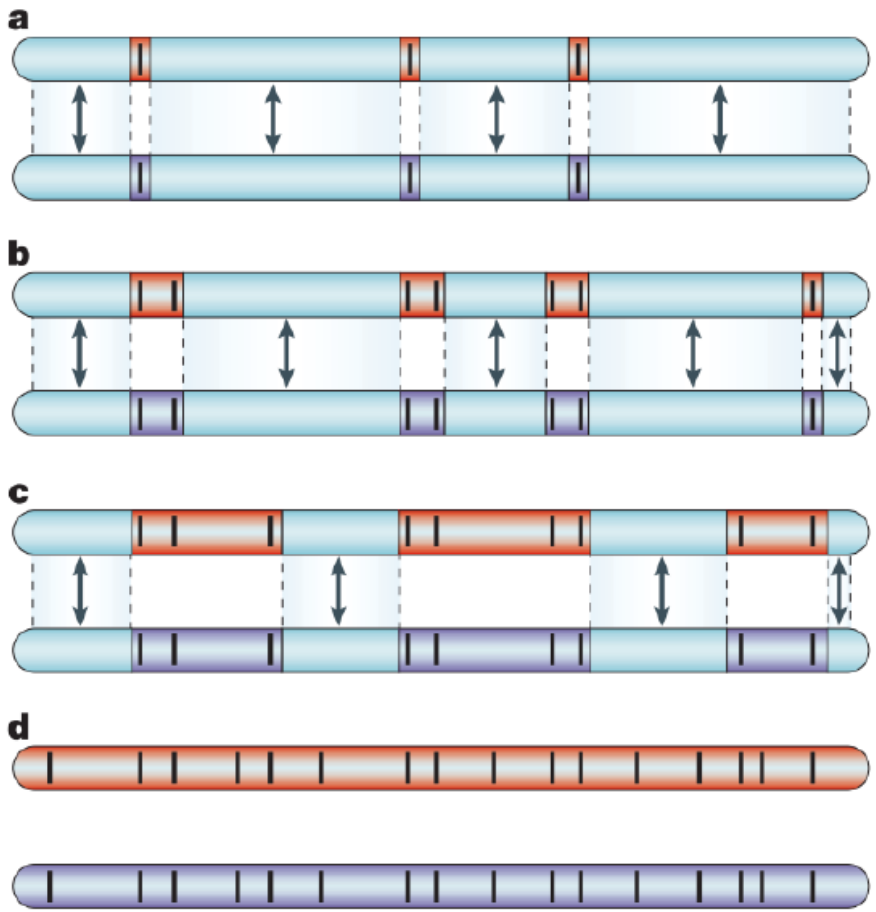
The genic concept of speciation



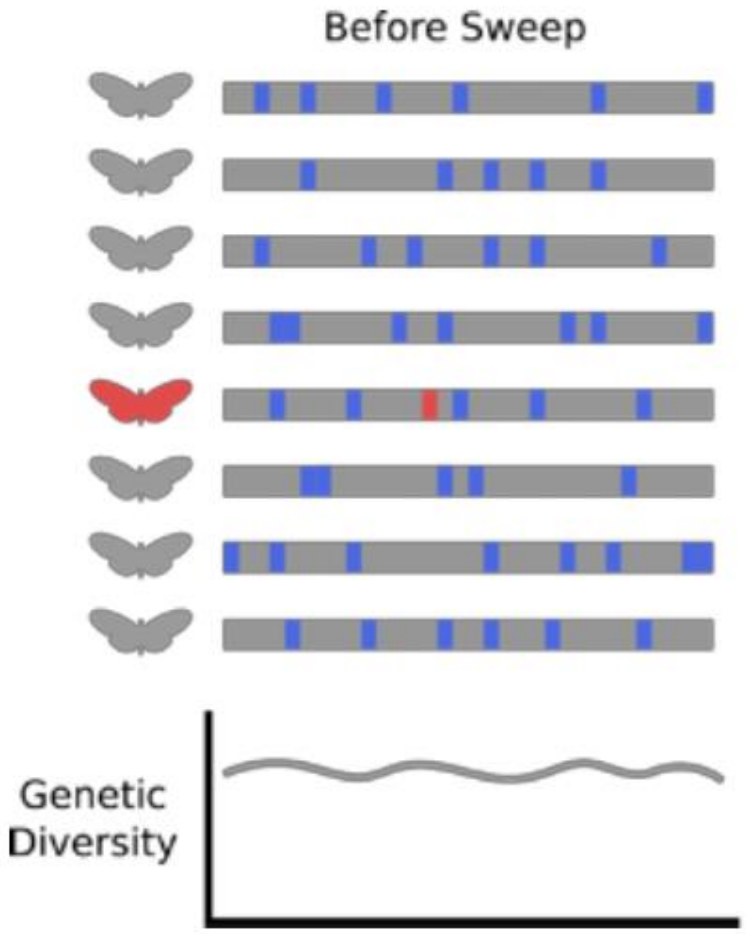
Divergent loci resist gene flow

Gene flow continues but linkage builds and divergent regions grow

Complete reproductive isolation evolves

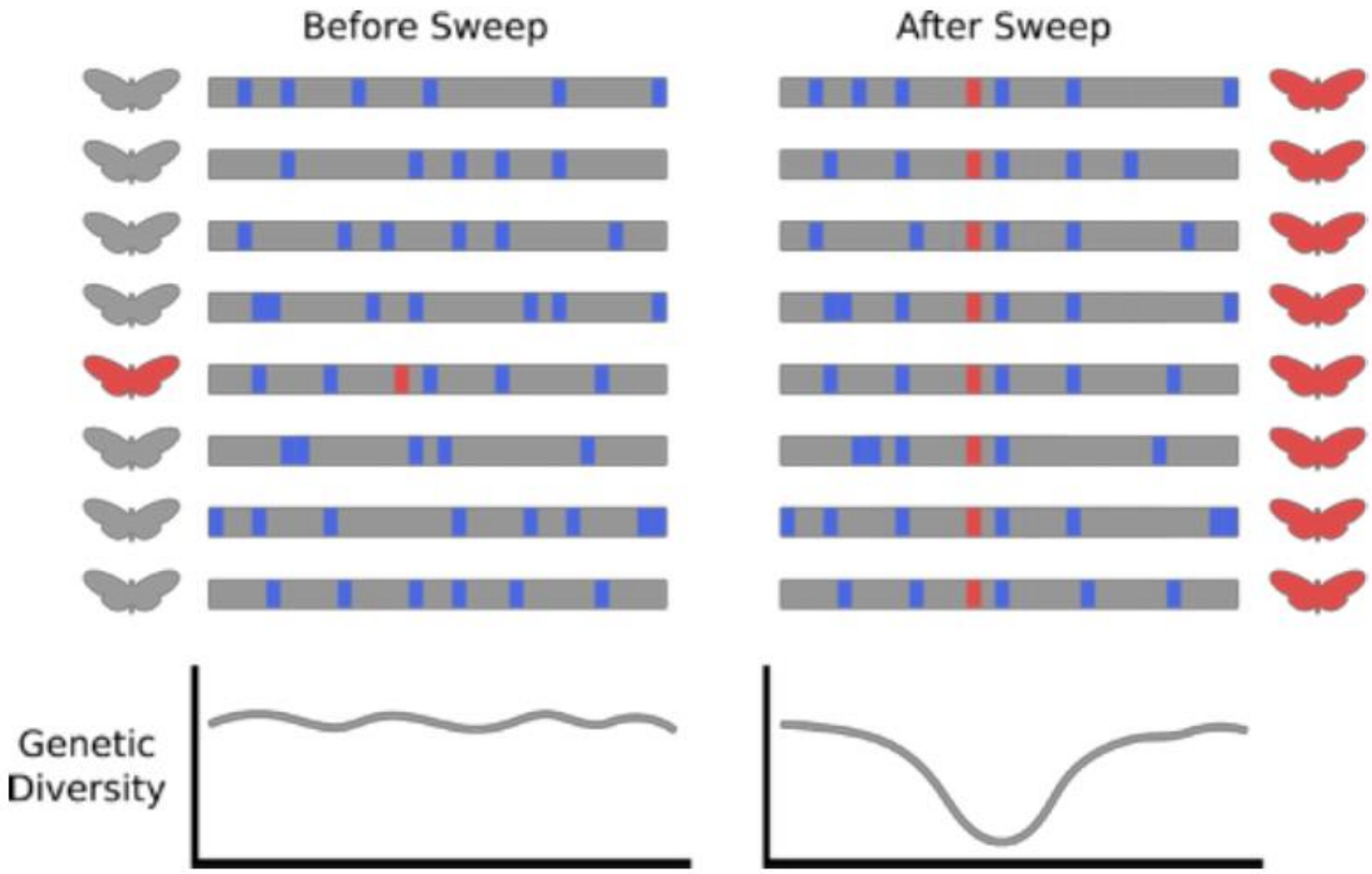
Wu (2001) **J Evo Bio**
Wu & Ting (2001) **Nat Rev Gen**

Selective sweep signatures



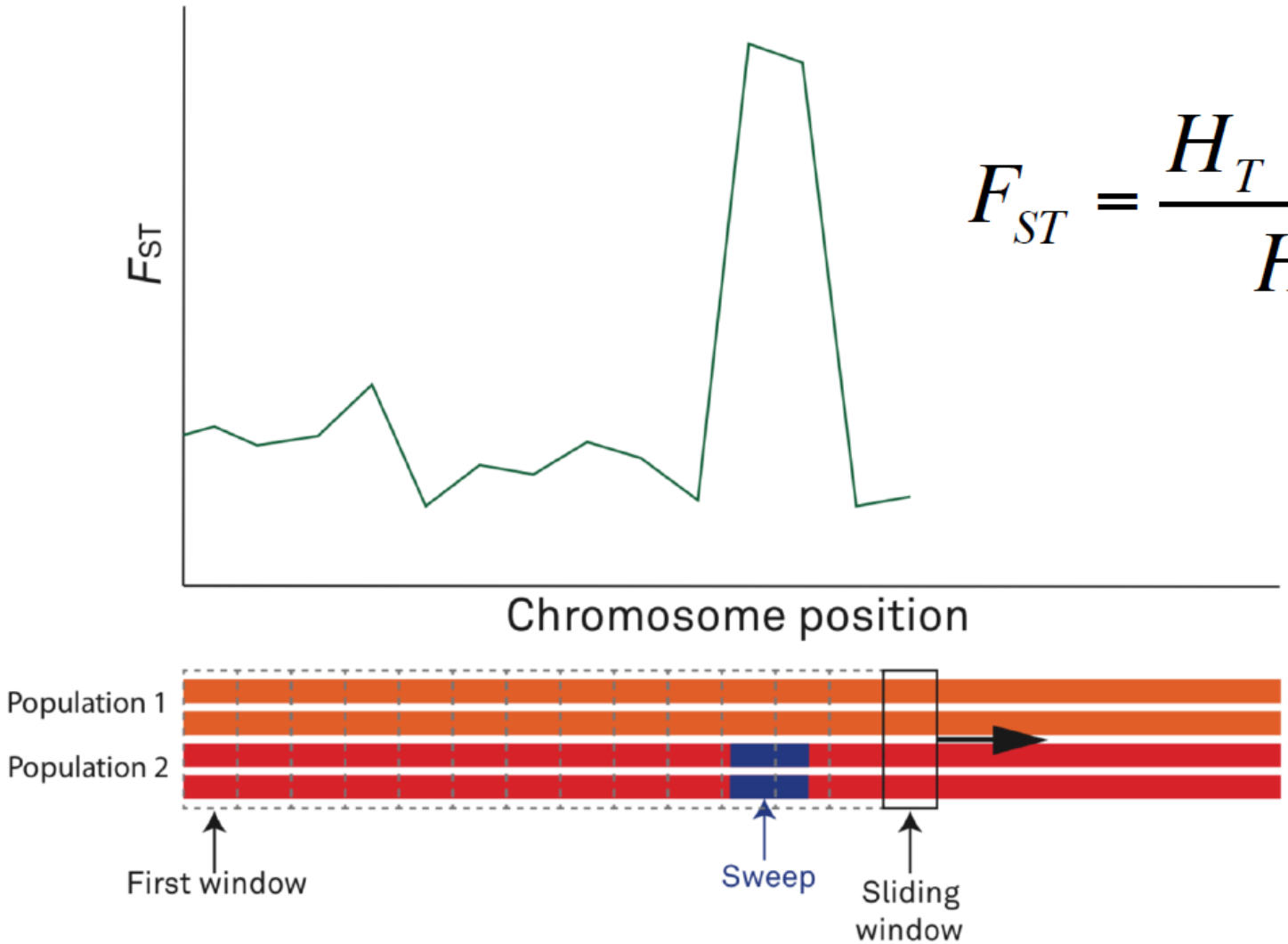
- **Increased genetic differentiation to another population**
- **Reduced genetic variation**
- **Increased haplotype length**

Selective sweep signatures



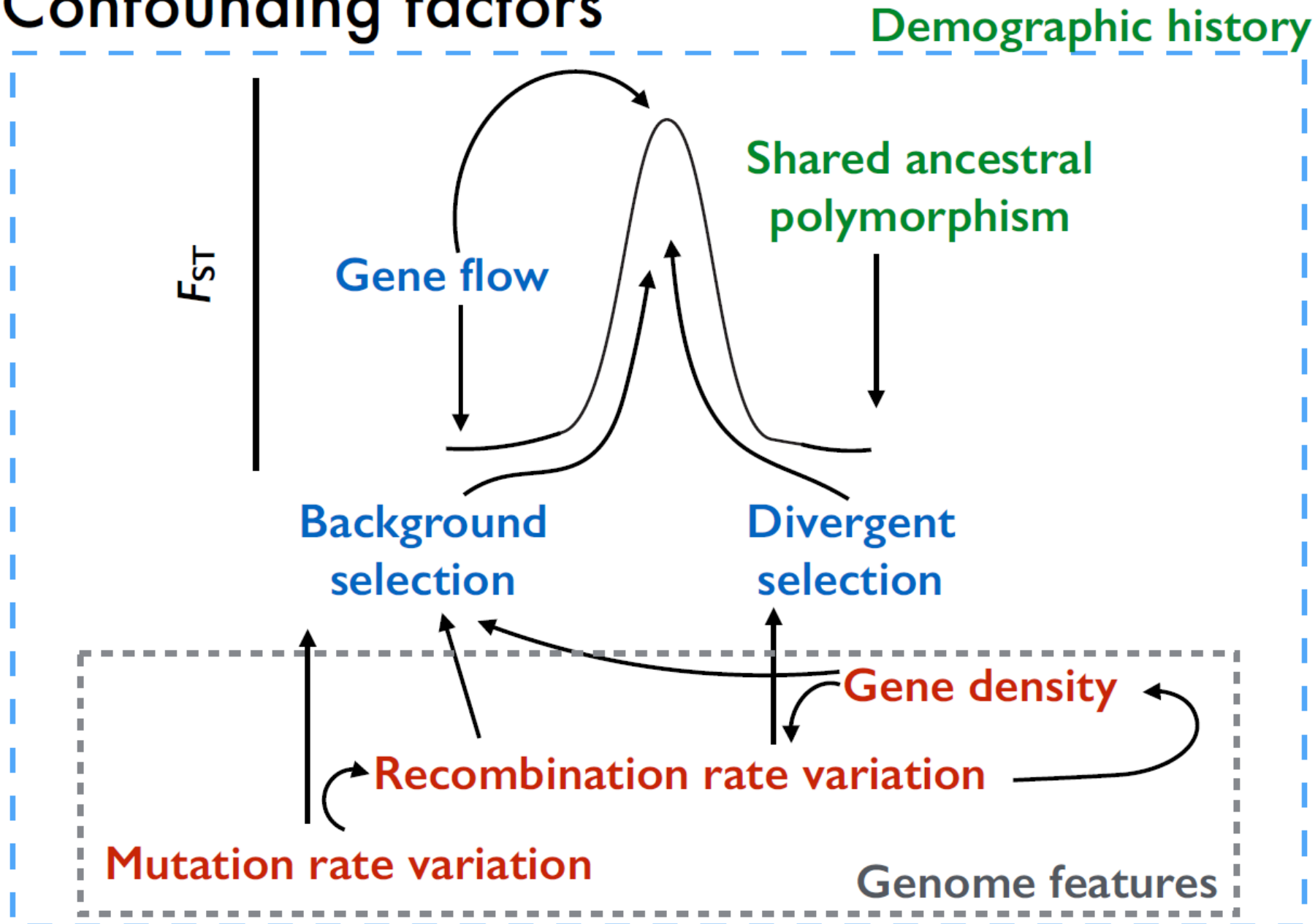
- **Increased genetic differentiation to another population**
- **Reduced genetic variation**
- **Increased haplotype length**

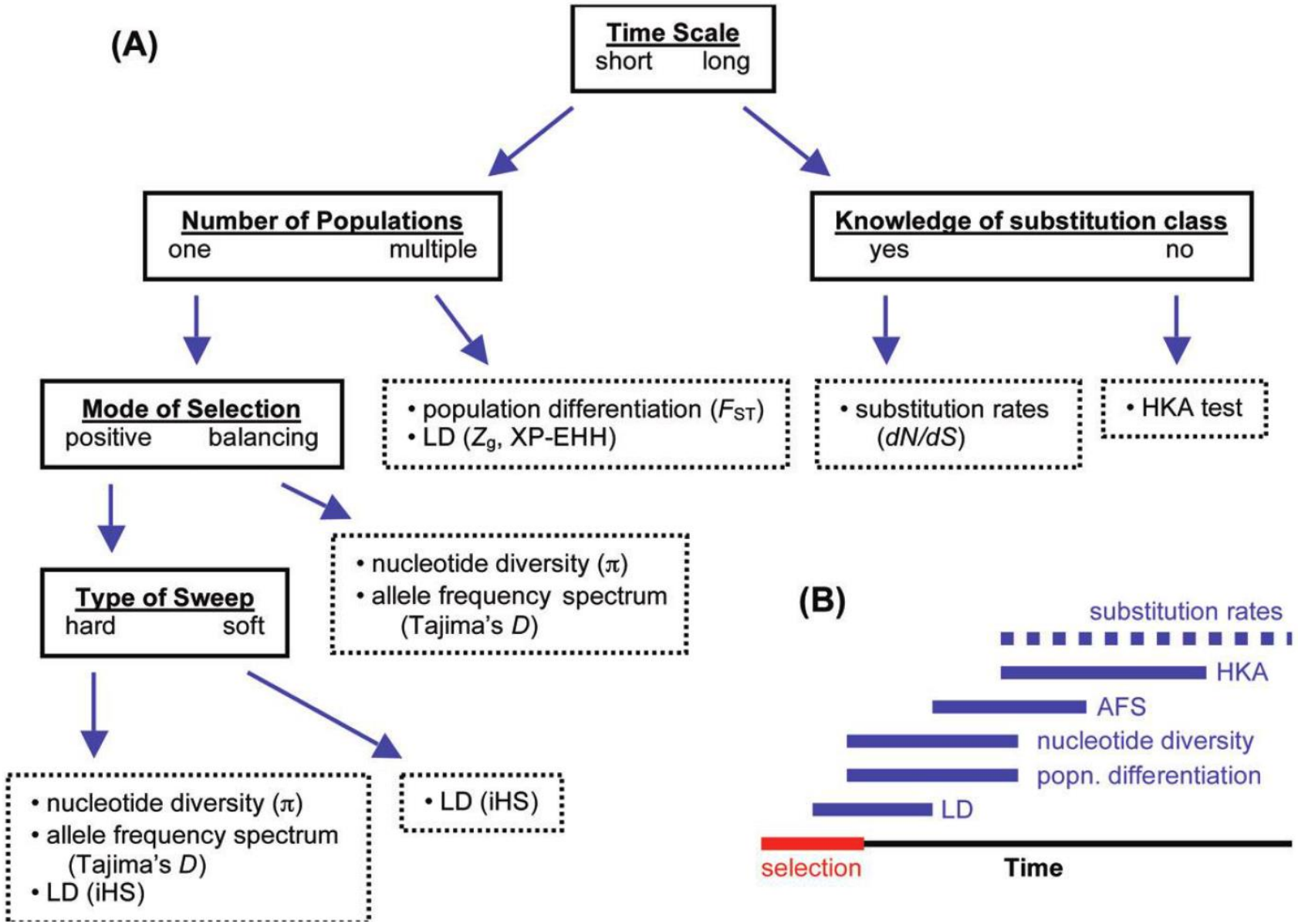
Sliding window estimates to detect selection



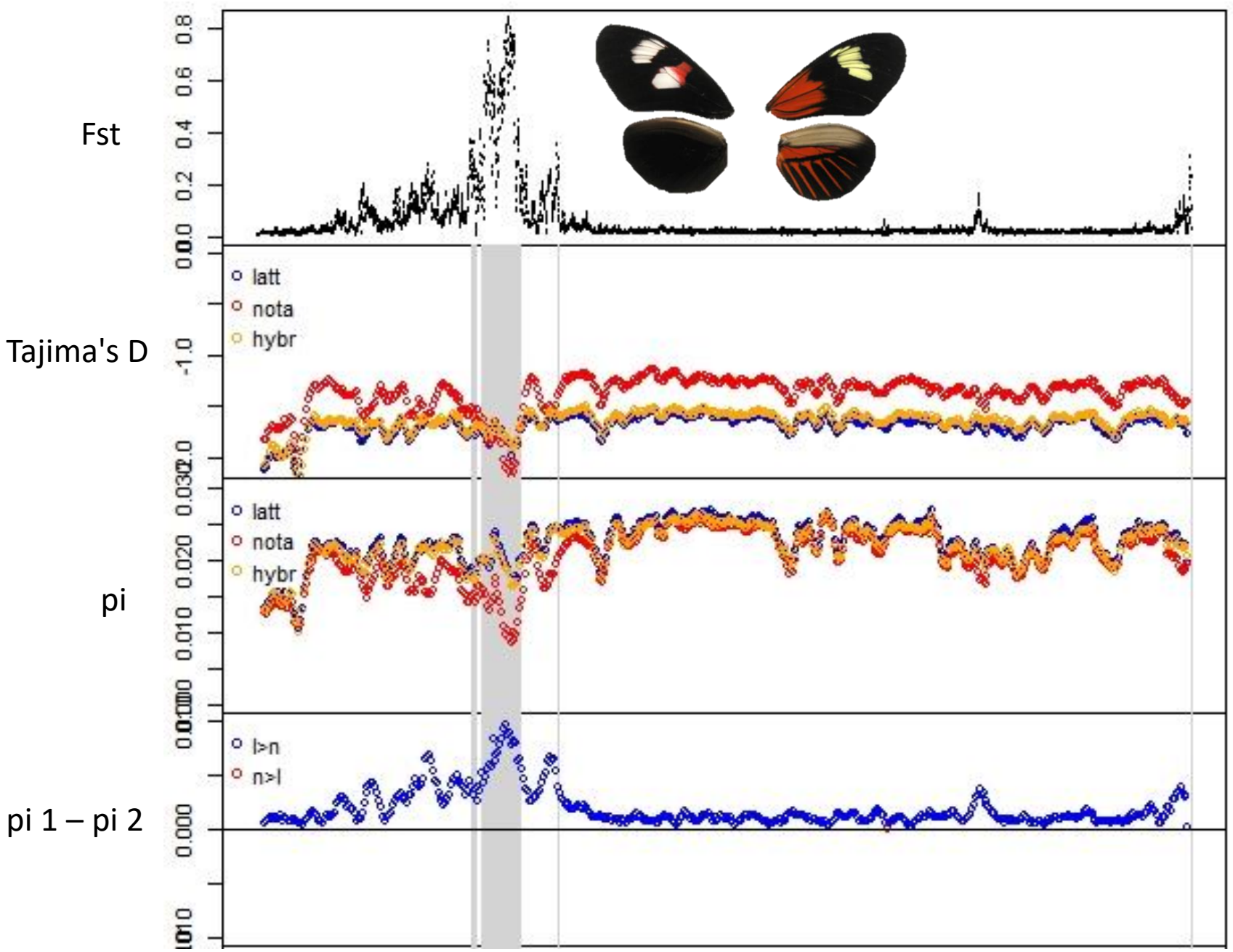
$$F_{ST} = \frac{H_T - H_S}{H_T}$$

Confounding factors





Hohenlohe et al. 2010 Int J Plant Sci



It is always best to combine Different approaches to identify genomic regions under divergent selection.

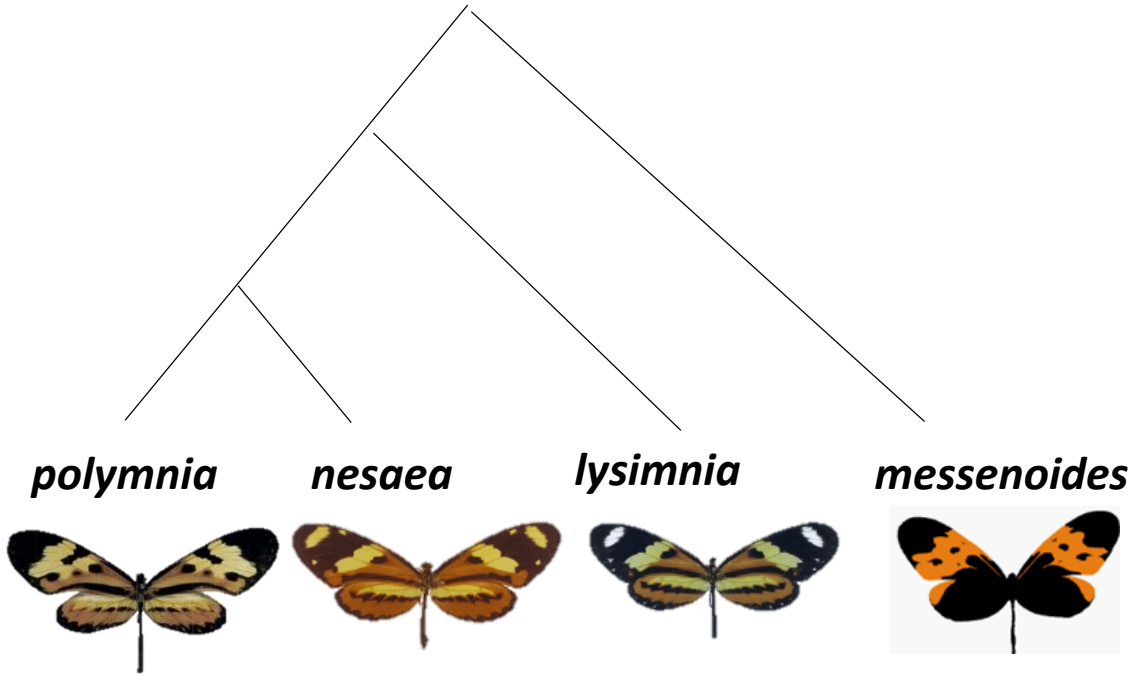
Meier et al. 2021, PNAS

Signatures and statistics to detect candidate barrier loci

- Locally restricted gene flow
  - Reduced f_d
  - Gmin
- Increased differentiation and potentially divergence
  - Increased Fst
  - Inreased dxy
  - increased Δpi
- Selective sweep signals in one or both populations
  - Increased Haplotype Length, e.g. iHS and XP-EHH
  - Reduced pi
  - Negative Tajima’s D

We will now apply genome scans to our *Mechanitis* dataset

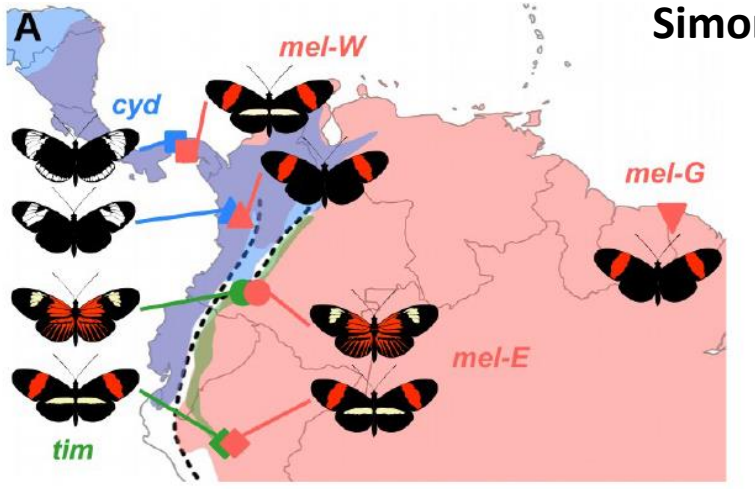
- Divergence (d_{xy}) and differentiation (F_{ST}) between *M. polymnia* and *M. nesaea*
- Introgression (f_d) between *M. lysimnia* and *M. neseaea*



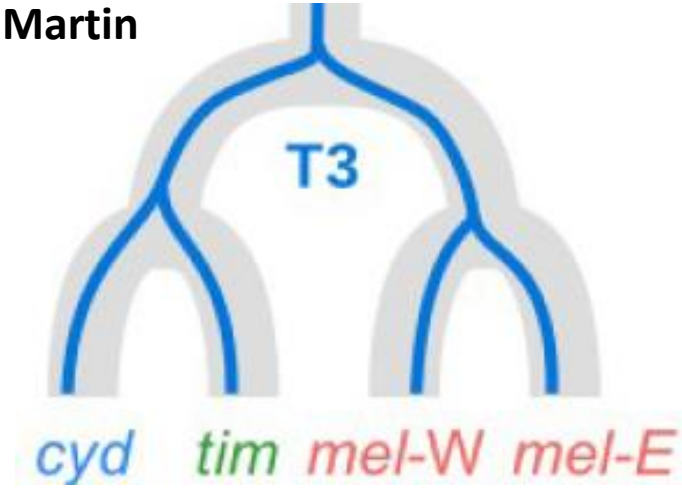
Supplementary slides for those interested

Detecting regions under divergent selection and barriers to gene flow

- If high gene flow -> $\mathbf{F_{ST}}$ is a good measure for detecting regions under divergent selection or barrier loci
- If the taxa are divergent enough -> $\mathbf{d_{xy}}$ may work best, particularly if levels of gene flow are not very high and in cases of secondary contact Ideally correct for differences in pi with an outgroup.
- If the taxa are very young and gene flow is not very high, $\mathbf{f_d}$ or TWISST might help if allopatric and parapatric populations were sequenced



Simon Martin



Detecting regions under divergent selection and barriers to gene flow

- If rates of gene flow between the two taxa compared is high, $\mathbf{F_{ST}}$ is a good measure for detecting regions under divergent selection or barrier loci
- If the taxa are divergent enough, $\mathbf{d_{xy}}$ may work best, particularly if levels of gene flow are not very high and in cases of secondary contact. Ideally correct for differences in pi with an outgroup.
- If the taxa are very young and gene flow is not very high, $\mathbf{f_d}$ or TWISST might help if allopatric and parapatric populations were sequenced
- If there is no gene flow, it is better to search for signatures of selective sweeps (e.g. iHS, XP-EHH, Tajima's D). However, inferring if these regions are involved in speciation is difficult.

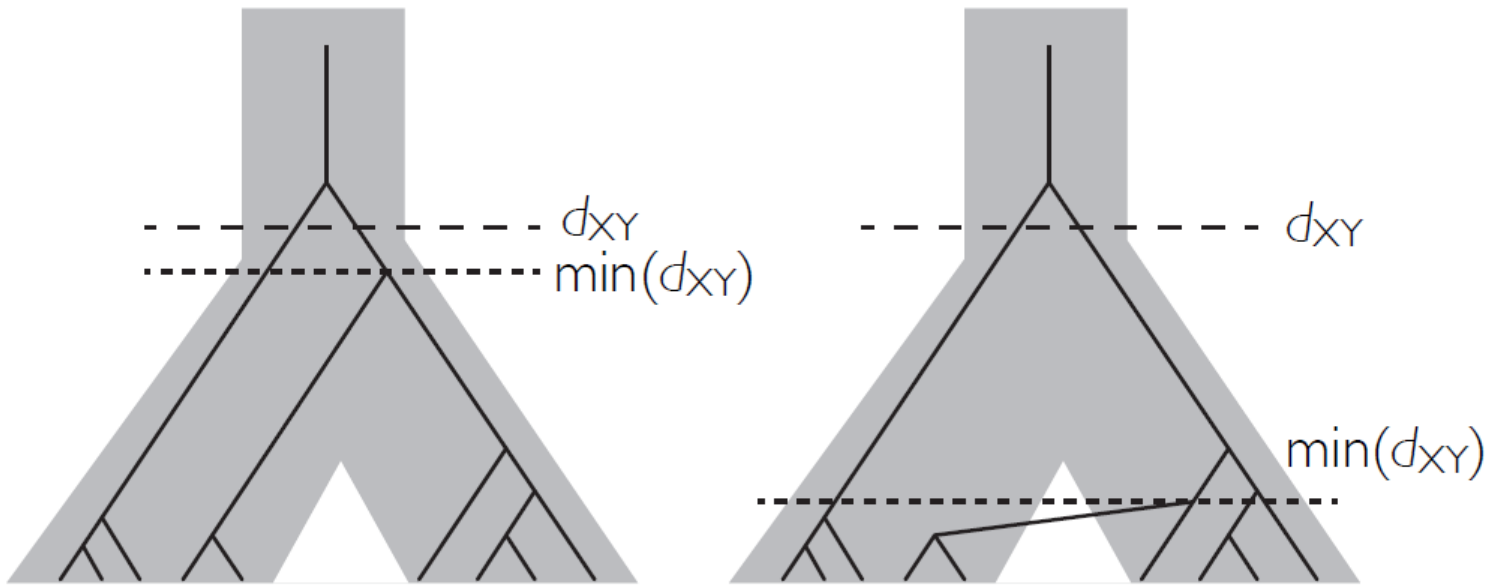
Absolute measures of divergence

$$d_{XY} = \sum_{ij} x_i y_j d_{ij}$$

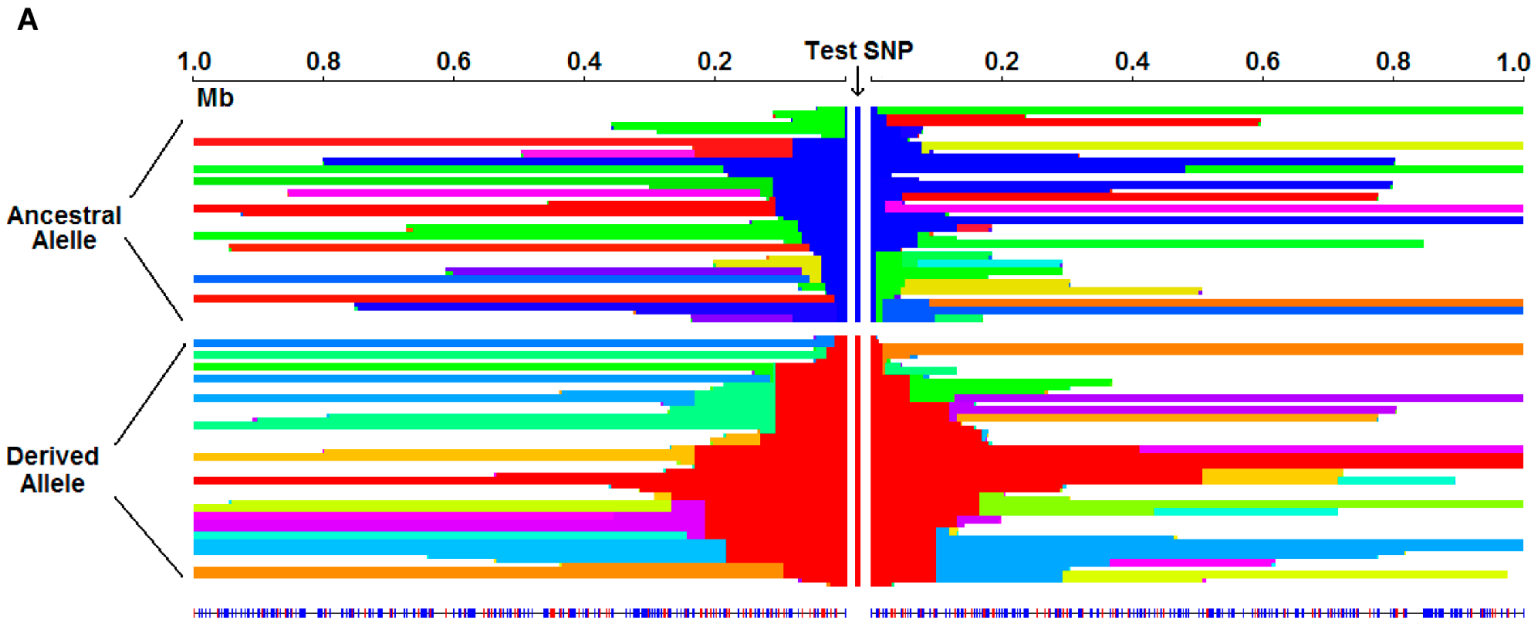
Average number of pairwise differences between two populations

Pop A	Pop B
ACTGTC	ATTAGC
ATTGTC	ACTGGC
ACTGTC	ACTAGC
ATTGTC	ATTAGC

Here d_{XY} is 0.375



iHS and XP-EHH



iHS: within a population

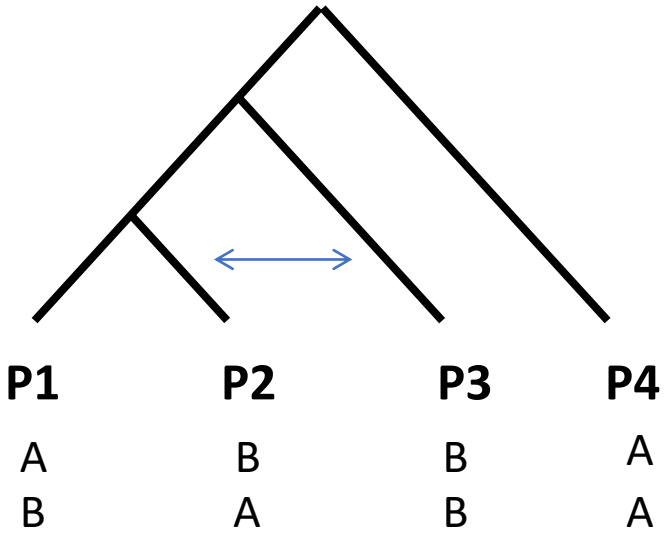
iHS (integrated haplotype score) compares haplotype lengths **within a population**
-> an allele under selection will lead to increased haplotype length relative to other haplotypes in the same region
-> useful to detect **ongoing/incomplete sweeps**

XP-EHH: beetween populations

XP-EHH (cross population extended haplotype homozygosity) compares haplotype lengths **between populations**
-> a population that had a sweep has increased haplotype lengths relative to the haplotypes in the other population in the same region
-> most powerful with **complete sweeps** restricted to one population

Sliding window introgression: f_d

f_d can be applied to smaller number of ABBA and BABA sites than D and is thus ideal for sliding windows.
ABBA and BABA patterns are computed from allele frequencies and the f test of the four populations is standardized by the maximum value it could get which would be the scenario of complete mixing between P2 and P3. P2 and P3 are thus both set to PD which is the taxon with higher derived allele frequency of P2 and P3.



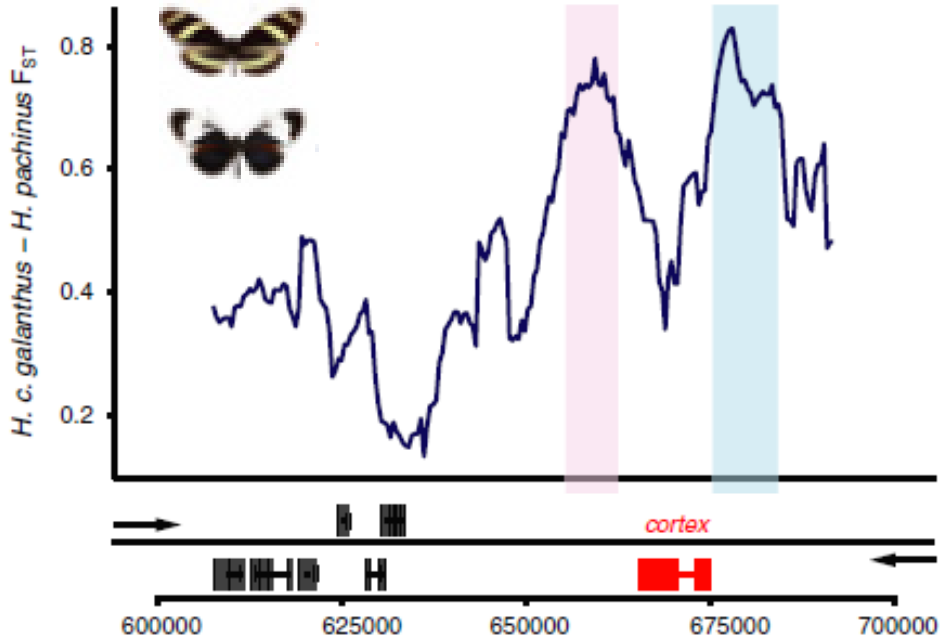
$$C_{ABBA}(i) = (1 - \hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3}(1 - \hat{p}_{i4})$$

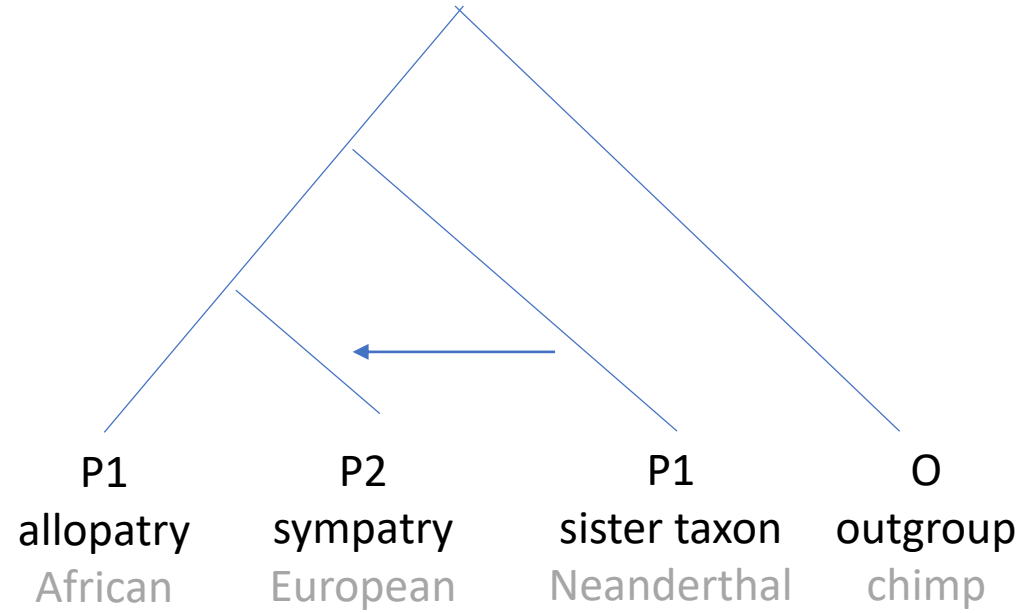
$$C_{BABA}(i) = \hat{p}_{i1}(1 - \hat{p}_{i2})\hat{p}_{i3}(1 - \hat{p}_{i4})$$

$$\hat{f}_d = \frac{S(P_1, P_2, P_3, O)}{S(P_1, P_D, P_D, O)}$$

PD=P2 or P3
(taxon with higher
derived allele frequency)

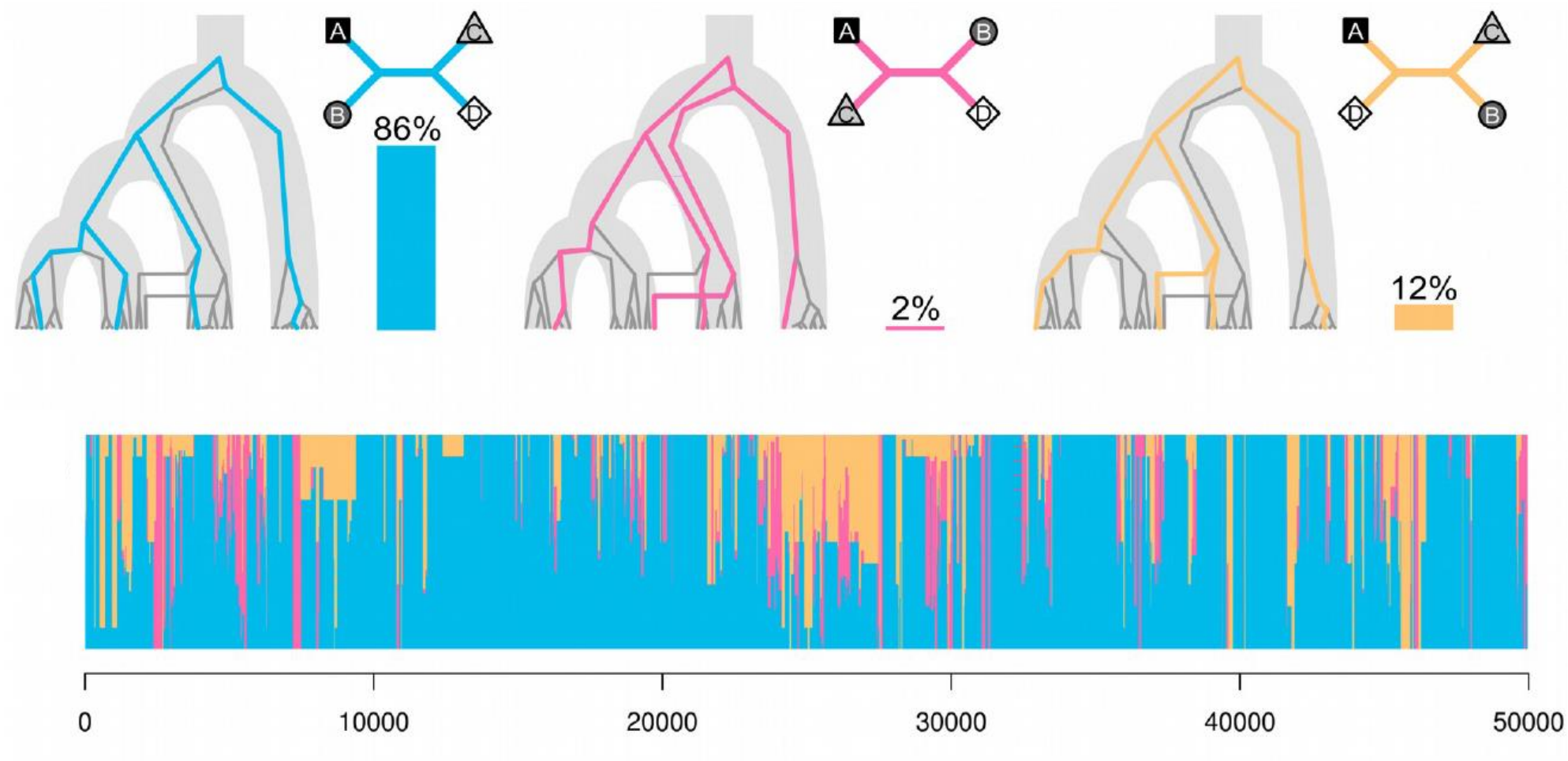
f_d can be used to find regions of reduced gene flow if allopatric and sympatric populations exist or alternatively, of adaptive introgression





Enciso-Romero, 2017, MolEcol

TWISST: Visualizing gene trees across the genome



See tutorial on https://github.com/simonhmartin/twisst

Martin & van Belleghem, 2017