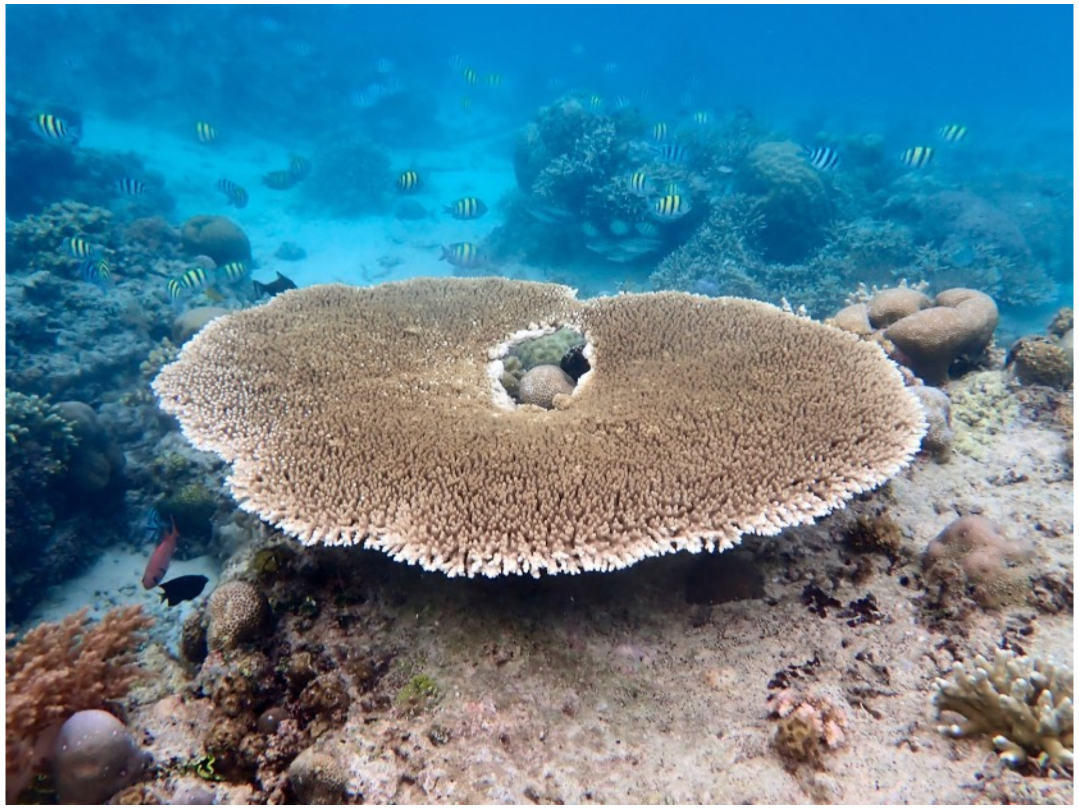
Exercises in Marine Ecological Genetics

06. Population genomics and genetic diversity

- Summarize and filter VCF files
- Reduce linkage disequilibrium in SNP data
- Assess genetic diversity in SNP data

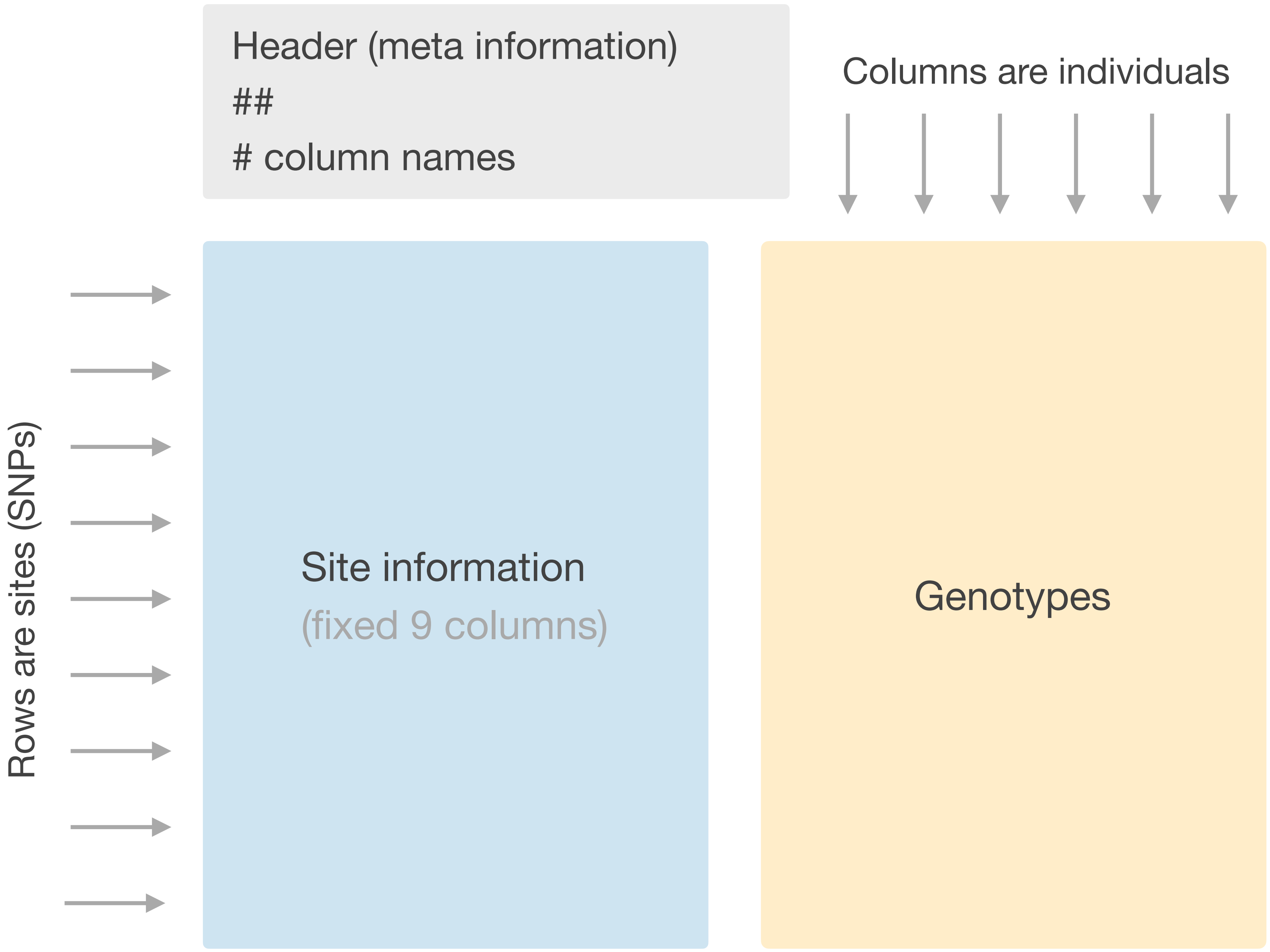


Martin Helmkampf

https://github.com/mhelmkampf/meg25

Variant call format (VCF)

Recap



Efficiently stores

genetic variation data

Optional data e.g:

- allele frequencies
- no. samples with data
- read depth
- genotype likelihood

Variant call format (VCF)

Recap

```
zcat < local/snps_hamlets_filtered.vcf.gz | head
```

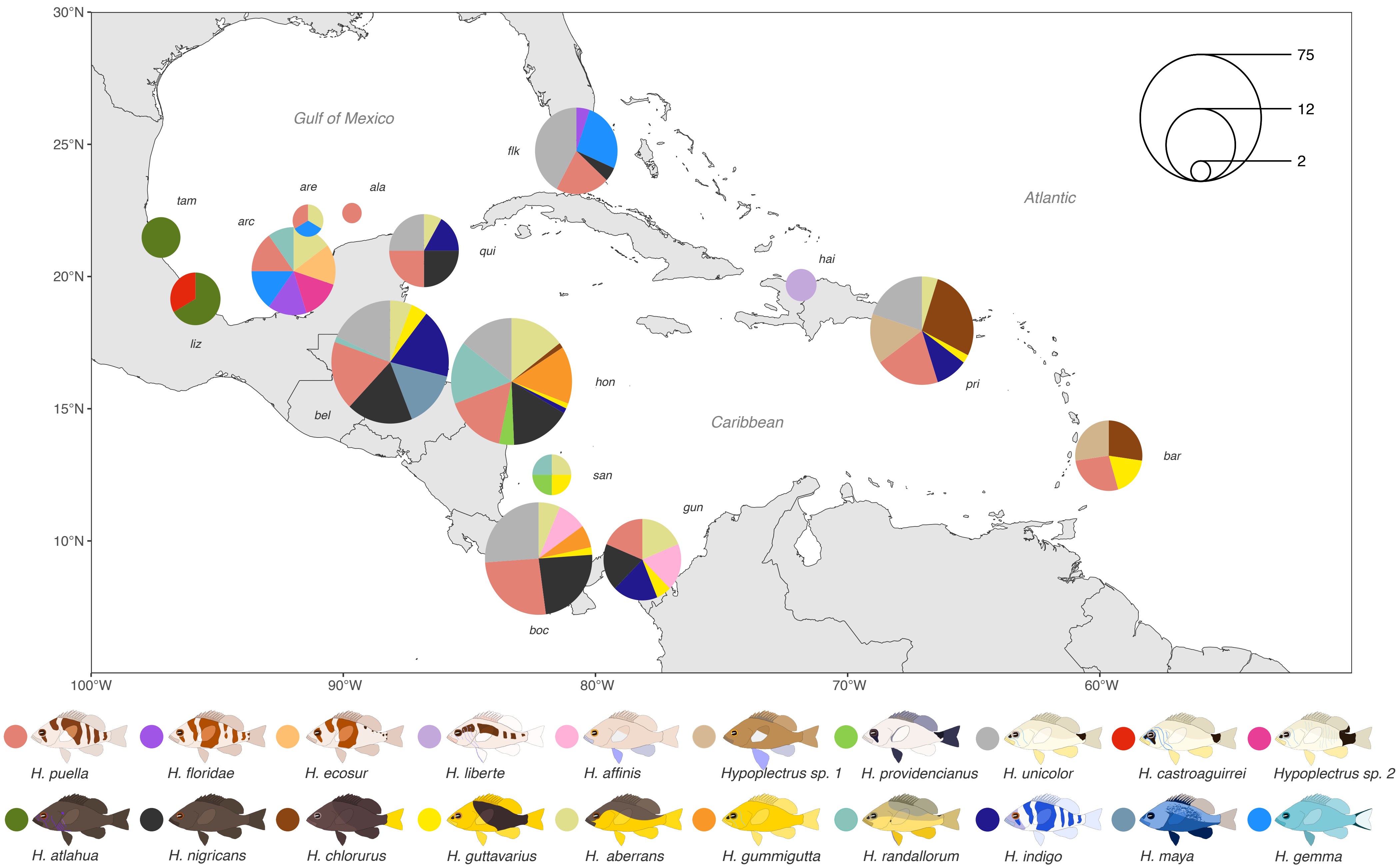
```
##fileformat=VCFv4.1
##fileDate=02012019_20h38m04s
##source=SHAPEIT2.v837
##log_file=shapeit_02012019_20h38m04s_959049fa-700a-4d37-a4ff-3b5db0353190.log
##FORMAT=<ID=GT,Number=1,Type=String,Description="Phased Genotype">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 18158nigbel 18159nigbel 18162nigbel …
LG12 4152 . T G . PASS . GT 0|0 0|0 0|0 …
LG12 4228 . C A . PASS . GT 0|1 0|0 0|1 …
LG12 4262 . A G . PASS . GT 1|0 0|1 1|0 …
LG12 4263 . C T . PASS . GT 0|1 1|0 0|1 …
```

0|0 Homozygous for reference (1st) allele

1|1 Homozygous for alternate (2nd) allele

0|1 and 1|0 Heterozygous

Example dataset



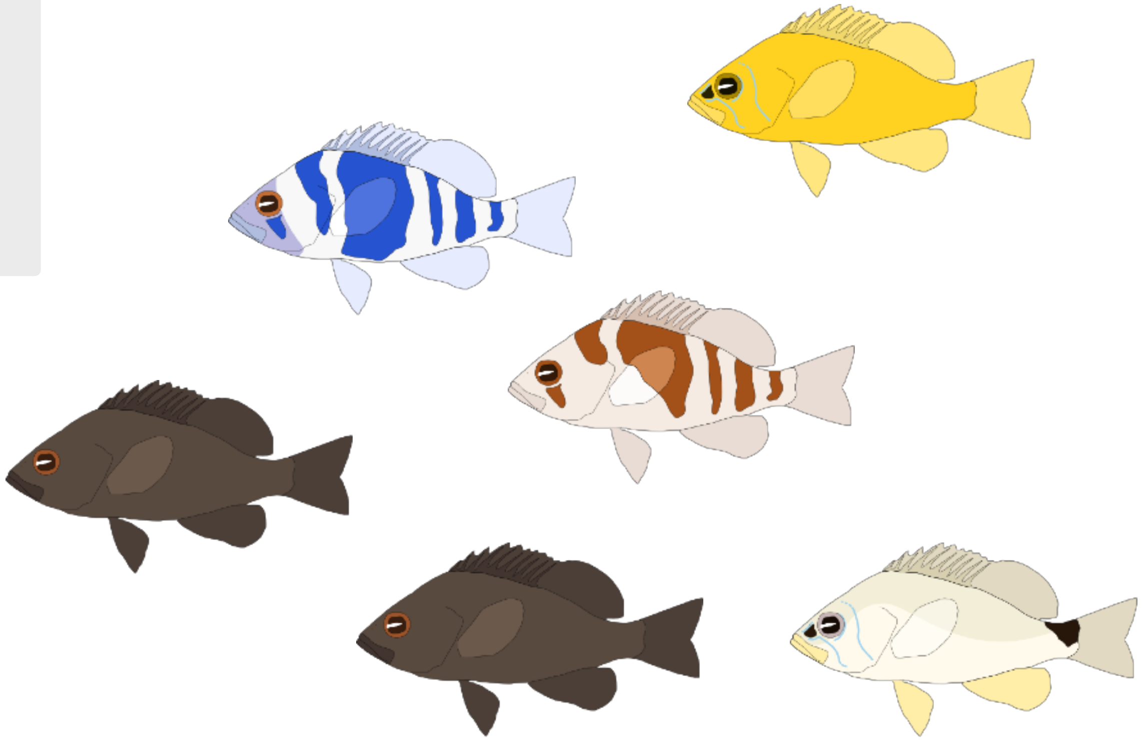
Example dataset

- 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

LD statistics

Exercise 2

$$D_{AB} = p_{AB} - p_A\, p_B$$

Product of allele frequencies

Haplotype frequency

Coefficient of linkage disequilibrium between two alleles

0 to ±1, but constrained by allele frequencies

$$D' = D \,/\, D_{max}$$

Max value given allele frequencies

D normalized with respect to allele frequencies

0 to ±1, full range (0: no association, ±1: perfect LD)

$$r^2 = \frac{D^2}{p_A\,(1 - p_A) p_B (1 - p_B)}$$

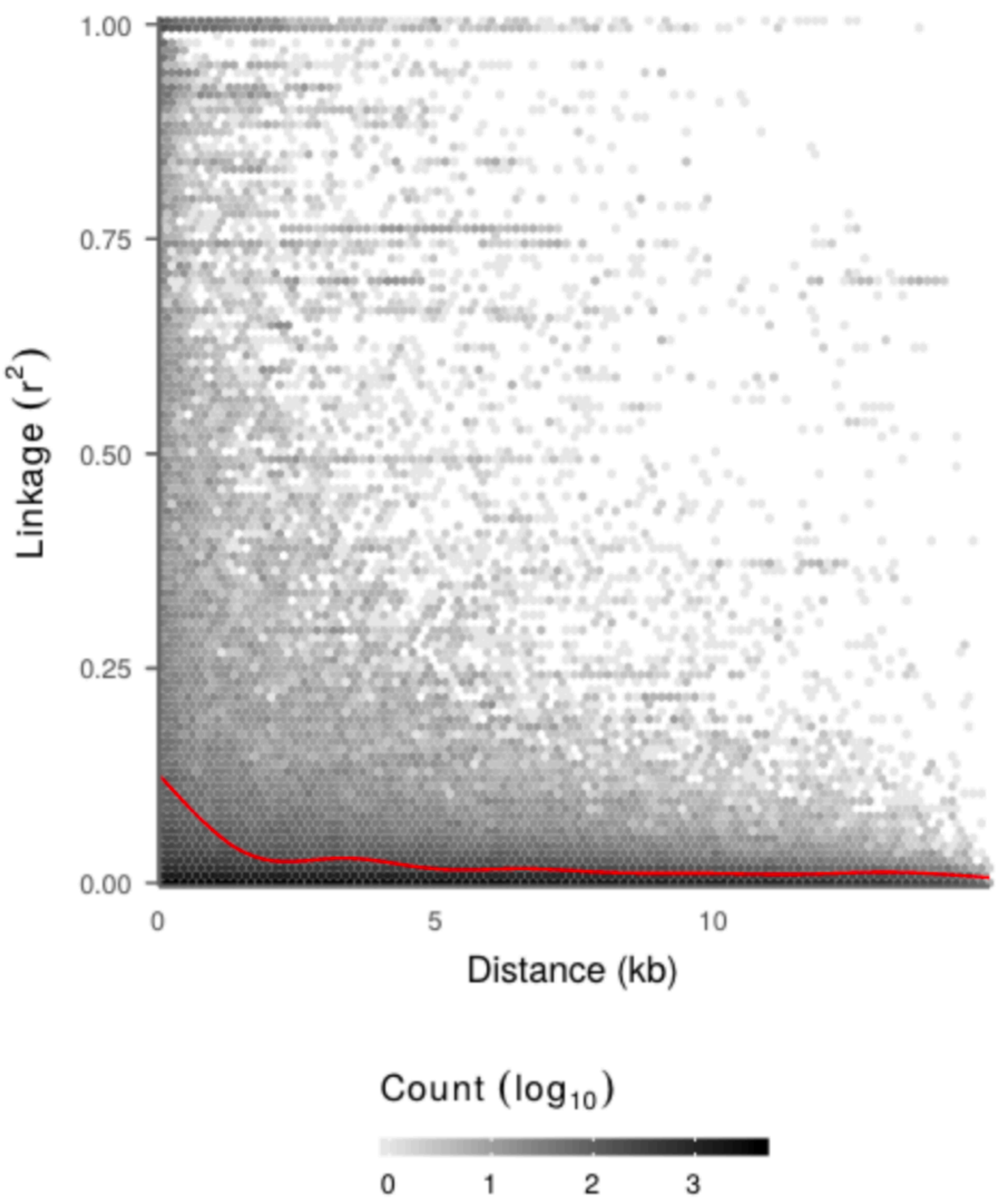
Correlation coefficient of linkage disequilibrium

0 to 1, but constrained by allele frequencies

a.k.a. *ρ* (*rho*)

Decay of linkage with physical distance

Exercise 2



Hench et al. 2019 (Nat Ecol Evol)

Nucleotide diversity π

Exercise 3

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

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

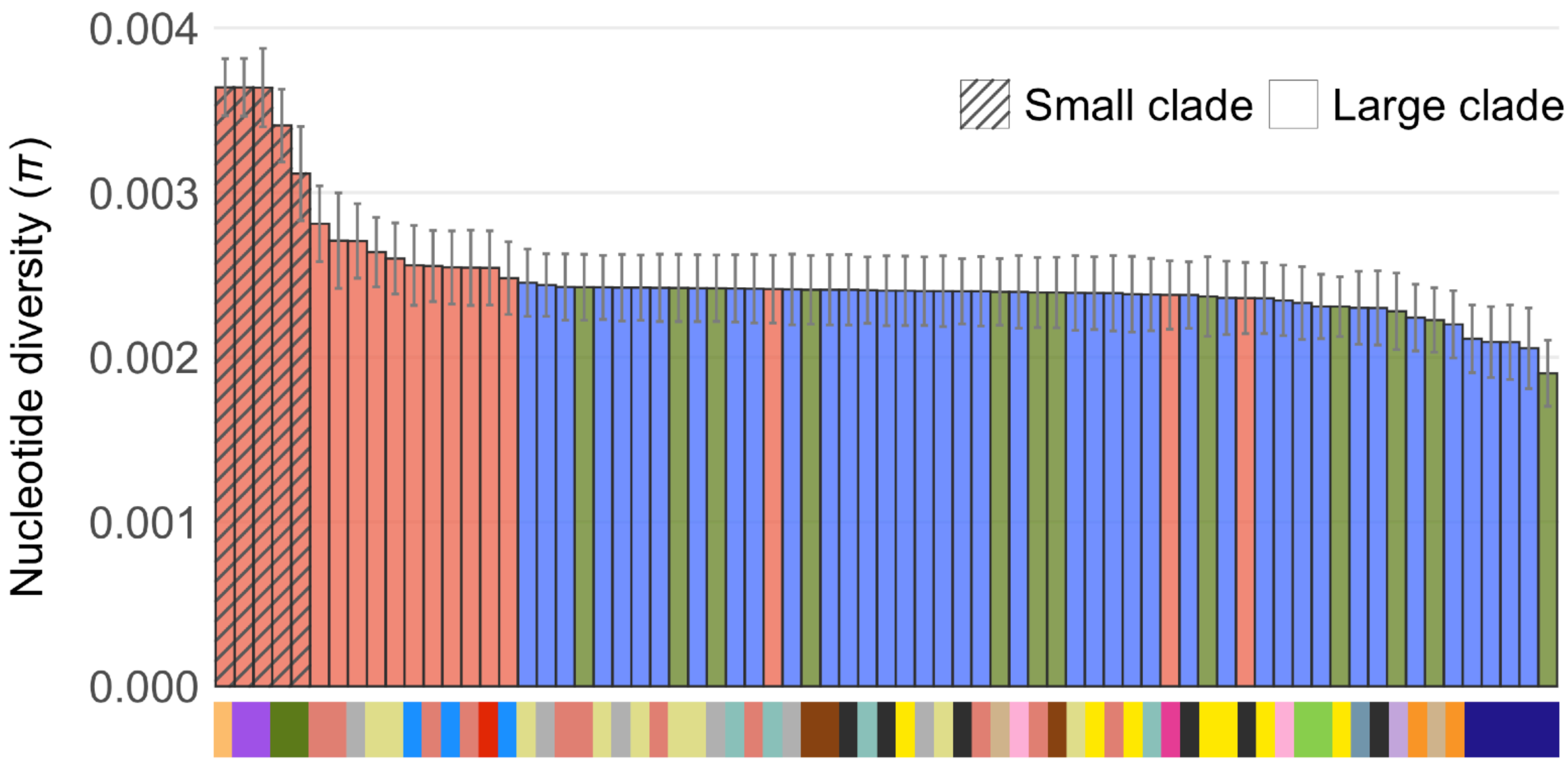
- 2 chromosomes per individual
- 15 pairwise comparisons
- π per site: no. diff / comparisons

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
π per site:	0.53	0	0.60	0.33	0

Average π = 0.29 ←

Nucleotide diversity π

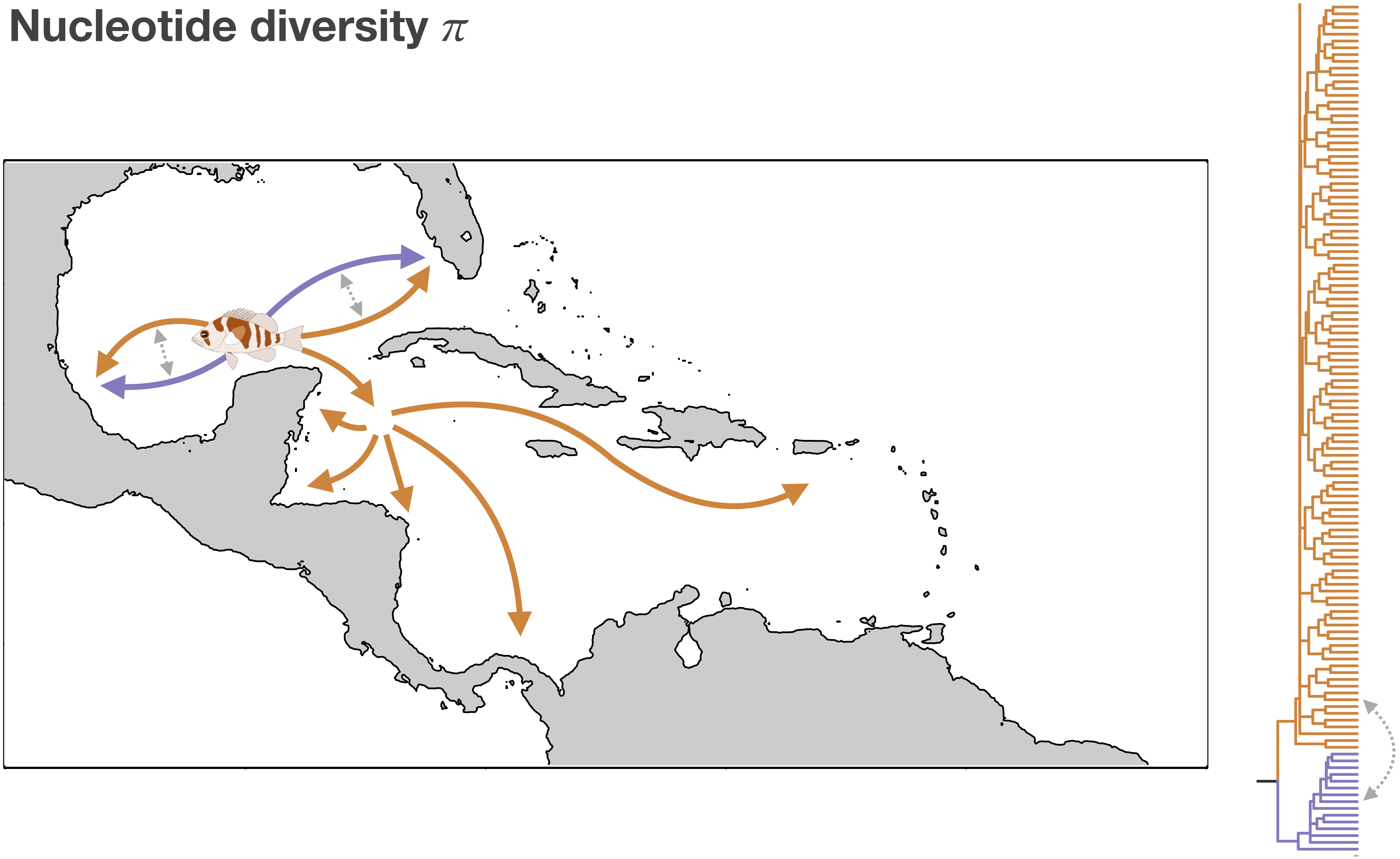
Exercise 3



Summer 2025

Carl von Ossietzky Universität Oldenburg

Nucleotide diversity π



Map data provided by NOAA

Population genomics and genetic diversity

Summary

```
vcftools --gzvcf … --mac 2 --thin 2000 --rcode …   # Filter by MAC and distance

vcftools --gzvcf … --het --stdout > …                # Calculate heterozygosity

vcftools --gzvcf … --keep <pop.txt> --site-pi --out …            # Calculate pi
```

- SNPs must be filtered carefully, e.g. with respect to minor allele count or missing data, to ensure high-quality results

- Removing SNPs in linkage disequilibrium is important for analyses that assume independence between loci

- Genome-wide statistics like heterozygosity or nucleotide diversity (pi) provide valuable information about genetic variation within populations