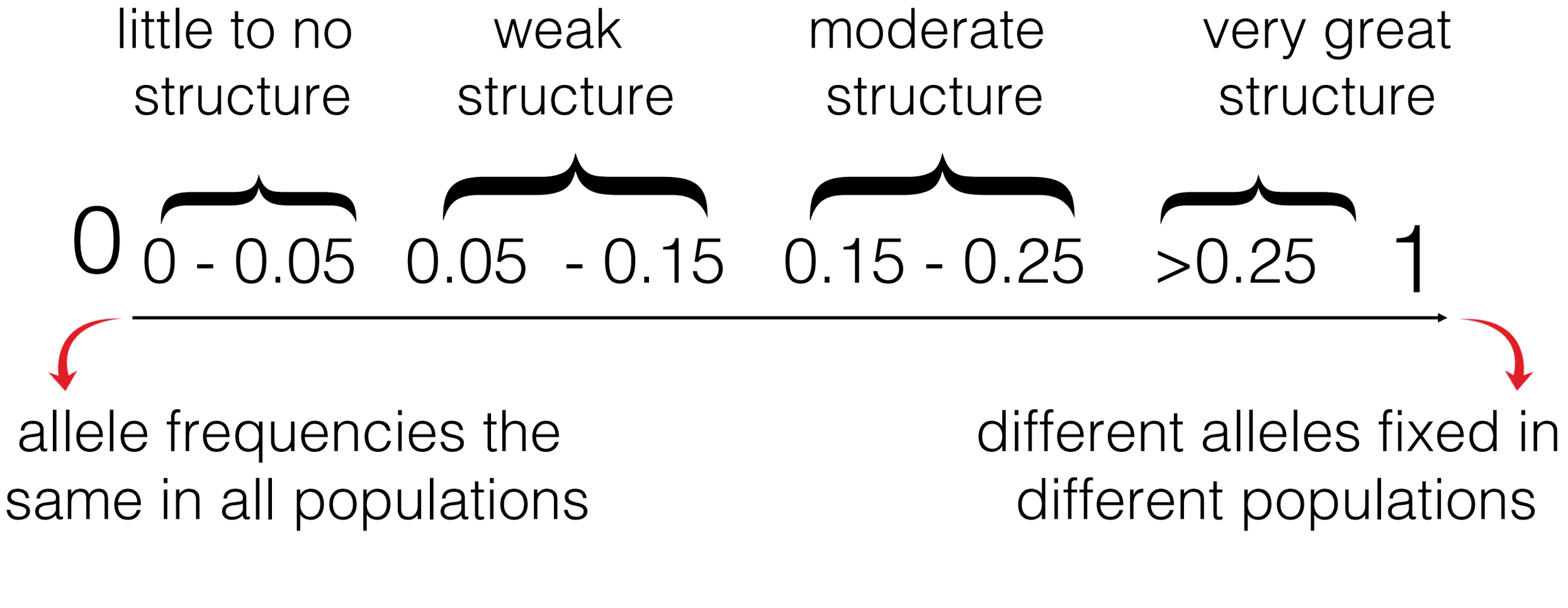
Quantifying population structure: beyond PCA

PCA	Dimension-reduction (no population model)	What are the major axes of genetic variation?
STRUCTURE MODEL (STRUCTURE, Admixture, Entropy)	Model-based ancestry inference	What proportion of each individual's genome comes from K ancestral populations?
SNMF	Sparse non-negative matrix factorization	Can we infer ancestry proportions like ADMIXTURE, but faster and with fewer assumptions?

F_{ST}: Quantifying Pop Subdivision/ genetic differentiation

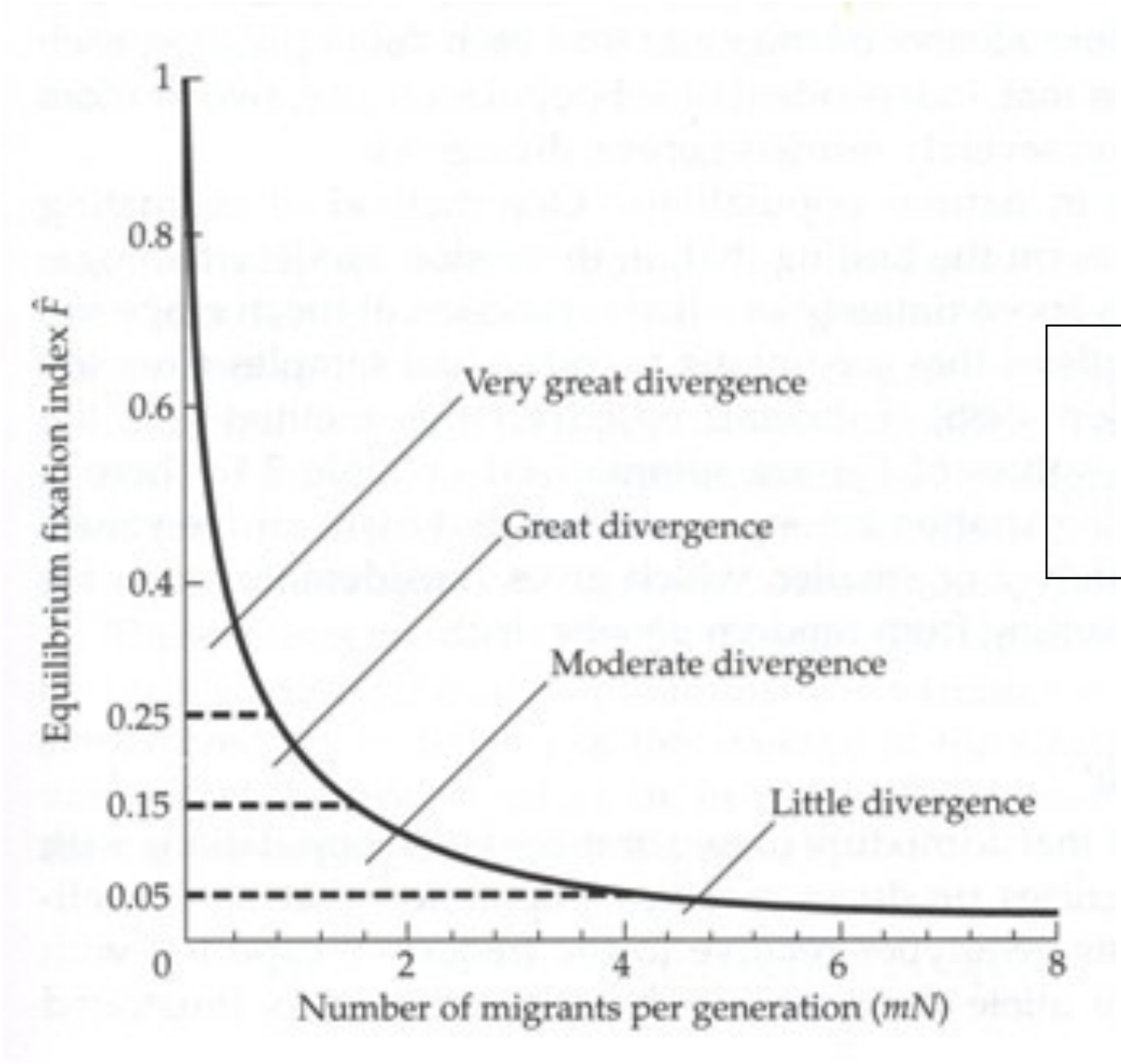
$$F_{ST} = \frac{H_T - H_S}{H_T}$$

$$F_{ST} = \frac{\text{variation among populations}}{\text{total variation}}$$



Migration has a huge effect on F_{ST}

F_{ST}



> 1-2 migrants per generation = NO differentiation!

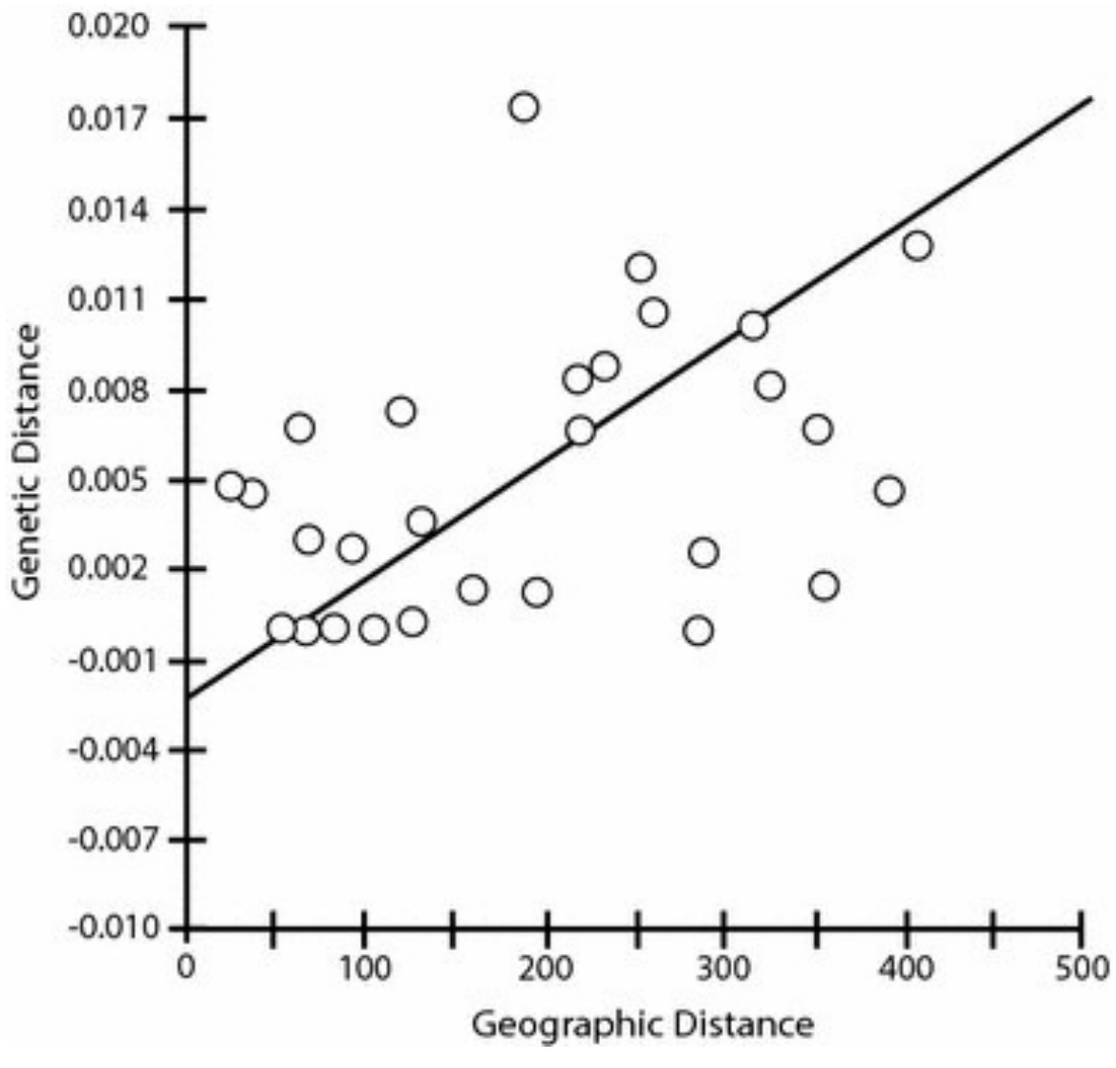
This result extends to many other models

Hartl 2000

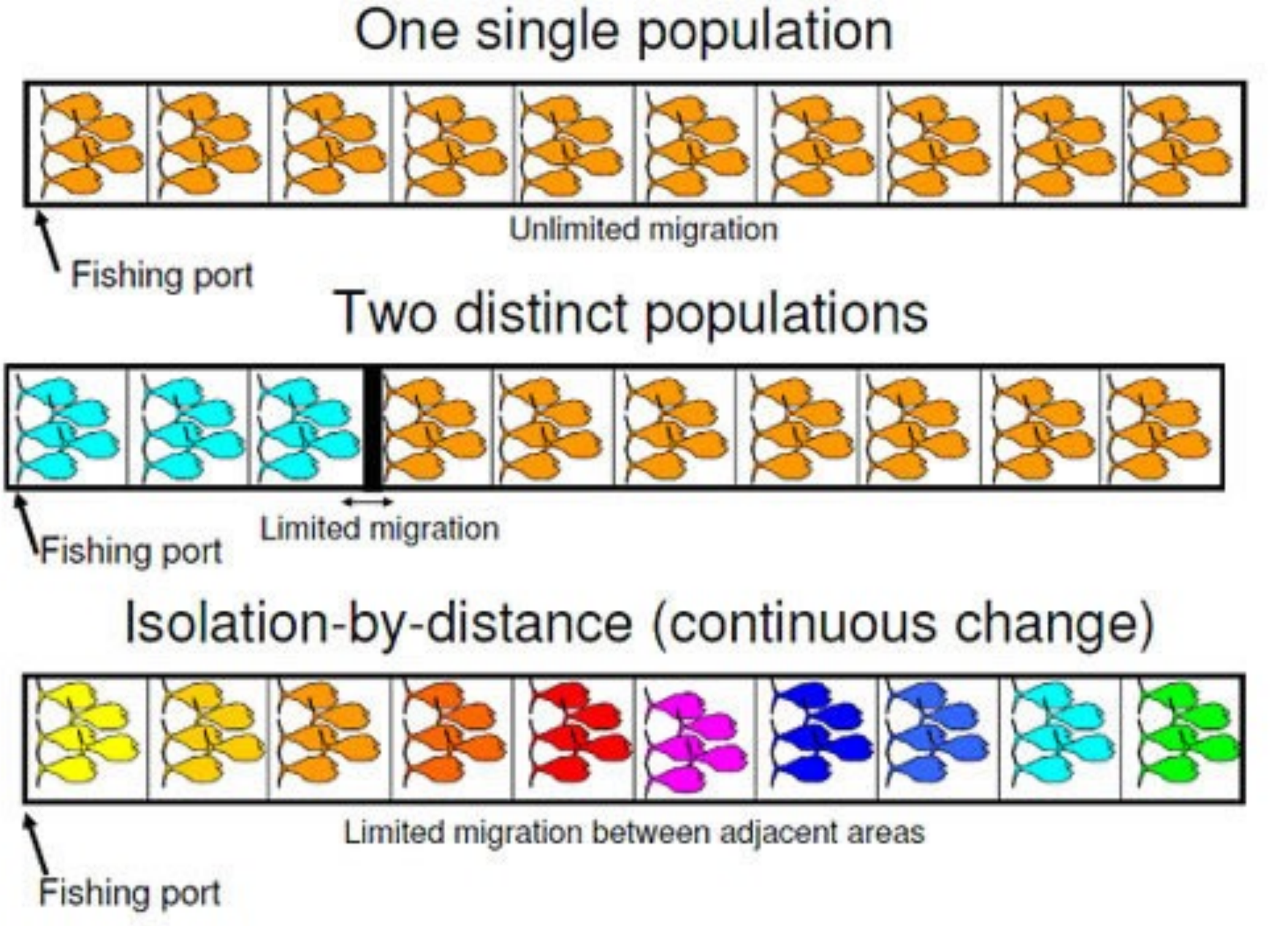
Isolation by Distance

The degree of population subdivision often increases with increasing geographic distance.

F_{ST}



Population structure and isolation by distance



“One very important aspect of all the FST-like statistics described above is that they are strongly influenced by within-subpopulation levels of variation (Charlesworth 1998; Jakobsson, Edge, and Rosenberg 2013). Because of this, we refer to them as relative measures of differentiation. In contrast, absolute measures of population differentiation are mostly independent of levels of within-population diversity; absolute measures are also known as genetic distances.”

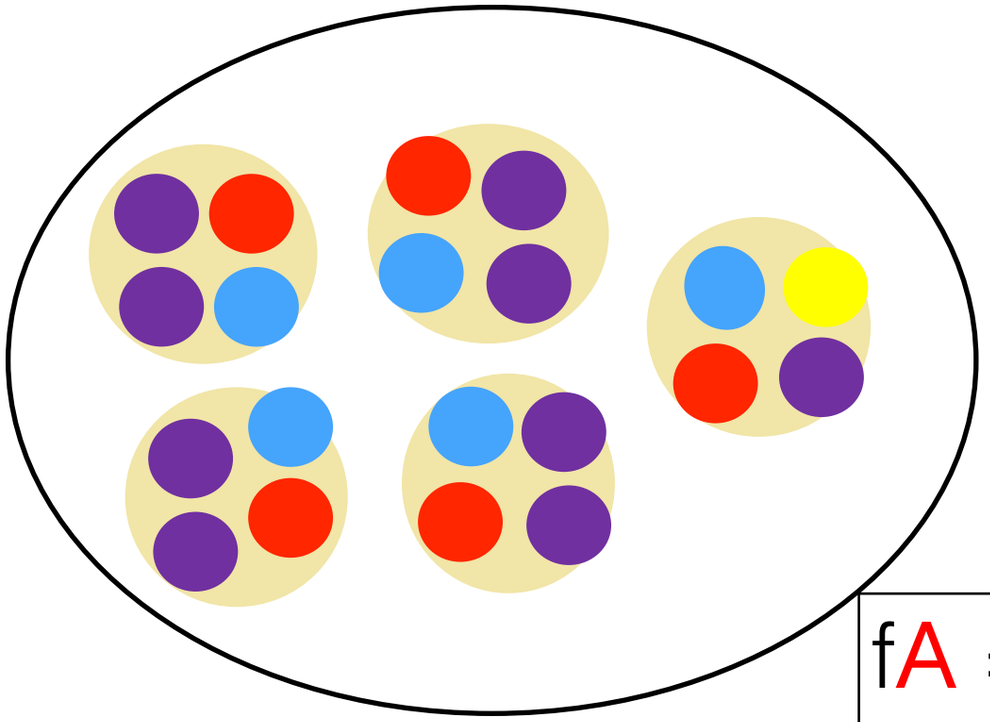
There are several different estimators of Fst and these are implemented in R in various packages and scripts.

Think about *how* calculating Fst makes sense in your situation — do you want genetic distances between populations? Between individuals? Across loci (as a means of detecting selection)?

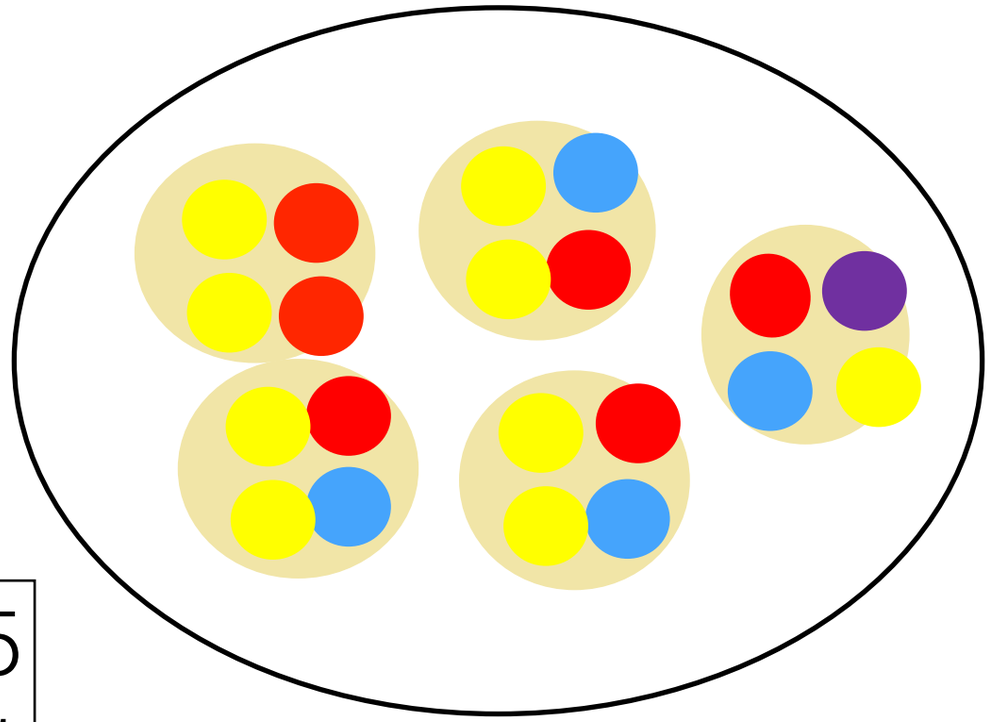
When accurately defining populations, we expect to maximize Hardy-Weinburg equilibrium

"...one way to find the best assignment of individuals to populations—or to find the most likely number of populations—is to attempt to minimize the amount of Hardy-Weinberg disequilibrium...The optimal assignment is then the configuration that results in the least amount of Hardy-Weinberg disequilibrium in each population."

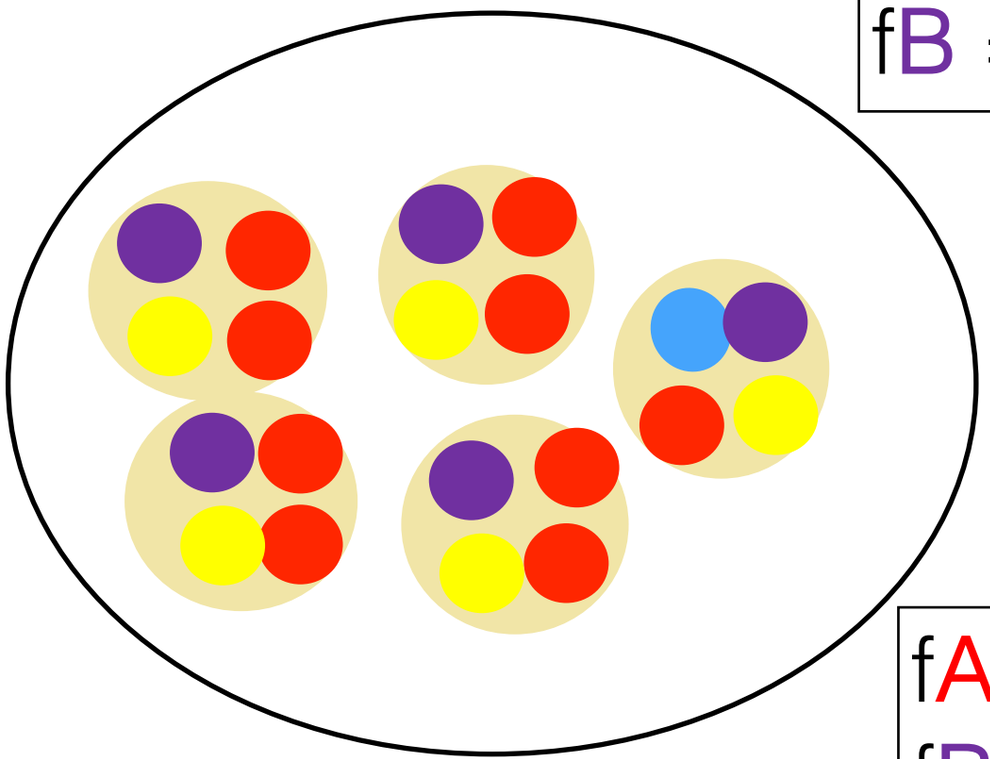
We can calculate the probability that an individual comes from a particular population based on its genotype, using differences in allele frequencies among populations



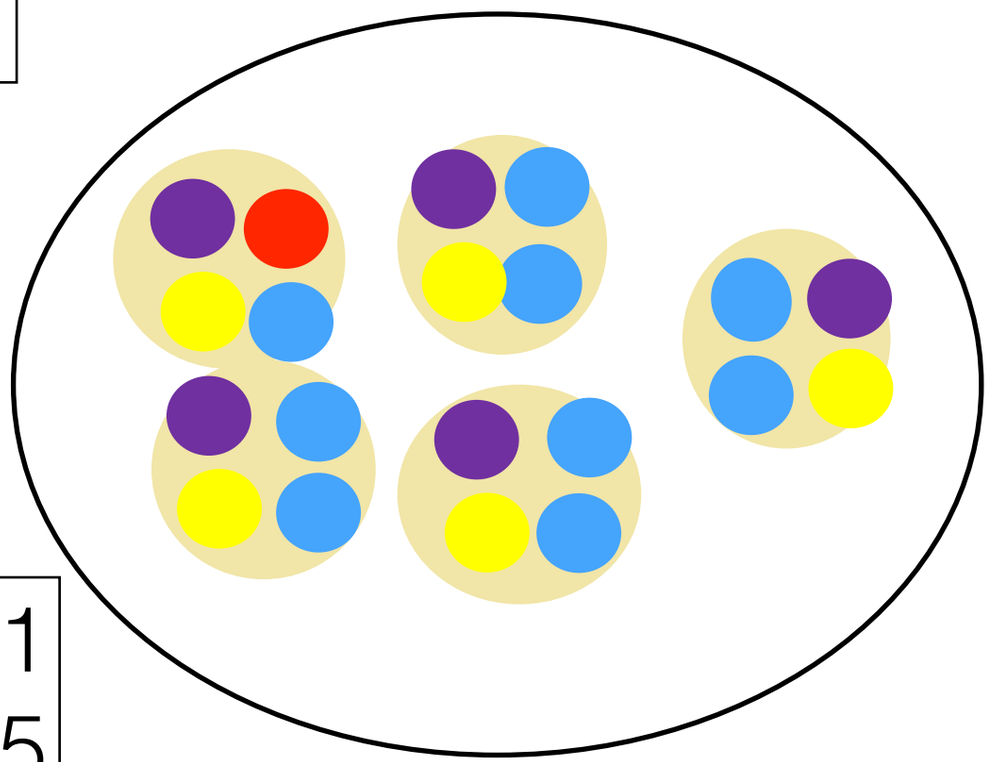
fA = 0.5
fB = 0.9



fA = 0.5
fB = 0.1



fA = 0.9
fB = 0.5



fA = 0.1
fB = 0.5

Inference of Population Structure Using Multilocus Genotype Data

Jonathan K. Pritchard, Matthew Stephens and Peter Donnelly

Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom

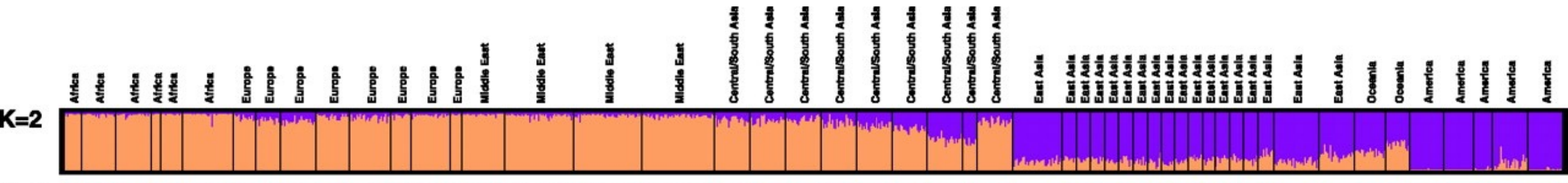
Manuscript received September 23, 1999
Accepted for publication February 18, 2000

ABSTRACT

We describe a model-based clustering method for using multilocus genotype data to infer population structure and assign individuals to populations. We assume a model in which there are K populations (where K may be unknown), each of which is characterized by a set of allele frequencies at each locus. Individuals in the sample are assigned (probabilistically) to populations, or jointly to two or more populations if their genotypes indicate that they are admixed. Our model does not assume a particular mutation process, and it can be applied to most of the commonly used genetic markers, provided that they are not closely linked. Applications of our method include demonstrating the presence of population structure, assigning individuals to populations, studying hybrid zones, and identifying migrants and admixed individuals. We show that the method can produce highly accurate assignments using modest numbers of loci—*e.g.*, seven microsatellite loci in an example using genotype data from an endangered bird species. The software used for this article is available from http://www.stats.ox.ac.uk/~pritch/home.html.

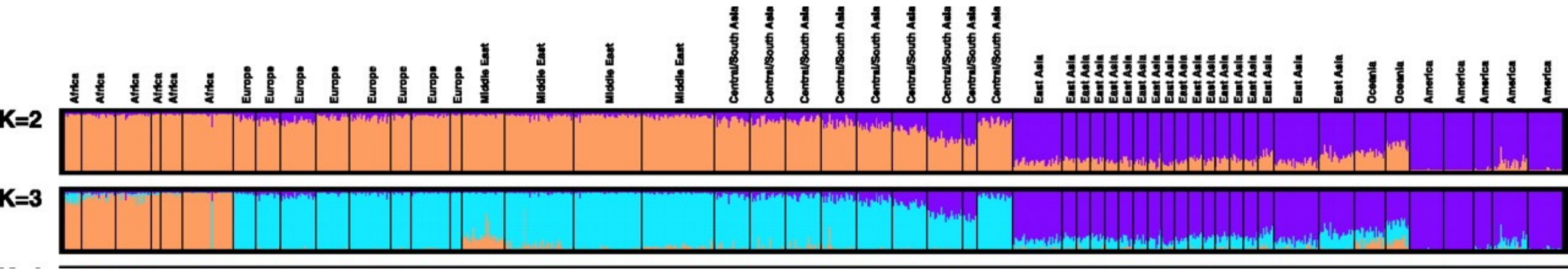
Prichard 2000 — cited 40,000+ times!

STRUCTURE: proportion of an individual's ancestry that comes from each of *k* populations



Pritchard et al 2000, Rosenberg et al 2002

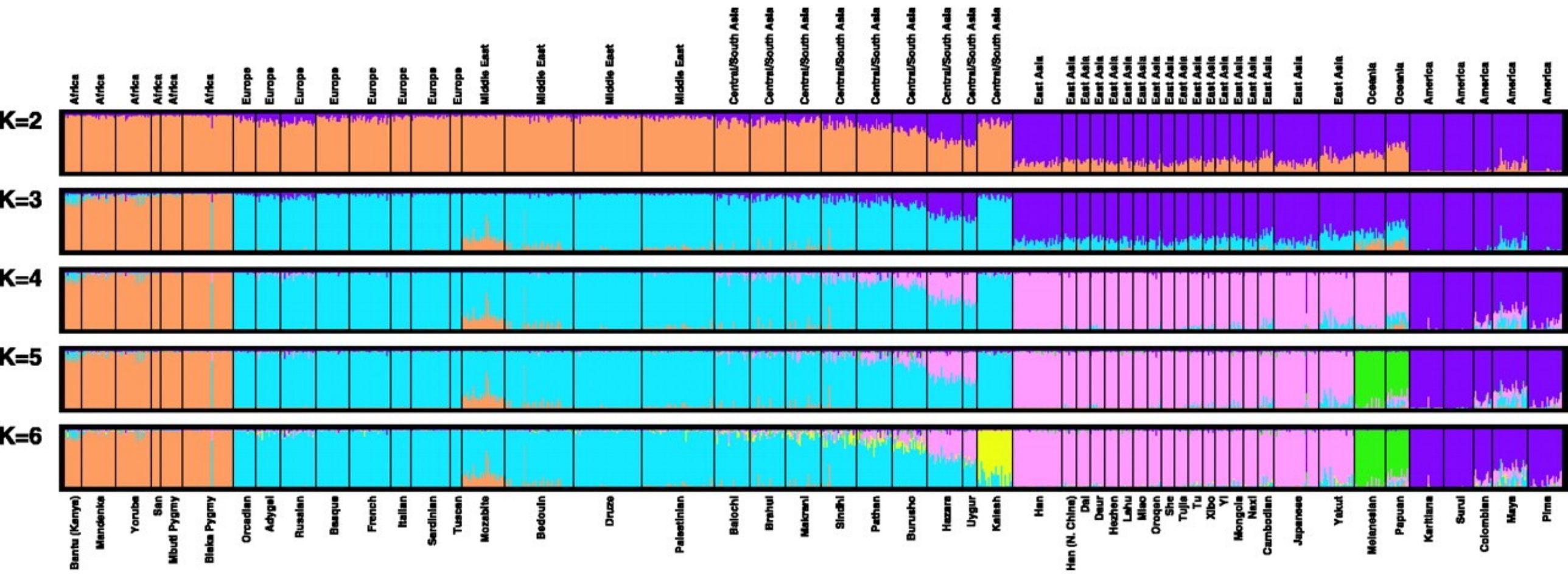
STRUCTURE: proportion of an individual's ancestry that comes from each of *k* populations



Increasing the value of K can reveal finer-scale population structure

Pritchard et al 2000, Rosenberg et al 2002

STRUCTURE: proportion of an individual's ancestry that comes from each of *k* populations

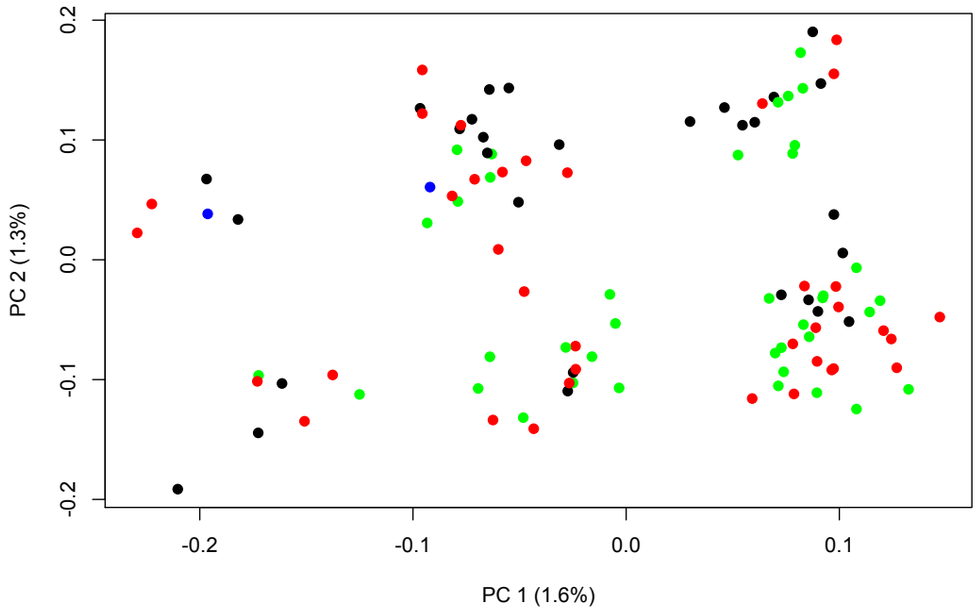


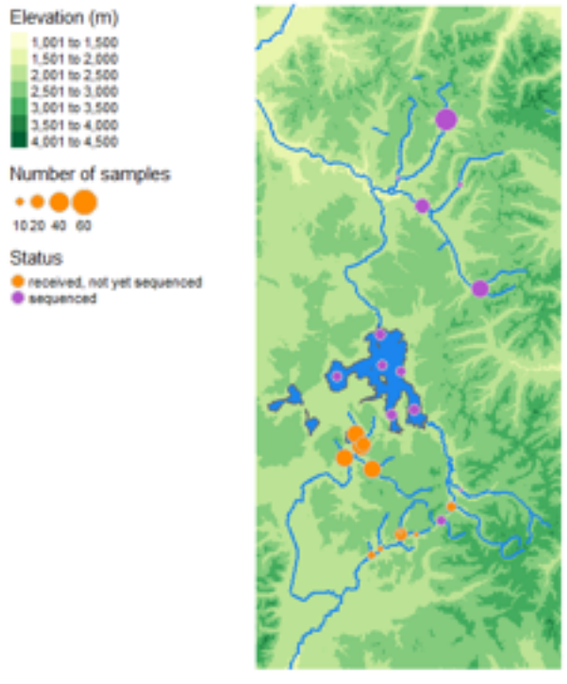
Increasing the value of K can reveal finer-scale population structure

Pritchard et al 2000, Rosenberg et al 2002

PCA is also useful!

"…Patterson, Price, and Reich (2006) showed that the number of significant eigenvalues in a PCA study is equal to K – 1."





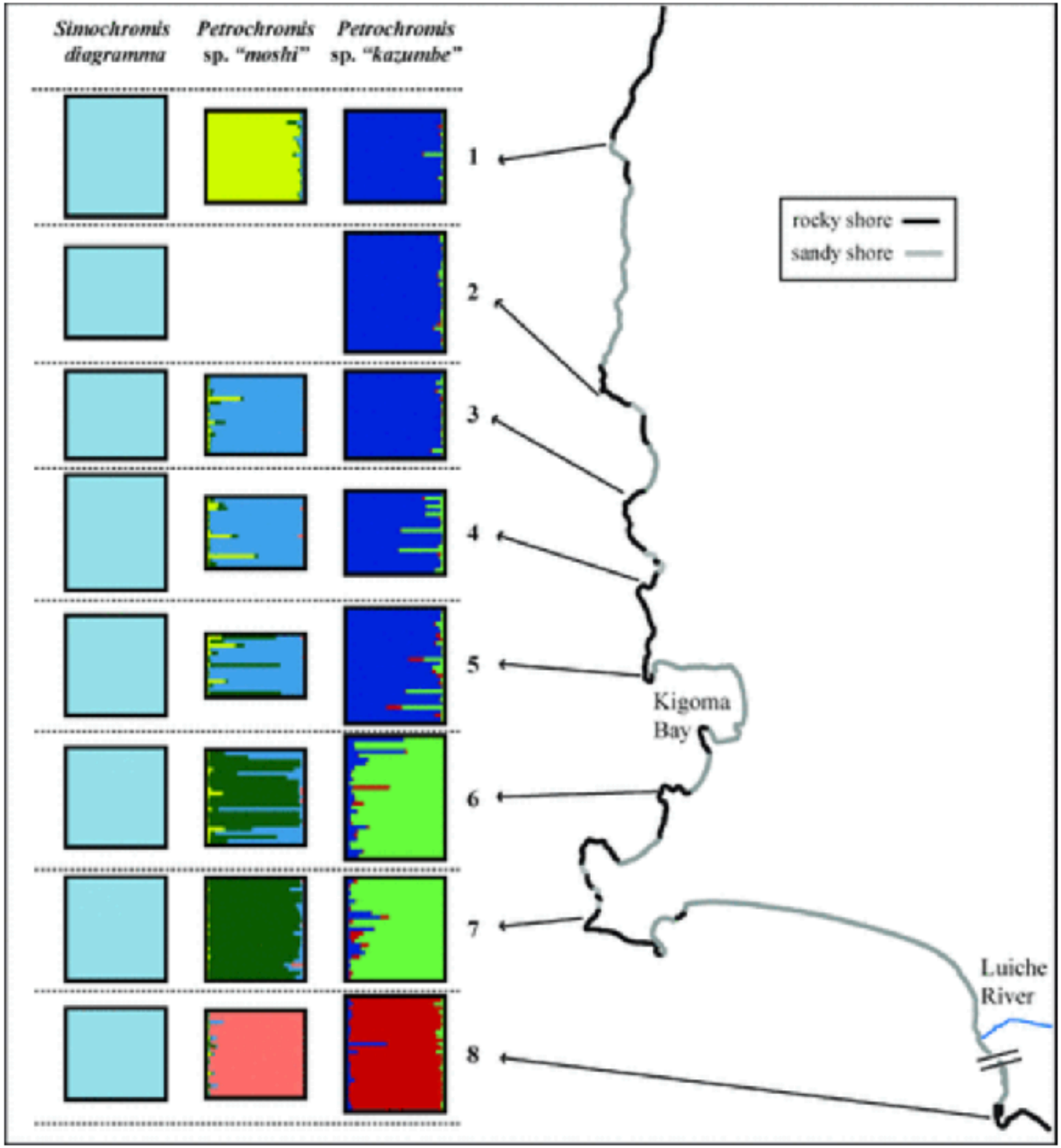
Variations on the STRUCTURE model:

Admixture: Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome research 19(9): 1655–1664

accommodates large genomic datasets

Entropy: Gompert et al 2014, Shastry et al 2021

incorporates genotype likelihoods



Wagner and McCune 2009

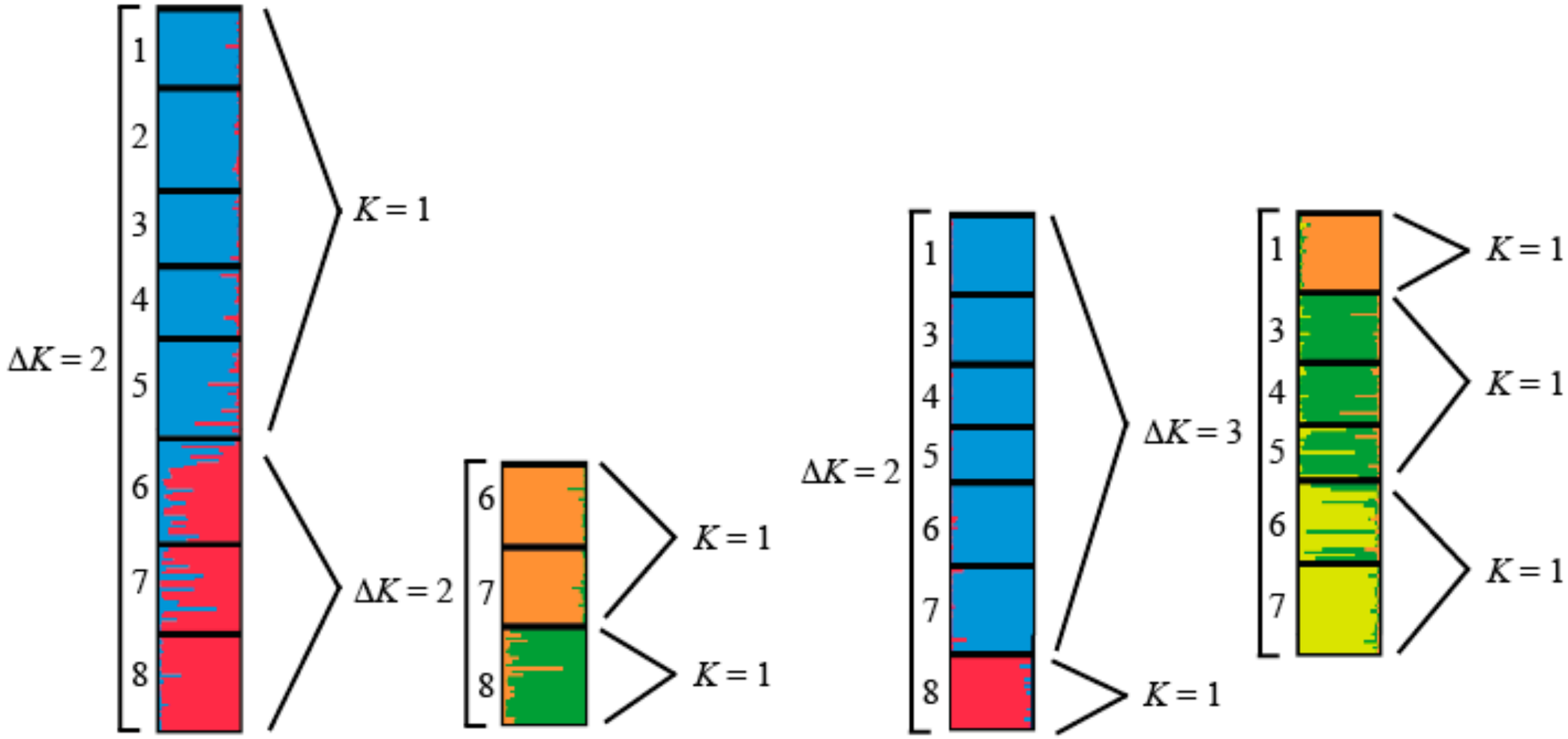


Figure S4. Results of fully hierarchical Δ*K* STRUCTURE analyses for a) *P*. sp. "kazumbe" and b) *P*. sp. "moshi". At each round, datasets were divided into *K* subsets, where *K* is the number of genetic groups supported by Δ*K* analyses (Evanno et al 2005) of the dataset from the previous round. Individuals were assigned into subsets if their assignment probabilities were 0.6 or higher for a given group. Assignment probabilities used for subsetting were the consensus of 10 runs at the *K* indicated by Δ*K* analyses, generated using CLUMPP (Jakobsson and Rosenberg 2007).

In STRUCTURE, LnP(D) is an approximation of the marginal likelihood

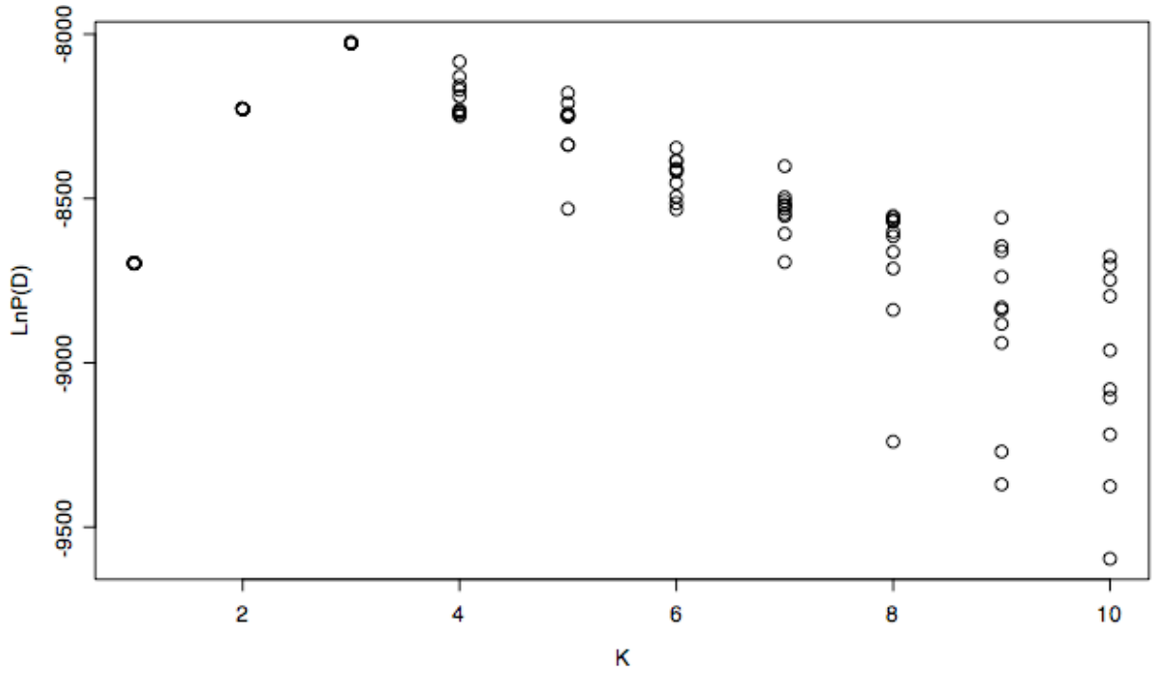


Figure S1. Log probability of the data (LnP(D)) for 10 STRUCTURE runs at eack K value, for K 1 through 10 for *P*. sp. "kazumbe".

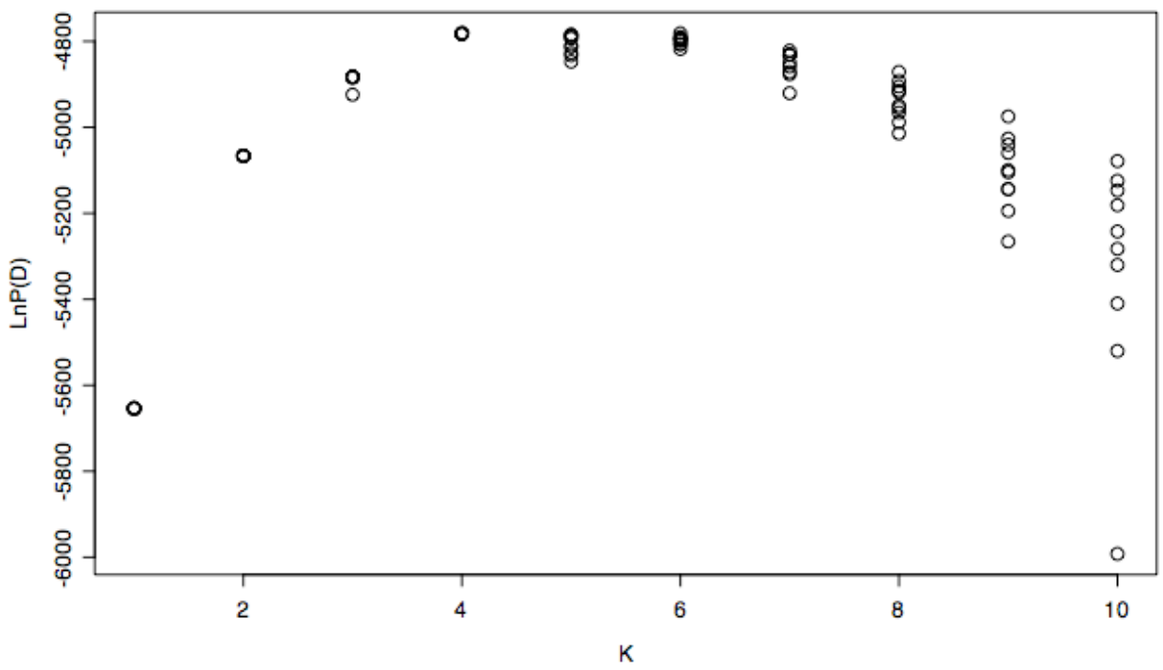


Figure S2. Log probability of the data (LnP(D)) for 10 STRUCTURE runs at eack K value, for K 1 through 10 for *P*. sp. "moshi".

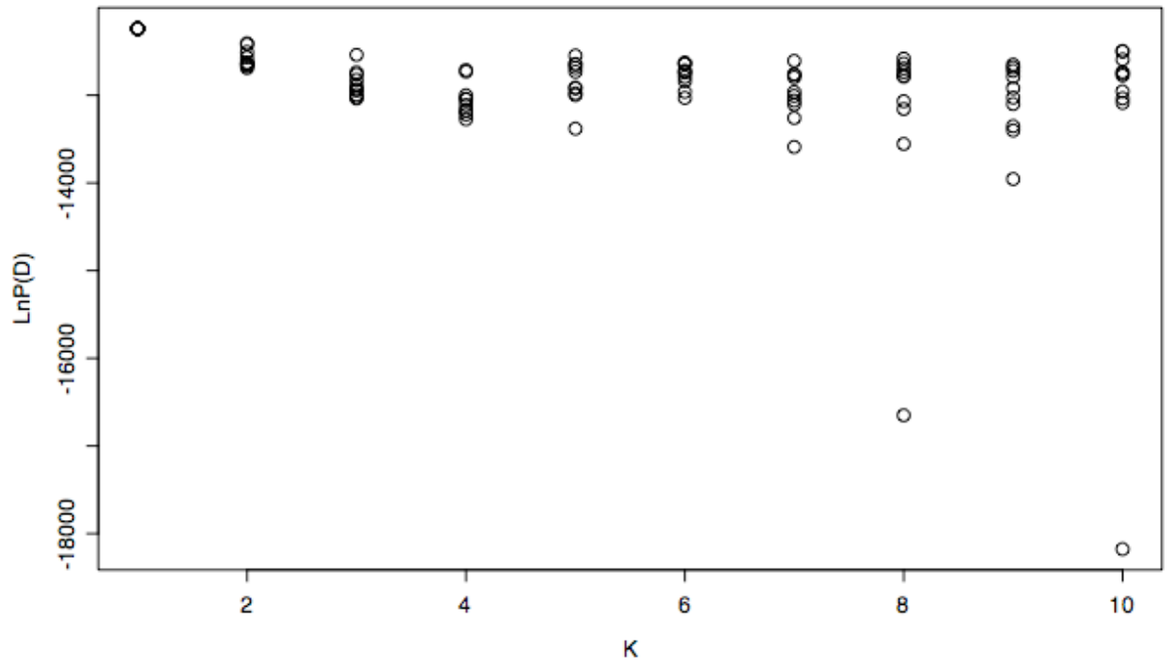
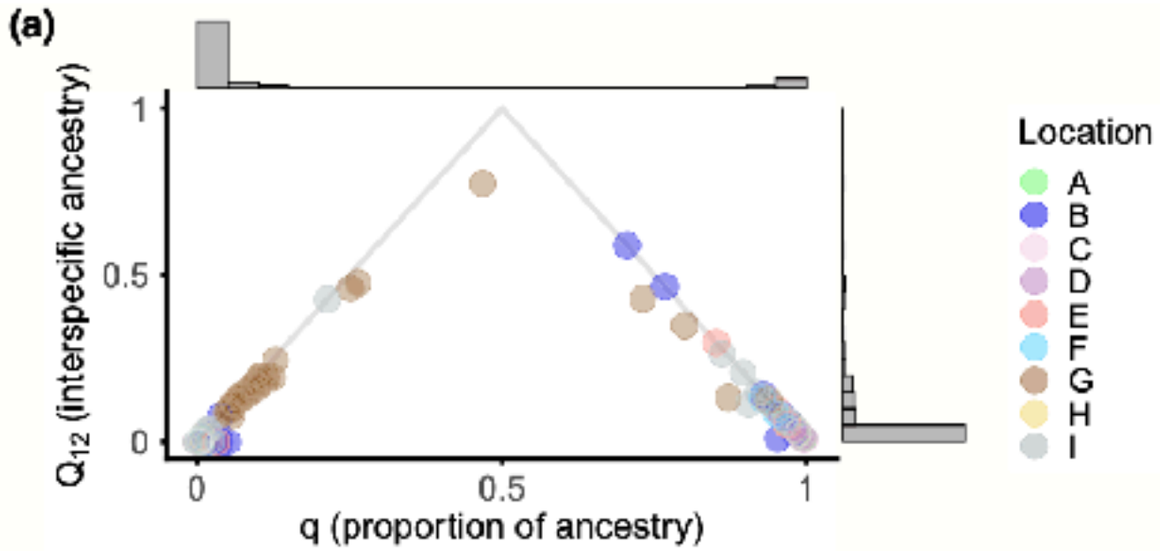
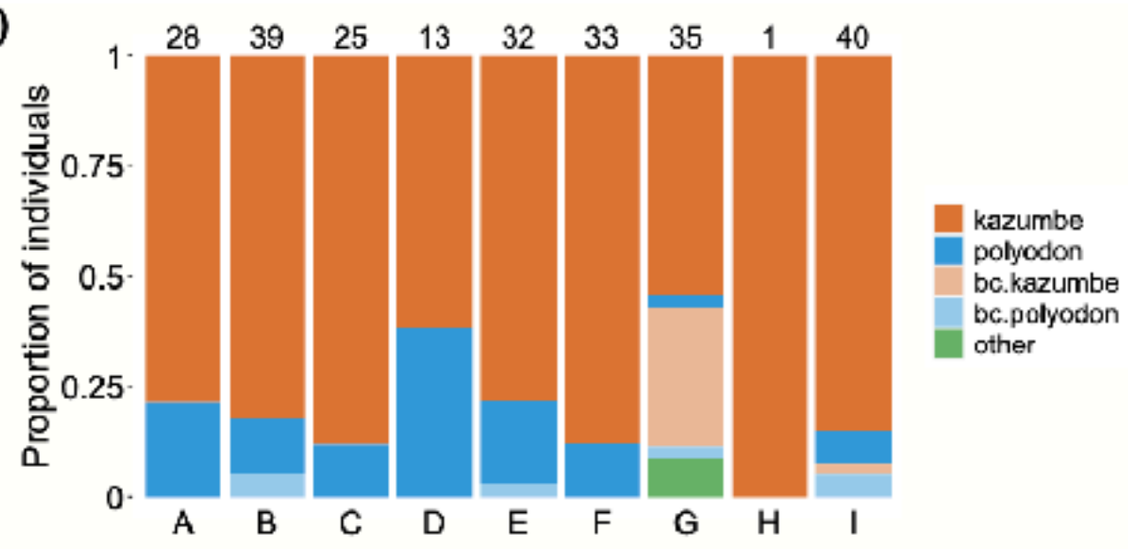


Figure S3. Log probability of the data (LnP(D)) for 10 STRUCTURE runs at eack K value, for K 1 through 10 for *S. diagramma*.

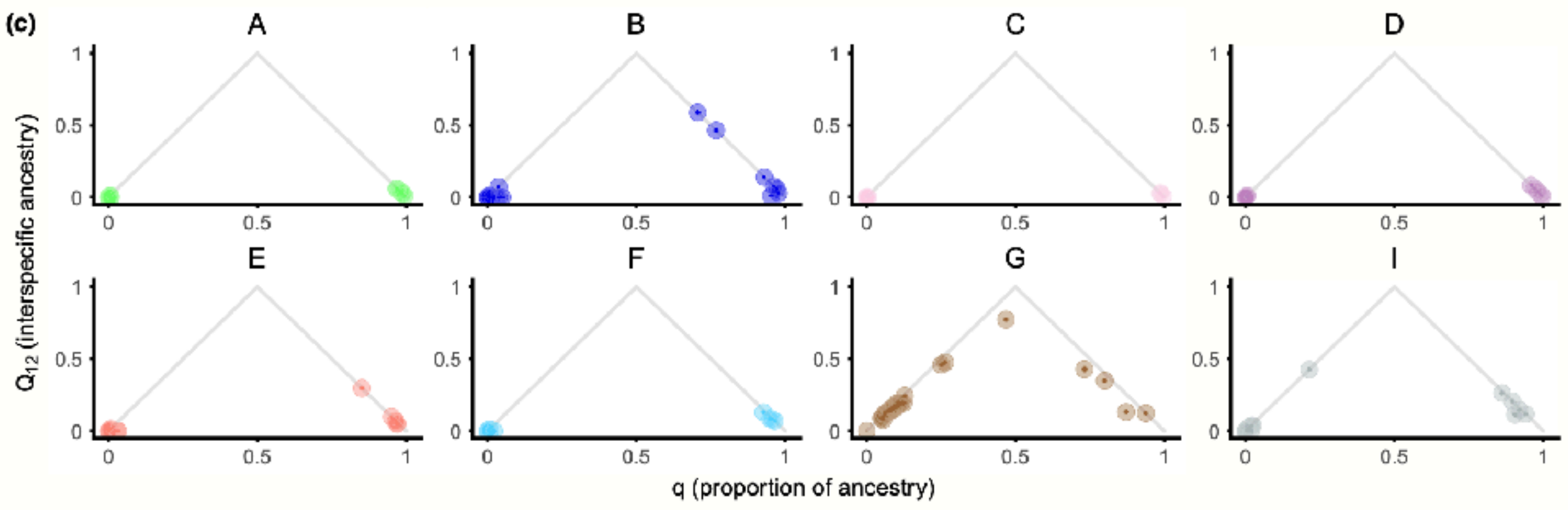
(a)


Q_{12} (interspecific ancestry)
q (proportion of ancestry)
0
0.5
1
Location
A
B
C
D
E
F
G
H
I

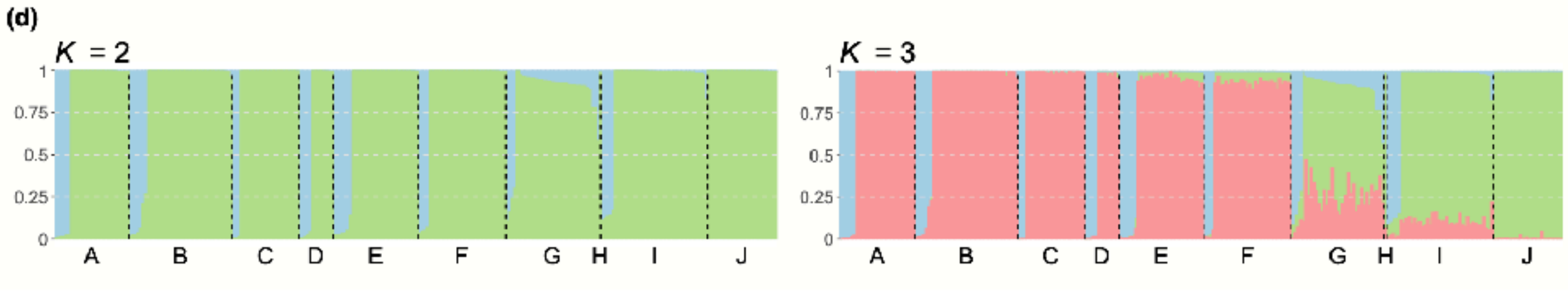
(b)


28
39
25
13
32
33
35
1
40
Proportion of individuals
0
0.25
0.5
0.75
1
A
B
C
D
E
F
G
H
I
kazumbe
polyodon
bc.kazumbe
bc.polyodon
other

(c)

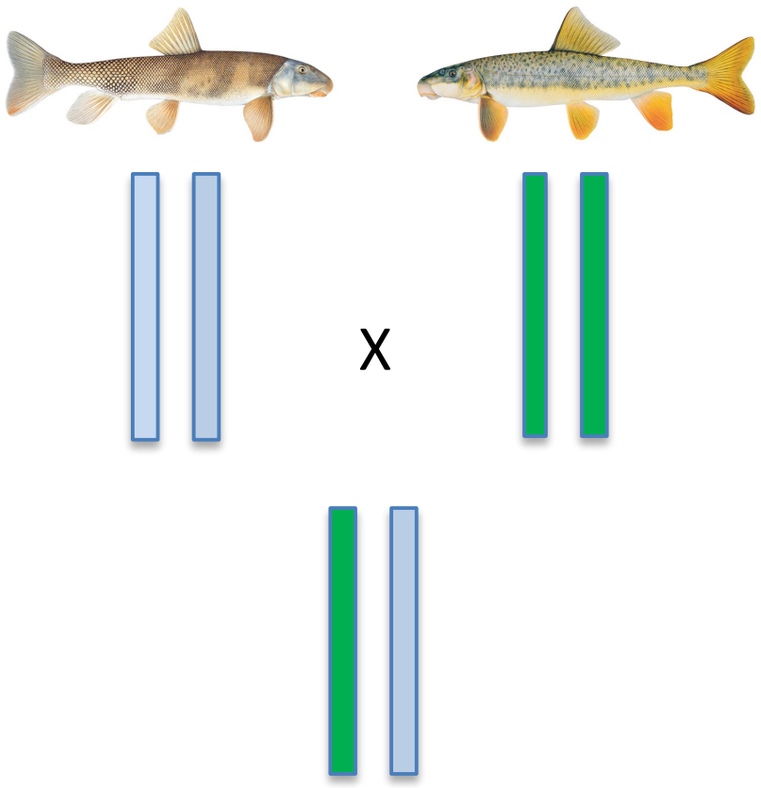

A
B
C
D
E
F
G
I
Q_{12} (interspecific ancestry)
q (proportion of ancestry)
0
0.5
1

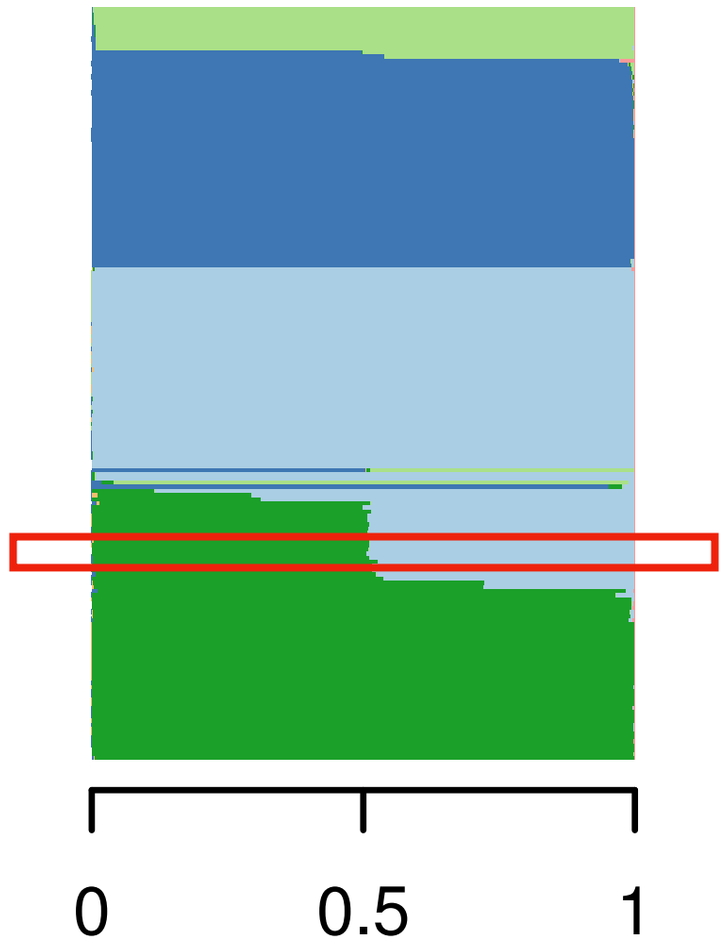
(d)


K = 2
K = 3
0
0.25
0.5
0.75
1
A
B
C
D
E
F
G
H
I
J

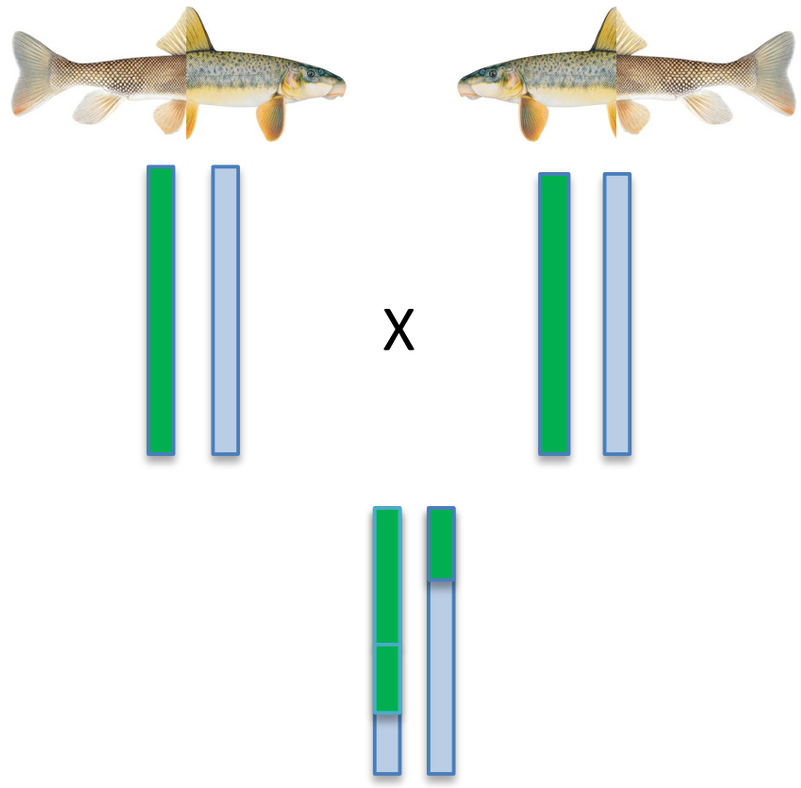
Interspecific ancestry helps differentiate among hybrids

First generation (F_1)



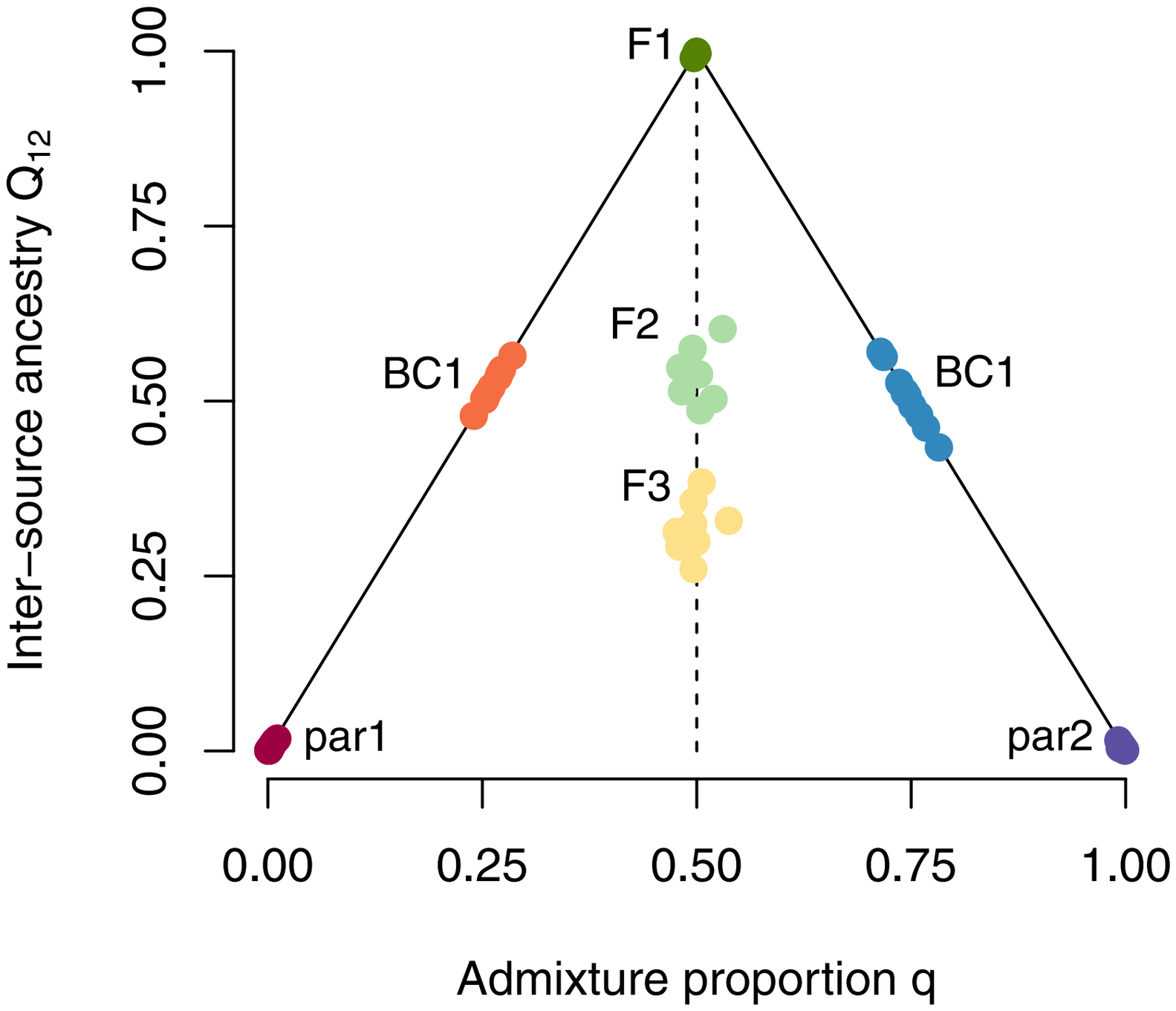


Later-generation ($F_2..F_n$)



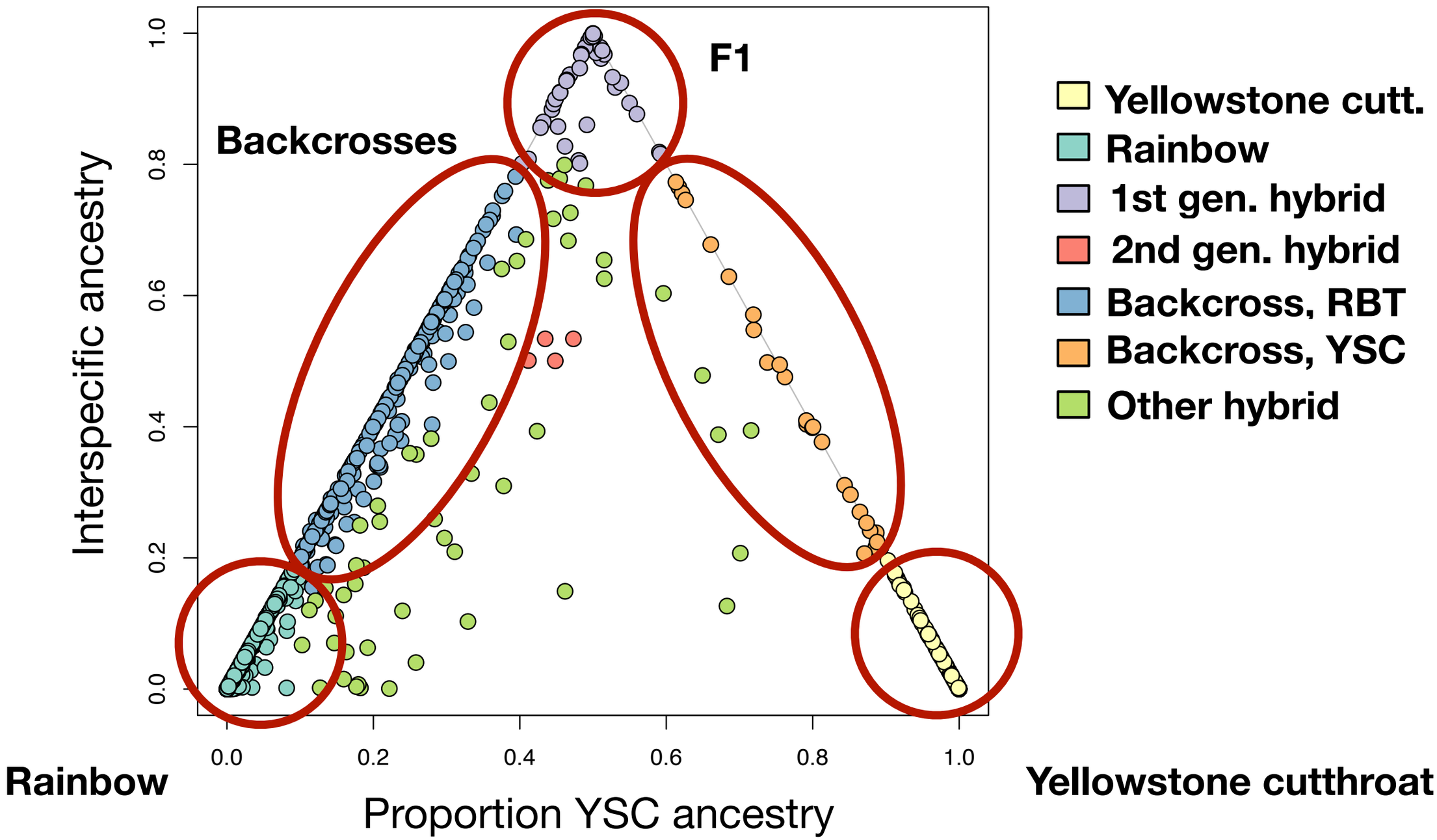
Proportion of ancestry=0.5
Interspecific ancestry=1

Proportion of ancestry=0.5
Interspecific ancestry=0.5



Shastry et al 2021 Figure 2

Interspecific ancestry reveals history of hybridization



Mandeville *et al.* 2019
Molecular Ecology