



Morning

Inferring population structure and admixture histories

Afternoon

Genome-wide genealogies: 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

Lunch

13:30-14:15: Lecture

14:15-14:45: Coffee

14:45-16:30: Relate practical

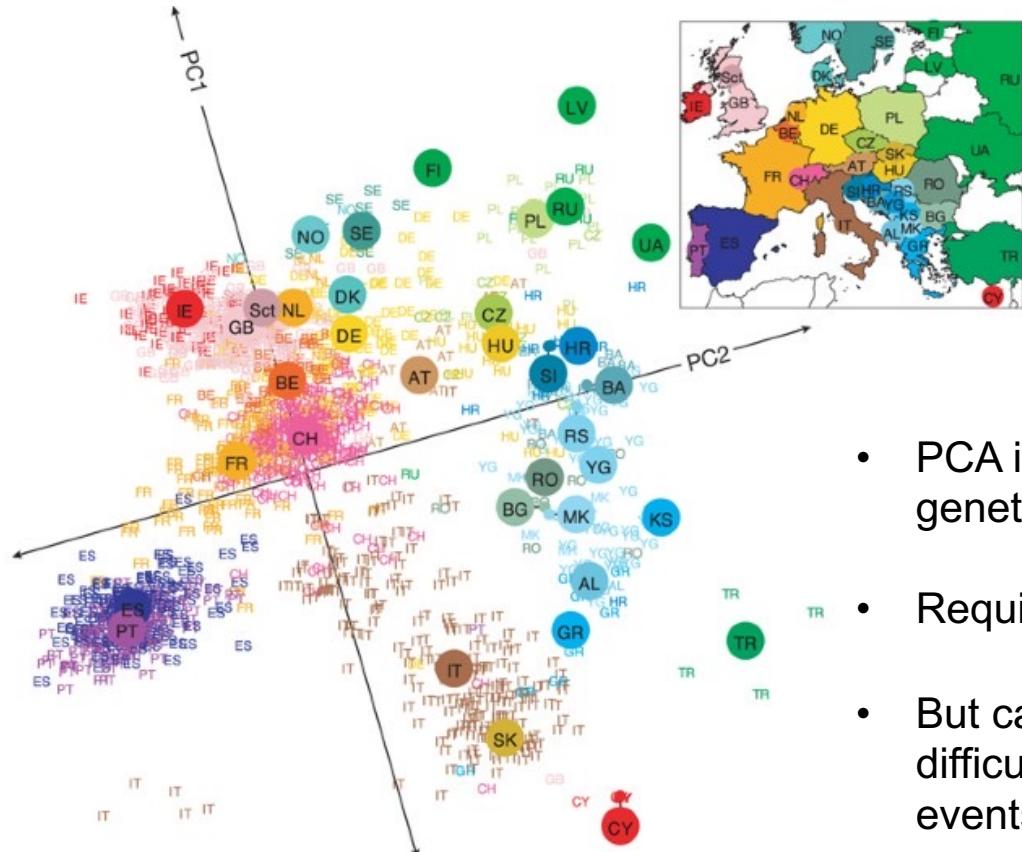


What does genetic diversity within & across (human) groups look like?

- Genome-wide patterns of variation across individuals provide a powerful source of data
- Genetic diversity is underpinned by the key ancestral processes
 - 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.

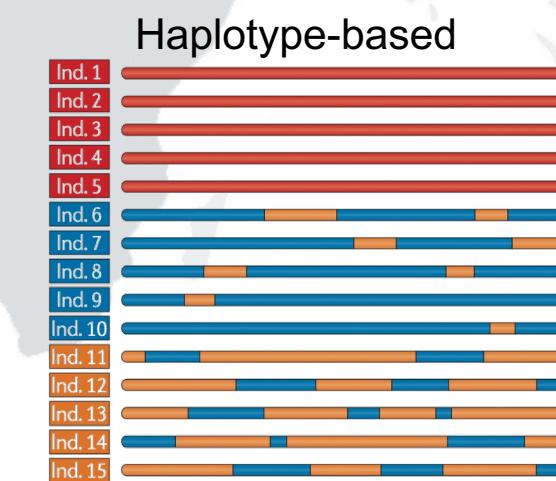
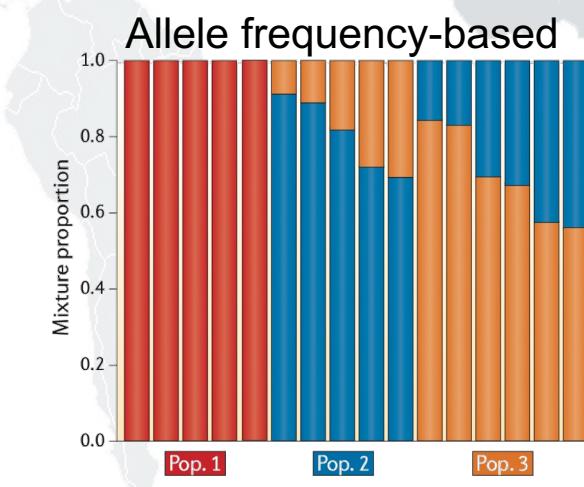
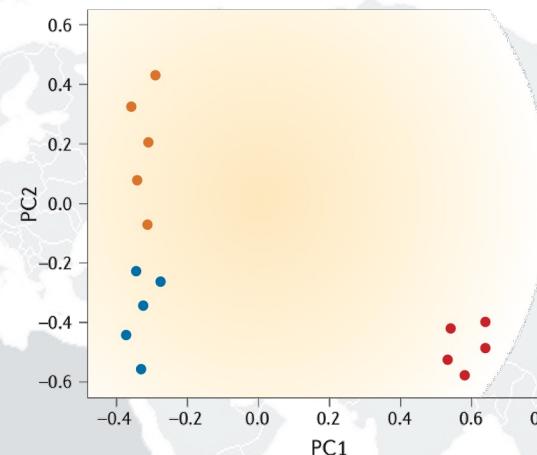
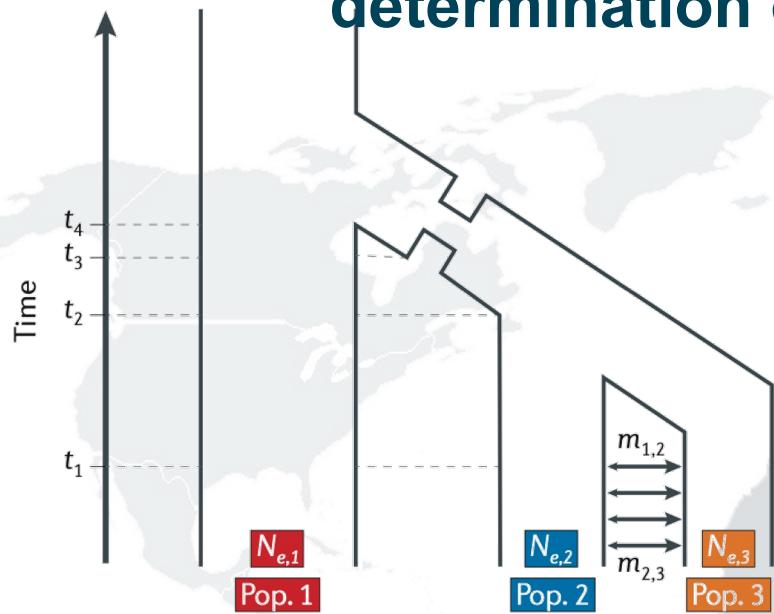
Genetic structure: spotting patterns in genetic variation

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

Different approaches for determination of population structure



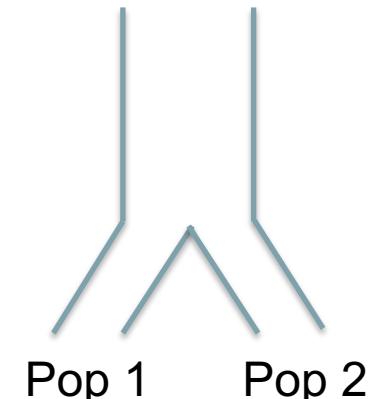
Different approaches for determination of population structure

Name	Data type	Inference	Notes	Ref
STRUCTURE	Unlinked multi-allelic genotypes	Population structure, admixture	User friendly GUI; can be computationally demanding	Pritchard, Stephens & Donnelly. 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.

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?

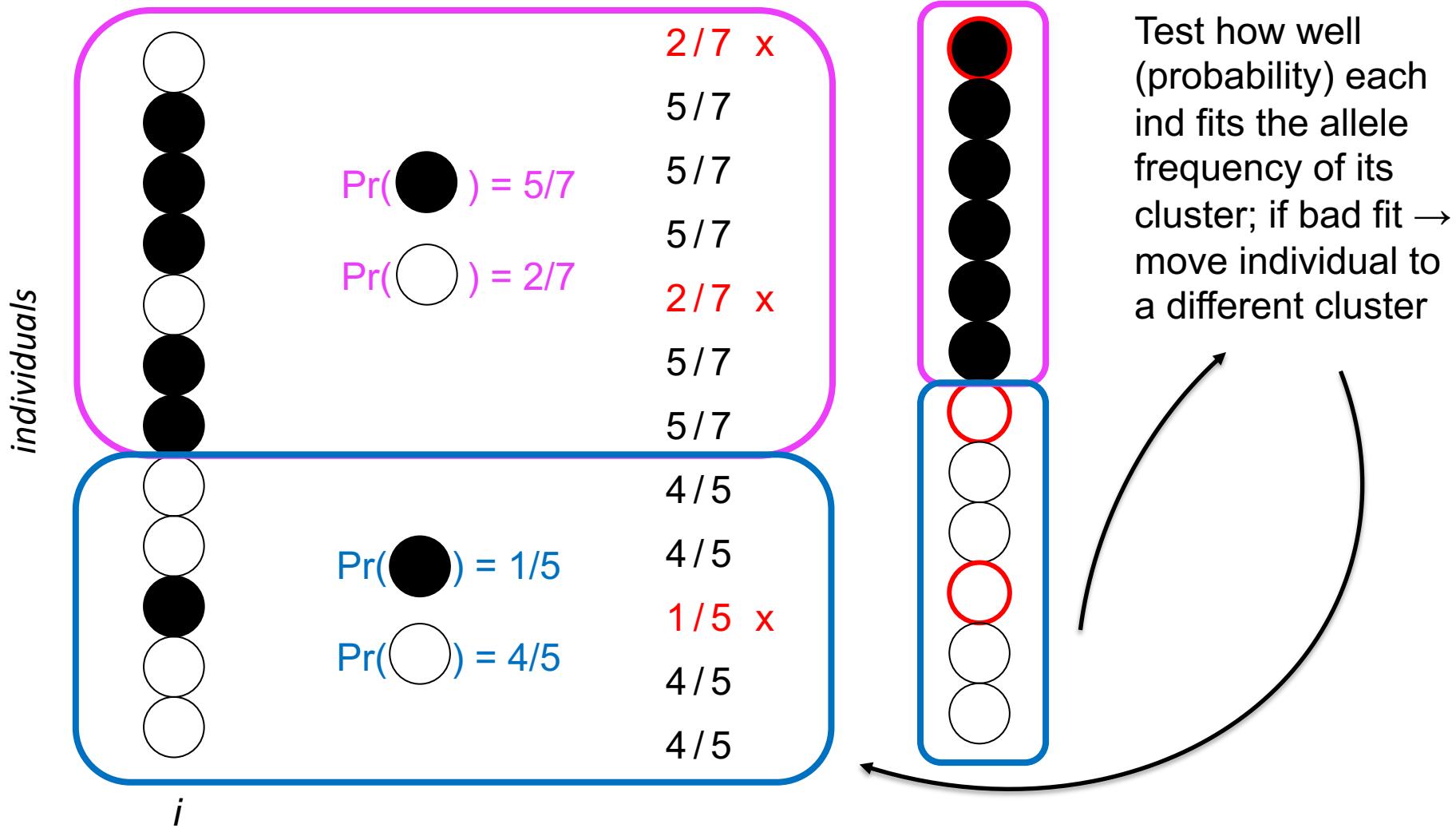


	Allele frequency in Pop1	Pop2	Genotype of individual	
SNP1	0.8	0.4	1	→ Pop1?
SNP2	0.3	0.7	0	→ Pop1?
SNP3	0.4	0.6	1	→ Pop2? (Pop1?)
SNP4	0.9	0.1	1	→ Pop1?

Allele frequency-based clustering

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

One locus:

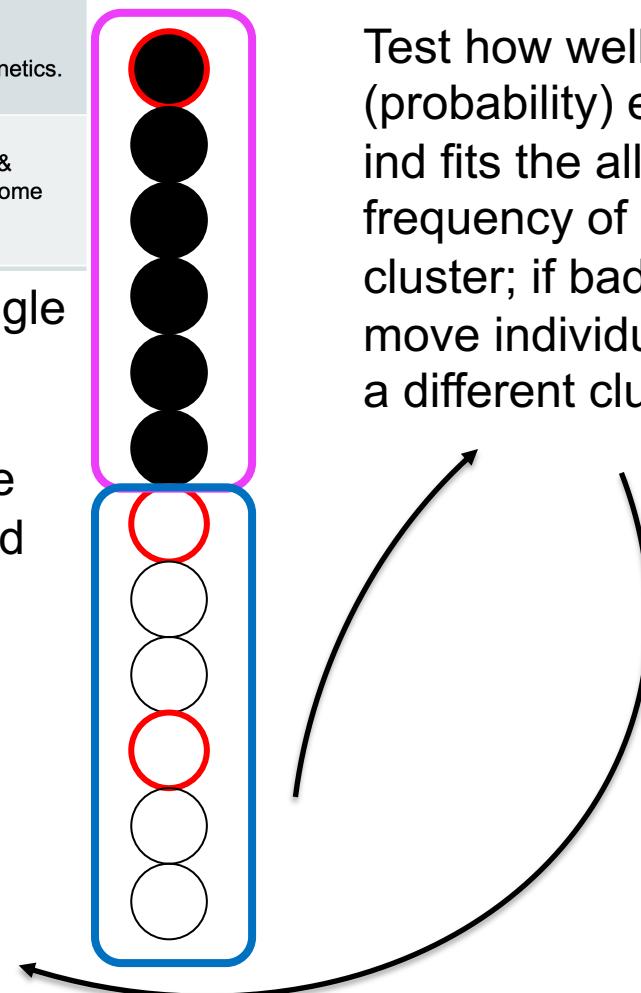


LD prune

Allele frequency-based clustering

Name	Data type	Inference	Notes	Ref
STRUCTURE	Unlinked multi-allelic genotypes	Population structure, admixture	User friendly GUI; can be computationally demanding	Pritchard, Stephens & Donnelly. 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.

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

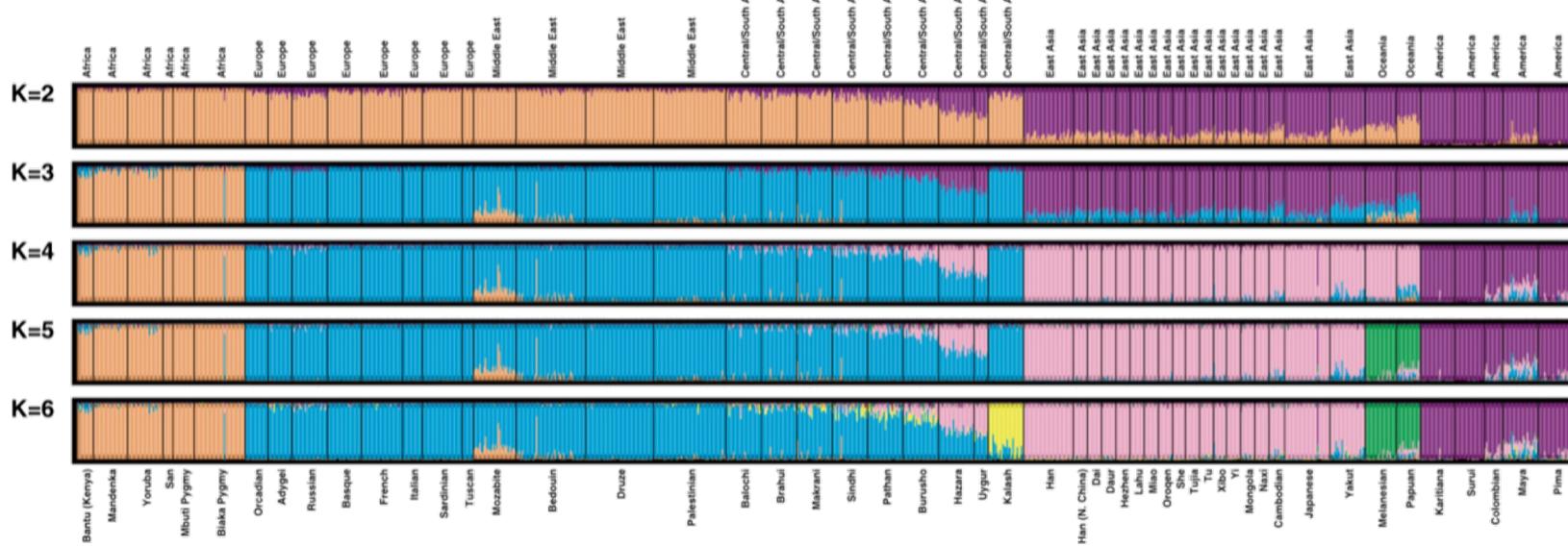


- “no admixture model” – assign each ind i to single cluster k
- “admixture model” – assign each ind to multiple clusters (i.e. infer % of ind i ’s genome assigned to clusters $1, \dots, K$)
- “linkage model” – can identify regions of ind i assigned to each cluster (Falush et al 2003, *Genetics* **164**:1567)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

Allele frequency-based clustering

Human Genome Diversity Panel

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

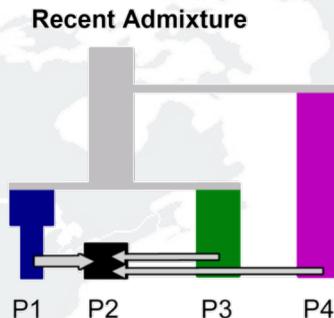


- Accuracy of population identification depends on the number of individuals included, the number of loci used and the allele frequency differences among those populations.
- Population identification can be strongly affected by the amount of admixture within individuals and the proportion of admixed individuals in the dataset – all populations are admixed....

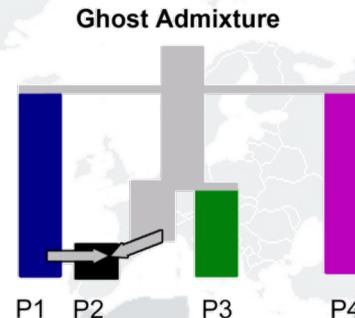
Allele frequency-based clustering

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

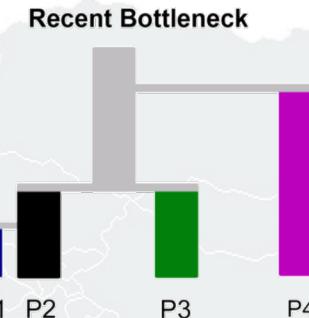
a



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

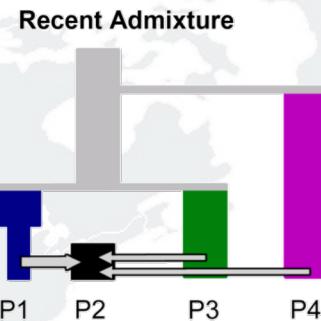
Taken from :

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

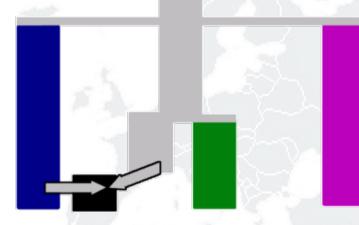
Allele frequency-based clustering

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

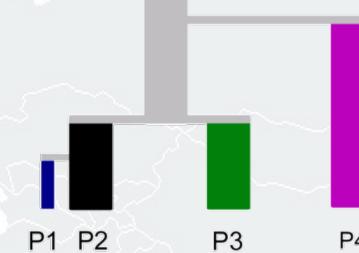
a



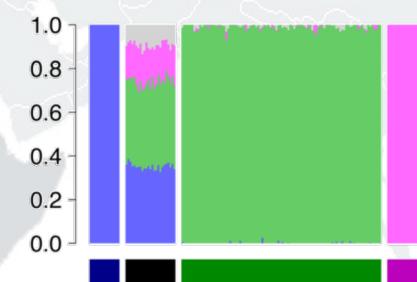
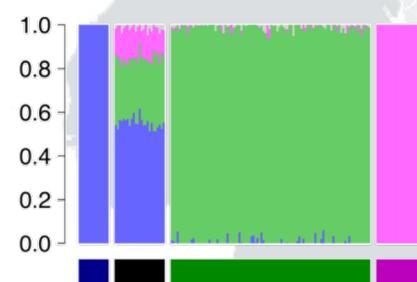
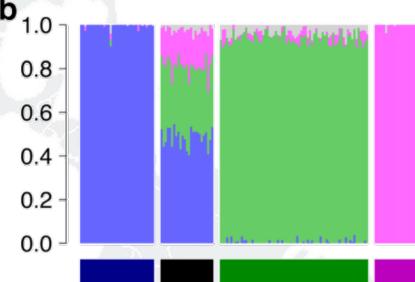
Ghost Admixture



Recent Bottleneck



b

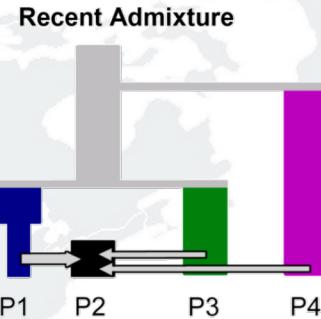


Lawson, van Dorp & Falush. *Nat Comms.* (2018).

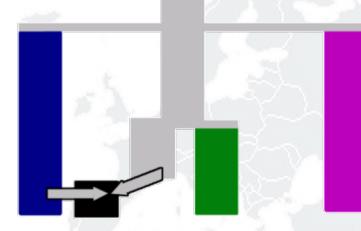
Allele frequency-based clustering

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

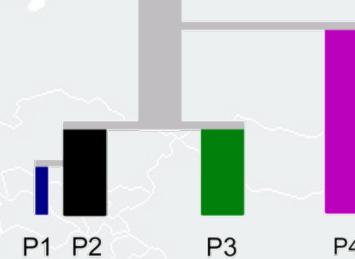
a



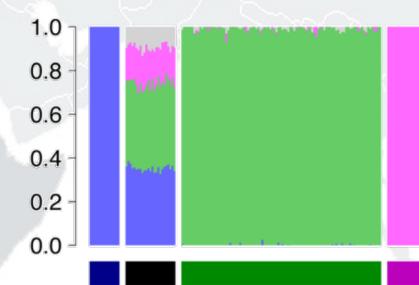
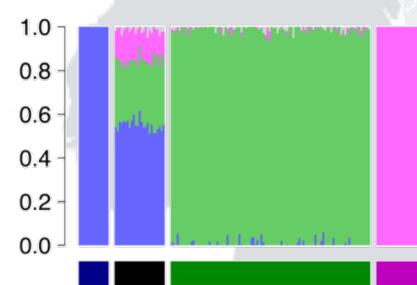
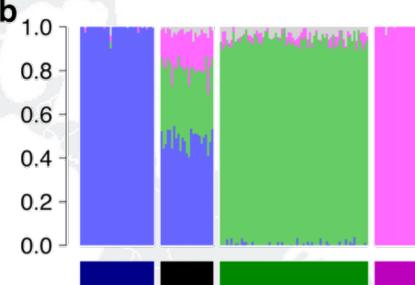
Ghost Admixture



Recent Bottleneck



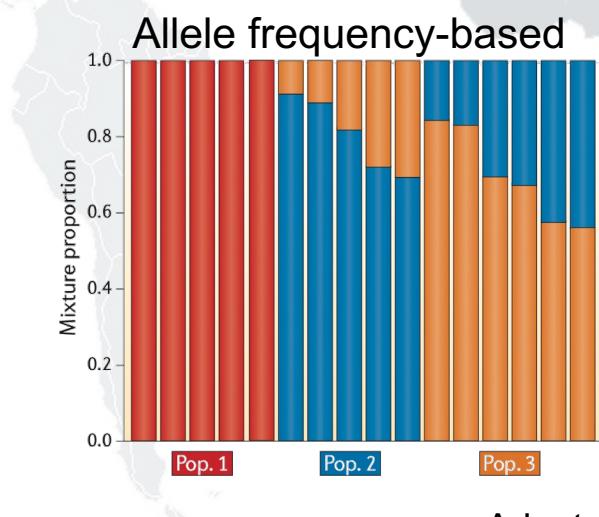
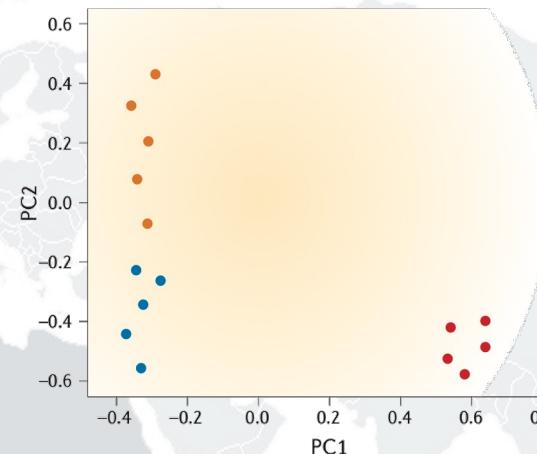
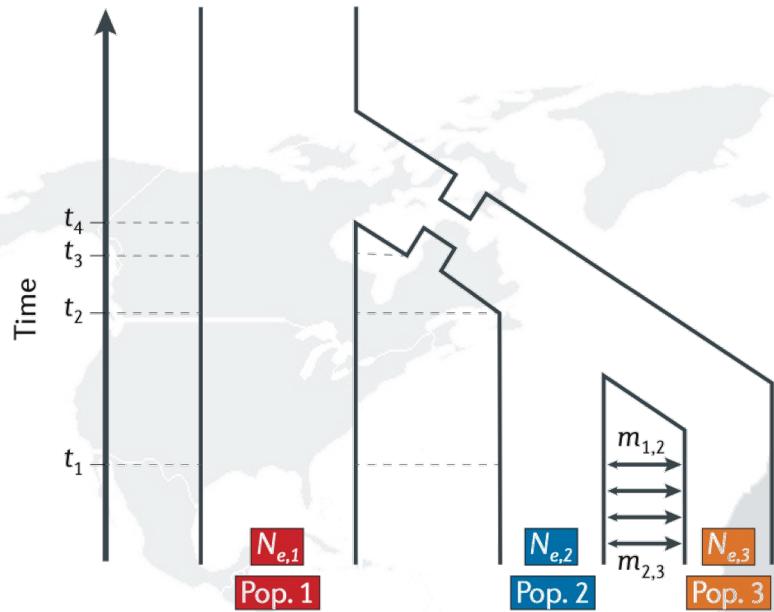
b



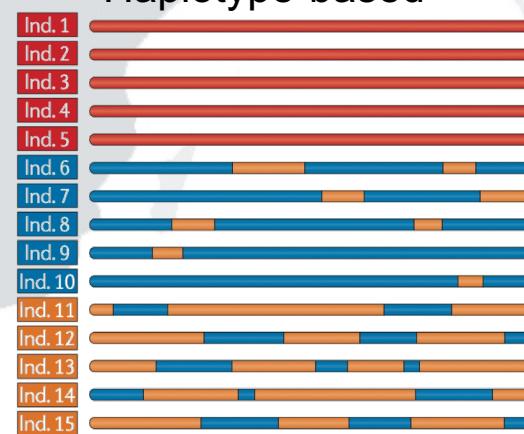
Lawson, van Dorp & Falush. *Nat Comms.* (2018).

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

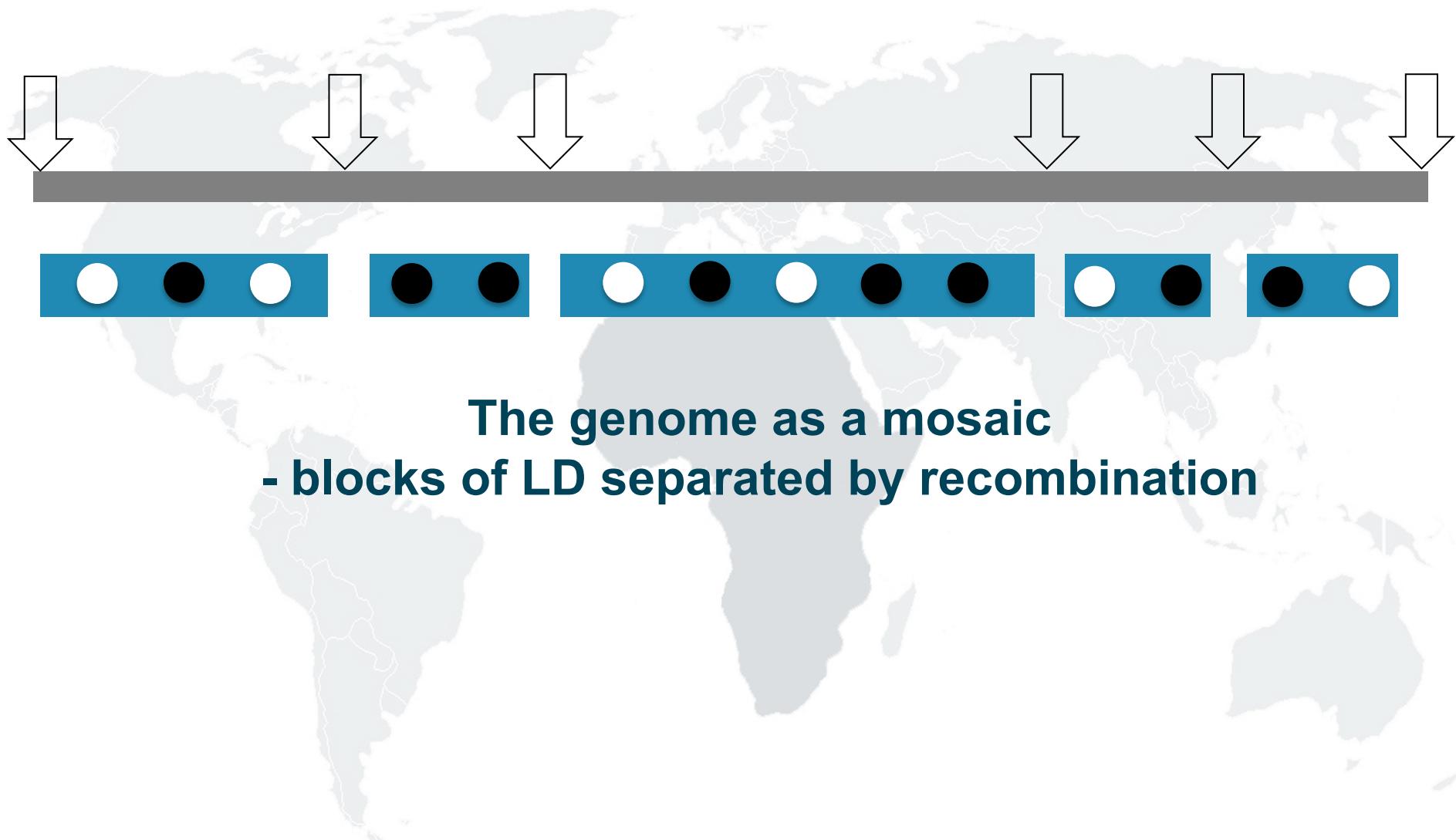


Haplotype-based



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

Incorporating haplotype information

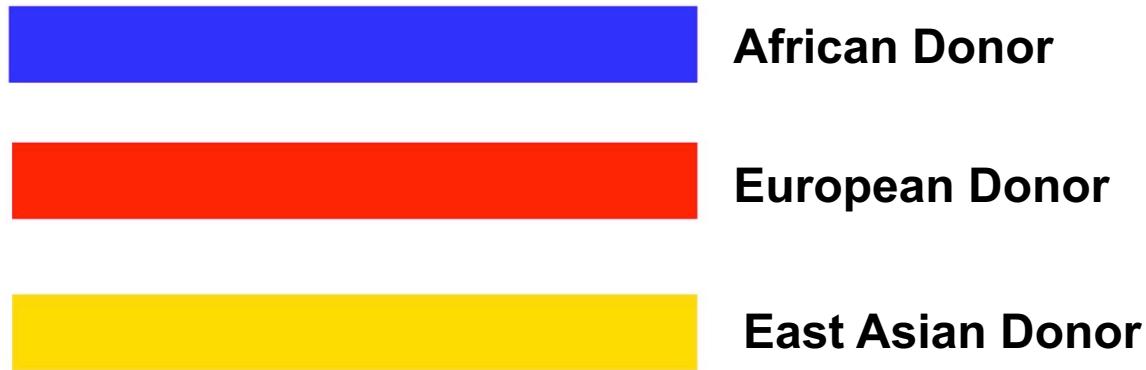


Incorporating haplotype information

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.

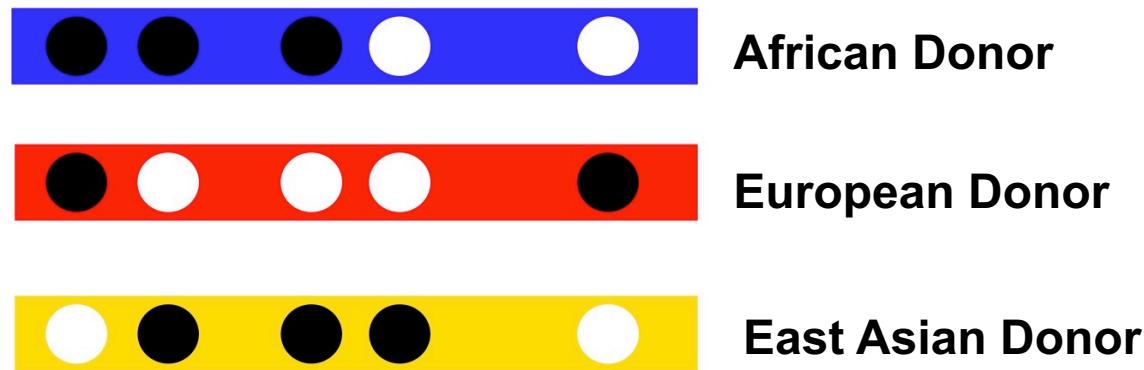
Incorporating haplotype information

Chromosome painting



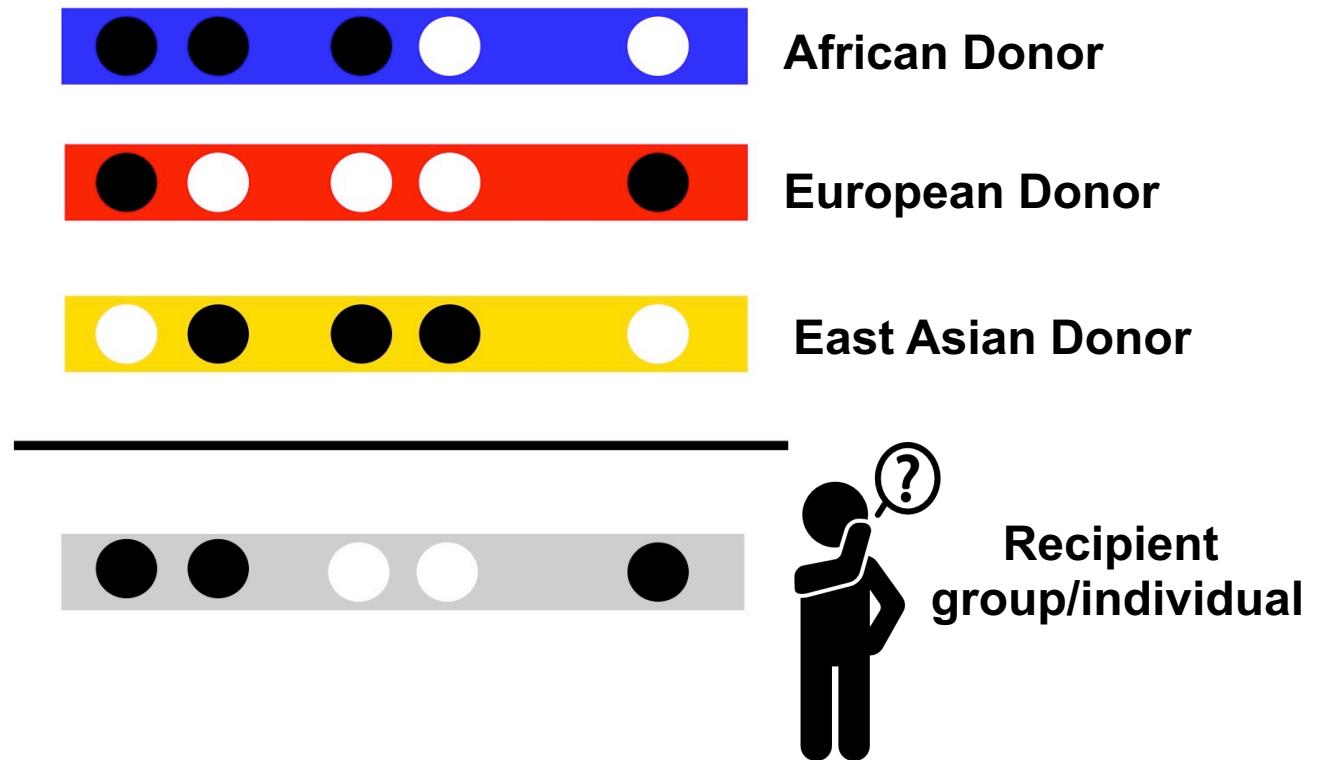
Incorporating haplotype information

Chromosome painting



Incorporating haplotype information

Chromosome painting

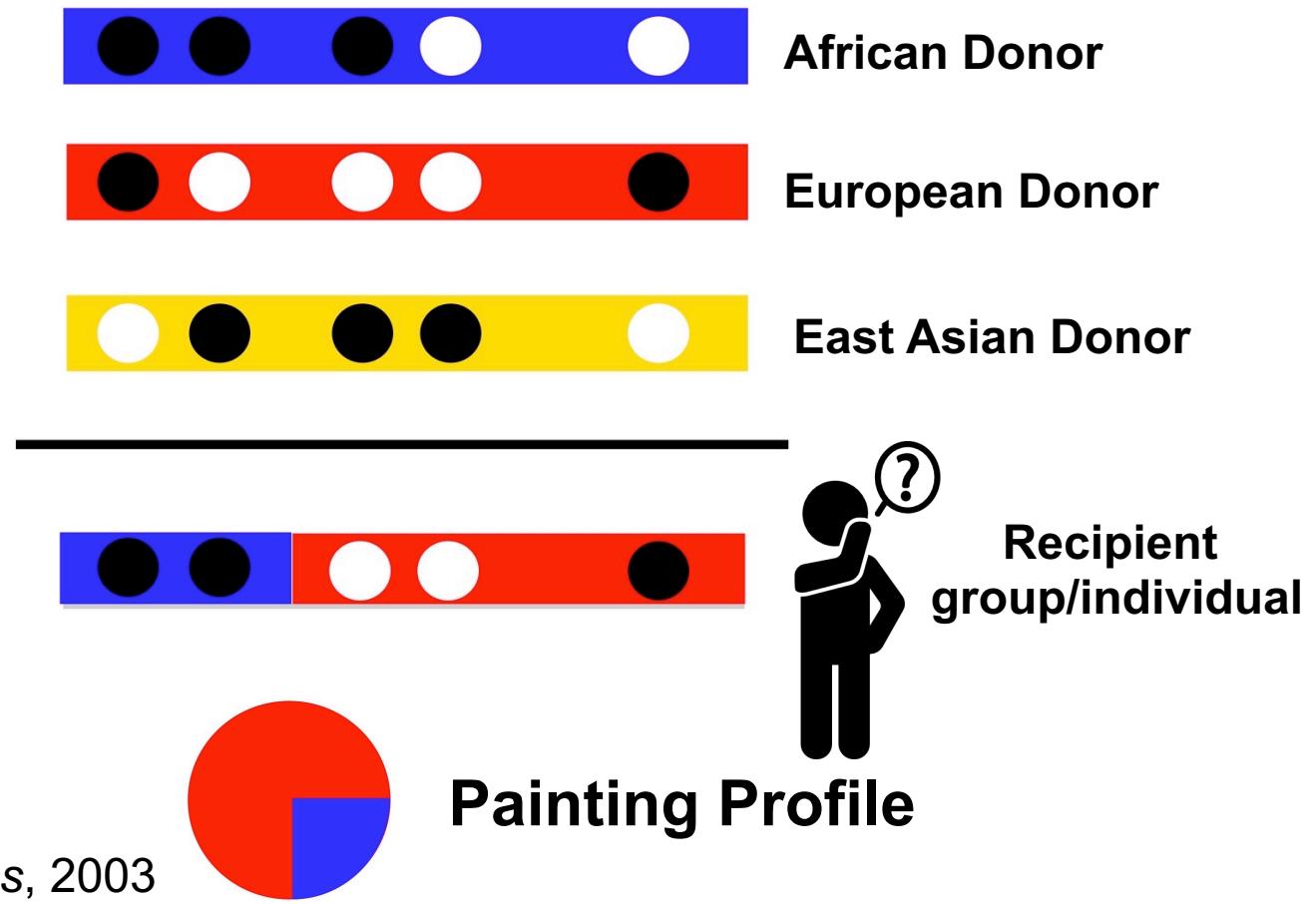


Li & Stephens, *Genetics*, 2003

Lawson et al, *PLoS Genetics*, 2012

Incorporating haplotype information

Chromosome painting

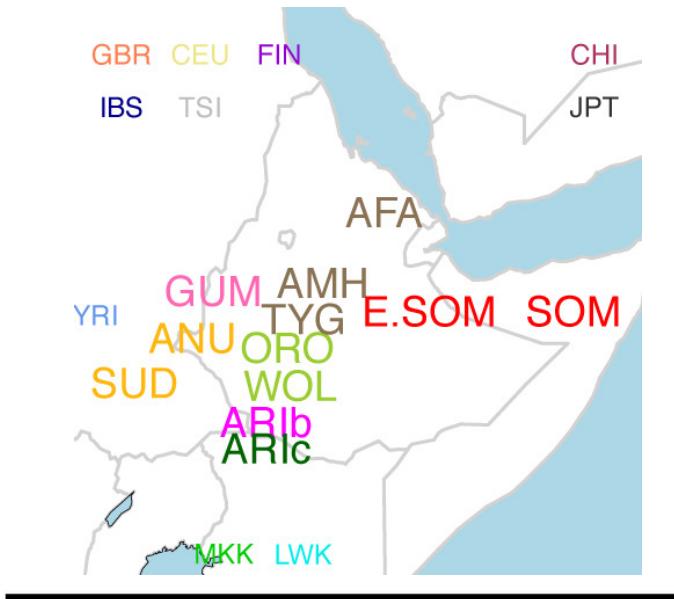


Li & Stephens, *Genetics*, 2003

Lawson et al, *PLoS Genetics*, 2012



Incorporating haplotype information Chromosome painting

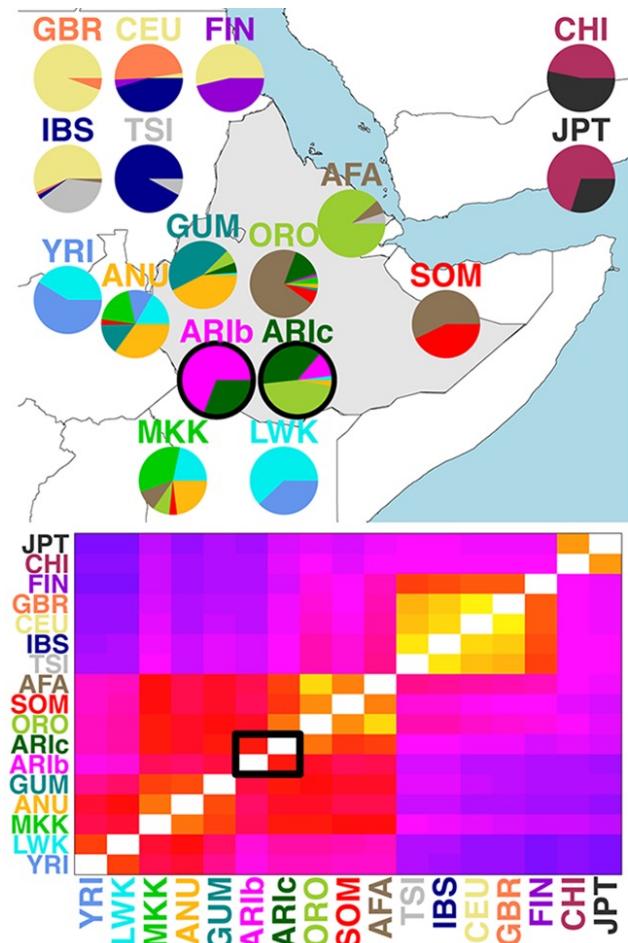


All-donors



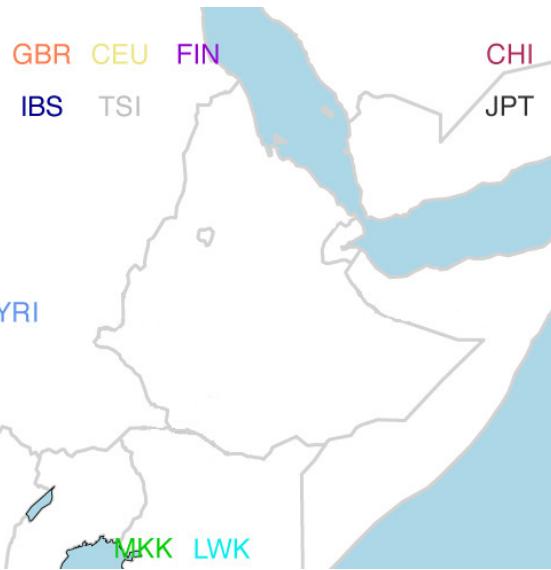
**ARI blacksmith/
ARI cultivators**

Average haplotype length shared = 2.76cM



Incorporating haplotype information

Chromosome painting



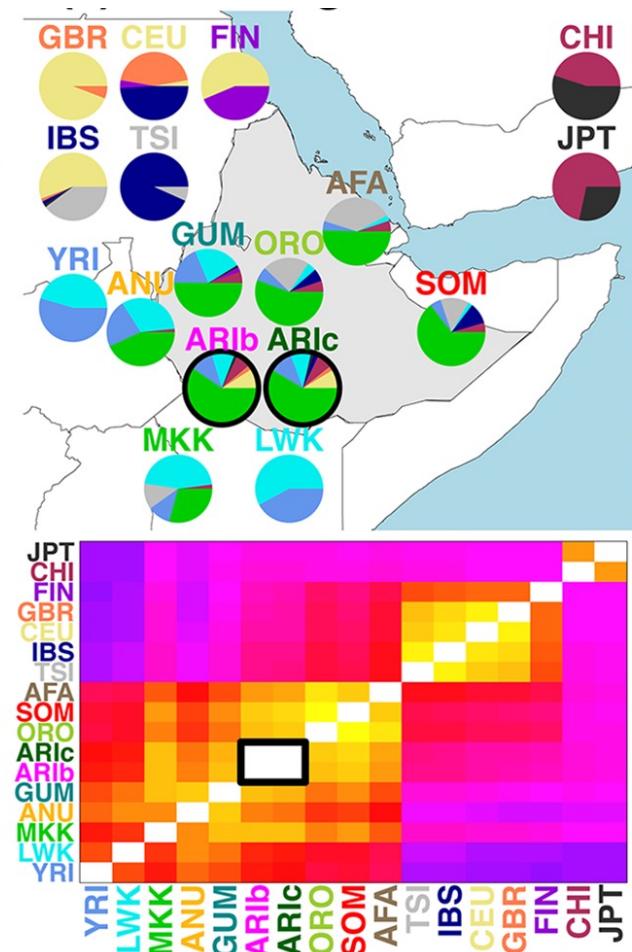
No Ethiopia



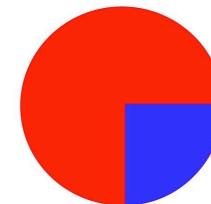
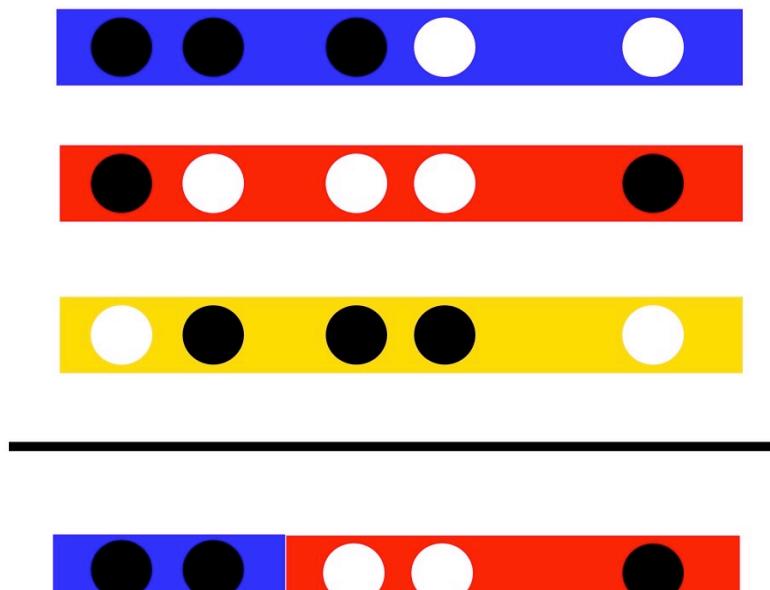
ARI blacksmith/
ARI cultivators

Average haplotype length shared = 0.65cM

van Dorp et al, *PLoS Genetics*, 2015



Haplotype clustering



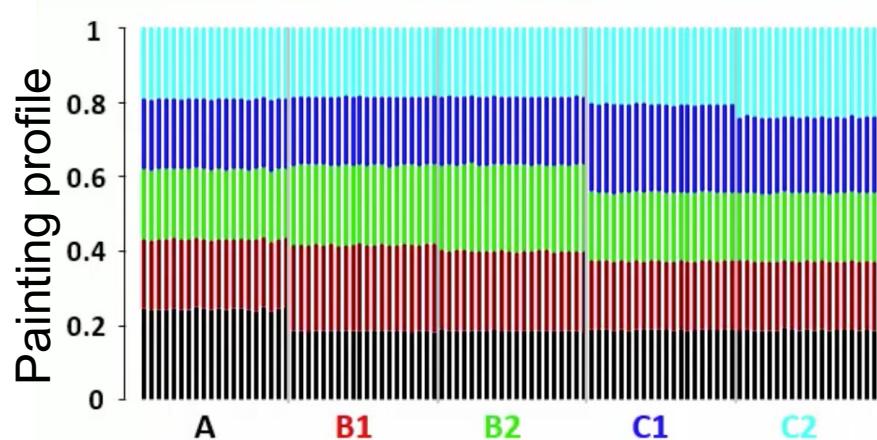
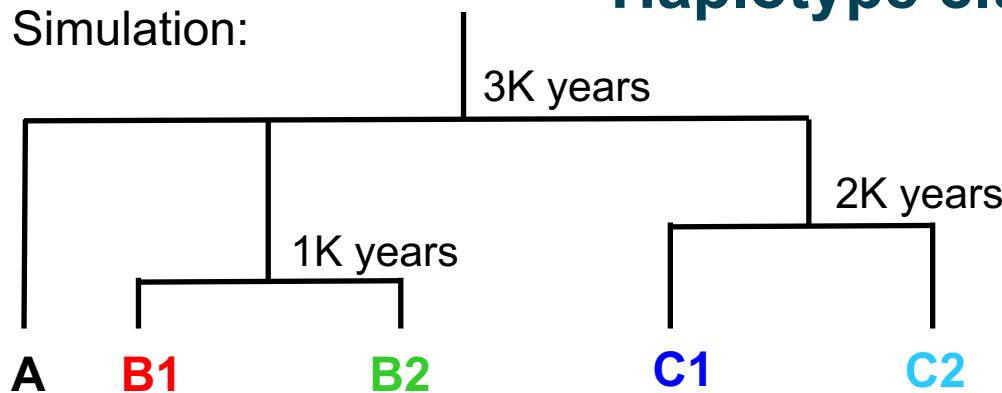
Clustering
fineSTRUCTURE



Painting Profile

Haplotype clustering

Simulation:

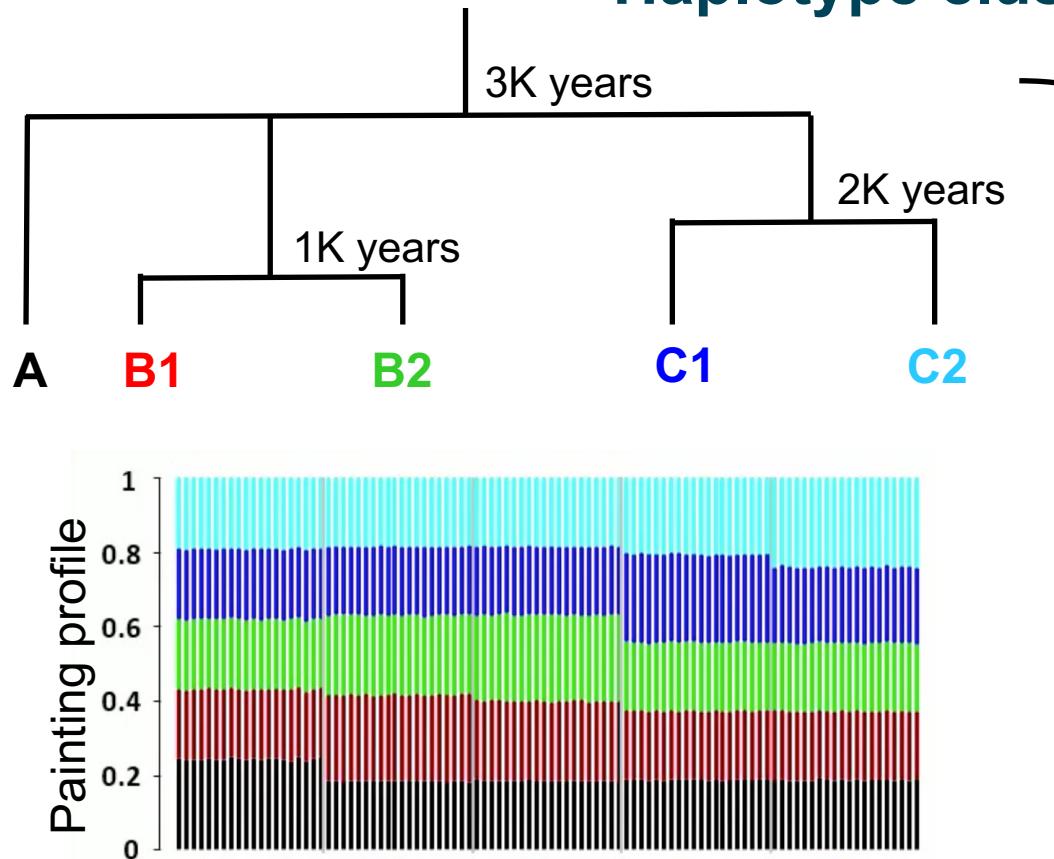


Donor individual = j

Recipient individual = i

Number of chunks of DNA by which individual i is painted by individual j = Y_{ij}

Haplotype clustering



Assign individuals to a random set of clusters k

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

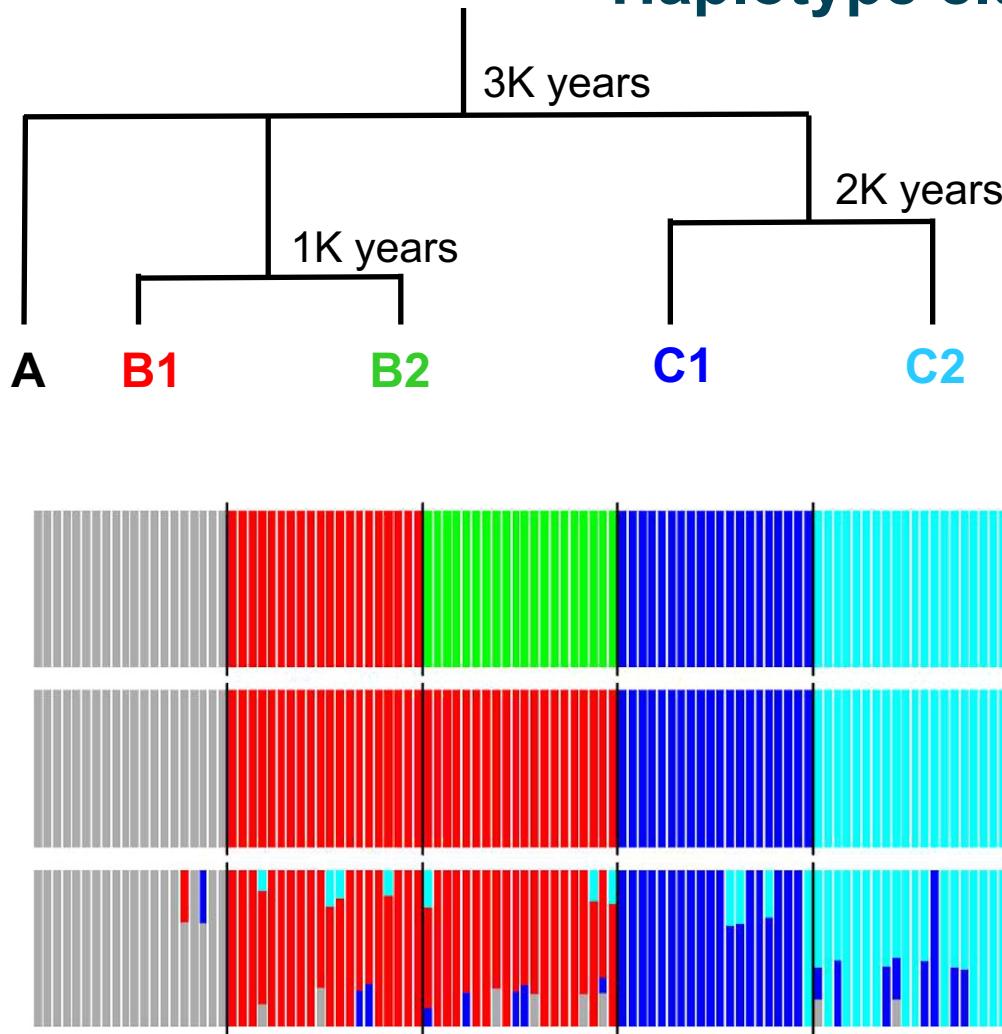
1. Infer P_{Ak} – number of chunks individuals in cluster A are painted with individuals in cluster k
2. Test if $\Pr(Y_{i1}, \dots, Y_{ik})$ is low – if so change cluster and repeat!

Donor individual = j

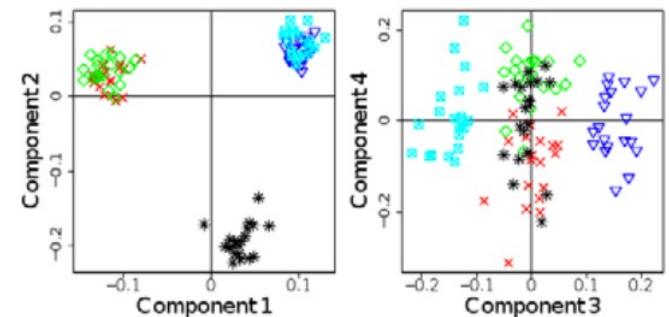
Recipient individual = i

Number of chunks of DNA by which individual I is painted by individual j = Y_{ij}

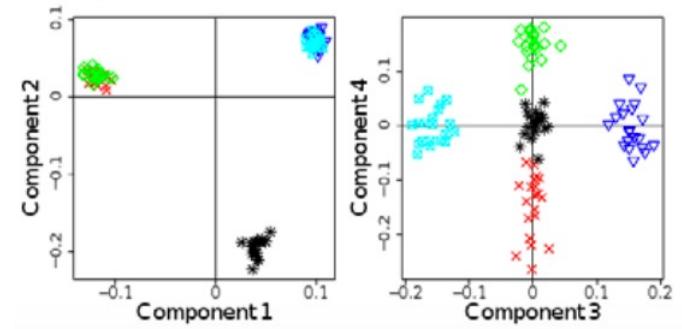
Haplotype clustering



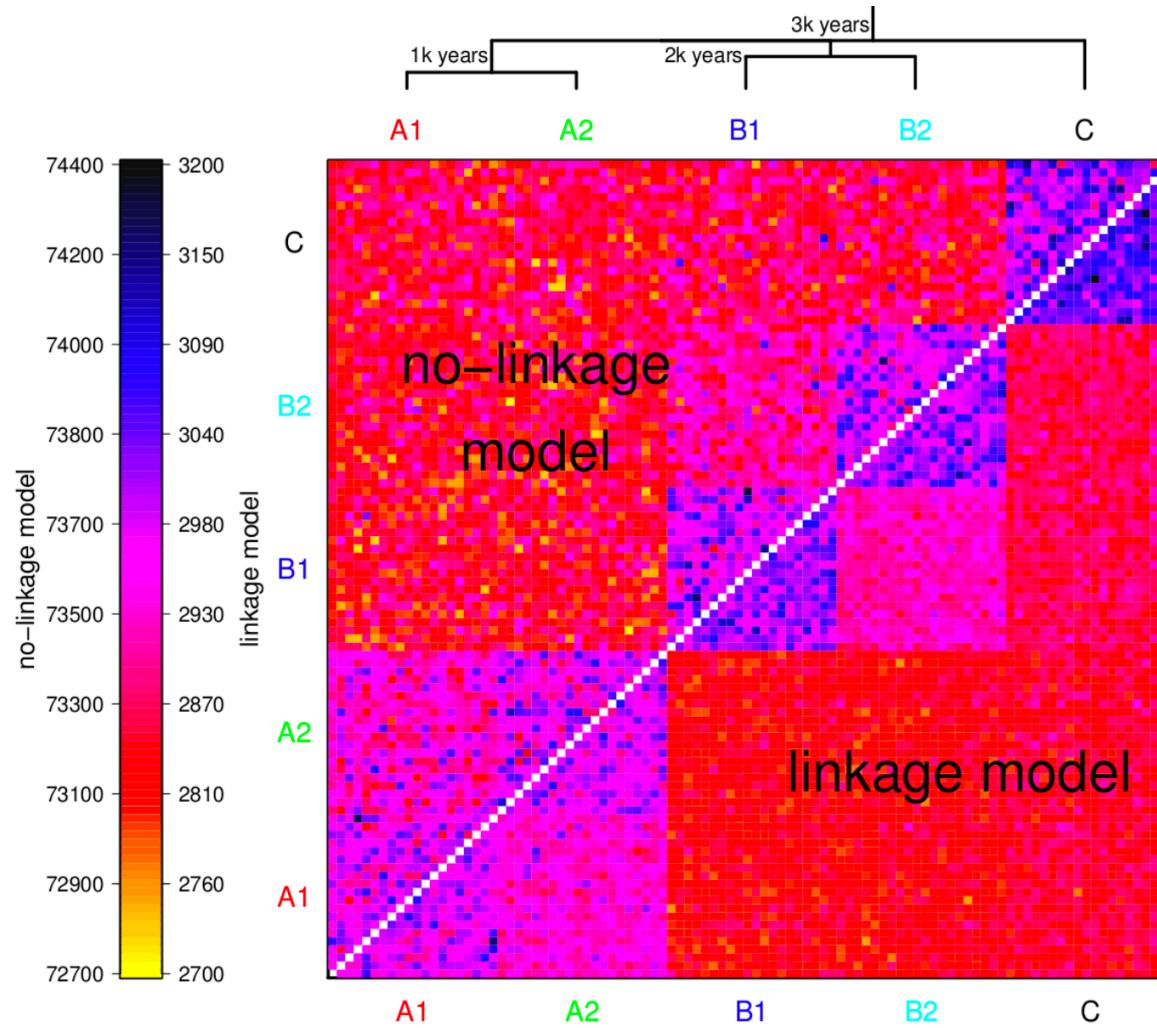
E) Unlinked model PCA



F) Linked model PCA

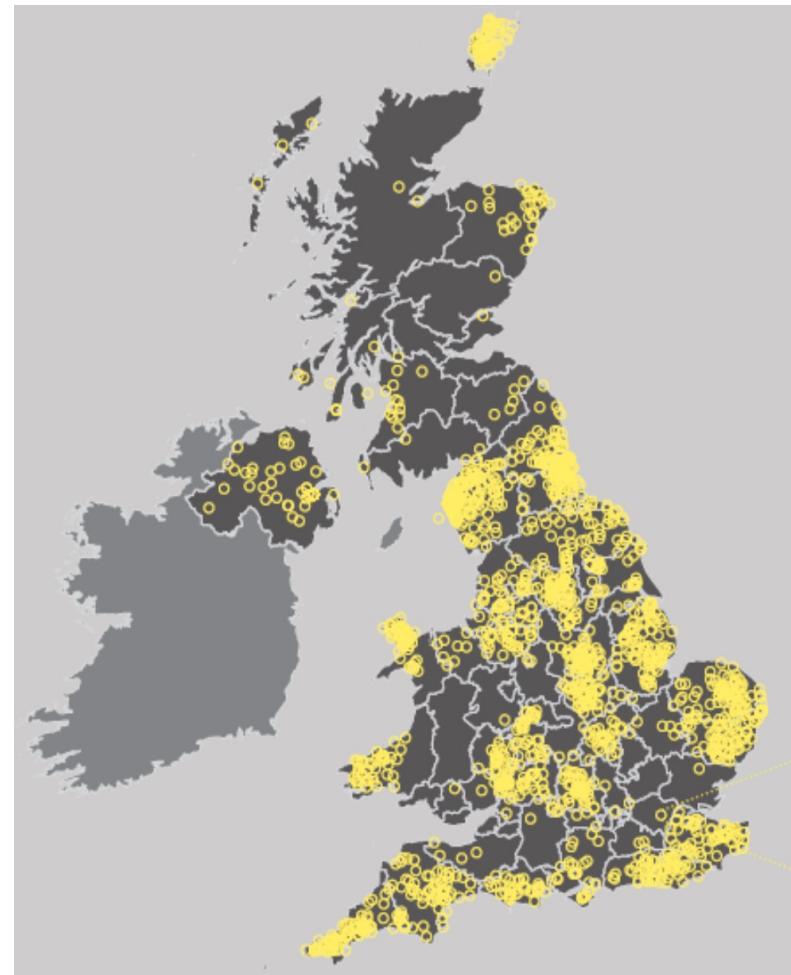


Haplotype clustering

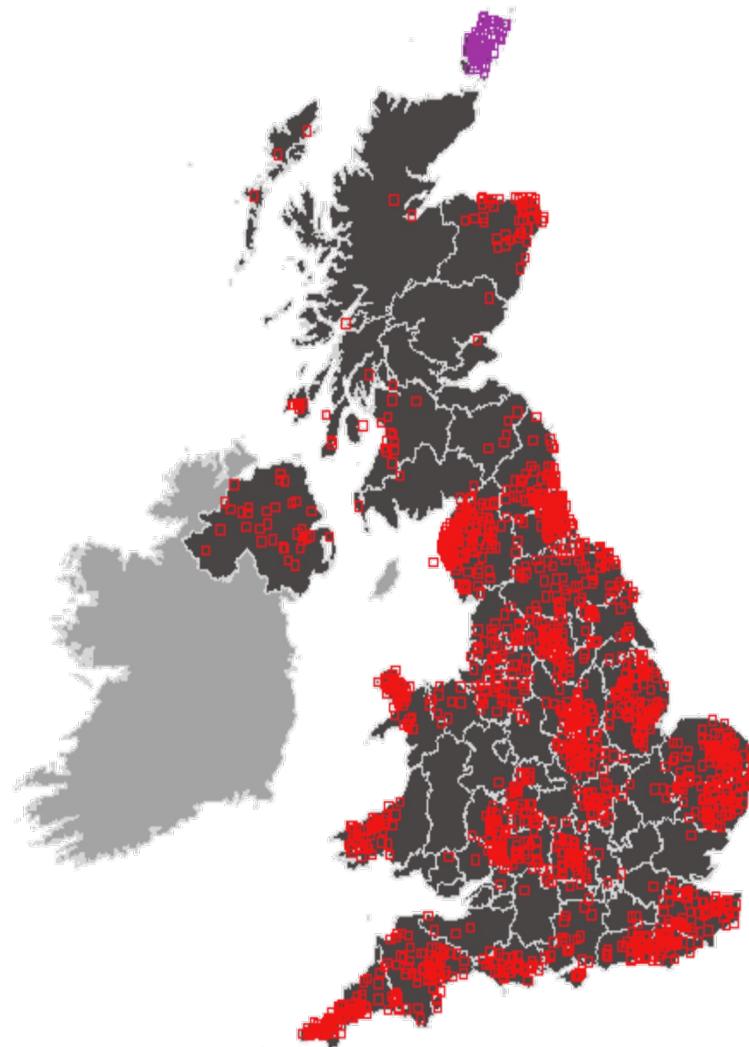


Haplotype clustering in the British Isles

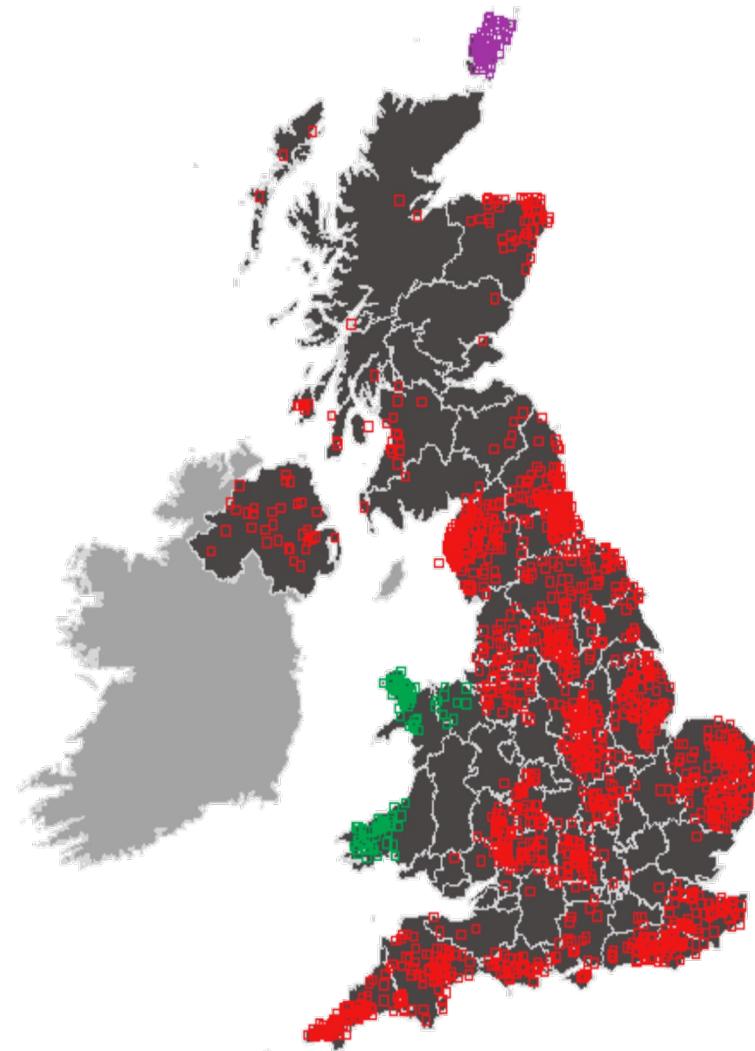
- 2,039 individuals collected across rural areas of UK
- All four-grandparents born within 80km of one another
- Inferred DNA matching patterns amongst all individuals
- Grouped individuals into hierarchical groups based on these DNA patterns



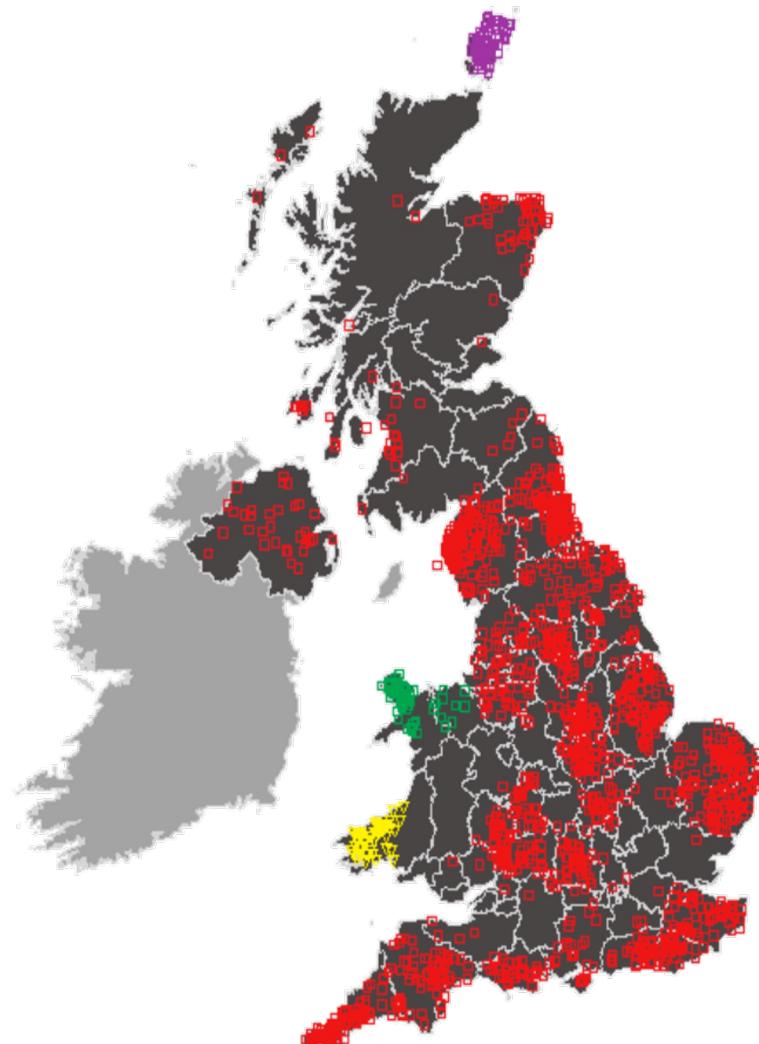
Haplotype clustering in the British Isles



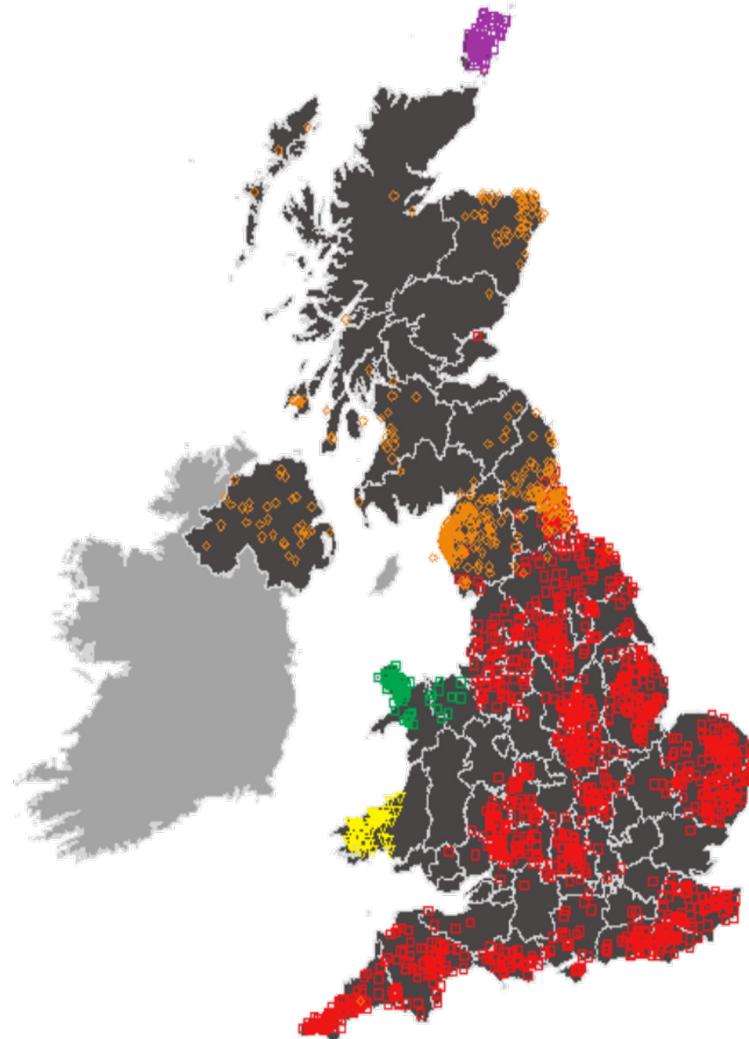
Haplotype clustering in the British Isles



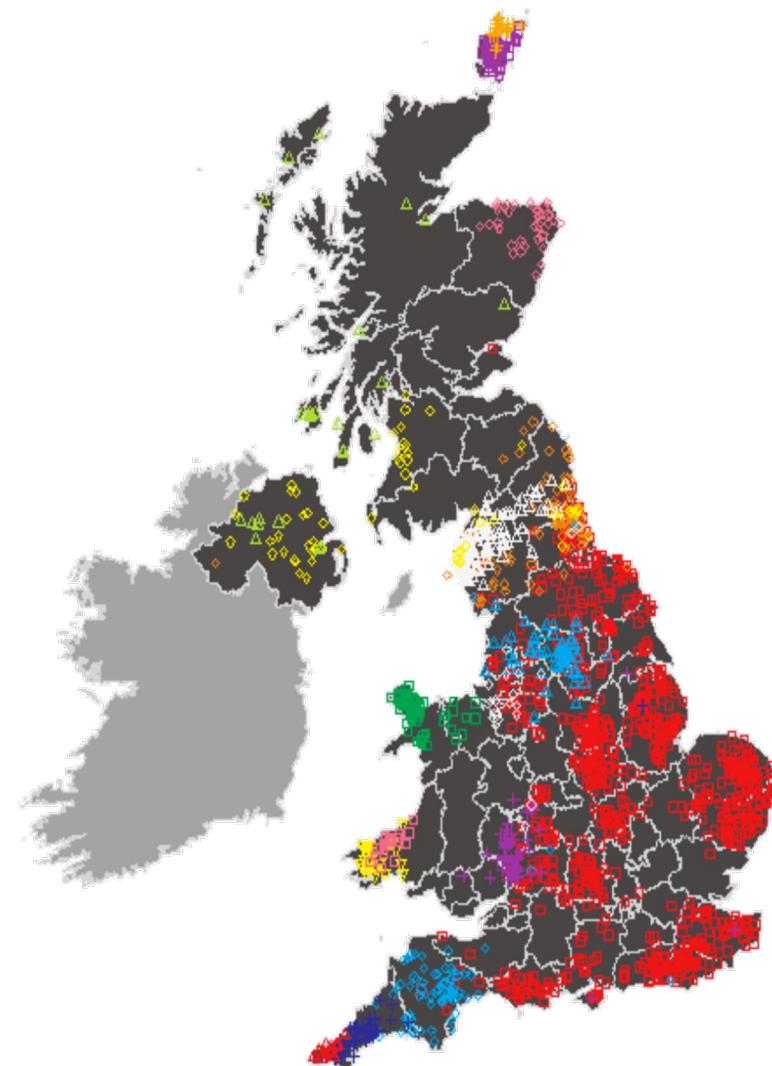
Haplotype clustering in the British Isles



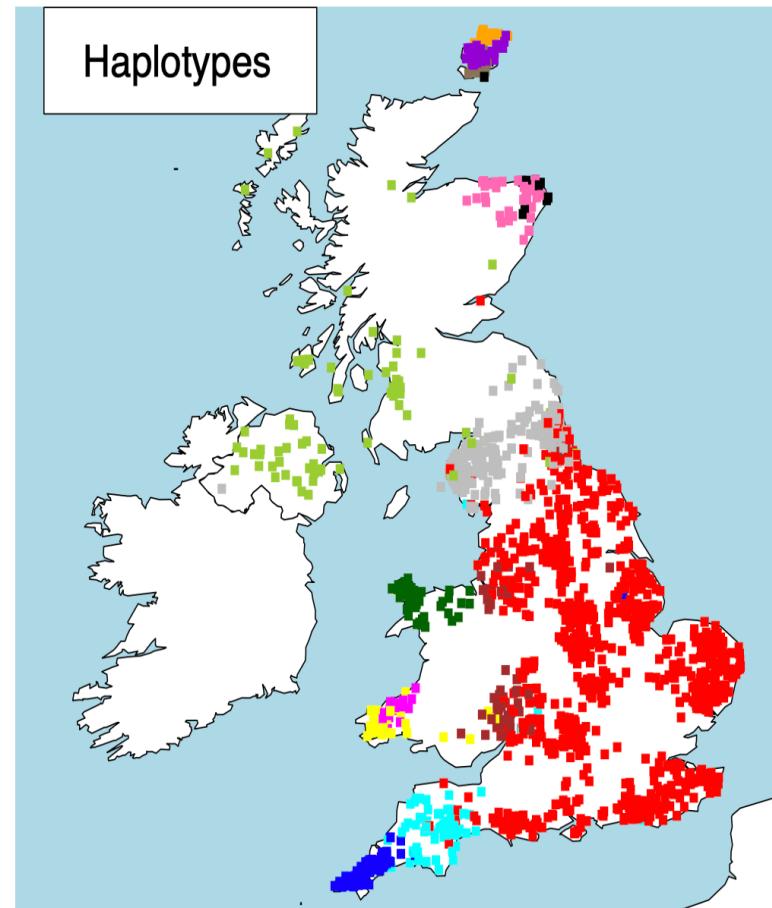
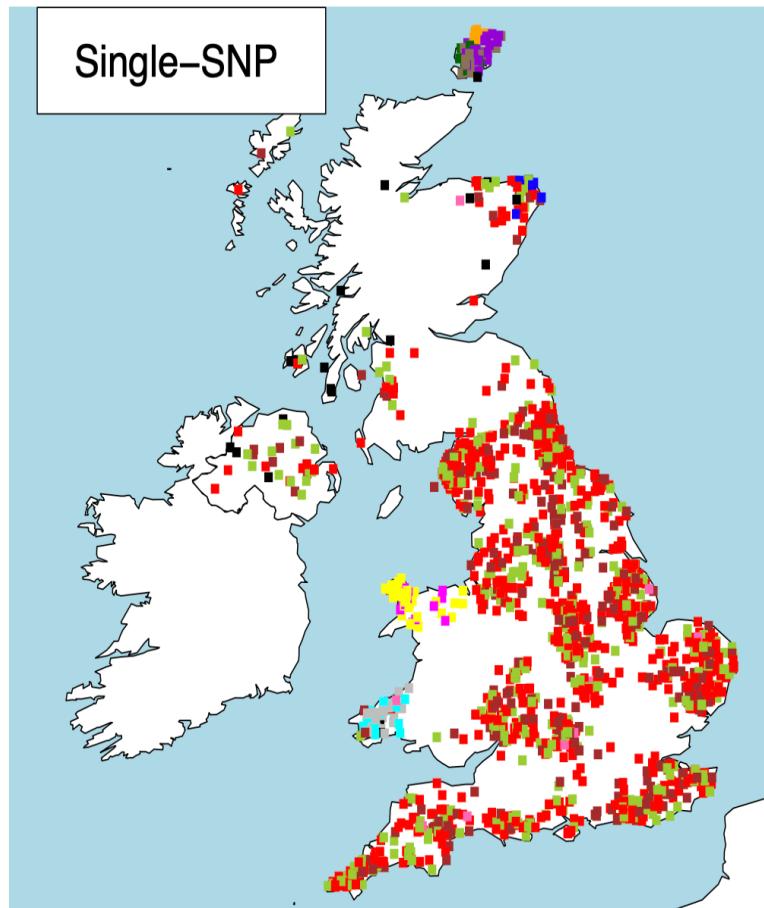
Haplotype clustering in the British Isles



Haplotype clustering in the British Isles



Haplotype clustering in the British Isles



Haplotype clustering - fineSTRUCTURE

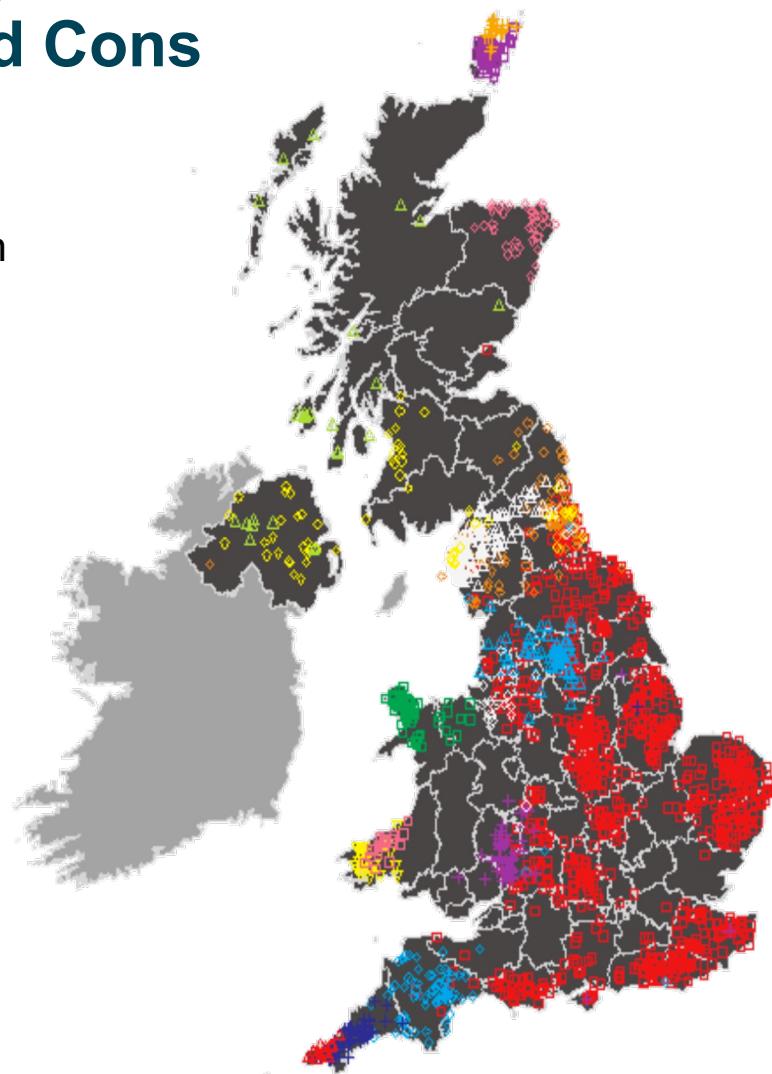
Pros and Cons

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

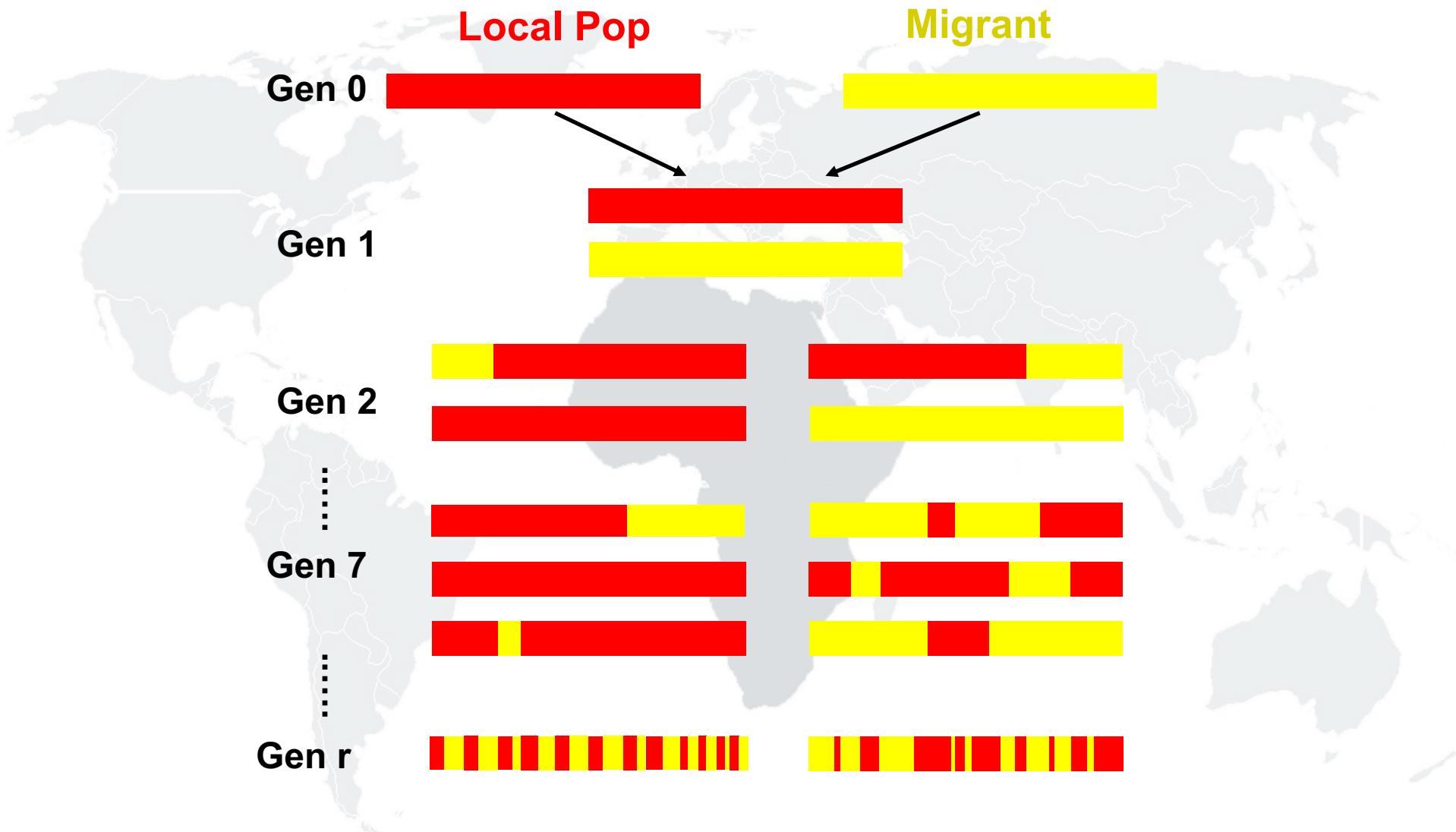


Haplotype sharing patterns can also be used to infer admixture events in human populations

- 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.

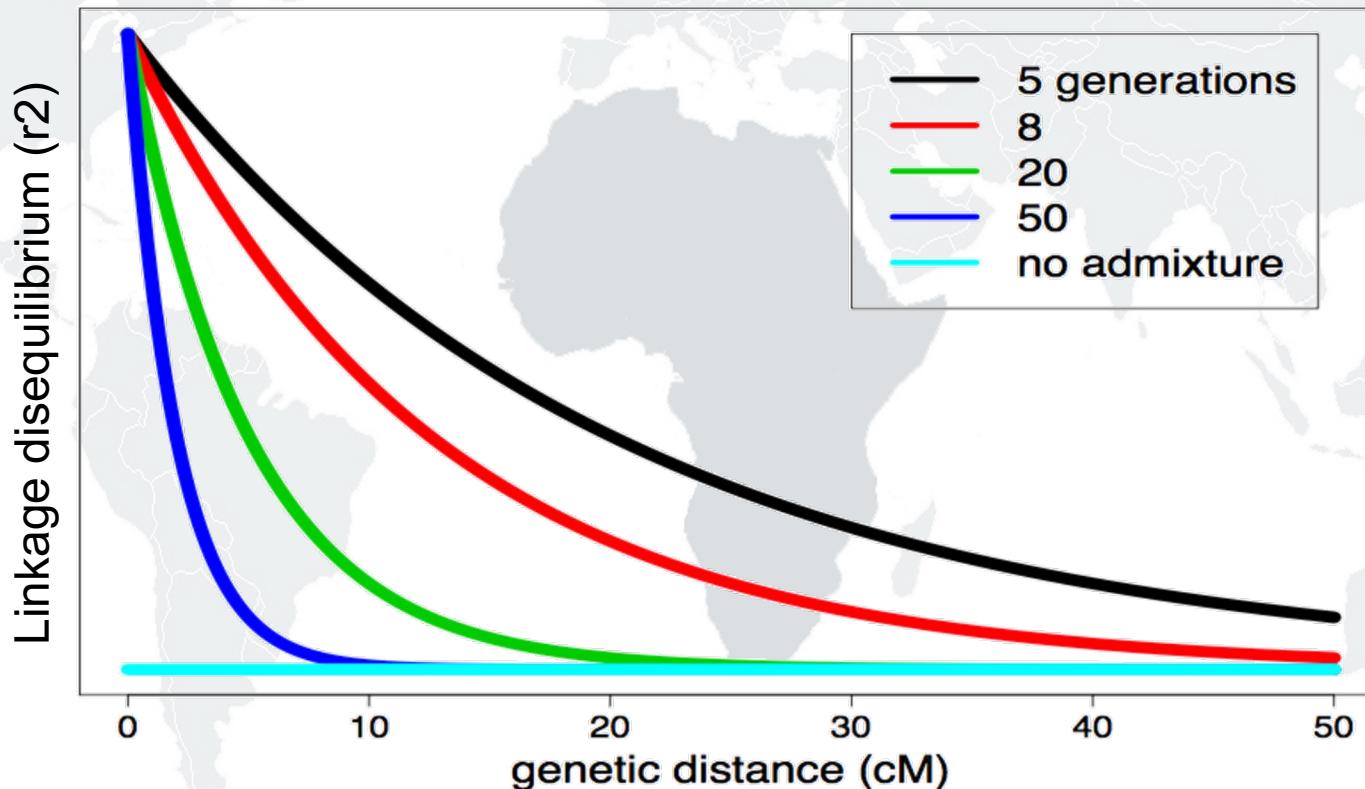


- The length of red and yellow depends on the number of recombination's that have occurred.
 - Eg. a continuous tract is one in which no ancestry-breaking recombinations have occurred.
 - Thus, tract length distribution should be exponential.
- Can we use this distribution to approximate the number of generations since admixture?

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

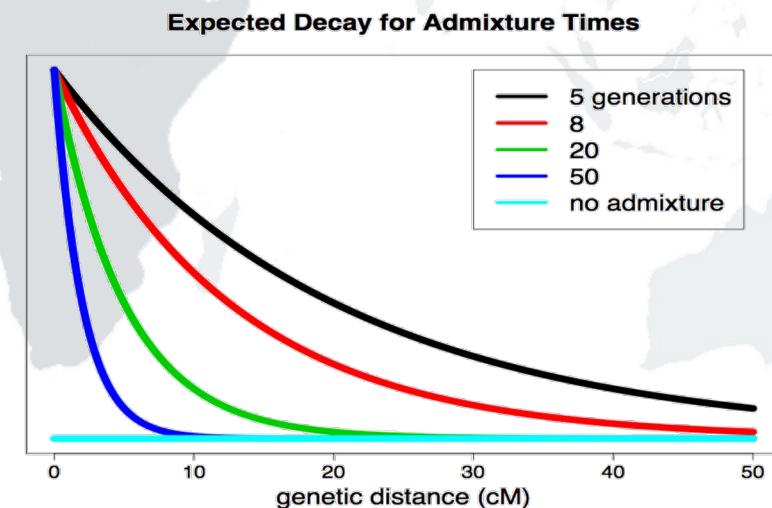
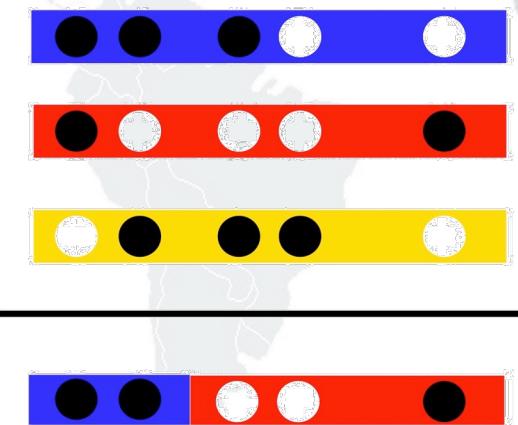


Expected Decay for Admixture Times

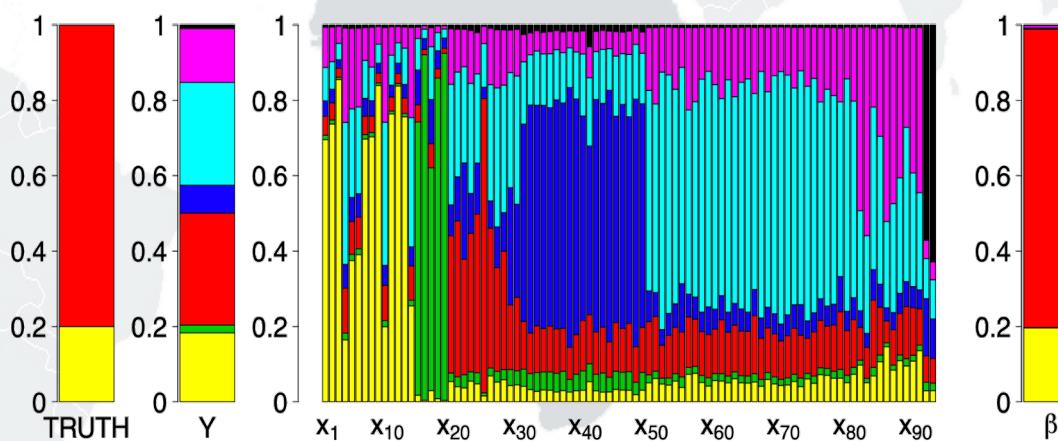
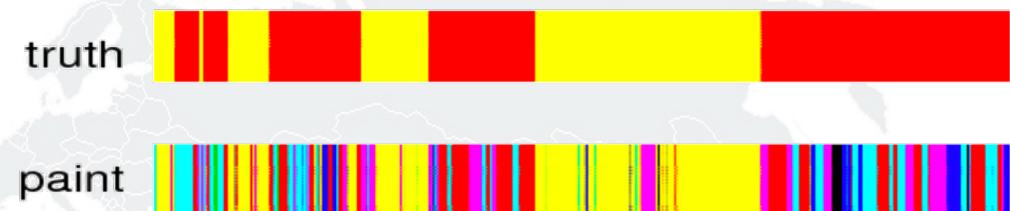
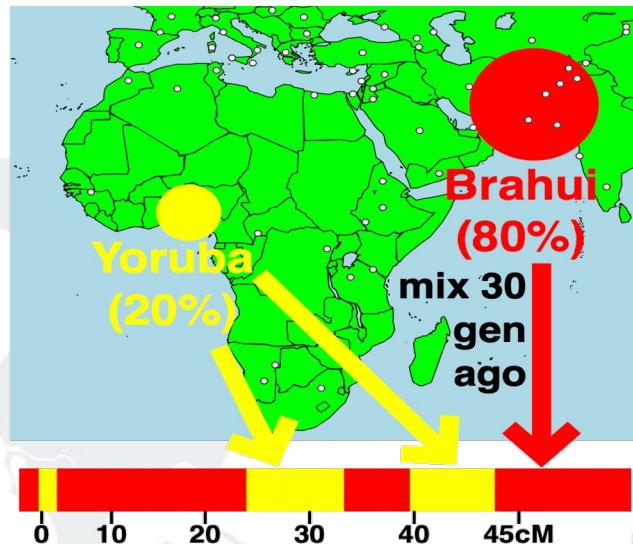


Inferring admixture events in human populations

- 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



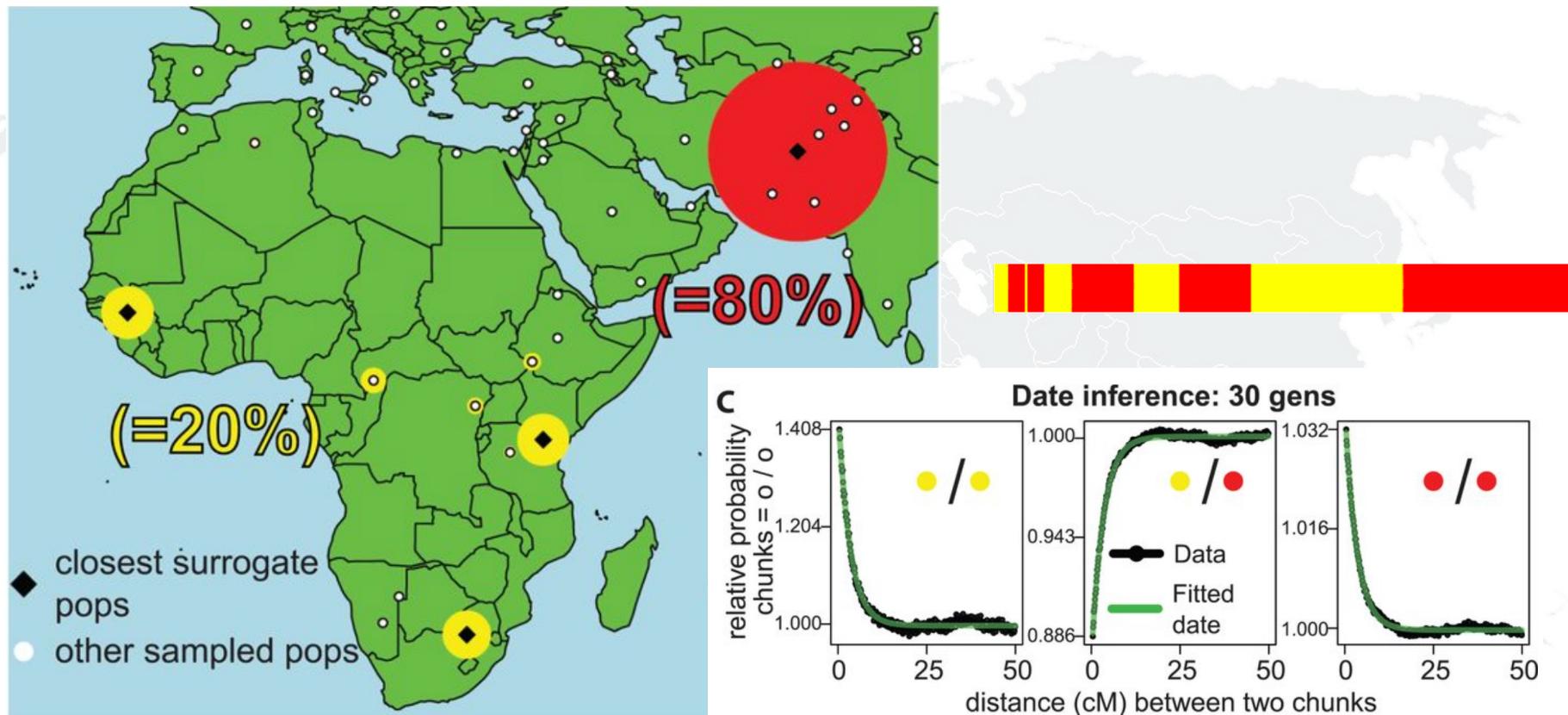
Inferring admixture events - Globetrotter



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

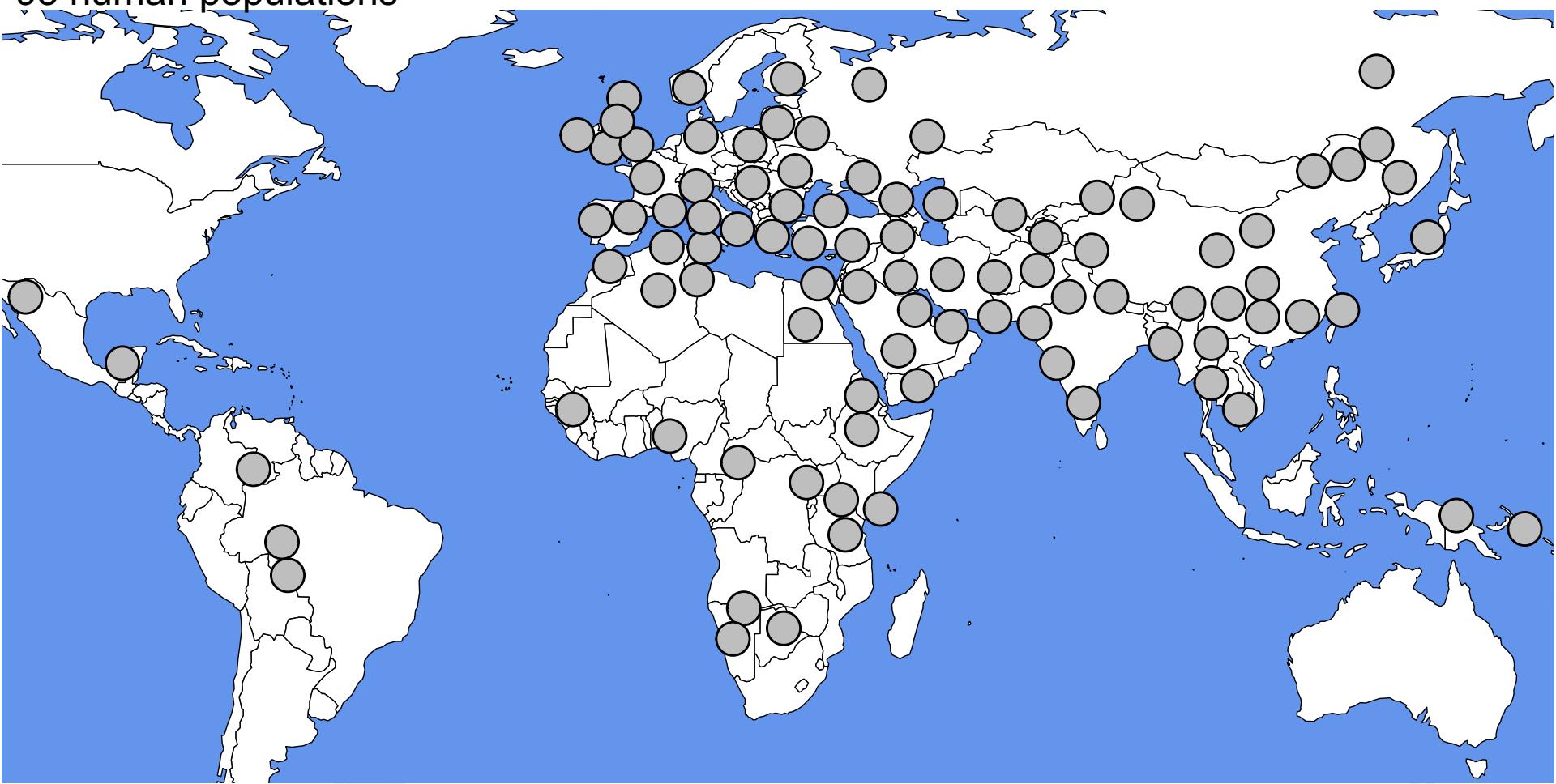
Hellenthal et al, Science, 2014

Inferring admixture events - Globetrotter



Genetic mixing over the past 4000 years

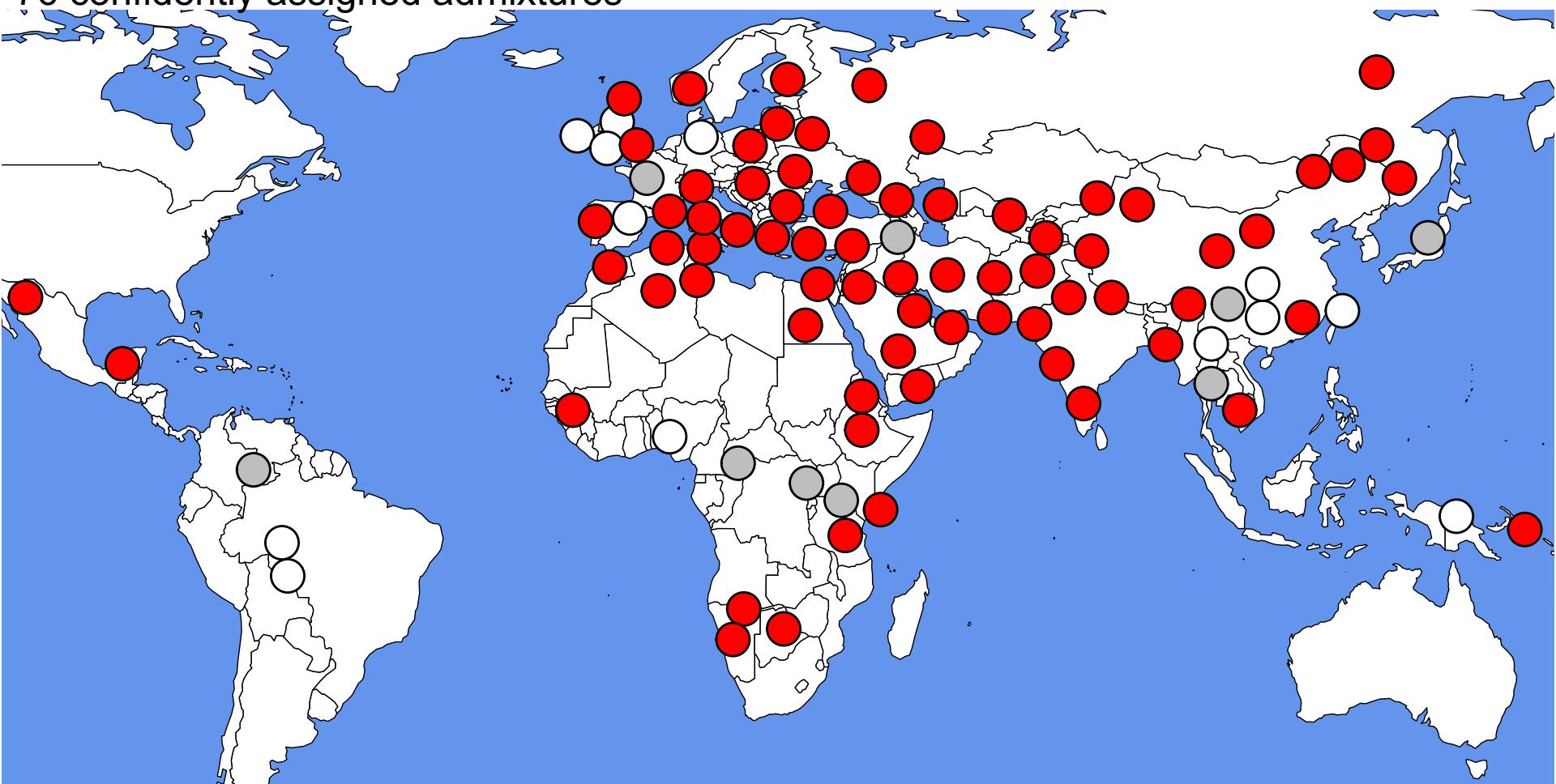
93 human populations



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

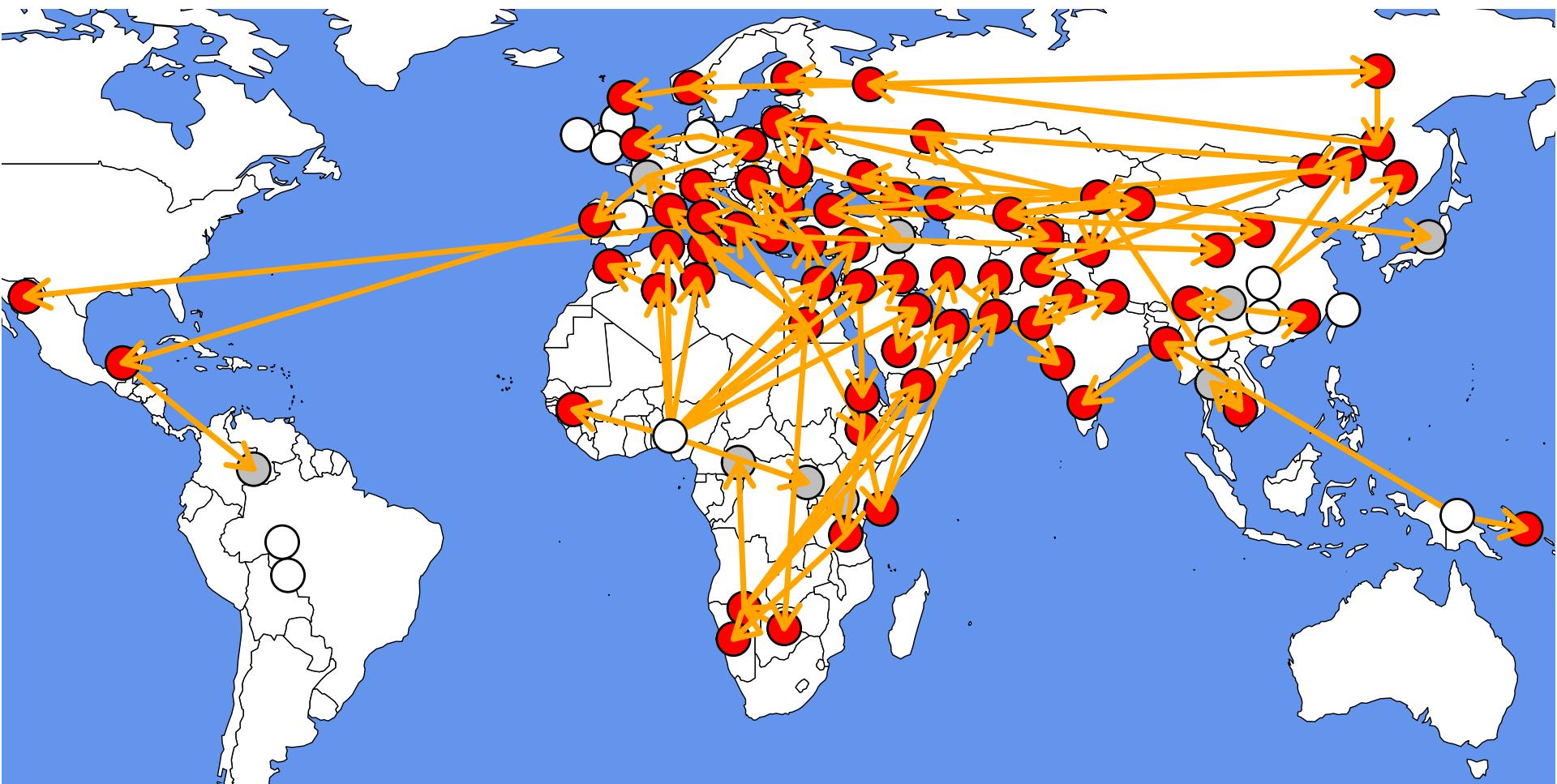
Genetic mixing over the past 4000 years

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 - instructions

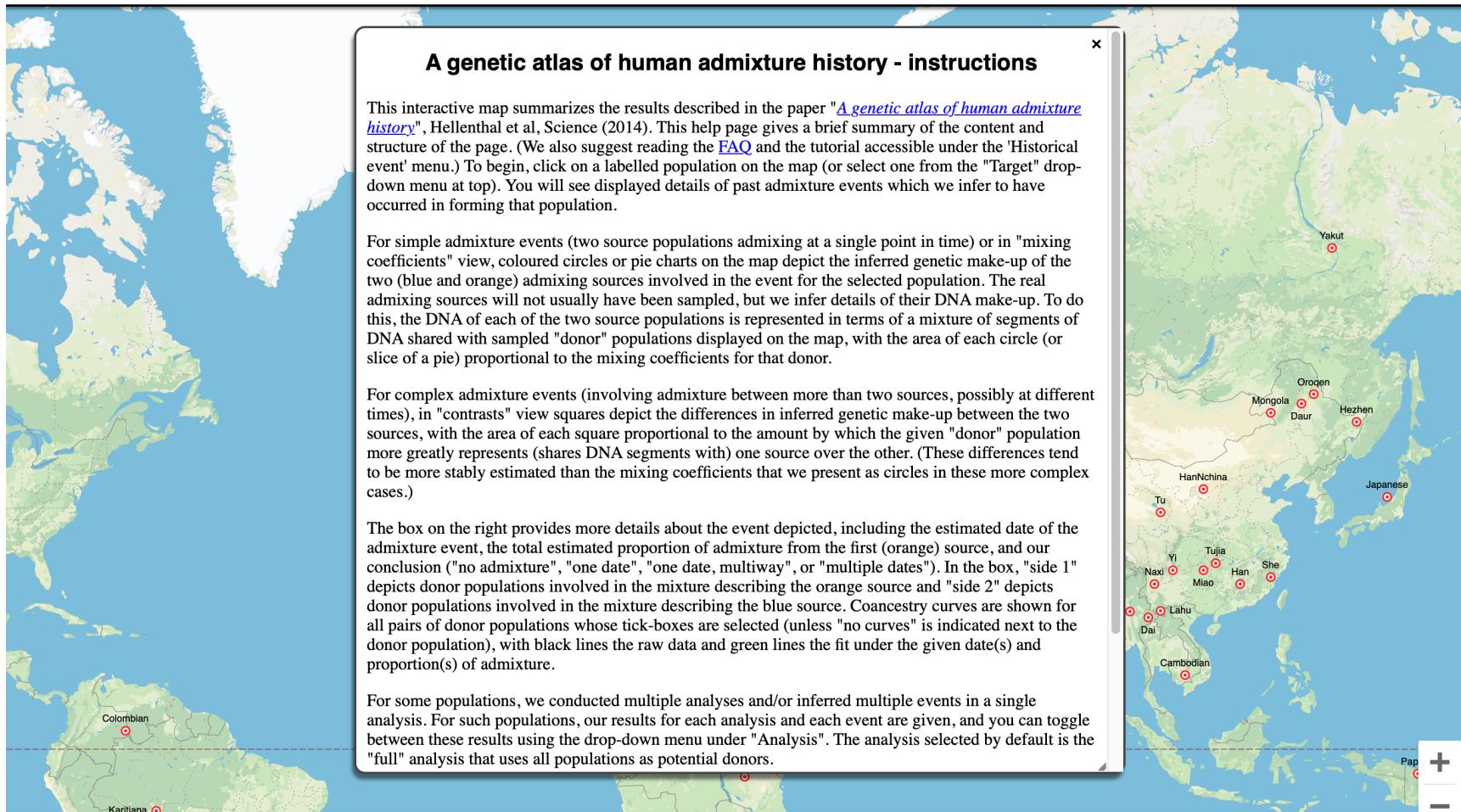
This interactive map summarizes the results described in the paper "[A genetic atlas of human admixture history](#)", Hellenthal et al, Science (2014). This help page gives a brief summary of the content and structure of the page. (We also suggest reading the [FAQ](#) and the tutorial accessible under the 'Historical event' menu.) To begin, click on a labelled population on the map (or select one from the "Target" drop-down menu at top). You will see displayed details of past admixture events which we infer to have occurred in forming that population.

For simple admixture events (two source populations admixing at a single point in time) or in "mixing coefficients" view, coloured circles or pie charts on the map depict the inferred genetic make-up of the two (blue and orange) admixing sources involved in the event for the selected population. The real admixing sources will not usually have been sampled, but we infer details of their DNA make-up. To do this, the DNA of each of the two source populations is represented in terms of a mixture of segments of DNA shared with sampled "donor" populations displayed on the map, with the area of each circle (or slice of a pie) proportional to the mixing coefficients for that donor.

For complex admixture events (involving admixture between more than two sources, possibly at different times), in "contrasts" view squares depict the differences in inferred genetic make-up between the two sources, with the area of each square proportional to the amount by which the given "donor" population more greatly represents (shares DNA segments with) one source over the other. (These differences tend to be more stably estimated than the mixing coefficients that we present as circles in these more complex cases.)

The box on the right provides more details about the event depicted, including the estimated date of the admixture event, the total estimated proportion of admixture from the first (orange) source, and our conclusion ("no admixture", "one date", "one date, multiway", or "multiple dates"). In the box, "side 1" depicts donor populations involved in the mixture describing the orange source and "side 2" depicts donor populations involved in the mixture describing the blue source. Coancestry curves are shown for all pairs of donor populations whose tick-boxes are selected (unless "no curves" is indicated next to the donor population), with black lines the raw data and green lines the fit under the given date(s) and proportion(s) of admixture.

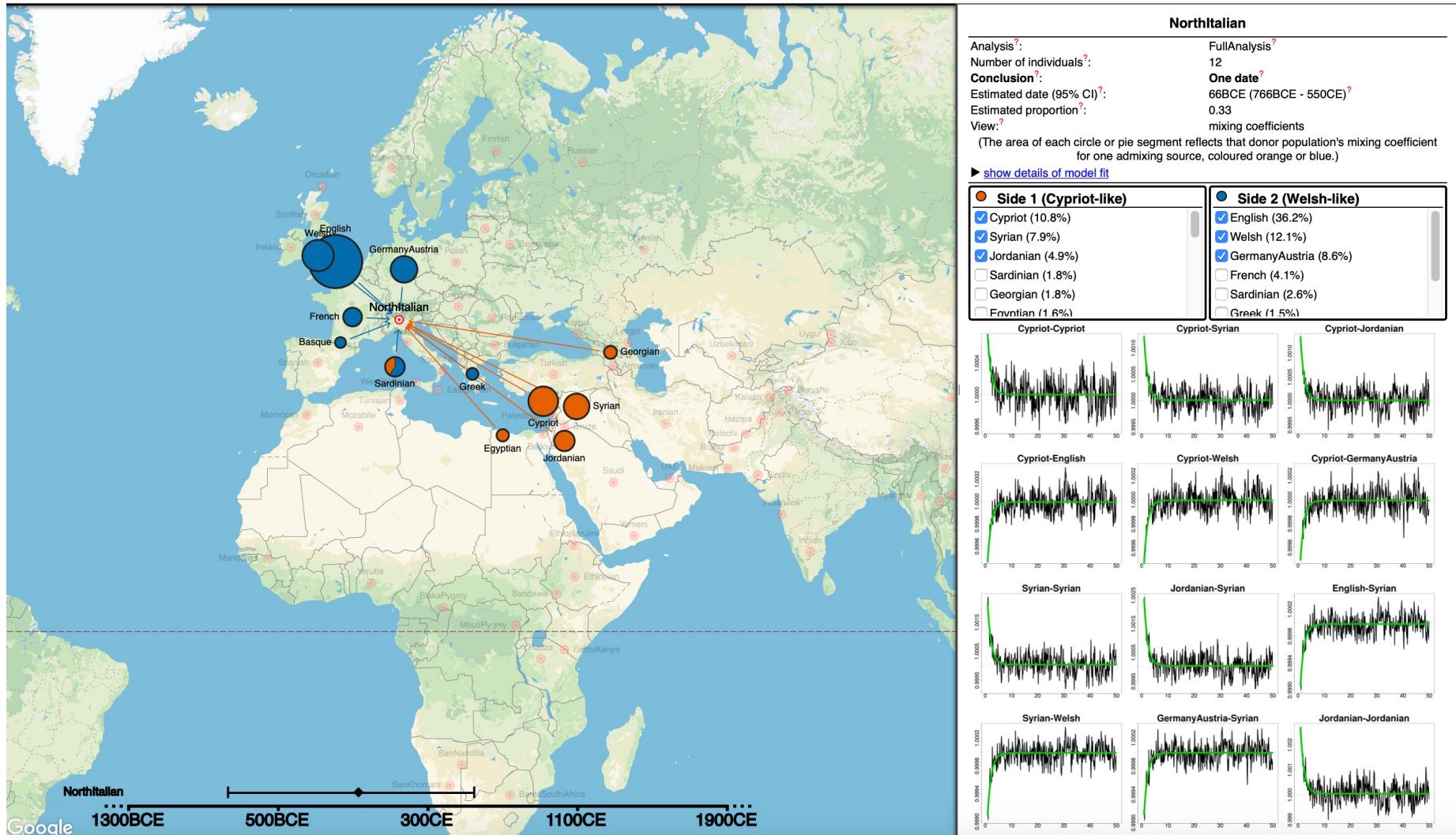
For some populations, we conducted multiple analyses and/or inferred multiple events in a single analysis. For such populations, our results for each analysis and each event are given, and you can toggle between these results using the drop-down menu under "Analysis". The analysis selected by default is the "full" analysis that uses all populations as potential donors.



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 ↓



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!