Stochastic methods for forensic lineage markers

Aug 28, 2018

Mikkel M Andersen* and David J Balding
* mikl@math.aau.dk

Advanced Statistical and Stochastic Methods in Forensic Genetics

University of Cologne, Germany

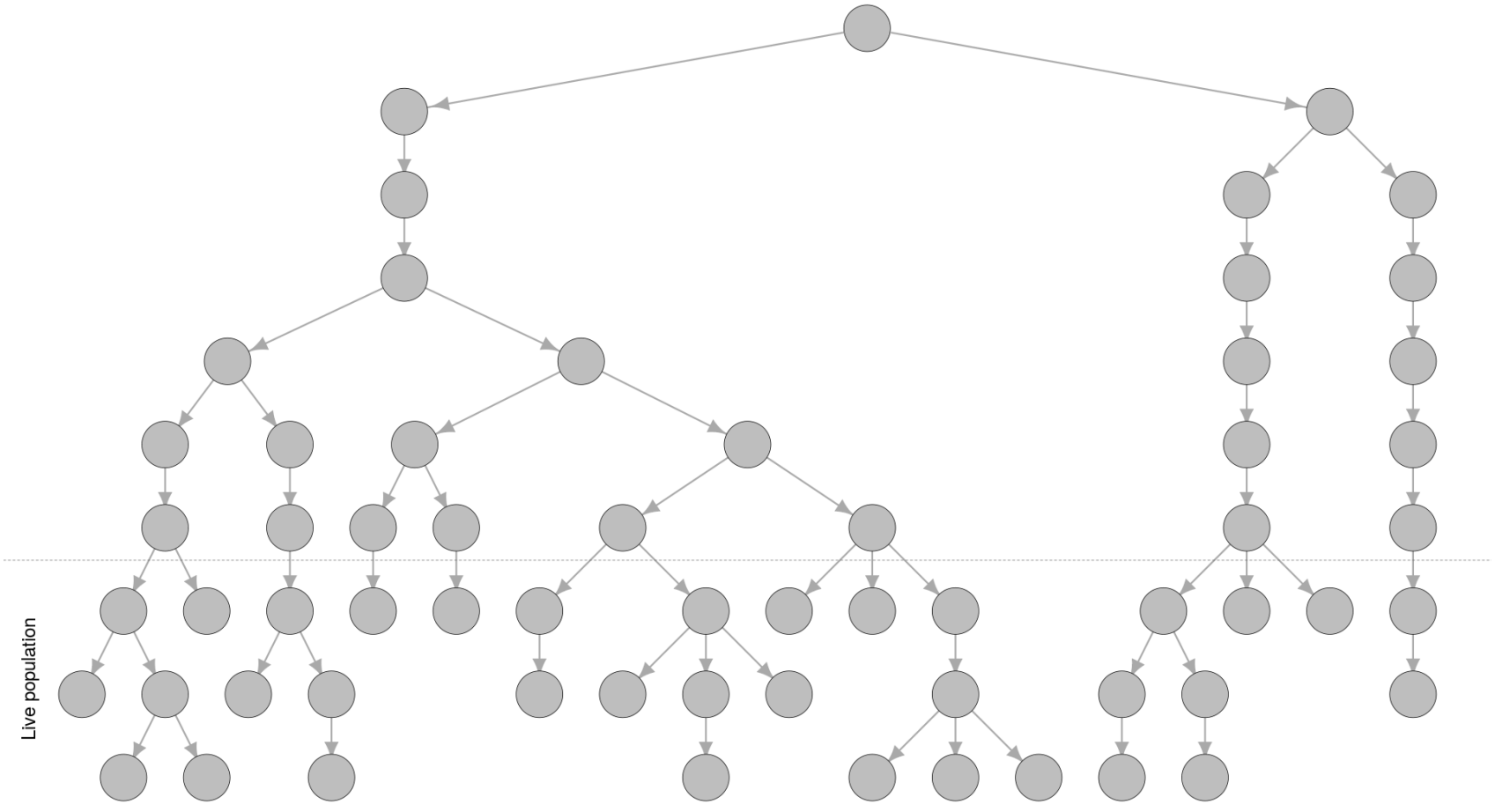
AALBORG UNIVERSITY
DENMARK

Outline

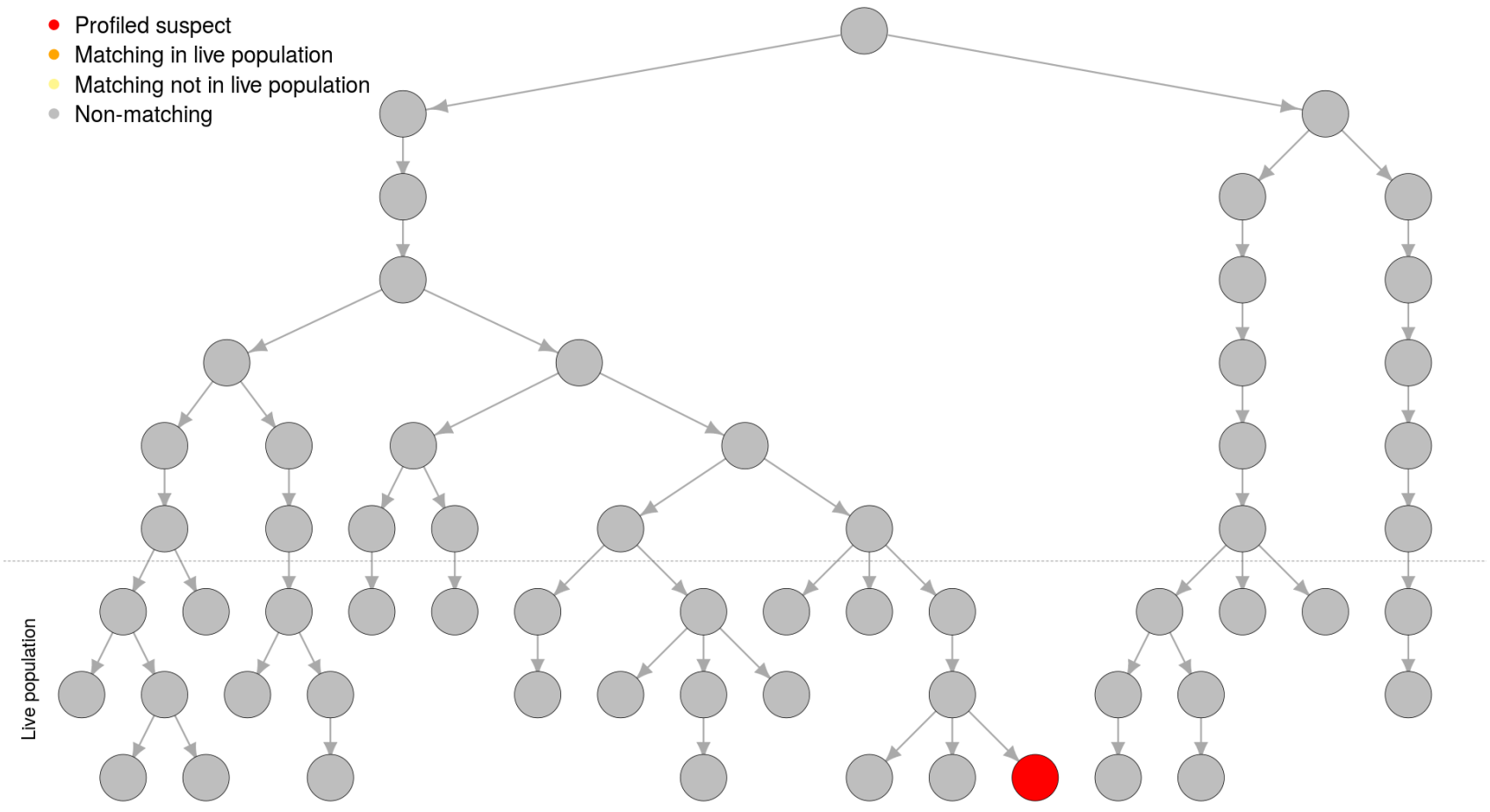
Joint work with David Balding (Isaac Newton Institute, 2016):

- 2017: "How convincing is a matching Y-chromosome profile?" (PLoS Genetics)
  - Modern kits, many loci: all profiles are rare and share by close relatives
  - Match probability – what population?
- Today
  - Recap
  - "Y-profile evidence: close paternal relatives and mixtures"
    - `https://www.biorxiv.org/content/early/2018/07/20/373423`
  - "How many individuals share a mitochondrial genome?"
    - `https://www.biorxiv.org/content/early/2018/07/23/374686`
- All results: Yfiler Plus (PowerPlex Y23 in papers)

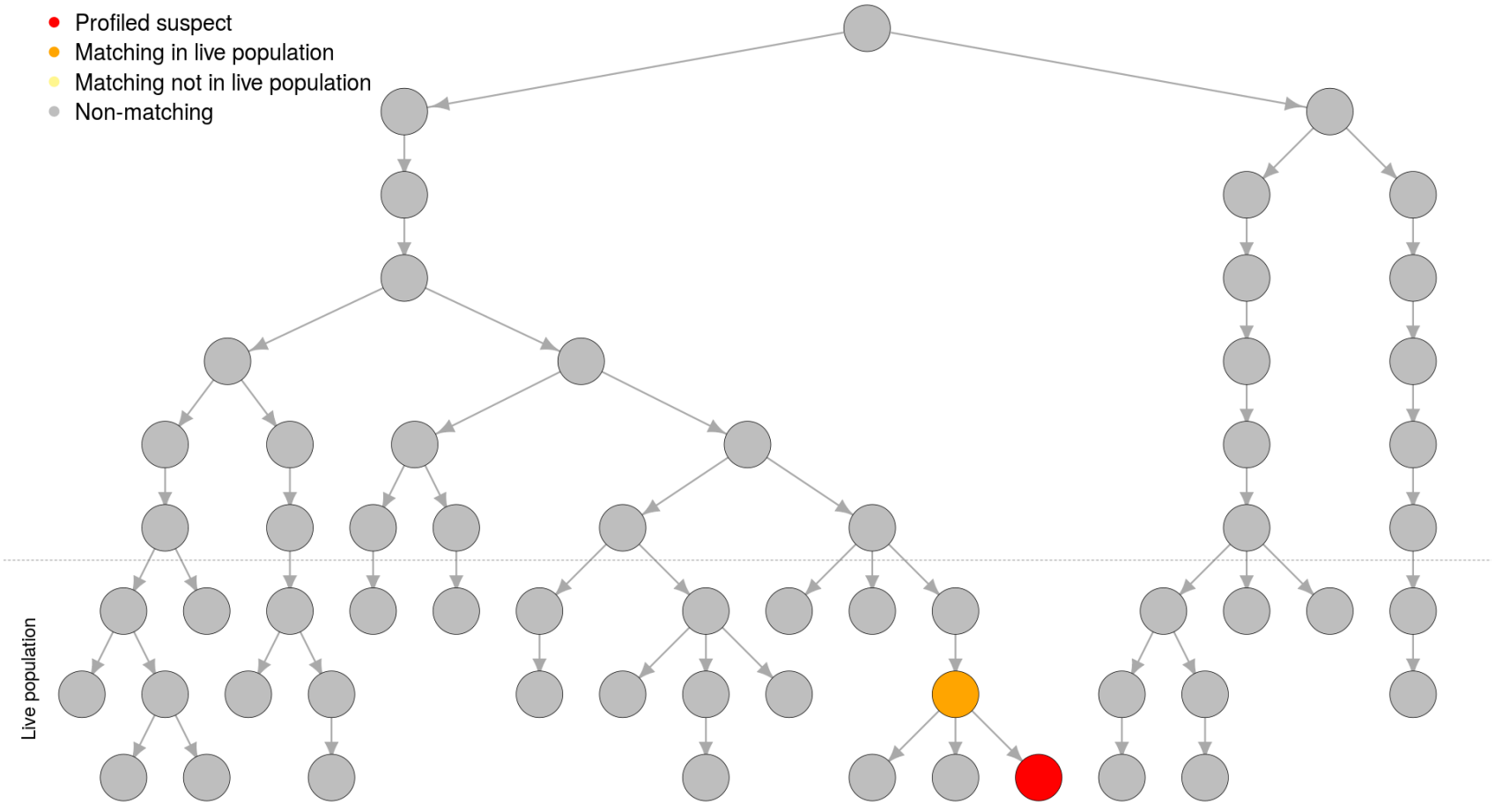
Recap



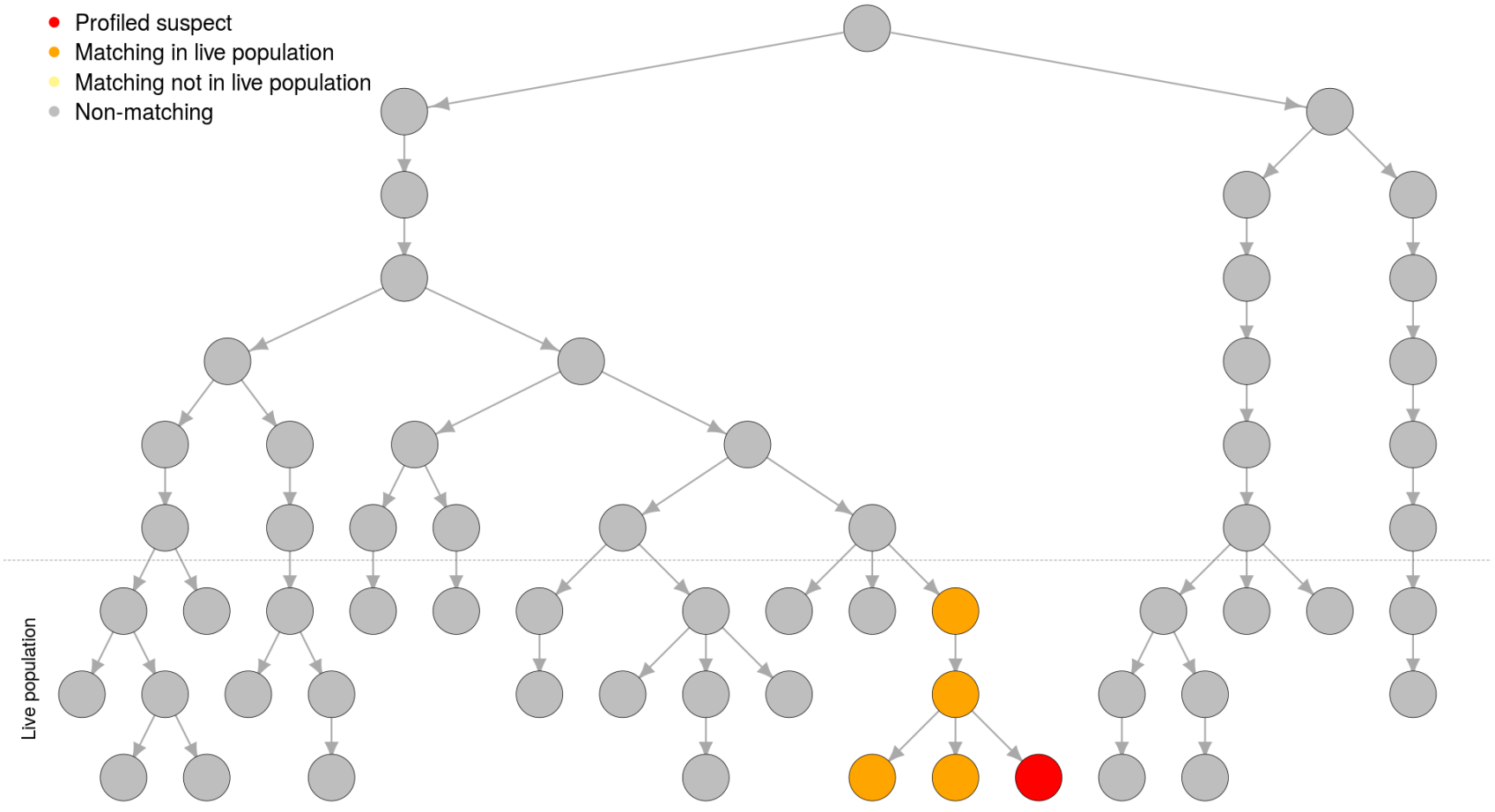
Recap


Profiled suspect
Matching in live population
Matching not in live population
Non-matching
Live population

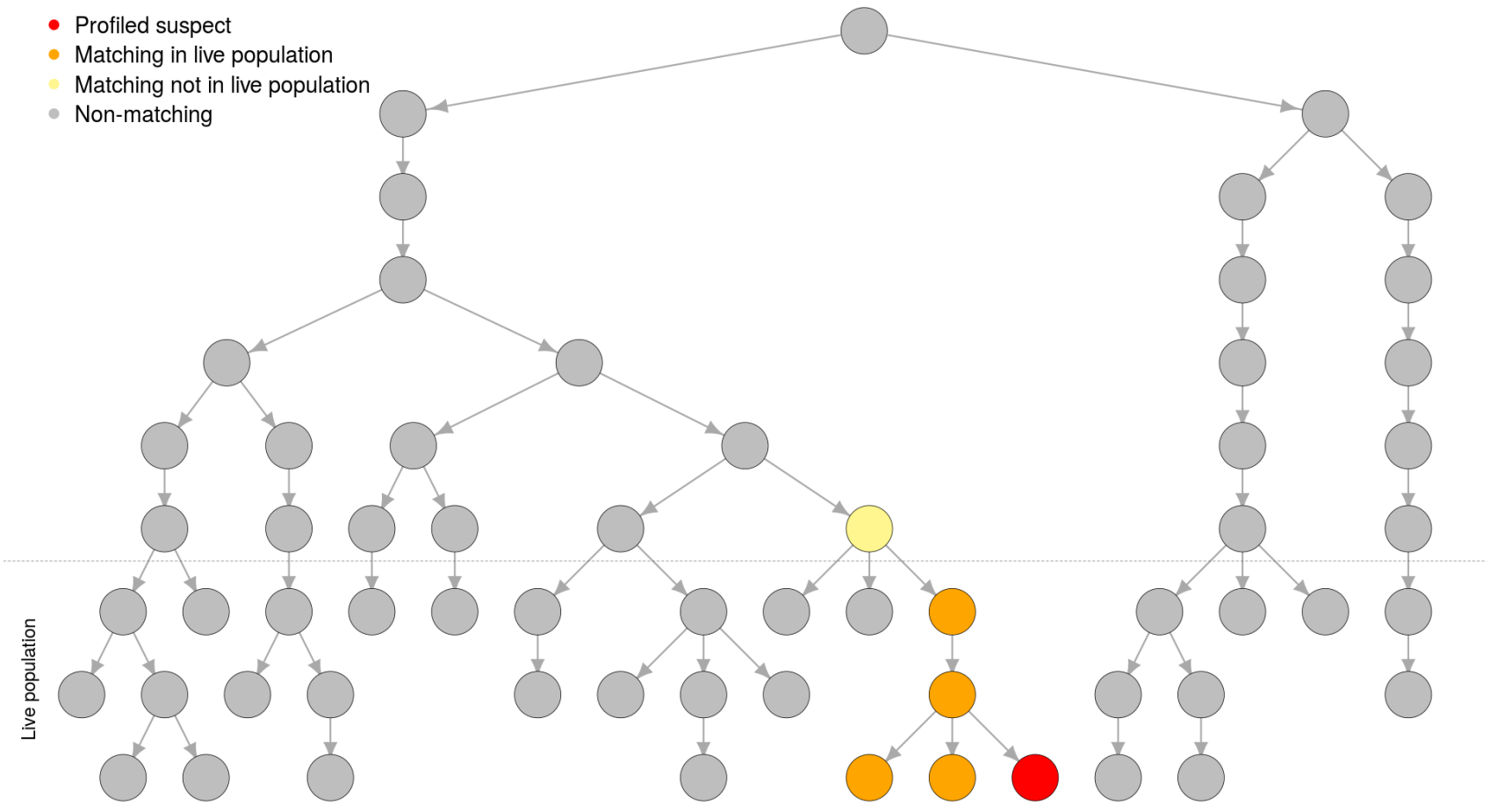
Recap


Profiled suspect
Matching in live population
Matching not in live population
Non-matching
Live population

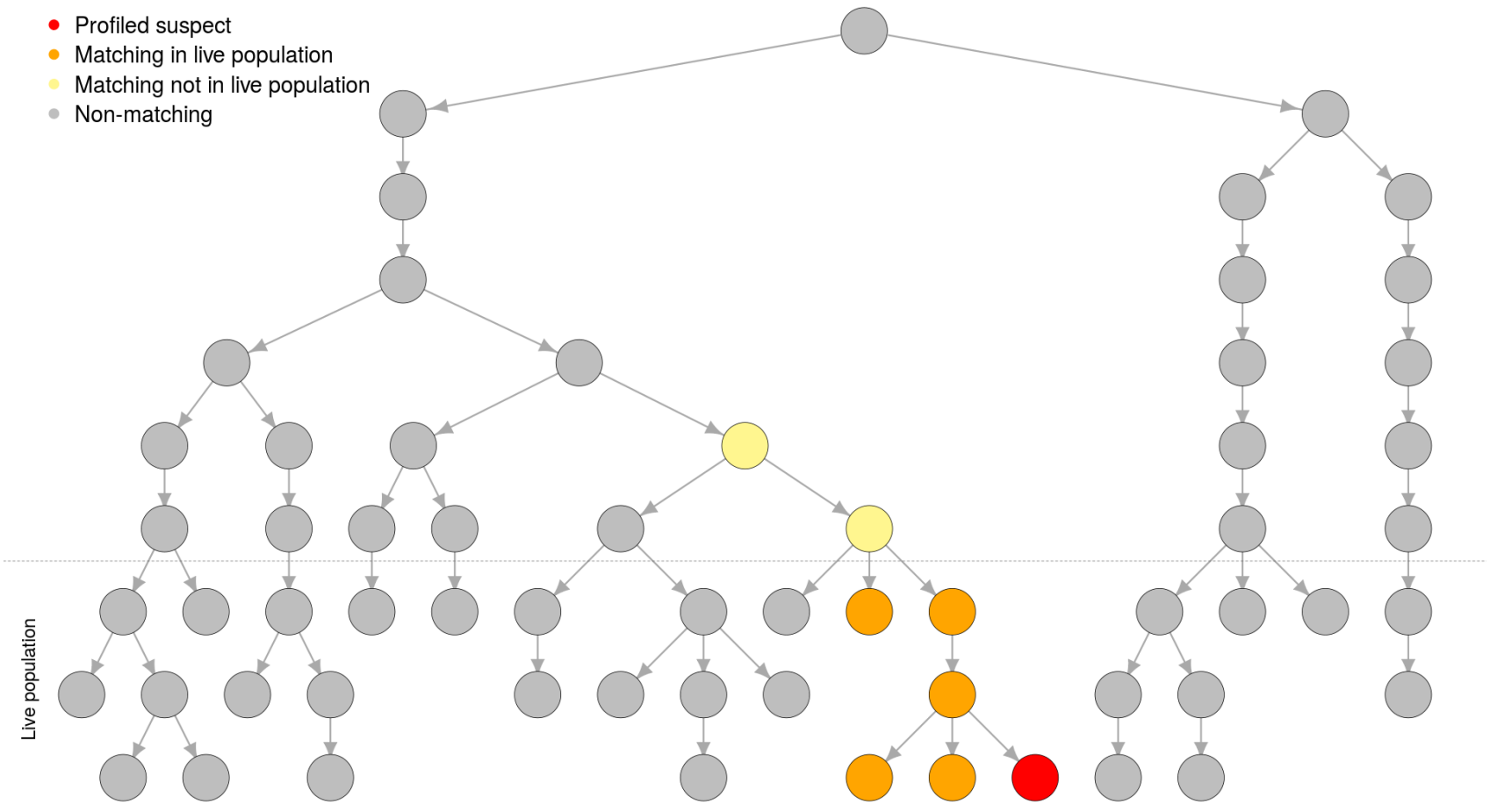
Recap


Profiled suspect
Matching in live population
Matching not in live population
Non-matching
Live population

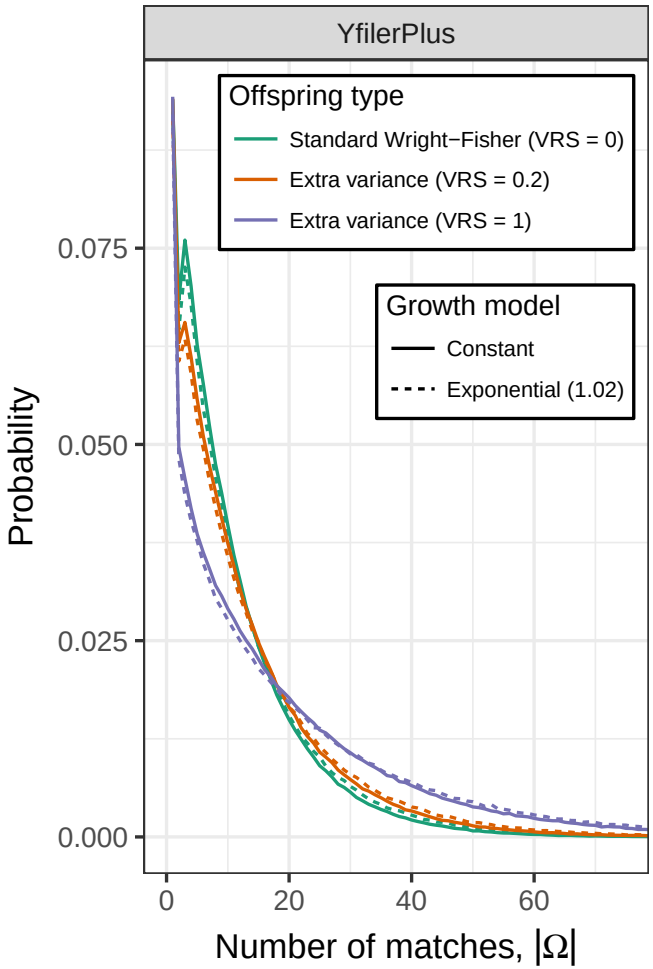
Recap


Profiled suspect
Matching in live population
Matching not in live population
Non-matching
Live population

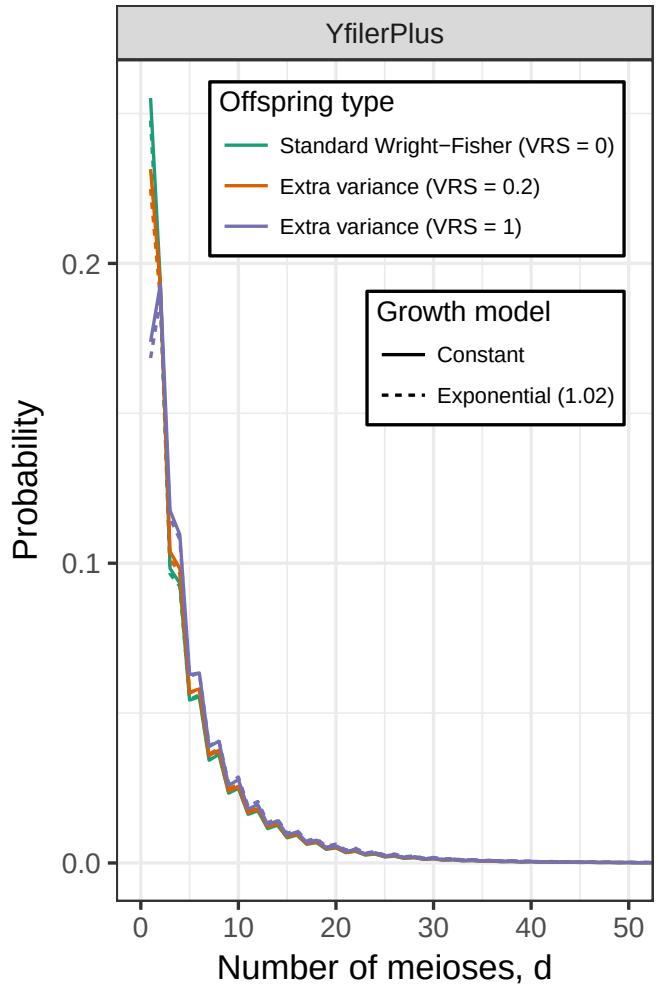
Recap


Profiled suspect
Matching in live population
Matching not in live population
Non-matching
Live population

Recap



95% quantile ≈ 40



95% quantile ≈ 20

Recap: Evidence

[. . .] **the number of males in the population with a matching Y profile is very unlikely (probability $<$ 5%) to exceed 40** [. . .]
They are all **paternal-line relatives** of Q, but the relationship may extend over many father-son steps, well beyond the known relatives of Q. [. . .] **similar in ethnic identity, language, religion, physical appearance and place of residence**.

Recap: Results

- Likelihood ratio:
  - Match probability: Number of meioses from queried contributor Q to the particular individual X
  - Known population of interest (???) with size N: $LR = 1/(40/N) = N/40$
  - "Unrelated man" vs "Random man"
  - What population? Varying population frequency.
- Report: Number of males with matching Y profiles
  - Previously done for autosomal DNA profiles (in mid-1990s, the England and Wales Court of Appeal recommended this instead of a match probability)

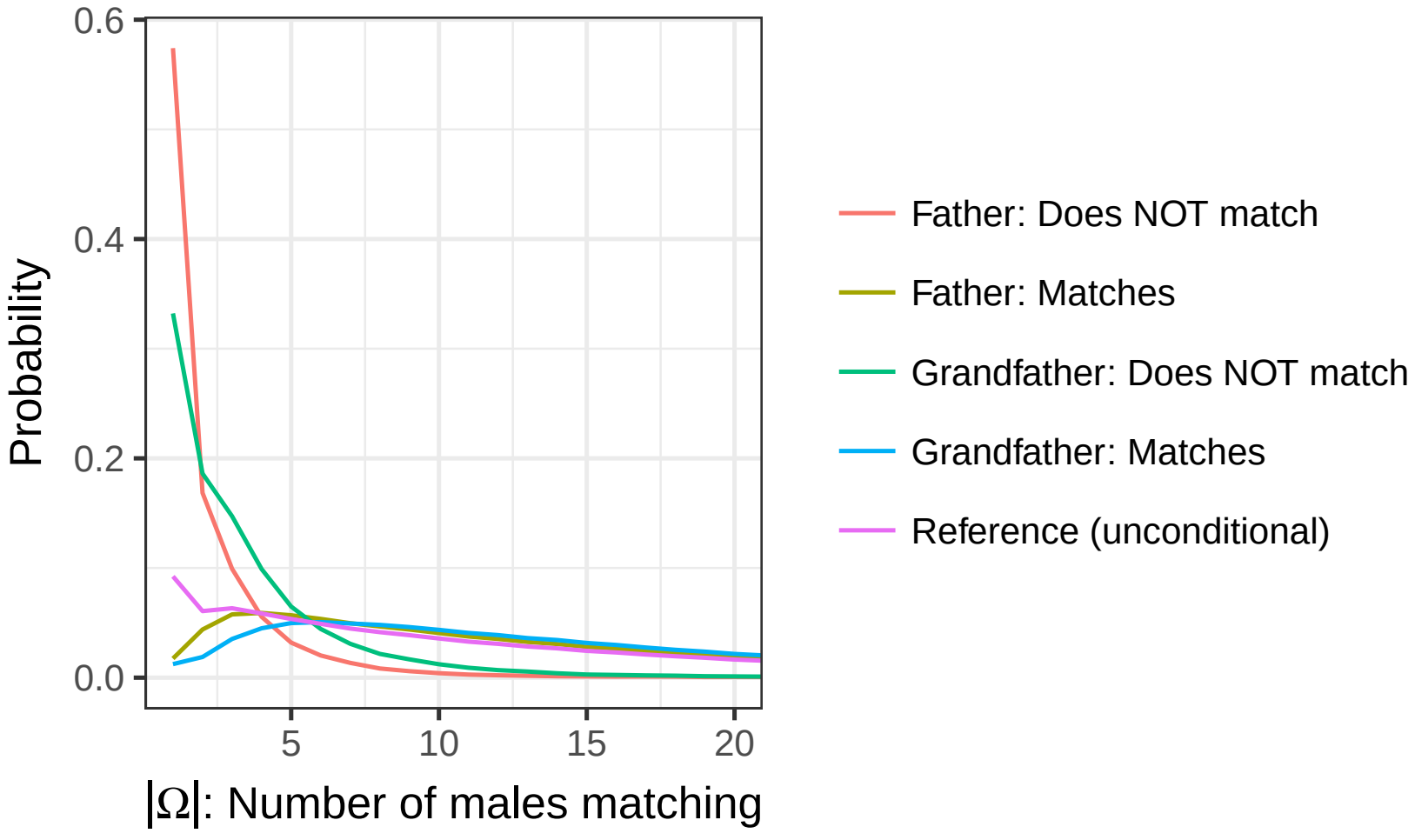
Extensions

- Database information
  - Representative sample?
  - Importance sampling reweighting: number of matching males conditional on a database frequency
  - 95% quantiles:

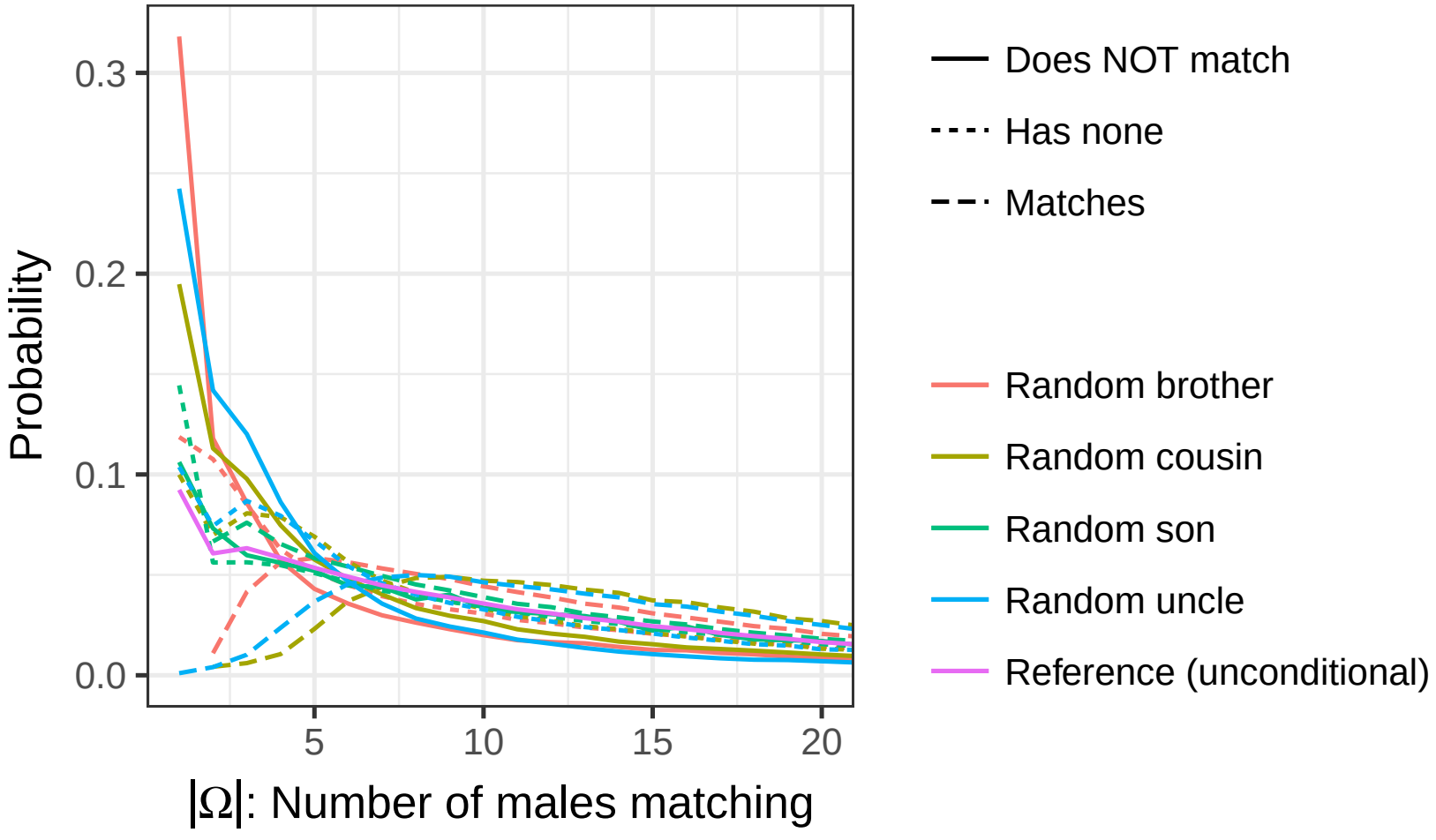
$$\begin{aligned} \text{Unconditional} &: 37\ (\approx 40) \\ \text{DB size 1,000 has } \mathbf{0} \text{ copies} &: 36 \\ \text{DB size 1,000 has } \mathbf{1} \text{ copies} &: 56 \\ \text{DB size 1,000 has } \mathbf{2} \text{ copies} &: 74 \end{aligned}$$

- Close paternal relatives
- Mixtures

Close paternal relatives



Close paternal relatives



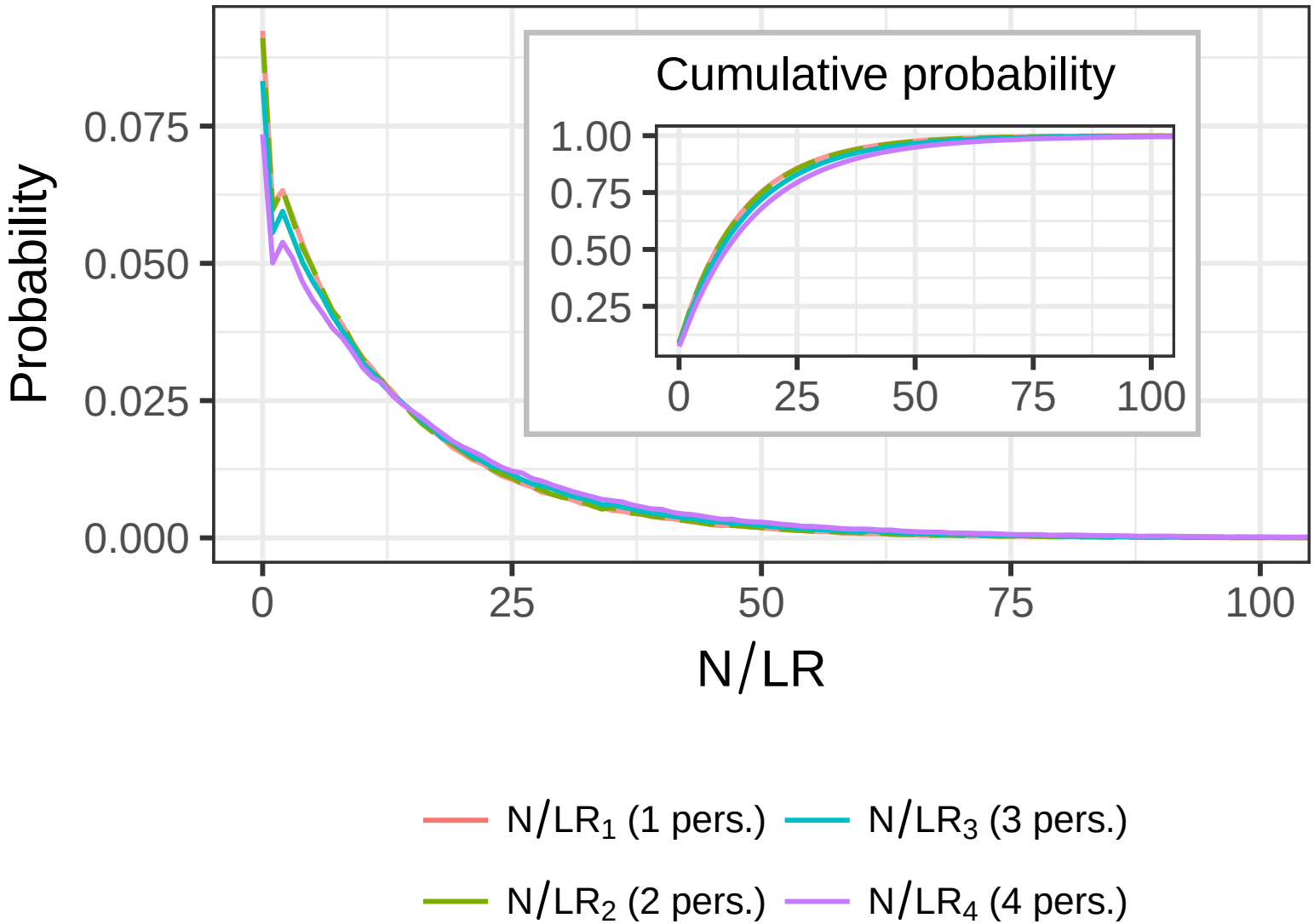
Mixtures

$$\mathrm{LR}_2 = \frac{P(m \mid H_p)}{P(m \mid H_d)} = \frac{n_u/N}{\sum_{(r,s)=m}(n_r/N)(n_s/N)} = \frac{Nn_u}{\sum_{(r,s)=m} n_r n_s}$$

$$N/\mathrm{LR}_2 = \frac{\sum_{(r,s)=m} n_r n_s}{n_u}$$

- Q: Queried contributor
- q: Profile of Q
- m: Observed mixture
- N: population size
- n_a: Population count of profile a
- $u = m - q$: profile of U, the unknown in H_p ($H_p : Q + U$)
- (r, s): Pairs that make up m ($H_d : R + S$)

Mixtures


Probability
0.075
0.050
0.025
0.000
0
25
50
75
100
N/LR
Cumulative probability
1.00
0.75
0.50
0.25
0
25
50
75
100
N/LR1 (1 pers.)
N/LR3 (3 pers.)
N/LR2 (2 pers.)
N/LR4 (4 pers.)

Mixtures

- Inclusion in a (2 pers.) Y-mixture: Almost as strong evidence as a single-contributor evidence profile.

Number of (r, s) sets	Count	%
0	15	0.00%
1	482,896	96.58%
2	16,411	3.28%
3-49	678	0.14%
≥ 50	0	0%

$$N/\mathrm{LR}_2 = \frac{\sum_{(r,s)=m} n_r n_s}{n_u} \quad \underset{=}{|\{(r,s)=m\}|=1} \quad \frac{n_u n_q}{n_u} = n_q = N/\mathrm{LR}_1$$

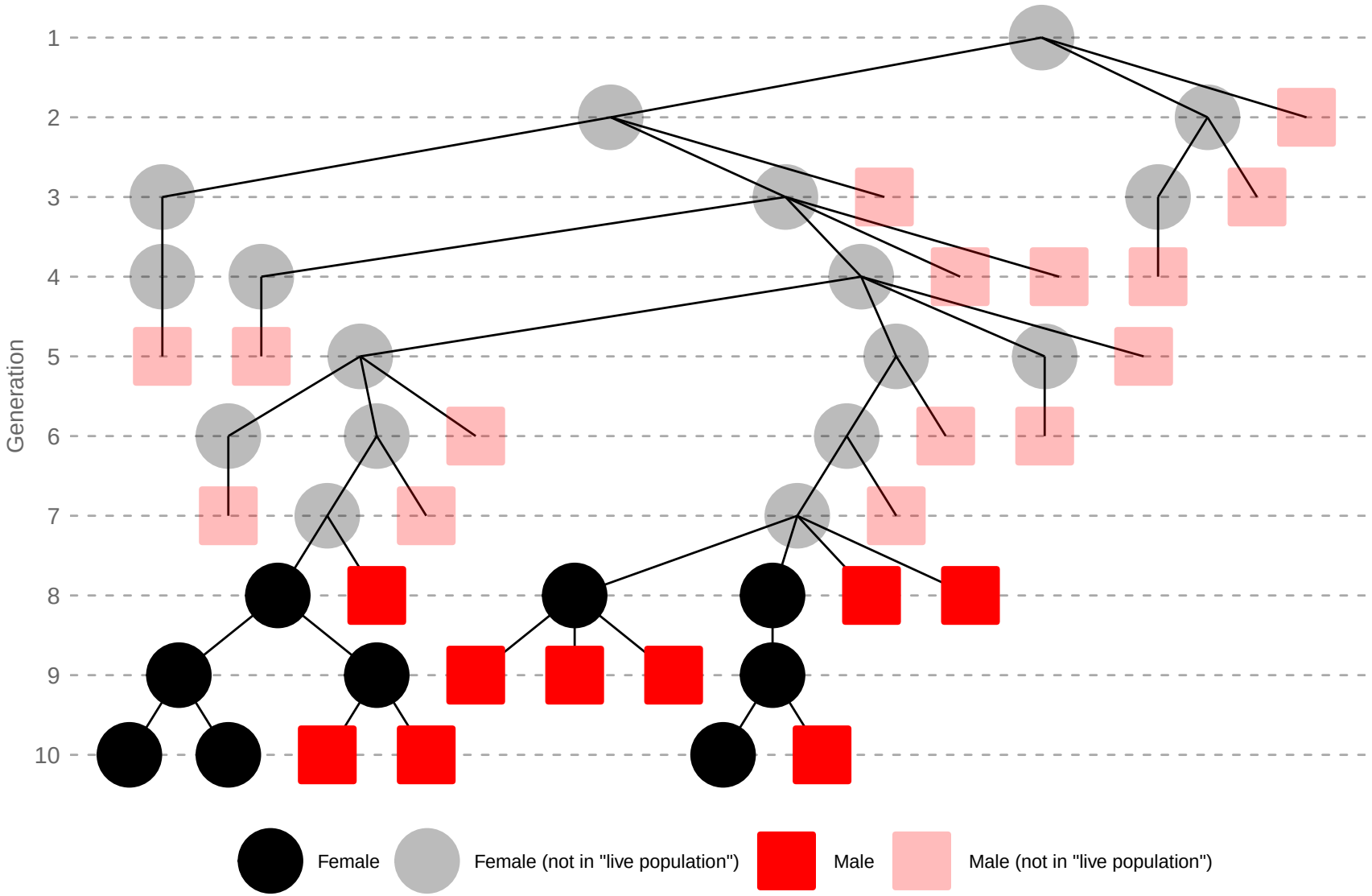
Simulating Y-profiles: Malan

- Open source R package with C++ code through Rcpp
  - `https://github.com/mikldk/malan`
  - `http://joss.theoj.org/papers/10.21105/joss.00684`
- 2017: "How convincing is a matching Y-chromosome profile?" (PLoS Genetics)
  - `https://doi.org/10.1371/journal.pgen.1007028`
- "Y-profile evidence: close paternal relatives and mixtures"
  - `https://www.biorxiv.org/content/early/2018/07/20/373423`

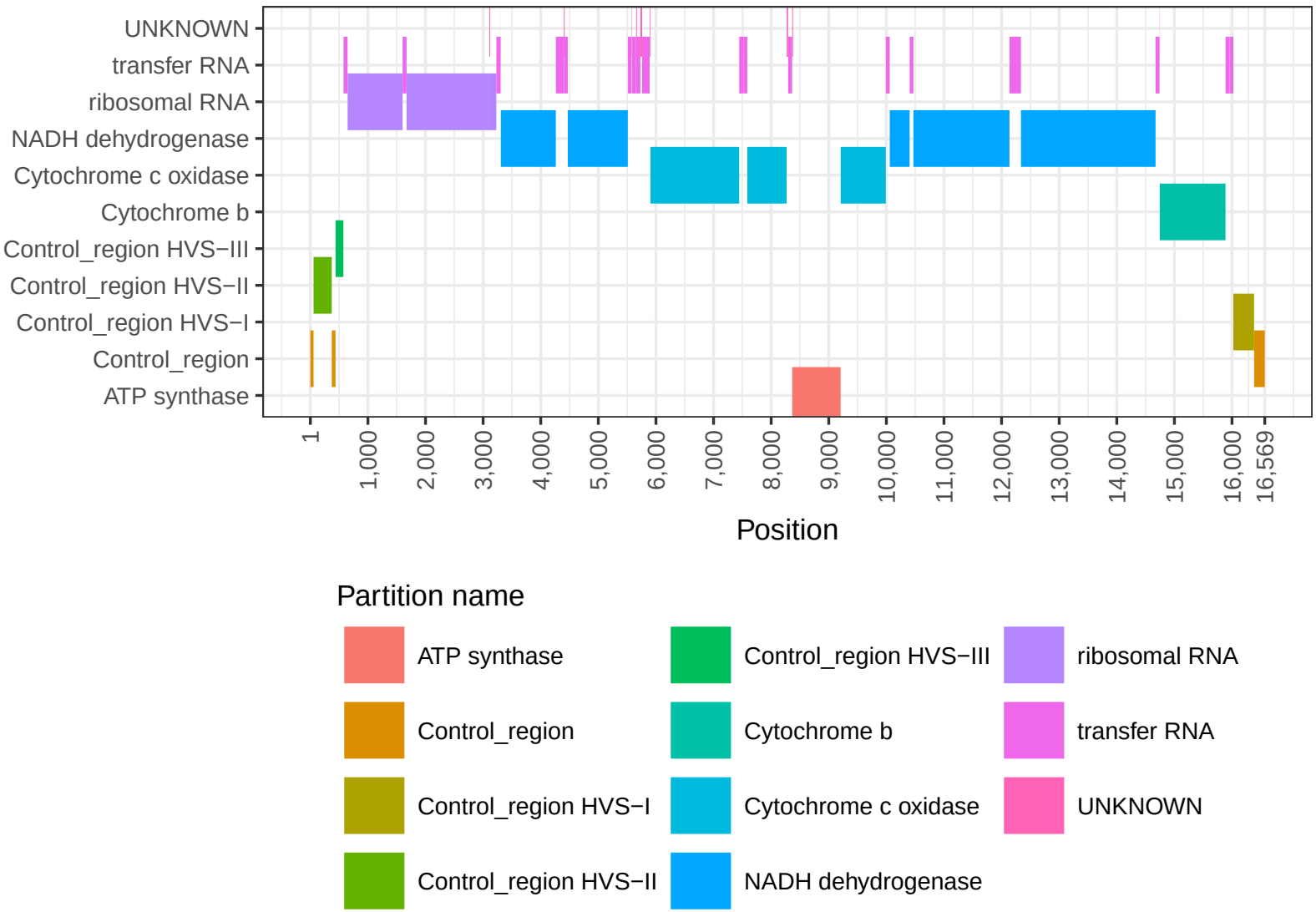
Mitochondrial genomes (mitogenomes)

- Whole mitochondrial genome (*mitogenome*)
- Similar to Y-profiles, but lower mutation rate
- Different genetics
- Mutation rates (phylotree/pedigree)

Simulating mitogenomes



Mutation schemes



Mutation schemes

Number of sites:

Region	Rieux (2014)	Översti (2017)
HVS1 + HVS2	698	1,122
PC1 + PC2	7,565	—
PC3	3,776	—
rRNA + tRNA	4,031	—
—		
Mitogenome	16,070	16,494

Mutation schemes

- Mutation rate per site class
- Generation time: 25 years

Region	Rieux (2014)	Översti (2017)
Entire mitegenome	0.0110 (1:90)	0.0135 (1:74)

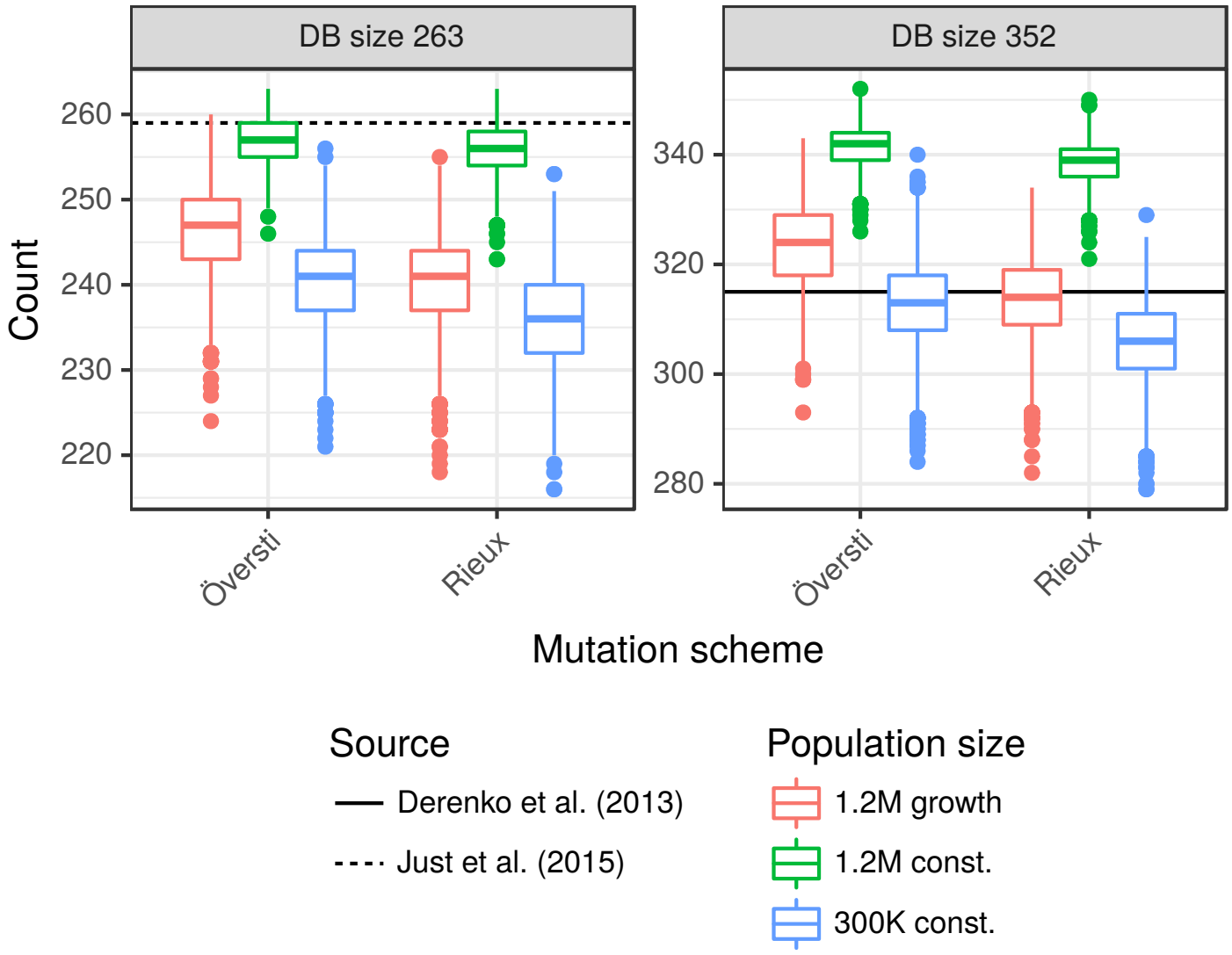
- Sample from 95% highest posterior density interval for each class

Population sizes

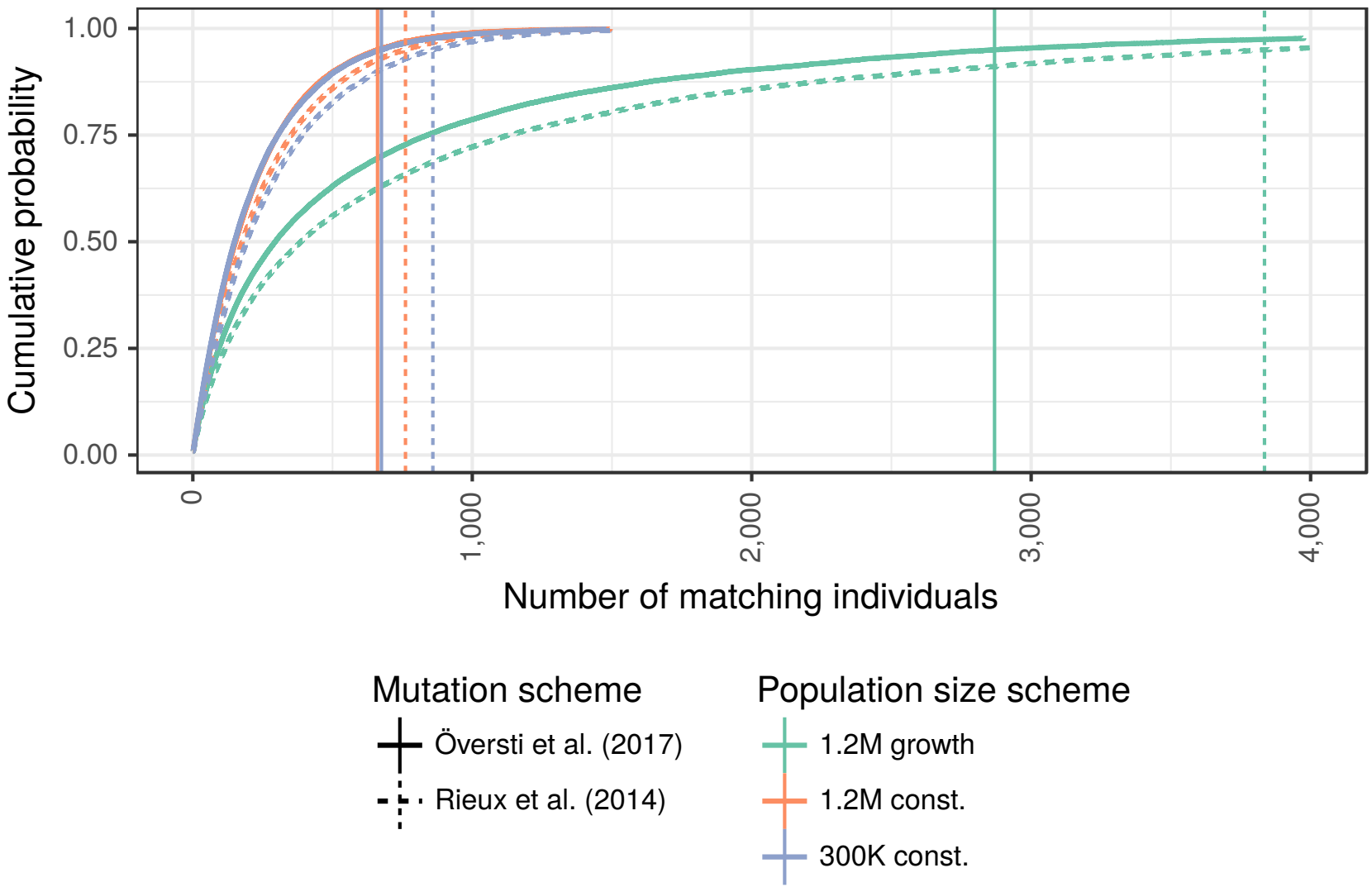
Three demographic scenarios yielding a live population of **300K**, **1.2M**, and **1.2M**:

- Constant-size Wright-Fisher populations
  - **50K** females per generation for 1,200 generations; live population of $2 \times 3 \times 50{,}000 = \mathbf{300K}$
  - **200K** females per generation for 1,200 generations; live population of $2 \times 3 \times 300{,}000 = \mathbf{1.2M}$
- Growth
  - 10,000 for 1,000 generations
  - Growth: 2% per generation for 150 generations
  - Final generation with 200K females
  - Live population of $2 \times 3 \times 300{,}000 = \mathbf{1.2M}$

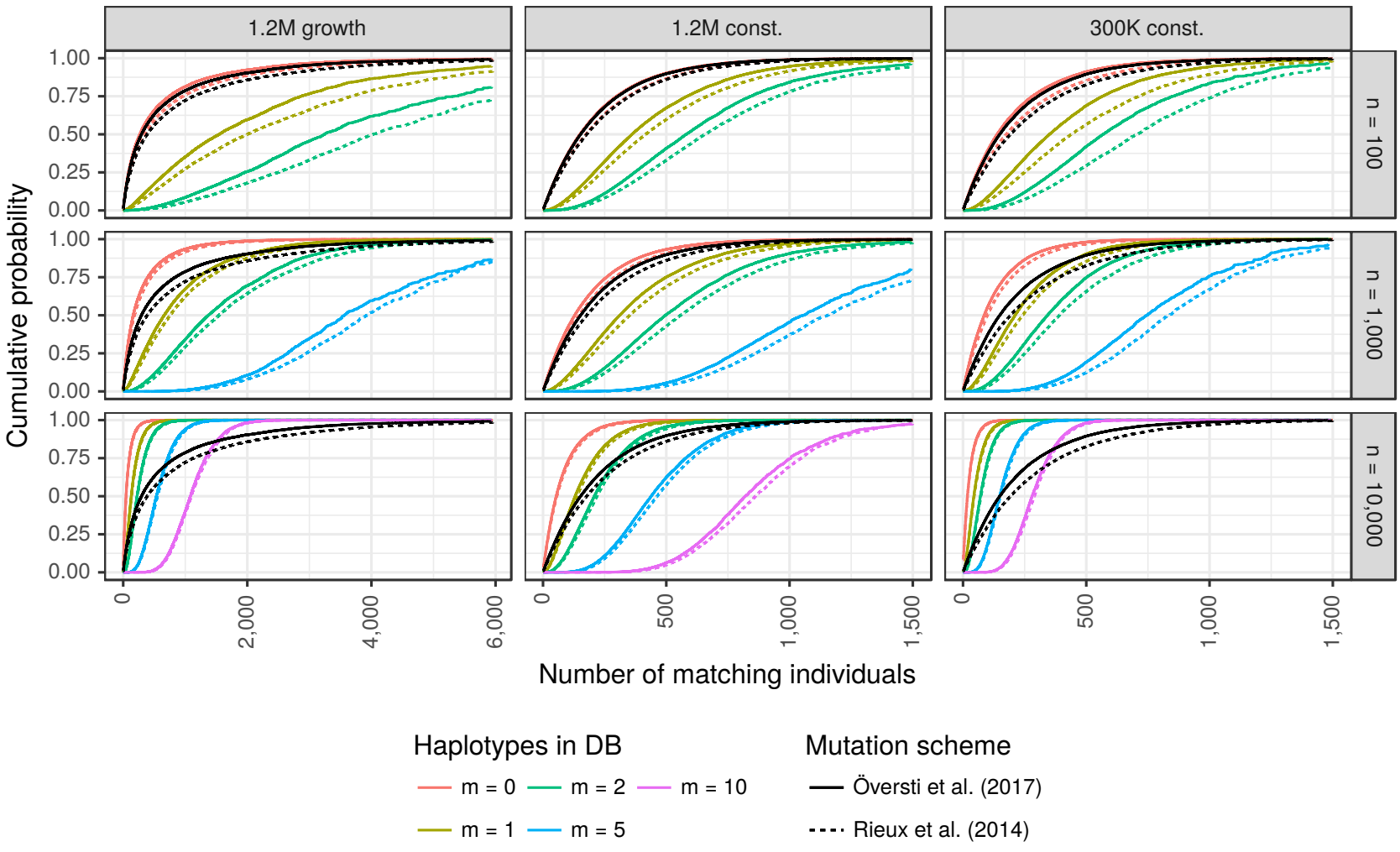
Comparing to real DBs – number of different mitogenomes



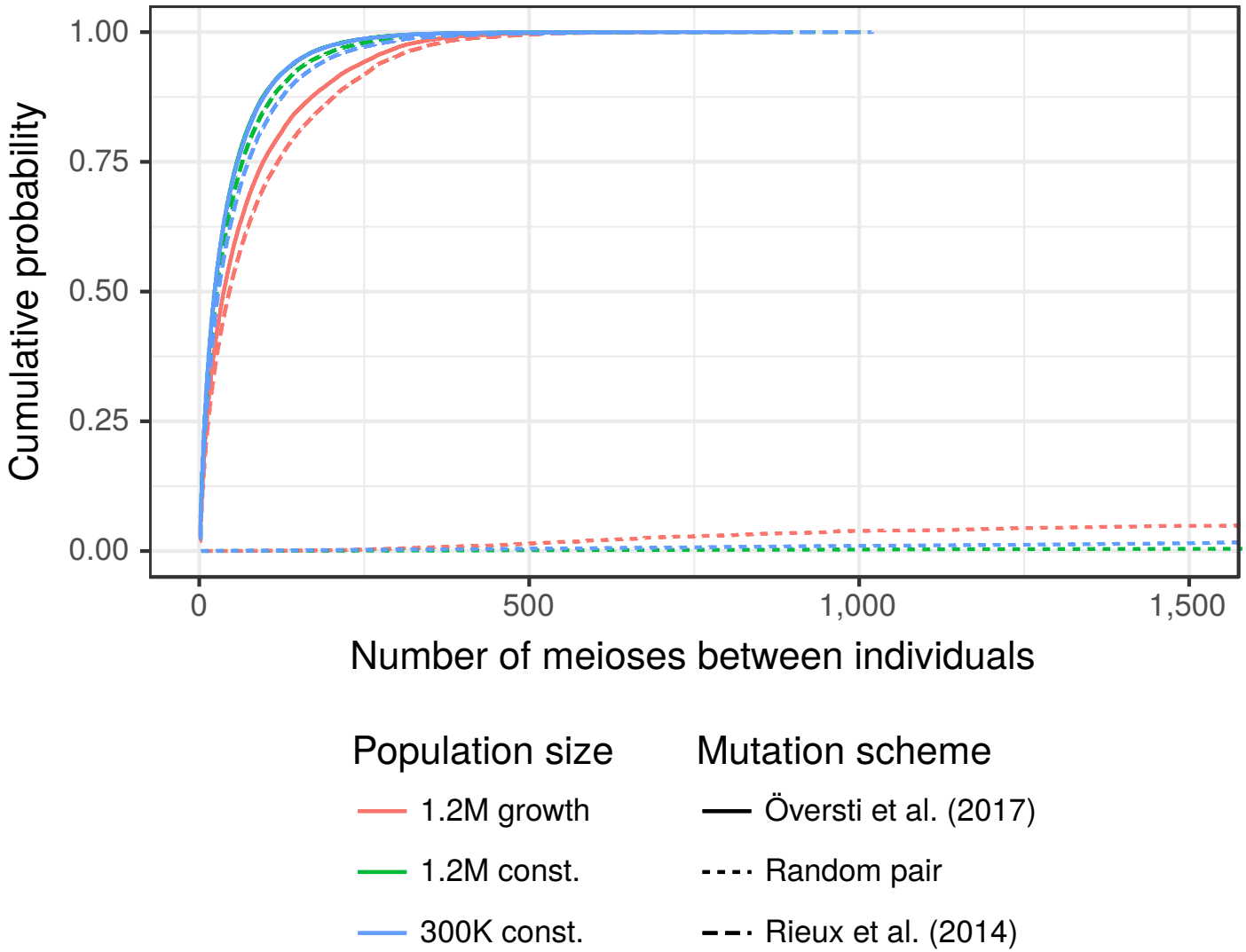
Mitogenomes



Mitogenomes



Mitogenomes



Mitogenomes

- Not as few matching individuals as with modern Y-profiles
- Simulation tool for further research
- Exploit database information?
  - DBs randomly sampled?

Simulating mitogenomes: Mitolina

- Open source R package with C++ code through Rcpp
  - `https://github.com/mikldk/mitolina`
- "How many individuals share a mitochondrial genome?"
  - `https://www.biorxiv.org/content/early/2018/07/23/374686`