Analysis of Y-Chromosomal STR Population Data Using the Discrete Laplace Model

9th ICFIS
Leiden University, August 2014

Mikkel Meyer Andersen, Poul Svante Eriksen and Niels Morling

AALBORG UNIVERSITY
DENMARK

Themes

- The discrete Laplace method and its applications
- Comparing methods for calculating *LR* for Y-STR data

MM Andersen
mikl@math.aau.dk

Introduction

Evidential weight

H_p (prosecutor's hypothesis): 'The suspect left the Y-chromosome DNA in the crime stain.'

H_d (defence attorney's hypothesis): 'A random man left the Y-chromosome DNA in the crime stain.'

E: Evidence (e.g. DNA profile from crime scene)

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

Non-match:

$$LR = \frac{0}{P(E \mid H_d)}$$

Match:

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

(Ideal situation, no errors, etc.)

Y-STR: Loci are <u>not</u> independent $\Rightarrow$ <u>No</u> product rule

MM Andersen
mikl@math.aau.dk

Sparsity of Y-STRs

19,630 samples	Forensic marker set				
	MHT 9 loci	SWGDAM 11 loci	PPY12 12 loci	Yfiler 17 loci	PPY23 23 loci
$n = 1$ (singletons)	6,083 (31.0%)	8,495 (43.3%)	9,092 (46.3%)	15,263 (77.8%)	18,237 (92.9%)
$n = 2$ (doubletons)	1,131	1,227	1,260	1,064	531
$n = 3$	435	436	416	256	64
$n = 4$	226	199	196	94	16
$n = 5$	114	101	106	63	6
$n = 6$	86	85	85	21	2
$n = 7$	63	51	50	12	2
$n = 8$	43	50	41	12	1
$n = 9$	29	29	34	9	
$n = 10$	31	21	24	4	
$n = 11$	22	24	28	5	1
...					
$n \in (30, 40]$	13	11	7	1	
...					
$n \in (100, 515]$	8	4	4		

Purps J, Siegert S, et al. (2014). A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. Forensic Science International: Genetics, Volume 12, 2014, p. 12-23.

MM Andersen
mikl@math.aau.dk

Estimators

Estimators

- Forensic 'conservatism' (innocent suspect): For whom – what about paternity, immigration, etc.?
- Precise (low prediction error) – how do we measure this (more later)?
- Does it work for all datasets, also for those only consisting of singletons?
- Statistical model: Guaranteed behaviour (e.g. probabilities sum to 1)
  - Assign probability to all possible haplotypes (e.g. for mixture *LR*)
  - Probability mass 1 to be distributed among possible haplotypes

Estimators

- Match probability $\approx$ DNA profile population frequency
- Count method (works for any trait, e.g. blood type)
  - n: Dataset size
  - n_x: Number of times x is observed in the dataset
  - $P(X = x) = n_x/n$

Estimators

- Include in dataset (new observation)
  - Additional information: Under H_d, suspect considered as a random (wrongly accused) individual from the population; the haplotype is just another random sample
- Old dataset: D^- of size n
- New dataset: D of size $n+1$
- $P(X = x) = (n_x + 1)/(n + 1)$
  - $n_x = 0$: $P(X = x) = \frac{1}{n+1}$
- $\sum_{x \in D} \frac{n_x}{n+1} = \frac{1}{n+1} \sum_{x \in D} n_x = \frac{n+1}{n+1} = 1$, hence $P(X = x) = 0$ for $x \notin D$
- Corrected count estimators:
  - Brenner's κ (CH Brenner (2010) / HE Robbins (1968))
  - Generalised Good (IJ Good (1953), G Cereda/R Gill)

The Discrete Laplace method

Motivation

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

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

  - $P(x) > 0$ 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 $X \sim DL(p, \mu)$:

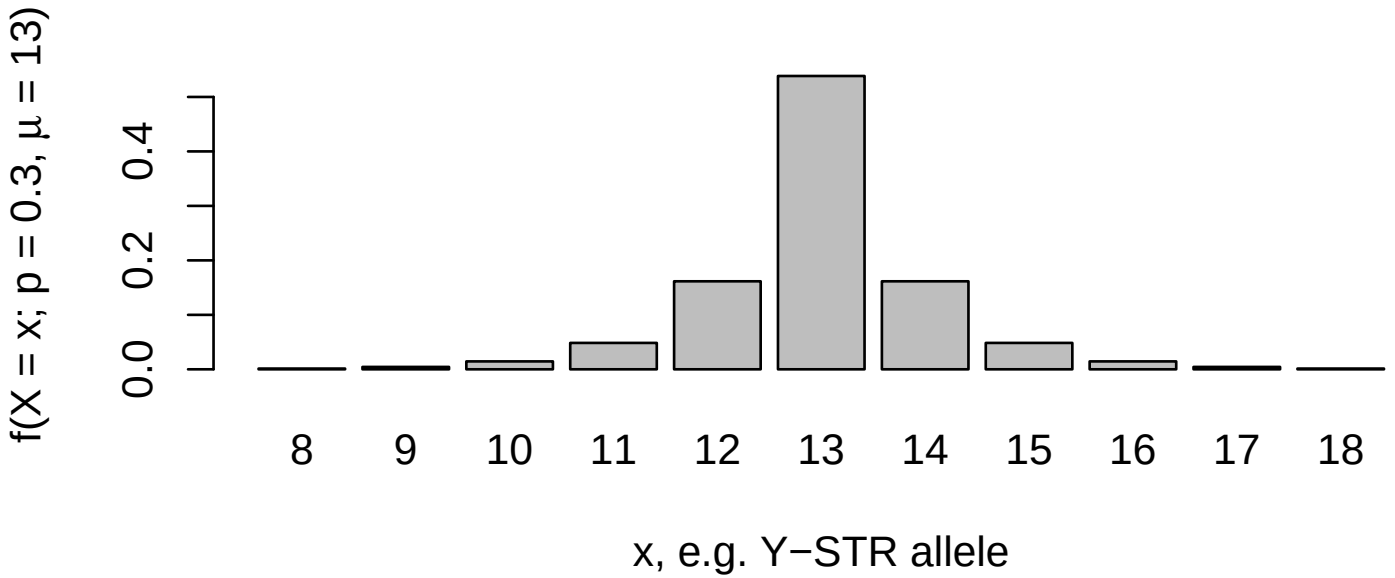
- Dispersion parameter $0 < p < 1$ and
- Location parameter $\mu \in \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$

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 $\vec{\mu}$)

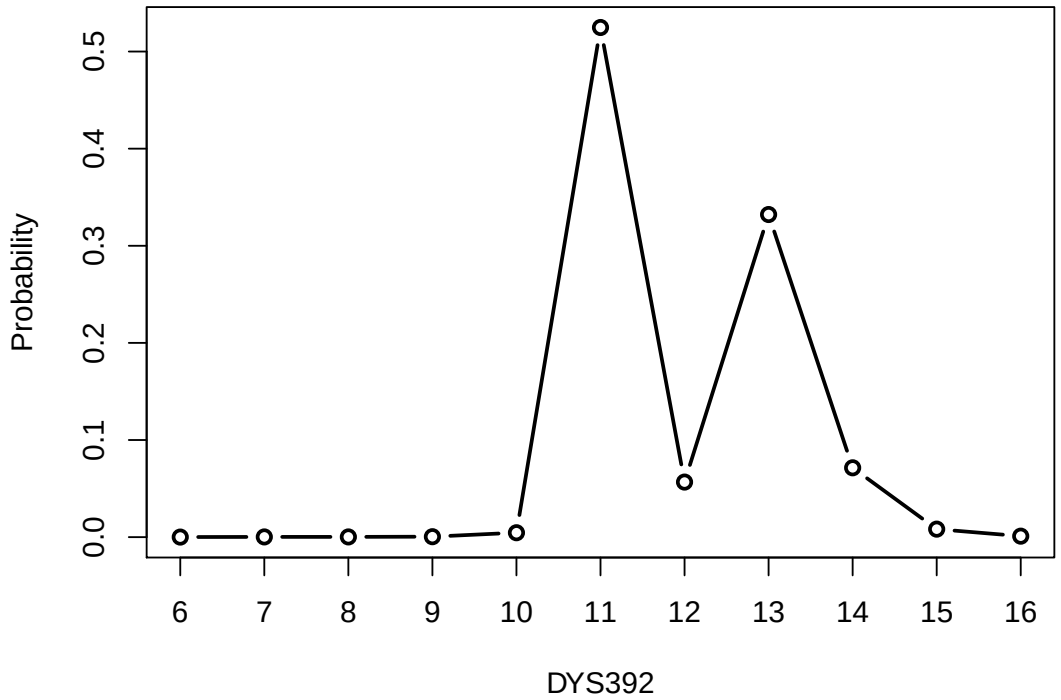
Statistical model for Y-STR haplotypes

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`

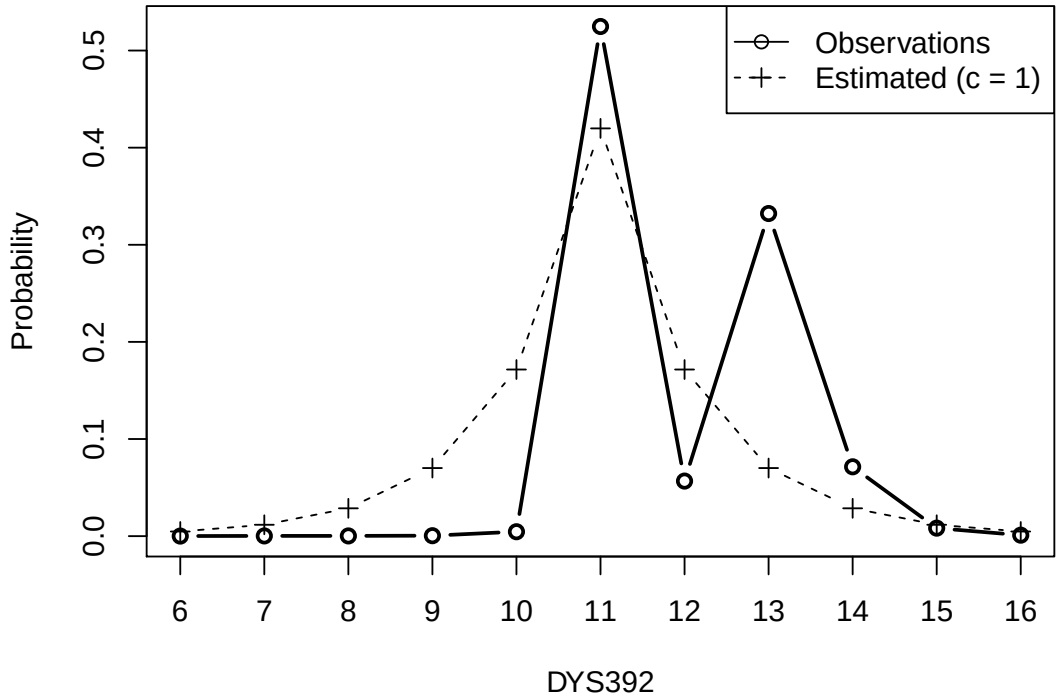
Data and fit



c: Number of subpopulations

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

Data and fit

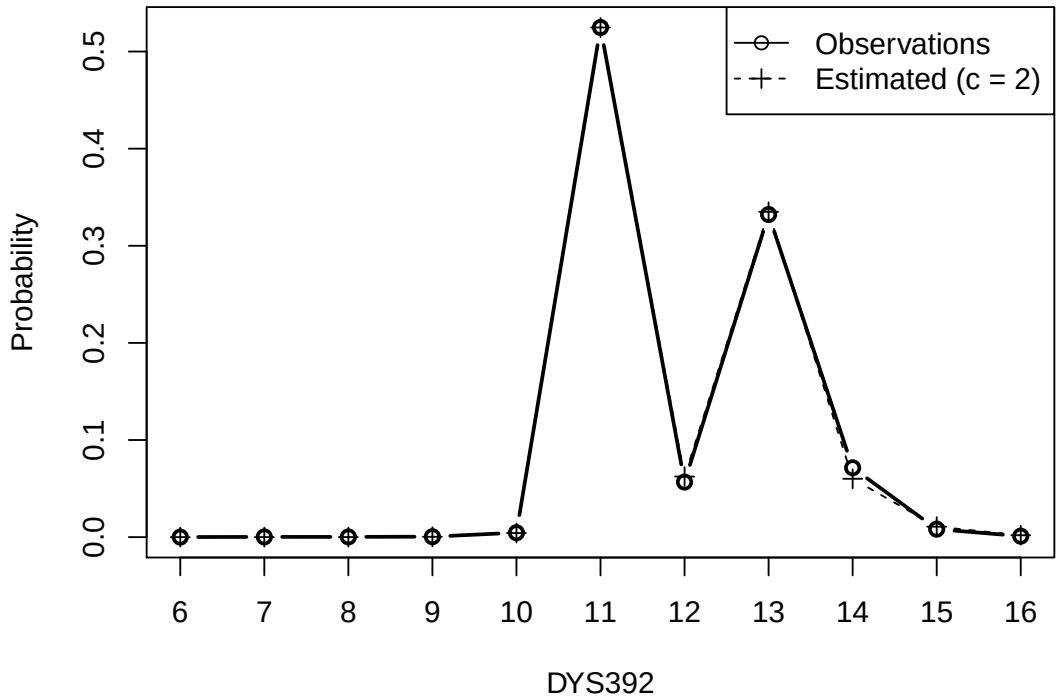


c: Number of subpopulations

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

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

Data and fit



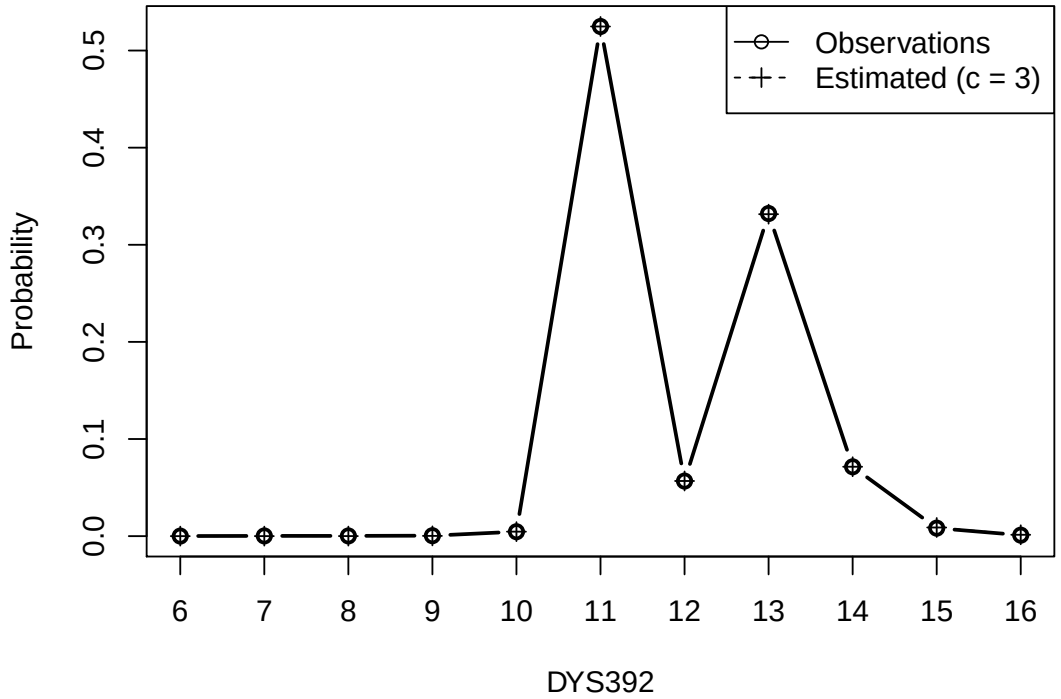
c: Number of subpopulations

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

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

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

Data and fit

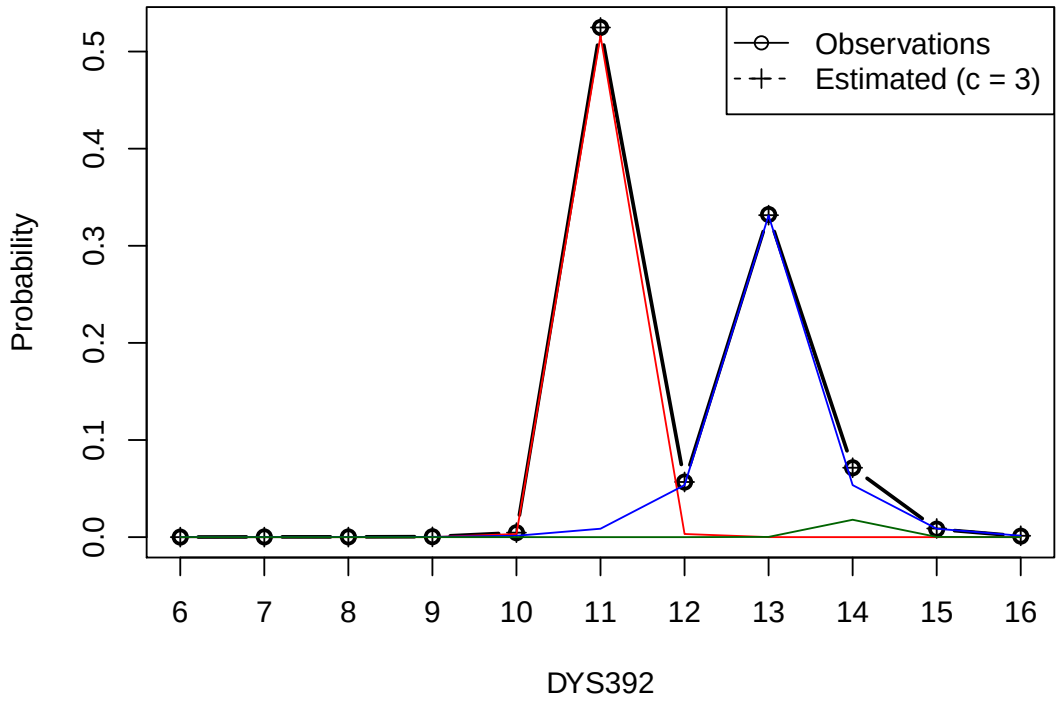


c: Number of subpopulations

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

MM Andersen
mikl@math.aau.dk

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

MM Andersen
mikl@math.aau.dk

MM Andersen
mikl@math.aau.dk

Match probability

Simulation study

Simulation study:

- Simulate populations (each 7 loci and 20 mio individuals)
- Draw random datasets
- Estimate haplotype frequencies of all singletons and compare with the true values
- Result: Smaller prediction error than those with count estimator and Brenner's κ method

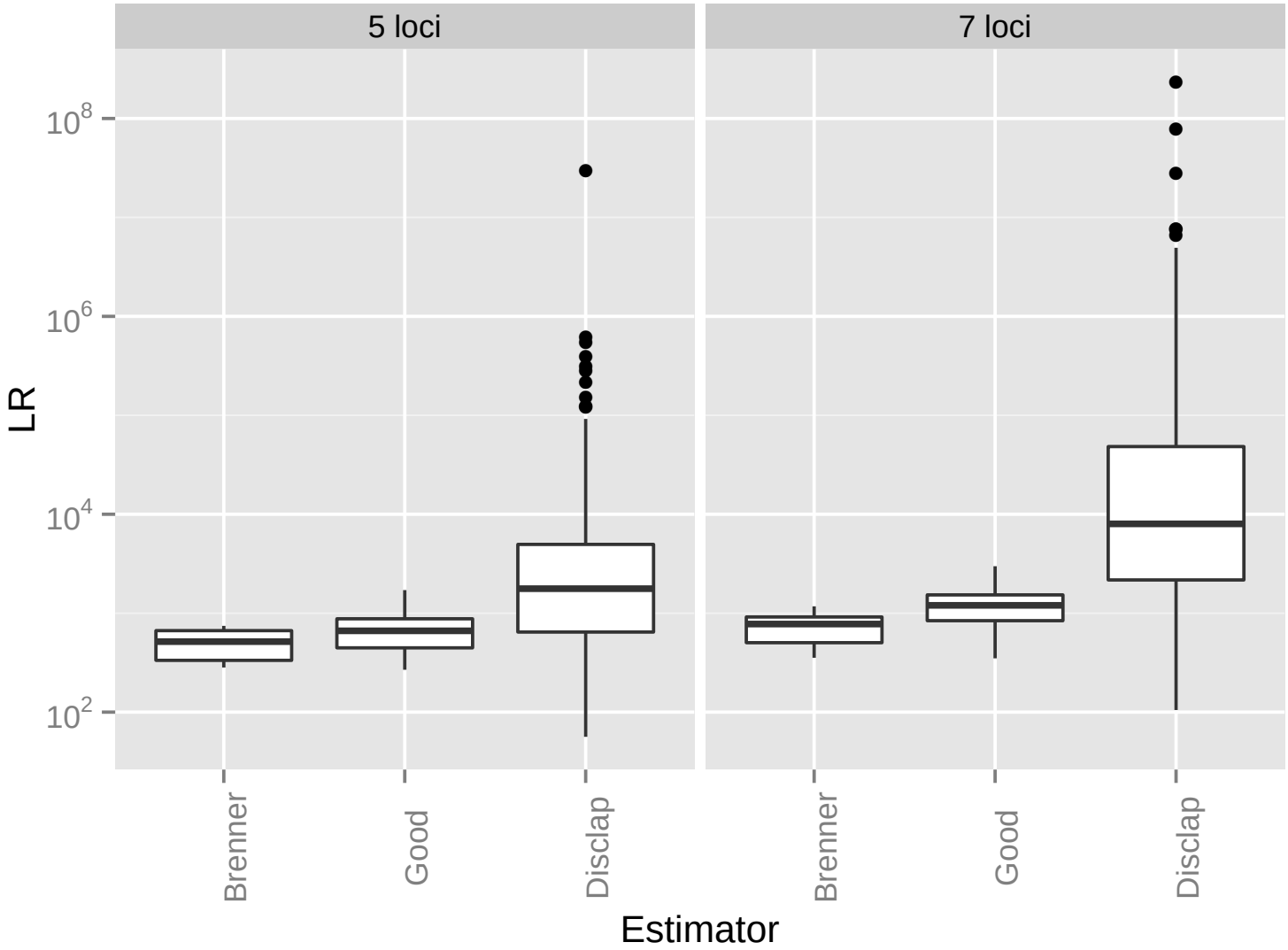
Estimate match probability

Real data (Y23 dataset)

		Singleton proportions	
Population	**Size, n**	**5 loci**	**7 loci**
World	18,925	0.026	0.108
Europe	11,664	0.029	0.101

- Dataset sizes: 200 and 500
- Sampling cases with singleton haplotype:
  1. Draw dataset, D, from population
  2. Draw an extra observation, h
  3. If $h \in D$, skip and go to next sample
  4. If $h \notin D$: Estimate frequency and compare to n_h/n ('true')
- 100 cases for each dataset size, population and locus count
- Compare to Brenner's κ method and Generalised Good

Estimate match probability



MM Andersen
mikl@math.aau.dk

Prediction error

	Probability	LR	LR inflation
Case 1	$p_1 = 0.01$ $\hat{p}_1 = 0.00995$	$LR = 100.0$ $\hat{LR} = 100.5$	$\hat{LR}/LR = 1.005$
Case 2	$p_2 = 0.0001$ $\hat{p}_2 = 0.00015$	$LR = 10{,}000$ $\hat{LR} = 6{,}667$	$\hat{LR}/LR = 0.667$

MM Andersen
mikl@math.aau.dk

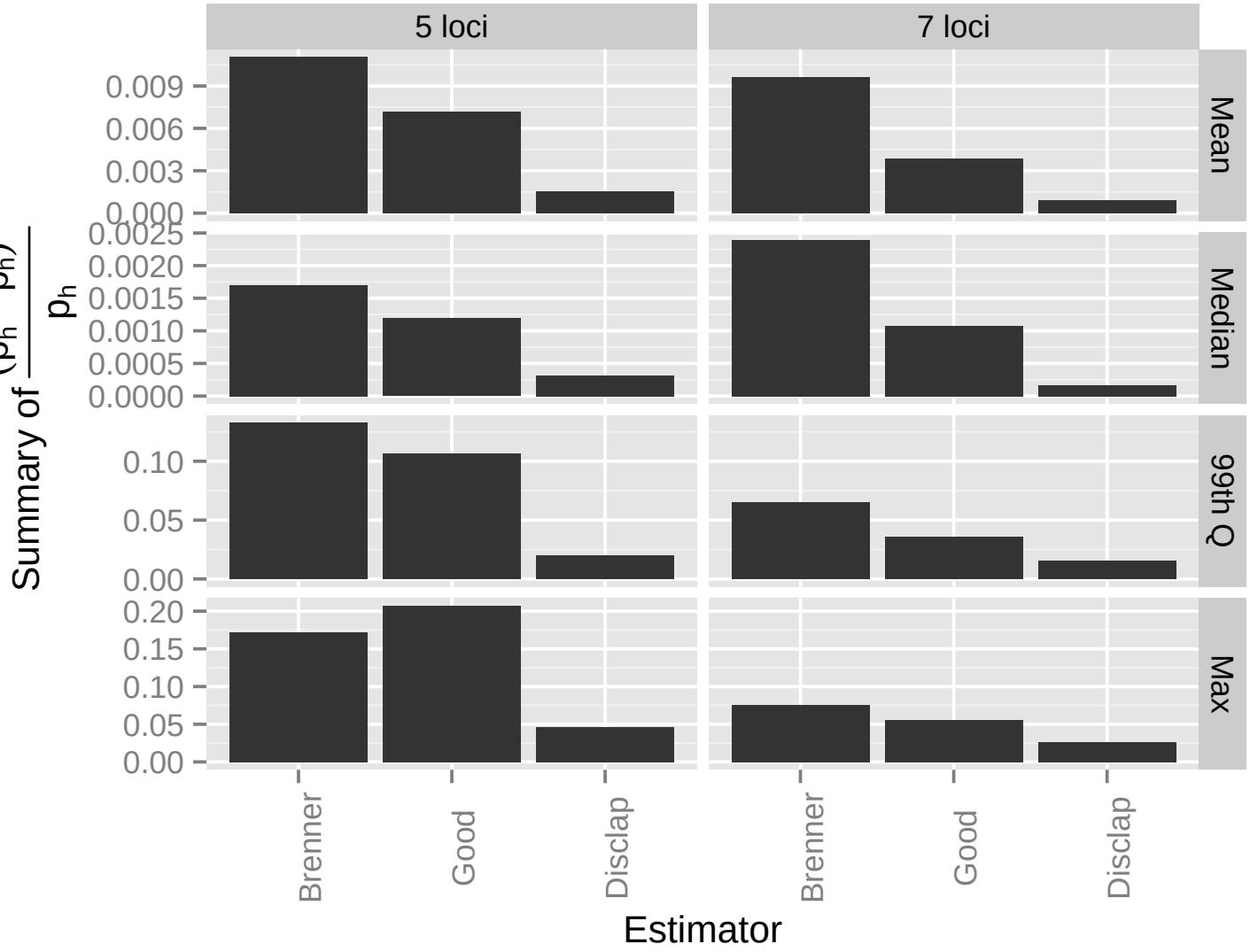
Prediction error

	Probability	LR	LR inflation
Case 1	$p_1 = 0.01$ $\hat{p}_1 = 0.00995$	$LR = 100.0$ $\hat{LR} = 100.5$	$\hat{LR}/LR = 1.005$
Case 2	$p_2 = 0.0001$ $\hat{p}_2 = 0.00015$	$LR = 10{,}000$ $\hat{LR} = 6{,}667$	$\hat{LR}/LR = 0.667$

	Error type				
	$\hat{p}_i - p_i$	$\frac{\hat{p}_i - p_i}{p_i}$	$(\hat{p}_i - p_i)^2$	$\frac{(\hat{p}_i - p_i)^2}{p_i}$	$\log_{10}\left(\frac{\hat{p}_i}{p_i}\right)$
Case 1	-0.00005	-0.005	$2.5 \cdot 10^{-9}$	$2.5 \cdot 10^{-7}$	-0.002
Case 2	0.00005	0.5	$2.5 \cdot 10^{-9}$	$2.5 \cdot 10^{-5}$	0.176

Taking summary (sum, mean, median, ...): What is 0?

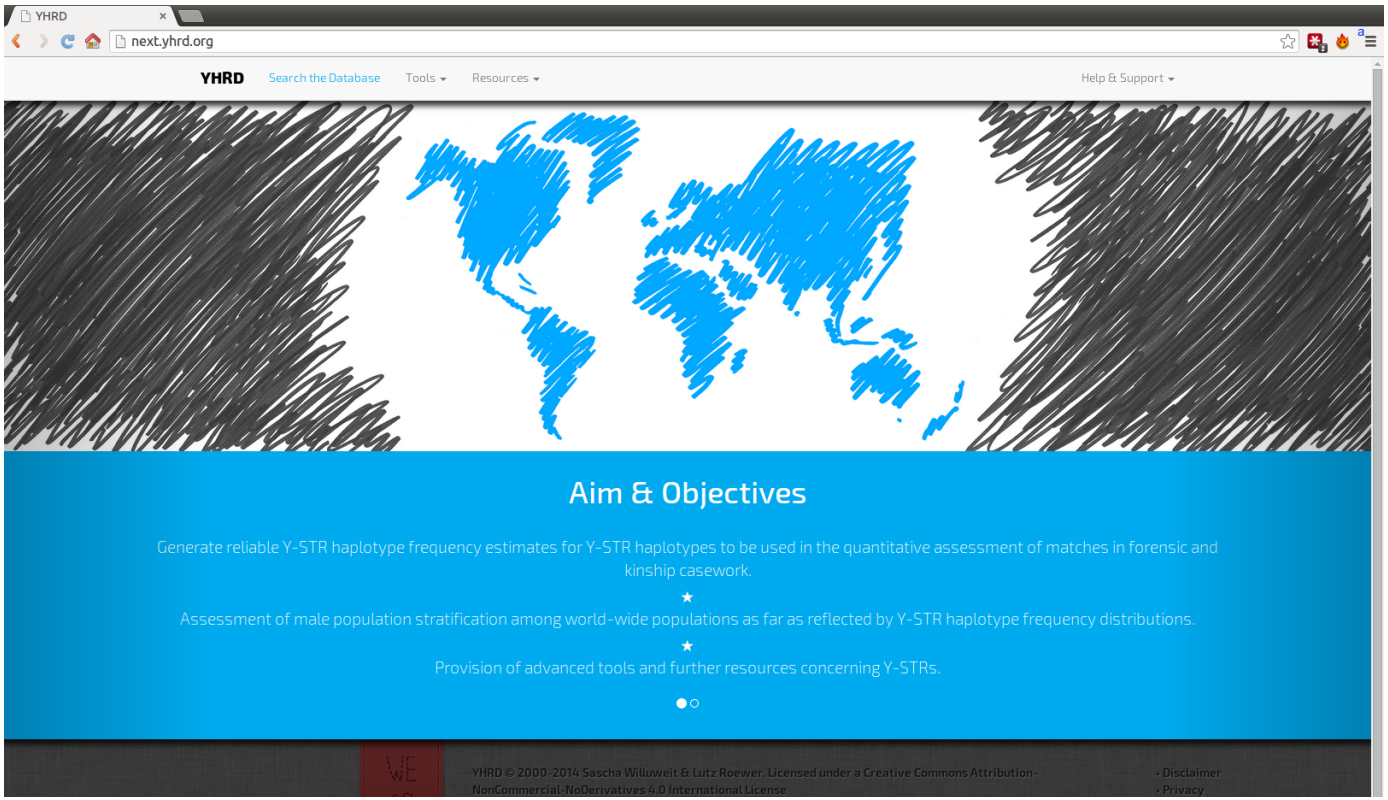
Estimate match probability



MM Andersen
mikl@math.aau.dk

YHRD

- The discrete Laplace method and Brenner's κ is implemented in upcoming version of `http://www.yhrd.org`
- The discrete Laplace method helped finding haplotypes with wrong metapopulation assignments



Mixture separation

Mixture separation

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

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

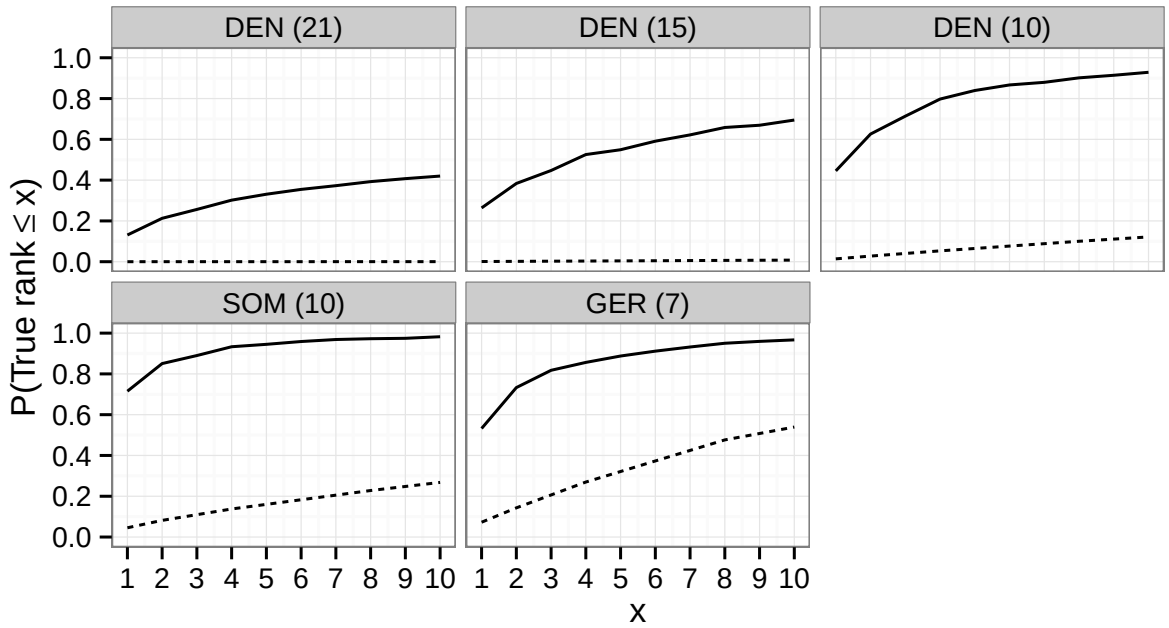
Mixture separation

	Danish			Somali	German
	DEN (21)	**DEN (15)**	**DEN (10)**	**SOM (10)**	**GER (7)**
Loci	21	15	10	10	7
n	181	181	181	201	3,443
Singletons	181 (100%)	164 (90.6%)	112 (61.9%)	56 (27.9%)	662 (19.2%)

- 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

Probabiliy	DEN (21)	DEN (15)	DEN (10)	SOM (10)	GER (7)
Rank $\leq$ 1	13%	26%	45%	72%	53%
Rank $\leq$ 5	33%	55%	84%	94%	89%
Rank $\leq$ 10	42%	69%	93%	98%	97%
Random $\leq$ 10	0.03%	0.78%	12.15%	26.79 %	53.93%

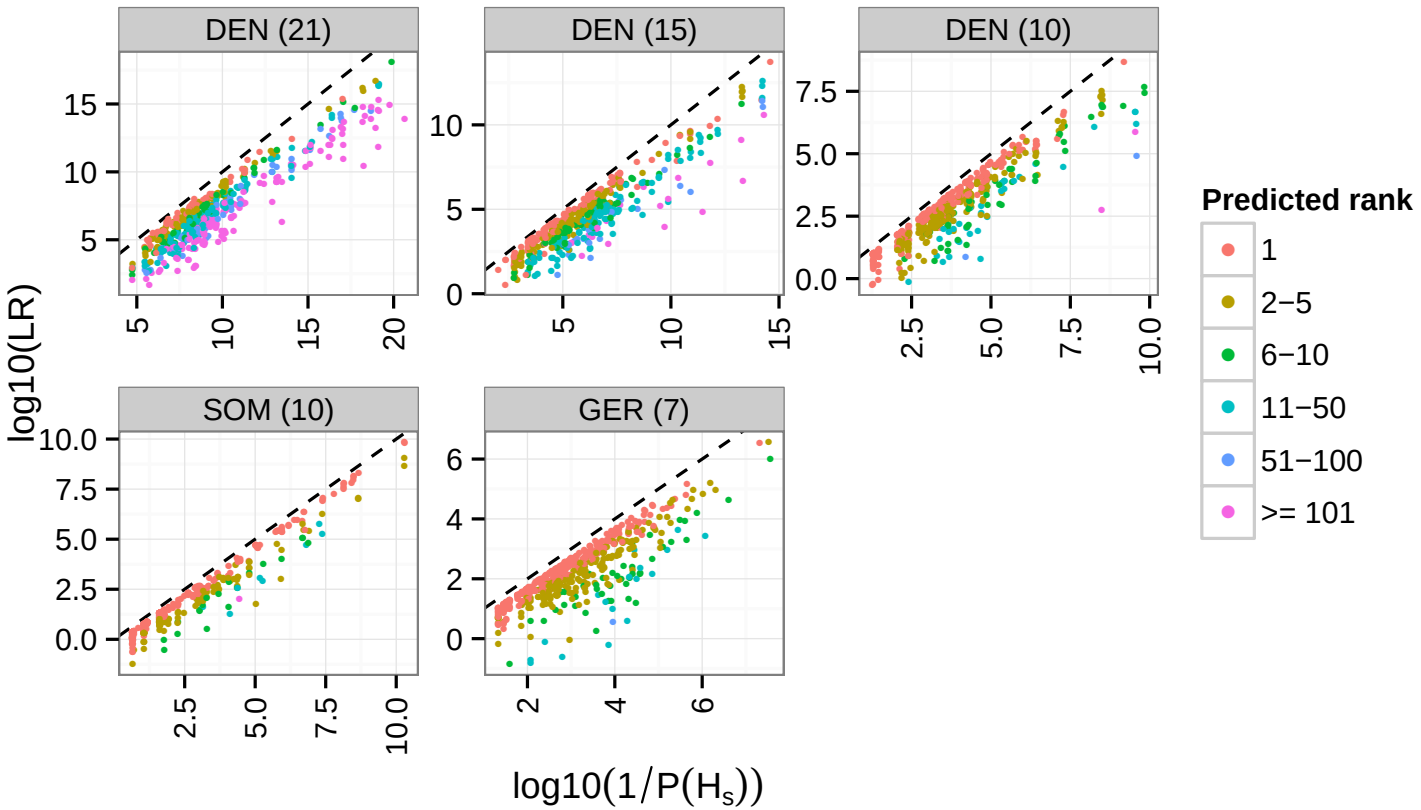


MM Andersen
mikl@math.aau.dk

Mixture *LR*

$$H_p : S + U$$
$$H_d : U_1 + U_2$$

$$LR = \frac{P(H_U)}{\sum_{(H_{U_1}, H_{U_2})} P(H_{U_1})P(H_{U_2})}$$



MM Andersen
mikl@math.aau.dk

Cluster analysis

Cluster analysis

- $\tau_j = P(\text{From subpopulation } j)$
- Haplotype frequency by summing the contributions from each subpopulation:

$$P(\text{Haplotype} = x) = \sum_{j=1}^{c} \tau_j \cdot \underbrace{P(\text{Haplotype} = x \mid \text{From subpopulation } j)}_{\text{Discrete Laplace model}}.$$

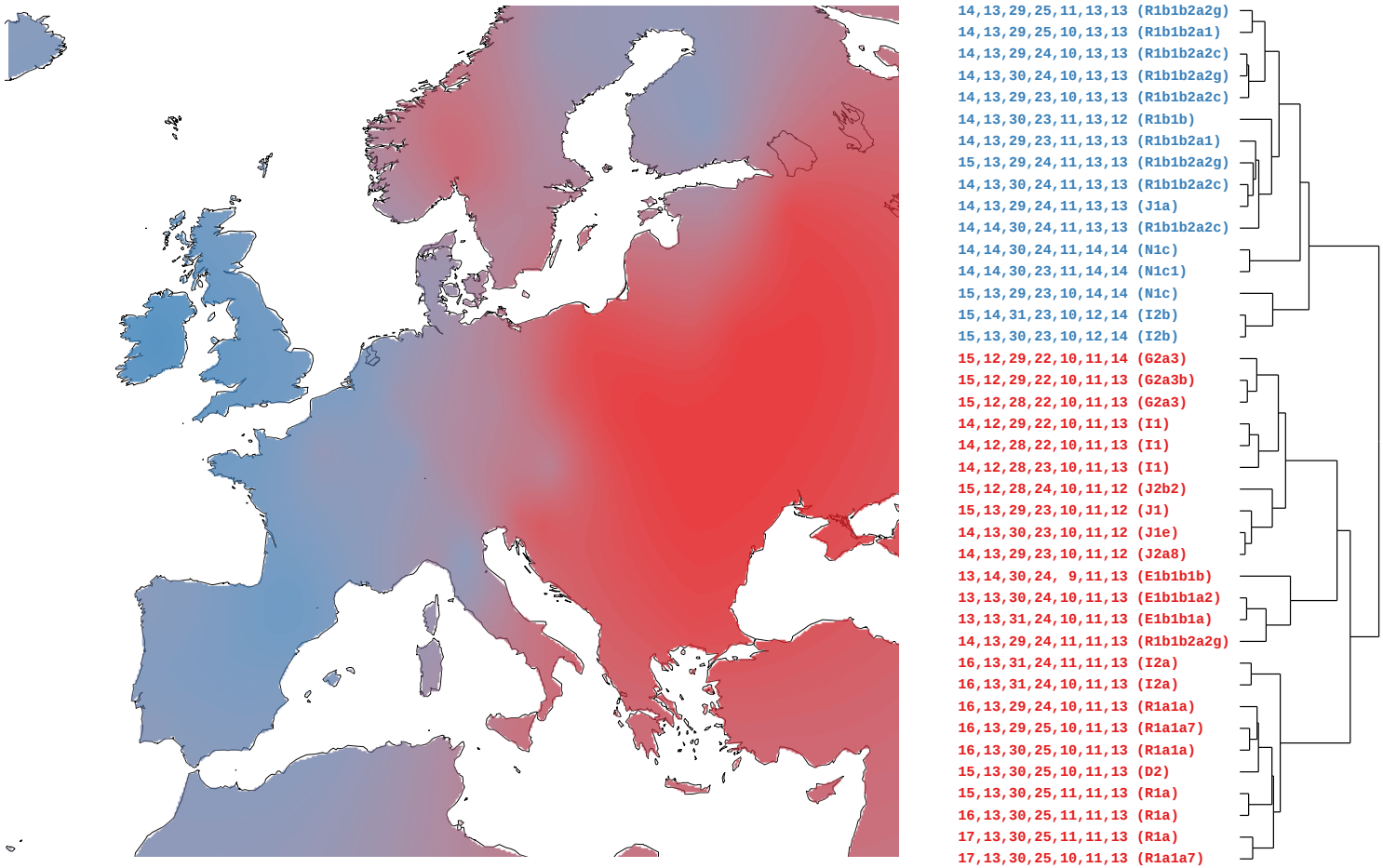
- Bayes theorem:

$$P(\text{From subpopulation } j \mid \text{Haplotype} = x) = \frac{\tau_j \cdot P(\text{Haplotype} = x \mid \text{From subpopulation } j)}{P(\text{Haplotype} = x)}$$

MM Andersen
mikl@math.aau.dk

Cluster analysis of European data

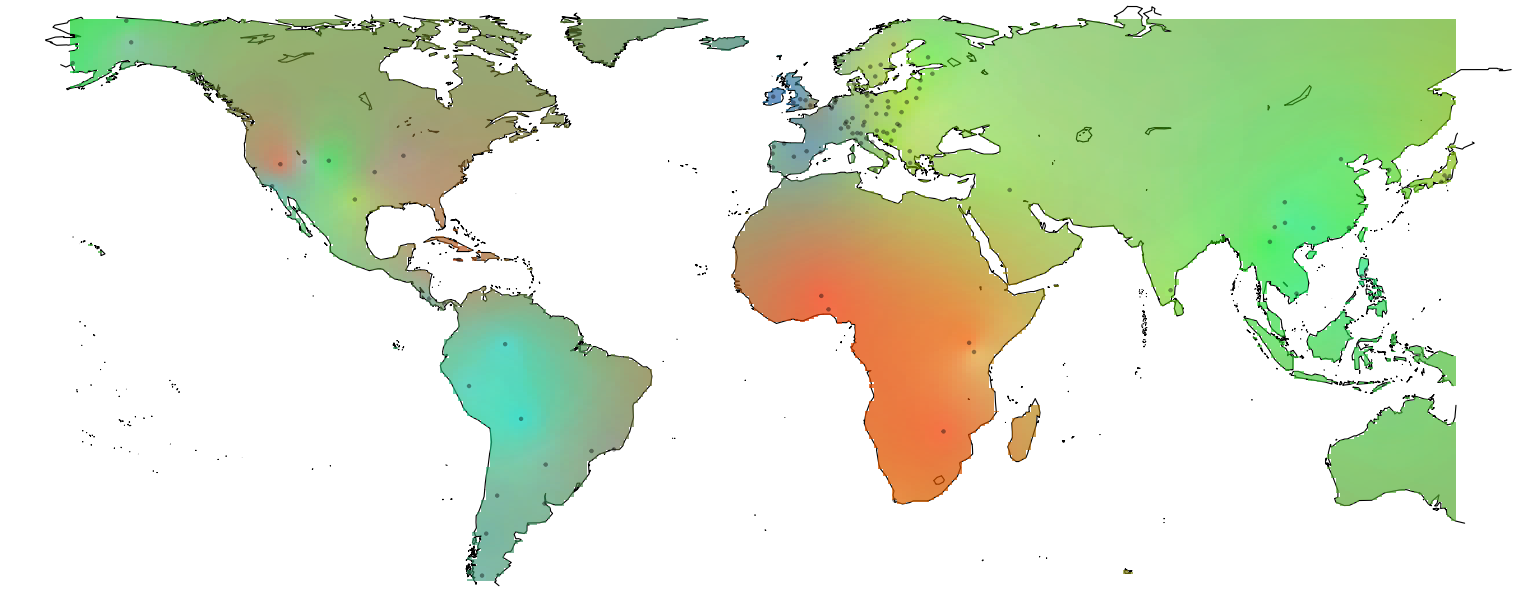
7 loci

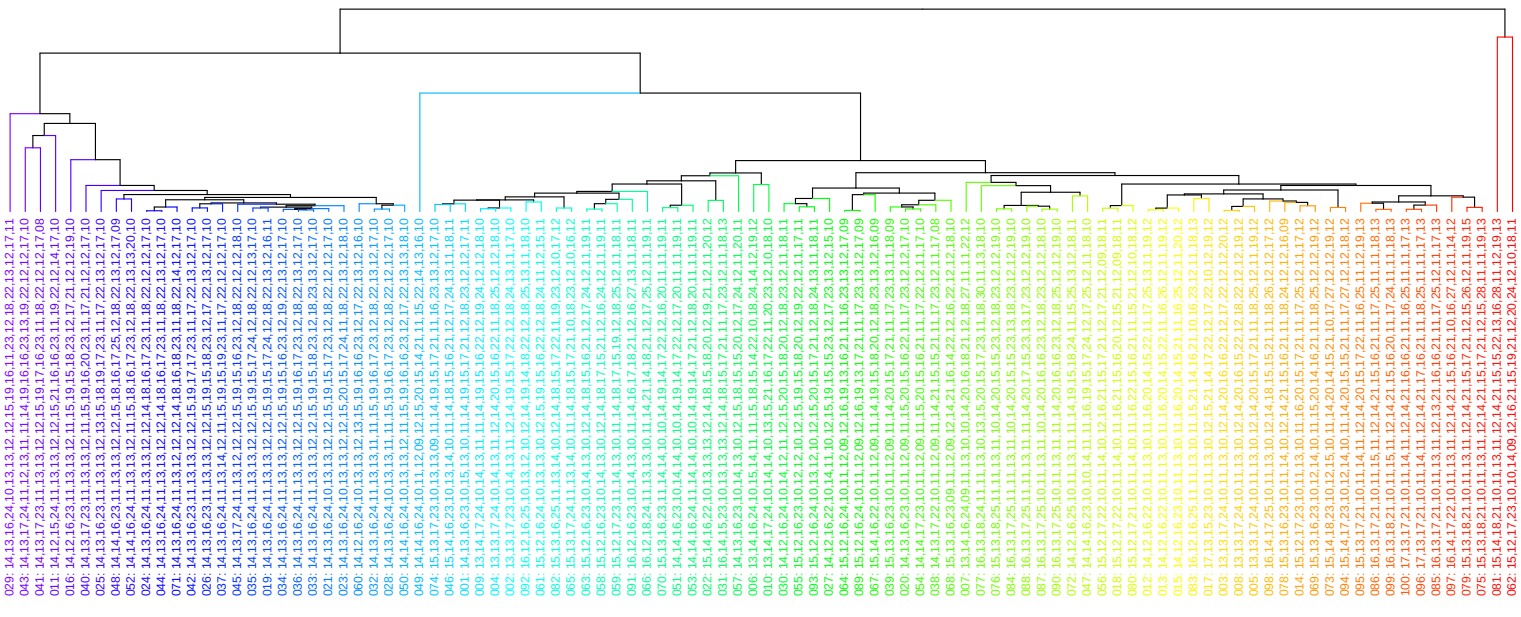


First analysed in 'Signature of recent historical events in the European Y-chromosomal STR haplotype distribution' by Roewer et al. in 2005

Cluster analysis of Y23

21 loci (from Purps J, Siegert S, et al. (2014))





Pairwise population distances

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
- 21-locus (Y23), 18,925 males (129 locations): Correlation(AMOVA, discrete Laplace) = 0.78

MM Andersen
mikl@math.aau.dk

Concluding remarks

The discrete Laplace method

- Sound statistical properties
- Applications
  - Estimation of Y-STR haplotype population frequencies
  - Mixture analysis
  - Cluster analysis
- Computationally feasible
- Open source software: `R` libraries `disclap` and `disclapmix` (and `fwsim` for simulating populations)
- Criticism
  - Intermediate alleles (e.g. 10.2)
  - Duplications (e.g. DYS385a/b)
  - Copy number variation (e.g. Yfiler Plus)
  - $\hat{\mu}$'s difficult to estimate (curse of dimensionality)

Conclusion

- Match probability is of great interest and is difficult
- Validation of methods
  - Open source software (e.g. `R` library, C++ program): Compare to your own method
  - Availability of real data (PPY23)
    - Purps J, Siegert S, et al. (2014). *A global analysis of Y-chromosomal haplotype diversity for 23 STR loci*. Forensic Science International: Genetics, Volume 12, 2014, p. 12-23. `http://dx.doi.org/10.1016/j.fsigen.2014.04.008`
    - `R`-object: `http://people.math.aau.dk/~mikl/?p=y23`
  - Battery of simulated populations
  - Measure of prediction error
- For a matching profile (e.g. Y23 or Yfiler Plus), use only subset (e.g. 7 or 10 loci) for *LR* calculations?
  - Easier to validate

MM Andersen
mikl@math.aau.dk

Thank you for your attention