Identifying clusters of related individuals with data on dominant genetic markers

Karl W Broman

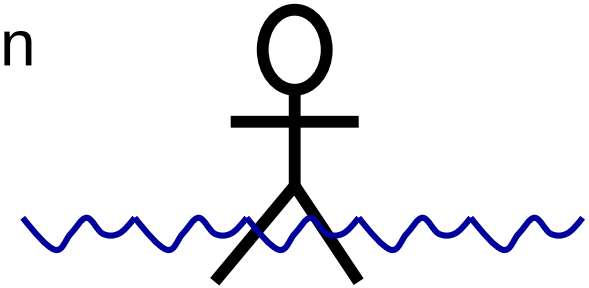
Dept of Biostatistics
Johns Hopkins University

Joint work with Laura Plantinga and Clive Shiff

http://biosun01.biostat.jhsph.edu/~kbroman

Schistosoma haematobium

Sexual reproduction within humans



Excrete eggs

Miracidia

Penetrate snail

Asexual reproduction within snails

Cercariae

Penetrate skin

Samples

- ~ 4 populations of people
- ~ 10 subjects per population
- Eggs from a urine sample from each subject
- Egg $\rightarrow$ snail $\rightarrow$ many clones $\rightarrow$ DNA
- For each egg (worm), binary data at ~40 RAPD loci
- Final data: ~ 200 worms $\times$ ~ 40 loci

Imagine: Worms composed of clusters of siblings; groups of siblings are unrelated

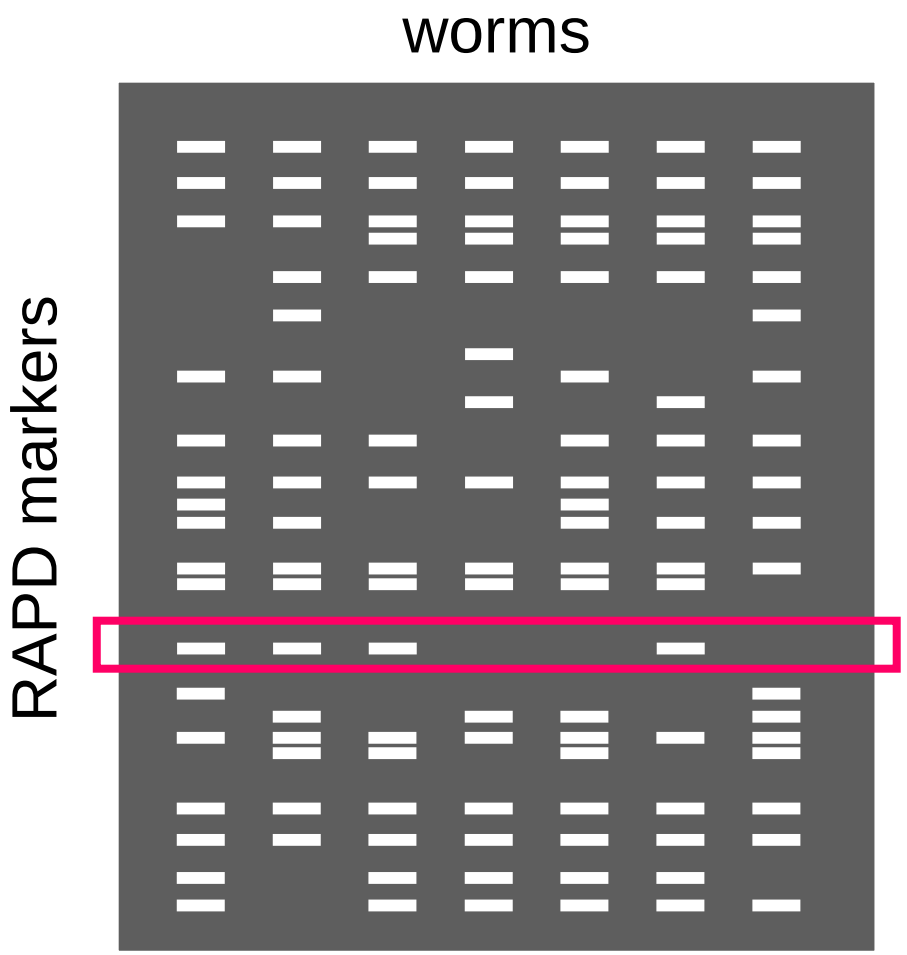
Goals

Major goal: Understand the population genetics of Schistosome

Minor goal: Identify clusters of related individuals via genetic marker data

Current goal: Inference of full-sibling families [After Apostol et al. 1993]

RAPD genetic markers



- Short “random” primers + PCR → DNA fragments
- Gel electrophoresis + stain DNA → bands
- RAPD marker: presence or absence of a particular band
- At each locus:

  B = band allele
  N = no-band allele

  BB/BN → band
  NN → no band

A bit of genetics

- Worms (like humans) have 2 copies of each chromosome
- **Hardy-Weinberg equilibrium**:

  A worm's genotype is a random union of alleles

  p = Pr(band allele); q = 1 – p

  Pr(no band) = q^2; Pr(band) = 1 – q^2

- For a pair of *unrelated* worms:

  [N = neither; B = both; D = mismatch]

  Pr(N) = q^4 Pr(B) = $(1 – q^2)^2$

- For a pair of *siblings*:

  $Pr(N) = (1/4)\ p^2\ q^2 + p\ q^3 + q^4$

  $Pr(B) = p^2 + 2\ p^3\ q + (13/4)\ p^2\ q^2 + p\ q^3$

The procedure

- **Form a measure of distance**

  Proportion of mismatches
  (drop less-informative loci)

- **Hierarchical clustering:** distance $\rightarrow$ dendrogram

  "UPGMA" (agglomerative w/ average distance)

- **Pick a cutoff:** dendrogram $\rightarrow$ clusters

  Average distance between siblings

- **[Assess the quality of the results]**

  [Rand index]

Pairwise distances

- **Proportion of mismatches**
  - Simple, fast
  - 1/1 and 0/0 treated the same
  - Varying frequencies not taken into account
- **Log likelihood ratio**
  - log{ Pr(data | unrelated) / Pr(data | siblings) }
  - Re-center so distance > 0
  - Slower; better?
- **How to deal with missing data?**

An example

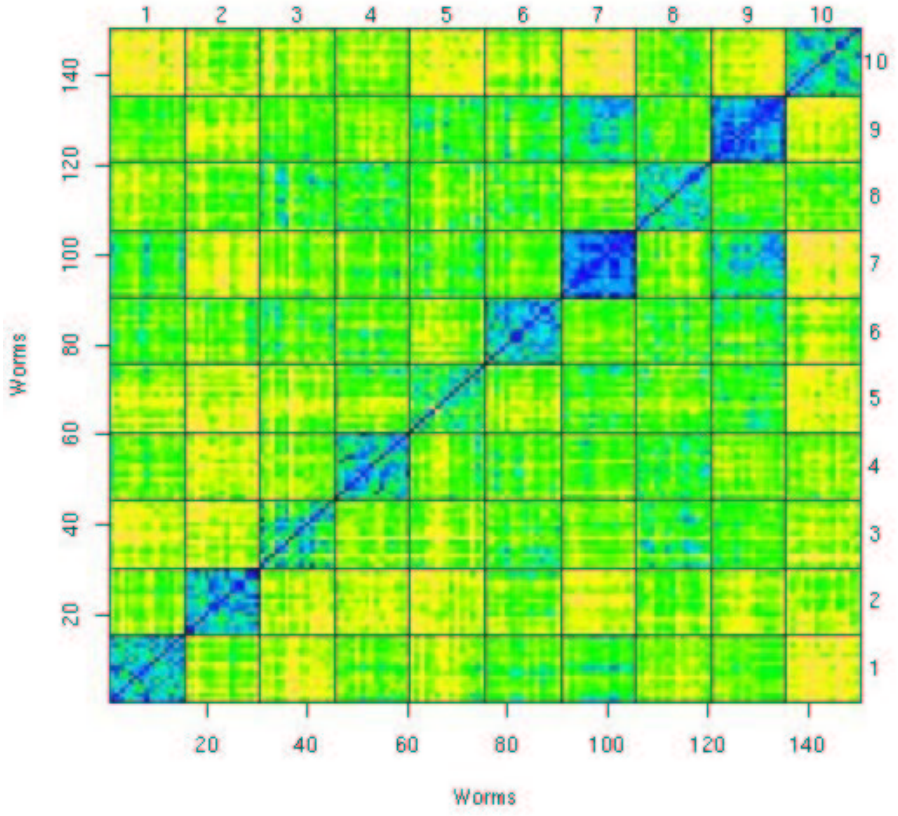
Apostol et al. (1993) Theor Appl Genet **86**: 991–1000

- Mosquito: *Aedes aegypti*
- 10 groups of 15 *known siblings*
- Data on 40 RAPD markers

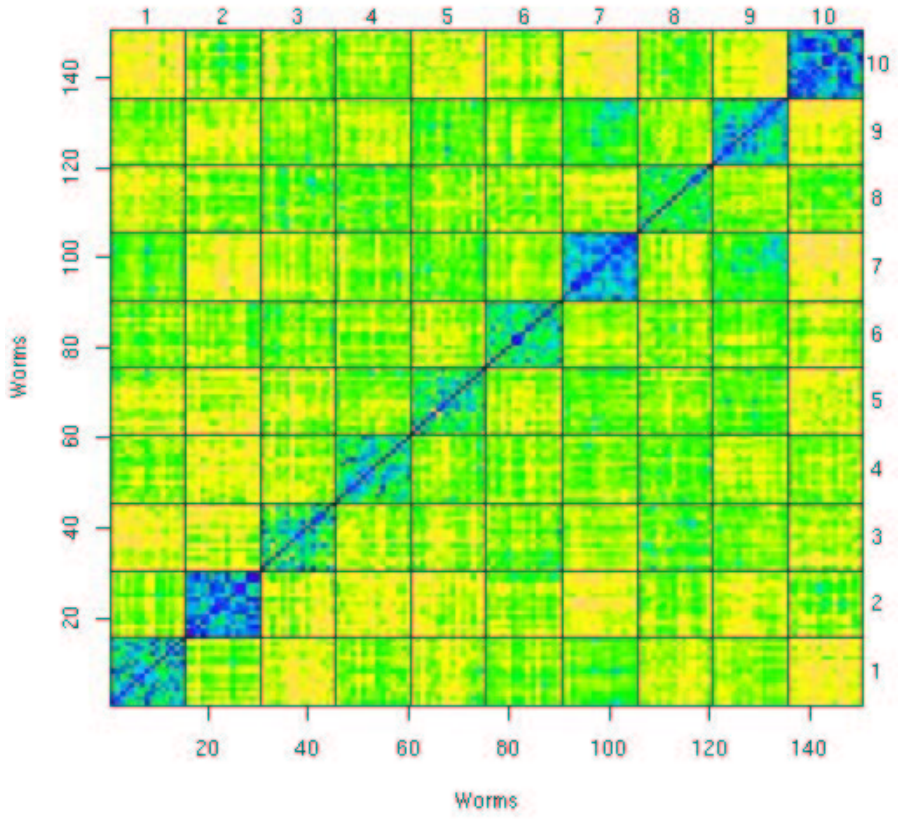
[Insufficient Schistosome data at this point.]

Pairwise distances in example

Proportion mismatches

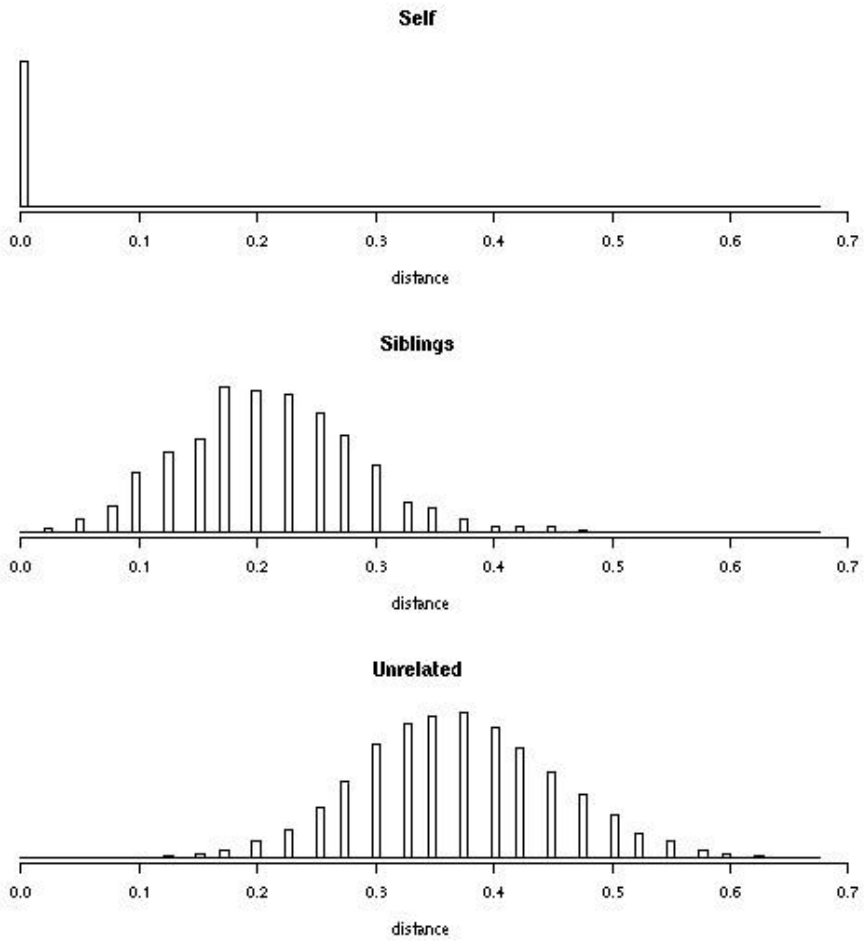


Log likelihood ratio

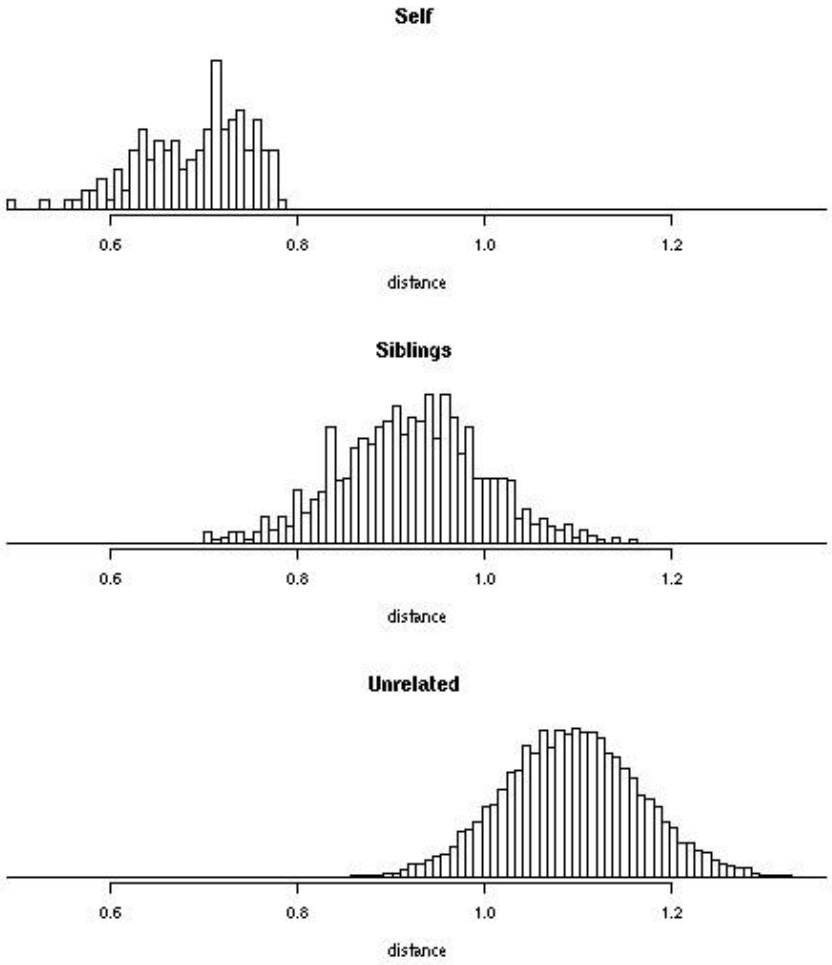


Distributions of distances

Proportion mismatches



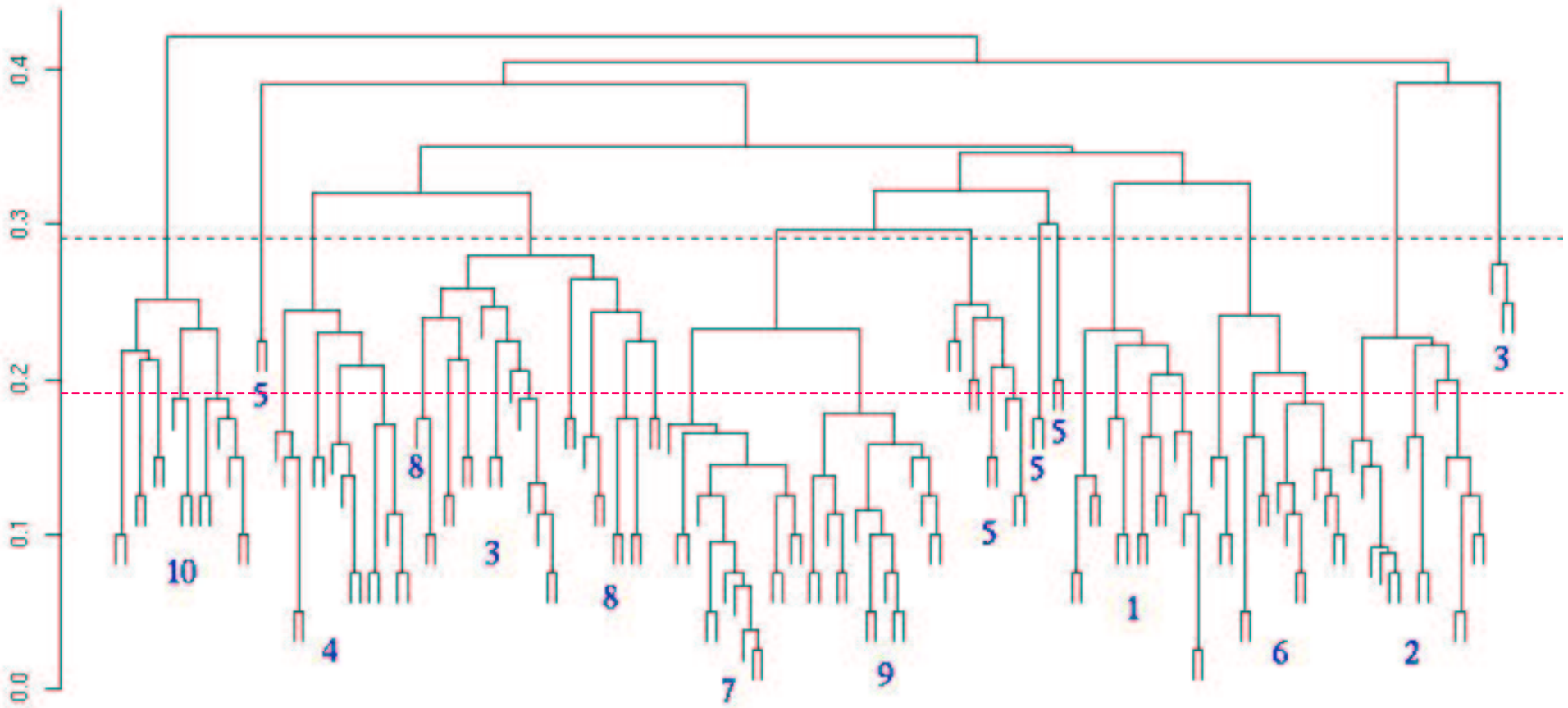
Log likelihood ratio



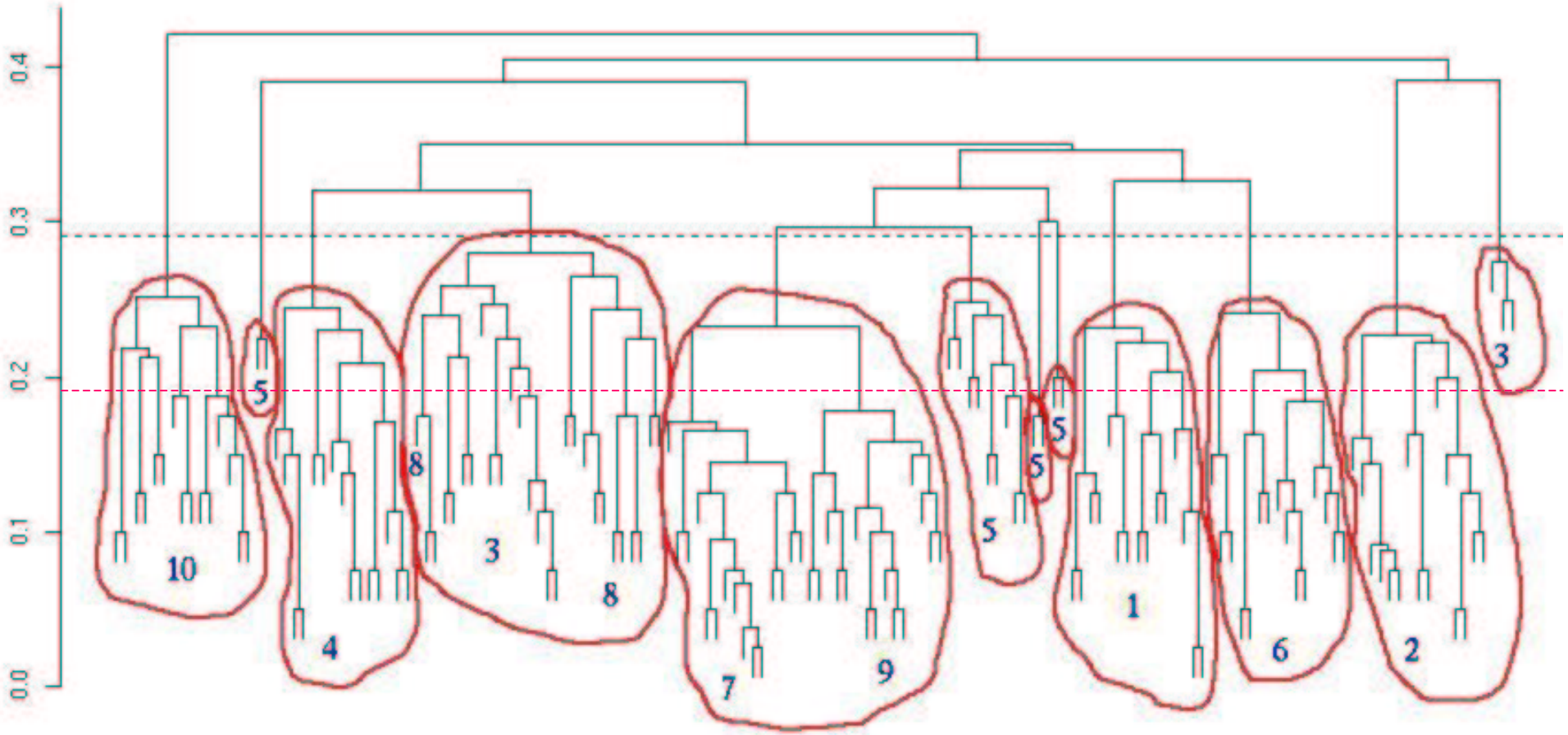
Cutoffs for dendrogram

- Consider distribution of distances between siblings or distribution of distances between unrelateds
  - Easy to calculate mean and SD for each
  - Normal approximation appropriate in case of many loci
- Consider some attribute of these distributions
  - Mean among siblings
  - 20^{th} percentile among unrelateds
  - Something Bayesian

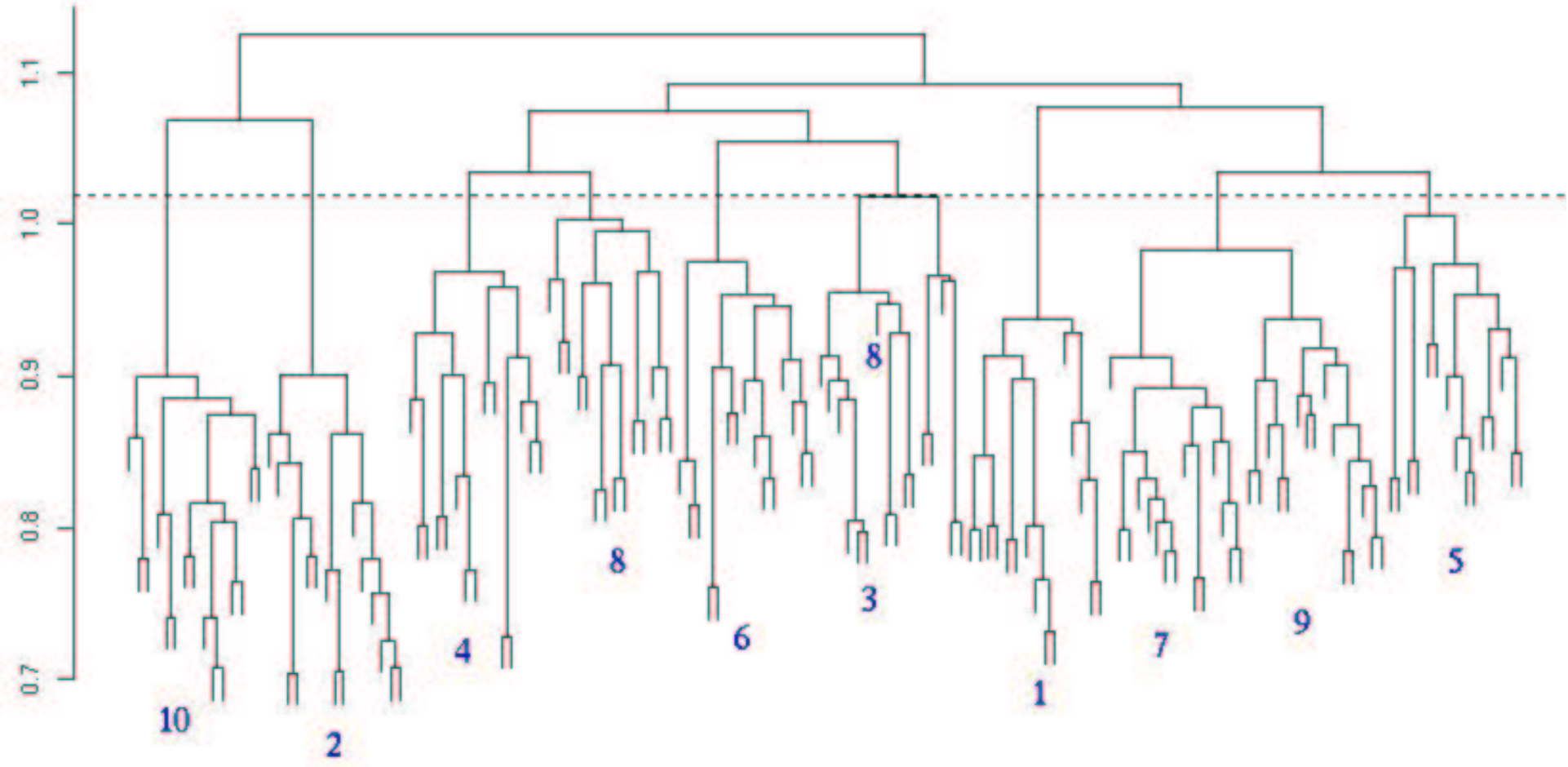
Dendrogram w/ prop'n mismatches



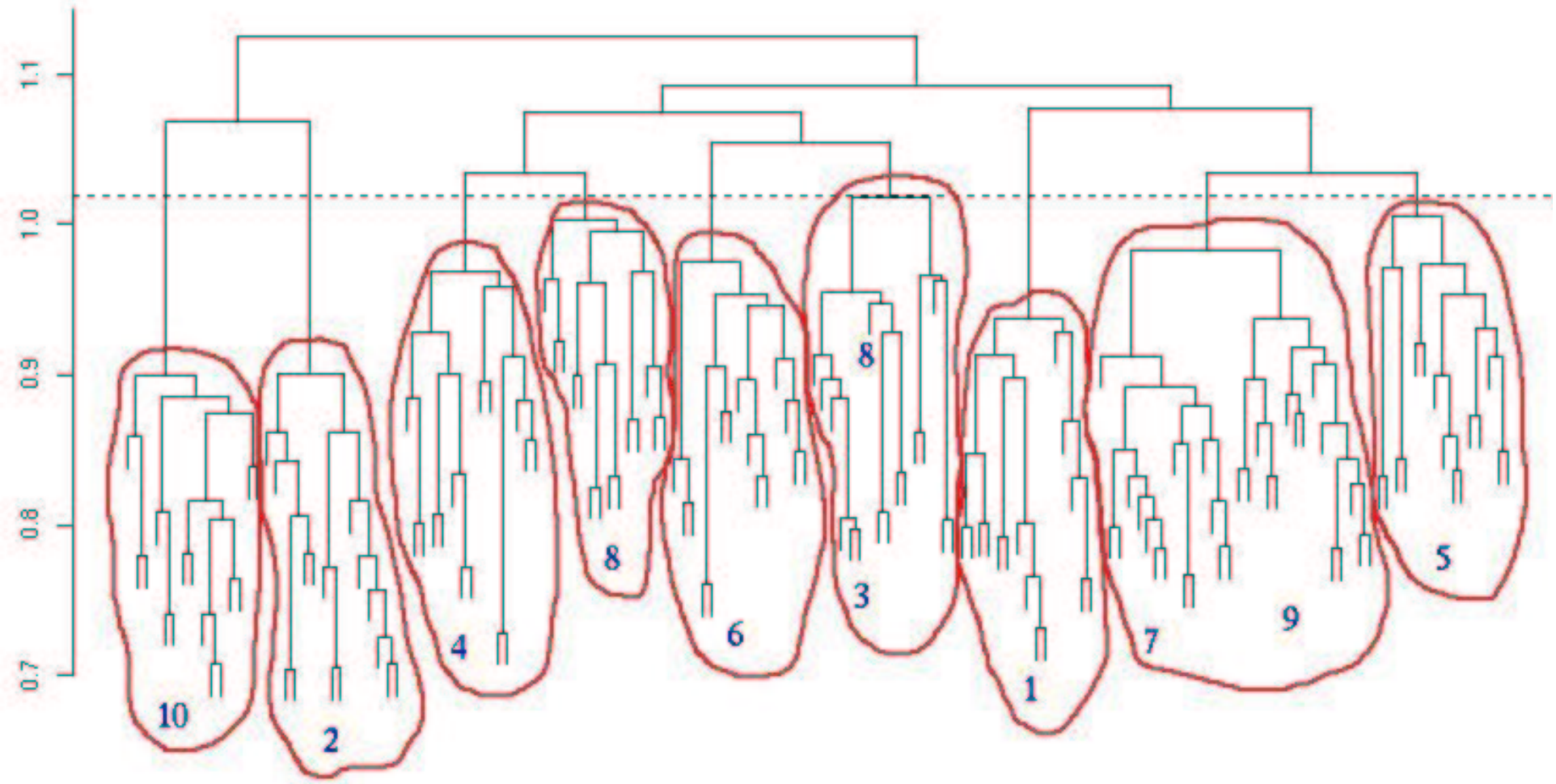
Dendrogram w/ prop'n mismatches



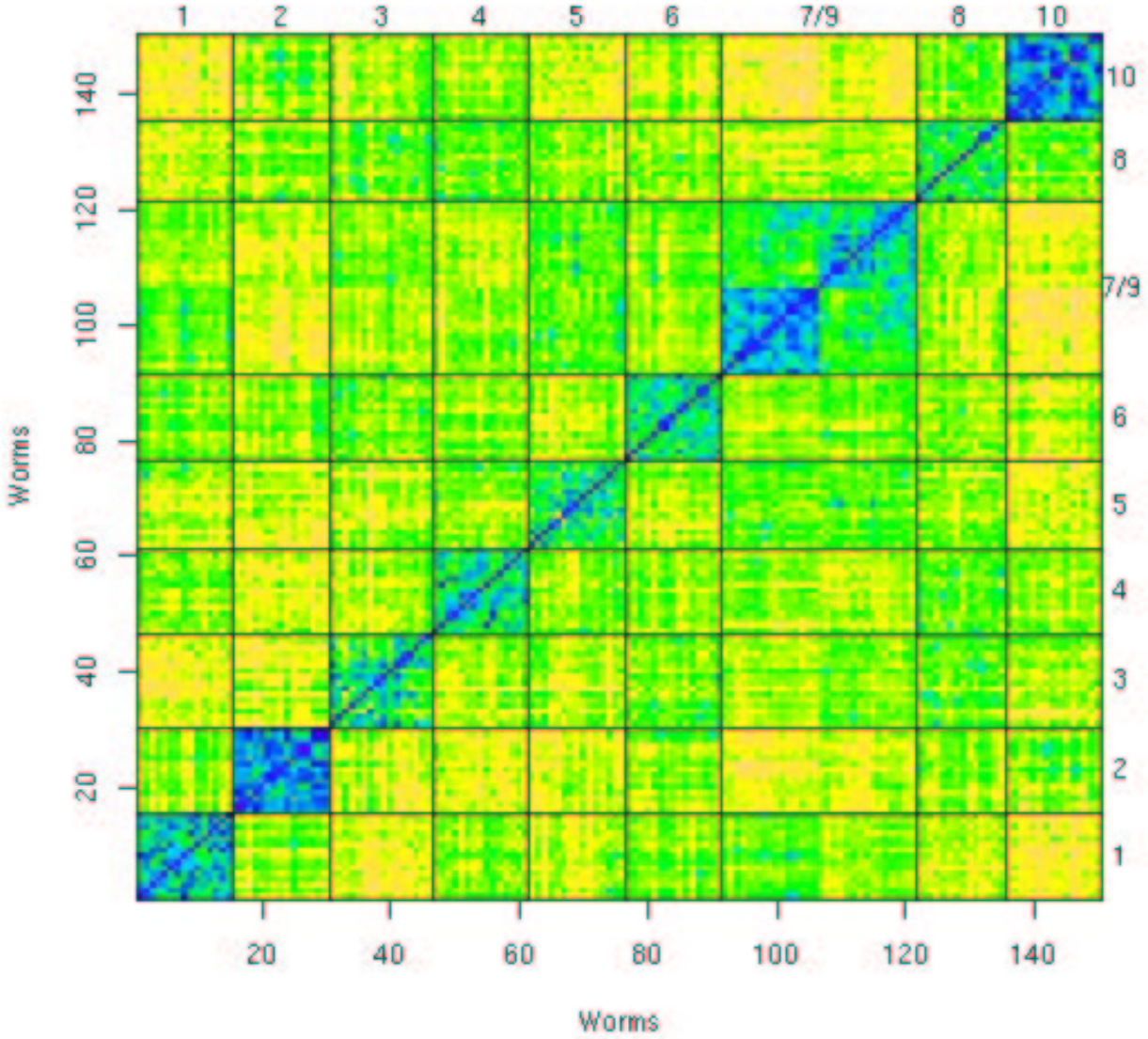
Dendrogram w/ LLR distance



Dendrogram w/ LLR distance



Last picture on the example



Simulations

Goal: Compare distances and cutoffs

Simulation: Varying numbers of families of equal size

120 individuals total

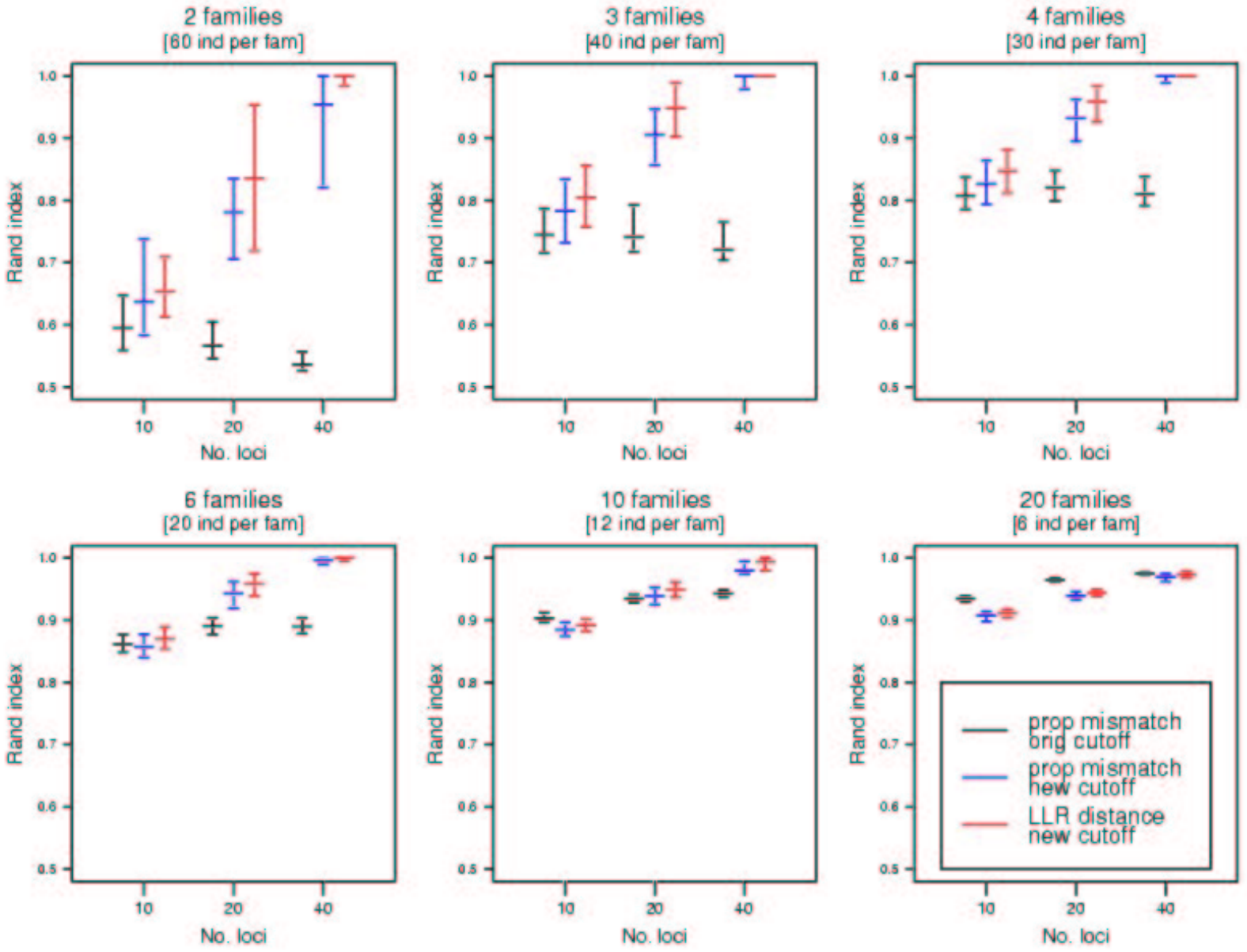
10, 20 or 40 RAPD markers

Allele frequencies fixed

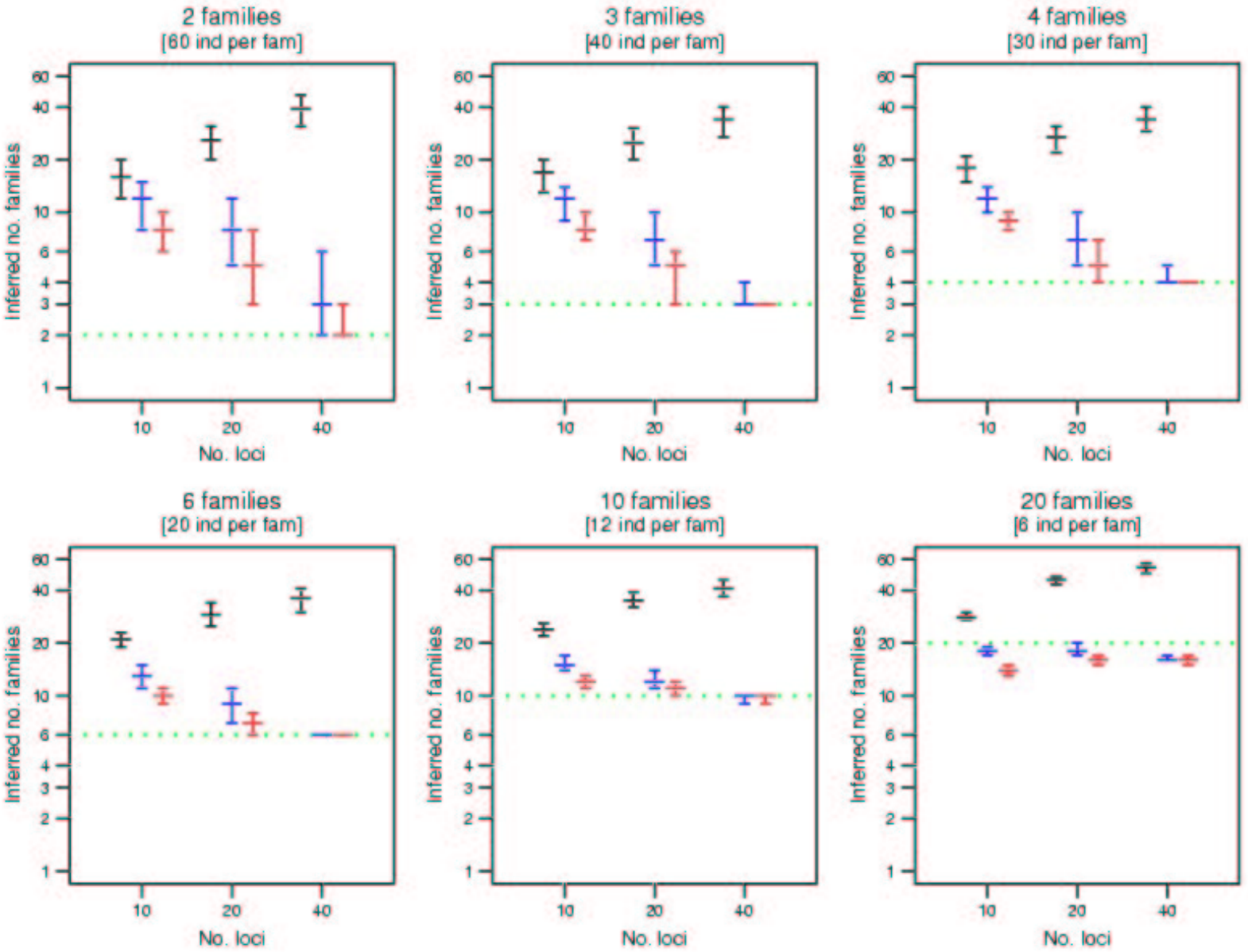
Methods: Distance = proportion of mismatches or LLR

Cutoff = mean distance among siblings or
20^{th} %ile of dist. between unrelateds

Results: Rand index



Results: Inferred no. families



Summary

- Interesting application of cluster analysis
- The Apostol et al. (1993) mosquito data is beautiful
- Simple changes can give great improvements
- Simulations + performance criteria are important
- Value of approach, for understanding Schistosome population genetics, is still uncertain
- We have lots of work to do