Exercises in Marine Ecological Genetics

08. Recombination and linkage disequilibrium

- Visualize genetic structure in SNP dataset using PCA
- Calculate heterozygosity measures with SNP data
- Filter SNPs by linkage
- Estimate N_e using linkage disequilibrium

Martin Helmkampf

Summer 2023

Carl von Ossietzky
Universität
Oldenburg

Example dataset

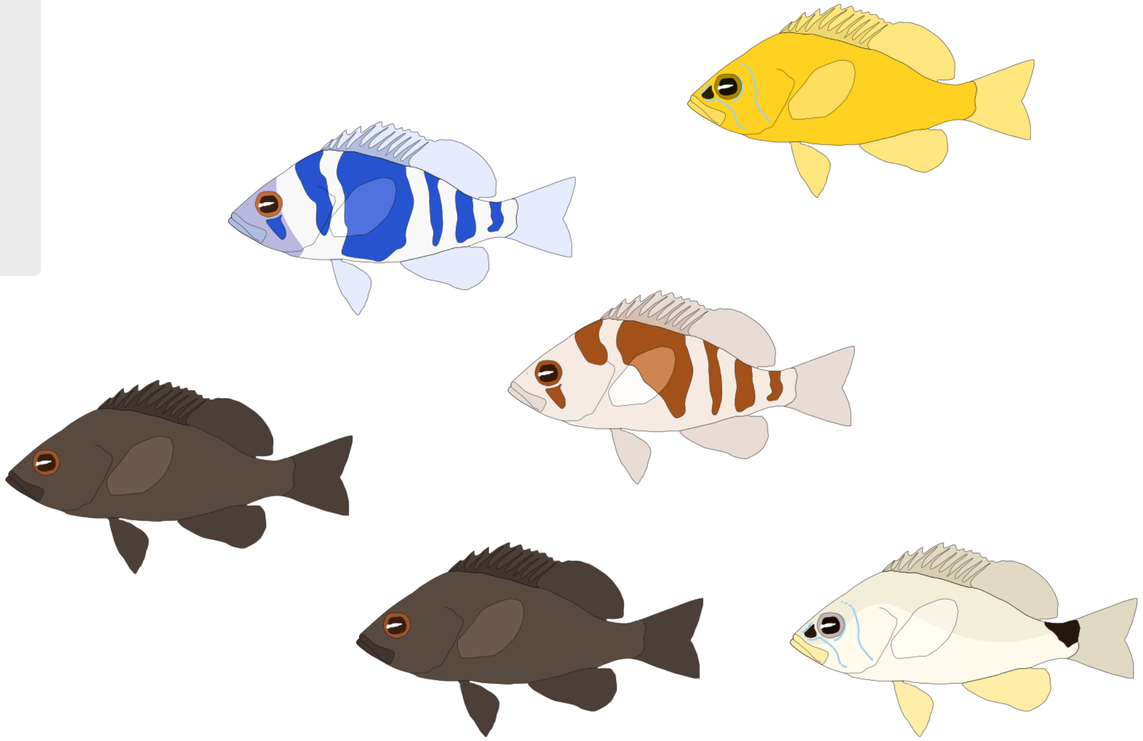
Recap

- 8 species of hamlet (genus *Hypoplectrus*)
- 3 Caribbean sites: Belize, Honduras, Panama
- 167 hamlet samples total
- Illumina short-read resequencing (mean depth 17×)
- Genotyping with GATK
- High-quality reference genome of *H. puella*

↓

snps_hamlets_lg12.vcf.gz

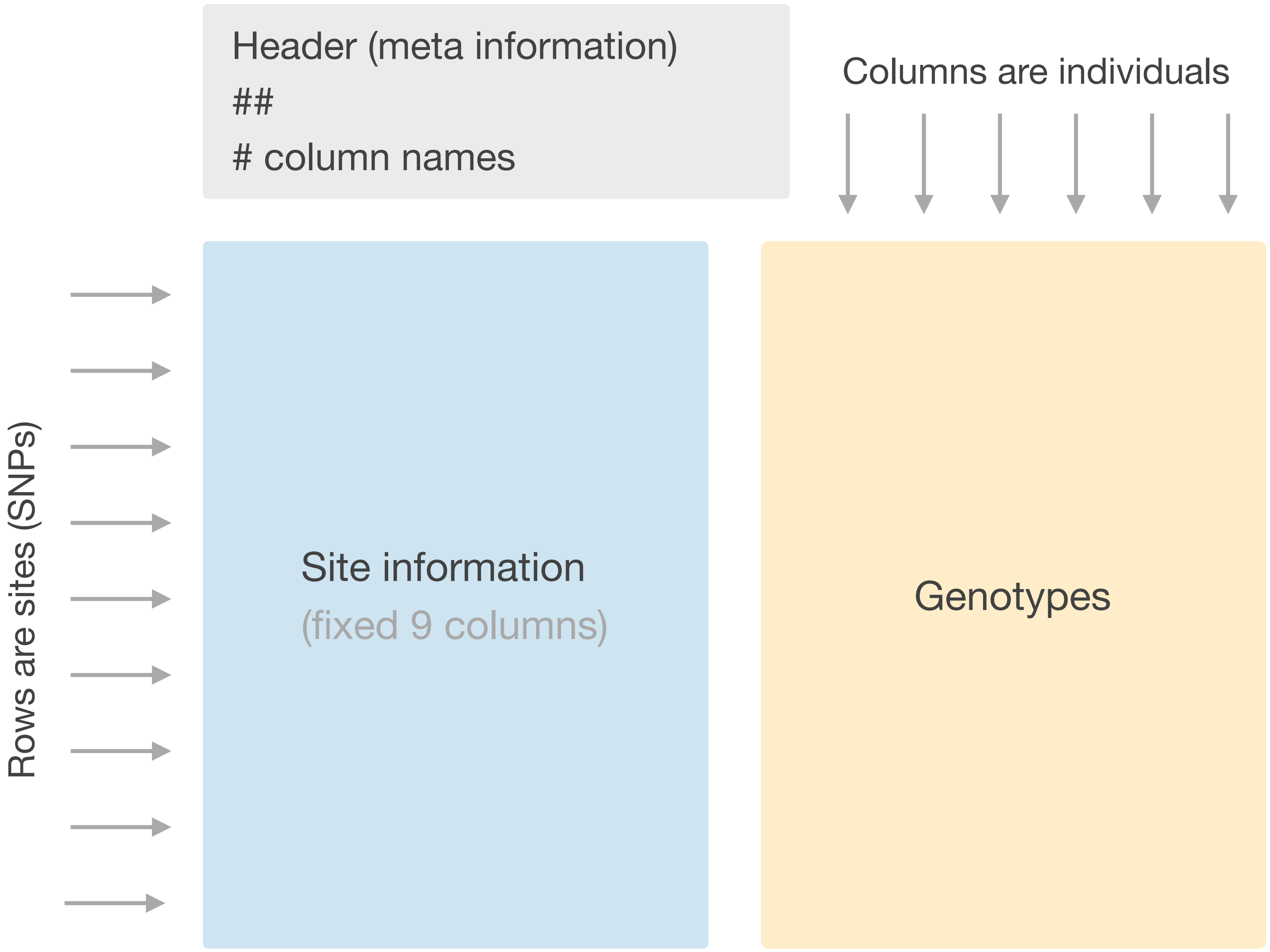
- Chromosome 12 only
- Subset to 36 samples from 6 populations



Illustrations by Kosmas Hench

Variant call format (VCF)

Recap



Header (meta information)

##

# column names

Columns are individuals

Rows are sites (SNPs)

Site information

(fixed 9 columns)

Genotypes

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

VCFtools

Recap

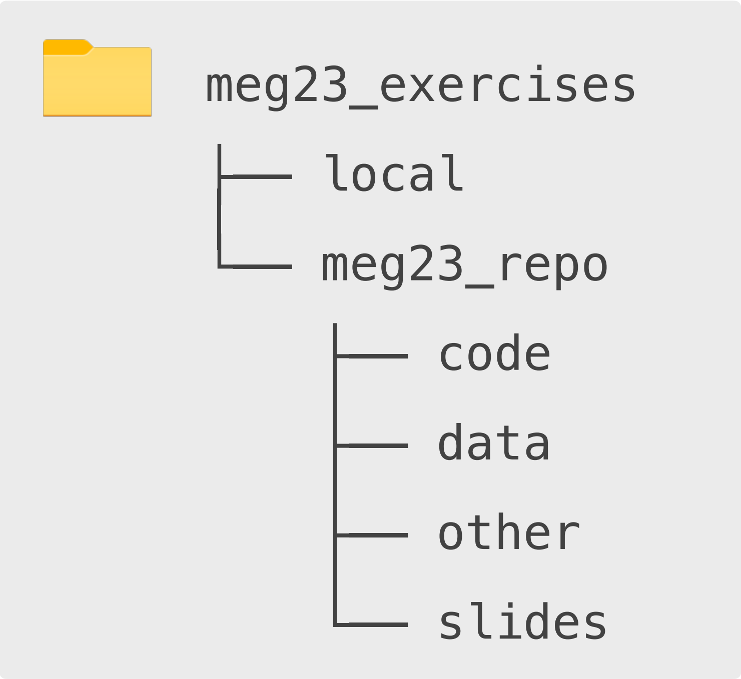
```
### Filter sites (rows)
vcftools \
    --gzvcf snps_hamlets_lg12.vcf.gz \
    --max-missing 1 \
    --mac 2 \
    --recode \
    --stdout | bgzip > snps_hamlets_filtered.vcf.gz

### Calculate heterozygosity and Fis for each individual
vcftools \
  --gzvcf snps_hamlets_filtered.vcf.gz \
  --het \
  --stdout > Het_hamlets_snps.tsv
```

Download course materials using git

Go to project directory

```
cd dir         # e.g. Documents/meg23_exercises
ls -l     # view directory contents, long format
```



```
meg23_exercises
  ├── local
  └── meg23_repo
        ├── code
        ├── data
        ├── other
        └── slides
```

Update course repository

```
cd meg23_repo
git pull
```

In case of an error message

```
cd ..                                      # go back to project directory
rm -rf meg23_repo                                # delete old repository
git clone https://github.com/mhelmkampf/meg23_repo.git
```

Avoiding version conflict

Please do not save over files in the course repository. Instead, save your own scripts to the local subdirectory (including copies of course scripts you would like to edit), e.g with

```
cp code/08_recomb.sh ../local/08_recomb_lc.sh        # cp [source] [destination]
```

Get set up on the HPC cluster

Connect to login node

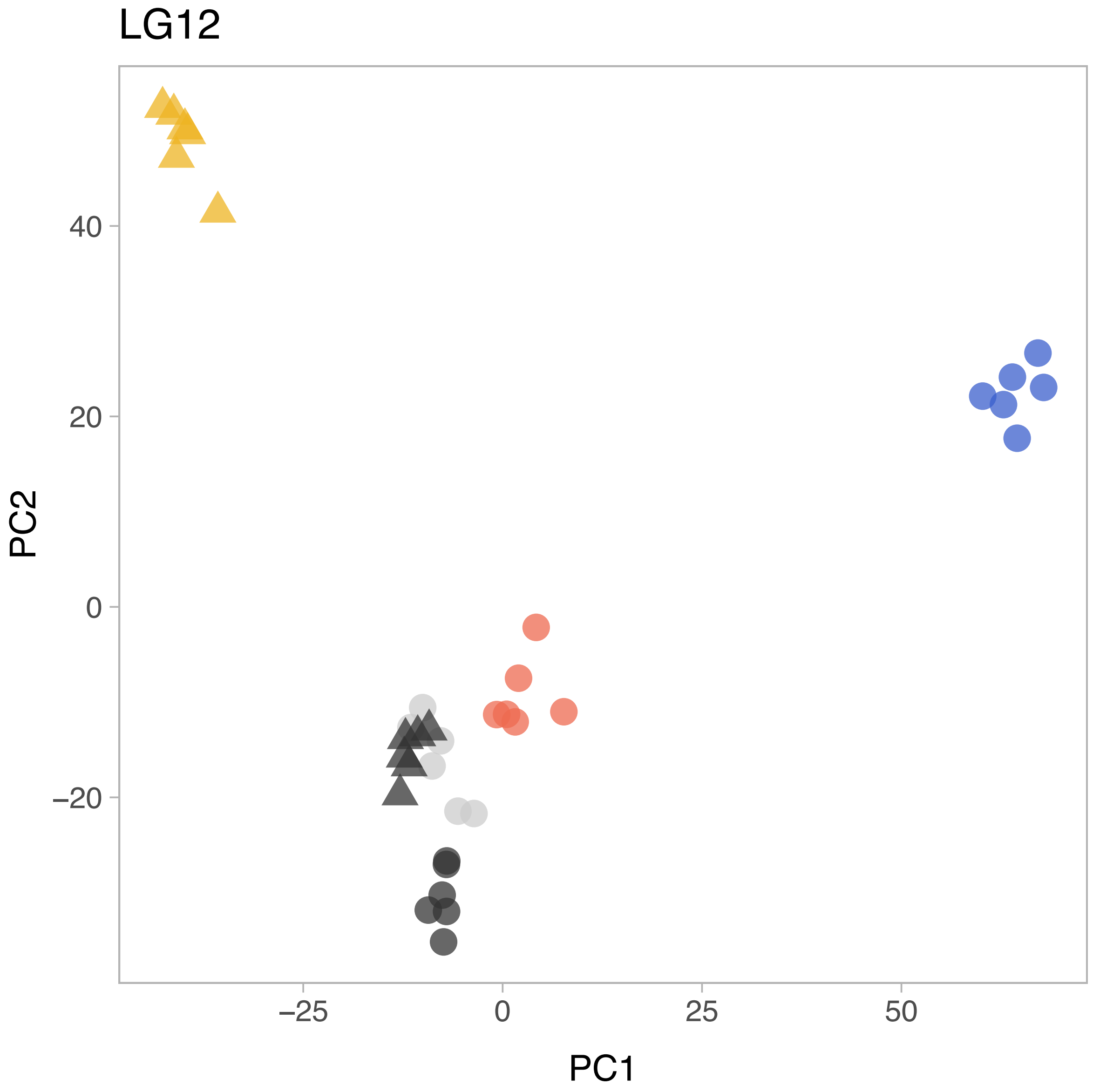
```
ssh <account>@carl.hpc.uni-oldenburg.de
# Account ids and passwords can be found on StudIP in Files | course_accounts.csv
```

Download course materials to cluster account using git

```
cd meg23_repo
git pull
# first time: git clone https://github.com/mhelmkampf/meg23_repo.git
```

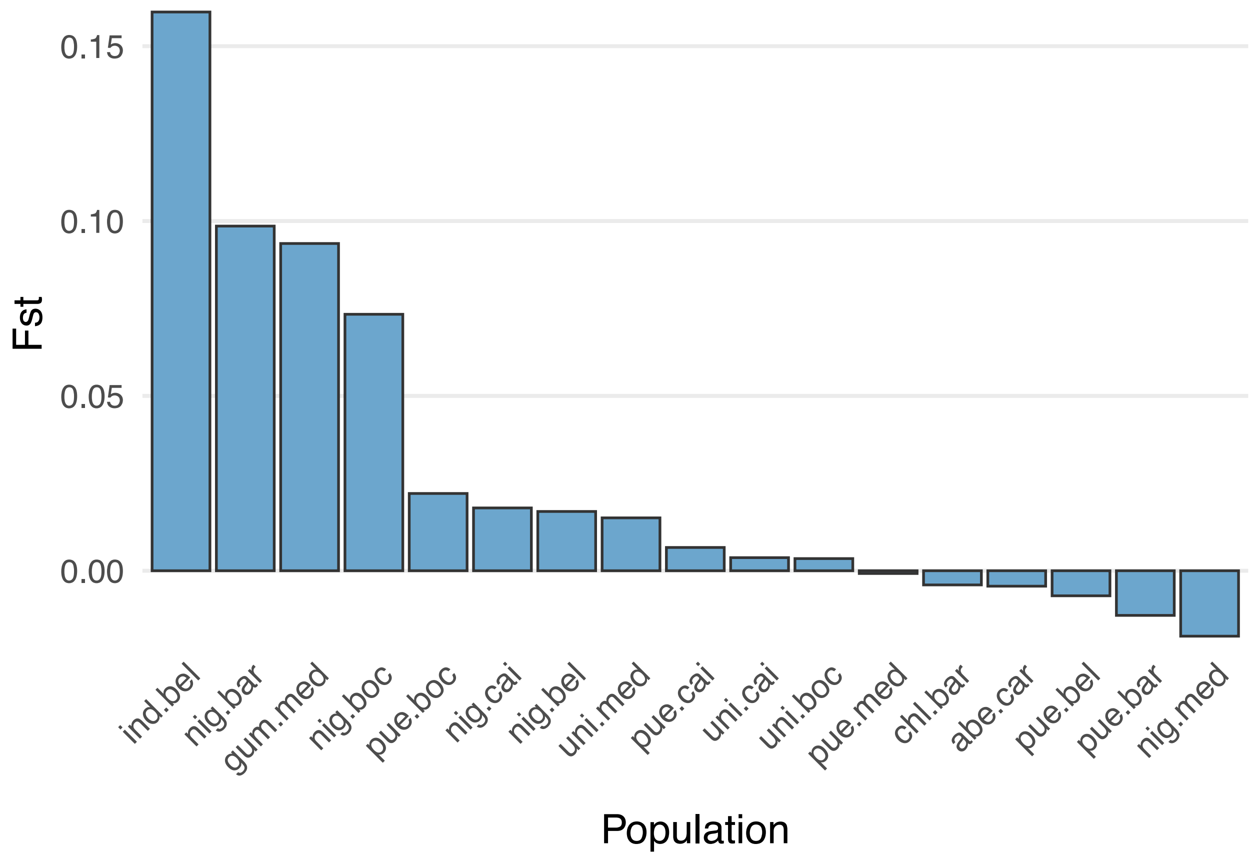
Compare PCA with microsatellite-based F_{ST}

Exercise 1



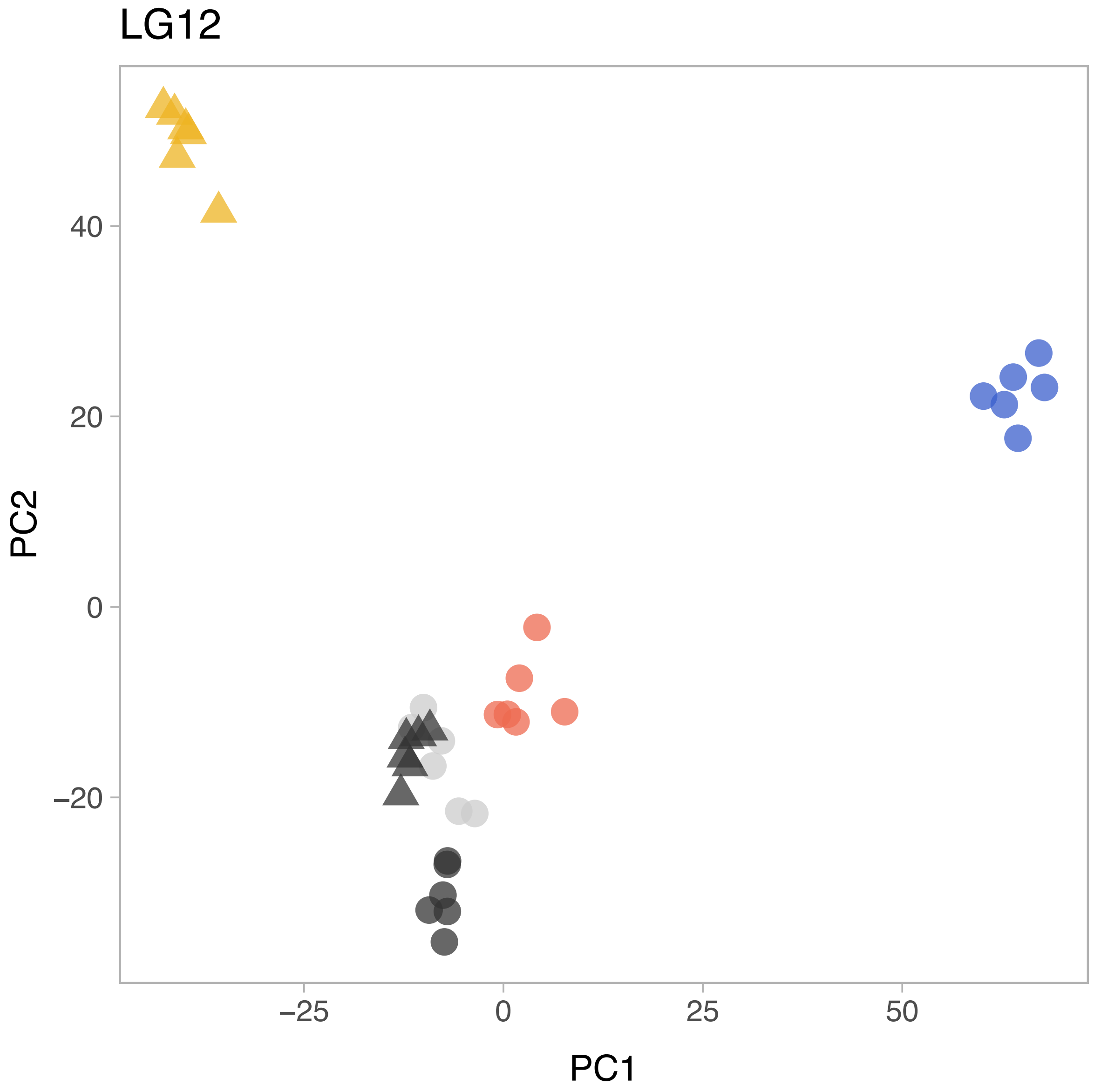
boc = pan

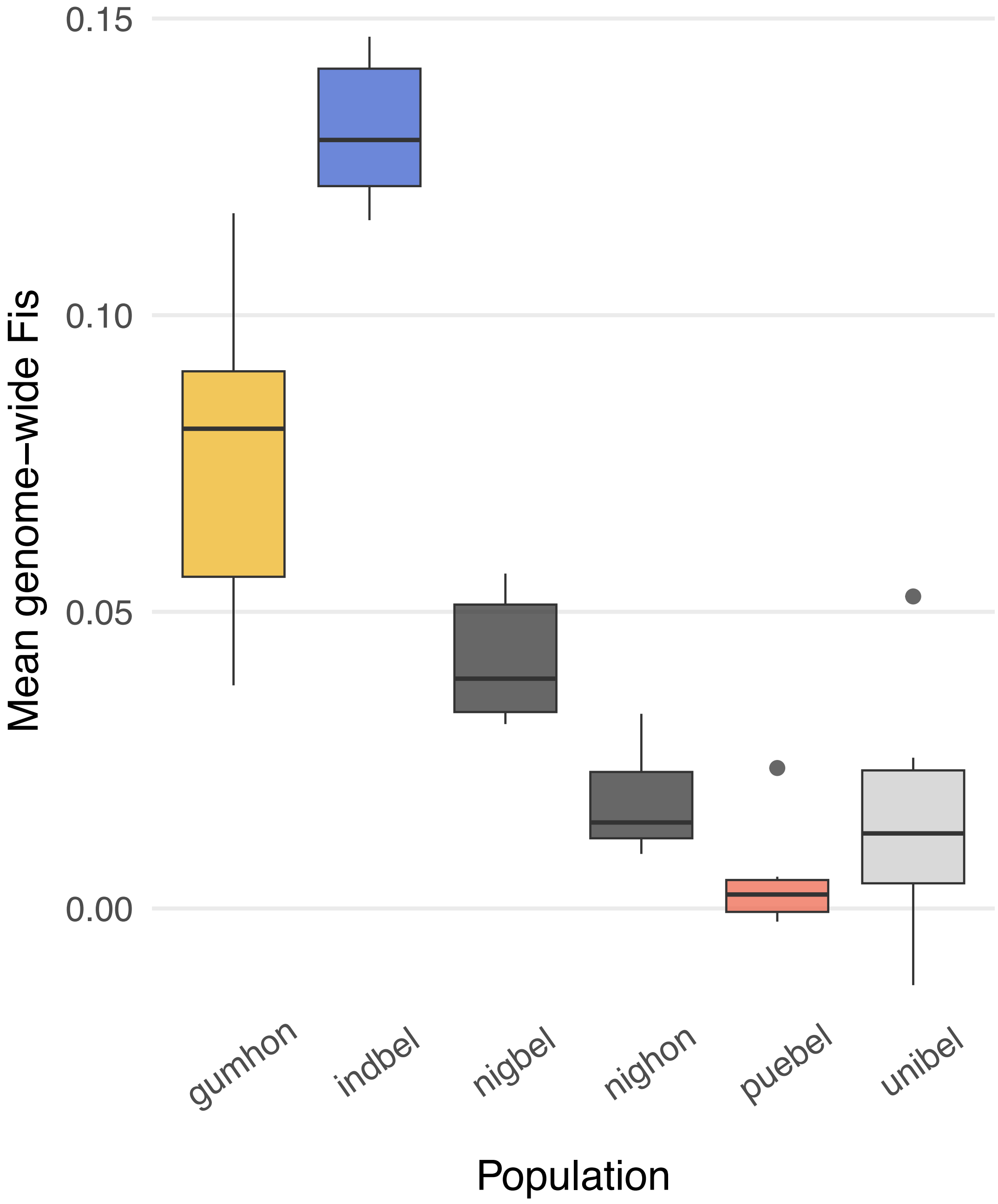
med ~ hon



Compare PCA with genome-wide F_{IS}

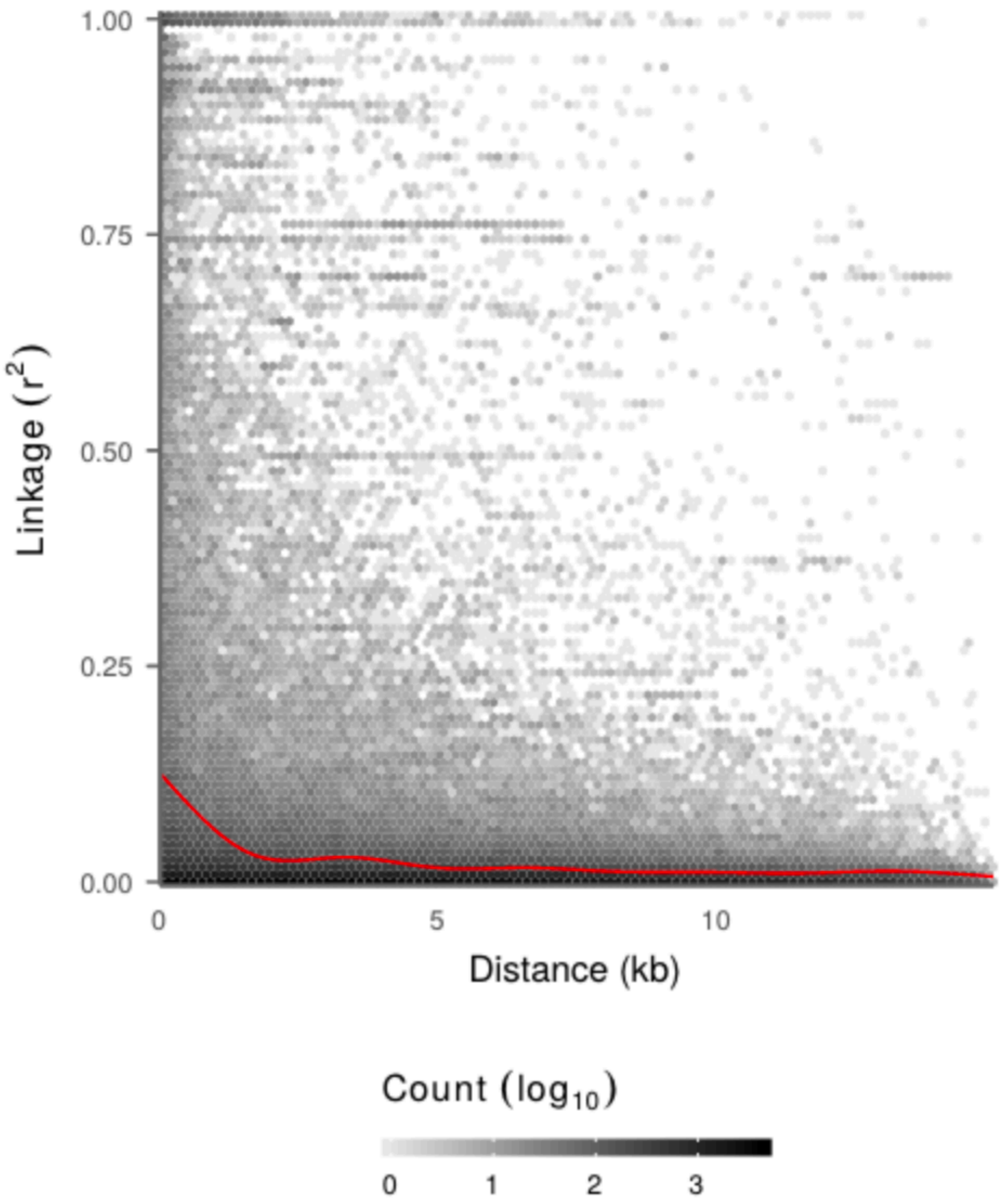
Exercise 1/2





Decay of linkage with physical distance

Exercise 3



Hench et al. 2019 (Nat Ecol Evol)

LD statistics

Exercise 3

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

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)

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

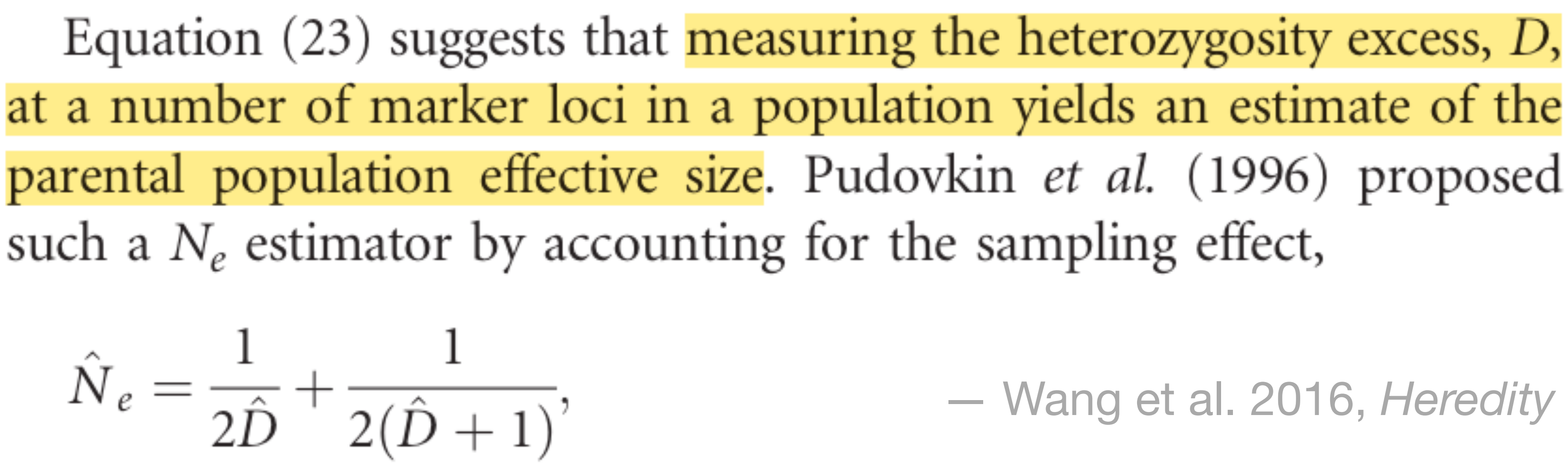
Correlation coefficient of linkage disequilibrium

0 to 1, but constrained by allele frequencies

a.k.a. ρ (*rho*)

Recap: Estimate N_e using heterozygote excess

Exercise 4



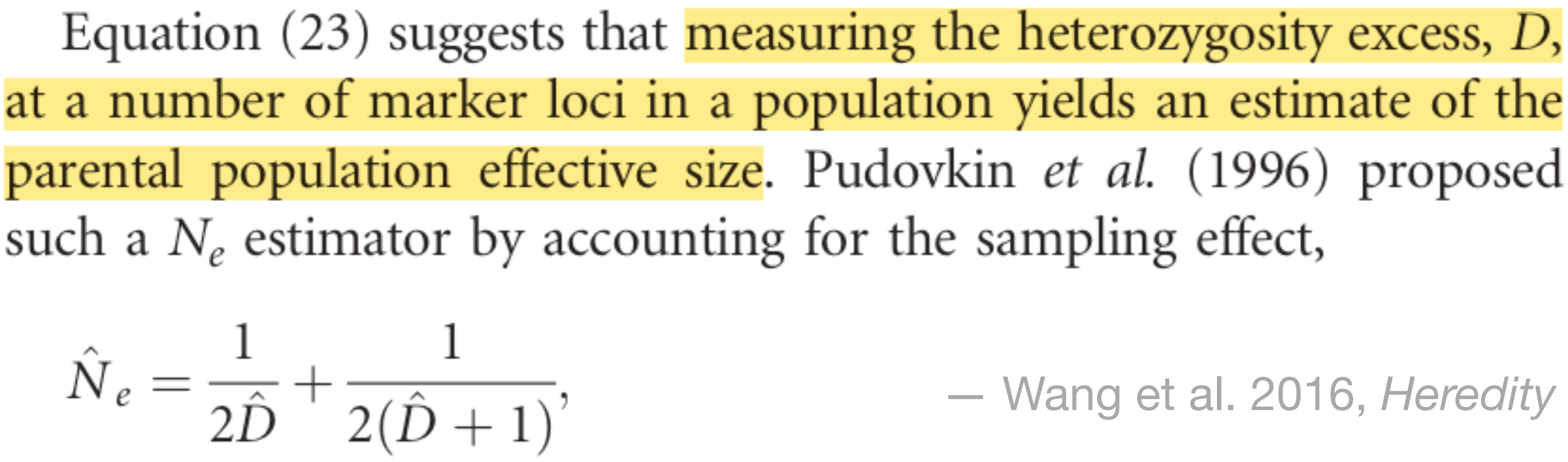
Equation (23) suggests that measuring the heterozygosity excess, *D*, at a number of marker loci in a population yields an estimate of the parental population effective size. Pudovkin *et al.* (1996) proposed such a N_e estimator by accounting for the sampling effect,

$$\hat{N}_e = \frac{1}{2\hat{D}} + \frac{1}{2(\hat{D}+1)},$$

— Wang et al. 2016, *Heredity*

Recap: Estimate *N*$_e$ using heterozygote excess

Exercise 4



Equation (23) suggests that measuring the heterozygosity excess, *D*, at a number of marker loci in a population yields an estimate of the parental population effective size. Pudovkin *et al.* (1996) proposed such a N_e estimator by accounting for the sampling effect,

$$\hat{N}_e = \frac{1}{2\hat{D}} + \frac{1}{2(\hat{D}+1)},$$

— Wang et al. 2016, *Heredity*

$$D_j(i) = \frac{H_j^{obs}(i) - H_j^{exp}(i)}{H_j^{exp}(i)}$$

Index of heterozygote excess

H$_j$(i): observed / expected frequency of heterozygotes having allele *i* at locus *j*

Estimate N_e using LD

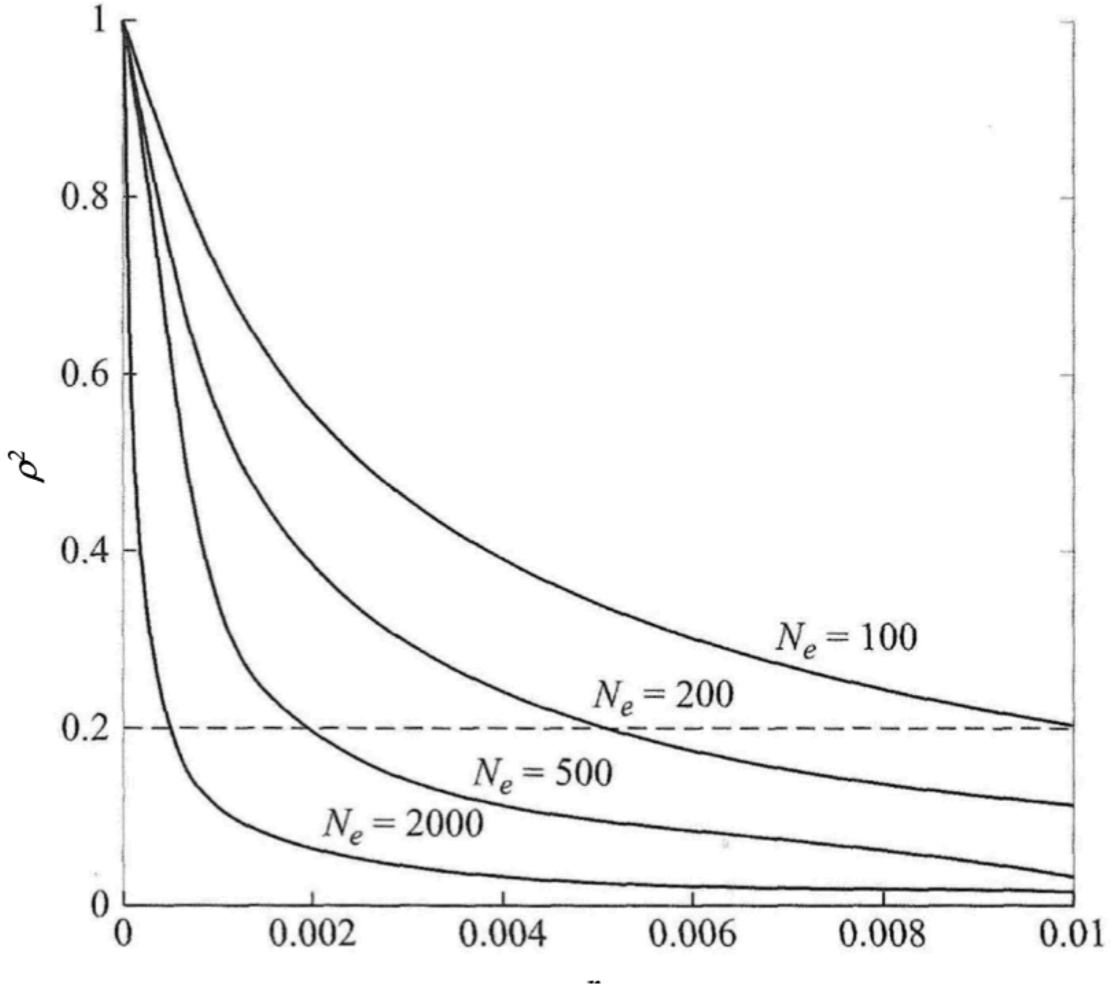
Exercise 4

Sampling error in finite populations may result in deviations from independent segregation (imagine the four gamete types as one locus with four alleles)

A_1 B_1 A_1 B_2 A_2 B_1 A_2 B_2

At equilibrium between drift and recombination, it can be shown that

$$\rho^2 \approx \frac{1}{1 + 4N_e r}$$



Drift generates linkage disequilibrium

Linkage can therefore be used to estimate effective population size!

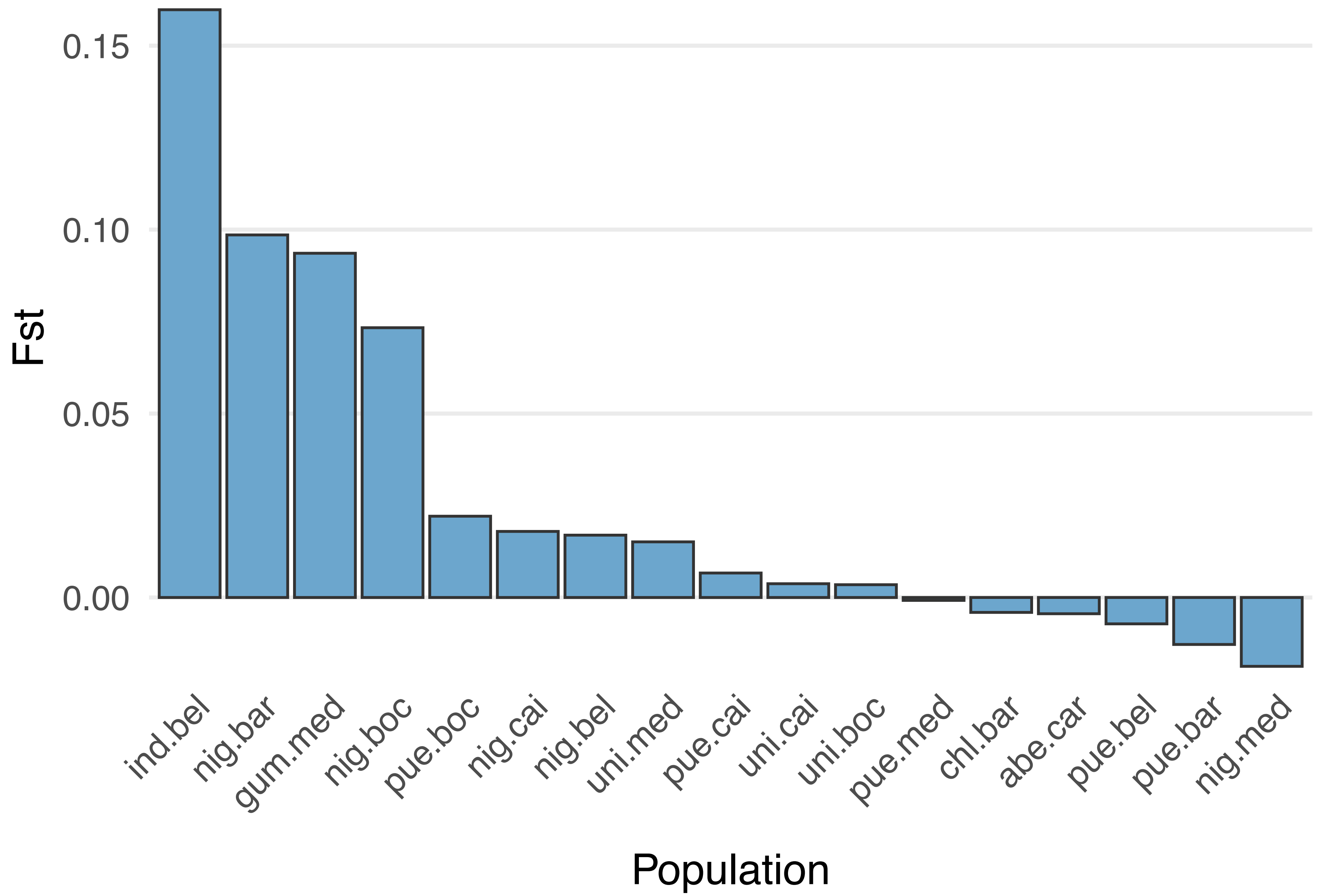
Course outline

May be subject to change

Class	Date	Topics	Script
01	Apr 14	Introduction, software installation	01_intro.R
02	Apr 21	Hardy-Weinberg equilibrium	02_hwe.R
03	Apr 28	Genetic drift and effective population size	03_drift.R
04	May 05	Population structure and gene flow	04_structure.R
05	May 12	Isolation by distance (lecture online, exercises in person)	05_ibd.R
–	May 19	Himmelfahrt break	–
06	May 26	Genome sequencing and assembly	06_genseq.sh
07	Jun 02	Genotyping, SNPs and population genomics	07_snps.R
08	Jun 09	Recombination and linkage disequilibrium	08_linkage.R
–	Jun 16	Student presentations	–
09	Jun 23	Selection and mutation	09_selection.R
10	Jun 30	DNA barcoding	10_barcode.sh
11	Jul 07	Metabarcoding	11_meta.sh
–	Jul 14	To be determined	–

Compare PCA with population-specific F_{ST}

Exercise 3



In PCA:

boc = pan

med ~ hon