Y Chromosomal STR Markers: Assessing Evidential Value

FOSW03: Statistical Modelling of Scientific Evidence

Isaac Newton Institute in Probability and Statistics in Forensic Science, Cambridge, Nov 2016

Mikkel Meyer Andersen, Poul Svante Eriksen and Niels Morling

... and many others!

AALBORG UNIVERSITY
DENMARK

Outline

- The discrete Laplace model (quick*ish* recap)
- Comparing match probability estimators
- Population substructure

Motivation

of the discrete Laplace model

Statistical model

$$P(h) \geq 0 \quad \text{and} \quad \sum_{h \in \mathcal{H}} P(h) = 1$$

(Forensic genetic) applications:

- $LR = \frac{P(E|H_p)}{P(E|H_d)}$
- $P(h)$
- $\theta + (1 - \theta)P(h)$
- Mixture deconvolution
- *LR* for mixtures (qualitative/quantitative)
- Cluster analysis (not shown)
- Not a new ad-hoc tool for each task

Model

- Y-STR: Loci not statistically independent
- Our approach: Condition on [something] to obtain independence between loci

The Discrete Laplace model

for Y-STR haplotypes (MM Andersen *et al.*, 2013)

$$f(x; p, \mu) = \frac{1-p}{1+p} \cdot p^{|x-\mu|} \quad \text{for } x \in \mathbb{Z},$$

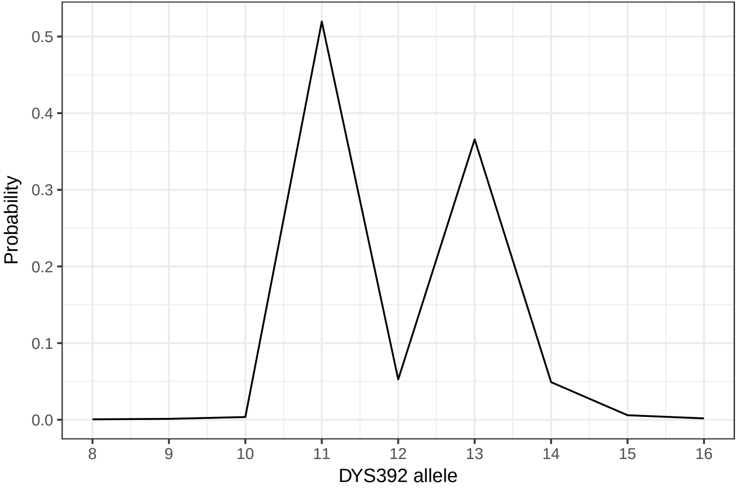
$$P(X = \vec{x} = (x_1, x_2, \ldots, x_r)) = \sum_{j=1}^{c} \tau_j g\left(\vec{x}; \vec{p_j}, \vec{\mu_j}\right) = \sum_{j=1}^{c} \tau_j \prod_{k=1}^{r} f\left(x_k; p_{jk}, \mu_{jk}\right),$$

$$p_{jk} = \exp\left(\alpha_j + \beta_k\right).$$

- Estimation: EM algorithm w/ GLM heavily exploiting structure of design matrix
- Parameter estimation from observations using `R` library `disclapmix`

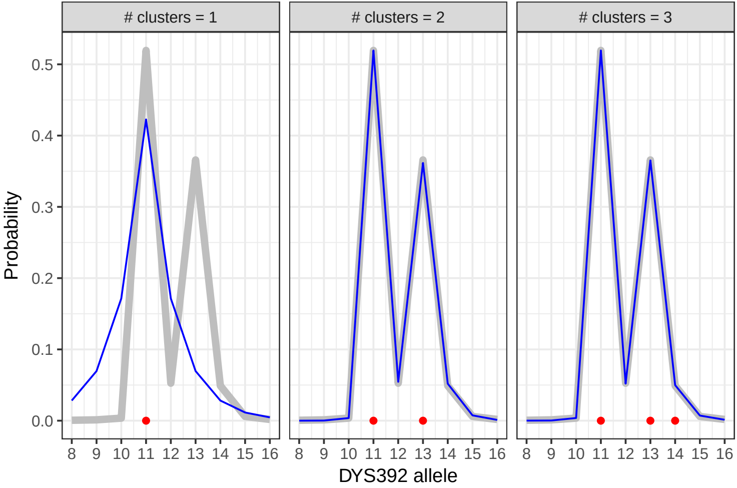
Data and fit

1,692 Germans from Purps (2014) Y23



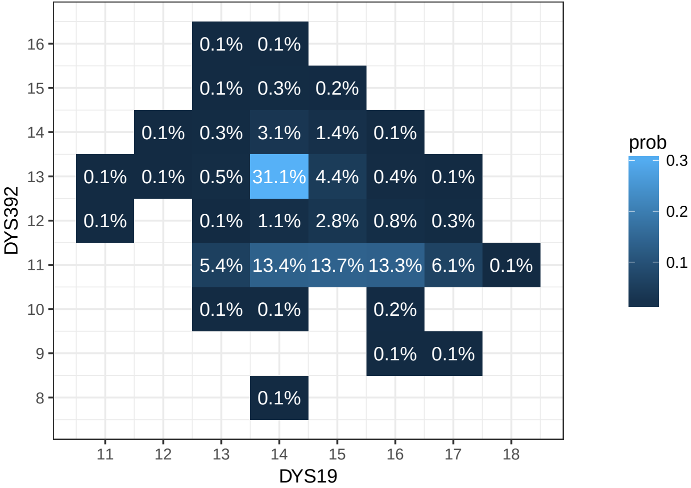
Data and fit

1,692 Germans from Purps (2014) Y23



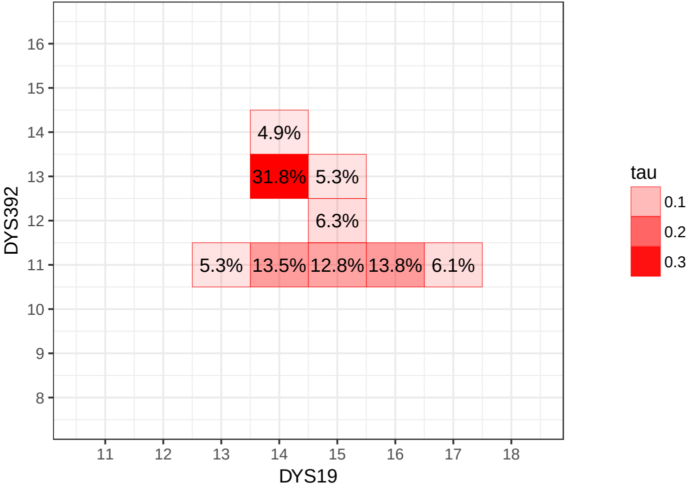
Data and fit

1,692 Germans from Purps (2014) Y23



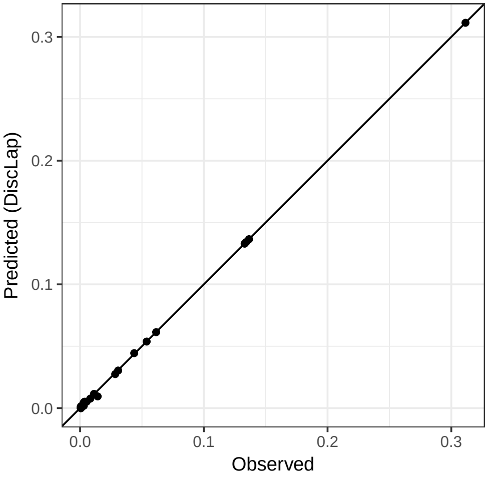
Data and fit

1,692 Germans from Purps (2014) Y23



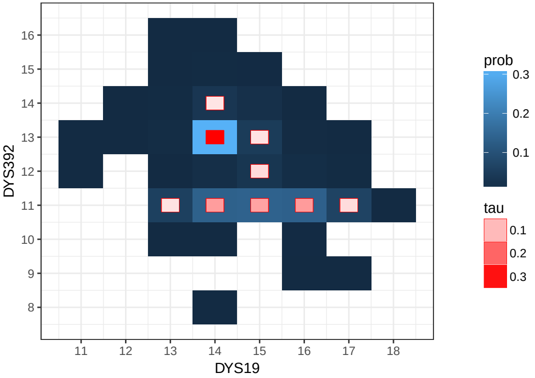
Data and fit

1,692 Germans from Purps (2014) Y23



Data and fit

1,692 Germans from Purps (2014) Y23



- $(\text{rows} - 1)(\text{columns} - 1) = (9 - 1)(8 - 1) = 8 \cdot 7 = 56$
- $(r \cdot c) + (c - 1) + (r + c - 1) = (2 \cdot 9) + 8 + (2 + 8) = 36$
  - $p_{jk} = \exp(\alpha_j + \beta_k)$, $\beta_1 = 0$

Estimator validation

Estimator validation

- Single source stain
- No errors
- No peak heights
- Compare/validate/investigate estimators estimating different population quantities
  - Data reduction

LR and donorship

- Case
  - Profile from donor to crime scene stain, h_{donor}
  - Profile from suspect, h_{suspect}
  - Reference database
- Decision problem: Is the suspect the donor? Tried solved by *LR*
  - Simple case: $LR = 1/\text{match probability} = 1/\text{population frequency}$
- Higher *LR*, more evidence that the suspect is the donor
- Trade-off: Conservative (when possible) vs informative
  - Data reduction
  - Non-match $\Rightarrow LR = 0$ and match $\Rightarrow LR = 1$

Simulate cases

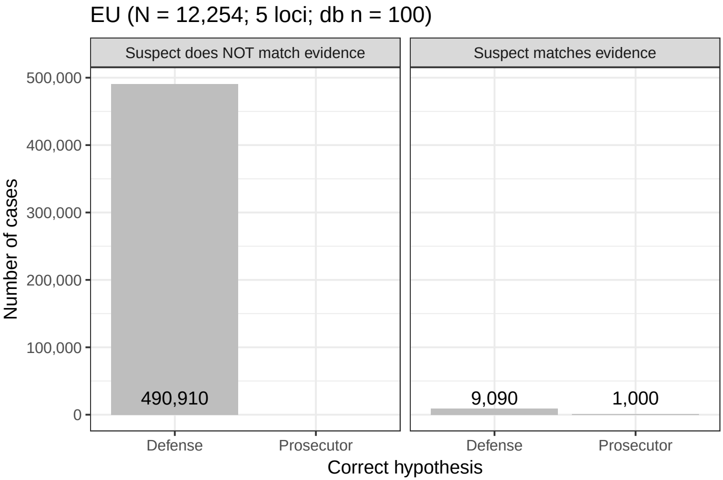
Population:

- EU on 5 loci from Purps (2014) Y23 dataset ($N = 12{,}254$)
  - 2.9% of haplotypes are singletons

Cases:

- Simulate cases under H_p (suspect is the donor), $k_p = 1{,}000$
  - Simulate reference database ($n = 100$)
  - Simulate the suspect's/donor's haplotype
- Simulate cases under H_d (suspect is not the donor), $k_d = 500{,}000$
  - Simulate reference database ($n = 100$)
  - Simulate the suspect's haplotype, h_{suspect}
  - Simulate the donor's haplotype, h_{donor}
  - Often, $h_{\text{suspect}} \neq h_{\text{donor}} \Rightarrow LR = 0$

Simulate cases



Estimators

Database size n:

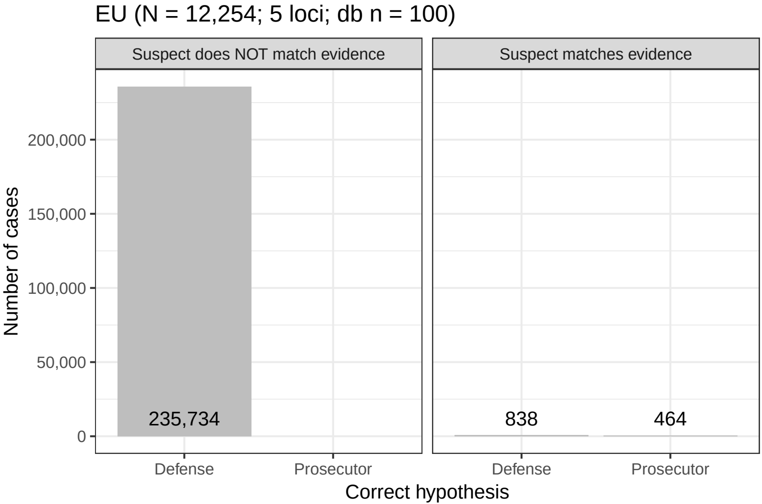
- n_1 = # singletons
- n_2 = # doubletons
- $\kappa = n_1/n$

Estimators:

- Kappa (CH Brenner): $LR_{\text{rare}} = n/(1-\kappa) = n \cdot \frac{n}{n-n_1} > n$
- Generalised Good (G Cereda): $LR_{\text{rare}} = (n \cdot n_1)/(2 \cdot n_2) = n \cdot \frac{n_1}{2n_2}$ also LR for non-rare
- Discrete Laplace (MM Andersen)
- (~~Coalescent: Not included due to computational requirements, could be interesting~~)
- (~~Chinese restaurant (G Cereda): Work on including it is in progress~~)

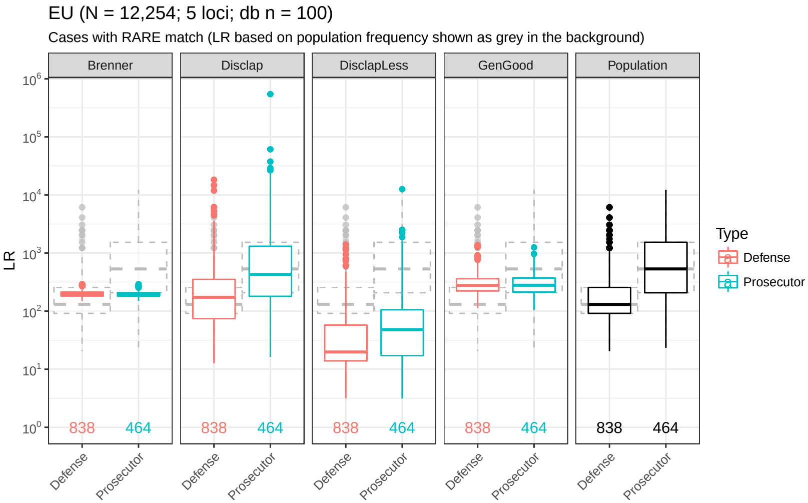
Cases

Rare/unobserved haplotypes



LR distribution

Rare/unobserved haplotypes

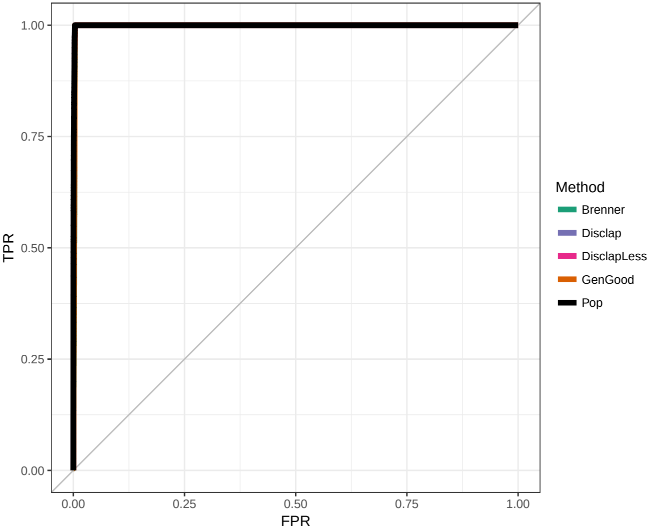


ROC

Rare/unobserved haplotypes

EU (N = 12,254; 5 loci; db n = 100)

All cases with k = 0 of sus.hap. in db. Grey line is the identity line, y = x.

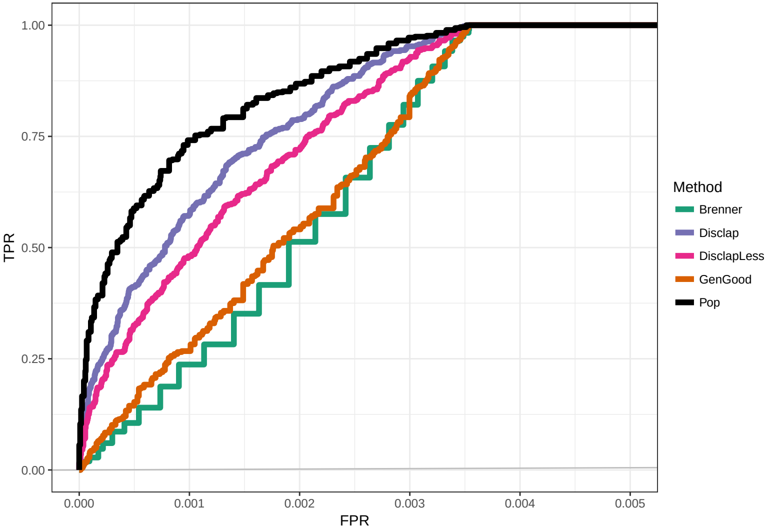


ROC

Rare/unobserved haplotypes

EU (N = 12,254; 5 loci; db n = 100)

All cases with k = 0 of sus.hap. in db. Grey line is the identity line, y = x.

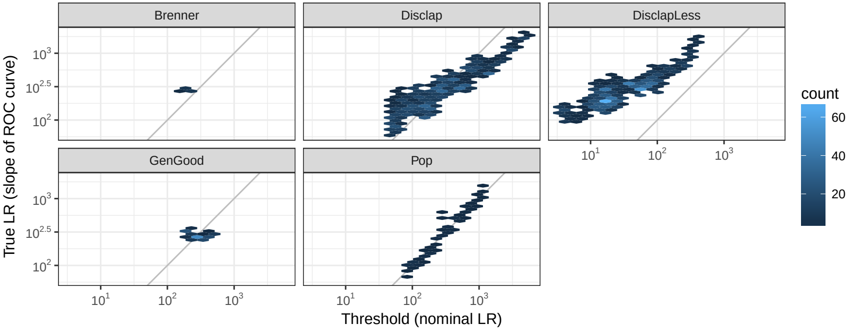


ROC

Rare/unobserved haplotypes

EU (N = 12,254; 5 loci; db n = 100)

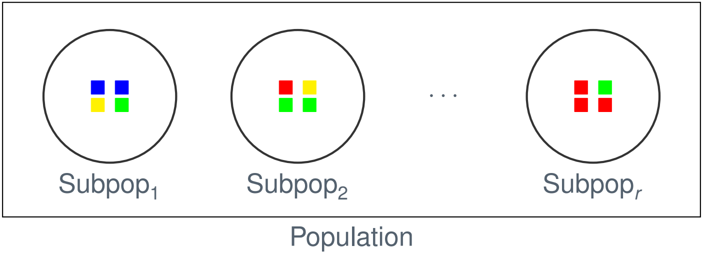
All cases with k = 0 of sus.hap. in db. Grey line is the identity line, y = x.



Slope of the tangent line at a point on the ROC curve gives the *LR* ('True *LR*') for that value/threshold of the test ('Threshold (nominal LR)').

Population substructure

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: quantifies this; a remedy for not knowing the population substructure.

Estimating θ (theta)

Bruce Weir, pers. com. Assumptions apply.

- r: Number of subpopulations
- n_i: Size of reference database from i'th subpopulation ($i = 1, 2, \dots, r$)
- n_{ih}: Number of times haplotype h is observed in reference database from i'th subpopulation

$$m_i = \frac{1}{n_i(n_i - 1)} \sum_h n_{ih}(n_{ih} - 1) \quad \text{and} \quad m_{ij} = \frac{1}{n_i n_j} \sum_h n_{ih} n_{jh}$$

$$m_W = \frac{1}{r} \sum_{i=1}^{r} m_i \quad \text{and} \quad m_B = \frac{2}{r(r-1)} \sum_{i=1}^{r-1} \sum_{j=i+1}^{r} m_{ij}$$

$$\hat{\theta} = \frac{\frac{r-1}{r} \frac{m_W - m_B}{1 - m_B}}{1 - \frac{1}{r} \frac{m_W - m_B}{1 - m_B}} \overset{\text{large } r}{\approx} \frac{m_W - m_B}{1 - m_B}$$

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 with unknown population substructure:
  - One approach (based on 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

Match probability

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

Note, that

$$P^{BN}(E \mid H_d) \geq \theta$$

and

$$P^{BN}(E \mid H_d) \geq p_h$$

- p_h really small compared to $\theta \Rightarrow P^{BN}(E \mid H_d) \approx \theta$
- p_h really large compared to $\theta \Rightarrow P^{BN}(E \mid H_d) \approx p_h$

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

Population substructure: Examples

Example 1: English reference database. We assume no population substructure (haplotype distribution same in the entire population).

- H_d: 'A random Englishman left the Y-chromosome DNA in the crime stain.'
- Use (estimate of) population frequency, p_h, based on English reference database (and no θ correction)

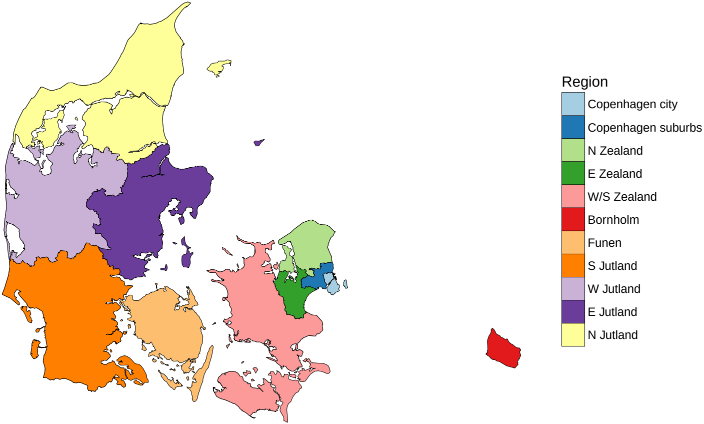
Example 2: English reference database. We assume population substructure (such that haplotype distribution may differ e.g. between regions).

- H_d: 'A random Englishman originating from the same region as the suspect left the Y-chromosome DNA in the crime stain.'
- Use θ correction: $\theta + (1 - \theta)p_h$ with θ estimated in advance using reference databases from comparable regions and (estimate of) population frequency, p_h, based on English reference database

Denmark

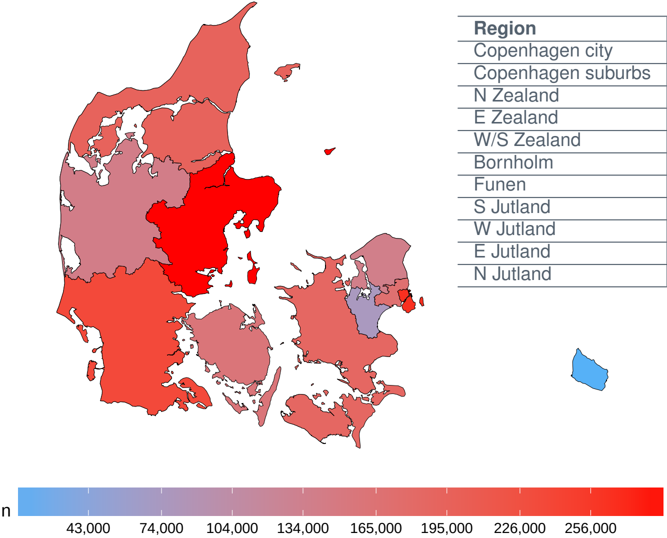
11 NUTS-3 (*nomenclature des unités territoriales statistiques*) regions

- The Danish Family Relations Database: ≈ 9,300,000
- Males: ≈ 4,700,000
- Men, alive, 15-65 years, known last residence: ≈ 1,900,000



Denmark

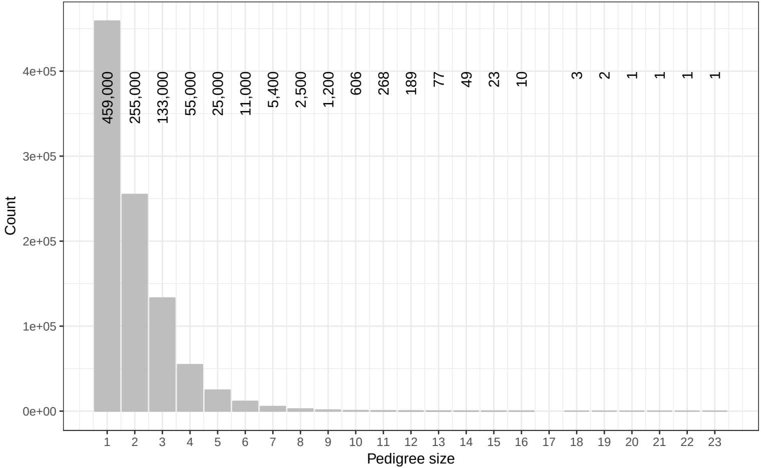
Region population sizes



Region	$n_{\text{males}} \approx$	**Area [km^2]**
Copenhagen city	250,000	150
Copenhagen suburbs	175,000	350
N Zealand	150,000	1,500
E Zealand	75,000	800
W/S Zealand	200,000	6,000
Bornholm	15,000	550
Funen	150,000	3,500
S Jutland	250,000	9,000
W Jutland	150,000	7,000
E Jutland	300,000	6,000
N Jutland	200,000	8,000

Denmark

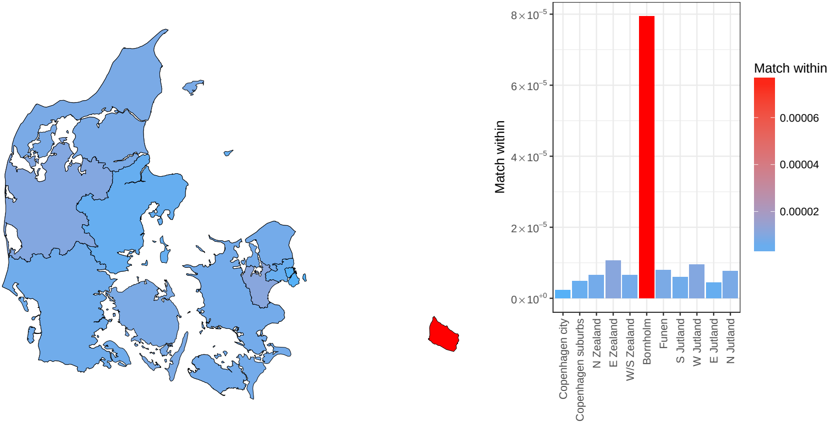
Pedigrees (males)



Denmark

Match within

$$\frac{1}{n_i(n_i-1)}\sum_h n_{ih}(n_{ih}-1)$$



θ (theta) Danish subdivisions

θ (theta) for 11 Danish NUTS-3 regions:

$$\theta = 1.2 \cdot 10^{-5}$$
$$\theta_{\text{weighted}} = 5.5 \cdot 10^{-6}$$

θ (theta) for 99 Danish local authorities/municipalities:

$$\theta = 4.1 \cdot 10^{-4}$$
$$\theta_{\text{weighted}} = 4.2 \cdot 10^{-5}$$

θ_{weighted}: means weighted by subpopulation sizes.

(Based on – possibly incomplete – pedigree information, no genetic information.)

Thank you for your attention

[ACJ+13] Mikkel Meyer Andersen, Amke Caliebe, Arne Jochens, Sascha Willuweit, and Michael Krawczak. Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. *Forensic Science International: Genetics*, 7:264–271, 2013.

[AEM13] Mikkel Meyer Andersen, Poul Svante Eriksen, and Niels Morling. The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology*, 329:39–51, 2013.

[AEM14] Mikkel Meyer Andersen, Poul Svante Eriksen, and Niels Morling. Cluster analysis of European Y-chromosomal STR haplotypes using the discrete Laplace method. *Forensic Science International: Genetics*, 11:182–194, 2014.

[AEMM15] Mikkel Meyer Andersen, Poul Svante Eriksen, Helle Smidt Mogensen, and Niels Morling. Identifying the most likely contributors to a Y-STR mixture using the discrete Laplace method. *Forensic Science International: Genetics*, 15:76–83, 2015.

[Bre10] Charles H. Brenner. Fundamental problem of forensic mathematics – The evidential value of a rare haplotype. *Forensic Science International: Genetics*, 4(5):281–291, 2010.

[Cer15] Giulia Cereda. Non parametric Bayesian approach to LR assessment in case of rare haplotype match. *arXiv:1506.08444*, (in preparation), 2015.

[Cer17] Giulia Cereda. Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach). *Scandinavian Journal of Statistics*, (to appear), 2017.