Chapter 11

Human Identity Testing

DNA can be used to establish paternity, match forensic samples to suspects, identify human remains, confirm or disconfirm genealogical relationships, and perform the gruesome task of identifying body parts in accidents and war zones. Together, these tasks fall under the rubric of *human identity testing* because the purpose is to use DNA to identify a person.

In this chapter, we will examine two examples of human identity testing. The first is the *Combined DNA Index System* (CODIS) used in forensics in the US. The second is genealogical testing.

11.1 CODIS

As the number of detectable polymorphisms exploded in the 1980s, it became apparent that DNA could be used to identify perpetrators of criminal acts. The classic case was the sexual assault of a woman where the rapist left semen or other biological material at the crime scene. With PCR, one could obtain a large amount of DNA for analysis from a very tiny biological sample. With a large number of polymorphisms, one could arrive at astonishing low probabilities someone other that a suspect with DNA alleles that matched those of the crime scene DNA could have committed the act.

In the United States this forensic technology was initially developed by individual law enforcement agencies in individual states. This created a bureaucratic problem. Different districts genotyped different polymorphisms and used different software for storage. Hence, if a DNA profile were developed from, say, a sexual assault, it was very difficult and time consuming–sometimes even impossible–to search the myriad number of data bases for a match.

In 1990, the U.S. Federal Bureau of Investigation (FBI) began a pilot project with the aim of establishing a national, searchable data base of DNA profiles for crime scene biological material and for some people arrested for crimes. In 1994, the DNA Identification Act (42 U.S.C. §14132) established a national data base of DNA and authorized the FBI to maintain it. Together, this data base

(formally called the National DNA Index System or NDIS), the software used to enter samples into it and search it, and the standards for quality control became known as the Combined DNA Index System or CODIS. It became operational in 1998. Originally, only forensic samples and convicted offenders were included, but soon DNA samples that permitted the identification of missing persons were added as were the DNA profiles of some types of arrestees.

Contrary to popular belief, the CODIS database does not contain names or any other identifying information on a person.[1] That information is kept at a local level, so while CODIS provide matches for DNA profiles, identification sometimes requires communication among different local agencies.

An overview of how CODIS works will reinforce knowledge about previous topics such as tandem repeat polymorphisms and electrophoresis.

11.1.1 The CODIS loci

The loci currently used in CODIS are 13 tandem repeats called *short tandem repeats* or *STR*s plus the AMEL locus. They are depicted in Figure 11.1. Some of the loci are genes that code for proteins (TPOX, FGA, TH01, VWA, and AMEL). The others are STRs in noncoding region. They are given by the notation DnS$xxxxx$ where n is chromosome number and $xxxxx$ is the order in which the polymorphism was identified.

Genotyping for these loci requires appropriate robots to perform the PCR and scan and record the results as well as a laboratory kit containing the reagents and other chemicals necessary to separate and purify the DNA and to select the sections of DNA for amplification in the PCR process.

Let's imagine a crime with three prime suspects: Moe, Larry and Curly. DNA left at the crime scene would be analyzed and a biological specimen (usually a swab of the inner cheek) would be obtained from the three suspects. The first matter to test for is the sex of the person who left the DNA at the crime scene. Moe, Larry and Curly are all males so if the crime scene DNA is from a female, we can eliminate all three from further consideration.

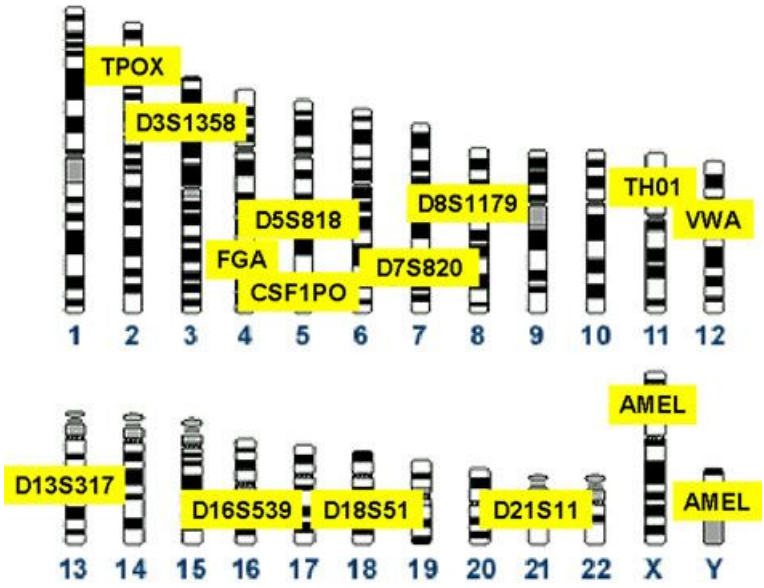
One could test for sex by using a karyotype, but that is both expensive and time consuming. Instead, CODIS uses a gene (AMEL in Figure 11.1) that is located on both the X and the Y chromosomes. The number of tandem repeats for the AMEL gene on the X, however, differs from that on the Y. Hence, if the sample contains only the AMEL from the X, the person is a female. If it contains both the X and the Y AMEL, then it comes from a male. We assume that the crime scene DNA indicates a male perp.

11.1.1.1 A genotyping example

Let's examine a single CODIS locus, TH01, and go through the process of genotyping in detail. TH01 is an acronym for the gene that codes for the tyrosine hydroxylase. This enzyme converts the animo acid tyrosine into DOPA which

[1] See `http://www.fbi.gov/about-us/lab/biometric-analysis/codis/codis-and-ndis-fact-sheet`

Figure 11.1: The CODIS loci.



from `http://www.cstl.nist.gov/div831/strbase/fbicore.htm`

serves as a precursor for melanin in skin cells and for the neurotransmitters dopamine and norepinephrine in neurons. The tandem repeat is the four nucleotide sequence `AATG` that occurs in the first intron of the gene. There are 21 different known types of repeats, although many are rare (`http://www.cstl.nist.gov/div831/strbase/str_TH01.htm`). To make this example simple, let us consider those repeats with a frequency greater than 1% in the population. They are: 6, 7, 8, 9, 9.3, and 10 repeats.

The first thing that you may wonder about is this 9.3 business. What is that? Nature is seldom obliging to logic and neatness and this is an example. The notation 9.3 means that there are 9 repeats of `AATG` and one case in which only three nucleotides are present instead of all four. In this allele, `AATG` is repeated 6 times, then comes the three nucleotide sequence `ATG` followed by three more `AATG`s.

The procedure for genotyping is to use PCR to obtain large quantities of DNA from the region of the TH01 gene containing the repeats. Next the PCR products are subjected to electrophoresis which will separate the DNA according to size which in the case of an STR is tantamount to the number of repeats. There will be two lanes in the electrophoresis. The first uses DNA from the CODIS kit. It contains DNA corresponding to all of the known alleles for the gene. The second is from the crime scene DNA or from the suspect.

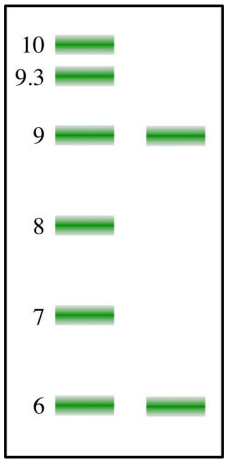
The electrophoresis is performed on single stranded DNA. If we simply looked at the medium after electrophoresis, we would see nothing because DNA is

colorless. To highlight the fragments we must "bathe" the medium in probes– single-stranded DNA that is complementary to the strands used in the PCR and that carries a "lightbulb." The lightbulb is an analogy for a radioactive label or fluorescent dye that allows it to be visible.

The probe is allowed to bind (aka hybridize) with its complementary section in the medium. Then special procedures are used to wash away any remaining single stranded probe that has not bound to the DNA from the crime scene or suspect. An example of the result is provided in Figure 11.2.

Here, the left hand panel is for reference. It has electrophoretic bands for all of the major alleles. The right hand panel is a sample either from the crime scene or a suspect. In this case, the bands indicate that the person is a heterozygote having a six repeat allele and a nine repeat allele. Assume that this is from the biological specimen at the crime scene.

Figure 11.2: Major TH01 alleles after electrophoresis and probe identification.



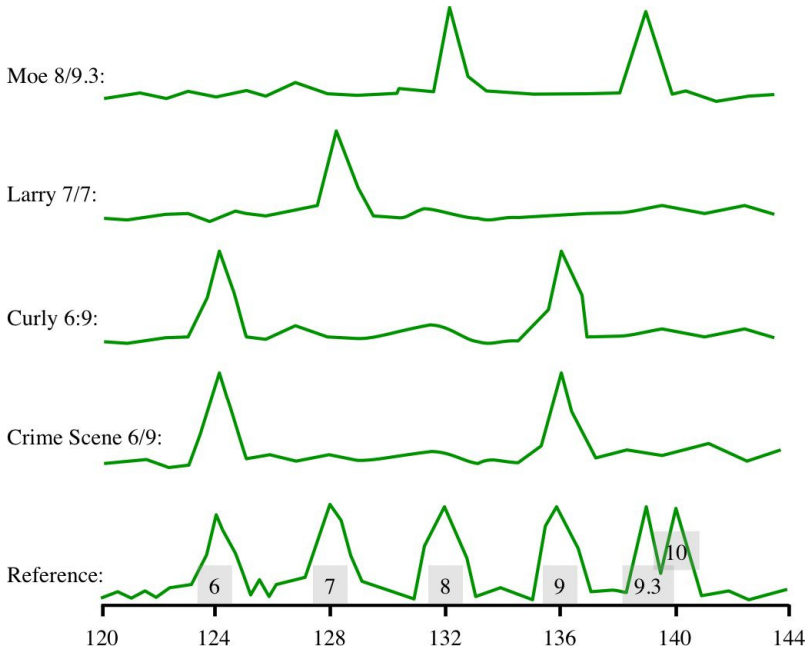
We now want to genotype Moe, Larry and Curly. We could do so and examine the electrophoretic bands as in Figure 11.2, but CODIS, as well as all other modern genotyping technologies, automates this procedure. Imagine a laser that is sensitive to the green wavelength. It starts at the bottom of the results from electrophoresis in Figure 11.2 and then proceeds to the top while computers record the intensity of the green signal. There be some background noise, but as the laser approaches the first band (the six repeat allele), the intensity of the signal increases, reaches a peak in the middle of the band and then recedes as the laser moves away from the band. There will be some low intensity signal from background noise until the laser begins to sense the green from the seven repeat allele.

The result of this scanning is the bottom panel in Figure 11.3. The profile from the crime scene is immediately above it showing once again that the perpetrator has genotype 6/9 (for a six repeat allele and a nine repeat allele). The results for Moe (8/9.3 genotype), Larry (homozygote for the 7 repeat allele) and Curly (6/9 genotype) are given in the top three panels of the figure.

Faced with the observation that Curly has the same genotype as the perpetrator, many people conclude that Curly is the guilty party. But that would be wrong. The correct inference is that the data show that neither Moe nor Larry is the perpetrator because their DNA differs from the crime scene DNA.

The problem with Curly is that the 6/9 genotype may be common in the general population. Curly could be innocent but by chance have the same genotype as the perp. From the Promega website (Promega is one manufacturer

Figure 11.3: Laser intensity profiles for genotyping at the TH01 locus.



of CODIS DNA kits), the probability of this genotype in American Causasian population is .06.

To examine Curly further, we must compare his and the crime scene DNA on all of the other loci. If there is a match, then we multiply the probabilities. For example, suppose that Curly and the crime scene DNA match on a second locus where the population base rate is .17. The probability of randomly picking a person who is a match on both the TH01 and the second locus would be the product of these two probabilities or $.06 \times .17 = .01$. Thew probability of a random match at a third locus is then .01 times the base rate for that genotype. By multiplying small probabilities over all 13 of the CODIS loci, the probability of a match from a person from the general public is usually one in many tens of millions to several billions.

11.1.2 CODIS in reality

Although the account above accurately describes the logic of CODIS, it omits many of the gory–but important–details. To achieve efficiency, the genotyping of all STRs for all loci is performed simultaneously after one series of PCR amplifications. The peaks from laser scanning have much greater resolution than those depicted in Figure 11.3. Matching of DNA sample peaks with their reference counterparts is performed by complex mathematical algorithms that also output statistics about the certainty of a match and flag potential problems.

Finally, all results are examined and vetted by a professional specially trained in the CODIS protocols.

11.2 Genealogy

X.X To be completed.