Cluster analysis of Y-chromosomal STR population data using discrete Laplace distributions

25th ISFG Congress, Melbourne, 2013

Mikkel Meyer Andersen[1,*], Poul Svante Eriksen[1] and Niels Morling[2]

[1] Department of Mathematical Sciences, Aalborg University, Denmark

[2] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

[*] mikl@math.aau.dk

AALBORG UNIVERSITY
DENMARK

Introduction

- Count method
- Frequency surveying (2000, 2001, 2010)
- Brenner (2010)
- Coalescent method (2013)

Statistical model for Y-STR haplotypes

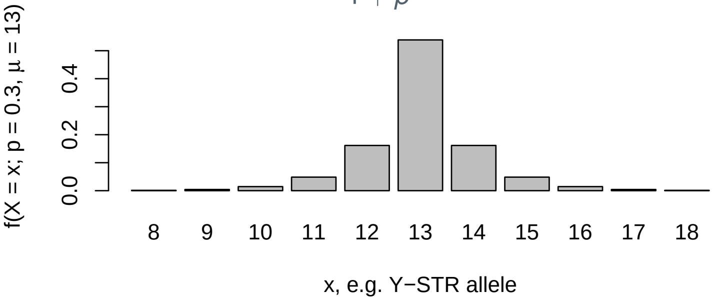
...based on the discrete Laplace distribution

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:

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



Statistical model for Y-STR haplotypes

...based on the discrete Laplace distribution

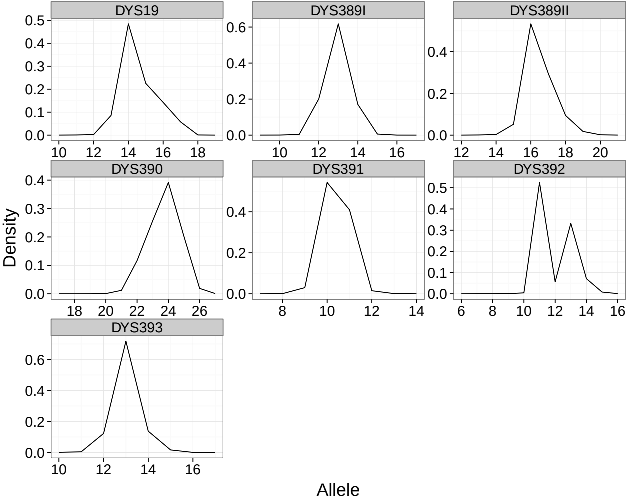
Perfectly homogeneous population with a 1-locus haplotype:

$$P(X = x) = f(X = x; p, \mu) = \frac{1-p}{1+p} \cdot p^{|x-\mu|}$$

- p depends on e.g. mutation rate and population growth
- μ is central haplotype/allele
- Motivated by theoretical result for Fisher-Wright population with neutral single step mutation model

Statistical model for Y-STR haplotypes

...based on the discrete Laplace distribution



Statistical model for Y-STR haplotypes

...based on the discrete Laplace distribution

- L loci: Neutral mutations across loci are assumed independent
- S subpopulations/clusters: Mixture of distributions
  - τ_s is the a priori probability for originating from the s'th subpopulation ($\sum_{s=1}^{S} \tau_s = 1$)
- Parameter estimation from database of Y-STR haplotype (`R` library `disclapmix`)

Statistical model for Y-STR haplotypes

...based on the discrete Laplace distribution

Simulation study

- 60 different (growth, mutation rate, initial size) populations of size 20,000,000
- From each population: 50 data sets of 500, 1,000 and 5,000 Y-STR-profiles sampled (total: 9,000 data sets)
- Smaller prediction error than existing estimators (like naïve count and Brenner's method)

Statistical model for Y-STR haplotypes

...based on the discrete Laplace distribution

- Let $\{x_i\}_{i=1}^{n}$ be a database of n Y-STR haplotypes (one x_i per observation)
- Subpopulation membership:

$$v_{is} = \begin{cases} 1 & \text{if individual } i \text{ originates from subpopulation } s \\ 0 & \text{otherwise} \end{cases}$$

- An individual can originate from only one subpopulation
- Membership not observed, infer the probability of each outcome:

$$\hat{v}_{is} = P(v_{is} = 1 \mid x_i)$$

Danish data

- Example with 7-locus Y-STR haplotypes from $n = 63$ Danes (43 are singletons)
- Sum of observed probability

$$\underbrace{1 - \frac{43}{63+1}}_{\text{Robbins (1968)}} = 0.328,$$

- Discrete Laplace method: $\hat{S} = 3$ and sum of observed probability

$$\sum_{x \in \mathrm{DB}_u} P(X = x) = 0.318$$

where DB_u is the different (unique) haplotypes in the database.

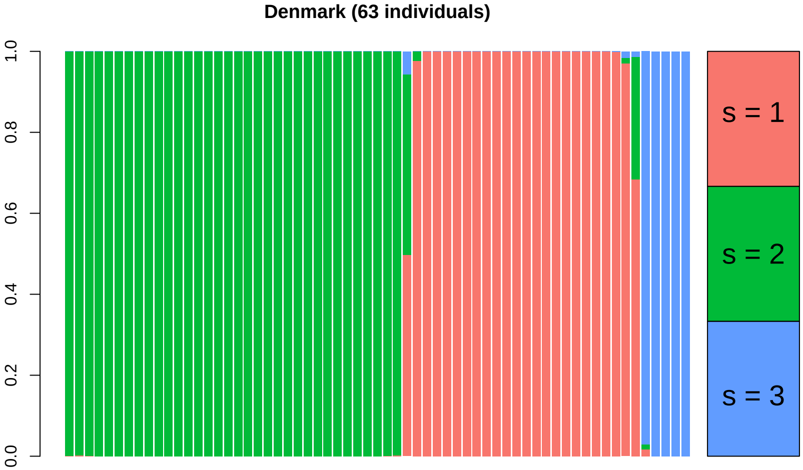
Danish data

Example with 7-locus Y-STR haplotypes from $n = 63$ Danes (43 are singletons)

		$s = 1$	$s = 2$	$s = 3$
	$\hat{\tau}_s$ (a priori probability)	0.37	0.55	0.08
Central haplotype, $\hat{\mu}_s$	DYS19	14	14	15
	DYS389I	12	13	14
	DYS389II	28	29	32
	DYS390	22	24	23
	DYS391	10	11	10
	DYS392	11	13	12
	DYS393	13	13	14
	$n_{\hat{\mu}_s}$ (# in database)	4	4	1
	Haplogroup (`yhrd.org`)	I	R1b	I2/I2b

Danish data

3 clusters/subpopulations: $\hat{v}_{is}$ values for each individual



Columns: Individuals. Rows: Subpopulations. Bar at column i, row s: $\hat{v}_{is}$

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 (including the optimal number of subpopulations, $\hat{S}$)
  - Parameters (genetic information) versus known sample locations
  - Discrete Laplace model does not know about sample locations, it infers ’genetic’ subpopulations (or clusters)

European data

- Sample locations: $r = 1, 2, \ldots, R$ ($R = 91$)
- Subpopulations: $s = 1, 2, \ldots, S$ ($\hat{S} = 40$)
- w_{rs}: Fraction of individuals from location r originating from subpopulation s

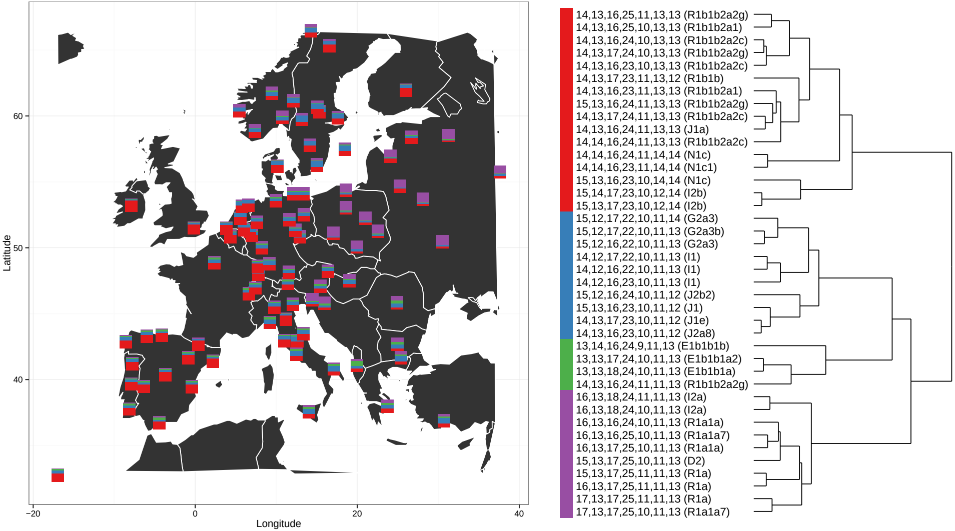
w_{rs} values for selected subpopulations and regions:

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

Empty cell means 0.0.

European data

Collapsed w_{rs} values for 4 mega clusters



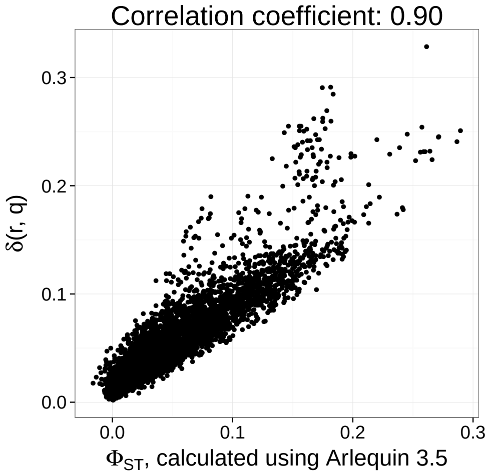
European data

- AMOVA (Excoffier, 1992): Pairwise Φ_{ST} distances calculated with Arlequin version 3.5
- Distance between sample location r and sample location q

$$\delta(r, q) = \sum_{s=1}^{S} (w_{rs} - w_{qs})^2$$

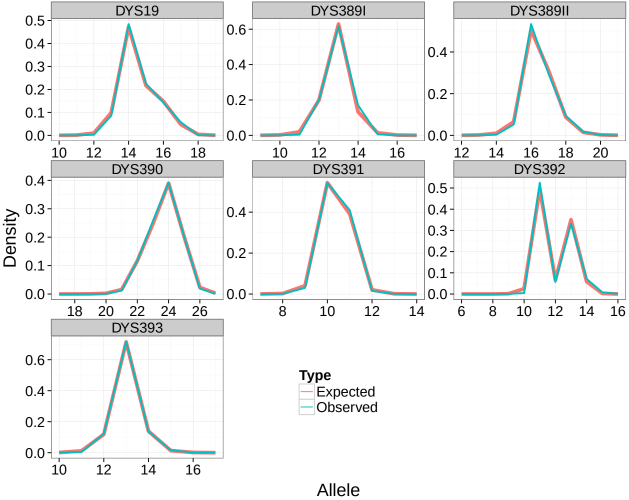
  - Squared Euclidean distance between vector $(w_{r1}, w_{r2}, \dots, w_{rS})$ and vector $(w_{q1}, w_{q2}, \dots, w_{qS})$

European data



European data

Marginal distribution



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 study and AMOVA
- Computationally feasible
- Open source software: `R` library `disclapmix` (tutorial available online)