Dating admixture and detecting selection in admixed individuals

Garrett Hellenthal
g.hellenthal@ucl.ac.uk

University College London

Population genetics summer course, Denmark
August 23, 2024

Introduction

This lecture/practical will cover:

- methods for detecting and dating admixture events between two or more source groups
  1. *ALDER / MALDER* (Loh et al 2013, Pickrell et al 2014)
  2. *GLOBETROTTER, fastGLOBETROTTER*
     (Hellenthal et al 2014, Wangkumhang et al 2021)
  3. *MOSAIC* (Salter-Townshend & Myers 2019)

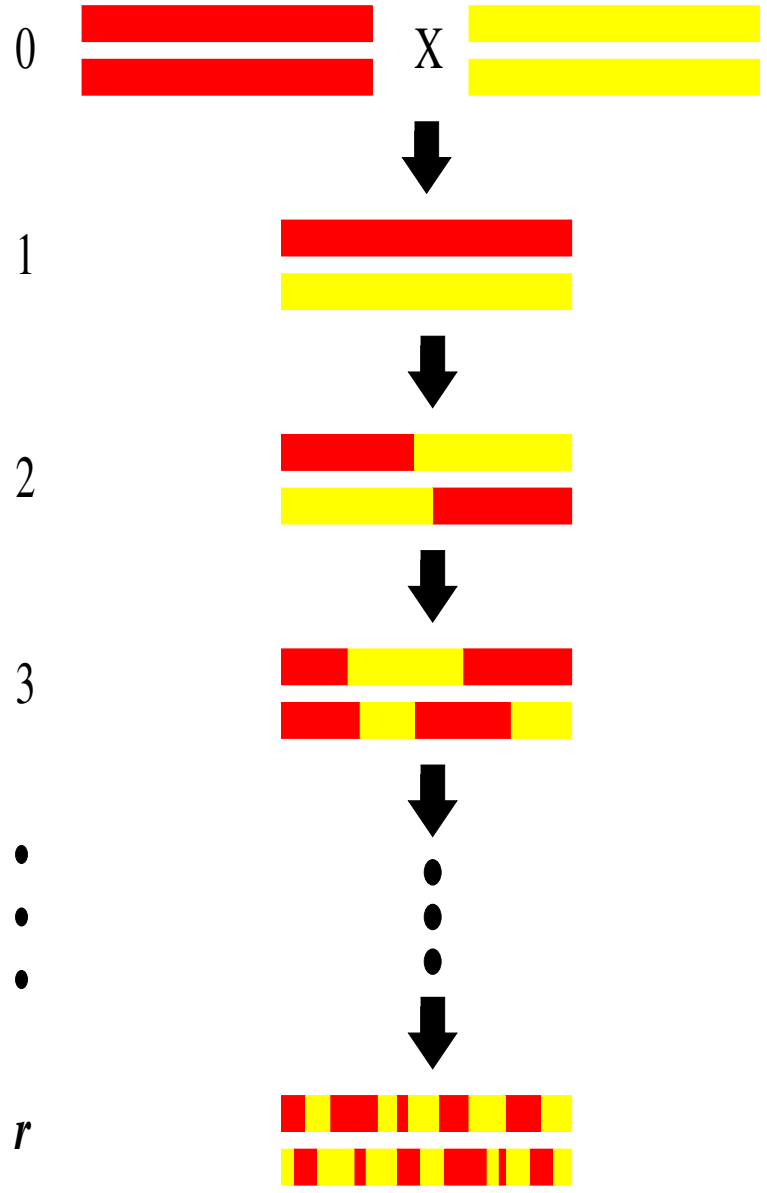
- methods for detecting selection in admixed individuals
  1. *AdaptMix* (Mendoza-Revilla 2022)

Outline

Outline

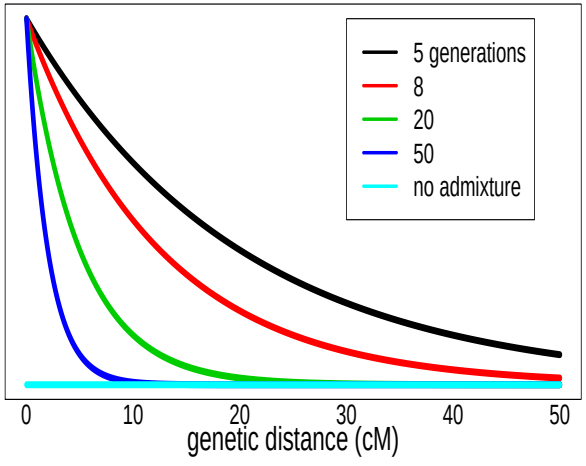
Inferring & dating admixture using autosomal DNA

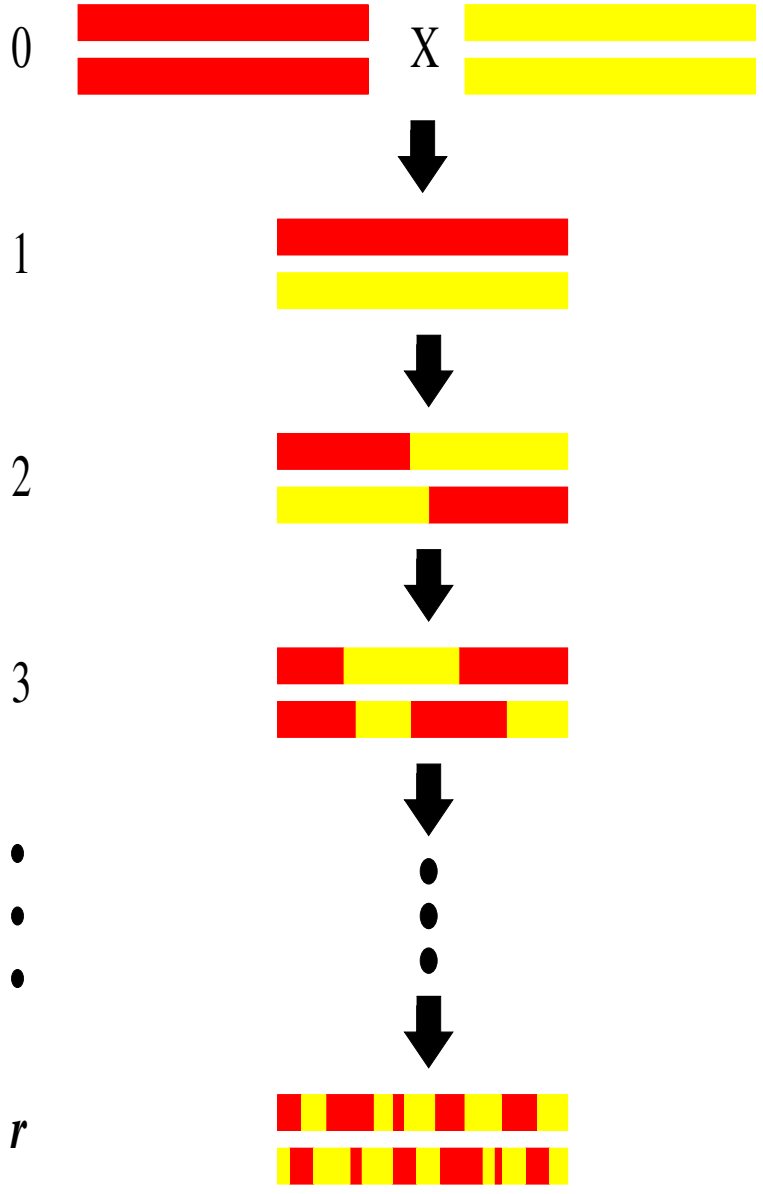
- two populations (**red**,**yellow**) admix *r* generations ago; then the admixed pop randomly mates
- genetic pieces from each population get smaller each subsequent generation due to recombination



Inferring & dating admixture using autosomal DNA

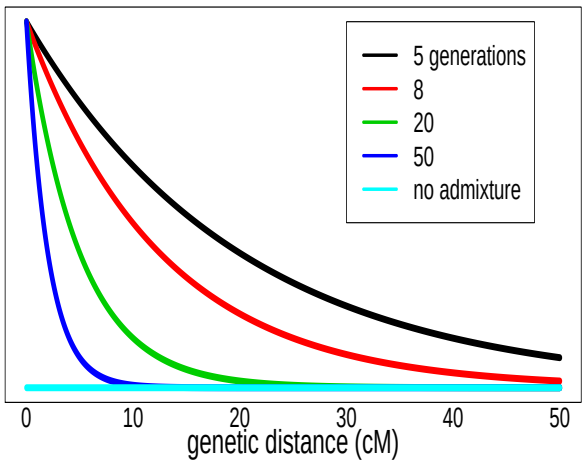
- two populations (**red**,**yellow**) admix *r* generations ago; then the admixed pop randomly mates
- genetic pieces from each population get smaller each subsequent generation due to recombination
- cM size of contiguous **red** and **yellow** segments follow exponential distribution with rate *r*:

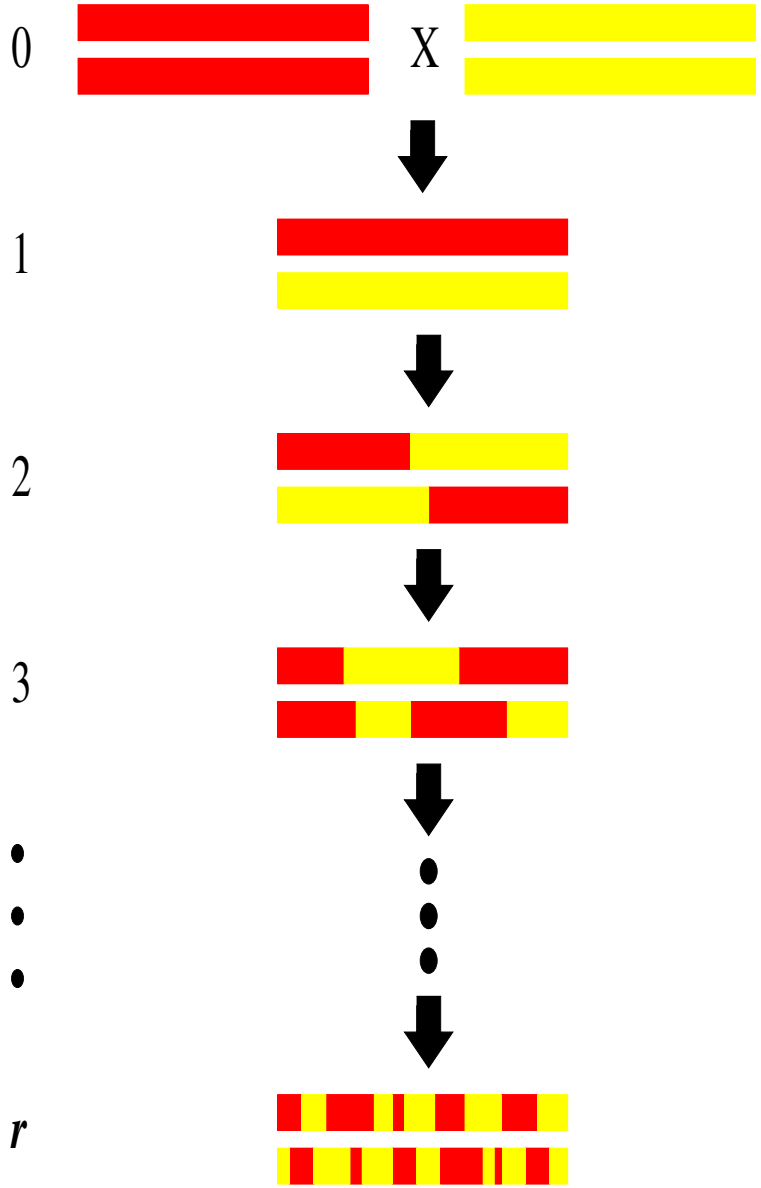




Inferring & dating admixture using autosomal DNA

- two populations (**red**,**yellow**) admix *r* generations ago; then the admixed pop randomly mates
- genetic pieces from each population get smaller each subsequent generation due to recombination
- cM size of contiguous **red** and **yellow** segments follow exponential distribution with rate *r*:





inferring exact **yellow** and **red** segments is difficult $\longrightarrow$ use *admixture LD*

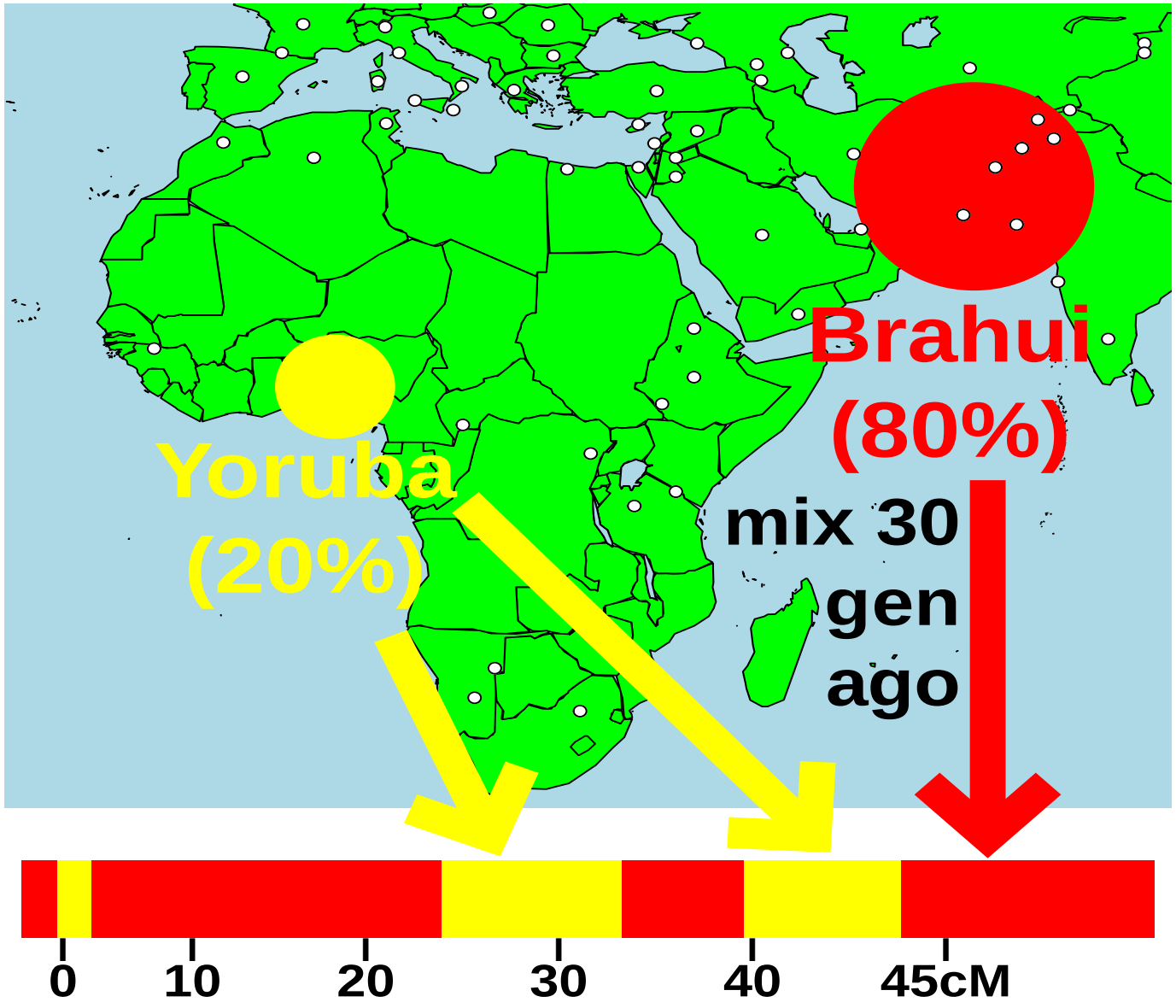
Simulated Example

We will use a simulation to illustrate each approach.

- 20 admixed “individuals”
- descend from admixture occurring 30 generations ago

  (Price et al 2009, Hellenthal et al 2014)
- **80%** of DNA from **Brahui** from Pakistan
- **20%** from **Yoruba** from Nigeria
- will use proxy (surrogate) populations to represent each admixing source:
  - **Brahui** $\longrightarrow$ Balochi
  - **Yoruba** $\longrightarrow$ BantuKenya, BantuSouthAfrica, Mandenka

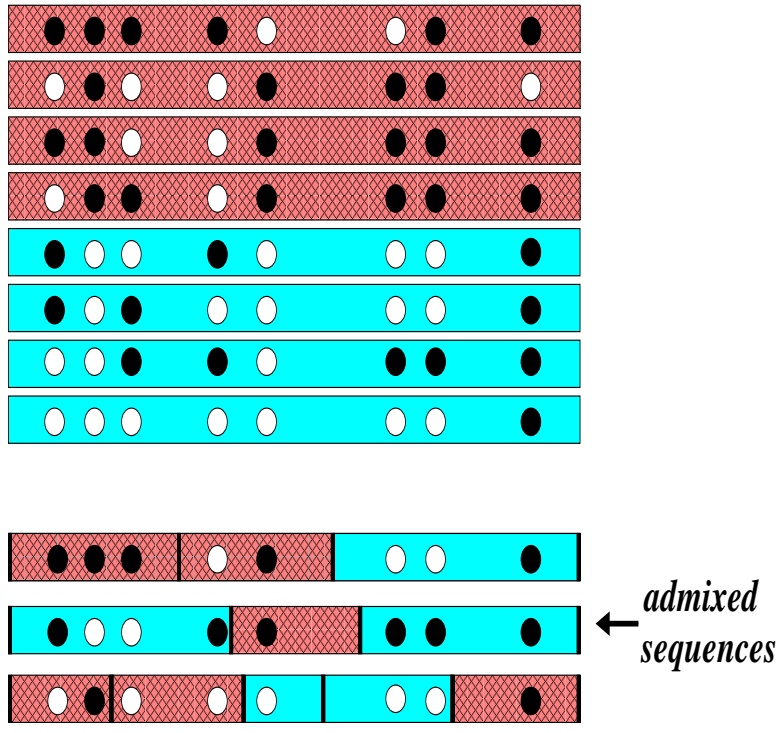
Simulation: 80% **Brahui** + 20% **Yoruba**, 30gen



(Hellenthal et al 2014, *Science* **343**:747)

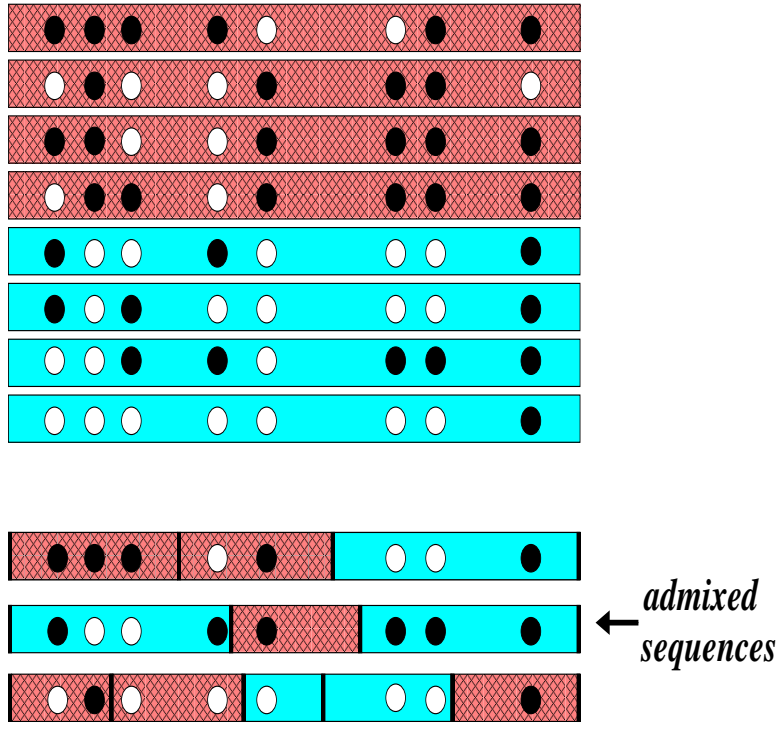
Outline

ALDER/MALDER (Loh et al 2013, Pickrell et al 2014)



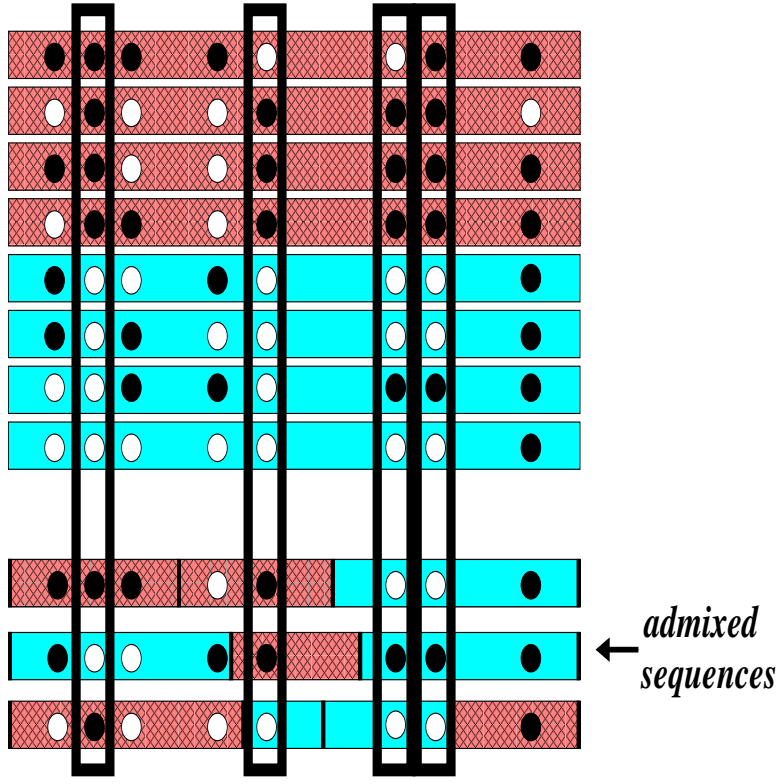
- pick two **surrogate populations (*A*, *B*)** (**red**,**cyan**)
- find covariance (cov($\vec{x}$, $\vec{y}$)) among unlinked SNP pairs (x, y) in admixed pop

ALDER/MALDER (Loh et al 2013, Pickrell et al 2014)



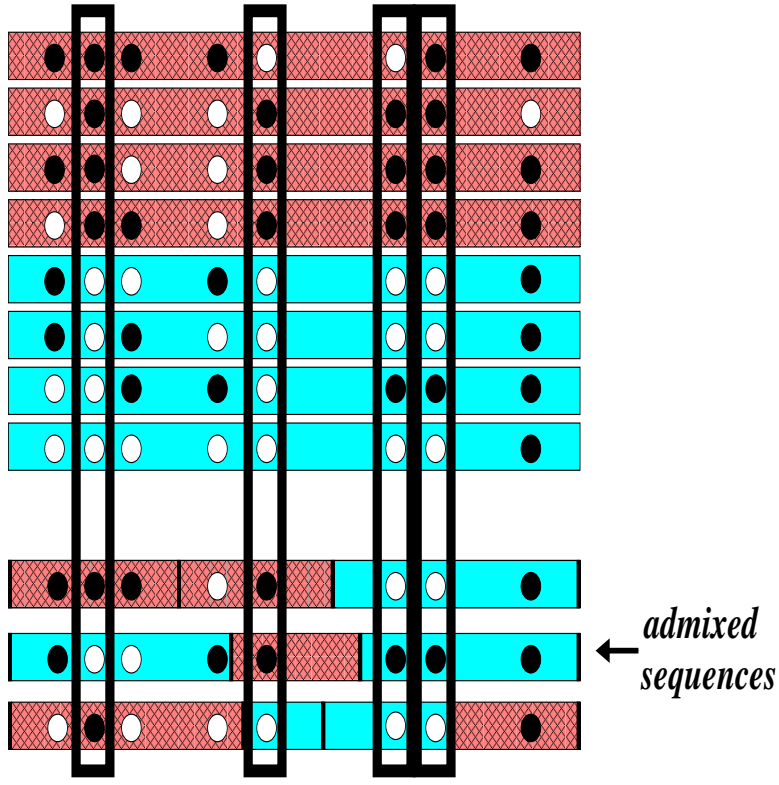
- pick two **surrogate populations (*A*, *B*)** (**red**,**cyan**)
- find covariance ($\mathrm{cov}(\vec{x}, \vec{y})$) among unlinked SNP pairs (x, y) in admixed pop
- upweight SNPs where surrogates are distinguishable
  $\longrightarrow$ i.e. where $|x_A - x_B|$ is big, with x_A the allele frequency for pop A at SNP x

ALDER/MALDER (Loh et al 2013, Pickrell et al 2014)



- pick two **surrogate populations (*A*, *B*)** (**red**,**cyan**)
- find covariance ($\mathrm{cov}(\vec{x}, \vec{y})$) among unlinked SNP pairs (x, y) in admixed pop
- upweight SNPs where surrogates are distinguishable
  $\longrightarrow$ i.e. where $|x_A - x_B|$ is big, with x_A the allele frequency for pop A at SNP x

ALDER/MALDER (Loh et al 2013, Pickrell et al 2014)



- pick two **surrogate populations** (*A*, *B*) (**red**,**cyan**)
- find covariance (cov($\vec{x}$, $\vec{y}$)) among unlinked SNP pairs (x, y) in admixed pop
- upweight SNPs where surrogates are distinguishable
  $\longrightarrow$ i.e. where $|x_A - x_B|$ is big, with x_A the allele frequency for pop A at SNP x

$$\text{cov}(\vec{x}, \vec{y}) = 2\alpha(1-\alpha)(x_A - x_B)(y_A - y_B)\exp^{-rg} \qquad \longrightarrow \text{infer } (r, \alpha)$$

running *ALDER*

Input files: (e.g. specified in "BrahuiYorubaSimulation.alder.par")

1. "BrahuiYorubaSimulation.smartpca.geno" – 1,466 individuals & 26,157 SNPs (chr 20-22)

2. "BrahuiYorubaSimulation.smartpca.snp" – 26,157 SNPs (row = rsID,chromo,cM,bp,allele0,allele1)

3. "BrahuiYorubaSimulation.smartpca.ind" – 1,466 individuals (row = ID,sex,population)

4. "BrahuiYorubaSimulation.alder.par":

```
genotypename:    data/BrahuiYorubaSimulation.smartpca.geno
snpname:         data/BrahuiYorubaSimulation.smartpca.snp
indivname:       data/BrahuiYorubaSimulation.smartpca.ind
raw_outname:     data/BrahuiYorubaSimulation.smartpca.LDoutput
admixpop: BrahuiYorubaSimulation
refpops: Balochi;Mandenka
```

running *MALDER*

- MALDER can assign different dates to different surrogate pairs
  - Different pairings may have different admixture dates (multiple waves of admixture)

- Input files: (e.g. specified in "`BrahuiYorubaSimulation.malder.par`")
  1. "`BrahuiYorubaSimulation.smartpca.geno`" – 1,466 individuals & 26,157 SNPs (chr 20-22)
  2. "`BrahuiYorubaSimulation.smartpca.snp`" – 26,157 SNPs (row = rsID,chromo,cM,bp,allele0,allele1)
  3. "`BrahuiYorubaSimulation.smartpca.ind`" – 1,466 individuals (row = ID,sex,population)
  4. "`BrahuiYorubaSimulation.malder.par`":

```
genotypename:    data/BrahuiYorubaSimulation.smartpca.geno
snpname:         data/BrahuiYorubaSimulation.smartpca.snp
indivname:       data/BrahuiYorubaSimulation.smartpca.ind
raw_outname:     data/BrahuiYorubaSimulation.malder.LDoutput
admixpop: BrahuiYorubaSimulation
refpops: Balochi;Mandenka;BantuSouthAfrica
```

ALDER/MALDER output plots

`ALDER/MALDER` output files, e.g.:

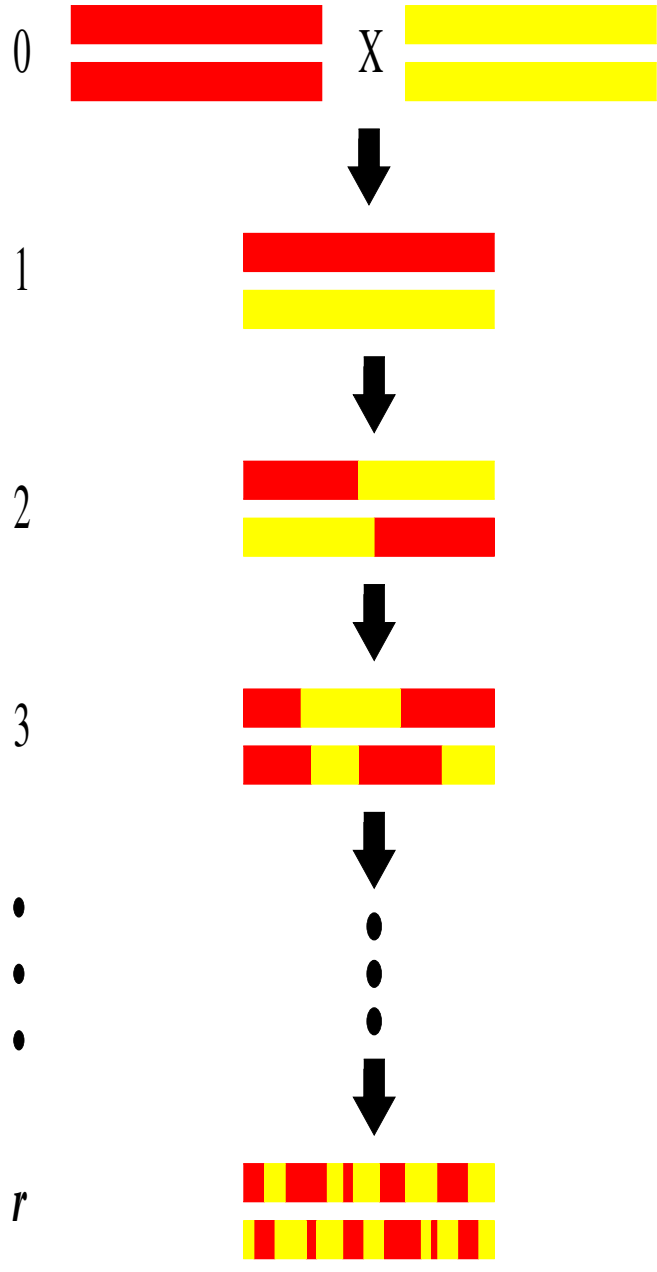
1. "`[screen output]`" – details of the admixture fit (inferred date and standard errors, significance) plus plots of the LD decay curves

2. "`[raw_outname]`" specified in `.par` file – gives data used to generate plots of the LD decay curves

Outline

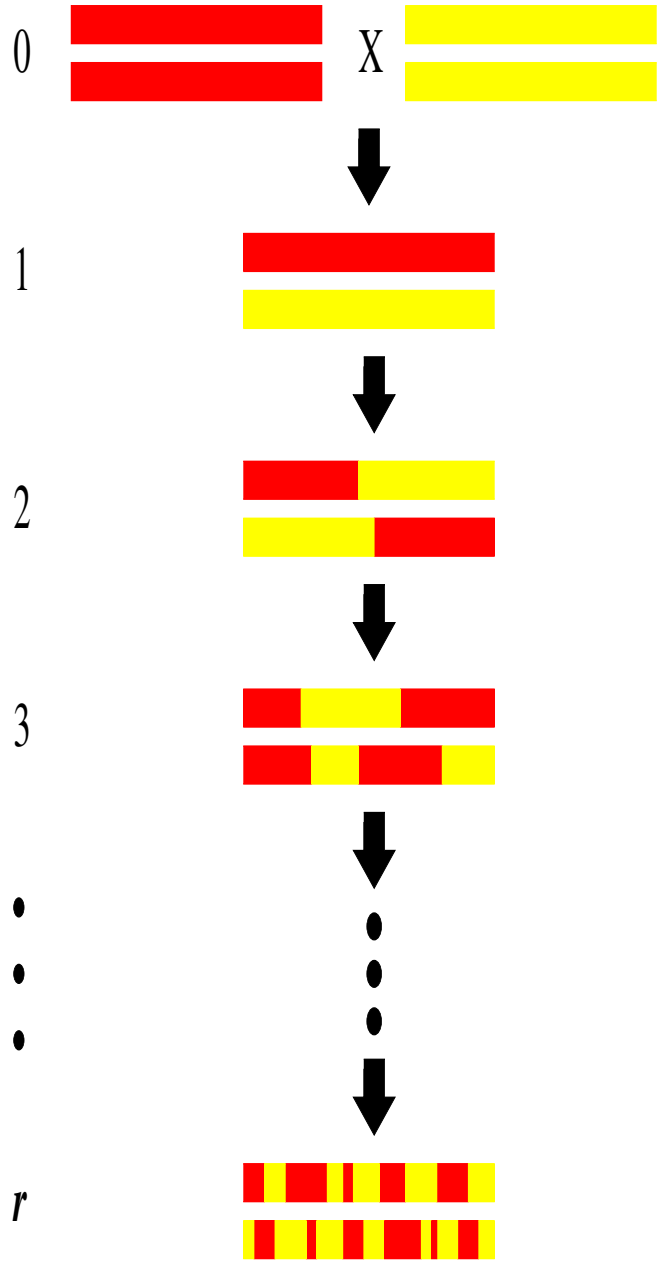
Inferring & dating admixture using autosomal DNA

- two pops (**“red”** and **“yellow”**) intermix *r* generations ago



Inferring & dating admixture using autosomal DNA

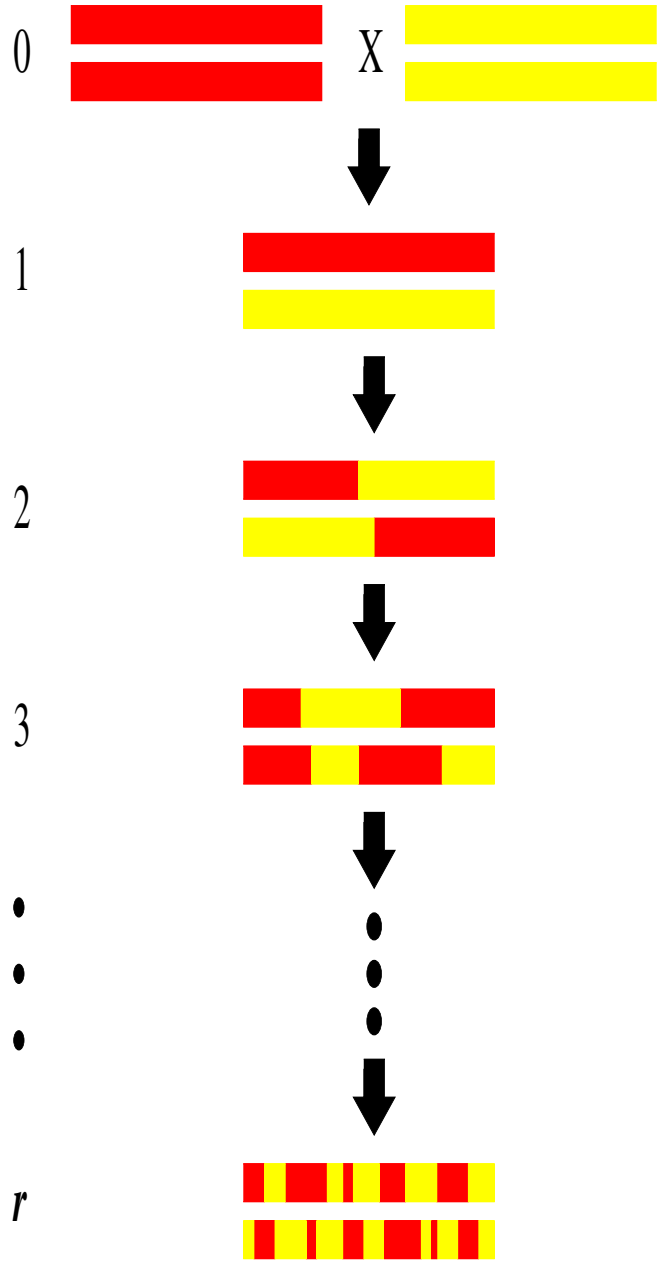
- two pops (**"red"** and **"yellow"**) intermix r generations ago
- random mating since admixture $\longrightarrow$ Poisson(r/Morgan) process of ancestry switches ($\bullet \rightarrow \bullet$) at gen r



Inferring & dating admixture using autosomal DNA

- two pops (“**red**” and “**yellow**”) intermix r generations ago
- random mating since admixture $\longrightarrow$ Poisson(r/Morgan) process of ancestry switches ($\bullet \rightarrow \bullet$) at gen r
- α = % DNA from **red** population
- d = distance (M) between two genomic locations at generation r

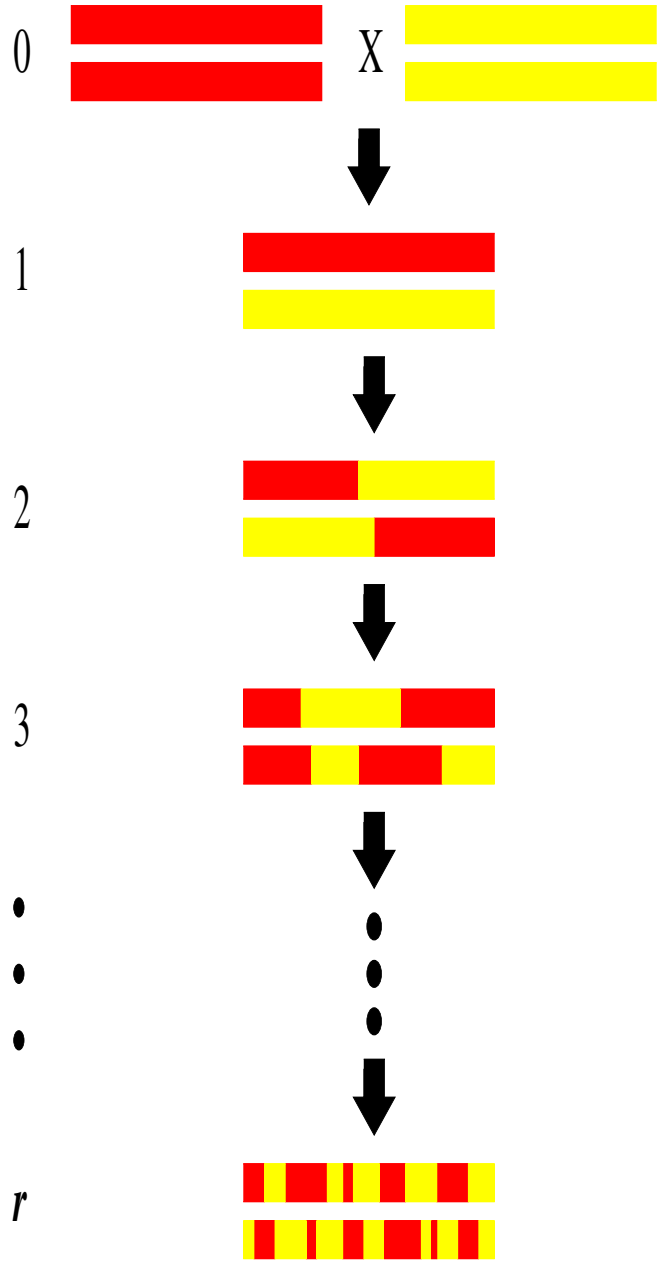
$$\Pr(\bullet \rightarrow \bullet; d) = \alpha \exp^{-dr} + \alpha^2(1 - \exp^{-dr})$$
$$\Pr(\bullet \rightarrow \bullet; d) = \alpha(1 - \alpha)(1 - \exp^{-dr})$$



Inferring & dating admixture using autosomal DNA

- two pops (“**red**” and “**yellow**”) intermix r generations ago
- random mating since admixture $\longrightarrow$ Poisson(r/Morgan) process of ancestry switches ($\bullet \rightarrow \bullet$) at gen r
- α = % DNA from **red** population
- d = distance (M) between two genomic locations at generation r

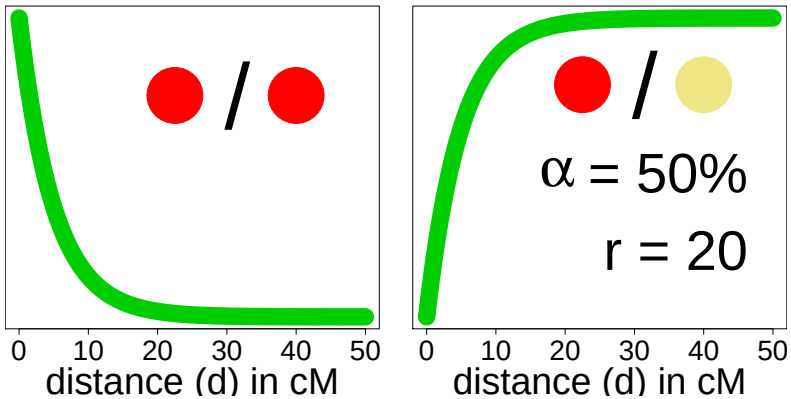
$$\Pr(\bullet \rightarrow \bullet; d) = \alpha^2 + \alpha(1-\alpha)\exp^{-dr}$$
$$\Pr(\bullet \rightarrow \bullet; d) = \alpha(1-\alpha)(1-\exp^{-dr})$$

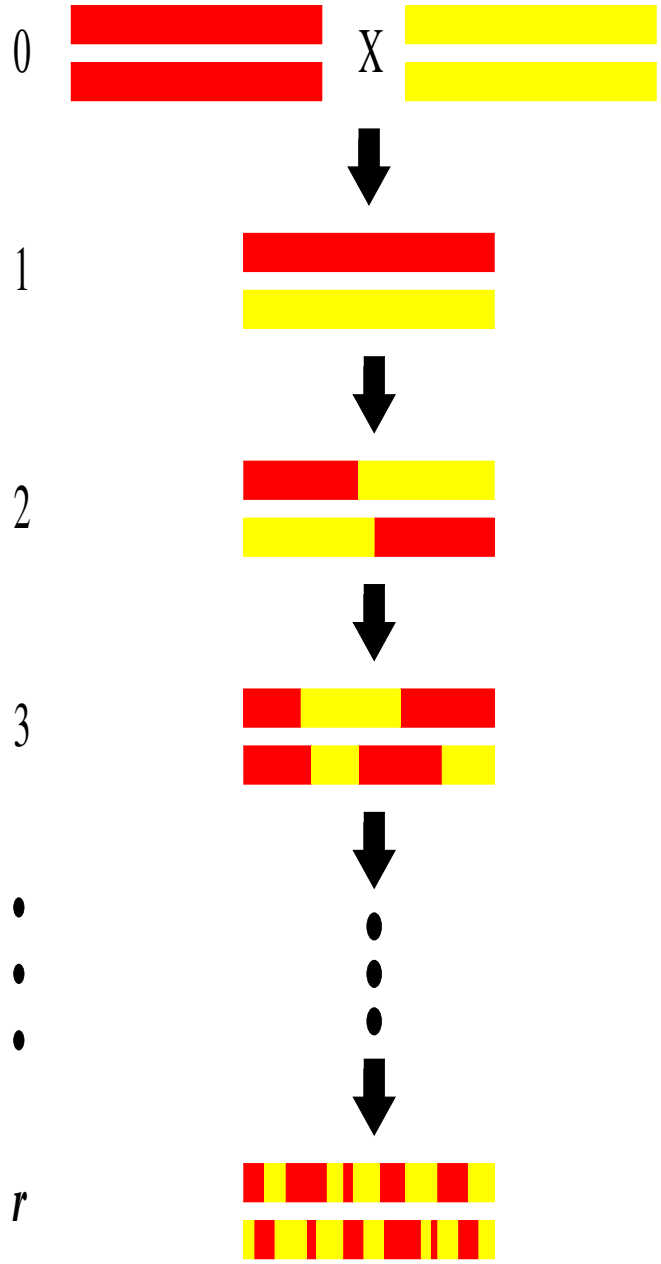


Inferring & dating admixture using autosomal DNA

- two pops (“**red**” and “**yellow**”) intermix r generations ago
- random mating since admixture $\longrightarrow$ Poisson(r/Morgan) process of ancestry switches ($\bullet \rightarrow \bullet$) at gen r
- α = % DNA from **red** population
- d = distance (M) between two genomic locations at generation r

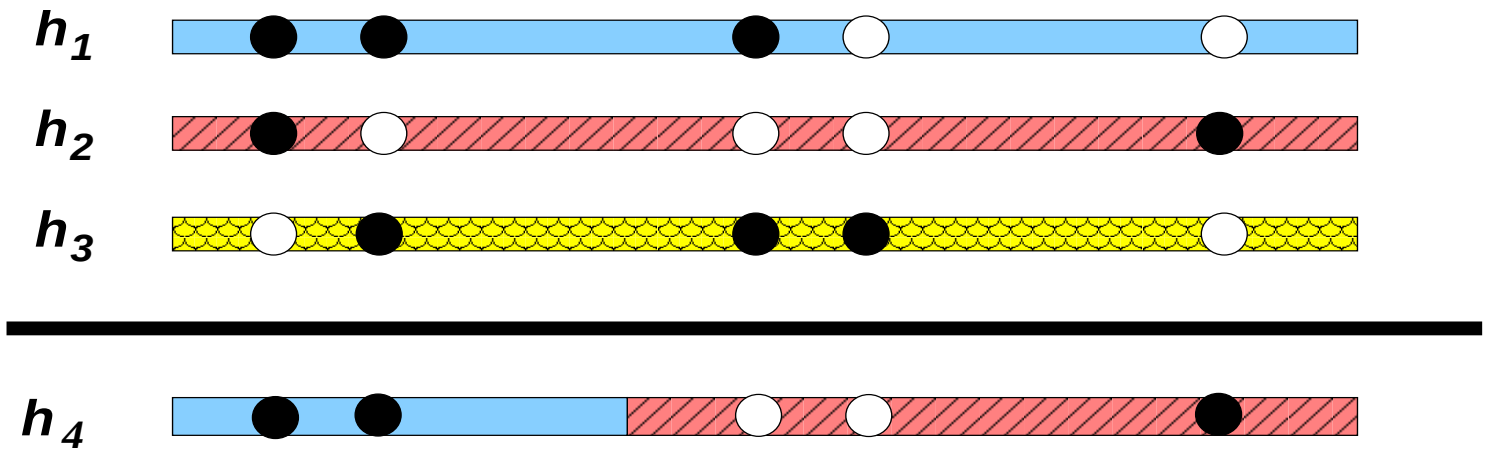
$$\Pr(\bullet \rightarrow \bullet; d) = \alpha^2 + \alpha(1-\alpha)\exp^{-dr}$$
$$\Pr(\bullet \rightarrow \bullet; d) = \alpha(1-\alpha)(1-\exp^{-dr})$$





GLOBTROTTER / fastGLOBETROTTER

(Hellenthal et al 2014, Wangkumhang et al 2022)

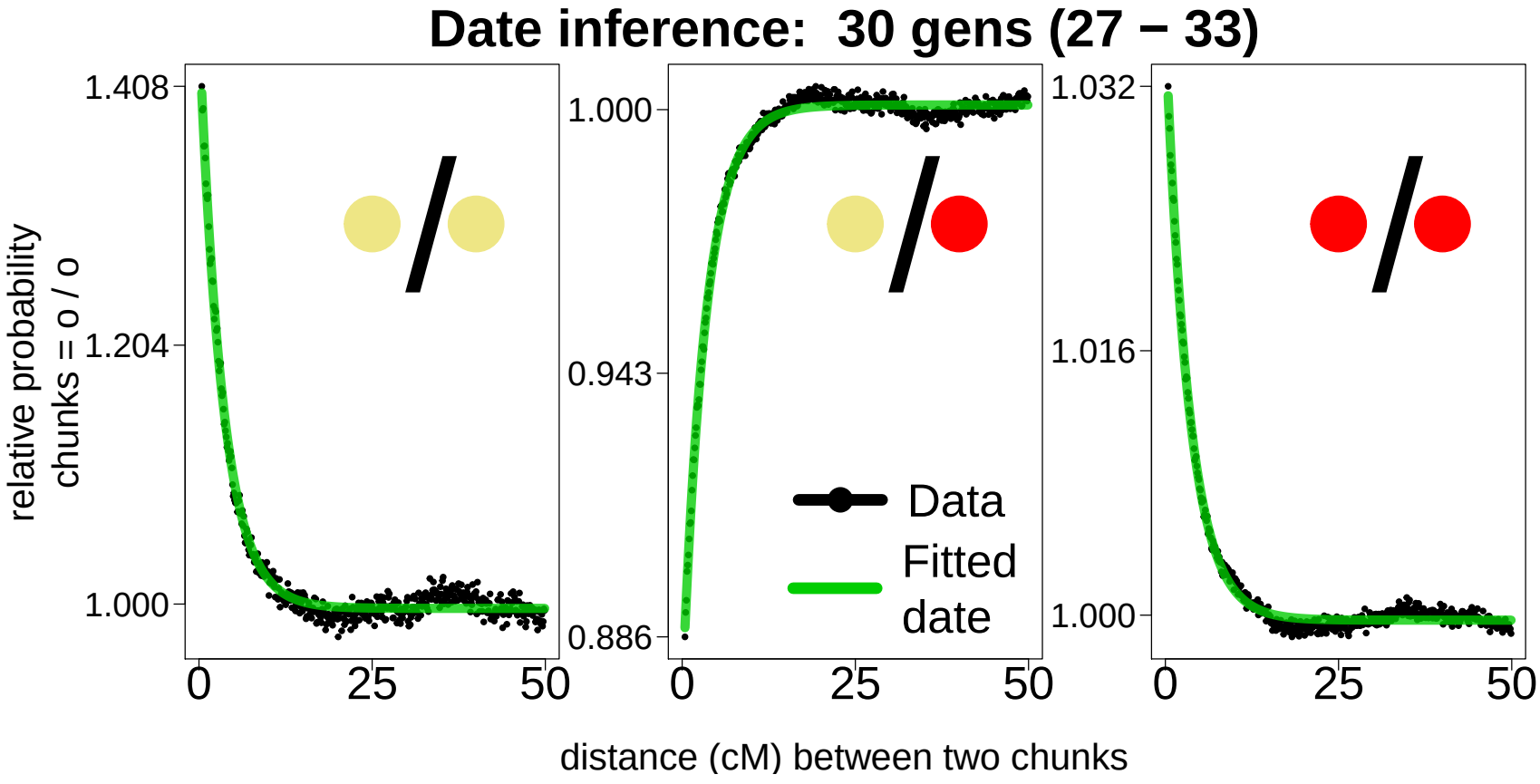


- match (or "paint") admixed individuals' sequences (e.g. h_4) as mixtures of those from surrogate populations (e.g. h_1, h_2, h_3)
  $\longrightarrow$ *CHROMOPAINTER* (Lawson et al 2012)

- $\Pr(\bullet \rightarrow \bullet; d)$ informs on:
  1. evidence of admixture $\longrightarrow$ is there an exponential curve?
  2. date of admixture $\longrightarrow$ rate of exponential curve?
  3. do $(\bullet, \bullet)$ represent same admixing source$\longrightarrow$ does curve decrease?
  4. do $(\bullet, \bullet)$ represent different sources $\longrightarrow$ does curve increase?

- incorporates haplotype information, which may increase power

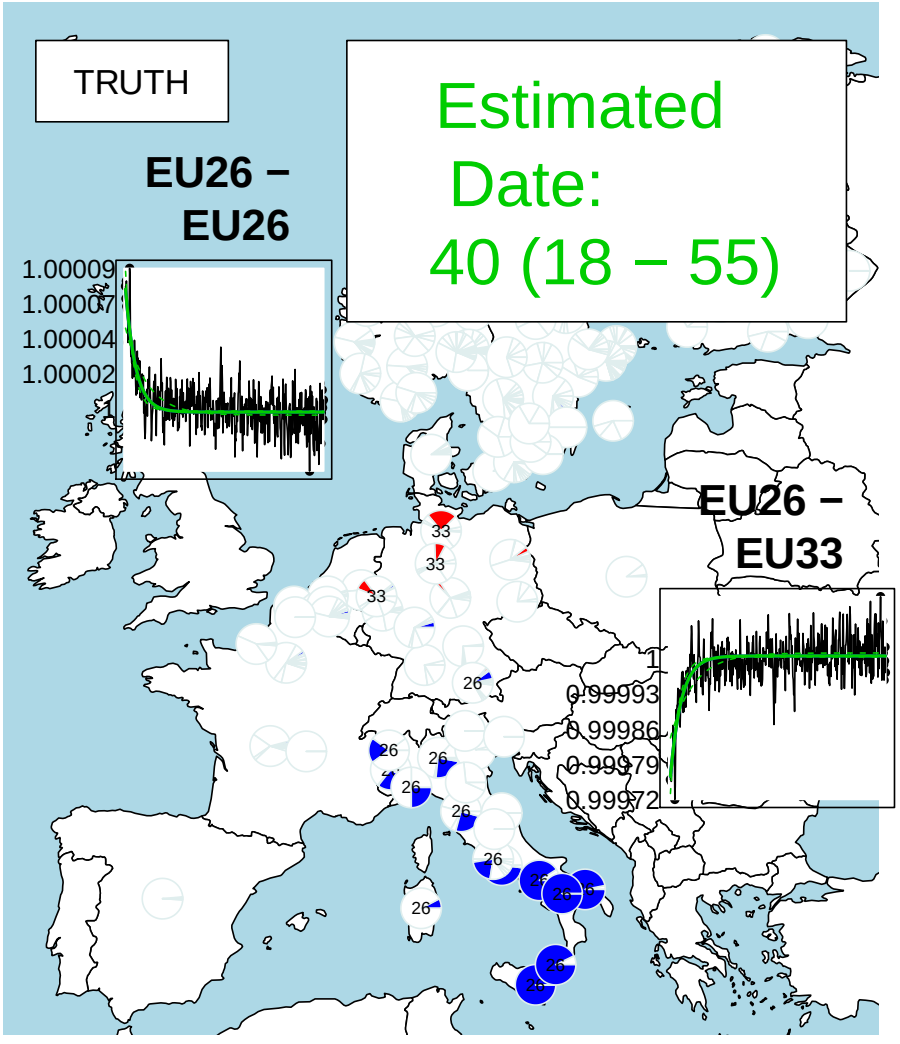
Dating Admixture (Sim: 80% **Brahui** + 20% **Yoruba**, 30gen)

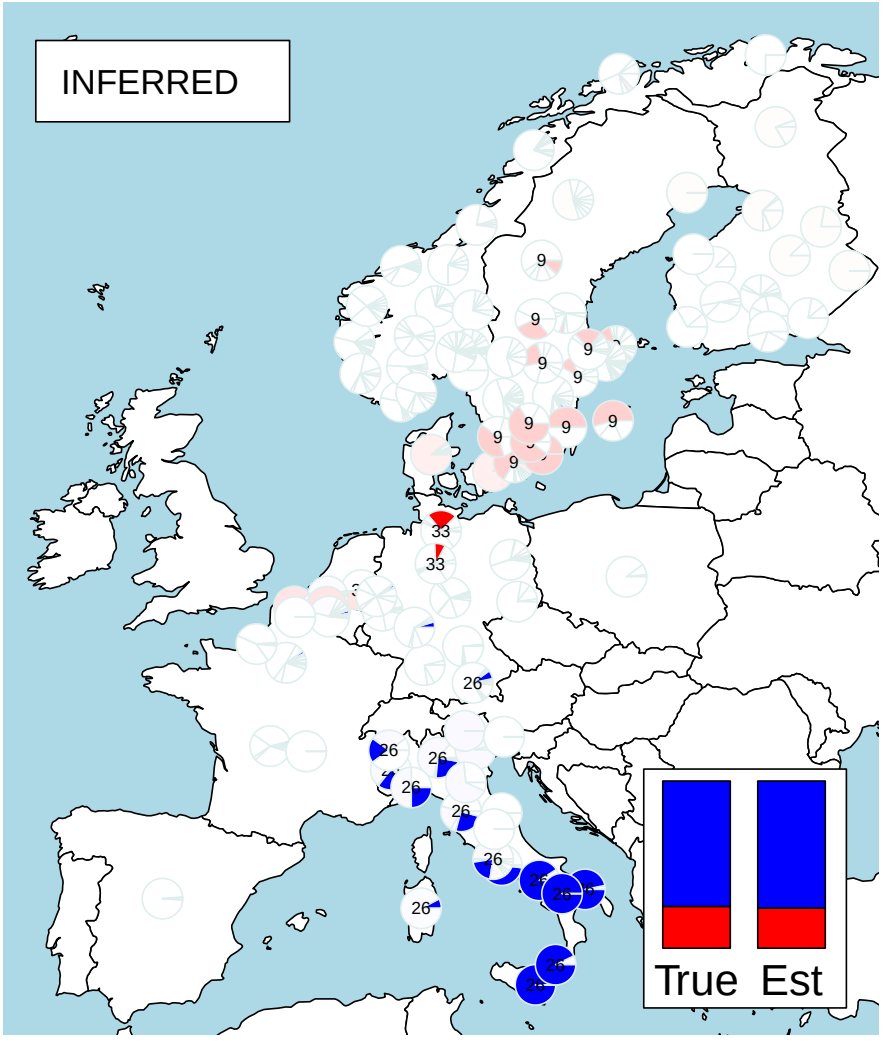


Coancestry curves:

- **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)

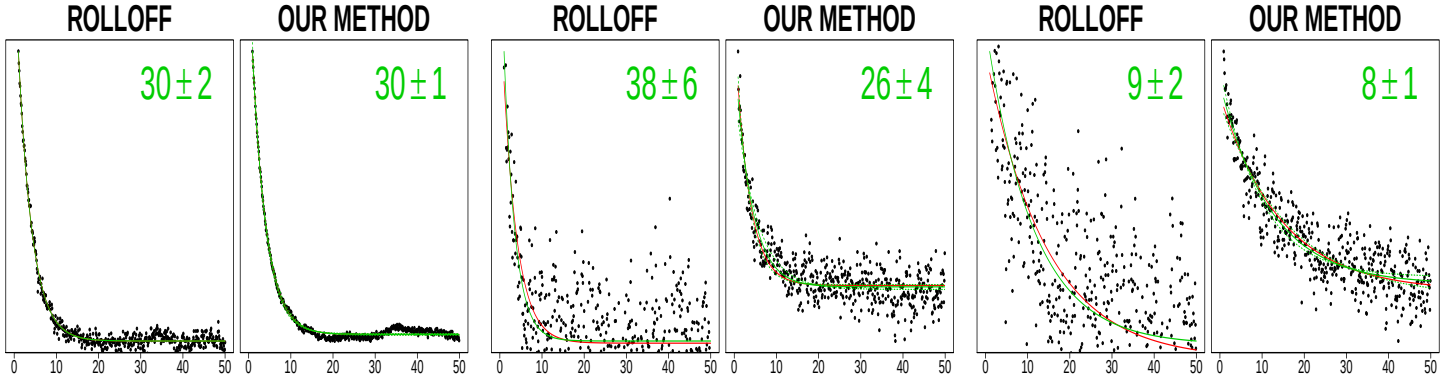




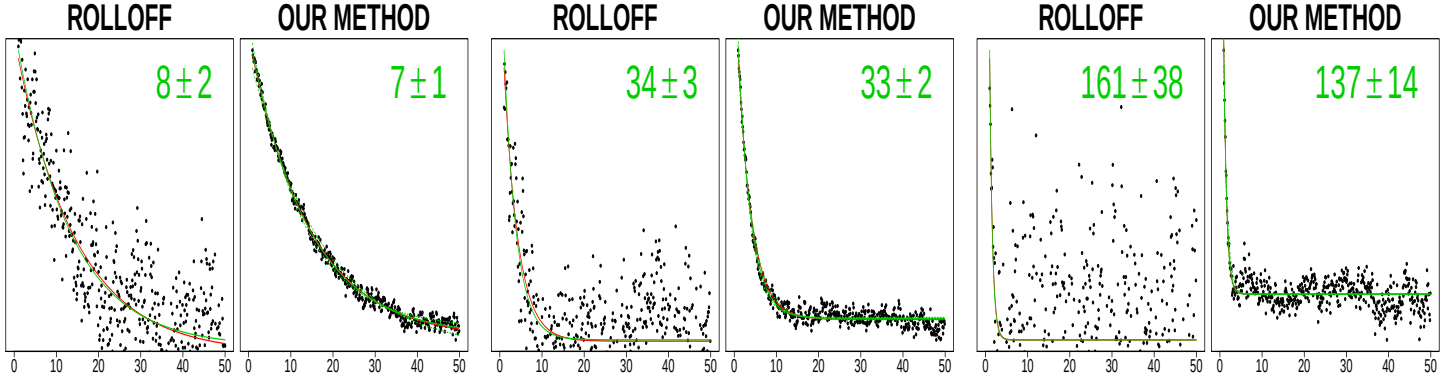
(Leslie et al 2015, *Nature* **519**:309)

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

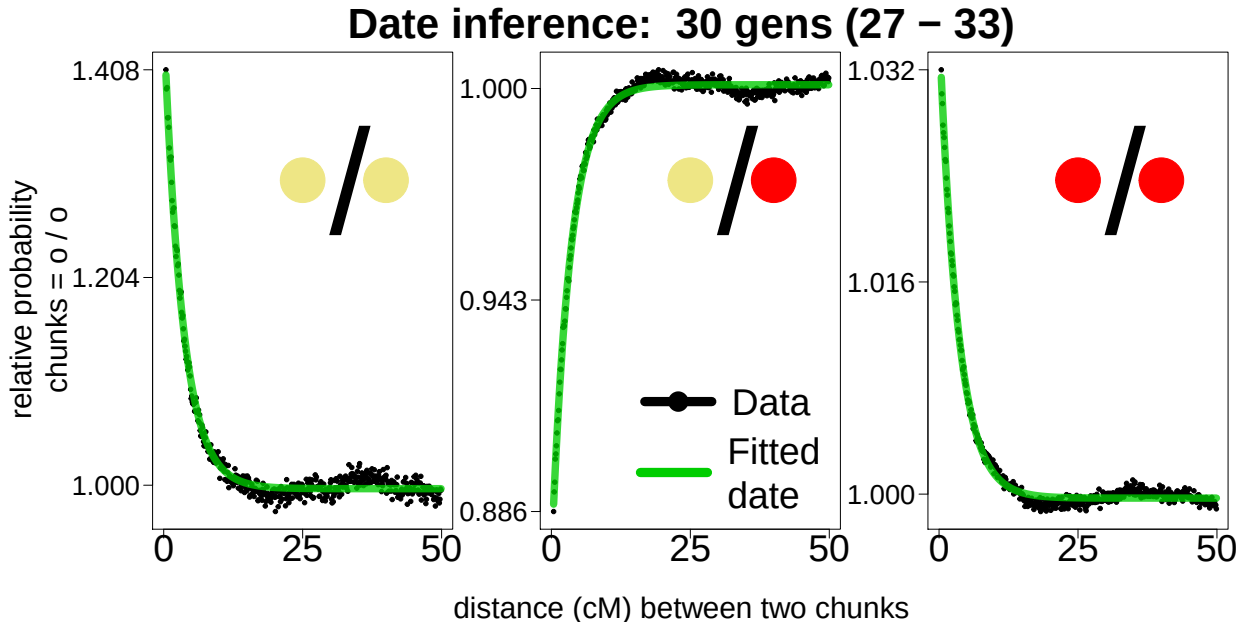


Brahui-Yoruba (20%) French-Brahui (50%) French-Brahui (20%)



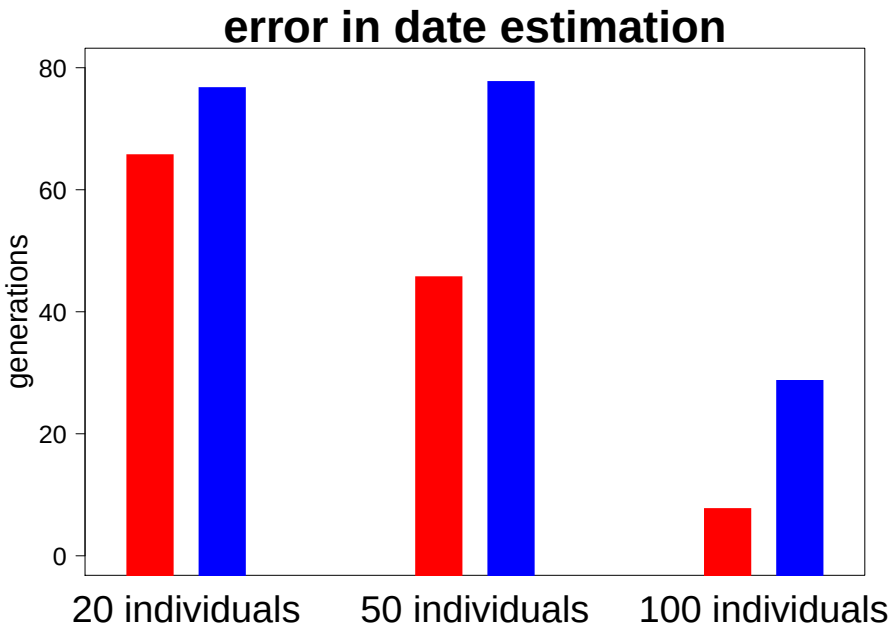
Colombia-Han (20%) Colombia-Han (20%) Colombia-Han (50%)

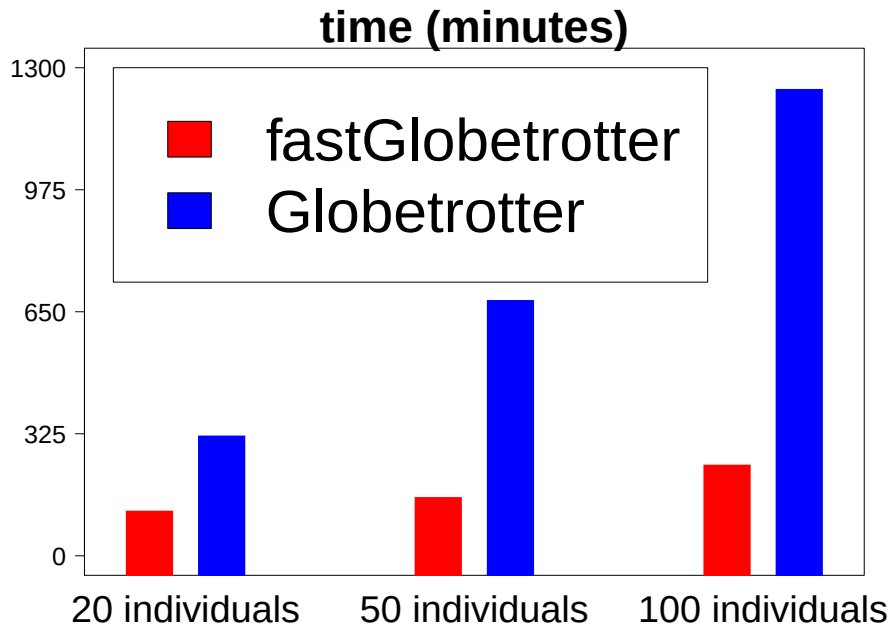
efficient dating – *fastGLOBETROTTER*



- Most curve information is at left end (noise at right)
- `fastGLOBETROTTER` preferentially fits left parts of curves

fastGLOBETROTTER vs *GLOBETROTTER*





simulated admixture 150gen ago

French + **Brahui** ($F_{ST} \approx 0.02$)

running *GLOBETROTTER (fastGLOBETROTTER)*

Running GLOBETROTTER requires three steps:

1. use *CHROMOPAINTER* to paint surrogate populations against each other
2. use *CHROMOPAINTER* to paint target (admixed) populations against surrogates
3. run *GLOBETROTTER* using combined results from (1)-(2)

We already discussed steps (1)-(2) in the previous lecture.

running *GLOBETROTTER (fastGLOBETROTTER)*

Running GLOBETROTTER requires three steps:

1. use *CHROMOPAINTER* to paint surrogate populations against each other
2. use *CHROMOPAINTER* to paint target (admixed) populations against surrogates
3. run *GLOBETROTTER* using combined results from (1)-(2)

Step (2) – run *CHROMOPAINTER* on admixed target pops:

```
./ChromoPainterv2 -g [].haplotypes
-r [].recomrates -t [].idfile.txt
-f [].poplist.txt 0 0 -s 10 -o [outputname]
```

- two output files of interest:
  - `[outputname].chunklengths.out` (previous lecture)
  - `[outputname].samples.out`

```
EM_iter = 0 (N_e = 0 / copy_prop = 0 / mutation = 0 / mutationGLOBAL
HAP 1 BrahuiYorubaSimulation1
1 977 1483 1483 396 396 396 396 396 396 396 396 396 1417 1417 1417 14
2 67 386 386 1464 1464 1464 1464 1464 1464 1464 1464 1464 1464 1417 1417 1
3 1479 1297 1297 1747 1309 1309 1309 1309 1309 1309 1309 1309 1309 1417 14
4 136 1442 1442 2305 102 102 102 102 102 102 102 102 102 1417 1417 1417 1
5 128 128 128 128 128 128 128 128 128 128 128 128 128 128 1417 1417 1417
6 975 1823 1823 2292 2292 2292 2292 2292 2292 2292 2292 2292 2292 1417 141
7 1544 2291 2291 2291 2291 2291 2291 2291 2291 2291 2291 2291 2291 984 200
8 991 133 133 133 133 133 133 133 133 133 133 133 133 1827 1827 1827 1827
9 1759 1759 1759 1759 1759 1759 1759 1759 1759 1759 1759 1759 1759 1417 14
10 1819 1819 1819 2318 1819 1819 1819 1819 1819 1819 1819 1819 1819 1417 1
HAP 2 BrahuiYorubaSimulation1
1 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 88 8
2 94 94 94 94 1623 1623 1623 1623 1623 1623 1623 1623 1623 1472 1472 1472
```

running *GLOBETROTTER (fastGLOBETROTTER)*

Running GLOBETROTTER requires three steps:

1. use *CHROMOPAINTER* to paint surrogate populations against each other
2. use *CHROMOPAINTER* to paint target (admixed) populations against surrogates
3. run *GLOBETROTTER* using combined results from (1)-(2)

Step (3) – run *GLOBETROTTER*:

```
R < GLOBETROTTER.R [].paramfile.txt
[].samplesfile.txt [].recomfiles.txt –no-save >
[screenoutput]
```

- Input files:

1. "`[].samplefiles.inp`" – contains rows pointing to the *CHROMOPAINTER* output -s files made in step (2)

```
example/BrahuiYorubaSimulationAdmixtureChrom20.samples.out
example/BrahuiYorubaSimulationAdmixtureChrom21.samples.out
example/BrahuiYorubaSimulationAdmixtureChrom22.samples.out
```

running *GLOBETROTTER (fastGLOBETROTTER)*

Running GLOBETROTTER requires three steps:

1. use *CHROMOPAINTER* to paint surrogate populations against each other
2. use *CHROMOPAINTER* to paint target (admixed) populations against surrogates
3. run *GLOBETROTTER* using combined results from (1)-(2)

Step (3) – run *GLOBETROTTER*:

```
R < GLOBETROTTER.R [].paramfile.txt
[].samplesfile.txt [].recomfiles.txt –no-save >
[screenoutput]
```

- Input files:

2. "`[].recomfiles.inp`" – contains rows pointing to the *CHROMOPAINTER* input -r files

```
example/BrahuiYorubaSimulationChrom20.recomrates
example/BrahuiYorubaSimulationChrom21.recomrates
example/BrahuiYorubaSimulationChrom22.recomrates
```

running *GLOBETROTTER (fastGLOBETROTTER)*

Running GLOBETROTTER requires three steps:

1. use *CHROMOPAINTER* to paint surrogate populations against each other
2. use *CHROMOPAINTER* to paint target (admixed) populations against surrogates
3. run *GLOBETROTTER* using combined results from (1)-(2)

Step (3) – run *GLOBETROTTER*:

```
R < GLOBETROTTER.R [].paramfile.txt
[].samplesfile.txt [].recomfiles.txt –no-save >
[screenoutput]
```

- Input files:

3. "`[].paramfile.txt`" – info for running *GLOBETROTTER*

```
prop.ind: 1
bootstrap.date.ind: 1
null.ind: 1
input.file.ids: example/BrahuiYorubaSimulation.idfile.txt
input.file.copyvectors: BrahuiYorubaSimulationAllVersusAllChrom22.chunklengths.out
save.file.main: example/BrahuiYorubaSimulationAdmixed.fastGT.main
save.file.bootstraps: example/BrahuiYorubaSimulationAdmixed.fastGT.boot
copyvector.popnames: Balochi BantuKenya BantuSouthAfrica Burusho English HanNchina
surrogate.popnames: Balochi BantuKenya BantuSouthAfrica Burusho English HanNchina K
target.popname: BrahuiYorubaSimulation
num.mixing.iterations: 5
props.cutoff: 0.001
bootstrap.num: 20
num.admixdates.bootstrap: 1
num.surrogatepops.perplot: 3
curve.range: 1 30
bin.width: 0.1
xlim.plot: 0 30
prop.continue.ind: 0
haploid.ind: 0
```

running *GLOBETROTTER (fastGLOBETROTTER)*

Running GLOBETROTTER requires three steps:

1. use *CHROMOPAINTER* to paint surrogate populations against each other
2. use *CHROMOPAINTER* to paint target (admixed) populations against surrogates
3. run *GLOBETROTTER* using combined results from (1)-(2)

Step (3) – *GLOBETROTTER* output files:

1. `[filename].txt` – results, including `GLOBETROTTER`'s "best-guess" conclusion regarding admixture and the inferred admixture dates and proportions

2. `[filename].pdf` – plotted "coancestry curves" for every combination of surrogate populations that are inferred to have contributed $>$0.1% ancestry to the target population

3. `[filename]_curves.txt` – raw data used to produce the curves in (2), in case you want to make your own plots

4. `[filename.boot].txt` – re-estimated dates using bootstrap re-sampling, so that you can calculate confidence intervals for the full date.

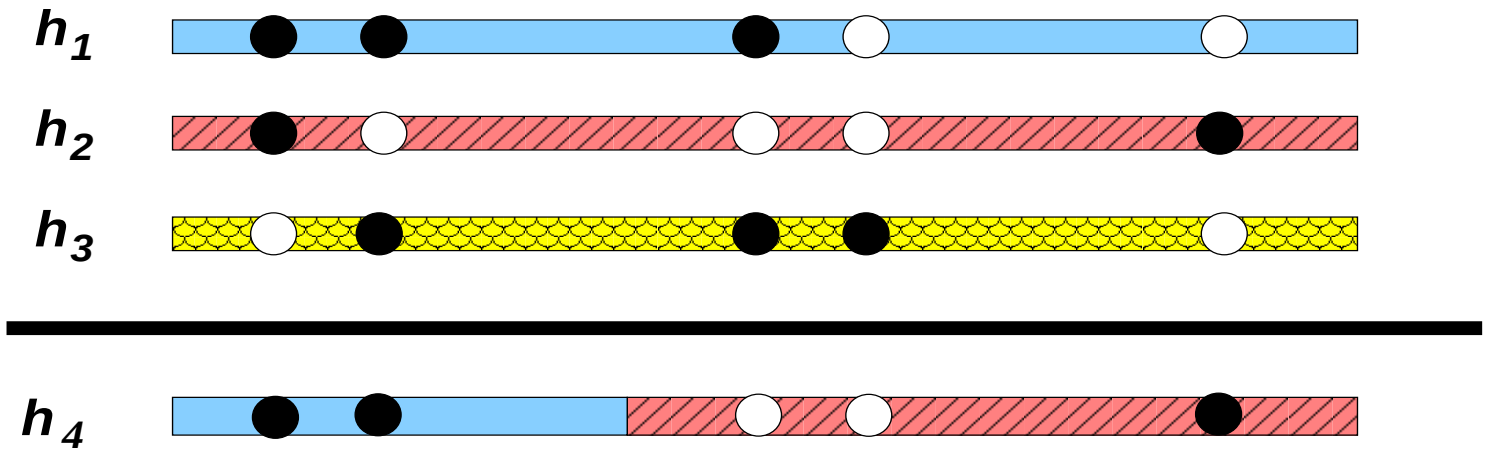
GLOBETROTTER output file – `[filename].txt`

```
### INFERRED SOURCES AND DATES ('best-guess' conclusion: one-date)
########################################
### 1-DATE FIT EVIDENCE, DATE ESTIMATE, SINGLE BEST-FITTING DONORS
gen.1date proportion.source1 maxR2fit.1date fit.quality.1event fit.qu
rce2
31.6058171831469 0.26 0.955048938359772 0.999997240697297 0.99999995
########################################
### 2-DATE FIT EVIDENCE, DATE ESTIMATES, SINGLE BEST-FITTING DONORS
gen.2dates.date1 gen.2dates.date2 maxScore.2events proportion.date1.s
5.9359285476493 31.5070427682761 0.0587005354447215 0.16 BantuSouthAf
########################################
### 1-DATE FIT SOURCES, PC1:
proportion MbutiPygmy Mandenka BantuSouthAfrica
0.26 0.0298070304917474 0.118563691512838 0.851629277995415
proportion Balochi
0.74 1
########################################
### 1-DATE FIT SOURCES, PC2:
proportion MbutiPygmy Mandenka Balochi
0.47 0.159504599211661 0.279662944331825 0.560832456456514
proportion Kalash Burusho BantuKenya Makrani
0.53 0.0201280106537808 0.0274288380541541 0.468655191971574 0.48378
########################################
### 2-DATE FIT SOURCES, DATE1-PC1:
proportion MbutiPygmy BantuKenya Mandenka
0.16 0.170597115424694 0.355969820024231 0.473433064551075
proportion Orcadian Sardinian BantuKenya Pathan Balochi
0.84 0.00539413524165527 0.0105267789145734 0.0351623301070365 0.043
########################################
### 2-DATE FIT SOURCES, DATE2-PC1:
proportion Orcadian BantuKenya
0.14 0.00252527099828796 0.997474729001712
proportion Orcadian Sardinian MbutiPygmy Pathan Mandenka Balochi
0.86 0.00361178625366078 0.00870760210388043 0.0116292468756023 0.035
```

- use "`1-DATE`" results if: "conclusion: one-date" (PC1)
  "conclusion: one-date-multiway" (PC1,PC2)
- use "`2-DATE`" results if: "conclusion: multiple-dates"

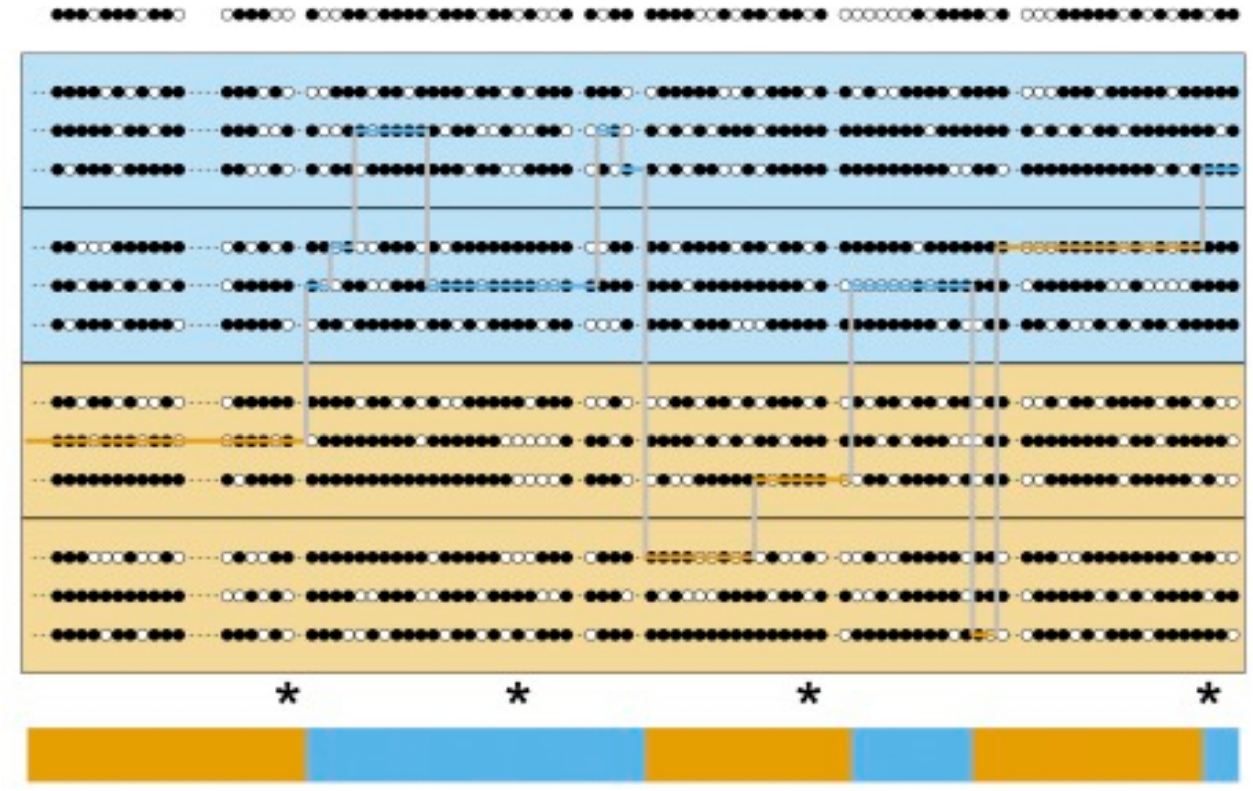
Outline

MOSAIC (Salter-Townshend & Myers 2019)



- uses haplotype information
- **Advantage:** can also identify specific segments of DNA inherited from each source

MOSAIC (Salter-Townshend & Myers 2019)



- combination of `GLOBETROTTER` and `HAPMIX` (Price et al 2009)
- when painting target individual (top sequence):
  - HMM within each of the **Orange** and **Blue** sources
  - HMM that models “ancestry switches” **between** the **Orange** and **Blue** sources

running *MOSAIC*

- Input files:

1. "`BrahuiYorubaSimulationgenofile.20`" – chromosome 20 data for simulated individuals (rows = SNPs, columns = all haplotypes)
   - repeat for each surrogate {Balochi,BantuSouthAfrica,BantuKenya,Mandenka,...} for each chromosome 1...22

2. "`rates.20`" – contains bp (row 2) and cM (row 3) position for each SNP on chromosome 20
   - repeat for each chromosome 1...22

3. "`snpfile.20`" – (row = rsID,chromo,cM,bp,allele0,allele1)
   - repeat for each chromosome 1...22

4. "`samples.names`" – 1,466 individuals (row = population,ID)

running *MOSAIC*

To run `MOSAIC` admixture test on (e.g.) Brahui-Yoruba simulation:

1. Run:
   ```
   Rscript mosaic.R -c 1:22 -p "Balochi BantuKenya
   BantuSouthAfrica ...." BrahuiYorubaSimulation
   -a 2 data/
   ```
   - `-a 2`: specifies you want 2 sources (can do more)
   - `-p ...`: lists surrogate populations
   - `data/`: location of all input files
2. Results will be in three folders (`MOSAIC_RESULTS`, `MOSAIC_PLOTS`, `FREQS`) in directory you run from

MOSAIC output plots

`MOSAIC` will generate the following plots in `MOSAIC_PLOTS`:

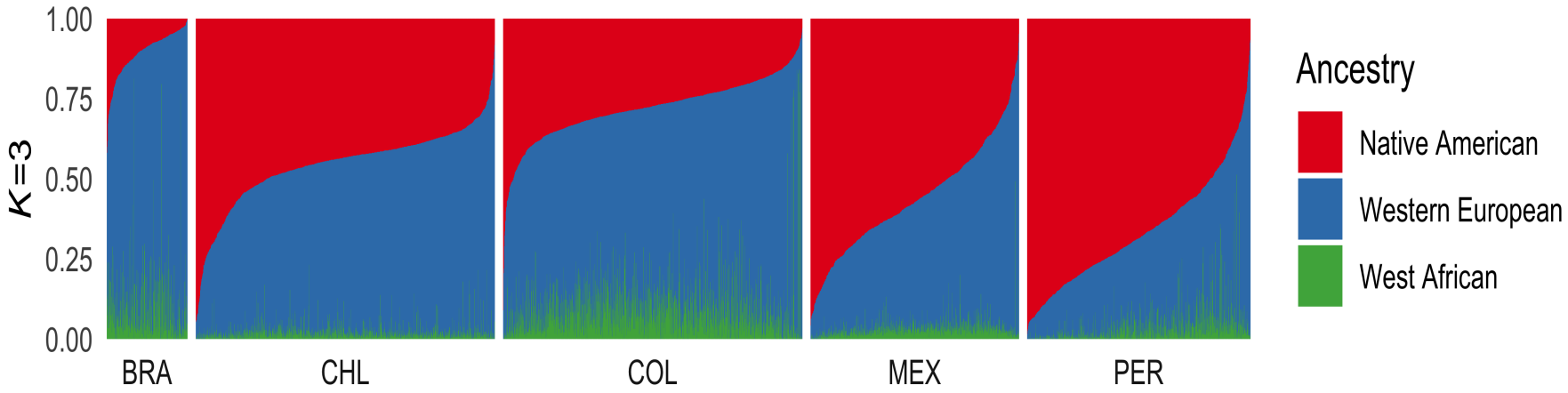
1. "`[filename]_2way_40_20-22_552_60_acoanc.pdf`" – co-ancestry curves among sources
2. "`[filename]_2way_40_20-22_552_60_EMlog.pdf`" – convergence of E-M algorithm
3. "`[filename]_2way_40_20-22_552_60_Fst.pdf`" – F_{ST} between each inferred source and various surrogate populations
4. "`[filename]_2way_40_20-22_552_60_Mu.pdf`" – inferred genetic make-up of each source
5. "`[filename]_2way_40_20-22_552_60_karyograms.pdf`" – inferred "painting" per chromosome per target haplotype

SUMMARY: techniques to date admixture

- in theory can detect & date admixture between genetically very similar sources
- have demonstrated that recent admixture appears to be ubiquitous among human groups (Loh et al 2013, Hellenthal et al 2014)
- Limitations:
  - assume "pulse(s)" of admixture, followed by random mating $\longrightarrow$ almost certainly a major simplification (Liang & Nielsen 2014)
  - continuous migration gives similar signal to pulses (Hellenthal et al 2014)
  - segments from each source will be too small to identify after a certain number of generations ($\approx$ 5,000 years)

Outline

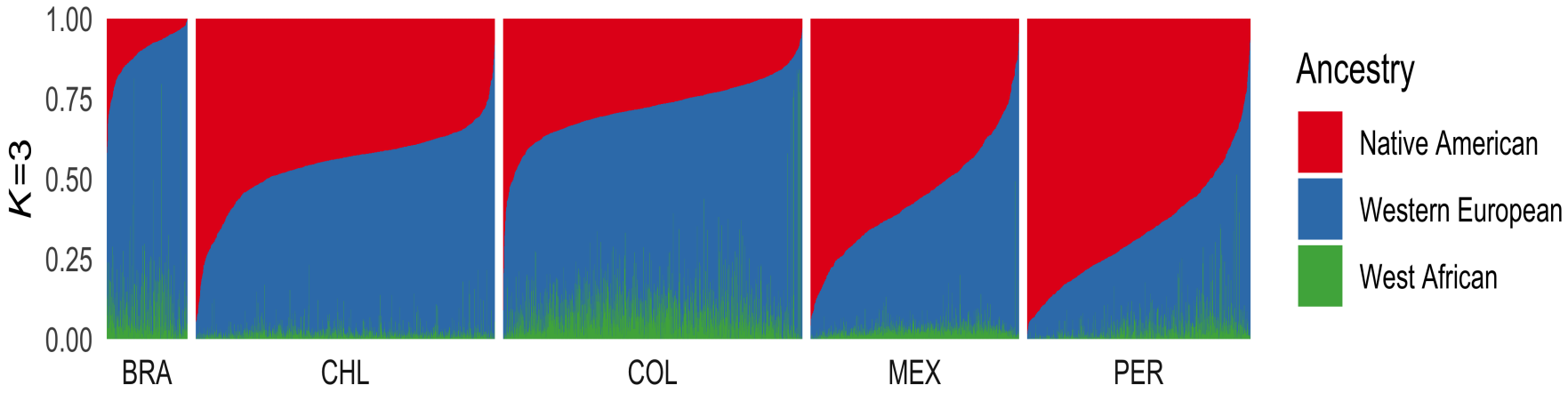
Testing for selection in admixed populations



- many selection tests assume populations are genetically homogeneous
- but many pops (e.g. Latin Americans) are admixed

(Mendoza-Revilla et al 2022, *Mol Biol Evol* **39:**msac076)

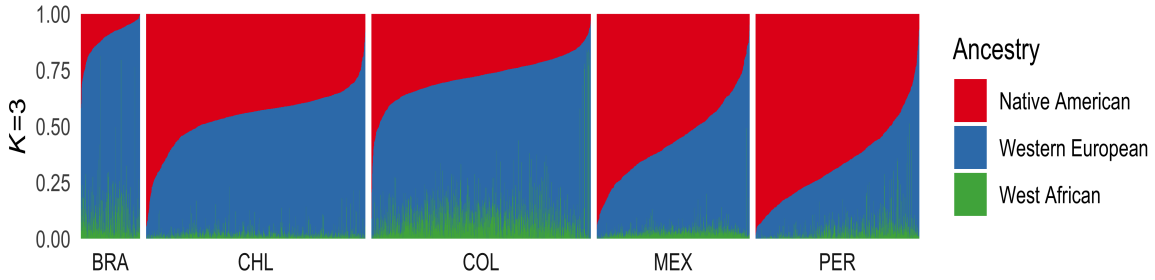
Testing for selection in admixed populations

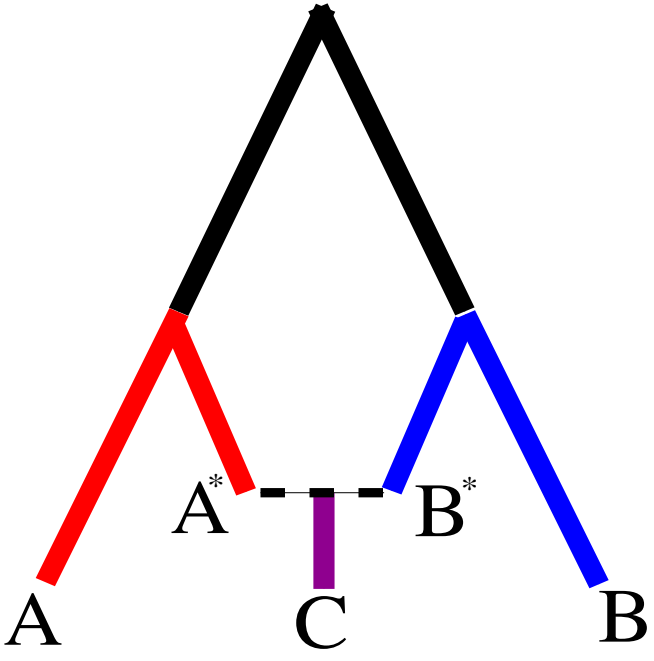


- many selection tests assume populations are genetically homogeneous
- but many pops (e.g. Latin Americans) are admixed
- *AdaptMix* – tests for selection in a population where individuals have **varying admixture proportions**
- In such cases, *AdaptMix* also aims to determine whether selection occurred:
  - in one of **Natives**, **Europeans** or **Africans** **before** admixture
  - **after** admixture in the mixed population

(Mendoza-Revilla et al 2022, *Mol Biol Evol* **39:**msac076)

Testing a SNP for selection in mixed populations





Example:

C = **target pop (e.g. Peruvians)**

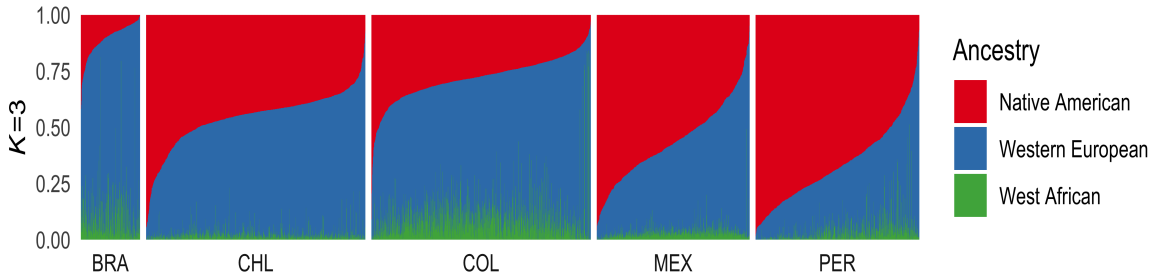
A* = **true source1 (e.g. Native American)**

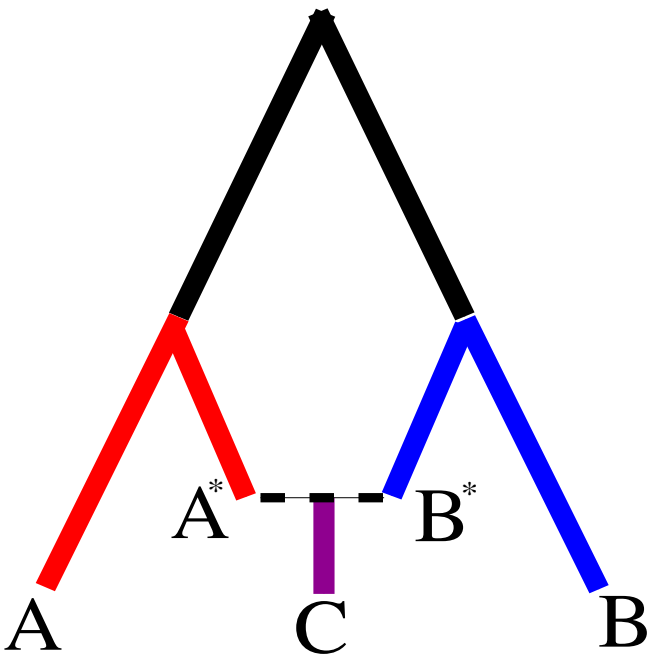
A = **surrogate for source1 (e.g. Native groups)**

B* = **true source2 (e.g. European)**

B = **surrogate for source2 (e.g. Iberians)**

Testing a SNP for selection in mixed populations





Example:

C = **target pop (e.g. Peruvians)**

A* = **true source1 (e.g. Native American)**

A = **surrogate for source1 (e.g. Native groups)**

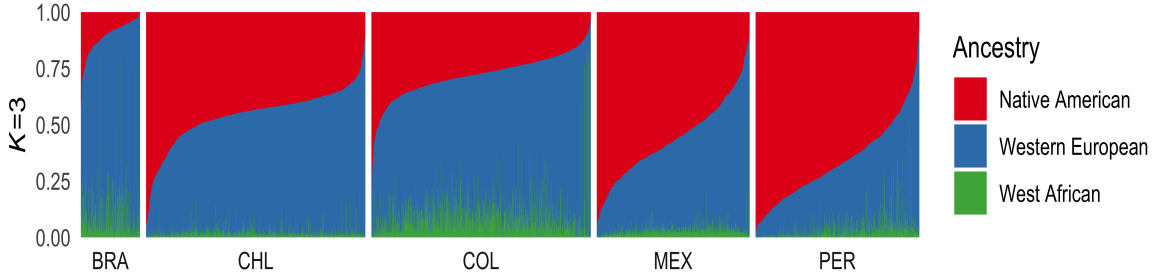
B* = **true source2 (e.g. European)**

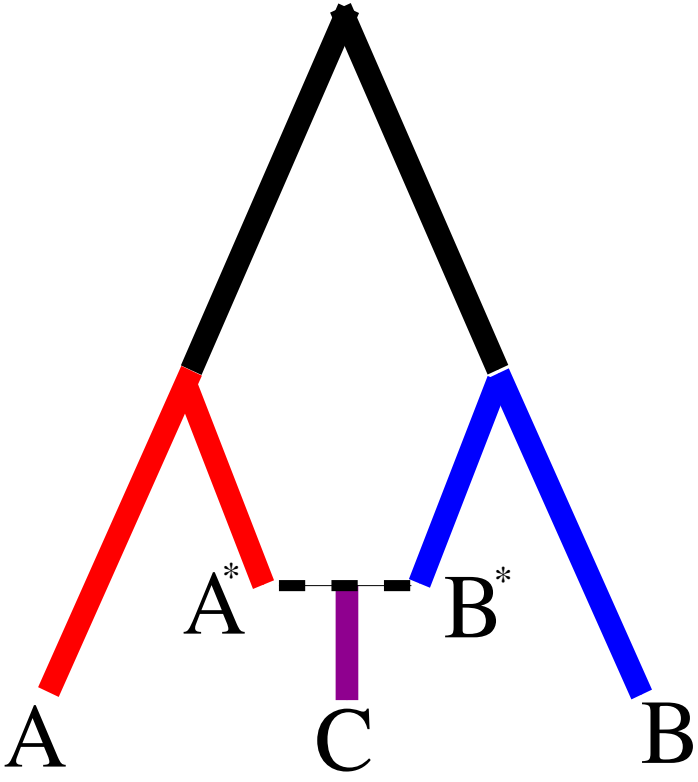
B = **surrogate for source2 (e.g. Iberians)**

AdaptMix **has two steps**:

1. test SNP for *any* selection
2. determine whether selection in **red**, **blue** or **purple** branch

(1) Testing a SNP for selection in mixed populations





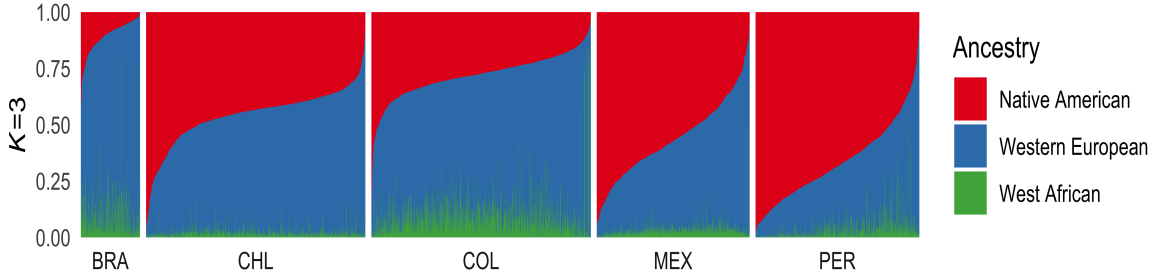
α = prop of DNA that C inherited from A^*

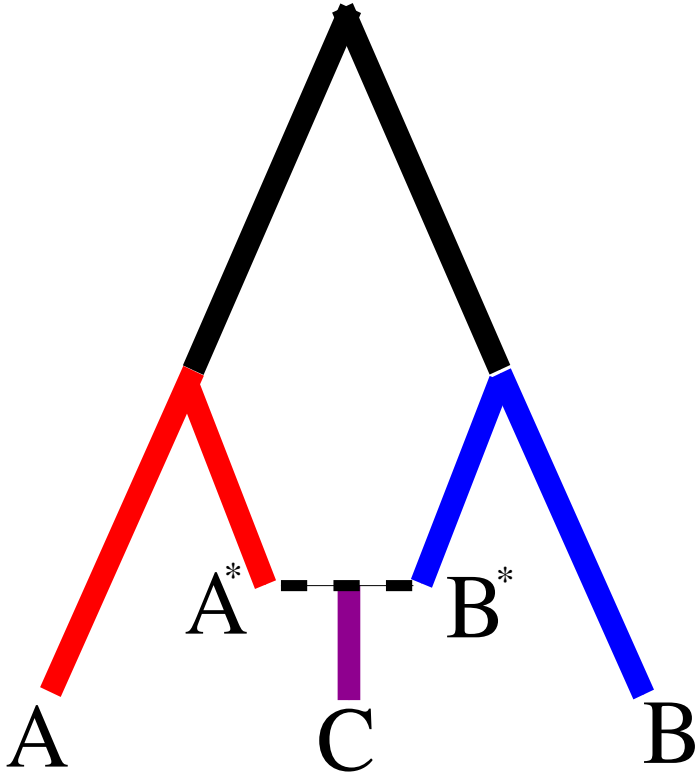
s = selection strength $\quad$ f_c = frequency in C

$$E[f_C \mid s = 0] = \alpha f_{A^*} + (1-\alpha) f_{B^*}$$
$$= \alpha f_A + (1-\alpha) f_B$$

(Mathieson et al 2015, *Nature* **528**:499)

(1) Testing a SNP for selection in mixed populations





α = prop of DNA that C inherited from A^*

s = selection strength f_c = frequency in C

$$E[f_C \mid s = 0] = \alpha f_{A^*} + (1-\alpha) f_{B^*}$$
$$= \alpha f_A + (1-\alpha) f_B$$

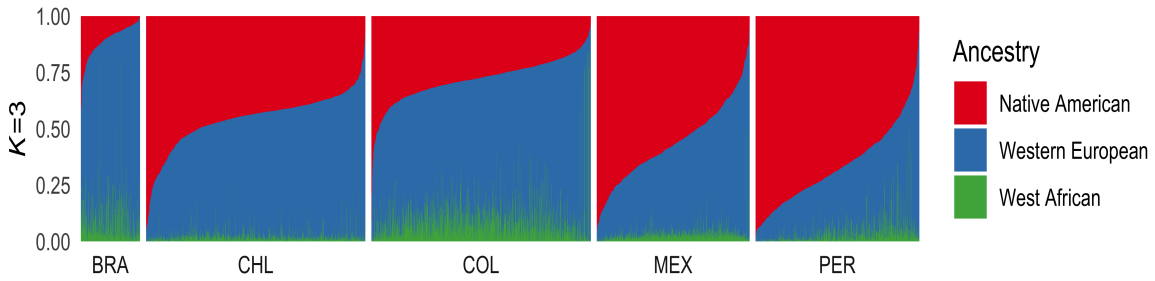
(Mathieson et al 2015, *Nature* **528**:499)

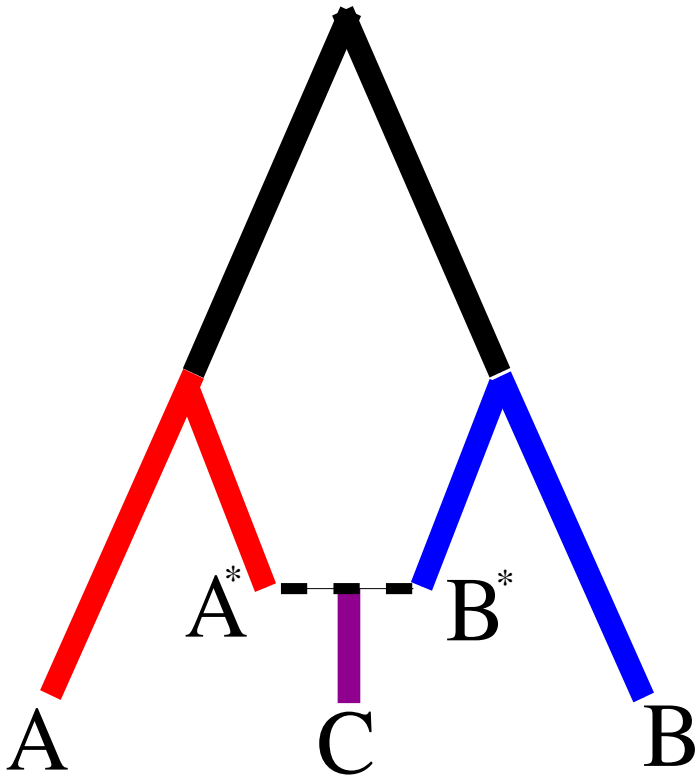
$$f_C \sim \text{Beta}(\text{mean} = \alpha f_A + (1-\alpha) f_B, \text{var} = \delta)$$

(δ = "drift")

(Number of • in C) $\sim$ Binomial(n_C, f_C)

(1) Testing a SNP for selection in mixed populations





α = prop of DNA that C inherited from A^*

s = selection strength $\quad$ f_c = frequency in C

$$E[f_C \mid s = 0] = \alpha f_{A^*} + (1-\alpha) f_{B^*}$$
$$= \alpha f_A + (1-\alpha) f_B$$

(Mathieson et al 2015, *Nature* **528**:499)

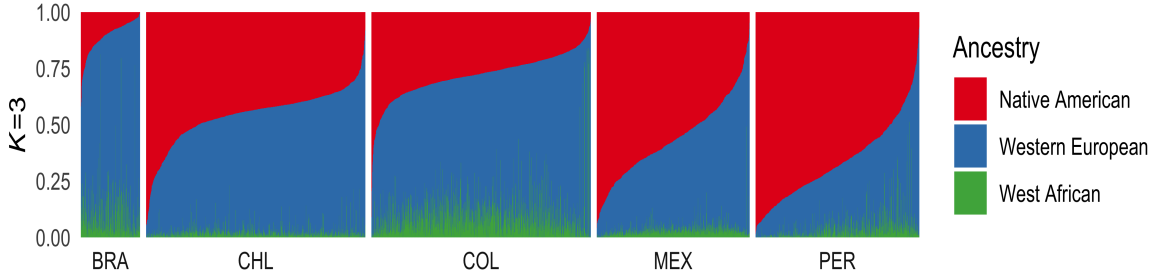
$$f_C \sim \text{Beta}(\text{mean} = \alpha f_A + (1-\alpha) f_B, \ \text{var} = \delta)$$

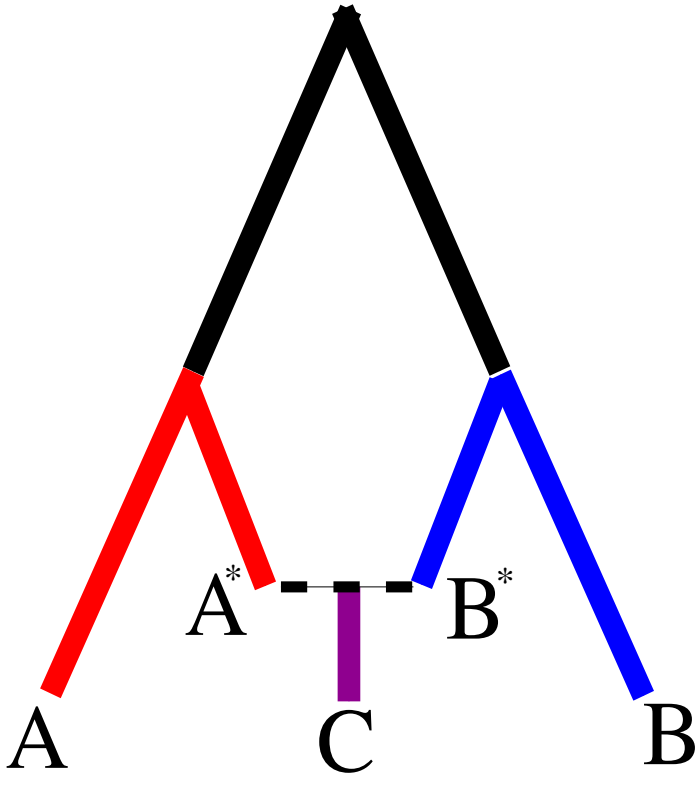
(δ = "drift")

(Number of • in C) $\sim$ Binomial(n_C, f_C)

$\longrightarrow$ **p-value for neutral test**

(2) Was selection on **purple**, **red** or **blue** branch?





α = prop of DNA that C inherited from A^*

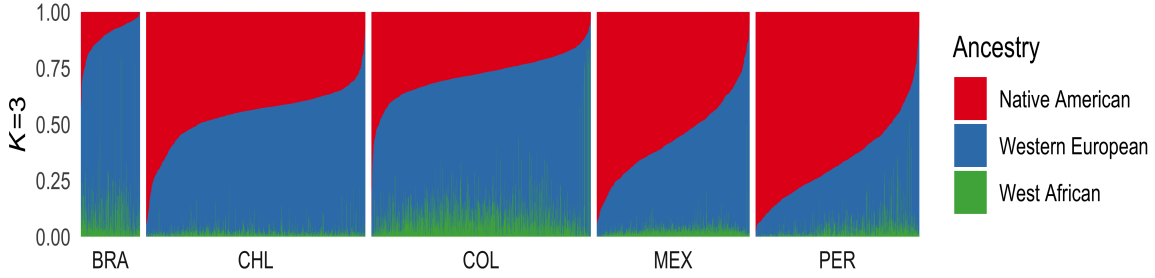
s = selection strength $\quad f_c$ = frequency in C

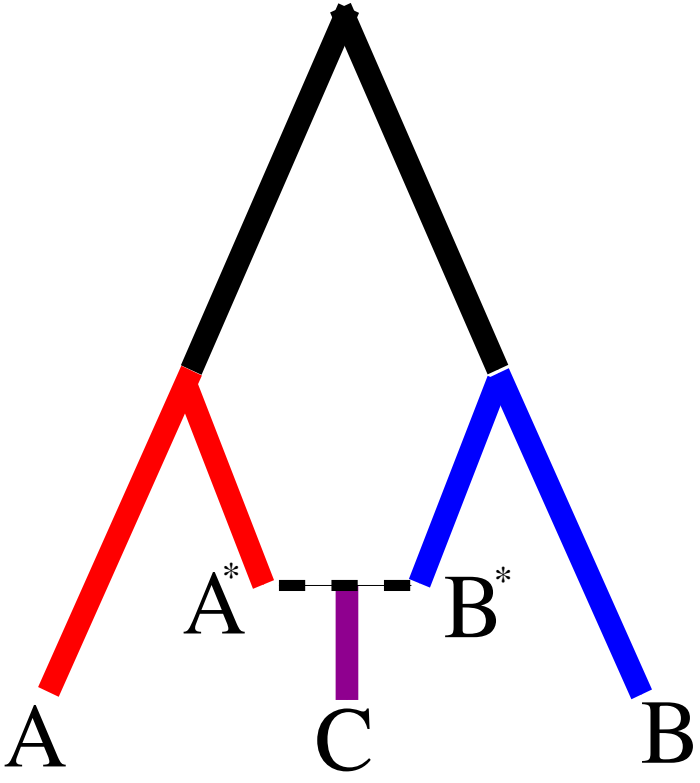
$$E[f_C \mid s > 0] = \frac{(1+s)(\alpha f_A + (1-\alpha) f_B)}{1+s(\alpha f_A + (1-\alpha) f_B)}$$

$$E[f_C \mid s > 0] = \alpha \left[\frac{(1+s) f_A}{1+s f_A} \right] + (1-\alpha) f_B$$
$$= f_B + k(f_A, f_B, s)\alpha$$

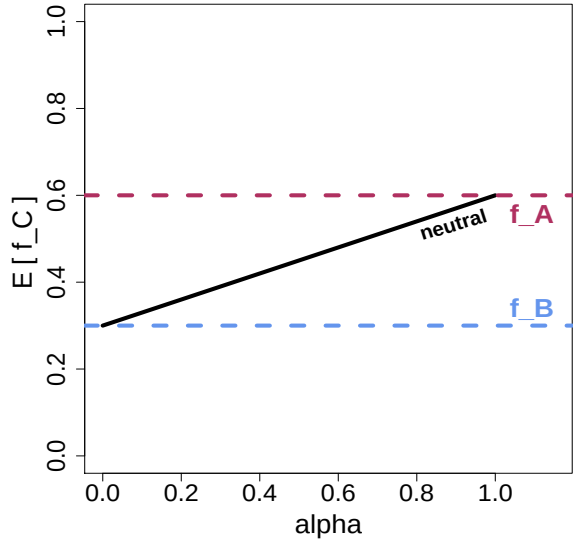
$$E[f_C \mid s > 0] = \alpha f_A + (1-\alpha) \left[\frac{(1+s) f_B}{1+s f_B} \right]$$
$$= f_A + k(f_A, f_B, s)(1-\alpha)$$

(2) Was selection on **purple**, **red** or **blue** branch?



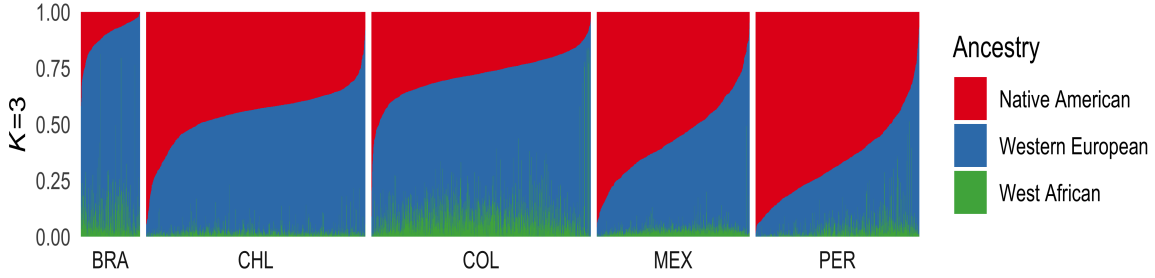


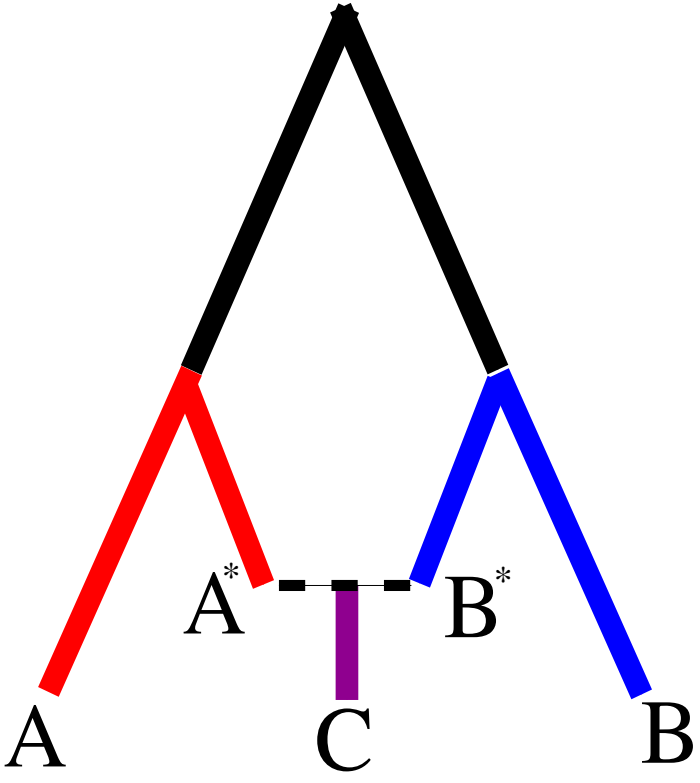
α = prop DNA inherited from A^* s = selection strength



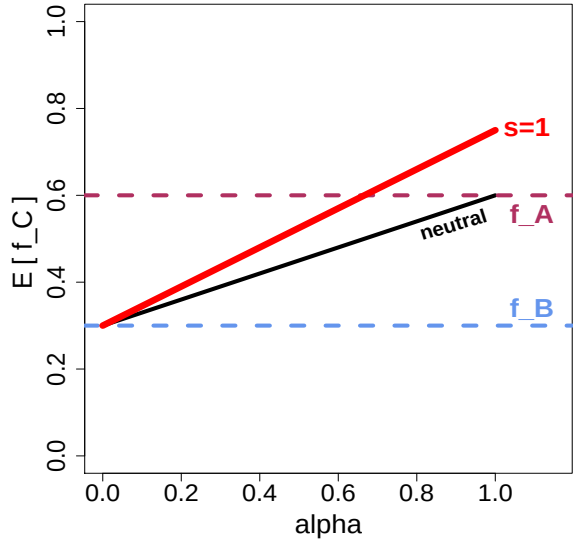
$$E[f_C \mid s = 0] = \alpha f_A + (1 - \alpha) f_B$$

(2) Was selection on **purple**, **red** or **blue** branch?



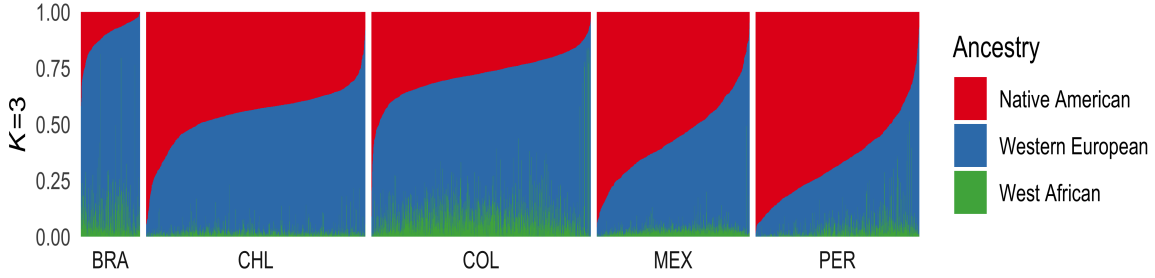


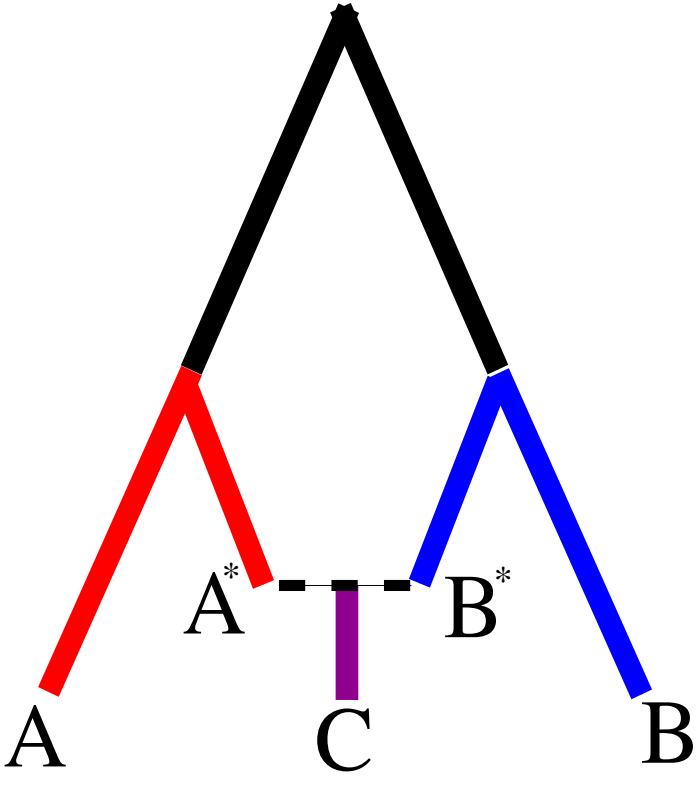
α = prop DNA inherited from A^* s = selection strength



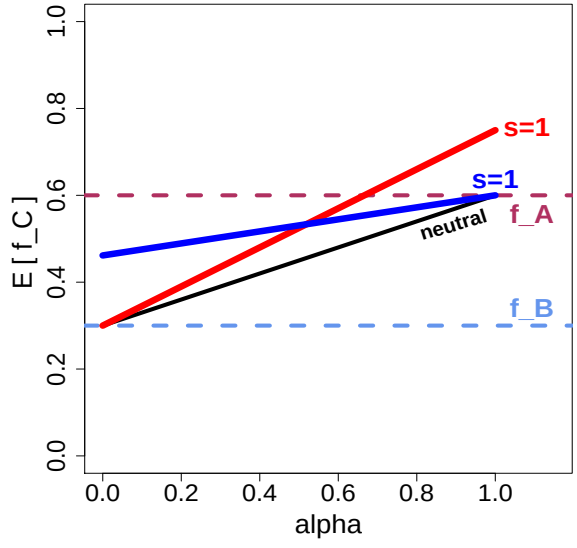
$$E[f_C \mid s > 0] = \alpha\left[\frac{(1+s)f_A}{1+sf_A}\right] + (1-\alpha)f_B$$

(2) Was selection on **purple**, **red** or **blue** branch?



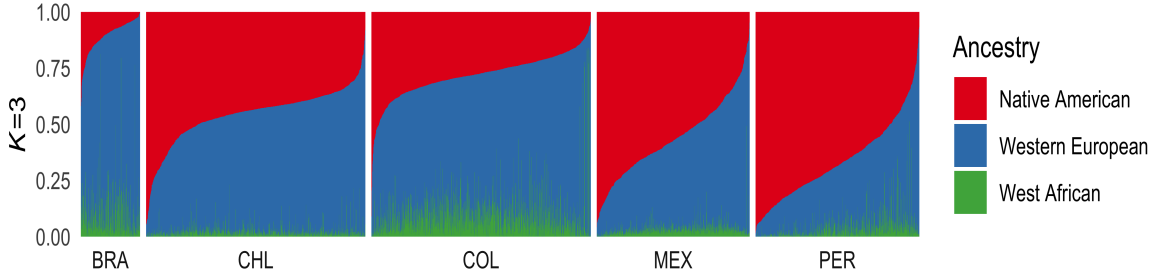


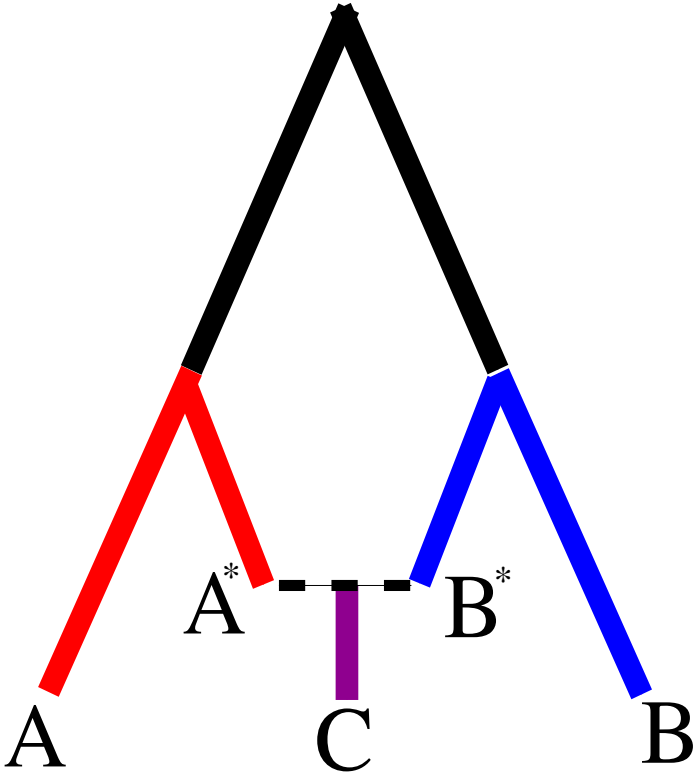
α = prop DNA inherited from A^* s = selection strength



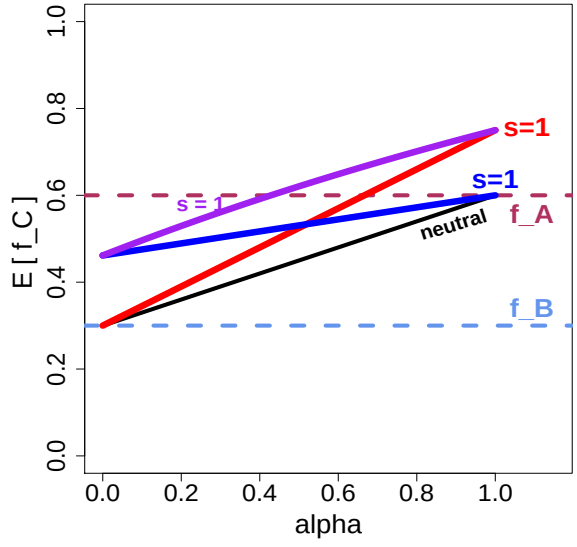
$$E[f_C \mid s > 0] = \alpha f_A + (1-\alpha)\left[\frac{(1+s)f_B}{1+sf_B}\right]$$

(2) Was selection on **purple**, **red** or **blue** branch?



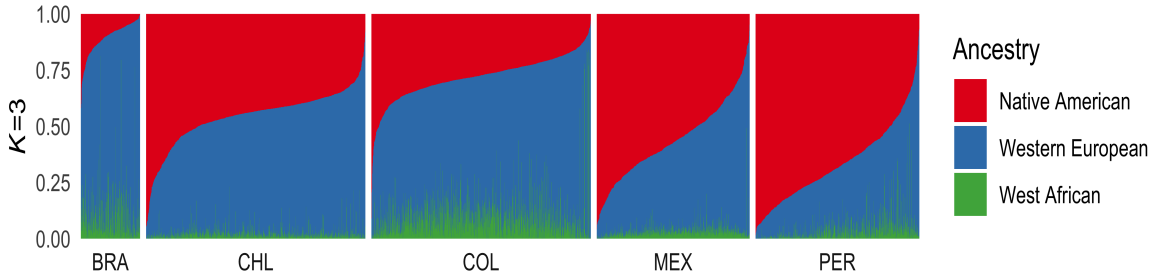


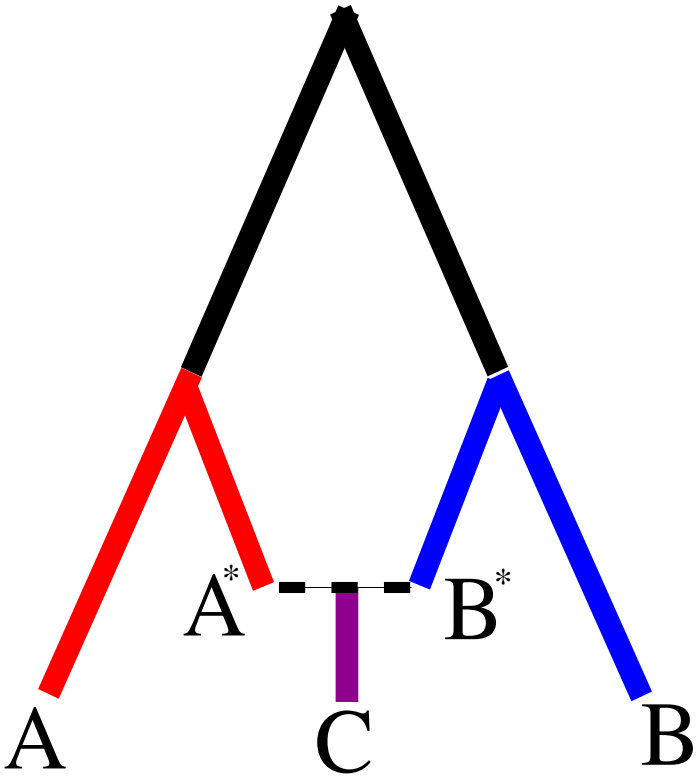
α = prop DNA inherited from A^* s = selection strength



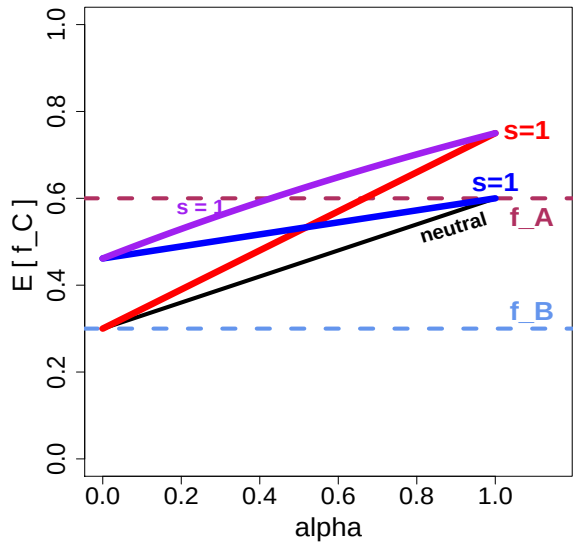
$$E[f_C \mid s > 0] = \frac{(1+s)(\alpha f_A + (1-\alpha) f_B)}{1+s(\alpha f_A + (1-\alpha) f_B)}$$

(2) Was selection on **purple**, **red** or **blue** branch?





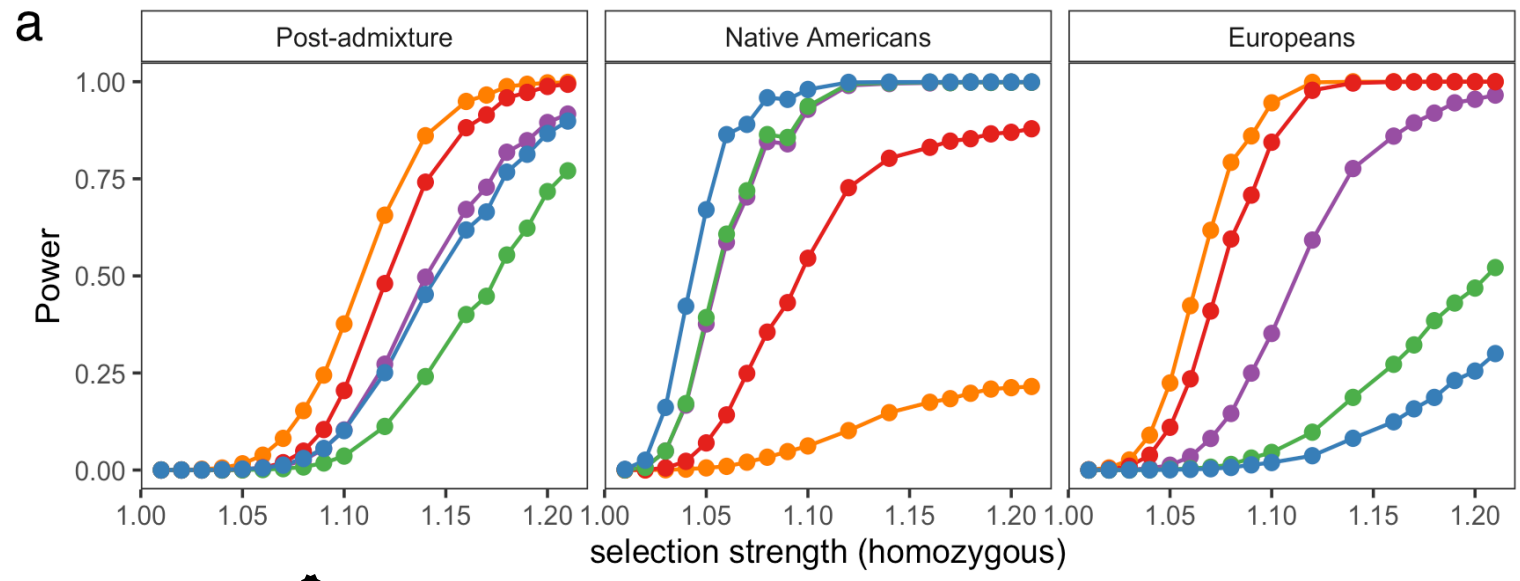
α = prop DNA inherited from A^* s = selection strength

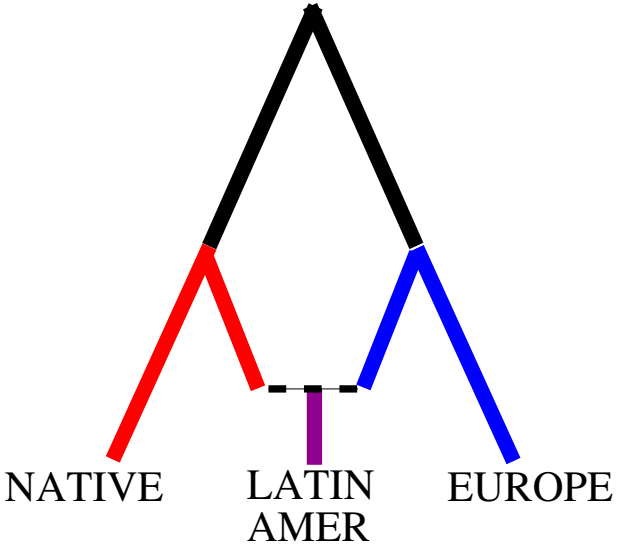


$$E[f_C \mid s > 0] = \frac{(1+s)(\alpha f_A + (1-\alpha) f_B)}{1+s(\alpha f_A + (1-\alpha) f_B)}$$

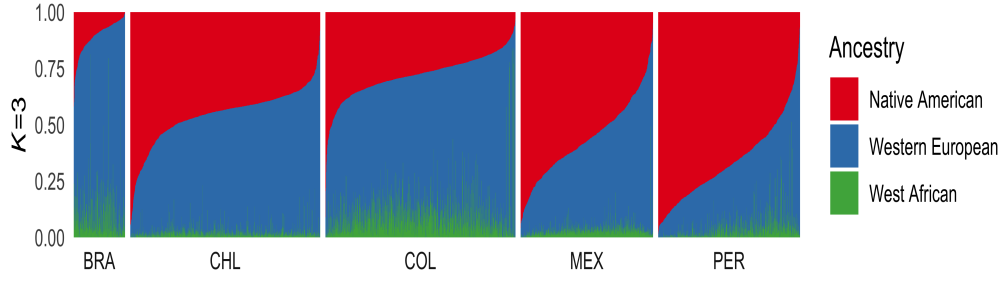
$\longrightarrow$ use **AIC** to determine best fitting model

Simulations: *AdaptMix* step (1) – detecting *any* selection





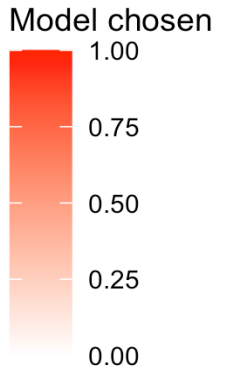
(Mendoza-Revilla et al 2022, *Mol Biol Evol* **39:**msac076)

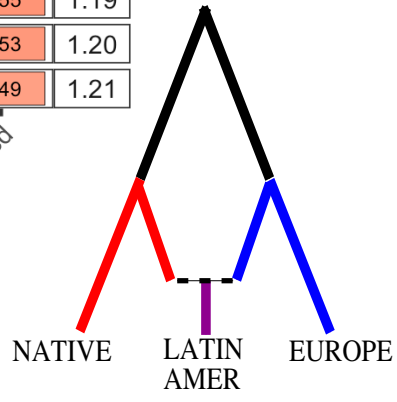


Simulations: *AdaptMix* step (2) – which branch?

b

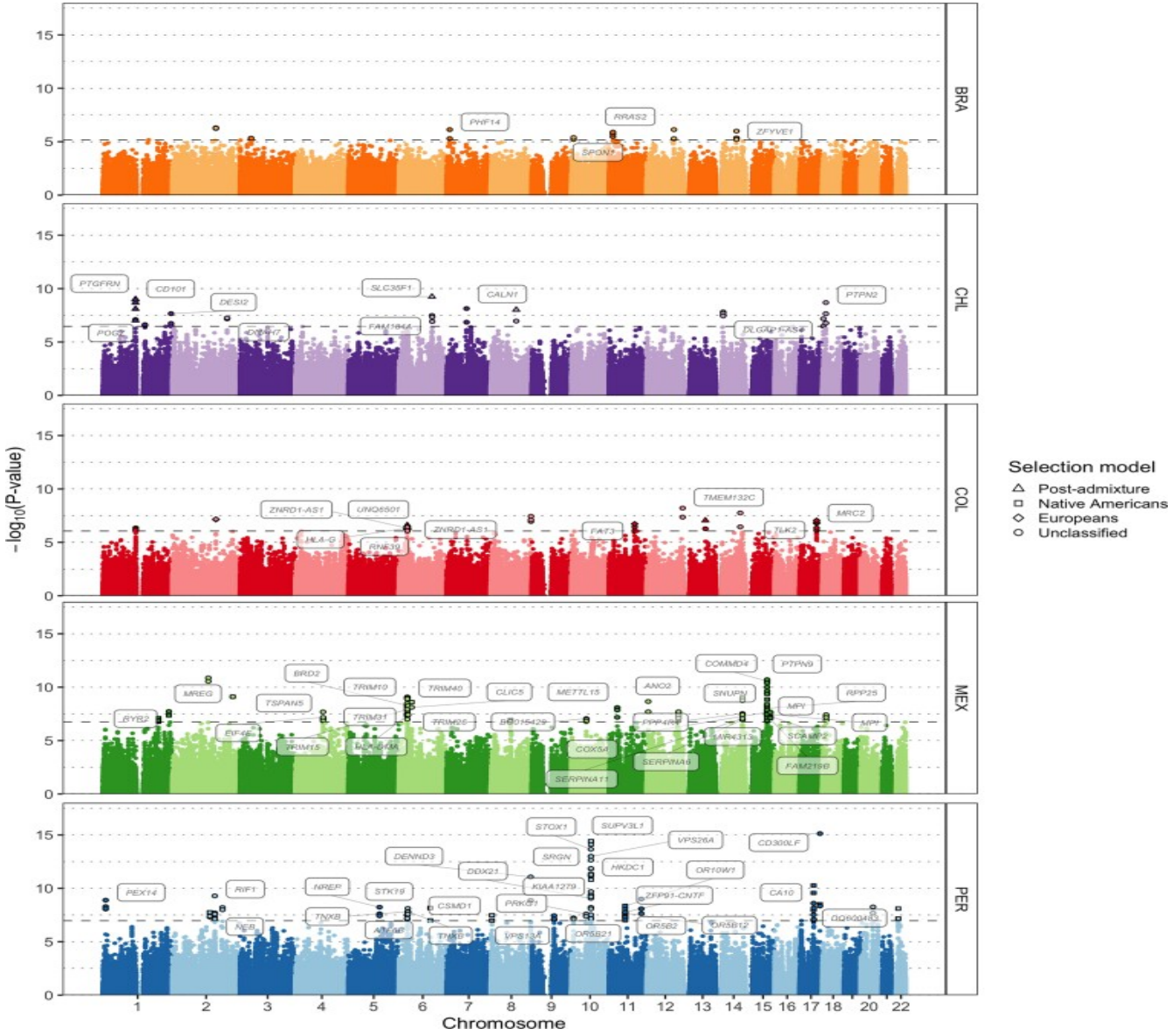
Post-admixture					Native Americans					Europeans					
NA	NA	NA	NA	NA	0.01	0.06	0.00	0.00	0.93	NA	NA	NA	NA	NA	1.01
0.01	0.00	0.00	0.00	0.99	0.01	0.12	0.00	0.00	0.87	0.01	0.00	0.00	0.00	0.99	1.02
0.01	0.01	0.00	0.00	0.98	0.01	0.17	0.00	0.00	0.82	0.01	0.00	0.00	0.00	0.99	1.03
0.02	0.00	0.00	0.00	0.97	0.01	0.24	0.00	0.00	0.75	0.01	0.00	0.01	0.00	0.98	1.04
0.04	0.01	0.00	0.00	0.95	0.01	0.36	0.00	0.00	0.63	0.01	0.00	0.01	0.00	0.98	1.05
0.07	0.00	0.00	0.00	0.92	0.01	0.45	0.00	0.00	0.55	0.01	0.00	0.02	0.00	0.97	1.06
0.12	0.01	0.00	0.00	0.87	0.01	0.56	0.00	0.00	0.44	0.01	0.00	0.04	0.00	0.95	1.07
0.16	0.01	0.00	0.00	0.83	0.00	0.62	0.00	0.00	0.37	0.01	0.00	0.06	0.00	0.93	1.08
0.22	0.01	0.00	0.00	0.77	0.00	0.67	0.00	0.00	0.32	0.01	0.00	0.10	0.00	0.89	1.09
0.28	0.01	0.00	0.00	0.71	0.00	0.72	0.00	0.00	0.27	0.02	0.00	0.13	0.00	0.86	1.10
0.43	0.00	0.00	0.00	0.57	0.00	0.81	0.00	0.00	0.18	0.02	0.00	0.20	0.00	0.78	1.12
0.54	0.00	0.00	0.00	0.46	0.00	0.84	0.00	0.00	0.15	0.02	0.00	0.28	0.00	0.70	1.14
0.60	0.00	0.00	0.00	0.39	0.00	0.86	0.00	0.00	0.14	0.02	0.00	0.34	0.00	0.64	1.16
0.63	0.00	0.00	0.00	0.37	0.00	0.86	0.00	0.00	0.14	0.02	0.00	0.37	0.00	0.61	1.17
0.66	0.00	0.00	0.00	0.34	0.01	0.87	0.00	0.00	0.13	0.02	0.00	0.40	0.00	0.58	1.18
0.68	0.00	0.00	0.00	0.31	0.01	0.87	0.00	0.00	0.13	0.02	0.00	0.43	0.00	0.55	1.19
0.70	0.00	0.00	0.00	0.30	0.01	0.87	0.00	0.00	0.12	0.02	0.00	0.46	0.00	0.53	1.20
0.72	0.00	0.00	0.00	0.27	0.01	0.87	0.00	0.00	0.12	0.02	0.00	0.49	0.00	0.49	1.21
Post-admixture	Native Americans	Europeans	Africans	Unclassified	Post-admixture	Native Americans	Europeans	Africans	Unclassified	Post-admixture	Native Americans	Europeans	Africans	Unclassified	



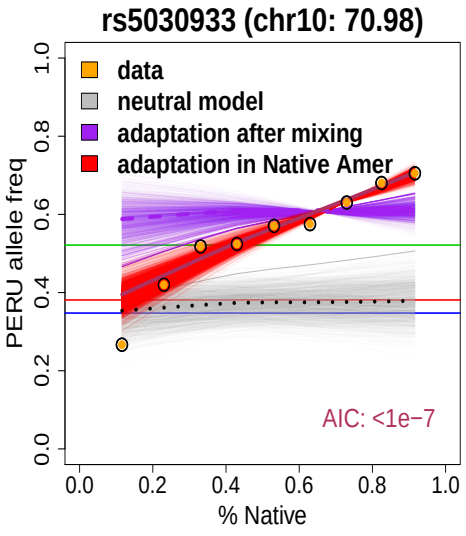


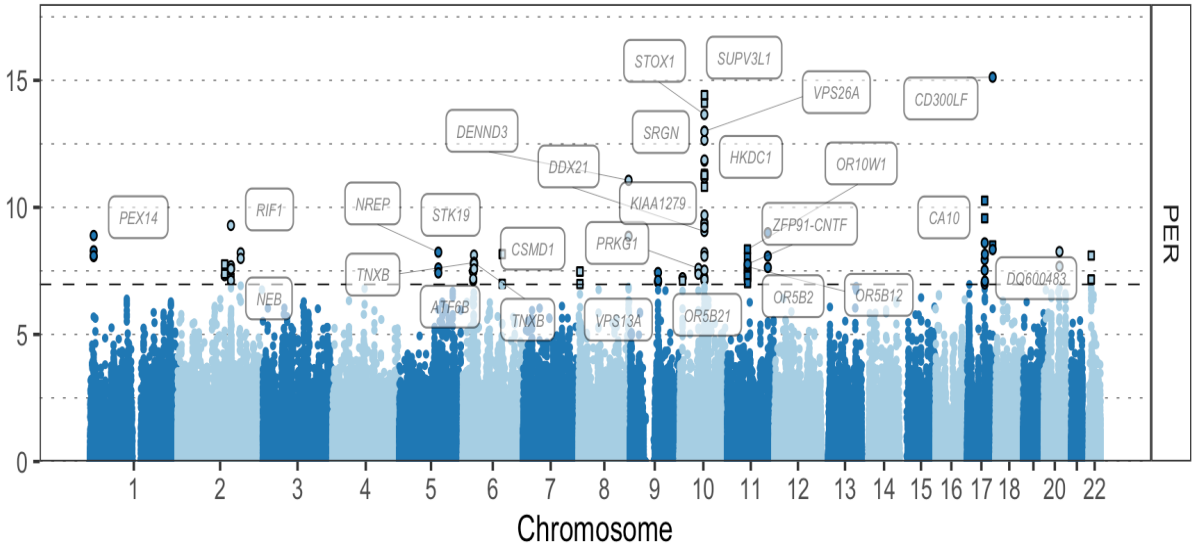
(Mendoza-Revilla et al 2022, *Mol Biol Evol* **39:**msac076)

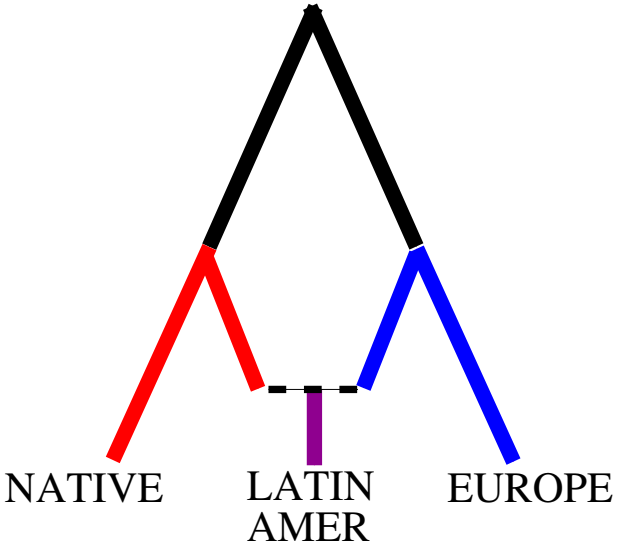
Latin American analysis: *AdaptMix*

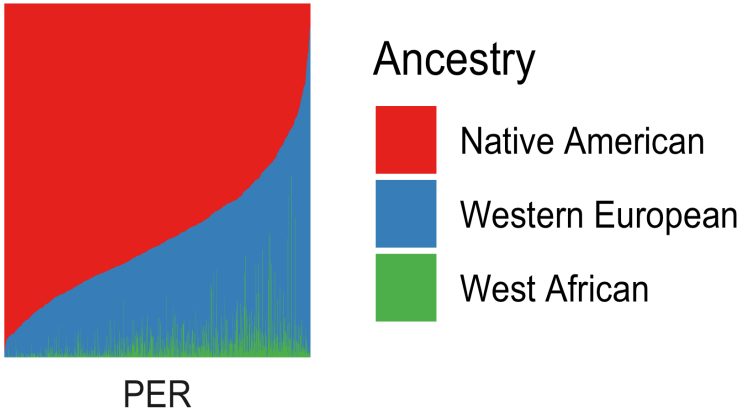


Latin American analysis: *AdaptMix*



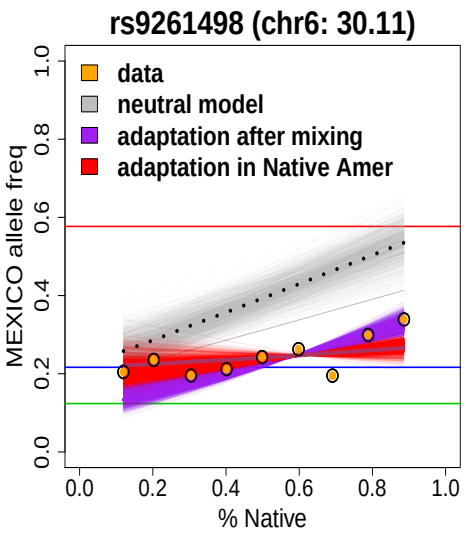


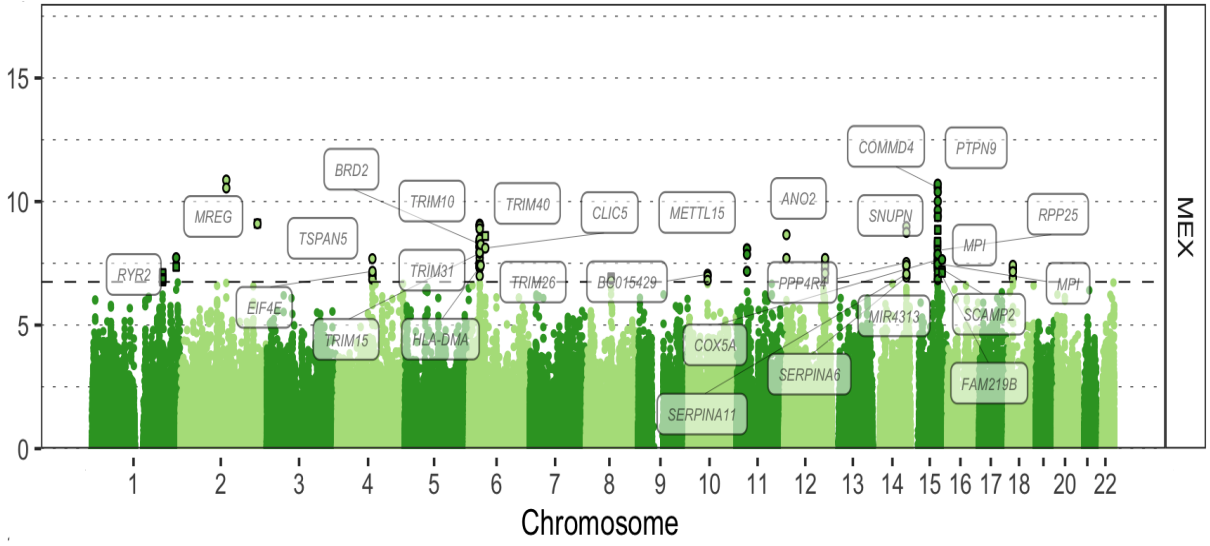


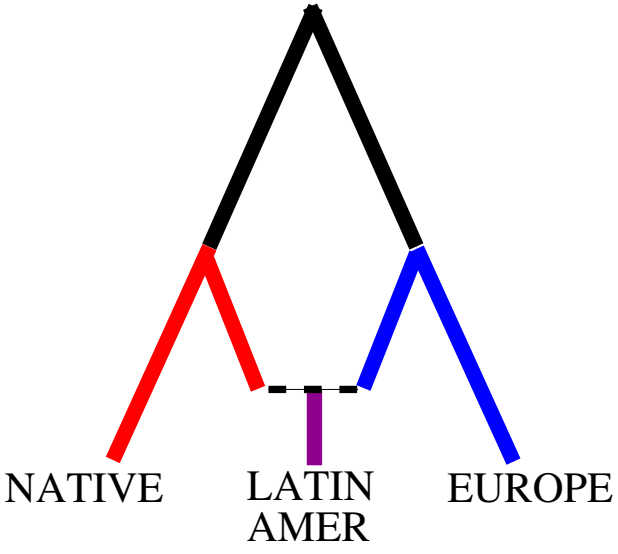


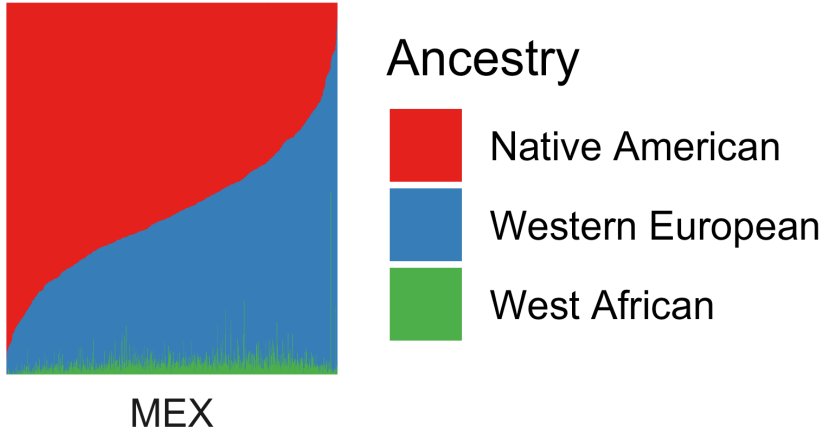
(Mendoza-Revilla et al 2022, *Mol Biol Evol* **39:**msac076)

Latin American analysis: *HLA*









(Mendoza-Revilla et al 2022, *Mol Biol Evol* **39:**msac076)

Running *AdaptMix*

```
R < run_AdaptMix.R [paramfile.inp] [datafile.inp]
[sample.idfile.inp] [outfilename] -no-save >
[screen_output]
```

Input Files:

- `datafile.inp` – lists location of haplotype file(s); these files must be in `CHROMOPAINTER -g` format
- `sample.idfile.inp` – lists each individual (row), giving {ID,pop} and then admixture fractions from *K* sources (e.g. inferred by `ADMIXTURE`, `GLOBETROTTER`,...)
- `paramfile.inp`:

```
pop.vec: PEL
surrogate.vec: CHB IBS YRI
drift.maf.bins: 0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50
min.allele.freq.shift: 0.01
```

(`surrogate.vec` should correspond to last *K* columns of `sample.idfile.inp`)

Running *AdaptMix*

```
R < run_AdaptMix.R [paramfile.inp] [datafile.inp]
[sample.idfile.inp] [outfilename] –no-save >
[screen_output]
```

Output File:

```
drift.est [0,0.05) [0.05,0.1) [0.1,0.15) [0.15,0.2) [0.2,0.25) [0.25,0.3) [0.3,0.35) [0.35,0.4) [0.4,0.45) [
PEL 0.000259 0.003634 0.006538 0.009982 0.008339 0.012232 0.014434 0.018024 0.019075 0.016548
file pos log10.pval.target.1 obs.freq.target.1 exp.freq.target.1 AIC.neutral.target.1 AIC.postadmix.target.1
1 1:848849 0.928 1 0.9313 24.235 2.34 2.34 26.165 26.227 102.3 102.3 209715.1 26214.3
1 1:884529 0 1 0.9989 0.367 2.34 2.346 2.354 2.34 1.5 838860.7 1677721.5 6.3
1 1:1464806 0 1 0.999 0.34 2.34 2.34 2.34 2.34 0.1 0.1 0.1 0.3
1 1:1608229 0.195 0.7647 0.6931 138.866 136.547 134.841 140.361 140.662 0.441 0.87 0.541 104857.5
1 1:1648405 0.778 0.3471 0.1867 192.558 170.093 168.791 181.159 193.796 1.32 1.756 11.95 13.7
1 1:1845358 0 1 0.998 0.677 2.34 2.62 2.654 2.34 19.1 6710886.3 3355443.1 76.7
1 1:1990341 0 1 0.9989 0.367 2.34 2.346 2.354 2.34 1.5 838860.7 1677721.5 6.3
1 1:2302911 0.145 0.6588 0.596 172.17 171.298 170.848 173.534 174.164 0.312 0.419 1.988 -0.588
1 1:2627555 0 1 0.9896 3.54 2.34 2.973 4.605 5.533 12.7 53687091.1 26843545.5 209715.1
1 1:2852825 0.01 0.5176 0.5102 187.997 189.959 189.736 189.121 189.97 0.03 0.103 -0.478 -0.444
```

- `log10.pval.target.1` – selection test (i.e. step 1)
- `obs.freq.target.1` – observed target pop frequency
- `exp.freq.target.1` – expected target pop frequency under neutral model
- `AIC.postadmix.target.1` – AIC score for selection **post-admixture** (i.e. step 2)
- `AIC.insurr.source1target.1` – AIC score for selection **pre-admixture in source 1** ... (i.e. step 2)
- `sel.postadmix.target.1` – inferred selection coefficient (post-admixture) ... (step 2)

AdaptMix – simulations to decide significance thresholds (and power to detect selection)

```
R < AdaptMixSimulator.R [paramfileSIM.inp]
[datafile.inp] [sample.idfile.inp] [outfilename]
-no-save > [screen_output]
```

Input Files:

- `datafile.inp` – same to be run in `AdaptMix`
- `sample.idfile.inp` – same to be run in `AdaptMix`
- `paramfileSIM.inp`:

```
selection.post-admixture?: 0
sel.coeff: 0.1
sel.type: additive
target.pop: PEL
surrogate.pops: CHB IBS YRI
sources.with.selection.preadmixture: 1 0 0
generations.selection.each.source: 150 0 0
pop.size.sources: 10000 10000 10000
drift.btwn.surrogates.and.sources: 0.1 0.1 0.2
num.neutral.snps: all
range.startfrequency.selected.snp: 0.05 0.1
infer.source.freq.using.target.data: 1
divide.into.runs.ofXX.inds(to.reduce.RAM): 700
```

AdaptMix – strategy to decide significance thresholds

1. run `AdaptMixSimulator.R` on your data, with `sel.coeff:0`
2. at bottom of `[screen_output]`, compare values of `cor.sims` to `cor.truth`
3. toggle `drift.btwn.surrogates.and.sources` until `cor.sims` ≈ `cor.truth` for each surrogate
4. use p-value and AIC scores in simulated data to choose thresholds for real data

(`cor.sims` = correlation *r* between real surrogate allele frequencies and simulated source AFs)
(`cor.truth` = *r* between real surrogate AFs and source AFs inferred using real target inds)

**Possible additional steps:

3b. once finished with **(3)**, run `AdaptMix` on your real data and simulated data

3c. compare row two (`drift.est`) of `AdaptMix` output between your real data and simulated data

3d. multiply final `drift.btwn.surrogates.and.sources` from **(3)** by constant, and repeat simulations until `AdaptMix` row two output for simulated data (roughly) aligns with that for real data

Outline

Summary

- Detecting and dating admixture:

  1. *ALDER / MALDER* – uses LD decay related to admixture event to date the event (Loh et al 2013, Pickrell et al 2014)

  2. *GLOBETROTTER, fastGLOBETROTTER* – leverages haplotype information to increase power

     (Hellenthal et al 2014, Wangkumhang et al 2021)

  3. *MOSAIC* – can also identify specific segments of DNA inherited from each admixing source (Salter-Townshend & Myers 2019)

- detecting selection in admixed individuals:

  1. *ADAPTMIX* – models allele frequencies in admixed people (with varying admixture fractions) to test for selection, and to determine whether selection occurred before or after admixture

References

- Alexander et al 2009, *Genome Research*, 19:1655-64.
- Chacon-Duque et al 2018, *Nature Commun*, 9:5388.
- Falush et al 2003, *Genetics*, 164:1567-87.
- Hellenthal et al 2014, *Science*, 343:747-51.
- Lawson et al 2012, *PLoS Genet*, 8:e1002453.
- Liang & Nielsen 2014, *Genetics*, 197:953-67.
- Loh et al 2013, *Genetics*, 193:1233-54.
- Mendoza-Revilla et al 2022, *Mol Biol Evol*, 39:msac076.
- Patterson et al 2012, *Genetics*, 192:1065-93.
- Pickrell et al 2014, *Proc Natl Acad Sci USA*, 111:2632-37.
- Price et al 2009, *PLoS Genetics*, 5:e1000519.
- Salter-Townshend & Myers 2019, *Genetics*, 212:869-89.
- Wangkumhang et al 2022, *Genome Research*, 32:1553-64.