Statistical methods in forensic genetics: Haplotype evidence

Train the trainers – Workshop Part II
Copenhagen, May 21, 2014

Mikkel Meyer Andersen

Department of Mathematical Sciences
Aalborg University
Denmark

AALBORG UNIVERSITY
DENMARK

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

(We assume that no errors, e.g. typing errors, have happened.)

- Non-match: $P(E \mid H_p) = 0$, suspect is probably not the perpetrator (provided that all DNA material is found at the crime scene etc.)
- Match: $P(E \mid H_p) = 1$, the suspect *might* be the perpetrator:

$$LR = \frac{1}{P(E \mid H_d)},$$

Evidential weight

ISFG recommendations of Y-STR usage from 2006 (`http://www.isfg.org/Publication;Gusmao2006`):

> *Recommendations on the estimation of Y-STR haplotype frequencies and estimation of the weight of the evidence of Y-STR typing will be presented separately as guidelines for the interpretation of forensic genetic evidence.*

- Highly wanted guidelines
- Problem: Singletons (haplotypes only observed once) are common (a lot of rare variants)

Evidential weight

Calculate match probability:

1. Use random population sample (database)
2. Calculate match probability (in practise, estimate population frequency)

Estimators

- Lineage markers, e.g. Y-STR and mtDNA
- Match probability $\approx$ DNA profile frequency
- Count method (works for any trait, e.g. blood type)
  - n: Database size
  - n_x: Number of times x is observed in the database
  - $P(X = x) = \frac{n_x}{n}$
- Examples
  - $n_x = 0$: $P(X = x) = \frac{0}{n} = 0$
  - $n_x \gg 0$ (e.g. $n_x = 9$, $n = 99$): $P(X = x) = \frac{9}{99} = \frac{1}{11}$

Estimators

- Include in database (new observation, conservative)
  - Additional information: Under H_d, suspect considered as a random (wrongly accused) individual from the population; haplotype just yet another random sample
  - In favor of the suspect: Increases the match probability, meaning that it decreases the *LR*
- Old database: D^- of size n
- New database: D of size $n+1$
- Corrected count method:
  - $n_x = 0$: $P(X = x) = \frac{1}{n+1}$
  - $n_x \gg 0$ (e.g. $n_x = 9$, $n = 99$): $P(X = x) = \frac{9+1}{99+1} = \frac{1}{10}$
- $\sum_{x \in D} \frac{n_x}{n+1} = \frac{1}{n+1} \sum_{x \in D} n_x = \frac{n+1}{n+1} = 1$, hence $P(X = x) = 0$ for $x \notin D$

Estimators

Brenner (2014): *'[...] structural neighborhood of a haplotype offers little information about the probability to match the haplotype. That isn't to say the structure contains no information at all.'*

- Count method
- Confidence interval upper bound (Clopper and Pearson (1934))
- Frequency surveying (Roewer, Krawczak, *et al.* in 2000, 2001, 2010): Deprecated in new `www.YHRD.org`
- Coverage, saturation and PCA (Egeland and Salas, 2008)
- Brenner's kappa (2010), based on work by Robbins (1968)
- Coalescent method (Andersen *et al.*, 2013)
- Discrete Laplace method (Andersen *et al.*, 2013, 2014)

Brenner's kappa

- h is the haplotype of the suspect and also found at crime scene
- h is previously unobserved (never seen before)
- Database D with n observations (h not present)
- α = #singletons in D
- $\kappa = \frac{\alpha+1}{n+1}$ (add h to D)
- Match probability = $\frac{1-\kappa}{n+1}$
- $LR = \frac{1}{\text{Match probability}} = \frac{n+1}{1-\kappa} > n+1$ obtained from count estimator
- LR inflated by $\frac{1}{1-\kappa}$

DEMO

DEMO

Parametric model

- Sample from population of peoples' height
- Inference questions:
  - Percentage of population being heigher than x cm.?
  - Percentage of population being between y cm. and $y + \delta$ cm.?
- Non-parametric vs. parametric model

Motivation

- Haplotype probability distribution (statistical model)
- Enables a wide range of inferences using one model:
  - Haplotype frequency estimation (observed and unobserved)
  - Cluster analysis
  - Mixtures (e.g. separation and *LR*)
  - ...
- Not a new ad-hoc tool for each task
- Statistical model gives desirable properties:
  - $P(x)$: probability mass function
  - Consistent:

$$\sum_{x \in \mathcal{H}} P(x) = 1$$

- ..but also challanges in being conservative:
  - $\sum_{x \in \text{DB}} P(x) < 1$
  - $P(x) > 0$ for all $x \in \mathcal{H}$

Model

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

Discrete Laplace distribution

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

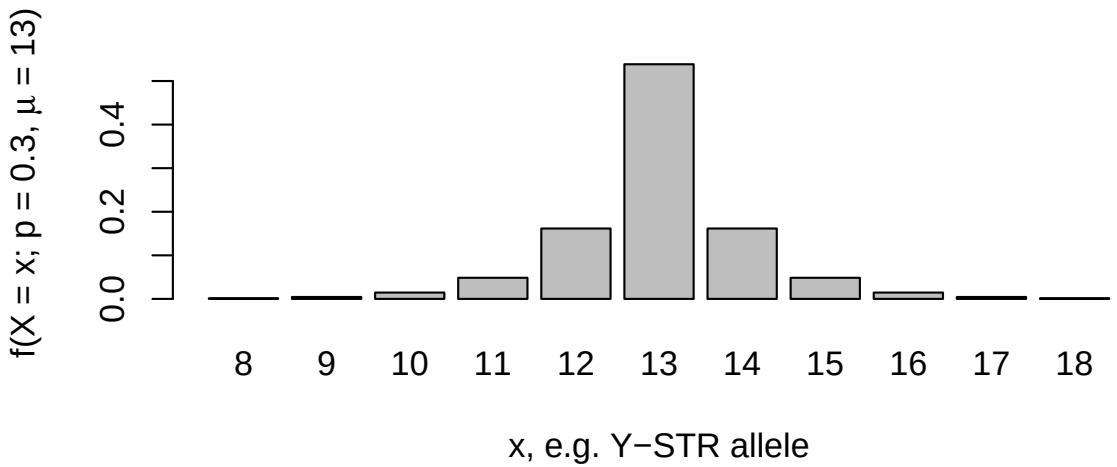
- Dispersion parameter $0 < p < 1$ and
- Location parameter $\mu \in \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$

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:

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



Statistical model for Y-STR haplotypes

Perfectly homogeneous population with r-locus haplotypes:

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

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

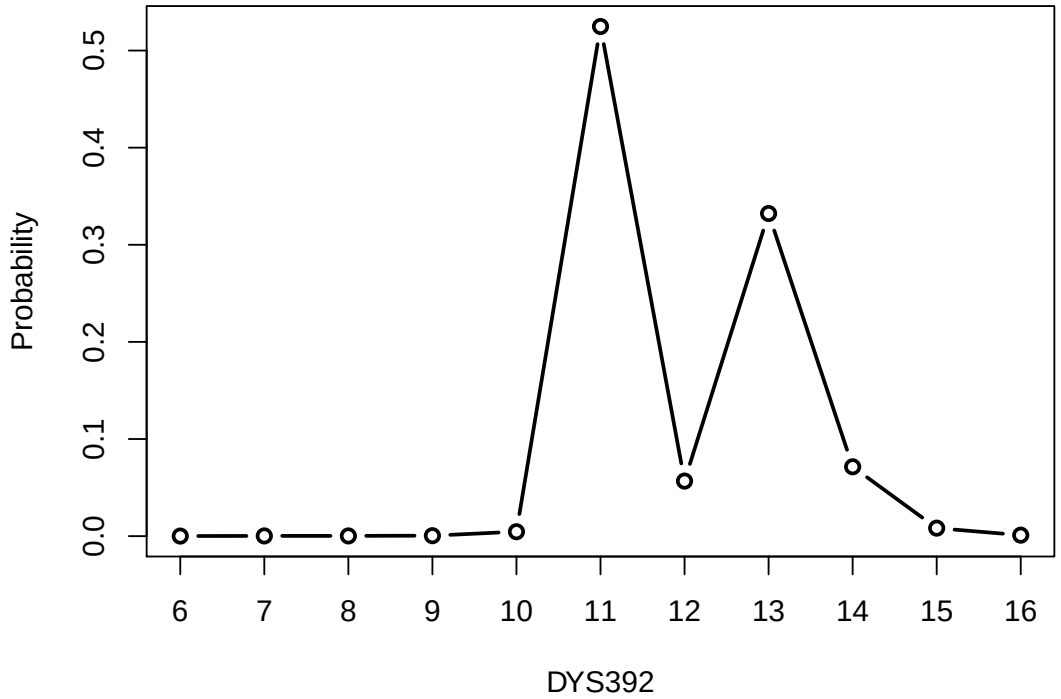
Statistical model for Y-STR haplotypes

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

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

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

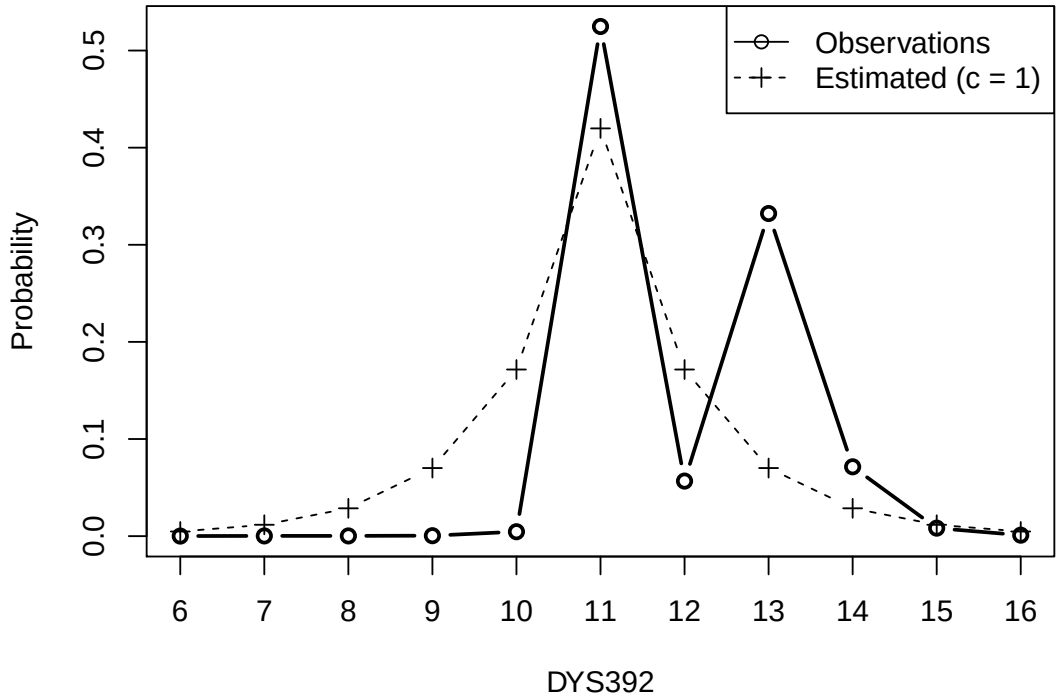
Data and fit



c: Number of subpopulations

$P(X = x) = \sum_{j=1}^{c} \tau_j f(x; p_j, \mu_j)$

Data and fit

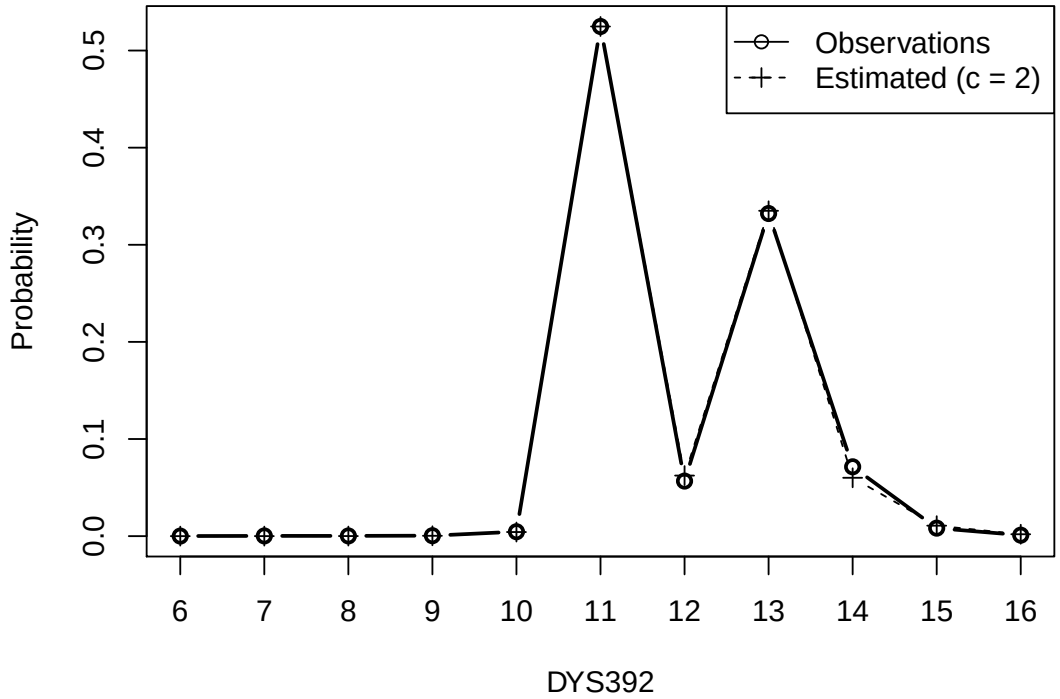


c: Number of subpopulations

$P(X = x) = \sum_{j=1}^{c} \tau_j f(x; p_j, \mu_j)$

$P(\texttt{DYS392} = x) = 1 \cdot f(x; p = 0.41, \mu = 11)$

Data and fit



c: Number of subpopulations

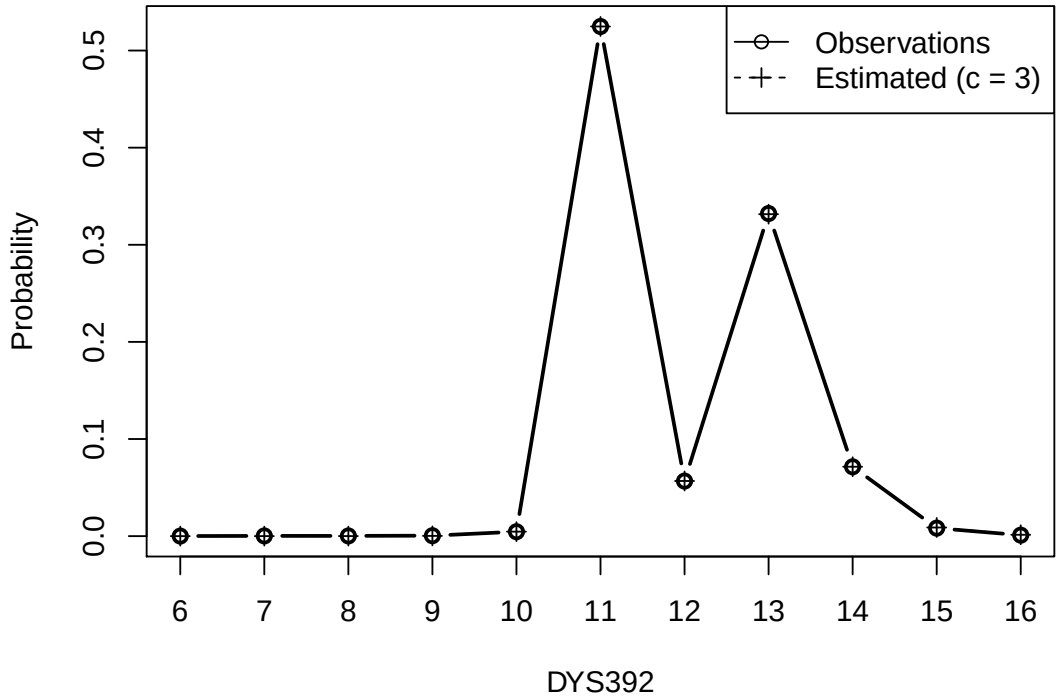
$P(X = x) = \sum_{j=1}^{c} \tau_j f(x; p_j, \mu_j)$

$P(\texttt{DYS392} = x) =$

$0.519 \cdot f(x; p = 0.004, \mu = 11) + 0.481 \cdot f(x; p = 0.179, \mu = 13)$

Dept. of Mathematical Sciences
Aalborg University
Denmark

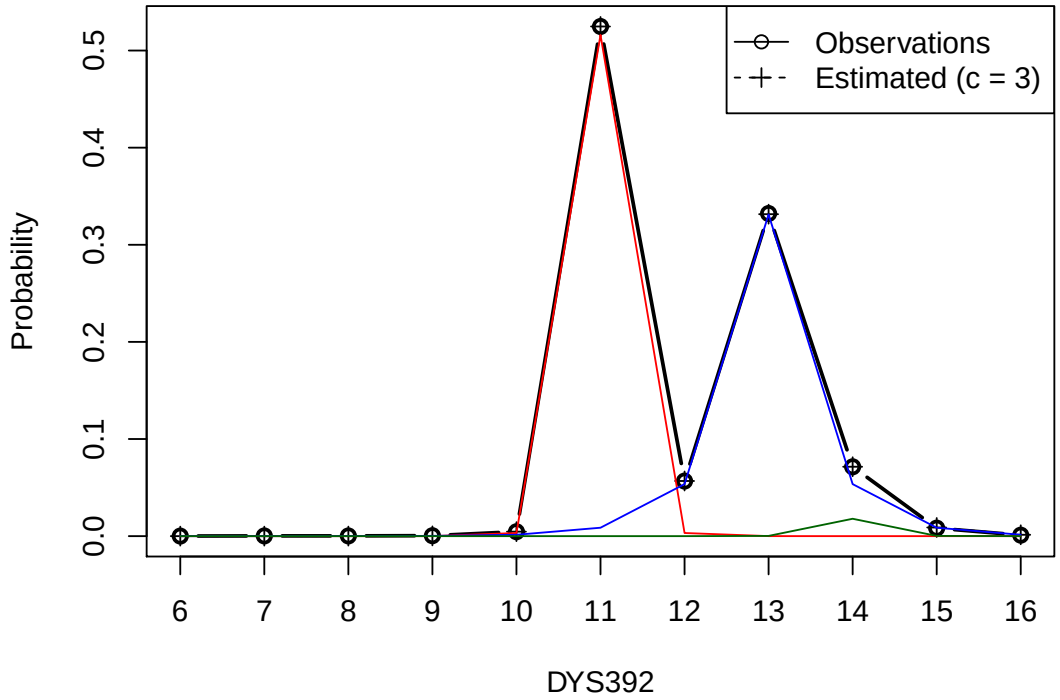
Data and fit



c: Number of subpopulations

$P(X = x) = \sum_{j=1}^{c} \tau_j f(x; p_j, \mu_j)$

Data and fit



c: Number of subpopulations

$P(X = x) = \sum_{j=1}^{c} \tau_j f(x; p_j, \mu_j)$

- 3 subpopulations:

$\hat{\mu}_j$	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

Dept. of Mathematical Sciences
Aalborg University
Denmark

Estimate match probability

1. Simulate population (e.g. 20 mio. individuals)
2. Draw random database of individuals (e.g. 1,000)
3. Estimate haplotype frequency and compare to true value

Result: smaller prediction error than those with existing estimators

DEMO

DEMO

Note: Haplotype of interest must be added to database on which model is fitted

Mixture separation

Yfiler trace (DYS385a/b removed), 15 loci left:

Locus	Alleles
DYS19	14, 15
DYS389I	13, 14
DYS389II'	16, 17
DYS390	24, 26
DYS391	10, 11
DYS392	11, 13
DYS393	13
DYS438	11, 12
DYS439	10, 11
DYS437	14, 15
DYS448	19, 20
DYS456	15, 16
DYS458	14, 18
DYS635	23
Y GATA H4	12, 13

$2^{13-1} = 4{,}096$ possible contributor pairs.

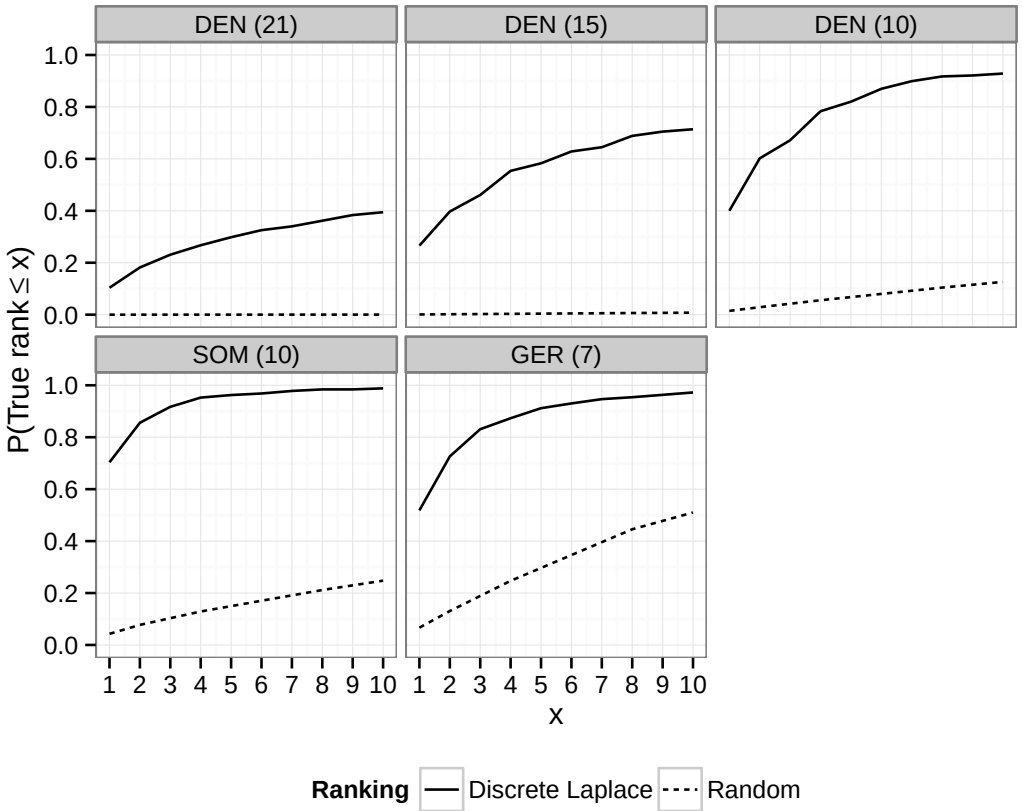
Mixture separation

	DEN (21)	DEN (15)	DEN (10)	SOM (10)	GER (7)
Dataset	All three the same Danish dataset			Somali	Germany
Loci (w/o DYS385a/b)	21	15	10	10	7
Observations	181	181	181	201	3,443
Singletons	181	164	112	56	662
Singleton proportion	1	0.906	0.619	0.279	0.192
Median loci w/ 2 alleles	14	10	6	3	4
Median #pairs	8,192	512	32	4	8

- For each dataset, 550 mixtures were simulated.
- i'th contributor pair $c_i = \{h_{i,1}, h_{i,2}\}$, find $\hat{p}_i = \hat{P}(h_{i,1})\hat{P}(h_{i,2})$
- Order all pairs according to the $\hat{p}_i$ values (highest to lowest)

Mixture separation

	DEN (21)	DEN (15)	DEN (10)	SOM (10)	GER (7)
P(Rank $\leq$ 1)	10%	27%	40%	70%	52%
P(Rank $\leq$ 5)	30%	58%	82%	96%	91%
P(Rank $\leq$ 10)	39%	71%	93%	99%	97%
P(RandomRank $\leq$ 10)	0.03%	0.78%	12.62%	24.76%	51.01%



Dept. of Mathematical Sciences
Aalborg University
Denmark

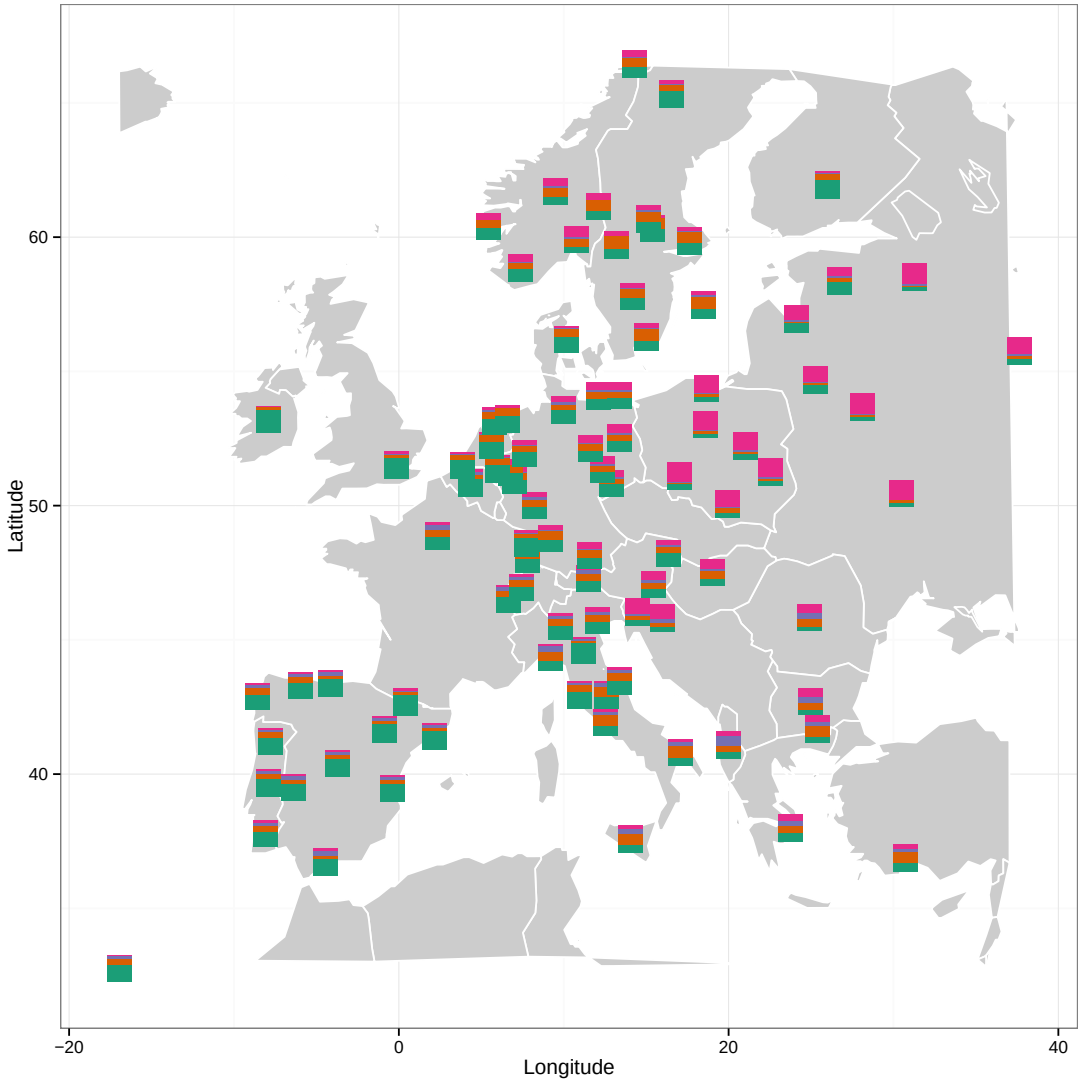
DEMO

DEMO

Cluster analysis of European data

- 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



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

Pairwise population distances:

- 7-locus, 12,727 European males (91 locations): Correlation(AMOVA, discrete Laplace) = 0.90
- 10-locus, 2,736 African males (26 locations): Correlation(AMOVA, discrete Laplace) = 0.82

Conclusion

- Sound statistical properties
- Applications
  - Estimation of Y-STR haplotype population frequencies
  - Mixture separation (new) – even for many loci
  - Cluster analysis
    - Many analyses possible (also those of e.g. AMOVA)
    - Gives results similar to those of previous studies
- Computationally feasible
- Open source software: `R` libraries `disclap` and `disclapmix` (and `fwsim` for simulating populations)

Conclusion

- Match probability is of great interest and is difficult
- Most promising methods for match probability:
  - Count method (but only if $n_h \gg 0$)
  - Brenner's method (but only if $n_h = 0$ and $\alpha < n$, maybe even $\alpha \ll n$)
  - Discrete Laplace method (no restrictions on n_h nor α; assigns probability mass to all possible haplotypes, both good and bad)
- Surveying (previously on `www.YHRD.org`) deprecated
- Other applications
  - Mixture analysis (separation, *LR*): only discrete Laplace at the moment
- Difficult to be both conservative and useful for unobserved haplotypes