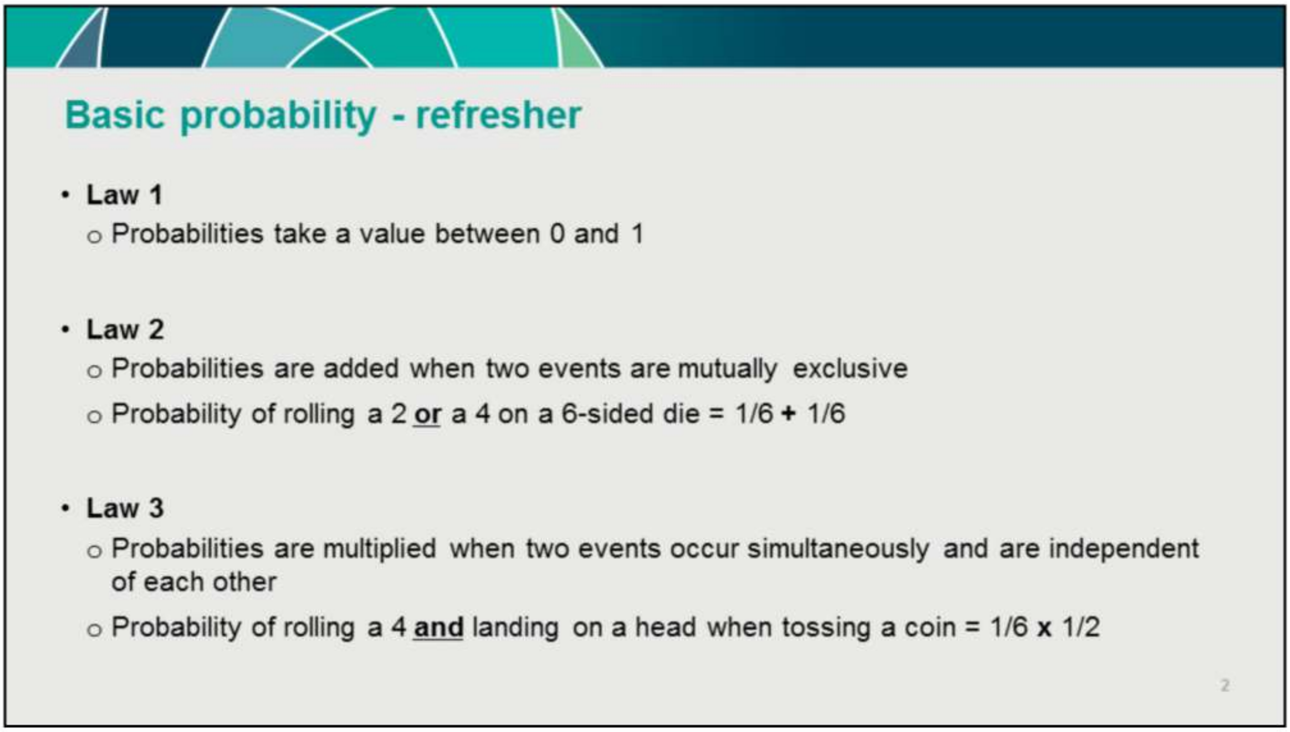
Introduction to DNA profile interpretation

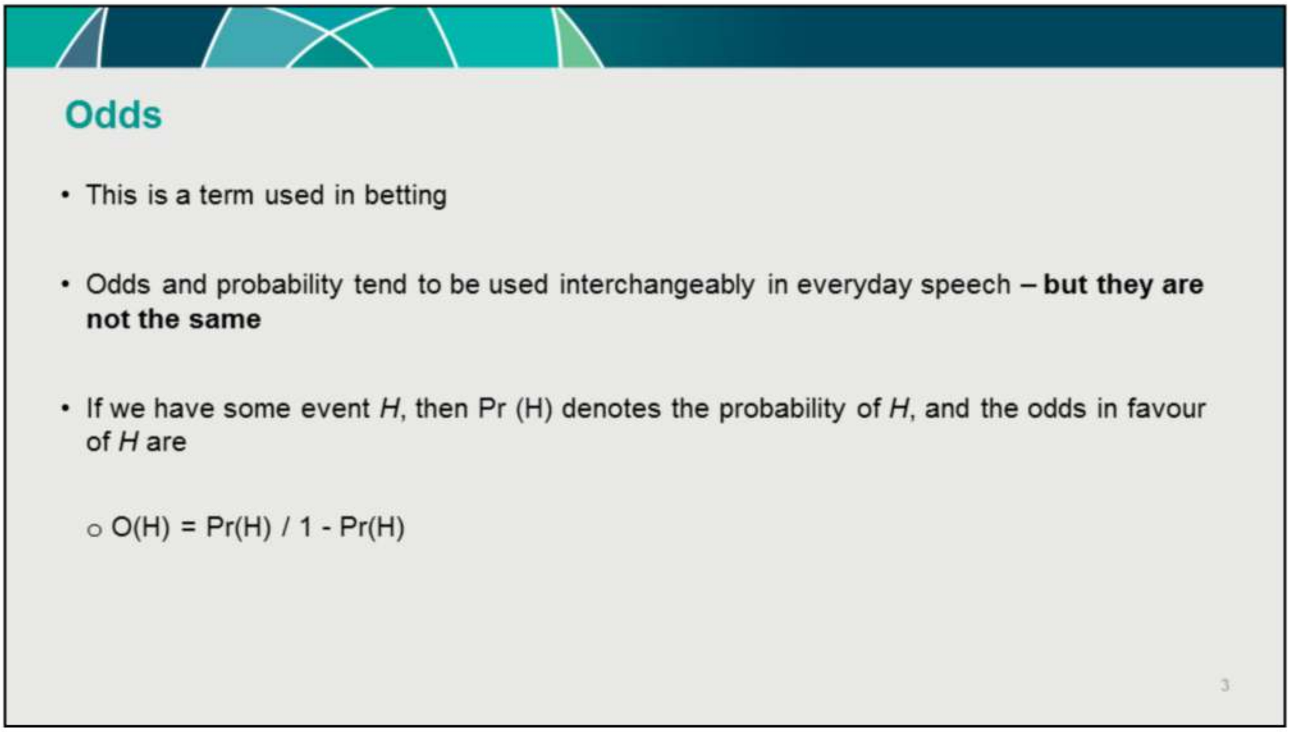


Basic probability - refresher

- **Law 1**
  - Probabilities take a value between 0 and 1

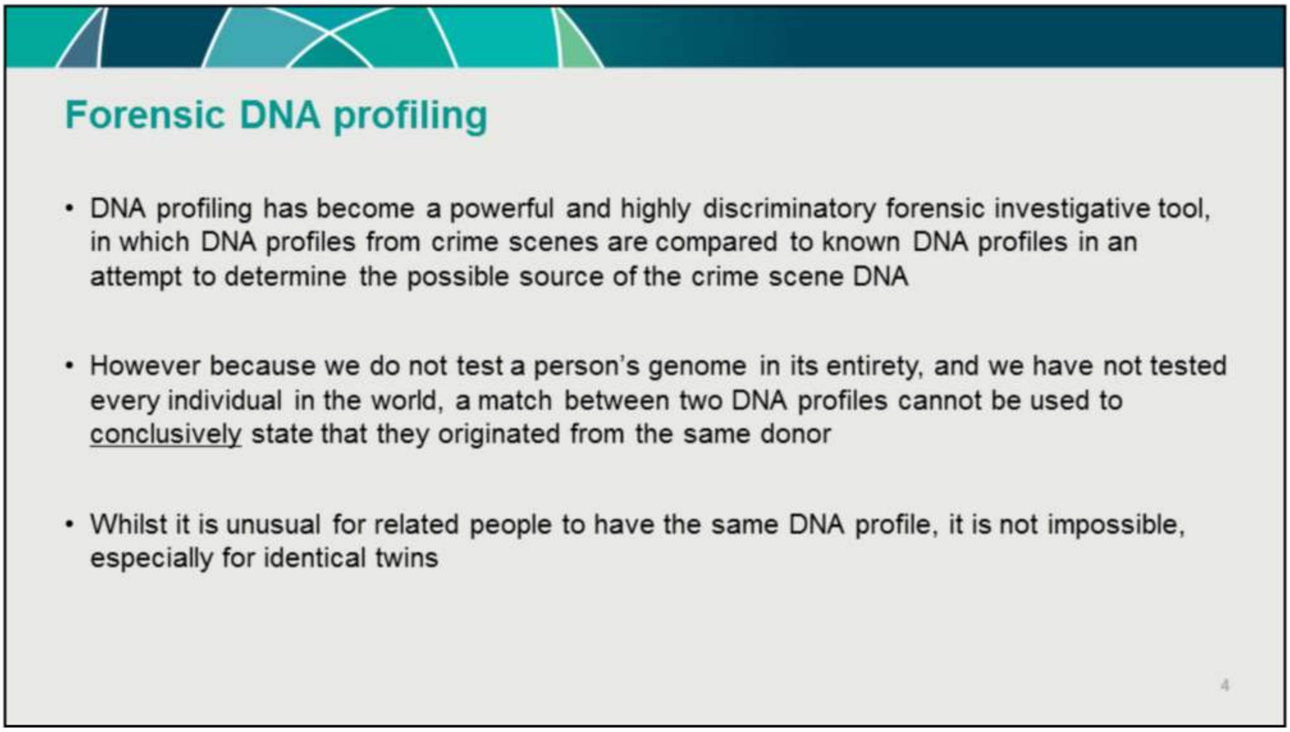
- **Law 2**
  - Probabilities are added when two events are mutually exclusive
  - Probability of rolling a 2 **<u>or</u>** a 4 on a 6-sided die = 1/6 **+** 1/6

- **Law 3**
  - Probabilities are multiplied when two events occur simultaneously and are independent of each other
  - Probability of rolling a 4 **<u>and</u>** landing on a head when tossing a coin = 1/6 **x** 1/2



Odds

- This is a term used in betting
- Odds and probability tend to be used interchangeably in everyday speech – **but they are not the same**
- If we have some event *H*, then Pr (H) denotes the probability of *H*, and the odds in favour of *H* are
    - $O(H) = Pr(H) / 1 - Pr(H)$

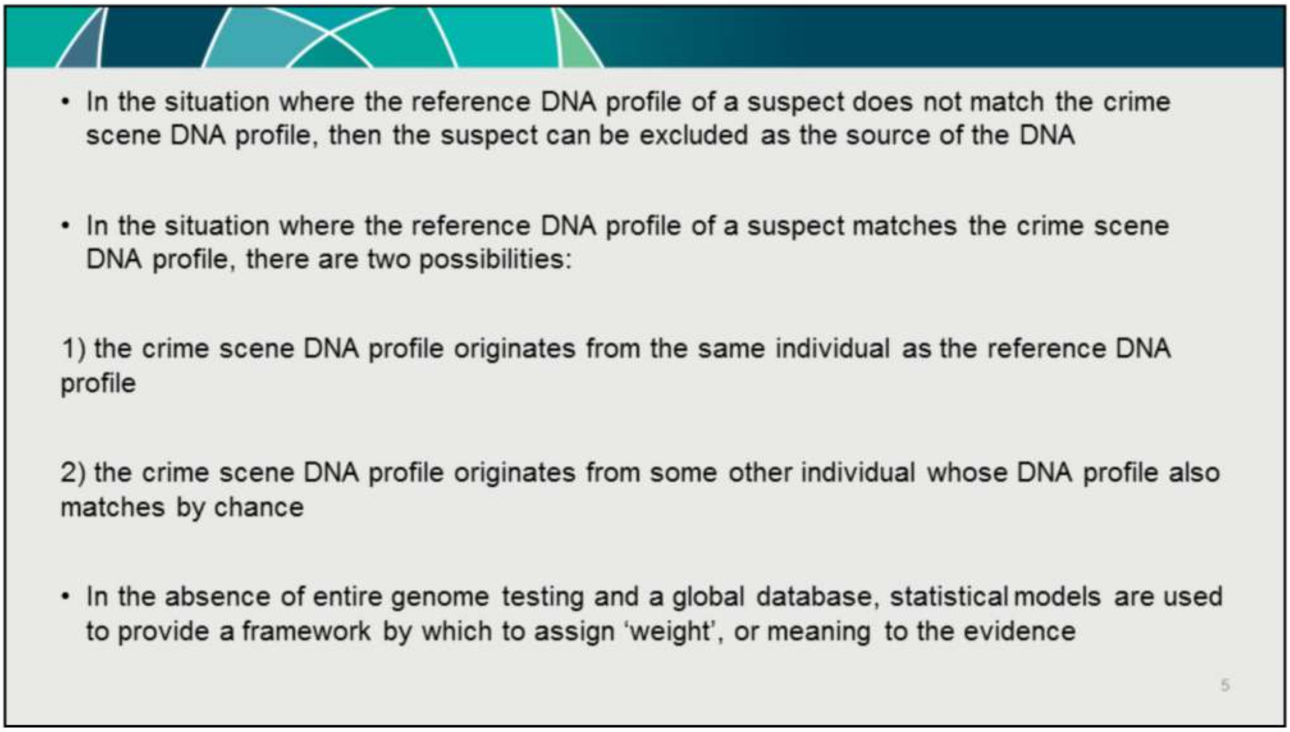


Forensic DNA profiling

- DNA profiling has become a powerful and highly discriminatory forensic investigative tool, in which DNA profiles from crime scenes are compared to known DNA profiles in an attempt to determine the possible source of the crime scene DNA

- However because we do not test a person's genome in its entirety, and we have not tested every individual in the world, a match between two DNA profiles cannot be used to conclusively state that they originated from the same donor

- Whilst it is unusual for related people to have the same DNA profile, it is not impossible, especially for identical twins



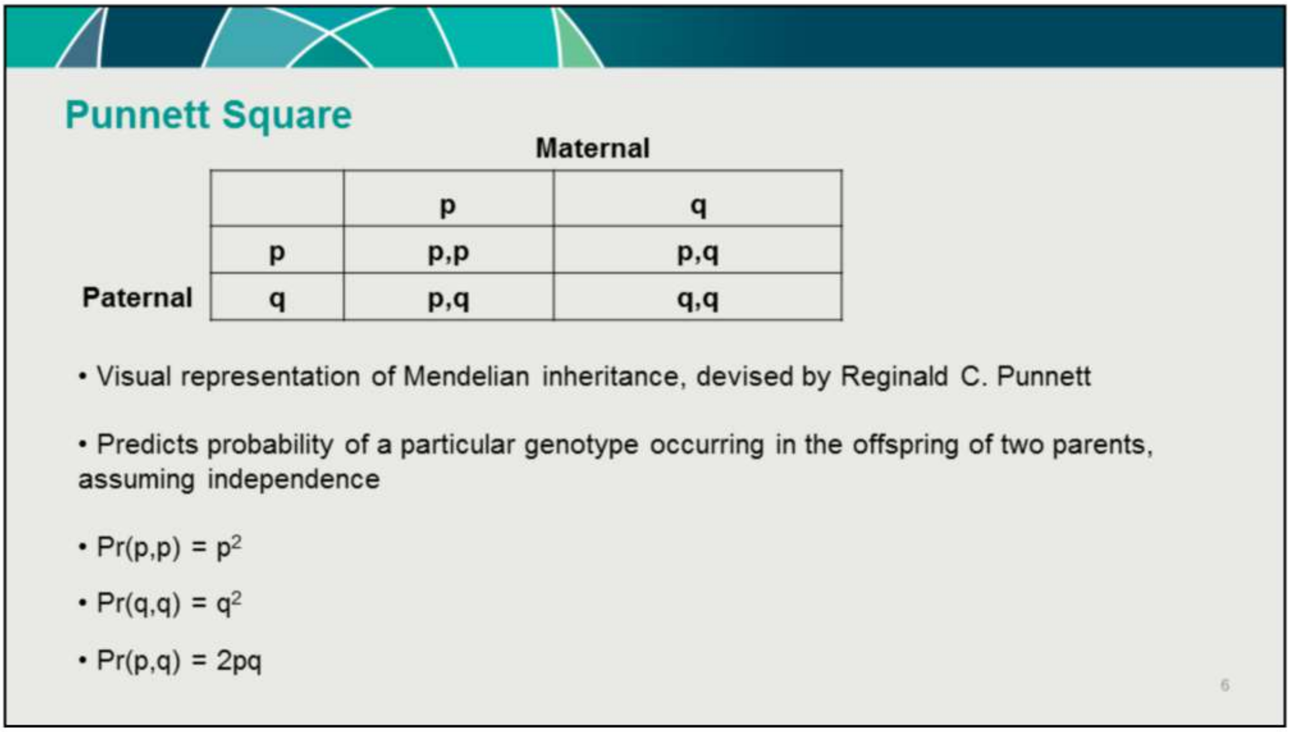
- In the situation where the reference DNA profile of a suspect does not match the crime scene DNA profile, then the suspect can be excluded as the source of the DNA

- In the situation where the reference DNA profile of a suspect matches the crime scene DNA profile, there are two possibilities:

1) the crime scene DNA profile originates from the same individual as the reference DNA profile

2) the crime scene DNA profile originates from some other individual whose DNA profile also matches by chance

- In the absence of entire genome testing and a global database, statistical models are used to provide a framework by which to assign ‘weight’, or meaning to the evidence



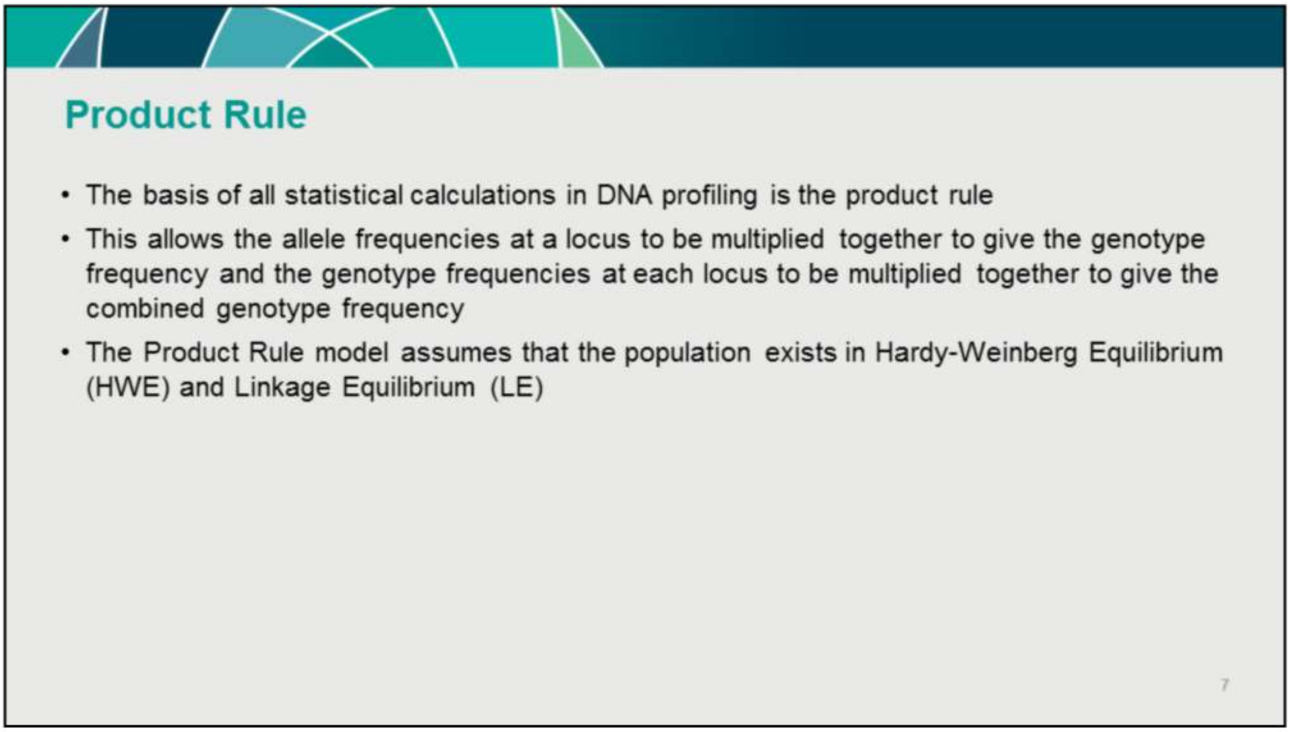
Punnett Square

Maternal

		p	**q**
	p	**p,p**	**p,q**
Paternal	**q**	**p,q**	**q,q**

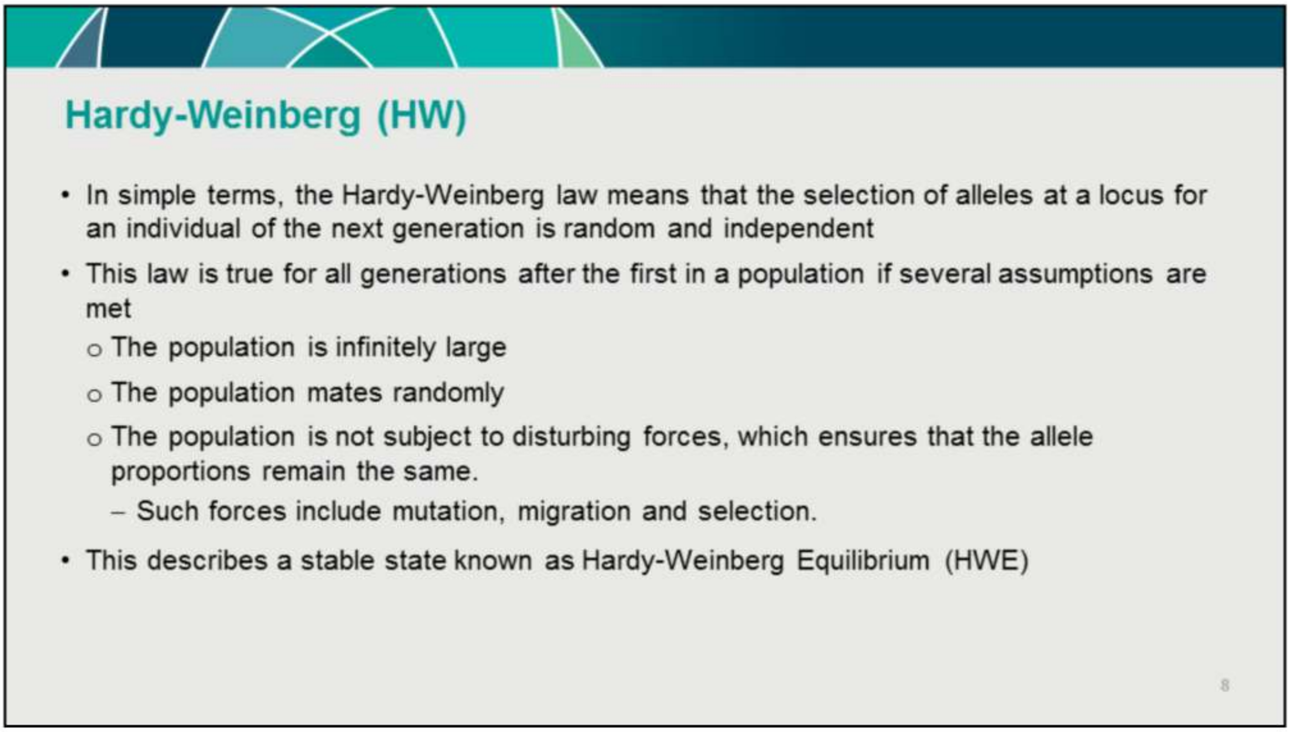
- Visual representation of Mendelian inheritance, devised by Reginald C. Punnett
- Predicts probability of a particular genotype occurring in the offspring of two parents, assuming independence
- Pr(p,p) = p^2
- Pr(q,q) = q^2
- Pr(p,q) = 2pq

6



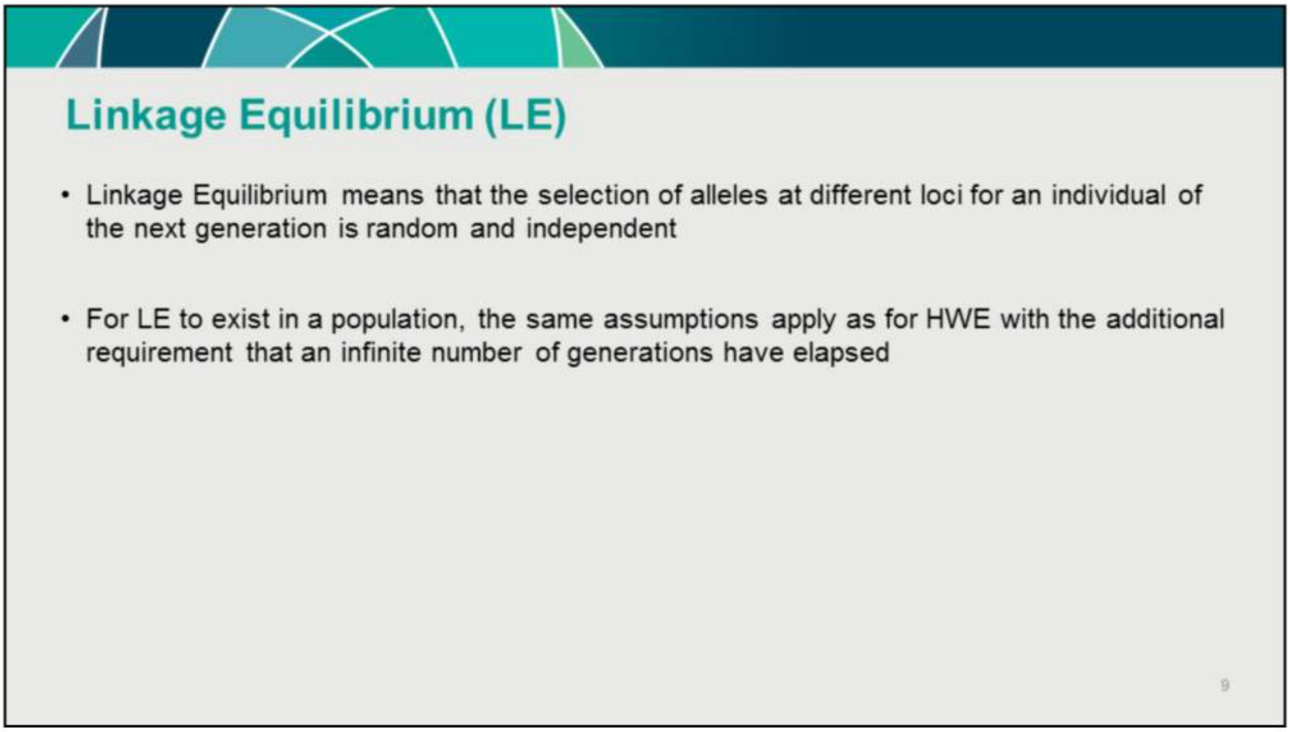
Product Rule

- The basis of all statistical calculations in DNA profiling is the product rule
- This allows the allele frequencies at a locus to be multiplied together to give the genotype frequency and the genotype frequencies at each locus to be multiplied together to give the combined genotype frequency
- The Product Rule model assumes that the population exists in Hardy-Weinberg Equilibrium (HWE) and Linkage Equilibrium (LE)



Hardy-Weinberg (HW)

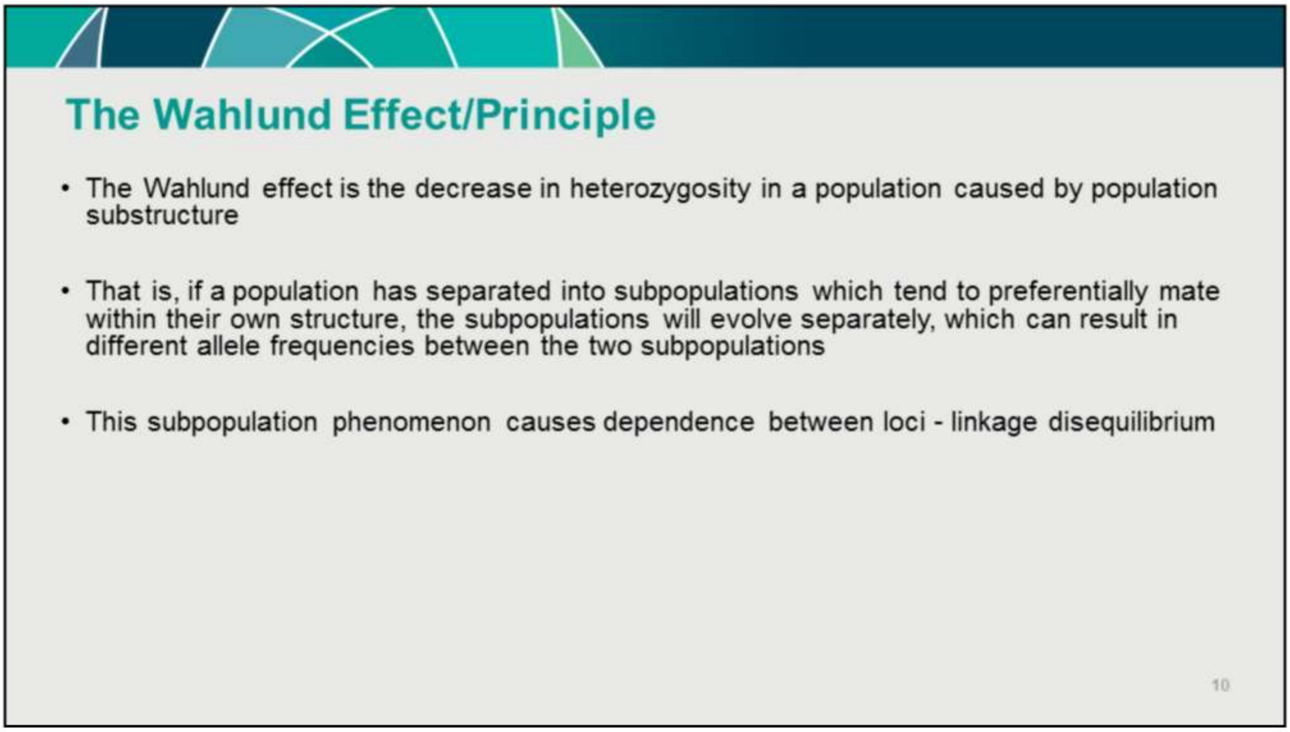
- In simple terms, the Hardy-Weinberg law means that the selection of alleles at a locus for an individual of the next generation is random and independent
- This law is true for all generations after the first in a population if several assumptions are met
  - The population is infinitely large
  - The population mates randomly
  - The population is not subject to disturbing forces, which ensures that the allele proportions remain the same.
    - Such forces include mutation, migration and selection.
- This describes a stable state known as Hardy-Weinberg Equilibrium (HWE)



Linkage Equilibrium (LE)

- Linkage Equilibrium means that the selection of alleles at different loci for an individual of the next generation is random and independent

- For LE to exist in a population, the same assumptions apply as for HWE with the additional requirement that an infinite number of generations have elapsed



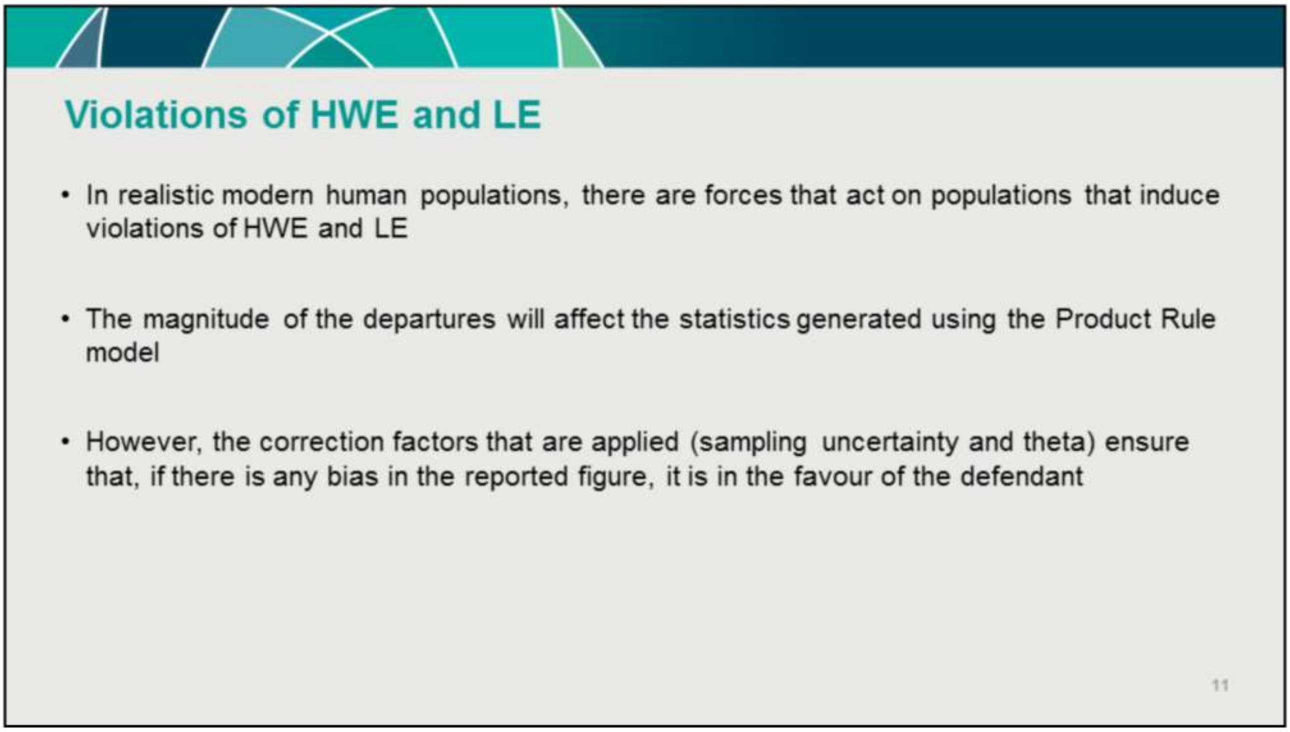
The Wahlund Effect/Principle

- The Wahlund effect is the decrease in heterozygosity in a population caused by population substructure

- That is, if a population has separated into subpopulations which tend to preferentially mate within their own structure, the subpopulations will evolve separately, which can result in different allele frequencies between the two subpopulations

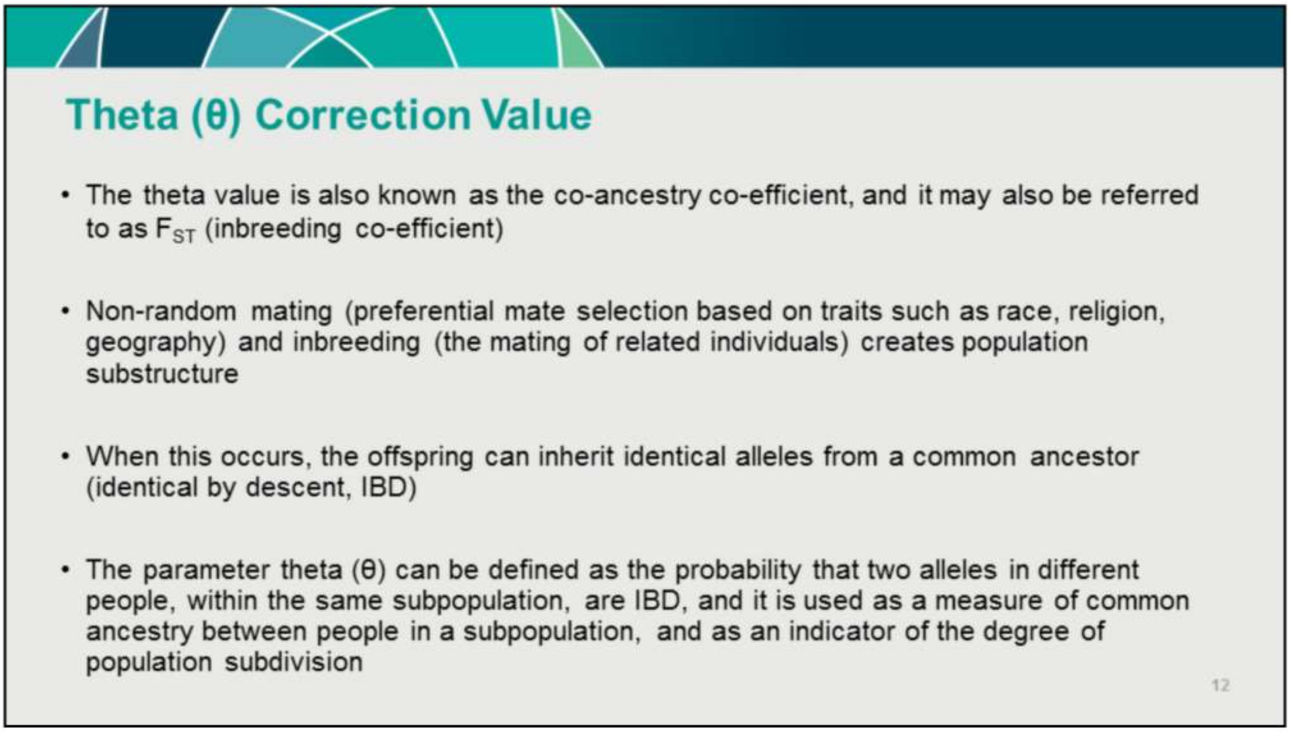
- This subpopulation phenomenon causes dependence between loci - linkage disequilibrium

10



Violations of HWE and LE

- In realistic modern human populations, there are forces that act on populations that induce violations of HWE and LE
- The magnitude of the departures will affect the statistics generated using the Product Rule model
- However, the correction factors that are applied (sampling uncertainty and theta) ensure that, if there is any bias in the reported figure, it is in the favour of the defendant



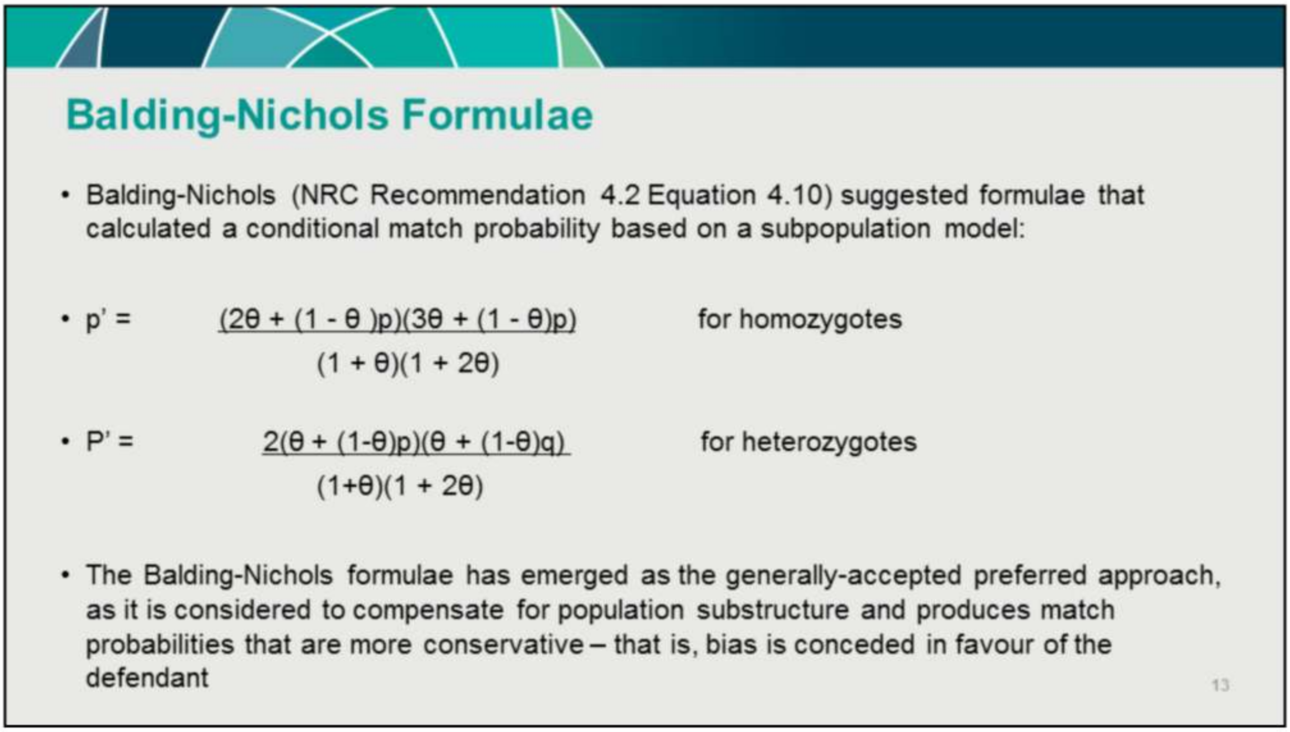
Theta (θ) Correction Value

- The theta value is also known as the co-ancestry co-efficient, and it may also be referred to as F_{ST} (inbreeding co-efficient)

- Non-random mating (preferential mate selection based on traits such as race, religion, geography) and inbreeding (the mating of related individuals) creates population substructure

- When this occurs, the offspring can inherit identical alleles from a common ancestor (identical by descent, IBD)

- The parameter theta (θ) can be defined as the probability that two alleles in different people, within the same subpopulation, are IBD, and it is used as a measure of common ancestry between people in a subpopulation, and as an indicator of the degree of population subdivision



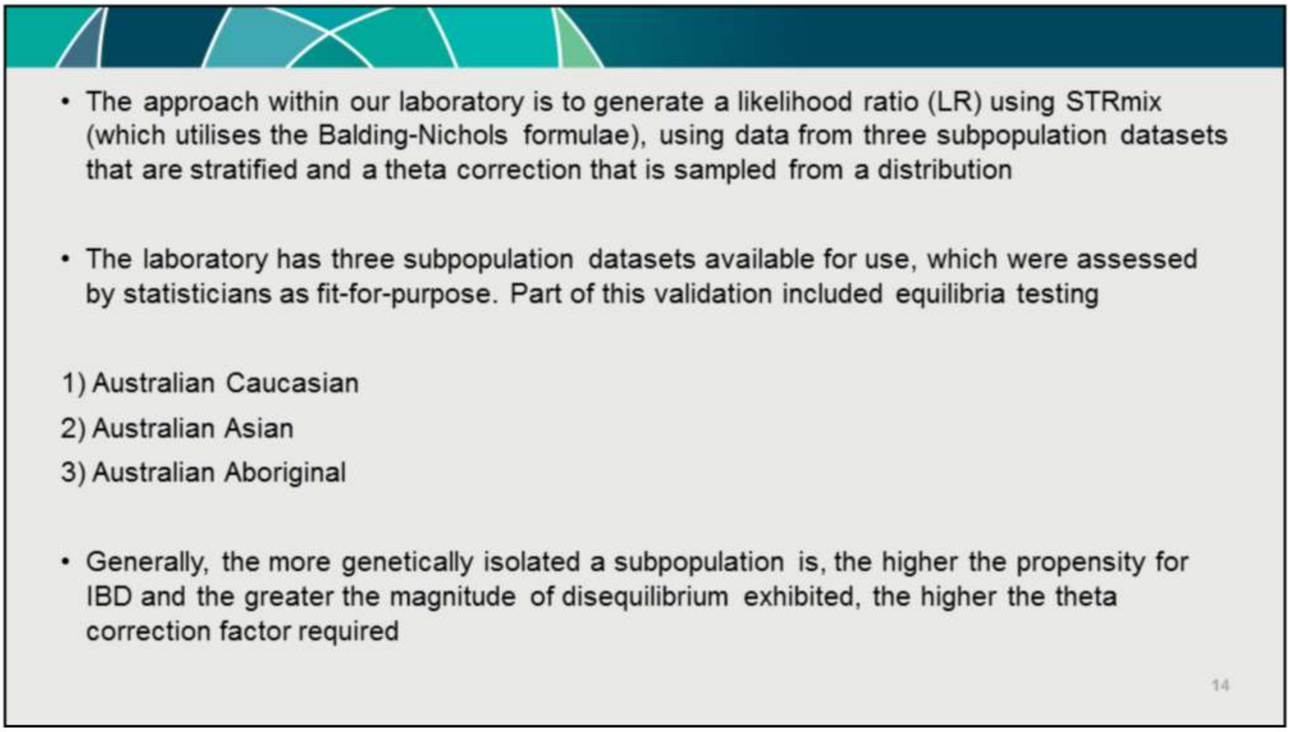
Balding-Nichols Formulae

- Balding-Nichols (NRC Recommendation 4.2 Equation 4.10) suggested formulae that calculated a conditional match probability based on a subpopulation model:

- $p' = \dfrac{(2\theta + (1 - \theta)p)(3\theta + (1 - \theta)p)}{(1 + \theta)(1 + 2\theta)}$ for homozygotes

- $P' = \dfrac{2(\theta + (1-\theta)p)(\theta + (1-\theta)q)}{(1+\theta)(1 + 2\theta)}$ for heterozygotes

- The Balding-Nichols formulae has emerged as the generally-accepted preferred approach, as it is considered to compensate for population substructure and produces match probabilities that are more conservative – that is, bias is conceded in favour of the defendant

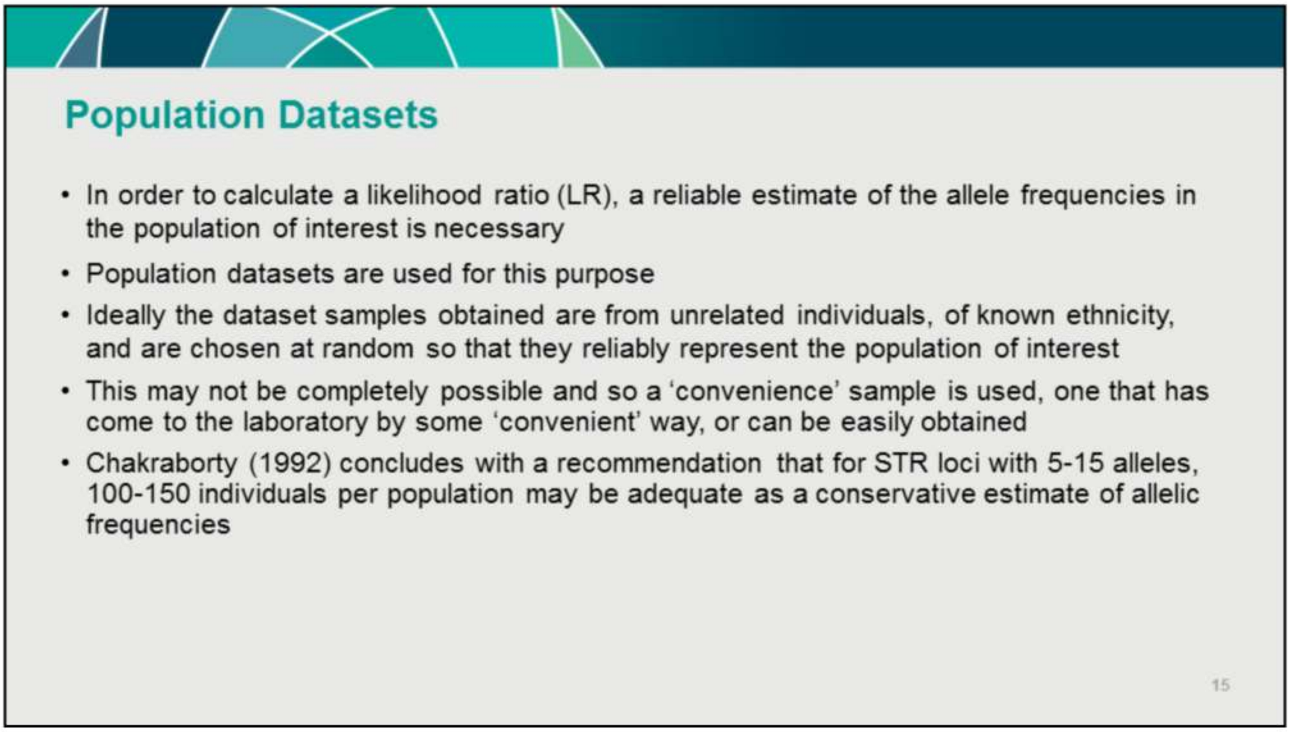


- The approach within our laboratory is to generate a likelihood ratio (LR) using STRmix (which utilises the Balding-Nichols formulae), using data from three subpopulation datasets that are stratified and a theta correction that is sampled from a distribution

- The laboratory has three subpopulation datasets available for use, which were assessed by statisticians as fit-for-purpose. Part of this validation included equilibria testing

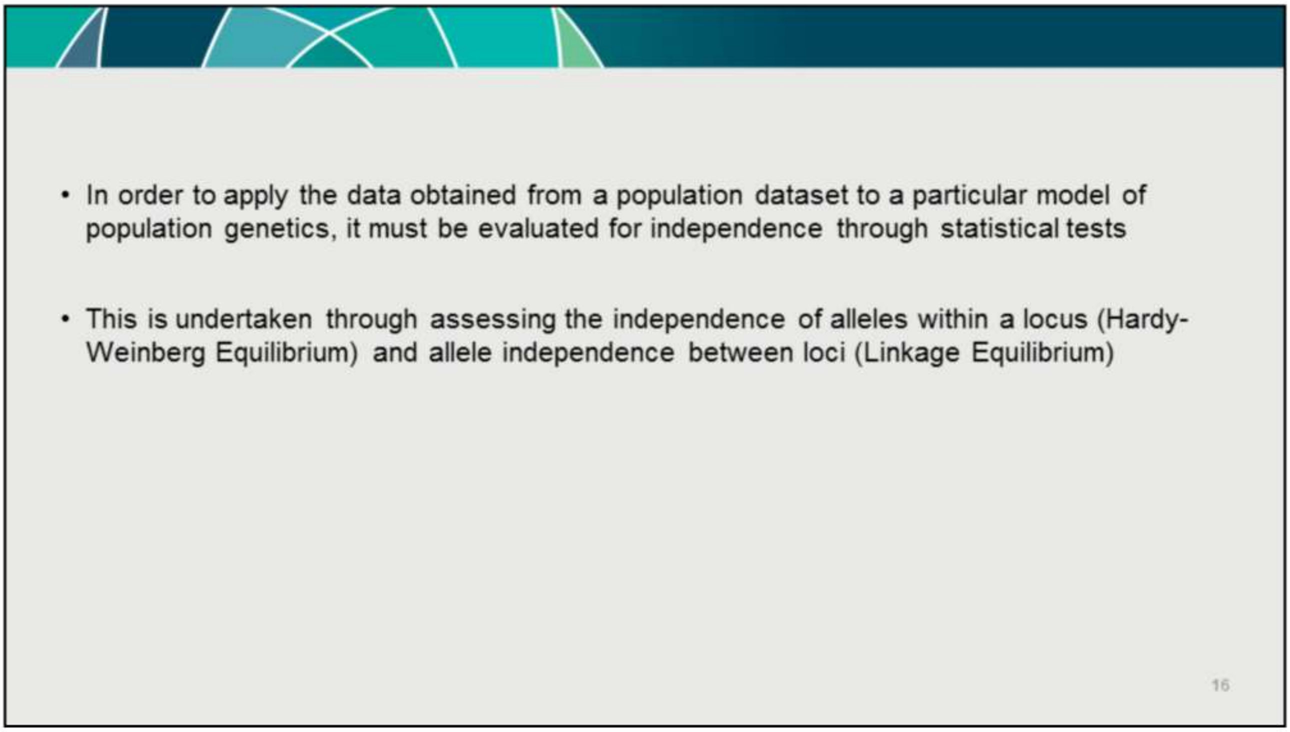
1) Australian Caucasian
2) Australian Asian
3) Australian Aboriginal

- Generally, the more genetically isolated a subpopulation is, the higher the propensity for IBD and the greater the magnitude of disequilibrium exhibited, the higher the theta correction factor required



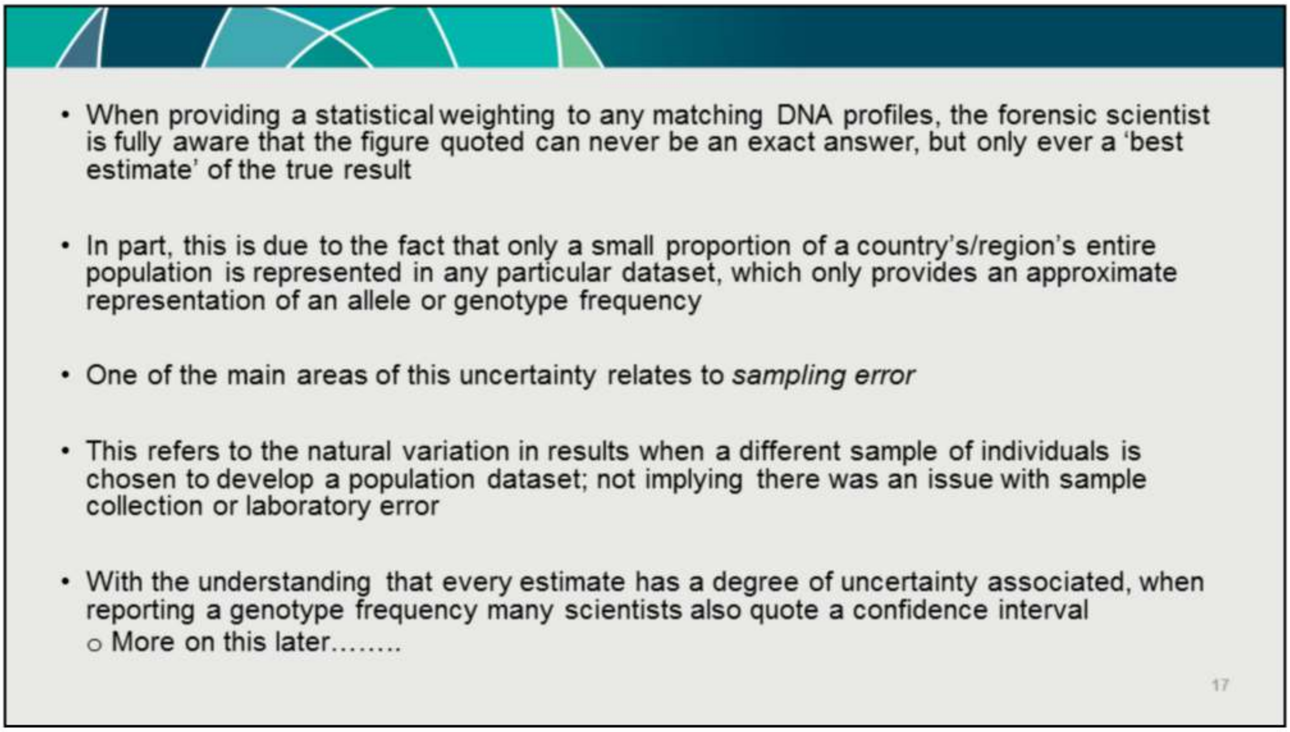
Population Datasets

- In order to calculate a likelihood ratio (LR), a reliable estimate of the allele frequencies in the population of interest is necessary
- Population datasets are used for this purpose
- Ideally the dataset samples obtained are from unrelated individuals, of known ethnicity, and are chosen at random so that they reliably represent the population of interest
- This may not be completely possible and so a ‘convenience’ sample is used, one that has come to the laboratory by some ‘convenient’ way, or can be easily obtained
- Chakraborty (1992) concludes with a recommendation that for STR loci with 5-15 alleles, 100-150 individuals per population may be adequate as a conservative estimate of allelic frequencies



- In order to apply the data obtained from a population dataset to a particular model of population genetics, it must be evaluated for independence through statistical tests

- This is undertaken through assessing the independence of alleles within a locus (Hardy-Weinberg Equilibrium) and allele independence between loci (Linkage Equilibrium)



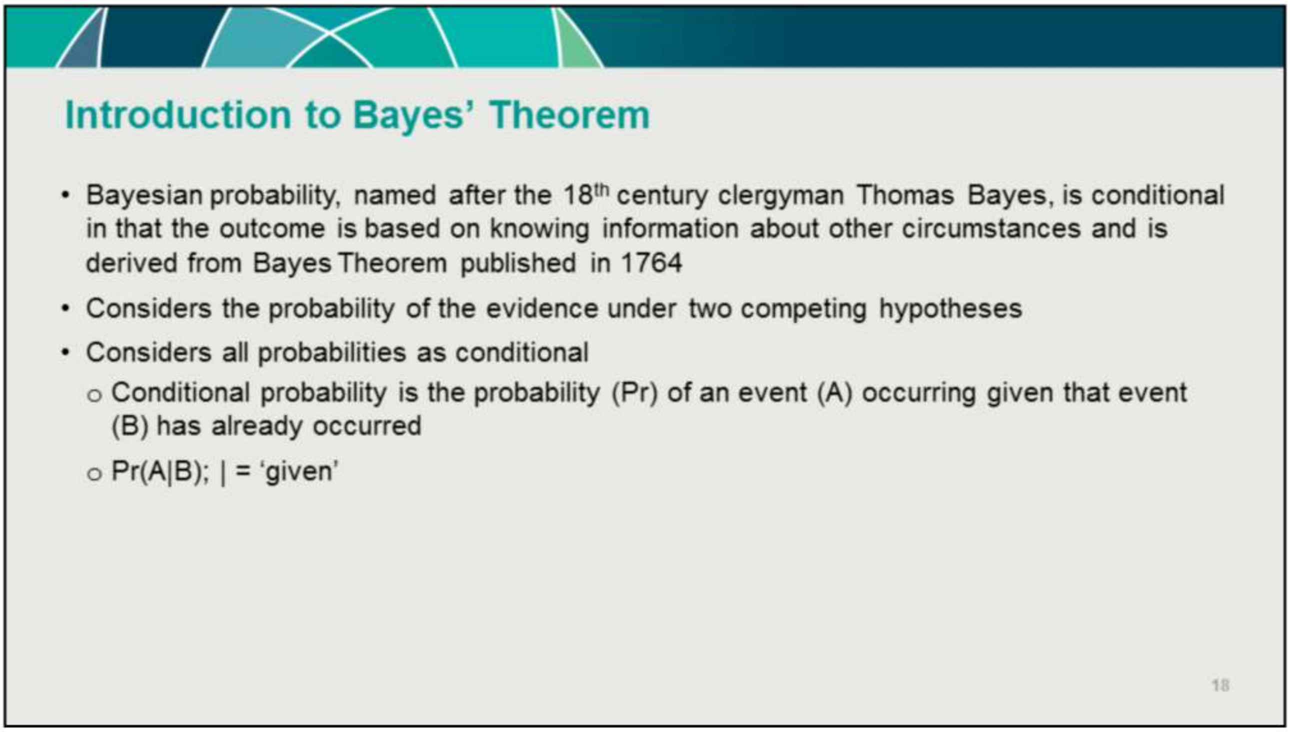
- When providing a statistical weighting to any matching DNA profiles, the forensic scientist is fully aware that the figure quoted can never be an exact answer, but only ever a ‘best estimate’ of the true result

- In part, this is due to the fact that only a small proportion of a country’s/region’s entire population is represented in any particular dataset, which only provides an approximate representation of an allele or genotype frequency

- One of the main areas of this uncertainty relates to *sampling error*

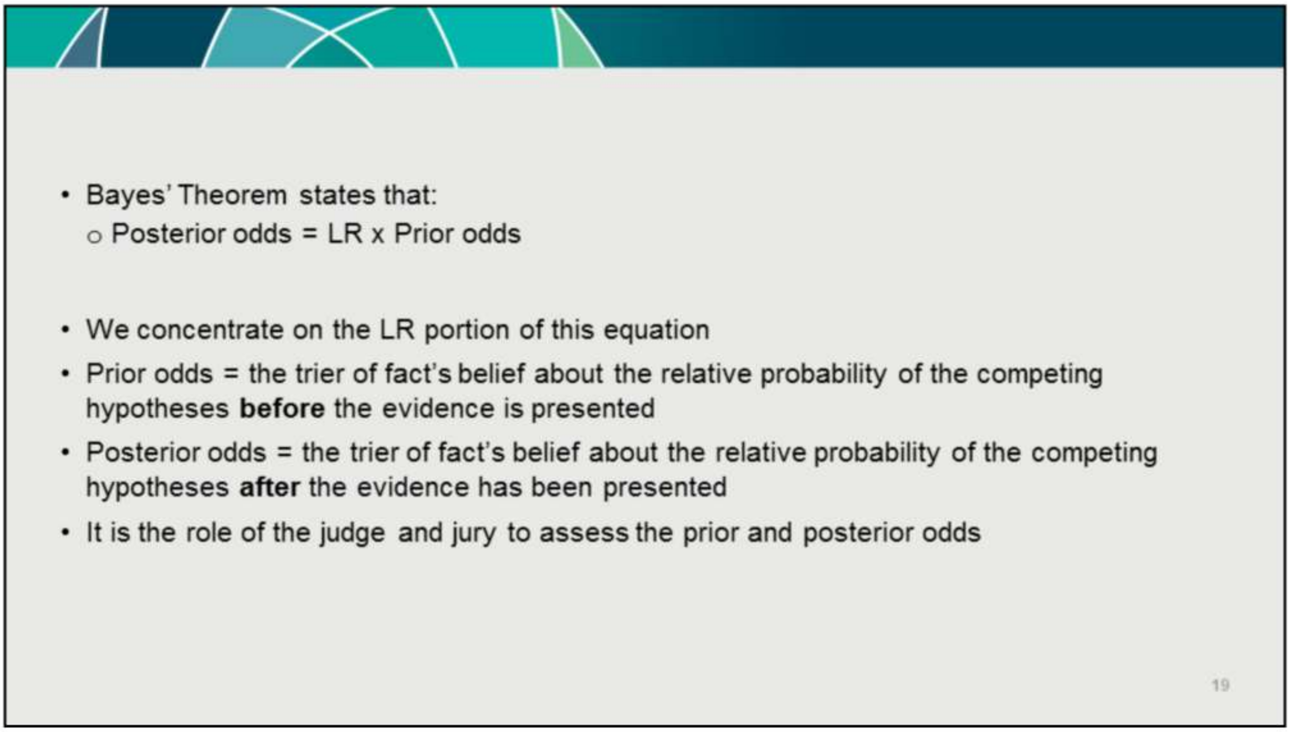
- This refers to the natural variation in results when a different sample of individuals is chosen to develop a population dataset; not implying there was an issue with sample collection or laboratory error

- With the understanding that every estimate has a degree of uncertainty associated, when reporting a genotype frequency many scientists also quote a confidence interval
  - o More on this later……..



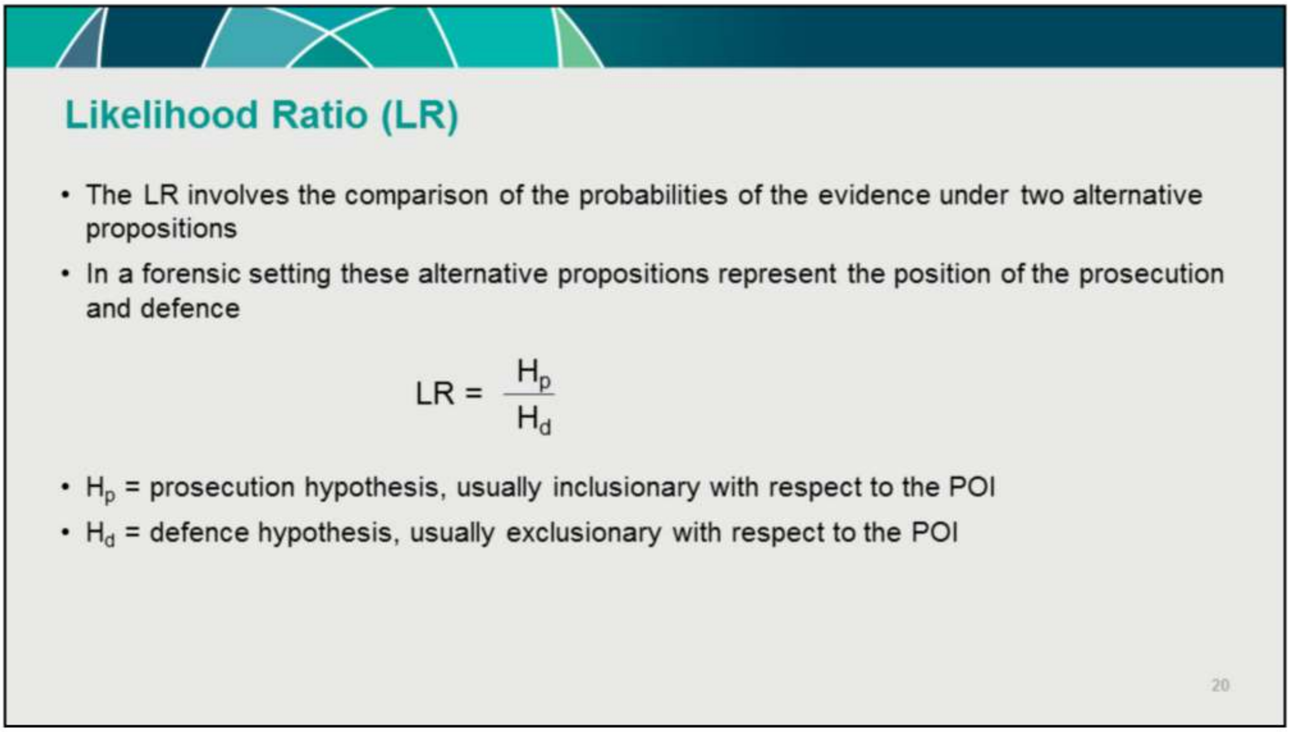
Introduction to Bayes' Theorem

- Bayesian probability, named after the 18th century clergyman Thomas Bayes, is conditional in that the outcome is based on knowing information about other circumstances and is derived from Bayes Theorem published in 1764
- Considers the probability of the evidence under two competing hypotheses
- Considers all probabilities as conditional
  - Conditional probability is the probability (Pr) of an event (A) occurring given that event (B) has already occurred
  - Pr(A|B); | = 'given'



- Bayes' Theorem states that:
  - Posterior odds = LR x Prior odds

- We concentrate on the LR portion of this equation
- Prior odds = the trier of fact's belief about the relative probability of the competing hypotheses **before** the evidence is presented
- Posterior odds = the trier of fact's belief about the relative probability of the competing hypotheses **after** the evidence has been presented
- It is the role of the judge and jury to assess the prior and posterior odds

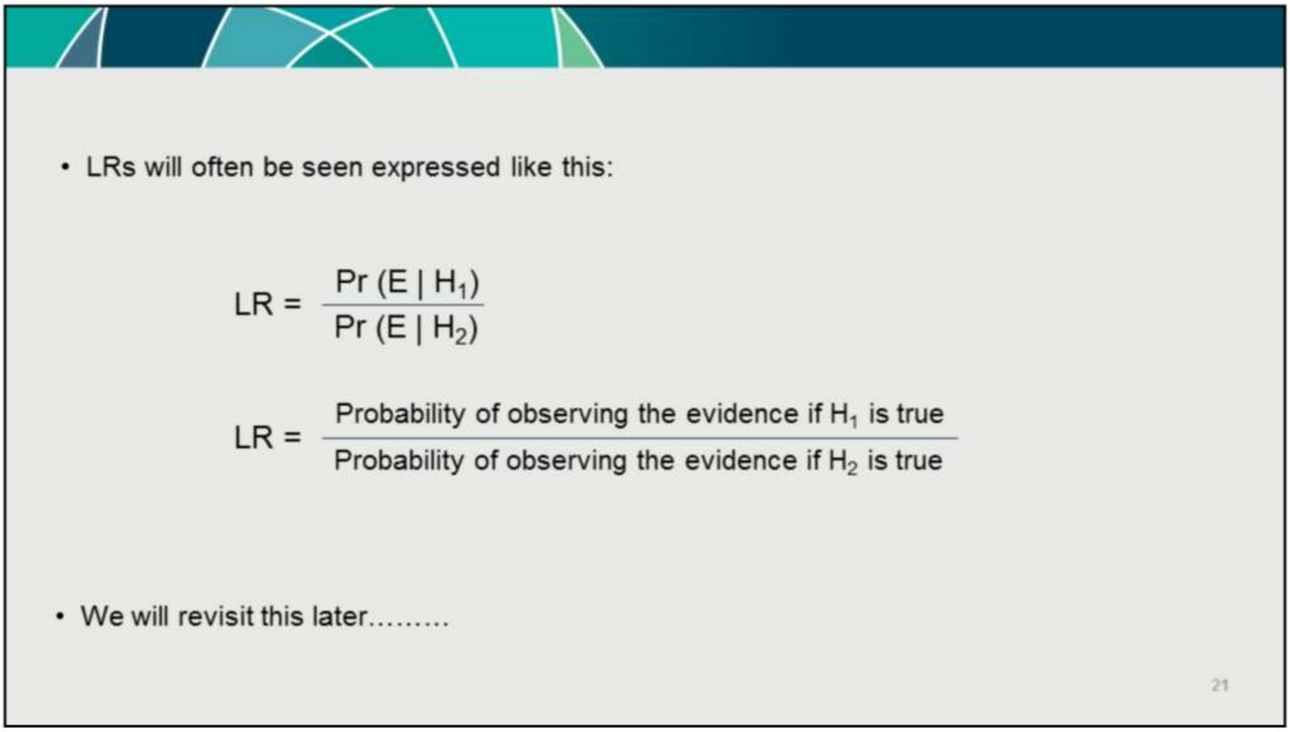


Likelihood Ratio (LR)

- The LR involves the comparison of the probabilities of the evidence under two alternative propositions
- In a forensic setting these alternative propositions represent the position of the prosecution and defence

$$LR = \frac{H_p}{H_d}$$

- H_p = prosecution hypothesis, usually inclusionary with respect to the POI
- H_d = defence hypothesis, usually exclusionary with respect to the POI

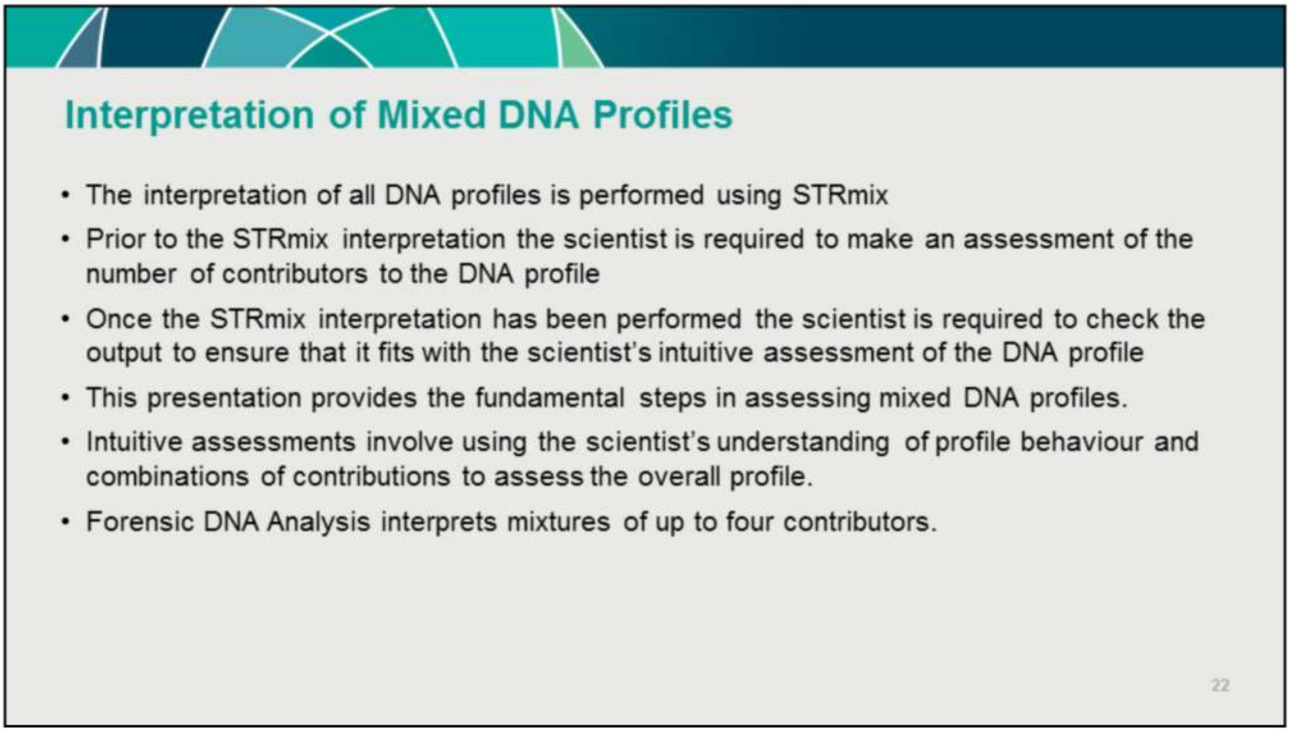


- LRs will often be seen expressed like this:

$$LR = \frac{Pr\ (E \mid H_1)}{Pr\ (E \mid H_2)}$$

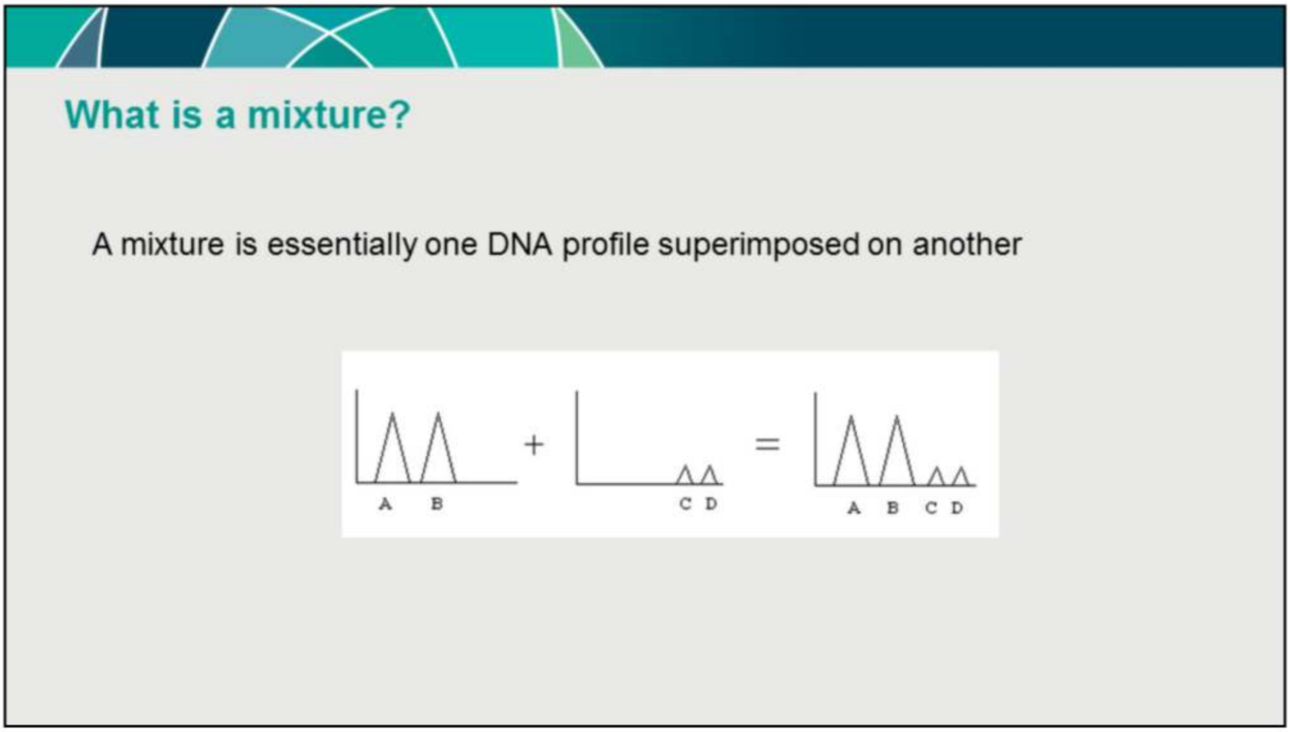
$$LR = \frac{\text{Probability of observing the evidence if } H_1 \text{ is true}}{\text{Probability of observing the evidence if } H_2 \text{ is true}}$$

- We will revisit this later………



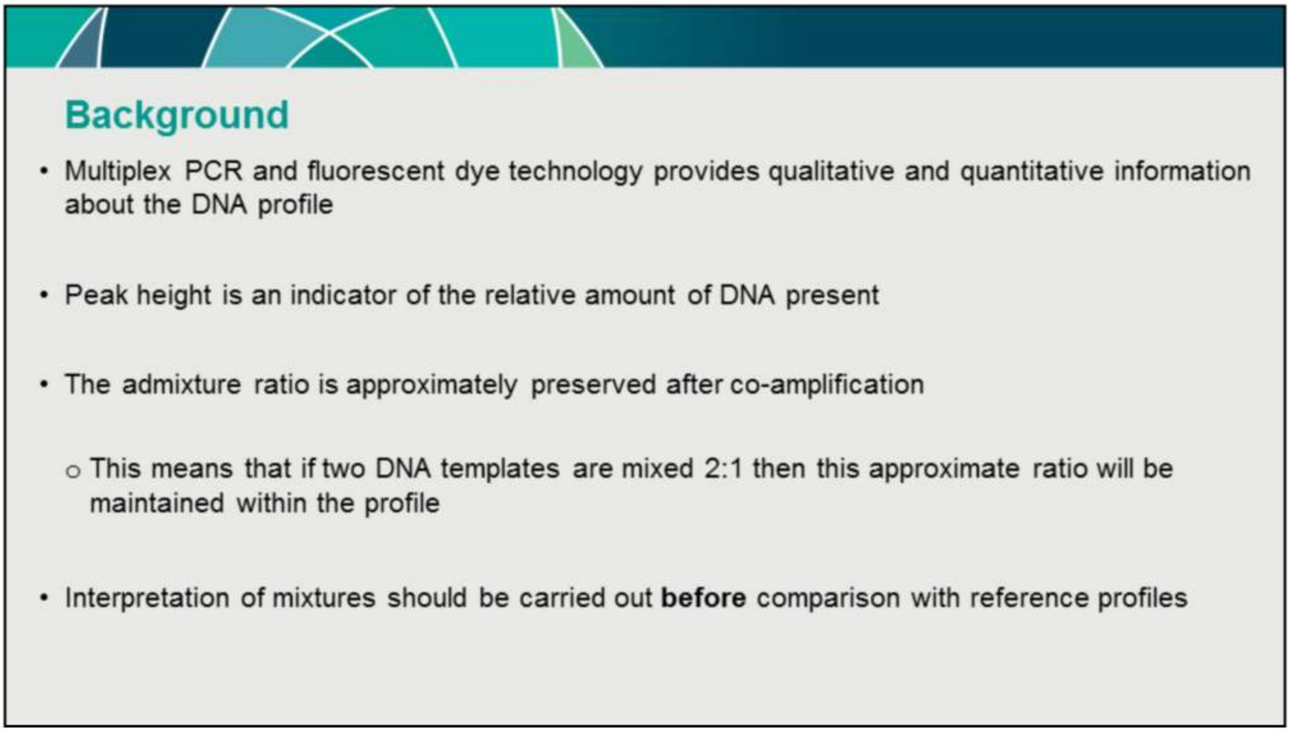
Interpretation of Mixed DNA Profiles

- The interpretation of all DNA profiles is performed using STRmix
- Prior to the STRmix interpretation the scientist is required to make an assessment of the number of contributors to the DNA profile
- Once the STRmix interpretation has been performed the scientist is required to check the output to ensure that it fits with the scientist's intuitive assessment of the DNA profile
- This presentation provides the fundamental steps in assessing mixed DNA profiles.
- Intuitive assessments involve using the scientist's understanding of profile behaviour and combinations of contributions to assess the overall profile.
- Forensic DNA Analysis interprets mixtures of up to four contributors.



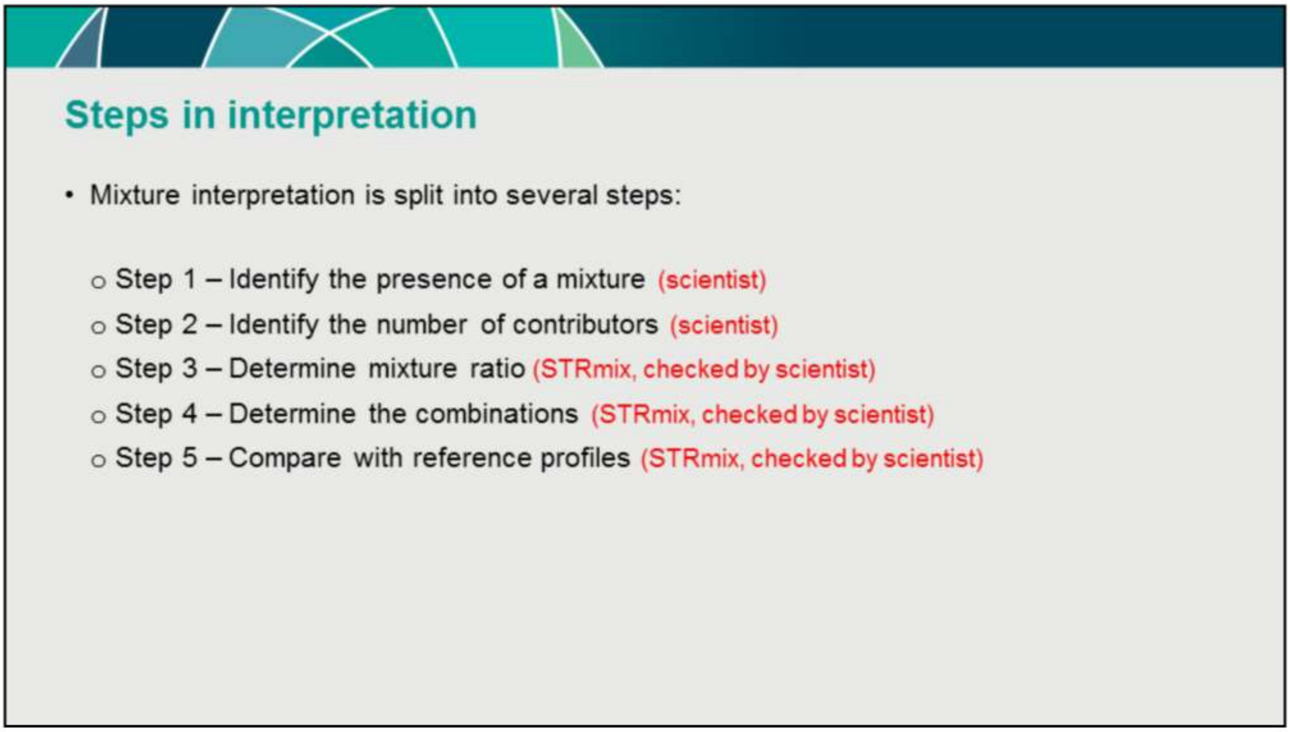
What is a mixture?

A mixture is essentially one DNA profile superimposed on another



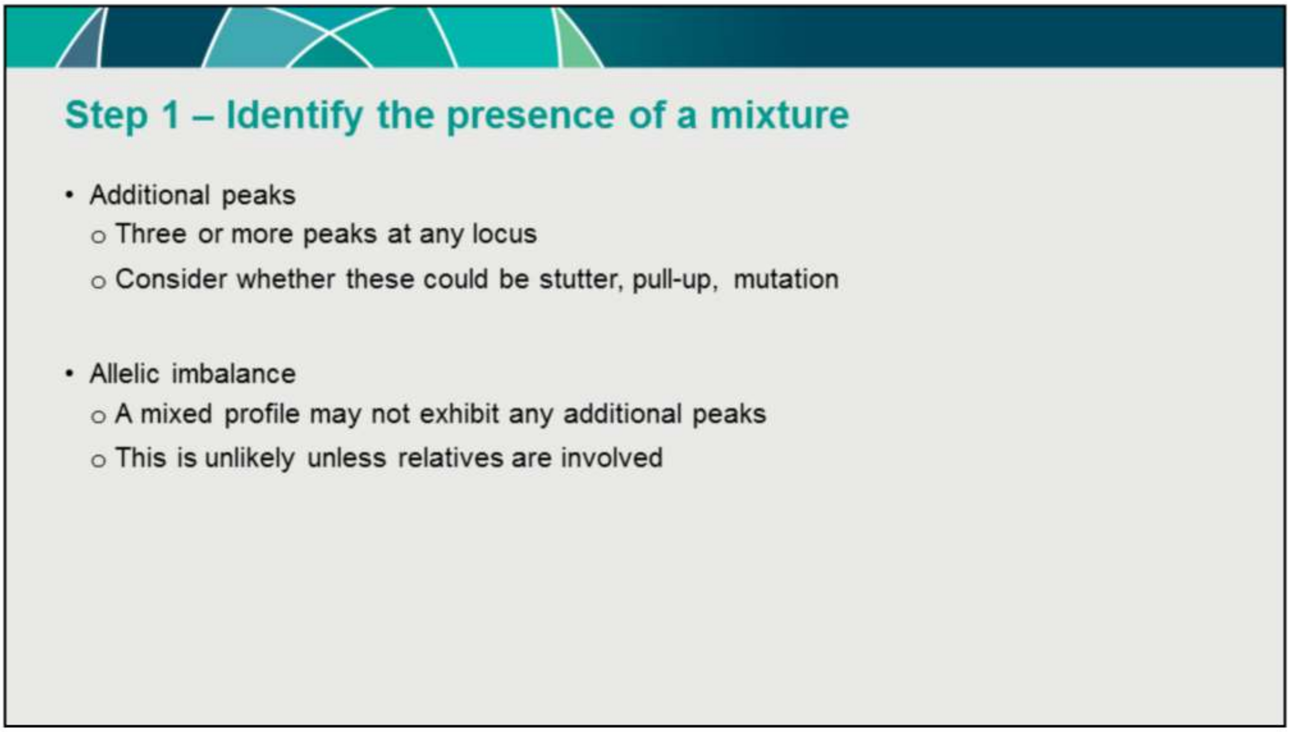
Background

- Multiplex PCR and fluorescent dye technology provides qualitative and quantitative information about the DNA profile
- Peak height is an indicator of the relative amount of DNA present
- The admixture ratio is approximately preserved after co-amplification
    - This means that if two DNA templates are mixed 2:1 then this approximate ratio will be maintained within the profile
- Interpretation of mixtures should be carried out **before** comparison with reference profiles



Steps in interpretation

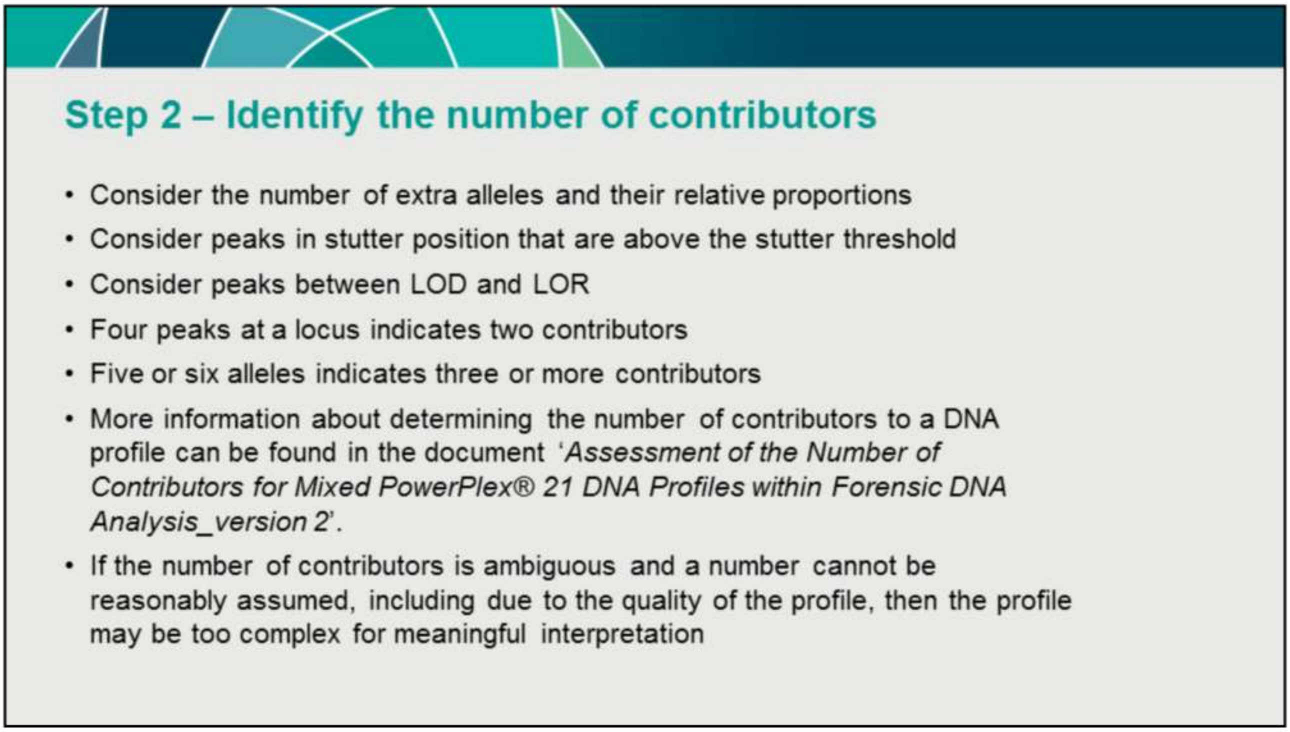
- Mixture interpretation is split into several steps:
  - Step 1 – Identify the presence of a mixture (scientist)
  - Step 2 – Identify the number of contributors (scientist)
  - Step 3 – Determine mixture ratio (STRmix, checked by scientist)
  - Step 4 – Determine the combinations (STRmix, checked by scientist)
  - Step 5 – Compare with reference profiles (STRmix, checked by scientist)



Step 1 – Identify the presence of a mixture

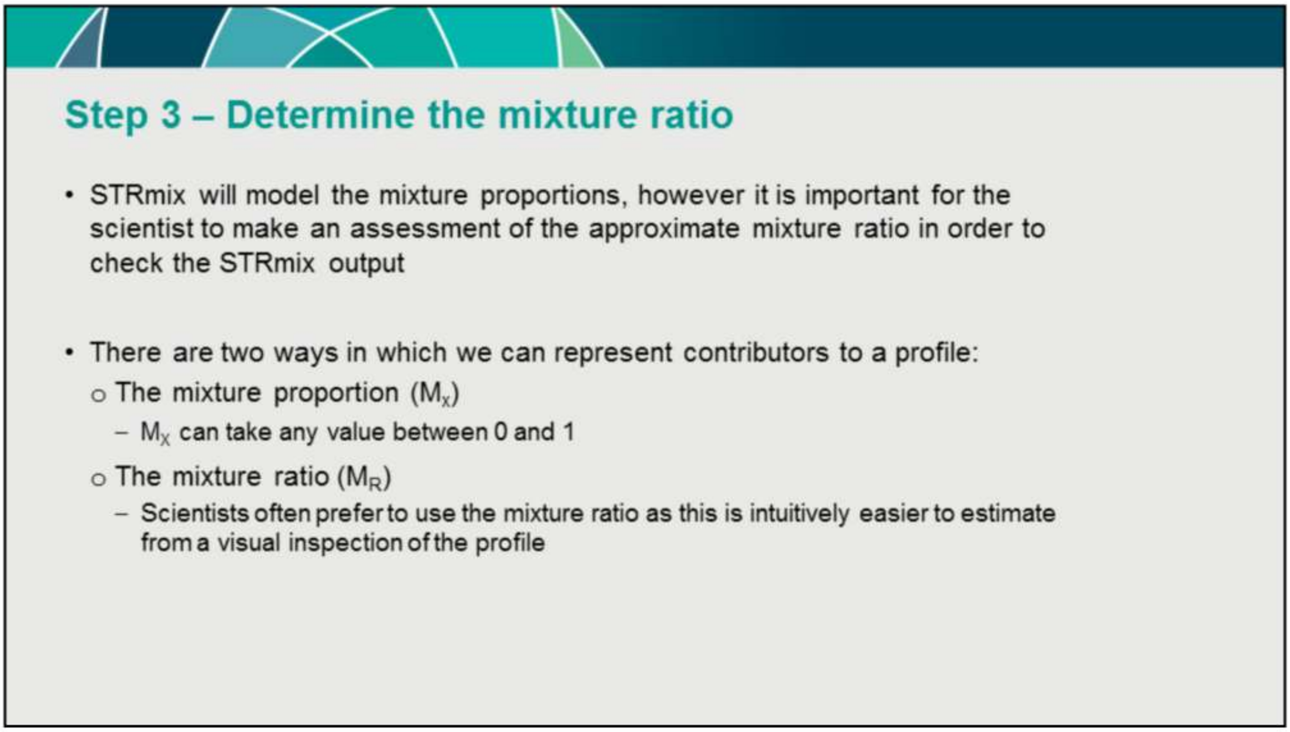
- Additional peaks
  - Three or more peaks at any locus
  - Consider whether these could be stutter, pull-up, mutation

- Allelic imbalance
  - A mixed profile may not exhibit any additional peaks
  - This is unlikely unless relatives are involved



Step 2 – Identify the number of contributors

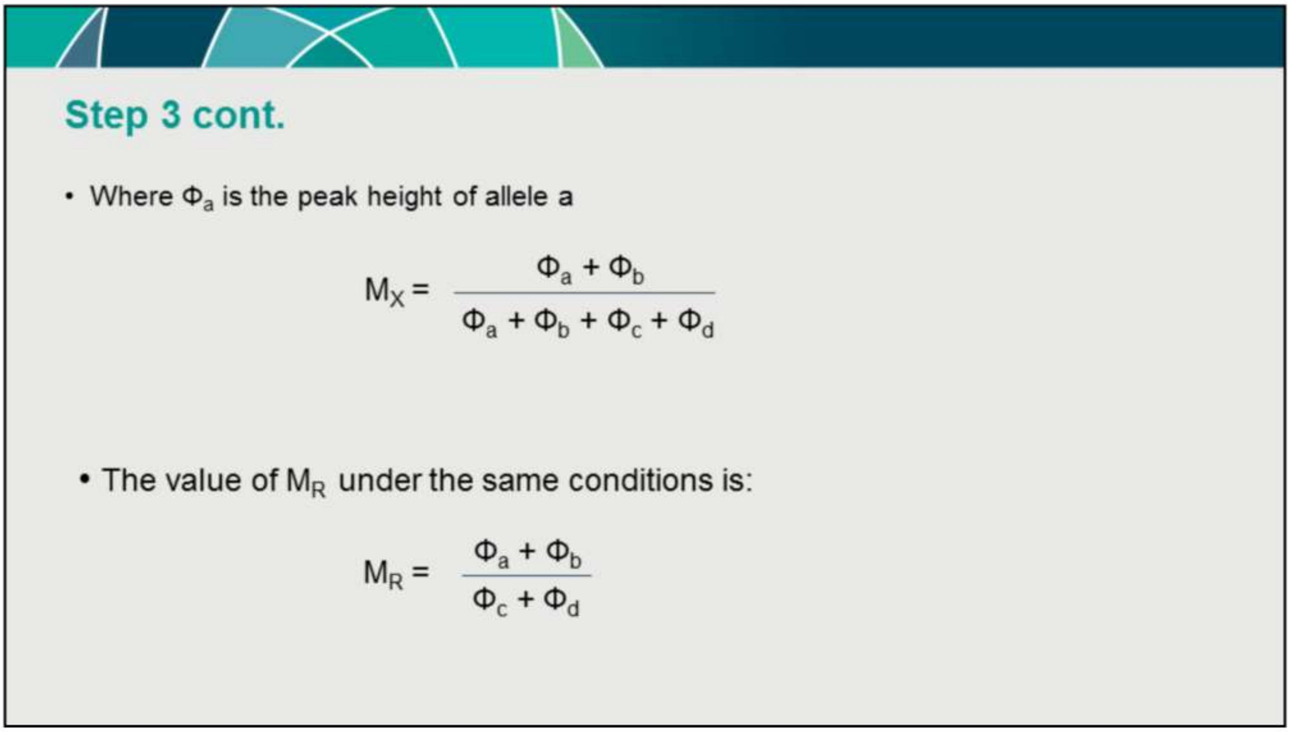
- Consider the number of extra alleles and their relative proportions
- Consider peaks in stutter position that are above the stutter threshold
- Consider peaks between LOD and LOR
- Four peaks at a locus indicates two contributors
- Five or six alleles indicates three or more contributors
- More information about determining the number of contributors to a DNA profile can be found in the document '*Assessment of the Number of Contributors for Mixed PowerPlex® 21 DNA Profiles within Forensic DNA Analysis_version 2*'.
- If the number of contributors is ambiguous and a number cannot be reasonably assumed, including due to the quality of the profile, then the profile may be too complex for meaningful interpretation



Step 3 – Determine the mixture ratio

- STRmix will model the mixture proportions, however it is important for the scientist to make an assessment of the approximate mixture ratio in order to check the STRmix output

- There are two ways in which we can represent contributors to a profile:
  - o The mixture proportion (M_x)
    - – M_X can take any value between 0 and 1
  - o The mixture ratio (M_R)
    - – Scientists often prefer to use the mixture ratio as this is intuitively easier to estimate from a visual inspection of the profile



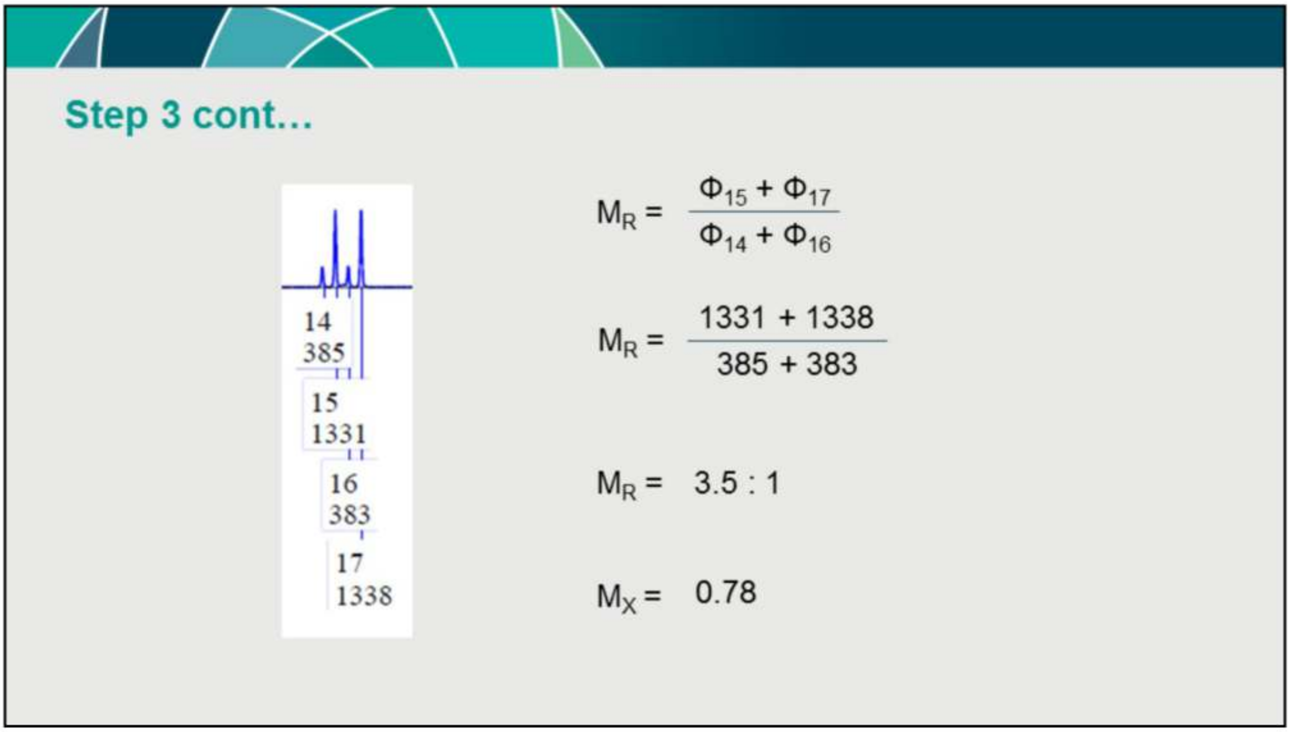
Step 3 cont.

- Where Φ_a is the peak height of allele a

$$M_X = \frac{\Phi_a + \Phi_b}{\Phi_a + \Phi_b + \Phi_c + \Phi_d}$$

- The value of M_R under the same conditions is:

$$M_R = \frac{\Phi_a + \Phi_b}{\Phi_c + \Phi_d}$$



Step 3 cont…

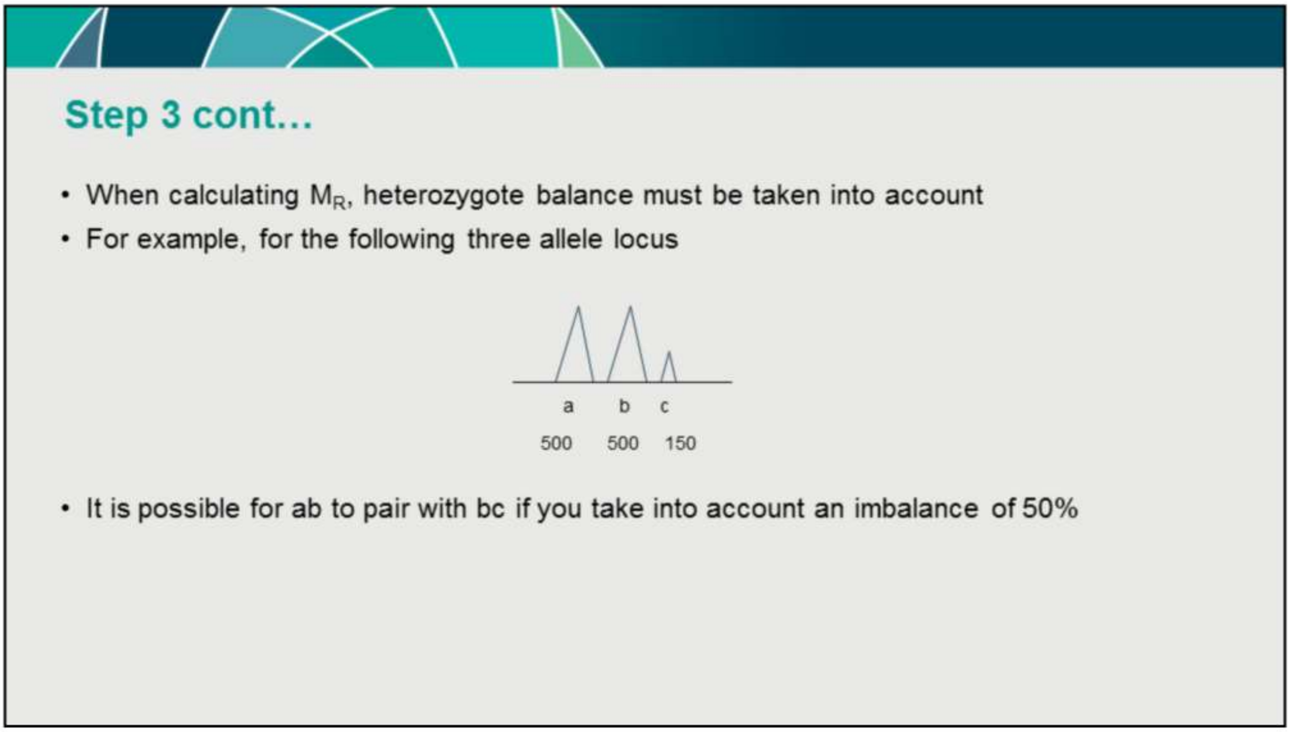
14
385
15
1331
16
383
17
1338

$$M_R = \frac{\Phi_{15} + \Phi_{17}}{\Phi_{14} + \Phi_{16}}$$

$$M_R = \frac{1331 + 1338}{385 + 383}$$

$M_R =$ 3.5 : 1

$M_X =$ 0.78

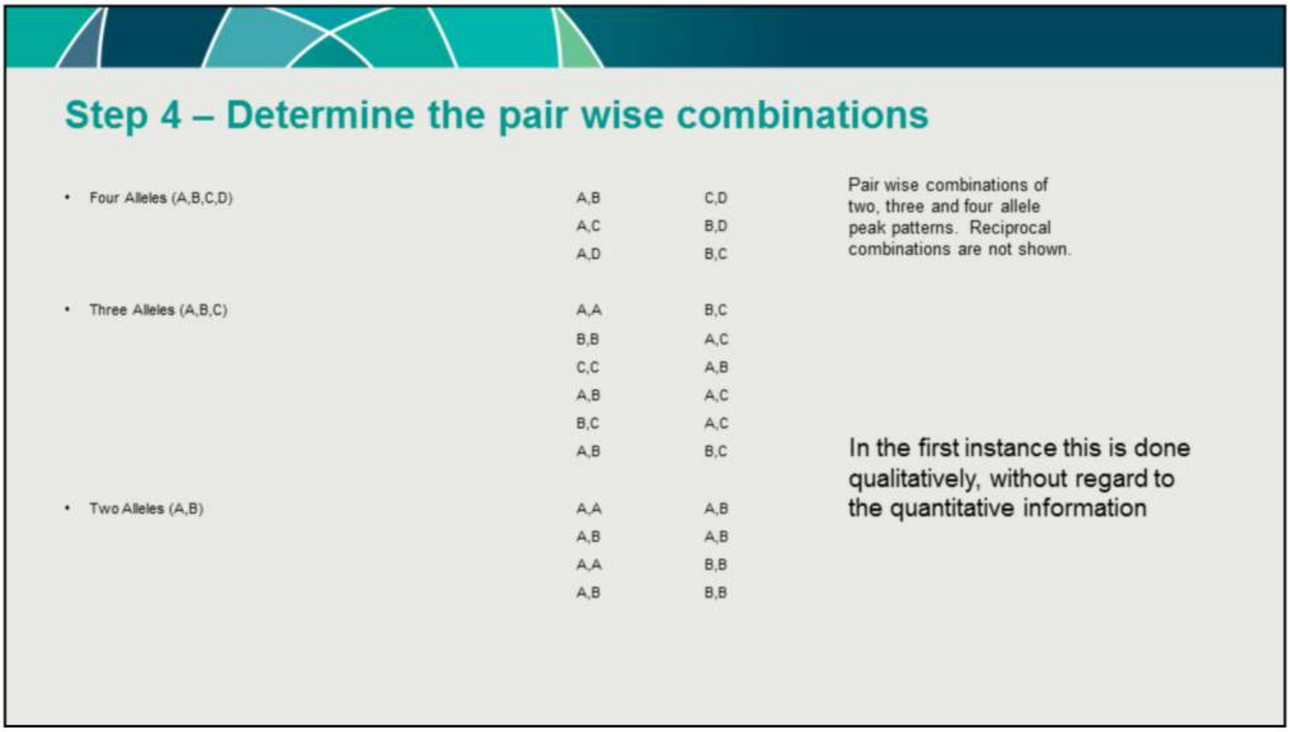


Step 3 cont…

- When calculating M_R, heterozygote balance must be taken into account
- For example, for the following three allele locus

a
b
c
500
500
150

- It is possible for ab to pair with bc if you take into account an imbalance of 50%

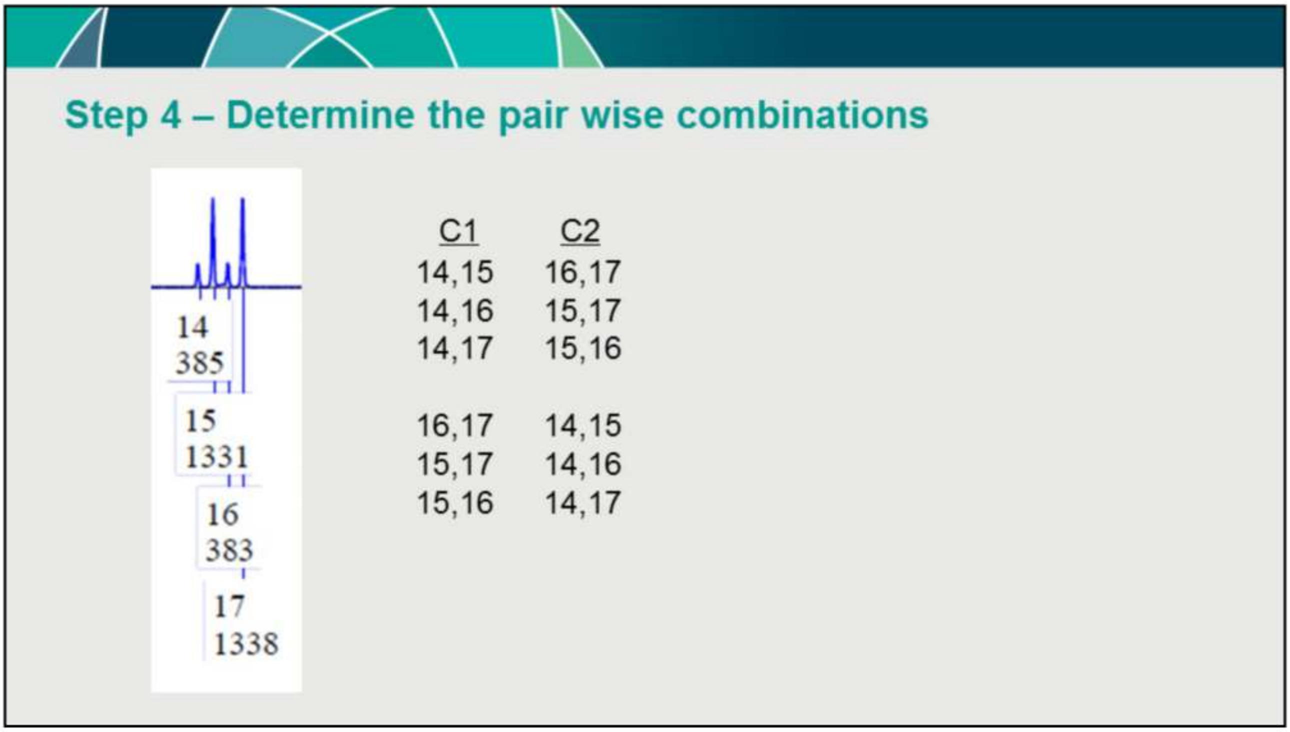


Step 4 – Determine the pair wise combinations

Pair wise combinations of two, three and four allele peak patterns. Reciprocal combinations are not shown.

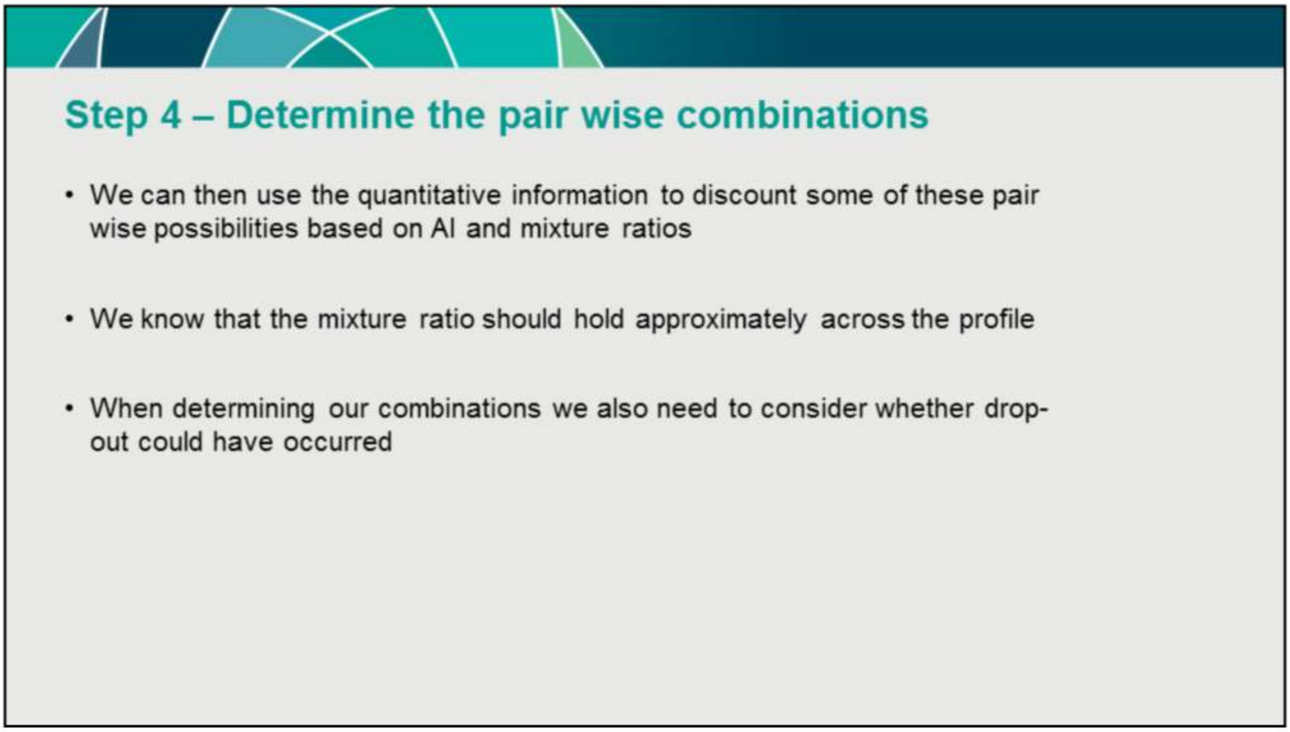
• Four Alleles (A,B,C,D)	A,B	C,D
	A,C	B,D
	A,D	B,C
• Three Alleles (A,B,C)	A,A	B,C
	B,B	A,C
	C,C	A,B
	A,B	A,C
	B,C	A,C
	A,B	B,C
• Two Alleles (A,B)	A,A	A,B
	A,B	A,B
	A,A	B,B
	A,B	B,B

In the first instance this is done qualitatively, without regard to the quantitative information



Step 4 – Determine the pair wise combinations

C1	C2
14,15	16,17
14,16	15,17
14,17	15,16
16,17	14,15
15,17	14,16
15,16	14,17

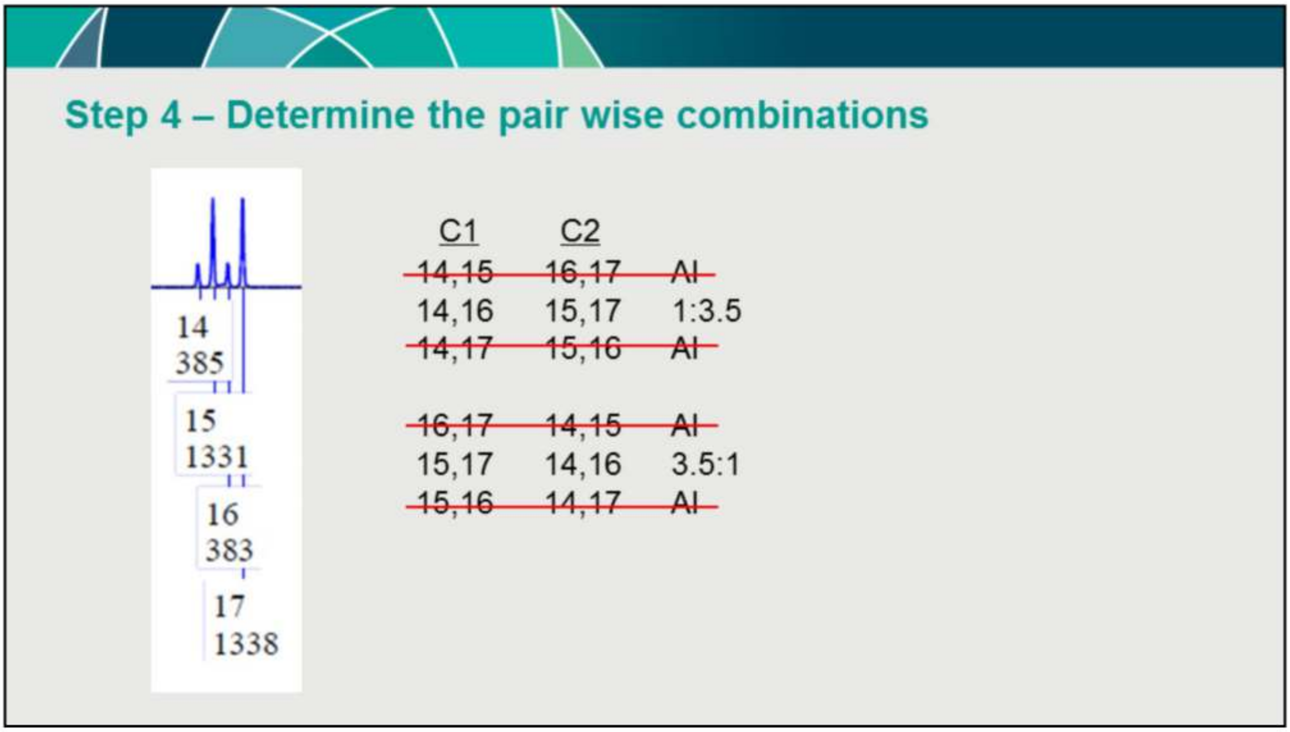


Step 4 – Determine the pair wise combinations

- We can then use the quantitative information to discount some of these pair wise possibilities based on AI and mixture ratios

- We know that the mixture ratio should hold approximately across the profile

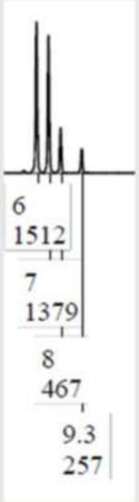
- When determining our combinations we also need to consider whether drop-out could have occurred

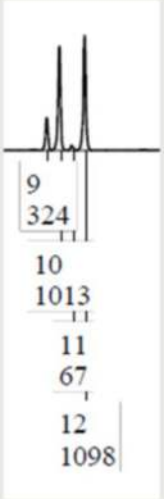


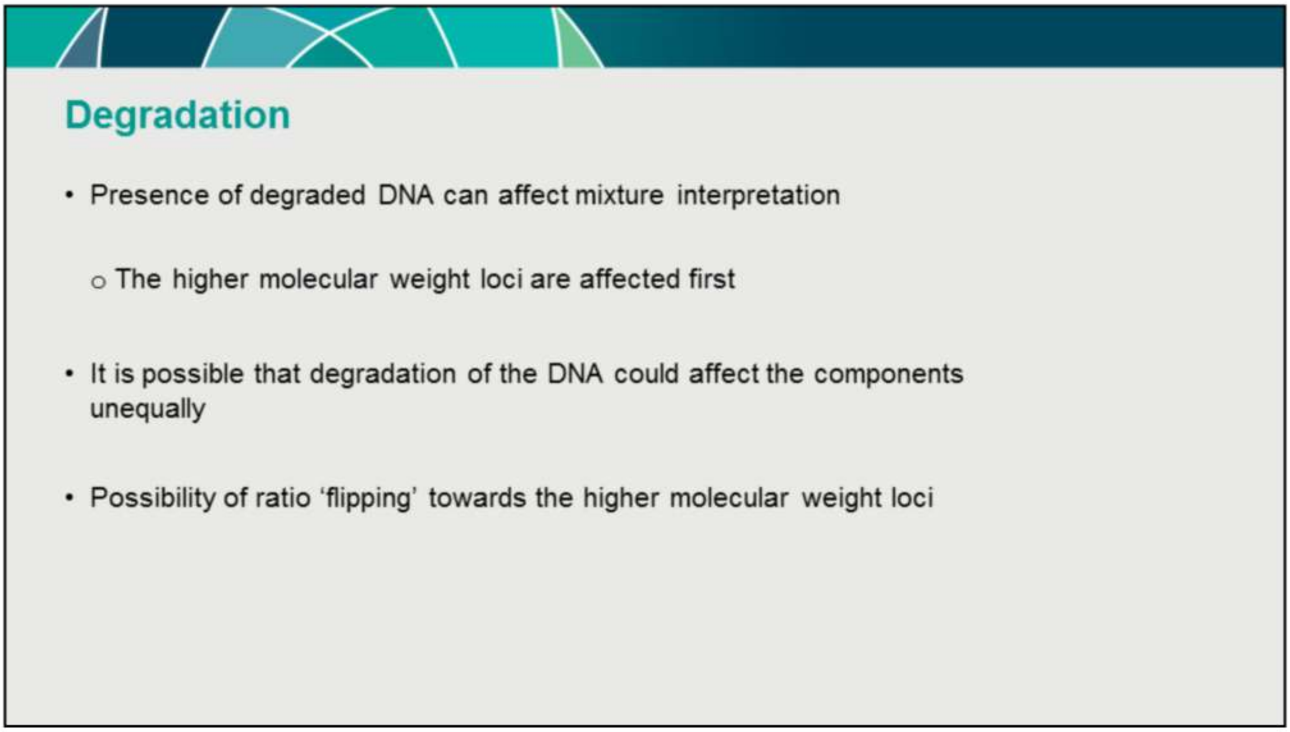
Step 4 – Determine the pair wise combinations

C1	C2	
~~14,15~~	~~16,17~~	~~AI~~
14,16	15,17	1:3.5
~~14,17~~	~~15,16~~	~~AI~~
~~16,17~~	~~14,15~~	~~AI~~
15,17	14,16	3.5:1
~~15,16~~	~~14,17~~	~~AI~~

Step 4 – Determine the pair wise combinations





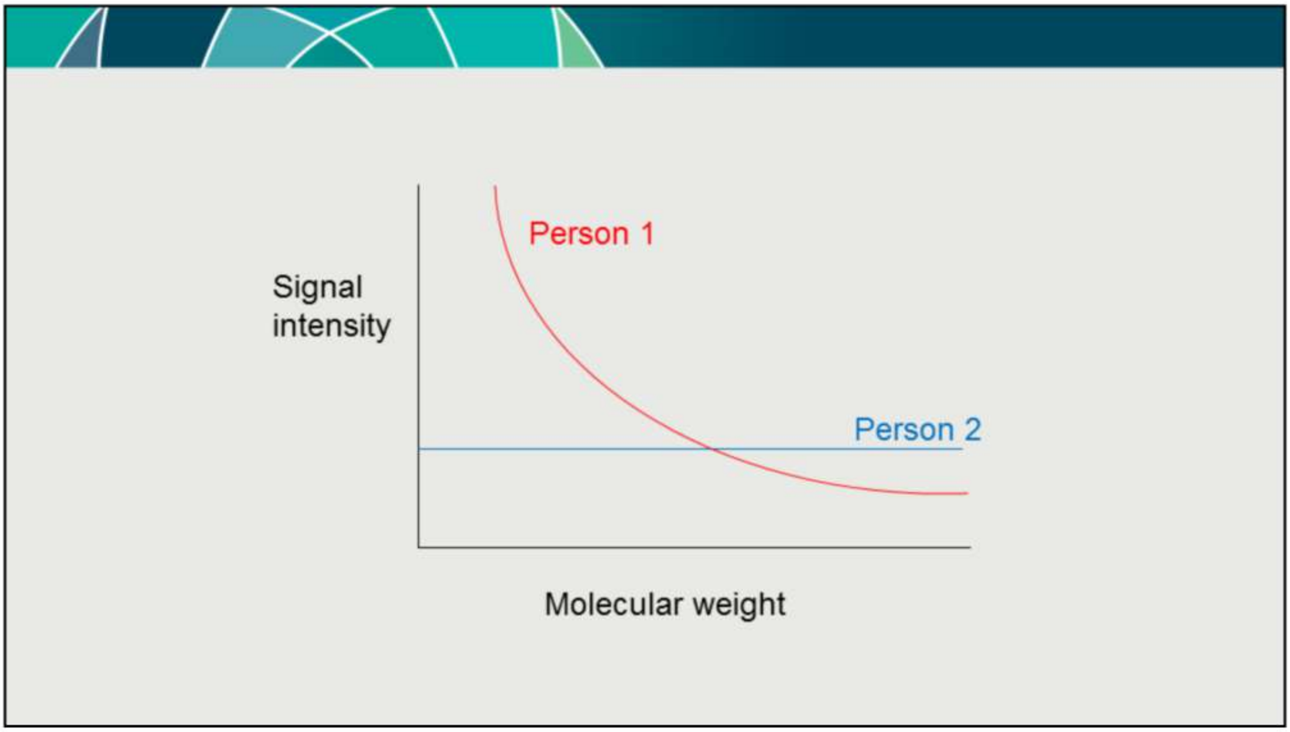


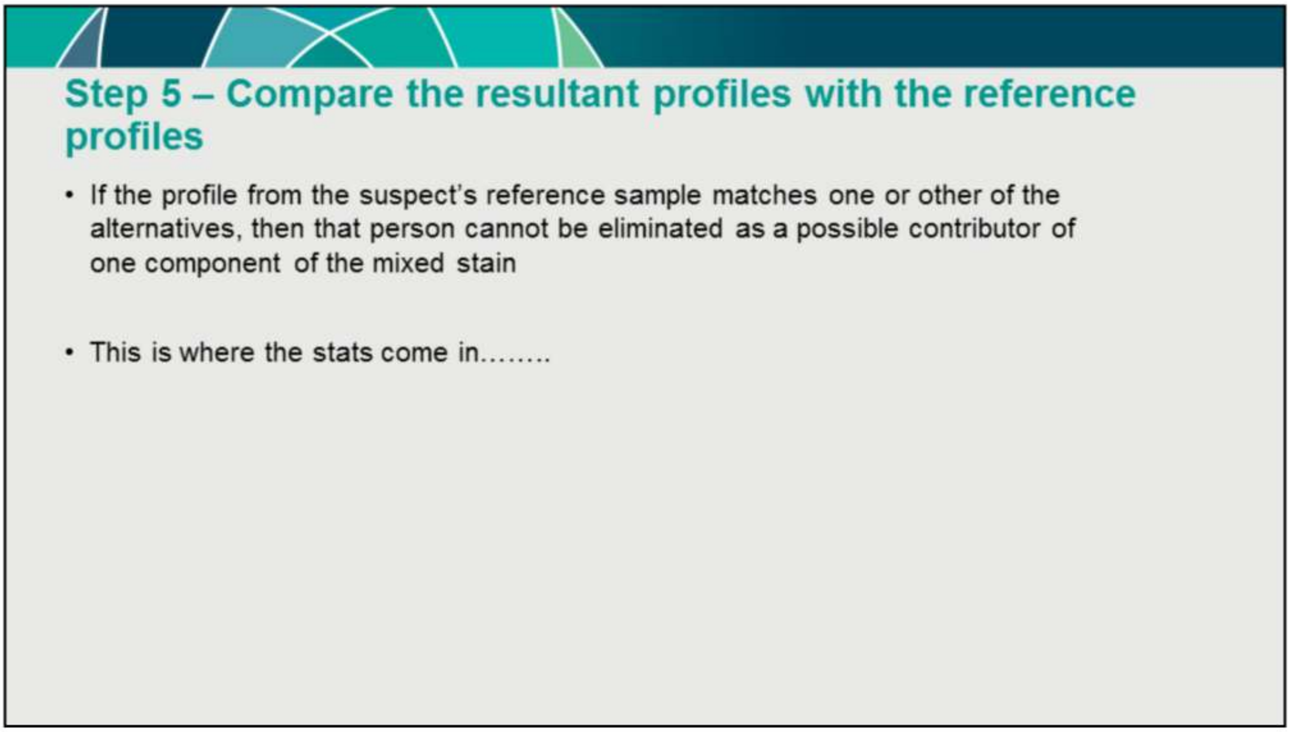
Degradation

- Presence of degraded DNA can affect mixture interpretation
  - The higher molecular weight loci are affected first

- It is possible that degradation of the DNA could affect the components unequally

- Possibility of ratio ‘flipping’ towards the higher molecular weight loci

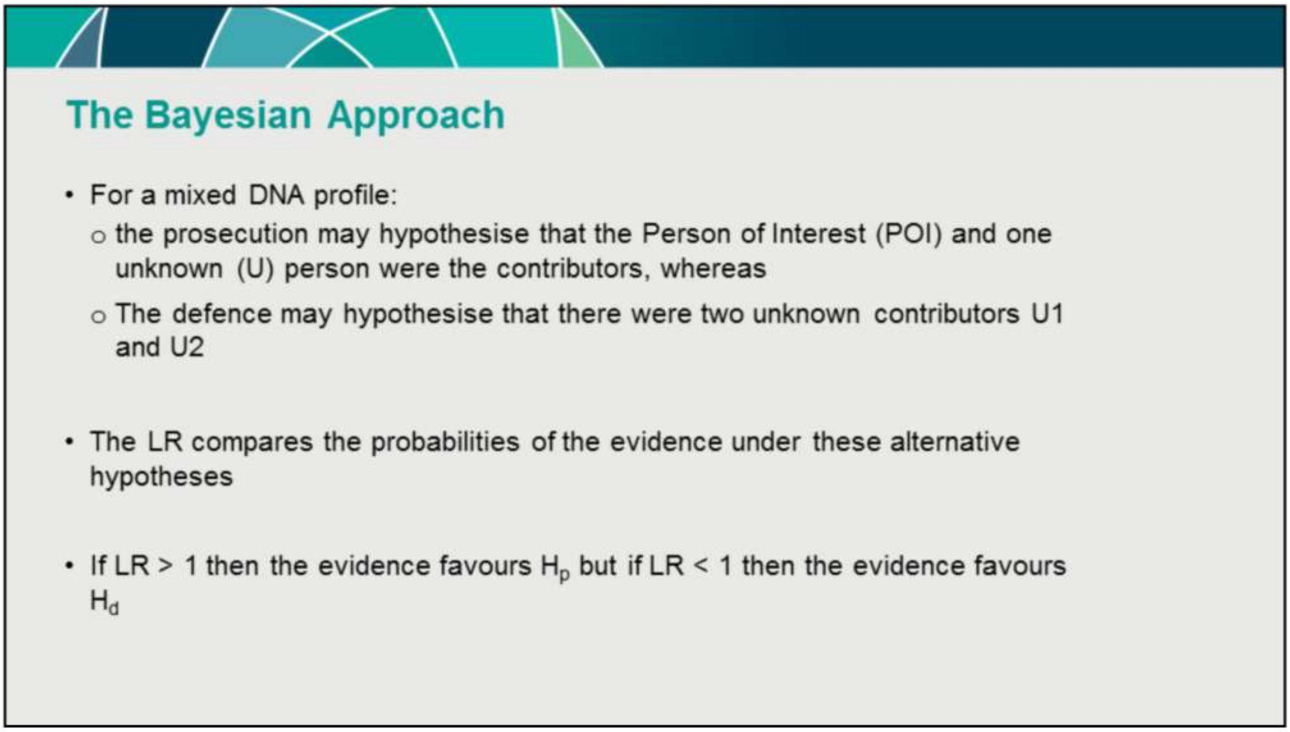

Person 1
Signal intensity
Person 2
Molecular weight



Step 5 – Compare the resultant profiles with the reference profiles

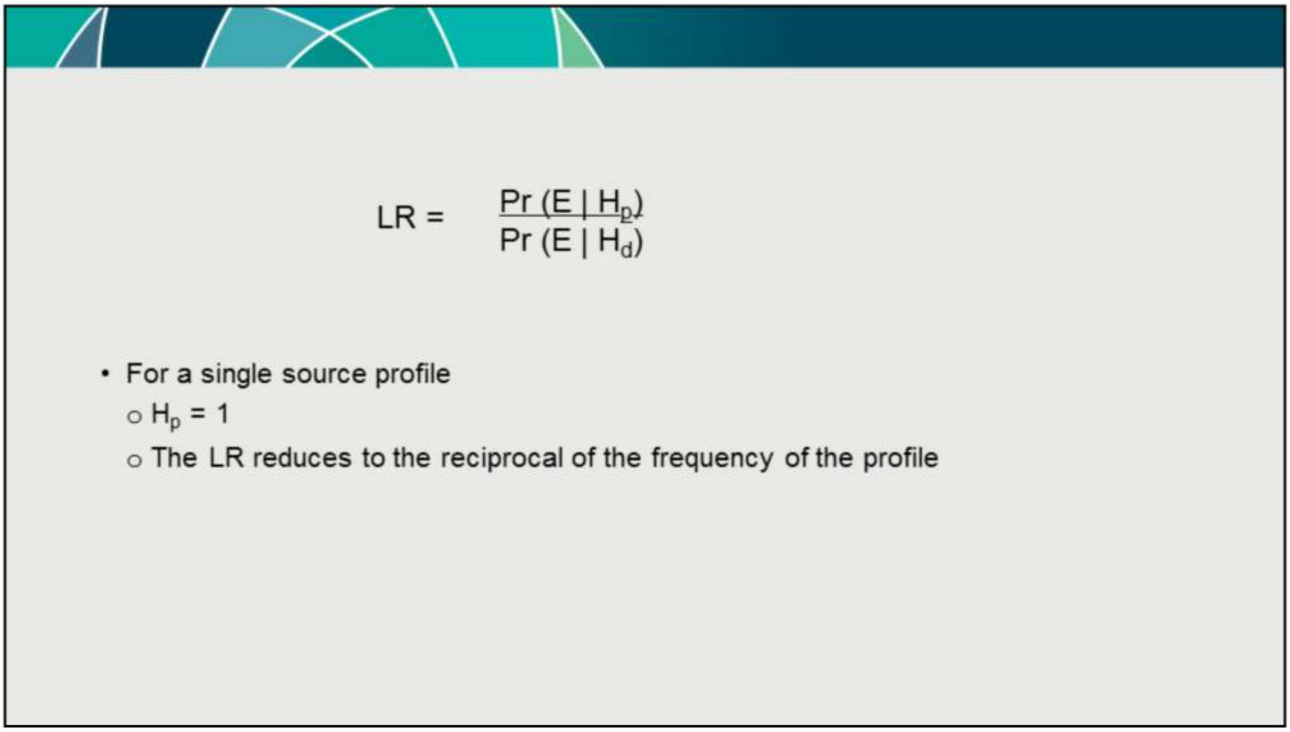
- If the profile from the suspect's reference sample matches one or other of the alternatives, then that person cannot be eliminated as a possible contributor of one component of the mixed stain

- This is where the stats come in……..



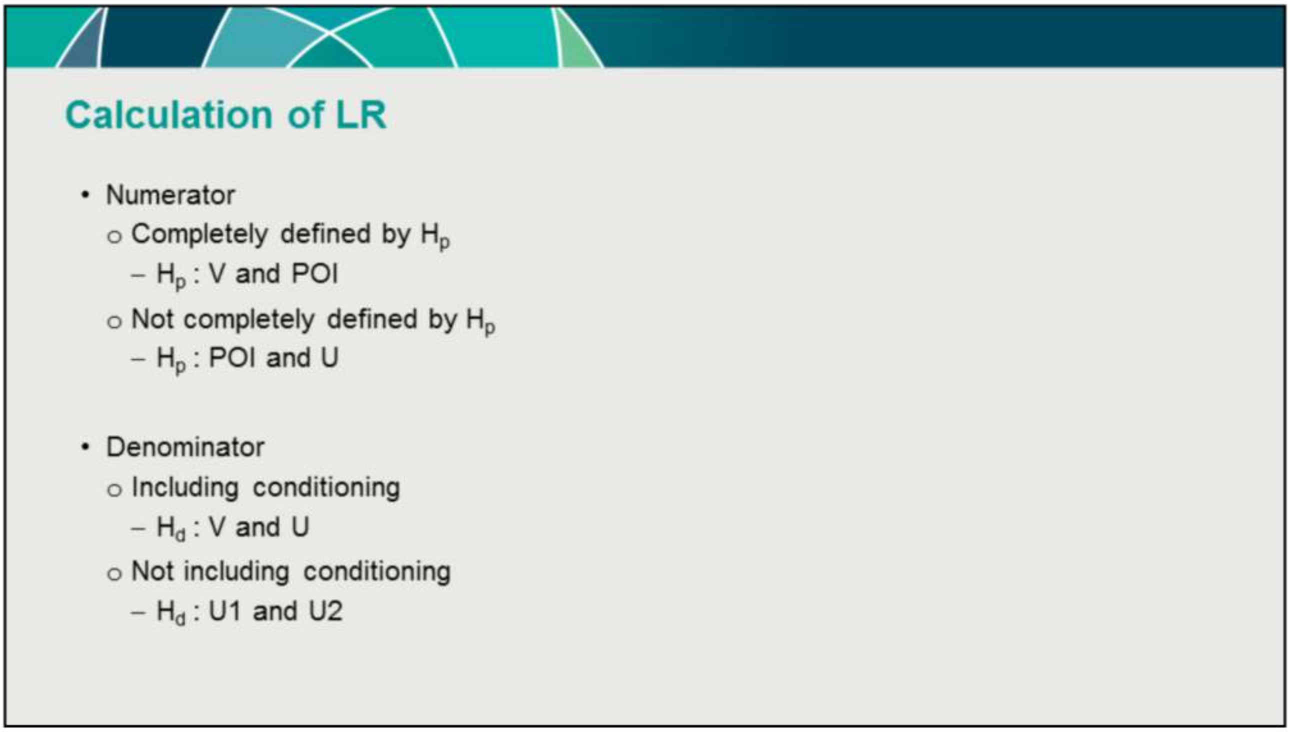
The Bayesian Approach

- For a mixed DNA profile:
  - the prosecution may hypothesise that the Person of Interest (POI) and one unknown (U) person were the contributors, whereas
  - The defence may hypothesise that there were two unknown contributors U1 and U2
- The LR compares the probabilities of the evidence under these alternative hypotheses
- If LR > 1 then the evidence favours H_p but if LR < 1 then the evidence favours H_d



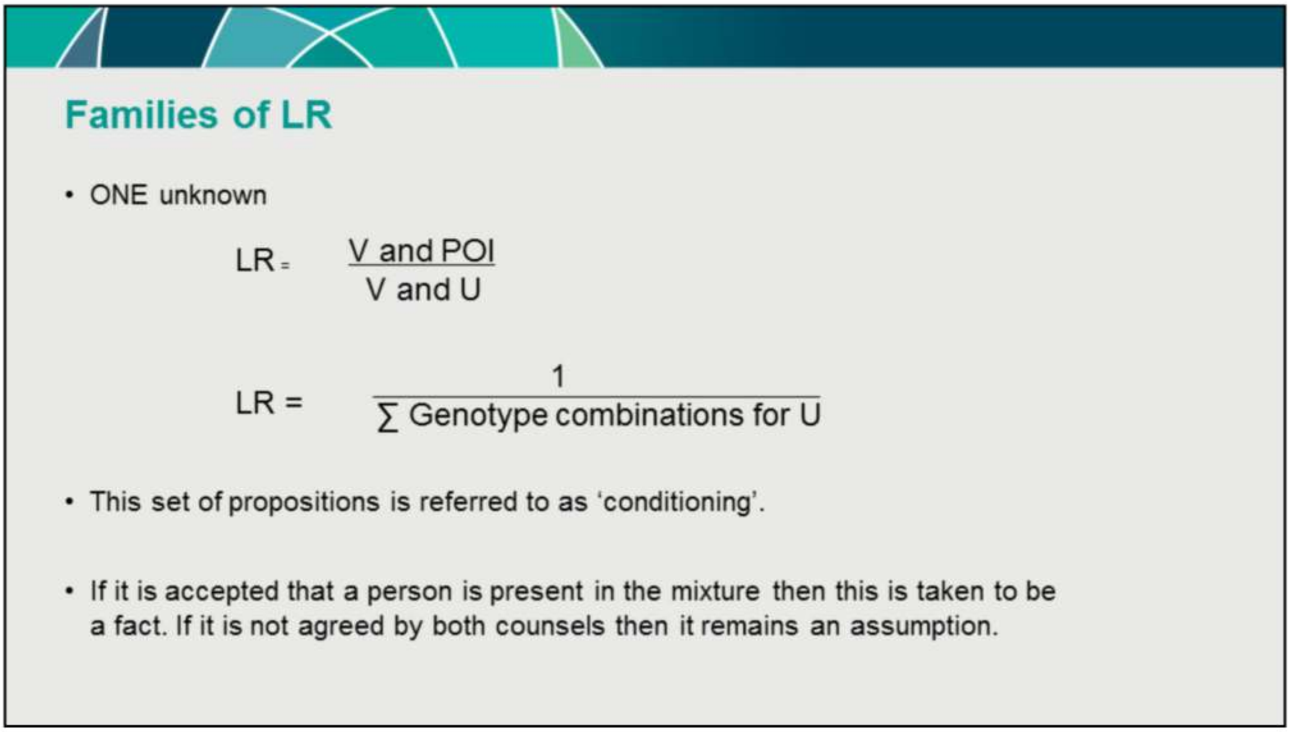
$$LR = \frac{Pr\ (E \mid H_p)}{Pr\ (E \mid H_d)}$$

- For a single source profile
  - $H_p = 1$
  - The LR reduces to the reciprocal of the frequency of the profile



Calculation of LR

- Numerator
  - o Completely defined by H_p
    - – H_p : V and POI
  - o Not completely defined by H_p
    - – H_p : POI and U
- Denominator
  - o Including conditioning
    - – H_d : V and U
  - o Not including conditioning
    - – H_d : U1 and U2



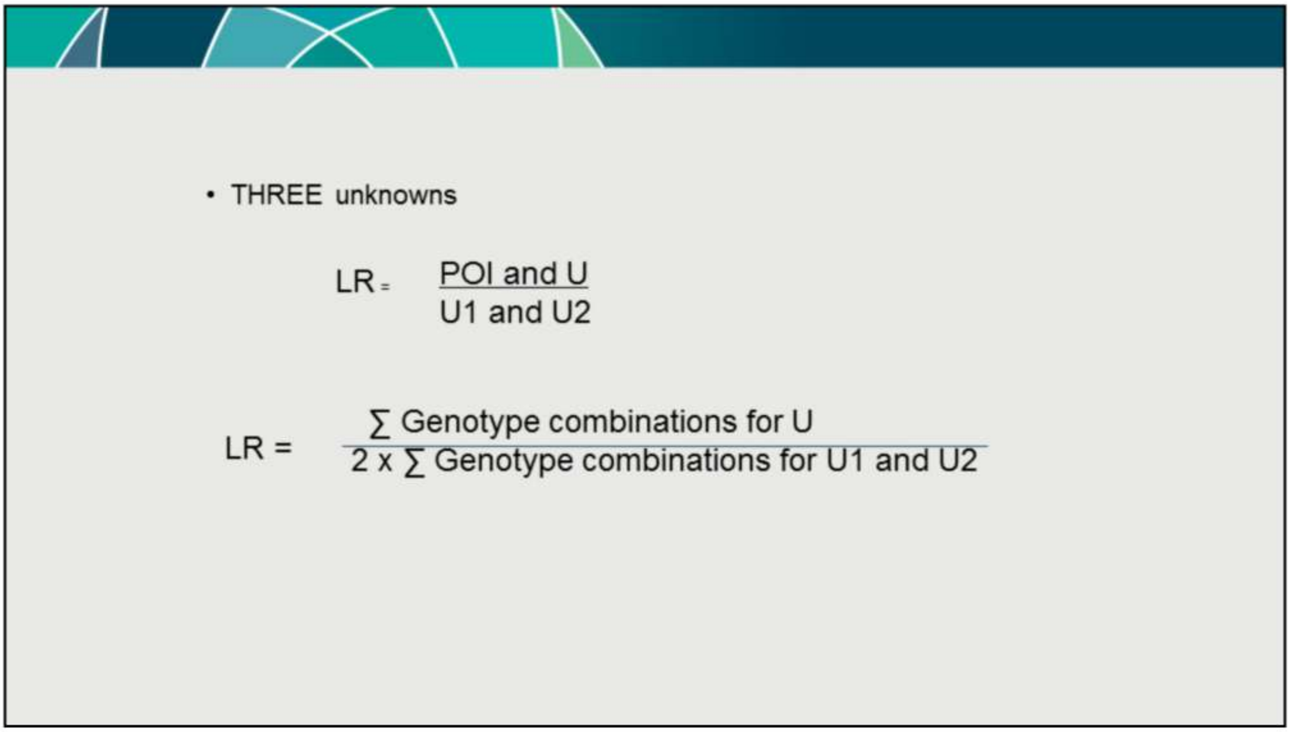
Families of LR

- ONE unknown

$$LR = \frac{\text{V and POI}}{\text{V and U}}$$

$$LR = \frac{1}{\sum \text{Genotype combinations for U}}$$

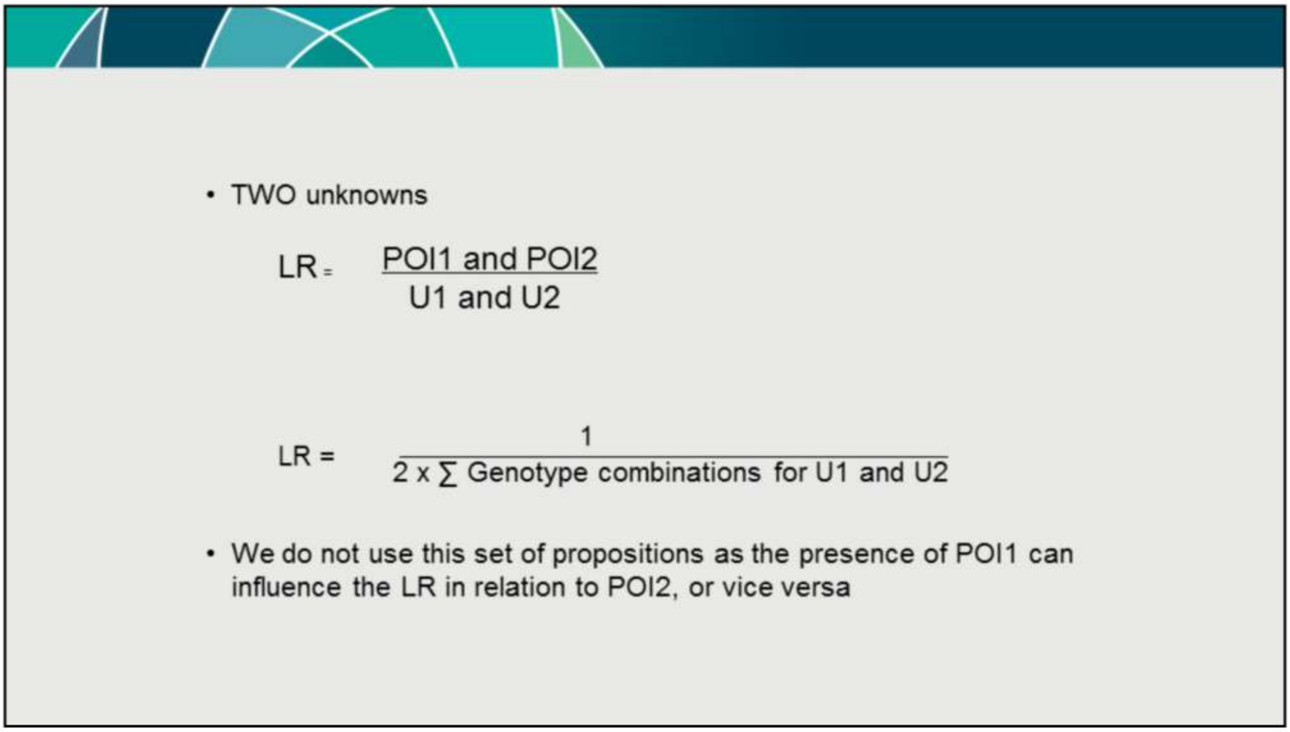
- This set of propositions is referred to as 'conditioning'.
- If it is accepted that a person is present in the mixture then this is taken to be a fact. If it is not agreed by both counsels then it remains an assumption.



- THREE unknowns

$$LR = \frac{\text{POI and U}}{\text{U1 and U2}}$$

$$LR = \frac{\sum \text{Genotype combinations for U}}{2 \times \sum \text{Genotype combinations for U1 and U2}}$$

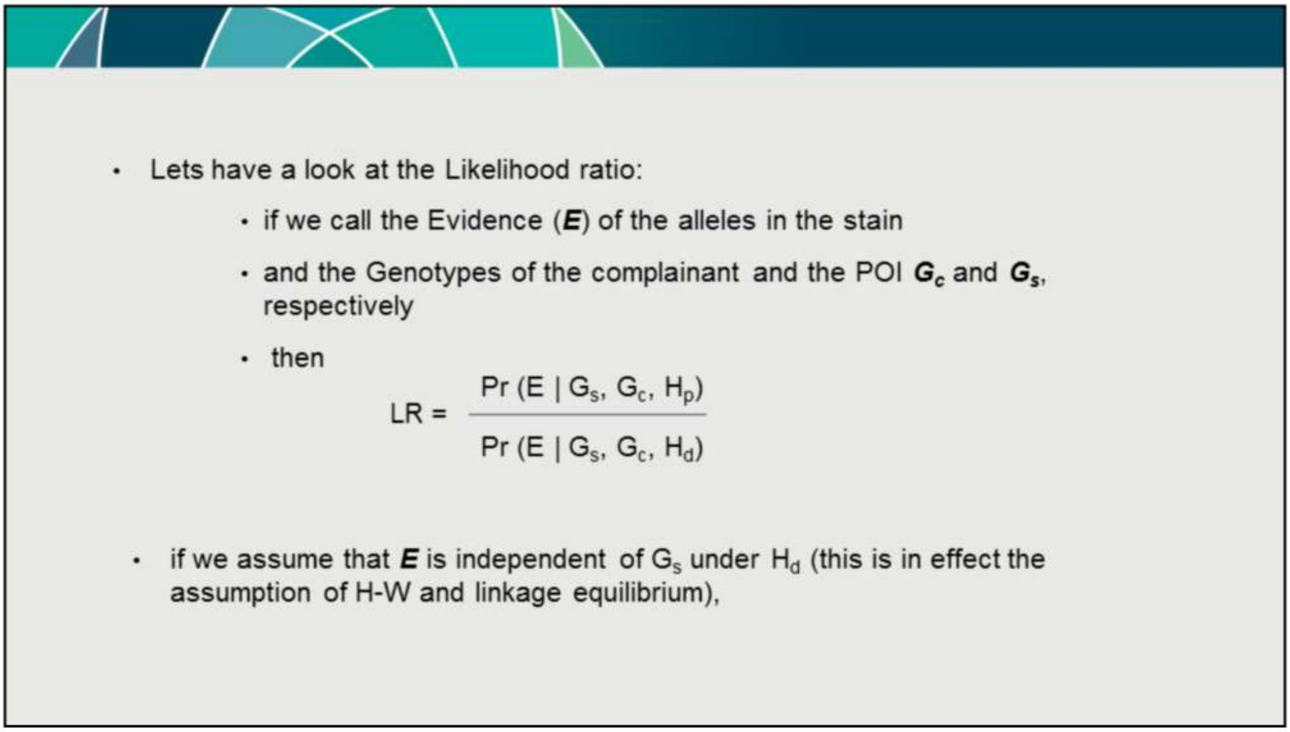


- TWO unknowns

$$\text{LR} = \frac{\text{POI1 and POI2}}{\text{U1 and U2}}$$

$$\text{LR} = \frac{1}{2 \text{ x } \sum \text{Genotype combinations for U1 and U2}}$$

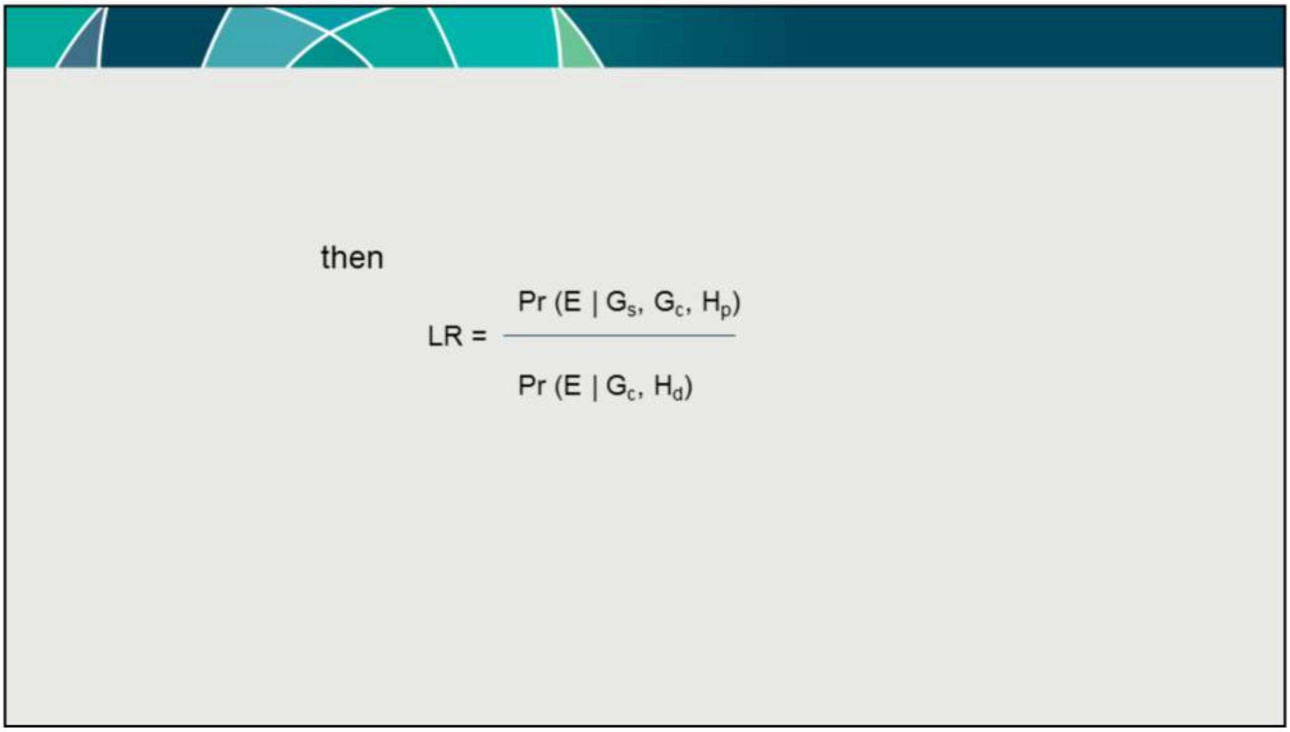
- We do not use this set of propositions as the presence of POI1 can influence the LR in relation to POI2, or vice versa



- Lets have a look at the Likelihood ratio:
  - if we call the Evidence (***E***) of the alleles in the stain
  - and the Genotypes of the complainant and the POI $\boldsymbol{G_c}$ and $\boldsymbol{G_s}$, respectively
  - then

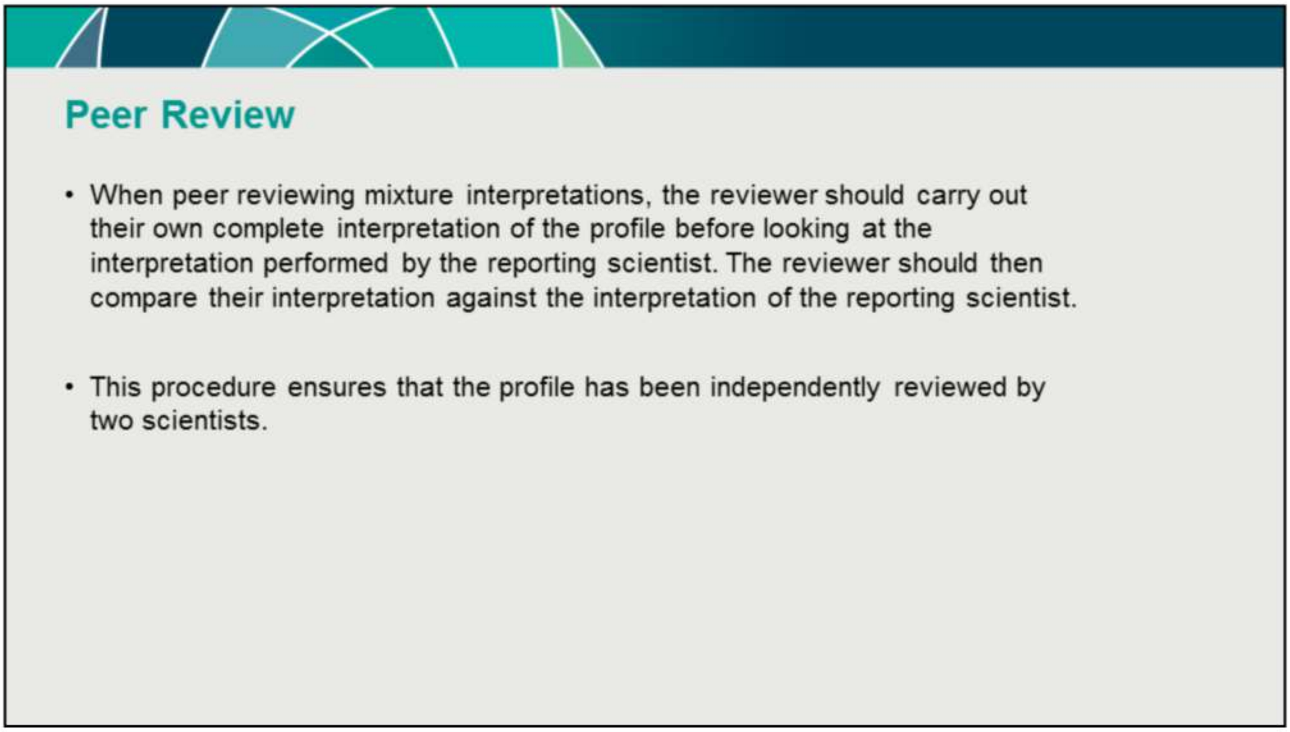
$$LR = \frac{Pr\ (E \mid G_s,\ G_c,\ H_p)}{Pr\ (E \mid G_s,\ G_c,\ H_d)}$$

- if we assume that ***E*** is independent of G_s under H_d (this is in effect the assumption of H-W and linkage equilibrium),



then

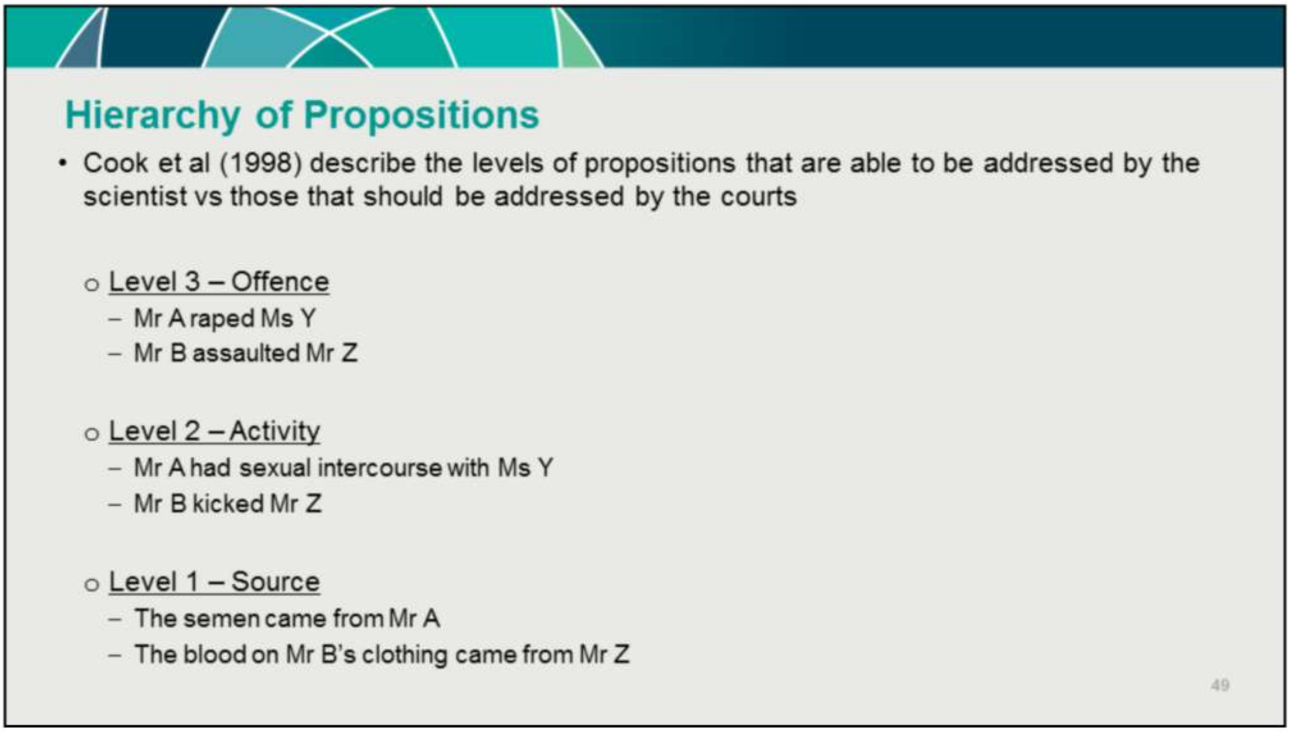
$$LR = \frac{Pr\ (E \mid G_s, G_c, H_p)}{Pr\ (E \mid G_c, H_d)}$$



Peer Review

- When peer reviewing mixture interpretations, the reviewer should carry out their own complete interpretation of the profile before looking at the interpretation performed by the reporting scientist. The reviewer should then compare their interpretation against the interpretation of the reporting scientist.

- This procedure ensures that the profile has been independently reviewed by two scientists.



Hierarchy of Propositions

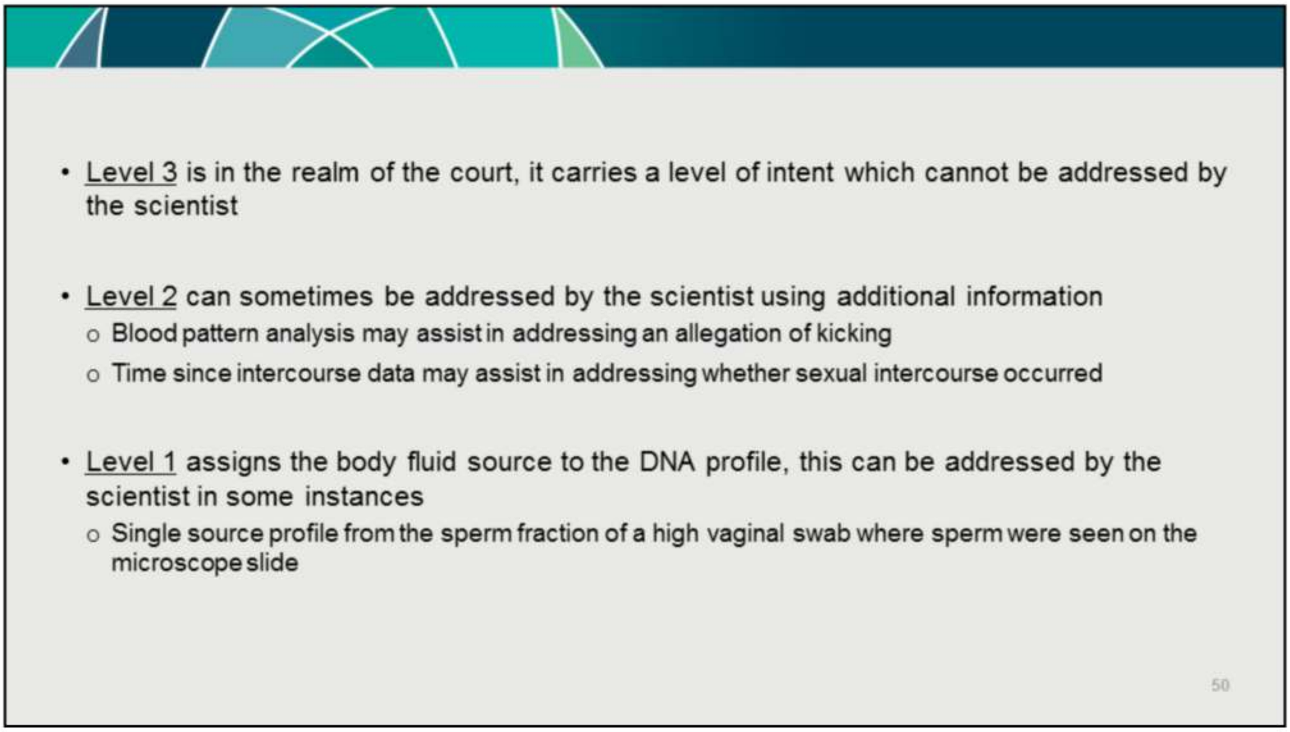
- Cook et al (1998) describe the levels of propositions that are able to be addressed by the scientist vs those that should be addressed by the courts

  - Level 3 – Offence
    - Mr A raped Ms Y
    - Mr B assaulted Mr Z

  - Level 2 – Activity
    - Mr A had sexual intercourse with Ms Y
    - Mr B kicked Mr Z

  - Level 1 – Source
    - The semen came from Mr A
    - The blood on Mr B's clothing came from Mr Z

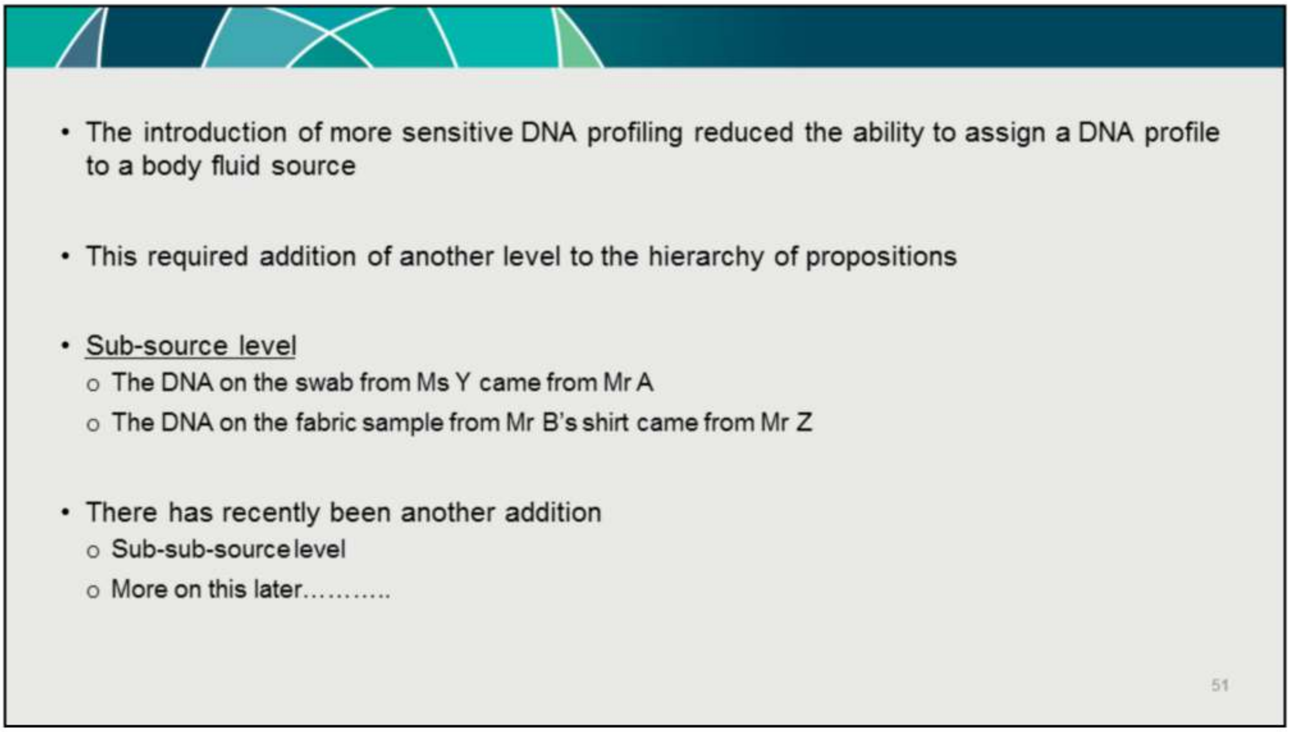
49



- Level 3 is in the realm of the court, it carries a level of intent which cannot be addressed by the scientist

- Level 2 can sometimes be addressed by the scientist using additional information
  - Blood pattern analysis may assist in addressing an allegation of kicking
  - Time since intercourse data may assist in addressing whether sexual intercourse occurred

- Level 1 assigns the body fluid source to the DNA profile, this can be addressed by the scientist in some instances
  - Single source profile from the sperm fraction of a high vaginal swab where sperm were seen on the microscope slide

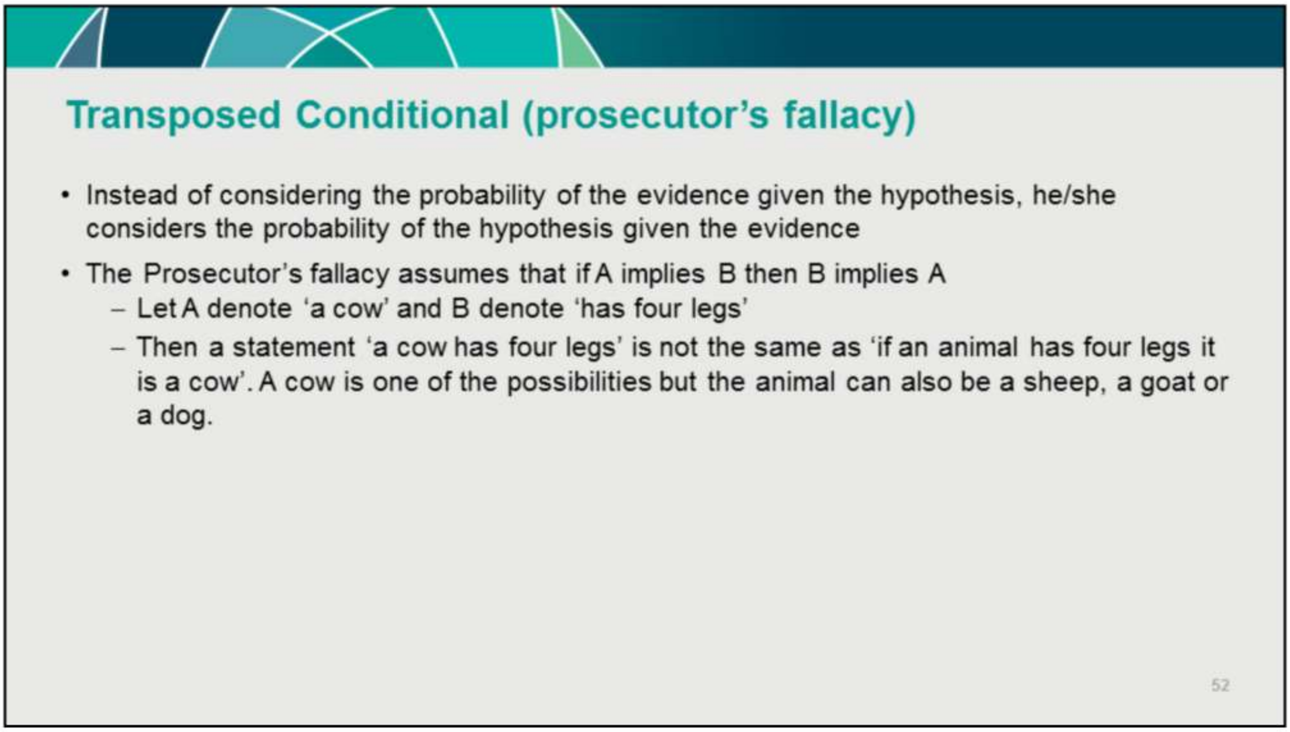


- The introduction of more sensitive DNA profiling reduced the ability to assign a DNA profile to a body fluid source

- This required addition of another level to the hierarchy of propositions

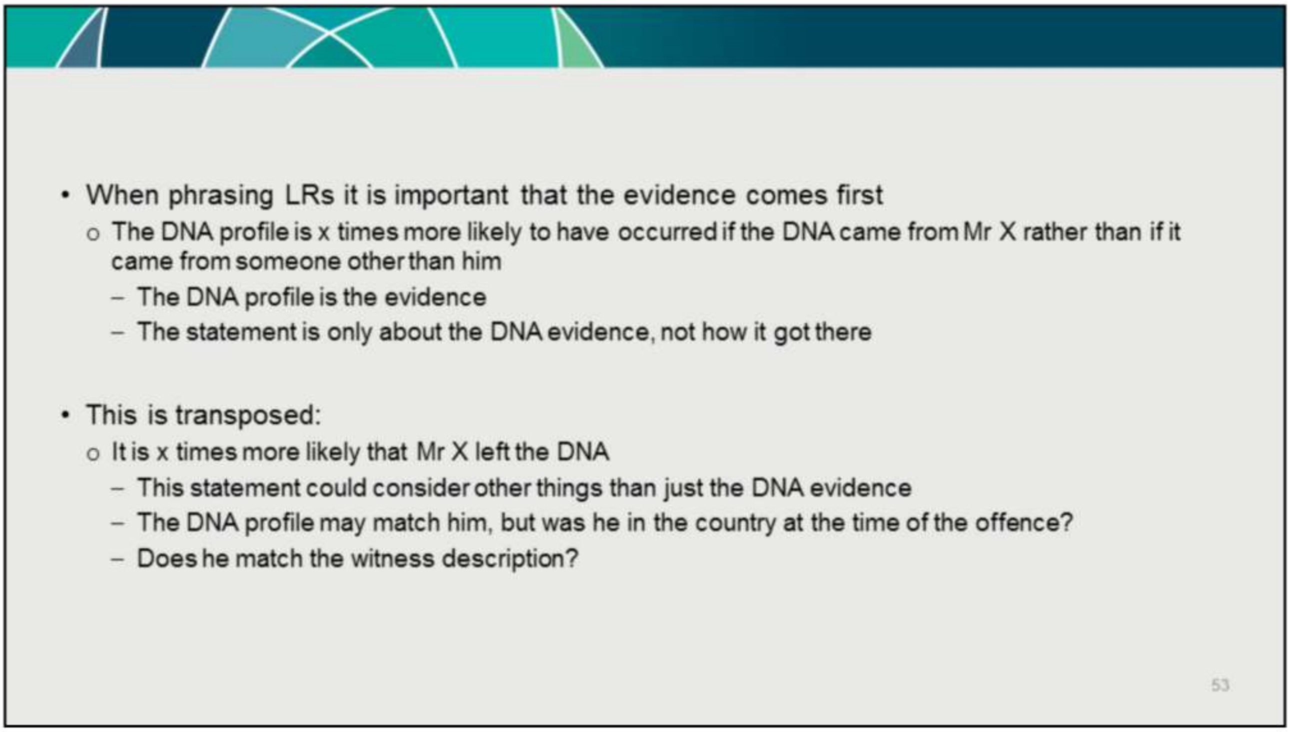
- Sub-source level
  - The DNA on the swab from Ms Y came from Mr A
  - The DNA on the fabric sample from Mr B's shirt came from Mr Z

- There has recently been another addition
  - Sub-sub-source level
  - More on this later………..



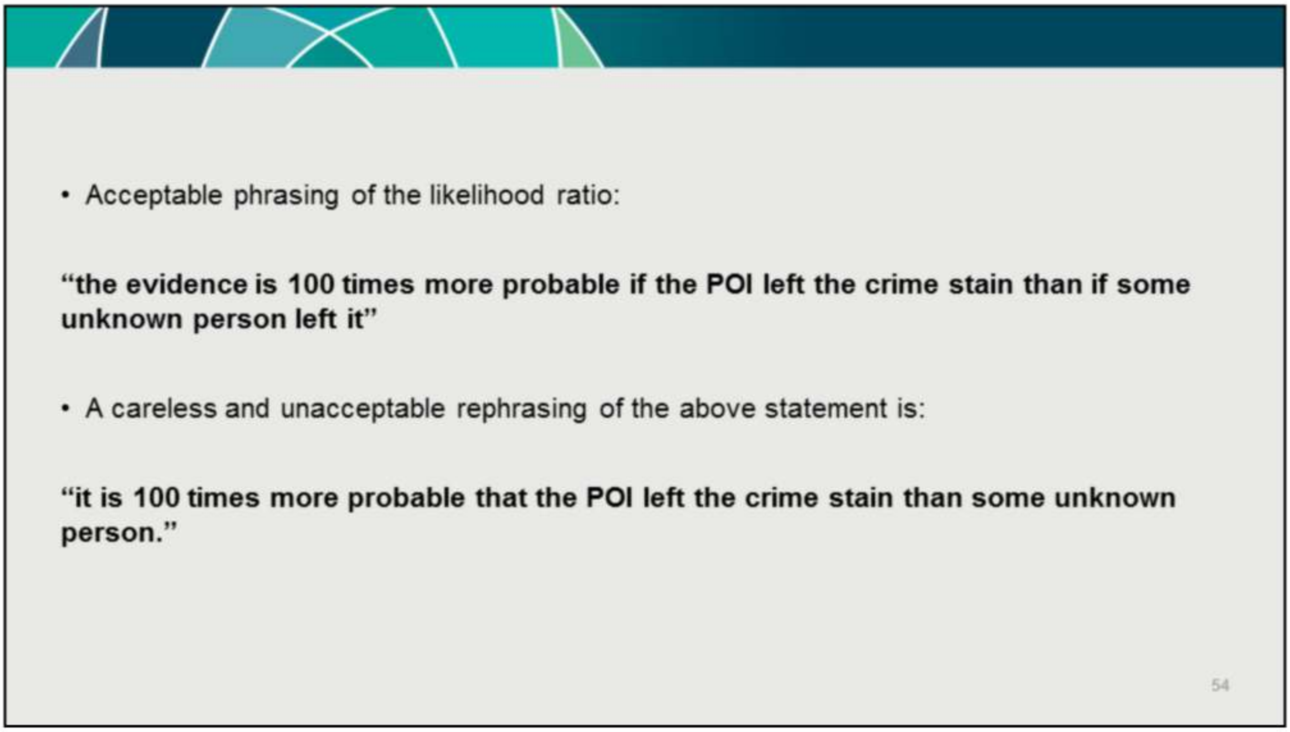
Transposed Conditional (prosecutor’s fallacy)

- Instead of considering the probability of the evidence given the hypothesis, he/she considers the probability of the hypothesis given the evidence
- The Prosecutor’s fallacy assumes that if A implies B then B implies A
  - Let A denote ‘a cow’ and B denote ‘has four legs’
  - Then a statement ‘a cow has four legs’ is not the same as ‘if an animal has four legs it is a cow’. A cow is one of the possibilities but the animal can also be a sheep, a goat or a dog.



- When phrasing LRs it is important that the evidence comes first
  - The DNA profile is x times more likely to have occurred if the DNA came from Mr X rather than if it came from someone other than him
    - The DNA profile is the evidence
    - The statement is only about the DNA evidence, not how it got there

- This is transposed:
  - It is x times more likely that Mr X left the DNA
    - This statement could consider other things than just the DNA evidence
    - The DNA profile may match him, but was he in the country at the time of the offence?
    - Does he match the witness description?

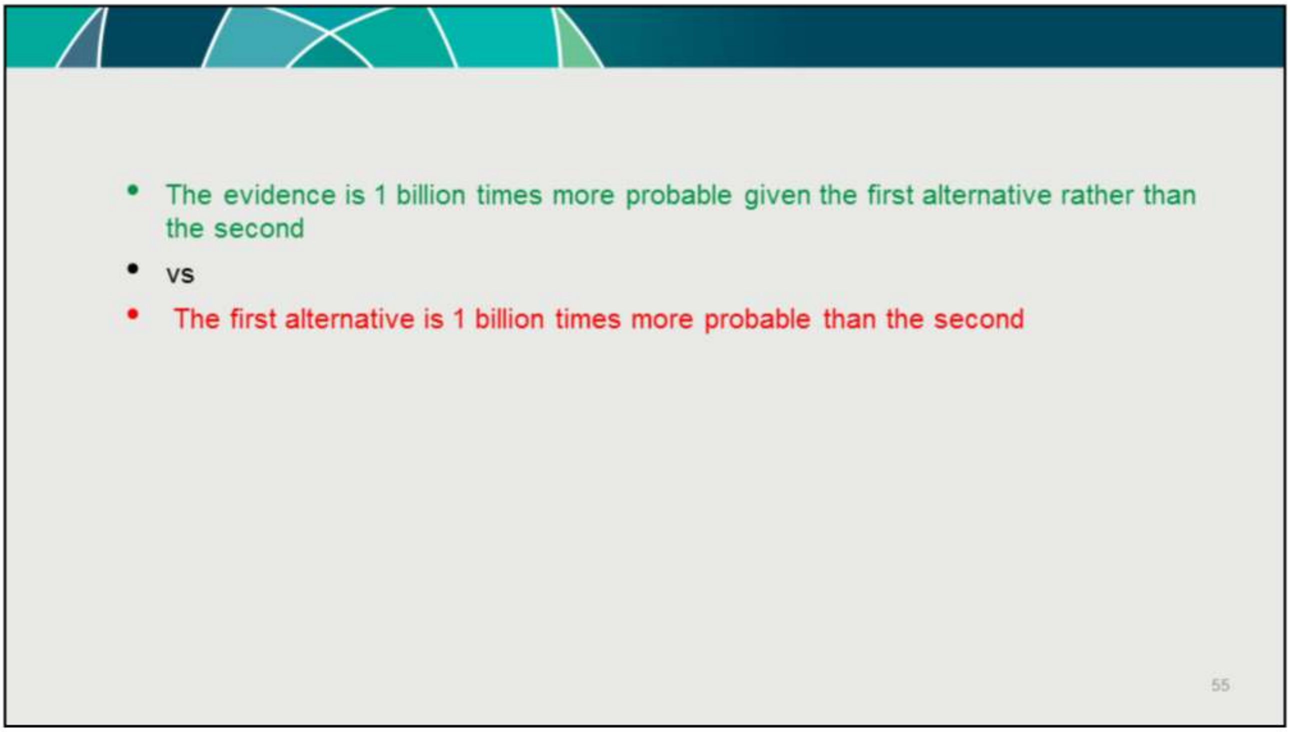


- Acceptable phrasing of the likelihood ratio:

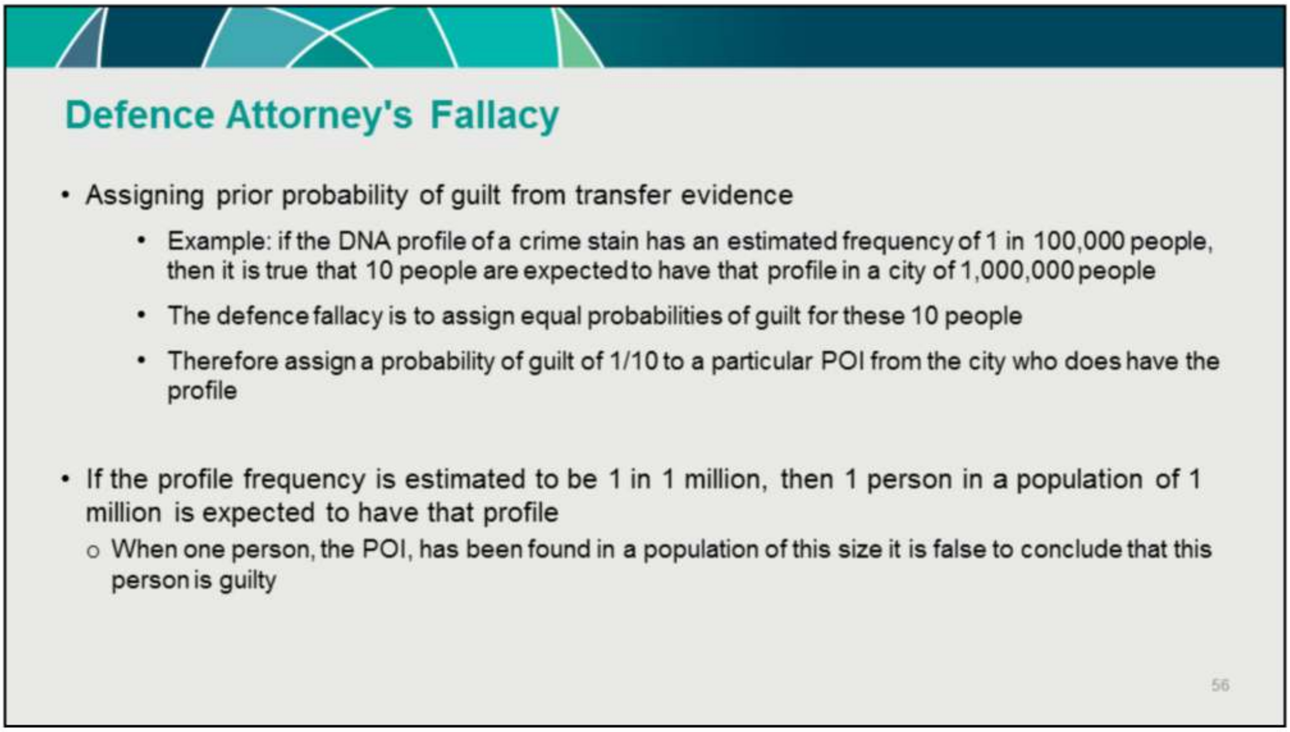
“the evidence is 100 times more probable if the POI left the crime stain than if some unknown person left it”

- A careless and unacceptable rephrasing of the above statement is:

“it is 100 times more probable that the POI left the crime stain than some unknown person.”



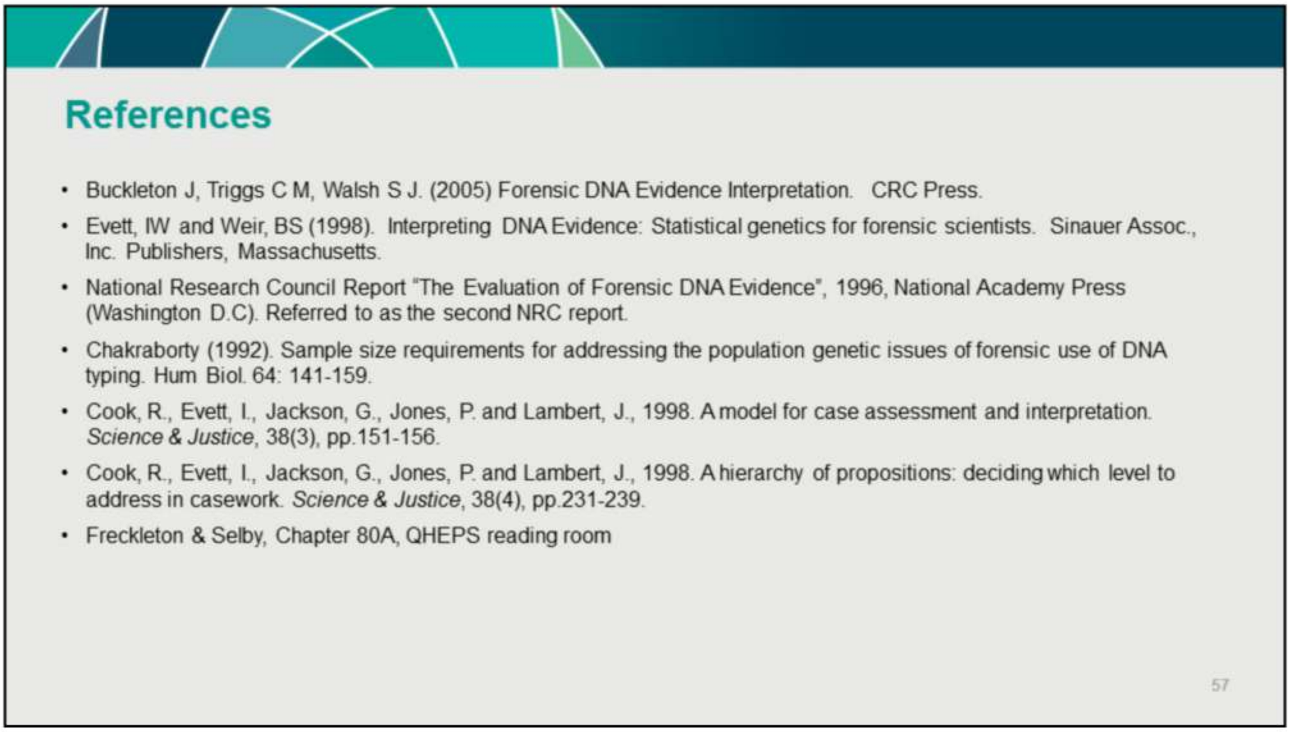
- The evidence is 1 billion times more probable given the first alternative rather than the second
- vs
- The first alternative is 1 billion times more probable than the second



Defence Attorney's Fallacy

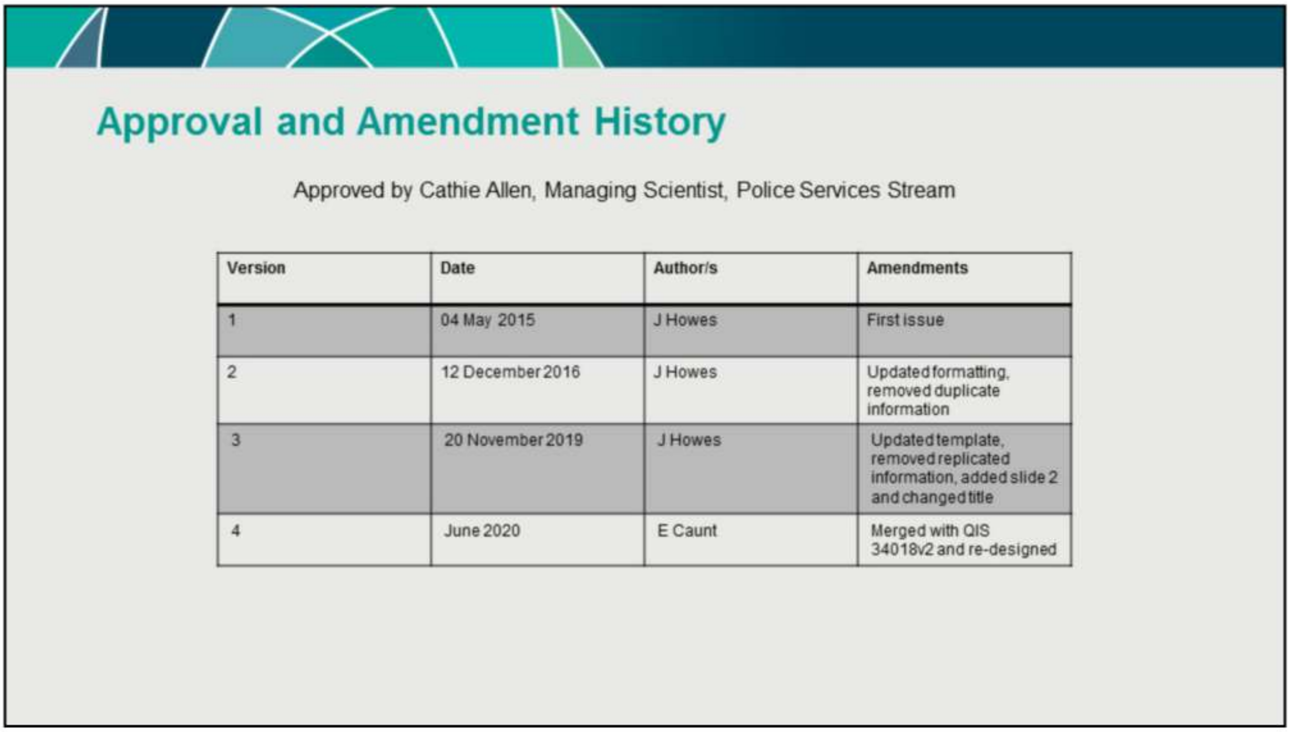
- Assigning prior probability of guilt from transfer evidence
  - Example: if the DNA profile of a crime stain has an estimated frequency of 1 in 100,000 people, then it is true that 10 people are expected to have that profile in a city of 1,000,000 people
  - The defence fallacy is to assign equal probabilities of guilt for these 10 people
  - Therefore assign a probability of guilt of 1/10 to a particular POI from the city who does have the profile

- If the profile frequency is estimated to be 1 in 1 million, then 1 person in a population of 1 million is expected to have that profile
  - o When one person, the POI, has been found in a population of this size it is false to conclude that this person is guilty



References

- Buckleton J, Triggs C M, Walsh S J. (2005) Forensic DNA Evidence Interpretation. CRC Press.
- Evett, IW and Weir, BS (1998). Interpreting DNA Evidence: Statistical genetics for forensic scientists. Sinauer Assoc., Inc. Publishers, Massachusetts.
- National Research Council Report "The Evaluation of Forensic DNA Evidence", 1996, National Academy Press (Washington D.C). Referred to as the second NRC report.
- Chakraborty (1992). Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. Hum Biol. 64: 141-159.
- Cook, R., Evett, I., Jackson, G., Jones, P. and Lambert, J., 1998. A model for case assessment and interpretation. *Science & Justice*, 38(3), pp.151-156.
- Cook, R., Evett, I., Jackson, G., Jones, P. and Lambert, J., 1998. A hierarchy of propositions: deciding which level to address in casework. *Science & Justice*, 38(4), pp.231-239.
- Freckleton & Selby, Chapter 80A, QHEPS reading room



Approval and Amendment History

Approved by Cathie Allen, Managing Scientist, Police Services Stream

Version	Date	Author/s	Amendments
1	04 May 2015	J Howes	First issue
2	12 December 2016	J Howes	Updated formatting, removed duplicate information
3	20 November 2019	J Howes	Updated template, removed replicated information, added slide 2 and changed title
4	June 2020	E Caunt	Merged with QIS 34018v2 and re-designed