How convincing is a matching Y-chromosome profile?

ISFG2017
27th Congress of the International Society for Forensic Genetics

Friday, September 1, Session 5

Mikkel M Andersen and David J Balding

AALBORG UNIVERSITY
DENMARK

Motivation

YfilerPlus match (27 loci).

What can we say?

Results for YfilerPlus match

All (YfilerPlus) haplotypes are rare. In fact:

- The **matching males are** **very likely** (prob. $\geq$ 95%) to be
  - **less than 40 in number** (not dependent on population size)
  - **less than 20 meioses from the suspect**
    - may be well beyond the known relatives of the suspect
- The matching boys or men could also be similar to the suspect in ethnic identity, language, religion, physical appearance, and place of residence

Reporting the number of matching males

- Forensic weight-of-evidence is often best quantified using a **likelihood ratio; we support that** approach in general
- Difficult/impossible: Match probability for Ychr **depends strongly on number of meioses** from queried contributor Q to the particular individual X
- **Report the number of males with matching Y profiles** (an estimate of); has been recommended in the past for autosomal DNA profiles
- In mid-1990s (not as rare autosomal DNA profiles), the England and Wales Court of Appeal recommended this instead of a match probability [Steele and Balding, 2015]
- Recommendation was followed until autosomal match probabilities became too small for the approach to be helpful to jurors

Reporting the number of matching males
Notes

- **Older Y-profiling kits** with lower profile mutation rate or partial Y-profiles: May be appropriate to use a **standard match probability approach** as it is **less sensitive to number of meioses** and there will be **many matching individuals** in the population

Documentation

Documentation of

Match probability depends strongly on number of meioses between Q and X.

Mutations

PowerPlex Y23 data from Purps J, Siegert S, et al. (2014):

	$m = 1$ *singletons*	$m = 2$ *doubletons*	$m = 3$	$m = 4$	$m \geq 5$
n_m	18,226	531	64	16	12
n_m/n	**96.7 %**	2.8 %	0.3 %	0.08 %	0.06 %

Mutation data from YHRD (Jan 19, 2017):

Kit	**Markers**	**Probability of at least one mutation** (per meiosis per generation)
Yfiler	17	**4.4 %**
PowerPlex Y23	23	**8.3 %**
YfilerPlus	27	**13.5 %**

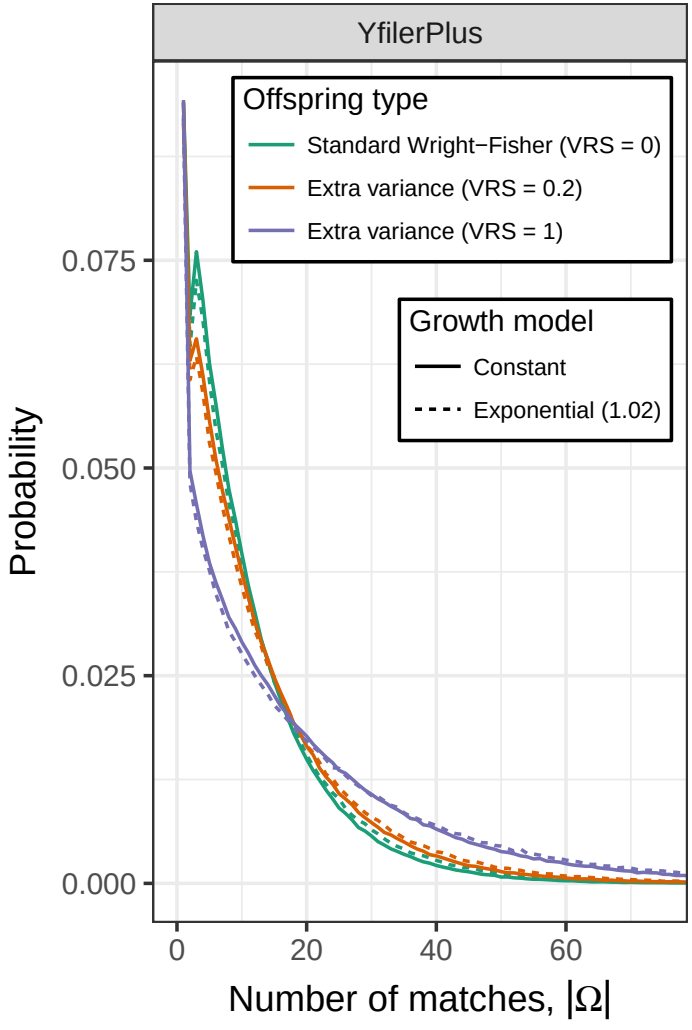
DYS385 and DYF387S1 were split into two loci each.

Simulation study

- **Big simulation study**
  - $N \in \{10^6, 10^5\}$
  - $G = 250$
  - Each density estimated using 500,000 simulated suspects and information about their matches.
- **Various population parameters**
  - Growth rate (1 and 1.02)
  - Variance in reproductive success, VRS (0, 0.2, 1)
    - Wright-Fisher (growth rate 1 and VRS 0)
- **Live population: last three generations**

Results

Distribution of number of matches

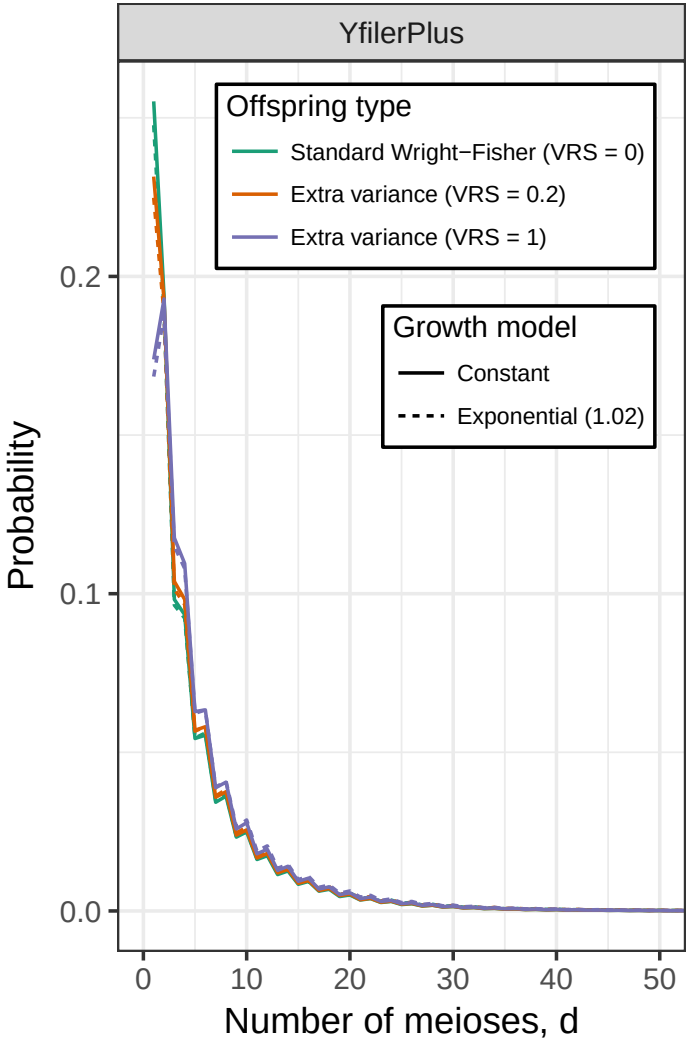


Kit	Growth rate	VRS	Median	95% quantile
YfilerPlus	1	0	8	32
YfilerPlus	1	0.2	9	37
YfilerPlus	1	1	13	59
YfilerPlus	1.02	0	8	35
YfilerPlus	1.02	0.2	9	41
YfilerPlus	1.02	1	14	66

$$P(\#\text{ of matches} \leq \mathbf{37}) \geq 0.95$$

Results

Distribution of meiotic distance



Kit	Growth rate	VRS	Median	95% quantile
YfilerPlus	1	0	3	17
YfilerPlus	1	0.2	3	18
YfilerPlus	1	1	4	18
YfilerPlus	1.02	0	3	19
YfilerPlus	1.02	0.2	3	19
YfilerPlus	1.02	1	4	20

$$P(\#\text{ meioses to matches} \leq \mathbf{18}) \geq 0.95$$

Results

Distribution of number of matches with database information

Use of database information:

- All haplotypes are rare (YfilerPlus, ...)
- Most profiles are absent from a database, most/all in the database occur only once
- Profiles present/absent from the **database is largely ”noise“** and not very informative (too many loci, structure difficult to find)
- More information about population frequencies from mutation rates and population model
- Database info can refine this through conditioning – usually not much impact
- One advantage: **take into account a database count of 0** – tells us little as we know a priori that most profiles won't be in the database

Results

Distribution of number of matches with database information

YfilerPlus for VRS = 0.2 and constant population size ($N = 10^5$):

n	m	Median	95% quantile
-	-	9	37
100	0	9	37
100	1	21	58
100	2	33	77
1,000	0	9	36
1,000	1	20	56
1,000	2	32	74
10,000	0	6	27
10,000	1	15	41
10,000	2	23	55

$$P(\#\text{ matches} \leq \mathbf{37}) \geq 0.95$$

$$P(\#\text{ matches} \leq \mathbf{36} \mid \text{db size 1,000 has 0 copies}) \geq 0.95$$

$$P(\#\text{ matches} \leq \mathbf{56} \mid \text{db size 1,000 has 1 copies}) \geq 0.95$$

$$P(\#\text{ matches} \leq \mathbf{74} \mid \text{db size 1,000 has 2 copies}) \geq 0.95$$

Summary

- The distributions vary (slightly) with population parameters (VRS and population growth rate)
- Parameters cannot be known exactly for a specific court case
- But the distributions are **insensitive to** major changes in **parameter values** relative to the precision necessary for a juror's reasoning
- Number of matching individuals in the population: **40 or 50 or 60** is unlikely to have much impact on a juror's decision, but **orders of magnitude may well be important**

Presentation in court

E.g. for YfilerPlus:

> *"A Y-chromosome profile was recovered from the crime scene. Mr Q has a matching Y profile and so is not excluded as a contributor of DNA.* ***Using population genetics theory and data, we conclude that the number of males in the population with a matching Y profile is probably less than 20, and is very unlikely (probability*** $<$ ***5%) to exceed 40****. These men or boys span a wide range of ages and we don't know where they live.* ***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****. Since these individuals share paternal-line ancestry with Q, they could also be* ***similar to Q in ethnic identity, language, religion, physical appearance and place of residence****."*

Presentation in court

With database frequency information, e.g.

> *"The Y profile of Q was not observed in a database of 1,000 profiles. Because the database does not represent a scientific random sample and because paternal-line relatives may tend to be clustered in geographic and social groups that are not well sampled in the database, it is difficult to interpret this information.* ***If the database were a random sample from the population, its effect would be to reduce the 95% upper limit on the number of matching males from 40 to 39****."*

(Impact of this information may be minimal, and it could perhaps be omitted except that courts may be expecting to hear database information.)

Presentation in court

Depending on the circumstances of the case, a judge might further instruct members of the jury:

> *"If you consider that there may be up to 40 males of different ages with a Y profile matching that of Q, and that these males may tend to resemble Q in some characteristics more than random members of the population, your task is to decide whether all the evidence that has been presented to you is enough to convince you that Q is the source of the crime scene DNA, and not one of these other males with the same Y profile."*

Future work

- Conditional distributions for known profiles of relatives
- Mixtures
- mtDNA whole genome

Summary

- "How Convincing Is A Matching Y-Chromosome Profile?": `www.biorxiv.org/content/early/2017/08/28/131920`
- R package `malan` used to perform all simulations: `www.github.com/mikldk/malan`

Thank you for your attention