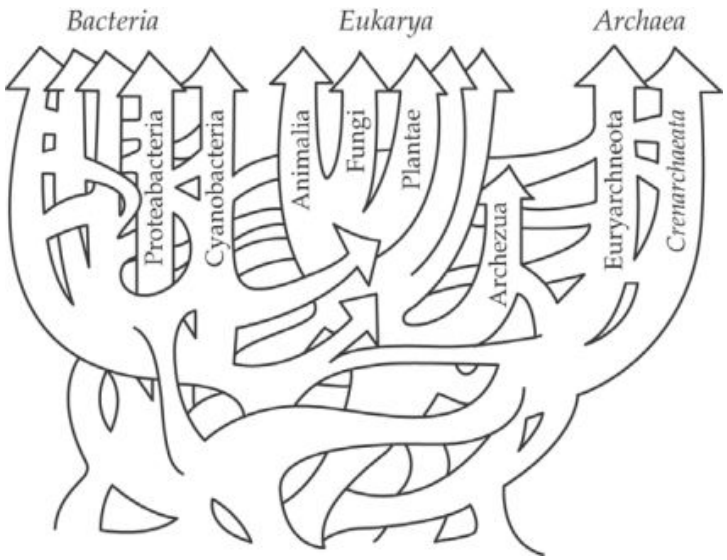
Using ABBA-BABA tests to infer gene flow and genomic introgression from genome data (WGS, radSeq, etc.)

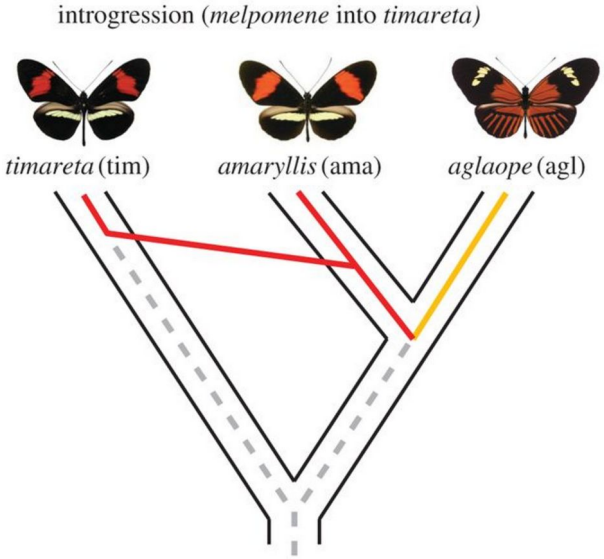
Hannes Svardal, Milan Malinsky

Hybridisation, gene flow, admixture and introgression

- **In short:** they all kind of mean the same thing
- Genetic admixture ~ gene flow … previously isolated populations (species) interbreed (hybridise) resulting in a genetic mixture of the two original populations
- Genomic introgression … incorporation of alleles from one population (species) into the gene pool of a second, divergent population (species);



Arnold and Fogarty Int J Mol Sci 2009

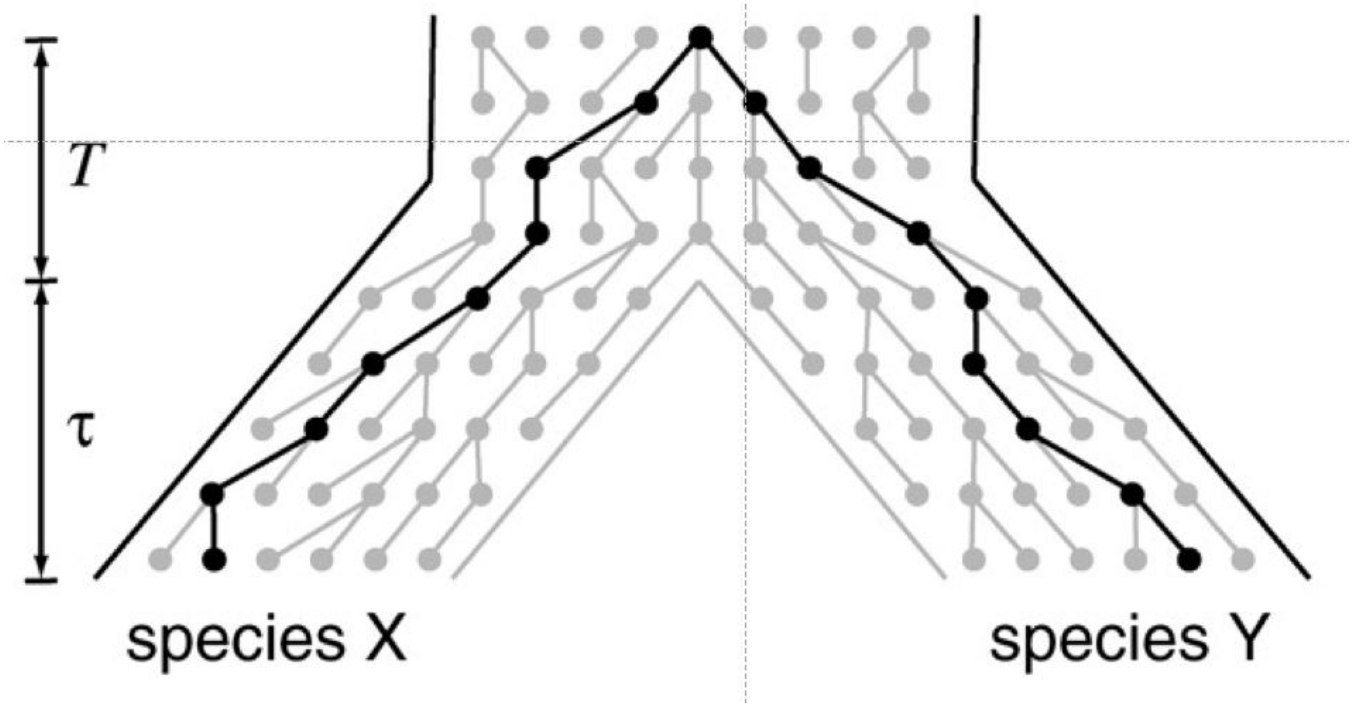


Smith and Kronforst Biology Letters 2013

Primer

- In the following, we use the word **species** to refer to genetically separated groups
- However, these groups could just as well be separate **populations** of the same species

Cross-species coalescent

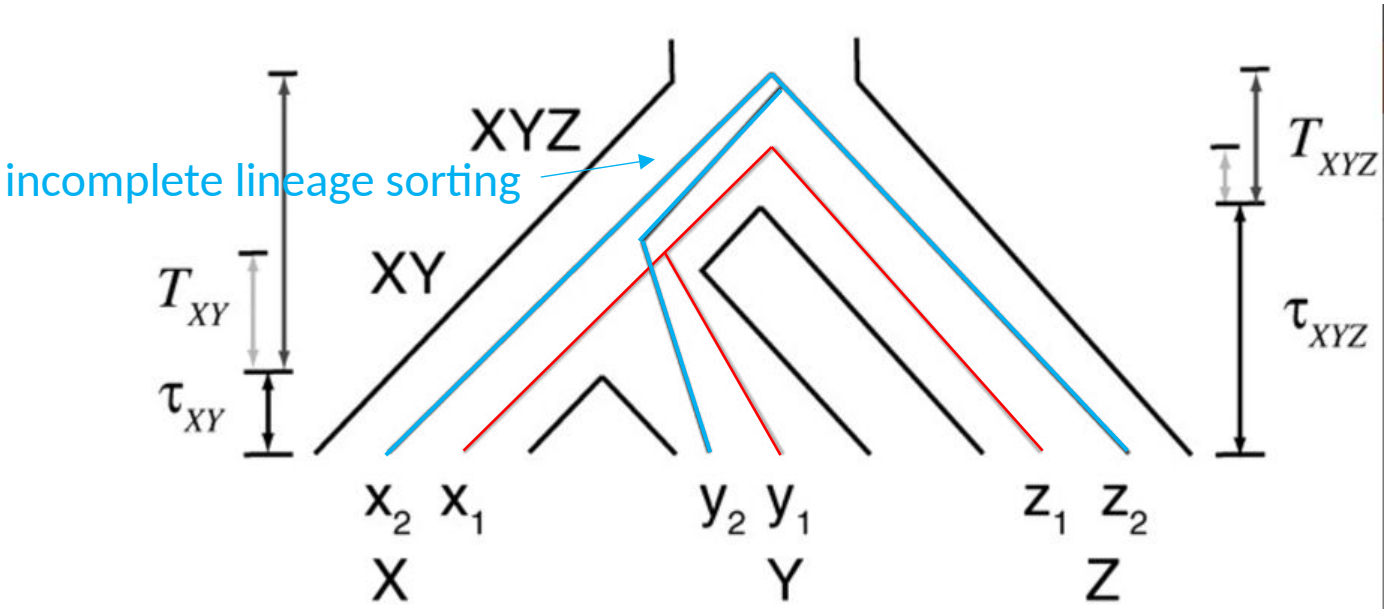


- τ generations before the present the precursors of species X and Y became genetically isolated
- Chromosomes from species X and Y can only find a common ancestor in the ancestral population (after τ generations), but not immediately (here after τ + T generations)

Siepel A. Genome Res. 2009

Incomplete lineage sorting

- Three-species phylogeny for species *X*, *Y*, and *Z*, with ancestral species *XY* and *XYZ*



- At locus 1, samples x_1, y_1, and z_1 have a genealogy that reflects the species tree (red), but at locus 2, samples x_2, y_2, and z_2 have a genealogy with a discordant topology (blue) = **incomplete lineage sorting**

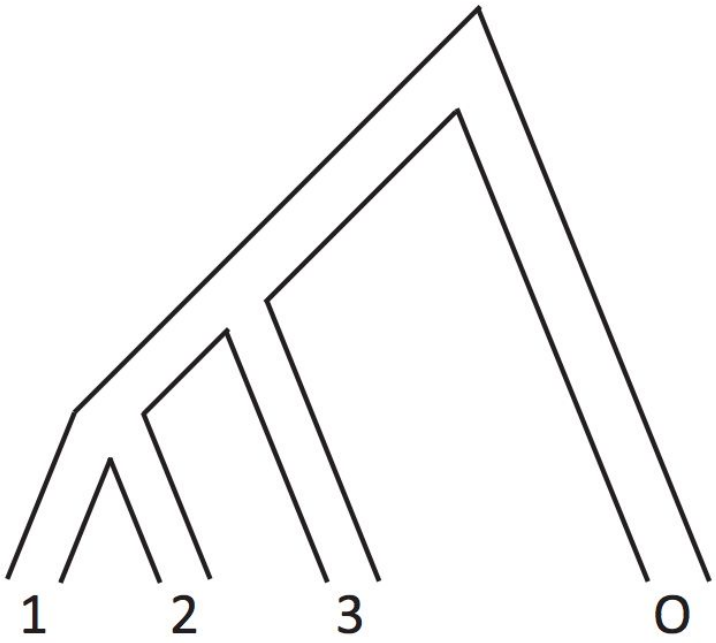
Siepel A. Genome Res. 2009

Incomplete lineage sorting

- Incomplete lineage sorting can lead to patterns that resemble signals of gene flow:
  - as some loci, samples of non-sister species are closer to each other than samples from sister species
- Hence, tests for gene flow need to be robust against incomplete lineage sorting

D statistic (ABBA-BABA test)

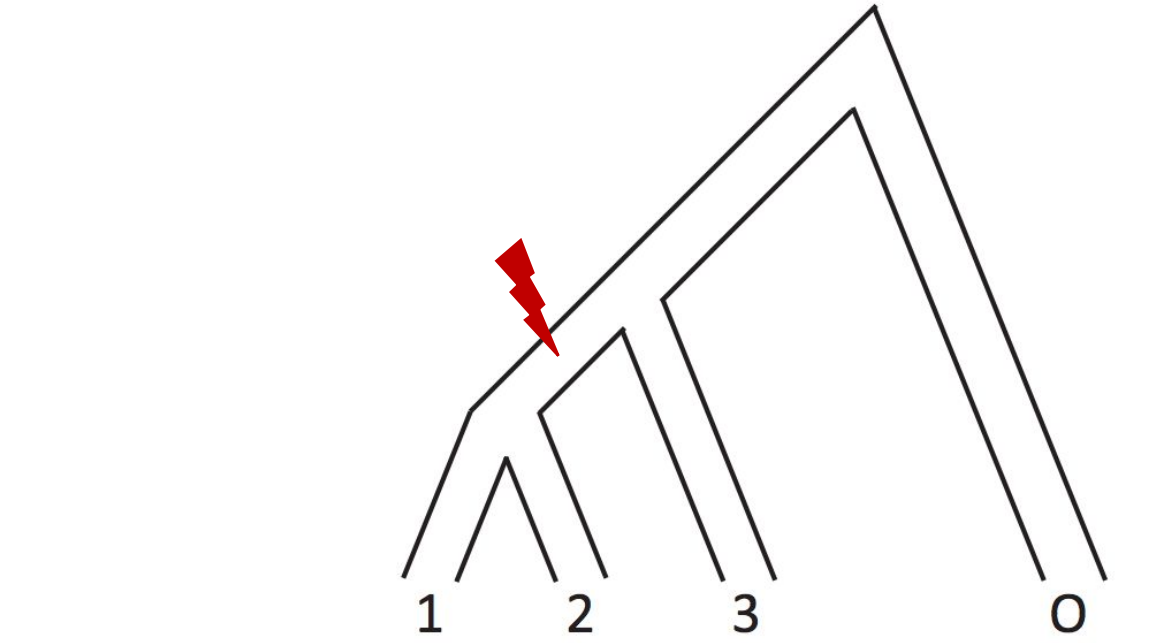
- Patterson, Reich and others developed a powerful and widely used set of statistics that test for genomic introgression
- Assume tree of three populations (African, European, Neanderthal) and an outgroup (Chimpanzee)
- Call ancestral allele A
- Consider sites where you observe each allele twice
- Three possible configurations



	1	2	3	O
Configuration 1	B	B	A	A
Configuration 2	A	B	B	A
Configuration 3	B	A	B	A

Gaut et al. Trends in Genetic 2015

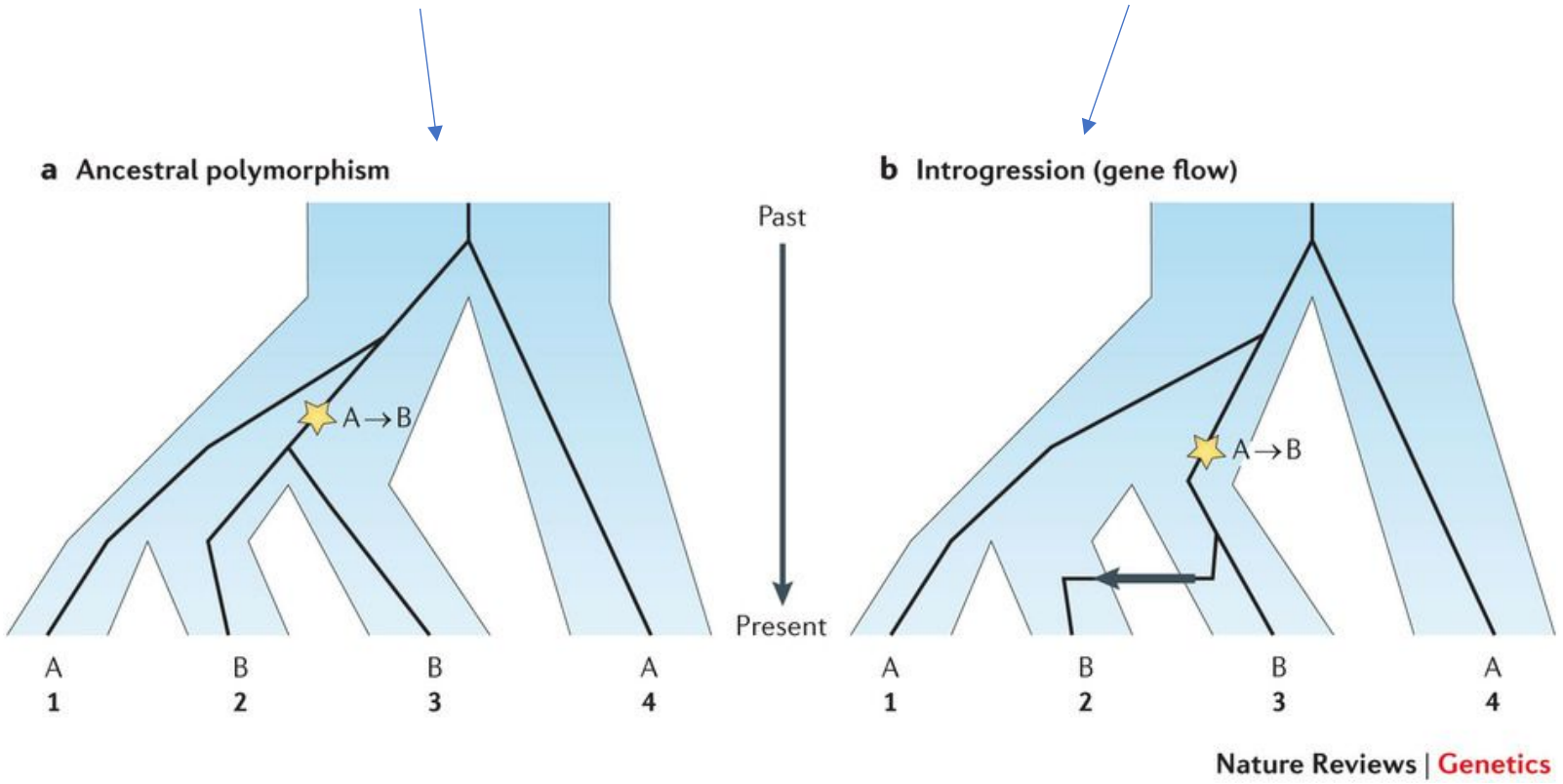
D statistic (ABBA-BABA test)



- Mutations on terminal branches are not expected to give two As and two Bs
- Would you expect to observe Configurations 2 and 3?

ABBA and BABA

- Configurations 2 and 3 are expected due to *incomplete lineage sorting (ILS)* or *introgression*

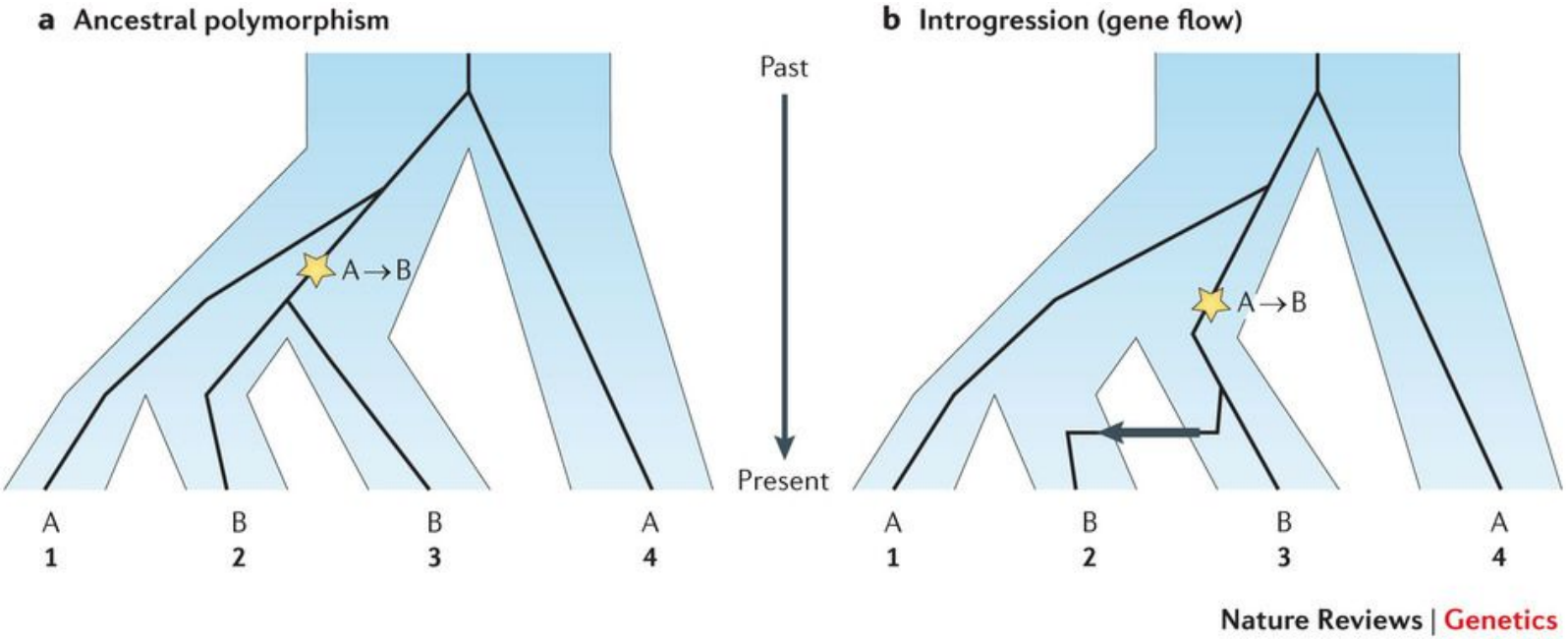


- How to distinguish these two alternatives?

Sousa and Hey *Nature Reviews Genetics 2013*

Incomplete lineage sorting vs introgression

- Under a model with no gene flow, we expect that the pattern ABBA is as frequent as BABA because there is 50% chance that either the lineage from population 1 or from population 2 coalesces with lineage from population 3 in the population ancestral to populations 1, 2 and 3.
- Under introgression 3 -> 2, we expect more ABBA than BABA.



D statistic (ABBA-BABA test; Pattersons D)

$$D(1,2,3,4) = \frac{\sum C_{\mathrm{A}ABBA} - \sum C_{\mathrm{B}ABA}}{\sum C_{\mathrm{A}ABBA} + \sum C_{\mathrm{B}ABA}}$$

- C_{ABBA} is the count of SNPs with the ABBA pattern
- Equivalently, D can be calculated from allele frequencies

$$D(1,2,3,4) = \frac{\sum_{\mathrm{loci}} (1-\hat{p}_1)\hat{p}_2\hat{p}_3(1-\hat{p}_4) - \sum_{\mathrm{loci}} \hat{p}_1(1-\hat{p}_2)\hat{p}_3(1-\hat{p}_4)}{\sum_{\mathrm{loci}} (1-\hat{p}_1)\hat{p}_2\hat{p}_3(1-\hat{p}_4) + \sum_{\mathrm{loci}} \hat{p}_1(1-\hat{p}_2)\hat{p}_3(1-\hat{p}_4)}$$

Where $\hat{p}_i$ is the estimated allele frequency in population i (this formula is only valid for $\hat{p}_4 = 0$)

D statistic (ABBA-BABA test) (II)

- A significant D > 0 (excess of ABBA) would be a sign of introgression between populations 2 and 3
- A significant D < 0 (excess of BABA) would be a sign of introgression between populations 1 and 3
- D measures excess allele sharing of population 3 with populations 1 or 2
- It is difficult to infer the directionality of introgession (from 3 or into 3)

Testing significance

- If all loci (SNPs) were independent, one could test significance using a binomial test
- However, nearby loci are in linkage disequilibrium
- In that case, significance can be assessed using a block jackknife
  - partition the data in *m* blocks larger than the LD structure
  - recalculate D_i, $i = 1, \ldots, m$, by leaving out block i
  - then $\sqrt{mVar[D_i]}$ is an approximately normally distributed standard error
  - $Z = \frac{D}{\sqrt{mVar[D_i]}}$ measures how many standard deviations D is from its expectation (which is 0)

Testing significance (II)

- A p-value can be obtained for a given Z as the quantile of a standard normal distribution
- p = 1 – CDF(|Z|)
- In practice, p-values can be obtained from statistical software or tables
- Example: Z = 3 → p = 0.00135

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$

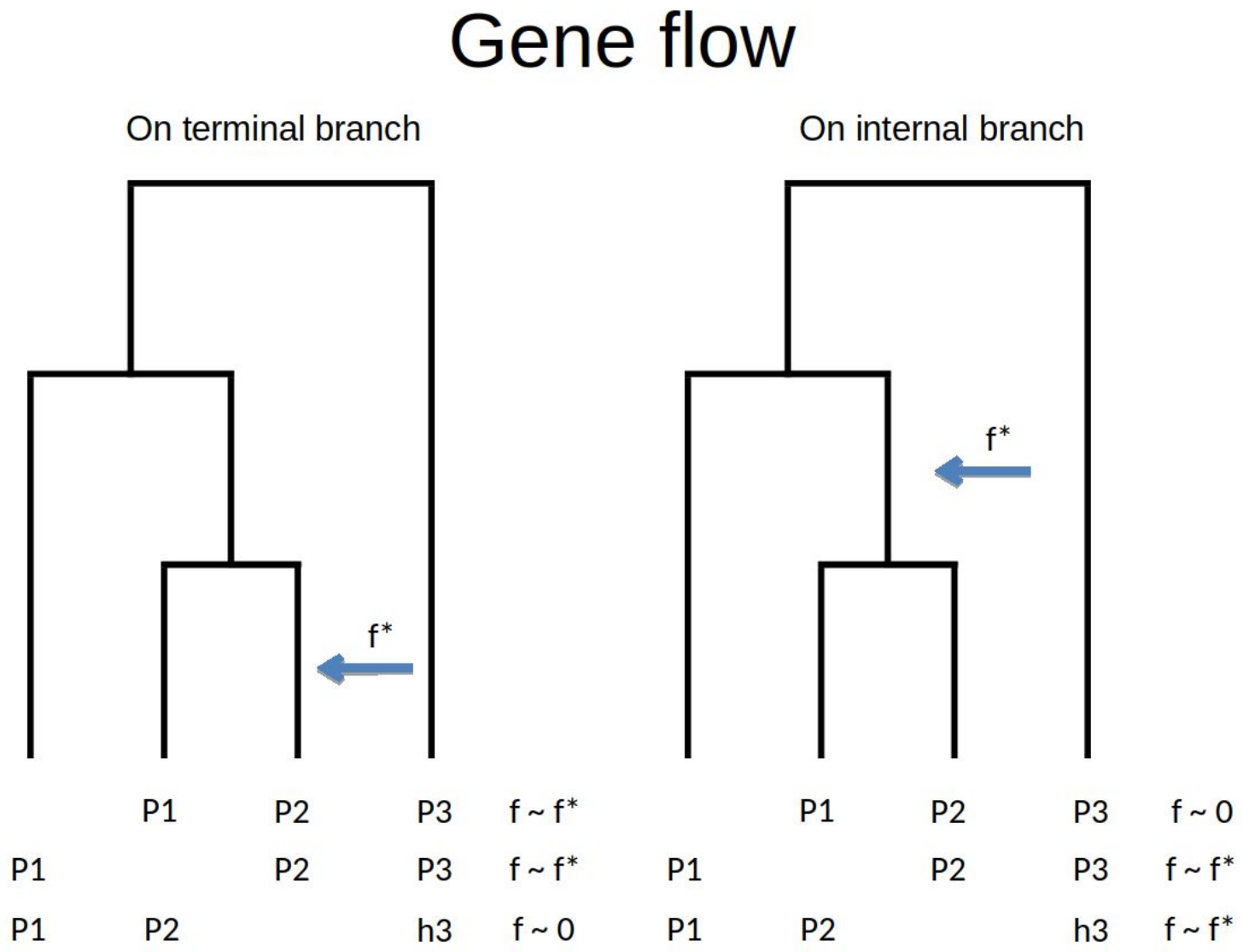
F4 admixture ratio, etc.

- The original **F4 statistic** is just the numerator of the D statistic (not normalised)
- The **F4 admixture ratio** has the same numerator as the D statistic but a different normalisation.
- Under some assumptions, the **F4 admixture ratio** is a direct estimate of the **proportion of admixture.**
- (In simulations it often is an under-estimate).
- For more information about F3, F4, D statistic etc., see Patterson et al. 2012 Genetics

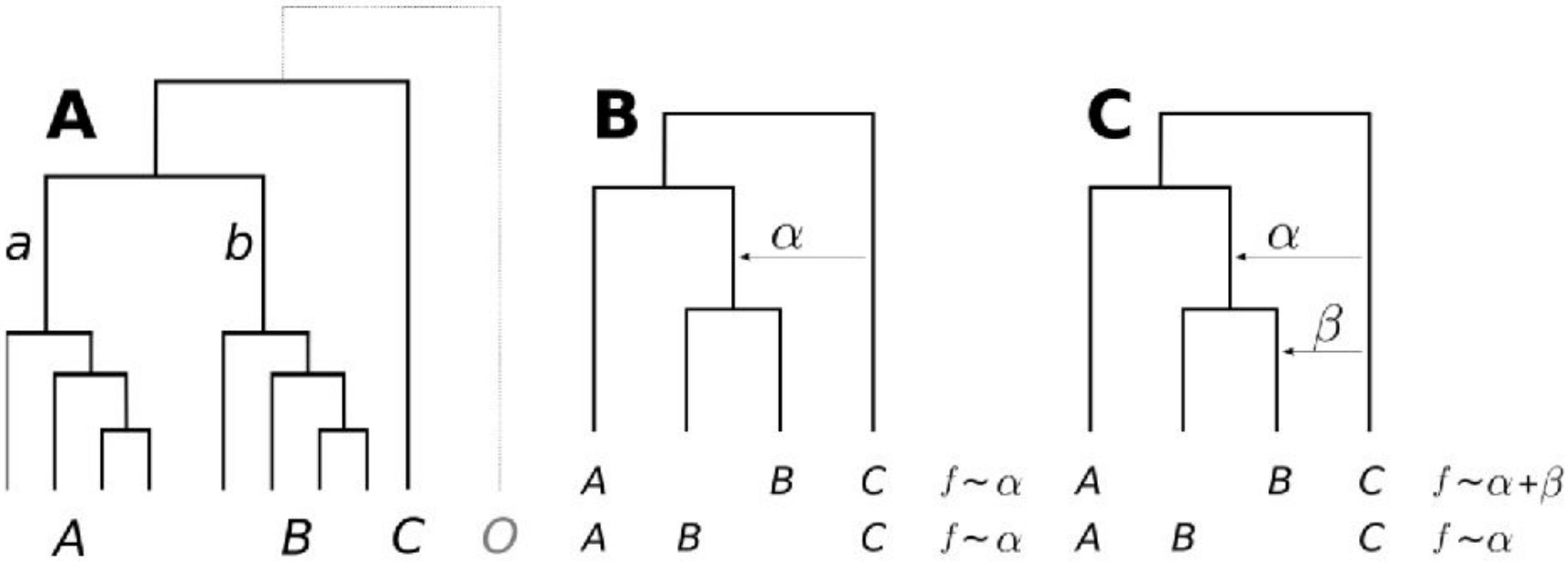
f branch

- D/f statistics are not independent
- They can show highly correlated values when the share a branch (internal or external)
- A single gene flow event can lead to a large number of non-zero D/f statistics
- f branch addresses this (partly) by calculating and visualising f statistics along internal and external branches of a tree

f branch – motivation



f branch – motivation



Supplementary Figure 26

The calculation of *f* scores on trees. (A) A schematic illustration accompanying the methods section which explains how the reduced branch specific $f_b(C)$ scores are calculated. **(B, C)** Illustration of the way in which interdependences between different *f* scores can be informative about the timing of introgression. As shown in (B), gene flow into the common ancestor of two species is expected to be equally detectable in both of them. Additional gene flow into only one species after their split will add to the *f* statistic in that species, but not in its sister species.