The number of distinct alleles in mixed DNA profiles when contributors are related

Maarten V. Kruijver[1]
James M. Curran[2]

[1]ESR
[2]University of Auckland

November 2022

DNA Mixtures

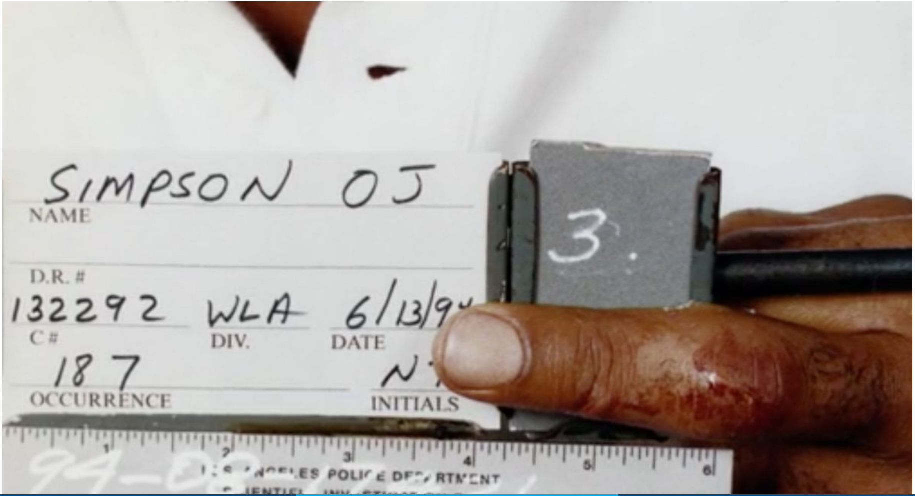
- Much of our forensic statistics work relates to the interpretation of DNA mixtures.
- DNA mixtures occur when (somewhat obviously) DNA from two or more people is mixed.

A real-life example

In 1994 famous NFL star and actor, OJ Simpson, was arrested for the murder of Nicole Brown Simpson (his ex-wife) and Ron Goldman (her partner).

A real-life example

In 1994 famous NFL star and actor, OJ Simpson, was arrested for the murder of Nicole Brown Simpson (his ex-wife) and Ron Goldman (her partner).



Evaluation of DNA mixtures

- We quantify the weight of DNA evidence using the likelihood ratio

$$LR = \frac{\Pr\left(Evidence|H_p\right)}{\Pr\left(Evidence|H_d\right)}.$$

- We can use this update our belief about the hypotheses

$$\frac{\Pr(H_p|Evidence)}{\Pr(H_d|Evidence)} = \frac{\Pr(Evidence|H_p)}{\Pr(Evidence|H_d)} \times \frac{\Pr(H_p)}{\Pr(H_d)}.$$

$$\text{Posterior Odds} = \text{Likelihood Ratio} \times \text{Prior Odds}$$

  where H_p and H_d are two mutually exclusive explanations, or hypotheses, for the presence of the evidence.

- We could spend several days talking about how to evaluate the LR, but today I will focus on just one small part of it: the number of contributors.

The hypotheses under consideration

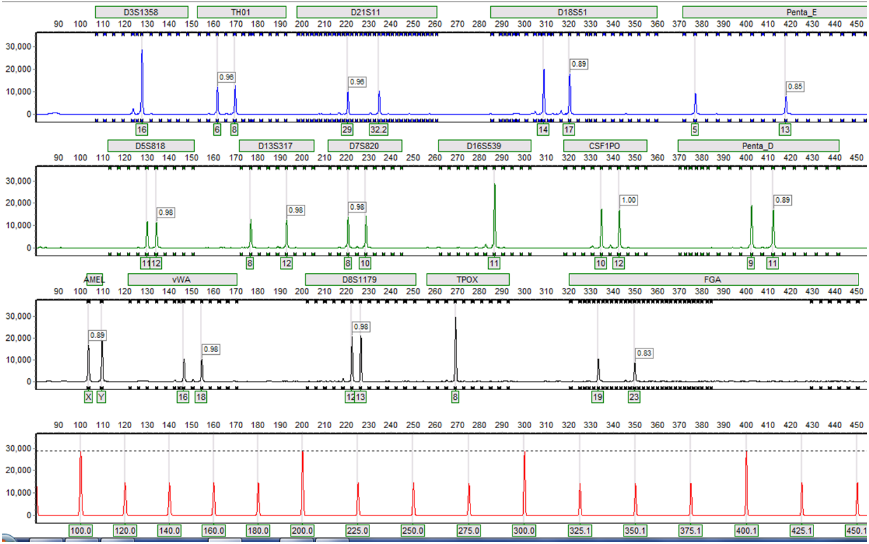
- The hypotheses (or propositions), H_p and H_d, speculate on the contributors to a mixture.

- For example:
  H_p: Person of Interest + Victim
  H_d: Person of Interest + Unknown

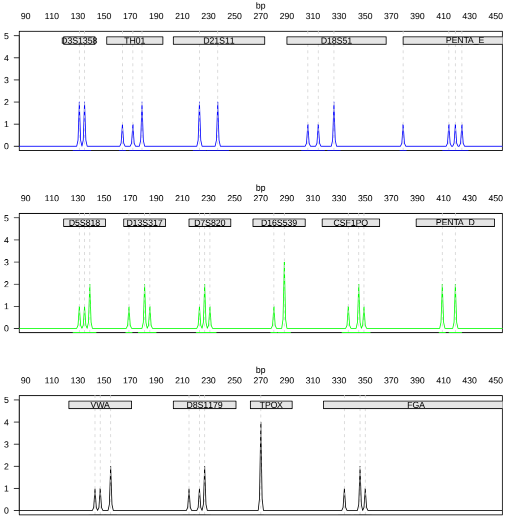
- Implicit in these hypotheses is an assumption about the number of contributors.

- The maximum allele count (MAC) across loci and total allele count (TAC) can inform one's choice about the number of contributors.

What do DNA profiles look like?



What does a mixed DNA profile look like?



$MAC = 4, \; TAC = 40$

A statistician's view

- We have L discrete-valued distributions with probability functions $p_1, p_2, \ldots, p_L$.

- We sample $N = 2k$ items with replacement for each distribution, yielding X_ℓ unique items, $\ell = 1, \ldots, L$

- Let

$$MAC = \max_\ell X_\ell$$

  and

$$TAC = \sum_{\ell=1}^{L} X_\ell$$

- What is $\Pr(TAC = tac|N)$ or $\Pr(MAC = mac|N)$?

Previous work

- This problem was explored through naïve Monte Carlo simulation by Paoletti et al. (2005), and later extended by Buckleton, Curran and Gill (2007), Curran and Buckleton (2014), and Coble et al. (2015).
- With the benefit of hindsight, this work would have been vastly better if we had used importance sampling.
- Tvedebrink (2013) derived an elegant and exact solution, now very efficiently implemented in the R package `DNAtools`.
- It should be noted that all solutions assume that the X_ℓ's are independent of each other.

Independent and distinct alleles

- For a given number, M, of *independent* alleles at a locus, we can obtain the probability distribution of the number of *distinct* alleles, N.

- For example, considering the genotype of one person (two independent alleles), we have $N = 2$, and

$$\Pr(N = 1|M = 2) = \sum_{a} {p_a}^2$$

$$\Pr(N = 2|M = 2) = \sum_{a \neq b} p_a p_b = 1 - \sum_{a} {p_a}^2$$

- We implemented an efficient algorithm to compute the probability distribution N given M.

What about relatives?

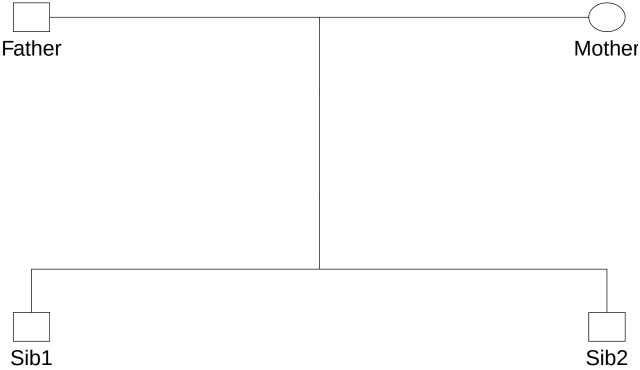
Lawyers, especially defence lawyers, have learned that life becomes more complicated when relatives are involved.

We can do this!

- If contributors are related, then the number of independent alleles M is no longer fixed.
- If contributors are related, then M is often constrained to be much smaller if they are not related.
- We exploit the Identical by Descent (IBD) pattern distribution for a set of pedigree members to obtain the probability distribution of M.
- The IBD pattern distribution generalises the three well known IBD states for pairwise relationships to an arbitrary number of persons.

Example: two full siblings

Consider the two siblings in the following pedigree



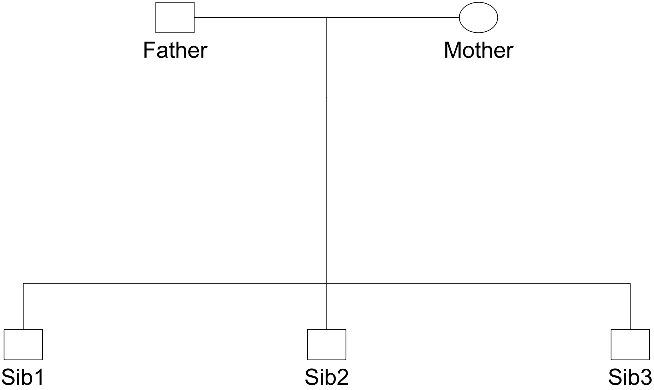
Example: two full siblings

V	$\Pr(v_i)$	S_1	S_2	M
v_1	$1/4$	1 2	1 2	2
v_2	$1/2$	1 2	1 3	3
v_3	$1/4$	1 2	3 4	4

IBD pattern distribution (V) for two full siblings (S_1, S_2)

Example: three full siblings

Now consider the three siblings in the following pedigree



Example: three full siblings

v_i	$\Pr(v_i)$	S_1	S_2	S_3	M
v_1	$1/16$	1 2	1 2	1 2	2
v_2	$1/8$	1 2	1 2	1 3	3
v_3	$1/16$	1 2	1 2	3 4	4
v_4	$1/8$	1 2	1 3	1 2	3
v_5	$1/8$	1 2	1 3	1 3	3
v_6	$1/8$	1 2	1 3	2 4	4
v_7	$1/8$	1 2	1 3	3 4	4
v_8	$1/16$	1 2	3 4	1 2	4
v_9	$1/8$	1 2	3 4	1 3	4
v_{10}	$1/16$	1 2	3 4	3 4	4

IBD pattern distribution (V) for three full siblings (S_1, S_2 and S_3)

Recap

- In general we are interested in $\Pr(N = n)$ where N is the number of distinct alleles observed in a mixed DNA profile.

- If there is a pedigree, then we need to take into account the number of independent alleles, M. We do this by marginalising over the distribution of M.

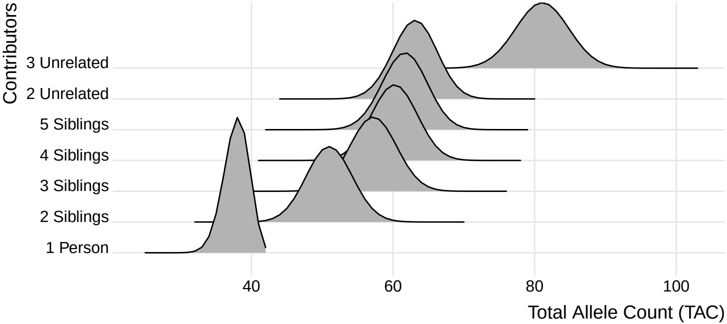
$$\Pr(N = n) = \sum_m \Pr(N = n | M = m) \Pr(M = m).$$

- The IBD pattern allows us to compute $\Pr(M = m)$, i.e.

$$\begin{aligned}\Pr(M = m) &= \sum_v \Pr(M = m, V = v) \\ &= \sum_v \Pr(M = m | V = v) \Pr(V = v).\end{aligned}$$

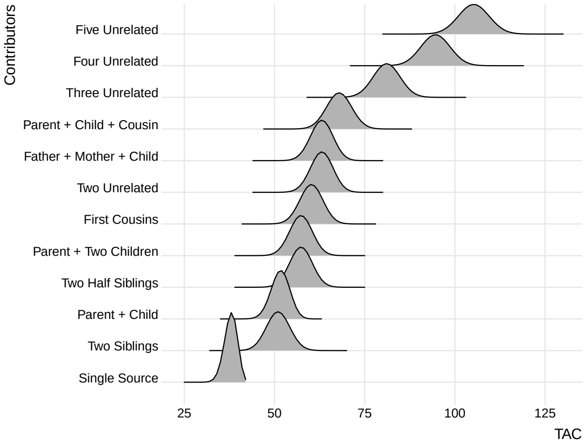
Results for sets of full siblings

We computed the distributions for a variety of pedigrees using a 21 locus multiplex and allele frequencies from US Caucasians.



Results for further combinations of relatives

We computed the distributions for a variety of pedigrees using a 21 locus multiplex and allele frequencies from US Caucasians.



Conclusions

- We have developed a method for predicting the total number of distinct alleles in a mixed DNA profile.
  - Contributors may be related according to a pedigree.
  - The effect of dropout can optionally be modelled (not shown).
- It is generally possible to identify mixtures of relatives based on the Total Allele Count.
- It is challenging to correctly assign the number of contributors to a mixture if the donors are related.
- An R-package named `numberofalleles` implements the methods.

Acknowledgements

This work was supported in part by grant NIJ 2020-DQ-BX-0022 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of their organisations.