Meeting 17
Forensic applications, part I

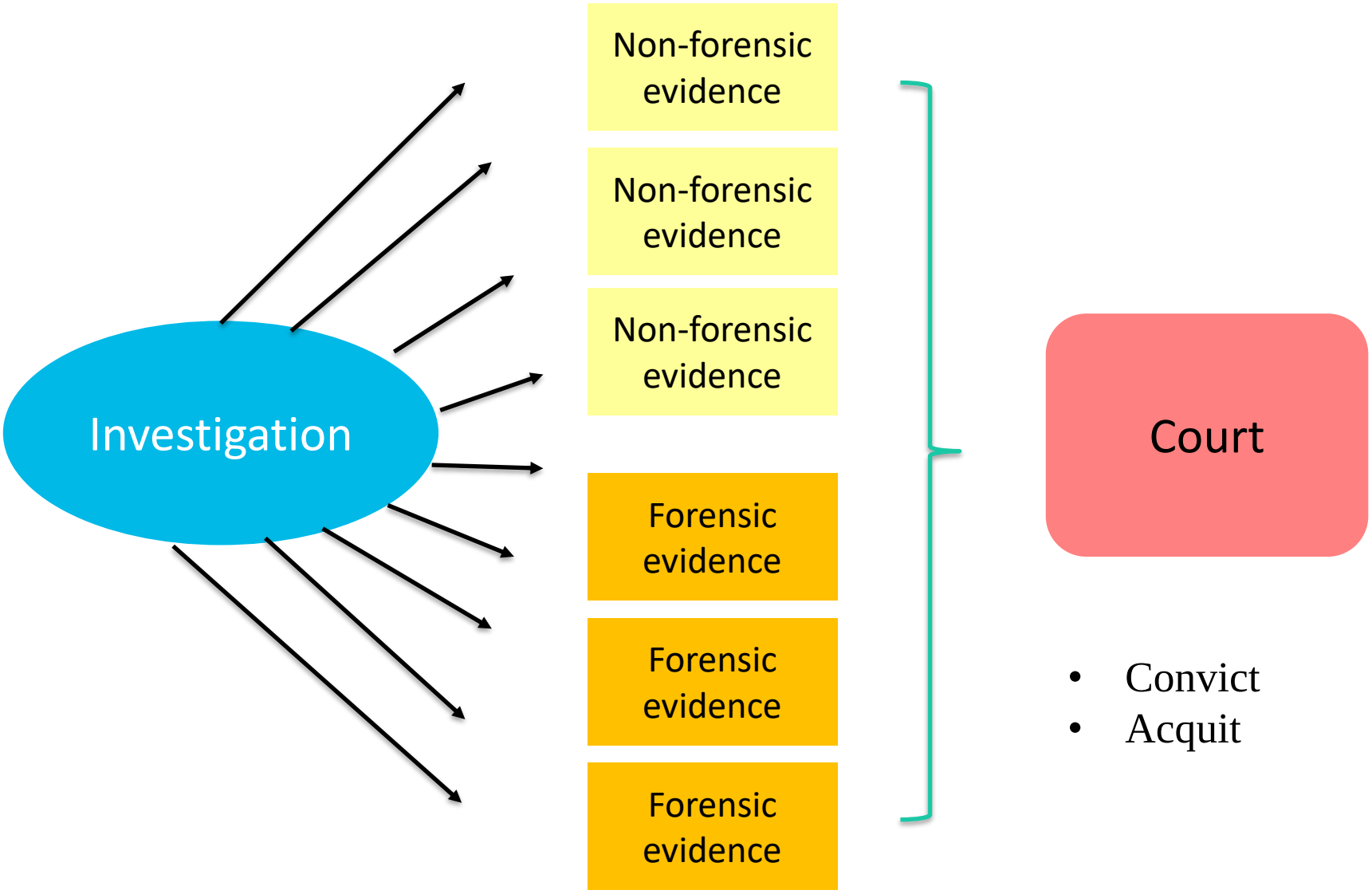
What is forensic science?

"Scientific investigations and their outcome for use in law enforcement and legal disputes."

- assists in the preliminary investigation of a crime
- constitutes part of the evidence in court
  - ✓ criminal law
  - ✓ civil law
- assists in sorting out (disputed) kinships between individuals
- …

- criminalistics
- forensic chemistry
- IT-forensics
- forensic genetics
- forensic pathology and entomology
- forensic toxicology
- forensic psychiatry
- …

Where is the decision problem?



The ultimate hypotheses in crime cases

$\boldsymbol{H}$: The defendant is guilty as charged
$\neg\boldsymbol{H}$: The defendant is not guilty as charged

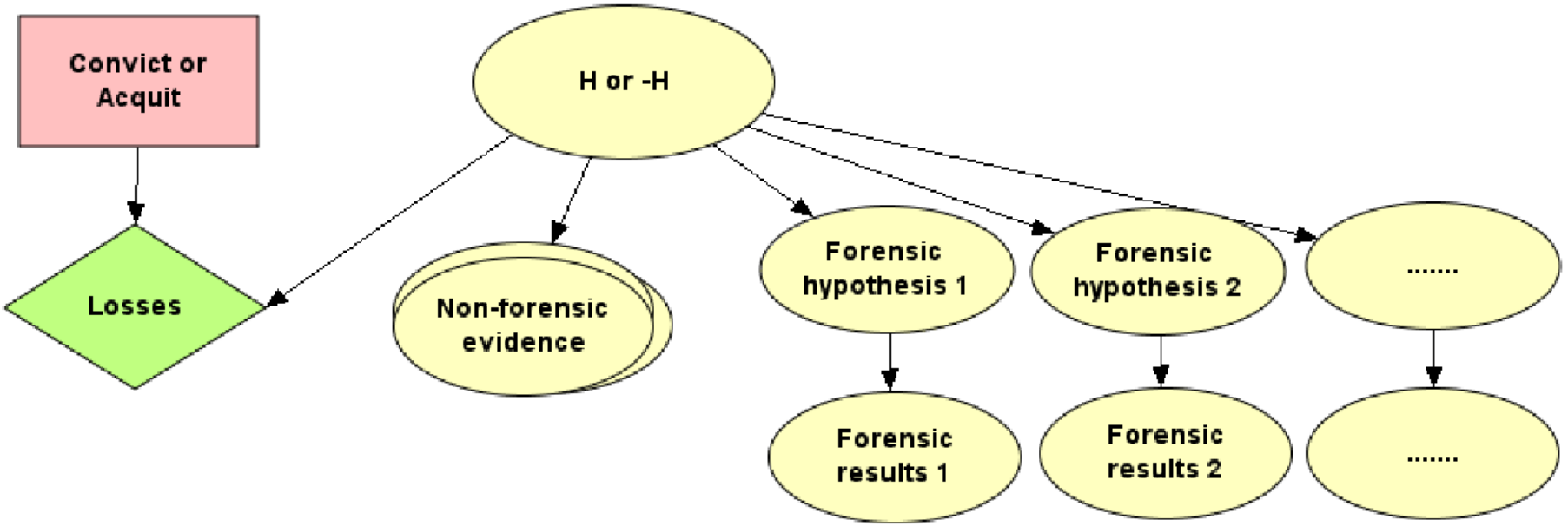
Since it is practically impossible to learn whether the defendant is guilty or not, the hypotheses can be technically reformulated as

$\boldsymbol{H}$: The prosecutor has proven their case against the defendant
$\neg\boldsymbol{H}$: The prosecutor has not proven their case against the defendant

The court's decision problem:

	H is true	$\neg H$ is true
Convict	0	L_{II}
Acquit	L_{I}	0

How the evidence may come in following a Bayesian model for evidence interpretation and evaluation.



Examples of forensic questions in crime cases

- Was the recovered saliva stain left by the suspect?
- Was the recovered shoe mark left by the seized shoe?
- Does the white powder contain a narcotic substance?
- How did the perpetrator make entrance to the premises?
- Was the suspect involved with the shooting incident?
- Are there traces of ignitable liquids in the fire debris (suspected arson)?
- Did the oil spill come from the vessel?
- Was the suspect’s pullover in contact with the car seat?
- Has the laptop been used to distribute child pornography?

Was the recovered saliva stain left by the suspect?

Main forensic hypothesis (forwarded by the prosecutor)**:**

$\mathcal{H}_m$**:** The recovered saliva stain was left by the suspect.

Main hypothesis *reformulated* for forensic investigation purposes:

H_m**:** The suspect is the source of the DNA extracted from the saliva stain.

Alternative hypothesis:

H_a**:** Someone else than the suspect is the source of the DNA extracted from the saliva stain.

DNA-analysis

The human genome consists of 23 chromosome pairs.

Each chromosome is a double helix consisting of nucleobase pairs *Guanine-Cytosine* (**G-C)** and *Adenine-Thymine* (**A-T**)

At conception, one chromosome in a pair is inherited from the mother and the other from the father, but at the meiosis phase upon conception so-called recombinations (random) between the chromosomes result in new chromosomes for that individual.

Hence, it is not possible to know which parts of a chromosome come from the mother, and which come from the father.

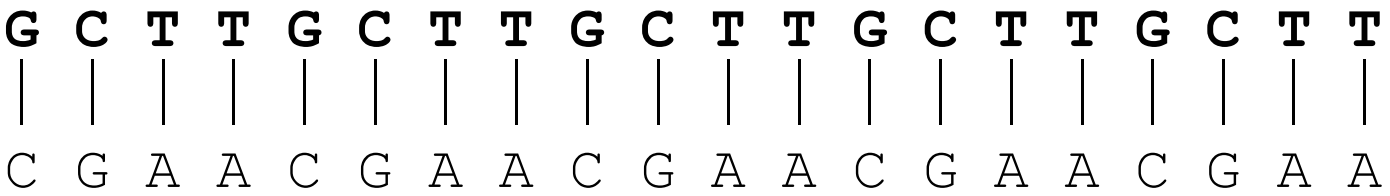
Most of the genome (the DNA) of humans (more than 90 %) has no coding function (evolutionary rest).

The non-coding part of the genome shows high degrees of polymorphism (due to that it has no effect on mating preferences).

$\Rightarrow$

- The whole genome of a human being can be considered as unique (with exception from identical twins, triplets etc. with no mutations)
- Parts of the genome constitute genetic fingerprints

A genetic *autosomal marker* (or *locus*) is a specified section of a chromosome pair where sequences (of lengths from 1 to 100s) of nucleobases are observed.



Dominating types of markers used today:

- Short Tandem Repeat (STR) markers (*Jeffreys, 1990*)
- Single Nucleotide Polymorphisms (SNP)

observed using Polymerase Chain Reaction (PCR) technique combined with Capillary Electrophoresis

In a typical STR marker, on each chromosome a sequence of nucleobases is typically repeated a number of times to form an *allele*.

But the sequence and the number of repetitions can be different between the two chromosomes.

Example:

Chromosome

1 --- CGATCGATCGATCGATCGATCGATCGATCGATCGATCGAT --- *10 repetitions of "CGAT"*
GCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTA

2 --- AATAATAATAATAATAAT --- *6 repetitions of "AAT"*
TTATTATTATTATTATTA

STR alleles are entirely inherited from the mother and the father, but one cannot deduce which is which (unless genetic information from the mother and father is known and they do not share any alleles).

The two alleles of a marker is referred to as the marker's *genotype*.

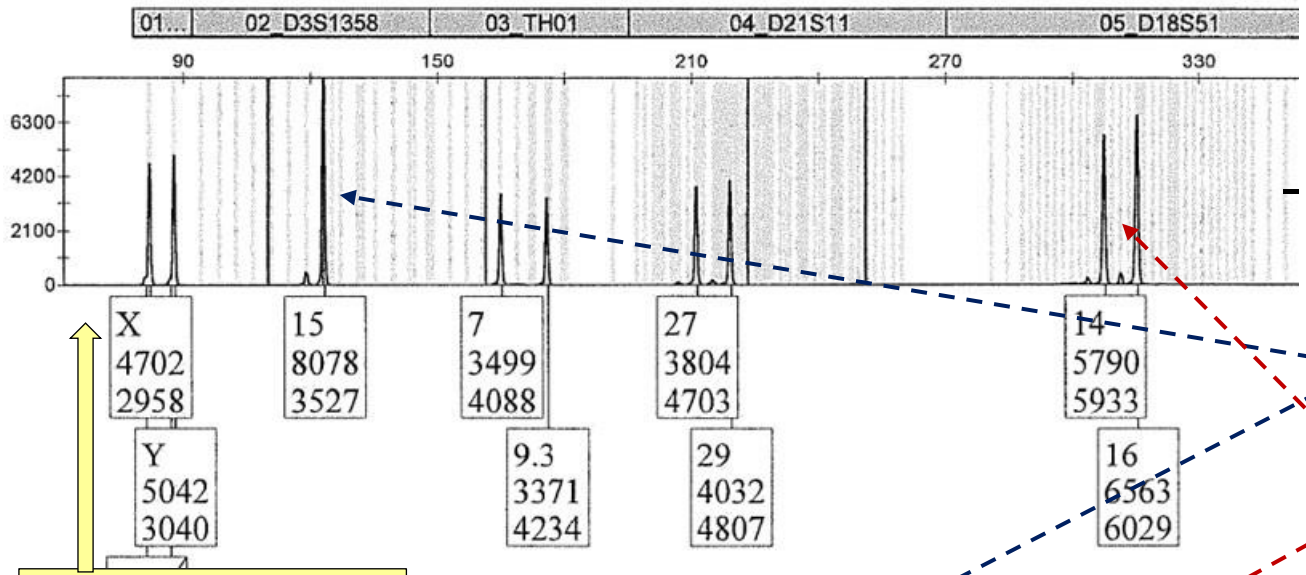
PCR (*Mullis & Smith, 1983*)

Commercial biotech companies provide so-called *kits* to be used with PCR to amplify and recovered DNA (usually very small amounts) in a number of STR markers.
(Today's standard in Europe is 23 autosomal markers, one sex-defining marker and 3 *Y-chromosomal* markers.)

Capillary electrophoresis

Takes the output from PCR and separates the chemical compounds (with respect to alleles lengths) into signals that can be visualised in an *electropherogram*.

Example: Electropherogram used with a kit of 15 autosomal markers and *typing*.



Sex chromosome marker

Homozygote genotype (two identical alleles- one peak)

Heterozygote genotype (two different alleles- two peaks)

Marker	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16
Allele 1	15	7	27	14	16	17.3	18	11	15	15	12	21	11	17.3	15
Allele 2	15	9.3	29	16	16	18.3	25	12	16	19	13	22.2	11	19	16

The allele code in a marker is simply the number of repeats of a certain sequence.

A complete set of 15 genotypes is referred to as a *DNA profile*.

H_m: The suspect is the source of the DNA extracted from the saliva stain.
H_a: Someone else than the suspect is the source of the DNA extracted from the saliva stain.

Now, assume it was possible to type all 15 markers in the DNA recovered from the saliva stain, and the profile was

Marker	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16
Allele 1	15	7	27	14	16	17.3	18	11	15	15	12	21	11	17.3	15
Allele 2	15	9.3	29	16	16	18.3	25	12	16	19	13	22.2	11	19	16

The suspect was swabbed, and his DNA profile was typed to

Marker	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16
Allele 1	15	7	27	14	16	17.3	18	11	15	15	12	21	11	17.3	15
Allele 2	15	9.3	29	16	16	18.3	25	12	16	19	13	22.2	11	19	16

Hence, identical profiles – a *match*.

What does this mean? How should the match be evaluated?

How rare is a particular genotype in a particular marker? *Population genetics* models must be used.

Marker	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16
Allele 1	15	7	27	14	16	17.3	18	11	15	15	12	21	11	17.3	15
Allele 2	15	9.3	29	16	16	18.3	25	12	16	19	13	22.2	11	19	16

An allele (coded as the number of repetitions of a nucleobase sequence) has a marker-specific relative frequency in the population of interest.

For instance, in the profile above, the relative frequency of allele 15 in marker 02 is different from the relative frequency of allele 15 in marker 11.

For two alleles, *A* and *B* let f_A and f_B denote their relative frequencies in a particular marker.

Assuming random mating (so-called *Hardy-Weinberg equilibrium*) the *genotype* frequencies of genotypes (*A*, *A*) (homozygote) and (*A*, *B*) (heterozygote) can be calculated as

$$f_{A,A} = f_A^2 \qquad f_{A,B} = 2 \cdot f_A \cdot f_B$$

Many national populations almost satisfies Hardy-Weinberg (HW) equilibrium (at least such hypothesis is hard to reject on basis of collected data)

Adjustment (Balding & Nichols, 1994) to take into account so-called *subpopulation effects* (meaning that mating is not random, but alleles are structurally inherited along "lines" in the population):

$$f_{A,A} = \frac{(2 \cdot F_{ST} + (1 - F_{ST}) \cdot f_A) \cdot (3 \cdot F_{ST} + (1 - F_{ST}) \cdot f_A)}{(1 + F_{ST}) \cdot (1 + 2 \cdot F_{ST})}$$

$$f_{A,B} = \frac{2 \cdot (F_{ST} + (1 - F_{ST}) \cdot f_A) \cdot (F_{ST} + (1 - F_{ST}) \cdot f_B)}{(1 + F_{ST}) \cdot (1 + 2 \cdot F_{ST})}$$

where F_{ST} is the *co-ancestry coefficient* measuring the subpopulation effects (to what extent the mating is non-random).

In Sweden F_{ST} is close to 0.01.

Example

A study was made in a population where the coancestry coefficient is estimated to be around 3 % . The following results were obtained for marker TH01:

Allele	Relative frequency
6	0.295
7	0.147
8	0.184
9	0.232
9.3	0.026
10	0.116

Relative frequencies for the genotypes (7,8) and (8,8):

$$f_{7,8} = 2 \cdot 0.147 \cdot 0.184 \approx 0.054$$
$$f_{8,8} = 0.184^2 \approx 0.034$$

Assuming Hardy-Weinberg equilibrium

$$f_{7,8} = \frac{2 \cdot (0.03 + (1 - 0.03) \cdot 0.147) \cdot (0.03 + (1 - 0.03) \cdot 0.184)}{(1 + 0.03) \cdot (1 + 2 \cdot 0.03)} \approx 0.066$$
$$f_{8,8} = \frac{(2 \cdot 0.03 + (1 - 0.03) \cdot 0.184) \cdot (3 \cdot 0.03 + (1 - 0.03) \cdot 0.184)}{(1 + 0.03) \cdot (1 + 2 \cdot 0.03)} \approx 0.059$$

Assuming substructures

Marker	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16
Allele 1	15	7	27	14	16	17.3	18	11	15	15	12	21	11	17.3	15
Allele 2	15	9.3	29	16	16	18.3	25	12	16	19	13	22.2	11	19	16

How rare is the entire profile?

Linkage equilibrium:

Genotypes at different markers become less statistical dependent with the distance them between in the double helix – due the recombinations at the meiosis phase.

Independence is empirically proven for markers situated on different chromosomes.

Markers chosen in forensic kits for typing short tandem repeats (STR) markers satisfy the assumption of (approximate) independence and are said to be in *linkage equilibrium* (LE).

With linkage equilibrium the relative frequency of a DNA profile can be calculated from the genotype relative frequencies:

$$f_{\text{profile}} = f_{A_1,B_1} \cdot f_{A_2,B_2} \cdot \ldots \cdot f_{A_L,B_L}$$

L = number of markers in the kit

(A_i , B_i) is the genotype of locus i ($A_i \neq B_i$ or $A_i = B_i$)

Linkage equilibrium implies that a profile relative frequency at a very fast rate goes towards zero when the number of markers used increases.

With a full 15-marker profile typical relative frequencies are of magnitude less than 10^{-14}.

Are these actually to be considered as relative frequencies?

Locus	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Allele 1	15	7	27	14	16	17.3	18	11	15	15	12	21	11	17.3	15
Allele 2	15	9.3	29	16	16	18.3	25	12	16	19	13	22.2	11	19	16
$f_{A,B}$	0.085	0.140	0.016	0.051	0.020	0.028	0.026	0.192	0.254	0.017	0.099	0.011	0.152	0.008	0.026

The genotype relative frequencies have been calculated using allele relative frequencies obtained from a database from an average modern Swedish population and assuming subpopulation effects with $F_{ST} = 0.01$

The relative frequency of this profile is calculated to $4 \cdot 10^{-21}$

With a population of almost 10 million inhabitants this cannot be a profile belonging to that population if the value is to be taken for an observed relative frequency.

Actually, one estimates that just above $100 \cdot 10^9$ human beings have ever existed on earth. Even in this population the value cannot be an observed relative frequency.

$\boldsymbol{H_m}$: The suspect is the source of the DNA extracted from the saliva stain.
$\boldsymbol{H_a}$: Someone else than the suspect is the source of the DNA extracted from the saliva stain.

The evaluation in this case:

Evidence:

$\boldsymbol{E}$: "A match in DNA profile (matches in all 15 autosomal markers of an ESX16-profile and match in the sex-defining marker) "

Value of evidence: $V = \frac{P(\boldsymbol{E}|\boldsymbol{H_m})}{P(\boldsymbol{E}|\boldsymbol{H_a})}$ *A Bayes factor*

How to find (estimates of) the numerator and the denominator?

$\boldsymbol{H_m}$: The suspect is the source of the DNA extracted from the saliva stain.
$\boldsymbol{H_a}$: Someone else than the suspect is the source of the DNA extracted from the saliva stain.

$$V = \frac{P(\boldsymbol{E}|\boldsymbol{H_m})}{P(\boldsymbol{E}|\boldsymbol{H_a})}$$

$$P(\boldsymbol{E}|\boldsymbol{H_m})$$

If the suspect actually is the source we expect to obtain matches in all markers.

There is no genetic reason for any variation (besides mutations, but such interventions can usually be controlled).

There could be variation due to deficiencies with the equipment or with the operators (reading off the wrong values).

However, it is generally non-debatable to set this probability to 1.

$\boldsymbol{H_m}$: The suspect is the source of the DNA extracted from the saliva stain.

$\boldsymbol{H_a}$: Someone else than the suspect is the source of the DNA extracted from the saliva stain.

$$V = \frac{P(\boldsymbol{E}|\boldsymbol{H_m})}{P(\boldsymbol{E}|\boldsymbol{H_a})}$$

$P(\boldsymbol{E}|\boldsymbol{H_a})$

If someone else is the source of the DNA, what is the probability of obtaining the match?

Sometimes things become clearer if we formulate the evidence in terms of the variables

$\boldsymbol{E_c}$: DNA profile of the saliva stain

$\boldsymbol{E_s}$: DNA profile of the suspect

The evidence can then be written

$$\boldsymbol{E} = (\boldsymbol{E_c} = \Gamma, \boldsymbol{E_s} = \Gamma)$$

where Γ is the profile obtained both from the saliva stain and from the suspect.

H_m: The suspect is the source of the DNA extracted from the saliva stain.
H_a: Someone else than the suspect is the source of the DNA extracted from the saliva stain.

E_c : DNA profile of saliva stain
E_s : DNA profile of suspect

$$\Rightarrow$$

$$V = \frac{P(E|H_m)}{P(E|H_a)} = \frac{P(E_c = \Gamma, E_s = \Gamma|H_m)}{P(E_c = \Gamma, E_s = \Gamma|H_a)}$$

$$= \frac{P(E_c = \Gamma|E_s = \Gamma, H_m) \cdot P(E_s = \Gamma|H_m)}{P(E_c = \Gamma|E_s = \Gamma, H_a) \cdot P(E_s = \Gamma|H_a)} =$$

$$= \left\langle \begin{matrix} \text{Suspect's profile (isolated) does} \\ \text{not depend on the hypotheses} \end{matrix} \right\rangle = \frac{P(E_c = \Gamma|E_s = \Gamma, H_m) \cdot P(E_s = \Gamma)}{P(E_c = \Gamma|E_s = \Gamma, H_a) \cdot P(E_s = \Gamma)} =$$

$$\left\langle \begin{matrix} \text{If someone else is the source, the suspect's profile} \\ \text{cannot have any impact on the profile of the stain} \end{matrix} \right\rangle = \frac{P(E_c = \Gamma|E_s = \Gamma, H_m)}{P(E_c = \Gamma|H_a)}$$

Now, the denominator is the probability of obtaining the profile Γ of the stain if the source is someone else than the suspect.

This probability should account for the rarity of this profile in the population of potential sources of the stain.

H_m: The suspect is the source of the DNA extracted from the saliva stain.
H_a: Someone else than the suspect is the source of the DNA extracted from the saliva stain.

E_c : DNA profile of crime stain
E_s : DNA profile of suspect

$$V = \frac{P(E_c = \Gamma | E_s = \Gamma, H_m)}{P(E_c = \Gamma | H_a)}$$

$$P(E_c = \Gamma | H_a)$$

Is this probability higher for certain groups of the population of potential sources (i.e. is the population stratified with respect to the occurrence of this profile)?

Note! Since the DNA is from a male (due to the sex defining marker) the population only consists of males.

What about

- an identical twin of the suspect?
- a full brother of the suspect?
- the suspect's father?
- a son of the suspect?
- a half-brother of the suspect?
- the grand-fathers of the suspect?
- an uncle or a male cousin of the suspect?

If stratification should be taken into account, we need to use a so-called full Bayesian approach and compute the value of evidence as the Bayes factor

$$B = \frac{P\left(\boldsymbol{E_c} = \Gamma \middle| \boldsymbol{E_s} = \Gamma, \boldsymbol{H_m}\right)}{\sum P(\boldsymbol{E_c} = \Gamma | \text{Individual } i \text{ is the source}, \boldsymbol{H_a}) \cdot P\left(\text{Individual } i \text{ is the source} \middle| \boldsymbol{H_a}\right)}$$

…where the sum is over all individuals in the population of possible sources except for the suspect.

However, this will need knowledge about the prior probabilities

$$P\left(\text{Individual } i \text{ is the source} \middle| \boldsymbol{H_a}\right),$$
$$i = 1,2, \ldots$$

of which the forensic scientist has no opinion (and should not have).

Hence, the evidentiary strength cannot be assessed without prior opinions about which persons could have been involved.

To be able to report measures of evidentiary strength, NFC (and laboratories/institutes in other countries) formulate a different alternative hypotheses.

First choice: $\boldsymbol{H_a}$: "Someone else, not closely related to the suspect, is the source"

$$V = \frac{P\left(\boldsymbol{E_c} = \Gamma \middle| \boldsymbol{E_s} = \Gamma, \boldsymbol{H_m}\right)}{P\left(\boldsymbol{E_c} = \Gamma \middle| \boldsymbol{H_a}\right)}$$

The denominator of V can now be estimated from a random sample of individuals from the population to which the source is assumed to belong.

Such a random sample is (today) a kind of panel, i.e. several persons from a general population (covering the population of potential sources with negligible effects of over coverage)

$\Rightarrow$ DNA population database

Hence, $P(\boldsymbol{E_c} = \Gamma | \boldsymbol{H_a})$ is estimated by calculating the relative frequency of this profile using the database.

Less problematic that this relative frequency is not possible to physically obtain in the population, it is used to estimate a probability through a *model* of the population.

For the current profile we previously obtained a calculated relative frequency of $4 \cdot 10^{-21}$.

$$V = \frac{P(\boldsymbol{E_c} = \Gamma | \boldsymbol{E_s} = \Gamma, \boldsymbol{H_m})}{P(\boldsymbol{E_c} = \Gamma | \boldsymbol{H_a})} = \frac{1}{4 \cdot 10^{-21}} = 2.5 \cdot 10^{20}$$

The match is thus $2.5 \cdot 10^{20}$ times more probable to obtain if the suspect is the source than if someone else, not closely related to the suspect, is the source.

Was it him?

Another alternative hypothesis may be

$\boldsymbol{H_{a,2}}$: "The source of the DNA is a full brother of the suspect"

We then need more population genetics to calculate the probability

$$P\left(\boldsymbol{E_c} = \Gamma \middle| \boldsymbol{H_{a,2}}\right)$$

For the current profile an estimate of this probability becomes $1.82 \cdot 10^{-7}$

Hence, the value of evidence is

$$V^{(2)} = \frac{P\left(\boldsymbol{E_c} = \Gamma \middle| \boldsymbol{E_s} = \Gamma, \boldsymbol{H_m}\right)}{P\left(\boldsymbol{E_c} = \Gamma \middle| \boldsymbol{H_{a,2}}\right)} = \frac{1}{1.82 \cdot 10^{-07}} = 5.5 \cdot 10^6$$

The match is thus 5.5 million times more probable to obtain if the suspect is the source than if a full brother of the suspect is the source.

Besides identical twins, full siblings of the same sex are the closest related individuals.

Changing the alternative hypothesis to something like

“The source of the DNA is a father or a son of the suspect”

will also render a higher relative frequency (however lower than with a full brother) – and as a consequence a lower value of evidence (against the suspect) than with no close relatives in the alternative hypothesis.

It has become more and more common for a suspect to “blame the brother”. The most obvious way to handle this situation is to swab the brother.

- A mismatch directly excludes the brother.
- However, with a (utterly unexpected) match the two brothers cannot be separated by the current DNA evidence

Challenges with DNA evidence

With today's technique very small amounts of DNA can be recovered and typed (with PCR: LCN-analysis (*L*ow *C*opy *N*umber))

Small amounts of DNA is typical for so-called touch-DNA (contact with skin)

Since several persons may have been in contact with a surface of interest (someone's garments, doorhandle, table, …) it is common to observe DNA from more than one person in a sample – so-called *DNA mixtures*.

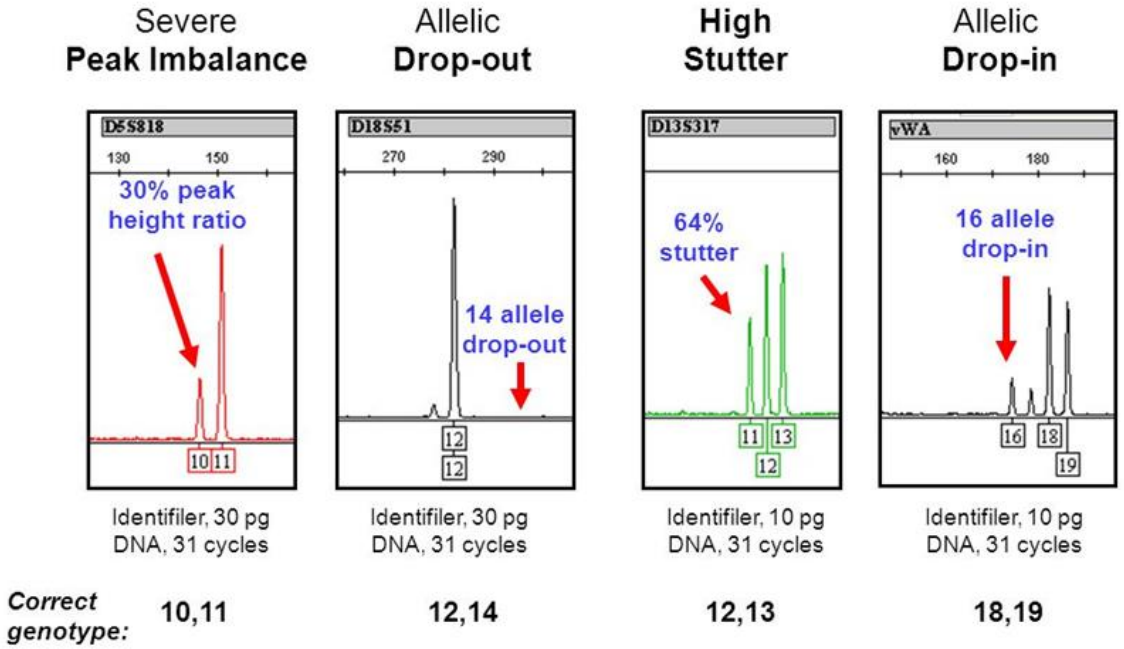
This is also often the case in sex crimes where body fluid samples contain DNA both from both the perpetrator and the victim (but sometimes also from a third or fourth person).

The hypothesis would comprise more than one person, e.g.

$\boldsymbol{H_m}$: The DNA originates from the victim and the suspect
$\boldsymbol{H_a}$: The DNA originates from the victim and an unknown person

When (very) small amounts of DNA are analysed, there is appreciable risks that…

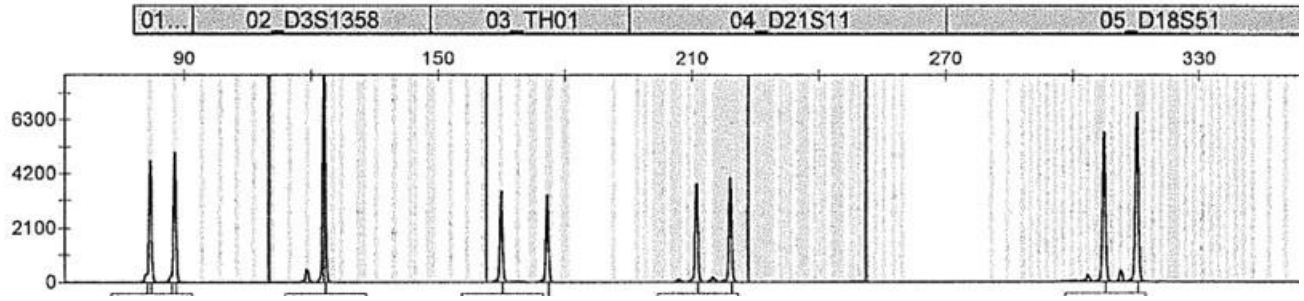
- alleles in one or several markers are not detected at all in the electropherogram (so-called *drop-out* alleles)
- peaks in a marker (if more than one) has substantially different heights – is it a heterozygote marker or alleles from more than one person?
- artefacts in forms of extra peaks (so-called *stutters*) aside the true peaks (a multiplying effect)
- residues from previous analyses – despite cleaning – may cause extra peaks in a marker (so-called *drop-in alleles*).



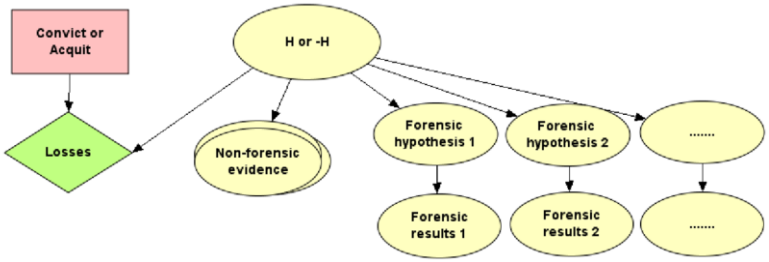
Several (commercial and non-commercial) software have been developed to handle these problems, especially for samples with DNA from more than one person (e.g. STRmix™, TrueAllele®, *EuroForMix*, *DNAxs*…)

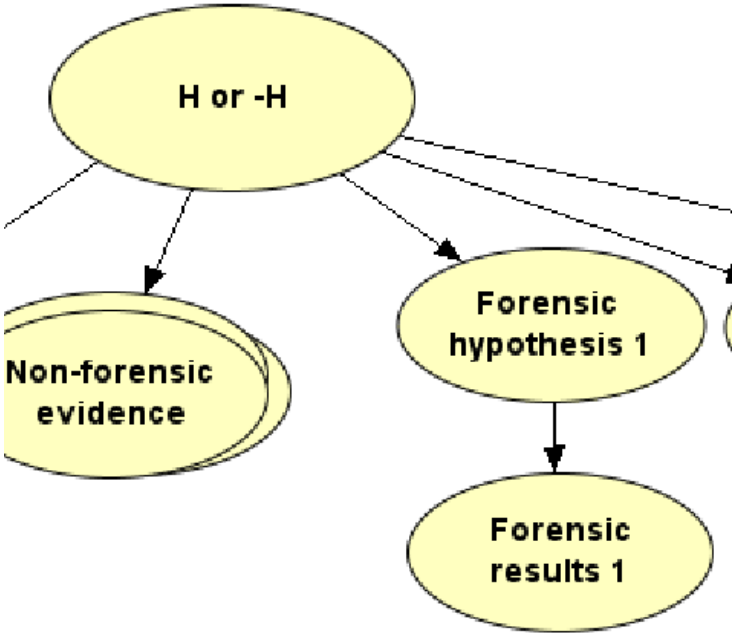
As an example, the modelling behind *EuroForMix* (Bleka, Gill) and *DNAxs* (Bishop et al) is semicontinuous.

- Typed alleles are modelled with population frequencies
- Probabilities of drop-outs and artefacts are assigned from fitting gamma distributions to the peak heights in the electropherogram (*Cowell et al, 2007*)



But how do the evaluation of the forensic results propagate back to the ultimate hypotheses in court?





H_m: The suspect is the source of the DNA extracted from the saliva stain.

H_a: Someone else than the suspect is the source of the DNA extracted from the saliva stain.

E_c : DNA profile of the saliva stain

E_s : DNA profile of the suspect

With an extremely high value of the Bayes factor, it can in principle be taken for proven that the suspect is the source of the DNA

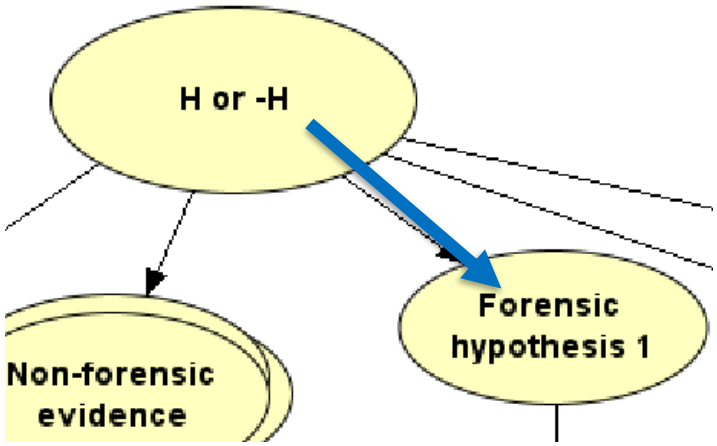
$$\frac{P(\boldsymbol{H_m}|\boldsymbol{E})}{P(\boldsymbol{H_a}|\boldsymbol{E})} = \frac{P(\boldsymbol{E}|\boldsymbol{H_m})}{P(\boldsymbol{E}|\boldsymbol{H_a})} \times \frac{P(\boldsymbol{H_m})}{P(\boldsymbol{H_a})}$$

$$2.5 \cdot 10^{20}$$

The prior odds must be extremely low for $P(\boldsymbol{H_m}|\boldsymbol{E})$ to be that low that it is disputable whether $\boldsymbol{H_m}$ is true or not.

But...what about

$$P(\boldsymbol{H}|\boldsymbol{H_m} \text{ is true})$$



A challenge is when there is no longer a dispute on ***who's*** DNA it is.

Infancy of DNA evidence evaluation: Often sufficient to confront the suspect like "We've got your DNA!"

In course of time, culprits have learnt that there are loopholes in the interpretation of DNA evidence.

- *Blaming on a close relative* – will be less efficient as the amount of DNA analysed is increasing (more STR-markers, sequencing).

- *Questioning **how** the DNA was deposited* – Claiming a secondarily or even tertiarily DNA transfer from an innocent contact.

  Particularly common in sex crimes where the suspect denies having sexually assaulted the victim but claims they had only social contact (e.g. drinking and/or dancing together).

In such cases the hypothesis are no longer about the source of the recovered DNA (since there is no dispute on that).

They must address *activities* (be formulated at *activity level*), e.g.

H_m : The suspect had sexual intercourse with the victim
H_a : The suspect and the victim had social contact during the party

When the evidence material is recovered body fluid (for instance DNA from the victim is found in the suspect's underwear), the amount of DNA recovered is important.

The probability of recovering substantial amounts of DNA cannot be well explained by the hypothesis H_a (e.g. pointing towards secondary transfer of saliva), but very well by hypothesis H_m (vaginal secretion).

When the amounts recovered are small it is not possible to discriminate between H_m and H_a with sufficient confidence. Sometimes it is possible to find out which type of cells (saliva secretion or vaginal secretion) the sample consisted of, but a potential secondary transfer cannot be easily rejected.

Note that the immensely high likelihood ratios obtained with the source attribution of the DNA are completely worthless in this dispute.