Y Chromosomal STR Markers: Assessing Evidential Value

10th International Y Chromosome workshop
Berlin, May, 2016

Mikkel Meyer Andersen, Poul Svante Eriksen and Niels Morling

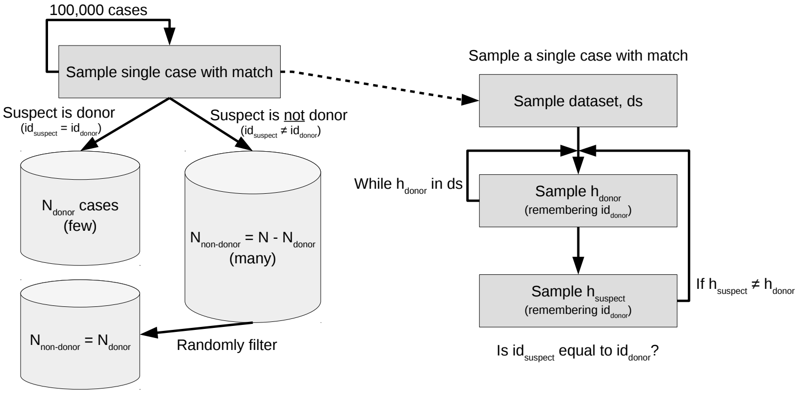
AALBORG UNIVERSITY
DENMARK

LR and donorship

- Profile from donor to crime scene stain, h_{donor}
- Profile from suspect, $h_{\text{suspect}} = h_{\text{donor}}$ (we have a match)
- Reference database
- Decision problem: Is the suspect the donor? Answer based on h_{suspect} and reference database
- Simple case: $LR = 1/\text{match probability} = 1/\text{population frequency}$
- Decision problem tried solved by LR
- Higher LR, more evidence that the suspect is the donor

Simulate cases with known donor

- Simulate population (simple) of approx 2,000,000 individuals
  - FW, 100,000 in 300 generations w/ growth rate 1.01
  - 7 loci, neutral single-step mutation model ($\mu = 0.003$)


100,000 cases
Sample single case with match
Suspect is donor
($id_{suspect} = id_{donor}$)
Suspect is not donor
($id_{suspect} \neq id_{donor}$)
N_{donor} cases
(few)
$N_{non-donor} = N - N_{donor}$
(many)
$N_{non-donor} = N_{donor}$
Randomly filter
Sample a single case with match
Sample dataset, ds
While h_{donor} in ds
Sample h_{donor}
(remembering id_{donor})
Sample $h_{suspect}$
(remembering id_{donor})
If $h_{suspect} \neq h_{donor}$
Is $id_{suspect}$ equal to id_{donor}?

Estimators problem

Database size n:

- n_1: singletons
- n_2: doubletons
- κ: n_1/n

Estimators:

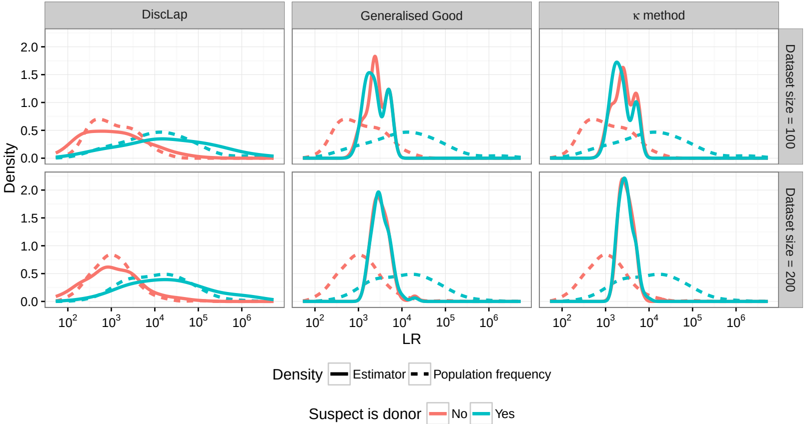
- Kappa (Brenner, 2010): $LR = n/(1-\kappa) = n \cdot \frac{n}{n-n_1} > n$
- Generalised Good (Cereda, 2015): $LR = (n \cdot n_1)/(2 \cdot n_2) = n \cdot \frac{n_1}{2n_2}$
- Discrete Laplace (Andersen, 2013): Statistical model using genetic information

  ```
  fit <- disclapmix(db, 5L)
  ```

  $LR = 1/$`predict(fit, `h`)`

LR distributions

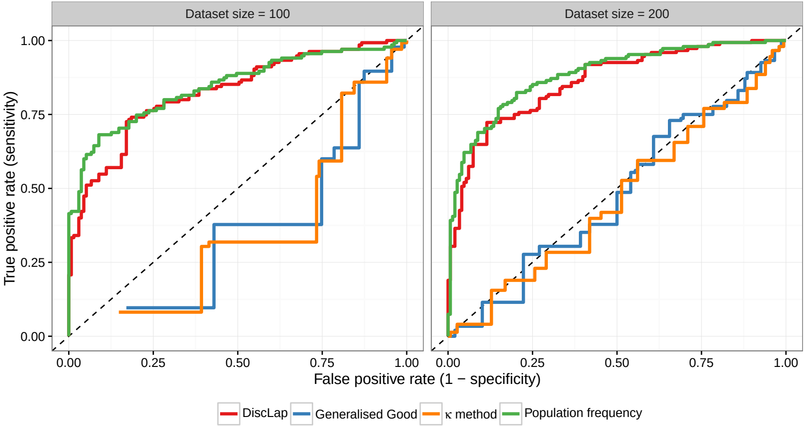
- Dataset size $n = 100$: $N_{\text{donor}} = N_{\text{non-donor}} = 135$
- Dataset size $n = 200$: $N_{\text{donor}} = N_{\text{non-donor}} = 148$
- [Larger dataset, greater *LR*, and greater $P(\text{suspect} = \text{donor} \mid \text{match})$ hence more cases where it happens.]



ROC curve of decision problem

- $LR_{\text{case}} \geq LR_{\text{threshold}}$: Suspect is donor
- $LR_{\text{case}} < LR_{\text{threshold}}$: Suspect is not donor

For all possible $LR_{\text{threshold}}$'s:



Best *LR* threshold

Dataset size 200

$$LR_{\text{threshold}} = \text{argmax}_t \left(\text{sensitivity}(t) + \text{specificity}(t)\right)$$

$$LR_{\text{threshold}}(r) = \text{argmax}_t \left(\text{sensitivity}(t) + r \cdot \text{specificity}(t)\right)$$

Dataset size 200:

Estimator	r	$LR_{\text{threshold}}(r)$	TP	TN	FP	FN	FPR	FNR
DiscLap	1	3,440	107	131	17	41	0.11	0.28
Generalised Good	1	2,630	107	52	96	41	0.65	0.28
Kappa	1	2,550	88	65	83	60	0.56	0.41
Population frequency	1	2,220	122	118	30	26	0.20	0.18
DiscLap	50	113,650	28	148	0	120	0.00	0.81
Generalised Good	50	20,570	0	148	0	148	0.00	1.00
Kappa	50	20,570	0	148	0	148	0.00	1.00
Population frequency	50	104,770	11	148	0	137	0.00	0.93

Discussion

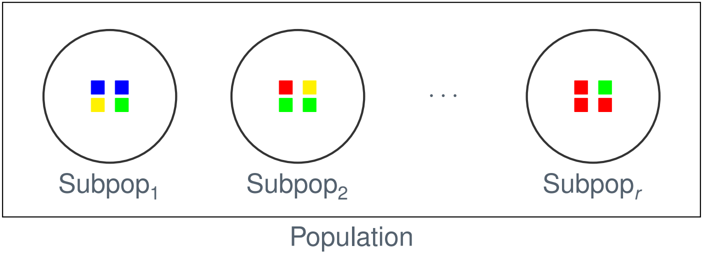
Validation of estimators:

- Fisher-Wright population too simple
- Single-step mutation model too simple
- More realistic reference population simulation schemes agreed upon by community

Discrete Laplace workings in progress:

- Working on quantifying statistical error of estimate
- C++ library for faster estimation
- ’`fit <- disclapmix(db)`’ (more automatic and user-friendly; maybe using several number of clusters, e.g. a weighted average of 3-5 best)

Population substructure



Coloured squares represent haplotypes.

Random man (donor) and suspect belong to same subpopulation: Expected to share a haplotype more often than a random database sample from the whole population would represent.

θ (theta) correction seeks to quantify this.

Match probability

H_d: 'A random man **– that originate from the same subpopulation as the suspect –** left the Y-chromosome DNA in the crime stain.'

- Reference database from this subpopulation exists
  - Subpopulation is now the population
  - Use this reference database and no θ correction!
- Reference database from population containing subpopulation (as well as other subpopulations, and structure unknown):
  - One approach (the Balding-Nichols model):
    $P(E \mid H_d) \overset{BN}{=} \theta + (1-\theta)p_h$
  - θ (theta) $(0 \leq \theta \leq 1)$
    - Population parameter (related to how much haplotype frequencies vary in different subpopulations)
    - Most simple model – many extensions possible
  - p_h: Population frequency of h $(0 \leq p_h \leq 1)$

Match probability

Note, that

$$P(E \mid H_d) \overset{BN}{=} \theta + (1-\theta)p_h \geq \theta$$

and

$$P(E \mid H_d) \overset{BN}{=} \theta + (1-\theta)p_h \geq p_h$$

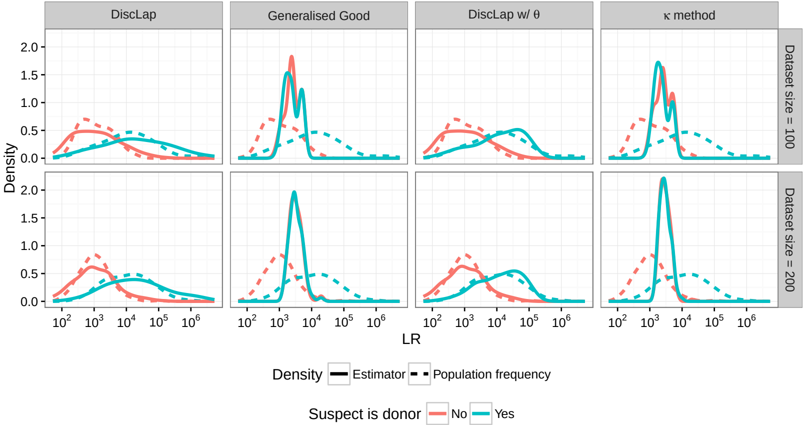
- If p_h is really small compared to θ: $\theta + (1-\theta)p_h \approx \theta$
- If p_h is really large compared to θ: $\theta + (1-\theta)p_h \approx p_h$

	$p_h = 1/100{,}000$ = 0.00001	$p_h = 1/100 = 0.01$
$\theta = 0.001$	$P(E \mid H_d)$ = 0.0010099	$P(E \mid H_d)$ = 0.01099
$\theta = 0.003$	$P(E \mid H_d)$ = 0.0030099	$P(E \mid H_d)$ = 0.01297

Best *LR* threshold

Dataset size 200

$$\theta = 0.00001 = 10^{-5} = 1/10^5$$



Best *LR* threshold

Dataset size 200

$$\theta = 0.00001 = 10^{-5} = 1/10^5$$

Estimator	r	$LR_{\text{threshold}}(r)$	TP	TN	FP	FN	FPR	FNR
DiscLap	1	3,440	107	131	17	41	0.11	0.28
DiscLap w/ theta	1	3,440	107	131	17	41	0.11	0.28
Generalised Good	1	2,630	107	52	96	41	0.65	0.28
Kappa	1	2,550	88	65	83	60	0.56	0.41
Population frequency	1	2,220	122	118	30	26	0.20	0.18
DiscLap	50	113,650	28	148	0	120	0.00	0.81
DiscLap w/ theta	50	54,630	26	148	0	122	0.00	0.82
Generalised Good	50	20,570	0	148	0	148	0.00	1.00
Kappa	50	20,570	0	148	0	148	0.00	1.00
Population frequency	50	104,770	11	148	0	137	0.00	0.93

Thank you for your attention

(Slides soon available at `http://people.math.aau.dk/~mikl/`)