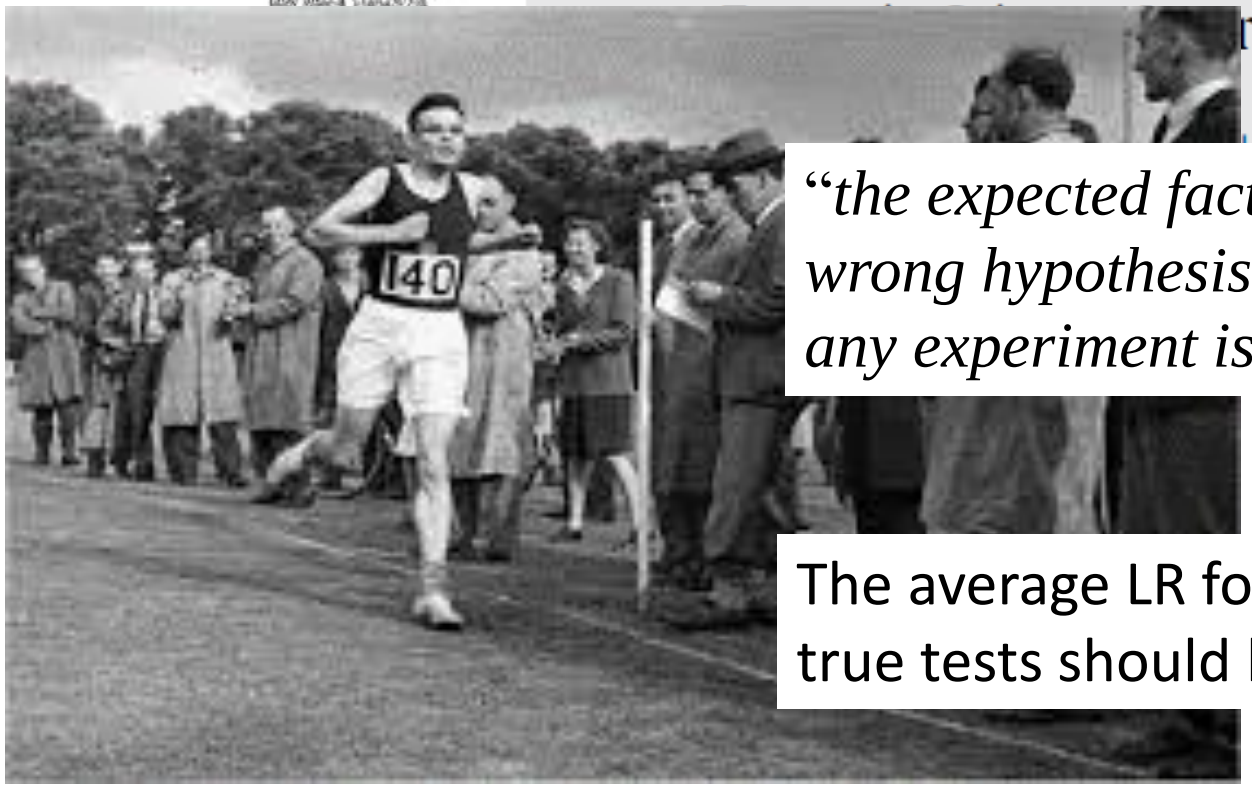
False donors and Importance sampling

Forensic Science International: Genetics 16 (2015) 165–171

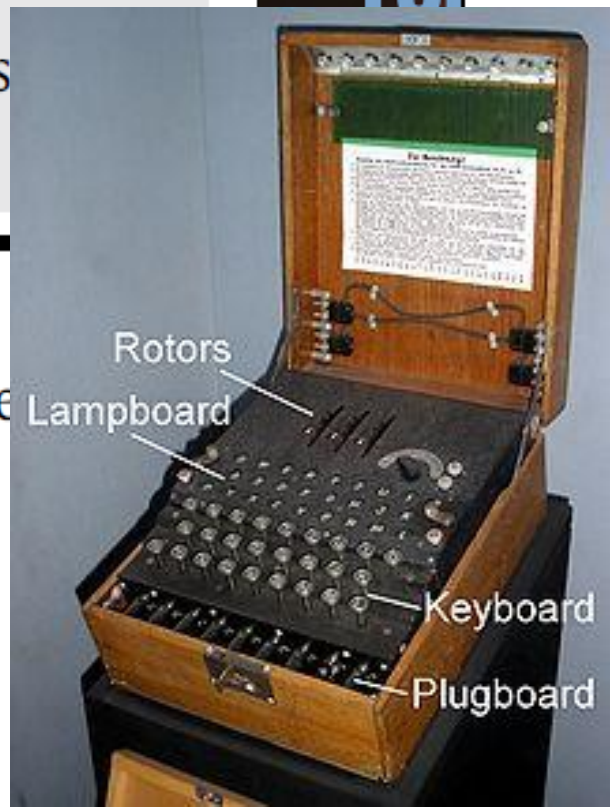
Contents lists available at ScienceDirect

rnational: Genetics



"the expected factor for a wrong hypothesis in virtue of any experiment is 1."

The average LR for the H_d true tests should be 1



From Turing we can infer that

$$\tilde{p} \leq \frac{1}{LR_{POI}}$$ Equation 2

The chance of an *LR* greater than or equal to LR_{POI} is less than $1/LR_{\mathrm{POI}}$

This is true for every *LR* not just LR_{POI}

False donor testing

- This tests known false donors against the profile
- Either use a database (say staff) or
- Simulate
- Run against the profile with your system,
- Record the results and present (?)
- Problem …. To test *LR* = x you need at least x

Forensic Science International: Genetics 16 (2015) 165–171

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Forensic population genetics – original research

Testing likelihood ratios produced from complex DNA profiles

Duncan Taylor [a,b,*], John Buckleton [c], Ian Evett [d]

Please consider a single source profile, $locus_1$ = ab

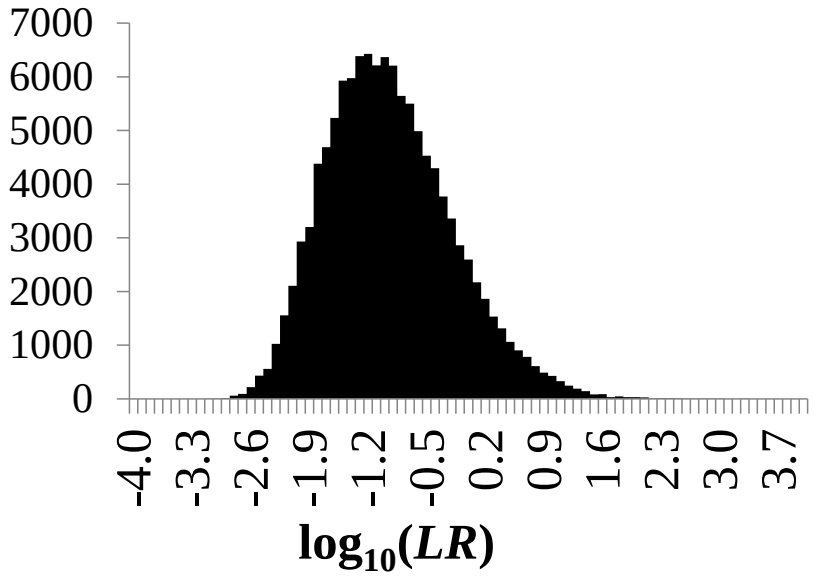
- Start sampling randomly according to allele probabilities,
- Every time you sample an allele that is not a or b you could stop,
- You are wasting most of your time,
- The *LR* for all of these is 0,
- Please mentally extend to 21 or 24 loci.

The distribution of Hd true

-the shape depends on the profile

-there will be a maximum,

- Not directly known to us but potentially calculable
- this is probably slightly bigger that the largest Hp true



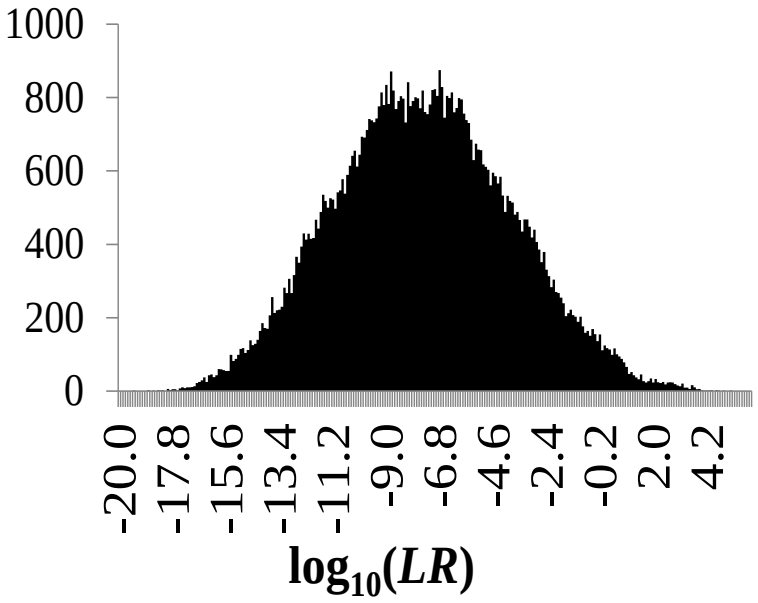
A low level four person mixture (4:3:2:1 pg) 12 loci where none of contributors are assumed. All H_p true LRs were low and again there were no instances of H_d true $LR = 0$. The average H_d true LR was 0.927, and equation 2 held for all H_p true LRs.

Equation 2

$$\tilde{p} \leq \frac{1}{LR_{POI}}$$

Average LR		LR for Hp true
0.927	C_1	4
	C_2	7
	C_3	5
	C_4	6

This profile was generated from three individuals (100:100:100pg 9 loci), who contained a lot of masking. Only two of the nine STR loci exhibited more than four allelic peaks. The result was a range of H_p true *LR*s. Equation 2 held true, with no observations of an H_d true *LR* appearing above the H_p true *LR* for C_1. Again the average H_d true *LR* was close to 1.



Average LR		LR for Hp true	Log(LR)
0.91	C_1	234,738	5.37
	C_2	2,530	3.40
	C_3	43	1.63

Importance sampling

- Modern PG software can produce a list of genotypes that have some chance of explaining the profile,
- This is called the “weight”,
- A high weight helps a high *LR*,
- A zero weight means a 0 *LR*,

Importance sampling

- We should sample at genotype probability, p_i
- But we sample at weight probability, w_i
- And reweight the answers by p_i/w_i
- Let us say we sample ab LR = 33, w_i = 1 p_{ab} = 0.03
- We score an LR of 33 but at a bias of 0.03/1 =0.03

Bias

- For each of the 'y' H_d true tests we produce a genotype set S_y and calculate a bias, b_y
- Ratio of the probability of the choice using an unbiased method to the probability of that choice had the biasing method been used

Average LR approximation

- Average *LR* (over the y tests) assuming a naïve simulator had been used is:

$$\overline{LR} = \frac{1}{Y}\sum_{y} LR_y b_y$$

Number of naïve simulations

- Number of simulations (I) assuming a naïve simulator was used is approximated by:

$$I = \frac{\sum_y LR_y}{\overline{LR}}$$

I find this version easier $$I = \frac{y\sum_y LR_y}{\sum_y LR_y b_y}$$

Effective count for $S_y = c_y$

- Adjusted count assuming naïve simulation method was used

$$c_y = b_y \times I$$

I find this version easier $c_y = \dfrac{y b_y \sum_y LR_y}{\sum_y LR_y b_y}$

Tail area probability, *p*

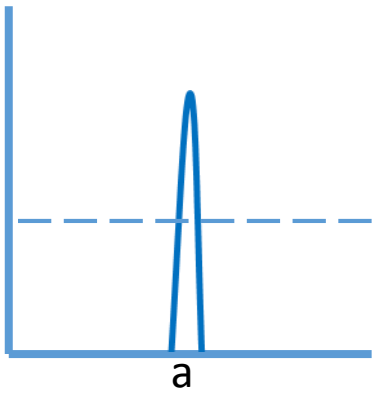
- The values for *p* (the proportion of non-donors who would yield a *LR* greater than or equal to that of the LR_j) is:

$$p = \frac{I}{Y} \times \sum_{i:LR_i \geq LR_j} c_y$$

LR_j could be anything you are interested in… i.e. 10^3, LR_{POI}

One locus example

Genotype	Weight
a,a	0.7
a,Q	0.3
Q,Q	0.0



The a allele is rare

Pr(a) = 0.02, Pr(Q) = 1-0.02 = 0.98

Pr(aa) = 0.0004 (0.02^2)

Pr(aQ)=0.0392 (2 x 0.02 x 0.98)

Pr(QQ)=0.9604 (0.98^2)

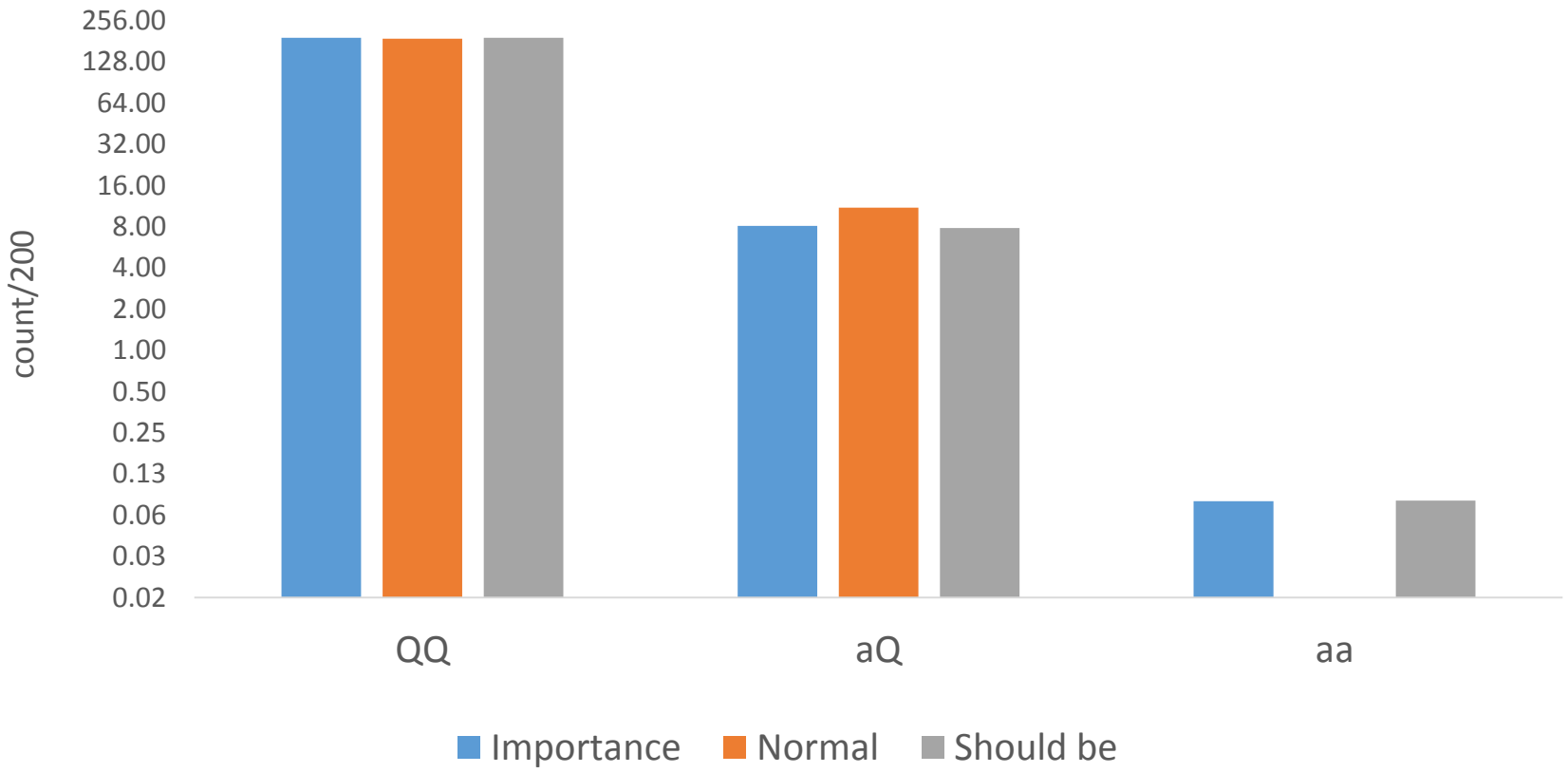
One locus example

Genotype	Weight
a,a	0.7
a,Q	0.3

- We sample from genotypes a,a and a,Q
- ~70% of the time it will be a,a ~30% a,Q, never Q,Q
- Calculate the LR
- Calculate the bias
- Calculate LR x bias
- Calculate $\overline{LR}$ and I

Plot of LRs

Importance sampling always better than naïve sampling



Forensic Science International: Genetics 27 (2017) 74–81

Contents lists available at ScienceDirect

Forensic Science International: Genetics

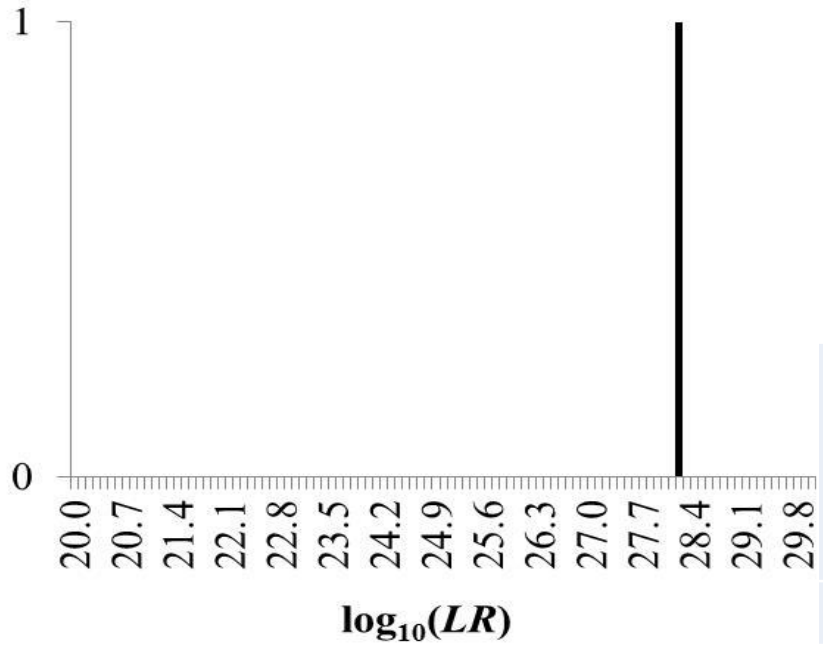
journal homepage: www.elsevier.com/locate/fsig

Research paper

Importance sampling allows H_d true tests of highly discriminating DNA profiles

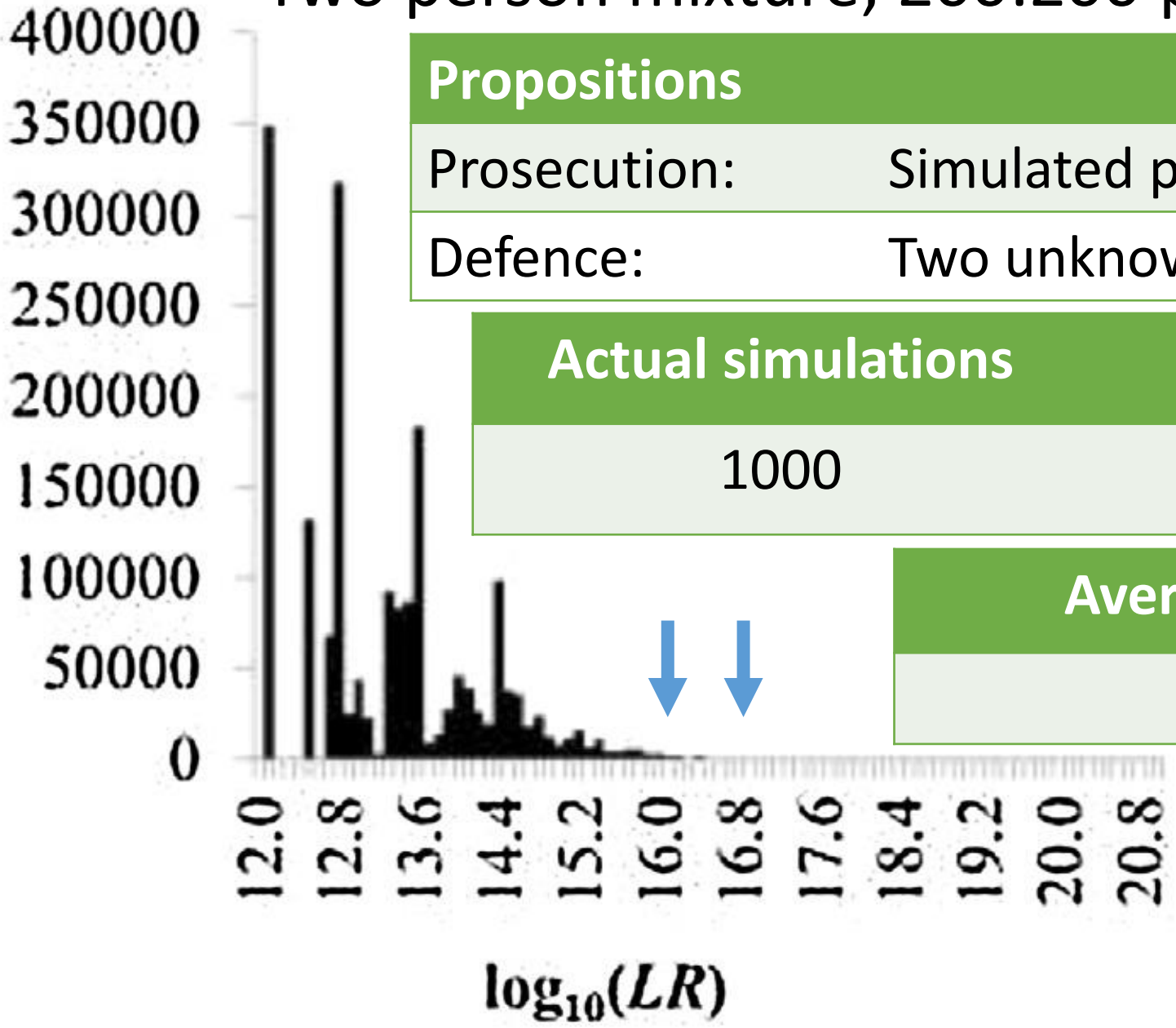
Duncan Taylor[a,b,*], James M. Curran[c], John Buckleton[d,e]

Globalfiler 400pg single source



equivalent # naïve simulations	Average LR		LR for Hp true	Log(LR)
1.45×10^{28}	1	C_1	1.45×10^{28}	28.2

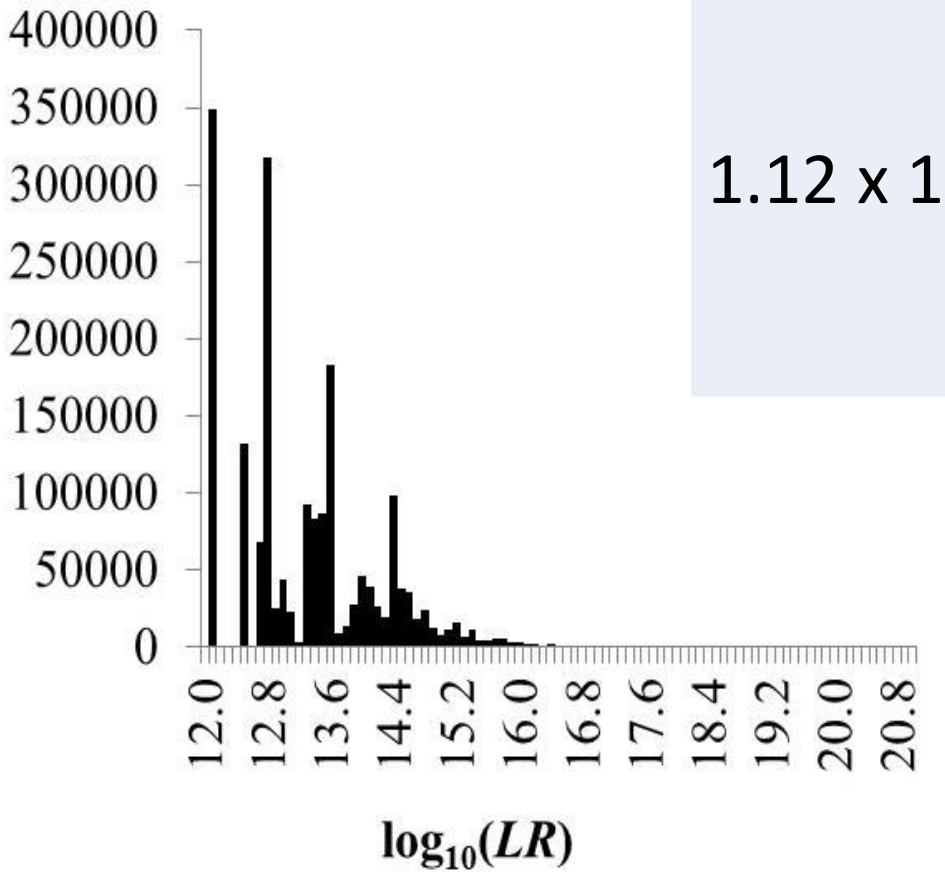
Two person mixture; 200:200 pg; GlobalFiler kit



Propositions	
Prosecution:	Simulated profile + Unknown
Defence:	Two unknowns

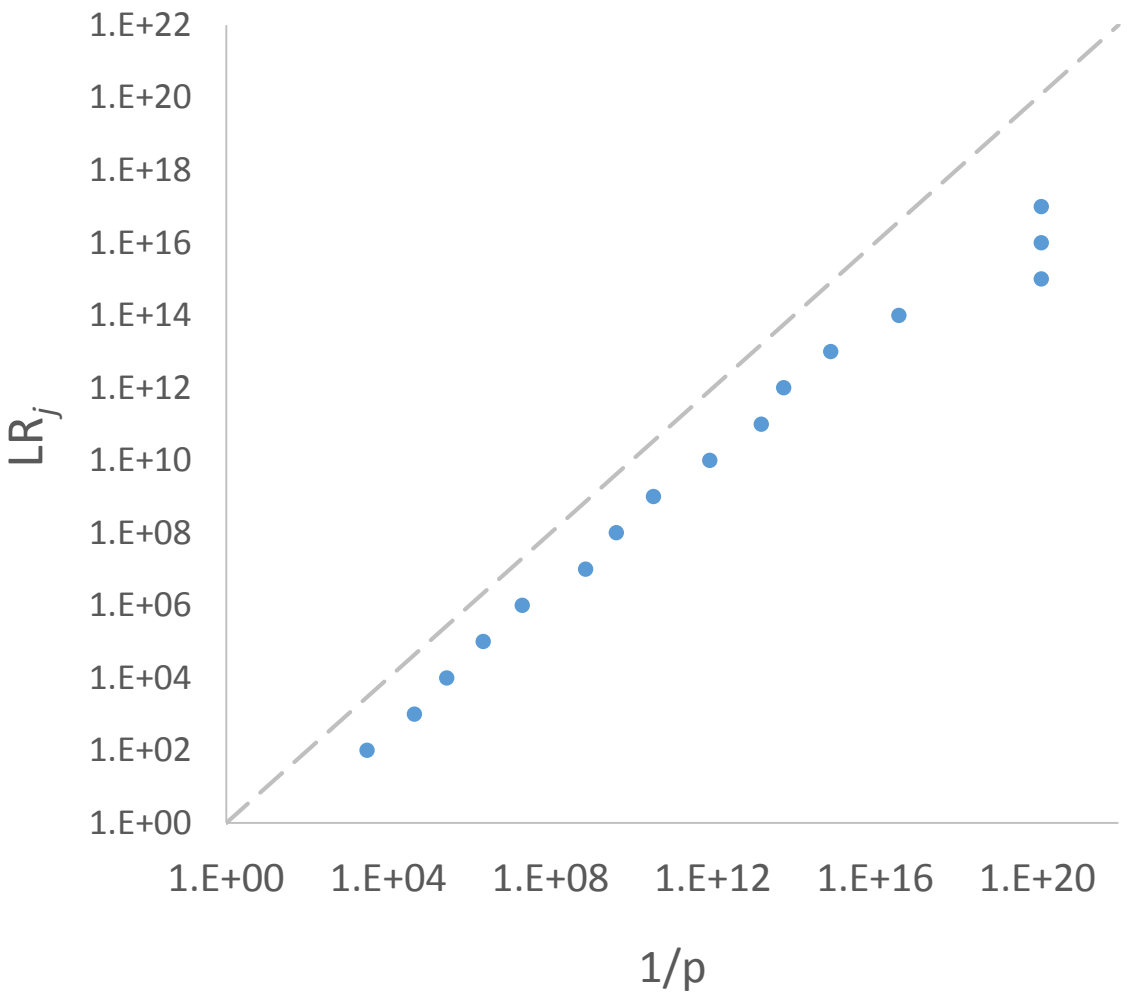
Actual simulations	Equivalent 'naïve'
1000	1.12×10^{21}

Average H_d true LR
1.12



equivalent # naïve simulations	Average LR	LR for Hp true	Log(LR)	1/P ≥
1.12×10^{21}	1.12	6.54×10^{16}	16.81	4.34×10^{17}
		1.22×10^{16}	16.08	9.35×10^{16}

Results of H_d true tester



Conclusion

- This is a large H_d true test for the sample in question
- Close to real time validation of the exact case. On the fly validation…
- Provides additional support for the case
- Can be used to inform statements about the *LR*

Dr Jamieson had been sacked from the force in 1996 after he was found guilty of placing a female motorist in a state of fear and alarm on the M8.

He was alleged to have pulled the woman over to lecture her about her driving after flashing a police sign and was later fined £300 at Airdrie Sheriff Court

A senior police scientist has resigned following an internal inquiry into allegations of a conflict of interest.

Report of Professor Allan Jamieson in the case of Donte Lee

8th May 2017

Occupation: Director of The Forensic Institute

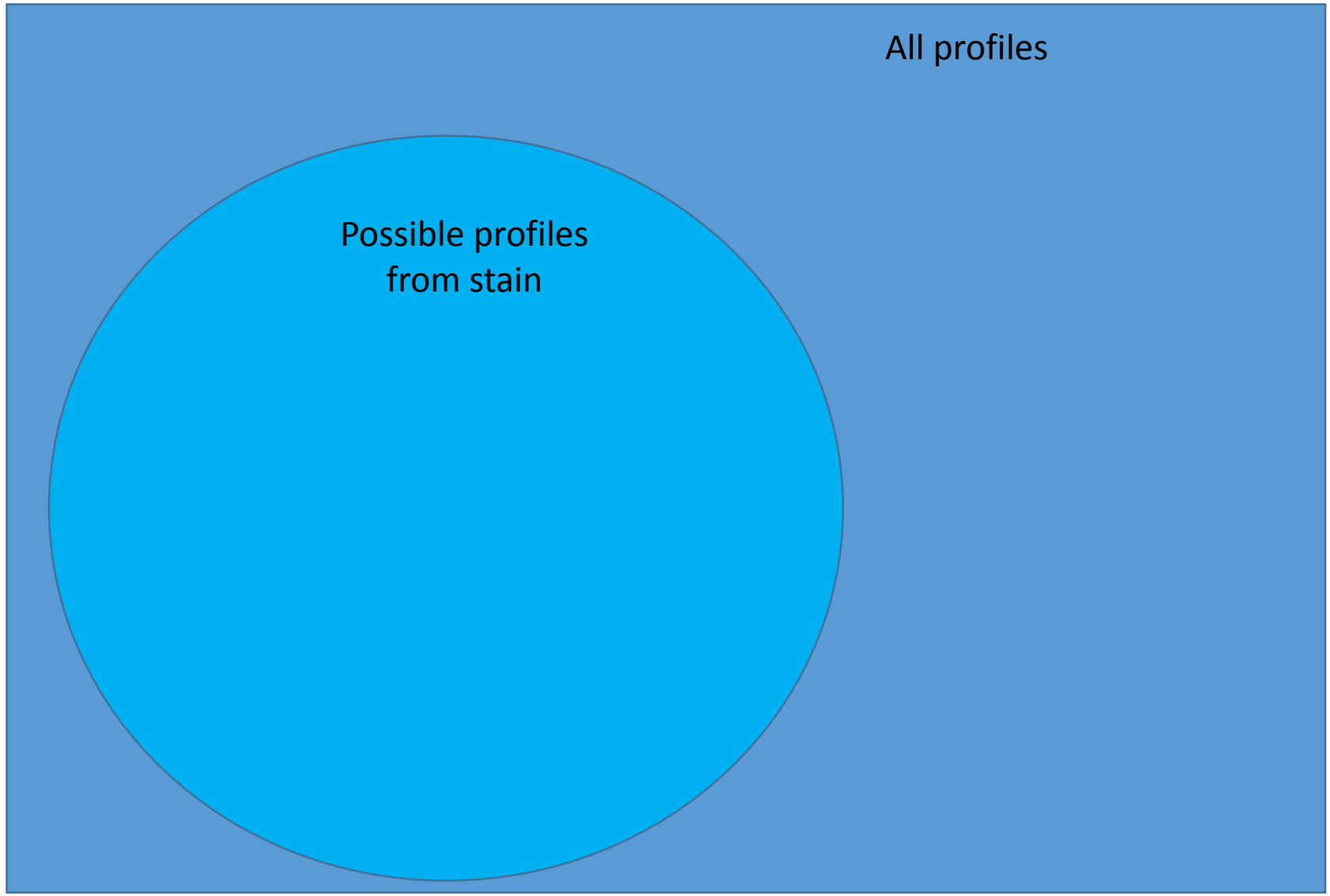
In my opinion, the problem with the LR is that it applies only to the suspect and does not give a true picture of the evidence.

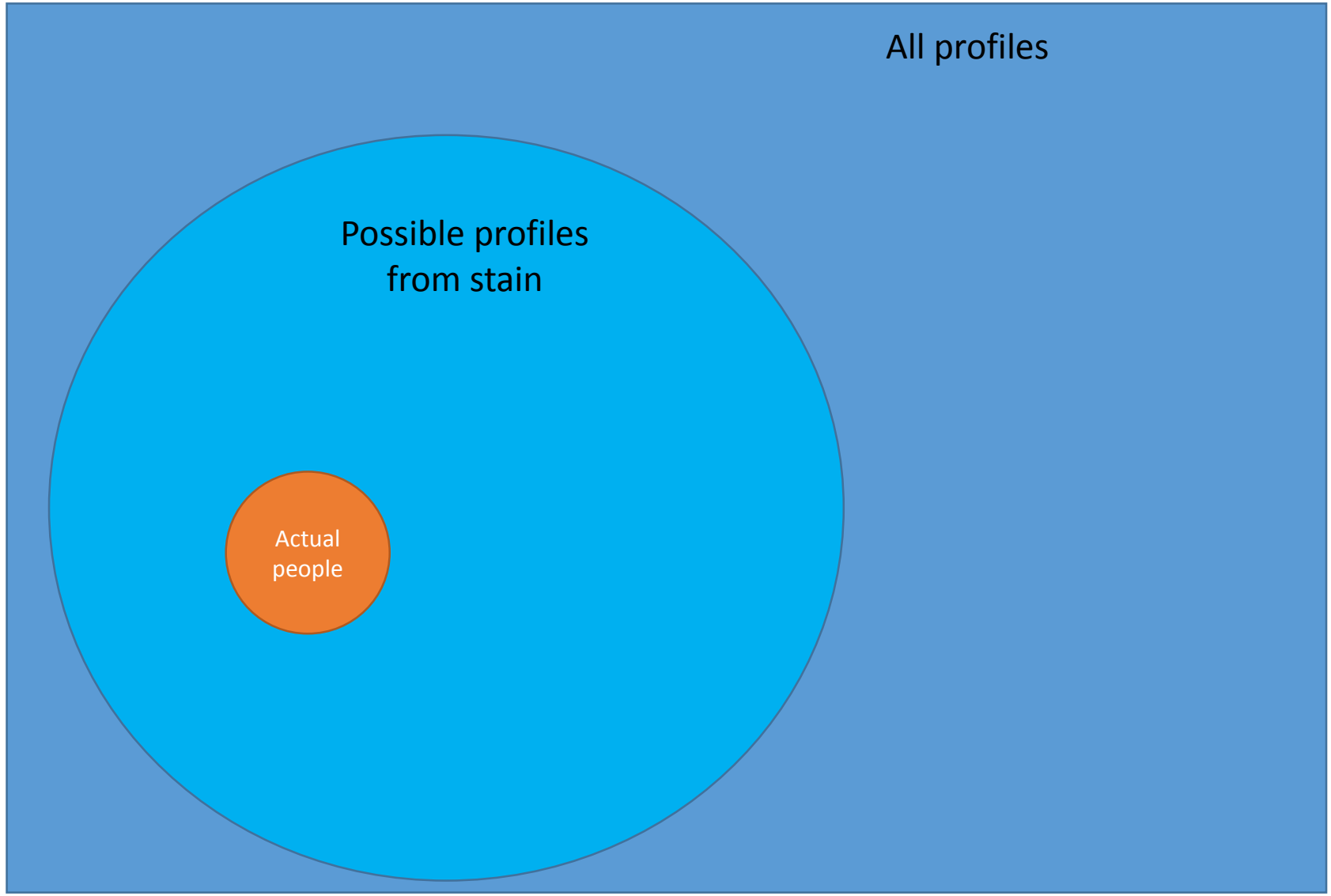
This illustrates that if the LRs of all the millions of potential genotypes from a mixture were calculated and then arranged in order of size, the suspect is unlikely to be the highest LR.

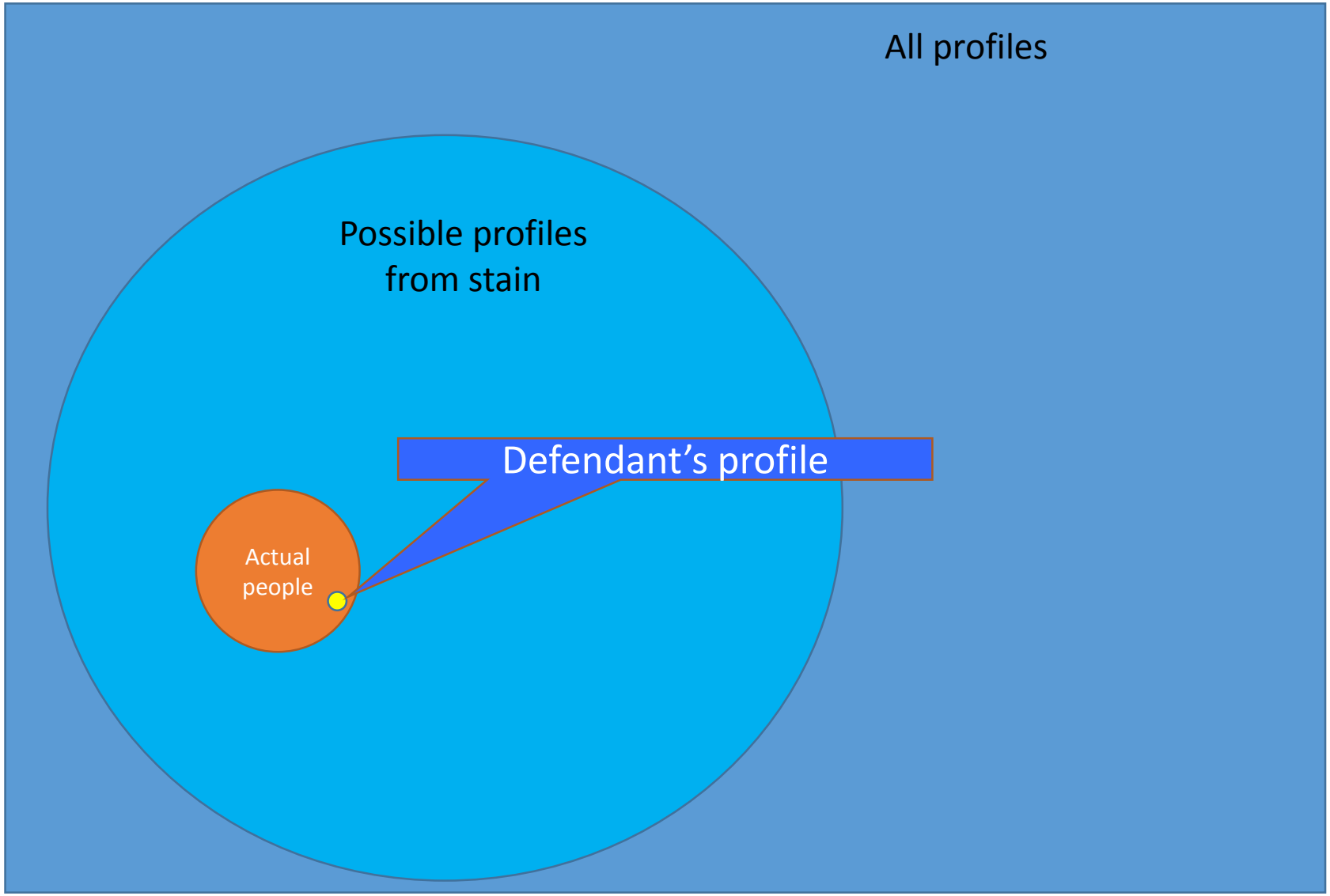
In other words, the LR provides only the weight of evidence against the specific defendant without reference to other people who would also have a LR greater that 1 (i.e. support for the prosecution hypothesis).

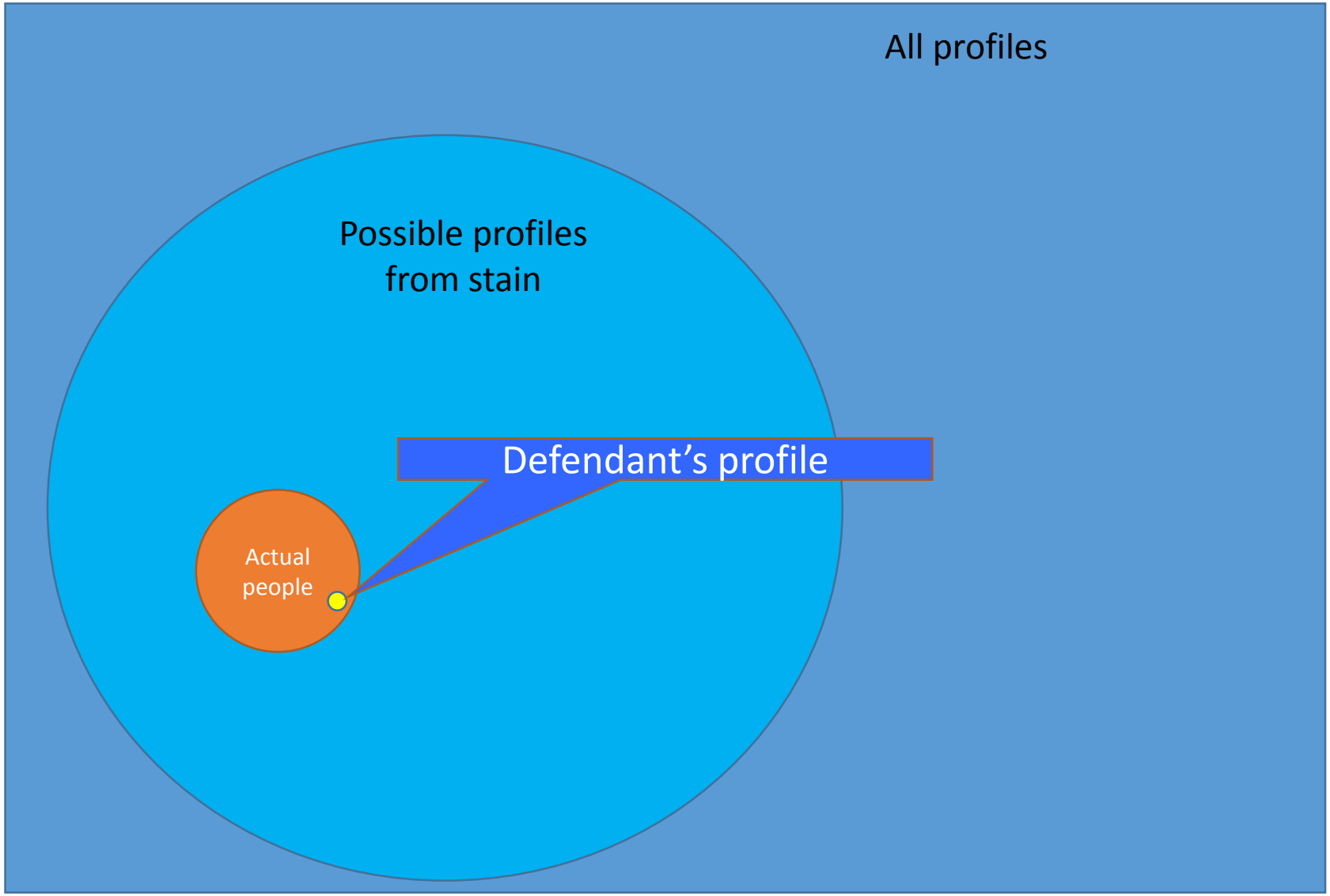
In effect, the LR is a sophisticated version of the disparaged ‘consistent with’ statement.

All profiles

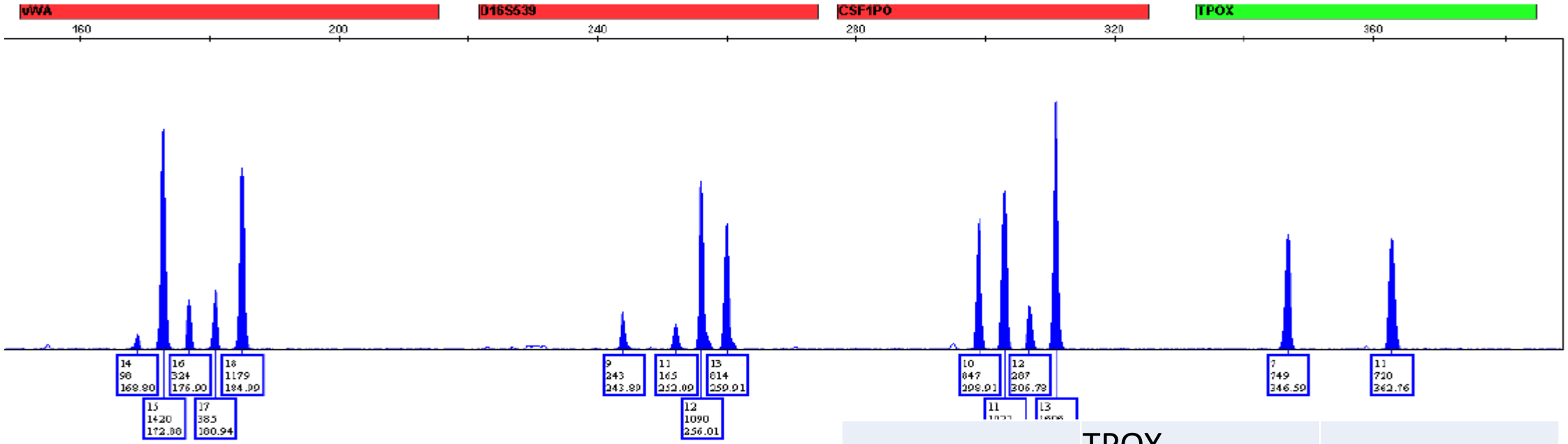

All profiles
Possible profiles
from stain


All profiles
Possible profiles
from stain
Actual
people

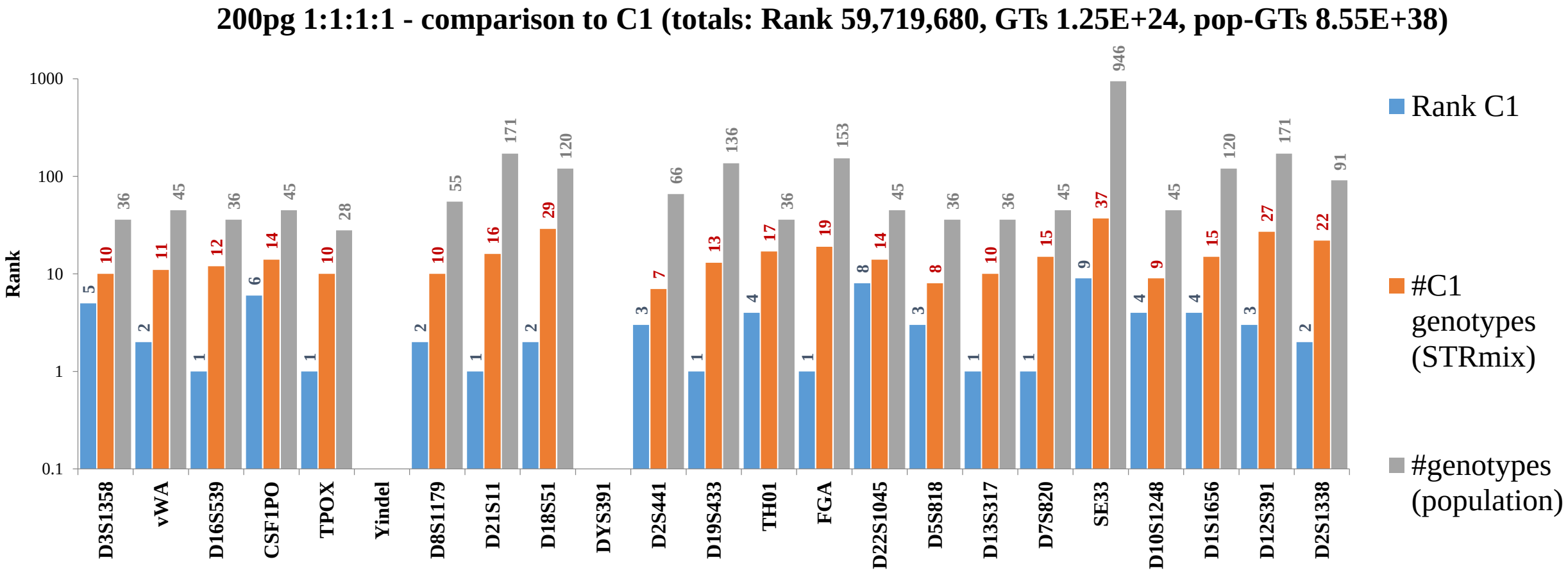

All profiles
Possible profiles
from stain
Defendant's profile
Actual
people


All profiles
Possible profiles
from stain
Defendant's profile
Actual
people

Weights and ranks



	TPOX	
C1	7,11	100.00%
C2	7,7	23.42%
	7,Q	5.85%
	7,11	37.80%
	11,11	25.73%
	11,Q	5.67%
	Q,Q	1.53%



Note that the true donor is not always rank 1

He would only be rank 1 everywhere in a very clear profile

Most genotypes do not exist

Weir, BS

In our example 8.55×10^{38} genotypes
7.5×10^{9} people

Only about 1 in 10^{29} genotypes exist

There are about 6×10^{7} genotypes above our rank

Hence potentially no actual people above our rank

Likelihood ratio

"The probability of observing this evidence is n times more likely if it arose from Mr X + an unknown person rather than two unknowns"

- Is NOT measuring the probability of Mr Lee being a contributor – many profiles will produce a high LR
- High LRs can be obtained for false propositions
- Depends on the number of contributors
- Does not test all of the possible explanations for the evidence

Why do I believe in the *LR*?

Let us start by thinking about what Jamieson wants, an exclusion probability. We cannot create this for loci with potential for drop-out… but let's pretend we could.

Let is say that is 10^{-9}

So maybe there are 7 ½ people in the world not excluded.

The crime is in Yakima, Wa.

The POI is a Yakamation (Yakamite), male, 38yo

Now what?

IF there are 7½ people then some are women, young, in China or India……

"I cannot think of anything less relevant than the population of the world." Dr Ian Evett.

LR

Enables better methods
Leads to higher discrimination of true and false donors

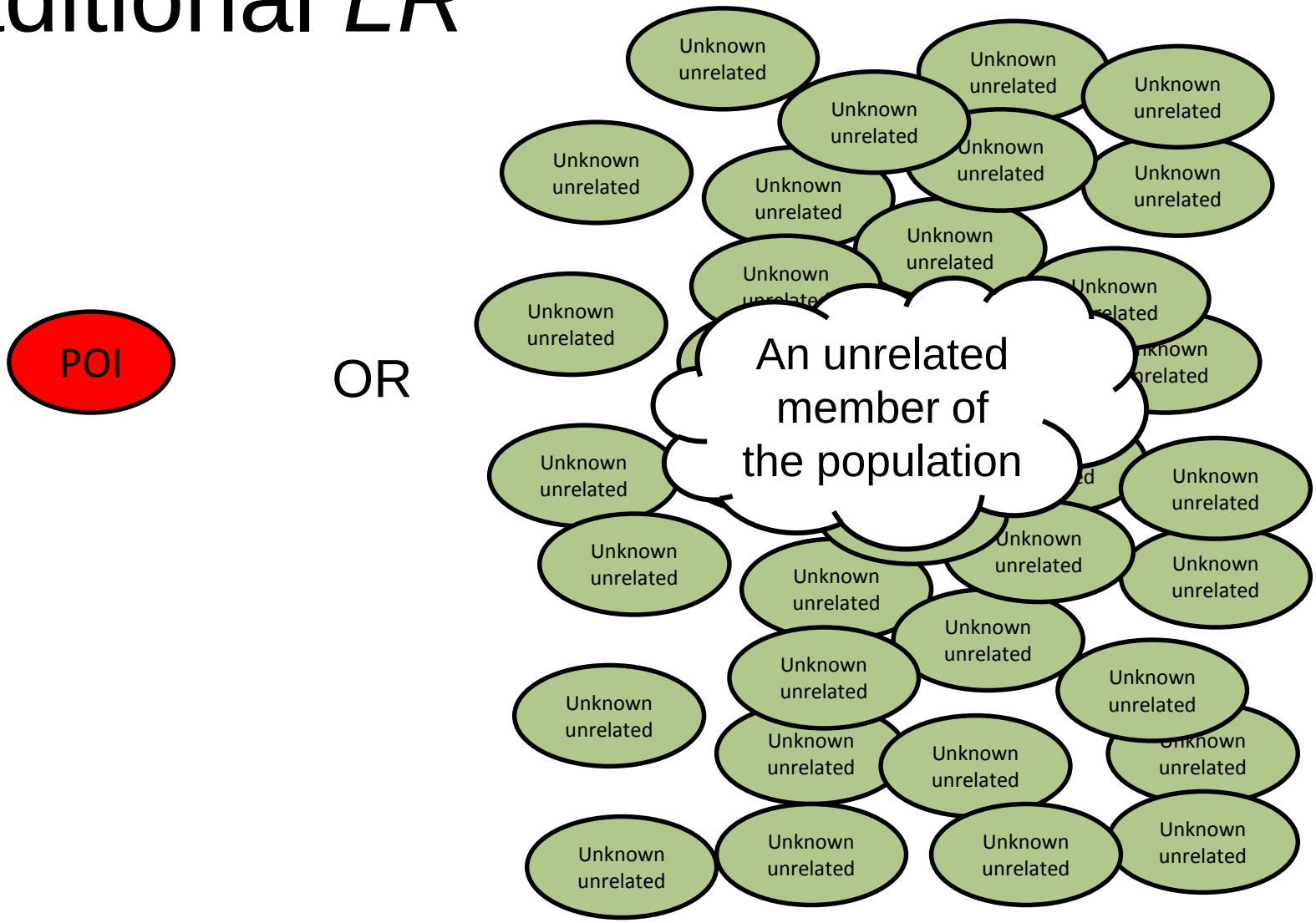
Because the LR actually relates to this POI

But, we have not connected with the judiciary

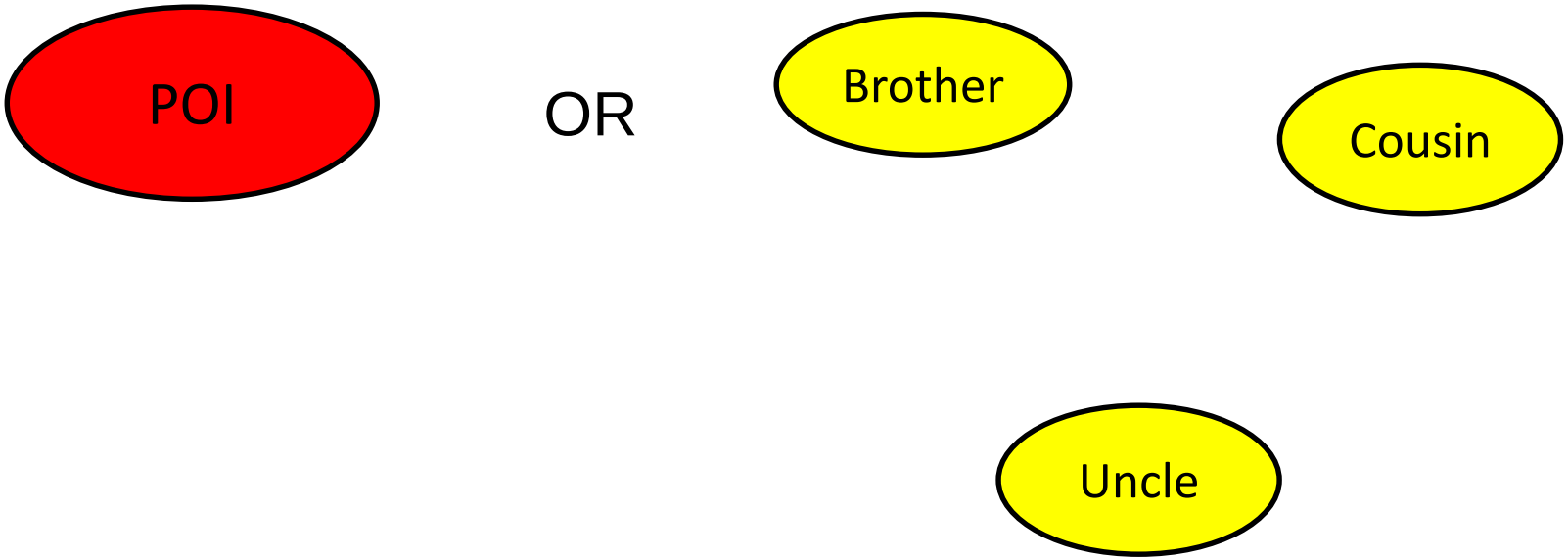
™

Calculating the LR considering relatives

Traditional *LR*



Relatives



$\Pr(E|H_p)$ The DNA originated from the POI
$\Pr(E|H_d)$ The DNA originated from a *brother* of the POI

Brother’s LR

This is easy for single source calculations where we have matching evidence and reference profiles

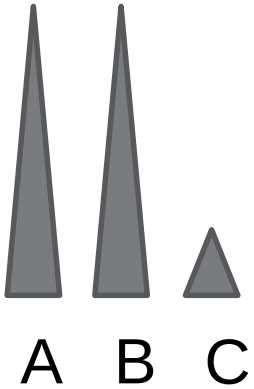
The size of the *LR* is just going to be the inverse of the probability of a brother of the POI having an identical reference profile

With mixtures it gets a little trickier… but not much

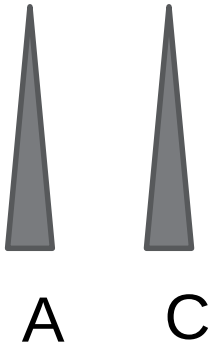
Brother’s LR example

You have the following data

V=AB



2p mixture

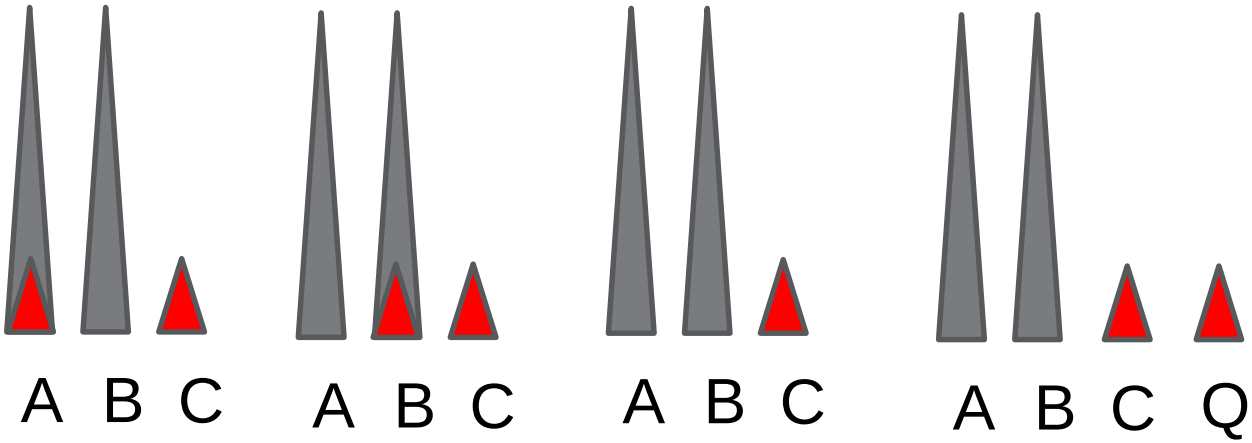


POI ref

And we consider the defence that a brother of the POI is a contributor and not the POI themselves

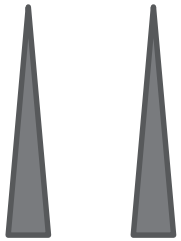
Brother’s LR example

You have the following data



2p mixture

V=AB



POI ref

And we consider the defence that a brother of the POI is a contributor and not the POI themselves

Brother's LR

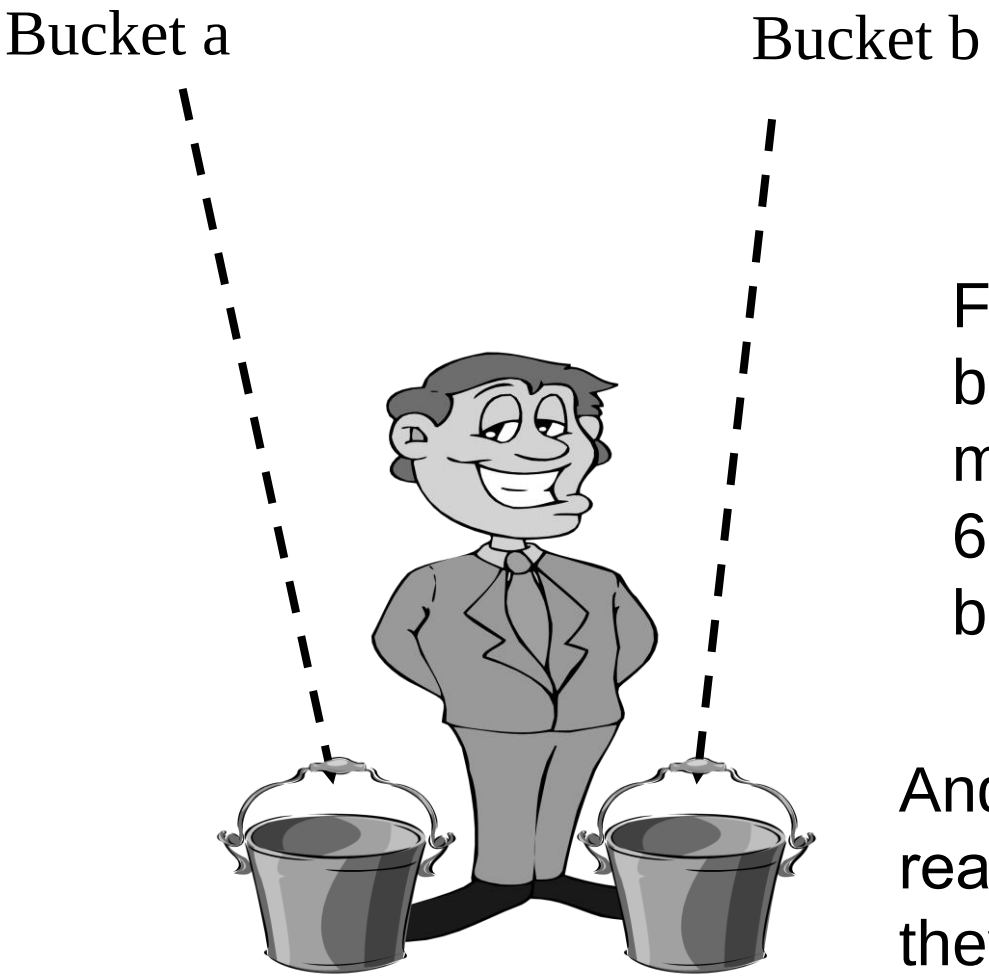
- The relationship type can be anything: parent/child/sibling/uncle/cousin/etc
- The more distant the relationship type the closer the value will become to the *LR* considering unrelated individuals
- But: STRmix can give you the relatives results in many but not all situations:
- Not Hp: $P_1 + P_2$ Hd: 2U

Duncan Taylor, Jo-Anne Bright and John Buckleton. Considering relatives when assessing the evidential strength of mixed DNA profiles. Forensic Science International: Genetics 13 (2014) 259–263

IBD

- Central to these calculations is the concept of IBD
- Two allele are IBD if they are copies of the same ancestral allele

I need to name alleles

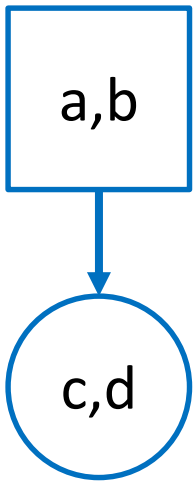


For example bucket a might have a 6 allele and bucket b a 7

And that would be really cool ‘cos then they would match me at TH01

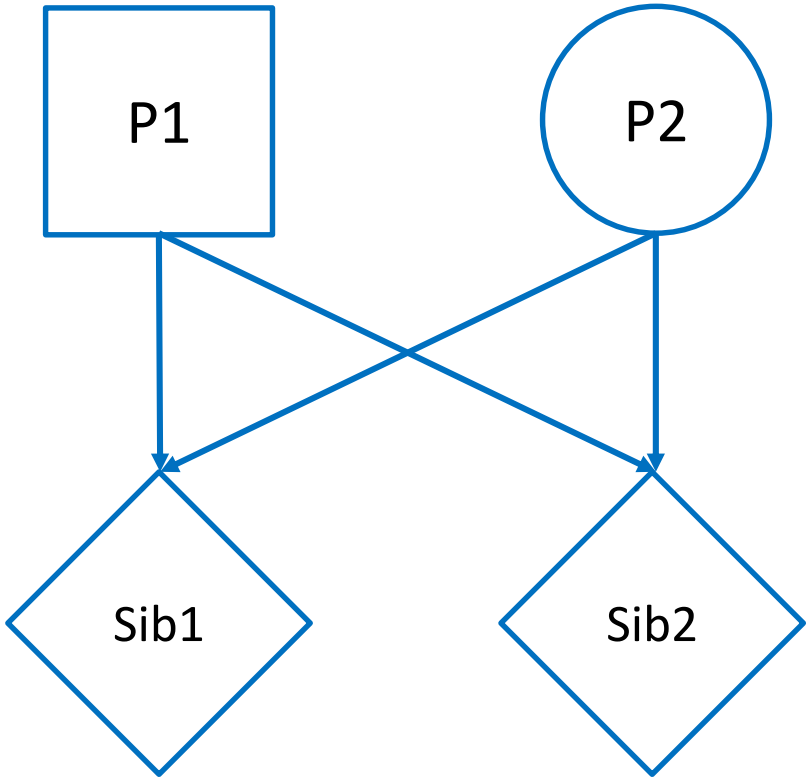
IBD parent/child

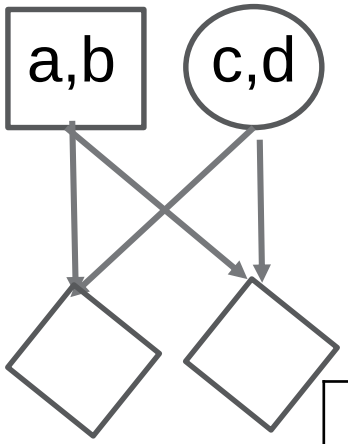
- Consider parent/child relationship
  - Mendel states that one of the alleles labelled c or d must be a copy of the a or b allele
  - If child received a c from parent, then a≡c or b≡c
  - Pr one allele IBD = 1
  - $Pr(Z_1)=1$



IBD siblings

- What about siblings?
- They can share either 2, 1 or 0 alleles IBD
- How?





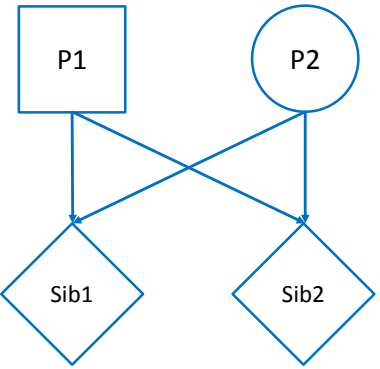
IBD siblings

Parent 1

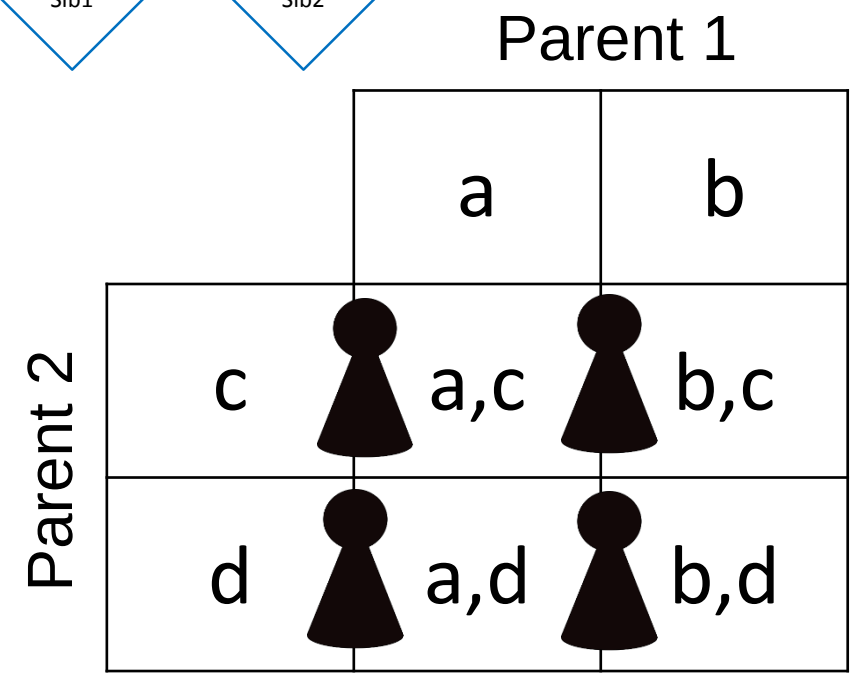
Parent 2	a	b
c	a,c	b,c
d	a,d	b,d

- Assume sib 1 is an a,c
- Sib 2 could be…

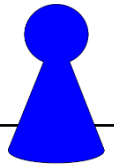
Sib1	Sib2	0	1	2
a,c				
Total				



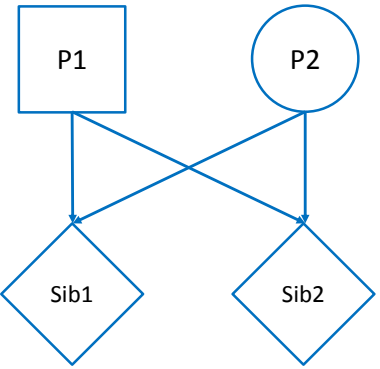
IBD siblings



- Assume sib 1 is an a,c
- Sib 2 could be…



Sib1	Sib2	Alleles IBD		
		0	1	2
a,c	a,c			
	b,c			
	a,d			
	b,d			
Total				

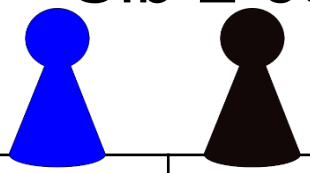


IBD siblings

Parent 1

Parent 2	a	b
c	a,c	b,c
d	a,d	b,d

- Assume sib 1 is an a,c
- Sib 2 could be…

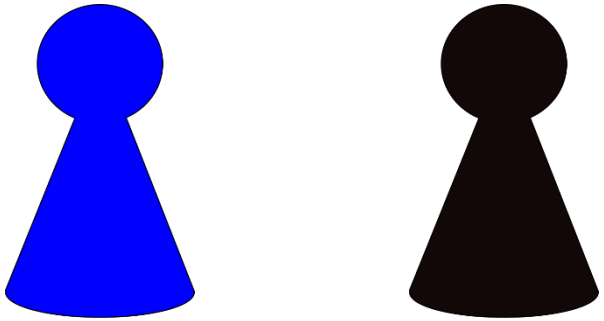


Sib1	Sib2	0	1	2
a,c	a,c			✓
	b,c		✓	
	a,d		✓	
	b,d	✓		
Total		¼	½	¼

Probability that two individuals share 0, 1, or 2 IBD alleles

Relationship	Z_0	Z_1	Z_2
Siblings	¼	½	¼
Parent/child	0	1	0
Half-siblings, Uncle/Aunt/ Grandparent /grandchild	½	½	0
Cousins	¾	¼	0

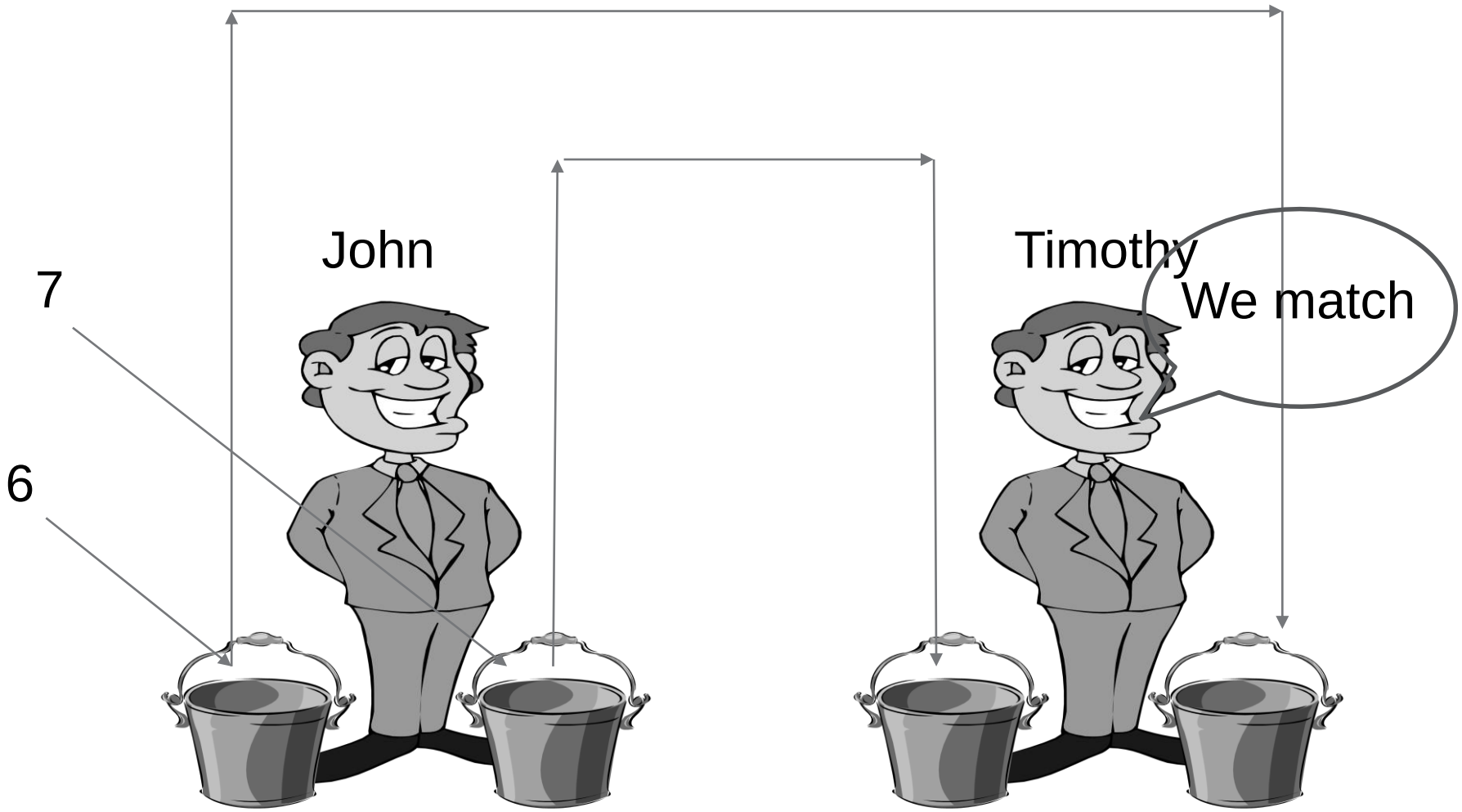
One example, matching sibs, product rule



POI=6,7 Sibling=6,7

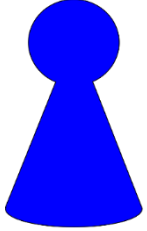
$$\Pr(sibling = ab \mid POI = ab)$$

If Z_2 is true



POI	Z state	Genotype given Z	
ab	Z_2	ab	
ab	$Z_1/2$	a?	p_b
ab	$Z_1/2$	?b	p_a
ab	Z_0	??	$2p_a p_b$

$$Z_2 + \frac{Z_1}{2} p_a + \frac{Z_1}{2} p_b + Z_0 2 p_a p_b$$



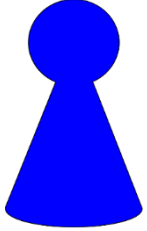
POI=aa

Sibling=aa?

$$\Pr(sibling = aa \mid POI = aa)$$

POI	Z state	Genotype given Z	
aa	Z_2	aa	
aa	$Z_1/2$	a?	p_a
aa	$Z_1/2$	?a	p_a
aa	Z_0	??	p_a^2

$$Z_2 + \frac{Z_1}{2} p_a + \frac{Z_1}{2} p_a + Z_0 p_a^2$$



POI=aa

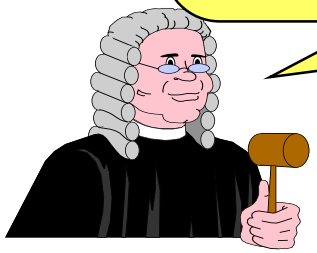
Sibling=aa?

$$\Pr(sibling = aa \mid POI = aa)$$

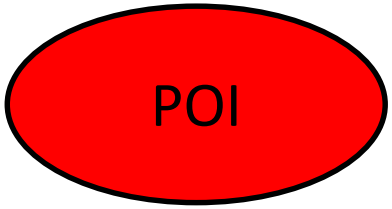
POI	Z state	Genotype given Z	
aa	Z_2		
aa	$Z_1/2$		
aa	$Z_1/2$		
aa	Z_0		

But I never know if a brother is a sensible alternative

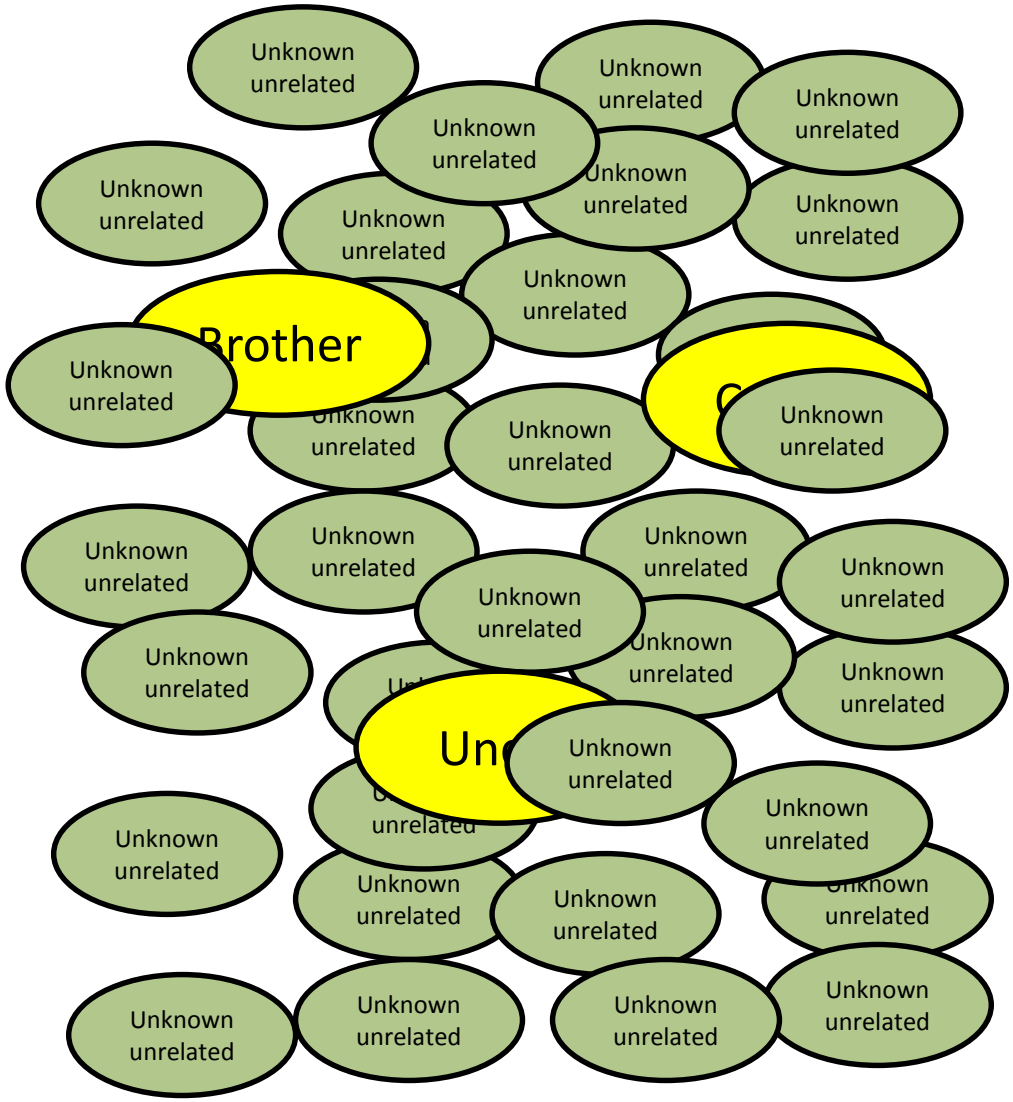
Could you model a full population with some brothers, some cousins etc?



Relatives



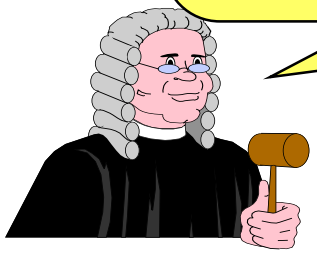
OR

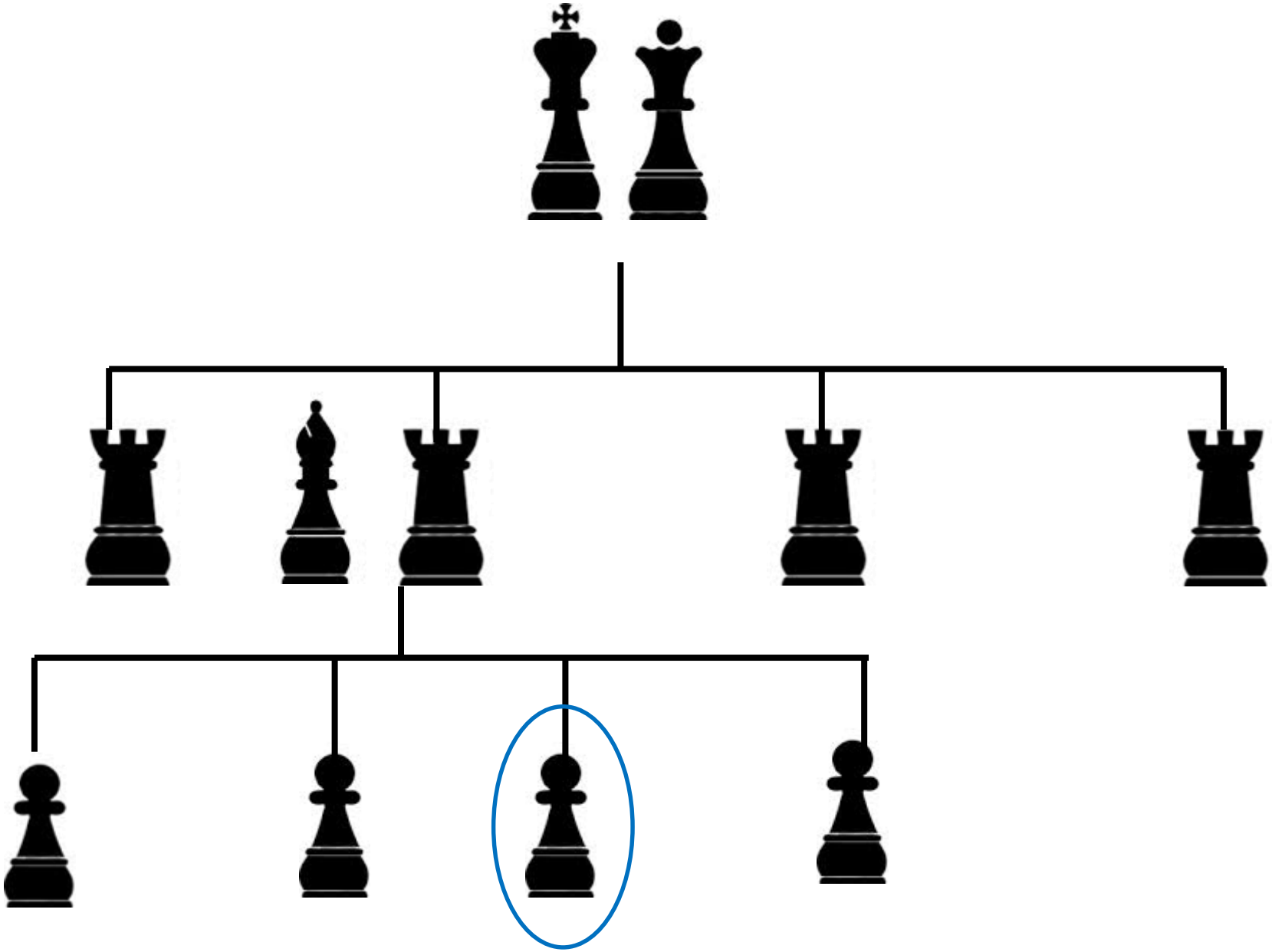


Relatives of the POI will only make up a small proportion of the population

Hmm. I'll need to think about how many brothers etc. That's not trivial.

What is a sensible top end?





	Mr Circled pawn
Parents	2
Siblings	3
Grandparents	4
Uncle/Aunt	6
Cousins	24
Unrelated	?

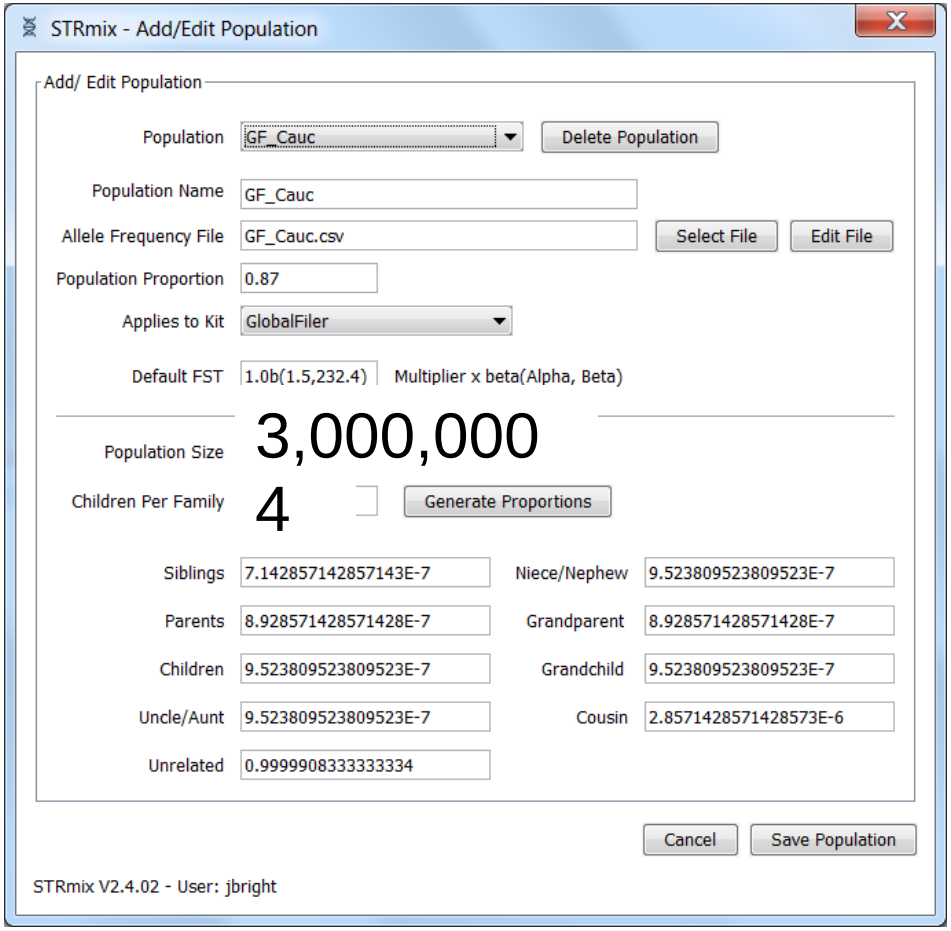
Set 4 children in population of N

Unified LR

- Takes into account sibling, parent, child, uncle/aunt, nephew/niece, grandparent/grandchild, cousins and unrelated
- This allows the use of propositions of the form
  - H_d: the donor is a member of the population

Calculation of priors

- We model these priors as a simple proportion in the population
- That population is constructed by specifying an average number of children and a population size



Pros and Cons

Con	Pro
Two more assumptions	Not really. We were already sort of implying no brothers/cousins etc. There is a safe upper side.
Another change for us/courts/prosecutors	Yes
	Simpler statement We can take out the word “unrelated”
	Probably high science option nearer what the courts want
	More important with new multiplexes

Average number of children?

https://www.cia.gov/library/publications/the-world-factbook/fields/2127.html

- Total Fertility Rates
  - …average number of children that would be born per woman if all women lived to the end of their childbearing years and bore children according to a given fertility rate at each age

Average number of children?

https://www.cia.gov/library/publications/the-world-factbook/fields/2127.html

Population	**children born/woman**
United States	1.87
World	2.42

Unified, single source Identifiler

			Number of children		
			0	2	4
population size	None	0	8.79E+18	-	-
	Washington DC	658,000	-	2.17E+12	7.23E+11
	DC Metro	6,000,000	-	1.98E+13	6.59E+12
	USA	319,000,000	-	1.05E+15	3.50E+14

Do you have to do it?

- Absolutely not
- Can we still do source attribution?
  - If you must do that then this is a stronger way.

- Balding DJ. Weight-of-evidence for forensic DNA profiles. Chichester: John Wiley and Sons; 2005

Reasonable scientific certainty

- National Commission on Forensic Science
- “to a reasonable scientific certainty”
- “In the courtroom setting, the phrase risks misleading or confusing the factfinder… It is the view of the NCFS that the scientific community should not promote or promulgate the use of this terminology.”

http://www.justice.gov/sites/default/files/ncfs/pages/attachments/2015/04/16/initial_draft_views_document_on_testimony_using_the_term_scientific_certainty.pdf

Pros

1. More holistic approach to dealing with uncertainty
   - No assumption in the proposition of “an unknown <u>unrelated</u> individual from the population”
   - Report “an unknown individual from the population”
2. Addresses one of the common lines of questioning in court:
   - Q: “What if someone who was related to the POI is the source of the DNA?”
   - A: “Our statistic already takes into account the possibility that an alternative source of DNA was someone from the population that is related to the POI”

Advanced report

SUMMARY OF LR

LR (population proportion)	GF Asian Hill.csv (0.06)	GF Caucasian ESR.csv (0.71)	GF EP ESR.csv (0.17)	GF WP ESR.csv (0.05)	Stratified
Total LR	1.06E22	1.23E23	3.34E21	**1.64E21**	1.15E23
Sibling	1.59E8	2.61E8	1.14E8	**1.35E8**	2.59E8
Parent/Child	2.42E13	1.12E14	3.02E13	**4.06E13**	1.12E14
Half sibs	2.51E16	1.28E17	1.68E16	**1.87E16**	1.28E17
Grandparent / Grandchild	2.51E16	1.28E17	1.68E16	**1.87E16**	1.28E17
Uncle or Aunt/Niece or Nephew	2.51E16	1.28E17	1.68E16	**1.87E16**	1.28E17
First Cousin	4.60E18	3.08E19	2.47E18	**2.07E18**	3.08E19
Unified	3.18E14	5.22E14	2.27E14	**2.70E14**	5.18E14

Court questions?

- The default LR is for unrelated
- That actually optimises the evidence for the prosecution

	HPD	MCMC	α	sides	unified
DC DFS	Y	Y	0.99	1	Maybe
Cal DOJ current	N	N			N
Cal DOJ planned	Y	Y	0.99	1	Maybe
USACIL	Y	Y	0.99	1	N
FBI					
SDPD (5p)					
SDPD (ss-4p)					
NYC OCME	Y	Y	0.99	1	Y
John Buckleton	Y	Y	0.99	1	Y
OSP	Y	Y	0.99	1	in file not report
SDSO	Y	Y	0.99		in file not report
Sacremento County Crime Lab	Y	Y	0.99		N
TriCounty	Y	Y	0.99	1	possible
Erie (NY)					Y
Scottsdale PD	Y	Y	0.99	1	N
Idaho SP	Y	Y	0.99	1	in file not report

™

End