

# Exercises in Marine Ecological Genetics

## 07. Genetic clustering

- Condense genetic variance into principal components
- Visualize admixture with model-based clustering
- Generate more complex plots in R with ggplot

Martin Helmkampf

<https://github.com/mhelmkampf/meg25>



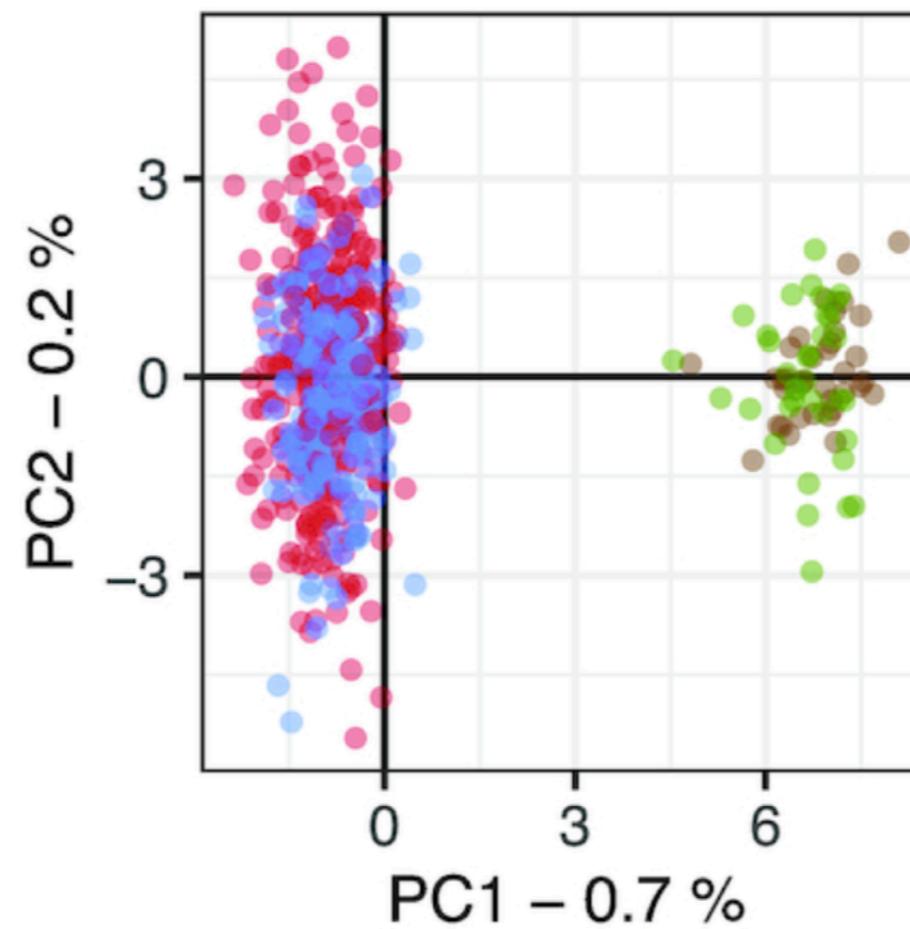
# Principal Component Analysis

- Reduces thousands of SNPs into small number of principal components that capture the most genetic variation among individuals
- Treats each SNP as a variable and encodes genotypes numerically (e.g., 0 = AA, 1 = AB, 2 = BB)
- Calculates covariance matrix and extracts PCs (via eigenvalue decomposition)
- Plots individuals based on PC scores in 2D space (sometimes 3–6D)  
➡ Relationship between individuals (relatedness), population structure, outliers

# Principal Component Analysis

Example: Atlantic mackerel (*Scomber scombrus*) in the Northwest Atlantic, > 10,000 SNPs

- NEA – Bay of Biscay
- NEA – Greenland
- NWA – Canada
- NWA – US



Bourret et al. 2023

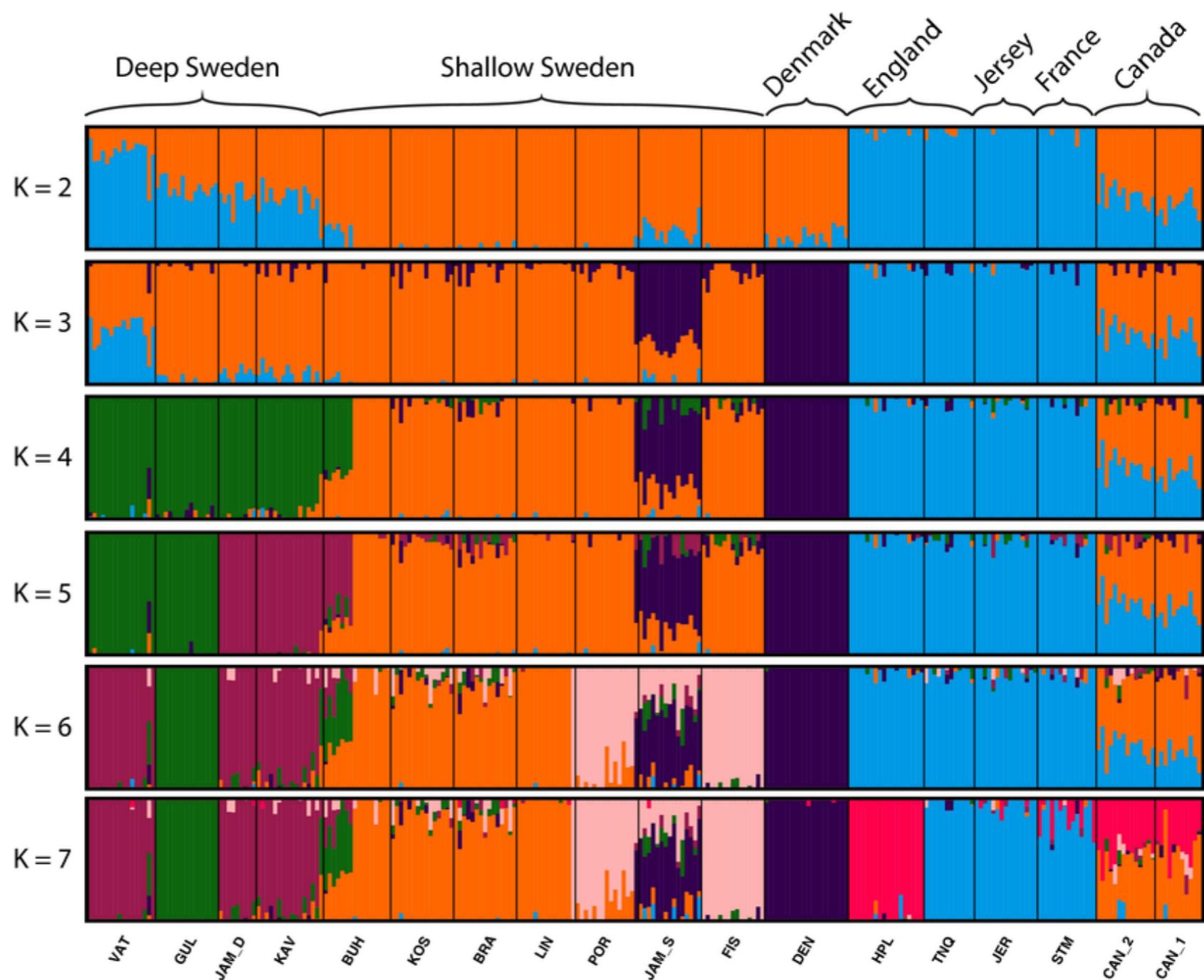
# Admixture analysis

- Estimates ancestry proportions for each individual by assuming  $k$  ancestral populations, based on genome-wide SNPs
  - Uses model-based clustering to estimates which SNPs likely came from which ancestral population
  - Calculates Q-values for each individual  
(e.g., 60% ancestry from population 1, 40% from population 2)
  - Visualizes these ancestry proportions as stacked bar plots, with one bar per individual
- Ancestry and admixture (i.e., mixed ancestry), population structure, assign individuals to populations

# Admixture analysis

Example:

Ancestry proportions of  
*Ciona intestinalis*,  
~ 1600 SNPs



Hudson et al. 2019, *Evolutionary Applications*

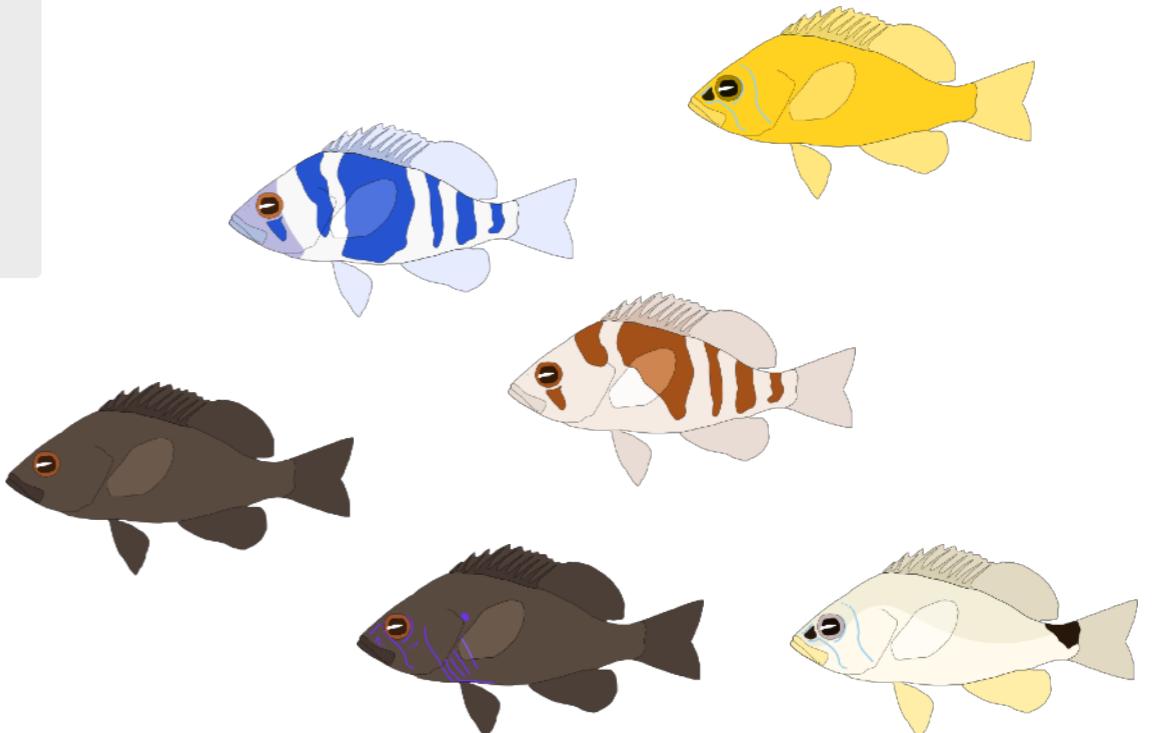
# Example dataset

- 19 species of hamlet (genus *Hypoplectrus*)
- 15 sites in Caribbean and Gulf of Mexico
- 327 hamlet samples total
- Illumina short-read resequencing (mean depth 17×)
- Genotyping with GATK
- High-quality reference genome of *H. puella*



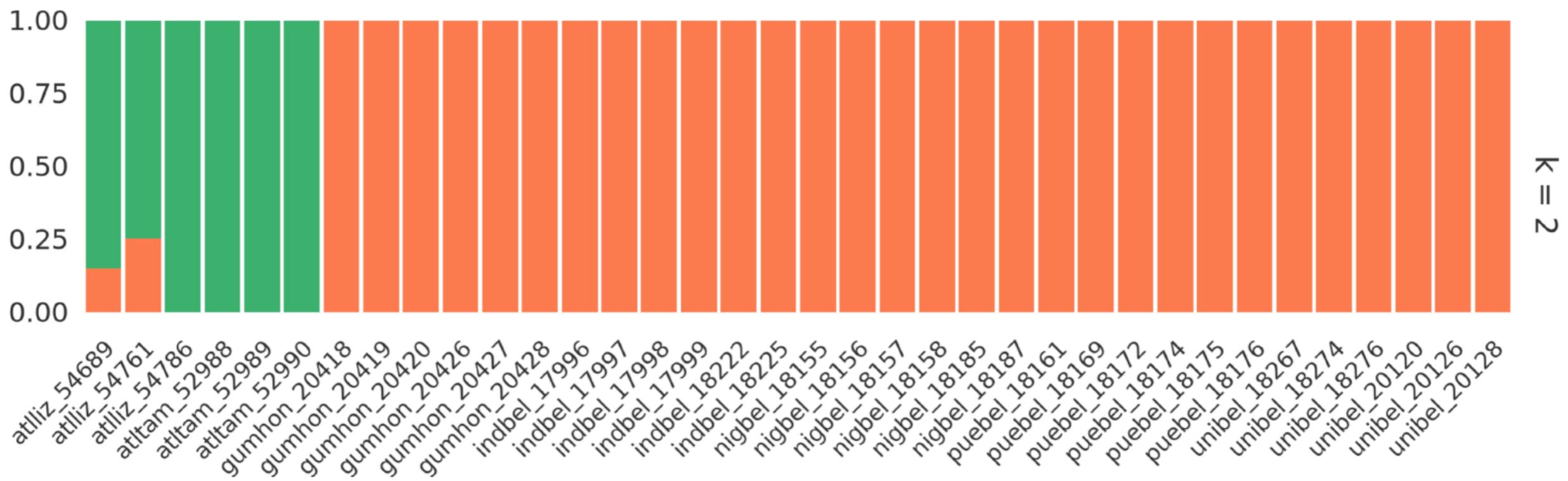
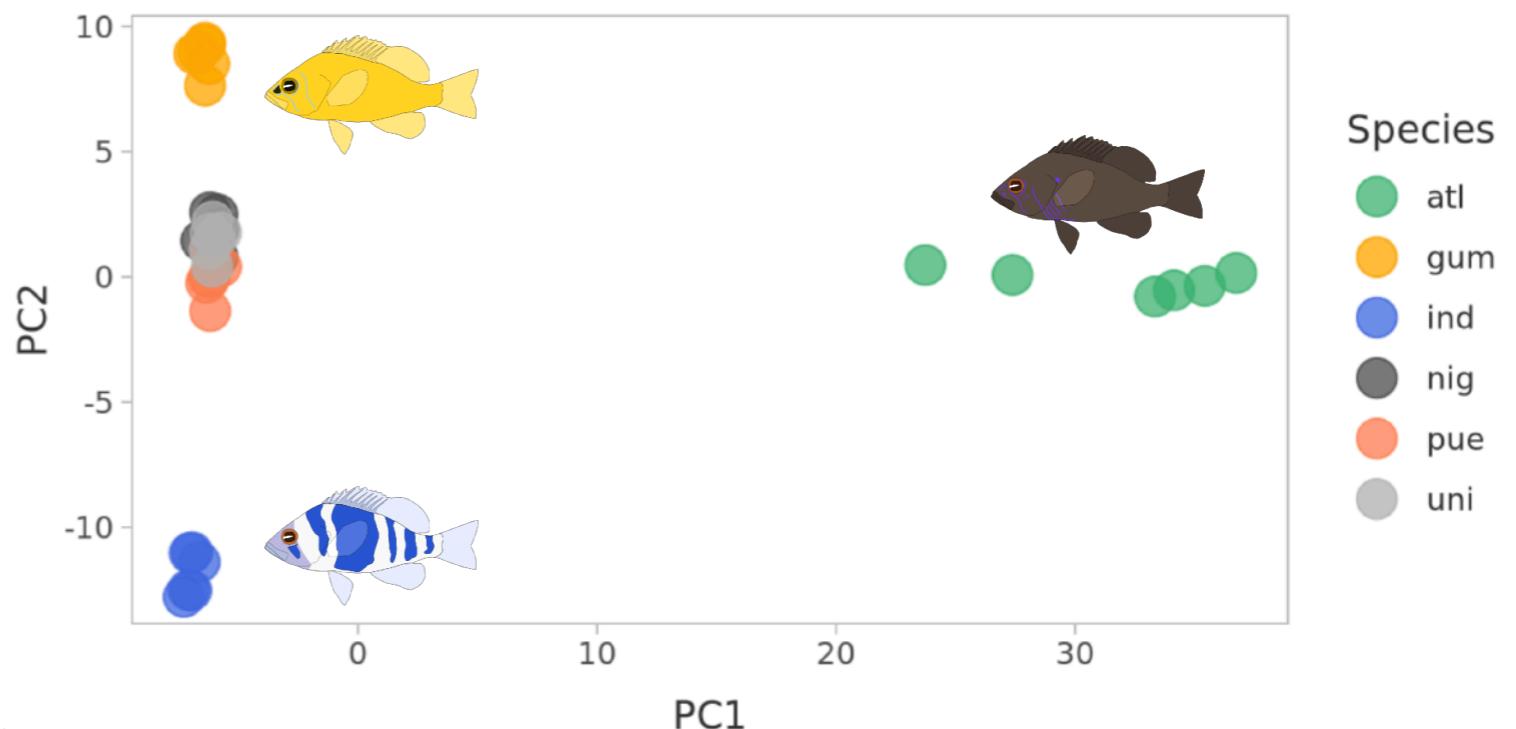
**hamlets\_LG12.snp.vcf.gz**

- Chromosome 12 only
- Subset to 36 samples from 6 populations

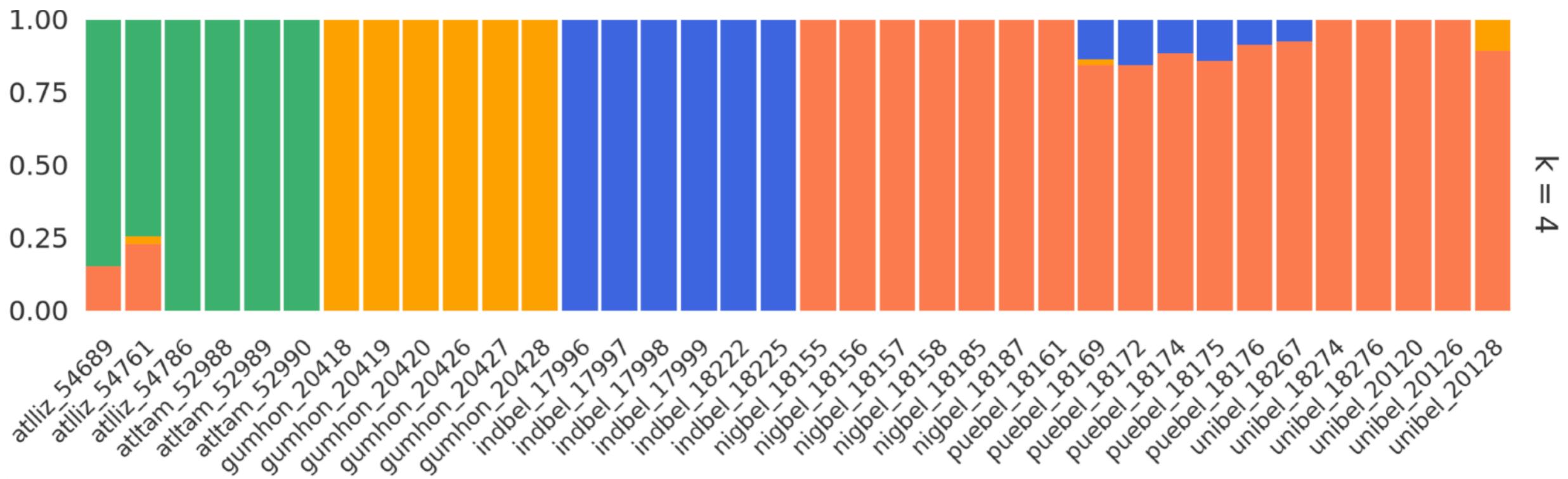
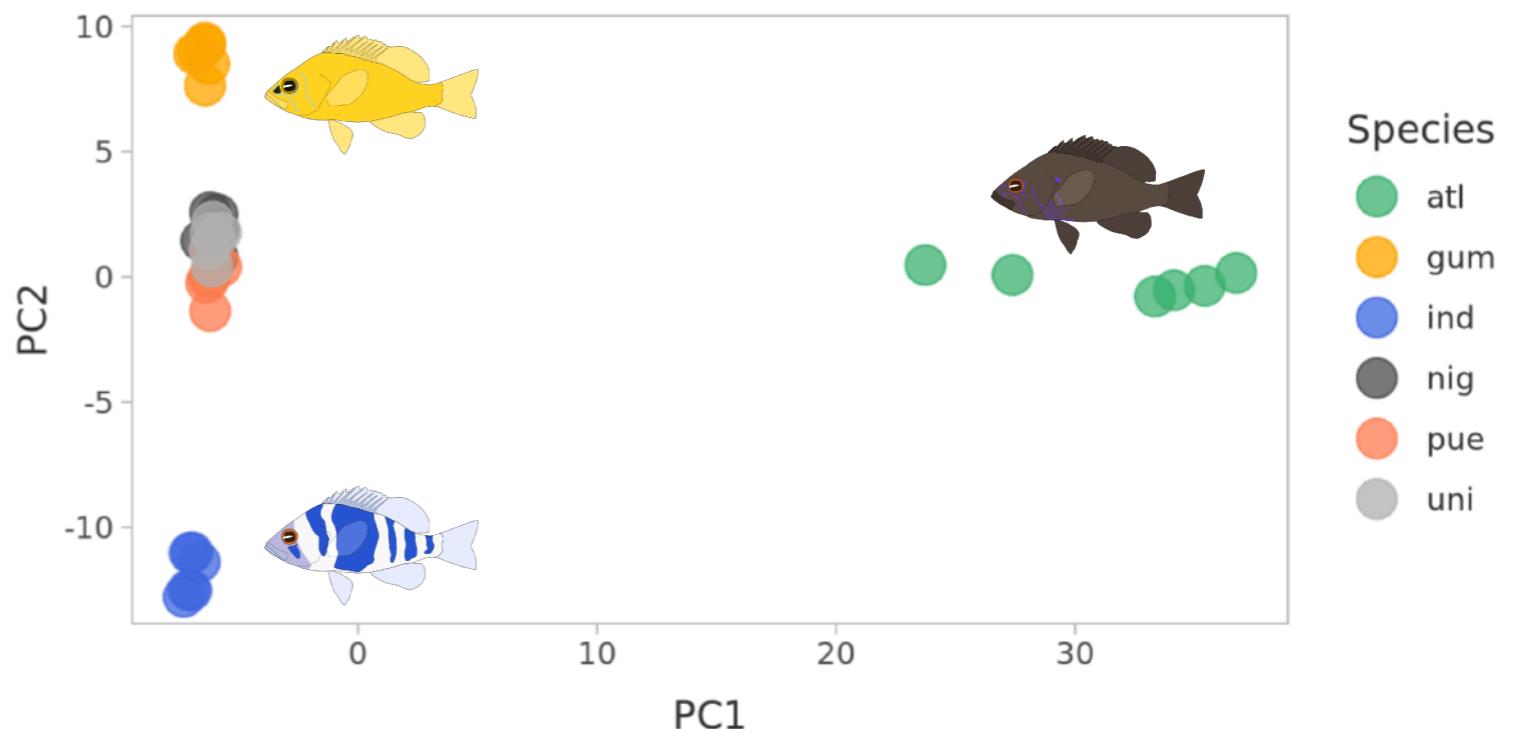


Illustrations by Kosmas Hench

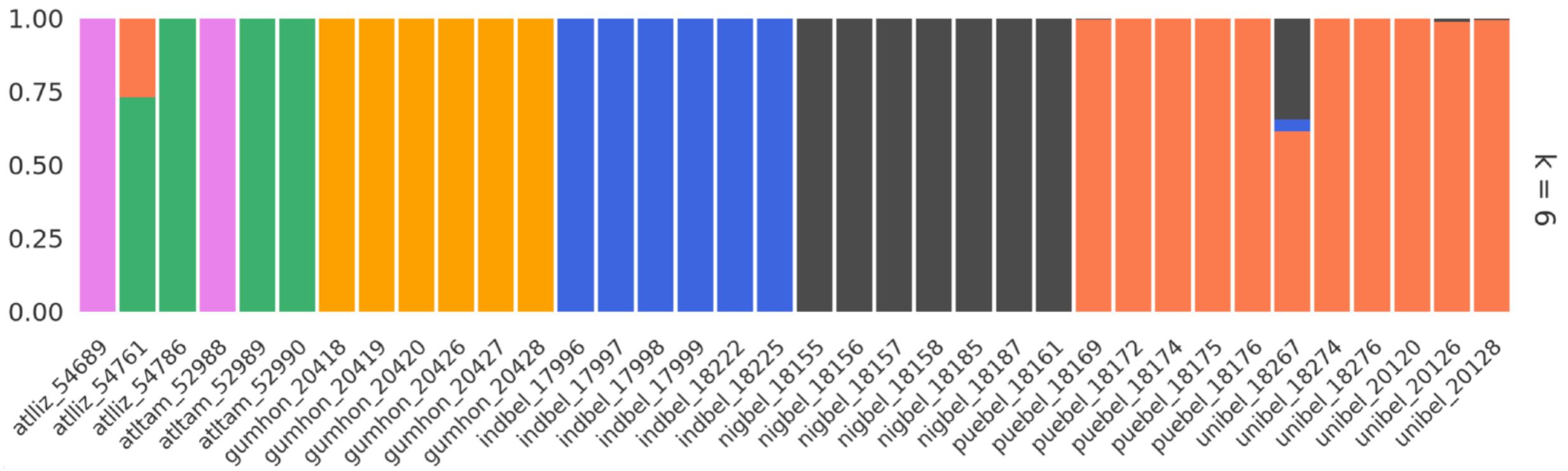
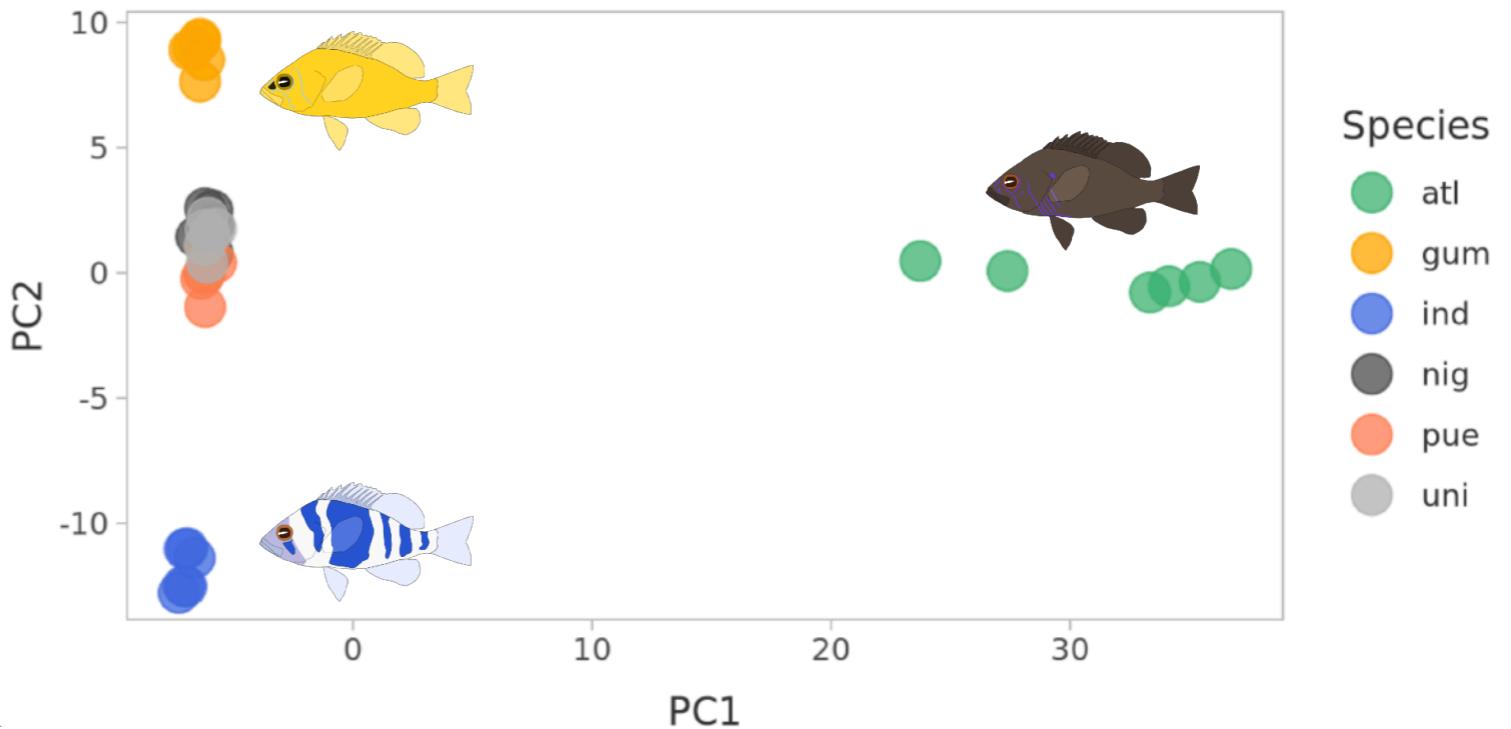
# PCA and admixture plot



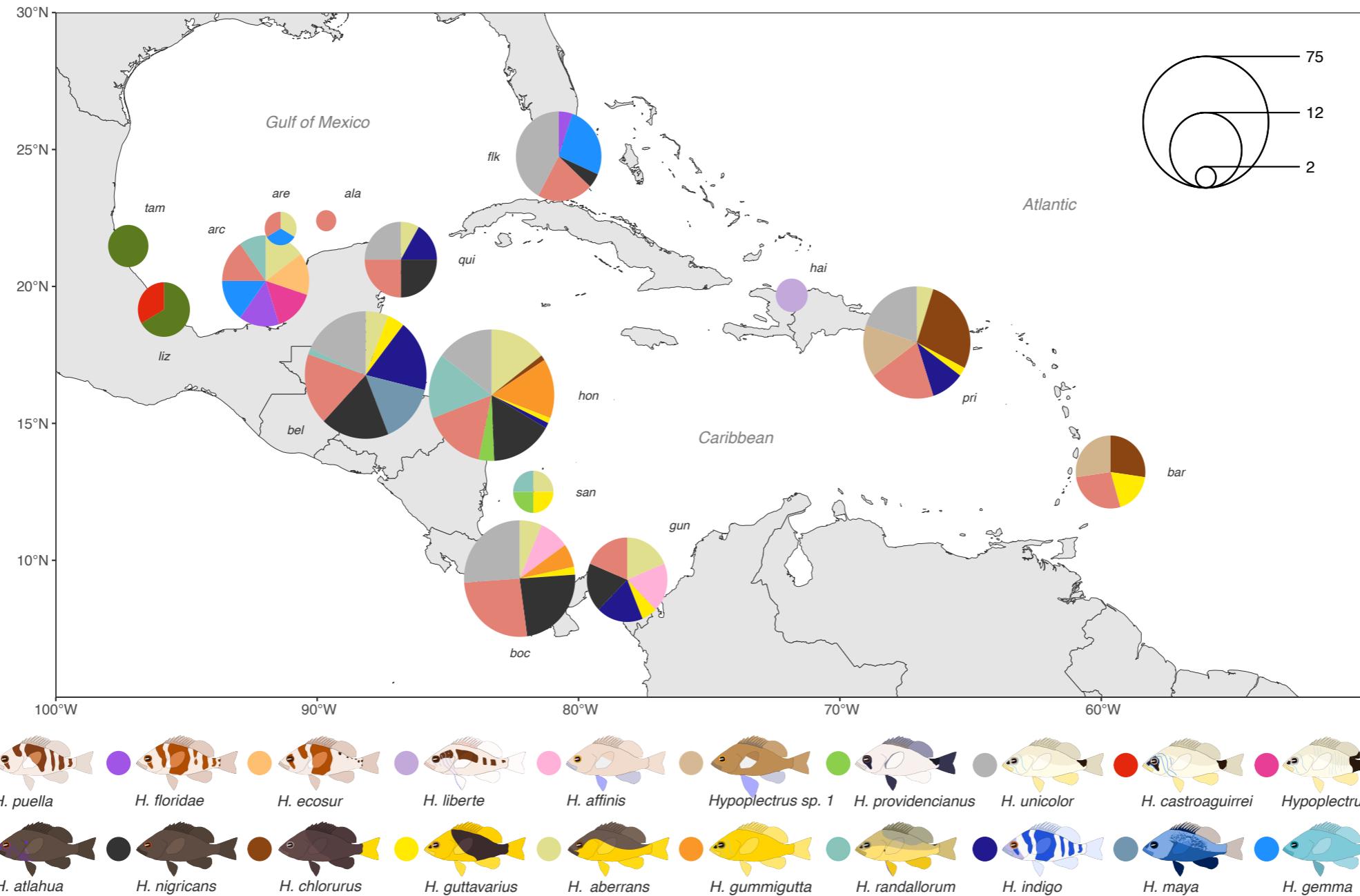
# PCA and admixture plot



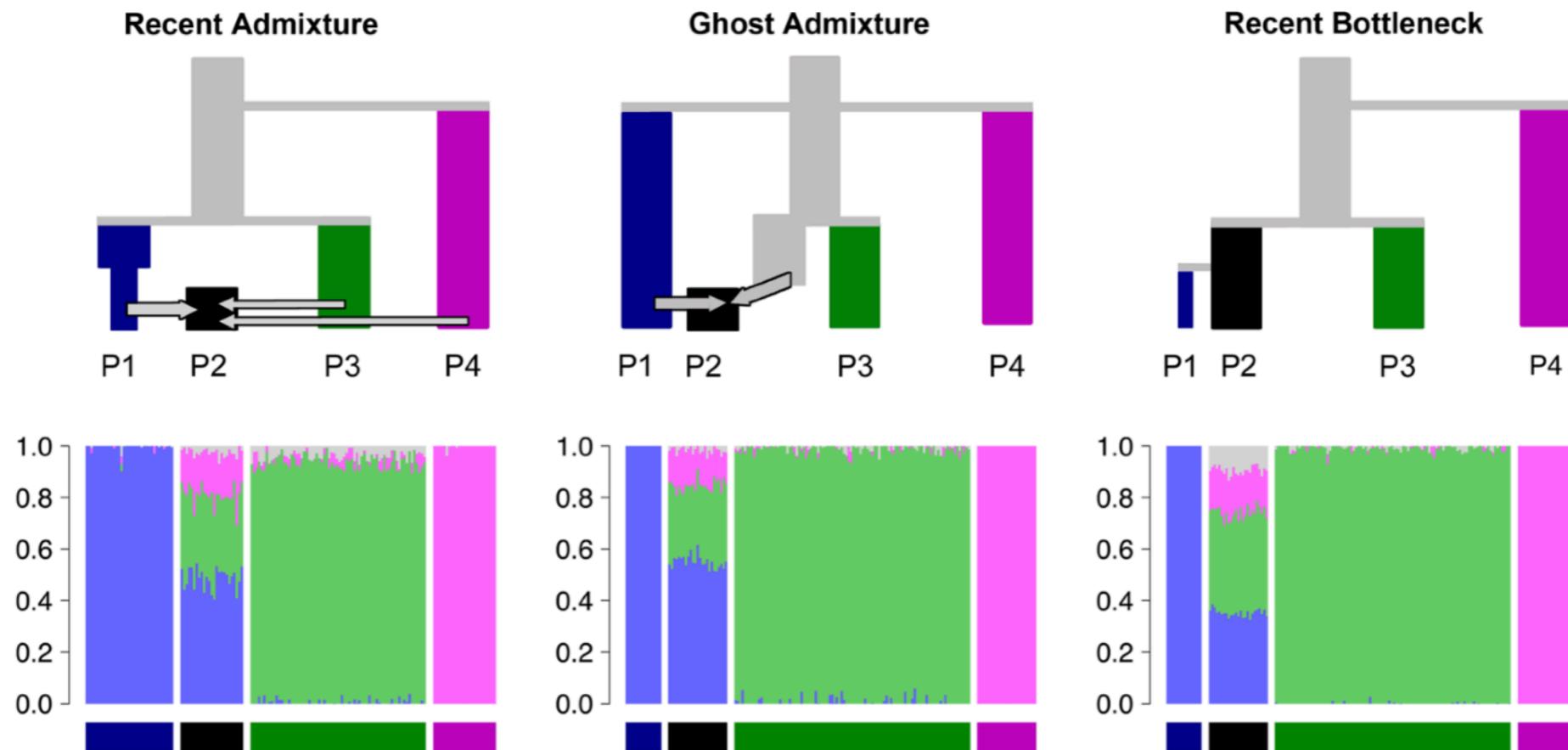
# PCA and admixture plot



# Example dataset



# Admixture caveats

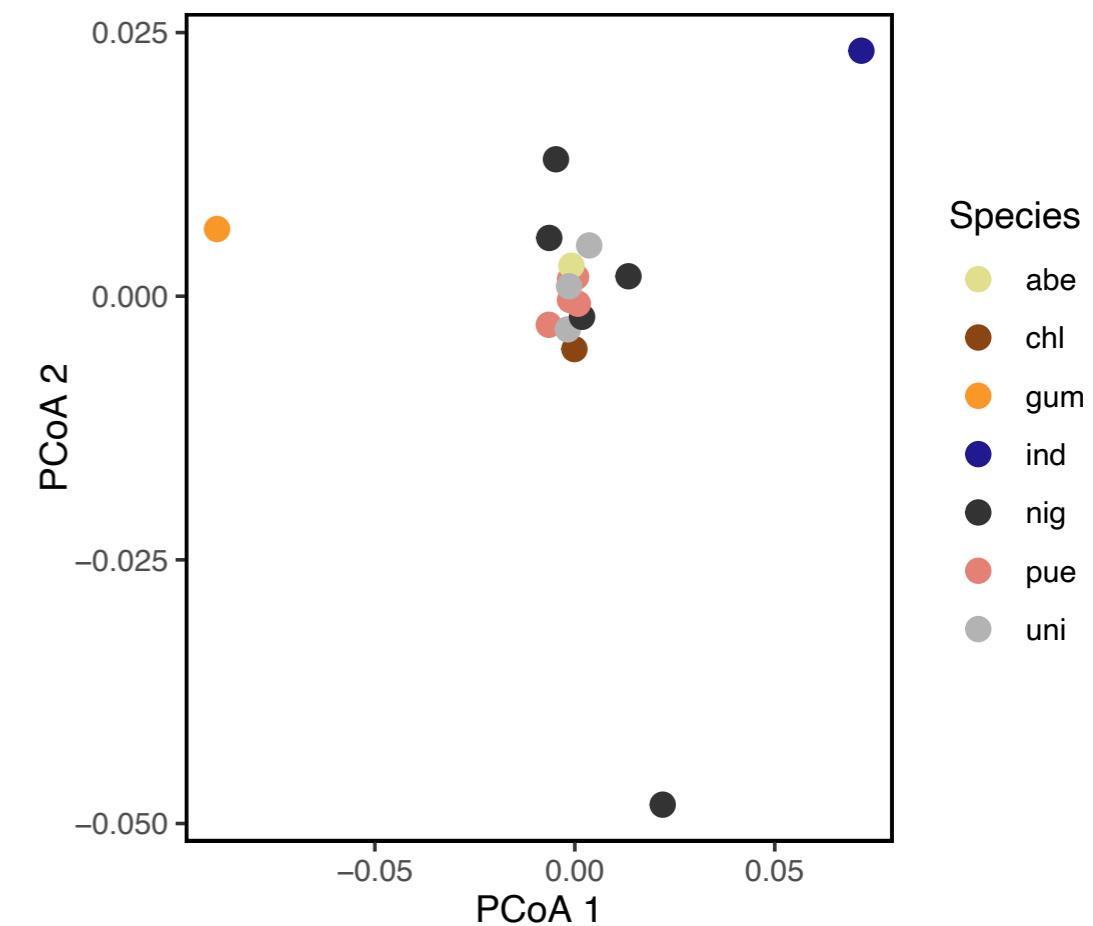
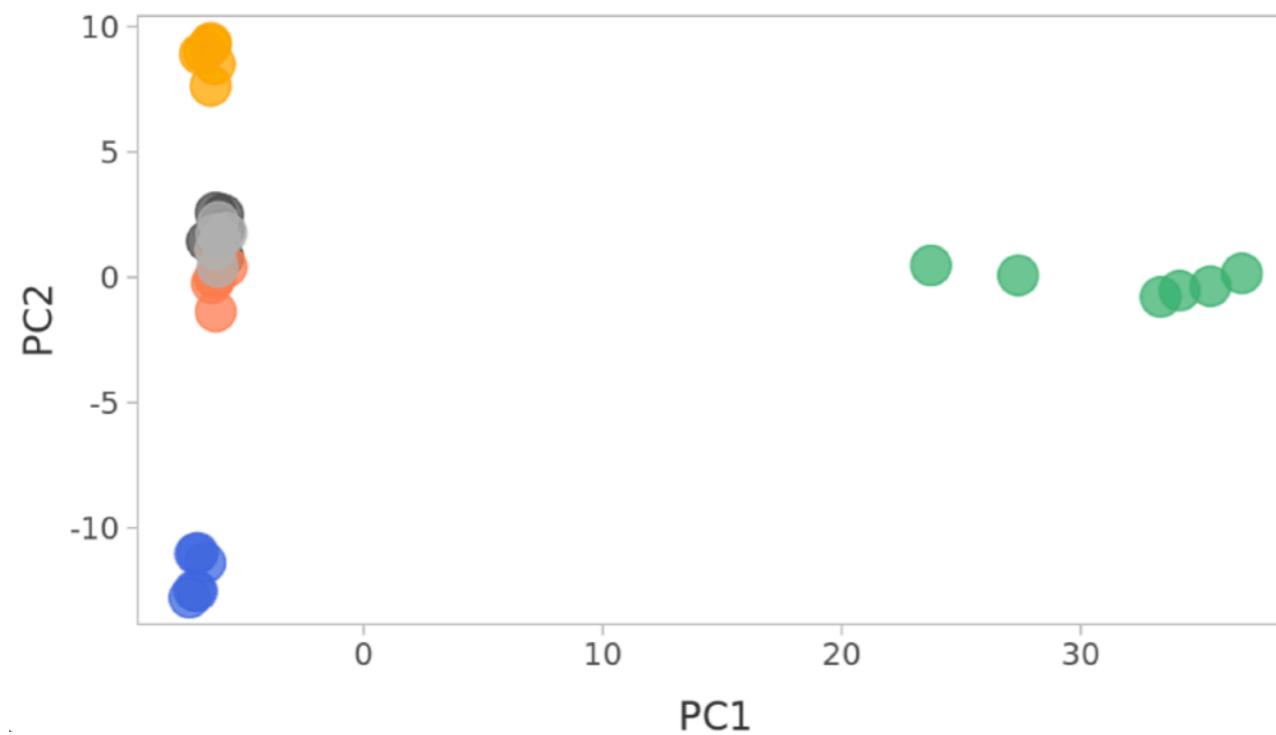


- Demographic histories other than recent admixture can lead to the same results
- Populations are assumed to be in HWE

Lawson et al. 2018, *Nature Communications*

# Pairwise $F_{ST}$ PCoA compared to PCA

- Based on few microsatellite markers, not genome-wide
- Per-population, no per-individual estimates / substructure



```
admixture --cv input.bed k          # Run admixture analysis (Linux command line)

read.vcfR(path_to_vcf_file)        # Read VCF file into R (vcfR package)

glPca(genlight_object)            # Calculate principal components (adegenet package)
```

- Principal Component Analysis of SNP data reveals genetic clusters, population structure and outliers at a glance by reducing complex variation to key axes
- Admixture analysis excels at detecting recent admixture and ancestry proportions, but requires choosing the right  $k$  and careful interpretation
- Both methods are complementary and can be used to assign individuals to populations without prior labels or grouping assumptions (unsupervised clustering)

# Nucleotide diversity $\pi$

Addendum

Average number of nucleotide differences per site between all possible pairs of sequences

$$\hat{\pi} = \frac{n}{n-1} \sum_{ij} x_i x_j \pi_{ij} :$$

- 2 chromosomes per individual
- 15 pairwise comparisons
- $\pi$  per site: no. diff / comparisons

Average  $\pi = 0.29$  ←

Site:	1	2	3	4	5
Sample A:	A	G	C	T	T
	A	G	C	T	T
Sample B:	A	G	T	T	T
	A	G	T	T	T
Sample C:	G	G	T	C	T
	G	G	C	T	T
No. diff:	8	0	9	5	0
$\pi$ per site:	0.53	0	0.60	0.33	0

# Genetic diversity

Addendum

