Populations in statistical genetics

What are they, and how can we infer them from whole genome data?

Daniel Lawson

Heilbronn Institute, University of Bristol
www.paintmychromosomes.com

Work with:

Garrett Hellenthal
UCL

Simon Myers
Oxford

Daniel Falush
Max Planck, Leipzig

January 2014, Cambridge

Overview

- Some views of 'population'
- A statistical definition
- Generative approaches
- Inference from sparse weakly linked data (STRUCTURE)
- Inference from dense linked data (fineSTRUCTURE)
- Selected results

What is a population in general?

Like species, populations are elusive objects.

- Definitions are a function of knowledge and application
- Some definitions - Need not be biologically meaningful
- Example: Individuals found on same island
- Example: Sample location due to clustered sampling procedure
- Example: Disease status for case/control studies
- In statistics: we generalize from samples to a population
- Share: individuals are equivalent under some measure

But do populations really exist? Does it matter if they are useful?

Diversion: some fuzzy definitions of species

Are definitions of species analogous to definitions of populations?

- **Typological species:** Organisms that share the same set of phenotypes
- **Ecological species:** Organisms that compete for the same environment
- **Phylogenetic species:** Organisms with a single common ancestor
- **Biological species:** Organisms that can share DNA via sexual reproduction

(Western/Eastern) Meadow Lark: http://evolution.berkeley.edu

What is a population? Motivation from genetics

We are interested in non-subjective definitions of populations.

- Individuals have equivalent genetic ancestry
- This depends on the data available ...
- ... and the model used
- Knowledge driven definition
- Requires defining equivalent!

The birds and the bees (vs the plants and bacteria)

- Sex dramatically affects genetic transmission
- Different population concept required
- **Sexual species**
  - Generalise the Biological Species concept
  - Random mating within a population
  - Try to detect deviations from random mating to identify populations
- **Asexual species**
  - Generalise the Ecological Species concept
  - Neutral competition within a population
  - Try to detect deviations from neutrality to identify populations
  - Not examined further in this talk

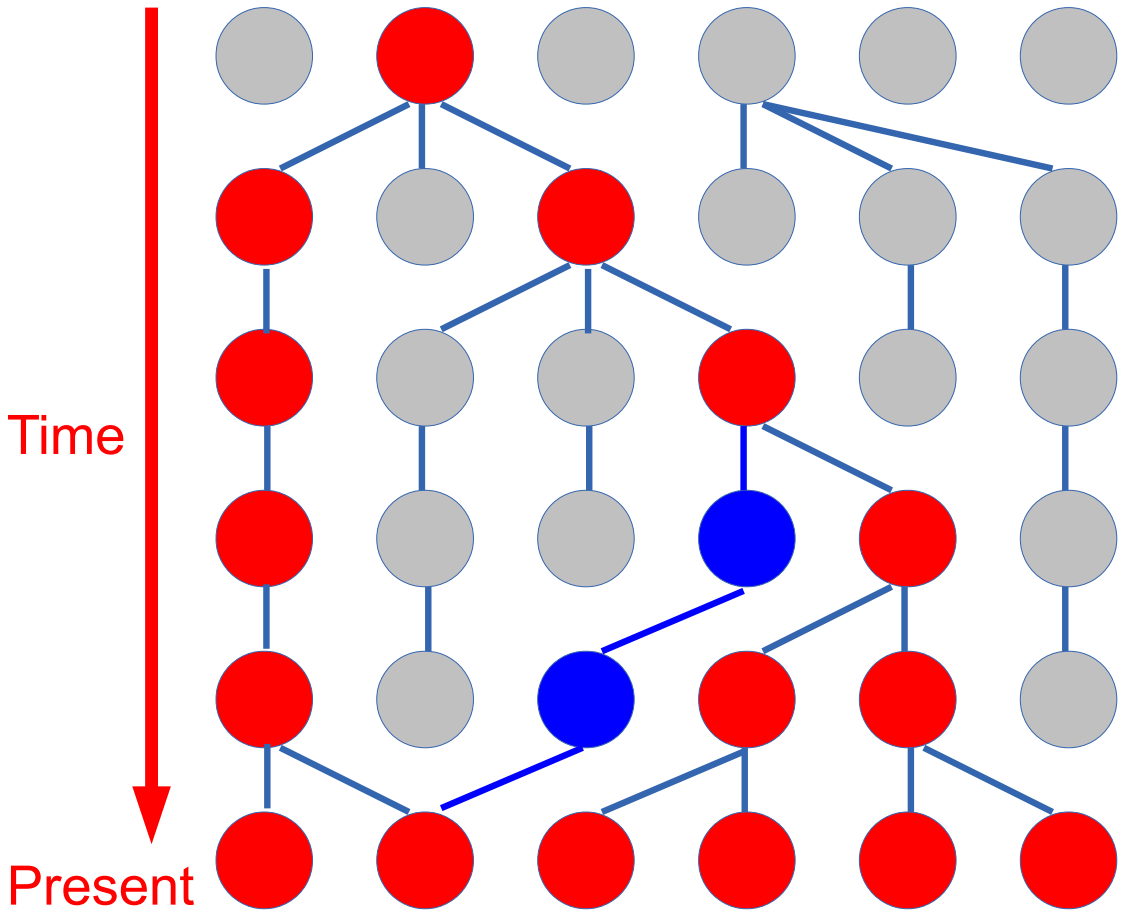
Top down approaches

Start from a theoretical generative model of reproduction.
For example:

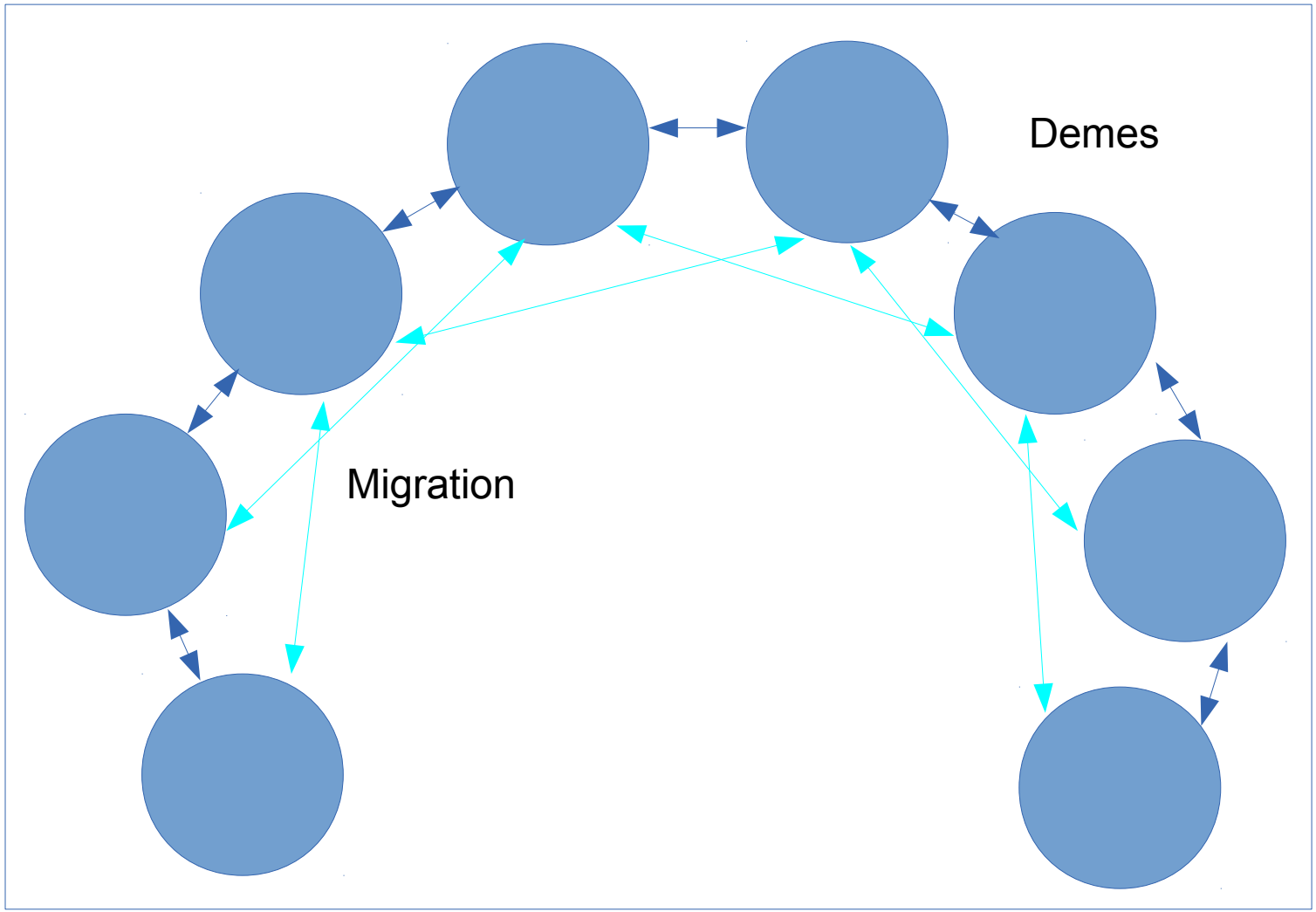
- Individuals move between populations via migration
- Discrete generations (for simplicity)
- Each generation randomly chooses parent(s) within a population

Population structure: individuals migrate between populations but mate randomly within each

Ancestry Process



Populations as Demes



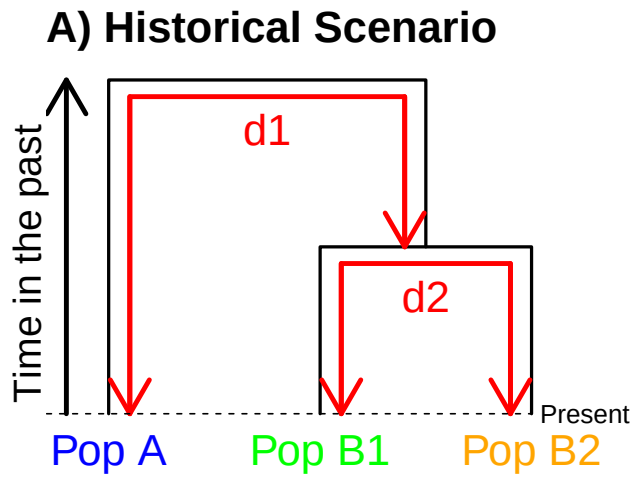
Ammerman and Cavalli-Sforza, 1984 The Neolithic Transition and the Genetics of Populations in Europe.

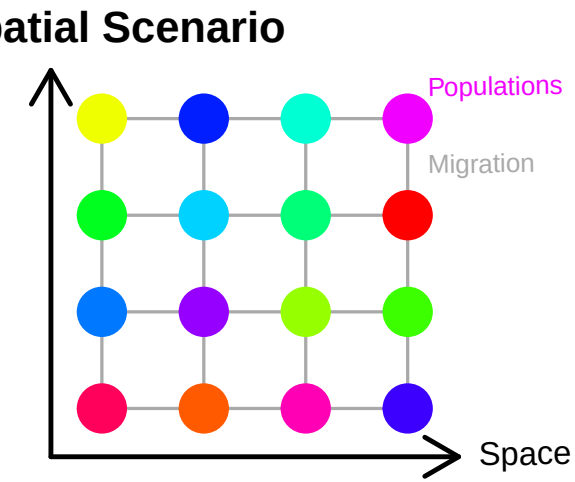
Diversion: Principal Components Analysis

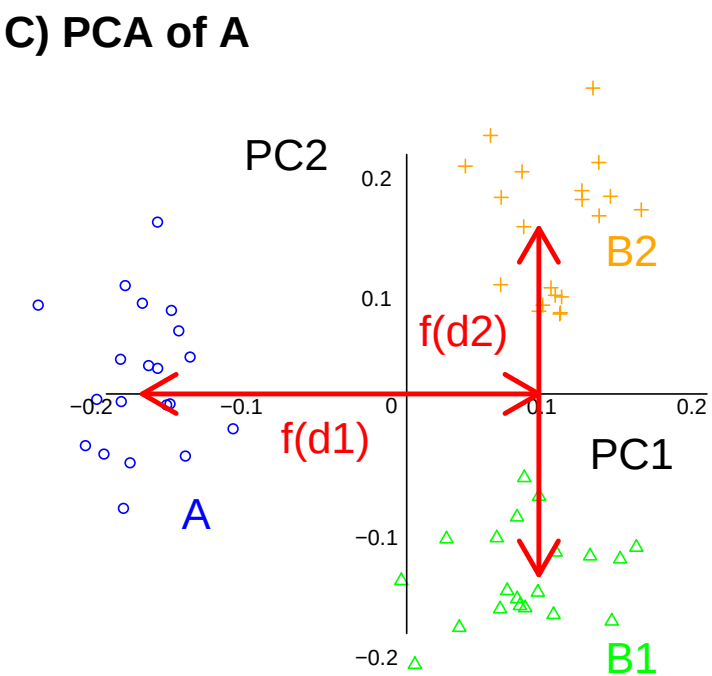
- PCA is widely used in genetics, and helpful for visualisation
- It does not make any explicit modelling assumptions...
- BUT to interpret the output, you do!
  - 'Just' rotating the data
  - Similar individuals tend to be close
  - Differences shared by many individuals tend to appear first
  - Each component describes a different direction of variation
  - If we are lucky, these correspond to real shared drift events
  - There is no unique interpretation of PCA (see McVean 2009)

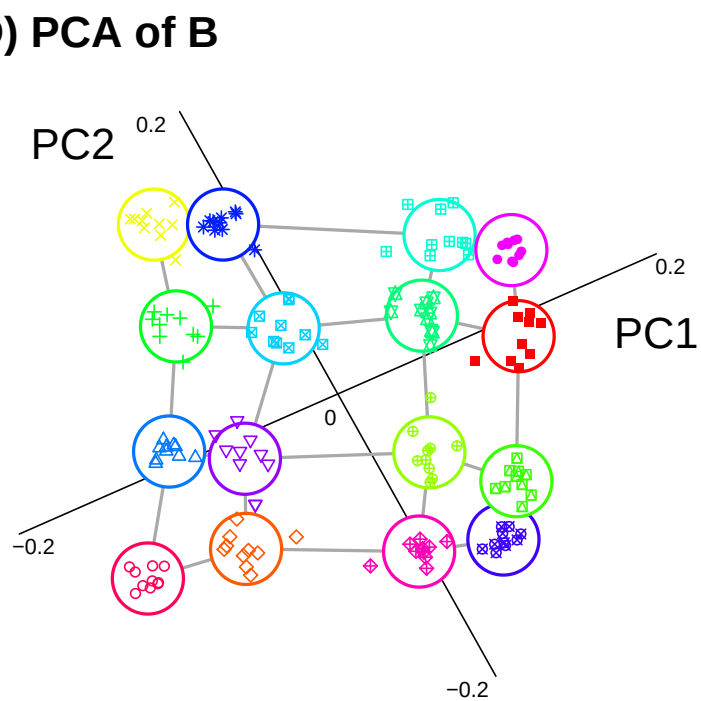
See Lawson & Falush 2012 for details.

Example


A) Historical Scenario
Time in the past
d1
d2
Present
Pop A
Pop B1
Pop B2


B) Spatial Scenario
Populations
Migration
Space


C) PCA of A
PC2
0.2
0.1
−0.1
−0.2
0
0.1
0.2
B2
f(d2)
f(d1)
PC1
A
B1

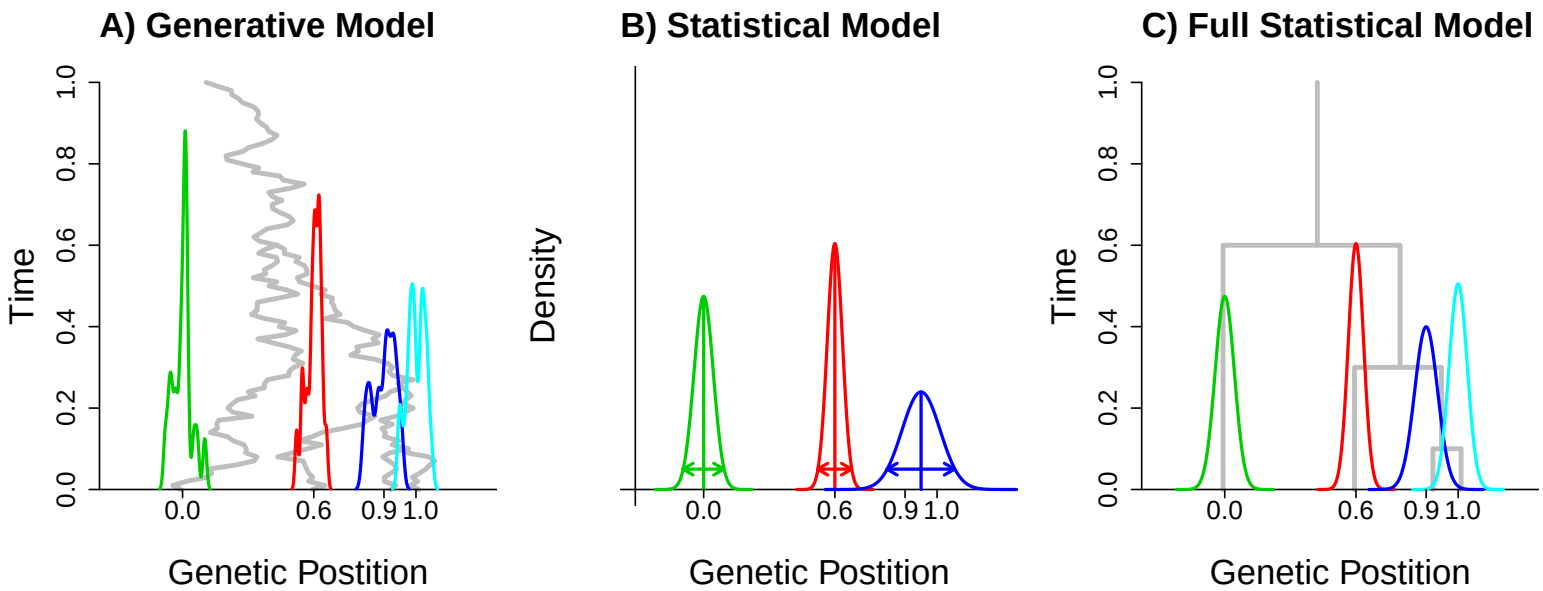

D) PCA of B
PC2
0.2
0.2
PC1
0
−0.2
−0.2

Bottom up approaches

Start from a definition of population as 'equivalent' individuals

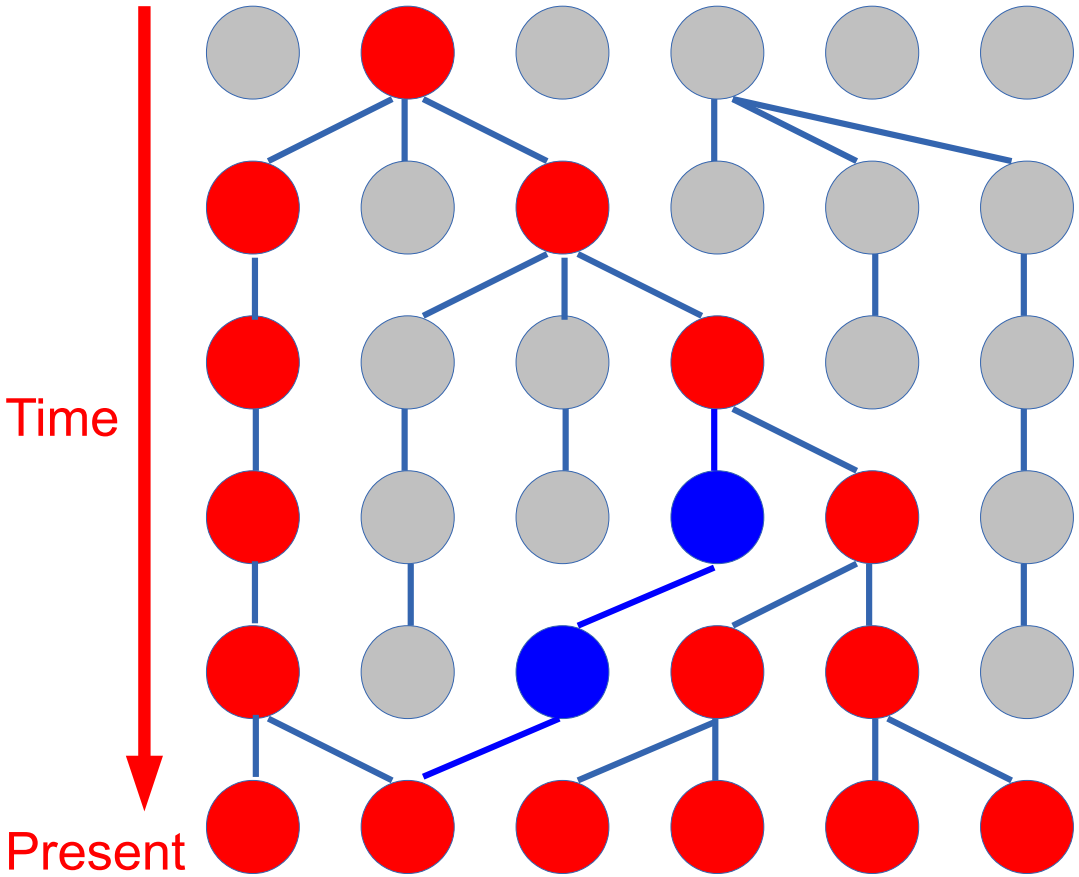
- Within a population, individuals are randomly mating
- Small samples of large populations: individuals are approximately independent
- Smaller populations: relationships must be accounted for
- What does random mating mean for the data?
- How are populations related?

Top down vs Bottom up



- A) Generative approaches are best in theory, if we can make the model match reality
- But, hard to use in practice - how to do inference?
- B) Bottom up approaches are approximate - might lose power
- C) But can be refined until they are close to the generating process

Ancestry Process - Ancestral Recombination Graph



- Take limit $N \to \infty$ with N/T constant
- Recombination rate ρ (for tract length)
- Ignore unbranching ancestors

Structure model

Pritchard, Stephens & Donnelly 2000.

- Populations are large and well mixed
- SNPs D_{il} are unlinked*
- loci have some ancestral frequency

$$p_{0l} \sim P(\cdot)$$

- Population k has frequency p_{kl} drifted from ancestral p_{0l}**

$$p_{kl} \sim \mathrm{Dirichlet}(p_{0l})$$

- Individual i is in population k if $Q_{ik} = 1$, assigned by

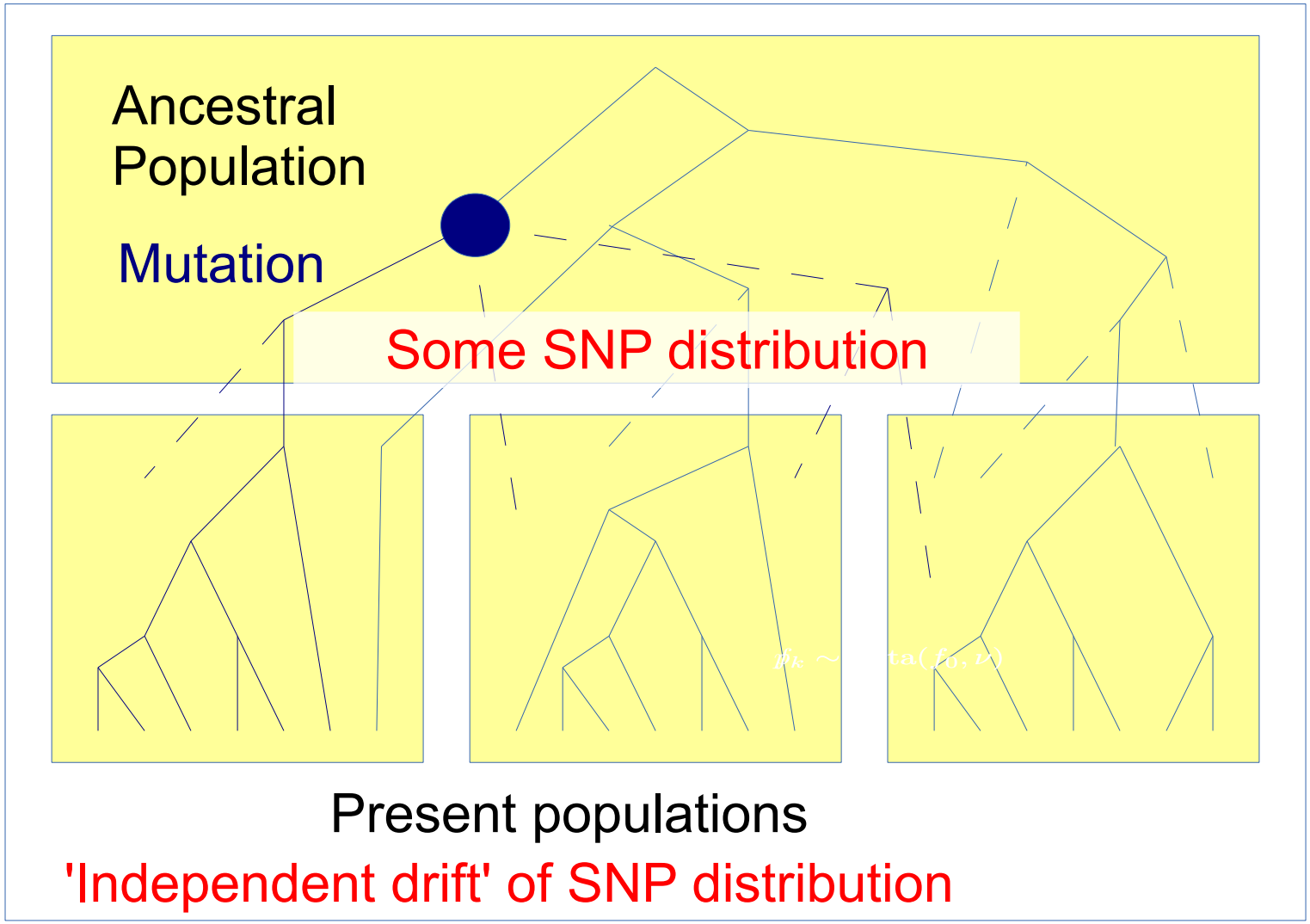
$$\prod_l P(D_{il}|p_{Q_{ik},l})$$

- Individuals in the same population are *exchangeable* with respect to the SNP frequencies

* Solutions to this have been explored (computationally inconvenient)

** Valid approximation, originally derived by Wright

Single SNP with populations



Scaling STRUCTURE

STRUCTURE approach has a parameter for every SNP. But:

- Assuming that drift is weak, $p_{0l} = E(D_{.l})$ and:

$$p(p_{kl}) \sim N\left(p_{0l}, p_{0l}(1 - p_{0l})\right)$$

- Probability SNP l is shared not by chance:

$$X_{ijl} = D_{il}D_{jl}/p_{0l} + (1 - D_{il})(1 - D_{jl})/(1 - p_{0l})$$

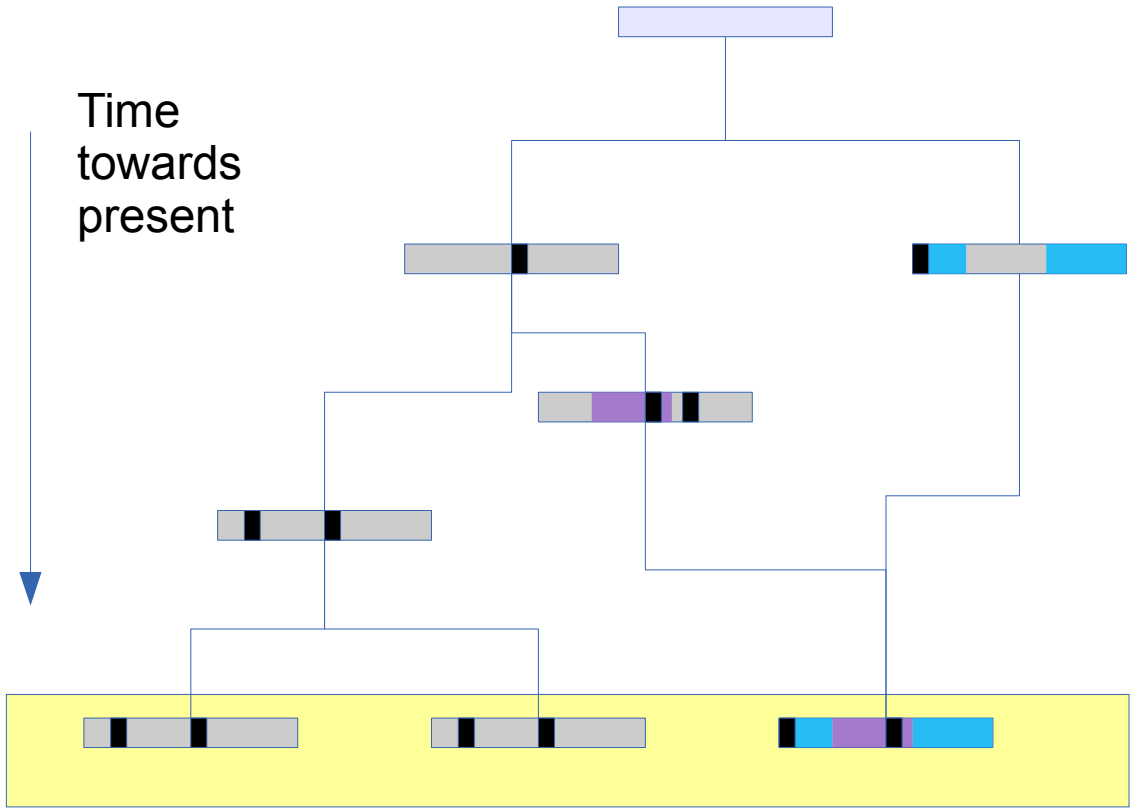
- Invoke the Central limit theorem:

$$X_{ij} = \sum_{l=1}^{L} X_{ijl} \sim N(\mu_{ij}, \sigma_{ij}^2)$$

- This is the Coancestry Matrix
- It is a sufficient statistic for $p(D|Q)$
- μ and σ known and same for all individuals in a population
- Exchangability again!

Now lets think about linkage...

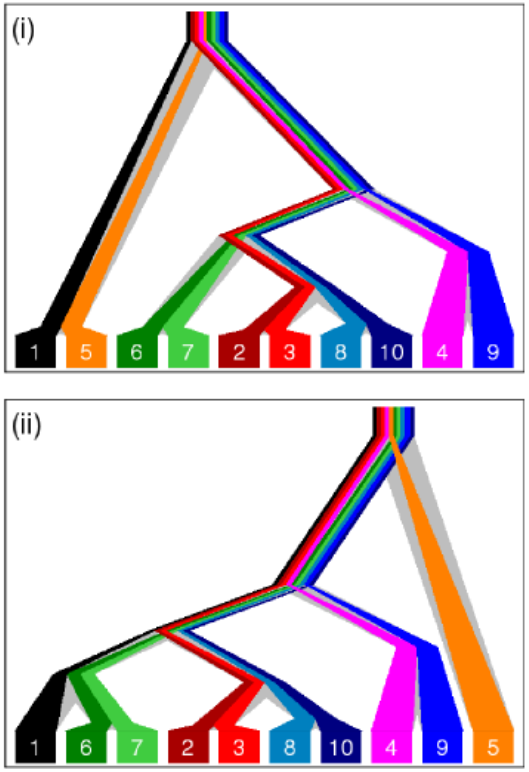
Ancestral Recombination Graph



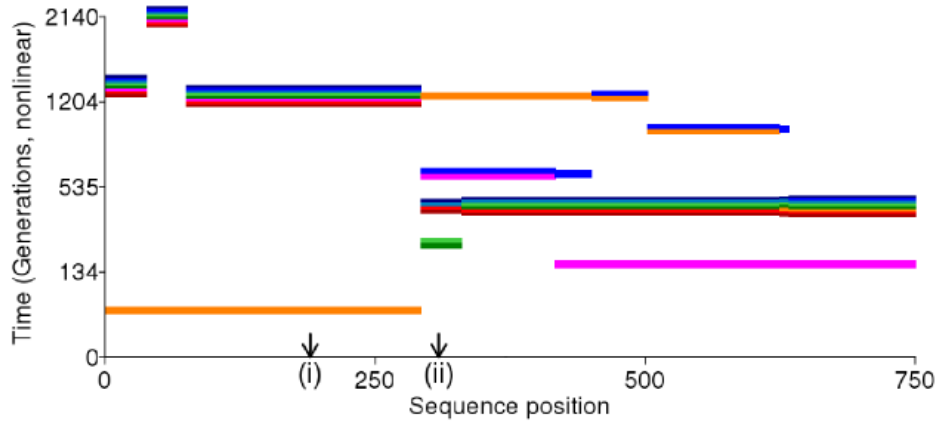
Hein, Schierup and Wiuf 'Gene Genealogies, Variation and Evolution', OUP 2005

ChromoPainter

Local genealogies



Time to MRCA with haplotype 1



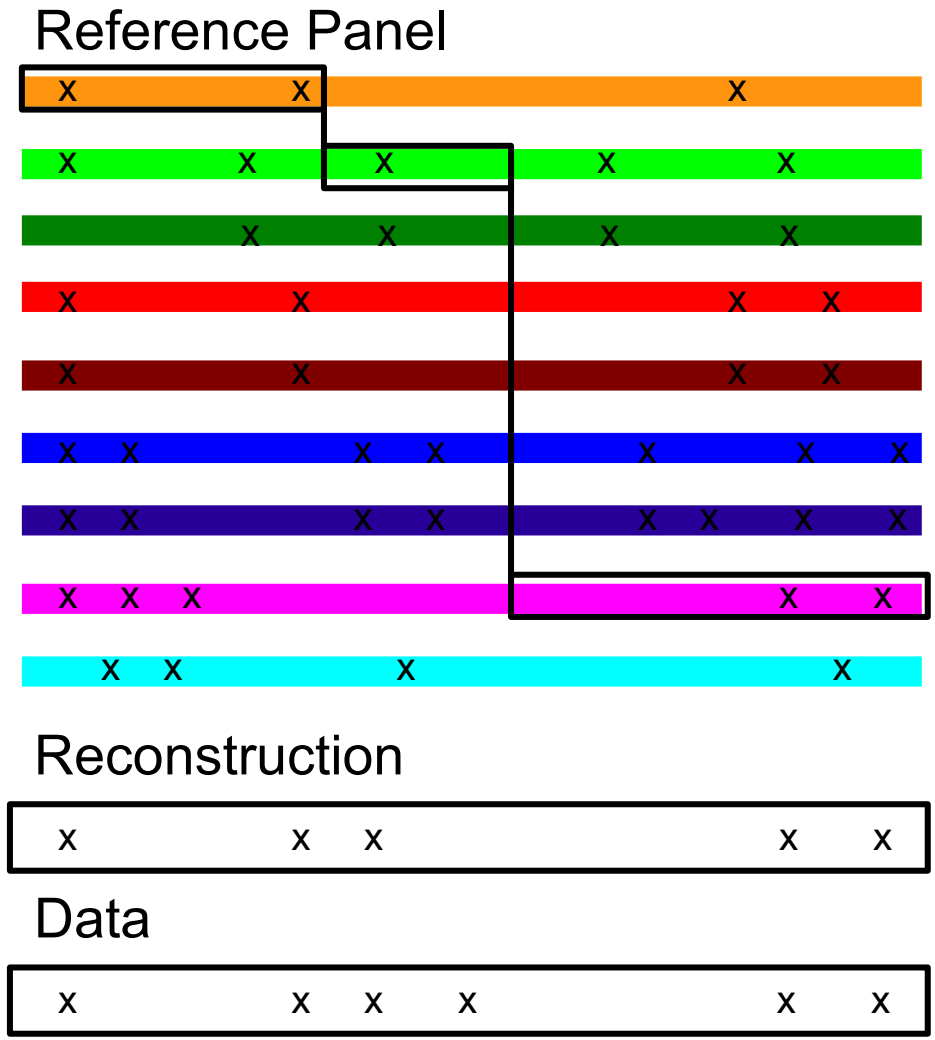
True 'nearest neighbour' distribution of haplotype 1

Mean painting of haplotype 1

Coancestry matrix row for haplotype 1

	Donor haplotype									
	1	2	3	4	5	6	7	8	9	10
Haplotype 1	0	0.08	0.09	1.1	1.24	0.52	0.52	0.06	0.01	0.06

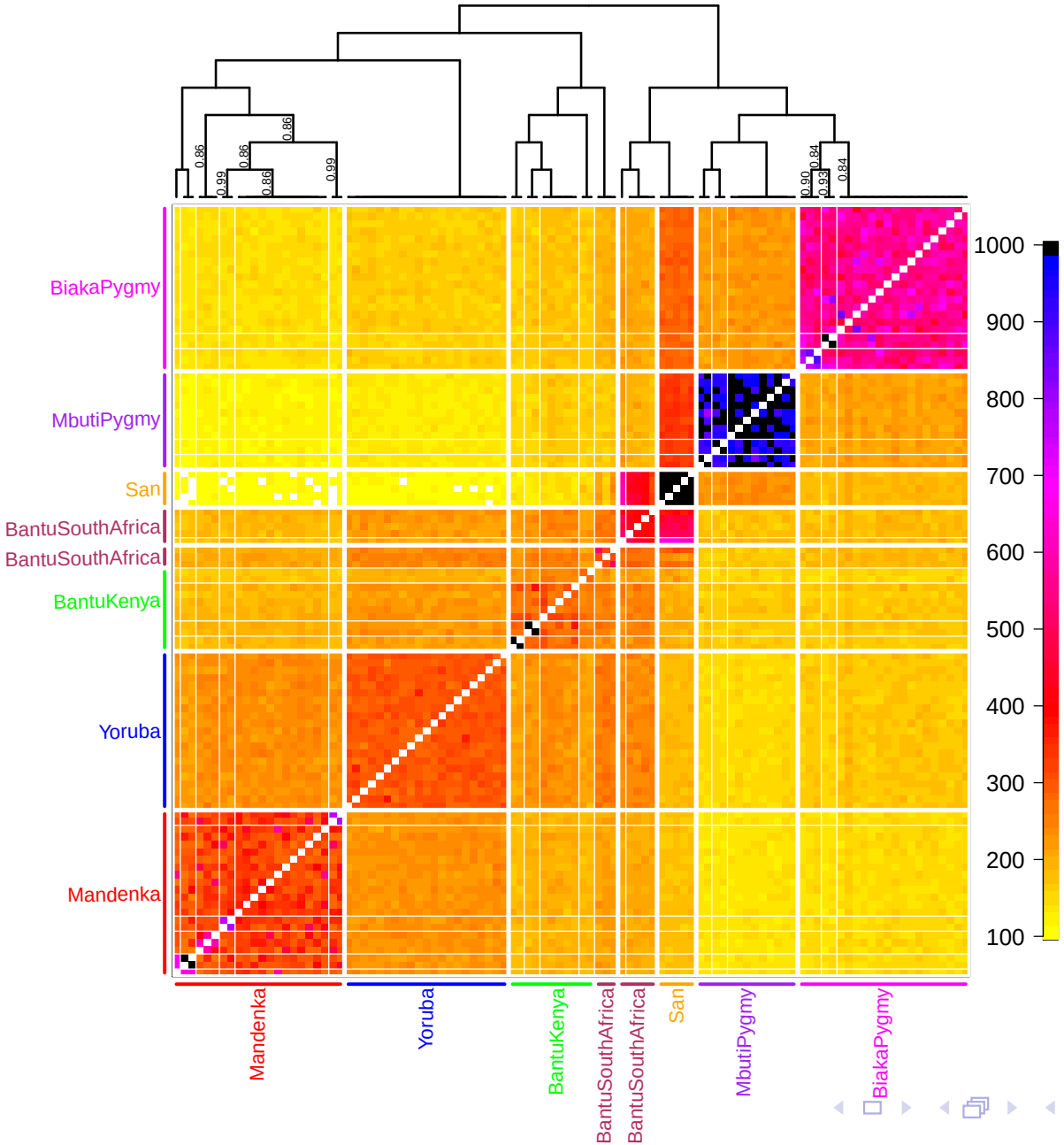
ChromoPainter Hidden Markov Model



FineSTRUCTURE model

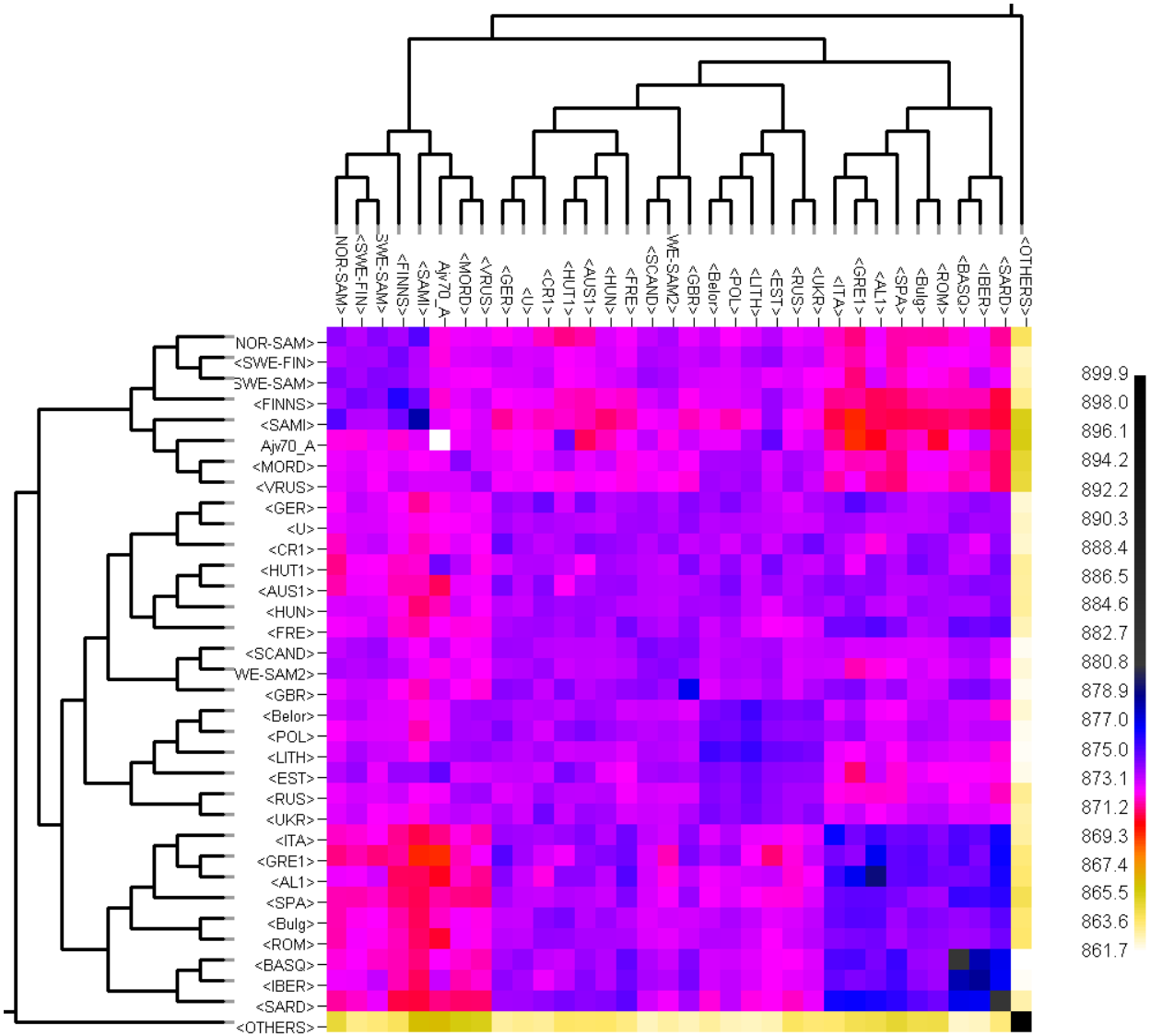
- Each population has a characteristic rate P_{ab} of sharing 'chunks' with each other population
- Individuals are again exchangeable within populations
- Each recombination event has probability $\hat{P}_{ab} = P_{ab}/\hat{n}_b$ when coming from an individual in population b into population a
- Dirichlet Process prior on the parameters P_a.
- We integrate out P, leaving no population level parameters
- We can put a meaningful prior on the variation of P between populations, and the number of populations
- In practice these details don't matter much

Example: Africa HGDP



Ancient Genomes: fennoscandia.blogspot.co.uk

Ajv70 Gotland hunter gatherer

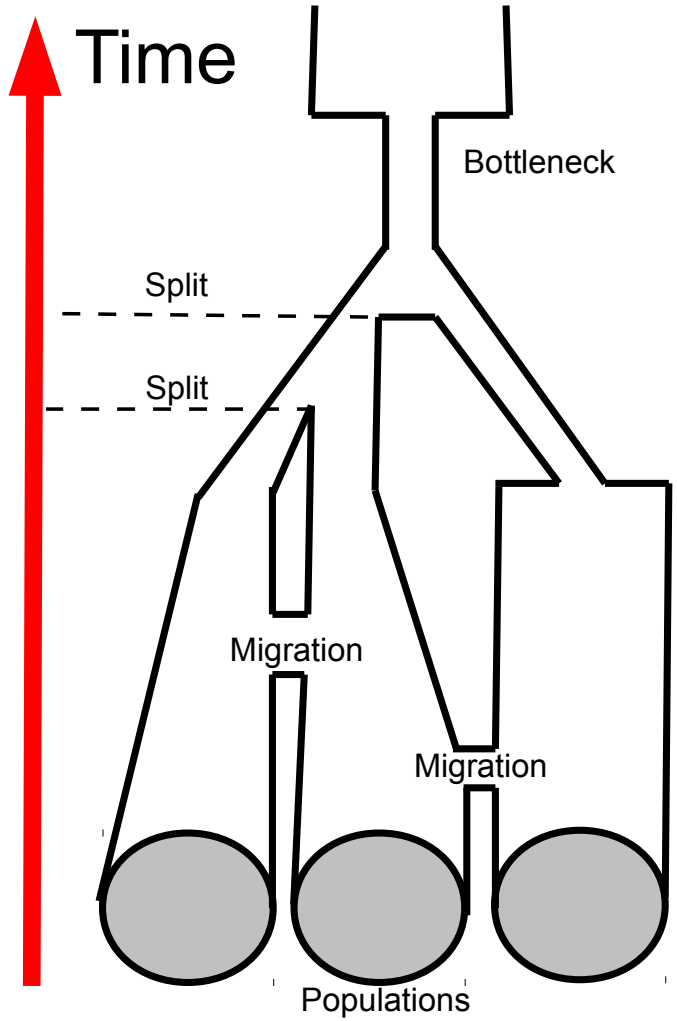


fennoscandia.blogspot.co.uk/2013/11/ajv70-and-modern-european-variation-ii.html

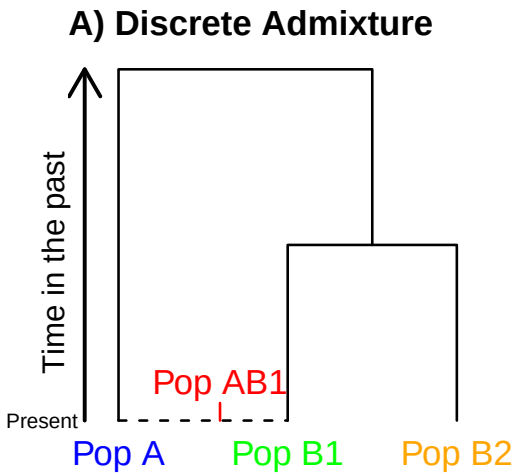
444k SNPs

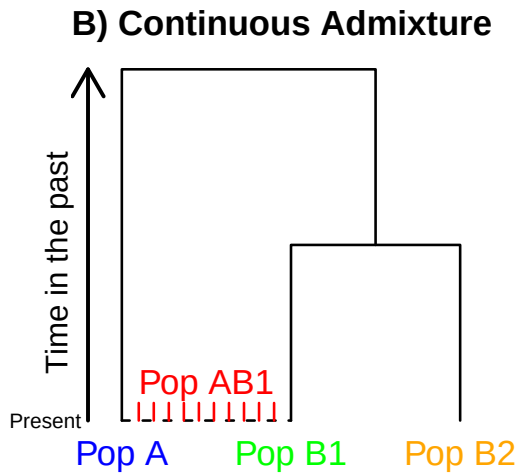
History of populations

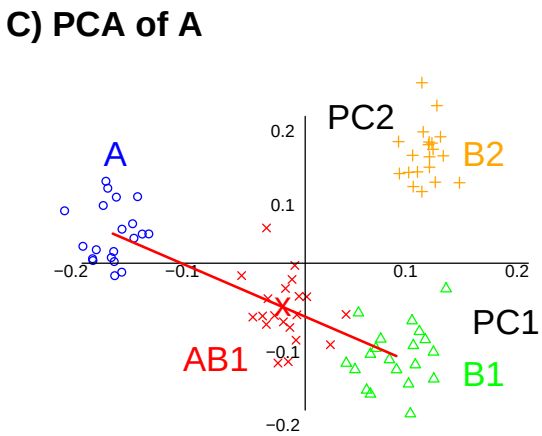
- We have assumed that we observe individuals from real populations
- Populations differ by genetic drift
- This works, even if there is historical migration, provided that the mixture fractions are equal (exchangeability)
- Real individuals are related by a combination of drift and admixture
- 'Ancestral population graph'

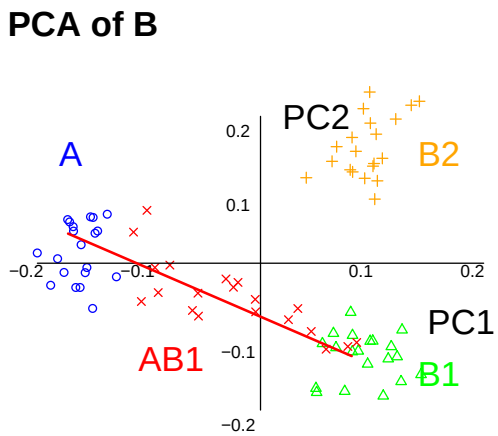


Admixture



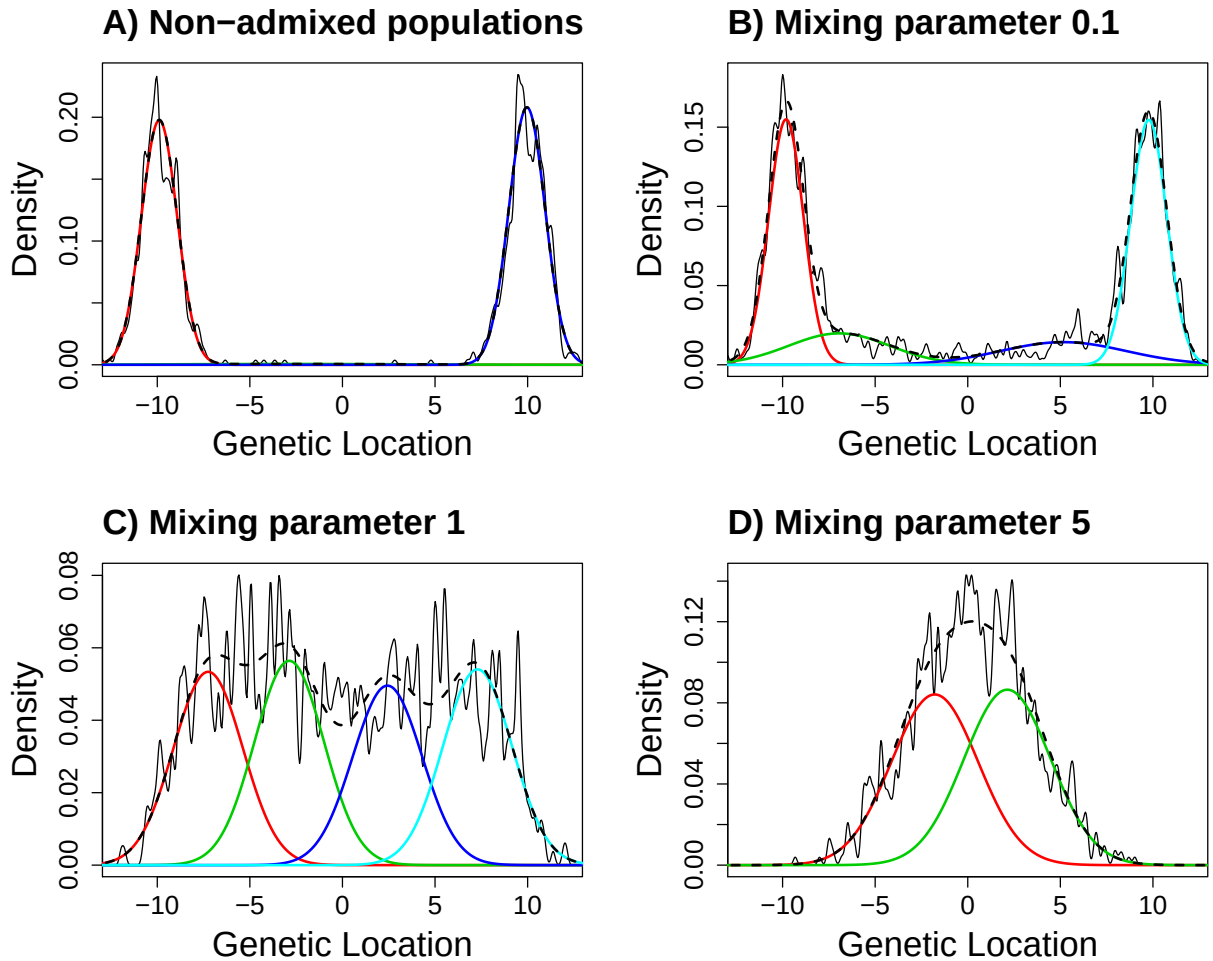






Admixture describes mixture without drift.

Admixture



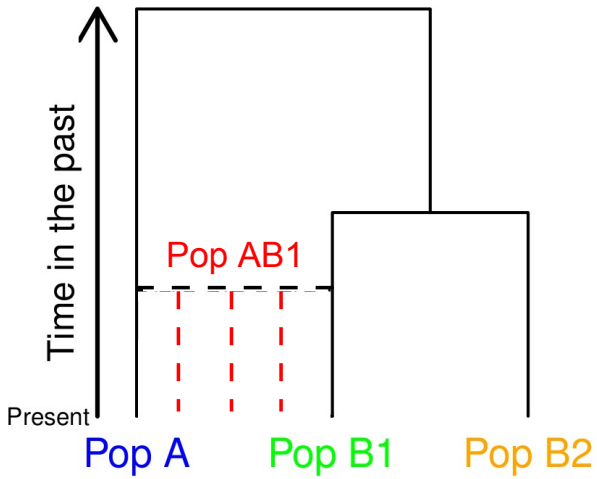
Continuous admixture is a problem.

Admixture models

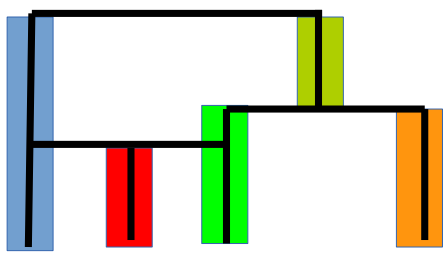
- STRUCTURE can infer 'pure' populations from admixed populations
- Population assignment Q_{ik} sums to 1 without requiring a single element
- Interpretation: Observed individuals are mixtures of pure populations, without drift
- How can we tell apart:
  - Large drift, mixed by admixture?
  - small drift without admixture?
- Solution: SNPs have fixed in some populations $p_{ik} = 0$ (or 1)
- FineSTRUCTURE cannot use this description, as we've integrated out these details

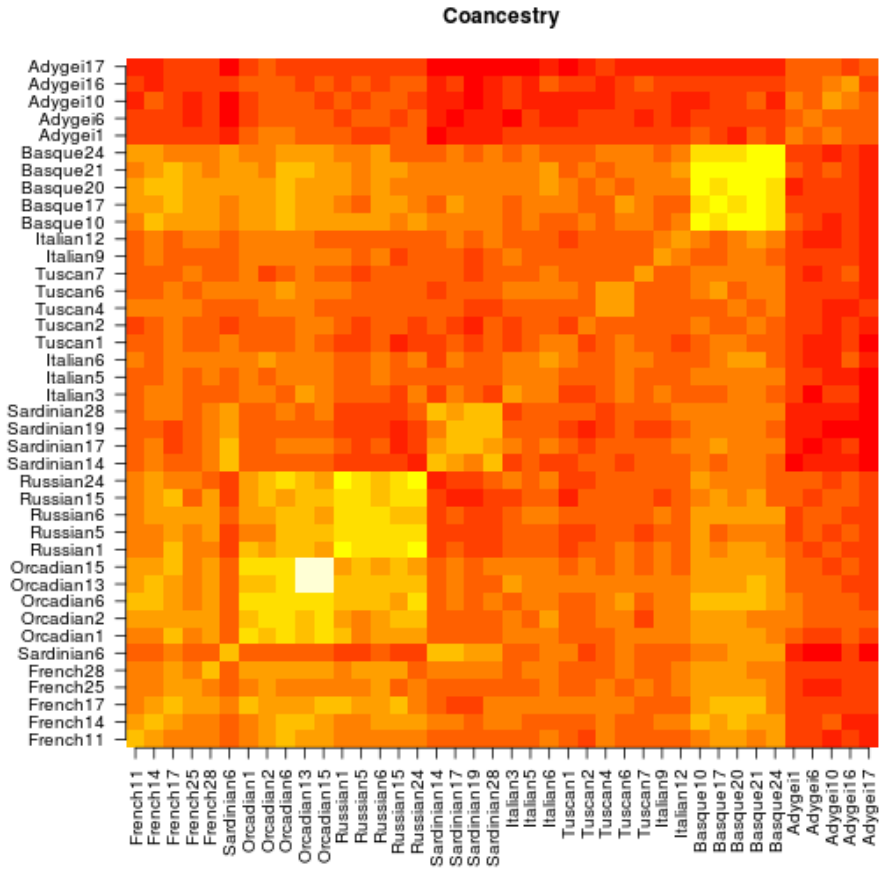
Drift model

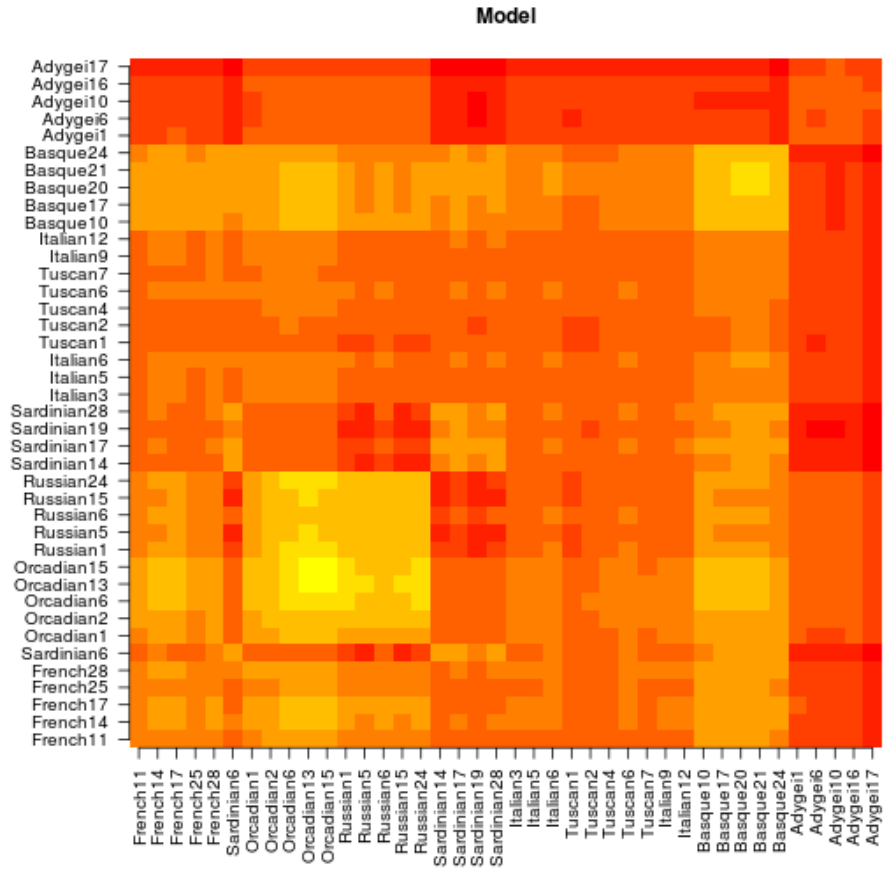
- See each drift event as independent
- Population assignment Q_{ik} now takes any value
- 'Amount' of each drift event retained by individual i
- Reconstructs the coancestry matrix
- Requires a strong prior to obtain a unique solution

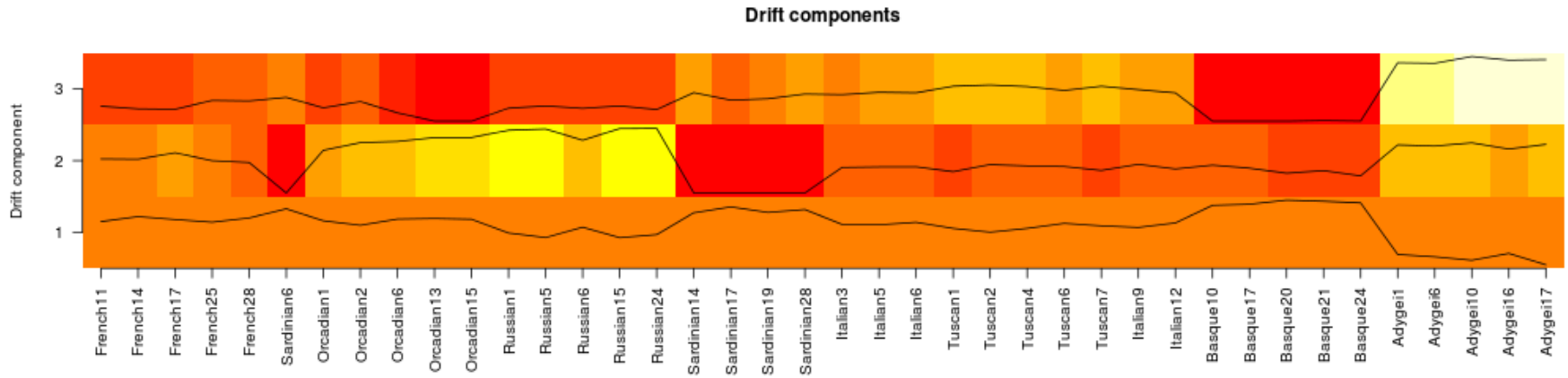


Drift Components




Coancestry
Adygei17
Adygei16
Adygei10
Adygei6
Adygei1
Basque24
Basque21
Basque20
Basque17
Basque10
Italian12
Italian9
Tuscan7
Tuscan6
Tuscan4
Tuscan2
Tuscan1
Italian6
Italian5
Italian3
Sardinian28
Sardinian19
Sardinian17
Sardinian14
Russian24
Russian15
Russian6
Russian5
Russian1
Orcadian15
Orcadian13
Orcadian6
Orcadian2
Orcadian1
Sardinian6
French28
French25
French17
French14
French11


Model
Adygei17
Adygei16
Adygei10
Adygei6
Adygei1
Basque24
Basque21
Basque20
Basque17
Basque10
Italian12
Italian9
Tuscan7
Tuscan6
Tuscan4
Tuscan2
Tuscan1
Italian6
Italian5
Italian3
Sardinian28
Sardinian19
Sardinian17
Sardinian14
Russian24
Russian15
Russian6
Russian5
Russian1
Orcadian15
Orcadian13
Orcadian6
Orcadian2
Orcadian1
Sardinian6
French28
French25
French17
French14
French11


Drift components
Drift component
1
2
3

Projects

- Siberian cold adaptation (Alexia & Toomas, in review)
- GlobeTrotter (Hellenthal et al: very accurate admixture dating using ChromoPainter, to appear in Science)
- Peopling of the British Isles (in review)
- UK10K, ALSPAC (4K whole genomes, use in genome-wide association studies)
- Highly recombining asexuals (fungus, bacteria)
- Model improvements:
  - Relatedness
  - Admixture/history of populations
  - Complete recoding for usability
  - Efficient computation (fastFineSTRUCTURE)

See www.paintmychromosomes.com

- Lawson, Hellenthal, Myers & Falush, 'Inference of population structure using dense haplotype data', 2012. PLoS Genetics.
- Lawson & Falush 'Similarity matrices and clustering algorithms for population identifcation using genetic data', 2012. ARHG.
- Lawson 2014 'Populations in statistical genetics modelling and inference', in 'Populations in the Human Sciences', Eds. Kreager, Capelli, Ulijaszek & Winney.

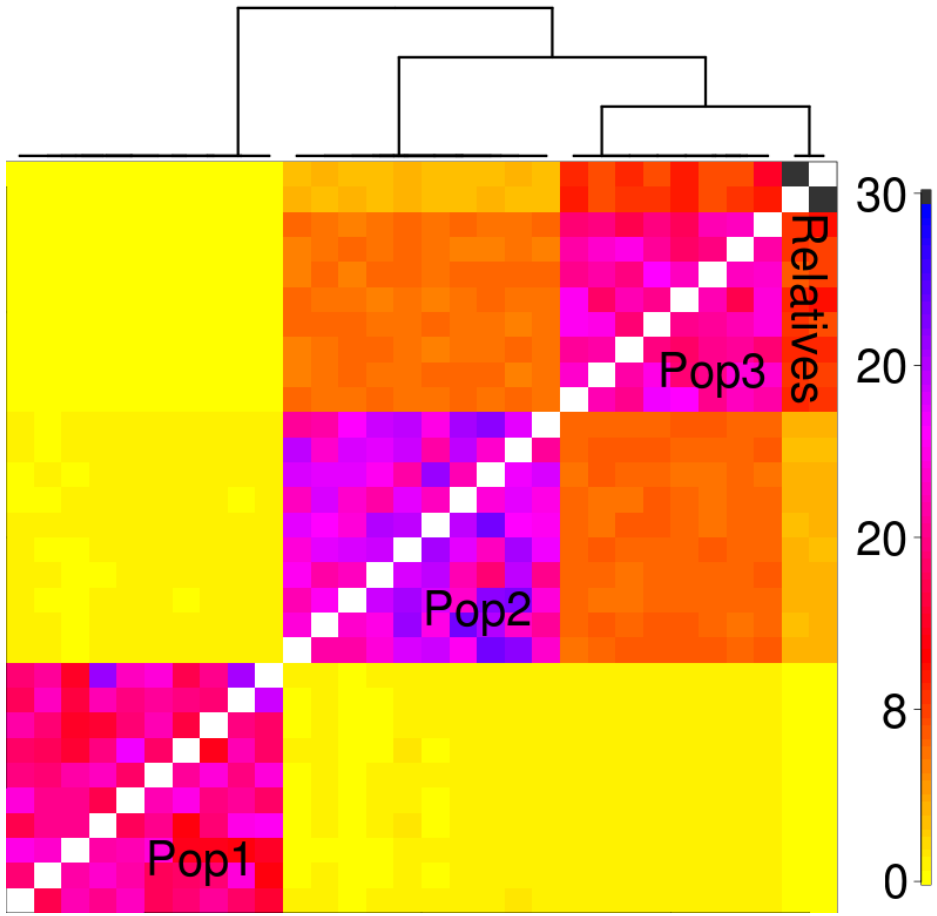
Garrett Hellenthal
UCL

Simon Myers
Oxford

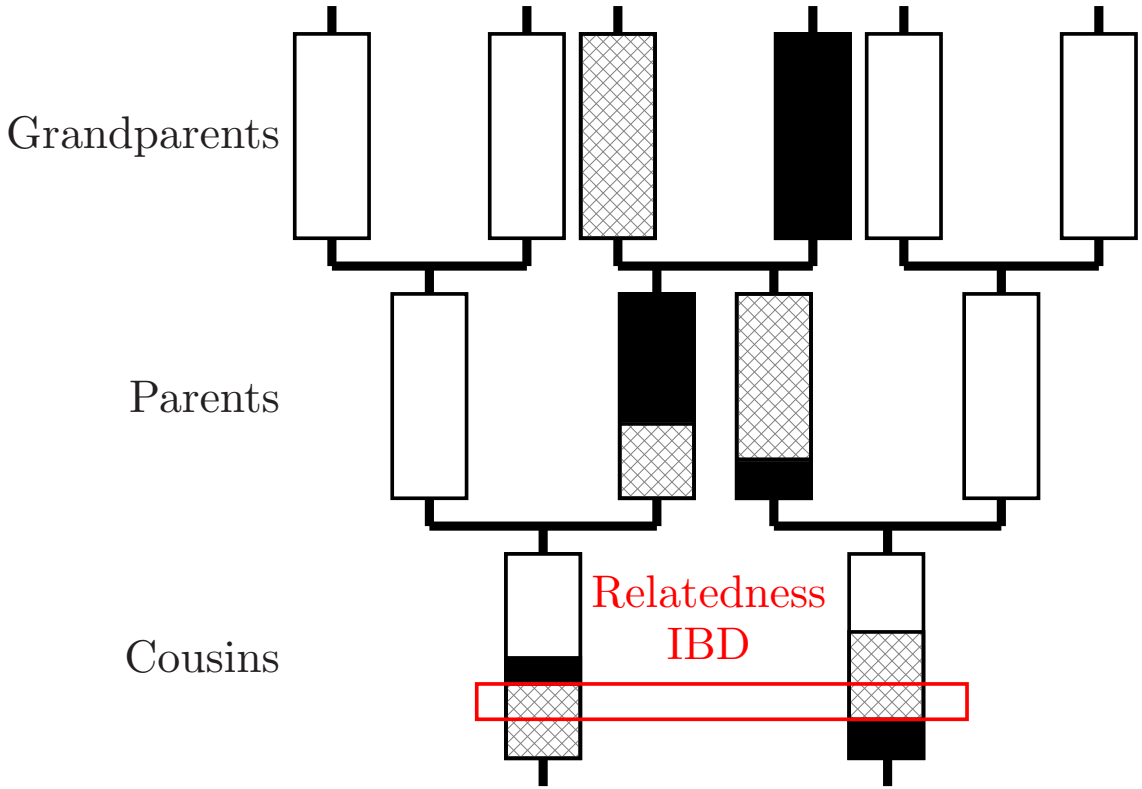
Daniel Falush
Max Planck, Leipzig

Relatedness

What if individuals within a population are related?

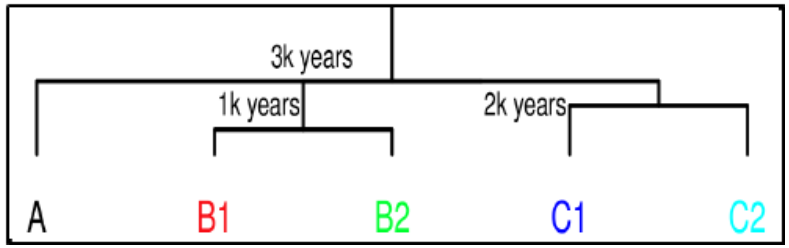


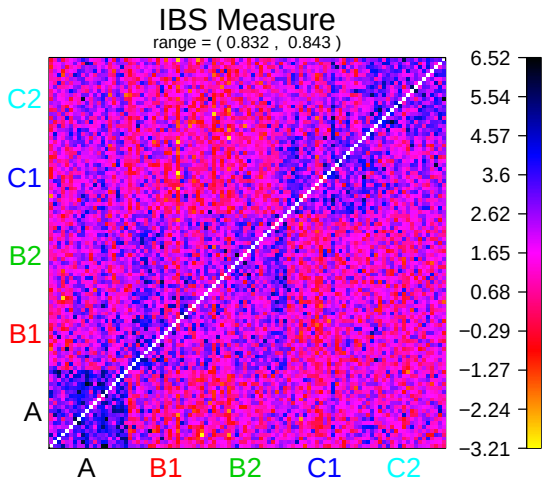
Relatedness

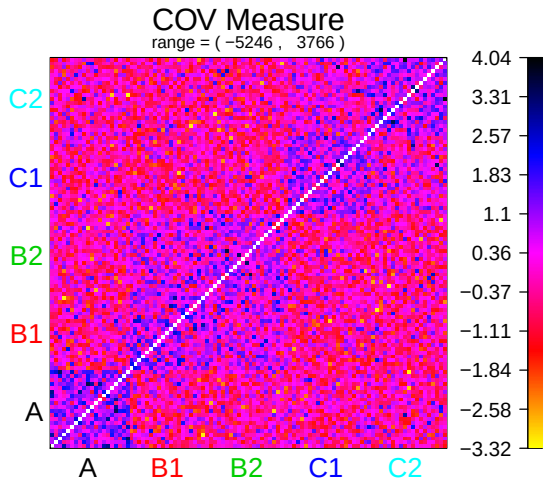


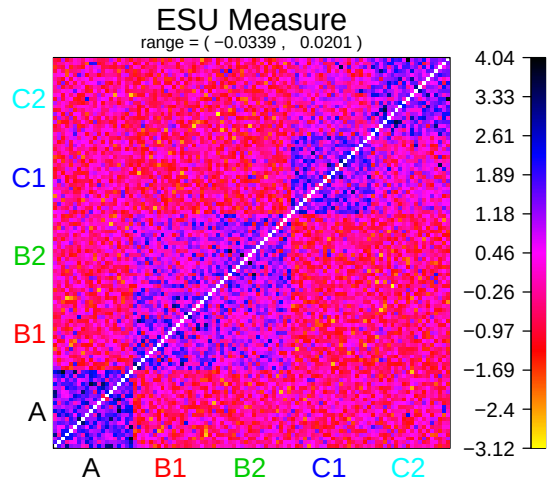
- Samples are cousins
- This is very easy to tell from their tract length distribution
- Excluding these tracts, we sample from the population distribution of chunks
- Multiple ordering model in progress

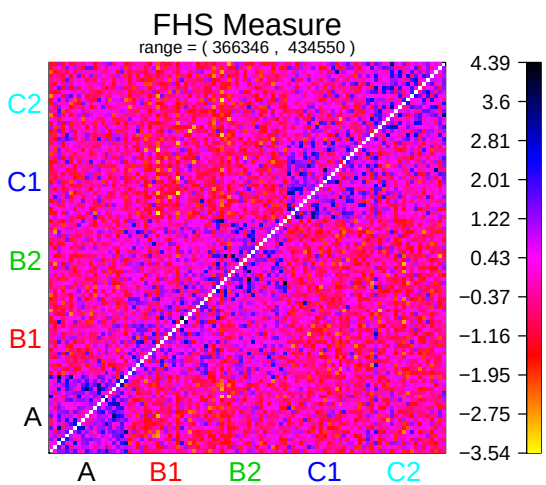
Choice of measure

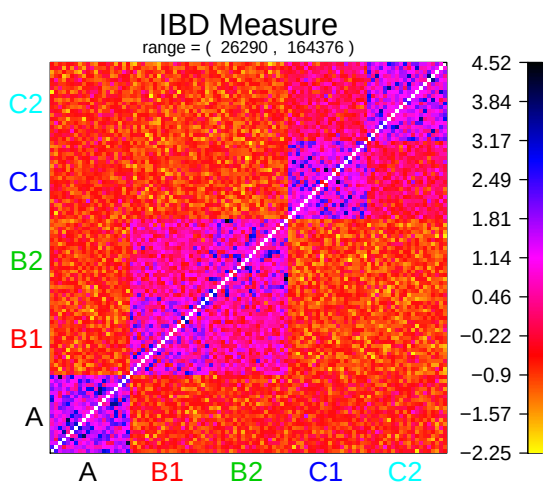


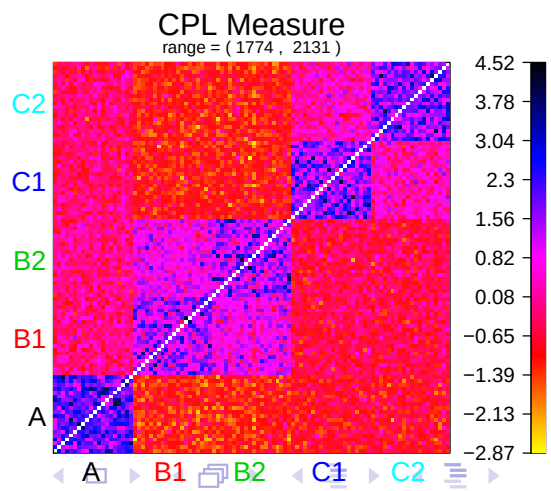












Choice of measure