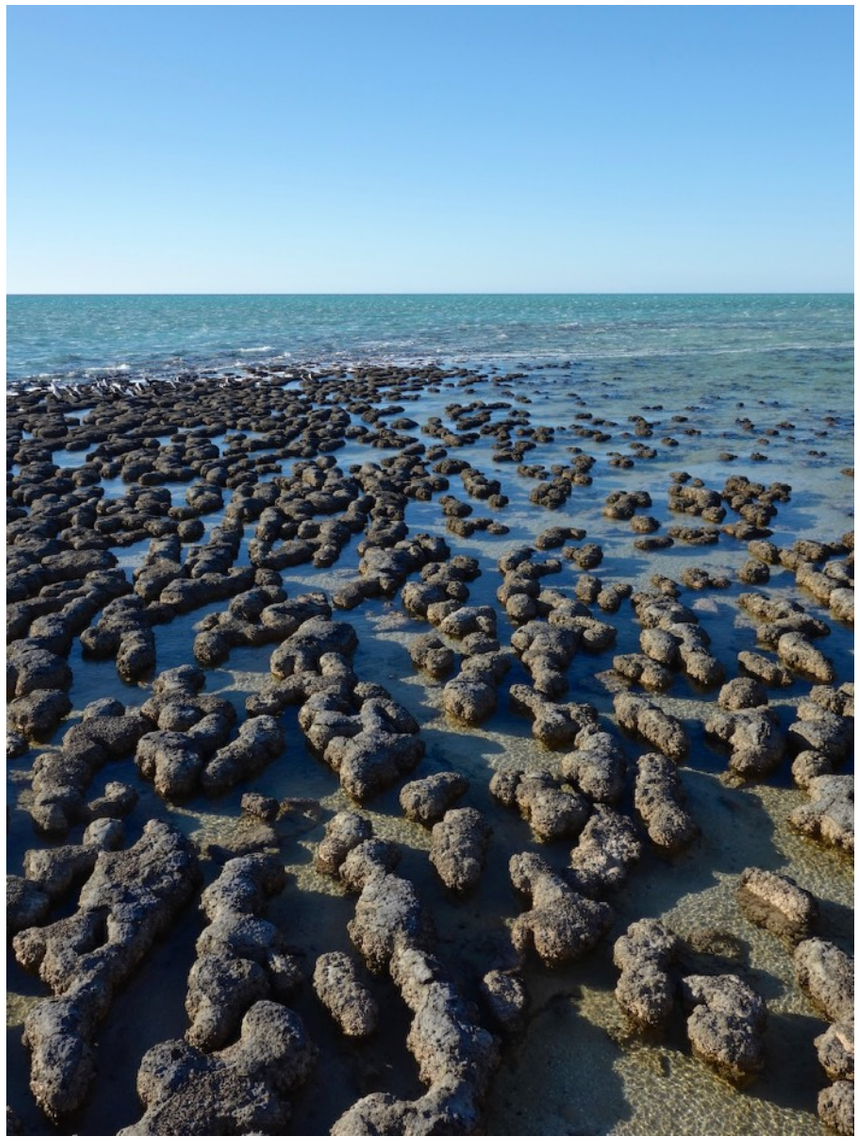
Exercises in Marine Ecological Genetics

03. Population structure

- Calculate *F*-statistics in R
- Test for genetic differentiation
- Use PCoA to visualize population structure

Martin Helmkampf

https://github.com/mhelmkampf/meg25



Useful R functions for population genetics

Import Genepop file as genind object

```
x <- read.genepop(path_to_file, ncode = n)                    # adegenet package
```

Object name

Name of Genepop file (including path)

No. of characters encoding each allele

Locus statistics (e.g. no. of alleles, Simpson’s index, heterozygosity, evenness)

```
locus_table(x)                                                # poppr package
```

Useful R functions for population genetics

Recap

Test for departure from Hardy-Weinberg equilibrium

```
hw.test(x)                                          # quick summary (pegas package)

test_HW(path_to_file, outputFile = new_file)        # detailed (genepop package)
```

Conclusions

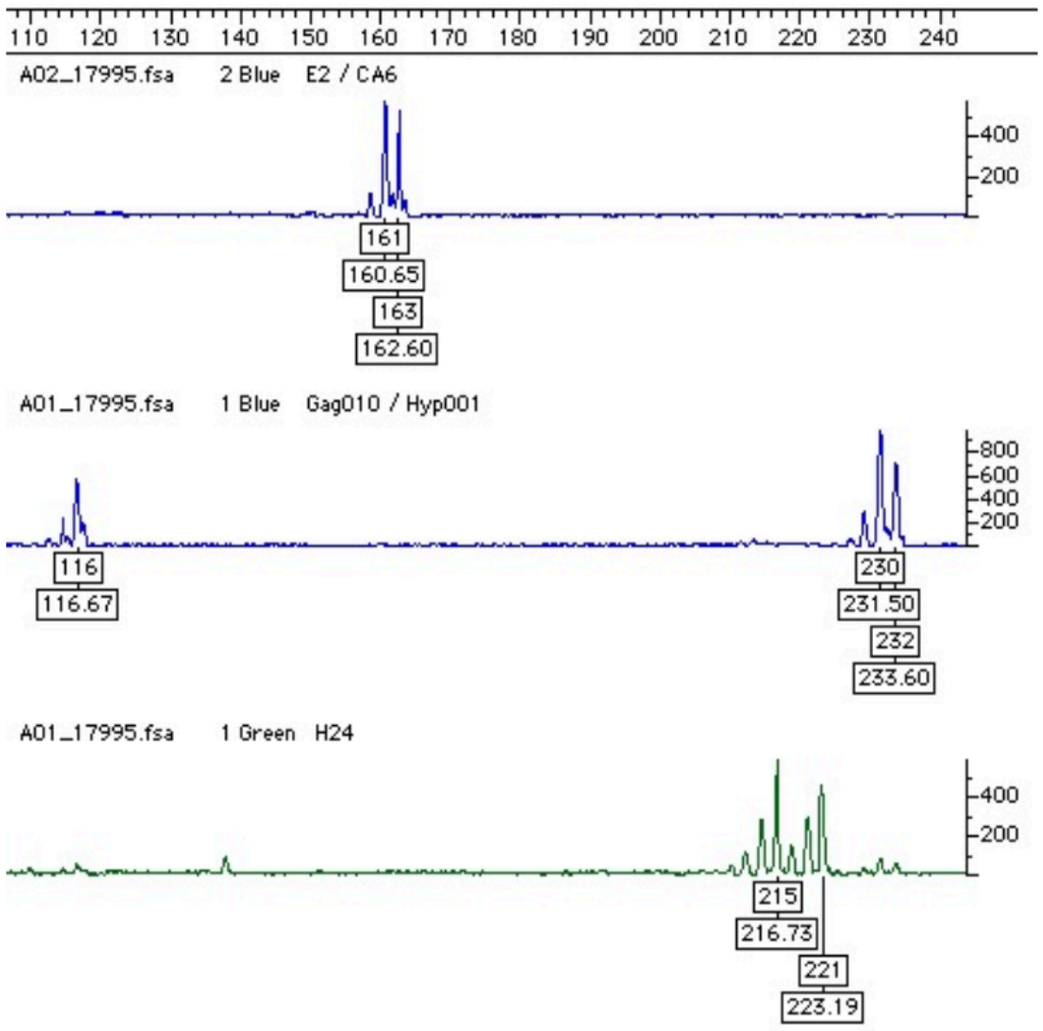
- Deviations from HW expectations may vary by locus
- Genome- or population-wide patterns can be assessed by analyzing multiple loci
- Correcting for multiple hypothesis testing may be necessary

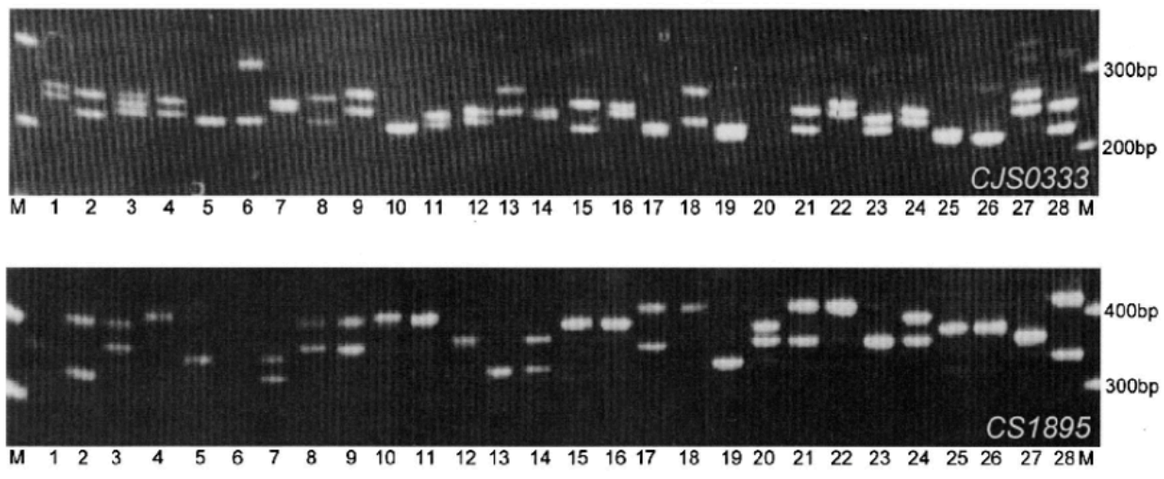
Microsatellites

ACAGTTACAGCAGGAAGTTC […] TATATATATATATATATATATA […] CGCTAATG ACAGCACGCTAAC

- Co-dominant
- Many alleles
- High mutation rates
- High heterozygosities

$(TA)_n$





Genepop format

puella_barbados.txt

```
Microsatellite genotypes of Hypoplectrus puella from Barbados
g2
gag010
h24
hyp001
hyp015
hyp018
e2
hyp008
hyp016
pam013
POP
barbados ind#735 , 203235 121127 210248 225231 126132 190192 157157 236244 194220 114128
barbados ind#736 , 205217 119119 204216 225229 126126 190192 157165 236236 216222 106144
barbados ind#737 , 203203 135147 224226 233235 126136 190198 157163 238250 202226 126160
barbados ind#738 , 211217 119121 228230 223231 126130 190190 157157 240244 220226 138160
barbados ind#739 , 205225 121125 208216 227231 132132 192192 158159 244256 192196 146146
barbados ind#740 , 217233 119121 216228 227233 126126 192198 157157 242256 194216 000000
barbados ind#741 , 203209 000000 216222 223229 126130 192198 157157 234236 188226 118130
barbados ind#742 , 203215 119119 222234 233233 130130 190190 157157 240246 182188 108126
barbados ind#743 , 203243 111121 216224 231231 126126 190192 157159 234249 188198 118126
barbados ind#744 , 203225 121135 206232 231231 126130 192198 157159 240249 182222 122156
barbados ind#745 , 203211 121123 224236 227231 126132 192192 157163 238238 182204 138148
barbados ind#746 , 227233 119123 216234 227233 126130 190190 157159 236245 186188 124142
barbados ind#747 , 203223 133141 236240 231231 126126 192198 157157 234236 194204 118126
barbados ind#748 , 203217 141145 222226 227235 126126 190192 145157 238240 182196 108108
barbados ind#749 , 205221 125147 210238 223231 126132 192198 157157 234234 188200 106148
barbados ind#750 , 203235 123123 230234 223231 126132 192198 157157 236256 188214 122136
barbados ind#751 , 203211 121143 212248 223229 126132 190196 157159 234242 214216 118128
barbados ind#752 , 213217 123169 204234 223231 126134 190198 157163 236248 193224 142148
```

Free text

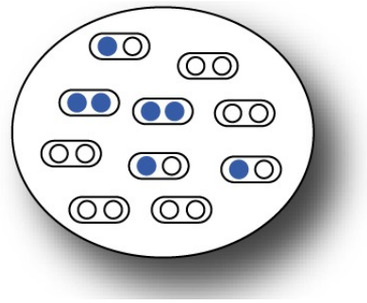
Locus names

Sample / population

Alleles (000 = missing data)

Hardy-Weinberg equilibrium

A *single generation of reproduction will result in a population that meets the expected Hardy-Weinberg frequencies*, i.e. is at *Hardy-Weinberg* (HW) *equilibrium*



Assuming an “ideal” population, i.e. :

- Diploid organisms
- Sexual reproduction (as opposed to clonal)
- Random mating (as opposed to e.g. assortative) with respect to genotype
- Random union of gametes
- Discrete, non-overlapping generations
- Very large (infinite) population
- No migration
- No population structure
- No natural selection
- Two alleles
- Identical allele frequencies in both sexes

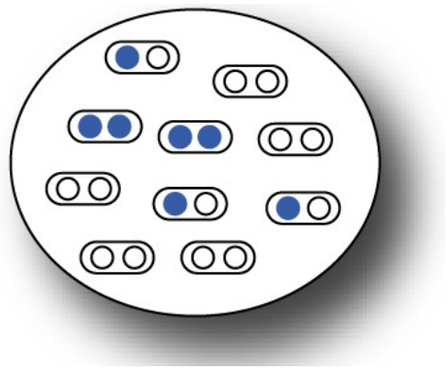
-> Departures from HW equilibrium may indicate:

- Inbreeding
- Assortative mating
- Self-fertilization
- Natural selection
- Population structure
- ...

Fixation index

HETEROZYGOSITY

In one population



H_o = proportion of heterozygote individuals, observed heterozygosity

$H_e = 2pq = 1 - p^2 - q^2$, expected heterozygosity (assuming HW equilibrium)

$$F = \frac{H_e - H_o}{H_e}$$

Fixation index: *proportion by which heterozygosity is reduced or increased relative to the heterozygosity of a population at HW equilibrium with the same allele frequencies*.

Divided by H_e −> *proportion* (of expected heterozygosity)

Varies between -1 and 1

F < 0: heterozygote excess

F > 0 heterozygote deficit (homozygote excess)

May be averaged over several loci −> reduces bias

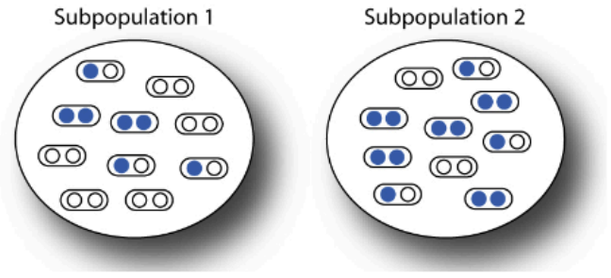
May be extended to *k* alleles

$$\frac{0.445 - 0.1}{0.445} = 0.78$$

Extension to multiple subpopulations

In *n* subpopulations

$$H_I = \frac{1}{n}\sum_{i=1}^{n}\hat{H}_i$$ Mean observed heterozygosity within subpopulations



$$H_S = \frac{1}{n}\sum_{i=1}^{n}2p_iq_i$$ Mean expected heterozygosity within subpopulations (assuming HW within subpops)

$$H_T = 2\overline{p}\overline{q}$$ Expected heterozygosity of the total population (assuming HW within the total pop)

$\hat{H}_i$ observed heterozygosity in population i $p_i(q_i)$ frequency of allele A(a) in population i

$\overline{p}(\overline{q})$ mean frequency of allele A(a) n: number of subpopulations

$$F_{IS} = \frac{H_S - H_I}{H_S} \qquad F_{ST} = \frac{H_T - H_S}{H_T} \qquad F_{IT} = \frac{H_T - H_I}{H_T}$$

F_{IS} Average difference between observed and Hardy–Weinberg expected heterozygosity within each subpopulation (due to non-random mating)

F_{ST} Reduction in heterozygosity due to subpopulation divergence of allele frequencies

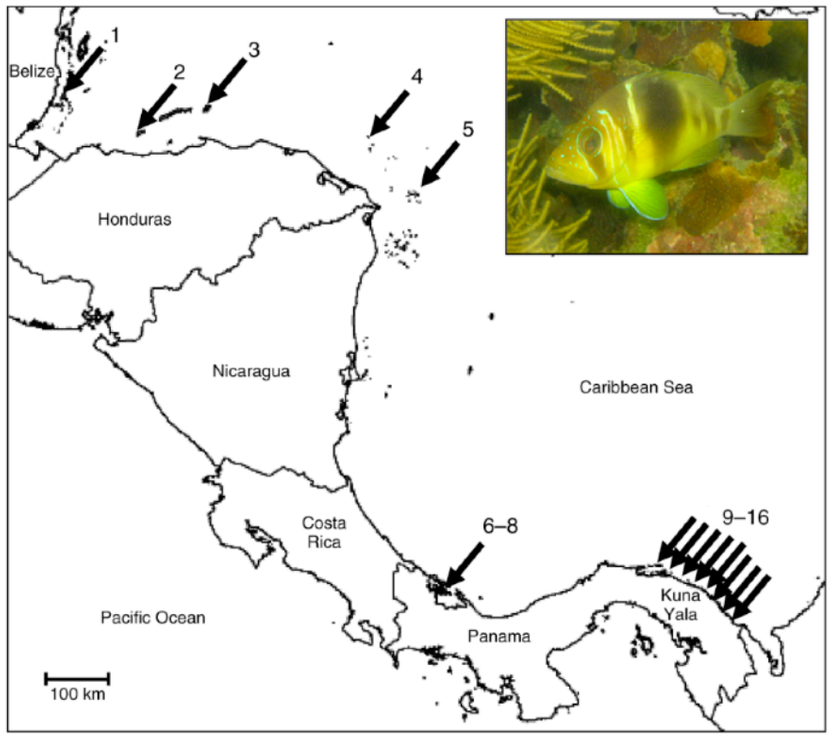
F_{IT} *Combined departure from HW expected genotype frequencies due to non-random mating within subpopulations and divergence of allele frequencies among subpopulations*

Calculate F_{ST} and test for genetic differentiation

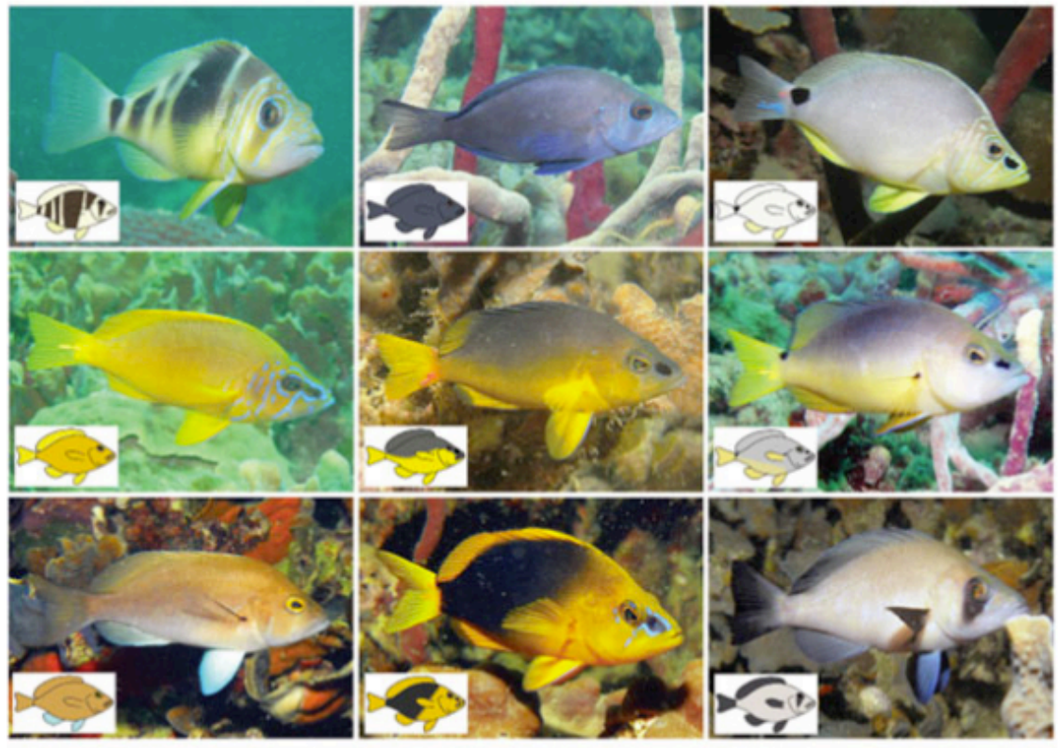
Exercises 1–3

Compare the amount of genetic structure in

- Caribbean populations of *H. puella* (puella_caribbean.gen)
- A time series of *H. puella* samples from Panama (puella_timeseries.gen)
- Several species of hamlets from various Caribbean locations (hamlets_caribbean.gen)



— Puebla et al. 2009, *Ecology*



— Puebla et al. 2012, *Proc R Soc B*

Calculate global F_{ST} and test for differentiation

Exercises 1–3

Species	$\hat{F}_{ST}$	$\widehat{N_e m}$	Reference
Amphibians			
Alytes muletansis (Mallorcan midwife toad)	0.12–0.53	1.8–0.2	Kraaijeveld-Smit et al. 2005
Birds			
Gallus gallus (broiler chicken breed)	0.19	1.0	Emara et al. 2002
Mammals			
Capreolus capreolus (roe deer)	0.097–0.146	2.2–1.4	
Homo sapiens (human)	0.03–0.05	7.8–4.6	
Microtus arvalis (common vole)	0.17	1.2	
Plants			
Arabidopsis thaliana (mouse-ear cress)	0.643	0.1	
Oryza officinalis (wild rice)	0.44	0.3	
Phlox drummondii (annual phlox)	0.17	1.2	
Prunus armeniaca (apricot)	0.32	0.5	
Fish			
Morone saxatilis (striped bass)	0.002	11.8	Brown et al. 2005
Sparisoma viride (stoplight parrotfish)	0.019	12.4	Geertjes et al. 2004
Insects			
Drosophila melanogaster (fruit fly)	0.112	2.0	Singh and Rhomberg 1987
Glossina pallidipes (tsetse fly)	0.18	1.1	Ouma et al. 2005
Heliconius charithonia (butterfly)	0.003	79.8	Kronforst and Flemming 2001
Corals			
Seriatopora hystrix	0.089–0.136	2.6–1.6	Maier et al. 200

F_{ST}	Subpopulations are …
< 0.15	very similar
0.15 – 0.25	similar
> 0.25	distinct

Visualize population structure

- Reduce high-dimensional genetic data to 2–6 axes for visualization
  - Calculate pairwise distances and project samples or subpopulations into a coordinate system that preserves these distances (**PCoA**)
  - Find axes (principal components) that capture the maximum variance in genotype data (**PCA**)
- Individuals or subpopulations cluster by genetic similarity, revealing structure across populations or species
- PCoA is suited for subpopulations and microsatellite data, while PCA can be used on the level of individuals and SNP genotypes

New R concepts: tidyverse and ggplot

Tidyverse pipes

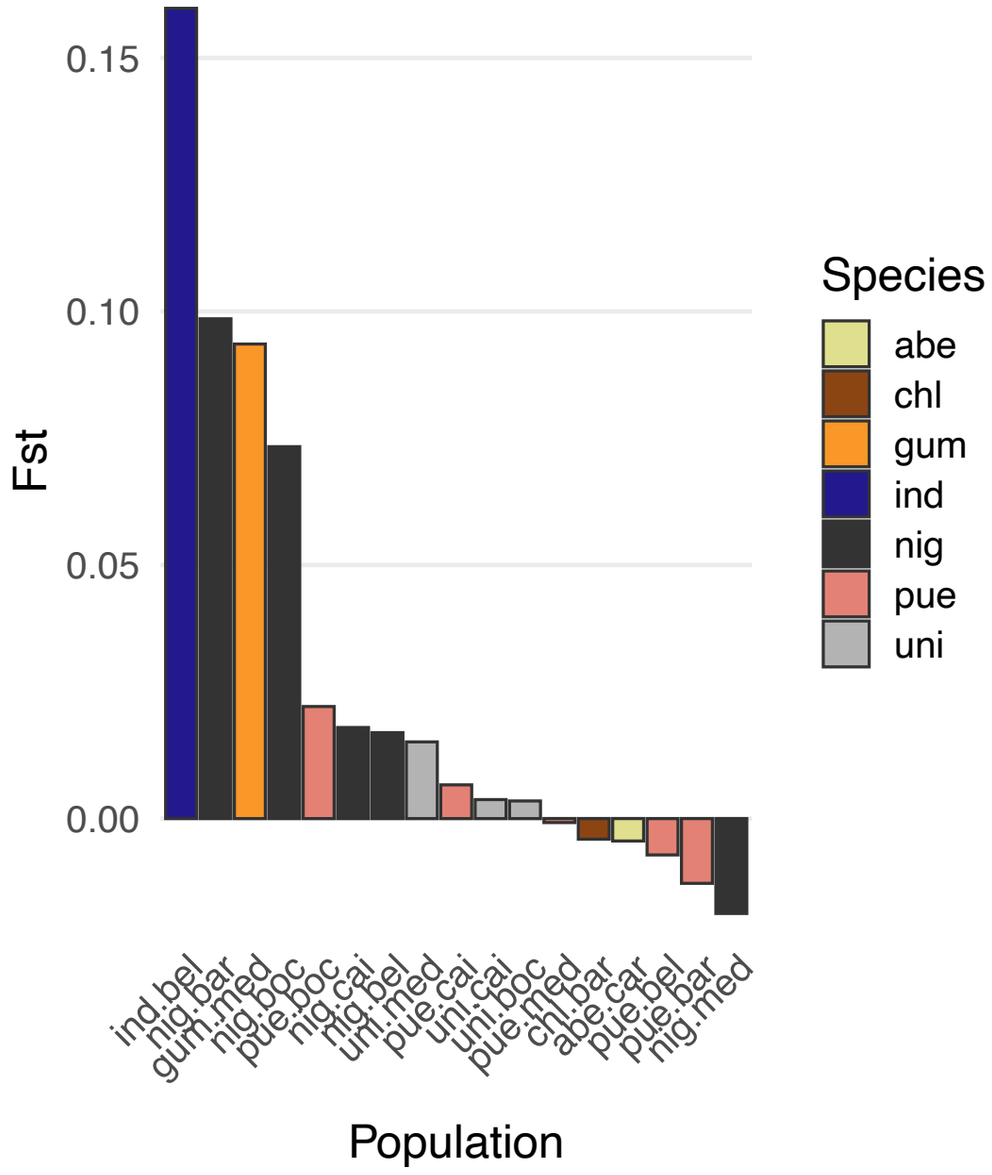
```
tb <- df %>%       # assign to new object (copy)
   mutate() %>%    # 1st operation
   arrange() %>%   # 2nd operation
   select()        # 3rd operation
```

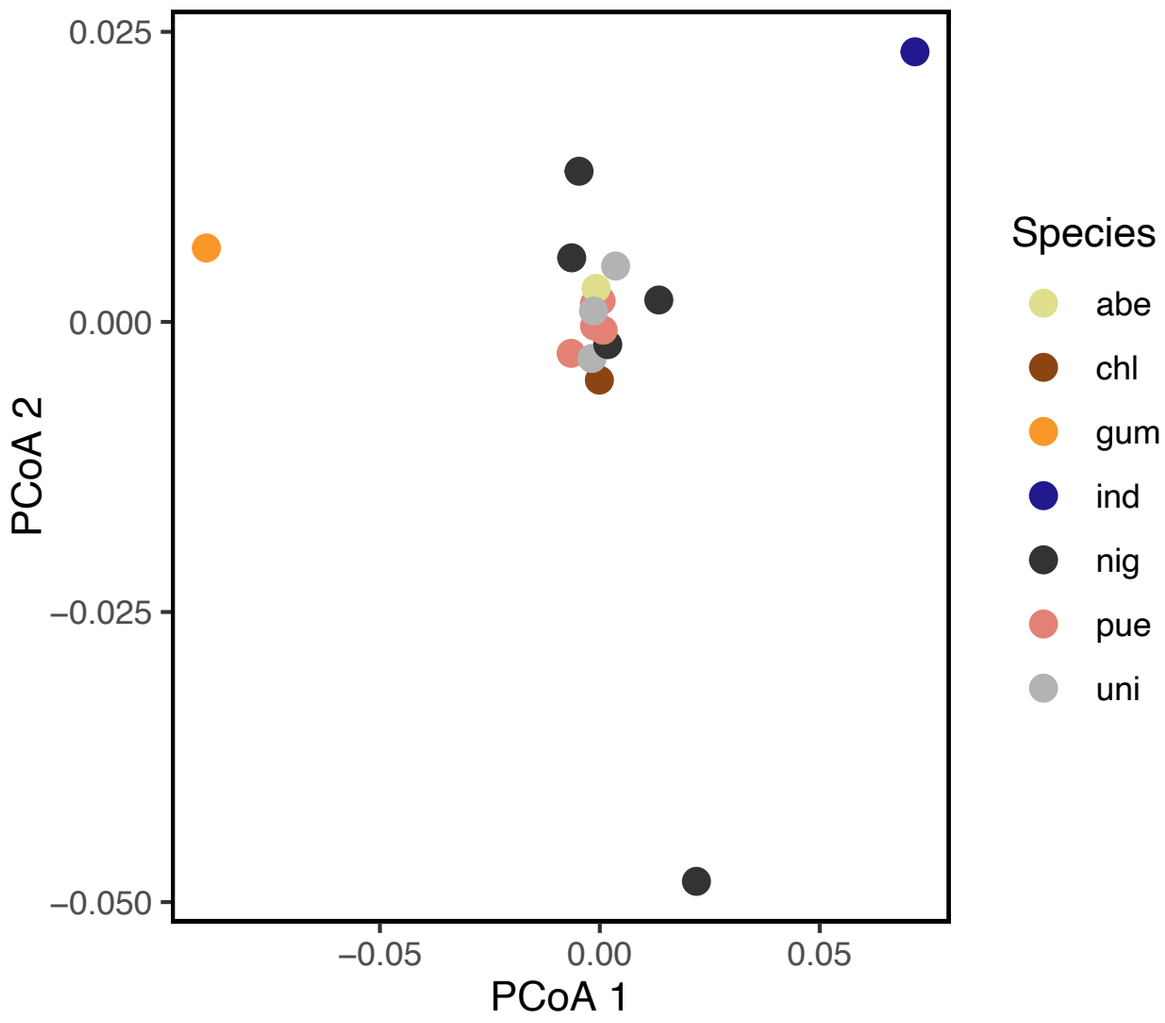
Basic ggplot syntax

```
ggplot(data = tb, aes(x = var1, y = var2)) +  # mapping variables
   geom_bar() +    # geometric shapes representing the data
   labs() +        # set plot and axis labels
   guides() +      # customize plot legend
   theme()         # customize non-data, e.g. fonts, gridlines
```

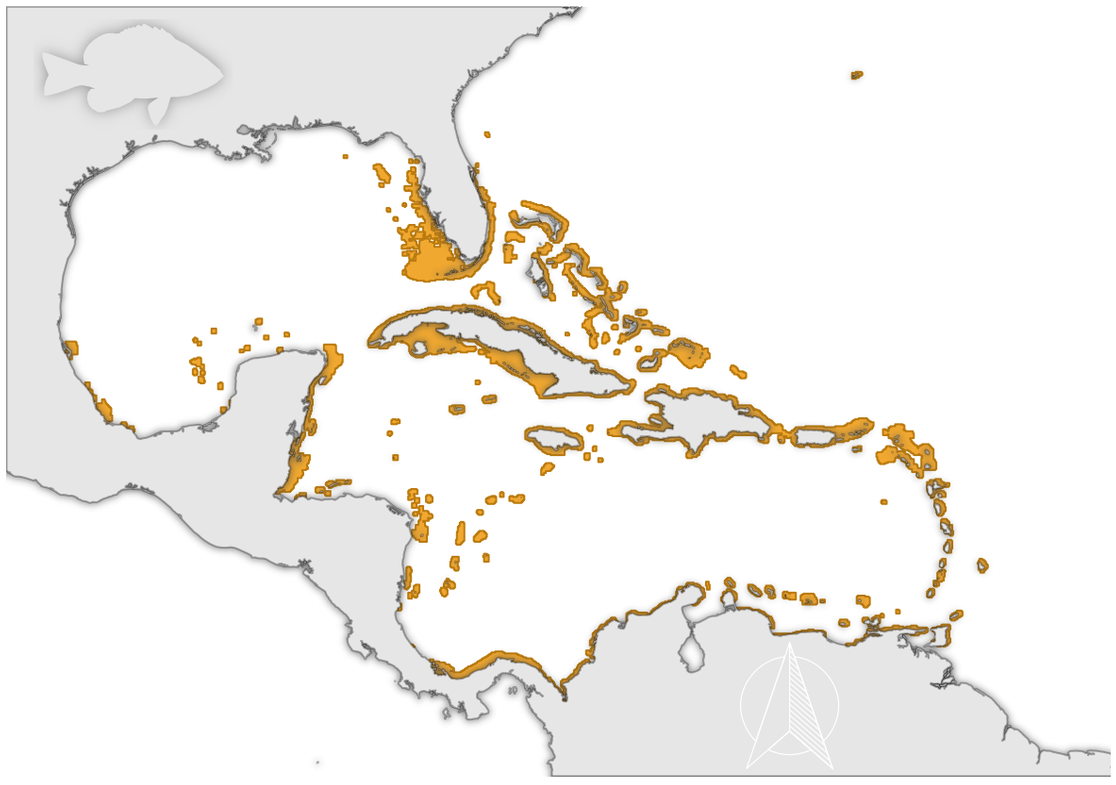
Visualize population structure

Exercises 2/3



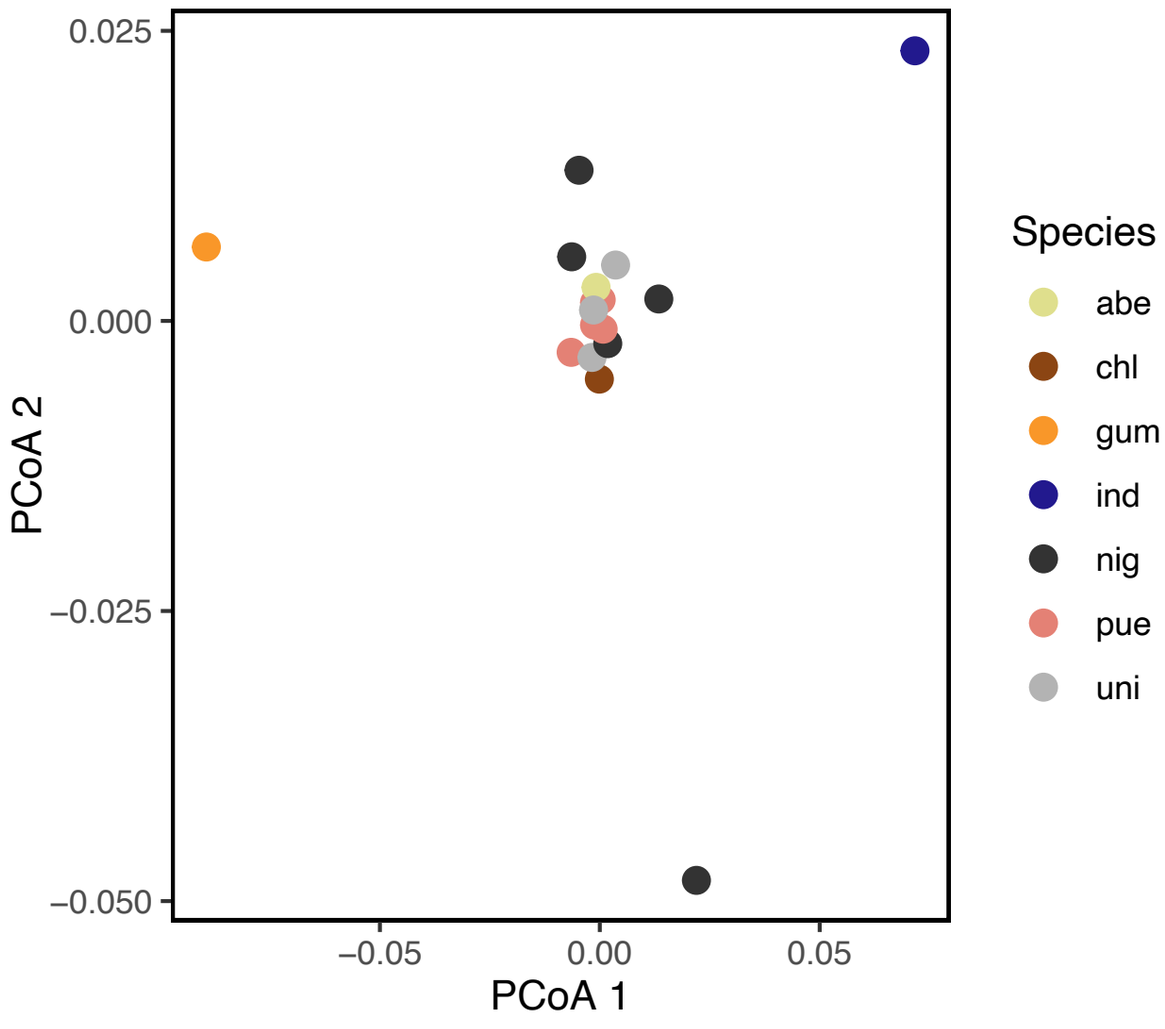


Visualize population structure



Map by Kosmas Hench

data from biogeodb.stri.si.edu (Robertson & Van Tassell 2019)



Take-home messages

- F_{ST} can be estimated globally (across all populations)
  or specifically for each subpopulation
- Genetic differentiation can be statistically tested using the G-test (but no effect size)
- Unsupervised methods, such as PCA and PCoA, help visualize population structure
- Investigating population structure helps disentangle the effects of selection, drift, migration and demographic history
- Identifying distinct genetic populations is essential for the conservation of genetic diversity

Barred hamlet (*H. puella*) microsatellite dataset

Bonus

Hamlets are simultaneous hermaphrodites with external fertilization. Discuss whether self-fertilization occurs regularly according to *F*-statistics.