Evidential strength of a Y-STR match

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

1 Dept. of Mathematical Sciences, Aalborg University, Denmark
2 Sect. of Forensic Genetics, Dept. of Forensic Medicine, University of Copenhagen, Denmark
* mikl@math.aau.dk

Introduction

Two important issues

1. Estimate Y-STR haplotype frequencies of rare and unseen Y-STR haplotypes
2. **Compensate for the possible lack of information concerning the relevant population referred to in the defence hypothesis,** H_d

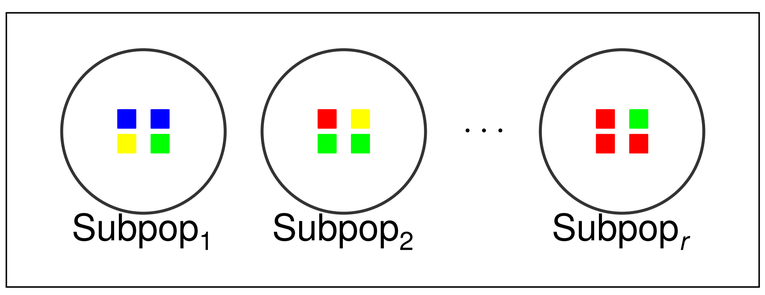
Elaborating H_d

- H_d: 'A random man left the Y-chromosome DNA in the crime stain'
- H_d: 'A random man **(from the population of which the reference database is a random sample)** left the Y-chromosome DNA in the crime stain'

Population? What is that? Does it matter? Yes

Population substructure

A population is a collection of subpopulations. Haplotypes are more common in some subpopulations than in others.



Population

Coloured squares represent haplotypes.

We have a sample from the population without complete substructure information.

Frequency of ■ in:

- Entire population? (4/12 = 1/3)
- Subpopulation 1? (0/4 – unseen)
- Subpopulation r? (3/4)

Match probability

Imagine that we have a random sample from the entire population.

1. H_d: 'A random man from the entire population left the Y-chromosome DNA in the crime stain'
2. H_d: 'A random man from **subpopulation** r left the Y-chromosome DNA in the crime stain'
3. H_d: 'A random man from **the same subpopulation as the suspect** left the Y-chromosome DNA in the crime stain'

Ad 3: A haplotype may be more frequent in a subpopulation than in the population. But no information about population substructure.

Balding-Nichols model with population structure:

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

- θ (theta) ($0 < \theta < 1$): Population structure parameter
- p_h: Population frequency of h ($0 < p_h < 1$)

θ (theta)

θ (theta) ($0 < \theta < 1$):

- Population parameter (related to the variability of haplotype frequencies in different subpopulations)
- Not subpopulation specific (an average)
- Not haplotype specific (an average)
- Can be estimated using databases from two or more subpopulations (assumed known structure)

The Balding-Nichols match probability, $\theta + (1-\theta)\mathbf{p_h}$, is larger than both θ and p_h.

If a random man and the suspect belong to the same subpopulation, they are expected to share the Y-STR haplotype more often than if they do not belong to the same subpopulation.

Estimating θ

- Use geographical information: Sample assumed subpopulations (populations without substructure), e.g. islands, cities (or even countries) separately (at the right level)
- θ between countries may be different from θ between cities/islands within a country

Examples

Example 1: Danish reference database. We assume no population substructure (haplotype distribution same in cities and small islands).

- H_d: 'A random Dane left the Y-chromosome DNA in the crime stain'
- Use population frequency, p_h, based on a Danish reference database (and no θ correction)

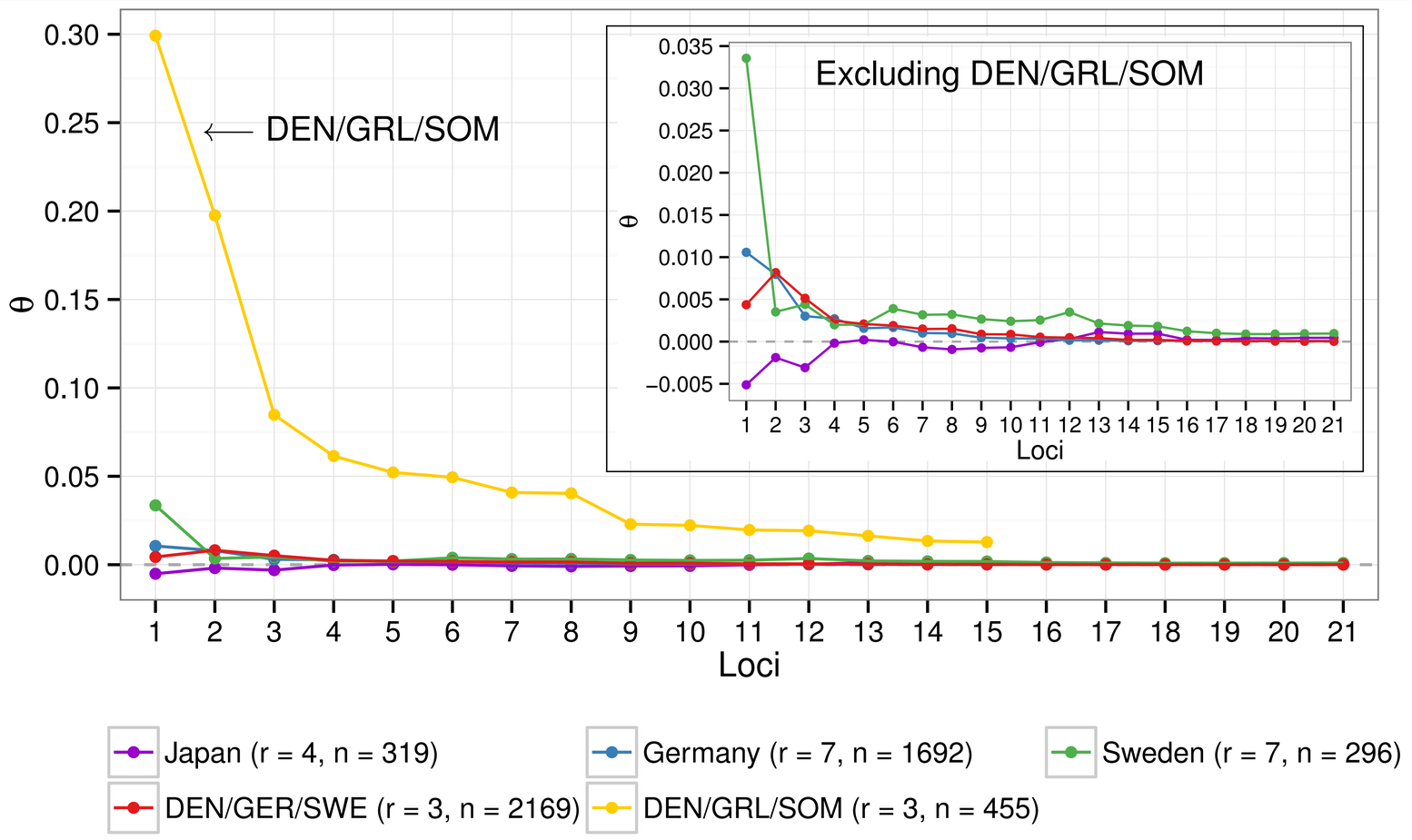
Example 2: Danish reference database. We assume population substructure such that the haplotype distributions may differ, e.g. in cities and small islands.

- H_d: 'A random Dane originating from the same small island, Bornholm, as the suspect left the Y-chromosome DNA in the crime stain'
- Use θ correction: $\theta + (1-\theta)p_h$ with known θ and population frequency, p_h, based on a Danish reference database

Example 3: Reference database from Bornholm (small Danish island). We assume no population substructure (haplotype distributions same in cities and rural areas).

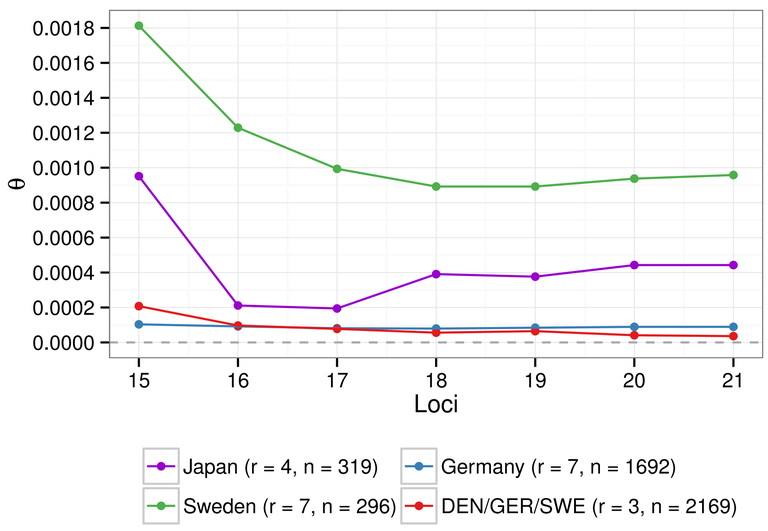
- H_d: 'A random man from Bornholm left the Y-chromosome DNA in the crime stain'
- Use population frequency, p_h, based on a reference database from Bornholm (and no θ correction)

Examples of θ values (1)



Estimated with a method by Bruce Weir (pers. comm.).

Examples of θ values (2)



Considerations

- θ (theta) correction in the form presented is a remedy for using a 'wrong' reference database, i.e. not taking population substructure into account
- How can we identify subpopulations?
- Can population data of ethnic groups be used to estimate θ in large cities?
- Many loci: Use θ as the match probability – θ dominates because most haplotypes are very rare?

Thanks to Bruce Weir and John Buckleton for discussions about θ.