Forensic Statistics of Lineage DNA Markers

PhD defence

Feb 28, 2014

Mikkel Meyer Andersen

Department of Mathematical Sciences

AALBORG UNIVERSITY
DENMARK

Outline

1. Introduction (to forensic genetics)
2. Overview of PhD work
3. Details of parts of the PhD work (discrete Laplace method)

Introduction

Forensic genetics

- Aims: Identify people and investigate legal issues using genetic evidence
- Unbiased evidence evaluation (using statistics, not subjective assessments)
- Rule out suspects (like innocents on death row)

Trace found at crime scene

1. Trace of genetic evidence from the perpetrator found at crime scene
2. Suspect arrested
3. DNA profiles are compared

Evidential weight

- E: evidence (e.g. DNA profile from crime scene)
- Weight of the evidence (likelihood ratio):

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)},$$

- H_p (prosecutor's hypothesis) is 'the suspect is the donor of the genetic data' (often assumed equal to 1)
- H_d (defence attorney's hypothesis) is 'the suspect is unconnected to the crime'
- $P(E \mid H_d)$: Match probability $\approx$ match by chance $\approx$ 'How probable it is that some random man's DNA profile matches the DNA profile found at the crime scene?' (population frequency)

Evidential weight interpretation

- E: evidence (e.g. DNA profile from crime scene)
- H_p (prosecutor's hypothesis) is 'the suspect is the donor of the genetic data'
- H_d (defence attorney's hypothesis) is 'the suspect is unconnected to the crime'
- Ideal usage of LR:

$$\underbrace{\frac{P(H_p \mid E)}{P(H_d \mid E)}}_{\text{Posterior odds}} = \underbrace{\frac{P(E \mid H_p)}{P(E \mid H_d)}}_{\text{LR}} \times \underbrace{\frac{P(H_p)}{P(H_d)}}_{\text{Prior odds}},$$

- Toss a coin 10 times to obtain $E = \{4 \text{ heads}, 6 \text{ tails}\}$
- $H_1 : \theta = 0.5$ vs $H_2 : \theta = 0.9$ ($\theta = P(\text{heads})$)
- $P(\theta = 0.5 \mid E) / P(\theta = 0.9 \mid E)$?
- $P(E \mid \theta = 0.5) = 20.51\%$ and $P(E \mid \theta = 0.9) = 0.01\%$
- $LR = P(E \mid \theta = 0.5) / P(E \mid \theta = 0.9) = 1488$
- $P(H_1) / P(H_2)$ must be known to say anything about posterior odds

DNA

- Bases: A, T, C, G (A-T and C-G)
- 3.3 billion base pairs (3.3 billion = 3,300,000,000)
- 23 chromosome pairs
- In each pair: One chromosome inherited from mother and one from father

From www.wikimedia.org

DNA profiles
Based on short tandem repeats, STRs

- Method used today: short tandem repeat (STR)
- Locus (*loci* in plural): Location at a certain chromosome (e.g. D3S1358, DYS391)
- Allele: The number of times a *motif* (short sequence of 3-5 base pairs) repeats itself
- An example of an allele of 3:

$$\underbrace{AGAT}_{motif}\ AGAT\ AGAT = [AGAT]_3$$

- STR's can mutate during meiosis causing variation (e.g. $11 \rightarrow 10$)

DNA profiles
Based on short tandem repeats, STRs

- Traditional DNA profile: Based on autosomal (non-sex) chromosomes
- DNA profile consists of 10-20 loci
- Example of autosomal STR DNA profile (only three loci shown):

$$\texttt{D3S1358} = \{15, 18\},\ \texttt{D5S818} = \{12, 12\},\ \texttt{D7S820} = \{10, 11\}$$

- Other types (lineage markers): e.g. Y chromosomal
  - Y-STR haplotypes: DNA profiles from the Y chromosome using STR
  - Example of Y-STR DNA profile (only three loci shown):

$$\texttt{DYS391} = 10,\ \texttt{DYS437} = 15,\ \texttt{DYS635} = 22$$

DNA profiles: autosomal vs Y profiles

Why bother using anything else than traditional autosomal STR DNA profiles?

- Unbalanced mixture of female/male DNA (minor male component masked)
- Extract Y chromosomal DNA to obtain Y chromosomal DNA profile

DNA profiles: autosomal vs Y profiles

From www.wikimedia.org

Statistical properties (due to genetic inheritance)

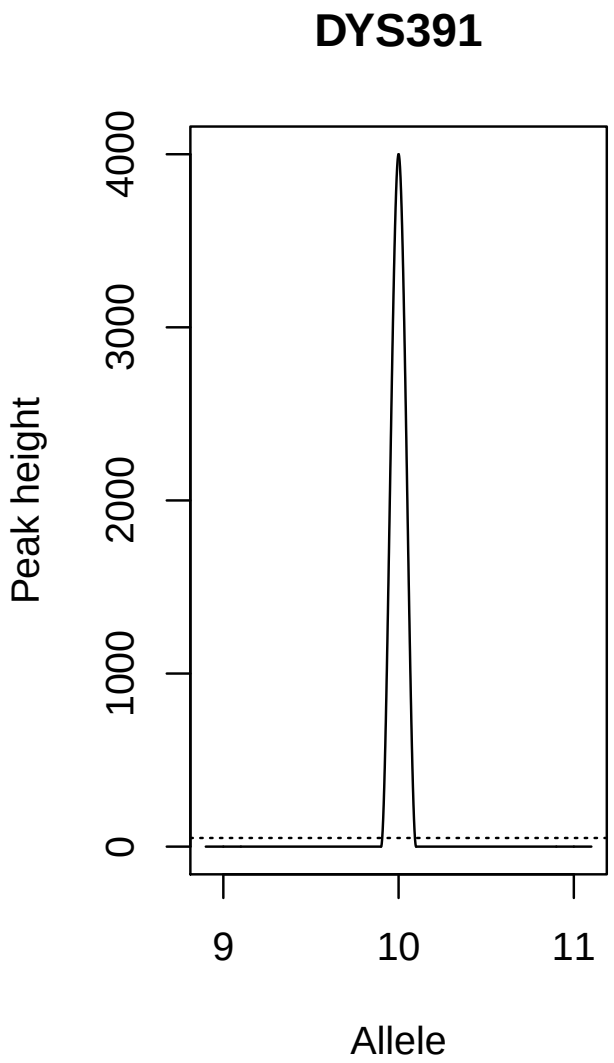
- Autosomal: 2 alleles per locus inherited independently between and within loci from each parent
  - Widely used and a lot of statistics for that area exist
  - Match probability of DNA profile: Product of the allele frequencies at each locus
- Y chromosomal: 1 allele per locus inherited as a whole from the father
  - Strong dependency between loci
  - Match probability of DNA profile: Very different than for autosomal DNA profiles (main focus of PhD thesis)

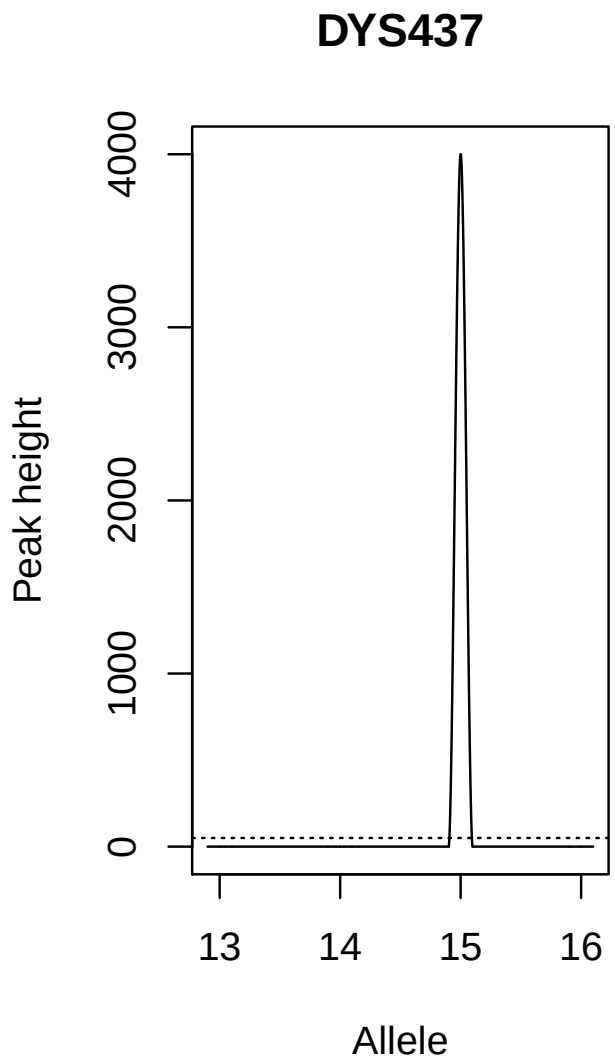
PhD work

PhD work

- Extraction of DNA profile from biological material
  - Paper I-III
- Haplotype distribution modelling
  - Paper IV, VI, VIII
- Utilities
  - Paper V, VII, IX

Example of Y-STR signal





Paper III: Modelling drop-out rates

Forensic Science International: Genetics

Alleles not showing up (no signal or signal indistinguishable from background noise)

- **Null alleles**: Alleles hidden due to molecular mechanism (e.g. mutation in primer region)
  - Unique *primer sequences* anchor the allele (here allele 13):

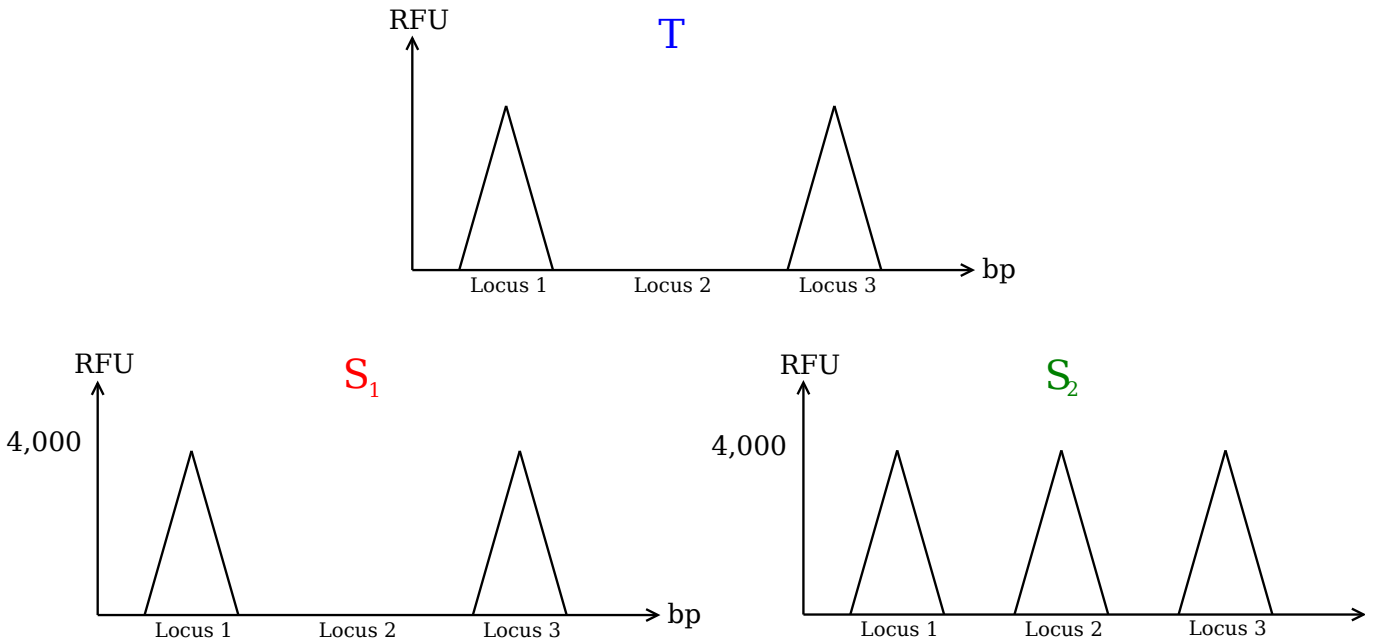
$$\underbrace{GCAGTCCAATCTGGGT}_{\text{Forward primer}}\underbrace{[AGAT]_{13}}_{\text{Allele}}\underbrace{CAAGCCTCTGTTGAT}_{\text{Reverse primer}}$$

  - Happens approx. 1:5,000 alleles (`http://www.yhrd.org`, release 39)
- **Drop-out**: Stochastic error (e.g. due to low amount of input DNA)
  - Simple logistic regression model: $P(\text{Drop-out})$ modelled by (mainly) signal strength
  - Peak height model: $\log x_j \sim N_{\log t}\left(\theta_j + \log S, \sigma^2\right)$
    - Truncation ($N_{\log t}$, $t = 50$ RFU) and interlocus balances (θ_j)
  - $P(\text{Drop-out} \mid S \approx 4{,}000 \text{ RFU}) \approx 1{:}100{,}000$
    - 20 times less likely than null allele
  - $P(\text{Drop-out} \mid S \approx 75 \text{ RFU}) \approx 1{:}5$
    - 1,000 times more likely than null allele

Paper III: Modelling drop-out rates

Forensic Science International: Genetics

- P(Null alleles) $=$ 1:5,000 (independent of signal strength)
- P(Drop-out | Signal strength $\approx$ 4,000 RFU) $\approx$ 1:100,000 (20 times less likely than null allele)
- P(Drop-out | Signal strength $\approx$ 75 RFU) $\approx$ 1:5 (1,000 times more likely than null allele)

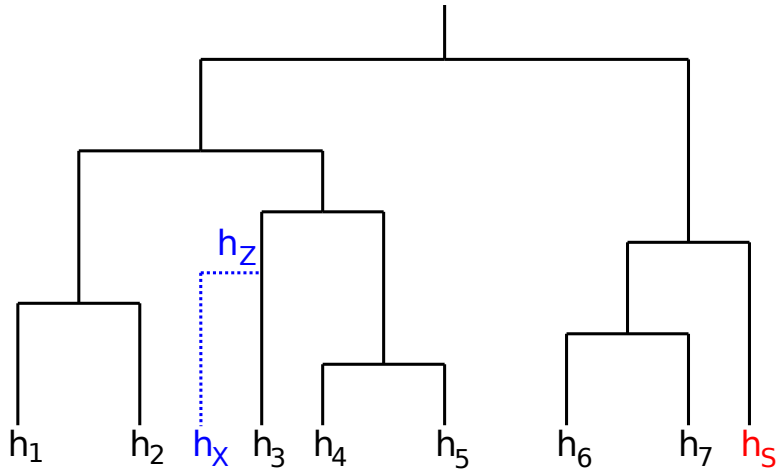


Match probability

- Match probability $\approx$ DNA profile frequency
- Count method (works for any trait, e.g. blood type)
  - n: Database (DB) size
  - n_x: Number of times x is observed in the database
  - $P(X = x) = n_x/n$
- Problem: Singletons (haplotypes only observed once) are common (a lot of rare variants)
  - $\sum_{x \in \text{DB}} n_x/n = 1$, hence $P(X = x) = 0$ for $x \notin$ DB
- Many suggestions (not probability distributions)

Paper IV: Coalescent method

Forensic Science International: Genetics



H: Database. S: Suspect. X: Trace donor. h_i: Haplotype of i.

- X: Unknown trace donor (random lineage in each tree)
- Z: Most recent common ancestor of X and closest from database
- $P(h_X = h_S \mid H, h_S, h_Z(i), t(i))$: Probability that h_Z mutates into h_S when passed down from Z to X

Paper IV: Coalescent method

Forensic Science International: Genetics

Results:

- Theoretically interesting approach
- Current method/software too slow
- Focuses on one haplotype (distribution only given implicitly)

Paper VI & VIII: Discrete Laplace method

VI: Journal of Theoretical Biology; VIII: Submitted to FSI: Genetics

Model the (multivariate) probability distribution of Y-STR haplotypes

Discrete Laplace distribution

Discrete Laplace distributed $X \sim DL(p, \mu)$:

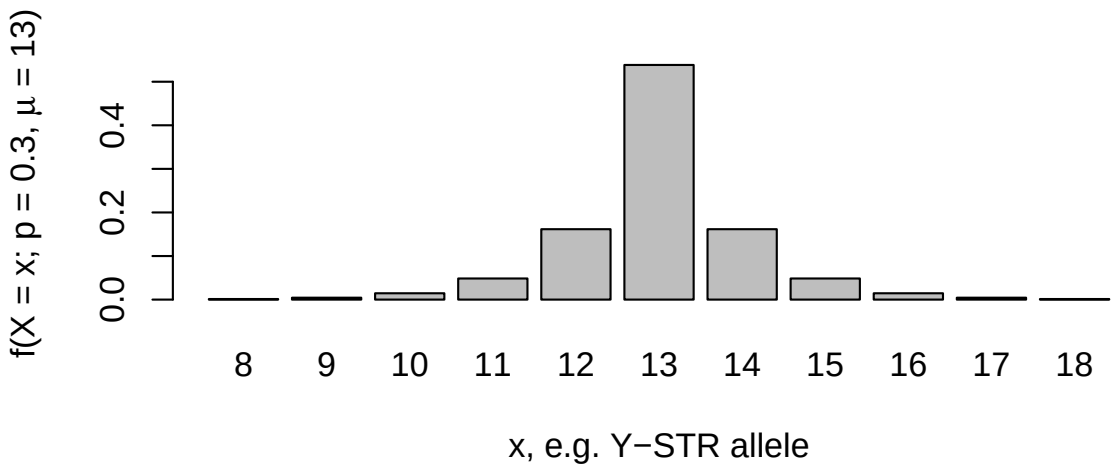
- Dispersion parameter $0 < p < 1$ and
- Location parameter $\mu \in \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$

Probability mass function:

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

Perfectly homogeneous population with 1-locus haplotypes:

$$P(X = x) = f(X = x; p, \mu)$$



Discrete Laplace exponential family

- Exponential family for known location parameter ($\theta = \log p$ and $d = x - \mu$):

$$f(d;\theta) = \exp\left(\theta|d| - A(\theta)\right) \quad \text{with } A(\theta) = \log\left(\frac{1+e^{\theta}}{1-e^{\theta}}\right).$$

- `R` family object for generalized linear model implemented in `R` library `disclap` (also {`d`, `p`, `r`}`disclap`)
- `glm(d ~ 1, dat, family = DiscreteLaplace())`

Statistical model for Y-STR haplotypes

Perfectly homogeneous population with r-locus haplotypes:

$$P(X = (x_1, x_2, \ldots, x_r)) = \prod_{k=1}^{r} f(x_k - \mu_k; p_k)$$

- $\mu = (\mu_1, \mu_2, \ldots, \mu_r)$: central haplotype
- $p = (p_1, p_2, \ldots, p_r)$: discrete Laplace parameters (one for each locus)
- Mutations happen independently across loci (relative to μ)

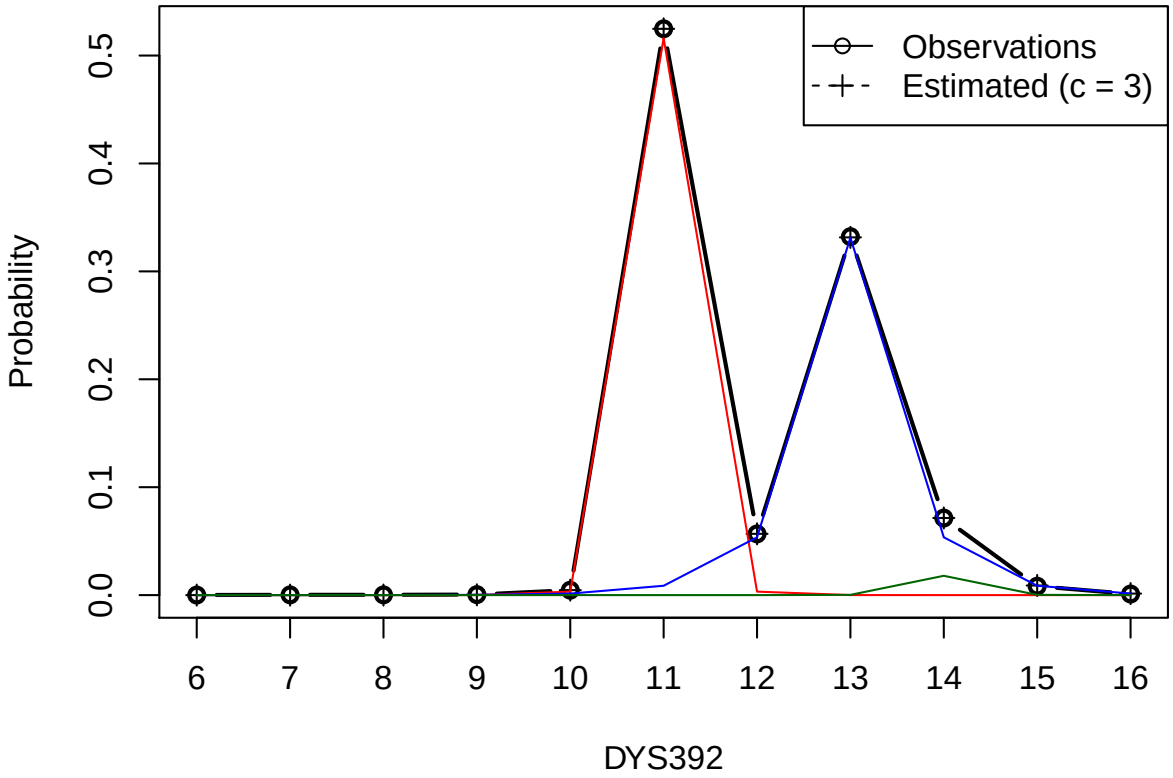
Statistical model for Y-STR haplotypes

Non-homogeneous population with c subpopulations and r-locus haplotypes:

$$P(X = (x_1, x_2, \ldots, x_r)) = \sum_{j=1}^{c} \tau_j \prod_{k=1}^{r} f(x_k - \mu_{jk}; p_{jk})$$

- τ_j: a priori probability for originating from the j'th subpopulation ($\sum_{j=1}^{c} \tau_j = 1$)
- $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jr})$: central haplotype for j'th subpopulation
- $p_j = (p_{j1}, p_{j2}, \ldots, p_{jr})$: parameters for all loci at j'th subpopulation
- Parameter estimation from observations using `R` library `disclapmix`
- Software tutorial on using the discrete Laplace method software (paper VII)

Data and fit



- 3 clusters:

$\hat{\mu}_j$	11	13	14
$\hat{\tau}_j$	52%	46%	2%

- Observed vs expected:

Allele	11	12	13	14	15
Observed	0.5248	0.0567	0.3322	0.0714	0.0083
Expected	0.5248	0.0567	0.3315	0.0715	0.0089

Estimate match probability

Paper VI: Journal of Theoretical Biology

- Estimate haplotype frequency and compare to true value
- Simulate population (e.g. 20 mio. individuals)
- Draw random database of individuals (e.g. 1,000)
- Paper V: Efficient simulation of populations (simulate haplotypes, not individuals)
- Result: smaller prediction error than existing estimators

Cluster analysis of European data

Paper VIII: Submitted to FSI: Genetics

- European 7-loci Y-STR database from 2004 consisting of 12,727 individuals in 91 European sample locations
- First analysed in 'Signature of recent historical events in the European Y-chromosomal STR haplotype distribution' by Roewer *et al.* in 2005
- Our study
  - Fit a discrete Laplace model
  - Parameters (genetic information) versus known sample locations
  - Discrete Laplace model does not know about sample locations, it infers 'genetic' subpopulations (or clusters)

Cluster analysis of European data

- Sample locations: $s = 1, 2, \ldots, S$ $(S = 91)$
- Subpopulations: $j = 1, 2, \ldots, c$ $(c = 40)$
- w_{sj}: Fraction of individuals from location s originating from subpopulation j
- $w_{s+} = \sum_{j=1}^{c} w_{sj} = 1$

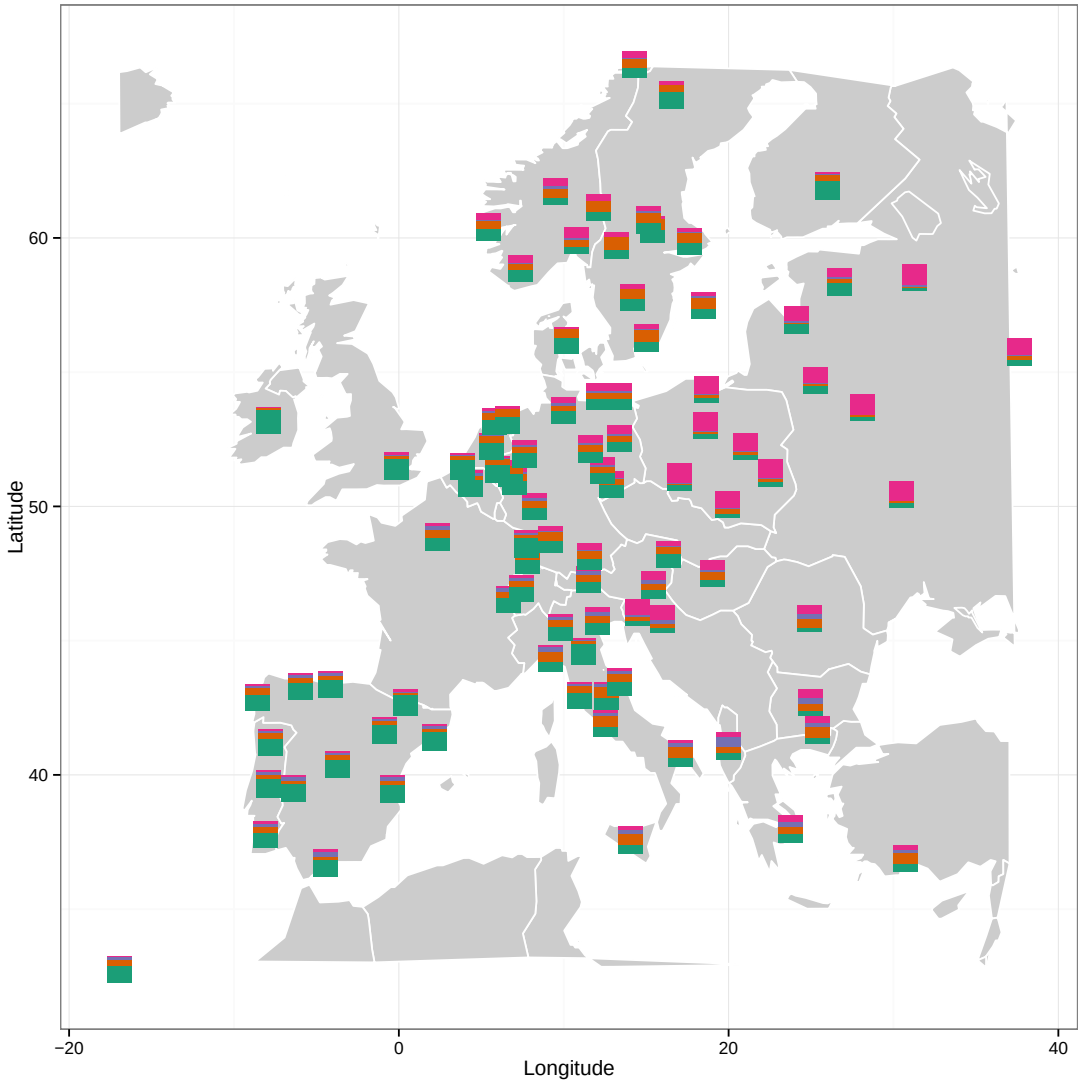
w_{sj} values for selected subpopulations and regions:

s	Location	$j = 1$	$j = 4$	$j = 14$	$j = 17$	$j = 27$	$j = 40$
1	Croatia	0.13		0.19			
2	Denmark		0.13			0.17	
3	Finland				0.39		
4	Northern Poland			0.09			0.14

Empty cell means 0.0.

Cluster analysis of European data

Collapsed w_{sj} values for 4 mega clusters:

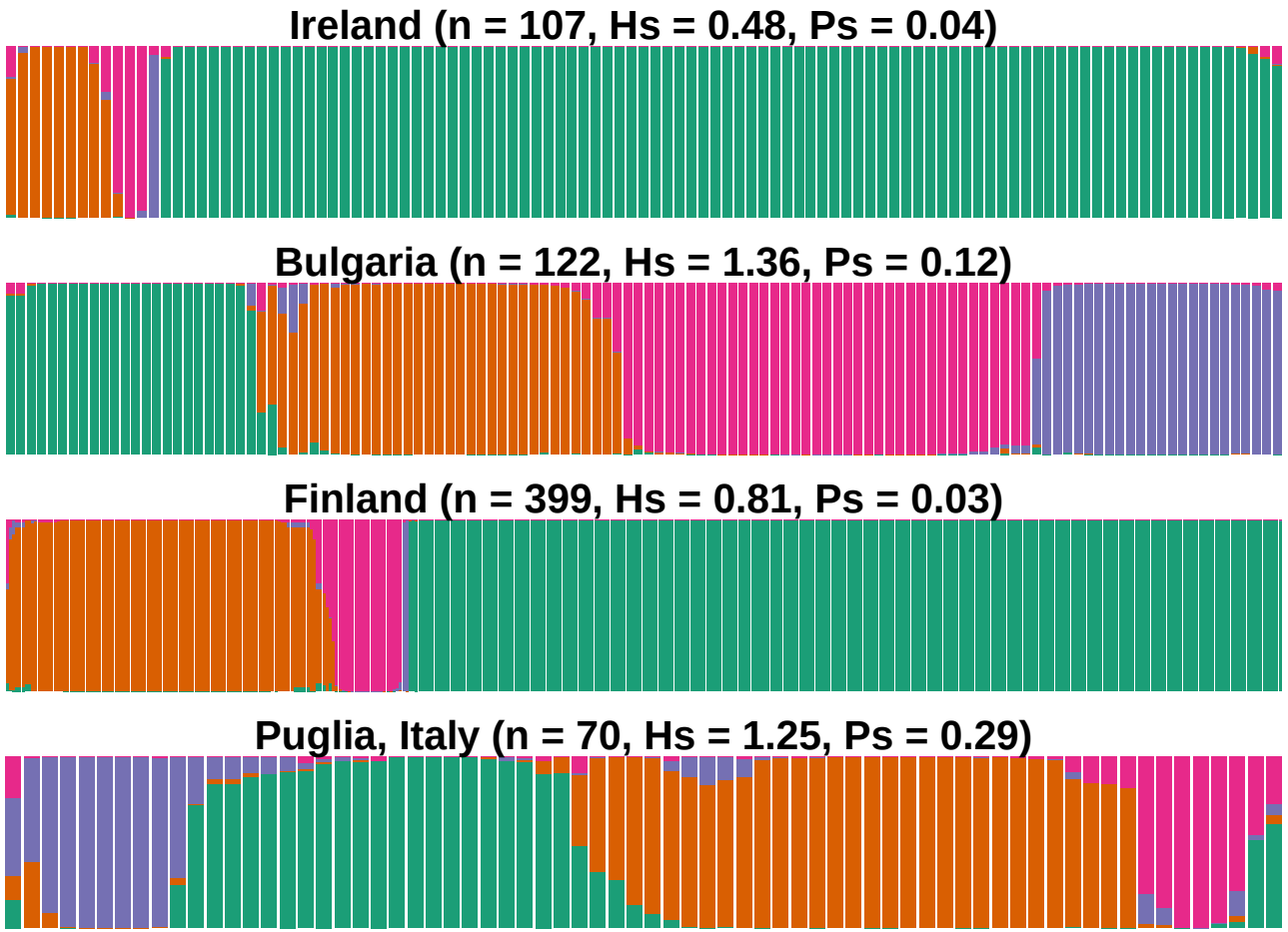


14,13,16,25,11,13,13 (R1b1b2a2g)
14,13,16,25,10,13,13 (R1b1b2a1)
14,13,16,24,10,13,13 (R1b1b2a2c)
14,13,17,24,10,13,13 (R1b1b2a2g)
14,13,16,23,10,13,13 (R1b1b2a2c)
14,13,17,23,11,13,12 (R1b1b)
14,13,16,23,11,13,13 (R1b1b2a1)
15,13,16,24,11,13,13 (R1b1b2a2g)
14,13,17,24,11,13,13 (R1b1b2a2c)
14,13,16,24,11,13,13 (J1a)
14,14,16,24,11,13,13 (R1b1b2a2c)
14,14,16,24,11,14,14 (N1c)
14,14,16,23,11,14,14 (N1c1)
15,13,16,23,10,14,14 (N1c)
15,14,17,23,10,12,14 (I2b)
15,13,17,23,10,12,14 (I2b)
15,12,17,22,10,11,14 (G2a3)
15,12,17,22,10,11,13 (G2a3b)
15,12,16,22,10,11,13 (G2a3)
14,12,17,22,10,11,13 (I1)
14,12,16,22,10,11,13 (I1)
14,12,16,23,10,11,13 (I1)
15,12,16,24,10,11,12 (J2b2)
15,13,16,23,10,11,12 (J1)
14,13,17,23,10,11,12 (J1e)
14,13,16,23,10,11,12 (J2a8)
13,14,16,24,9,11,13 (E1b1b1b)
13,13,17,24,10,11,13 (E1b1b1a2)
13,13,18,24,10,11,13 (E1b1b1a)
14,13,16,24,11,11,13 (R1b1b2a2g)
16,13,18,24,11,11,13 (I2a)
16,13,18,24,10,11,13 (I2a)
16,13,16,24,10,11,13 (R1a1a)
16,13,16,25,10,11,13 (R1a1a7)
16,13,17,25,10,11,13 (R1a1a)
15,13,17,25,10,11,13 (D2)
15,13,17,25,11,11,13 (R1a)
16,13,17,25,11,11,13 (R1a)
17,13,17,25,11,11,13 (R1a)
17,13,17,25,10,11,13 (R1a1a7)

Dept. of Mathematical Sciences
Aalborg University
Denmark

Cluster analysis of European data

- H_s: Uncertainty about subpopulation origin at sample location level (entropy of means)
- P_s: Uncertainty about subpopulation origin at individual level (mean of entropies)



Columns: Individuals. Rows: Mega clusters. Bar at column *i*, row *m*: *P*(Indiv. *i* orig. *m*)

Dept. of Mathematical Sciences
Aalborg University
Denmark

Conclusion
Capabilities of the discrete Laplace method

- Estimation of Y-STR haplotype population frequencies
  - Sound statistical properties
  - Simulation study showed smaller prediction error than existing estimators
- Cluster analysis
  - Many analyses possible
  - Gives results similar to previous studies
- Computationally feasible
- Open source software: `R` library `disclap` and `disclapmix`

Forensic Statistics of Lineage DNA Markers

Mikkel Meyer Andersen

Thank you for your attention

AALBORG UNIVERSITY
DENMARK