Inference about relationships from DNA mixtures

Julia Mortera

Università Roma Tre

Based on joint work with Peter Green

Outline

- Motivating example
- DNA mixtures
- Methods for inference about relationships
- Results from two real casework examples

A real case from the Forensic Institute Sapienza Università Roma

A famous Italian singer B, while working on the set of an operetta, met a young dancer and began a secret relationship. One of the dancer's sons, A, as an adult learns that he is probably the singer's son. Some years after the singer's death, A claims his share of B's substantial inheritance. After his mother's death and almost 20 years after B's death, B's body is exhumed and DNA is extracted from a bone.

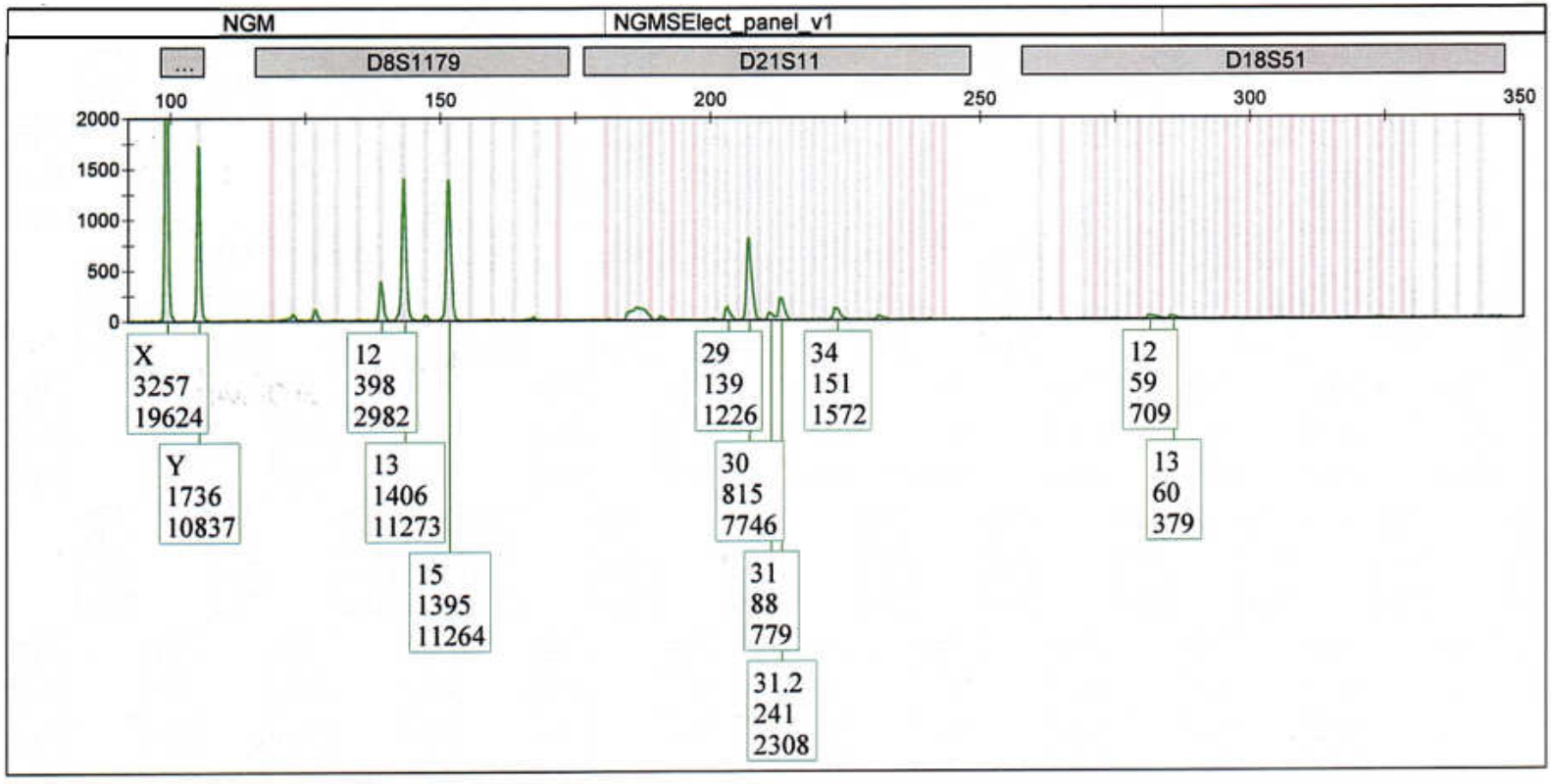
This is to be used to establish if A could be the son of B.

The DNA is highly contaminated and appears to be mixture of at least 2 individuals.

Extract of Paternity Testing Data

Alleged son			Data from B's bone	
A's genotype		Marker	Alleles	Peak height
15	16	D8S1179	12	398
			13	1406
			15	1395
30	32	D21S11	29	139
			30	815
			31	88
			31.2	241
			34	151
...	...	...	...	...
14	20	SE33	20	139

Electropherogram (EPG)



Peak height is roughly proportional to the amount of the allele pre-PCR.

Statistical model for relationships from DNA mixtures

Joint model for EPG, genotypes and relationships.

Combine

1. *model for peak heights* for fixed genotypes;
2. *model for genotypes;*
3. models for inference about relationships.

The gamma model for peak heights JRSS (2016)

Consider a mixture of DNA from $i \in I$ contributors.

- Contribution H_{ia} to peak at a is roughly proportional to the *amount of DNA* contributed by individual i;
- H_{ia} are *independent and gamma-distributed*:

$$H_{ia} \sim \Gamma(\rho\phi_i n_{ia}, \eta)$$

ρ is proportional to the *total amount of DNA* (before PCR)

ϕ_i is the *proportion of DNA* from individual i. This is constant across markers.

n_{ia} is the *number of alleles* of type a that individual i has.

η is a scale parameter

The gamma model for peak heights

$H_{ia} \sim \Gamma(\rho\phi_i n_{ia}, \eta)$ implies

$$\mathbb{E}(H_{ia}) = \rho\phi_i n_{ia}\eta$$
$$\mathbb{V}(H_{ia}) = \rho\phi_i n_{ia}\eta^2 \propto \mathbb{E}(H_{ia})$$

For a single heterozygous contributor:
$\mu = \rho\eta$ is the average peak height;
$\sigma = 1/\sqrt{\rho}$ is the coefficient of variation for peak heights.

The gamma model for peak heights

The individual peak heights H_{ia} are not observable. Ignoring artefacts (stutter, dropout, and noise), we observe

$$H_{+a} = \sum_{i \in I} H_{ia},$$

combining all the contributions from the individuals to the peak at allele a for a given marker. Using the additivity properties of the gamma distribution we also have that

$$H_{+a} \sim \Gamma\{\rho B_a(\phi, \mathbf{n}), \eta\}, \text{ where } B_a(\phi, \mathbf{n}) = \sum_i \phi_i n_{ia}$$

is the *effective number of alleles* at a. The model is extended to take into account artefacts like *stutter* and *dropout*.

Likelihood function

For given genotypes $\mathbf{n}$, given ϕ and fixed values of parameters (ρ, ξ, η) all observed peaks are independent. Thus *the conditional likelihood function* based on observations $\mathbf{z} = \{z_{ma}\}_{m \in M, a \in A_m}$ is

$$L(\rho, \xi, \phi, \eta \,|\, \mathbf{z}, \mathbf{n}) = \prod_m \prod_a L_{ma}(z_{ma})$$

where

$$L_{ma}(z_{ma}) = \begin{cases} g\{z_{ma}; \rho D_a(\phi, \xi, \mathbf{n}), \eta\} & \text{if } z_{ma} > C \\ G\{C; \rho D_a(\phi, \xi, \mathbf{n}), \eta\} & \text{otherwise.} \end{cases}$$

with g and G denoting the gamma density and cdf respectively, and where $D_a(\phi, \xi, \mathbf{n}) = (1 - \xi) B_a(\phi, \mathbf{n}) + \xi B_{a+1}(\phi, \mathbf{n})$ is the effective number of alleles of type a after stutter, where ξ denotes the mean stutter proportion.

Full likelihood

For a given hypothesis $\mathcal{H}$ on the number of contributors, the full likelihood is obtained by *summing over all possible combinations of genotypes*

$$L(\mathcal{H}) = \Pr(E \,|\, \mathcal{H}) = \sum_{\mathbf{n}} L(\rho, \xi, \phi, \eta \,|\, \mathbf{z}, \mathbf{n}) P(\mathbf{n} \,|\, \mathcal{H}).$$

with probabilities associated with the hypothesis, $P(\mathbf{n} \,|\, \mathcal{H})$. *This sum is astronomical in size* for any hypothesis which potentially involves unknown contributors to the mixture.
However, the sum *can be calculated efficiently by appropriate use of Bayesian network (BN) techniques*. It is in fact the normalising constant of the BN.

Estimating unknown parameters

The likelihood function $L(H)$ involves a number of parameters $(\rho, \xi, \theta, \eta)$ which may be completely or partially unknown. One way of dealing with this is to choose parameters associated with the hypothesis that make the likelihood as large as possible and thus calculate

$$\hat{L}(H) = \sup_{\rho,\xi,\theta,\eta} \sum_{\mathbf{n}} L(\rho, \xi, \theta, \eta \,|\, \mathbf{z}, \mathbf{n}) P(\mathbf{n} \,|\, H)$$

corresponding to using *maximum likelihood estimates for the unknown parameters.*

Computational aspects

The main bottleneck is the need to *sum over possible genotypes.*
We use two tricks:

1. Markov genotype representation.
2. Computation by auxiliary nodes in a Bayesian network.

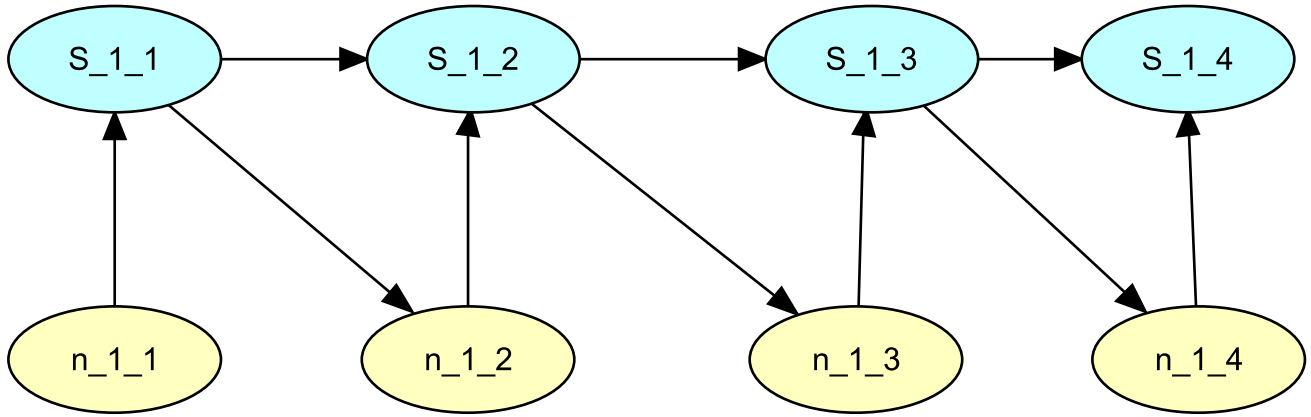
Benefits

- *Exact evaluation of likelihood function* and other quantities by probability propagation.
- We can introduce *auxiliary likelihood nodes to also represent family relationships.*
- *No approximations* are introduced to be able to perform computations (apart from numeric maximisation and differentiation for computing MLEs)
- We can currently handle up to *6 unknown contributors.*

Genotype representation

Imagine the 2 alleles of a genotype being allocated sequentially. Let S_{ia} be the partial sums $S_{ia} = \sum_{b \le a} n_{ib}$. Then $n_{i1} \sim \mathrm{Bin}(2, q_1)$ and

$$n_{i,a+1} \,|\, (n_{i1}, n_{i2}, \cdots, n_{ia}) \sim \mathrm{Bin}\left(2 - S_{ia}, q_{a+1} / \sum_{b>a} q_b\right)$$



Hence a genotype is modelled by a Markov structure and computations can be done linearly in number of alleles.

Relationship Identification from DNA mixtures

Examples are:

- Is A the son of a contributor to a DNA mixture? A's mother's DNA might or not be available.
- Is a contributor to the mixture the son of individual A?
- is A a family relative of one or both contributors to a mixture?

The evidence is E=$\{$DNA mixture, genotype of measured individuals$\}$

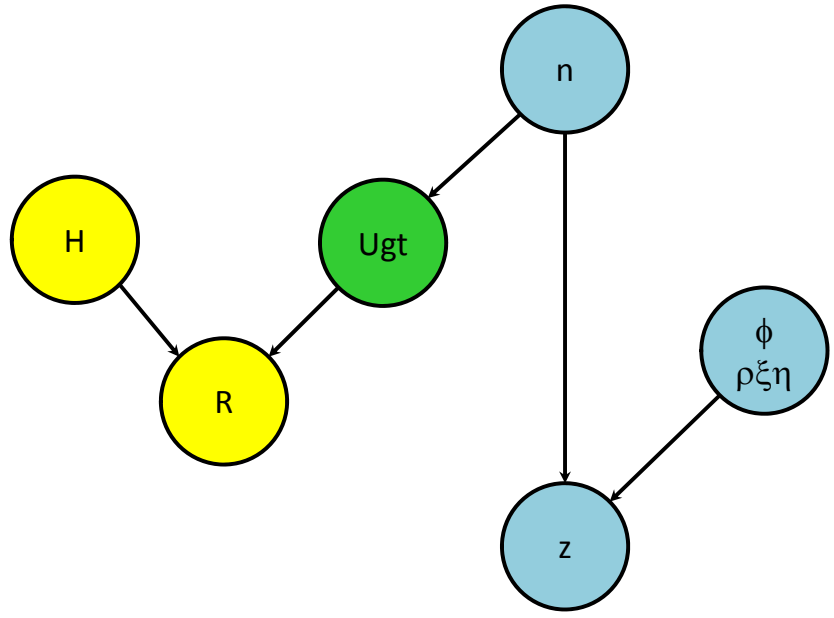
General Setup

Let $U_i = U$ be a specified contributor to the mixture, and let *Ugt* denote the genotype of U.

We are interested in assessing a potential relationship between U and one or more other individuals *who have a known relationship to each other*. Let R denote the genotype of these individuals.

$\mathcal{H}_1 : U$ has the specified relationship with individuals whose genotypes are in R
$\mathcal{H}_0 : U$ is unrelated with individuals whose genotypes are in R

DAG for potential relationship with U



Blue nodes: The gamma model;

Yellow nodes: Putative relationship between mixture contributor *Ugt* and relatives with genotypes R under $\mathcal{H}$.

Also, $R \perp\!\!\!\perp z \mid Ugt$.

Examples for Disputed Paternity

- R is *cgt*
- R is *cgt* and *mgt*

and we are testing if U is the father of the child.

Conditional on *Ugt*

$$\mathsf{LR}_{Ugt} = \frac{P(R|\mathcal{H}_1, Ugt)}{P(R|\mathcal{H}_0, Ugt)} = \frac{P(R|\mathcal{H}_1, Ugt)}{P(R|\mathcal{H}_0)}$$

Since $Ugt \perp\!\!\!\perp R \mid \mathcal{H}_0$.

Likelihood Ratio

The LR for $\mathcal{H}_1 : U$ has the specified relationship with individuals whose genotypes are in R; against the contrary hypothesis $\mathcal{H}_0$ is

$$\begin{aligned}
\mathsf{LR} &= \frac{P(R, z|\mathcal{H}_1)}{P(R, z|\mathcal{H}_0)} = \frac{P(R, z|\mathcal{H}_1)}{P(R|\mathcal{H}_0)p(z|\mathcal{H}_0)} \\
&= \frac{\sum_{\mathit{Ugt}} P(R|\mathcal{H}_1, \mathit{Ugt})p(z|\mathit{Ugt})P(\mathit{Ugt})}{P(R|\mathcal{H}_0)p(z)} \\
&= \sum_{\mathit{Ugt}} \mathsf{LR}_{\mathit{Ugt}} \times p(\mathit{Ugt}|z)
\end{aligned}$$

Note that

$$\mathsf{LR} \leq \max_{\mathit{Ugt}} \mathsf{LR}_{\mathit{Ugt}},$$

so inference is always less incriminating than if the most probable *Ugt* were directly observed.

Methods for inference on relationships: e.g. for paternity

$$\mathsf{LR} = \sum_{Ugt} \mathsf{LR}_{Ugt} \times p(Ugt|z)$$

Weighted likelihood ratio (WLR) The set of possible genotypes from a mixture deconvolution together with their probability ranking are used to approximate $p(Ugt|z)$ and hence to compute the likelihood ratio for paternity.

Additional likelihood node (ALN) uses the Markov genotype BN and adds a single likelihood node as a child of relevant paternal allele counts $\{n_{i,a}\}$ and the LR is computed by probability propagation.

Only child genotyped

$$\mathsf{LR}_{Ugt} = \frac{P(cgt|Ugt, \mathcal{H}_1)}{P(cgt|\mathcal{H}_0)} = \begin{cases} \frac{n_{ia}}{2q_a} & \text{if } cgt = aa, \\ \frac{n_{ia}}{4q_a} + \frac{n_{ib}}{4q_b} & \text{if } cgt = ab \end{cases}$$

The only allele counts for U_i needed as parents for defining the CPT for the likelihood node *are those (a and/or b) in cgt*.

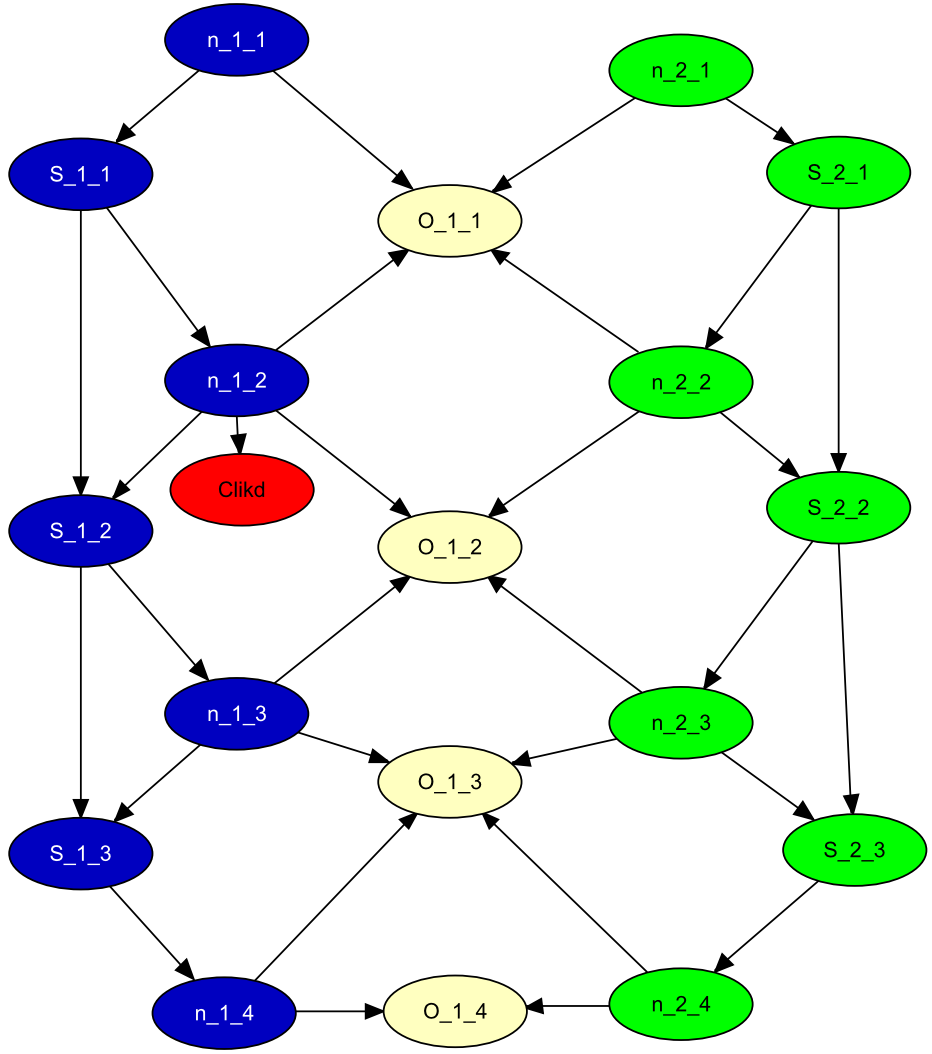
Mother and Child genotyped

$$\mathsf{LR}_{Ugt} =$$

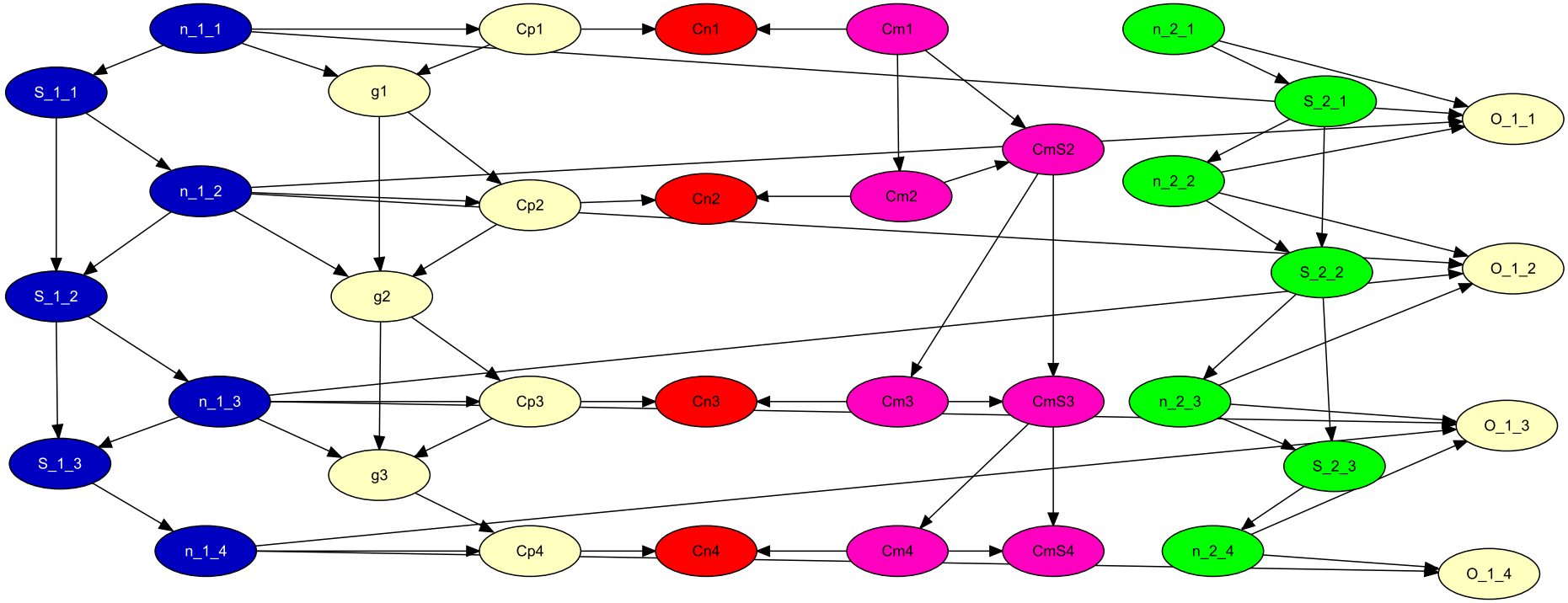
$$\frac{P(cgt|mgt, Ugt, \mathcal{H}_1)}{P(cgt|mgt, \mathcal{H}_0)} = \begin{cases} \frac{n_{ia}}{2q_a} & \text{if } cgt = aa, mgt = \{aa, ab\} \\ \frac{n_{ib}}{2q_b} & \text{if } cgt = ab, mgt = \{aa, ac\} \\ \frac{n_{ia}+n_{ib}}{2(q_a+q_b)} & \text{if } cgt = ab, mgt = ab \end{cases}$$

Again, *the only allele counts for U_i* needed as parents for defining the CPT for the likelihood node *are those (a and/or b) in cgt.*

Bayesian network for ALN (homozygous cgt $= (2, 2)$)



Meiosis Bayesian net (MBN)



Under $\mathcal{H}_1$, meiosis is captured for the child's paternal allele counts using segregation indicator g nodes. We represent meiosis in the sequential-over-alleles notation.

Replacing the probability tables (RPT)

Replace the default $P(Ugt)$ tables (based on Hardy-Weinberg equilibrium in the assumed population) in the network, with tables for $P(Ugt|\mathcal{H}_1, R)$, e.g. the father's genotype tables, given *cgt*. The Markov genotype structure is maintained.

For example, if $cgt = (a^*, a^*)$, then the binomial distribution is replaced by

$$n_{i,a+1}|S_{ia} \sim \delta_{a+1,a^*} + \text{Bin}\left(1 - S_{ia}, q_{a+1}/\sum_{b>a} q_b\right)$$

For $a = a^*$, the table for S_{ia} are re-defined to be the cumulative sums of the n_{ia} *excluding the IBD allele*.

Results: Disputed Paternity Case

Condition on information that U_1 is male.

Extract top-ranking genotypes and their probability

Marker	Top-ranked genotype		Alleged son's genotype, *cgt*		Probabilities without and with maleness
D8	13	15	15	16	0.7279, 0.7292
D21	30	31.2	30	32	0.3528, 0.3531
...	...	...	...	...	..., ...
SE33	20	20	14	20	0.9925, 0.9926

U_1's top ranking predicted genotype is compatible with *cgt* on all markers.

Deconvolution for marker D21

Rank	*Ugt*		Prob.	$\Pr(cgt \mid Ugt, \mathcal{H}_p)$	$\Pr(cgt \mid \mathcal{H}_0)$
1	30	31.2	0.353	0.0055	0.0051
2	30	34	0.258	0.0055	0.0051
3	30	31	0.190	0.0055	0.0051
4	31	31.2	0.079	0	0.0051
...	...	...	...	...	...
9	29	34	0.0016	0	0.0051

Recall that *cgt*'s genotype is (30,32), so the predicted genotypes from rank 1–3 are all compatible with A. From rank 4 on they are incompatible with *cgt*.

Marker-wise LRs

Marker	Top ranked *Ugt*		Alleged *cgt*		Likelihood ratios Top ranked	Likelihood ratios ALN MBN & RPT
D8	13	*15*	*15*	16	1.76	1.51
D21	*30*	31.2	*30*	32	1.08	0.869
...	...	...	...	...	...	...
SE33	*20*	*20*	14	*20*	10.18	10.14
$\log_{10} LR$					5.671	5.425

WLR is a good approximation to ALN, MBN and RPT *which give exact answers*. If *Ugt* has the top-ranked genotype we get a $\log_{10}$ LR $= 5.67$ which is reduced by a factor 1.75 based on the mixture.

For a prior probability > 0.01 the posterior probability of paternity is > 0.9994.

LR for Paternity with and without mgt

Method	$\log_{10} LR$	
	without *mgt*	with *mgt*
Exact methods	5.43	8.16

The information on *mgt* *increases the LR by a factor of about 540*.

Computational times

Method	Time (seconds)
ALN	1.32
RPT	1.66
MBN	2.82
WLR	46.90

Italian Criminal Case

A murder case where with 3 mixed crime traces T_1, T_2 and T_3. We also have the genotypes of victim V and alleged mother *mgt* of a contributor to the mixture were also available. We assume that there are at most 3 contributors to each mixture, the victim V *and two unknowns* U_1 *and* U_2.

The extra contributor U_2 *is included to account for dropin*.

MLEs based on combined information from T_1, T_2, T_3

Parameter	T_1	T_2	T_3
μ	3857	1289	1836
σ	0.408	0.671	0.562
ξ	0.127	0.048	0
ϕ_V	0.22	0.53	0.63
ϕ_{U_1}	0.71	0.45	0.37
ϕ_{U_2}	0.067	0.026	0

In T_1 the major contributes $\phi_{U_1} = 0.71 > \phi_V = 0.22$, whereas they each contribute about the same proportion to T_2, and in T_3 $\phi_V = 0.63 > \phi_{U_1} = 0.37$.

U_2 contributes a negligble amount in all traces.

Have we found the culprit?

In this case we want to compare the hypotheses:

$\mathcal{H}_\mathsf{p}$: U_1 is the child of *mgt* *vs.*

$\mathcal{H}_0$: no unknown contributors are related to *mgt*

Using the exact methods the likelihood ratio in favour of $\mathcal{H}_\mathsf{p}$ is $\log_{10} \mathsf{LR} = 4.26$

Comparison with single trace analysis

	separate traces			combined traces
	T_1	T_2	T_3	$T_1 \& T_2 \& T_3$
LR	17156	103.7	238.64	18046
$\log_{10}$LR	4.23	2.02	2.38	4.26

Using solely T_1, (ϕ_{U_1}=70%) the LR is reduced only by a factor of 1.05 the LR based on $T_1 \& T_2 \& T_3$.

Using trace T_2/T_3 alone yield LR about 174/76 times smaller than the LR based on $T_1 \& T_2 \& T_3$.

Leaving the contributor unspecified

Exactly the same method could be used to evaluate similar hypotheses referring to U_2. (The contributors are numbered in decreasing order of $\phi_{U_1} \geq \phi_{U_2} \geq \cdots$.)

Using a Bayesian interpretation of the likelihood ratio we suggest

$$\mathsf{LR} = \frac{\sum_k \mathsf{LR}_k P(\mathcal{H}_k)}{\sum_k P(\mathcal{H}_k)}.$$

If specifying the relative priors is difficult, appropriate bounds could be:

- For criminal trial, report $\min_k \mathsf{LR}_k$, erring on side of caution;
- For a civil court, report the range of LR over a reasonable range of priors.

Software

Calculations are done in R using a suite of functions `KinMix`, freely available.

This needs to call functions in the `RHugin` package to augment the capabilities of the `DNAmixtures` package.

Final Remarks

- The methods can be readily extended to analyse different scenarios. They are not limited to particular details of the genotyping kits, allele frequencies, number of contributors, or hypotheses in these examples.
- Which method is better suited will depend on the complexity of the relationship. For example, ALN in paternity case is linked to 1 or 2 allele counts. In more complex relationships one could need more links and it would be slower.
- Method could also be useful for identifying disaster victims.