

Detecting hybridization with ADMIXTOOLS

Patterson's D statistics to identify hybridization

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

A phylogenetic tree diagram is shown above the SNP data. It has four terminal nodes labeled P1, P2, P3, and P4 from left to right. Node P1 is at the bottom left, node P2 is above it to the right, node P3 is further to the right, and node P4 is at the bottom right. Node P2 is the root of the tree, indicated by a horizontal line segment connecting it to the other three nodes.

	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: $D = (ABBA - BABA) / (BABA + ABBA)$

A phylogenetic tree diagram with four leaves labeled P1, P2, P3, and P4. The root splits into two branches, one leading to P1 and P2, and the other to P3 and P4. P1 and P2 are sister taxa, as are P3 and P4.

	P1	P2	P3	P4	
T	T	A	A		
G	G	G	T		
C	T	T	T		
C	A	C	A		BABA
A	T	T	A		ABBA
C	G	G	C		ABBA

Concordant SNPs

ABBA

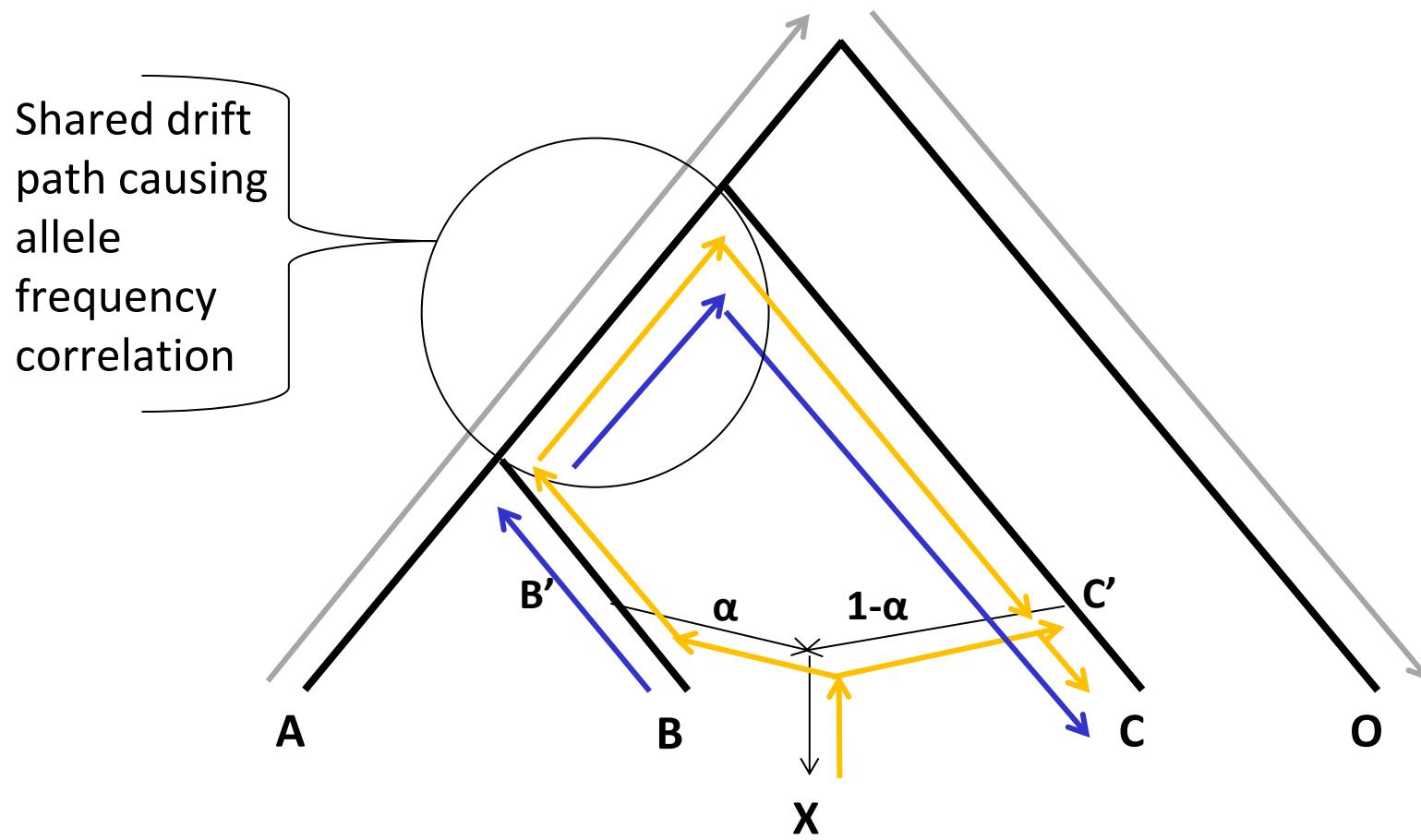
BABA

ABBA

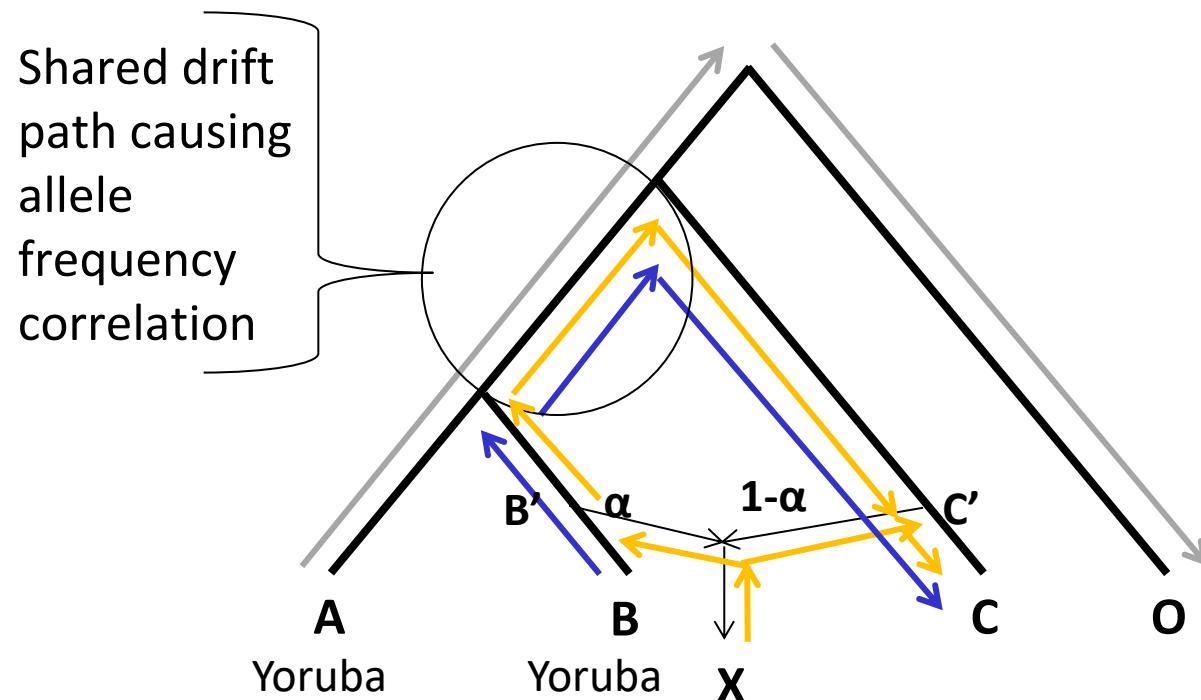
$$D = \frac{BABA - ABBA}{ABBA + BABA}$$

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

f4 ratio test to infer the amount of admixture



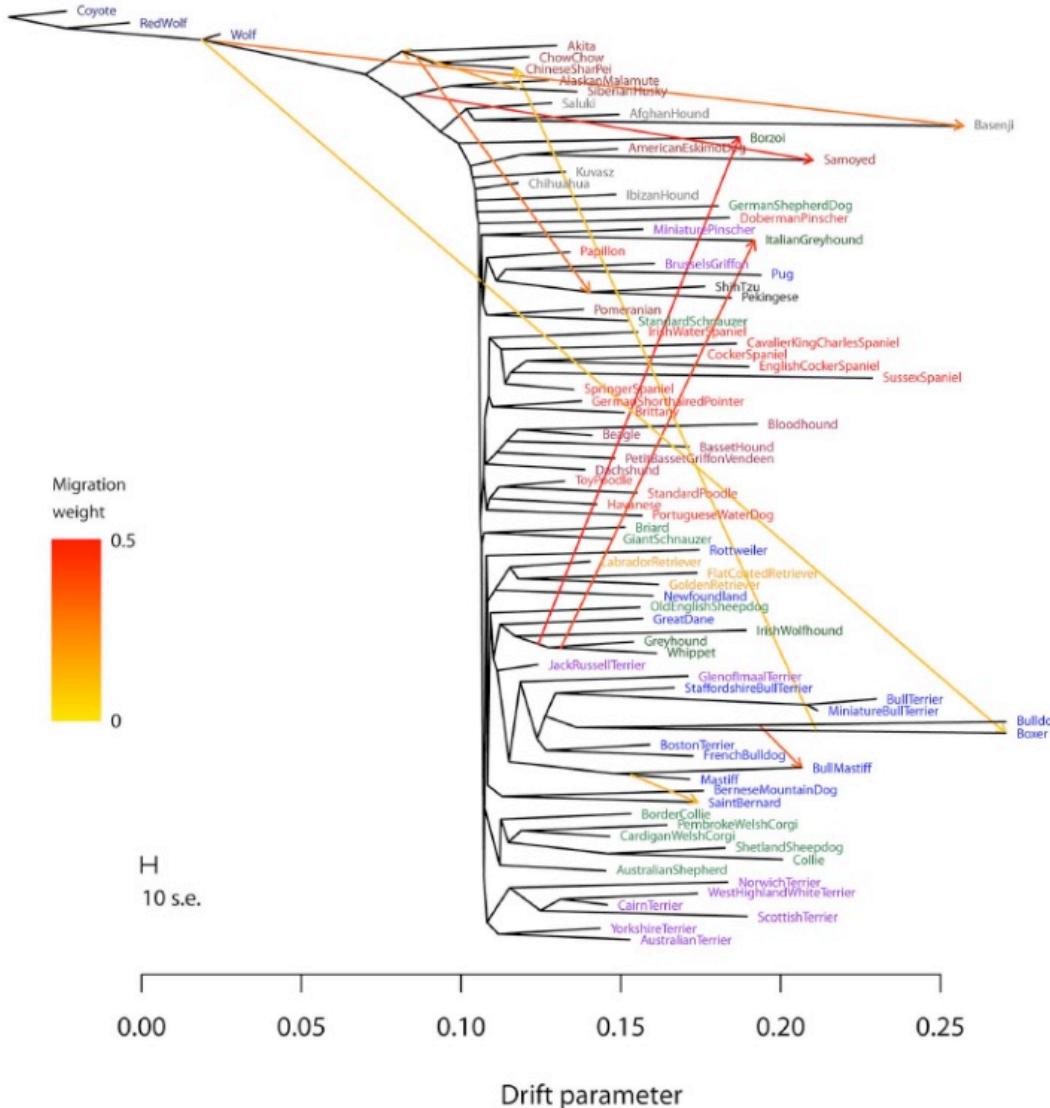
f4 ratio test to infer the amount of admixture



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

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

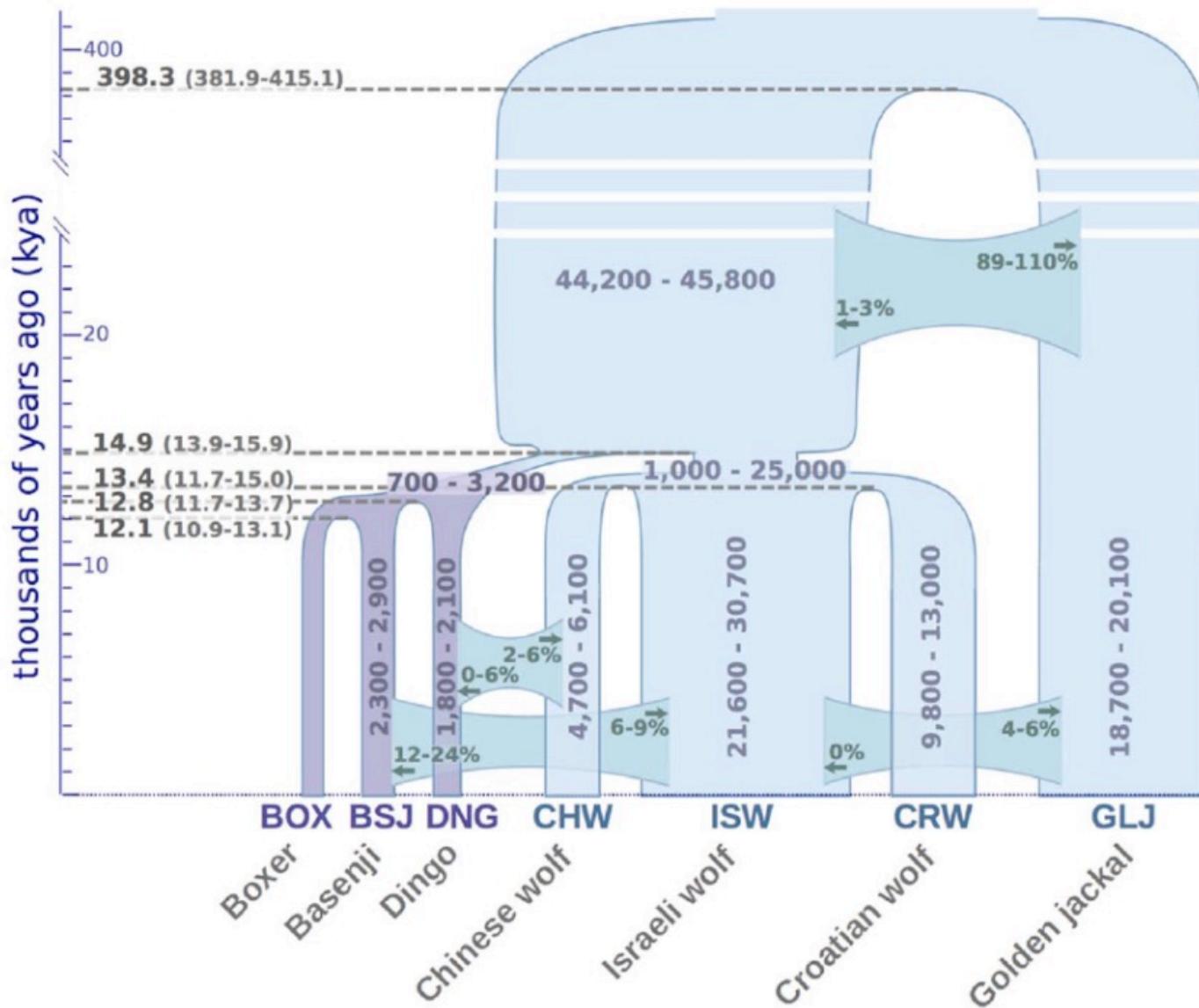
Treemix (Pickrell & Pritchard, 2012, Plos Genetics)



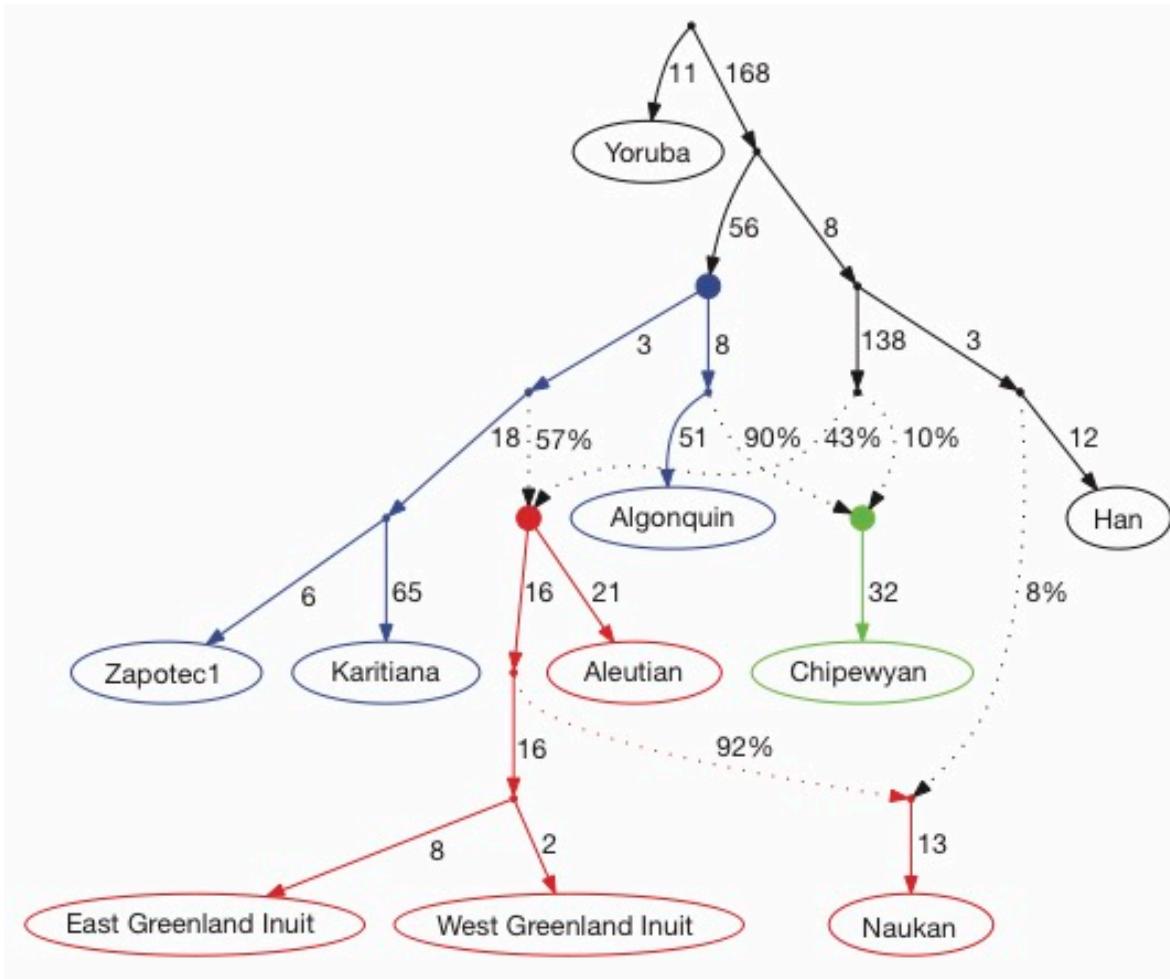
- Treemix uses covariances of allele frequencies to infer a maximum likely tree with a user-specified number of migration edges
- Gene flow edges are added sequentially to account for the greatest errors in the fit. This format makes TreeMix well-suited to handling very large trees: the entire fitting 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...

Freedman et al., 2014, PLoS Genetics: GPhoCS

Demographic modeling of population splits and gene flow «bands»



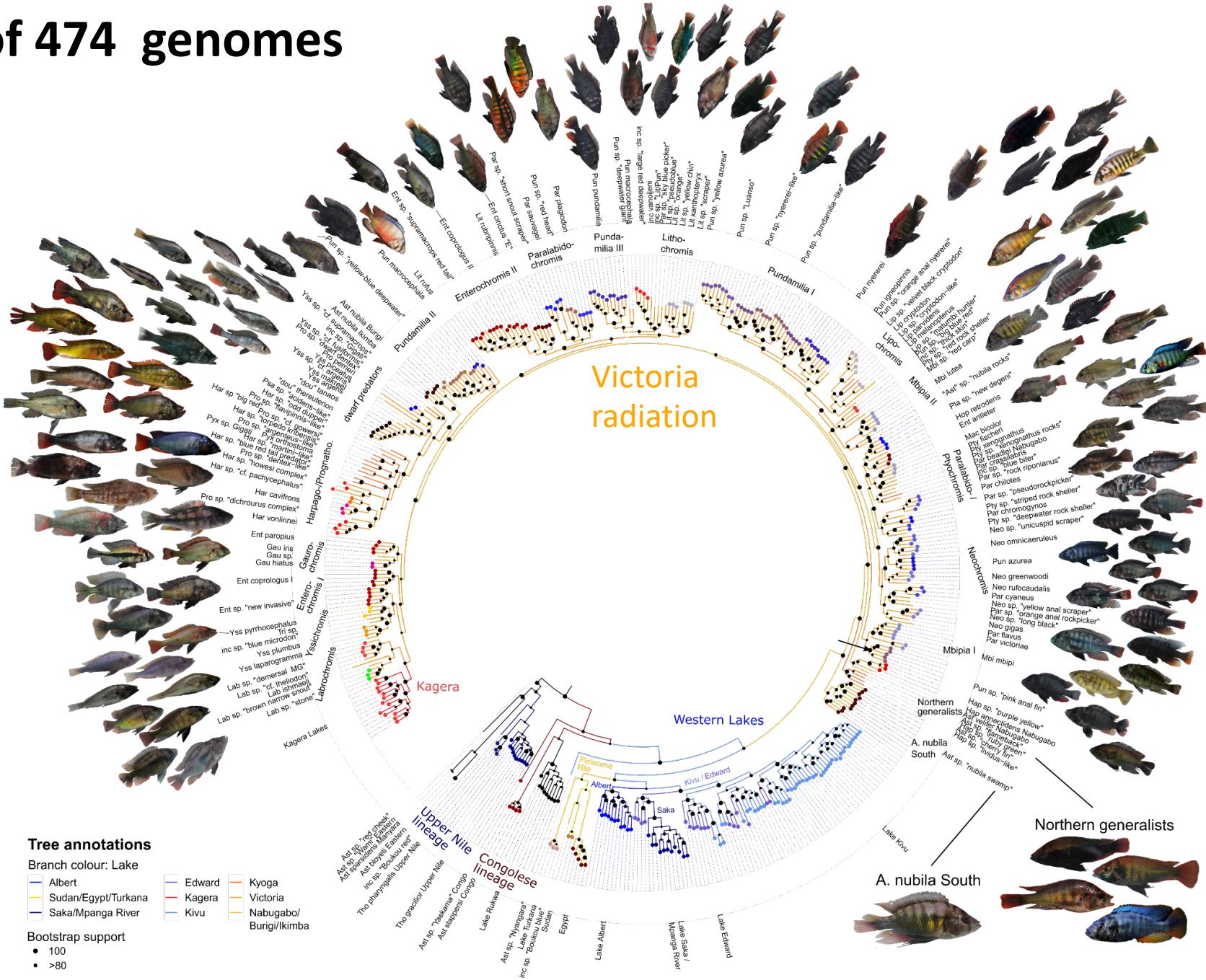
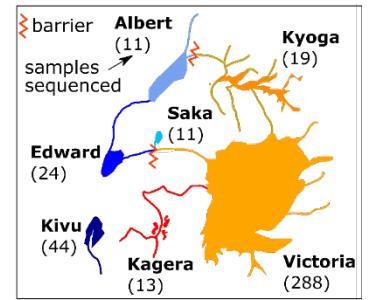
Admixture graphs (Reich et al. 2012)



New admixtools R package has a good implementation of admixture graphs but it is still quite badly documented and has some bugs.

<https://uqrmaie1.github.io/admixtools/articles/admixtools.html>

Phylogeny of 474 genomes



Putative hybrid origin of pelagic dwarf predators

fineSTRUCTURE (genetic similarity matrix)

