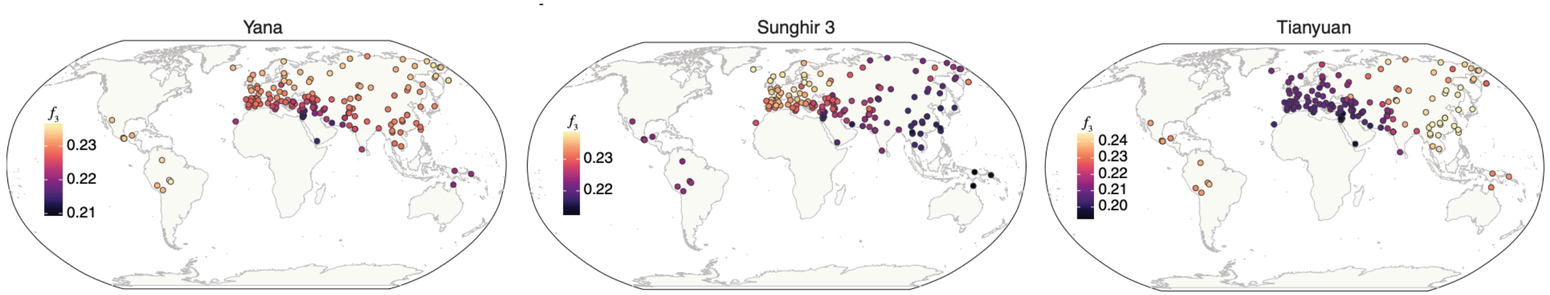
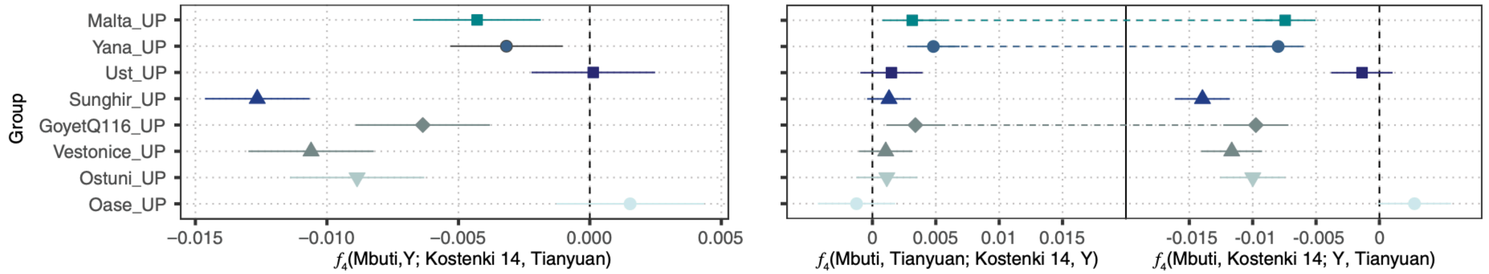
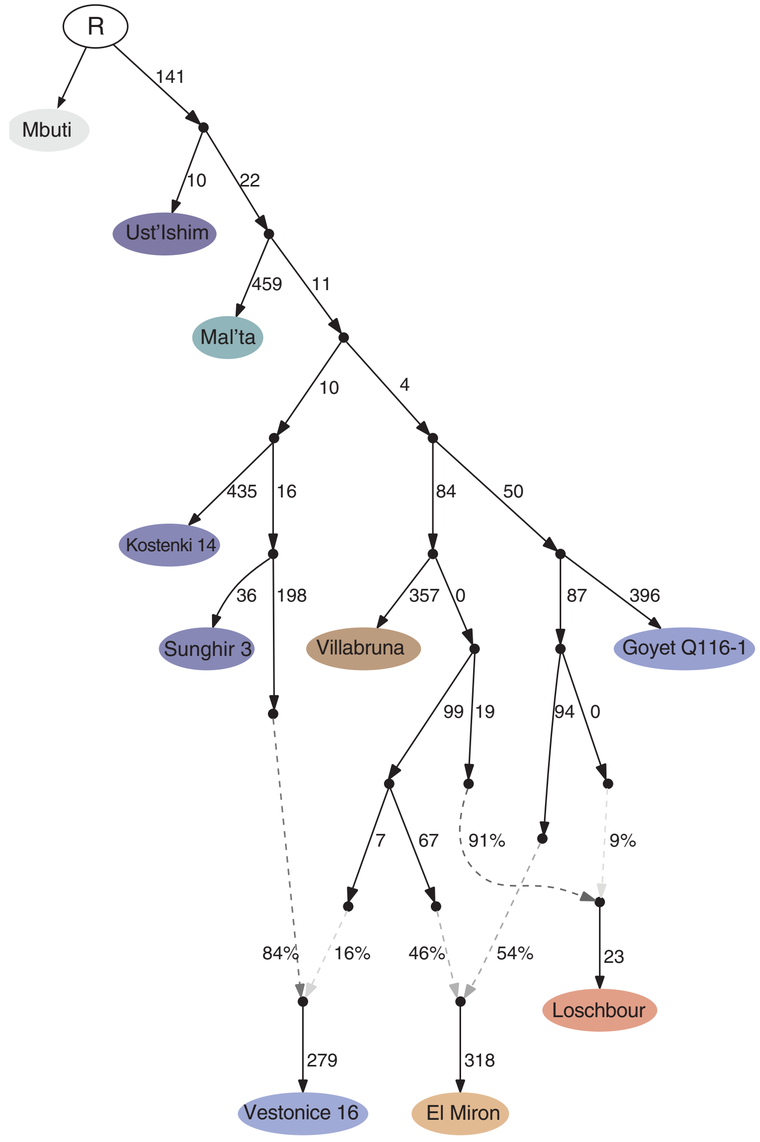
# D/f statistics and ancient gene flow

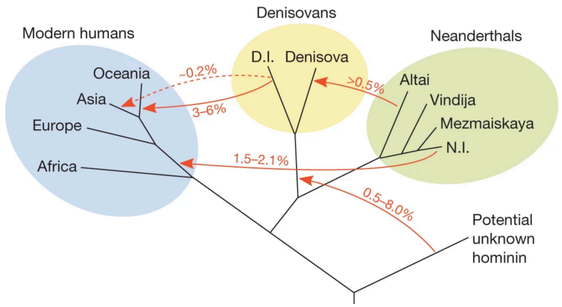
Summer course in analysis of high throughput data for population genetics 2024
August 21st 2024

Martin Sikora
Globe Institute
University of Copenhagen
martin.sikora@sund.ku.dk



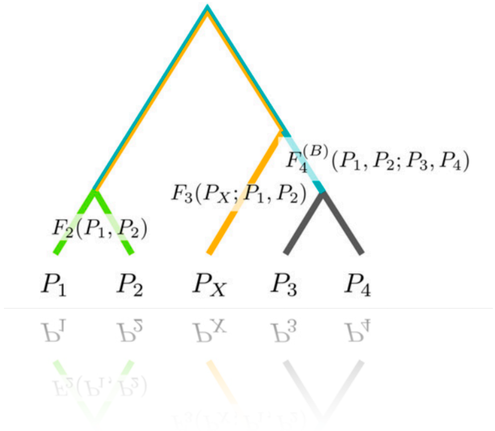





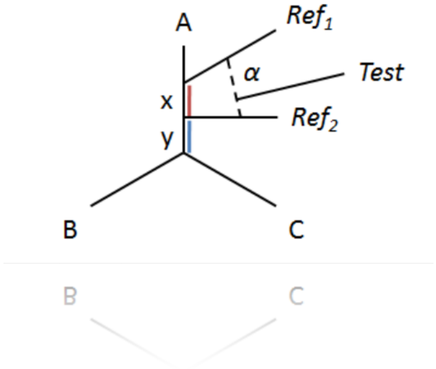

## Background

What are f-statistics and why do we use them?



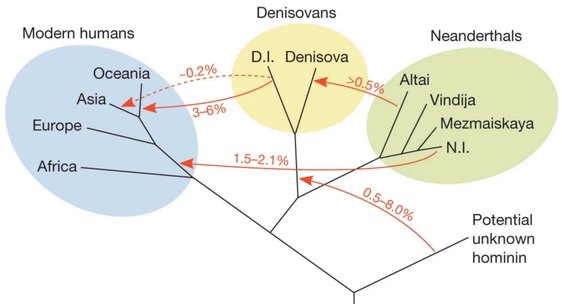

## Definitions

$F_2$, $F_3$ and $F_4$ statistics




## Applications

Testing hypotheses about population admixture




# Background

What are f-statistics and why do we use them?

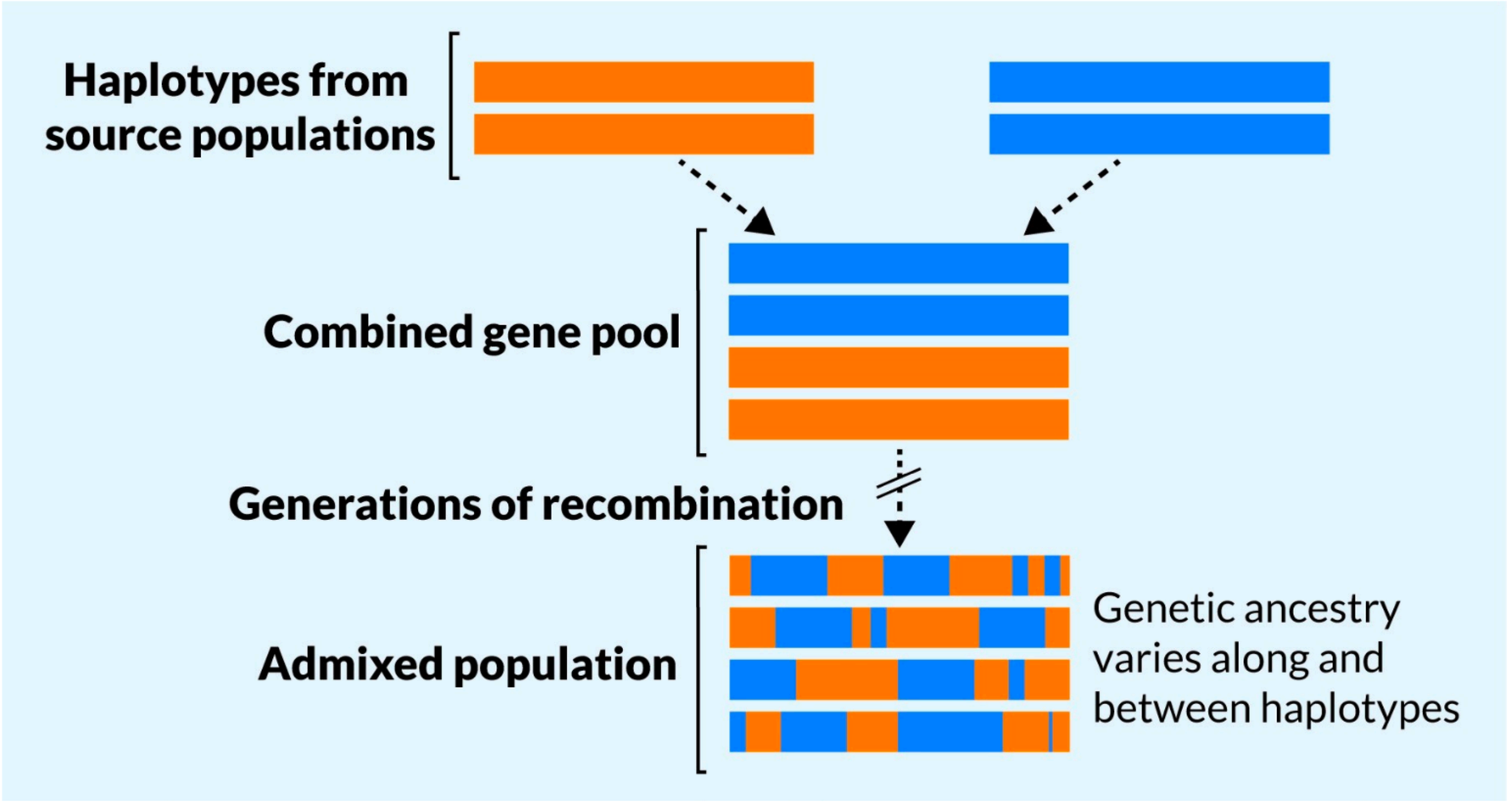
## Definitions

$F_2$, $F_3$ and $F_4$ statistics

## Applications

Testing hypotheses about population admixture

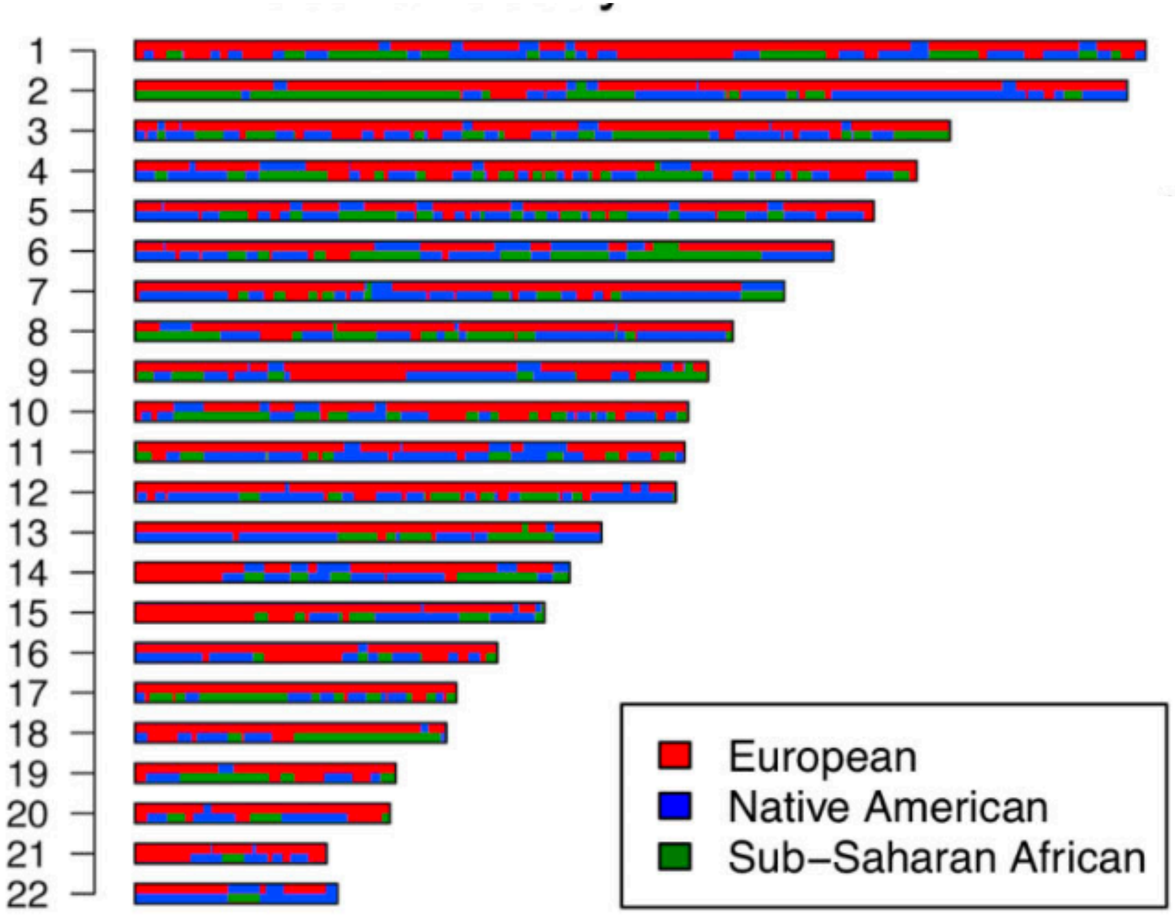
# Genomic signature of admixture




Admixture between two previously isolated populations

Korunes and Goldberg (2021) PLoS Genet

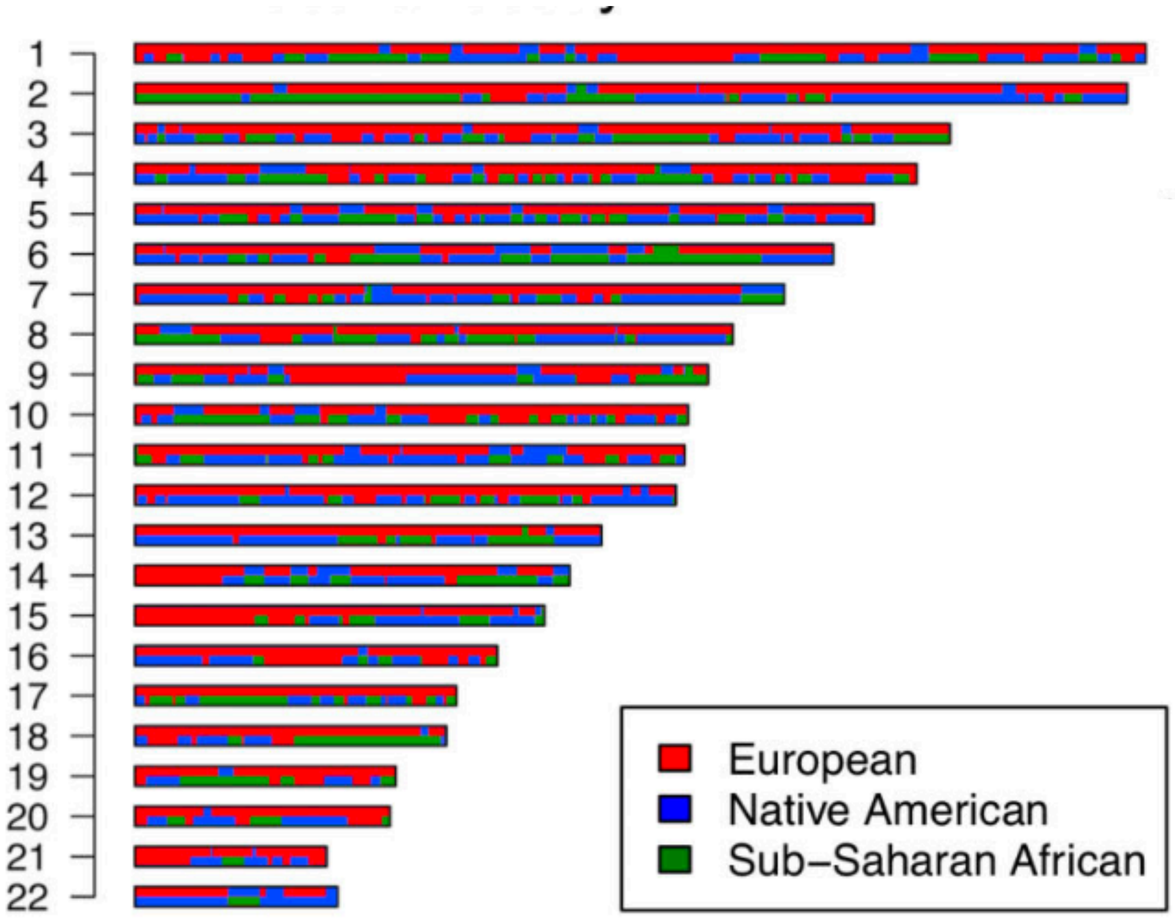
# Modelling admixture with local ancestry tracts

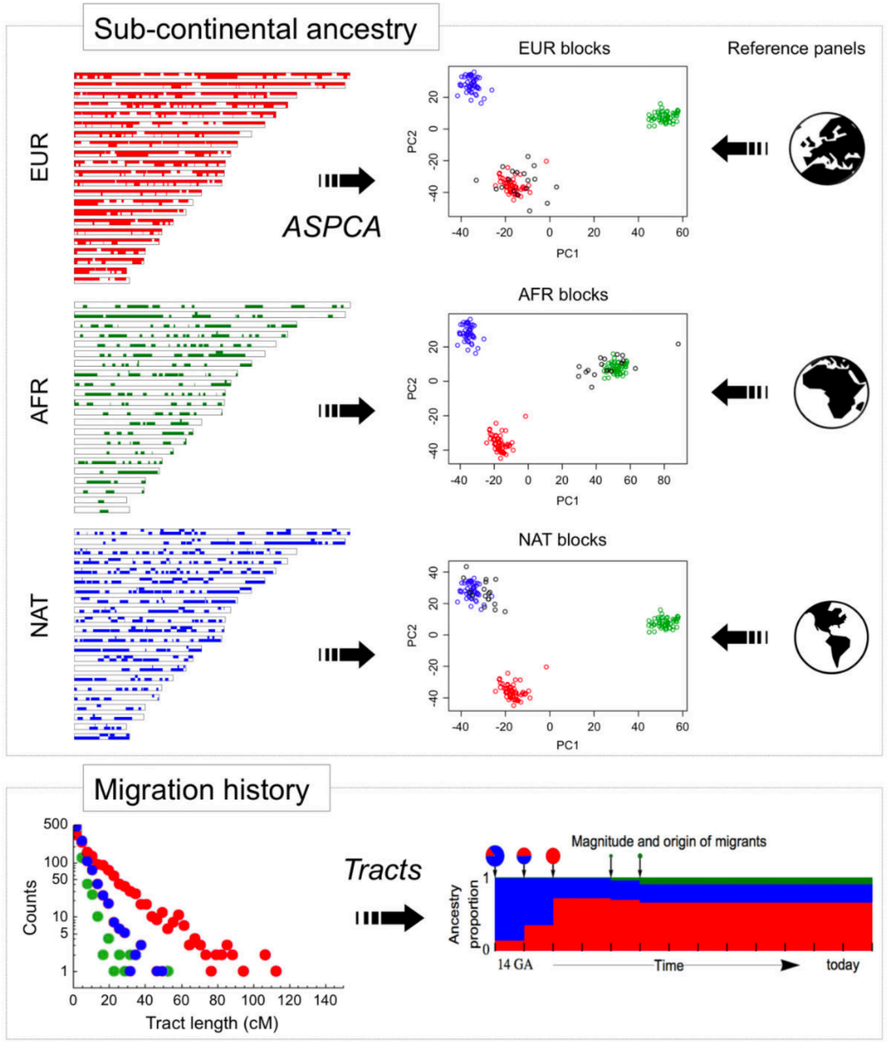



Local ancestry tracts in a Caribbean individual

Moreno-Estrada et al (2013) PLoS Genet

# Modelling admixture with local ancestry tracts

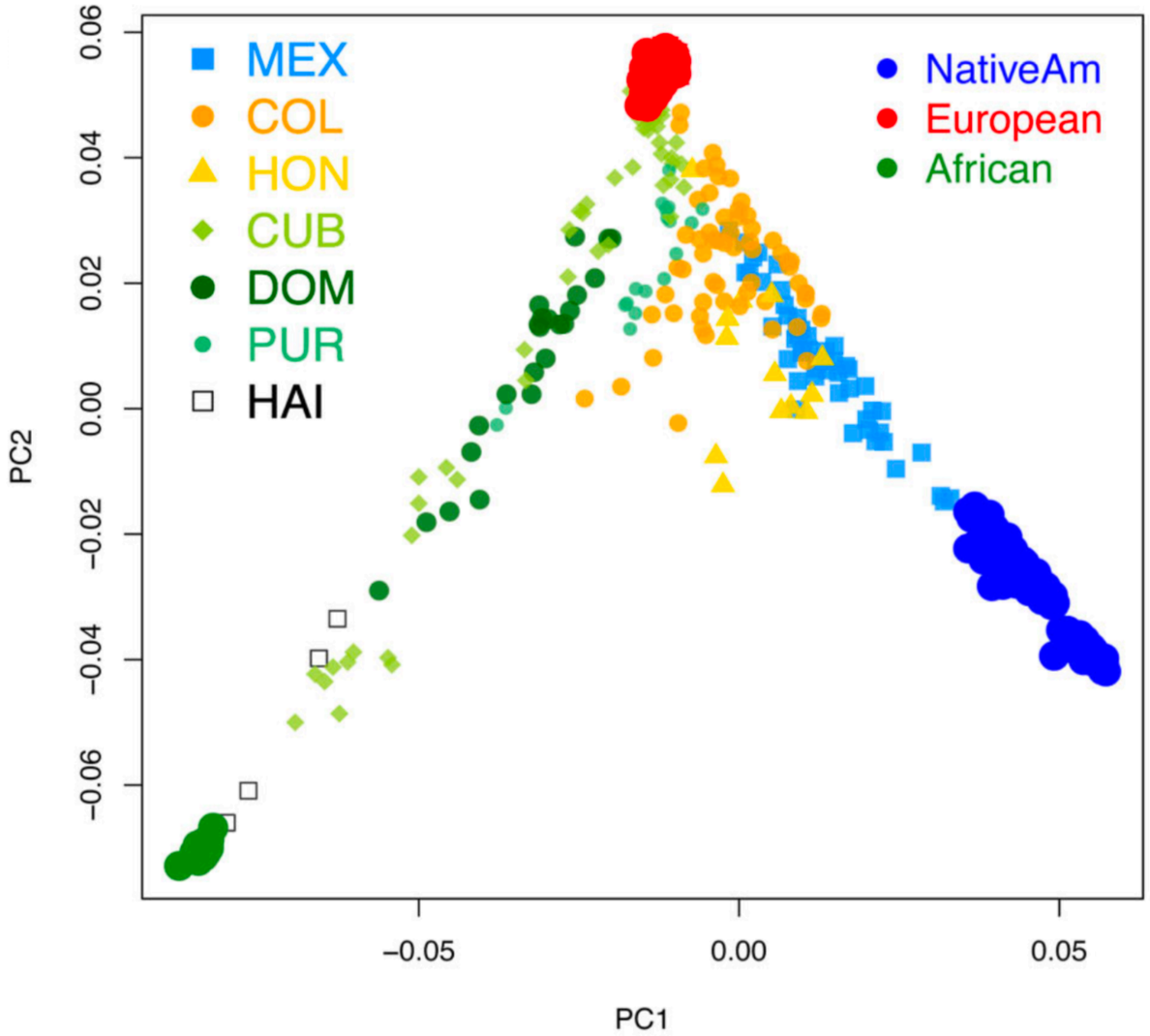



Fitting complex models of admixture

Moreno-Estrada et al (2013) PLoS Genet

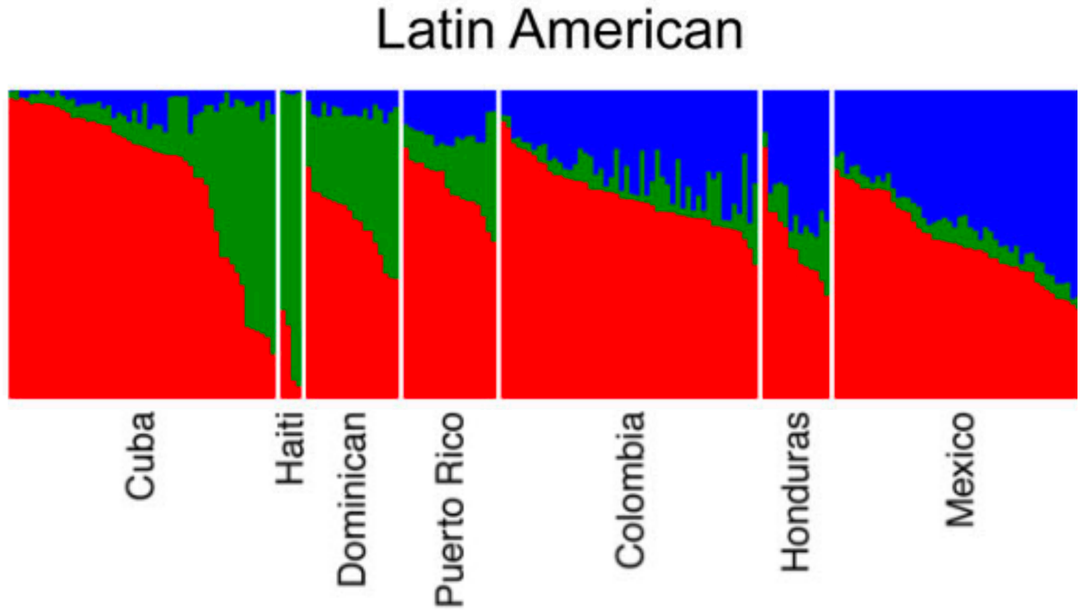
# Signatures of admixture in PCA


Recently admixed populations on a cline between source populations in PCA space

Moreno-Estrada et al (2013) PLoS Genet

# Model-based clustering

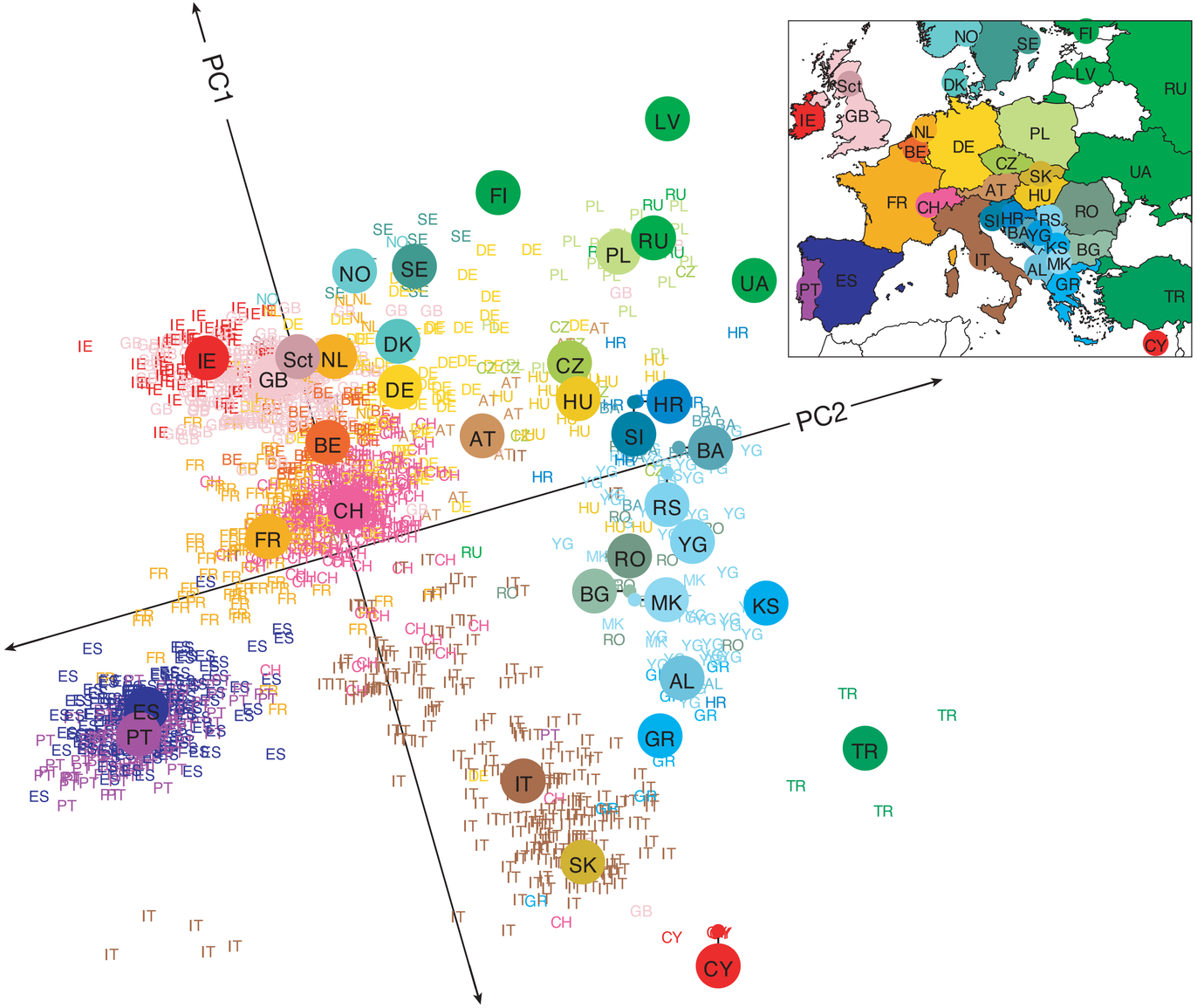





Admixture proportions can be accurately inferred when admixture is recent and between well differentiated sources

Moreno-Estrada et al (2013) PLoS Genet

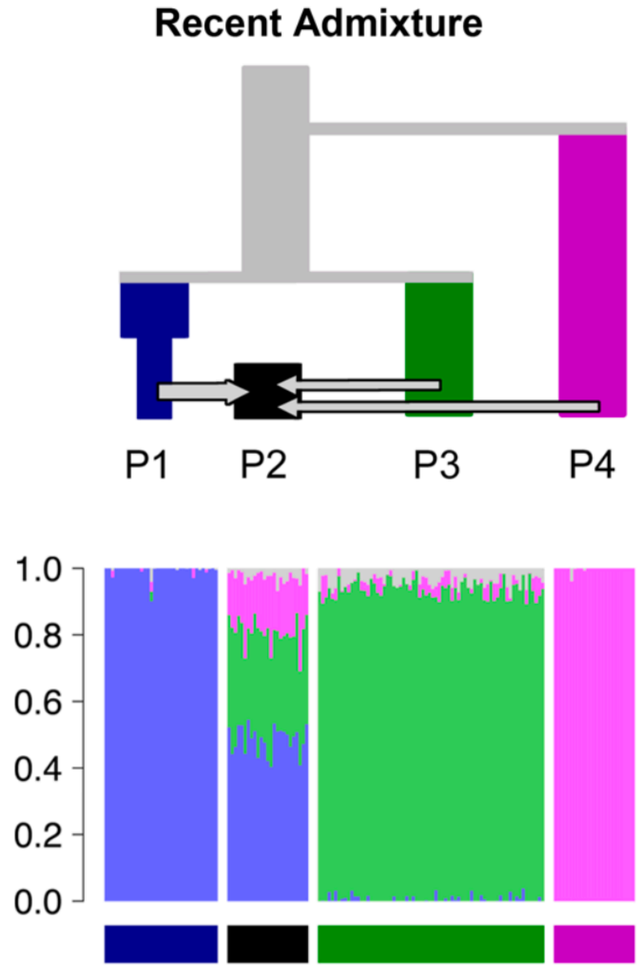
# Challenges in interpreting PCA clines




Admixture or isolation-by-distance?

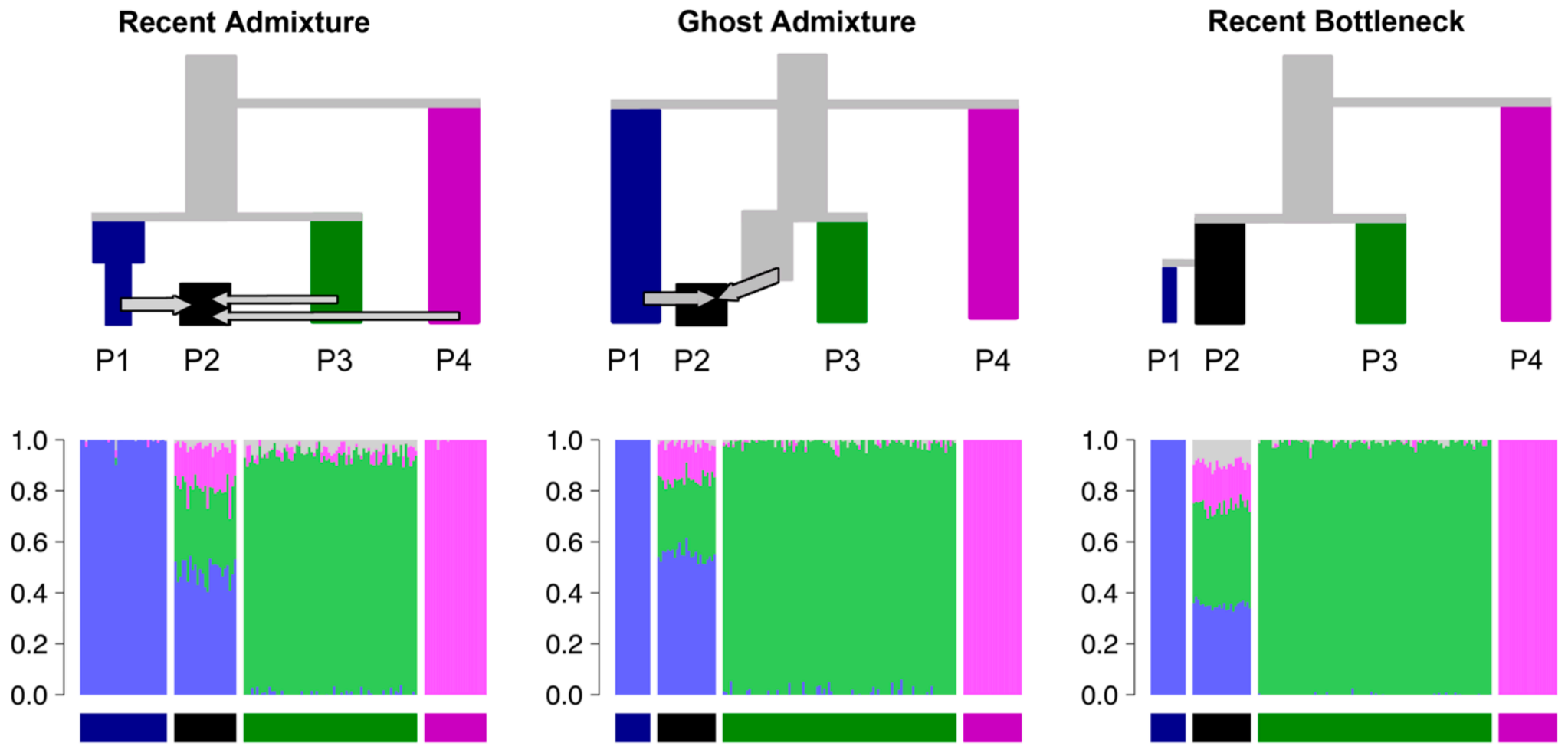
Novembre et al (2008) Nature

# Challenges in interpreting model-based clustering




Lawson et al (2018) Nat Commun

# Challenges in interpreting model-based clustering




Different demographic scenarios can produce indistinguishable results in model-based clustering

Lawson et al (2018) Nat Commun

*Am. J. Hum. Genet. 63:1237–1240, 1998*

## On the Probability of Neanderthal Ancestry

*To the Editor:*

The controversial relationship between Neanderthals and modern humans recently received much attention, owing to the recovery of a Neanderthal mtDNA fragment, the analysis of which indicated that the most-recent common ancestor (MRCA) of Neanderthal and modern-human mitochondria was several times more ancient than that of modern humans only (Krings et al. 1997; fig. 1). This finding was considered to be strong evidence that Neanderthals and anatomically modern humans are separate species, the latter having replaced the former without interbreeding ("In our genes?" 1997; Kahn and Gibbons 1997; Lindahl 1997; Wade 1997; Ward and Stringer 1997). Here, I investigate the strength of this evidence by considering the probability of erroneous rejection of interbreeding (i.e., the probability of a type I error). I demonstrate that, although completely random mating clearly can be rejected, more-relevant models of interbreeding cannot.

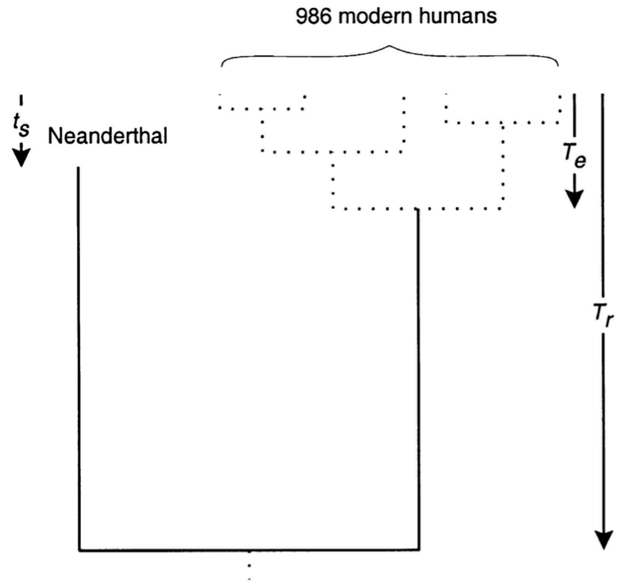



**Figure 1** Schematic genealogy of the 986 modern-human mtDNAs and a single Neanderthal mtDNA (the carrier of which lived at time $t_s$ before the present). The MRCA of the entire sample was inferred to be at least four times more ancient than the MRCA of the modern sample—that is, $T_r \geq 4T_e$ (Krings et al. 1997).

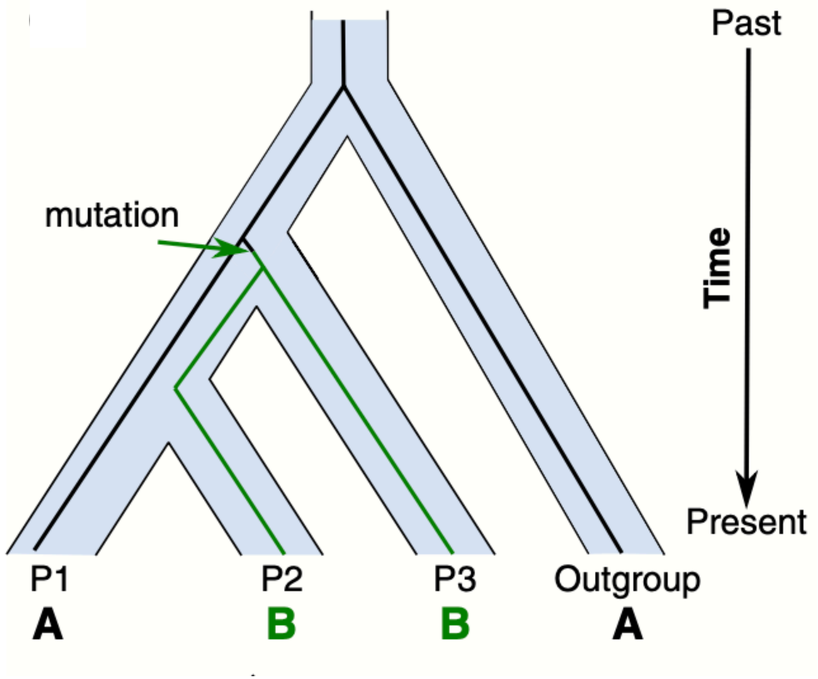
RESEARCH ARTICLE

# A Draft Sequence of the Neandertal Genome

Richard E. Green,[1]*†‡ Johannes Krause,[1]†§ Adrian W. Briggs,[1]†§ Tomislav Maricic,[1]†§ Udo Stenzel,[1]†§ Martin Kircher,[1]†§ Nick Patterson,[2]†§ Heng Li,[2]† Weiwei Zhai,[3]†‖ Markus Hsi-Yang Fritz,[4]† Nancy F. Hansen,[5]† Eric Y. Durand,[3]† Anna-Sapfo Malaspinas,[3]† Jeffrey D. Jensen,[6]† Tomas Marques-Bonet,[7,13]† Can Alkan,[7]† Kay Prüfer,[1]† Matthias Meyer,[1]† Hernán A. Burbano,[1]† Jeffrey M. Good,[1,8]† Rigo Schultz,[1] Ayinuer Aximu-Petri,[1] Anne Butthof,[1] Barbara Höber,[1] Barbara Höffner,[1] Madlen Siegemund,[1] Antje Weihmann,[1] Chad Nusbaum,[2] Eric S. Lander,[2] Carsten Russ,[2] Nathaniel Novod,[2] Jason Affourtit,[9] Michael Egholm,[9] Christine Verna,[21] Pavao Rudan,[10] Dejana Brajkovic,[11] Željko Kucan,[10] Ivan Gušic,[10] Vladimir B. Doronichev,[12] Liubov V. Golovanova,[12] Carles Lalueza-Fox,[13] Marco de la Rasilla,[14] Javier Fortea,[14]¶ Antonio Rosas,[15] Ralf W. Schmitz,[16,17] Philip L. F. Johnson,[18]† Evan E. Eichler,[7]† Daniel Falush,[19]† Ewan Birney,[4]† James C. Mullikin,[5]† Montgomery Slatkin,[3]† Rasmus Nielsen,[3]† Janet Kelso,[1]† Michael Lachmann,[1]† David Reich,[2,20]*† Svante Pääbo[1]*†

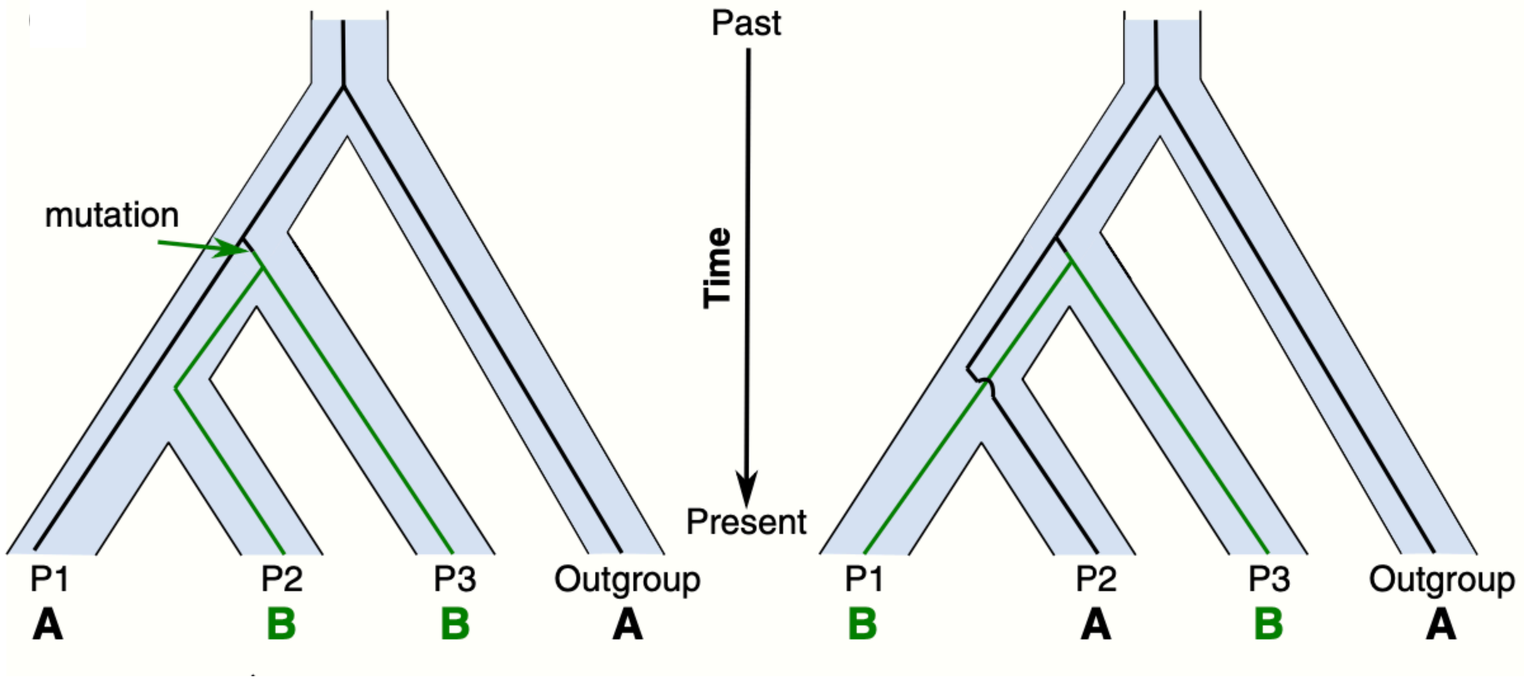
Neandertals, the closest evolutionary relatives of present-day humans, lived in large parts of Europe and western Asia before disappearing 30,000 years ago. We present a draft sequence of the Neandertal genome composed of more than 4 billion nucleotides from three individuals. Comparisons of the Neandertal genome to the genomes of five present-day humans from different parts of the world identify a number of genomic regions that may have been affected by positive selection in ancestral modern humans, including genes involved in metabolism and in cognitive and skeletal development. We show that Neandertals shared more genetic variants with present-day humans in Eurasia than with present-day humans in sub-Saharan Africa, suggesting that gene flow from Neandertals into the ancestors of non-Africans occurred before the divergence of Eurasian groups from each other.

# The D-statistic (ABBA-BABA test)



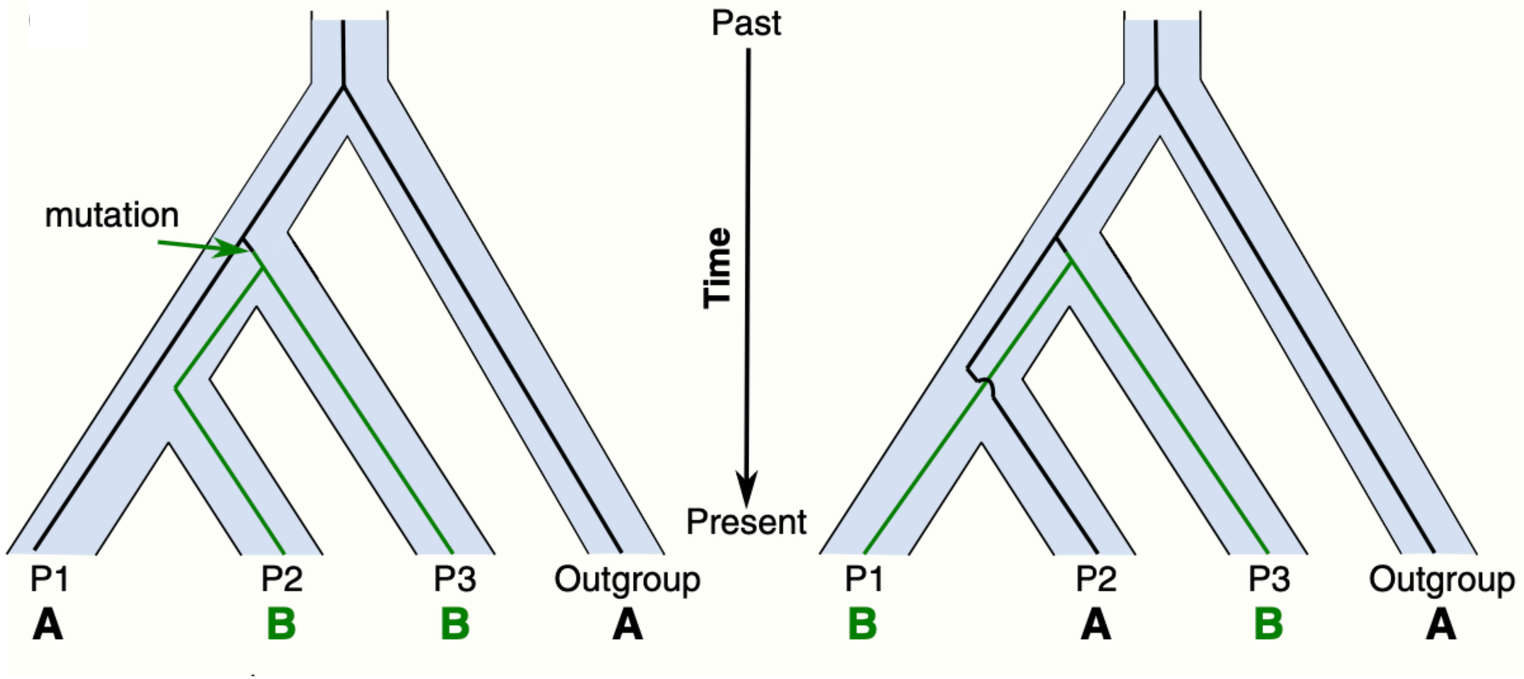

Malinsky et al (2021) Mol Ecol Resour

# The D-statistic (ABBA-BABA test)




Malinsky et al (2021) Mol Ecol Resour

# The D-statistic (ABBA-BABA test)

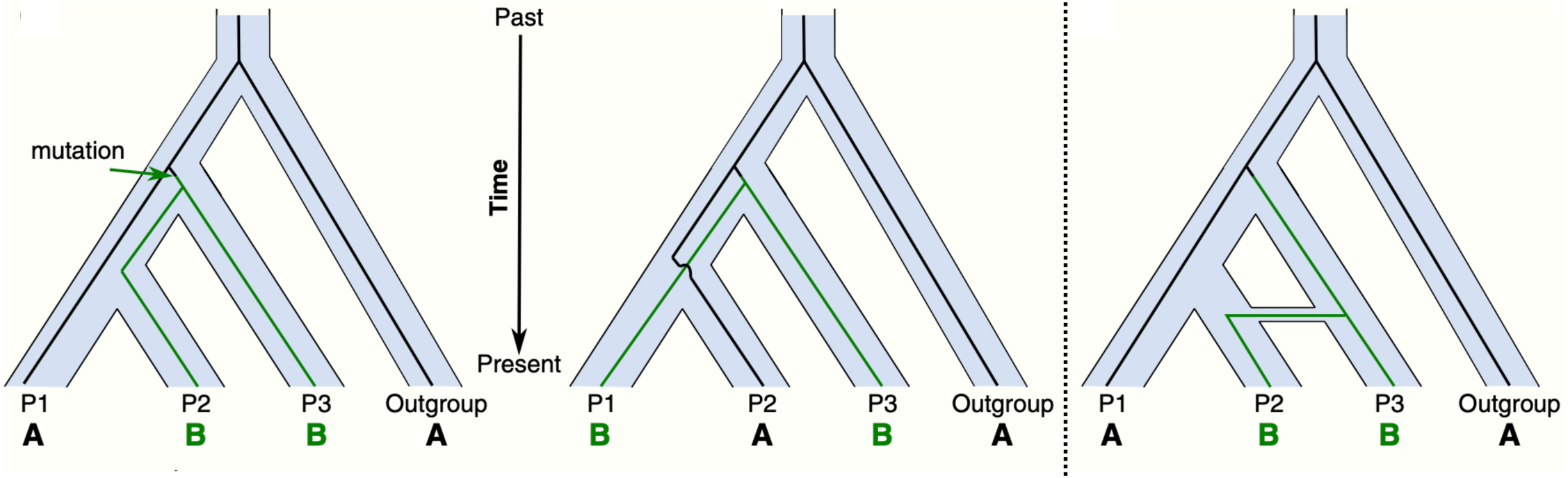



$$D = \frac{nABBA - nBABA}{nABBA + nBABA}$$

Ancestral polymorphism (D=0)

Malinsky et al (2021) Mol Ecol Resour

# The D-statistic (ABBA-BABA test)




$$D = \frac{nABBA - nBABA}{nABBA + nBABA}$$

Ancestral polymorphism (D=0)

$$D = \frac{nABBA - nBABA}{nABBA + nBABA}$$

Gene flow (D>0)

Malinsky et al (2021) Mol Ecol Resour

# The D-statistic (ABBA-BABA test)

| $H_1$ | $H_2$ | $H_3$ | Number ABBA sites | Number BABA sites | D=(ABBA-BABA)/(ABBA+BABA) Value% | Std.err% | Z-score for D≠0 | Interpretation (" indicates same as previous row) |
|---|---|---|---|---|---|---|---|---|
| San | Yoruba | Neandertal | 99,515 | 99,788 | -0.1 | ± 0.3 | **-0.4** | Neandertal equally close to Africans |
| French | Han | Neandertal | 74,477 | 73,089 | 0.9 | ± 0.5 | **1.7** | Neandertal equally close to non-Africans |
| French | Papuan | Neandertal | 70,094 | 70,093 | 0 | ± 0.5 | **0.0** | " |
| Han | Papuan | Neandertal | 67,022 | 68,260 | -0.9 | ± 0.6 | **-1.4** | " |

Green et al (2010) Science; Durand et al (2011) Mol Biol Evol

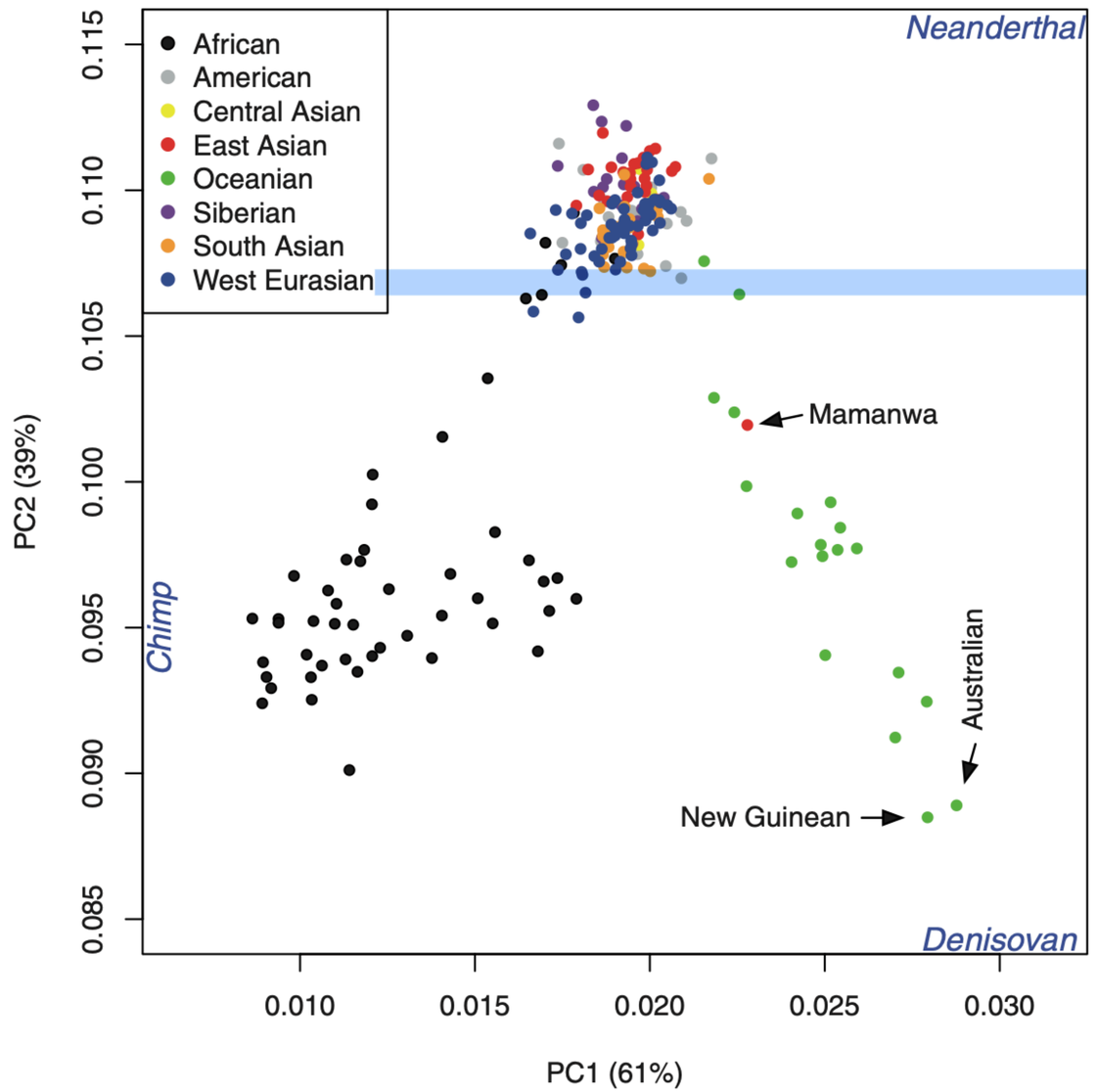
# The D-statistic (ABBA-BABA test)

| $H_1$ | $H_2$ | $H_3$ | Number ABBA sites | Number BABA sites | D=(ABBA-BABA)/(ABBA+BABA) Value% | Std.err% | Z-score for D≠0 | Interpretation (" indicates same as previous row) |
|---|---|---|---|---|---|---|---|---|
| San | Yoruba | Neandertal | 99,515 | 99,788 | -0.1 | ± 0.3 | **-0.4** | Neandertal equally close to Africans |
| French | Han | Neandertal | 74,477 | 73,089 | 0.9 | ± 0.5 | **1.7** | Neandertal equally close to non-Africans |
| French | Papuan | Neandertal | 70,094 | 70,093 | 0 | ± 0.5 | **0.0** | " |
| Han | Papuan | Neandertal | 67,022 | 68,260 | -0.9 | ± 0.6 | **-1.4** | " |
| French | San | Neandertal | 95,347 | 103,612 | -4.2 | ± 0.5 | -9.3 | Neandertal gene flow with non-Africans |
| French | Yoruba | Neandertal | 84,025 | 92,066 | -4.6 | ± 0.4 | -10.5 | " |
| Han | San | Neandertal | 94,029 | 103,590 | -4.8 | ± 0.5 | -9.9 | " |
| Han | Yoruba | Neandertal | 82,575 | 91,872 | -5.3 | ± 0.5 | -10.5 | " |
| Papuan | San | Neandertal | 90,059 | 97,088 | -3.8 | ± 0.5 | -7.0 | " |
| Papuan | Yoruba | Neandertal | 79,529 | 86,570 | -4.2 | ± 0.6 | -7.5 | " |

ABBA-BABA test provides evidence for Neanderthal gene flow into modern non-Africans

Green et al (2010) Science; Durand et al (2011) Mol Biol Evol

# Archaic gene flow in PCA

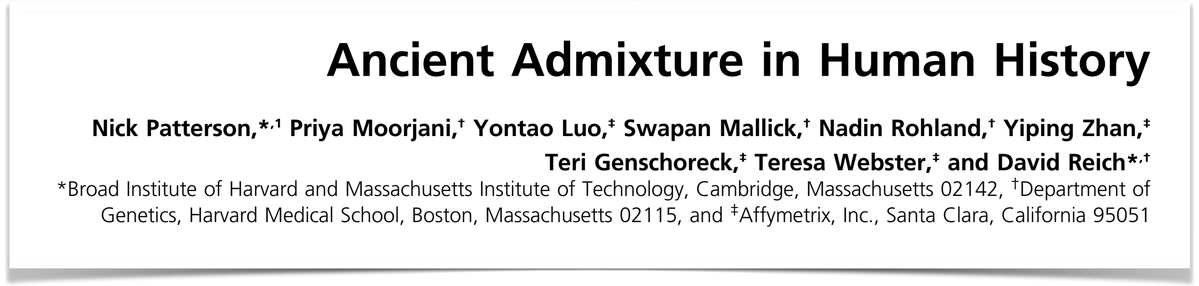



Projection of modern human individuals on PCs from archaic humans and chimp

Prüfer et al (2014) Nature

# F-statistics background

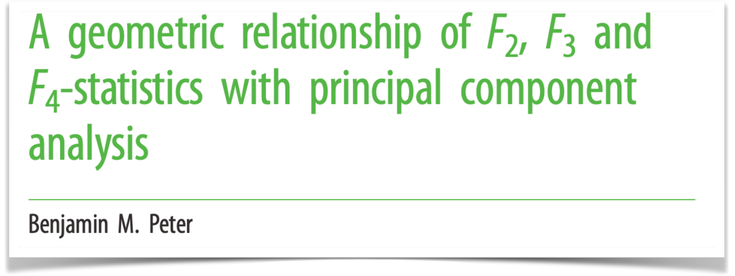
- A framework for admixture inference using allele frequency correlations / shared genetic drift among sets of populations
- Components appeared earlier in the literature, but seminal paper is Patterson et al. 2012
- Recent theoretic work relating f-statistics to population genetics theory and PCA
- Has become a standard toolset to test hypotheses about population history and admixture, efficient to calculate and applicable to lower quality data (e.g. ancient DNA)



**Ancient Admixture in Human History**

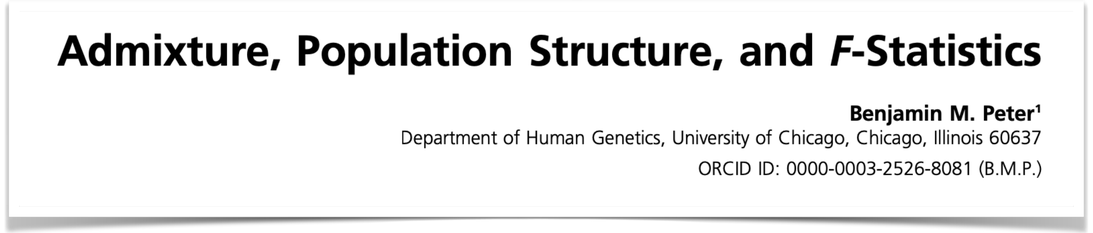
**Nick Patterson,*,[1] Priya Moorjani,[†] Yontao Luo,[‡] Swapan Mallick,[†] Nadin Rohland,[†] Yiping Zhan,[‡] Teri Genschoreck,[‡] Teresa Webster,[‡] and David Reich*,[†]**

*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



A geometric relationship of $F_2$, $F_3$ and $F_4$-statistics with principal component analysis

Benjamin M. Peter



**Admixture, Population Structure, and *F*-Statistics**

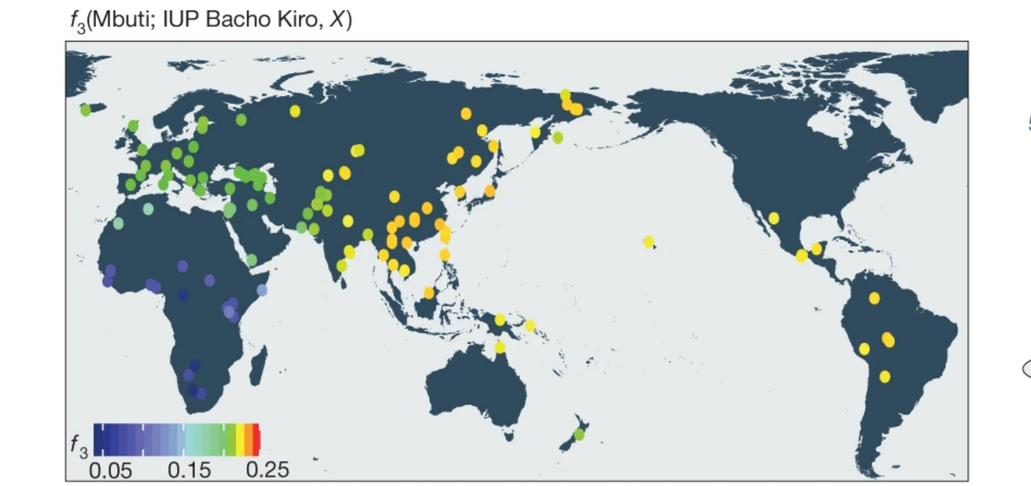
**Benjamin M. Peter[1]**

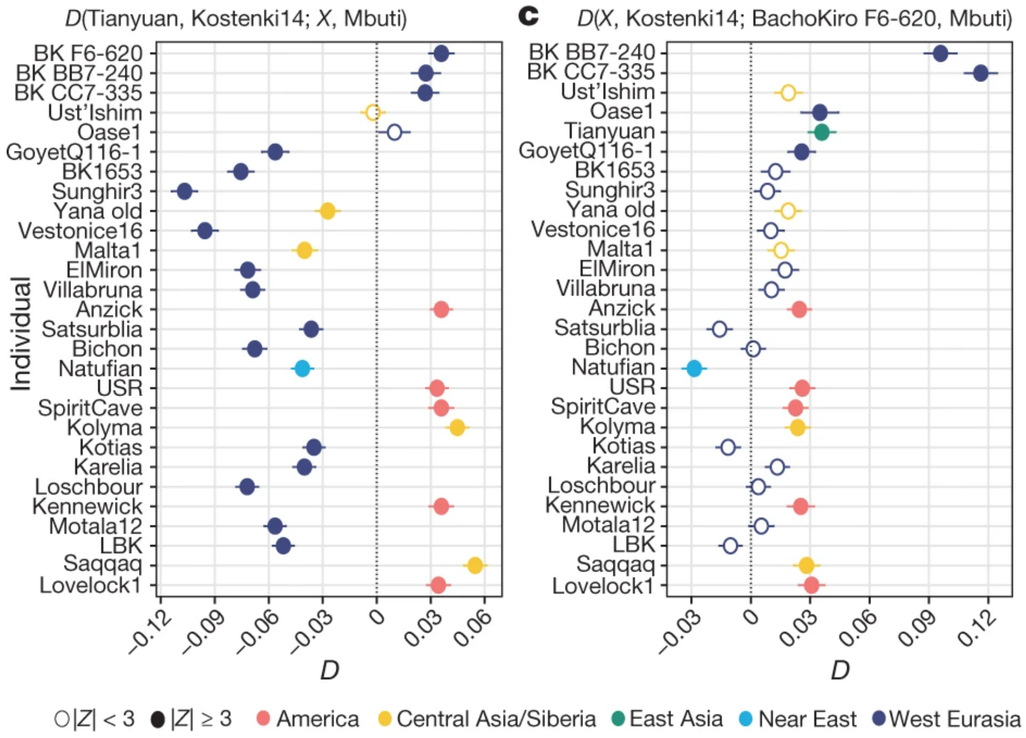
Department of Human Genetics, University of Chicago, Chicago, Illinois 60637

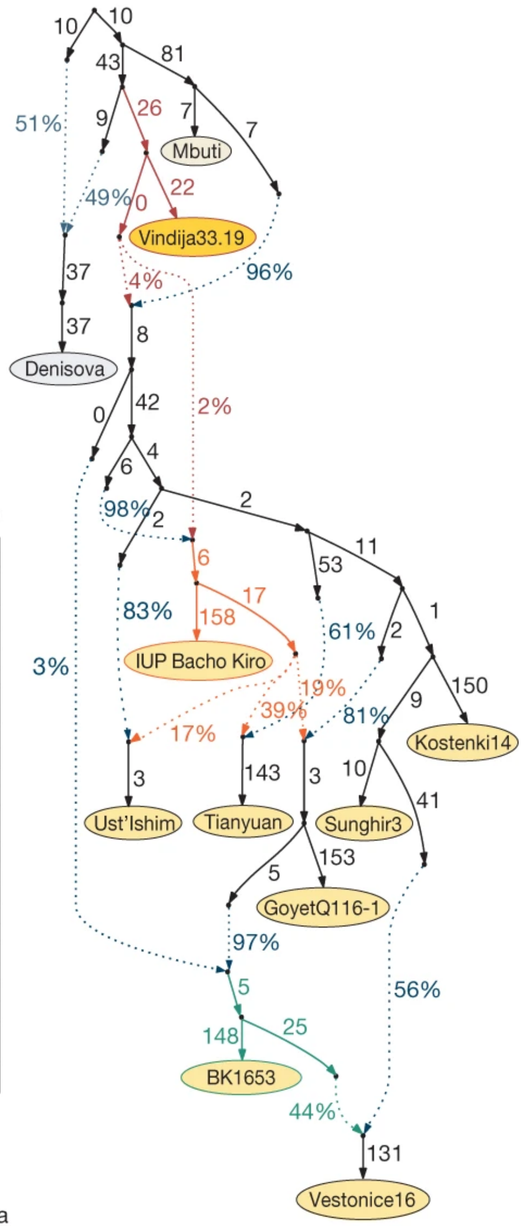
ORCID ID: 0000-0003-2526-8081 (B.M.P.)

Patterson et al (2012) Genetics; Peter (2016) Genetics; Peter (2022) Phil Trans R Soc B

# F-statistics are the standard toolkit for aDNA analysis



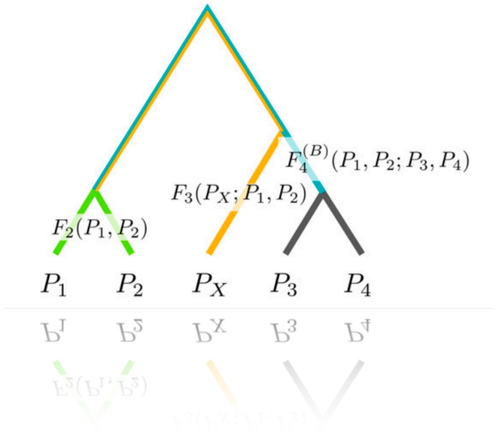




Hajdinjak et al (2021) Nature

## Background

What are f-statistics and why do we use them?




## Definitions

$F_2$, $F_3$ and $F_4$ statistics



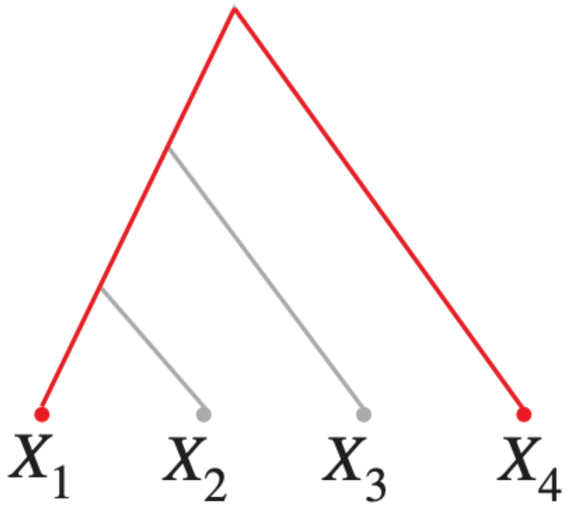

## Applications

Testing hypotheses about population admixture

# F-statistics definitions

Sums of products of allele frequency differences between pairs of 2, 3, or 4 populations

$$F_2(X_1, X_4) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{4l})^2$$



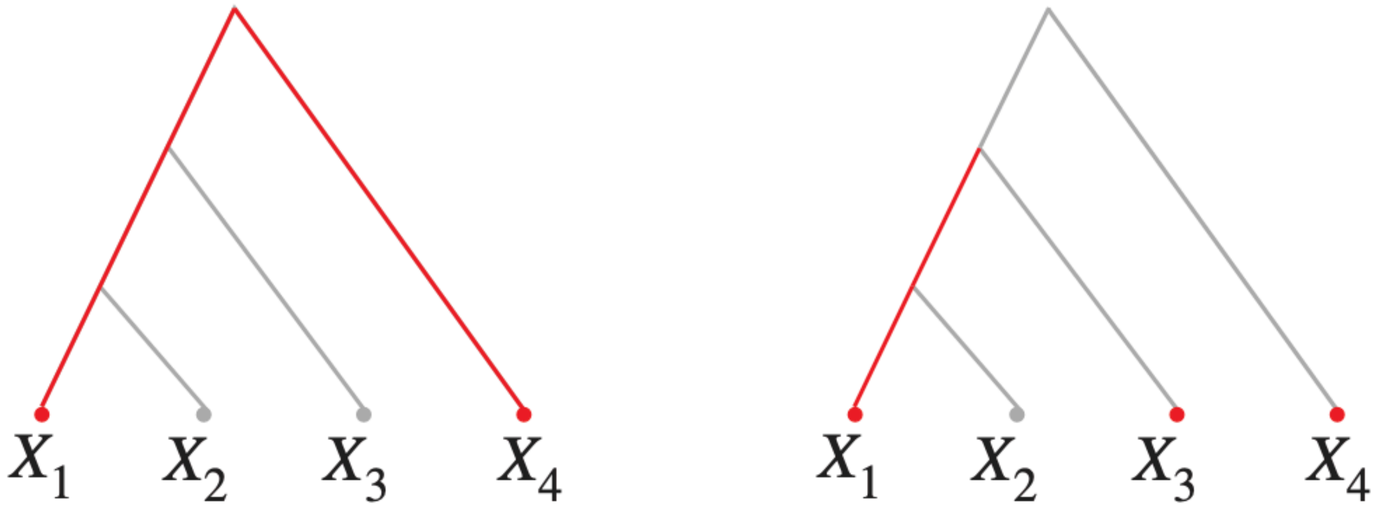

Interpretation as branch lengths in a population tree

Peter (2022) Phil Trans R Soc B

# F-statistics definitions

Sums of products of allele frequency differences between pairs of 2, 3, or 4 populations

$$F_2(X_1, X_4) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{4l})^2 \qquad F_3(X_1; X_2, X_3) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{1l} - x_{3l})$$



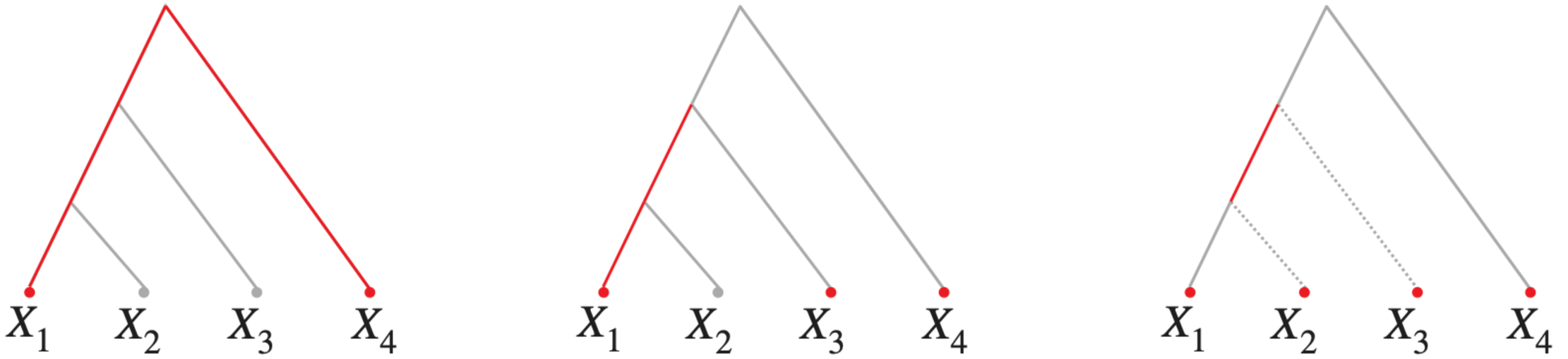

Interpretation as branch lengths in a population tree

Peter (2022) Phil Trans R Soc B

# F-statistics definitions

Sums of products of allele frequency differences between pairs of 2, 3, or 4 populations

$$F_2(X_1, X_4) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{4l})^2 \qquad F_3(X_1; X_2, X_3) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{1l} - x_{3l}) \qquad F_4(X_1, X_4; X_3, X_2) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{4l})(x_{3l} - x_{2l}).$$




Interpretation as branch lengths in a population tree

Peter (2022) Phil Trans R Soc B

# F-statistics definitions

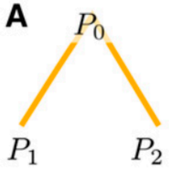
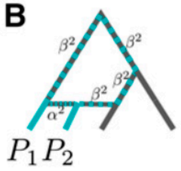
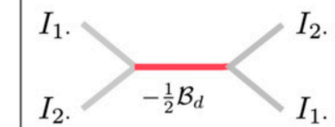
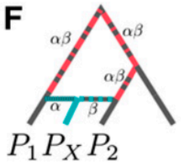
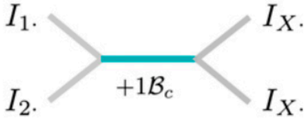
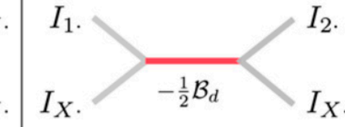
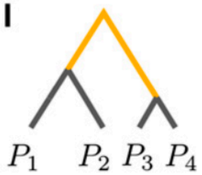
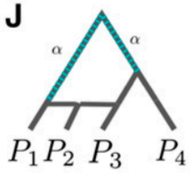
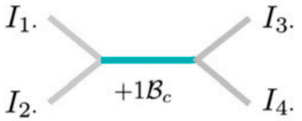
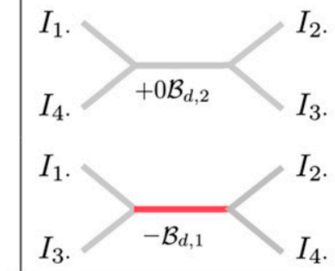
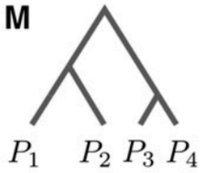
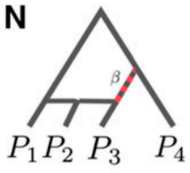
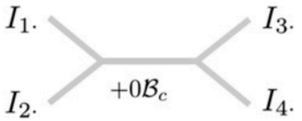
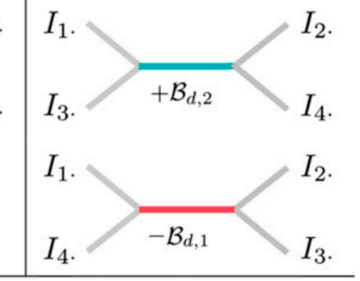
$$2F_3(X_1; X_2, X_3) = F_2(X_1, X_2) + F_2(X_1, X_3) - F_2(X_2, X_3)$$

$$2F_4(X_1, X_2; X_3, X_4) = F_2(X_1, X_3) + F_2(X_2, X_4) - F_2(X_1, X_4) - F_2(X_2, X_3).$$

$F_3$ and $F_4$ can be expressed as sums of $F_2$ statistics

Peter (2022) Phil Trans R Soc B

# Other interpretations of F-statistics

| | Branch length | Path | Gene tree: concordant | Gene tree: discordant |
|---|---|---|---|---|
| $F_2(P_1, P_2)$ | **A**   | **B**   | **C**   | **D**   |
| $F_3(P_X; P_1, P_2)$ | **E**   | **F**   | **G**   | **H**   |
| $F_4^{(B)}(P_1; P_2; P_3, P_4)$ $F_4(P_1; P_3; P_2, P_4)$ (internal branch) | **I**   | **J**   | **K**   | **L**   |
| $F_4^{(T)}(P_1; P_2; P_3, P_4)$ $F_4(P_1; P_2; P_3, P_4)$ (branch absent) | **M**   | **N**   | **O**   | **P**   |

Peter (2016) Genetics

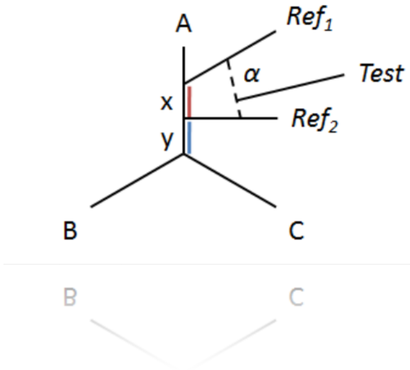
## Background

What are f-statistics and why do we use them?




## Definitions

$F_2$, $F_3$ and $F_4$ statistics




## Applications

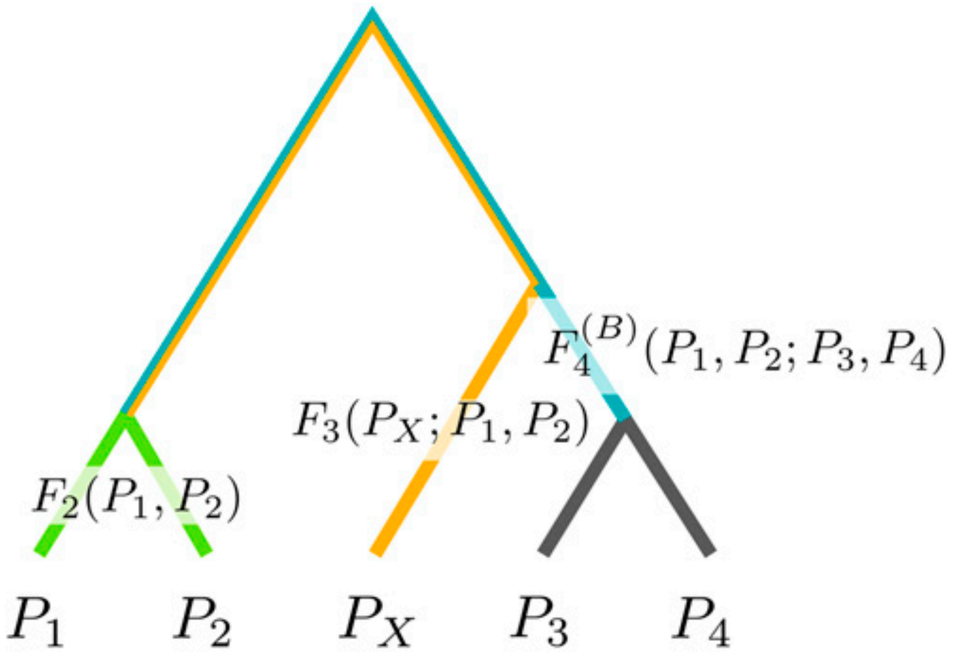
Testing hypotheses about population admixture

# Applications of F-statistics

| 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 |

# Admixture graphs

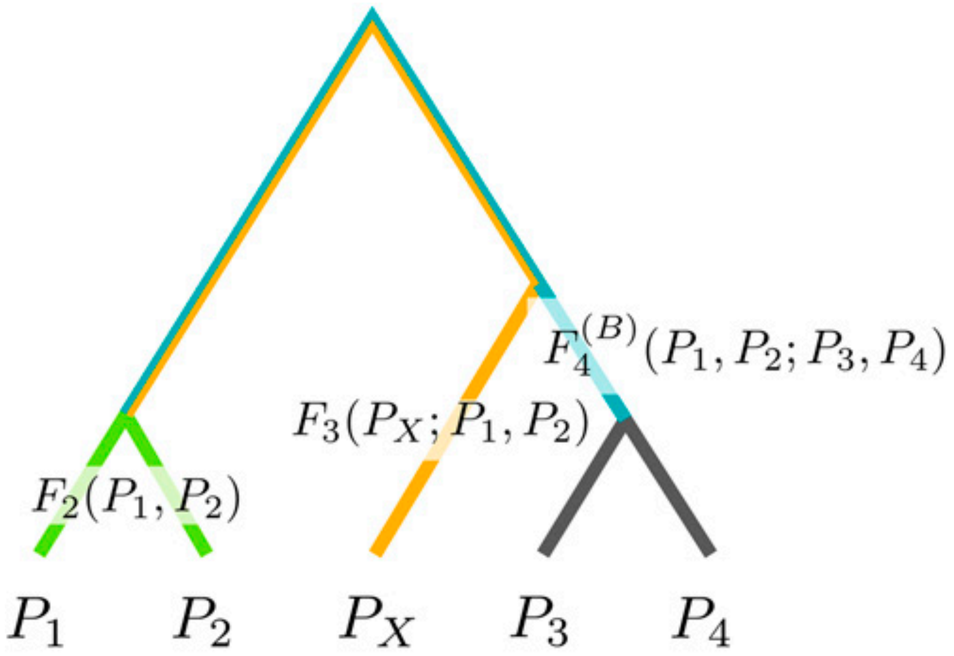
Population phylogeny



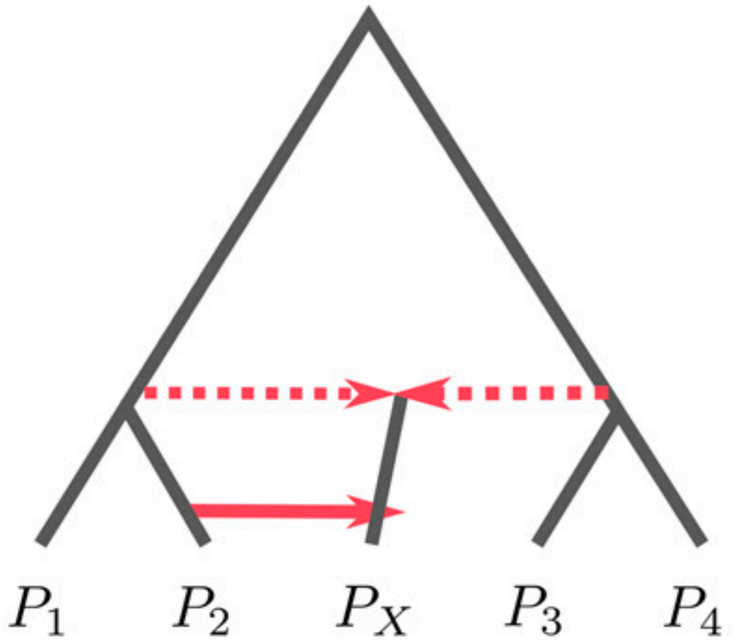

Peter (2016) Genetics

# Admixture graphs

Population phylogeny



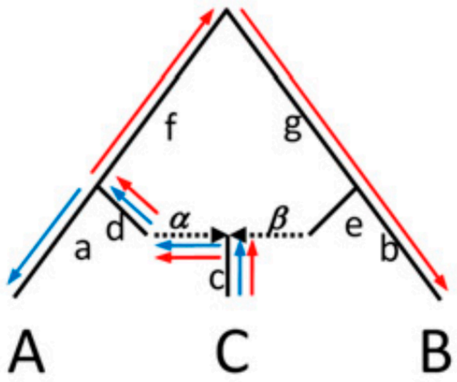

Admixture graph




Admixture graphs extend population phylogenies by allowing for gene flow

Peter (2016) Genetics

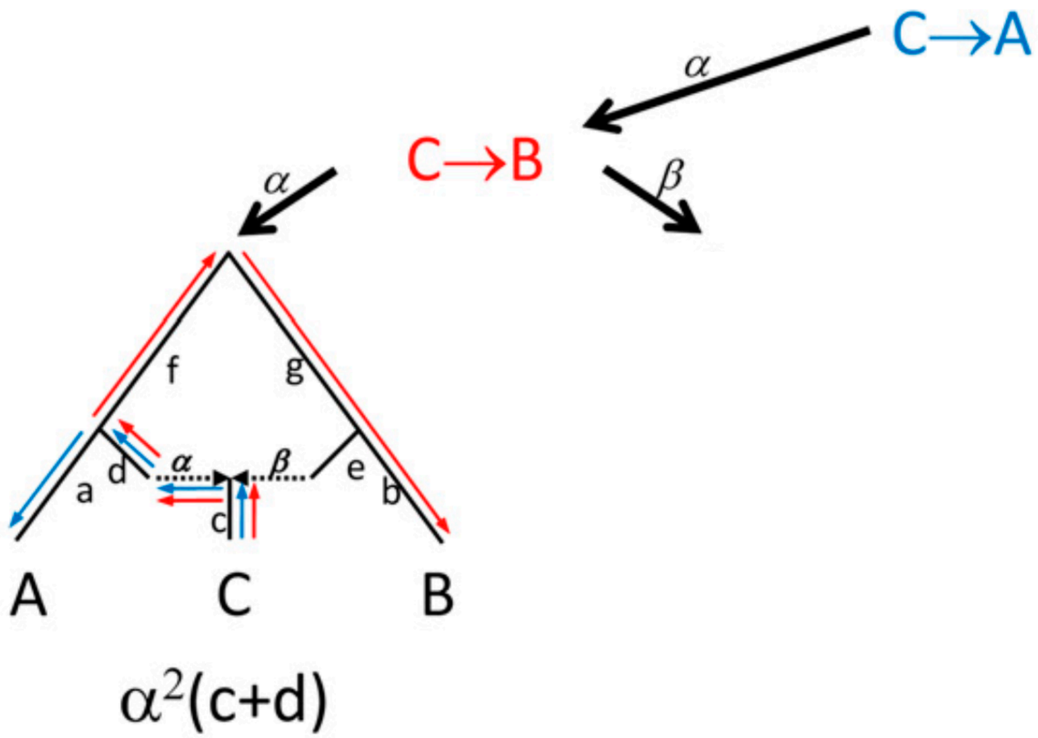
# Admixture $F_3$




Deriving expectation of $F_3$(C;A,B) using path overlap in admixture graph

Patterson et al (2012) Genetics

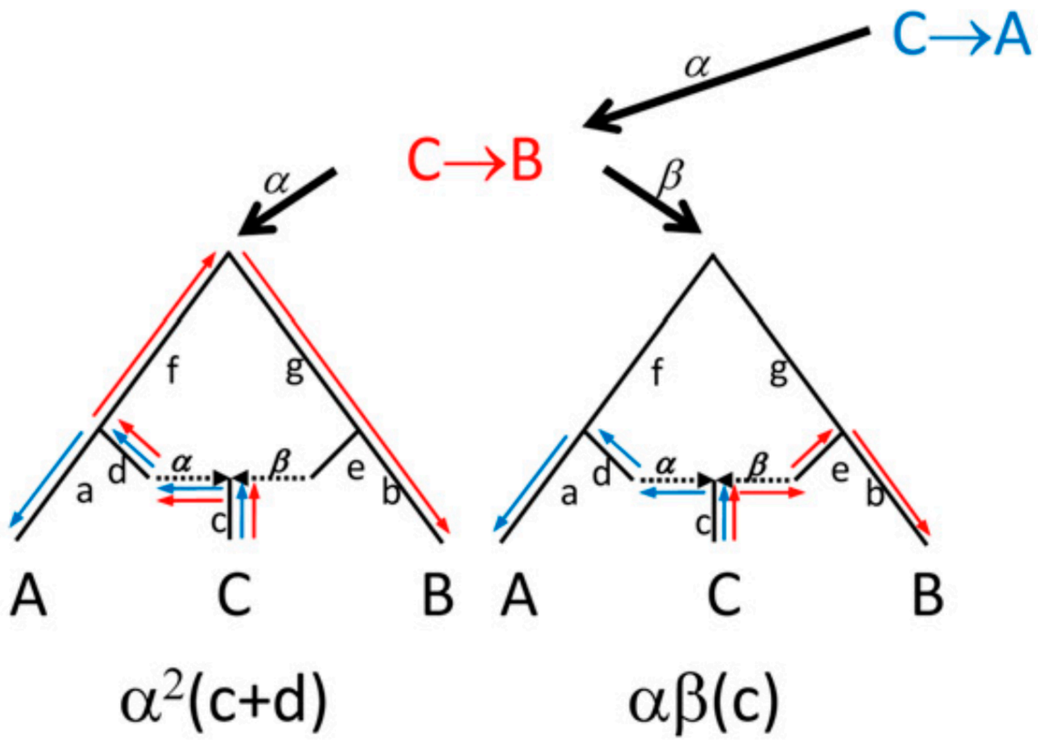
# Admixture $F_3$




Deriving expectation of $F_3$(C;A,B) using path overlap in admixture graph

Patterson et al (2012) Genetics

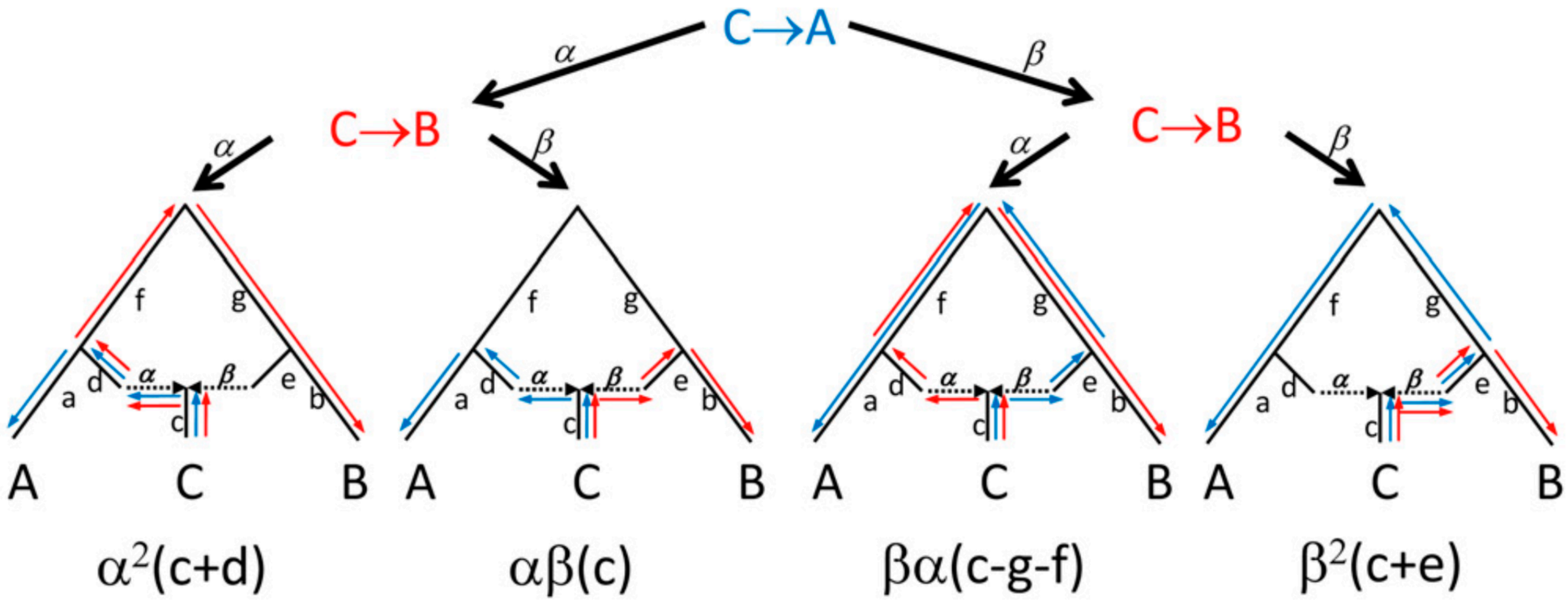
# Admixture $F_3$




Deriving expectation of $F_3$(C;A,B) using path overlap in admixture graph

Patterson et al (2012) Genetics

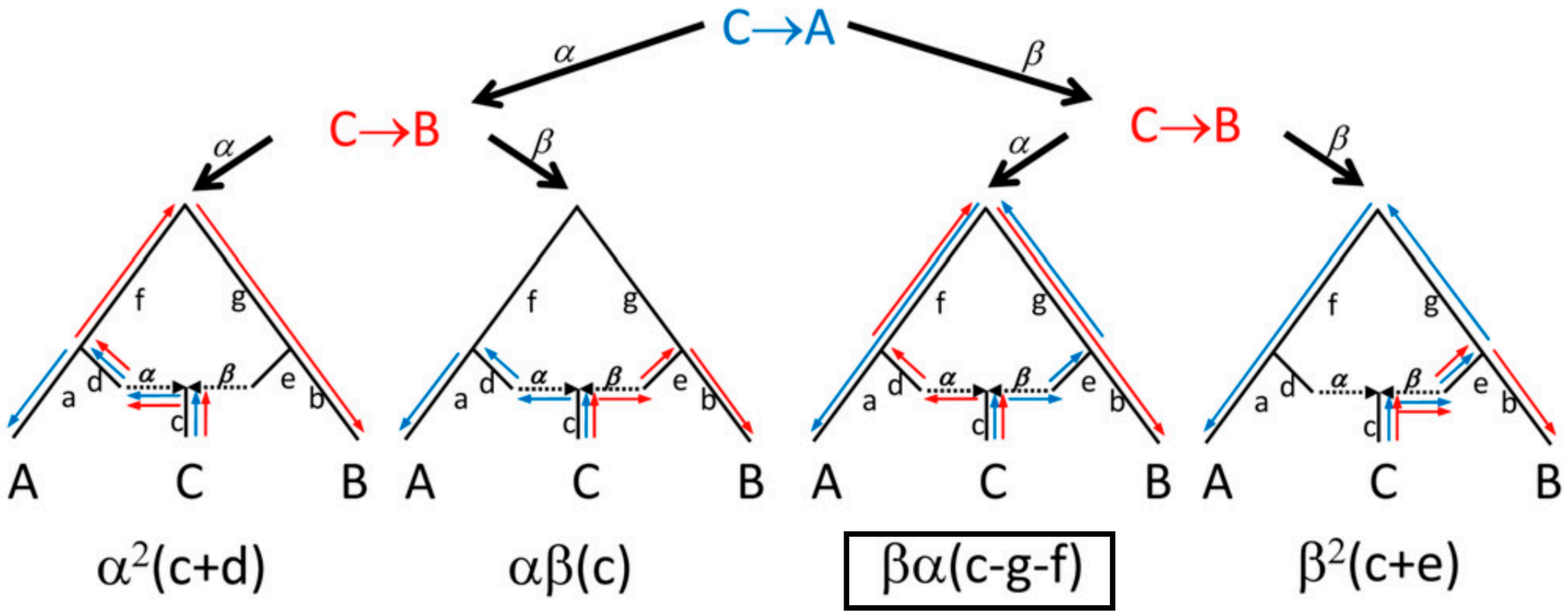
# Admixture $F_3$




Deriving expectation of $F_3$(C;A,B) using path overlap in admixture graph

Patterson et al (2012) Genetics

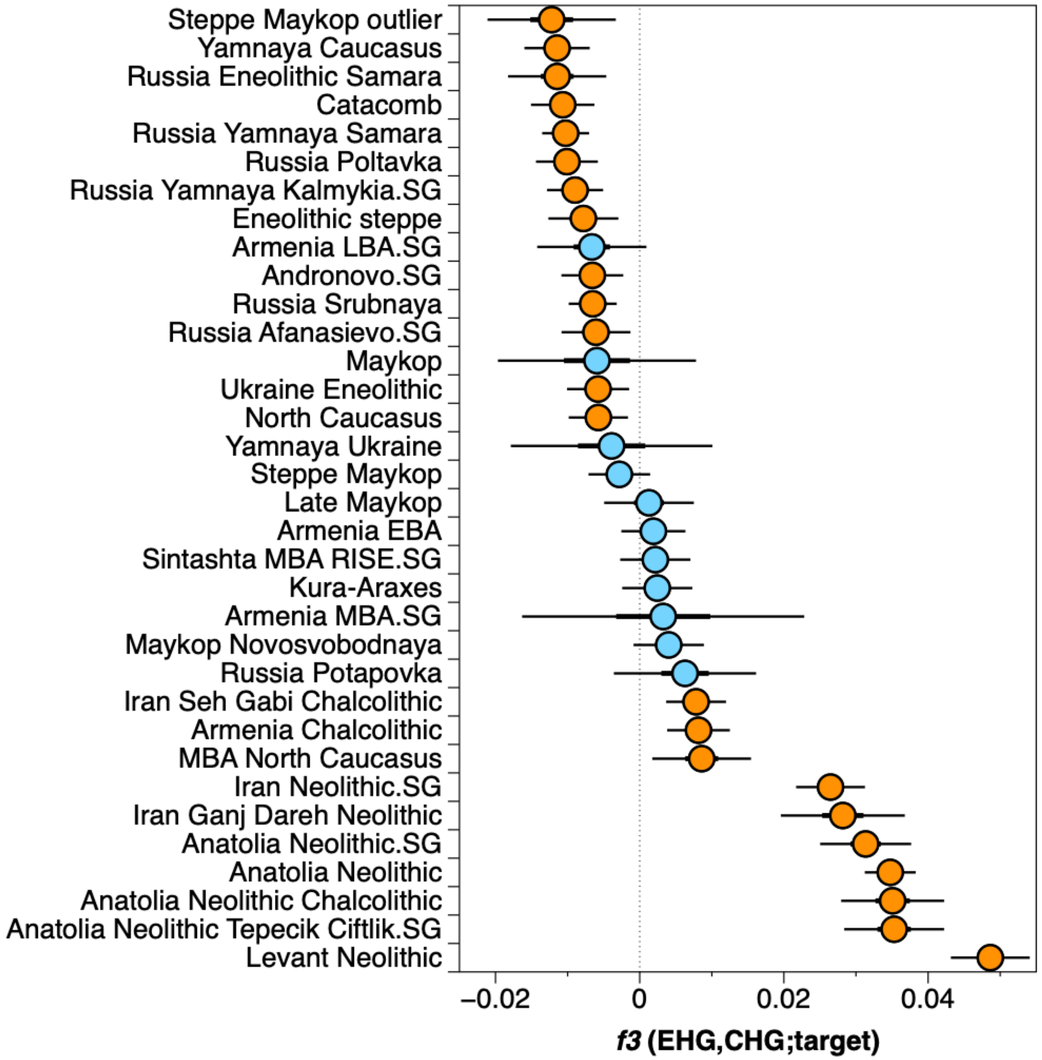
# Admixture $F_3$




$F_3$ (C; A,B) can only negative if C is a mixture of populations related (possibly distantly!) to A and B

Patterson et al (2012) Genetics

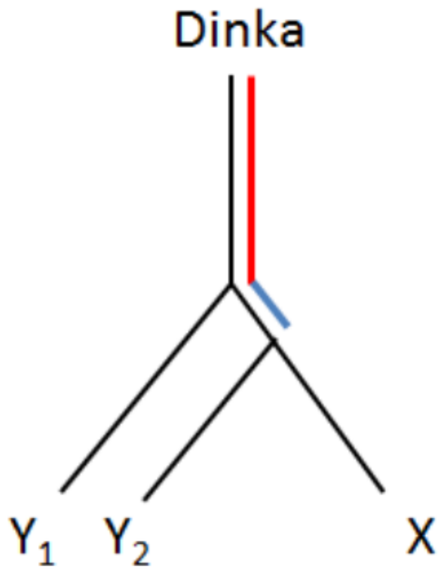
# Example of admixture $F_3$ in real data

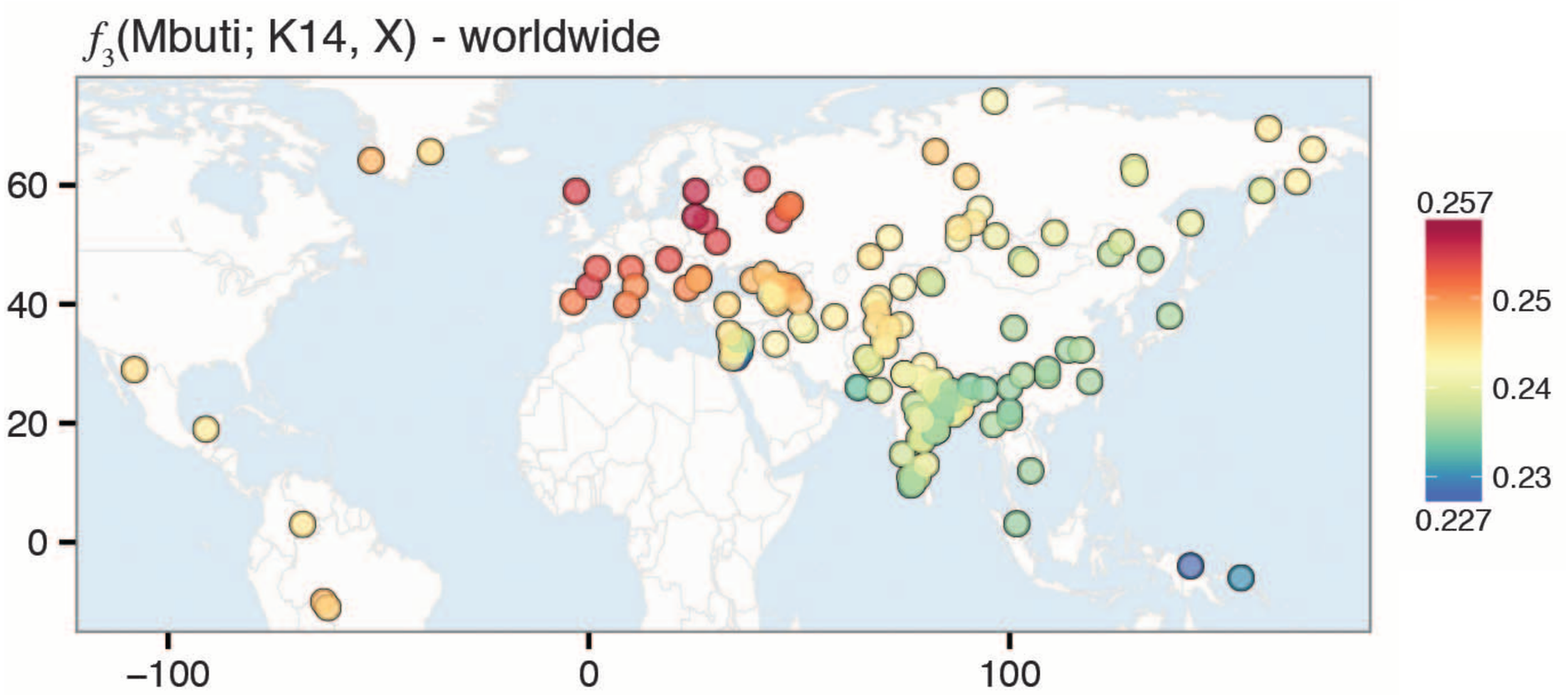



$F_3$ test indicates that ancient Steppe populations are admixed related to Eastern and Caucasus hunter-gatherers

Wang et al (2019) Nat Commun

# Outgroup $F_3$



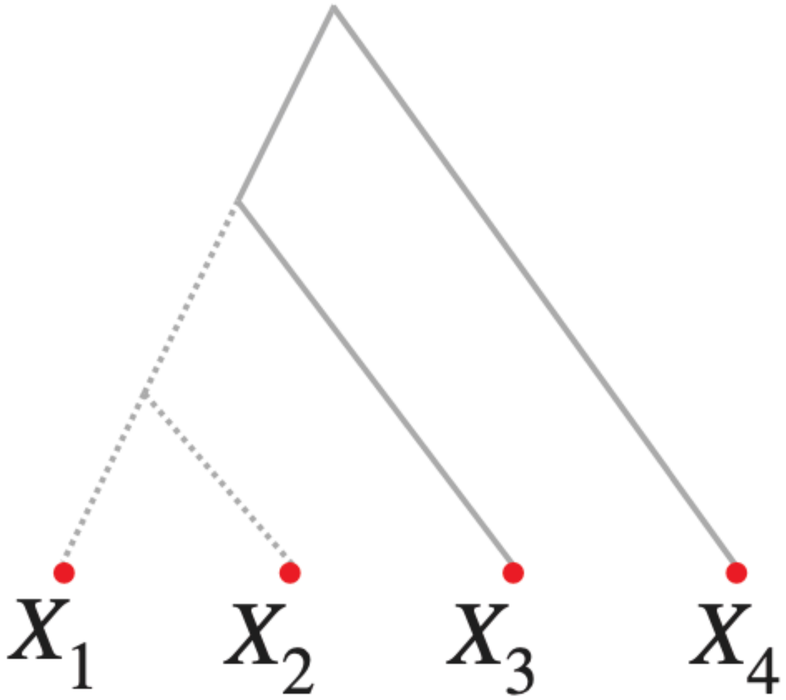




Outgroup - $F_3$ statistics measure the shared drift between an outgroup and the divergence of two test populations

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

# Testing for treeness with $F_4$

$F_4(X_1,X_2;X_3,X_4) = 0$

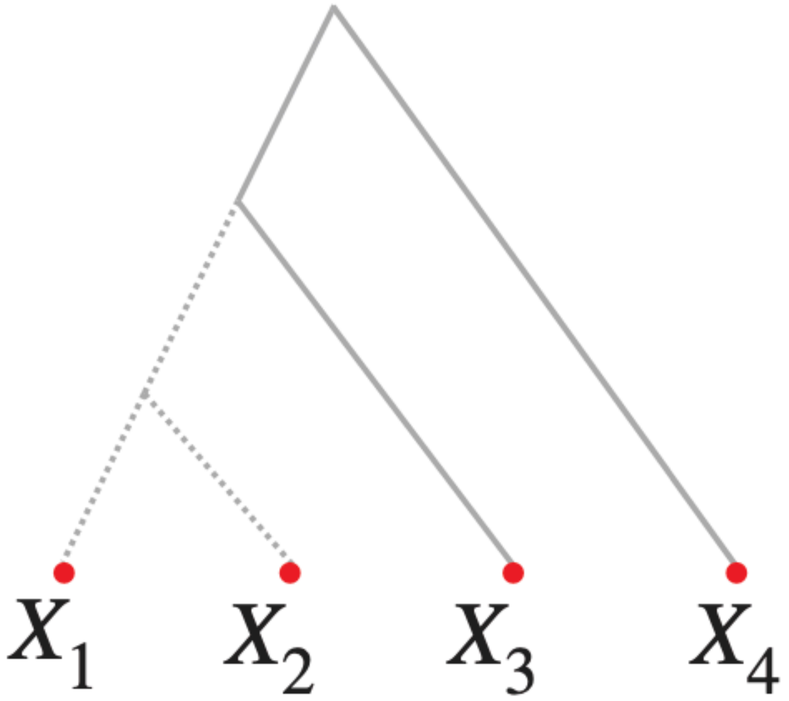



$F_4(X_1,X_2;X_3,X_4) = 0$ if populations $(X_1,X_2)$ form a clade with respect to $(X_3,X_4)$

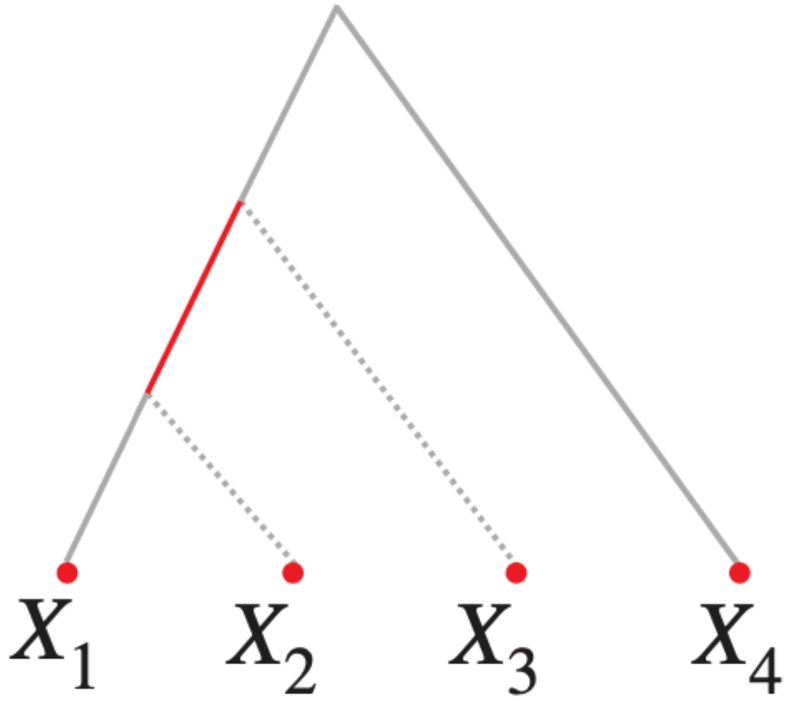
Peter (2022) Phil Trans R Soc B

# Testing for treeness with $F_4$

$F_4(X_1,X_2;X_3,X_4) = 0$



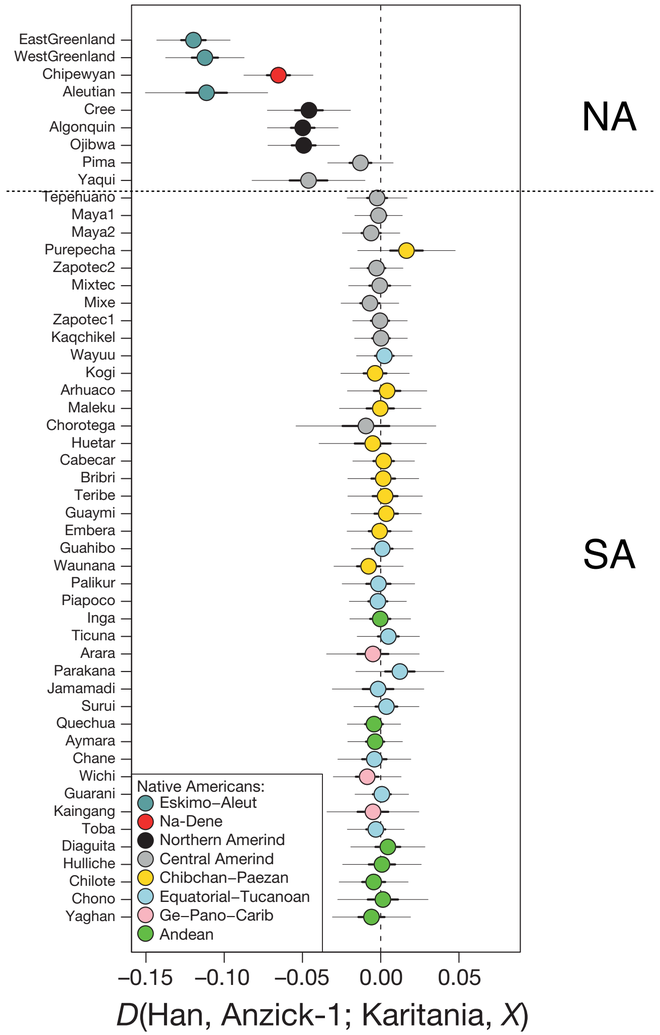

$F_4(X_1,X_3;X_2,X_4) > 0$




Remaining configurations have to also consistently group $(X_1,X_2)$ and $(X_3,X_4)$

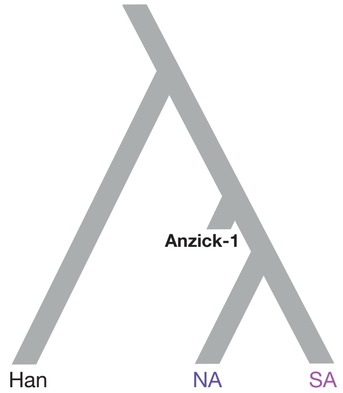
Peter (2022) Phil Trans R Soc B

# Testing for treeness with $F_4$




NA

SA

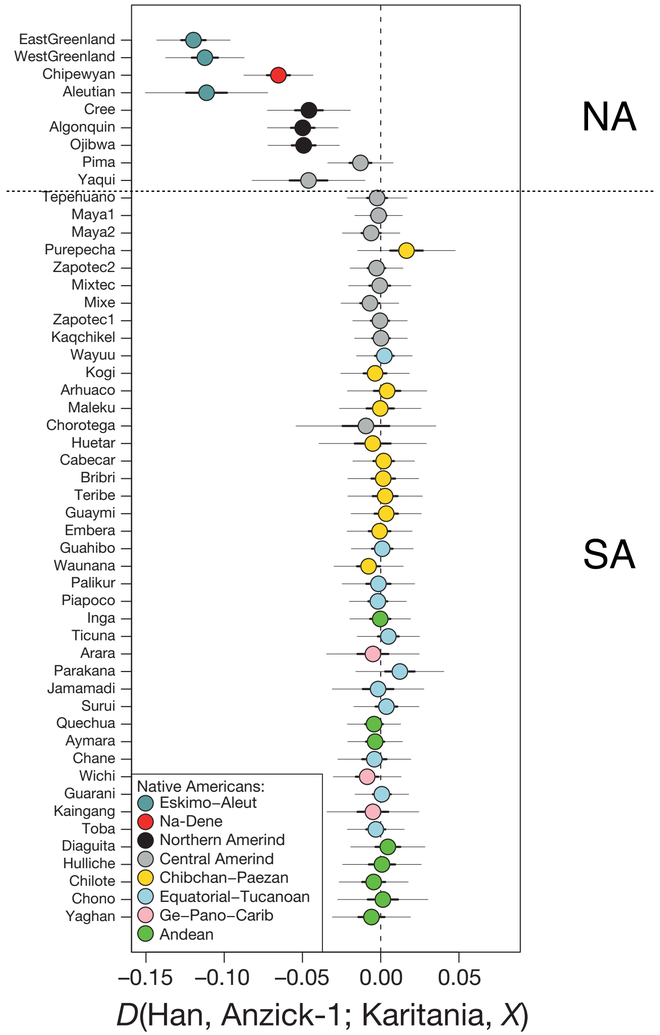



Rejected

Failed treeness test can be interpreted in multiple ways!

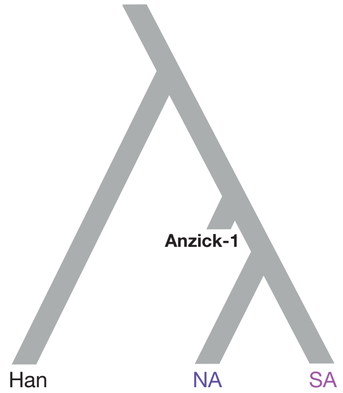
Rasmussen et al (2014) Nature

# Testing for treeness with $F_4$

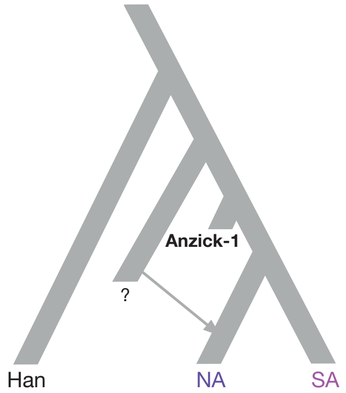



NA

SA

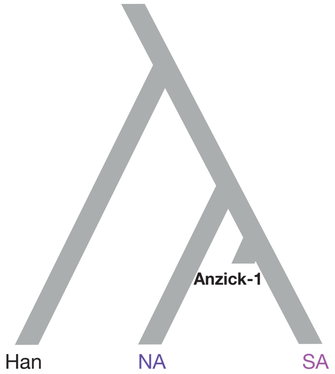



Rejected

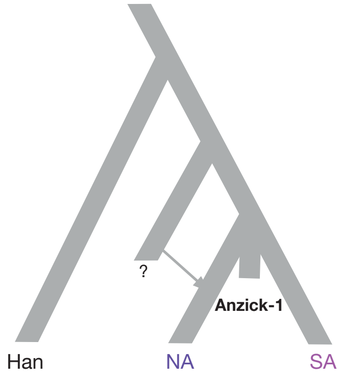



Not supported by the data

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




Consistent with the data

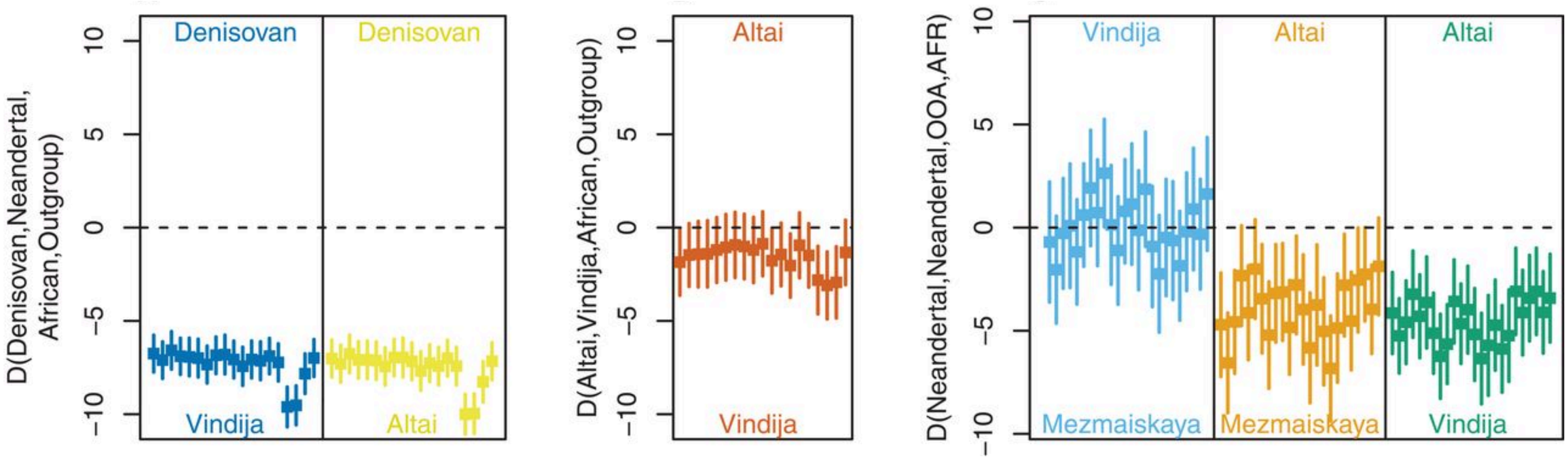



Consistent with the data

Failed treeness test can be interpreted in multiple ways!

Rasmussen et al (2014) Nature

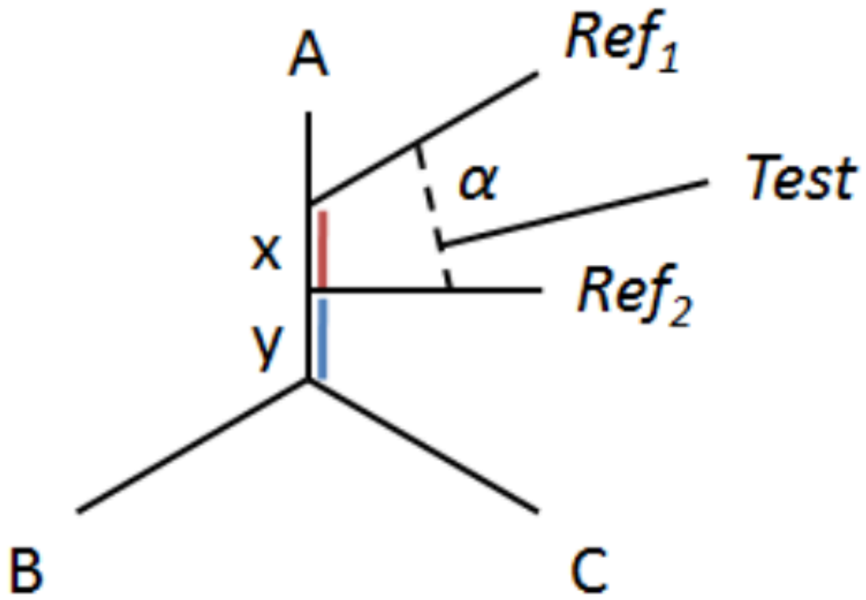
# Symmetry test with $F_4$ / D




Testing for symmetry with either D- or $F_4$-statistic by using known outgroup as one population

Prüfer et al. (2017) Science

# Estimating admixture proportions - qpAdm

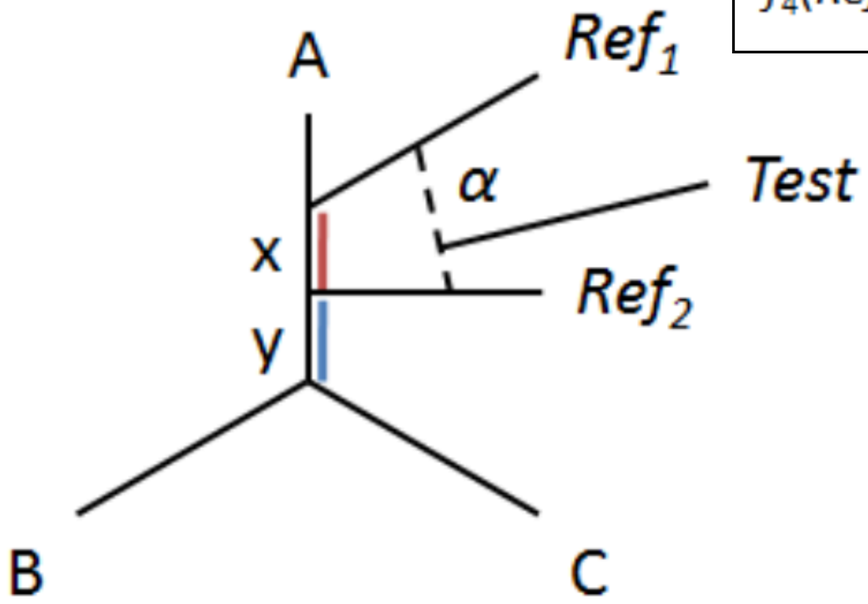



Differential drift sharing with sets of outgroup (or "right") populations A,B,C is leveraged to infer mixture proportions

Haak et al (2015) Nature

# Estimating admixture proportions - qpAdm

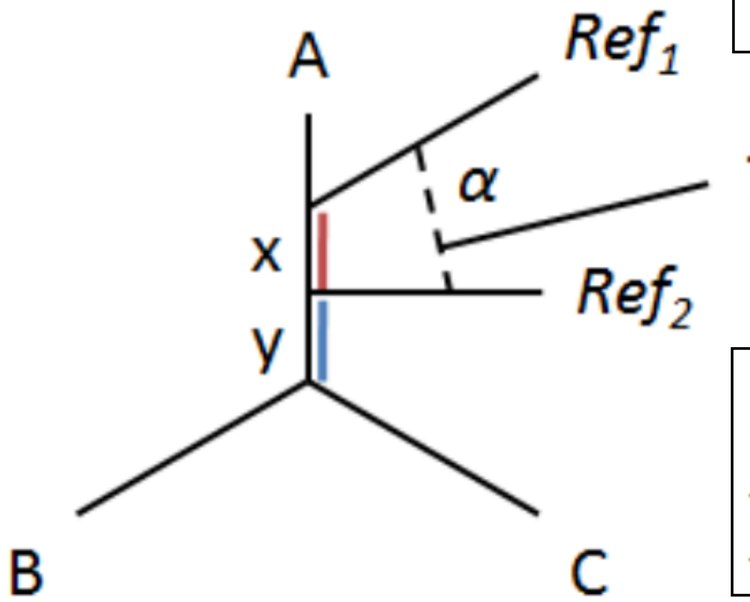
$f_4(Ref_1, A; B, C) = 0$
$f_4(Ref_1, B; A, C) = x+y$
$f_4(Ref_1, C; A, B) = x+y$




Differential drift sharing with sets of outgroup (or "right") populations A,B,C is leveraged to infer mixture proportions

Haak et al (2015) Nature

# Estimating admixture proportions - qpAdm



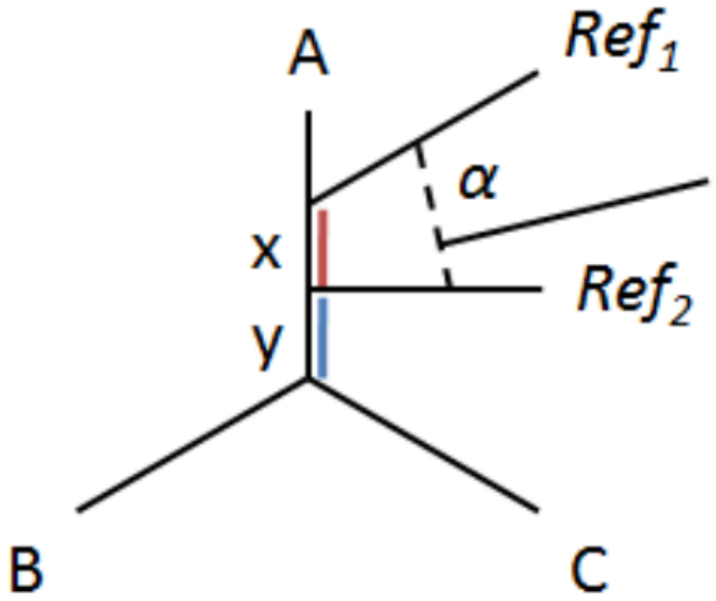

$f_4(Ref_1, A; B, C) = 0$
$f_4(Ref_1, B; A, C) = x+y$
$f_4(Ref_1, C; A, B) = x+y$

$f_4(Ref_2, A; B, C) = 0$
$f_4(Ref_2, B; A, C) = y$
$f_4(Ref_2, C; A, B) = y$

Differential drift sharing with sets of outgroup (or "right") populations A,B,C is leveraged to infer mixture proportions

Haak et al (2015) Nature

# Estimating admixture proportions - qpAdm




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

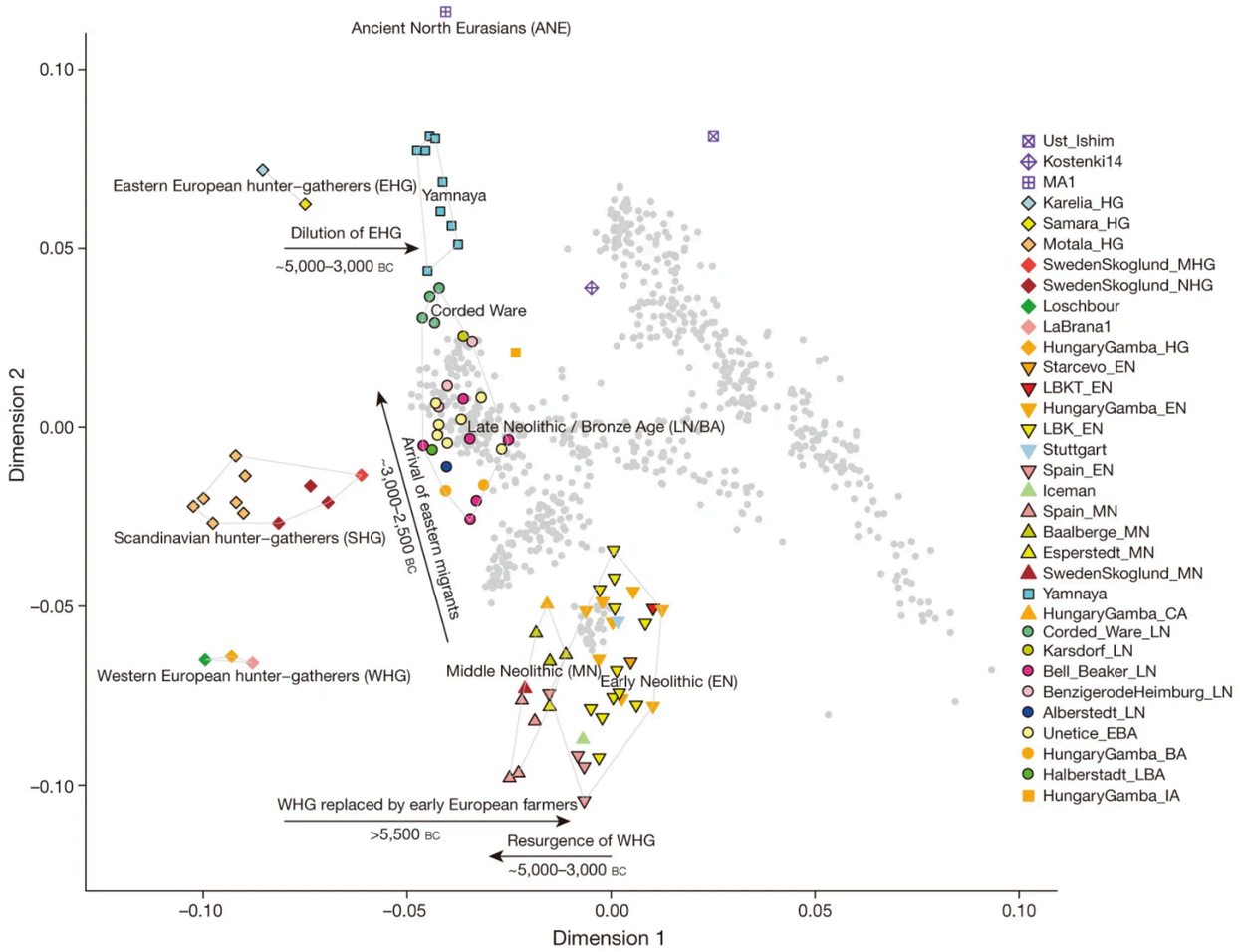
$f_4(Test, A; B, C) = 0$
$f_4(Test, B; A, C) = y+\alpha x=(1-\alpha) f_4(Ref_2, B; A, C)+\alpha f_4(Ref_1, B; A, C)$
$f_4(Test, C; A, B) = y+\alpha x=(1-\alpha) f_4(Ref_2, C; A, B)+\alpha f_4(Ref_1, C; A, B)$

$f_4(Ref_2, A; B, C) = 0$
$f_4(Ref_2, B; A, C) = y$
$f_4(Ref_2, C; A, B) = y$

Differential drift sharing with sets of outgroup (or "right") populations A,B,C is leveraged to infer mixture proportions

Haak et al (2015) Nature

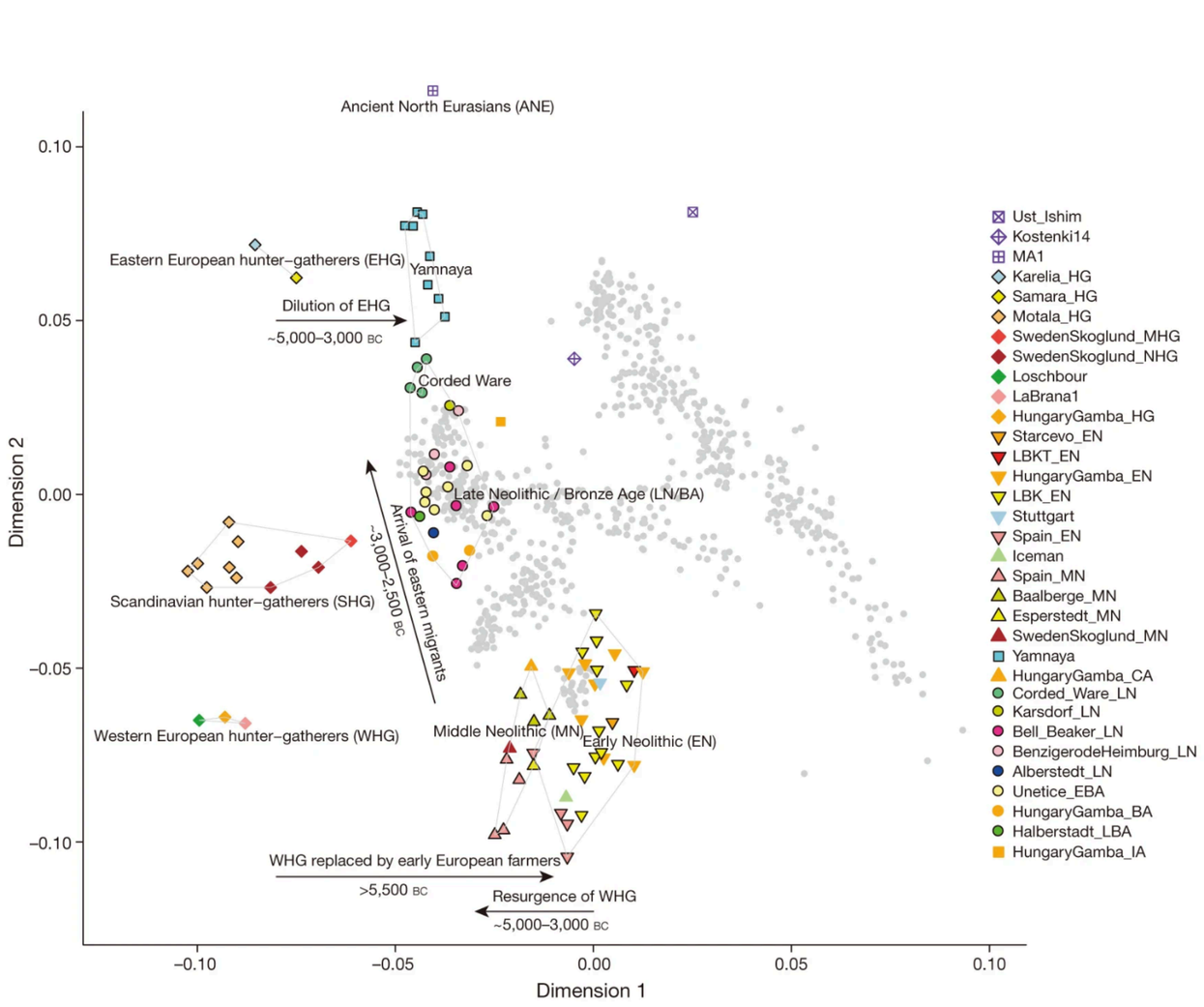
# Estimating admixture proportions - qpAdm

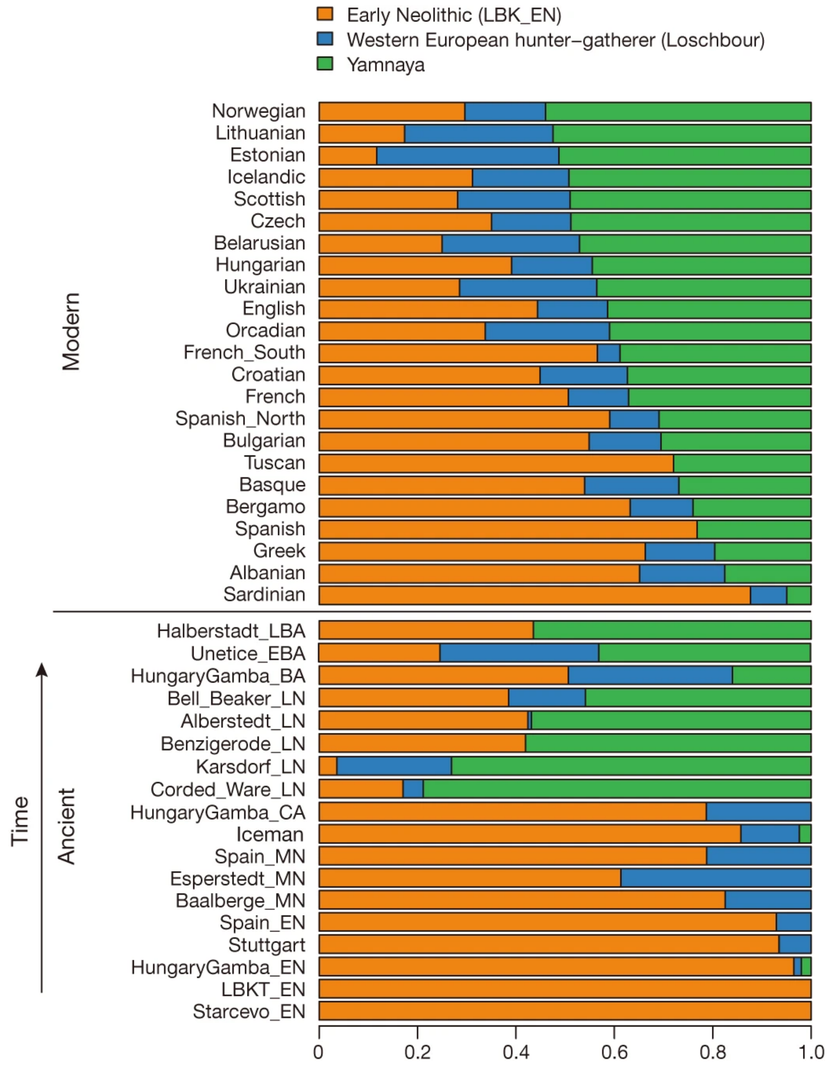



Introduction of “Steppe” ancestry into Europe during the Late Neolithic / Early Bronze Age

Haak et al (2015) Nature

# Estimating admixture proportions - qpAdm




Introduction of "Steppe" ancestry into Europe during the Late Neolithic / Early Bronze Age

Haak et al (2015) Nature