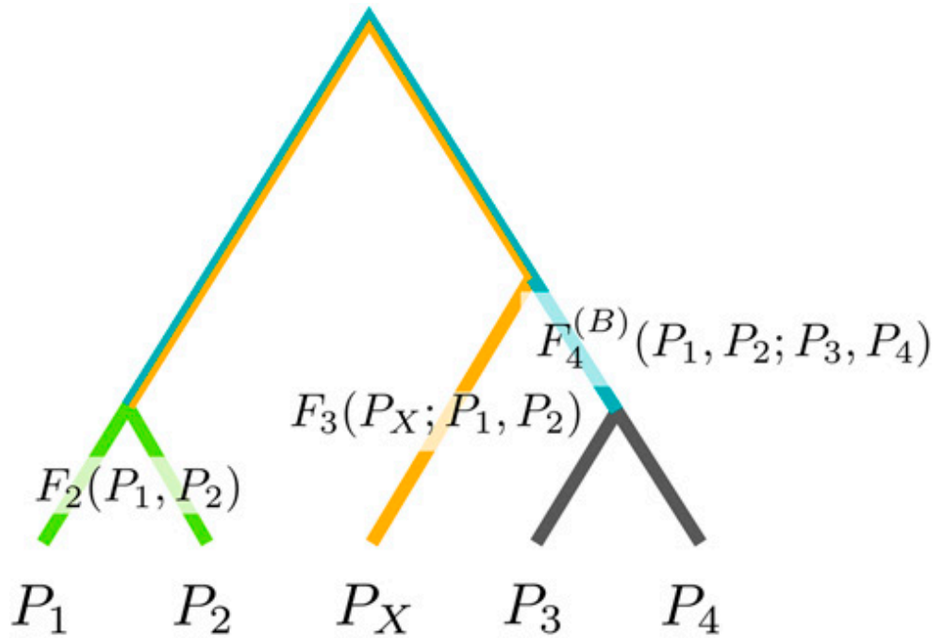
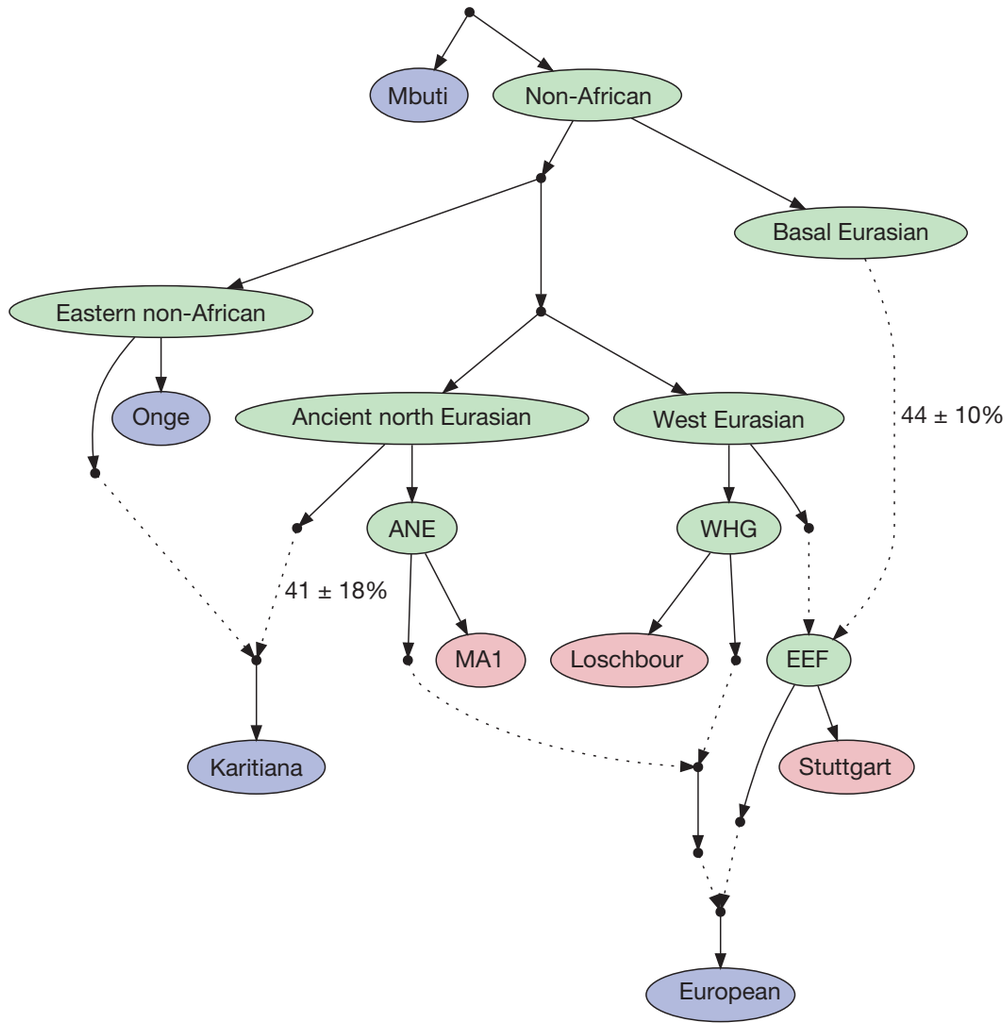
Inferring admixture from genomic data

EMBO Workshop Population Genomics, Napoli 2017





Martin Sikora, PhD

Centre for Geogenetics, Natural History Museum of Denmark, University of Copenhagen

Outline of today's lecture

- Motivation and Background
  - Admixture in human history
  - Overview of methods

Outline of today's lecture

- Motivation and Background
  - Admixture in human history
  - Overview of methods

- Testing for admixture using f-statistics
  - Basic concepts
  - What can we test with f-statistics?
  - Estimation and interpretation
  - Admixture graph fitting

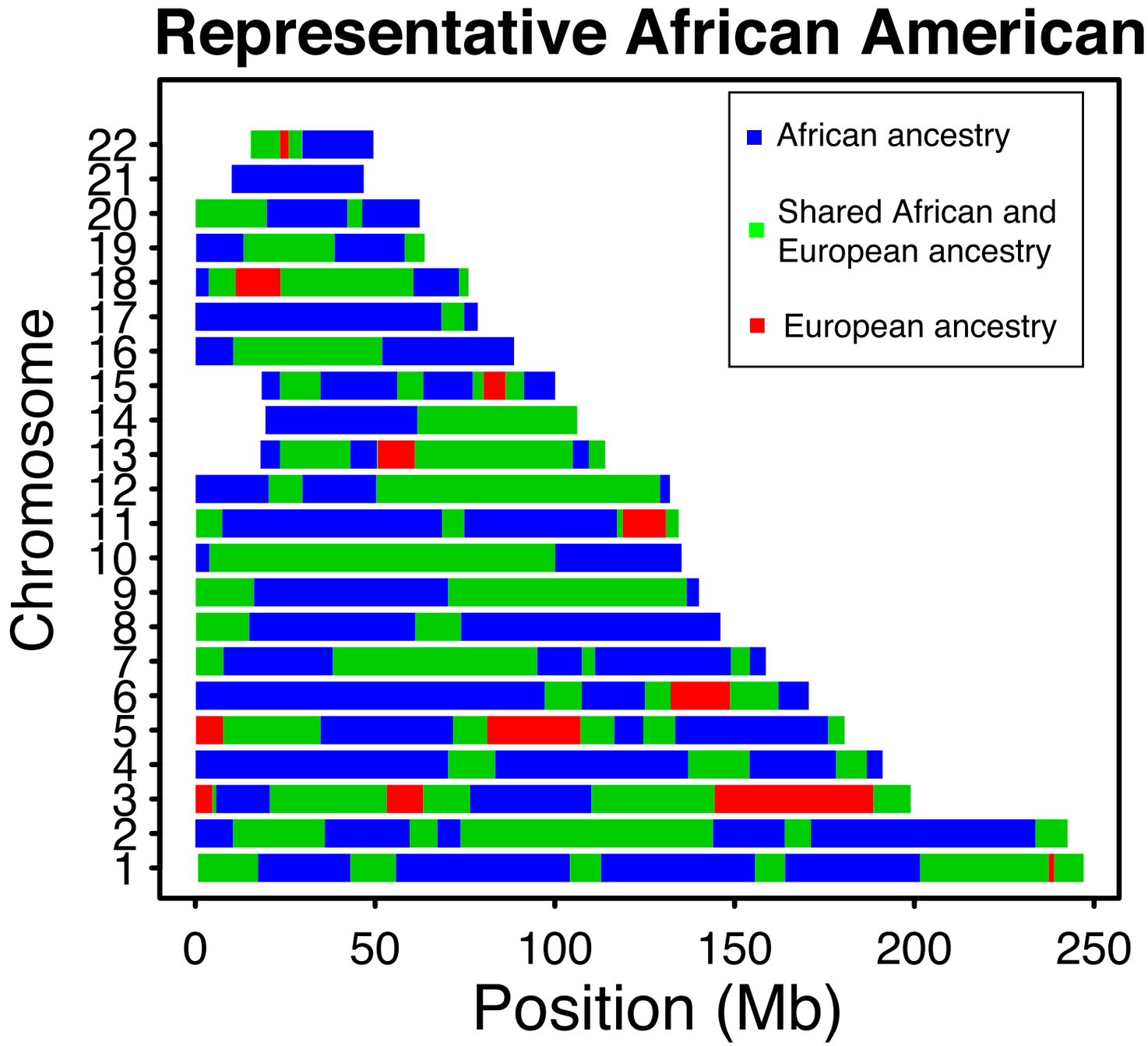
Outline of today's lecture

- Motivation and Background
  - Admixture in human history
  - Overview of methods

- Testing for admixture using f-statistics
  - Basic concepts
  - What can we test with f-statistics?
  - Estimation and interpretation
  - Admixture graph fitting

- Admixture dating
  - Basic concepts
  - ALDER

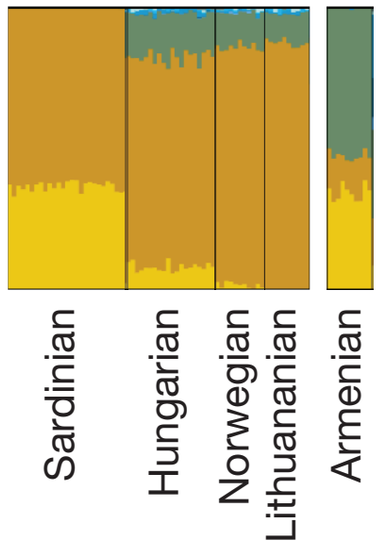
Pervasive admixture in human history



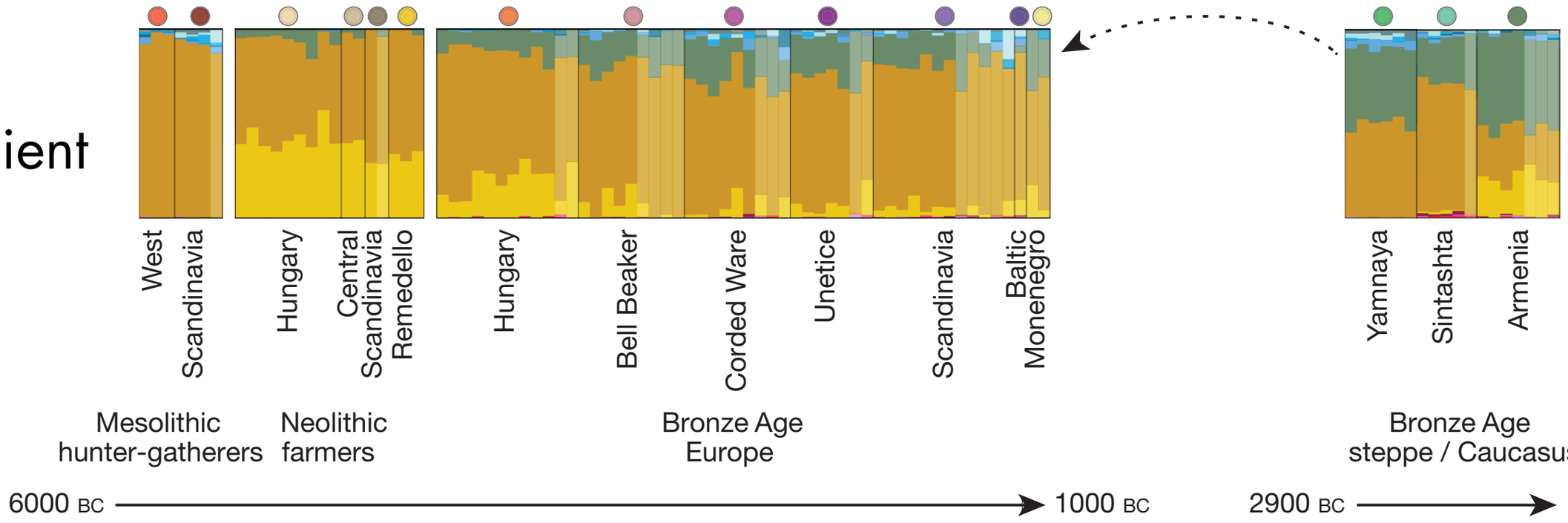
Bryc et al. (2009 PNAS)

Pervasive admixture in human history

Modern

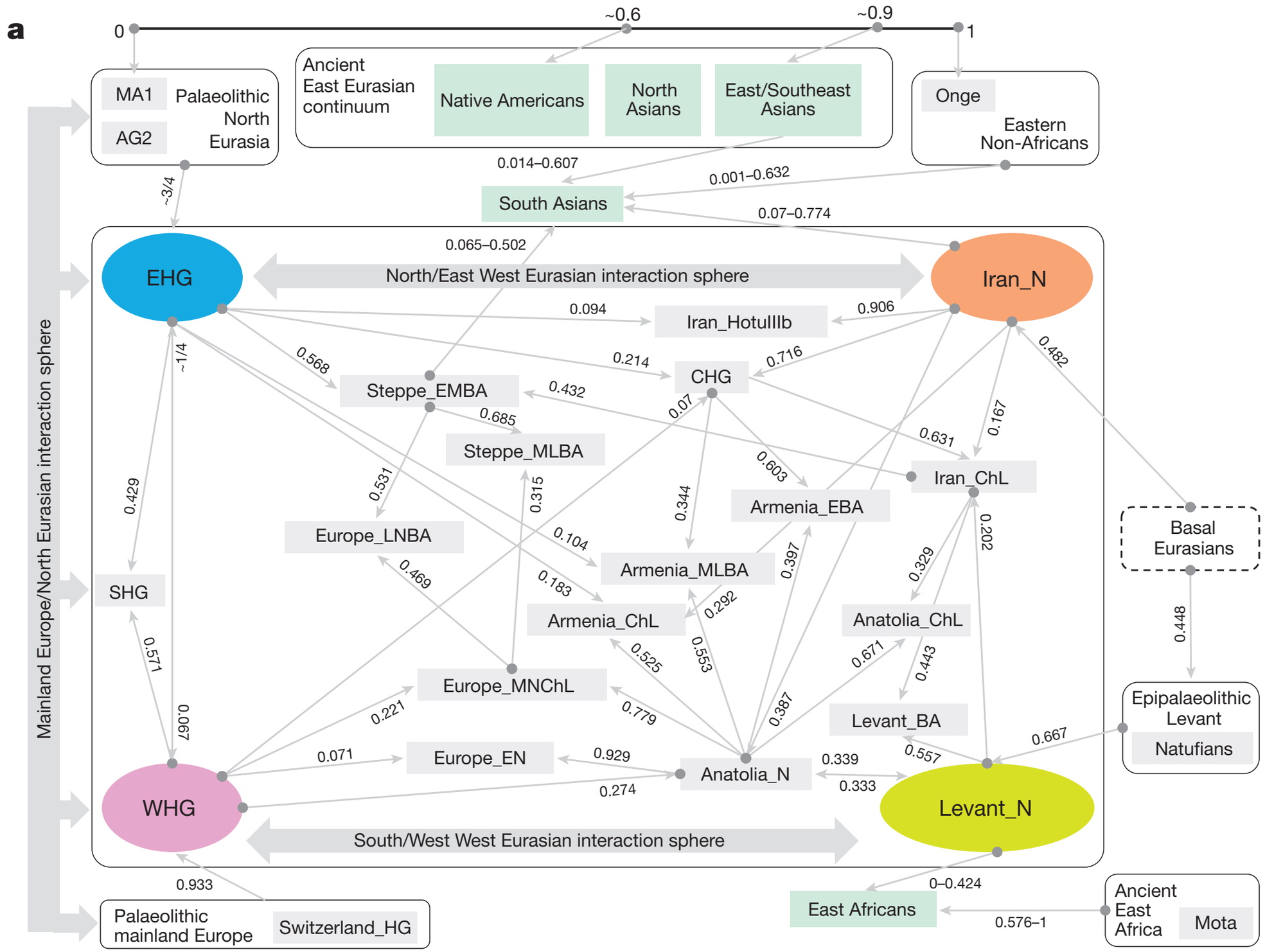


Ancient



Allentoft*, Sikora* et al. (2015) Nature

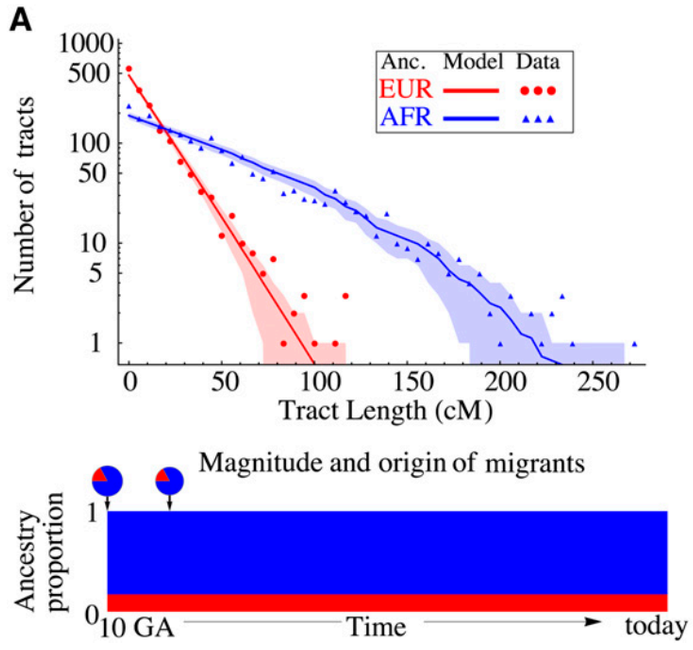
Pervasive admixture in human history



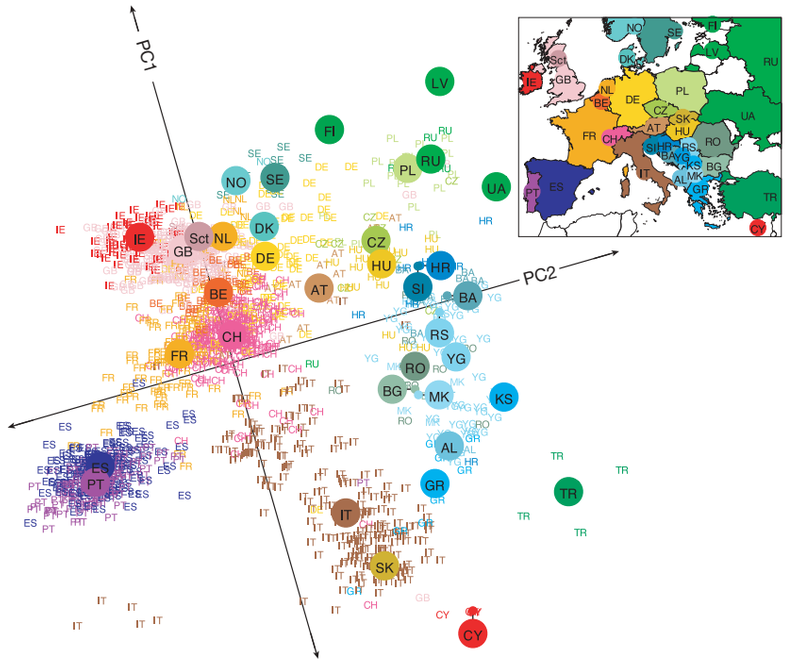
Lazaridis et al. (2017) Nature

Methods to detect admixture

- Local methods
  - Local ancestry tracts (HAPMIX, TRACTS…)
    - Allow for complex admixture models
    - Need phased data
    - Less powerful for older admixture events



- Global methods
  - PCA, Model-based clustering (ADMIXTURE, STRUCTURE…)
    - Efficient and powerful method to detect structure in the data
    - Better at handling missing and lower-quality data
    - Difficult to interpret, or easy to over-interpret



Gravel (2012) Genetics; Novembre (2008) Nature

Outline of today's lecture

- Motivation and Background
  - Admixture in human history
  - Overview of methods

- Testing for admixture using f-statistics
  - Basic concepts
  - What can we test with f-statistics?
  - Estimation and interpretation
  - Admixture graph fitting

- Admixture dating
  - Basic concepts
  - ALDER

F-statistics

Ancient Admixture in Human History

Nick Patterson,*,[1] Priya Moorjani,[†] Yontao Luo,[‡] Swapan Mallick,[†] Nadin Rohland,[†] Yiping Zhan,[‡] Teri Genschoreck,[‡] Teresa Webster,[‡] and David Reich*,[†]

*Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, [†]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, and [‡]Affymetrix, Inc., Santa Clara, California 95051

Admixture, Population Structure, and *F*-Statistics

Benjamin M. Peter[1]

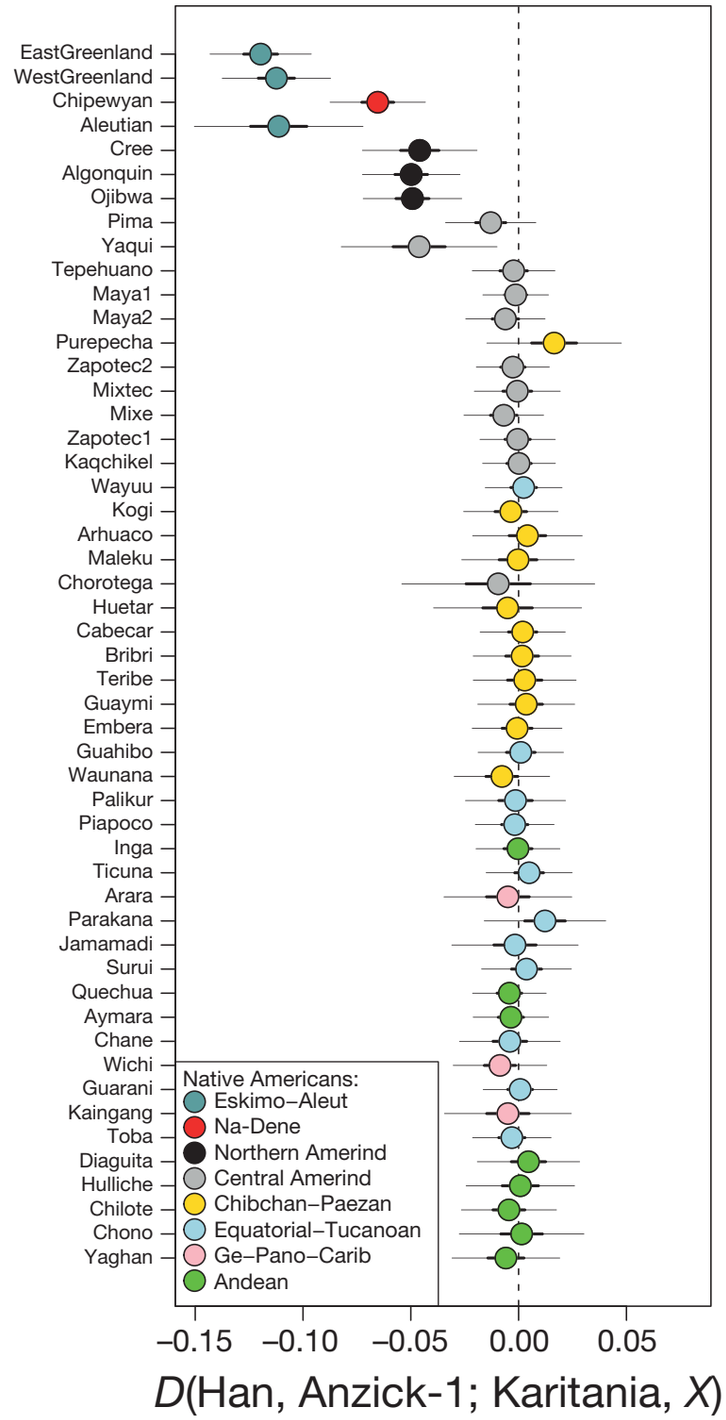
Department of Human Genetics, University of Chicago, Chicago, Illinois 60637

ORCID ID: 0000-0003-2526-8081 (B.M.P.)

- Framework for admixture inference using allele frequency correlations / shared genetic drift among sets of populations
- Parts appeared earlier in various forms in the literature, but seminal paper summarising is Patterson et al. (2012) *Genetics*
- More recently additional theoretic work by Peter (2014) *Genetics*, interpreting f-statistics in the context of population genetics theory
- Has become a standard toolset to test hypotheses about population history and admixture

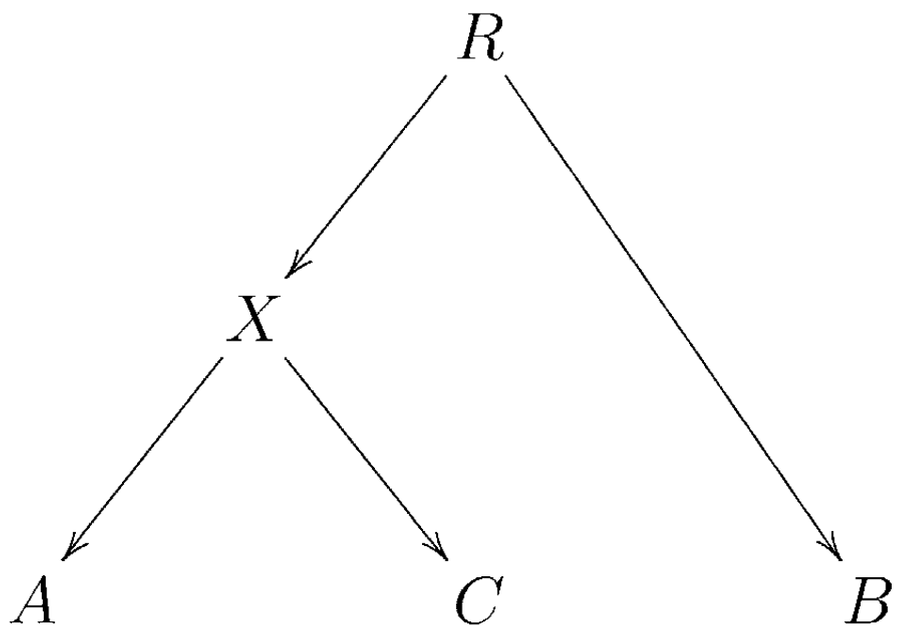
What can f-statistics be used for?

- Formal test for admixture (f_3)
- Test for treeness (f_4)
- Estimation of admixture proportions (f_4 ratio)
- Admixture graph fitting
- Finding of closest related populations (outgroup f_3)
- Test for number of migration waves in a region (qpWave)
- Phylogeny-free admixture fitting (qpAdm)

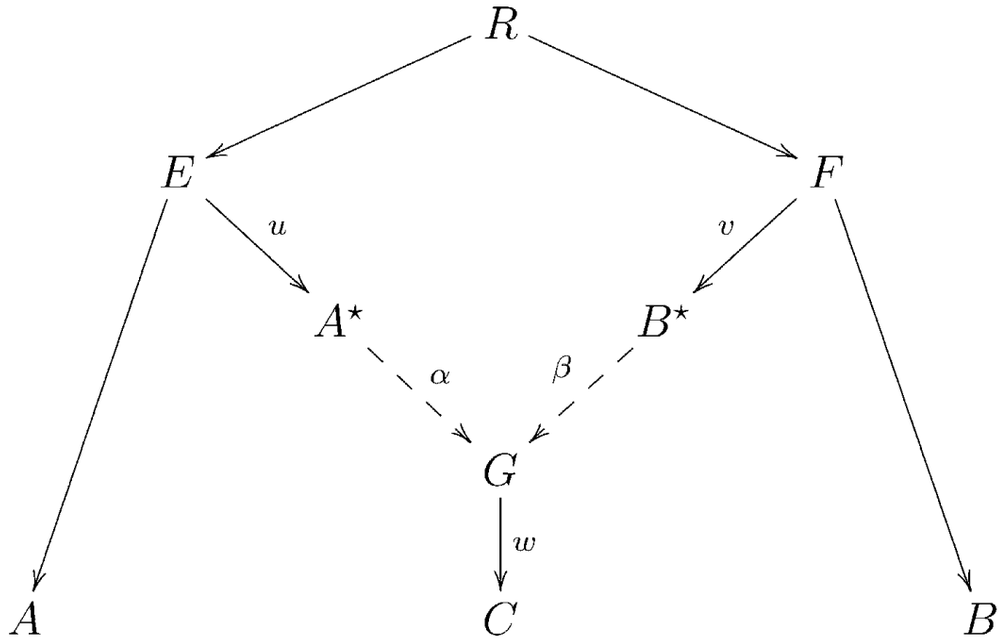


Rasmussen (2014) Nature

Definitions



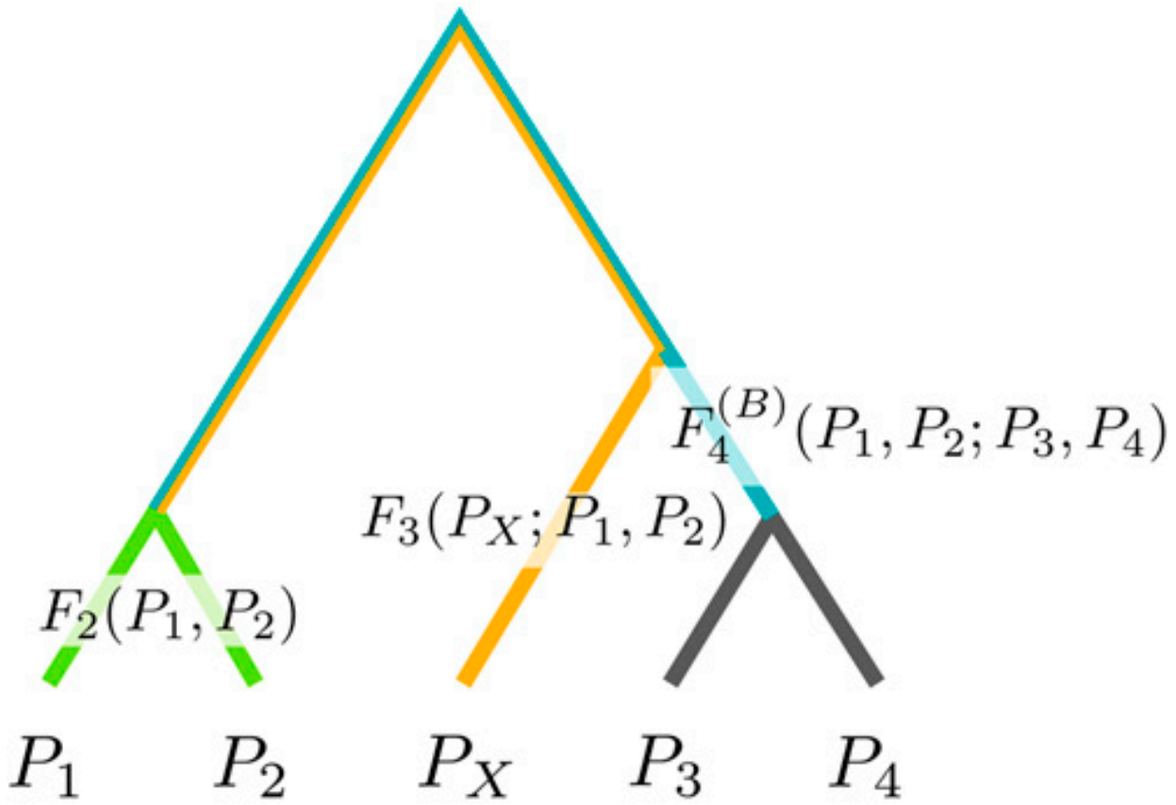
Population phylogeny



Admixture graph

Patterson et al (2012) Genetics

Definitions



$$\frac{F_2(P_1, P_2)}{\mathbb{E}[(p_1 - p_2)^2]}$$

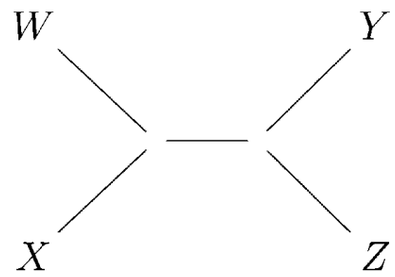
$$\frac{F_3(P_X; P_1, P_2)}{\mathbb{E}(p_X - p_1)(p_X - p_2)}$$

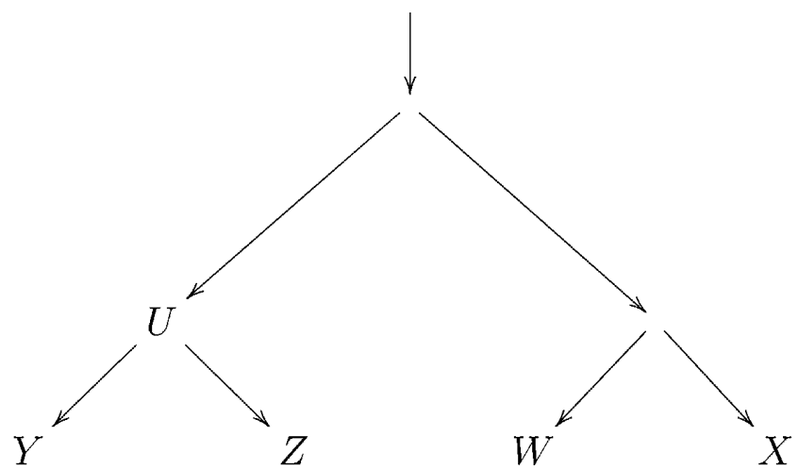
$$\frac{F_4(P_1, P_2, P_3, P_4)}{\mathbb{E}(p_1 - p_2)(p_3 - p_4)}$$

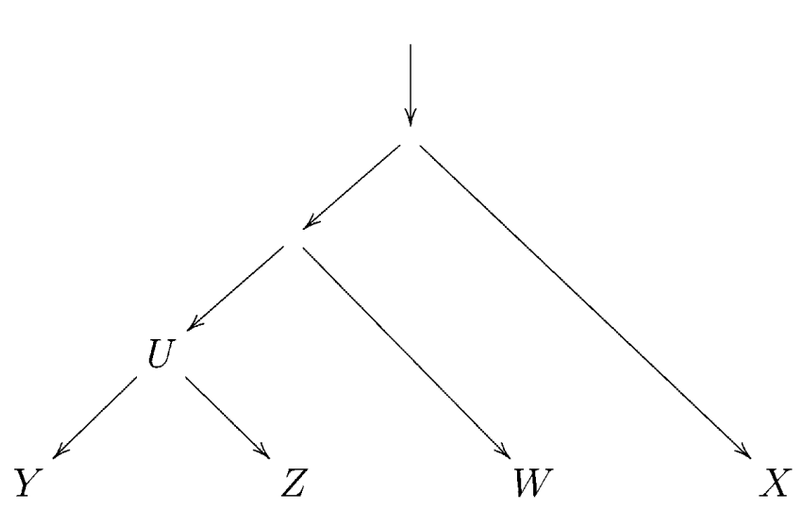
Interpretation in terms of branch lengths on a population phylogeny

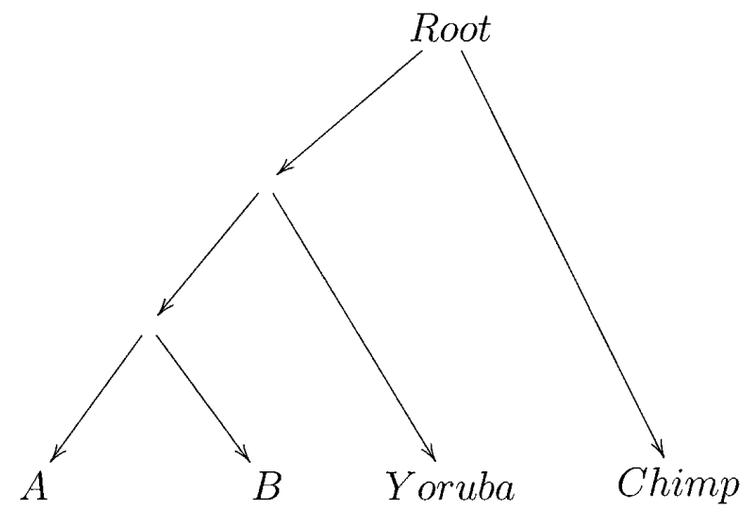
Peter (2014) Genetics

Expectations and usage - f_4









f_4 tests whether four populations are related through a simple tree

Test for "treeness"

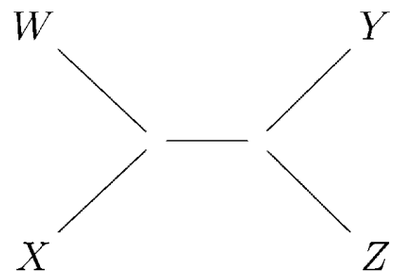
Patterson et al. (2012) Genetics

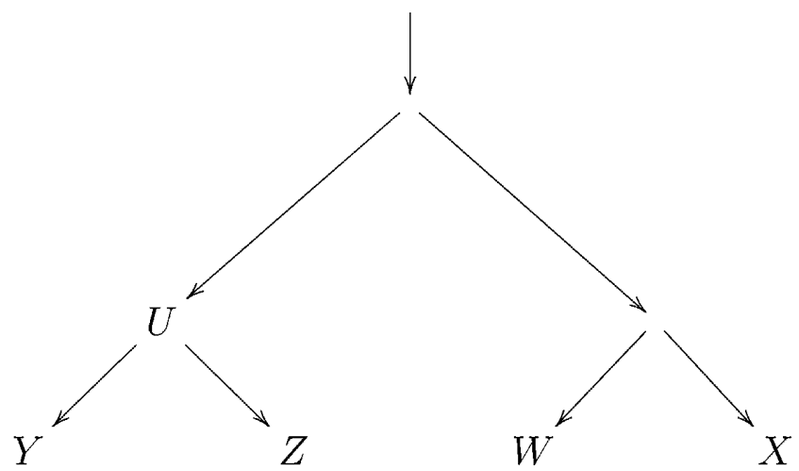
Expectations and usage - f_4

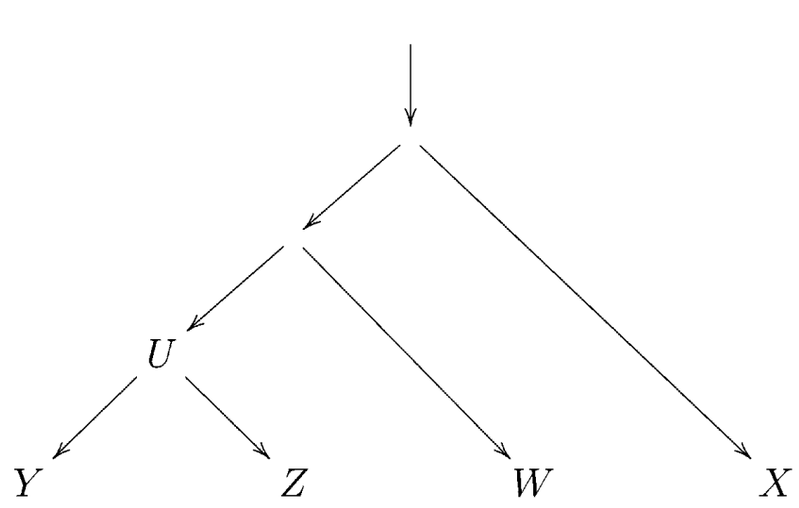
f_4 expectations on whiteboard

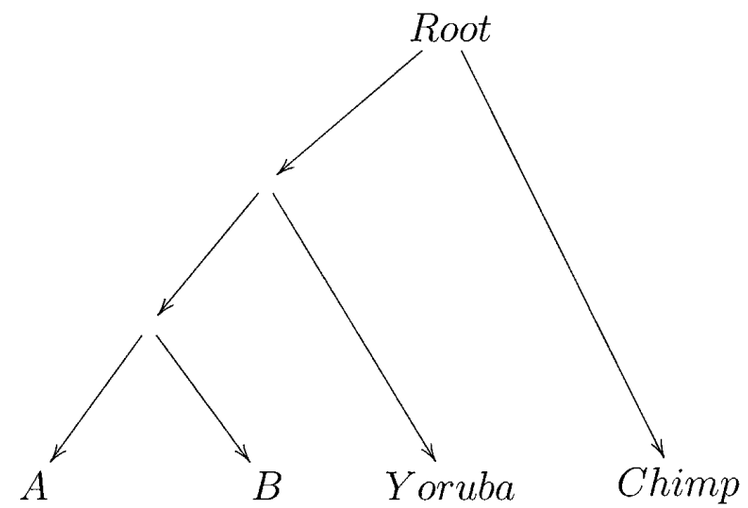
Patterson et al. (2012) Genetics

Expectations and usage - f_4







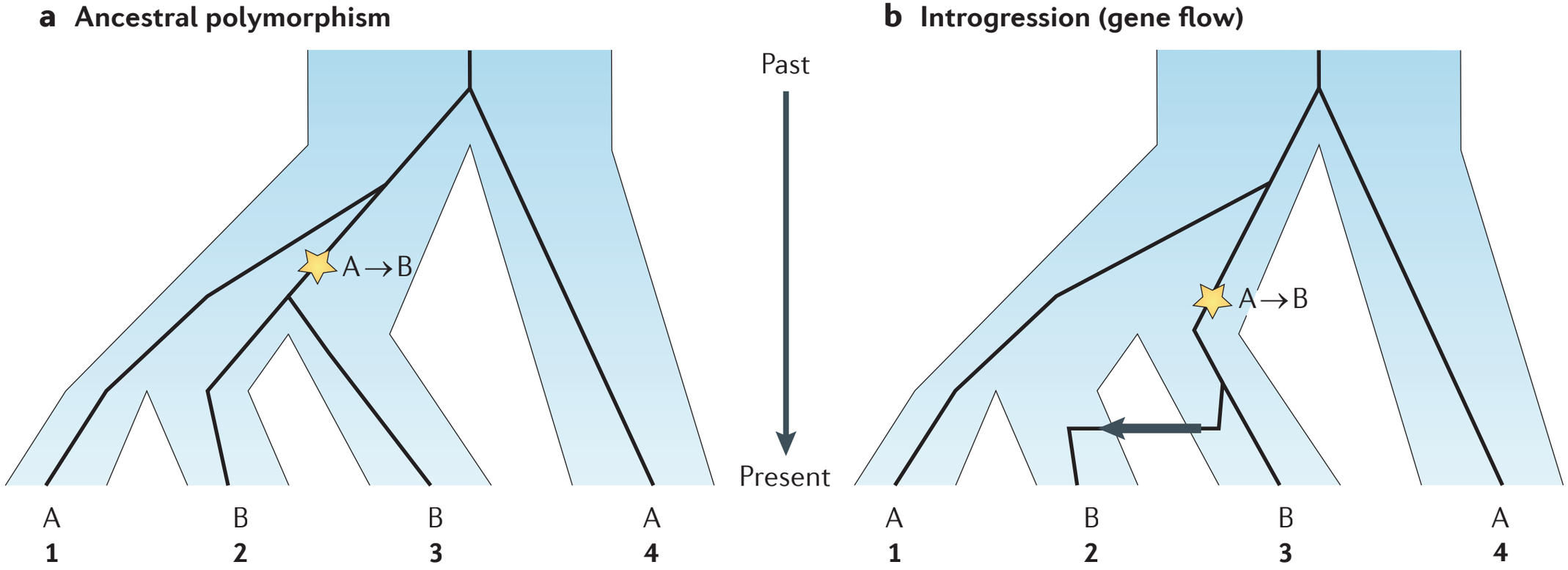


f_4 tests whether four populations are related through a simple tree

Test for "treeness"

Patterson et al. (2012) Genetics

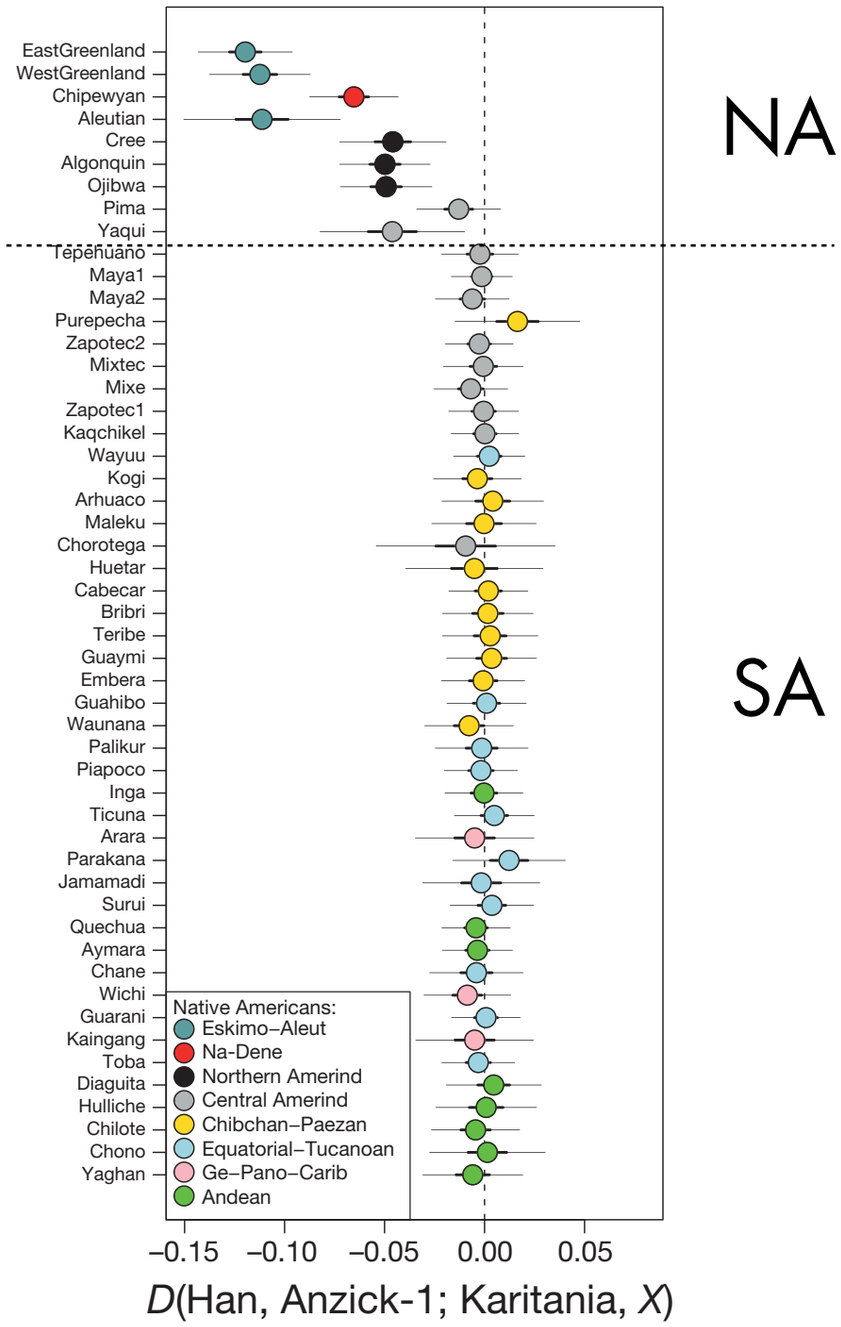
f_4 is closely related to D-test / ABBA-BABA test



Test for Neanderthal admixture is test of treeness of modern humans

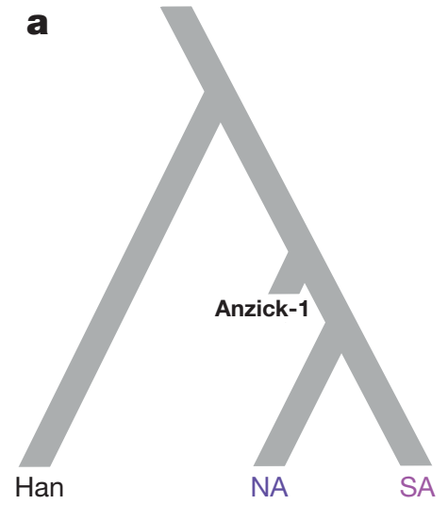
Sousa and Hey (2013) Nat Rev Genet

Interpretation of failed treeness tests

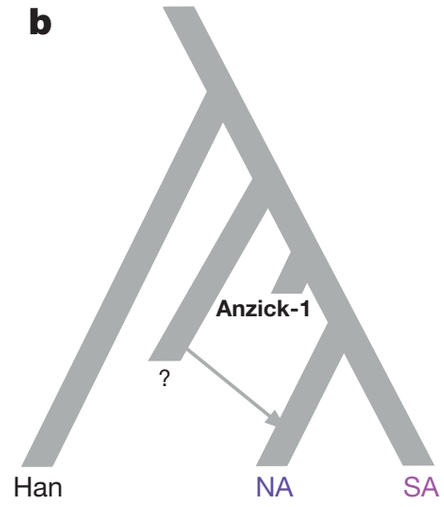


NA

SA

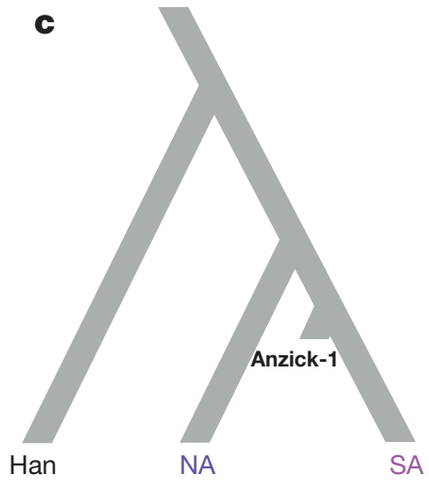


Rejected

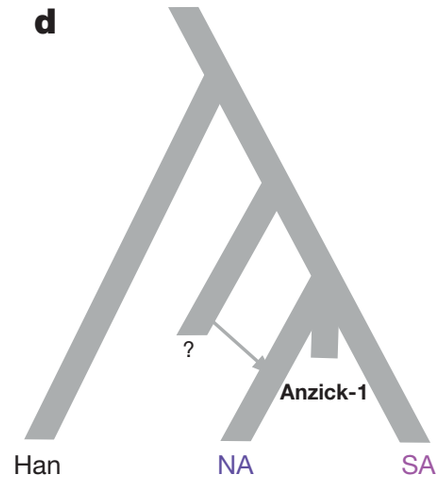


Not supported by the data

D(Han, NA; **Anzick-1**, SA) = 0.005 ± 0.0060, *Z* = 0.87



Consistent with the data



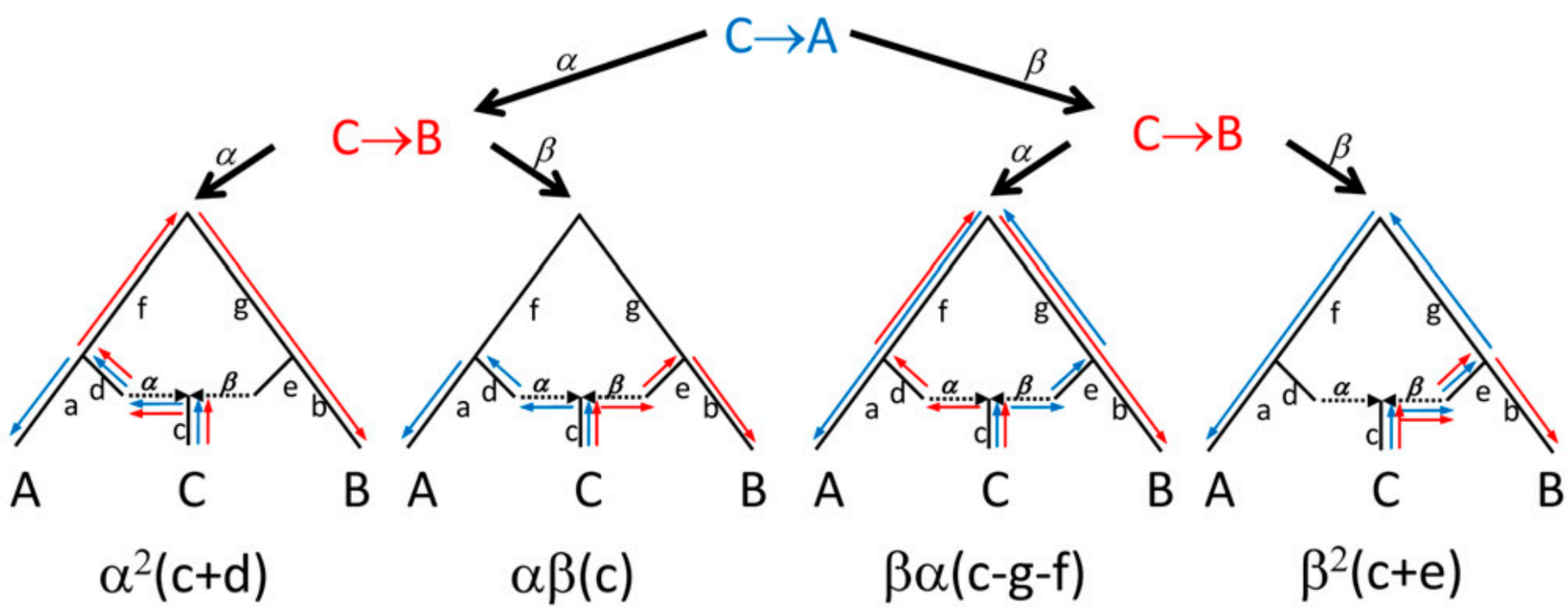
Consistent with the data

Failed treeness test can be interpreted in multiple ways!

Rasmussen (2014) Nature

Expectations and usage - f_3

$$F_3(C;A,B) = c + \alpha^2 d + \beta^2 e - \alpha\beta(g+f)$$



f_3 tests whether a target population C is the product of admixed from two source populations related (possibly distant) to sample populations A and B

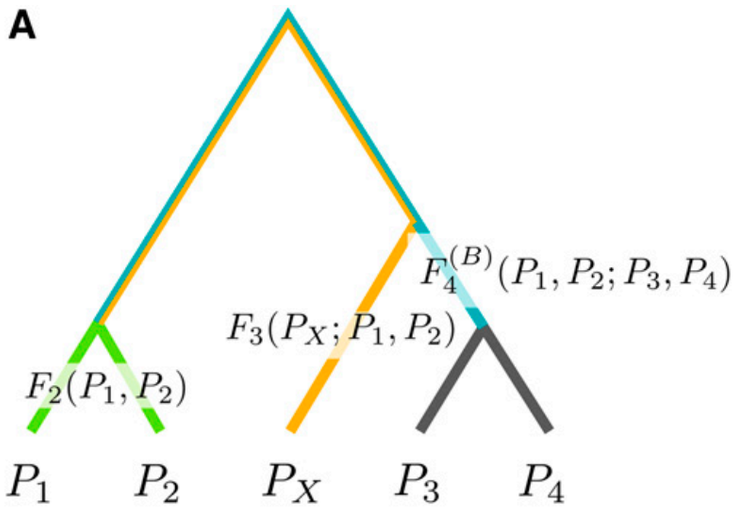
Patterson et al. (2012) Genetics

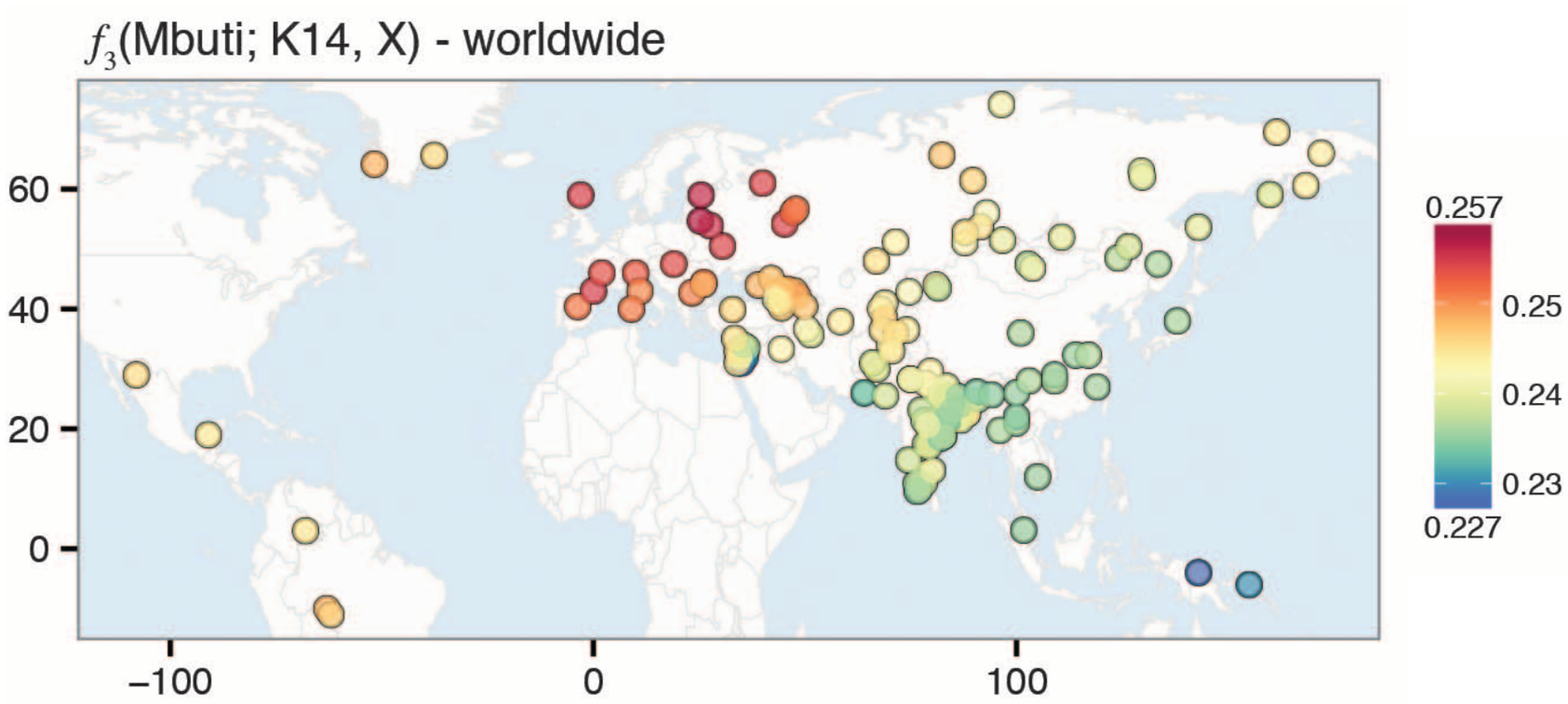
Statistical testing

F-statistic	Test	Interpretation
$f_3(X;A,B)$	$f_3 < 0$	X is admixed related to A,B
$f_4(A,B;C,D)$	$f_4 = 0$	(A,B) form a clade with respect to (C,D)
f_4-ratio	$\alpha > 0$	Admixture proportion > 0

Standard errors are estimated using a weighted block jackknife and converted into a Z-score to determine significance

Outgroup f_3 statistics



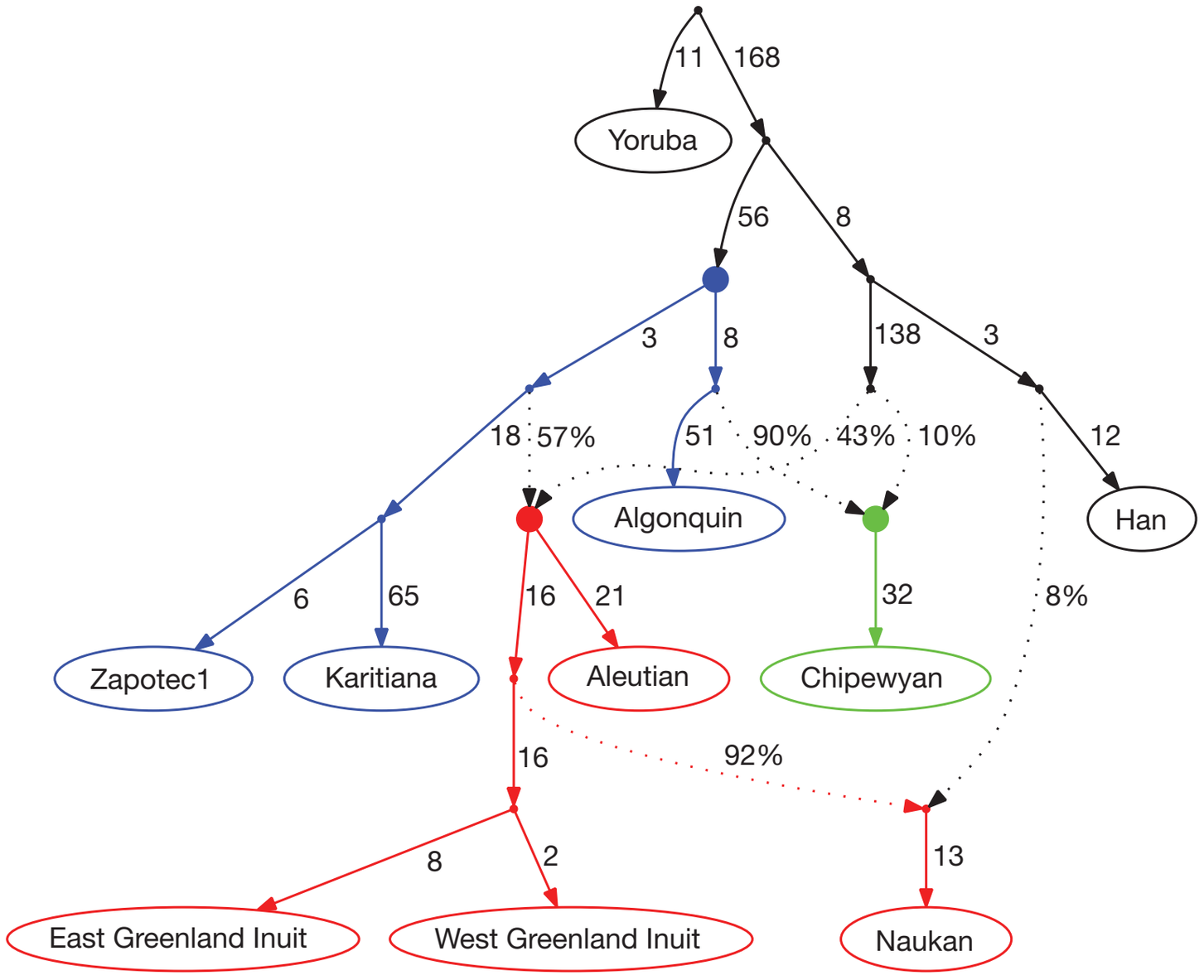


In "outgroup" f_3 statistics, the target admixed population is replaced by an outgroup population

It measures the shared drift between the outgroup and the divergence of the two other populations

Peter (2014) Genetics; Seguin-Orlando et al. (2014) Science

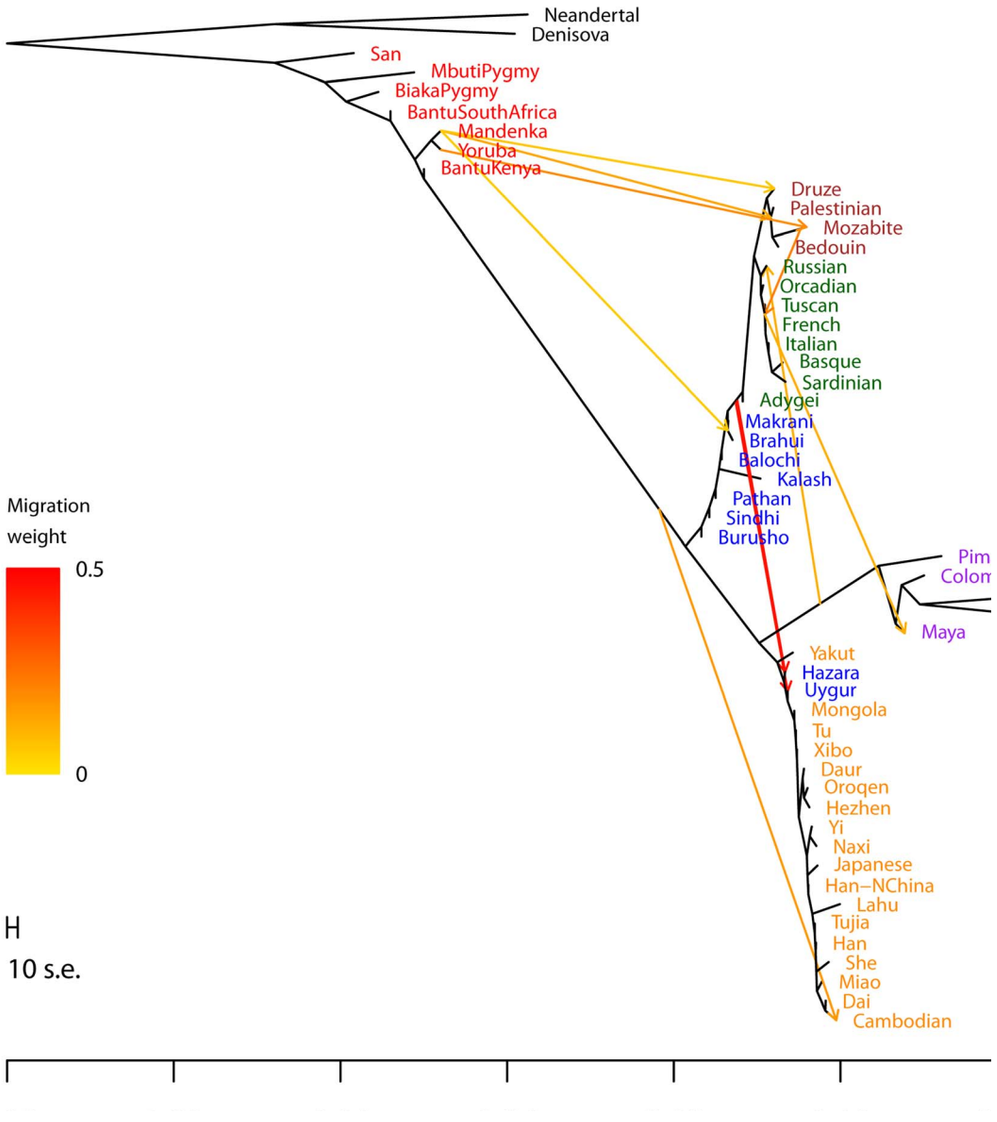
Admixture graph fitting



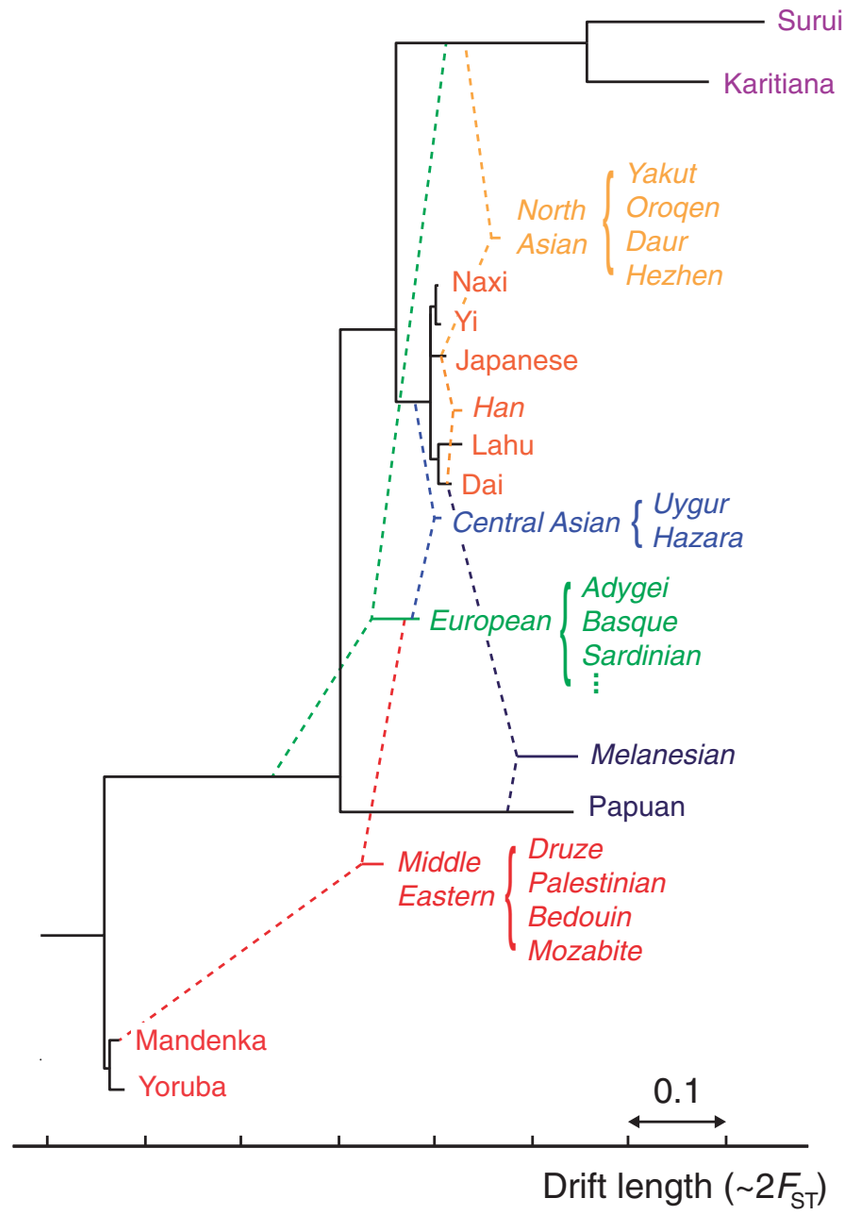
Given an admixture graph, calculate expected drift and admixture terms, and compare fit with observed values

Reich et al. (2011) Nature

Related methods



Treemix

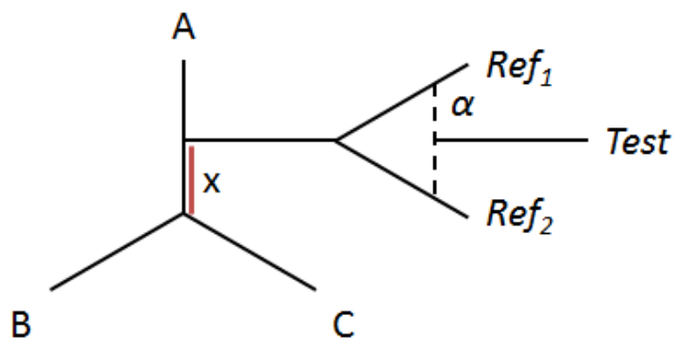


Mixmapper

Pickrell and Pritchard (2012) PLoS Genetics; Lipson (2013) Mol Biol Evol

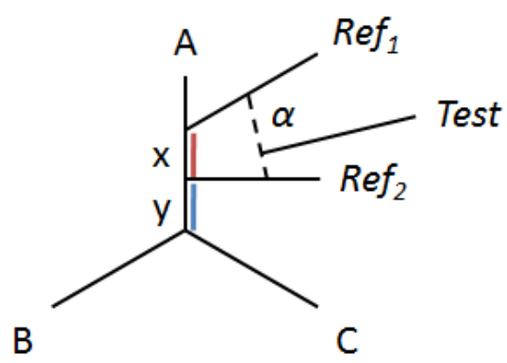
Phylogeny-free admixture fitting (qpAdm)

(a)



$f_4(Test \text{ or } Ref_1 \text{ or } Ref_2, A; B, C) = 0$
$f_4(Test \text{ or } Ref_1 \text{ or } Ref_2, B; A, C) = x$
$f_4(Test \text{ or } Ref_1 \text{ or } Ref_2, C; A, B) = x$

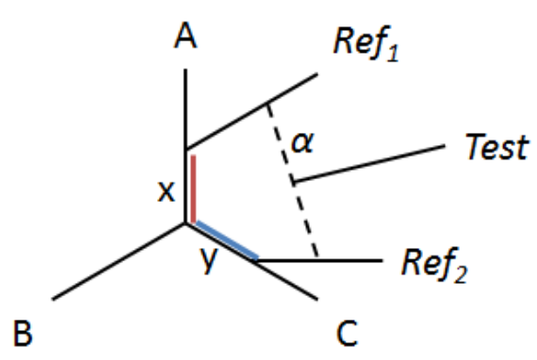
(b)



$f_4(Ref_1, A; B, C) = 0$ $\quad$ $f_4(Ref_2, A; B, C) = 0$
$f_4(Ref_1, B; A, C) = x+y$ $\quad$ $f_4(Ref_2, B; A, C) = y$
$f_4(Ref_1, C; A, B) = x+y$ $\quad$ $f_4(Ref_2, C; A, B) = y$

$f_4(Test, A; B, C) = 0$
$f_4(Test, B; A, C) = y+\alpha x=(1-\alpha) f_4(Ref_2, B; A, C)+\alpha f_4(Ref_1, B; A, C)$
$f_4(Test, C; A, B) = y+\alpha x=(1-\alpha) f_4(Ref_2, C; A, B)+\alpha f_4(Ref_1, C; A, B)$

(c)



$f_4(Ref_1, A; B, C) = 0$ $\quad$ $f_4(Ref_2, A; B, C) = -y$
$f_4(Ref_1, B; A, C) = x$ $\quad$ $f_4(Ref_2, B; A, C) = -y$
$f_4(Ref_1, C; A, B) = x$ $\quad$ $f_4(Ref_2, C; A, B) = 0$

$f_4(Test, A; B, C) = -(1-\alpha)y = (1-\alpha) f_4(Ref_2, A; B, C)+\alpha f_4(Ref_1, A; B, C)$
$f_4(Test, B; A, C) = \alpha x-(1-\alpha)y = \alpha f_4(Ref_1, A; B, C)+(1-\alpha) f_4(Ref_2, A; B, C)$
$f_4(Test, C; A, B) = \alpha x = \alpha f_4(Ref_1, A; B, C)+(1-\alpha) f_4(Ref_2, A; B, C)$

$$f_4(Test; A; B, C) \approx \sum_{i=1}^{N} \alpha_i f_4(Ref_i; A; B, C)$$

Differential drift sharing with sets of outgroup populations is leveraged to infer mixture proportions

Haak et al (2015) Nature

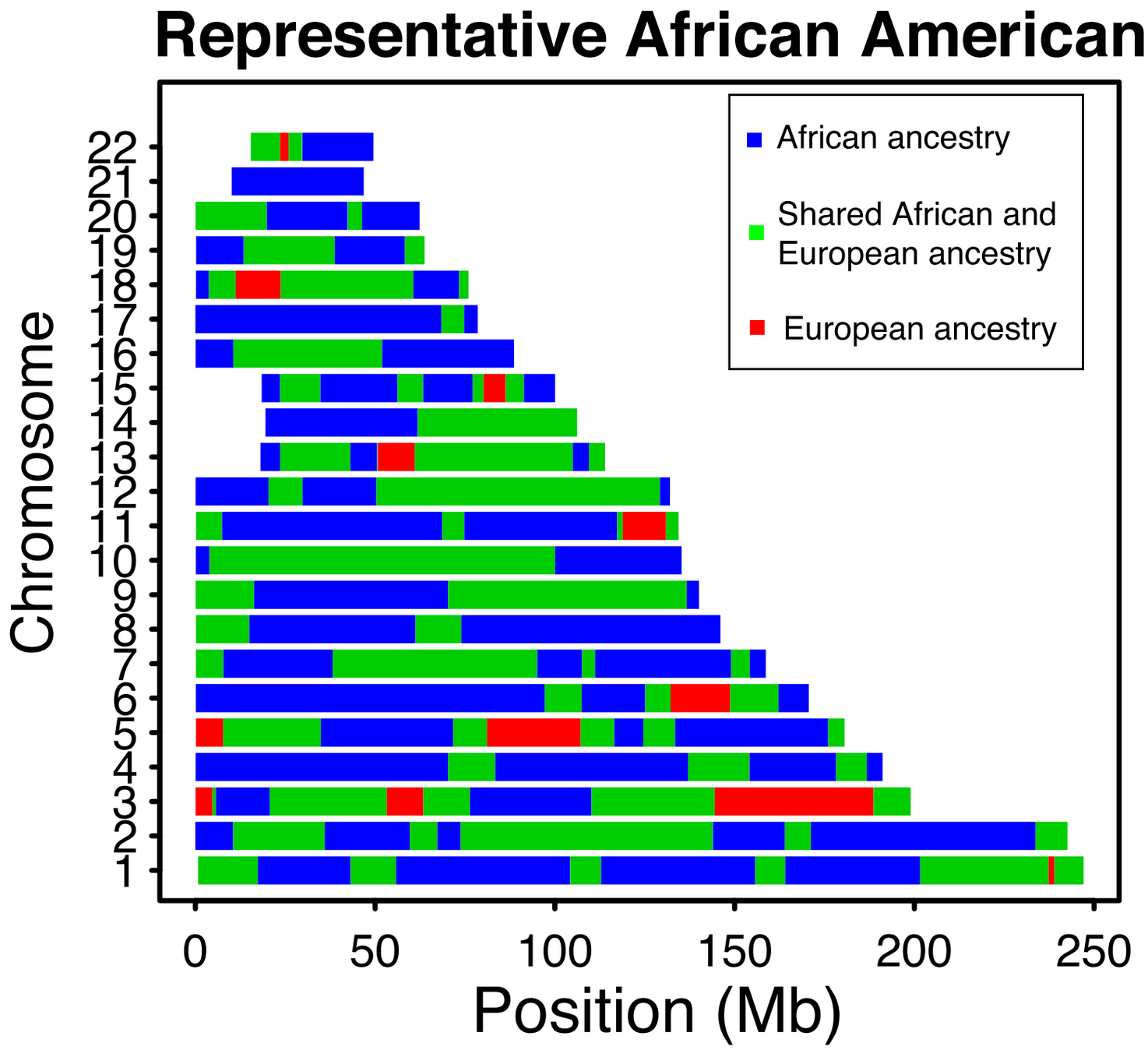
Outline of today's lecture

- Motivation and Background
  - Admixture in human history
  - Overview of methods

- Testing for admixture using f-statistics
  - Basic concepts
  - What can we test with f-statistics?
  - Estimation and interpretation
  - Admixture graph fitting

- Admixture dating
  - Basic concepts
  - ALDER

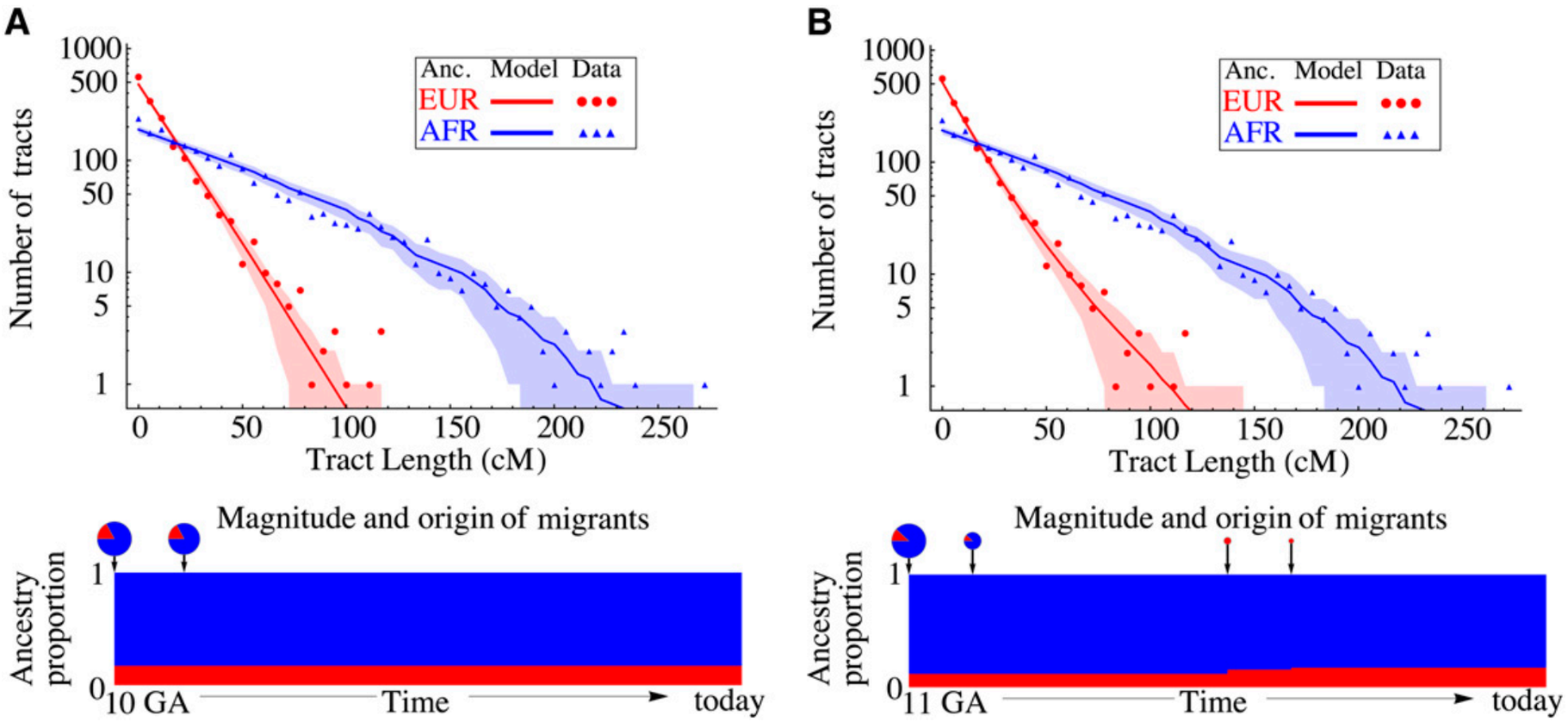
Admixture Linkage disequilibrium



The length distribution of ancestry tracts is informative about time since admixture

Bryc et al. (2009 PNAS)

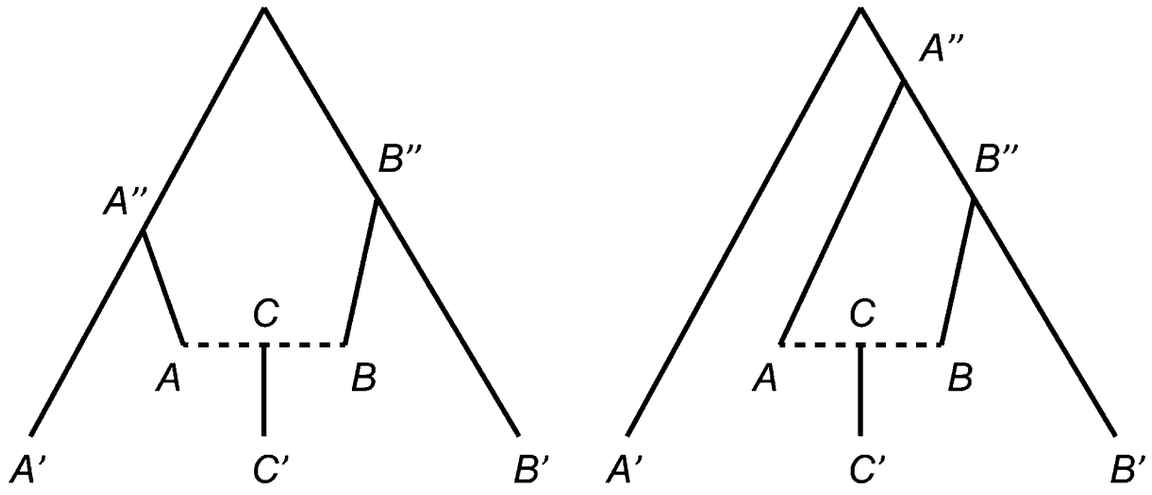
Local methods

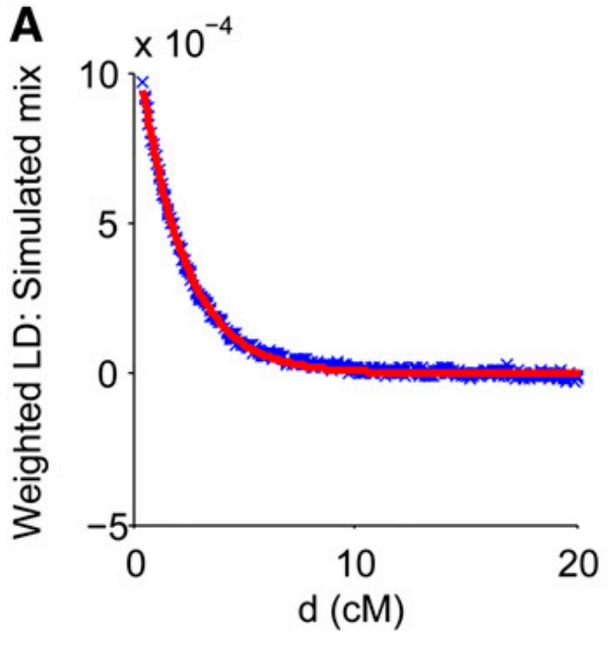


- Allow for complex admixture models
- Need phased data
- Less powerful for older admixture events

Gravel (2009) Genetics

ALDER weighted LD model





$$D_n = e^{-nd} D_0 = e^{-nd} \alpha\beta\delta(x)\delta(y)$$

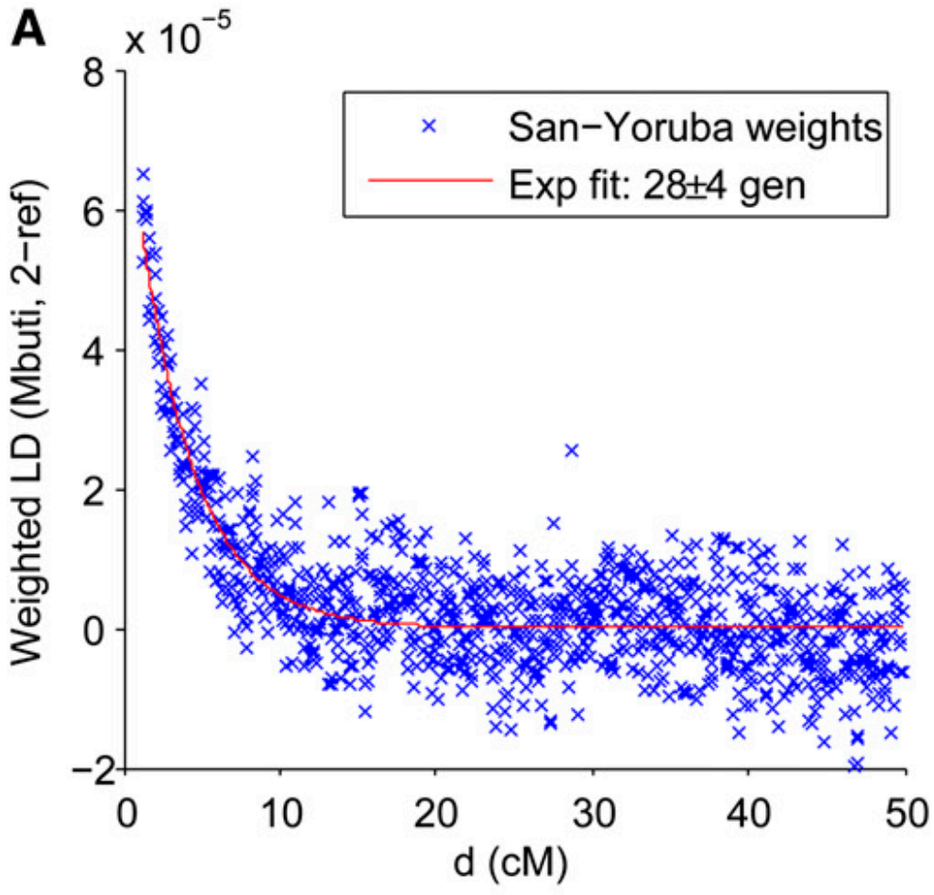
$$a(d) := \frac{\sum_{\mathcal{S}(d)} z(x,y) w(x) w(y)}{|\mathcal{S}(d)|}. \qquad E[a(d)] = 2\alpha\beta F_2\left(A'', B''\right)^2 e^{-nd},$$

LD between markers weighted by product of allele frequencies shows exponential decay with rate of generations since admixture

Loh et al (2013) Genetics

ALDER features

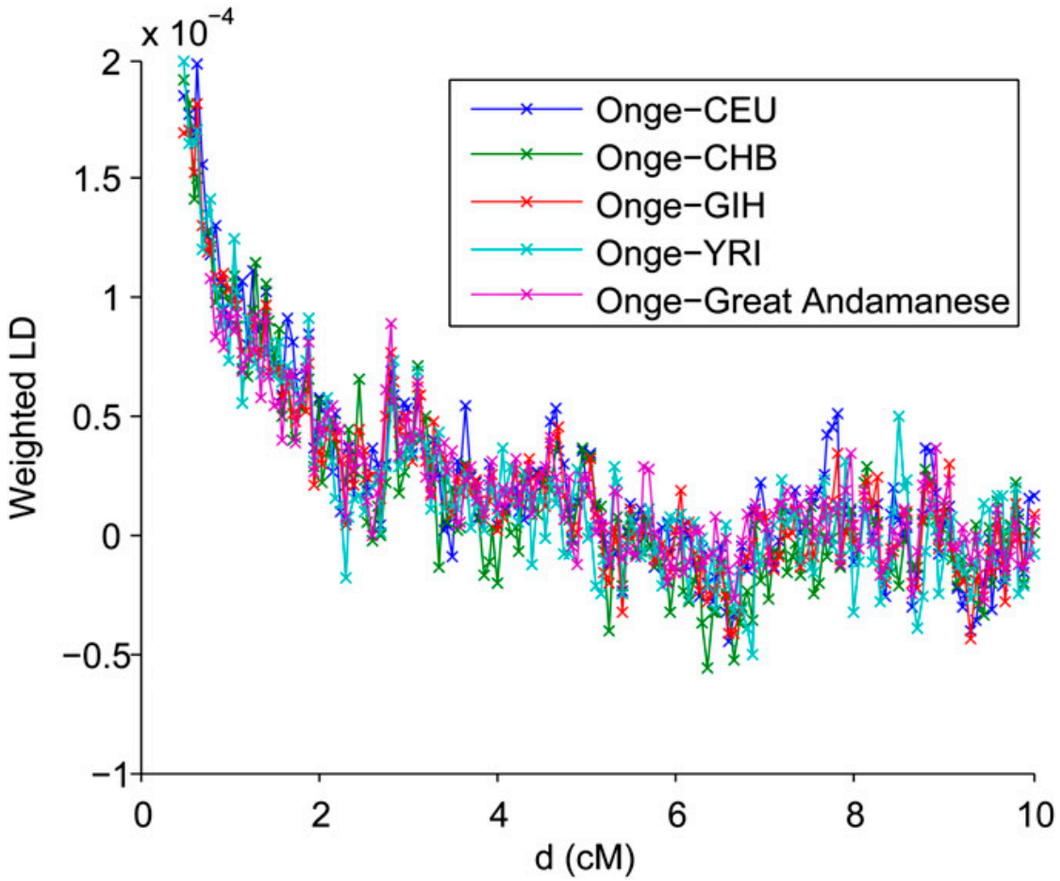
- Estimate time since admixture (decay of exponential)
- Infer phylogenic placement of different source populations (amplitude)
- Can use admixed population itself as one reference
- Phasing not needed
- Single-pulse model, although implementation for multiple pulses exists (MALDER)

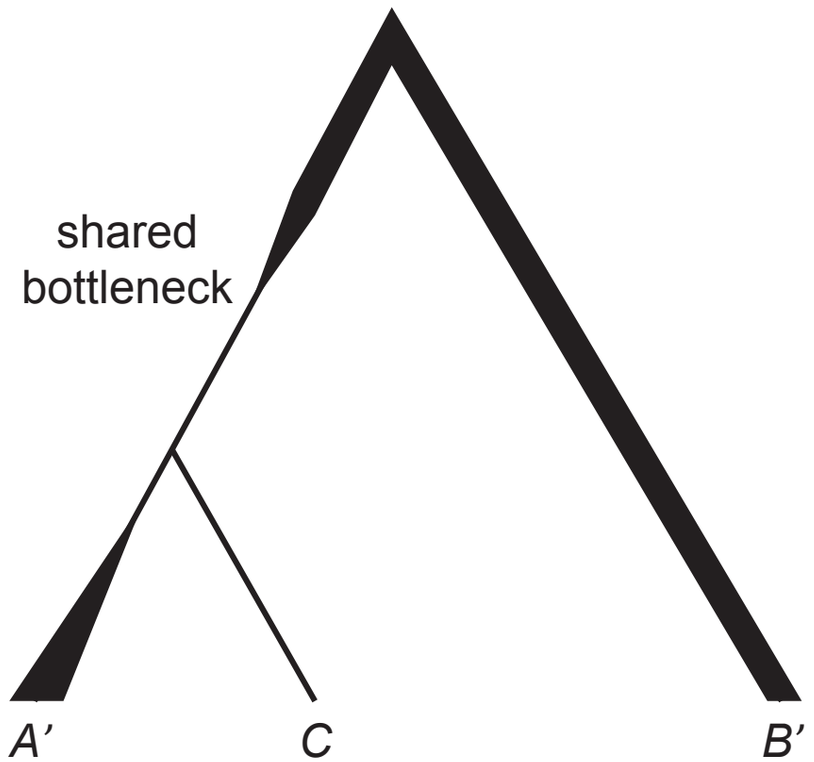


$$E[a(d)] = 2\alpha\beta F_2\left(A'', B''\right)^2 e^{-nd},$$

Loh et al (2013) Genetics

Weighted LD curves can occur without admixture



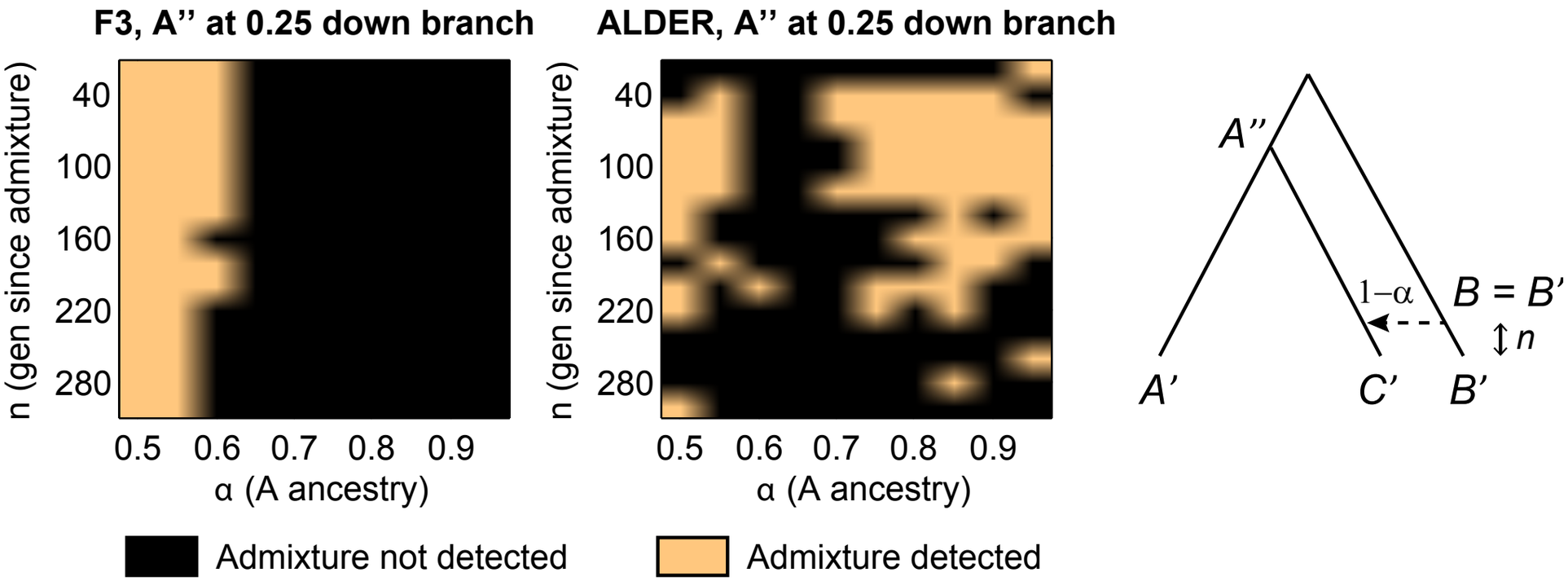


Single reference population curves have same amplitude

Two reference population curves more robust

Loh et al (2013) Genetics

ALDER vs f_3-statistics



Complementary for detecting admixture

ALDER less affected by drift since admixture or very diverged source populations

f_3 more power for older events

Loh et al (2013) Genetics

Take - home messages

- f-statistics / ALDER are versatile frameworks for testing admixture hypotheses, including quite complex scenarios
- Suitable for low quality data (e.g. aDNA)
- Care has to be taken as f-statistics are susceptible to batch effects that induce spurious allele frequency correlations

Practicals

- Practical 1
  - f3 / f4 statistics

- Practical 2
  - qpGraph
  - ALDER