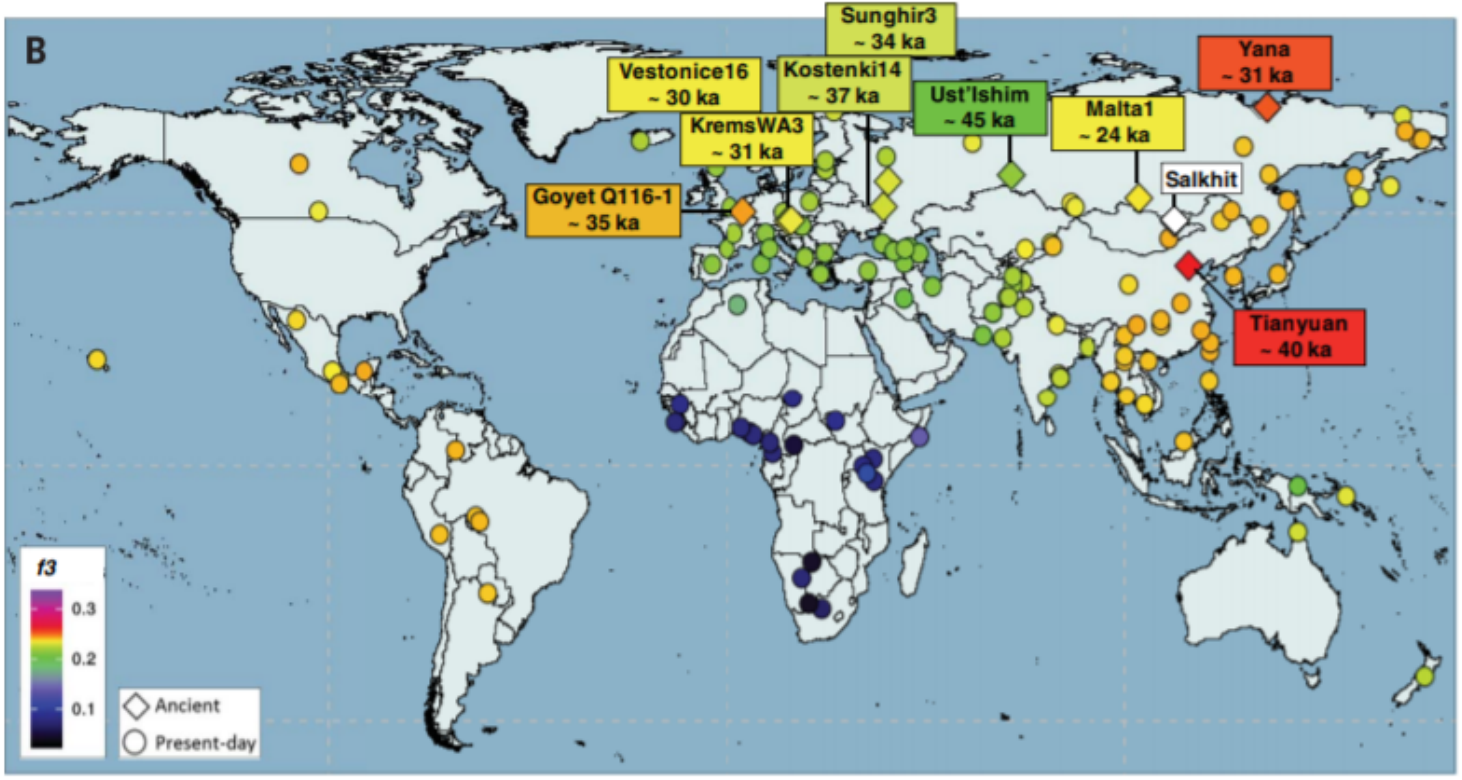
F-statistics and Population Structure

Benjamin Peter, MPI for Evolutionary Anthropology

Motivation: Ancient DNA

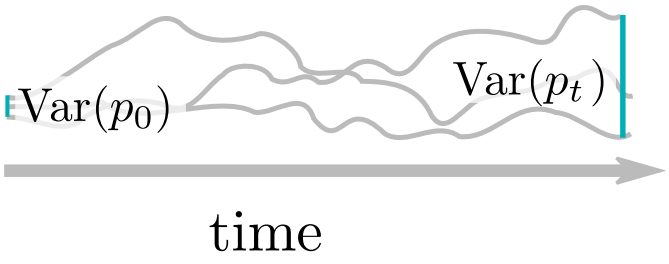


Setup

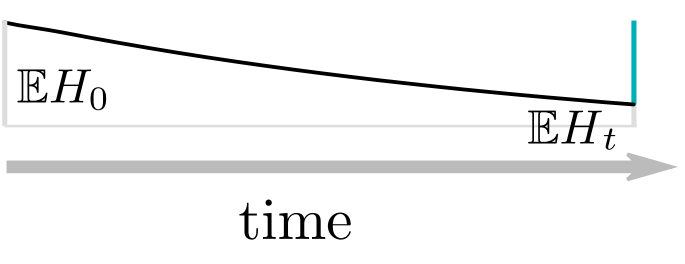
- Today: Theory of F-statistics and Computations
- Tomorrow: Using F-statistics to build more complex models

Measuring Genetic Drift

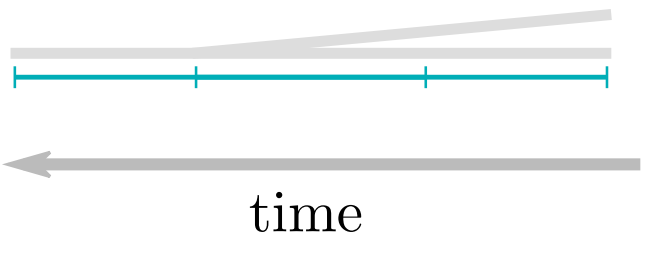
Change in Allele Frequency



Decay of Heterozygosity



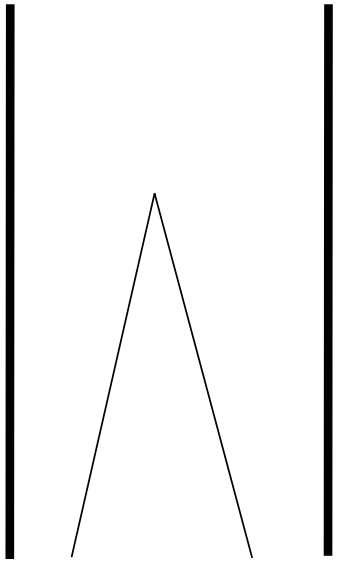
Coalescence rates

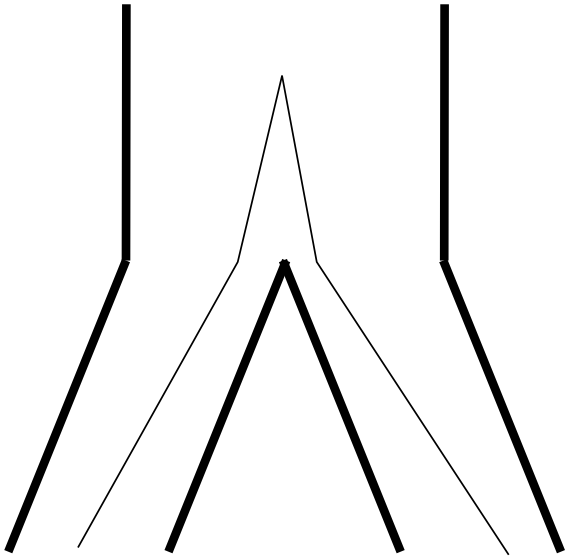


Pairwise differences

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

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



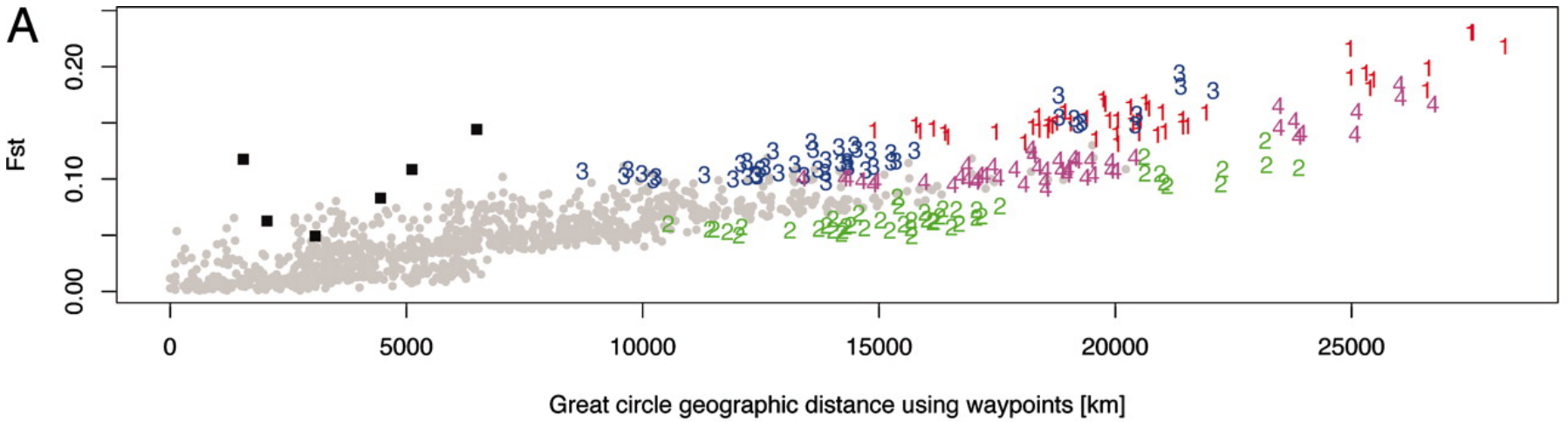


Fixation Index F_{ST}

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

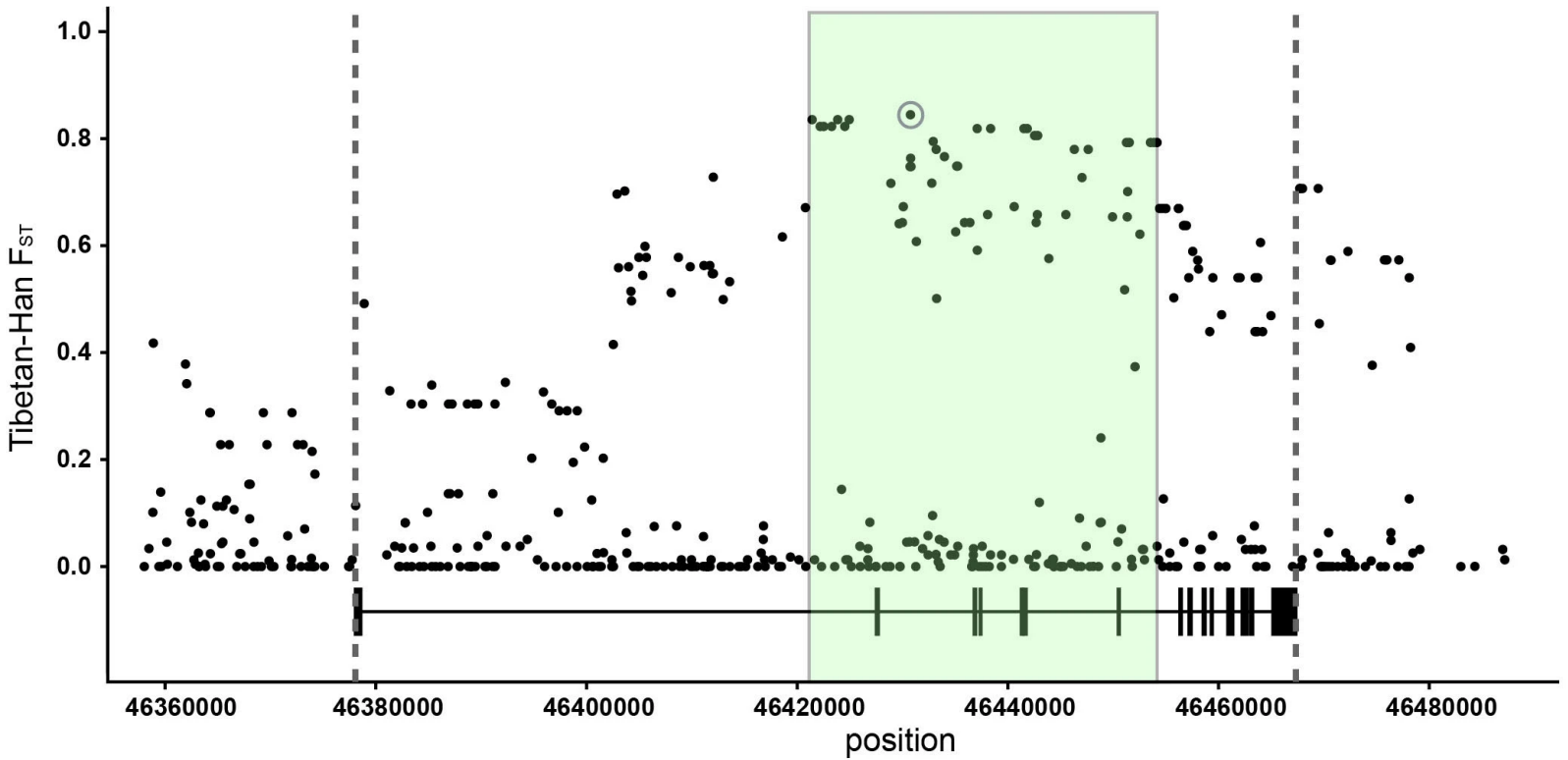
- F_{ST} is a correlation coefficient
- Between 0 and 1
- Hierarchical partitioning (AMOVA)
- Many estimators exist
  - Hudson (1991)
  - Weir & Cockerham (1984)

Fixation Index F_{ST}



Ramachandran et al. 2005

F_{ST} Outliers



Huerta-Sanchez et al. 2014

F_2-statistic

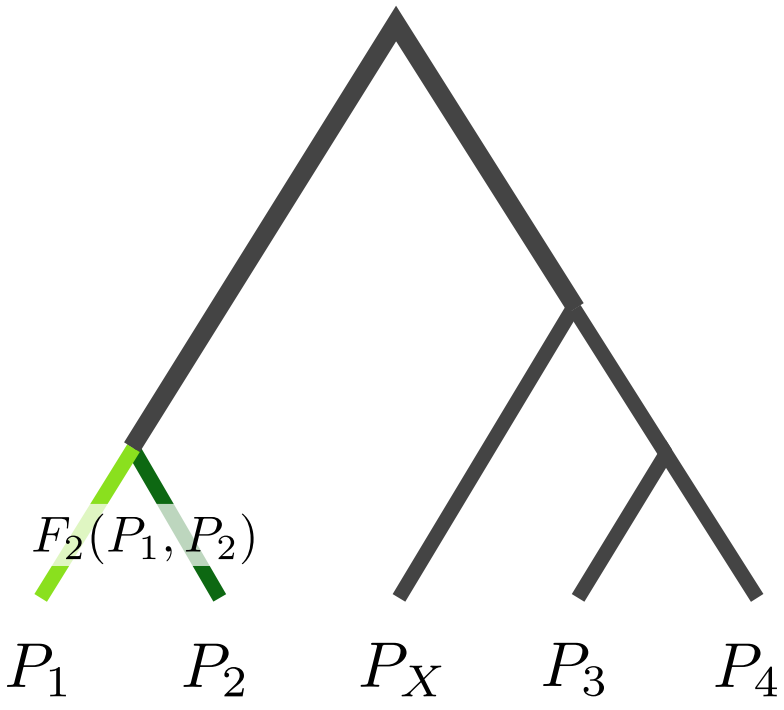
$$F_{ST}(P_1, P_2) = \frac{\pi_{12} - \frac{\pi_1 + \pi_2}{2}}{\pi_{12}}$$

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

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

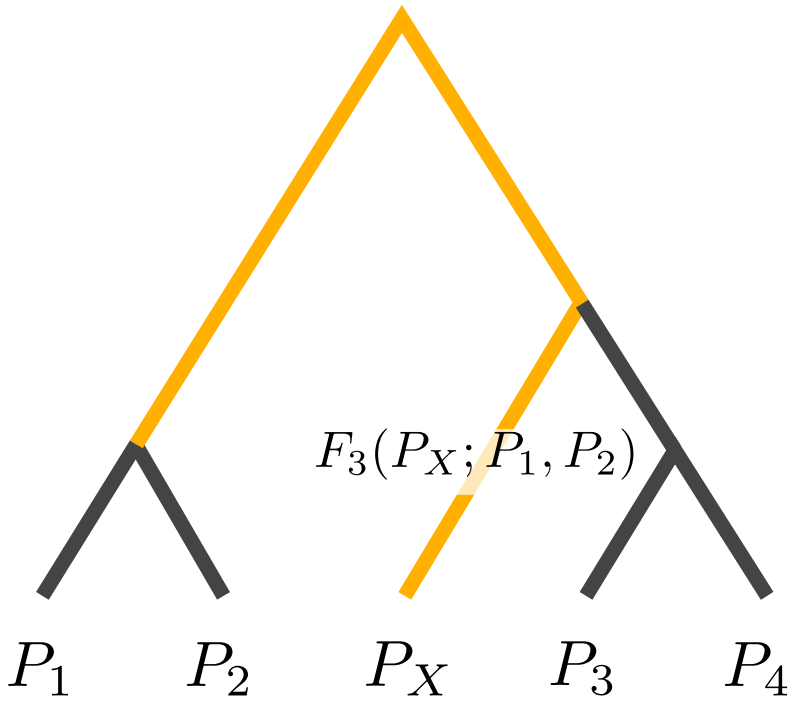
- F_2 is a covariance
- Bigger than 0
- Tree-additive
- Testing for treeness

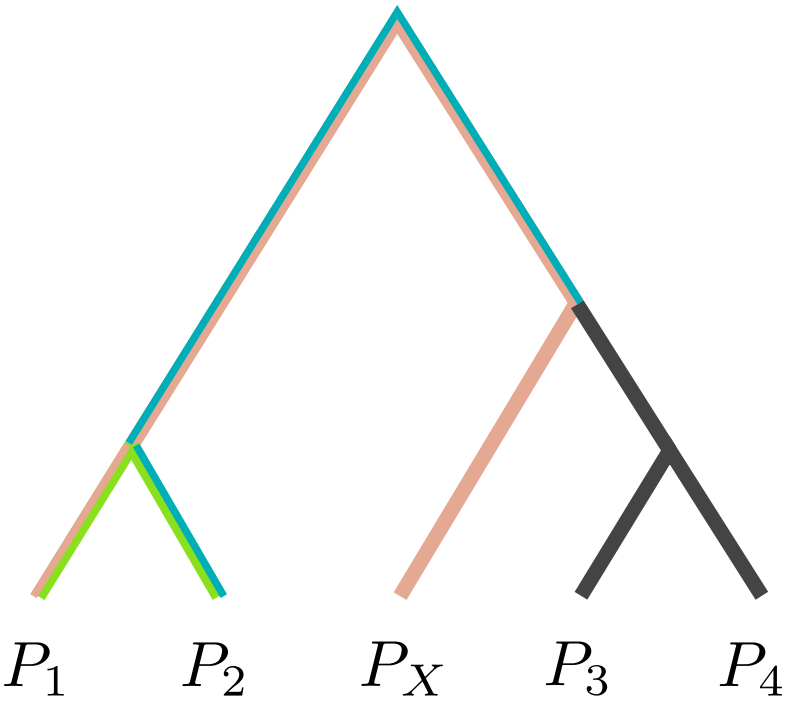
Tree-additive



F_3-statistic

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





F_3-statistic equations

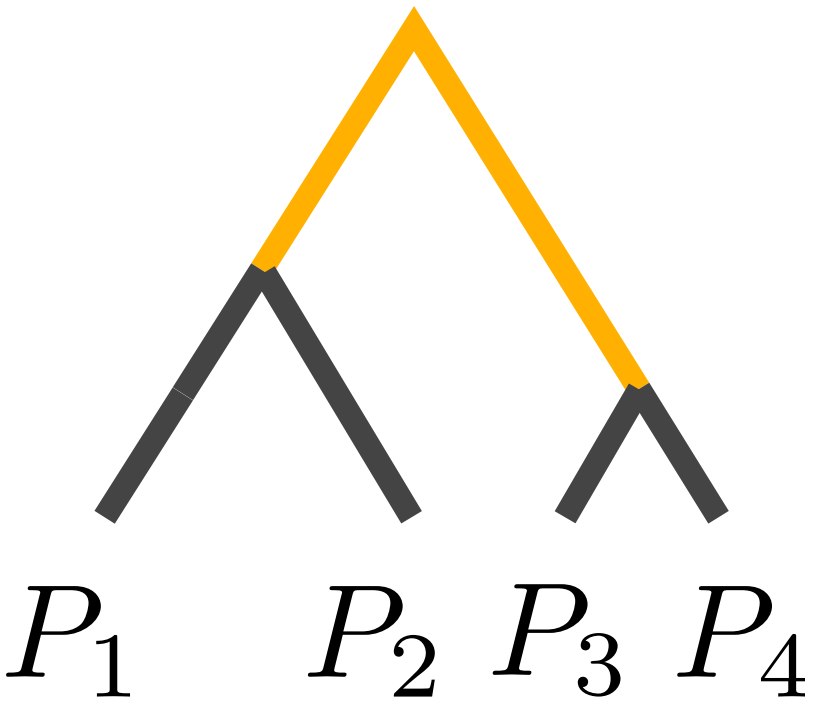
$$F_3(P_X; P_1, P_2) = \frac{1}{2}\Big(F_2(P_X, P_1) + F_2(P_X, P_2) - F_2(P_1, P_2) \Big)$$

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

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

(Branch)-F_4-statistic

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

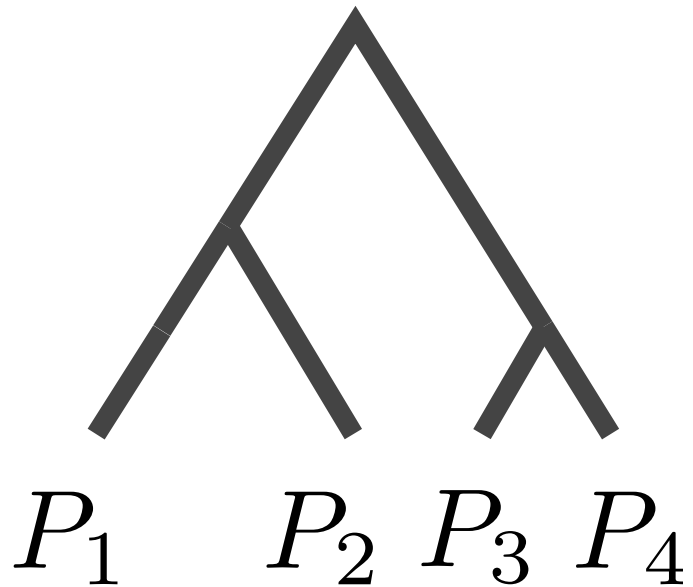


What if we reorder the arguments?

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

(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

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

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

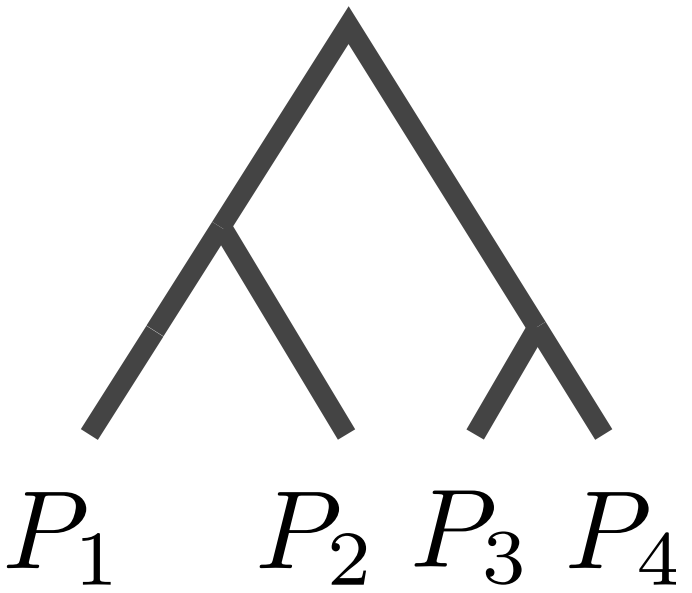
$$F_4^{(T)}(P_1, P_2; P_3, P_4) = \pi_{13} + \pi_{24} - \pi_{14} - \pi_{23}$$

Testing Treeness

If data is generated from a tree:

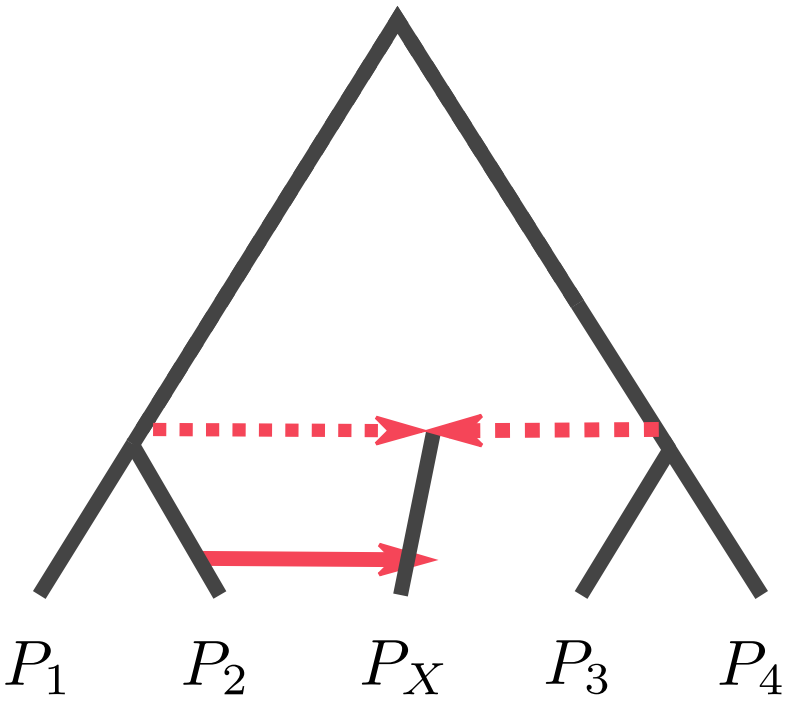
$$F_3(P_3; P_1; P_2) \geq 0$$

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



Buneman 1974

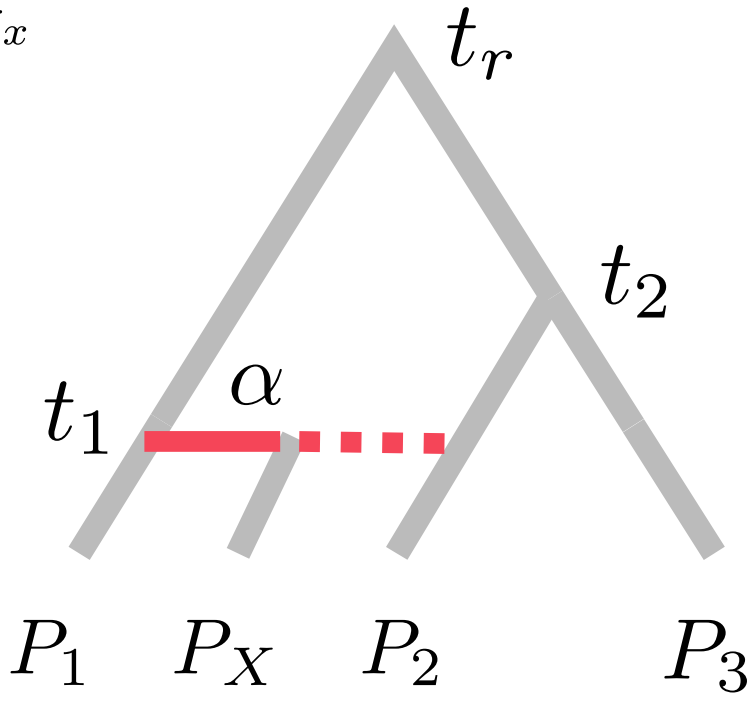
Admixture Graphs



F3 in an admixture graph

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

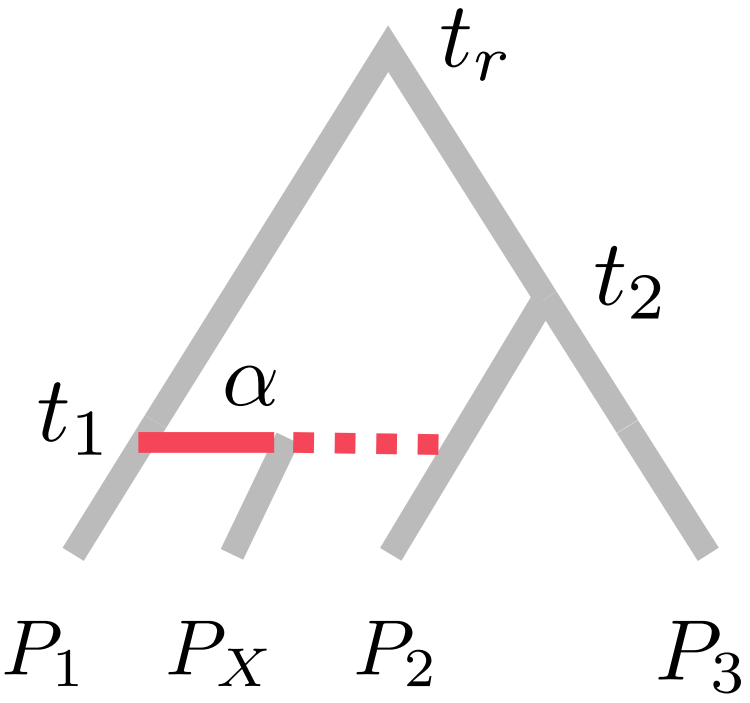
$$F_3(P_X; P_1, P_2) \approx \theta\left[t_1 - 2\alpha(1-\alpha)t_r\right]$$



F4 in an admixture graph

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

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

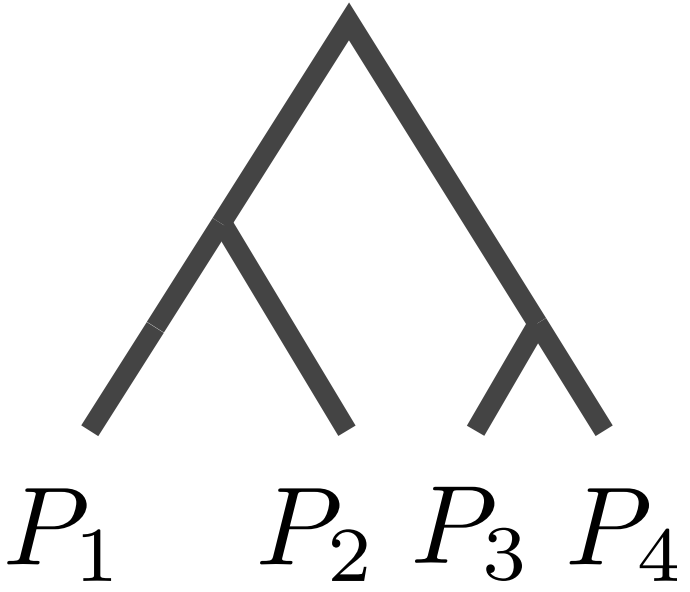


D-statistic

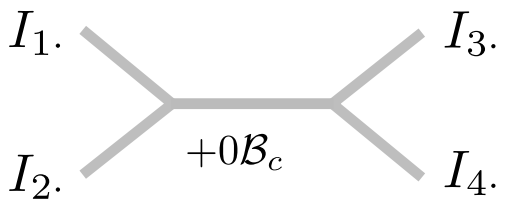
$$D = \frac{\mathrm{ABBA} - \mathrm{BABA}}{\mathrm{BABA} + \mathrm{ABBA}}$$

- D-statistic and F4 are closely related

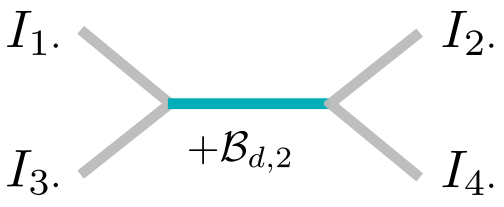
$$F_4^{(T)}(P_1, P_2; P_3, P_4) = \pi_{13} + \pi_{24} - \pi_{14} - \pi_{23}$$



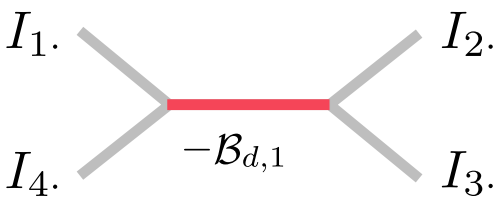
concordant



ABBA-genealogy



BABA-genealogy



F4-ratio

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