Section 1: Overview

HISTORY OF IDENTIFICATION

Legal v. Scientific Thinking

“The very goals of science and law differ.
Science searches for the truth and seeks to increase knowledge by formulating and testing theories. Law seeks justice by resolving individual conflicts, although this search often coincides with one for truth.”

“Rules of decision that are not tailored to individual cases, such as those that turn on statistical reasoning, are often viewed as suspect.”

Feinberg SE (Editor). 1989. *The Evolving Role of Statistical Assessments as Evidence in the Courts.* Springer.

Forensic Science Approach

“The central problem of the criminal investigator is the establishment of personal identity – usually of the criminal, sometimes of the victim.”

Need to distinguish between identity and individualization. **Identity** refers to unique existence – no two different things can be identical. The DNA profiles from a suspect and a crime scene are different things.

Individualization points to a specific person. A fingerprint from a crime scene is not identical to a suspect’s recorded fingerprint, but can be used to identify him and prove his individuality.

Kirk PL. 1974. *Crime Investigation, (Second Edition).* Krieger,

Uniqueness

“no two objects can ever be identical. They can and often do have properties that are not distinguishable. If enough of these properties exist ... *identity of source* is established.”

“The criminalist of the future may well be able to individualize the criminal directly through the hair he has dropped, the blood he has shed, or the semen he has deposited. All these things are unique to the individual, just as his fingerprints are unique to him.”

Kirk PL. 1974. *Crime Investigation, (Second Edition).* Krieger,

Forensic science question

Not: “Is this profile unique?” (it is).

Not: “Are these two profiles identical?” (they can’t be).

But: “ Is there sufficient evidence to demonstrate that these two profiles originate from the same source?”

Bertillonage

Alphonse Bertillon (1853-1914), French anthropometrist. Son and brother of statisticians. Used 11 measurements:

1. Standing height
2. Arm reach
3. Sitting height

4.* Head length
5.* Head breadth

6. Length of right ear
7. Cheek width

8.* Length of left foot
9.* Length of left middle finger

10. Length of left little finger
11. Length of the left forearm and hand to the tip of extended middle finger

Bertillonage

Searching was done on four categories 4, 5, 8, 9. Each measurement divided into three subdivisions (large, medium, small) i.e. $3^4 = 81$ categories per person. Filing cabinets with 81 drawers used.

Using all 11 characters, plus 7 eye colors, the number of possible profiles is $3^{11} \times 7 = 1,240,029$.

Wikipedia entry for Alphonse Bertillon

“Being an orderly man, he was dissatisfied with the ad hoc methods used to identify the increasing number of captured criminals who had been arrested before. This, together with the steadily rising recidivism rate in France since 1870, motivated his invention of anthropometrics. His road to fame was a protracted and hard one, as he was forced to do his measurements in his spare time. He used the famous La Sant Prison in Paris for his activities, facing jeers from the prison inmates as well as police officers.

He is also the inventor of the mug shot. Photographing of criminals began in the 1840s only a few years after the invention of photography, but it was not until 1888 that Bertillon standardized the process.”

https://en.wikipedia.org/wiki/Alphonse_Bertillon

Coincidental match

Two different men at Leavenworth in 1903 had very similar Bertillon dimensions (lengths in mm):

	Will West	William West
1	19.7	19.8
2	15.8	15.9
3	12.3	12.2
4	28.2	27.5
5	50.2	50.3
6	178.5	177.5
7	9.7	9.6
8	91.3	91.3
9	187.0	188.0
10	6.6	6.6
11	14.8	14.8

http://www.globalsecurity.org/security/systems/biometrics-history.htm

Fingerprints

"The arrangement of skin ridges is never duplicated in two persons."
J.C.A. Mayer, 1783.

J.E. Purkinje established categories of fingerprints in early 19th century.

W. Herschel, a British administrator, used fingerprints in India in 1850's.

H. Faulds, a British physician, used fingerprints in Japan.

Francis Galton wrote the book "Fingerprints" in 1892, and gave some probabilities for coincidental matches.

Fingerprints

Galton considered that the chance that a random fingerprint would match a specified print was 2^{-36}. For a population of size 1.6×10^9, the odds were 1 to 39 that the print of any single finger would be exactly like the same finger of any other person.

[This is based on the probability of not finding the print in a sample of size 1.6 billion.]

Heritability of fingerprints

Galton looked at 105 sib-pairs:

Second	First sib		
sib	Arches	Loops	Whorls
Arches	**5**	12	2
Loops	4	**42**	15
Whorls	1	14	**10**

Galton noticed that the diagonal counts of 5, 42, 10 are larger than those (2, 40, 6) expected if the sibs had independent fingerprints, but not as great as they could be (10, 68, 27). He did not have the chi-square test available in 1892, but did conclude that there was an association.

He did not find racial differences.

Uniqueness of fingerprints

Probability arguments not used. By 1924, textbooks would say "No two fingerprints are identical in pattern." In 1939 J.Edgar Hoover wrote that fingerprints were "a certain and quick means of identification."

Acceptance of uniqueness probably followed from "(i) striking visual appearance of fingerprints in court, (ii) a few dramatically successful cases, and (iii) a long period in which they were used without a single case being noted where two different individuals exhibited the same pattern."

Stigler SM. 1995. Galton and identification by fingerprints. Genetics 140:857-860.

Stigler anticipated the same growing acceptance of DNA profiles being unique.

Misuse of Fingerprints

Oregon attorney Brandon Mayfield was wrongly identified by the FBI as the source of a fingerprint on an item of evidence in the 2004 Madrid train bombings.

https://en.wikipedia.org/wiki/Brandon_Mayfield

A subsequent report by the FBI admitted the error

https://www.fbi.gov/about-us/lab/forensic-science-communications
/fsc/jan2005/special_report/2005_special_report.htm

Accuracy of Fingerprints

A subsequent study by Ulerya et al "Accuracy and reliability of forensic latent fingerprint decisions" was published

"169 latent print examiners each compared approximately 100 pairs of latent and exemplar fingerprints from a pool of 744 pairs. ...Five examiners made false positive errors for an overall false positive rate of 0.1%. Eighty-five percent of examiners made at least one false negative error for an overall false negative rate of 7.5%."

Ulerya BT, Hicklina RA, Buscagliab J, Roberts A. 2011. Proc Natl Acad Sci USA 108: 77337738.

Statistical approach

Partial transfer evidence: physical material or impressions transferred from crime scene to perpetrator (or perpetrator's possessions), or vice versa.

PTE is characterized and assigned to an identity-set. Does a particular person (or their type) belong to the set? Does anyone else belong to the set?

"If it is *highly improbable* that another member could be found, we would be *reasonably sure* that the correct origin has been located. But if it is *quite probable* that other members exist, we would *not be so sure* that we have the correct origin."

Kingston CR. 1965. J Am Stat Assoc 60:70-80, 1028-1034.

Blood Typing

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

ABO System

The possible offspring blood groups for each pair of parents:

	Mother			
Father	A	B	AB	O
A	A,O	A,B,AB,O	A,B,AB	A,O
B	A,B,AB,O	B,O	A,B,AB	B,O
AB	A,B,AB	A,B,AB	A,B,AB	A,B
O	A,O	B,O	A,B	O

ABO System

Blood group	Antigens in red blood cells	Antibodies in serum
O	None	Anti-A and Anti-B
A	A	Anti-B
B	B	Anti-A
AB	A and B	None

http://www.redcrossblood.org/learn-about-blood/blood-types

ABO System

For blood transfusions, recipient should not produce antibodies to the donor's antigens:

	Donor			
Recipient	O	A	B	AB
O	OK			
A	OK	OK		
B	OK		OK	
AB	OK	OK	OK	OK

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

California Court of Appeals, 1946

“Concerning the immutability of the scientific law of blood-grouping, which we have no reason to question ...”

“Whatever claims the medical profession may make for blood tests to determine paternity, no evidence is by law made conclusive or unanswerable unless so declared by the Code of Civil Procedure of the State of California ”

74 Cal.App.2d 652 (1946)

Outcome of Chaplin Trial

“The brouhaha surrounding Chaplin’s case and similar paternity suits (like 1937’s Arais v. Kalensnikoff and 1951’s Hill v. Johnson) led to the reformation of paternity laws in the state of California, with other states eventually following suit. In 1953, along with Oregon and New Hampshire, California drafted the Uniform Act on Blood Tests to Determine Paternity, which in legalese states that: ‘If the court finds that the conclusions of all the experts as disclosed by the evidence based upon the tests are that the alleged father is not the father of the child, the question of paternity shall be resolved accordingly.’ ”

http://mentalfloss.com/article/63158/how-charlie-chaplin-changed-paternity-laws-america

Spencer v Commonwealth of Virginia

From the Supreme Court of Virginia, September 22, 1989.

“Timothy Wilson Spencer was indicted for the capital murder of Susan Tucker, i.e., the willful, deliberate, and premeditated murder during the commission of, or subsequent to, rape. Spencer also was indicted for the rape of Tucker. ... a jury convicted Spencer of capital murder and fixed his punishment at death. The jury also convicted Spencer of rape and fixed his punishment at life imprisonment. Following a sentencing hearing, the trial court imposed the sentences fixed by the jury and entered judgments on the jury verdicts.

... We have considered all of Spencer’s assignments of error and find no reversible error. We also have made the review of the death sentence mandated by Code 17-110.1 and conclude that the sentence should be affirmed. Accordingly, the judgments of the trial court will be affirmed.”

Spencer v Commonwealth of Virginia (contd.)

"The parties stipulated that Spencer does not have an identical twin and that none of his blood relatives had committed the murder. Therefore, the chances that anyone other than Spencer produced the semen stains was one in 135 million. There are approximately 10 million adult black males in the United States."

Spencer was the first person executed after a conviction based on DNA evidence.

SPENCER v. COM 384 S.E.2d 775 (Va. 1989)
aw.justia.com/cases/virginia/supreme-court/1989/890579-1.html

Extreme Numbers: Robinson v. Mandell, 1868.

Two signatures matched at 30 downstrokes. The probability of a coincidental match was estimated to be 1 in 5. The probability of 30 coincidences in one pair of signatures was “once in 2,666 millions of millions of millions.” (Mathematics professor Benjamin Pierce.)

“This number far transcends human experience. So vast an improbability is practically an impossibility. Such evanescent shadows of probability cannot belong to actual life. They are unimaginably less than those least things which the law cares not for.”

Refers to chance of a coincidental match between two handwriting samples.

https://en.wikipedia.org/wiki/Howland_will_forgery_trial

No Extreme Numbers in Minnesota

"Schwartz contends that any probative value of statistical frequency evidence is outweighed by its prejudicial effect, as illustrated by the media exposure forensic DNA typing has received implying its infallibility. In dealing with complex technology, like DNA testing, we remain convinced that juries in criminal cases may give undue weight and deference to presented statistical evidence and are reluctant to take that risk."

447 N.W.2d 422 (1989)

Refers to matching DNA profile with a frequency reported as 1 in 33 billion.

Extreme numbers: Fingerprints

Chance of a match for a single finger print estimated to be less than 1 in 64 thousand million.

“When two fingers of each of two persons are compared, and found to have the same minutiae, the improbability [of 1 in 2^{36}] becomes squared, and reaches a figure altogether beyond the range of the imagination.”

Galton F. 1892. Fingerpints.

DNA Profiling

Human Genome has about 3,000,000,000 elements (base pairs).

Any two people differ at about 3,000,000 of these.

Forensic profiles use 20 STR markers. Each of these markers as at least 10 variant forms, or at least 55 different combinations. Therefore there are about $55^{20} = 6.4 \times 10^{34}$ different profiles possible.

Only 1 in 10^{24} of the possible profiles exists in the whole world.

Beyond Reasonable Doubt?

After forensic evidence is presented, a jury or judge may have to make a decision, based on the concept of “beyond reasonable doubt.” What does that mean? A survey found:

Probability	Judges	Jurors	Students
0–50%	0	5	3
50%	1	6	2
55%	2	2	1
60%	8	4	1
65%	2	1	0
70%	14	2	1
75%	23	2	1
80%	58	8	9
85%	21	2	3
90%	68	9	20
95%	44	3	17
100%	106	25	30
Total	347	69	88

Source unknown.

People v. Collins

Another attempt to introduce probabilities into court:

Characteristic	Frequency
Girl with blond hair	1 in 3
Girl with ponytail	1 in 10
Man with mustache	1 in 4
Black man with beard	1 in 10
Yellow car	1 in 10
Interracial couple in car	1 in 1000
All six characteristics	1 in 12 million

https://en.wikipedia.org/wiki/People_v._Collins

Alec Jeffreys

For forensic applications, the work of Alec Jeffreys with on Restriction Fragment Length Polymorphisms (RFLPs) or Variable Number of Tandem Repeats (VNTRs) 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.

“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

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.

In Freeze 6: 800 million Single Nulceotide Variants.

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 on 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

$$0 \leq \text{Pr}(A|I) \leq 1$$

$$\text{Pr}(A|A, I) = 1$$

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

$$\begin{aligned} \text{Pr}(A \text{ or } B|I) &= \text{Pr}(A|I) + \text{Pr}(B|I) \\ \text{so } \text{Pr}(\bar{A}|I) &= 1 - \text{Pr}(A|I) \end{aligned}$$

($\bar{A}$ 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:

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

Therefore, for independent events

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

This may be written as

$$\text{Pr}(AB) = \text{Pr}(A)\,\text{Pr}(B)$$

Law of Total Probability

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

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

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.

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:

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

This can be rearranged to give

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

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:

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

Odds Form of Bayes' Theorem

From the third law of probability

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

Taking the ratio of these two equations:

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

Posterior odds = likelihood ratio × prior odds.

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

$$\begin{aligned} P_M &= 1-\{1(1-P)(1-2P)\cdots[1-(N-1)P]\} \\ &\approx 1-\prod_{i=0}^{N-1} e^{-iP} \approx 1-e^{-N^2P/2} \end{aligned}$$

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?

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.

Evett IW, Weir BS. 1998. “Interpreting DNA Evidence.” Can be downloaded from
`www.biostat.washington.edu/ bsweir/InterpretingDNAEvidence`

Updating Odds

From the odds form of Bayes' theorem:

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

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

where

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

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 B, Vignaux GA, Berger CEH. 2016.*Interpreting Evidence (Second Edition)*. Wiley.

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

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

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):

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

and then likelihood ratio becomes

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

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.

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.

Section 2: STR Typing Characteristics

STR Typing

Forensic DNA interpretation has been centered on the analysis of STRs (*short tandem repeats*), i.e. short DNA sequences that are repeated several times. These repeat patterns are located in areas called *loci* and vary among individuals. Variants for a given locus are called *alleles* and it is this variation (called polymorphism) that allows us to associate a particular DNA sample with an individual person.

To effectively interpret DNA evidence, we need to understand STR typing characteristics such as

- the PCR process
- anomalies (like mutations, stutter, and drop-ins/drop-outs)
- peak height variability

Understanding PCR

To produce an STR profile from a biological sample many identical copies of the DNA molecules within the target region (i.e. the DNA *template*) are needed. PCR (*polymerase chain reaction*) can be used to copy, or amplify, DNA through the following steps:

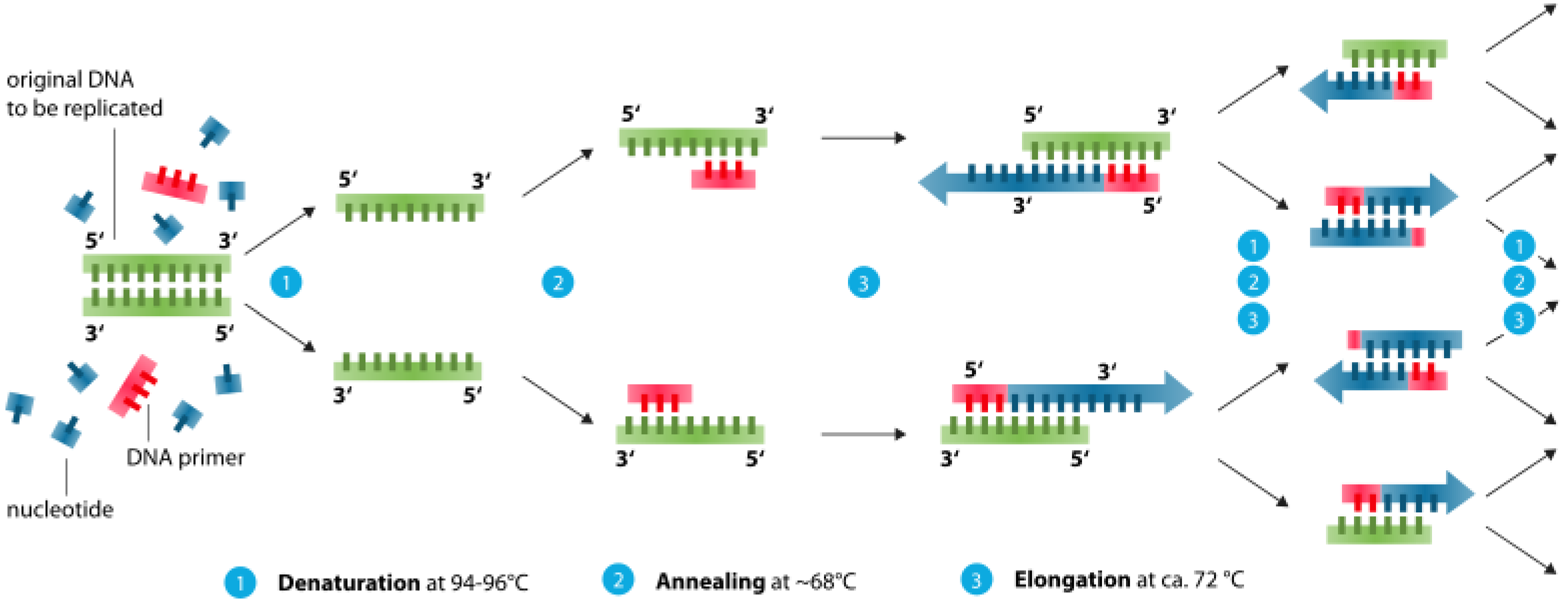
- **Denaturation:** Melting DNA such that the double-stranded template separates into two single-stranded DNA molecules.

- **Annealing:** Cooling the mixture to let *primers* bind to the strands.

- **Elongation:** DNA polymerase (a special copier molecule) completes missing sequences using available nucleotides.

Understanding PCR

These basic steps constitute one cycle, so by repeating this process, the DNA target gets amplified to millions of copies.



Source: `https://en.wikipedia.org/wiki/Polymerase_chain_reaction`

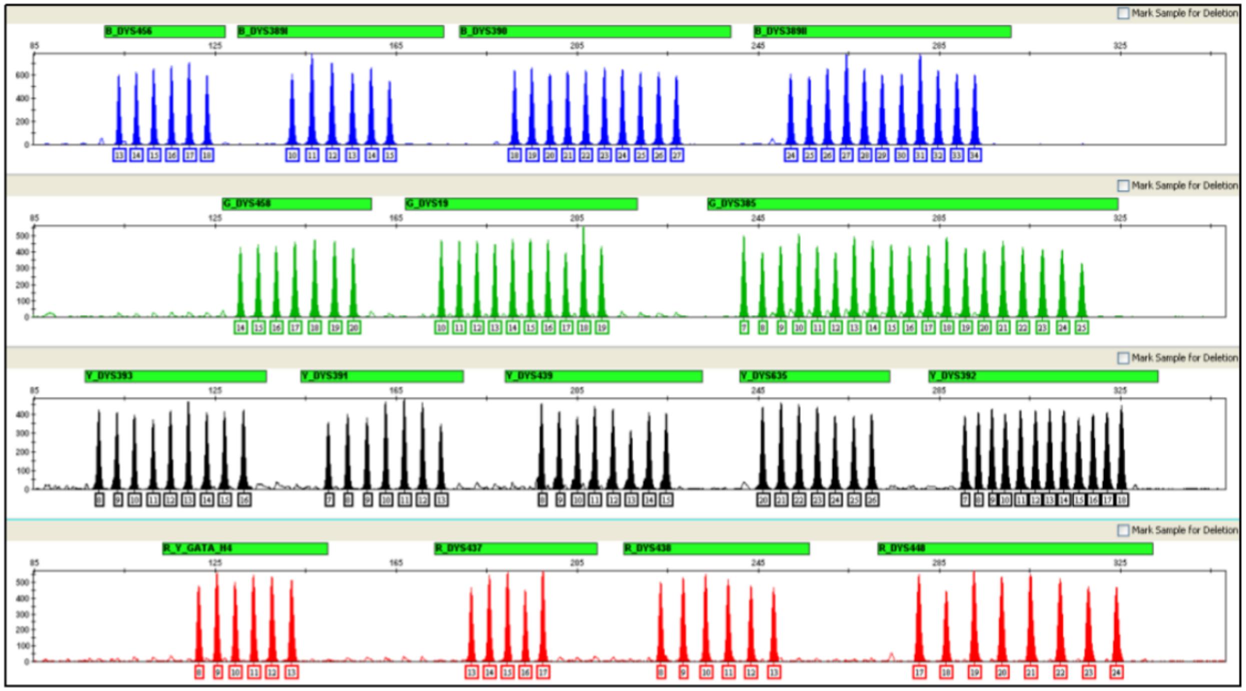
Capillary Electrophoresis

To obtain meaningful results from the PCR process, *capillary electrophoresis* (CE) has traditionally been used, allowing forensic scientists to gain access to the allele numbers contained in a DNA sample.

- DNA products are injected into the capillary where they travel in the direction of a positive charge;
- The travel time depends on the fragment size and can thus be used to infer the number of repeats;
- Primers are labeled with fluorescent dye, which will emit visible light at the detector window of the capillary.
- The fluorescence, measured in relative fluorescence units (RFU), is recorded over time and can be visualized with an *electropherogram* (epg).

Allelic Ladders

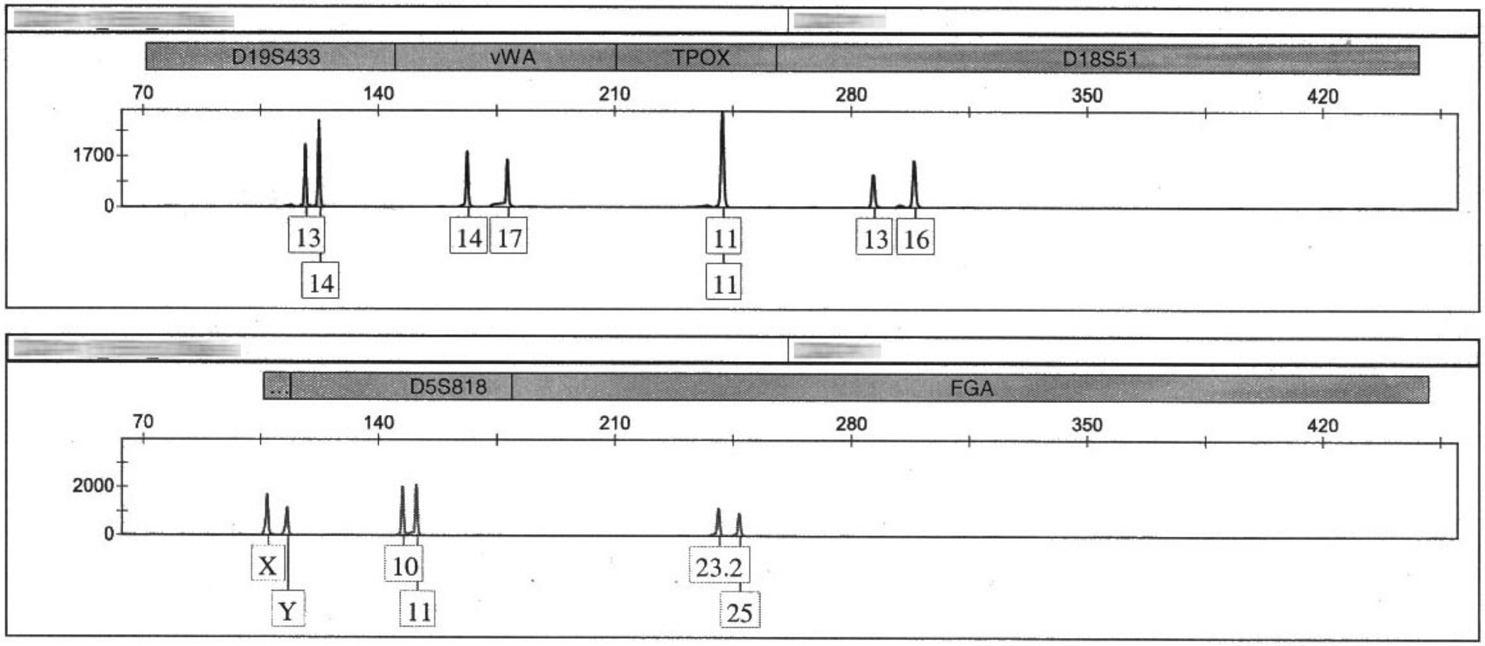
PCR-CE output can be compared to *allelic ladders* to determine allele designations.



Source: *AmpFℓSTR Yfiler PCR Amplification Kit User Guide*.

Example of an Electropherogram

An epg shows allelic designations, represented by peaks, with integer values indicating the number of complete repeat motifs and additional nucleotides separated by a decimal point.



Source: `https://en.wikipedia.org/wiki/Microsatellite`

STR Classes

STR loci may be categorized in three different classes, based on how well alleles conform to the core repeat pattern:

- **Simple STRs**: only show variation in the number of repeats without additional sequence variation.

- **Compound STRs**: consist of several adjacent repeats of the same repeat unit length.

- **Complex STRs**: contain repeats of variable length as well as sequences.

Examples of STR Classes

STR loci may be categorized in three different classes, based on how well alleles conform to the core repeat pattern:

Class	Locus	Allele sequence
Simple	CSF1PO	$[\mathrm{TCTA}]_8$
Simple	Penta D	$[\mathrm{AAAGA}]_{12}$
Compound	vWA	$[\mathrm{TCTA}][\mathrm{TCTG}]_4[\mathrm{TCTA}]_{13}$
Compound	D22S1045	$[\mathrm{ATT}]_7\mathrm{ACT}[\mathrm{ATT}]_2$
Complex	FGA	$[\mathrm{TTTC}]_3\mathrm{TTTTTT}[\mathrm{CTTT}]_{11}\mathrm{CTCC}[\mathrm{TTCC}]_2$
Complex	D1S1656	$[\mathrm{TAGA}]_4\mathrm{TGA}[\mathrm{TAGA}]_{13}\mathrm{TAGG}[\mathrm{TG}]_5$

Anomalies

If DNA profiling technologies were flawless, and no other (human) errors have been introduced, an STR profile would provide a perfect representation.

For good-quality samples, this is a reasonable assumption and STR allele calling is usually pretty straightforward.

However, a number of anomalies may still arise. And more importantly, crime scene profiles rarely belong to this category and usually consist of low template samples that may be contaminated and/or degraded, making them even more prone to typing errors.

PCR-CE Method Response Categories

PCR-CE method response can be classified into several categories:

- **Analyte signal**: peaks corresponding to one or both authentic alleles at a locus.

- **Molecular artifacts**: peaks identifiable as systematic method error, such as stutter.

- **Background noise**: method response resulting from negative controls or that cannot be classified as analyte signal or molecular artifact.

Mutations

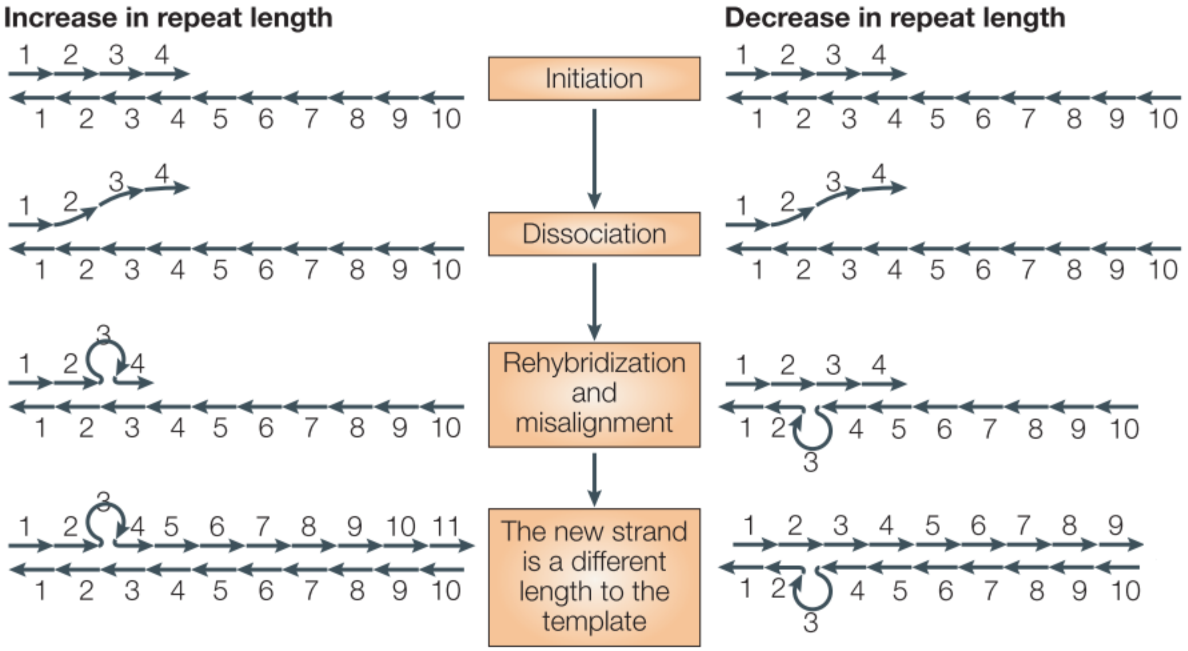
Mutations are the cause of the variation encountered in DNA and one of the reasons that STR loci render highly informative markers in forensic genetics. Most of the mutations are caused by an error during DNA replication (although other mechanisms and external influences can also lead to a change in DNA sequence).

Examples of mutations:

- **Substitutions:** A point mutation where one base is substituted for another, such as a SNP.
- **Indels:** Small insertions/deletions due to the addition of one or more extra nucleotides into the DNA or the loss of a section of DNA.

Slipped Strand Mispairing

STR polymorphisms derive mainly from variability in length. A proposed mechanism for these genetic variations is the *slipped strand mispairing* (SSM) mechanism: the dissociation of replicating DNA strands followed by misaligned re-association.



Increase in repeat length
Decrease in repeat length
Initiation
Dissociation
Rehybridization and misalignment
The new strand is a different length to the template
1 2 3 4
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10 11
1 2 3 4 5 6 7 8 9

Source: Microsatellites: simple sequences with complex evolution (Ellengren, 2004).

Mutations

The average in vivo mutational rates of the core STR loci are estimated to be between 0.01% and 0.64%, although the exact mutation rate of a locus is associated with the base composition of the repeats and the length of the allele.

Meiotic mutations, occurring in the process of transmitting an allele from a parent to a child, can cause the child's allele to differ from its parental type and can be important for paternity and other relatedness testing.

Mitotic mutations, or somatic mutations, occur within an individual and are of importance for identification and, although rare, could result in different profiles being recorded from the same individual (and hence possibly lead to a false exclusion).

Copy Number Variants

When mutations affect relatively large segments of DNA (1 kb or larger) the resulting difference is called a *copy number variant* (CNV).

This can lead to difficulties in forensic applications when:

- a deletion or duplication leads to an unusual pattern of peak heights (single peak with height similar to heterozygote alleles or unbalanced heights)

- a single-contributor profile contains a locus displaying three peaks (and may be confused with a low-template second contributor)

Micro-variants and sequence variations

An allele that contains an incomplete repeat unit is called a *micro-variant*. They are usually rare, with the exception of allele 9.3 at THO1, and can be reliably distinguished if the variant alters the allele length.

However, same-length variants (i.e. *isoalleles*) will be recorded as matching alleles even if they differ at sequence level. This means that CE-based methods have less discriminatory capability than is potentially available via sequencing techniques.

Locus	**Allele number**	**Allele sequence**
D3S1358	15	$[\mathrm{TCTA}][\mathrm{TCTG}]_3[\mathrm{TCTA}]_{11}$
D3S1358	15	$[\mathrm{TCTA}][\mathrm{TCTG}]_2[\mathrm{TCTA}]_{12}$
D18S51	20	$[\mathrm{AGAA}]_{20}$
D18S51	20	$[\mathrm{AGAA}]_{16}\mathrm{GGAA}[\mathrm{AGAA}]_3$

Stutter

Since STR typing methods make use of the PCR process, which relies on DNA replication characteristics, replication slippage also exists during DNA amplification of STRs in vitro.

This phenomenon manifests itself in an epg in the form of a *stutter* peak, i.e. a non-allelic peak that differs in size from the main product, usually by multiples of the length of the repeat unit, appearing adjacent to an allelic peak.

As a consequence, most profiling techniques cannot be used to study in vivo mutational dynamics.

Stutter Characteristics

The characteristics shared by mutations and stutter are considerable:

- Rates increase with the number of repeat units (i.e. less stutter for shorter alleles, more stutter for longer alleles);

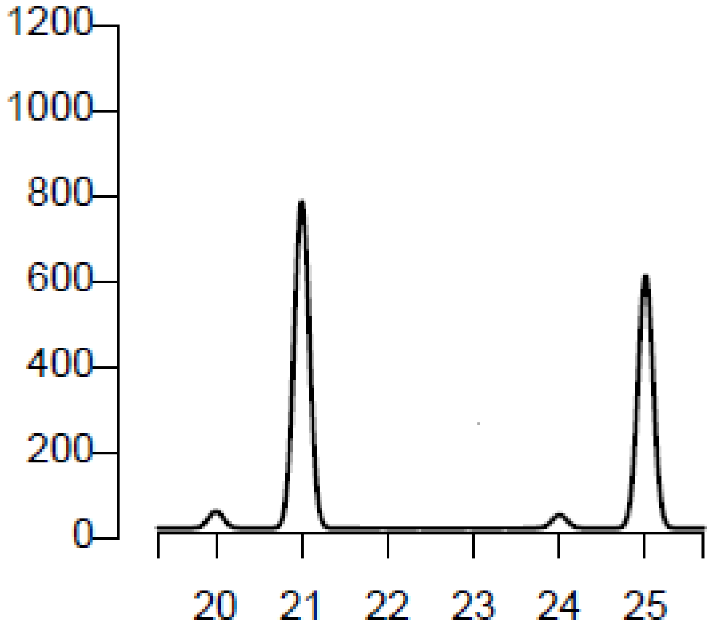
- Are inversely correlated with repeat unit length (i.e more stutter for dinucleotide repeats, less stutter for tetranucleotide repeats);

- And typically involve the insertion or deletion of a complete repeat unit.

Stutter Categories

Stutter can be primarily recognized as peaks whose length places them in 'stutter position' of other peaks present within a sample.

- Back stutter
- Forward stutter
- Double back stutter
- Two bp stutter



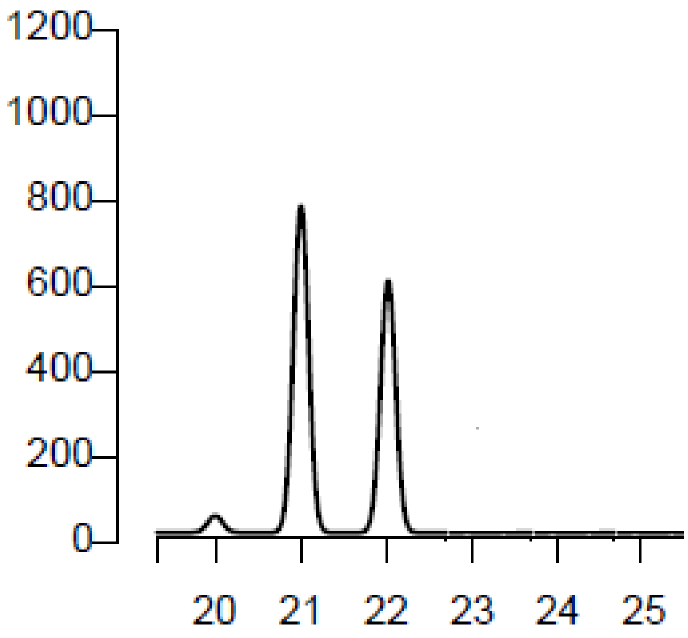
Stutter Difficulties

It is not always possible to distinguish stutter from other molecular artifacts or analyte signal:

- Stutter affected heterozygous genotypes;
- Composite stutter;
- Increase in repeat motif canceled out by a contraction;
- Compound repeats differing one nucleotide in repeat motif.

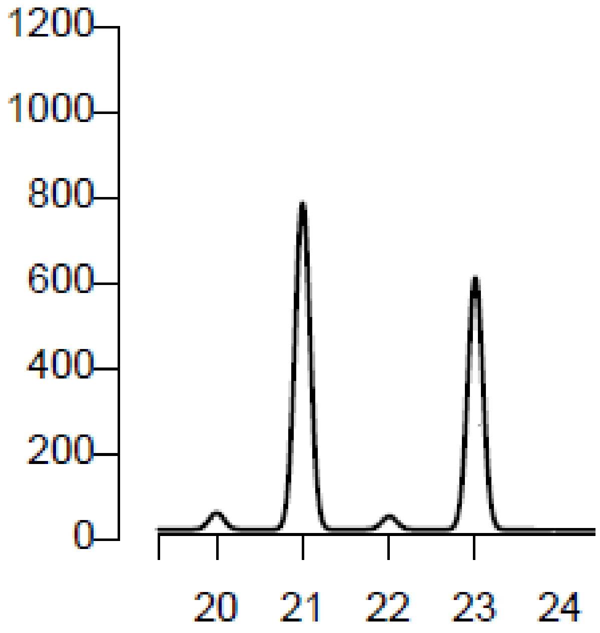
Stutter Affected Heterozygotes

Stutter affected heterozygous genotypes occur when two authentic alleles are separated by one repeat, and the total peak heights are a combination of analyte signal and stutter.



Composite Stutter

Composite stutter arises when the difference between two authentic alleles consist of two repeats and forward stutter of the low molecular weight allele coincides with back stutter of the high molecular weight allele.



Stutter Expansion and Contraction

In rare situations, an increase in repeat motif may cancel out a repeat contraction. This artifact would not be in stutter position and can only be recognized if the expansion and contraction involve different repeats.

$$[\mathrm{TCTA}][\mathrm{TCTG}]_3[\mathrm{TCTA}]_{11} \quad \Leftrightarrow \quad [\mathrm{TCTA}][\mathrm{TCTG}]_2[\mathrm{TCTA}]_{12}$$

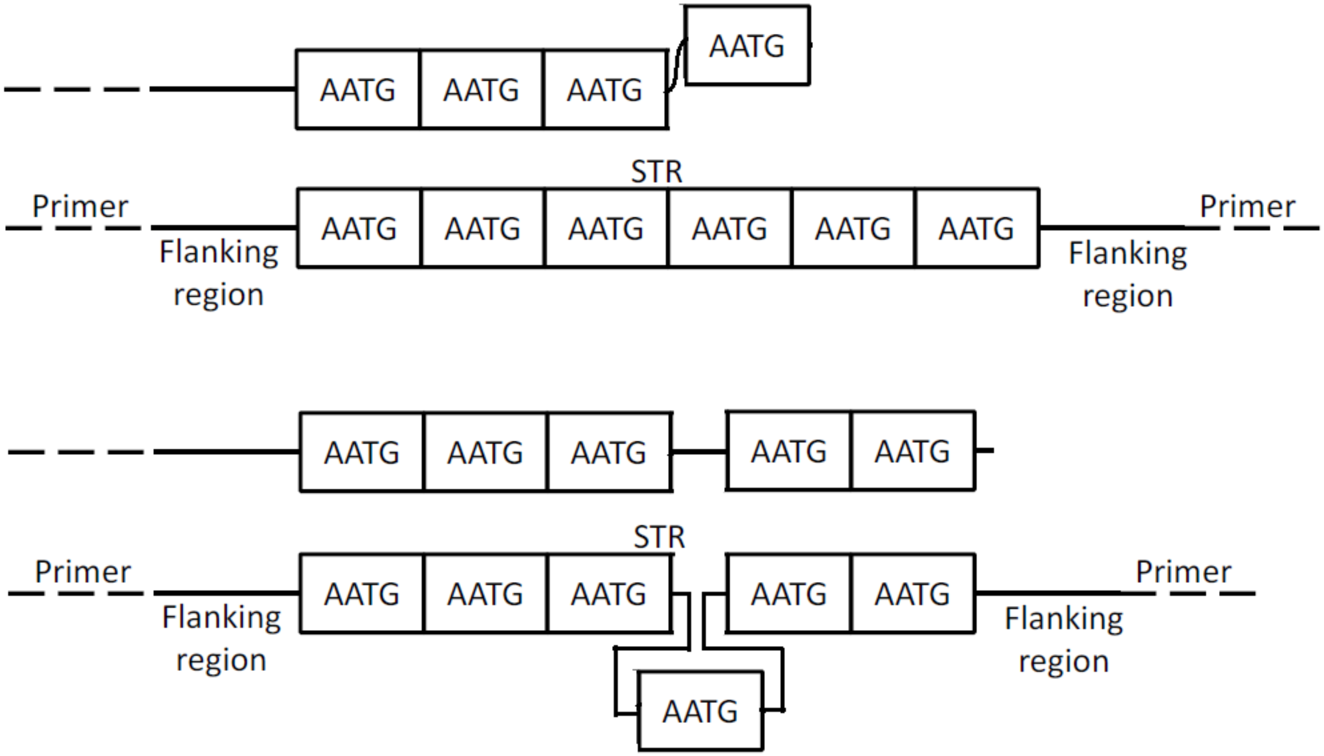
Stutter vs. Substitutions

If adjacent repeats of compound STR loci differ by a single nucleotide and are repeated only once or twice, stutter products can possibly not be distinguished from substitution errors.

$$\text{[TCTA][TCTG]}_{1-2}\text{[TCTA]}_{13-15} \quad \Rightarrow \quad \text{[TCTA]}_{13-17}$$

Back Stutter

Back stutter is the most prevailing type of stutter, suggesting a preference for repeat contractions over expansions (which are energetically less favorable).



Double Back Stutter

There are two possible mechanisms for the creation of double back stutter:

- Direct creation caused by a double loop during slipped strand mispairing;
- Stutter of a previously formed stutter product.

It is suggested that a double loop is more likely than stutter of stutter, at least for Y-STR data.

$$\begin{aligned}[\mathrm{AATG}]_6\mathrm{ATG}[\mathrm{AATG}]_3 \quad &\Rightarrow \quad [\mathrm{AATG}]_4\mathrm{ATG}[\mathrm{AATG}]_3 \\ &\Rightarrow \quad [\mathrm{AATG}]_5\mathrm{ATG}[\mathrm{AATG}]_2\end{aligned}$$

Source: Modelling PowerPlex Y stutter and artefacts (Bright et al., 2011).

Drop-ins

Allelic peaks that do not come from any of the assumed contributors to a DNA sample are termed *drop-ins*.

- Drop-ins may arise from airborne DNA fragments in a laboratory, or due to environmental exposure at the crime scene, and can typically not be reproduced on subsequent analysis of the same DNA extract.

- Verification of the source of drop-ins is not usually possible, although the existence of drop-ins can be confirmed through negative controls.

- As techniques become more sensitive, more drop-ins will occur, and potential difficulties may arise when they are incorrectly classified as analyte signal.

Contamination

Drop-ins are related to the concept of *contamination*.

- Contamination is one of the causes for drop-ins, as a result of DNA that got into a sample during collection or subsequent analysis.

- Databases of lab and scene staff can facilitate the identification of certain kinds of contamination.

- The most dangerous form of contamination is between different evidence samples, from either the same or different crime scenes.

- The observation of a more complete profile resulting from contamination is referred to as gross contamination.

Drop-outs

A *drop-out* occurs when an allele from a contributor to the crime scene sample is not reported in the STR profile.

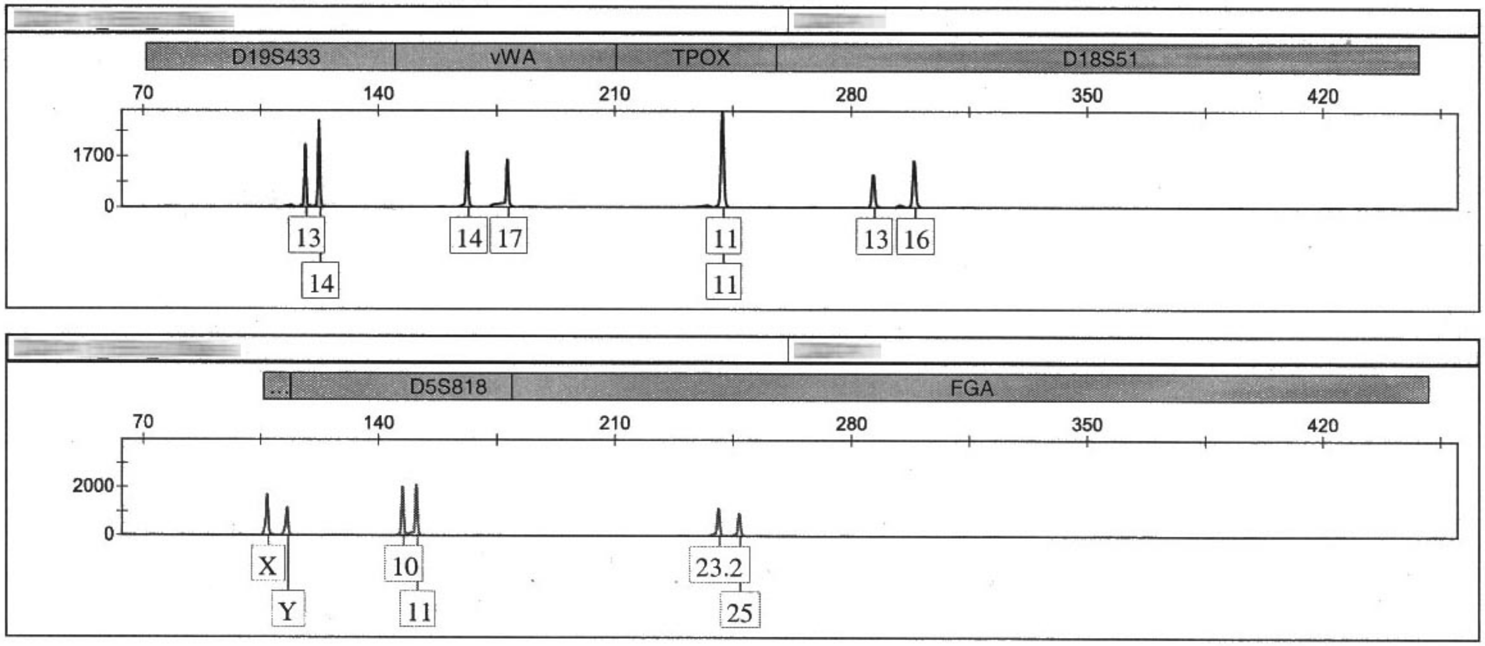
- This happens when a peak fails to reach the detection threshold, meaning that they cannot be reliably distinguished from background noise.

- Low template DNA samples and degradation increase the drop-out rate, which is believed to be associated with DNA fragment length.

- Drop-outs should not be confused with *silent alleles*, in which a system is unable to visualize an allele.

Peak Height Variability

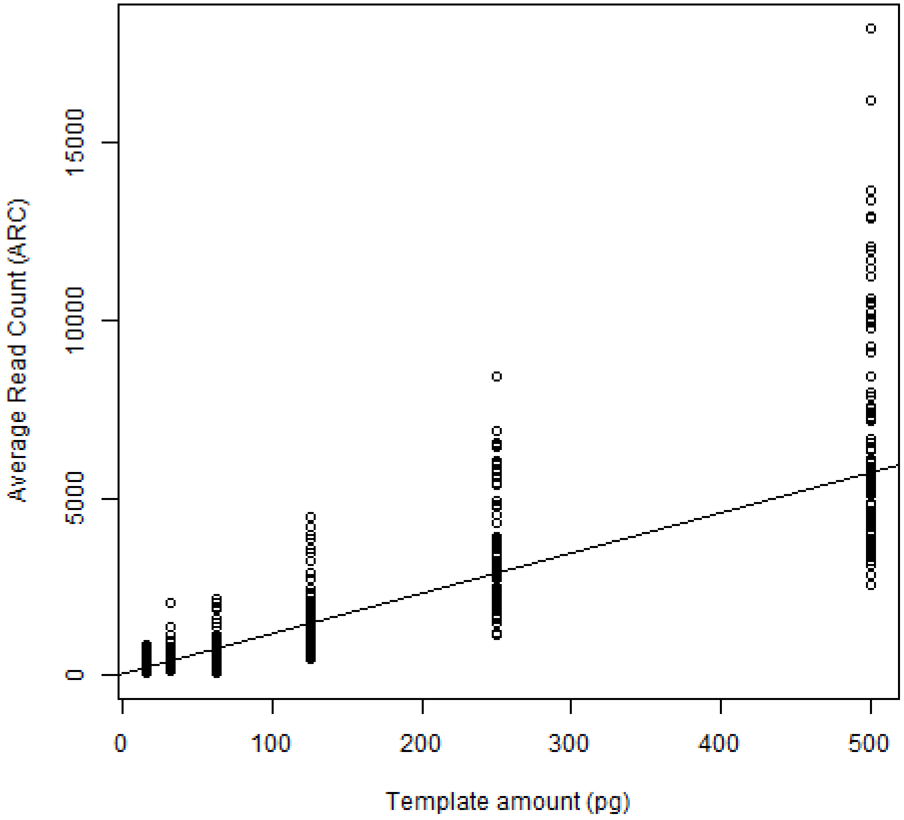
Besides the anomalies already discussed, several other factors play a role in observed variations within STR profiles.



Source: `https://en.wikipedia.org/wiki/Microsatellite`

Template

In theory, peak heights from a single contributor are expected to be approximately proportional to the amount of undegraded DNA template.



Template

The amount of DNA for each contributor to a sample will therefore directly relate to the peak height of contributors.

In practice, there exists some stochastic variation in peak height.

Nowadays, only a couple of picograms of DNA is sufficient to produce results. However, for these *low template DNA* (LTDNA) samples, stochastic effects can play a major role and will invariably influence the analysis (and likely decrease the statistical weight of the evidence).

Degradation

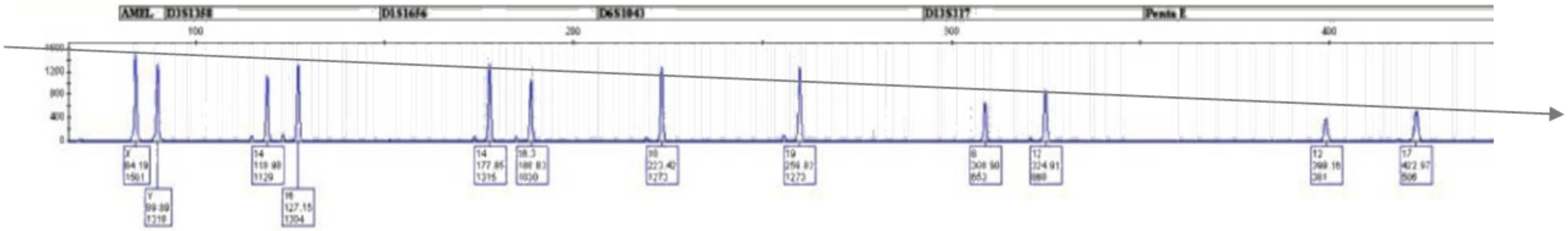
DNA evidence is prone to degradation due to a variety of mechanisms and circumstances, including chemical processes and environmental conditions, causing breakage of previously intact DNA molecules.

If breakage occurs in regions where primers anneal, or between the forward and reverse primers, target regions may not amplify efficiently or fail to amplify at all.

Degradation

Studies suggest that degradation leads to peak heights showing a downward trend with increasing molecular weight, supposedly because smaller alleles are more resistant to degradation.

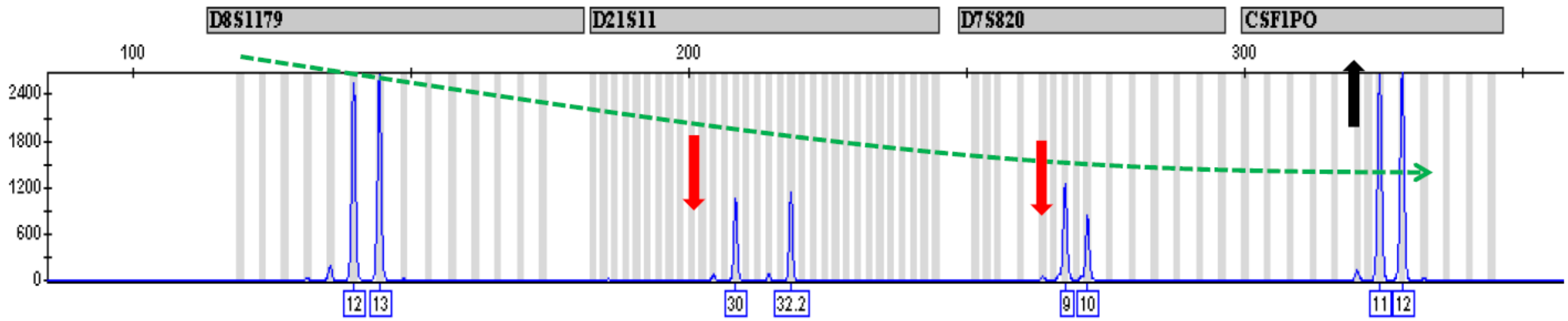
This observation is sometimes referred to as the degradation slope or the ski slope.



Locus Specific Amplification Efficiency

Additional variability arises from differences in amplification efficiency per locus. Observations show that some loci amplify more efficiently than others, and that these differences appear to vary over time.

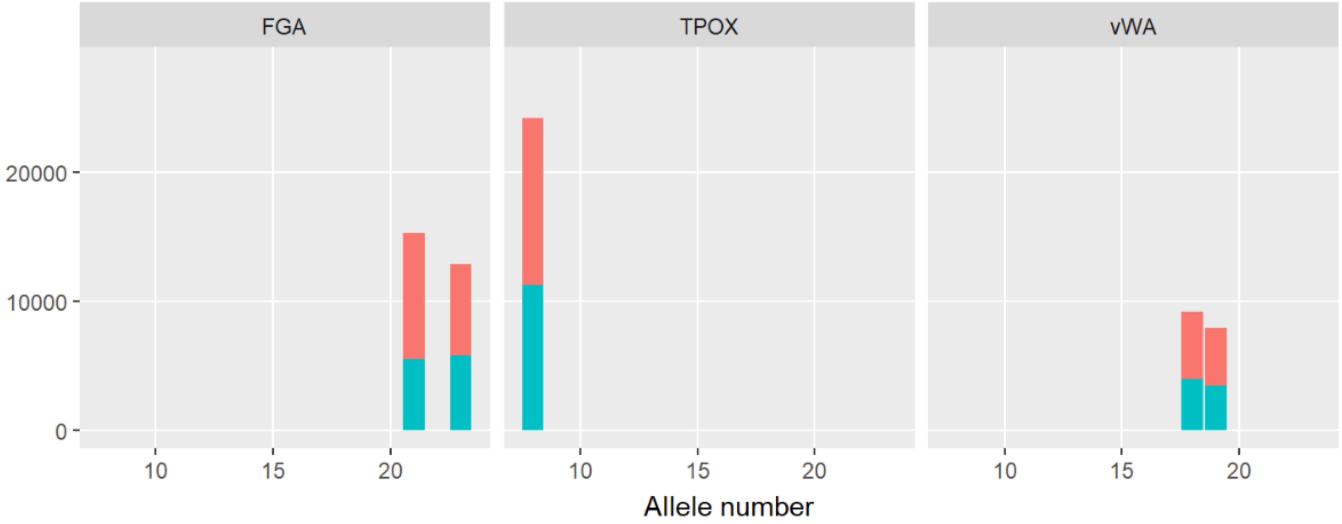
Amplification bias is thought to be a result of the large variation in target loci length.



Replicates

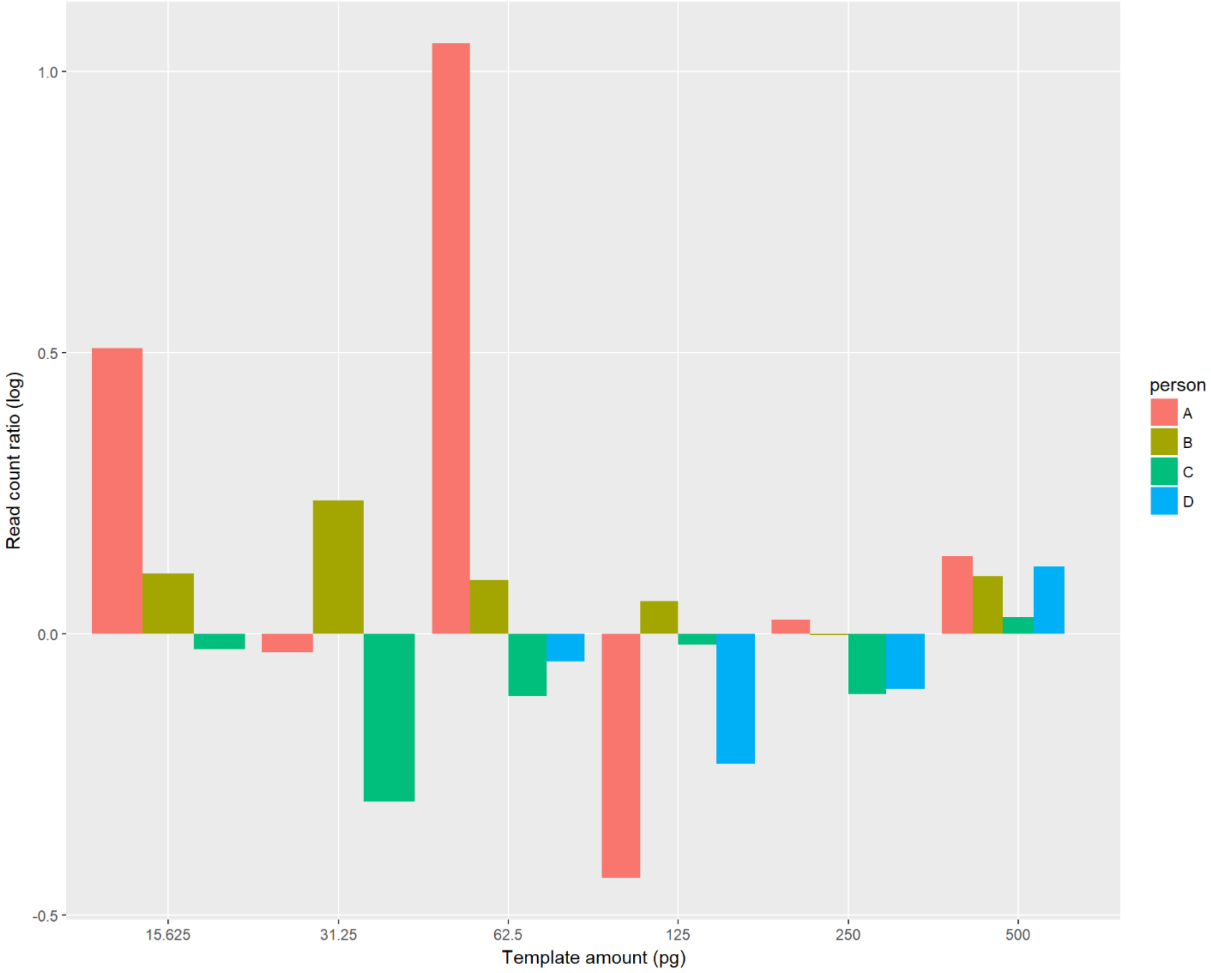
Replicates will show different replicate amplification efficiencies, but can be consolidated into a single analysis, even for different amounts of template DNA. As long as replicates originate from the same DNA extract, they can be used to obtain a more accurate genotype profile.

Replication is not always possible, and in case of a LTDNA sample it would probably be preferable to use as much as possible of the available DNA to give the best possible single-run profile.



Replicate Consistency vs. Template Amount

Higher template amounts result in more balanced peak heights between replicates.



Heterozygote Balance

A consequence of all the stochastic variations that have been introduced into the process, is that the two peaks of heterozygous alleles will also show variability, termed the *heterozygote balance*.

The difference is thought to be affected by the number of repeat sequences, since high molecular weight alleles:

- Stutter more;
- And amplify less.

Heterozygote Balance

Understanding the variability in heterozygous balance is important for the interpretation of mixed profiles and low template DNA:

- For LTDNA, peaks may be so imbalanced that it leads to alleles not exceeding the allelic threshold or even a drop-out.

- It may be used to classify combinations of alleles (or genotypes) as possible or impossible when considering a mixture.

STR Typing Characteristics

To effectively interpret DNA evidence, phenomena and factors like **mutations, CNVs, contamination, template amount, replicates, amplification efficiency, and degradation** should be considered.

These lead to observations in the form of **stutter, drop-ins, drop-outs, peak height variability and heterozygote balance**, that may need to be incorporated in weight-of-evidence calculations.

Section 3:
Allelic Independence and Matching

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:

Allele	<12	12	13	14	15	16	17	18	19	>19	Total
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."

Lander ES. 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 marriages 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:

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

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:

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

Exact HWE Test

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

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

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	$\Pr(n_{AA}, n_{Aa}, n_{aa} \vert n_A, n_a, \text{HWE})$
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:

			Exact	
n_{AA}	n_{Aa}	n_{aa}	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 smaller probability. We do not reject the HWE hypothesis in this case.

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.

However, if two profiles match at several loci this may be because they are from the same, or related, people and so are likely to match at additional loci.

Allele Matching

Forensic genetics is concerned with matching of genetic profiles from evidence and from persons of interest. Profile match probabilities rest on the probabilities of matching among the alleles constituting the profiles.

Allele matching can refer to alleles within an individual (inbreeding), between individuals within a population (relatedness) and between populations (population structure). In all these cases there are parameters that describe profile match probabilities, and these parameters can be estimated by comparing the observed matching for a target set of alleles with that between a comparison set.

Allele Matching Within Individuals

The inbreeding coefficient for an individual is the probability it receives two alleles at a locus, one from each parent, that are *identical by descent.*

What can be observed, however, is identity in state. An individual is either homozygous or heterozygous at a locus: the two alleles either match or miss-match at that locus. The proportion of matching alleles at a locus is either zero or one, not a very informative statistic, but the proportion of an individual's loci that are homozygous may be informative for their inbreeding status.

There is still a need for a reference: for a locus such as a SNP with a small number of alleles many loci will be homozygous even for non-inbred individuals. Therefore we compare the proportion of loci with matching alleles for an individual with the matching proportion for pairs of alleles taken one from each of two individuals: is allele matching higher within than between individuals?

Inbreeding

If $\tilde{M}_j$ is the observed proportion of loci with matching alleles (i.e. homozygous) for individual j, and if $\tilde{M}_S$ is the observed proportion of matching alleles, one from each of two individuals in the population, then the within-population inbreeding coefficient f_j is estimated as

$$\hat{f}_j \;=\; \frac{\tilde{M}_j - \tilde{M}_S}{1 - \tilde{M}_S}$$

Note that this can be negative for individuals with high degrees of heterozygosity.

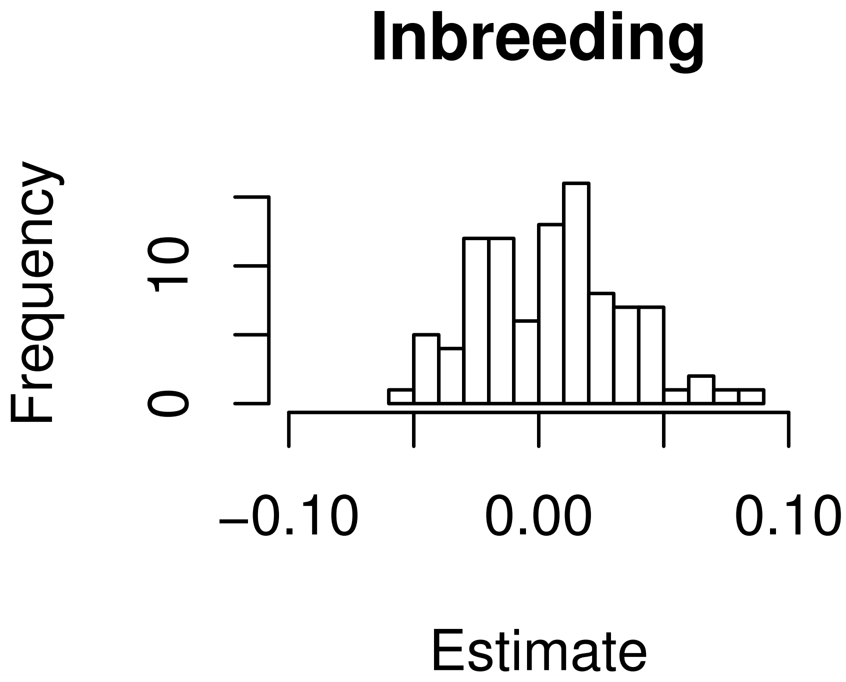
The average of these estimates over all the individuals in a sample from a population estimates the within-population inbreeding coefficient f:

$$\hat{f} \;=\; \frac{\tilde{M}_I - \tilde{M}_S}{1 - \tilde{M}_S}$$

where $\tilde{M}_I = \sum_{j=1}^{n} \tilde{M}_j / n$. Hardy-Weinberg equilibrium corresponds to $f = 0$.

SNP-based Inbreeding

From 400,000 SNPs on Chromosome 22 of the 1000 Genomes ACB populations (96 Afro-Caribbeans in Barbados);



Allele Matching Between Individuals

How can we tell if a pair of individuals has a high degree of allele matching? What does “high” mean?

We assess relatedness of individuals within a population by comparing their degree of allele matching with the degree for pairs of individuals with one from each of two different populations.

Allele Matching Between Individuals

If $\tilde{M}_{jj'}$ is the observed proportion of loci with matching alleles, one from each of individuals j and j', and if $\tilde{M}_S$ is the average of all the $\tilde{M}_{jj'}$'s, then the within-population kinship coefficient $beta_{jj'}$ is estimated as

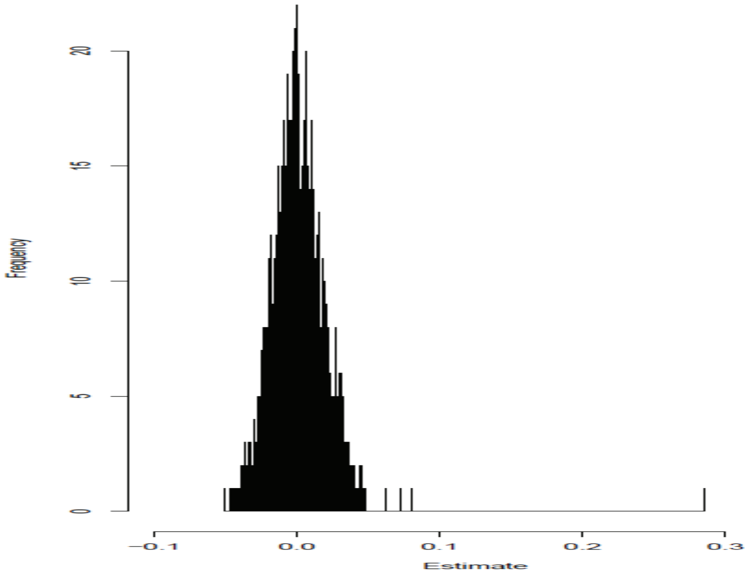
$$\hat{\beta}_{jj'} = \frac{\tilde{M}_{jj'} - \tilde{M}_S}{1 - \tilde{M}_S}$$

Note that this can be negative for pairs of individuals less related than the average pair-matching in the sample.

The average of these estimates over all pairs of individuals in a sample is zero, but this doesn't allow us to compare populations.

SNP-based Coancestry

From 400,000 SNPs on Chromosome 22 of the 1000 Genomes ACB populations (4560 pairs of Afro-Caribbeans in Barbados);



Allele Matching Between Populations

We calibrated allele matching within individuals by comparison with matching between pairs of individuals.

We calibrate the allele matching between pairs of individuals by comparison with matching between pairs of populations. If $\tilde{M}^{ii'}$ is the observed proportion of loci with matching alleles, one from each of populations i and i', and if $\tilde{M}_B$ is the average of all the $\tilde{M}^{ii'}$'s, then the total kinship coefficient $\beta_{jj'}$ is estimated as

$$\hat{\beta}_{jj'} = \frac{\tilde{M}_{jj'} - \tilde{M}^B}{1 - \tilde{M}^B}$$

The average of these estimates over all pairs of individuals in a sample from a population is

$$\hat{\beta} = \frac{\tilde{M}_S - \tilde{M}^B}{1 - \tilde{M}^B}$$

This is the "θ" needed for the "theta correction" discussed below.

Within-population Matching

We can get some empirical matching proportions when we have a set of profiles. To simplify this initial discussion, consider the following data for the Y-STR locus DYS390 from the NIST database:

	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 between individual-pair 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.

Section 4: DNA Interpretation and Modeling

DNA Interpretation and Modeling

- Thresholds

- Weight of evidence

- LR calculations

- LR modeling

Thresholds

The most straightforward way to interpret an STR profile is with the use of thresholds.

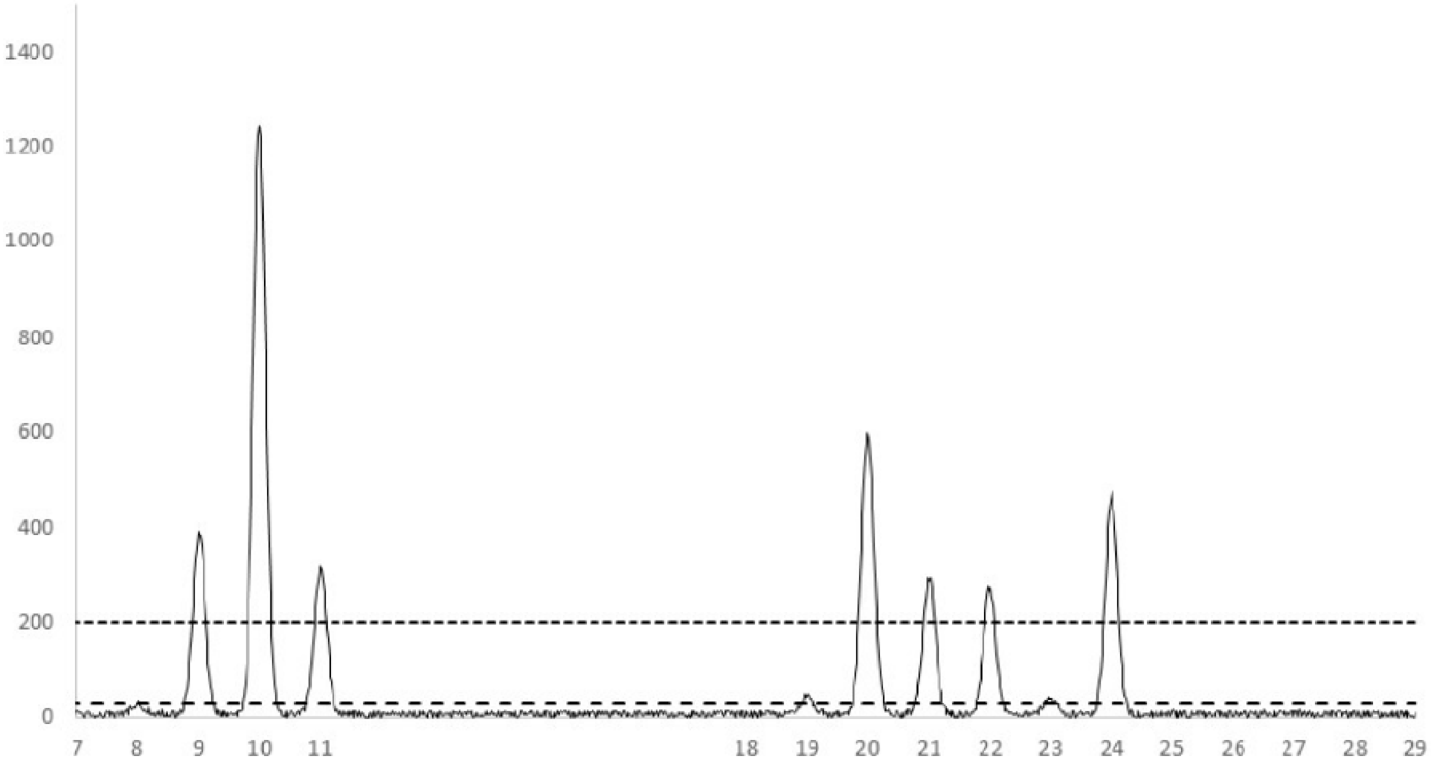
- **High thresholds**: will reduce the number of artifacts and remove a lot of background noise. However, it may potentially lead to a number of drop-outs.

- **Low thresholds**: will detect more authentic alleles, but have a higher probability of showing drop-ins.

Thresholds

An *analytical threshold* (AT) is usually set as a limit above which method response is interpreted as an authentic allele.

Additional stutter thresholds can help improve mixture profile interpretation (e.g. $5 - 15\%$ of the main allele).



Weight of Evidence

An STR profile obtained from a crime scene sample can be compared to a person of interest, and it may be found that this person cannot be excluded. An ‘inclusion’ may be reported, but is practically worthless without some expression on the strength of this evidence.

The Island Problem

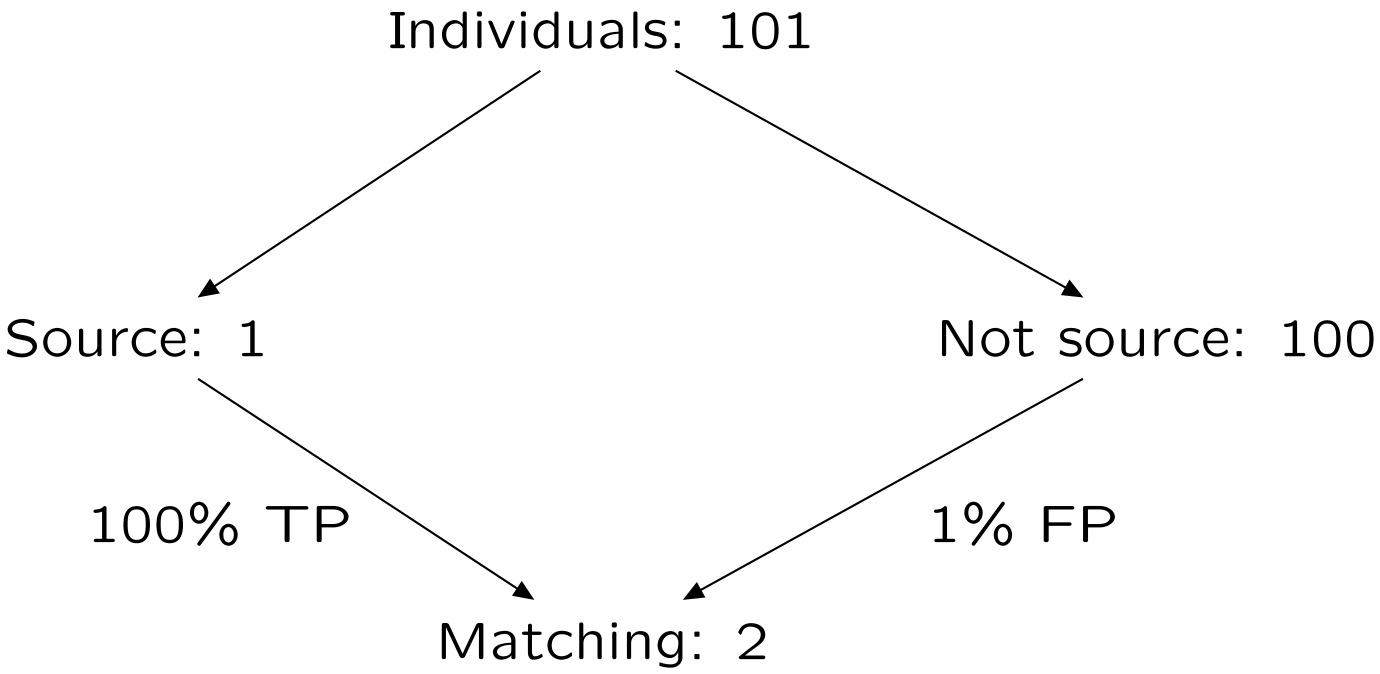
Suppose there is a crime committed on a remote island with a population of size 101. A suspect Q is found to match the crime scene profile. What is the probability that Q is the source of the profile, assuming that:

- All individuals are equally likely to be the source.
- The DNA profiles of all the other individuals are unknown.
- We expect 1 person in 100 to possess this observed profile.

Source: Weight-of-Evidence for Forensic DNA Profiles (Balding & Steele, 2015)

The Island Problem - Solution

In addition to Q, we expect one other individual on the island to match. So, even though the profile is rare, there is only a 50% chance that Q is the source.



Individuals: 101

Source: 1

Not source: 100

100% TP

1% FP

Matching: 2

The Island Problem - Odds Version

Recalling the odds form of Bayes' theorem:

$$\frac{P(H_p|E)}{P(H_d|E)} = \frac{P(E|H_p)}{P(E|H_d)} \times \frac{P(H_p)}{P(H_d)},$$

with

$$P(H_p) = \frac{1}{101} \qquad\qquad P(E|H_p) = 100\%$$

$$P(H_d) = \frac{100}{101} \qquad\qquad P(E|H_d) = 1\%,$$

yielding prior odds of $\frac{1}{100}$ and a likelihood ratio of 100. Combining this gives posterior odd of 1, or equivalently, a 50%/50% chance.

The Island Problem - Odds Version

More generally,

$$P(H_p|E) = \frac{1}{1 + Np},$$

with N the number of individuals on the island other than the suspect, and p the profile probability of the observed DNA sample.

Extreme oversimplification of assessing the weight of evidence:

- Uncertainty about N and p
- Effect of searches, typing errors, other evidence
- Population structure and relatives

The Island Problem - Searches

Now suppose Q was identified through a search, with the suspect being the only one among 21 tested individuals who matches the crime scene profile.

- How does this knowledge affect the probability of being the source?
- What is the general expression for the probability of being the source, using k for the number of individuals who have been excluded?

The Island Problem - Searches

In this case we can exclude individuals from our pool of possible donors, such that our prior odds will slightly increase.

Out of the $N - k = 80$ individuals, we expect another 0.8 matches, yielding a probability of being the source of $1/1.8 \approx 56\%$. Or, in formula:

$$P(H_p|E) = \frac{1}{1 + (N-k)p},$$

where setting $k = 0$ gives the original expression and $k = N$ gives $P(H_p|E) = 1$.

Likelihood Ratio

As seen previously, the forensic scientist is concerned with assigning the likelihood ratio

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

which is equivalent to the reciprocal of the *profile probability* for the island problem:

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

although we observed that the *match probability* is a more relevant quantity:

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

Match Probabilities

Recall the match probabilities for homozygotes:

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

$$= p_A^2 \qquad (\text{if } \theta = 0),$$

and for heterozygotes:

$$P(AB|AB) = \frac{2[\theta + (1-\theta)p_A][\theta + (1-\theta)p_B]}{(1+\theta)(1+2\theta)}$$

$$= 2p_A p_B \qquad (\text{if } \theta = 0).$$

LR for a Single Locus

Consider a simple two-person mixture profile (e.g. contributors are unrelated, ignoring population structure, no drop-outs/drop-ins), where $G_C = ABCD$. What is the appropriate single-locus LR (assuming HWE and p_A, p_B, p_C and p_D are known) when:

- $G_S = AB$ and $G_K = CD$, with

  $$H_p : \text{K + POI (S)} \quad \text{and} \quad H_d : \text{K + Unknown (U)}$$

- $G_S = AA$, with:

  $$H_p : \text{K + S} \quad \text{and} \quad H_d : \text{K + U}$$

- $G_S = AB$ and the second contributor is unknown

  $$H_p : \text{S + U} \quad \text{and} \quad H_d : \text{2U}$$

LR for a Single Locus

Consider a simple two-person mixture profile (e.g. contributors are unrelated, ignoring population structure, no drop-outs/drop-ins), where $G_C = ABCD$. What is the appropriate single-locus LR (assuming HWE and p_A, p_B, p_C and p_D are known) when:

- LR $= \frac{P(ABCD|AB,CD,H_p)}{P(ABCD|CD,H_d)} = \frac{1}{2p_Ap_B}$;
- LR $= \frac{P(ABCD|AA,CD,H_p)}{P(ABCD|CD,H_d)} = 0$;
- LR $= \frac{P(ABCD|AB,H_p)}{P(ABCD|H_d)} = \frac{2p_Cp_D}{6 \cdot 4p_Ap_Bp_Cp_D} = \frac{1}{12p_Ap_B}$.

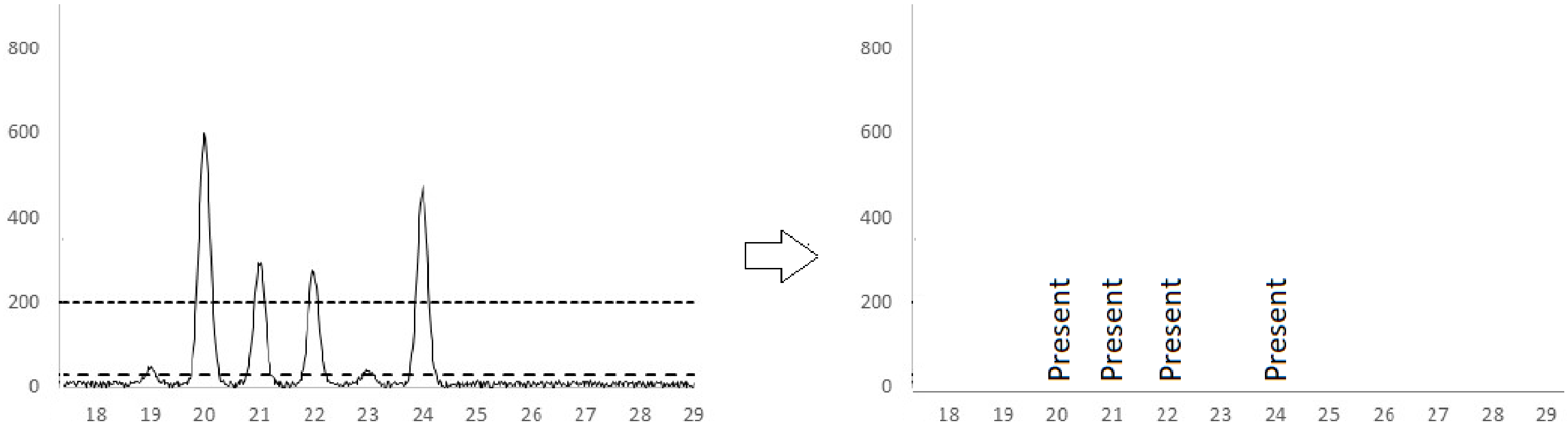
LR Modeling

Different approaches can be used to assess the likelihood ratio:

- Binary model
- Semi-continuous model
- Continuous model

Binary Model

A binary model limits interpretation of DNA profiles to qualitative allele callings only, without any attempt to infer the underlying genotypes (i.e. each are regarded as equally likely).



Just as in our previous example, single-locus LRs can be calculated and combined across loci via multiplication.

Semi-continuous Model

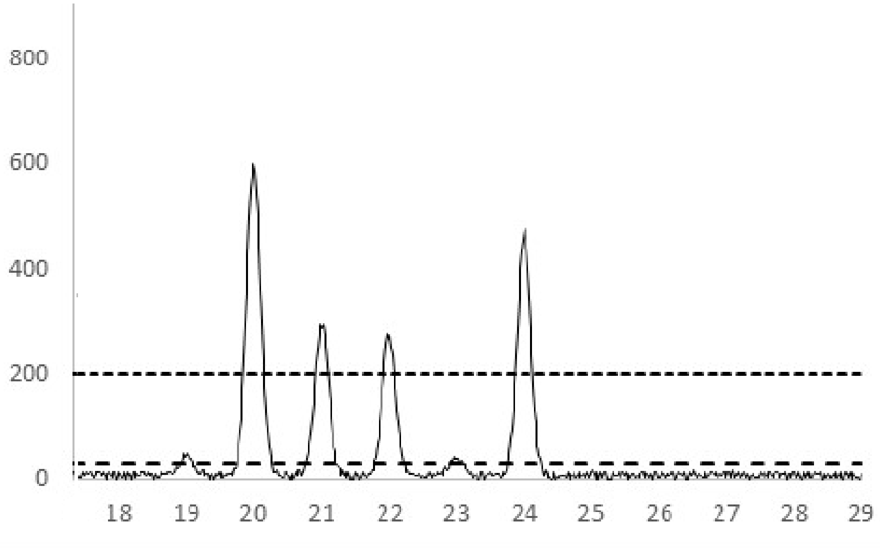
A semi-continuous model retains the simplicity of binary methods, but combines this with probabilistic modeling of known phenomena such as drop-ins and drop-outs.

Since these models still suffer from a significant loss of information, a more quantitative approach might be preferred.

Ideally, a statistical framework utilizes as much available quantitative information as possible, while maintaining comprehensibility.

Continuous Model

The key point of a fully continuous model is that it considers peak heights as a continuous variable.



Donor 1	Donor 2	Weights (Qualitative)	Weights (Quantitative)
20, 21	22, 24	1	0.05
20, 22	21, 24	1	0.05
20, 24	21, 22	1	0.75
21, 22	20, 24	1	0.05
21, 24	20, 22	1	0.05
22, 24	20, 21	1	0.05

Peak Height Modeling

Peak heights can be modeled by defining the *total allelic product* (TAP), which will be a function of

- the template amount t_n;
- a measure of degradation d_n;
- a locus-specific amplification efficiency A^l;
- a replicate multiplier R_r;
- and allele dosage X^l_{an}.

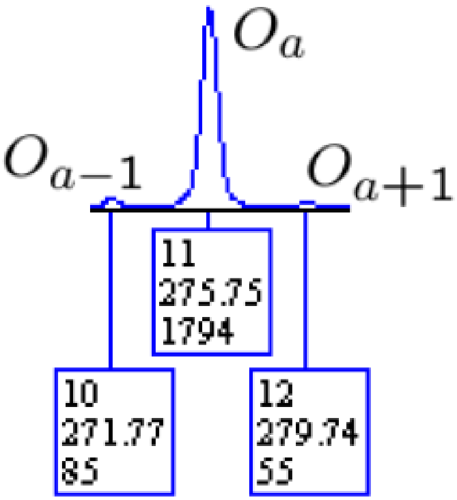
T^l_{arn} then describes the TAP of allele a at locus l, for replicate r from contributor n.

TAP Modeling

Theoretically, the previous slide models the peak heights, but in practice, we will observe slightly different values. This is because we haven't incorporated the concept of stutter yet.

If we allow for back stutter and forward stutter, we can write:

$$T_a = O_{a-1} + O_a + O_{a+1}.$$



Stacking

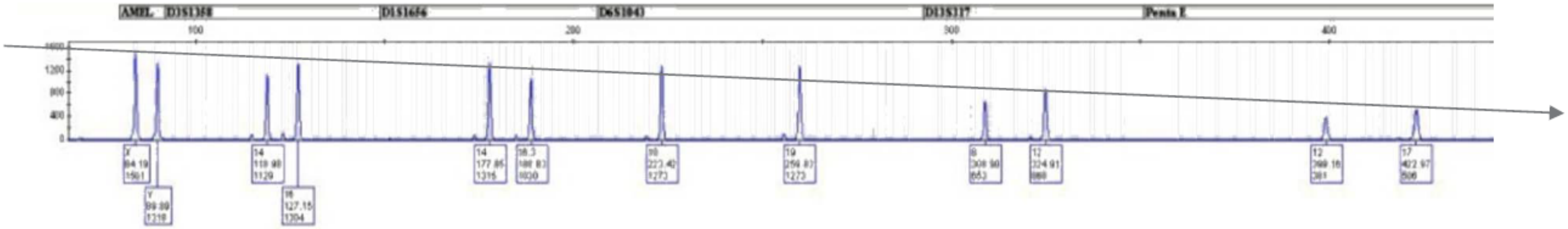
Note that we assume that expected peak heights are additive, i.e. if there are multiple sources of a single allele, the height of that allele will equal the sum of the individual expected heights from each source.

This assumption of additivity is called *stacking*.

Recent talks (Rudin, AAFS 2017) emphasize that this assumption has not been validated. To determine if this practice is scientifically supportable, it would be good to obtain a large set of mixtures from known profiles to look at the expected combined versus observed combined peak heights.

Modeling Degradation

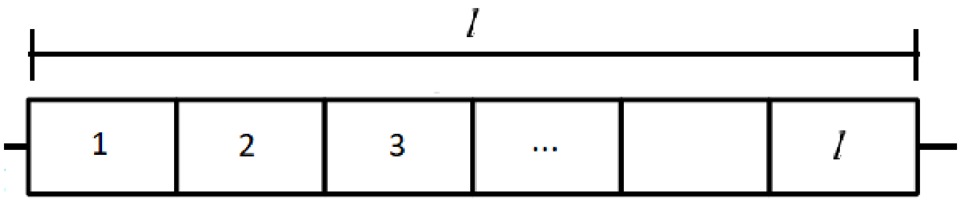
A simple model for degradation would be a linear model, i.e. peak heights decline constantly with respect to molecular weight.



If we assume that the breakdown of a DNA strand is random with respect to location, an exponential model seems more reasonable.

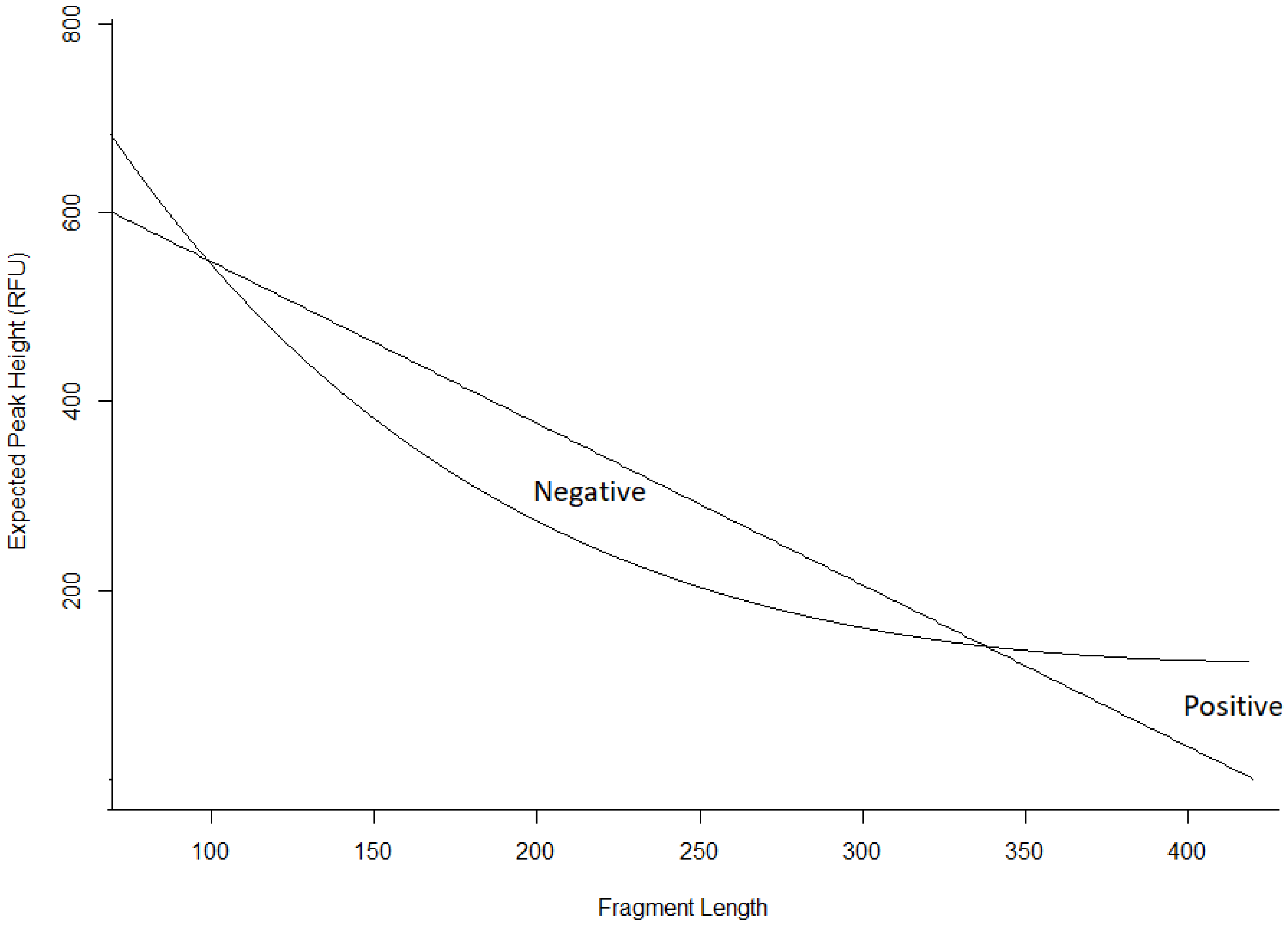
Modeling Degradation

- Consider a DNA fragment of length l.
- Let p be the probability of a break at any of the locations $1, \ldots, l$.
- The chance of the full fragment being amplified is $(1-p)^l$.
- This describes an exponential decline in peak heights.

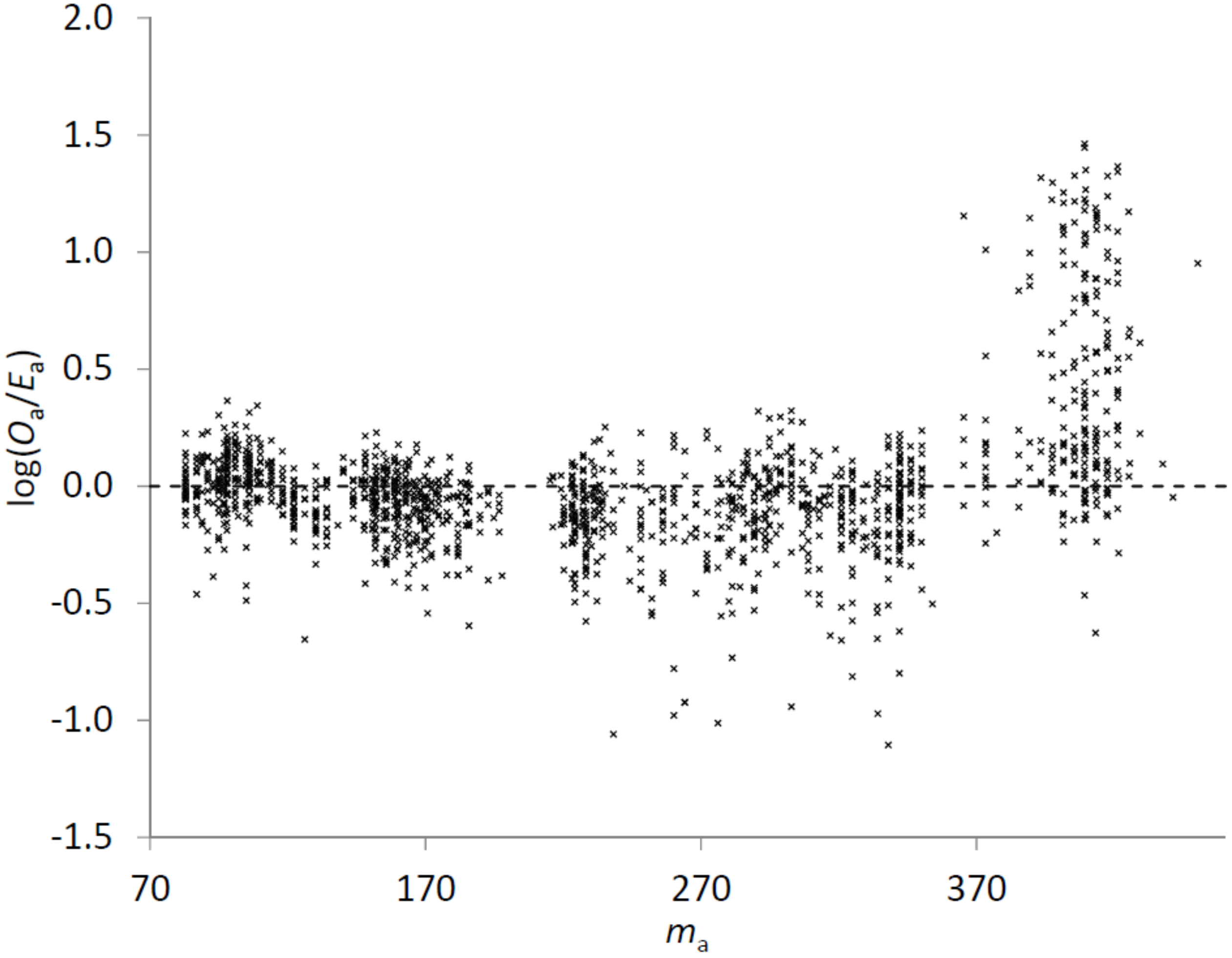


Source: Forensic DNA Evidence Interpretation (Buckleton et al., 2016).

Modeling Degradation

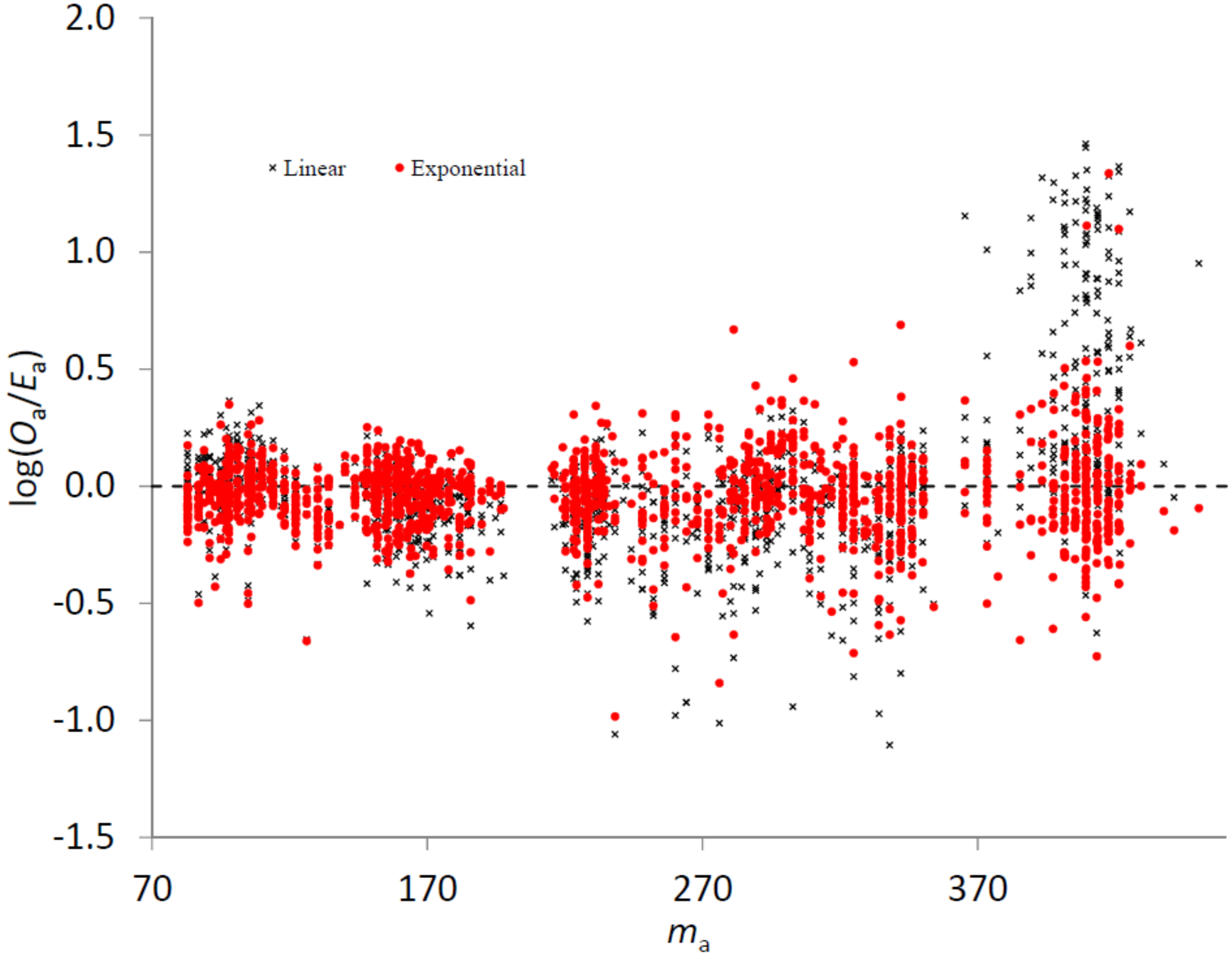


Modeling Degradation



Source: Degradation of Forensic DNA Profiles (Bright et al., 2013).

Modeling Degradation



Source: Degradation of Forensic DNA Profiles (Bright et al., 2013).

Modeling Heterozygote Balance

The heterozygote balance (Hb) is usually expressed as a peak height ratio, i.e. the ratio of two heterozygote peaks at a locus.

There are two common definitions:

$$\mathrm{Hb}_1 = \frac{O_{\mathrm{HMW}}}{O_{\mathrm{LMW}}}, \qquad \text{and} \qquad \mathrm{Hb}_2 = \frac{O_{\mathrm{smaller}}}{O_{\mathrm{larger}}},$$

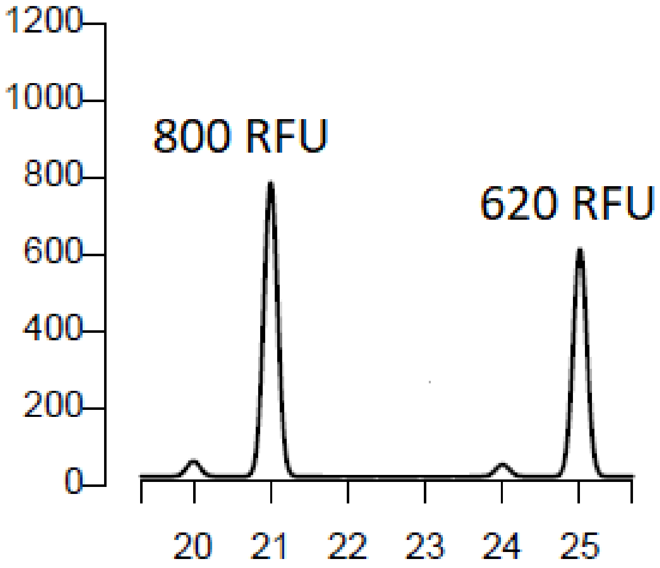
where O is the observed peak height; *smaller* and *larger* refer to the height of the alleles, and HMW and LMW refer to the higher and lower molecular weight allele, respectively.

Modeling Hb

$$\mathrm{Hb}_1 = \frac{O_{\mathrm{HMW}}}{O_{\mathrm{LMW}}}$$

$$= \frac{620}{800} = 0.775$$

$$= \mathrm{Hb}_2$$

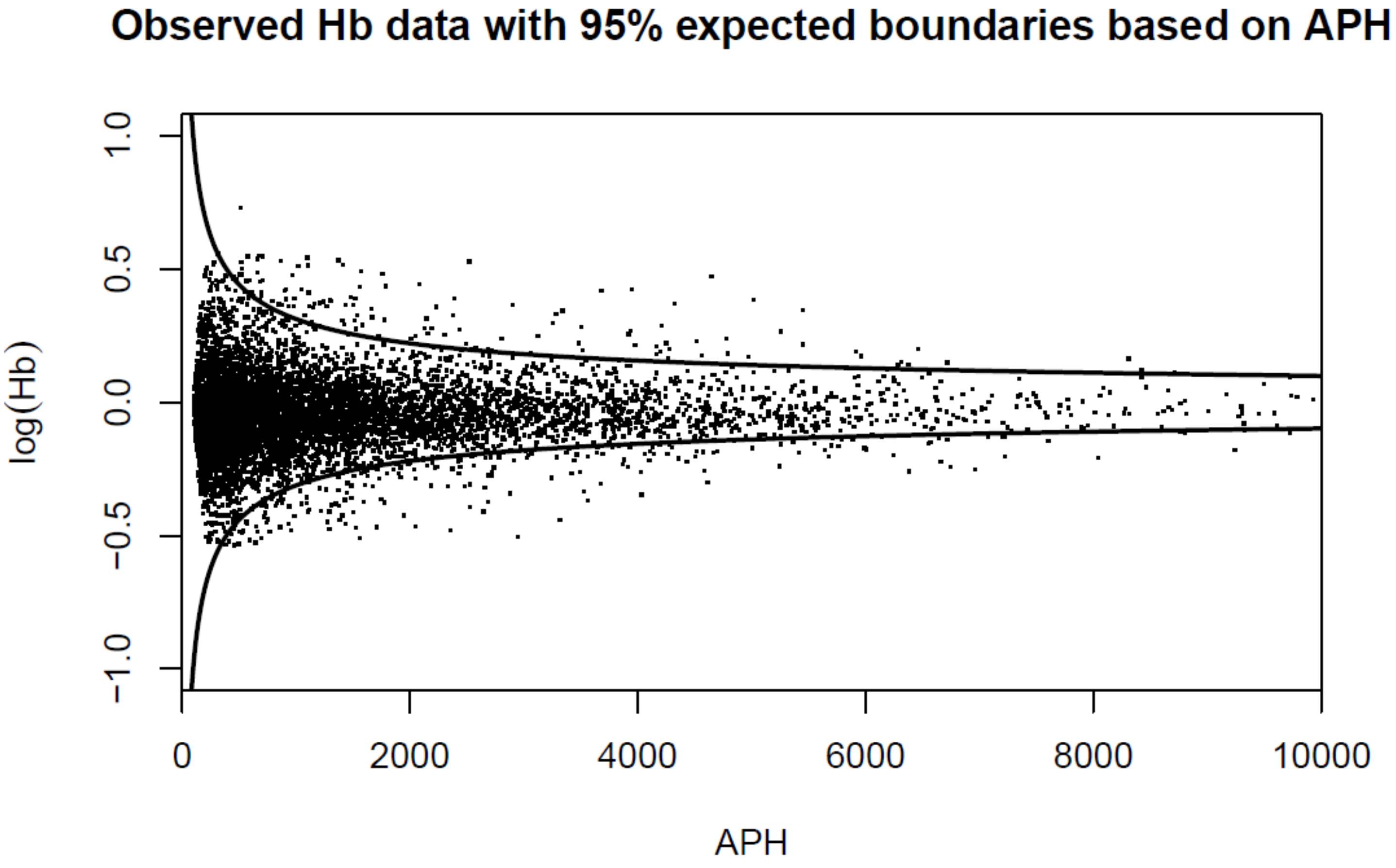


- Hb_1 has the highest information content, because it maintains peak order.

- Hb_2 may be obtained from Hb_1, but not vice versa.

Modeling Hb

The following figure shows Hb rates versus the *average peak height* (APH), which is simply the average of two observed heterozygote alleles at a locus.



Stutter Modeling

Stutter modeling becomes especially important in case of mixtures, when a true (minor) contributor's alleles are approximately the same height as stutter products from the major contributor.

Stutter is typically modeled by a stutter ratio (SR):

$$SR = \frac{O_{a-1}}{O_a},$$

where O_{a-1} refers to the observed peak height of the back stutter of parent peak O_a.

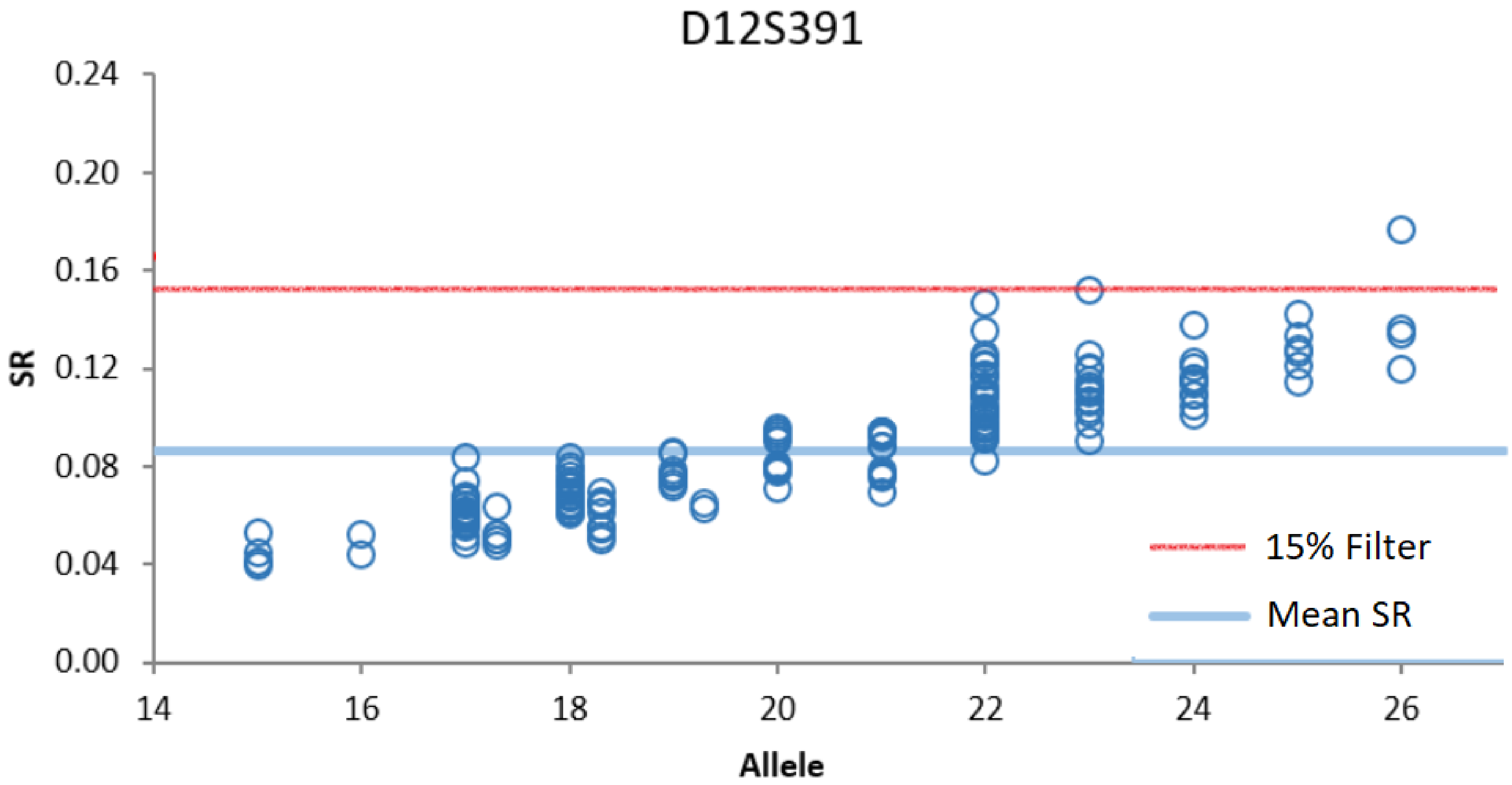
Stutter Modeling

As we've seen earlier, stutter thresholds can be set to help interpret a mixture profile. Locus-specific thresholds account for the variability observed between loci. Traditionally, fixed rates of around 15% are used to remove stutter.

Locus	Stutter Filter (%)
TH01	5
D2S441	9
vWA	11
FGA	11.5
SE33	15
D22S1045	17

However, fixed stutter thresholds have the disadvantage that they do not incorporate the well-known stutter characteristics (such as the correlation with the number of repeats).

Stutter Modeling - Locus Specific Thresholds



Source: Implementation and validation of an improved allele specific stutter filtering method for epg interpretation (Buckleton et al., 2017).

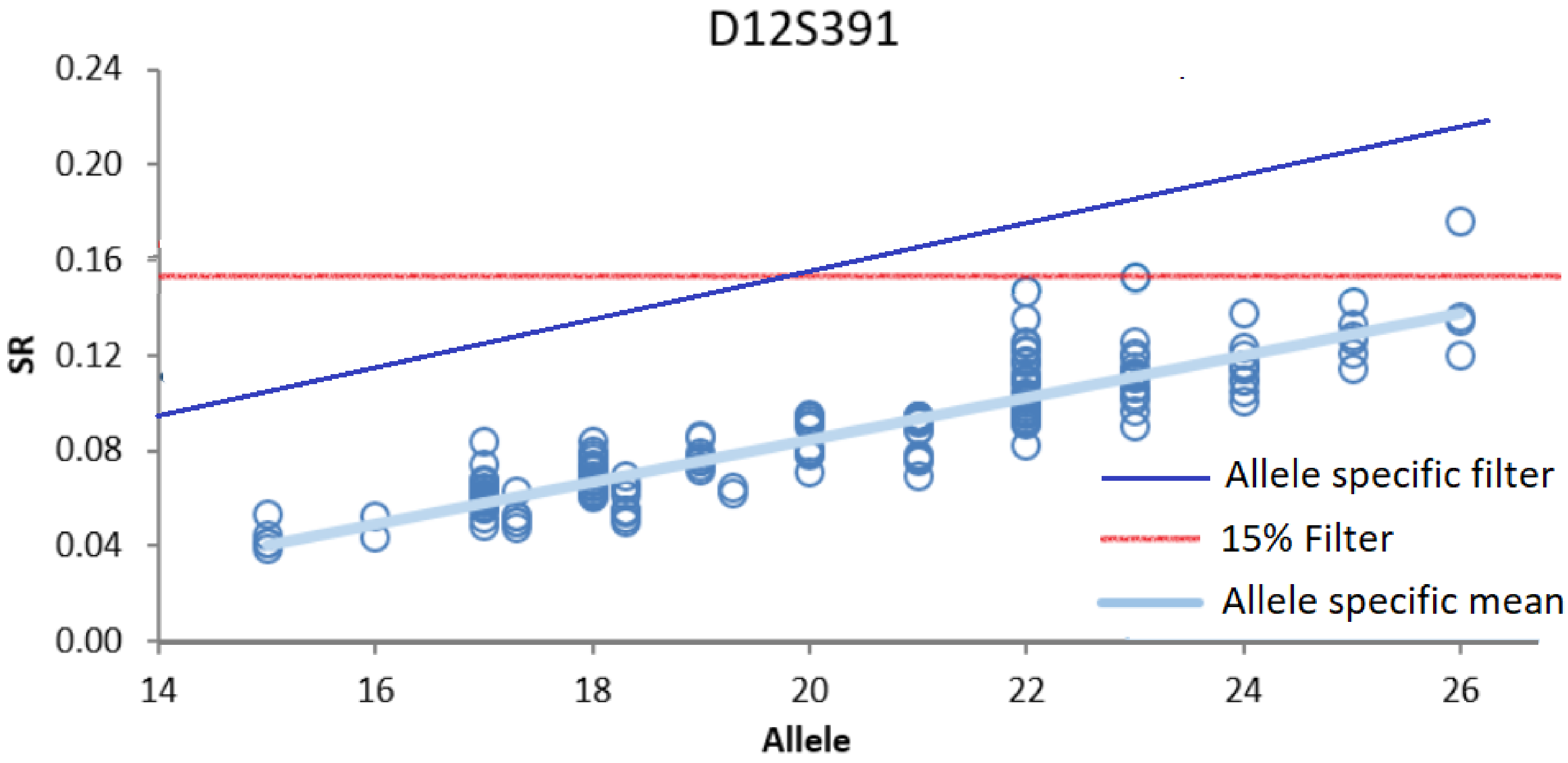
Stutter Modeling - Locus Specific Thresholds

Fixed stutter thresholds lead to over filtering and under filtering:

- **Over filtering**: leads to potential data loss and difficulties in interpretation when true allelic peaks of a minor contributor get filtered.

- **Under filtering**: leads to the possibility that stutter peaks are treated as allelic, and difficulties in determining genotypes for a minor contributor and the number of contributors.

Stutter Modeling - Allele Specific Thresholds



Source: Implementation and validation of an improved allele specific stutter filtering method for epg interpretation (Buckleton et al., 2017).

Stutter Modeling - Thresholds

These observations suggest that stutter thresholds should not only be locus-based, but at a minimum also allele-based. Moreover:

- Thresholds do not account for more complex situations such as composite stutter;
- And still result in a binary decision (i.e. the peak is either ignored or labeled as allelic).

Fully continuous models have the potential to overcome such problems, since there is no need for thresholds within a probabilistic approach.

Stutter Modeling - Allele Model

A simple linear, allele specific, model can be fitted for each locus:

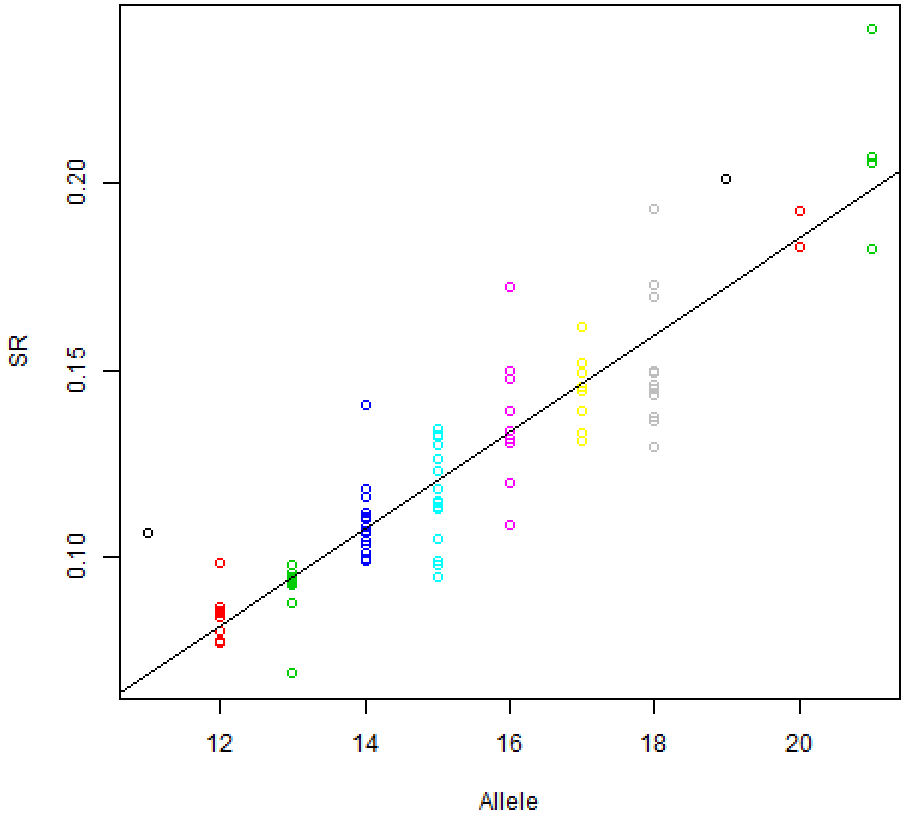
$$SR \sim \text{Allele number} \qquad \Rightarrow \qquad SR = ma + c,$$

with a the allele number, and m and c are constants that can be fitted to the data.

An R-squared measure (R^2) can be used to measure how close the data are fitted to the regression line.

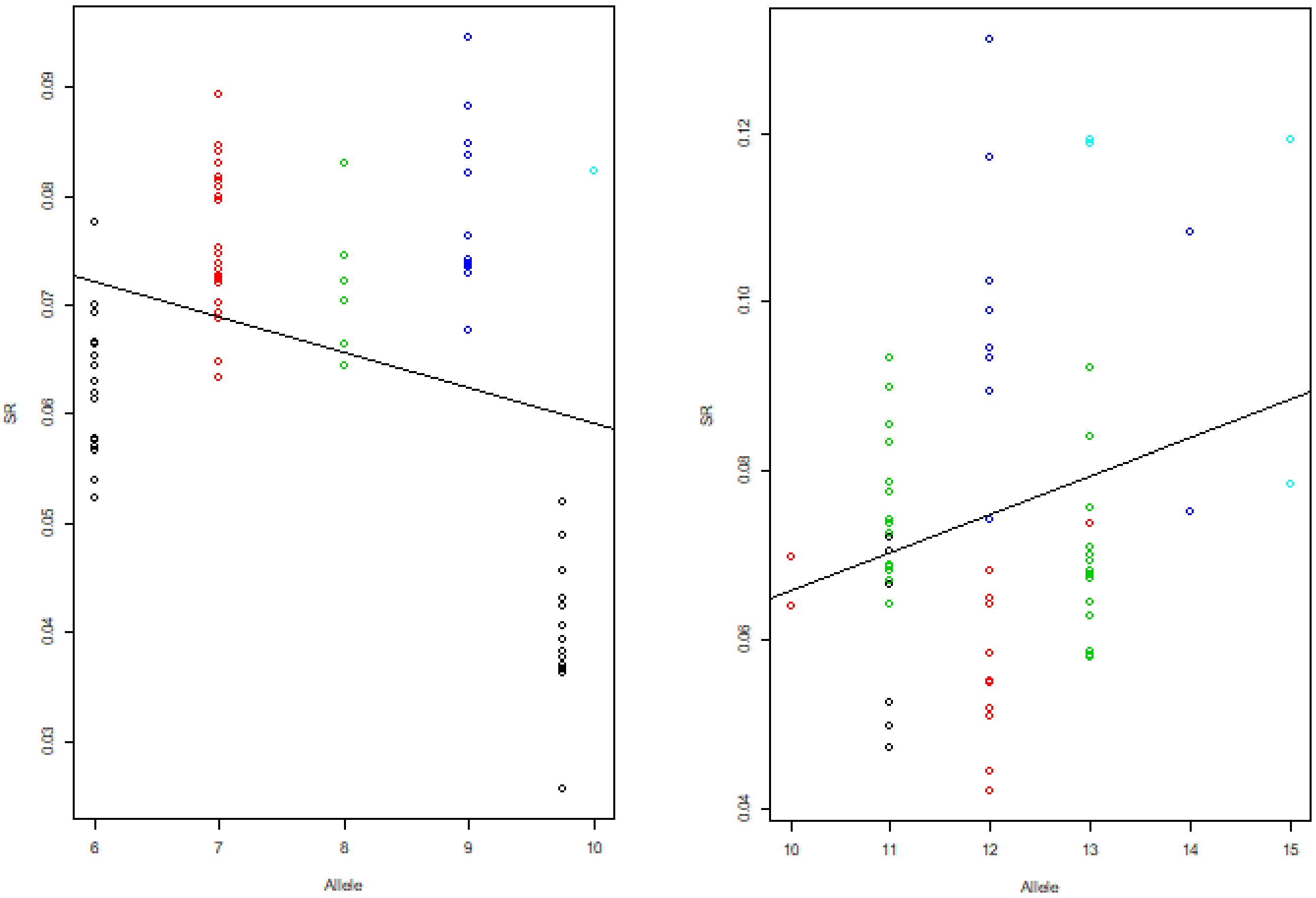
Stutter Modeling - Allele Model

The following figure shows locus D18S51 with a fitted model of $SR = 0.013a - 0.073$ ($R^2 = 82\%$).



Stutter Modeling - Allele Model

But this does not seem to work for all loci:



Locus TH01 and D9S1122.

Stutter Modeling - LUS

These observations suggest that there exists a linear relationship between stutter ratio and the *longest uninterrupted stretch* (LUS).

Repeat motif	Allele	LUS
$[AATG]_6$	6	6
$[AATG]_7$	7	7
$[AATG]_8$	8	8
$[AATG]_9$	9	9
$[AATG]_6ATG[AATG]_3$	9.3	6

Common TH01 allele sequences.

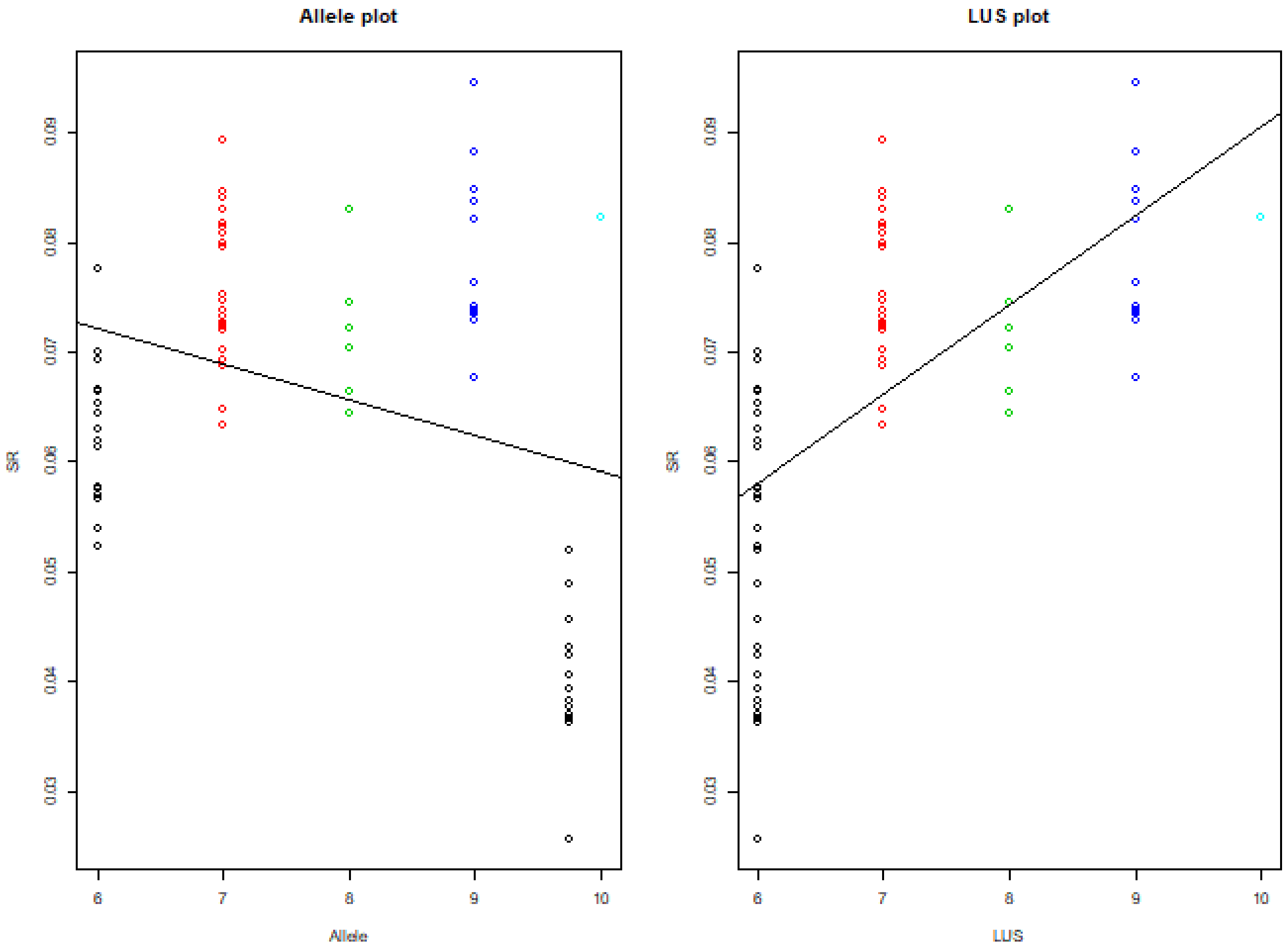
Stutter Modeling - LUS Model

A model based on the LUS can be fitted as follows:

$$SR \sim \text{LUS} \qquad \Rightarrow \qquad SR = ml + c,$$

with l the LUS, and m and c are constants that can be fitted to the data.

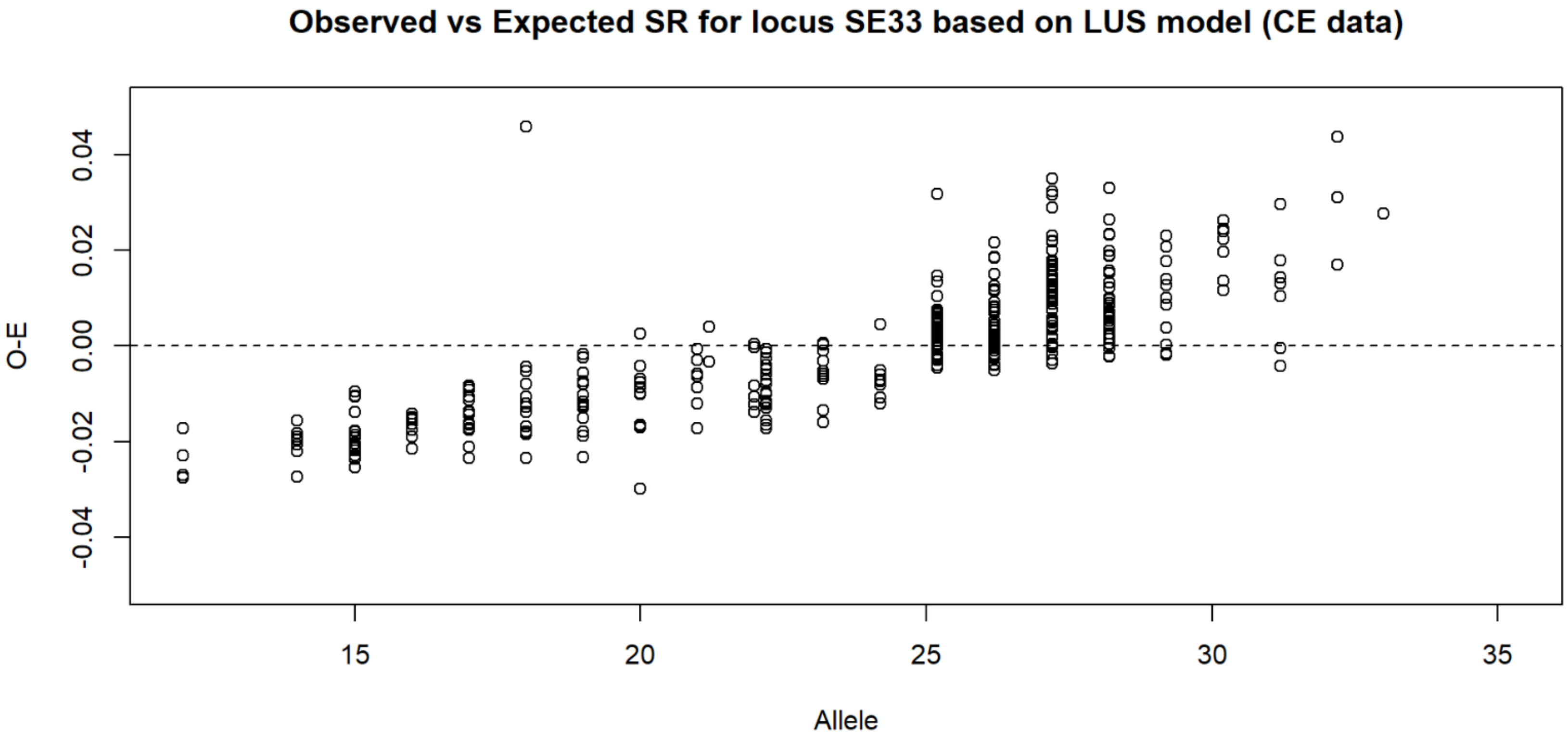
Stutter Modeling - LUS Model



Locus TH01 allele vs. LUS.

Stutter Modeling - LUS Model

What about more complex loci?



Stutter Modeling - AUS

It seems like the LUS still leaves some of the stutter variation unexplained. A multi-sequence model takes into account all uninterrupted stretches (AUS) as potentially contributing to stuttering.

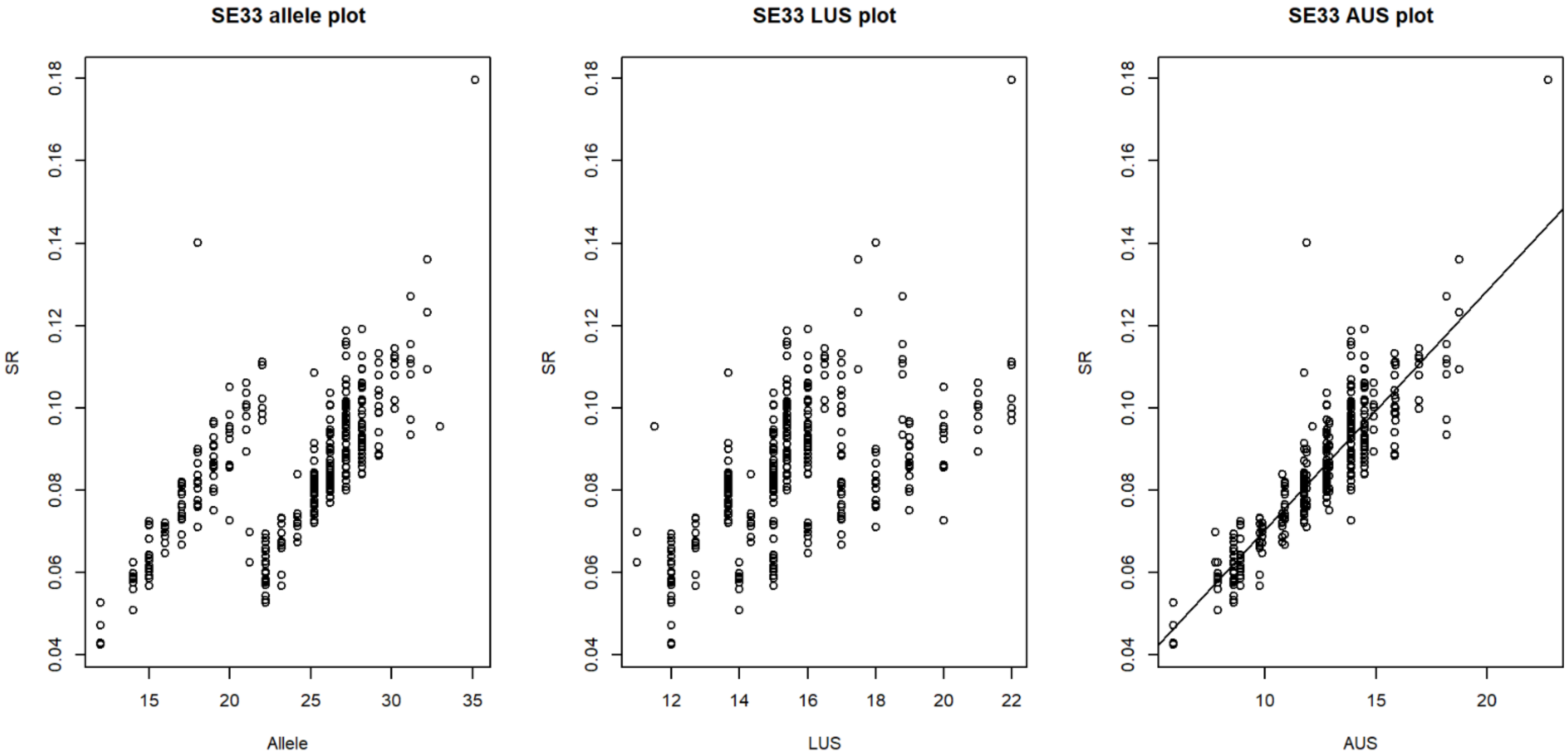
Allele	**Repeat motif**
21.2	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_9$AA AAAG[AAAG]$_{11}$G AAGG[AAAG]$_2$AG
21.2	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{11}$AA AAAG[AAAG]$_9$G AAGG[AAAG]$_2$AG
22	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{22}$G[AAAG]$_3$AG
22.2	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_7$AA AAAG[AAAG]$_{14}$GAAGG[AAAG]$_2$AG
22.2	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_8$[AG]$_5$[AAAG]$_{12}$GAAGG[AAAG]$_2$AG
22.2	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_9$AA AAAG[AAAG]$_{12}$GAAGG[AAAG]$_2$AG

Examples of locus SE33 sequences.

$$SR \sim \text{AUS} \qquad \Rightarrow \qquad SR = m \sum_i \max\left(l_i - x, 0\right) + c,$$

where l_i is the length of sequence i, and m, c and x are constants. The term x is called the lag, and can be interpreted as the number of repeats before stuttering begins.

Stutter Modeling - AUS Model



$$SR = m \sum_i \max(l_i - 6.11, 0) + c$$

Stutter Modeling

- Note that for simple repeats there is no difference between the three approaches:

$$[\text{AATG}]_8 \quad \Rightarrow \quad \text{Allele nr} = \text{LUS} = \text{AUS} = 8$$

- What about other stutter products?

We can model forward stutter as well, and can now use these expectations to decompose peak heights (e.g. for composite stutter or stutter affected heterozygotes).

However, the occurrence of artifacts such as double back and 2bp stutter is likely to be so rare that modeling them statistically can hardly be justified.

Forward Stutter Modeling

Forward stutter can be quantified by a stutter ratio as well (FSR):

$$FSR = \frac{O_{a+1}}{O_a},$$

where O_{a+1} refers to the observed peak height of the forward stutter of parent peak O_a.

Forward stutter is observed less often than back stutter, and peaks are more likely to fall below the limit of detection:

Locus	**Stutter Filter (%)**
TH01	0.06
vWA	0.33
FGA	0.30
D2S441	0.55
SE33	0.59
D10S1248	1.28

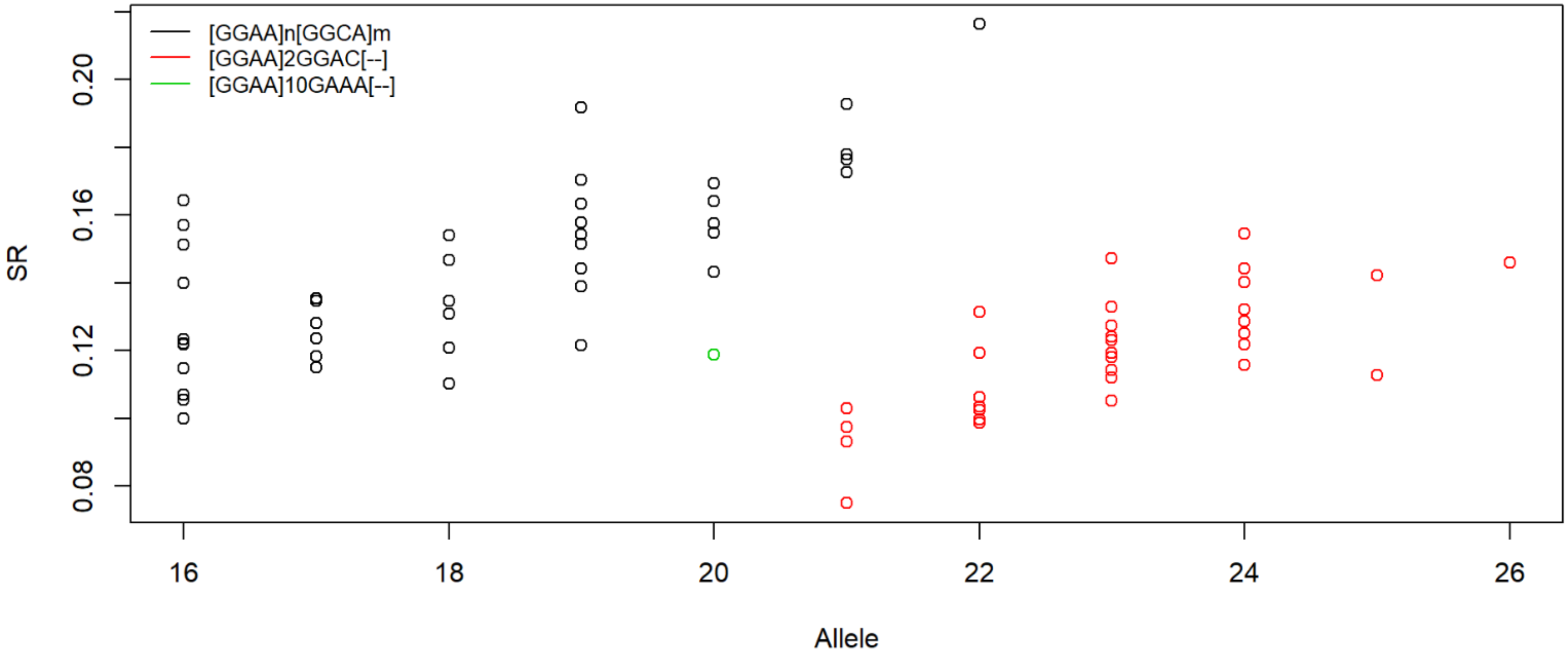
Stutter Modeling - Discussion

How to determine the sequence length for CE data?

Allele	Repeat motif
21.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{9}$AA AAAG[AAAG]$_{11}$G AAGG[AAAG]$_{2}$AG
21.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{11}$AA AAAG[AAAG]$_{9}$G AAGG[AAAG]$_{2}$AG
22	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{22}$G[AAAG]$_{3}$AG
22.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{7}$AA AAAG[AAAG]$_{14}$GAAGG[AAAG]$_{2}$AG
22.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{8}$[AG]$_{5}$[AAAG]$_{12}$GAAGG[AAAG]$_{2}$AG
22.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{9}$AA AAAG[AAAG]$_{12}$GAAGG[AAAG]$_{2}$AG

Stutter Modeling - Discussion

What about variation that is suggested to be attributable to sequence motif?



Stutter ratios for locus D2S1338.

Models fitted based on AUS still left some variability unexplained for some loci.

LR Modeling

The LR can now be assessed by writing the ratio in the form:

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

$$= \frac{\sum_j P(G_C|S_j)P(S_j|H_p)}{\sum_{j'} P(G_C|S_{j'})P(S_{j'}|H_d)}$$

$$= \frac{\sum_j w_j P(S_j|H_p)}{\sum_{j'} w_{j'} P(S_{j'}|H_d)}.$$

The two propositions each define sets of genotypes S, and the weights w describe how well these sets fit our observed data G_C. Under H_p all the genotype sets S_j usually include G_S.

LR Modeling

The full profile weight can be obtained as a product of the weights at each locus:

$$w_j = \prod_l w_j^l.$$

In case of the binary model, the weights are set either as 1 or 0, depending on whether or not the crime scene profile can be explained based on the genotype set under consideration.

		Weights	**Weights**
Donor 1	**Donor 2**	**(Binary)**	**(Continuous)**
$20, 21$	$22, 24$	1	0.05
$20, 22$	$21, 24$	1	0.05
$20, 24$	$21, 22$	1	0.75
$21, 22$	$20, 24$	1	0.05
$21, 24$	$20, 22$	1	0.05
$22, 24$	$20, 21$	1	0.05

Modeling Strategies

Now that a model has been developed, we require information about the input parameters.

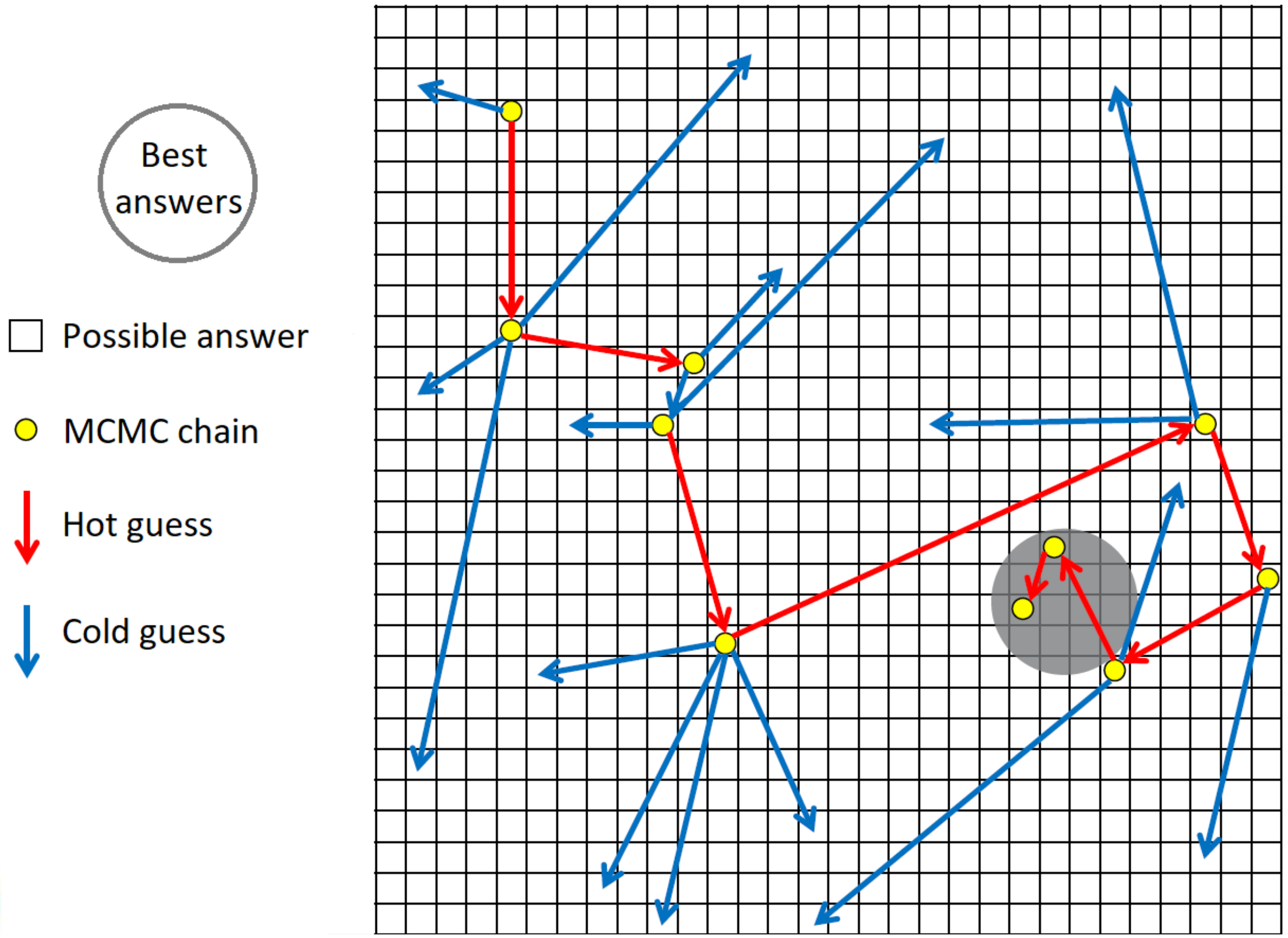
- **Maximization**: Parameters can be chosen that maximize the likelihood of the observations under each hypothesis.

- **Integration**: Rather than knowing the true values of the parameters, we need to know the effect they have on the probability of the observed data.

- **Markov chain Monte Carlo**: Instead of testing every possible combination of parameters, only a small distribution of parameter values and genotype sets will accurately describe the data.

Markov chain Monte Carlo (MCMC)

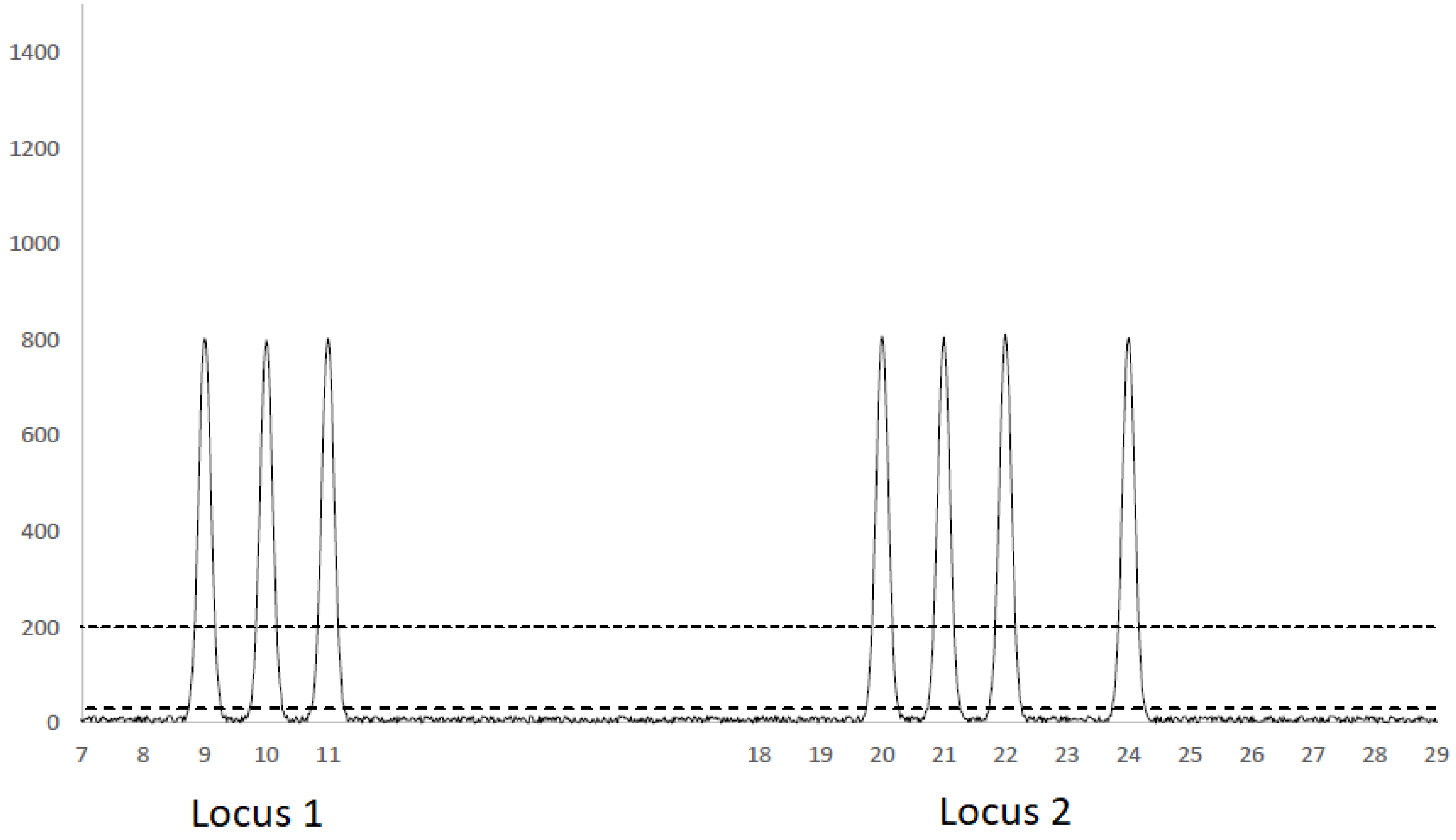
MCMC will start by choosing parameter values at random, eventually leading to more sensible options, until it has reached an equilibrium state.



Expected Peak Heights

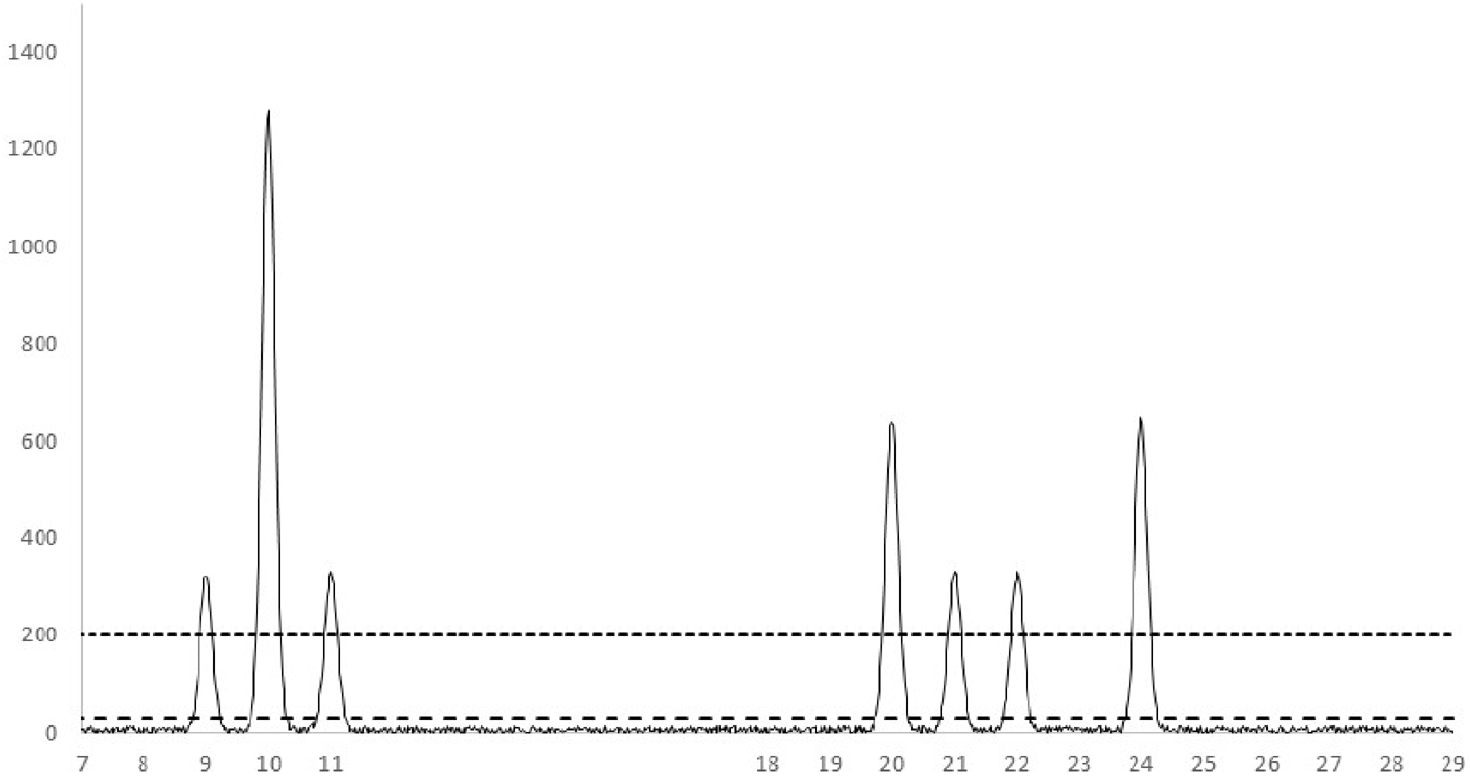
Based on a set of input parameters, an expected profile can be generated.

Step 1: Genotypes are chosen.



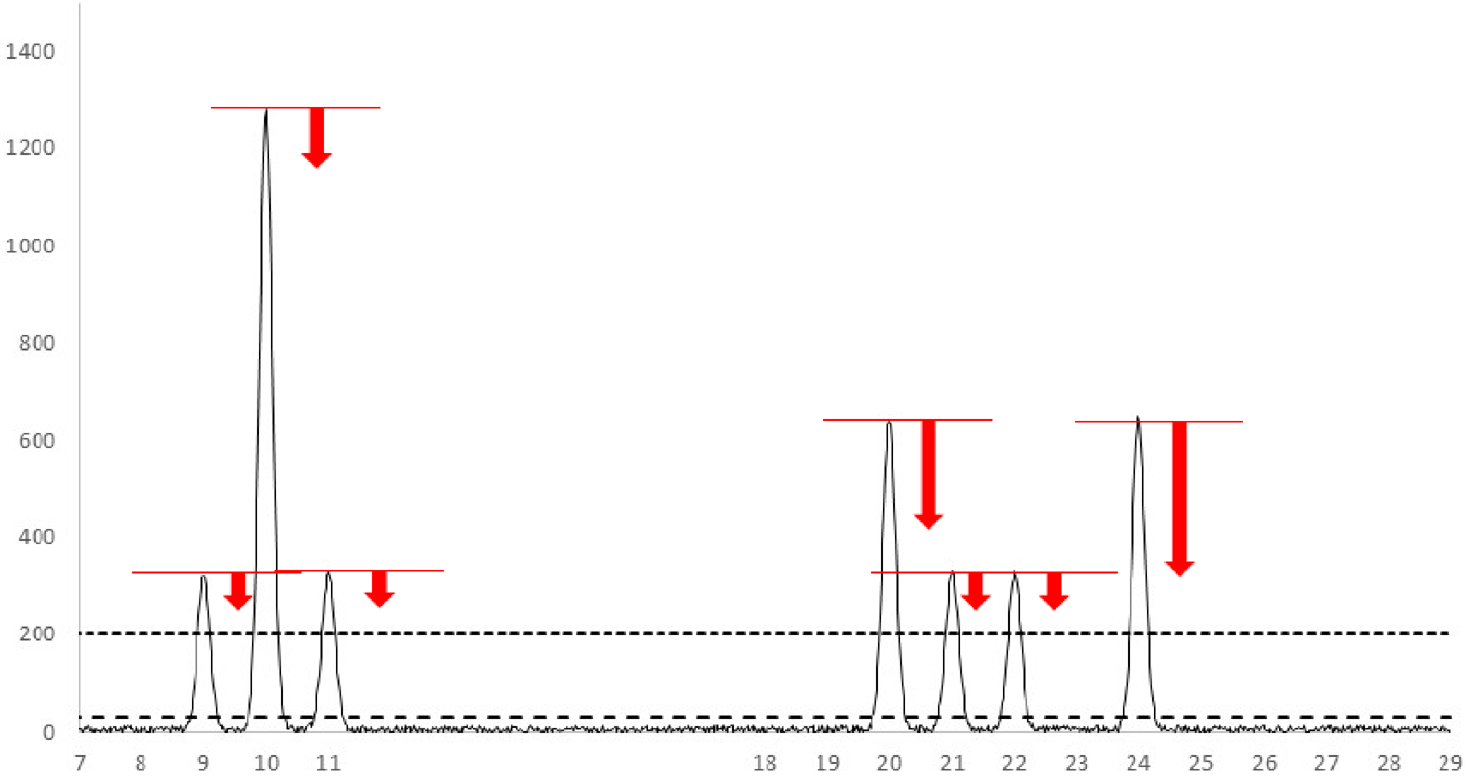
Expected Peak Heights

Step 2: Template amounts per contributor are incorporated.



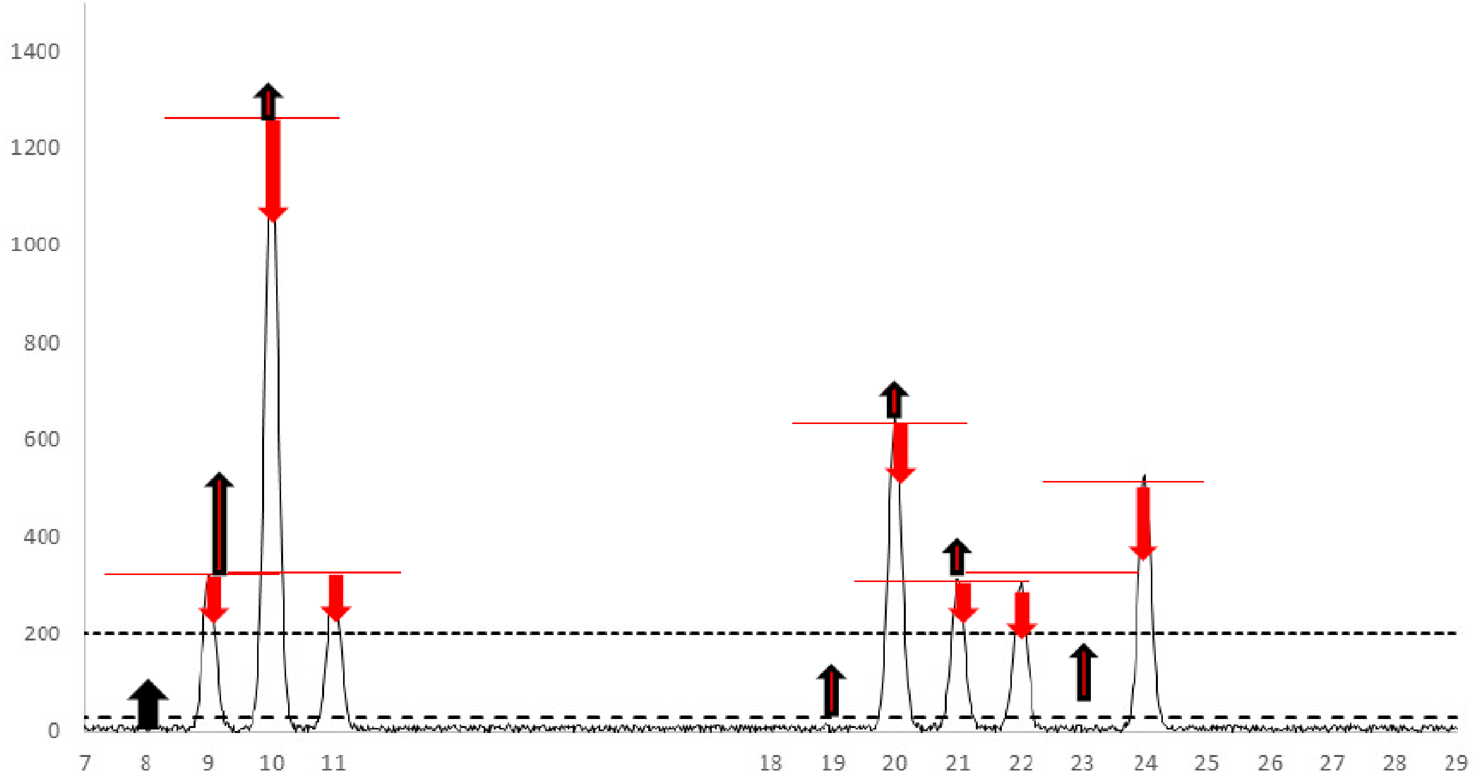
Expected Peak Heights

Step 3: Degradation is taken into account.



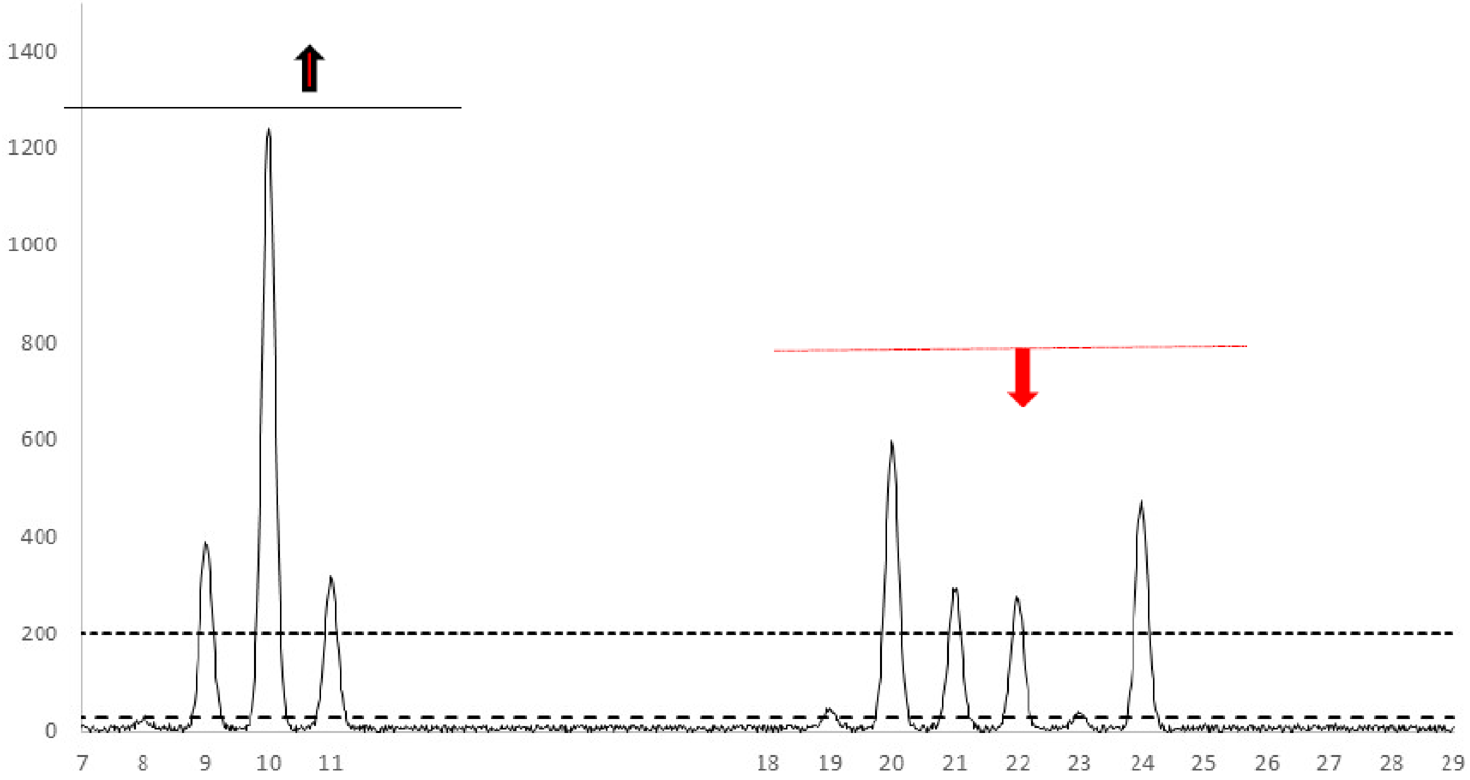
Expected Peak Heights

Step 4: Stutter is taken into account.



Expected Peak Heights

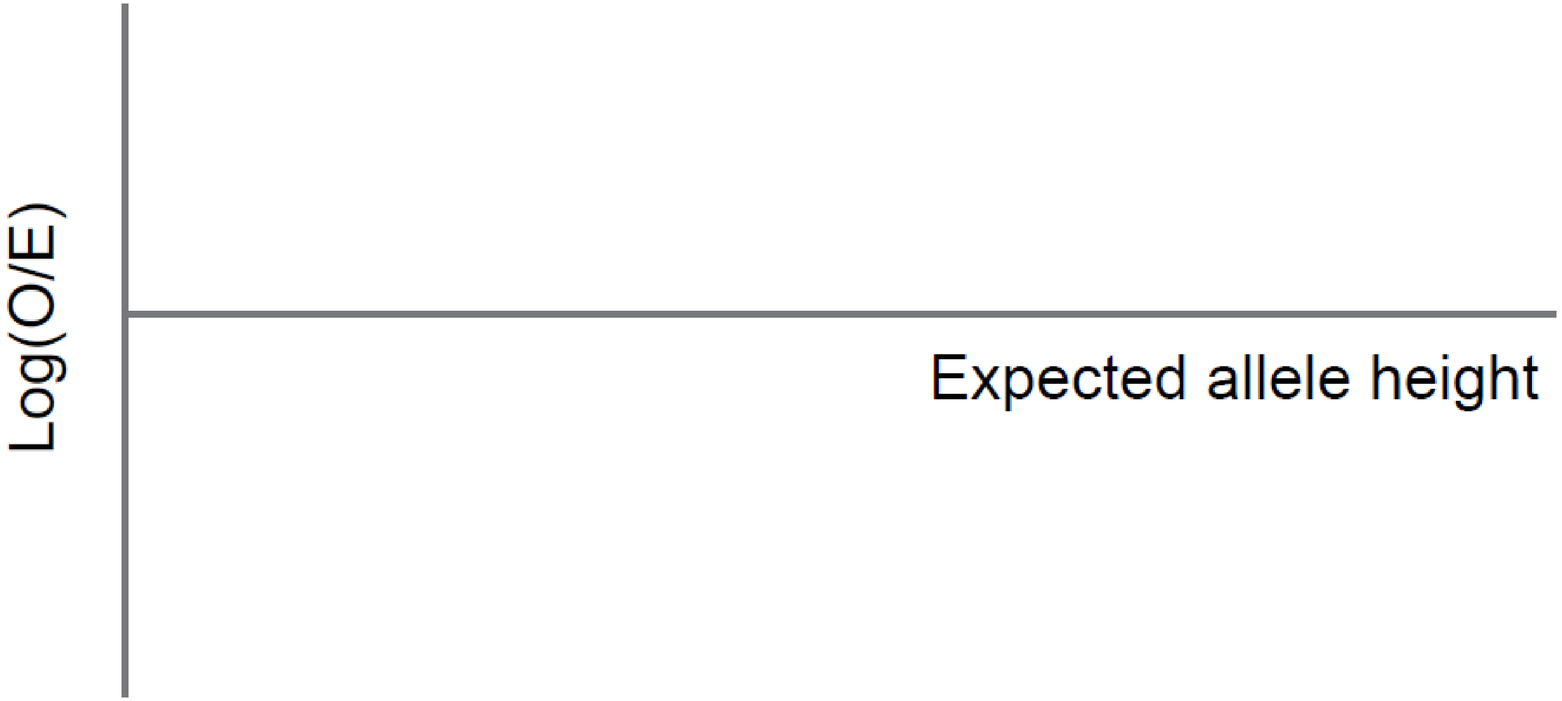
Step 5: Locus specific amplification efficiencies are introduced.



The Perfect Model

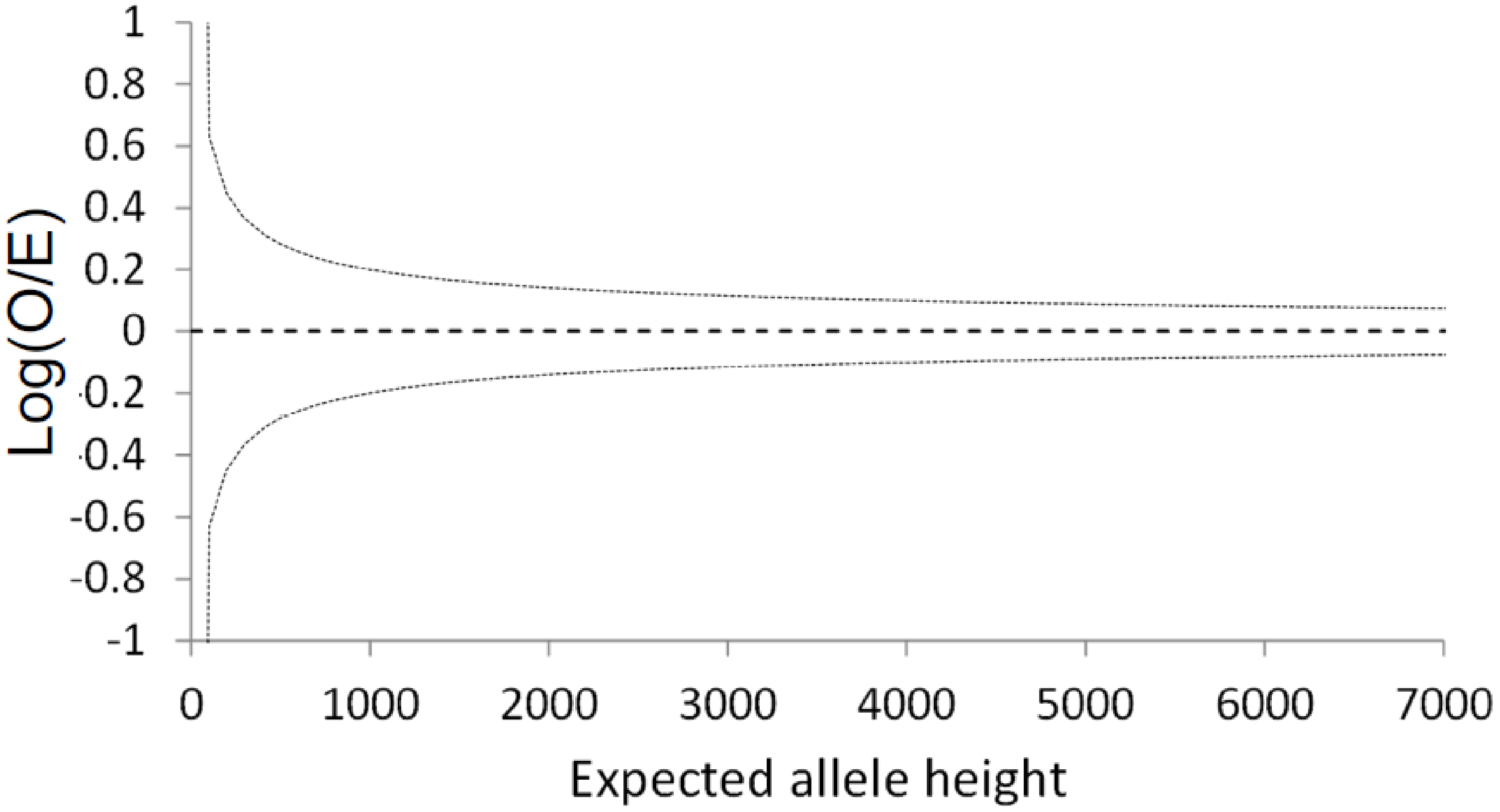
We can now compare our expected profile with the observed STR profile.

What would a perfect model look like?



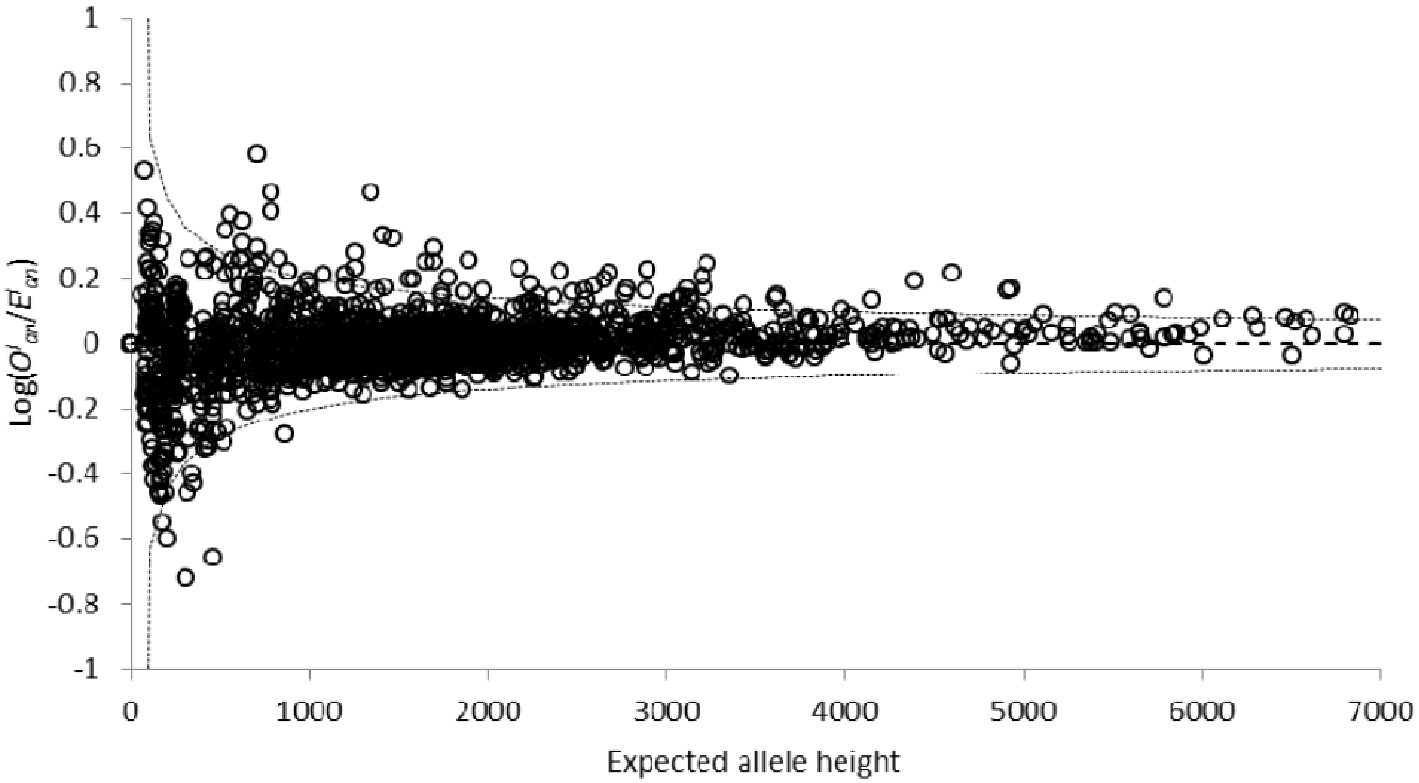
The Perfect Model

Observations show that the relative variance of small peaks is large and the relative variance of large peaks is small. This suggests that the variance is inversely proportional to the expected peak height.



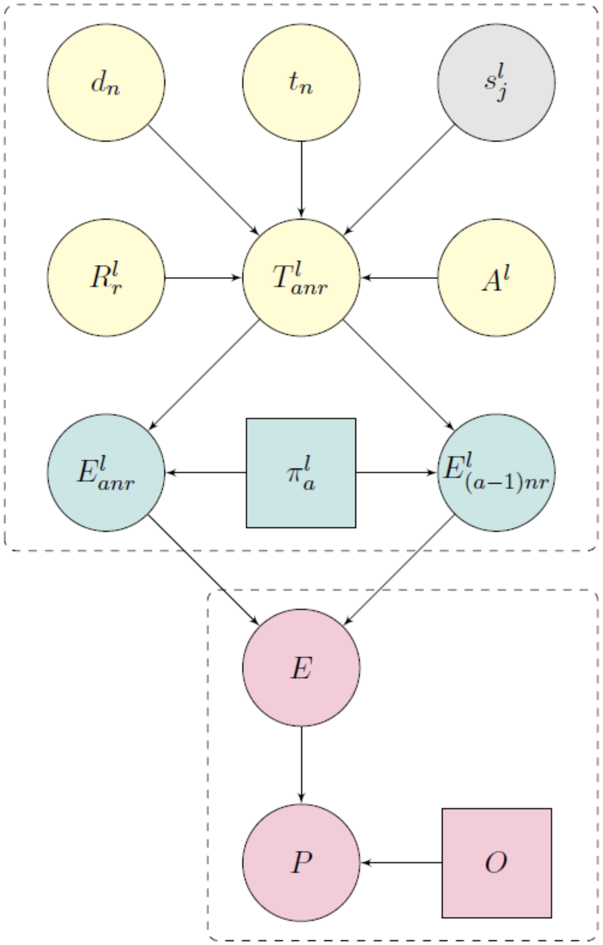
Generating Weights

The weights can now be calculated by considering the ratio of the observed and expected peak heights, assuming the log of this ratio has mean 0 and variance proportional to $1/E$.



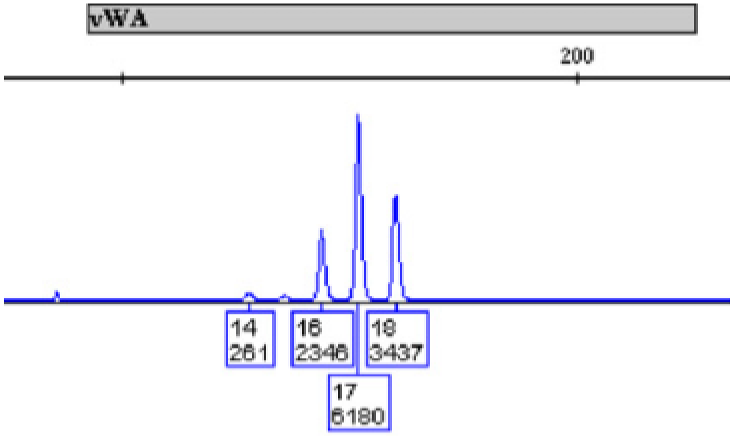
Continuous Model Network

Combining all elements leads to an overall continuous model network:



Worked Example for the Continuous Model

The epg for a 3-person mixture at locus vWA is as follows:



We would like to assess the LR under the hypothesis that:

H_p : $G_C = 17, 18$ and 2U are the source of the sample.

H_d : 3U are the source of the sample.

Source: The interpretation of single source and mixed DNA profiles (Taylor et al., 2013).

Worked Example for the Continuous Model

The LR can now be assessed by writing the ratio in the form:

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

$$= \frac{\sum_j P(G_C|S_j)P(S_j|H_p)}{\sum_{j'} P(G_C|S_{j'})P(S_{j'}|H_d)}$$

$$= \frac{\sum_j w_j P(S_j|H_p)}{\sum_{j'} w_{j'} P(S_{j'}|H_d)}.$$

The two propositions each define sets of genotypes S, and the weights w describe how well these sets fit our observed data G_C. Under H_p all the genotype sets S_j usually include G_S.

Worked Example for the Continuous Model

Suppose the following weights have been established for locus vWA:

Genotype Set	Donor 1	Donor 2	Donor 3	Weight
S_1	$16, 18$	$17, 17$	$14, 14$	0.00045
S_2	$16, 18$	$17, 17$	$14, 15$	0.00017
S_3	$16, 16$	$17, 17$	$14, 16$	0.00008
S_4	$16, 18$	$17, 17$	$14, 17$	0.00002
S_5	$16, 18$	$17, 17$	$14, 18$	0.00054
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_{15}	$16, 17$	$17, 18$	$14, 15$	0.15800
S_{16}	$16, 17$	$17, 18$	$14, 16$	0.28700
S_{17}	$16, 17$	$17, 18$	$14, 17$	0.21000
S_{18}	$16, 17$	$17, 18$	$14, 18$	0.11400
S_{19}	$17, 17$	$17, 18$	$14, 16$	0.00016

The actual reference profiles of the three known contributors are:

Locus	Donor 1	Donor 2	Donor 3
vWA	$16, 17$	$17, 18$	$14, 16$

Source: The interpretation of single source and mixed DNA profiles (Taylor et al., 2013).

Worked Example for the Continuous Model

Under H_p only the genotype sets containing G_C are relevant:

Set	Donor 1	Donor 2	Donor 3	Weight	$P(S_j \mid H_p)$
S_1	$16, 18$	$17, 17$	$14, 14$	0.00045	0
S_2	$16, 18$	$17, 17$	$14, 15$	0.00017	0
S_3	$16, 16$	$17, 17$	$14, 16$	0.00008	0
S_4	$16, 18$	$17, 17$	$14, 17$	0.00002	0
S_5	$16, 18$	$17, 17$	$14, 18$	0.00054	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_{15}	$16, 17$	$17, 18$	$14, 15$	0.15800	$2p_{16}p_{17} \cdot 2p_{14}p_{15}$
S_{16}	$16, 17$	$17, 18$	$14, 16$	0.28700	$2p_{16}p_{17} \cdot 2p_{14}p_{16}$
S_{17}	$16, 17$	$17, 18$	$14, 17$	0.21000	$2p_{16}p_{17} \cdot 2p_{14}p_{17}$
S_{18}	$16, 17$	$17, 18$	$14, 18$	0.11400	$2p_{16}p_{17} \cdot 2p_{14}p_{18}$
S_{19}	$17, 17$	$17, 18$	$14, 16$	0.00016	$p_{16}^2 \cdot 2p_{14}p_{16}$

Note that these calculations can be modified to allow for population substructure.

Multiplication of the weights with the probabilities, and summing over them, results in the numerator of the LR $P(E \mid H_p)$.

Source: The interpretation of single source and mixed DNA profiles (Taylor et al., 2013).

Worked Example for the Continuous Model

Using allele frequencies (in this case from an Australian Caucasian sub-population):

Allele	Frequency
14	0.1146
15	0.1071
16	0.2044
17	0.2726
18	0.2090

yields: $P(E|H_p) = 4.4 \times 10^{-3}$. Similarly, we can calculate the probabilities under H_d, now considering all genotype sets and corresponding donors, we get: $P(E|H_d) = 5.0 \times 10^{-4}$.

Combining this gives us the LR for this specific locus:

$$\text{LR} = \frac{P(E|H_p)}{P(E|H_d)} = \frac{4.4 \times 10^{-3}}{5.0 \times 10^{-4}} = 8.8$$

Source: The interpretation of single source and mixed DNA profiles (Taylor et al., 2013).

Worked Example for the Continuous Model

The overall LR is a combination of all loci (here compared with a binary model):

Locus	**LR**$_B$	**LR**$_C$
D10S1248	0.97	4.69
vWA	1.24	8.21
D16S539	0.45	5.32
D2S1338	2.27	31.22
D8S1179	0.51	7.79
D21S11	0.94	9.98
D18S51	3.85	52.08
D22S1045	4.32	59.18
D19S433	0.92	7.17
TH01	0.97	13.31
FGA	1.39	21.14
D2S441	0.65	4.84
D3S1358	0.93	13.22
D1S1656	5.55	106.14
D12S391	1.42	21.34
SE33	6.23	69.53
Overall LR	356	3.13×10^{19}

Source: The interpretation of single source and mixed DNA profiles (Taylor et al., 2013).

Available Software

The Scientific Working Group on DNA Analysis Methods (SWGDAM) defines probabilistic genotyping as

> *"...the use of biological modeling, statistical theory, computer algorithms, and probability distributions to calculate likelihood ratios (LRs) and/or infer genotypes for the DNA typing results of forensic samples ("forensic DNA typing results")".*

Over the years, several probabilistic genotyping programs have been developed across the globe, ranging from commercial packages to open-source platforms, with the main goal to interpret complex DNA mixtures for CE data.

Available Software

Not all models as published in literature have been translated into software.

Software	Class	Availability	Optimization
LRmix Studio	semi-continuous	open-source	ML
Lab Retriever	semi-continuous	open-source	ML
DNA LiRA	semi-continuous	open-source	Bayes
likeLTD	(semi-)continuous	open-source	ML
STRmix	continuous	commercial	Bayes
TrueAllele	continuous	commercial	Bayes
DNA·VIEW	continuous	commercial	ML
DNAmixtures	continuous	open-source*	ML
EuroForMix	continuous	open-source	ML or Bayes

Available Software - Discussion

There are no ground truths for probabilistic genotyping calculations. Moreover, the 2016 PCAST (President's Council of Advisors on Science and Technology) report stated:

> *"[w]hile likelihood ratios are a mathematically sound concept, their application requires making a set of assumptions about DNA profiles that require empirical testing. Errors in the assumptions can lead to errors in the results"*.

- Under what circumstances have the methods been validated? What are their limitations?

- Commercial software has received criticism regarding their black-box nature. Should source code be made accessible (to the defense)?

Available Software - Discussion

What about the consistency between software programs when they examine the same evidence?

Method	**Sample A**	**Sample B**	**Sample C**
LRmix Studio	1.29	1.85×10^{14}	0.0212
Lab Retriever	1.20	1.89×10^{14}	0.0241
DNA·VIEW	1.09×10^{-14}	4.66×10^{11}	2.24×10^{8}
Combined	Inconclusive	Support to H_p	Inconclusive

Another example can be found in the *People v. Hillary* (NY) case: TrueAllele reported no statistical support for a match (LR < 0), whereas STRmix inculpated the defendant with a likelihood ratio of 360 000. The evidence consisted of an LTDNA sample with an extreme mixture ratio.

Source: An alternative application of the consensus method to DNA typing interpretation for Low Template-DNA mixtures (Garofano et al., 2015).

Section 5:
Population Structure and Relatedness

Human Populations: History and Structure

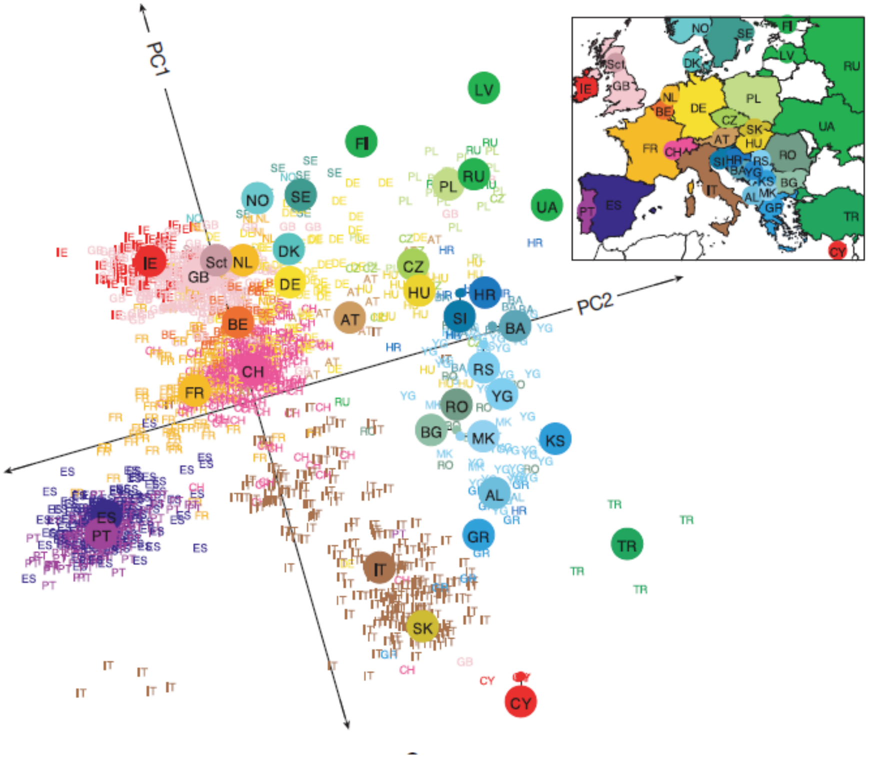
In the paper

> Novembre J, Johnson, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann A, Nelson MB, Stephens M, Bustamante CD. 2008. Genes mirror geography within Europe. Nature 456:98

there is quite dramatic evidence that our genetic profiles contain information about where we live, suggesting that these profiles reflect the history of our populations.

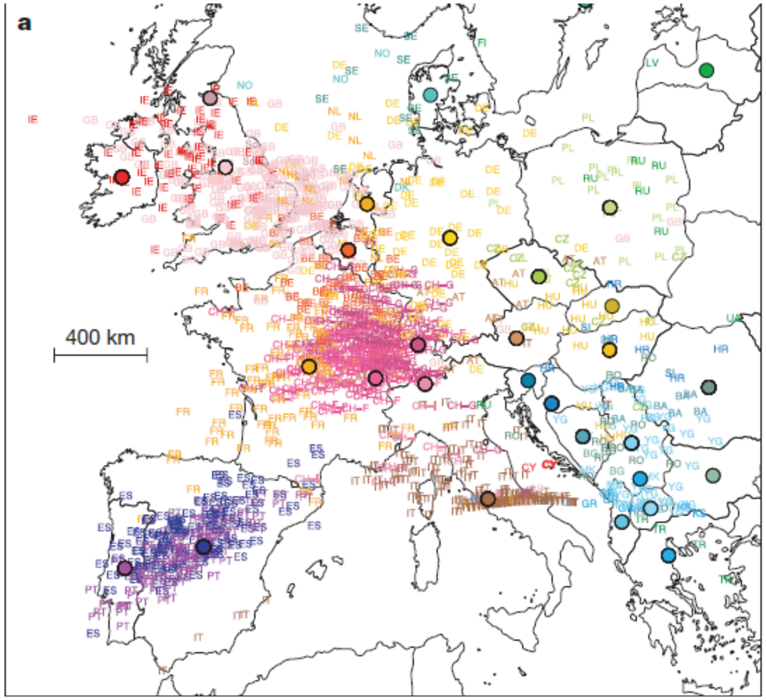
The authors collected "SNP" (single nucleotide polymorphism) data on over people living in Europe. Either the country of origin of the people's grandparents or their own country of birth was known. On the next slide, these geographic locations were used to color the location of each of 1,387 people in "genetic space." Instead of latitude and longitude on a geographic map, their first two principal components were used: these components summarize the 500,000 SNPs typed for each person.

Novembre et al., 2008



Novembre et al., 2008

As a follow-up, the authors took the genetic profile of each person and used it to predict their latitude and longitude, and plotted these on a geographic map. These predicted positions are colored by the country of origin of each person.



Y SNP Data Haplogroups

Another set of SNP data, this time from around the world, is available for the Y chromosome. These data were collected for the 1000 Genomes project (http://www.1000genomes.org/): there are 26 populations:

East Asia: CDX. Chinese Dai in Xishuangbanna; CHB. Han Chinese in Beijing; JPT. Japanese in Tokyo; KHV. Kinh in Ho Chi Minh City; CHS. Southern Han Chinese.

South Asian: BEB. Bengali in Bangladesh; GIH. Gujarati Indian in Houston; ITU. India Telugi in UK; PJL. Punjabi in Lahore; STU. Sri Lankan Tamil in UK.

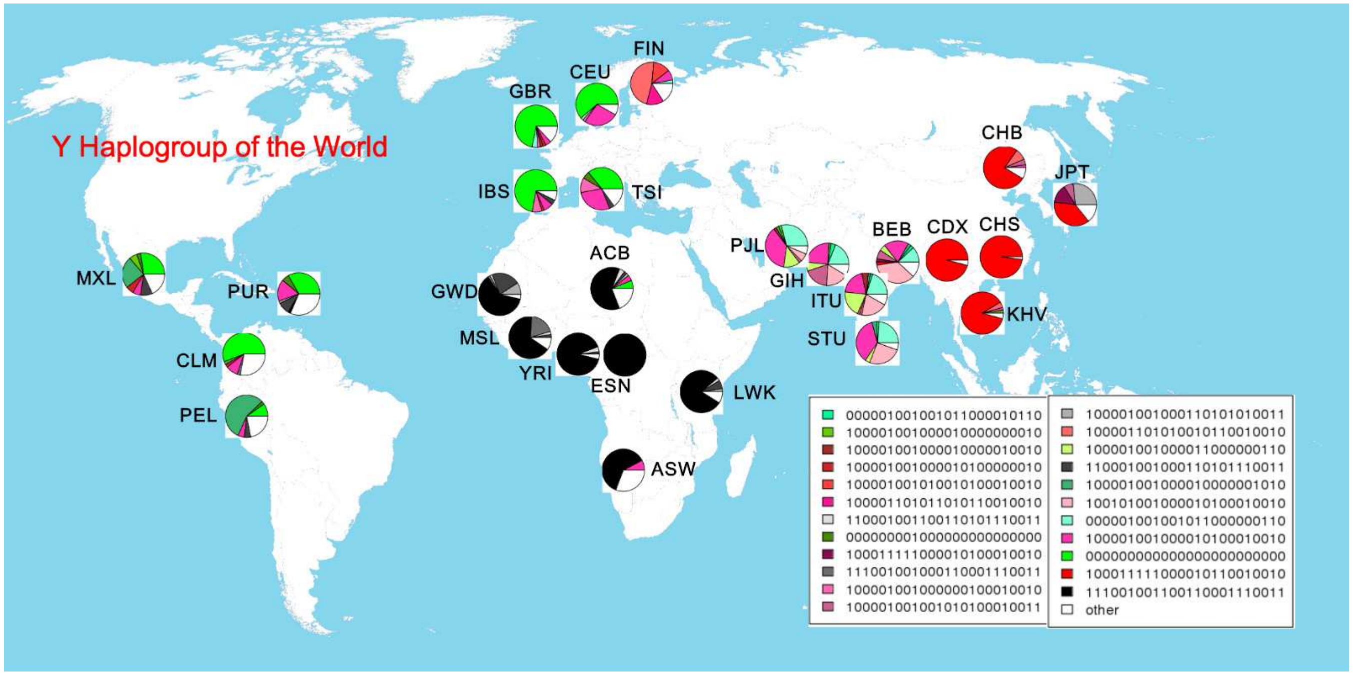
Y SNP Data Haplogroups

African: ASW. African Ancestry in Southwest US; ACB. African Caribbean in Barbados; ESN. Esan in Nigeria; GWD. Gambian in the Gambia; LWK. Luthya in Kenya; MSL. Mende in Sierra Leone; YRI. Yoruba in Nigeria.

European: GBR. British in UK; FIN. Finnish in Finland; IBS. Iberian in Spain; TSI. Toscani in Italy; CEU. Utah residents with European ancestry.

Americas: CLM. Columbian in Medellin; MXL. Mexican in Los Angeles; PEL. Peruvian in Lima, PUR. Puerto Rican in Puerto Rico.

Y SNP Data Haplogroups



Migration History of Early Humans

An interesting video of the migration of early humans is available at:

http://www.bradshawfoundation.com/journey/

Migration Map of Early Humans

https://genographic.nationalgeographic.com/human-journey/

This map summarizes the migration patterns of early humans.

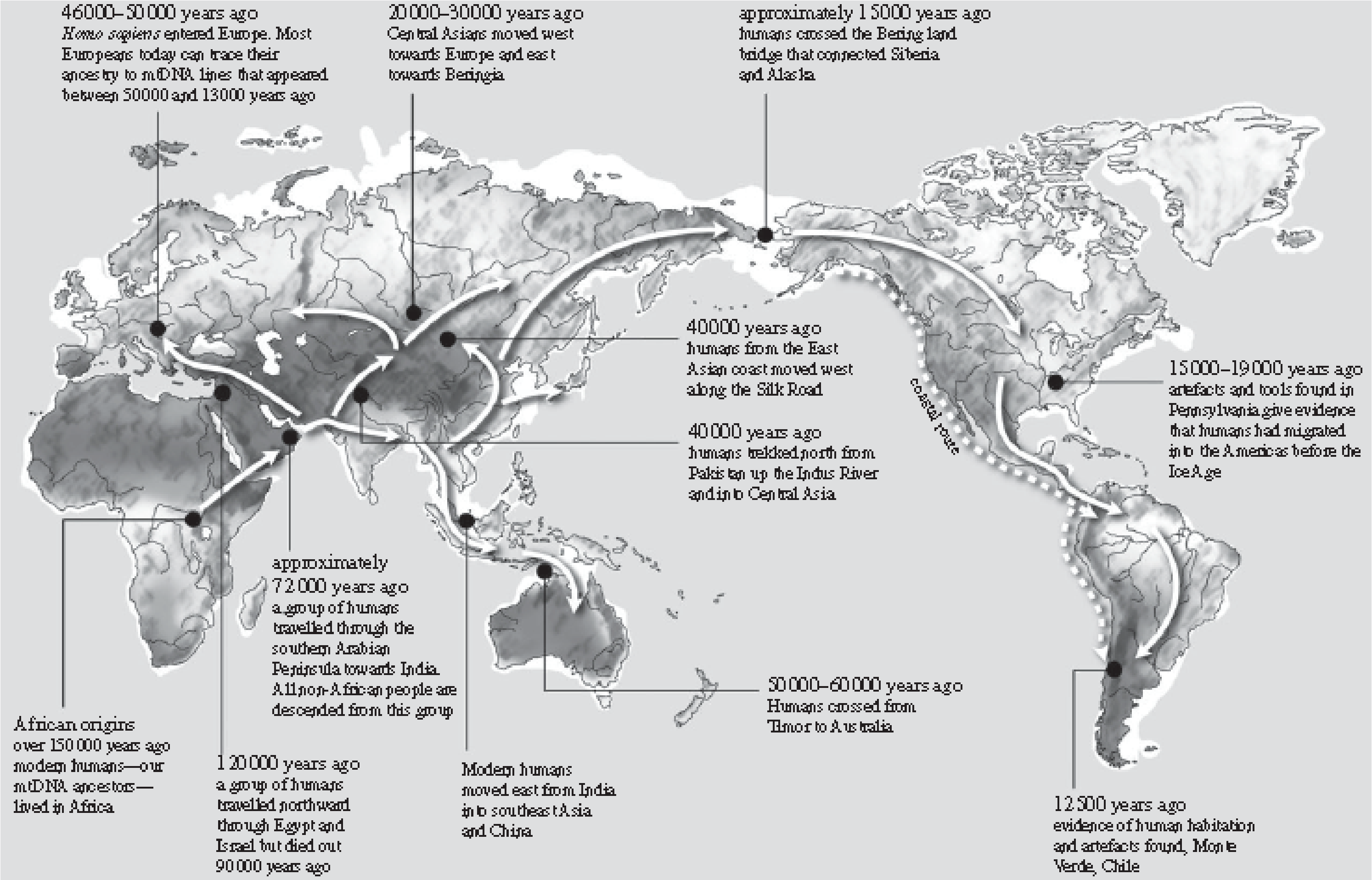
Migration Map of Early Humans

The map on the next slide, based on mitochondrial genetic profiles, is taken from:

Oppenheimer S. 2012. Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. Phil. Trans. R. Soc. B (2012) 367, 770-784 doi:10.1098/rstb.2011.0306.

The first two pages of this paper give a good overview, and they contain this quote: "The finding of a greater genetic diversity within Africa, when compared with outside, is now abundantly supported by many genetic markers; so Africa is the most likely geographic origin for a modern human dispersal."

Migration Map of Early Humans



46000–50000 years ago
Homo sapiens entered Europe. Most Europeans today can trace their ancestry to mtDNA lines that appeared between 50000 and 13000 years ago

20000–30000 years ago
Central Asians moved west towards Europe and east towards Beringia

approximately 15000 years ago
humans crossed the Bering land bridge that connected Siberia and Alaska

40000 years ago
humans from the East Asian coast moved west along the Silk Road

15000–19000 years ago
artefacts and tools found in Pennsylvania give evidence that humans had migrated into the Americas before the Ice Age

40000 years ago
humans trekked north from Pakistan up the Indus River and into Central Asia

approximately 72000 years ago
a group of humans travelled through the southern Arabian Peninsula towards India. All non-African people are descended from this group

50000–60000 years ago
Humans crossed from Timor to Australia

African origins
over 150000 years ago modern humans—our mtDNA ancestors—lived in Africa

120000 years ago
a group of humans travelled northward through Egypt and Israel but died out 90000 years ago

Modern humans moved east from India into southeast Asia and China

12500 years ago
evidence of human habitation and artefacts found, Monte Verde, Chile

Forensic Implications

What does the theory about the spread of modern humans tell us about how to interpret matching profiles?

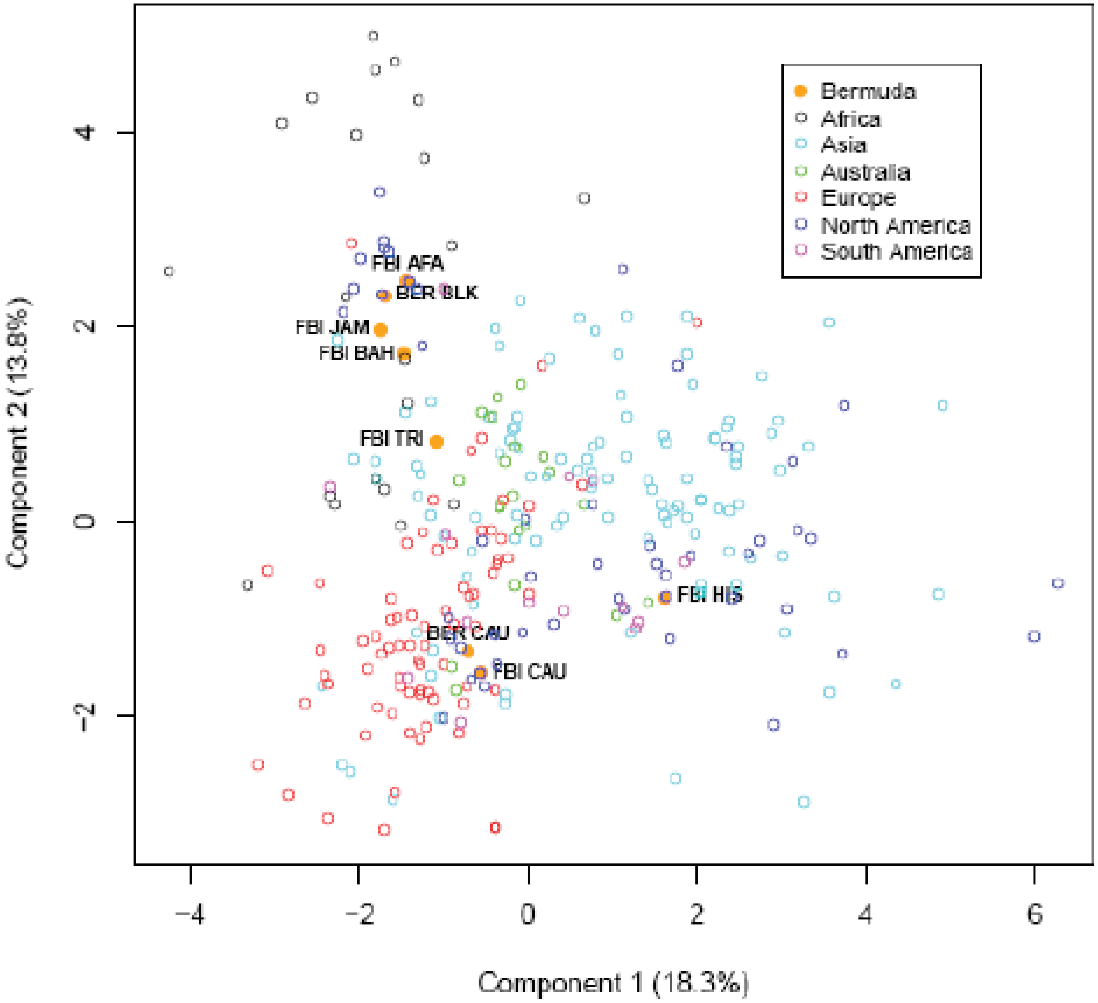
Matching probabilities should be bigger within populations, and more similar among populations that are closer together in time.

Forensic allele frequencies are consistent with the theory of human migration patterns.

Forensic STR PCA Map

A large collection of forensic STR allele frequencies was used to construct the principal component map on the next page. Also shown are some data collected by forensic agencies in the Caribbean, and by the FBI. The Bermuda police has been using FBI data - does this seem to be reasonable?

Forensic STR PCA Map



Genetic Distances

Forensic allele frequencies were collected from 21 populations. The next slides list the populations and show allele frequencies for the Gc marker. This has only three alleles, A, B, C.

The matching proportions within each population, and between each pair of populations, were calculated. These allow distances ("theta" or β) to be calculated for each pair of populations, say 1 and 2: $\hat{\beta}_{12} = ([\tilde{M}_1 + \tilde{M}_2]/2 - \tilde{M}_{12})/(1 - \tilde{M}_{12})$.

$\tilde{M}_1$: two alleles taken randomly from population 1 are the same type.
$\tilde{M}_1$: two alleles taken randomly from population 1 are the same type.
$\tilde{M}_{12}$: an allele taken randomly from population 1 matches an allele taken randomly from population 2.

Published Gc frequencies

Symbol	Description	Symbol	Description
AFA	FBI African-American	IT4	Italian
AL1	North Slope Alaskan	KOR	Korean
AL2	Bethel-Wade Alaskan	NAV	Navajo
ARB	Arabic	NBA	North Bavarian
CAU	FBI Caucasian	PBL	Pueblo
CBA	Coimbran	SEH	FBI Southeastern Hispanic
DUT	Dutch Caucasian	SOU	Sioux
GAL	Galician	SPN	Spanish
HN1	Hungarian	SWH	FBI Southwestern Hispanic
HN2	Hungarian	SWI	Swiss Caucasian
IT2	Italian		

Gc allele frequencies

Popn.	Sample size	A	B	C	Popn.	Sample size	A	B	C
AFA	145	.338	.237	.423	IT4	200	.302	.163	.535
AL1	96	.177	.489	.334	KOR	116	.310	.422	.267
AL2	112	.236	.451	.313	NAV	81	.105	.240	.654
ARB	94	.133	.441	.425	NBA	150	.133	.383	.484
CAU	148	.114	.456	.429	PBL	103	.102	.374	.524
CBA	119	.159	.533	.306	SEH	94	.165	.447	.389
DUT	155	.106	.422	.471	SOU	64	.055	.422	.524
GAL	143	.140	.448	.413	SPN	132	.118	.474	.409
HN1	345	.106	.457	.438	SWH	96	.156	.437	.407
HN2	163	.097	.448	.454	SWI	100	.135	.465	.400
IT2	374	.139	.454	.408					

Distances based on Gc

	AFA	AL1	AL2	ARB	CAU	CBA	DUT	GAL	HN1	HN2
AL1	.201									
AL2	.163	.000								
ARB	.224	.002	.016							
CAU	.303	.020	.046	.008						
CBA	.309	.017	.034	.022	.009					
DUT	.341	.039	.070	.021	.000	.017				
GAL	.295	.015	.037	.007	.000	.004	.002			
HN1	.339	.040	.072	.025	.001	.013	.000	.002		
HN2	.348	.041	.073	.024	.000	.016	.000	.003	.000	
IT2	.304	.023	.048	.015	.000	.004	.002	.000	.001	.002

Distances based on Gc

	AFA	AL1	AL2	ARB	CAU	CBA	DUT	GAL	HN1	HN2
IT4	.088	.029	.022	.032	.085	.098	.111	.081	.120	.117
KOR	.074	.051	.026	.082	.139	.122	.175	.128	.179	.179
NAV	.242	.060	.080	.028	.054	.103	.063	.061	.075	.070
NBA	.278	.017	.041	.002	.000	.018	.004	.001	.007	.006
PBL	.178	.033	.044	.015	.051	.085	.067	.053	.077	.073
SEH	.254	.001	.015	.000	.002	.005	.014	.000	.014	.015
SOU	.294	.035	.062	.008	.010	.046	.012	.015	.020	.016
SPN	.315	.022	.048	.012	.000	.005	.000	.000	.000	.000
SWH	.269	.004	.022	.000	.000	.004	.008	.000	.009	.009
SWI	.298	.013	.035	.007	.000	.002	.002	.000	.002	.003

Distances based on Gc

	IT2	IT4	KOR	NAV	NBA	PBL	SEH	SOU	SPN	SWH
IT4	.098									
KOR	.145	.026								
NAV	.072	.048	.143							
NBA	.005	.067	.127	.034						
PBL	.066	.016	.088	.003	.032					
SEH	.004	.052	.089	.054	.003	.038				
SOU	.021	.067	.148	.011	.001	.021	.019			
SPN	.000	.093	.144	.066	.002	.061	.003	.016		
SWH	.001	.060	.102	.053	.000	.040	.000	.014	.000	
SWI	.000	.079	.125	.062	.001	.054	.000	.016	.000	.000

Clustering populations

Populations can be clustered on the basis of the genetic distances between them. For short-term evolution (among human populations) the simple UPGMA method performs satisfactorily. The closest pair of populations are clustered, and then distances recomputed from each other population to this cluster. Then the process continues.

Look at four of the populations:

	AFA	CAU	SEH	NAV
AFA	–			
CAU	0.303	–		
SEH	0.254	0.002	–	
NAV	0.242	0.054	0.054	–

Clustering populations

The closest pair is CAU/SEH. Cluster them, and compute distances from the other two to this cluster:

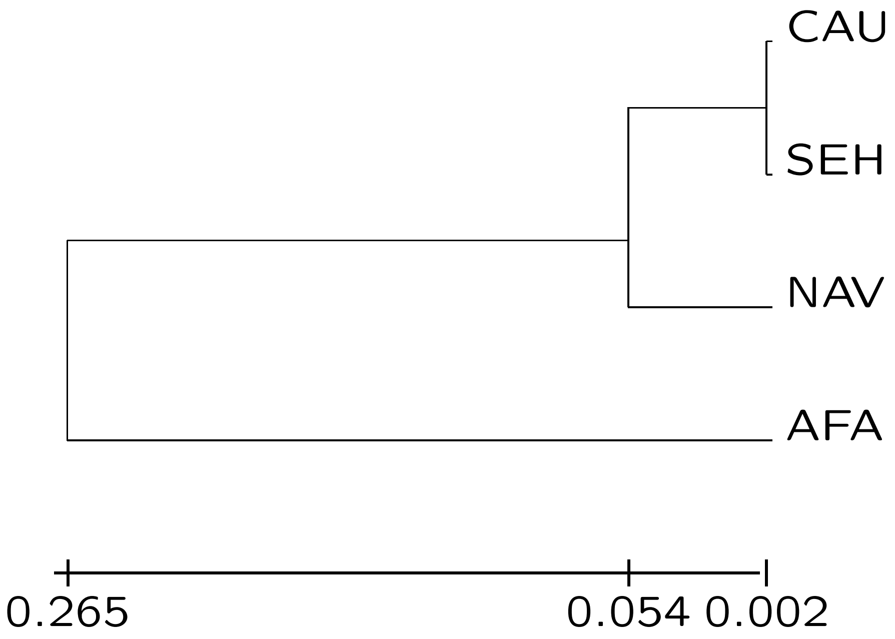
AFA	distance = (0.303+0.254)/2 = 0.278
NAV	distance = (0.054+0.054)/2 = 0.054

The new distance matrix is

	AFA	CAU/SEH	NAV
AFA	–		
CAU/SEH	0.278	–	
NAV	0.242	0.054	–

and the next shortest distance is between NAV and CAU/SEH.

Gc UPGMA Dendrogram



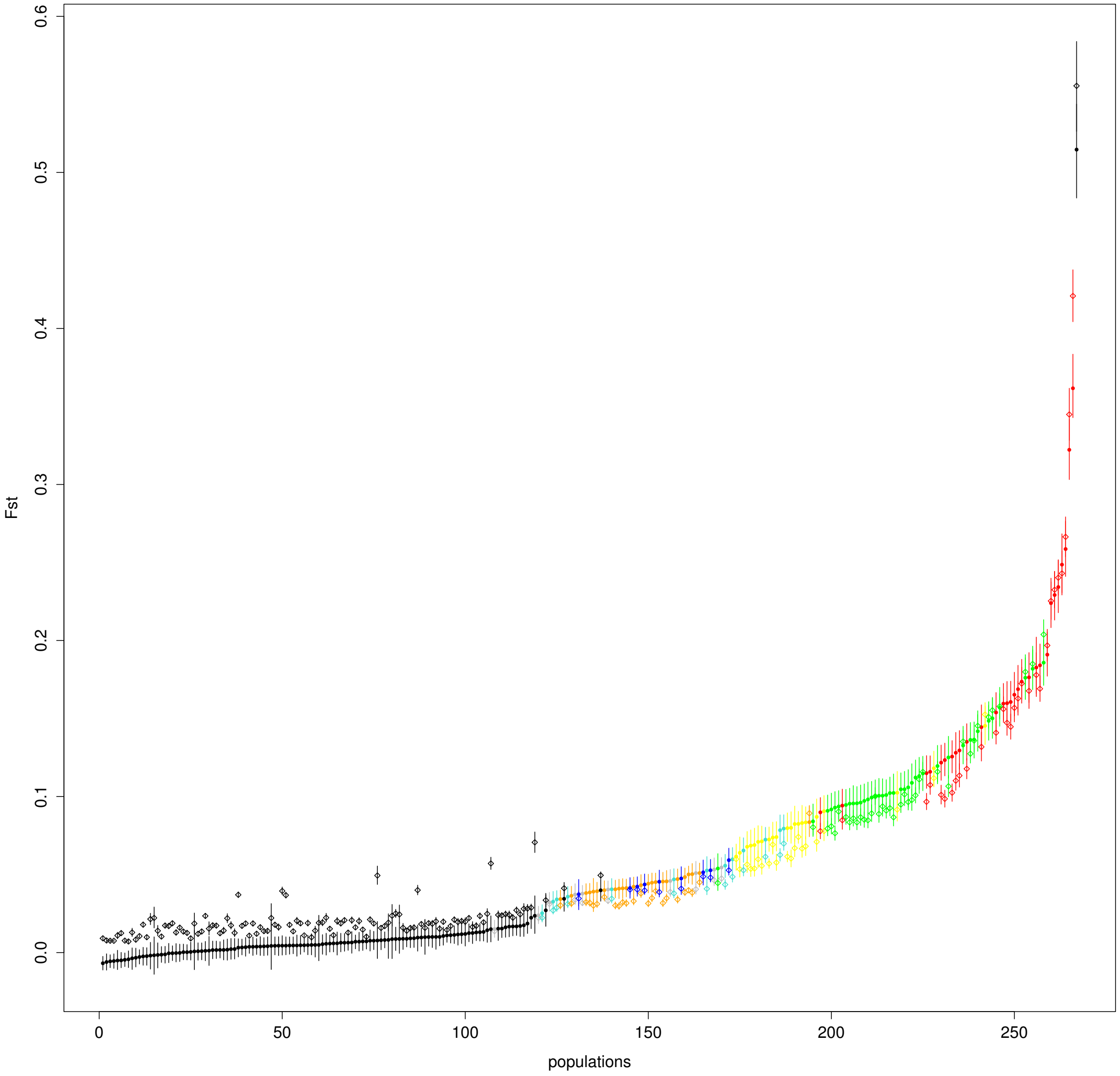
Worldwide Survey of STR Data

Published allele frequencies for 24 STR loci were obtained for 446 populations. For each population i, the within-population matching proportion $\tilde{M}_i$ was calculated. Also the average $\tilde{M}_B$ of all the between-population matching proportions. The "θ" for each population is calculated as $\hat{\beta}_i = (\tilde{M}_i - \tilde{M}_B)/(1 - \tilde{M}_B)$. These are shown on the next slide, ranked from smallest to largest and colored by continent.

Africa: black; America: red; South Asia: orange; East Asia: yellow; Europe: blue; Latino: turquoise; Middle East: grey; Oceania: green.

Buckleton JS, Curran JM, Goudet J, Taylor D, Thiery A, Weir BS. 2016. Forensic Science International: Genetics 23:91-100.

Worldwide Survey of STR Data



Match Probabilities

The β estimates for population structure provide numerical values to substitute for θ into the Balding-Nichols match probabilities.

For AA homozygotes:

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

and for AB heterozygotes

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

These match probabilities are greater than the profile probabilities $\Pr(AA), \Pr(AB)$.

Balding DJ, Nichols RA. 1994. Forensic Science International 64:125-140.

Partial Matching

For autosomal markers, two profiles may be:

Match:	AA, AA or AB, AB
Partially Match:	AA, AB or AB, AC
Mismatch:	AA, BB or AA, BC or AB, CD

How likely are each of these?

Database Matching

If every profile in a database is compared to every other profile, each pair can be characterized as matching, partially matching or mismatching without regard to the particular alleles. We find the probabilities of these events by adding over all allele types.

The probability P_2 that two profiles match (at two alleles) is

$$
\begin{aligned}
P_2 &= \sum_A \mathrm{Pr}(AA, AA) + \sum_{A \neq B} \mathrm{Pr}(AB, AB) \\
&= \frac{\sum_A p_A[\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A][3\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)} \\
&\quad + \frac{2\sum_{A \neq B}[\theta + (1-\theta)p_A][\theta + (1-\theta)p_B]}{(1+\theta)(1+2\theta)}
\end{aligned}
$$

Database Matching

This approach leads to probabilities P_2, P_1, P_0 of matching at 2,1,0 alleles:

$$\begin{aligned} P_2 &= \frac{1}{D}[6\theta^3 + \theta^2(1-\theta)(2+9S_2) + 2\theta(1-\theta)^2(2S_2+S_3) \\ &\quad + (1-\theta)^3(2S_2^2 - S_4)] \\[2ex] P_1 &= \frac{1}{D}[8\theta^2(1-\theta)(1-S_2) + 4\theta(1-\theta)^2(1-S_3) \\ &\quad + 4(1-\theta)^3(S_2 - S_3 - S_2^2 + S_4)] \\[2ex] P_0 &= \frac{1}{D}[\theta^2(1-\theta)(1-S_2) + 2\theta(1-\theta)^2(1-2S_2+S_3) \\ &\quad + (1-\theta)^3(1-4S_2+4S_3+2S_2^2-3S_4)] \end{aligned}$$

where $D = (1+\theta)(1+2\theta)$, $S_2 = \sum_A p_A^2$, $S_3 = \sum_A p_A^3$, $S_4 = \sum_A p_A^4$. For any value of θ we can predict the matching, partially matching and mismatching proportions in a database.

FBI Caucasian Matching Counts

One-locus matches in FBI Caucasian data (18,721 pairs of 13-locus profiles).

		θ				
Locus	Observed	.000	.001	.005	.010	.030
D3S1358	.077	.075	.075	.077	.079	.089
vWA	.063	.062	.063	.065	.067	.077
FGA	.036	.036	.036	.038	.040	.048
D8S1179	.063	.067	.068	.070	.072	.083
D21S11	.036	.038	.038	.040	.042	.051
D18S51	.027	.028	.029	.030	.032	.040
D5S818	.163	.158	.159	.161	.164	.175
D13S317	.076	.085	.085	.088	.090	.101
D7S820	.062	.065	.066	.068	.070	.080
CSF1PO	.122	.118	.119	.121	.123	.134
TPOX	.206	.195	.195	.198	.202	.216
THO1	.074	.081	.082	.084	.086	.096
D16S539	.086	.089	.089	.091	.094	.105

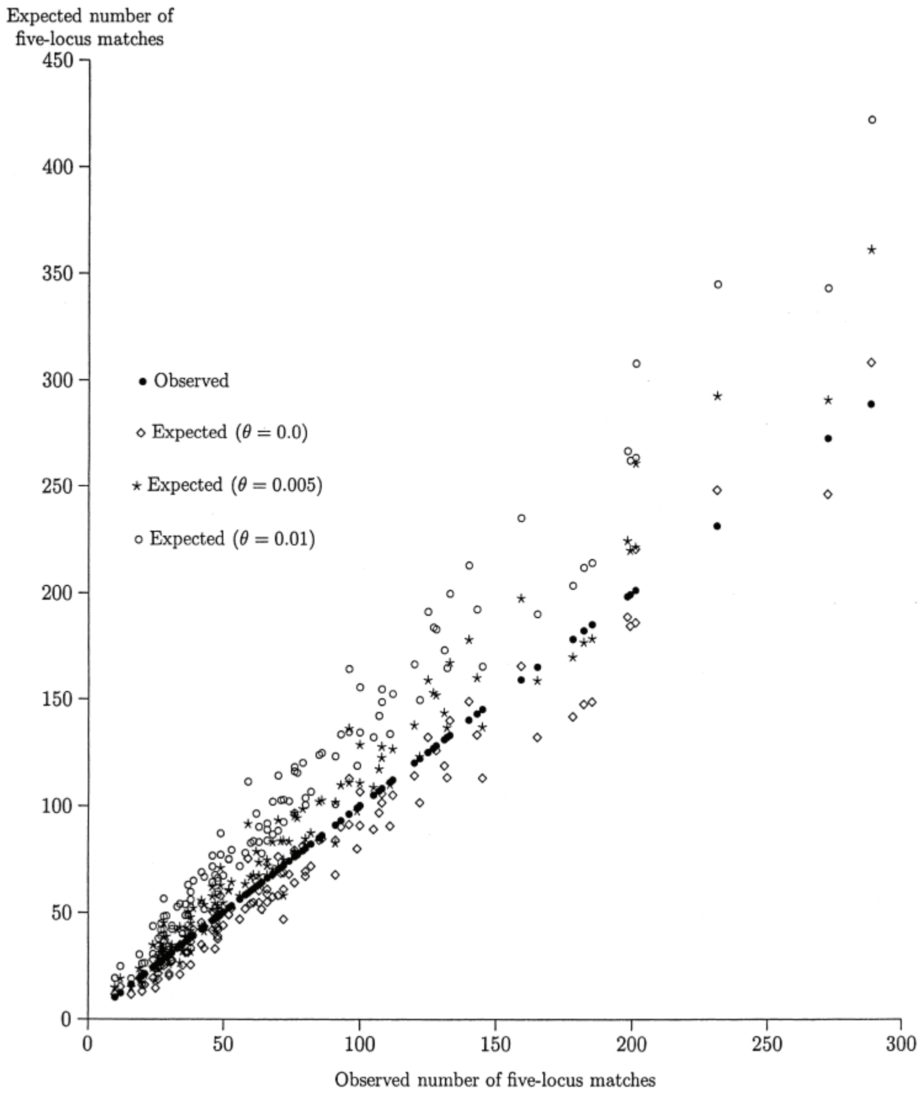
FBI Database Matching Counts

Matching		Number of Partially Matching Loci												
loci	θ	0	1	2	3	4	5	6	7	8	9	10	11	12
0	Obs.	0	3	18	92	249	624	1077	1363	1116	849	379	112	25
	.000	0	2	19	90	293	672	1129	1403	1290	868	415	134	26
	.010	0	2	14	70	236	566	992	1289	1241	875	439	148	30
1	Obs.	0	12	48	203	574	1133	1516	1596	1206	602	193	43	3
	.000	0	7	50	212	600	1192	1704	1768	1320	692	242	51	5
	.010	0	5	40	178	527	1094	1637	1779	1393	767	282	62	6
2	Obs.	0	7	61	203	539	836	942	807	471	187	35	2	
	.000	1	9	56	210	514	871	1040	877	511	196	45	5	
	.010	1	8	50	193	494	875	1096	969	593	239	57	6	
3	Obs.	0	6	33	124	215	320	259	196	92	16	1		
	.000	1	7	36	116	243	344	334	220	94	23	3		
	.010	0	6	35	117	256	380	387	268	120	32	4		
4	Obs.	1	5	17	29	54	82	67	16	6	0			
	.000	0	3	15	40	70	81	61	29	8	1			
	.010	0	3	15	44	81	98	78	40	12	1			
5	Obs.	0	1	2	6	12	14	6	5	0				
	.000	0	1	4	9	13	11	6	2	0				
	.010	0	1	4	11	16	15	9	3	0				
6	Obs.	0	1	0	2	2	0	0	0					
	.000	0	0	1	1	1	1	0	0					
	.010	0	0	1	2	2	1	1	0					

Predicted Matches when $n = 65,493$

Matching	Number of partially matching loci							
loci	0	1	2	3	4	5	6	7
6	4,059	37,707	148,751	322,963	416,733	319,532	134,784	24,125
7	980	7,659	24,714	42,129	40,005	20,061	4,150	
8	171	1,091	2,764	3,467	2,153	530		
9	21	106	198	163	50			
10	2	7	8	3				
11	0	0	0					
12	0	0						
13	0							

Multi-locus Matches



STR Survey: $\hat{\beta}$ Values for Groups and Loci

	Geographic Region								
Locus	Africa	AusAb	Asian	Cauc	Hisp	IndPK	NatAm	Poly	Aver.
CSF1PO	0.003	0.002	0.008	0.008	0.002	0.007	0.055	0.026	0.011
D1S1656	0.000	0.000	0.000	0.002	0.003	0.000	0.000	0.000	0.011
D2S441	0.000	0.000	0.002	0.003	0.021	0.000	0.000	0.000	0.020
D2S1338	0.009	0.004	0.011	0.017	0.013	0.003	0.023	0.005	0.031
D3S1358	0.004	0.010	0.009	0.006	0.012	0.040	0.079	0.001	0.025
D5S818	0.002	0.013	0.009	0.008	0.014	0.018	0.044	0.007	0.029
D6S1043	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.016
D7S820	0.004	0.021	0.010	0.007	0.007	0.046	0.030	0.005	0.026
D8S1179	0.003	0.007	0.012	0.006	0.002	0.031	0.020	0.008	0.019
D10S1248	0.000	0.000	0.000	0.002	0.004	0.000	0.000	0.000	0.007
D12S391	0.000	0.000	0.000	0.003	0.020	0.000	0.000	0.000	0.010
D13S317	0.015	0.016	0.013	0.008	0.014	0.025	0.050	0.014	0.038
D16S539	0.007	0.002	0.015	0.006	0.009	0.005	0.048	0.004	0.021
D18S51	0.011	0.012	0.014	0.006	0.004	0.010	0.033	0.003	0.018
D19S433	0.009	0.001	0.009	0.010	0.014	0.000	0.022	0.014	0.023
D21S11	0.014	0.012	0.013	0.007	0.006	0.023	0.067	0.018	0.021
D22S1045	0.000	0.000	0.007	0.001	0.000	0.000	0.000	0.000	0.015
FGA	0.002	0.009	0.012	0.004	0.007	0.016	0.021	0.006	0.013
PENTAD	0.008	0.000	0.012	0.012	0.002	0.017	0.000	0.000	0.022
PENTAE	0.002	0.000	0.017	0.006	0.003	0.012	0.000	0.000	0.020
SE33	0.000	0.000	0.012	0.001	0.000	0.000	0.000	0.000	0.004
TH01	0.022	0.001	0.022	0.016	0.018	0.014	0.071	0.017	0.071
TPOX	0.019	0.087	0.016	0.011	0.007	0.018	0.064	0.031	0.035
VWA	0.009	0.007	0.017	0.007	0.012	0.022	0.028	0.005	0.023
All Loci	0.006	0.014	0.010	0.007	0.008	0.018	0.043	0.011	0.022

Buckleton JS, Curran JM, Goudet J, Taylor D, Thiery A, Weir BS. 2016. Forensic Science International: Genetics 23:91-100.

Predicted Kinship Values

$$\begin{tikzcd}
 & & A \arrow[dl] \arrow[dr] & & \\
{} \arrow[dr] & \vdots \arrow[d] & & \vdots \arrow[d] & {} \arrow[dl] \\
 & X \arrow[dr] & & Y \arrow[dl] & \\
 & & I & &
\end{tikzcd}$$

Identify the path linking the parents X, Y of I to their common ancestor(s).

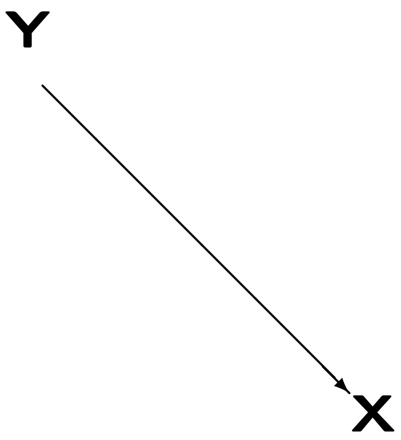
Path Counting

If the parents X, Y of an individual I have ancestor A in common, and if there are n individuals (including X, Y, I) in the path linking the parents through A, then the inbreeding coefficient of I, or the kinship of X and Y, is

$$F_I = \theta_{XY} \quad = \quad \left(\frac{1}{2}\right)^n (1 + F_A)$$

If there are several ancestors, this expression is summed over all the ancestors.

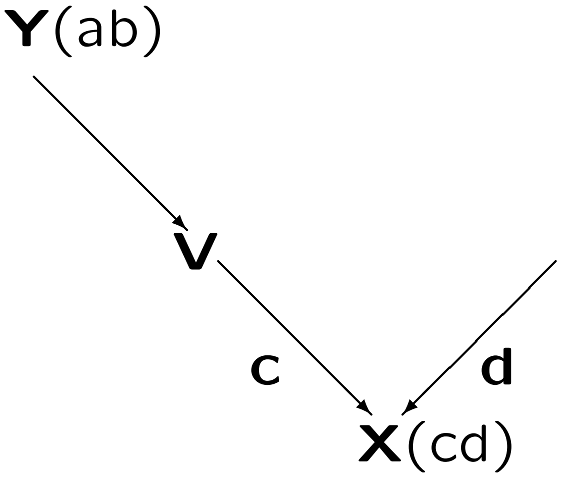
Parent-Child



The common ancestor of parent X and child Y is X. The path linking X, Y to their common ancestor is YX and this has $n = 2$ individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

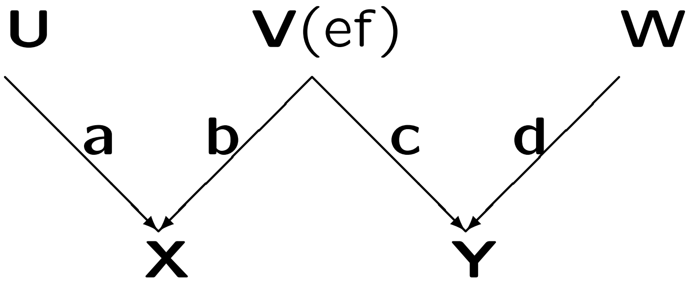
Grandparent-grandchild



The path joining X to Y is XVY with $n = 3$:

$$\theta_{XY} \;=\; \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

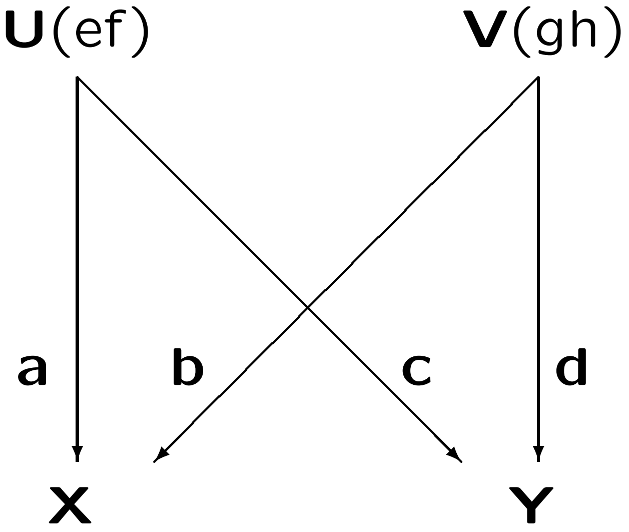
Half sibs



There is one path joining X to Y: XVY with $n = 3$:

$$\theta_{XY} \;=\; \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

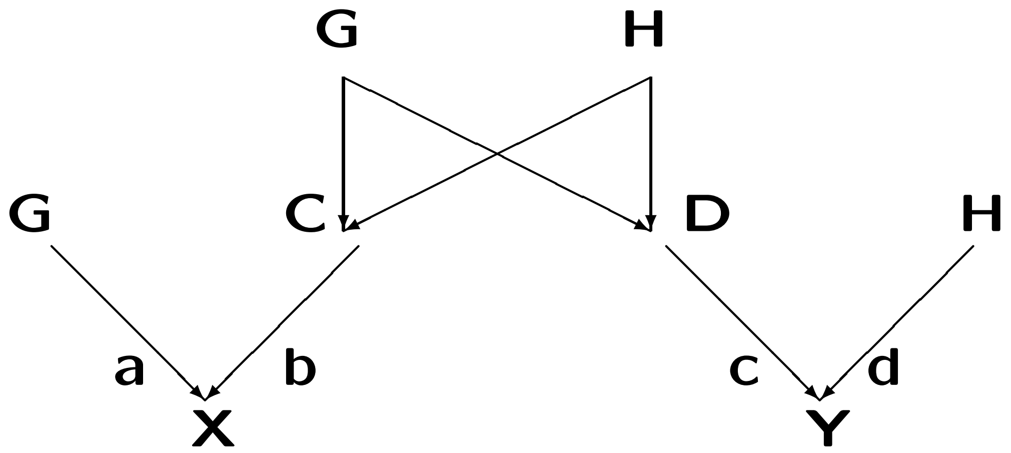
Full sibs



There are two paths joining X to Y: XUY and XVY each with $n = 3$:

$$\theta_{XY} \;=\; \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 = \frac{1}{4}$$

First cousins



Common Relatives

Relationship	Kinship
Identical Twins	0.5
Parent Child	0.25
Full Sibs	0.25
Half Sibs	0.125
Double First Cousins	0.125
First Cousins	0.0625
Uncle Niece	0.0625
Unrelated	0

Comparing Hypothesized Relationships

Current practise is to compare the likelihoods of two profiles under alternative hypotheses about their degrees of relatedness.

On the verge now of being able to estimate the degree of relatedness, especially with very large numbers of markers..

Estimating Kinship

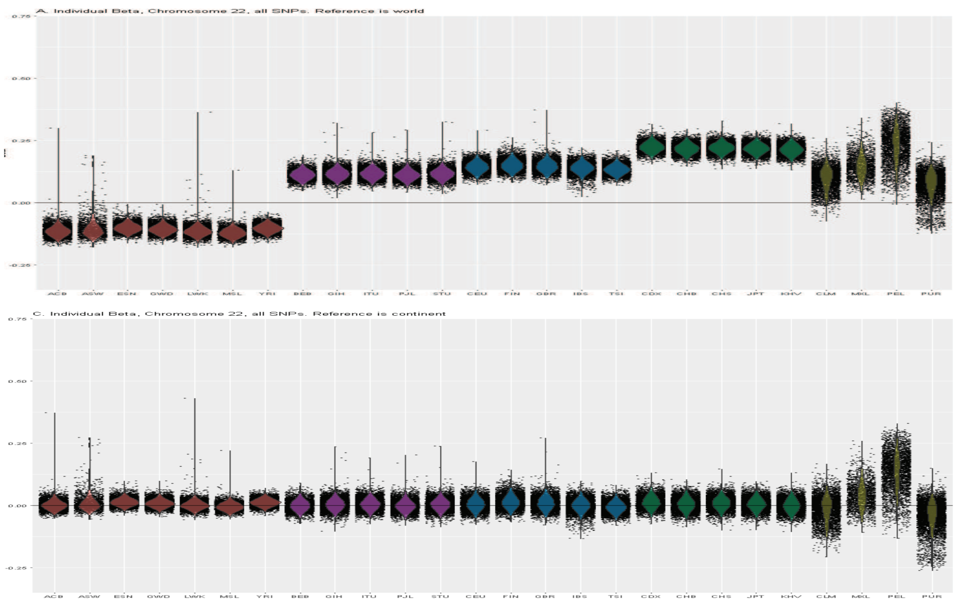
The proportion $\tilde{M}_{XY}$ of pairs of alleles, one from individual X and one from individual Y, that match is 0, 0.5 or 1:
Proportion=1: AA and AA
Proportion=0.5: AA and AB or AB and AB
Proportion=0: AA and BB or AA and BC or AB and CD

Averaging over all pairs of individuals, one er population, the observed proportion is $\tilde{M}^B$. The kinship of individuals X, Y, relative to that of all individuals in different populations is

$$\hat{\theta}_{XY} = \frac{\tilde{M}_{XY} - \tilde{M}^B}{1 - \tilde{M}^B}$$

Kinship is relative, not absolute

Top row: Whole world reference. Bottom row: Continental group reference.



Beta estimates

Chromosome 22 data from 1000 Genomes.

Continents (left to right): AFR, SAS, EUR, EAS, AMR

Populations (l to r):**AFR**: ACB, ASW, ESN, GWD, LWK, MSL, YRI;
SAS: BEB, GIH, ITU, PJL, STU; **EUR**: CEU, FIN, GBR, IBS, TSI;
EAS: CDX, CHB, CHS, JPT; **AMR**: KHV, CLM, MXL, PEL, PUR

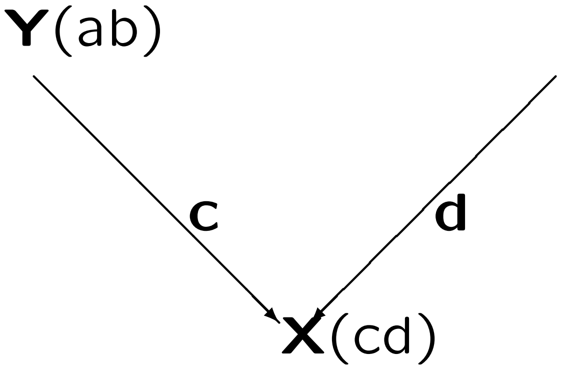
k-coefficients

The coancestry coefficient is the probability of a pair of alleles being ibd.

For joint genotypic frequencies, and for a more detailed characterization of relatedness of two non-inbred individuals, we need the probabilities that they carry 0, 1, or 2 pairs of ibd alleles. For example: their two maternal alleles may be ibd or not ibd, and their two paternal alleles may be ibd or not.

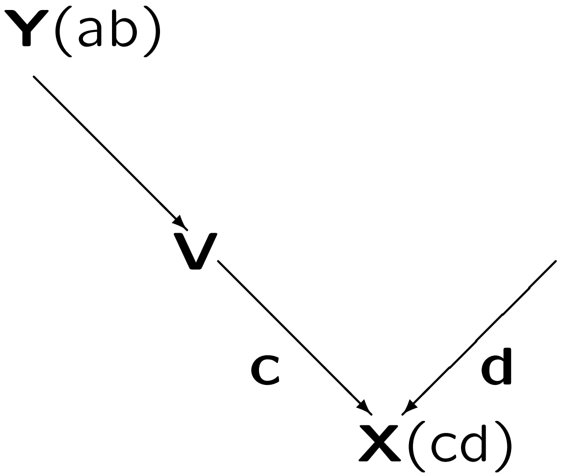
The probabilities of two individuals having 0, 1 or 2 pairs of ibd alleles are written as k_0, k_1, k_2 and $\theta = \frac{1}{2}k_2 + \frac{1}{4}k_1$.

Parent-Child



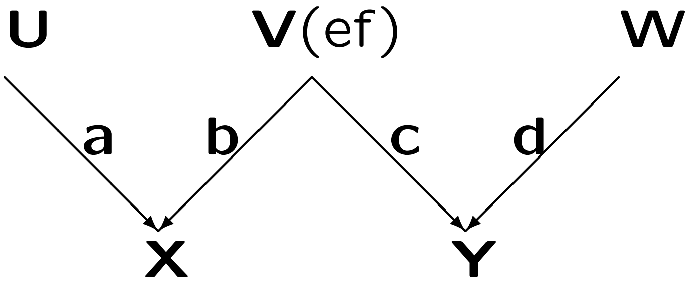
$$\text{Pr}(c \equiv a) = 0.5, \quad \text{Pr}(c \equiv b) = 0.5, \quad k_1 = 1$$

Grandparent-grandchild



$$\mathrm{Pr}(c \equiv a) = 0.25, \quad \mathrm{Pr}(c \equiv b) = 0.25, \quad k_1 = 0.5\& k_0 = 0.5$$

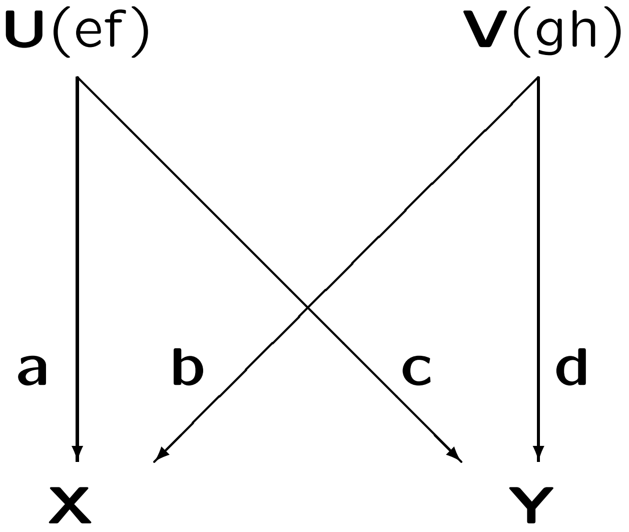
Half sibs



		0.5	0.5
		$c \equiv e$	$c \equiv f$
0.5	$b \equiv e$	0.25	0.25
0.5	$b \equiv f$	0.25	0.25

Therefore $k_1 = 0.5$ so $k_0 = 0.5$.

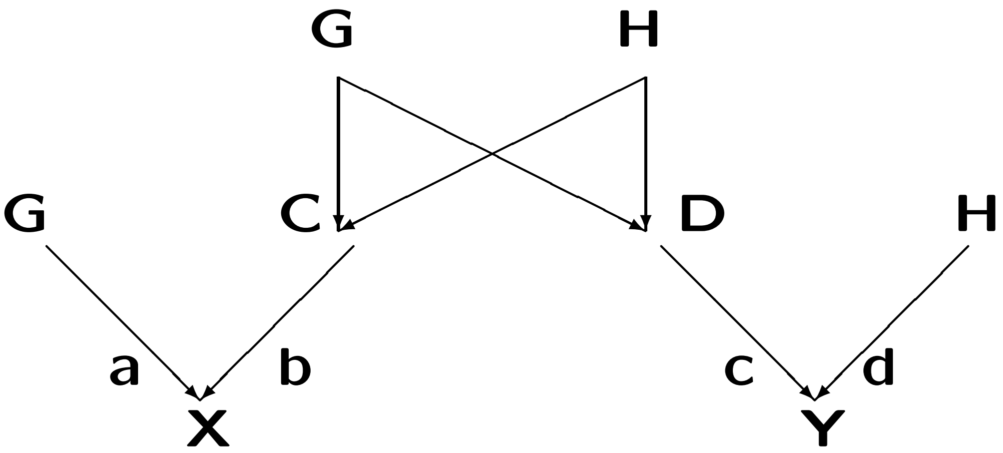
Full sibs



		0.5	0.5
		$b \equiv d$	$b \not\equiv d$
0.5	$a \equiv c$	0.25	0.25
0.5	$a \not\equiv c$	0.25	0.25

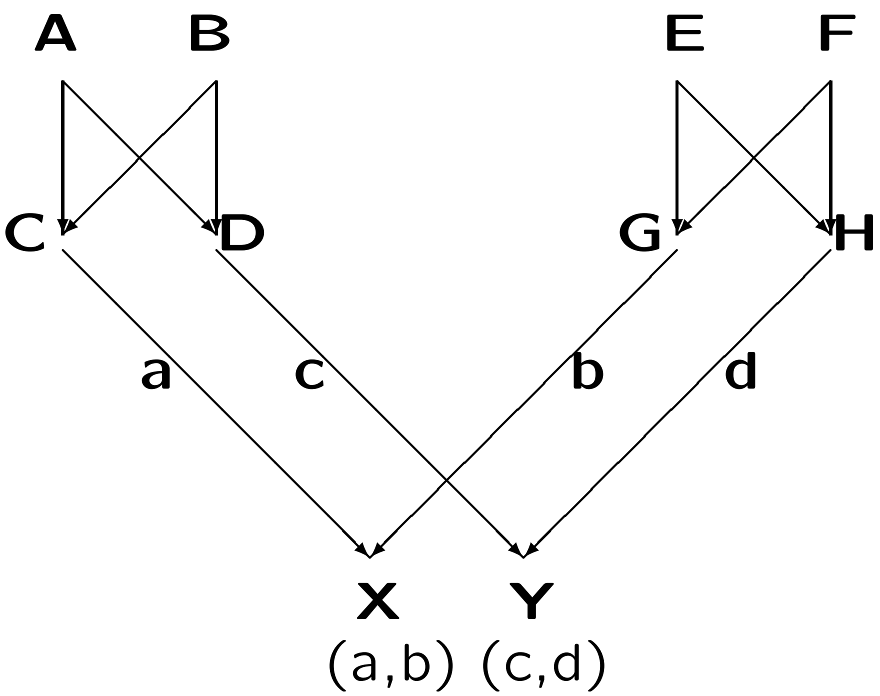
$$k_0 = 0.25, k_1 = 0.50, k_2 = 0.25$$

First cousins



Double First Cousins

What are the k's for double first cousins?

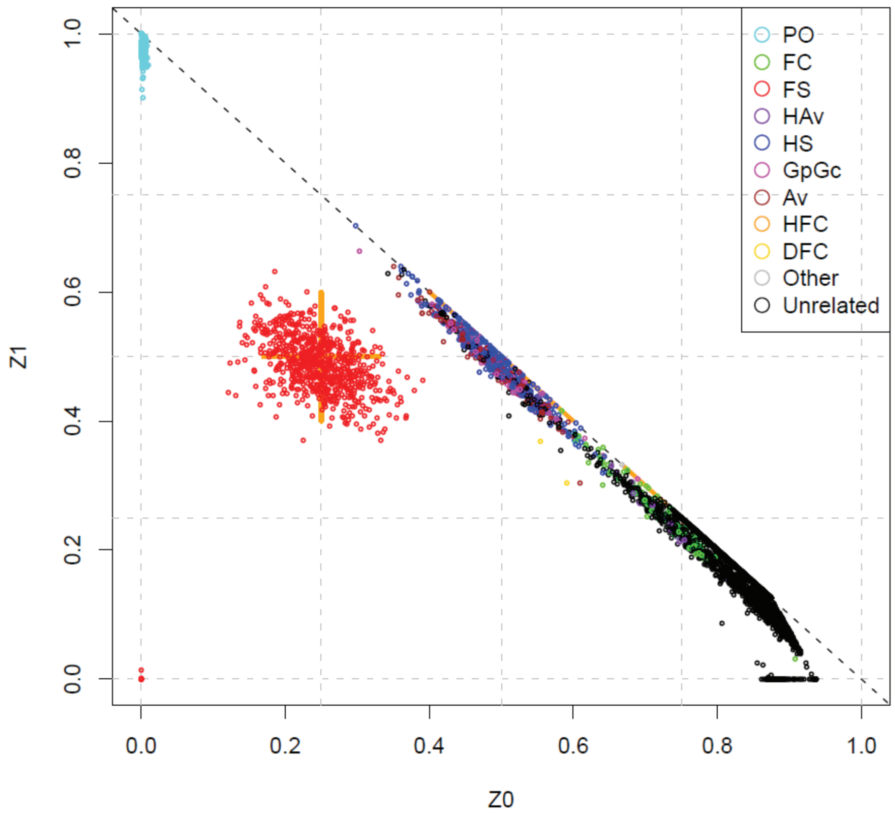


Non-inbred Relatives

Relationship	k_2	k_1	k_0	$\theta = \frac{1}{2}k_2 + \frac{1}{4}k_1$
Identical twins	1	0	0	$\frac{1}{2}$
Full sibs	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
Parent-child	0	1	0	$\frac{1}{4}$
Double first cousins	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$	$\frac{1}{8}$
Half sibs*	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{8}$
First cousins	0	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{16}$
Unrelated	0	0	1	0

* Also grandparent-grandchild and avuncular (e.g. uncle-niece).

PLINK Example



Joint genotypic probabilities

Genotypes	Probability
ii, ii	$k_2 p_i^2 + k_1 p_i^3 + k_0 p_i^4$
ii, jj	$k_0 p_i^2 p_j^2$
ii, ij	$k_1 p_i^2 p_j + 2 k_0 p_i^3 p_j$
ii, jk	$2 k_0 p_i^2 p_j p_k$
ij, ij	$2 k_2 p_i p_j + k_1 p_i p_j (p_i + p_j)$ $+ 4 k_0 p_i^2 p_j^2$
ij, ik	$k_1 p_i p_j p_k + 4 k_0 p_i^2 p_j p_k$
ij, kl	$4 k_0 p_i p_j p_k p_l$

Example: Non-inbred full sibs

Genotypes	Probability
ii, ii	$p_i^2(1+p_i)^2/4$
ii, jj	$p_i^2p_j^2/4$
ii, ij	$p_ip_j(p_i+p_j)/2$
ii, jk	$p_i^2p_jp_k/2$
ij, ij	$p_ip_j(1+p_i+p_j+2p_ip_j)/2$
ij, ik	$p_ip_jp_k(1+2p_i)/2$
ij, kl	$p_ip_jp_kp_l$

Match Probabilities with θ for Relatives

$$\begin{aligned}
\text{Pr(Match)} &= k_2 + k_1[\sum_i \text{Pr}(A_iA_iA_i) + \sum_i \sum_{j\neq i} \text{Pr}(A_iA_jA_j)] \\
&\quad + k_0P_2 \\
&= k_2 + k_1[\theta + (1-\theta)S_2] + k_0P_2
\end{aligned}$$

$$\begin{aligned}
\text{Pr(Partial Match)} &= k_1[2\sum_i \sum_{j\neq i} \text{Pr}(A_iA_iA_j) + \sum_i \sum_{j\neq i} \sum_{k\neq i,j} \text{Pr}(A_iA_jA_k)] \\
&\quad + k_0P_1 \\
&= k_1(1-\theta)(1-S_2) + k_0P_1
\end{aligned}$$

$$\text{Pr(Mismatch)} = k_0P_0$$

Quantities P_0, P_1, P_2 are given on Slide 29.

Match probabilities with $\theta = 0.03$

Locus	Not related	First-cousins	Parent -child	Full-sibs
D3S1358	.089	.124	.229	.387
vWA	.077	.111	.213	.376
FGA	.048	.078	.166	.345
D8S1179	.083	.119	.227	.384
D21S11	.051	.081	.172	.349
D18S51	.040	.068	.150	.335
D5S818	.175	.216	.339	.463
D13S317	.101	.139	.252	.401
D7S820	.080	.115	.219	.379
CSF1PO	.134	.173	.288	.428
TPOX	.216	.261	.397	.503
THO1	.096	.133	.241	.395
D16S539	.105	.143	.256	.404
Total	2×10^{-14}	2×10^{-12}	6×10^{-9}	5×10^{-6}

Arizona Matches: Mueller Analysis

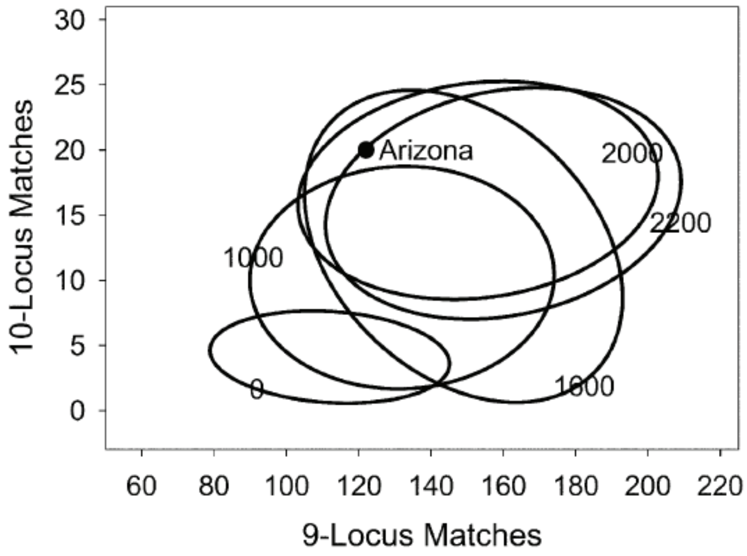


Figure 8. 95% confidence ellipsoids for simulations in which θ was set to 0.015 and the number of full sibs varied. The number on each ellipsoid corresponds to the number of pairs of sibs present in the simulated databases.

Mueller LD. 2008. Journal of Genetics 87:101-107.

Mueller Comment

“The product rule with some minor modification is the most common method for computing the frequency of DNA profiles in forensic laboratories. This method relies critically on the assumption that there is statistical independence between loci. The empirical support for this method comes mainly from tests of independence between pairs of loci (Budowle et al. 1999). However, recent research on finite populations, with mutation and a monogamous mating system shows that departures from the product rule get worse as one looks at more loci (Dr Yun Song, personal communication). Thus, rigorous testing of the product rule predictions at many loci may yield different results than prior work at only two loci. Perhaps the most important qu1ality control issue in forensic DNA typing is determining the adequacy of the methods for computing profile frequencies.”

Mueller LD. 2008. Journal of Genetics 87:101-107.

“RELPAIR” calculations

This approach compares the probabilities of two genotypes under alternative hypotheses; H_0: the individuals have a specified relationship, versus H_1: the individuals are unrelated. The alternative is that $k_0 = 1, k_1 = k_2 = 0$ so the likelihood ratios for the two hypotheses are:

$$\begin{aligned}
\mathrm{LR}(MM, MM) &= k_0 + k_1/p_M + k_2/p_M^2 \\
\mathrm{LR}(mm, mm) &= k_0 + k_1/p_m + k_2/p_m^2 \\
\mathrm{LR}(Mm, Mm) &= k_0 + k_1/(4p_Mp_m) + k_2/(2p_Mp_m)
\end{aligned}$$

$$\begin{aligned}
\mathrm{LR}(MM, Mm) &= k_0 + k_1/(2p_M) \\
\mathrm{LR}(mm, Mm) &= k_0 + k_1/(2p_m)
\end{aligned}$$

$$\mathrm{LR}(MM, mm) = k_0$$

Section 6: Reporting Likelihood Ratios

Components

- Hierarchy of propositions

- Formulating propositions

- Communicating LRs

Likelihood Ratio

The LR assigns a numerical value in favor or against one proposition over another:

$$\text{LR} = \frac{P(E|H_p, I)}{P(E|H_d, I)},$$

where H_p typically aligns with the prosecution case, H_d is a reasonable alternative consistent with the defense case, and I is the relevant background information.

Setting Propositions

- The value for the LR will depend on the propositions chosen: different sets of propositions will lead to different LRs.

- Choosing the appropriate pair of propositions can therefore be just as important as the DNA analysis itself.

Hierarchy of Propositions

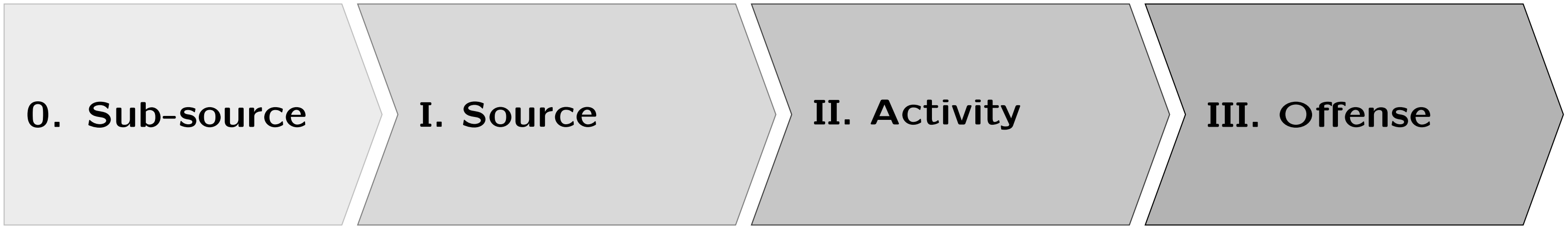
Evett & Cook (1998) established the following hierarchy of propositions:

Level	Scale	Example
III	Offense	H_p: The suspect raped the complainant. H_d: Some other person raped the complainant.
II	Activity	H_p: The suspect had intercourse with the complainant. H_d: Some other person had intercourse with the complainant.
I	Source	H_p: The semen came from the suspect. H_d: The semen came from an unknown person.
0	Sub-source	H_p: The DNA in the sample came from the suspect. H_d: The DNA in the sample came from an unknown person.

Hierarchy of Propositions

- The *offense* level deals with the ultimate issue of guilt/ innocence, which are outside the domain of the forensic scientist.
- The *activity* level associates a DNA profile or evidence source with the crime itself, and there may be occasions where a scientist can address this level.
- The *source* level associates a DNA profile or evidence item with a particular body fluid or individual source.
- The *sub-source* level refers to the strength of the evidence itself. This is usually the level a DNA reporting analyst will spend most of their time.

Hierarchy of Propositions



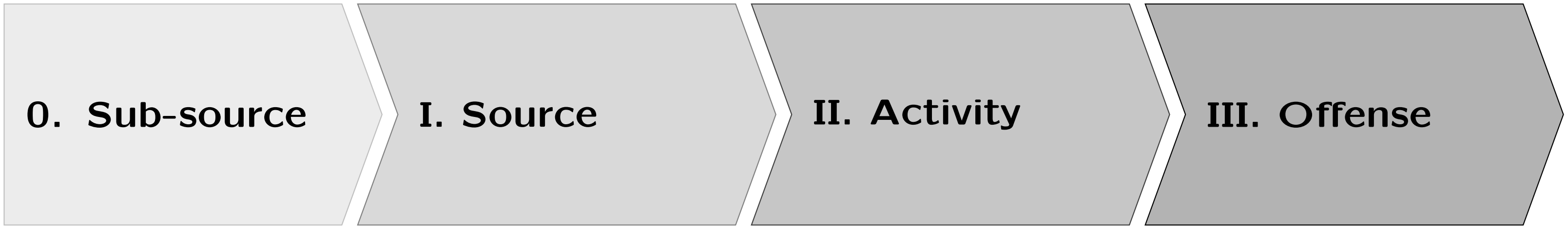
0. Sub-source **I. Source** **II. Activity** **III. Offense**

- A forensic scientist can provide information in relation to propositions which are intermediate to the ultimate issue.

- The higher the level of propositions, the more information is needed on the framework of circumstances.

- Since different levels rely on different assumptions to consider, strength of the evidence estimates will change significantly at each level.

Hierarchy of Propositions



0. Sub-source

I. Source

II. Activity

III. Offense

- Probabilistic genotyping is (usually) centered around sub-source level.

- Transition from sub-source to source or even activity level may be possible, e.g. by considering contamination, secondary transfer, timing, etc.

Setting Propositions

Some useful principles for setting hypotheses:

- Propositions should address the issue of interest;

- Propositions should be based on relevant case information;

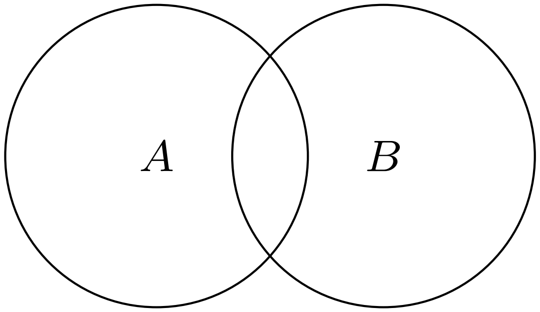
- Propositions should not include irrelevant details;

- Propositions should be (close to) MECE.

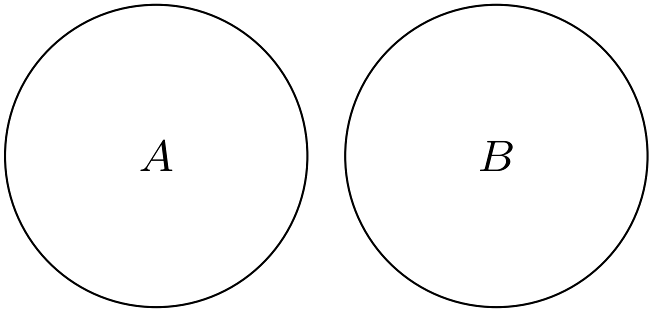
MECE Definition

Mutually exclusive

(i.e. non-overlapping)



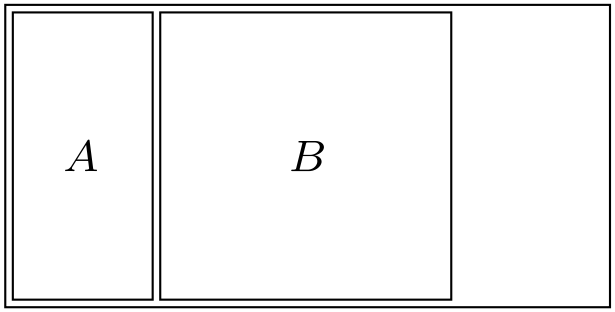
Not exclusive



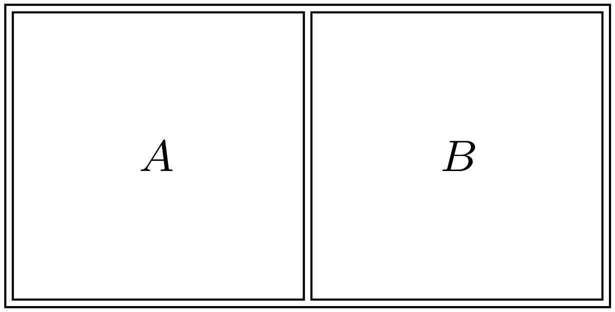
Exclusive

Collectively exhaustive

(i.e. covers all outcomes)



Not exhaustive



Exhaustive

Background Information

- **Relevant background information** can help set appropriate propositions. E.g. the origin of clothing or intimate vs. non-intimate swab can help determine if it is reasonable to assume a known contributor.

- **Irrelevant background information** is not needed and may contribute to bias decision making (e.g. criminal history, confession, presence or lack of other evidence).

Formulating Propositions

- The prosecution hypothesis (H_p) is usually known, or more or less straightforward to set.

- However, the defense are usually under no requirement to offer a proposition, and often they do not.

- If a defense stance is not available, a sensible proposition can be chosen.

Formulating Propositions - Example 1

An individual is discovered looking into a house one night. The police are called and find a single cigarette butt under the window where the incident occurred. No one in the family smokes. The police have a person of interest captured on a neighbor's CCTV.

A single-source profile is obtained from the cigarette butt and the reference profile of a person of interest (POI) matches.

Formulating Propositions - Example 1

An individual is discovered looking into a house one night. The police are called and find a single cigarette butt under the window where the incident occurred. No one in the family smokes. The police have a person of interest captured on a neighbor's CCTV.

A single-source profile is obtained from the cigarette butt and the reference profile of a person of interest (POI) matches.

H_p : The evidence came from the POI.

H_d : The evidence came from an unknown person.

Or, for simplicity:

H_p : POI

H_d : Unknown (U)

Formulating Propositions - Example 2

A complainant calls 911 to report a sexual assault in her home. She is taken to a hospital where an intimate swab is collected.

A POI is identified from the investigation and the obtained profile from the swab is fully explained by a mixture of the complainant (K) and the POI.

Formulating Propositions - Example 2

A complainant calls 911 to report a sexual assault in her home. She is taken to a hospital where an intimate swab is collected.

A POI is identified from the investigation and the obtained profile from the swab is fully explained by a mixture of the complainant (K) and the POI.

$$\begin{aligned} H_p: &\quad \text{K} + \text{POI} \\ H_d: &\quad \text{K} + \text{U} \end{aligned}$$

Formulating Propositions - Example 3

A complainant is cut with a knife during an altercation. Based upon eyewitness testimony, a POI is identified.

A stain on the clothing of the POI is tested for blood, and a DNA profile is developed that is consisted with a mixture of the POI and the complainant.

Formulating Propositions - Example 3

A complainant is cut with a knife during an altercation. Based upon eyewitness testimony, a POI is identified.

A stain on the clothing of the POI is tested for blood, and a DNA profile is developed that is consisted with a mixture of the POI and the complainant.

$$\begin{aligned} H_p: &\quad \text{POI} + \text{K} \\ H_d: &\quad \text{POI} + \text{U} \end{aligned}$$

Note how the direction of transfer provides important information.

Formulating Propositions - Example 4

Molotov cocktails have been thrown at random cars. An unexploded container is found in the street, and a 2 person mixture is developed from the evidence.

Two persons of interest are arrested.

Formulating Propositions - Example 4

Molotov cocktails have been thrown at random cars. An unexploded container is found in the street, and a 2 person mixture is developed from the evidence.

Two persons of interest are arrested.

$$\begin{aligned} H_p &: \quad \text{POI 1} + \text{POI 2} \\ H_{d1} &: \quad \text{POI 1} + \text{U} \\ H_{d2} &: \quad \text{POI 2} + \text{U} \\ H_{d3} &: \quad \text{2U} \end{aligned}$$

What if circumstances indicate that they cannot both be present?

Formulating Propositions - Example 5

A complainant walking through a city park is attacked from behind and is sexually assaulted on a blanket. She didn't get a good look at the perpetrator. The police recognize the blanket as possibly belonging to a vagrant known to live near the park.

A profile obtained from the blanket is fully explained by mixing of K and POI's DNA.

Formulating Propositions - Example 5

A complainant walking through a city park is attacked from behind and is sexually assaulted on a blanket. She didn't get a good look at the perpetrator. The police recognize the blanket as possibly belonging to a vagrant known to live near the park.

A profile obtained from the blanket is fully explained by mixing of K and POI's DNA.

$$\begin{aligned} H_p &: \quad \text{K} + \text{POI} \\ H_{d1} &: \quad \text{POI} + \text{U} \\ H_{d2} &: \quad \text{K} + \text{U} \\ H_{d3} &: \quad \text{2U} \end{aligned}$$

Formulating Propositions

What if multiple alternative hypotheses are relevant?

- Report the ‘most relevant’ LR (and provide the rest in the appendix);
- Provide all considered propositions and corresponding LRs;
- Report only the lowest LR to provide a lower bound for the LR.

Note that if K is a true source of the profile, but not considered under H_d, the LR will be larger than when assuming K as a known profile under both hypotheses. This is because K will explain many of the observed alleles (especially in case of being a major donor).

The Effect of Propositions on the LR

Consider a simple two-person mixture profile (e.g. contributors are unrelated, ignoring population structure, no drop-outs/drop-ins), where $G_C = ABCD$. Let K denote a known contributor with observed profile $G_K = CD$, and S the POI with profile $G_S = AB$.

- LR $= \frac{P(ABCD|H_p\colon\ K{+}S)}{P(ABCD|H_d\colon\ K{+}U)} = \frac{1}{2p_Ap_B}$;

- LR $= \frac{P(ABCD|H_p\colon\ K{+}S)}{P(ABCD|H_d\colon\ 2U)} = \frac{1}{6{\cdot}4p_Ap_Bp_Cp_D} = \frac{1}{24p_Ap_Bp_Cp_D}$;

- LR $= \frac{P(ABCD|H_p\colon\ S{+}U)}{P(ABCD|H_d\colon\ 2U)} = \frac{2p_Cp_D}{6{\cdot}4p_Ap_Bp_Cp_D} = \frac{1}{12p_Ap_B}$.

For $p_A = p_B = p_C = p_D = 0.1$, this yields LRs of 50, 417 and 8, respectively.

Formulating Propositions

What about relatives?
The LR can accommodate for this, which we will see in the next section.

What if the DNA got there by some other means?
This indicates a different level of propositions. The discussion will likely move to transfer and contamination.

Be prepared to change:
Propositions are formed based on information available at that time. If this information changes, or the defense want any other propositions considered, it may be necessary to update or add LR calculations.

Reporting LRs

As can be seen from the definition of the likelihood ratio

$$\text{LR} = \frac{P(E|H_p)}{P(E|H_d)},$$

- an LR > 1 supports the prosecution hypothesis, meaning that the evidence is more likely if H_p is true than if H_d is true;
- an LR < 1 supports the defense hypothesis;
- an LR $= 1$ is consistent with the observations being equally likely under the considered hypotheses.

Reporting LRs

The likelihood ratio is usually reported using phrases such as:

> *"The evidence is . . . more likely if the suspect is the donor of the sample than if someone else is the donor of the sample"*.

It is important to note that the LR is not an absolute measure of the weight of evidence, but is dependent on the underlying hypotheses.

How to express the LR in terms of a verbal 'equivalent'?

Verbal Scales

A verbal scale for evidence interpretation, applied to the prosecution proposition:

Likelihood Ratio	**Verbal Equivalent**
$1 <$ LR ≤ 10	Limited support (for H_p)
$10 <$ LR ≤ 100	Moderate support (for H_p)
$100 <$ LR $\leq 1\,000$	Moderately strong support (for H_p)
$1\,000 <$ LR $\leq 10\,000$	Strong support (for H_p)
$10\,000 <$ LR $\leq 1\,000\,000$	Very strong support (for H_p)
$1\,000\,000 <$ LR	Extremely strong support (for H_p)

The equivalent for H_d is given by taking the reciprocal.

Verbal Scales

The association of words with numbers is subjective and arbitrary.

LR	1	$1-10$	$10-10^2$	10^2-10^3	10^3-10^4	10^4-10^6	$>10^6$
Evett & Weir (1998)	–	l	l	m	s	vs	vs
Evett (2000)	–	l	m	ms	s	vs	vs
Martire (2015)	–	w or l	m	ms	s	vs	es
Taroni (2016)	n	l	m	s	vs	es	es

Using verbal scales of neutral (n), weak (w), limited (l), moderate (m), moderately strong (ms), strong (s), very strong (vs) and extremely strong (es).

Verbal Scales

Should we report a verbal equivalent for the LR?

Verbal Scales

Should we report a verbal equivalent for the LR?

- **Yes**: The verbal scale is helpful for the jury to put the LR into perspective.

- **No**: The verbal scale is not the responsibility of the forensic scientist.

Presenting Evidence

There are a lot of difficult issues that arise in interpreting DNA samples and presenting complex scientific evidence to non-expert judges and juries.

A sufficiently deep understanding of the principles can help an expert witness to make well-informed judgments and find good solutions to the problem of satisfying goals such as clarity, precision and simplicity.

> *"How forensic evidence is presented is at least as important as what is presented"*.

> *". . . it is not only what forensic experts say but how they say it that must be considered"*.

Heuristics and Biases

Valid probabilistic reasoning is not easy, so people often use various tricks, rules of thumb, habits, etc., to reason in daily life. These are called *heuristics*.

Heuristics may suffice for most practical situations, but can lead to systematic errors in probabilistic reasoning (i.e. fallacies).

Case Study 1

Quickly read/say the colors of the word:

RED
ORANGE
YELLOW
GREEN
BLUE
PURPLE

Case Study 1

Quickly read/say the colors of the word:

RED
ORANGE
YELLOW
GREEN
BLUE
PURPLE

Automatic cognitive processes are unintentional and involuntary, and occur outside awareness, probably controlling us more than we want to admit.

Case Study 2

Which option has the most paths? What is the difference?

Option A	**Option B**
XXXXXXXX	XX
XXXXXXXX	XX
XXXXXXXX	XX
	XX
	XX
	XX
	XX
	XX
	XX

Case Study 2

Option A	**Option B**
XXXXXXXX	XX
XXXXXXXX	XX
XXXXXXXX	XX
	XX
	XX
	XX
	XX
	XX
	XX

The number of paths is the same for both options:

$$8^3 = 2^9 = 512$$

In a study (Tversky and Kahneman) 85% of respondents found more paths in option A (median: 40) than in option B (median: 18).

This is an example of *availability heuristic*, i.e. the likelihood of an event is estimated as the ease with which examples of such events can be retrieved from memory.

Case Study 3

An unusual disease is expected to kill 600 people. Two alternative programs to combat the disease have been proposed:

- If program A is adopted, 200 people will be saved.

- If program B is adopted, there is a 1/3 chance that all 600 people will be saved and a 2/3 chance that nobody will be saved.

Which program would you choose?

Case Study 3

An unusual disease is expected to kill 600 people. Two alternative programs to combat the disease have been proposed:

- If program C is adopted, 400 people will die.
- If program D is adopted, there is a 1/3 chance that nobody will die and a 2/3 chance that all 600 people will die.

Which program would you choose?

Case Study 3

All four programs have the same expected outcome: 200 people will live, 400 will die.

When framed in terms of gains, 72% choose program A (risk-averse). When framed in terms of losses, 78% choose program D (risk-taking).

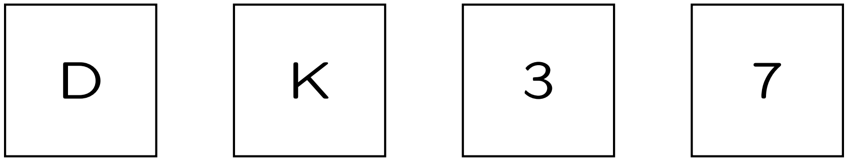
Certain gain is preferred over possible gain, while possible loss is preferred over certain loss.

This is an example of the *framing effect*.

Case Study 4

Four cards, each with a letter on one side and a number on the other, are placed on a table. The following hypothesis is proposed:

Every card that has a D on one side has a 3 on the other.

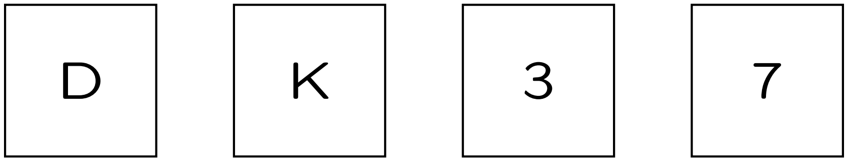


D K 3 7

Which card(s) need to be turned over to determine whether the hypothesis is true?

Case Study 4

Hypothesis: *Every card that has a D on one side has a 3 on the other.*



D K 3 7

The correct answer is D and 7. Selecting D and 3 is indicative of *confirmation bias*, i.e. the tendency to search for or interpret information in a way that confirms one's preexisting beliefs or hypotheses, but $P(3|D) \neq P(D|3)$.

Case Study 5

Estimate the number resulting from the following expression:

$$8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2$$

Case Study 5

Estimate the number resulting from the following expression:

$$2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$$

Case Study 5

Estimate the number resulting from the following expression:

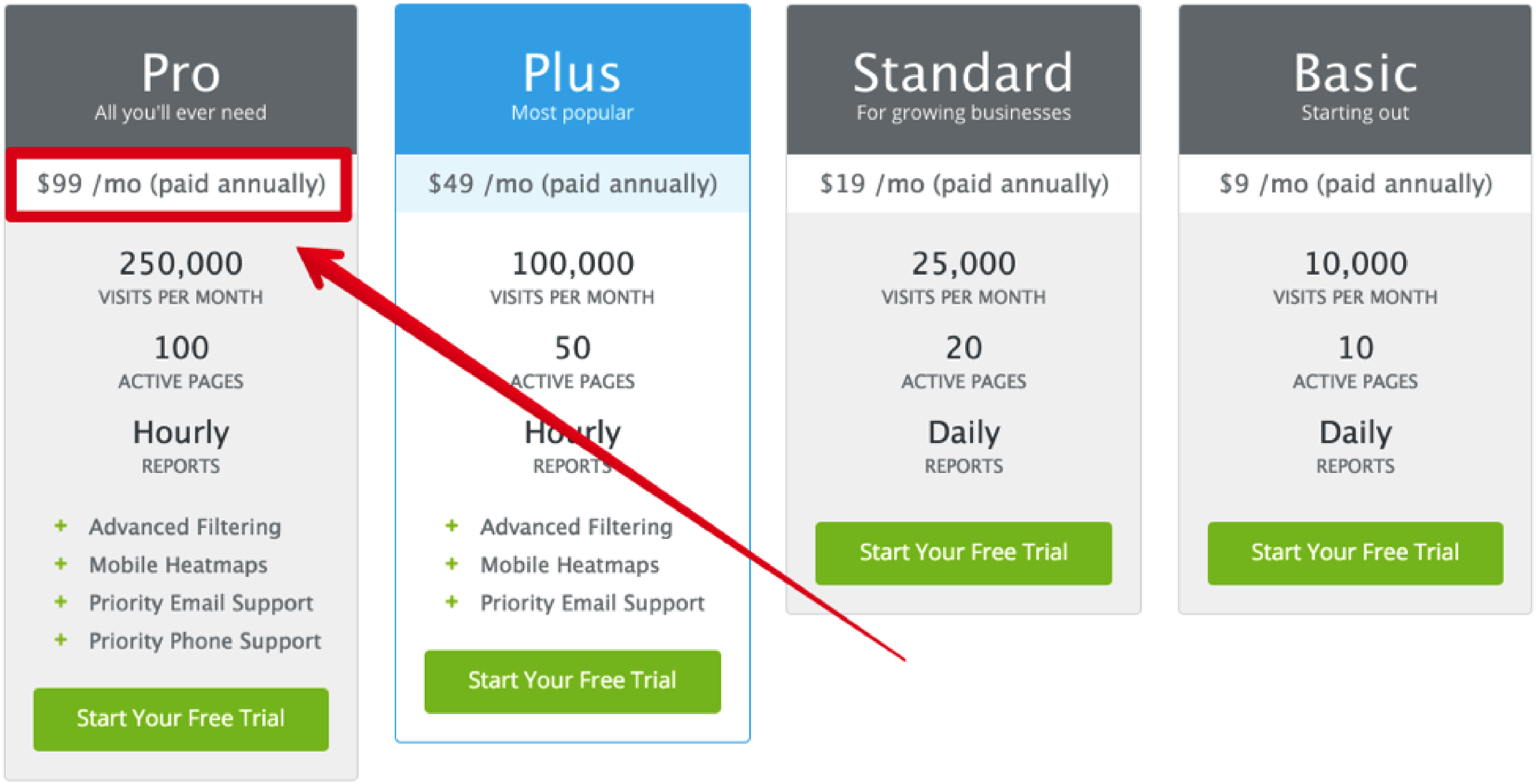
$$8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2$$

$$2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$$

Subjects gave a median estimate of 2 250 in the first case, while the second case had a median of 512. The true answer is of course $8! = 40\,320$.

This is an example of *anchoring*, i.e. estimates may depend too much on an initial number.

Case Study 5b



Pro	Plus	Standard	Basic
All you'll ever need	Most popular	For growing businesses	Starting out
$99 /mo (paid annually)	$49 /mo (paid annually)	$19 /mo (paid annually)	$9 /mo (paid annually)
250,000 VISITS PER MONTH	100,000 VISITS PER MONTH	25,000 VISITS PER MONTH	10,000 VISITS PER MONTH
100 ACTIVE PAGES	50 ACTIVE PAGES	20 ACTIVE PAGES	10 ACTIVE PAGES
Hourly REPORTS	Hourly REPORTS	Daily REPORTS	Daily REPORTS
+ Advanced Filtering + Mobile Heatmaps + Priority Email Support + Priority Phone Support	+ Advanced Filtering + Mobile Heatmaps + Priority Email Support		
Start Your Free Trial	Start Your Free Trial	Start Your Free Trial	Start Your Free Trial

Case Study 6

- Of the women complaining of painful hardening of the breast, 1% have a malignant tumor: $P(C) = 0.01$.

- The accuracy ($+$ or $-$) of a mammography is 90%: $P(+|C) = P(-|C') = 0.9$.

- Estimate $P(C|+)$ to decide whether or not to order a biopsy.

Case Study 6

Most physicians estimate $P(C|+) \approx 0.75$, while the correct answer is:

$$P(C|+) = \frac{P(+|C)P(C)}{P(+|C)P(C) + P(+|C')P(C')} = 0.0833.$$

Representativeness leads people to neglect the base rate, by assessing a conditional probability by the 'degree of similarity' $(P(A|B) \neq P(B|A))$. This is known as the *base rate fallacy*.

Bias in Forensic Science

- *Attractiveness bias*: Attractive criminals get lower sentences.
- *Target/suspect driven bias*: Using a reference profile to resolve drop-outs.
- *Base rate expectation*: Routinely pairing of examiners and reviewers, high verification rates.
- *Anchoring*: A dice throw influencing sentencing decisions[1].

[1] Playing Dice With Criminal Sentences (Englich, 2006).

Bias in Forensic Science

Cognitive bias (i.e. unintentional bias) affects forensic decision-making:

- Biases lead to differences between and within (forensic) experts;
- Bias doesn't necessarily translate into an error in interpretation;
- But cognitive contamination should be avoided just as physical contamination.

This, relatively new, area is often called *cognitive forensics*.

Avoiding Bias

The first step in avoiding cognitive bias is *awareness*: appreciate that it exists, and identify where it resides and affects interpretation, through training and education.

Awareness is necessary, but is insufficient to reduce cognitive bias and contamination: active steps must be taken as mere will power does not control bias.

Several methods have been proposed that can help manage bias sources, such as *Linear Sequential Unmasking*[1].

[1] Strengthening forensic DNA decision making through a better understanding of the influence of cognitive bias (Dror, 2017).

Bias in Forensic Science

What about probabilistic genotyping software?

- Interpretation software can reduce variation in interpretation among examiners.

- It does *not* make interpretation bias free;

- Subjectivity is also involved in software development (and underlying modeling).

- Different software can show LRs varying over 10 logs for the same DNA profiles.

Fallacies

Biases can lead to potential fallacies in the courtroom, and may even lead to a miscarriage of justice.

- Prosecutor's fallacy
- Defendant's fallacy
- Uniqueness fallacy
- Association fallacy

Prosecutor's Fallacy

One of the most common errors is to transpose the conditional:

$$P(A|B) \neq P(B|A),$$

e.g. saying that there is a very high probability that an animal has four legs *if* it is an elephant, is not the same as the probability that an animal is an elephant *if* it has four legs.

$$P(\text{4 legs} \mid \text{Elephant}) \neq P(\text{Elephant} \mid \text{4 legs}).$$

Prosecutor's Fallacy

This example may seem obvious, but it's often not so easy in court proceedings:

$$P(E|H_p) \neq P(H_p|E),$$

or, alternatively,

$$\begin{aligned} P(\text{Evidence} \mid \text{Proposition}) &\neq P(\text{Proposition} \mid \text{Evidence}) \\ &\neq P(\text{Proposition}) \end{aligned}$$

Prosecutor's Fallacy - Exercise

- The evidence is much more likely if the DNA profile came from the suspect.
- The probability of this DNA profile if it came from someone else is very low.
- The probability that this DNA profile came from someone else is very low.
- The probability of someone else having this DNA profile is very low.
- The probability of someone else leaving DNA of this type is very low.
- The evidence strongly supports the hypothesis that the DNA profile came from the suspect.

Prosecutor's Fallacy - Exercise

- The evidence is much more likely if the DNA profile came from the suspect.
- The probability of this DNA profile if it came from someone else is very low.
- The probability that this DNA profile came from someone else is very low.
- The probability of someone else having this DNA profile is very low.
- The probability of someone else leaving DNA of this type is very low.
- The evidence strongly supports the hypothesis that the DNA profile came from the suspect.

Prosecutor's Fallacy

- Subtle misstatements can lead (and have led) to misunderstandings.

- Forensic scientists should be (and are trained to be) very careful about the wording of probability statements.

Defendant's Fallacy

Suppose $P(E|H_d)$ is reported as 1 in 1 000. The defendant's fallacy is a logical error that usually favors the defendant:

- The city where the crime occurred has population size 100 000;
- So there are 100 people with a matching profile;
- This means that $P(H_p|E)$ is only 1 in 100 or 1%.

Defendant's Fallacy

Suppose $P(E|H_d)$ is reported as 1 in 1 000. The defendant's fallacy is a logical error that usually favors the defendant:

- The city where the crime occurred has population size 100 000;
- So there are 100 people with a matching profile;
- This means that $P(H_p|E)$ is only 1 in 100 or 1%.

Defendant's Fallacy

Suppose $P(E|H_d)$ is reported as 1 in 1 000. The defendant's fallacy is a logical error that usually favors the defendant:

- The city where the crime occurred has population size 100 000;
- So we *expect* 100 people with a matching profile;
- $P(H_p|E)$ is 1 in 100 or 1% *only* if each of these individuals has the same prior probability.

Uniqueness Fallacy

Suppose $P(E|H_d)$ is reported as 1 in 100 000. The uniqueness fallacy argues:

- The city where the crime occurred has population size 100 000;
- So there is only one individual with a matching profile;
- This means that this DNA profile is unique in this city and must come from the suspect.

Uniqueness Fallacy

Suppose $P(E|H_d)$ is reported as 1 in 100 000. The uniqueness fallacy argues:

- The city where the crime occurred has population size 100 000;
- So there is only one individual with a matching profile;
- This means that this DNA profile is unique in this city and must come from the suspect.

Uniqueness Fallacy

Suppose $P(E|H_d)$ is reported as 1 in 100 000.

- The city where the crime occurred has population size 100 000;
- So we *expect* 1 *other* individual with a matching profile;
- This usually also incorporates the belief that DNA profiles yield unique identification, which is untrue in light of LTDNA, often leading to complex mixtures and partial profiles (and ignores relatives, coancestry and phenomena such as drop-in).

Association Fallacy

An association fallacy occurs when a probability statement is transposed from one scale of the hierarchy of propositions to a higher level.

This is usually a result from assuming that there is a dependency between two observations or events, e.g.:

- Statements about evidence samples (sub-source) that are interpreted as the 'evidence being more likely if the suspect is the *source* of the crime stain';

- Or even on *activity* level as 'the evidence is more likely if the suspect left the crime stain'.

Fallacies in Practice - Case Example

The *People v. Nelson* (CA) court's decision report contains the following statements:

> *"In 2002, investigators compared evidence from a 1976 murder scene with defendant's deoxyribonucleic acid (DNA) profile and identified him as a possible donor of that evidence. He was then tried for and convicted of that murder. The prosecution presented evidence that the odds that a random person unrelated to defendant from the population group that produced odds most favorable to him could have fit the profile of some of the crime scene evidence are one in 930 sextillion (93 followed by 22 zeros)."*

> *"Because the worlds total population is only about seven billion (seven followed by nine zeros), this evidence is tantamount to saying that defendant left the evidence at the crime scene."*

> *". . . We also conclude that the jury properly heard evidence that it was virtually impossible that anyone other than defendant could have left the evidence found at the crime scene."*

Fallacies in Practice - Case Example

The *People v. Nelson* (CA) court's decision report contains the following statements:

> *"... Specifically, [the defendant] contends the evidence regarding the odds that the crime scene evidence could have come from some other person was inadmissible because the statistical method used to calculate those odds has not achieved general scientific acceptance under the standard stated in [...] People v. Kelly (1976) 17 Cal.3d 24 (sometimes referred to as the Kelly test)."*

> *"... Defendant agrees that using the product rule to calculate the random match probability makes sense when comparing one suspects profile with the crime scene evidence because, as he explains, the random match probability "estimates the chance that any single, random person drawn from the relevant population would have the same DNA profile as that of the unknown person whose DNA was found at the crime scene.""*

> *"... It is already settled that the product rule reliably shows the rarity of the profile in the relevant population. [...] To this extent, the product rule has already passed the Kelly test."*

Fallacies in Practice - Case Example

The *People v. Nelson* (CA) court's decision report contains the following statements:

> *"The Court of Appeal in this case and other courts that have considered this question have concluded that use of the product rule in a cold hit case is not the application of a new scientific technique subject to a further Kelly (or Kelly-like) test."*

> *"We agree. Jenkins explained its reasoning: "At the heart of this debate is a disagreement over the competing questions to be asked, not the methodologies used to answer those questions. [. . .] [T]here is no controversy in the relevant scientific community as to the accuracy of the various formulas. In other words, the math that underlies the calculations is not being questioned. [. . .] [T]he debate . . . is one of relevancy, not methodology . . . .""*

> *". . . The debate that exists is solely concerned with which number – rarity, database match probability, Balding-Donnelly, or some combination of the above is most relevant in signifying the importance of a cold hit. "*

Fallacies in Practice - Case Example

The *People v. Nelson* (CA) court's decision report contains the following statements:

> *"The database match probability ascertains the probability of a match from a given database. "But the database is not on trial. Only the defendant is". Thus, the question of how probable it is that the defendant, not the database, is the source of the crime scene DNA remains relevant. The rarity statistic addresses this question."*

> *"The fact that the match ultimately came about by means of a database search does not deprive the rarity statistic of all relevance. It remains relevant for the jury to learn how rare this particular DNA profile is within the relevant populations and hence how likely it is that someone other than defendant was the source of the crime scene evidence. Accordingly, the trial court correctly admitted the evidence, and the Court of Appeal correctly upheld that admission."*

Miscarriage of Justice - Case Example 1

Adam Scott was arrested, accused of rape and incarcerated on the basis of a DNA profile match, which was eventually traced back to a contamination incident.

> *"It is estimated that the chance of obtaining matching DNA components if the DNA came from someone else unrelated to Adam Scott is approximately one in 1 billion. In my opinion the DNA matching that of Adam Scott has most likely originated from semen. [. . .] In my opinion these findings are what I would expect if Adam Scott had some form of sexual activity with [the victim]. In order to assess the overall findings in this case I have therefore considered the following propositions:*
>
> - *Adam Scott had vaginal intercourse with [the victim]*
> - *Adam Scott has never been to Manchester and does not know [the victim]"*

Source: Misleading DNA Evidence (Gill, 2014).

Miscarriage of Justice - Case Example 1

- The perpetrator DNA was absent (hidden perpetrator effect and false inclusion error).

- The DNA match was falsely associated with the presence of sperm (association fallacy).

- The ‘presence’ of sperm was associated with sexual intercourse (association fallacy).

- Exculpatory evidence was ignored (base rate fallacy and confirmation bias).

Different biases/effects resulted in a *compounded error* or *snowball effect*.

Miscarriage of Justice - Case Example 2

The association fallacy assumes a dependency between two observation or events. The opposite version may also lead to errors, i.e. assuming independence where non exists.

Sally Clark was arrested and convicted for the murder of her two infant sons. In this case (UK, 1999) it was assumed that two sudden infant death syndrome (SIDS) deaths in a single family were independent events. A consulting pediatrician estimated the likelihood of a cot death as 1 in 8 500, and calculated the combined probability by squaring this number (i.e. yielding a likelihood of 1 in 73 million).

Miscarriage of Justice - Case Example 2

It was later found that her second son might have died from natural causes, and moreover, assuming independence of these events is unreasonable, due to possible underlying genetic causes:

$$P(A, B) = P(A|B)P(B) \neq P(A)P(B).$$

Sally Clark was released from prison after having served more than three years of her sentence.

Section 7: Y-STR Profiles

Y-chromosome Profiles

[Work of Taryn Hall, University of Washington.]

The Y-chromosome has several STR markers that are useful in forensic science. In one respect, the profiles are easier to interpret as each man has only one allele at an STR locus. Otherwise interpretation is made more complicated by the lack of recombination on the Y chromosome, meaning that alleles at different loci are not independent. Or are they?

We expect that mutations act independently at different loci and this may counter the lack of recombination to some extent.

Y-STR Databases

There are three public databases of Y-STR profiles:

- Y-Chromosome Haplotype Reference Database (YHRD) FSI: Genetics 15:43-48 (2013)

- Human Genome Diversity Project (HGDP) Science 296:262-262 (2002)

- Data published by Xu et al. (XU) Mol Genet Genomics 290:1451-150 (2014)

Two-locus LD for Y-STR Loci

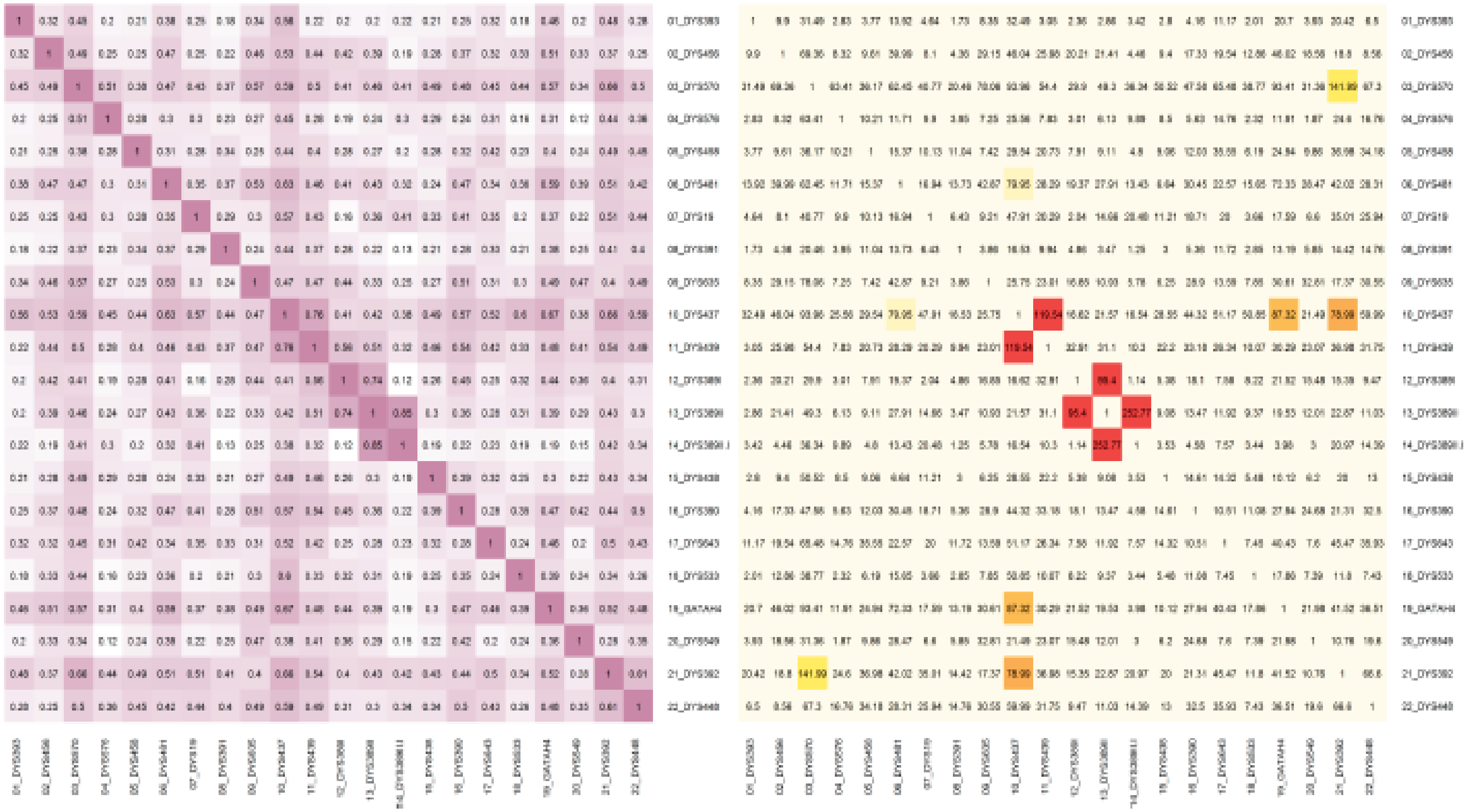


Figure D. Measures of linkage disequilbrium calculated between Y chromosome markers, European populations, Y-Chromosome Haplotype Reference Database.

Multi-locus Disequilibria: Entropy

It is difficult to describe associations among alleles at several loci. One approach is based on information theory.

For a locus with sample frequencies $\tilde{p}_u$ for alleles A_u the entropy is

$$H_A \;=\; -\sum_u \tilde{p}_u \ln(\tilde{p}_u)$$

For independent loci, entropies are additive: if haplotypes A_uB_v have sample frequencies $\tilde{P}_{uv}$ the two-locus entropy is

$$H_{AB} \;=\; -\sum_u\sum_v \tilde{P}_{uv} \ln(\tilde{P}_{uv}) = -\sum_u\sum_v \tilde{p}_u\tilde{p}_v[\ln(\tilde{p}_u) + \ln(\tilde{p}_v)] = H_A + H_B$$

so if $H_{AB} \neq H_A + H_B$ there is evidence of dependence. This extends to multiple loci.

Conditional Entropy

If the entropy for a multi-locus profile A is H_A then the conditional probability of another locus B, given A, is $H_{B|A} = H_{AB} - H_A$.

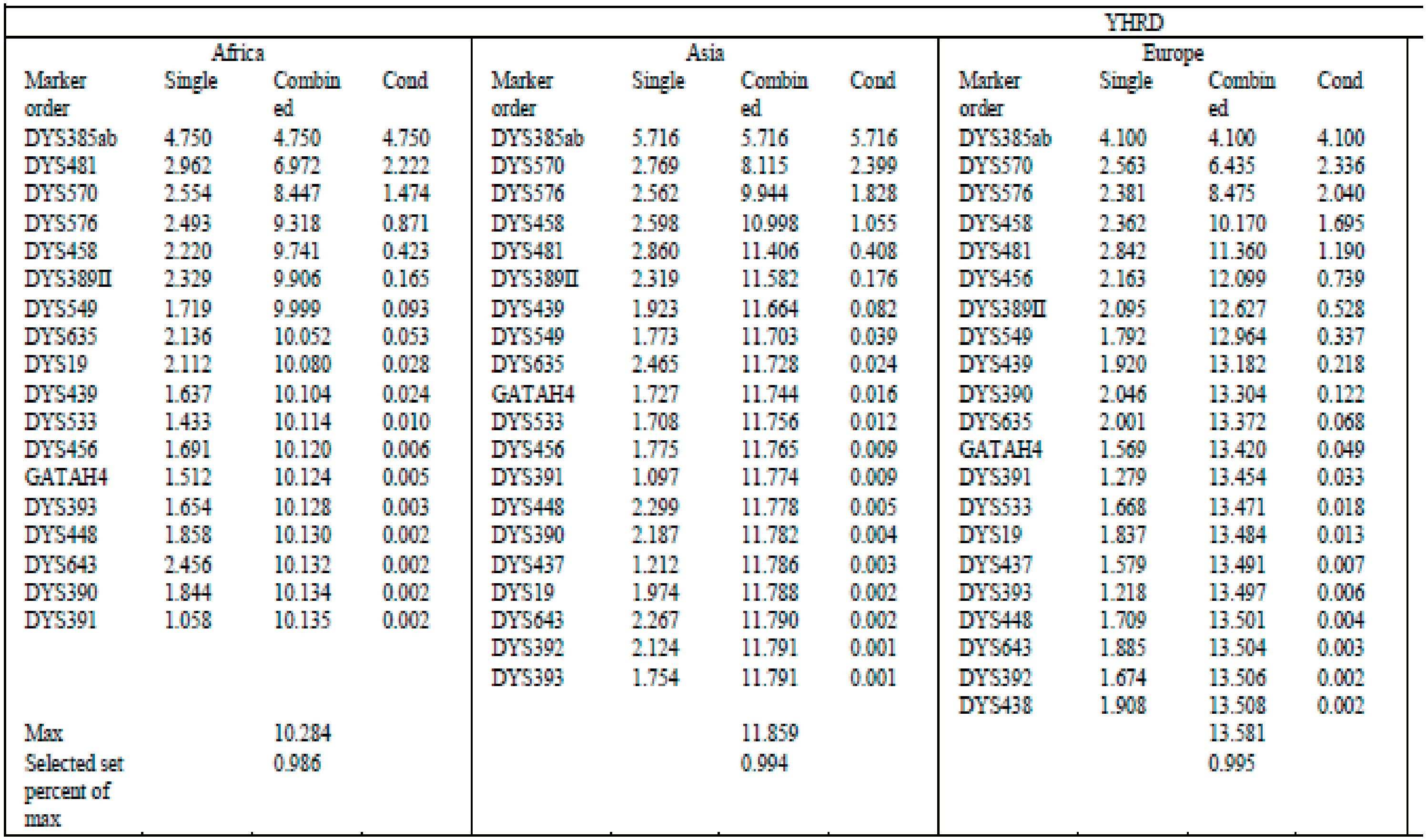
In performing meaningful calculations for Y-STR profiles, this suggests choosing a set of loci by an iterative procedure. First choose locus L_1 with the highest entropy. Then choose locus L_2 with the largest conditional entropy $H(L_2|L_1)$. Then choose L_3 with the highest conditional entropy with the haplotype L_1L_2, and so on.

Conditional Entropy: YHRD Data

Added	Entropy		
Marker	Single	Multi	Cond.
YS385ab	4.750	4.750	4.750
DYS481	2.962	6.972	2.222
DYS570	2.554	8.447	1.474
DYS576	2.493	9.318	0.871
DYS458	2.220	9.741	0.423
DYS389II	2.329	9.906	0.165
DYS549	1.719	9.999	0.093
DYS635	2.136	10.05	0.053
DYS19	2.112	10.08	0.028
DYS439	1.637	10.10	0.024
DYS533	1.433	10.11	0.010
DYS456	1.691	10.12	0.006
GATAH4	1.512	10.12	0.005
DYS393	1.654	10.13	0.003
DYS448	1.858	10.13	0.002
DYS643	2.456	10.13	0.002
DYS390	1.844	10.13	0.002
DYS391	1.058	10.13	0.002

This table shows that the most-discriminating loci may not contribute to the most-discriminating haplotypes. Furthermore, there is little additional discriminating power from Y-STR haplotypes beyond 10 loci.

Examples



									YHRD		
	Africa				Asia				Europe		
Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond
DYS385ab	4.750	4.750	4.750	DYS385ab	5.716	5.716	5.716	DYS385ab	4.100	4.100	4.100
DYS481	2.962	6.972	2.222	DYS570	2.769	8.115	2.399	DYS570	2.563	6.435	2.336
DYS570	2.554	8.447	1.474	DYS576	2.562	9.944	1.828	DYS576	2.381	8.475	2.040
DYS576	2.493	9.318	0.871	DYS458	2.598	10.998	1.055	DYS458	2.362	10.170	1.695
DYS458	2.220	9.741	0.423	DYS481	2.860	11.406	0.408	DYS481	2.842	11.360	1.190
DYS389II	2.329	9.906	0.165	DYS389II	2.319	11.582	0.176	DYS456	2.163	12.099	0.739
DYS549	1.719	9.999	0.093	DYS439	1.923	11.664	0.082	DYS389II	2.095	12.627	0.528
DYS635	2.136	10.052	0.053	DYS549	1.773	11.703	0.039	DYS549	1.792	12.964	0.337
DYS19	2.112	10.080	0.028	DYS635	2.465	11.728	0.024	DYS439	1.920	13.182	0.218
DYS439	1.637	10.104	0.024	GATAH4	1.727	11.744	0.016	DYS390	2.046	13.304	0.122
DYS533	1.433	10.114	0.010	DYS533	1.708	11.756	0.012	DYS635	2.001	13.372	0.068
DYS456	1.691	10.120	0.006	DYS456	1.775	11.765	0.009	GATAH4	1.569	13.420	0.049
GATAH4	1.512	10.124	0.005	DYS391	1.097	11.774	0.009	DYS391	1.279	13.454	0.033
DYS393	1.654	10.128	0.003	DYS448	2.299	11.778	0.005	DYS533	1.668	13.471	0.018
DYS448	1.858	10.130	0.002	DYS390	2.187	11.782	0.004	DYS19	1.837	13.484	0.013
DYS643	2.456	10.132	0.002	DYS437	1.212	11.786	0.003	DYS437	1.579	13.491	0.007
DYS390	1.844	10.134	0.002	DYS19	1.974	11.788	0.002	DYS393	1.218	13.497	0.006
DYS391	1.058	10.135	0.002	DYS643	2.267	11.790	0.002	DYS448	1.709	13.501	0.004
				DYS392	2.124	11.791	0.001	DYS643	1.885	13.504	0.003
				DYS393	1.754	11.791	0.001	DYS392	1.674	13.506	0.002
								DYS438	1.908	13.508	0.002
Max		10.284				11.859				13.581	
Selected set percent of max		0.986				0.994				0.995	

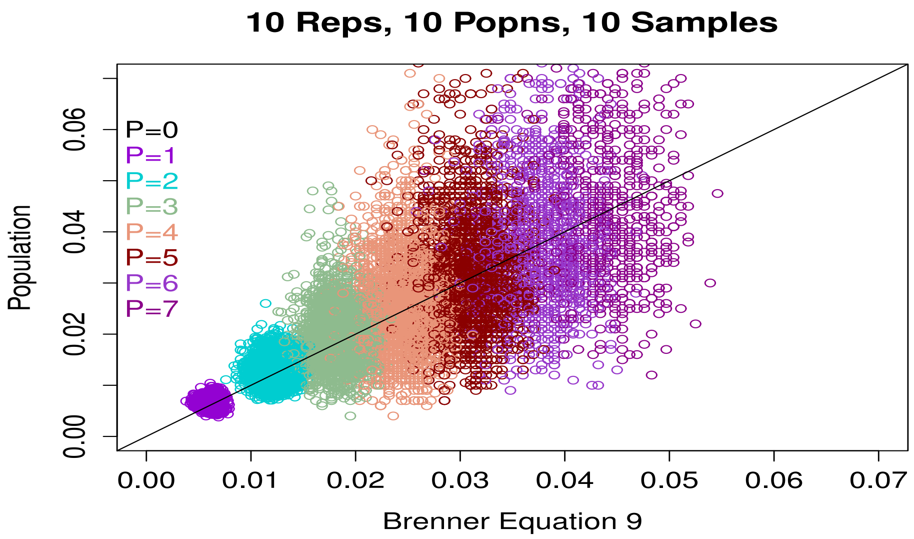
Brenner's Method

Brenner (2010) proposed the use of the proportion κ of profiles that occurred only once in a database that had been augmented by the evidentiary profile. His approach did not require a genetic model, although κ values can be predicted for some genetic models. The probability of a person taken randomly from a population would have the same profile as the evidentiary type when that type was not present in a sample of size $(n-1)$ (i.e. occurred once in the sample augmented by the evidentiary profile) was given by $(1-\kappa)/n$.

For profiles that occur p times in the augmented sample (those with "popularity" p), Brenner suggested a modification to $p(1-\kappa)/n$ that approaches the sample proportion $\tilde{p}$ when the proportion of singletons in the database becomes small.

Brenner's Method

Here we compare Brenner's estimates for every profile in the augmented database with the proportion of profiles of that type in the population from which the sample was drawn. Brenner's values appear better than the sample proportions for profiles not seen in the sample before it was augmented, as desired by Brenner. The quality decreases as the sample proportion of the evidentiary profile increases.



Brenner's Method

Brenner's estimate uses only the number of times a profile occurs ((popularity") in a database. It was not intended to do well for profiles that are seen more than a small number of times. Actual databases do have some profiles in high frequency. In Table 1 we show PPY23 haplotype counts for the YHRD database.

Popul.	Count	Popul.	Count	Popul.	Count	Popul.	Count
1	9004	14	12	28	1	53	1
2	1254	15	4	29	1	54	1
3	416	16	5	30	2	57	1
4	196	17	2	33	2	58	3
5	105	18	7	35	1	61	1
6	85	19	4	36	1	62	1
7	50	20	3	37	2	68	1
8	41	21	3	38	1	91	1
9	34	22	2	41	3	118	1
10	24	24	4	42	3	126	1
11	28	25	4	43	2	170	1
12	16	26	1	45	1	242	1
13	9	27	2	48	2		

Genetic Model

A genetic approach can be built on the notion of identity by descent. For large numbers of loci, profiles of the same type are likely to match because they have a common ancestral haplotype. If θ_i is the probability of identity by descent of two random haplotypes in population i, the probability a random profile in population i is of type A given the evidentiary profile, also from population i, is that type is $\Pr(A|A)_i = \theta_i + (1 - \theta_i)p_{Ai}$.

As profile proportions p_{Ai} become small the matching probabilities approach θ_i. These quantities, in turn, decrease as the number of loci increases. Kimura and Ohta (1968) showed that, for single-step mutations, STR loci have predicted θ values of $1/\sqrt{1 + 4N\mu}$. For L loci undergoing independent mutation we could replace μ by $1 - (1 - \mu)^L \approx L\mu$.

Y-STR Matches

The chance of a random man having Y-STR haplotype A is written as p_A, the profile probability.

The chance that two men have haplotype A is written as P_{AA}.

The chance that a man has haplotype A given that another man has been seen to have that profile is $P_{A|A}$, the match probability. The three quantities are related by $P_{A|A} = P_{AA}/p_A$.

A major difficulty is that we generally do not have samples from the relevant (sub)population to give us estimates of p_A or P_{AA}. Instead we have a database of profiles that may represent a larger population.

Interpreting Evidence

Two hypotheses for observed match between suspect and evidence:

H_P: Suspect is source of evidence.
H_D: Suspect is not source of evidence.

Then

$$\frac{\Pr(H_P|\text{Match})}{\Pr(H_D|\text{Match})} = \frac{\Pr(\text{Match}|H_P)}{\Pr(\text{Match}|H_D)} \times \frac{\Pr(H_P)}{\Pr(H_D)}$$

Interpreting Evidence

Suppose matching Y-STR profile is type A. The likelihood ratio reduces to

$$\begin{aligned}\frac{\text{Pr}(\text{Match}|H_P)}{\text{Pr}(\text{Match}|H_D)} &= \frac{\text{Pr}(A|A,H_P)}{\text{Pr}(A|A,H_D)} \\ &= \frac{1}{\text{Pr}(A|A)}\end{aligned}$$

A population genetic model introduces the quantity θ:

$$\begin{aligned}\text{Pr}(AA) &= \theta p_A + (1-\theta)p_A^2 \\ \text{Pr}(A|A) &= \theta + (1-\theta)p_A\end{aligned}$$

where θ is the probability that two profiles are identical by descent.

Within- and Between-population Matching

If the sample from population i has within-population matching proportion of $\tilde{M}_i$, the average over populations is:

$$\tilde{M}_W \;=\; \frac{1}{r}\sum_{i=1}^{r}\tilde{M}_i$$

If the sample between-population matching proportion for populations i and i' is $\tilde{M}_{ii'}$, the average over pairs of populations is:

$$\tilde{M}_B \;=\; \frac{1}{r(r-1)}\sum_{i=1}^{r}\sum_{\substack{i'=1\\ i\neq i'}}^{r}\tilde{M}_{ij}$$

We estimate theta as $\beta_W = (\tilde{M}_W - \tilde{M}_B)/(1-\tilde{M}_B)$.

Use of Database Frequencies

If data (database) from the population of interest are available they should be used directly.

For haplotype A, the database proportion $\tilde{p}_A$ is unbiased for the population proportion p_A. A confidence interval can be constructed, using properties of the binomial distribution. The $100(1-\alpha)\%$ upper confidence limit p_U when a database of size n has x copies of the target haplotype satisfies

$$\sum_{k=0}^{x} \binom{n}{k} p_U^k (1-p_U)^{n-k} \;\; \geq \;\; \alpha$$

If $x = 0$, then $(1-p_U)^n \geq \alpha$ or $p_U \leq 1 - \alpha^{1/n}$ and this is 0.0295 if $n = 100, \alpha = 0.05$. More generally $p_U \approx 3/n$ when $x = 0$ is the upper 95% confidence limit.

Use of θ-based Match Probabilities

If data are not available from the population of interest, but are available from a larger population (e.g. ethnic group), then the match-probability can be used with θ assigned or estimated from a set of subpopulations from the database population. The match probabilities use the database fequencies and β_W (for θ) and apply on average for any subpopulation.

θ for any subpopulation, or for the average over subpopulations, cannot be estimated from a single database. For example, a value for Native Amricans cannot be estimated from a Native American database.

One-locus NIST Y-STR Estimates

Locus	$\tilde{M}_W$	$\tilde{M}_B$	$\hat{\beta}_W$
DYS19	0.32571062	0.24309148	0.10915340
DYS385a/b	0.07982377	0.04427420	0.03719640
DYS389I	0.41279418	0.38319082	0.04799436
DYS389II	0.26072434	0.23741323	0.03056847
DYS390	0.28981997	0.18813203	0.12525182
DYS391	0.52191425	0.48517426	0.07136392
DYS392	0.39961865	0.35168087	0.07394164
DYS393	0.50285122	0.48769253	0.02958906
DYS437	0.46400112	0.38595032	0.12710828
DYS438	0.36817530	0.23212655	0.17717601
DYS439	0.35507469	0.34990863	0.00794667
DYS448	0.30091326	0.22640195	0.09631787
DYS456	0.33444029	0.32578009	0.01284478
DYS458	0.21642167	0.19701369	0.02416976
DYS481	0.18867019	0.14121936	0.05525373
DYS533	0.39365769	0.37177174	0.03483757
DYS549	0.33976578	0.30691346	0.04740003
DYS570	0.21298105	0.20775666	0.00659442
DYS576	0.20955290	0.18125443	0.03456321
DYS635	0.27720127	0.20653182	0.08906400
DYS643	0.28394262	0.20058158	0.10427710
Y-GATA-H4	0.40667782	0.39899963	0.01277568

Multiple-locus US-YSTR Estimates

No. Loci	Added Locus	$\tilde{M}_W$	$\tilde{M}_B$	$\hat{\beta}_W$
1	DYS_438	0.37903281	0.27283973	0.14603806
2	DYS_392	0.22353526	0.10233258	0.13501958
3	DYS_19	0.11294942	0.05471374	0.06160639
4	DYS_390	0.05923470	0.02393636	0.03616398
5	DYS_643	0.04798422	0.02456341	0.02401059
6	YGATA_C4	0.03119210	0.01541060	0.01602851
7	DYS_533	0.01979150	0.00777794	0.01210774
8	DYS_393	0.01482393	0.00650531	0.00837309
9	DYS_456	0.01073170	0.00396487	0.00679377
10	DYS_438	0.00889934	0.00287761	0.00603912
11	DYS_549	0.00524369	0.00123093	0.00401770
12	DYS_481	0.00317518	0.00055413	0.00262250
13	DYS_389I	0.00240161	0.00031517	0.00208710
14	DYS_391	0.00200127	0.00017039	0.00183119
15	DYS_576	0.00106995	0.00005877	0.00101124
16	DYS_ 389II	0.00089896	0.00004205	0.00085695
17	DYS_385	0.00065020	0.00002729	0.00062293
18	YGATA_H4	0.00063652	0.00002427	0.00061227
19	DYS_448	0.00055062	0.00000713	0.00054349
20	DYS_458	0.00051100	0.00000423	0.00050677
21	DYS_570	0.00043010	0.00000423	0.00042587
22	DYS_439	0.00038612	0.00000423	0.00038189

Combining Y & Autosomal Match Probabilities

Although autosomal and Y STR loci are unlinked, matching at autosomal and Y loci are not independent (matching in one system implies some degree of kinship and therefore matching in the other system).

N	μ	$\widehat{\theta}_Y$	$\widehat{\theta}_{AY}$	$\widehat{\theta}_A$	$\widehat{\theta}_{A\|Y}$	$\widehat{\theta}_{A\|Y} - \widehat{\theta}_A$	Walsh	$\widehat{\theta}_{AY}/(\widehat{\theta}_A\widehat{\theta}_Y)$
10^4	10^{-2}	0.00040	0.00001270	0.00123	0.03143	0.03020	0.03025	25.5580
10^4	10^{-3}	0.00447	0.00007101	0.01233	0.01587	0.00355	0.00361	1.2878
10^4	10^{-4}	0.04343	0.00483898	0.11110	0.11142	0.00032	0.00038	1.0029
10^5	10^{-2}	0.00004	0.00000123	0.00012	0.03036	0.03024	0.03024	246.6184
10^5	10^{-3}	0.00045	0.00000217	0.00125	0.00483	0.00359	0.00359	3.8785
10^5	10^{-4}	0.00452	0.00005742	0.01234	0.01271	0.00036	0.00037	1.0293
10^6	10^{-2}	0.00000	0.00000012	0.00001	0.03025	0.03024	0.03024	2457.2222
10^6	10^{-3}	0.00004	0.00000017	0.00012	0.00372	0.00359	0.00359	29.7852
10^6	10^{-4}	0.00045	0.00000073	0.00125	0.00161	0.00037	0.00037	1.2928

Y-STR matching has little effect on autosomal coancestry when θ_A, θ_Y are large but the effects can be substantial not when when θ_A, θ_Y are small.

Section 8: Incorporating Relatives

LR Problems

- A traditional LR considers an alternative proposition with unrelated individuals (which usually favors the prosecution).
  - Where does this individual come from? From the same population and sub-population, from a different sub-population, or a different population?
  - What if someone who is related to the suspect is the source of the DNA sample?
- The LR applies only to one specific defendant.

True Donor LRs

- What if there are genotypes that will result in higher LRs?

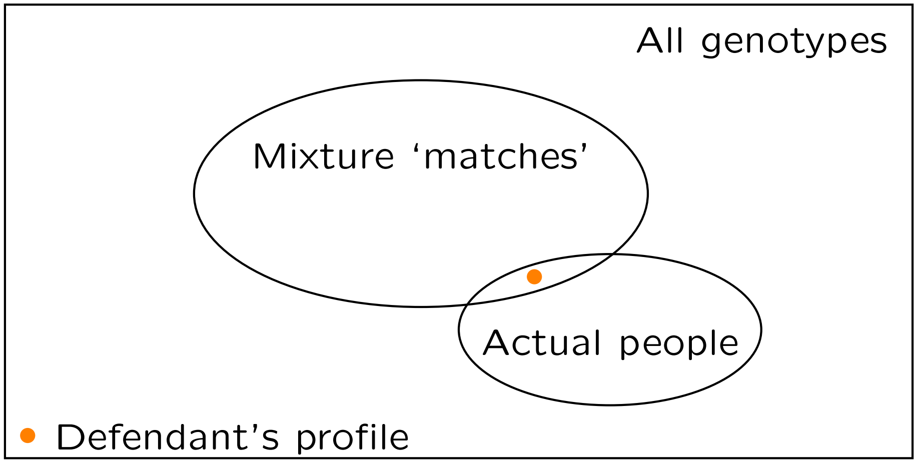
- Only in case of a very clear DNA profile will the true donor result in the highest LR (but such profiles are rarely observed from crime scene samples).

- There are possibly millions of other genotypes that are concordant with a mixture.

- If we would rank the LRs, the suspect is unlikely to produce the highest LR.

- This means that there are other genotypes that fit the data better, and provide more support for the prosecution hypothesis.

Most Genotypes Do Not Exist



But since most genotypes do not exist, there is potentially no living individual with a genotype that would produce a higher LR. Even if there are, their corresponding priors are likely low (e.g. for children, women, individuals living on a different continent).

Relatives

Because DNA profiles are inherited, relatives are more likely to share a DNA profile than unrelated individuals.

H_p: The DNA in the sample came from the suspect.
H_d: The DNA in the sample came from an unrelated individual.

H_p: The DNA in the sample came from the suspect.
H_d: The DNA in the sample came from a brother of the suspect.

The relationship type can be anything: parent, child, sibling, uncle, cousin, etc.

The more distant the relationship, the closer the value will become to the LR considering unrelated individuals.

Mendel's Laws

Mendel laid down the basic principles of heredity, even though DNA was not yet discovered.

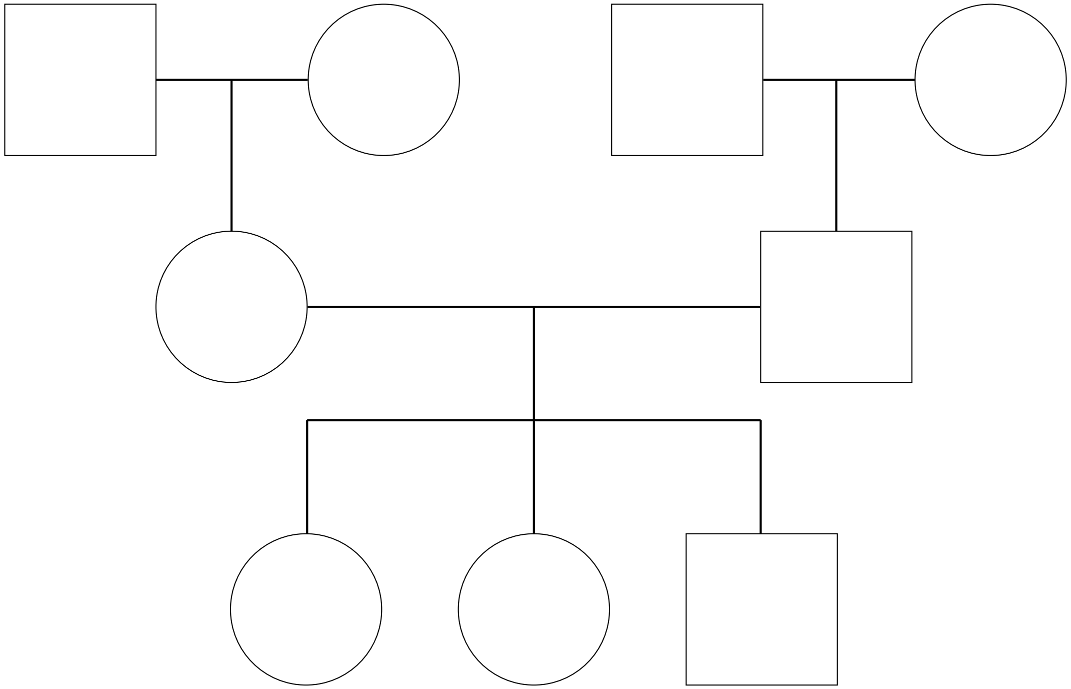
1. **The law of segregation**: An individual will pass down one of their two alleles to each offspring.

2. **The law of independent assortment**: Alleles for different traits segregate independently.

3. **The law of dominance**: If an individual's two alleles are different, one will be dominant.

Pedigrees

Pedigrees provide a graphical representation of relationships.



Individuals are said to be related if they share a common ancestor. Relationships can be unilateral (one-sided) or bilateral (two-sided).

Identity By Descent

- Relatives are similar because they share alleles that are *identical by descent* (IBD).

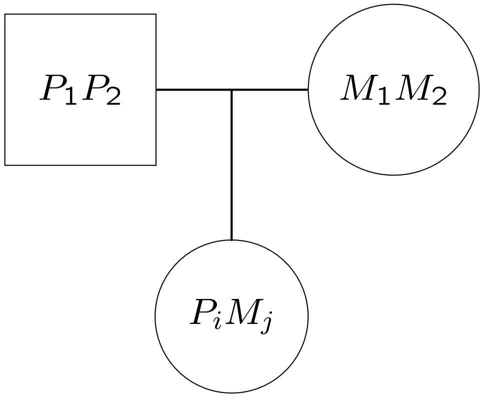
- IBD alleles are copies of the same allelic type inherited through a common ancestor (and ignores mutation).

- A pedigree or relationship determines IBD probabilities, which determine probabilities of joint genotypes.

IBD for Parent-Child Relationships

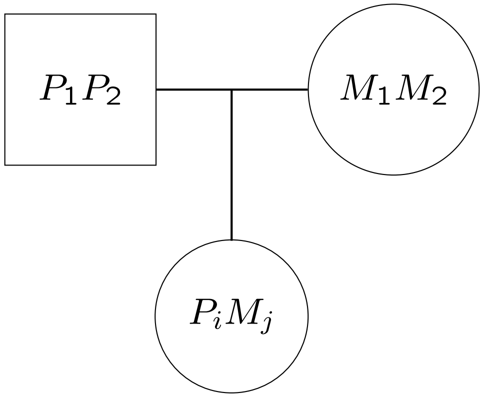
- Mendel's law states that one of the two alleles from a parent will be passed down to a child;

- Both alleles have equal probability $\frac{1}{2}$ of being passed down.



IBD for Parent-Child Relationships

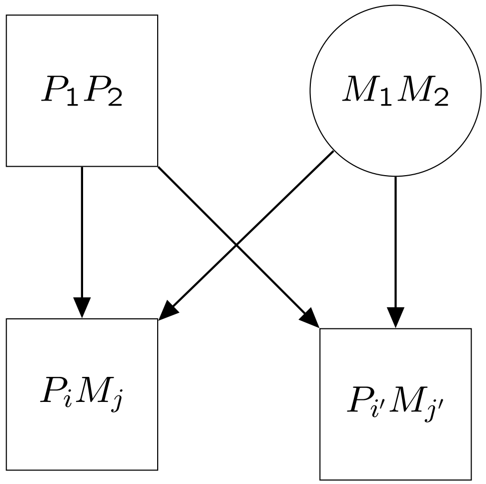
The child will always have exactly 1 allele that is IBD to an allele from a specific parent (the other allele will be IBD to an allele from the other parent).



		Parent 1	
		a	b
Parent 2	c	ac	bc
	d	ad	bd

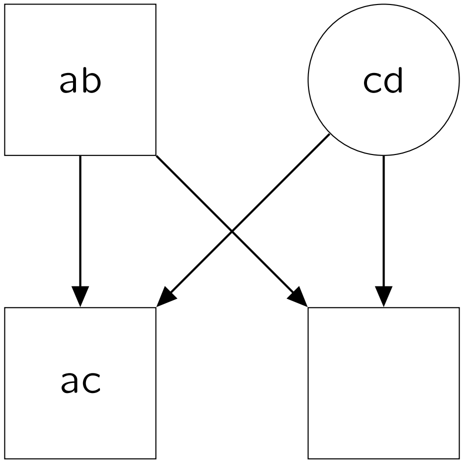
IBD for Sibling Relationships

What about siblings?



IBD for Sibling Relationships

They share either both, one or none of the alleles IBD.

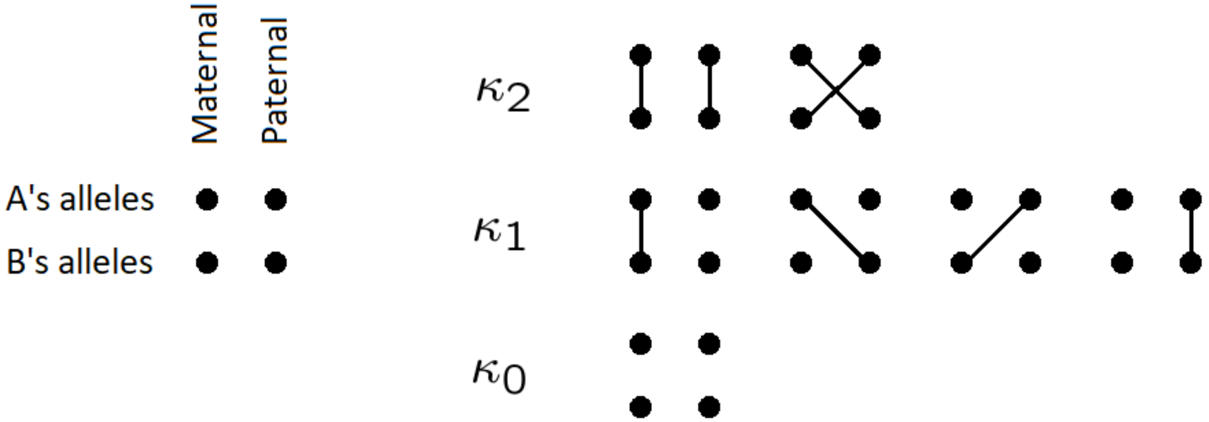


		Alleles IBD		
Sib 1	**Sib 2**	0	1	2
ac	ac	✓		
	bc		✓	
	ad		✓	
	bd			✓
Total		1/4	1/2	1/4

IBD Coefficients

For *non-inbred* relatives, there are three IBD classes. We write κ_i to denote the IBD probabilities:

$$\kappa_i = P(i \text{ alleles IBD})$$



What IBD classes are relevant for unrelated individuals?

IBD Coefficients

For parent-child relationships we saw that:

$$\kappa_1 = P(1 \text{ allele IBD}) = 1, \qquad \text{and} \qquad \kappa_0 = \kappa_2 = 0,$$

while for siblings we have:

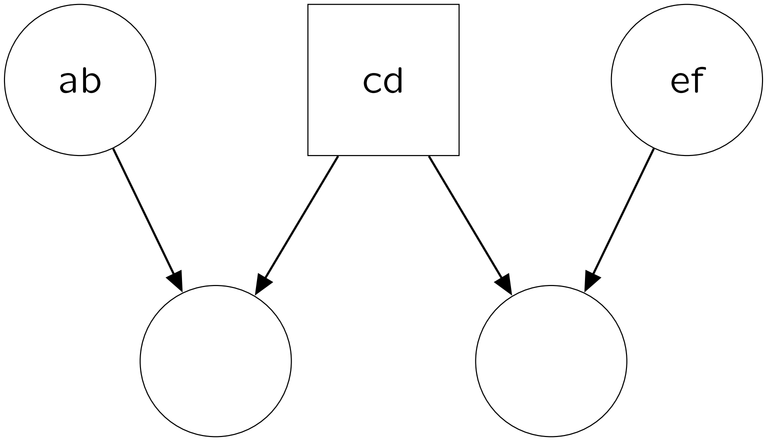
$$\kappa_0 = P(\overline{\text{IBD}_M}) \times P(\overline{\text{IBD}_P}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$\begin{aligned}\kappa_1 &= P(\text{IBD}_M) \times P(\overline{\text{IBD}_P}) + P(\overline{\text{IBD}_M}) \times P(\text{IBD}_P) \\ &= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}\end{aligned}$$

$$\kappa_2 = P(\text{IBD}_M) \times P(\text{IBD}_P) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

IBD Coefficients for Half-sibs

What are the IBD coefficients for half-sibs?



IBD Coefficients

The following table shows IBD probabilities for common relationships:

Relationship	κ_0	κ_1	κ_2
Unrelated	1	0	0
Parent/child	0	1	0
Identical twins	0	0	1
Siblings	1/4	1/2	1/4
Half-sibs	1/2	1/2	0
First cousins	3/4	1/4	0

These IBD probabilities give the expected relatedness between individuals (the realized relatedness is variable).

Match Probabilities for Relatives

If $\kappa_0 = 1$, we are in the original situation and write M_2 for the appropriate match probability:

$$M_2 = \begin{cases} p_A^2, & \text{for homozygous loci } AA, \\ 2p_A p_B, & \text{for heterozygous loci } AB. \end{cases}$$

If $\kappa_1 = 1$, the match probability M_1 changes to:

$$M_1 = \begin{cases} p_A, & \text{for homozygous loci } AA, \\ \frac{1}{2}(p_A + p_B), & \text{for heterozygous loci } AB. \end{cases}$$

If $\kappa_2 = 1$, both alleles are IBD and the match probability is 1.

Match Probabilities for Relatives

Combining the terms leads to the overall single-locus match probability for relatives:

$$\kappa_2 + \kappa_1 M_1 + \kappa_0 M_2,$$

which yields a standard match probability of M_2 for unrelated individuals.

Match Probabilities for Relatives - Exercise

Consider a simple single-source crime scene sample with genotype $G_C = AA$, and a suspect that matches at that locus. Calculate the LR, using $p_A = 4\%$, and alternative hypotheses:

- The DNA in the sample came from an unrelated individual;
- The DNA in the sample came from a half-brother of the suspect;
- The DNA in the sample came from a brother of the suspect;
- The DNA in the sample came from an identical twin of the suspect.

Match Probabilities for Relatives - Exercise

Consider a simple single-source crime scene sample with genotype $G_C = AA$, and a suspect that matches at that locus. Calculate the LR, using $p_A = 4\%$:

- LR $= \frac{P(AA|AA,H_p)}{P(AA|AA,H_d)} = \frac{1}{p_A^2} = 625$;

- LR $= \frac{1}{\kappa_0 M_2 + \kappa_1 M_1 + \kappa_2} = \frac{1}{0.5p_A^2 + 0.5p_A} \approx 48$;

- LR $= \frac{1}{0.25p_A^2 + 0.5p_A + 0.25} \approx 3.7$;

- LR $= 1$.

LRs for Relatives

With this approach we can incorporate specific relatives. But what if no specific alternative is available?

H_d : The DNA in the sample came from an unrelated individual.

H_d : The DNA in the sample came from a brother of the suspect.

H_d : The DNA in the sample came from an unknown individual from the population.

LRs Including Relatives

- We can model a situation where relatives of the suspect make up a small proportion of the total population.

- It is however not trivial to set the number of siblings, uncles/aunts, cousins, etc.

- An overall LR can be calculated by modeling these priors as simple population proportions.

- This requires specifying an average number of children (e.g. using fertility rates) and population size.

The Island Problem

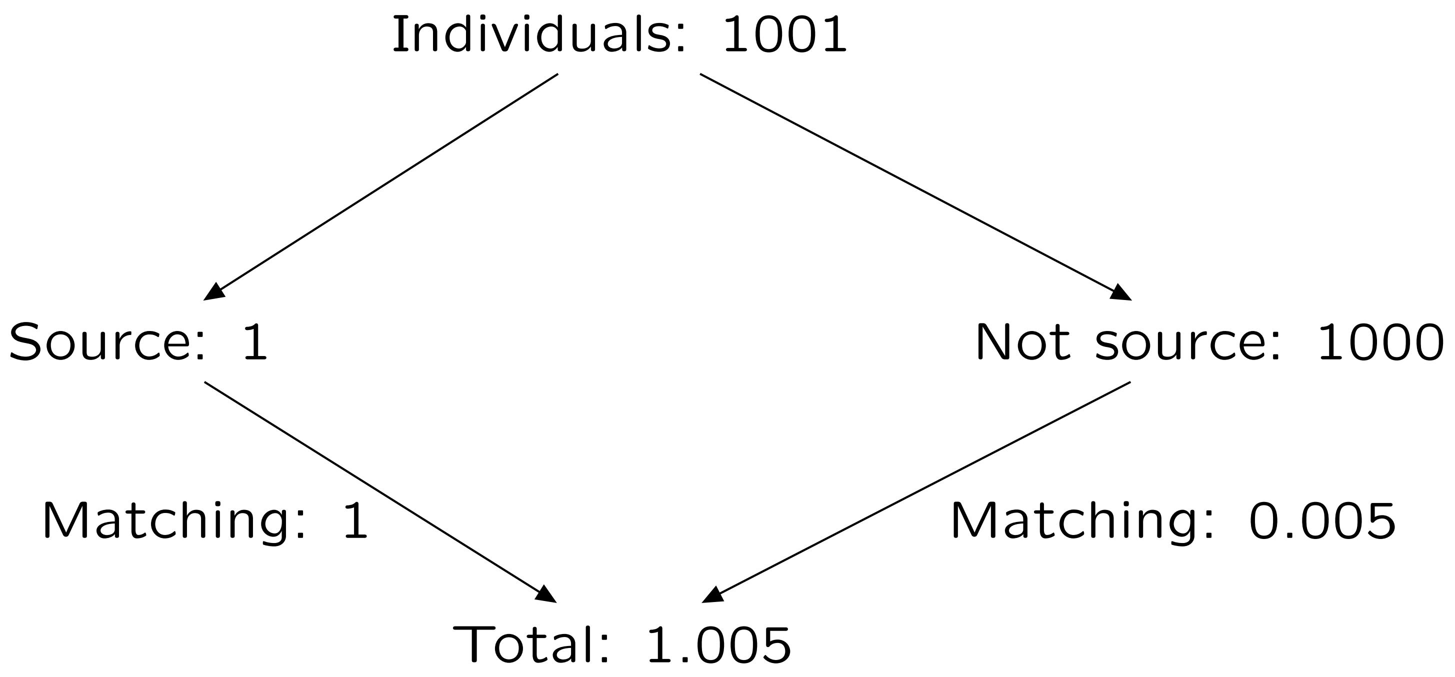
Suppose there is a crime committed on a remote island with a population of size 1001. A suspect Q is found to match the crime scene profile. What is the probability that Q is the source of the profile, assuming that:

- All individuals are equally likely to be the source.
- The DNA profiles of all the other individuals are unknown.
- The match probability for unrelated individuals is 5×10^{-6}.

Source: Weight-of-Evidence for Forensic DNA Profiles (Balding, 2015)

The Island Problem - Solution

Assuming Q has no relatives on the island, there is a $\frac{1}{1.005} \approx 99.5\%$ chance that Q is the source.



Individuals: 1001

Source: 1

Not source: 1000

Matching: 1

Matching: 0.005

Total: 1.005

The Island Problem - Relatives

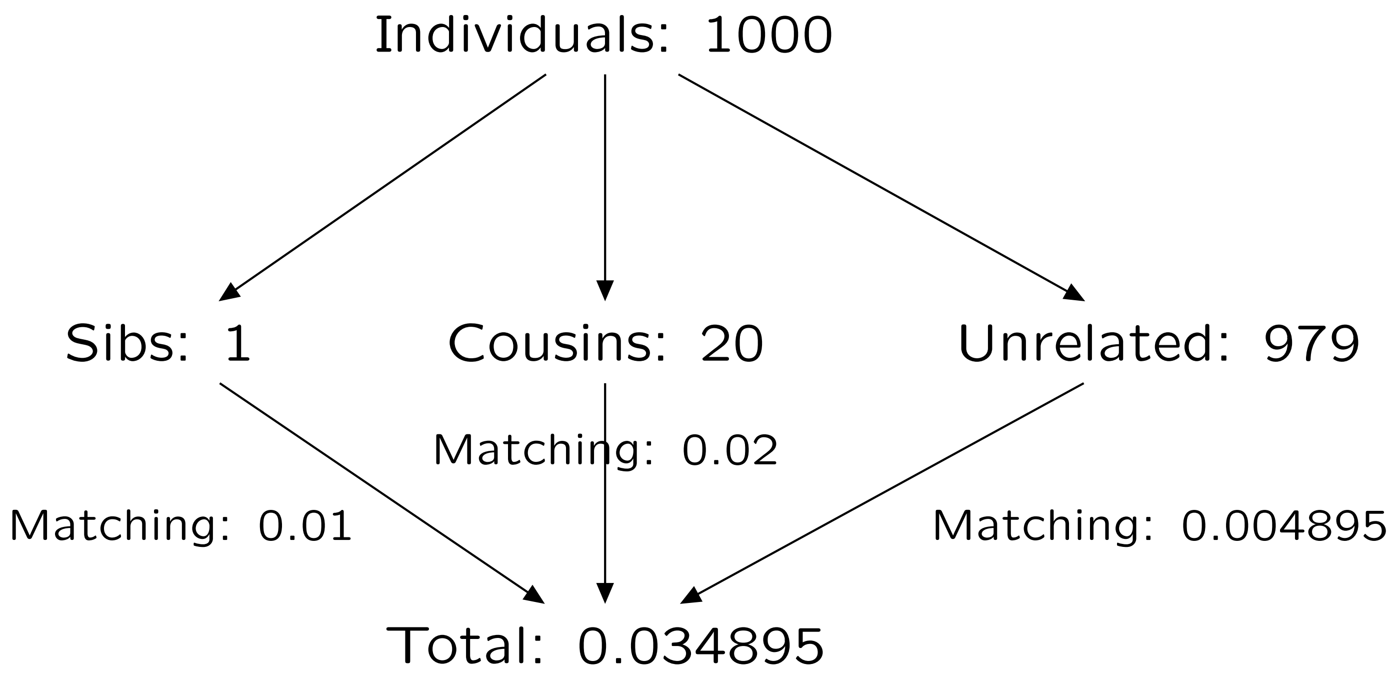
Now suppose that Q has one sibling and 20 cousins on the island, and no other relatives. What is now the probability that Q is the source, using match probabilities of:

- 1 in 1000 for a cousin;
- 1 in 100 for a sibling;
- and 5×10^{-6} for unrelated individuals.

Source: Weight-of-Evidence for Forensic DNA Profiles (Balding, 2015)

The Island Problem - Solution for Relatives

In this case the probability that Q is the source decreases to $\frac{1}{1.034895} \approx 96.6\%$.



The Island Problem - Solution for Relatives

Note how the LR for unrelated individuals ($\text{LR}_U = 200\,000$), the LR for cousins ($\text{LR}_C = 1\,000$), and the LR for siblings ($\text{LR}_S = 100$), can be combined as a weighted average of the match probabilities:

$$\left(\frac{979}{1000} \times 5 \times 10^{-6} + \frac{20}{1000} \times \frac{1}{1000} + \frac{1}{1000} \times \frac{1}{100}\right)^{-1} \approx 28\,650.$$

With prior odds of $\frac{1}{1000}$, the probability that Q is not the source decreases from $\frac{1}{201} \approx 0.5\%$ to $\frac{1}{29.65} \approx 3.4\%$.

What if we were not given any information about the relatives of Q?

The Island Problem - Relatives

What if we were not given any information about the relatives of Q?

In this case, background information may be used to assess plausible values for the priors, specifying the numbers of relatives in each category.

LRs can be calculated for each plausible set of values, and the resulting weight-of-evidence may be averaged over the sets.

In practice, it is often satisfactory to consider only an upper bound on the plausible number of relatives in each category.

Other Applications

The concept of relatedness is important for, and benefits, other applications as well:

- Familial searching
- Paternity testing
- Missing persons
- Inference of ethnicity
- Inference of phenotype

Familial Searching

- A database may be used to compare crime scene profiles to known offenders when investigators lack a suspect.

- A high stringency search requires a full match of the DNA profiles, and might not always return a hit.

- Lowering the search stringency level may lead to a partial match, and has the potential to identify close relatives.

- *Familial searching* refers to the process where investigators look for close relatives in the DNA database in order to open up new investigative leads.

Familial Searching - Case Example

A serial killer nicknamed the Grim Sleeper (due to a 14-year break) was responsible for the death of at least 10 young women in Los Angeles dating back to the mid 1980s.

A search in April 2010 with DNA evidence from one of the crime scene samples showed a potential match with a recently convicted young man. Together with other evidence this led to the suspicion of the father.

The L.A. police was notified by investigators and got a DNA sample from a discarded piece of pizza. Lonnie Franklin was found to match, leading to an arrest in July 2010 and eventual conviction in May 2016.

Familial Searching - Strategies

A certain strategy is required to select a potential relative of the unknown donor from the database. Two general methods are available, both resulting in a ranked list of candidates to investigate further:

- **IBS method**: simply counts the number of shared alleles between two DNA profiles.

- **LR method**: likelihood under two competing hypothesis (als in this context also called a kinship index (KI):

$$\text{KI} = \frac{\sum_{i=0,1,2} P(G_C, G_R|\text{IBD} = i)P(\text{IBD} = i|\text{relationship})}{\sum_{i=0,1,2} P(G_C, G_R|\text{IBD} = i)P(\text{IBD} = i|\text{unrelated})}$$

Familial Searching - Performance

Familial searching is typically focused on parent-child and sibling relationships, as more distant relatives are usually harder to identify and differentiate from unrelated individuals.

The following table shows the performance of the methods using simulated 10-locus profiles in the New Zealand database:

Method	Rank 1 (%)	Rank 1 – 100 (%)
IBS: Siblings	24	72
IBS: Parent-child	8	68
LR: Siblings	31	78
LR: Parent-child	25	99

Source: Effectiveness of familial searches (Curran, 2008).

Familial Searching - Effectiveness

- LR methods outperform the IBS method.
- It is slightly easier to locate parent-child relationships, although siblings more often obtain a number one ranking.
- More loci improve the effectiveness of familial searching, especially in case of extra highly polymorphic loci.
- Ranked lists can be refined based on lineage markers.
- The LR method can be extended to the conditional method, by incorporating priors.

It is important to note that the effectiveness depends on the assumption that a true close relative of the donor is actually present in the database.

Familial Searching - Considerations

Familial searching has proven to be a successful tool in several cases, but it also raises privacy and legal policy concerns:

- Disproportional attention to members of populations that are over-represented in the database.
- False positives may lead to the investigation of innocent people.
- Might reveal the presence of a family member in the database.
- Might reveal the presence of a previously unknown genetic link.
- Might reveal the absence of a genetic link.
- Crimes might go unreported (in case of searches against victim profiles).

Familial Searching - SNPs

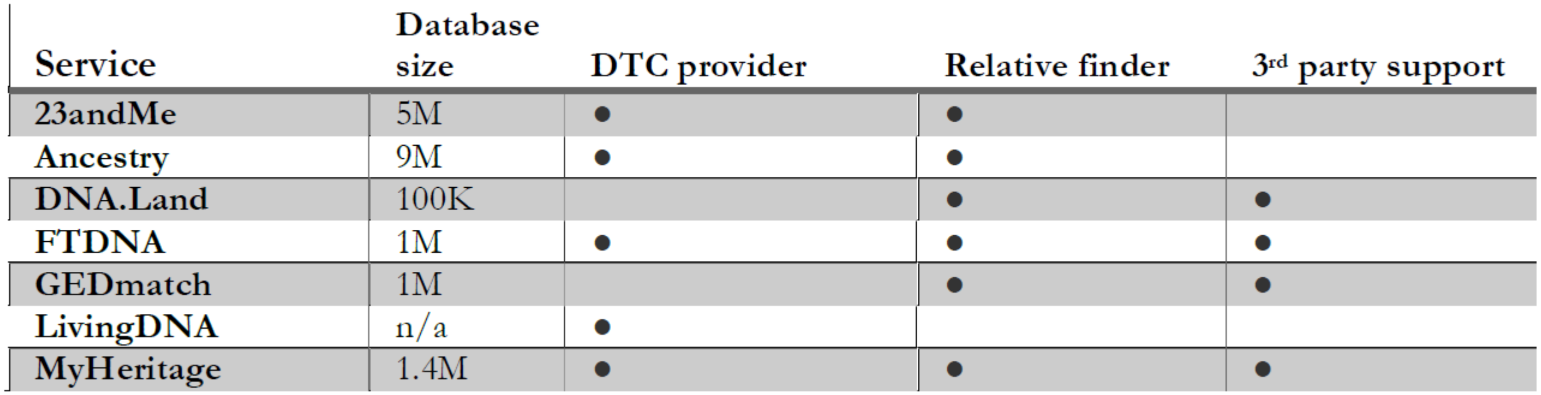
Instead of looking for a (partial) match in one database, it is also possible to combine different databases, even with no overlapping genetic markers. Provided that sufficiently strong LD exists, SNP and STR profiles can be associated with the same individual or distinct but closely related individuals.

Software can be used to infer STR genotypes from a SNP dataset, making it possible to compute match scores for pairs of individuals between databases. This means that CODIS profiles can possibly be connected to a SNP profile, collected for e.g. biomedical or genealogical research, and this cross-database record matching extends to relatives.

Familial Searching - Consumer Genomics Tools

With the emergence of consumer genomics tools, familial searching has become far more powerful. The limited set of STR markers does not allow for finding relatives beyond first and second degree relationships. Furthermore, policies largely restrict or even prohibit the practice completely.

These limitations, however, do not explicitly restrict the use of crime scene samples with civilian DNA databases.



Service	Database size	DTC provider	Relative finder	3rd party support
23andMe	5M	●	●	
Ancestry	9M	●	●	
DNA.Land	100K		●	●
FTDNA	1M	●	●	●
GEDmatch	1M		●	●
LivingDNA	n/a	●		
MyHeritage	1.4M	●	●	●

Source: Re-identification of genomic data using long range familial searches (Erlich, 2018).

Familial Searching - Case Example

To trace the Golden State Killer, crime scene evidence was used to obtain a profile that mimicked the format of regular DTC providers in order to upload it to GEDmatch.

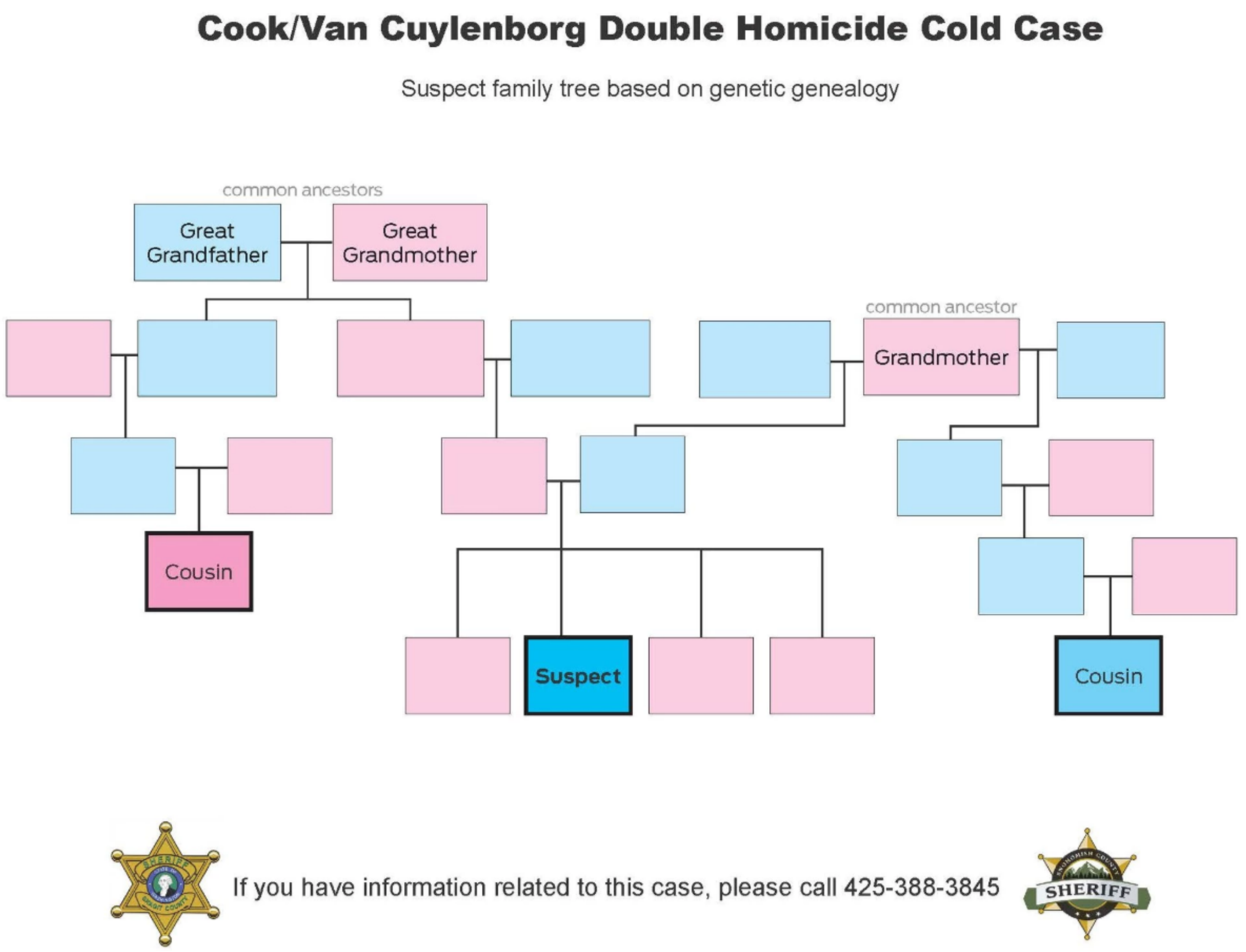
A search, based on IBD matching, identified a third degree cousin of the perpetrator, which eventually led to the arrest of Joseph James DeAngelo. It took five genealogists four months to trace back the identity of the suspected perpetrator.

Familial Searching - Case Example

Even more recently, the Snohomish County sheriff's office announced that they arrested a suspect in the killing of a young couple while they were vacationing in Washington State in 1987.

A GEDmatch led to two second cousins, which could be tied together through a marriage of two descendants from their great-grandparents. The only son from this marriage was investigated further and found to match the crime scene evidence.

Familial Searching - Case Example



Cook/Van Cuylenborg Double Homicide Cold Case

Suspect family tree based on genetic genealogy

If you have information related to this case, please call 425-388-3845

Snohomish County Sheriff's Office

Source: Technique Used to Find Golden State Killer Leads to a Suspect in 1987 Murders (Murphy, 2018).

Paternity Testing

Paternity and familial identification can provide evidence in criminal context and during civil litigation. For a paternity case, the two propositions could be:

H_p: The alleged father (AF) is the true father.

H_d: Some other (unrelated) man is the father.

The likelihood ratio is in this case often referred to as the paternity index (PI).

Paternity Testing - Exercise

Suppose a child has genotype $G_C = AB$. What are the LR values when:

- $G_M = AA$ and $G_{AF} = BB$;
- $G_M = AA$ and $G_{AF} = CD$;
- $G_M = AA$ and $G_{AF} = BC$;
- $G_M = AB$ and $G_{AF} = AA$.

Paternity Testing - Exercise

Suppose a child has genotype $G_C = AB$. The LR values are:

- LR $= \frac{P(G_C{=}AB|G_M{=}AA,G_{AF}{=}BB,H_p)}{P(G_C{=}AB|G_M{=}AA,H_d)} = \frac{1}{p_B}$;
- LR $= \frac{P(G_C{=}AB|G_M{=}AA,G_{AF}{=}CD,H_p)}{P(G_C{=}AB|G_M{=}AA,H_d)} = 0$;
- LR $= \frac{P(G_C{=}AB|G_M{=}AA,G_{AF}{=}BC,H_p)}{P(G_C{=}AB|G_M{=}AA,H_d)} = \frac{\frac{1}{2}}{p_B} = \frac{1}{2p_B}$;
- LR $= \frac{P(G_C{=}AB|G_M{=}AB,G_{AF}{=}AA,H_p)}{P(G_C{=}AB|G_M{=}AA,H_d)} = \frac{\frac{1}{2}}{\frac{1}{2}p_A+\frac{1}{2}p_B} = \frac{1}{p_A+p_B}$.

Paternity Testing - Exercise

Calculate the weight of the evidence for the following data:

Locus	G_C	G_M	G_{AF}
TPOX	(6,9)	(6,12)	(8,9)
vWA	(17,17)	(17,16)	(17,17)
TH01	(7,9)	(9,10)	(7,9)

Locus	**Allele**	**Frequency**
TPOX	6	0.006
	8	0.506
	9	0.094
	12	0.038
vWA	16	0.276
	17	0.300
TH01	7	0.147
	9	0.232
	10	0.116

Source: Introduction to Statistics for Forensic Scientist (Lucy, 2005).

Paternity Testing - Exercise

Calculate the weight of the evidence for the following data:

Locus	G_C	G_M	G_{AF}
TPOX	(6,9)	(6,12)	(8,9)
vWA	(17,17)	(17,16)	(17,17)
TH01	(7,9)	(9,10)	(7,9)

We calculate single-locus LRs and combine these results through multiplication:

- TPOX: $\text{LR} = \frac{0.25}{0.5p_9} = \frac{1}{2\times 0.094} = 5.32$;
- vWA: $\text{LR} = \frac{1}{p_{17}} = \frac{1}{0.3} = 3.33$;
- TH01: $\text{LR} = \frac{0.25}{0.5p_7} = \frac{1}{2\times 0.147} = 3.40$.

Our overall LR is in this case 60.23, yielding evidence in favor of H_p.

Paternity Testing

These cases can be extended to allow for more complex situations:

- Unavailability of the mother;
- Relatedness between the mother and alleged father;
- A relative of the alleged father is the true father;
- Incorporating profiles of (alleged) relatives (e.g. for half-sibs or when alleged father is unavailable);
- Multiple children;
- Incorporating mutations, substructure, silent alleles, non-autosomal DNA, etc.

Missing Persons

The discussed methods for evidence evaluation are also applicable to other situations, such as disaster victim identification and immigration cases.

A comparison must in these cases be carried out between a profile obtained from unidentified remains, or an applicant, and a missing person's profile.

It is, however, often the case that a sample from the missing person is not available, in which case it might be possible to make use of surrogate samples (e.g. obtained through a medical institution).

Alternatively, relatives can be used for testing purposes.

Missing Persons

For a missing person case, the two propositions could be:

H_p: The sample is from the missing person.

H_d: The sample is from some unknown person.

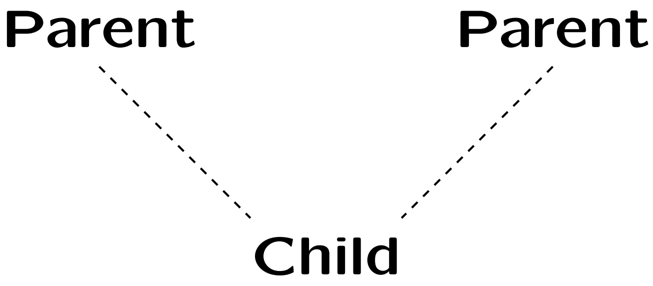
The following likelihood ratios are obtained for a sample with alleged mother (AM) and alleged father (AF), compared to the paternity index, for $p_A = p_B = 0.1$:

(A)M	AF	Sample	LR	Value	PI	Value
AA	BB	AB	$\frac{1}{2p_Ap_B}$	50	$\frac{1}{p_B}$	10
AA	BC	AB	$\frac{1}{4p_Ap_B}$	25	$\frac{1}{2p_B}$	5
AB	AA	AB	$\frac{1}{4p_Ap_B}$	25	$\frac{1}{p_A+p_B}$	5

Source: Interpreting DNA Evidence (Evett & Weir, 1998).

Missing Persons

In the previous case the genetic evidence E consists of the genotype from a sample that has come from some person X who may be the missing person, together with the genotypes from the parents of the missing person.



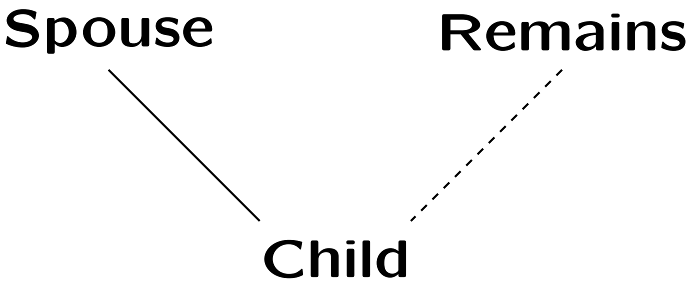
Parent

Parent

Child

If, instead, the genotypes of the spouse S and child C of the missing person are available, the situation is similar to evidence evaluation in case of paternity testing.

Missing Persons



Spouse

Remains

Child

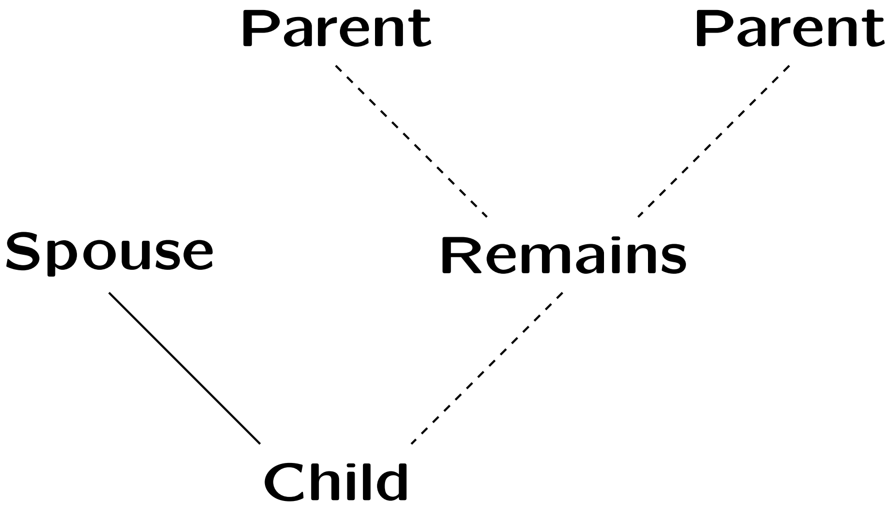
The likelihood ratios are the same as in the paternity case where X is the alleged father of child C who has mother S:

$$\begin{aligned}
\text{LR} &= \frac{P(E|H_p)}{P(E|H_d)} \\
&= \frac{P(G_C, G_S, G_X|H_p)}{P(G_C, G_S, G_X|H_d)} \\
&= \frac{P(G_C|G_S, G_X, H_p)P(G_S, G_X|H_p)}{P(G_C|G_S, G_X, H_d)P(G_S, G_X|H_d)} \\
&= \frac{P(G_C|G_S, G_X, H_p)}{P(G_C|G_S, H_d)}
\end{aligned}$$

Missing Persons

It may be the case that people apart from the spouse and child of the missing person are typed. The general procedure is the same: the probabilities of the set of observed genotypes under two explanations are compared.

Suppose the parents P and Q as well as the child C and spouse S of the missing person are typed, and that a sample is available that has come from some person X thought under H_p to be the missing person.



Missing Persons

Under explanation H_d, the sample from X did not come from the missing person, and therefore the genotype of X does not depend on the genotypes of P and Q and the genotype of C does not depend on the genotype of X.

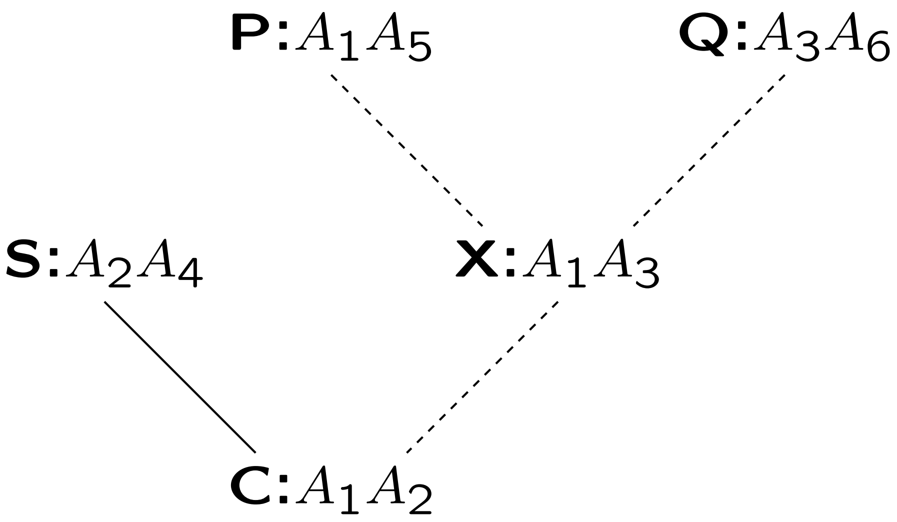
The likelihood ratio is arranged to involve probabilities of genotypes conditional on previous generations. If both parents of an individual have been typed, there is no need to condition on the grandparents of that individual.

In the following slides, C, S, X, P and Q represent the genotypes of the child, the remains, the spouse and the parents of the missing person.

Missing Persons

$$\begin{aligned}
\text{LR} &= \frac{P(E|H_p)}{P(E|H_d)} \\
&= \frac{P(C,S,X,P,Q|H_p)}{P(C,S,P,X,Q|H_d)} \\
&= \frac{P(C|S,X,P,Q,H_p)P(S,X,P,Q|H_p)}{P(C|S,X,P,Q,H_d)P(S,X,P,Q|H_d)} \\
&= \frac{P(C|S,X,H_p)P(S,X|P,Q,H_p)P(P,Q|H_p)}{P(C|S,P,Q,H_d)P(S,X|P,Q,H_d)P(P,Q|H_p)} \\
&= \frac{P(C|S,X,H_p)P(S|H_p)P(X|P,Q,H_p)}{P(C|S,P,Q,H_d)P(S|H_d)P(X|H_d)} \\
&= \frac{P(C|S,X,H_p)P(X|P,Q,H_p)}{P(C|S,P,Q,H_d)P(X|H_d)}
\end{aligned}$$

Missing Persons - Example



P:A_1A_5

Q:A_3A_6

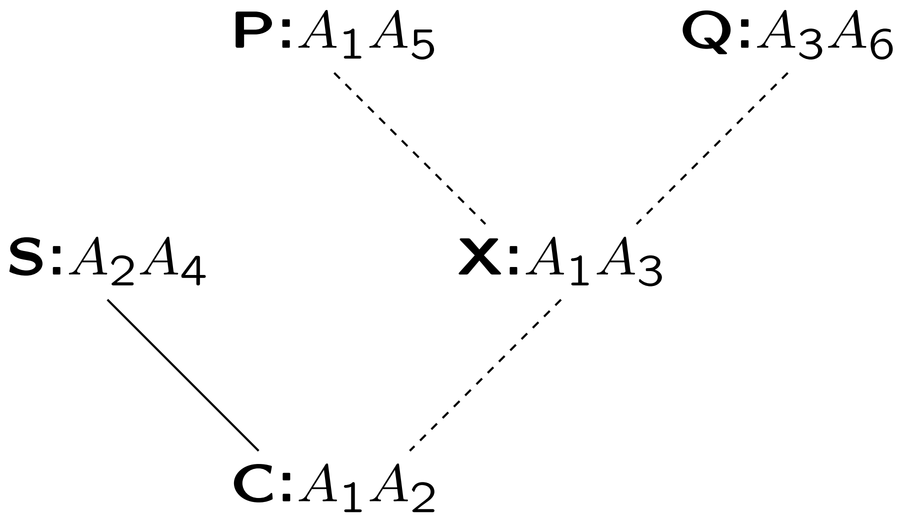
S:A_2A_4

X:A_1A_3

C:A_1A_2

$$\text{LR} = \frac{P(C|S, X, H_p)P(X|P, Q, H_p)}{P(C|S, P, Q, H_d)P(X|H_d)}$$

Missing Persons - Example



P:A_1A_5 **Q**:A_3A_6

S:A_2A_4 **X**:A_1A_3

C:A_1A_2

$$\begin{aligned}
P(C|S,X,H_p) &= 1/4\\
P(X|P,Q,H_p) &= 1/4\\
P(C|S,P,Q,H_d) &= 1/8\\
P(X|H_d) &= 2p_1p_3\\
\text{LR} &= \frac{1}{4p_1p_3}
\end{aligned}$$

Inference of Ethnicity

Suppose that a population can be classified into K groups. The probability of a DNA sample with profile D coming from group k, can be written as:

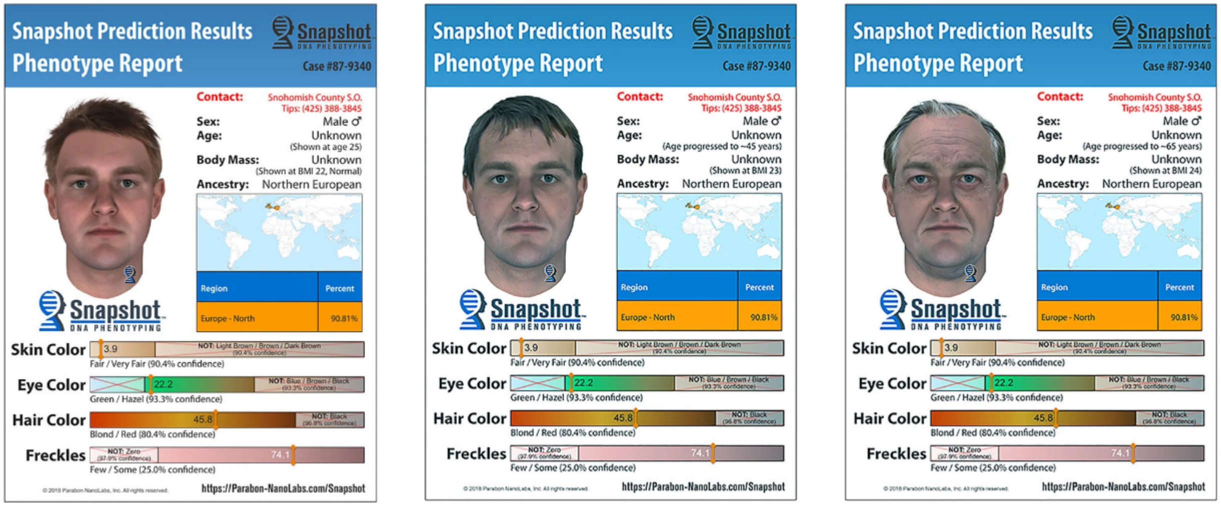
$$P(\text{group } k|D) = \frac{P(D|\text{group } k)P(\text{group } k)}{\sum_{j=1}^{K} P(D|\text{group } j)P(\text{group } j)}.$$

STR profiles can give some information, although they provide limited discriminatory power in this context. Instead, SNP sets (so-called ancestry informative markers) have been demonstrated to be useful for distinguishing individuals from certain (sub-)populations.

Inference of Phenotype

SNPs may be linked to some visual phenotypes, including hair color and eye color. Other facial characteristics can now also be predicted from genotypes with some accuracy.

These SNP associations can potentially be used in forensic settings, e.g. in combination with a description of an eyewitness of a target individual.



Snapshot Prediction Results
Phenotype Report
Case #87-9340

Contact: Snohomish County S.O.
Tips: (425) 388-3845
Sex: Male ♂
Age: Unknown
(Shown at age 25)
Body Mass: Unknown
(Shown at BMI 22, Normal)
Ancestry: Northern European

Region	Percent
Europe - North	90.81%

Skin Color: 3.9 — NOT: Light Brown / Brown / Dark Brown (90.4% confidence) — Fair / Very Fair (90.4% confidence)
Eye Color: 22.2 — NOT: Blue / Brown / Black (93.3% confidence) — Green / Hazel (93.3% confidence)
Hair Color: 45.8 — NOT: Black (96.8% confidence) — Blond / Red (80.4% confidence)
Freckles: 74.1 — NOT: Zero (97.9% confidence) — Few / Some (25.0% confidence)

© 2018 Parabon NanoLabs, Inc. All rights reserved. https://Parabon-NanoLabs.com/Snapshot

Snapshot Prediction Results
Phenotype Report
Case #87-9340

Contact: Snohomish County S.O.
Tips: (425) 388-3845
Sex: Male ♂
Age: Unknown
(Age progressed to ~45 years)
Body Mass: Unknown
(Shown at BMI 23)
Ancestry: Northern European

Region	Percent
Europe - North	90.81%

Skin Color: 3.9 — NOT: Light Brown / Brown / Dark Brown (90.4% confidence) — Fair / Very Fair (90.4% confidence)
Eye Color: 22.2 — NOT: Blue / Brown / Black (93.3% confidence) — Green / Hazel (93.3% confidence)
Hair Color: 45.8 — NOT: Black (96.8% confidence) — Blond / Red (80.4% confidence)
Freckles: 74.1 — NOT: Zero (97.9% confidence) — Few / Some (25.0% confidence)

© 2018 Parabon NanoLabs, Inc. All rights reserved. https://Parabon-NanoLabs.com/Snapshot

Snapshot Prediction Results
Phenotype Report
Case #87-9340

Contact: Snohomish County S.O.
Tips: (425) 388-3845
Sex: Male ♂
Age: Unknown
(Age progressed to ~65 years)
Body Mass: Unknown
(Shown at BMI 24)
Ancestry: Northern European

Region	Percent
Europe - North	90.81%

Skin Color: 3.9 — NOT: Light Brown / Brown / Dark Brown (90.4% confidence) — Fair / Very Fair (90.4% confidence)
Eye Color: 22.2 — NOT: Blue / Brown / Black (93.3% confidence) — Green / Hazel (93.3% confidence)
Hair Color: 45.8 — NOT: Black (96.8% confidence) — Blond / Red (80.4% confidence)
Freckles: 74.1 — NOT: Zero (97.9% confidence) — Few / Some (25.0% confidence)

© 2018 Parabon NanoLabs, Inc. All rights reserved. https://Parabon-NanoLabs.com/Snapshot

Picture rendered by Parabon Nanolabs.

Source: Technique Used to Find Golden State Killer Leads to a Suspect in 1987 Murders (Murphy, 2018).

Section 9:
Profile and Match Probabilities; CPI/CPE

Balding's Sampling Formula

if we have examined n alleles, and have seen n_A of type A, what is the probability the next allele is type A?

$$\Pr(A|n_A, n) = \frac{n_A\theta + (1-\theta)p_A}{1 + (n-1)\theta}$$

This implies the result for seeing a previously-unseen allele type B:

$$\Pr(B|n_B = 0, n) = \frac{(1-\theta)p_B}{1 + (n-1)\theta}$$

Examples of Balding's Formula

n	n_A	$\text{Pr}(A\vert n_A, n)$
0	0	p_A
1	0	$(1-\theta)p_A$
	1	$\theta + (1-\theta)p_A$
2	0	$(1-\theta)p_A/(1+\theta)$
	1	$[\theta + (1-\theta)p_A]/(1+\theta)$
	2	$[2\theta + (1-\theta)p_A]/(1+\theta)$
3	0	$(1-\theta)p_A/(1+2\theta)$
	1	$[\theta + (1-\theta)p_A]/(1+2\theta)$
	2	$[2\theta + (1-\theta)p_A]/(1+2\theta)$
	3	$[3\theta + (1-\theta)p_A]/(1+2\theta)$

Match Probability

Balding's formula lets the genotype match probabilities be found very easily from the third law of probability:

$$\begin{aligned}\Pr(AA|AA) &= \Pr(A|AA)\Pr(A|AAA)\\ &= \frac{2\theta+(1-\theta)p_A}{1+\theta}\times\frac{3\theta+(1-\theta)p_A}{1+2\theta}\end{aligned}$$

$$\begin{aligned}\Pr(AB|AB) &= \Pr(B|AB)\Pr(A|ABB)+\Pr(A|AB)\Pr(B|AAB)\\ &= \frac{2[\theta+(1-\theta)p_A][\theta+(1-\theta)p_B]}{(1+\theta)(1+2\theta)}\end{aligned}$$

Paternity Calculation

Balding's formula also lets paternity calculations be done very easily. In the case where the mother, child and alleged father are all homozygous AA, the paternity index is

$$\begin{aligned}
\text{LR} &= \frac{\Pr(\text{M}, \text{C}, \text{AF}|\text{AF is father})}{\Pr(\text{M.C.AF}|\text{AF not father})} \\
&= \frac{\Pr(\text{C}|\text{M}, \text{AF})\Pr(\text{M}, \text{AF})}{\Pr(\text{C}|\text{M})\Pr(\text{M}, \text{AF})} \\
&= \frac{1}{\Pr(A|AAAA)} \\
&= \frac{(1+3\theta)}{4\theta+(1-\theta)p_A}
\end{aligned}$$

The paternal allele is A, and four A alleles have been seen already.

Profile Probabilities

For a single autosmal locus, the probability a random person has genotypes AA or AB is written as Pr(AA) or Pr(BB).

If Hardy-Weinberg Equilibrium is assumed (NRC 4.1a,b):

$$\begin{aligned}\Pr(AA) &= p_A^2 \\ \Pr(AB) &= 2p_A p_B\end{aligned}$$

If a random individual has probability F of being inbred, then the probabilities become (NRC 4.2a,b):

$$\begin{aligned}\Pr(AA) &= p_A^2 + p_A(1-p_A)F \\ \Pr(AB) &= 2p_A p_B - 2p_A p_B F\end{aligned}$$

The probability of a homozygote is greater than the HWE value, and the probability of a heterozygote is less than the HWE value. Here F is the pedigree-value that follows from the path-counting method and it is greater than zero. p_A, p_B are the total population allele frequencies as can be estimated from a database.

NRC Equation

The National Research Council recommended using

$$\begin{aligned}\mathrm{Pr}(AA) &= p_A^2 + p_A(1-p_A)F \\ \mathrm{Pr}(AB) &= 2p_Ap_B\end{aligned}$$

in the interest of being conservative.

Single-allele Profile

An STR profile may show only one allele A at a locus. The true genotype may be homozygous AA or heterozygous AB where allele B is not detected or not called. The HWE probability for the profile allele is

$$\begin{aligned} \Pr(A) &= \Pr(AA) + \sum_{B \neq A} \Pr(AB) \\ &= p_A^2 + \sum_{B \neq A} 2p_A p_B \\ &= p_A^2 + 2p_A(1 - p_A) \\ &= 2p_A - p_A^2 \end{aligned}$$

The "2p" rule approximates this by the conservative value $2p_A$ (NRC Page 105).

Single-allele Profile

For inbred individuals, the value would be

$$\begin{aligned}
\text{Pr}(A) &= \text{Pr}(AA) + \sum_{B\neq A} \text{Pr}(AB) \\
&= p_A^2 + p_A(1-p_A)F + \sum_{B\neq A} 2p_Ap_B(1-F) \\
&= p_A^2 + p_A(1-p_A)F + 2p_A(1-p_A) - 2p_A(1-p_A)F \\
&= 2p_A - p_A^2 - p_A(1-p_A)F
\end{aligned}$$

which also has $2p_A$ as a (conservative) upper bound.

Match Probability for Relatives

For unilineal relatives, $k_2 = 0, k_1 > 0 : \theta = k_1/4$: (NRC 4.8a,b)

$$\begin{aligned}
\Pr(AA|AA) &= \frac{\Pr(AAAA)}{\Pr(AA)} = \frac{k_2 p_A^2 + k_1 p_A^3 + k_0 p_A^4}{p_A^2} \\
&= 4\theta p_A + (1-4\theta)p_A^2 \\
&= p_A^2 + 4p_A(1-p_A)\theta \\
\Pr(AB|AB) &= \frac{\Pr(ABAB)}{\Pr(AB)} = \frac{2k_2 p_A p_B + k_1 p_A p_B(p_A + p_B) + 4k_0 p_A^2 p_B^2}{2p_A p_B} \\
&= 2\theta(p_A + p_B) + 2(1-4\theta)p_A p_B \\
&= 2p_A p_B + 2(p_A + p_B - 4p_A p_B)\theta
\end{aligned}$$

Match Probability for Full Sibs

For full sibs, $k_2 = 1/4, k_1 = 1/2, k_0 = 1/4$: (NRC 4.9a,b)

$$\begin{aligned}
\Pr(AA|AA) &= \frac{k_2 p_A^2 + k_1 p_A^3 + k_0 p_A^4}{p_A^2} \\
&= \frac{1}{4}(1 + 2p_A + p_A^2) \\
\Pr(AB|AB) &= \frac{2k_2 p_A p_B + k_1 p_A p_B (p_A + p_B) + 4k_0 p_A^2 p_B^2}{2p_A p_B} \\
&= \frac{1}{4}(1 + p_A + p_B + 2p_A p_B)
\end{aligned}$$

Probability of Exclusion

The Principles of Evidence Interpretation, leading to the likelihood ratio for the probabilities of the evidence under alternative hypotheses, allow all situations to be addressed. The prosecution and defense perspectives are explicitly taken into account.

The probability of exclusion considers only the evidence profile and ignores prosecution and defense perspective. It does not inform the court.

For a single-contributor stain with genotype AA, anyone not of that type is excluded. The probability of exclusion is $(1 - p_A^2)$. For type AB the probability is $(1 - 2p_Ap_B)$. If many loci are typed, the combined probability of exclusion is the probability a person is excluded for at least one locus - i.e. one minus the probability of no exclusions:

$$\text{CPI} \;=\; 1 - \prod_{\text{loci}\,l} [1 - \text{Pr(Excluded at locus } l)]$$

Exclusion for Mixtures

The Probability of Exclusion understates the strength of mixture evidence. If a crime stain is observed to have four alleles A, B, C, D at a locus the probability of exclusion is $\text{Pr}(AA) + \text{Pr}(BB) + \text{Pr}(CC) + \text{Pr}(DD) + \text{Pr}(AB) + \text{Pr}(AC) + \text{Pr}(AD) + \text{Pr}(BS) + \text{Pr}(BD) + \text{Pr}(CD)$. This is $(p_A + p_B + p_C + p_D)^2$.

If the prosecution says the evidence represents the victim of type AB and the defendant of type CD then the evidence has probability of 1.

If the defense says the evidence (e.g. bedding) is not associated with either the victim or the defendant then the evidence has probability $\text{Pr}(AB, CD) + \text{Pr}(AC, BD) + \text{Pr}(AD, BC) = 24 p_A p_B p_C p_D$.

If all allele frequencies are 0.1, the PE is $0.4^2 = 0.16$ ("1 in 6") and the LR is $1/0.0024 = 416$.

Will match probabilities keep decreasing?

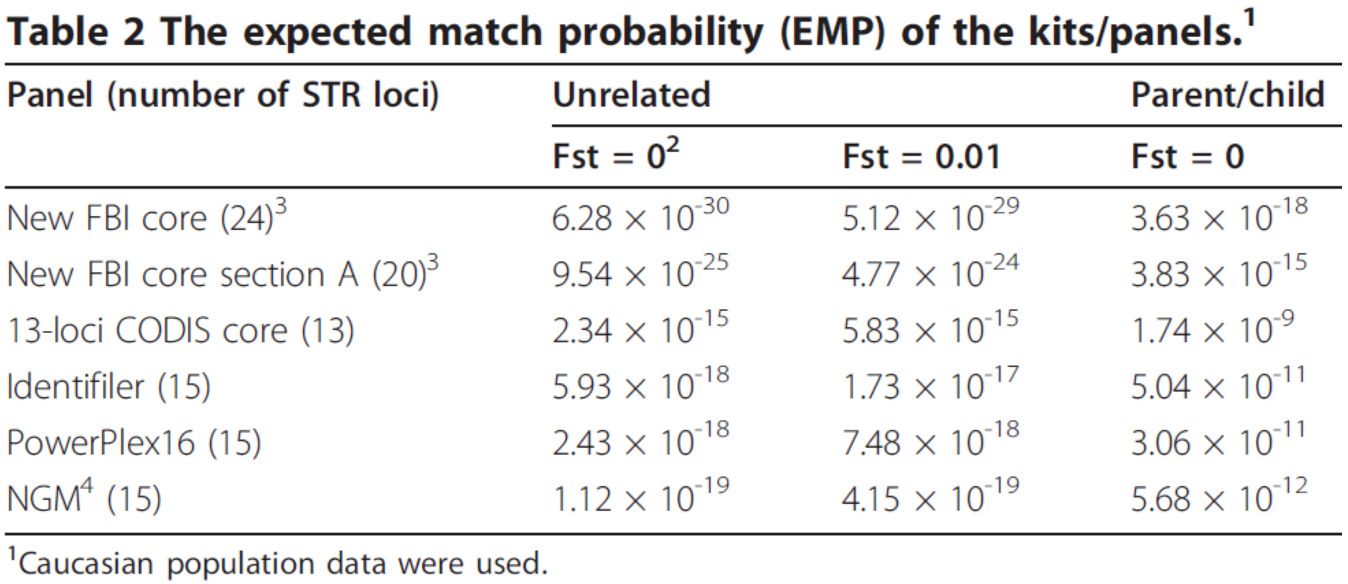


Table 2 The expected match probability (EMP) of the kits/panels.[1]

Panel (number of STR loci)	Unrelated		Parent/child
	Fst = 0[2]	Fst = 0.01	Fst = 0
New FBI core (24)[3]	6.28×10^{-30}	5.12×10^{-29}	3.63×10^{-18}
New FBI core section A (20)[3]	9.54×10^{-25}	4.77×10^{-24}	3.83×10^{-15}
13-loci CODIS core (13)	2.34×10^{-15}	5.83×10^{-15}	1.74×10^{-9}
Identifiler (15)	5.93×10^{-18}	1.73×10^{-17}	5.04×10^{-11}
PowerPlex16 (15)	2.43×10^{-18}	7.48×10^{-18}	3.06×10^{-11}
NGM[4] (15)	1.12×10^{-19}	4.15×10^{-19}	5.68×10^{-12}

[1] Caucasian population data were used.

Ge et al, Investigative Genetics 3:1-14, 2012.

Will match probabilities keep decreasing?

How do these match probabilities address the observation of Donnelly:

> "after the observation of matches at some loci, it is relatively much more likely that the individuals involved are related (precisely because matches between unrelated individuals are unusual) in which case matches observed at subsequent loci will be less surprising. That is, knowledge of matches at some loci will increase the chances of matches at subsequent loci, in contrast to the independence assumption."

Donnelly P. 1995. Heredity 75:26-64.

Are match probabilities independent over loci?

Is the problem that we keep on multiplying match probabilities over loci under the assumption they are independent? Can we even test that assumption for 10 or more loci?

Or is our standard "random match probability" not the appropriate statistic to be reporting in casework? Is it actually appropriate to report statements such as

> The approximate incidence of this profile is 1 in 810 quintillion Caucasians, 1 in 4.9 sextillion African Americans and 1 in 410 quadrillion Hispanics.

Putting “match” back in “match probability”

Let’s reserve “match” for a statement we make about two profiles and take “match probability” to mean the probability that *two profiles match*. This requires calculations about *pairs of profiles*.

If the source of an evidence profile is unknown (e.g. is not the person of interest), then the match probability is the probability this unknown person has the profile *already seen in the POI.* No two profiles are truly independent, and their dependence affects match probabilities across loci.

Likelihood ratios use match probabilities

As with many other issues on forensic genetics, the issue of multi-locus match probability dependencies is best addressed by comparing the probabilities of the evidence under alternative propositions:

> H_p: the person of interest is the source of the evidence DNA profile.
> H_d: an unknown person is the source of the evidence DNA profile.

Write the profiles of the POI and the source of the evidence as G_s and G_c. The evidence is the pair of profiles G_c, G_c.

Likelihood ratios use match probabilities

The likelihood ratio is

$$\begin{aligned}\mathrm{LR} &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)}\\ &= \frac{\Pr(G_c, G_s|H_p)}{\Pr(G_c, G_s|H_d)}\\ &= \frac{1}{\Pr(G_c|G_s, H_d)}\\ &= \frac{1}{\text{Match probability}}\end{aligned}$$

providing $G_c = G_s$ under H_p. The match probability is the chance an unknown person has the evidence profile given that the POI has the profile: this is not the profile probability.

Special Cases: Use of Sample Allele Frequencies

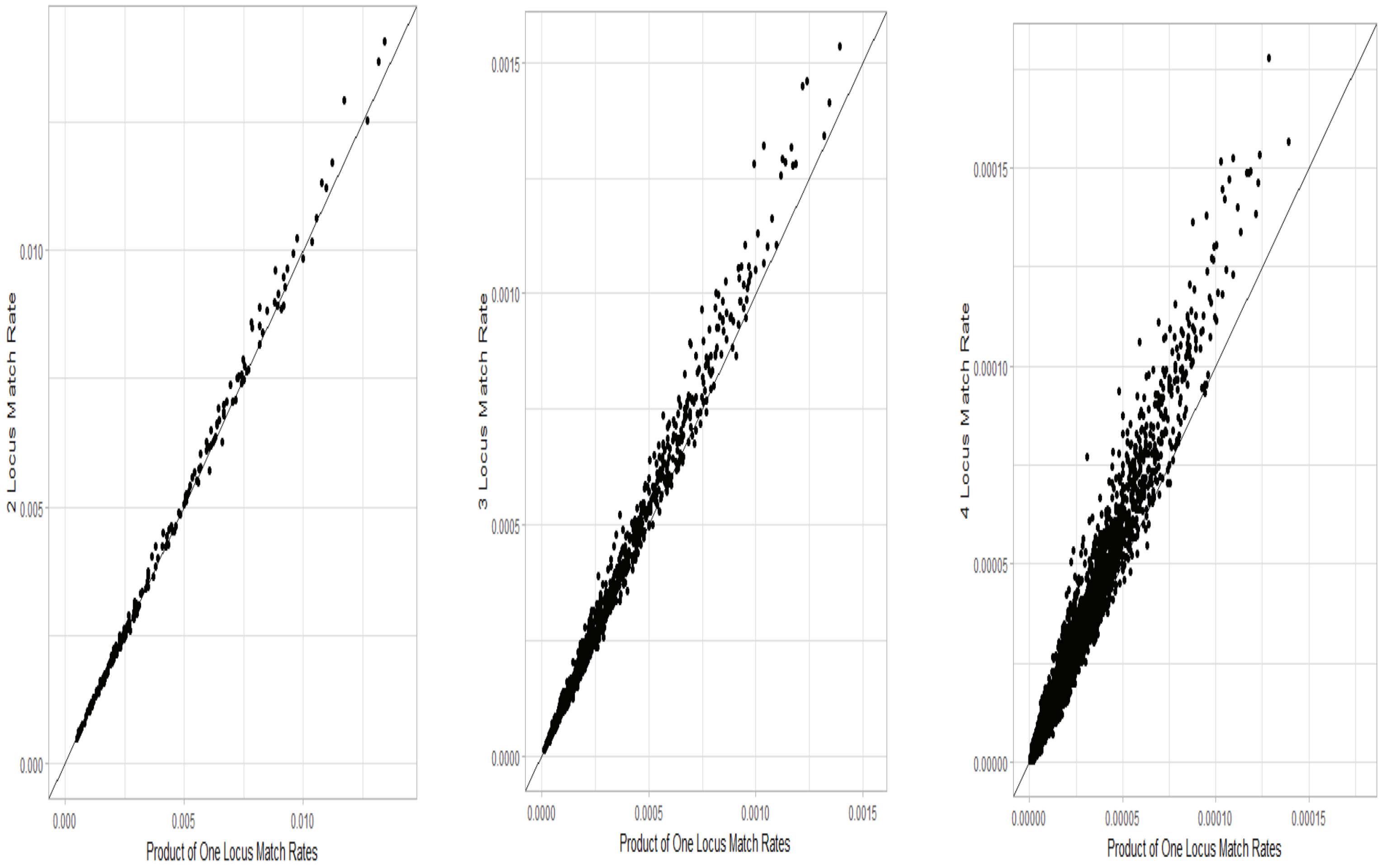
The match probability is usually estimated using allele frequencies from a database representing some broad class of people, such as "Caucasian" or "African American" or "Hispanic."

The population relevant for a particular crime may be a narrower class of people. There is population structure. If p are the allele frequencies in the database, the match probabilities are estimated as

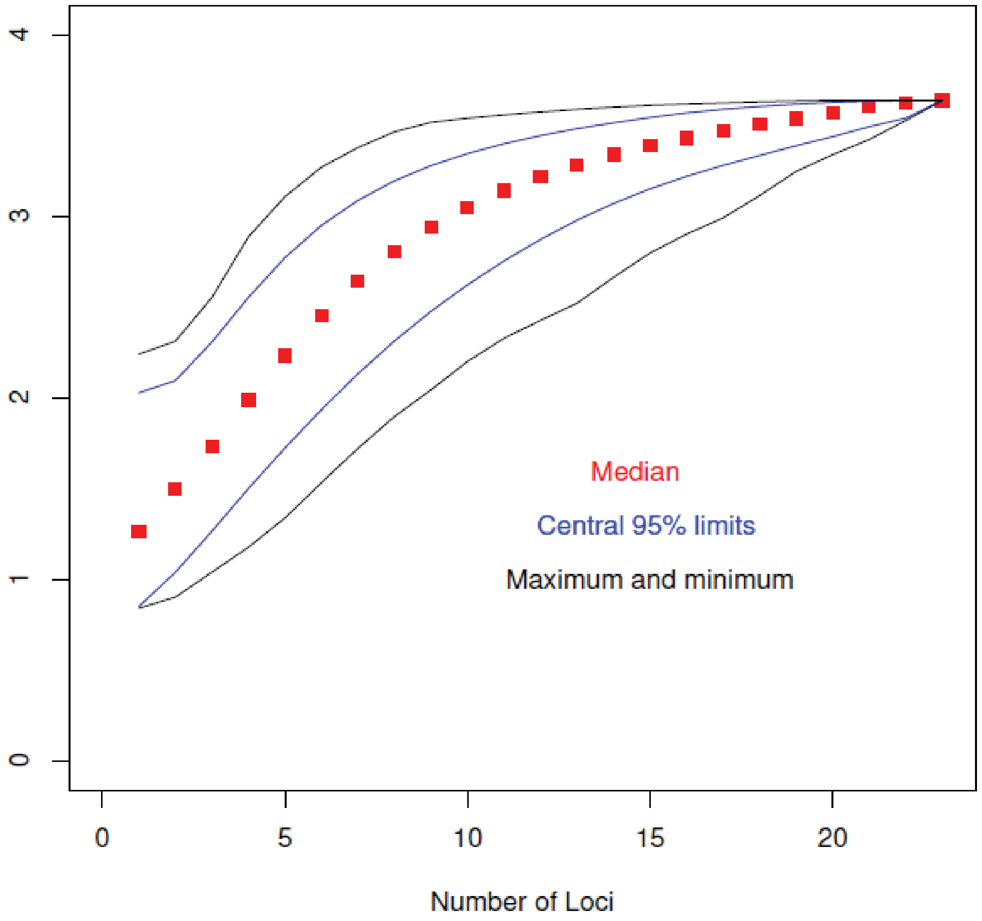
$$\begin{aligned}\mathrm{Pr}(AA|AA) &= \frac{[3\theta+(1-\theta)p_A][2\theta+(1-\theta)p_A]}{(1+\theta)(1+2\theta)}\\ \mathrm{Pr}(AB|AB) &= \frac{2[\theta+(1-\theta)p_A][\theta+(1-\theta)p_B]}{(1+\theta)(1+2\theta)}\end{aligned}$$

Can these be multiplied over loci?

Empirical dependencies: 2849 20-locus profiles



Empirical dependencies: Y-STR profiles



Plot of negative log of match probabilities for YHRD database.

Theoretical dependencies: No mutation

The probability an individual is homozygous $AABB$ at loci **A,B** is

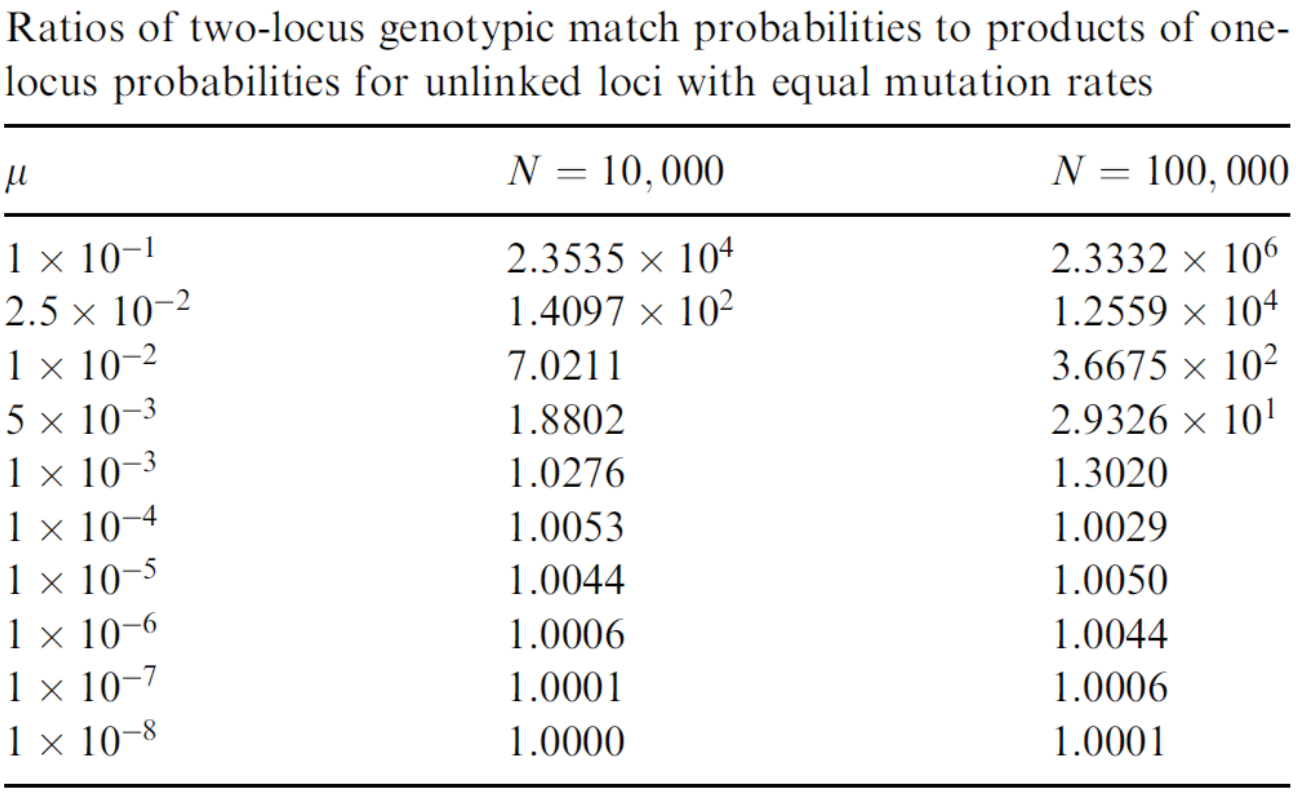
$$\begin{aligned} \text{Pr}(AABB) &= \text{Pr}(AA)\,\text{Pr}(BB) + p_A(1-p_A)p_B(1-p_B)\eta \\ &\geq \text{Pr}(AA)\,\text{Pr}(BB) \end{aligned}$$

where η is the *identity disequilibrium*. It can non-zero even for pairs of loci that are unlinked and/or in linkage equilibrium.

> Sampling among parents or gametes and/or the inclusion of random elements in the uniting gametes leads to a correlation in identity by descent even between unlinked loci because genes at both loci are of necessity included in each gamete.

Weir & Cockerham, Genetics 63:711-742, 1969.

Theoretical dependencies: Mutation



Ratios of two-locus genotypic match probabilities to products of one-locus probabilities for unlinked loci with equal mutation rates

μ	$N = 10,000$	$N = 100,000$
1×10^{-1}	2.3535×10^{4}	2.3332×10^{6}
2.5×10^{-2}	1.4097×10^{2}	1.2559×10^{4}
1×10^{-2}	7.0211	3.6675×10^{2}
5×10^{-3}	1.8802	2.9326×10^{1}
1×10^{-3}	1.0276	1.3020
1×10^{-4}	1.0053	1.0029
1×10^{-5}	1.0044	1.0050
1×10^{-6}	1.0006	1.0044
1×10^{-7}	1.0001	1.0006
1×10^{-8}	1.0000	1.0001

Laurie CA, Weir BS. 2003. Theoretical Population Biology 63:207-219.

Theoretical dependencies: Mutation

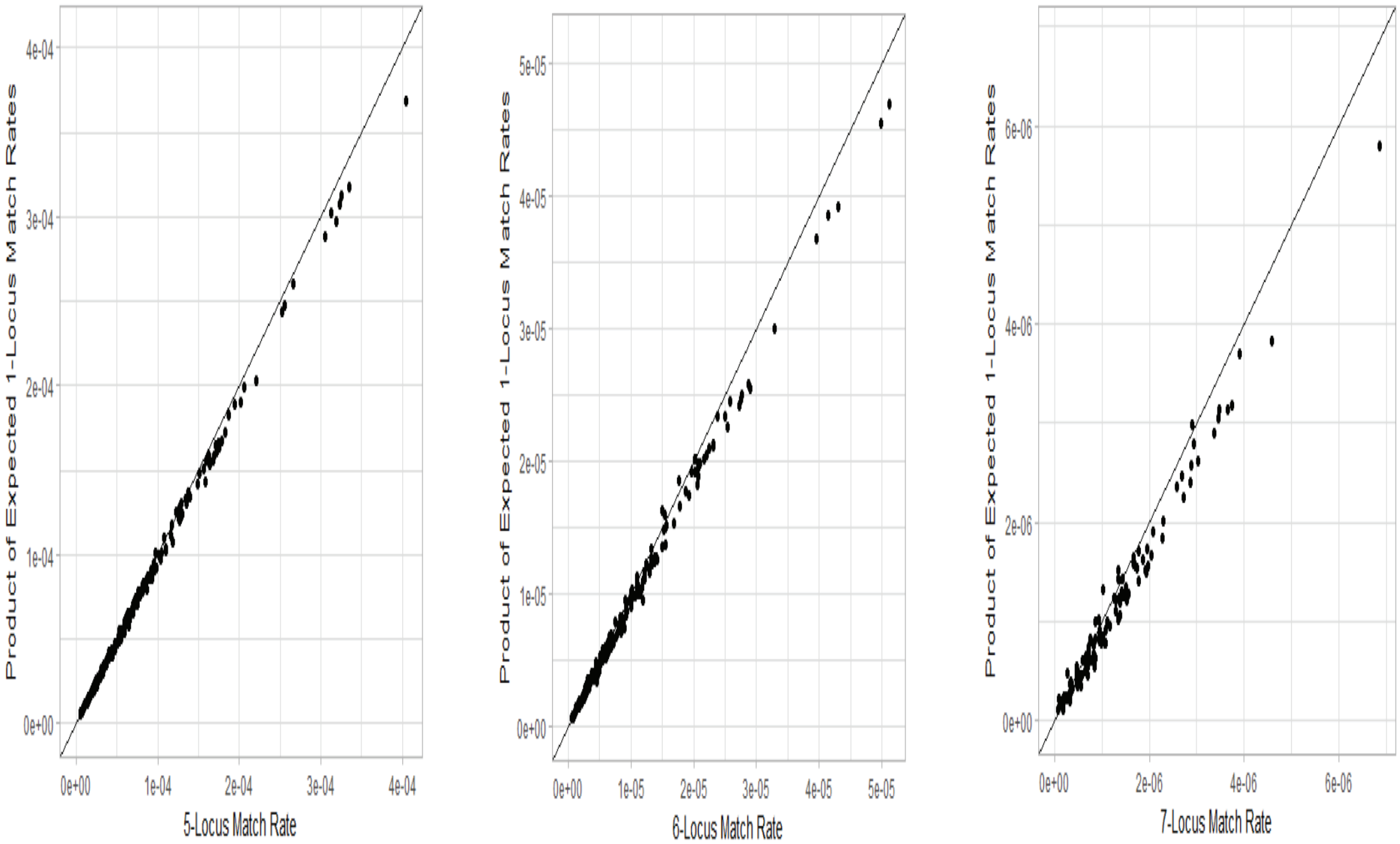
"Between-locus dependencies in finite populations can lead to under-estimates of genotypic match probabilities when using the product rule, even for unlinked loci.

The three-locus ratio is greater than one and is greater than the corresponding two-locus ratio for large mutation rates. These results provide evidence that between-locus dependency effects are magnified when considering more loci.

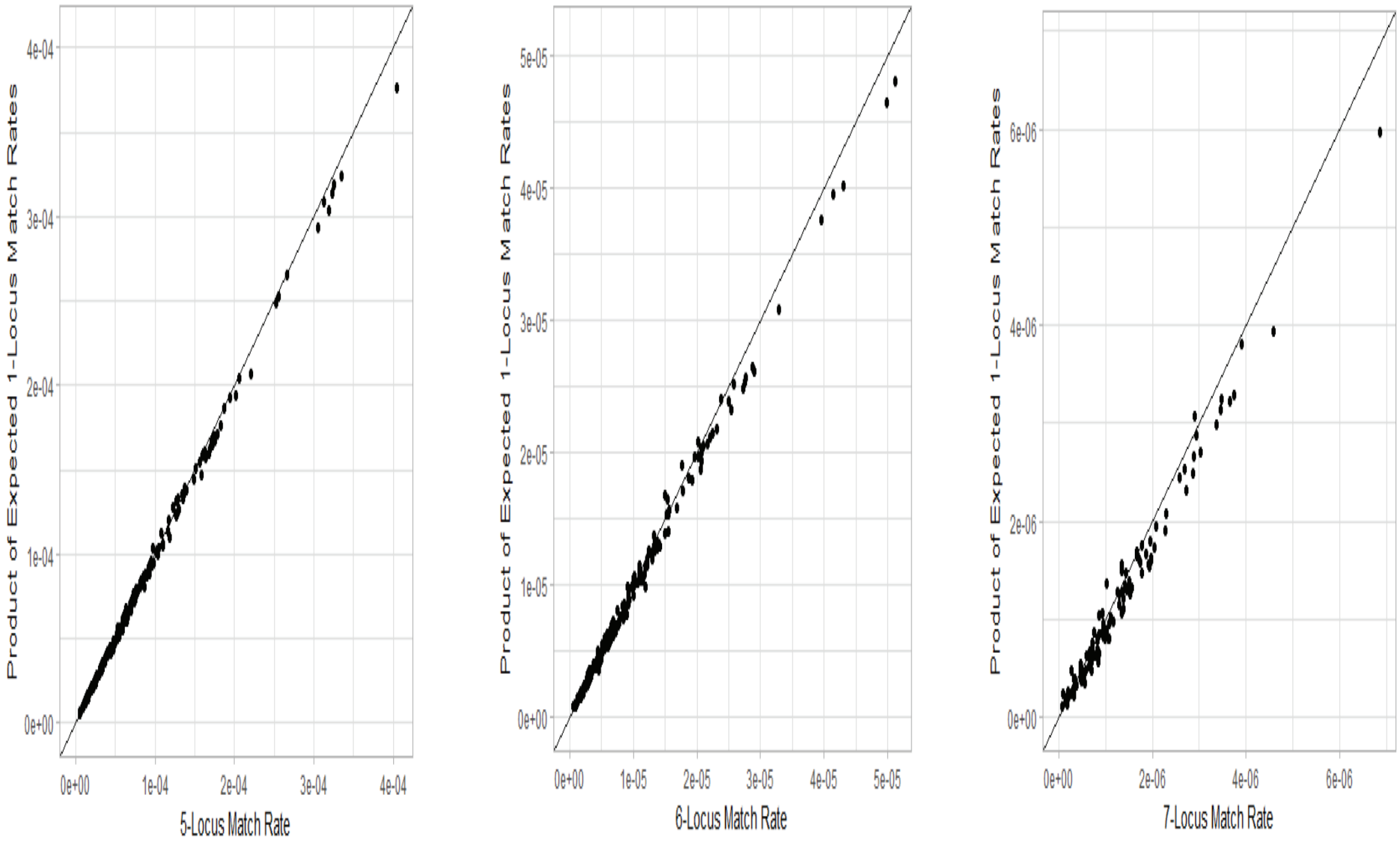
High mutation rates mean that specific mutants are likely to be recent and rare. Hence, if two individuals share alleles at one locus, they are more likely to be related through recent pedigree, and hence more likely to share alleles at a second locus."

Laurie CA, Weir BS. 2003. Theoretical Population Biology 63:207-219, 2003.

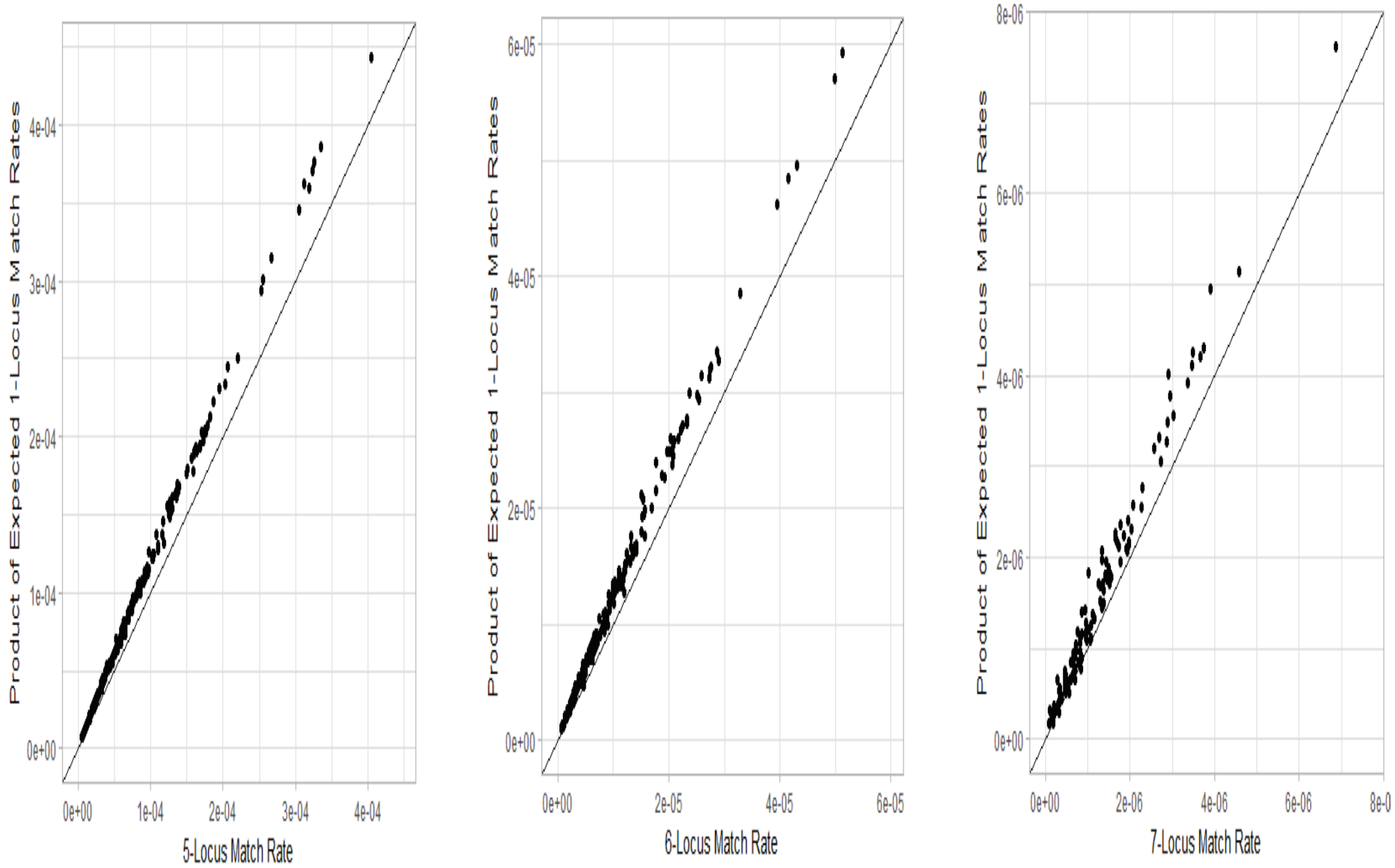
One population simulated data: $\theta = 0$



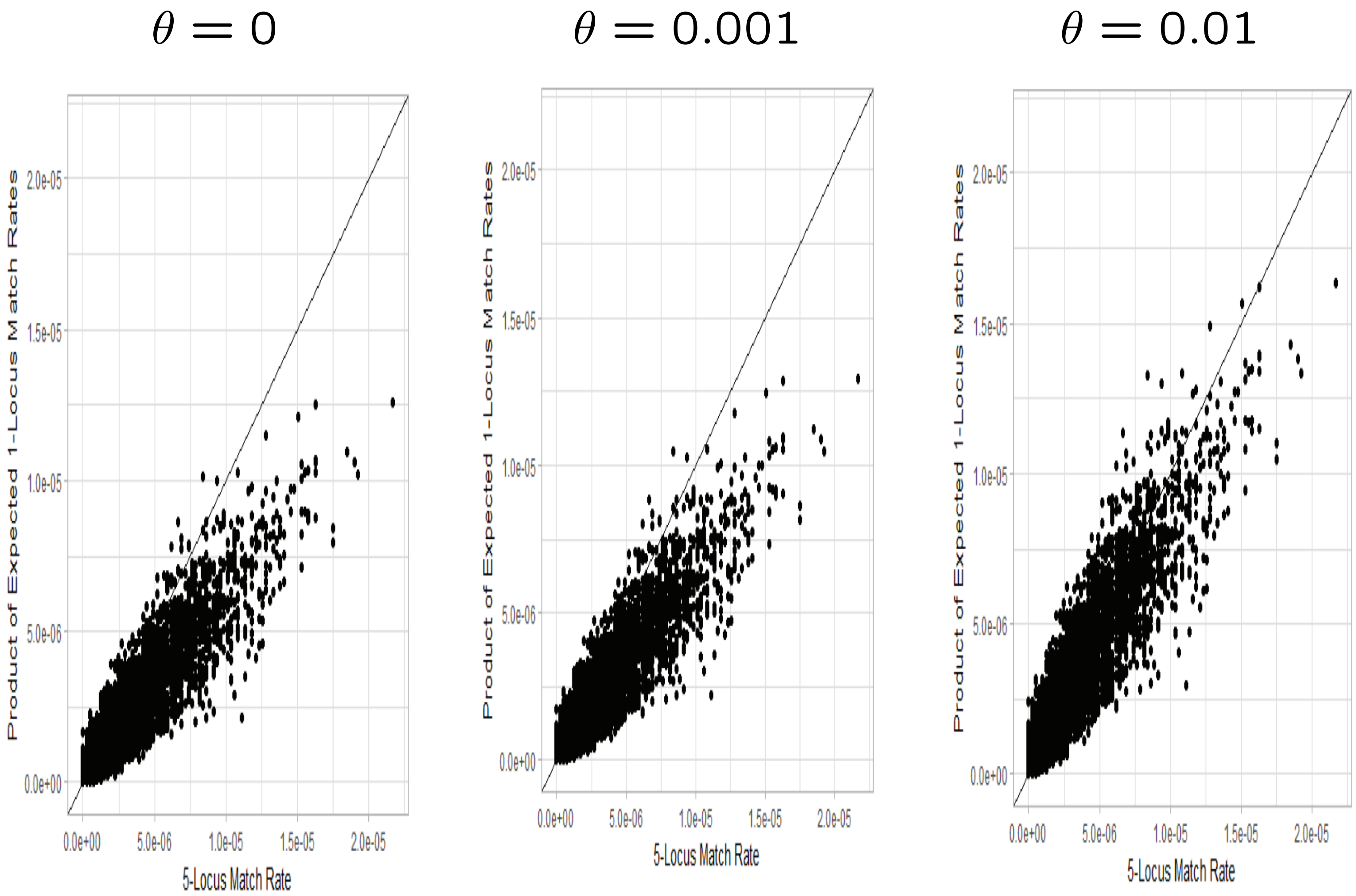
One population simulated data: $\theta = 0.001$



One population simulated data: $\theta = 0.01$

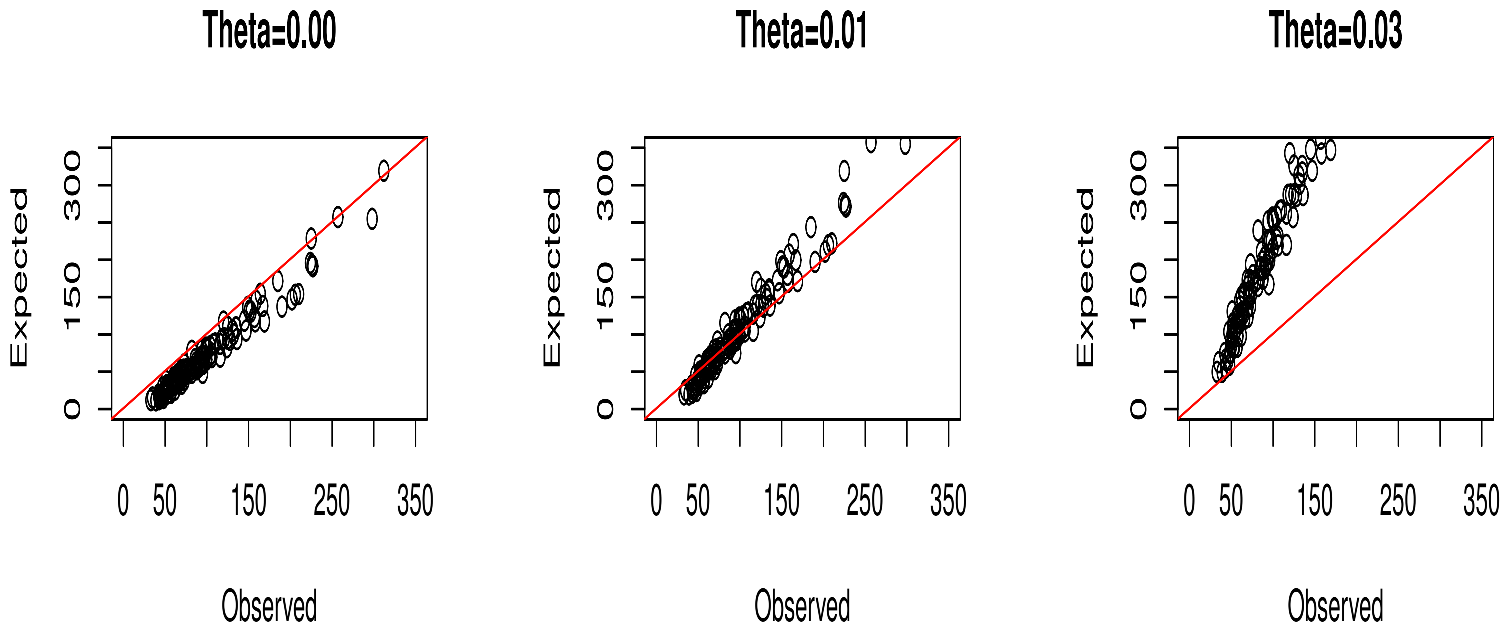


2849 US profiles



$\theta = 0$ $\theta = 0.001$ $\theta = 0.01$

15,000 Australian Profiles



Numbers of five-locus matches among nine-locus profiles.

Weir BS. 2004. Journal of Forensic Sciences 49:1009-1014, 2004.

Conclusions

- Profile probabilities decrease at the same rate as number of loci increases.
- Match probabilities are not profile probabilities.
- Match probabilities decrease more slowly as number of loci increases.
- "Theta correction" may accommodate multi-locus dependencies.
- Empirical studies need much larger databases.

Section 10: Other Techniques

Next Generation Sequencing

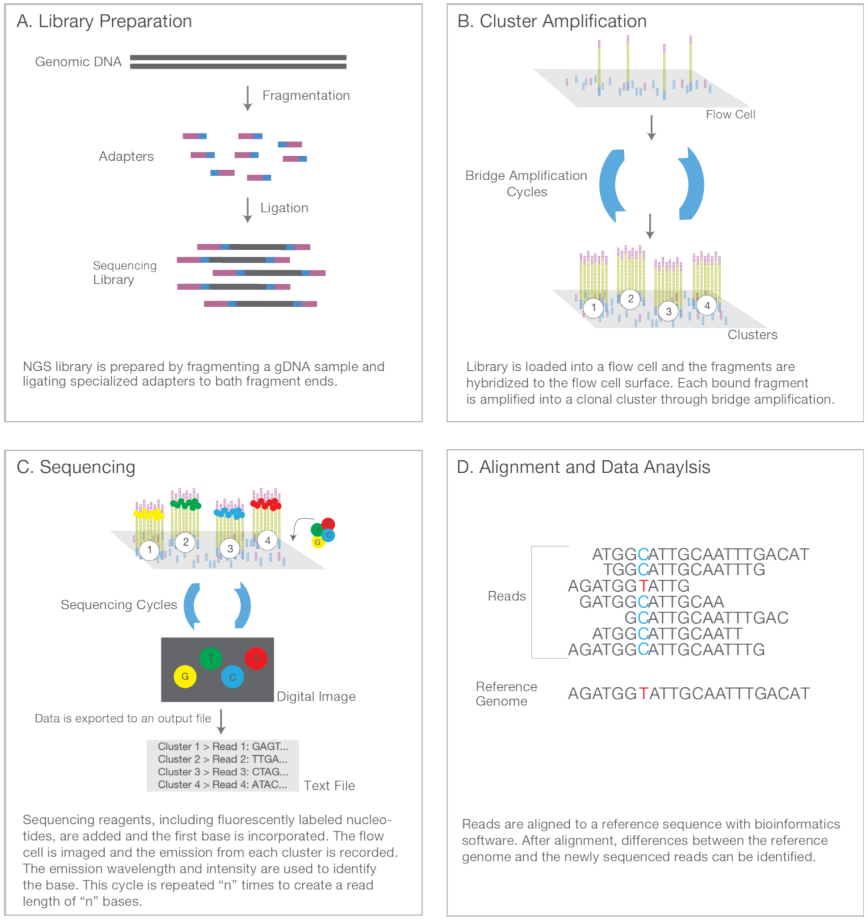
The introduction of *Next Generation Sequencing (NGS)* added a new dimension to the field of forensic genetics, providing distinct advantages over traditional CE systems in terms of captured information.

Locus	**Allele number**	**Allele sequence**
D3S1358	15	$[\mathrm{TCTA}][\mathrm{TCTG}]_3[\mathrm{TCTA}]_{11}$
D3S1358	15	$[\mathrm{TCTA}][\mathrm{TCTG}]_2[\mathrm{TCTA}]_{12}$
D18S51	20	$[\mathrm{AGAA}]_{20}$
D18S51	20	$[\mathrm{AGAA}]_{16}\mathrm{GGAA}[\mathrm{AGAA}]_3$

NGS is also referred to as Massively Parallel Sequencing (MPS), Second Generation Sequencing (SGS) or High-Throughput (HTP) sequencing.

NGS Workflow

By far the biggest player in the field of sequencing instruments is Illumina, which workflow includes four basic steps:



A. Library Preparation

NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

B. Cluster Amplification

Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

C. Sequencing

```
Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...
```

Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated “n” times to create a read length of “n” bases.

D. Alignment and Data Anaylsis

```
Reads       ATGGCATTGCAATTTGACAT
             TGGCATTGCAATTTG
          AGATGGTATTG
           GATGGCATTGCAA
               GCATTGCAATTTGAC
            ATGGCATTGCAATT
          AGATGGCATTGCAATTTG

Reference
Genome    AGATGGTATTGCAATTTGACAT
```

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

Source: An Introduction to NGS Technology (Illumina, 2015).

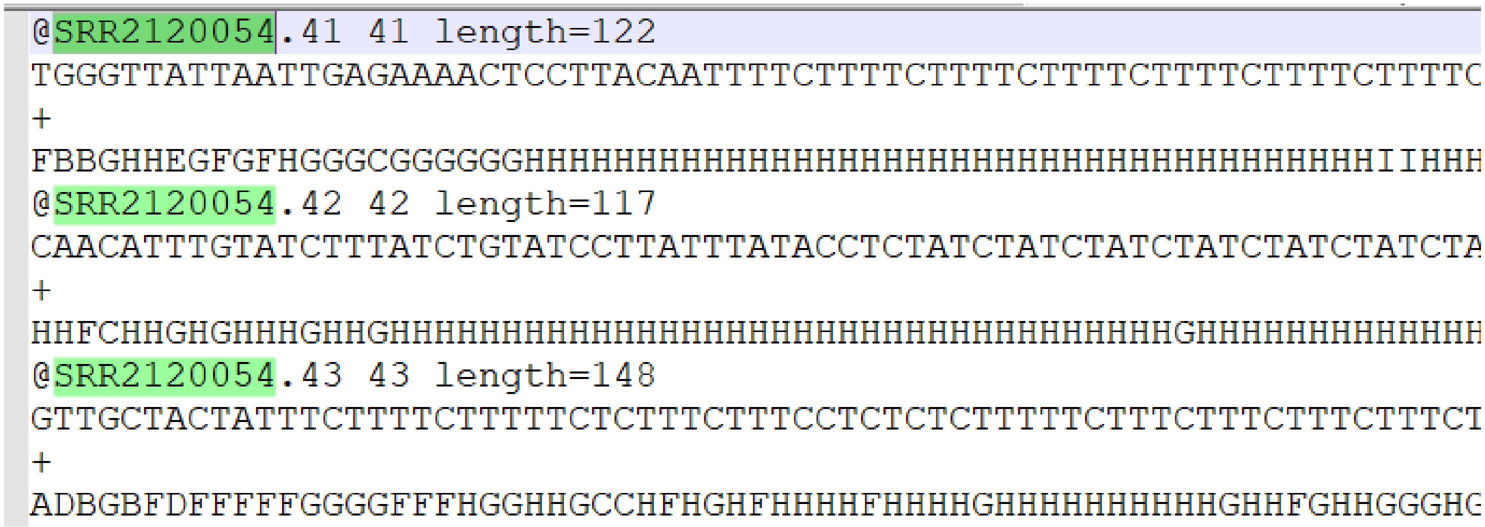
NGS Workflow

The first three steps of the workflow consist of:

- **Library preparation:** A DNA sample gets fragmented and adapters are added to both fragment ends, after which a library is obtained through PCR amplification.
- **Cluster generation:** Each fragment bounds to the surface of a flow cell and is amplified through bridge amplification, resulting in a cluster that will produce a single sequencing read.
- **Sequencing:** Base calls are made per cluster using fluorescently labeled and reversible terminator-bound nucleotides.

NGS Data Output

The most common format for storing the output of NGS instruments is a text-based FASTQ file. In addition to the observed sequence string, the file also lists its corresponding quality score, representing an estimate by the base calling software of the potential error at each sequence position.



```
@SRR2120054.41 41 length=122
TGGGTTATTAATTGAGAAAACTCCTTACAATTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTC
+
FBBGHHEGFGFHGGGCGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIIHHH
@SRR2120054.42 42 length=117
CAACATTTGTATCTTTATCTGTATCCTTATTTATACCTCTATCTATCTATCTATCTATCTATCTA
+
HHFCHHGHGHHHGHHGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGHHHHHHHHHHHH
@SRR2120054.43 43 length=148
GTTGCTACTATTTCTTTTCTTTTTCTCTTTCTTTCCTCTCTCTTTTTCTTTCTTTCTTTCTTTCT
+
ADBGBFDFFFFFGGGGFFFHGGHHGCCHFHGHFHHHHFHHHHGHHHHHHHHHHGHHFGHHGGGHG
```

NGS Data

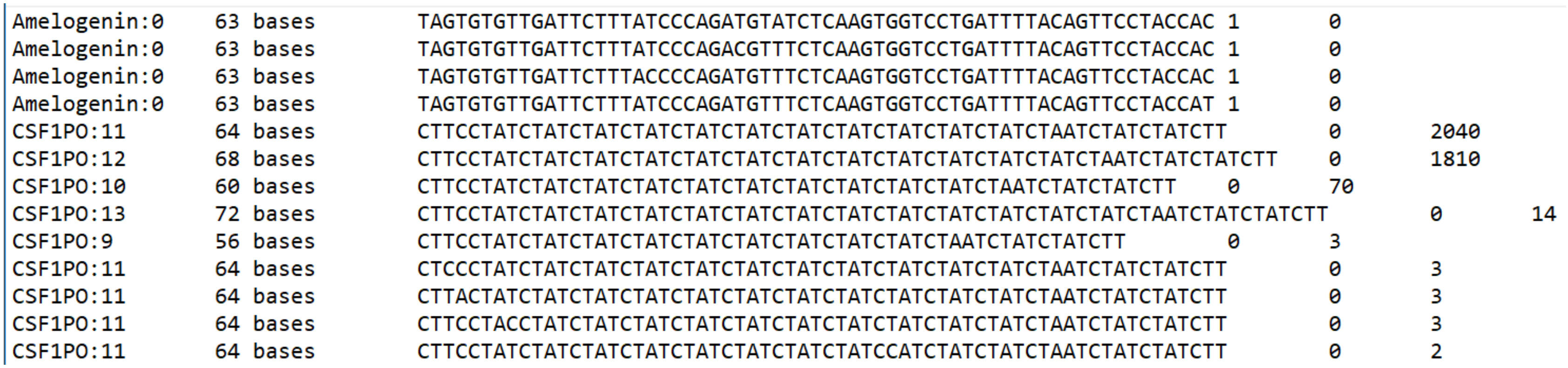
Results from sequencing platforms usually entail raw data, and need to be translated into information suitable for further (statistical) analysis.

- Software tools are available that align the reads to a reference sequence (**alignment**);
- Detect variations in the individual's genome (**variant calling**);
- And annotate the data using external information, resulting in a summarized data structure (**annotation**).

Instead of aligning to a reference sequence, sequence-searching techniques can be used that will use flanking sequences to detect STRs.

NGS Data Output

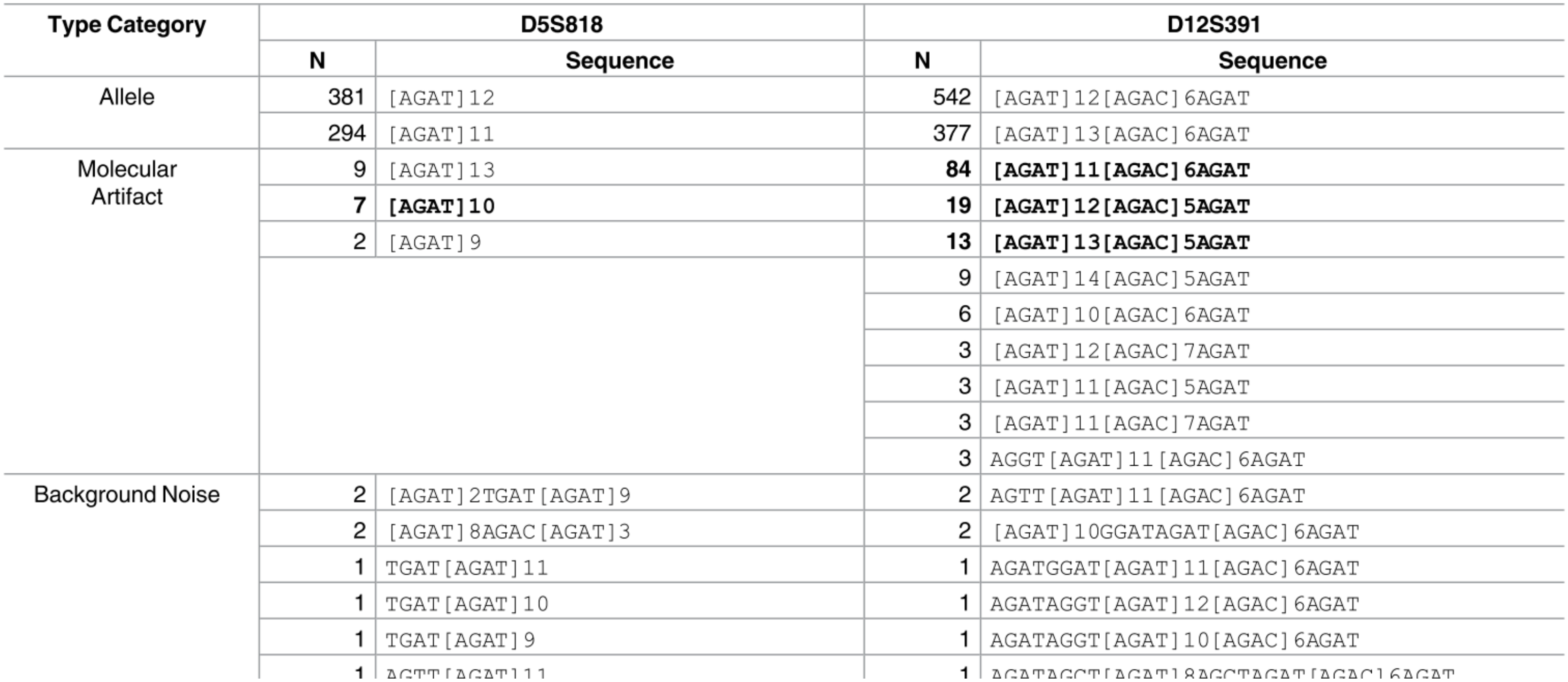
STRait Razor is an example of a sequence-searching technique, and produces output that looks as follows:



```
Amelogenin:0    63 bases        TAGTGTGTTGATTCTTTATCCCAGATGTATCTCAAGTGGTCCTGATTTTACAGTTCCTACCAC 1       0
Amelogenin:0    63 bases        TAGTGTGTTGATTCTTTATCCCAGACGTTTCTCAAGTGGTCCTGATTTTACAGTTCCTACCAC 1       0
Amelogenin:0    63 bases        TAGTGTGTTGATTCTTTACCCCAGATGTTTCTCAAGTGGTCCTGATTTTACAGTTCCTACCAC 1       0
Amelogenin:0    63 bases        TAGTGTGTTGATTCTTTATCCCAGATGTTTCTCAAGTGGTCCTGATTTTACAGTTCCTACCAT 1       0
CSF1PO:11       64 bases        CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT        0       2040
CSF1PO:12       68 bases        CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT    0       1810
CSF1PO:10       60 bases        CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT    0       70
CSF1PO:13       72 bases        CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT        0       14
CSF1PO:9        56 bases        CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT        0       3
CSF1PO:11       64 bases        CTCCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT        0       3
CSF1PO:11       64 bases        CTTACTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT        0       3
CSF1PO:11       64 bases        CTTCCTACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTT        0       3
CSF1PO:11       64 bases        CTTCCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCTATCTAATCTATCTATCTT        0       2
```

NGS Data Output

NGS output can be annotated further based on method response categories:



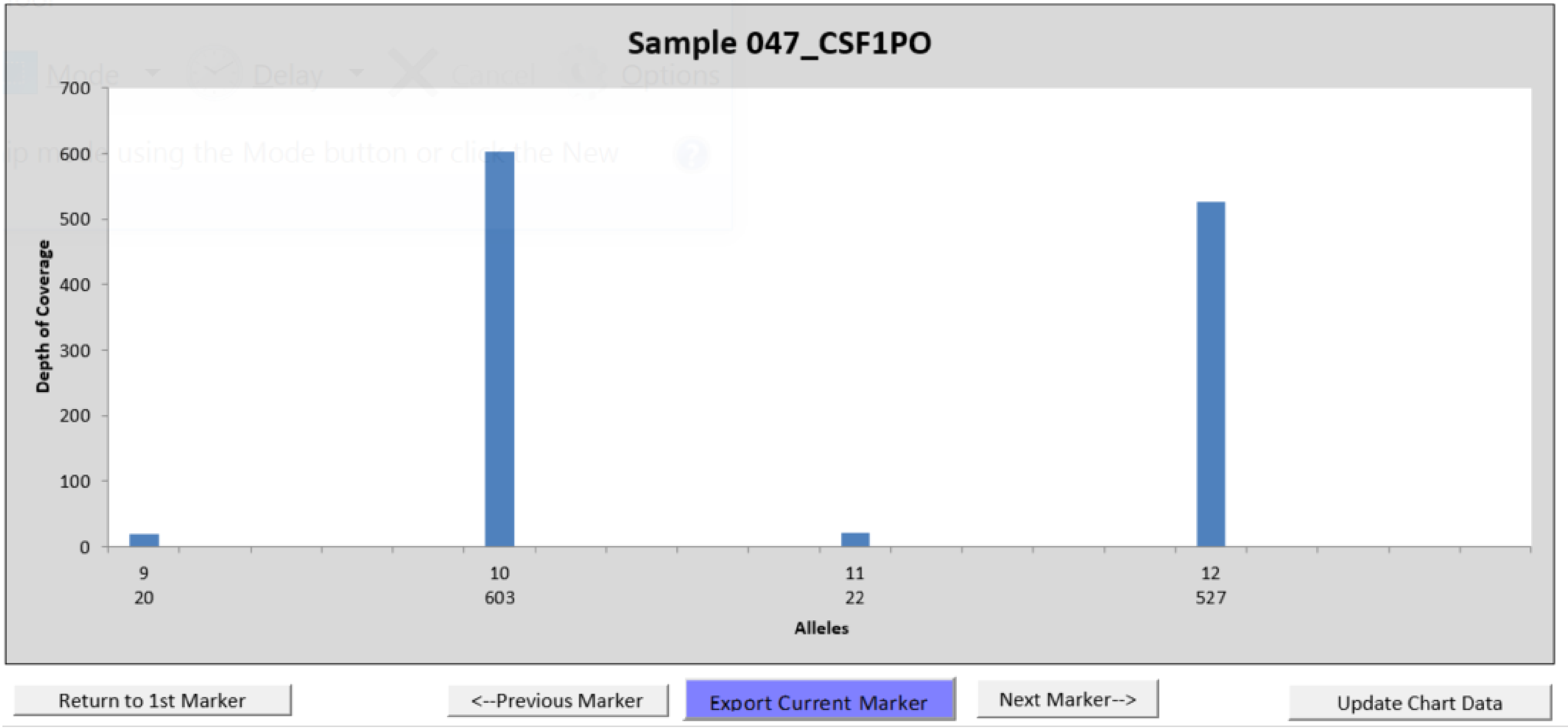
Type Category	D5S818		D12S391	
	N	Sequence	N	Sequence
Allele	381	[AGAT]12	542	[AGAT]12[AGAC]6AGAT
	294	[AGAT]11	377	[AGAT]13[AGAC]6AGAT
Molecular Artifact	9	[AGAT]13	**84**	**[AGAT]11[AGAC]6AGAT**
	7	**[AGAT]10**	**19**	**[AGAT]12[AGAC]5AGAT**
	2	[AGAT]9	**13**	**[AGAT]13[AGAC]5AGAT**
			9	[AGAT]14[AGAC]5AGAT
			6	[AGAT]10[AGAC]6AGAT
			3	[AGAT]12[AGAC]7AGAT
			3	[AGAT]11[AGAC]5AGAT
			3	[AGAT]11[AGAC]7AGAT
			3	AGGT[AGAT]11[AGAC]6AGAT
Background Noise	2	[AGAT]2TGAT[AGAT]9	2	AGTT[AGAT]11[AGAC]6AGAT
	2	[AGAT]8AGAC[AGAT]3	2	[AGAT]10GGATAGAT[AGAC]6AGAT
	1	TGAT[AGAT]11	1	AGATGGAT[AGAT]11[AGAC]6AGAT
	1	TGAT[AGAT]10	1	AGATAGGT[AGAT]12[AGAC]6AGAT
	1	TGAT[AGAT]9	1	AGATAGGT[AGAT]10[AGAC]6AGAT
	1	AGTT[AGAT]11	1	AGATAGCT[AGAT]8AGCTAGAT[AGAC]6AGAT

Source: A technique for setting analytical thresholds in MPS-based forensic DNA analysis (Young et al., 2017).

NGS data makes it easier to classify products, when compared with CE data.

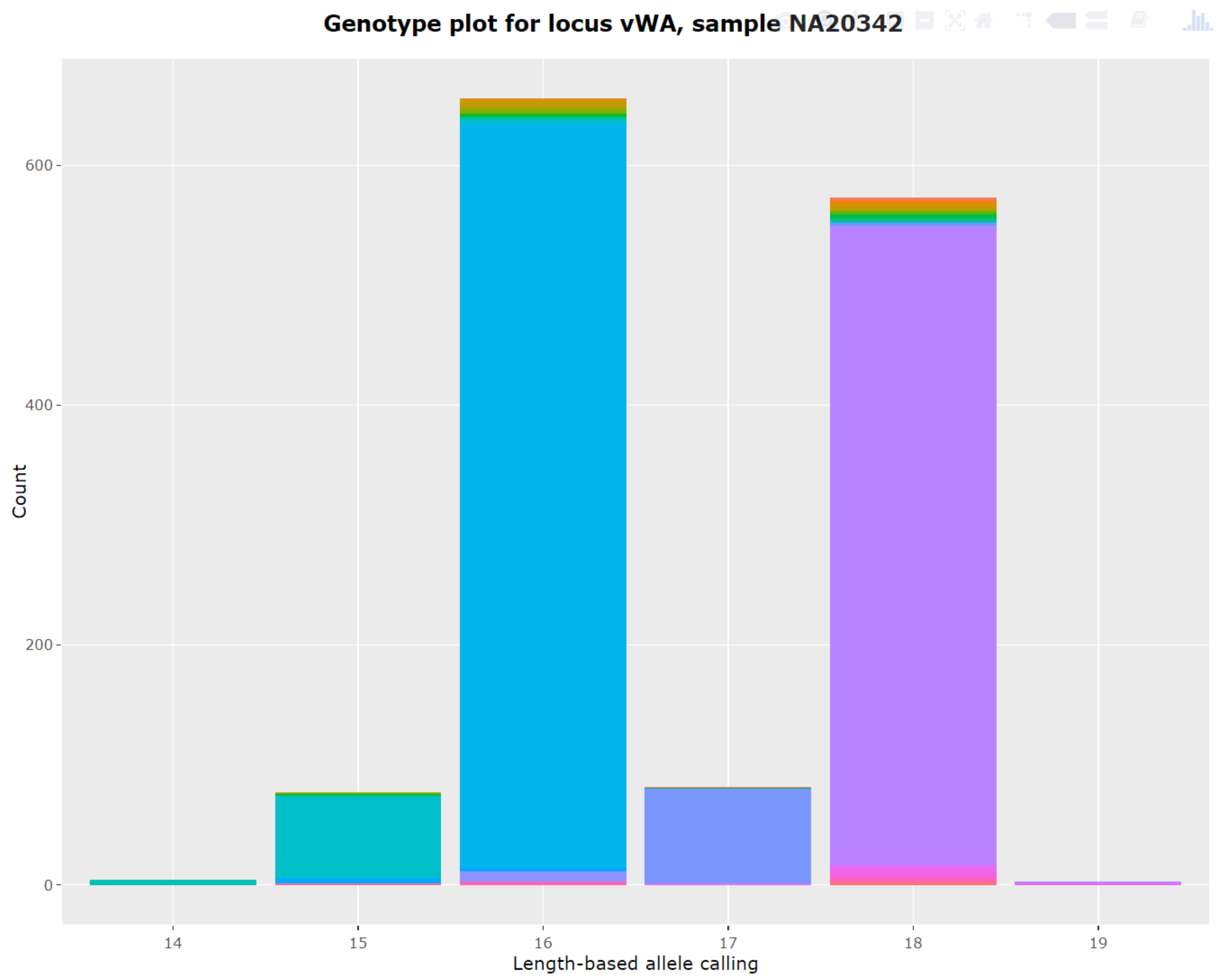
NGS Data Output

A DNA profile can be visualized similar to an epg:



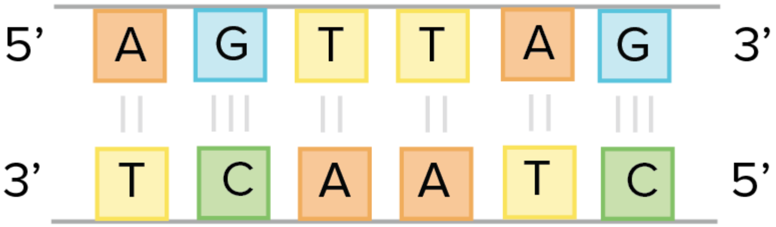
NGS Data Output

A DNA profile can be visualized similar to an epg:



NGS Considerations

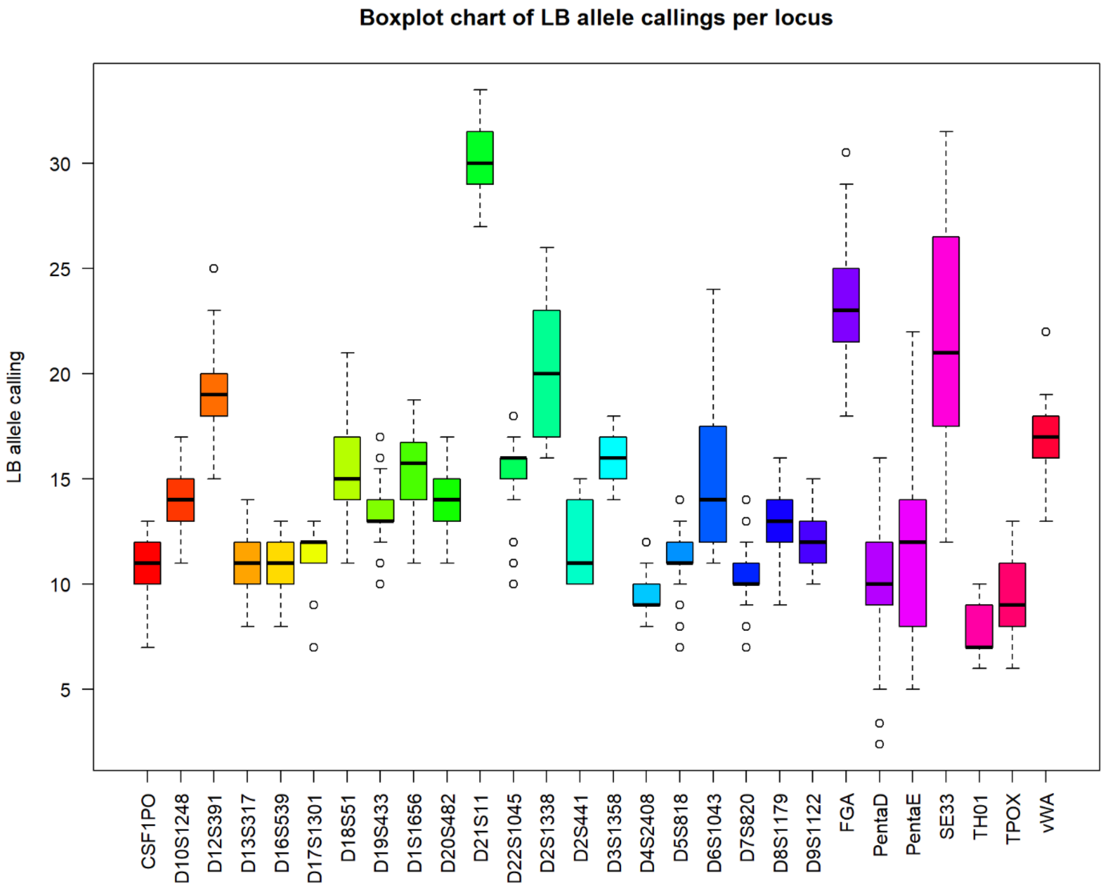
- Reads vs. peaks (discrete vs. continuous data)
- Discovery of previously unknown alleles and more variability
- New system of nomenclature needed
- Direction of strand reporting



Source: `https://www.khanacademy.org/science/biology/dna-as-the-genetic-material/dna-replication/a/molecular-mechanism-of-dna-replication`.

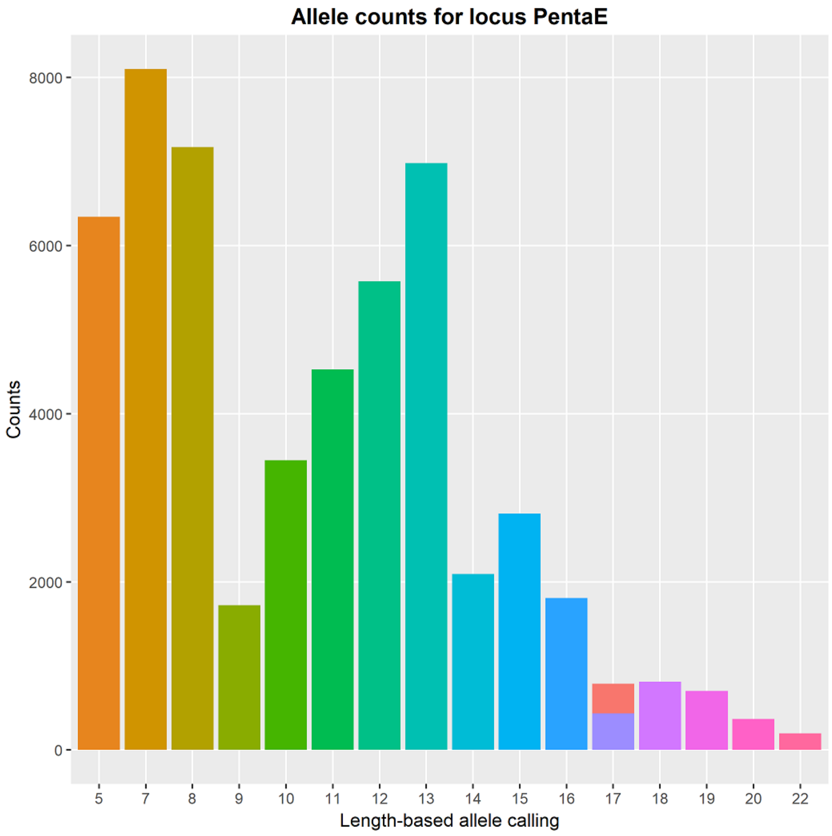
Length-based Allele Callings

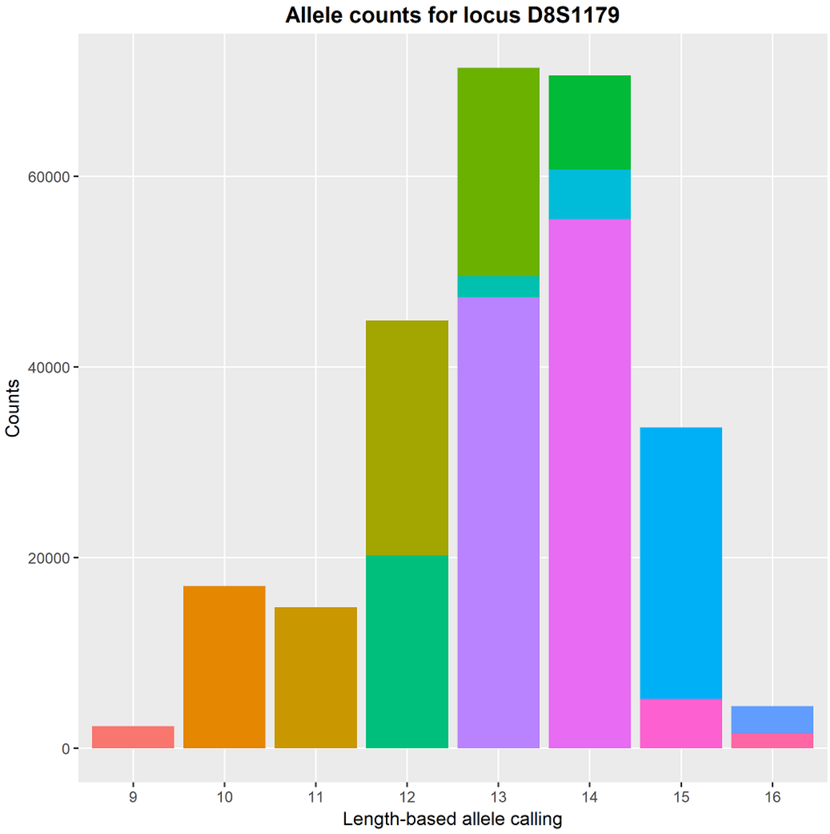
NGS data mainly leads to a gain in discrimination for compound and complex STRs, although this will be minimal for already highly polymorphic loci.



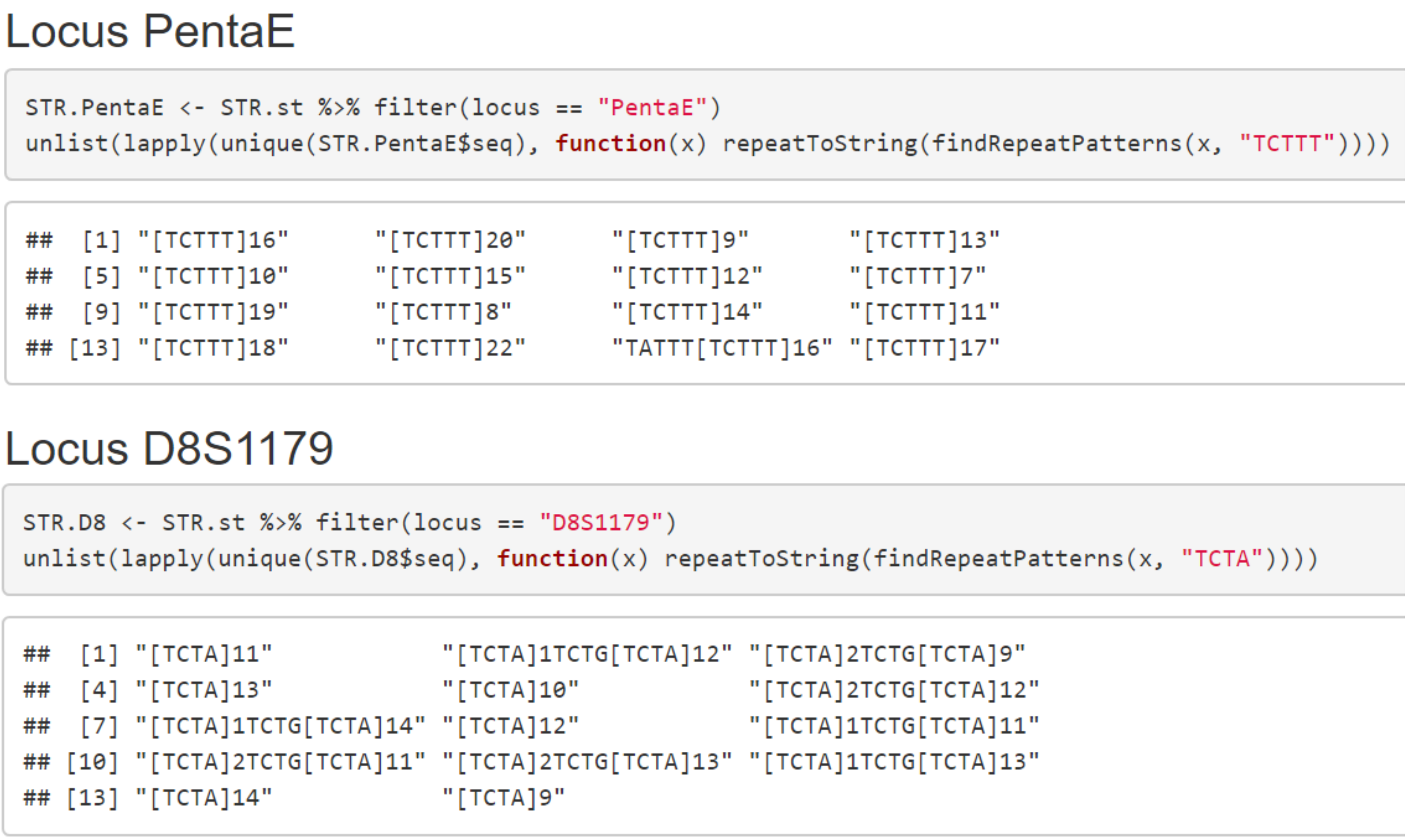
LB vs. SB Allele Callings

Locus Penta E is already quite polymorphic, so NGS data does not lead to significant improvements. For locus D8S1179, sequencing leads to a substantial increase in variability.





Sequence-based Allele Callings



Locus PentaE

```
STR.PentaE <- STR.st %>% filter(locus == "PentaE")
unlist(lapply(unique(STR.PentaE$seq), function(x) repeatToString(findRepeatPatterns(x, "TCTTT"))))
```

```
##  [1] "[TCTTT]16"      "[TCTTT]20"      "[TCTTT]9"       "[TCTTT]13"
##  [5] "[TCTTT]10"      "[TCTTT]15"      "[TCTTT]12"      "[TCTTT]7"
##  [9] "[TCTTT]19"      "[TCTTT]8"       "[TCTTT]14"      "[TCTTT]11"
## [13] "[TCTTT]18"      "[TCTTT]22"      "TATTT[TCTTT]16" "[TCTTT]17"
```

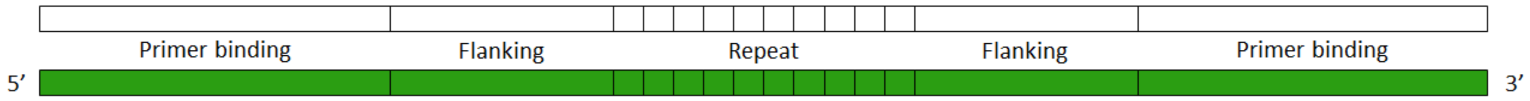
Locus D8S1179

```
STR.D8 <- STR.st %>% filter(locus == "D8S1179")
unlist(lapply(unique(STR.D8$seq), function(x) repeatToString(findRepeatPatterns(x, "TCTA"))))
```

```
##  [1] "[TCTA]11"             "[TCTA]1TCTG[TCTA]12" "[TCTA]2TCTG[TCTA]9"
##  [4] "[TCTA]13"             "[TCTA]10"            "[TCTA]2TCTG[TCTA]12"
##  [7] "[TCTA]1TCTG[TCTA]14" "[TCTA]12"            "[TCTA]1TCTG[TCTA]11"
## [10] "[TCTA]2TCTG[TCTA]11" "[TCTA]2TCTG[TCTA]13" "[TCTA]1TCTG[TCTA]13"
## [13] "[TCTA]14"             "[TCTA]9"
```

Flanking Region SNPs

Additional variation has been found in the flanking regions adjacent to repeat motifs.



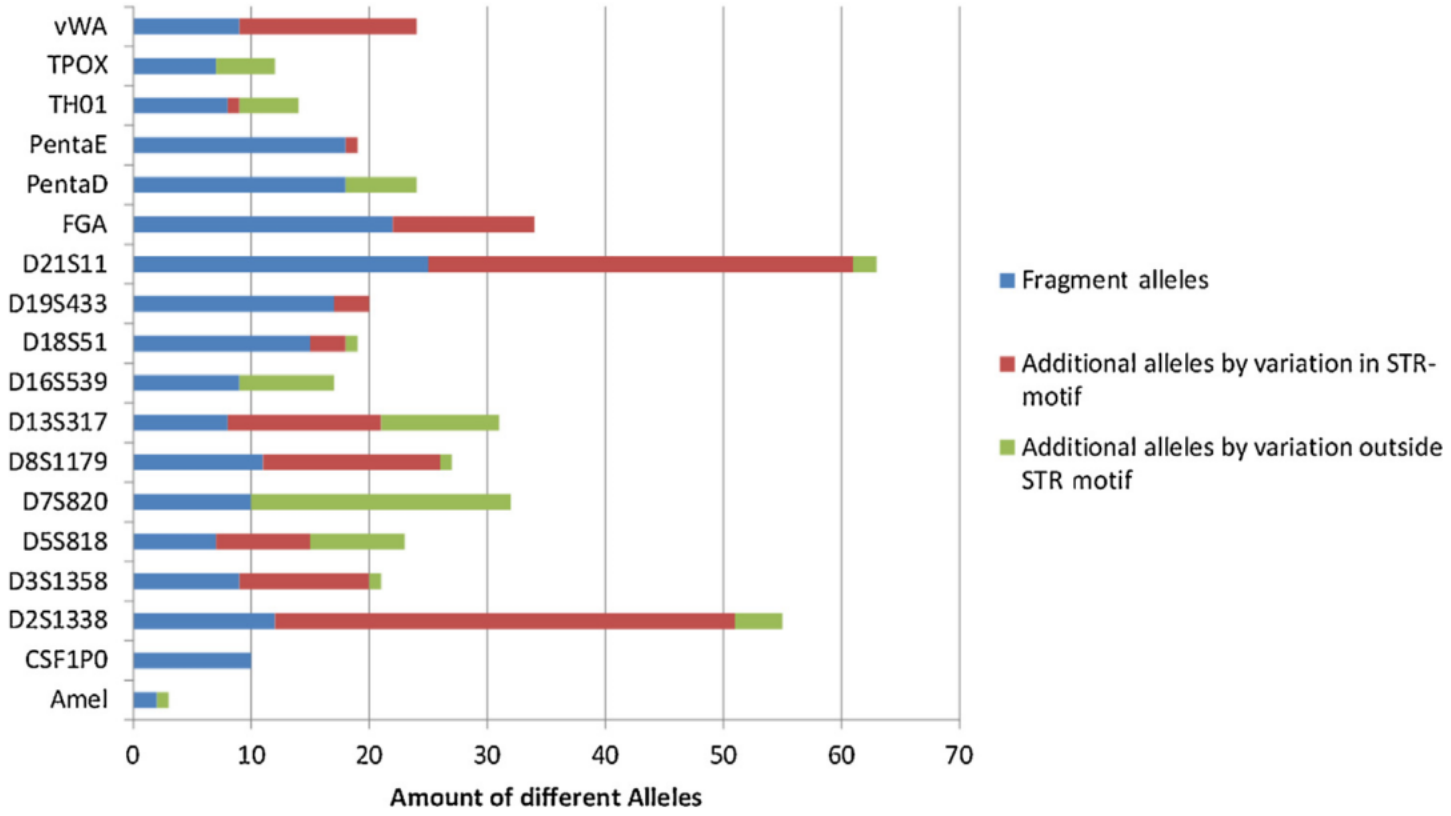
Source: Forensic DNA Evidence Interpretation (Buckleton et al., 2016).

For STR loci in which repeat regions do not display sequence differences, flanking region SNPs may still add substantial variability. Knowledge of these variants can be utilized in primer design to ensure optimal positioning during the PCR process.

Locus	**LB Allele**	**SB Allele**	**SB Allele with SNPs**
D16S539	11	$[\text{GATA}]_{11}$	$[\text{GATA}]_{11}$rs11642858[A]
D16S539	11	$[\text{GATA}]_{11}$	$[\text{GATA}]_{11}$rs11642858[C]

Observed Sequence Variation

STR sequence variation divided in length variation, additional sequence variation, and SNP variation:



Source: Massively parallel sequencing of short tandem repeats (van der Gaag et al., 2016).

NGS Modeling

New models need to be developed and implemented to accommodate NGS data, with the ultimate goal of developing a probabilistic approach for NGS mixture interpretation.

CE-based models can be used as a basis for NGS modeling. Both methods make use of the PCR process, so it is expected that artifacts such as stutter are similar.

However, peak heights need to be substituted with read counts and the remaining biological processes differ. This will materially affect the modeling parameters.

NGS Stutter Modeling

NGS data generally show higher stutter percentages than CE data. Illumina's ForenSeq uses the following thresholds (compared with Thermo Fisher's NGM Select Kit for CE data):

	Stutter Filter (%)	
Locus	**CE**	**NGS**
TH01	5	10
D2S441	9	7.5
vWA	11	22
FGA	11.5	25
D12S391	15	33
D22S1045	17	20

Multi-sequence Stutter Model

A multi-sequence model takes into account all uninterrupted stretches (AUS) as potentially contributing to stuttering.

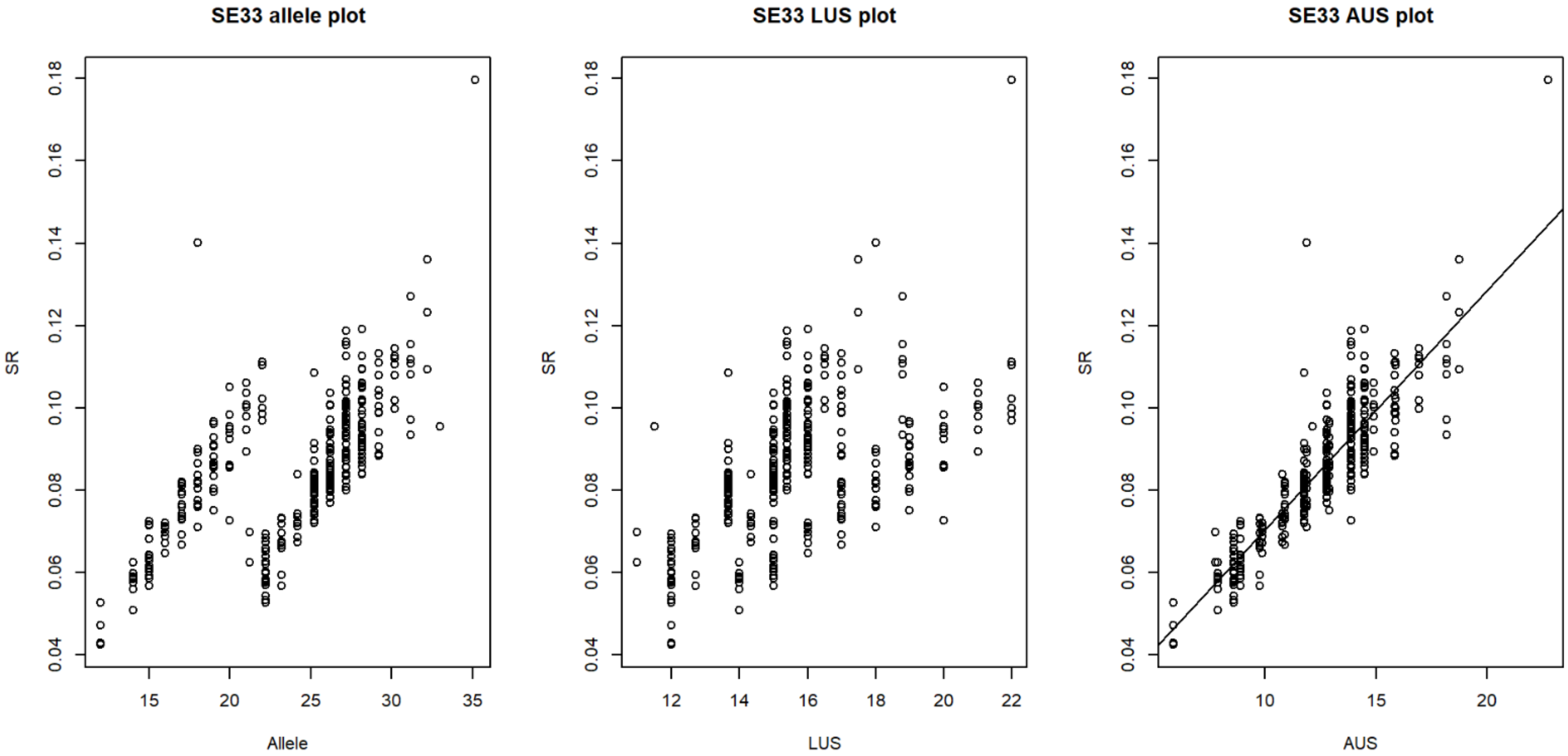
Allele	**Repeat motif**
21.2	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_9$AA AAAG[AAAG]$_{11}$G AAGG[AAAG]$_2$AG
21.2	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{11}$AA AAAG[AAAG]$_9$G AAGG[AAAG]$_2$AG
22	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{22}$G[AAAG]$_3$AG
22.2	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_7$AA AAAG[AAAG]$_{14}$GAAGG[AAAG]$_2$AG
22.2	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_8$[AG]$_5$[AAAG]$_{12}$GAAGG[AAAG]$_2$AG
22.2	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_9$AA AAAG[AAAG]$_{12}$GAAGG[AAAG]$_2$AG

Examples of locus SE33 sequences.

$$SR \sim \text{AUS} \qquad \Rightarrow \qquad SR = m\sum_i \max\left(l_i - x, 0\right) + c,$$

where l_i is the length of sequence i, and m, c and x are constants. The term x is called the lag, and can be interpreted as the number of repeats before stuttering begins.

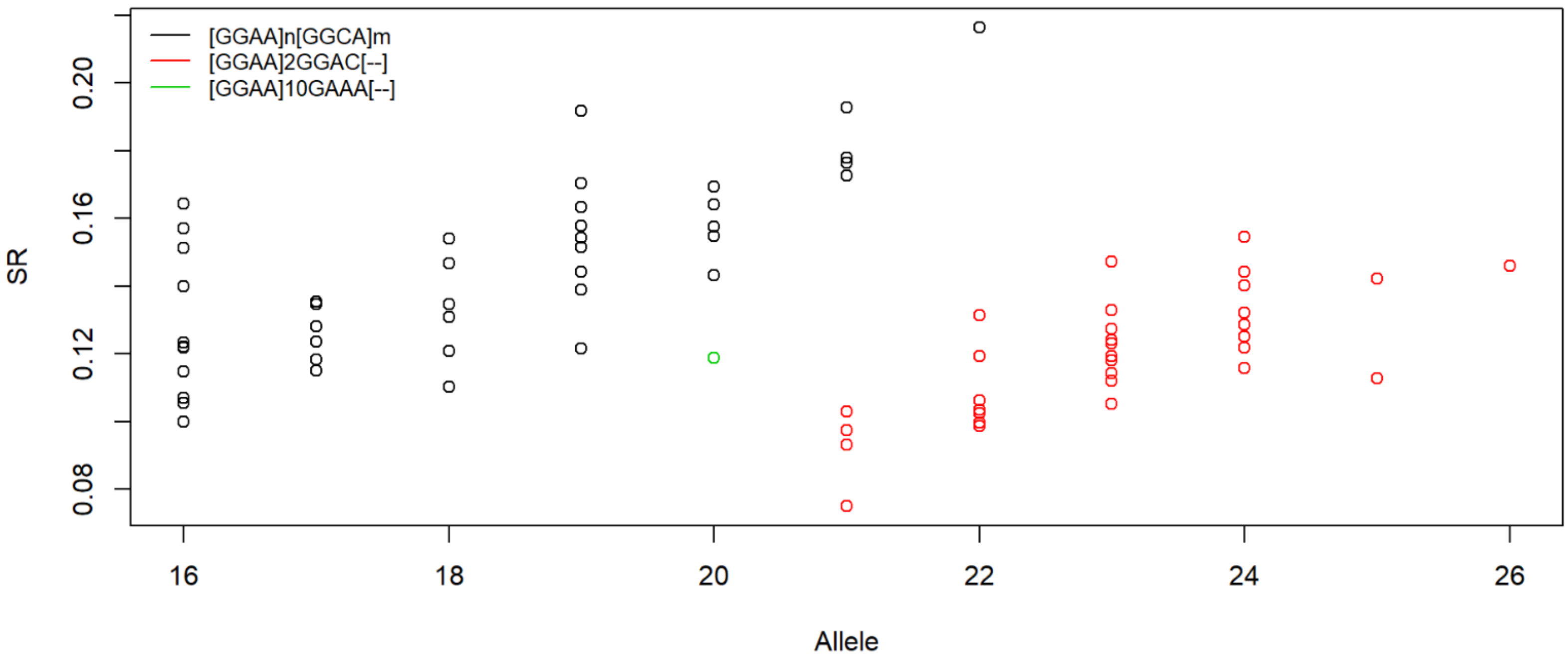
Multi-sequence Stutter Model for SE33



$$SR = m\sum_{i} \max\left(l_i - 6.11, 0\right) + c$$

Stutter Modeling and Sequence Variation

What about variation that is suggested to be attributable to sequence motif?



Stutter ratios for locus D2S1338.

Models fitted based on AUS still left some variability unexplained for some loci.

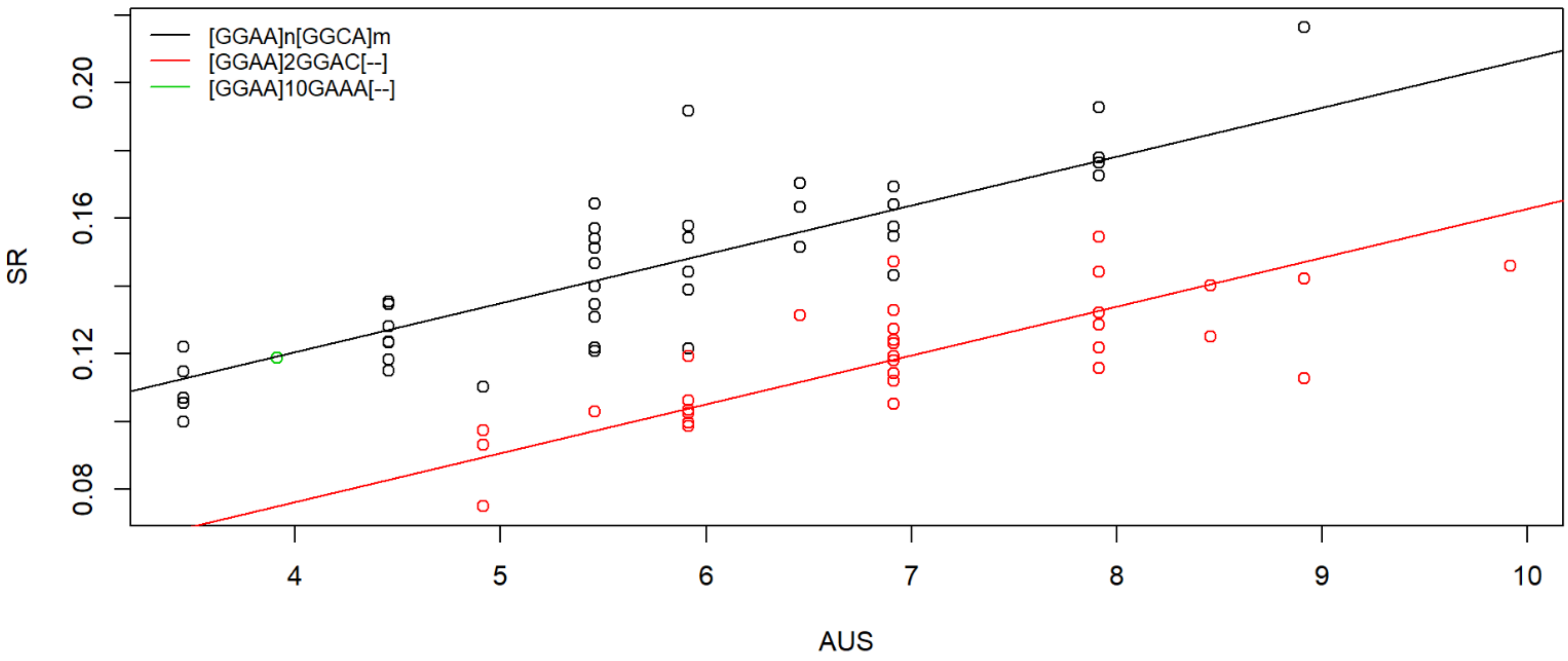
NGS Stutter Modeling - Sequence Variation

A slightly different model can allow for the sequence variations:

$$SR \sim \text{AUS} + \text{motif} \quad \Rightarrow \quad SR = m \sum_i \max(l_i - x, 0) + (c + b_j),$$

with b_j a constant for sequence variation (or motif) j. This effectively scales the regression line somewhat up or down.

NGS Stutter Modeling - Sequence Variation



Stutter ratio model for locus D2S1338.

A better fit is now obtained, from $R^2 = 0.20$ for the AUS model to values of 0.74 and 0.55 when including motif.

NGS Stutter Modeling - Sequence Variation

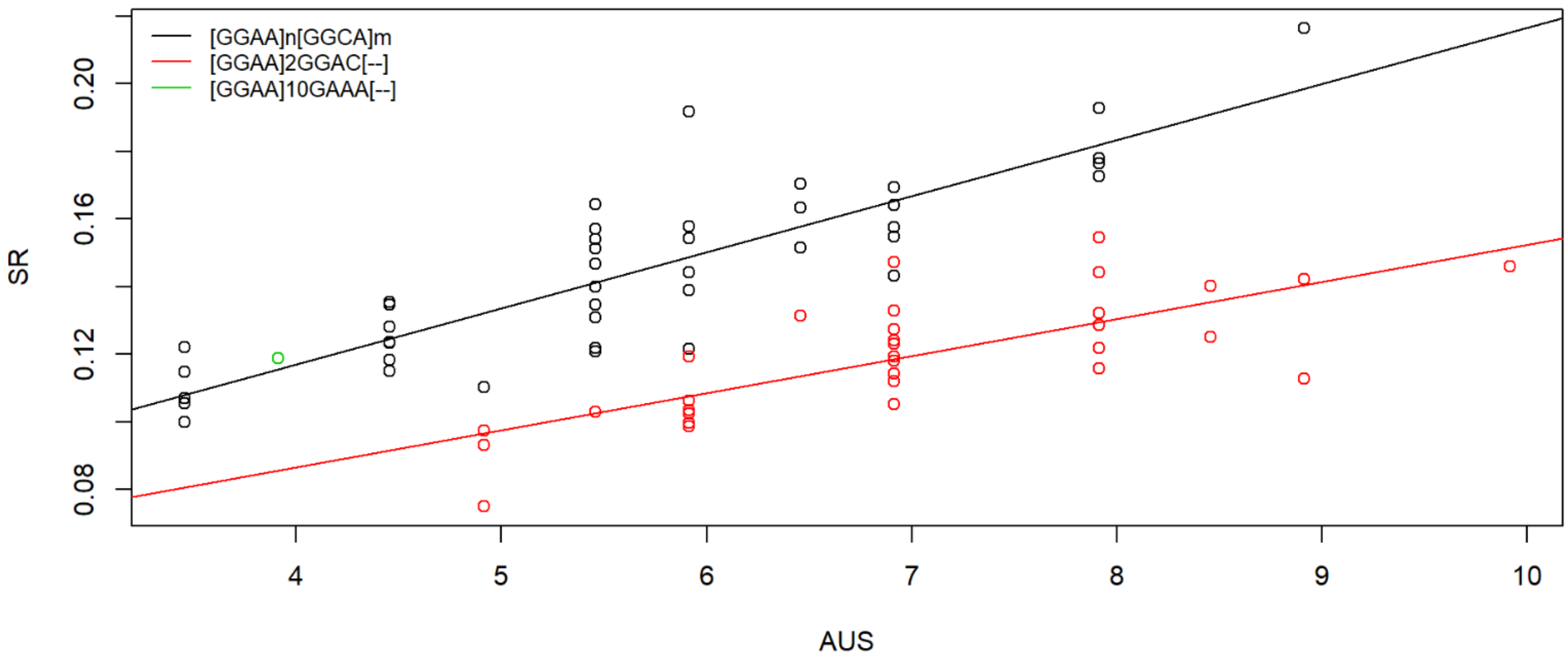
Alternatively, an interaction term can be introduced to allow for different slopes per motif:

$$SR \sim \text{AUS} \times \text{motif}$$

$$SR = (m + f_j) \sum_i \text{max}\,(l_i - x, 0) + (c + b_j),$$

with b_j and f_j constants depending on the motif.

NGS Stutter Modeling - Sequence Variation



Stutter ratio model for locus D2S1338.

The added value seems only marginal at the expense of a more complicated model.

NGS Stutter Modeling

With the sequence variations now in hand, it is possible to decompose certain stutter affected heterozygotes, composite stutter and regular stutter products.

For locus TH01, for example, there are two possible (back) stutter products:

Product	**LB Allele**	**SB Allele**
A	8.3	[AATG]$_6$ATG[AATG]$_2$
B	8.3	[AATG]$_5$ATG[AATG]$_3$

NGS Stutter Modeling

The total expected stutter count is now the sum of the two stutter products:

Product	**LB Allele**	**SB Allele**
A	8.3	$[\mathrm{AATG}]_6\mathrm{ATG}[\mathrm{AATG}]_2$
B	8.3	$[\mathrm{AATG}]_5\mathrm{ATG}[\mathrm{AATG}]_3$

$$E_{(a-1)} = \phi_A E_A + \phi_B E_B,$$

with ϕ_A and ϕ_B the proportion of stutter product A and B, respectively.

These proportions will likely reflect previous observations (e.g. longer sequences stutter more, but not all stutter come from the LUS).

NGS Stutter Modeling - Discussion

- How to determine motif?

- What about micro-variants?

- What about the possible influence from flanking variation?

- What about the effect of A-T content?

Duplex Sequencing

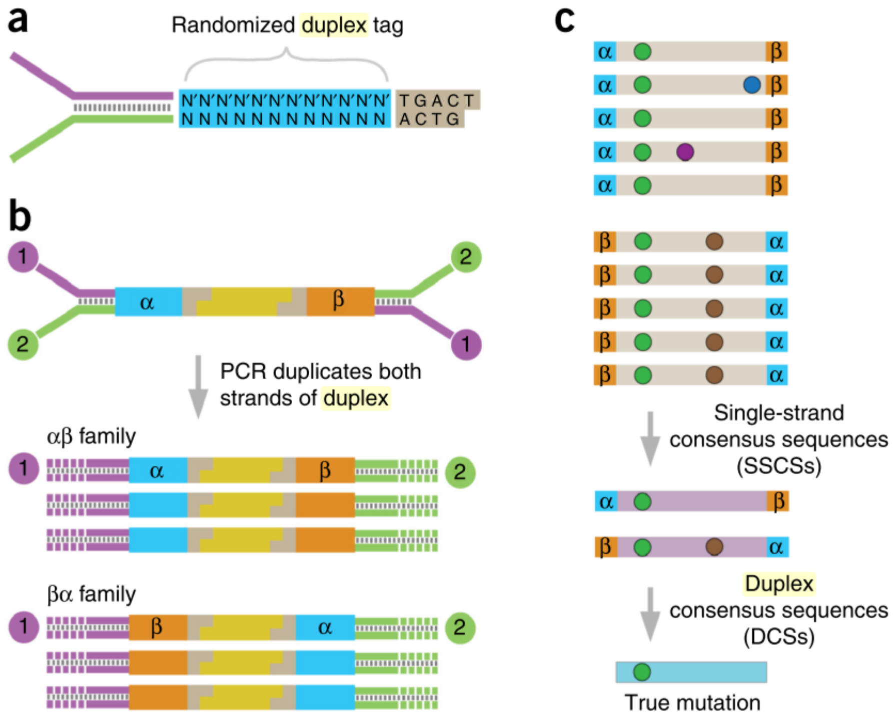
Most NGS approaches have a relatively high error rate and are therefore not suitable for detecting in vivo mutations. To overcome this limitation, a highly sensitive sequencing methodology termed *Duplex Sequencing (DS)* has been developed.

- DNA fragments get labeled with their own unique tag;
- After PCR amplification, each group yields one consensus sequence;
- Two complementary consensus sequences, derived from the same fragment, are then compared to yield a ‘duplex consensus sequence’.

Source: Detecting ultralow-frequency mutations by Duplex Sequencing (Kennedy et al., 2014).

Duplex Sequencing

Only true mutations will appear in both duplex sequences, while PCR-related artifacts will be eliminated when establishing the final consensus sequence.


a
Randomized duplex tag
N'N'N'N'N'N'N'N'N'N'N'N' TGACT
NNNNNNNNNNNN ACTG
b
1
2
α
β
2
1
PCR duplicates both strands of duplex
αβ family
1
α
β
2
βα family
1
β
α
2
c
α
β
Single-strand consensus sequences (SSCSs)
Duplex consensus sequences (DCSs)
True mutation

Source: Detecting ultralow-frequency mutations by Duplex Sequencing (Kennedy et al., 2014).

Microhaplotypes

Instead of looking at individual SNPs, it has been suggested that combining multiple SNPs into a microhap that renders highly informative for forensic purposes.

Although microhaps are more sensitive, the absence of stutter yields an increase in potential for mixture deconvolution. SNPs are also shown to be correlated with physical phenotypic traits, information the STRs cannot provide.

To make the use of microhaps feasible for forensic purposes, however, backward compatibility is required with CE data. This might be established through record linkage, based on STR inference from SNP data.

Source: Criteria for selecting microhaplotypes: mixture detection and deconvolution (Kidd & Speed, 2015).

Protein-based HID

Whereas DNA is prone to degradation, protein is chemically more robust and can persist for longer periods. Protein contains genetic variation in the form of single amino acid polymorphisms (SAPs), which can be used to infer non-synonymous SNPs (nsSNPs).

Hair is often a forensically relevant component of crime scenes and archaeological sites, where it persists under a wide range of environmental conditions. It is a poor source of nuclear DNA template, but retains a high protein content.

Genetically variant peptides (GVPs) containing SAPs can be identified and may thus be used to infer (SNP) profiles, regardless of the presence of DNA template in the sample, providing the potential for a complementary and, if necessary, alternative method for use in forensic practice.

Source: Demonstration of Protein-Based Human Identification Using the Hair Shaft Proteome (Parker et al., 2016).