Morning

Inferring population structure and admixture histories

Afternoon

Building coalescent trees: genetic ancestry, effective population sizes, mutation rates, selection

Leo Speidel

UCL Genetics Institute and the Francis Crick Institute

leo.speidel@outlook.com

Today's schedule:

09:30-10:15: Lecture

10:15-10:45: Coffee

10:45-12:30: ADMIXTURE, CHROMOPAINTER and GLOBETROTTER practical

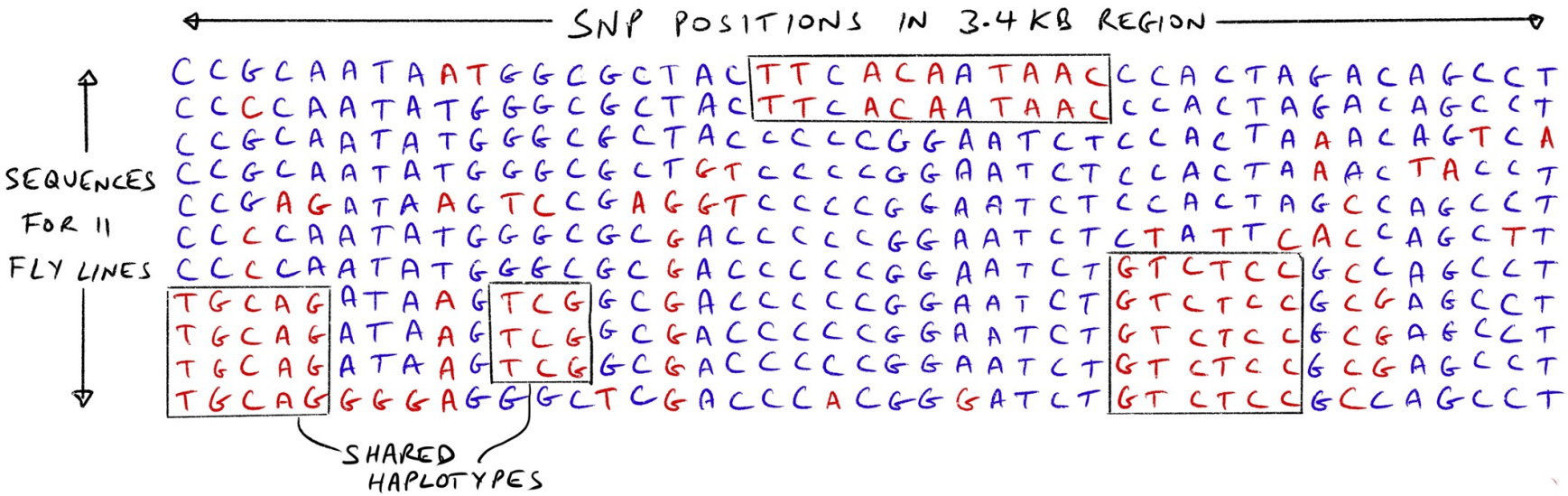
Roman Villa!

15:30-16:15: Lecture

16:15-16:30: Coffee

16:45-18:30: Relate practical

Spotting patterns in genetic variation



An Owner's Guide to the Human Genome, Jonathan Pritchard

What does genetic diversity within & across (human) groups look like?

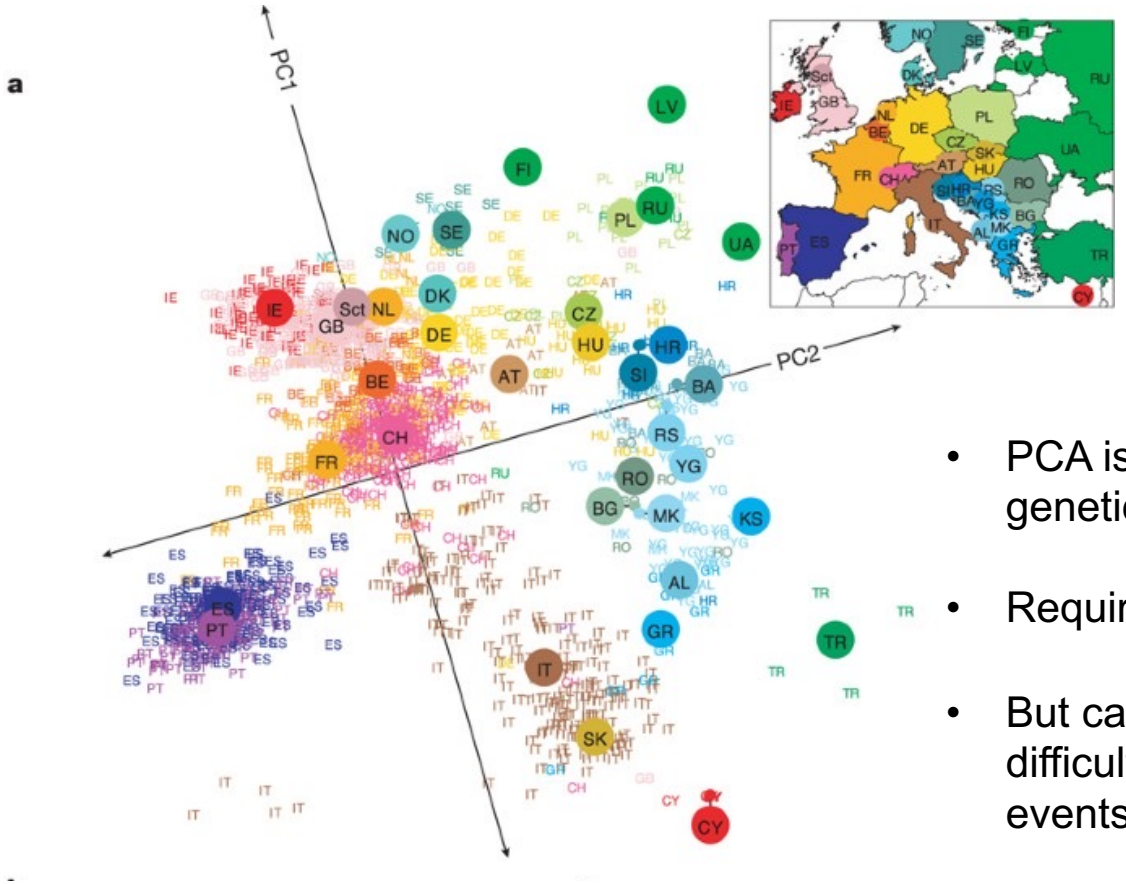
- Genome-wide patterns of variation across individuals provide a powerful source of data

- What evolutionary forces are important in shaping our genomes?
    - Genetic Drift
    - Mutation
    - Recombination
    - Admixture
    - Selection
    - etc

- There are many different types of data, from **allele frequencies** at single polymorphic sites (or multiple unlinked sites) to **haplotypic sequences**, to now **genome-wide genealogical tree** representations.

Genetic structure: spotting patterns in genetic variation

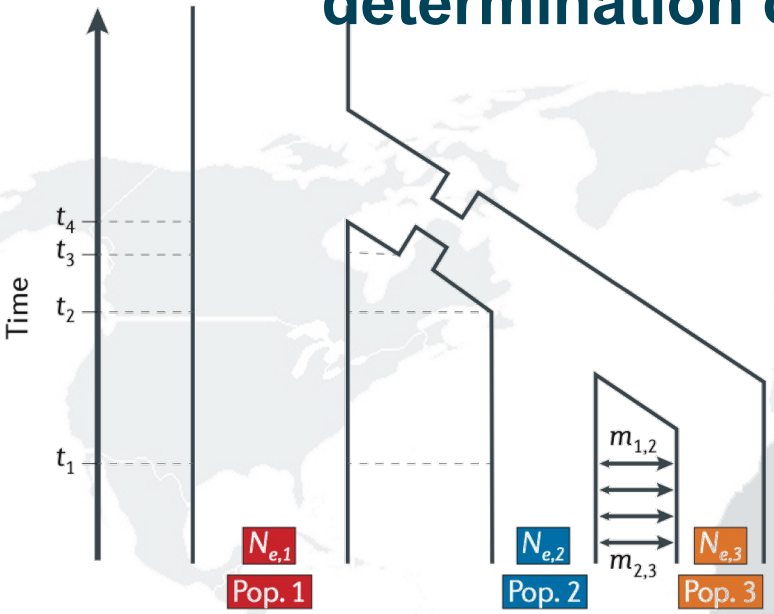
Genetic variation correlates with geography (Genes mirror geography)

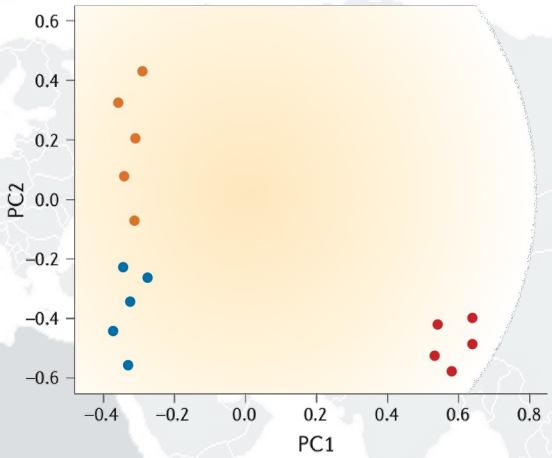


- PCA is a great & easy way to visualise genetic structure
- Requires no prior knowledge (model-free)
- But cannot quantify finer-scale structure and difficulties attributing to specific evolutionary events

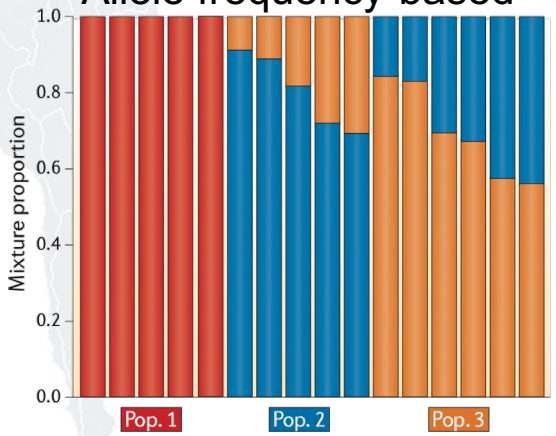
Novembre et al, Nature, 2008

Different approaches for determination of population structure

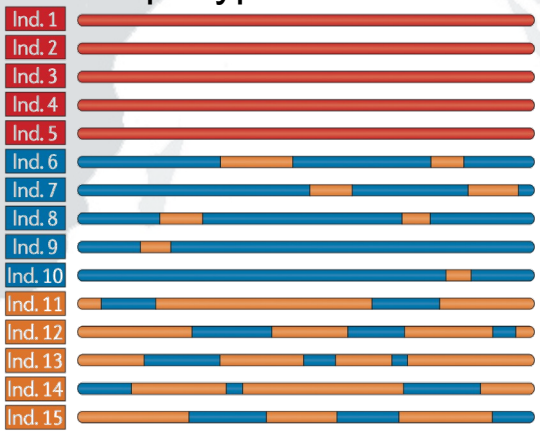




Allele frequency-based



Haplotype-based



Adapted from Schraiber & Akey. *Nature Review Genetics.* 2015.

Different approaches for determination of population structure

Name	Data type	Inference	Notes	Ref
STRUCTURE	Unlinked multi-allelic genotypes	Population structure, admixture	User friendly GUI; can be computationally demanding	Pritchard, Stephens & Donelly. Genetics. 2000.
ADMIXTURE	Unlinked bi-allelic SNPs	Population structure, admixture	Estimates the number of populations via cross-validation error	Alexander, Novembre, & Lange. Genome Res. 2009.
fineSTRUCTURE	Phased haplotypes	Population structure, admixture, chromosome painting	Can be used to identify the number and identity of populations	Lawson, Hellenthal, Myers & Falush. PLoS Genet. 2012.
GLOBETROTTER	Phased haplotypes	Population structure, admixture, chromosome painting	Estimate unsampled ancestral populations and admixture times	Hellenthal et al. Science. 2014.

Adapted from Schraiber & Akey. *Nature Review Genetics.* 2015.

Clustering algorithms

- We want to cluster n individuals into k clusters
  - Z_i: Cluster of individual i, e.g. $Z_1 = 2$ means individual 1 is in cluster 2
  - X: genotype matrix

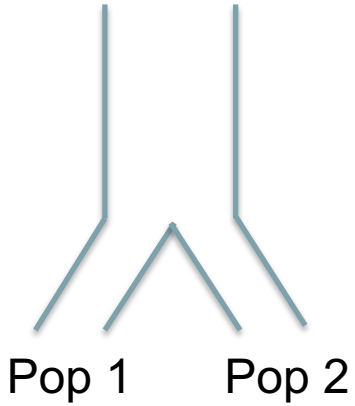
- **Modelling challenge**: How do we model $P(X \mid Z_1, Z_2, Z_3, ...)$?
  - E.g., allele frequencies, haplotypes

- **Computational challenge**: there are k^n possible assignments of individuals to clusters!
- E.g., 2^{100} is larger than the number of stars in the universe
- Solution: MCMC

Allele frequency-based clustering

Thought experiment:

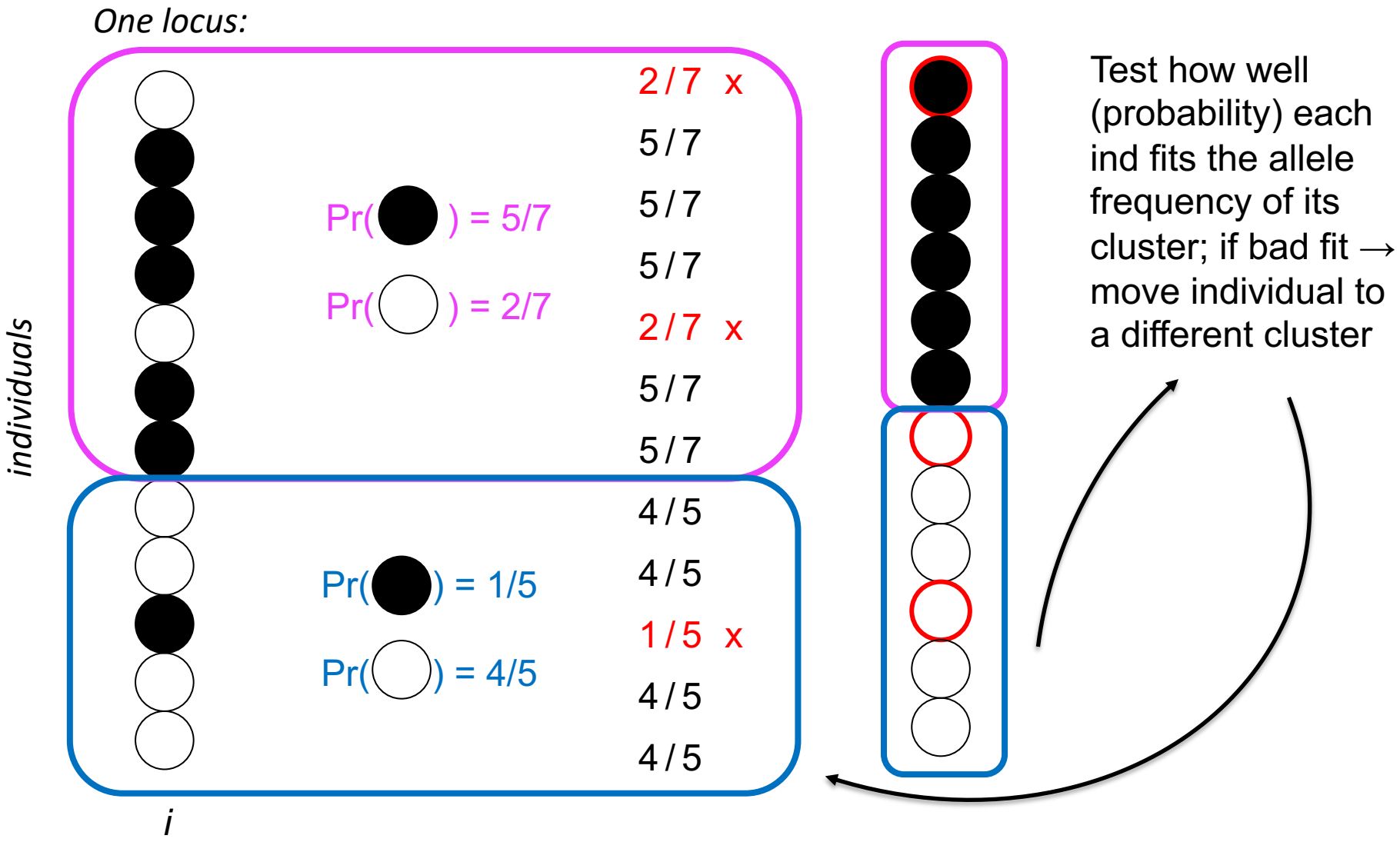
- Assume we sequenced individuals we knew are from Pop 1 and Pop 2
- We now sequence another individual, where we are unsure whether they are from Pop1 or Pop 2
- How could we try to assign this individual to Pop1 or Pop 2?



	Allele frequency in Pop1	Pop2	Genotype of individual	
SNP1	0.8	0.4	1	→ Pop1?
SNP2	0.3	0.7	0	→ Pop1?
SNP3	0.4	0.6	1	→ Pop2? (Pop1?)
SNP4	0.9	0.1	1	→ Pop1?

Allele frequency-based clustering

We can do the same **without** knowing allele frequencies in Pop1 and Pop2 by clustering

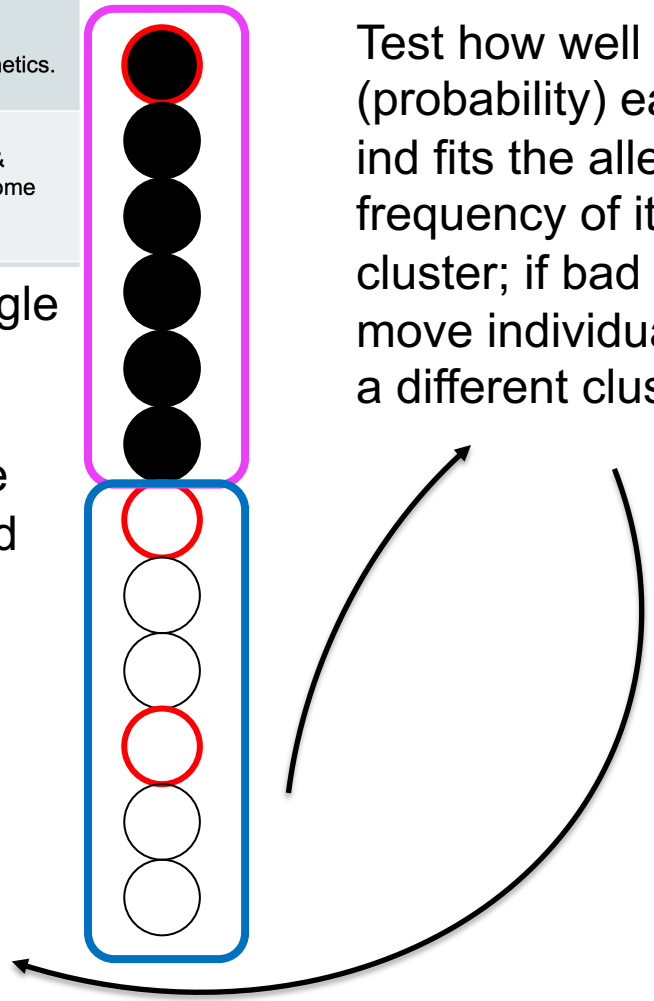


Allele frequency-based clustering

LD prune

Name	Data type	Inference	Notes	Ref
STRUCTURE	Unlinked multi-allelic genotypes	Population structure, admixture	User friendly GUI; can be computationally demanding	Pritchard, Stephens & Donelly. Genetics. 2000.
ADMIXTURE	Unlinked bi-allelic SNPs	Population structure, admixture	Estimates the number of populations via cross-validation error	Alexander, Novembre, & Lange. Genome Res. 2009.

- *"no admixture model"* – assign each ind *i* to single cluster *k*
- *"admixture model"* – assign each ind to multiple clusters (i.e. infer % of ind *i* 's genome assigned to clusters 1, ..., *K*)
- *"linkage model"* – can identify regions of ind *i* assigned to each cluster (Falush et al 2003, *Genetics* **164**:1567)
- Spatial priors on cluster membership.

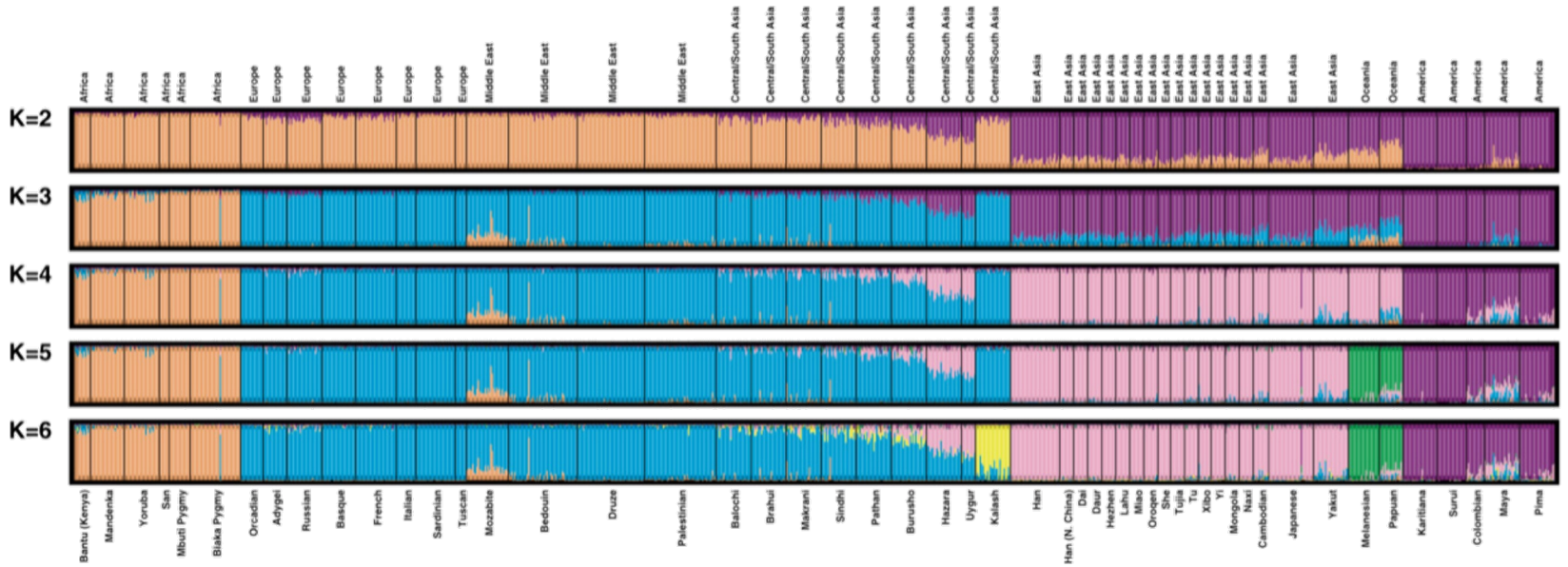


Test how well (probability) each ind fits the allele frequency of its cluster; if bad fit → move individual to a different cluster

Allele frequency-based clustering

Human Genome Diversity Panel

Rosenberg et al. *Science*. (2002), 298, 5602, 2381-2385.

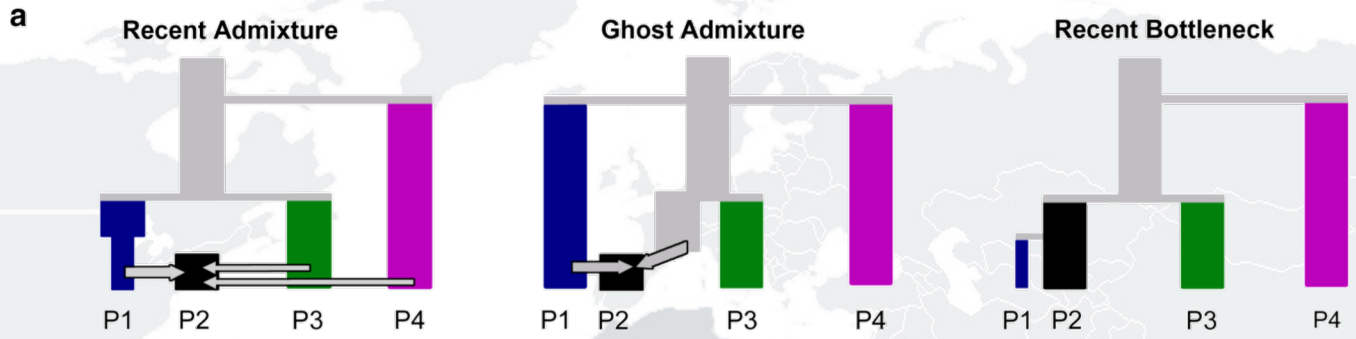


- Accuracy of population identification depends on the number of individuals included, the number of loci used and the allele frequency differences among those populations.

- Population identification can be strongly affected by the amount of admixture within individuals and the proportion of admixed individuals in the dataset – all populations are admixed….

Allele frequency-based clustering

Different combinations of drift, admixture and ancient relatedness can give similar signals



Recent admixture of P1, P3 & P4
Some parallels to African Americans

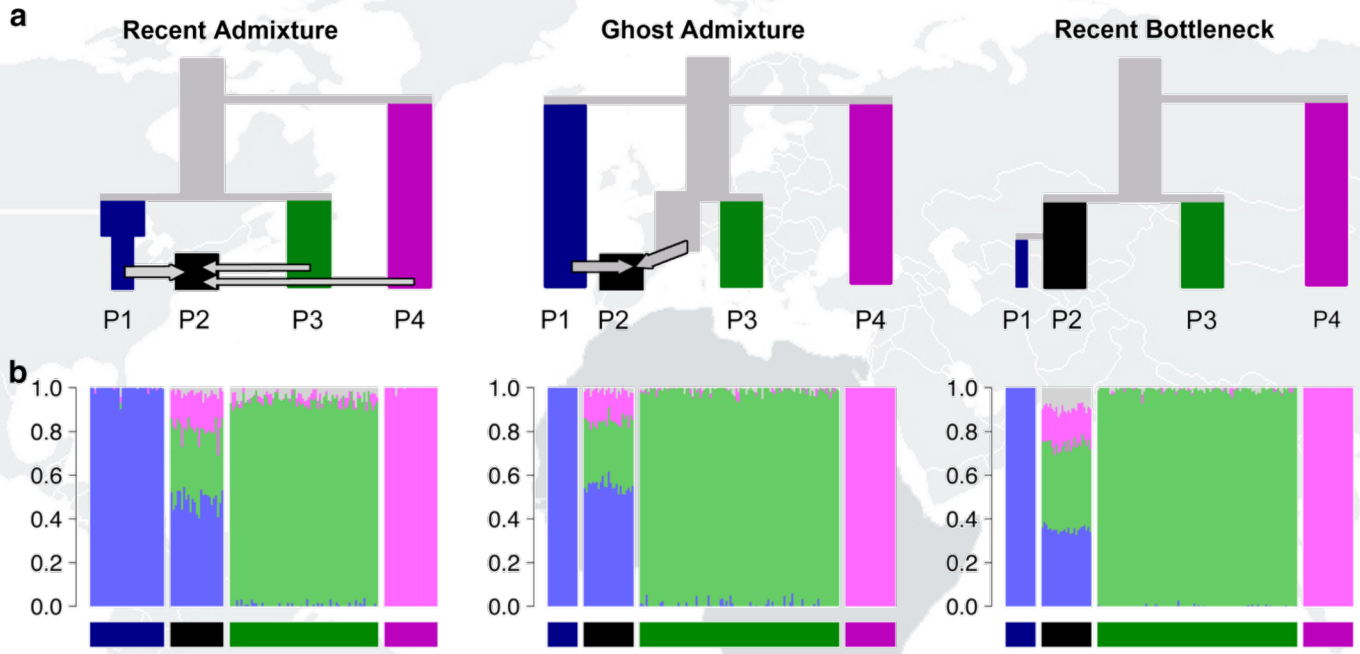
P2 50:50 admixture of P1 and unsampled population most closely related to P3

P1 – a sister population to P2 that underwent a strong recent bottleneck

Taken from :
A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots
Lawson, van Dorp, Falush, Nat Commun 2018

Allele frequency-based clustering

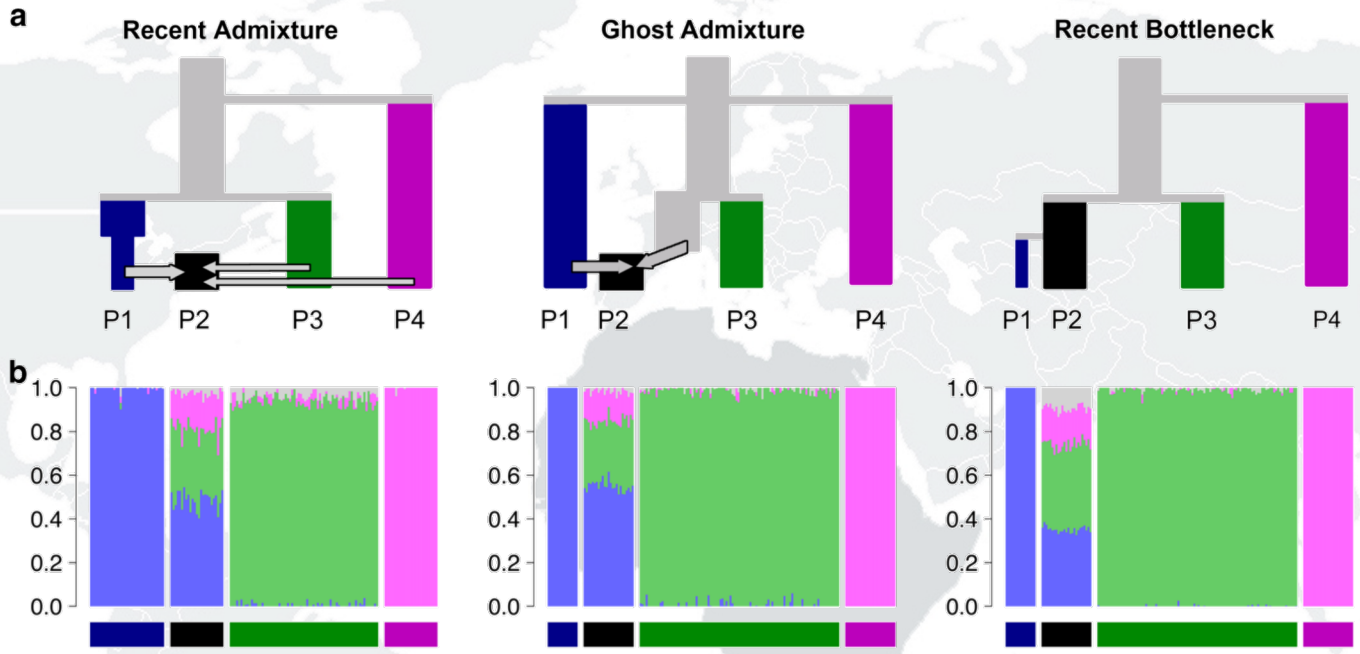
Different combinations of drift, admixture and ancient relatedness can give similar signals



Lawson, van Dorp & Falush. *Nat Comms*. (2018).

Allele frequency-based clustering

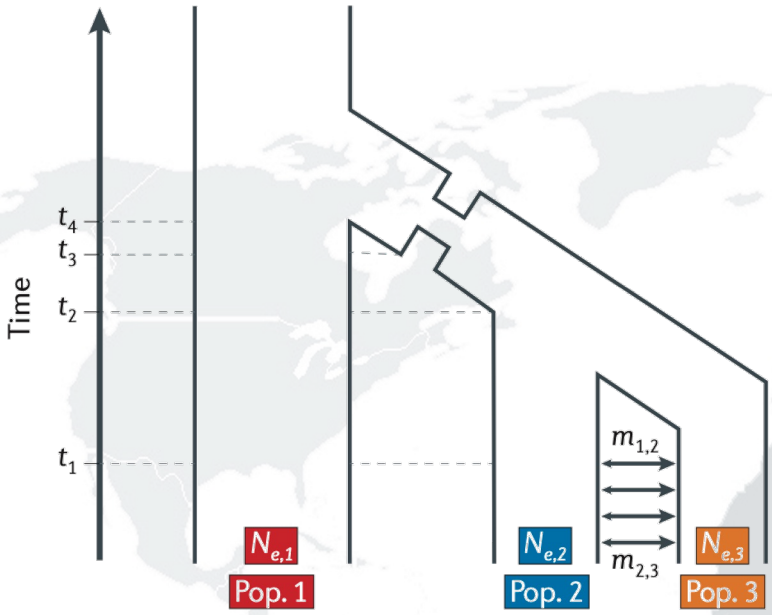
Different combinations of drift, admixture and ancient relatedness can give similar signals

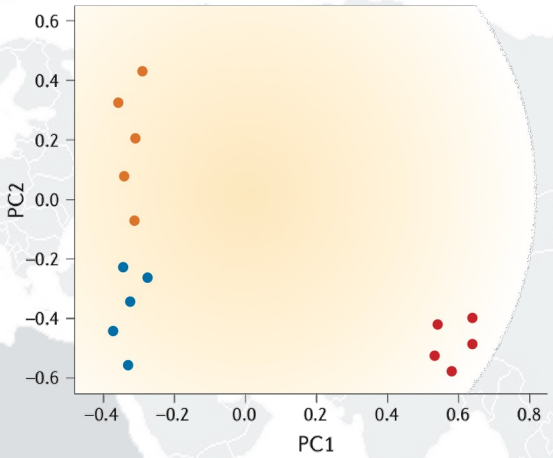


Lawson, van Dorp & Falush. *Nat Comms*. (2018).

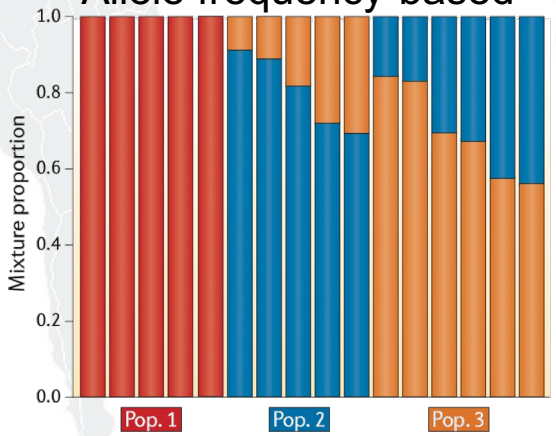
Ancestry assignments in STRUCTURE/ADMIXTURE often termed admixture proportions they do not identify where admixture has occurred → they do not alone indicate a history of admixture.

Determination of population structure

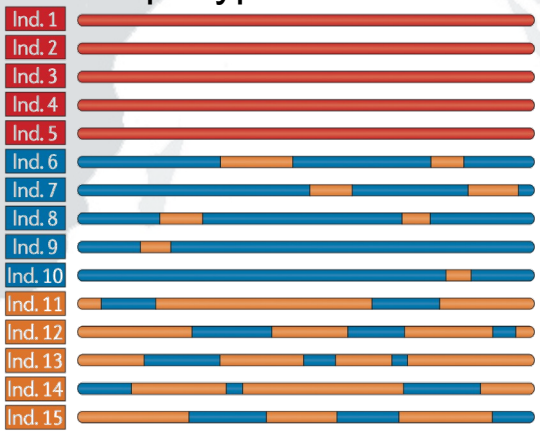




Allele frequency-based

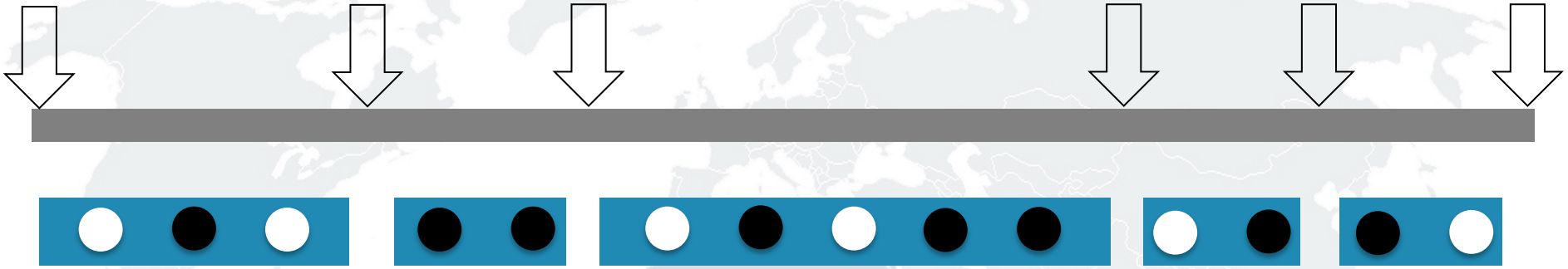


Haplotype-based



Adapted from Schraiber & Akey. *Nature Review Genetics.* 2015.

Incorporating haplotype information



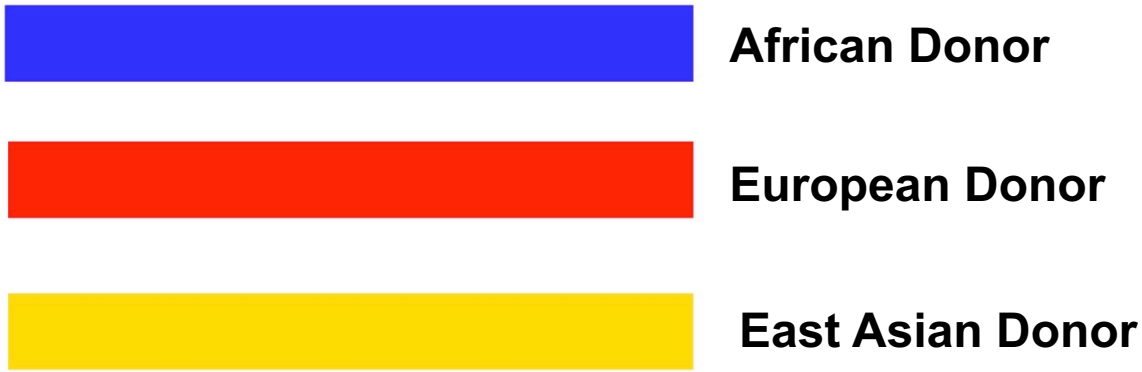
The genome as a mosaic
- blocks of LD separated by recombination

Incorporating haplotype information

Name	Data type	Inference	Notes	Ref
HAPMIX	Phased haplotypes, reference panel	Chromosome painting	Requires populations to be pre-specified	Price et al. PLoS Genetics. 2009
LAMP	Phased haplotypes	Chromosome painting	Identifies local ancestry in windows, rather than using an HMM	Sankararaman et al. Am. J. Hum. Gen. 2008.
PCAdmix	Phased haplotypes	Chromosome painting, population structure	Uses PCA in small chunks followed by HMM to estimate local ancestry	Brisbin et al. Hum. Biol. 2012.
fineSTRUCTURE	Phased haplotypes	Population structure, admixture, chromosome painting	Can be used to identify the number and identity of populations	Lawson, Hellenthal, Myers & Falush. PLoS Genet. 2012.
GLOBETROTTER	Phased haplotypes	Population structure, admixture, chromosome painting	Estimate unsampled ancestral populations and admixture times	Hellenthal et al. Science. 2014.

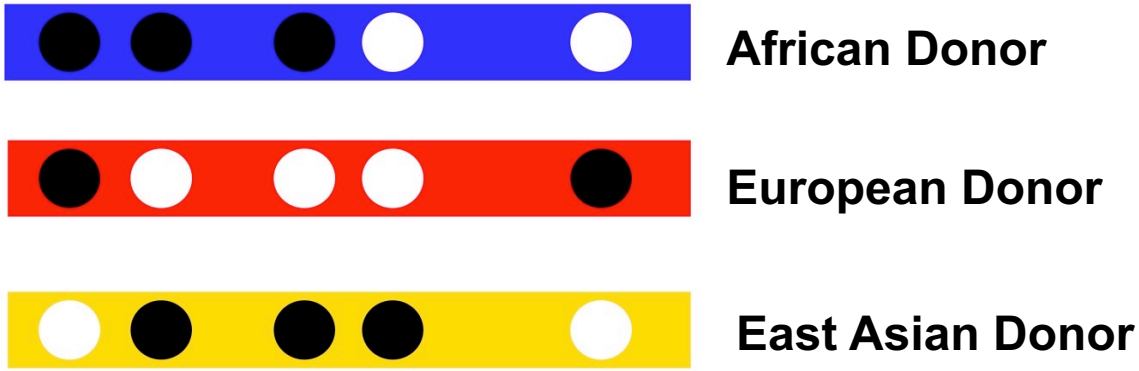
Adapted from Schraiber & Akey. *Nature Review Genetics.* 2015.

Incorporating haplotype information
Chromosome painting



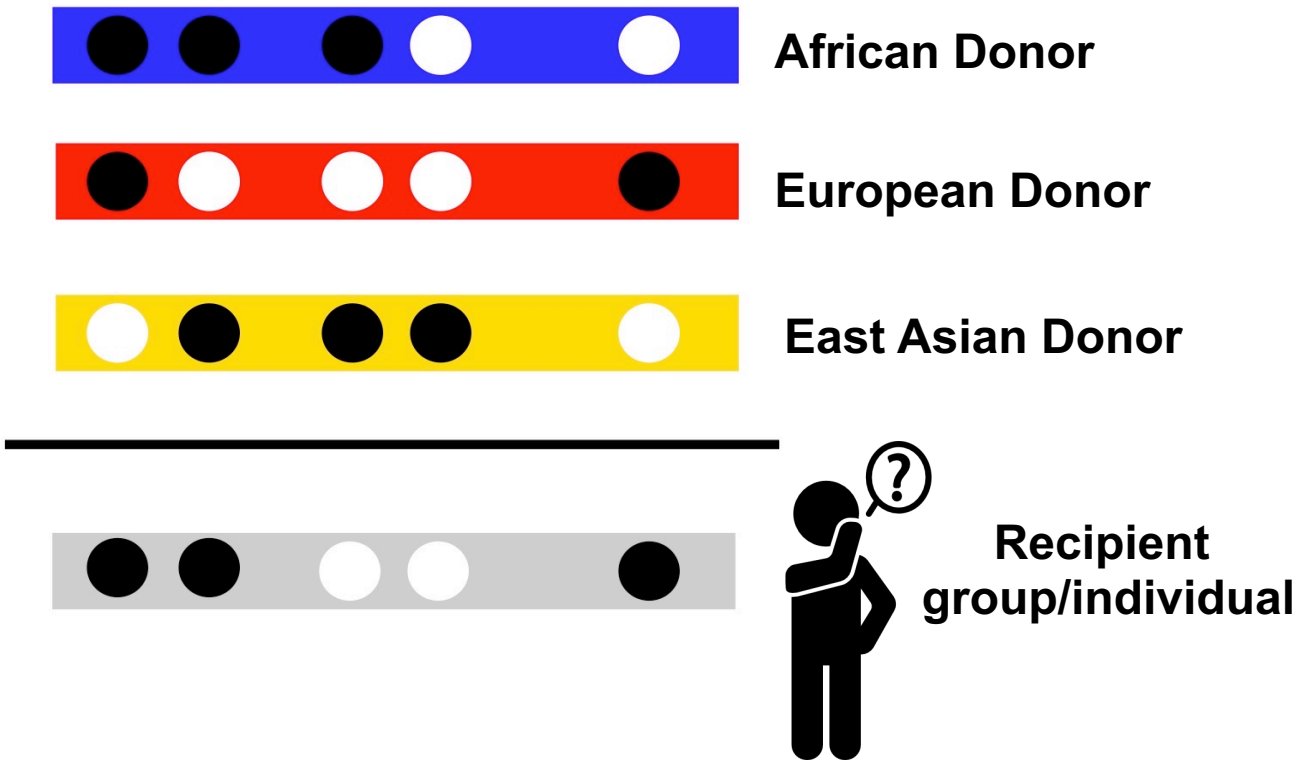
Li & Stephens, *Genetics*, 2003
Lawson et al, *PLoS Genetics*, 2012

Incorporating haplotype information
Chromosome painting



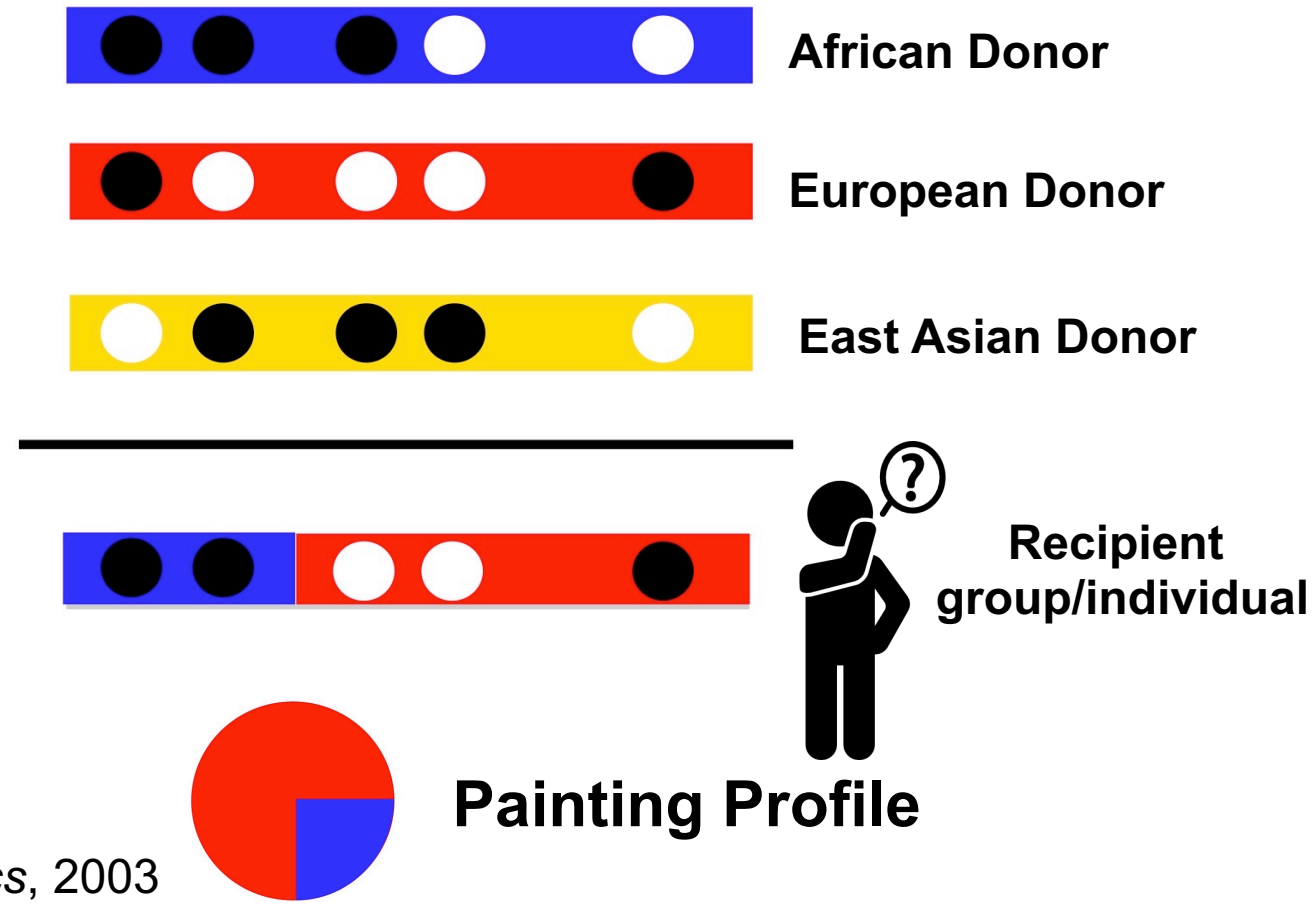
Li & Stephens, *Genetics*, 2003
Lawson et al, *PLoS Genetics*, 2012

Incorporating haplotype information
Chromosome painting



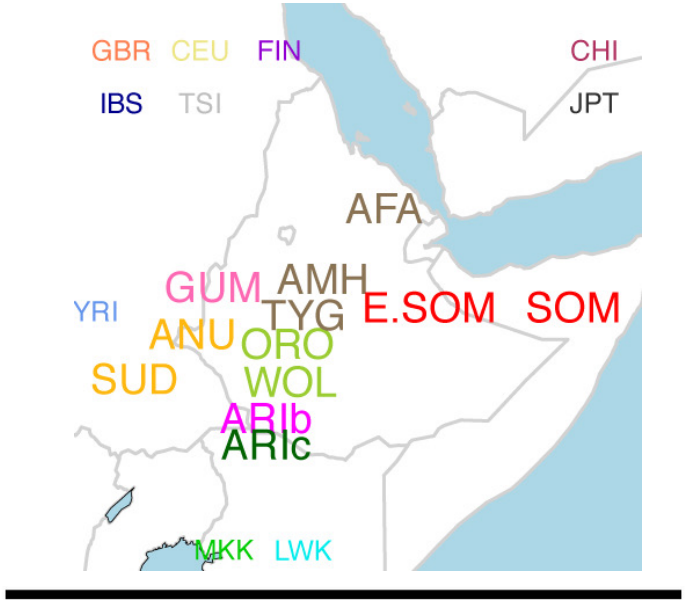
Li & Stephens, *Genetics*, 2003
Lawson et al, *PLoS Genetics*, 2012

Incorporating haplotype information
Chromosome painting



Li & Stephens, *Genetics*, 2003
Lawson et al, *PLoS Genetics*, 2012

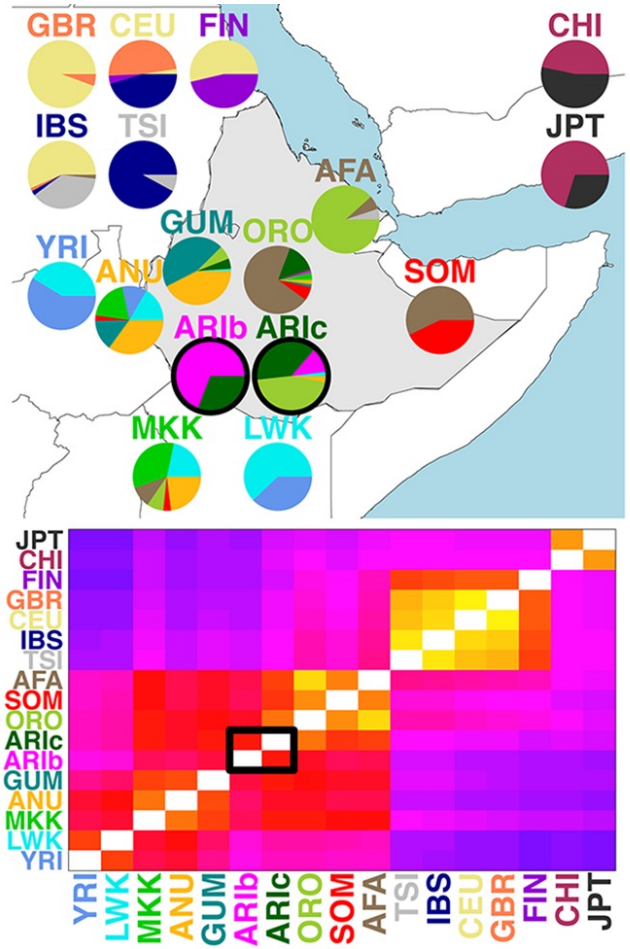
Incorporating haplotype information
Chromosome painting



All-donors

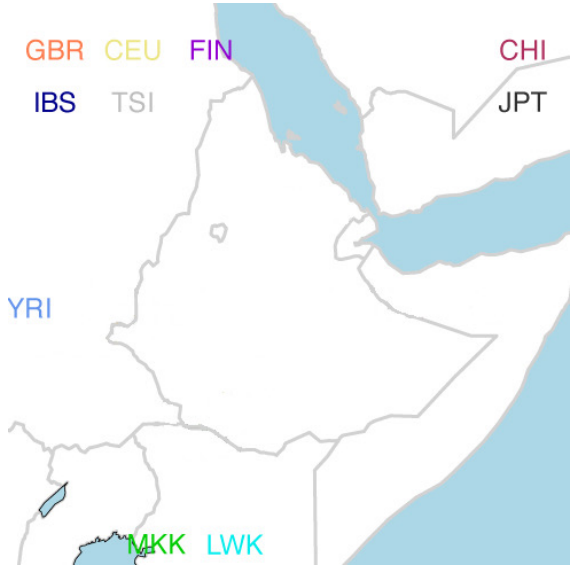
ARI blacksmith/
ARI cultivators

Average haplotype length shared = 2.76cM

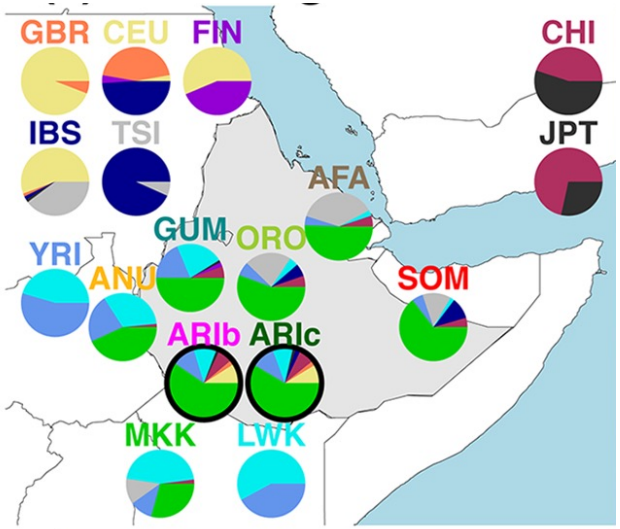


van Dorp et al, *PLoS Genetics*, 2015

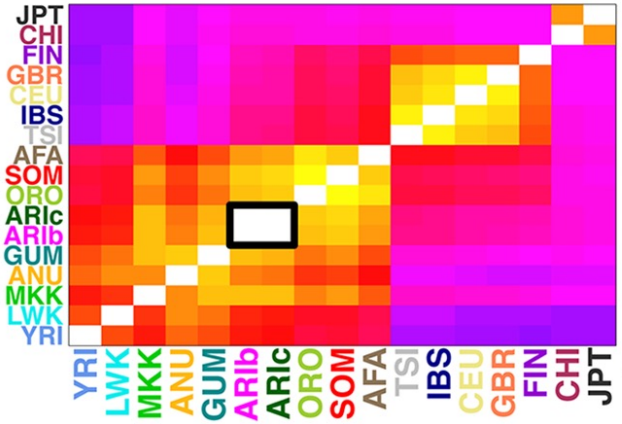
Incorporating haplotype information
Chromosome painting



No Ethiopia



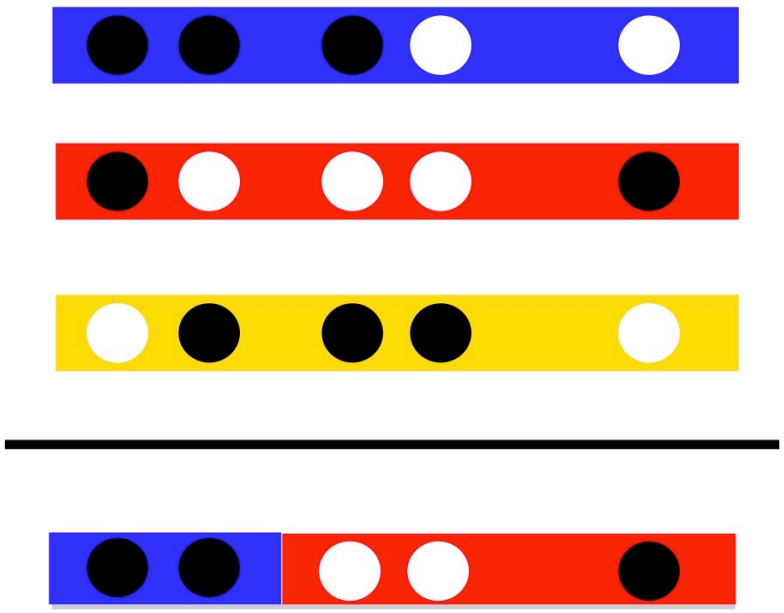
ARI blacksmith/
ARI cultivators

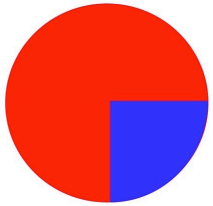


Average haplotype length shared = 0.65cM

van Dorp et al, *PLoS Genetics*, 2015

Haplotype clustering



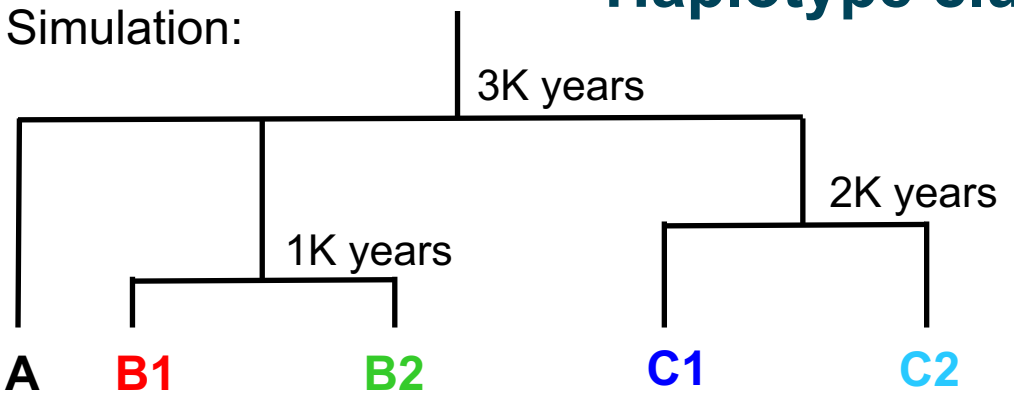


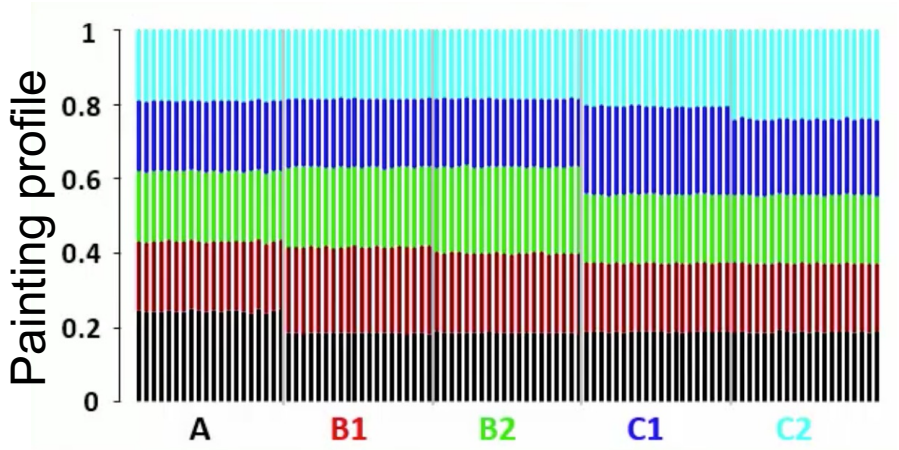
Clustering
fineSTRUCTURE

Painting Profile

Li & Stephens, *Genetics*, 2003
Lawson et al, *PLoS Genetics*, 2012

Haplotype clustering

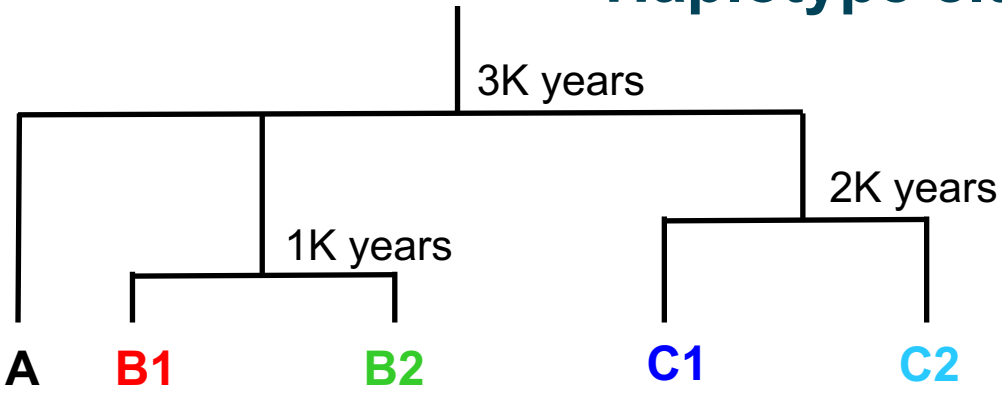


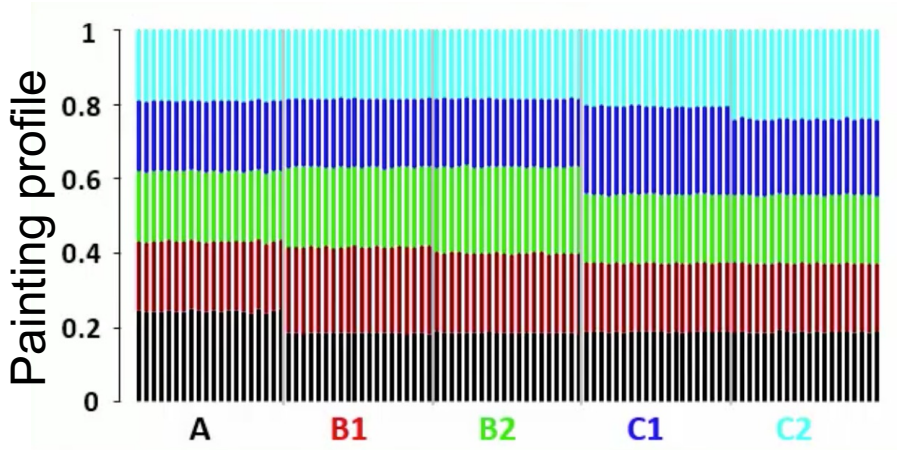


Donor individual = *j*
Recipient individual = *i*
Number of chunks of DNA by which individual i is painted by individual j = *Yij*

Lawson et al. *PLoS Genetics*, 2012

Haplotype clustering





Assign individuals to a random set of clusters *k*

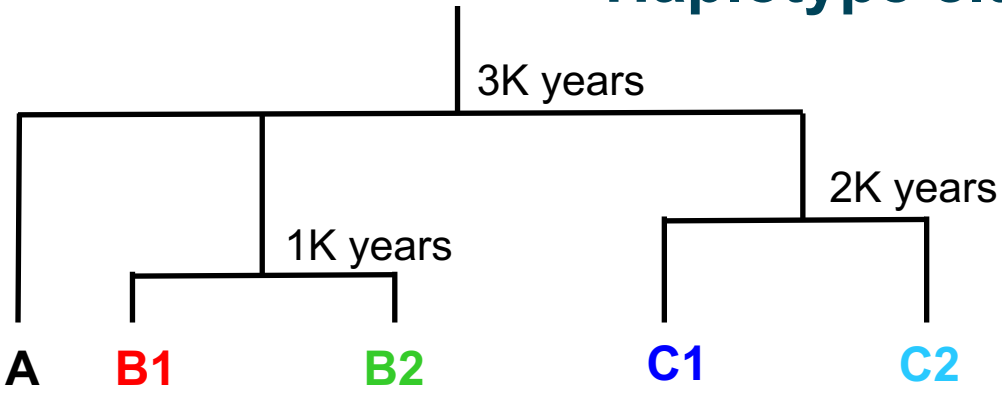
$(Y_{i1},\ldots,Y_{ik}) \sim \text{Mult}(P_{A1},\ldots,P_{Ak})$ for ind i assigned to cluster A

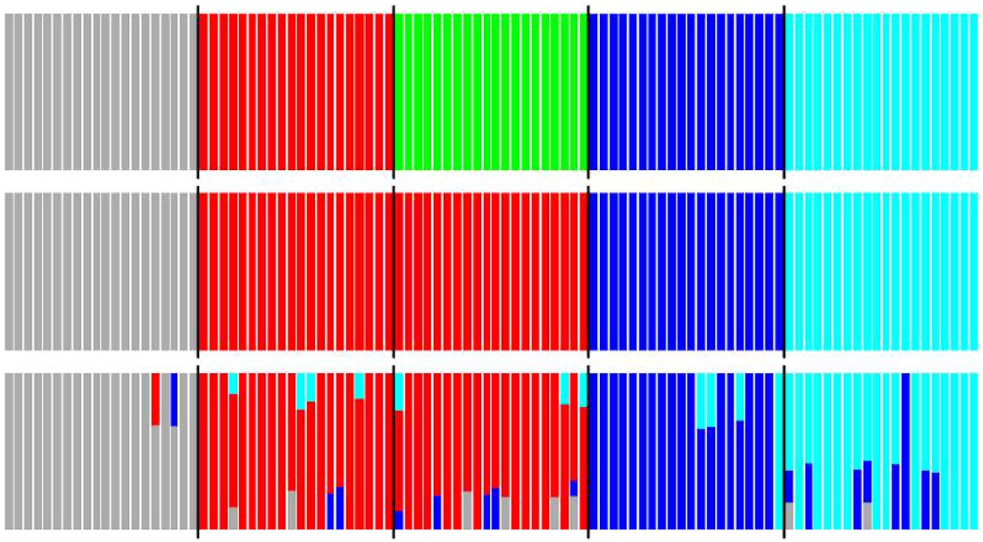
1. Infer P_{AK} – number of chunks individuals in cluster A are painted with individuals in cluster *k*
2. Test if $\Pr(Y_{i1},\ldots,Y_{ik})$ is low – if so change cluster and repeat!

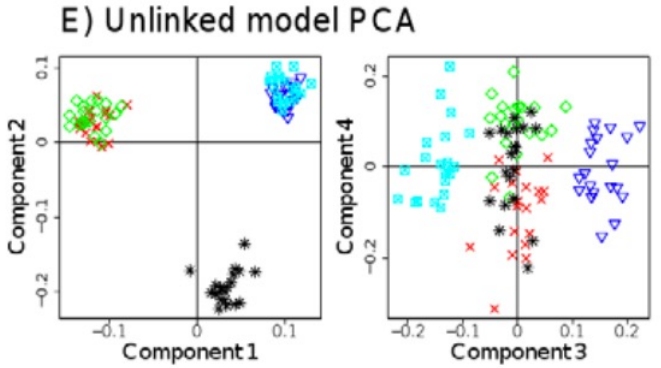
Donor individual = *j*
Recipient individual = *i*
Number of chunks of DNA by which individual I is painted by individual j = *Yij*

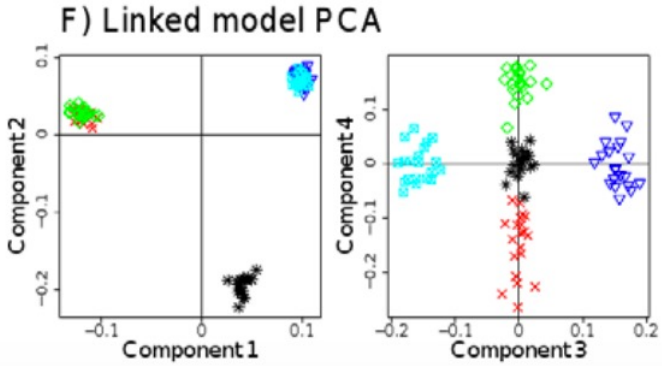
Lawson et al. *PLoS Genetics*, 2012

Haplotype clustering



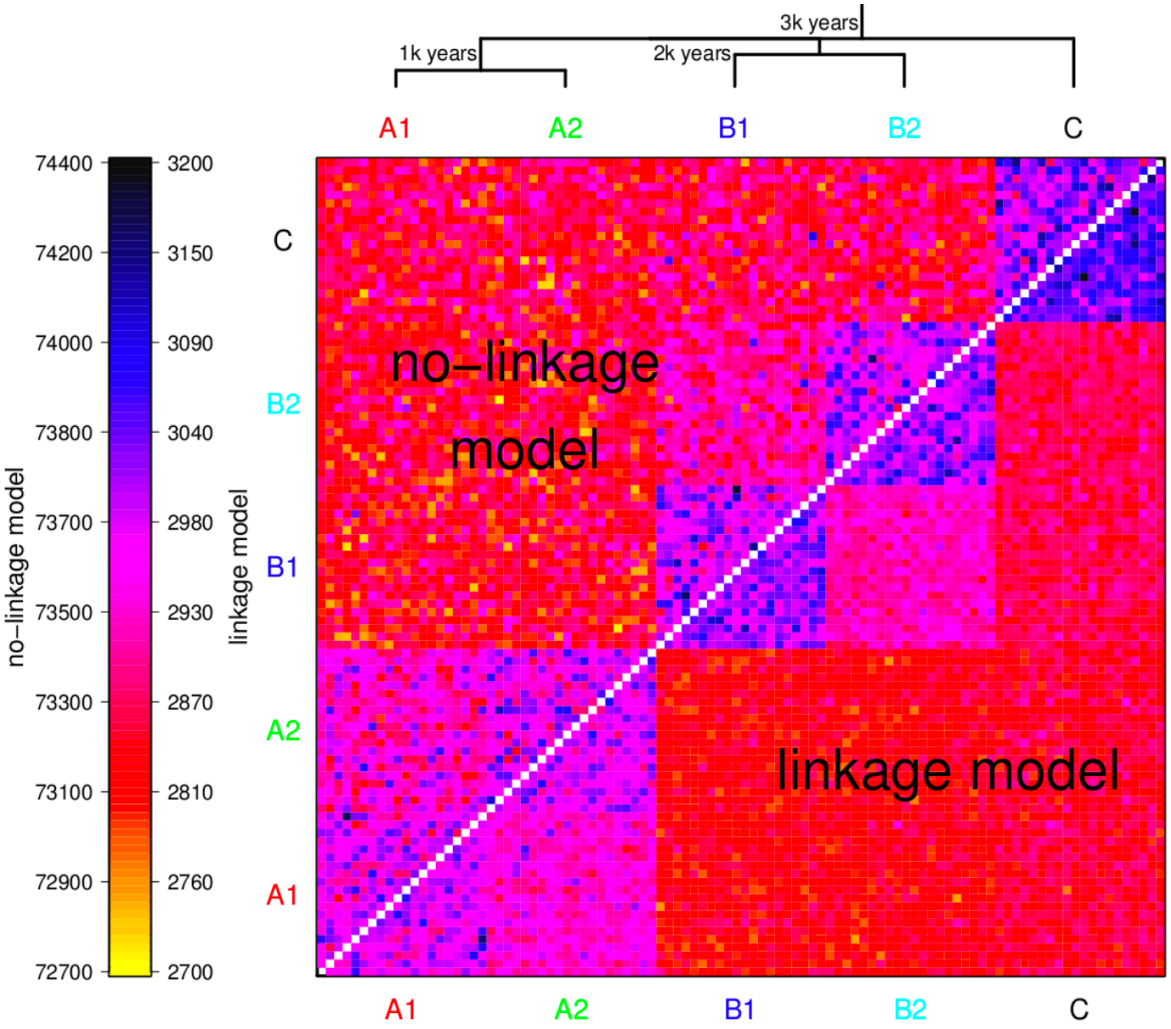






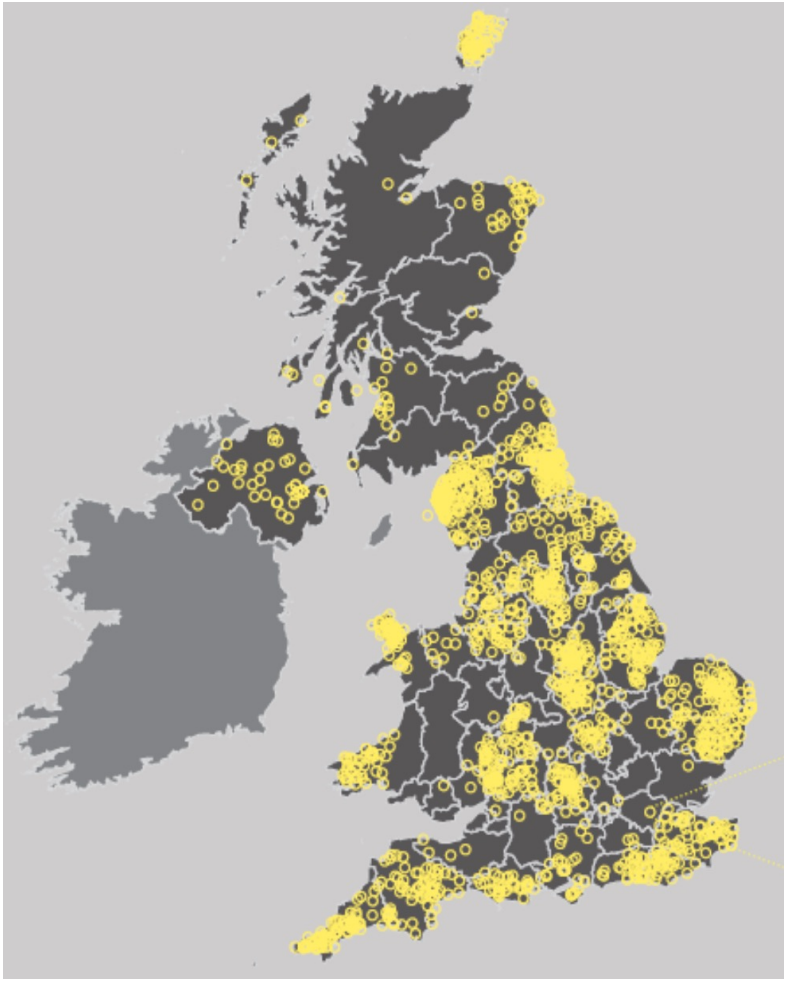
Lawson et al. *PLoS Genetics*, 2012

Haplotype clustering



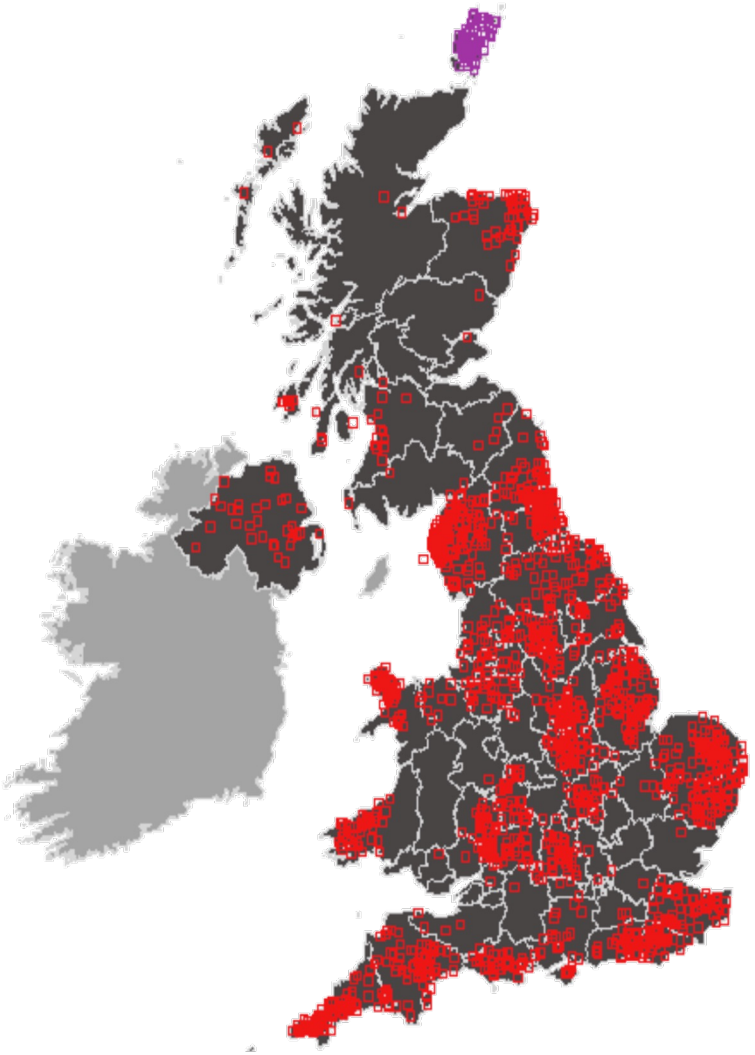
Haplotype clustering in the British Isles

- 2,039 individuals collected across rural areas of UK
- All four-grandparents born within 80km of one another
- Inferred DNA matching patterns amongst all individuals
- Grouped individuals into hierarchical groups based on these DNA patterns



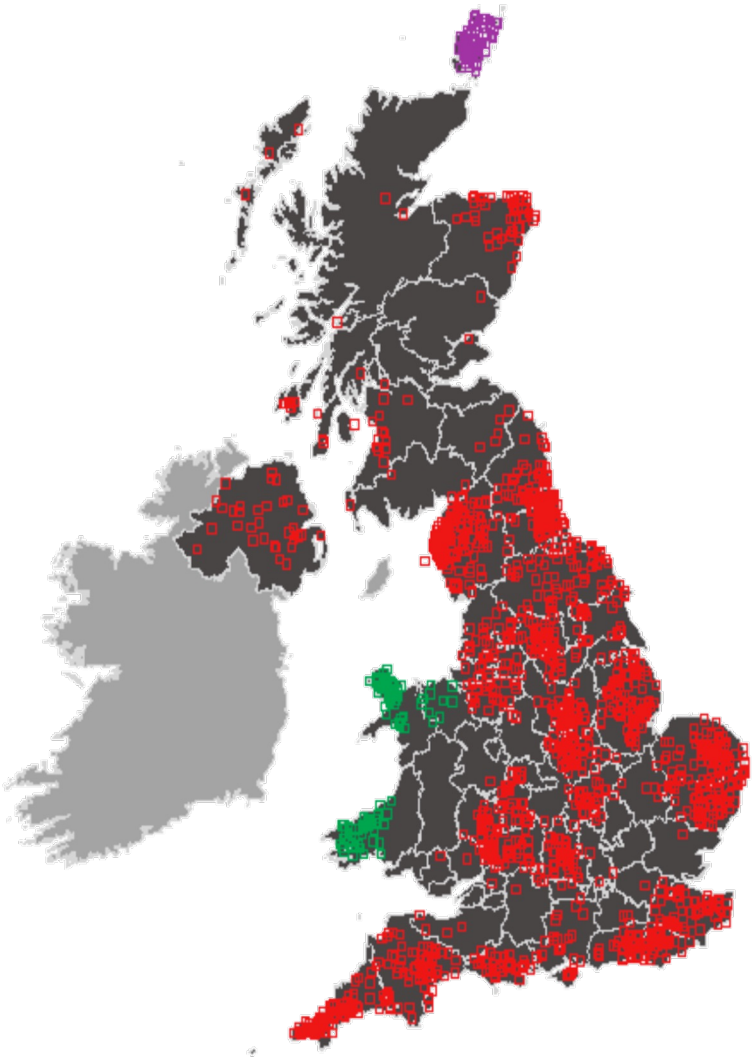
Leslie et al. *Nature*, 2015

Haplotype clustering in the British Isles



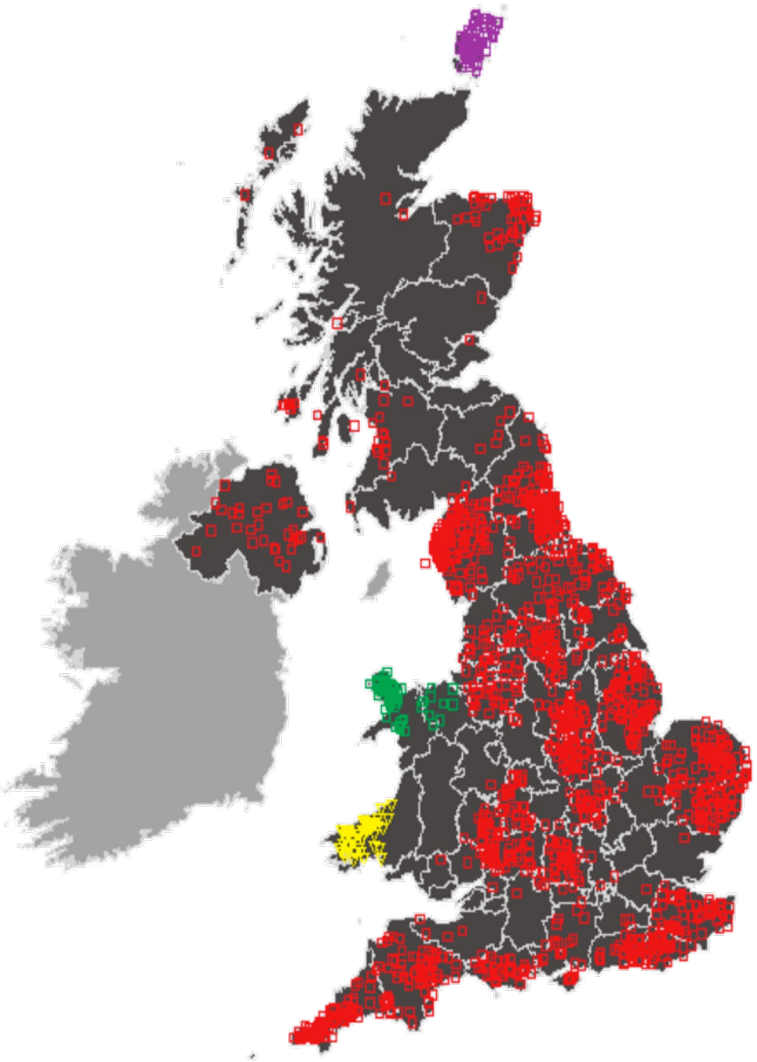
Leslie et al. *Nature*, 2015

Haplotype clustering in the British Isles



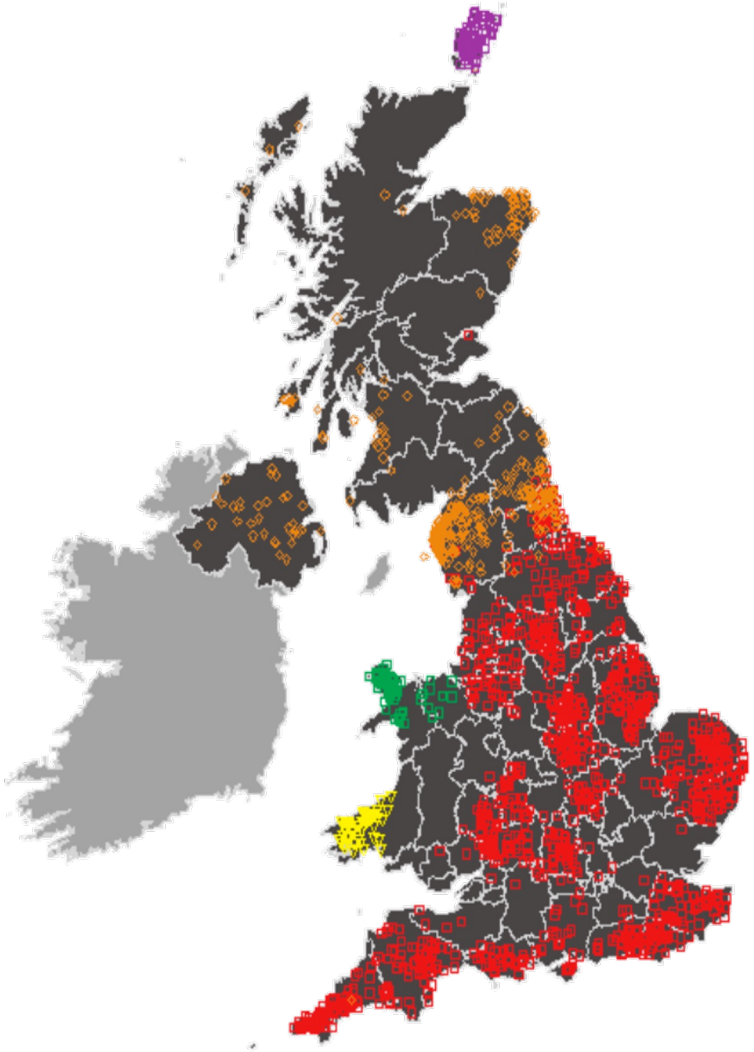
Leslie et al. *Nature*, 2015

Haplotype clustering in the British Isles



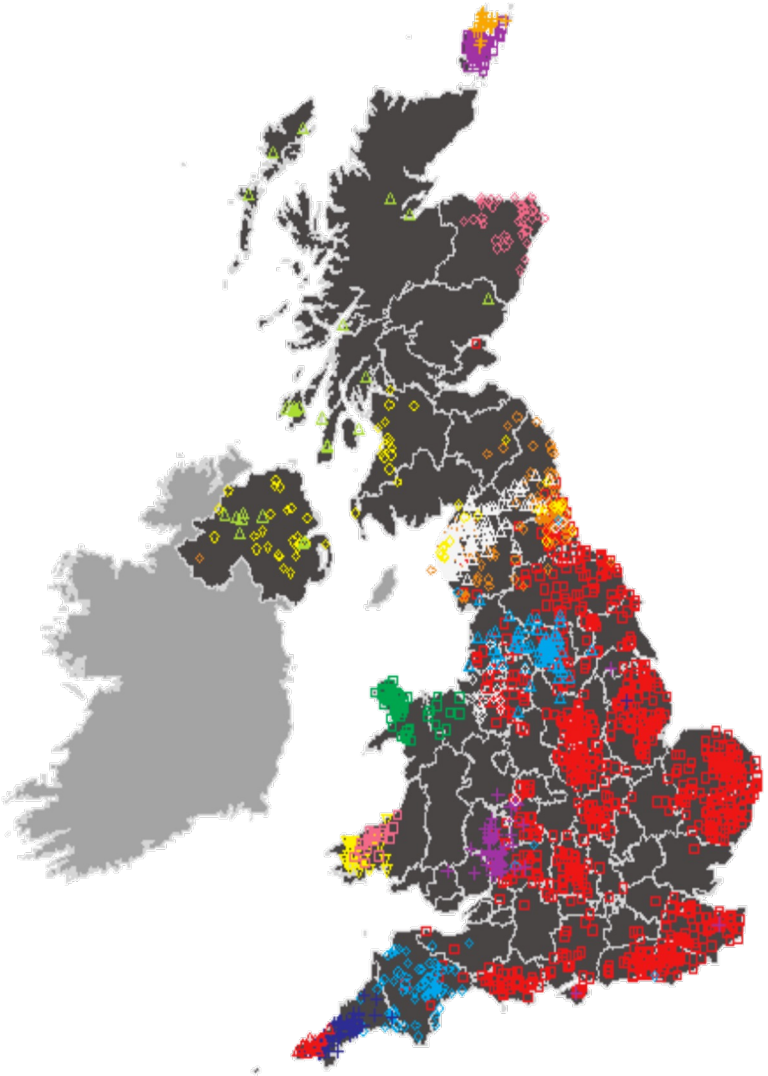
Leslie et al. *Nature*, 2015

Haplotype clustering in the British Isles



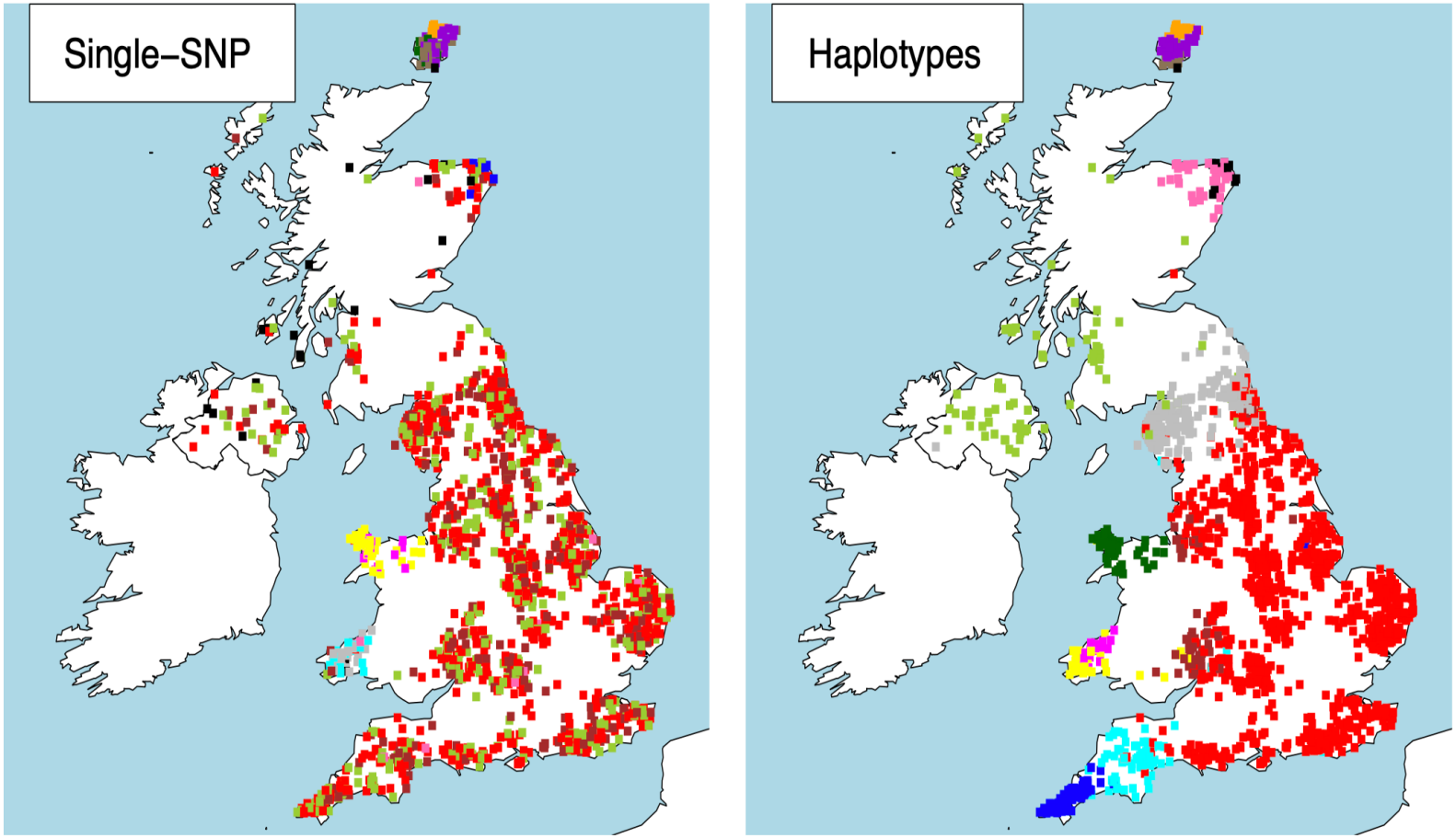
Leslie et al. *Nature*, 2015

Haplotype clustering in the British Isles



Leslie et al. *Nature*, 2015

Haplotype clustering in the British Isles



Leslie et al. *Nature*, 2015

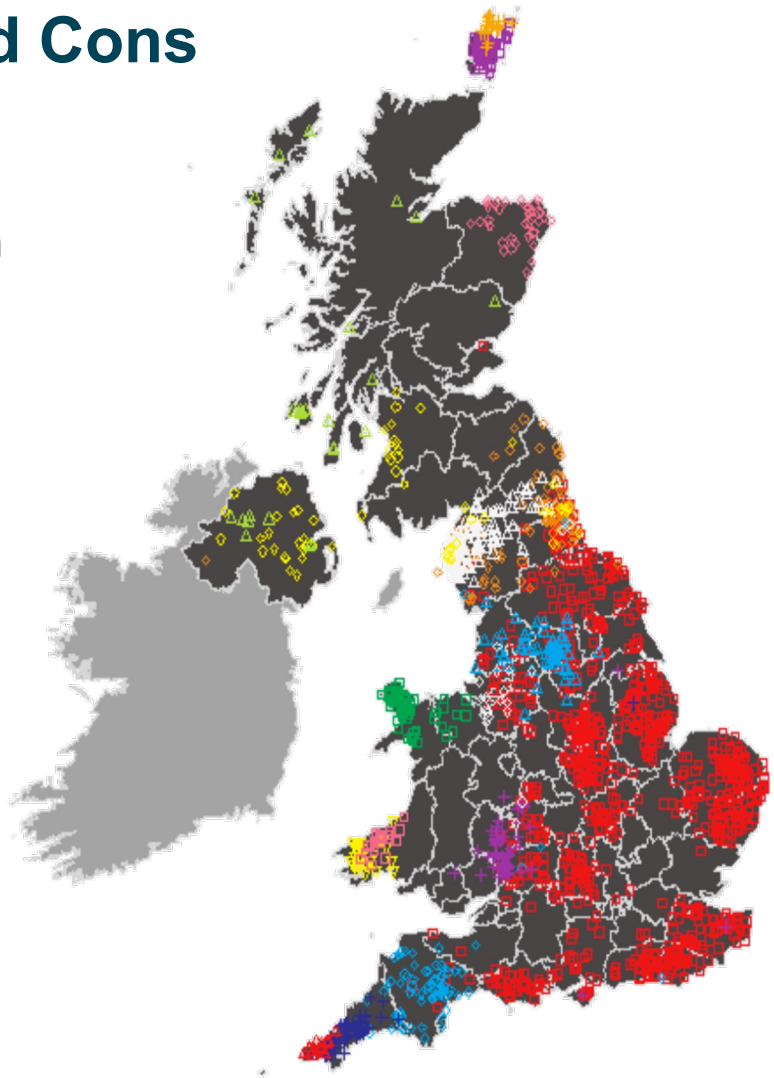
Haplotype clustering - fineSTRUCTURE Pros and Cons

Advantages:

- Increased power to find more subtle population structure
- Heatmaps demonstrate presence of structure/admixture
- Current implementation infers number of clusters K automatically and builds a tree

Disadvantages

- Challenges of interpretation – drift/admixture/other
- Requires phased data
- Computationally slow compared to ADMIXTURE

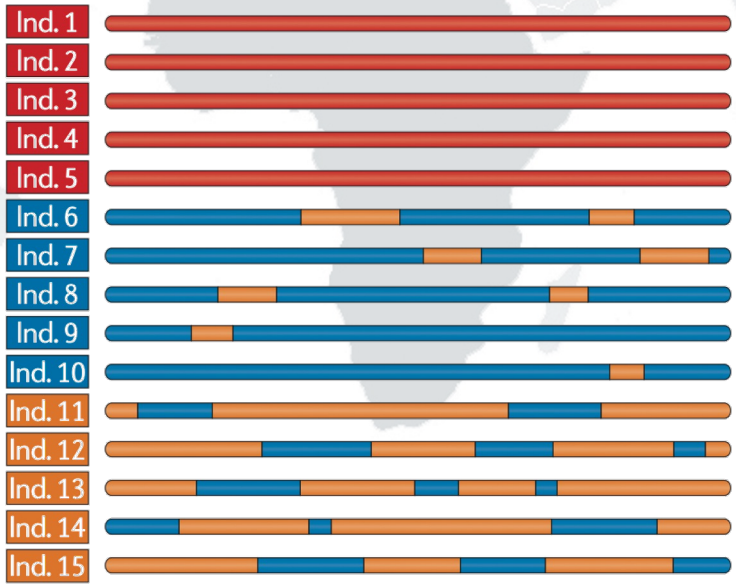


Leslie et al. *Nature*, 2015

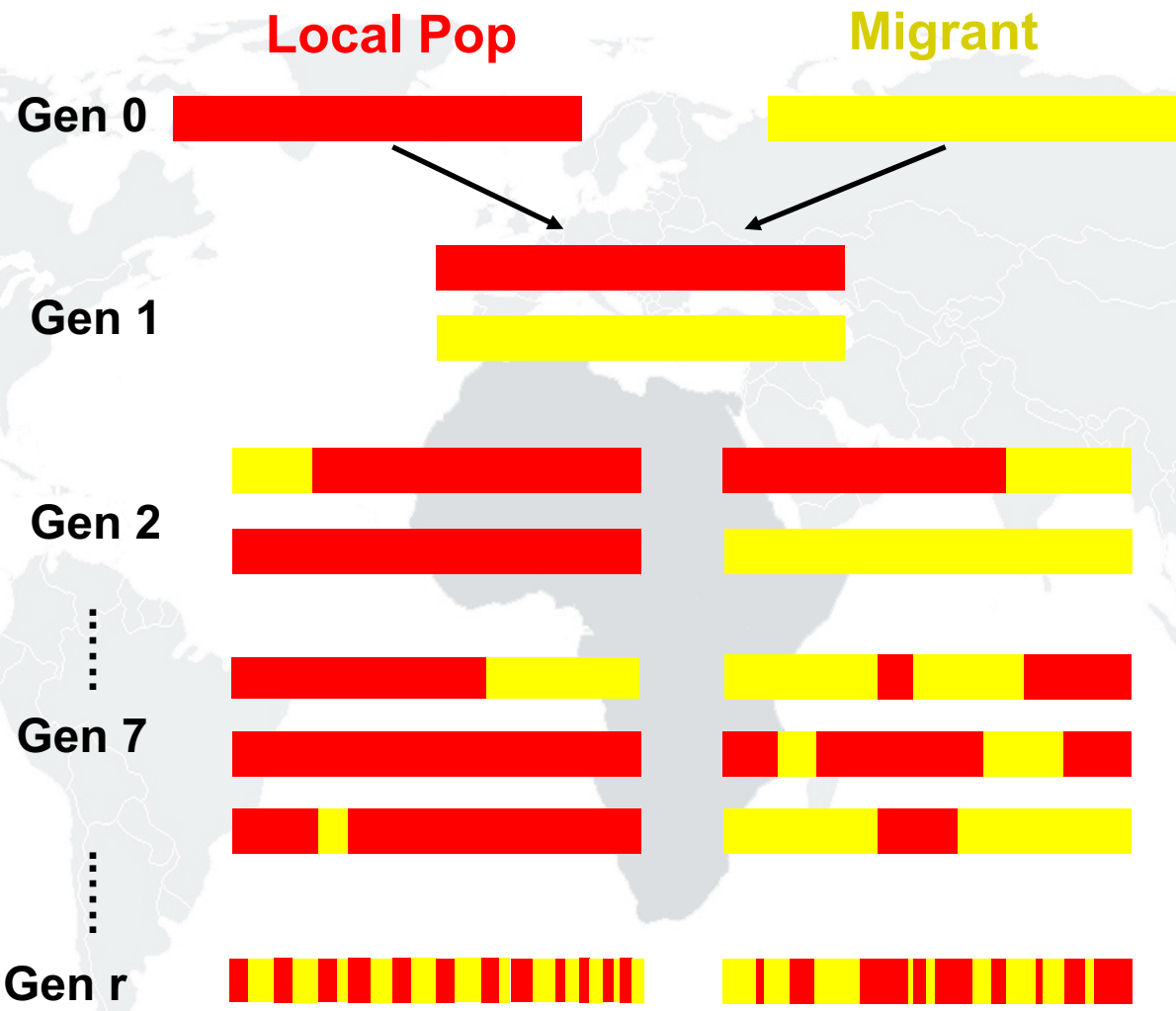
Haplotype sharing patterns can also be used to infer admixture events in human populations

- Admixture = mixing/interbreeding of two genetically distinguishable groups

- We can detect admixture in the genome by exploring patterns of decay of LD over genetic distance using chunks inferred from chromosome painting



Haplotype sharing patterns can also be used to infer admixture events in human populations



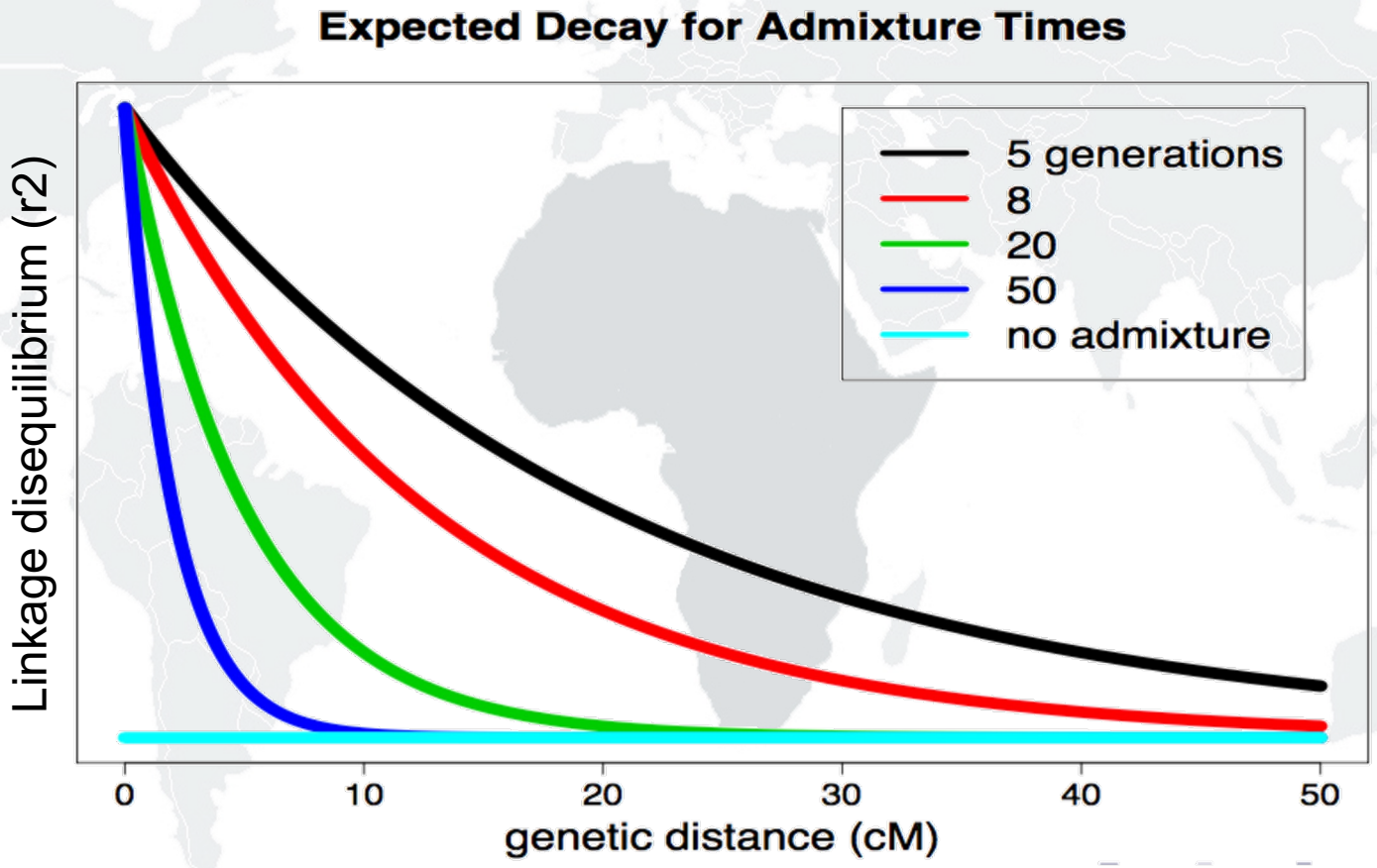
LD decays with genetic distance and this can be used for admixture inference

- Genetic recombinations are well modeled by a Poisson random variable (Falush et al, *Genetics*, 2003) → where the time between 'arrivals' of a Poisson process follows an exponential distribution.

- The length of red and yellow depends on the number of recombination's that have occurred.
  - Eg. a continuous tract is one in which no ancestry-breaking recombinations have occurred.
  - Thus, tract length distribution should be exponential.

- Can we use this distribution to approximate the number of generations since admixture?

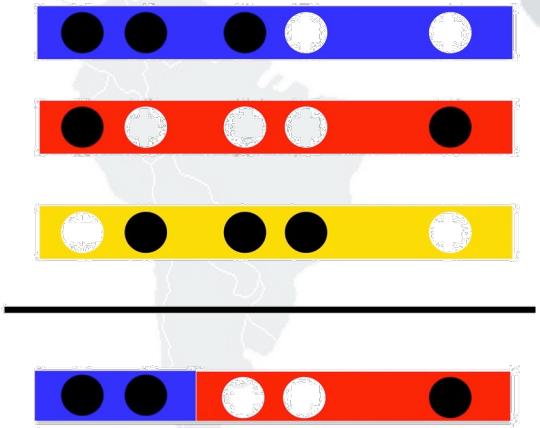
LD decays with genetic distance and this can be used for admixture inference

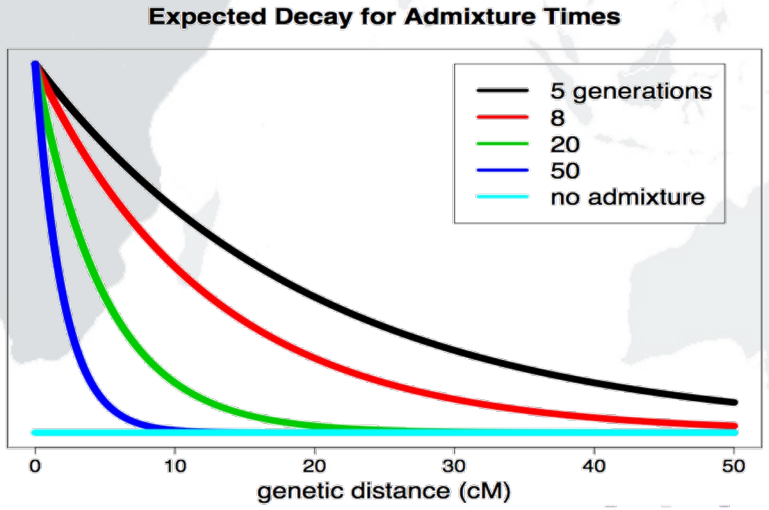


Inferring admixture events in human populations

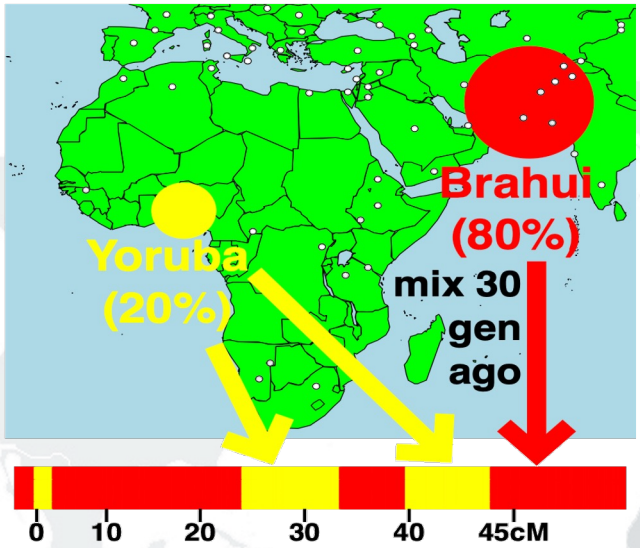
- Several methods that use admixture LD to infer, identify and date admixture events eg. ROLLOFF (Moorjani et al, AJHG, 2013), ALDER (Loh et al, PLoS Genetics, 2013), and GLOBETROTTER (Hellenthal et al, Science, 2014).

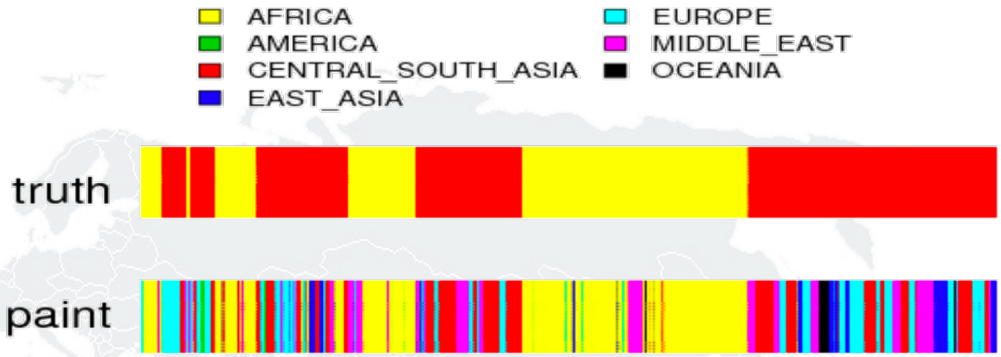
- DNA → chromosome painting (infer which groups) → size of segments → fit exponentials

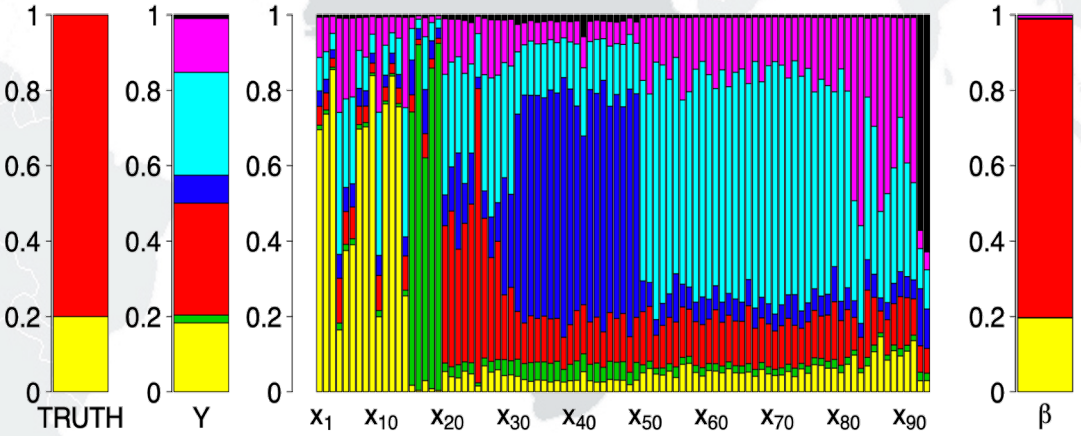




Inferring admixture events - Globetrotter



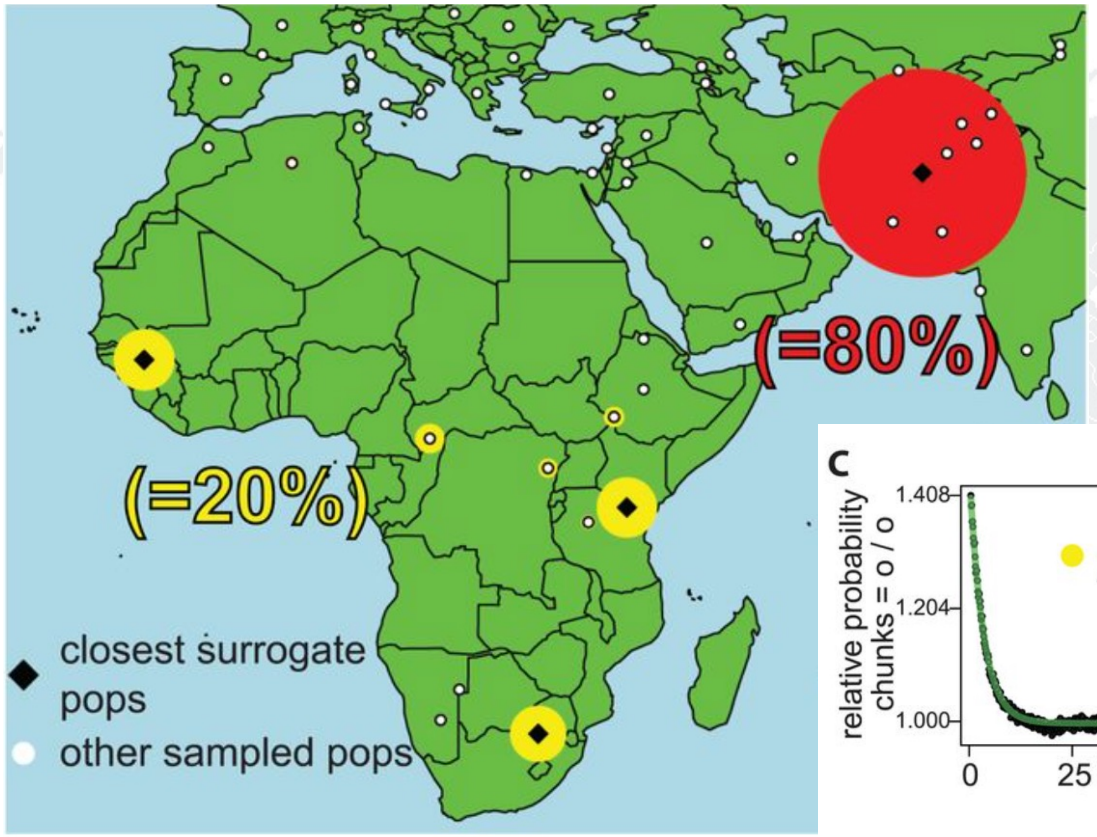


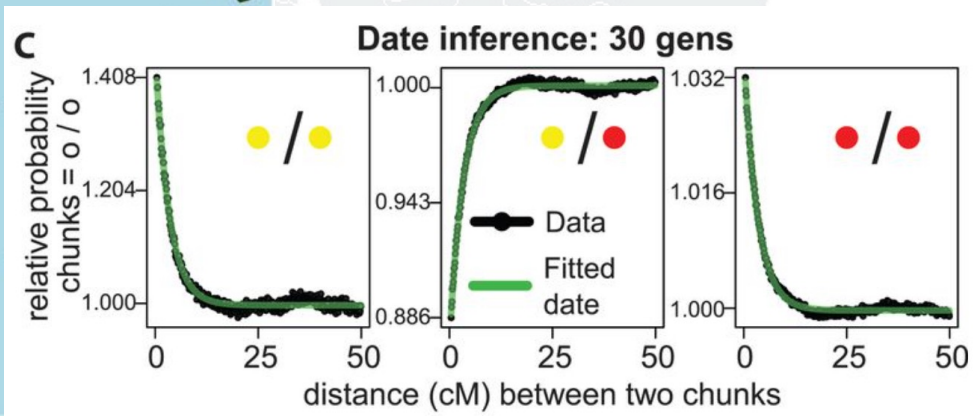


$$E[Y] = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{93} X_{93}$$

Hellenthal et al, Science, 2014

Inferring admixture events - Globetrotter

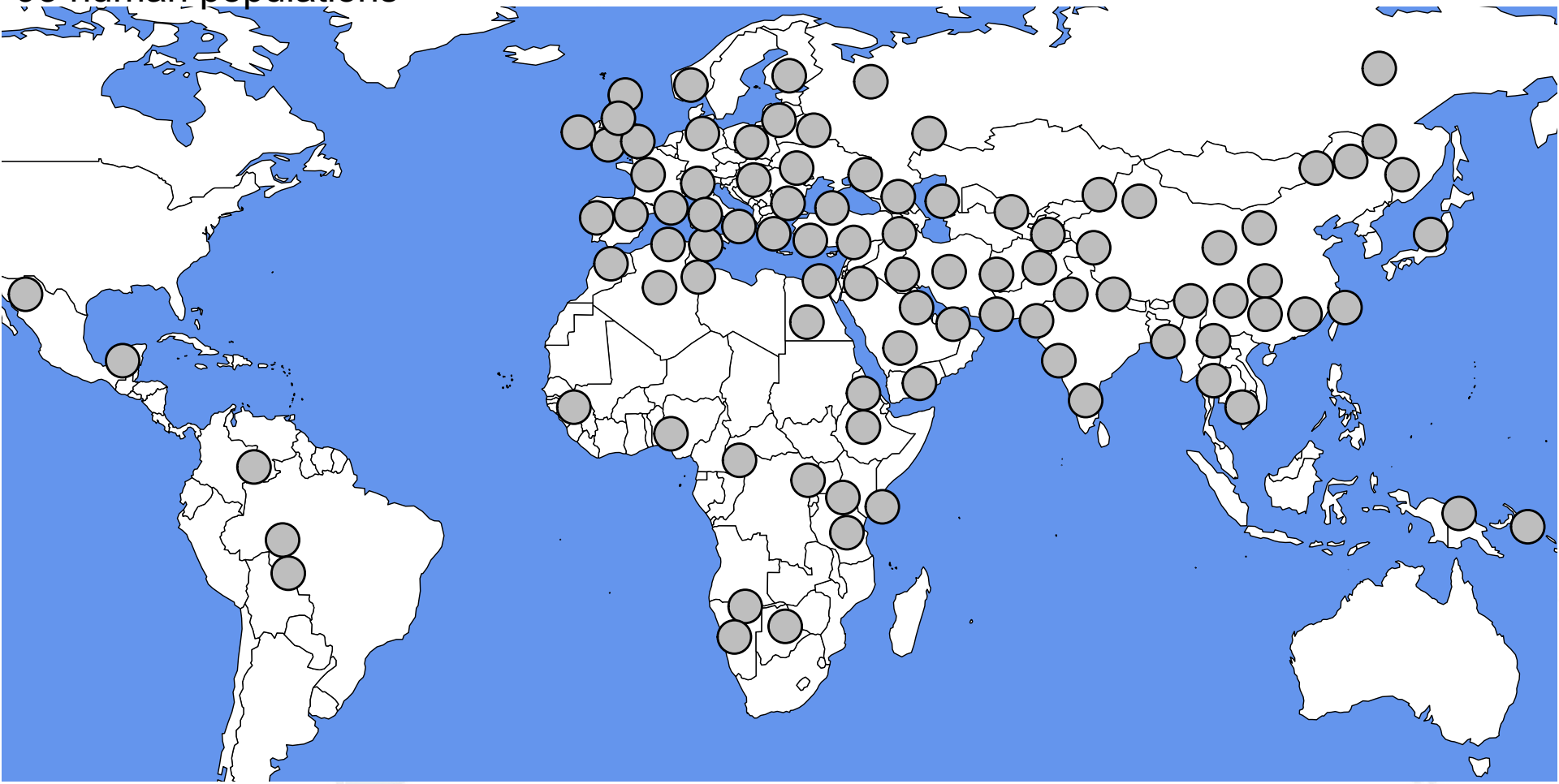




Hellenthal et al, Science, 2014

Genetic mixing over the past 4000 years

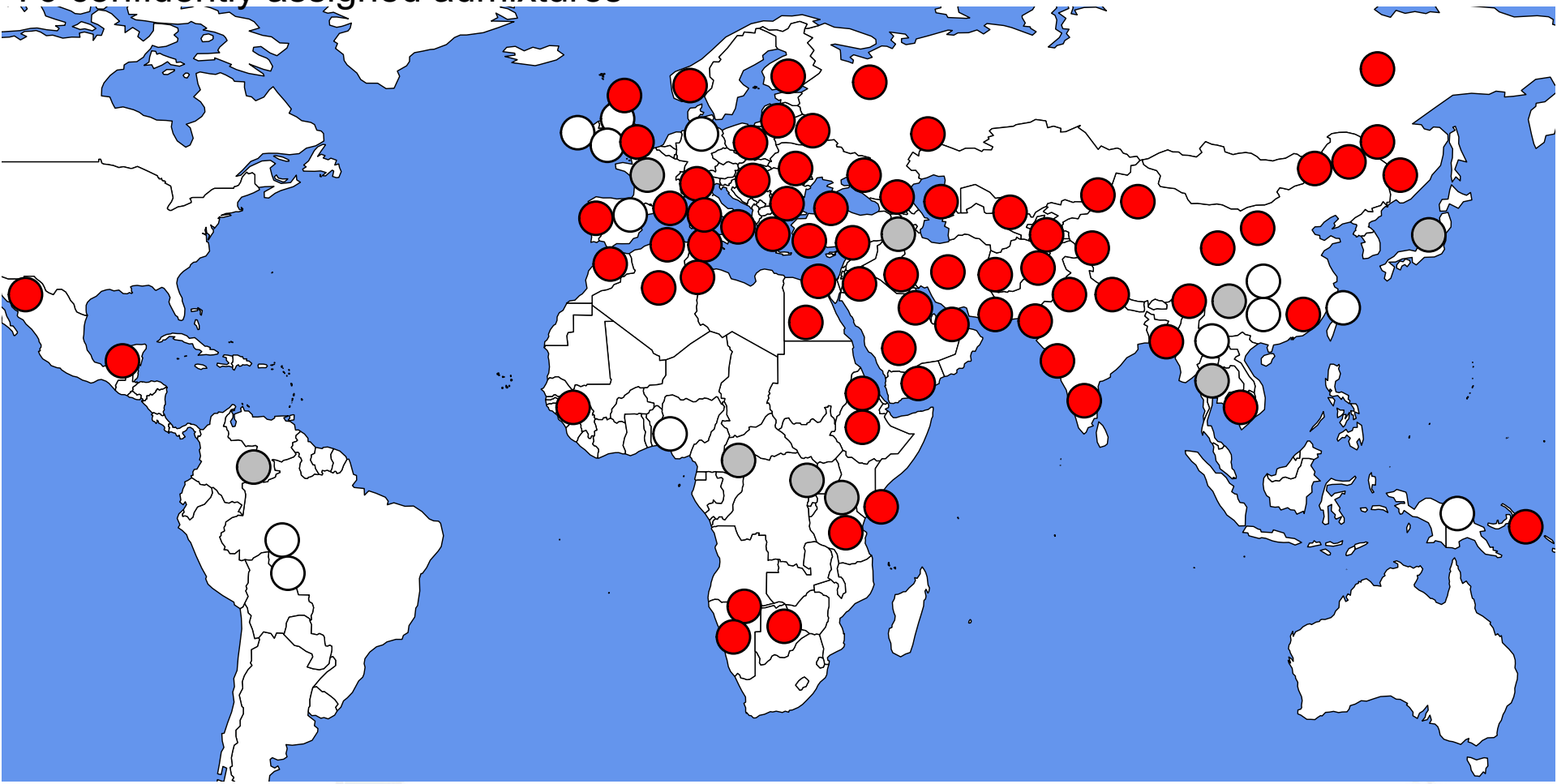
93 human populations



Hellenthal et al, *Science*, 2014
Image credit G Hellenthal

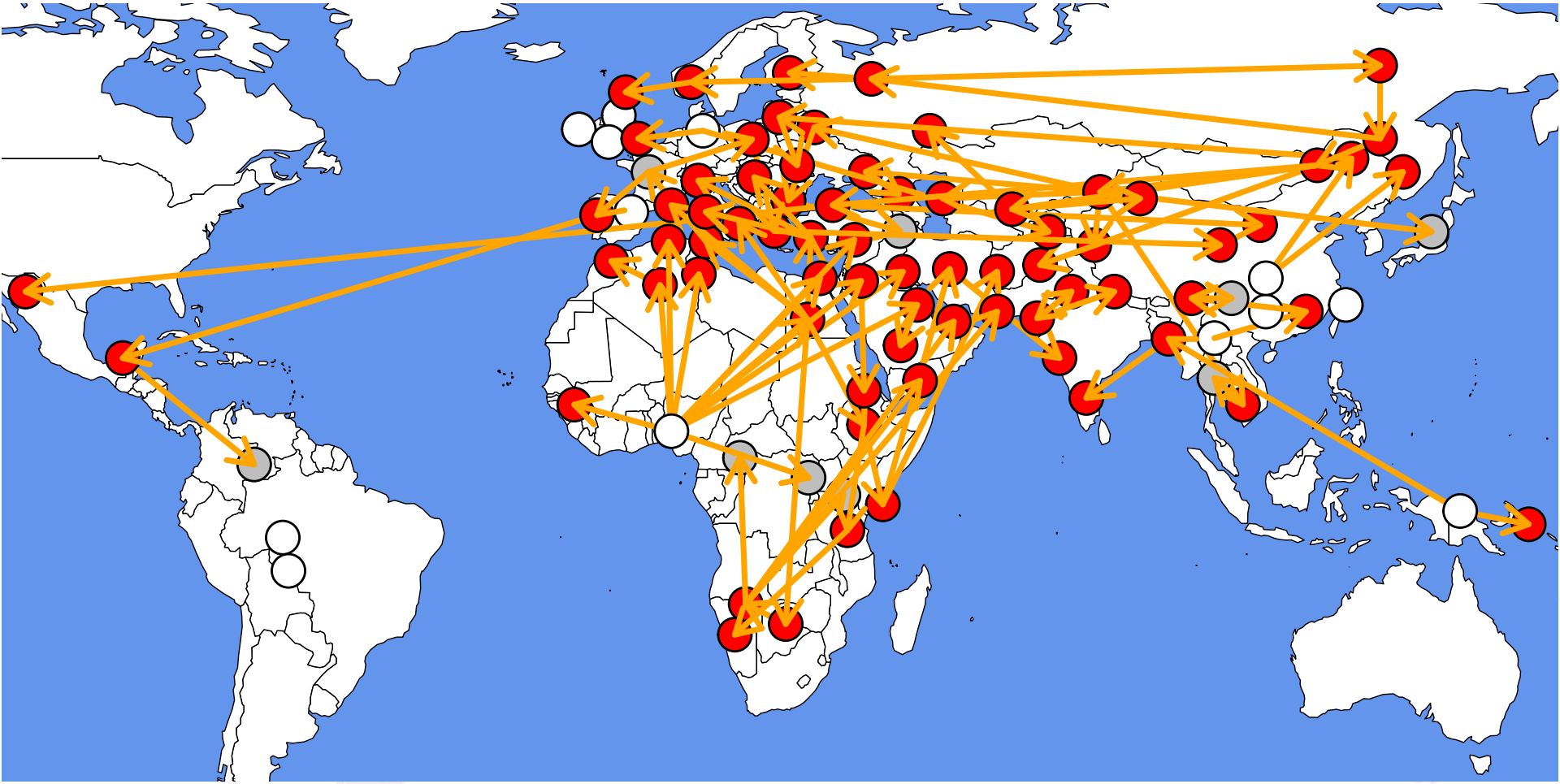
Genetic mixing over the past 4000 years

76 confidently assigned admixtures



Hellenthal et al, *Science*, 2014
Image credit G Hellenthal

Genetic mixing over the past 4000 years



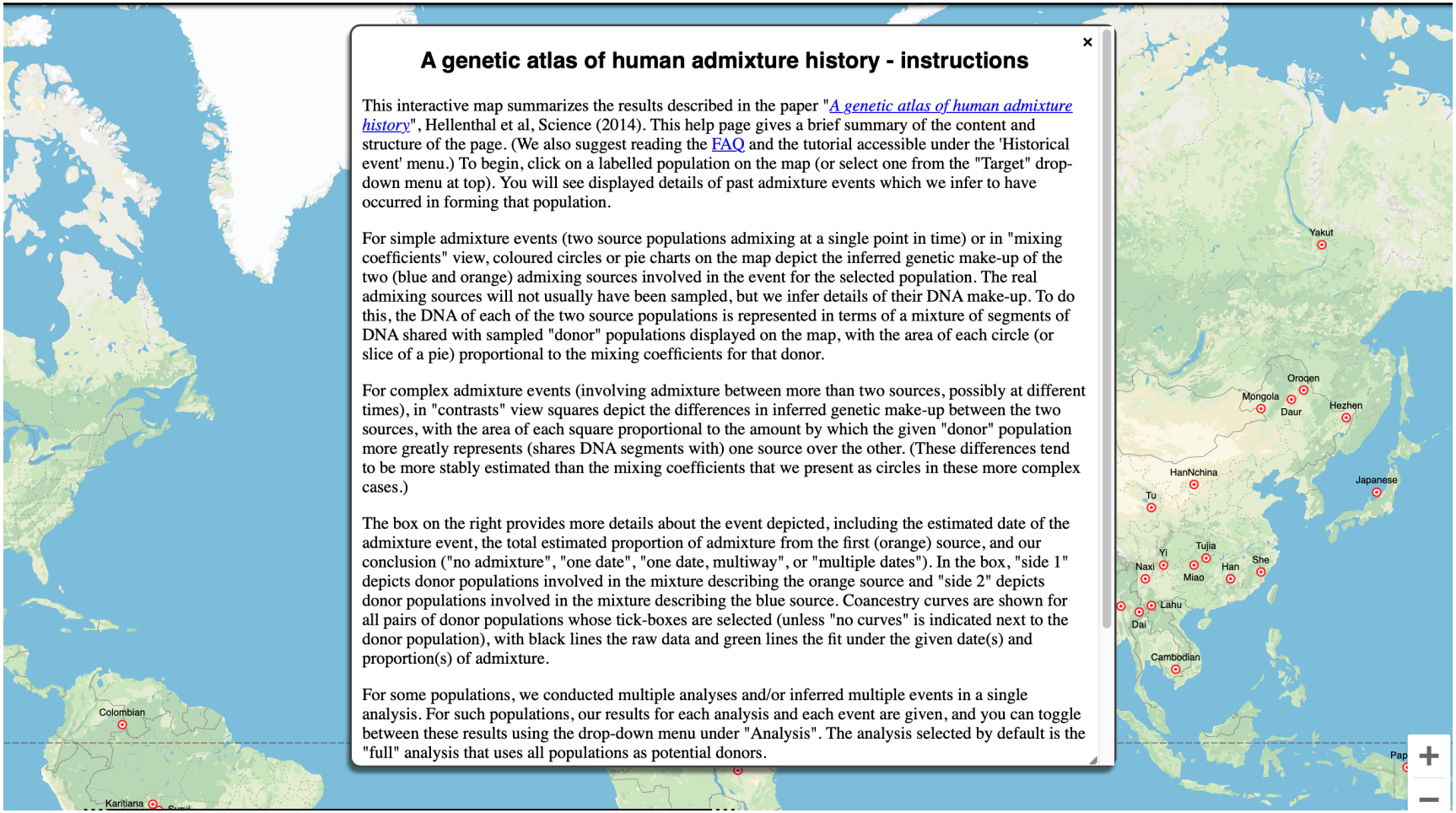
Hellenthal et al, *Science*, 2014
Image credit G Hellenthal

http://admixturemap.paintmychromosomes.com

A genetic atlas of human admixture history

Companion website for "*A genetic atlas of human admixture history*", Hellenthal et al, Science (2014).

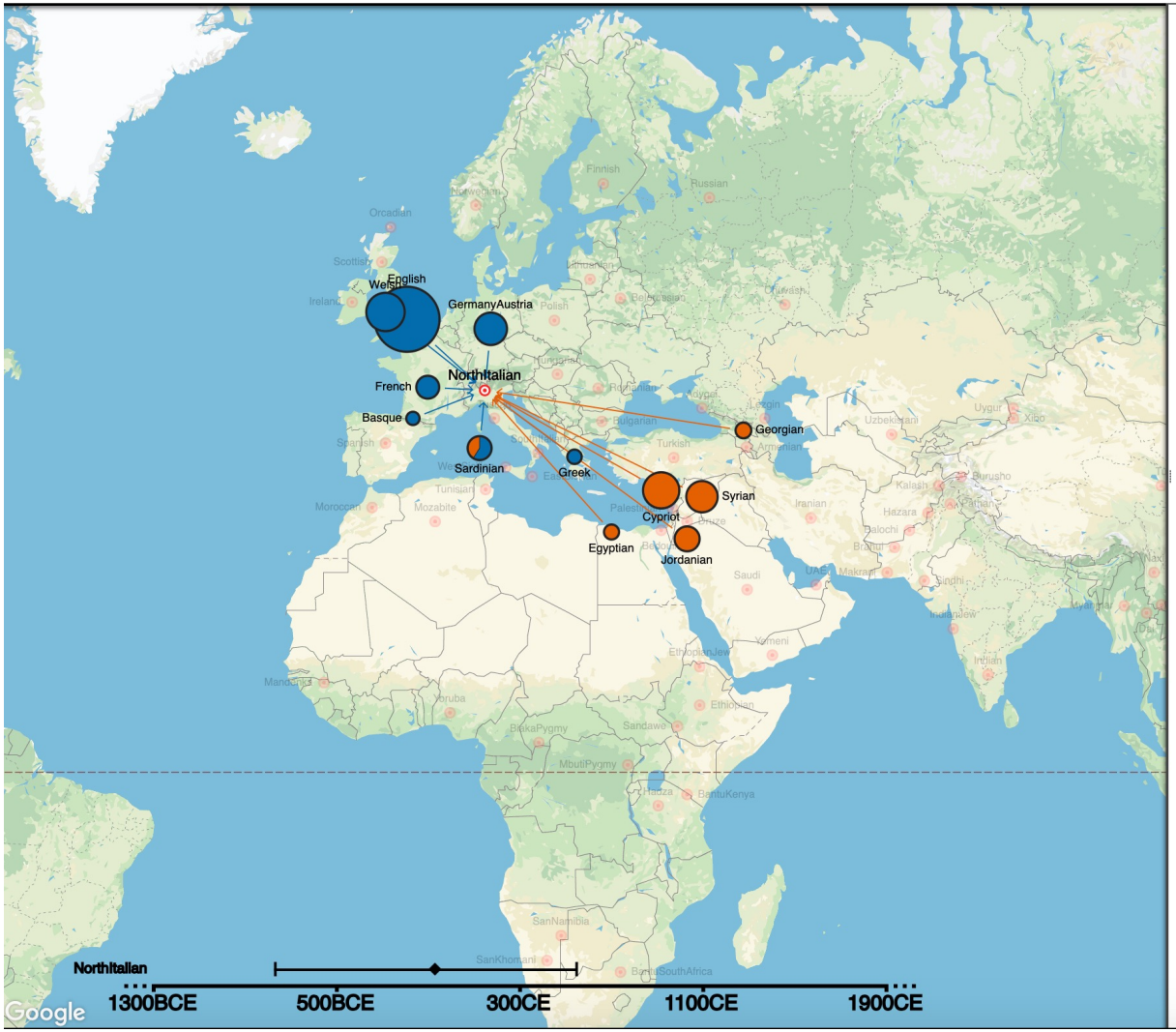
Historical event ↓ Target population ↓ Data ↓ Preferences ↓ Help and FAQ ↓



A genetic atlas of human admixture history

Companion website for "*A genetic atlas of human admixture history*", Hellenthal et al, Science (2014).

Historical event ↓ Target population ↓ Data ↓ Preferences ↓ Help and FAQ ↓



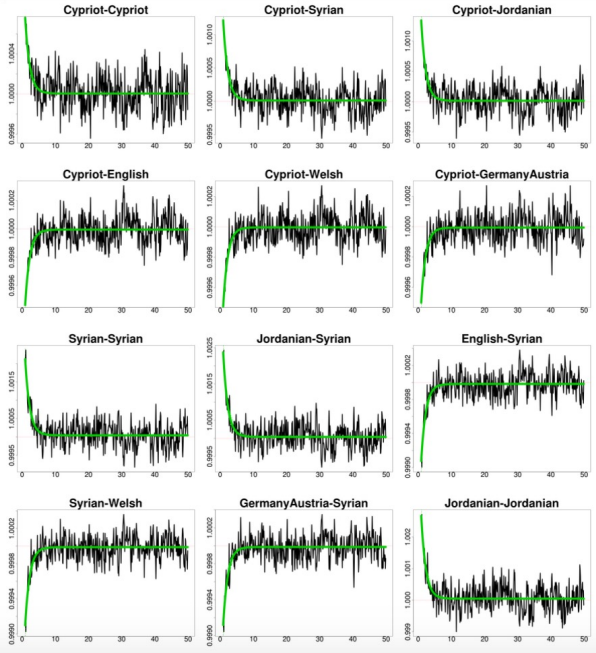
NorthItalian

Analysis[?]:	FullAnalysis[?]
Number of individuals[?]:	12
Conclusion[?]:	**One date**[?]
Estimated date (95% CI)[?]:	66BCE (766BCE - 550CE)[?]
Estimated proportion[?]:	0.33
View:[?]	mixing coefficients

(The area of each circle or pie segment reflects that donor population's mixing coefficient for one admixing source, coloured orange or blue.)

▶ show details of model fit

Side 1 (Cypriot-like)	Side 2 (Welsh-like)
☑ Cypriot (10.8%)	☑ English (36.2%)
☑ Syrian (7.9%)	☑ Welsh (12.1%)
☑ Jordanian (4.9%)	☑ GermanyAustria (8.6%)
☐ Sardinian (1.8%)	☐ French (4.1%)
☐ Georgian (1.8%)	☐ Sardinian (2.6%)
☐ Egyptian (1.6%)	☐ Greek (1.5%)



Summary

Many programs for inferring and determining population structure – central to understanding demographic history

PCA and model-based clustering make use of unlinked allele frequencies. Fast and tractable, though interpretation can be challenging.

Haplotype-based methods – a toolkit for exploring local ancestry, LD, recombination and subsequently fine-scale patterns of population structure at different scales.

Chromopainter and fineSTRUCTURE as tools for resolving subtle differences between populations.

Admixture LD can be utilized together with chromosome painting to identify when admixture has occurred, its sources and its timings.

Such analyses are highlighting that admixture is common place within human populations over the past 4000 years.

Suggested Reading/Software

Overview of methods in PopGen:
Schraiber & Akey. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*. (2015) 6(12):727-40.

How not to overinterpret ADMIXTURE plots:
Lawson, van Dorp & Falush. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications.* (2018) 9,3258.

Chromopainter/fineSTRUCTURE:
Lawson, Hellenthal, Myers & Falush. Inference of Population Structure using Dense Haplotype Data. PLoS Genetics. (2012) 8(1): e1002453.
https://people.maths.bris.ac.uk/~madjl/finestructure-old/chromopainter_info.html

Fine-scale structure of the British Isles:
Lesie, Winney & Hellenthal et al. The fine-scale genetic structure of the British population. Nature. (201) 519,309-314.

GLOBETROTTER:
Hellenthal et al. A genetic atlas of human admixture history. Science. 14; 343(6712):747-751.
https://people.maths.bris.ac.uk/~madjl/finestructure/globetrotter.html

Thank you to Lucy van Dorp and Garrett Hellenthal!