

Haplotype-based methods for demography

Garrett Hellenthal
g.hellenthal@ucl.ac.uk

University College London

EMBO – Population Genomics: background and tools
May 25, 2017

Introduction

- ▶ this lecture/practical will cover haplotype-based methods for demography, in particular:
 1. inferring admixture events between two or more source groups (*GLOBETROTTER*)
 2. inferring past population size changes over time (*PSMC/MSMC*)
- ▶ we will use SNP data throughout, with *PSMC/MSMC* requiring sequencing data

Outline

inferring admixture (*GLOBETROTTER*)

inferring pop size changes (*PSMC, MSMC*)

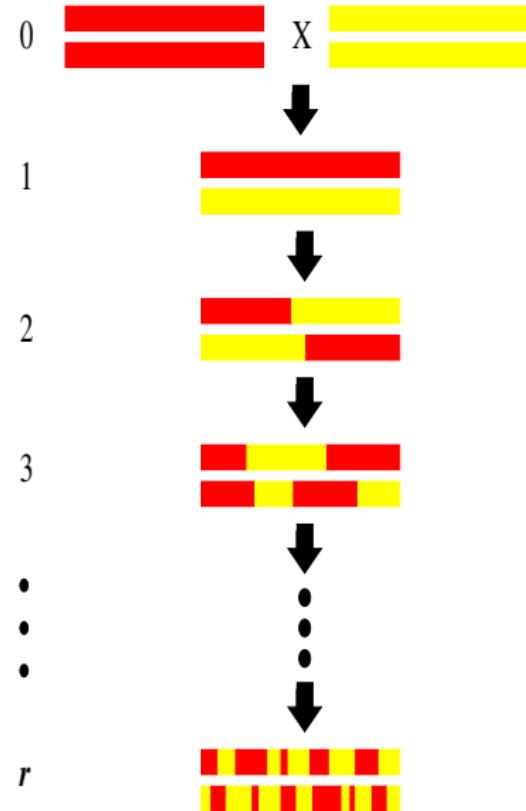
Outline

inferring admixture (*GLOBETROTTER*)

inferring pop size changes (*PSMC, MSMC*)

Inferring admixture using autosomal DNA

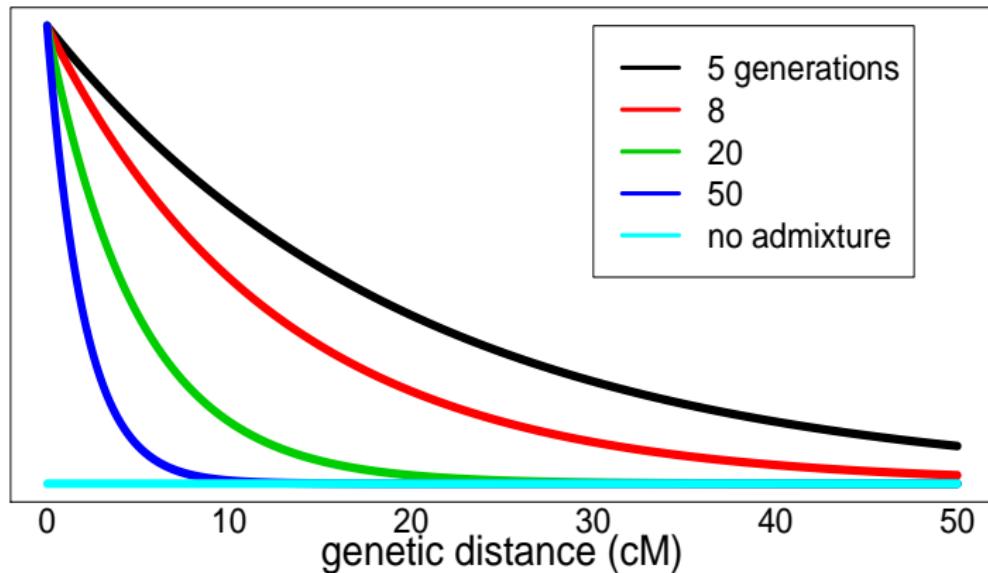
- ▶ two populations (**red**,**yellow**) admix r generations ago
- ▶ genetic pieces from each population get smaller each subsequent generation due to recombination
- ▶ assuming recombination occurs as Poisson process, size of contiguous **red** and **yellow** segments in present-day DNA follow exponential distribution with rate r



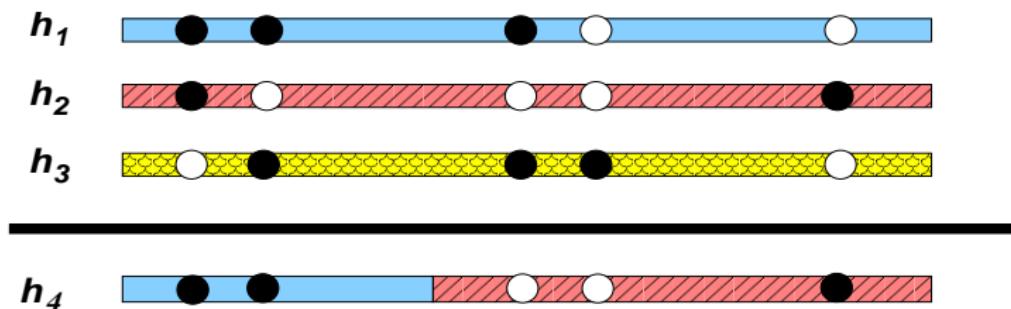
(Falush et al 2003, *Genetics* 164:1567)

Dating admixture events using autosomal DNA

distribution of **red** and **yellow** segment sizes, based on when the two groups mixed:

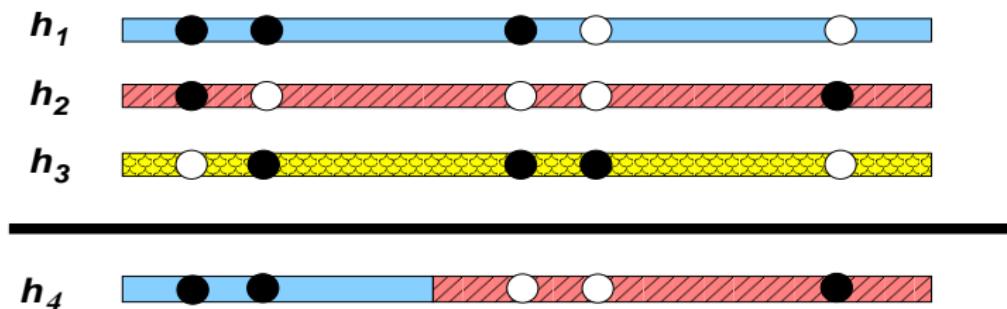


Inferring admixture – chromosome painting



- ▶ “paint” admixed individual h_4 using donors h_1 , h_2 , h_3
- ▶ e.g. here h_4 appears to be a mixture of **blue**, **red** pops
- ▶ size of **blue/red** segments tell us when mixture occurred

Inferring admixture – chromosome painting



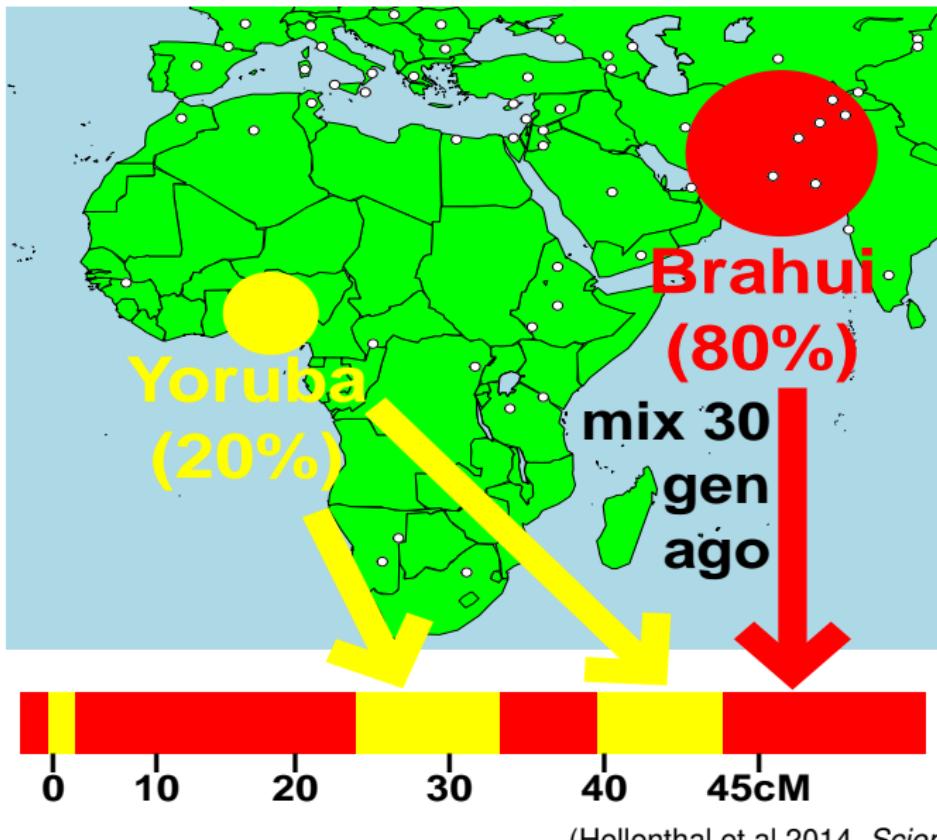
- ▶ “paint” admixed individual h_4 using donors h_1, h_2, h_3
- ▶ e.g. here h_4 appears to be a mixture of **blue**, **red** pops
- ▶ size of **blue/red** segments tell us when mixture occurred

Issue: Accurate painting is challenging. Instead:

1. identify SNPs/haplotypes that *likely* derive from admixing groups
2. plot decay of LD versus cM distance between such SNPs/haplotypes

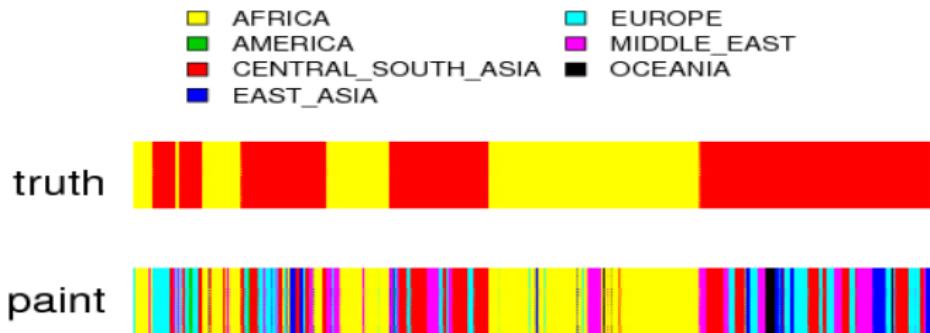
Disadvantage: can be computationally demanding

Simulation: 80% Brahui + 20% Yoruba, 30gen



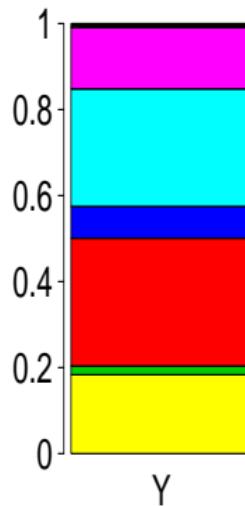
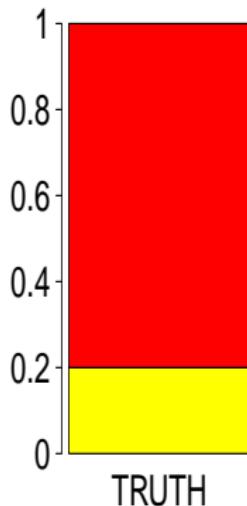
Infer mixing groups (Sim: 80% Brahui + 20% Yoruba, 30gen)

(1) paint admixed inds using 93 world groups (*CHROMOPAINTER*)



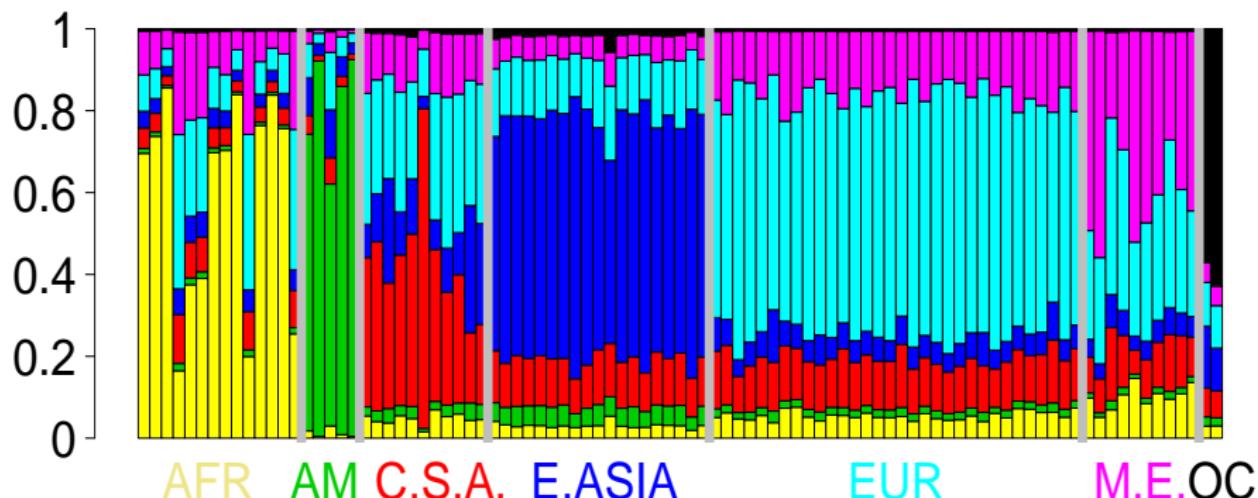
Infer mixing groups (Sim: 80% Brahui + 20% Yoruba, 30gen)

(1) paint admixed inds using 93 world groups (*CHROMOPAINTER*)



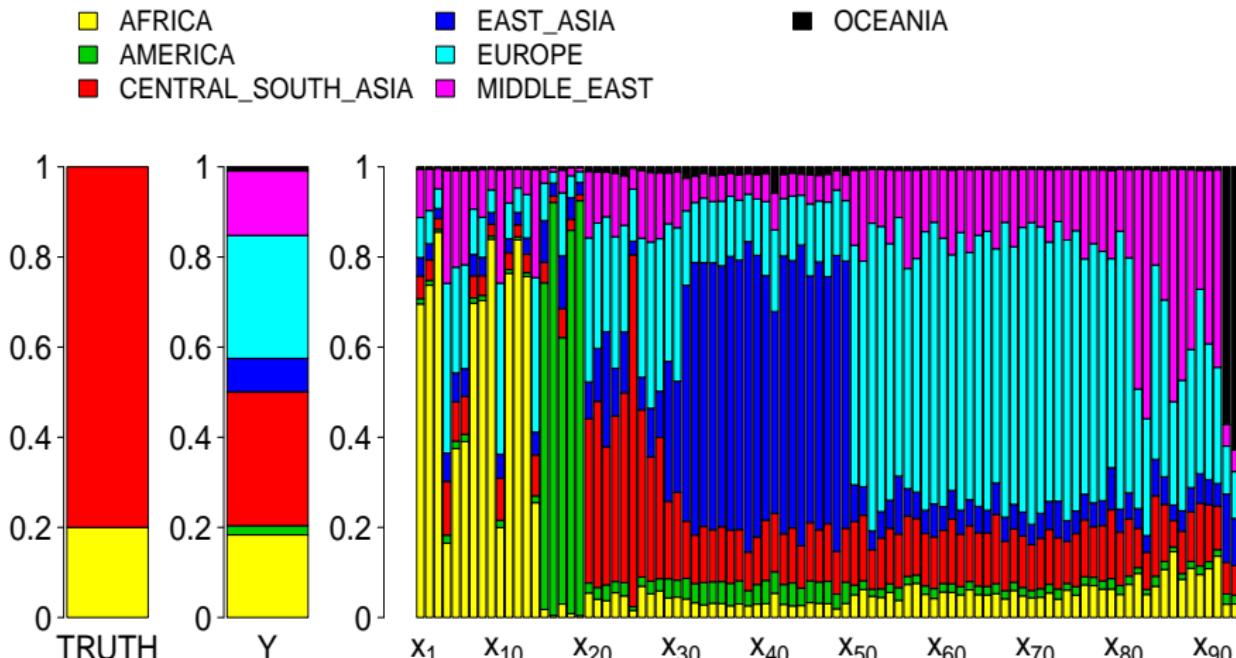
Infer mixing groups (Sim: 80% Brahui + 20% Yoruba, 30gen)

(1) paint admixed inds using 93 world groups (*CHROMOPAINTER*)



Infer mixing groups (Sim: 80% Brahui + 20% Yoruba, 30gen)

(1) paint admixed inds using 93 world groups (*CHROMOPAINTER*)

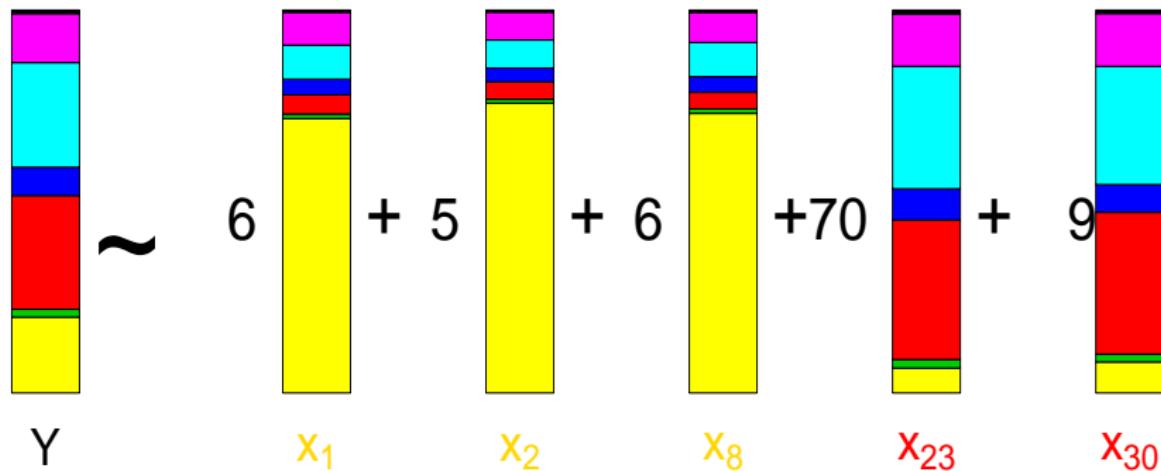


(2) form admixed ind's painting as mixture of that for all 93 groups:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{93} X_{93} \quad (\text{with } \sum_i \beta_i = 1.0)$$

Infer mixing groups (Sim: 80% Brahui + 20% Yoruba, 30gen)

(1) paint admixed inds using 93 world groups (*CHROMOPAINTER*)

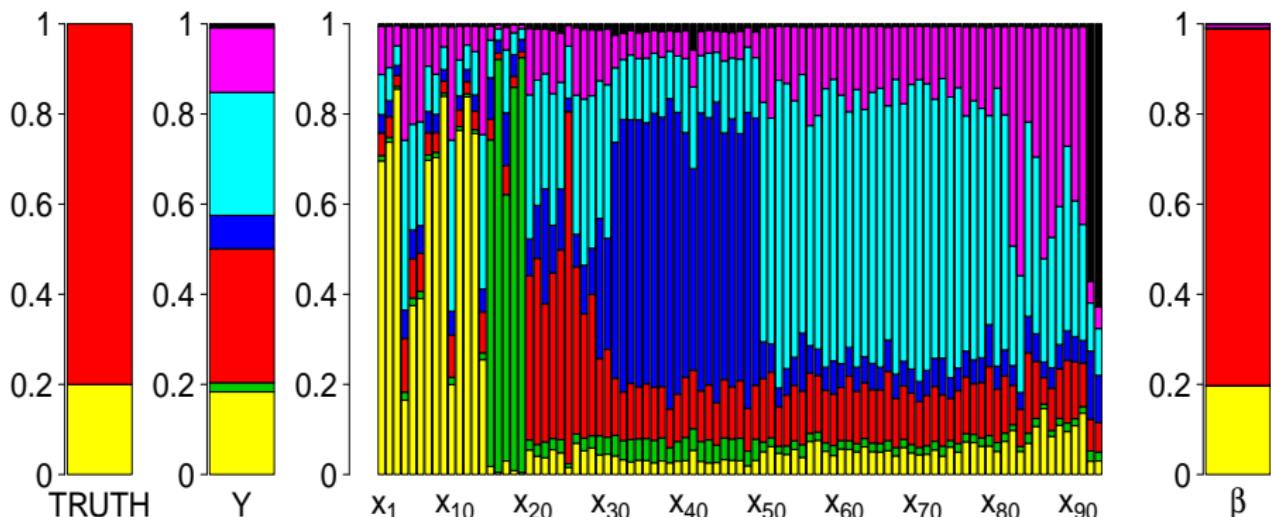


(2) form admixed ind's painting as mixture of that for all 93 groups:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{93} X_{93} \quad (\text{with } \sum_i \beta_i = 1.0)$$

Infer mixing groups (Sim: 80% Brahui + 20% Yoruba, 30gen)

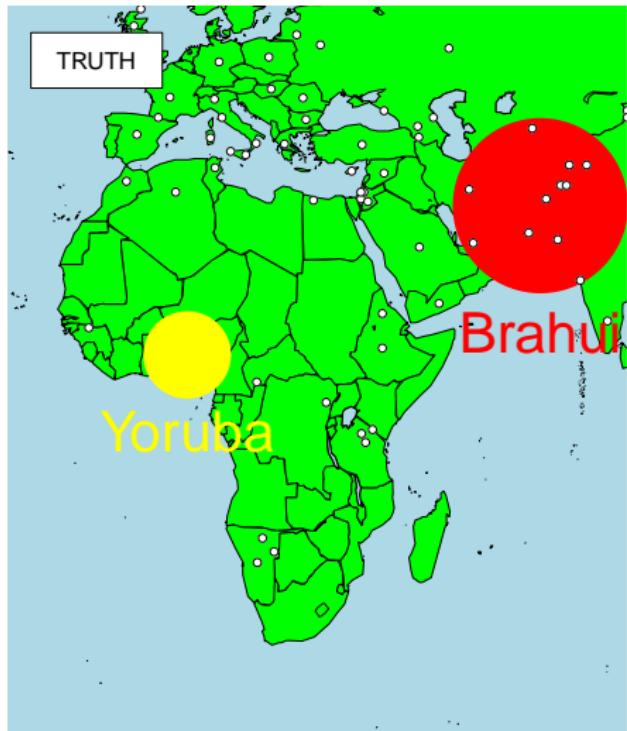
(1) paint admixed inds using 93 world groups (*CHROMOPAINTER*)



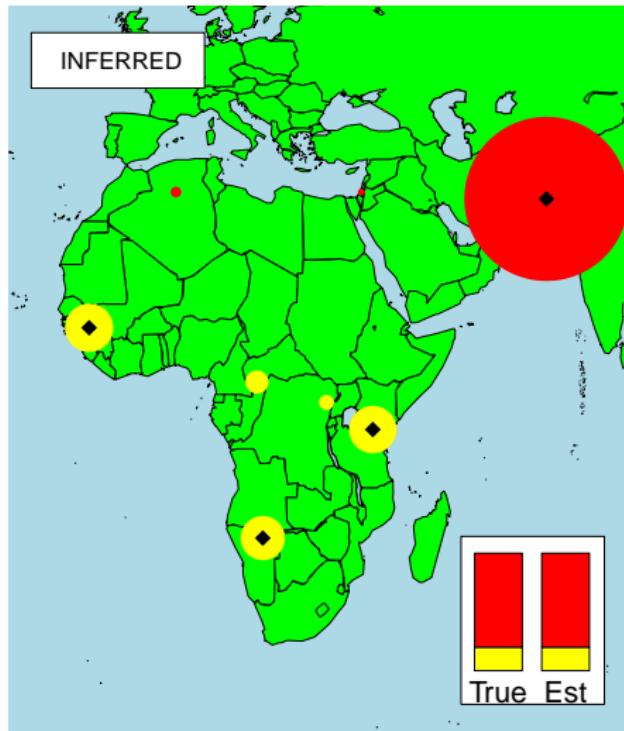
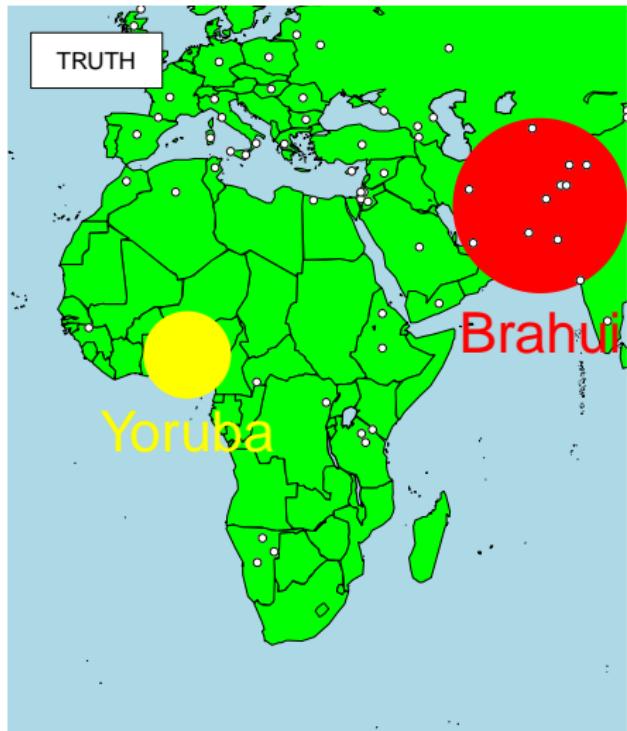
(2) form admixed ind's painting as mixture of that for all 93 groups:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{93} X_{93} \quad (\text{with } \sum_i \beta_i = 1.0)$$

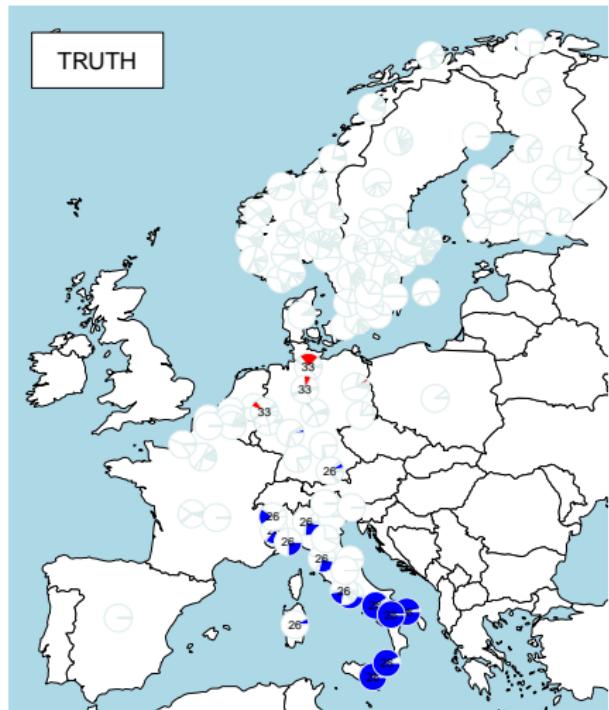
Infer mixing groups (Sim: 80% Brahui + 20% Yoruba, 30gen)



Infer mixing groups (Sim: 80% Brahui + 20% Yoruba, 30gen)

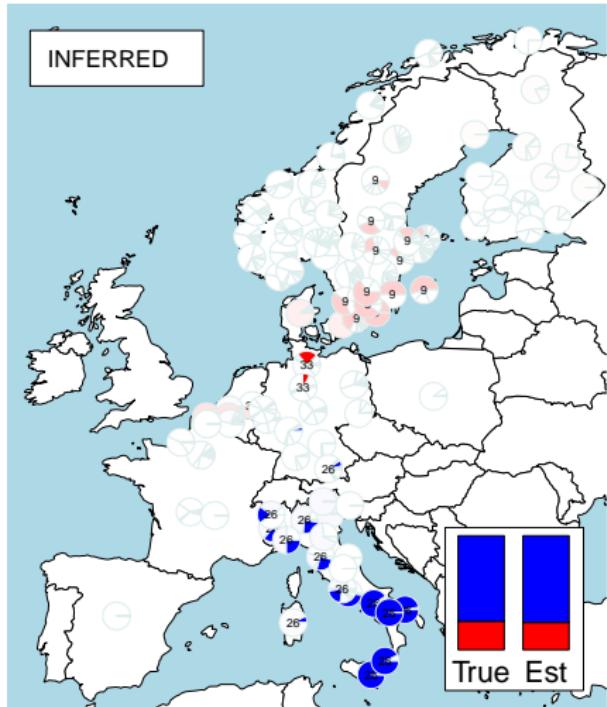
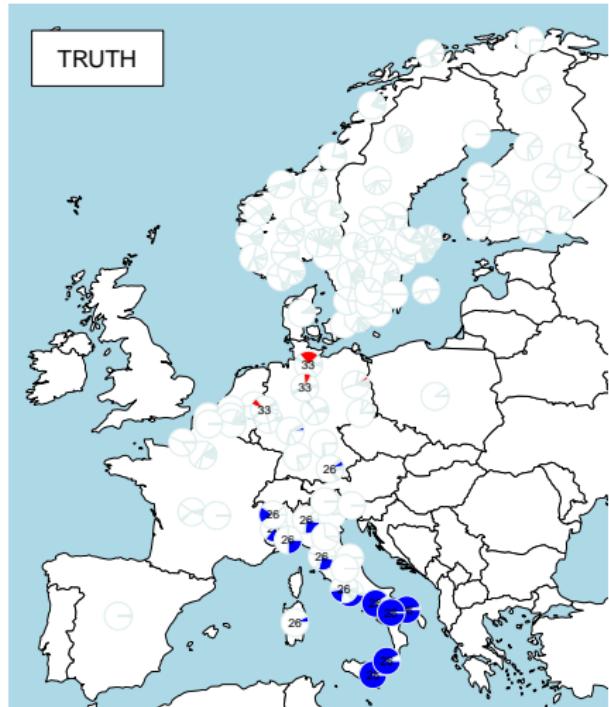


Simulated Europe Admixture: 75% Italy + 25% N.Germany, 40gen



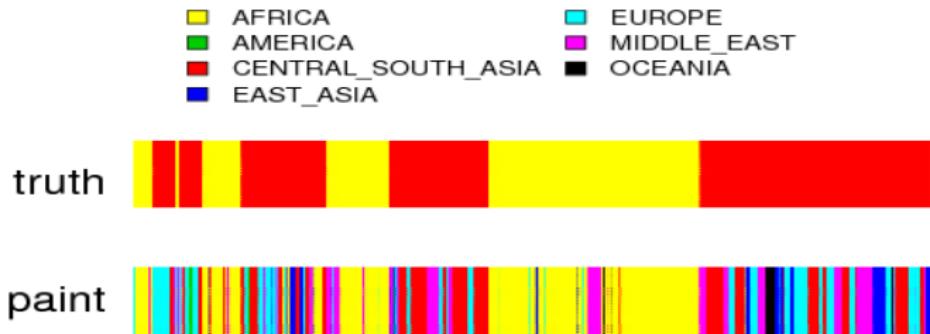
(Leslie et al 2015, *Nature* 519:309)

Simulated Europe Admixture: 75% Italy + 25% N.Germany, 40gen



(Leslie et al 2015, *Nature* 519:309)

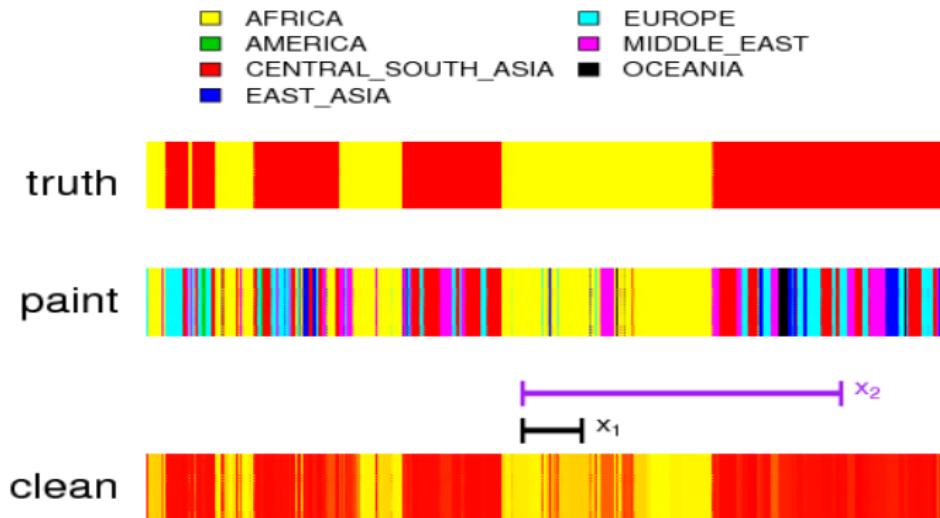
Dating Admixture (Sim: 80% Brahui + 20% Yoruba, 30gen)



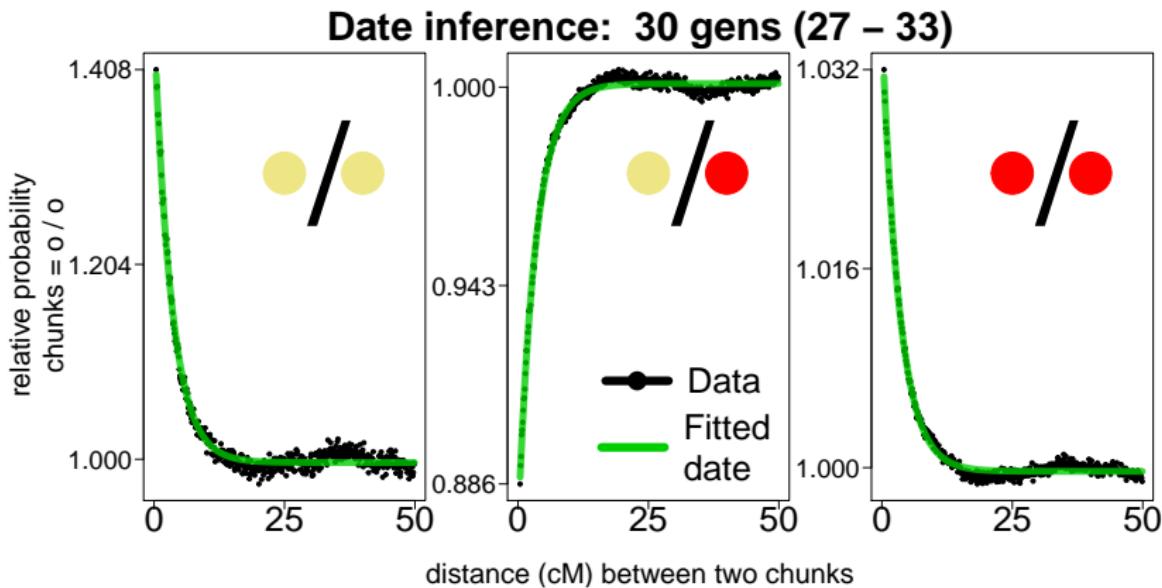
Dating Admixture (Sim: 80% Brahui + 20% Yoruba, 30gen)



Dating Admixture (Sim: 80% Brahui + 20% Yoruba, 30gen)



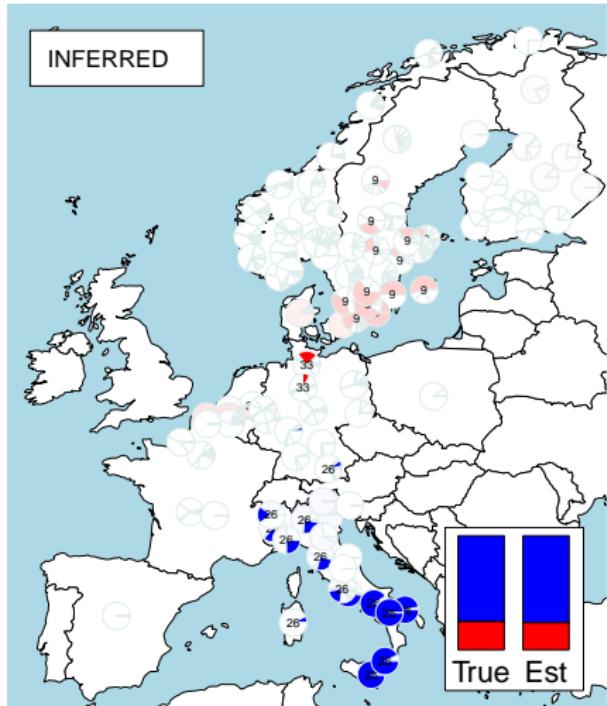
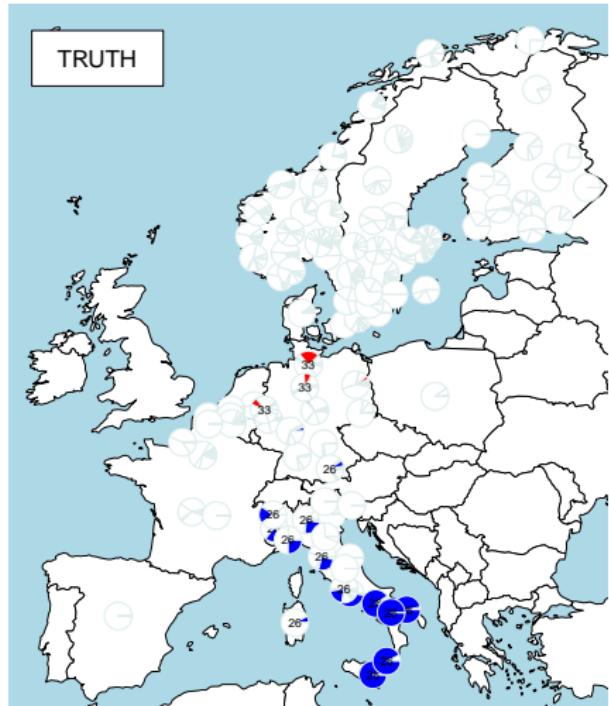
Dating Admixture (Sim: 80% Brahui + 20% Yoruba, 30gen)



Coancestry curves:

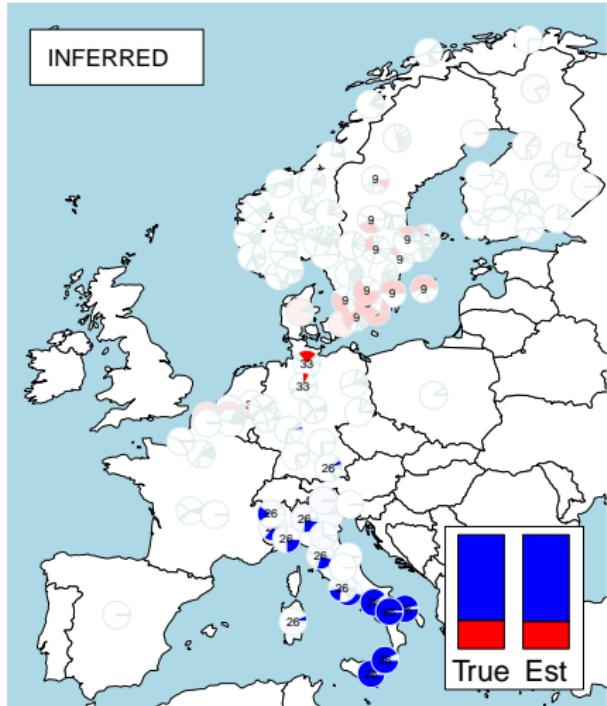
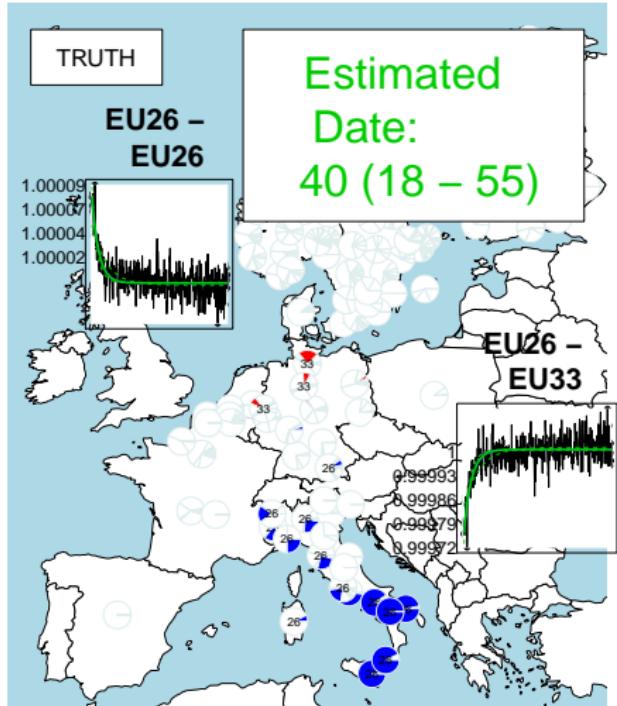
- ▶ **left:** (scaled) probability that two DNA segments ("chunks") separated by cM distance X are both from **yellow** source
- ▶ **middle:** probability one chunk is from **yellow** source, one chunk from **red** source
- ▶ **right:** probability both chunks are from **red** source

Dating Admixture (Sim: 75% Italy + 25% N.Germany, 40gen)



(Leslie et al 2015, *Nature* 519:309)

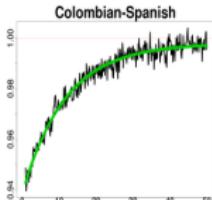
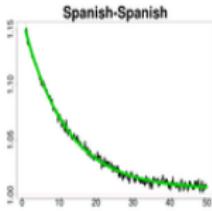
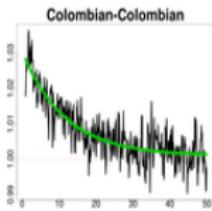
Dating Admixture (Sim: 75% Italy + 25% N.Germany, 40gen)



(Leslie et al 2015, *Nature* 519:309)

Inferring admixture – *GLOBETROTTER* (example)

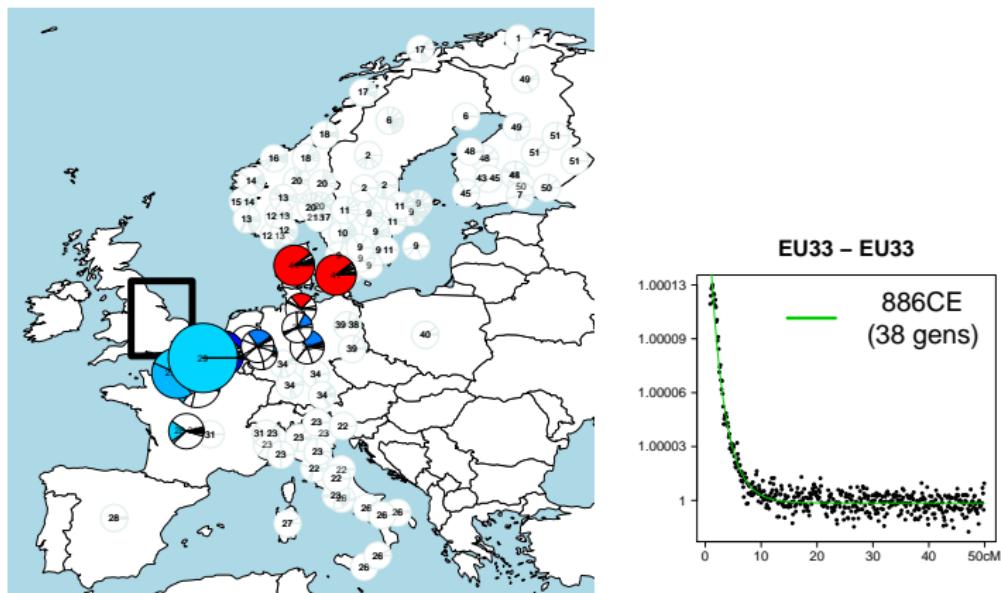
- ▶ **Maya from Mexico** show evidence for admixture between **Spain-like** and **Native American-like** groups
- ▶ dated to 1642-1726AD
- ▶ corresponds to colonial-era migration of Europeans to Americas



Identifying/Dating admixture in United Kingdom

“SE.England” cluster (Y) as mixture of 51 Europe clusters (X_1, \dots, X_{51}):

$$Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{51} X_{51}$$



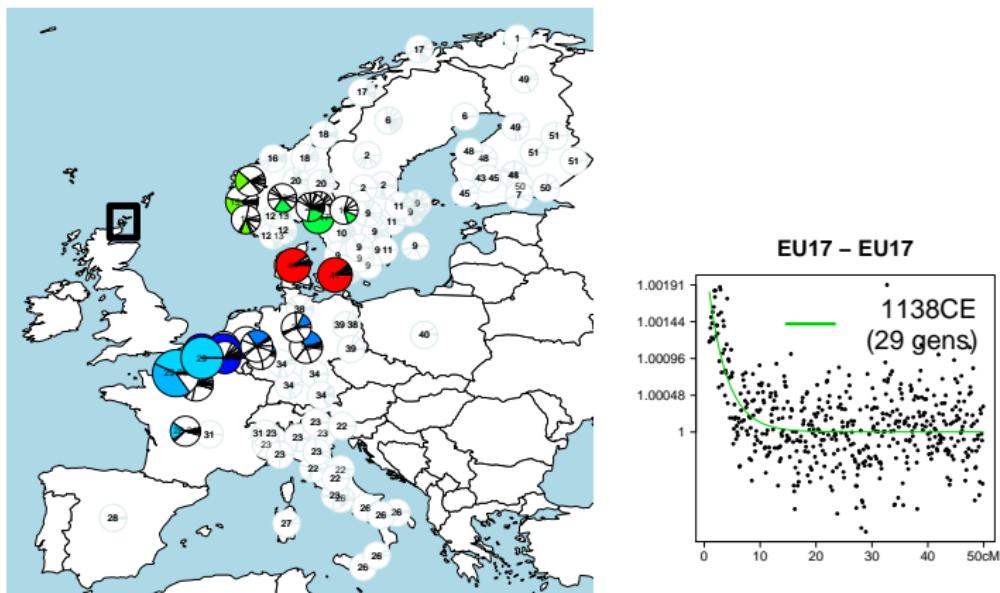
DNA matched to **N.Germany** and **Denmark** → **Anglo-Saxons?**

(Leslie et al 2015, *Nature* 519:309)

Identifying/Dating admixture in United Kingdom

“Orkney” cluster (Y) as mixture of 51 Europe clusters (X_1, \dots, X_{51}):

$$Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{51} X_{51}$$

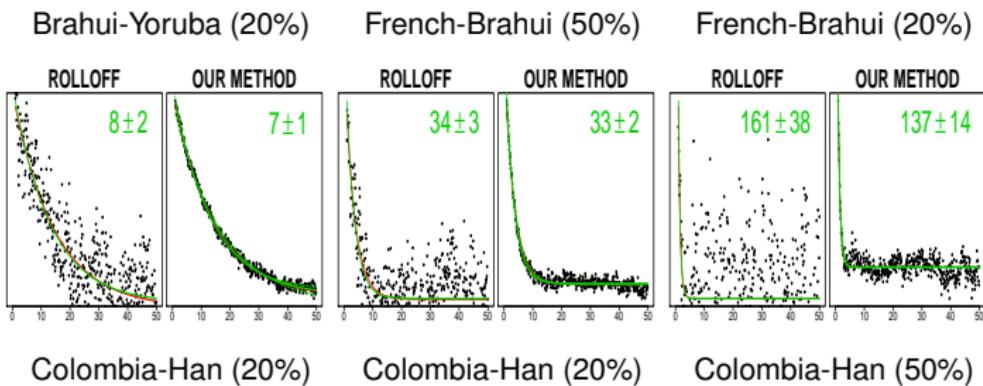
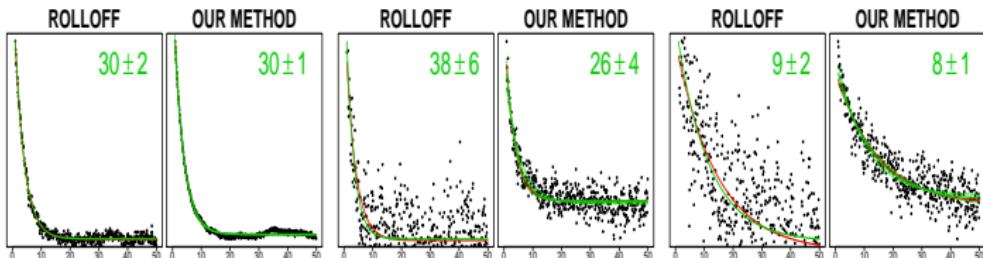


DNA matched to Norway → Norse Vikings?

(Leslie et al 2015, *Nature* 519:309)

Comparison to *ROLLOFF* (e.g. Patterson et al 2012, *Genetics* 192:1065)

- ▶ compared to *ROLLOFF*, fixing two “known” admixing sources
- ▶ increased power/precision when using haplotype information



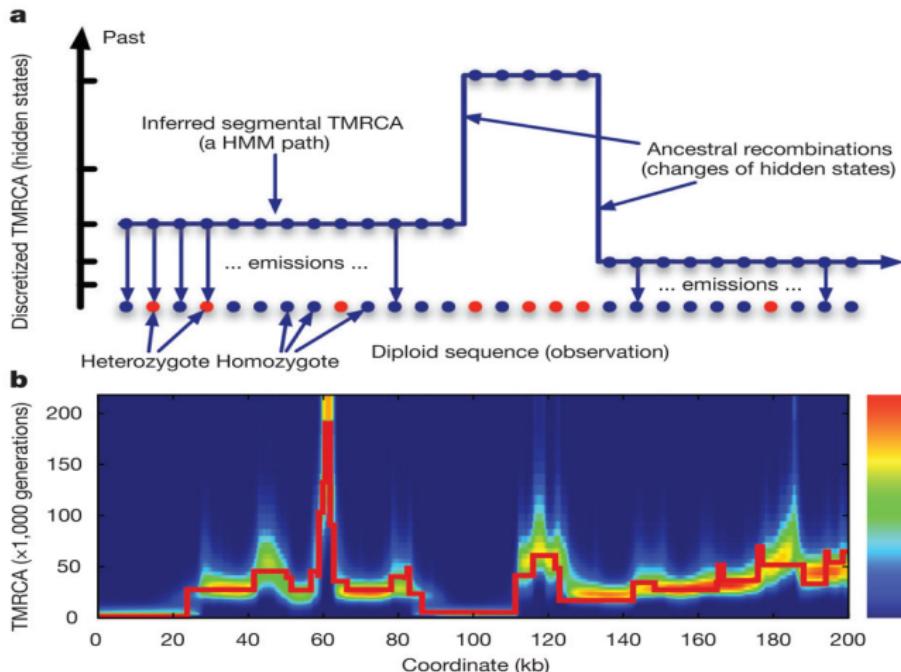
Outline

inferring admixture (*GLOBETROTTER*)

inferring pop size changes (*PSMC, MSMC*)

Inferring population size changes over time

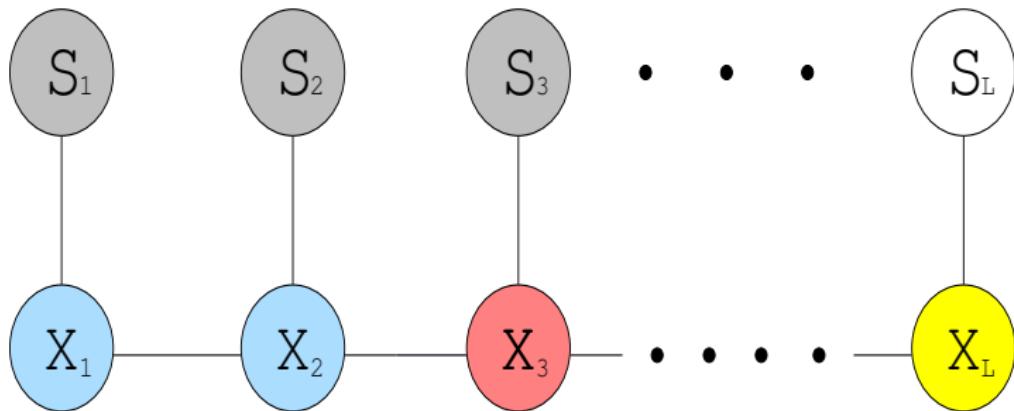
- ▶ aim is to infer how population sizes changed over time (e.g. due to expansions/bottlenecks)
- ▶ powerful technique to do so: Pairwise Sequential Markovian Coalescent (PSMC) (Li & Durbin 2011, *Nature* **475**:493)
 - ▶ requires sequencing data → use mutations as a “clock” to infer time
 - ▶ to convert to years, must assume mutation rate per generation → some debate over appropriate rate to use (e.g. Scally & Durbin 2012, *Nature Rev Genet* **13**:745)



- ▶ consider a single diploid individual: infer TMRCA (The Most Recent Common Ancestor) of individuals' two haplotypes
- ▶ **switches in TMRCA** depends on recombination rate
- ▶ **How to determine TMRCA?** lots of heterozygotes in region → higher TMRCA

PSMC – HMM (Li & Durbin 2011, *Nature* 475:493)

- ▶ S_l = (observed state) whether 100bp bin l is heterozygous or not
- ▶ X_l = (hidden state) time to most recent common ancestor (TMRCA) at 100bp bin l



- ▶ each TMRCA is depicted with a unique color here
- ▶ assume “switches” in TMRCA depend on recombination rate ρ
- ▶ “switches” also depend on TMRCA

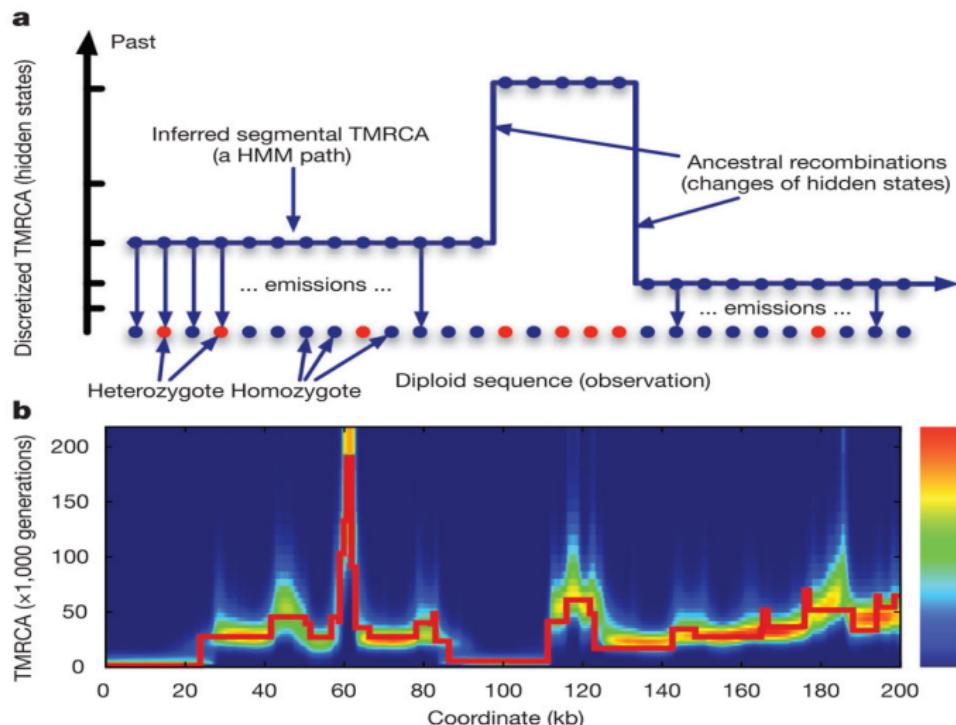
PSMC – HMM (Li & Durbin 2011, *Nature* 475:493)

Observed State (for given 100bp bin):

- ▶ “missing”: $\geq 90\text{bp}$ uncalled (SAMtools) or filtered
- ▶ “heterozygous”: $> 10\text{bp}$ called and ≥ 1 heterozygote
- ▶ “homozygous”: else
- ▶ $\Pr(\text{missing} \mid t) = 1$
- ▶ $\Pr(\text{heterozygous} \mid t) = \exp^{-\theta t}$ (??)
- ▶ $\Pr(\text{homozygous} \mid t) = 1 - \exp^{-\theta t}$ (??)
- ▶ $\theta = 4N_0\mu$ (with $\mu = 2.5 \times 10^{-8}$)

Hidden State: (switches in TMRCA from $s \rightarrow t$ between two 100bp bins; ρ =recombination rate):

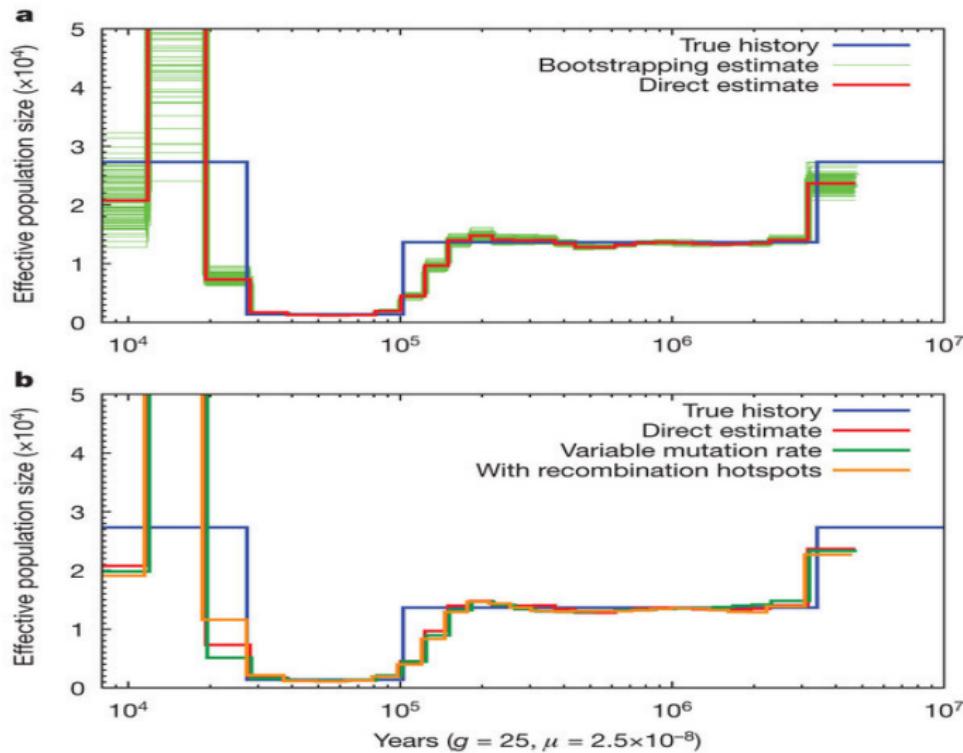
- ▶ $\Pr(t \mid s) = (1 - \exp^{-\rho t})q(t \mid s) + \exp^{-\rho s} \delta(t - s)$
- ▶ $q(t \mid s) = \Pr(t \mid s, \text{recomb}) = \frac{1}{\lambda(t)} \int_0^{\min(s,t)} \left[\frac{1}{s} \exp^{-\int_u^t \frac{1}{\lambda(v)} dv} \right] du$
- ▶ $\lambda(t) = \frac{N_e(t)}{N_0}$
- ▶ $\delta(t - s) = \text{Dirac-delta function}$



Aim: infer (effective) population size at different times in the past

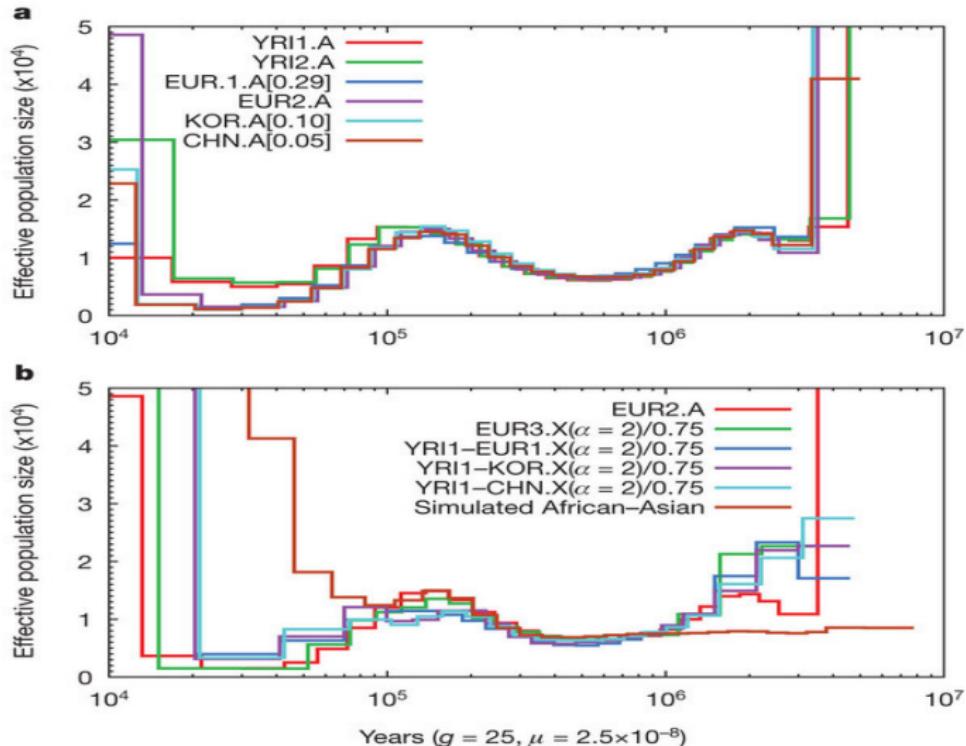
How? if lots of genome segments have the same TMRCA t — decrease in population size at t

PSMC – simulations (Li & Durbin 2011, *Nature* 475:493)

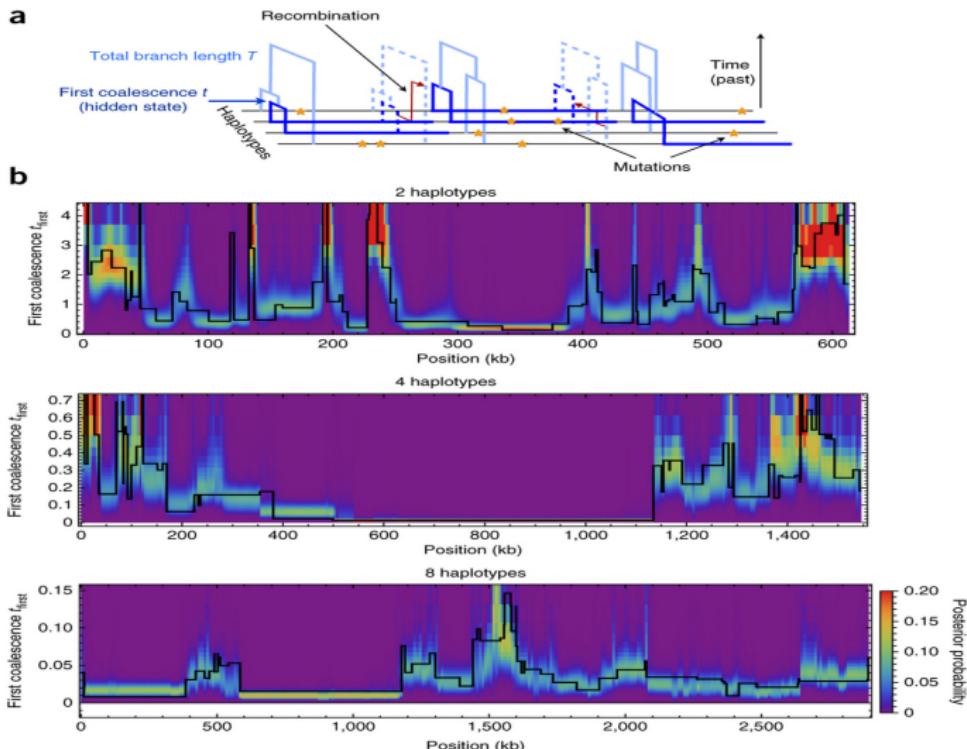


- ▶ nicely tracks population size over time (ms simulation), except at recent dates
- ▶ **green lines** – bootstrap re-sampling of 5Mb genetic regions

PSMC – results (Li & Durbin 2011, *Nature* 475:493)



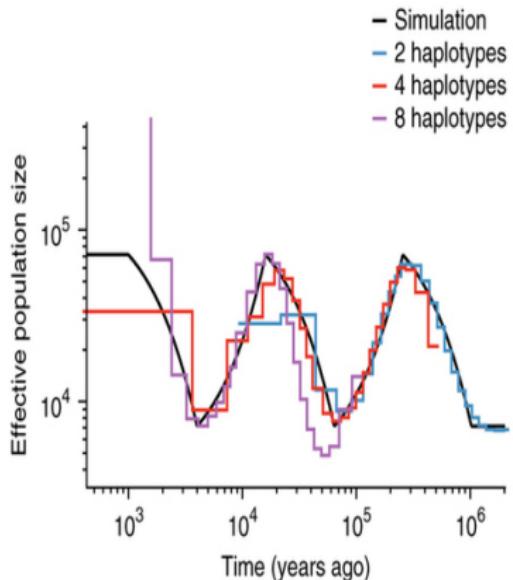
- (top:) dip in population size from around 100-150kya to about 40-20kya
- (bottom:) N_e between groups (e.g. Africa-Europe) very low relative to **simulation** where populations split 60kya



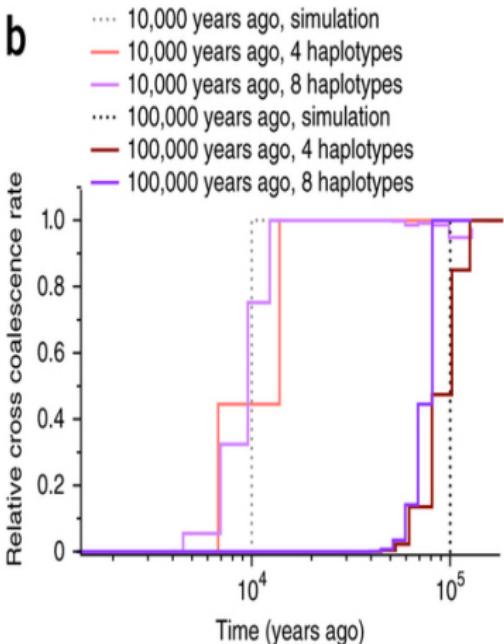
- ▶ consider 2-4 individuals, and find *first coalescence* in each genetic region
- ▶ first coalescence becomes more recent (and spans longer regions) as number of (phased) haplotypes increases (i.e. $E[t] = \frac{2}{n(n-1)}$)

MSMC – Simulations (Schiffels & Durbin 2014, *Nature Genetics* 46:919)

a

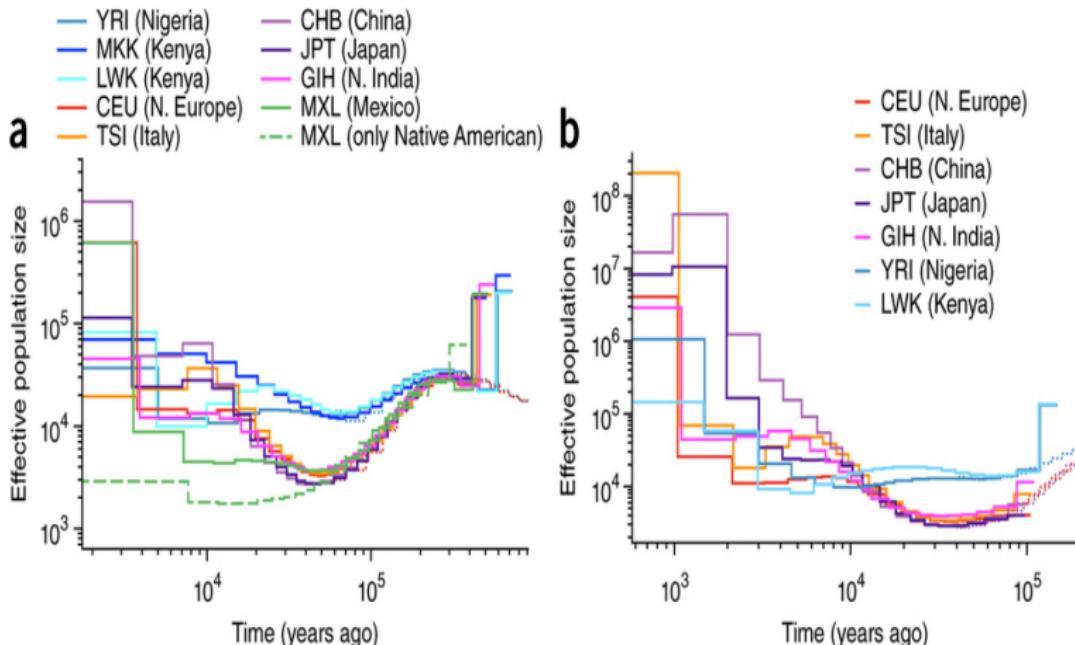


b



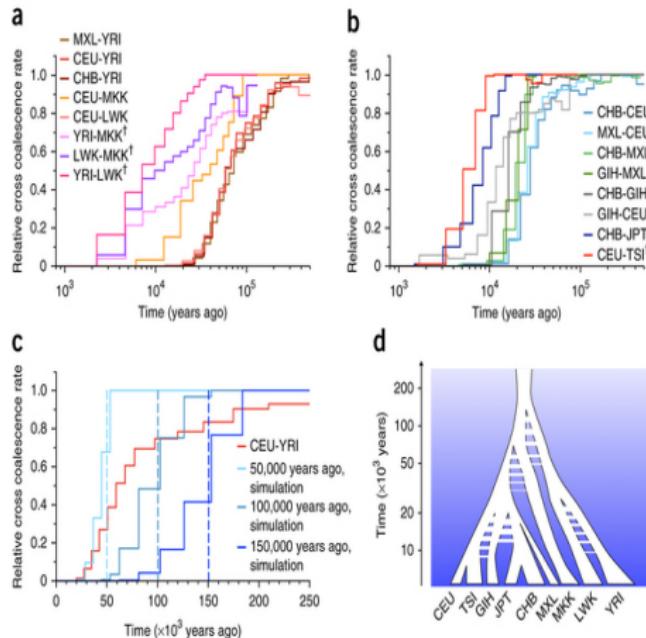
- ▶ (left:) using 4,8 haplotypes is more accurate for recent population sizes
- ▶ (right:) 8 haplotypes gives better estimates for more recent split; 4 haplotypes better for older split

MSMC – Results (Schiffels & Durbin 2014, *Nature Genetics* 46:919)



- ▶ (left:) estimates using 4 haplotypes per population
- ▶ (right:) estimates using 8 haplotypes per population

MSMC – Results (Schiffels & Durbin 2014, *Nature Genetics* 46:919)



- ▶ **(top left:)** cross-coalescence rates between YRI and non-Africans drop 40-200Kya
- ▶ **(bottom left:)** gradual decline not consistent with “clean” split

Summary

- ▶ using haplotypes can increase power, e.g. for:
 1. inferring admixture (*GLOBETROTTER*)
 2. inferring population size changes over time and population splits (*PSMC, MSMC*)
- ▶ can be computationally demanding; must infer haplotypes (phase)
- ▶ other programs (instead of *PSMC/MSMC*):
 - ▶ *diCal* (Sheehan et al 2013, *Genetics* **194**:647)
 - ▶ *SMC++* (Terhorst et al 2017, *Nature Genetics* **49**:303)