

# Exercises in Marine Ecological Genetics

## 07. Genotyping, SNPs and population genomics

- Get overview of whole-genome genotyping
- View and filter VCF files storing SNP data
- Run population genetic calculations on SNP data
- Work on a high performance computing cluster

Martin Helmkamp



# Download course materials using git

Go to project directory

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

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/07_snps.sh ../local/07_snps_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

```
git pull  
# first time: git clone https://github.com/mhelmkamp/meg23\_repo.git
```

# Genome assembly recap

- Reconstructing long, continuous sequence from millions of overlapping **reads**
- Reads can be very short (e.g. Illumina) or long (e.g. PacBio)
- Segments of assembled sequence are called **contigs**, which may be combined into **scaffolds**
- Scaffolds or PacBio contigs can be up to chromosome-length



# Genome assembly recap

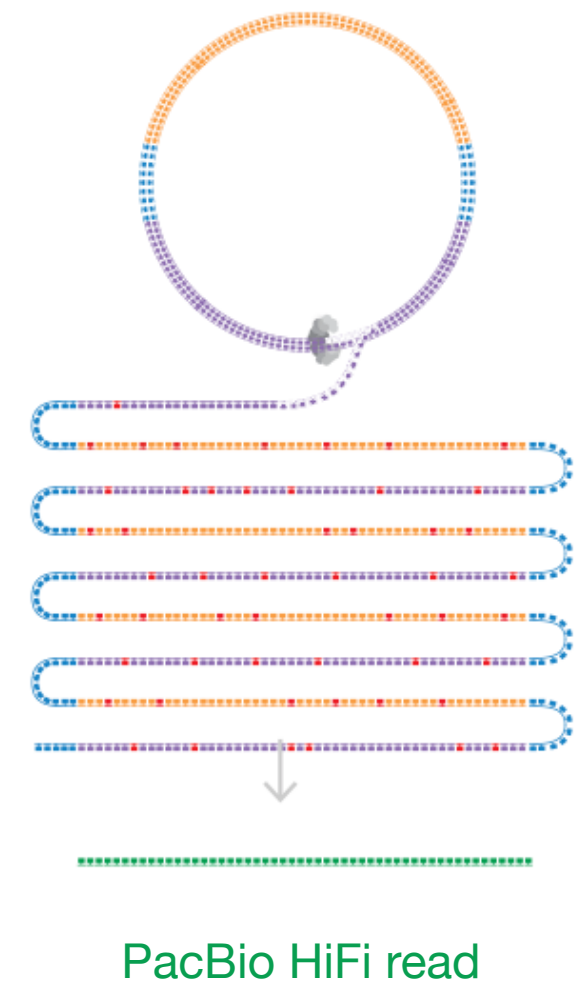
- Reconstructing long, continuous sequence from millions of overlapping **reads**
- Reads can be very short (e.g. Illumina) or long (e.g. PacBio)
- Segments of assembled sequence are called **contigs**, which may be combined into **scaffolds**
- Scaffolds or PacBio contigs can be up to chromosome-length



Genome

Reads

Contigs / scaffolds



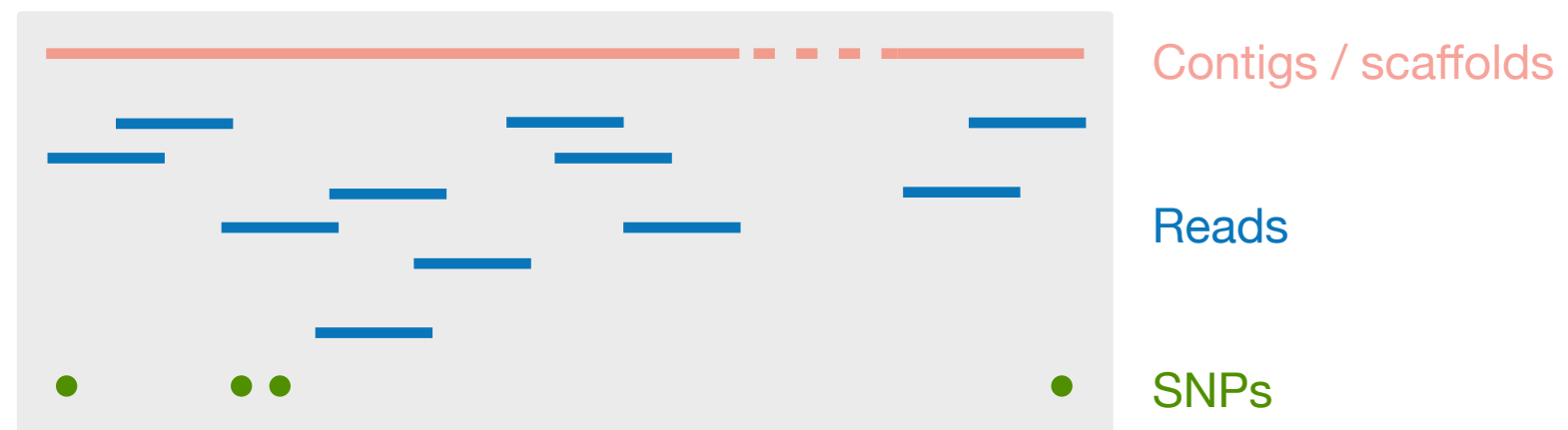
PacBio HiFi read

# Genome sequencing strategies

*De novo*



Re-sequencing



~ Reduced representation sequencing, e.g. RADseq

# Sequencing reads in FASTQ format

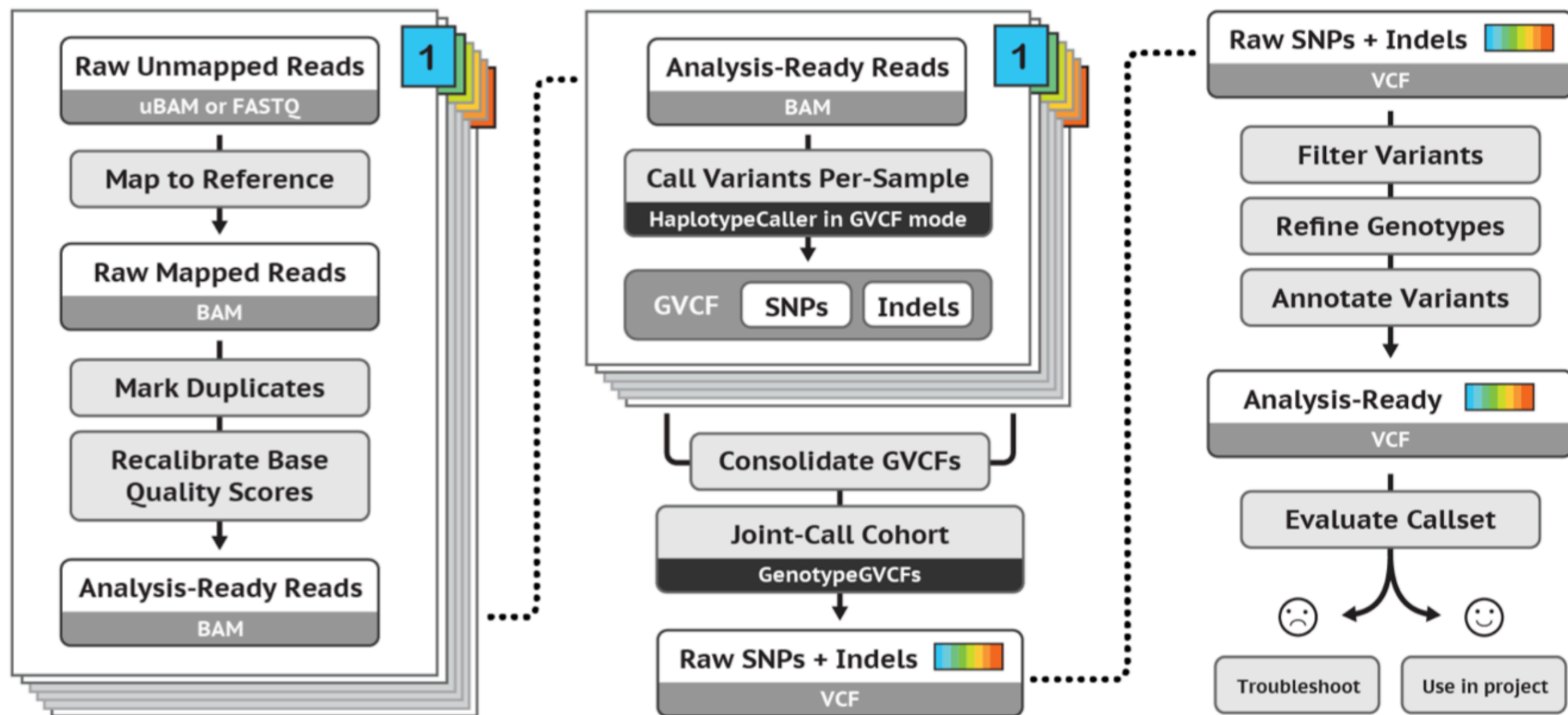
```
head -n 4 HypPue1_illumina_raw_F.fastq # display first 4 lines of file
```

```
@HWI-ST1293:199:HA9JHADXX:1:1101:1044:1603 1:N:0:CGATGT  
NCCCTGTTAAAGGATCATCTCTGACCTATCATTGTGGTGTAAATCACATTTAACACAATCACGATGTGCTTTACCTGCAGC  
ATCTTTACAGCAGGGCTGGGAGATATGACC AAAACAGTTATGATAATATGTTTATTTCTATTGAAAATCA  
+  
#1=DDFFHHHHHHJJJJJJJJJJJJJJJJJJJJJJGHJJJJJJIIJJJJHJJJJJJJJJJJJJJJJJJJJHHHHH  
FFFFFFEEEEEEEDDDDDDDDDD@DDDEDEDDBDDDDDDDCDCEDDEEDEEEDDEECDEDEDEEDDDDDDDDC
```

1. @ followed by sequence id and optional info (e.g. instrument/run id, barcode)
2. DNA sequence
3. +, sometimes followed by sequence id
4. base quality score (same length as sequence)

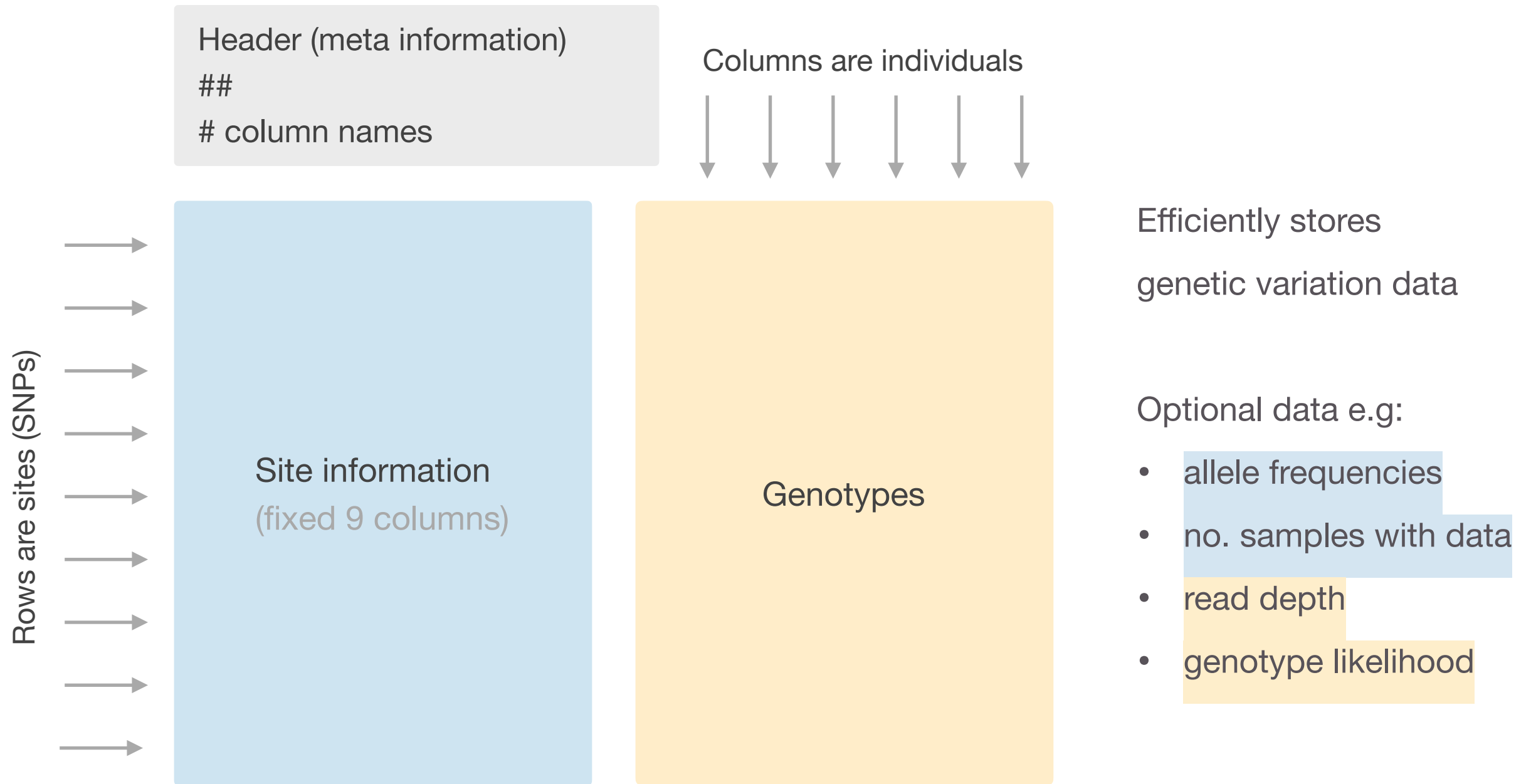


# Whole-genome genotyping workflow with GATK



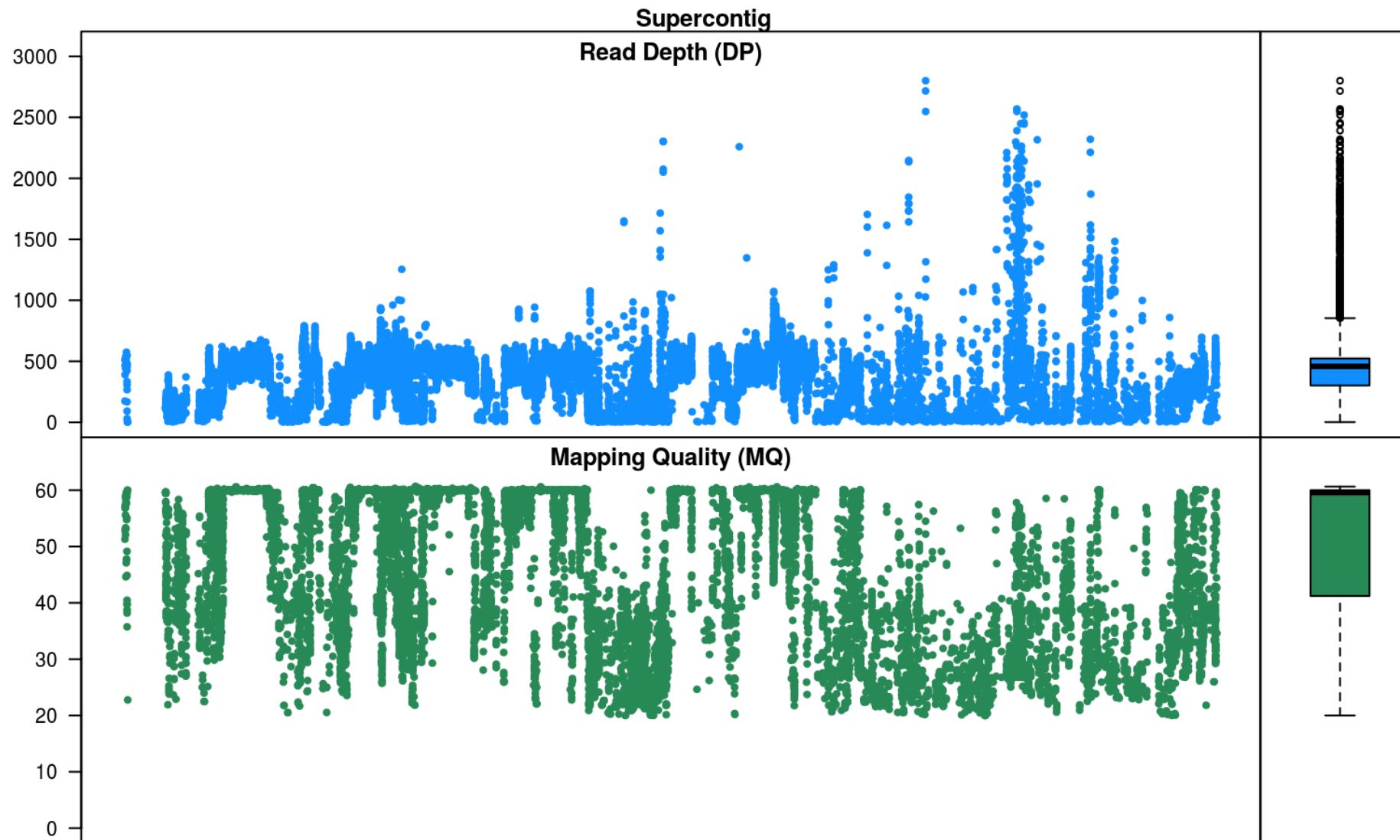
[gatk.broadinstitute.org](https://gatk.broadinstitute.org)

# Variant call format (VCF)



# Read depth and mapping quality

## Exercise 3



vcfR documentation

# Example dataset

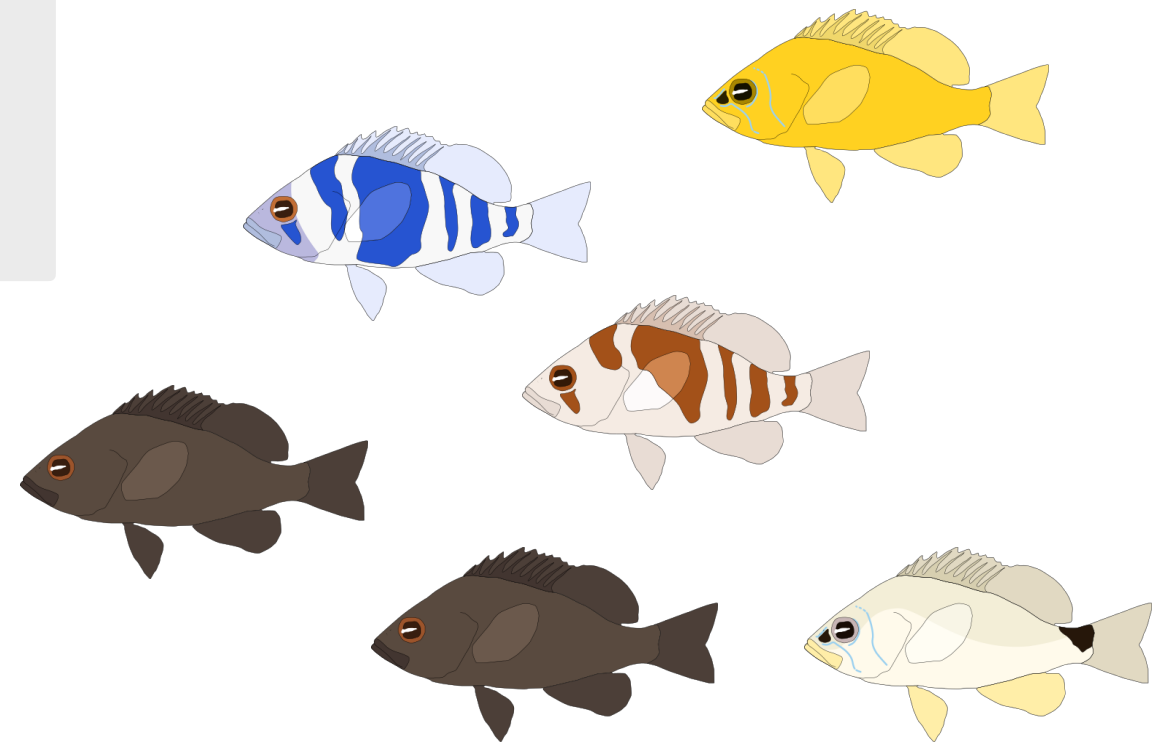
## Exercise 1–4

- 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**

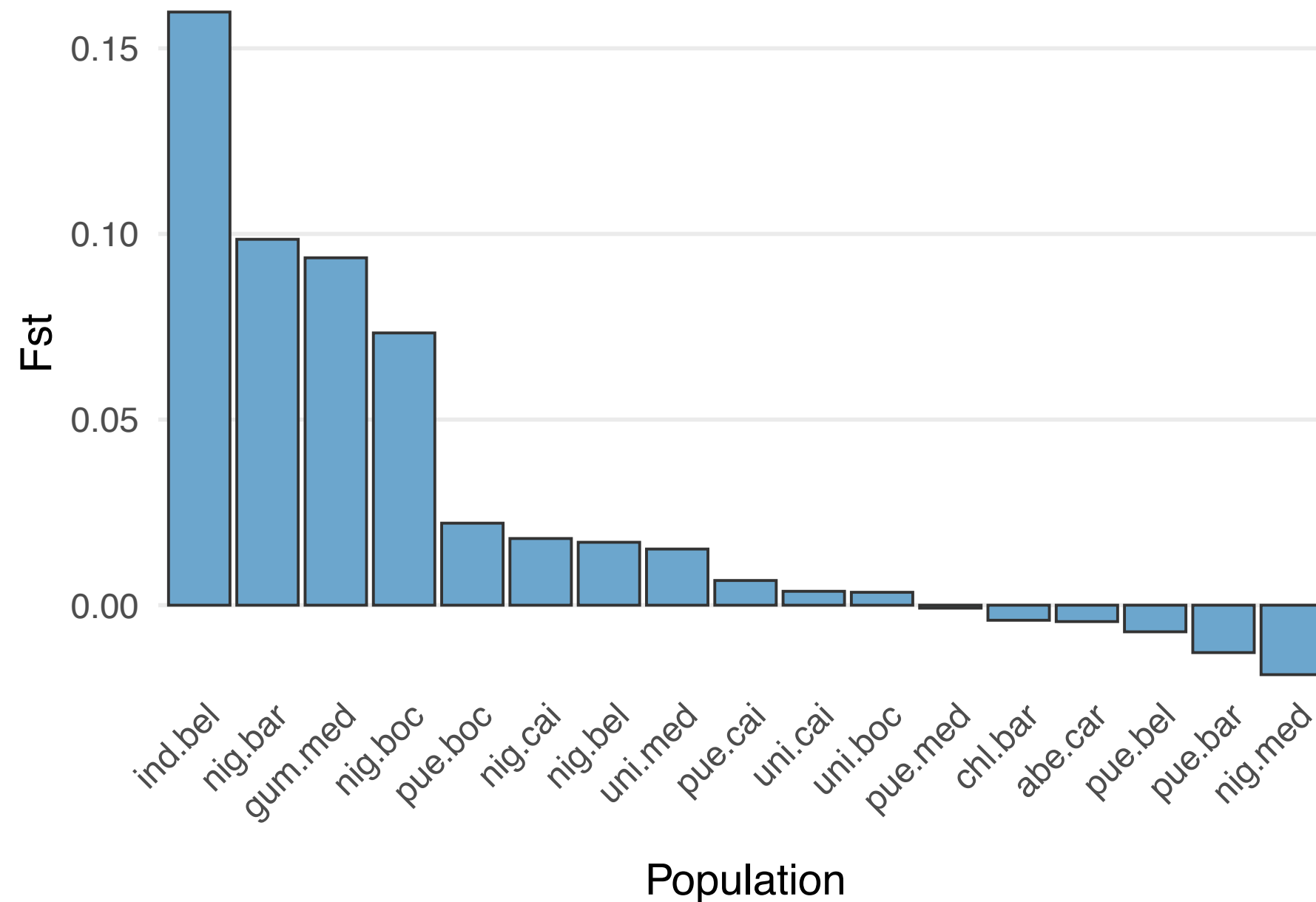
- Chromosome 12 only
- Subset to 36 samples from 6 populations



Illustrations by Kosmas Hensch

# Compare PCA with population-specific $F_{ST}$

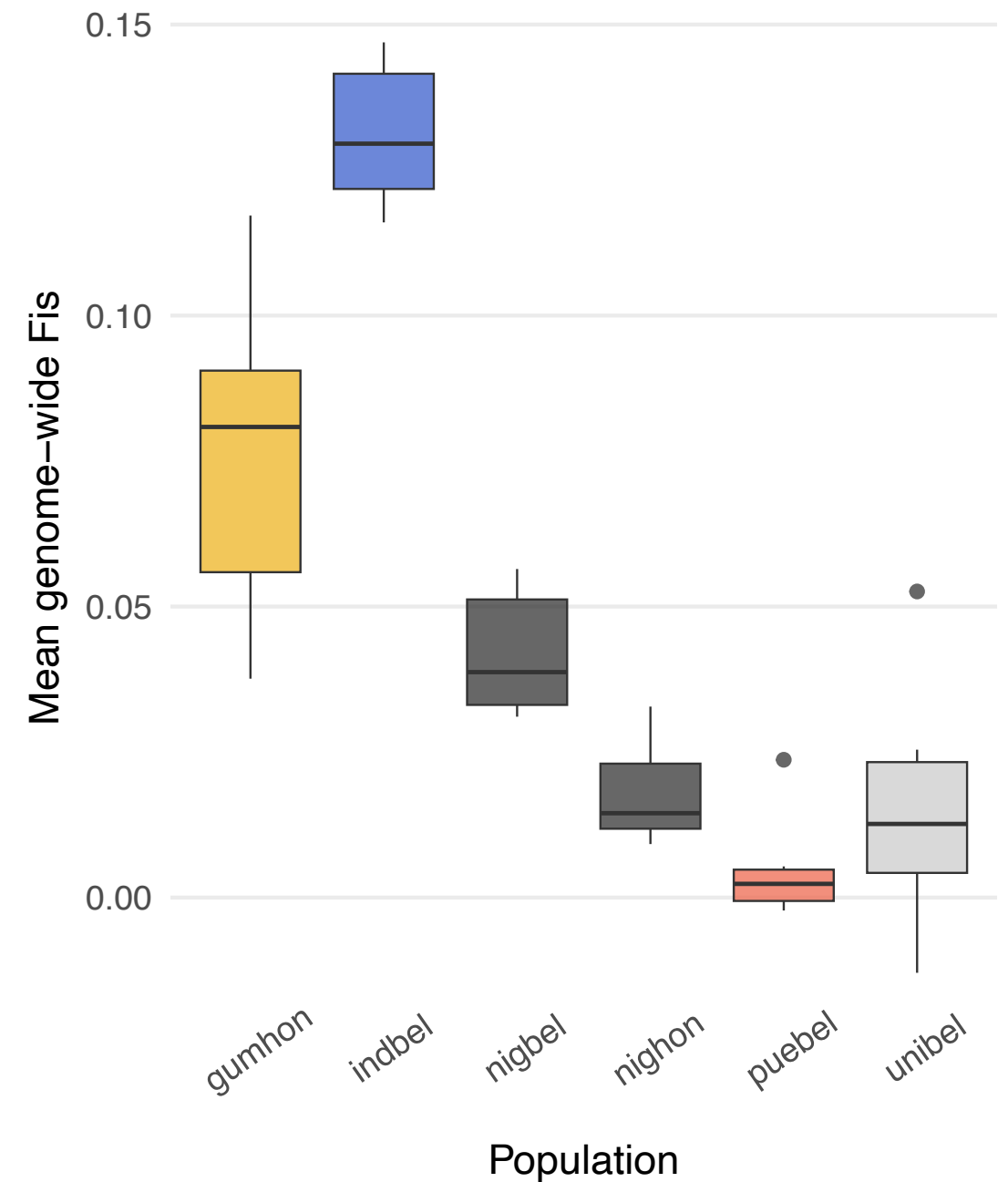
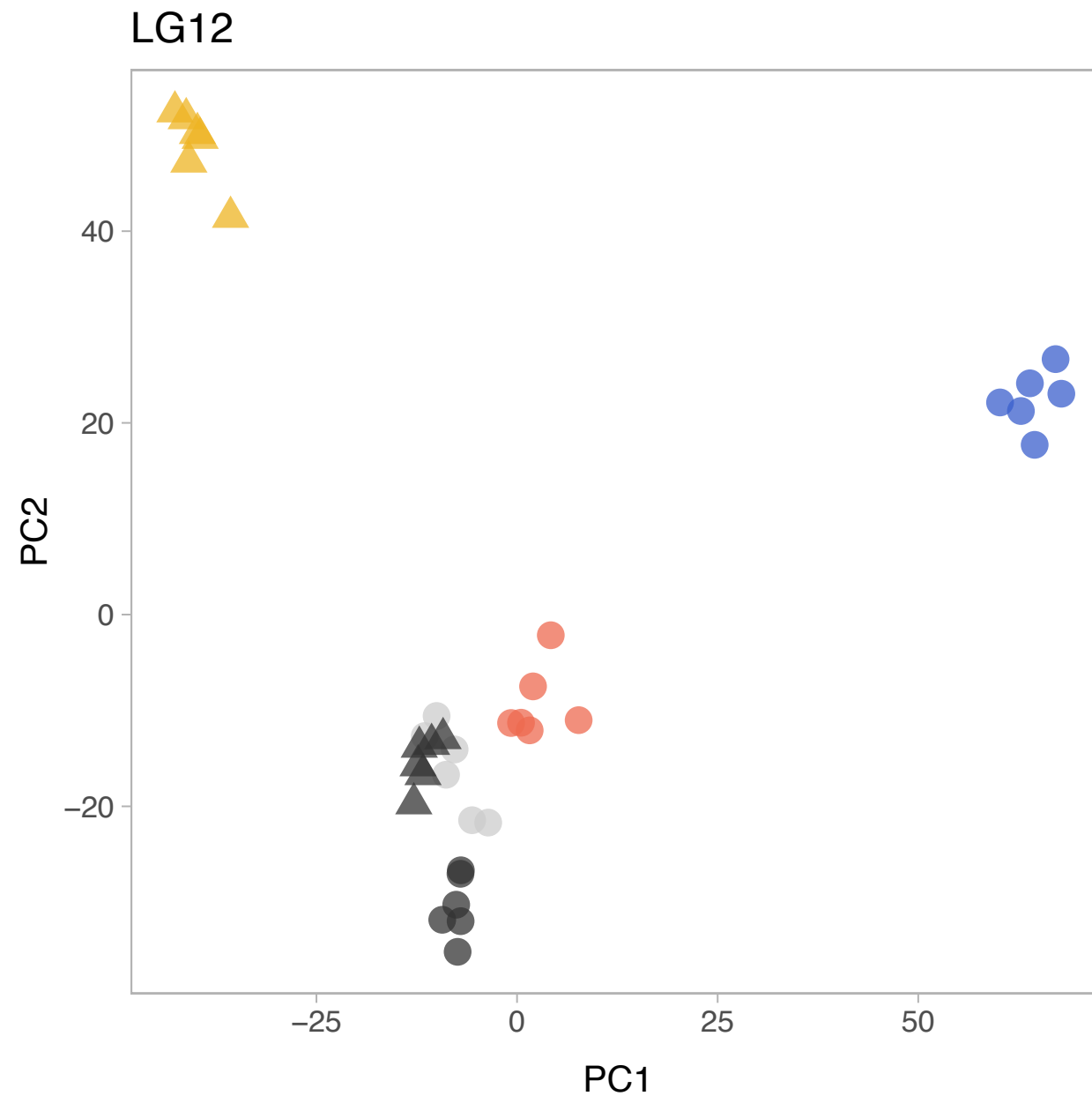
## Exercise 3



In PCA:  
boc = pan  
med ~ hon

# Compare PCA with population-specific $F_{ST}$

Exercise 3/4



# 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	<a href="#">Himmelfahrt break</a>	–
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	<a href="#">Student presentations</a>	–
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	–