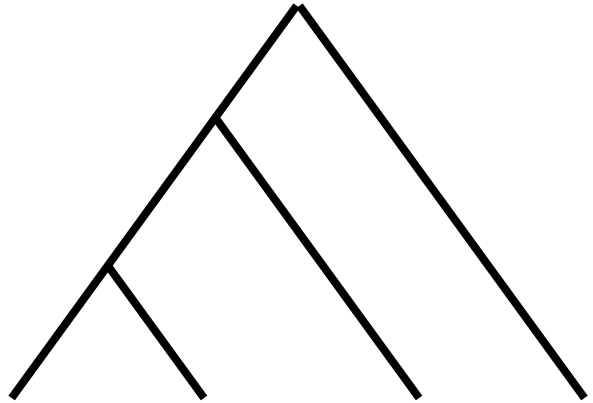
Detecting hybridization with ADMIXTOOLS

Patterson's D statistics to identify hybridization

- Also called ABBA-BABA test: D=(ABBA-BABA)/(BABA+ABBA)

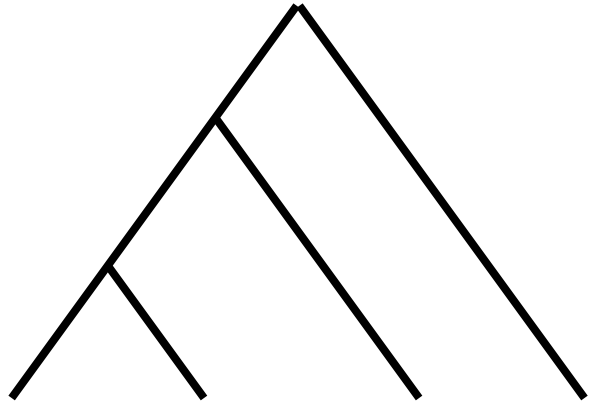


$$D = \frac{ABBA - BABA}{ABBA + BABA}$$

P1	P2	P3	P4	
T	**T**	A	A	
G	**G**	**G**	T	Concordant SNPs
C	T	T	T	
C	A	**C**	A	
A	**T**	**T**	A	Discordant SNPs

Patterson's D statistics to identify hybridization

- Also called ABBA-BABA test: D=(ABBA-BABA)/(BABA+ABBA)



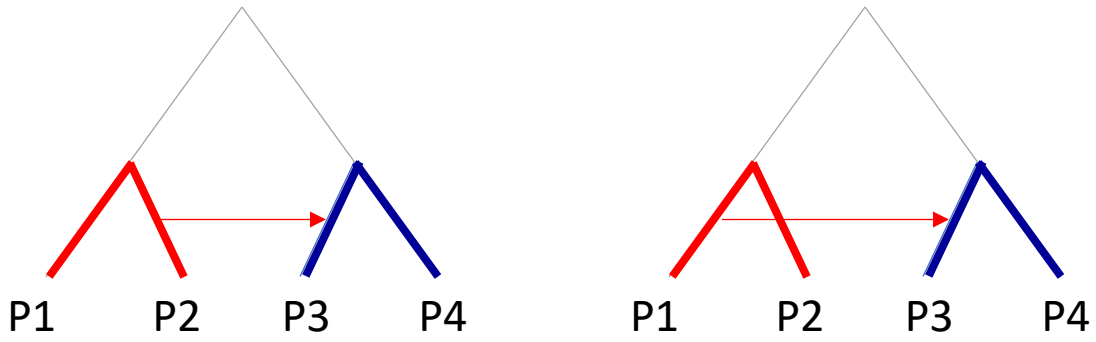
P1	P2	P3	P4	
T	**T**	A	A	
G	**G**	**G**	T	Concordant SNPs
C	T	T	T	
C	A	**C**	A	← BABA
A	**T**	**T**	A	← ABBA
C	**G**	**G**	C	← ABBA

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

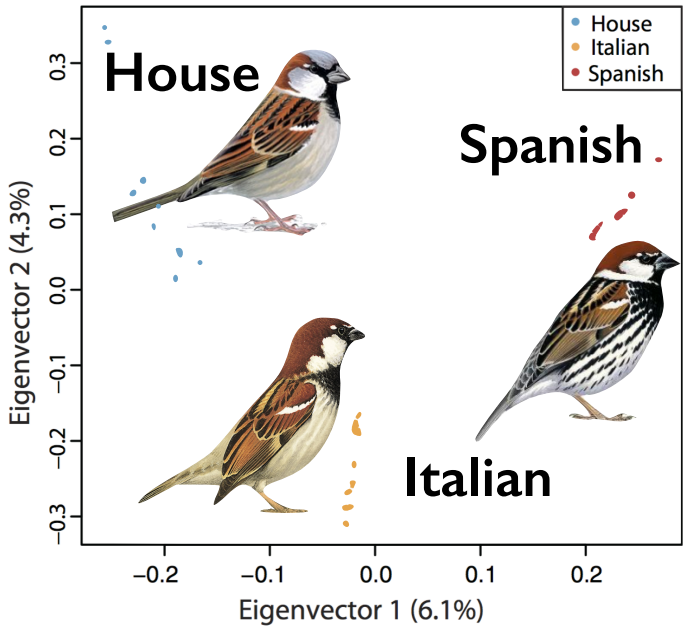
$$D = \frac{2-1}{2+1} = 1/3$$

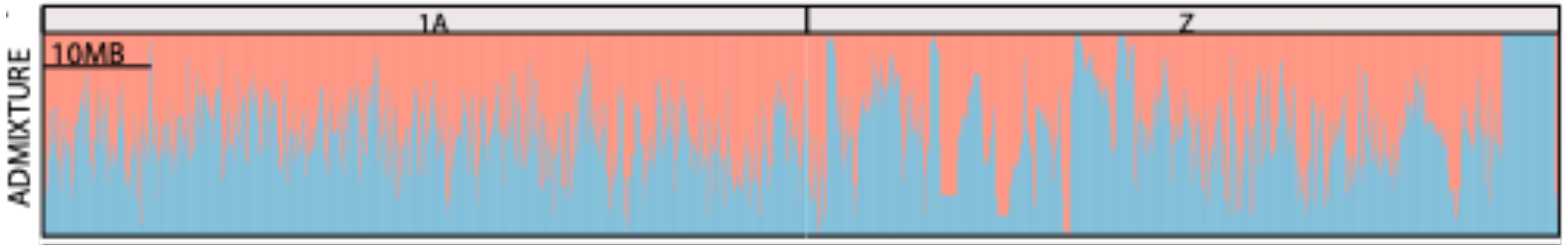
f4 tests to identify hybridization

- F statistics are based on drift indices or variance in allele frequencies between 4 populations
- F4(P1,P2:P3,P4) should be zero as no shared drift path without hybridization
- If ((P1,P2),(P3,P4)) is correct, the allele frequency differences between P1 and P2 should be uncorrelated with those between P3 and P4, which we can assess by averaging the quantity (p1-p2)(p3-p4) across SNPs.
- D statistic: Normalized f4 statistic with polarized alleles (conditions on derived allele frequency using an outgroup)

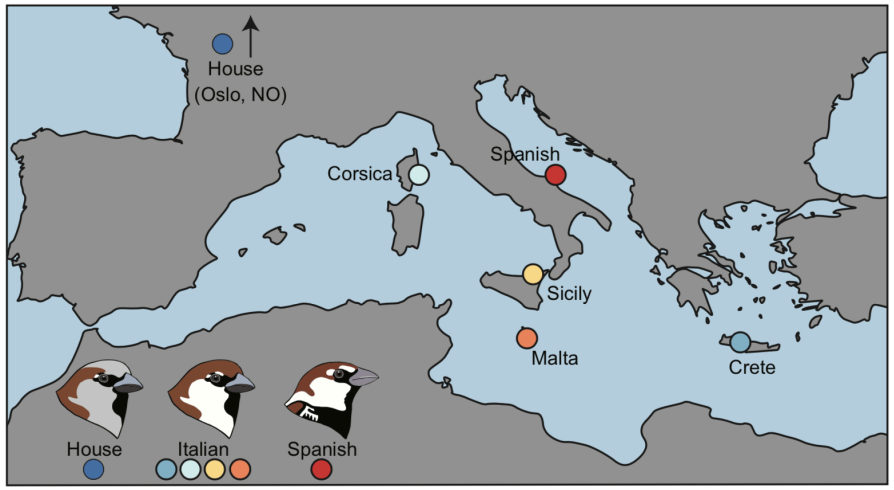


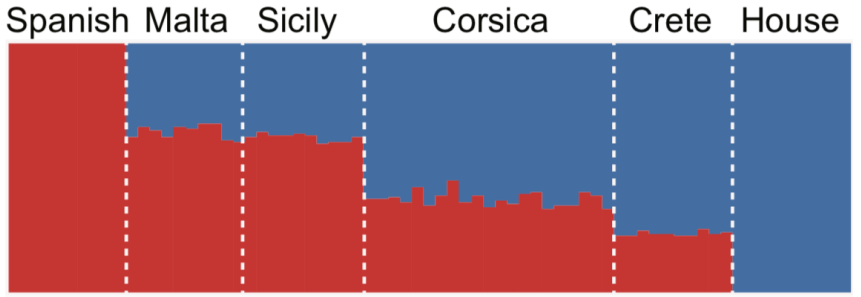
Are Italian sparrows admixed?





Elgvin et al (2017) **Science Advances**



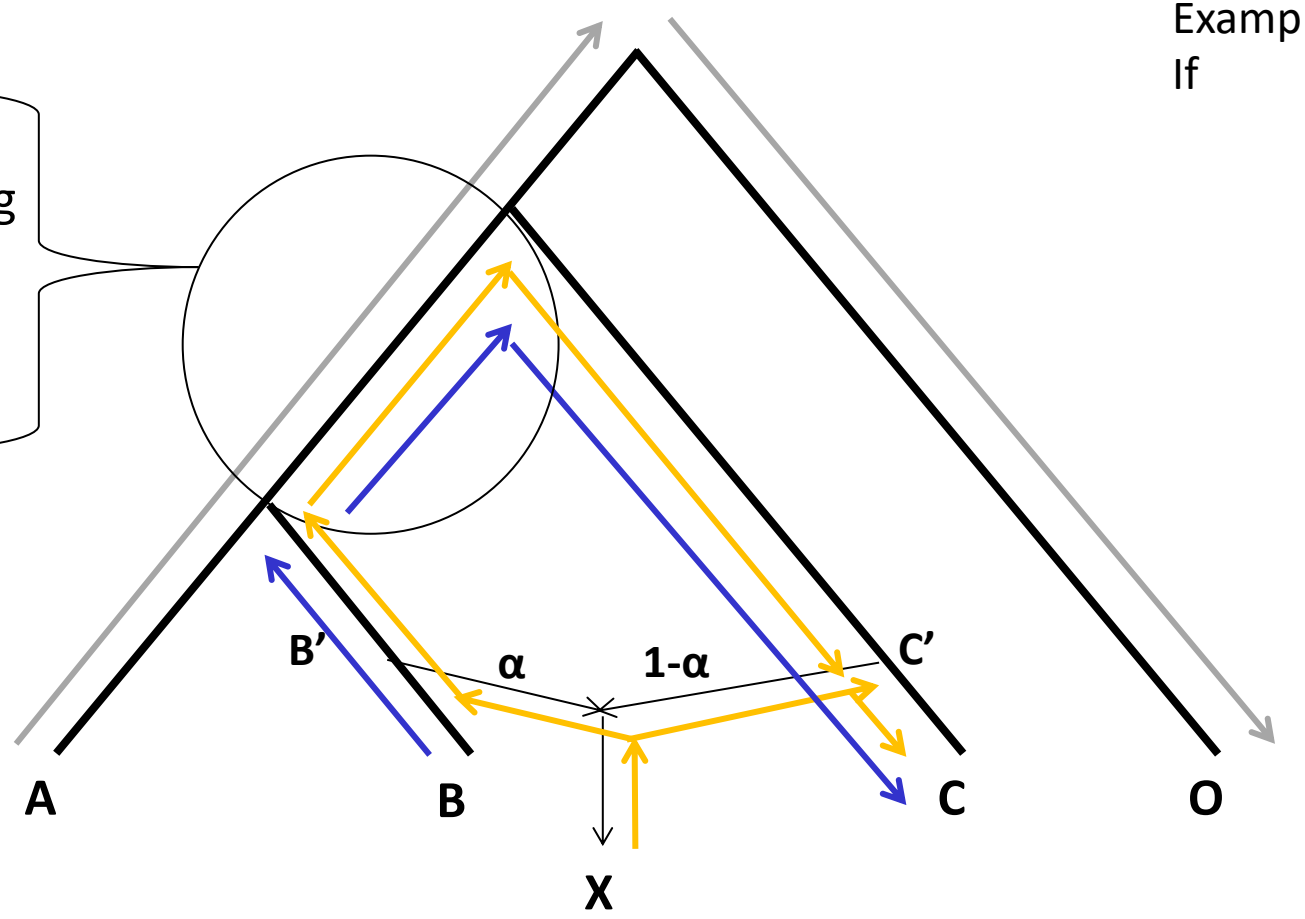


Multiple hybridisation events?

Runemark et al (2017) **Nature Ecology & Evolution**

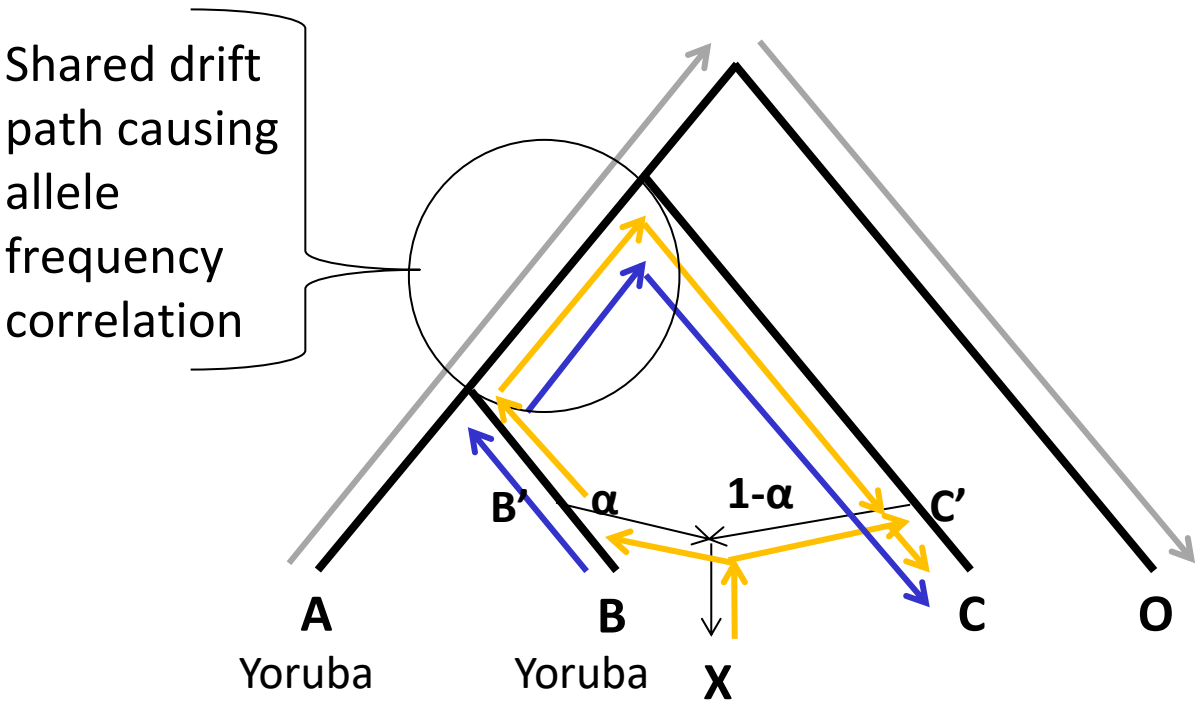
f4 ratio test to infer the amount of admixture

Shared drift path causing allele frequency correlation



Example correlation:
If

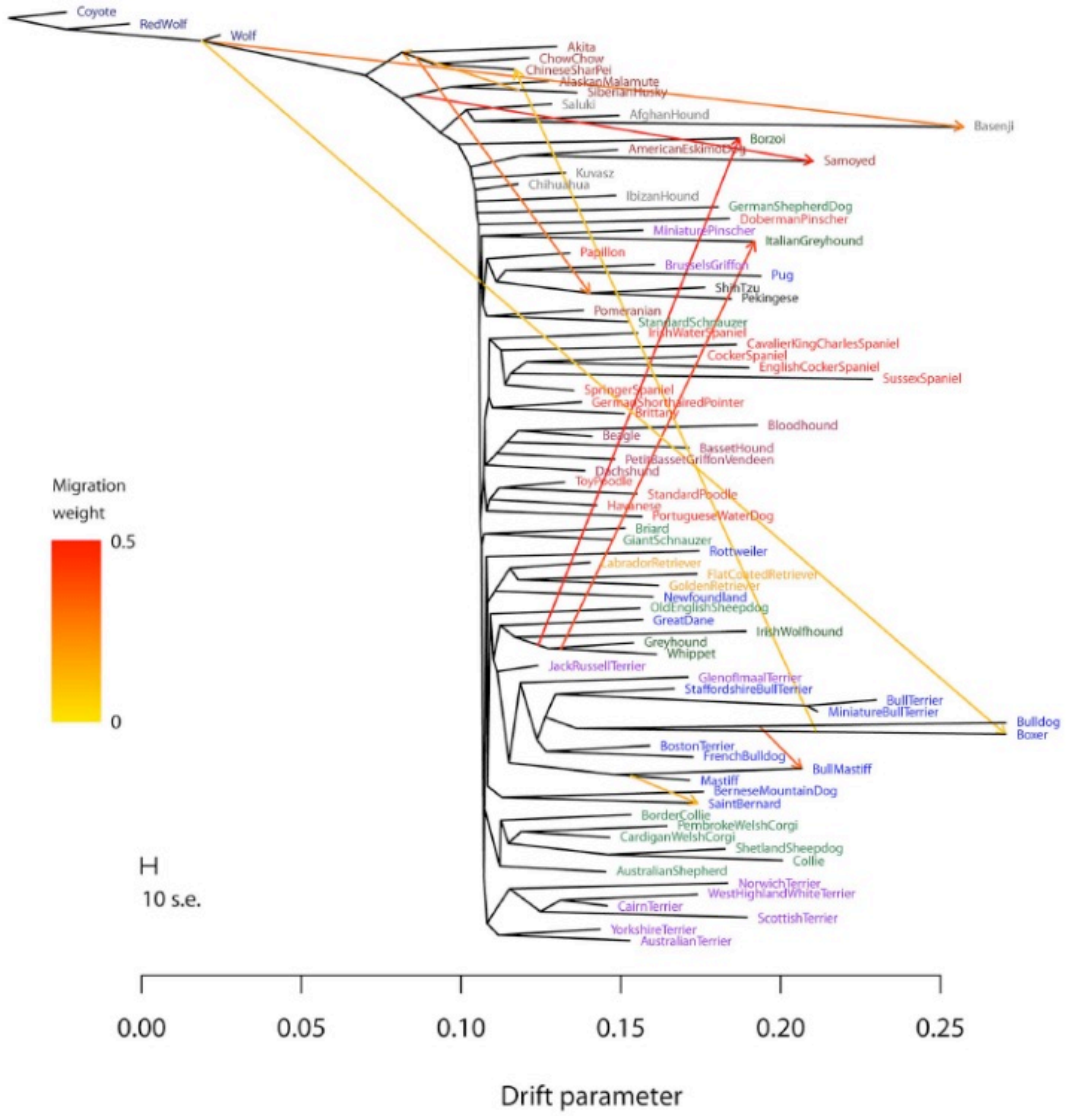
f4 ratio test to infer the amount of admixture



The shared drift path generates correlation between A-O and B-C. X-C share the same drift path by the proportion α.

$$\alpha = \frac{F_4(A,O;\,X,C)}{F_4(A,O;\,B,C)} = \frac{E[(a-o)(x-c)]}{E[(a-o)(b-c)]} = \frac{\text{Correlation of allele frequency difference between a-o and x-c}}{\text{Correlation of allele frequency difference between a-o and b-c}}$$

Treemix (Pickrell & Pritchard, 2012, Plos Genetics)



- Treemix uses covariances of allele frequencies to infer a maximum likely tree with a user-specified number of migration edges
- Gene flow edges are added sequentially to account for the greatest errors in the fit. This format makes TreeMix well-suited to handling very large trees: the entire fitting process is automated and can include arbitrarily many admixture events simultaneously.
- Treemix has most problems with inferring the direction of gene flow and sometimes places migration edges to taxa that are closely related but not involved in the hybridization event
- Treemix should not be used as only source of information and may sometimes not work at all…

Alternatives to treemix

- MixMapper (Lipson et al., 2013, MBE): partially supervised, first scaffold tree building without the putatively admixed taxon/taxa, then fitting the user can specify to fit a likely admixed taxon as combinations of two taxa in the tree (runs in Matlab)
- Admixturegraph R package: Full flexibility (and most work). Admixturegraph uses a set of f4 statistics (e.g. from ADMIXTOOLS) to infer the likelihood for a user-specified admixture graph
- G-PhoCS (Generalized Phylogenetic Coalescent Sampler, Gronau et al., 2011, Nature Genetics): Demographic modeling software designed for modeling population splits and gene flow «bands»