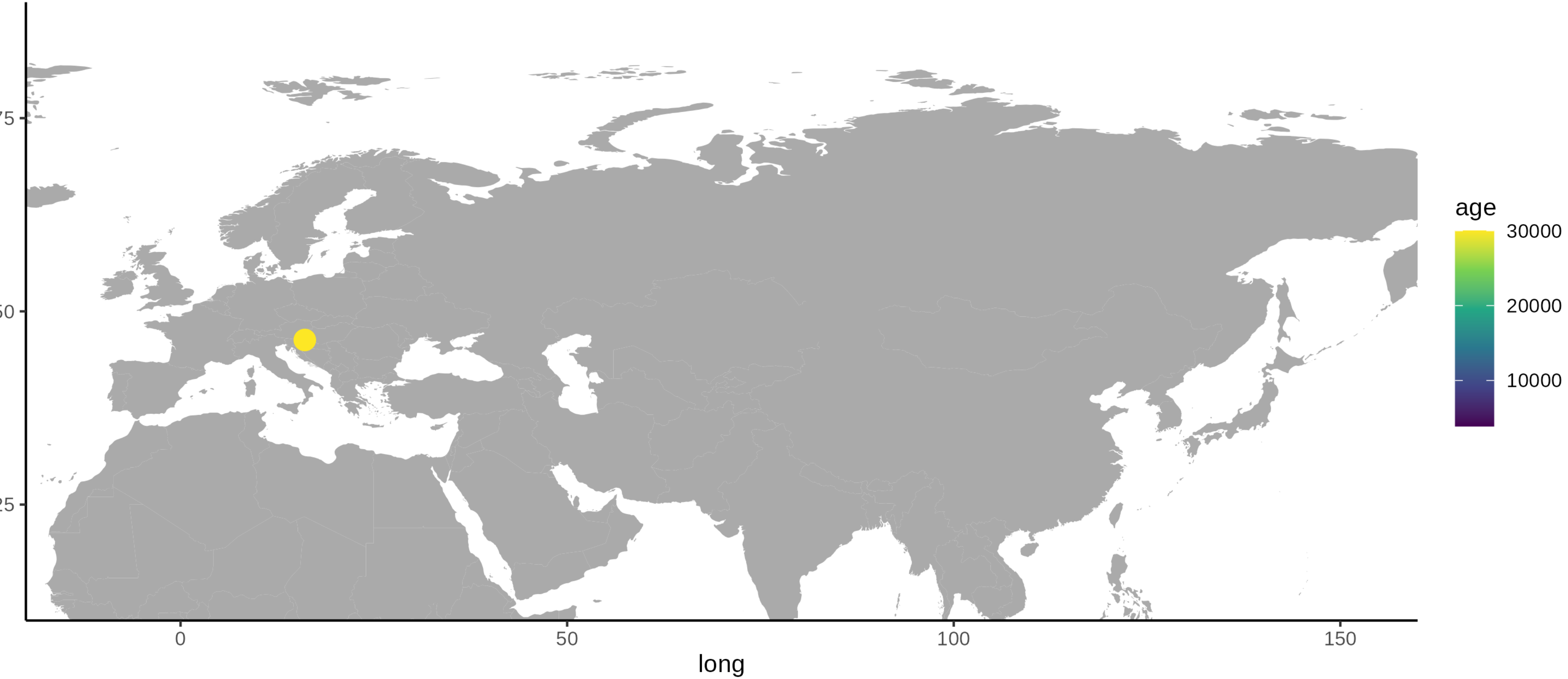
F-statistics and Population Structure

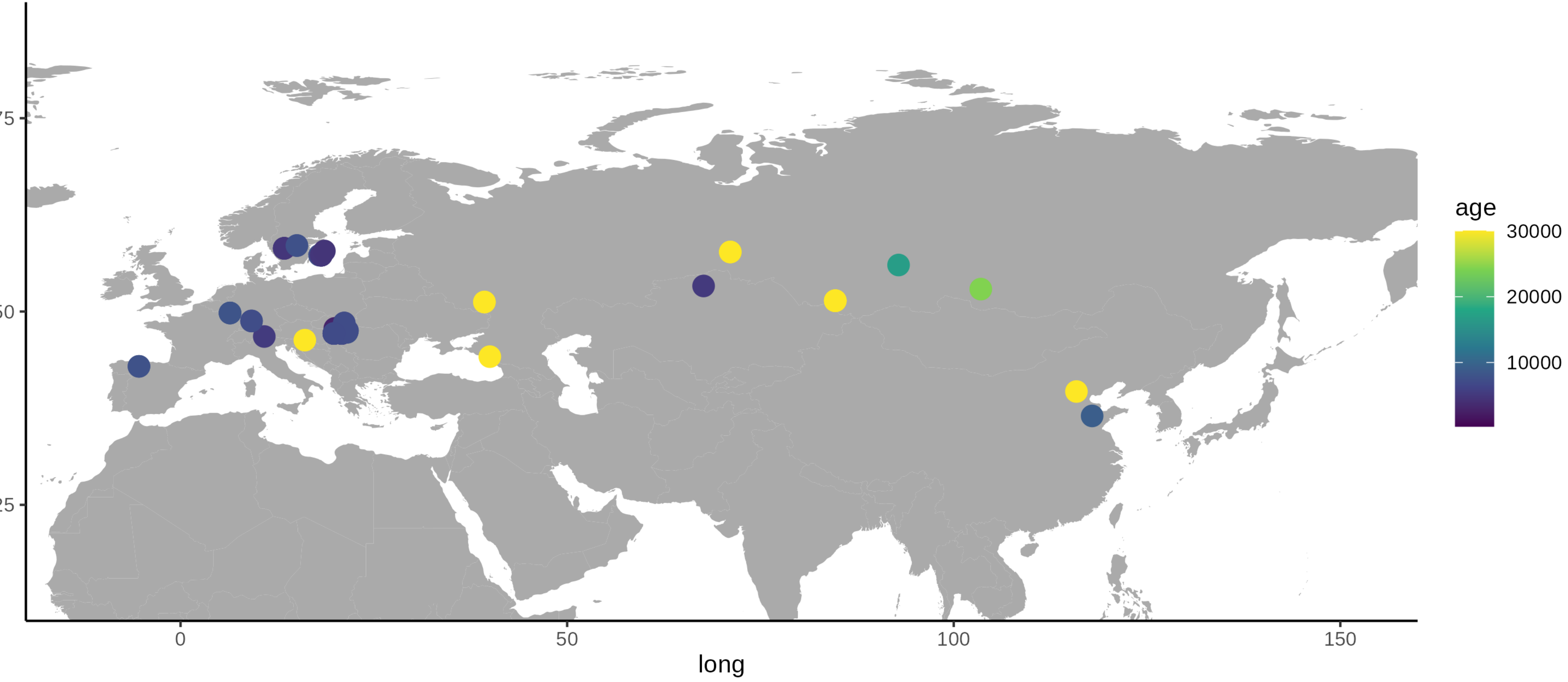
Benjamin Peter, MPI for Evolutionary Anthropology

Hominin Ancient DNA (2010)



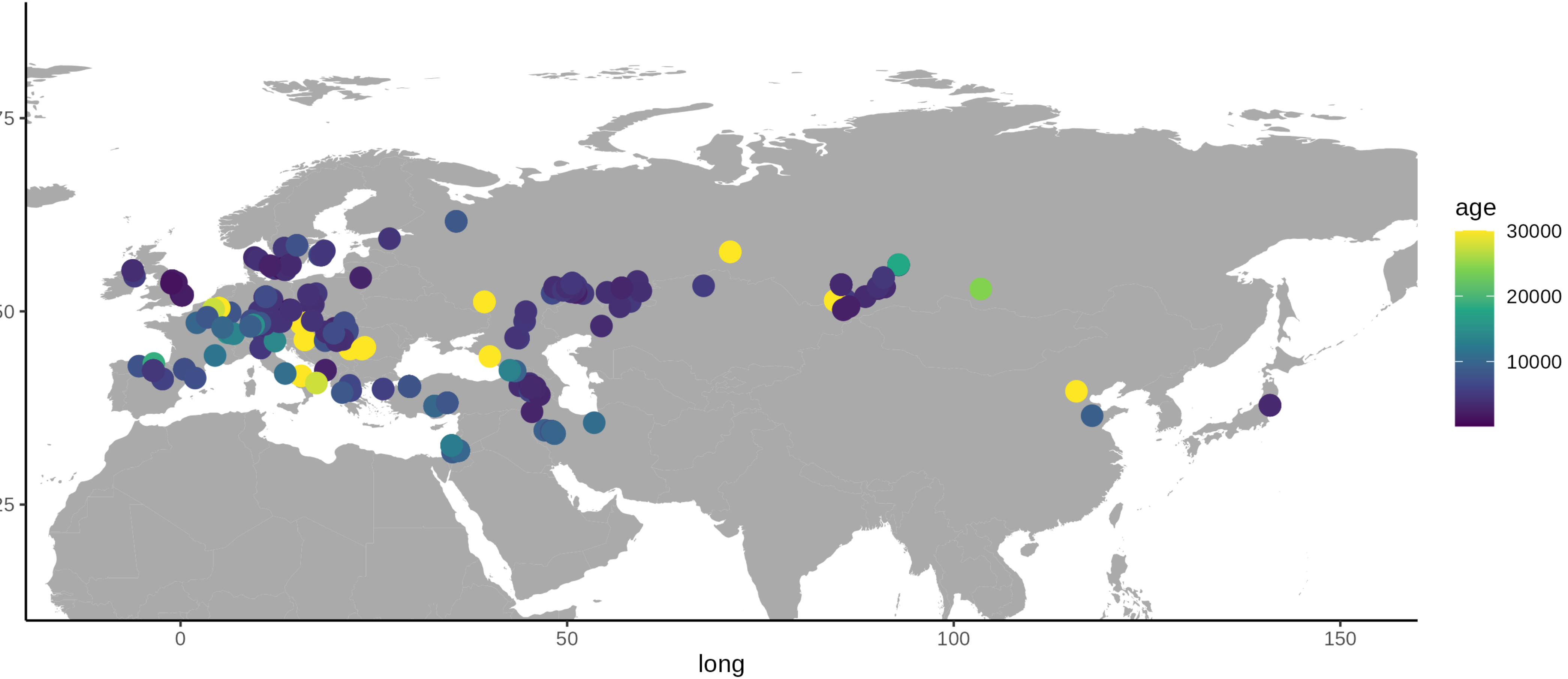
Data compiled by D. Reich et al.

Hominin Ancient DNA (2014)



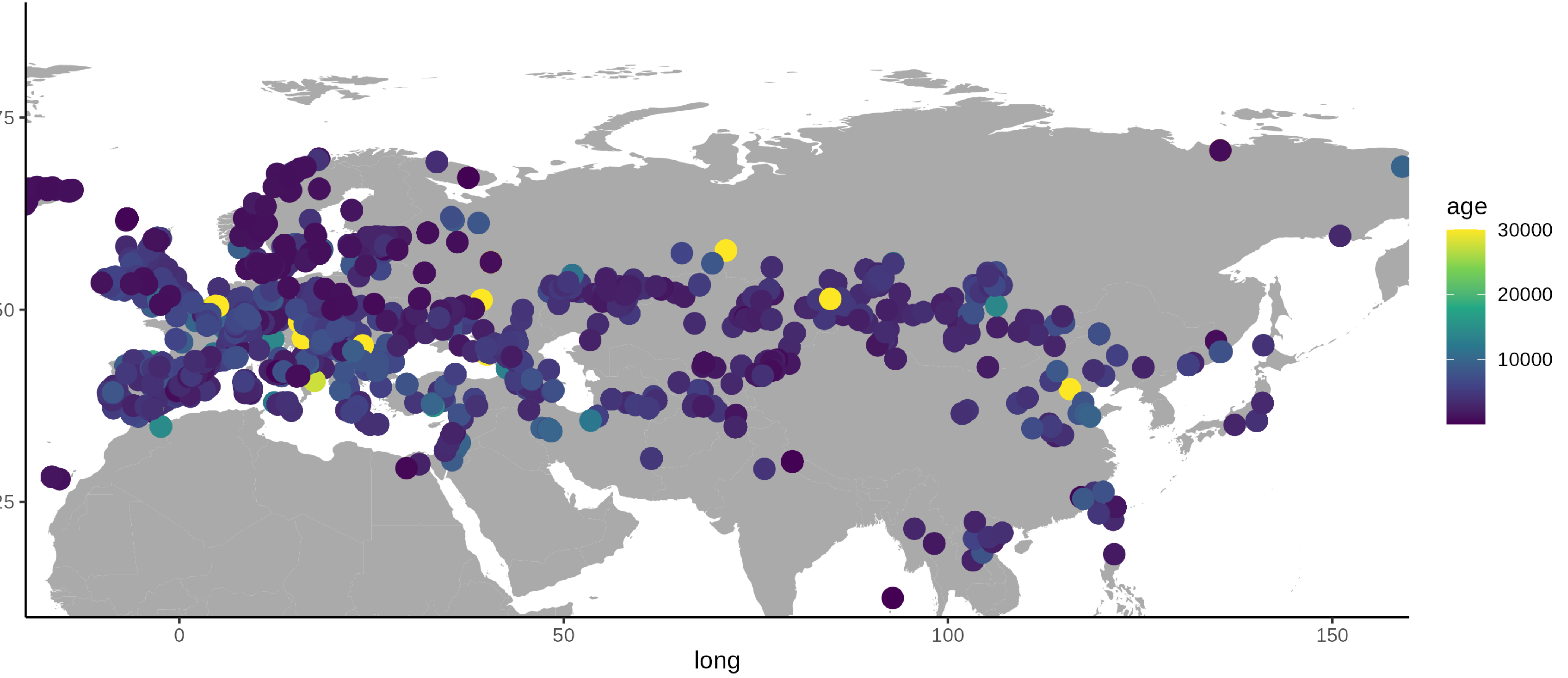
Data compiled by D. Reich et al.

Hominin Ancient DNA (2016)



Data compiled by D. Reich et al.

Hominin Ancient DNA (2020)



Data compiled by D. Reich et al.

Main Reference

GENETICS | INVESTIGATION

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.)

https://doi.org/10.1534/genetics.115.183913

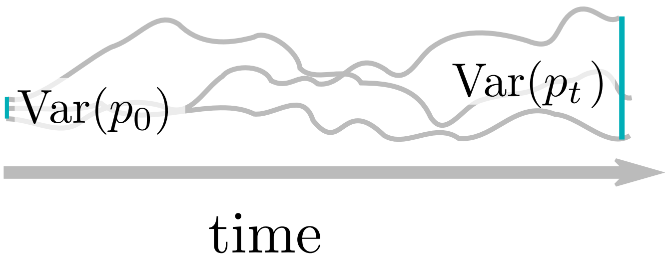
Setup

- Today: Theory of F-statistics and Computations
- Tomorrow: Using F-statistics to build more complex models

Measuring Genetic Drift

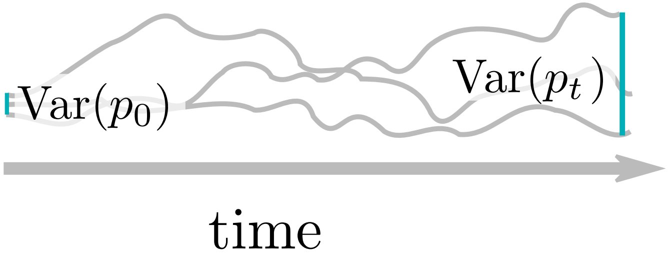
Measuring Genetic Drift

Change in Allele Frequency

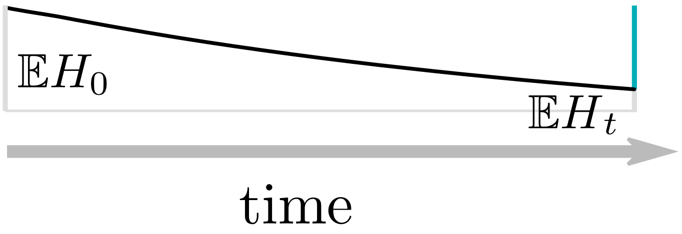


Measuring Genetic Drift

Change in Allele Frequency

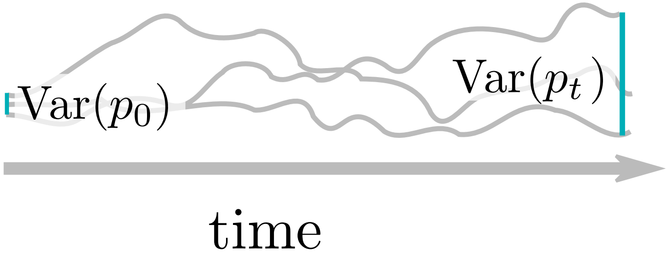


Decay of Heterozygosity

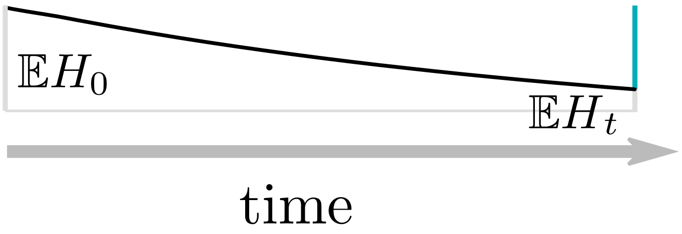


Measuring Genetic Drift

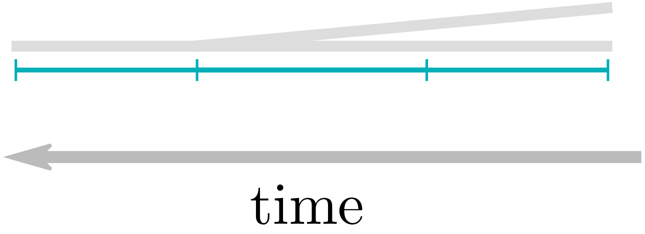
Change in Allele Frequency



Decay of Heterozygosity



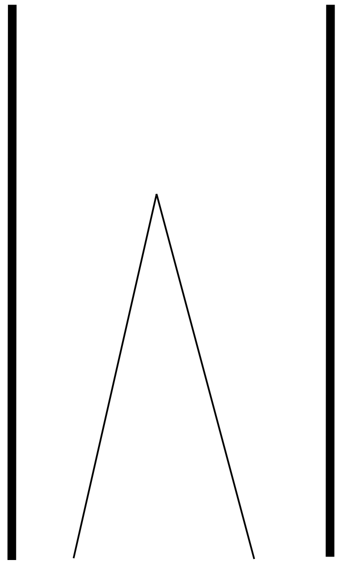
Coalescence rates

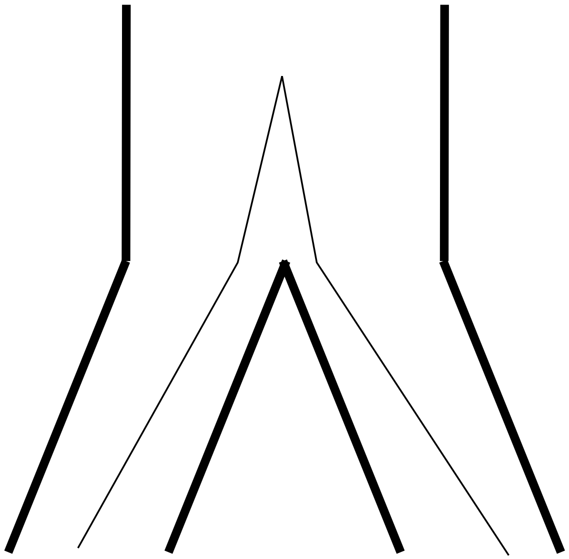


Pairwise differences

$$\mathbb{E}[\pi] = 4N\mu = \theta$$

$$\mathbb{E}[\pi_{12}] = t_{12} + 4N_{anc}\mu = t_{12} + \theta$$



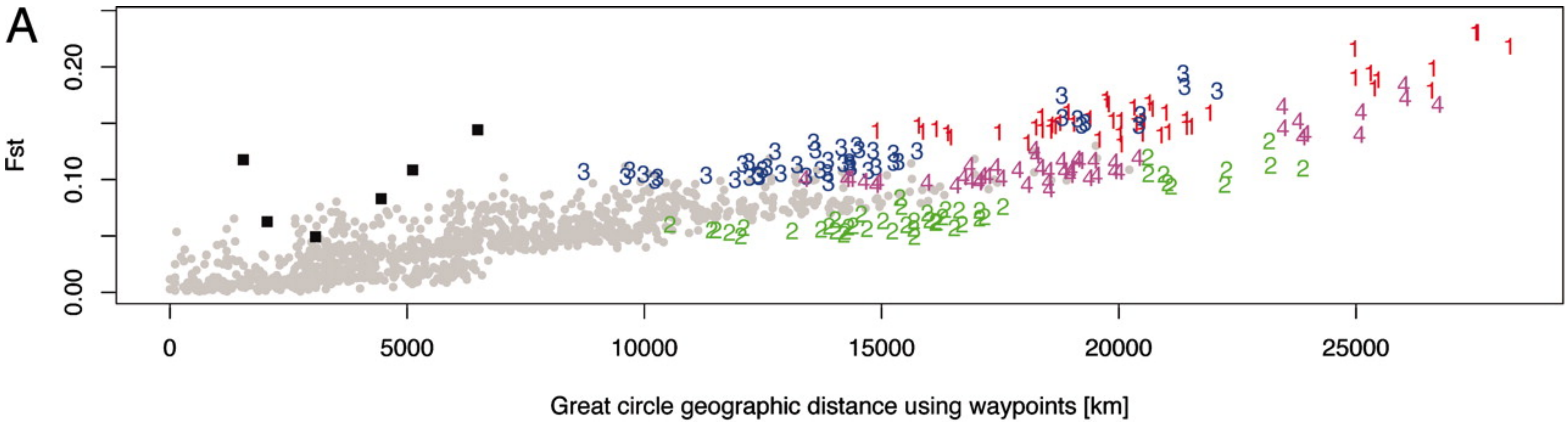


Fixation Index F_{ST}

$$F_{ST}(P_1, P_2) = \frac{\pi_{12} - \frac{\pi_1 + \pi_2}{2}}{\pi_{12}}$$

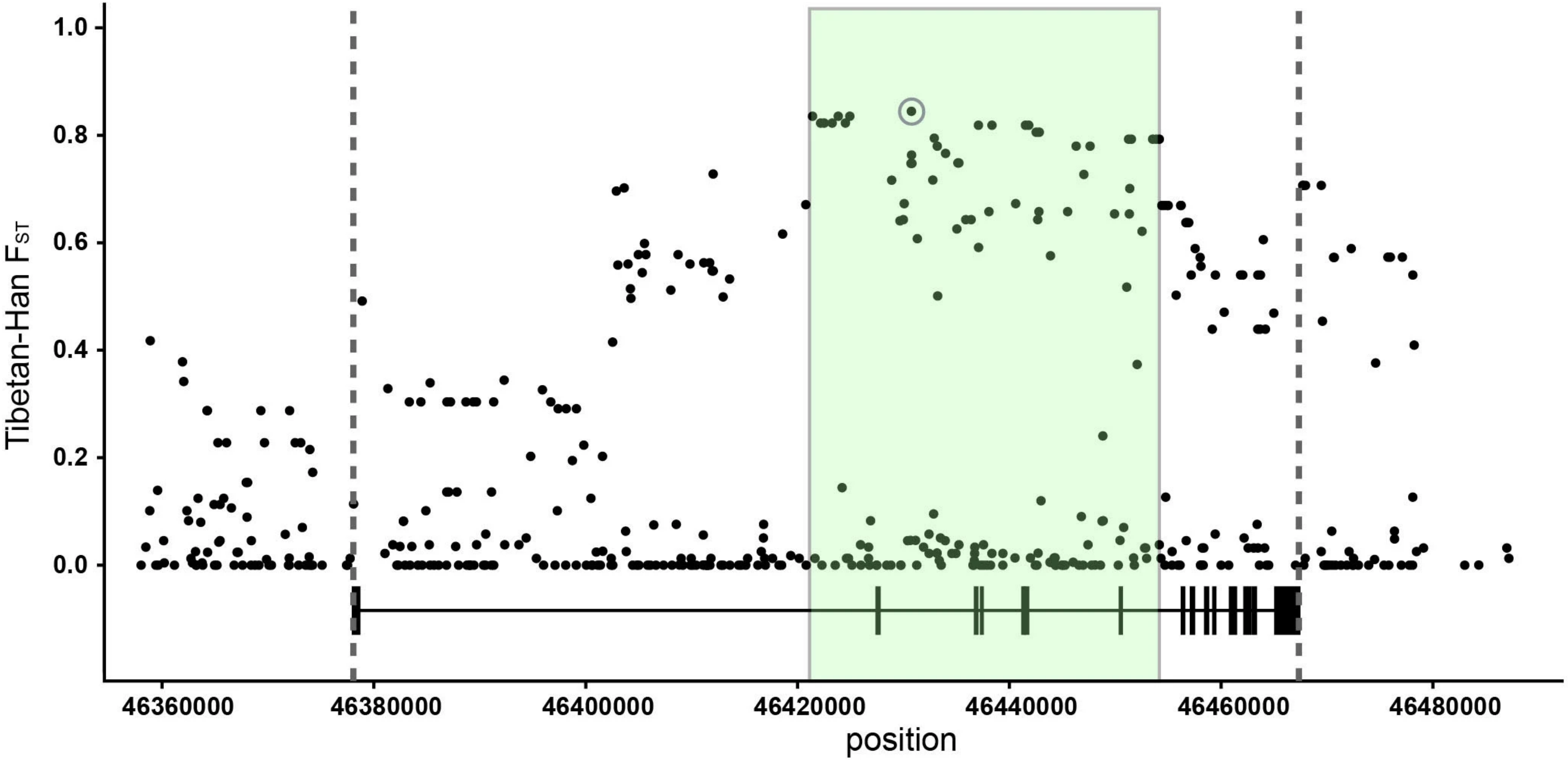
- F_{ST} is a correlation coefficient
- Between 0 and 1
- Hierarchical partitioning (AMOVA)
- Many estimators exist
  - Hudson (1991)
  - Weir & Cockerham (1984)

Fixation Index F_{ST}



Ramachandran et al. 2005

F_{ST} Outliers



Huerta-Sanchez et al. 2014

F_2-statistic

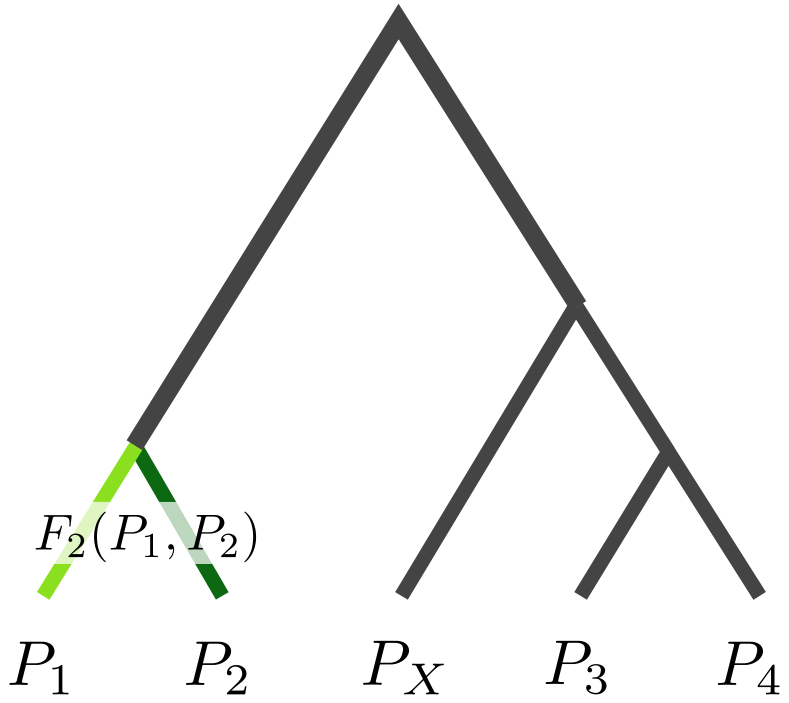
$$F_{ST}(P_1, P_2) = \frac{\pi_{12} - \frac{\pi_1 + \pi_2}{2}}{\pi_{12}}$$

- F_{ST} is a correlation coefficient
- Between 0 and 1
- Hierarchical partitioning (AMOVA)
- Many estimators exist
  - Hudson (1991)
  - Weir & Cockerham (1984)

$$\begin{aligned} F_2(P_1, P_2) &= 2\pi_{12} - \pi_1 - \pi_2 \\ &= \sum_l (p_{1l} - p_{2l})^2 \end{aligned}$$

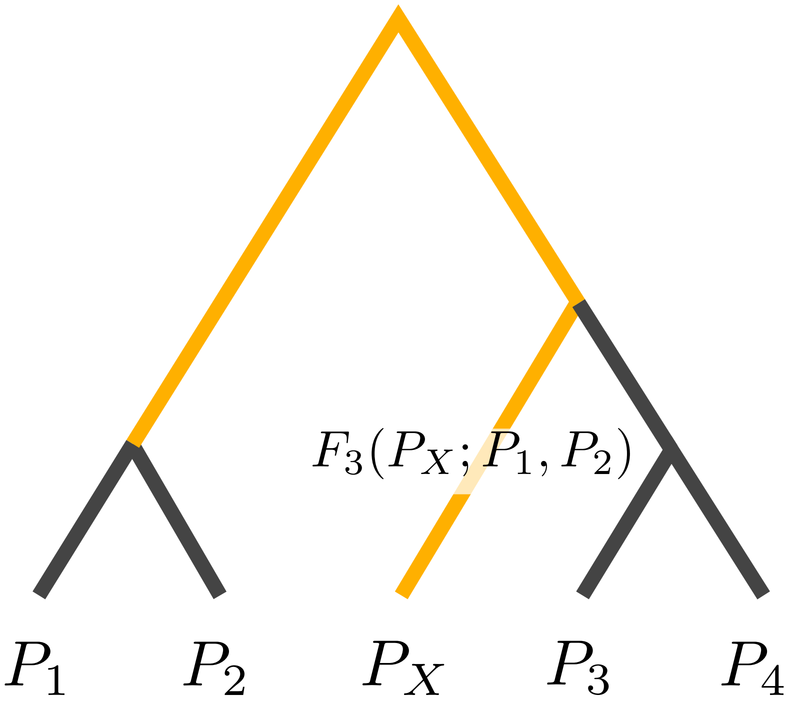
- F_2 is a covariance
- Bigger than 0
- Tree-additive
- Testing for treeness

Tree-additive



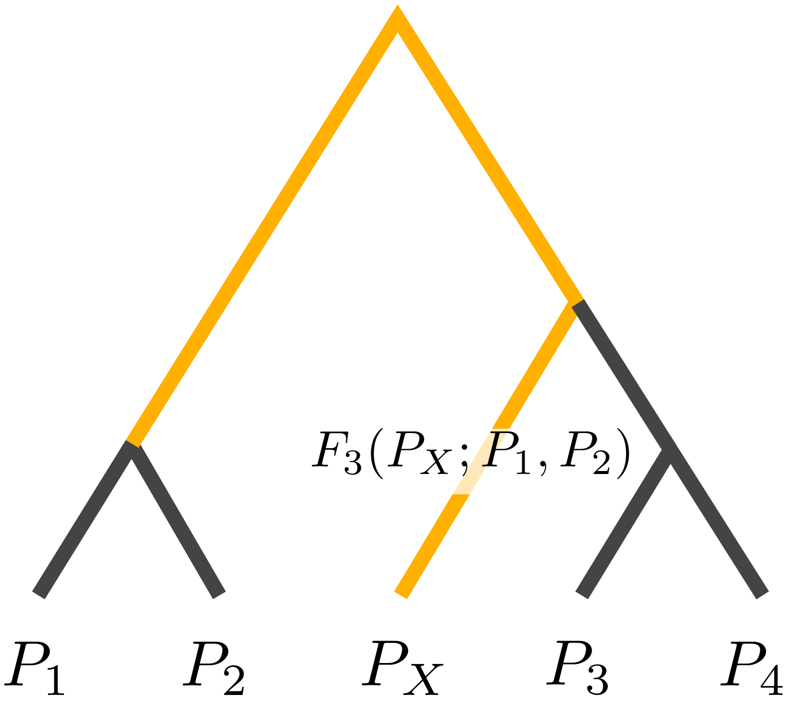
F_3-statistic

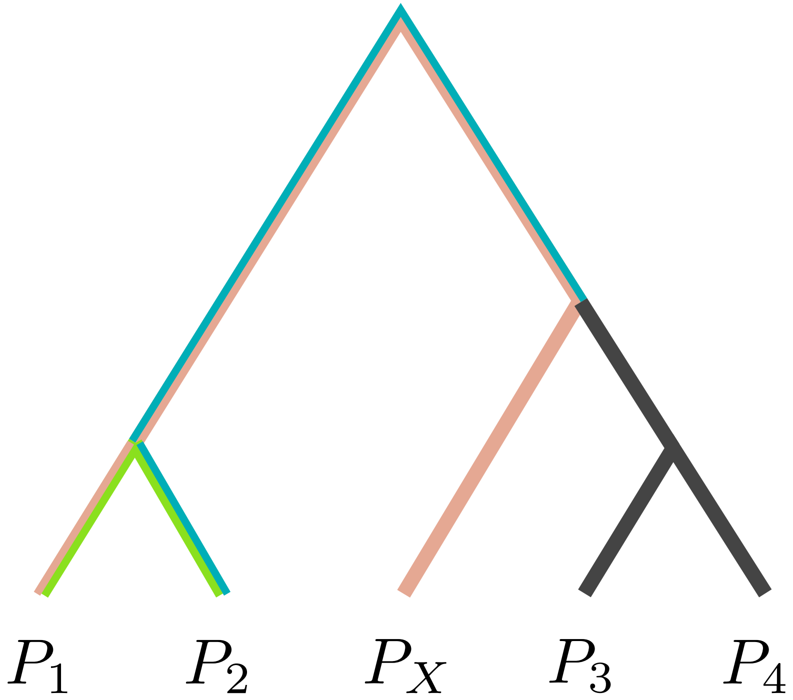
Given all F2-values, how can we calculate the yellow branch length?



F_3-statistic

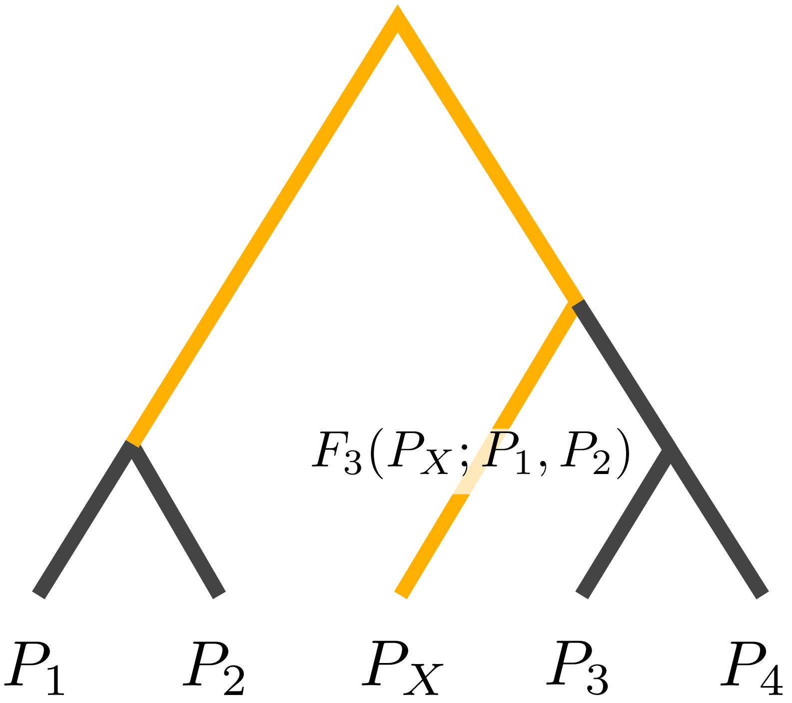
Given all F2-values, how can we calculate the yellow branch length?

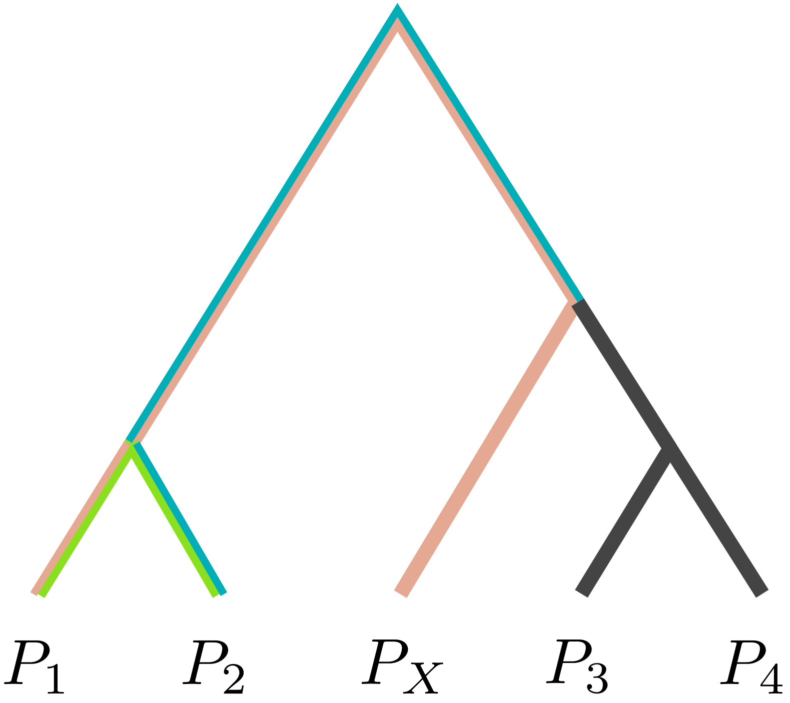




F_3-statistic

$$F_3(P_X; P_1, P_2) = \frac{1}{2}\left(F_2(P_X, P_1) + F_2(P_X, P_2) - F_2(P_1, P_2) \right)$$





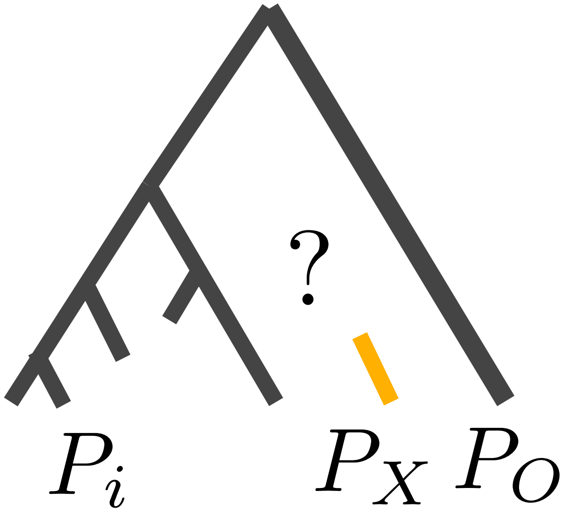
F_3-statistic equations

$$F_3(P_X; P_1, P_2) = \frac{1}{2}\left(F_2(P_X, P_1) + F_2(P_X, P_2) - F_2(P_1, P_2)\right)$$

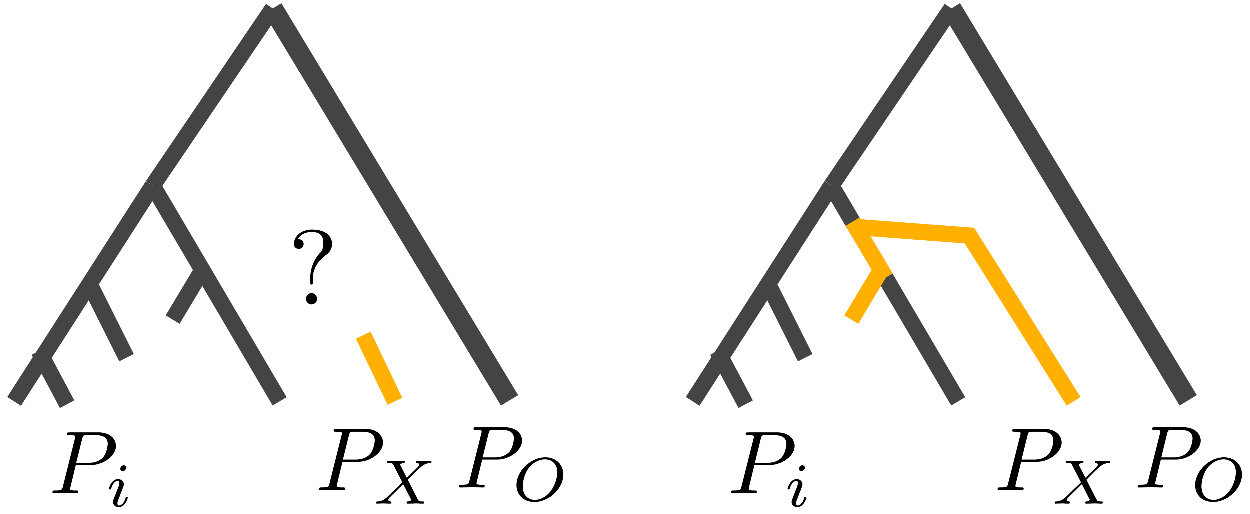
$$F_3(P_X; P_1, P_2) = \sum_l (p_{xl} - p_{x1})(p_{xl} - p_{x2})$$

$$F_3(P_X; P_1, P_2) = \pi_{1x} + \pi_{2x} - \pi_{12} - \pi_x$$

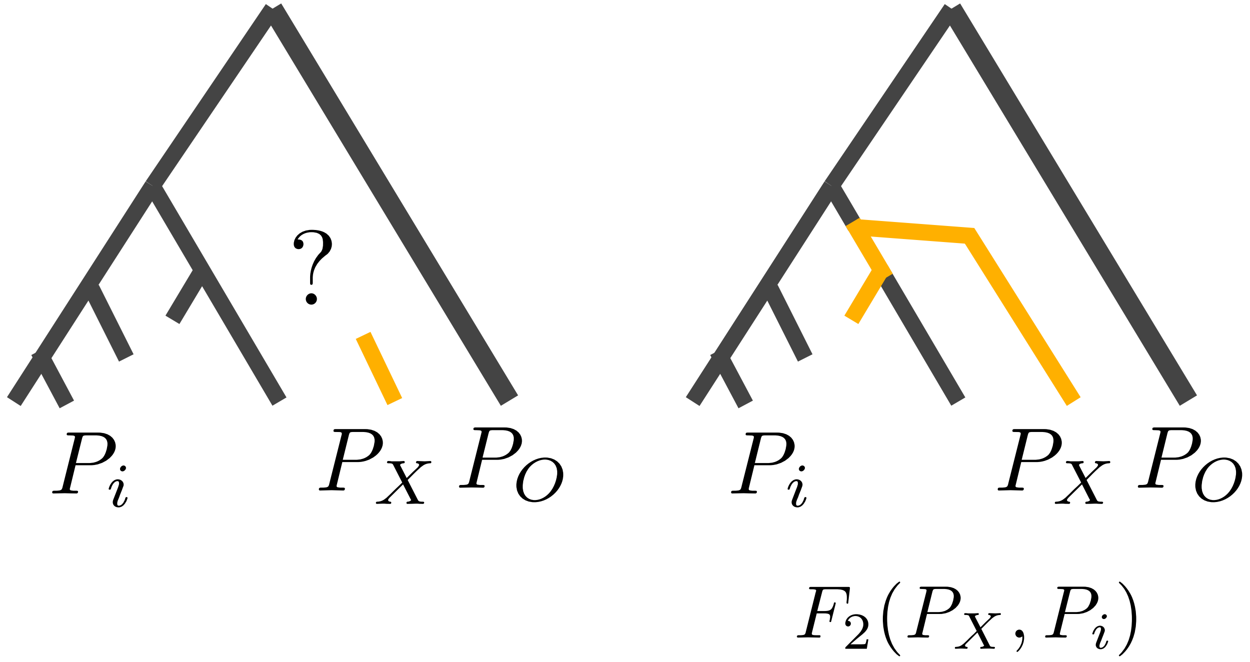
Outgroup-F_3-statistic



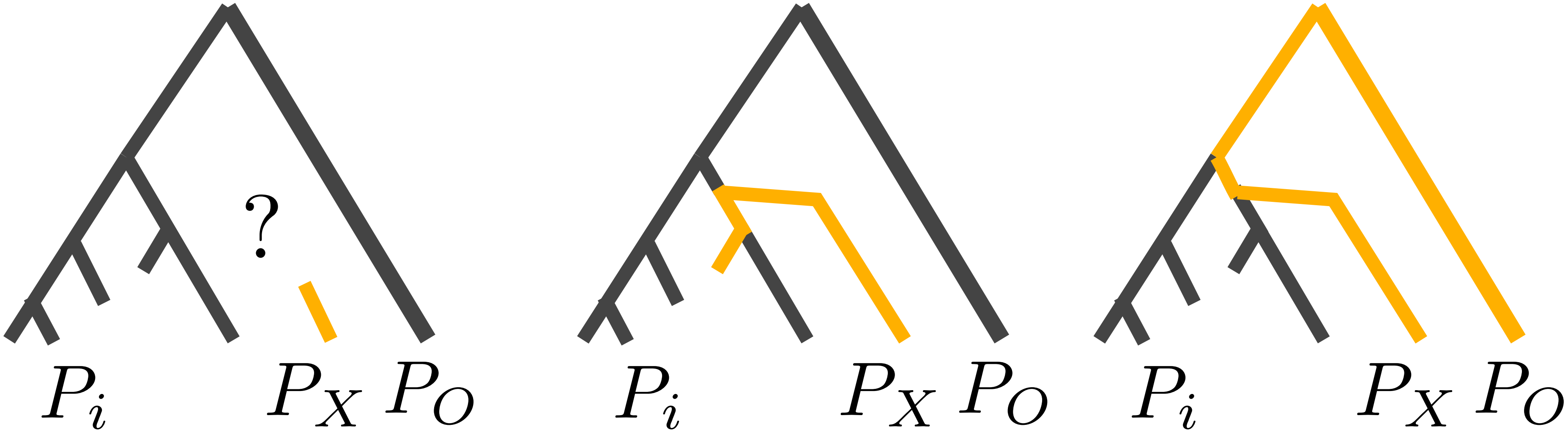
Outgroup-F_3-statistic



Outgroup-F_3-statistic



Outgroup-F_3-statistic

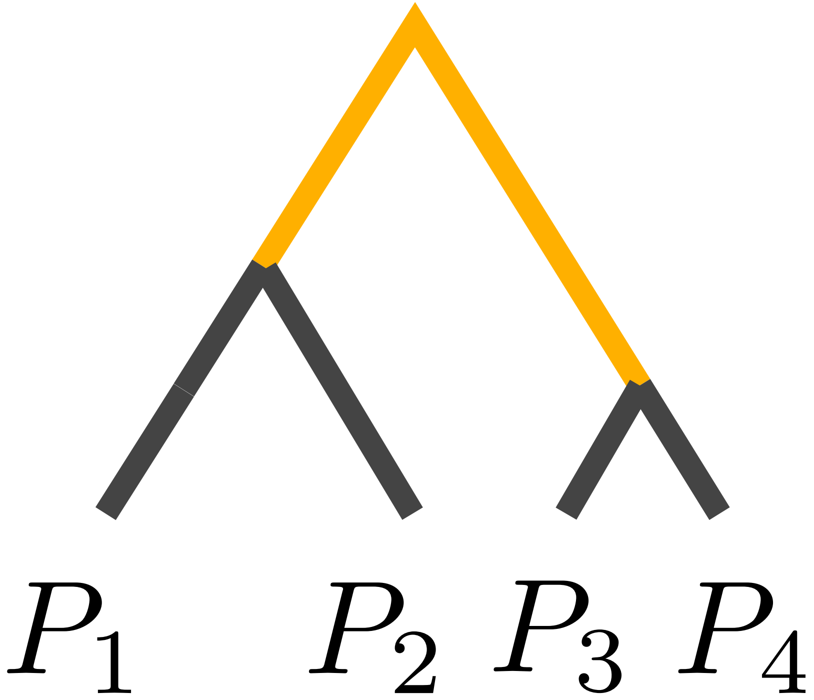


$F_2(P_X, P_i)$

$F_3(P_O; P_X, P_i)$

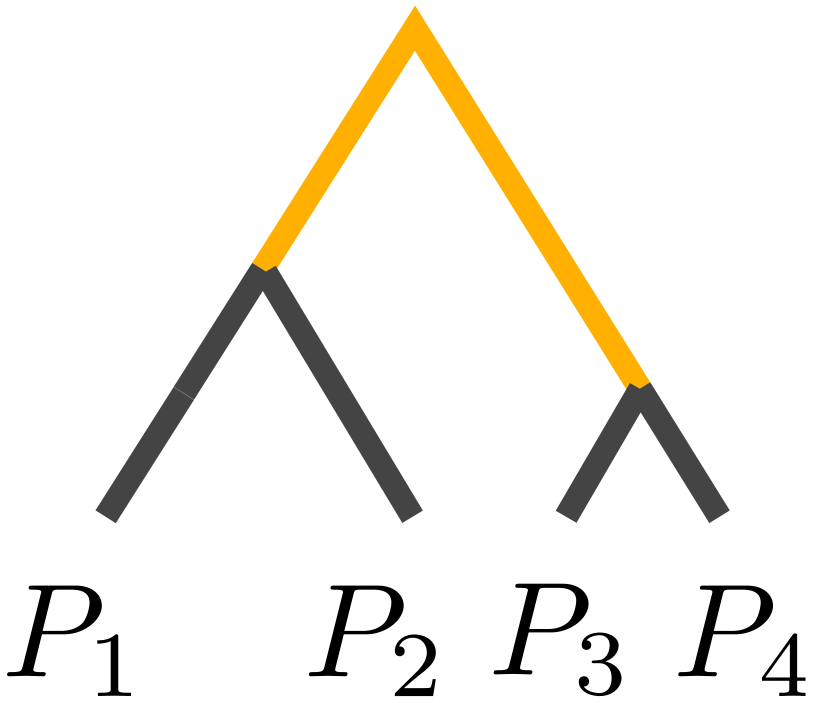
(Branch)-F_4-statistic

$$F_4^{(B)}(P_1, P_2; P_3, P_4) = \frac{1}{2}\Big(F_2(P_1, P_3) + F_2(P_2, P_4) - F_2(P_1, P_2) - F_2(P_3, P_4)\Big)$$



(Branch)-F_4-statistic

$$F_4^{(B)}(P_1, P_2; P_3, P_4) = \frac{1}{2}\Big(F_2(P_1, P_3) + F_2(P_2, P_4) - F_2(P_1, P_2) - F_2(P_3, P_4)\Big)$$

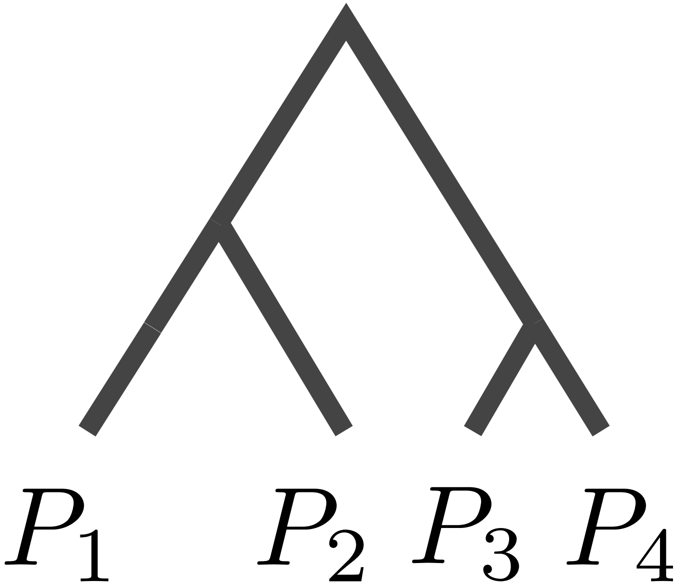


What if we reorder the arguments?

$$F_4^{(T)}(P_1, P_2; P_3, P_4) = F_4^{(B)}(P_1, P_4; P_3, P_2)$$

(Treeness)-F_4-statistic

$$F_4^{(T)}(P_1, P_2; P_3, P_4) = \frac{1}{2}\Big(F_2(P_1, P_3) + F_2(P_2, P_4) - F_2(P_1, P_4) - F_2(P_2, P_3)\Big)$$



F_4 -statistic-equations

$$F_4^{(T)}(P_1, P_2; P_3, P_4) = \frac{1}{2}\Big(F_2(P_1, P_3) + F_2(P_2, P_4) - F_2(P_1, P_4) - F_2(P_2, P_3)\Big)$$

$$F_4^{(T)}(P_1, P_2; P_3, P_4) = \sum_l (p_{l1} - p_{l2})(p_{l3} - p_{l4})$$

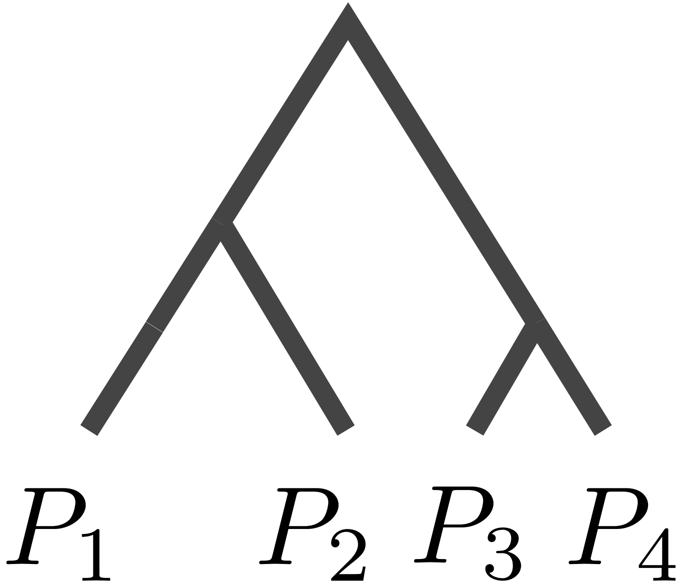
$$F_4^{(T)}(P_1, P_2; P_3, P_4) = \pi_{13} + \pi_{24} - \pi_{14} - \pi_{23}$$

Testing Treeness

If data is generated from a tree:

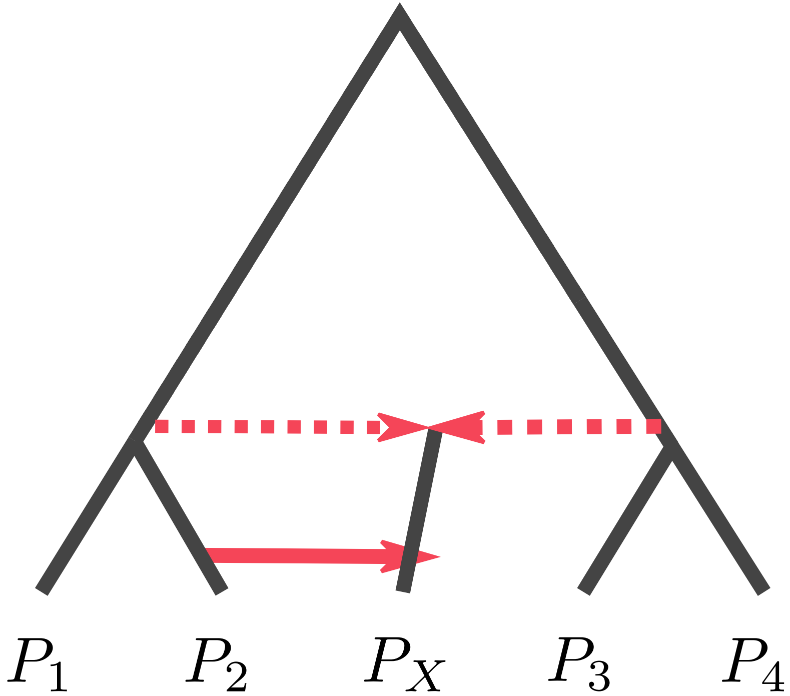
$$F_3(P_3; P_1; P_2) \geq 0$$

$$F_4^{(T)}(P_1, P_2; P_3, P_4) = 0$$



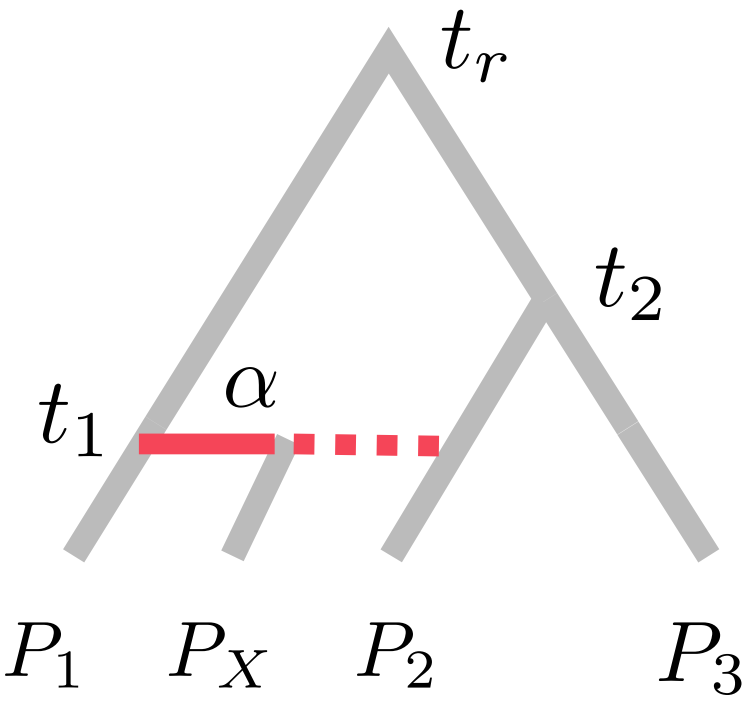
Buneman 1974

Admixture Graphs



F3 in an admixture graph

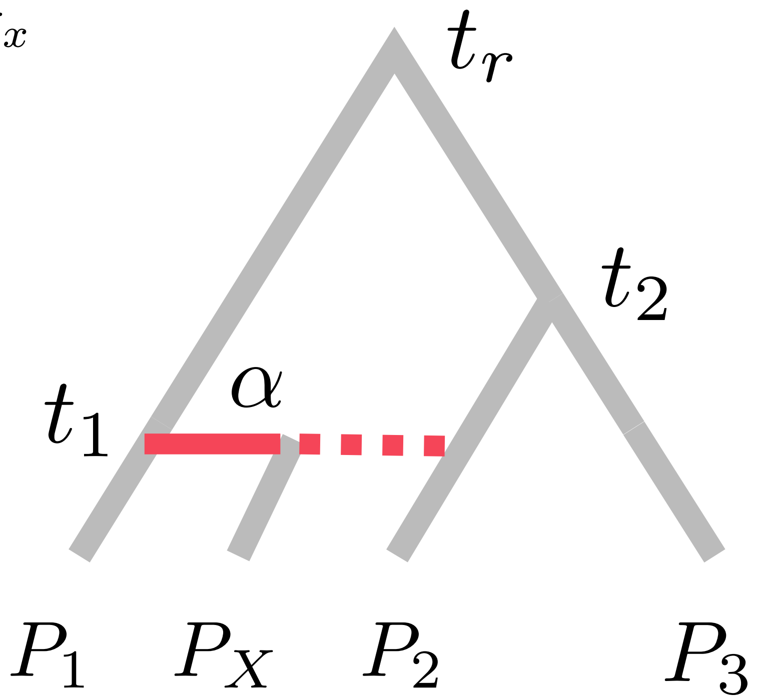
$$F_3(P_X; P_1, P_2) \approx \theta\left[t_1 - 2\alpha(1-\alpha)t_r\right]$$



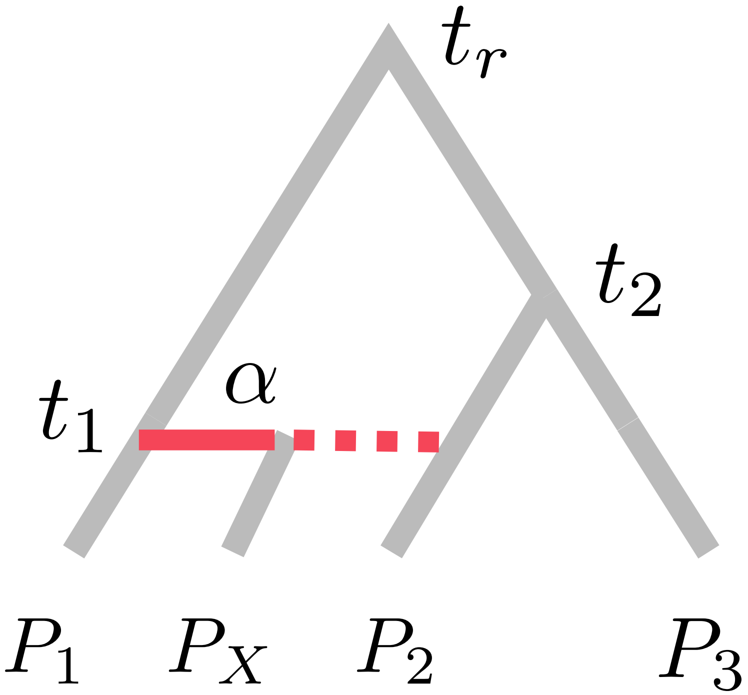
F3 in an admixture graph

$$F_3(P_X; P_1, P_2) = \pi_{1x} + \pi_{2x} - \pi_{12} - \pi_x$$

$$F_3(P_X; P_1, P_2) \approx \theta\left[t_1 - 2\alpha(1-\alpha)t_r\right]$$

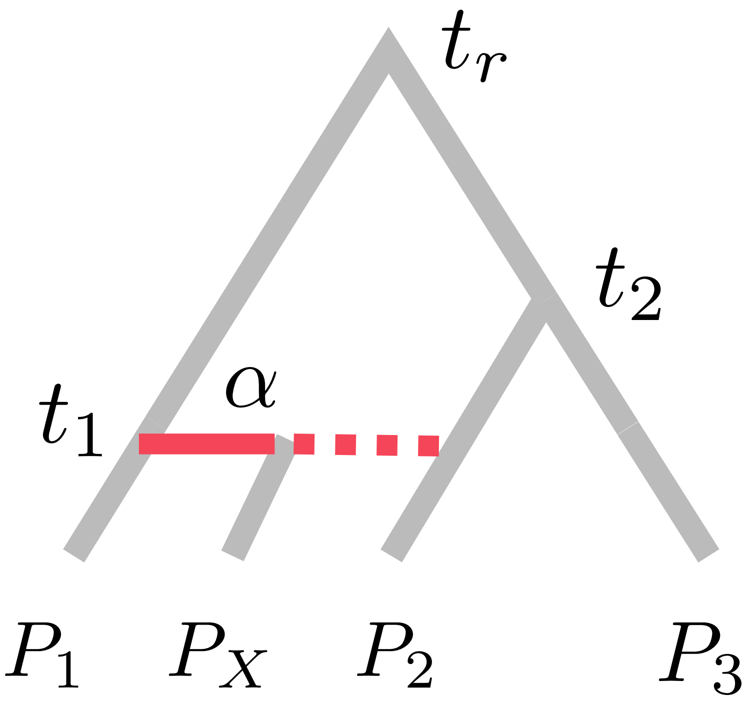


F4 in an admixture graph



F4 in an admixture graph

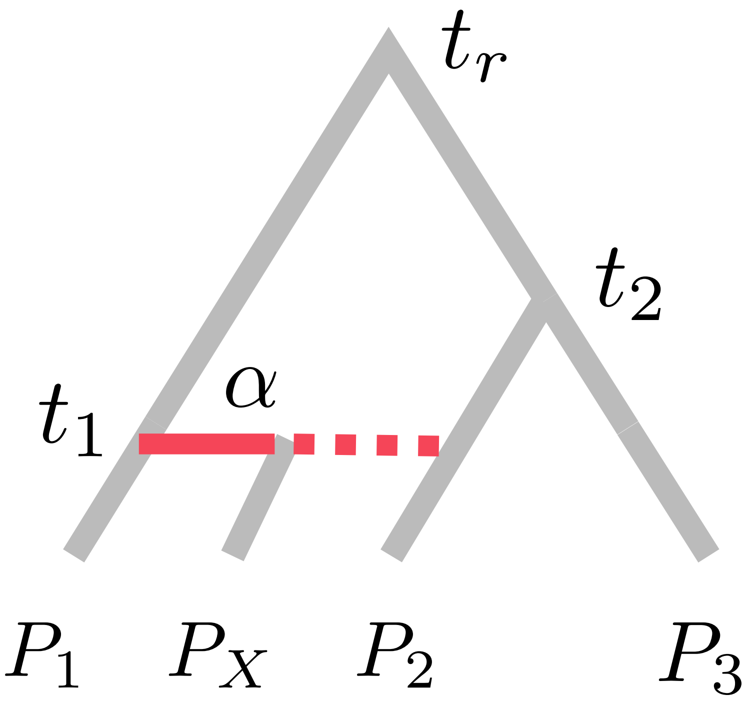
$$F_4^{(T)}(P_1, P_X; P_2, P_3) = \pi_{12} + \pi_{3x} - \pi_{13} - \pi_{2x}$$



F4 in an admixture graph

$$F_4^{(T)}(P_1, P_X; P_2, P_3) = \pi_{12} + \pi_{3x} - \pi_{13} - \pi_{2x}$$

$$F_4^{(T)}(P_1, P_X; P_2, P_3) = (1-\alpha)(t_2 - t_1) \neq 0$$

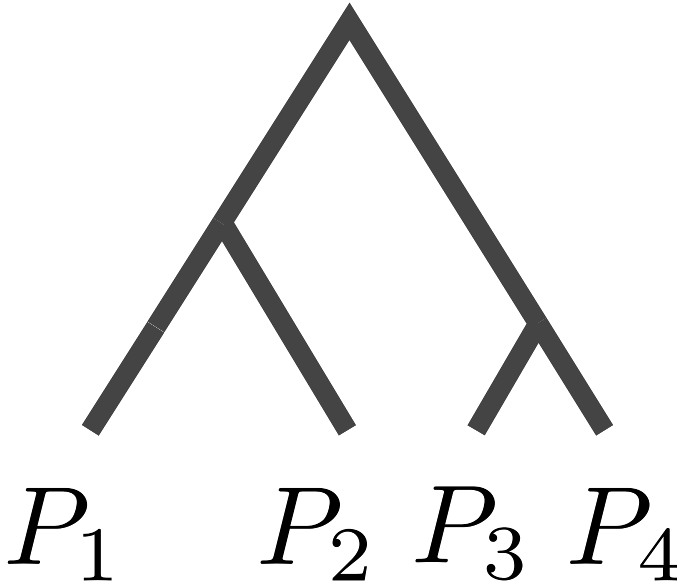


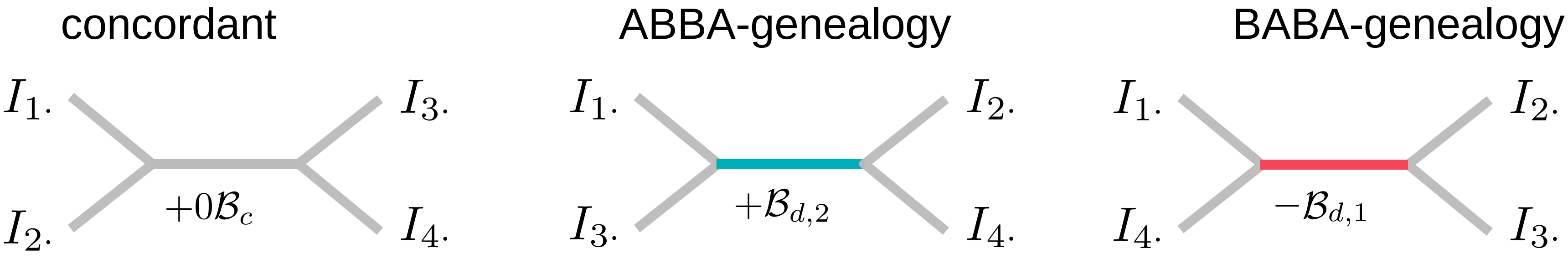
D-statistic

$$D = \frac{\mathrm{ABBA} - \mathrm{BABA}}{\mathrm{BABA} + \mathrm{ABBA}}$$

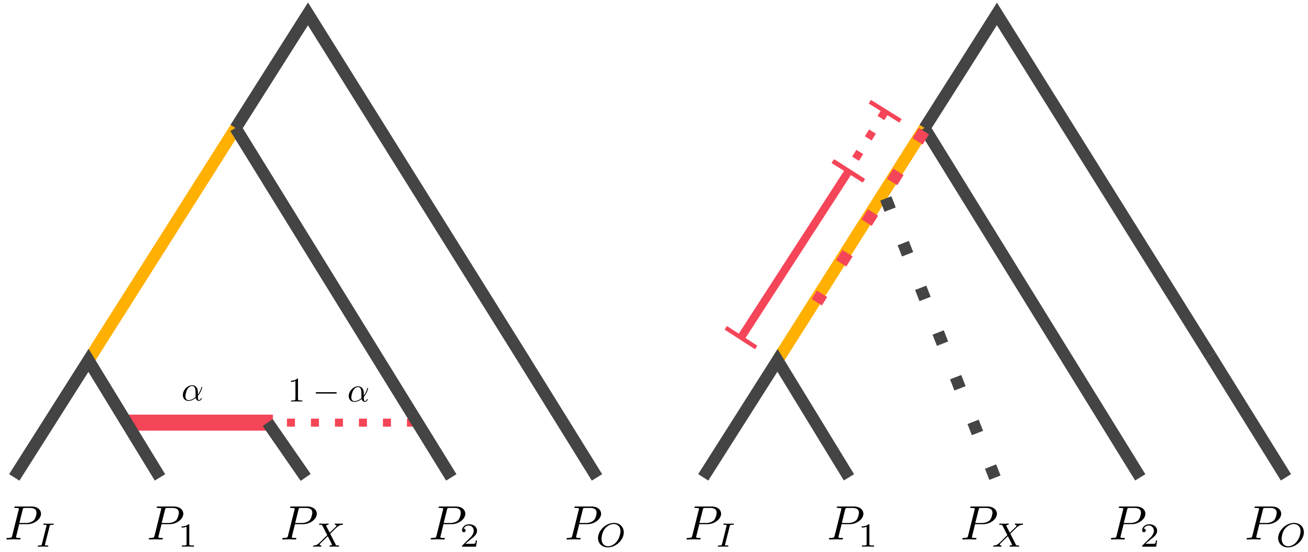
- D-statistic and F4 are closely related

$$F_4^{(T)}(P_1, P_2; P_3, P_4) = \pi_{13} + \pi_{24} - \pi_{14} - \pi_{23}$$



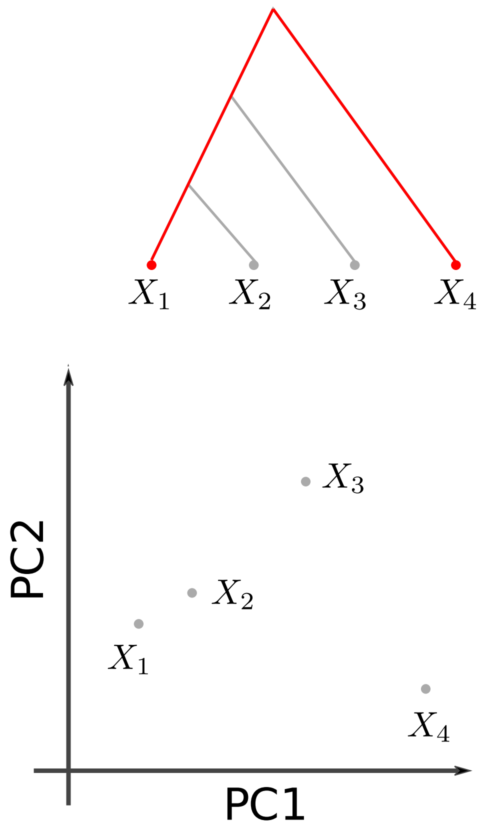


F4-ratio



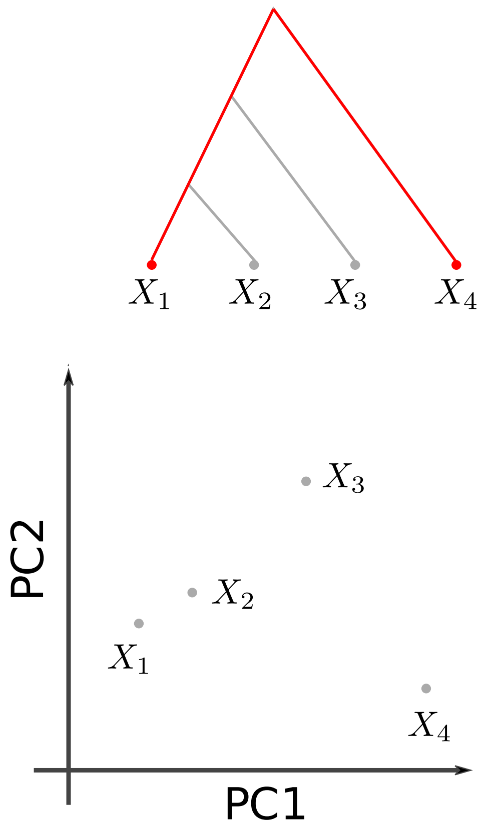
$$\alpha = 1 - \frac{F_4^{(B)}(P_I, P_1; P_X, P_O)}{F_4^{(B)}(P_I, P_1; P_2, P_O)}$$

A $F_2(X_1; X_4)$



Peter (2022): doi 10.1098/rstb.2020.0413

A $F_2(X_1; X_4)$



PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Check for updates

Cite this article: Peter BM. 2022 A geometric relationship of F_2, F_3 and F_4-statistics with principal component analysis. *Phil. Trans. R. Soc. B* **377**: 20200413. https://doi.org/10.1098/rstb.2020.0413

Received: 7 July 2021

Accepted: 12 February 2022

One contribution of 15 to a theme issue 'Celebrating 50 years since Lewontin's

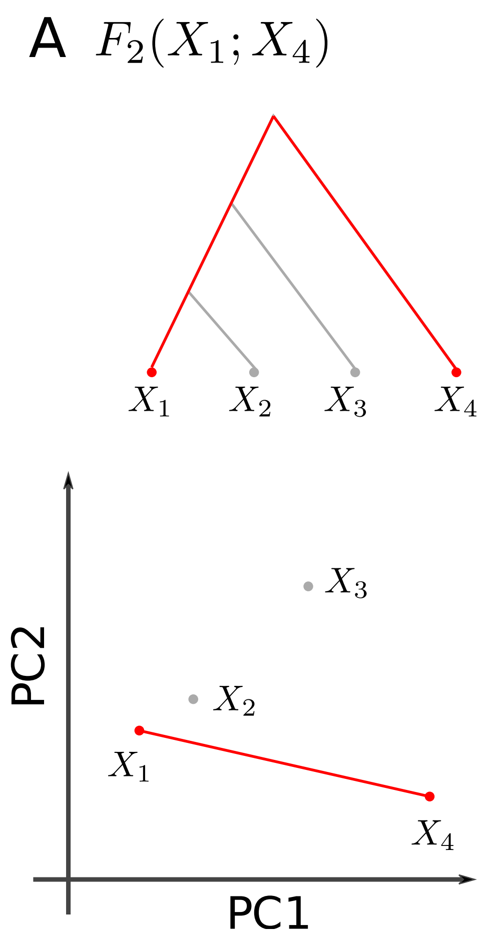
A geometric relationship of F_2, F_3 and F_4-statistics with principal component analysis

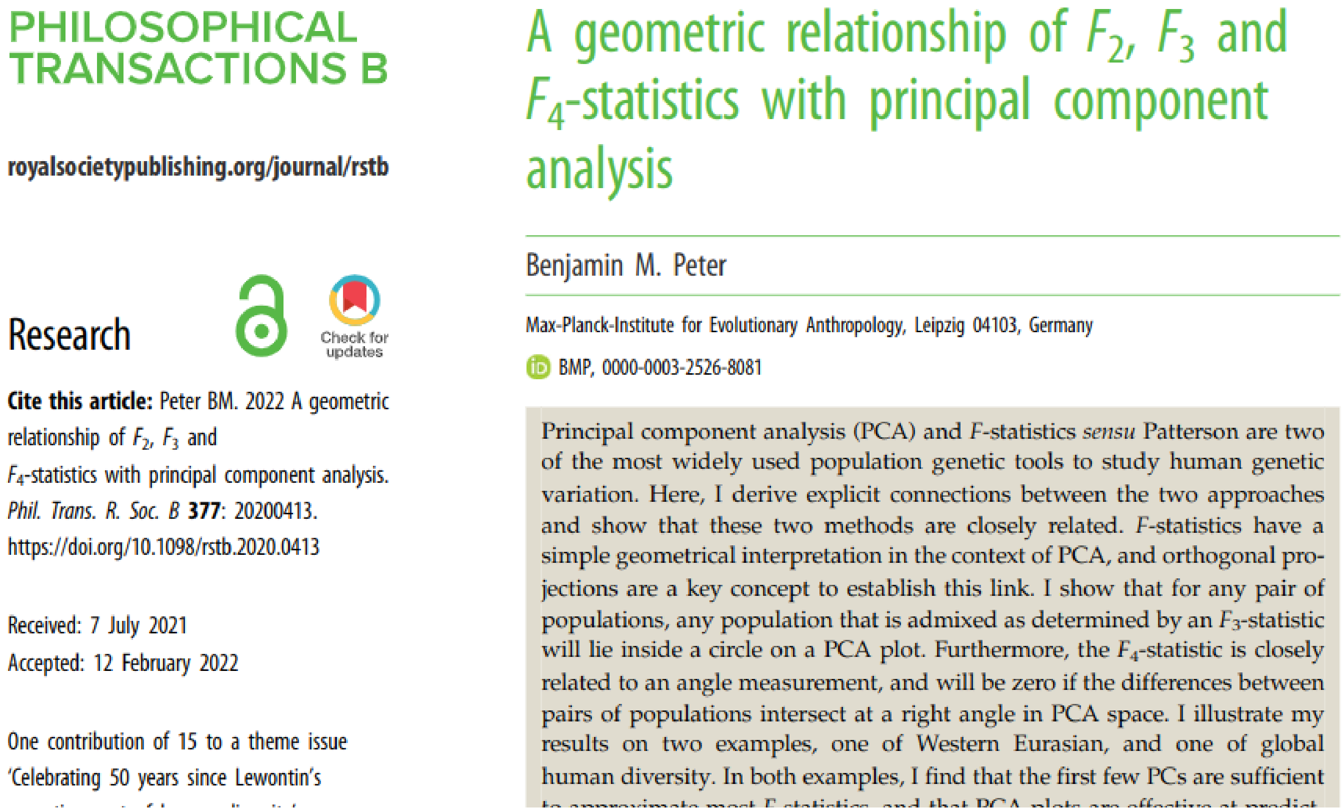
Benjamin M. Peter

Max-Planck-Institute for Evolutionary Anthropology, Leipzig 04103, Germany

BMP, 0000-0003-2526-8081

Principal component analysis (PCA) and *F*-statistics *sensu* Patterson are two of the most widely used population genetic tools to study human genetic variation. Here, I derive explicit connections between the two approaches and show that these two methods are closely related. *F*-statistics have a simple geometrical interpretation in the context of PCA, and orthogonal projections are a key concept to establish this link. I show that for any pair of populations, any population that is admixed as determined by an F_3-statistic will lie inside a circle on a PCA plot. Furthermore, the F_4-statistic is closely related to an angle measurement, and will be zero if the differences between pairs of populations intersect at a right angle in PCA space. I illustrate my results on two examples, one of Western Eurasian, and one of global human diversity. In both examples, I find that the first few PCs are sufficient




PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Check for updates

Cite this article: Peter BM. 2022 A geometric relationship of F_2, F_3 and F_4-statistics with principal component analysis. *Phil. Trans. R. Soc. B* **377**: 20200413. https://doi.org/10.1098/rstb.2020.0413

Received: 7 July 2021
Accepted: 12 February 2022

One contribution of 15 to a theme issue 'Celebrating 50 years since Lewontin's

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

Benjamin M. Peter

Max-Planck-Institute for Evolutionary Anthropology, Leipzig 04103, Germany

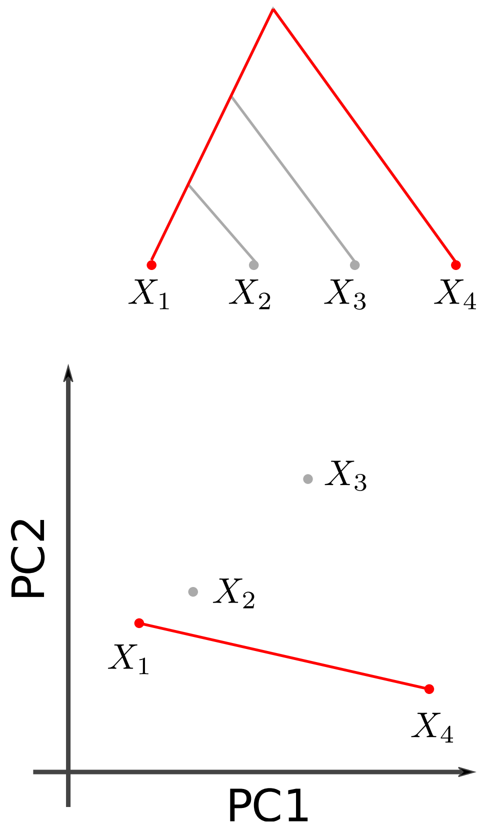
BMP, 0000-0003-2526-8081

Principal component analysis (PCA) and *F*-statistics *sensu* Patterson are two of the most widely used population genetic tools to study human genetic variation. Here, I derive explicit connections between the two approaches and show that these two methods are closely related. *F*-statistics have a simple geometrical interpretation in the context of PCA, and orthogonal projections are a key concept to establish this link. I show that for any pair of populations, any population that is admixed as determined by an F_3-statistic will lie inside a circle on a PCA plot. Furthermore, the F_4-statistic is closely related to an angle measurement, and will be zero if the differences between pairs of populations intersect at a right angle in PCA space. I illustrate my results on two examples, one of Western Eurasian, and one of global human diversity. In both examples, I find that the first few PCs are sufficient

* true if all PCs are used, optimal approximation otherwise

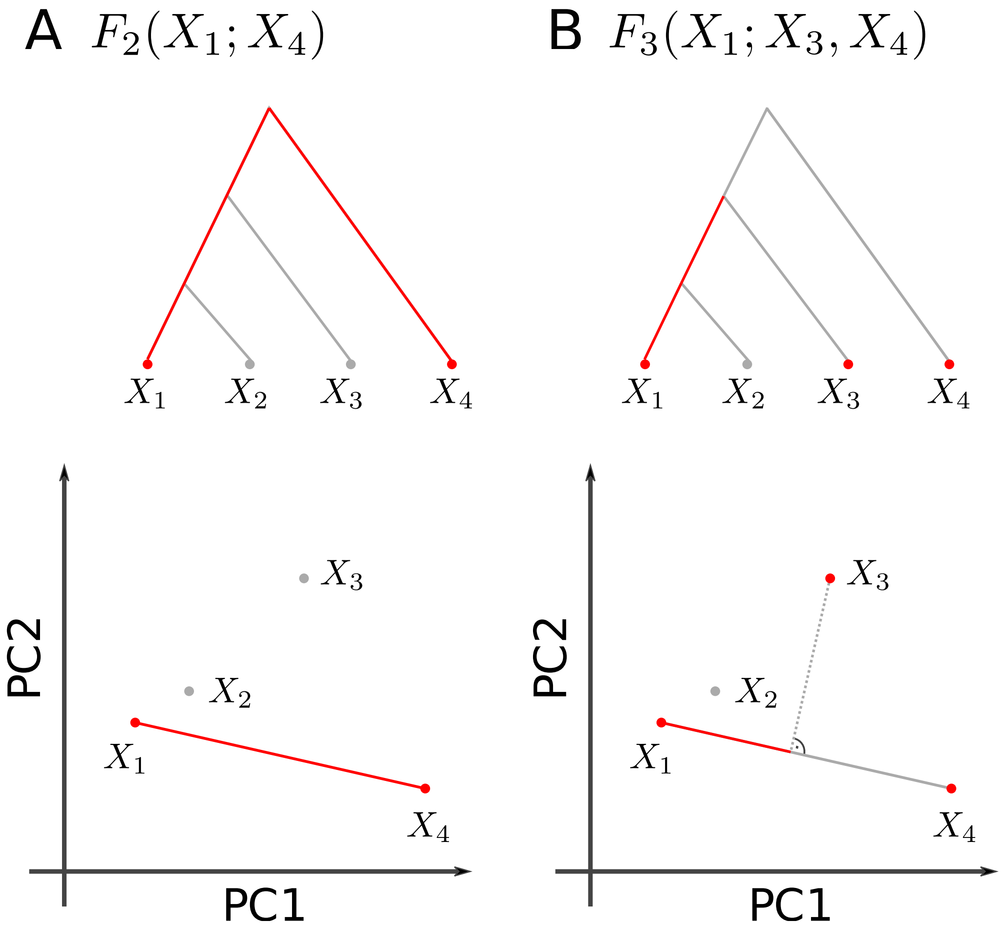
Peter (2022): doi 10.1098/rstb.2020.0413

A $F_2(X_1; X_4)$



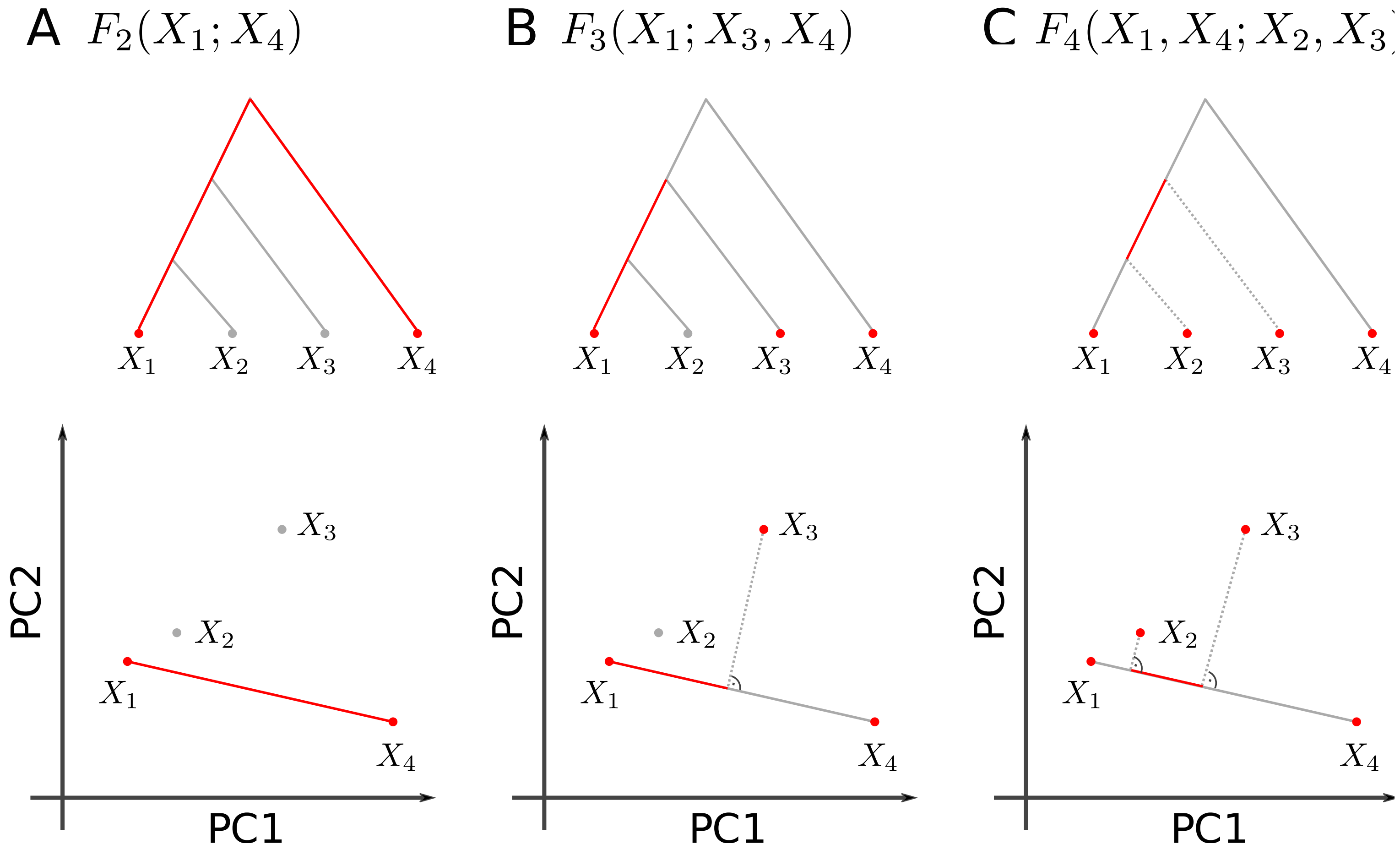
* true if all PCs are used, optimal approximation otherwise

Peter (2022): doi 10.1098/rstb.2020.0413



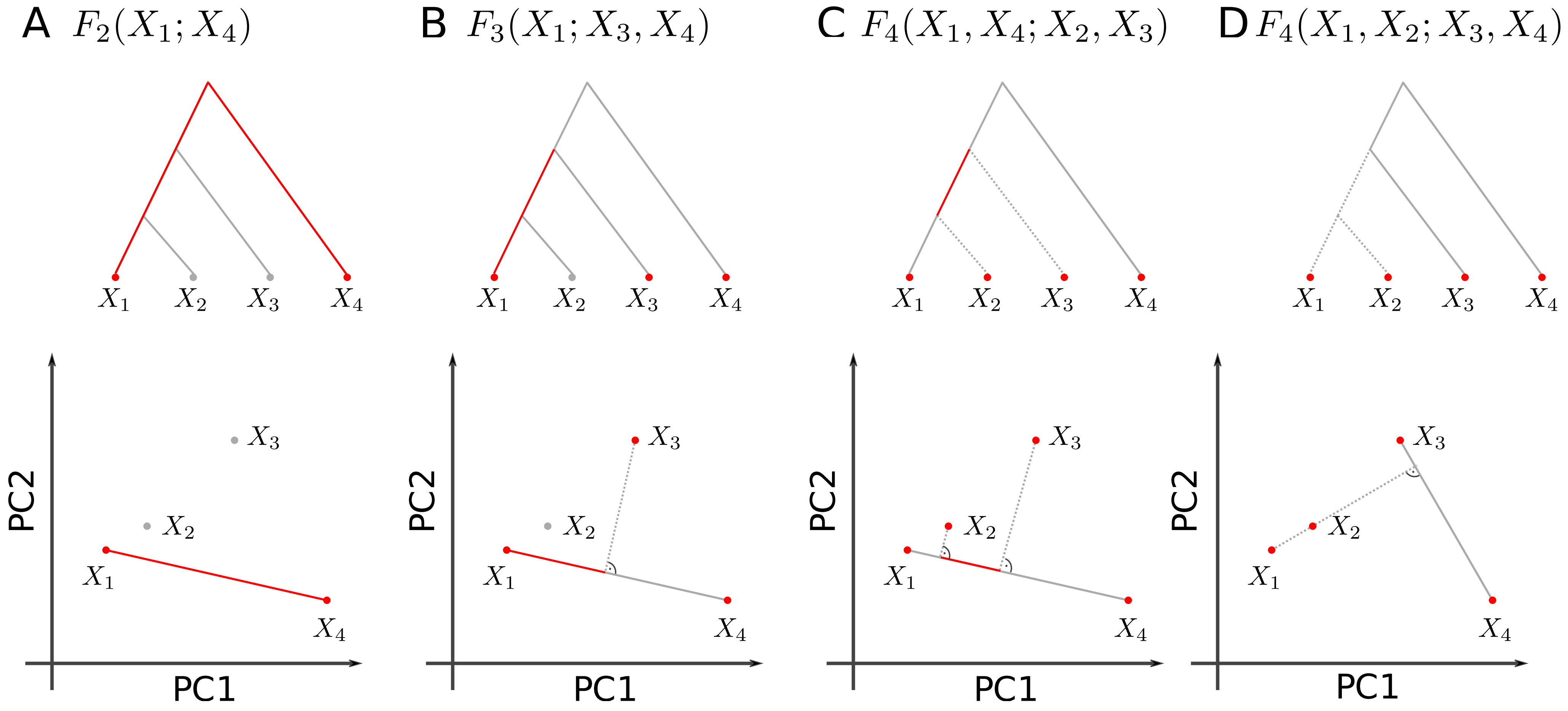
* true if all PCs are used, optimal approximation otherwise

Peter (2022): doi 10.1098/rstb.2020.0413



* true if all PCs are used, optimal approximation otherwise

Peter (2022): doi 10.1098/rstb.2020.0413



* true if all PCs are used, optimal approximation otherwise

Peter (2022): doi 10.1098/rstb.2020.0413