Netherlands Forensic Institute
Ministry of Justice

Forensic DNA analysis

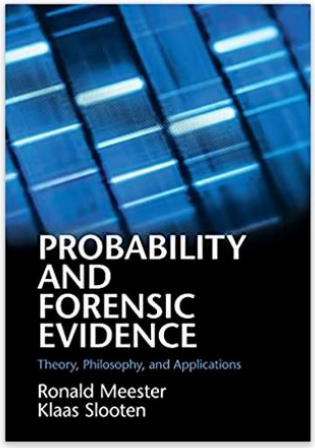
Symposium "Solving Crime in No Time"

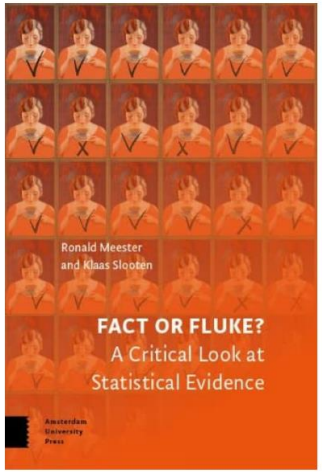
Delft, 18 October 2022

Prof dr K Slooten
NFI/VU

Work at NFI/VU

- Development and implementation of statistical models for the evaluation of DNA profiles
- Kinship analysis (disaster victim identification, e.g. MH17)
- Training of forensic DNA experts
- Casework (criminal cases, civil cases for police)
- At the VU: collaboration with Ronald Meester, courses "Forensic Probability and Statistics" (MasterMath, with Marjan Sjerps) and "a Likelihood approach to Evidence" (Honours class)





What is forensic mathematics?

Forensic: pertaining to legal issues

Goal of NFI is to give the evidential value of findings within the context of a criminal case.

Provide mathematical assistance/modeling/explanations.

Usually, probability/statistics.

Some aspects of forensics lend themselves to mathematical modeling, e.g. DNA profiling, inheritance, population genetics

For other aspects this is not so obvious.

Forensic statistics and the ground truth

In many areas one is directly interested in the ground truth: e.g. effectiveness of a medical treatment: does it work or not?

According to criminal law, the fact finder (judge/jury) must convict on the basis of the available *evidence*; it becomes critically important to know precisely how strong the evidence is.

A forensic lab therefore is not really concerned with the ground truth (which events happened) but in the first place with *evidence*: determining how well certain data can help to distinguish between various possible ground truths.

Forensic statistics: individuals vs. populations

In many areas of statistics error rates are important, or more generally the effect in a population as a whole.

In forensics, case specific assessments; judicial system makes a decision in a individual case, not for a whole group.

The evidence in any particular case must be interpreted within the context of that particular case.

Science can go on in principle forever; legal procedures must finish in a reasonable time frame.

What is evidence?

If there are various hypotheses (about a crime, a trace,..) then data E are evidence if the probabilities of the hypotheses change due to E, i.e., if there are H such that $P(H|E) \neq P(H)$.

What is meant by $P(H)$, e.g. the probability that a suspect committed a certain crime?

Interpretation as degree of belief is best fitting here; this makes the notion of probability personal and epistemic in nature: an observer defines their probabilities based on their knowledge of the world, not based on the world.

The idea is now that a forensic expert's probability assessment can also be used by the fact finder.

Bayes rule

$$\frac{P(H_1|E)}{P(H_2|E)} = \frac{P(E|H_1)}{P(E|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

"posterior odds = likelihood ratio x prior odds"

The idea is that the hypotheses are those from prosecution and defence; a forensic lab computes the LR and a decision maker determines their posterior odds. Separation between value of evidence (LR) and inference about what happened (posterior odds) thus realized.

It's actually not all simple now; we'll pretend it is today.

Remaining slides: mixed traces

- DNA profiles of crime related traces must be compared with reference profiles of (possible) suspects
- However, many traces contain DNA of several persons (e.g. clothing): mixed traces
- Analysis of mixed traces quickly runs into computational difficulties
- Labs cannot evaluate traces that contain DNA of too many persons (say, 4 or more)
- Came up with a (quick and dirty type) solution

Goal: compute Likelihood Ratio

- H_p: the PoI (person of interest) *S* with profile *g* has contributed DNA to the trace profile *M*

Versus

- H_d: PoI *S* did not contribute.

Likelihood ratio:

$$LR(M, g \mid I) = \frac{P(M \mid H_p, g, I)}{P(M \mid H_d, g, I)}$$

But how do we do that? In general, it is clear the PoI cannot have been the only contributor. There must be other, unknown contributors in addition to (and for H_d also instead of) the PoI

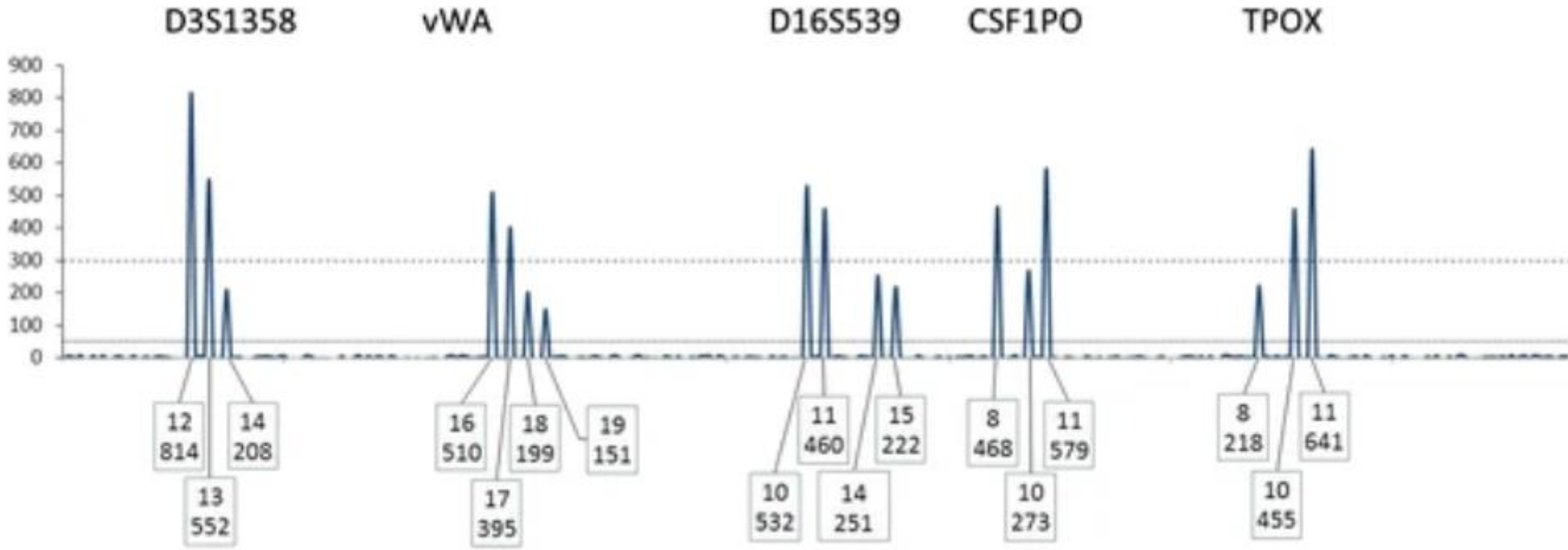
DNA profiling

- Inspect about 20 positions (“loci”) on human DNA
- On these loci there exist many DNA-variants (“alleles”) in the population: allows for distinguishing persons from each other
- Every person has every (non-gender) chromosome twice (one copy from each parent). Thus, if *n* persons are considered together, they have up to *2n* different alleles (fewer than *2n* if they share some alleles)

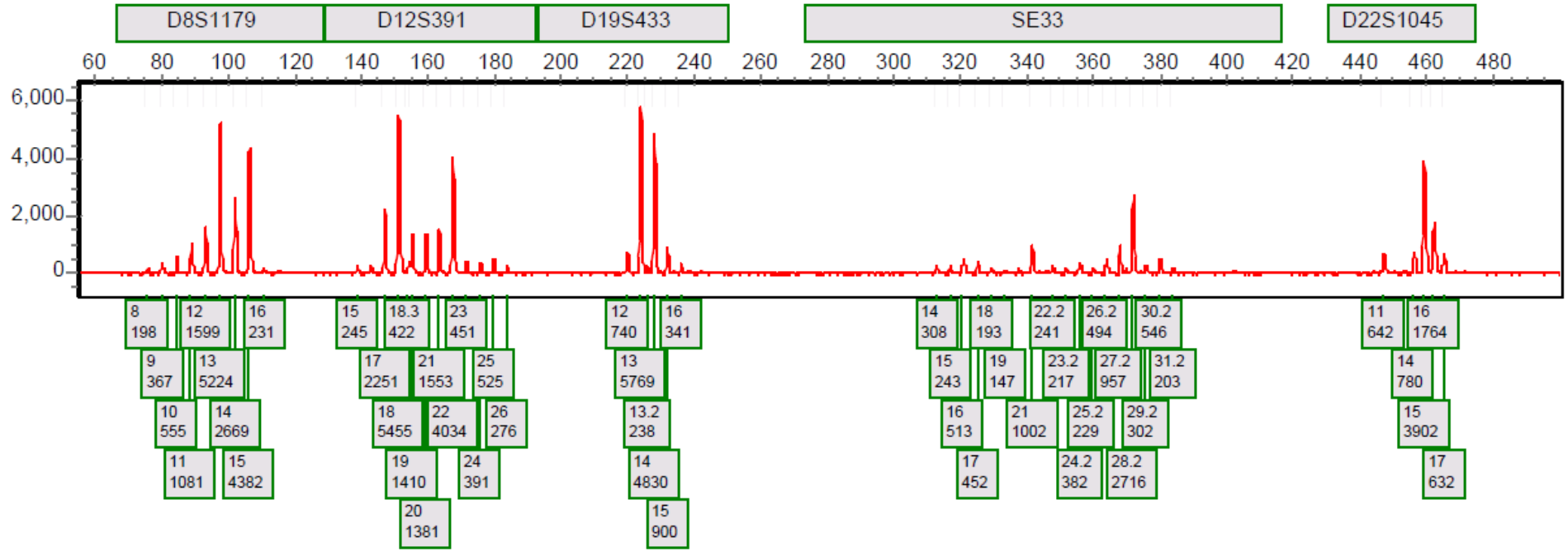
DNA profiling

- PCR reaction to copy the DNA: stochastic process, imperfect
- After PCR, for each allele, a quantification is obtained
- This can be plotted as a "peak profile": the alleles on the x-axis, the peaks on the y-axis
- In addition to the peaks coming from alleles of persons in the sample, there are other mechanisms possibly leading to peaks included in the profile, e.g. stutter (erroneous copies as a result of the PCR process) or sporadic contamination (drop-in)
- For small contributions, not all alleles will always be detected ("dropout")
- In general a trace profile contains incomplete information on an unknown number of persons, and some noise not coming from persons at all.

Tractable trace: 2 persons



Not tractable: large number (unknown)



Discrete model

Only considers presence/absence of alleles. Take into account that,

- for contributor i, there is a probability d_i for each of their alleles not to contribute to the registered profile.
- Alleles may drop-in sporadically, Poisson process with parameter c determines how many; then random sample.

Suppose contributor i has $n_{i,a}$ copies of allele a.

Then the probability that an allele a is registered in the profile is

$$1 - e^{-cp_a} \prod_{i=1}^{n} {d_i}^{n_{i,a}}$$

This means: an allele is registered unless (i) it drops out for everyone who has it, and (ii) it does not drop in.

The discrete model

This allows to define $P_\theta(M|g_1,..,g_n)$, with θ the vector of d_i.

Then choose $f(\theta)$ a uniform density and integrate.

$$P(M \mid H, g, I) = \int_\theta \sum_{g_1,\ldots,g_n} P_\theta(M \mid g_1, \ldots, g_n) P(g_1, \ldots, g_n \mid H, g, I) f(\theta) d\theta$$

This gives:

A computationally cheap model well suited for traces with few contributors.

Not enough power for many contributors (1, 2 fine, 3 problematic, 4 and more hardly ever really useful).

Continuous models

Take into account peak heights; requires more modeling.

E.g. consider peaks come from Gamma-distributions. Two parameters for Gamma distribution, but also for mixture proportions, degradation, and others.

Need to either integrate over the parameters or find MLE values for them.

Better model but more computationally expensive. Runs into trouble with >4 (unknown) contributors

The number of contributors (NoC)

- Lots of research about how to estimate the NoC and how bad it is if we don't get it right.

Estimating the **number** of **contributors** to forensic **DNA mixtures**: does maximum likelihood perform better than maximum allele count?
H Haned, L Pene, JR Lobry, AB Dufour... - Journal of forensic ..
Determining the **number** of **contributors** to a forensic **DNA mix** count is a common practice in many forensic laboratories. In this method to a maximum likelihood estimator, previously proposed
☆ 99 Geciteerd door 64 Verwante artikelen Alle 9 versie

Towards understanding the effect of uncertainty in the **number** of **contributors** to **DNA** stains
JS Buckleton, JM Curran, P Gill - Forensic Science International: Genetics, 2007 - Elsevier
... The effect of masking on the **number** of alleles presenting at a multiplex was investigated by ... We ocus and varying **numbers** of **contributors** ... and lleles of different **contributors** (also termed ...
len Alle 8 versies

[HTML] Estimating the **number** of **contributors** to a **DNA** profile
T Egeland, I Dalen, PF Mostad - International journal of legal medicine, 2003 - Springer
... It is possible to obtain a posterior distributio be seen what **numbers** will be available and if on conventional markers (see Evett and Weir 1

Bounding the **number** of **contributors** to mixed **DNA** stains
SL Lauritzen, J Mortera - Forensic science international, 2002 - Elsevier
argument leading to this is inexact and the bound is not universally true for small **numbers** of

Inference about the **number** of **contributors** to a **DNA mixture**: comparative analyses of a Bayesian network approach and the maximum allele count method
A Biedermann, S Bozza, K Konis, F Taroni - Forensic Science International ..., 2012 - Elsevier
In the forensic examination of **DNA mixtures**, th **contributors** (N) presents a topic of ongoing int issues of bias, in particular when assessments c
☆ 99 Geciteerd door 38 Verwante artikele

the **number** of **contributors** for a ... [1] CH tains when the **number** of donors ...
ersies »

NOCIt: A computational method to infer the **number** of **contributors** to **DNA** samples analyzed by STR genotyping
H Swaminathan, CM Grgicak, M Medard ... - Forensic Science ..., 2015 - Elsevier

[PDF] Estimating the **number** of **contributors** to two-, three-, and four-person **mixtures** containing **DNA** in high template and low template amounts
J Perez, AA Mitchell, N Ducasse, J Tamariz... - Croatian medical ..., 2
Methods Seven hundred and twenty-eight purposeful two-, three-, and composed of 85 individuals of various ethnicities with template amoun 500 pg were examined. The **number** of alleles labeled at each locus a

h one **contributor** in testing set 1 ... Supplementary in testing set 1 ...

The effect of varying the **number** of **contributors** on likelihood ratios for **DNA mixtures**
CCG Benschop, H Haned, L Jeurissen, PD Gill... - Forensic Science ..., 2015 - Elsevier
Abstract Interpretation of **DNA mixtures** with three or more **contributors**, defined here as

But is the NoC really so important?

It depends on the model that is used.

- Initial models supposed equal contribution of all contributors. Then indeed the NoC matters: the more there are, the less DNA there is per person.

- Newer models suppose equal opportunity for contribution: then the NoC matters less, because contributions may be estimated at zero. Working with too many is then, conceptually, not such a big problem: the superfluous ones will not do much in the model

- But the computation time explodes since it's exponential in NoC.
- More than four unknown contributors is generally untractable

(NFI software: NOC=1 or 2: <1 s; NOC=3: ~minutes; NOC=4: ~hours)

Idea: Top-down approach

Go after the contributors sequentially, starting with the most prominently present.

- Is there evidence that the person of interest is the most prominent contributor? If yes, we are done. If not,
- Is there evidence that the person of interest is among the two most prominent contributors? If yes, we are done. If not,
- Is there evidence that the person of interest is among the three most prominent contributors? If yes, we are done. If not,
- etc.

Advantage

Computational:

If we can compute the likelihood ratios for being in the top-k of donors as fast as traditionally for a k-person trace, then we save a lot of time.

Practical:

There is no limit on the total number of contributors that the trace can have; only a limit on how many of them we can identify.

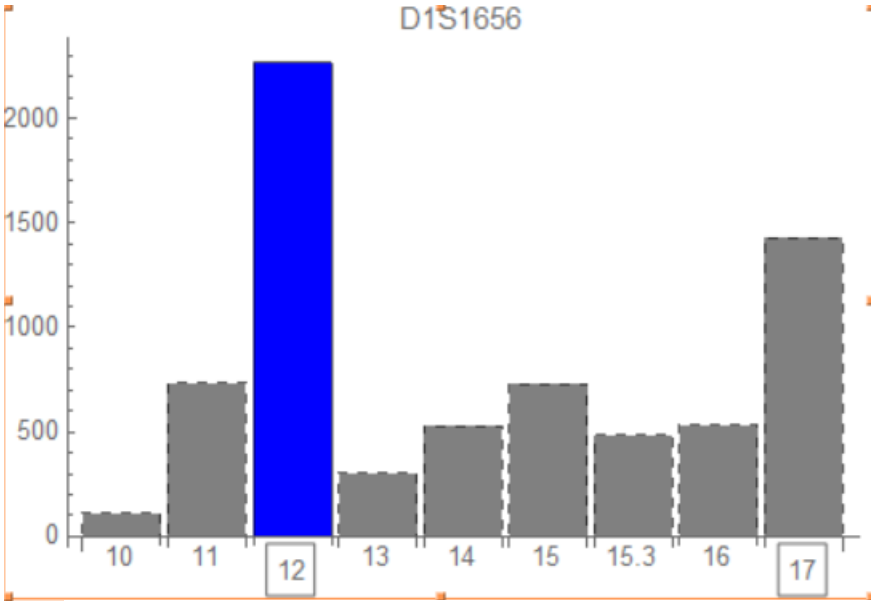
A top-down algorithm: create sub-profiles

Principle: more DNA contribution leads to higher peaks

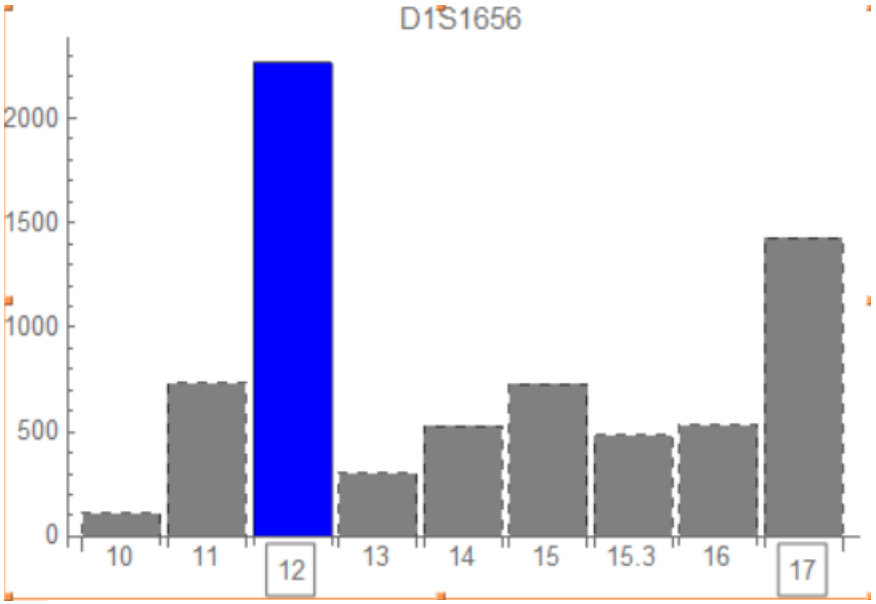
Target the more prominent contributors by making a subprofile that contains the largest peaks, until a prescribed relative proportion of the total sum of peak heights is taken in.

subprofile targeting at least 10 %

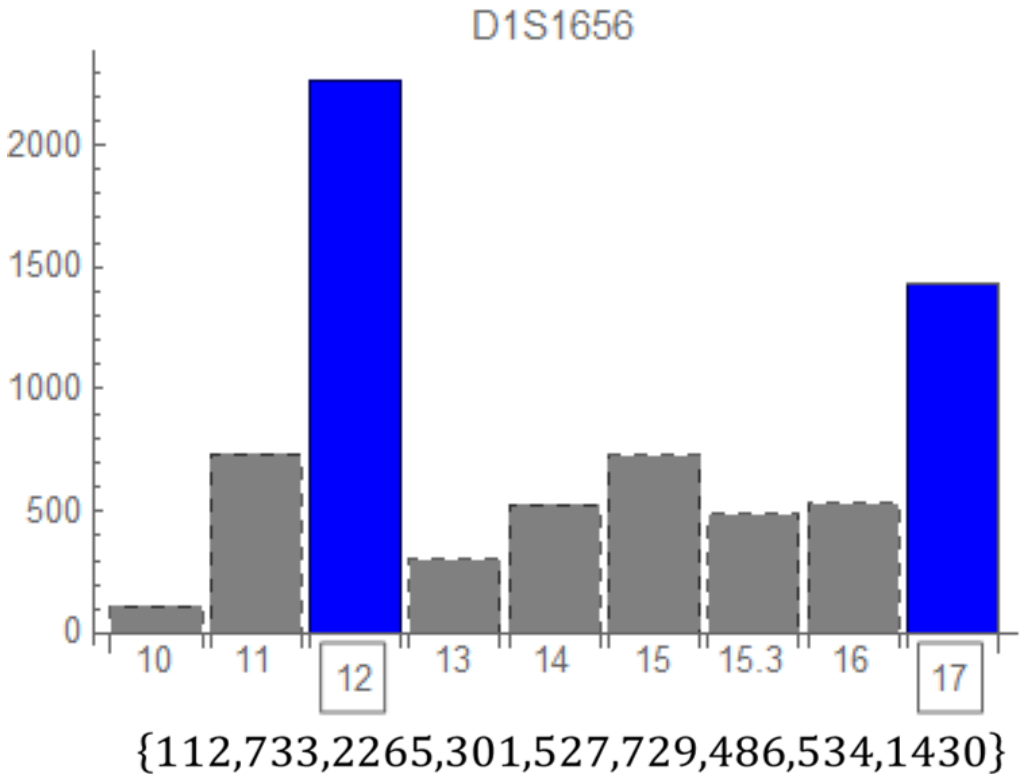
Peak heights: sum is 7117, largest peak (2265) already >10% of total



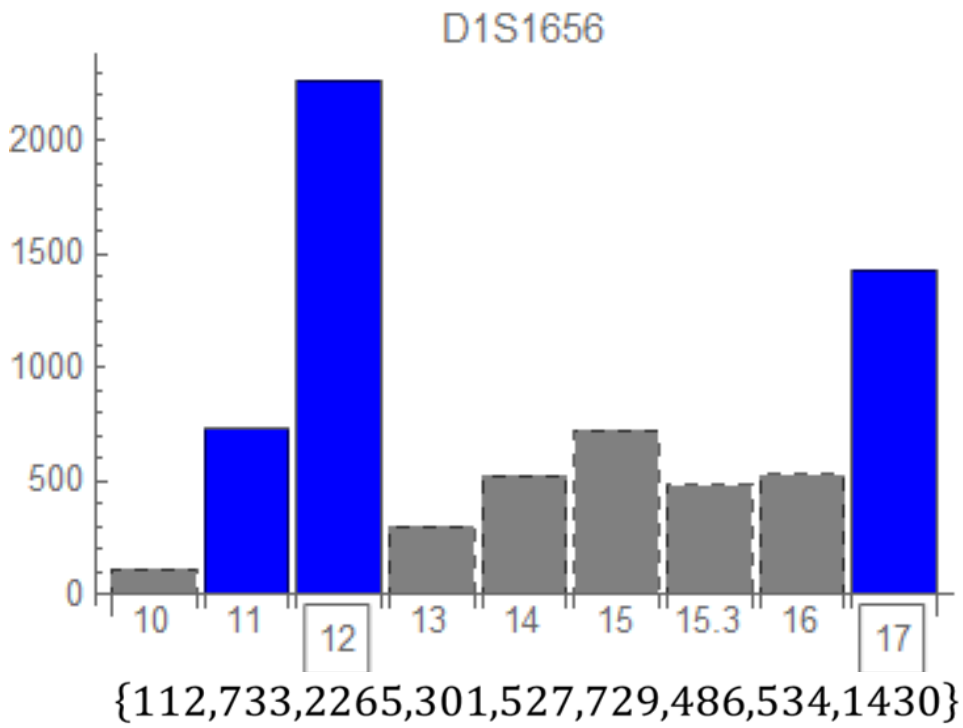
20% and 30%: still only largest peak needed



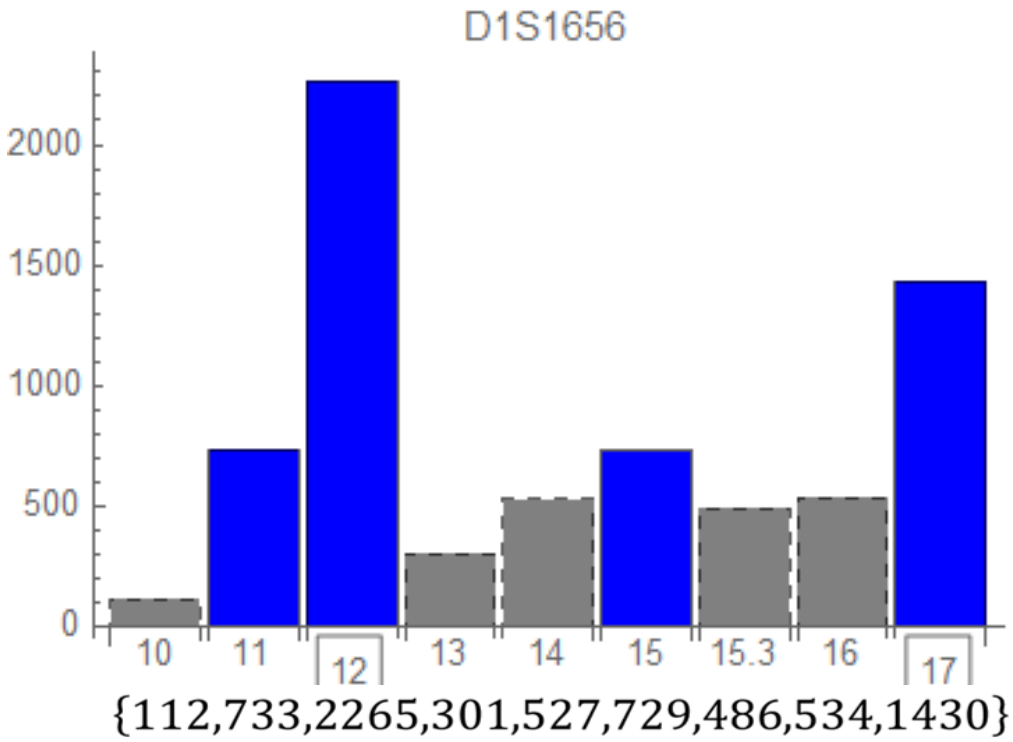
40% and 50%: second peak also needed



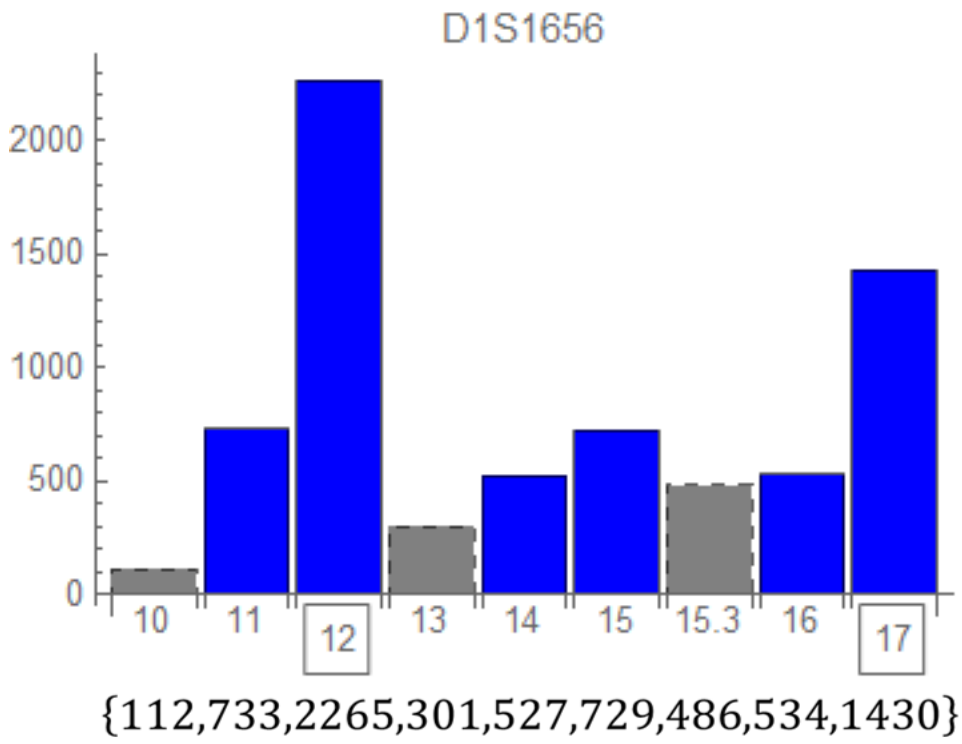
60%


D1S1656
2000
1500
1000
500
0
10
11
12
13
14
15
15.3
16
17
{112,733,2265,301,527,729,486,534,1430}

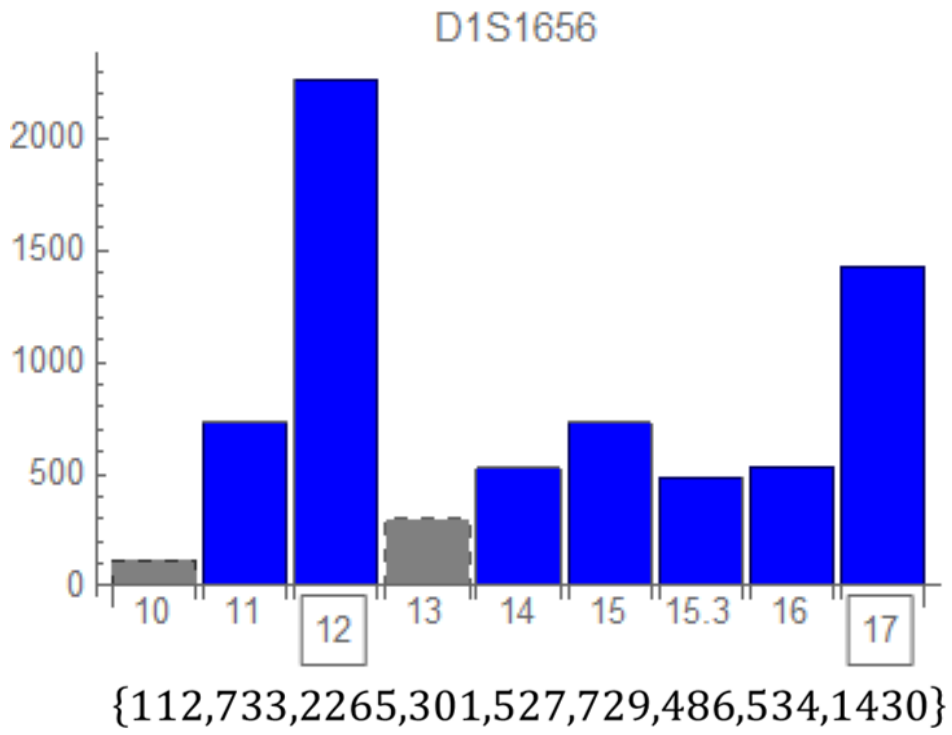
70%


D1S1656
2000
1500
1000
500
0
10
11
12
13
14
15
15.3
16
17
{112,733,2265,301,527,729,486,534,1430}

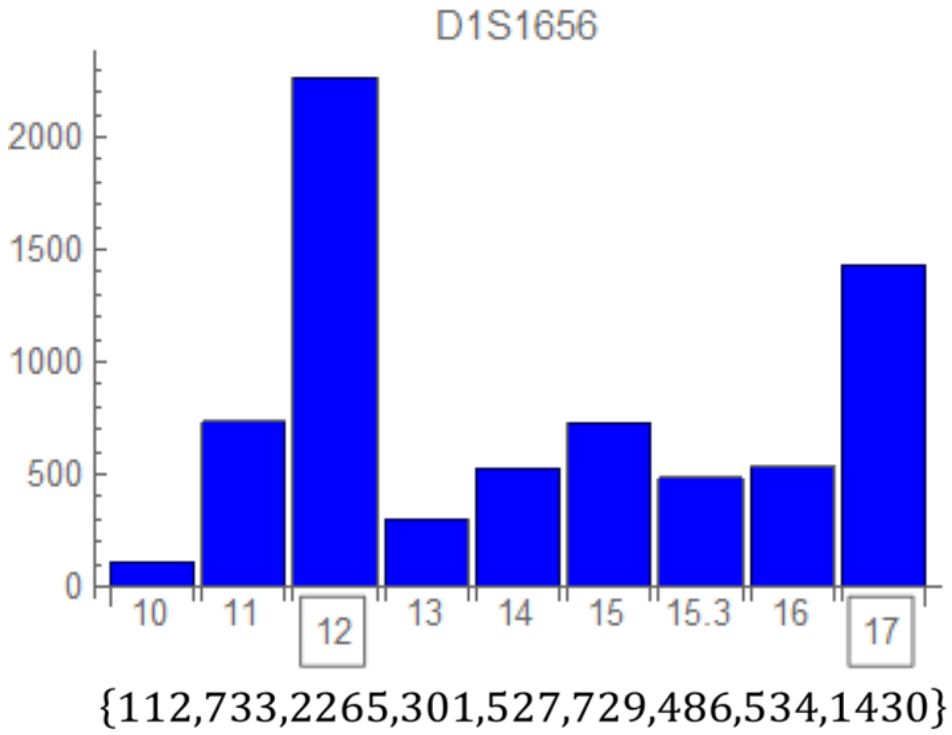
80%


D1S1656
2000
1500
1000
500
0
10
11
12
13
14
15
15.3
16
17
{112,733,2265,301,527,729,486,534,1430}

90%


D1S1656
2000
1500
1000
500
0
10
11
12
13
14
15
15.3
16
17
{112,733,2265,301,527,729,486,534,1430}

100%: full profile



Top-down LR: quick-and-dirty algorithm

- We obtain a series of subprofiles, each contained in the next one.
- Calculate the LR on all subprofiles using the discrete method (fast).
- Take the *largest* LR encountered on a subprofile as top-down LR.

Why would this work?

- In any subprofile, we may not have all alleles from the targeted contributors and we may have some from non-targeted people
- But this also holds for the full trace where we call that dropout and drop-in

- Imagine we sequentially 'reveal' the profile by lowering the detection threshold. If a PoI is the *k*-th contributor, we expect the evidence to be strongest if we have revealed the alleles of the first *k* contributors.
- If we continue revealing more alleles, we add more computational complexity but the evidence doesn't really change;

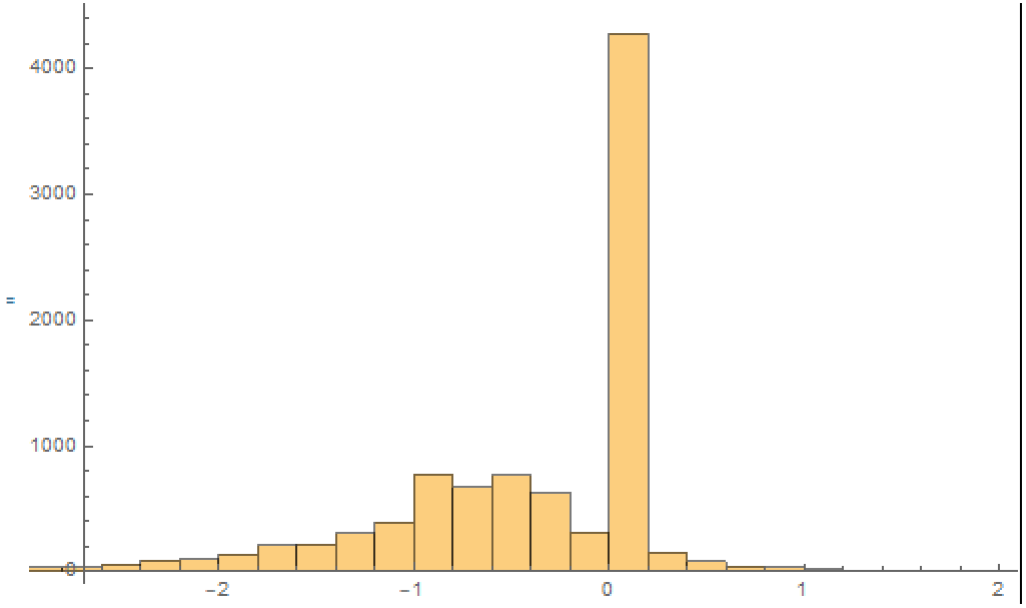
Validation: lab traces

Set of profiles with 2,3,4,5 contributors.

Can compare top-down LR with the LR of continuous model (quick'n'dirty versus accurate)

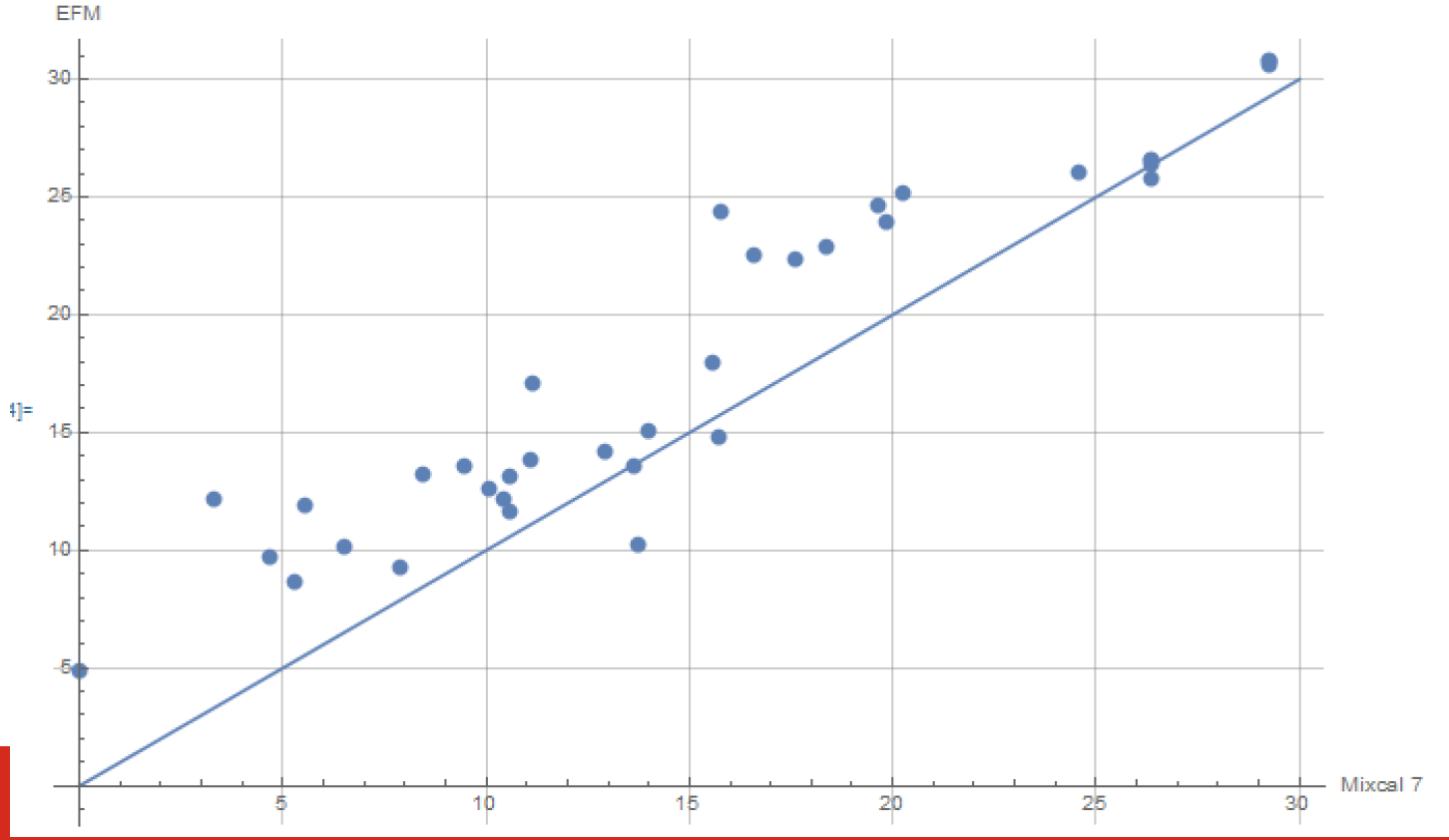
Non-contributors

- 9.385 combinations
- Largest top-down LR was 82
- Most top-down LR's are at most 1. (93%).



4 person traces, top-2 contributors

- Comparison of top down (x-axis) with continuous method (y-axis)
- Top-down gives slightly less evidence, not of much practical importance
- However, top-down in seconds, full analysis in hours.



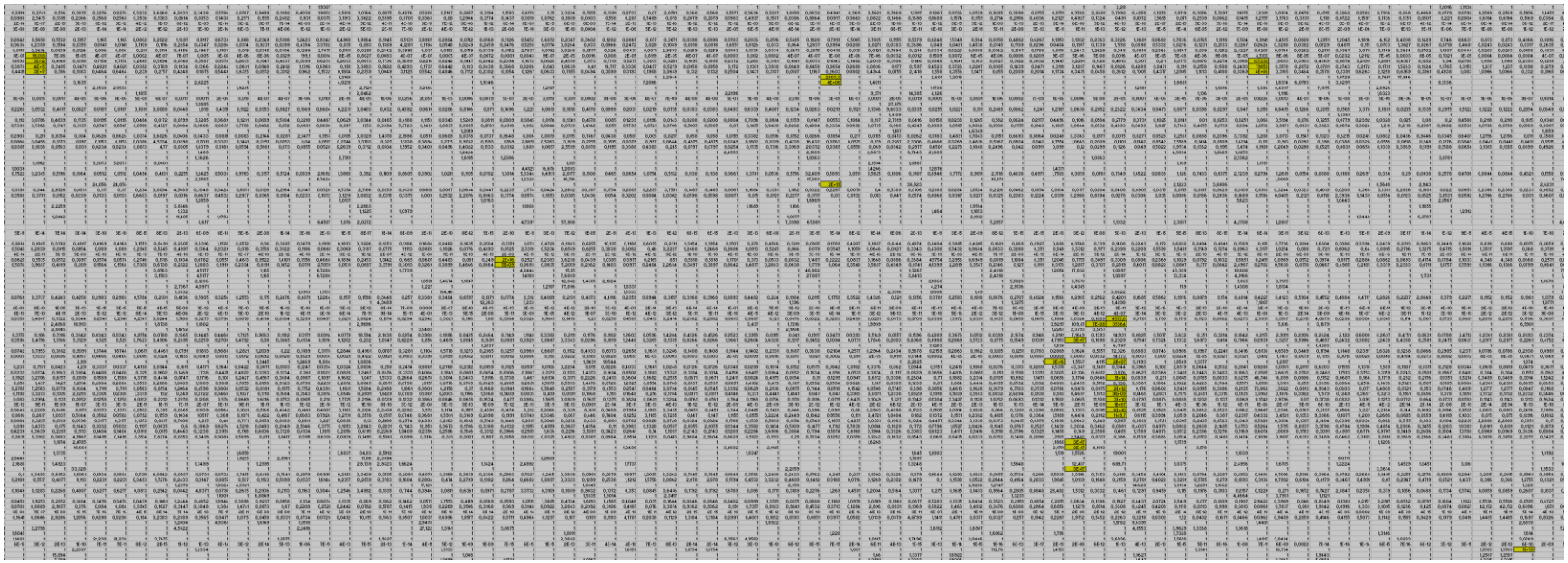
Validation conclusion

The top-down method turns out to **cheaply** calculate a LR that is **conservative** w.r.t. the LR of a more refined model.

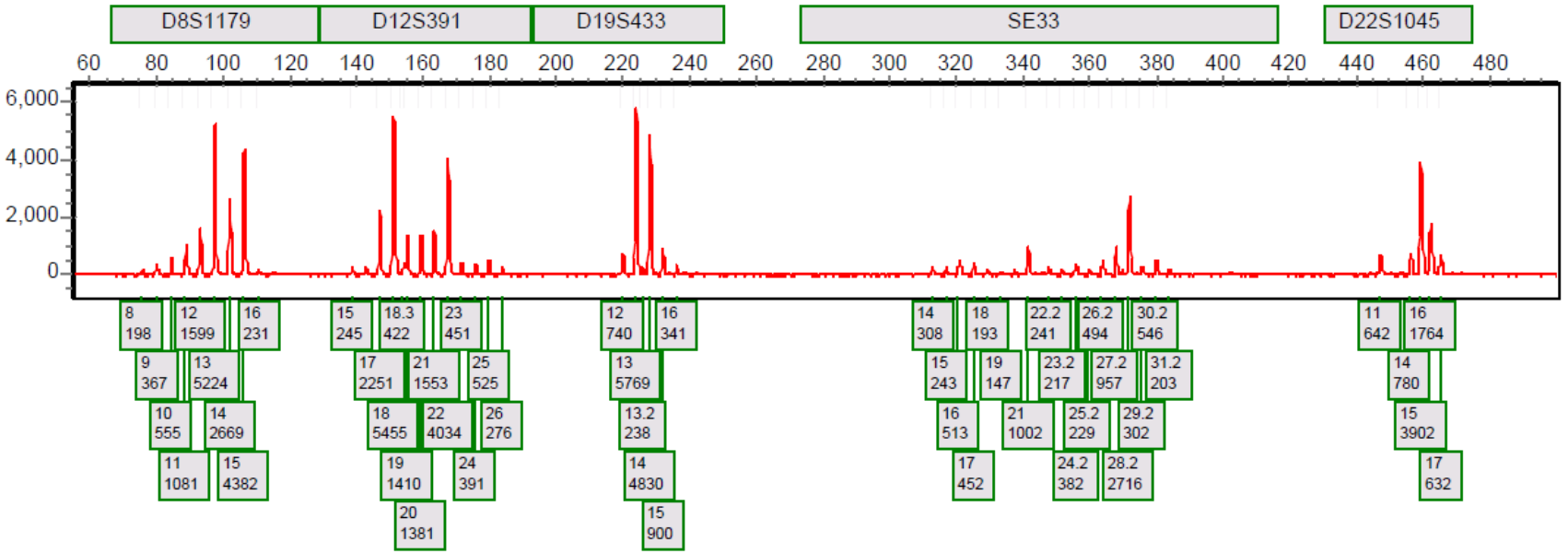
The method can be applied **regardless** of the number of contributors.

Application: **screening** of large numbers of traces with large numbers of possible suspects, to connect more cases with each other.

Screening excerpt (yellow: ‘matches’)



What about this one?



RFU Percentage	MAC (all reps)	MAC (per rep)	NoC	Log(LR)
5%	1	{1}	1	7.54818
10%	1	{1}	1	7.54818
15%	1	{1}	1	7.54818
20%	1	{1}	1	7.54818
25%	1	{1}	1	7.54818
30%	2	{2}	1	14.1673
35%	2	{2}	1	16.9138
40%	2	{2}	1	17.4141
45%	3	{3}	2	17.8321
50%	3	{3}	2	17.7564
55%	4	{4}	2	18.8395
60%	5	{5}	3	Not calculated
65%	6	{6}	3	Not calculated
70%	7	{7}	4	Not calculated
75%	8	{8}	4	Not calculated
80%	9	{9}	5	Not calculated
85%	11	{11}	6	Not calculated
90%	13	{13}	7	Not calculated
95%	15	{15}	8	Not calculated
100%	17	{17}	9	Not calculated

Very strong evidence to be in the top-2. Out of how many? We'll never know.

Top-down implementations

Currently in use at NFI.

Also picked up by others

Research paper

Validation of a top-down DNA profile analysis for database searching using a fully continuous probabilistic genotyping model

Duncan Taylor [a, b], Jo-Anne Bright [c], Lenara Scandrett [a], Damien Abarno [a], Shan-I Lee [c], Richard Wivell [c], Hannah Kelly [c], John Buckleton [c, d]

Show more

Thank you for your attention

Contact: k.slooten@(vu/nfi).nl