Forensic Genetics

Summer Institute in Statistical Genetics
July 26-28, 2017
University of Washington

John Buckleton: John.Buckleton@esr.cri.nz

Bruce Weir: bsweir@uw.edu

Contents

Sources of Genetic 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.

An introduction 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.

MPS also called NGS (Next Generation Sequencing.)

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, from 18,000 whole-genome sequences:

Total number of SNPs	86,974,704
Singletons	35,883,567
% Singletons	41.3%
Number in dbSNP	43,141,144
% 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 conditionalon some information I, so should write Pr($A|I$) for "the probability that A is true given that I 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

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

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

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

Second Law of Probability

If A, B are mutually exclusive given I

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

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

($\bar{A}$ 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

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

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.

Will generally omit the I from now on.

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:

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

Odds Form of Bayes' Theorem

From the third law of probability

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

Taking the ratio of these two equations:

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

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

AIDS Example

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

Suppose a test procedure has two outcomes: A (positive) and $\bar{A}$ (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 $\Pr(B|A)$ when $\Pr(A|B)$ is known. This requires Bayes' theorem, and the term $\Pr(A)$ follows from the Law of Total Probability.

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

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 $p^0 = 1$.

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

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

Apply laws of probability to change this into

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

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.

When we talk about DNA types, by "selecting a person at random" we mean choosing him in such a way as to be as uncertain as possible about their 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 ethnic group? (What is an ethnic group?)

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.