A STANDARD EXAMPLE

- Match evidence: the genetic profile at the crime scene matches the genetic profile of the defendant

- Prosecution hypothesis (source hypothesis): the defendant and the person who left the traces at the scene are the same person

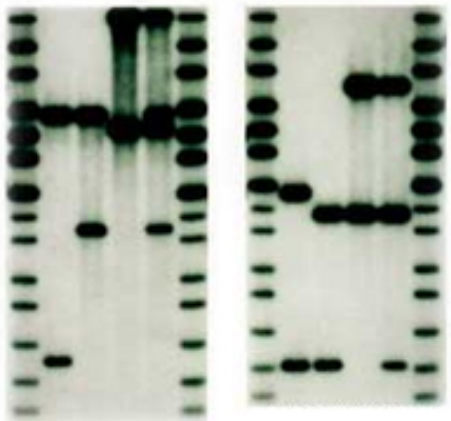
THE NEED TO QUANTIFY UNCERTAINTY

- Match evidence supports the source hypothesis to some extent, but never without doubt

DNA evidence

It is used to answer an **identification question**:

accused = person who left a DNA trace on the crime scene?



If the two DNA profiles do **not match**, then the answer is NO.

If the two DNA profiles **match**, then....what?

In addition, **the DNA profile must be rare enough**, then ... the answer is YES.

DNA evidence: two important features

match between two DNA profiles

frequency of the DNA profile in question

Rarity and uniqueness

How **rare** should the DNA profile in question be to warrant a positive answer to the identification question?

Ideally, the DNA profile must be **unique**.

But when are uniqueness claims justified?

If DNA profile has a frequency of 1 in a billion...

There is 10^{-9} chance that a random individual has it.

$$10^{-9} = 0.000000001$$

Earth population is 7 billion.
There are 7 people on earth with the same DNA profile.

If DNA profile has a frequency of 1 in 10 billion...

There is 10^{-10} chance that a random individual has it.

$$10^{-10} = 0.0000000001$$

Earth population is 7 billion. Is the DNA profile, then, unique?

If DNA profile has a frequency of 1 in 10 billion...

There is 10^{-10} chance that a random individual has it.

$$10^{-10} = 0.0000000001$$

Earth population is 7 billion. Is the DNA profile, then, unique?

NO – there is a **fifty percent chance** that another person on earth shares the DNA profile.

$$1 - (\tfrac{10{,}000{,}000{,}000-1}{10{,}000{,}000{,}000})^{7{,}000{,}000{,}000} = 1 - 0.9999999999^{7{,}000{,}000{,}000} \approx 0.5.$$

Matters of statistics $\neq$ matters of law

A positive answer to the identification would be warranted **if the DNA profile were unique**.

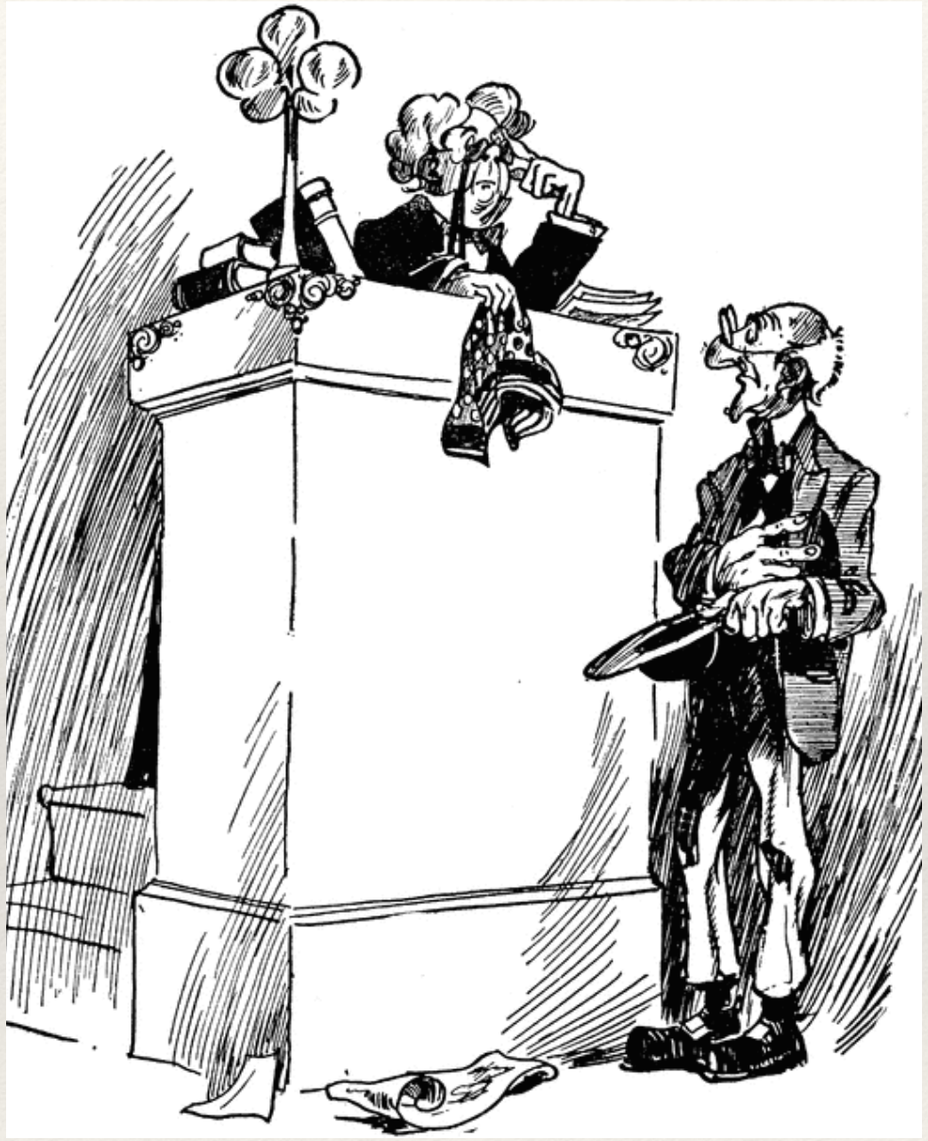
But statistics and genetics alone cannot make uniqueness claims.

Hence, establishing when a positive answer to the identification question is warranted is **not a matter of statistics**.

It is ultimately a **matter of law**.

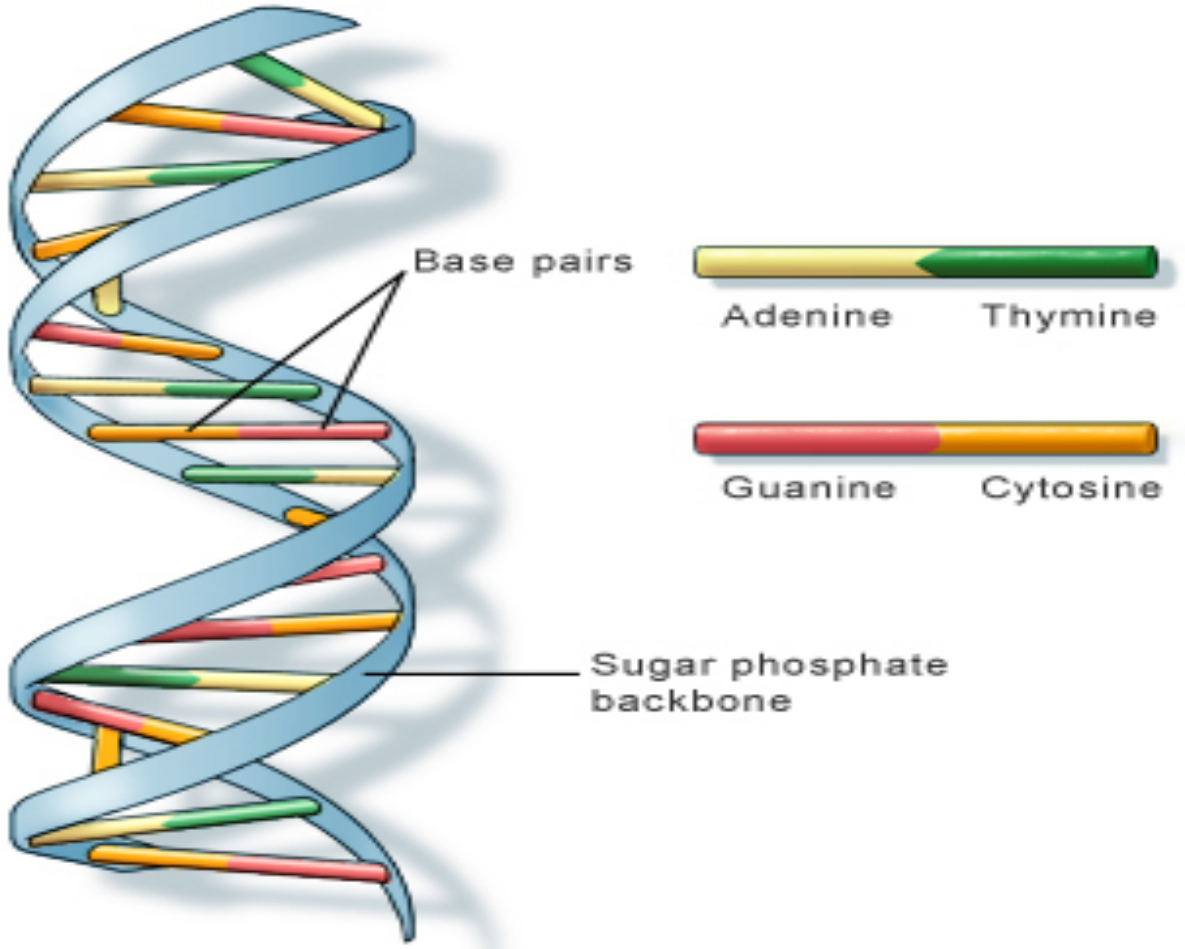
Judge Hardwick - Missouri Ct. App.

"We conclude that where, as here, DNA material is found in a location, quantity, and type **inconsistent with casual contact** and there is one in **one quintillion likelihood** that some else was the source of the material, the evidence is **legally sufficient to support a guilty verdict"**



Missouri v. Abdelmalik, 273 S.W.3d, 61, 66 (Mo. Ct. App. 2008)

The molecule of DNA



U.S. National Library of Medicine

DNA as a double-series of letters

...ATTAAGGAATAAGAGGGAAATTAAAAGG...
...TAATTCCTTATTCTCCCTTTAATTTTCC...

DNA as a double-series of letters

...ATTAAGGAATAAGAGGGAAATTAAAAGG...
...TAATTCCTTATTCTCCCTTTAATTTTCC...

Combinatorics: With roughly 3 billion sites on the human DNA, there are $4^{3,000,000,000}$ combinations. Greater than 10^{100}.

Alec Jeffreys discovered DNA fingerprinting on the morning of September 1984.

Variable Number Tandem Repeats (VNTRs)

VNTRs are regions of the DNA in which a given sequence of letters is repeated a number of times.

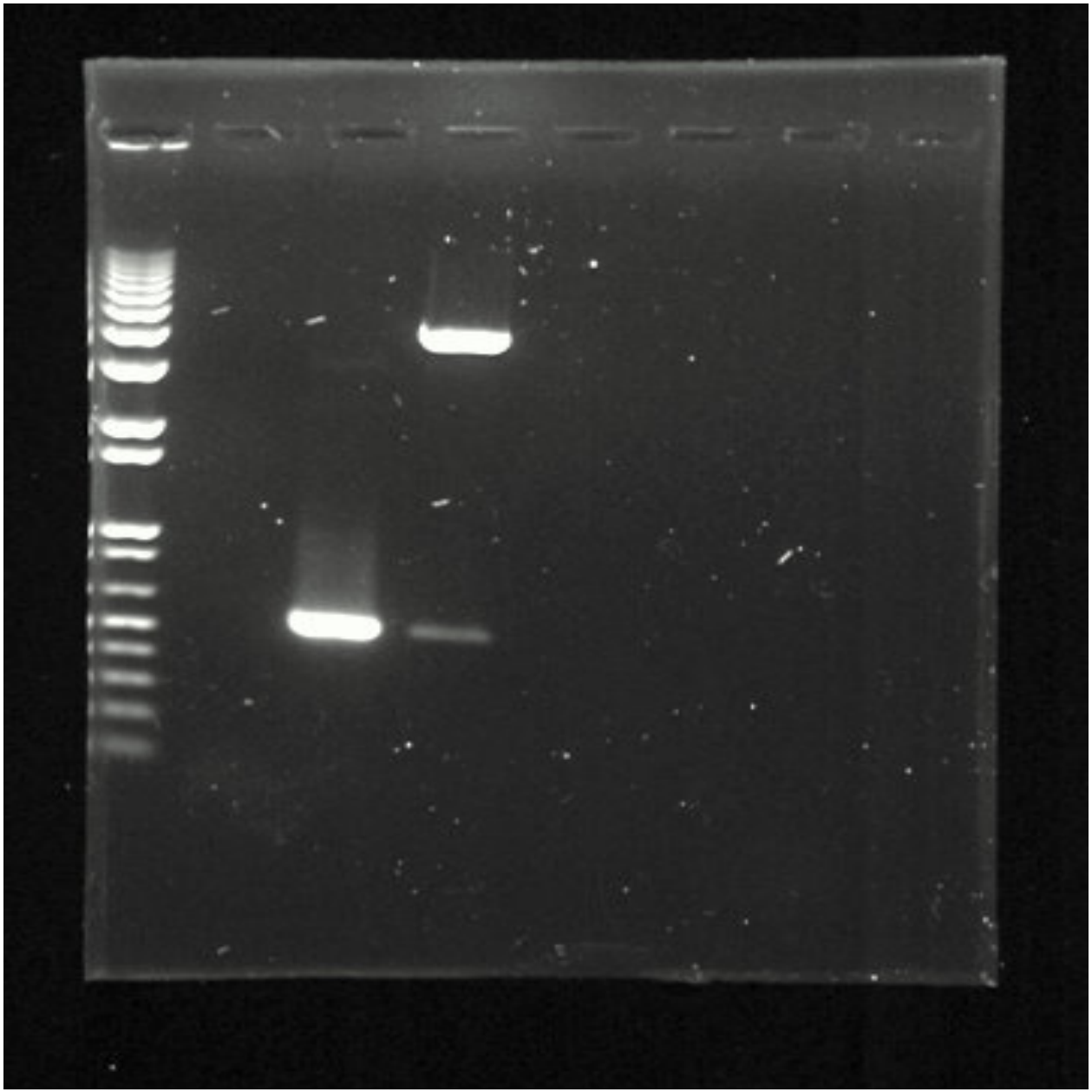
E.g. sequence AGA is repeated five times:

———————...AGA-AGA-AGA-AGA-AGA...———————

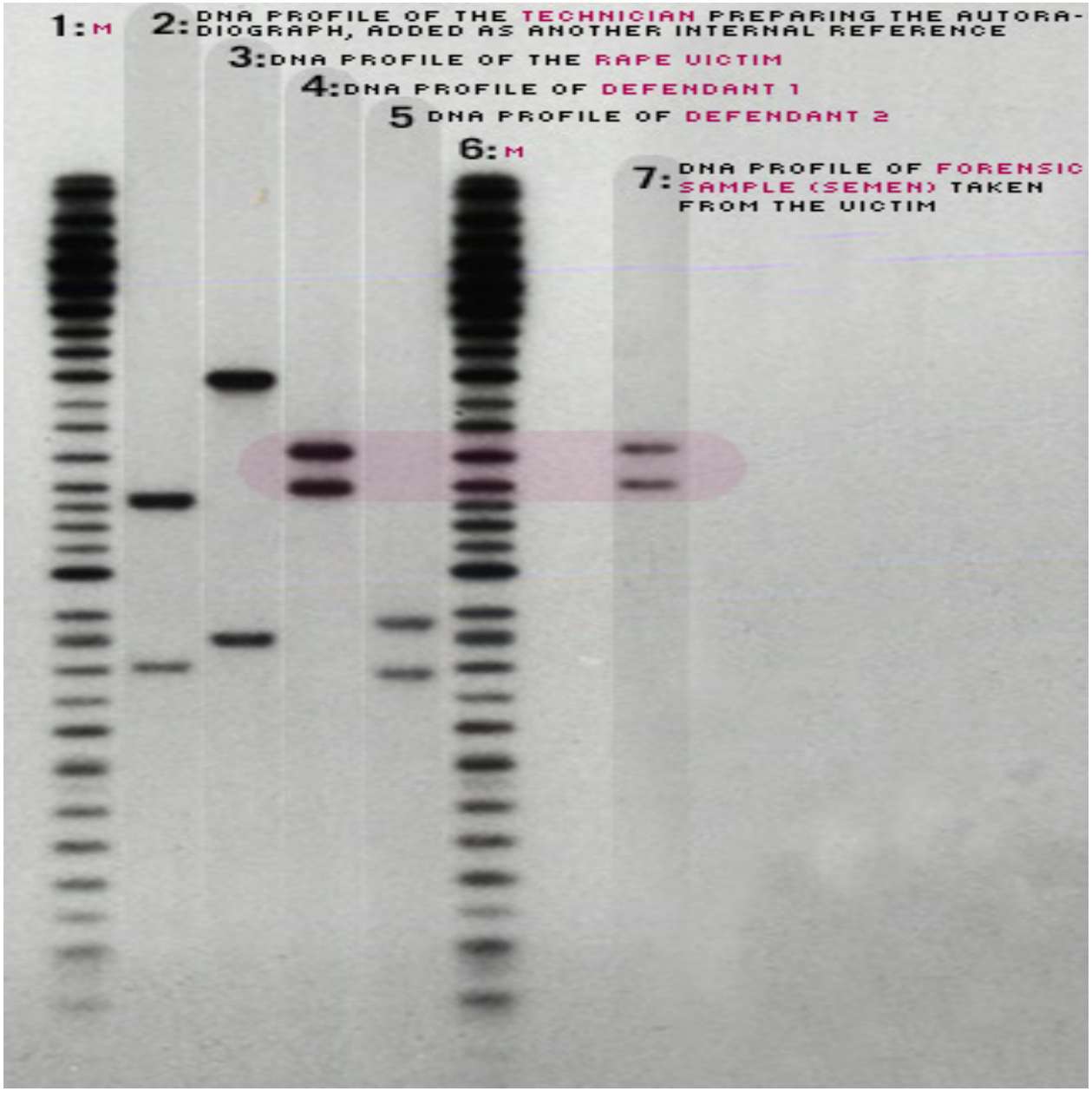
NB: Currently Short Tandem Repeats (STRs) are used since they are easier to analyze. The principle is however the same as VNTRs

VNTR regions	repetitions
D3S1258	16,18
VWA	15,20
FGA	24,26
D8S1179	28,30
D2S11	10,16
D18S51	10,13
D5S818	11,11
D72820	12,15

Number of repetitions = length of the VNTR region



A match for a single VNTR region



Sources of error in declaring a match

- The distance traversed by the VNTR fragments might not be indicative of their respective length.
- The VNTR fragments in question might have been misplaced or mishandled.
- The expert witness testifying as to the match might misreport the findings.

QUESTION: Can we quantify precisely the probability of all these errors occurring?

Recall .... DNA evidence: two important features

match between two DNA profiles

frequency of the DNA profile in question

- **Question: How do we calculate the random match probability θ?**

=> That is, the probability/ frequency of a genetic profile of interest?

- **In short: random match probability θ is the output of a genetic model paired with the frequency data available**

CALCULATING THE RANDOM MATCH PROBABILITY

- We want to know the value of

$$\theta = \Pr(Match \mid H_d)$$

the probability that a random person (someone who is not the source) would have the matching profile of interest

- A genetic profile usually consists of 15 loci or markers. Think of a locus as a marked placeholder in the genome
- Each locus will have a particular locus-specific genotype. Think of a locus-specific genotype as a combination of letters that occupies the placeholder
- Each locus-specific genotype has a probability of occurring at random, call it p_i, the locus-specific genotype probability

DNA Profile		Allele frequency from database				Genotype frequency for locus	
Locus	Alleles	times allele observed	size of database	Frequency		formula	number
CSF1PO	10	109	432	*p*=	0.25	2*pq*	0.16
	11	134		*q*=	0.31		
TPOX	8	229	432	*p*=	0.53	p^2	0.28
	8						
THO1	6	102	428	*p*=	0.24	2*pq*	0.07
	7	64		*q*=	0.15		
vWA	16	91	428	*p*=	0.21	p^2	0.05
	16						
			profile frequency=				**0.00014**

Source:

Charles H. Brenner's "Forensic mathematics of DNA matching"

https://dna-view.com/profile.htm

- **The product of the 15 individual locus-specific genotype probabilities**

$$p_1 \times p_2 \times \ldots \times p_{15}$$

is the 15-loci genotype probability, or random match probability

(or the probability/ frequency of a genetic profile of interest)

***Question*: How to calculate the locus-specific genotype probabilities p_1, p_2, etc?**

- **Since each locus-specific genotype consists of two alleles, the locus-specific genotype probability is the product of the allele probabilities**
- **Allele frequencies in a database—obtained simply by counting occurrences—are used as approximation of allele probabilities**

Determining the frequency of a DNA profile

STEP 1: frequency of each VNTR individual fragment;

$\Rightarrow$ this is done by counting the number of VNTR fragments with the same length in a given database, yielding f_i.

STEP 2: frequency of the VNTR genotype (from both parents).

$\Rightarrow F_1 = 2 \times (f_i \times f_j)$

STEP 3: frequency of the entire DNA profile (with 10 VNTR genotypes).

$\Rightarrow F = F_1 \times F_2 \times F_3 \times F_4 \times ... \times F_10$

NOTE:
Each step depends on assumptions about population genetics.

Why be cautions about DNA evidence

Three sources of uncertainty in DNA evidence:

1. Uniqueness claims are not warranted by statistics alone.
   We only have an estimate of the frequency of the DNA profile.
2. Frequency estimates themselves could be wrong.
3. Declarations of a match could be wrong.

Why be cautions about DNA evidence

Three sources of uncertainty in DNA evidence:

1. Uniqueness claims are not warranted by statistics alone.
   We only have an estimate of the frequency of the DNA profile.
2. Frequency estimates themselves could be wrong.
3. Declarations of a match could be wrong.

Even if those sources of uncertainty were completely removed,
DNA evidence would help us answer only question (Q1):

(Q1) accused = person who left a DNA trace on the crime scene ?

But there are two more important identification questions:

(Q2) accused = person who committed the criminal act?

(Q3) accused = guilty party ?

A further difficulty: The significance of a match

Standard case:
The accused is identified through non-DNA evidence and—only afterwards—his DNA is matched with the DNA from the crime scene.

Cold hit case:
The accused is identified through a database search of existing DNA profiles.

A further difficulty: The significance of a match

Standard case:
The accused is identified through non-DNA evidence and—only afterwards—his DNA is matched with the DNA from the crime scene.

Cold hit case:
The accused is identified through a database search of existing DNA profiles.

Some argue that a DNA match in the cold hit case is less significant than a DNA match in the standard case.

Argument: There is a very high chance of getting, say, 10 consecutive heads if one makes a sufficient number of attempts at tossing a coin. Likewise, there is a very high chance of getting a match if the database is sufficiently large.