Forensic Genetics

Summer Institute in Statistical Genetics, July 20-22, 2016

Bruce Weir: bsweir@uw.edu

Contents

Sources of Data

Phenotype	Mendel's peas
	Blood groups
DNA	Restriction sites, RFLPs
	Length variants, VNTRs, STRs
	SNPs
	Nucleotide sequences

Mendel's Data

Dominant Form		Recessive Form	
	Seed characters		
5474	Round	1850	Wrinkled
6022	Yellow	2001	Green
	Plant characters		
705	Grey-brown	224	White
882	Simply inflated	299	Constricted
428	Green	152	Yellow
651	Axial	207	Terminal
787	Long	277	Short

ABO System

Human ABO blood groups discovered in 1900. ABO gene on human chromosome 9 has 3 alleles: A, B, O. Six genotypes but only four phenotypes (blood groups):

Genotypes	Phenotype
AA, AO	A
BB, BO	B
AB	AB
OO	O

Charlie Chaplin and ABO Testing

Relationship	Person	Blood Group	Genotype
Mother	Joan Berry	A	AA or AO
Child	Carol Ann Berry	B	BB or BO
Alleged Father	Charles Chaplin	O	OO

The obligate paternal allele was B, so the true father must have been of blood group B or AB.

Berry v. Chaplin, 74 Cal. App. 2d 652

Electrophoretic Detection

Charge differences among alleles ("allozymes") of soluble proteins lead to separation on electrophoretic gels. Protein loaded at one end of a slab gel and an electric current is passed through the gel. Allozymes migrate according to their net charge: separation of alleles depends on how far they migrate in a given amount of time.

This techniques was the first to allow large-scale collection of genetic marker data. The data in this case reflected variation in the amino acid sequences of soluble proteins.

Alec Jeffreys

For forensic applications, the work of Alec Jeffreys with on Restriction Fragment Length Polymorphisms (RFLPs) or Variable Number of Tandem Repeats (VNTRs) also used electrophoresis. Different alleles now represented different numbers of repeat units and therefore different length molecules. Smaller molecules move faster through a gel and so move further in a given amount of time.

Initial work was on mini-satellites, where repeat unit lengths were in the tens of bases and fragment lengths were in thousands of bases. Jeffrey's multi-locus probes detected regions from several pats of the genome and resulted in many detectable fragments per individual. This gave high discrimination but difficulty in assigning numerical strength to matching profiles.

Jeffreys et al. 1985. Nature 316:76-79 and 317: 818-819.

Single-locus Probes

Next development for gel-electrophoresis used probes for single mini-satellites. Only two fragments were detected per individual, but there was difficulty in determining when two profiles matched.

The technology also required "large" amounts of DNA and was not suitable for degraded samples.

PCR-based STR Markers

The ability to increase the amount of DNA in a sample by the Polymerase Chain Reaction (PCR) was of substantial benefit to forensic science. The typing technology changed to the use of capillary tube electrophoresis, where the time taken by a DNA molecule to pass a fixed point was measured and used to infer the number of repeat units in an allele.

A good source is "Following multiplex PCR amplification, DNA samples containing the length-variant STR alleles are typically separated by capillary electrophoresis and genotyped by comparison to an allelic ladder supplied with a commercial kit. "
Butler JM. Short tandem repeat typing technologies used in human identity testing. BioTechniques 43:Sii-Sv (October 2007) doi 10.2144/000112582

STR markers: CTT set

(http://www.cstl.nist.gov/biotech/strbase/seq_info.htm)

Locus	Structure	Chromosome	Usual No. of repeats
CSF1PO	$[AGAT]_n$	5q	6–16
TPOX	$[AATG]_n$	2p	5–14
TH01*	$[AATG]_n$	11p	3–14

* "9.3" is $[AATG]_6ATG[AATG]_3$

Length variants detected by capillary electrophoresis.

"CTT" Data - Forensic Frequency Database

CSF1P0		TPOX		TH01	
11	12	8	11	7	8
11	13	8	8	6	7
11	12	8	11	6	7
10	12	8	8	6	9
11	12	8	12	9	9.3
10	12	9	11	6	7
10	13	8	11	6	6
11	12	8	8	6	9.3
9	10	8	9	7	9.3
11	12	8	8	6	8
11	13	8	11	7	9
11	12	8	11	6	9.3
10	11	8	8	7	9.3
10	10	8	11	7	9.3
9	10	8	8	6	9.3
11	12	9	11	9	9.3
9	11	9	11	9	9.3
11	12	8	8	6	7
10	10	9	11	6	9.3
10	13	8	8	8	9.3

Sequencing of STR Alleles

"STR typing in forensic genetics has been performed traditionally using capillary electrophoresis (CE). Massively parallel sequencing (MPS) has been considered a viable technology in recent years allowing high-throughput coverage at a relatively affordable price. Some of the CE-based limitations may be overcome with the application of MPS ... generate reliable STR profiles at a sensitivity level that competes with current widely used CE-based method."

Zeng XP, King JL, Stoljarova M, Warshauer DH, LaRue BL, Sajantila A, Patel J, Storts DR, Budowle B. 2015. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Science International: Genetics 16:38-47.

Single Nucleotide Polymorphisms (SNPs)

"Single nucleotide polymorphisms (SNPs) are the most frequently occurring genetic variation in the human genome, with the total number of SNPs reported in public SNP databases currently exceeding 9 million. SNPs are important markers in many studies that link sequence variations to phenotypic changes; such studies are expected to advance the understanding of human physiology and elucidate the molecular bases of diseases. For this reason, over the past several years a great deal of effort has been devoted to developing accurate, rapid, and cost-effective technologies for SNP analysis, yielding a large number of distinct approaches. "

Kim S. Misra A. 2007. SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng. 2007;9:289-320.

Phase 3 1000Genomes Data

- 84.4 million variants
- 2504 individuals
- 26 populations

www.1000Genomes.org

Whole-genome Sequence Studies

One current study is the NHLBI Trans-Omics for Precision Medicine (TOPMed) project. www.nhlbiwgs.org
In the first data freeze of Phase 1 of this study:

Description	Variants Passing Filters
Total Number of SNPs	86,974,704
Singletons	35,883,567
% Singletons	41.3%
Nonsynonymous	599,883
Singletons	305,479
% Singletons	50.9%
Stop Gains	13,436
Singletons	8,067
% Singletons	60.0%
# in dbSNP (142)	43,141,344
% in dbSNP	49.6%

Abecasis et al. 2016. ASHG Poster

Probability Theory

We wish to attach probabilities to different kinds of events (or hypotheses or propositions):

- Event A: the next card is an Ace.
- Event R: it will rain tomorrow.
- Event C: the suspect left the crime stain.

Probabilities

Assign probabilities to events: Pr(A) or p_A or even p means "the probability that event A is true." All probabilities are conditional, so should write Pr($A|E$) for "the probability that A is true given that E is known."

No matter how probabilities are defined, they need to follow some mathematical laws in order to lead to consistent theories.

First Law of Probability

$$0 \leq \Pr(A|E) \leq 1$$

$$\Pr(A|A) = 1$$

If A is the event that a die shows an even face (2, 4, or 6), what is E? What is $\Pr(A|E)$?

Second Law of Probability

If A, B are mutually exclusive given E

$$\Pr(A \text{ or } B|E) \;=\; \Pr(A|E) + \Pr(B|E)$$

$$\text{so } \Pr(\bar{A}|E) \;=\; 1 - \Pr(A|E)$$

($\bar{A}$ means not-A).

If A is the event that a die shows an even face, and B is the event that the die shows a 1, verify the Second Law.

Third Law of Probability

$$\Pr(A \text{ and } B|E) \;=\; \Pr(A|B,E) \times \Pr(B|E)$$

If A is event that die shows an even face, and B is the event that the die shows a 1, verify the Third Law.

Independent Events

Events A and B are independent if knowledge of one does not affect probability of the other:

$$\begin{aligned} \Pr(A|B) &= \Pr(A) \\ \Pr(B|A) &= \Pr(B) \end{aligned}$$

Therefore, for independent events

$$\Pr(A \text{ and } B) = \Pr(A)\Pr(B)$$

This may be written as

$$\Pr(AB) = \Pr(A)\Pr(B)$$

Law of Total Probability

Because B and $\bar{B}$ are mutually exclusive and exhaustive:

$$\Pr(A) = \Pr(A|B)\Pr(B) + \Pr(A|\bar{B})\Pr(\bar{B})$$

If A is the event that die shows a 3, B is the event that the die shows an even face, and $\bar{B}$ the event that the die shows an odd face, verify the Law of Total Probability.

IF B_1, B_2, B_3 are mutually exclusive and exhaustive:

$$\begin{aligned}\Pr(A) = &\Pr(A|B_1)\Pr(B_1) + \Pr(A|B_2)\Pr(B_2) \\ &+ \Pr(A|B_3)\Pr(B_3)\end{aligned}$$

Odds

The odds $O(A)$ of an event A are the probability of the event being true divided by the probability of the event not being true:

$$O(A) = \frac{\Pr(A)}{\Pr(\bar{A})}$$

This can be rearranged to give

$$\Pr(A) = \frac{O(A)}{1 + O(A)}$$

Odds of 10 to 1 are equivalent to a probability of 10/11.

Bayes' Theorem

The third law of probability can be used twice to reverse the order of conditioning:

$$\begin{aligned} \Pr(E|A) &= \frac{\Pr(E \text{ and } A)}{\Pr(A)} \\ &= \frac{\Pr(A|E)\Pr(E)}{\Pr(A)} \end{aligned}$$

Odds Form of Bayes' Theorem

From the third law of probability

$$\begin{aligned} \Pr(E|A) &= \Pr(A|E)\Pr(E)/\Pr(A) \\ \Pr(\bar{E}|A) &= \Pr(A|\bar{E})\Pr(\bar{E})/\Pr(A) \end{aligned}$$

Taking the ratio of these two equations:

$$\frac{\Pr(E|A)}{\Pr(\bar{E}|A)} = \frac{\Pr(A|E)}{\Pr(A|\bar{E})} \times \frac{\Pr(E)}{\Pr(\bar{E})}$$

Posterior odds = likelihood ratio $\times$ prior odds.

AIDS Example

Suppose the event E of AIDS occurs 1 in 10,000 people chosen at random.

Suppose a test procedure has two outcomes: A (positive) and B (negative). The probability of a positive result is 0.99 if the person has AIDS, and 0.05 if the person does not have AIDS. What is the probability that a person has AIDS if she tests positive?

AIDS Example

The problem is to determine $\Pr(E|A)$ when $\Pr(A|E)$ is known. This requires Bayes' theorem, and the term $\Pr(A)$ follows from the Law of Total Probability.

$$
\begin{aligned}
\Pr(E) &= \\
\Pr(\bar{E}) &= \\
\Pr(A|E) &= \\
\Pr(A|\bar{E}) &= \\
\Pr(A) &= \\
\Pr(E|A) &=
\end{aligned}
$$

Birthday Problem

Forensic scientists in Arizona looked at the 65,493 profiles in the Arizona database and reported that two profiles matched at 9 loci out of 13. They reported a "match probability" for those 9 loci of 1 in 754 million. Are the numbers 65,493 and 754 million inconsistent?
(Troyer et al., 2001. Proc Promega 12th Int Symp Human Identification.)

To begin to answer this question suppose that every possible profile has the same profile probability P and that there are N profiles in a database (or in a population). The probability of at least one pair of matching profiles in the database is one minus the probability of no matches.

Birthday Problem

Choose profile 1. The probability that profile 2 does not match profile 1 is $(1-P)$. The probability that profile 3 does not match profiles 1 or 2 is $(1-2P)$, etc. So, the probability P_M of at least one matching pair is

$$P_M = 1-\{1(1-P)(1-2P)\cdots[1-(N-1)P]\}$$

$$\approx 1-\prod_{i=0}^{N-1} e^{-iP} \approx 1-e^{-N^2P/2}$$

If $P = 1/365$ and $N = 23$, then $P_M = 0.51$. So, approximately, in a room of 23 people there is greater than a 50% probability that two people have the same birthday.

Birthday Problem

If $P = 1/(754 \text{ million})$ and $N = 65,493$, then $P_M = 0.98$ so it is highly probable there would be a match. There are other issues, having to do with the four non-matching loci, and the possible presence of relatives in the database.

If $P = 10^{-16}$ and $N = 300$ million, then $P_M =$ is essentially 1. It is almost certain that two people in the US have the same rare DNA profile.

Statistics

- Probability: For a given model, what do we expect to see?

- Statistics: For some given data, what can we say about the model?

- Example: A marker has an allele A with frequency p_A.
  - Probability question: If $p_A = 0.5$, and if alleles are independent, what is the probability of AA?
  - Statistics question: If a sample of 100 individuals has 23 AA's, 48 Aa's and 29 aa's, what is an estimate of p_A?

Binomial distribution

Imagine tossing a coin n times, when every toss has the same chance p of giving a head:

The probability of x heads in a row is

$$p \times p \times \ldots \times p \;=\; p^x$$

The probability of $n - x$ tails in a row is

$$(1-p) \times (1-p) \times \ldots \times (1-p) \;=\; (1-p)^{n-x}$$

The number of ways of ordering x heads and $n - x$ tails among n outcomes is $n!/[x!(n-x)!]$.

Binomial distribution

Combining the probabilities of x successive heads, $n-x$ successive trials, and the number of ways of ordering x heads and $n-x$ tails: the binomial probability of x successes (heads) in n trials (tosses) is

$$\Pr(x|p) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

Binomial distribution

The probabilities of x heads in $n = 4$ tosses of a coin when the chance of a head is 1/2 at each toss:

No. heads x	Probability $\Pr(x\|p)$
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16

Note that $0! = 1$ and $p^0 = 1$.

Binomial distribution

Find the binomial probabilities, for a sample of size $n = 4$ alleles, when the chance that each allele is of type A is 1/10.

No. A's	Probability
0	
1	
2	
3	
4	

TRANSFER EVIDENCE

Relevant Evidence

Rule 401 of the US Federal Rules of Evidence:

"Relevant evidence" means evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence.

Single Crime Scene Stain

Suppose a blood stain is found at a crime scene, and it must have come from the offender. A suspect is identified and provides a blood sample. The crime scene sample and the suspect have the same (DNA) "type."

The prosecution subsequently puts to the court the proposition (or hypothesis or explanation):

H_p: The suspect left the crime stain.

The symbol H_p is just to assist in the formal analysis. It need not be given in court.

Transfer Evidence Notation

G_S, G_C are the DNA types for suspect and crime sample. $G_S = G_C$. I is non-DNA evidence.

Before the DNA typing, probability of H_p is conditioned on I.

After the typing, probability of H_p is conditioned on G_S, G_C, I.

Updating Uncertainty

Method of updating uncertainty, or changing $\Pr(H_p|I)$ to $\Pr(H_p|G_S, G_C, I$ uses Bayes' theorem:

$$\begin{aligned}\Pr(H_p|G_S, G_C, I) &= \frac{\Pr(H_p, G_S, G_C|I)}{\Pr(G_S, G_C|I)} \\ &= \frac{\Pr(G_S, G_C|H_p, I)\Pr(H_p|I)}{\Pr(G_S, G_C|I)}\end{aligned}$$

We can't evaluate $\Pr(G_S, G_C|I)$ without additional information, and we don't know $\Pr(H_p|I)$.

Can proceed by introducing alternative to H_p.

First Principle of Evidence Interpretation

To evaluate the uncertainty of a proposition, it is necessary to consider at least one alternative proposition.

The simplest alternative explanation for a single stain is:

H_d: Some other person left the crime stain.

Updating Odds

From the odds form of Bayes' theorem:

$$\frac{\Pr(H_p|G_S, G_C, I)}{\Pr(H_d|G_S, G_C, I)} = \frac{\Pr(G_S, G_C|H_p, I)}{\Pr(G_S, G_C|H_d, I)} \times \frac{\Pr(H_p|I)}{\Pr(H_d|I)}$$

i.e. Posterior odds = LR $\times$ Prior odds

where

$$\text{LR} = \frac{\Pr(G_S, G_C|H_p, I)}{\Pr(G_S, G_C|H_d, I)}$$

Questions for a Court to Consider

The trier of fact needs to address questions of the kind

- What is the probability that the prosecution proposition is true given the evidence,
  $\Pr(H_p|G_C, G_S, I)$?

- What is the probability that the defense proposition is true given the evidence,
  $\Pr(H_d|G_C, G_S, I)$?

Questions for Forensic Scientist to Consider

The forensic scientist must address different questions:

- What is the probability of the DNA evidence if the prosecution proposition is true,
  $\Pr(G_C, G_S | H_p, I)$?

- What is the probability of the DNA evidence if the defense proposition is true,
  $\Pr(G_C, G_S | H_d, I)$?

Important to articulate H_p, H_d. Also important not to confuse the difference between these two sets of questions.

Second Principle of Evidence Interpretation

Evidence interpretation is based on questions of the kind 'What is the probability of the evidence given the proposition.'

This question is answered for alternative explanations, and the ratio of the probabilities presented. It is not necessary to use the words "likelihood ratio". Use phrases such as:
'The probability that the crime scene DNA type is the same as the suspect's DNA type is one million times higher if the suspect left the crime sample than if someone else left the sample.'

Third Principle of Evidence Interpretation

Evidence interpretation is conditioned not only on the alternative propositions, but also on the framework of circumstances within which they are to be evaluated.

The circumstances may simply be the population to which the offender belongs so that probabilities can be calculated. Forensic scientists must be clear in court about the nature of the non-DNA evidence I, as it appeared to them when they made their assessment. If the court has a different view then the scientist must review the interpretation of the evidence.

Example

"In the analysis of the results I carried out I considered two alternatives: either that the blood samples originated from Pengelly or that the ... blood was from another individual. I find that the results I obtained were at least 12,450 times more likely to have occurred if the blood had originated from Pengelly than if it had originated from someone else."

Example

Question: "Can you express that in another way?"

Answer: "It could also be said that 1 in 12,450 people would have the same profile ... and that Pengelly was included in that number ... very strongly suggests the premise that the two blood stains examined came from Pengelly."

[Testimony of M. Lawton in R. v Pengelly 1 NZLR 545 (CA), quoted by Robertson & Vignaux, "Interpreting Evidence", Wiley 1995.]

Likelihood Ratio

$$\text{LR} = \frac{\Pr(G_C, G_S | H_p, I)}{\Pr(G_C, G_S | H_d, I)}$$

Apply laws of probability to change this into

$$\text{LR} = \frac{\Pr(G_C | G_S, H_p, I) \Pr(G_S | H_p, I)}{\Pr(G_C | G_S, H_d, I) \Pr(G_S | H_d, I)}$$

Likelihood Ratio

Whether or not the suspect left the crime sample (i.e. whether or not H_p or H_d is true) provides no information about his genotype:

$$\Pr(G_S|H_p, I) = \Pr(G_S|H_d, I) \;=\; \Pr(G_S|I)$$

so that

$$\text{LR} \;=\; \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)}$$

Likelihood Ratio

$$\text{LR} \;=\; \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)}$$

When $G_C = G_S$, and when they are for the same person (H_p is true):

$$\Pr(G_C|G_S, H_p, I) \;=\; 1$$

so the likelihood ratio becomes

$$\text{LR} \;=\; \frac{1}{\Pr(G_C|G_S, H_d, I)}$$

This is the reciprocal of the probability of the *match probability*, the probability of profile G_C, conditioned on having seen profile G_S in a different person (i.e. H_d) and on I.

Likelihood Ratio

$$\text{LR} \;=\; \frac{1}{\Pr(G_C|G_S, H_d, I)}$$

The next step depends on the circumstances I. If these say that knowledge of the suspect's type does not affect our uncertainty about the offender's type when they are different people (i.e. when H_d is true):

$$\Pr(G_C|G_S, H_d, I) \;=\; \Pr(G_C|H_d, I)$$

and then likelihood ratio becomes

$$\text{LR} \;=\; \frac{1}{\Pr(G_C|H_d, I)}$$

The LR is now the reciprocal of the *profile probability* of profile G_C.

Profile and Match Probabilities

Dropping mention of the other information I, the quantity $\Pr(G_C)$ is the probability that a person randomly chosen from a population will have profile type G_C. This profile probability usually very small and, although it is interesting, it is not the most relevant quantity.

Of relevance is the match probability, the probability of seeing the profile in a randomly chosen person after we have already seen that profile in a typed person (the suspect). The match probability is bigger than the profile probability. Having seen a profile once there is an increased chance we will see it again. This is the genetic essence of DNA evidence.

Likelihood Ratio

The estimated probability in the denominator of LR is determined on the basis of judgment, informed by I. Therefore the nature of I (as it appeared to the forensic scientist at the time of analysis) must be explained in court along with the value of LR. If the court has a different view of I, then the scientist will need to review the interpretation of the DNA evidence.

Random Samples

The circumstances I may define a population or racial group. The probability is estimated on the basis of a sample from that population. If the probability is written as P, then the likelihood ratio is $1/P$. If P is estimated to be 1 in a million, then LR is 1 million.

When we talk about DNA types, by "selecting a man at random" we mean choosing him in such a way as to be as uncertain as possible about his DNA type.

Convenience Samples

The problem with a formal approach is that of defining the population: if we mean the population of a town, do we mean *every* person in the town at the time the crime was committed? Do we mean some particular area of the town? One sex? Some age range?

It seems satisfactory instead to use a convenience sample, i.e. a set of people from whom it is easy to collect biological material in order to determine their DNA profiles. These people are not a random sample of people, but they have not been selected on the basis of their DNA profiles.

Meaning of Likelihood Ratios

There is a personal element to interpreting DNA evidence, and there is no "right" value for the LR. (There is a right answer to the question of whether the suspect left the crime stain, but that is not for the forensic scientist to decide.)

The denominator for LR is conditioned on the stain coming from an unknown person, and "unknown" may be hard to define. A relative? Someone in that town? Someone in the same racial group? (What is a race?)

Errors and Fallacies

Once the numerical strength of the evidence has been calculated, it is important that it be presented in a way that does not distort its meaning.

There are a series of common fallacies that can be avoided by careful application of the Principles of Evidence Interpretation.

Transposed conditional

Correct: *The evidence is 1000 times more likely if the suspect left the crime stain than if some unknown person left it.*

Incorrect: *It is 1000 times more likely that the suspect left the crime stain than some unknown person.*

The second statement is true only if the prior odds are 1.

Prosecution Fallacy

From the O.J. Simpson trials:

"You testify that there is ... a 1 in 71 chance that a pair of contributors at random could have left the stain." (Defense attorney at transcript p. 33,242.)

"The chances are at least 1-in-170 million that anybody else's DNA besides Simpson's could be contained in a blood drop found near the bodies of Nicole Brown Simpson and Ronald Goldman, testified Robin Cotton, director of Cellmark Diagnostics in Maryland." (Associated Press, 11/14/96)

Defense Fallacy

The matching DNA profile has a probability of 1 in 100,000. The crime was committed in a city of 1,000,000 people.

Correct: Therefore 10 people in the city are expected to have that profile.

Incorrect: Therefore the suspect (who has the profile) has a probability of 1 in 10 of being guilty.

Defense Fallacy

The fallacy is to assign equal (prior) probabilities to all 10 people who are expected to have the profile. Also note that the actual number of people in the city with the profile could be any number from 0 to 1,000,000. Expected numbers are not actual numbers.

Uniqueness Fallacy

The matching DNA profile has a probability of 1 in 1,000,000. The crime was committed in a city of 1,000,000 people.

Correct: Therefore 1 person in the city is expected to have that profile.

Incorrect: Therefore the suspect (who has the profile) is the guilty person.

Uniqueness Fallacy

The fallacy is not to recognize that the actual number of people in the city with the profile could be any number from 0 to 1,000,000.

Expected numbers are not actual numbers.

Database Fallacy

Probabilities needed in LR are estimated on the basis of a sample from a population. Ideally this sample is drawn from the population defined by H_d and I. This is not practical.

If H_d is true, the racial background of the suspect does not define the population to be sampled. Do not need a database of (exactly) the same ethnicity as the suspect.

Aitken and Taroni, Science and Justice 1998; 38:165–177

The laboratory results and the expert's testimony were described in 12 cases.

1. Ross v. State of Indiana (Indiana Court of Appeal, May 13, 1996). The DNA expert knew that the frequency of the DNA profile found in the vaginal swab sample and in the suspect's blood sample was 1 in 80,000. He said that Ross was the source of the seminal fluid.

2. State of Washington v. Gentry (125 Wash. 2d 570, 888 P.2d 1105 (1995)]. The matching DNA profile had a frequency of 0.18%. The expert said that the percentage of the population from which the blood found on the defendant's shoelaces could have originated is 0.18%.

Aitken and Taroni, Science and Justice 1998; 38:165–177

3. R v. Deen (Court of Appeal, Criminal Division, December 21, 1993). The matching DNA profile had a frequency of 1 in 3 million. The Prosecutor and Expert had this exchange: Q (Prosecutor) "So the likelihood of this being any other man but Andrew Deen is one in 3 million?" A (Expert): "In 3 million, yes." Q: "On the figure which you have established according to your research, the probability of it being anybody else being one in 3 million what is your conclusion?" Expert: "My conclusion is that the semen originated from Andrew Deen." Q. Are you sure of that?" A. "Yes."

Aitken and Taroni, Science and Justice 1998; 38:165–177

4. U.S. v. Jakobetz [955 F. 2d 786 (2nd Cir. 1992)]. In that case the FBI expert knew that the frequency of the matching DNA profile in the population is 1 in 300 million. The expert testified that the DNA profiles from the two samples constituted a match and calculated there was one chance in 300 million that the DNA from the semen sample could have come from someone in the Caucasian sample other than Jakobetz.

Aitken and Taroni, Science and Justice 1998; 38:165–177

5. Gordon (Court of Appeal, November 22, 1993, April 22, May 26, 1994). The DNA profile had a frequency of 1 in 10,500,000. The expert agreed that there was a visual match between the critical samples and the appellant's sample which showed a likelihood that the appellant was the rapist in each case.

6. Lonsdale (Court of Appeal, March 9, 16 1995.) The DNA profile frequency was 1 in 1,000,000. The expert said that the chances of a sample from another Afro-Caribbean (the relevant population) matching the crime sample were one in a million.

Aitken and Taroni, Science and Justice 1998; 38:165–177

7. U.S. v. Bonds [12 F. 3d 540 (6th Cir. 1993)] The DNA profile frequency was 1 in 270,000. The FBI calculated a probability of 1 in 270,000 that an unrelated individual selected randomly from the Caucasian population (the relevant population) would have a DNA profile matching that of Bonds.

8. U.S. v. Martinez [13 F.3d 1191 (8th Cir. 1993)] The FBI expert knew that the population frequency of the matching DNA profile is 1 in 2,600. The expert testified that only 1 in 2,600 American Indians (relevant population) would be expected to produce the identical genetic characteristics as Martinez.

Aitken and Taroni, Science and Justice 1998; 38:165–177

9. Arizona v. Johnson [905 P. 2d 1002; 192 Ariz. Rep. 19 91995)] The expert knew that the matching DNA profile has a frequency of 1 in 312 million. The expert testified that the victim's shirt was examined and found to contain human blood and semen. Testing performed showed that DNA extracted from these stains matched Johnson's blood at five different chromosome locations or loci. The expert testified that the possibility of a random match - two unrelated individuals having the same DNA pattern across five loci - was 1 in 312 million.

Aitken and Taroni, Science and Justice 1998; 38:165–177

10. Ross v. State [B14-90-00659 (Tex. App. Feb. 13, 1992)] The DNA profile frequency was 1 in 209,100,000. The expert said that he has a database of blood samples from all over the country and he asks the question "How many people would we have to look at before we saw another person like this?" The answer is 209,100,000.

11. Harrison v. Indiana [Supreme Court of Indiana (Jan. 4, 1995)] The DNA profile frequency was 7.4 in 100. The expert said that although 92.6% of all white males could be excluded as the source of the specimen, the defendant had not been excluded. She acknowledged that for 13,000 white men (the size of the city where the crime was committed) the specimen could have come from any 962 [7.4% of 13,000] of them. Hence, the suspect is one of 962 men who might have committed the crime.

Aitken and Taroni, Science and Justice 1998; 38:165–177

12. R v. Montella [1 NZLR High Court (1992) 63-68]. The DNA profile frequency was 8.06 $\times 10^{-5}$. The expert said "A DNA profiling examination of the samples strongly supports a contention that the semen stain on the underpants of the complainants came from the accused. It is said that the likelihood of obtaining such DNA profiling results is at least 12,400 times greater if the semen stain originated from the accused than from another individual."

Nevada v Troy Brown

In the early morning hours of January 29, 1994, Jane Doe was sexually assaulted in the bedroom of her trailer home at 1637 Pruett Street in Carlin. Jane Doe and her four-year-old sister were home alone while their mother, Pam, was drinking at a bar, and their step-father, Wayne, was working the night shift at his job. Troy was arrested, tried, and convicted for the crime.

Nevada v Troy Brown

At trial, Renee Romero testified that she had conducted a DNA test on stains found on Jane Doe's underwear. Romero explained in detail what DNA is and how it is tested. Romero testified that the DNA sample tested from Jane Doe's underwear matched Troy's and that only 1 in 3,000,000 people had the same DNA code as the one tested. Troy's counsel cross-examined Romero regarding how she conducted the tests, the amount of DNA required to run the tests, and the databases against which the DNA tests were compared to determine the statistical probability that others would have the same DNA code. However, Troy's counsel did not call his own expert DNA witness even though the court provided funds for such a witness.

Nevada Supreme Court, February 26, 1997

Appellant Troy Brown was tried and convicted of sexually assaulting Jane Doe, a nine-year-old girl. Troy was convicted of two counts of sexual assault of a child under fourteen years of age, and one count of child abuse by sexual abuse. He was acquitted of one count of attempted murder. Troy claims on appeal that (1) he was improperly denied bail; (2) the DNA evidence was improperly admitted because no evidentiary hearing was held; (3) sufficient evidence did not exist to support his conviction; (4) double jeopardy barred his convictions for both sexual assault and child abuse by sexual abuse; and (5) the district judge abused his discretion during the sentencing phase of the trial.

Nevada Supreme Court, February 26, 1997

We conclude that the district judge properly denied bail for Troy, that the DNA evidence was properly admitted at trial, and that sufficient evidence existed to support Troy's conviction. However, we conclude that Troy's conviction for both sexual assault and child abuse by sexual abuse violated the double jeopardy provision of the Constitution and that the conviction for child abuse must be vacated. Finally, we conclude that the district judge abused his discretion during the sentencing phase of the trial and the case must be remanded to the district court for a new sentencing hearing on the remaining sexual assault conviction.

US District Court, February 6, 2004

On February 6, 2004, Troy filed his federal petition for writ of habeas corpus pursuant to 28 U.S.C. 2254, arguing, inter alia, violations of due process and ineffective assistance of counsel. Judge Pro permitted Troy to expand the record, admitting, among other things, an uncontested report discrediting Romero's testimony by Dr. Laurence Mueller (the "Mueller Report"), a professor of Ecology and Evolutionary Biology at the University of California, Irvine.

US District Court, February 6, 2004

The district court granted Troy's petition. First, the district court concluded that, in light of the Mueller Report, Romero's testimony was unreliable. Absent that testimony, no rational trier of fact could conclude beyond a reasonable doubt that Troy was guilty of each and every element of the offenses with which he was charged. The district court also concluded that Troy's attorney's failure to diligently defend against Respondents' DNA testimony, as well as his failure to investigate the alibi of Henle, a potential suspect, amounted to ineffective assistance of counsel. Respondents [the State of Nevada] timely appealed.

US Court of Appeals, May 8, 2008

At trial, Respondents presented the testimony of DNA expert Renee Romero of the Washoe County Sheriff's Office Crime Lab. Romero testified that, among other things, there was a 99.99967 percent chance that Troy was the assailant.

At Petitioner Troy Brown's trial for sexual assault, the Warden and State's ("Respondents") deoxyribonucleic acid ("DNA") expert provided critical testimony that was later proved to be inaccurate and misleading. Respondents have conceded at least twice that, absent this faulty DNA testimony, there was not sufficient evidence to sustain Troy's conviction. In light of these extraordinary circumstances, we agree with District Judge Philip M. Pro's conclusions that Troy was denied due process, and we affirm the district court's grant of Troy's petition for writ of habeas corpus.

US Court of Appeals, May 8, 2008

Troy asserts that there was insufficient evidence to convict him. His argument rests on the admission of Romero's later discredited testimony regarding the DNA evidence, which was introduced without rebuttal at trial. Respondents have conceded that absent introduction of Romero's DNA evidence, the remaining evidence is insufficient to sustain Troy's conviction. Having reviewed the record ourselves, we affirm the district court's conclusion that, had Romero's inaccurate and unreliable testimony on the DNA evidence been excluded, there would have been insufficient evidence to convict Troy on each essential element of the offenses beyond a reasonable doubt. We further agree with the district court's conclusion that the Nevada Supreme Court's decision was both "contrary to" and an "unreasonable application of" established United States Supreme Court precedent.

US Court of Appeals, May 8, 2008

Here, Romero initially testified that Troy's DNA matched the DNA found in Jane's underwear, and that 1 in 3,000,000 people randomly selected from the population would also match the DNA found in Jane's underwear (random match probability). After the prosecutor pressed her to put this another way, Romero testified that there was a 99.99967 percent chance that the DNA found in Jane's underwear was from Troy's blood (source probability). This testimony was misleading, as it improperly conflated random match probability with source probability. In fact, the former testimony (1 in 3,000,000) is the probability of a match between an innocent person selected randomly from the population; this is not the same as the probability that Troy's DNA was the same as the DNA found in Jane's underwear, which would prove his guilt. Statistically, the probability of guilt given a DNA match is based on a complicated formula known as Bayes's Theorem, see id. at 170-71 n. 2, and the 1 in 3,000,000 probability described by Romero is but one of the factors in this formula.

US Court of Appeals, May 8, 2008

Because we affirm the district court's grant of Troy Brown's habeas petition on due process grounds, we need not reach his arguments regarding ineffective assistance of counsel. The district court's grant of Troy's petition for writ of habeas corpus and reversal of his conviction is AFFIRMED. Respondents shall retry Troy within 180 days or shall release him from custody.

US Supreme Court, January 11, 2010

The prosecutor's fallacy is the assumption that the random match probability is the same as the probability that the defendant was not the source of the DNA sample. In other words, if a juror is told the probability a member of the general population would share the same DNA is 1 in 10,000 (random match probability), and he takes that to mean there is only a 1 in 10,000 chance that someone other than the defendant is the source of the DNA found at the crime scene (source probability), then he has succumbed to the prosecutors fallacy. It is further error to equate source probability with probability of guilt, unless there is no explanation other than guilt for a person to be the source of crime-scene DNA. This faulty reasoning may result in an erroneous statement that, based on a random match probability of 1 in 10,000, there is a .01% chance the defendant is innocent or a 99.99% chance the defendant is guilty.

US Supreme Court, January 11, 2010

In sum, the two inaccuracies upon which this case turns are testimony equating random match probability with source probability, and an underestimate of the likelihood that one of Troys brothers would also match the DNA left at the scene.

We have stated before that "DNA testing can provide powerful new evidence unlike anything known before."

The State acknowledges that Romero committed the prosecutor's fallacy. Regardless, ample DNA and non-DNA evidence in the record adduced at trial supported the jury's guilty verdict. Accordingly, the judgment of the Court of Appeals is reversed, and the case is remanded for further proceedings consistent with this opinion.

Hierarchy of Propositions

(Evett et al., 2002. Journal of Forensic Sciences 47:520–523.)

Third level: Offense level propositions:

The defendant raped the victim.

Some unknown person raped the victim.

Second level: Activity level propositions:

The defendant smashed the window.

The defendant has never been at the scene.

First level: Source level propositions:

The glass on the defendant's clothing came from the broken window.

The glass on the defendant's clothing is from some other source.

Meaning of Frequencies

What is meant by "the frequency of the matching profile is 1 in 57 billion"?

It is an estimated probability, obtained by multiplying together the allele frequencies, and refers to an infinite random mating population. It has nothing to do with the size of the world's population.

Meaning of Frequencies

With 13 STR loci having (at least) 10 alleles each, there are $55^{13} = 4.2 \times 10^{22}$ possible genotypes, even though there are only 6 billion people. The total world population is itself a sample from all possible genotypes. Almost all the possible genotypes are not in the present population, and have expected frequencies that are very small: e.g. if all 26 alleles were independent, and had frequency of 0.1, we could quote an estimated frequency of 8.2×10^{-23} for a completely heterozygous. We don't *expect* that anyone living will have that profile – but of course we know that someone does.

Meaning of Frequencies

The question is really whether we would see the profile in two people, given that we have already seen it in one person. This conditional probability may be very low, but has nothing to do with the size of the population.

ALLELIC INDEPENDENCE

Testing for Allelic Independence

What is the probability a person has a particular DNA profile? What is the probability a person has a particular profile if it has already been seen once?

The first question is a little easier to think about, but difficult to answer in practice: it is very unlikely that a profile will be seen in any sample of profiles. Even for one STR locus with 10 alleles, there are 55 different genotypes and most of those will not occur in a sample of a few hundred profiles.

For locus D3S1358 in the African American population, the FBI frequency database shows that 31 of the 55 genotype counts are zero. Estimating the population frequencies for these 31 types as zero doesn't seem sensible.

D3S1358 Genotype Counts

Observed	<12	12	13	14	15	16	17	18	19	>19
<12	0									
12	0	0								
13	0	0	0							
14	0	0	0	2						
15	0	0	1	19	15					
16	1	1	1	15	39	19				
17	0	0	2	10	26	24	9			
18	1	0	1	2	6	10	3	0		
19	0	0	0	1	0	0	1	0	0	
>19	0	0	0	0	1	0	0	0	0	0

Hardy-Weinberg Law

A solution to the problem is to assume that the Hardy-Weinberg Law holds. For a random mating population, expect that genotype frequencies are products of allele frequencies.

For a locus with two alleles, A, a:

$$\begin{aligned} P_{AA} &= (p_A)^2 \\ P_{Aa} &= 2p_A p_a \\ P_{aa} &= (p_a)^2 \end{aligned}$$

For a locus with several alleles A_i:

$$\begin{aligned} P_{A_iA_i} &= (p_{A_i})^2 \\ P_{A_iA_j} &= 2p_{A_i}p_{A_j} \end{aligned}$$

D3S1358 Hardy-Weinberg Calculations

The allele counts for D3S1358 in the African-American sample are:

										Total
Allele	<12	12	13	14	15	16	17	18	19	>19	
Count	2	1	5	51	122	129	84	23	2	1	420

If the Hardy-Weinberg Law holds, then we would expect to see $n\tilde{p}_{13}^2 = 210 \times (5/420)^2 = 0.03$ individuals of type 13,13 in a sample of 210 individuals.

Also, we would expect to see $2n\tilde{p}_{13}\tilde{p}_{14} = 420 \times (5/420) \times (51/420) = 0.61$ individuals of type 13,14 in a sample of 210 individuals.

Other values are shown on the next slide.

D3S1358 Observed and Expected Counts

		<12	12	13	14	15	16	17	18	19	>19
<12	Obs.	0									
	Exp.	0.0									
12	Obs.	0	0								
	Exp.	0.0	0.0								
13	Obs.	0	0	0							
	Exp.	0.0	0.0	0.0							
14	Obs.	0	0	0	2						
	Exp.	0.2	0.1	0.6	3.1						
15	Obs.	0	0	1	19	15					
	Exp.	0.6	0.3	1.5	14.8	17.7					
16	Obs.	1	1	1	15	39	19				
	Exp.	0.6	0.3	1.5	15.7	37.5	19.8				
17	Obs.	0	0	2	10	26	24	9			
	Exp.	0.4	0.2	1.0	10.2	24.4	25.8	8.4			
18	Obs.	1	0	1	2	6	10	3	0		
	Exp.	0.1	0.1	0.3	2.8	6.7	7.1	4.6	0.6		
19	Obs.	0	0	0	1	0	0	1	0	0	
	Exp.	0.0	0.0	0.0	0.2	0.6	0.6	0.4	0.1	0.0	
>19	Obs.	0	0	0	0	1	0	0	0	0	0
	Exp.	0.0	0.0	0.0	0.1	0.3	0.3	0.2	0.1	0.0	0.0

Testing for Hardy-Weinberg Equilibrium

A test of the Hardy-Weinberg Law will somehow decide if the observed and expected numbers are sufficiently similar that we can proceed as though the law can be used.

In one of the first applications of Hardy-Weinberg testing in a US forensic setting:

> "To justify applying the classical formulas of population genetics in the Castro case the Hispanic population must be in Hardy-Weinberg equilibrium. Applying this test to the Hispanic sample, one finds spectacular deviations from Hardy-Weinberg equilibrium."

E.S. Lander. 1989. DNA fingerprinting on trial. Nature 339: 501-505.

VNTR "Coalescence"

Forensic DNA profiling initially used minisatellites, or VNTR loci, with large numbers of alleles. Heterozygotes would be scored as homozygotes if the two alleles were so similar in length that they coalesced into one band on an autoradiogram. Small alleles often not detected at all, and this is the cause of Lander's finding.

Considerable debate in early 1990s on alternative "binning" strategies for reducing the number of alleles (Science 253:1037-1041, 1991).

Typing has moved to microsatellites with fewer and more easily distinguished alleles, but testing for Hardy-Weinberg equilibrium continues. There are still reasons why the law may not hold.

Population Structure can Cause Departure from HWE

If a population consists of a number of subpopulations, each in HWE but with different allele frequencies, there will be a departure from HWE at the population level. This is the Wahlund effect.

Suppose there are two equal-sized subpopulations, each in HWE but with different allele frequencies, then

	Subpopn 1	Subpopn 2	Total Popn
p_A	0.6	0.4	0.5
p_a	0.4	0.6	0.5
P_{AA}	0.36	0.16	$0.26 > (0.5)^2$
P_{Aa}	0.48	0.48	$0.48 < 2(0.5)(0.5)$
P_{aa}	0.16	0.36	$0.26 > (0.5)^2$

Population Structure

Effect of population structure taken into account with the "theta-correction." Matching probabilities allow for a variance in allele frequencies among subpopulations.

$$\Pr(AA|AA) = \frac{[3\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)}$$

where p_A is the average allele frequency over all subpopulations. We will come back to this expression.

Population Admixture

A population might represent the recent admixture of two parental populations. With the same two populations as before but now with 1/4 of marriages within population 1, 1/2 of marrieages between populations 1 and 2, and 1/4 of marriages within population 2. If children with one or two parents in population 1 are considered as belonging to population 1, there is an excess of heterozygosity in the offspring population.

If the proportions of marriages within populations 1 and 2 are both 25% and the proportion between populations 1 and 2 is 50%, the next generation has

	Population 1	Population 2
P_{AA}	$0.09 + 0.12 = 0.21$	0.04
P_{Aa}	$0.12 + 0.26 = 0.38$	0.12
P_{aa}	$0.04 + 0.12 = 0.16$	0.09
	0.75	0.25

Exact HWE Test

The preferred test for HWE is an "exact" one. The test rests on the assumption that individuals are sampled randomly from a population so that genotype counts have a multinomial distribution:

$$\Pr(n_{AA}, n_{Aa}, n_{aa}) = \frac{n!}{n_{AA}! n_{Aa}! n_{aa}!} (P_{AA})^{n_{AA}} (P_{Aa})^{n_{Aa}} (P_{aa})^{n_{aa}}$$

This equation is always true, and when there is HWE ($P_{AA} = p_A^2$ etc.) there is the additional result that the allele counts have a binomial distribution:

$$\Pr(n_A, n_a) = \frac{(2n)!}{n_A! n_a!} (p_A)^{n_A} (p_a)^{n_a}$$

Exact HWE Test

Putting these together gives the conditional probability of the genotypic data given the allelic data and given HWE:

$$\Pr(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a, \text{HWE}) = \frac{\frac{n!}{n_{AA}! n_{Aa}! n_{aa}!}(p_A^2)^{n_{AA}}(2p_A p_a)^{n_{Aa}}(p_a^2)^{n_{aa}}}{\frac{(2n)!}{n_A! n_a!}(p_A)^{n_A}(p_a)^{n_a}}$$

$$= \frac{n!}{n_{AA}! n_{Aa}! n_{aa}!}\frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}$$

Reject the Hardy-Weinberg hypothesis if this probability is unusually small.

Exact HWE Test Example

Reject the HWE hypothesis if the probability of the genotypic array, conditional on the allelic array, is among the smallest probabilities for all the possible sets of genotypic counts for those allele counts.

As an example, consider $(n_{AA} = 1, n_{Aa} = 0, n_{aa} = 49)$. The allele counts are $(n_A = 2, n_a = 98)$ and there are only two possible genotype arrays:

AA	Aa	aa	$\Pr(n_{AA}, n_{Aa}, n_{aa} \mid n_A, n_a, \text{HWE})$
1	0	49	$\frac{50!}{1!0!49!}\frac{2^0 2!98!}{100!} = \frac{1}{99}$
0	2	48	$\frac{50!}{0!2!48!}\frac{2^2 2!98!}{100!} = \frac{98}{99}$

Exact HWE Test

The probability of the data on the previous slide, conditional on the allele frequencies and on HWE, is $1/99 = 0.01$. This is less than the conventional 5% significance level.

In general, the p-value is the (conditional) probability of the data plus the probabilities of all the less-probable datasets. The probabilities are all calculated assuming HWE is true.

Exact HWE Test

Still in the two-allele case, for a sample of size $n = 100$ with minor allele frequency of 0.07, there are only 8 sets of genotype counts:

n_{AA}	n_{Aa}	n_{aa}	Exact Prob.	p-value
93	0	7	0.0000	0.0000*
92	2	6	0.0000	0.0000*
91	4	5	0.0000	0.0000*
90	6	4	0.0002	0.0002*
89	8	3	**0.0051**	**0.0053***
88	10	2	0.0602	0.0654
87	12	1	0.3209	0.3863
86	14	0	0.6136	1.0000

So, for a nominal 5% significance level, the actual significance level is 0.0053 for an exact test that rejects when $n_{Aa} \leq 8$.

Permutation Test

For large sample sizes and many alleles per locus, there are too many genotypic arrays for a complete enumeration and a determination of which are the least probable 5% arrays.

A large number of the possible arrays is generated by permuting the alleles among genotypes, and calculating the proportion of these permuted genotypic arrays that have a smaller conditional probability than the original data. If this proportion is small, the Hardy-Weinberg hypothesis is rejected.

Permutation Test

Mark a set of five index cards to represent five genotypes:

Card 1:	A	A
Card 2:	A	A
Card 3:	A	A
Card 4:	a	a
Card 5:	a	a

Tear the cards in half to give a deck of 10 cards, each with one allele. Shuffle the deck and deal into 5 pairs, to give five genotypes.

Permutation Test

The permuted set of genotypes fall into one of four types:

AA	Aa	aa	Number of times
3	0	2	
2	2	1	
1	4	0	

Permutation Test

Check the following theoretical values for the proportions of each of the three types, from the expression:

$$\frac{n!}{n_{AA}!n_{Aa}!n_{aa}!} \times \frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}$$

AA	Aa	aa	Conditional Probability
3	0	2	$\frac{1}{21} = 0.048$
2	2	1	$\frac{12}{21} = 0.571$
1	4	0	$\frac{8}{21} = 0.381$

These should match the proportions found by repeating shufflings of the deck of 10 allele cards.

Permutation Test for D3S1358

For a STR locus, where $\{n_g\}$ are the genotype counts and $n = \sum_g n_g$ is the sample size, and $\{n_a\}$ are the alleles counts with $2n = \sum_a n_a$, the exact test statistic is

$$\Pr(\{n_g\}|\{n_a\}, \text{HWE}) = \frac{n! 2^H \prod_a n_a!}{\prod_g n_g! (2n)!}$$

where H is the count of heterozygotes.

This probability for the African American genotypic counts at D3S1358 is 0.6163×10^{-13}, which is a very small number. But it is not unusually small if HWE holds: a proportion 0.81 of 1000 permutations have an even small probability.

Linkage Disequilibrium

This term is generally reserved for association between pairs of alleles – one at each of two loci. In the present context, it may simply mean some lack of independence of profile or match probabilities at different loci.

Unlinked loci are expected to be almost independent.

Allelic Matching

Within-population Matching

The key forensic genetic issue is that of matching profiles. What is the probability that two people have the same STR profile?

We can get some empirical estimate of this when we have a set of profiles. For the African -American sample of 210 profiles for D3S1358, how many pairs of profiles match? Only those genotypes that occur more than once in the sample provide matches. To simplify this initial discussion, consider the following data for the Y-STR locus DYS390 from the NIST database:

Allele Counts in NIST Data for DYS390

	Population				
Allele	Afr.Am.	Cauc.	Hisp.	Asian	Total
20	4	1	1	0	6
21	176	4	17	1	198
22	43	45	14	17	119
23	36	116	50	17	219
24	56	145	129	21	351
25	23	46	21	36	126
26	3	2	2	4	11
27	0	0	2	0	2
Total	341	359	236	96	1032

Within- and Between-population Matching for DYS390

Within the African-American sample there are $341 \times 340 = 115,940$ pairs of profiles and the number of matches is

$$4 \times 3 + 176 \times 175 + 43 \times 42 + 36 \times 35 + 56 \times 55 + 23 \times 22 + 3 \times 2 = 37,470$$

so the within-population matching proportion is $37,470/115,940 = 0.323$.

Between the African-American and Caucasian samples, there are $341 \times 359 = 122,419$ pairs of profiles and the number of matches is

$$4 \times 1 + 176 \times 4 + 43 \times 45 + 36 \times 116 + 56 \times 145 + 23 \times 4 + 3 \times 2 = 12,403$$

so the between-population matching proportion is $12,403/122,419 = 0.101$.

Allele Counts in NIST Data for DYS391

	Population				
Allele	Afr.Am.	Cauc.	Hisp.	Asian	Total
7	0	0	1	0	1
8	0	1	0	1	2
9	2	12	16	3	33
10	238	162	128	79	607
11	93	175	89	13	370
12	7	9	2	0	18
13	1	0	0	0	1
Total	341	359	236	96	1032

The within-population matching proportion for the African-American sample is 65,006/115,940=0.561.

The between-population matching proportion for the African-American and Caucasian samples is 54,918/122,419=0.449.

Two-locus counts in NIST African-American Data for DYS390, DYS391

DYS390	DYS391	Count n_g	$n_g(n_g-1)$
22	10	34	1122
22	11	9	72
24	10	15	210
24	11	39	1482
24	12	1	0
24	9	1	0
23	10	19	342
23	11	14	182
23	12	3	6
21	10	157	24492
21	11	15	210
21	12	2	2
21	9	1	0
21	13	1	0
25	10	11	110
25	11	12	132
26	10	1	0
26	11	2	2
20	10	1	0
20	11	2	2
20	12	1	0

Two-locus counts in NIST Caucasian Data for DYS390, DYS391

DYS390	DYS391	Count n_g	$n_g(n_g-1)$
22	10	43	1806
22	11	1	0
22	9	1	0
24	10	48	2256
24	11	88	7656
24	12	4	12
24	9	5	20
23	10	50	2450
23	11	60	3540
23	12	2	2
23	9	3	6
23	8	1	0
21	10	3	6
21	11	1	0
25	10	18	306
25	11	22	462
25	12	3	6
25	9	3	6
26	11	2	2
20	11	1	0

Two-locus Matches

The within-population matching proportion for the African-American sample is 28,366/115,940=0.245.

The within-population matching proportion for the Caucasian sample is 18,536/128,522=0.144.

The between-population matching proportion for the African-American and Caucasian samples is 8,347/122,419=0.068.

There is a clear decrease in matching between populations from within populations. We can establish some theory that describes these proportions.