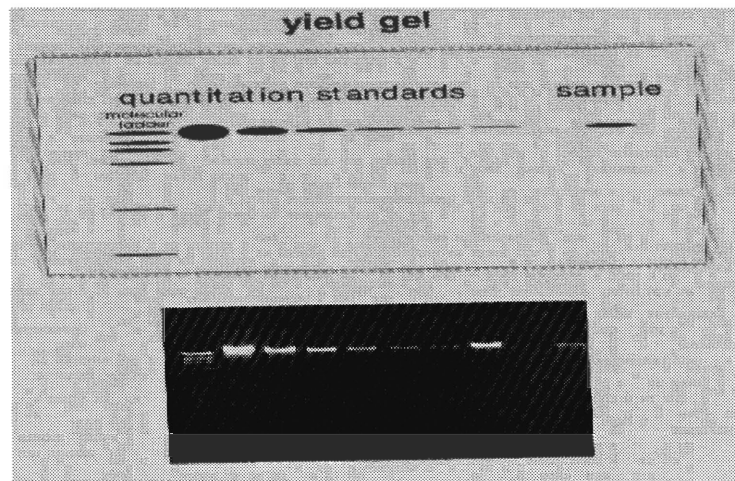


Figure 2.5: A picture of yield gel. The last lane is the sample. The intensity of the band is compared to the standards to quantify the amount of DNA in the sample. (From Inman Keith, and Norah Rudin. *An Introduction to Forensic DNA Analysis*. Copyright 1997 by CRC Press.)

Another method, called a yield gel (Figure 2.5), is used for quantitation of DNA. However, it requires double-stranded DNA in higher quantity for a result. Thus, it is an

alternative way to check the quantity of DNA when Chelex extraction is not used. Unknown samples, along with standard samples of known quantities, are loaded into separate wells of agarose gel and they are run for less than an hour. The gel is then stained with ethidium bromide. This substance binds to DNA and fluoresces under UV light. DNA viewed in this way is seen as a blob or a smear, depending on its state of degradation. Large, intact DNA molecules will form a compact band near the origin of the gel, similar in placement and shape to the standards. Degraded DNA of smaller sizes will migrate further in the gel and will form smear. Extremely degraded DNA might not be seen at all. Additional data on the state of degradation of DNA can be obtained by performing the slot blot together with a yield gel.

Analysis of DNA

a) RFLP Analysis

RFLP stands for restriction fragment length **polymorphism**. As the name implies, the technique makes use of different **restriction enzymes** to cut DNA at specific sites in the loci. It measures the size of DNA fragments produced by restriction enzymes. In order to be successful in this kind of analysis, the sample must contain sufficient HMW DNA. This technique requires more DNA than the other method called **PCR** (polymerase chain reaction).

To begin, the amounts of DNA, restriction enzyme(s), and other components are carefully calculated and combined in a small tube, which is incubated in a warm bath (usually at 37°C) overnight. Then it can be tested whether all specified sites are cut by using a digest gel (Figure 2.6). A comparison is made of the distance traveled by the

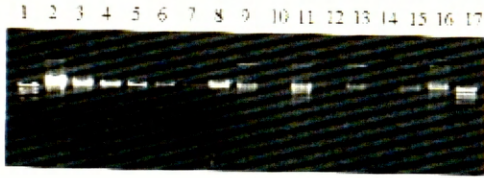


Figure 2.6: A yield gel. Lanes 1 and 17 contain a molecular ladder, that is, a marker. Lanes 2 to 7 contain HMW quantitation standards in decreasing concentration from left to right. Lane 8 contains the control sample. The samples to be observed are in lanes 9, 11, 13, 15, and 16. (From Inman Keith, and Norah Rudin. *An Introduction to Forensic DNA Analysis*. Copyright 1997 by CRC Press.)

completely digested standard samples and the uncut DNA. If the DNA is partially digested, meaning some sites are left uncut, another dose of enzyme is applied.

A blue dye is then added to the solution containing cut (completely digested) DNA. Then, each sample along with markers are loaded into the wells of an **agarose** gel (Figure

2.7). This agarose gel contains different sizes of microscopic pores depending on the concentration of agarose, a gel-like material. The bigger size the fragment is, the slower it travels through the agarose.

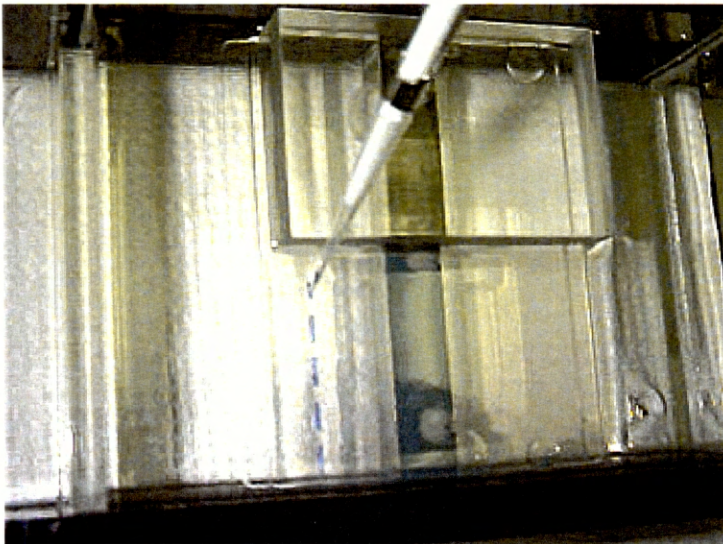


Figure 2.7: A gel and wells of agarose. Here, samples plus blue dye are loaded into the well. (From Mehrrens, Brad, Magaret Timme, Ray Zielinski, and Deanna Raineri. 2000. (<http://www.life.uiuc.edu/molbio/geldigest/electro.html#run>.)

Since DNA carries an overall negative (-) charge, an electric field is applied so that the DNA will move towards the positive (+) end. This technique is called **gel electrophoresis**. A blue dye is added to the DNA to allow the front edge of DNA to be seen as it travels (Figure 2.8). The gel is then stained in a chemical dye, ethidium bromide which stains all DNA, and observed under ultraviolet (UV) light.

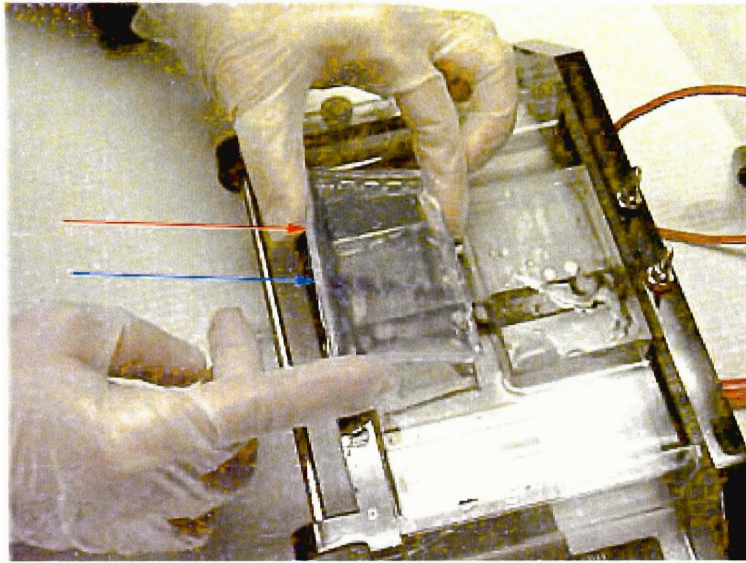


Figure 2.8: A finished gel with visible bands as indicated by the arrows. (From Mehrrens, Brad, Magaret Timme, Ray Zielinski, and Deanna Raineri. 2000. <http://www.life.uiuc.edu/molbio/geldigest/electro.html#run.>)

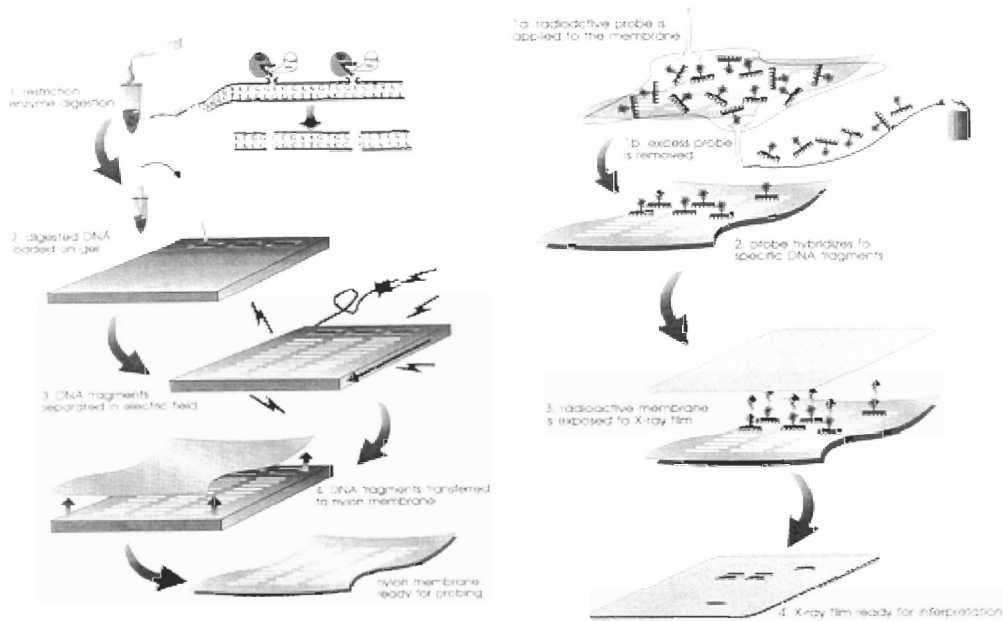


Figure 2.9: Process of RFLP (From Inman, Keith *An introduction to forensic DNA analysis*. Copyright 1997 by CRC Press LLC)

However, in order to detect specific polymorphic fragments of interest, the DNA double helix must be separated into single strands and transferred to a solid support called nylon membrane (Figure 2.9). This procedure is called **Southern blotting**.

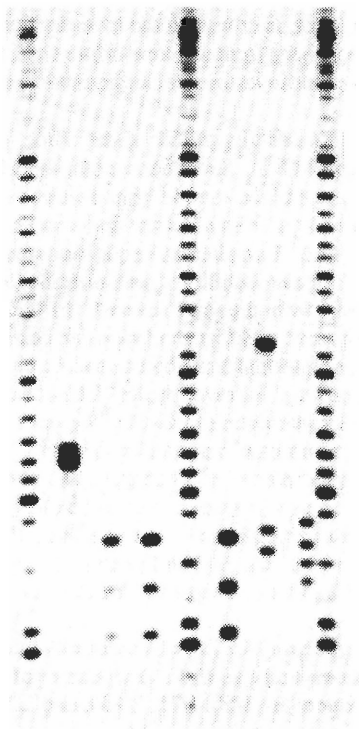


Figure 2.10: Autoradiograms of DNA-PRINT test results from a rape case. The lanes from the left to right are (1) marker, (2) suspect 1, (3) evidence from vaginal swabs, (4) suspect 2, (5) marker, (6) evidence from semen stains, (7) victim, (8) control, and (9) marker. (From Kirby, Lorne T. *DNA Fingerprinting*. Copyright 1990 by Stockton Press.)

The membrane is then bathed in a solution containing short, single-stranded DNA fragments called “probes.” These probes are labeled with a radioactive tag and they contain VNTR locus sequences. Under the right conditions, the probes and DNA fragments match and hybridize to yield double-stranded DNA of specific polymorphic fragments of interest. The signal from labeled probes allows the DNA to be visualized on a sheet of X-ray film, called autorad, short for autoradiogram or autoradiograph (Figure 2.10).

In each lane on the autorad, one to two bands will be present. If a person is **homozygous** for a particular locus (that is, he/she has inherited the same DNA sequence from both parents), only one band will show up. On the other hand, two bands will be detected if a person is **heterozygous** for that particular locus.

Forensic investigations commonly use different probes for four or five different loci on a single sample to improve the validity of the test. However, one probe must be completely stripped from the membrane before the next one is applied. The entire process may take anywhere from 3 weeks to 3 months to complete; a nerve-wracking time for defendant, prosecutors and defense attorneys. Unfortunately,

the RFLP requires many sample cells, like several strands of hair or large splatters of blood. The cells have to be “fresh” too, that is, undamaged and recently dead.

Although these techniques are standard practice in many laboratories, great care must be taken in carrying out DNA typing tests. Forensic samples of DNA are rarely pure. DNA from bacteria or fungi may show up in the fingerprint; dyes from denim can interfere with restriction enzymes; and proteins in the evidence sample can retard the migration of DNA fragments in gels, a problem known as “band shift”.

b) PCR Amplification

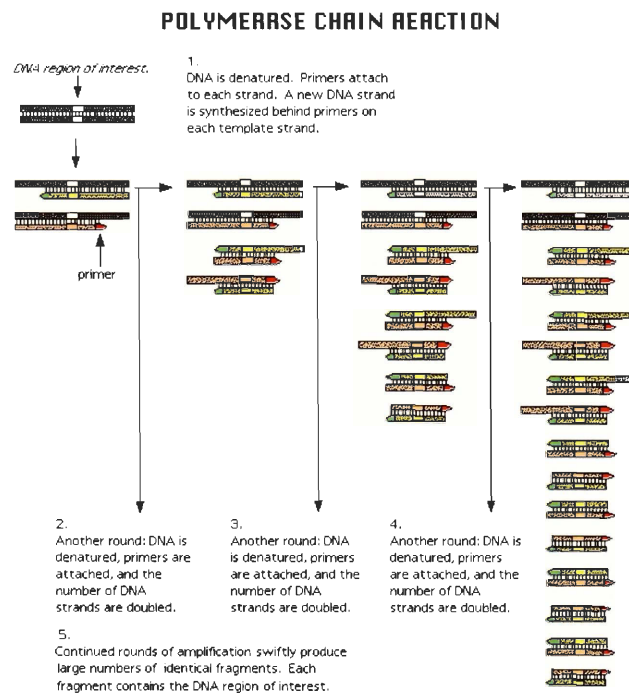


Figure 2. 11: Process of PCR. (1999. Graphic Gallery: Polymerase Chain Reaction. <http://www.accessexcellence.org/AB/GG/polymerase.html>)

If the forensic sample is too minuscule, or highly degraded to give reliable RFLP results, the polymerase chain reaction (PCR) can be used to obtain a DNA profile.

PCR is a molecular copying process used to amplify specific DNA sequences (Figure 2.11). With this technique it is possible to obtain enough DNA from a single hair follicle or a single sperm cell to determine an individual's DNA profile. The remarkable sensitivity of PCR is the procedure's main advantage.

To begin, DNA prepared by either Chelex or organic extraction is denatured (that is, separated into two single strands) by putting fairly high heat so that each strand can serve as a template for synthesis of a new strand. The next step involves annealing of DNA primers, short DNA sequences that have bases complementary to a starting location of the synthesis of the new DNA. The primers, together with *Taq* polymerase, an enzyme that adds nucleotides to the DNA chain, extends the DNA strands from 5' to 3' end. The end result is the duplicated segment of DNA of interest. This process is then repeated over and over again to generate million of copies identical to the original. The product can be checked using the gel electrophoresis method.

PCR can be used to selectively amplify DNA fragments containing either length or sequence polymorphisms. Sequence polymorphisms, as occur within the genes of the highly polymorphic HLA complex, are the result of single nucleotide base changes. This HLA DQ(A1 is the name of the locus in which some of its bases show differences between people. This variation in this region is detected using probes. These probes

(E) Final form of Manufacturer's Reverse Dot Blot DQα Typing Strips

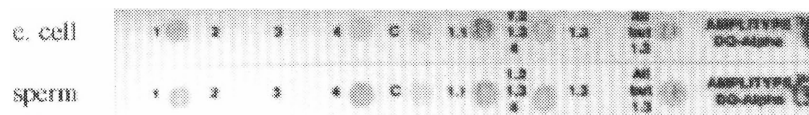


Figure 2.12: A picture of commercial reverse dot blot strips. Here, the upper strip is from the victim's vaginal cell and the lower strip is from the suspect's sperm. (From Inman Keith, and Norah Rudin. *An Introduction to Forensic DNA Analysis*. Copyright 1997 by CRC Press.)

have already been attached to a nylon strip. The strips are commercially available and contain specific DNA sequences. The probes used are able to detect six common DQ alleles. The final results are seen as a series of blue dots on a nylon strip, the format known as **reverse dot blot** (Figure 2.12). A comparison of the pattern of dots between the strips indicates whether two samples may have come from the same source. The strips however are usually discarded because the blue color fades when exposed to light. Thus, the results are photographed before the strips are thrown away.

Length polymorphisms are exemplified by the variable number of tandem repeat loci (VNTR). Length variation at a given VNTR locus is detected by size-fractionation of PCR products in polyacrylamide gels. This type of gel is more appropriate than agarose gel used in RFLP analysis because of the small size of PCR products. PCR-amplified DNA products along with standard markers can be directly visualized by staining after electrophoresis, eliminating the need for radioactive probes. The gel is then dried to be kept as a permanent record.

In each lane of the samples, like RFLP analysis only one to two bands will be present. The standard lane will result in a ruler-like picture in which all the bands represent all possible alleles within a particular locus.

In DNA samples of length polymorphisms, the most common VNTR loci used are called D1S80, STRs, and amelogenin. Amelogenin is the gene for tooth pulp and it shows a length variation between the sexes. By examining it, a forensic investigator will know whether the samples are from male or female.

DNA Fingerprinting using VNTR's

On some human chromosomes, a short sequence of DNA has been repeated a number of times. In any particular chromosomes the repeat number may vary from one to thirty repeats. Since these repeat regions are usually bounded by specific restriction enzyme sites, it is possible to cut out the

segment of the chromosome containing this variable number of tandem repeats or VNTR's, run the total DNA on a gel, and identify the VNTR's by hybridization with a probe specific for the DNA sequence of the repeat.

Shown in Figure 2.13 at the top panel are the chromosomes of the two parental individuals of the pedigree from the lower panel. The first individual has one chromosome with 4 repeated sequences and one chromosome with 6 repeated sequences. The other individual has one chromosome with 3 repeated sequences and one chromosome with 5 repeated sequences.

At the bottom of the figure is a pedigree of the mating between these two individuals and their four children. The DNA of each of the individuals has been analyzed for the VNTR repeat number, and the gels are shown below each individual along with the genotype for each individual. Notice that each of the six children are distinguishable from each other by the VNTR's at this one genetic locus.

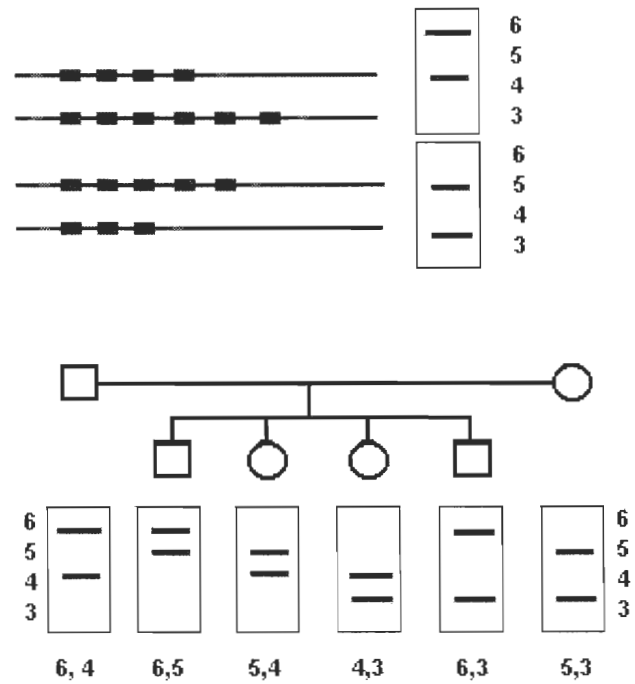


Figure 2.13: Sample Chromosomes. (Huskey, Robert. June, 1998. www.people.virginia.edu/~rjh9u/gif/vntr1.gif)

Although the PCR test isn't quite as accurate, it takes much less time to complete, a week at most. The test can be performed with minute crime scene samples too, which helps investigators who have little physical evidence. The DNA doesn't have to be recently collected, the PCR test can still be performed even years or decades after the fact, and still be just as accurate. This is because the DNA can be partially degraded and still have fragments long enough to give a PCR signal.

CHAPTER 3: APPLICATIONS OF DNA FINGERPRINTING

Nearly a decade has passed since DNA typing methods were first used in criminal investigations and trials. Law enforcement agencies have committed substantial resources to the technology; prosecutors, defense counsel, and judges have struggled with the terminology and ideas of molecular biology, genetics, and statistics. In 1992, a broad-ranging report released by the National Research Council attempted to explain the basics of the relevant science and technology, to offer suggestions for improving forensic DNA testing and its use in law enforcement, and to quiet the controversy that had followed the introduction of DNA profiling in court. Yet, the report did not eliminate all controversy. Indeed, in propounding what the committee regarded as a moderate position the ceiling principle, the report itself became the target of criticism from scientists and lawyers on both sides of the debate on DNA evidence in the courts. Moreover, some of the statements in the 1992 report have been misinterpreted or misapplied in the courts.

DNA analysis promises to be the most important tool for human identification since Francis Galton developed the use of fingerprints for that purpose. We can confidently predict that, in the not-distant future, persons as closely related as brothers will be routinely distinguished, and DNA profiles will be as fully accepted as fingerprints now are. But that time has not yet arrived, and the winds of controversy have not been stilled. In the early days there was doubt, both as to the reproducibility and reliability of the methods, and as to the appropriateness of simplistic calculations that took no account of possible subdivision of the population. Despite the potential power of the technique, there were serious reservations about its actual use.

Ceiling Principal

In 1989, the National Research Council formed the Committee on DNA Technology in Forensic Science to study this new technique. The committee issued its report in 1992. The report resolved a number of questions, and several of its recommendations were widely adopted. Nevertheless, it generated controversy and criticism. Much of that centered on the "interim ceiling principle," a procedure intended to provide an estimate of a profile frequency that is highly conservative (i.e., favorable to the defendant) and independent of the racial origins of the DNA. The principle was criticized as being arbitrary and unnecessarily conservative, as not taking population genetic theory into account, and as being subject to misuse.

Since the DNA can be used to both free and convict suspects, the accuracy of this type of evidence is a crucial issue to address, and is often looked at from varying viewpoints. For example, if a guilty suspect left incriminating DNA evidence, then the defense would attack the credibility of such evidence, highlighting the fact that there is a chance that the evidence is false. The test used to obtain information about the crime scene DNA will be chosen according to age of the sample, amount available, purity of sample, etc. Upon obtaining results of the given test, the next responsibility is to decide the innocence or guilt of the suspect. In accordance with our judicial standard of "innocent until proven guilty," forensic DNA analysts take into account the probability that two samples match coincidentally and follow this basic principle: If crime scene sample does not match reference material from suspect, then the individual is acquitted. If crime scene sample does match reference material from suspect, then there exists a possible culprit. The guilt must be further verified through other criminal evidence.

An easy misconception to fall into when studying forensic DNA is the idea that in DNA fingerprinting, the forensic scientists compare entire DNA strands. As Dr. Bruce Weir states, "We know that even genetically identical twins have different fingerprints. In contrast, DNA typing, because it uses a very small fraction of the DNA, is certainly not unique."

Geneticists do not match a person's entire genetic code to the code contained in a crime scene sample. That's because scientists have not yet mapped the entire human genome. And even if they had, crime labs would not have the computing power to run such complete tests.

The lawyers' caution reflects the DNA statistical ambiguities, but the accuracy of a DNA identification rarely is directly challenged anymore. Instead, defense attorneys more frequently contend that lab results are skewed by the crime scene sample being mishandled, degraded or contaminated to the point where test results are not trustworthy.

Following are the factors that might influence the estimated accuracy of DNA fingerprinting: What if scientists are given a blood sample that has been mixed with dog hair, gum, dirt, bacteria, and other factors that might throw off DNA analysis? Well, it turns out that with proper skill (and the use of probes), scientists can look specifically for human DNA and ignore other DNA. However, the problem that arises is that many DNA labs do not have scientists or equipment of this caliber: "...only one-third of 60 police department DNA labs have been accredited by the American Society of Crime Directors Laboratory Accreditation Board. That's more than the commercial DNA labs, which almost all operate without accreditation."

It is, of course, possible to make a mistake in the laboratory. Since the procedures

are fairly straightforward, however, the chances of this happening are minimal. Instead of attempting to ascertain an estimation of how much laboratory error affects DNA fingerprinting, the National Research Council's report proposed recommendations to improve laboratory performance and accountability.

The two procedures used to identify DNA, PCR and RFLP, have a large impact on the accuracy of DNA comparison. Although PCR is a quick, overnight process, it allows typically only three or four variants, while RFLP has 30 distinguishable types. Furthermore, PCR is more prone to contamination than RFLP because of the small samples that are used.

Still, geneticists do not declare that the DNA in a crime scene sample and the genetic material provided by a suspect absolutely are one and the same. Rather, they describe it as a random-match probability. That is an estimate of the chance that a randomly selected person will share the DNA pattern present in both the defendant and the crime scene sample. In most cases, the random-match probability is one in a million, one in a billion or higher that a crime scene sample originated from someone other than the suspect whose DNA matched. DNA evidence can help put someone at the scene of a crime, but it is usually one small piece of a large puzzle, a puzzle of evidence. It is possible to get absolutely no results whatsoever.

DNA Database

Huge databases storing DNA sequences are maintained with the hopes that statistical DNA accuracy estimations can be improved. The National Center for Biotechnology Information (NCBI) controls GenBank--a database that has gathered DNA

from around the world.

Intuitively, one may think that DNA structures within races and ethnic groups might be more similar than those of different racial and ethnic backgrounds. The ceiling principle was a concept that conservatively embodied this concern. At the time, scientists were unsure of the extent of the role that this principle played in matching DNA sequences. The Research Council's 1992 report on DNA technology in forensic science concluded that scientists did not know enough about these variations to apply them with confidence when calculating the likelihood of a match. The report advanced a conservative, interim formula to be applied to any suspect, regardless of that person's racial or ethnic background. This formula was designed to be favorable to the defendant, to lessen the chances of a false conviction.

However, during the years between 1992 and the present, statistical data has been obtained and stored in databases that allows for better estimations of the effects of racial and ethnic backgrounds in DNA matching. Consequently, the latest report has specific formulas for calculating the accuracy of DNA within a given racial domain. The report states that the technology for DNA profiling and the methods for estimating frequencies and related statistics have progressed to the point where the reliability and validity of properly collected and analyzed DNA data should not be in doubt.

CHAPTER 4: DNA IN THE COURTROOM

With the advancement of DNA typing techniques, the courtroom soon became the proving ground of the feasibility and accuracy of DNA analysis, as this technique is frequently used in murder and fraternity cases. In some instances DNA evidence is the sole determinant in the ruling of the case, for DNA analysis often effectively shows the accused is guilty or not guilty. In other instances, the DNA typing results are either insufficient or flawed in ways such that it cannot be used as appropriate evidence in constructing a verdict. In these cases, the DNA evidence is thrown out by the court. In the following chapter we examine several cases where the use of DNA evidence are both effective and ineffective.

Colin Pitchfork

The first reported use of DNA as evidence in criminal case was the case that happened in Leicestershire, England, in 1983. A schoolgirl disappeared while she was making her way home. She was raped and murdered. There was an investigation, but with the lack of sufficient evidence, the case remained unsolved. Three years later, in July 1986, another schoolgirl was also raped and murdered. Shortly after the second murder, the local police arrested a seventeen-year-old boy who worked for the local mental institution. He confessed to the second murder, but denied any knowledge of the first. However, the police believed that both murders were committed by the same man. Thus, they needed to find some way to establish the suspect's guilt. The police then decided to send forensic samples from the victims, and a blood sample from the suspect, to Dr. Jeffreys who invented the technique of DNA fingerprinting. DNA fingerprinting

patterns of sperm DNA from semen recovered from the victims did not match the pattern from a blood sample from the suspect. However, it was found that both semen samples appeared to have derived from the same man. As a result, the police released the boy. He was the first person to be proved innocent by DNA fingerprinting.

As for the police, the search for the murderer went on. They had the evidence that suggested that the murderer was a young male who lived in the district where the girls had been murdered. With this information in hand, they began the world's first DNA-based manhunt, and requested, on a voluntary basis, a small sample of blood from all men in the vicinity between the ages of 18 and 30. More than 5,000 blood samples were collected from three villages. People who refused to test, however, were investigated by the police. Of the total samples, approximately 40 percent could not be excluded by conventional blood typing and were DNA profiled. The fingerprint results were discouraging because none of them matched those from the stains found on the victims.

Around the town, rumors about the killer were well circulated. In a local pub, a man was drinking and mentioned his feat that he had donated two blood samples, one in the name of his coworker, Colin Pitchfork, who was unable to donate. The police were notified, and Pitchfork was arrested in 1987. His DNA was analyzed and found to perfectly match those recovered from the victims' bodies. The suspect confessed to both crimes and is now serving two life sentences. He became the first person to be convicted using DNA.

Ronald Cotton

In two separate incidents in July 1984, an assailant broke into an apartment, destroyed the phone wires, raped a woman, and searched through her belongings, taking money and other things.

On August 1, 1984, Ronald Cotton was arrested for both incidents. After two trials, an Alamance County Superior Court sentenced Cotton to life plus 54 years. Cotton tried to prove that he was elsewhere while the crimes happened. His words were supported by family members. In addition, the jury was not allowed to hear evidence that the second victim failed to pick Cotton out of either a photo array or a police lineup. Therefore, the prosecution on the case was based on: a photo identification made by one of the victims, a police lineup identification made by one of the victims, a flashlight in Cotton's home that resembled the one used by the assailant, and rubber from Cotton's tennis shoe that was consistent with the rubber found at one of the crime scenes.

However, Cotton's attorney filed an appeal and the North Carolina Supreme Court overturned the conviction because the second victim had picked another man out of the lineup and the trial court did not allow this evidence to be heard by the jury. Thus, in November 1987, Cotton was retried. This time it was for both rapes and the second victim had picked Cotton as the assailant. Before this second trial though, a man already in prison said to his inmate that he did the two rapes that Cotton now was tried for. However, the superior court judge refused to allow this information into evidence, and Cotton was convicted of both rapes and sentenced to life.

In 1994, two new lawyers took over Cotton's case. They filed a request to have a DNA test on Cotton. The request was granted in October 1994. In the spring of 1995,

the Burlington Police Department turned over all evidence that contained the rapist's semen for DNA testing. The samples from one victim were too deteriorated to be conclusive, but the samples from the other victim's vaginal swab and underwear were submitted to PCR testing. The fingerprint did not match that of Cotton's. At the defense attorney's request, the results were sent to the State Bureau of Investigation's DNA database containing the DNA patterns of convicted, violent felons in North Carolina prisons. The state's DNA database showed a match with the convict who had earlier confessed to the crime. Thus, on June 30, 1995, Cotton was officially cleared of all charges and released from prison.

Edward Honaker

In the early morning of June 23, 1984, a man pretending to be a police officer approached a woman and her boyfriend who were sleeping in their car. He told the boyfriend to run into the woods. Then, he forced the woman into his truck and drove into a secluded area. He repeatedly raped her. While the police were trying to sketch the picture of the rapist from the victim and her boyfriend, a woman was also raped 100 miles away, near Edward Honaker's house. She said the attacker looked like Honaker but Honaker had witness where he was during the second rape. Thus, he wasn't charged for the second one. The detective, however, took a picture of Honaker and showed it to the first victim and her boyfriend.

Edward Honaker was tried and the jury found him convicted of seven counts of sexual assault, sodomy, and rape. He was sentenced to three life sentences plus 34 years. The prosecution was based on several points, including the one that a State laboratory

forensic specialist testified that hair found on the woman's shorts was "unlikely to match anyone" other than Honaker.

Honaker's lawyers, however, filed a motion with the State of Virginia to release evidence for DNA tests because Honaker's 1976 vasectomy was barely mentioned in the trial (and not known by the prosecution's criminalist). The evidence was sent to Forensic Science Associates (FSA) for PCR testing. The complication came about because the victim claimed that she had a secret lover at the time of the original incident. Thus, the DNA tests had to prove that one of the stains was not from Honaker or either boyfriend in order to establish Honaker's innocence.

The first report from FSA came out on January 13, 1994. The report included DQ alpha typing of a vaginal swab from the rapekit, an oral swab from the victim, a semen stain from the victim's shorts, and a blood sample from Honaker. The report indicated that there were two different seminal deposits (the one on the swab and the one from the shorts did not match). FSA then asked for blood samples from the victim and the boyfriend. The report stated that even if Honaker was able to produce sperm, he was eliminated as the source of sperm from both deposits.

On March 15, 1994, the second report was written. It included the boyfriend's typing and verified the victim's DQ alpha. Honaker and the boyfriend were both eliminated as the source of sperm on the vaginal swabs but the boyfriend could not be eliminated as a sperm donor found on the shorts. The Virginia State laboratory then tested the secret lover, or second boyfriend, and could not exclude him as the sperm source on the vaginal swab.

FSA then repeated the DQ alpha typing from all the evidence. This time five additional polymorphic genes were typed, too. The report indicated that neither the boyfriends nor Honaker could have accounted for the sperm from the vaginal swab. Therefore, Honaker was released on October 21, 1994. He had served 10 years of his sentence.

The O. J. Simpson Trial

O. J. Simpson was one of the most notorious figures of American history. With a natural talent for football, he was the recipient of the Heisman Trophy of 1968 as the best college football player of the year. He played professional football until 1979 when injuries forced him to retire. This great American hero also cast his fame as a sportscaster and movie actor. For most people his achievements would be more than gratifying for a lifetime. However, for Simpson the highlight of his life did not come until age 47 (1994) when he was charged with the brutal murder of his wife Nicole Brown Simpson and her friend Ronald Goldman, and underwent years of controversial trials and media frenzy. This “trial of the century” became an important landmark for the controversial discussion in the quality of DNA sample handling.

Orenthal James Simpson was born in San Francisco, California, in 1947. He attended University of Southern California, where he built his future career as a professional football star by being selected as an All-American in 1967 and 1968. As the first person selected in the professional football draft of 1969, he played for the Buffalo Bills until 1977. He then went to play for the San Francisco 49ers for two years until his retirement due to injuries.

Simpson was among the greatest heroes in American sports. With a total of 11,236 yards rushing during his career, he set season records for most yards gained (2,003; 1973) and most touchdowns (23; 1975). He still holds the season record for yards rushing per game (143; 1973). Although Simpson was frequently selected to the Pro Bowl, he never attended the Super Bowl with the Bills. His frustration led to his trade to the 49ers in 1978 where he stayed for two years. Then he retired from Professional football and became a sportscaster and an actor². In 1985 Simpson was inducted into the Pro Football hall of Fame.

In 1994, the world was thrown into media frenzy by the murder of Simpson's former wife Nicole and a friend of hers, Ronald Goldman. Simpson was immediately charged with the murders. A trial was opened in January of 1995. Besides the racial relations, police procedures, and domestic abuse issues involved, the effectiveness of DNA analysis used by both the defense and prosecutors was put on trial, and perhaps for the first time in history the feasibility and accuracy of DNA tests were exhaustively investigated by a court of law.

The Murder Scene (A Simple Debriefing)

On June 12, 1994, Nicole Brown Simpson and Ronald Goldman were stabbed to death. Nicole Simpson was found on the ground lying on her left side facing down (Figure 3.1), just outside the threshold of her open gate. Her buttocks were positioned against the first riser of the four steps leading up to the level of her condo, with a large pool of blood on the ground directly under her head. Ron Goldman was found not far away next to a fence (Figure 3.2). The crime scene examinations determined that both

² * O. J. Simpson was well known for his role as a police investigator in The Naked Gun series.



Figure 3.1: Simpson Crime Scene. (CNN; www.cnn.com/us/oj/index.html)

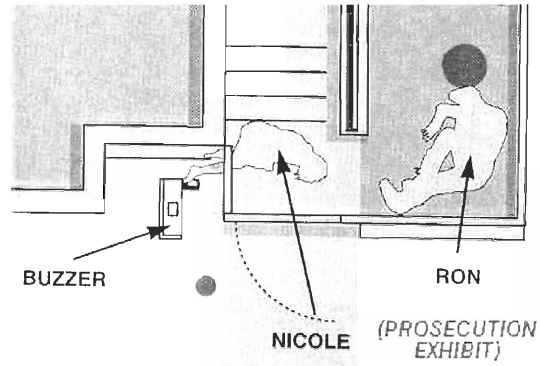


Figure 3.2: Top View of the Simpson crime scene. (www.wagnerandson.com/oj.oj.html)

Nicole and Ron died of open slash wound on their throats. There was a great pool of blood from her slashed throat. Some of the blood flowed down the small inclined surface away from the stairs toward the street. There was also a large amount of blood behind her toward the stairs. Blood was found on the first step, indicating that Nicole's throat might have been slashed on or over the first step.

The Arrest and Charge (CNN reports)

On the date of the murders, O.J. Simpson fled to Chicago for promotional events just before midnight on Sunday. The bodies were found after midnight, Monday morning. Simpson was contacted by the Los Angeles Police Department and fled back to LA. Thirty minutes after he arrived at home Simpson was arrested by police and taken in for questioning. On Friday, June 17, O.J. Simpson was officially charged with two counts of murder with special circumstances.

The Evidence (CNN reports)

Various photographs of Nicole and Simpson were presented as evidence in the

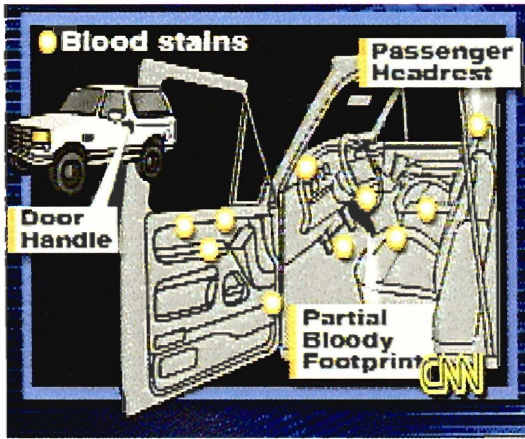


Figure 3.3: Inside Simpson's Vehicle. (CNN; www.cnn.com/us/oj/index.html)

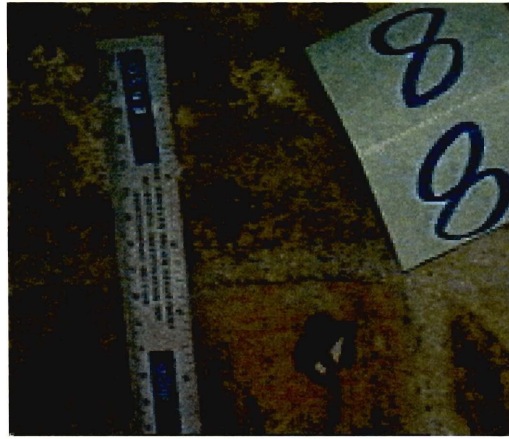


Figure 3.4: Blood found on O.J. Simpson's sidewalk. (CNN; www.cnn.com/us/oj/index.html)

trial. Audio tapes of Nicole's 911 call to police in October 1993, complaining of Simpson attacking her were also used to solidify Simpson's motive for killing. There were a total of 57 items, all of which had no direct physical connection to the murder.

Blood stains were found on O.J. Simpson's white Ford Bronco (Figure 3.3), a pair of socks in his bedroom, a pair of black gloves with one found at the crime scene and one at Simpson's estate, Simpson's shoes, Simpson's driveway (Figure 3.4), and in his house (Figure 3.5). All of this evidence was enough for the LA district attorney to charge Simpson with first-degree murder.



Figure 3.5: Blood stains in Simpson's foyer. (CNN; www.cnn.com/us/oj/index.html)

Importance of Blood

The most important aspect of the investigation was the blood pool at Nicole's estate and the bloodstains found on O.J. Simpson's properties. The result of blood typing



Figure 3.6: Blood sample in test tubes.
(CNN; www.cnn.com/us/oj/index.html)

provided the police of a match that linked O.J. Simpson to the crime scene, and caused beyond a reasonable doubt for the arrest of O.J. Simpson. Samples of blood from each of the blood-related evidence were sent to a Los Angeles crime lab for DNA testing (Figure 3.6). The results from these tests were used as evidence by the prosecution to incriminate Simpson.

During the investigation of the murder, Dr. Robin Cotton, lab director of Cellmark Diagnostics, tested blood samples found at the bottom of Simpson's shoe. Dr. Cotton used RFLP typing to identify the DNA. According to the transcripts³ the testing provided positive results for the prosecution.

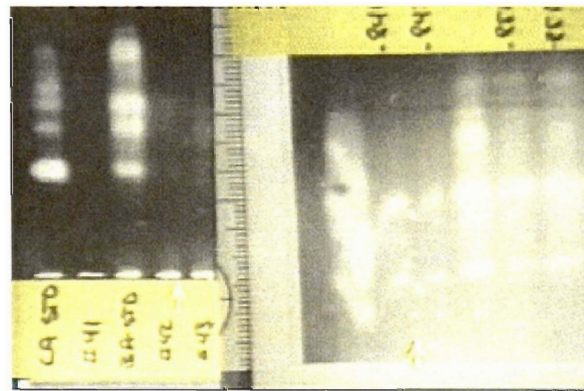


Figure 3.7: Defense Exhibit of DNA autoradiograph.
(CNN; www.cnn.com/us/oj/index.html)

MR. CLARKE: Based on this first test or first typing process, using this RFLP method, what conclusions were you then able to reach?

DR. COTTON: For that sample, the conclusion would be that you do have two people there and the possible contributors from the known individuals that we have would be the major amount of DNA coming from or consistent with Nicole Brown, and the minor amount of DNA

³ All excerpts from the trial transcripts to the O.J. Simpson's murder trial was obtained on the world wide web at <http://207.175.199.183/~walraven/simpson/#transcripts>.

consistent with some of the bands in Mr. Goldman's pattern. There are not all of his bands there in this cocktail; there are only three.

Mr. Clarke was an outside prosecutor and a DNA expert with the Deputy District Attorney. Above he asks Dr. Cotton of the results of the DNA analysis using RFLP typing method, in which the doctor replies with a confirmation that the blood on the bottom of O.J. Simpson's shoe matches those of Nicole and Mr. Goldman. According to the transcripts Dr. Cotton also performed other types of DNA typing methods. One was PCR testing, which he claims shows similar results as that of RFLP testing.

MR. CLARKE: With what results as to this particular item and based on PCR typing?

DR. COTTON: The results are basically the same as the RFLP, that Nicole Brown and Mr. Goldman can be included; Mr. Simpson is excluded.

Based on Dr. Cotton's testimony, the DNA contained in the blood stain found on the bottom of O.J. Simpson's shoe was an undeniable match to that of Nicole and Ronald Goldman's DNA. In any trial such confirmation should be more than enough for the jury to construct the verdict of guilty. However, this murder case is one with special circumstances in which such a simple conclusion cannot be the sole determinant factor in the final verdict.

It was established in previous chapters that DNA fingerprinting is only as accurate as the samples and test procedures allow. If either the samples or the tests are flawed or "impure" then the DNA results are not accurate.

John Gerdes, a Denver microbiologist testifying for the defense, said he wouldn't trust the results of tests on blood found in Simpson's Ford

Bronco because of sloppy handling and lax security, cast doubt on the quality of the evidence collected.⁴

This constriction of DNA typing gave the Defense a good opportunity to over throw the DNA results as proper evidence in this trial. After days of argument and examination, the Defense provided adequate proof that the blood and DNA samples were not properly handled from the crime scene to the lab and during the test process. The Court ruled that the DNA evidence may not be submitted as evidence.

RULING: DNA EVIDENCE HEARING

The court has read and considered the defendant's Notice of Objections to Testimony Concerning DNA Evidence and Memorandum in Support Thereof, and the People's Response to Notice of Objections to Testimony Concerning DNA Evidence and Memorandum in Support Thereof. In his notice, defendant seeks to notify the court of "defendant's position regarding the legal requirements for admissibility of DNA evidence, the foundational showings that are necessary to meet those requirements, and the types of testimony concerning DNA evidence that are legally improper and inadmissible."

Due to this ruling, which might possibly be the most important factor in this case, along with the lack of other evidence, the Supreme Court found O.J. Simpson was not guilty on the two counts of first-degree murder.

⁴ Duetsch, Linda. (1995, August 4). The Detroit News.

JonBenet Ramsey Murder Case



Figure 3.8 : ABC news. 2000, April.
(http://abcnews.go.com/sections/us/DailyNews/jonbenetgrandjury_feature.html)

Nearly 2,000 children were murdered in the United States last year, but only the Christmas night killing of a blond 6-year-old from Boulder has transcended obscurity to become obsession. In the 12 months since JonBenet Ramsey's (Figure 3.8) death, careers and reputations have been made or tarnished or lost. Family friends have become famous or

infuriated, media darlings or door slammers.

John and Patsy Ramsey and their 10-year-old son, Burke, never returned to the house where JonBenet died. In the weeks after the murder, they grieved behind other people's doors, and finally left Colorado for a new home near family and longtime friends in Atlanta. Suspicion and speculation have followed them, along with reporters and photographers for supermarket tabloids, which have seen sales rise every time JonBenet's face appears on a cover.

In the wake of JonBenet's murder, attorneys, cops and journalists have become regulars on national television, analyzing every new tidbit of information. A veritable army of specialists in criminal profiling, forensic pathology and handwriting have spent hours speculating on the JonBenet mystery and second-guessing every move the Boulder police have made. Nowhere has the fallout over the murder been felt more than in the Boulder Police Department, which has so far failed to solve the case.

Chief Tom Koby has become one of America's most watched and criticized cops—a dubious honor he has shared with John Eller, who headed the JonBenet murder investigation until Koby replaced him in October. Both Eller and Koby announced that they would leave the department in 1998. Eller left first, in February. Koby has said he'll be gone by year's end. He said it's time to move on, and then adding that he intended to stay long enough to see the JonBenet case turned over to the district attorney's office.

The police union voted no confidence in Koby in May. The department recently paid a \$10,000 out-of-court settlement with Sgt. Larry Mason, who had threatened to sue because Eller had removed him from the JonBenet investigation in its early weeks, falsely claiming he had leaked information about the case. Detective Linda Arndt, one of the first officers on the scene, was removed from the case in May and went on medical leave to recover from the stress. She's back on the job now, but is not involved in the Ramsey case.

In the Ramsey's former Boulder neighborhood, people gawk at the 15-room tudor home where JonBenet's battered, abused body was found. In Christmases past, it was a holiday showplace with a different elaborate tree in every room. Now it is a tourist mecca. Neighbors got tired of the stream of strangers on their street and the reporters asking questions. "I don't want to participate in this sport," one man said, abruptly closing the door.

Fleet White, one of the family friends with John Ramsey when he found his daughter's body, at one time posted a "Journalists Keep Out" sign outside his house. In a letter to The New York Times this month, White, who has refused to speak to reporters, asked that everyone leave Boulder alone. White requested that Boulder be given the

opportunity to enjoy this sacred season in peace without interruption from anyone who intends to further erode our community's privacy and exploit our misfortunes for the purpose of profit or personal gain.

KOA radio reporter Carol McKinley found the story a welcome challenge; now it has led to a new career. In mid-January, McKinley will become the Denver-based correspondent for Fox News network, a move that will quadruple her radio salary. She had a constant headache from this story and finally had to go see someone to calm her down. After a while, this story really took a bite out of her.

Ramsey family friends who have spoken out have had the same experience. In the months after JonBenet's death, Jim Marino grieved for John Ramsey, the man he had known for 20 years as a client, a boss and a friend. Jim Marino was devastated over it, emotionally distraught. He was bothered by the possibility that his best friend in life was a murderer. That's what everyone was saying. And he just didn't believe it. And he was telling everyone, there's no way that he could do such a thing. But, he didn't go public with this at first. But, in September, he decided to change that. He allowed his name to be used in a Vanity Fair article about the case, and he has come to regret it. At least half of his day was spent talking to reporters, storybook tellers, spin doctors, TV personalities, radio personalities, friends and acquaintances who are curious. It affected not only his personality, but emotionally he was drained.

Are they innocent? Friends say the Ramseys couldn't have killed JonBenet. At first, John and Patsy Ramsey were grieving parents mourning the murder of their 6-year-old daughter. At first, everyone grieved with them. And then, everyone suspected them when they hired attorneys to protect themselves. The initial outpouring of sympathy for

the Ramseys after the discovery of JonBenet's bound and battered body the day after Christmas quickly turned into ugly innuendo, relentless examination and screaming supermarket tabloid headlines. Yet no charges have been filed, and through it all the couple have maintained their innocence. The people who have known them the best and longest and most intimately believed them. In the merciless court of public opinion, John and Patsy Ramsey have been tried and convicted thousands of times over. Authorities say the Ramseys are the focus of their investigation. No one else is known to have been in the house the night of the murder, and Patsy Ramsey has not been ruled out as the author of the ransom note. However, the people who know them say the Ramseys most of the world have criticized and condemned—the billion-dollar businessman and his beauty queen wife, living in some rarefied stratosphere above the law—are cartoonish caricatures. They often bring up the name Richard Jewell—the innocent man convicted first by the FBI of bombing the Olympics in Atlanta, then by the press. The Ramseys have been criticized for what they've done—going on CNN, hiring attorneys and publicists—and what they haven't done: mainly, talking to police immediately after JonBenet's body was found. They have been criticized for their privileged lifestyle, for putting JonBenet in beauty pageants, for doing things most of middle America doesn't. They have been criticized for launching their own investigation, for the way they have grieved, for the way they looked on television. Patsy, 40, was too emotional, people said after they appeared on CNN; John, 53, was not emotional enough.

Even at the funeral, John Andrew had to put himself between photographers and friends. It was only the beginning. Back in Boulder, the Ramseys moved out of their house and in with friends to escape reporters and photographers. They spent the six

weeks after the funeral with Jay Elowsky, who had to steel himself each time he returned to them. Instead, Elowsky found himself swinging at paparazzi with a baseball bat. John and Patsy, out in Atlanta on an errand, saw a photographer pull up next to them and try to take their picture. When they pulled into a bank, the photographer jumped out of his car, ran over to the Ramseys and pushed a camera in John's face. "Why did you kill your daughter?" he screamed, snapping pictures.

John Andrew, alone now in Boulder, always looks over his shoulder when he goes out. The Ramseys have moved back to Atlanta, a place where they feel at home. Still, there will be no unlocked doors. Their new home has state-of-the-art security. They have hired guards to keep them safe from prying cameras and reporters. Jeff Ramsey said he was tired of waiting for the system to work, because it's not working. He was tired of hearing all these innuendos. Somebody out there murdered a member of his family. It's not John or Patsy that murdered her, so whoever did it is still out there he claims.

Could the parents have done it? No, say the Ramseys. No, say their family. No, say their friends. There may never be an answer to what happened to JonBenet. Many people will always suspect the Ramseys. Their friends and family say they never will.

Forensic Evidence

There were no witnesses, no bloody glove, and no confession. Only a little girl's body covered by a blanket in the basement of her family's mansion and a handwritten ransom note found upstairs. To solve the murder of JonBenet Ramsey, Boulder police must rely to a great extent on the results of forensic tests being conducted in crime

laboratories. The question is whether there is a sufficient amount of physical evidence—including body fluids, fingerprints, hair fibers (Figure 3.9), and handwriting—to conclusively determine who sexually assaulted and strangled the 6-year-old beauty queen and youngest child of John and Patsy Ramsey. And the looming problem for police and prosecutors, according to forensics experts, is whether the evidence is in good condition.



Figure 3.9: The murder weapon used to strangle JonBenet Ramsey. 2000, March 20. (<http://www.ramseyfamily.com/garrote.html>)

certainly didn't help anything. It's certainly in the realm of possibility that he contaminated the scene. The JonBenet case, he said, is fascinating from a forensics perspective because so many basic questions remain unanswered and so many possibilities still exist about what happened in the Ramsey house Christmas night. Given the career-damaging cross-examinations that forensic investigators endured in the O.J. Simpson criminal and civil trials, Nelson Jennette, a Montrose forensics investigator, said scientists who undertake laboratory tests on the evidence in the Ramsey case must take extra care to report just what you find.

Or whether lax procedures, including John Ramsey's search of the house eight hours after police were called, his discovery of his slain daughter and his handling of the body as he carried it upstairs, resulted in key evidence being hopelessly contaminated. Many crimes, especially major ones, are solved in the laboratory. Him moving the body

Boulder Police have disclosed little during its investigation into the Ramsey murder. JonBenet died by asphyxiation caused by strangulation (Figure 3.10). Her skull was fractured, and she was sexually assaulted and that blood and possibly semen were found at the scene. No one has been publicly named or eliminated as a suspect. JonBenet's parents and siblings have provided police with blood, hair and handwriting samples for comparison. So have some family friends and employees, as well as some employees at Ramsey's firm. Boulder police spent 10 days collecting physical evidence at the Ramsey's sprawling home near Chautauqua Park. Investigators removed doors, carpet sections and other large items, as well as suitcases believed to contain clothing, bedding and other personal effects. They photographed a maze of footprints in the snowy yard.



Figure 3.10: Artist's recreation from information in new autopsy reports and police sources. 2000, March 29. (<http://members.xoom.com/jbramsey/case/crimephotos.html>)

That suggests a very thorough collection process, forensic experts said. It also means there is a lot of evidence to sift through, much of it under microscopes. Generally speaking, more specimens are better than a few. But that's not necessarily true if evidence has been disturbed or commingled, or if it hasn't been

meticulously handled and catalogued. Sorting through it all and completing dozens of sophisticated tests, from genetic fingerprinting to dissolving fibers in acid to determine

their contents and origins, could take weeks or months. Then there is the matter of what the results mean to the case. With the exception of fingerprints, forensics evidence rarely is conclusive.

A forensics match, even with DNA, doesn't prove a person committed a crime; additional evidence almost always is needed for a conviction. Forensics testing is much better at demonstrating who didn't do it and who wasn't there. Investigators are using genetic testing to sort through potential suspects in the JonBenet case and try to zero in on the killer. DNA is likely being extracted and analyzed on crime scene samples of hair, blood and possibly semen. They would be compared to genetic material in blood and hair samples provided by family and friends.

After years of debate over family variability's, genetic patterns of ethnic groups and reliability of the technology, lawyers and scientists now agree that DNA testing can match the blood, semen or tissue recovered at a crime scene to a given individual. Assuming, of course, that technicians properly collected the samples and the testing procedures were correctly followed. Still, geneticists don't declare that the DNA in a crime scene sample and the genetic material provided by a suspect absolutely are one and the same.

In the home, skin, hair, clothing fibers and other microscopic traces of JonBenet are likely to be everywhere in the house, as well as traces of every family member and perhaps frequent visitors. If no clear physical evidence of an intruder is present, and police suspicions narrow to the people closest to the girl, it could be difficult to determine what traces are crime-related and what are not. With the crime scene being in the basement, it makes it very difficult to collect "clean" evidence for forensic analysis. If

investigators used powerful vacuum sweepers to collect everything that might have been overlooked, the process could jumble the crime scene dirt with older, unimportant debris layers.

After John Ramsey discovered his daughter's body, removing the duct tape from her mouth and carrying the body upstairs has been described as the natural reaction of a grief-stricken father. But from a forensics perspective, it was a major blunder. Fibers, hair and debris left by the killer might have been dislodged in the process. Unless Ramsey was wearing an astronaut suit or something else along those lines that neither sheds nor attracts, Jennette and others believe his actions probably resulted in some materials being transferred between him and the victim. When a crime is committed, the person leaves something behind and takes something away. It can be microscopic in size. Moving the body can contaminate things in a major way. Only the future will show whether DNA fingerprinting comes to the forefront of this case.

Chapter 5: Probability and Statistics

Determination of the probability of specimen match and estimation of population allele frequency distributions are two key areas of DNA profiling requiring probabilistic and statistical analyses. Statistical calculations can be tedious, and slight changes in the wording of probability statements can result in vastly different meanings. It may, therefore, be prudent for the fingerprint analyst to seek the advice of a qualified statistician before assembling population frequency data or submitting probability statements in a court of law.

Statistical methods provide powerful tools to assist with decision-making. Because of easy access to powerful statistical software on personal computers, the methods are relatively easy to use. However, for this same reason they can also often be misused, and questionable data presented in a favorable light. One must always be aware of those skilled in the misleading use of statistics, or those who simply make incorrect statistical statements even though the quantity of base data may be considerable and the quality good.

The objective of this chapter is to review methods of probability and statistics relevant to DNA fingerprint analysis.

STATISTICAL METHODOLOGY

Random Sampling

Basic statistical data are usually derived from samples drawn from large populations. A population is a collection of individuals having stated features in common, such as all Orientals in the United States. A simple random example is a subset

of individuals selected from a population using a random choice mechanism (such as a random number generator), which guarantees that all members of the population are equally likely to be chosen. Usually the sample size is denoted by n . Values in a sample are called observations and denoted by X_1, X_2, \dots, X_n .

The features of interest in a population such as the true average IQ, or the true proportion of a specific allele at a given locus, are called parameters, while statistics are numerical summaries of data such as means or standard deviations. The science of statistics is concerned with inferring information about population parameters based on sample statistics.

Summary Statistics

Measure of central tendency: The mean, median, and mode are measures of the center of the data. The (arithmetic) mean is the most common measure; it is computed by summing the data and dividing by the number of observations n . The Greek letter μ denotes a population mean and (\bar{x}) a sample mean. When the population is finite, μ is the average computed by integration. The quantity (\bar{x}) is the average over all observations in the sample. It is computed by:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

The median is the middle value among the sample observations in the population. Fifty percent of the values are above and 50% are below this point. If the number of data points is odd, the median is the middle value; if even, it is the average of the two central values.

The median is found by arranging the data in ascending order and choosing the central value or the average of the two central values. The mode is the value in a sample or population that occurs most frequently. Some distributions may be bimodal or multimodal. In fingerprinting practice, the mode has little practical significance. For data that are symmetric, the mean and median are good preferable. When data are categorical, important summary statistics include the proportion p of observations in each category.

Measures of dispersion

It is important to calculate the dispersion within a distribution as well as the mean or median because many distributions might have the same mean, but the variation of observations around the mean might differ markedly. Measures of dispersion include the range, percentiles, standard deviation (SD), and the coefficient of variation (CV).

The range gives the smallest and largest numbers in a sample or population. The range provides a measure of spread but is overly sensitive to disparate measurements because it is based on only the two extreme values.

A percentile is the point in a sample or population below which a specific percent of values lie and above which the complementary percentage lies. The 95th percentile is defined to be the value which lies above 95% of the observations, and below 5% of the observations.

The standard deviation is the most important measure of dispersion in a population or sample. It is the square root of the average of the sum of squared deviations from the mean. The population standard deviation (in a finite population) is defined by :

$$\alpha = \sqrt{\frac{\sum_{i=1}^n (X_i - \mu)^2}{n}}$$

And the *sample standard deviation* by

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

In a symmetric population (or sample), roughly 95% of the observations lie within two standard deviations of the mean, and in any population (or sample), 75% of the observations lie within two standard deviations of the mean.

The coefficient of variation (CV) is a relative measure of dispersion expressed as a percentage. The sample coefficient of variation is given by

$$CV = \frac{s}{\bar{X}} \times 100$$

The other measures of variability, such as the SD, have magnitudes dependent on the scale of the data, whereas the CV is a ratio and is dimensionless.

Statistical Inference

The purpose of almost all statistics is to infer something about a population based on observations in a sample. This is carried out through point and interval estimation, and hypothesis testing.

Point estimation: Sample quantities such as sample means, sample standard deviation, and sample proportions are used as point estimates of the analogous population quantities. They are called point estimates because they are single values with no associated measures of precision. Desirable properties of such estimators are that they use the data efficiently and are certain to be sufficiently close to the true parameter value when a sufficiently large sample is selected.

Interval estimation: Interval estimates or confidence intervals are useful for determining the precision of a point estimate. They are determined so that if an experiment or sampling procedure were repeated many times, the true parameter value would lie in the interval a pre-specified proportion of the time. For example, a 95% confidence interval would be computed from a rule that ensured that in 95% of the samples, the interval would contain the actual parameter value.

When computing confidence intervals for population means, a key quantity is the standard error of the mean. It is often referred to as the standard error or SEM and is usually computed using the formula $S/(\text{square root of } n)$. It provides an idea of how variable a sample mean is as an estimate of the population mean. This quantity should not be confused with the sample standard deviation, which describes the variability of a single observation. The standard error is an important component in computing a $(1-\alpha) \times 100\%$ confidence interval. When the sample size is large enough (say greater than 50), confidence intervals are given by the form:

$$\bar{X} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

In this expression, the quantity $z_{\alpha/2}$ is the value a standard normal random variable exceeds with probability $\alpha/2$. For 90% confidence intervals $z_{0.05} = 1.645$, and for 95% confidence intervals $z_{0.025} = 1.96$. A 95% confidence interval for the population mean is approximately the sample mean plus or minus two standard errors.

Confidence intervals for a population proportion π are obtained from the sample proportion p using the formula

$$p - z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

This formula applies only when the sample is sufficiently large so that $n \times p$ and $n(1-p)$ both exceed five. If p were 0.02 and n were 100 so that $n \times p = 2$, the interval would not be guaranteed to contain the true population frequency in $(1-\alpha) \times 100\%$ of the samples. As a rule of thumb, such confidence intervals are invalid when one wishes to estimate low frequencies. Methods based on the Poisson distribution or exact procedures are then preferred.

Calculation of allele frequency distributions: Allele frequencies for the specific population in question must be determined before the proportion of the population that are potential contributors of an evidentiary specimen can be calculated. Natural populations are often characterized by differences in allele and genotype frequencies in various geographic regions. Probability calculations, as discussed under the Hardy Weinberg and Wahlund principles, must take account of this fact; otherwise, the resultant predictions will not be valid.

Depending on the confidence limits chosen, hundreds or even thousands of control tissue samples may have to be analyzed to determine the frequency of rare alleles.

These specimens usually constitute only small samples from the specific populations; thus, an indicator of the confidence one has in the data estimated proportions must also be determined. A 95% confidence interval based on a rare allele frequency of $p=1/50=0.02$ and $n=1000$ specimens is

$$0.02 - 1.96\sqrt{\frac{0.02 \times 0.98}{1000}} \leq \pi \leq 0.02 + 1.96\sqrt{\frac{0.02 \times 0.98}{1000}}$$

or equivalently, 0.02 ± 0.0086 , or 0.0114 to 0.0286. In this population, the true frequency of this rare allele could range from 114/10000 to 286/10000. If the sample size n were increased to 5000, this range would be reduced to 161/10000 to 239/10000.

Sample size determination: The confidence interval formulae can be used to determine the sample size required to attain a pre-specified degree of precision. If one desires an estimate that is accurate to plus or minus L with $(1-\alpha) \times 100\%$ confidence and the population proportion is expected to be p' then the required sample size is

$$n = \left(\frac{z_{\alpha/2}}{L}\right)^2 p'(1-p')$$

Suppose the allele frequency is expected to be (0.01), a 95% confidence interval is sought, and a precision of plus or minus 0.001 is desired. The number of analysis specimens (n) requires is $(1.96/0.001)^2(0.01 \times 0.99) = 38031$.

Simultaneous confidence intervals: Often the analyst is interested in estimating more than one allele frequency. For example, if several alleles are present at one locus, then estimates of both frequency and precision are required. If individual confidence

intervals, such as those described above, were obtained for each estimate, the overall confidence statement (that pertaining to several allele frequencies), would not hold at the predetermined confidence level. For example, with two independent 95% confidence intervals, the probability that both contain the true allele frequencies would be $(0.95)^2 = 0.9025$. If instead we wished to determine five allele frequencies, the probability that all five individual 95% confidence intervals would contain the true frequencies is $(0.95)^5 = 0.7738$.

To ensure that the probability that all confidence intervals contain the true parameter values is as specified, for example 95%, then each individual interval must be made wider. Bonferroni's method is the easiest to use and applies in many situations (Goodman, 1965). If K quantities, such as allele frequencies, are to be estimated and the goal is to have an overall $(1-\alpha) \times 100\%$ confidence interval, then choosing each interval to be a $(1-\alpha/K) \times 100\%$ interval achieves this objective. For example, if we wish to obtain an overall 95% confidence interval, and two allele frequencies are being estimated, then each should be estimated with confidence $(1-\alpha/2) \times 100\%$ or 97.5%. If five frequencies are being estimated, then using a 99% confidence interval, $(1-0.05/5) \times 100\%$, for each ensures an overall 95% confidence interval.

Suppose we estimate K frequencies, say π_1, \dots, π_k , by p_1, \dots, p_k based on samples of size n_1, \dots, n_k . Then the following individual confidence intervals will give an overall $(1-\alpha) \times 100\%$ confidence interval.

$$p_i - z_{\frac{\alpha}{2}} \sqrt{\frac{p_i(1-p_i)}{n_i}} \leq \pi_i \leq p_i + z_{\frac{\alpha}{2}} \sqrt{\frac{p_i(1-p_i)}{n_i}}$$

Simultaneous allele frequency determination: Suppose in sample of size 100, we observe the frequency of two alleles to be 0.05 and 0.10 and we seek confidence intervals that will contain both true frequencies at a confidence level of 95%. As described above, we compute 97.5% confidence intervals for each. In the above formula, $\alpha=0.05$ and $K=2$, so we obtain $z_{0.0125}=2.24$. These intervals are

$$0.05 - 2.24\sqrt{\frac{0.05 \times 0.95}{100}} \leq \pi_1 \leq 0.05 + 2.24\sqrt{\frac{0.05 \times 0.95}{100}}$$

and

$$0.10 - 2.24\sqrt{\frac{0.10 \times 0.90}{100}} \leq \pi_2 \leq 0.10 + 2.24\sqrt{\frac{0.10 \times 0.90}{100}}$$

This gives the intervals 0.05 \pm 0.049 for the first allele and 0.10 \pm 0.067 for the second allele.

Binning: Using a panel of VNTR probes, with each probe recognizing an independent and hyper variable locus, the procedure used to produce a composite DNA profile unique to each individual. This approach currently provides the best route for characterization of body tissues for excluding an individual falsely associated with an evidence sample. For many profiles, the analyst must determine the proportion of the population that could have contributed the specimen. This calculation must take into consideration the limitations of the technology and the available population data.

Resolution of alleles that differ by only one or a few repeat units is not possible by present gel electrophoresis and autoradiography techniques. This is especially true when an allele is large and the tandem repeat units are small. The true number of alleles associated with a hyper variable locus may be extremely difficult to ascertain, especially

if a number of the alleles are infrequent in the population. The D1S7 locus alleles range from approximately 1 to 20 kb, or 110 to 2220 tandem repeats. Theoretically, over 2100 alleles could exist for this D1S7 locus.

Binning is used by the Federal Bureau of Investigation to determine population allele frequencies and in turn to determine the frequency of an allele in an evidentiary sample (Budowle, 1989). The approach compensates for limited fragment resolution. The system is conservative; consequently it is unlikely that an allele from a sample specimen would be assessed a frequency of occurrence that is lower than the true frequency in the population of unrelated individuals.

Bins are designed with boundaries defined by size standards, such as restriction digests of viral DNA. The differences in the sizes of two fragments that define each bin must be greater than the measurement imprecision of the analytical system. Sample population alleles are placed in the bins according to size and a frequency of occurrence for each bin is calculated. An allele from a suspect or crime sample is assigned the frequency of the bin the allele falls within.

Allele's are said to match only if they are of the same size not if they fall in the same bin. If an allele overlaps two bins (i.e. it could be placed in either of two adjacent bins), the bin with the larger frequency is chosen. It should be noted that closely sized alleles reside within the same bin, and although they do not match exactly, each will be assigned the same frequency.

Lastly, how can low frequency alleles be used in the binning system? This is an important consideration because it is these alleles that can provide match probabilities measuring in the billionths. Until adequate population data are available, Budowle and

Monson (1989) recommend that bins be combined to contain a minimum of five alleles each.

A floating bin approach (as opposed to fixed bins as described above) has been used by Balasz et al (1989). Whereby the size assigned to an evidentiary sample allele is assumed to be the mean size. Because, with current assay methodology, it cannot be demonstrated that the value assigned is the true mean, the Federal Bureau of Investigation avoids using this approach altogether.

PROBABILITY

A critical question arises in forensic DNA profile analysis: What is the chance or probability that suspect and crime samples have the same source given that they match? For paternity testing, a similar question must be asked: What is the probability that a putative father is the biological father when all of the child's alleles are found in the mother and putative father?

These are some of the questions that can be answered using probability theory. Probability is a mean of quantifying uncertainty. Probabilities are numbers between 0 and 1 that indicate the degree of likelihood of various outcomes (events) of experiments. They can be determined either subjectively, objectively, or empirically.

Subjective probabilities are based on one's feelings, preferably based on experience, about the likelihood an event will or will not occur. If this subjective approach is based on solid experience (expert opinion), the conclusions could be sound. An example of a subjective probability calculation is a determination of the probability that it will rain more than two centimeters two weeks from now.

Objective probability calculations are relevant when an experiment has several equally likely outcomes such as occurs in games of chance. The probability that an event E occurs, $P(E)$ equals the number of outcomes corresponding to E , n_E , divided by the total number of possible outcomes of the experiment n_{TOT} . that is, $P(E) = n_E / n_{TOT}$. Clearly $0 \leq P(E) \leq 1$. For example, when rolling a die, there are six possible outcomes so the probability of each is $1/6$. To compute the probability of the event E “the outcome is odd”, we note that if any of the three outcomes 1, 3 or 5, occurs, then E occurs so that $n_E = 3$, $n_{TOT} = 6$, and $P(E) = 3/6 = 1/2$.

Empirical probability calculations are based on information acquired from data collection. For example, to determine frequency of individuals possessing a certain allele in a population, choose a sample and define probability of the allele to be the number of individuals exhibiting that allele divided by the number of individuals in the sample. If the sample size is large, the empirically determined frequency would be a very good approximation of the “true” frequency. Confidence intervals determine the precision of such calculations.

Probability of Combined Events: Addition and Multiplication Rules

Two events A and B are said to be *mutually exclusive* if the occurrence of one excludes the occurrence of the other. For example, the events “an individual is Caucasian” and “an individual is Oriental” are mutually exclusive, while the events “an individual is Caucasian” and “an individual is a male” are not. The probability that one of several mutually exclusive events occurs is the sum of the probability that the individual events occur. That is:

$$P(A \text{ or } B \text{ or } C \dots) = P(A) + P(B) + P(C) + \dots$$

In this calculation, it is essential that the events contain no items in common. If not, the calculation becomes more complex, in the case of two events:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

The event “A or B” is sometimes referred to as the union of A and B, while the event “A and B” is referred to as the intersection of A and B.

The probability that both events A and B occur is computed by the formula

$$P(A \text{ and } B) = P(A/B)P(B) \text{ or } P(A \text{ and } B) = P(B/A)P(A)$$

Where $P(A/B)$ is the conditional probability that A occurs given that B occurred, while $P(B/A)$ is the conditional probability that B occurs given that A occurred. Conditional probabilities are probabilities that take evidence into account. Two events A and B are *independent* if $P(B/A) = P(B)$ and $P(A/B) = P(A)$. Independence means that evidence that A occurred does not influence our estimate of $P(B)$. If two (or more events) are independent, the joint probability that they all occur is the product of the probability that each occurs. That is

$$P(A \text{ and } B \text{ and } C \dots) = P(A)P(B)P(C)\dots$$

Examples of the addition and multiplication rules follow: If at locus M the probability of detecting allele A is 0.2, and at locus R the probability of detecting allele B is 0.3, and occurrence of alleles at the loci are independent, then the probability of detecting both A and B is $P(A \text{ and } B) = P(A)P(B) = 0.2 \times 0.3 = 0.06$. The probability of detecting either allele A or allele B or both is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 0.2 + 0.3 - 0.06 = 0.44$$

When matching individual genotypes to forensic samples, the phrases “probability of a match” and “probability of a match given one of the set to be matched” are quite different. If the frequency of a given allele, A, in a population is 1/100, the probability of independently drawing two individuals are random from the population each with an A is $(1/100)(1/100) = 1/10000$. If a crime specimen is known to contain an A allele, the probability of drawing an individual at random from the population matching this allele is 1/100. The former, P(match), is a case of an unconditional probability, while the latter P(match one allele is A), is a conditional probability.

Bayes’ theorem and revision of Probabilities

Bayes’ Theorem provides a method for revision probabilities based on acquired information or evidence. We begin with a prior estimate of the probability (prior probability) of event A and combine it with the known conditional probability of the evidence B, given A has occurred, to obtain the revised or posterior probability of A give the evidence B. Thus we are given the prior P(A) and the conditional P(B/A) and wish to obtain the posterior P(A/B). Bayes’ Theorem can be expressed by the formula

$$P(B | A) = \frac{P(A | B)P(B)}{P(A | B)P(B) + P(A | notB)P(notB)}$$

Where *not* B is the event known as the complement of B and corresponds to all outcomes for which B does not occur. Note that $P(B/A) + P(not B/A) = 1$.

In determine whether an individual is a carrier for a specific allele, such as female for an X-linked defect, one might wish to compute the probability that a woman, whose mother is a confirmed carrier, is herself a carrier, when it is known that she has two non-

affected sons. Let B be the event the mother is a carrier and A be the event that she has two non-affected sons. Suppose $P(B) = P(\text{not } B) = 0.5$. The conditional probability $P(A|B)$ is the chance of having two non-affected sons if the mother is a carrier, which is $(0.5)(0.5)=0.25$, which $P(A|\text{not } B)$ is the chance of having two non-affected sons if she is not a carrier, which is 1. Then:

$$P(B | A) = \frac{P(A | B)P(B)}{P(A | B)P(B) + P(A | \text{not } B)P(\text{not } B)} = \frac{0.25 \times 0.5}{0.25 \times 0.5 + 1 \times 0.5} = 0.2$$

Similarly, $P(\text{not } B|A) = 0.8$, so that the probability that the mother of two non-affected sons is a carrier is 0.2, while the probability she is not is 0.8. Thus, the original assumption of 0.5 (since her mother was a confirmed carrier) has been modified to 0.2 based on the information that she has two non-affected sons. If her mother had not been a confirmed carrier, we would use the population frequency of carriers for the prior probability and repeat the above calculation.

In paternity cases, Bayes' theorem can be applied as follows. Let B be the event that the putative father is the biological father. Then $P(B)$ is the prior probability that he is the biological father and $P(\text{not } B)$ is the probability he is not. Let A be the event that the putative father could transmit a specific allele (one present in the offspring). Then $P(A|B)$ is the probability he could transmit the specific allele given he is the father and $P(A|\text{not } B)$ is the probability that the putative father could transmit this allele given he is a randomly selected individual. If the putative father is homozygous for that allele, the conditional probability $P(A|B)$ is 1 while the $P(A|\text{not } B)$ is the frequency of that allele in the population. If the putative father is heterozygous, $P(A|B) = 0.5$ and $P(A|\text{not } B)$ is the frequency of the allele multiplied by 0.5, since there is only a 50% chance the allele would be transmitted to an offspring. $P(B|A)$ is the probability that the man is the father

given he transmitted the specified allele. It can be computed by direct substitution into the above formula. In paternity and forensic laboratory testing, the prior probabilities are often considered equal and cancel from the genetic marker analysis calculations because of the above formula. Inclusion of the prior estimates is the prerogative of the judge or jury at a latter stage of the proceedings.

Random variables and distributions

Random variables (denoted by capital letter such as X and Z) are numerical value outcomes of random experiment. For example, the number of dots observed when rolling a die is a random variable, as is the height of an individual selected at random from some population. A probability distribution assigns a probability to each possible value of the random variable. If the set of outcomes is discrete, such as when rolling a die, the probability distribution can be given by a list: $P(X=1) = 1/6$, $P(X=2) = 1/6$, ... $P(X=6) = 1/6$. The possible values together with the probability distribution for the random variable can also be determined theoretically, or a frequency distribution of the actual observations can be used as an approximation. Important discrete probability distributions include the binomial, Poisson, and hypergeometric; they will not be discussed here.

Probability distributions of continuous random variables, such as height of an individual, are best represented graphically or through formulae. The most important continuous distribution in statistical analysis is the normal distribution. Many quantities, such as IQ and male height, have distributions resembling the normal curve. Characteristics of the curve include a symmetric bell shape with the mean, median, and

mode at the same central location, and the curve tails approaching the horizontal axis. Probabilities are found by computing areas under the curve. The total area under the curve is designated as 1.0 or 100 percent.

The normal distribution is characterized by two parameters: its mean μ and its standard deviation α . The areas or probabilities corresponding to intervals $\mu \pm \alpha$, $\mu \pm 2\alpha$, $\mu \pm 3\alpha$ are depicted in the normal curve. They are 68.24%, 95.45% and 99.73% respectively. To calculate other probabilities, a general normal random variable is transformed to one with mean 0 and standard deviation 1. This transformation is given by

$$Z = \frac{X - \mu}{\alpha}$$

This is referred to as a *standard normal* and is usually denoted by Z .

Example:

Suppose one wished to calculate the probability that a randomly selected individual has an IQ between 110 and 125. IQ scores have a mean of 100 and standard deviation of 15. Let X represent the IQ of the randomly selected individual. The desired probability is $P(110 \leq X \leq 125)$ and it is calculated as follows:

$$\begin{aligned} P\left(\frac{110-100}{15} \leq \frac{X-100}{15} \leq \frac{125-100}{15}\right) &= P(0.67 \leq Z \leq 1.67) \\ &= P(0 \leq Z \leq 1.67) - P(0 \leq Z \leq 0.67) = 0.4525 - 0.2486 = 0.2039 \end{aligned}$$

Frequently, especially in statistical applications, it is necessary to determine a numerical value that Z exceeds with probability α , that is the $(1-\alpha)$ th percentile of Z . This quantity is denoted by z_α and can also be obtained from the normal curve.

Genetic Applications of Probability

Hardy-Weinberg Law. The Hardy-Weinberg law states that in a large random mating population, where no disturbances by outside influences such as mutation, migration, or selection exist, the relative proportions of the different genotypes remain constant from generation to generation. In a two-allele (A and a) system with the frequency of allele A equal to p and that of a equal to q, the genotype proportions of AA:2Aa:aa, or $p^2:2pq:q^2$ are valid under the Hardy-Weinberg law. The frequency of the allele A plus the frequency of allele a must equal 1 ($p + q=1$). If the frequency of A in a given population is 0.8 and that of a is 0.2, the genotype frequencies pp, 2pq and qq are $(0.8)(0.8)=0.64$, $2(0.8)(0.2)=0.32$, and $(0.2)(0.2)=0.04$. In a three allele system, the expansion of $(p+q+r)^2$ or $p^2+q^2+r^2+2pq+2pr+2qr$ gives the relative frequencies.

Nothing that $p+q=1$ or $q=1-p$, the genotype proportions at one locus in a two-allele system in Hardy-Weinberg equilibrium can be expressed as

Genotype	Proportion
AA	$p^2 = (1-q)^2$
Aa	$2pq = 2q - 2q^2$
aa	$q^2 = (1-p)^2$

The frequency of allele a in the population is the sum of the frequencies of genotypes aa and Aa which, in this notation, equals $2q-q^2$. Similarly, the frequency of allele A is $1-q^2$.

Genotype frequencies for populations in Hardy-Weinberg equilibrium can be determined using a Punnett square. Examples for loci with two and three alleles are given in *Table 5.1*. In a two-allele system, the probability that an A-bearing sperm fertilizes an A-bearing egg is p^2 . The probability that an a-bearing sperm fertilizes an a-bearing egg is q^2 . The probability that an a-bearing sperm fertilizes an A-bearing egg is pq and the probability that an A-bearing sperm fertilizes an a-bearing egg is pq . The frequency of Aa is $pq+pq=2pq$.

Valid use of the multiplication rule in probability calculations dictates that each event is random and independent of the others. In population genetics, the procedure of determining the combined power of identity from a series of individual powers is valid only if the loci profiled are not linked and Hardy-Weinberg equilibrium exists. If the observed genotype frequencies deviate significantly from those expected, then Hardy-Weinberg equilibrium has not been attained. The chi-square goodness of fit test described earlier can be used to compare observed and theoretical frequencies.

Non-random mating and migration can exert a considerable influence on Hardy-Weinberg equilibrium. A founder effect may occur in population isolates where a single gene spreads through the population in a few generations. In isolates, the phenomenon of random genetic drift can result in the fixation of one gene and loss of another in a single generation. The loss could occur if only a very few individuals have a specific gene that is not transmitted. This phenomenon could have occurred with human blood groups. The majority of North American Indians are group O; however, in the Blood and Blackfoot populations, type A is frequent. Perhaps this is due to immigrant ancestors

carrying A, B, and O from Asia, but in a small isolate, A became common due to genetic drift.

Table 5.1: Hardy-Weinberg Genotype Frequency Determination

Two Alleles ($p = q = 0.5$)

		Sperm	
		A(p)	a(q)
Egg	A(p)	AA(p^2)	Aa(pq)
Egg	a(q)	Aa(qp)	aa(q^2)
		Offspring	
		Genotype	Frequency
		AA	$p^2 = 0.25$
		Aa	$2pq = 0.50$
		aa	$q^2 = 0.25$
		Total	1.00

Three Alleles ($p = q = r = 0.333$)

		Sperm		
		$A_1(p)$	$A_2(q)$	$A_3(r)$
Egg	$A_1(p)$	$A_1A_1(pp)$	$A_1A_2(pq)$	$A_1A_3(pr)$
Egg	$A_2(q)$	$A_2A_1(qp)$	$A_2A_2(qq)$	$A_2A_3(qr)$
Egg	$A_3(r)$	$A_3A_1(rp)$	$A_3A_2(rq)$	$A_3A_3(rr)$

Offspring

	Genotype	Frequency
	A_1A_1	$p^2 = 0.111$
	A_2A_2	$q^2 = 0.111$
	A_3A_3	$r^2 = 0.111$
	A_1A_2	$2pq = 0.222$
	A_1A_3	$2pr = 0.222$
	A_2A_3	$2qr = 0.222$
	Total*	= 1.000

* Note individual frequencies do not sum to 1.000 because of roundoff.

Gene frequencies can be considerably affected by migration of population groups. Blood type O is frequent in American Indians but rare in Asiatic, even though there is good evidence that American Indians are descended from Asiatic ancestors who migrated

across the Bering Straits several thousand years ago. Lastly, it has been possible to estimate gene flow between Caucasians and Blacks in the United States by determining blood group gene frequencies in these groups as well as Blacks in West Africa. The rate is from a low of 4% in the southern U.S. to a high of 25% in the north.

Wahlund's Principle. The observation that the rate of homozygosity in a population consisting of distinct subgroups is greater than that expected under Hardy-Weinberg equilibrium is referred to as Wahlund's principle. The subgroups must be in Hardy-Weinberg equilibrium and have different allele frequencies. The pooled population allele will equal the average of those in the subpopulations.

Observed allele and genotype frequencies in two equal sized subpopulations of a specific population, together with the calculated genotype frequencies under Hardy-Weinberg equilibrium in each subpopulation and in a combined population with the average allele frequencies, are given in *Table 5.2*.

The calculations show that because the population consists of distinct subgroups, the population composed of a 50:50 mixture of these two subgroups will have genotype frequencies 17:46:37. These differ from the Hardy-Weinberg frequencies of 16:48:36. Wahlund's principle is the observation that in this (or 54%) exceeds that in a population in Hardy-Weinberg equilibrium (in the example, 52%). If the subgroups randomly mate, the population will achieve Hardy-Weinberg proportions in one generation.

Table 5.2:

Hardy-Weinberg Equilibrium in Each Subpopulation.

	Alleles		Genotypes		
	A	a	AA	Aa	aa
Subpopulation 1	0.3	0.7	0.09	0.42	0.49
Subpopulation 2	0.5	0.5	0.25	0.50	0.25
Mixed population (50 : 50)	0.4	0.6	0.17	0.46	0.37

Hardy-Weinberg Equilibrium in the Mixed Population

	Alleles		Genotypes		
	A	a	AA	Aa	aa
Frequency	0.4	0.6	0.16	0.48	0.36

Forensic probability determination. As mentioned earlier, a fundamental question in forensic science concerns the probability that a specific crime tissue specimen is derived from a specific suspect. If a multilocus multiallele DNA analysis approach is used, as with good quality fingerprints, a match of two sets of DNA fingerprints normally implies that they are from the same source because each person's DNA is unique. If the prints derive from the same source, the probability of a match is 1; if they derive from different sources, the probability of a match, from experience to date, is zero. Population studies can be carried out to estimate the probability that any specific fragment in a multiband profile present in individual A is also present in another individual B randomly selected from the population. Provided the loci analyzed are not linked, the probability that all fragments present in A are also present in B can be calculated using methods described and applied in the section "estimation of probabilities for multilocus multiallele fingerprint systems" below.

If single-locus multiallele DNA analysis is used, this is comparable to, but considerably more specific than conventional blood classification typing. When a match is observed, the probability that the match could have arisen by chance in the population must be calculated. Population allele frequencies and the frequencies of the loci genotypes must be known. Provided Hardy-Weinberg and linkage equilibria apply, the probabilities for the loci matched can be multiplied to determine the composite profile probability. The value of match evidence, in conjunction with other evidence, can be very incriminating especially if a number of loci are analyzed and rare alleles are present.

A match between a crime and suspect sample for a specific allele does not equate to the same specimen source. The probability of a match between suspect and crime specimens when the source is the same is 1. If the sources are different, the probability of a match, given the crime allele size is known, equals the specific population allele frequency. With a specific allele frequency of 0.01, the match probability would be 0.01. The probability of a profile match, with alleles at a number of loci, is the product of the individual locus genotype frequencies. For example, if five independent loci with genotype frequencies of 0.05, 0.1, 0.15, 0.3, and 0.4 are tested, the probability of a match is $(0.05)(0.1)(0.15)(0.3)(0.4) = 0.00009$.

The incrimination value (IV) of match evidence defined by

$$IV \frac{\text{probability of match with the same source}}{\text{probability of match with the different source}}$$

is directly dependent on the probability of a match when the suspect and crime specimens are from different sources. If a genotype frequency is 0.01 and there is a match between the suspect and crime specimen for this genotype, then $IV=1/(0.01)=100$. If the genotype frequency were ten fold greater, then $IV=1/(0.1)=10$.

The FBI prefers to determine the percent of the population that could contributed the evidentiary sample for which the suspect cannot be excluded as a contributor. This approach is especially important when several mixed profiles or partial profiles must be considered.

Likelihood of paternity. During the past decade, statistical analysis of conventional blood-typing results applied to paternity testing has involved a degree of controversy. Representative journal articles include “Basic Fallacies in the Formulation of the Paternity Index” (Li, 1985), and “No Fallacies in the Formulation on the Paternity Index” (Bauer, 1986). The problem centers on the method of calculating and reporting the probability of paternity when exclusion of the alleged father is not possible.

Measures of the chance of paternity include statistical frequency, the paternity index, and the probability of paternity. (The chance of paternity refers to the probability of transmitting the offspring’s paternal alleles.) The statistical frequency (SF) is the chance that a randomly selected male from the population is the father (δ). The paternity index (PI) or likelihood ratio (LR) is the ratio of the chance of paternity for the putative father (β), assuming he is the biological father, to the chance of paternity for a randomly selected male (δ), $PI = \beta / \delta$. The probability of paternity (PP) expresses the paternity index as a percentage and equals $[\beta / (\beta + \delta)] \times 100$.

These basic expressions, as applied in practice, employ only genetic markers identified in the mother, offspring, and putative father. The same calculation techniques can be applied when DNA markers are used but the system is simpler because only genotypes need be considered. Size measurement of the DNA marker bands (alleles)

may create problems if the fragment sizes are similar. The simplest solution is to use only enzyme and probe systems where the alleles can be readily separated on gels and the sizes easily and accurately determined. A typical calculation process follows:

1. List the mother, child, and putative father genotypes.
2. Determine the alleles inherited from the father. (First define the maternal alleles; those remaining must be paternal.)
3. In (2) above, list the allele frequencies from the population in question. These data are used to calculate the statistical frequency of the paternal allele profile in the population.
4. Using the data from (3) above, determine the frequency of the paternal allele profile in the population. This value is the statistical frequency δ .
5. Calculate PI and PP.

This procedure is illustrated below. Given an offspring DNA profile consisting of paternal alleles X, Y, Z, and W from non-linked loci A, C, O, and P with population frequency of these alleles equal to 1/25, 1/50, 1/100, and 1/125; and, provided the putative father is homozygous at each locus, the calculations proceed as follows. Since the putative father is homozygous at each locus, $\beta=1$.

$$SF = \delta = (0.04)(0.02)(0.01)(0.008) = 6.4 \times 10^{-8}$$

$$PI = \beta/\delta = 1/6.4 \times 10^{-8} = 1.56 \times 10^7$$

$$PP = [1/(1 + 6.4 \times 10^{-8})] \times 100 = 100\%$$

If the putative father is heterozygous at any locus, the allele frequency at that locus must be multiplied by 0.5 since there is only a 50% chance the allele would be

transmitted to an offspring. If the four loci in the above example were instead heterozygous,

$$SF = (0.04/2)(0.02/2)(0.01/2)(0.008/2) = 4 \times 10^{-9}$$

$$\beta = (0.5)(0.5)(0.5)(0.5) = 6.25 \times 10^{-2}$$

$$PI = 6.25 \times 10^{-2} / 4 \times 10^{-9} = 1.56 \times 10^7$$

$$PP = [0.0625 / (0.0625 + 4 \times 10^{-9})] \times 100 = 100\%$$

Note that the statistical frequency is effected by heterozygosity while PI and PP are not.

Estimation of probabilities for multilocus multiallele fingerprint systems. The following method was proposed by Jeffreys (1985, 1987a) and used by Georges (1988) for estimating the probability that a specific allele is present in individual B given that individual A is known to possess the allele. As described in the discussion of conditional probability, this is the probability that a DNA fingerprint fragment (allele) is present in the population. This calculation serves as the basis for estimating the probability that two individuals have identical DNA fingerprints as the result of applying a minisatellites probe.

Suppose q is the frequency of a specific allele in the population. Then the probability (x) that an individual selected at random from this population contains this allele is $2q - q^2$. (See the section on the Hardy-Weinberg law.) When q is sufficiently small so that q^2 is much smaller than q , x can be approximated by $2q$. The “heterozygosity” or proportion of individuals possessing a specific allele who are in the heterozygous state is given by $h = (2q - 2q^2) / (2q - q^2) = 2(1 - q) / (2 - q)$.

Georges and Jeffreys use the following approach for estimating x . A sample of individuals is selected, DNA is isolated, and profiles are prepared. From this sample of fingerprints, distinct resolvable bands are unverified and the proportion of the sample possessing each is recorded. The average of these sample proportions is the estimate of the “true” average probability x . From this estimated average probability x , the allele frequency is estimated by solving $x=2q-q^2$ (Georges, 1988) or $x=2q$ (Jeffreys, 1985, 1987a).

This method is best illustrated through a simple example with a small number of individuals and a small number of resolvable bands. Suppose there are six individuals in the sample and five bands are resolvable. The data are presented in the following table with an “X” indicating that the band is present in the individual. The column “Proportion” indicates the fraction of individuals in the sample who possess that band in their “fingerprint”.

Bands	1	2	3	4	5	6	Proportion
2.3kb	X		X	X			0.5
2.9kb		X		X	X	X	0.67
4.1kb	X		X				0.33
8.2kb	X	X	X		X	X	0.83
11.4kb	X		X	X			0.5
Total Number of bands	4	2	4	3	2	2	

The estimated average probability of a match, x' , is the average of the “Proportion” column, which equals 0.57. From this, the average allele frequency is

estimated by solving $2q' - q'^2 = 0.57$ which gives $q' = 0.34$. The estimated heterozygosity is $h = 0.80$. Using the formulas from the section on descriptive statistics shows that the mean number of bands per individual, m is $17/6 = 2.83$ and the standard deviation of the number of bands is 0.98.

Using the probe that produced the above result, the probability that an individual selected at random has an identical profile to a specific individual is estimated by the method of Jeffreys and Georges to be $x^m = (0.57)^{2.83} = 0.20$.

Jeffreys et al. (1985, 1987) calculated that the mean level of band sharing between unrelated individuals is 0.25 in both the North European and the Indian subcontinent populations when their Hinf I endonuclease-digested DNA is hybridized with minisatellites probes 33.15 or 33.6. The Home Office Forensic Science Service in Britain uses a factor of 0.26 with probe 33.15 and Hinf I-digested DNA. This is not a mean value. It is the most conservative figure taken from the 4 kb to 6 kb region where bands are most common. Examination of over 700 profiles shows the figure to be conservative, that is, the frequency of bands at any given position does not exceed 0.26. If one person has a profile consisting of 10 resolvable bands, the probability of an unrelated person having the identical pattern is approximately $(1/4)^{10} = 1/1048576$. If the profile consisted of 18 bands, the probability would be $(1/4)^{18} = 1/68719475200$. With a world population of 5.2 billion, and increasing at the rate of three people per second, there is little question that 1 in 69 billion is significantly small.

Band sharing is considerable more common in biologically related individuals. the Probability that a band in sibling A is also present in sibling B is approximately 0.5. If 18 bands are resolvable in sibling A when the digested DNA is hybridized with a

minisatellites probe, the probability that the same bands will be detected in the DNA profile from sibling B is $(1/2)^{18} = 1/262144$ or approximately 4×10^{-6} .

A critical review together with suggested improvements of some methods of statistical analysis of data on DNA fingerprinting has been carried out by Cohen (1990). The approach used by Alex Jeffreys received detailed attention in this article. Cohen's paper exemplifies a number of possible pitfalls in terms of data collection and handling in DNA profiling. The points of contention include (1) the sampling procedures used, (2) the small number of well defined populations sampled, (3) the suggestion that some loci analyzed may not be independent, that is, they may be linked, (4) the assumptions that DNA fragments occur independently and with constant frequency within a size class, and (5) use of the geometric mean instead of the arithmetic mean in some calculations. The degree to which the points made are real or perceived in terms of the final results in forensic identification cases will, in all likelihood, be resolved only when the databases are sufficiently expanded to provide large enough sample numbers to verify or reject the statistical approaches used.

GLOSSARY

Adenine: A purine base, occurring in ribonucleic acid and deoxyribonucleic acid.

Agarose: The gelling component of agar; possesses a double-helical structure which forms a three-dimensional framework capable of holding water molecules in the interstices.

Alu: Alu is a family of repeat DNA sequences, cleaved by the restriction enzyme Alu I, dispersed throughout genomes of many animal species. The family consists of about 500,000 copies, at 300 bp each, per human genome. (DNA Fingerprinting, Kirby, 1990)

Allele: One of a pair of genes, or of multiple forms of a gene, located at the same locus of homologous chromosomes.

Anneal: To recombine strands of denatured deoxyribonucleic acid that were separated.

Antiparallel: Pertaining to parallel molecules that point in opposite direction as the strands of deoxyribonucleic acid.

Base Pair: Two nitrogenous bases, one purine and one pyrimidine, that pair in double-stranded deoxyribonucleic acid.

Chromosome: Any of the complex, threadlike structures seen in animal and plant nuclei.

Cytosine: A pyrimidine occurring as a fundamental unit or base of nucleic acids.

Deoxyribonucleic Acid (DNA): A linear polymer made up of deoxyribonucleotide repeating units (composed of the sugar 2-deoxyribose, phosphate, and a purine or pyrimidine base); most molecules are double-stranded and antiparallel, resulting in a right handed helix structure; carrier of genetic information, which is encoded in the sequence of bases.

DNA fingerprinting: A forensic identification technique that enables virtually 100% discrimination between individuals from small samples of blood or semen, using probes for hypervariable minisatellite deoxyribonucleic acid.

Double Helix: The shape of two linear strands of DNA assumed when bonded together.

Electrophoresis: The process of separating charged molecules, for example, negatively charged DNA fragments, in a porous medium such as agarose, by the application of an electric field. DNA separates according to size with the small fragments moving most rapidly.

Forensic Sciences: The application of scientific facts to legal problems

Gene: The basic unit of inheritance.

Guanine: A purine base, occurs naturally as a fundamental component of nucleic acids.

Heterozygous: The presence of different alleles at corresponding homologous chromosome loci.

Homozygous: The presence of identical alleles at corresponding homologous chromosome loci.

Loci: see locus

Locus: (plural, loci) The specific position on a chromosome.

Marker: see size marker

Nucleic Acid: A large, acidic, chainlike molecule containing phosphoric acid, sugar, and purine and pyrimidine bases.

Nucleotide: A building block unit of nucleic acid.

Polymerase Chain Reaction (PCR): A technique for copying and amplifying the complementary strands of a target deoxyribonucleic acid molecule.

Polymorphism: The Presence of multiple alleles of a gene in a population.

Primer: A short ribonucleic acid (RNA) sequence that is complementary to a sequence of DNA, providing an initiation point for addition of deoxyribonucleotides in DNA replication.

Probe: A single-stranded DNA, or mRNA, capable of being tagged with a tracer, such as ^{32}P , and hybridized to its complementary sequence.

Purine: A heterocyclic compound containing fused pyrimidine and imidazole rings; adenine and guanine are the purine components of nucleic acids and coenzymes.

Pyrimidine: A heterocyclic organic compound containing nitrogen atoms at positions 1 and 3; naturally occurring derivatives are components of nucleic acids and coenzymes.

Restriction Enzyme: see restriction endonucleases.

Restriction Endonucleases: Enzymes (molecular scissors) that cleave double-stranded DNA at specific palindromic base recognition sequences.

Restriction Fragment Length Polymorphism (RFLP): Variation in the length of DNA fragments produced by a restriction endonuclease (an enzyme) that cuts at a polymorphic

locus. The polymorphism may be either in the restriction enzyme site or in the number of tandem repeat between the cut points.

Reverse Dot Blot: Specific detection method used for DNA amplified by the polymerase chain reaction (PCR). The probe is bound to the typing strip, and the PCR product applied after.

Size Marker: DNA fragments of known molecular weight and base pair length.

Slot Blot: A diagnostic tool used in DNA analysis to determine how much human DNA has been extracted from a sample. Useful in making decisions about how much sample to use in various typing procedures.

Southern Blotting: A technique developed by E. Southern for the direct transfer of DNA fragments from an agarose gel onto a solid support such as a nylon membrane. The transfer occurs by salt solution capillary action.

Thymine: A pyrimidine component of nucleic acid.

Variable Number Tandem Repeats (VNTR): The variable number of repeat core base pair sequences at specific loci in the genome. The variation in length of the alleles formed from the repeats provides the basis for unique individual identification.

Bibliography

- ABC News, 1999. What's Next in Ramsey Case? [On-line]. Available:
http://abcnews.go.com/sections/us/DailyNews/jonbenetgrandjury_feature.html
- Adler, Jerry, and John McCormick. The DNA Detectives. [On-line]. Available WWW:
http://www.newsweek.com/nw-srv/20_98b/printed/us/so/front/htm.
- Balazs I, Baird M, Clyne M, and Meade E. 1989. Human population genetic studies of five hypervariable DNA loci. *Am. j. Hum. Genet.* 44:182-198.
- Baur MP, Elston RC, Gurtler H, Henningsen K, Hummel K, Matsumoto H, Mayr W, Moris JW, Niejenhais L, Polesky H, Salmon D, Valentin J, and Walker R. 1986. No fallacies in the formation of the paternity index. *Am. J. Hum. Genet.* 39:528-536.
- Beckner, Mark 1997. The Jonbenet Ramsey Case. [On-line]. Available WWW:
<http://www.courtvt.com/casefiles/jonbenet/links.html>
- Bellamy, Patric 1999. The Murder of Jonbenet Ramsey. [On-line}. Available
<http://va.crimelibrary.com/ramsey/ramseymain.htm>
- Budowle B, and Monson KL. 1989. A statistical approach for VNTR analysis. *In Proceedings-DNA symposium*. International Symposium on the Forensic Aspects of DNA Analysis. Government Publishing Office, Washington, D.C. (in press)
- CNN. (1996, November 14). DNA Experts link sock blood to Nicole. [On-line]. Available WWW: <http://europe.cnn.com/US/9611/14/simpson.thursday/>
- CNN O.J. Simpson Trial News: The Evidence. 1995. [On-line]. Available WWW:
<http://www.cnn.com/US/OJ/evidence/index.html>
- Cohen JE. 1990. DNA fingerprinting for forensic identification: Potential effects on data interpretation of subpopulation heterogeneity and band number variability. *Am. J. Hum. Genet.* 46:358-368.
- Cooper, Geoffrey M. 1997. *The Cell: A Molecular Approach*. Sunderland: Sinauer Associates, Inc.
- Court TV Library. (1997, January 28). Court TV Casefiles: O. J. Simpson Transcript. [On-line]. Available WWW:
<http://www.courttv.com/casefiles/simpson/transcripts/jan/jan28.html>
- FRONTLINE: Cotton's Wrongful Conviction. (1998). [On-line]. Available WWW:
<http://www.ps.org/wgbh/pages/frontline/shows/dna/cotton/summary.html>.

- Garrett, Reginald H., and Charles M. Grisham. 1999. *Biochemistry*, 2nd ed. Fort Worth: Saynders College Publishing.
- Georges M, Lequarre AS, Castelli M, Hanset R, and Vassart G. 1988. DNA fingerprinting in domestic animals using four different minisatellite probes. *Cytogenet. Cell Genet.* 47:127-131
- Graphic Gallery: Polymerase Chain Reaction. 1999. [On-line]. Available WWW: <http://www.accessexcellence.org/AB/GG/polymerase.html>.
- Griffiths, Anthony J. F., Jeffrey H. Miller, David T. Suzuki, Richard C. Lewontin, William M. Gilbert. 1996. *An Introduction to Genetic Analysis*, 6th ed. New York: W. H. Freeman and Company.
- How was a murderer traced through blood samples? [On-line]. Available WWW: <http://www.faseb.org/genetics/gsa/careeres/bro-05.htm>.
- Inman, Keith., and Norah Rudin. 1997. *An Introduction to Forensic DNA Analysis*. Boca Ration: CRC Press.
- Jeffreys AJ, Wilson V, and thein SL. 1985. Individual-specific 'fingerprints' of human DNA. *Nature* 316:76-79
- Jeffreys AJ. 1987. Highly variable minisatellites and DNA fingerprints. *Biochem. Soc. Trans.* 15:309-317.
- Jeffreys AJ and Morton DB. 1987a. DNA fingerprints of dogs and cats. *Anim. Genet.* 18:115
- Kirby, Lorne T. 1990. *DNA Fingerprinting*. New York: Stockton Press.
- Li CC and Chakravarti A. 1985. Basic fallacies in the formulation of the paternity index. *Am. J. Hum. Genet.* 37:809-818.
- McCullen, Kevin 1997. Ramseys Decline Offer to Observe DNA Testing. [On-line]. Available WWW: <http://www.sunflower.org/%7Emapek/mar97.html>
- Parker, Sybil P. 1997. *Dictionary of Bioscience*. New York: McGraw-Hill.
- Prescott, Lansing M., John P. Harley, and Donald A. Klein. 1999. *Microbiology*, 4th ed. Boston: WCB/McGraw-Hill
- Raven & Johnson. 1997. *Biology*, 4th ed. New York: Then McGraw-Hill Companies, Inc.
- Rocky Mountain News 1998. Jonbenet Ramsey Case. [On-line]. Available <http://insidedenver.com/extra/ramsey.html>

- Science: Alec Jeffreys – Genetic Evidence. 1999. [On-line]. Available
<http://www.britcoun.org/science/science/personalities/text/ukperson/jeffreys.htm>
- The Actual Cases: Chapter IV. [On-line]. Available:
<http://crimemagazine.com/dnapart3.htm>
- The Detroit News. (1995, August 4). Microbiologist calls O.J. DNA test unreliable. [On-line]. Available WWW: <http://www.detnews.com/menu/stories/12706.htm>
- The O.J. Simpson Case. 1999. [On-line]. Available WWW:
<http://www.wagnerandson.com/oj/OJ.htm>
- The Seattle Times, 1997. Stun Gun May Have Been Used to Kill Jonbenet Ramsey. [On-line]. Available
http://www.seattletimes.com/extra/browse/html97/altjonb_122197.html
- Thompson, Molly 1999. If Not Brother, then whom? [On-line]. Available
<http://www.channel3000.com/townpulse/townpulse-stories-990526-105929.html>
- Tsai, Catherine 1999. Ramsey's Lie Detector Arrangements Hit Snag. [On-line]. Available
<http://cnews.tribune.com/news/story/0,1162,wbdc-nation-41077,00.html>
- Walraven, Jack. (1999, March 20). The Simpson Trial Transcripts. [On-line]. Available
WWW: <http://207.175.199.183/~walraven/simpson/#transcripts>
- USA Today. (1997, February 12). O. J. Simpson Civil Trial. [On-line]. Available WWW:
<http://www.usatoday.com/news/index/nns0.htm>