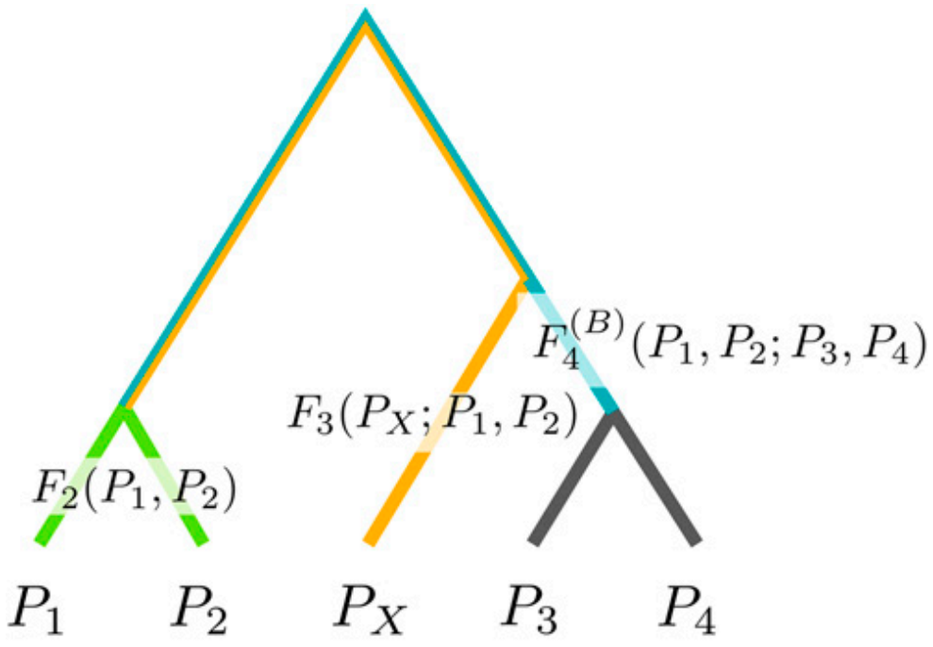
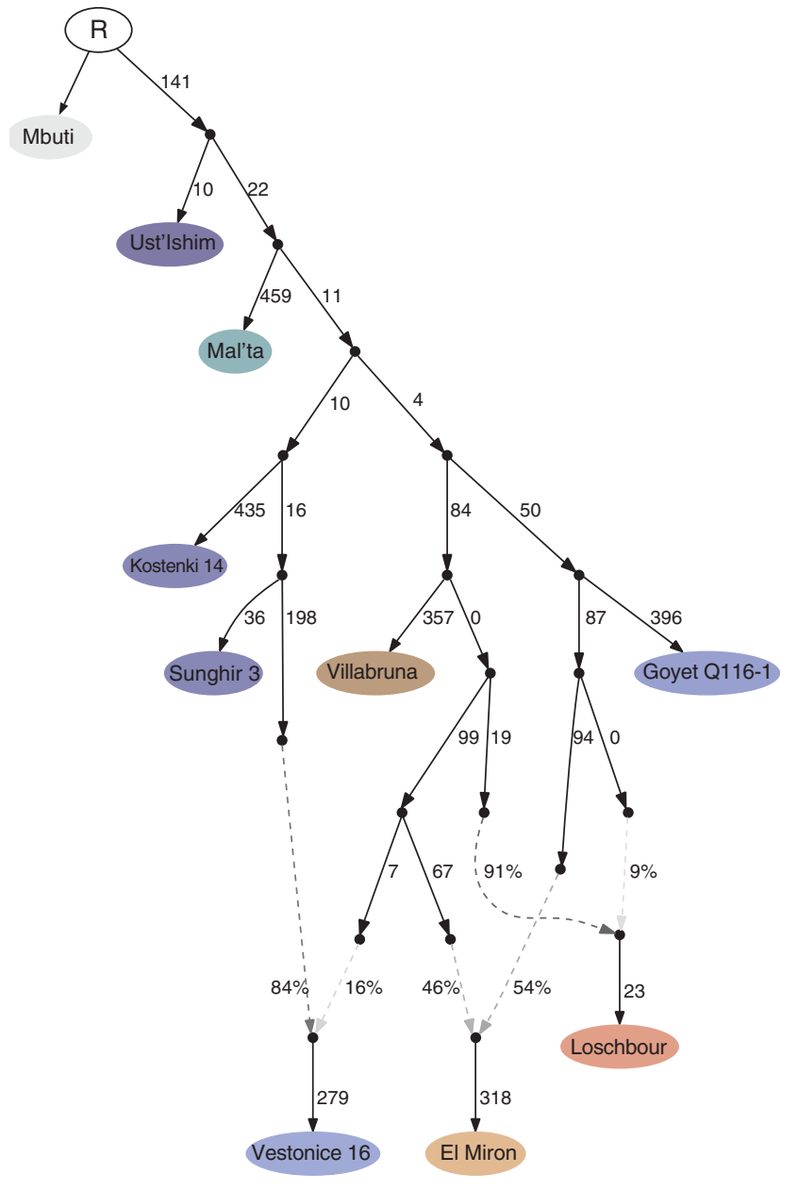
Inference of admixture from genomic data using f-statistics

Elixir Workshop "Population Genomics: Background and Tools"
Napoli 2018





Martin Sikora, PhD
Centre for Geogenetics, Natural History Museum of Denmark, University of Copenhagen

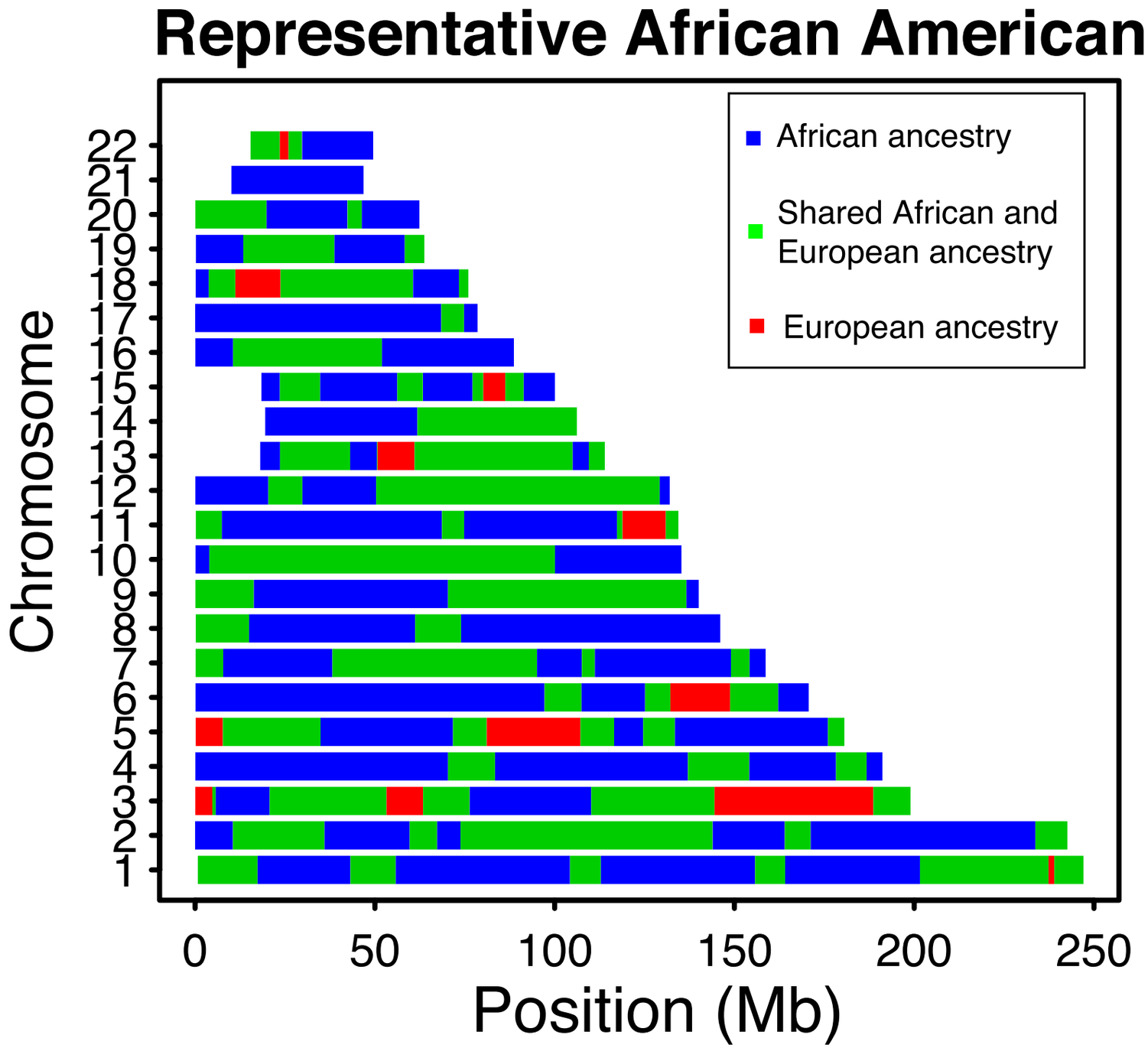
Outline of today's session

- Motivation and Background
  - Overview of methods to detect admixture

- Testing for admixture using f-statistics
  - Basic concepts
  - Estimation and usage

- Practicals
  - f3 / f4 statistics
  - qpAdm /qpWave and qpGraph

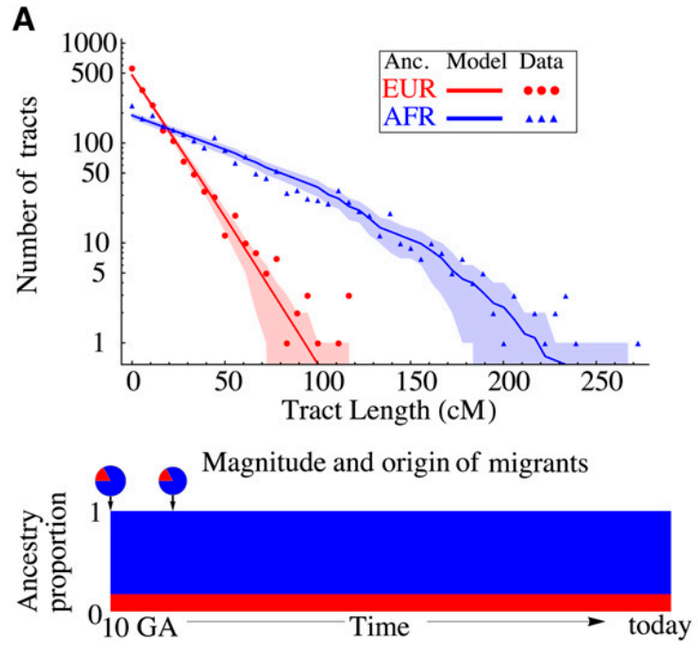
Genomic signatures of admixture



Bryc (2009) PNAS

Methods to detect admixture

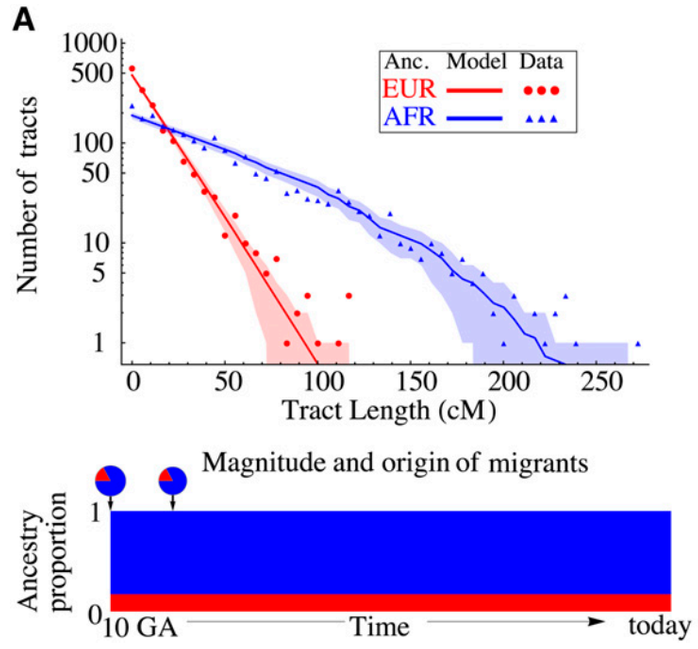
- Local methods (HAPMIX, TRACTS, Chromopainter, …)
  - Allow for complex admixture models
  - Need phased data
  - Less powerful for older admixture events



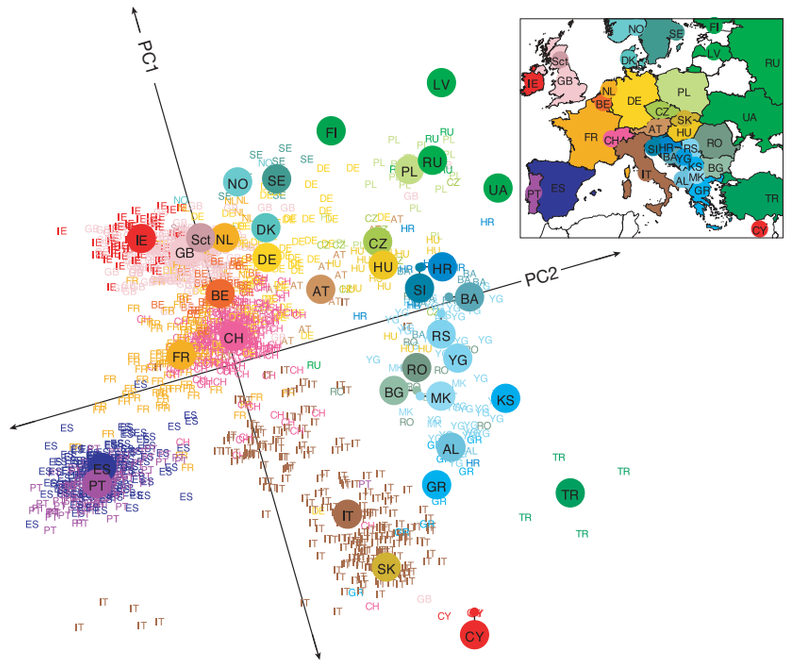
Gravel (2012) Genetics; Novembre (2008) Nature

Methods to detect admixture

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

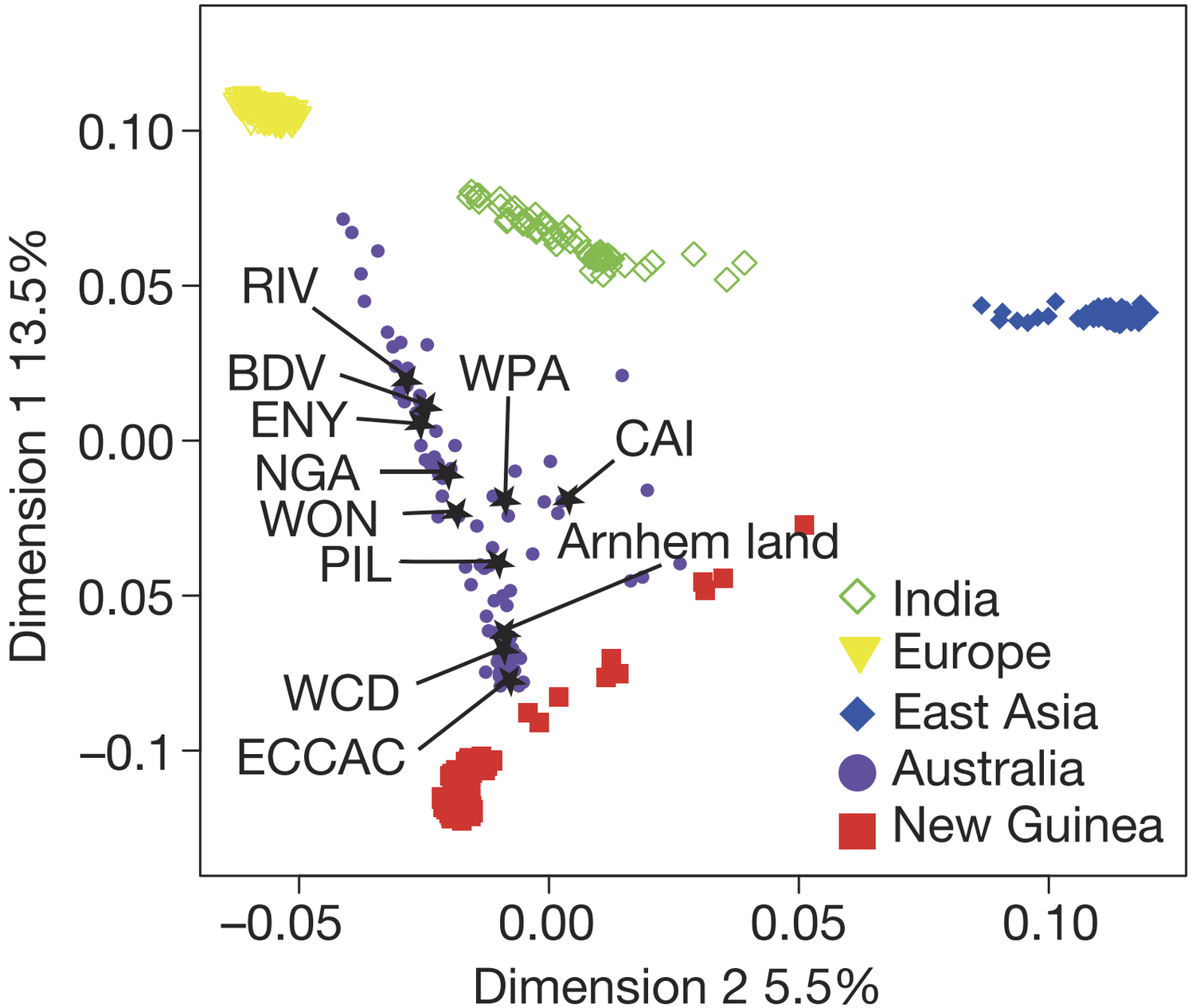


- Global methods (PCA, ADMIXTURE, STRUCTURE, …)
  - Efficient and powerful method to detect structure in the data
  - Better at handling missing and lower-quality data
  - Difficult to interpret, or easy to over-interpret



Gravel (2012) Genetics; Novembre (2008) Nature

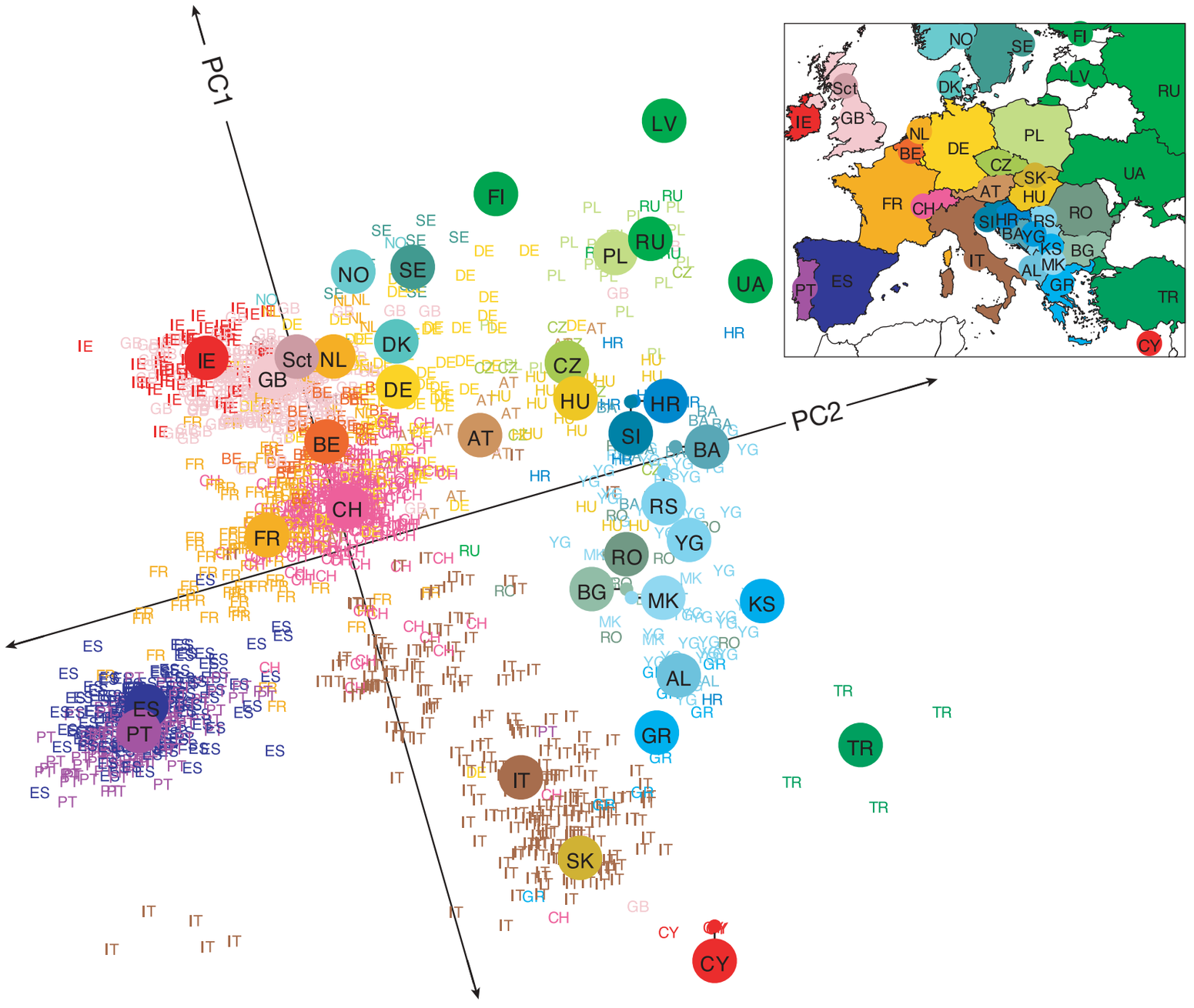
Patterns of admixture in PCA



Recent admixture between highly differentiated source populations can be detected relatively easily

Malaspinas et al. (2016 Nature)

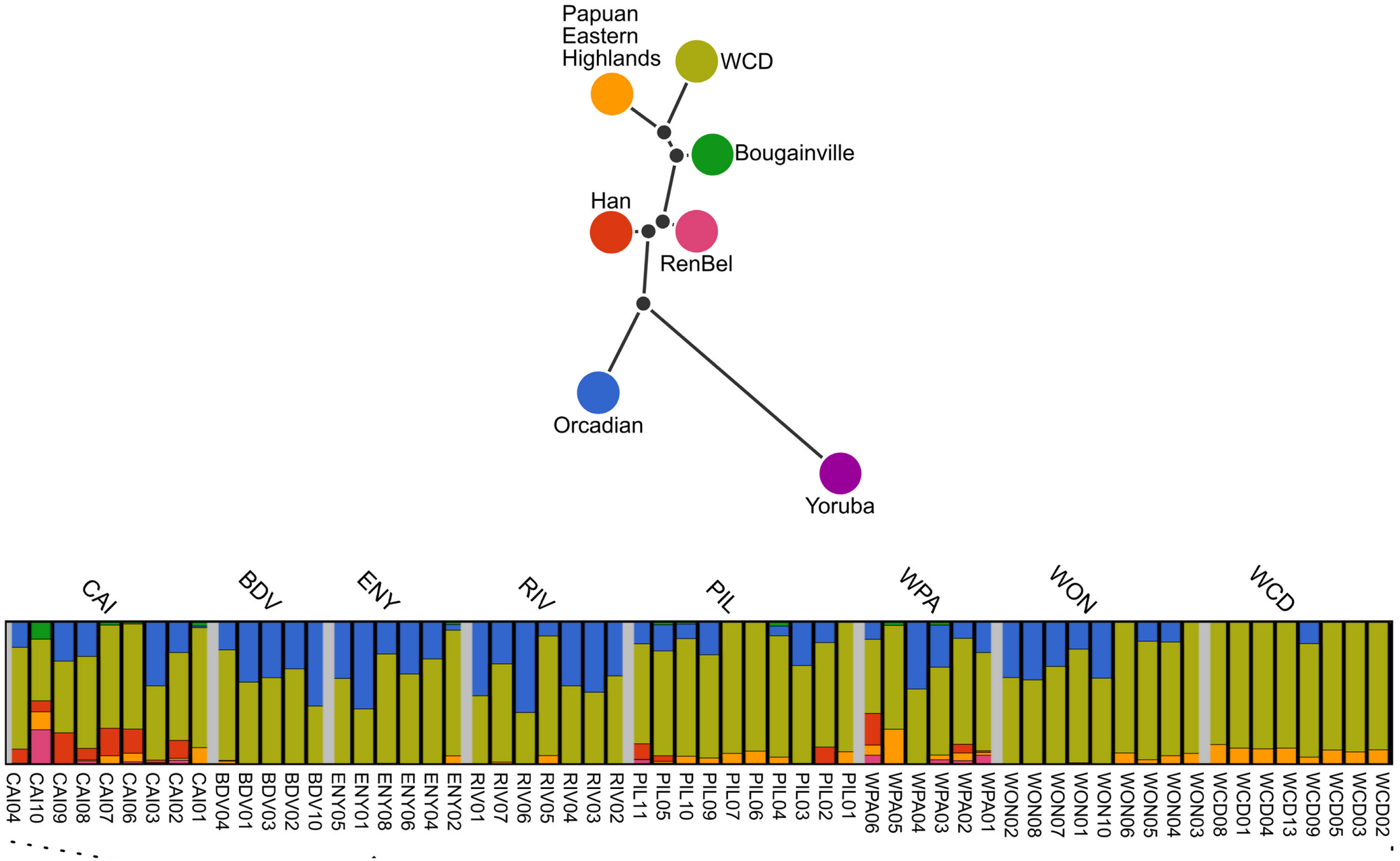
Patterns of admixture in PCA



Admixture or isolation-by-distance?

Novembre et al. (2008 Nature)

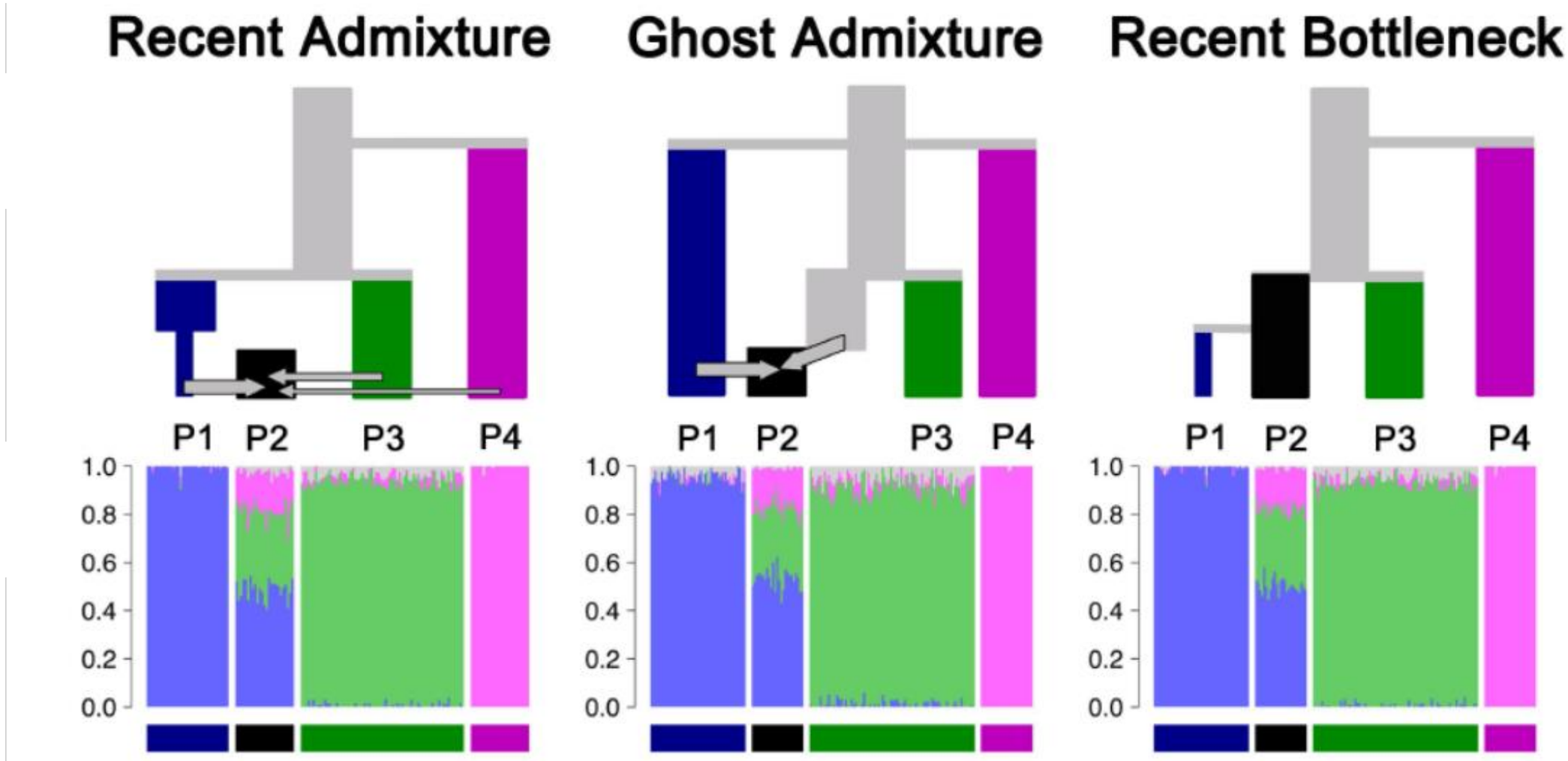
Patterns of admixture in model-based clustering



Recent admixture between highly differentiated source populations can be detected relatively easily

Malaspinas et al. (2016 Nature)

Patterns of admixture in model-based clustering



Other demographic scenarios can mimic admixture in model-based clustering

Lawson et al. (2016 biorxiv)

F-statistics - a framework for testing for admixture

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

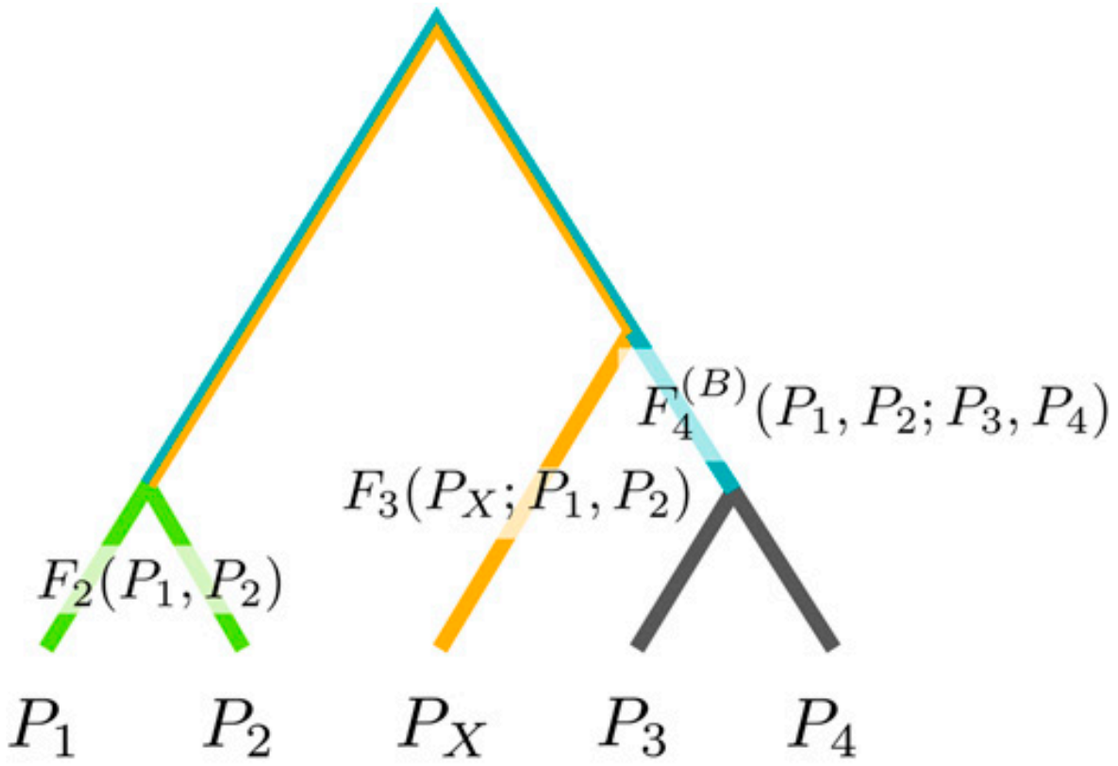
Benjamin M. Peter[1]

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

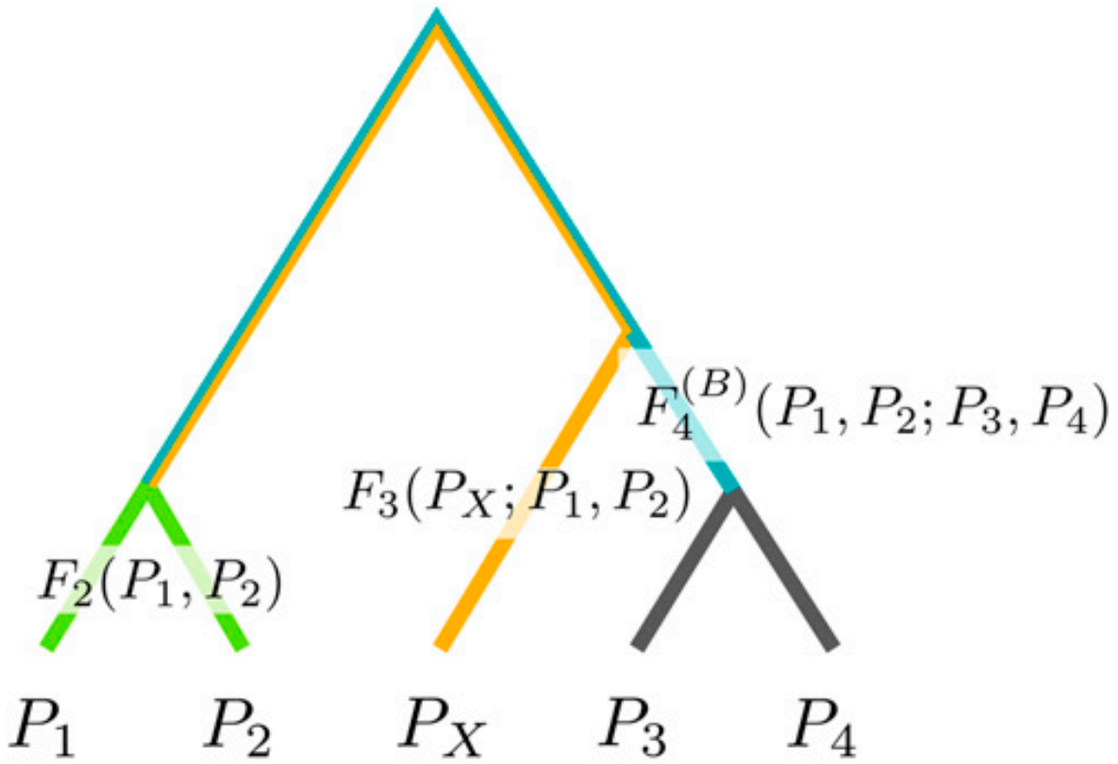
F-statistics - Definitions



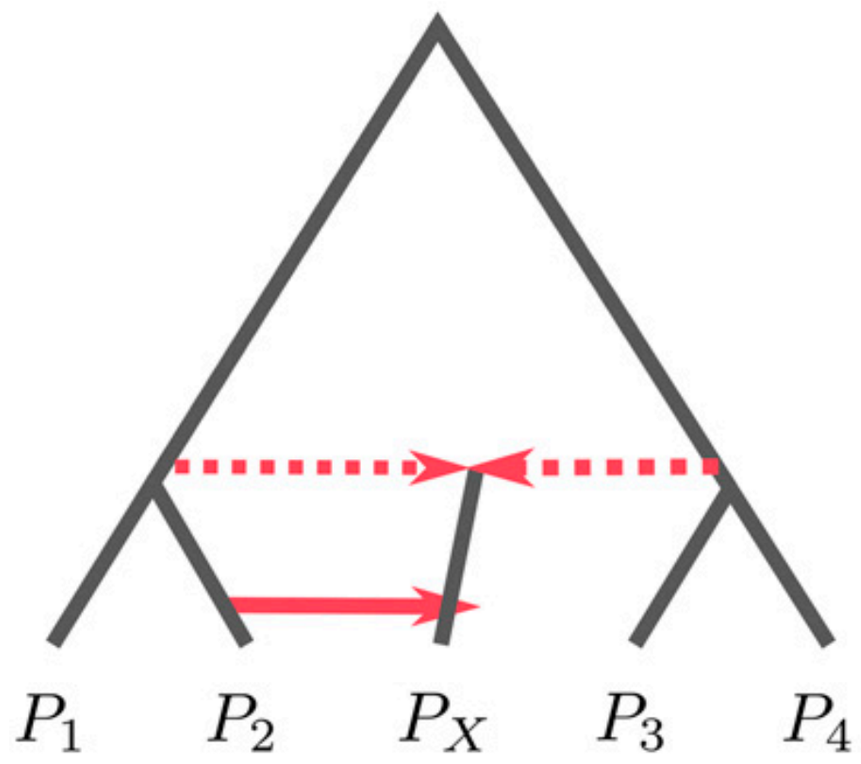
Population phylogeny

Peter (2014) Genetics

F-statistics - Definitions



Population phylogeny

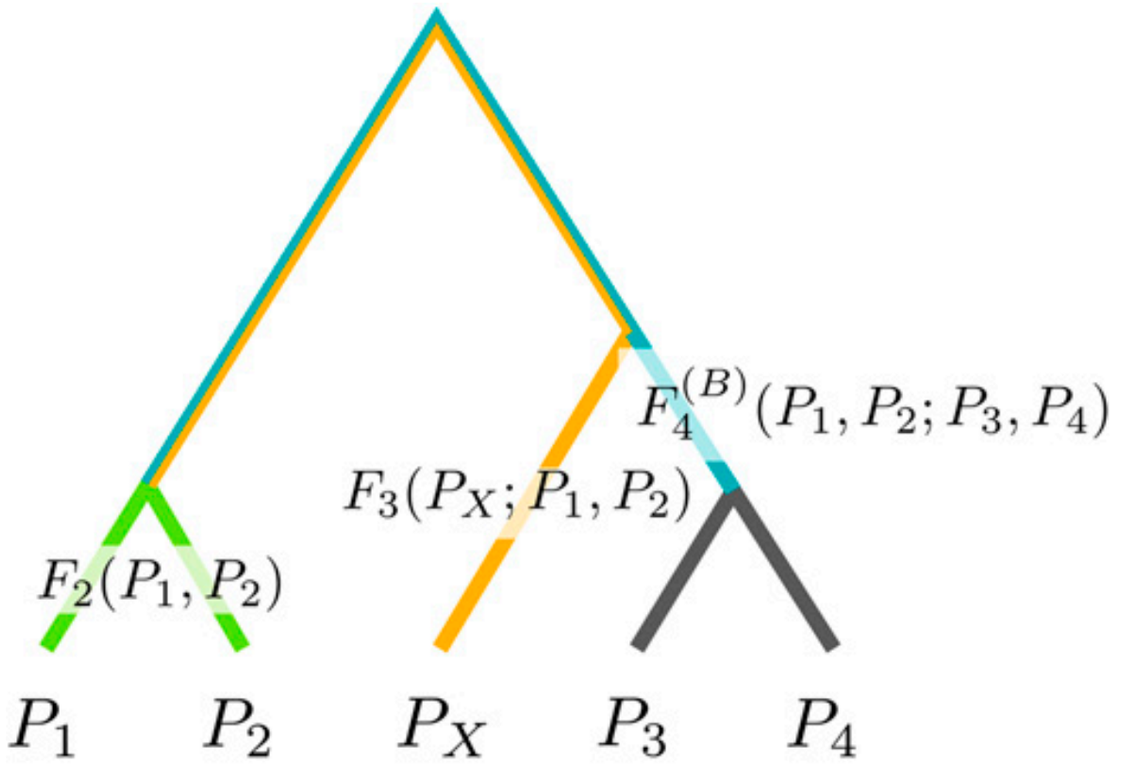


Admixture graph

Peter (2014) Genetics

F-statistics - Definitions

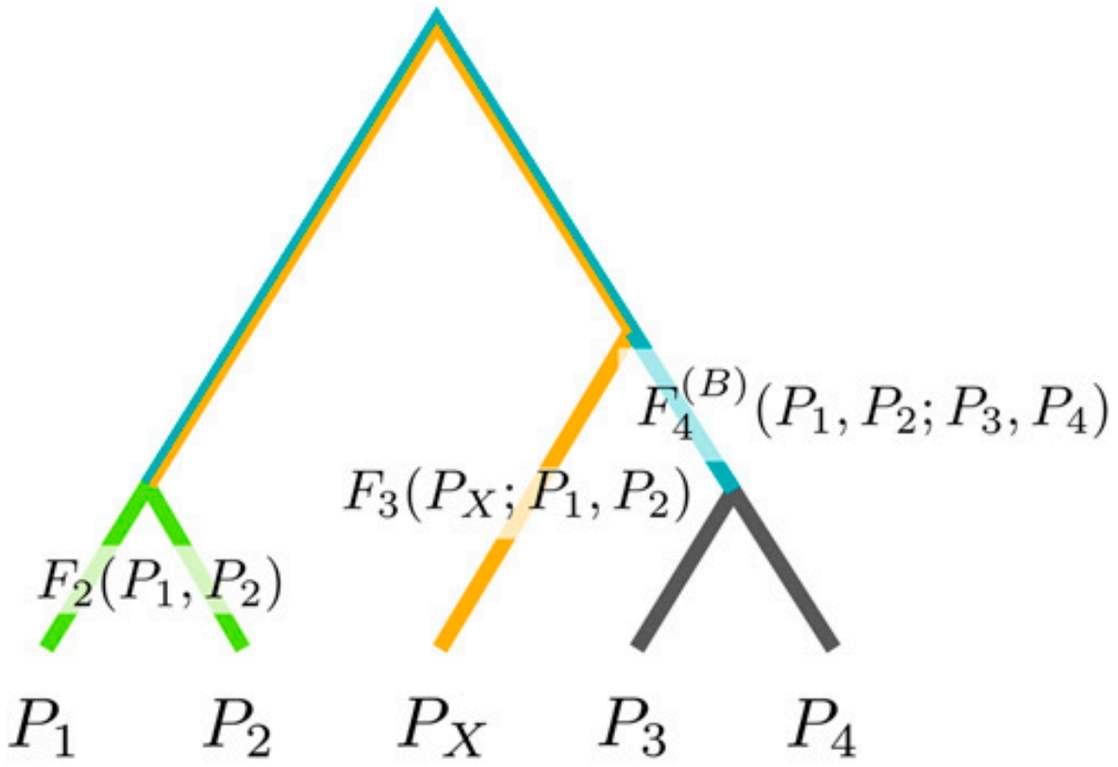
$$= \frac{F_2(P_1, P_2)}{\mathbb{E}[(p_1 - p_2)^2]}$$



Population phylogeny

Peter (2014) Genetics

F-statistics - Definitions



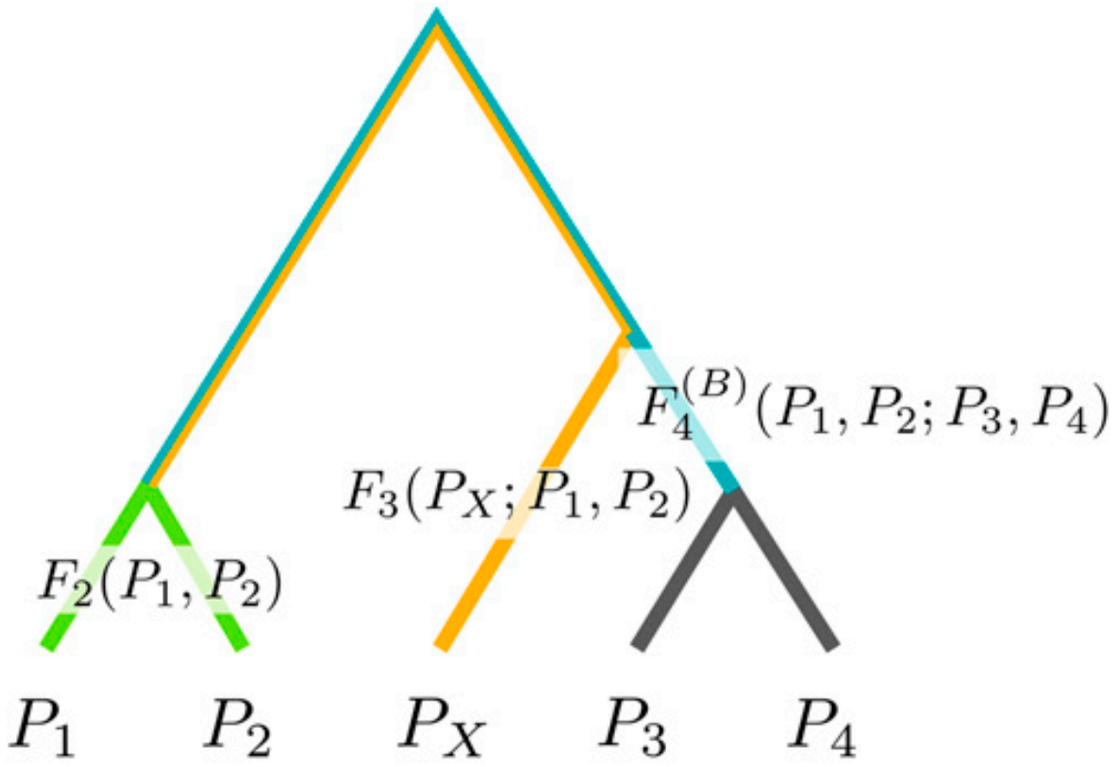
Population phylogeny

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

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

Peter (2014) Genetics

F-statistics - Definitions



Population phylogeny

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

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

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

Interpretation in terms of branch lengths on a population phylogeny

Peter (2014) Genetics

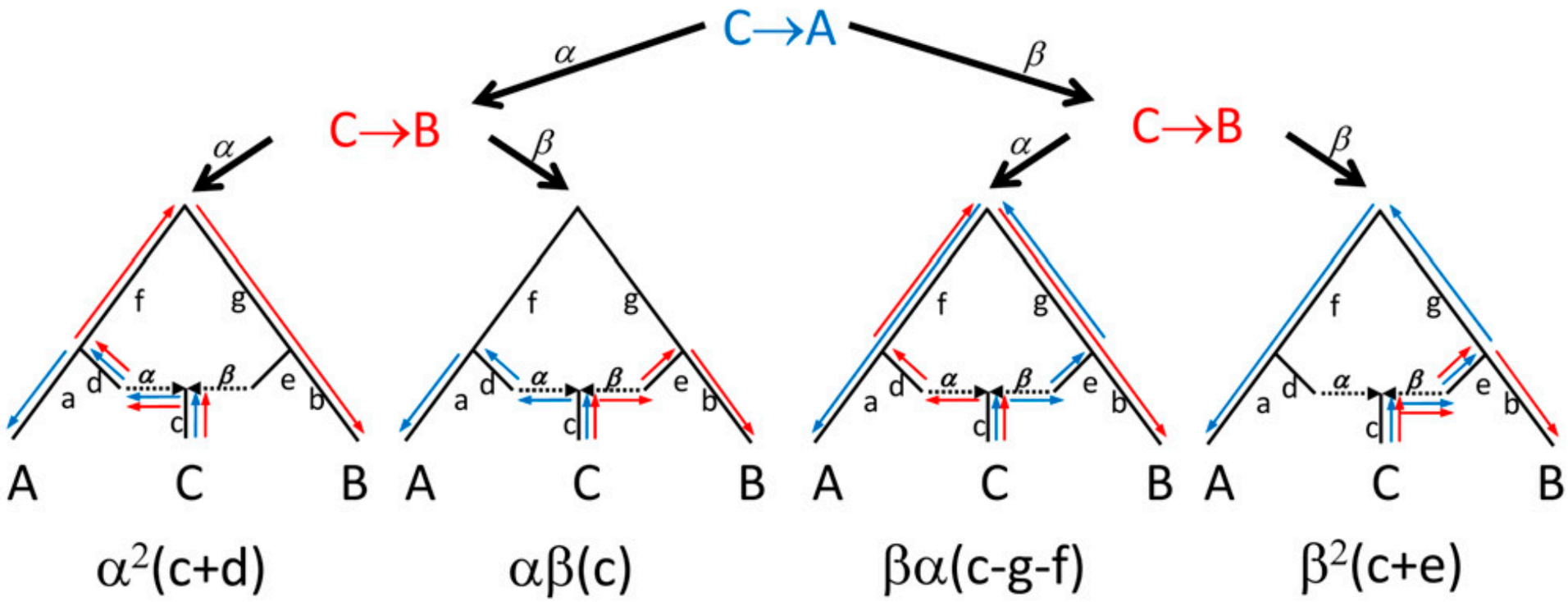
F-statistics - Applications

F-statistic	Application	Test	Interpretation
f_2(A,B)	Branch length		
f_3(X;A,B)	Admixture f_3 - test	$f_3 < 0$	X is admixed related to A,B
	Outgroup - f_3		If X is outgroup to (A,B), f_3 proportional to shared drift between X and divergence of (A,B)
D(A,B;C,D)	D - test	D = 0	(A,B) form a clade with respect to (C,D)
D(O,B;C,D)	Symmetry test	D = 0	If O is outgroup to (B,C,D), tests for symmetry of B with respect to (C,D)
f_4(A,B;C,D)	f_4 - ratio test	$\alpha > 0$	Admixture proportion > 0
	Number of distinct ancestry streams between sets of outgroup and target populations (*qpWave*)		If rank of f4 - matrix is m, target populations are carry at least m + 1 streams of ancestry differentially related to the outgroup set
	Phylogeny-free estimation of admixture proportions (*qpAdm*)		Admixture proportions and fit for a target population as a mixture of N source populations
	Admixture graph fitting (*qpGraph*)		Goodness of fit of f-statistics predicted for specific graph topology

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

F-statistics - Expectations and usage

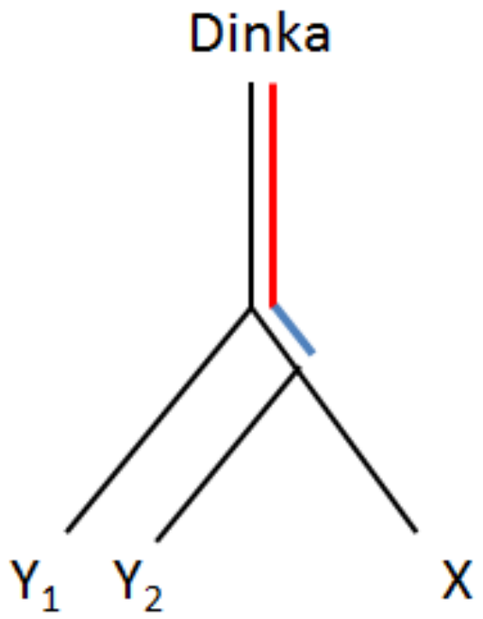
$$F_3(C;A,B) = c + \alpha^2 d + \beta^2 e - \alpha\beta(g+f)$$

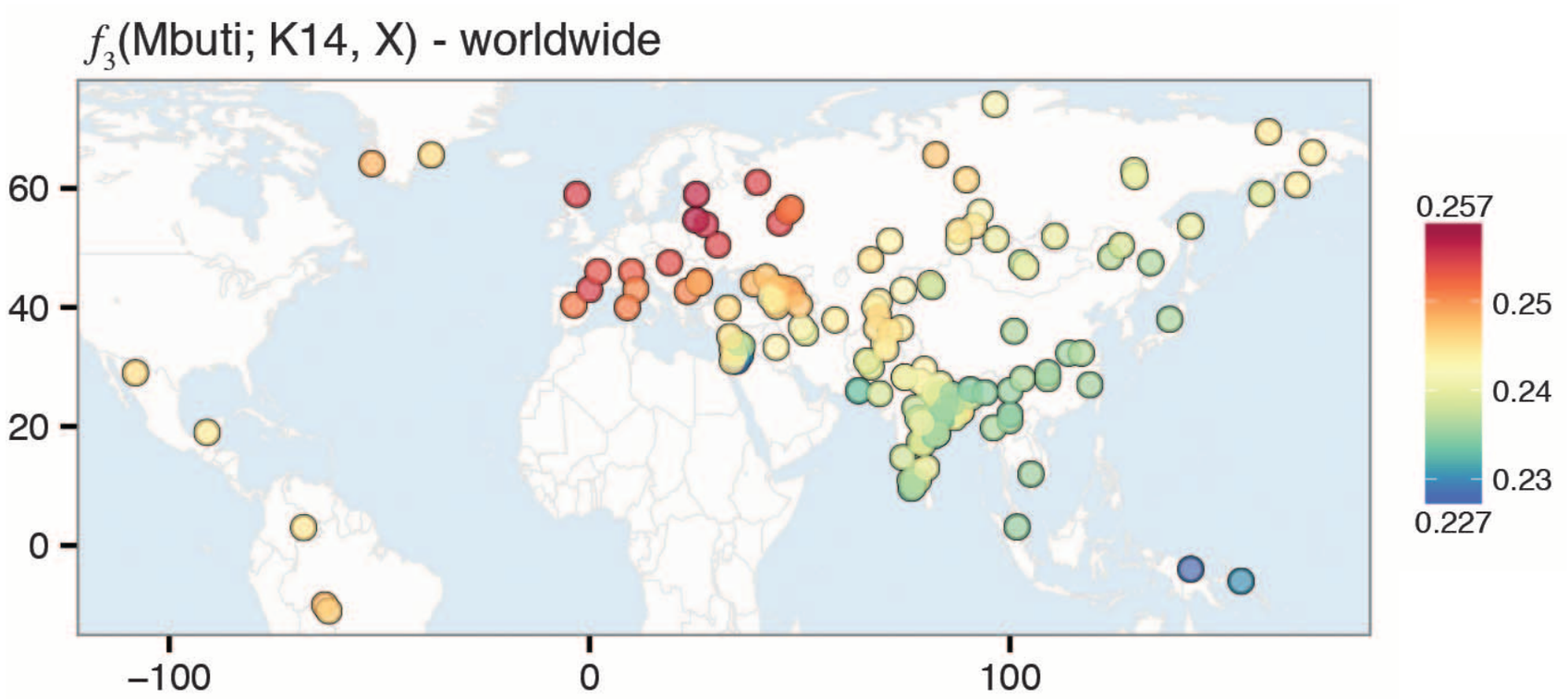


Admixture f_3 tests whether a target population C shows evidence for admixture from two source populations related (possibly distant) to sample populations A and B

Patterson et al. (2012) Genetics

F-statistics - Expectations and usage

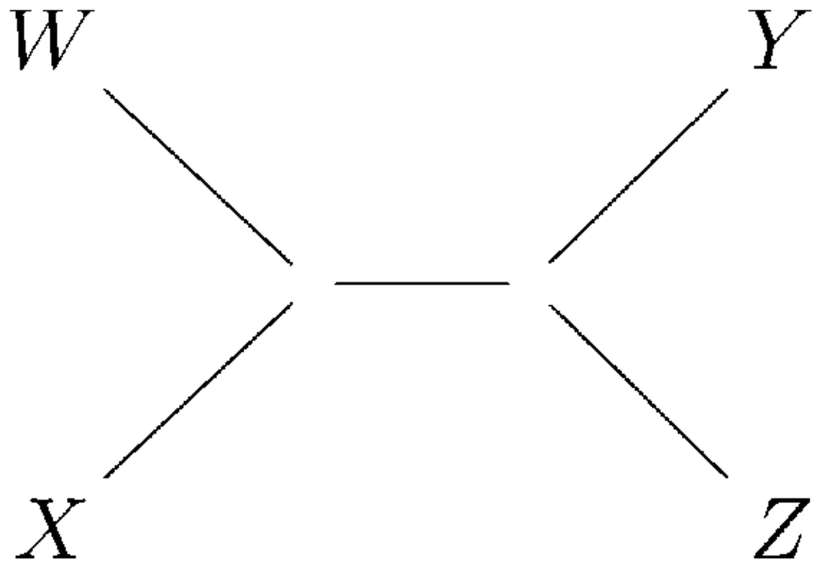




Outgroup - f_3 statistics measure the shared drift between an outgroup and the divergence of two test populations

Haak (2015) Nature; Seguin-Orlando et al. (2014) Science

F-statistics - Expectations and usage



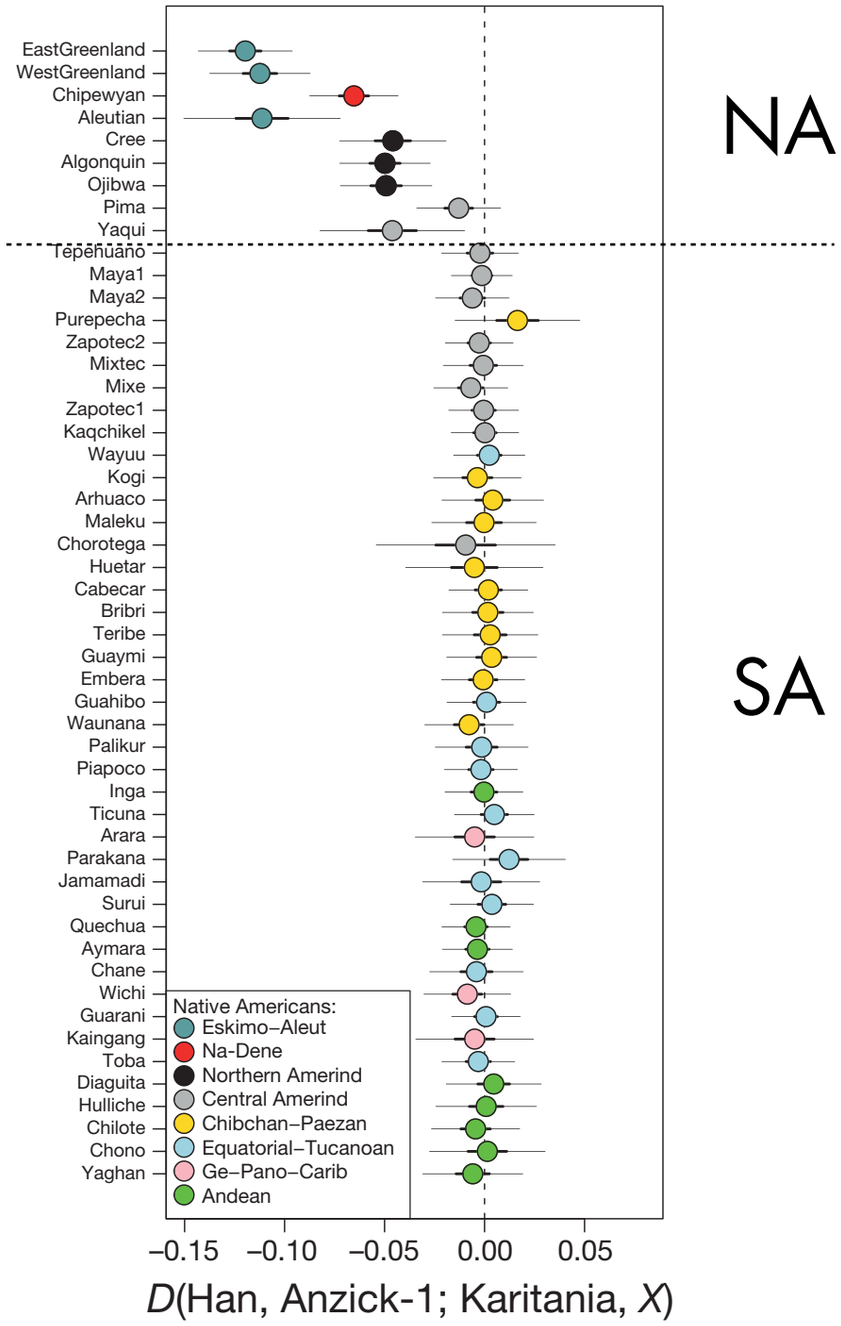
$$\hat{Num}_i = (w - x)(y - z)$$
$$\hat{Den}_i = (w + x - 2wx)(y + z - 2yz)$$

$$D = \frac{\sum_i \hat{Num}_i}{\sum_i \hat{Den}_i},$$

D - test is used to test whether four populations are related through a simple tree

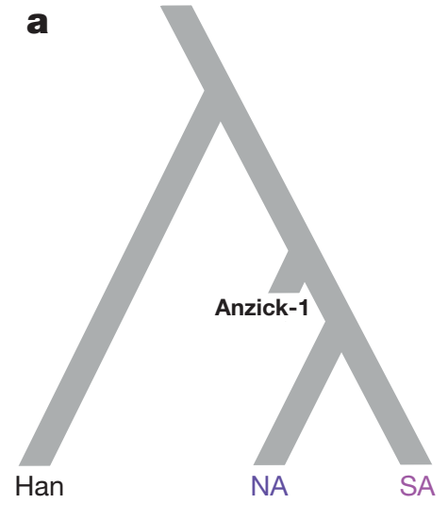
Patterson (2011) Genetics

F-statistics - Expectations and usage

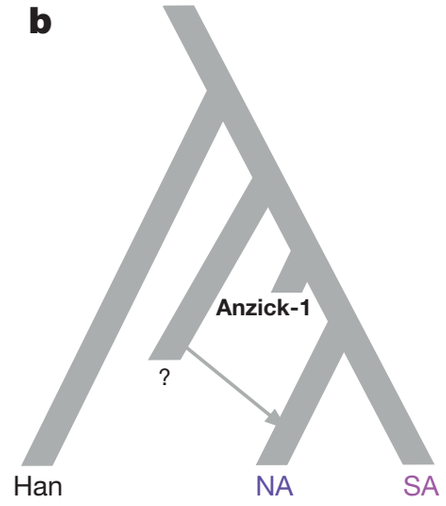


NA

SA

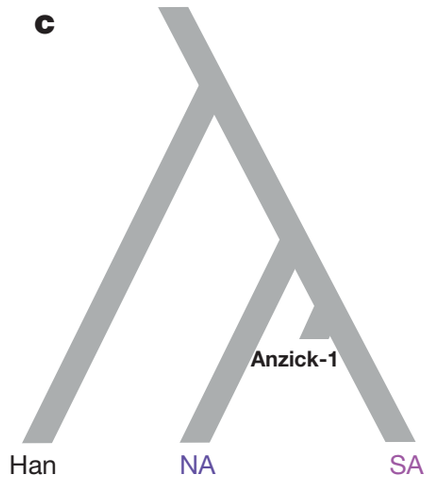


Rejected

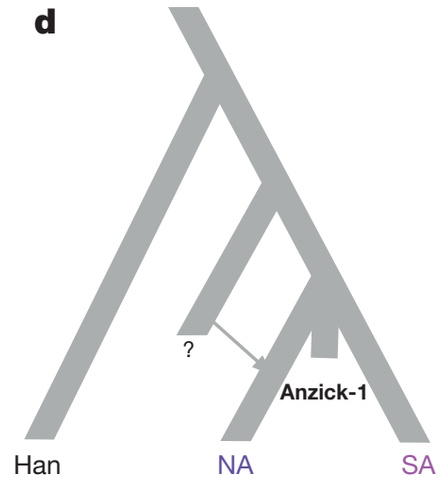


Not supported by the data

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



Consistent with the data



Consistent with the data

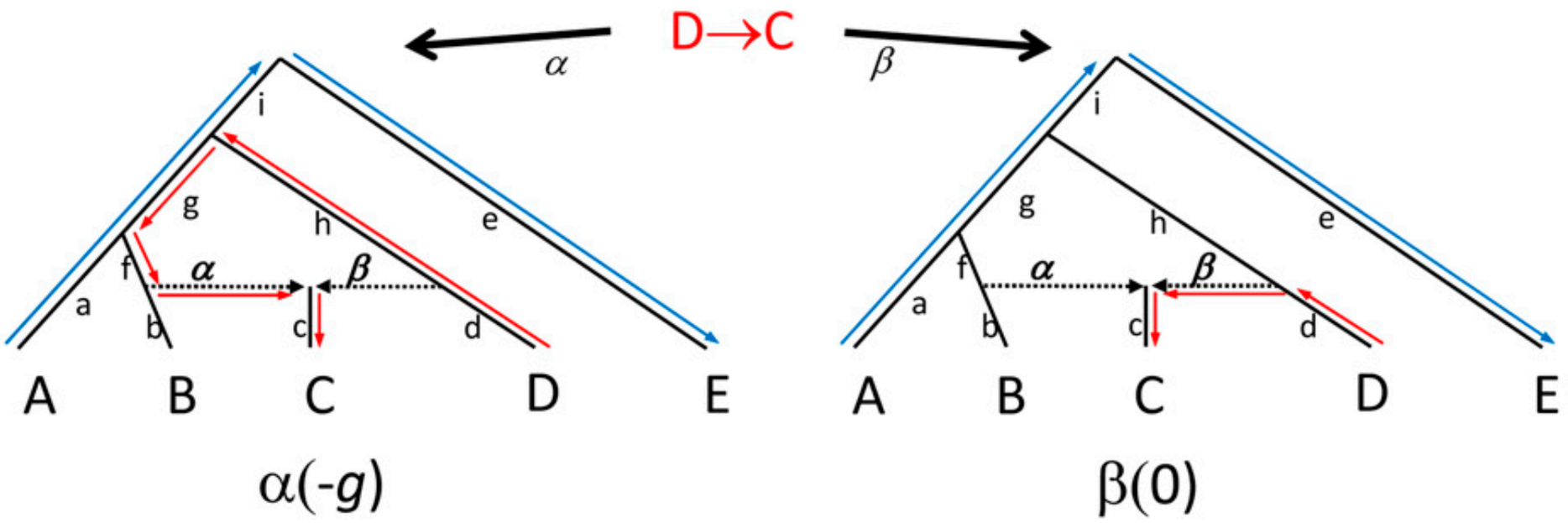
Failed treeness test can be interpreted in multiple ways!

Rasmussen (2014) Nature

F-statistics - Expectations and usage

$F_4(A,E;D,C) = -\alpha g$

$$F_4\ ratio = \frac{F_4(A,E;D,C)}{F_4(A,E;D,B)} = \frac{-\alpha g}{-g} = \alpha$$

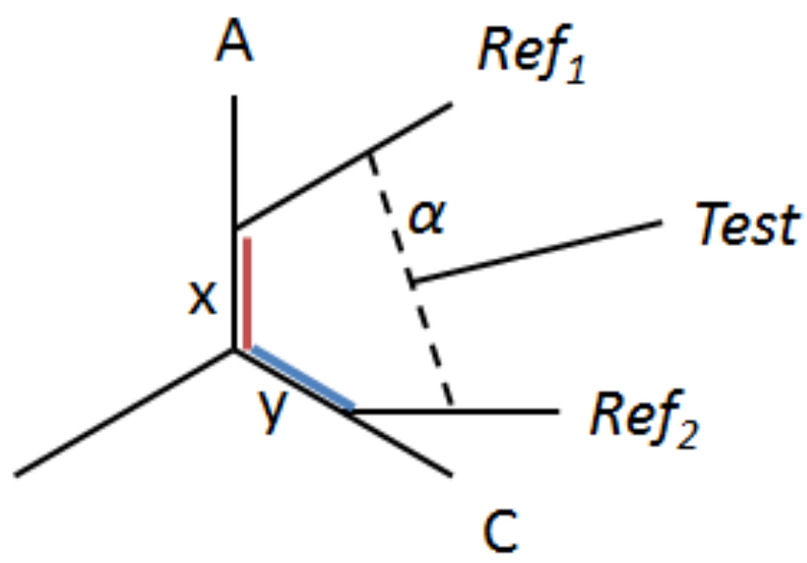


f_4 - ratio estimates admixture proportions under the assumption of a specific population phylogeny

Patterson et al. (2012) Genetics

Phylogeny-free admixture fitting (qpAdm)

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



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

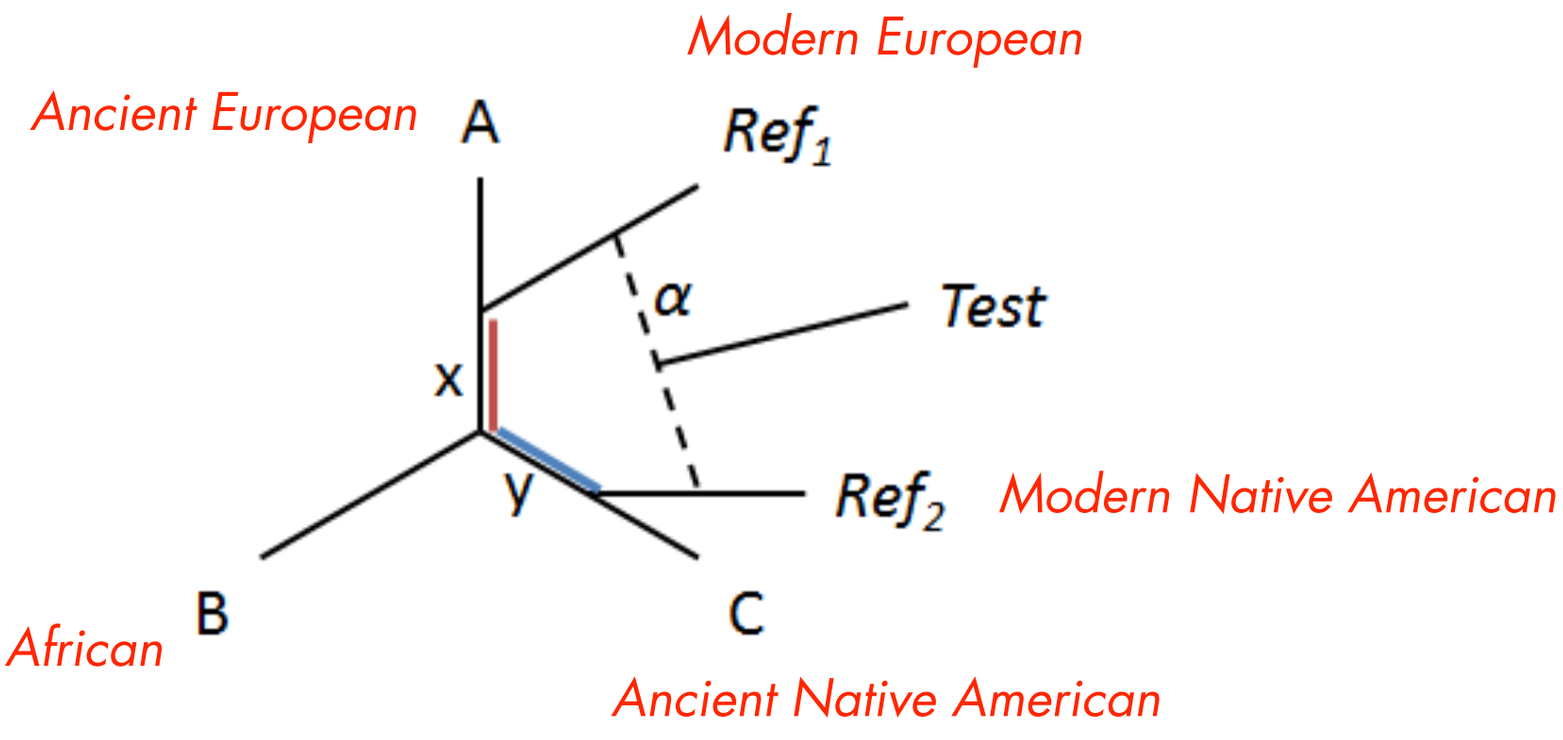
$f_4(Test, A; B, C) = -(1-\alpha)y = (1-\alpha) f_4(Ref_2, A; B, C) + \alpha f_4(Ref_1, A; B, C)$
$f_4(Test, B; A, C) = \alpha x - (1-\alpha)y = \alpha f_4(Ref_1, A; B, C) + (1-\alpha) f_4(Ref_2, A; B, C)$
$f_4(Test, C; A, B) = \alpha x = \alpha f_4(Ref_1, A; B, C) + (1-\alpha) f_4(Ref_2, A; B, C)$

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

Haak et al (2015) Nature

Phylogeny-free admixture fitting (qpAdm)

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



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

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

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

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

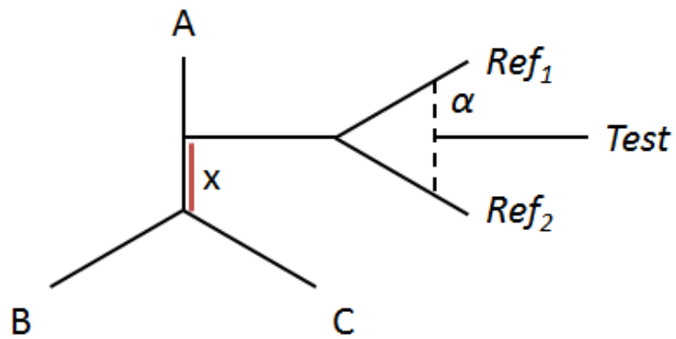
$f_4(Test, B; A, C) = \alpha x - (1-\alpha)y = \alpha f_4(Ref_1, A; B, C) + (1-\alpha) f_4(Ref_2, A; B, C)$

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

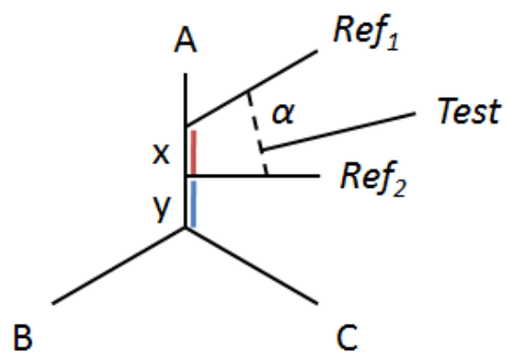
Differential drift sharing with sets of outgroup populations is leveraged to infer mixture proportions

Haak et al (2015) Nature

Phylogeny-free admixture fitting (qpAdm)

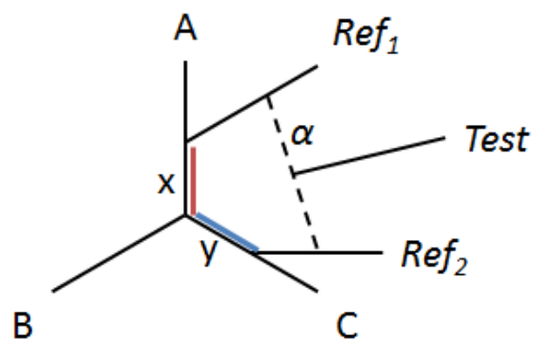


$f_4(Test \text{ or } Ref_1 \text{ or } Ref_2, A; B, C) = 0$
$f_4(Test \text{ or } Ref_1 \text{ or } Ref_2, B; A, C) = x$
$f_4(Test \text{ or } Ref_1 \text{ or } Ref_2, C; A, B) = x$



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

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



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

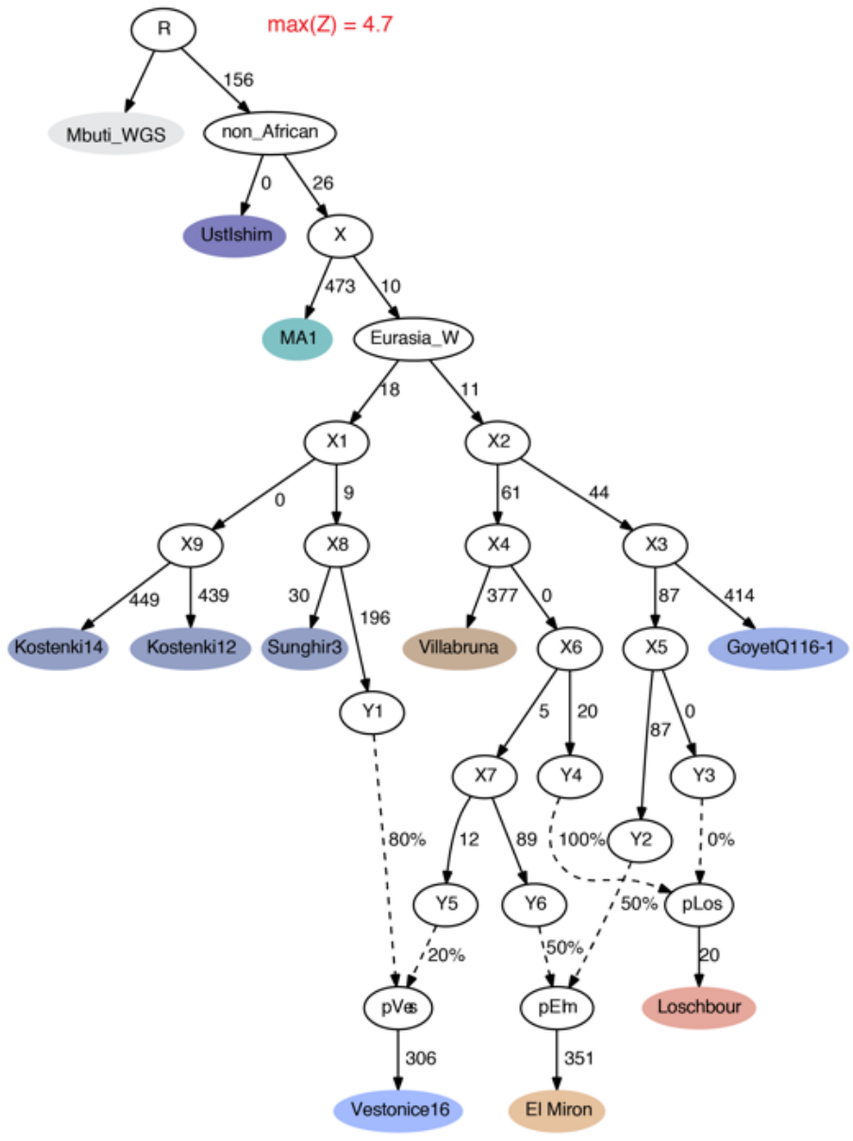
$f_4(Test, A; B, C) = -(1-\alpha)y = (1-\alpha) f_4(Ref_2, A; B, C)+\alpha f_4(Ref_1, A; B, C)$
$f_4(Test, B; A, C) = \alpha x-(1-\alpha)y = \alpha f_4(Ref_1, A; B, C)+(1-\alpha) f_4(Ref_2, A; B, C)$
$f_4(Test, C; A, B) = \alpha x = \alpha f_4(Ref_1, A; B, C)+(1-\alpha) f_4(Ref_2, A; B, C)$

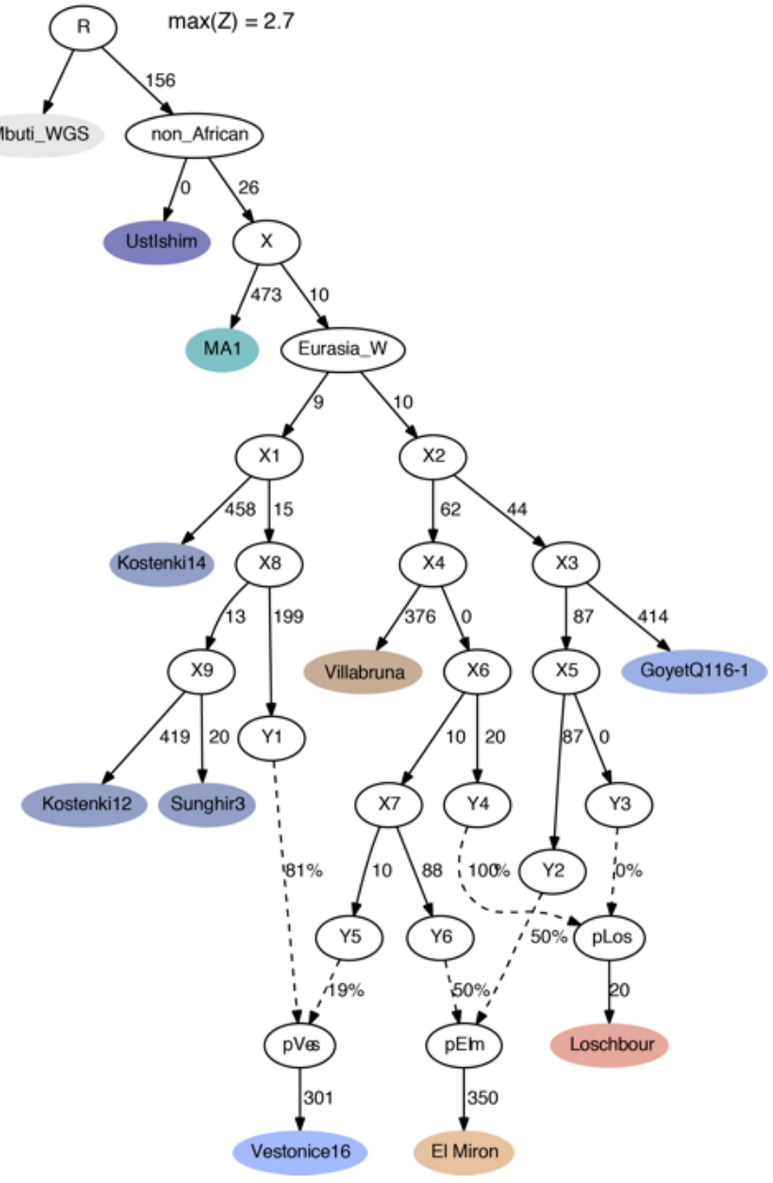
$$f_4(Test; A; B, C) \approx \sum_{i=1}^{N} \alpha_i f_4(Ref_i; A; B, C)$$

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

Haak et al (2015) Nature

Admixture graph fitting

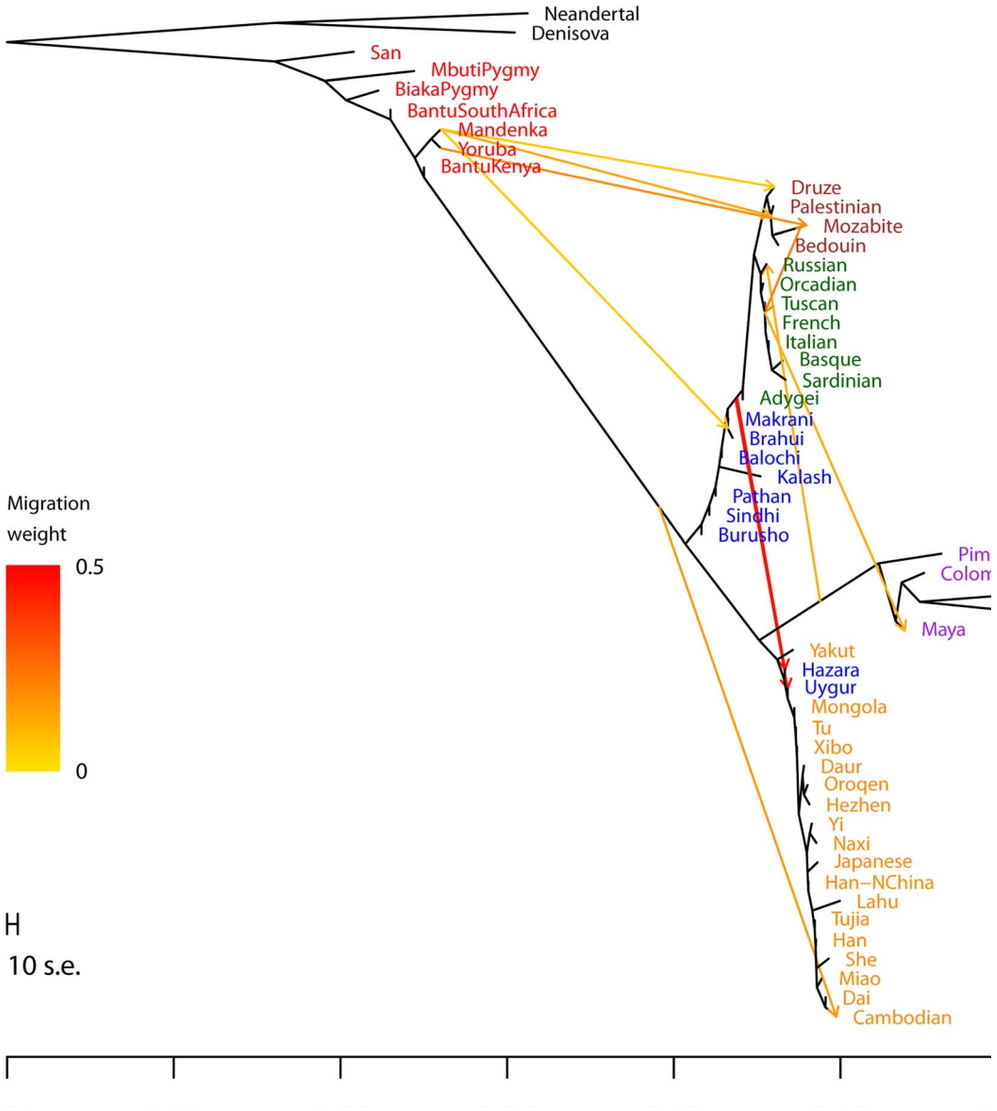




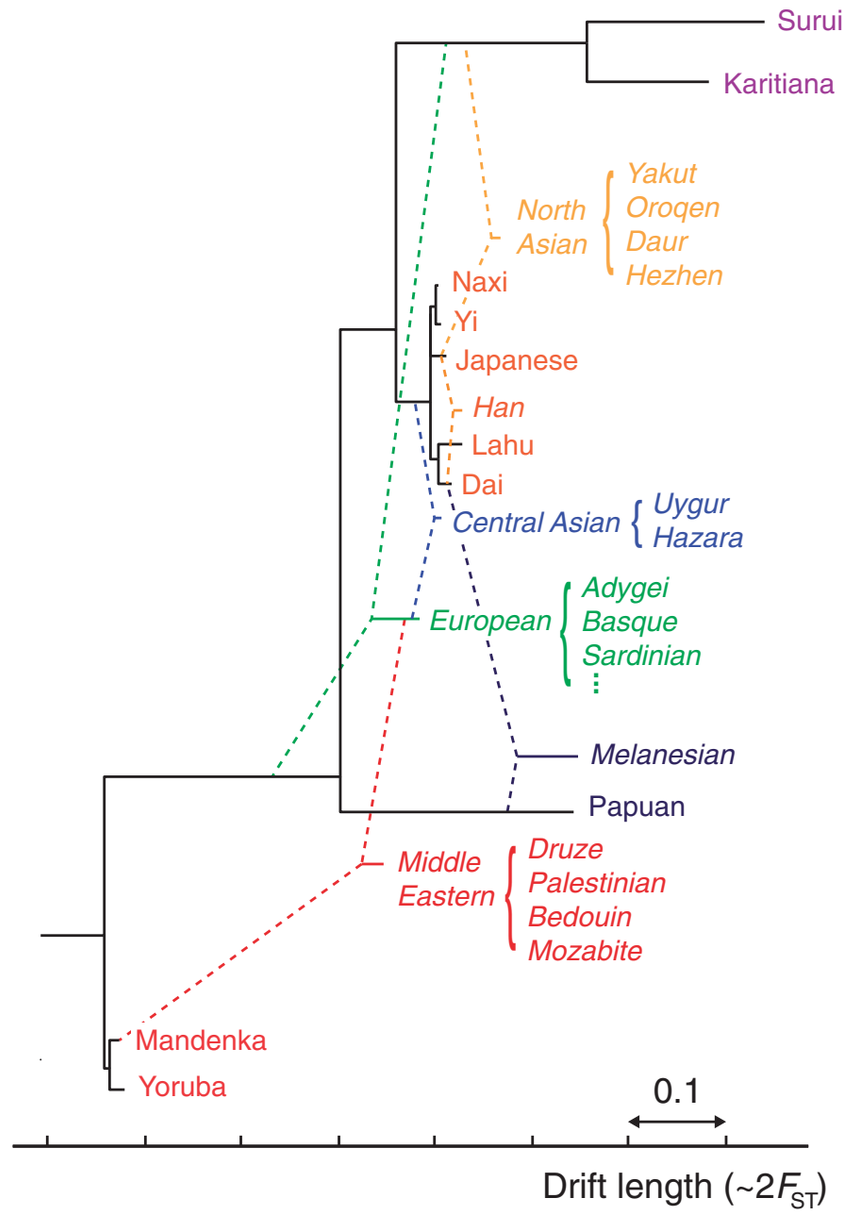
Given an admixture graph, calculate expected drift and admixture terms, and compare fit with observed values

Sikora et al. (2017) Science

Admixture graph fitting - related methods



Treemix



Mixmapper

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

Take - home messages

- F-statistics are a versatile framework that have become a standard tool for testing hypotheses about population history
- Straightforward and efficient to estimate, but applicable to quite complex scenarios
- Suitable for lower quality data (e.g. aDNA), although 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
  - qpAdm /qpWave
  - qpGraph

F-statistics overview

F-statistic	Application	Test	Interpretation
f_2(A,B)	Branch length		
f_3(X;A,B)	Admixture f_3 - test	$f_3 < 0$	X is admixed related to A,B
	Outgroup - f_3		If X is outgroup to (A,B), f_3 proportional to shared drift between X and divergence of (A,B)
D(A,B;C,D)	D - test	D = 0	(A,B) form a clade with respect to (C,D)
D(O,B;C,D)	Symmetry test	D = 0	If O is outgroup to (B,C,D), tests for symmetry of B with respect to (C,D)
f_4(A,B;C,D)	f_4 - ratio test	$\alpha > 0$	Admixture proportion > 0
	Number of distinct ancestry streams between sets of outgroup and target populations (*qpWave*)		If rank of f4 - matrix is m, target populations are carry at least m + 1 streams of ancestry differentially related to the outgroup set
	Phylogeny-free estimation of admixture proportions (*qpAdm*)		Admixture proportions and fit for a target population as a mixture of N source populations
	Admixture graph fitting (*qpGraph*)		Goodness of fit of f-statistics predicted for specific graph topology

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