

AGTGAAGCTACTTAAAGAAAGTGACTGCTACTGGTGAAAAT

SISG  
2014

## SISG Module 2: Forensic Genetics

**19th Summer Institute in Statistical Genetics**

**W** UNIVERSITY *of* WASHINGTON

(This page left intentionally blank.)

## **Forensic Genetics**

Summer Institute in Statistical Genetics, July 7-9, 2014

Simone Gittelson: simone.gittelson@gmail.com

and

Bruce Weir: bsweir@uw.edu

1

## **Forensic Genetics**

Summer Institute in Statistical Genetics, July 7-9, 2014

Simone Gittelson: simone.gittelson@gmail.com

and

Bruce Weir: bsweir@uw.edu

1

## Contents

---

---

Topic	Slide
Sources of Data	3
Probability Theory	14
Interpreting Evidence	29
Statistics	84
Allelic Independence	105
Matching Profiles	130
Mixtures: Binary Treatment	178
Inbreeding and Relatedness	209
Parentage Testing	247
Mixtures: Semi-continuous Treatment	274
Missing Persons	314
Mixtures: Continuous Treatment	328

---

---

2

## Contents

---

---

Topic	Slide
Sources of Data	3
Probability Theory	14
Interpreting Evidence	29
Statistics	84
Allelic Independence	105
Matching Profiles	130
Mixtures: Binary Treatment	178
Inbreeding and Relatedness	209
Parentage Testing	247
Mixtures: Semi-continuous Treatment	274
Missing Persons	314
Mixtures: Continuous Treatment	328

---

---

2

## Sources of Data

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

3

## Sources of Data

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

3

## Mendel's Data

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

4

## Mendel's Data

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

4

## ABO System

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

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

5

## ABO System

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

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

5

## 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

6

## 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

6

## **Electrophoretic Detection**

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

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

7

## **Electrophoretic Detection**

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

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

7

## Alec Jeffreys

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

Initial work was on mini-satellites, where repeat unit lengths were in the tens of bases and fragment lengths were in thousands of bases. Jeffrey's multi-locus probes detected regions from several parts 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.

8

## Alec Jeffreys

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

Initial work was on mini-satellites, where repeat unit lengths were in the tens of bases and fragment lengths were in thousands of bases. Jeffrey's multi-locus probes detected regions from several parts 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.

8

## **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.

9

## **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.

9

## **PCR-based STR Markers**

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

A good source is “Following multiplex PCR amplification, DNA samples containing the length-variant STR alleles are typically separated by capillary electrophoresis and genotyped by comparison to an allelic ladder supplied with a commercial kit. ”

Butler JM. Short tandem repeat typing technologies used in human identity testing. *BioTechniques* 43:Sii-Sv (October 2007)  
doi 10.2144/000112582

10

## **PCR-based STR Markers**

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

A good source is “Following multiplex PCR amplification, DNA samples containing the length-variant STR alleles are typically separated by capillary electrophoresis and genotyped by comparison to an allelic ladder supplied with a commercial kit. ”

Butler JM. Short tandem repeat typing technologies used in human identity testing. *BioTechniques* 43:Sii-Sv (October 2007)  
doi 10.2144/000112582

10

## **Additional STR Information**

An excellent resource is provided by the National Institute of Standards and Technology (NIST), especially for training materials:

<http://www.cstl.nist.gov/strbase/training.htm>

Look especially at

[http://www.cstl.nist.gov/strbase/ppt/AAFS2006\\_1\\_STR\\_Biology.pps](http://www.cstl.nist.gov/strbase/ppt/AAFS2006_1_STR_Biology.pps)

11

## **Additional STR Information**

An excellent resource is provided by the National Institute of Standards and Technology (NIST), especially for training materials:

<http://www.cstl.nist.gov/strbase/training.htm>

Look especially at

[http://www.cstl.nist.gov/strbase/ppt/AAFS2006\\_1\\_STR\\_Biology.pps](http://www.cstl.nist.gov/strbase/ppt/AAFS2006_1_STR_Biology.pps)

11

## **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.

12

## **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.

12

## **Current 1000Genomes Data**

Tuesday June 24, 2014

The Initial Phase 3 variant list and phased genotypes.

The initial call set from the 1000 Genomes Project Phase 3 analysis is now available on our ftp site in the directory release/20130502/.

These release contains more than 79 million variant sites and includes not just biallelic snps but also indels, deletions, complex short substitutions and other structural variant classes. It is based on data from 2535 individuals from 26 different populations around the world.

[www.1000genomes.org](http://www.1000genomes.org)

13

## **Current 1000Genomes Data**

Tuesday June 24, 2014

The Initial Phase 3 variant list and phased genotypes.

The initial call set from the 1000 Genomes Project Phase 3 analysis is now available on our ftp site in the directory release/20130502/.

These release contains more than 79 million variant sites and includes not just biallelic snps but also indels, deletions, complex short substitutions and other structural variant classes. It is based on data from 2535 individuals from 26 different populations around the world.

[www.1000genomes.org](http://www.1000genomes.org)

13

## 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.

14

## 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.

14

## Probabilities

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

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

15

## Probabilities

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

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

15

## First Law of Probability

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

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

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

16

## First Law of Probability

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

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

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

16

## Second Law of Probability

If  $A, B$  are mutually exclusive given  $E$

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

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

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

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

17

## Second Law of Probability

If  $A, B$  are mutually exclusive given  $E$

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

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

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

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

17

## **Third Law of Probability**

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

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

18

## **Third Law of Probability**

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

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

18

## Independent Events

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

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

Therefore, for independent events

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

This may be written as

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

19

## Independent Events

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

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

Therefore, for independent events

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

This may be written as

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

19

## Law of Total Probability

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

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

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

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

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

20

## Law of Total Probability

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

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

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

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

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

20

## Odds

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

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

This can be rearranged to give

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

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

21

## Odds

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

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

This can be rearranged to give

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

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

21

## Bayes' Theorem

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

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

22

## Bayes' Theorem

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

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

22

## Odds Form of Bayes' Theorem

From the third law of probability

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

Taking the ratio of these two equations:

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

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

23

## Odds Form of Bayes' Theorem

From the third law of probability

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

Taking the ratio of these two equations:

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

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

23

## AIDS Example

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

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

24

## AIDS Example

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

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

24

## AIDS Example

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

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

25

## AIDS Example

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

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

25

## 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.

26

## 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.

26

## Birthday Problem

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

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

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

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

27

## Birthday Problem

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

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

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

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

27

## Birthday Problem

If  $P = 1/(574 \text{ million})$  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  $N = 300 \text{ million}$ , then  $P_M =$  is essentially 1. It is almost certain that two people in the US have the same rare DNA profile.

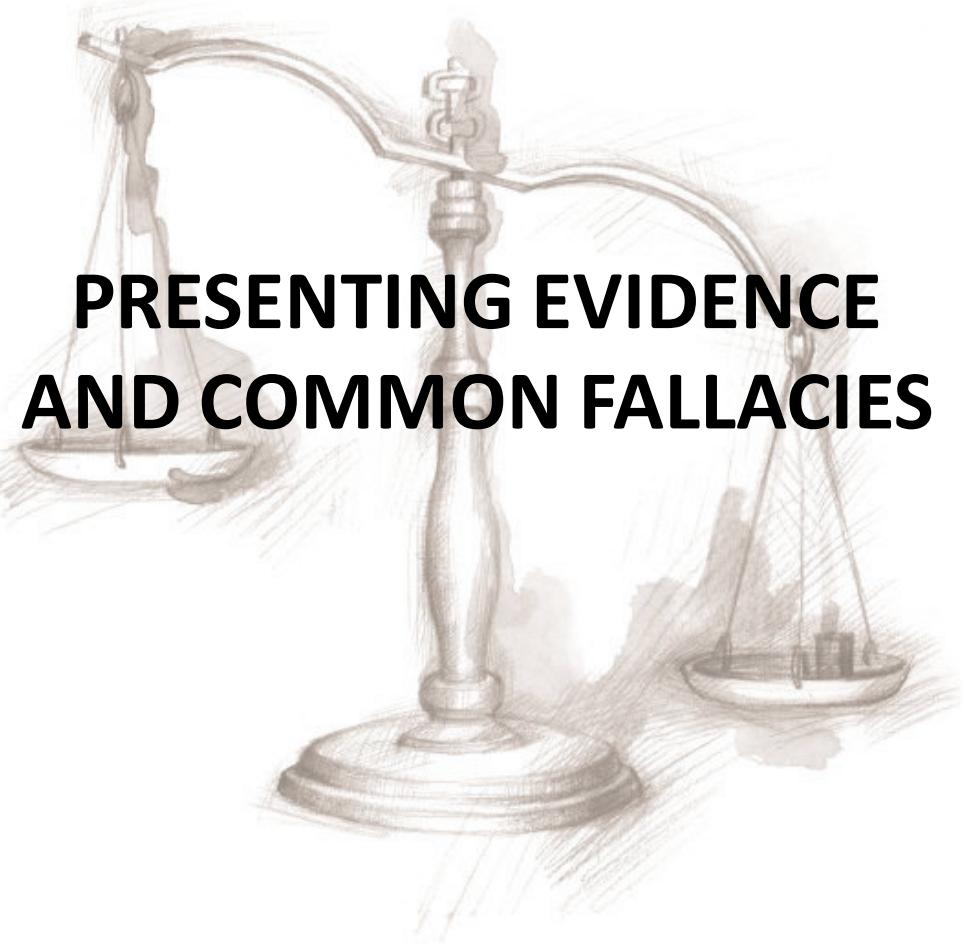
28

## Birthday Problem

If  $P = 1/(574 \text{ million})$  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.

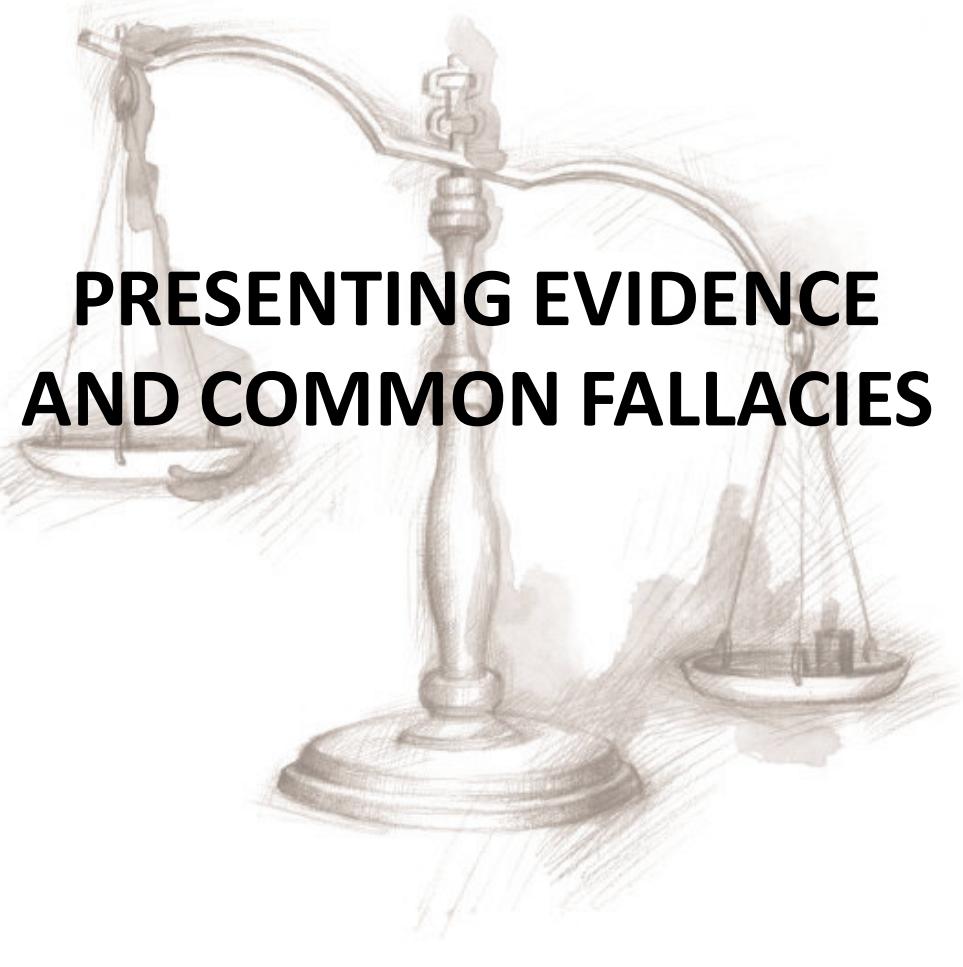
If  $P = 10^{-16}$  and  $N = 300 \text{ million}$ , then  $P_M =$  is essentially 1. It is almost certain that two people in the US have the same rare DNA profile.

28



## **PRESENTING EVIDENCE AND COMMON FALLACIES**

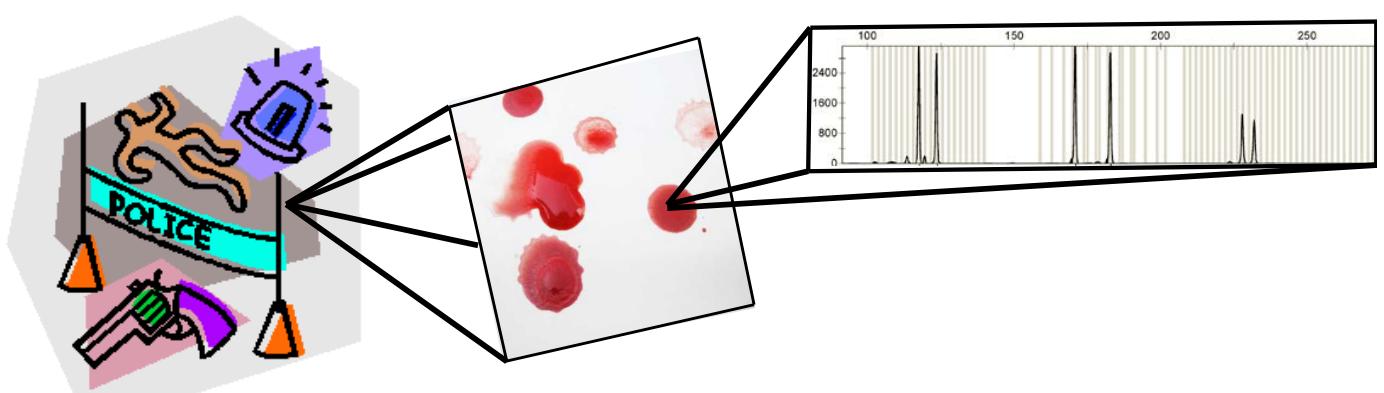
29



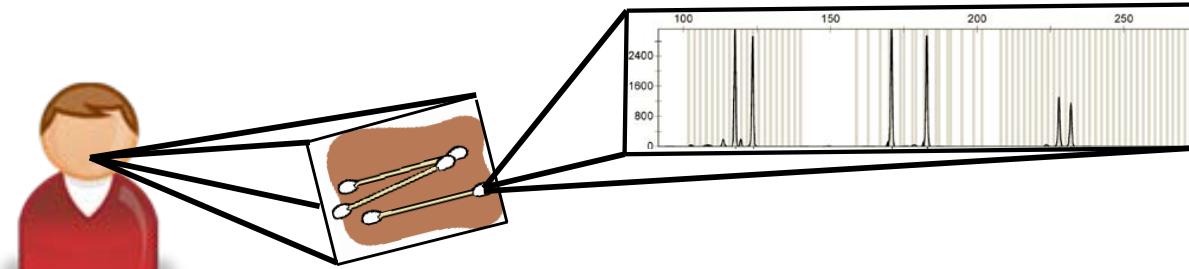
## **PRESENTING EVIDENCE AND COMMON FALLACIES**

29

# Single Contributor Stain



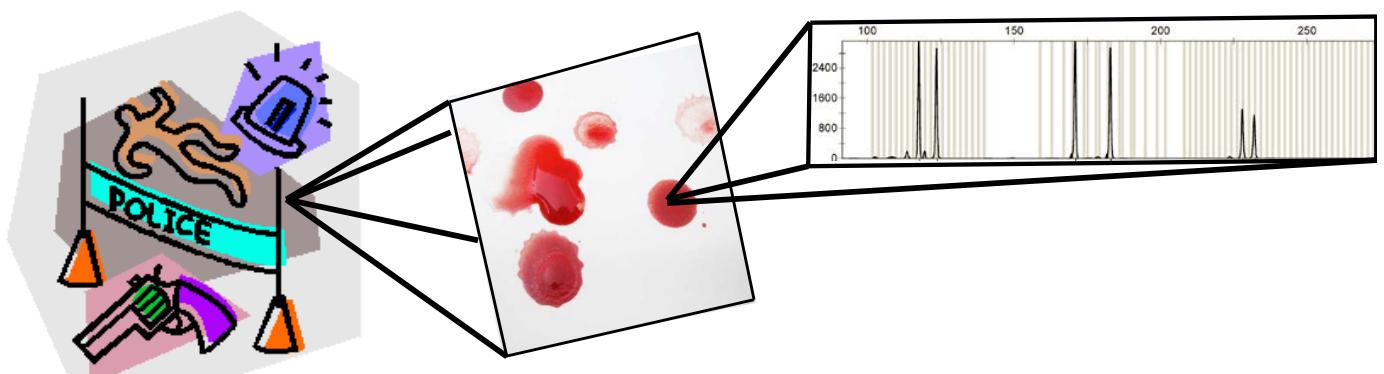
crime scene



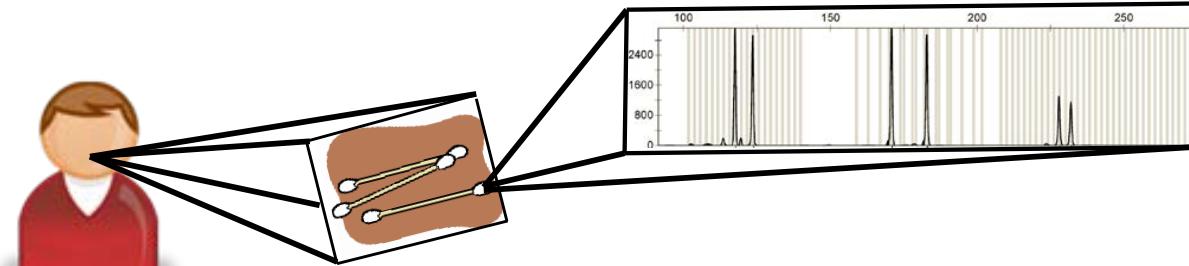
suspect

30

# Single Contributor Stain



crime scene

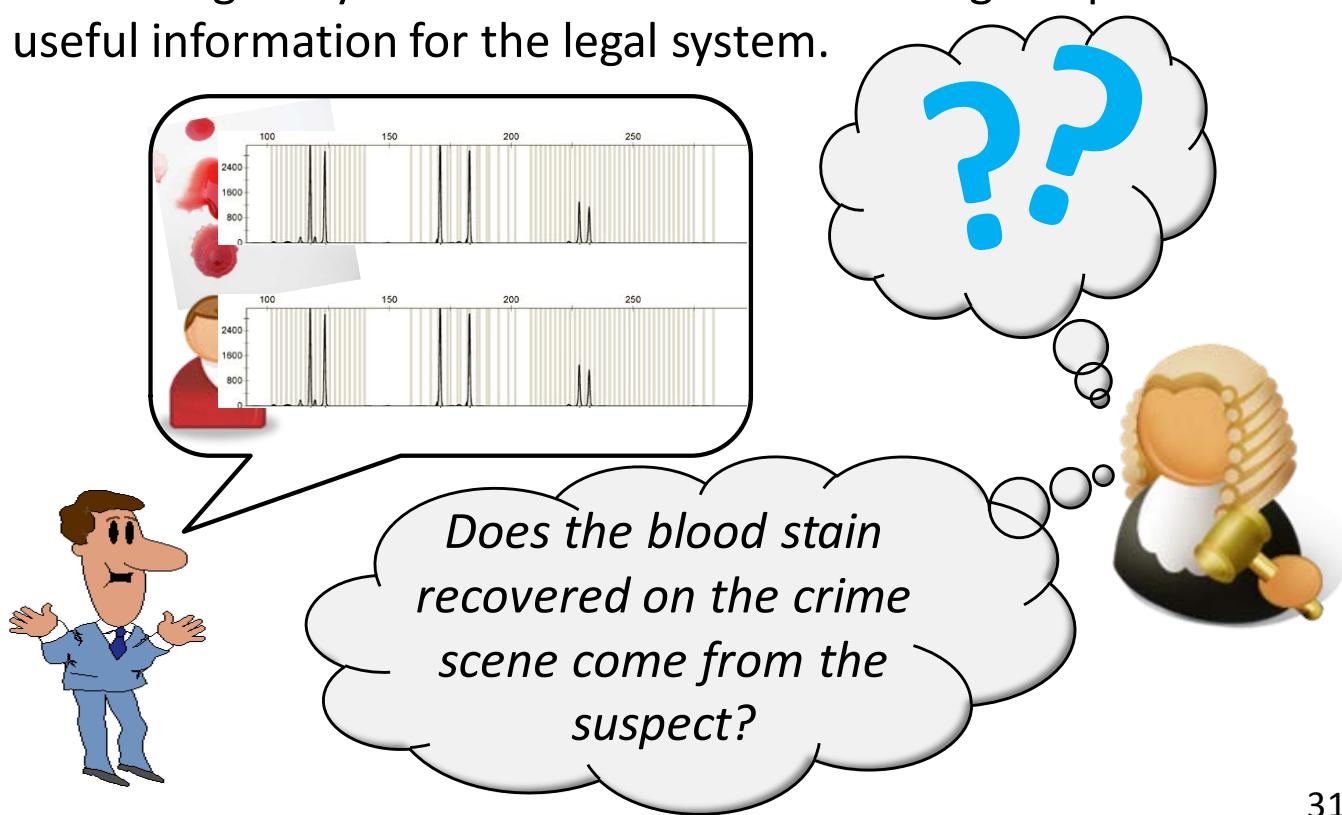


suspect

30

# Single Contributor Stain

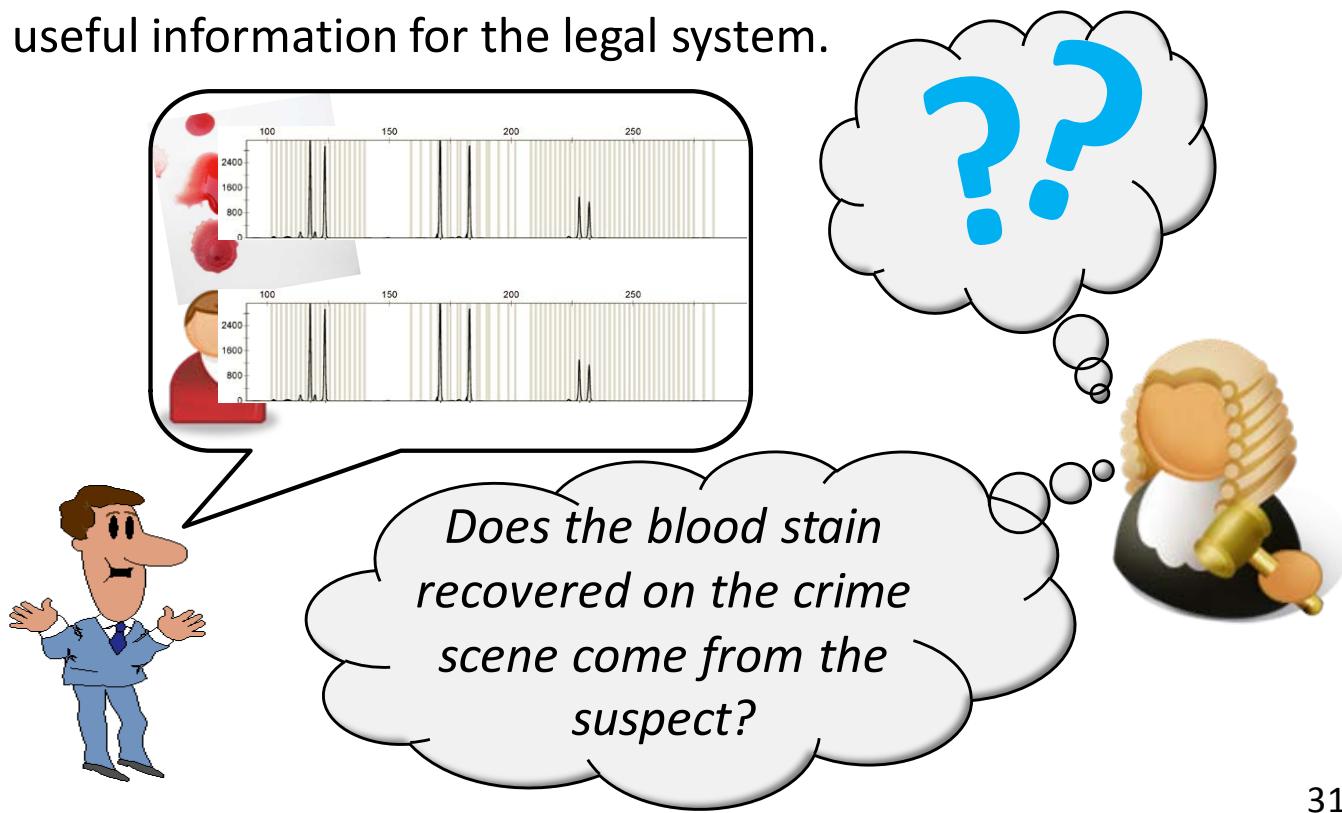
Presenting analytical results alone is not enough to provide useful information for the legal system.



31

# Single Contributor Stain

Presenting analytical results alone is not enough to provide useful information for the legal system.



31

# Relevant Evidence

For evidence to be admissible, it must be relevant.

## 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.



32

# Relevant Evidence

For evidence to be admissible, it must be relevant.

## 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.



32

# Relevant Evidence

The blood stain  
comes from the  
suspect.

## 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.

33

# Relevant Evidence

The blood stain  
comes from the  
suspect.

## 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.

33

# Relevant Evidence

**UNCERTAINTY**

The blood stain  
comes from the  
suspect.

## 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.

34

# Relevant Evidence

**UNCERTAINTY**

The blood stain  
comes from the  
suspect.

## Rule 401 of the US Federal Rules of Evidence

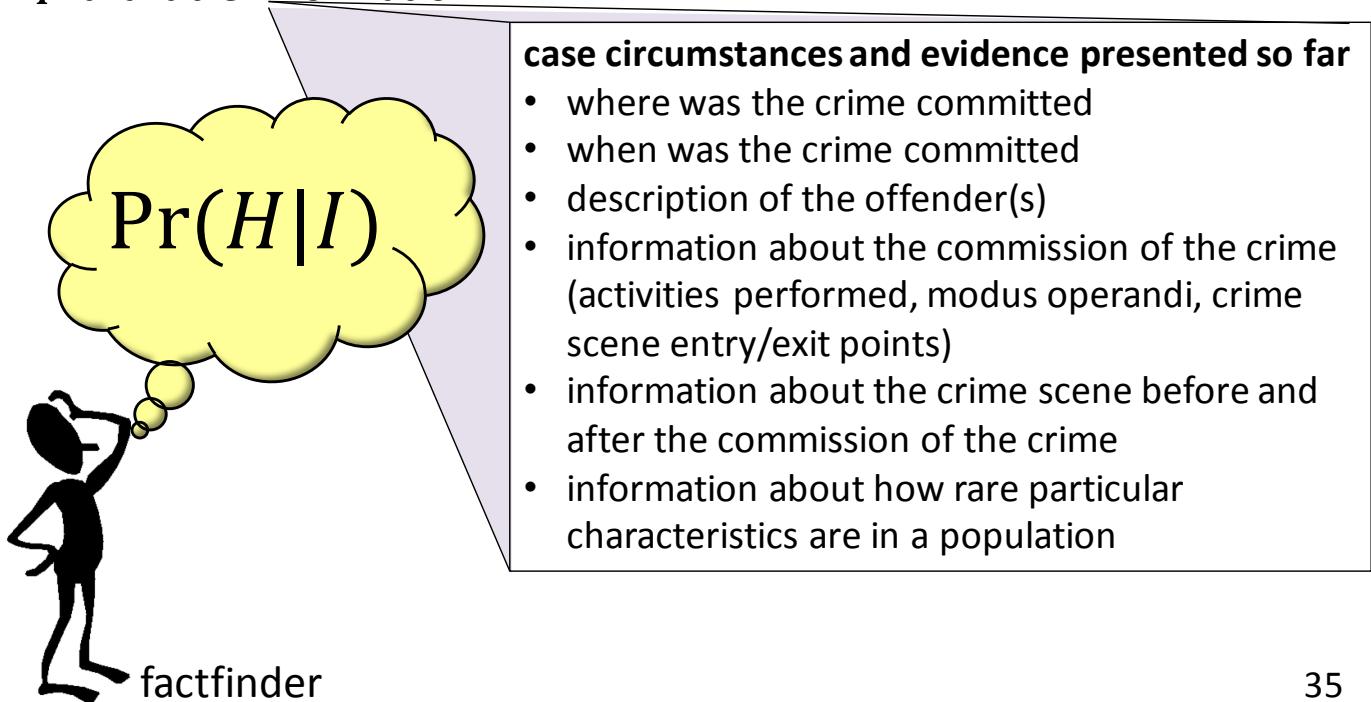
“Relevant evidence” means evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action **more probable or less probable** than it would be without the evidence.

34

# Logical Framework for Updating Uncertainty

$H$ : The crime stain comes from the suspect.

$I$ : available information

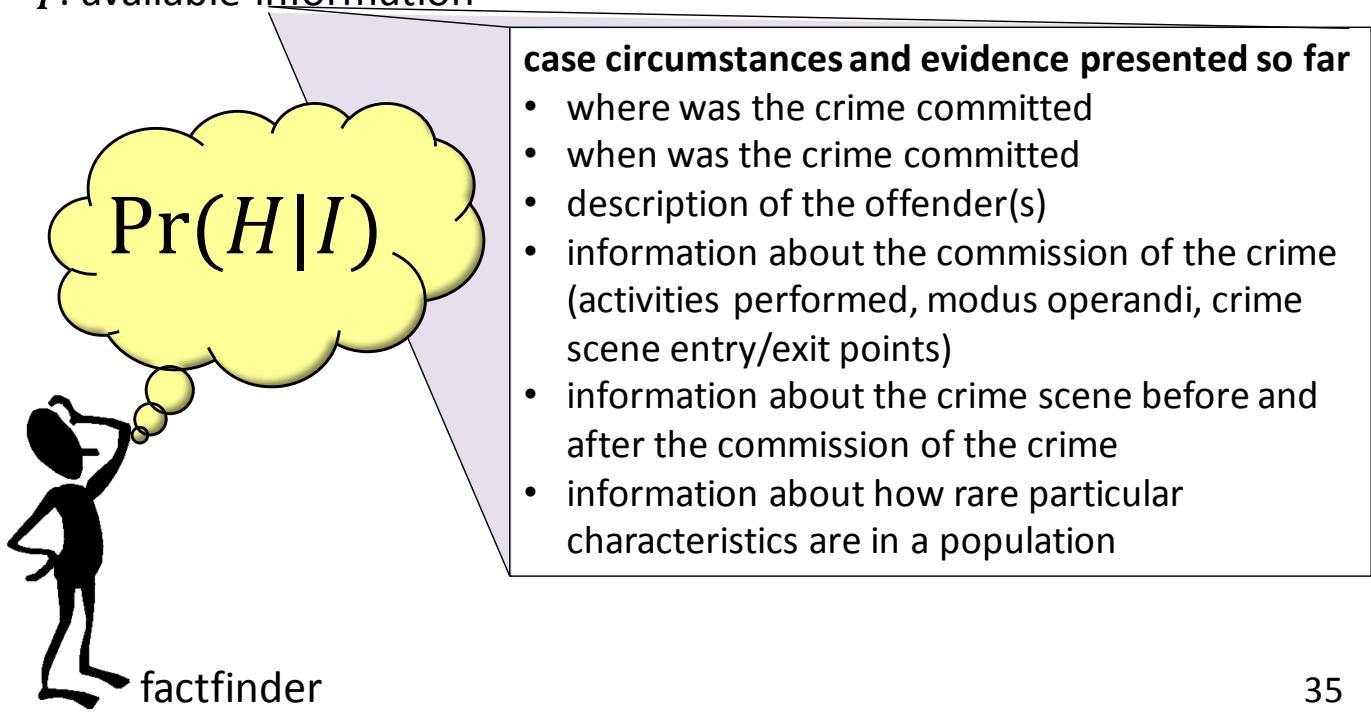


35

# Logical Framework for Updating Uncertainty

$H$ : The crime stain comes from the suspect.

$I$ : available information



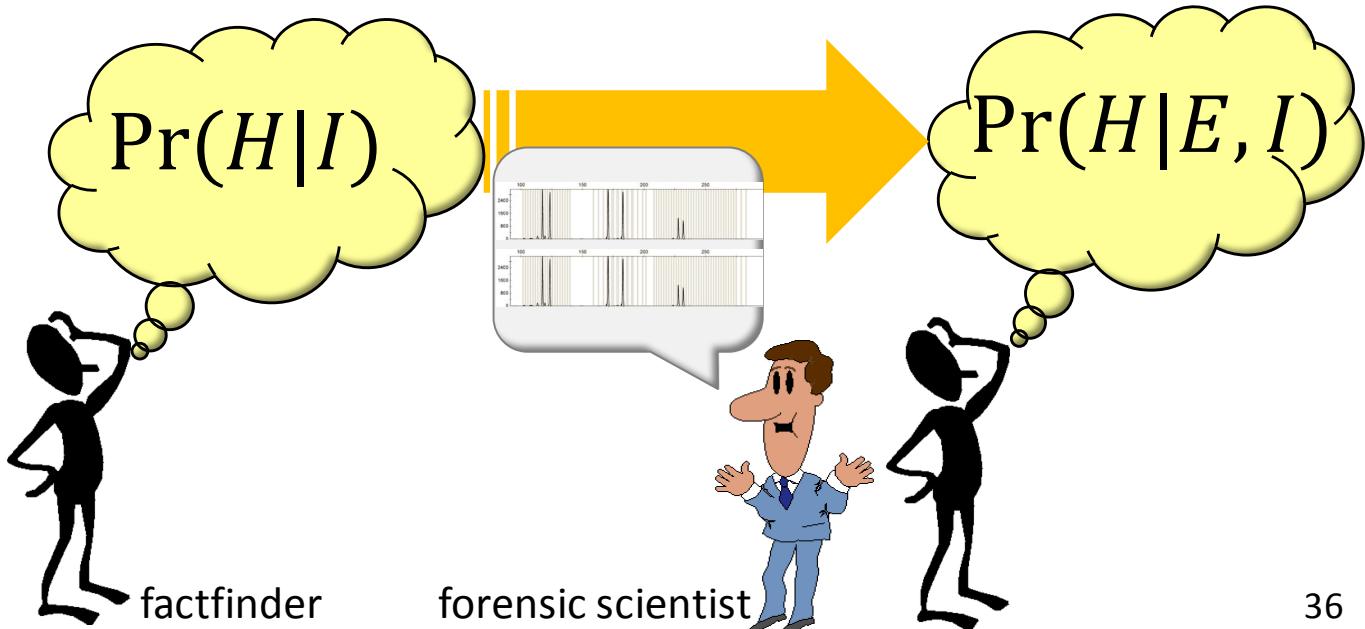
35

# Logical Framework for Updating Uncertainty

$H$ : The crime stain comes from the suspect.

$I$ : available information (case circumstances, evidence presented so far)

$E$ : observed analytical results



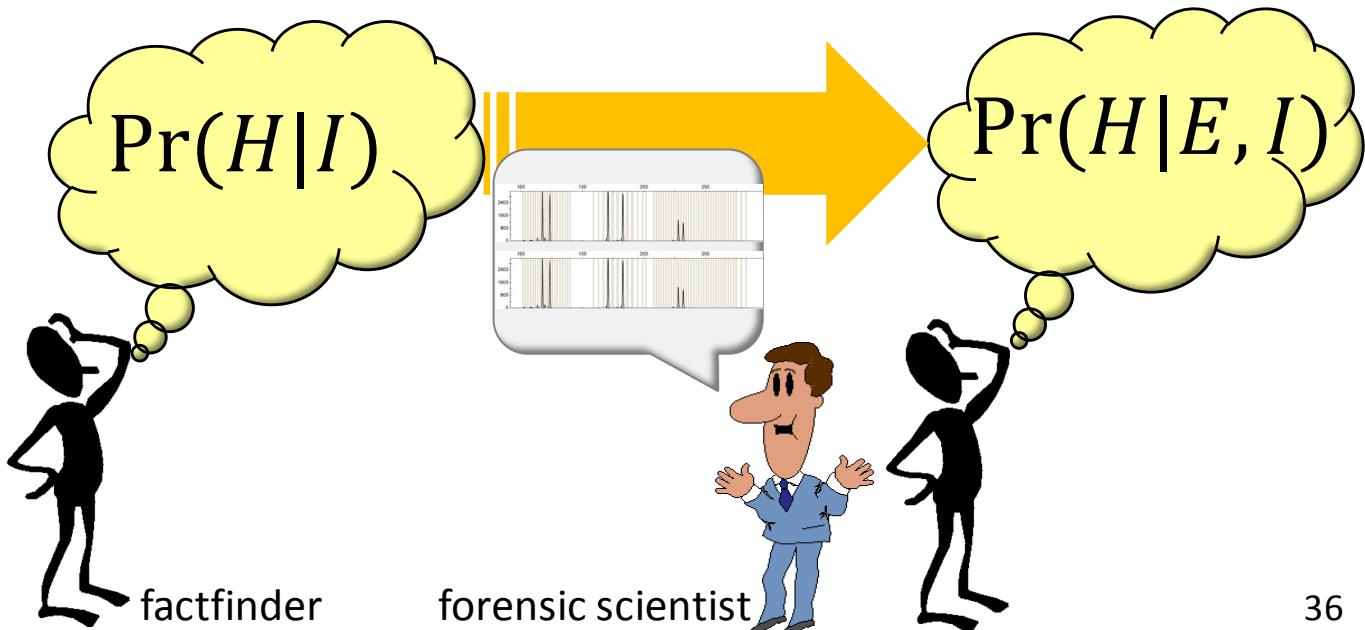
36

# Logical Framework for Updating Uncertainty

$H$ : The crime stain comes from the suspect.

$I$ : available information (case circumstances, evidence presented so far)

$E$ : observed analytical results



36

# Logical Framework for Updating Uncertainty

$H$ : The crime stain comes from the suspect.

$I$ : available information (case circumstances, evidence presented so far)

$E$ : observed analytical results

## Bayes' theorem:

$$\begin{aligned}\Pr(H|E, I) &= \frac{\Pr(H, E|I)}{\Pr(E|I)} \\ &= \frac{\Pr(E|H, I) \Pr(H|I)}{\Pr(E|I)}\end{aligned}$$

37

# Logical Framework for Updating Uncertainty

$H$ : The crime stain comes from the suspect.

$I$ : available information (case circumstances, evidence presented so far)

$E$ : observed analytical results

## Bayes' theorem:

$$\begin{aligned}\Pr(H|E, I) &= \frac{\Pr(H, E|I)}{\Pr(E|I)} \\ &= \frac{\Pr(E|H, I) \Pr(H|I)}{\Pr(E|I)}\end{aligned}$$

37

# Logical Framework for Updating Uncertainty

## Bayes' theorem:

$$\Pr(H|E, I) = \frac{\Pr(H, E|I)}{\Pr(E|I)}$$
$$= \frac{\Pr(E|H, I) \Pr(H|I)}{\Pr(E|I)}$$

I don't know  $\Pr(H|I)$ , and I need additional information to assign  $\Pr(E|I)$ .



38

# Logical Framework for Updating Uncertainty

## Bayes' theorem:

$$\Pr(H|E, I) = \frac{\Pr(H, E|I)}{\Pr(E|I)}$$
$$= \frac{\Pr(E|H, I) \Pr(H|I)}{\Pr(E|I)}$$

I don't know  $\Pr(H|I)$ , and I need additional information to assign  $\Pr(E|I)$ .



38

# Principles of Evidence Interpretation

- 1. To evaluate the uncertainty of a proposition, it is necessary to consider at least one alternative proposition.**
- 2.
- 3.

I.W. Evett and B.S. Weir. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer Associates, Sunderland, 1998: page 29.

39

# Principles of Evidence Interpretation

- 1. To evaluate the uncertainty of a proposition, it is necessary to consider at least one alternative proposition.**
- 2.
- 3.

I.W. Evett and B.S. Weir. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer Associates, Sunderland, 1998: page 29.

39

# Consider At Least One Alternative Proposition

$H_p$ : The crime stain comes from the suspect.



prosecution's proposition

$H_d$ : The crime stain does not come from the suspect. It comes from some other person.



defense's proposition

40

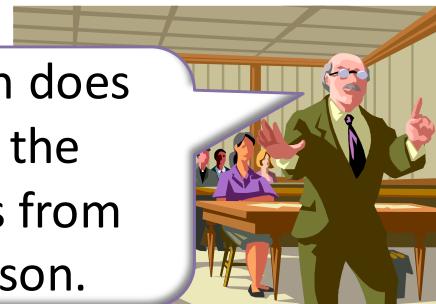
# Consider At Least One Alternative Proposition

$H_p$ : The crime stain comes from the suspect.



prosecution's proposition

$H_d$ : The crime stain does not come from the suspect. It comes from some other person.



defense's proposition

40

## Consider At Least One Alternative Proposition

$$\Pr(H_p|E, I) = \frac{\Pr(E|H_p, I)\Pr(H_p|I)}{\Pr(E|I)}$$

$$\Pr(H_d|E, I) = \frac{\Pr(E|H_d, I)\Pr(H_d|I)}{\Pr(E|I)}$$

41

## Consider At Least One Alternative Proposition

$$\Pr(H_p|E, I) = \frac{\Pr(E|H_p, I)\Pr(H_p|I)}{\Pr(E|I)}$$

$$\Pr(H_d|E, I) = \frac{\Pr(E|H_d, I)\Pr(H_d|I)}{\Pr(E|I)}$$

41

## Consider At Least One Alternative Proposition

$$\Pr(H_p|E, I) = \frac{\Pr(E|H_p, I)\Pr(H_p|I)}{\Pr(E|I)}$$

---

$$\Pr(H_d|E, I) = \frac{\Pr(E|H_d, I)\Pr(H_d|I)}{\Pr(E|I)}$$

42

## Consider At Least One Alternative Proposition

$$\Pr(H_p|E, I) = \frac{\Pr(E|H_p, I)\Pr(H_p|I)}{\Pr(E|I)}$$

---

$$\Pr(H_d|E, I) = \frac{\Pr(E|H_d, I)\Pr(H_d|I)}{\Pr(E|I)}$$

42

# Consider At Least One Alternative Proposition

**Bayes' theorem in the form of odds:**

$$\frac{\Pr(H_p|E, I)}{\Pr(H_d|E, I)} = \underbrace{\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}}_{\text{posterior odds}} \times \underbrace{\frac{\Pr(H_p|I)}{\Pr(H_d|I)}}_{\text{prior odds}}$$

**Likelihood Ratio  
(or Bayes Factor)**



43

# Consider At Least One Alternative Proposition

**Bayes' theorem in the form of odds:**

$$\frac{\Pr(H_p|E, I)}{\Pr(H_d|E, I)} = \underbrace{\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}}_{\text{posterior odds}} \times \underbrace{\frac{\Pr(H_p|I)}{\Pr(H_d|I)}}_{\text{prior odds}}$$

**Likelihood Ratio  
(or Bayes Factor)**



43

# Principles of Evidence Interpretation

1. To evaluate the uncertainty of a proposition, it is necessary to consider at least one alternative proposition.
2. **Scientific interpretation is based on questions of the kind “What is the probability of the evidence given the proposition?”**
- 3.

I.W. Evett and B.S. Weir. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer Associates, Sunderland, 1998: page 29.

44

# Principles of Evidence Interpretation

1. To evaluate the uncertainty of a proposition, it is necessary to consider at least one alternative proposition.
2. **Scientific interpretation is based on questions of the kind “What is the probability of the evidence given the proposition?”**
- 3.

I.W. Evett and B.S. Weir. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer Associates, Sunderland, 1998: page 29.

44

# Role of the Forensic Scientist

# Bayes' theorem in the form of odds:

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

court

expert witness  
(forensic scientist)

court

# Role of the Forensic Scientist

# Bayes' theorem in the form of odds:

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

court

court

# Likelihood Ratio (LR)

The probability of observing the analytical results given the available information and that the prosecution's proposition is true

$$\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}$$

divided by

the probability of observing the analytical results given the available information and that the defense's proposition is true.

46

# Likelihood Ratio (LR)

The probability of observing the analytical results given the available information and that the prosecution's proposition is true

$$\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}$$

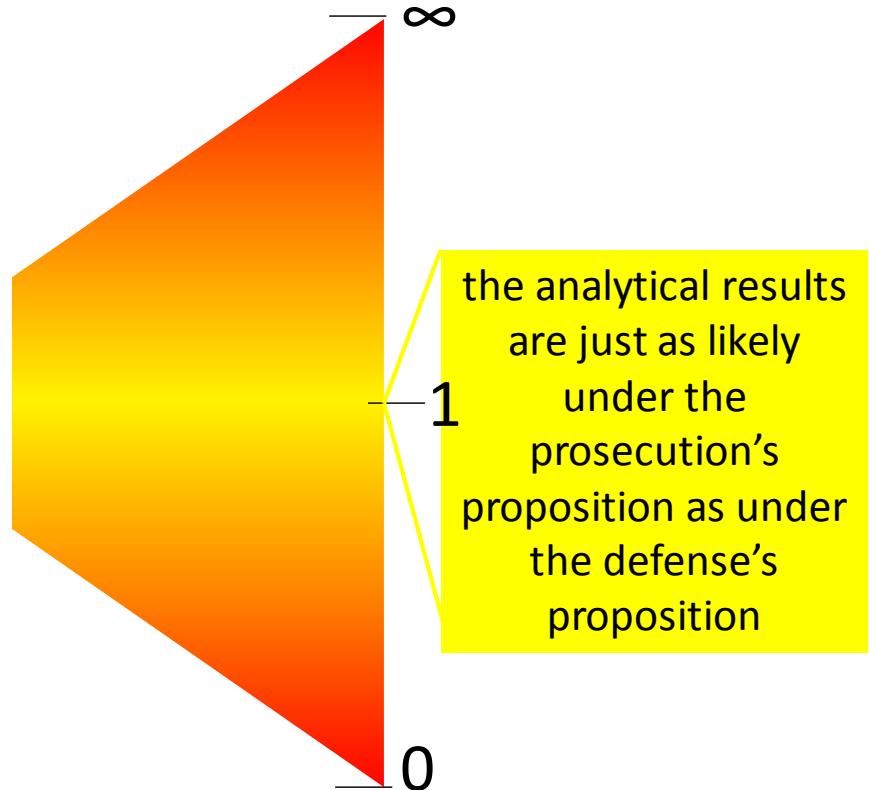
divided by

the probability of observing the analytical results given the available information and that the defense's proposition is true.

46

# Likelihood Ratio (LR)

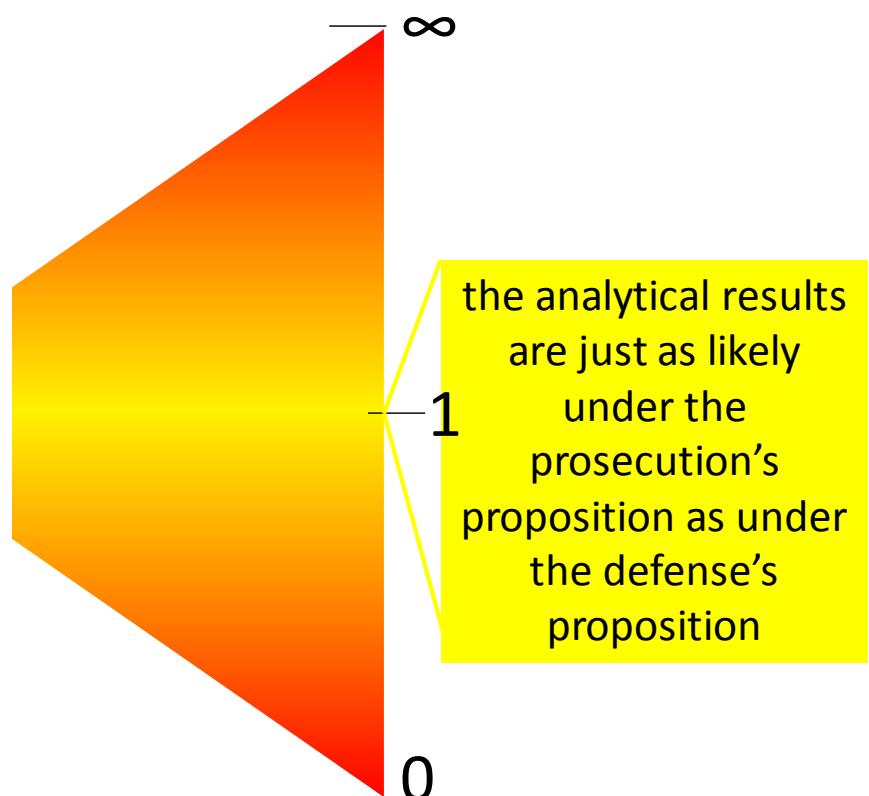
$$\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}$$



47

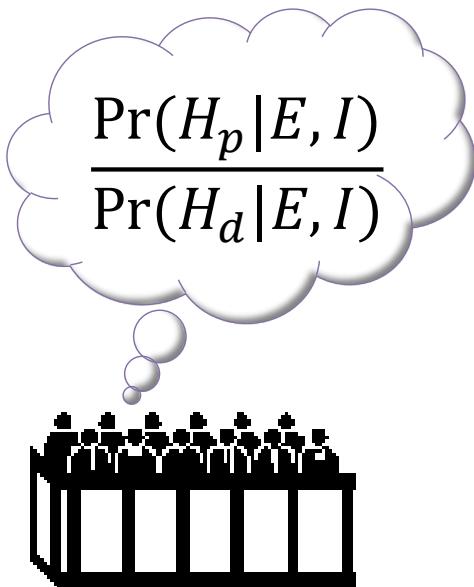
# Likelihood Ratio (LR)

$$\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}$$



47

# Role of the Factfinder

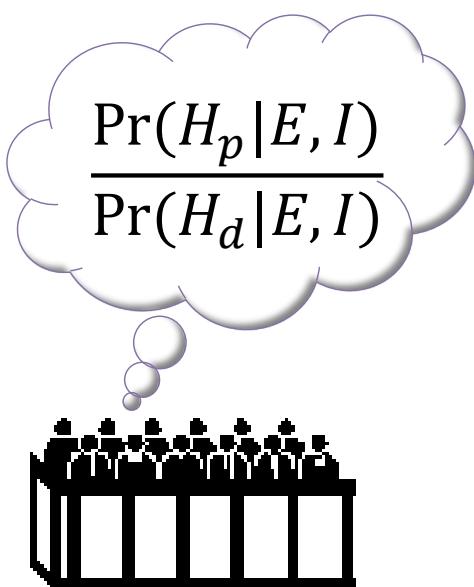


Given the evidence, what is the probability that the prosecution's proposition is true?

Given the evidence, what is the probability that the defense's proposition is true?

48

# Role of the Factfinder



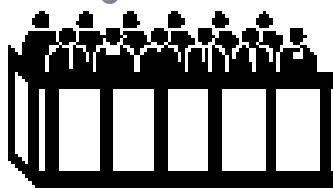
Given the evidence, what is the probability that the prosecution's proposition is true?

Given the evidence, what is the probability that the defense's proposition is true?

48

# Role of the Factfinder

$$\frac{\Pr(H_p | E, I)}{\Pr(H_d | E, I)}$$



The chance that the crime stain comes from the suspect is 0.9.

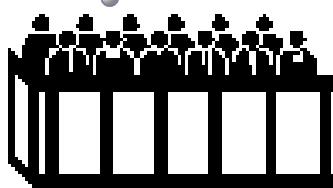
The probability that the crime stain comes from someone other than the suspect is 0.1.

There is a 1 in 10 chance that the crime stain does not come from the suspect.

49

# Role of the Factfinder

$$\frac{\Pr(H_p | E, I)}{\Pr(H_d | E, I)}$$



The chance that the crime stain comes from the suspect is 0.9.

The probability that the crime stain comes from someone other than the suspect is 0.1.

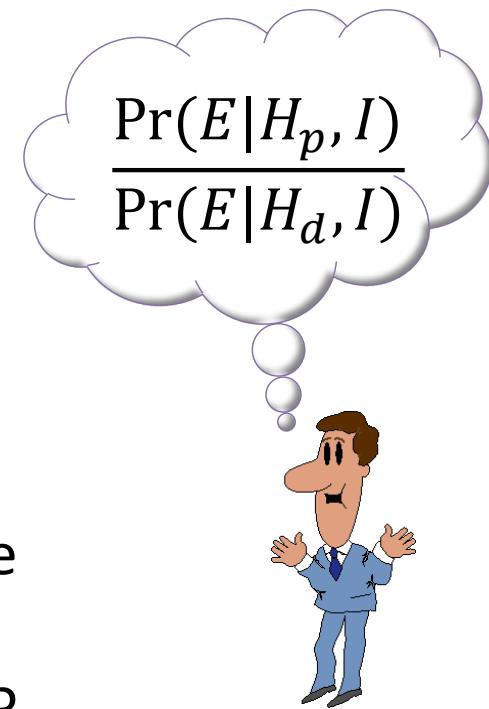
There is a 1 in 10 chance that the crime stain does not come from the suspect.

49

# Role of the Forensic Scientist

What is the probability of the analytical results if the prosecution's proposition is true?

What is the probability of the analytical results if the defense's proposition is true?

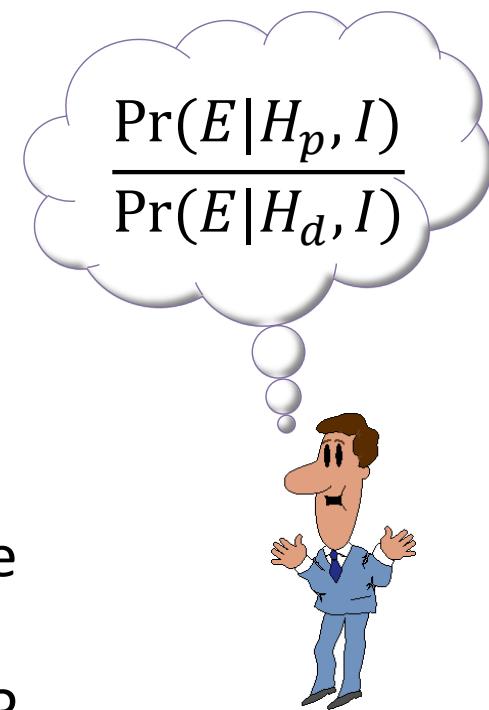


50

# Role of the Forensic Scientist

What is the probability of the analytical results if the prosecution's proposition is true?

What is the probability of the analytical results if the defense's proposition is true?



50

# Role of the Forensic Scientist

The probability of obtaining these DNA results if the crime stain comes from the suspect is very close to 1.

$$\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}$$

The chance of obtaining these DNA results if the crime stain comes from some other person, unrelated to the suspect, is 1 in 1 million.



51

# Role of the Forensic Scientist

The probability of obtaining these DNA results if the crime stain comes from the suspect is very close to 1.

$$\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}$$

The chance of obtaining these DNA results if the crime stain comes from some other person, unrelated to the suspect, is 1 in 1 million.



51

# Principles of Evidence Interpretation

1. To evaluate the uncertainty of a proposition, it is necessary to consider at least one alternative proposition.
2. Scientific interpretation is based on questions of the kind “What is the probability of the evidence given the proposition?”
3. **Scientific interpretation is conditioned not only by the competing propositions, but also by the framework of circumstances within which they are to be evaluated.**

I.W. Evett and B.S. Weir. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer Associates, Sunderland, 1998: page 29.

52

# Principles of Evidence Interpretation

1. To evaluate the uncertainty of a proposition, it is necessary to consider at least one alternative proposition.
2. Scientific interpretation is based on questions of the kind “What is the probability of the evidence given the proposition?”
3. **Scientific interpretation is conditioned not only by the competing propositions, but also by the framework of circumstances within which they are to be evaluated.**

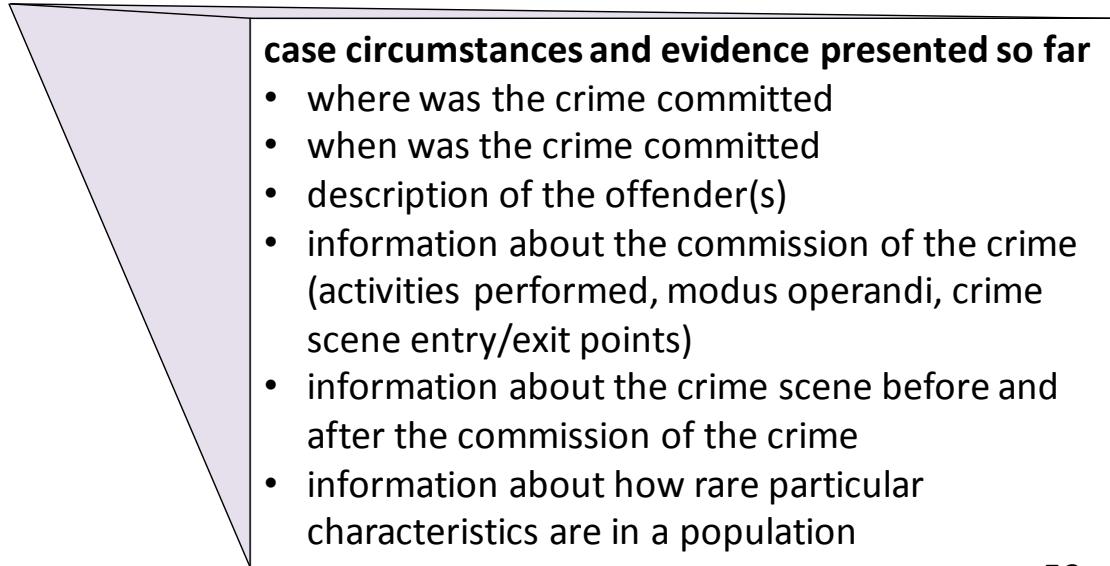
I.W. Evett and B.S. Weir. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer Associates, Sunderland, 1998: page 29.

52

# Conditioning on the Framework of Circumstances

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

with  $I$ : available information

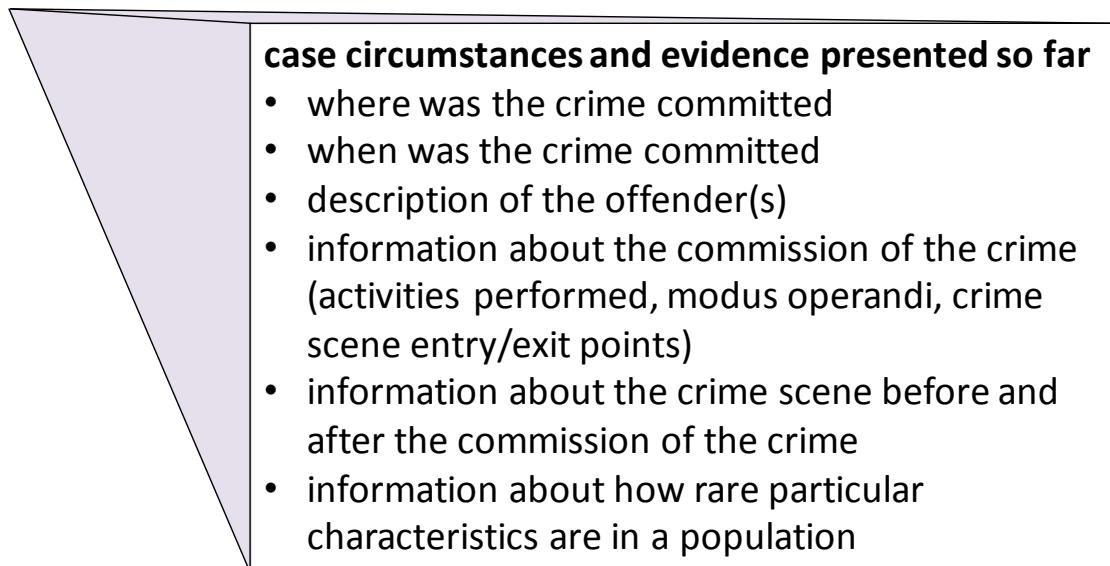


53

# Conditioning on the Framework of Circumstances

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

with  $I$ : available information



53

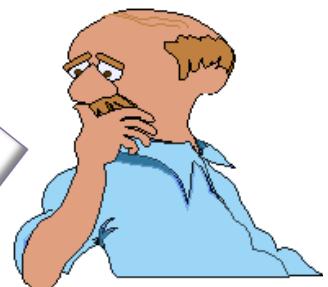
# Conditioning on the Framework of Circumstances

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

with  $I$ :

What do we know or assume about the offender? What population does the offender come from?

What are the genetic properties of this population? What do we know about the rarity of the observed genotype in this population?



54

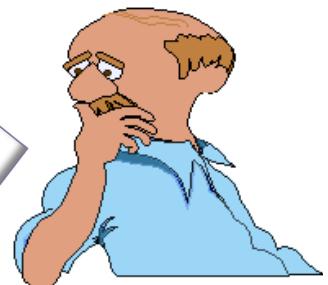
# Conditioning on the Framework of Circumstances

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

with  $I$ :

What do we know or assume about the offender? What population does the offender come from?

What are the genetic properties of this population? What do we know about the rarity of the observed genotype in this population?



54

## Conditioning on the Framework of Circumstances

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

The LR will vary according to the information in  $I$ .



It is therefore imperative for the forensic scientist to make explicit to the court what information makes up the  $I$  in his/her LR. If the court disagrees, or new information becomes available, the forensic scientist must re-assign the probabilities forming the LR conditioned on the new  $I$ .

55

## Conditioning on the Framework of Circumstances

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

The LR will vary according to the information in  $I$ .



It is therefore imperative for the forensic scientist to make explicit to the court what information makes up the  $I$  in his/her LR. If the court disagrees, or new information becomes available, the forensic scientist must re-assign the probabilities forming the LR conditioned on the new  $I$ .

55

# True or False?

$$\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)} = 100 \text{ , or abbreviated as } LR = 100$$



Given the available information, the probability of observing these analytical results is 100 times greater if the prosecution's proposition is true than if the defense's proposition is true.

Given the available information, the observation of the analytical results indicate that the probability of the prosecution's proposition being true is 100 times greater than the probability of the defense's proposition being true.

56

# True or False?

$$\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)} = 100 \text{ , or abbreviated as } LR = 100$$



Given the available information, the probability of observing these analytical results is 100 times greater if the prosecution's proposition is true than if the defense's proposition is true.

Given the available information, the observation of the analytical results indicate that the probability of the prosecution's proposition being true is 100 times greater than the probability of the defense's proposition being true.

56

# True or False?

$$\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)} = 100 \text{ , or abbreviated as } LR = 100$$



Given the available information, the probability of observing these analytical results is 100 times greater if the prosecution's proposition is true than if the defense's proposition is true.

Given the available information, the observation of the analytical results indicate that the probability of the prosecution's proposition being true is 100 times greater than the probability of the defense's proposition being true.

57

# True or False?

$$\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)} = 100 \text{ , or abbreviated as } LR = 100$$



Given the available information, the probability of observing these analytical results is 100 times greater if the prosecution's proposition is true than if the defense's proposition is true.

Given the available information, the observation of the analytical results indicate that the probability of the prosecution's proposition being true is 100 times greater than the probability of the defense's proposition being true.

57

# Transposed Conditional

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

100                                    100

Given the available information, the probability of observing these analytical results is 100 times greater if the prosecution's proposition is true than if the defense's proposition is true.

Given the available information, the observation of the analytical results indicate that the probability of the prosecution's proposition being true is 100 times greater than the probability of the defense's proposition being true.

58

# Transposed Conditional

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

100                                    100

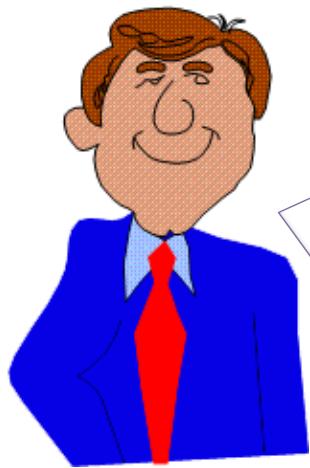
Given the available information, the probability of observing these analytical results is 100 times greater if the prosecution's proposition is true than if the defense's proposition is true.

Given the available information, the observation of the analytical results indicate that the probability of the prosecution's proposition being true is 100 times greater than the probability of the defense's proposition being true.

58

# Prosecutor's Fallacy

$$\Pr(E|H_d, I) = 1 \text{ in 7 million}$$



In layman's terms, just so I get this right, are you saying that the probability that the DNA that was found in the question samples came from anyone else besides A.L. is one in 7 million (...)?



Yes, approximately.

State v. Lee, No. 90CA004741 (Ohio App. Dec. 5, 1990), transcript at 464

59

# Prosecutor's Fallacy

$$\Pr(E|H_d, I) = 1 \text{ in 7 million}$$



In layman's terms, just so I get this right, are you saying that the probability that the DNA that was found in the question samples came from anyone else besides A.L. is one in 7 million (...)?



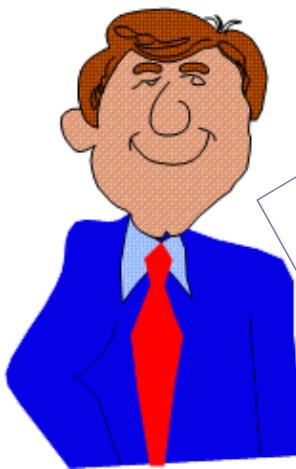
Yes, approximately.

State v. Lee, No. 90CA004741 (Ohio App. Dec. 5, 1990), transcript at 464

59

# Prosecutor's Fallacy

The frequency of the genetic profile in the reference population is 1 in 10 million



The witness concludes that the genetic profile of the two analyzed samples match perfectly, and he deduces that the probability of someone other than the suspect being the source of the trace found on the victim's cloths is 1 in 10 million.

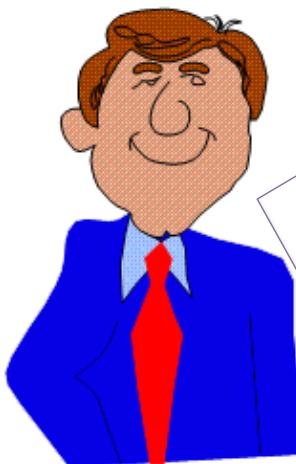


modified from:  
State of Arizona v. Michael Steven Gallegos [178 Ariz. 1; 870 P.2d 1097 (1994)]

60

# Prosecutor's Fallacy

The frequency of the genetic profile in the reference population is 1 in 10 million



The witness concludes that the genetic profile of the two analyzed samples match perfectly, and he deduces that the probability of someone other than the suspect being the source of the trace found on the victim's cloths is 1 in 10 million.

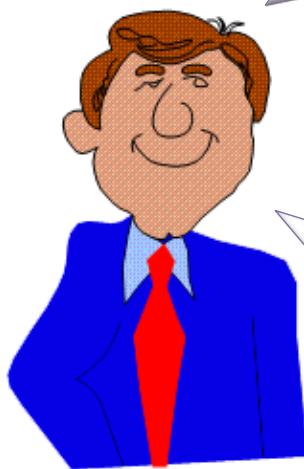


modified from:  
State of Arizona v. Michael Steven Gallegos [178 Ariz. 1; 870 P.2d 1097 (1994)]

60

# Prosecutor's Fallacy

Are you able to (...) determine what is the likelihood of the DNA found in B.G. just randomly occurring in some other DNA sample?



Yes (...) and the final number comes out as 1 in 18 billion



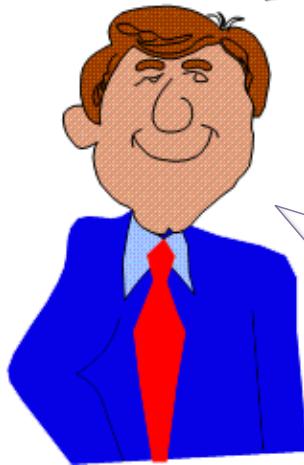
So the likelihood that DNA belongs to someone other than B.G. is one in 18 billion?

State v. Glover, 825 S.W.2d 127 (Tex. Crim.App. 1992), transcript at 413

61

# Prosecutor's Fallacy

Are you able to (...) determine what is the likelihood of the DNA found in B.G. just randomly occurring in some other DNA sample?



Yes (...) and the final number comes out as 1 in 18 billion



So the likelihood that DNA belongs to someone other than B.G. is one in 18 billion?

State v. Glover, 825 S.W.2d 127 (Tex. Crim.App. 1992), transcript at 413

61

# Prosecutor's Fallacy

The fallacy is to transpose the conditional:

- $\Pr(E|H_d, I) = \Pr(H_d|E, I)$

or

- $\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)} = \frac{\Pr(H_p|E, I)}{\Pr(H_d|E, I)}$

so that a low  $\Pr(E|H_d, I)$  is expressed as a low  $\Pr(H_d|E, I)$  or high  $\frac{\Pr(H_p|E, I)}{\Pr(H_d|E, I)}$  when the prior odds are not necessarily equal to 1.

62

# Prosecutor's Fallacy

The fallacy is to transpose the conditional:

- $\Pr(E|H_d, I) = \Pr(H_d|E, I)$

or

- $\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)} = \frac{\Pr(H_p|E, I)}{\Pr(H_d|E, I)}$

so that a low  $\Pr(E|H_d, I)$  is expressed as a low  $\Pr(H_d|E, I)$  or high  $\frac{\Pr(H_p|E, I)}{\Pr(H_d|E, I)}$  when the prior odds are not necessarily equal to 1.

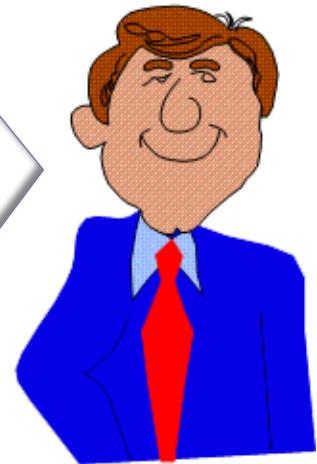
62

# Defense Attorney's Fallacy



$$\Pr(E|H_d, I) = 1 \text{ in } 1,000$$

The city where the crime occurred has a population of 200,000. In this city, this genotype would be found in 200 people. Therefore the evidence merely shows that the suspect is one of 200 people in the city who might have committed the crime.



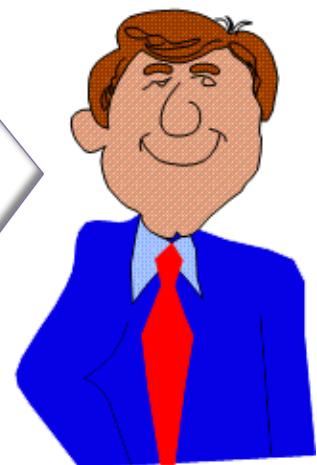
modified from: W.C. Thompson and E.L. Schumann. Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and defence attorney's fallacy. *Law and Human Behaviour*, 11: 167-187, 1987. 63

# Defense Attorney's Fallacy



$$\Pr(E|H_d, I) = 1 \text{ in } 1,000$$

The city where the crime occurred has a population of 200,000. In this city, this genotype would be found in 200 people. Therefore the evidence merely shows that the suspect is one of 200 people in the city who might have committed the crime.



modified from: W.C. Thompson and E.L. Schumann. Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and defence attorney's fallacy. *Law and Human Behaviour*, 11: 167-187, 1987. 63

# Defense Attorney's Fallacy

The fallacy is:

1. To assume that each of these 200 individuals has the **same prior probability** of being the source of the crime stain
2. To assume that the **actual number** of individuals in this city having the genotype in question is equal to the **expected number** of individuals having this genotype. The actual number could be anywhere between 1 and 200,000.

64

# Defense Attorney's Fallacy

The fallacy is:

1. To assume that each of these 200 individuals has the **same prior probability** of being the source of the crime stain
2. To assume that the **actual number** of individuals in this city having the genotype in question is equal to the **expected number** of individuals having this genotype. The actual number could be anywhere between 1 and 200,000.

64

# Correct



$$\Pr(E|H_d, I) = 1 \text{ in } 1,000$$

The city where the crime occurred has a population of 200,000. In this city, we would **expect** to find this genotype in 200 people.



65

# Correct



$$\Pr(E|H_d, I) = 1 \text{ in } 1,000$$

The city where the crime occurred has a population of 200,000. In this city, we would **expect** to find this genotype in 200 people.



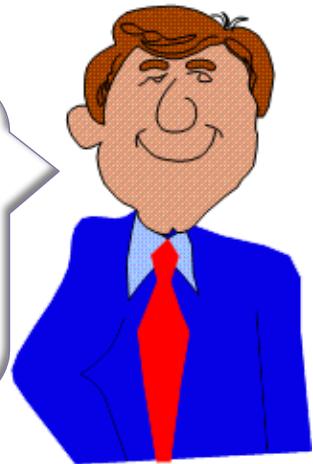
65

# Uniqueness Fallacy



$$\Pr(E|H_d, I) = 1 \text{ in } 200,000$$

The city where the crime occurred has a population of 200,000. In this city, this genotype can therefore only come from the suspect.



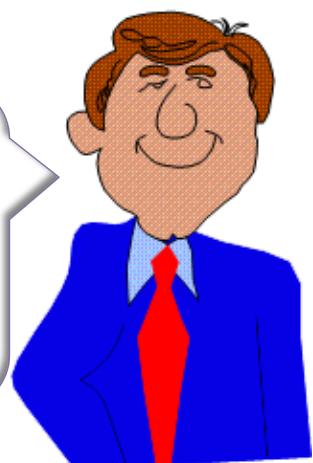
66

# Uniqueness Fallacy



$$\Pr(E|H_d, I) = 1 \text{ in } 200,000$$

The city where the crime occurred has a population of 200,000. In this city, this genotype can therefore only come from the suspect.



66

# Uniqueness Fallacy

The fallacy is to assume that the **actual number** of individuals in this city having the genotype in question is equal to the **expected number** of individuals having this genotype. The actual number could be anywhere between 1 and 200,000.

67

# Uniqueness Fallacy

The fallacy is to assume that the **actual number** of individuals in this city having the genotype in question is equal to the **expected number** of individuals having this genotype. The actual number could be anywhere between 1 and 200,000.

67

# Correct



$$\Pr(E|H_d, I) = 1 \text{ in } 200,000$$

The city where the crime occurred has a population of 200,000. In this city, we would **expect** to find this genotype in 1 person.



68

# Correct

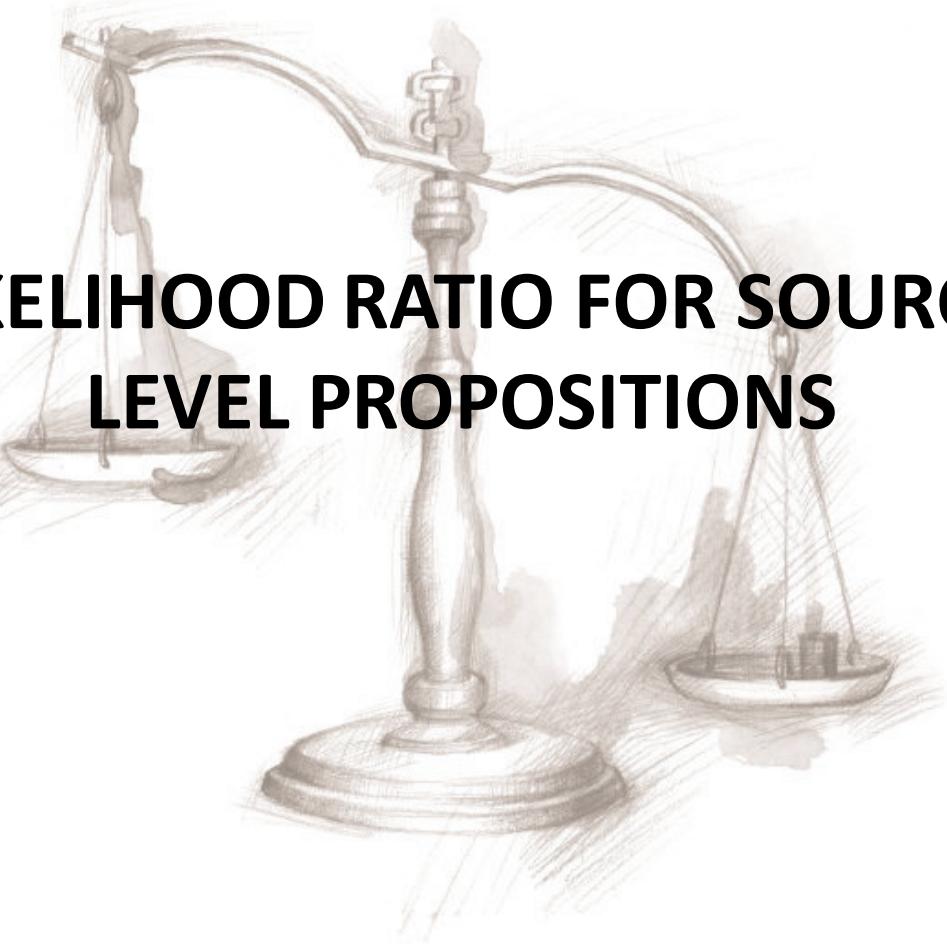


$$\Pr(E|H_d, I) = 1 \text{ in } 200,000$$

The city where the crime occurred has a population of 200,000. In this city, we would **expect** to find this genotype in 1 person.



68



## **LIKELIHOOD RATIO FOR SOURCE LEVEL PROPOSITIONS**

69



## **LIKELIHOOD RATIO FOR SOURCE LEVEL PROPOSITIONS**

69

# LR for Source Level Propositions

Source level propositions:

$H_p$ : The crime stain **comes from** the suspect.

$H_d$ : The crime stain **comes from** some other person.

70

# LR for Source Level Propositions

Source level propositions:

$H_p$ : The crime stain **comes from** the suspect.

$H_d$ : The crime stain **comes from** some other person.

70

# LR for Source Level Propositions

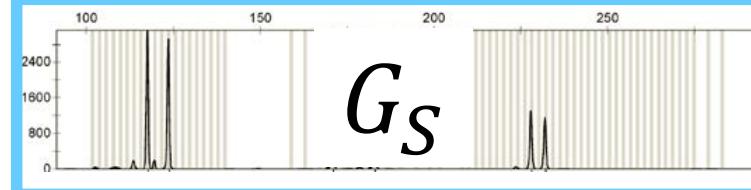
Source level propositions:

$H_p$ : The crime stain **comes from** the suspect.

$H_d$ : The crime stain **comes from** some other person.



$G_C$



$G_S$

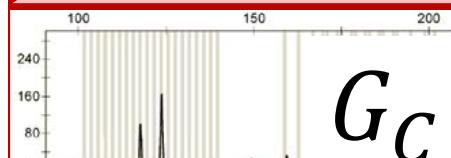
71

# LR for Source Level Propositions

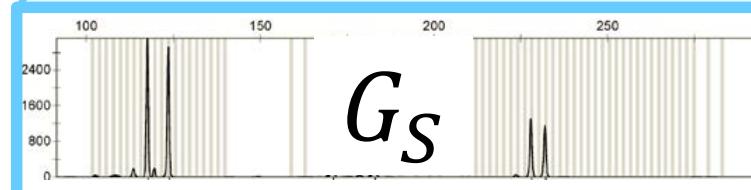
Source level propositions:

$H_p$ : The crime stain **comes from** the suspect.

$H_d$ : The crime stain **comes from** some other person.



$G_C$



$G_S$

71

# LR for Source Level Propositions

$$\begin{aligned}
 LR &= \frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)} \\
 &= \frac{\Pr(G_C, G_S|H_p, I)}{\Pr(G_C, G_S|H_d, I)} \\
 &= \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)} \times \frac{\Pr(G_S|H_p, I)}{\Pr(G_S|H_d, I)} \\
 &= \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)} \times \underbrace{\frac{\Pr(G_S|I)}{\Pr(G_S|I)}}_1
 \end{aligned}$$

The suspect's genotype does not depend on  $H_p$  being true or  $H_d$  being true.

72

# LR for Source Level Propositions

$$\begin{aligned}
 LR &= \frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)} \\
 &= \frac{\Pr(G_C, G_S|H_p, I)}{\Pr(G_C, G_S|H_d, I)} \\
 &= \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)} \times \frac{\Pr(G_S|H_p, I)}{\Pr(G_S|H_d, I)} \\
 &= \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)} \times \underbrace{\frac{\Pr(G_S|I)}{\Pr(G_S|I)}}_1
 \end{aligned}$$

The suspect's genotype does not depend on  $H_p$  being true or  $H_d$  being true.

72

# LR for Source Level Propositions

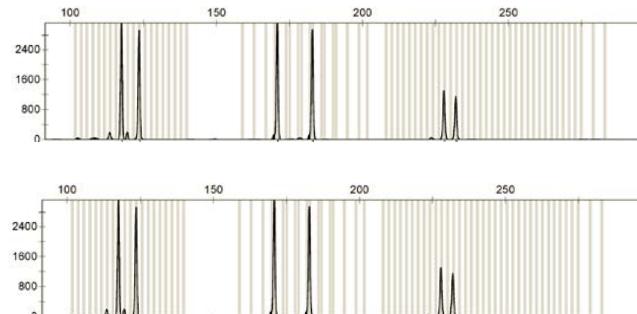
$$LR = \frac{\Pr(G_C | G_S, H_p, I)}{\Pr(G_C | G_S, H_d, I)}$$

## Numerator

the probability of observing the analytical results of the crime stain if the crime stain comes from the suspect and given the analytical results of the suspect's sample and other available information

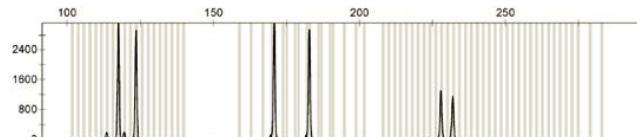


$G_C$ :



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

$G_S$ :



73

# LR for Source Level Propositions

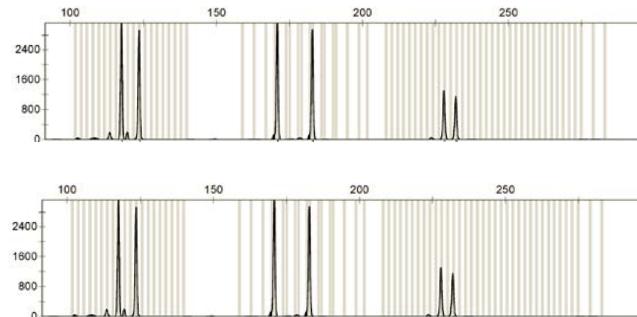
$$LR = \frac{\Pr(G_C | G_S, H_p, I)}{\Pr(G_C | G_S, H_d, I)}$$

## Numerator

the probability of observing the analytical results of the crime stain if the crime stain comes from the suspect and given the analytical results of the suspect's sample and other available information



$G_C$ :



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

$G_S$ :

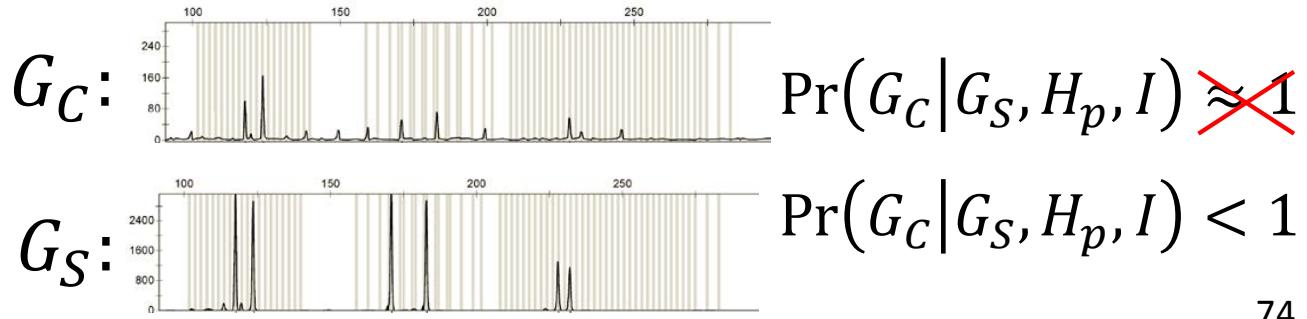
73

# LR for Source Level Propositions

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

## Numerator

the probability of observing the analytical results of the crime stain if the crime stain comes from the suspect and given the analytical results of the suspect's sample and other available information



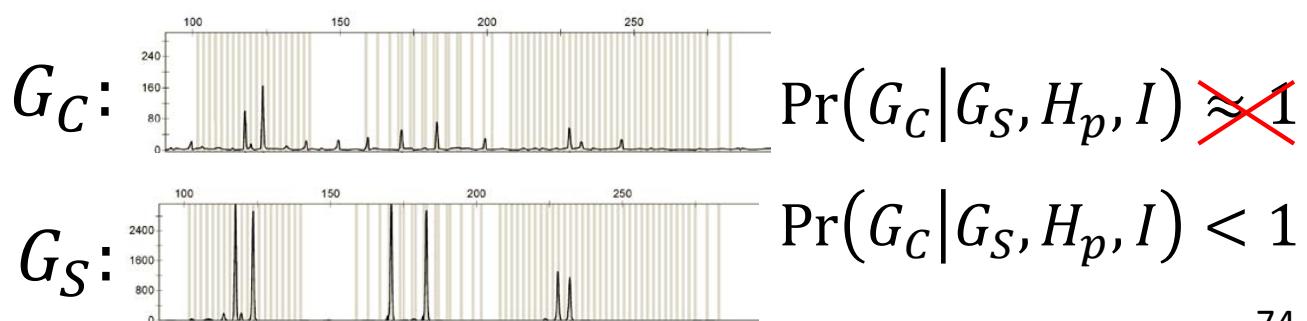
74

# LR for Source Level Propositions

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

## Numerator

the probability of observing the analytical results of the crime stain if the crime stain comes from the suspect and given the analytical results of the suspect's sample and other available information



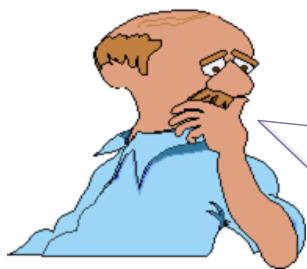
74

# LR for Source Level Propositions

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

## Denominator

the probability of observing the analytical results of the crime stain if the crime stain comes from some other person and given the analytical results of the suspect's sample and the available information



What is the probability of observing a second person with this genotype given that we have already observed one person with this genotype?

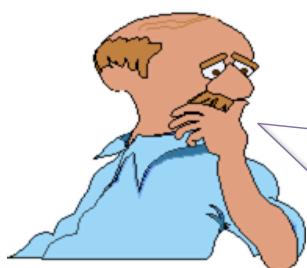
75

# LR for Source Level Propositions

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

## Denominator

the probability of observing the analytical results of the crime stain if the crime stain comes from some other person and given the analytical results of the suspect's sample and the available information



What is the probability of observing a second person with this genotype given that we have already observed one person with this genotype?

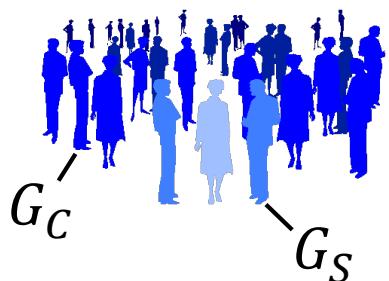
75

# LR for Source Level Propositions

Denominator

ASSUMPTION:

The probability of observing  $G_C$  is independent of the genotype observed for  $G_S$ .



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

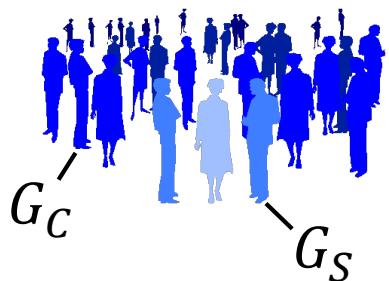
76

# LR for Source Level Propositions

Denominator

ASSUMPTION:

The probability of observing  $G_C$  is independent of the genotype observed for  $G_S$ .



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

76

# LR for Source Level Propositions

Denominator



ASSUMPTION:

The probability of observing  $G_C$  is not independent of the genotype observed for  $G_S$ . There is a probability  $\theta > 0$  that the crime stain's donor and the suspect share an allele passed down from a common ancestor.

The **coancestry coefficient  $\theta$**  is the probability that two individuals have an allele **identical by descent (ibd)**.

$$\Pr(G_C | G_S, H_d, I) \neq \Pr(G_C | H_d, I)$$

77

# LR for Source Level Propositions

Denominator



ASSUMPTION:

The probability of observing  $G_C$  is not independent of the genotype observed for  $G_S$ . There is a probability  $\theta > 0$  that the crime stain's donor and the suspect share an allele passed down from a common ancestor.

The **coancestry coefficient  $\theta$**  is the probability that two individuals have an allele **identical by descent (ibd)**.

$$\Pr(G_C | G_S, H_d, I) \neq \Pr(G_C | H_d, I)$$

77

# Dirichlet Distribution

ASSUMPTION:

A random mating population has reached an evolutionary equilibrium.

The allele proportions in this population satisfy a Dirichlet distribution. The probability of observing allele  $A$  if a total of  $n$  alleles containing  $n_A$  occurrences of allele  $A$  have already been observed is:

$$\Pr(A|n_A, n) = \frac{n_A \theta + (1 - \theta)p_A}{1 + (n - 1)\theta}$$

78

# Dirichlet Distribution

ASSUMPTION:

A random mating population has reached an evolutionary equilibrium.

The allele proportions in this population satisfy a Dirichlet distribution. The probability of observing allele  $A$  if a total of  $n$  alleles containing  $n_A$  occurrences of allele  $A$  have already been observed is:

$$\Pr(A|n_A, n) = \frac{n_A \theta + (1 - \theta)p_A}{1 + (n - 1)\theta}$$

78

# Dirichlet Distribution

$$\Pr(A|n_A, n) = \frac{n_A \theta + (1 - \theta)p_A}{1 + (n - 1)\theta}$$

gives us:

$$\Pr(A) = p_A$$

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

$$\Pr(A|A) = \theta + (1 - \theta)p_A$$

$$\Pr(A|AAA) = \frac{3\theta + (1 - \theta)p_A}{1 + 2\theta}$$

and applying the third law of probability, we obtain:

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

$$\begin{aligned}\Pr(AAA) &= \Pr(AA) \Pr(A|AA) \\ &= \frac{p_A[\theta + (1 - \theta)p_A][2\theta + (1 - \theta)p_A]}{1 + \theta}\end{aligned}$$

79

# Dirichlet Distribution

$$\Pr(A|n_A, n) = \frac{n_A \theta + (1 - \theta)p_A}{1 + (n - 1)\theta}$$

gives us:

$$\Pr(A) = p_A$$

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

$$\Pr(A|A) = \theta + (1 - \theta)p_A$$

$$\Pr(A|AAA) = \frac{3\theta + (1 - \theta)p_A}{1 + 2\theta}$$

and applying the third law of probability, we obtain:

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

$$\begin{aligned}\Pr(AAA) &= \Pr(AA) \Pr(A|AA) \\ &= \frac{p_A[\theta + (1 - \theta)p_A][2\theta + (1 - \theta)p_A]}{1 + \theta}\end{aligned}$$

79

# Dirichlet Distribution

$$\Pr(A|n_A, n) = \frac{n_A \theta + (1 - \theta)p_A}{1 + (n - 1)\theta}$$

gives us:

$$\Pr(A) = p_A$$

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

$$\Pr(A|A) = \theta + (1 - \theta)p_A$$

$$\Pr(A|AAA) = \frac{3\theta + (1 - \theta)p_A}{1 + 2\theta}$$

and applying the third law of probability, we obtain:

$$\begin{aligned} \Pr(AAAA) &= \Pr(AAA) \Pr(A|AAA) \\ &= \frac{p_A[\theta + (1 - \theta)p_A][2\theta + (1 - \theta)p_A][3\theta + (1 - \theta)p_A]}{(1 + \theta)(1 + 2\theta)} \end{aligned}$$

80

# Dirichlet Distribution

$$\Pr(A|n_A, n) = \frac{n_A \theta + (1 - \theta)p_A}{1 + (n - 1)\theta}$$

gives us:

$$\Pr(A) = p_A$$

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

$$\Pr(A|A) = \theta + (1 - \theta)p_A$$

$$\Pr(A|AAA) = \frac{3\theta + (1 - \theta)p_A}{1 + 2\theta}$$

and applying the third law of probability, we obtain:

$$\begin{aligned} \Pr(AAAA) &= \Pr(AAA) \Pr(A|AAA) \\ &= \frac{p_A[\theta + (1 - \theta)p_A][2\theta + (1 - \theta)p_A][3\theta + (1 - \theta)p_A]}{(1 + \theta)(1 + 2\theta)} \end{aligned}$$

80

## Match probability for a homozygote

$$\Pr(AA|AA) = \frac{\Pr(AAAA)}{\Pr(AA)}$$

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

81

## Match probability for a homozygote

$$\Pr(AA|AA) = \frac{\Pr(AAAA)}{\Pr(AA)}$$

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

81

# Match probability for a heterozygote

$$\Pr(AB|AB) = 2 \frac{\Pr(ABAB)}{\Pr(AB)}$$

with:

$$\begin{aligned}\Pr(AB) &= \Pr(A) \Pr(B|A) \\ &= p_A[(1 - \theta)p_B]\end{aligned}$$

$$\begin{aligned}\Pr(ABAB) &= \Pr(A) \Pr(B|A) \Pr(A|AB) \Pr(B|ABA) \\ &= \frac{p_A[(1 - \theta)p_B][\theta + (1 - \theta)p_A][\theta + (1 - \theta)p_B]}{(1 + \theta)(1 + 2\theta)}\end{aligned}$$

82

# Match probability for a heterozygote

$$\Pr(AB|AB) = 2 \frac{\Pr(ABAB)}{\Pr(AB)}$$

with:

$$\begin{aligned}\Pr(AB) &= \Pr(A) \Pr(B|A) \\ &= p_A[(1 - \theta)p_B]\end{aligned}$$

$$\begin{aligned}\Pr(ABAB) &= \Pr(A) \Pr(B|A) \Pr(A|AB) \Pr(B|ABA) \\ &= \frac{p_A[(1 - \theta)p_B][\theta + (1 - \theta)p_A][\theta + (1 - \theta)p_B]}{(1 + \theta)(1 + 2\theta)}\end{aligned}$$

82

## Match probability for a heterozygote

$$\Pr(AB|AB) = 2 \frac{\Pr(ABAB)}{\Pr(AB)}$$

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

83

## Match probability for a heterozygote

$$\Pr(AB|AB) = 2 \frac{\Pr(ABAB)}{\Pr(AB)}$$

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

83

## 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  $p_A$ ?

84

## 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  $p_A$ ?

84

## **Binomial distribution**

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

The probability of  $x$  heads in a row is

$$p \times p \times \dots \times p = p^x$$

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

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

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

85

## **Binomial distribution**

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

The probability of  $x$  heads in a row is

$$p \times p \times \dots \times p = p^x$$

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

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

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

85

## **Binomial distribution**

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

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

86

## **Binomial distribution**

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

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

86

## Binomial distribution

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

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

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

87

## Binomial distribution

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

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

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

87

## **Binomial distribution**

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

No. A's	Probability
0	
1	
2	
3	
4	

88

## **Binomial distribution**

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

No. A's	Probability
0	
1	
2	
3	
4	

88

## Binomial Likelihood

The quantity  $\Pr(x|p)$  is the *probability of the data*,  $x$  successes in  $n$  trials, when each trial has probability  $p$  of success.

The same quantity, written as  $L(p|x)$ , is the *likelihood of the parameter*,  $p$ , when the value  $x$  has been observed.

Each value of  $x$  gives a different likelihood curve, and each curve points to a  $p$  value with maximum likelihood. This leads to *maximum likelihood estimation*.

89

## Binomial Likelihood

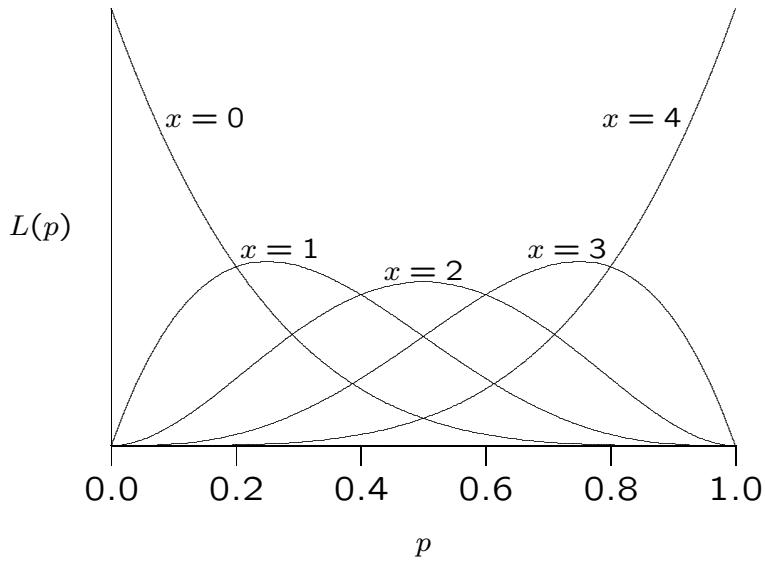
The quantity  $\Pr(x|p)$  is the *probability of the data*,  $x$  successes in  $n$  trials, when each trial has probability  $p$  of success.

The same quantity, written as  $L(p|x)$ , is the *likelihood of the parameter*,  $p$ , when the value  $x$  has been observed.

Each value of  $x$  gives a different likelihood curve, and each curve points to a  $p$  value with maximum likelihood. This leads to *maximum likelihood estimation*.

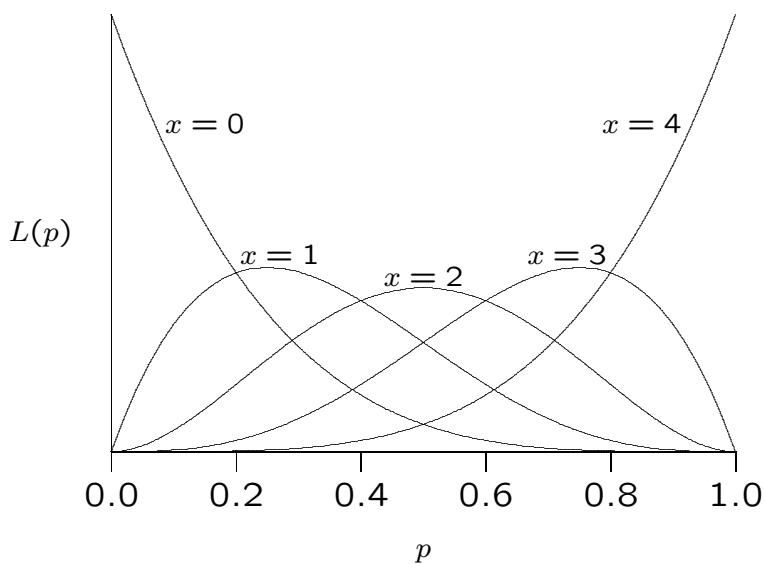
89

**Likelihood**  $L(p|x, n = 4)$



90

**Likelihood**  $L(p|x, n = 4)$



90

## Mean of Binomial

If there are  $n$  trials, each of which has probability  $p$  of giving a success, the *expected number* of successes is  $np$ .

The *sample proportion* of successes is

$$\tilde{p} = \frac{x}{n}$$

(This is the maximum likelihood estimate of  $p$ .)

The expected, or *mean*, value of  $\tilde{p}$  is  $p$ .

$$\mathcal{E}(\tilde{p}) = p$$

The expected value of  $x$  is  $np$ .

91

## Mean of Binomial

If there are  $n$  trials, each of which has probability  $p$  of giving a success, the *expected number* of successes is  $np$ .

The *sample proportion* of successes is

$$\tilde{p} = \frac{x}{n}$$

(This is the maximum likelihood estimate of  $p$ .)

The expected, or *mean*, value of  $\tilde{p}$  is  $p$ .

$$\mathcal{E}(\tilde{p}) = p$$

The expected value of  $x$  is  $np$ .

91

## Variance of Binomial

The expected value of  $(x - np)^2$  is  $np(1 - p)$ . This is the *variance* of the number of successes in  $n$  trials, and indicates the spread of the distribution.

The variance of the sample proportion  $\tilde{p}$  is

$$\text{Var}(\tilde{p}) = \frac{p(1 - p)}{n}$$

The variance of the sample count  $x$  is  $np(1 - p)$ .

92

## Variance of Binomial

The expected value of  $(x - np)^2$  is  $np(1 - p)$ . This is the *variance* of the number of successes in  $n$  trials, and indicates the spread of the distribution.

The variance of the sample proportion  $\tilde{p}$  is

$$\text{Var}(\tilde{p}) = \frac{p(1 - p)}{n}$$

The variance of the sample count  $x$  is  $np(1 - p)$ .

92

## Normal Approximation to Binomial

Provided  $np$  is not “too small”, the binomial distribution can be approximated by the normal distribution with the same mean and variance. In particular:

$$\tilde{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

If the number  $x$  of heads in 100 tosses has a binomial distribution

$$x \sim B(100, 1/2)$$

then the proportion  $\tilde{p}$  of heads is approximately normally distributed:

$$\tilde{p} \sim N\left(\frac{1}{2}, \frac{\frac{1}{2} \times \frac{1}{2}}{100} = \frac{1}{400}\right)$$

93

## Normal Approximation to Binomial

Provided  $np$  is not “too small”, the binomial distribution can be approximated by the normal distribution with the same mean and variance. In particular:

$$\tilde{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

If the number  $x$  of heads in 100 tosses has a binomial distribution

$$x \sim B(100, 1/2)$$

then the proportion  $\tilde{p}$  of heads is approximately normally distributed:

$$\tilde{p} \sim N\left(\frac{1}{2}, \frac{\frac{1}{2} \times \frac{1}{2}}{100} = \frac{1}{400}\right)$$

93

## Normal Approximation to Binomial

To use the normal distribution in practice, change to the *standard normal* variable  $z$  with a mean of 0, and a variance of 1.

Make this change by

$$z = \frac{\tilde{p} - p}{\sqrt{p(1-p)/n}}$$

For a standard normal, 95% of the values lie between  $\pm 1.96$ . The normal approximation to the binomial therefore implies that 95% of the values of  $\tilde{p}$  lie in the range

$$p \pm 1.96\sqrt{p(1-p)/n}$$

94

## Normal Approximation to Binomial

To use the normal distribution in practice, change to the *standard normal* variable  $z$  with a mean of 0, and a variance of 1.

Make this change by

$$z = \frac{\tilde{p} - p}{\sqrt{p(1-p)/n}}$$

For a standard normal, 95% of the values lie between  $\pm 1.96$ . The normal approximation to the binomial therefore implies that 95% of the values of  $\tilde{p}$  lie in the range

$$p \pm 1.96\sqrt{p(1-p)/n}$$

94

## Sampling Variation

Imagine drawing 10 beads from a jar, and noting the number  $n_b$  of black beads. This number follows the binomial distribution:

$$n_b \sim B(10, p_b)$$

The maximum likelihood estimate of  $p_b$  is

$$\hat{p}_b = \tilde{p}_b = \frac{n_b}{10}$$

but the same value of  $n_b$  could have arisen for different values of  $p_b$ .

Along with the *point estimate*  $\hat{p}_b$ , a *confidence interval* for  $p_b$  can be given.

95

## Sampling Variation

Imagine drawing 10 beads from a jar, and noting the number  $n_b$  of black beads. This number follows the binomial distribution:

$$n_b \sim B(10, p_b)$$

The maximum likelihood estimate of  $p_b$  is

$$\hat{p}_b = \tilde{p}_b = \frac{n_b}{10}$$

but the same value of  $n_b$  could have arisen for different values of  $p_b$ .

Along with the *point estimate*  $\hat{p}_b$ , a *confidence interval* for  $p_b$  can be given.

95

## Confidence Intervals

A 95% confidence interval is a variable quantity. It has endpoints which vary with the sample. It is expected that 95% of samples will lead to an interval that includes the unknown true value  $p_b$ .

The standard normal variable  $z$  has 95% of its values between  $-1.96$  and  $+1.96$ . This suggests that a 95% confidence interval for  $p_b$  is

$$\tilde{p}_b \pm 1.96 \sqrt{\frac{\tilde{p}_b(1 - \tilde{p}_b)}{n}}$$

96

## Confidence Intervals

A 95% confidence interval is a variable quantity. It has endpoints which vary with the sample. It is expected that 95% of samples will lead to an interval that includes the unknown true value  $p_b$ .

The standard normal variable  $z$  has 95% of its values between  $-1.96$  and  $+1.96$ . This suggests that a 95% confidence interval for  $p_b$  is

$$\tilde{p}_b \pm 1.96 \sqrt{\frac{\tilde{p}_b(1 - \tilde{p}_b)}{n}}$$

96

## Confidence Intervals

These normal-theory confidence intervals for  $n = 10$  are:

$\tilde{p}_c$	Confidence Interval
0.0	$0.0 \pm 1.96\sqrt{0.000}$ 0.00, 0.00
0.1	$0.1 \pm 1.96\sqrt{0.009}$ 0.00, 0.29
0.2	$0.2 \pm 1.96\sqrt{0.016}$ 0.00, 0.45
0.3	$0.3 \pm 1.96\sqrt{0.021}$ 0.02, 0.58
0.4	$0.4 \pm 1.96\sqrt{0.024}$ 0.10, 0.70
0.5	$0.5 \pm 1.96\sqrt{0.025}$ 0.19, 0.81
0.6	$0.6 \pm 1.96\sqrt{0.024}$ 0.30, 0.90
0.7	$0.7 \pm 1.96\sqrt{0.021}$ 0.42, 0.98
0.8	$0.8 \pm 1.96\sqrt{0.016}$ 0.55, 1.00
0.9	$0.9 \pm 1.96\sqrt{0.009}$ 0.71, 1.00
1.0	$1.0 \pm 1.96\sqrt{0.000}$ 1.00, 1.00

97

## Confidence Intervals

These normal-theory confidence intervals for  $n = 10$  are:

$\tilde{p}_c$	Confidence Interval
0.0	$0.0 \pm 1.96\sqrt{0.000}$ 0.00, 0.00
0.1	$0.1 \pm 1.96\sqrt{0.009}$ 0.00, 0.29
0.2	$0.2 \pm 1.96\sqrt{0.016}$ 0.00, 0.45
0.3	$0.3 \pm 1.96\sqrt{0.021}$ 0.02, 0.58
0.4	$0.4 \pm 1.96\sqrt{0.024}$ 0.10, 0.70
0.5	$0.5 \pm 1.96\sqrt{0.025}$ 0.19, 0.81
0.6	$0.6 \pm 1.96\sqrt{0.024}$ 0.30, 0.90
0.7	$0.7 \pm 1.96\sqrt{0.021}$ 0.42, 0.98
0.8	$0.8 \pm 1.96\sqrt{0.016}$ 0.55, 1.00
0.9	$0.9 \pm 1.96\sqrt{0.009}$ 0.71, 1.00
1.0	$1.0 \pm 1.96\sqrt{0.000}$ 1.00, 1.00

97

## Confidence Intervals

To be 95% sure that the estimate is no more than 0.01 from the true value,  $1.96\sqrt{p(1-p)/n}$  should be less than 0.01. The widest confidence interval is when  $p = 0.5$ , and then need

$$0.01 \geq 1.96\sqrt{0.5 \times 0.5/n}$$

which means that

$$n \geq 10,000$$

If the true value of  $p$  was about 0.05, however,

$$\begin{aligned} 0.01 &\geq 2\sqrt{0.05 \times 0.95/n} \\ n &\geq 1,900 \approx 2,000 \end{aligned}$$

98

## Confidence Intervals

To be 95% sure that the estimate is no more than 0.01 from the true value,  $1.96\sqrt{p(1-p)/n}$  should be less than 0.01. The widest confidence interval is when  $p = 0.5$ , and then need

$$0.01 \geq 1.96\sqrt{0.5 \times 0.5/n}$$

which means that

$$n \geq 10,000$$

If the true value of  $p$  was about 0.05, however,

$$\begin{aligned} 0.01 &\geq 2\sqrt{0.05 \times 0.95/n} \\ n &\geq 1,900 \approx 2,000 \end{aligned}$$

98

## Exact Confidence Intervals: One-sided

The normal-based confidence intervals are constructed to be symmetric about the sample value, unless the interval goes outside the interval from 0 to 1. They are therefore less satisfactory the closer the true value is to 0 or 1.

More accurate confidence limits follow from the binomial distribution exactly. For events with low probabilities  $p$ , how large could  $p$  be for there to be at least a 5% chance of seeing at most as many as  $x$  (i.e.  $0, 1, 2, \dots, x$ ) occurrences of that event among  $n$  events.

99

## Exact Confidence Intervals: One-sided

The normal-based confidence intervals are constructed to be symmetric about the sample value, unless the interval goes outside the interval from 0 to 1. They are therefore less satisfactory the closer the true value is to 0 or 1.

More accurate confidence limits follow from the binomial distribution exactly. For events with low probabilities  $p$ , how large could  $p$  be for there to be at least a 5% chance of seeing at most as many as  $x$  (i.e.  $0, 1, 2, \dots, x$ ) occurrences of that event among  $n$  events.

99

## Exact Confidence Intervals: One-sided

If this upper bound is  $p_U$ ,

$$\sum_{k=0}^x \Pr(k) \geq 0.05$$
$$\sum_{k=0}^x \binom{n}{k} p_U^k (1-p_U)^{n-k} \geq 0.05$$

If  $x = 0$ , then  $(1 - p_U)^n \geq 0.05$  or  $p_U \leq 1 - 0.05^{1/n}$  and this is 0.0295 if  $n = 100$ .

More generally  $p_U \approx 3/n$  for 95% upper confidence limit.

100

## Exact Confidence Intervals: One-sided

If this upper bound is  $p_U$ ,

$$\sum_{k=0}^x \Pr(k) \geq 0.05$$
$$\sum_{k=0}^x \binom{n}{k} p_U^k (1-p_U)^{n-k} \geq 0.05$$

If  $x = 0$ , then  $(1 - p_U)^n \geq 0.05$  or  $p_U \leq 1 - 0.05^{1/n}$  and this is 0.0295 if  $n = 100$ .

More generally  $p_U \approx 3/n$  for 95% upper confidence limit.

100

## Explicit Equations for Confidence Intervals

The Central Limit Theorem (the average of independent variables has a normal distribution) can be used for logarithms of profile frequencies:

$$\begin{aligned}\hat{P} &= \prod_l \tilde{P}_l \\ \ln(\hat{P}) &= \sum_l \ln(\tilde{P}_l) \sim N(\ln(P), \text{Var}[\ln(\hat{P})])\end{aligned}$$

A confidence interval for  $\ln(P)$  is, therefore,

$$\ln(\hat{P}) \pm 1.96 \sqrt{\text{Var}(\ln \hat{P})}$$

Taking anti-logs (i.e. raising  $e$  to this power) gives the confidence interval for  $P$ :

$$\hat{P} e^{\pm 1.96 \sqrt{\text{Var}(\ln \hat{P})}}$$

because  $e^{(\ln \hat{P})} = \ln(\hat{P})$ . The interval therefore has the form  $(\hat{P}/C, C\hat{P})$ , where  $C = e^{1.96 \sqrt{\text{Var}(\ln \hat{P})}}$ .

101

## Explicit Equations for Confidence Intervals

The Central Limit Theorem (the average of independent variables has a normal distribution) can be used for logarithms of profile frequencies:

$$\begin{aligned}\hat{P} &= \prod_l \tilde{P}_l \\ \ln(\hat{P}) &= \sum_l \ln(\tilde{P}_l) \sim N(\ln(P), \text{Var}[\ln(\hat{P})])\end{aligned}$$

A confidence interval for  $\ln(P)$  is, therefore,

$$\ln(\hat{P}) \pm 1.96 \sqrt{\text{Var}(\ln \hat{P})}$$

Taking anti-logs (i.e. raising  $e$  to this power) gives the confidence interval for  $P$ :

$$\hat{P} e^{\pm 1.96 \sqrt{\text{Var}(\ln \hat{P})}}$$

because  $e^{(\ln \hat{P})} = \ln(\hat{P})$ . The interval therefore has the form  $(\hat{P}/C, C\hat{P})$ , where  $C = e^{1.96 \sqrt{\text{Var}(\ln \hat{P})}}$ .

101

## Explicit Equations for Confidence Intervals

For samples of  $n$  individuals

$$\text{Var}[\ln(\tilde{p}_i^2)] \approx \frac{2(1-p_i)}{np_i}$$

$$\text{Var}[\ln(2\tilde{p}_i\tilde{p}_j)] \approx \frac{p_i + p_j - 4p_ip_j}{2np_ip_j}$$

These variances are used to construct normal-theory confidence intervals on the log-scale. Then anti-logs are taken to get back to the original scale.

102

## Explicit Equations for Confidence Intervals

For samples of  $n$  individuals

$$\text{Var}[\ln(\tilde{p}_i^2)] \approx \frac{2(1-p_i)}{np_i}$$

$$\text{Var}[\ln(2\tilde{p}_i\tilde{p}_j)] \approx \frac{p_i + p_j - 4p_ip_j}{2np_ip_j}$$

These variances are used to construct normal-theory confidence intervals on the log-scale. Then anti-logs are taken to get back to the original scale.

102

## Explicit Equations for Confidence Intervals

As an example, consider this Polymarker profile:

Locus	Genotype	$\tilde{P}$	$\ln(\tilde{P})$	$\text{Var}(\ln \tilde{P})$
LDLR	AB	0.4921	-0.7092	0.0003
GYPA	BB	0.2125	-1.5487	0.0227
HBGG	BC	0.0044	-5.4330	0.9626
D7S8	AB	0.4961	-0.7009	0.0002
Gc	BC	0.2185	-1.5210	0.0138
		Product	Sum	Sum
		0.000,050	-9.9128	0.9996

The 95% CI on the log product is  $-9.9128 \pm 1.96\sqrt{0.9996}$  or  $(-11.8724, -7.9532)$ . Taking antilogs, this interval becomes  $(0.000,007 \text{ to } 0.000,352)$  – a factor 7 in each direction from the estimate.

103

## Explicit Equations for Confidence Intervals

As an example, consider this Polymarker profile:

Locus	Genotype	$\tilde{P}$	$\ln(\tilde{P})$	$\text{Var}(\ln \tilde{P})$
LDLR	AB	0.4921	-0.7092	0.0003
GYPA	BB	0.2125	-1.5487	0.0227
HBGG	BC	0.0044	-5.4330	0.9626
D7S8	AB	0.4961	-0.7009	0.0002
Gc	BC	0.2185	-1.5210	0.0138
		Product	Sum	Sum
		0.000,050	-9.9128	0.9996

The 95% CI on the log product is  $-9.9128 \pm 1.96\sqrt{0.9996}$  or  $(-11.8724, -7.9532)$ . Taking antilogs, this interval becomes  $(0.000,007 \text{ to } 0.000,352)$  – a factor 7 in each direction from the estimate.

103

## Multinomial Distribution

Toss two coins  $n$  times. For each double toss, the probabilities of the three outcomes are:

$$\begin{array}{ll} \text{2 heads} & p_{HH} = 1/4 \\ \text{1 head, 1 tail} & p_{HT} = 1/2 \\ \text{2 tails} & p_{TT} = 1/4 \end{array}$$

The probability of  $x$  lots of 2 heads is  $(p_{HH})^x$ , etc.

The numbers of ways of ordering  $x, y, z$  occurrences of the three outcomes is  $n!/[x!y!z!]$  where  $n = x + y + z$ .

The multinomial probability for  $x$  of  $HH$ , and  $y$  of  $HT$  and  $z$  of  $TT$  in  $n$  trials is:

$$\Pr(x, y, z) = \frac{n!}{x!y!z!}(p_{HH})^x(p_{HT})^y(p_{TT})^z$$

104

## Multinomial Distribution

Toss two coins  $n$  times. For each double toss, the probabilities of the three outcomes are:

$$\begin{array}{ll} \text{2 heads} & p_{HH} = 1/4 \\ \text{1 head, 1 tail} & p_{HT} = 1/2 \\ \text{2 tails} & p_{TT} = 1/4 \end{array}$$

The probability of  $x$  lots of 2 heads is  $(p_{HH})^x$ , etc.

The numbers of ways of ordering  $x, y, z$  occurrences of the three outcomes is  $n!/[x!y!z!]$  where  $n = x + y + z$ .

The multinomial probability for  $x$  of  $HH$ , and  $y$  of  $HT$  and  $z$  of  $TT$  in  $n$  trials is:

$$\Pr(x, y, z) = \frac{n!}{x!y!z!}(p_{HH})^x(p_{HT})^y(p_{TT})^z$$

104

## **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.

105

## **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.

105

## 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

106

## 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

106

## 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  $A_i$ :

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

107

## 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  $A_i$ :

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

107

## D3S1358 Hardy-Weinberg Calculations

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

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

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

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

Other values are shown on the next slide.

108

## D3S1358 Hardy-Weinberg Calculations

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

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

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

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

Other values are shown on the next slide.

108

## 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
	Exp.	0.0	0.0	0.0	0.1	0.3	0.3	0.2	0.1	0.0

109

## 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
	Exp.	0.0	0.0	0.0	0.1	0.3	0.3	0.2	0.1	0.0

109

## **Testing for Hardy-Weinberg Equilibrium**

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

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

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

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

110

## **Testing for Hardy-Weinberg Equilibrium**

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

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

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

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

110

## **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.

111

## **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.

111

## 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	$0.26 > (0.5)^2$

112

## 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	$0.26 > (0.5)^2$

112

## 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.

113

## 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.

113

## Population Admixture

A population might represent the recent admixture of two parental populations. People with one or two parents in population 1 may be considered as belonging to population 1. This causes excess heterozygosity in that 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

114

## Population Admixture

A population might represent the recent admixture of two parental populations. People with one or two parents in population 1 may be considered as belonging to population 1. This causes excess heterozygosity in that 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

114

## Exact HWE Test

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

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

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

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

115

## Exact HWE Test

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

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

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

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

115

## Exact HWE Test

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

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

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

116

## Exact HWE Test

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

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

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

116

## 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:

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

117

## 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:

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

117

## 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.

118

## 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.

118

## Exact HWE Test

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	<b>0.0051</b>	<b>0.0053*</b>
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  $n_{Aa} \leq 8$ .

119

## Exact HWE Test

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	<b>0.0051</b>	<b>0.0053*</b>
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  $n_{Aa} \leq 8$ .

119

## **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.

120

## **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.

120

## 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.

121

## 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.

121

## 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	

122

## 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	

122

## Permutation Test

Find the 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	
2	2	1	
1	4	0	

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

123

## Permutation Test

Find the 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	
2	2	1	
1	4	0	

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

123

## 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  $0.6163 \times 10^{-13}$ , which is a very small number. But it is not unusually small if HWE holds: a proportion 0.81 of 1000 permutations have an even small probability.

124

## 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  $0.6163 \times 10^{-13}$ , which is a very small number. But it is not unusually small if HWE holds: a proportion 0.81 of 1000 permutations have an even small probability.

124

## Multiple Testing

When multiple tests are performed, each at significance level  $\alpha$ , a proportion of the tests are expected to cause rejection even if all the hypotheses are true.

Bonferroni correction makes the overall (experimentwise) significance level equal to  $\alpha$  by adjusting the level for each individual test to  $\alpha'$ . If  $\alpha$  is the probability that at least one of the  $L$  tests causes rejection, it is also 1 minus the probability that none of the tests causes rejection:

$$\begin{aligned}\alpha &= 1 - (1 - \alpha')^L \\ &\approx L\alpha'\end{aligned}$$

provided the  $L$  tests are independent.

If  $L = 15$ , need  $\alpha' = 0.0033$  in order for  $\alpha = 0.05$ .

125

## Multiple Testing

When multiple tests are performed, each at significance level  $\alpha$ , a proportion of the tests are expected to cause rejection even if all the hypotheses are true.

Bonferroni correction makes the overall (experimentwise) significance level equal to  $\alpha$  by adjusting the level for each individual test to  $\alpha'$ . If  $\alpha$  is the probability that at least one of the  $L$  tests causes rejection, it is also 1 minus the probability that none of the tests causes rejection:

$$\begin{aligned}\alpha &= 1 - (1 - \alpha')^L \\ &\approx L\alpha'\end{aligned}$$

provided the  $L$  tests are independent.

If  $L = 15$ , need  $\alpha' = 0.0033$  in order for  $\alpha = 0.05$ .

125

## QQ-Plots

An alternative approach to considering multiple-testing issues is to use QQ-plots. If all the hypotheses being tested are true then the resulting  $p$ -values are uniformly distributed between 0 and 1.

For a set of  $n$  tests, we would expect to see  $p$  values at  $1/(n + 1), 2/n, \dots, n/(n + 1), 1$ . We plot the observed  $p$ -values against these expected values: the smallest against  $1/(n + 1)$  and the largest against 1. It is more convenient to transform to  $-\log_{10}(p)$  to accentuate the extremely small  $p$  values. The point at which the observed values start departing from the expected values is an indication of “significant” values in a way that takes into account the number of tests.

126

## QQ-Plots

An alternative approach to considering multiple-testing issues is to use QQ-plots. If all the hypotheses being tested are true then the resulting  $p$ -values are uniformly distributed between 0 and 1.

For a set of  $n$  tests, we would expect to see  $p$  values at  $1/(n + 1), 2/n, \dots, n/(n + 1), 1$ . We plot the observed  $p$ -values against these expected values: the smallest against  $1/(n + 1)$  and the largest against 1. It is more convenient to transform to  $-\log_{10}(p)$  to accentuate the extremely small  $p$  values. The point at which the observed values start departing from the expected values is an indication of “significant” values in a way that takes into account the number of tests.

126

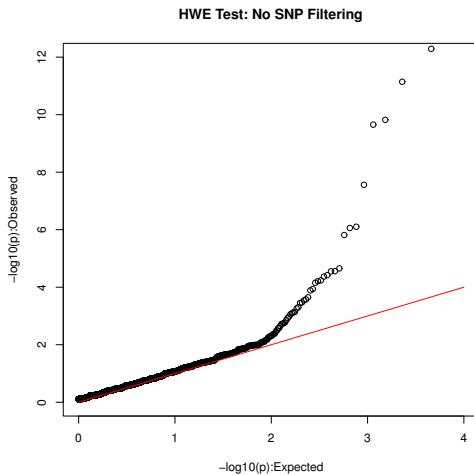
## QQ-Plots



The results for 9208 SNPs on human chromosome 1. Bonferroni would suggest rejecting HWE when  $p \leq 0.05/9205 = 5.4 \times 10^{-6}$  or  $-\log_{10}(p) \geq 5.3$ .

127

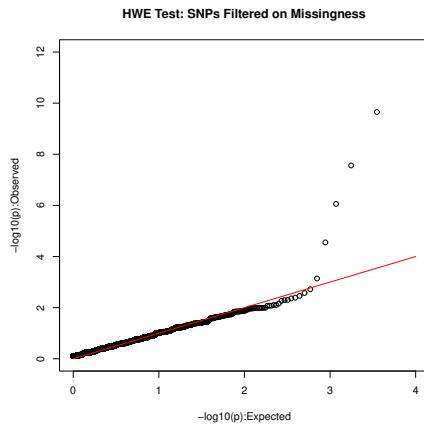
## QQ-Plots



The results for 9208 SNPs on human chromosome 1. Bonferroni would suggest rejecting HWE when  $p \leq 0.05/9205 = 5.4 \times 10^{-6}$  or  $-\log_{10}(p) \geq 5.3$ .

127

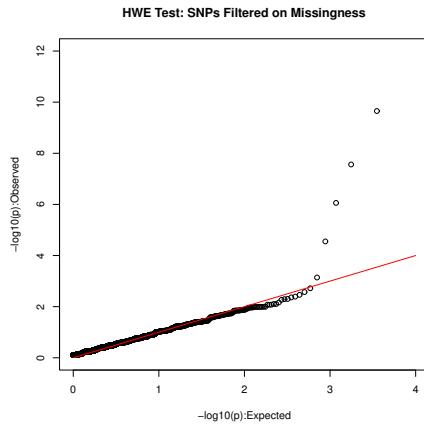
## QQ-Plots



The same set of results as on the previous slide except now that any SNP with any missing data was excluded. Now 7446 SNPs and Bonferroni would reject if  $-\log_{10}(p) \geq 5.2$ . All five outliers had zero counts for the minor allele homozygote and at least 32 heterozygotes in a sample of size 50.

128

## QQ-Plots



The same set of results as on the previous slide except now that any SNP with any missing data was excluded. Now 7446 SNPs and Bonferroni would reject if  $-\log_{10}(p) \geq 5.2$ . All five outliers had zero counts for the minor allele homozygote and at least 32 heterozygotes in a sample of size 50.

128

## **Linkage Disequilibrium**

This term reserved for association between pairs of alleles – one at each of two loci.

When gametic data are available, could refer to gametic disequilibrium.

When genotypic data are available, but gametes can be inferred, can make inferences about gametic and non-gametic pairs of alleles.

When genotypic data are available, but gametes cannot be inferred, can work with composite measures of disequilibrium.

129

## **Linkage Disequilibrium**

This term reserved for association between pairs of alleles – one at each of two loci.

When gametic data are available, could refer to gametic disequilibrium.

When genotypic data are available, but gametes can be inferred, can make inferences about gametic and non-gametic pairs of alleles.

When genotypic data are available, but gametes cannot be inferred, can work with composite measures of disequilibrium.

129

## **Matching Within and Between Populations**

130

## **Matching Within and Between Populations**

130

## **Within-population Matching**

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

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

131

## **Within-population Matching**

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

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

131

## NIST Data for DYS390

Allele	Population					Total
	Afr.Am.	Cauc.	Hisp.	Asian		
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	

132

## NIST Data for DYS390

Allele	Population					Total
	Afr.Am.	Cauc.	Hisp.	Asian		
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	

132

## **Within- and Between-population Matching for DYS390**

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

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

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

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

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

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

133

## **Within- and Between-population Matching for DYS390**

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

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

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

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

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

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

133

## Within- and Between-population Matching for DYS391

Allele	Population					Total
	Afr.Am.	Cauc.	Hisp.	Asian		
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  $54,918/122,419=0.449$ .

134

## Within- and Between-population Matching for DYS391

Allele	Population					Total
	Afr.Am.	Cauc.	Hisp.	Asian		
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  $54,918/122,419=0.449$ .

134

## DYS390, DYS391 African-AmericanData

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	

135

## DYS390, DYS391 African-AmericanData

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	

135

## DYS390, DYS391 Caucasian Data

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	

136

## DYS390, DYS391 Caucasian Data

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	

136

## **Two-locus Matches**

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

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

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

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

137

## **Two-locus Matches**

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

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

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

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

137

## 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.

138

## 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.

138

## 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

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

139

## 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

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

139

## Interpreting Evidence

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

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

A population genetic model:

$$\Pr(AA) = \theta p_A + (1 - \theta)p_A^2$$

$$\Pr(A|A) = \theta + (1 - \theta)p_A$$

where  $\theta$  is the probability that two profiles are identical by descent.

140

## Interpreting Evidence

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

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

A population genetic model:

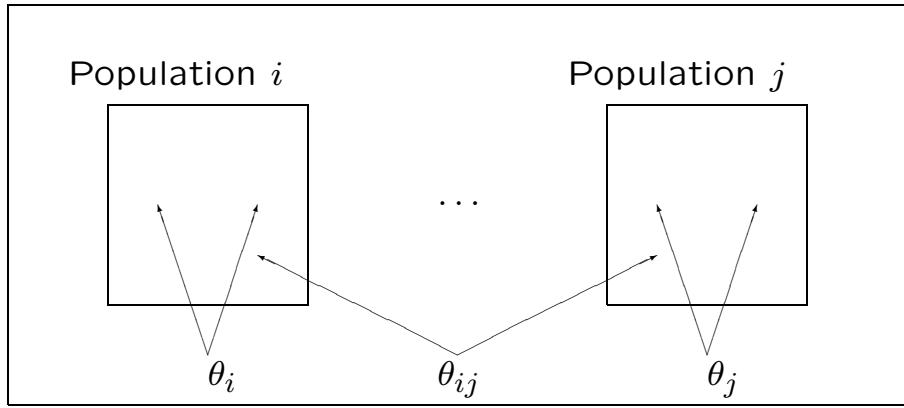
$$\Pr(AA) = \theta p_A + (1 - \theta)p_A^2$$

$$\Pr(A|A) = \theta + (1 - \theta)p_A$$

where  $\theta$  is the probability that two profiles are identical by descent.

140

## Coancestries

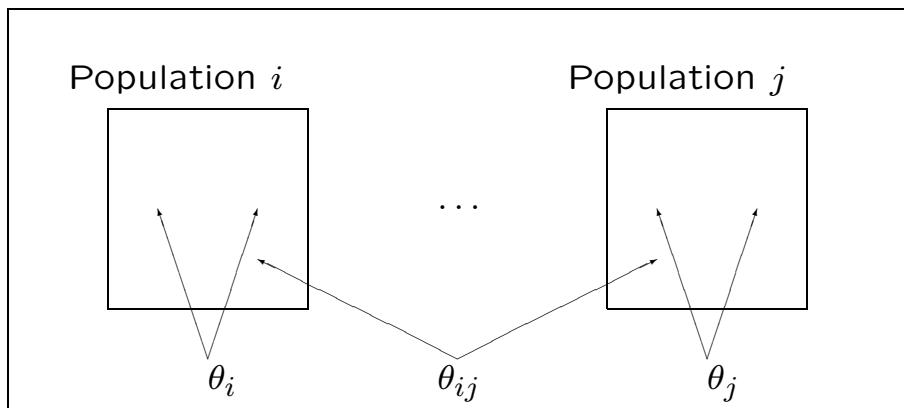


Average within-population coancestry:  $\theta_W = \frac{1}{r} \sum_{i=1}^r \theta_i$ .

Average between-population coancestry:  $\theta_B = \frac{1}{r(r-1)} \sum_{i \neq j} \theta_{ij}$ .

141

## Coancestries

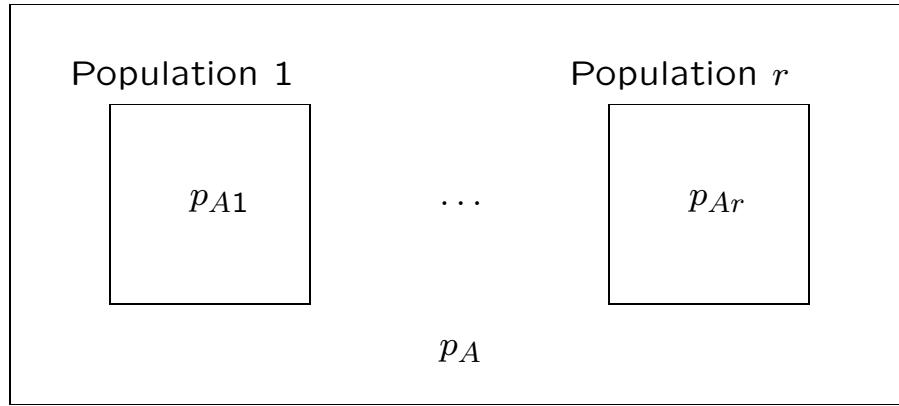


Average within-population coancestry:  $\theta_W = \frac{1}{r} \sum_{i=1}^r \theta_i$ .

Average between-population coancestry:  $\theta_B = \frac{1}{r(r-1)} \sum_{i \neq j} \theta_{ij}$ .

141

## Haplotype Frequencies

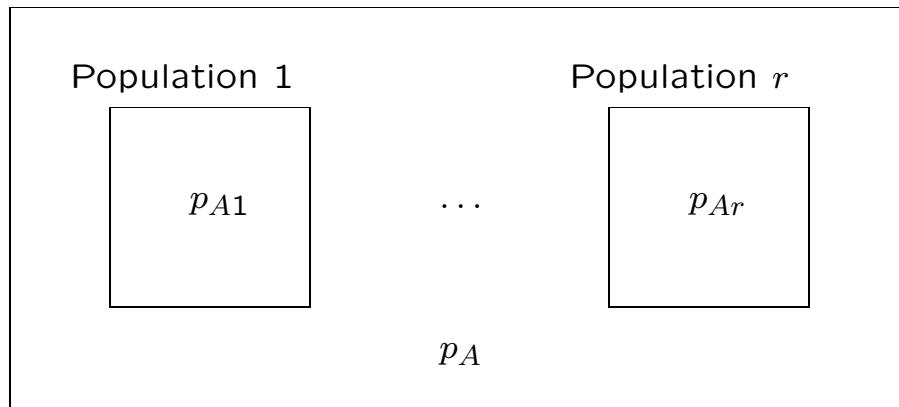


Within population  $i$ :  $P_{AA} = \theta_i p_A + (1 - \theta_i)p_A^2$ .

Between populations  $i, j$ :  $P_{AA} = \theta_{ij} p_A + (1 - \theta_{ij})p_A^2$ .

142

## Haplotype Frequencies



Within population  $i$ :  $P_{AA} = \theta_i p_A + (1 - \theta_i)p_A^2$ .

Between populations  $i, j$ :  $P_{AA} = \theta_{ij} p_A + (1 - \theta_{ij})p_A^2$ .

142

## Sample Within-population Matching

If the sample from population  $i$  has  $n_{Ai}$  copies of allele (or haplotype)  $A$ , and these sum to  $n_i$ , the sample within-population matching proportion for this population is

$$\begin{aligned}\tilde{M}_{Wi} &= \frac{1}{n_i(n_i - 1)} \sum_A n_{Ai}(n_{Ai} - 1) \\ &= \frac{n_i}{n_i - 1} \sum_A \frac{n_{Ai}}{n_i} \left( \frac{n_{Ai}}{n_i} - \frac{1}{n_i} \right) \\ &= \frac{n_i}{n_i - 1} \left( \sum_A \tilde{p}_{Ai}^2 - \frac{1}{n_i} \right)\end{aligned}$$

Averaging over populations:

$$\tilde{M}_W = \frac{1}{r} \sum_{i=1}^r \tilde{M}_{Wi}$$

143

## Sample Within-population Matching

If the sample from population  $i$  has  $n_{Ai}$  copies of allele (or haplotype)  $A$ , and these sum to  $n_i$ , the sample within-population matching proportion for this population is

$$\begin{aligned}\tilde{M}_{Wi} &= \frac{1}{n_i(n_i - 1)} \sum_A n_{Ai}(n_{Ai} - 1) \\ &= \frac{n_i}{n_i - 1} \sum_A \frac{n_{Ai}}{n_i} \left( \frac{n_{Ai}}{n_i} - \frac{1}{n_i} \right) \\ &= \frac{n_i}{n_i - 1} \left( \sum_A \tilde{p}_{Ai}^2 - \frac{1}{n_i} \right)\end{aligned}$$

Averaging over populations:

$$\tilde{M}_W = \frac{1}{r} \sum_{i=1}^r \tilde{M}_{Wi}$$

143

## Sample Between-population Matching

The sample between-population matching proportion for populations  $i$  and  $j$  is

$$\begin{aligned}\tilde{M}_{Bij} &= \frac{1}{n_i n_j} \sum_A n_{Ai} n_{Aj} \\ &= \sum_A \frac{n_{Ai}}{n_i} \frac{n_{Aj}}{n_j} \\ &= \sum_A \tilde{p}_{Ai} \tilde{p}_{Aj}\end{aligned}$$

Averaging over pairs of populations:

$$\tilde{M}_B = \frac{1}{r(r-1)} \sum_{i \neq j}^r \tilde{M}_{Bij}$$

144

## Sample Between-population Matching

The sample between-population matching proportion for populations  $i$  and  $j$  is

$$\begin{aligned}\tilde{M}_{Bij} &= \frac{1}{n_i n_j} \sum_A n_{Ai} n_{Aj} \\ &= \sum_A \frac{n_{Ai}}{n_i} \frac{n_{Aj}}{n_j} \\ &= \sum_A \tilde{p}_{Ai} \tilde{p}_{Aj}\end{aligned}$$

Averaging over pairs of populations:

$$\tilde{M}_B = \frac{1}{r(r-1)} \sum_{i \neq j}^r \tilde{M}_{Bij}$$

144

## Expected Values of Sample Matching Proportions

From the way in which  $\theta_i$  and  $\theta_{ij}$  were defined, the expected values of the sample matching proportions are

$$\mathcal{E}(1 - \tilde{M}_{Wi}) = H(1 - \theta_i)$$

$$\mathcal{E}(1 - \tilde{M}_{Bij}) = H(1 - \theta_{ij})$$

The quantities  $(1 - \tilde{M})$  are often called sample heterozygosities, and  $H = 1 - \sum_A p_A^2$ .

Taking averages

$$\mathcal{E}(1 - \tilde{M}_W) = H(1 - \theta_W)$$

$$\mathcal{E}(1 - \tilde{M}_B) = H(1 - \theta_B)$$

145

## Expected Values of Sample Matching Proportions

From the way in which  $\theta_i$  and  $\theta_{ij}$  were defined, the expected values of the sample matching proportions are

$$\mathcal{E}(1 - \tilde{M}_{Wi}) = H(1 - \theta_i)$$

$$\mathcal{E}(1 - \tilde{M}_{Bij}) = H(1 - \theta_{ij})$$

The quantities  $(1 - \tilde{M})$  are often called sample heterozygosities, and  $H = 1 - \sum_A p_A^2$ .

Taking averages

$$\mathcal{E}(1 - \tilde{M}_W) = H(1 - \theta_W)$$

$$\mathcal{E}(1 - \tilde{M}_B) = H(1 - \theta_B)$$

145

## Estimates

We take ratios of sample matching proportions:

$$\begin{aligned}\hat{\beta}_i &= \frac{\tilde{M}_{Wi} - \tilde{M}_B}{1 - \tilde{M}_B} \\ \hat{\beta}_W &= \frac{\tilde{M}_W - \tilde{M}_B}{1 - \tilde{M}_B} \\ \hat{\beta}_{ij} &= \frac{\tilde{M}_{Bij} - \tilde{M}_B}{1 - \tilde{M}_B}\end{aligned}$$

and these have expected values of

$$\begin{aligned}\mathcal{E}(\hat{\beta}_i) &= \frac{\theta_i - \theta_B}{1 - \theta_B} \\ \mathcal{E}(\hat{\beta}_W) &= \frac{\theta_W - \theta_B}{1 - \theta_B} \\ \mathcal{E}(\hat{\beta}_{ij}) &= \frac{\theta_{ij} - \theta_B}{1 - \theta_B}\end{aligned}$$

146

## Estimates

We take ratios of sample matching proportions:

$$\begin{aligned}\hat{\beta}_i &= \frac{\tilde{M}_{Wi} - \tilde{M}_B}{1 - \tilde{M}_B} \\ \hat{\beta}_W &= \frac{\tilde{M}_W - \tilde{M}_B}{1 - \tilde{M}_B} \\ \hat{\beta}_{ij} &= \frac{\tilde{M}_{Bij} - \tilde{M}_B}{1 - \tilde{M}_B}\end{aligned}$$

and these have expected values of

$$\begin{aligned}\mathcal{E}(\hat{\beta}_i) &= \frac{\theta_i - \theta_B}{1 - \theta_B} \\ \mathcal{E}(\hat{\beta}_W) &= \frac{\theta_W - \theta_B}{1 - \theta_B} \\ \mathcal{E}(\hat{\beta}_{ij}) &= \frac{\theta_{ij} - \theta_B}{1 - \theta_B}\end{aligned}$$

146

## 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

147

## 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

147

## 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

148

## 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

148

## Y-STR Match Probabilities

Within subpopulation  $i$ , if we knew the haplotype frequencies and if we assumed random mating, the chance that two unrelated men have haplotype  $A$  is  $p_{Ai}^2$  and the match probability is just the profile probability  $p_{Ai}$ .

If we allow for the evolutionary variation that led to the current subpopulation, then the total population allele frequencies  $p_A$  can be used when the specific population frequencies  $p_{Ai}$  are not known:

$$P_{AAi} = \theta_i p_A + (1 - \theta_i) p_A^2$$

149

## Y-STR Match Probabilities

Within subpopulation  $i$ , if we knew the haplotype frequencies and if we assumed random mating, the chance that two unrelated men have haplotype  $A$  is  $p_{Ai}^2$  and the match probability is just the profile probability  $p_{Ai}$ .

If we allow for the evolutionary variation that led to the current subpopulation, then the total population allele frequencies  $p_A$  can be used when the specific population frequencies  $p_{Ai}$  are not known:

$$P_{AAi} = \theta_i p_A + (1 - \theta_i) p_A^2$$

149

## Average Match Probabilities

Within subpopulation  $i$  the match probability for haplotype  $A$  is

$$P_{A|Ai} = \theta_i + (1 - \theta_i)p_A$$

and, if we don't know the subpopulations, we may use the average over subpopulations:

$$P_{A|A} = \theta_W + (1 - \theta_W)p_A$$

The probability that two men within the same subpopulation match, without regard to the particular matching type is

$$M_W = \theta_W + (1 - \theta_W) \sum_A p_A^2$$

150

## Average Match Probabilities

Within subpopulation  $i$  the match probability for haplotype  $A$  is

$$P_{A|Ai} = \theta_i + (1 - \theta_i)p_A$$

and, if we don't know the subpopulations, we may use the average over subpopulations:

$$P_{A|A} = \theta_W + (1 - \theta_W)p_A$$

The probability that two men within the same subpopulation match, without regard to the particular matching type is

$$M_W = \theta_W + (1 - \theta_W) \sum_A p_A^2$$

150

## Estimating Match Proportions

The predicted match proportion within subpopulations is

$$M_W = \theta_W + (1 - \theta_W) \sum_A p_A^2$$

For estimates  $\tilde{p}_A$  of frequencies from the whole population:

$$\mathcal{E}(\sum_A \tilde{p}_A^2) = \sum_A p_A^2 + (1 - \sum_A p_A^2) \left( \theta_B + \frac{\theta_W - \theta_B}{r} \right)$$

If  $\tilde{p}_A$  is used for  $p_A$ :

$$\begin{aligned} \mathcal{E}[\theta_W + (1 - \theta_W) \sum_A \tilde{p}_A^2] &\approx \theta_W + (1 - \theta_W) [\sum_A p_A^2 + (1 - \sum_A p_A^2) \theta_B] \\ &= [\theta_W + (1 - \theta_W) \theta_B] + (1 - \theta_W)(1 - \theta_B) \sum_A p_A^2 \end{aligned}$$

151

## Estimating Match Proportions

The predicted match proportion within subpopulations is

$$M_W = \theta_W + (1 - \theta_W) \sum_A p_A^2$$

For estimates  $\tilde{p}_A$  of frequencies from the whole population:

$$\mathcal{E}(\sum_A \tilde{p}_A^2) = \sum_A p_A^2 + (1 - \sum_A p_A^2) \left( \theta_B + \frac{\theta_W - \theta_B}{r} \right)$$

If  $\tilde{p}_A$  is used for  $p_A$ :

$$\begin{aligned} \mathcal{E}[\theta_W + (1 - \theta_W) \sum_A \tilde{p}_A^2] &\approx \theta_W + (1 - \theta_W) [\sum_A p_A^2 + (1 - \sum_A p_A^2) \theta_B] \\ &= [\theta_W + (1 - \theta_W) \theta_B] + (1 - \theta_W)(1 - \theta_B) \sum_A p_A^2 \end{aligned}$$

151

## Estimating Match Proportions

To take account of what  $\sum_A \tilde{p}_A^2$  is actually estimating, we form an estimate of the within subpopulation matching as

$$\hat{M}_W = \beta_W + (1 - \beta_W) \sum_A \tilde{p}_A^2$$

where

$$\beta_W = \frac{\theta_W - \theta_B}{1 - \theta_B}$$

This is the “theta” for the “theta-correction” expression for match probabilities.

When there are data from the subpopulations, it has a simple estimate:

$$\hat{\beta}_W = \frac{\tilde{M}_W - \tilde{M}_B}{1 - \tilde{M}_B}$$

152

## Estimating Match Proportions

To take account of what  $\sum_A \tilde{p}_A^2$  is actually estimating, we form an estimate of the within subpopulation matching as

$$\hat{M}_W = \beta_W + (1 - \beta_W) \sum_A \tilde{p}_A^2$$

where

$$\beta_W = \frac{\theta_W - \theta_B}{1 - \theta_B}$$

This is the “theta” for the “theta-correction” expression for match probabilities.

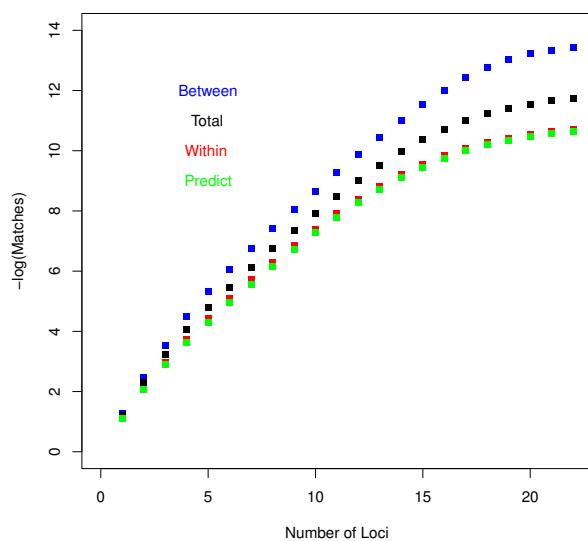
When there are data from the subpopulations, it has a simple estimate:

$$\hat{\beta}_W = \frac{\tilde{M}_W - \tilde{M}_B}{1 - \tilde{M}_B}$$

152

## US-YSTR Predictions

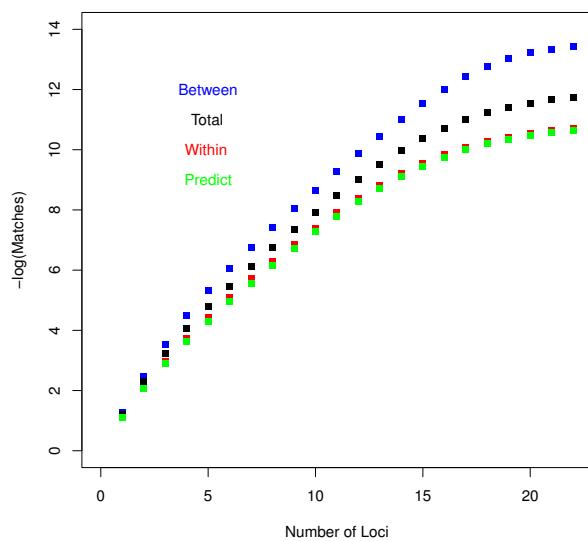
PP23



153

## US-YSTR Predictions

PP23



153

## **Matching and Partial Matching**

154

## **Matching and Partial Matching**

154

## Y-matching Review

If  $A$  is a Y-STR profile, then the match probability is

$$\Pr(A|A) = \theta + (1 - \theta)p_A$$

To allow for population structure, a value is assigned to  $\theta$  that reflects the variation in haplotype frequencies among sampled populations.

This result is an example of the “sampling formula” for seeing type  $A$  given that it has been seen  $n_A$  times previously in a sample of size  $n$ :

$$\Pr(A|n_A \text{ of } n) = \frac{n_A\theta + (1 - \theta)p_A}{1 + (n - 1)\theta}$$

for the situation of  $n_A = n = 1$ .

155

## Y-matching Review

If  $A$  is a Y-STR profile, then the match probability is

$$\Pr(A|A) = \theta + (1 - \theta)p_A$$

To allow for population structure, a value is assigned to  $\theta$  that reflects the variation in haplotype frequencies among sampled populations.

This result is an example of the “sampling formula” for seeing type  $A$  given that it has been seen  $n_A$  times previously in a sample of size  $n$ :

$$\Pr(A|n_A \text{ of } n) = \frac{n_A\theta + (1 - \theta)p_A}{1 + (n - 1)\theta}$$

for the situation of  $n_A = n = 1$ .

155

## Autosomal Profiles

For an autosomal marker, an individual has two alleles and matching of profiles means that two alleles match: for example  $AA, AA$  or  $AB, AB$ . The match probabilities are found from the sampling formula on the previous slide.

The probability that an allele is of type  $A$  is  $p_A$ .

The probabilities that a second allele is either  $A$  or  $B$  given that the first is  $A$  are

$$\begin{aligned}\Pr(A|A) &= \theta + (1 - \theta)p_A, \quad n = 1, n_A = 1 \\ \Pr(B|A) &= (1 - \theta)p_B, \quad n = 1, n_B = 0\end{aligned}$$

156

## Autosomal Profiles

For an autosomal marker, an individual has two alleles and matching of profiles means that two alleles match: for example  $AA, AA$  or  $AB, AB$ . The match probabilities are found from the sampling formula on the previous slide.

The probability that an allele is of type  $A$  is  $p_A$ .

The probabilities that a second allele is either  $A$  or  $B$  given that the first is  $A$  are

$$\begin{aligned}\Pr(A|A) &= \theta + (1 - \theta)p_A, \quad n = 1, n_A = 1 \\ \Pr(B|A) &= (1 - \theta)p_B, \quad n = 1, n_B = 0\end{aligned}$$

156

## Autosomal Profiles

The probabilities that a third allele is  $A$  after the first two alleles are  $AA$  or  $AB$  are

$$\begin{aligned}\Pr(A|AA) &= [2\theta + (1 - \theta)p_A]/(1 + \theta), \quad n = 2, n_A = 2 \\ \Pr(A|AB) &= [\theta + (1 - \theta)p_A]/(1 + \theta), \quad n = 2, n_A = 1\end{aligned}$$

The probabilities that a fourth allele is  $A$  or  $B$  after the first three alleles are  $AAA$  or  $ABA$  are

$$\begin{aligned}\Pr(A|AAA) &= [3\theta + (1 - \theta)p_A]/(1 + 2\theta), \quad n = 3, n_A = 3 \\ \Pr(B|ABA) &= [\theta + (1 - \theta)p_B]/(1 + 2\theta), \quad n = 3, n_B = 1\end{aligned}$$

157

## Autosomal Profiles

The probabilities that a third allele is  $A$  after the first two alleles are  $AA$  or  $AB$  are

$$\begin{aligned}\Pr(A|AA) &= [2\theta + (1 - \theta)p_A]/(1 + \theta), \quad n = 2, n_A = 2 \\ \Pr(A|AB) &= [\theta + (1 - \theta)p_A]/(1 + \theta), \quad n = 2, n_A = 1\end{aligned}$$

The probabilities that a fourth allele is  $A$  or  $B$  after the first three alleles are  $AAA$  or  $ABA$  are

$$\begin{aligned}\Pr(A|AAA) &= [3\theta + (1 - \theta)p_A]/(1 + 2\theta), \quad n = 3, n_A = 3 \\ \Pr(B|ABA) &= [\theta + (1 - \theta)p_B]/(1 + 2\theta), \quad n = 3, n_B = 1\end{aligned}$$

157

## Autosomal Profiles

From the third law of probability, we can get the required match probabilities

$$\begin{aligned}\Pr(AA|AA) &= \Pr(AAAA)/\Pr(AA) \\&= [\Pr(A|AAA)\Pr(AAA)]/\Pr(AA) \\&= [\Pr(A|AAA)\Pr(A|AA)\Pr(AA)]/\Pr(AA) \\&= \frac{[3\theta + (1 - \theta)p_A][2\theta + (1 - \theta)p_A]}{(1 + \theta)(1 + 2\theta)} \\ \\ \Pr(AB|AB) &= \Pr(ABAB)/\Pr(AB) \\&= [\Pr(B|ABA)\Pr(ABA) + \Pr(A|BAB)\Pr(BAB)]/\Pr(AB) \\&= [\Pr(B|ABA)\Pr(A|AB) + \Pr(A|BAB)\Pr(B|AB)]\Pr(AB) \\&= \frac{2[\theta + (1 - \theta)p_A][\theta + (1 - \theta)p_B]}{(1 + \theta)(1 + 2\theta)}\end{aligned}$$

158

## Autosomal Profiles

From the third law of probability, we can get the required match probabilities

$$\begin{aligned}\Pr(AA|AA) &= \Pr(AAAA)/\Pr(AA) \\&= [\Pr(A|AAA)\Pr(AAA)]/\Pr(AA) \\&= [\Pr(A|AAA)\Pr(A|AA)\Pr(AA)]/\Pr(AA) \\&= \frac{[3\theta + (1 - \theta)p_A][2\theta + (1 - \theta)p_A]}{(1 + \theta)(1 + 2\theta)} \\ \\ \Pr(AB|AB) &= \Pr(ABAB)/\Pr(AB) \\&= [\Pr(B|ABA)\Pr(ABA) + \Pr(A|BAB)\Pr(BAB)]/\Pr(AB) \\&= [\Pr(B|ABA)\Pr(A|AB) + \Pr(A|BAB)\Pr(B|AB)]\Pr(AB) \\&= \frac{2[\theta + (1 - \theta)p_A][\theta + (1 - \theta)p_B]}{(1 + \theta)(1 + 2\theta)}\end{aligned}$$

158

## **Autosomal Matching**

These match probabilities are for a single autosomal marker when the profile is homozygous  $AA$  or heterozygous  $AB$ . The value of  $\theta$  is assigned, and the value is based on estimates from a set of sampled populations.

$\theta$  can still be regarded as the probability of identity by descent of two alleles, but it may be better to use the concept of heterozygosity rather than matching when considering alleles (rather than genotypes) for autosomal markers.

159

## **Autosomal Matching**

These match probabilities are for a single autosomal marker when the profile is homozygous  $AA$  or heterozygous  $AB$ . The value of  $\theta$  is assigned, and the value is based on estimates from a set of sampled populations.

$\theta$  can still be regarded as the probability of identity by descent of two alleles, but it may be better to use the concept of heterozygosity rather than matching when considering alleles (rather than genotypes) for autosomal markers.

159

## Within-Population Heterozygosity

For a sample of  $n_i$  alleles from population  $i$ , the within-population heterozygosity  $\tilde{H}_i$  is

$$\tilde{H}_i = \frac{n_i}{1 - n_i} \sum_A \tilde{p}_{Ai}(1 - \tilde{p}_{Ai}) = \frac{n_i}{n_i - 1} (1 - \sum_A \tilde{p}_{Ai}^2)$$

Averaging over populations

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

160

## Within-Population Heterozygosity

For a sample of  $n_i$  alleles from population  $i$ , the within-population heterozygosity  $\tilde{H}_i$  is

$$\tilde{H}_i = \frac{n_i}{1 - n_i} \sum_A \tilde{p}_{Ai}(1 - \tilde{p}_{Ai}) = \frac{n_i}{n_i - 1} (1 - \sum_A \tilde{p}_{Ai}^2)$$

Averaging over populations

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

160

## Heterozygosity and Matching

Recall from the previous section that the allelic matching proportion within the sample from population  $i$  is

$$\tilde{M}_{Wi} = \frac{n_i}{n_i - 1} \left( \sum_A \tilde{p}_{Ai}^2 - \frac{1}{n_i} \right)$$

so we have that  $\tilde{H}_i = 1 - \tilde{M}_{Wi}$  or  $\tilde{M}_{Wi} = 1 - \tilde{H}_i$ .

161

## Heterozygosity and Matching

Recall from the previous section that the allelic matching proportion within the sample from population  $i$  is

$$\tilde{M}_{Wi} = \frac{n_i}{n_i - 1} \left( \sum_A \tilde{p}_{Ai}^2 - \frac{1}{n_i} \right)$$

so we have that  $\tilde{H}_i = 1 - \tilde{M}_{Wi}$  or  $\tilde{M}_{Wi} = 1 - \tilde{H}_i$ .

161

## Between-Population Heterozygosity

For samples from populations  $i$  and  $j$ , the between-population heterozygosity is

$$\tilde{H}_{ij} = 1 - \sum_A \tilde{p}_{Ai} \tilde{p}_{Aj} = 1 - \tilde{M}_{Bij}$$

162

## Between-Population Heterozygosity

For samples from populations  $i$  and  $j$ , the between-population heterozygosity is

$$\tilde{H}_{ij} = 1 - \sum_A \tilde{p}_{Ai} \tilde{p}_{Aj} = 1 - \tilde{M}_{Bij}$$

162

## **Autosomal Estimates of $\theta$**

Using the same logic as for Y-STR profiles, the estimates of  $\theta$  are

$$\tilde{\beta}_i = 1 - \frac{\tilde{H}_i}{\tilde{H}_B}$$

$$\tilde{\beta}_W = 1 - \frac{\tilde{H}_W}{\tilde{H}_B}$$

$$\tilde{\beta}_{ij} = 1 - \frac{\tilde{H}_{ij}}{\tilde{H}_B}$$

163

## **Autosomal Estimates of $\theta$**

Using the same logic as for Y-STR profiles, the estimates of  $\theta$  are

$$\tilde{\beta}_i = 1 - \frac{\tilde{H}_i}{\tilde{H}_B}$$

$$\tilde{\beta}_W = 1 - \frac{\tilde{H}_W}{\tilde{H}_B}$$

$$\tilde{\beta}_{ij} = 1 - \frac{\tilde{H}_{ij}}{\tilde{H}_B}$$

163

## Autosomal Estimates of $\theta$

These estimates have expectations

$$\mathcal{E}(\tilde{\beta}_i) = \frac{\theta_i - \theta_B}{1 - \theta_B}$$

$$\mathcal{E}(\tilde{\beta}_W) = \frac{\theta_W - \theta_B}{1 - \theta_B}$$

$$\mathcal{E}(\tilde{\beta}_{ij}) = \frac{\theta_{ij} - \theta_B}{1 - \theta_B}$$

164

## Autosomal Estimates of $\theta$

These estimates have expectations

$$\mathcal{E}(\tilde{\beta}_i) = \frac{\theta_i - \theta_B}{1 - \theta_B}$$

$$\mathcal{E}(\tilde{\beta}_W) = \frac{\theta_W - \theta_B}{1 - \theta_B}$$

$$\mathcal{E}(\tilde{\beta}_{ij}) = \frac{\theta_{ij} - \theta_B}{1 - \theta_B}$$

164

## 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?

165

## 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?

165

## Probability of AA, AB

From the sampling formula:

$$\begin{aligned}\Pr(AA, AB) &= 4 \Pr(AAAAB) \\&= 4 \Pr(B|AAA) \Pr(AAA) \\&= 4 \Pr(B|AAA) \Pr(A|AA) \Pr(AA) \\&= 4 \Pr(B|AAA) \Pr(A|AA) \Pr(A|A) \Pr(A) \\&= \frac{4p_A p_B [\theta + (1 - \theta)p_A][2\theta + (1 - \theta)p_A](1 - \theta)p_B}{(1 + \theta)(1 + 2\theta)}\end{aligned}$$

166

## Probability of AA, AB

From the sampling formula:

$$\begin{aligned}\Pr(AA, AB) &= 4 \Pr(AAAAB) \\&= 4 \Pr(B|AAA) \Pr(AAA) \\&= 4 \Pr(B|AAA) \Pr(A|AA) \Pr(AA) \\&= 4 \Pr(B|AAA) \Pr(A|AA) \Pr(A|A) \Pr(A) \\&= \frac{4p_A p_B [\theta + (1 - \theta)p_A][2\theta + (1 - \theta)p_A](1 - \theta)p_B}{(1 + \theta)(1 + 2\theta)}\end{aligned}$$

166

## Probability of $AB, AC$

From the sampling formula:

$$\begin{aligned}\Pr(AB, AC) &= 4 \Pr(AABC) \\&= 4 \Pr(C|AAB) \Pr(AAB) \\&= 4 \Pr(C|AAB) \Pr(B|AA) \Pr(AA) \\&= 4 \Pr(C|AAB) \Pr(B|AA) \Pr(A|A) \Pr(A) \\&= \frac{4p_A[\theta + (1 - \theta)p_A](1 - \theta)p_B(1 - \theta)p_C}{(1 + \theta)(1 + 2\theta)}\end{aligned}$$

167

## Probability of $AB, AC$

From the sampling formula:

$$\begin{aligned}\Pr(AB, AC) &= 4 \Pr(AABC) \\&= 4 \Pr(C|AAB) \Pr(AAB) \\&= 4 \Pr(C|AAB) \Pr(B|AA) \Pr(AA) \\&= 4 \Pr(C|AAB) \Pr(B|AA) \Pr(A|A) \Pr(A) \\&= \frac{4p_A[\theta + (1 - \theta)p_A](1 - \theta)p_B(1 - \theta)p_C}{(1 + \theta)(1 + 2\theta)}\end{aligned}$$

167

## Probability of AA, BC

From the sampling formula:

$$\begin{aligned}\Pr(AA, BC) &= 4 \Pr(AABC) \\&= 4 \Pr(C|AAB) \Pr(AAB) \\&= 4 \Pr(C|AAB) \Pr(B|AA) \Pr(AA) \\&= 4 \Pr(C|AAB) \Pr(B|AA) \Pr(A|A) \Pr(A) \\&= \frac{4p_A[\theta + (1 - \theta)p_A](1 - \theta)p_B(1 - \theta)p_C}{(1 + \theta)(1 + 2\theta)}\end{aligned}$$

168

## Probability of AA, BC

From the sampling formula:

$$\begin{aligned}\Pr(AA, BC) &= 4 \Pr(AABC) \\&= 4 \Pr(C|AAB) \Pr(AAB) \\&= 4 \Pr(C|AAB) \Pr(B|AA) \Pr(AA) \\&= 4 \Pr(C|AAB) \Pr(B|AA) \Pr(A|A) \Pr(A) \\&= \frac{4p_A[\theta + (1 - \theta)p_A](1 - \theta)p_B(1 - \theta)p_C}{(1 + \theta)(1 + 2\theta)}\end{aligned}$$

168

## Probability of $AB, CD$

From the sampling formula:

$$\begin{aligned}\Pr(AB, AC) &= 4 \Pr(ABCD) \\&= 4 \Pr(D|ABC) \Pr(ABC) \\&= 4 \Pr(D|ABC) \Pr(C|AB) \Pr(AB) \\&= 4 \Pr(C|ABC) \Pr(C|AB) \Pr(B|A) \Pr(A) \\&= \frac{4p_A(1-\theta)p_B(1-\theta)p_C(1-\theta)p_D}{(1+\theta)(1+2\theta)}\end{aligned}$$

169

## Probability of $AB, CD$

From the sampling formula:

$$\begin{aligned}\Pr(AB, AC) &= 4 \Pr(ABCD) \\&= 4 \Pr(D|ABC) \Pr(ABC) \\&= 4 \Pr(D|ABC) \Pr(C|AB) \Pr(AB) \\&= 4 \Pr(C|ABC) \Pr(C|AB) \Pr(B|A) \Pr(A) \\&= \frac{4p_A(1-\theta)p_B(1-\theta)p_C(1-\theta)p_D}{(1+\theta)(1+2\theta)}\end{aligned}$$

169

## Database Matching

The previous slides gave the probabilities of specific pairs of genotypes (sets of four alleles).

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

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

170

## Database Matching

The previous slides gave the probabilities of specific pairs of genotypes (sets of four alleles).

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

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

170

## Database Matching

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

$$P_2 = \frac{1}{D} [6\theta^3 + \theta^2(1-\theta)(2+9S_2) + 2\theta(1-\theta)^2(2S_2+S_3) + (1-\theta)^3(2S_2^2-S_4)]$$

$$P_1 = \frac{1}{D} [8\theta^2(1-\theta)(1-S_2) + 4\theta(1-\theta)^2(1-S_3) + 4(1-\theta)^3(S_2-S_3-S_2^2+S_4)]$$

$$P_0 = \frac{1}{D} [\theta^2(1-\theta)(1-S_2) + 2\theta(1-\theta)^2(1-2S_2+S_3) + (1-\theta)^3(1-4S_2+4S_3+2S_2^2-3S_4)]$$

where  $D = (1+\theta)(1+2\theta)$ ,  $S_2 = \sum_A p_A^2$ ,  $S_3 = \sum_A p_A^3$ ,  $S_4 = \sum_A p_A^4$ .

171

## Database Matching

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

$$P_2 = \frac{1}{D} [6\theta^3 + \theta^2(1-\theta)(2+9S_2) + 2\theta(1-\theta)^2(2S_2+S_3) + (1-\theta)^3(2S_2^2-S_4)]$$

$$P_1 = \frac{1}{D} [8\theta^2(1-\theta)(1-S_2) + 4\theta(1-\theta)^2(1-S_3) + 4(1-\theta)^3(S_2-S_3-S_2^2+S_4)]$$

$$P_0 = \frac{1}{D} [\theta^2(1-\theta)(1-S_2) + 2\theta(1-\theta)^2(1-2S_2+S_3) + (1-\theta)^3(1-4S_2+4S_3+2S_2^2-3S_4)]$$

where  $D = (1+\theta)(1+2\theta)$ ,  $S_2 = \sum_A p_A^2$ ,  $S_3 = \sum_A p_A^3$ ,  $S_4 = \sum_A p_A^4$ .

171

## Database Matching

If population structure is ignored,  $\theta = 0$ :

$$P_2 = 2S_2^2 - S_4$$

$$P_1 = 4S_2 - 4S_3 - 4S_2^2 + 4S_4$$

$$P_0 = 1 - 4S_2 + 4S_3 + 2S_2^2 - 3S_4$$

where  $S_2 = \sum_A p_A^2$ ,  $S_3 = \sum_A p_A^3$ ,  $S_4 = \sum_A p_A^4$ .

For any value of  $\theta$  we can predict the matching, partially matching and mismatching proportions in a database.

172

## Database Matching

If population structure is ignored,  $\theta = 0$ :

$$P_2 = 2S_2^2 - S_4$$

$$P_1 = 4S_2 - 4S_3 - 4S_2^2 + 4S_4$$

$$P_0 = 1 - 4S_2 + 4S_3 + 2S_2^2 - 3S_4$$

where  $S_2 = \sum_A p_A^2$ ,  $S_3 = \sum_A p_A^3$ ,  $S_4 = \sum_A p_A^4$ .

For any value of  $\theta$  we can predict the matching, partially matching and mismatching proportions in a database.

172

## FBI Caucasian Matching Counts

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

Locus	Obs.	$\theta = 0.000$	$\theta = 0.001$	$\theta = 0.005$	$\theta = 0.010$	$\theta = 0.030$
D3S1358	1443	<b>1397</b>	<b>1406</b>	<b>1441</b>	1485	1669
vWA	1179	<b>1168</b>	<b>1177</b>	1212	1256	1440
FGA	679	<b>668</b>	<b>675</b>	705	743	903
D8S1179	1188	1256	1266	1305	1354	1555
D21S11	677	710	718	749	789	955
D18S51	509	530	537	564	599	749
D5S818	3054	<b>2960</b>	<b>2971</b>	<b>3012</b>	3065	3279
D13S317	1414	1588	1598	1639	1689	1897
D7S820	1170	1222	1231	1267	1312	1499
CSF1PO	2290	<b>2212</b>	<b>2222</b>	<b>2260</b>	2309	2509
TPOX	3860	<b>3646</b>	<b>3659</b>	<b>3712</b>	<b>3777</b>	4038
THO1	1393	1522	1531	1568	1614	1805
D16S539	1614	1658	1668	1708	1758	1963

Boldface when observed number is greater than expected number.

173

## FBI Caucasian Matching Counts

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

Locus	Obs.	$\theta = 0.000$	$\theta = 0.001$	$\theta = 0.005$	$\theta = 0.010$	$\theta = 0.030$
D3S1358	1443	<b>1397</b>	<b>1406</b>	<b>1441</b>	1485	1669
vWA	1179	<b>1168</b>	<b>1177</b>	1212	1256	1440
FGA	679	<b>668</b>	<b>675</b>	705	743	903
D8S1179	1188	1256	1266	1305	1354	1555
D21S11	677	710	718	749	789	955
D18S51	509	530	537	564	599	749
D5S818	3054	<b>2960</b>	<b>2971</b>	<b>3012</b>	3065	3279
D13S317	1414	1588	1598	1639	1689	1897
D7S820	1170	1222	1231	1267	1312	1499
CSF1PO	2290	<b>2212</b>	<b>2222</b>	<b>2260</b>	2309	2509
TPOX	3860	<b>3646</b>	<b>3659</b>	<b>3712</b>	<b>3777</b>	4038
THO1	1393	1522	1531	1568	1614	1805
D16S539	1614	1658	1668	1708	1758	1963

Boldface when observed number is greater than expected number.

173

## FBI Database Matching Proportions

Locus	Observed	$\theta$				
		.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
TH01	.074	.081	.082	.084	.086	.096
D16S539	.086	.089	.089	.091	.094	.105

174

## FBI Database Matching Proportions

Locus	Observed	$\theta$				
		.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
TH01	.074	.081	.082	.084	.086	.096
D16S539	.086	.089	.089	.091	.094	.105

174

## ESR 10-locus Matching Counts

		$\theta = 0.03$								
$x$		$y = 0$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$	$y = 7$	
0	e	64503	758792	3985493	12307140	24740935	33828661	31856963	20399527	
	o	230893	2131120	8848782	21782292	35224130	39151651	30314540	16177419	
1	e	115278	1202561	5531046	14719859	24977273	28020130	20778516	9819915	
	o	268277	2211969	8169809	17606306	24465883	22796532	14257072	5786109	
2	e	90106	822966	3261626	7325604	10196919	9006180	4928044	1527043	
	o	137454	1018744	3297466	6125791	7163409	5402925	2576700	710754	
3	e	40536	318975	1066732	1965066	2153123	1402933	503204	76621	
	o	42289	274281	767534	1204489	1144898	661756	215478	30539	
4	e	11616	77117	211489	306613	247786	105802	18641	0	
	o	8690	48338	114443	147273	107214	42242	7205	0	
5	e	2214	12052	26006	27801	14718	3086	0	0	
	o	1244	5888	11316	11155	5648	1262	0	0	
6	e	284	1216	1935	1355	352	0	0	0	
	o	120	539	747	531	213	0	0	0	
7	e	24	76	80	27	0	0	0	0	
	o	10	40	36	33	0	0	0	0	
8	e	1	3	1	0	0	0	0	0	
	o	0	0	11	0	0	0	0	0	
9	e	0	0	0	0	0	0	0	0	
	o	1	1	0	0	0	0	0	0	

175

## ESR 10-locus Matching Counts

		$\theta = 0.03$								
$x$		$y = 0$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$	$y = 7$	
0	e	64503	758792	3985493	12307140	24740935	33828661	31856963	20399527	
	o	230893	2131120	8848782	21782292	35224130	39151651	30314540	16177419	
1	e	115278	1202561	5531046	14719859	24977273	28020130	20778516	9819915	
	o	268277	2211969	8169809	17606306	24465883	22796532	14257072	5786109	
2	e	90106	822966	3261626	7325604	10196919	9006180	4928044	1527043	
	o	137454	1018744	3297466	6125791	7163409	5402925	2576700	710754	
3	e	40536	318975	1066732	1965066	2153123	1402933	503204	76621	
	o	42289	274281	767534	1204489	1144898	661756	215478	30539	
4	e	11616	77117	211489	306613	247786	105802	18641	0	
	o	8690	48338	114443	147273	107214	42242	7205	0	
5	e	2214	12052	26006	27801	14718	3086	0	0	
	o	1244	5888	11316	11155	5648	1262	0	0	
6	e	284	1216	1935	1355	352	0	0	0	
	o	120	539	747	531	213	0	0	0	
7	e	24	76	80	27	0	0	0	0	
	o	10	40	36	33	0	0	0	0	
8	e	1	3	1	0	0	0	0	0	
	o	0	0	11	0	0	0	0	0	
9	e	0	0	0	0	0	0	0	0	
	o	1	1	0	0	0	0	0	0	

175

## FBI Database Matching Counts

Matching loci	$\theta$	Number of Partially Matching Loci													
		0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	Obs.	0	3	18	92	249	624	1077	1363	1116	849	379	112	25	4
	.000	0	2	19	90	293	672	1129	1403	1290	868	415	134	26	2
	.010	0	2	14	70	236	566	992	1289	1241	875	439	148	30	3
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	0					
	.000	0	0	1	1	1	1	0	0	0					
	.010	0	0	1	2	2	1	1	0	0					

176

## FBI Database Matching Counts

Matching loci	$\theta$	Number of Partially Matching Loci													
		0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	Obs.	0	3	18	92	249	624	1077	1363	1116	849	379	112	25	4
	.000	0	2	19	90	293	672	1129	1403	1290	868	415	134	26	2
	.010	0	2	14	70	236	566	992	1289	1241	875	439	148	30	3
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	0					
	.000	0	0	1	1	1	1	0	0	0					
	.010	0	0	1	2	2	1	1	0	0					

176

## Predicted Matches when $n = 65,493$

Matching loci	Number of partially matching 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							

177

## Predicted Matches when $n = 65,493$

Matching loci	Number of partially matching 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							

177

## **INTERPRETATION OF MIXTURES**

178

## **INTERPRETATION OF MIXTURES**

178

# Contents

- Frequentist Approach
  - Cumulative Probability of Inclusion (CPI)
  - Random Man Not Excluded (RMNE)
- Likelihood Ratio
  - Binary Model
  - Semi-continuous Model
  - Continuous Model

179

# Contents

- Frequentist Approach
  - Cumulative Probability of Inclusion (CPI)
  - Random Man Not Excluded (RMNE)
- Likelihood Ratio
  - Binary Model
  - Semi-continuous Model
  - Continuous Model

179

# Assumptions

- There is only one population.
- This population is in Hardy-Weinberg equilibrium.
- All of the contributors to the mixtures are unrelated.

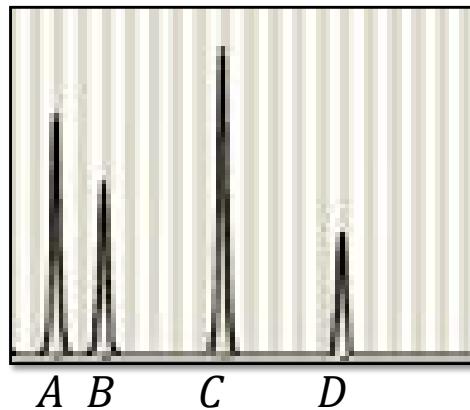
180

# Assumptions

- There is only one population.
- This population is in Hardy-Weinberg equilibrium.
- All of the contributors to the mixtures are unrelated.

180

# Cumulative Probability of Inclusion (CPI)

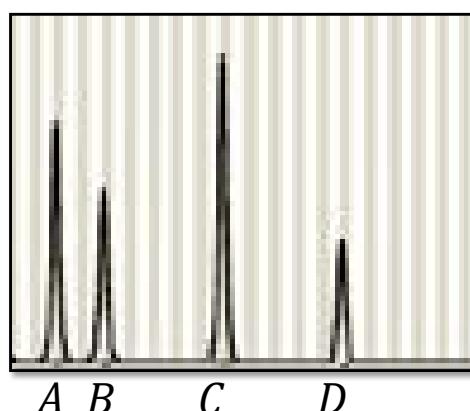


$$CPI = (p_A + p_B + p_C + p_D)^2$$

the probability that a random person would be included as a contributor to this mixture

181

# Cumulative Probability of Inclusion (CPI)

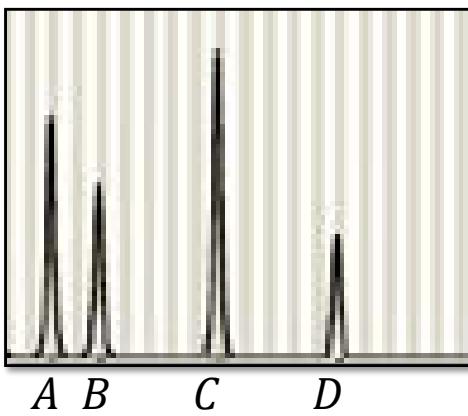


$$CPI = (p_A + p_B + p_C + p_D)^2$$

the probability that a random person would be included as a contributor to this mixture

181

# Cumulative Probability of Inclusion (CPI)



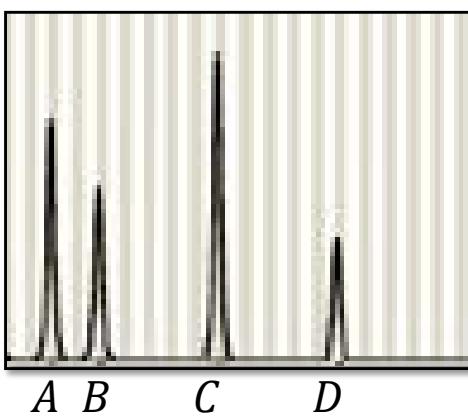
random person:  
AA  
AB  
AC  
AD  
BB  
BC  
BD  
CC  
CD  
DD

$$CPI = (p_A + p_B + p_C + p_D)^2$$

the probability that a random person would be included as a contributor to this mixture

182

# Cumulative Probability of Inclusion (CPI)



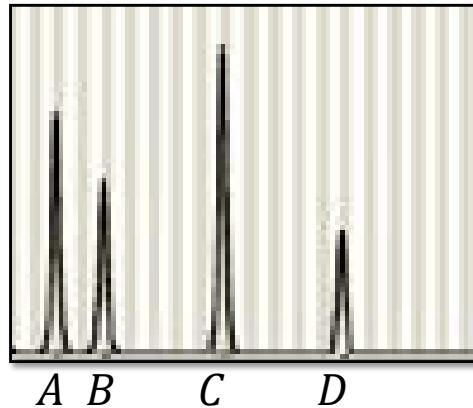
random person:  
AA  
AB  
AC  
AD  
BB  
BC  
BD  
CC  
CD  
DD

$$CPI = (p_A + p_B + p_C + p_D)^2$$

the probability that a random person would be included as a contributor to this mixture

182

# Random Man Not Excluded (RMNE)

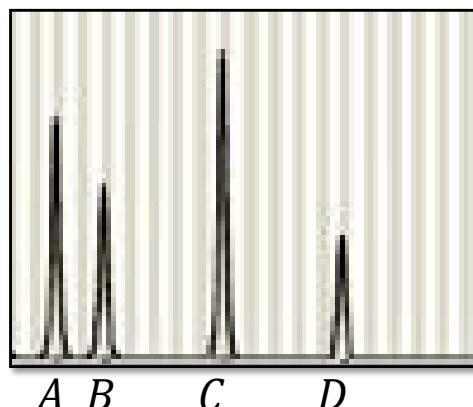


$$1 - CPI = 1 - (p_A + p_B + p_C + p_D)^2$$

the probability that a random person would be excluded as a contributor to this mixture

183

# Random Man Not Excluded (RMNE)



$$1 - CPI = 1 - (p_A + p_B + p_C + p_D)^2$$

the probability that a random person would be excluded as a contributor to this mixture

183

But CPI and RMNE don't tell me anything about the likelihood ratio for the person of interest (for example, the defendant).



$$\frac{\Pr(H_p|E, I)}{\Pr(H_d|E, I)} = \underbrace{\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}}_{\text{Likelihood Ratio (or Bayes Factor)}} \times \frac{\Pr(H_p|I)}{\Pr(H_d|I)}$$

184

But CPI and RMNE don't tell me anything about the likelihood ratio for the person of interest (for example, the defendant).



$$\frac{\Pr(H_p|E, I)}{\Pr(H_d|E, I)} = \underbrace{\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}}_{\text{Likelihood Ratio (or Bayes Factor)}} \times \frac{\Pr(H_p|I)}{\Pr(H_d|I)}$$

184

# Likelihood Ratio

$$\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}$$

$E$ : DNA typing results

$G_C$ : crime stain

$G_K$ : known contributors

$\begin{cases} G_V: \text{victim or complainant} \\ G_S: \text{suspect or defendant} \end{cases}$

$I$  will be omitted in the rest of this presentation to simplify the mathematical notations

185

# Likelihood Ratio

$$\frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}$$

$E$ : DNA typing results

$G_C$ : crime stain

$G_K$ : known contributors

$\begin{cases} G_V: \text{victim or complainant} \\ G_S: \text{suspect or defendant} \end{cases}$

$I$  will be omitted in the rest of this presentation to simplify the mathematical notations

185

# Likelihood Ratio

$$LR = \frac{\Pr(G_C | G_K^p, H_p)}{\Pr(G_C | G_K^d, H_d)}$$

where  $G_K^p$  = profiles of known contributors if  $H_p$  is true  
 $G_K^d$  = profiles of known contributors if  $H_d$  is true

186

# Likelihood Ratio

$$LR = \frac{\Pr(G_C | G_K^p, H_p)}{\Pr(G_C | G_K^d, H_d)}$$

where  $G_K^p$  = profiles of known contributors if  $H_p$  is true  
 $G_K^d$  = profiles of known contributors if  $H_d$  is true

186

# Likelihood Ratio

Let us consider **2-person mixtures.**

(Applying this approach to >2 persons follows the same principles as for 2-person mixtures.)

187

# Likelihood Ratio

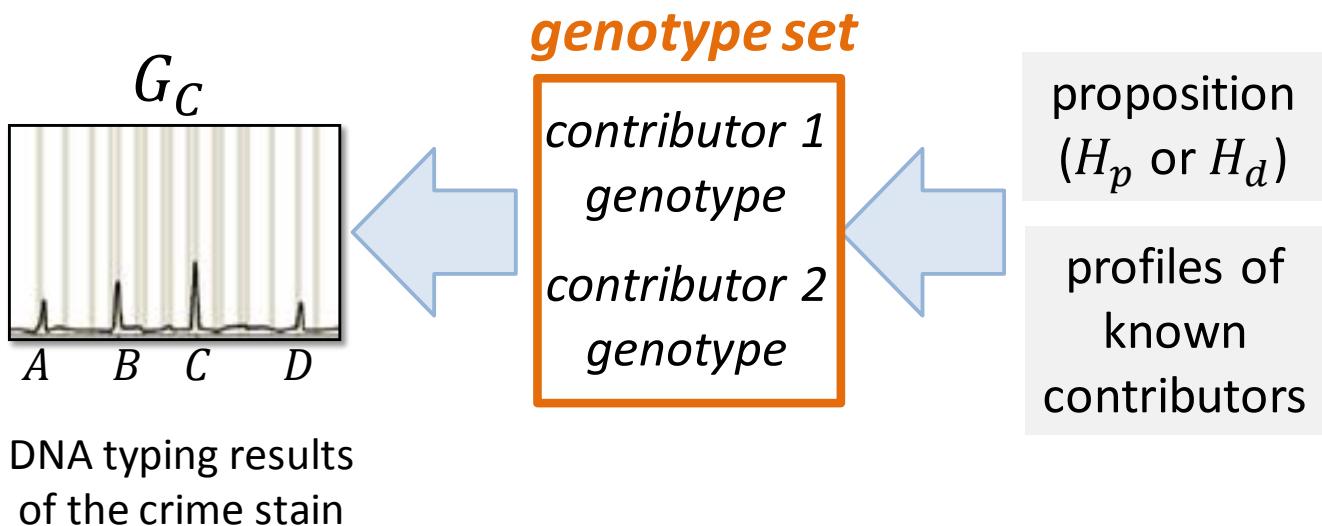
Let us consider **2-person mixtures.**

(Applying this approach to >2 persons follows the same principles as for 2-person mixtures.)

187

# Likelihood Ratio

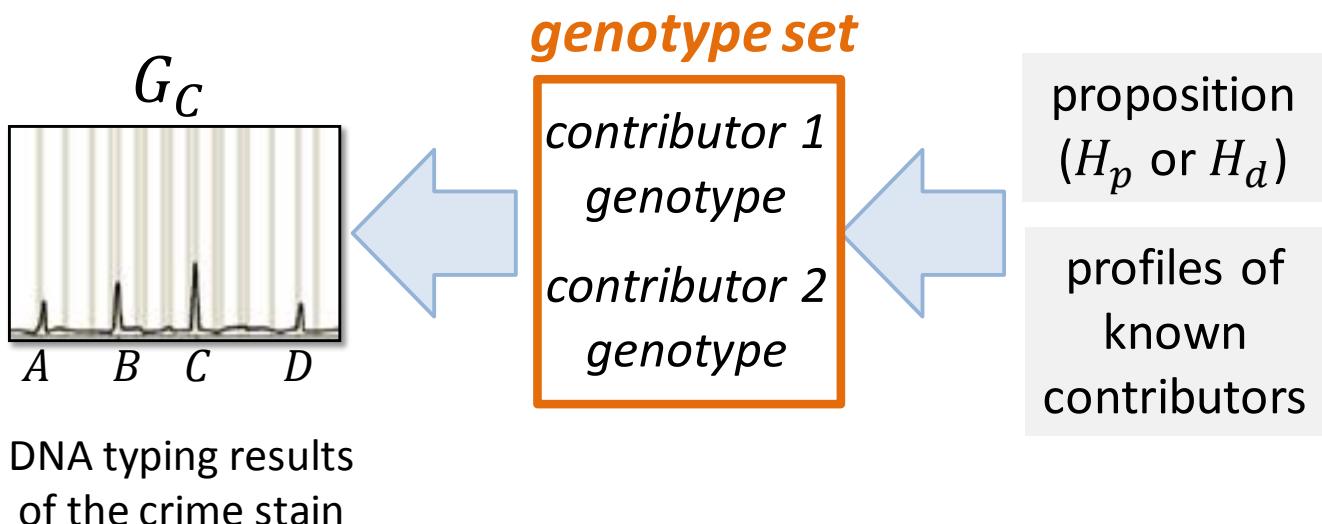
The probability of  $G_C$  depends on the genotypes considered for the contributors.



188

# Likelihood Ratio

The probability of  $G_C$  depends on the genotypes considered for the contributors.



188

# Likelihood Ratio

It is therefore reasonable to apply the law of total probability to represent the probability of  $G_C$  as the weighted sum of the probabilities of  $G_C$  given each of the possible genotype sets of the contributors:

$$LR = \frac{\sum_{j=1}^M \Pr(G_C | S_j) \Pr(S_j | G_K^p, H_p)}{\sum_{i=1}^N \Pr(G_C | S_i) \Pr(S_i | G_K^d, H_d)}$$

where  $M \leq N$ , and

- $S_1, \dots, S_M$ : the genotype sets considered for all of the contributors given the profiles of the known contributors under  $H_p$ ,
- $S_1, \dots, S_N$ : the genotype sets considered for all of the contributors given the profiles of the known contributors under  $H_d$ .

189

# Likelihood Ratio

It is therefore reasonable to apply the law of total probability to represent the probability of  $G_C$  as the weighted sum of the probabilities of  $G_C$  given each of the possible genotype sets of the contributors:

$$LR = \frac{\sum_{j=1}^M \Pr(G_C | S_j) \Pr(S_j | G_K^p, H_p)}{\sum_{i=1}^N \Pr(G_C | S_i) \Pr(S_i | G_K^d, H_d)}$$

where  $M \leq N$ , and

- $S_1, \dots, S_M$ : the genotype sets considered for all of the contributors given the profiles of the known contributors under  $H_p$ ,
- $S_1, \dots, S_N$ : the genotype sets considered for all of the contributors given the profiles of the known contributors under  $H_d$ .

189

# Likelihood Ratio

A genotype set  $S_i$  lists the genotypes of contributor 1 and contributor 2 to the mixture.

$$S_i = \text{genotype of contributor 1 and} \\ \text{genotype of contributor 2}$$

If peak heights are taken into account, then  $S_i$  lists the genotypes of the major contributor and the minor contributor to the mixture.

$$S_i = \text{genotype of major contributor and} \\ \text{genotype of minor contributor}$$

190

# Likelihood Ratio

A genotype set  $S_i$  lists the genotypes of contributor 1 and contributor 2 to the mixture.

$$S_i = \text{genotype of contributor 1 and} \\ \text{genotype of contributor 2}$$

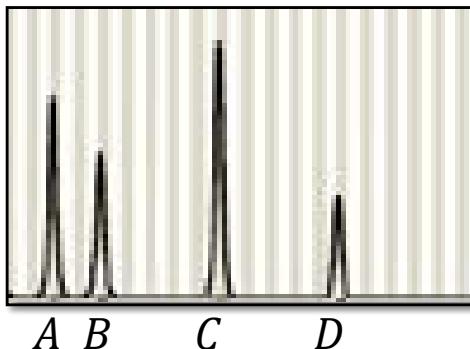
If peak heights are taken into account, then  $S_i$  lists the genotypes of the major contributor and the minor contributor to the mixture.

$$S_i = \text{genotype of major contributor and} \\ \text{genotype of minor contributor}$$

190

# Likelihood Ratio

If this is considered to be a mixture of two contributors without allele drop-ins, then some of the genotypes included in the frequentist approach may be unreasonable.



The likelihood ratio will consider the probabilities of  $G_C$  as a whole, and therefore focus on the more reasonable genotype sets.

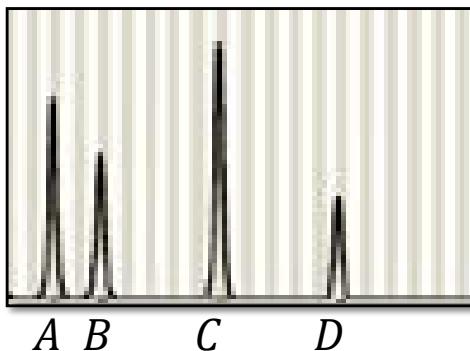
random person:

~~AA~~  
~~AB~~  
~~AC~~  
~~AD~~  
~~BB~~  
~~BC~~  
~~BD~~  
~~CC~~  
~~CD~~  
~~DD~~

191

# Likelihood Ratio

If this is considered to be a mixture of two contributors without allele drop-ins, then some of the genotypes included in the frequentist approach may be unreasonable.



The likelihood ratio will consider the probabilities of  $G_C$  as a whole, and therefore focus on the more reasonable genotype sets.

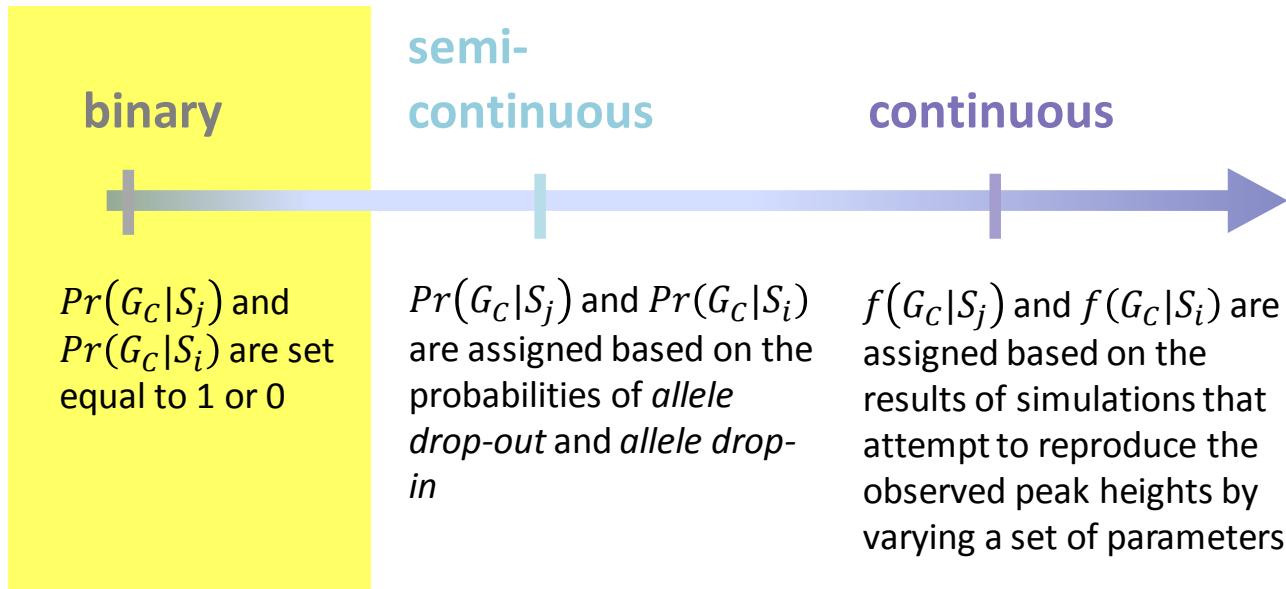
random person:

~~AA~~  
~~AB~~  
~~AC~~  
~~AD~~  
~~BB~~  
~~BC~~  
~~BD~~  
~~CC~~  
~~CD~~  
~~DD~~

191

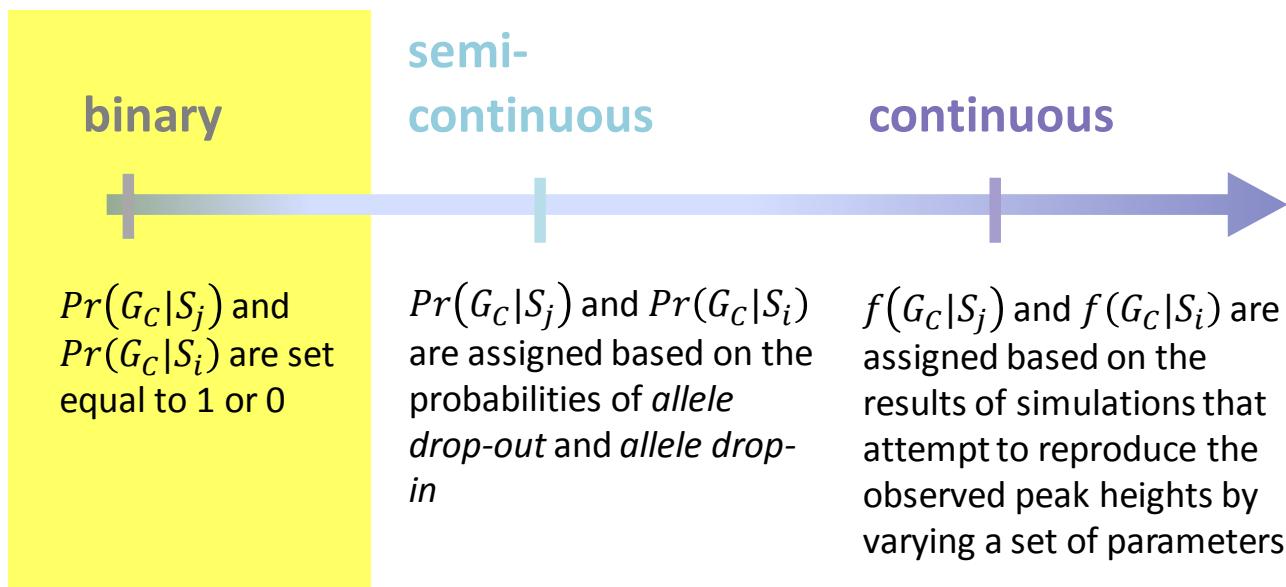
# Likelihood Ratio

$$LR = \frac{\sum_{j=1}^M \Pr(G_C|S_j) \Pr(S_j|G_K^p, H_p)}{\sum_{i=1}^N \Pr(G_C|S_i) \Pr(S_i|G_K^d, H_d)} \quad \text{where } M \leq N$$



# Likelihood Ratio

$$LR = \frac{\sum_{j=1}^M \Pr(G_C|S_j) \Pr(S_j|G_K^p, H_p)}{\sum_{i=1}^N \Pr(G_C|S_i) \Pr(S_i|G_K^d, H_d)} \quad \text{where } M \leq N$$



# Binary Model

- A peak is either present or absent. The peak heights are not taken into account.

- $$LR = \frac{\sum_{j=1}^M \Pr(G_C|S_j) \Pr(S_j|G_K^p, H_p)}{\sum_{i=1}^N \Pr(G_C|S_i) \Pr(S_i|G_K^d, H_d)}$$
 where  $M \leq N$

each of these probabilities is set equal to 0 or 1

193

# Binary Model

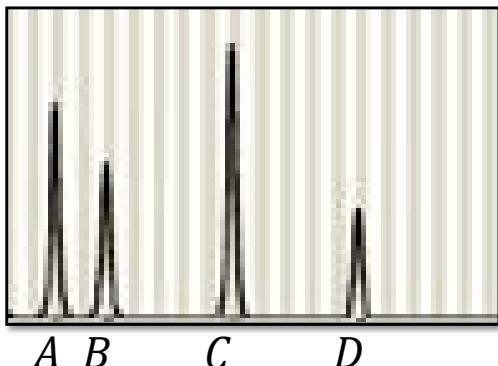
- A peak is either present or absent. The peak heights are not taken into account.

- $$LR = \frac{\sum_{j=1}^M \Pr(G_C|S_j) \Pr(S_j|G_K^p, H_p)}{\sum_{i=1}^N \Pr(G_C|S_i) \Pr(S_i|G_K^d, H_d)}$$
 where  $M \leq N$

each of these probabilities is set equal to 0 or 1

193

# Binary Model



Case 1

$$G_C = \{A, B, C, D\} \quad G_V = \{B, C\}$$

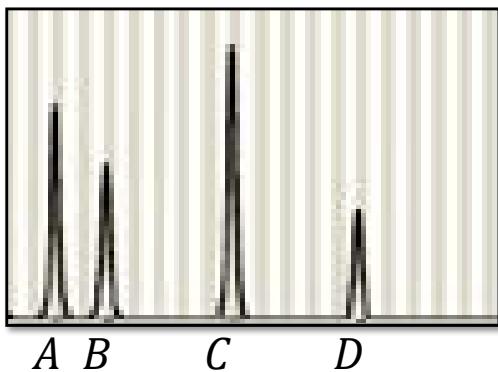
$$G_S = \{A, D\}$$

$H_p$ : The crime stain contains the DNA from the victim and the suspect ( $G_K^p = \{G_V, G_S\}$ ).

$H_d$ : The crime stain contains the DNA from the victim and an unknown unrelated person ( $G_K^d = \{G_V\}$ ).

194

# Binary Model



Case 1

$$G_C = \{A, B, C, D\} \quad G_V = \{B, C\}$$

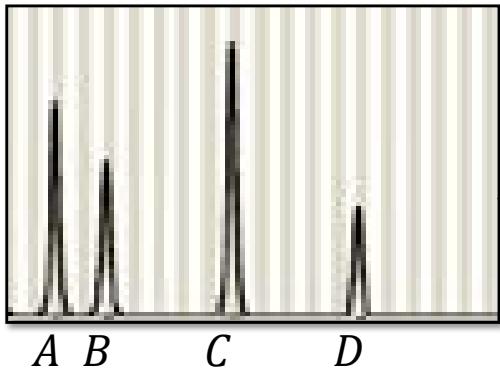
$$G_S = \{A, D\}$$

$H_p$ : The crime stain contains the DNA from the victim and the suspect ( $G_K^p = \{G_V, G_S\}$ ).

$H_d$ : The crime stain contains the DNA from the victim and an unknown unrelated person ( $G_K^d = \{G_V\}$ ).

194

# Binary Model



Case 1

$$G_C = \{A, B, C, D\} \quad G_V = \{B, C\}$$

$$G_S = \{A, D\}$$

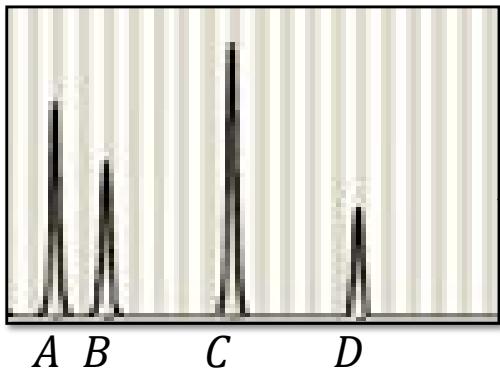
We set *contributor* 1 =  $G_V$ .  
 $S_1$ : BC and AD

$$LR = \frac{Pr(S_1 | G_S, G_V, H_p)}{Pr(S_1 | G_V, H_d)}$$

$$= \frac{1}{2p_A p_D}$$

195

# Binary Model



Case 1

$$G_C = \{A, B, C, D\} \quad G_V = \{B, C\}$$

$$G_S = \{A, D\}$$

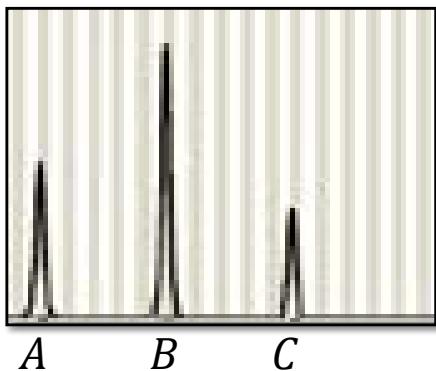
We set *contributor* 1 =  $G_V$ .  
 $S_1$ : BC and AD

$$LR = \frac{Pr(S_1 | G_S, G_V, H_p)}{Pr(S_1 | G_V, H_d)}$$

$$= \frac{1}{2p_A p_D}$$

195

# Binary Model



Case 2

$$G_C = \{A, B, C\}$$

$$G_V = \{B, C\}$$

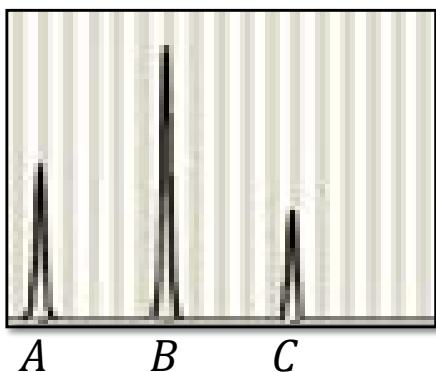
$$G_S = \{A, A\}$$

$H_p$ : The crime stain contains the DNA from the victim and the suspect ( $G_K^p = \{G_V, G_S\}$ ).

$H_d$ : The crime stain contains the DNA from the victim and an unknown unrelated person ( $G_K^d = \{G_V\}$ ).

196

# Binary Model



Case 2

$$G_C = \{A, B, C\}$$

$$G_V = \{B, C\}$$

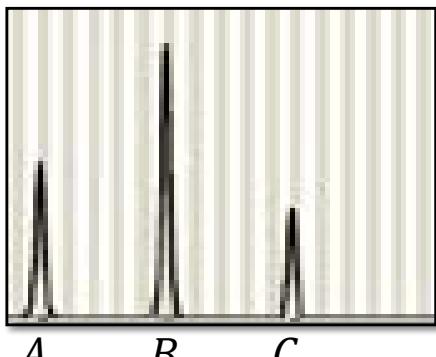
$$G_S = \{A, A\}$$

$H_p$ : The crime stain contains the DNA from the victim and the suspect ( $G_K^p = \{G_V, G_S\}$ ).

$H_d$ : The crime stain contains the DNA from the victim and an unknown unrelated person ( $G_K^d = \{G_V\}$ ).

196

# Binary Model



Case 2

$$G_C = \{A, B, C\}$$

$$G_V = \{B, C\}$$

$$G_S = \{A, A\}$$

$A \quad B \quad C$

We set *contributor* 1 =  $G_V$ .

$S_1$ : BC and AA

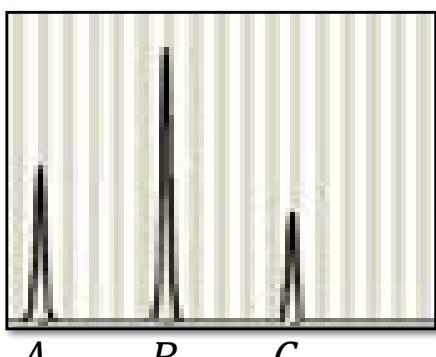
$S_2$ : BC and AB

$S_3$ : BC and AC

$$\begin{aligned} LR &= \frac{Pr(S_1|G_S, G_V, H_p)}{\sum_{i=1}^3 Pr(S_i|G_V, H_d)} \\ &= \frac{1}{p_A^2 + 2p_A p_B + 2p_A p_C} \end{aligned}$$

197

# Binary Model



Case 2

$$G_C = \{A, B, C\}$$

$$G_V = \{B, C\}$$

$$G_S = \{A, A\}$$

$A \quad B \quad C$

We set *contributor* 1 =  $G_V$ .

$S_1$ : BC and AA

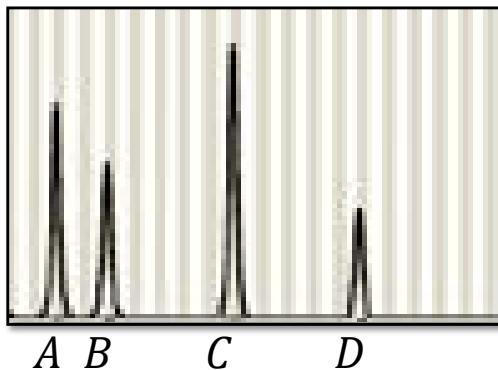
$S_2$ : BC and AB

$S_3$ : BC and AC

$$\begin{aligned} LR &= \frac{Pr(S_1|G_S, G_V, H_p)}{\sum_{i=1}^3 Pr(S_i|G_V, H_d)} \\ &= \frac{1}{p_A^2 + 2p_A p_B + 2p_A p_C} \end{aligned}$$

197

# Binary Model



Case 3

$$G_C = \{A, B, C, D\}$$

$$G_{V1} = \{B, C\}$$

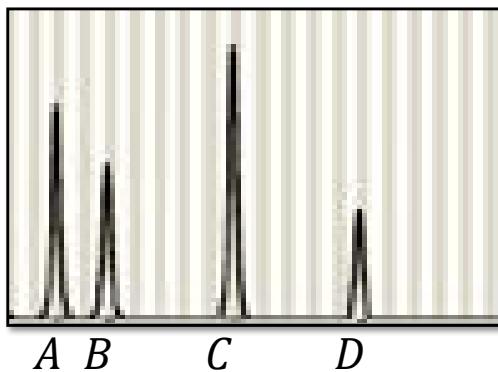
$$G_{V2} = \{A, D\}$$

$H_p$ : The crime stain contains the DNA from victim 1 and victim 2 ( $G_K^p = \{G_{V1}, G_{V2}\}$ ).

$H_d$ : The crime stain contains the DNA from two unknown and unrelated individuals ( $G_K^d = \{\}$ ).

198

# Binary Model



Case 3

$$G_C = \{A, B, C, D\}$$

$$G_{V1} = \{B, C\}$$

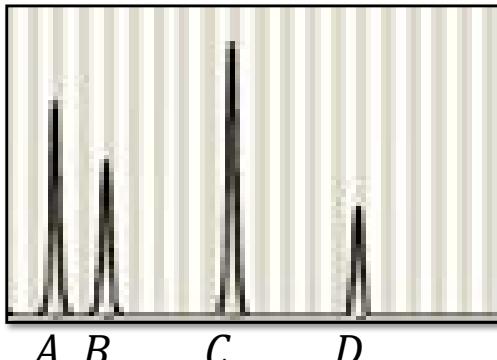
$$G_{V2} = \{A, D\}$$

$H_p$ : The crime stain contains the DNA from victim 1 and victim 2 ( $G_K^p = \{G_{V1}, G_{V2}\}$ ).

$H_d$ : The crime stain contains the DNA from two unknown and unrelated individuals ( $G_K^d = \{\}$ ).

198

# Binary Model



Case 3

$$G_C = \{A, B, C, D\} \quad G_{V1} = \{B, C\}$$

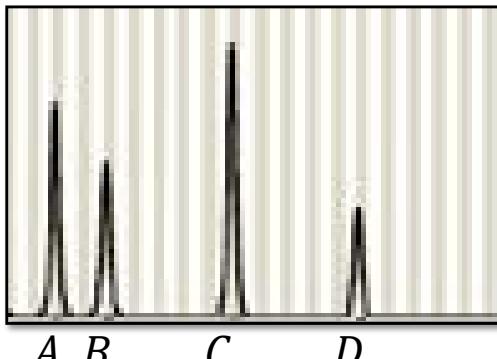
$$G_{V2} = \{A, D\}$$

- $S_1: BC$  and  $AD$
- $S_2: AD$  and  $BC$
- $S_3: AB$  and  $CD$
- $S_4: CD$  and  $AB$
- $S_5: AC$  and  $BD$
- $S_6: BD$  and  $AC$

$$\begin{aligned} LR &= \frac{Pr(S_1|G_{V1}, G_{V2}, H_p) + Pr(S_2|G_{V1}, G_{V2}, H_p)}{\sum_{i=1}^6 Pr(S_i|H_d)} \\ &= \frac{0.5 + 0.5}{6 \times 4p_A p_B p_C p_D} \\ &= \frac{1}{24p_A p_B p_C p_D} \end{aligned}$$

199

# Binary Model



Case 3

$$G_C = \{A, B, C, D\} \quad G_{V1} = \{B, C\}$$

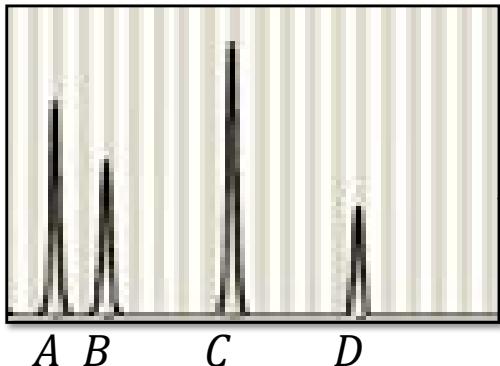
$$G_{V2} = \{A, D\}$$

- $S_1: BC$  and  $AD$
- $S_2: AD$  and  $BC$
- $S_3: AB$  and  $CD$
- $S_4: CD$  and  $AB$
- $S_5: AC$  and  $BD$
- $S_6: BD$  and  $AC$

$$\begin{aligned} LR &= \frac{Pr(S_1|G_{V1}, G_{V2}, H_p) + Pr(S_2|G_{V1}, G_{V2}, H_p)}{\sum_{i=1}^6 Pr(S_i|H_d)} \\ &= \frac{0.5 + 0.5}{6 \times 4p_A p_B p_C p_D} \\ &= \frac{1}{24p_A p_B p_C p_D} \end{aligned}$$

199

# Binary Model



Case 4

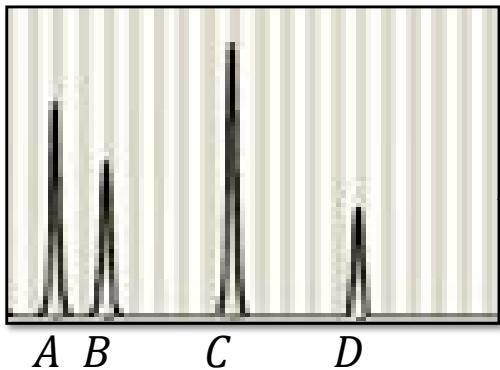
$$G_C = \{A, B, C, D\} \quad G_S = \{A, D\}$$

$H_p$ : The crime stain contains the DNA from the suspect and an unknown unrelated person ( $G_K^p = \{G_S\}$ ).

$H_d$ : The crime stain contains the DNA from two unknown and unrelated individuals ( $G_K^d = \{\}$ ).

200

# Binary Model



Case 4

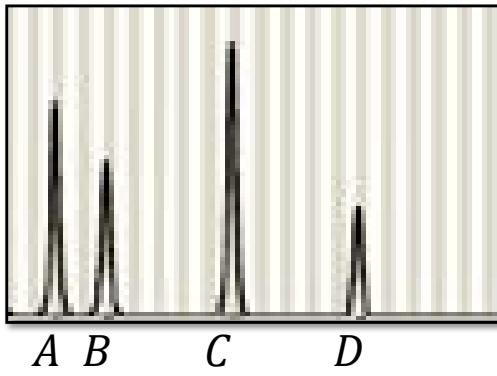
$$G_C = \{A, B, C, D\} \quad G_S = \{A, D\}$$

$H_p$ : The crime stain contains the DNA from the suspect and an unknown unrelated person ( $G_K^p = \{G_S\}$ ).

$H_d$ : The crime stain contains the DNA from two unknown and unrelated individuals ( $G_K^d = \{\}$ ).

200

# Binary Model



Case 4

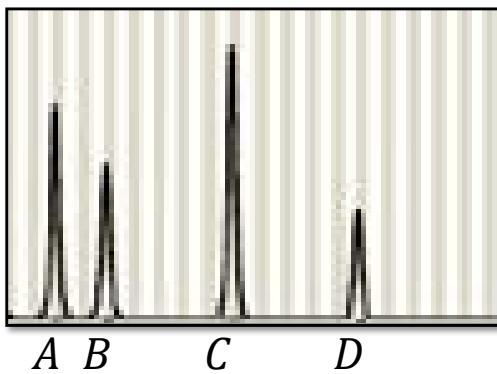
$$G_C = \{A, B, C, D\} \quad G_S = \{A, D\}$$

- $S_1: BC$  and  $AD$
- $S_2: AD$  and  $BC$
- $S_3: AB$  and  $CD$
- $S_4: CD$  and  $AB$
- $S_5: AC$  and  $BD$
- $S_6: BD$  and  $AC$

$$\begin{aligned} LR &= \frac{Pr(S_1|G_S, H_p) + Pr(S_2|G_S, H_p)}{\sum_{i=1}^6 Pr(S_i|H_d)} \\ &= \frac{0.5 \times 2p_B p_C + 0.5 \times 2p_B p_C}{24p_A p_B p_C p_D} \\ &= \frac{1}{12p_A p_D} \end{aligned}$$

201

# Binary Model



Case 4

$$G_C = \{A, B, C, D\} \quad G_S = \{A, D\}$$

- $S_1: BC$  and  $AD$
- $S_2: AD$  and  $BC$
- $S_3: AB$  and  $CD$
- $S_4: CD$  and  $AB$
- $S_5: AC$  and  $BD$
- $S_6: BD$  and  $AC$

$$\begin{aligned} LR &= \frac{Pr(S_1|G_S, H_p) + Pr(S_2|G_S, H_p)}{\sum_{i=1}^6 Pr(S_i|H_d)} \\ &= \frac{0.5 \times 2p_B p_C + 0.5 \times 2p_B p_C}{24p_A p_B p_C p_D} \\ &= \frac{1}{12p_A p_D} \end{aligned}$$

201

# Binary Model

General formula for LR:

1. write the numerator and the denominator each as an expression of type  $P_x(\text{unattributed alleles}|G_c)$  where  $x$  is the number of unknown contributors, and *unattributed alleles* lists the alleles not accounted for by the known contributors

Reference: B.S. Weir et al. Interpreting DNA Mixtures. *J Forensic Sci* 1997; 42(2): 213-222. 202

# Binary Model

General formula for LR:

1. write the numerator and the denominator each as an expression of type  $P_x(\text{unattributed alleles}|G_c)$  where  $x$  is the number of unknown contributors, and *unattributed alleles* lists the alleles not accounted for by the known contributors

Reference: B.S. Weir et al. Interpreting DNA Mixtures. *J Forensic Sci* 1997; 42(2): 213-222. 202

# Binary Model

General formula for LR:

2.

$$P_x(\text{unattributed alleles} | G_C) = (T_0)^{2x} - \sum_j (T_{1j})^{2x} + \sum_{j,k} (T_{2j,k})^{2x} - \sum_{j,k,l} (T_{3j,k,l})^{2x} + \dots$$

where:

$T_0$ : sum of the probabilities of all alleles in  $G_C$

$T_{1j}$ : sum of the probabilities of all alleles in  $G_C$  except the  $j$ th one in the *unattributed alleles*

$T_{2j,k}$ : sum of the probabilities of all alleles in  $G_C$  except the  $j$ th one and the  $k$ th one in the *unattributed alleles*

$T_{3j,k,l}$ : sum of the probabilities of all alleles in  $G_C$  except the  $j$ th one, the  $k$ th one and the  $l$ th one in the *unattributed alleles*

Reference: B.S. Weir et al. Interpreting DNA Mixtures. *J Forensic Sci* 1997; 42(2): 213-222. 203

# Binary Model

General formula for LR:

2.

$$P_x(\text{unattributed alleles} | G_C) = (T_0)^{2x} - \sum_j (T_{1j})^{2x} + \sum_{j,k} (T_{2j,k})^{2x} - \sum_{j,k,l} (T_{3j,k,l})^{2x} + \dots$$

where:

$T_0$ : sum of the probabilities of all alleles in  $G_C$

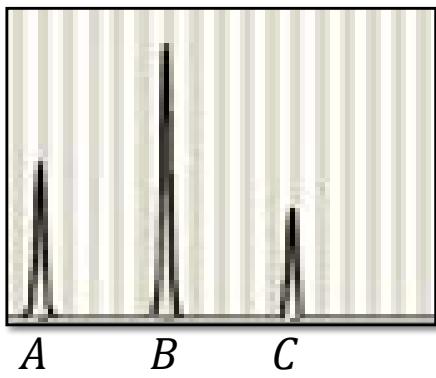
$T_{1j}$ : sum of the probabilities of all alleles in  $G_C$  except the  $j$ th one in the *unattributed alleles*

$T_{2j,k}$ : sum of the probabilities of all alleles in  $G_C$  except the  $j$ th one and the  $k$ th one in the *unattributed alleles*

$T_{3j,k,l}$ : sum of the probabilities of all alleles in  $G_C$  except the  $j$ th one, the  $k$ th one and the  $l$ th one in the *unattributed alleles*

Reference: B.S. Weir et al. Interpreting DNA Mixtures. *J Forensic Sci* 1997; 42(2): 213-222. 203

# Binary Model



Case 2

$$G_C = \{A, B, C\}$$

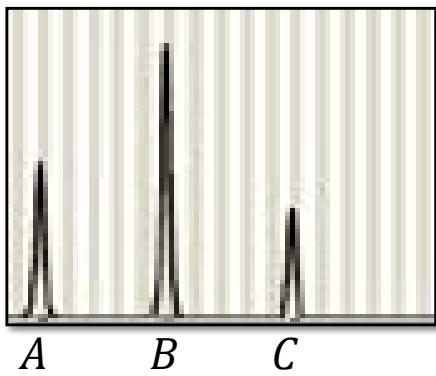
$$G_V = \{B, C\}$$

$$G_S = \{A, A\}$$

$$\begin{aligned} LR &= \frac{P_0(\emptyset | A, B, C)}{P_1(A | A, B, C)} \\ &= \frac{1}{(p_A + p_B + p_C)^2 - (p_B + p_C)^2} \end{aligned}$$

204

# Binary Model



Case 2

$$G_C = \{A, B, C\}$$

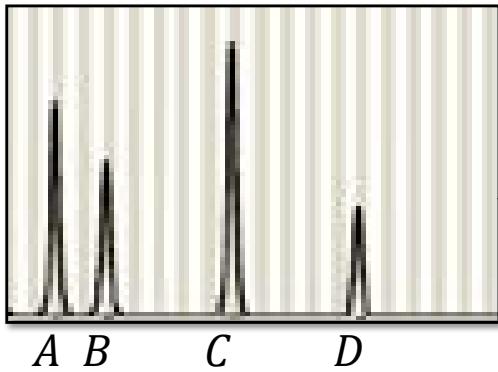
$$G_V = \{B, C\}$$

$$G_S = \{A, A\}$$

$$\begin{aligned} LR &= \frac{P_0(\emptyset | A, B, C)}{P_1(A | A, B, C)} \\ &= \frac{1}{(p_A + p_B + p_C)^2 - (p_B + p_C)^2} \end{aligned}$$

204

# Binary Model



Case 1

$$G_C = \{A, B, C, D\} \quad G_V = \{B, C\}$$

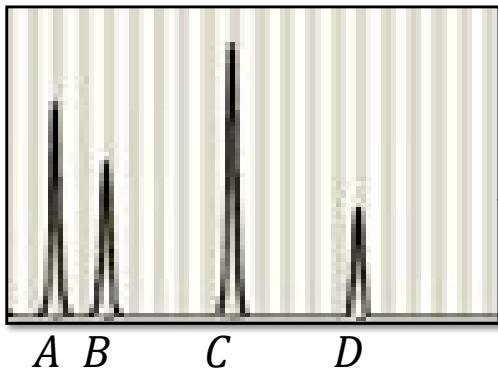
$$G_S = \{A, D\}$$

$$LR = \frac{P_0(\emptyset|A, B, C, D)}{P_1(A, D|A, B, C, D)}$$

$$= \frac{1}{(p_A + p_B + p_C + p_D)^2 - (p_B + p_C + p_D)^2 - (p_A + p_B + p_C)^2 + (p_B + p_C)^2}$$

205

# Binary Model



Case 1

$$G_C = \{A, B, C, D\} \quad G_V = \{B, C\}$$

$$G_S = \{A, D\}$$

$$LR = \frac{P_0(\emptyset|A, B, C, D)}{P_1(A, D|A, B, C, D)}$$

$$= \frac{1}{(p_A + p_B + p_C + p_D)^2 - (p_B + p_C + p_D)^2 - (p_A + p_B + p_C)^2 + (p_B + p_C)^2}$$

205

## Binary Model

That was without considering peak heights for assigning the probabilities  $Pr(G_C|S_i)$ .

It is also possible to apply this model while considering peak heights to decide whether to set each  $Pr(G_C|S_i)$  equal to 0 or 1.

This is usually done by comparing the observed peak heights to a set of heuristics.

206

## Binary Model

That was without considering peak heights for assigning the probabilities  $Pr(G_C|S_i)$ .

It is also possible to apply this model while considering peak heights to decide whether to set each  $Pr(G_C|S_i)$  equal to 0 or 1.

This is usually done by comparing the observed peak heights to a set of heuristics.

206

# Binary Model

Heterozygous balance ( $Hb$  or  $P\text{H}r$ ):

$$P\text{H}r = \frac{\text{peak height of higher molecular weight allele}}{\text{peak height of lower molecular weight allele}}$$

For a heterozygous genotype to be possible, it must satisfy:  $0.60 \leq P\text{H}r \leq 1.66$ .

207

# Binary Model

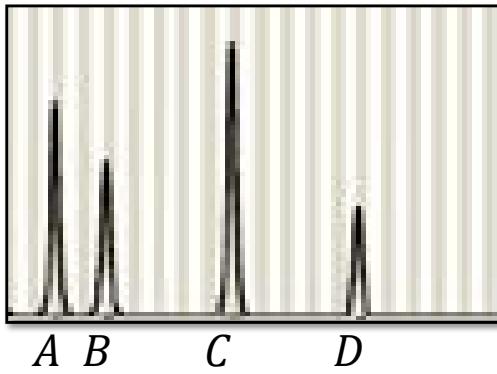
Heterozygous balance ( $Hb$  or  $P\text{H}r$ ):

$$P\text{H}r = \frac{\text{peak height of higher molecular weight allele}}{\text{peak height of lower molecular weight allele}}$$

For a heterozygous genotype to be possible, it must satisfy:  $0.60 \leq P\text{H}r \leq 1.66$ .

207

# Binary Model



Case 3

$$G_C = \{A, B, C, D\} \quad G_{V1} = \{B, C\}$$

$$G_{V2} = \{A, D\}$$

$S_1: BC$  and  $AD$

$S_2: AD$  and  $BC$

$S_3: \cancel{AB}$  and  $\cancel{CD}$

$S_4: \cancel{CD}$  and  $\cancel{AB}$

$S_5: AC$  and  $BD$

$S_6: BD$  and  $AC$

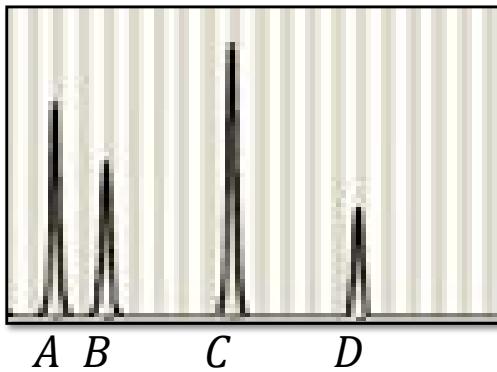
$$\frac{\text{peak height}_D}{\text{peak height}_C} = 0.5$$

$$LR = \frac{1}{4 \times 4 p_A p_B p_C p_D}$$

$$= \frac{1}{16 p_A p_B p_C p_D}$$

208

# Binary Model



Case 3

$$G_C = \{A, B, C, D\} \quad G_{V1} = \{B, C\}$$

$$G_{V2} = \{A, D\}$$

$S_1: BC$  and  $AD$

$S_2: AD$  and  $BC$

$S_3: \cancel{AB}$  and  $\cancel{CD}$

$S_4: \cancel{CD}$  and  $\cancel{AB}$

$S_5: AC$  and  $BD$

$S_6: BD$  and  $AC$

$$\frac{\text{peak height}_D}{\text{peak height}_C} = 0.5$$

$$LR = \frac{1}{4 \times 4 p_A p_B p_C p_D}$$

$$= \frac{1}{16 p_A p_B p_C p_D}$$

208

## **Inbreeding and Relatedness**

209

## **Inbreeding and Relatedness**

209

## Inbreeding

Two alleles that descend from the same allele are said to be **identical by descent (ibd)**.

The probability that two parents transmit ibd alleles to a child is the **inbreeding coefficient F** of the child.

Use small letters for alleles, and capital letters for their particular type: i.e. parents might transmit alleles  $a$  and  $b$  to their child, and these alleles might both be type  $A$ .

210

## Inbreeding

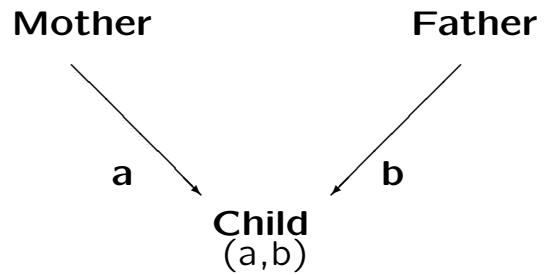
Two alleles that descend from the same allele are said to be **identical by descent (ibd)**.

The probability that two parents transmit ibd alleles to a child is the **inbreeding coefficient F** of the child.

Use small letters for alleles, and capital letters for their particular type: i.e. parents might transmit alleles  $a$  and  $b$  to their child, and these alleles might both be type  $A$ .

210

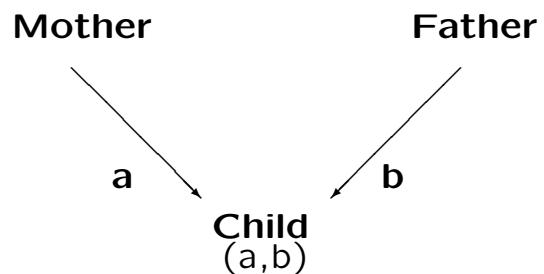
## Inbreeding



$$F_{\text{Child}} = \Pr(a, b \text{ ibd}) = \Pr(a \equiv b)$$

211

## Inbreeding



$$F_{\text{Child}} = \Pr(a, b \text{ ibd}) = \Pr(a \equiv b)$$

211

## Relatedness

Two individuals that have ibd alleles are said to be **related**.

The probability that an allele taken at random from one individual is ibd to an allele taken at random from another individual is the **coancestry coefficient**  $\theta$  of those two individuals.

212

## Relatedness

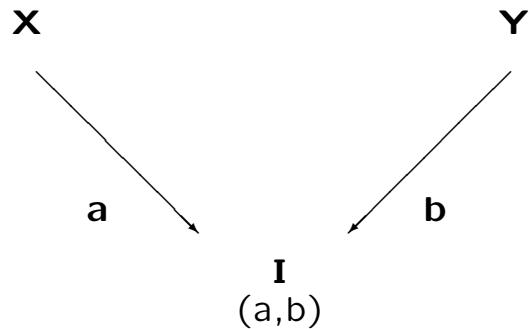
Two individuals that have ibd alleles are said to be **related**.

The probability that an allele taken at random from one individual is ibd to an allele taken at random from another individual is the **coancestry coefficient**  $\theta$  of those two individuals.

212

## Relatedness

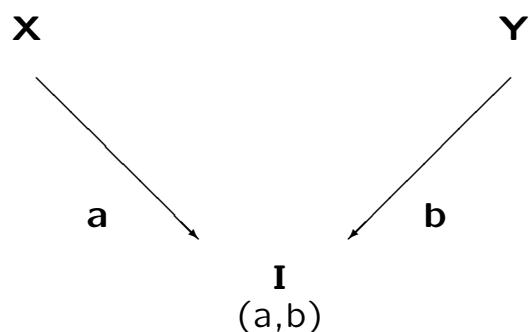
The inbreeding coefficient of an individual is the coancestry of its parents;  $F_I = \theta_{XY}$ .



213

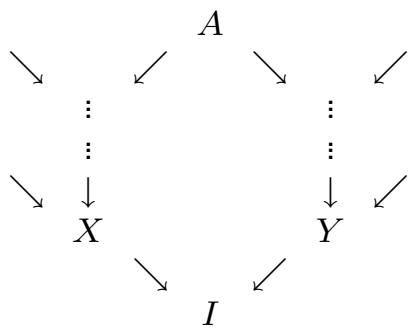
## Relatedness

The inbreeding coefficient of an individual is the coancestry of its parents;  $F_I = \theta_{XY}$ .



213

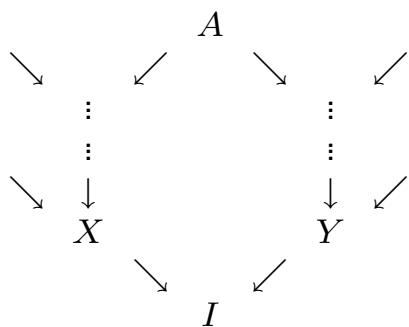
## Path Counting



Identify the path linking the parents of  $I$  to their common ancestor(s).

214

## Path Counting



Identify the path linking the parents of  $I$  to their common ancestor(s).

214

## Path Counting

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

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

If there are several paths to an ancestor, sum over all paths.

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

215

## Path Counting

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

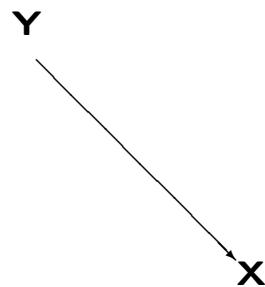
$$F_I = \theta_{XY} = \left(\frac{1}{2}\right)^n (1 + F_A)$$

If there are several paths to an ancestor, sum over all paths.

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

215

## Parent-Child

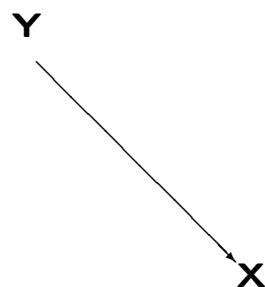


The common ancestor of parent  $X$  and child  $Y$  is  $X$ . The path linking  $X, Y$  to their common ancestor is  $YX$  and this has  $n = 2$  individuals. Therefore

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

216

## Parent-Child

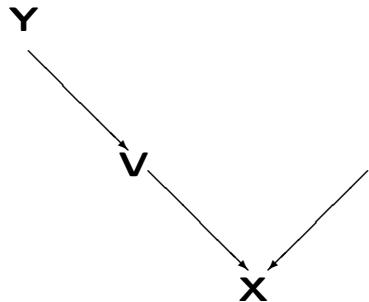


The common ancestor of parent  $X$  and child  $Y$  is  $X$ . The path linking  $X, Y$  to their common ancestor is  $YX$  and this has  $n = 2$  individuals. Therefore

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

216

## Grandparent-grandchild

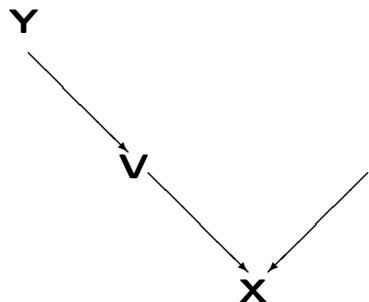


The common ancestor of grandparent  $X$  and grandchild  $Y$  is  $X$ .  
The path linking  $X, Y$  to their common ancestor is  $YVX$  and this  
has  $n = 3$  individuals. Therefore

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

217

## Grandparent-grandchild

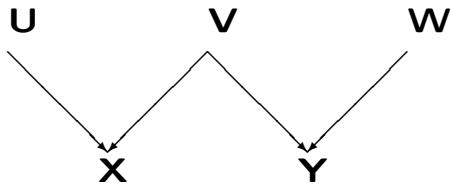


The common ancestor of grandparent  $X$  and grandchild  $Y$  is  $X$ .  
The path linking  $X, Y$  to their common ancestor is  $YVX$  and this  
has  $n = 3$  individuals. Therefore

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

217

## Half sibs

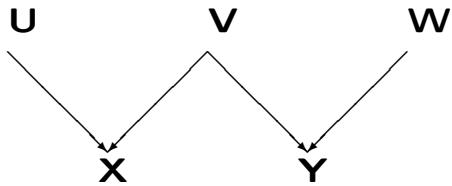


The common ancestor of half sibs  $X$  and  $Y$  is  $V$ . The path linking  $X, Y$  to their common ancestor is  $XVY$  and this has  $n = 3$  individuals. Therefore

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

218

## Half sibs

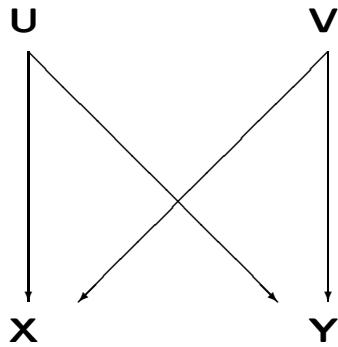


The common ancestor of half sibs  $X$  and  $Y$  is  $V$ . The path linking  $X, Y$  to their common ancestor is  $XVY$  and this has  $n = 3$  individuals. Therefore

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

218

## Full sibs

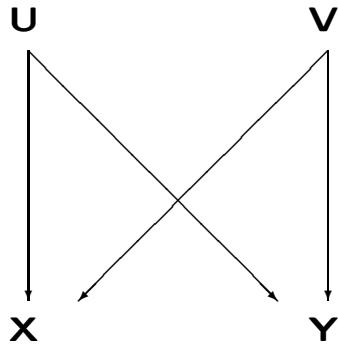


The common ancestors of full sibs  $X$  and  $Y$  are  $U$  and  $V$ . The paths linking  $X, Y$  to their common ancestors are  $XUY$  and  $XVY$  and these each have  $n = 3$  individuals. Therefore

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

219

## Full sibs

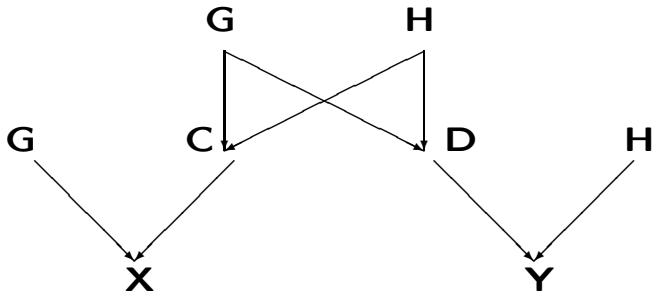


The common ancestors of full sibs  $X$  and  $Y$  are  $U$  and  $V$ . The paths linking  $X, Y$  to their common ancestors are  $XUY$  and  $XVY$  and these each have  $n = 3$  individuals. Therefore

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

219

## First cousins

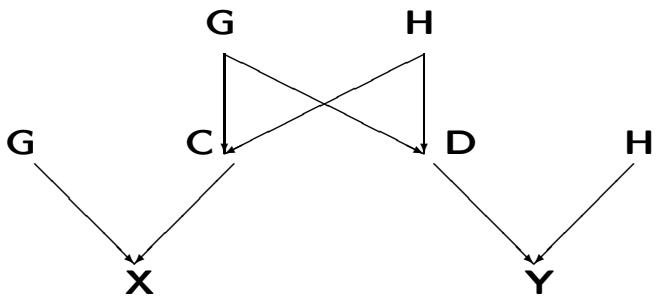


The common ancestors of cousins  $X$  and  $Y$  are  $U$  and  $V$ . The paths linking  $X, Y$  to their common ancestors are  $XCUDY$  and  $XCVDY$  and these each have  $n = 5$  individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^5 = \frac{1}{16}$$

220

## First cousins



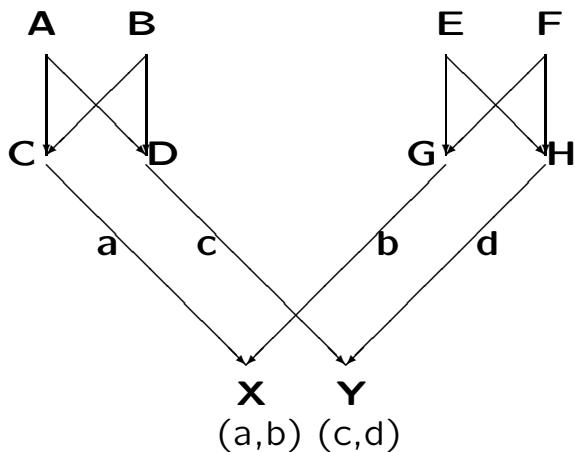
The common ancestors of cousins  $X$  and  $Y$  are  $U$  and  $V$ . The paths linking  $X, Y$  to their common ancestors are  $XCUDY$  and  $XCVDY$  and these each have  $n = 5$  individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^5 = \frac{1}{16}$$

220

## Double First Cousins

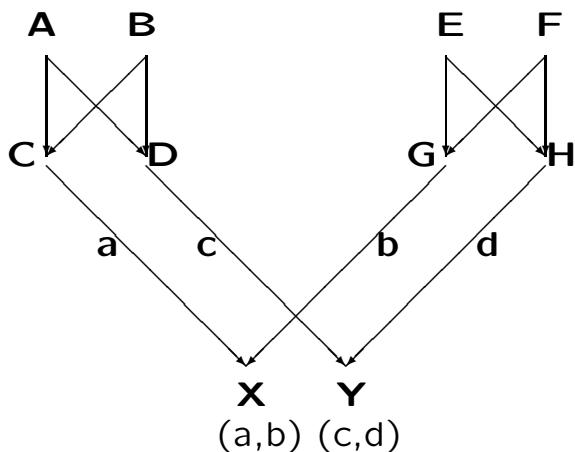
If two brothers  $C, D$  marry two sisters  $G, H$ , their children  $X, Y$  are both maternal and paternal first cousins: i.e. they are double first cousins. What is the coancestry coefficient of double first cousins?



221

## Double First Cousins

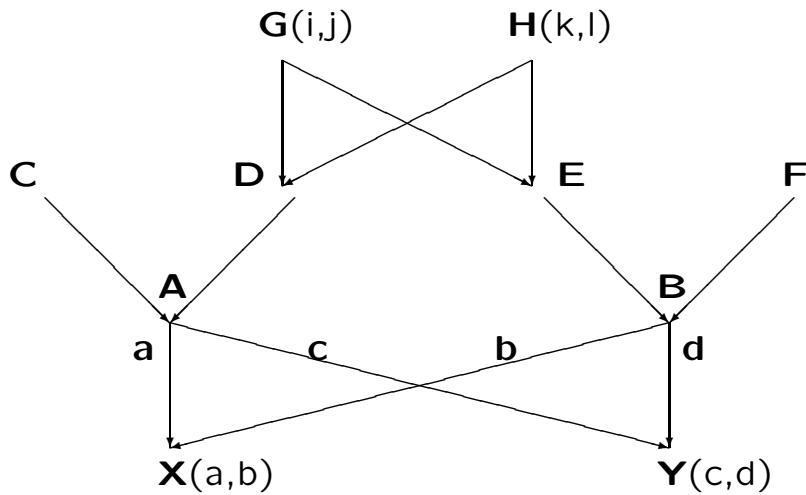
If two brothers  $C, D$  marry two sisters  $G, H$ , their children  $X, Y$  are both maternal and paternal first cousins: i.e. they are double first cousins. What is the coancestry coefficient of double first cousins?



221

## Siblings whose Parents are Cousins

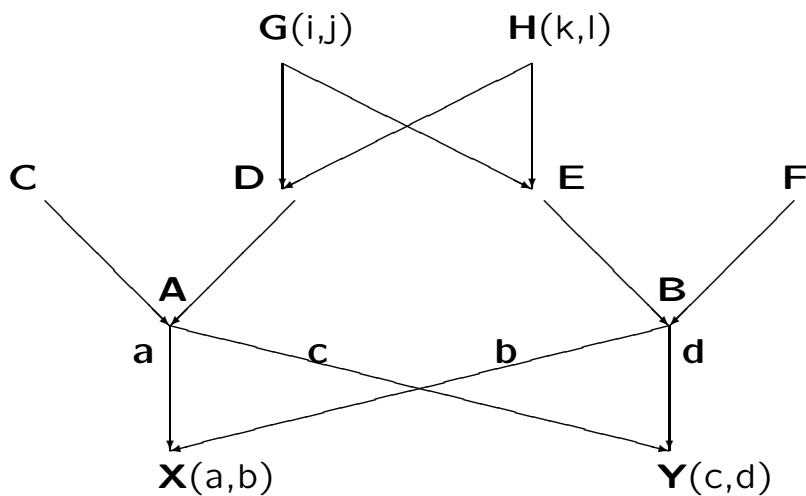
If two first cousins,  $A, B$ , marry and have two children  $X, Y$ , what is the coancestry coefficient of those children?



222

## Siblings whose Parents are Cousins

If two first cousins,  $A, B$ , marry and have two children  $X, Y$ , what is the coancestry coefficient of those children?



222

## Genotype frequencies

Suppose individuals in a population all have inbreeding coefficients  $F$ . The probability an individual has two ibd alleles is  $F$ , and the probability of two non-ibd alleles is  $(1 - F)$ . The probability that any allele is type  $A$  is  $p_A$ , the population allele frequency. So, the probability an individual is homozygous is

$$P_{AA} = F \times p_A + (1 - F) \times p_A^2$$

223

## Genotype frequencies

Suppose individuals in a population all have inbreeding coefficients  $F$ . The probability an individual has two ibd alleles is  $F$ , and the probability of two non-ibd alleles is  $(1 - F)$ . The probability that any allele is type  $A$  is  $p_A$ , the population allele frequency. So, the probability an individual is homozygous is

$$P_{AA} = F \times p_A + (1 - F) \times p_A^2$$

223

## Genotype frequencies

To emphasize the difference from Hardy-Weinberg:

$$P_{AA} = p_A^2 + Fp_A(1 - p_A)$$

Because heterozygous individuals must have non-ibd alleles:

$$\begin{aligned} P_{Aa} &= 2(1 - F)p_Ap_a \\ &= 2p_Ap_a - 2Fp_Ap_a \end{aligned}$$

224

## Genotype frequencies

To emphasize the difference from Hardy-Weinberg:

$$P_{AA} = p_A^2 + Fp_A(1 - p_A)$$

Because heterozygous individuals must have non-ibd alleles:

$$\begin{aligned} P_{Aa} &= 2(1 - F)p_Ap_a \\ &= 2p_Ap_a - 2Fp_Ap_a \end{aligned}$$

224

## First Cousin Example

If every person in the population had parents who were first cousins,  $F = 1/16 = 0.0625$ . For a locus with allele frequencies  $\{p_i\}$  that were all 0.10:

$$\begin{aligned}P_{ii} &= (0.1)^2 + 0.0625(0.1)(0.10) = 0.015625 \\P_{ij} &= 2(0.1)(0.1) - 2(0.0625)(0.1)(0.1) = 0.018750\end{aligned}$$

225

## First Cousin Example

If every person in the population had parents who were first cousins,  $F = 1/16 = 0.0625$ . For a locus with allele frequencies  $\{p_i\}$  that were all 0.10:

$$\begin{aligned}P_{ii} &= (0.1)^2 + 0.0625(0.1)(0.10) = 0.015625 \\P_{ij} &= 2(0.1)(0.1) - 2(0.0625)(0.1)(0.1) = 0.018750\end{aligned}$$

225

## Probabilities of Pairs of Relatives

The inbreeding coefficient  $F$  is a statement about a pair of alleles, and it provides genotypic frequencies – the frequencies of pairs of alleles.

What about pairs of individuals? Their joint genotypic frequencies must involve four-allele analogs of  $F$ .

226

## Probabilities of Pairs of Relatives

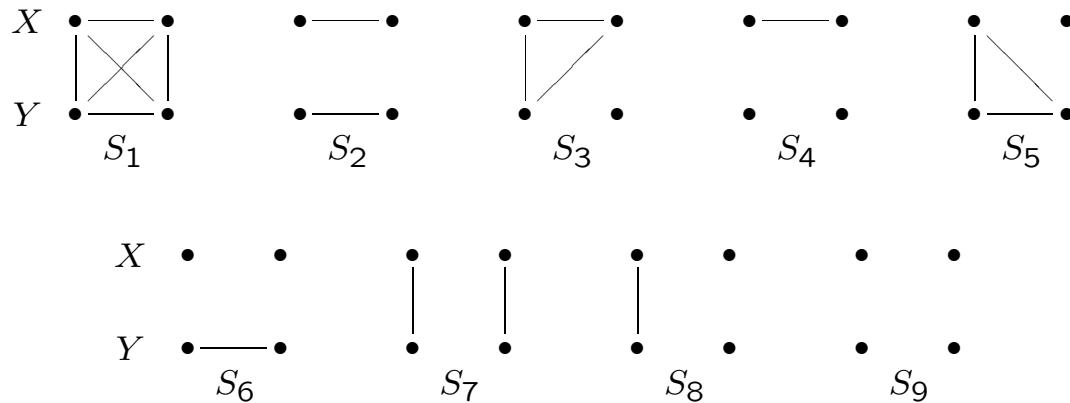
The inbreeding coefficient  $F$  is a statement about a pair of alleles, and it provides genotypic frequencies – the frequencies of pairs of alleles.

What about pairs of individuals? Their joint genotypic frequencies must involve four-allele analogs of  $F$ .

226

## Nine-parameter IBD Set

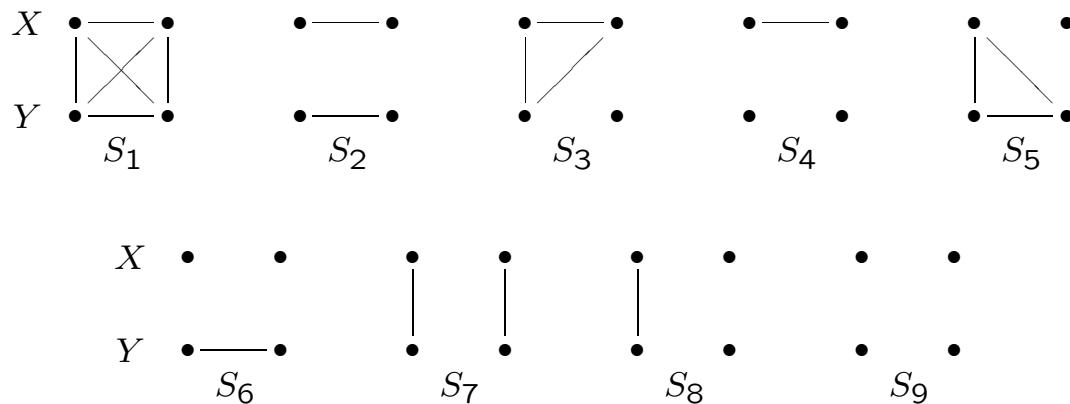
Solid lines join pairs of ibd alleles: top row is the pair of alleles for  $X$ , bottom row the pair of alleles for  $Y$ .



227

## Nine-parameter IBD Set

Solid lines join pairs of ibd alleles: top row is the pair of alleles for  $X$ , bottom row the pair of alleles for  $Y$ .



227

## Non-inbred Relatives

There is a reduction when neither individual is inbred, as then neither  $a, b$  nor  $c, d$  are ibd. There are then only three states and the three probabilities are often written as  $k_2 = \Delta_7, k_1 = \Delta_8$  or  $k_0 = \Delta_9$  to indicate the number of pairs of ibd alleles carried by the two individuals. Examples follow:

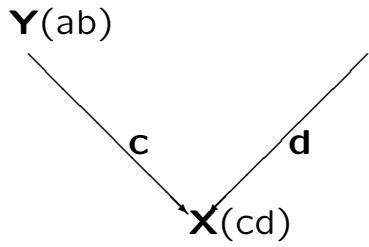
228

## Non-inbred Relatives

There is a reduction when neither individual is inbred, as then neither  $a, b$  nor  $c, d$  are ibd. There are then only three states and the three probabilities are often written as  $k_2 = \Delta_7, k_1 = \Delta_8$  or  $k_0 = \Delta_9$  to indicate the number of pairs of ibd alleles carried by the two individuals. Examples follow:

228

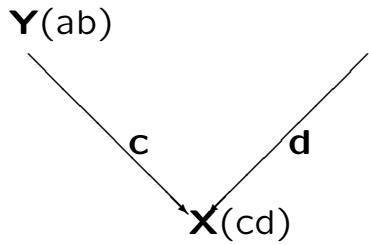
## Parent-Child



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

229

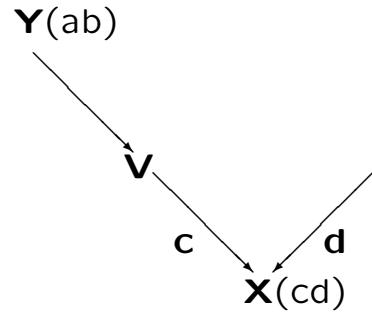
## Parent-Child



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

229

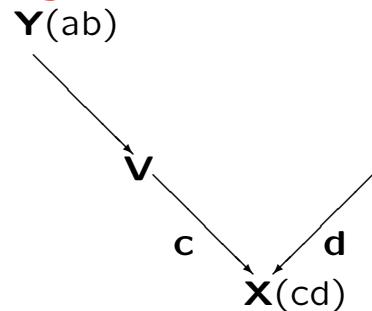
## Grandparent-grandchild



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

230

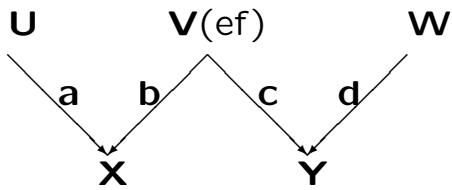
## Grandparent-grandchild



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

230

## Half sibs

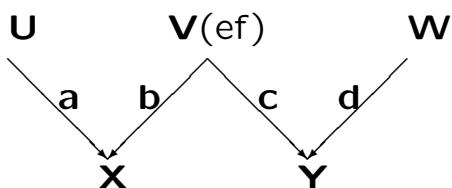


$$\begin{array}{cccc}
 & 0.5 & 0.5 \\
 & c \equiv e & c \equiv f \\
 \hline
 0.5 & b \equiv e & 0.25 & 0.25 \\
 0.5 & b \equiv f & 0.25 & 0.25
 \end{array}$$

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

231

## Half sibs

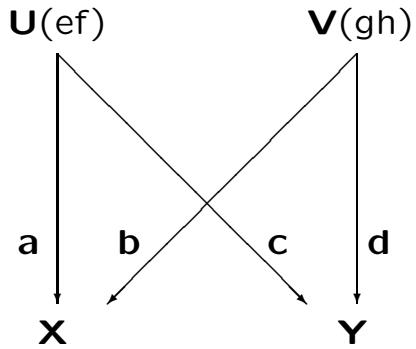


$$\begin{array}{cccc}
 & 0.5 & 0.5 \\
 & c \equiv e & c \equiv f \\
 \hline
 0.5 & b \equiv e & 0.25 & 0.25 \\
 0.5 & b \equiv f & 0.25 & 0.25
 \end{array}$$

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

231

## Full sibs

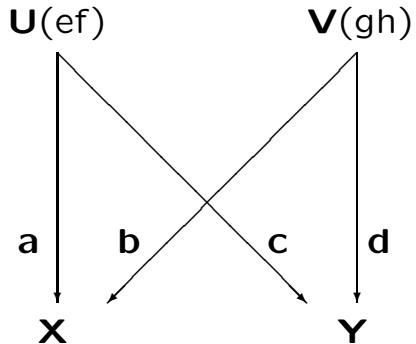


	0.5	0.5	
	$b \equiv d$	$b \not\equiv d$	
0.5	$a \equiv c$	0.25	0.25
0.5	$a \not\equiv c$	0.25	0.25

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

232

## Full sibs

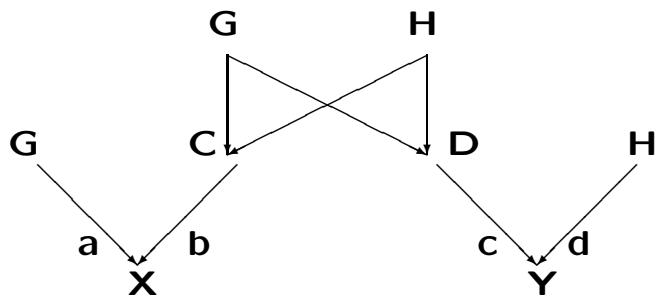


	0.5	0.5	
	$b \equiv d$	$b \not\equiv d$	
0.5	$a \equiv c$	0.25	0.25
0.5	$a \not\equiv c$	0.25	0.25

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

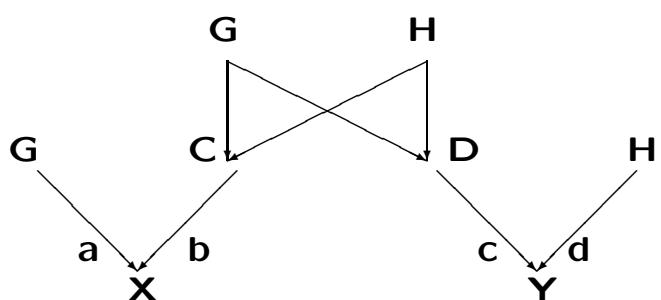
232

## First cousins



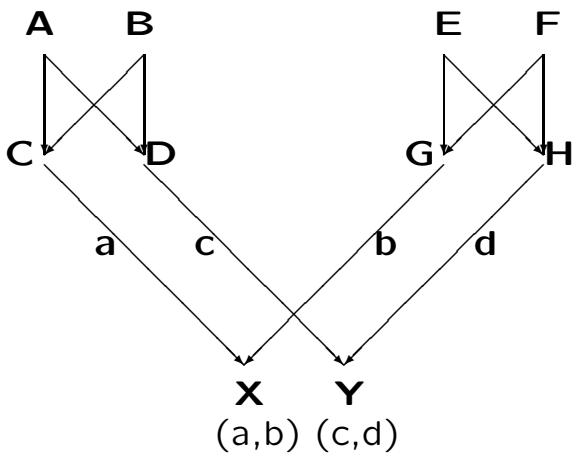
233

## First cousins



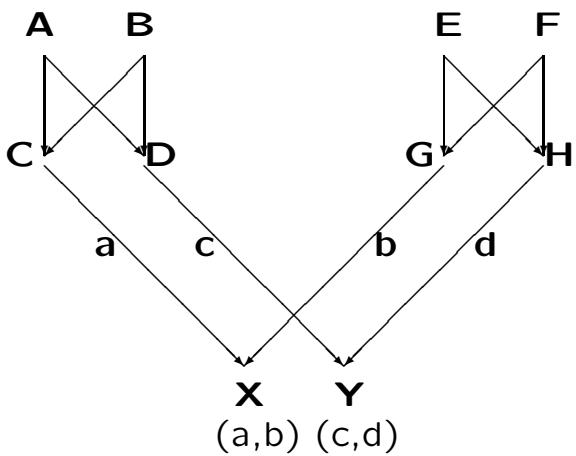
233

## Double First Cousins



234

## Double First Cousins



234

## Non-inbred Relatives

Values of the three probabilities for some common relationships between non-inbred relatives are:

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).

235

## Non-inbred Relatives

Values of the three probabilities for some common relationships between non-inbred relatives are:

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).

235

## Joint genotypic probabilities

For any specific pair of genotypes, the  $\Delta$ 's or  $k$ 's describe the identity-by-descent classes. For two  $A_iA_i$ :

- with probability  $k_2$  there are two pairs of ibd alleles, so two independent  $A_i$  alleles. These  $A_i$  with probability  $p_i^2$ .
- with probability  $k_1$  there is one pair of ibd alleles, so three independent  $A_i$  alleles. These are  $A_i$  with probability  $p_i^3$ .
- with probability  $k_0$  the no ibd alleles, so four independent  $A_i$  alleles. These are all  $A_i$  with probability  $p_i^4$ .

$$\Pr(A_iA_i, A_iA_i) = k_2p_i^2 + k_1p_i^3 + k_0p_i^4$$

236

## Joint genotypic probabilities

For any specific pair of genotypes, the  $\Delta$ 's or  $k$ 's describe the identity-by-descent classes. For two  $A_iA_i$ :

- with probability  $k_2$  there are two pairs of ibd alleles, so two independent  $A_i$  alleles. These  $A_i$  with probability  $p_i^2$ .
- with probability  $k_1$  there is one pair of ibd alleles, so three independent  $A_i$  alleles. These are  $A_i$  with probability  $p_i^3$ .
- with probability  $k_0$  the no ibd alleles, so four independent  $A_i$  alleles. These are all  $A_i$  with probability  $p_i^4$ .

$$\Pr(A_iA_i, A_iA_i) = k_2p_i^2 + k_1p_i^3 + k_0p_i^4$$

236

## Joint genotypic probabilities

Genotypes	General	Non – inbred
$ii, ii$	$\Delta_1 p_i + (\Delta_2 + \Delta_3 + \Delta_5 + \Delta_7) p_i^2 + (\Delta_4 + \Delta_6 + \Delta_8) p_i^3 + \Delta_9 p_i^4$	$k_2 p_i^2 + k_1 p_i^3 + k_0 p_i^4$
$ii, jj$	$\Delta_2 p_i p_j + \Delta_4 p_i p_j^2 + \Delta_6 p_i^2 p_j + \Delta_9 p_i^2 p_j^2$	$k_0 p_i^2 p_j^2$
$ii, ij$	$\Delta_3 p_i p_j + (2\Delta_4 + \Delta_8) p_i^2 p_j + 2\Delta_9 p_i^3 p_j$	$k_1 p_i^2 p_j + 2k_0 p_i^3 p_j$
$ii, jk$	$2\Delta_4 p_i p_j p_k + 2\Delta_9 p_i^2 p_j p_k$	$2k_0 p_i^2 p_j p_k$
$ij, ij$	$2\Delta_7 p_i p_j + \Delta_8 p_i p_j (p_i + p_j) + 4\Delta_9 p_i^2 p_j^2$	$2k_2 p_i p_j + k_1 p_i p_j (p_i + p_j) + 4k_0 p_i^2 p_j^2$
$ij, ik$	$\Delta_8 p_i p_j p_k + 4\Delta_9 p_i^2 p_j p_k$	$k_1 p_i p_j p_k + 4k_0 p_i^2 p_j p_k$
$ij, kl$	$4\Delta_9 p_i p_j p_k p_l$	$4k_0 p_i p_j p_k p_l$

237

## Joint genotypic probabilities

Genotypes	General	Non – inbred
$ii, ii$	$\Delta_1 p_i + (\Delta_2 + \Delta_3 + \Delta_5 + \Delta_7) p_i^2 + (\Delta_4 + \Delta_6 + \Delta_8) p_i^3 + \Delta_9 p_i^4$	$k_2 p_i^2 + k_1 p_i^3 + k_0 p_i^4$
$ii, jj$	$\Delta_2 p_i p_j + \Delta_4 p_i p_j^2 + \Delta_6 p_i^2 p_j + \Delta_9 p_i^2 p_j^2$	$k_0 p_i^2 p_j^2$
$ii, ij$	$\Delta_3 p_i p_j + (2\Delta_4 + \Delta_8) p_i^2 p_j + 2\Delta_9 p_i^3 p_j$	$k_1 p_i^2 p_j + 2k_0 p_i^3 p_j$
$ii, jk$	$2\Delta_4 p_i p_j p_k + 2\Delta_9 p_i^2 p_j p_k$	$2k_0 p_i^2 p_j p_k$
$ij, ij$	$2\Delta_7 p_i p_j + \Delta_8 p_i p_j (p_i + p_j) + 4\Delta_9 p_i^2 p_j^2$	$2k_2 p_i p_j + k_1 p_i p_j (p_i + p_j) + 4k_0 p_i^2 p_j^2$
$ij, ik$	$\Delta_8 p_i p_j p_k + 4\Delta_9 p_i^2 p_j p_k$	$k_1 p_i p_j p_k + 4k_0 p_i^2 p_j p_k$
$ij, kl$	$4\Delta_9 p_i p_j p_k p_l$	$4k_0 p_i p_j p_k p_l$

237

## Example: Non-inbred full sibs

Genotypes	Probability
$ii, ii$	$p_i^2(1 + p_i)^2/4$
$ii, jj$	$p_i^2 p_j^2/4$
$ii, ij$	$p_i p_j(p_i + p_j)/2$
$ii, jk$	$p_i^2 p_j p_k/2$
$ij, ij$	$p_i p_j(1 + p_i + p_j + 2p_i p_j)/2$
$ij, ik$	$p_i p_j p_k(1 + 2p_i)/2$
$ij, kl$	$p_i p_j p_k p_l$

238

## Example: Non-inbred full sibs

Genotypes	Probability
$ii, ii$	$p_i^2(1 + p_i)^2/4$
$ii, jj$	$p_i^2 p_j^2/4$
$ii, ij$	$p_i p_j(p_i + p_j)/2$
$ii, jk$	$p_i^2 p_j p_k/2$
$ij, ij$	$p_i p_j(1 + p_i + p_j + 2p_i p_j)/2$
$ij, ik$	$p_i p_j p_k(1 + 2p_i)/2$
$ij, kl$	$p_i p_j p_k p_l$

238

## **“It was my brother.”**

The defense hypothesis may be that the source of an evidentiary stain was a relative of the defendant. For example

$H_p$ : the defendant is the source of the crime stain.

$H_d$ : (an untyped) brother of the defendant is the source of the crime stain.

239

## **“It was my brother.”**

The defense hypothesis may be that the source of an evidentiary stain was a relative of the defendant. For example

$H_p$ : the defendant is the source of the crime stain.

$H_d$ : (an untyped) brother of the defendant is the source of the crime stain.

239

## “It was my brother.”

If the evidence profile is  $E : AB$  and the defendant has genotype  $G_S : AB$ , then the likelihood ratio is

$$\begin{aligned} LR &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\ &= \frac{1}{\Pr(AB|\text{brother of } AB \text{ person})} \\ &= \frac{1}{\Pr(AB, AB|\text{brothers})/\Pr(AB)} \\ &= \frac{1}{[p_A p_B (1 + p_A + p_B + 2p_A p_B)/2]/(2p_A p_B)} \\ &= \frac{1}{1 + p_A + p_B + 2p_A p_B} \end{aligned}$$

240

## “It was my brother.”

If the evidence profile is  $E : AB$  and the defendant has genotype  $G_S : AB$ , then the likelihood ratio is

$$\begin{aligned} LR &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\ &= \frac{1}{\Pr(AB|\text{brother of } AB \text{ person})} \\ &= \frac{1}{\Pr(AB, AB|\text{brothers})/\Pr(AB)} \\ &= \frac{1}{[p_A p_B (1 + p_A + p_B + 2p_A p_B)/2]/(2p_A p_B)} \\ &= \frac{1}{1 + p_A + p_B + 2p_A p_B} \end{aligned}$$

240

## Are These People Related?

Remains identification often involves the comparison of two profiles and comparing the hypotheses:

$H_1$ : These profiles are from two people with a specific relationship.

$H_2$ : These profiles are from two unrelated people.

If the profiles have genotypes  $ab$  and  $cd$  at a locus, then the likelihood ratio is

$$LR = \frac{\Pr(ab, cd|H_1)}{\Pr(ab, cd|H_2)}$$

241

## Are These People Related?

Remains identification often involves the comparison of two profiles and comparing the hypotheses:

$H_1$ : These profiles are from two people with a specific relationship.

$H_2$ : These profiles are from two unrelated people.

If the profiles have genotypes  $ab$  and  $cd$  at a locus, then the likelihood ratio is

$$LR = \frac{\Pr(ab, cd|H_1)}{\Pr(ab, cd|H_2)}$$

241

## Example: Full sibs vs Unrelated

Suppose two samples  $X, Y$  have genotypes  $AA$  and  $AB$  at a locus.

For

$H_p$ :  $X, Y$  are from full-sibs

$H_d$ :  $X, Y$  are from unrelated individuals

The likelihood ratio is

$$\begin{aligned} LR &= \frac{\Pr(AA, AB|Full\ sibs)}{\Pr(AA, AB|Unrelated)} \\ &= \frac{k_1 p_A^2 p_B + k_0 2p_A^3 p_B | k_1 = 0.5, k_0 = 0.25)}{k_1 p_A^2 p_B + k_0 2p_A^3 p_B | k_1 = 0.0, k_0 = 1.00)} \\ &= \frac{p_A^2 p_B + p_A^3 p_B}{4p_A^3 p_B} = \frac{1 + p_A}{4p_A} \end{aligned}$$

242

## Example: Full sibs vs Unrelated

Suppose two samples  $X, Y$  have genotypes  $AA$  and  $AB$  at a locus.

For

$H_p$ :  $X, Y$  are from full-sibs

$H_d$ :  $X, Y$  are from unrelated individuals

The likelihood ratio is

$$\begin{aligned} LR &= \frac{\Pr(AA, AB|Full\ sibs)}{\Pr(AA, AB|Unrelated)} \\ &= \frac{k_1 p_A^2 p_B + k_0 2p_A^3 p_B | k_1 = 0.5, k_0 = 0.25)}{k_1 p_A^2 p_B + k_0 2p_A^3 p_B | k_1 = 0.0, k_0 = 1.00)} \\ &= \frac{p_A^2 p_B + p_A^3 p_B}{4p_A^3 p_B} = \frac{1 + p_A}{4p_A} \end{aligned}$$

242

## Example: Full sib vs Half sibs

Suppose two samples  $X, Y$  have genotypes  $AB$  and  $AB$  at a locus.

For

$H_p$ :  $X, Y$  are from full-sibs

$H_d$ :  $X, Y$  are from half-sibs

The likelihood ratio is

$$\begin{aligned} LR &= \frac{\Pr(AA, AB|\text{Full sibs})}{\Pr(AA, AB|\text{Half sibs})} \\ &= \frac{2\frac{1}{4}p_A p_B + \frac{1}{2}p_A p_B(p_A + p_B) + 4\frac{1}{4}p_A^2 p_B^2}{\frac{1}{2}p_A p_B(p_A + p_B) + 4\frac{1}{2}p_A^2 p_B^2} \\ &= \frac{p_A p_B + p_A p_B(p_A + p_B) + 2p_A^2 p_B^2}{p_A p_B(p_A + p_B) + 2p_A^2 p_B^2} = \frac{1 + p_A + p_B + 2p_A p_B}{p_A + p_B + 2p_A p_B} \end{aligned}$$

243

## Example: Full sib vs Half sibs

Suppose two samples  $X, Y$  have genotypes  $AB$  and  $AB$  at a locus.

For

$H_p$ :  $X, Y$  are from full-sibs

$H_d$ :  $X, Y$  are from half-sibs

The likelihood ratio is

$$\begin{aligned} LR &= \frac{\Pr(AA, AB|\text{Full sibs})}{\Pr(AA, AB|\text{Half sibs})} \\ &= \frac{2\frac{1}{4}p_A p_B + \frac{1}{2}p_A p_B(p_A + p_B) + 4\frac{1}{4}p_A^2 p_B^2}{\frac{1}{2}p_A p_B(p_A + p_B) + 4\frac{1}{2}p_A^2 p_B^2} \\ &= \frac{p_A p_B + p_A p_B(p_A + p_B) + 2p_A^2 p_B^2}{p_A p_B(p_A + p_B) + 2p_A^2 p_B^2} = \frac{1 + p_A + p_B + 2p_A p_B}{p_A + p_B + 2p_A p_B} \end{aligned}$$

243

## Match Probabilities for Relatives

For relatives, described by  $k_0, k_1, k_2$  in a structured population described by  $\theta$ :

$$\begin{aligned}\Pr(A_u A_u, A_u A_u) &= k_0 \Pr(A_u A_u A_u A_u) + k_1 \Pr(A_u A_u A_u) + k_2 \Pr(A_u A_u) \\ \Pr(A_u A_v, A_u A_v) &= 4k_0 \Pr(A_u A_u A_v A_v) + k_1 [\Pr(A_u A_u A_v) + \Pr(A_u A_v A_v)] \\ &\quad + 2k_2 \Pr(A_u A_v), \quad u \neq v.\end{aligned}$$

The allelic-set probabilities in these equations refer to the generation to which the relatives' most recent common ancestors belong. The match probabilities become

$$\Pr(A_u A_v | A_u A_v) = \begin{cases} k_0 \frac{[2\theta + (1 - \theta)p_u][3\theta + (1 - \theta)p_u]}{(1 + \theta)(1 + 2\theta)} \\ \quad + k_1 \frac{2\theta + (1 - \theta)p_u}{1 + \theta} + k_2, & u = v, \\ k_0 \frac{2[\theta + (1 - \theta)p_u][\theta + (1 - \theta)p_v]}{(1 + \theta)(1 + 2\theta)} \\ \quad + k_1 \frac{2\theta + (1 - \theta)(p_u + p_v)}{2(1 + \theta)} + k_2, & u \neq v \end{cases}$$

Parameters  $p_u$  and  $\theta$  are assumed to have the same value in successive generations.

244

## Match Probabilities for Relatives

For relatives, described by  $k_0, k_1, k_2$  in a structured population described by  $\theta$ :

$$\begin{aligned}\Pr(A_u A_u, A_u A_u) &= k_0 \Pr(A_u A_u A_u A_u) + k_1 \Pr(A_u A_u A_u) + k_2 \Pr(A_u A_u) \\ \Pr(A_u A_v, A_u A_v) &= 4k_0 \Pr(A_u A_u A_v A_v) + k_1 [\Pr(A_u A_u A_v) + \Pr(A_u A_v A_v)] \\ &\quad + 2k_2 \Pr(A_u A_v), \quad u \neq v.\end{aligned}$$

The allelic-set probabilities in these equations refer to the generation to which the relatives' most recent common ancestors belong. The match probabilities become

$$\Pr(A_u A_v | A_u A_v) = \begin{cases} k_0 \frac{[2\theta + (1 - \theta)p_u][3\theta + (1 - \theta)p_u]}{(1 + \theta)(1 + 2\theta)} \\ \quad + k_1 \frac{2\theta + (1 - \theta)p_u}{1 + \theta} + k_2, & u = v, \\ k_0 \frac{2[\theta + (1 - \theta)p_u][\theta + (1 - \theta)p_v]}{(1 + \theta)(1 + 2\theta)} \\ \quad + k_1 \frac{2\theta + (1 - \theta)(p_u + p_v)}{2(1 + \theta)} + k_2, & u \neq v \end{cases}$$

Parameters  $p_u$  and  $\theta$  are assumed to have the same value in successive generations.

244

## Relatedness and Matching

What is the chance that two relatives, with relationship described by  $k_0, k_1, k_2$ , match?

$$\begin{aligned}\Pr(\text{Match}) &= k_2 + k_1 \left[ \sum_i \Pr(A_i A_i A_i) + \sum_i \sum_{j \neq i} \Pr(A_i A_j A_j) \right] + k_0 P_2 \\ &= k_2 + k_1 [\theta + (1 - \theta) S_2] + k_0 P_2 \\ \Pr(\text{Partial Match}) &= k_1 [2 \sum_i \sum_{j \neq i} \Pr(A_i A_i A_j) + \sum_i \sum_{j \neq i} \sum_{k \neq i, j} \Pr(A_i A_j A_k)] \\ &\quad + k_0 P_1 \\ &= k_1 (1 - \theta)(1 - S_2) + k_0 P_1 \\ \Pr(\text{Mismatch}) &= k_0 P_0\end{aligned}$$

where  $P_2, P_1, P_0$  are the match, partial match and mismatch probabilities for unrelated people. Setting  $\theta = 0$  gives the results for unstructured populations.

245

## Relatedness and Matching

What is the chance that two relatives, with relationship described by  $k_0, k_1, k_2$ , match?

$$\begin{aligned}\Pr(\text{Match}) &= k_2 + k_1 \left[ \sum_i \Pr(A_i A_i A_i) + \sum_i \sum_{j \neq i} \Pr(A_i A_j A_j) \right] + k_0 P_2 \\ &= k_2 + k_1 [\theta + (1 - \theta) S_2] + k_0 P_2 \\ \Pr(\text{Partial Match}) &= k_1 [2 \sum_i \sum_{j \neq i} \Pr(A_i A_i A_j) + \sum_i \sum_{j \neq i} \sum_{k \neq i, j} \Pr(A_i A_j A_k)] \\ &\quad + k_0 P_1 \\ &= k_1 (1 - \theta)(1 - S_2) + k_0 P_1 \\ \Pr(\text{Mismatch}) &= k_0 P_0\end{aligned}$$

where  $P_2, P_1, P_0$  are the match, partial match and mismatch probabilities for unrelated people. Setting  $\theta = 0$  gives the results for unstructured populations.

245

## Relatedness and Matching Data

If  $\theta = 0.03$ , using FBI allele frequencies for Caucasians.

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	$2 \times 10^{-14}$	$2 \times 10^{-12}$	$6 \times 10^{-9}$	$5 \times 10^{-6}$

246

## Relatedness and Matching Data

If  $\theta = 0.03$ , using FBI allele frequencies for Caucasians.

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	$2 \times 10^{-14}$	$2 \times 10^{-12}$	$6 \times 10^{-9}$	$5 \times 10^{-6}$

246

## **PARENTAGE TESTING**

247

## **PARENTAGE TESTING**

247

## **Parentage Testing**

Usual situation is that mother, child and alleged father are genotyped. The alleged father is declared “not excluded” if he carries an allele that is inferred to be the child’s paternal allele.

Legal requirement for calculation of a “paternity index” or of “probability of paternity.” Both depend on a likelihood ratio.

248

## **Parentage Testing**

Usual situation is that mother, child and alleged father are genotyped. The alleged father is declared “not excluded” if he carries an allele that is inferred to be the child’s paternal allele.

Legal requirement for calculation of a “paternity index” or of “probability of paternity.” Both depend on a likelihood ratio.

248

## Parentage Testing

Two explanations for genetic evidence  $E$ :

$H_p$ : the alleged father is the father.

$H_d$ : the alleged father is not the father.

249

## Parentage Testing

Two explanations for genetic evidence  $E$ :

$H_p$ : the alleged father is the father.

$H_d$ : the alleged father is not the father.

249

## Paternity Index

From the laws of probability

$$\begin{aligned}\Pr(H_p|E) &= \Pr(H_p, E)/\Pr(E) \\ &= \Pr(E|H_p)\Pr(H_p)/\Pr(E)\end{aligned}$$

$$\begin{aligned}\Pr(H_d|E) &= \Pr(H_d, E)/\Pr(E) \\ &= \Pr(E|H_d)\Pr(H_d)/\Pr(E)\end{aligned}$$

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

250

## Paternity Index

From the laws of probability

$$\begin{aligned}\Pr(H_p|E) &= \Pr(H_p, E)/\Pr(E) \\ &= \Pr(E|H_p)\Pr(H_p)/\Pr(E)\end{aligned}$$

$$\begin{aligned}\Pr(H_d|E) &= \Pr(H_d, E)/\Pr(E) \\ &= \Pr(E|H_d)\Pr(H_d)/\Pr(E)\end{aligned}$$

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

250

## Paternity Index

The paternity index is defined as the ratio

$$\text{P.I.} = \frac{\Pr(E|H_p)}{\Pr(E|H_d)}$$

The genetic evidence is P.I. times more likely if the alleged father is the father than if he is not the father.

251

## Paternity Index

The paternity index is defined as the ratio

$$\text{P.I.} = \frac{\Pr(E|H_p)}{\Pr(E|H_d)}$$

The genetic evidence is P.I. times more likely if the alleged father is the father than if he is not the father.

251

## Probability of Paternity

The probability of paternity and the probability of non-paternity add to 1:

$$\begin{aligned}\pi &= \Pr(H_p|E) = 1 - \Pr(H_d|E) \\ \pi_0 &= \Pr(H_p) = 1 - \Pr(H_d)\end{aligned}$$

where  $\pi_0$  is the “prior probability” of paternity.

252

## Probability of Paternity

The probability of paternity and the probability of non-paternity add to 1:

$$\begin{aligned}\pi &= \Pr(H_p|E) = 1 - \Pr(H_d|E) \\ \pi_0 &= \Pr(H_p) = 1 - \Pr(H_d)\end{aligned}$$

where  $\pi_0$  is the “prior probability” of paternity.

252

## Probability of Paternity

$$\frac{\pi}{1 - \pi} = \text{P.I.} \times \frac{\pi_0}{1 - \pi_0}$$

If the prior probability is set to fifty percent:

$$\frac{\pi}{1 - \pi} = \text{P.I.}$$
$$\pi = \frac{\text{P.I.}}{1 + \text{P.I.}}$$

Assuming a probability of paternity of  $\pi_0$  before genetic testing, the genetic evidence gives  $\pi$  as the probability that the alleged father is the father.

253

## Probability of Paternity

$$\frac{\pi}{1 - \pi} = \text{P.I.} \times \frac{\pi_0}{1 - \pi_0}$$

If the prior probability is set to fifty percent:

$$\frac{\pi}{1 - \pi} = \text{P.I.}$$
$$\pi = \frac{\text{P.I.}}{1 + \text{P.I.}}$$

Assuming a probability of paternity of  $\pi_0$  before genetic testing, the genetic evidence gives  $\pi$  as the probability that the alleged father is the father.

253

## **Effect of prior probabilities**

The assumption of 50% prior probability is difficult to defend. Better to show how large paternity index values can overcome even small prior probabilities. Construct a table showing effect of different prior probabilities with a specific paternity index.

254

## **Effect of prior probabilities**

The assumption of 50% prior probability is difficult to defend. Better to show how large paternity index values can overcome even small prior probabilities. Construct a table showing effect of different prior probabilities with a specific paternity index.

254

## Effect of prior probabilities

$\pi_0$	P.I.			
	1	10	100	1,000
0	0	0	0	0
.001	.001	.00991	.09099	.5002501
.010	.010	.09174	.50251	.9099181
.100	.100	.52631	.91743	.9910803
.500	.500	.90909	.99009	.9990010
.900	.900	.98901	.99889	.9998889
.990	.990	.99899	.99989	.9999899
.999	.999	.99989	.99999	.9999990
1	1	1	1	1

255

## Effect of prior probabilities

$\pi_0$	P.I.			
	1	10	100	1,000
0	0	0	0	0
.001	.001	.00991	.09099	.5002501
.010	.010	.09174	.50251	.9099181
.100	.100	.52631	.91743	.9910803
.500	.500	.90909	.99009	.9990010
.900	.900	.98901	.99889	.9998889
.990	.990	.99899	.99989	.9999899
.999	.999	.99989	.99999	.9999990
1	1	1	1	1

255

## Evaluation of probabilities

The P.I. can be expressed in terms of probability of genotype  $G_C$  of child, conditional on genotypes  $G_M, G_{AF}$  of mother and alleged father:

$$\begin{aligned} \text{P.I.} &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\ &= \frac{\Pr(G_C|G_M, G_{AF}, H_p) \Pr(G_M, G_{AF}|H_p)}{\Pr(G_C|G_M, G_{AF}, H_d) \Pr(G_M, G_{AF}|H_d)} \\ &= \frac{\Pr(G_C|G_M, G_{AF}, H_p)}{\Pr(G_C|G_M, G_{AF}, H_d)} \end{aligned}$$

256

## Evaluation of probabilities

The P.I. can be expressed in terms of probability of genotype  $G_C$  of child, conditional on genotypes  $G_M, G_{AF}$  of mother and alleged father:

$$\begin{aligned} \text{P.I.} &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\ &= \frac{\Pr(G_C|G_M, G_{AF}, H_p) \Pr(G_M, G_{AF}|H_p)}{\Pr(G_C|G_M, G_{AF}, H_d) \Pr(G_M, G_{AF}|H_d)} \\ &= \frac{\Pr(G_C|G_M, G_{AF}, H_p)}{\Pr(G_C|G_M, G_{AF}, H_d)} \end{aligned}$$

256

## Evaluation of probabilities

It is more convenient to work with the child's maternal and paternal alleles  $A_M, A_P$ , provided the mother and alleged father are not related (so that  $A_M, A_P$  are independent):

$$\Pr(A_M A_P | G_M, G_{AF}, H) = \Pr(A_M | G_M) \times \Pr(A_P | G_{AF}, H)$$

because  $\Pr(A_M | G_M)$  does not depend on  $H$ .

257

## Evaluation of probabilities

It is more convenient to work with the child's maternal and paternal alleles  $A_M, A_P$ , provided the mother and alleged father are not related (so that  $A_M, A_P$  are independent):

$$\Pr(A_M A_P | G_M, G_{AF}, H) = \Pr(A_M | G_M) \times \Pr(A_P | G_{AF}, H)$$

because  $\Pr(A_M | G_M)$  does not depend on  $H$ .

257

## Numerator Probabilities

Under  $H_p$  or  $H_d$ :

$$\begin{aligned}\Pr(A_M = A_i | G_M = A_i A_i) &= 1 \\ \Pr(A_M = A_i | G_M = A_i A_j) &= 0.5, \quad j \neq i\end{aligned}$$

Under explanation  $H_p$ , the alleged father has provided the paternal allele:

$$\Pr(A_P = A_i | G_{AF} = A_i A_i, H_p) = 1$$

$$\Pr(A_P = A_i | G_{AF} = A_i A_j, H_p) = 0.5, \quad j \neq i$$

$$\Pr(A_P = A_i | G_{AF} = A_j A_k, H_p) = 0, \quad j, k \neq i$$

258

## Numerator Probabilities

Under  $H_p$  or  $H_d$ :

$$\begin{aligned}\Pr(A_M = A_i | G_M = A_i A_i) &= 1 \\ \Pr(A_M = A_i | G_M = A_i A_j) &= 0.5, \quad j \neq i\end{aligned}$$

Under explanation  $H_p$ , the alleged father has provided the paternal allele:

$$\Pr(A_P = A_i | G_{AF} = A_i A_i, H_p) = 1$$

$$\Pr(A_P = A_i | G_{AF} = A_i A_j, H_p) = 0.5, \quad j \neq i$$

$$\Pr(A_P = A_i | G_{AF} = A_j A_k, H_p) = 0, \quad j, k \neq i$$

258

## Denominator Probabilities

Under explanation  $H_d$  some other man  $TF$  has provided the paternal allele:

$$\Pr(A_P|G_{AF}, H_d) = \Pr(A_P|H_d)$$

if the genotype of the alleged father has no effect on the paternal allele. The RHS is just the population proportion of allele  $A_P$ .

Paternal allele	Alleged father	Paternity index	When $p_i = 0.1$
$A_i$	$A_iA_i$	$\frac{1}{p_i}$	10
	$A_iA_j$	$\frac{1}{2p_i}$	5

259

## Denominator Probabilities

Under explanation  $H_d$  some other man  $TF$  has provided the paternal allele:

$$\Pr(A_P|G_{AF}, H_d) = \Pr(A_P|H_d)$$

if the genotype of the alleged father has no effect on the paternal allele. The RHS is just the population proportion of allele  $A_P$ .

Paternal allele	Alleged father	Paternity index	When $p_i = 0.1$
$A_i$	$A_iA_i$	$\frac{1}{p_i}$	10
	$A_iA_j$	$\frac{1}{2p_i}$	5

259

## **Alleged father related to father**

More generally, need to allow the paternal allele  $A_P$  to depend on the alleles in the alleged father. This is obviously the case if the two men are related. Need

$$\Pr(A_P|G_{AF}) = \frac{\Pr(A_P, G_{AF})}{\Pr(G_{AF})}$$

The numerator of the RHS requires information about three alleles (need 3-allele ibd probabilities), and the denominator is the proportion of genotype  $G_{AF}$  in the population.

260

## **Alleged father related to father**

More generally, need to allow the paternal allele  $A_P$  to depend on the alleles in the alleged father. This is obviously the case if the two men are related. Need

$$\Pr(A_P|G_{AF}) = \frac{\Pr(A_P, G_{AF})}{\Pr(G_{AF})}$$

The numerator of the RHS requires information about three alleles (need 3-allele ibd probabilities), and the denominator is the proportion of genotype  $G_{AF}$  in the population.

260

## Relatives in H-W populations

When the population is in Hardy-Weinberg equilibrium, and the alleged father is not inbred:

$$\Pr(A_i A_i, A_i) = p_i^2 [2\theta_{AT} + (1 - 2\theta_{AT})p_i]$$

$$\Pr(A_i A_i, A_j) = p_i^2 p_j (1 - 2\theta_{AT})$$

$$\Pr(A_i A_j, A_i) = 2p_i p_j [\theta_{AT} + (1 - 2\theta_{AT})p_i]$$

$$\Pr(A_i A_j, A_k) = 2p_i p_j p_k (1 - 2\theta_{AT})$$

so that only the coancestry coefficient  $\theta_{AT}$  of the two men is needed.

261

## Relatives in H-W populations

When the population is in Hardy-Weinberg equilibrium, and the alleged father is not inbred:

$$\Pr(A_i A_i, A_i) = p_i^2 [2\theta_{AT} + (1 - 2\theta_{AT})p_i]$$

$$\Pr(A_i A_i, A_j) = p_i^2 p_j (1 - 2\theta_{AT})$$

$$\Pr(A_i A_j, A_i) = 2p_i p_j [\theta_{AT} + (1 - 2\theta_{AT})p_i]$$

$$\Pr(A_i A_j, A_k) = 2p_i p_j p_k (1 - 2\theta_{AT})$$

so that only the coancestry coefficient  $\theta_{AT}$  of the two men is needed.

261

## “It was my brother”

If the alleged father and the true father are brothers,  $\theta_{AT} = 1/4$ ,

$$\begin{aligned}\Pr(A_i A_i, A_i) &= p_i^2(1 + p_i)/2 \\ \Pr(A_i A_i, A_j) &= p_i^2 p_j/2 \\ \Pr(A_i A_j, A_i) &= p_i p_j(1 + 2p_i)/2 \\ \Pr(A_i A_j, A_k) &= p_i p_j p_k\end{aligned}$$

so that

$$\begin{aligned}\Pr(A_i | A_i A_i) &= (1 + p_i)/2 \\ \Pr(A_j | A_i A_i) &= p_j/2 \\ \Pr(A_i | A_i A_j) &= (1 + 2p_i)/4 \\ \Pr(A_k | A_i A_j) &= p_k/2\end{aligned}$$

262

## “It was my brother”

If the alleged father and the true father are brothers,  $\theta_{AT} = 1/4$ ,

$$\begin{aligned}\Pr(A_i A_i, A_i) &= p_i^2(1 + p_i)/2 \\ \Pr(A_i A_i, A_j) &= p_i^2 p_j/2 \\ \Pr(A_i A_j, A_i) &= p_i p_j(1 + 2p_i)/2 \\ \Pr(A_i A_j, A_k) &= p_i p_j p_k\end{aligned}$$

so that

$$\begin{aligned}\Pr(A_i | A_i A_i) &= (1 + p_i)/2 \\ \Pr(A_j | A_i A_i) &= p_j/2 \\ \Pr(A_i | A_i A_j) &= (1 + 2p_i)/4 \\ \Pr(A_k | A_i A_j) &= p_k/2\end{aligned}$$

262

## **Alleged father a brother of the true father**

Paternity index for a non-inbred alleged father versus an untested brother:

Paternal allele	Alleged father	Paternity Index	When $p_i = 0.1$
$A_i$	$A_iA_i$	$\frac{2}{1+p_i}$	1.8
	$A_iA_j$	$\frac{2}{1+2p_i}$	1.7

"The genetic evidence is P.I. times more likely if the alleged father is the father of this child than if an untested brother of his is the father."

263

## **Alleged father a brother of the true father**

Paternity index for a non-inbred alleged father versus an untested brother:

Paternal allele	Alleged father	Paternity Index	When $p_i = 0.1$
$A_i$	$A_iA_i$	$\frac{2}{1+p_i}$	1.8
	$A_iA_j$	$\frac{2}{1+2p_i}$	1.7

"The genetic evidence is P.I. times more likely if the alleged father is the father of this child than if an untested brother of his is the father."

263

## Any pair of non-inbred relatives

General expression for a non-inbred alleged father  $AF$  versus an untested relative  $TF$ :

Paternal allele	Alleged father	Paternity Index
$A_i$	$A_i A_i$	$\frac{1}{2\theta_{AT} + (1 - 2\theta_{AT})p_i}$
	$A_i A_j$	$\frac{1}{2[\theta_{AT} + (1 - 2\theta_{AT})p_i]}$

For cousins,  $\theta_{AT} = 1/16$ . For uncle-nephew,  $\theta_{AT} = 1/8$  and for father-son,  $\theta_{AT} = 1/4$ .

264

## Any pair of non-inbred relatives

General expression for a non-inbred alleged father  $AF$  versus an untested relative  $TF$ :

Paternal allele	Alleged father	Paternity Index
$A_i$	$A_i A_i$	$\frac{1}{2\theta_{AT} + (1 - 2\theta_{AT})p_i}$
	$A_i A_j$	$\frac{1}{2[\theta_{AT} + (1 - 2\theta_{AT})p_i]}$

For cousins,  $\theta_{AT} = 1/16$ . For uncle-nephew,  $\theta_{AT} = 1/8$  and for father-son,  $\theta_{AT} = 1/4$ .

264

## Non-HW population

For populations in which there is a (low) level of relatedness of individuals, need to consider the relatedness of mother, alleged father and father.

This does not affect  $\Pr(G_C|G_M, G_{AF}, H_p)$ .

Suppose maternal allele  $A_M$  and paternal allele  $A_P$  can be determined from  $G_M, G_{AF}$ . Then, under  $H_d$ :

$$\begin{aligned}\Pr(G_C|G_M, G_{AF}) &= \Pr(A_M A_P|G_M, G_{AF}) \\ &= \Pr(A_M|A_P G_M G_{AF}) \times \Pr(A_P|G_M G_{AF}) \\ &= \Pr(A_M|G_M) \Pr(A_P|G_M, G_{AF})\end{aligned}$$

265

## Non-HW population

For populations in which there is a (low) level of relatedness of individuals, need to consider the relatedness of mother, alleged father and father.

This does not affect  $\Pr(G_C|G_M, G_{AF}, H_p)$ .

Suppose maternal allele  $A_M$  and paternal allele  $A_P$  can be determined from  $G_M, G_{AF}$ . Then, under  $H_d$ :

$$\begin{aligned}\Pr(G_C|G_M, G_{AF}) &= \Pr(A_M A_P|G_M, G_{AF}) \\ &= \Pr(A_M|A_P G_M G_{AF}) \times \Pr(A_P|G_M G_{AF}) \\ &= \Pr(A_M|G_M) \Pr(A_P|G_M, G_{AF})\end{aligned}$$

265

## Non-HW population

The quantity  $\Pr(A_M|G_M)$  is still 0, 1/2 or 1. Need the probability of  $A_P$  being a random allele in a man (*TF*) in the population given the genotypes of the mother and the alleged father.

$$\Pr(A_P|G_M, G_{AF}, H_d) = \frac{\Pr(A_P, G_M, G_{AF})}{\Pr(G_M, G_{AF})}$$

The numerator of the RHS requires the relationships among the paternal allele and the two alleles in each of the mother and the alleged father. The denominator requires the relationship among the two in each of the mother and the alleged father.

266

## Non-HW population

The quantity  $\Pr(A_M|G_M)$  is still 0, 1/2 or 1. Need the probability of  $A_P$  being a random allele in a man (*TF*) in the population given the genotypes of the mother and the alleged father.

$$\Pr(A_P|G_M, G_{AF}, H_d) = \frac{\Pr(A_P, G_M, G_{AF})}{\Pr(G_M, G_{AF})}$$

The numerator of the RHS requires the relationships among the paternal allele and the two alleles in each of the mother and the alleged father. The denominator requires the relationship among the two in each of the mother and the alleged father.

266

## Dirichlet distribution

If a random mating population can be assumed to have reached an evolutionary equilibrium, then allele proportions satisfy a Dirichlet distribution, and simple expressions are available for the joint probabilities of sets of alleles.

All results follow from a “sampling formula” which gives the probability of an  $A$  allele if a set of  $n$  alleles has already been seen to contain  $n_A$  of this allele:

$$\Pr(A|n_A, n) = \frac{n_a\theta + (1 - \theta)p_A}{1 + (n - 1)\theta}$$

267

## Dirichlet distribution

If a random mating population can be assumed to have reached an evolutionary equilibrium, then allele proportions satisfy a Dirichlet distribution, and simple expressions are available for the joint probabilities of sets of alleles.

All results follow from a “sampling formula” which gives the probability of an  $A$  allele if a set of  $n$  alleles has already been seen to contain  $n_A$  of this allele:

$$\Pr(A|n_A, n) = \frac{n_a\theta + (1 - \theta)p_A}{1 + (n - 1)\theta}$$

267

## Dirichlet distribution

This sampling result provides:

$$\begin{aligned}\Pr(A) &= p_A \\ \Pr(A|A) &= \theta + (1 - \theta)p_A \\ \Pr(A|AA) &= [2\theta + (1 - \theta)p_A]/(1 + \theta) \\ \Pr(A|AAA) &= [3\theta + (1 - \theta)p_A]/(1 + 2\theta)\end{aligned}$$

and, applying the third law of probability

$$\begin{aligned}\Pr(AA) &= p_A \Pr(A|A) = p_A[\theta + (1 - \theta)p_A] \\ \Pr(AAA) &= \Pr(AA) \Pr(A|AA) \\ &= \frac{p_A[\theta + (1 - \theta)p_A][2\theta + (1 - \theta)p_A]}{1 + \theta} \\ \Pr(AAAA) &= \Pr(AAA) \Pr(A|AAA) \\ &= \frac{p_A[\theta + (1 - \theta)p_A][2\theta + (1 - \theta)p_A][3\theta + (1 - \theta)p_A]}{(1 + \theta)(1 + 2\theta)}\end{aligned}$$

268

## Dirichlet distribution

This sampling result provides:

$$\begin{aligned}\Pr(A) &= p_A \\ \Pr(A|A) &= \theta + (1 - \theta)p_A \\ \Pr(A|AA) &= [2\theta + (1 - \theta)p_A]/(1 + \theta) \\ \Pr(A|AAA) &= [3\theta + (1 - \theta)p_A]/(1 + 2\theta)\end{aligned}$$

and, applying the third law of probability

$$\begin{aligned}\Pr(AA) &= p_A \Pr(A|A) = p_A[\theta + (1 - \theta)p_A] \\ \Pr(AAA) &= \Pr(AA) \Pr(A|AA) \\ &= \frac{p_A[\theta + (1 - \theta)p_A][2\theta + (1 - \theta)p_A]}{1 + \theta} \\ \Pr(AAAA) &= \Pr(AAA) \Pr(A|AAA) \\ &= \frac{p_A[\theta + (1 - \theta)p_A][2\theta + (1 - \theta)p_A][3\theta + (1 - \theta)p_A]}{(1 + \theta)(1 + 2\theta)}\end{aligned}$$

268

## Match probability for homozygotes

For two allele  $A$ :

$$\Pr(AA) = \Pr(A) \Pr(A|A) = p_A[\theta + (1 - \theta)p_A]$$

$$\Pr(AAAA) = \Pr(A) \Pr(A|A) \Pr(A|AA) \Pr(A|AAA)$$

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

so

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

269

## Match probability for homozygotes

For two allele  $A$ :

$$\Pr(AA) = \Pr(A) \Pr(A|A) = p_A[\theta + (1 - \theta)p_A]$$

$$\Pr(AAAA) = \Pr(A) \Pr(A|A) \Pr(A|AA) \Pr(A|AAA)$$

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

so

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

269

## Match probability for heterozygotes

For two different alleles,  $A, B$ :

$$\Pr(AB) = \Pr(A) \Pr(B|A) = p_A[(1 - \theta)p_B]$$

$$\begin{aligned}\Pr(ABAB) &= \Pr(A) \Pr(B|A) \Pr(A|AB) \Pr(B|ABA) \\ &= p_A[(1 - \theta)p_B] \frac{[\theta + (1 - \theta)p_A]}{1 + \theta} \frac{[\theta + (1 - \theta)p_B]}{1 + 2\theta}\end{aligned}$$

so, with a “2” to account for heterozygotes,

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

270

## Match probability for heterozygotes

For two different alleles,  $A, B$ :

$$\Pr(AB) = \Pr(A) \Pr(B|A) = p_A[(1 - \theta)p_B]$$

$$\begin{aligned}\Pr(ABAB) &= \Pr(A) \Pr(B|A) \Pr(A|AB) \Pr(B|ABA) \\ &= p_A[(1 - \theta)p_B] \frac{[\theta + (1 - \theta)p_A]}{1 + \theta} \frac{[\theta + (1 - \theta)p_B]}{1 + 2\theta}\end{aligned}$$

so, with a “2” to account for heterozygotes,

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

270

## Dirichlet Assumption for Paternity Index

If the mother is  $AA$  and her child is  $AB$  the paternal allele is  $B$ . If the alleged father is  $AB$  the numerator of the paternity index is the probability of mother and alleged father multiplied by 0.5 for the choice of paternal allele: this is  $\Pr(AAAB)$  and multiplied by 2 for the heterozygous alleged father. The denominator is the probability of mother, alleged father and paternal allele:  $AAABB$  multiplied by 2 for the heterozygous alleged father:

$$\begin{aligned} \text{PI} &= \frac{\Pr(AAAB)}{\Pr(AAABB)} \\ &= \frac{p_A[\theta+(1-\theta)p_A][2\theta+(1-\theta)p_A](1-\theta)p_B}{(1-\theta)(1)(1+\theta)(1+2\theta)} \\ &= \frac{2p_A[\theta+(1-\theta)p_A][2\theta+(1-\theta)p_A][(1-\theta)p_B][\theta+(1-\theta)p_B]}{(1-\theta)(1)(1+\theta)(1+2\theta)(1+3\theta)} \\ &= \frac{1+3\theta}{2[\theta+(1-\theta)p_B]} \end{aligned}$$

271

## Dirichlet Assumption for Paternity Index

If the mother is  $AA$  and her child is  $AB$  the paternal allele is  $B$ . If the alleged father is  $AB$  the numerator of the paternity index is the probability of mother and alleged father multiplied by 0.5 for the choice of paternal allele: this is  $\Pr(AAAB)$  and multiplied by 2 for the heterozygous alleged father. The denominator is the probability of mother, alleged father and paternal allele:  $AAABB$  multiplied by 2 for the heterozygous alleged father:

$$\begin{aligned} \text{PI} &= \frac{\Pr(AAAB)}{\Pr(AAABB)} \\ &= \frac{p_A[\theta+(1-\theta)p_A][2\theta+(1-\theta)p_A](1-\theta)p_B}{(1-\theta)(1)(1+\theta)(1+2\theta)} \\ &= \frac{2p_A[\theta+(1-\theta)p_A][2\theta+(1-\theta)p_A][(1-\theta)p_B][\theta+(1-\theta)p_B]}{(1-\theta)(1)(1+\theta)(1+2\theta)(1+3\theta)} \\ &= \frac{1+3\theta}{2[\theta+(1-\theta)p_B]} \end{aligned}$$

271

## P.I. for homozygous mother

$G_M$	$G_C$	$A_M$	$A_P$	$G_{AF}$	P.I.	$\theta = 0.03$
$A_i A_i$	$A_i A_i$	$A_i$	$A_i$	$A_i A_i$	$\frac{1+3\theta}{4\theta+(1-\theta)p_i}$	$p_i = 0.1$
				$A_i A_j$	$\frac{1+3\theta}{2[3\theta+(1-\theta)p_i]}$	3.0
				$A_i A_j$	$\frac{1+3\theta}{2\theta+(1-\theta)p_j}$	6.6
				$A_i A_j$	$\frac{1+3\theta}{2[\theta+(1-\theta)p_j]}$	4.5
				$A_j A_k$	$\frac{1+3\theta}{2[\theta+(1-\theta)p_j]}$	4.5

272

## P.I. for homozygous mother

$G_M$	$G_C$	$A_M$	$A_P$	$G_{AF}$	P.I.	$\theta = 0.03$
$A_i A_i$	$A_i A_i$	$A_i$	$A_i$	$A_i A_i$	$\frac{1+3\theta}{4\theta+(1-\theta)p_i}$	$p_i = 0.1$
				$A_i A_j$	$\frac{1+3\theta}{2[3\theta+(1-\theta)p_i]}$	3.0
$A_i A_j$	$A_i$	$A_j$	$A_j A_j$		$\frac{1+3\theta}{2\theta+(1-\theta)p_j}$	6.6
				$A_i A_j$	$\frac{1+3\theta}{2[\theta+(1-\theta)p_j]}$	4.5
				$A_j A_k$	$\frac{1+3\theta}{2[\theta+(1-\theta)p_j]}$	4.5

272

## P.I. for heterozygous mother

$G_M$	$G_C$	$A_M$	$A_P$	$G_{AF}$	P.I.	$\theta = 0.03$	$p_i = 0.1$
$A_iA_k$	$A_iA_i$	$A_i$	$A_i$	$A_iA_i$	$\frac{1+3\theta}{3\theta+(1-\theta)p_i}$		6.0
				$A_iA_k$	$\frac{1+3\theta}{2[2\theta+(1-\theta)p_i]}$	3.6	
				$A_iA_j$	$A_jA_j$	$\frac{1+3\theta}{2\theta+(1-\theta)p_j}$	6.6
				$A_iA_j$		$\frac{1+3\theta}{2[\theta+(1-\theta)p_j]}$	4.5
				$A_jA_l$		$\frac{1+3\theta}{2[\theta+(1-\theta)p_j]}$	4.5

273

## P.I. for heterozygous mother

$G_M$	$G_C$	$A_M$	$A_P$	$G_{AF}$	P.I.	$\theta = 0.03$	$p_i = 0.1$
$A_iA_k$	$A_iA_i$	$A_i$	$A_i$	$A_iA_i$	$\frac{1+3\theta}{3\theta+(1-\theta)p_i}$		6.0
				$A_iA_k$	$\frac{1+3\theta}{2[2\theta+(1-\theta)p_i]}$	3.6	
				$A_iA_j$	$A_jA_j$	$\frac{1+3\theta}{2\theta+(1-\theta)p_j}$	6.6
				$A_iA_j$		$\frac{1+3\theta}{2[\theta+(1-\theta)p_j]}$	4.5
				$A_jA_l$		$\frac{1+3\theta}{2[\theta+(1-\theta)p_j]}$	4.5

273



# **INTERPRETATION OF MIXTURES**

## **SEMI-CONTINUOUS MODEL**

274



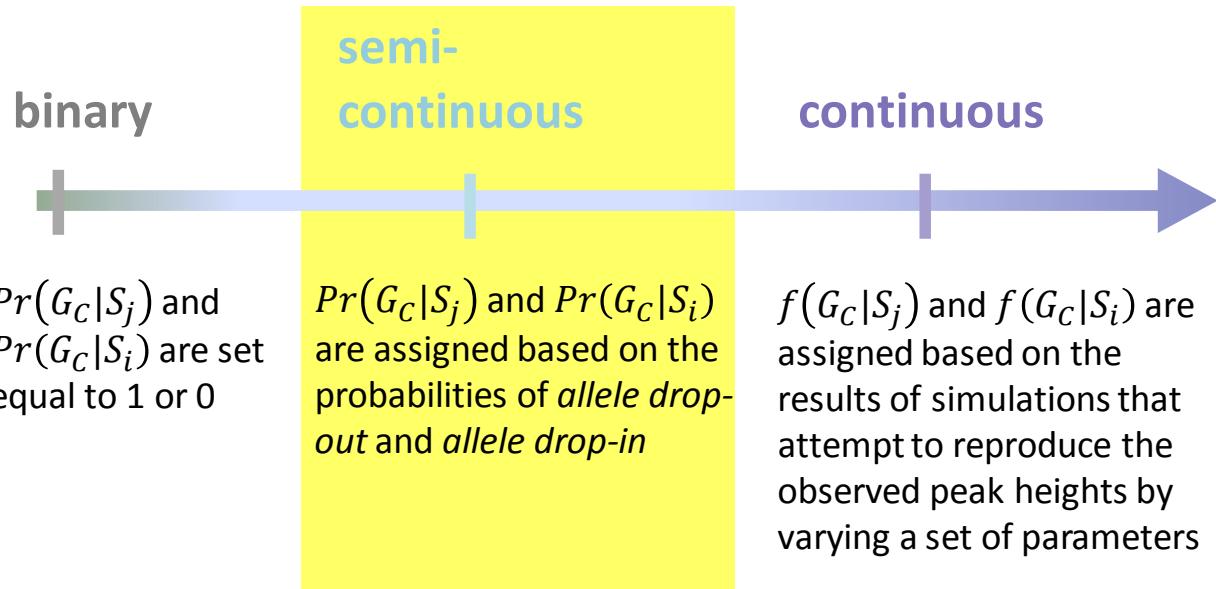
# **INTERPRETATION OF MIXTURES**

## **SEMI-CONTINUOUS MODEL**

274

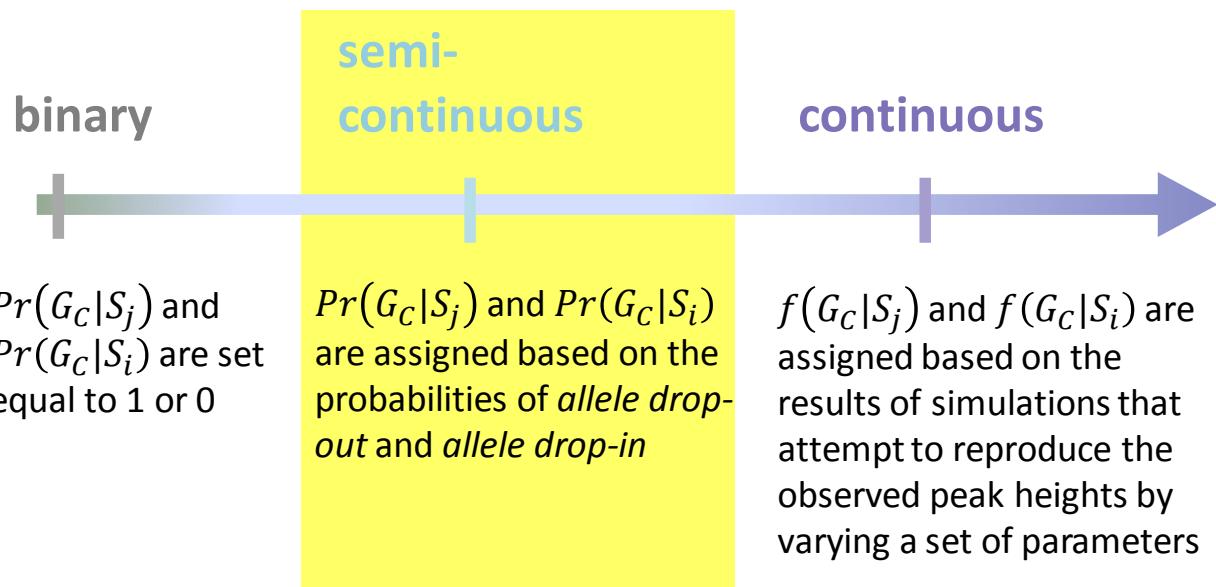
# Likelihood Ratio

$$LR = \frac{\sum_{j=1}^M \Pr(G_C|S_j) \Pr(S_j|G_K^p, H_p)}{\sum_{i=1}^N \Pr(G_C|S_i) \Pr(S_i|G_K^d, H_d)} \quad \text{where } M \leq N$$



# Likelihood Ratio

$$LR = \frac{\sum_{j=1}^M \Pr(G_C|S_j) \Pr(S_j|G_K^p, H_p)}{\sum_{i=1}^N \Pr(G_C|S_i) \Pr(S_i|G_K^d, H_d)} \quad \text{where } M \leq N$$



# Semi-continuous Model

- An allele is either present or absent. The peak heights may be taken in to account to assign probabilities of allele drop-out.
- $$LR = \frac{\sum_{j=1}^M \Pr(G_C | S_j) \Pr(S_j | G_K^p, H_p)}{\sum_{i=1}^N \Pr(G_C | S_i) \Pr(S_i | G_K^d, H_d)}$$

each of these probabilities is assigned based  
on the probabilities of *allele drop-out* and  
*allele drop-in*

276

# Semi-continuous Model

- An allele is either present or absent. The peak heights may be taken in to account to assign probabilities of allele drop-out.
- $$LR = \frac{\sum_{j=1}^M \Pr(G_C | S_j) \Pr(S_j | G_K^p, H_p)}{\sum_{i=1}^N \Pr(G_C | S_i) \Pr(S_i | G_K^d, H_d)}$$

each of these probabilities is assigned based  
on the probabilities of *allele drop-out* and  
*allele drop-in*

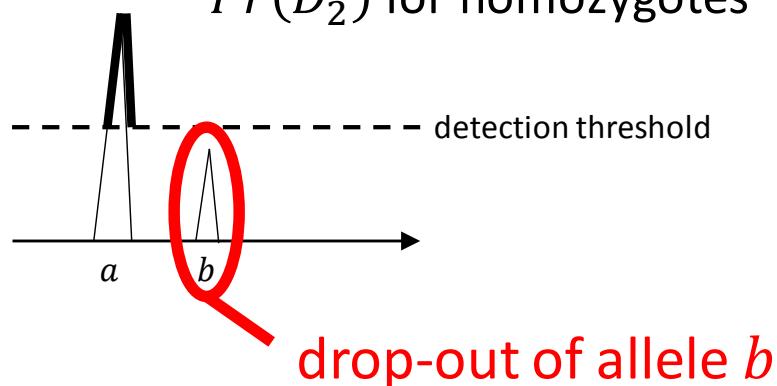
276

# Semi-continuous Model

## Allele drop-out

The donor is  $ab$ :  $Pr(D)$  for heterozygotes

$Pr(D_2)$  for homozygotes



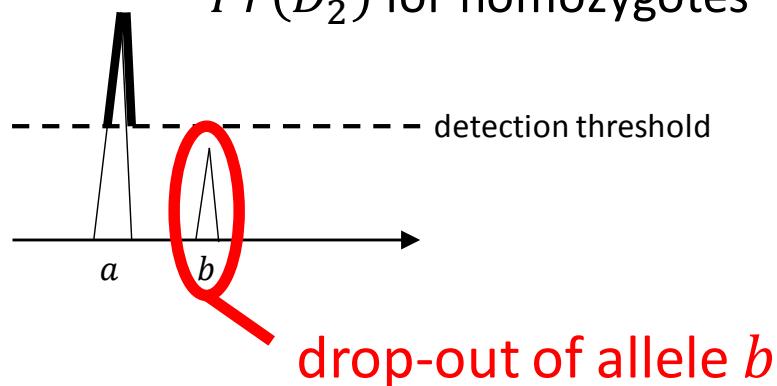
277

# Semi-continuous Model

## Allele drop-out

The donor is  $ab$ :  $Pr(D)$  for heterozygotes

$Pr(D_2)$  for homozygotes

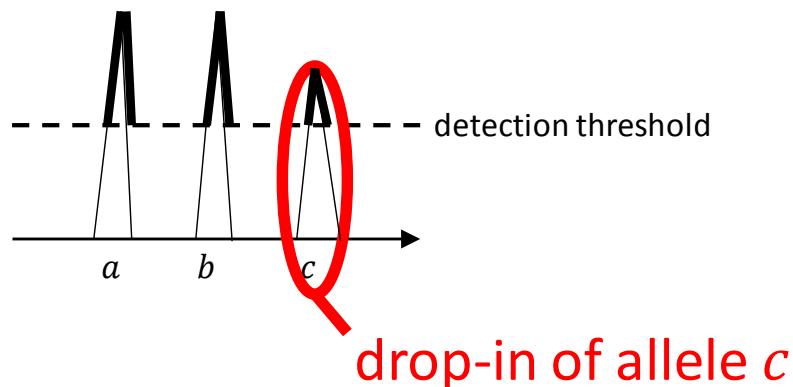


277

# Semi-continuous Model

## Allele drop-in

The donor is  $ab$ :  $Pr(C)p_c$

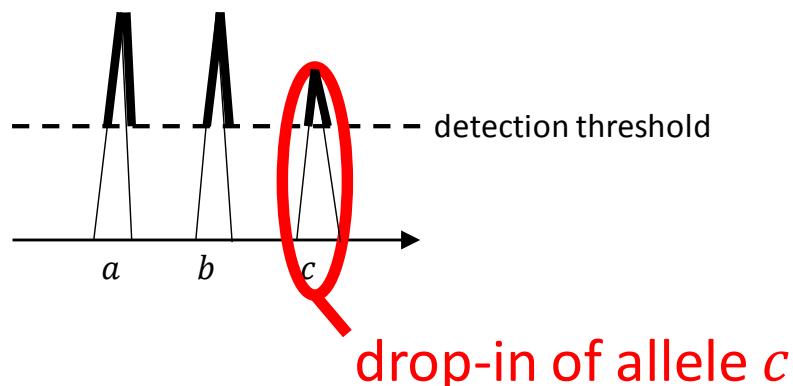


278

# Semi-continuous Model

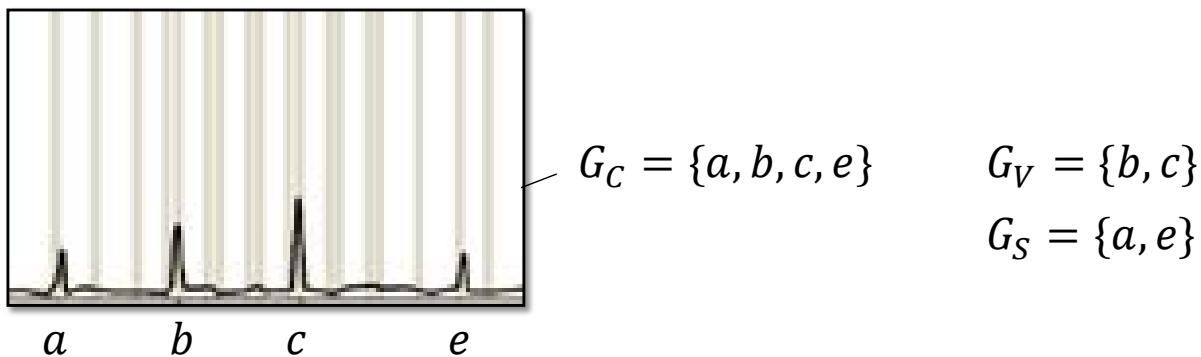
## Allele drop-in

The donor is  $ab$ :  $Pr(C)p_c$



278

# Semi-continuous Model

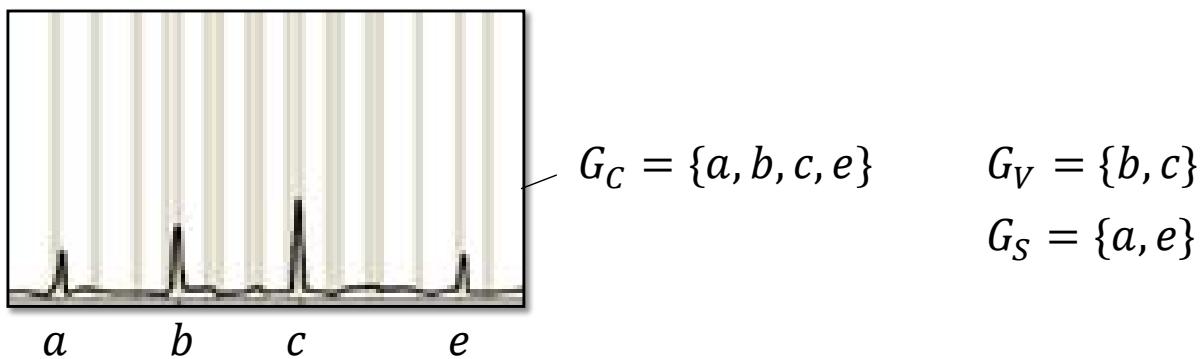


$H_p$ : The crime stain contains the DNA from the victim and the suspect ( $G_K^p = \{G_V, G_S\}$ ).

$H_d$ : The crime stain contains the DNA from the victim and an unknown unrelated person ( $G_K^d = \{G_V\}$ ).

279

# Semi-continuous Model

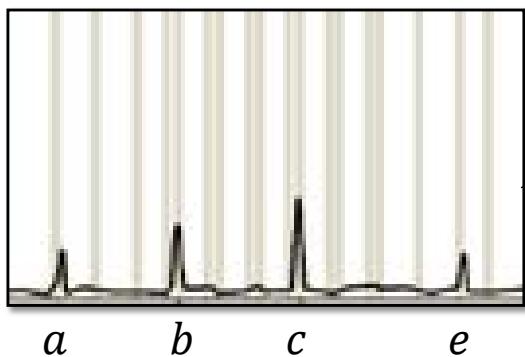


$H_p$ : The crime stain contains the DNA from the victim and the suspect ( $G_K^p = \{G_V, G_S\}$ ).

$H_d$ : The crime stain contains the DNA from the victim and an unknown unrelated person ( $G_K^d = \{G_V\}$ ).

279

# Semi-continuous Model



$$G_C = \{a, b, c, e\} \quad G_V = \{b, c\} \\ G_S = \{a, e\}$$

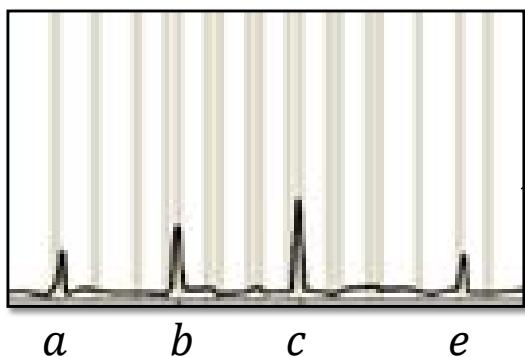
We set *contributor* 1 =  $G_V$ .

Numerator of the likelihood ratio:

genotype set $S_j$	$\Pr(S_j   G_K^p, H_p)$	$\Pr(G_C   S_j)$	product
$bc$ and $ae$	1	$Pr(\bar{D})^4 Pr(\bar{C})$	$Pr(\bar{D})^4 Pr(\bar{C})$

280

# Semi-continuous Model



$$G_C = \{a, b, c, e\} \quad G_V = \{b, c\} \\ G_S = \{a, e\}$$

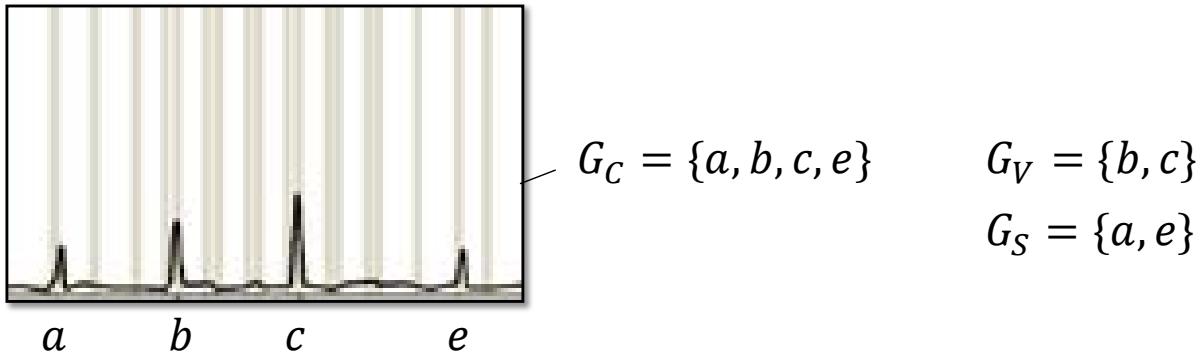
We set *contributor* 1 =  $G_V$ .

Numerator of the likelihood ratio:

genotype set $S_j$	$\Pr(S_j   G_K^p, H_p)$	$\Pr(G_C   S_j)$	product
$bc$ and $ae$	1	$Pr(\bar{D})^4 Pr(\bar{C})$	$Pr(\bar{D})^4 Pr(\bar{C})$

280

# Semi-continuous Model



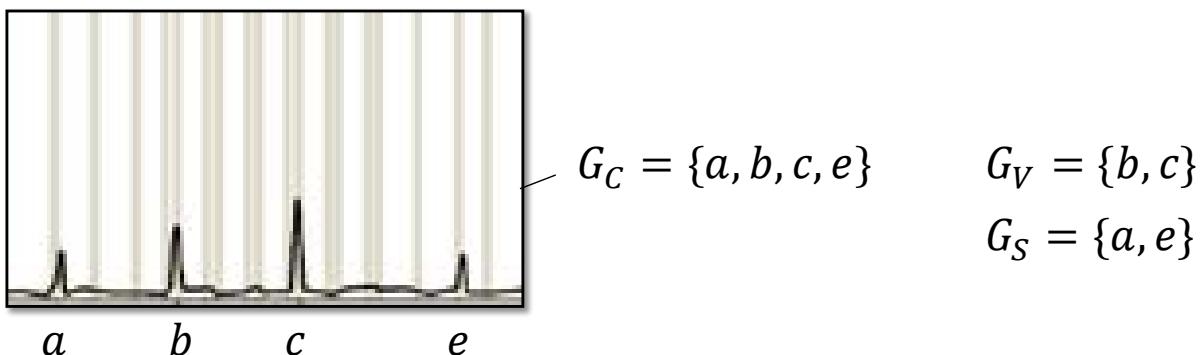
We set *contributor* 1 =  $G_V$ .

Numerator of the likelihood ratio:

$$Pr(G_C | S_1) Pr(S_1 | G_V, G_S, H_p) = Pr(\bar{D})^4 Pr(\bar{C})$$

281

# Semi-continuous Model



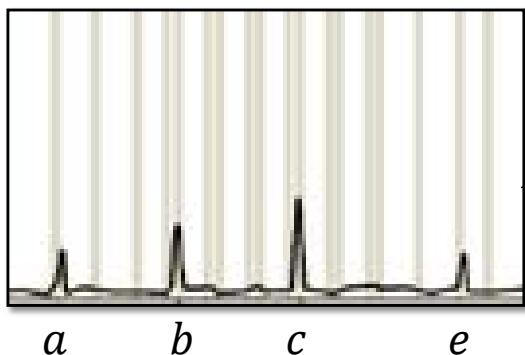
We set *contributor* 1 =  $G_V$ .

Numerator of the likelihood ratio:

$$Pr(G_C | S_1) Pr(S_1 | G_V, G_S, H_p) = Pr(\bar{D})^4 Pr(\bar{C})$$

281

# Semi-continuous Model



$$G_C = \{a, b, c, e\} \quad G_V = \{b, c\} \\ G_S = \{a, e\}$$

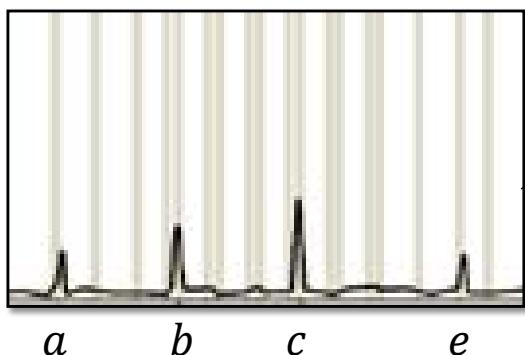
We set *contributor* 1 =  $G_V$ .

Denominator of the likelihood ratio:

$S_1$ :  $bc$  and  $ae$

282

# Semi-continuous Model



$$G_C = \{a, b, c, e\} \quad G_V = \{b, c\} \\ G_S = \{a, e\}$$

We set *contributor* 1 =  $G_V$ .

Denominator of the likelihood ratio:

$S_1$ :  $bc$  and  $ae$

282

# Semi-continuous Model

Denominator of the likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

genotype set $S_i$	$\Pr(S_i G_K^d, H_d)$	$\Pr(G_C S_i)$	product
$bc$ and $ae$	$2p_a p_e$	$Pr(\bar{D})^4 Pr(\bar{C})$	$Pr(\bar{D})^4 Pr(\bar{C}) 2p_a p_e$

283

# Semi-continuous Model

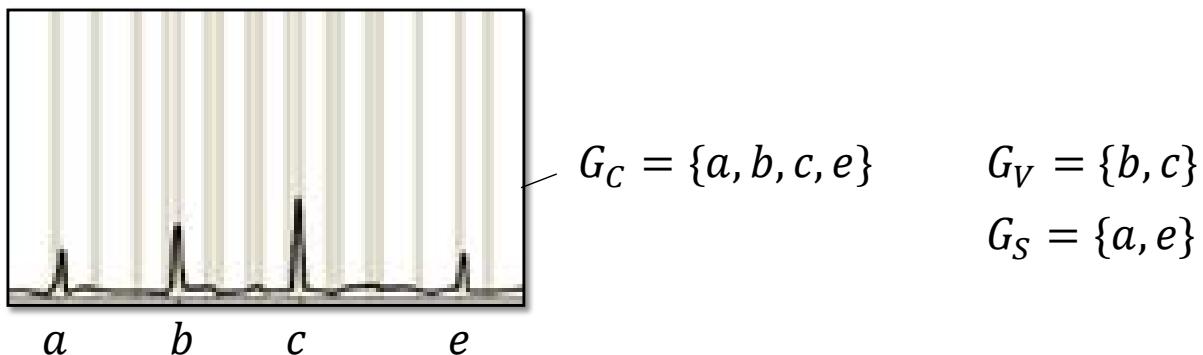
Denominator of the likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

genotype set $S_i$	$\Pr(S_i G_K^d, H_d)$	$\Pr(G_C S_i)$	product
$bc$ and $ae$	$2p_a p_e$	$Pr(\bar{D})^4 Pr(\bar{C})$	$Pr(\bar{D})^4 Pr(\bar{C}) 2p_a p_e$

283

# Semi-continuous Model



We set *contributor* 1 =  $G_V$ .

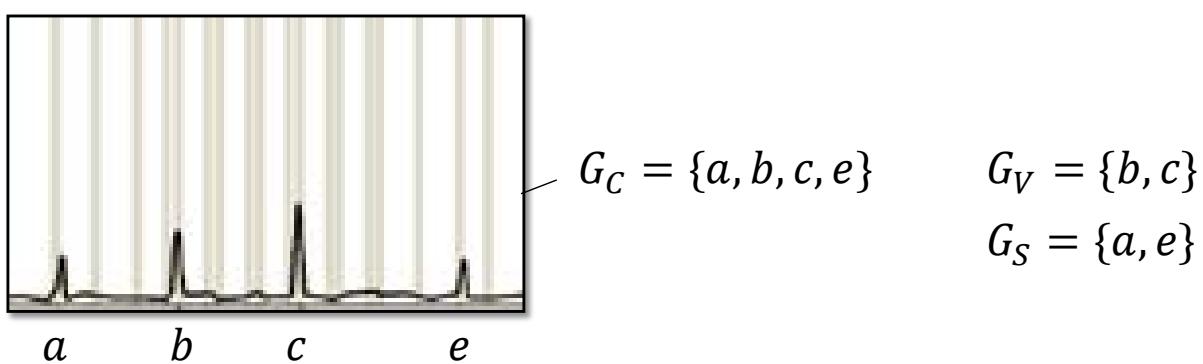
Denominator of the likelihood ratio:

$S_2$ : bc and aa

$S_3$ : bc and ee

284

# Semi-continuous Model



We set *contributor* 1 =  $G_V$ .

Denominator of the likelihood ratio:

$S_2$ : bc and aa

$S_3$ : bc and ee

284

# Semi-continuous Model

Denominator of the likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

genotype set $S_i$	$\Pr(S_i   G_K^d, H_d)$	$\Pr(G_C   S_i)$	product
bc and ae	$2p_a p_e$	$\Pr(\bar{D})^4 \Pr(\bar{C})$	$\Pr(\bar{D})^4 \Pr(\bar{C}) 2p_a p_e$
bc and aa	$p_a^2$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e p_a^2$
bc and ee	$p_e^2$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a p_e^2$

285

# Semi-continuous Model

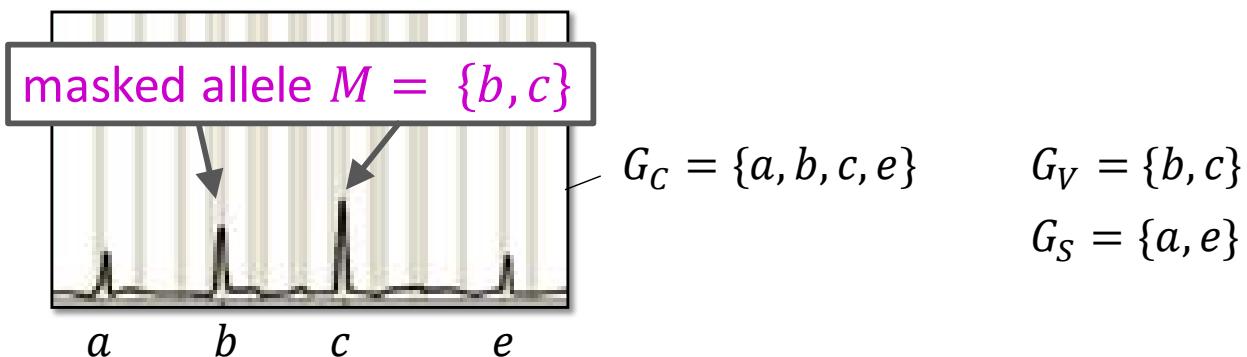
Denominator of the likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

genotype set $S_i$	$\Pr(S_i   G_K^d, H_d)$	$\Pr(G_C   S_i)$	product
bc and ae	$2p_a p_e$	$\Pr(\bar{D})^4 \Pr(\bar{C})$	$\Pr(\bar{D})^4 \Pr(\bar{C}) 2p_a p_e$
bc and aa	$p_a^2$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e p_a^2$
bc and ee	$p_e^2$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a p_e^2$

285

# Semi-continuous Model



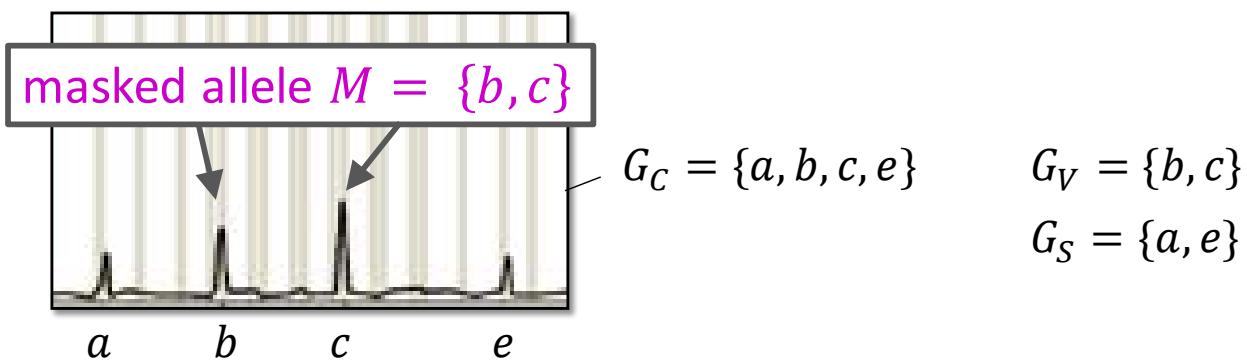
We set *contributor* 1 =  $G_V$ .

Denominator of the likelihood ratio:

$$\begin{array}{ll} S_4: bc \text{ and } aM & S_6: bc \text{ and } eM \\ S_5: & S_7: \end{array}$$

286

# Semi-continuous Model



We set *contributor* 1 =  $G_V$ .

Denominator of the likelihood ratio:

$$\begin{array}{ll} S_4: bc \text{ and } aM & S_6: bc \text{ and } eM \\ S_5: & S_7: \end{array}$$

286

# Semi-continuous Model

Denominator of the likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

genotype set $S_i$	$\Pr(S_i   G_K^d, H_d)$	$\Pr(G_C   S_i)$	product
bc and ae	$2p_a p_e$	$\Pr(\bar{D})^4 \Pr(\bar{C})$	$\Pr(\bar{D})^4 \Pr(\bar{C}) 2p_a p_e$
bc and aa	$p_a^2$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e p_a^2$
bc and ee	$p_e^2$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a p_e^2$
bc and aM	$2p_a(p_b + p_c)$	$\Pr(\bar{D})^4 \Pr(C) p_e$	$\Pr(\bar{D})^4 \Pr(C) p_e 2p_a(p_b + p_c)$
bc and eM	$2p_e(p_b + p_c)$	$\Pr(\bar{D})^4 \Pr(C) p_a$	$\Pr(\bar{D})^4 \Pr(C) p_a 2p_e(p_b + p_c)$

287

# Semi-continuous Model

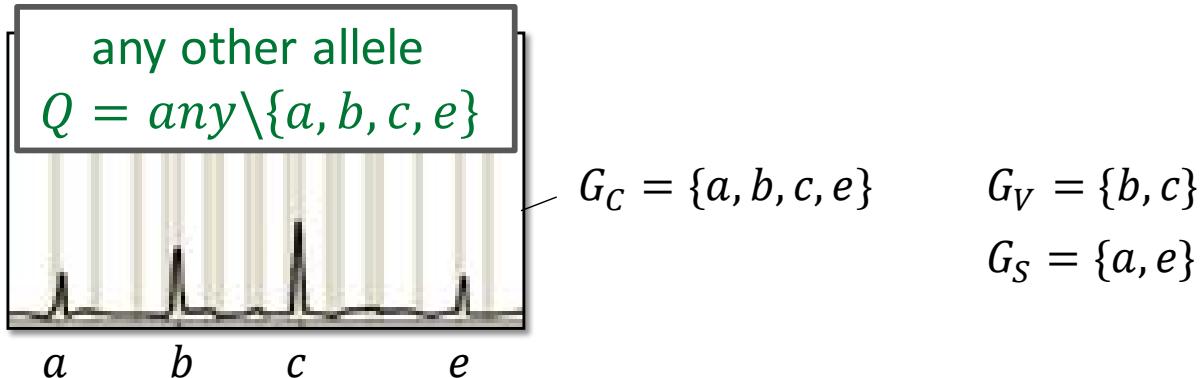
Denominator of the likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

genotype set $S_i$	$\Pr(S_i   G_K^d, H_d)$	$\Pr(G_C   S_i)$	product
bc and ae	$2p_a p_e$	$\Pr(\bar{D})^4 \Pr(\bar{C})$	$\Pr(\bar{D})^4 \Pr(\bar{C}) 2p_a p_e$
bc and aa	$p_a^2$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e p_a^2$
bc and ee	$p_e^2$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a p_e^2$
bc and aM	$2p_a(p_b + p_c)$	$\Pr(\bar{D})^4 \Pr(C) p_e$	$\Pr(\bar{D})^4 \Pr(C) p_e 2p_a(p_b + p_c)$
bc and eM	$2p_e(p_b + p_c)$	$\Pr(\bar{D})^4 \Pr(C) p_a$	$\Pr(\bar{D})^4 \Pr(C) p_a 2p_e(p_b + p_c)$

287

# Semi-continuous Model



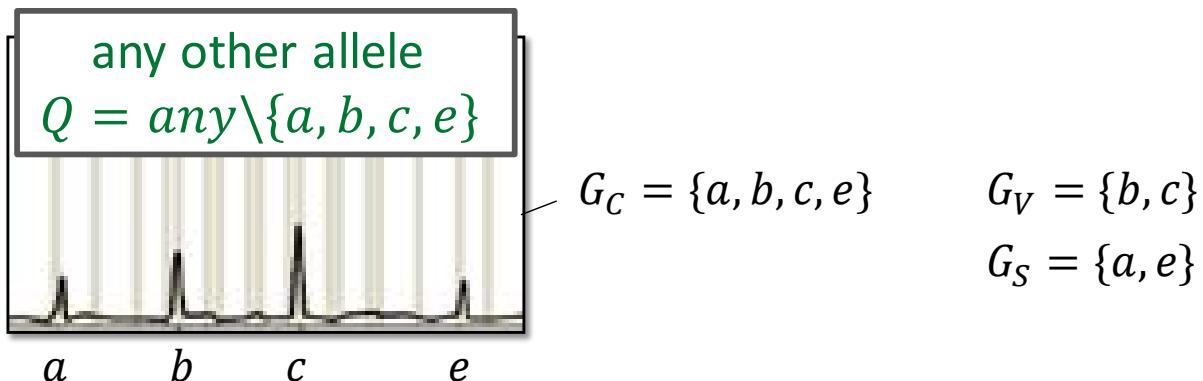
We set *contributor* 1 =  $G_V$ .

Denominator of the likelihood ratio:

$$\begin{array}{ll} S_8: bc \text{ and } aQ & S_{10}: bc \text{ and } eQ \\ S_9: & S_{11}: \end{array}$$

288

# Semi-continuous Model



We set *contributor* 1 =  $G_V$ .

Denominator of the likelihood ratio:

$$\begin{array}{ll} S_8: bc \text{ and } aQ & S_{10}: bc \text{ and } eQ \\ S_9: & S_{11}: \end{array}$$

288

# Semi-continuous Model

Denominator of the likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

genotype set $S_i$	$\Pr(S_i   G_K^d, H_d)$	$\Pr(G_C   S_i)$	product
bc and ae	$2p_a p_e$	$Pr(\bar{D})^4 Pr(\bar{C})$	$Pr(\bar{D})^4 Pr(\bar{C}) 2p_a p_e$
bc and aa	$p_a^2$	$Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_e$	$Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_e p_a^2$
bc and ee	$p_e^2$	$Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_a$	$Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_a p_e^2$
bc and aM	$2p_a(p_b + p_c)$	$Pr(\bar{D})^4 Pr(C) p_e$	$Pr(\bar{D})^4 Pr(C) p_e 2p_a(p_b + p_c)$
bc and eM	$2p_e(p_b + p_c)$	$Pr(\bar{D})^4 Pr(C) p_a$	$Pr(\bar{D})^4 Pr(C) p_a 2p_e(p_b + p_c)$
bc and aQ	$2p_a p_Q$	$Pr(\bar{D})^3 Pr(D) Pr(C) p_e$	$Pr(\bar{D})^3 Pr(D) Pr(C) p_e 2p_a p_Q$
bc and eQ	$2p_e p_Q$	$Pr(\bar{D})^3 Pr(D) Pr(C) p_a$	$Pr(\bar{D})^3 Pr(D) Pr(C) p_a 2p_e p_Q$

289

# Semi-continuous Model

Denominator of the likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

genotype set $S_i$	$\Pr(S_i   G_K^d, H_d)$	$\Pr(G_C   S_i)$	product
bc and ae	$2p_a p_e$	$Pr(\bar{D})^4 Pr(\bar{C})$	$Pr(\bar{D})^4 Pr(\bar{C}) 2p_a p_e$
bc and aa	$p_a^2$	$Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_e$	$Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_e p_a^2$
bc and ee	$p_e^2$	$Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_a$	$Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_a p_e^2$
bc and aM	$2p_a(p_b + p_c)$	$Pr(\bar{D})^4 Pr(C) p_e$	$Pr(\bar{D})^4 Pr(C) p_e 2p_a(p_b + p_c)$
bc and eM	$2p_e(p_b + p_c)$	$Pr(\bar{D})^4 Pr(C) p_a$	$Pr(\bar{D})^4 Pr(C) p_a 2p_e(p_b + p_c)$
bc and aQ	$2p_a p_Q$	$Pr(\bar{D})^3 Pr(D) Pr(C) p_e$	$Pr(\bar{D})^3 Pr(D) Pr(C) p_e 2p_a p_Q$
bc and eQ	$2p_e p_Q$	$Pr(\bar{D})^3 Pr(D) Pr(C) p_a$	$Pr(\bar{D})^3 Pr(D) Pr(C) p_a 2p_e p_Q$

289

# Semi-continuous Model

Denominator of the likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

$$\begin{aligned} \sum_{i=1}^N Pr(G_C|S_i)Pr(S_i|G_V, H_d) = & Pr(\bar{D})^4 Pr(\bar{C}) 2p_a p_e + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_e p_a^2 \\ & + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_a p_e^2 + Pr(\bar{D})^4 Pr(C) p_e 2p_a (p_b + p_c) \\ & + Pr(\bar{D})^4 Pr(C) p_a 2p_e (p_b + p_c) + Pr(\bar{D})^3 Pr(D) Pr(C) p_e 2p_a p_Q \\ & + Pr(\bar{D})^3 Pr(D) Pr(C) p_a 2p_e p_Q \end{aligned}$$

290

# Semi-continuous Model

Denominator of the likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

$$\begin{aligned} \sum_{i=1}^N Pr(G_C|S_i)Pr(S_i|G_V, H_d) = & Pr(\bar{D})^4 Pr(\bar{C}) 2p_a p_e + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_e p_a^2 \\ & + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_a p_e^2 + Pr(\bar{D})^4 Pr(C) p_e 2p_a (p_b + p_c) \\ & + Pr(\bar{D})^4 Pr(C) p_a 2p_e (p_b + p_c) + Pr(\bar{D})^3 Pr(D) Pr(C) p_e 2p_a p_Q \\ & + Pr(\bar{D})^3 Pr(D) Pr(C) p_a 2p_e p_Q \end{aligned}$$

290

# Semi-continuous Model

Likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

$$LR = \frac{Pr(\bar{D})^4 Pr(\bar{C})}{Pr(\bar{D})^4 Pr(\bar{C}) 2p_a p_e + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_e p_a^2 + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_a p_e^2 + Pr(\bar{D})^4 Pr(C) p_e 2p_a (p_b + p_c) + Pr(\bar{D})^4 Pr(C) p_a 2p_e (p_b + p_c) + Pr(\bar{D})^3 Pr(D) Pr(C) p_e 2p_a p_Q + Pr(\bar{D})^3 Pr(D) Pr(C) p_a 2p_e p_Q}$$

291

# Semi-continuous Model

Likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

$$LR = \frac{Pr(\bar{D})^4 Pr(\bar{C})}{Pr(\bar{D})^4 Pr(\bar{C}) 2p_a p_e + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_e p_a^2 + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_a p_e^2 + Pr(\bar{D})^4 Pr(C) p_e 2p_a (p_b + p_c) + Pr(\bar{D})^4 Pr(C) p_a 2p_e (p_b + p_c) + Pr(\bar{D})^3 Pr(D) Pr(C) p_e 2p_a p_Q + Pr(\bar{D})^3 Pr(D) Pr(C) p_a 2p_e p_Q}$$

291

# Semi-continuous Model

Likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

$$LR = \frac{\Pr(\bar{D})^4 \Pr(C)}{\Pr(\bar{D})^4 \Pr(C) 2p_a p_e + \cancel{\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e p_a^2} \\ + \cancel{\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a p_e^2} + \cancel{\Pr(\bar{D})^4 \Pr(C) p_e 2p_a (p_b + p_c)} \\ + \cancel{\Pr(\bar{D})^4 \Pr(C) p_a 2p_e (p_b + p_c)} + \cancel{\Pr(\bar{D})^3 \Pr(D) \Pr(C) p_e 2p_a p_Q} \\ + \cancel{\Pr(\bar{D})^3 \Pr(D) \Pr(C) p_a 2p_e p_Q}}$$

If  $\Pr(C) = 0$ , then this expression simplifies to:  $LR = \frac{1}{2p_a p_e}$

292

# Semi-continuous Model

Likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

$$LR = \frac{\Pr(\bar{D})^4 \Pr(C)}{\Pr(\bar{D})^4 \Pr(C) 2p_a p_e + \cancel{\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e p_a^2} \\ + \cancel{\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a p_e^2} + \cancel{\Pr(\bar{D})^4 \Pr(C) p_e 2p_a (p_b + p_c)} \\ + \cancel{\Pr(\bar{D})^4 \Pr(C) p_a 2p_e (p_b + p_c)} + \cancel{\Pr(\bar{D})^3 \Pr(D) \Pr(C) p_e 2p_a p_Q} \\ + \cancel{\Pr(\bar{D})^3 \Pr(D) \Pr(C) p_a 2p_e p_Q}}$$

If  $\Pr(C) = 0$ , then this expression simplifies to:  $LR = \frac{1}{2p_a p_e}$

292

# Semi-continuous Model

Likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

$$LR = \frac{Pr(\bar{D})^4 Pr(\bar{C})}{Pr(\bar{D})^4 Pr(\bar{C}) 2p_a p_e + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_e p_a^2 + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_a p_e^2 + Pr(\bar{D})^4 Pr(C) p_e 2p_a (p_b + p_c) + Pr(\bar{D})^4 Pr(C) p_a 2p_e (p_b + p_c) + \cancel{Pr(\bar{D})^3 Pr(D) Pr(C) p_e 2p_a p_0} + \cancel{Pr(\bar{D})^3 Pr(D) Pr(C) p_a 2p_e p_Q}}$$

if  $Pr(D) = 0$

293

# Semi-continuous Model

Likelihood ratio:

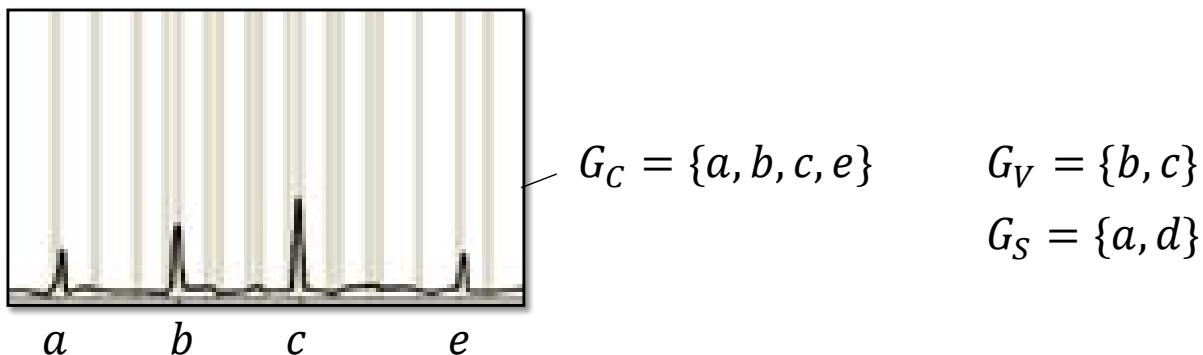
Assumption: The maximum number of drop-in alleles is 1.

$$LR = \frac{Pr(\bar{D})^4 Pr(\bar{C})}{Pr(\bar{D})^4 Pr(\bar{C}) 2p_a p_e + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_e p_a^2 + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_a p_e^2 + Pr(\bar{D})^4 Pr(C) p_e 2p_a (p_b + p_c) + Pr(\bar{D})^4 Pr(C) p_a 2p_e (p_b + p_c) + \cancel{Pr(\bar{D})^3 Pr(D) Pr(C) p_e 2p_a p_0} + \cancel{Pr(\bar{D})^3 Pr(D) Pr(C) p_a 2p_e p_Q}}$$

if  $Pr(D) = 0$

293

# Semi-continuous Model

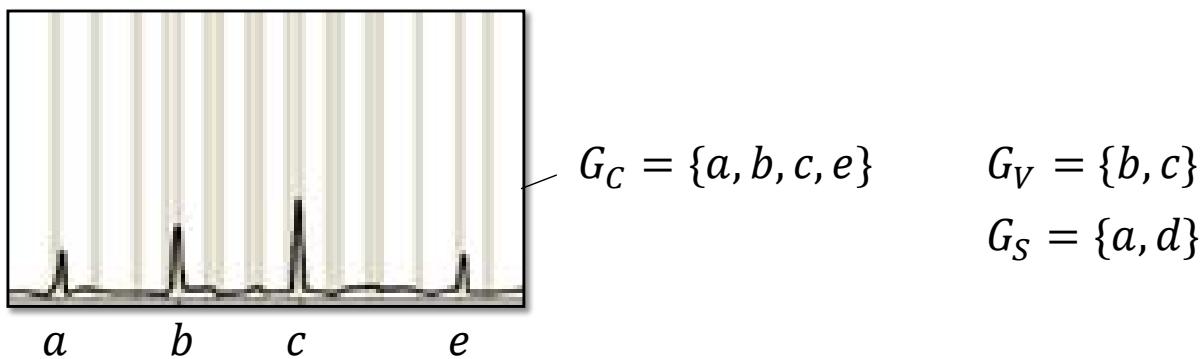


$H_p$ : The crime stain contains the DNA from the victim and the suspect ( $G_K^p = \{G_V, G_S\}$ ).

$H_d$ : The crime stain contains the DNA from the victim and an unknown unrelated person ( $G_K^d = \{G_V\}$ ).

294

# Semi-continuous Model

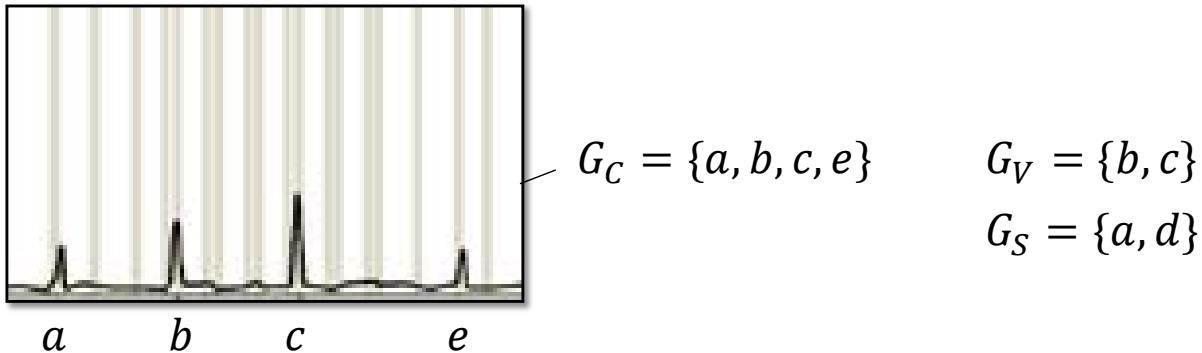


$H_p$ : The crime stain contains the DNA from the victim and the suspect ( $G_K^p = \{G_V, G_S\}$ ).

$H_d$ : The crime stain contains the DNA from the victim and an unknown unrelated person ( $G_K^d = \{G_V\}$ ).

294

# Semi-continuous Model



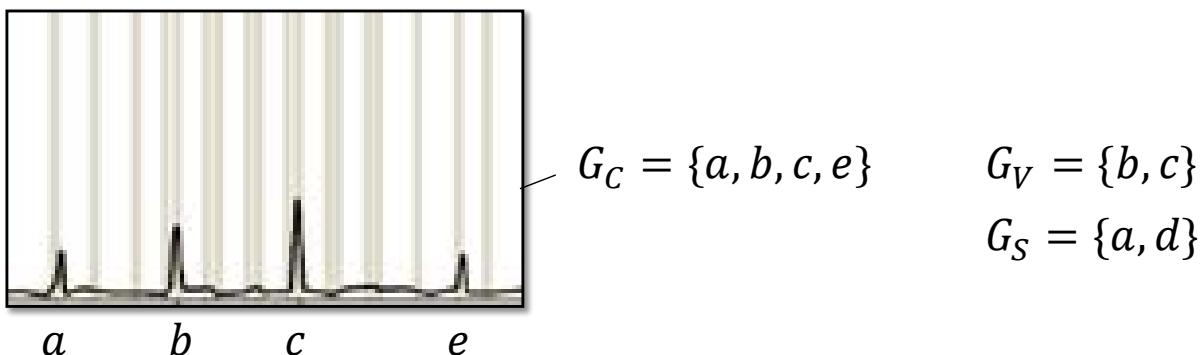
We set *contributor* 1 =  $G_V$ .

Numerator of the likelihood ratio:

genotype set $S_j$	$\Pr(S_j   G_K^p, H_p)$	$\Pr(G_C   S_j)$	product
$bc$ and $ad$	1	$Pr(\bar{D})^3 Pr(D) Pr(C) p_e$	$Pr(\bar{D})^3 Pr(D) Pr(C) p_e$

295

# Semi-continuous Model



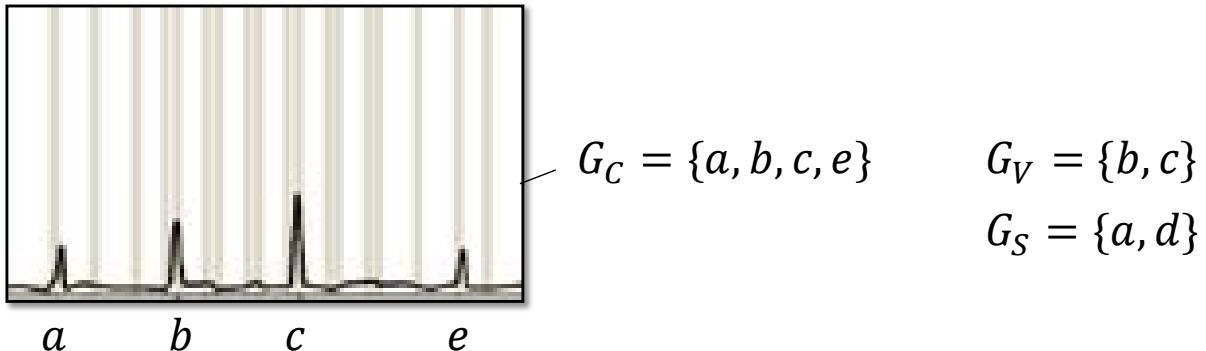
We set *contributor* 1 =  $G_V$ .

Numerator of the likelihood ratio:

genotype set $S_j$	$\Pr(S_j   G_K^p, H_p)$	$\Pr(G_C   S_j)$	product
$bc$ and $ad$	1	$Pr(\bar{D})^3 Pr(D) Pr(C) p_e$	$Pr(\bar{D})^3 Pr(D) Pr(C) p_e$

295

# Semi-continuous Model



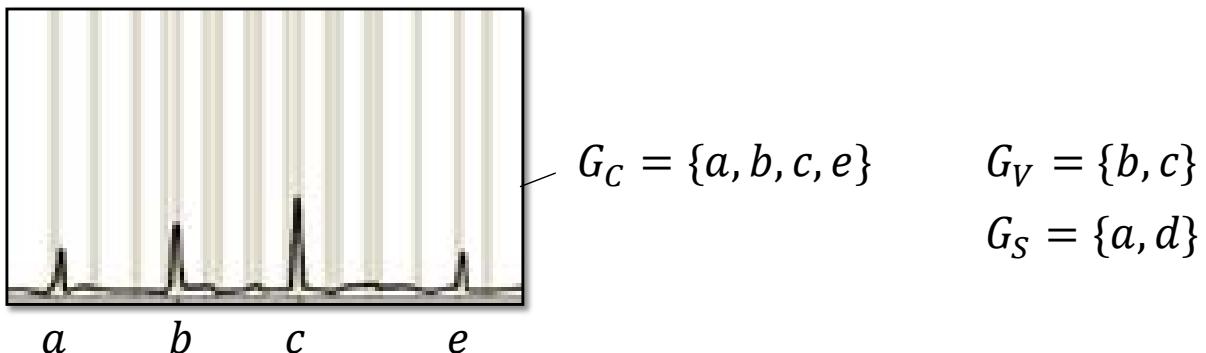
We set *contributor 1* =  $G_V$ .

Numerator of the likelihood ratio:

$$Pr(G_C|S_1)Pr(S_1|G_V, G_S, H_p) = Pr(\bar{D})^3 Pr(D) Pr(C) p_e$$

296

# Semi-continuous Model



We set *contributor 1* =  $G_V$ .

Numerator of the likelihood ratio:

$$Pr(G_C|S_1)Pr(S_1|G_V, G_S, H_p) = Pr(\bar{D})^3 Pr(D) Pr(C) p_e$$

296

# Semi-continuous Model

Denominator of the likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

genotype set $S_i$	$\Pr(S_i G_K^d, H_d)$	$\Pr(G_C S_i)$	product
bc and ae	$2p_a p_e$	$\Pr(\bar{D})^4 \Pr(\bar{C})$	$\Pr(\bar{D})^4 \Pr(\bar{C}) 2p_a p_e$
bc and aQ	$2p_a p_Q$	$\Pr(\bar{D})^3 \Pr(D) \Pr(C) p_e$	$\Pr(\bar{D})^3 \Pr(D) \Pr(C) p_e 2p_a p_Q$
bc and eQ	$2p_e p_Q$	$\Pr(\bar{D})^3 \Pr(D) \Pr(C) p_a$	$\Pr(\bar{D})^3 \Pr(D) \Pr(C) p_a 2p_e p_Q$
bc and aa	$p_a^2$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e p_a^2$
bc and ee	$p_e^2$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a p_e^2$
bc and aM	$2p_a(p_b + p_c)$	$\Pr(\bar{D})^4 \Pr(C) p_e$	$\Pr(\bar{D})^4 \Pr(C) p_e 2p_a(p_b + p_c)$
bc and eM	$2p_e(p_b + p_c)$	$\Pr(\bar{D})^4 \Pr(C) p_a$	$\Pr(\bar{D})^4 \Pr(C) p_a 2p_e(p_b + p_c)$

297

# Semi-continuous Model

Denominator of the likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

genotype set $S_i$	$\Pr(S_i G_K^d, H_d)$	$\Pr(G_C S_i)$	product
bc and ae	$2p_a p_e$	$\Pr(\bar{D})^4 \Pr(\bar{C})$	$\Pr(\bar{D})^4 \Pr(\bar{C}) 2p_a p_e$
bc and aQ	$2p_a p_Q$	$\Pr(\bar{D})^3 \Pr(D) \Pr(C) p_e$	$\Pr(\bar{D})^3 \Pr(D) \Pr(C) p_e 2p_a p_Q$
bc and eQ	$2p_e p_Q$	$\Pr(\bar{D})^3 \Pr(D) \Pr(C) p_a$	$\Pr(\bar{D})^3 \Pr(D) \Pr(C) p_a 2p_e p_Q$
bc and aa	$p_a^2$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_e p_a^2$
bc and ee	$p_e^2$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a$	$\Pr(\bar{D})^2 \Pr(\bar{D}_2) \Pr(C) p_a p_e^2$
bc and aM	$2p_a(p_b + p_c)$	$\Pr(\bar{D})^4 \Pr(C) p_e$	$\Pr(\bar{D})^4 \Pr(C) p_e 2p_a(p_b + p_c)$
bc and eM	$2p_e(p_b + p_c)$	$\Pr(\bar{D})^4 \Pr(C) p_a$	$\Pr(\bar{D})^4 \Pr(C) p_a 2p_e(p_b + p_c)$

297

# Semi-continuous Model

Likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

$$LR = \frac{Pr(\bar{D})^3 Pr(D) Pr(C) p_e}{Pr(\bar{D})^4 Pr(\bar{C}) 2p_a p_e + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_e p_a^2 + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_a p_e^2 + Pr(\bar{D})^4 Pr(C) p_e 2p_a (p_b + p_c) + Pr(\bar{D})^4 Pr(C) p_a 2p_e (p_b + p_c) + Pr(\bar{D})^3 Pr(D) Pr(C) p_e 2p_a p_Q + Pr(\bar{D})^3 Pr(D) Pr(C) p_a 2p_e p_Q}$$

If  $Pr(C) = 0$  and/or  $Pr(D) = 0$ , then  $LR = 0$ .

298

# Semi-continuous Model

Likelihood ratio:

Assumption: The maximum number of drop-in alleles is 1.

$$LR = \frac{Pr(\bar{D})^3 Pr(D) Pr(C) p_e}{Pr(\bar{D})^4 Pr(\bar{C}) 2p_a p_e + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_e p_a^2 + Pr(\bar{D})^2 Pr(\bar{D}_2) Pr(C) p_a p_e^2 + Pr(\bar{D})^4 Pr(C) p_e 2p_a (p_b + p_c) + Pr(\bar{D})^4 Pr(C) p_a 2p_e (p_b + p_c) + Pr(\bar{D})^3 Pr(D) Pr(C) p_e 2p_a p_Q + Pr(\bar{D})^3 Pr(D) Pr(C) p_a 2p_e p_Q}$$

If  $Pr(C) = 0$  and/or  $Pr(D) = 0$ , then  $LR = 0$ .

298

# Semi-continuous Model

→ example Lab Retriever

→ example LRmix

299

# Semi-continuous Model

→ example Lab Retriever

→ example LRmix

299

# Lab Retriever and LRmix

- free software
  - input: designation of alleles (without peak heights)
  - user defines the pair of propositions by specifying the number of contributors and the known contributors for each proposition
  - Fst
  - $Pr(C)$
  - $Pr(D)$
- Lab Retriever  $\neq$  LRmix**

300

# Lab Retriever and LRmix

- free software
  - input: designation of alleles (without peak heights)
  - user defines the pair of propositions by specifying the number of contributors and the known contributors for each proposition
  - Fst
  - $Pr(C)$
  - $Pr(D)$
- Lab Retriever  $\neq$  LRmix**

300

# Lab Retriever and LRmix

	Lab Retriever	LRmix
replicates?	no	yes
Fst	currently hard-coded: Fst = 0.01	set by user (default Fst = 0)
$Pr(C)$	set by user (default $Pr(C) = 0.01$ )	set by user (default $Pr(C) = 0.05$ )
$Pr(D)$	A logistic regression model defines $Pr(D)$ in function of the average allelic peak height in the crime stain's profile. $Pr(D_2) = \frac{1}{2}Pr(D)^2$	Monte Carlo simulations of the number of allele drop-outs for different values of $Pr(D)$ produce the most plausible range for $Pr(D)$ , and then within this range $Pr(D)$ is set equal to the value producing the smallest $LR$ .

301

# Lab Retriever and LRmix

	Lab Retriever	LRmix
replicates?	no	yes
Fst	currently hard-coded: Fst = 0.01	set by user (default Fst = 0)
$Pr(C)$	set by user (default $Pr(C) = 0.01$ )	set by user (default $Pr(C) = 0.05$ )
$Pr(D)$	A logistic regression model defines $Pr(D)$ in function of the average allelic peak height in the crime stain's profile. $Pr(D_2) = \frac{1}{2}Pr(D)^2$	Monte Carlo simulations of the number of allele drop-outs for different values of $Pr(D)$ produce the most plausible range for $Pr(D)$ , and then within this range $Pr(D)$ is set equal to the value producing the smallest $LR$ .

301

# $Pr(D)$ : Lab Retriever

$Pr(D)$  is determined from the average allelic peak height in the crime stain's profile.

$$Pr(D) = \frac{e^{\beta_0 + \beta_1 \hat{H}}}{1 + e^{\beta_0 + \beta_1 \hat{H}}}$$

where:

$\hat{H}$  = the average allelic peak height in the crime stain's profile

Reference: K.E. Lohmueller, N. Rudin, K. Inman. Analysis of allelic drop-out using Identifiler® and PowerPlex® 16 forensic STR typing systems. Forensic Sci Int. Genet. 2014; 12: 1-11.

302

# $Pr(D)$ : Lab Retriever

$Pr(D)$  is determined from the average allelic peak height in the crime stain's profile.

$$Pr(D) = \frac{e^{\beta_0 + \beta_1 \hat{H}}}{1 + e^{\beta_0 + \beta_1 \hat{H}}}$$

where:

$\hat{H}$  = the average allelic peak height in the crime stain's profile

Reference: K.E. Lohmueller, N. Rudin, K. Inman. Analysis of allelic drop-out using Identifiler® and PowerPlex® 16 forensic STR typing systems. Forensic Sci Int. Genet. 2014; 12: 1-11.

302

# $Pr(D)$ : Lab Retriever

$Pr(D)$  is determined from the average allelic peak height in the crime stain's profile.

$$Pr(D) = \frac{e^{\beta_0 + \beta_1 \hat{H}}}{1 + e^{\beta_0 + \beta_1 \hat{H}}}$$

where:

Kit:	PowerPlex 16		Identifiler	
Detection Thresh.:	30 rfu	50 rfu	30 rfu	50 rfu
$\beta_0$	1.864	3.756	3.286	4.3884
$\beta_1$	-0.0357	-0.0402	-0.0554	-0.0472

Reference: K.E. Lohmueller, N. Rudin, K. Inman. Analysis of allelic drop-out using Identifiler® and PowerPlex® 16 forensic STR typing systems. Forensic Sci Int. Genet. 2014; 12: 1-11.

303

# $Pr(D)$ : Lab Retriever

$Pr(D)$  is determined from the average allelic peak height in the crime stain's profile.

$$Pr(D) = \frac{e^{\beta_0 + \beta_1 \hat{H}}}{1 + e^{\beta_0 + \beta_1 \hat{H}}}$$

where:

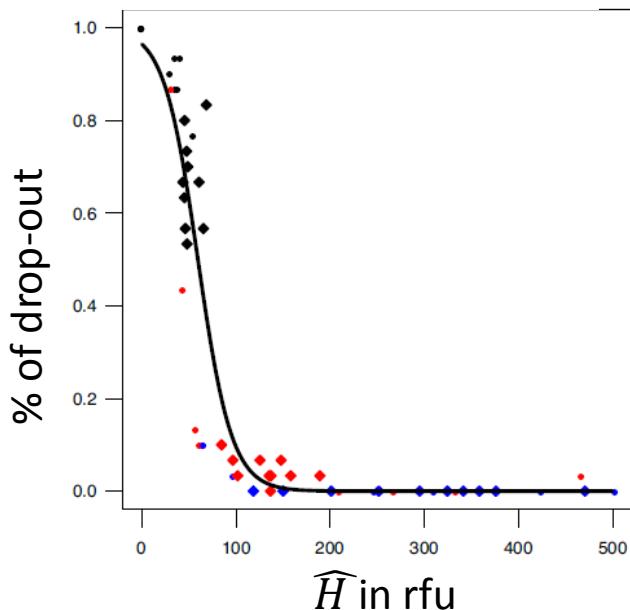
Kit:	PowerPlex 16		Identifiler	
Detection Thresh.:	30 rfu	50 rfu	30 rfu	50 rfu
$\beta_0$	1.864	3.756	3.286	4.3884
$\beta_1$	-0.0357	-0.0402	-0.0554	-0.0472

Reference: K.E. Lohmueller, N. Rudin, K. Inman. Analysis of allelic drop-out using Identifiler® and PowerPlex® 16 forensic STR typing systems. Forensic Sci Int. Genet. 2014; 12: 1-11.

303

# $Pr(D)$ : Lab Retriever

detection threshold = 30 rfu:



detection threshold = 50 rfu:

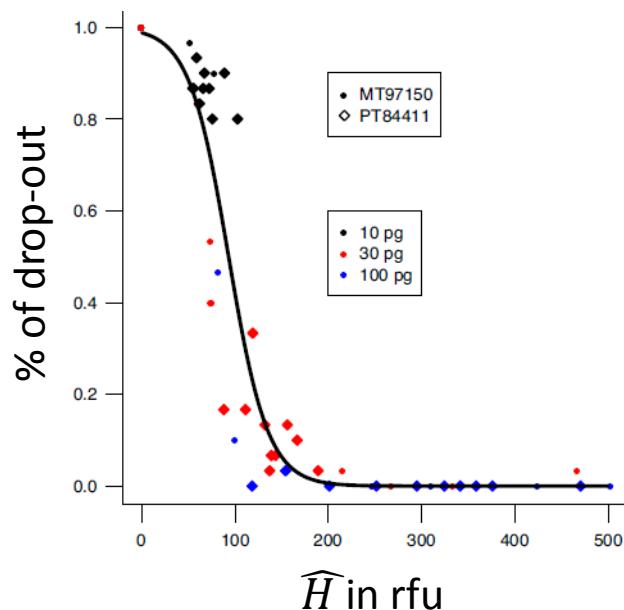
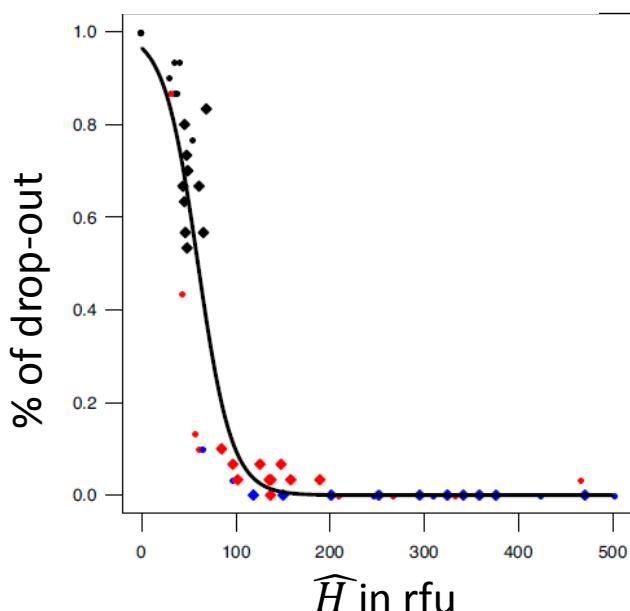


Figure 1 in K.E. Lohmueller, N. Rudin, K. Inman. Analysis of allelic drop-out using Identifiler® and PowerPlex® 16 forensic STR typing systems. Forensic Sci Int. Genet. 2014; 12: page 4.

304

# $Pr(D)$ : Lab Retriever

detection threshold = 30 rfu:



detection threshold = 50 rfu:

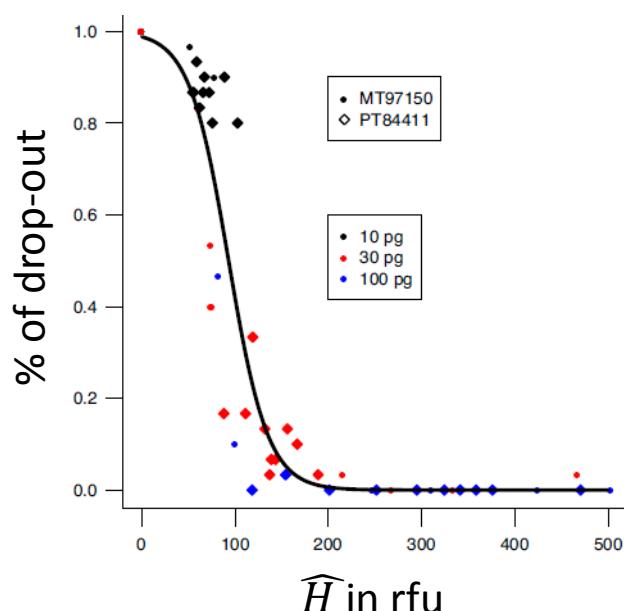
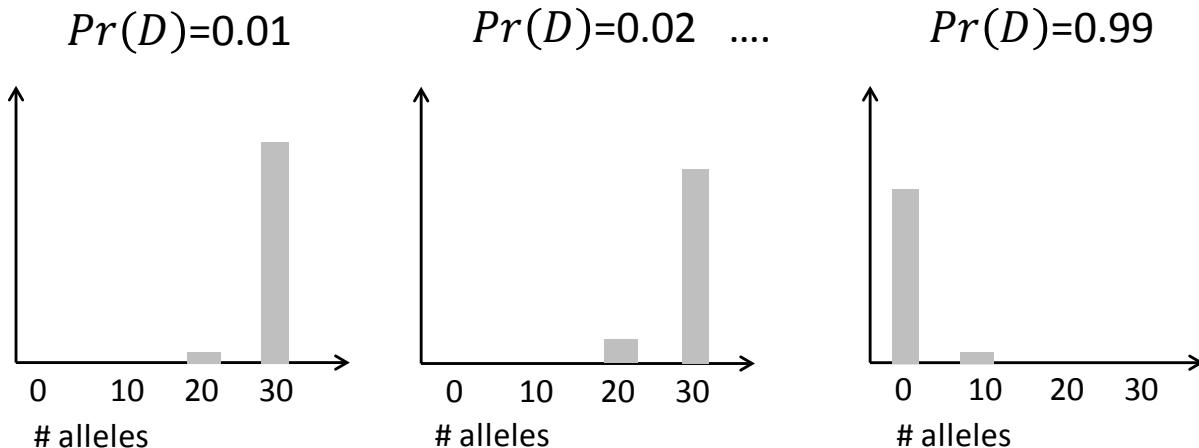


Figure 1 in K.E. Lohmueller, N. Rudin, K. Inman. Analysis of allelic drop-out using Identifiler® and PowerPlex® 16 forensic STR typing systems. Forensic Sci Int. Genet. 2014; 12: page 4.

304

# $Pr(D)$ : LRmix

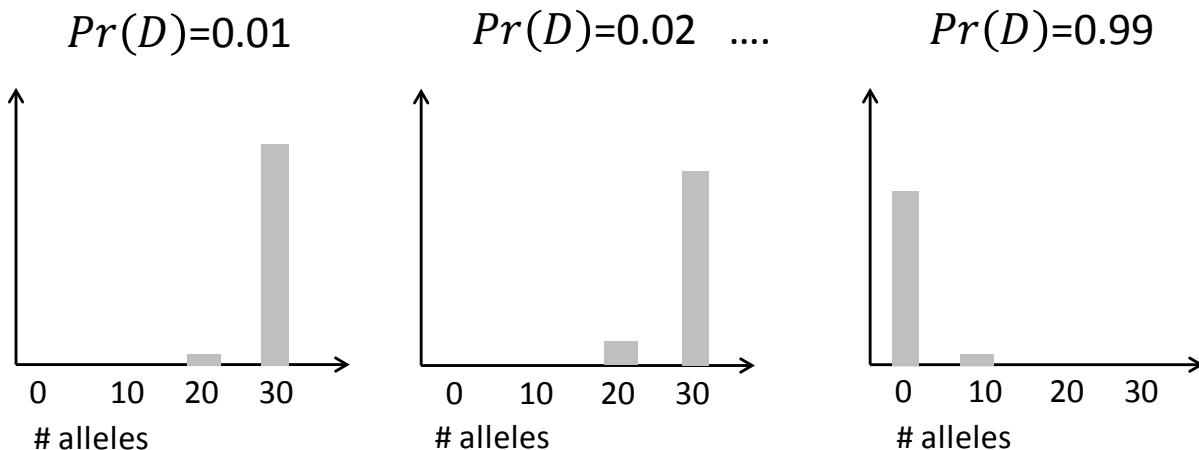
Monte Carlo simulations of the number of allele drop-outs for different values of  $Pr(D)$  produce the most plausible range for  $Pr(D)$ .



305

# $Pr(D)$ : LRmix

Monte Carlo simulations of the number of allele drop-outs for different values of  $Pr(D)$  produce the most plausible range for  $Pr(D)$ .

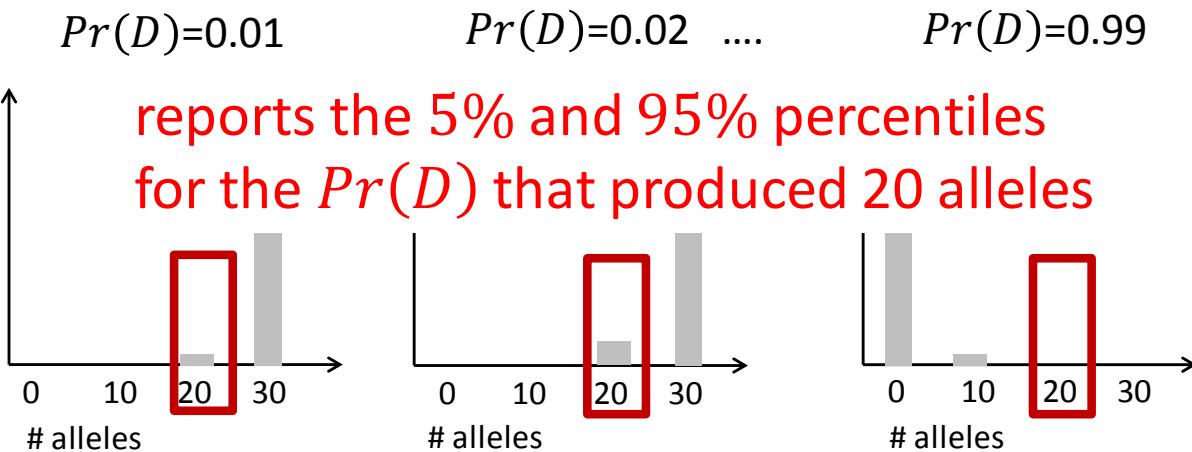


305

# $Pr(D)$ : LRmix

Monte Carlo simulations of the number of allele drop-outs for different values of  $Pr(D)$  produce the most plausible range for  $Pr(D)$ .

$G_C$  has 20 alleles

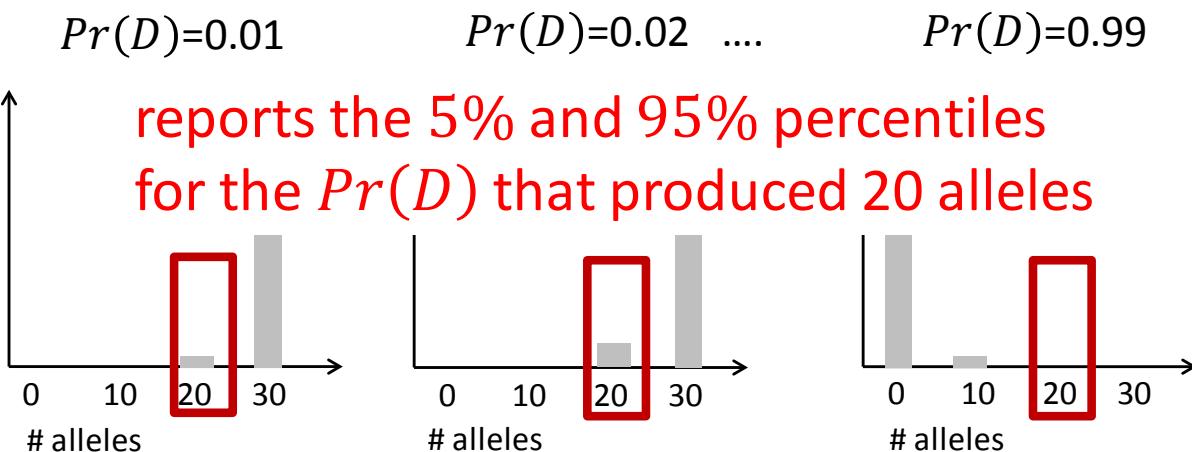


306

# $Pr(D)$ : LRmix

Monte Carlo simulations of the number of allele drop-outs for different values of  $Pr(D)$  produce the most plausible range for  $Pr(D)$ .

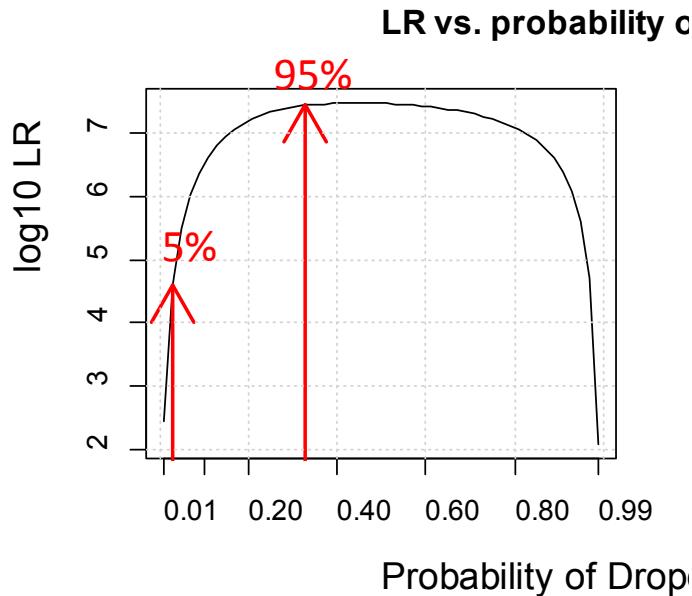
$G_C$  has 20 alleles



306

# $Pr(D)$ : LRmix

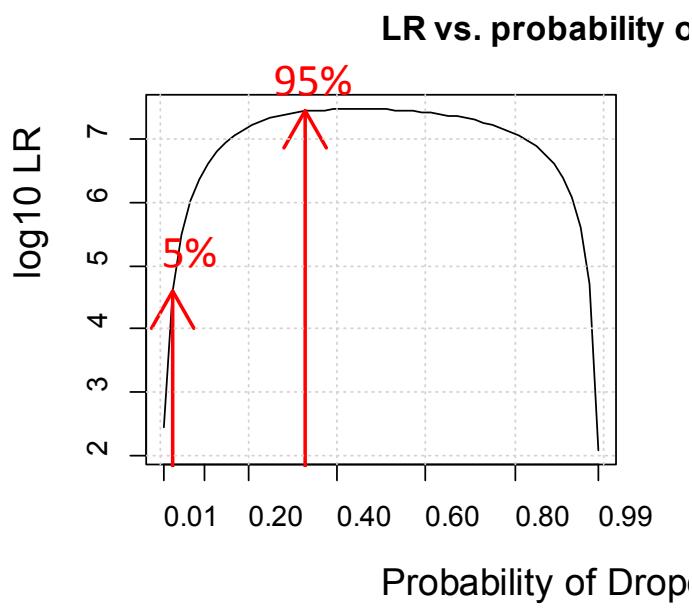
Within this range,  $Pr(D)$  is set equal to the value producing the smallest  $LR$ .



307

# $Pr(D)$ : LRmix

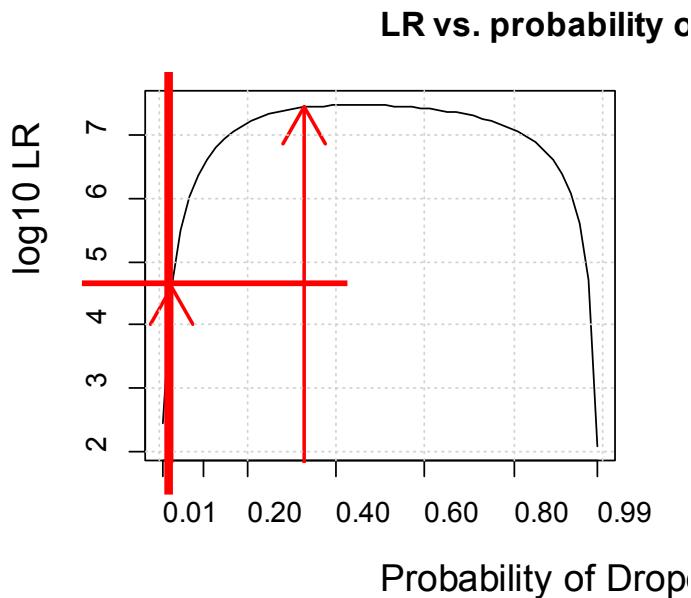
Within this range,  $Pr(D)$  is set equal to the value producing the smallest  $LR$ .



307

# $Pr(D)$ : LRmix

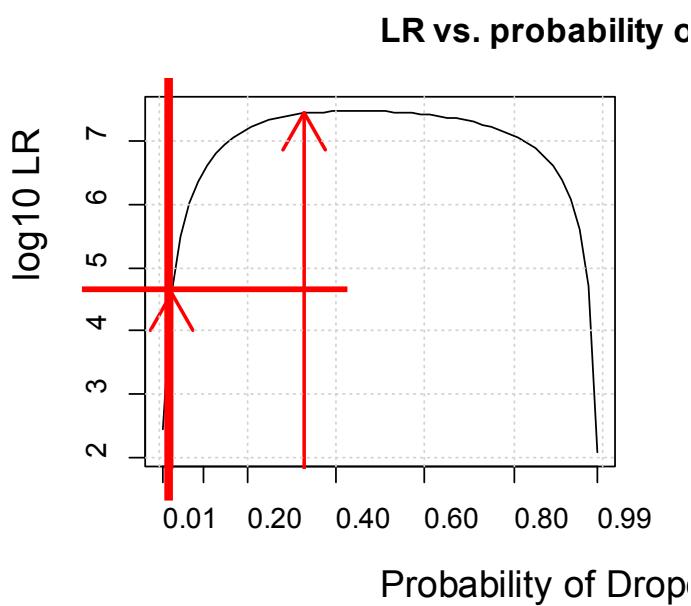
Within this range,  $Pr(D)$  is set equal to the value producing the smallest  $LR$ .



308

# $Pr(D)$ : LRmix

Within this range,  $Pr(D)$  is set equal to the value producing the smallest  $LR$ .



308

Lab Retriever and LRmix

Demo

309

Lab Retriever and LRmix

Demo

309



# Real Case Example

The victim died from stab wounds.



$G_C$ : bloodstains found as diluted vertical drips around the kitchen sink



2 suspects

Reference: K.E. Lohmueller and N. Rudin. Calculating the Weight of Evidence in Low-Template Forensic DNA Casework. *J Forensic Sci* 2013; 58: S243-S249. 310



# Real Case Example

The victim died from stab wounds.



$G_C$ : bloodstains found as diluted vertical drips around the kitchen sink



2 suspects

Reference: K.E. Lohmueller and N. Rudin. Calculating the Weight of Evidence in Low-Template Forensic DNA Casework. *J Forensic Sci* 2013; 58: S243-S249. 310

# Real Case Example



Locus	Major Peaks	Minor Peaks	Minor Peaks at Stutter Positions	Victim	Suspect	Alternate Suspect
D8S1179	13, 14	15	12	13, 14	13, 15	13, 15
D21S11	30, 31.2	32.2	29, 30.2	30, 31.2	30, 31.2	30, 32.2
D7S820	11, 12	9	10	11, 12	8, 11	9, 11
CSF1PO	10, 13	11, 12*	9	10, 13	12, 13	11, 12
D3S1358	15, 17	16†	14	15, 17	15	16
TH01	7, 8	9	6	7, 8	8, 9.3	8, 9
D13S317	11	8, 12	10	11	10, 11	8, 12
D16S539	11, 13	9	10, 12	11, 13	11, 14	9, 11
D2S1338	17, 21	19	16, 20	17, 21	17, 26	16, 19
D19S435	13	14	12	13	13	13, 14
vWA	17	15, 18	16‡	17	17	15, 18
TPOX	8, 9	10		8, 9	9, 11	9, 10
D18S51	13, 17		12, 16	13, 17	13, 17	17
D5S818	11, 12		10	11, 12	11	11, 12
FGA	20, 21	24	19	20, 21	21, 24	19, 24

**Table 1** in Lohmueller and Rudin. Calculating the Weight of Evidence in Low-Template Forensic DNA Casework. *J Forensic Sci* 2013; 58: S243-S249.

311

# Real Case Example

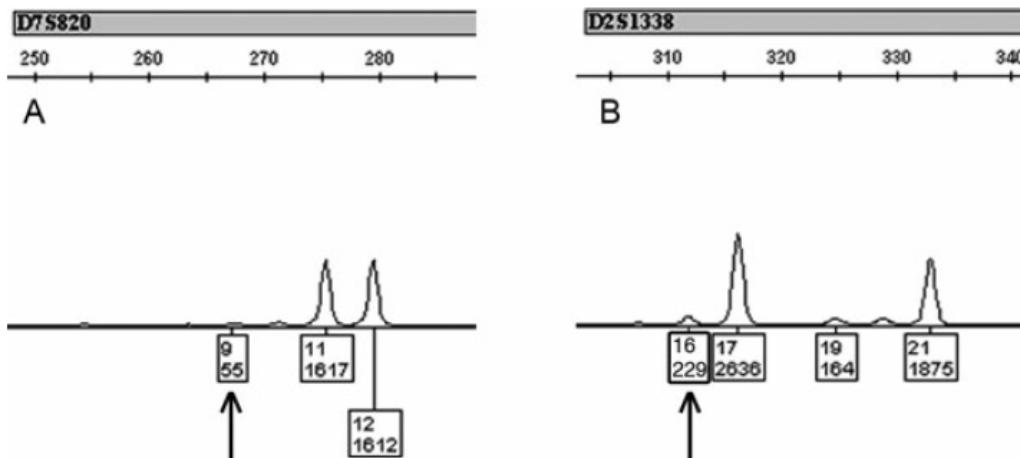


Locus	Major Peaks	Minor Peaks	Minor Peaks at Stutter Positions	Victim	Suspect	Alternate Suspect
D8S1179	13, 14	15	12	13, 14	13, 15	13, 15
D21S11	30, 31.2	32.2	29, 30.2	30, 31.2	30, 31.2	30, 32.2
D7S820	11, 12	9	10	11, 12	8, 11	9, 11
CSF1PO	10, 13	11, 12*	9	10, 13	12, 13	11, 12
D3S1358	15, 17	16†	14	15, 17	15	16
TH01	7, 8	9	6	7, 8	8, 9.3	8, 9
D13S317	11	8, 12	10	11	10, 11	8, 12
D16S539	11, 13	9	10, 12	11, 13	11, 14	9, 11
D2S1338	17, 21	19	16, 20	17, 21	17, 26	16, 19
D19S435	13	14	12	13	13	13, 14
vWA	17	15, 18	16‡	17	17	15, 18
TPOX	8, 9	10		8, 9	9, 11	9, 10
D18S51	13, 17		12, 16	13, 17	13, 17	17
D5S818	11, 12		10	11, 12	11	11, 12
FGA	20, 21	24	19	20, 21	21, 24	19, 24

**Table 1** in Lohmueller and Rudin. Calculating the Weight of Evidence in Low-Template Forensic DNA Casework. *J Forensic Sci* 2013; 58: S243-S249.

311

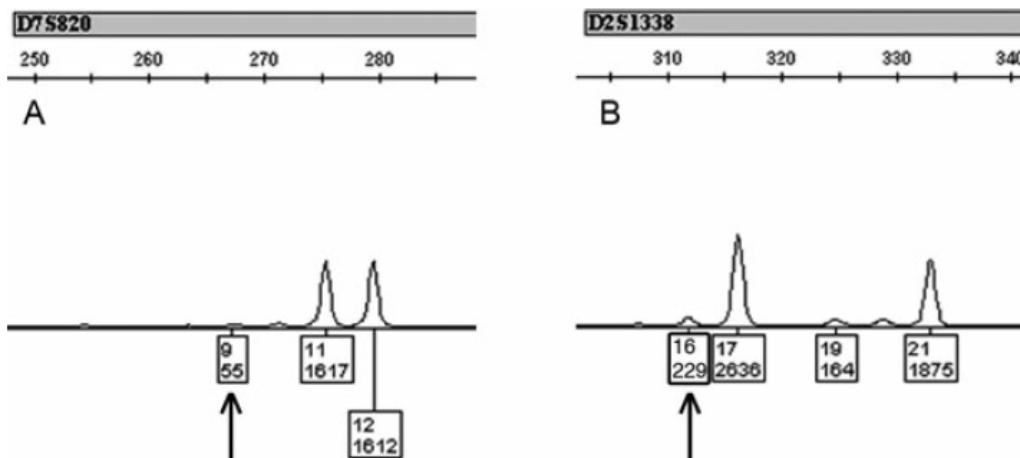
# Real Case Example



**Figure 1** in Lohmueller and Rudin. Calculating the Weight of Evidence in Low-Template Forensic DNA Casework. *J Forensic Sci* 2013; 58: S243-S249.

312

# Real Case Example



**Figure 1** in Lohmueller and Rudin. Calculating the Weight of Evidence in Low-Template Forensic DNA Casework. *J Forensic Sci* 2013; 58: S243-S249.

312

# Real Case Example

Conclusions of the laboratory:

1. Exclusion of suspect 1 and suspect 2 as contributors to the mixture
2. Suspect 2 is “included”: CPI = 0.16
3. LR for propositions:

$H_p$ : The bloodstain contains the DNA from the victim and suspect 2.

$H_d$ : The bloodstain contains the DNA from the victim and an unknown unrelated person.

- Binary model:  $LR = 10,000$
- Semi-continuous model:  $LR \geq 47 \text{ billion}$

Reference: Lohmueller and Rudin. Calculating the Weight of Evidence in Low-Template Forensic DNA Casework. *J Forensic Sci* 2013; 58: S243-S249.

313

# Real Case Example

Conclusions of the laboratory:

1. Exclusion of suspect 1 and suspect 2 as contributors to the mixture
2. Suspect 2 is “included”: CPI = 0.16
3. LR for propositions:

$H_p$ : The bloodstain contains the DNA from the victim and suspect 2.

$H_d$ : The bloodstain contains the DNA from the victim and an unknown unrelated person.

- Binary model:  $LR = 10,000$
- Semi-continuous model:  $LR \geq 47 \text{ billion}$

Reference: Lohmueller and Rudin. Calculating the Weight of Evidence in Low-Template Forensic DNA Casework. *J Forensic Sci* 2013; 58: S243-S249.

313

## **Missing Person Calculations**

Many of the issues involved in missing person calculations are the same as those for paternity. Instead of a paternal allele being known, a biological sample from the missing person is available.

Suppose a person is missing. 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 spouse  $S$  and child  $C$  of the missing person.

314

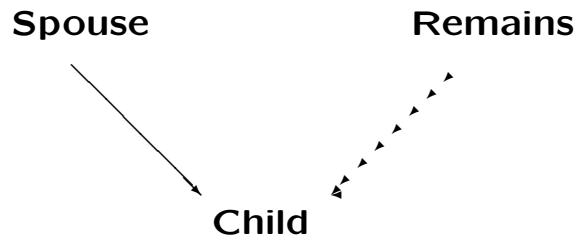
## **Missing Person Calculations**

Many of the issues involved in missing person calculations are the same as those for paternity. Instead of a paternal allele being known, a biological sample from the missing person is available.

Suppose a person is missing. 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 spouse  $S$  and child  $C$  of the missing person.

314

## Missing Person Calculations



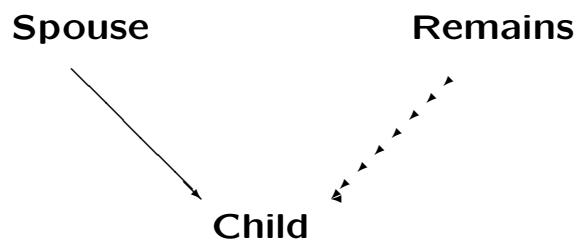
Two explanations of the evidence are:

$H_p$ : the remains are from the missing person.

$H_d$ : the remains are not from the missing person.

315

## Missing Person Calculations



Two explanations of the evidence are:

$H_p$ : the remains are from the missing person.

$H_d$ : the remains are not from the missing person.

315

## Missing Person Likelihood Ratio

Helpful to work with probabilities of genotypes conditional on those in the previous generation(s):

$$\begin{aligned} L &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\ &= \frac{\Pr(G_C, G_S, G_X|H_p)}{\Pr(G_C, G_S, G_X|H_d)} \\ &= \frac{\Pr(G_C|G_S, G_X, H_p) \Pr(G_S, G_X|H_p)}{\Pr(G_C|G_S, G_X, H_d) \Pr(G_S, G_X|H_d)} \\ &= \frac{\Pr(G_C|G_S, G_X, H_p)}{\Pr(G_C|G_S, H_d)} \end{aligned}$$

since the genotype of the child does not depend on that of  $X$  when  $H_d$  is true (any low-level relatedness within the population is being ignored).

316

## Missing Person Likelihood Ratio

Helpful to work with probabilities of genotypes conditional on those in the previous generation(s):

$$\begin{aligned} L &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\ &= \frac{\Pr(G_C, G_S, G_X|H_p)}{\Pr(G_C, G_S, G_X|H_d)} \\ &= \frac{\Pr(G_C|G_S, G_X, H_p) \Pr(G_S, G_X|H_p)}{\Pr(G_C|G_S, G_X, H_d) \Pr(G_S, G_X|H_d)} \\ &= \frac{\Pr(G_C|G_S, G_X, H_p)}{\Pr(G_C|G_S, H_d)} \end{aligned}$$

since the genotype of the child does not depend on that of  $X$  when  $H_d$  is true (any low-level relatedness within the population is being ignored).

316

## **Missing Person Likelihood Ratios**

The likelihood ratios are the same as in the paternity case where  $X$  is alleged to be the father of child  $C$  who has mother  $S$ .

Similar extensions can be made to allow for  $X$  to be a relative of the missing person, or to allow for relatedness among all members of a population.

317

## **Missing Person Likelihood Ratios**

The likelihood ratios are the same as in the paternity case where  $X$  is alleged to be the father of child  $C$  who has mother  $S$ .

Similar extensions can be made to allow for  $X$  to be a relative of the missing person, or to allow for relatedness among all members of a population.

317

## **Additional Family Typings**

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  $P, Q$  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.

318

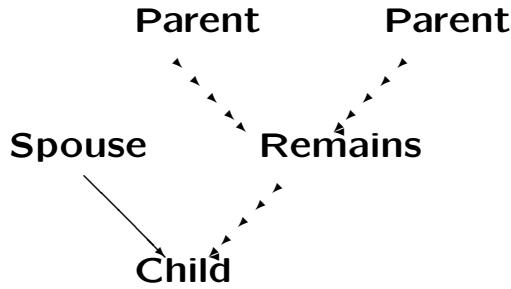
## **Additional Family Typings**

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  $P, Q$  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.

318

## Additional Missing Person Calculations



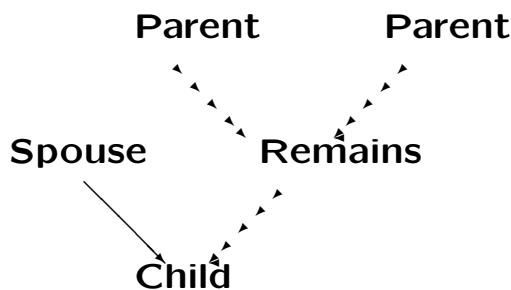
Two explanations of the evidence are:

$H_p$ : the remains are from the missing person.

$H_d$ : the remains are not from the missing person.

319

## Additional Missing Person Calculations



Two explanations of the evidence are:

$H_p$ : the remains are from the missing person.

$H_d$ : the remains are not from the missing person.

319

## Additional Family Typings

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,  $C, S, X, P, Q$  represent the genotypes of the child, the remains, the spouse and the parents of the missing person.

320

## Additional Family Typings

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,  $C, S, X, P, Q$  represent the genotypes of the child, the remains, the spouse and the parents of the missing person.

320

## Additional Family Typings

$$\begin{aligned}
L &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\
&= \frac{\Pr(C,S,X,P,Q|H_p)}{\Pr(C,S,X,P,Q|H_d)} \\
&= \frac{\Pr(C|S,X,P,Q,H_p) \Pr(S,X,P,Q|H_p)}{\Pr(C|S,X,P,Q,H_d) \Pr(S,X,P,Q|H_d)} \\
&= \frac{\Pr(C|S,X,H_p) \Pr(S|H_p) \Pr(X|P,Q,H_p)}{\Pr(C|S,P,Q,H_d) \Pr(S|H_d) \Pr(X|H_d)} \\
&= \frac{\Pr(C|S,X,H_p) \Pr(X|P,Q,H_p)}{\Pr(C|S,P,Q,H_d) \Pr(X)}
\end{aligned}$$

321

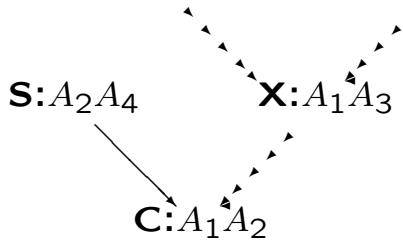
## Additional Family Typings

$$\begin{aligned}
L &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\
&= \frac{\Pr(C,S,X,P,Q|H_p)}{\Pr(C,S,X,P,Q|H_d)} \\
&= \frac{\Pr(C|S,X,P,Q,H_p) \Pr(S,X,P,Q|H_p)}{\Pr(C|S,X,P,Q,H_d) \Pr(S,X,P,Q|H_d)} \\
&= \frac{\Pr(C|S,X,H_p) \Pr(S|H_p) \Pr(X|P,Q,H_p)}{\Pr(C|S,P,Q,H_d) \Pr(S|H_d) \Pr(X|H_d)} \\
&= \frac{\Pr(C|S,X,H_p) \Pr(X|P,Q,H_p)}{\Pr(C|S,P,Q,H_d) \Pr(X)}
\end{aligned}$$

321

## Example of Additional Family Typings

$$\mathbf{P}:A_1A_5 \quad \mathbf{Q}:A_3A_6$$

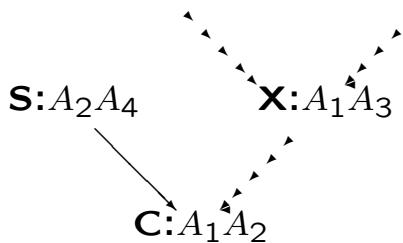


$$\begin{aligned}
 \Pr(C|S, X, H_p) &= 1/4 \\
 \Pr(X|P, Q, H_p) &= 1/4 \\
 \Pr(C|S, P, Q, H_d) &= 1/8 \\
 \Pr(X|H_d) &= 2p_1p_3 \\
 L &= \frac{1}{4p_1p_3}
 \end{aligned}$$

322

## Example of Additional Family Typings

$$\mathbf{P}:A_1A_5 \quad \mathbf{Q}:A_3A_6$$



$$\begin{aligned}
 \Pr(C|S, X, H_p) &= 1/4 \\
 \Pr(X|P, Q, H_p) &= 1/4 \\
 \Pr(C|S, P, Q, H_d) &= 1/8 \\
 \Pr(X|H_d) &= 2p_1p_3 \\
 L &= \frac{1}{4p_1p_3}
 \end{aligned}$$

322

## Missing Person Example

A DNA profile was extracted from a body  $X$  that had been burnt beyond recognition. The wife (LG), child (EW), and two brothers (JR, JW) of a missing man were typed.

$H_p$ :  $X$  is brother of JR and JW, and father of EW

$H_d$ :  $X$  is unrelated to JR, JW and EW

$E$ : the five profiles PM, EW, LG, JR, JW

The strength of the evidence is quantified by

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

and this is evaluated by conditioning the profile of EW on that of LG (and, under  $H_p$ , on that of  $X$ ), and conditioning the profiles of JR, JW (and, under  $H_p$ , of  $X$ ) on GM and GF.

323

## Missing Person Example

A DNA profile was extracted from a body  $X$  that had been burnt beyond recognition. The wife (LG), child (EW), and two brothers (JR, JW) of a missing man were typed.

$H_p$ :  $X$  is brother of JR and JW, and father of EW

$H_d$ :  $X$  is unrelated to JR, JW and EW

$E$ : the five profiles PM, EW, LG, JR, JW

The strength of the evidence is quantified by

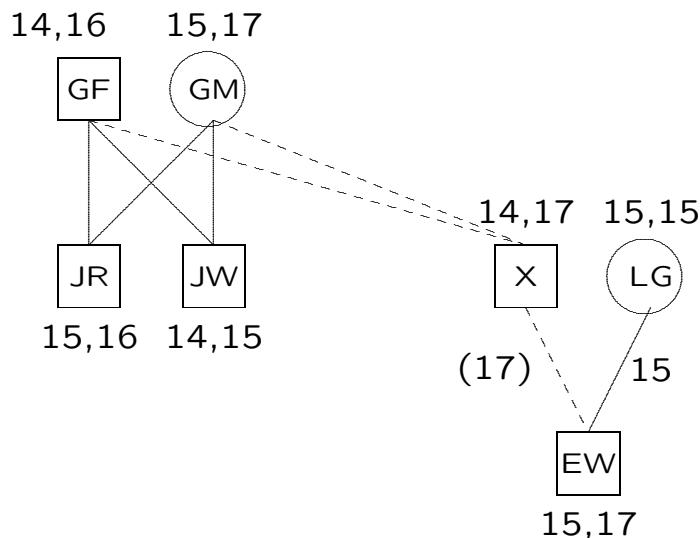
$$\text{LR} = \frac{\Pr(E|H_p)}{\Pr(E|H_d)}$$

and this is evaluated by conditioning the profile of EW on that of LG (and, under  $H_p$ , on that of  $X$ ), and conditioning the profiles of JR, JW (and, under  $H_p$ , of  $X$ ) on GM and GF.

323

## D3S1358 Evidence

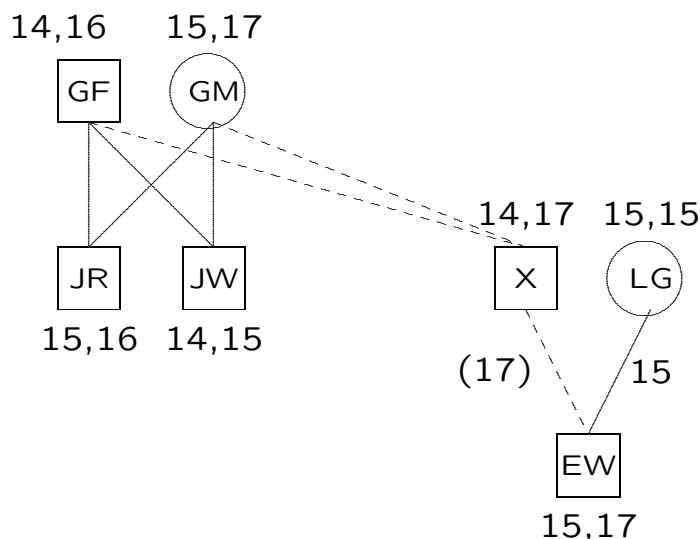
The profiles of JR, JM, EW determine those of GM, GF or of GF, GM.



324

## D3S1358 Evidence

The profiles of JR, JM, EW determine those of GM, GF or of GF, GM.



324

## D3S1358 Evidence

Under  $H_p$

$$\begin{aligned}\Pr(E|H_p) &= \Pr(X, EW, LG, JR, JW|H_p) \\ &= \Pr(EW|LG, X) \Pr(LG) \Pr(X, JR, JW|GF, GM) \Pr(GF, GM) \\ &= \left(\frac{1}{2}\right) \Pr(LG) \left(\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}\right) [2(2p_{14}p_{16})(2p_{15}p_{17})] \\ &= \Pr(LG)p_{14}p_{15}p_{16}p_{17}/16\end{aligned}$$

Under  $H_d$

$$\begin{aligned}\Pr(E|H_d) &= \Pr(X, EW, LG, JR, JW|H_d) \\ &= \Pr(EW|LG, GM, GF) \Pr(LG) \Pr(X) \Pr(JR, JW|GF, GM) \Pr(GF, GM) \\ &= \left(\frac{1}{4}\right) \Pr(LG) (2p_{14}p_{17}) \left(\frac{1}{4} \times \frac{1}{4}\right) [2(2p_{14}p_{16})(2p_{15}p_{17})] \\ &= \Pr(LG)p_{14}^2p_{15}p_{16}p_{17}^2/4\end{aligned}$$

so that

$$LR = \frac{1}{4p_{14}p_{17}}$$

325

## D3S1358 Evidence

Under  $H_p$

$$\begin{aligned}\Pr(E|H_p) &= \Pr(X, EW, LG, JR, JW|H_p) \\ &= \Pr(EW|LG, X) \Pr(LG) \Pr(X, JR, JW|GF, GM) \Pr(GF, GM) \\ &= \left(\frac{1}{2}\right) \Pr(LG) \left(\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}\right) [2(2p_{14}p_{16})(2p_{15}p_{17})] \\ &= \Pr(LG)p_{14}p_{15}p_{16}p_{17}/16\end{aligned}$$

Under  $H_d$

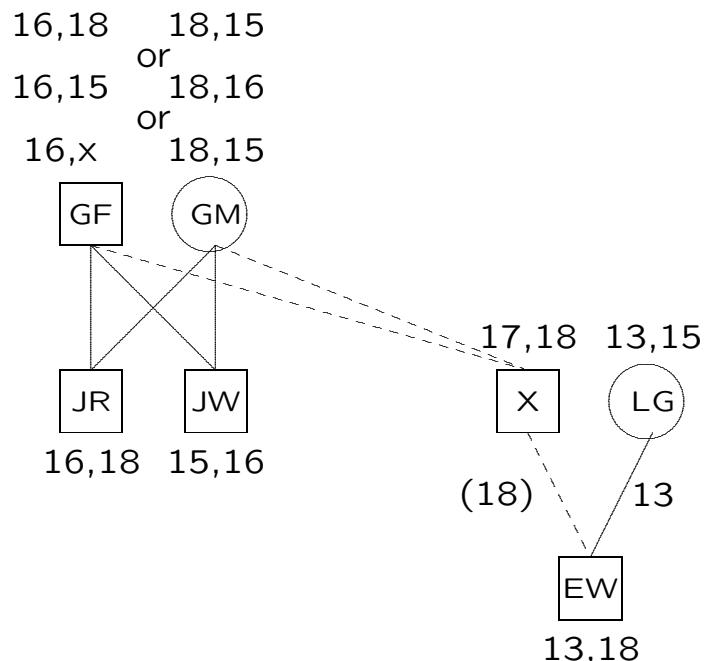
$$\begin{aligned}\Pr(E|H_d) &= \Pr(X, EW, LG, JR, JW|H_d) \\ &= \Pr(EW|LG, GM, GF) \Pr(LG) \Pr(X) \Pr(JR, JW|GF, GM) \Pr(GF, GM) \\ &= \left(\frac{1}{4}\right) \Pr(LG) (2p_{14}p_{17}) \left(\frac{1}{4} \times \frac{1}{4}\right) [2(2p_{14}p_{16})(2p_{15}p_{17})] \\ &= \Pr(LG)p_{14}^2p_{15}p_{16}p_{17}^2/4\end{aligned}$$

so that

$$LR = \frac{1}{4p_{14}p_{17}}$$

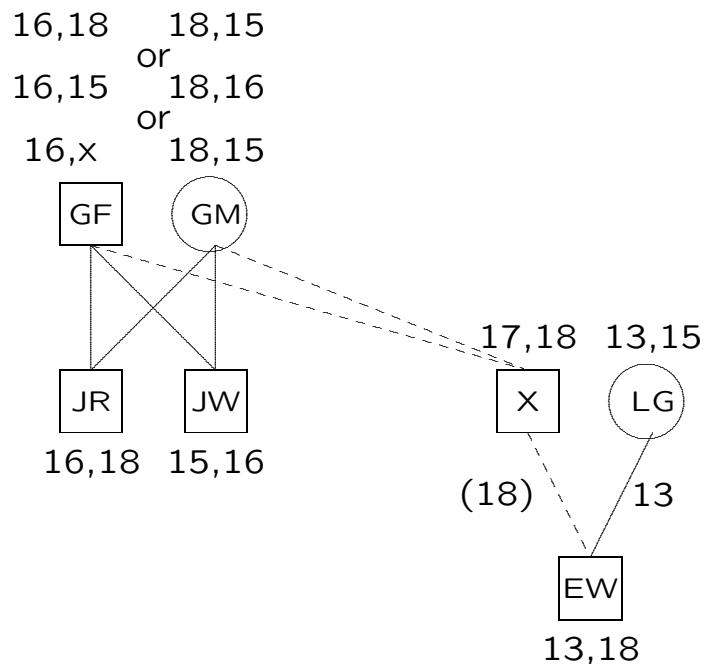
325

## vWA Evidence



326

## vWA Evidence



326

## vWA Evidence

Under  $H_p$ , the parents of X must have been of types 16,17 and 18,15 (or vice versa) so

$$\begin{aligned}\Pr(E) &= \Pr(X, EW, LG, JR, JW | H_p) \\ &= \Pr(EW | LG, X) \Pr(LG) \Pr(X, JR, JW | GF, GM) \Pr(GF, GM) \\ &= \left(\frac{1}{4}\right)\left(\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}\right) \Pr(LG)[2(2p_{16}p_{17})(2p_{18}p_{15})] \\ &= \Pr(LG)p_{15}p_{16}p_{17}p_{18}/32\end{aligned}$$

Under  $H_d$ , the grandparental genotypes are not completely specified so

$$\begin{aligned}\Pr(E) &= \Pr(X, EW, LG, JR, JW | H_p) \\ &= \Pr(X) \Pr(LG) \sum_{GM, GF} \Pr(EW | LG, GM, GF) \Pr(JR, JW | GF, GM) \Pr(GF, GM) \\ &= (2p_{17}p_{18}) \Pr(LG) 2[\frac{1}{4} \frac{1}{4} \frac{1}{8} (2p_{16}p_{15})(2p_{18}p_{16}) + \frac{1}{2} \frac{1}{4} \frac{1}{8} (p_{16}^2)(2p_{18}p_{15}) \\ &\quad + \frac{1}{4} \frac{1}{4} \frac{1}{4} (2p_{16}p_{18})(2p_{18}p_{15}) + \frac{1}{4} \frac{1}{4} \frac{1}{8} (2p_{16}(1 - p_{16} - p_{18}))(2p_{18}p_{15})] \\ &= \Pr(LG)p_{15}p_{16}p_{17}p_{18}^2(1 + 2p_{16} + p_{18})/8\end{aligned}$$

so that  $LR = 1/[4p_{17}p_{18}(1 + 2p_{16} + p_{18})]$ .

327

## vWA Evidence

Under  $H_p$ , the parents of X must have been of types 16,17 and 18,15 (or vice versa) so

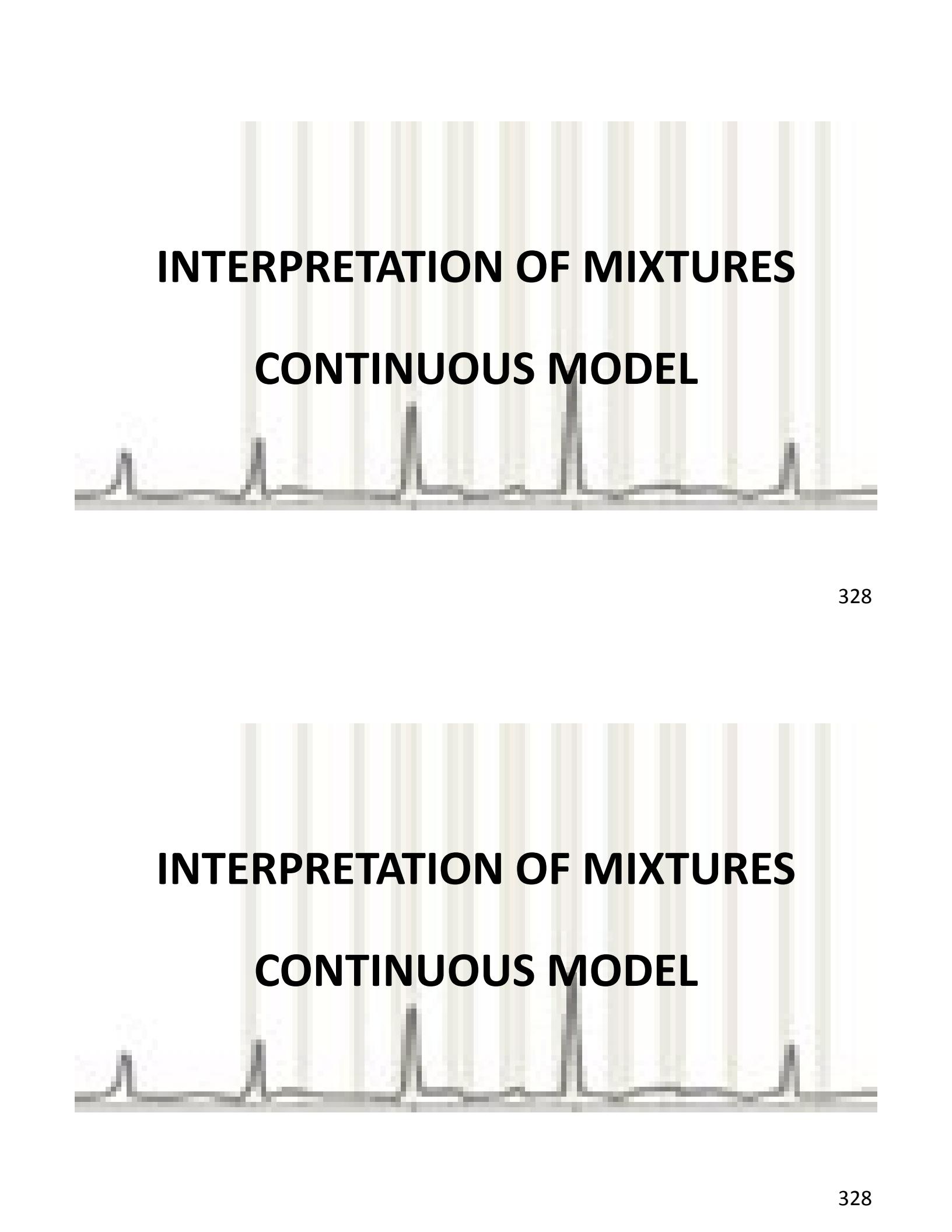
$$\begin{aligned}\Pr(E) &= \Pr(X, EW, LG, JR, JW | H_p) \\ &= \Pr(EW | LG, X) \Pr(LG) \Pr(X, JR, JW | GF, GM) \Pr(GF, GM) \\ &= \left(\frac{1}{4}\right)\left(\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}\right) \Pr(LG)[2(2p_{16}p_{17})(2p_{18}p_{15})] \\ &= \Pr(LG)p_{15}p_{16}p_{17}p_{18}/32\end{aligned}$$

Under  $H_d$ , the grandparental genotypes are not completely specified so

$$\begin{aligned}\Pr(E) &= \Pr(X, EW, LG, JR, JW | H_p) \\ &= \Pr(X) \Pr(LG) \sum_{GM, GF} \Pr(EW | LG, GM, GF) \Pr(JR, JW | GF, GM) \Pr(GF, GM) \\ &= (2p_{17}p_{18}) \Pr(LG) 2[\frac{1}{4} \frac{1}{4} \frac{1}{8} (2p_{16}p_{15})(2p_{18}p_{16}) + \frac{1}{2} \frac{1}{4} \frac{1}{8} (p_{16}^2)(2p_{18}p_{15}) \\ &\quad + \frac{1}{4} \frac{1}{4} \frac{1}{4} (2p_{16}p_{18})(2p_{18}p_{15}) + \frac{1}{4} \frac{1}{4} \frac{1}{8} (2p_{16}(1 - p_{16} - p_{18}))(2p_{18}p_{15})] \\ &= \Pr(LG)p_{15}p_{16}p_{17}p_{18}^2(1 + 2p_{16} + p_{18})/8\end{aligned}$$

so that  $LR = 1/[4p_{17}p_{18}(1 + 2p_{16} + p_{18})]$ .

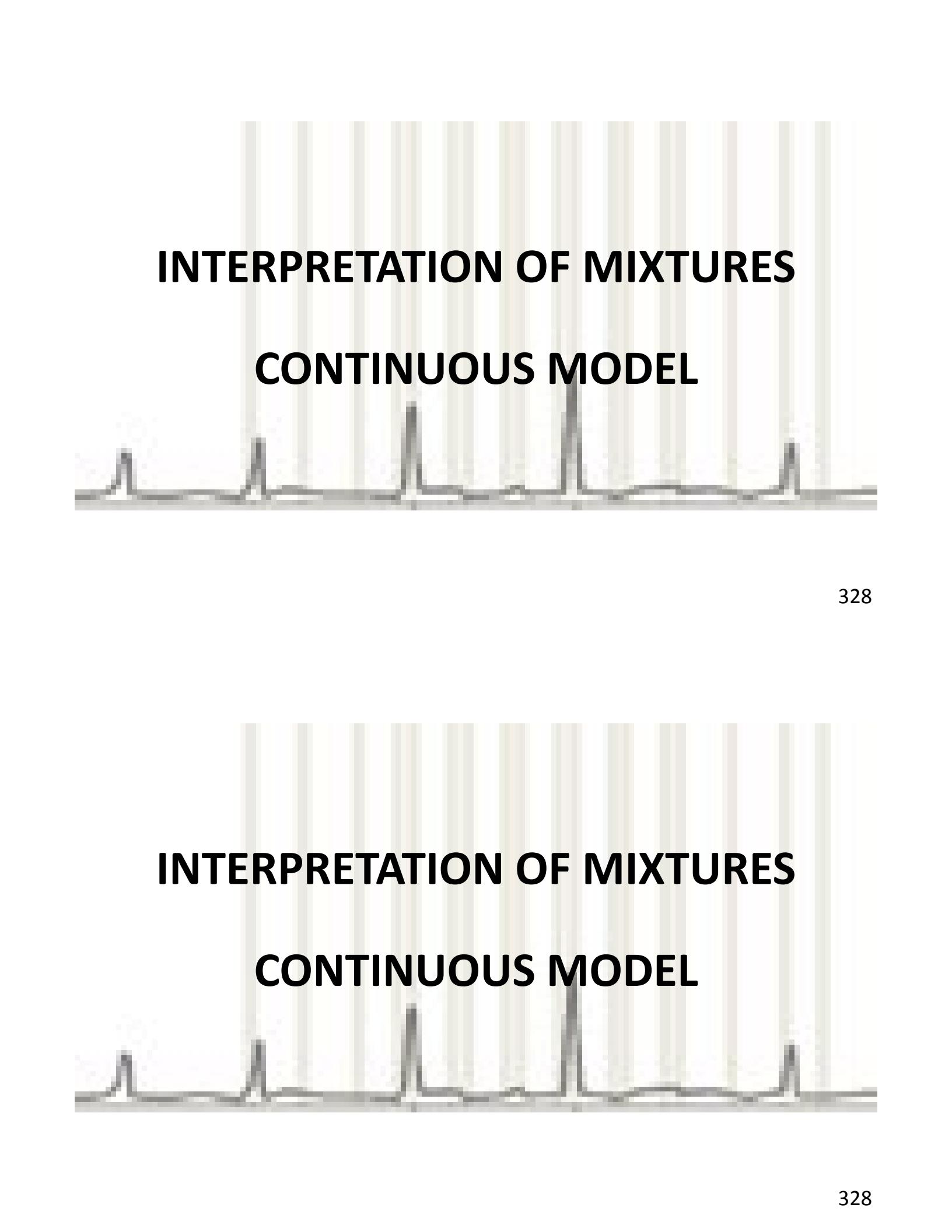
327



# **INTERPRETATION OF MIXTURES**

## **CONTINUOUS MODEL**

328



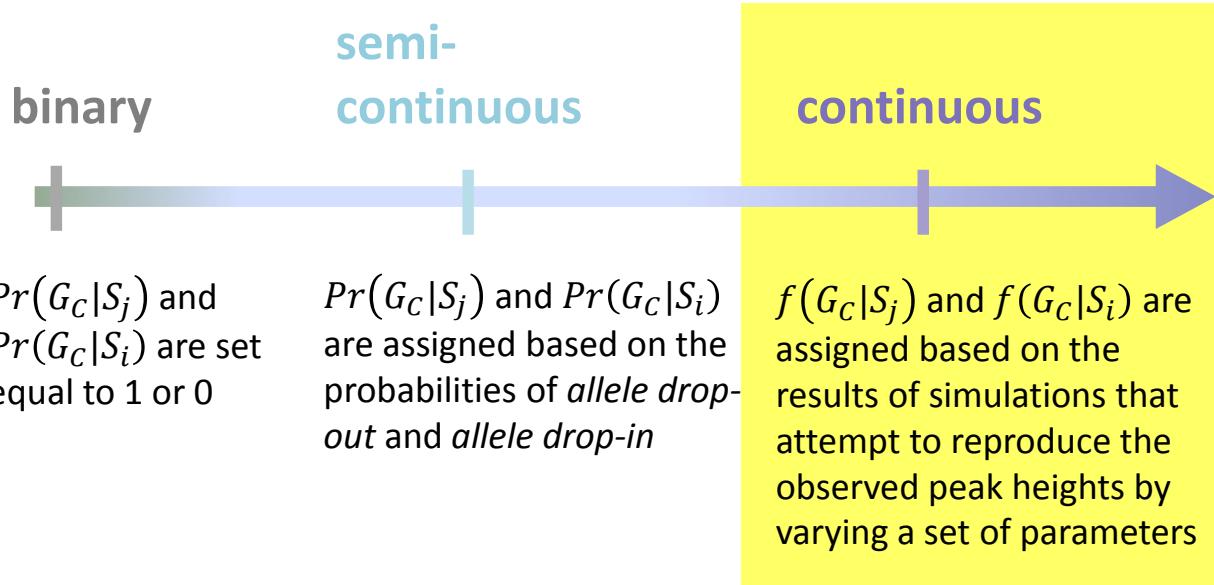
# **INTERPRETATION OF MIXTURES**

## **CONTINUOUS MODEL**

328

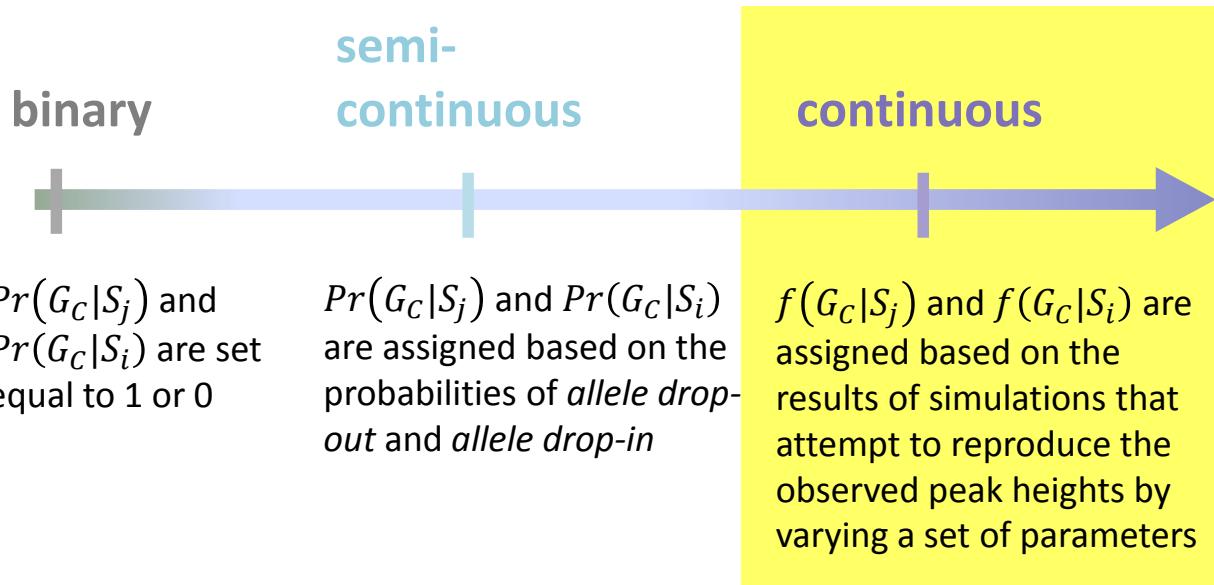
# Likelihood Ratio

$$LR = \frac{\sum_{j=1}^M f(G_C|S_j) \Pr(S_j|G_K^p, H_p)}{\sum_{i=1}^N f(G_C|S_i) \Pr(S_i|G_K^d, H_d)} \quad \text{where } M \leq N$$



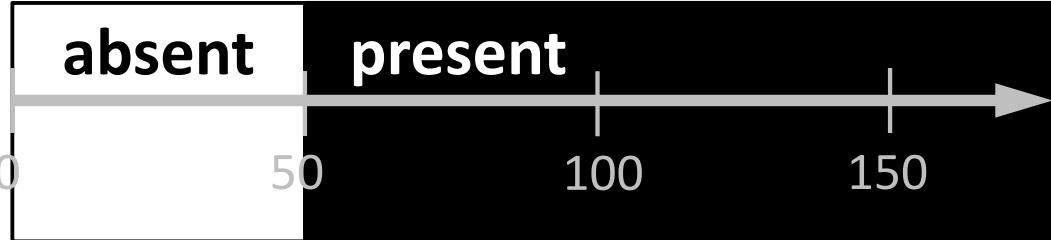
# Likelihood Ratio

$$LR = \frac{\sum_{j=1}^M f(G_C|S_j) \Pr(S_j|G_K^p, H_p)}{\sum_{i=1}^N f(G_C|S_i) \Pr(S_i|G_K^d, H_d)} \quad \text{where } M \leq N$$



# Discrete vs. Continuous

- The observed peaks as a **discrete** variable:



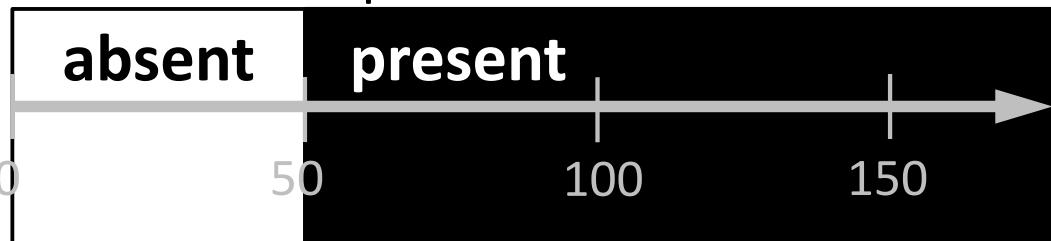
- The observed peaks as a **continuous** variable:



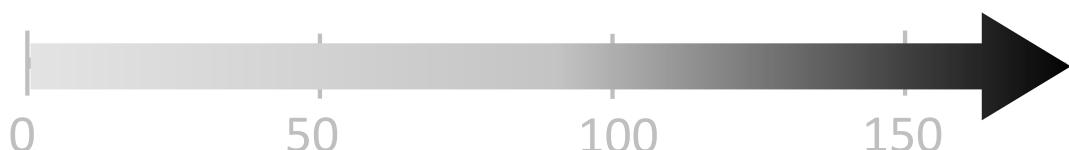
330

# Discrete vs. Continuous

- The observed peaks as a **discrete** variable:



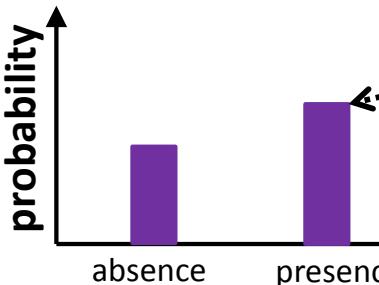
- The observed peaks as a **continuous** variable:



330

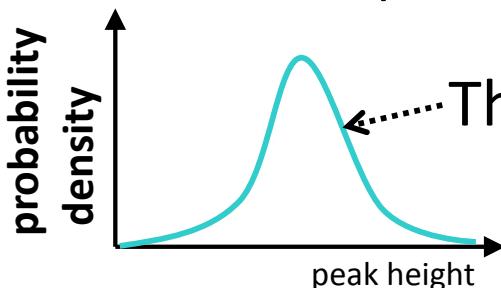
# Discrete vs. Continuous

- The observed peaks as a **discrete** variable:



This is a **probability**.

- The observed peaks as a **continuous** variable:

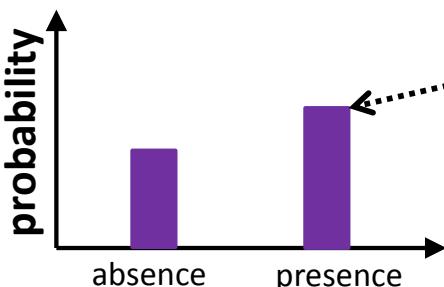


This is a **probability density**.

331

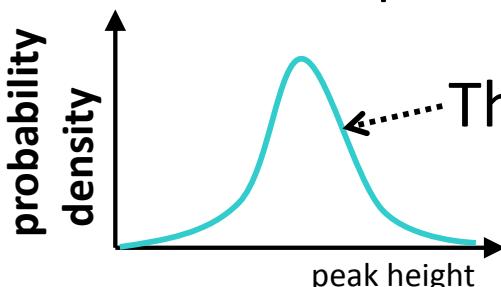
# Discrete vs. Continuous

- The observed peaks as a **discrete** variable:



This is a **probability**.

- The observed peaks as a **continuous** variable:



This is a **probability density**.

331

# Continuous Model

- The peak heights are a continuous variable.
- We therefore cannot speak of a probability  $Pr(G_C|S_i)$ , but must speak of a probability density  $f(G_C|S_i)$ .
- The probability densities  $f(G_C|S_i)$  are assigned based on the results of simulations that attempt to reproduce the observed peak heights by varying a set of parameters.

332

# Continuous Model

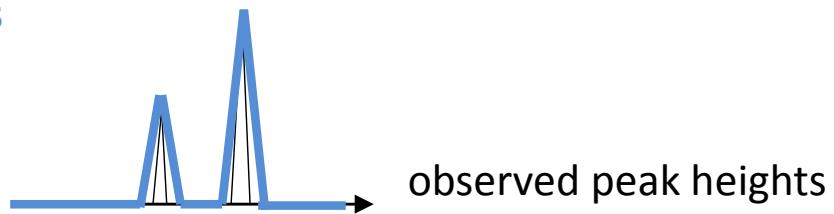
- The peak heights are a continuous variable.
- We therefore cannot speak of a probability  $Pr(G_C|S_i)$ , but must speak of a probability density  $f(G_C|S_i)$ .
- The probability densities  $f(G_C|S_i)$  are assigned based on the results of simulations that attempt to reproduce the observed peak heights by varying a set of parameters.

332

# Continuous Model

Simulations attempt to reproduce the observed peak heights by varying a set of parameters.

simulated peak heights



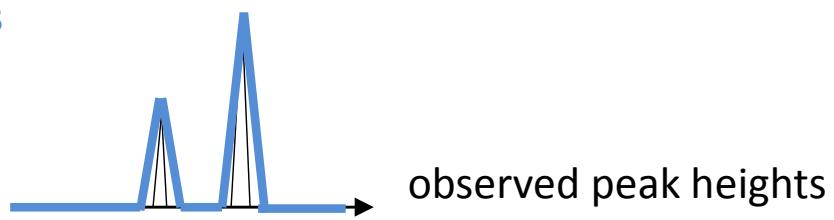
The better  $S_i$  and the set of parameters explain  $G_C$ , the greater the probability density  $f(G_C|S_i)$ .

333

# Continuous Model

Simulations attempt to reproduce the observed peak heights by varying a set of parameters.

simulated peak heights



The better  $S_i$  and the set of parameters explain  $G_C$ , the greater the probability density  $f(G_C|S_i)$ .

333

# Continuous Model

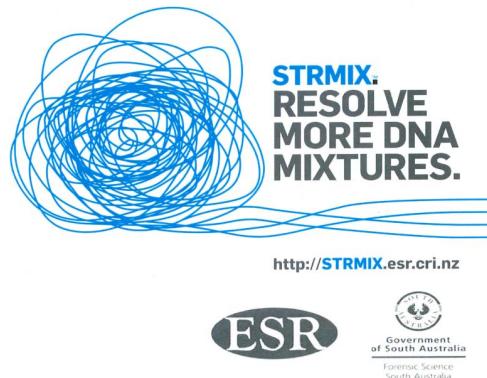
→ example STRmix™



Reference: D. Taylor, J. Bright, J. Buckleton. The interpretation of single source and mixed DNA profiles. *Forensic Sci Int Genet* 2013; 7: 516-528. 334

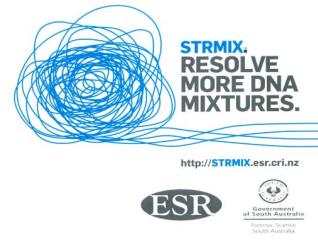
# Continuous Model

→ example STRmix™



Reference: D. Taylor, J. Bright, J. Buckleton. The interpretation of single source and mixed DNA profiles. *Forensic Sci Int Genet* 2013; 7: 516-528. 334

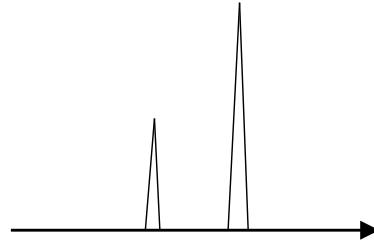
# Continuous Model



We have:

- a vector of observed peak heights

$$\mathbf{O} = \{O_{ar}^\ell\}$$



where  $a = \text{allele}$ ,  
 $\ell = \text{locus}$ ,  
 $r = \text{replicate}$ .

335

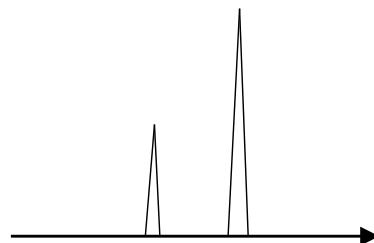
# Continuous Model



We have:

- a vector of observed peak heights

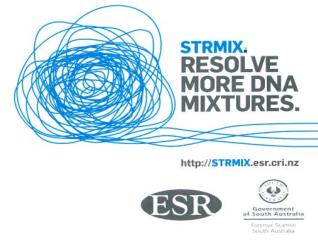
$$\mathbf{O} = \{O_{ar}^\ell\}$$



where  $a = \text{allele}$ ,  
 $\ell = \text{locus}$ ,  
 $r = \text{replicate}$ .

335

# Continuous Model



We have:

- a vector of expected peak heights obtained from the STRmix™ model in function of
  - genotype set  $S_i$
  - a set of mass parameters  $\mathbf{M}$

$$\mathbf{E} = \{E_{ar}^\ell\}$$



where  $a = \text{allele}$ ,  $\ell = \text{locus}$ ,  $r = \text{replicate}$ .

336

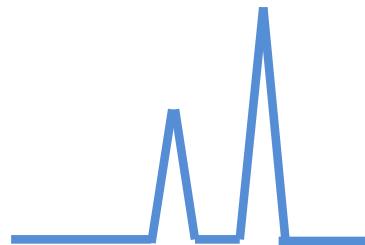
# Continuous Model



We have:

- a vector of expected peak heights obtained from the STRmix™ model in function of
  - genotype set  $S_i$
  - a set of mass parameters  $\mathbf{M}$

$$\mathbf{E} = \{E_{ar}^\ell\}$$



where  $a = \text{allele}$ ,  $\ell = \text{locus}$ ,  $r = \text{replicate}$ .

336

# Continuous Model



ESR  
Government of South Australia  
Executive Scientific Research Institute

Mathematical model:

- The probability density  $f(G_C|S_i)$  is equal to the probability density of the observed peak heights given the expected peak heights modeled in function of genotype set  $S_i$  and the set of mass parameters  $\mathbf{M}$ .

$$f(G_C|S_i) = \int_{\mathbf{M}} f(\mathbf{O}|S_i, \mathbf{M}) \Pr(\mathbf{M}) d\mathbf{M}$$

337

# Continuous Model



ESR  
Government of South Australia  
Executive Scientific Research Institute

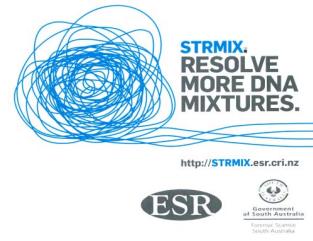
Mathematical model:

- The probability density  $f(G_C|S_i)$  is equal to the probability density of the observed peak heights given the expected peak heights modeled in function of genotype set  $S_i$  and the set of mass parameters  $\mathbf{M}$ .

$$f(G_C|S_i) = \int_{\mathbf{M}} f(\mathbf{O}|S_i, \mathbf{M}) \Pr(\mathbf{M}) d\mathbf{M}$$

337

# Continuous Model



## Assumptions:

1. The observed peak heights  $O_{ar}^\ell$  are conditionally independent given  $S_i$  and  $\mathbf{M}$ .
2. The probability density  $f(G_C|S_i)$  for the entire profile is the product of the probability densities at each locus.

$$f(G_C|S_i) = \prod_{\ell} \int_M \prod_a \prod_r f(O_{ar}^\ell | S_i, \mathbf{M}) \Pr(\mathbf{M}) d\mathbf{M}$$

338

# Continuous Model



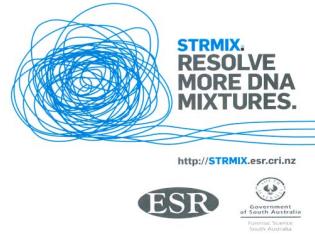
## Assumptions:

1. The observed peak heights  $O_{ar}^\ell$  are conditionally independent given  $S_i$  and  $\mathbf{M}$ .
2. The probability density  $f(G_C|S_i)$  for the entire profile is the product of the probability densities at each locus.

$$f(G_C|S_i) = \prod_{\ell} \int_M \prod_a \prod_r f(O_{ar}^\ell | S_i, \mathbf{M}) \Pr(\mathbf{M}) d\mathbf{M}$$

338

# Continuous Model



ESR  
Government of South Australia  
Forensic Science South Australia

The probability densities  $f(O_{ar}^\ell | S_i, \mathbf{M})$  are obtained by comparing the observed peak heights  $O_{ar}^\ell$  to the expected peak heights  $E_{ar}^\ell$  obtained from the STRmix™ model for genotype set  $S_i$  and mass parameters  $\mathbf{M}$ .

339

# Continuous Model



ESR  
Government of South Australia  
Forensic Science South Australia

The probability densities  $f(O_{ar}^\ell | S_i, \mathbf{M})$  are obtained by comparing the observed peak heights  $O_{ar}^\ell$  to the expected peak heights  $E_{ar}^\ell$  obtained from the STRmix™ model for genotype set  $S_i$  and mass parameters  $\mathbf{M}$ .

339

# Continuous Model



<http://STRMIX.esr.cri.nz>



1. Model the expected peak heights  $E_{ar}^\ell$  in function of genotype set  $S_i$  and mass parameters  $\mathbf{M}$ .
2. Compare the expected peak heights  $E_{ar}^\ell$  to the observed peak heights  $O_{ar}^\ell$ .
3. Perform a large number of simulations of steps 1. and 2. that randomly vary the genotype set  $S_i$  and mass parameters  $\mathbf{M}$  (this approximates the integration over  $\mathbf{M}$ ).
4. Assign  $f(G_C|S_i)$  based on the simulation results.

340

# Continuous Model



<http://STRMIX.esr.cri.nz>



1. Model the expected peak heights  $E_{ar}^\ell$  in function of genotype set  $S_i$  and mass parameters  $\mathbf{M}$ .
2. Compare the expected peak heights  $E_{ar}^\ell$  to the observed peak heights  $O_{ar}^\ell$ .
3. Perform a large number of simulations of steps 1. and 2. that randomly vary the genotype set  $S_i$  and mass parameters  $\mathbf{M}$  (this approximates the integration over  $\mathbf{M}$ ).
4. Assign  $f(G_C|S_i)$  based on the simulation results.

340

# Continuous Model



<http://STRMIX.esr.cri.nz>



1. Model the expected peak heights  $E_{ar}^\ell$  in function of genotype set  $S_i$  and mass parameters  $\mathbf{M}$ .
2. Compare the expected peak heights  $E_{ar}^\ell$  to the observed peak heights  $O_{ar}^\ell$ .
3. Perform a large number of simulations of steps 1. and 2. that randomly vary the genotype set  $S_i$  and mass parameters  $\mathbf{M}$  (this approximates the integration over  $\mathbf{M}$ ).
4. Assign  $f(G_C|S_i)$  based on the simulation results.

341

# Continuous Model



<http://STRMIX.esr.cri.nz>



1. Model the expected peak heights  $E_{ar}^\ell$  in function of genotype set  $S_i$  and mass parameters  $\mathbf{M}$ .
2. Compare the expected peak heights  $E_{ar}^\ell$  to the observed peak heights  $O_{ar}^\ell$ .
3. Perform a large number of simulations of steps 1. and 2. that randomly vary the genotype set  $S_i$  and mass parameters  $\mathbf{M}$  (this approximates the integration over  $\mathbf{M}$ ).
4. Assign  $f(G_C|S_i)$  based on the simulation results.

341

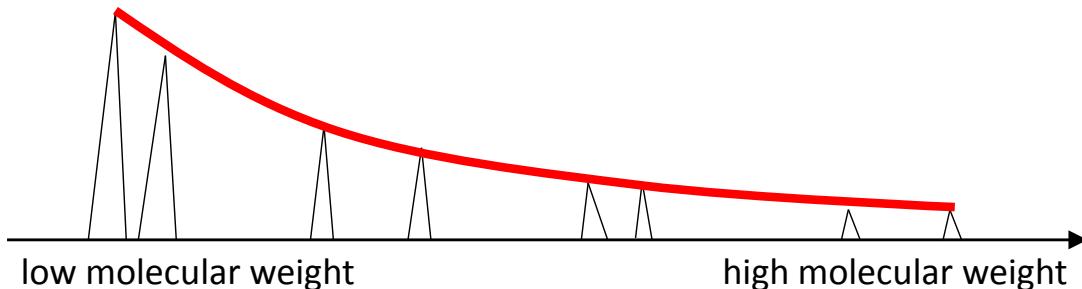
# 1. Expected Peak Heights



<http://STRMIX.esr.cri.nz>



The expected peak height of an allele is modeled as an exponential function of the allele's molecular weight.



$$y = \alpha_0 e^{-\alpha_1 x}$$

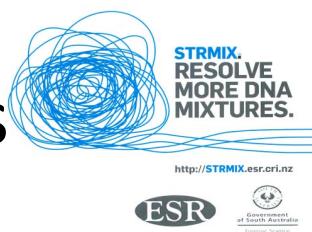
where:

$y$  = expected peak height of the allele

$x$  = the allele's molecular weight

342

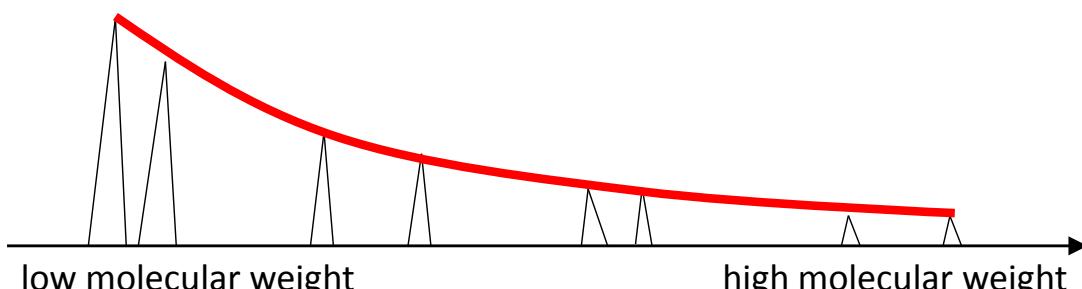
# 1. Expected Peak Heights



<http://STRMIX.esr.cri.nz>



The expected peak height of an allele is modeled as an exponential function of the allele's molecular weight.



$$y = \alpha_0 e^{-\alpha_1 x}$$

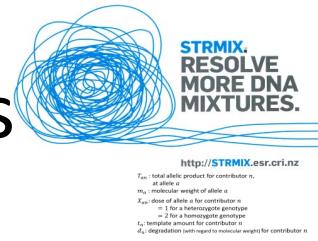
where:

$y$  = expected peak height of the allele

$x$  = the allele's molecular weight

342

# 1. Expected Peak Heights



1 replicate, 1 locus:

$$T_{an} = X_{an} t_n e^{-d_n m_a}$$

$T_{an}$  : total allelic product for contributor  $n$ ,  
at allele  $a$

$m_a$  : molecular weight of allele  $a$

$X_{an}$ : dose of allele  $a$  for contributor  $n$   
= 1 for a heterozygote genotype  
= 2 for a homozygote genotype

$t_n$ : template amount for contributor  $n$

$d_n$ : degradation (with regard to molecular weight) for contributor  $n$

343

# 1. Expected Peak Heights



1 replicate, 1 locus:

$$T_{an} = X_{an} t_n e^{-d_n m_a}$$

$T_{an}$  : total allelic product for contributor  $n$ ,  
at allele  $a$

$m_a$  : molecular weight of allele  $a$

$X_{an}$ : dose of allele  $a$  for contributor  $n$   
= 1 for a heterozygote genotype  
= 2 for a homozygote genotype

$t_n$ : template amount for contributor  $n$

$d_n$ : degradation (with regard to molecular weight) for contributor  $n$

343

# 1. Expected Peak Heights



ESR  
Government of South Australia  
Forensic Science South Australia

1 replicate:

$$T_{an}^\ell = A^\ell X_{an}^\ell t_n e^{-d_n m_a^\ell}$$

$T_{an}^\ell$  : total allelic product for contributor  $n$ ,  
at allele  $a$  of locus  $\ell$

$m_a^\ell$  : molecular weight of allele  $a$  of locus  $\ell$

$X_{an}^\ell$ : dose of allele  $a$  of locus  $\ell$  for contributor  $n$   
= 1 for a heterozygote genotype  
= 2 for a homozygote genotype

$t_n$ : template amount for contributor  $n$

$d_n$ : degradation (with regard to molecular weight) for contributor  $n$

$A^\ell$ : locus specific amplification efficiency

344

# 1. Expected Peak Heights



ESR  
Government of South Australia  
Forensic Science South Australia

1 replicate:

$$T_{an}^\ell = A^\ell X_{an}^\ell t_n e^{-d_n m_a^\ell}$$

$T_{an}^\ell$  : total allelic product for contributor  $n$ ,  
at allele  $a$  of locus  $\ell$

$m_a^\ell$  : molecular weight of allele  $a$  of locus  $\ell$

$X_{an}^\ell$ : dose of allele  $a$  of locus  $\ell$  for contributor  $n$   
= 1 for a heterozygote genotype  
= 2 for a homozygote genotype

$t_n$ : template amount for contributor  $n$

$d_n$ : degradation (with regard to molecular weight) for contributor  $n$

$A^\ell$ : locus specific amplification efficiency

344

# 1. Expected Peak Heights



ESR  
Government of South Australia  
Forensic Science South Australia

$$T_{anr}^{\ell} = R_r A^{\ell} X_{an}^{\ell} t_n e^{-d_n m_a^{\ell}}$$

$T_{anr}^{\ell}$  : total allelic product for contributor  $n$ ,

at allele  $a$  of locus  $\ell$  of replicate  $r$

$m_a^{\ell}$  : molecular weight of allele  $a$  of locus  $\ell$

$X_{an}^{\ell}$ : dose of allele  $a$  of locus  $\ell$  for contributor  $n$

= 1 for a heterozygote genotype

= 2 for a homozygote genotype

$t_n$ : template amount for contributor  $n$

$d_n$ : degradation (with regard to molecular weight) for contributor  $n$

$A^{\ell}$ : locus specific amplification efficiency

$R_r$ : parameter for taking into account variations between replicates

345

# 1. Expected Peak Heights



ESR  
Government of South Australia  
Forensic Science South Australia

$$T_{anr}^{\ell} = R_r A^{\ell} X_{an}^{\ell} t_n e^{-d_n m_a^{\ell}}$$

$T_{anr}^{\ell}$  : total allelic product for contributor  $n$ ,

at allele  $a$  of locus  $\ell$  of replicate  $r$

$m_a^{\ell}$  : molecular weight of allele  $a$  of locus  $\ell$

$X_{an}^{\ell}$ : dose of allele  $a$  of locus  $\ell$  for contributor  $n$

= 1 for a heterozygote genotype

= 2 for a homozygote genotype

$t_n$ : template amount for contributor  $n$

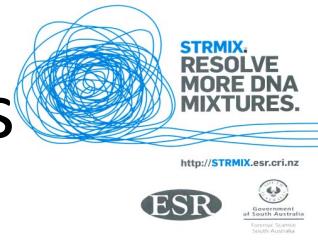
$d_n$ : degradation (with regard to molecular weight) for contributor  $n$

$A^{\ell}$ : locus specific amplification efficiency

$R_r$ : parameter for taking into account variations between replicates

345

# 1. Expected Peak Heights



total allelic product

$$T_{anr}^\ell$$

346

# 1. Expected Peak Heights

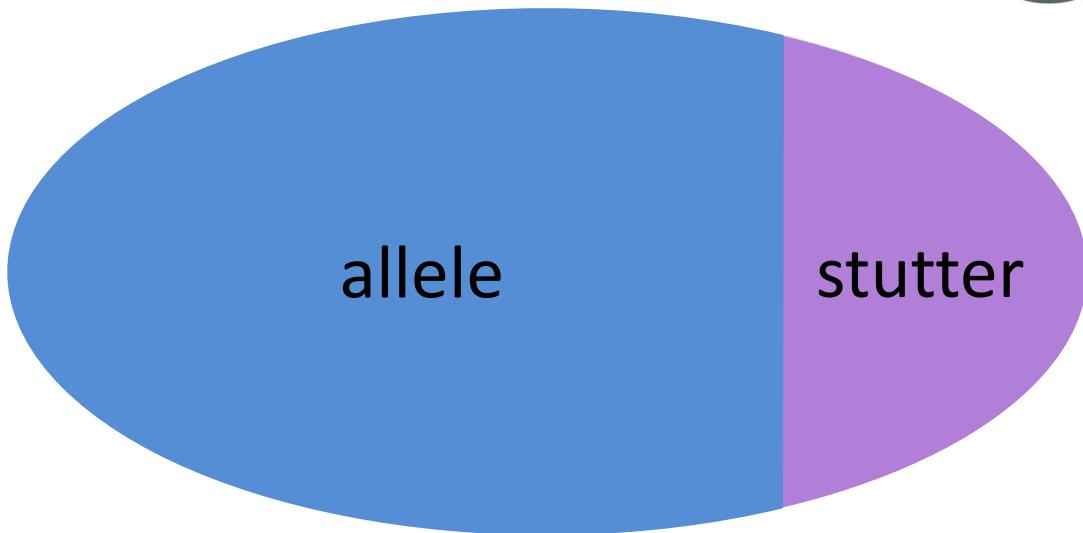


total allelic product

$$T_{anr}^\ell$$

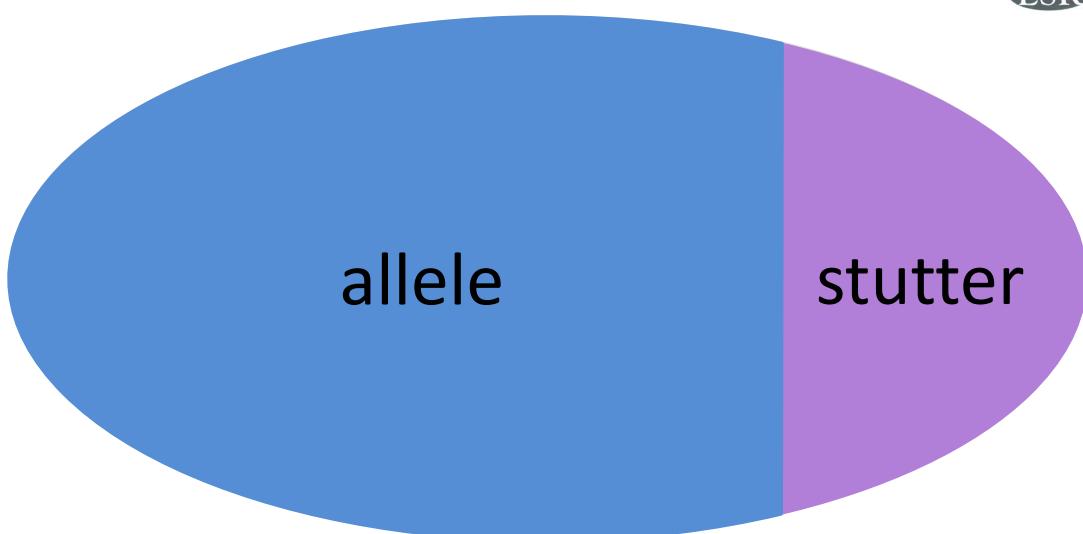
346

# 1. Expected Peak Heights



$$E_{anr}^\ell = \frac{T_{anr}^\ell}{1 + \pi_a^\ell}$$

$$E_{(a-1)nr}^\ell = \frac{\pi_a^\ell T_{anr}^\ell}{1 + \pi_a^\ell}$$



$$E_{anr}^\ell = \frac{T_{anr}^\ell}{1 + \pi_a^\ell}$$

$$E_{(a-1)nr}^\ell = \frac{\pi_a^\ell T_{anr}^\ell}{1 + \pi_a^\ell}$$

347

# 1. Expected Peak Heights



Add together the expected peak heights of the different contributors for each allele  $a$ , to obtain a single expected peak height for each allele, at each locus, and in each replicate.

$$E_{ar}^{\ell} = \sum_{n=1}^N E_{anr}^{\ell}$$

348

# 1. Expected Peak Heights



Add together the expected peak heights of the different contributors for each allele  $a$ , to obtain a single expected peak height for each allele, at each locus, and in each replicate.

$$E_{ar}^{\ell} = \sum_{n=1}^N E_{anr}^{\ell}$$

348

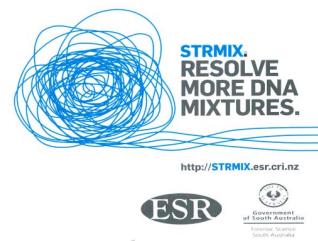
# Continuous Model



1. Model the expected peak heights  $E_{ar}^\ell$  in function of genotype set  $S_i$  and mass parameters  $\mathbf{M}$ .
2. Compare the expected peak heights  $E_{ar}^\ell$  to the observed peak heights  $O_{ar}^\ell$ .
3. Perform a large number of simulations of steps 1. and 2. that randomly vary the genotype set  $S_i$  and mass parameters  $\mathbf{M}$  (this approximates the integration over  $\mathbf{M}$ ).
4. Assign  $f(G_C|S_i)$  based on the simulation results.

349

# Continuous Model



1. Model the expected peak heights  $E_{ar}^\ell$  in function of genotype set  $S_i$  and mass parameters  $\mathbf{M}$ .
2. Compare the expected peak heights  $E_{ar}^\ell$  to the observed peak heights  $O_{ar}^\ell$ .
3. Perform a large number of simulations of steps 1. and 2. that randomly vary the genotype set  $S_i$  and mass parameters  $\mathbf{M}$  (this approximates the integration over  $\mathbf{M}$ ).
4. Assign  $f(G_C|S_i)$  based on the simulation results.

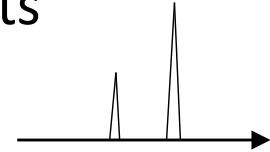
349

## 2. Compare $E$ to $O$

We have:

- a vector of observed peak heights

$$O = \{O_{ar}^{\ell}\}$$



- a vector of expected peak heights obtained from the STRmix™ model

$$E = \{E_{ar}^{\ell}\}$$



where  $a = \text{allele}$ ,  $\ell = \text{locus}$ ,  $r = \text{replicate}$ .

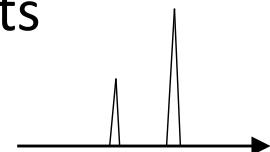
350

## 2. Compare $E$ to $O$

We have:

- a vector of observed peak heights

$$O = \{O_{ar}^{\ell}\}$$



- a vector of expected peak heights obtained from the STRmix™ model

$$E = \{E_{ar}^{\ell}\}$$



where  $a = \text{allele}$ ,  $\ell = \text{locus}$ ,  $r = \text{replicate}$ .

350

## 2. Compare $E$ to $O$

A continuous variable is defined:

$$\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right),$$

where

$O_{ar}^\ell$  = observed peak height for allele  $a$ , at locus  $\ell$ , in replicate  $r$

$E_{ar}^\ell$  = expected peak height obtained from the STRmix™ model for allele  $a$ , at locus  $\ell$ , in replicate  $r$

351

## 2. Compare $E$ to $O$

A continuous variable is defined:

$$\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right),$$

where

$O_{ar}^\ell$  = observed peak height for allele  $a$ , at locus  $\ell$ , in replicate  $r$

$E_{ar}^\ell$  = expected peak height obtained from the STRmix™ model for allele  $a$ , at locus  $\ell$ , in replicate  $r$

351

## 2. Compare $E$ to $O$

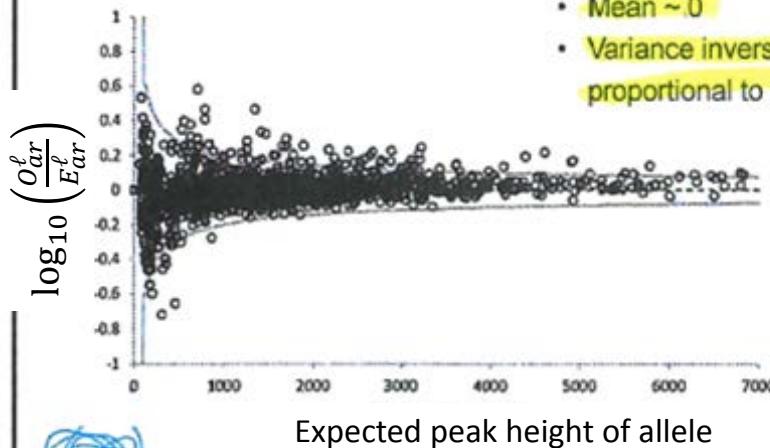
STRMIX.  
RESOLVE  
MORE DNA  
MIXTURES.

<http://STRMIX.esr.cri.nz>



### Variance of allele model

- Mean ~ 0
- Variance inversely proportional to  $E$



slide from Buckleton, Bright, McGovern's STRmix™ workshop in Phoenix, AZ, May 12-14, 2014

$$\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right) \text{ in function of } E_{ar}^\ell$$

352

## 2. Compare $E$ to $O$

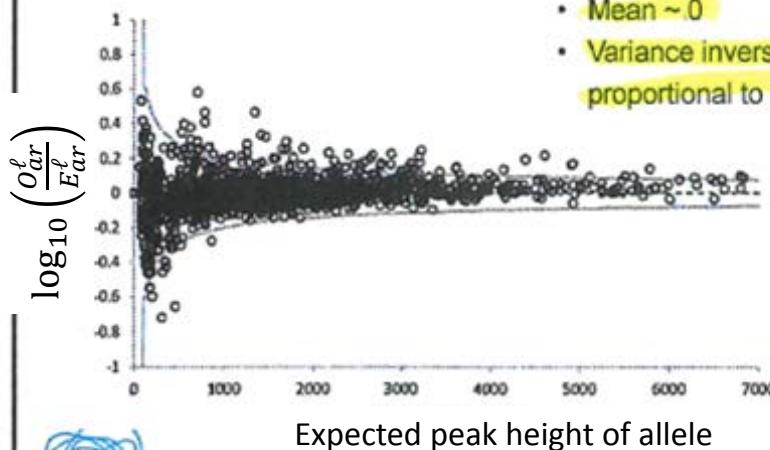
STRMIX.  
RESOLVE  
MORE DNA  
MIXTURES.

<http://STRMIX.esr.cri.nz>



### Variance of allele model

- Mean ~ 0
- Variance inversely proportional to  $E$



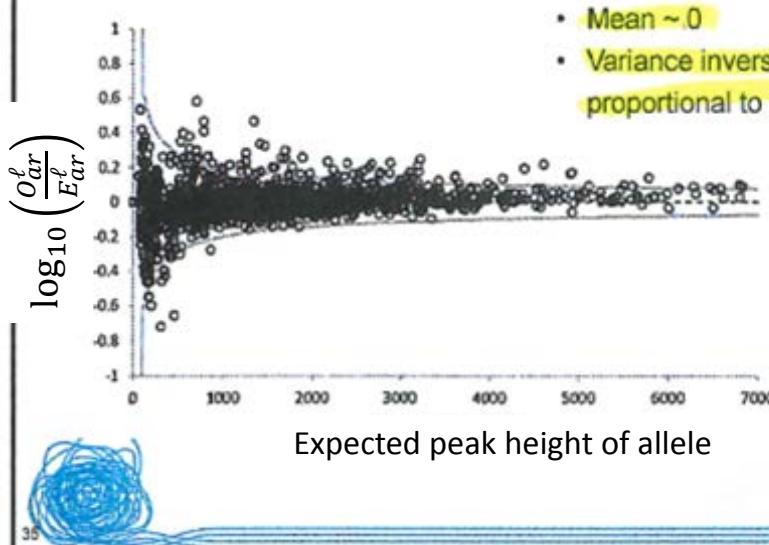
slide from Buckleton, Bright, McGovern's STRmix™ workshop in Phoenix, AZ, May 12-14, 2014

$$\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right) \text{ in function of } E_{ar}^\ell$$

352

## 2. Compare $E$ to $O$

### Variance of allele model



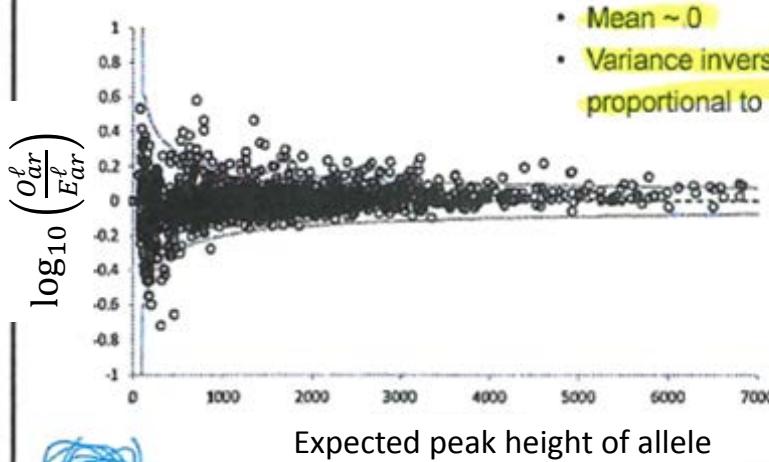
slide from Buckleton, Bright, McGovern's STRmix™ workshop in Phoenix, AZ, May 12-14, 2014

$$\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right) \sim N\left(0, \frac{c^2}{E_{ar}^\ell}\right)$$

353

## 2. Compare $E$ to $O$

### Variance of allele model

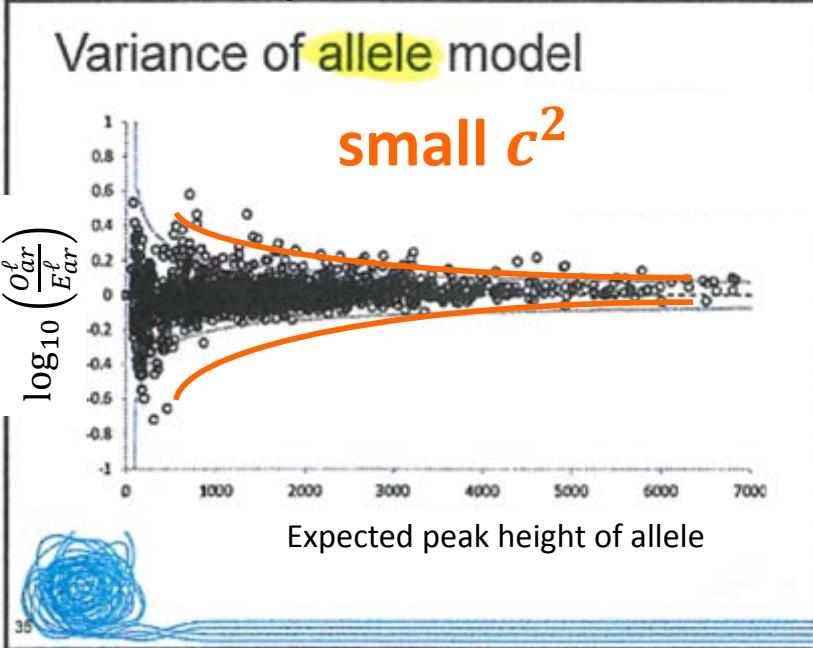


slide from Buckleton, Bright, McGovern's STRmix™ workshop in Phoenix, AZ, May 12-14, 2014

$$\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right) \sim N\left(0, \frac{c^2}{E_{ar}^\ell}\right)$$

353

## 2. Compare $E$ to $O$



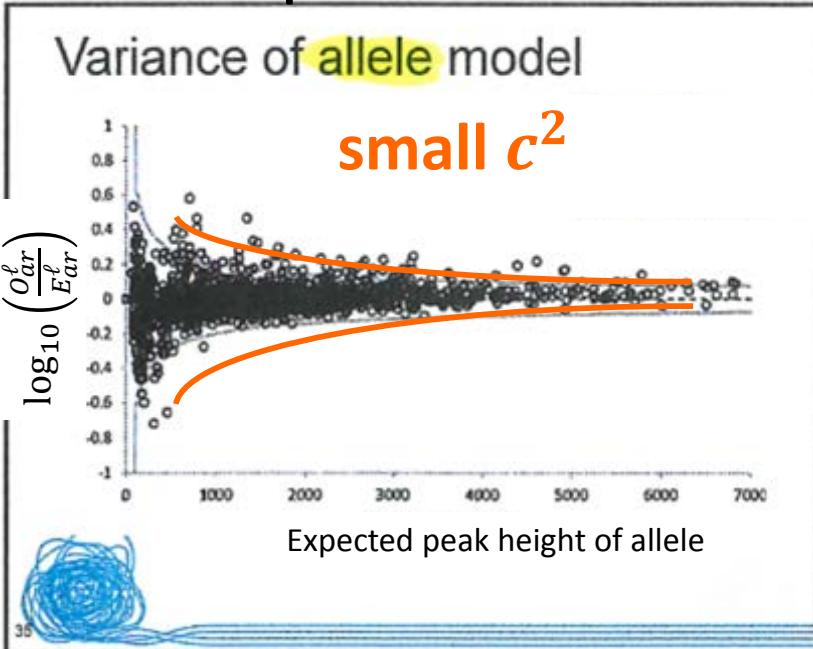
slide from Buckleton, Bright, McGovern's STRmix™ workshop in Phoenix, AZ, May 12-14, 2014

$$\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right) \sim N(0, \frac{c^2}{E_{ar}^\ell})$$

variance parameter

354

## 2. Compare $E$ to $O$



slide from Buckleton, Bright, McGovern's STRmix™ workshop in Phoenix, AZ, May 12-14, 2014

$$\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right) \sim N(0, \frac{c^2}{E_{ar}^\ell})$$

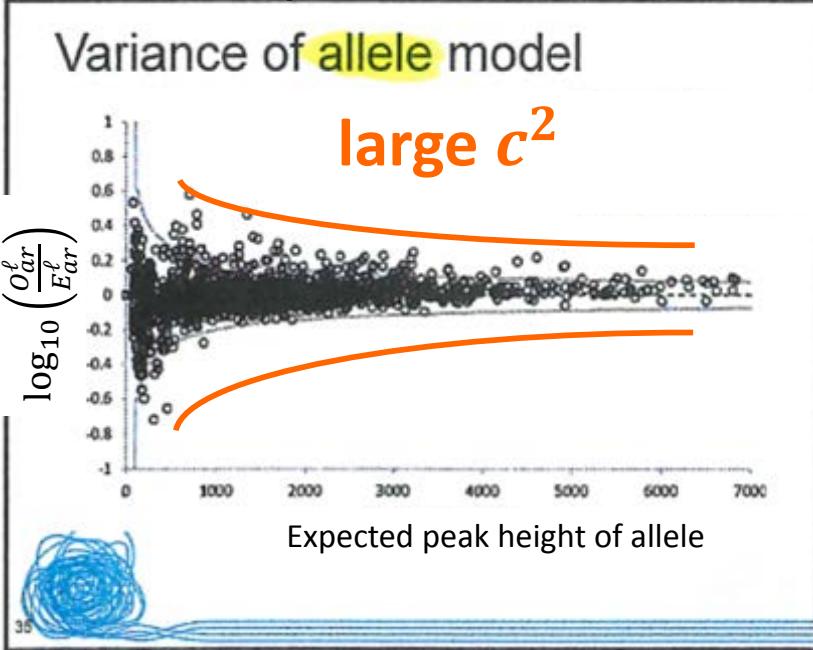
variance parameter

354

## 2. Compare $E$ to $O$



<http://STRMIX.esr.cri.nz>



slide from Buckleton, Bright, McGovern's STRmix™ workshop in Phoenix, AZ, May 12-14, 2014

$$\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right) \sim N(0, \frac{c^2}{E_{ar}^\ell})$$

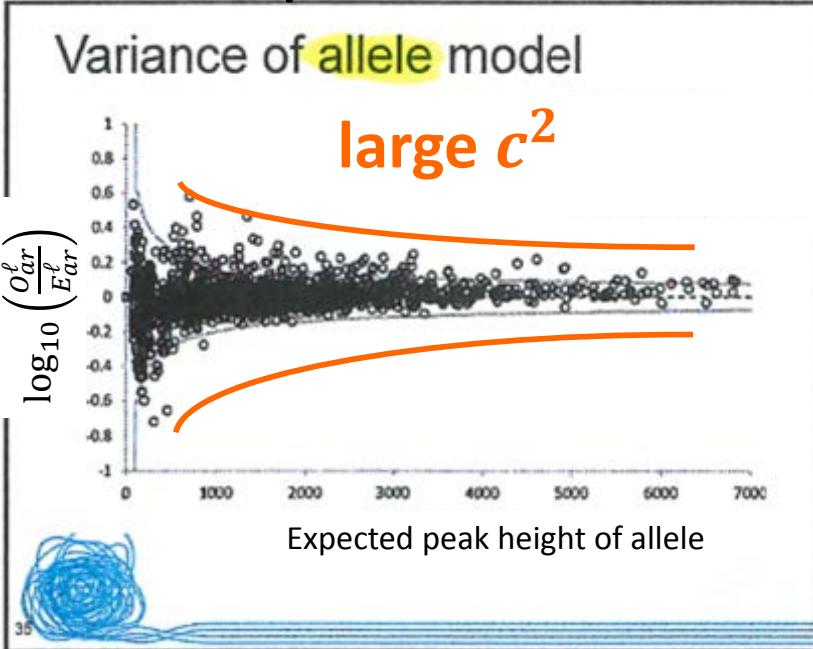
variance parameter

355

## 2. Compare $E$ to $O$



<http://STRMIX.esr.cri.nz>



slide from Buckleton, Bright, McGovern's STRmix™ workshop in Phoenix, AZ, May 12-14, 2014

$$\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right) \sim N(0, \frac{c^2}{E_{ar}^\ell})$$

variance parameter

355

## 2. Compare $E$ to $O$

Given the model

$$\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right) \sim N\left(0, \frac{c^2}{E_{ar}^\ell}\right),$$

we can obtain:

- $f(O_{ar}^\ell | S_i, M) = f\left(\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right)\right)$  for each allele  $a$ , locus  $\ell$ , replicate  $r$ ,
- $f(O | S_i, M) = f\left(\log_{10} \left( \frac{O}{E} \right)\right)$   
 $= \prod_\ell \prod_a \prod_r f\left(\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right)\right).$

356

## 2. Compare $E$ to $O$

Given the model

$$\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right) \sim N\left(0, \frac{c^2}{E_{ar}^\ell}\right),$$

we can obtain:

- $f(O_{ar}^\ell | S_i, M) = f\left(\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right)\right)$  for each allele  $a$ , locus  $\ell$ , replicate  $r$ ,
- $f(O | S_i, M) = f\left(\log_{10} \left( \frac{O}{E} \right)\right)$   
 $= \prod_\ell \prod_a \prod_r f\left(\log_{10} \left( \frac{O_{ar}^\ell}{E_{ar}^\ell} \right)\right).$

356

# Continuous Model



1. Model the expected peak heights  $E_{ar}^\ell$  in function of gentotype set  $S_i$  and mass parameters  $\mathbf{M}$ .
2. Compare the expected peak heights  $E_{ar}^\ell$  to the observed peak heights  $O_{ar}^\ell$ .
3. Perform a large number of simulations of steps 1. and 2. that randomly vary the gentotype set  $S_i$  and mass parameters  $\mathbf{M}$  (this approximates the integration over  $\mathbf{M}$ ).
4. Assign  $f(G_C|S_i)$  based on the simulation results.

357

# Continuous Model



1. Model the expected peak heights  $E_{ar}^\ell$  in function of gentotype set  $S_i$  and mass parameters  $\mathbf{M}$ .
2. Compare the expected peak heights  $E_{ar}^\ell$  to the observed peak heights  $O_{ar}^\ell$ .
3. Perform a large number of simulations of steps 1. and 2. that randomly vary the gentotype set  $S_i$  and mass parameters  $\mathbf{M}$  (this approximates the integration over  $\mathbf{M}$ ).
4. Assign  $f(G_C|S_i)$  based on the simulation results.

357

### 3. Simulations

$$f(G_C | S_i) = \int_M f(\mathbf{O} | S_i, \mathbf{M}) \Pr(\mathbf{M}) d\mathbf{M}$$

This integration is too complex to compute, but we can approximate the result using **Markov Chain Monte Carlo (MCMC)** sampling and the **Metropolis-Hastings** algorithm.

358

### 3. Simulations

$$f(G_C | S_i) = \int_M f(\mathbf{O} | S_i, \mathbf{M}) \Pr(\mathbf{M}) d\mathbf{M}$$

This integration is too complex to compute, but we can approximate the result using **Markov Chain Monte Carlo (MCMC)** sampling and the **Metropolis-Hastings** algorithm.

358

### 3. Simulations

#### Markov Chain Monte Carlo (MCMC) sampling:

For each simulation  $y$ , the model “randomly” picks values for the unknown parameters

- template amount  $t_n$
- degradation parameter  $d_n$
- locus specific amplification efficiency  $A^\ell$
- parameter for taking into account variations between replicates  $R_r$
- genotype set  $S_i$

359

### 3. Simulations

#### Markov Chain Monte Carlo (MCMC) sampling:

For each simulation  $y$ , the model “randomly” picks values for the unknown parameters

- template amount  $t_n$
- degradation parameter  $d_n$
- locus specific amplification efficiency  $A^\ell$
- parameter for taking into account variations between replicates  $R_r$
- genotype set  $S_i$

359

### 3. Simulations

Metropolis-Hastings algorithm:

- 1) For each simulation  $y$ , it calculates

$$f_y = f(\mathbf{O} | S_i, \mathbf{M}) = f\left(\log_{10}\left(\frac{\mathbf{O}}{E}\right)\right).$$

360

### 3. Simulations

Metropolis-Hastings algorithm:

- 1) For each simulation  $y$ , it calculates

$$f_y = f(\mathbf{O} | S_i, \mathbf{M}) = f\left(\log_{10}\left(\frac{\mathbf{O}}{E}\right)\right).$$

360

### 3. Simulations

#### Metropolis-Hastings algorithm:

2) It “accepts” the parameter values and genotype set:

- in every case where  $f_y \geq f_{y-1}$ , that is where the calculated probability density is **greater than or equal to** the probability density of the previous “accepted” parameter values and genotype set,
- in a fraction  $\frac{f_y}{f_{y-1}}$  of the cases where  $f_y < f_{y-1}$ , that is where the calculated probability density  $f_y$  is **smaller than** the probability density of the previous “accepted” parameter values and genotype set.

361

### 3. Simulations

#### Metropolis-Hastings algorithm:

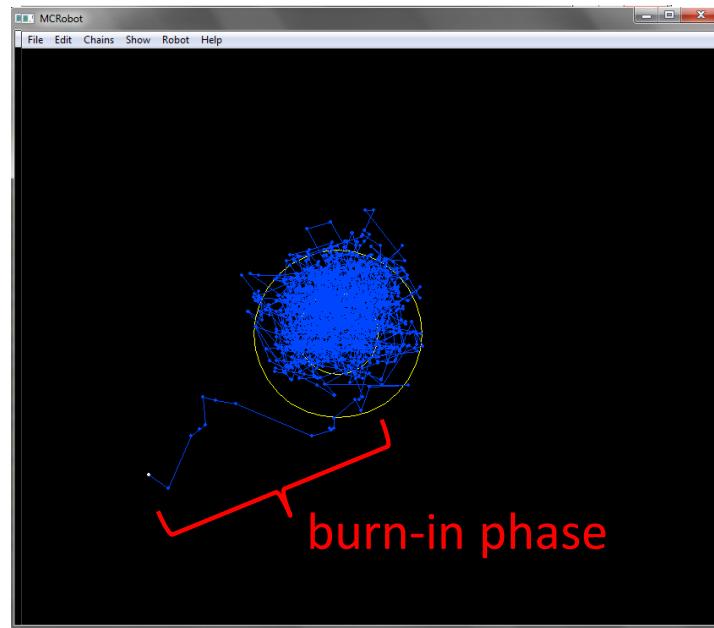
2) It “accepts” the parameter values and genotype set:

- in every case where  $f_y \geq f_{y-1}$ , that is where the calculated probability density is **greater than or equal to** the probability density of the previous “accepted” parameter values and genotype set,
- in a fraction  $\frac{f_y}{f_{y-1}}$  of the cases where  $f_y < f_{y-1}$ , that is where the calculated probability density  $f_y$  is **smaller than** the probability density of the previous “accepted” parameter values and genotype set.

361

### 3. Simulations

MCRobot-2.1:

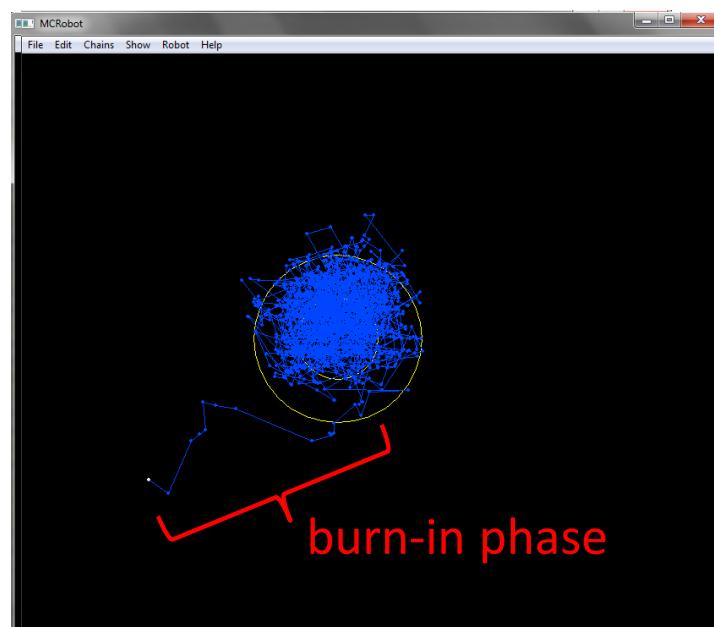


after taking 2100 samples

362

### 3. Simulations

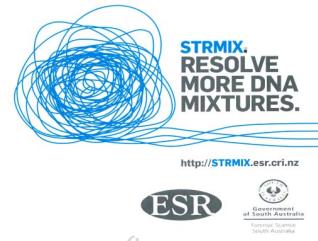
MCRobot-2.1:



after taking 2100 samples

362

# Continuous Model



<http://STRMIX.esr.cri.nz>



1. Model the expected peak heights  $E_{ar}^\ell$  in function of genotype set  $S_i$  and mass parameters  $\mathbf{M}$ .
2. Compare the expected peak heights  $E_{ar}^\ell$  to the observed peak heights  $O_{ar}^\ell$ .
3. Perform a large number of simulations of steps 1. and 2. that randomly vary the genotype set  $S_i$  and mass parameters  $\mathbf{M}$  (this approximates the integration over  $\mathbf{M}$ ).
4. Assign  $f(G_C|S_i)$  based on the simulation results.

363

# Continuous Model



<http://STRMIX.esr.cri.nz>



1. Model the expected peak heights  $E_{ar}^\ell$  in function of genotype set  $S_i$  and mass parameters  $\mathbf{M}$ .
2. Compare the expected peak heights  $E_{ar}^\ell$  to the observed peak heights  $O_{ar}^\ell$ .
3. Perform a large number of simulations of steps 1. and 2. that randomly vary the genotype set  $S_i$  and mass parameters  $\mathbf{M}$  (this approximates the integration over  $\mathbf{M}$ ).
4. Assign  $f(G_C|S_i)$  based on the simulation results.

363

# 4. Assign $f(G_C | S_i)$



Accepted parameter values and genotype sets:

discard burn-in

genotype set	$t_n$	$d_n$
16, 19 and 26, 29	2000 and 1500	0.0005 and 0.001
16, 26 and 19, 29	1500 and 1222	0.0009 and 0.0015
16, 19 and 26, 29	2050 and 1800	0.001 and 0.0017
19, 26 and 16, 29	1900 and 1400	0.0008 and 0.0012
19, 29 and 16, 26	1850 and 1600	0.0007 and 0.0014
$S_i$ , 29	2010 and 1300	0.0006 and 0.0013
19, 26 and 16, 29	1500 and 1900	0.0006 and 0.0012
$S_i$ , 29	1800 and 2100	0.0007 and 0.0012
$S_i$ , 29	1740 and 1990	0.0007 and 0.0013
19, 26 and 16, 29	1860 and 1850	0.0008 and 0.0013
:	:	:

364

# 4. Assign $f(G_C | S_i)$



Accepted parameter values and genotype sets:

discard burn-in

genotype set	$t_n$	$d_n$
16, 19 and 26, 29	2000 and 1500	0.0005 and 0.001
16, 26 and 19, 29	1500 and 1222	0.0009 and 0.0015
16, 19 and 26, 29	2050 and 1800	0.001 and 0.0017
19, 26 and 16, 29	1900 and 1400	0.0008 and 0.0012
19, 29 and 16, 26	1850 and 1600	0.0007 and 0.0014
$S_i$ , 29	2010 and 1300	0.0006 and 0.0013
19, 26 and 16, 29	1500 and 1900	0.0006 and 0.0012
$S_i$ , 29	1800 and 2100	0.0007 and 0.0012
$S_i$ , 29	1740 and 1990	0.0007 and 0.0013
19, 26 and 16, 29	1860 and 1850	0.0008 and 0.0013
:	:	:

364

# Likelihood Ratio

$$LR = \frac{\sum_{j=1}^M f(G_C|S_j) \Pr(S_j|G_K^p, H_p)}{\sum_{i=1}^N f(G_C|S_i) \Pr(S_i|G_K^d, H_d)} \quad \text{where } M \leq N$$

Insert the obtained probability densities for values  $f(G_C|S_i)$  and  $f(G_C|S_i)$  in the likelihood ratio.

365

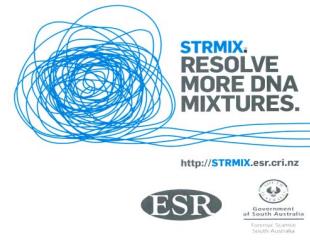
# Likelihood Ratio

$$LR = \frac{\sum_{j=1}^M f(G_C|S_j) \Pr(S_j|G_K^p, H_p)}{\sum_{i=1}^N f(G_C|S_i) \Pr(S_i|G_K^d, H_d)} \quad \text{where } M \leq N$$

Insert the obtained probability densities for values  $f(G_C|S_i)$  and  $f(G_C|S_i)$  in the likelihood ratio.

365

# STRmix™ Summary



## what it does:

- deconvolutes mixtures for a given number of contributors (1, 2, 3 or 4)
- compares reference profiles to the crime stain EPG data (this can consist of multiple replicates) and produces a likelihood ratio for a pair of propositions defined by the user
- searches the crime stain EPG data against a database to obtain a list of all the possible contributors with a likelihood ratio greater than or equal to threshold specified by the user

366

# STRmix™ Summary



## what it does:

- deconvolutes mixtures for a given number of contributors (1, 2, 3 or 4)
- compares reference profiles to the crime stain EPG data (this can consist of multiple replicates) and produces a likelihood ratio for a pair of propositions defined by the user
- searches the crime stain EPG data against a database to obtain a list of all the possible contributors with a likelihood ratio greater than or equal to threshold specified by the user

366

# STRmix™ Summary



## the model incorporates:

- Fst value for each population (Balding and Nichols, 1994, model)
- allele probabilities as point estimates or the Highest Posterior Density from a Dirichlet probability density function
- a model for allele drop-out
- a model for allele drop-in (exponential function or a constant)
- a laboratory-, locus- and allele-specific model for  $-1$  stutters
- laboratory-specific variance parameters for the model's normal distributions for alleles and stutters

367

# STRmix™ Summary

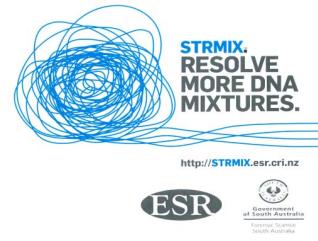


## the model incorporates:

- Fst value for each population (Balding and Nichols, 1994, model)
- allele probabilities as point estimates or the Highest Posterior Density from a Dirichlet probability density function
- a model for allele drop-out
- a model for allele drop-in (exponential function or a constant)
- a laboratory-, locus- and allele-specific model for  $-1$  stutters
- laboratory-specific variance parameters for the model's normal distributions for alleles and stutters

367

# STRmix™ Summary



## what it doesn't do:

- analyze data from lineage markers
- take into account the possibility of mutation events
- take into account triallelic loci
- kinship analyses
- assess scenarios with familial relationships
- take into account multiple EPGs generated with different instruments or kits
- deconvolute mixtures with >4 contributors

368

# STRmix™ Summary



## what it doesn't do:

- analyze data from lineage markers
- take into account the possibility of mutation events
- take into account triallelic loci
- kinship analyses
- assess scenarios with familial relationships
- take into account multiple EPGs generated with different instruments or kits
- deconvolute mixtures with >4 contributors

368

# STRmix™



# Demo

369

# STRmix™



# Demo

369