Section 8: Incorporating Relatives

LR Challenges

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

LR Validation

Probabilistic genotyping systems need to be validated. Part of such validation studies focus on sensitivity and specificity analyses, to explore the range of expected LRs.

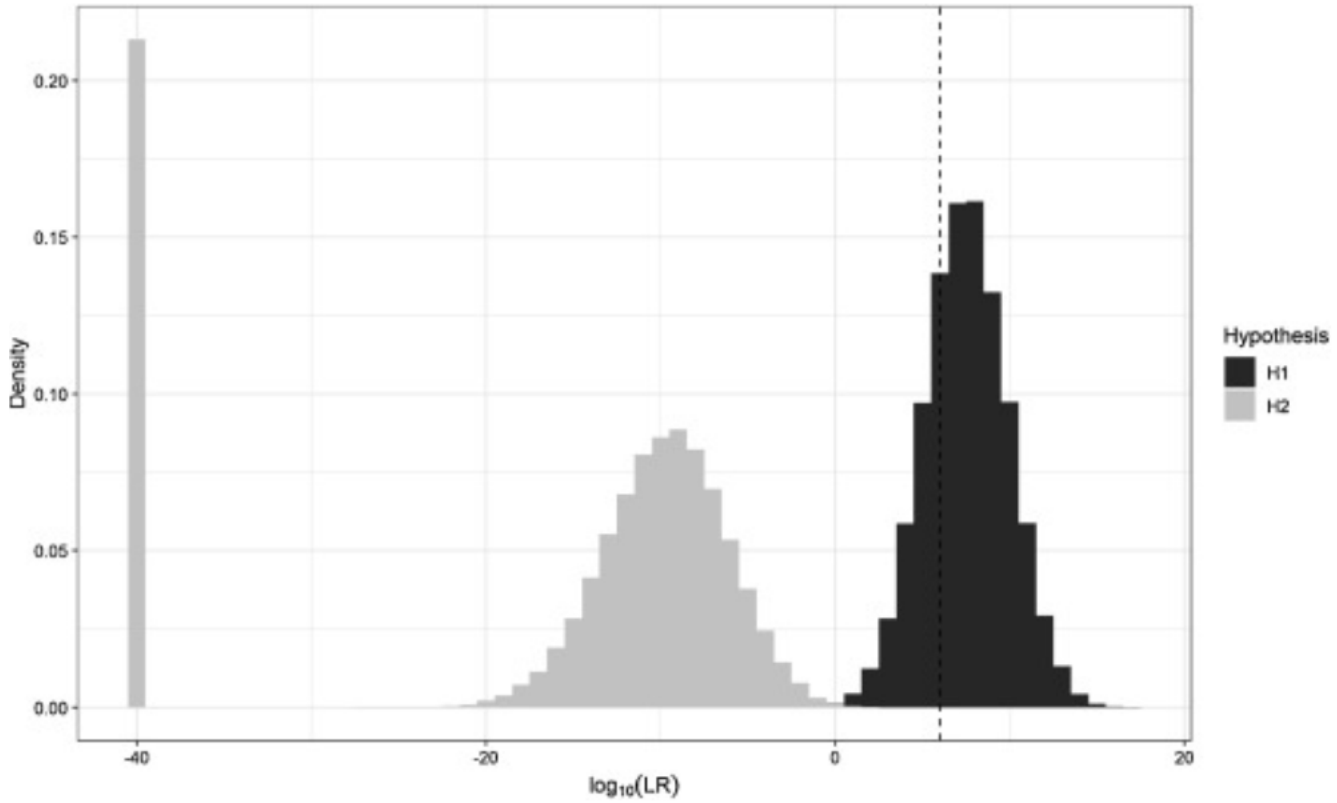
- **Sensitivity:** The ability to reliably associate a true contributor with a DNA profile (true donor LRs, H_p true tests).

- **Specificity:** The ability to reliable exclude non-contributors (false donor testing, H_d true tests).

These analyses can be based on real cases or by means of simulations.

LR Validation

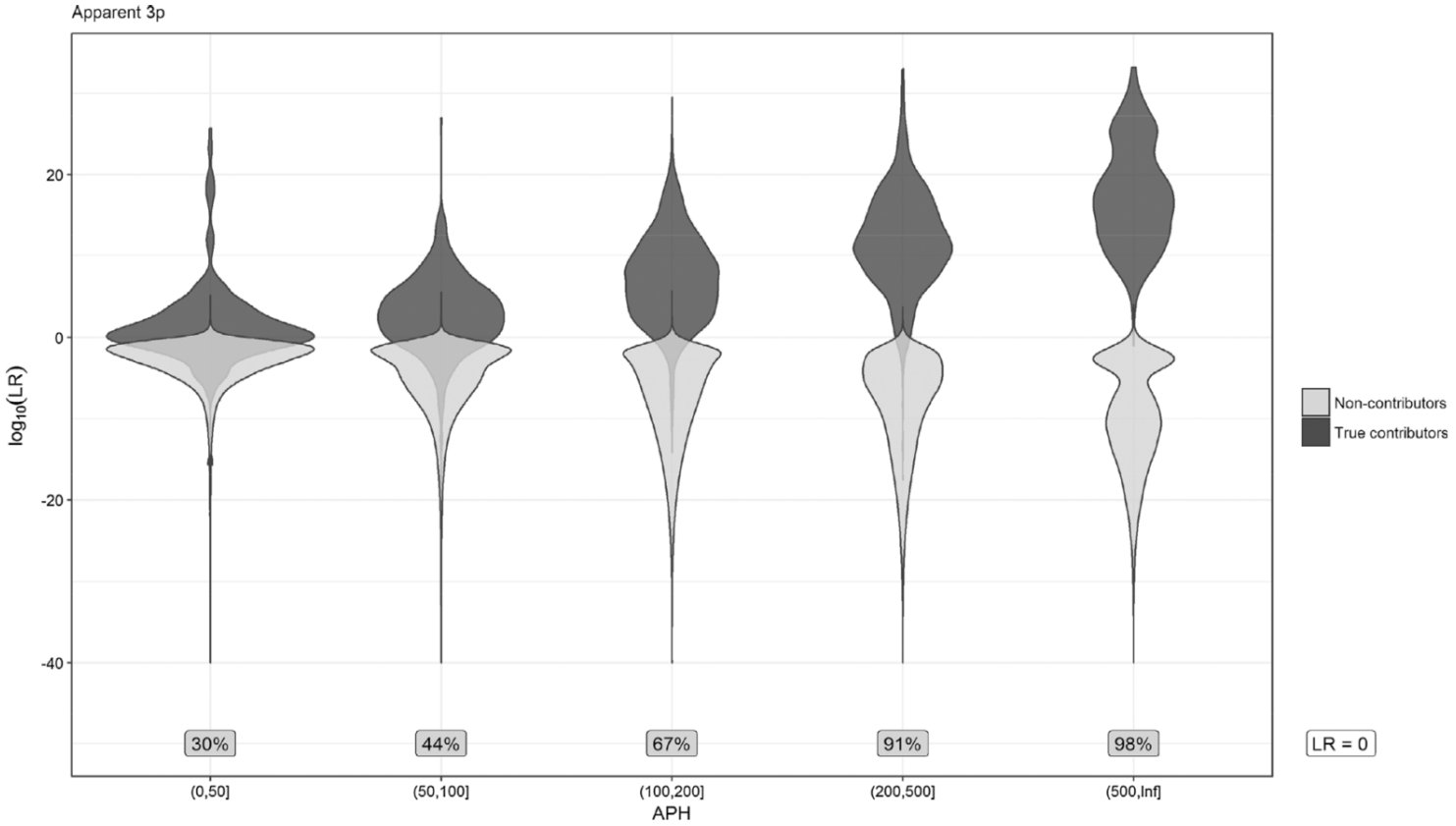
Ideally, the distribution of true contributor LRs is separate from the distribution of non-contributor LRs.



Source: Exploring the probative value of mixed DNA profiles (Kruijver et al., 2019).

LR Validation

The power to discriminate contributors from non-contributors depends on the quality and complexity of the sample.

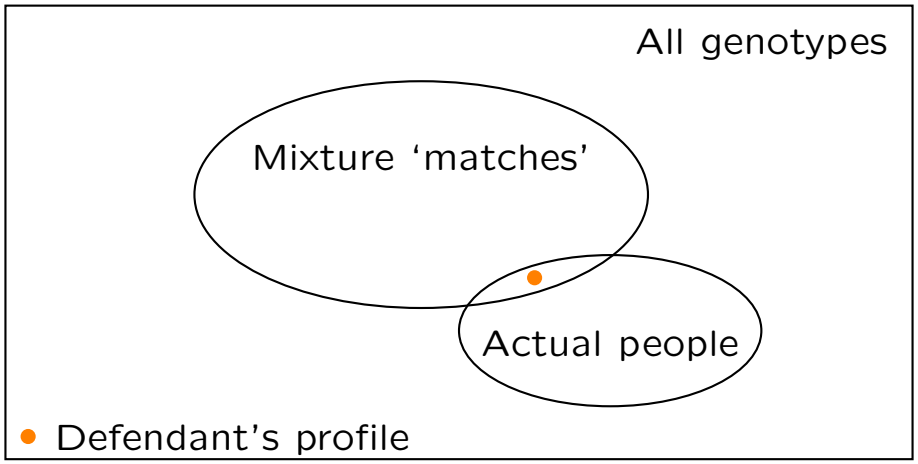


Source: Internal validation of STRmix^{TM} A multi laboratory response to PCAST (Bright et al., 2018).

Adventitious Matches

- What if there are genotypes that will result in high(er) LRs?
- Only in case of a very clear DNA profile will the true donor result in the highest LR (but such profiles are rarely observed from crime scene samples).
- There are possibly millions of other genotypes that are concordant with a mixture.
- If we would rank the LRs, the suspect is unlikely to produce the highest LR.
- This means that there are other genotypes that fit the data better, and provide more support for the prosecution hypothesis.

Most Genotypes Do Not Exist



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

Relatives

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

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

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

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

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

Mendel's Laws

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

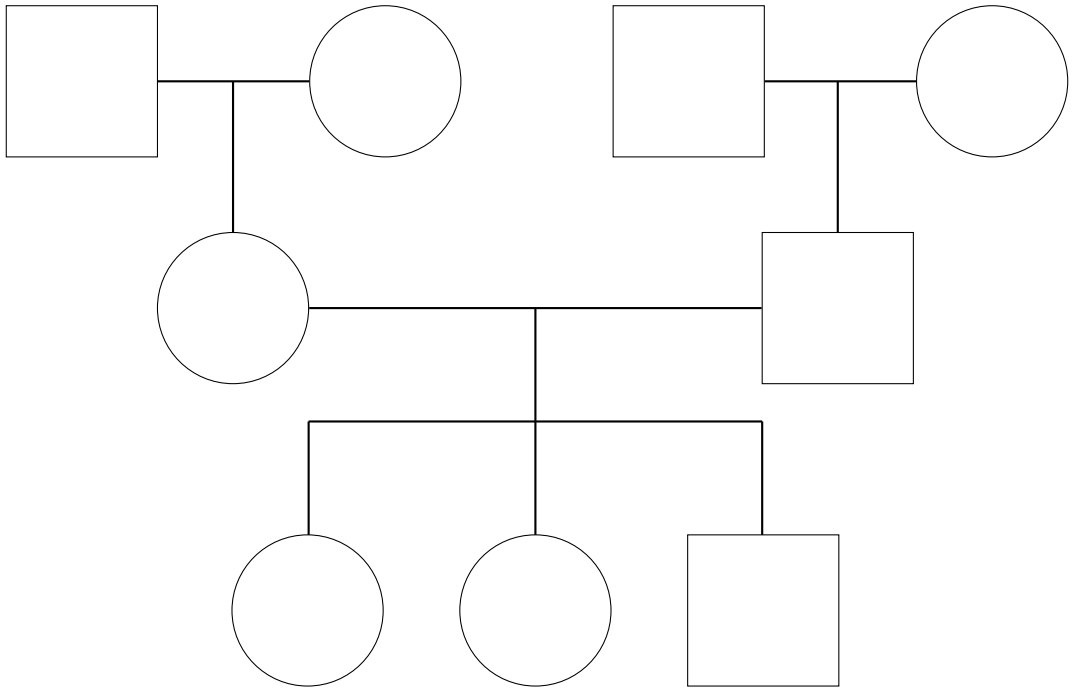
1. **The law of segregation**: An individual will pass down one of their two alleles to each offspring.

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

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

Pedigrees

Pedigrees provide a graphical representation of relationships.



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

Identity By Descent

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

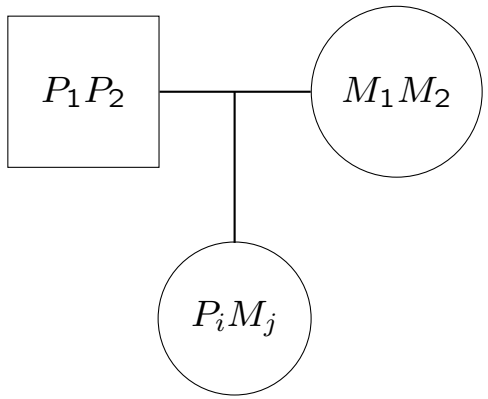
- IBD alleles are copies of the same allelic type inherited through a common ancestor (and ignores mutation).

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

IBD for Parent-Child Relationships

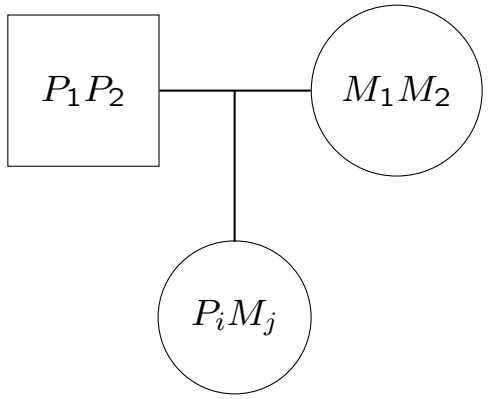
- Mendel's law states that one of the two alleles from a parent will be passed down to a child;

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



IBD for Parent-Child Relationships

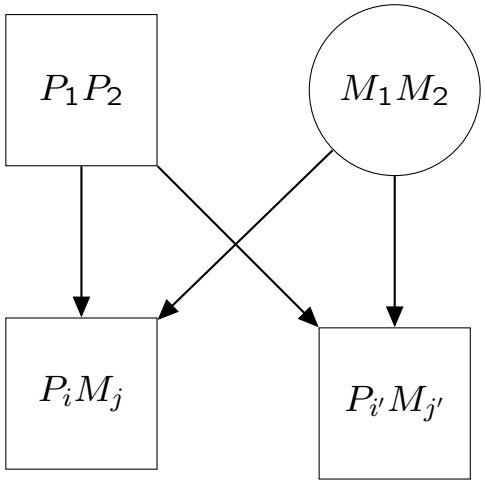
The child will always have exactly 1 allele that is IBD to an allele from a specific parent (the other allele will be IBD to an allele from the other parent).



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

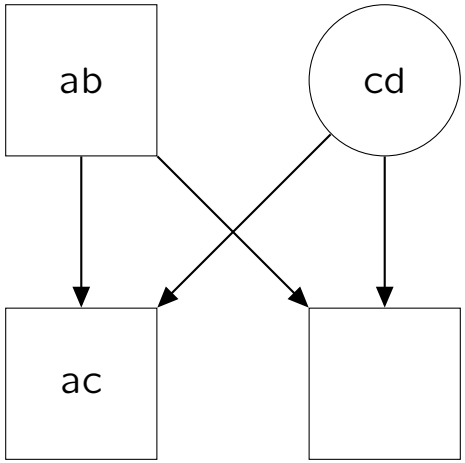
IBD for Sibling Relationships

What about siblings?



IBD for Sibling Relationships

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

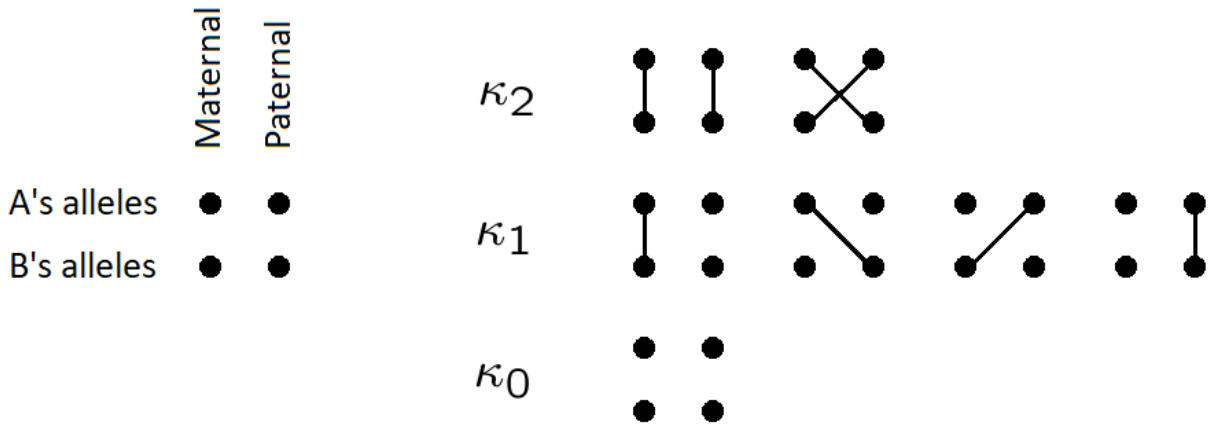


		Alleles IBD		
Sib 1	**Sib 2**	0	1	2
ac	ac	✓		
	bc		✓	
	ad		✓	
	bd			✓
Total		1/4	1/2	1/4

IBD Coefficients

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

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



What IBD classes are relevant for unrelated individuals?

IBD Coefficients

For parent-child relationships we saw that:

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

while for siblings we have:

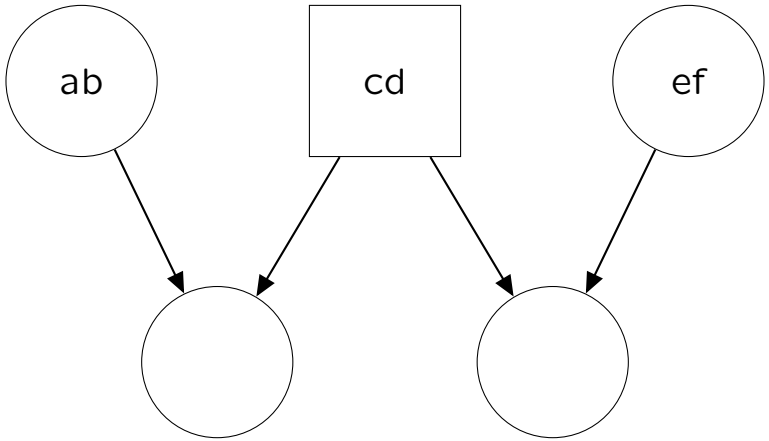
$$\kappa_0 = \Pr(\overline{\text{IBD}_M}) \times \Pr(\overline{\text{IBD}_P}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

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

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

IBD Coefficients for Half-sibs

What are the IBD coefficients for half-sibs?



IBD Coefficients

The following table shows IBD probabilities for common relationships:

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

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

Match Probabilities for Relatives

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

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

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

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

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

Match Probabilities for Relatives

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

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

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

Match Probabilities for Relatives - Exercise

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

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

Match Probabilities for Relatives - Exercise

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

- LR $= \frac{\Pr(AA|AA,H_p)}{\Pr(AA|AA,H_d)} = \frac{1}{p_A^2} = 625$;
- LR $= \frac{1}{\kappa_0 M_2 + \kappa_1 M_1 + \kappa_2} = \frac{1}{0.5p_A^2 + 0.5p_A} \approx 48$;
- LR $= \frac{1}{0.25p_A^2 + 0.5p_A + 0.25} \approx 3.7$;
- LR $= 1$.

LRs for Relatives

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

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

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

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

LRs Including Relatives

- We can model a situation where relatives of the suspect make up a small proportion of the total population.
- It is however not trivial to set the number of siblings, uncles/aunts, cousins, etc.
- An overall LR can be calculated by modeling these priors as simple population proportions.
- This requires specifying an average number of children (e.g. using fertility rates) and population size.

The Island Problem

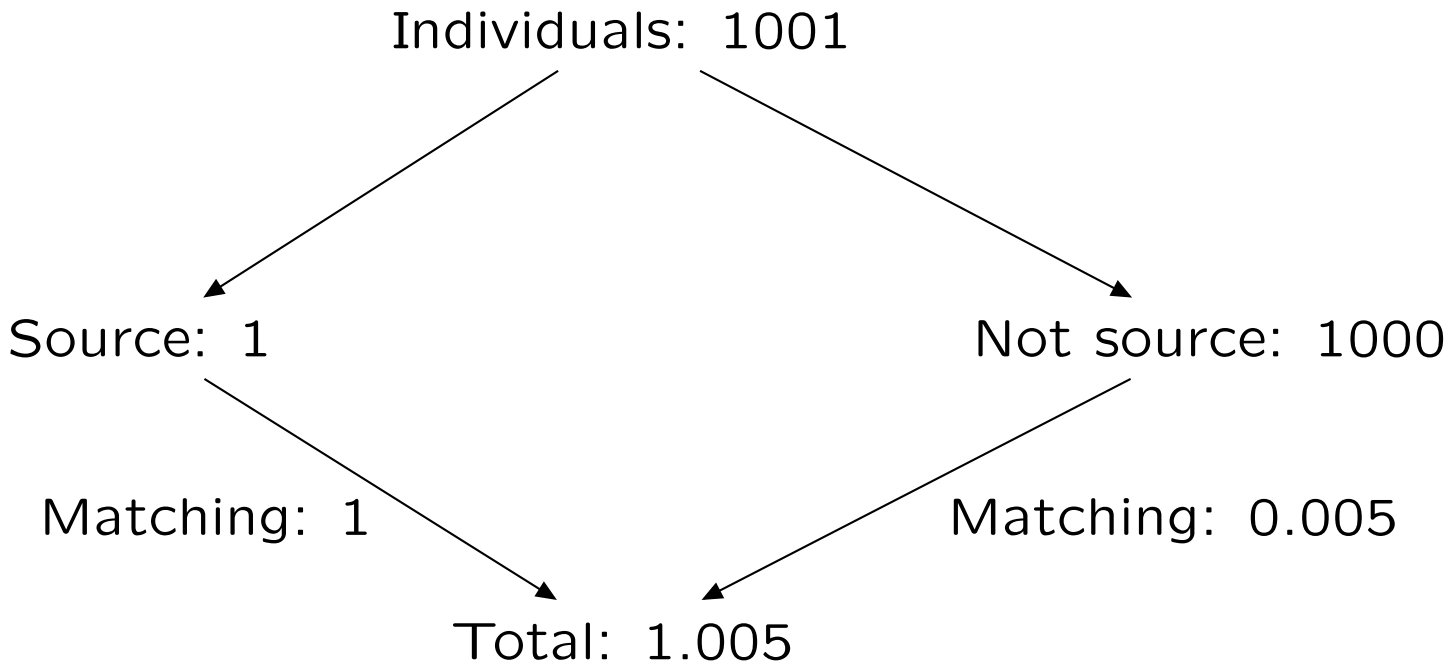
Suppose there is a crime committed on a remote island with a population of size 1001. A suspect Q is found to match the crime scene profile. What is the probability that Q is the source of the profile, assuming that:

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

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

The Island Problem - Solution

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



The Island Problem - Relatives

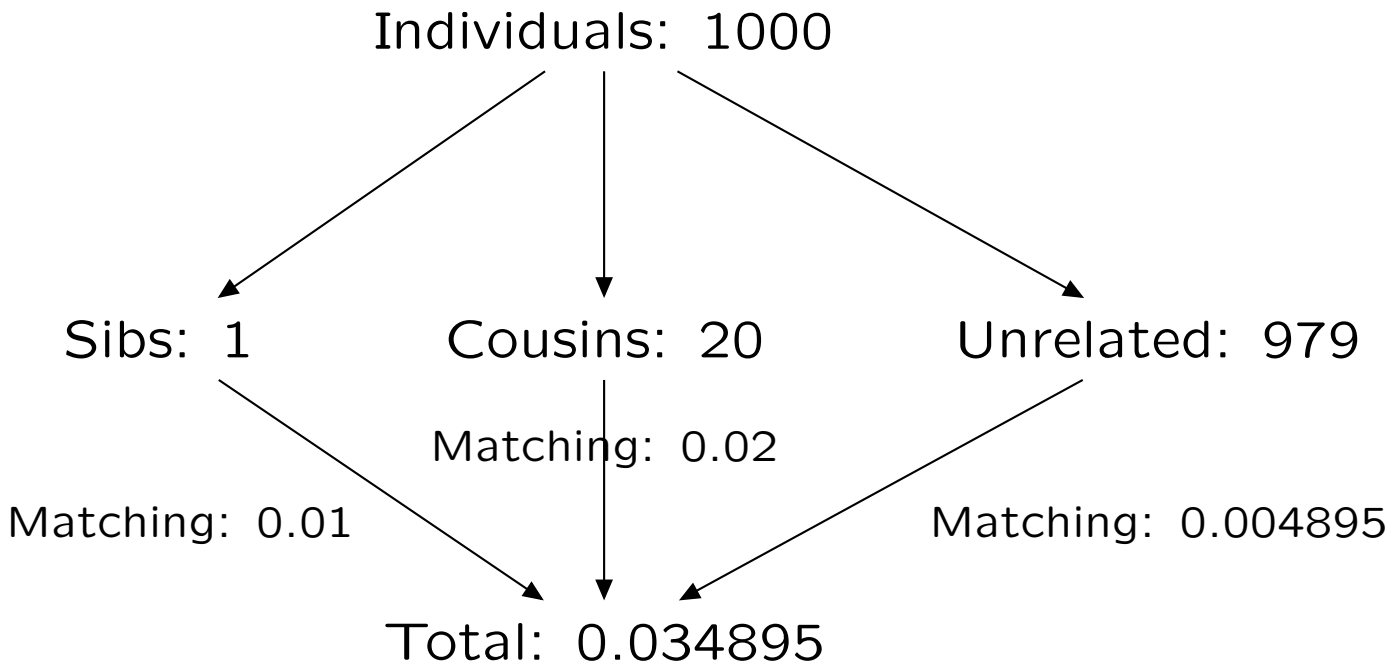
Now suppose that Q has one sibling and 20 cousins on the island, and no other relatives. What is now the probability that Q is the source, using match probabilities of:

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

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

The Island Problem - Solution for Relatives

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



The Island Problem - Solution for Relatives

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

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

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

What if we were not given any information about the relatives of Q?

The Island Problem - Relatives

What if we were not given any information about the relatives of Q?

In this case, background information may be used to assess plausible values for the priors, specifying the numbers of relatives in each category.

LRs can be calculated for each plausible set of values, and the resulting weight-of-evidence may be averaged over the sets.

In practice, it is often satisfactory to consider only an upper bound on the plausible number of relatives in each category.

Other Applications

The concept of relatedness is important for, and benefits, other applications as well:

- Paternity testing
- Missing persons
- Familial searching
- Inference of ethnicity
- Inference of phenotype

Paternity Testing

Paternity and familial identification can provide evidence in criminal context and during civil litigation. For a paternity case, the two propositions could be:

H_p: The alleged father (AF) is the true father.
H_d: Some other (unrelated) man is the father.

The likelihood ratio is in this case often referred to as the paternity index (PI).

Paternity Testing - Exercise

Suppose a child has genotype $G_C = AB$. What are the LR values when:

- $G_M = AA$ and $G_{AF} = BB$;
- $G_M = AA$ and $G_{AF} = CD$;
- $G_M = AA$ and $G_{AF} = BC$;
- $G_M = AB$ and $G_{AF} = AA$.

Paternity Testing - Exercise

Suppose a child has genotype $G_C = AB$. The LR values are:

- $\text{LR} = \frac{\Pr(G_C{=}AB|G_M{=}AA,G_{AF}{=}BB,H_p)}{\Pr(G_C{=}AB|G_M{=}AA,H_d)} = \frac{1}{p_B}$;
- $\text{LR} = \frac{\Pr(G_C{=}AB|G_M{=}AA,G_{AF}{=}CD,H_p)}{\Pr(G_C{=}AB|G_M{=}AA,H_d)} = 0$;
- $\text{LR} = \frac{\Pr(G_C{=}AB|G_M{=}AA,G_{AF}{=}BC,H_p)}{\Pr(G_C{=}AB|G_M{=}AA,H_d)} = \frac{\frac{1}{2}}{p_B} = \frac{1}{2p_B}$;
- $\text{LR} = \frac{\Pr(G_C{=}AB|G_M{=}AB,G_{AF}{=}AA,H_p)}{\Pr(G_C{=}AB|G_M{=}AA,H_d)} = \frac{\frac{1}{2}}{\frac{1}{2}p_A+\frac{1}{2}p_B} = \frac{1}{p_A+p_B}$.

Paternity Testing - Exercise

Calculate the weight of the evidence for the following data:

Locus	G_C	G_M	G_{AF}
TPOX	(6,9)	(6,12)	(8,9)
vWA	(17,17)	(17,16)	(17,17)
TH01	(7,9)	(9,10)	(7,9)

Locus	**Allele**	**Frequency**
TPOX	6	0.006
	8	0.506
	9	0.094
	12	0.038
vWA	16	0.276
	17	0.300
TH01	7	0.147
	9	0.232
	10	0.116

Source: Introduction to Statistics for Forensic Scientist (Lucy, 2005).

Paternity Testing - Exercise

Calculate the weight of the evidence for the following data:

Locus	G_C	G_M	G_{AF}
TPOX	(6,9)	(6,12)	(8,9)
vWA	(17,17)	(17,16)	(17,17)
TH01	(7,9)	(9,10)	(7,9)

We calculate single-locus LRs and combine these results through multiplication:

- TPOX: LR $= \frac{0.25}{0.5p_9} = \frac{1}{2\times 0.094} = 5.32$;
- vWA: LR $= \frac{1}{p_{17}} = \frac{1}{0.3} = 3.33$;
- TH01: LR $= \frac{0.25}{0.5p_7} = \frac{1}{2\times 0.147} = 3.40$.

Our overall LR is in this case 60.23, yielding evidence in favor of H_p.

Paternity Testing

These cases can be extended to allow for more complex situations:

- Unavailability of the mother;
- Relatedness between the mother and alleged father;
- A relative of the alleged father is the true father;
- Incorporating profiles of (alleged) relatives (e.g. for half-sibs or when alleged father is unavailable);
- Multiple children;
- Incorporating mutations, substructure, silent alleles, non-autosomal DNA, etc.

Missing Persons

The discussed methods for evidence evaluation are also applicable to other situations, such as disaster victim identification and immigration cases.

A comparison must in these cases be carried out between a profile obtained from unidentified remains, or an applicant, and a missing person's profile.

It is, however, often the case that a sample from the missing person is not available, in which case it might be possible to make use of surrogate samples (e.g. obtained through a medical institution).

Alternatively, relatives can be used for testing purposes.

Missing Persons

For a missing person case, the two propositions could be:

H_p: The sample is from the missing person.

H_d: The sample is from some unknown person.

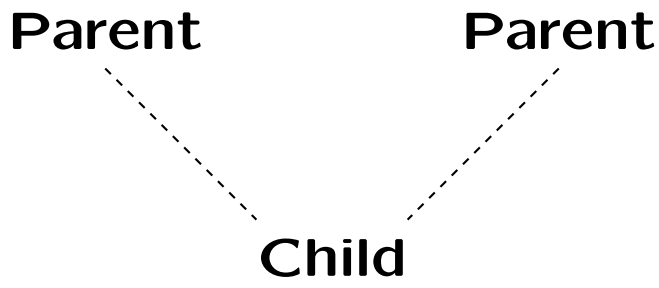
The following likelihood ratios are obtained for a sample with alleged mother (AM) and alleged father (AF), compared to the paternity index, for $p_A = p_B = 0.1$:

(A)M	AF	Sample	LR	Value	PI	Value
AA	BB	AB	$\frac{1}{2p_Ap_B}$	50	$\frac{1}{p_B}$	10
AA	BC	AB	$\frac{1}{4p_Ap_B}$	25	$\frac{1}{2p_B}$	5
AB	AA	AB	$\frac{1}{4p_Ap_B}$	25	$\frac{1}{p_A+p_B}$	5

Source: Interpreting DNA Evidence (Evett & Weir, 1998).

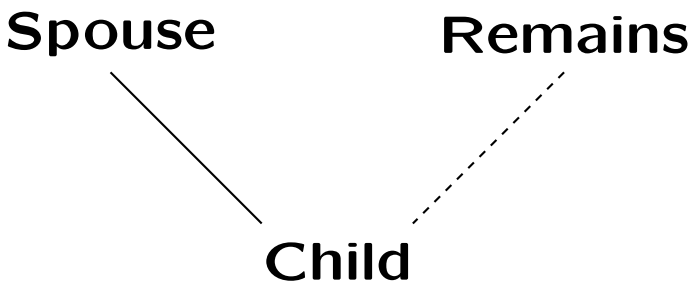
Missing Persons

In the previous case the genetic evidence E consists of the genotype from a sample that has come from some person X who may be the missing person, together with the genotypes from the parents of the missing person.



If, instead, the genotypes of the spouse S and child C of the missing person are available, the situation is similar to evidence evaluation in case of paternity testing.

Missing Persons



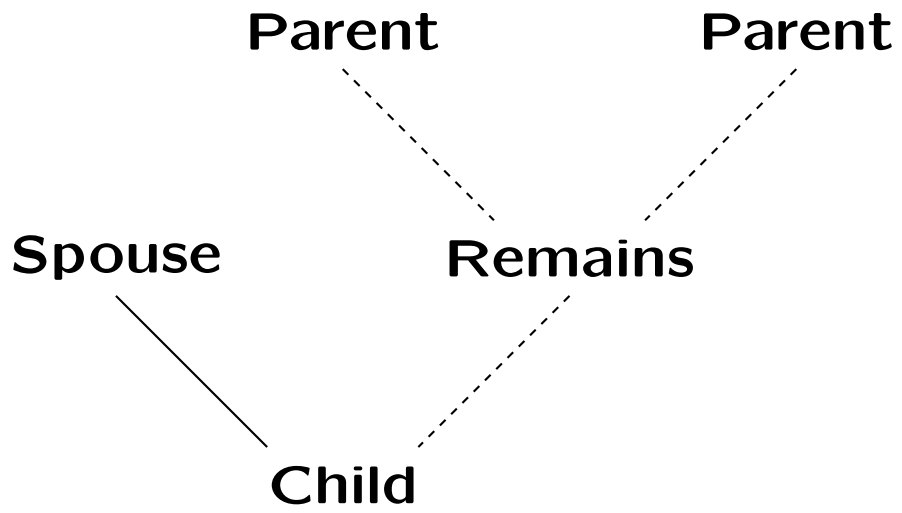
The likelihood ratios are the same as in the paternity case where X is the alleged father of child C who has mother S:

$$\begin{aligned}
\mathsf{LR} &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\
&= \frac{\Pr(G_C, G_S, G_X|H_p)}{\Pr(G_C, G_S, G_X|H_d)} \\
&= \frac{\Pr(G_C|G_S, G_X, H_p)\Pr(G_S, G_X|H_p)}{\Pr(G_C|G_S, G_X, H_d)\Pr(G_S, G_X|H_d)} \\
&= \frac{\Pr(G_C|G_S, G_X, H_p)}{\Pr(G_C|G_S, H_d)}
\end{aligned}$$

Missing Persons

It may be the case that people apart from the spouse and child of the missing person are typed. The general procedure is the same: the probabilities of the set of observed genotypes under two explanations are compared.

Suppose the parents P and Q as well as the child C and spouse S of the missing person are typed, and that a sample is available that has come from some person X thought under H_p to be the missing person.



Missing Persons

Under explanation H_d, the sample from X did not come from the missing person, and therefore the genotype of X does not depend on the genotypes of P and Q and the genotype of C does not depend on the genotype of X.

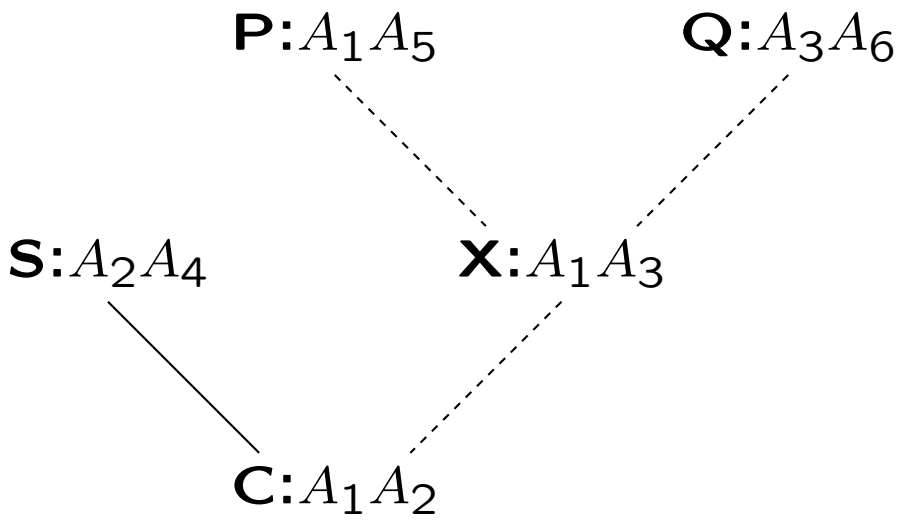
The likelihood ratio is arranged to involve probabilities of genotypes conditional on previous generations. If both parents of an individual have been typed, there is no need to condition on the grandparents of that individual.

In the following slides, C, S, X, P and Q represent the genotypes of the child, the remains, the spouse and the parents of the missing person.

Missing Persons

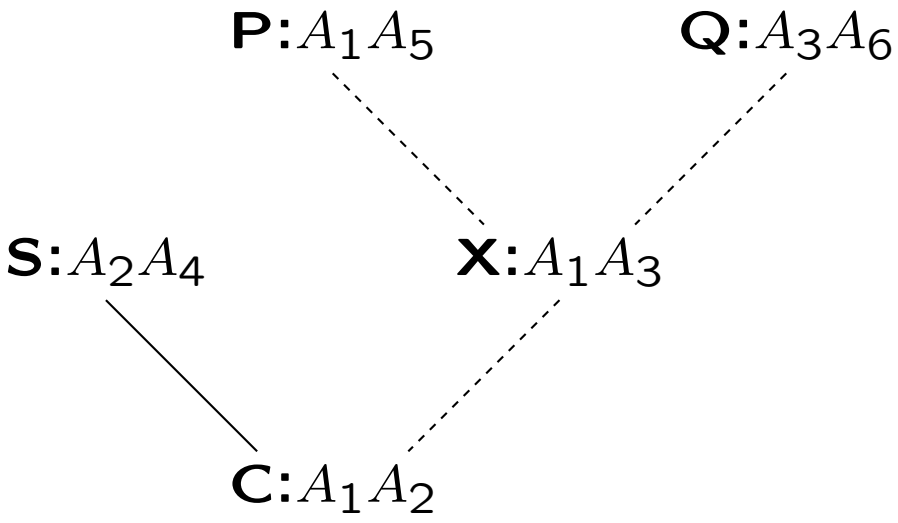
$$
\begin{aligned}
\mathsf{LR} &= \frac{\mathsf{Pr}(E|H_p)}{\mathsf{Pr}(E|H_d)} \\
&= \frac{\mathsf{Pr}(C,S,X,P,Q|H_p)}{\mathsf{Pr}(C,S,P,X,Q|H_d)} \\
&= \frac{\mathsf{Pr}(C|S,X,P,Q,H_p)\,\mathsf{Pr}(S,X,P,Q|H_p)}{\mathsf{Pr}(C|S,X,P,Q,H_d)\,\mathsf{Pr}(S,X,P,Q|H_d)} \\
&= \frac{\mathsf{Pr}(C|S,X,H_p)\,\mathsf{Pr}(S,X|P,Q,H_p)\,\mathsf{Pr}(P,Q|H_p)}{\mathsf{Pr}(C|S,P,Q,H_d)\,\mathsf{Pr}(S,X|P,Q,H_d)\,\mathsf{Pr}(P,Q|H_p)} \\
&= \frac{\mathsf{Pr}(C|S,X,H_p)\,\mathsf{Pr}(S|H_p)\,\mathsf{Pr}(X|P,Q,H_p)}{\mathsf{Pr}(C|S,P,Q,H_d)\,\mathsf{Pr}(S|H_d)\,\mathsf{Pr}(X|H_d)} \\
&= \frac{\mathsf{Pr}(C|S,X,H_p)\,\mathsf{Pr}(X|P,Q,H_p)}{\mathsf{Pr}(C|S,P,Q,H_d)\,\mathsf{Pr}(X|H_d)}
\end{aligned}
$$

Missing Persons - Example



$$\mathsf{LR} = \frac{\Pr(C|S,X,H_p)\Pr(X|P,Q,H_p)}{\Pr(C|S,P,Q,H_d)\Pr(X|H_d)}$$

Missing Persons - Example



$$
\begin{aligned}
\Pr(C|S,X,H_p) &= 1/4 \\
\Pr(X|P,Q,H_p) &= 1/4 \\
\Pr(C|S,P,Q,H_d) &= 1/8 \\
\Pr(X|H_d) &= 2p_1p_3 \\
\mathrm{LR} &= \frac{1}{4p_1p_3}
\end{aligned}
$$

Familial Searching

- A database may be used to compare crime scene profiles to known offenders when investigators lack a suspect.
- A high stringency search requires a full match of the DNA profiles, and might not always return a hit.
- Lowering the search stringency level may lead to a partial match, and has the potential to identify close relatives.
- *Familial searching* refers to the process where investigators look for close relatives in the DNA database in order to open up new investigative leads.

Familial Searching - Case Example

A serial killer nicknamed the Grim Sleeper (due to a 14-year break) was responsible for the death of at least 10 young women in Los Angeles between 1985 and 2007.

When traditional forensic methods failed, investigators turned to novel partial-match DNA search methods authorized in 2008, eventually leading to a positive result for a recently convicted young man. Together with other evidence this led to the suspicion of the father.

The L.A. police was notified by investigators and got a DNA sample from a discarded piece of pizza. Lonnie Franklin was found to match, leading to an arrest in July 2010 and eventual conviction in May 2016.

Familial Searching - Strategies

A certain strategy is required to select a potential relative of the unknown donor from the database. Two general methods are available, both resulting in a ranked list of candidates to investigate further:

- **IBS method**: simply counts the number of shared alleles between two DNA profiles.

- **LR method**: likelihood under two competing hypothesis (als in this context also called a kinship index (KI):

$$\mathrm{KI} = \frac{\sum_{i=0,1,2} \Pr(G_C, G_R|\mathrm{IBD} = i) \Pr(\mathrm{IBD} = i|\mathrm{relationship})}{\sum_{i=0,1,2} \Pr(G_C, G_R|\mathrm{IBD} = i) \Pr(\mathrm{IBD} = i|\mathrm{unrelated})}$$

Familial Searching - Performance

Familial searching is typically focused on parent-child and sibling relationships, as more distant relatives are usually harder to identify and differentiate from unrelated individuals.

The following table shows the performance of the methods using simulated 10-locus profiles in the New Zealand database:

Method	Rank 1 (%)	Rank $1-100$ (%)
IBS: Siblings	24	72
IBS: Parent-child	8	68
LR: Siblings	31	78
LR: Parent-child	25	99

Source: Effectiveness of familial searches (Curran & Buckleton, 2008).

Familial Searching - Effectiveness

- LR methods outperform the IBS method.
- It is slightly easier to locate parent-child relationships, although siblings more often obtain a number one ranking.
- More loci improve the effectiveness of familial searching, especially in case of extra highly polymorphic loci.
- Ranked lists can be refined based on lineage markers.
- The methods can be extended by using a combination of IBS and LR, setting thresholds, or using a weighted approach.

It is important to note that the effectiveness depends on the assumption that a true close relative of the donor is actually present in the database.

Familial Searching - Considerations

Familial searching has proven to be a successful tool in several cases, but it also raises privacy and legal policy concerns:

- Disproportional attention to members of populations that are over-represented in the database.
- False positives may lead to the investigation of innocent people.
- Might reveal the presence of a family member in the database.
- Might reveal the presence of a previously unknown genetic link.
- Might reveal the absence of a genetic link.
- Crimes might go unreported (in case of searches against victim profiles).

Consumer Genomics Tools

With the emergence of consumer genomics tools, familial searching has become far more powerful. The limited set of STR markers does not allow for finding relatives beyond first and second degree relationships. Furthermore, policies largely restrict or even prohibit the practice completely.

These limitations, however, do not explicitly restrict the use of crime scene samples with civilian DNA databases.

Service	Database size	DTC provider	Relative finder	3rd party support
23andMe	5M	●	●	
Ancestry	9M	●	●	
DNA.Land	100K		●	●
FTDNA	1M	●	●	●
GEDmatch	1M		●	●
LivingDNA	n/a	●		
MyHeritage	1.4M	●	●	●

Source: Re-identification of genomic data using long range familial searches (Erlich et al., 2018).

Genealogical Searching - Case Example

The Golden State Killer is a serial killer, rapist, and burglar who committed at least 13 murders, more than 50 rapes, and over 100 burglaries in California from 1974 to 1986. He is believed to be responsible for three crime sprees, each of which spawned a different nickname (the Visalia Ransacker, the East Area Rapist, and the Original Night Stalker) before it became evident that they were committed by the same person.

Crime scene evidence was used to obtain a profile that mimicked the format of regular direct-to-consumer (DTC) providers in order to upload it to GEDmatch.

Genealogical Searching - Case Example

A search identified 10 to 20 distant relatives of the perpetrator, which eventually led to the arrest of Joseph James DeAngelo. It took five genealogists four months to trace back the identity of the suspected perpetrator.

DeAngelo, a former police officer, was arrested in April 2018, after a DNA sample collected from the door handle of his car confirmed a match. In April 2019, the case prosecutors announced that they would seek the death penalty.

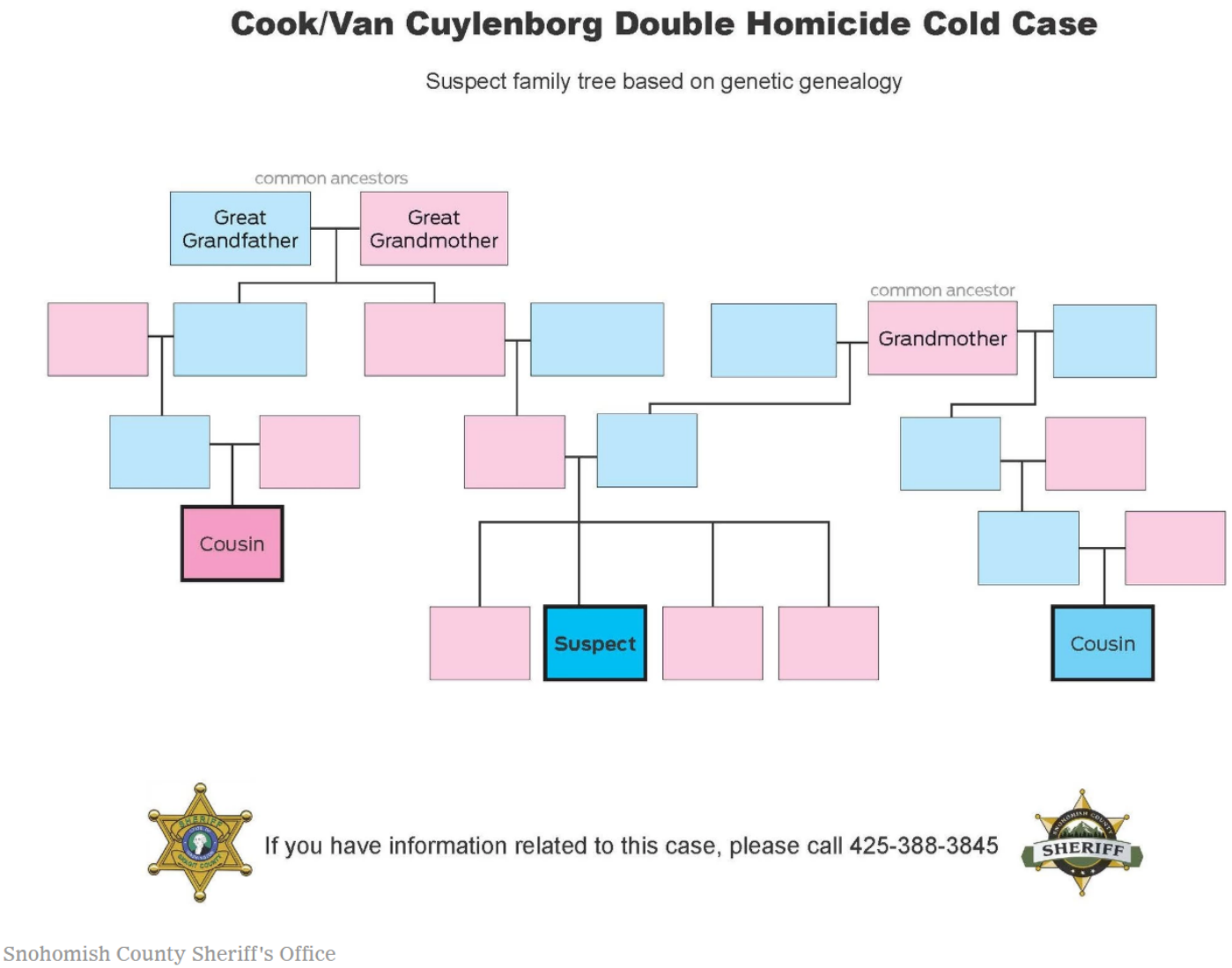
Genealogical Searching - Case Example

Even more recently, the Snohomish County sheriff's office announced that they arrested a suspect in the killing of a young couple while they were vacationing in Washington State in 1987.

A GEDmatch search led to two second cousins, which could be tied together through a marriage of two descendants from their great-grandparents. The only son from this marriage was investigated further and found to match the crime scene evidence, leading to his arrest in May 2018.

William Earl Talbott II is the first person convicted of murder as a result of using genetic genealogy searches. He was found guilty in June 2019 and sentencing is scheduled for July 24 in Snohomish County Superior Court.

Genealogical Searching - Case Example



Source: Technique Used to Find Golden State Killer Leads to a Suspect in 1987 Murders (Murphy, 2018).

Genealogical Searching - Case Example

The victim's mother insisted on genealogical searching, not because she believed the convicted man was innocent, but to find the others who were believed to be involved.

In an Apparent First, Genetic Genealogy Aids a Wrongful Conviction Case

An Idaho man who falsely confessed to a 1996 rape and murder is expected to have his name cleared on Wednesday.

By MIA ARMSTRONG

UPDATE 04:45 P.M. 07.17.2019

The wrongful conviction was mainly based on a (false) confession and outweighed a DNA mismatch and the absence of other physical evidence.

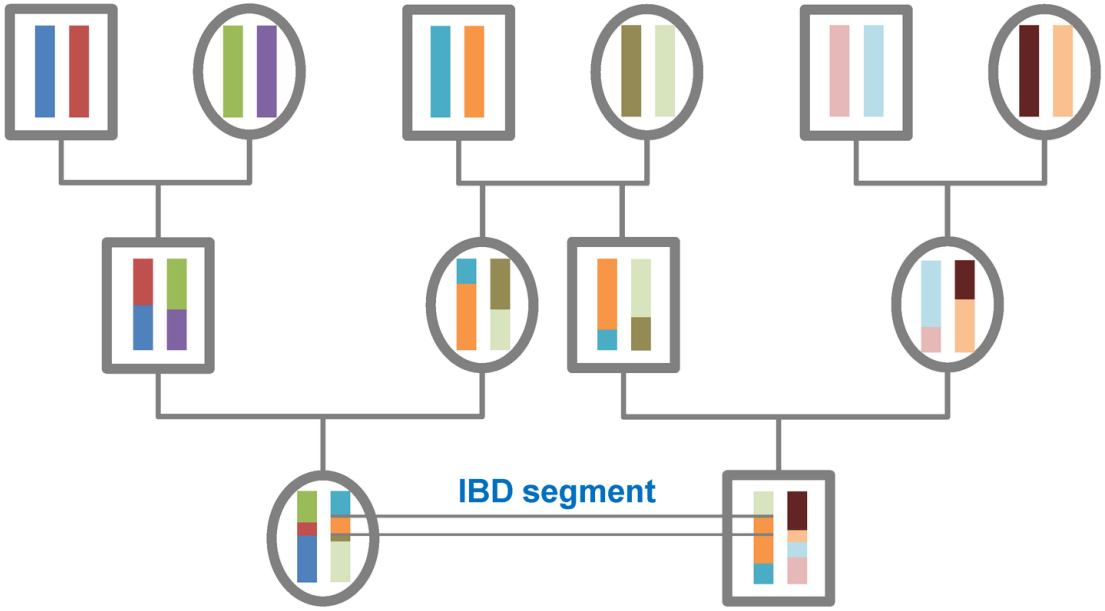
Genealogical Searching - Strategy

The limited set of STR markers does not allow for finding relatives beyond first and second degree relationships. SNP panels, with up to a million SNPs allow distinguishing even distant cousins.

A different statistical measure is used, that takes (lack of) recombination into account. Genetic linkage is measured in centimorgans (cM), which can be used to determine the level of relatedness between SNP profiles.

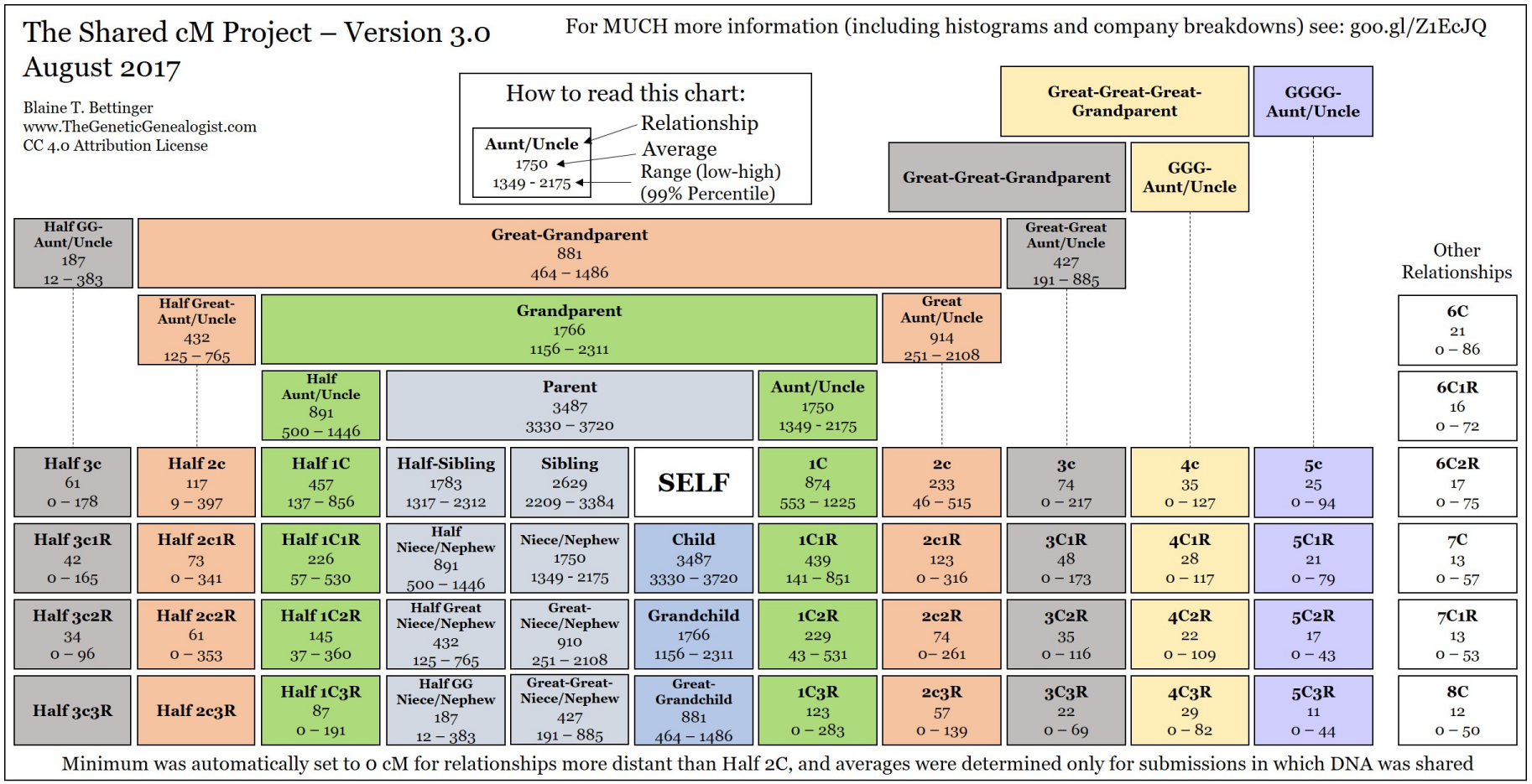
Genealogical Searching - Strategy

One Morgan is the length along a chromosome in which 1 recombination event is expected to occur. IBD segments occur when people share matching DNA segments that have been inherited from a common ancestor without any intervening recombination.



Source: `https://isogg.org/wiki/Identical_by_descent`

Genealogical Searching - Strategy



Genealogical Searching - Strategy

Suppose a GEDmatch search for an evidence profile E reveals two first cousins $C1$, $C2$.

E and $C1$ have two of their four grandparents in common. Think of the four grandparents of $C1$ and trace their descendants $D1$: there are the parents, uncles, aunts and cousins of $C1$.

E and $C2$ have two of their four grandparents in common. Think of the four grandparents of $C2$ and trace their descendants $D2$: there are the parents, uncles, aunts and cousins of $C2$.

The source of E belongs to both $D1$ and $D2$.

From STR to SNP Profile

Instead of looking for a (partial) match in one database, it is also possible to combine different databases, even with no overlapping genetic markers. Provided that sufficiently strong LD exists, SNP and STR profiles can be associated with the same individual or distinct but closely related individuals.

Software can be used to infer STR genotypes from a SNP dataset, making it possible to compute match scores for pairs of individuals between databases. This means that CODIS profiles can possibly be connected to a SNP profile, collected for e.g. biomedical or genealogical research, and this cross-database record matching extends to relatives.

Linkage disequilibrium connects genetic records of relatives typed with disjoint genomic marker sets (Rosenberg et al., 2018).

Genealogical Searching - Implications

About 60 percent of people of European descent who search genetic genealogy databases will find a match with a relative who is a third cousin or closer.

Genealogy databases could reveal the identity of most Americans

Keeping your DNA private is getting harder

BY TINA HESMAN SAEY 4:12PM, OCTOBER 12, 2018

It took the team a day to trace back anonymous data from the publicly available 1000 Genomes database to the right person.

Identity inference of genomic data using long-range familial searches (Erlich et al., 2018).

Genealogical Searching - Implications

In May 2019, in response to privacy concerns, GEDmatch updated their policy.

Users must now explicitly opt in to allow their profiles to be used in law enforcement investigations.

At the same time, the service authorized law enforcement to upload DNA data to identify a perpetrator of a violent crime against another individual. The policy defined a violent crime to include only homicide and sexual assault.

Inference of Ancestry and Ethnicity

Suppose that a population can be classified into K groups. The probability of a DNA sample with profile D coming from group k, can be written as:

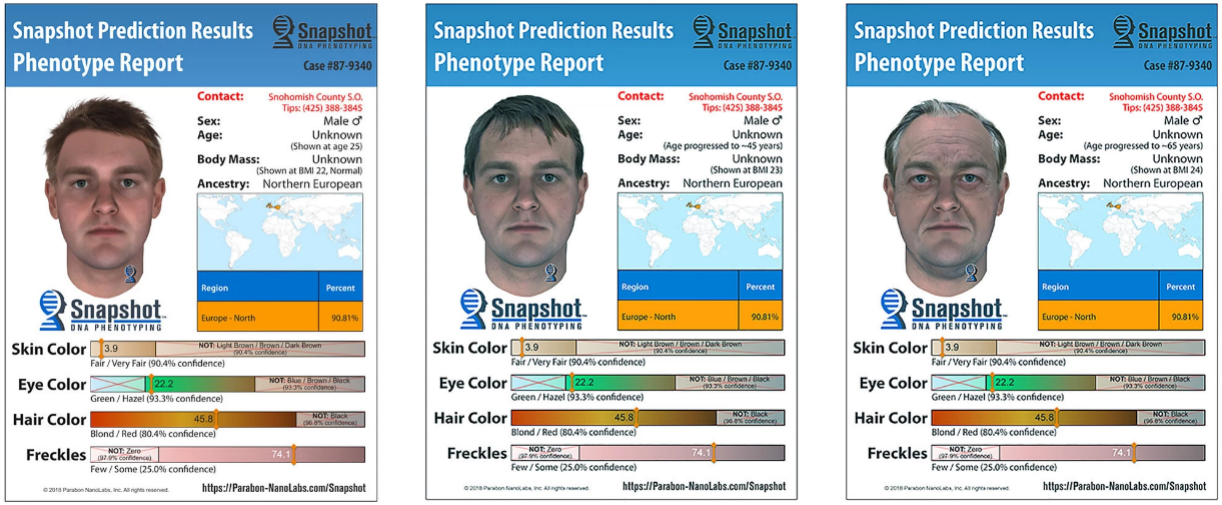
$$\Pr(\text{group } k|D) = \frac{\Pr(D|\text{group } k)\Pr(\text{group } k)}{\sum_{j=1}^{K}\Pr(D|\text{group } j)\Pr(\text{group } j)}.$$

STR profiles can give some information, although they provide limited discriminatory power in this context. Instead, SNP sets (so-called ancestry informative markers) have been demonstrated to be useful for distinguishing individuals from certain (sub-)populations.

Inference of Phenotype

SNPs may be linked to some visual phenotypes, including hair color and eye color. Other facial characteristics can now also be predicted from genotypes with some accuracy.

These SNP associations can potentially be used in forensic settings, e.g. in combination with a description of an eyewitness of a target individual.



Picture rendered by Parabon Nanolabs.

Source: Technique Used to Find Golden State Killer Leads to a Suspect in 1987 Murders (Murphy, 2018).