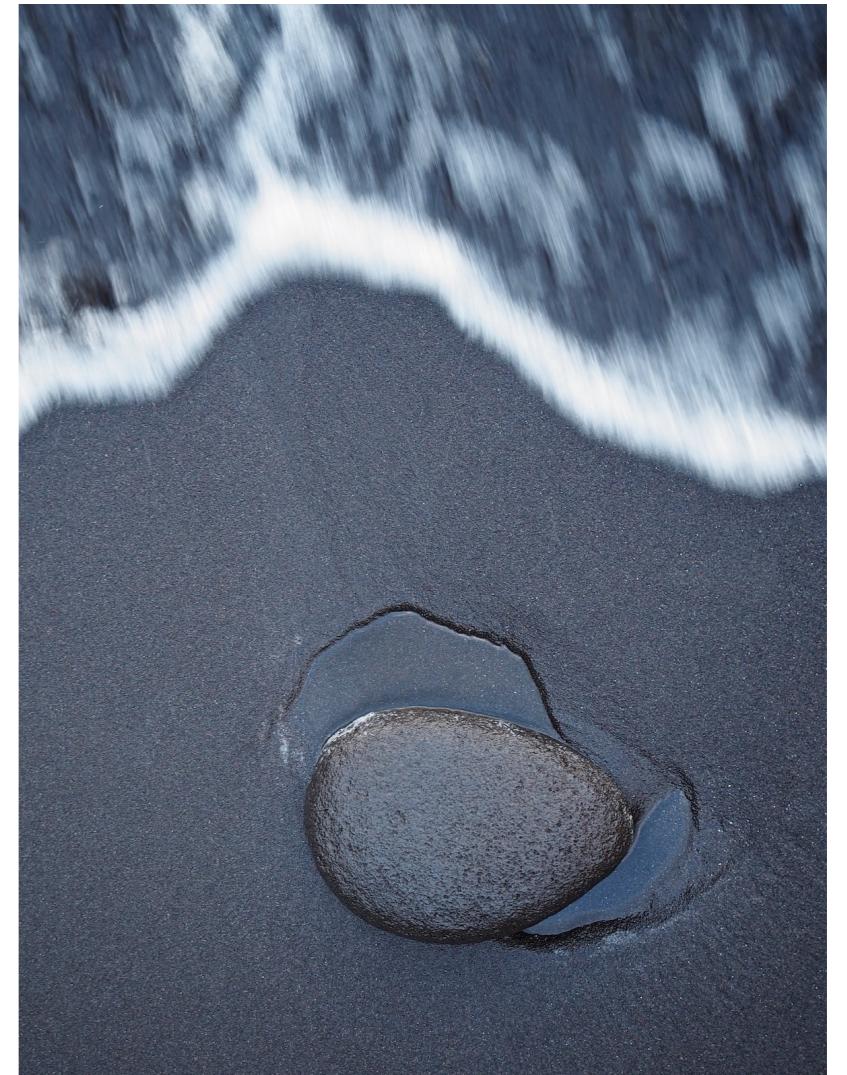


Exercises in Marine Ecological Genetics

06. Genome sequencing and assembly

- Become familiar with short and long sequencing reads
- Trim reads and assess read quality
- Calculate genome assembly metrics
- Learn to use a high performance computing cluster

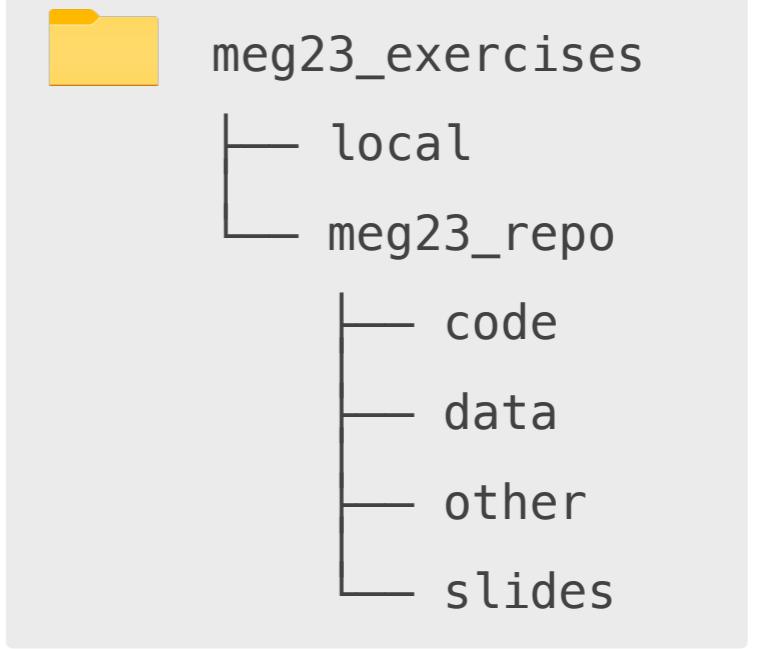
Martin Helmkampf



Download course materials using git

Go to project directory and check contents

```
cd dir          # e.g. Documents/meg23_exercises  
ls -l          # -l: long format
```



Download course materials using git

Go to project directory and check contents

```
cd dir          # e.g. Documents/meg23_exercises  
ls -l          # -l: long format
```

Update course repository

```
cd meg23_repo  
git pull
```



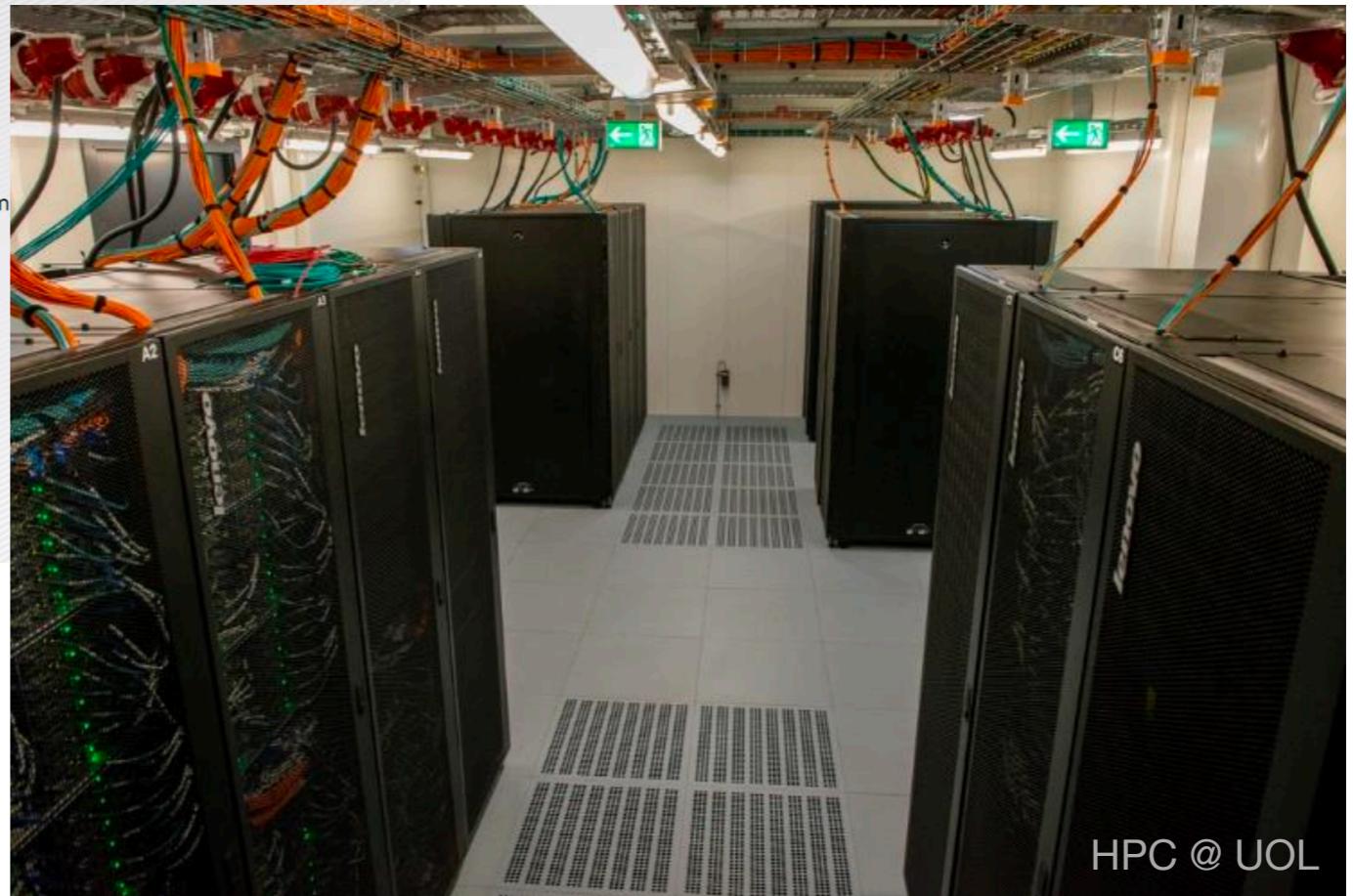
Please do not save over files in the course repository. Instead, save your own scripts to the local subdirectory (including copies of course scripts you would like to edit), e.g with

```
cp code/06_genseq.sh ../local/06_genseq_lc.sh      # cp [source] [destination]
```

High performance computing (HPC) at UOL



327 compute nodes
7640 cores (CPUs)
77 TB RAM total
271 TFlop/s



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
```

Get set up on the HPC cluster

Connect to login node

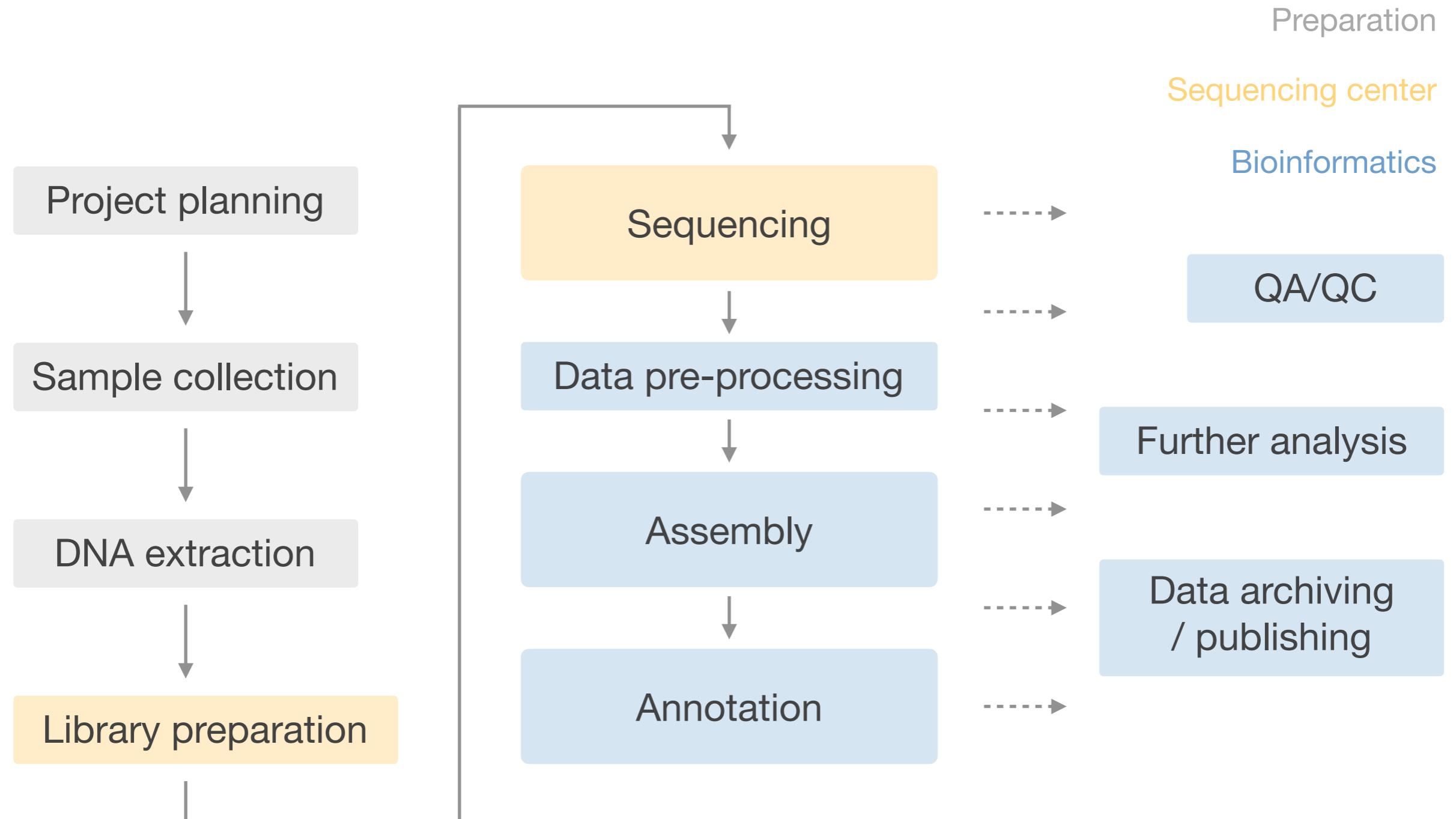
```
ssh <account>@carl.hpc.uni-oldenburg.de  
# Account ids and passwords can be found on StudIP in Files | course_accounts.csv
```

Download course materials to cluster account using git

```
git clone https://github.com/mhelmkampf/meg23\_repo.git
```

De novo genome sequencing workflow

Legend



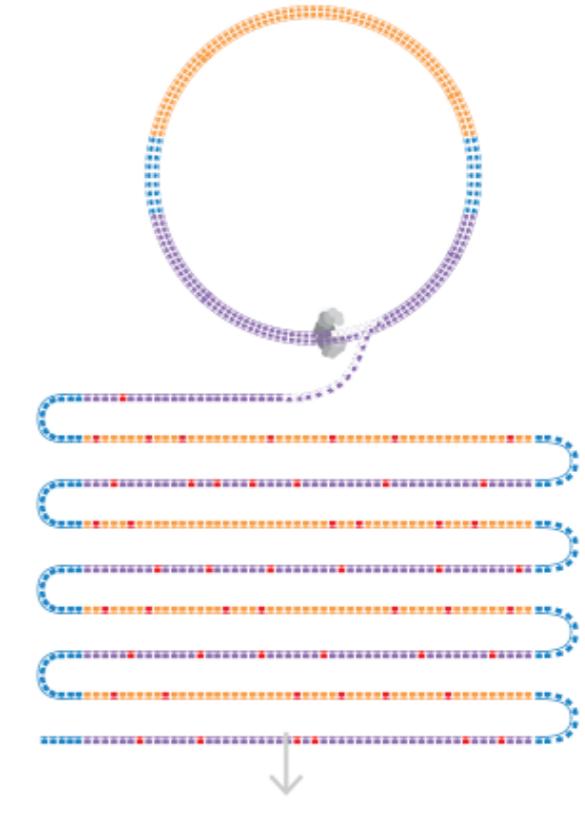
Genome assembly

- 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

- 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



Sequencing reads in FASTQ format

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

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)

Base quality

Phred quality score:

$$Q = -10 \log_{10} P$$

Common benchmark:

% bases with Q ≥ 30

Quality score	P incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

FASTQ encoding (Illumina 1.8+):

ASCII Symbol: !"#\$%&'()*)+,./0123456789:;=>?@ABCDEFGHIJ

Quality Score: 0.2.....10.....20.....30.....41

Assessing assembly quality

- Sequencing depth / coverage
- Assembly metrics: size distribution of contigs / scaffolds
- Average base accuracy (Q score)
- Percentage of assembly assigned to chromosomes
- Gene completeness
- Phasing information

Challenges

- Contamination
- Misassembled regions
- Presence of false duplications

Sequencing depth / coverage

- Average number of reads representing each position in the genome
- coverage or depth = **read count × read length / genome size**
- high coverage facilitates assembly, detection of sequencing errors
- Typical coverage: **50–100×** (or more) for *de novo* genome sequencing

10–30× for re-sequencing

```
Genome: CGTAATGGCATATCGCCTAGATTGAAACG  
Read 1: TAATGGCATATCGCCTAGAT  
Read 2: CATATGCCTAGATTGAAA  
Read 3: TATGCCTAGATTGAAACG  
Depth: 001111122333333333322222211
```

Assembly metrics

- Total size (compare to expected genome size)
- Number of contigs / scaffolds
- Largest scaffold
- **N50**: contig / scaffold size where 50% of assembly is found on contigs / scaffolds of equal or larger size (measure for sequence continuity)

Scaffolds: 530, 760, 1050, 610, 450, 800, 220, and 1200 kb

Reorder: 1200, 1050, 800, 760, 610, 530, 450, 220 kb

Sum/2: $5620/2 = 2810$

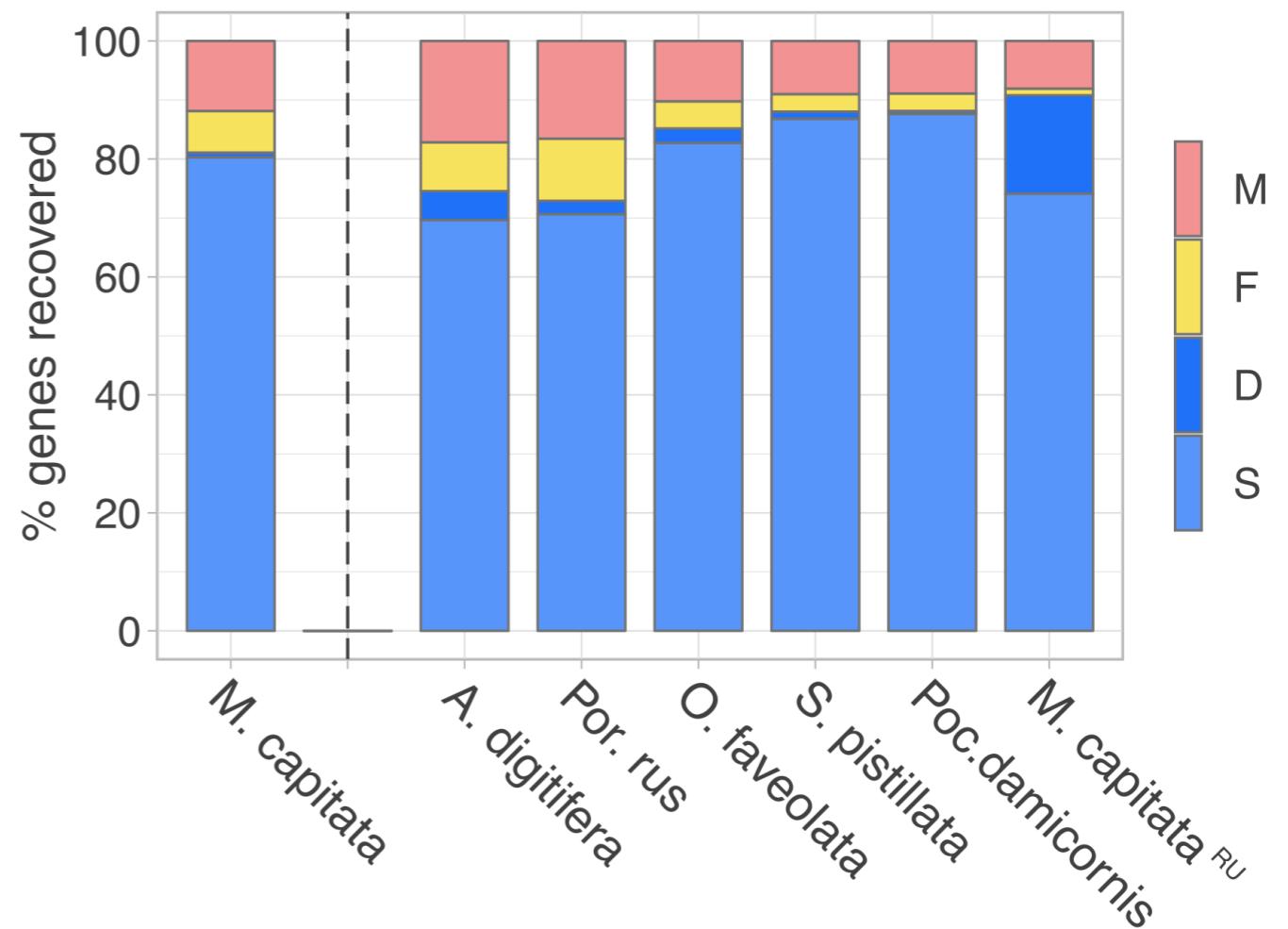
Add up until sum/2 is reached: $1200 + 1050 + 800 > 2810$

$N50 = 800$ kb

Gene completeness with BUSCO

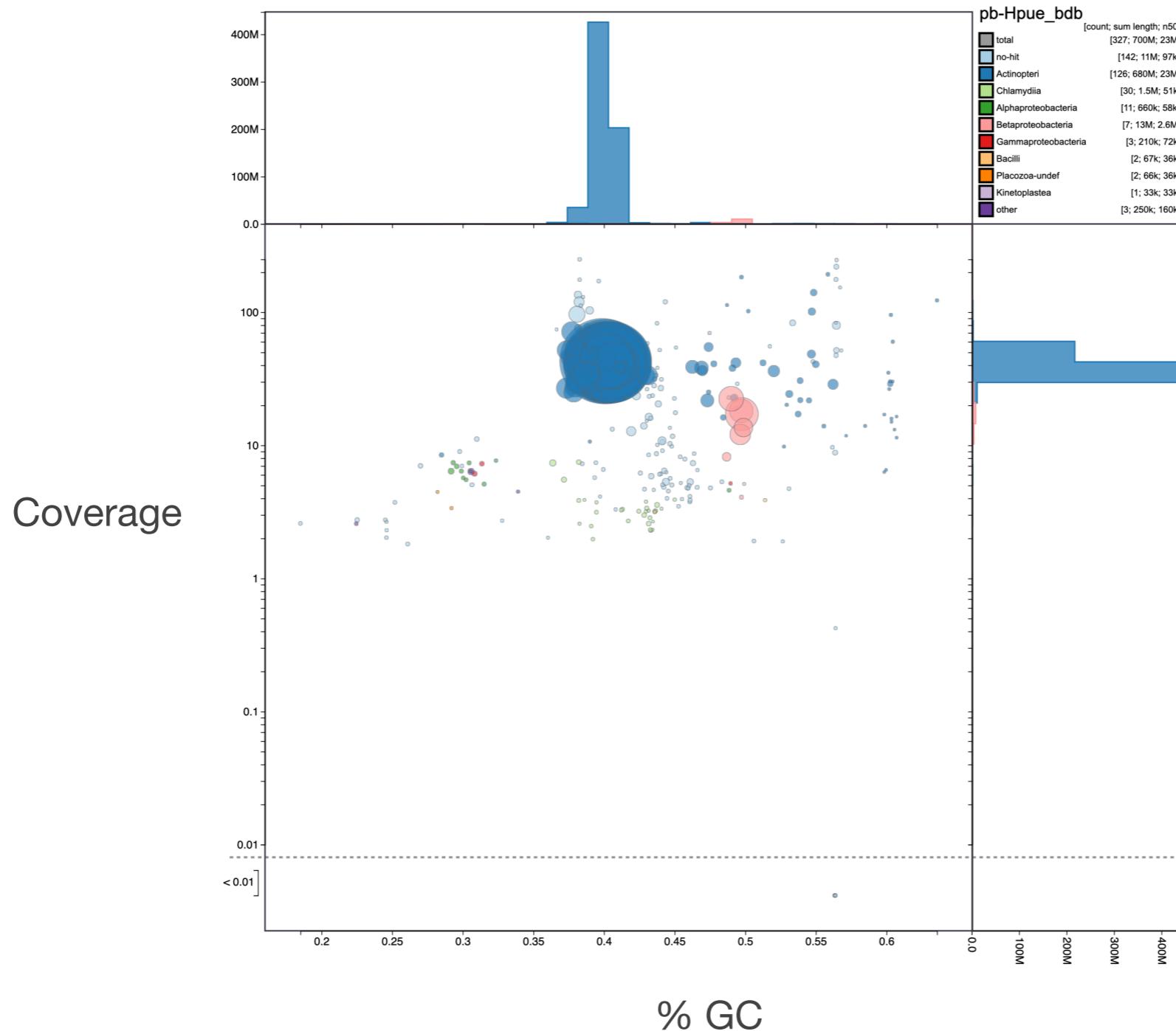
<https://busco.ezlab.org>

Quantifies assembly completeness based on presence of universal, highly conserved, single-copy genes (e.g. housekeeping genes)



Helmkampf et al. 2019 (Genome Biology and Evolution)

Contamination QC with BlobTools



Color:
Most similar
known taxon

HypPue2.1_pacbio_pctg.fas

Course outline

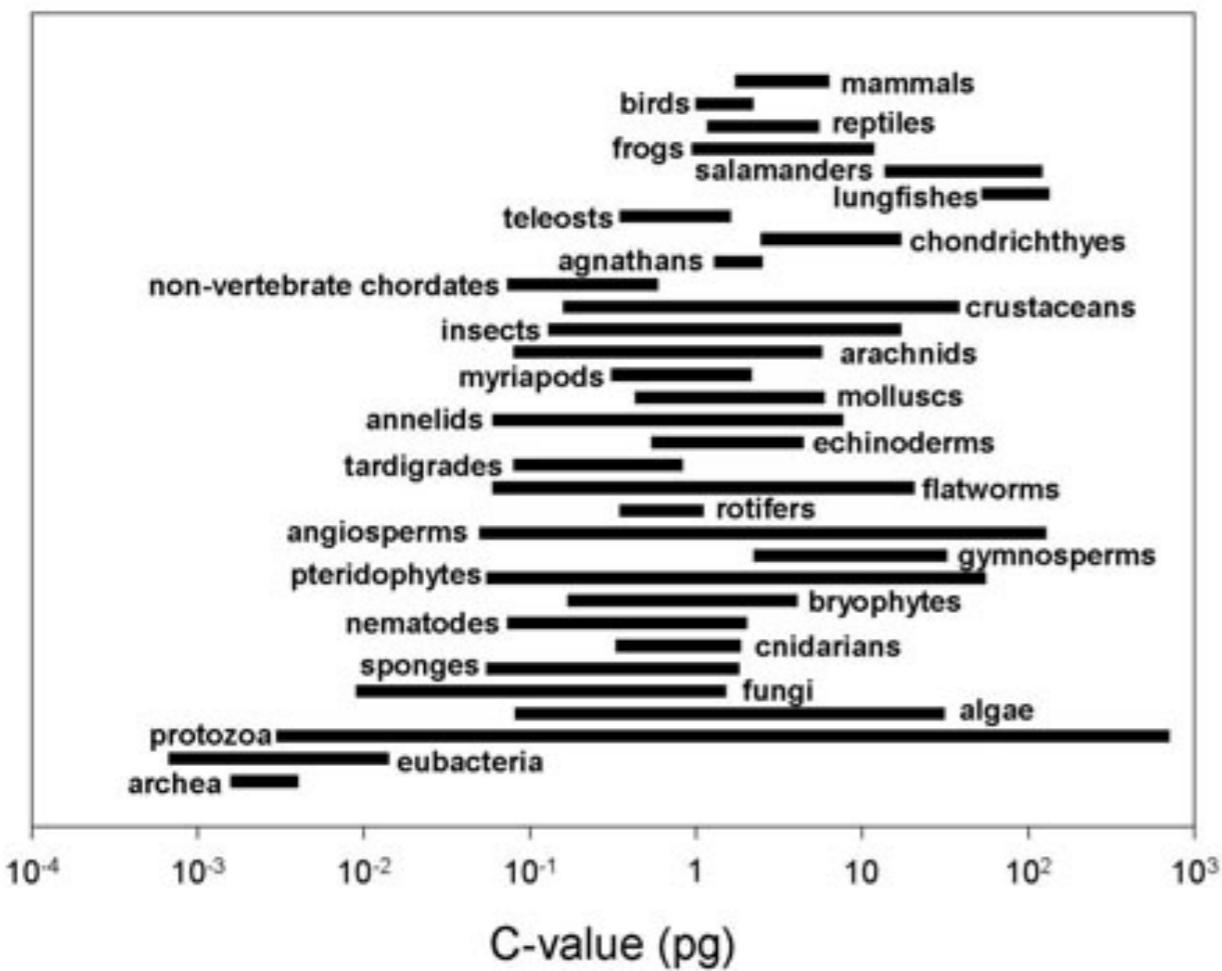
May be subject to change

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	-

Genome size

How large is the genome?

Search Animal Genome Size Database: www.genomesize.com



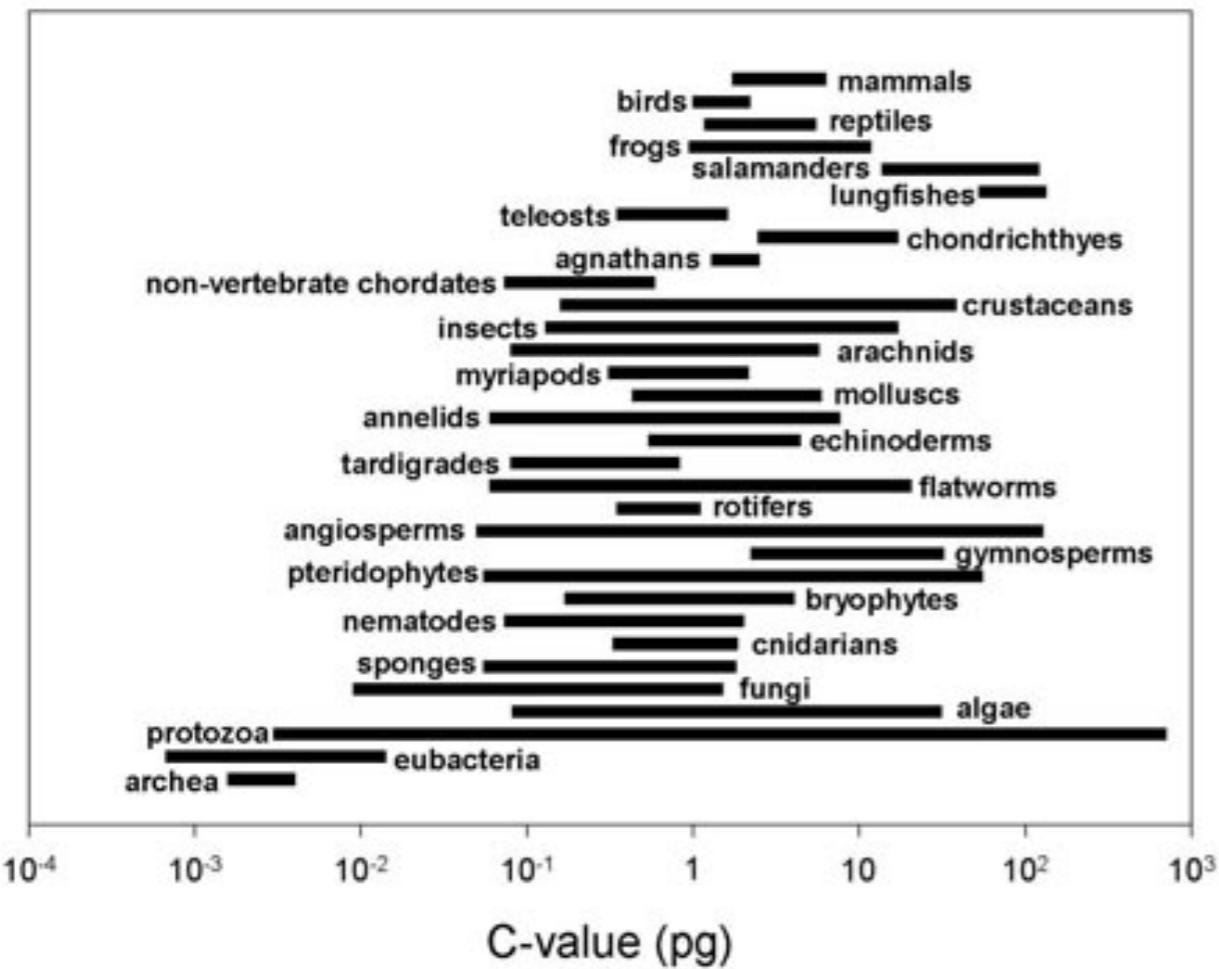
1 pg ~ 1 Gb

Gregory 2021, Animal Genome Size Database

Genome size

How large is the genome?

Search Animal Genome Size Database: www.genomesize.com



1 pg ~ 1 Gb

Genome size is often correlated
with repetitive DNA, which is
difficult to sequence and assemble

Gregory 2021, Animal Genome Size Database