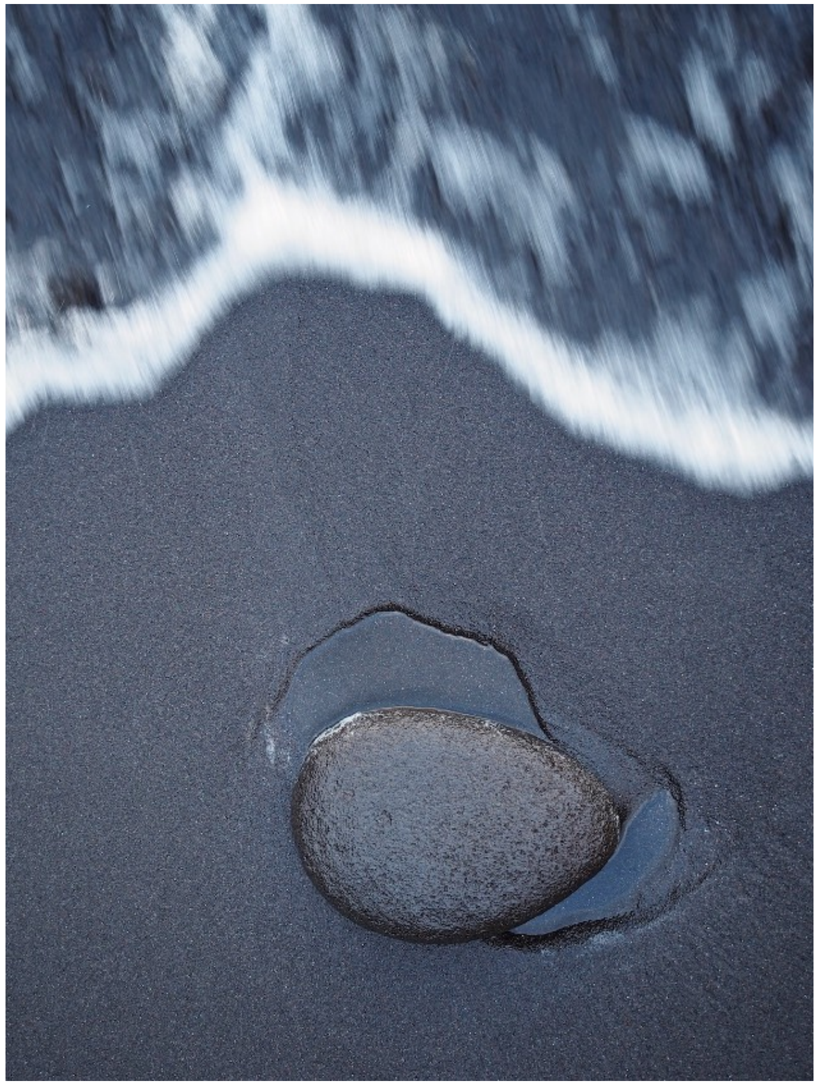
Exercises in Marine Ecological Genetics

04. Genome sequencing and assembly

- Become familiar with short and long read data
- Assess read quality before and after trimming
- Assemble PacBio HiFi reads
- Calculate genome assembly metrics



Martin Helmkampf

https://github.com/mhelmkampf/meg25

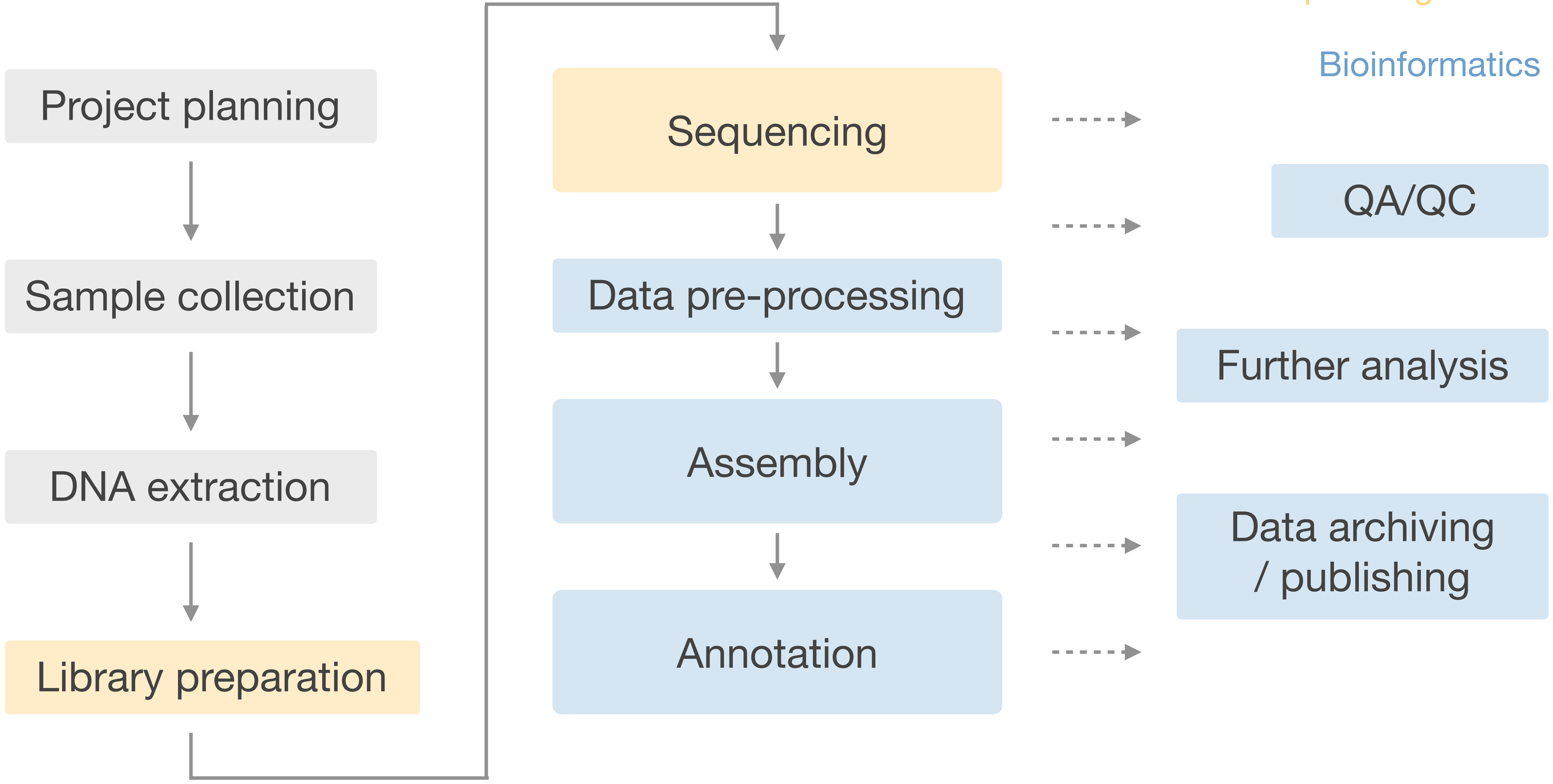
De novo genome sequencing workflow

Legend

Preparation

Sequencing center

Bioinformatics



Project planning

Sample collection

DNA extraction

Library preparation

Sequencing

Data pre-processing

Assembly

Annotation

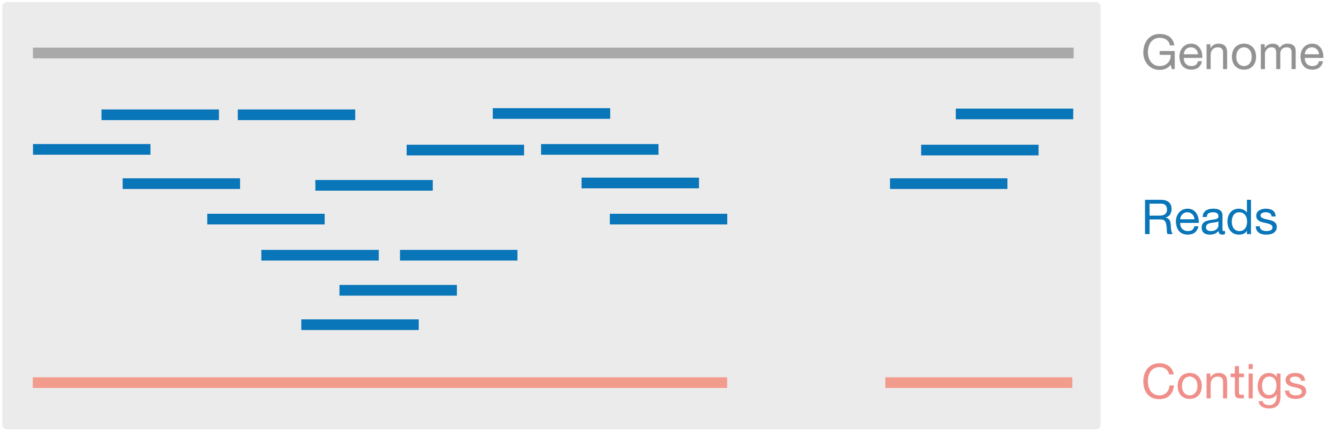
QA/QC

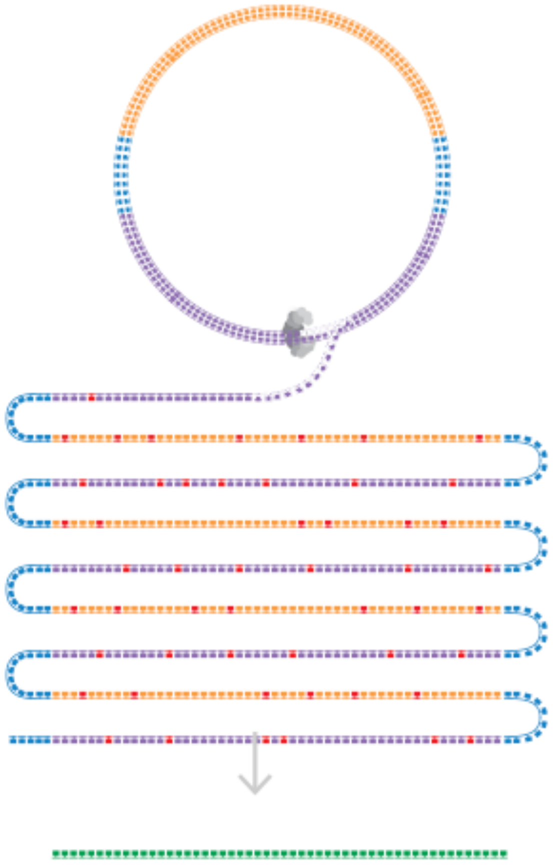
Further analysis

Data archiving / publishing

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




PacBio HiFi read

Sequencing technologies compared (2025)

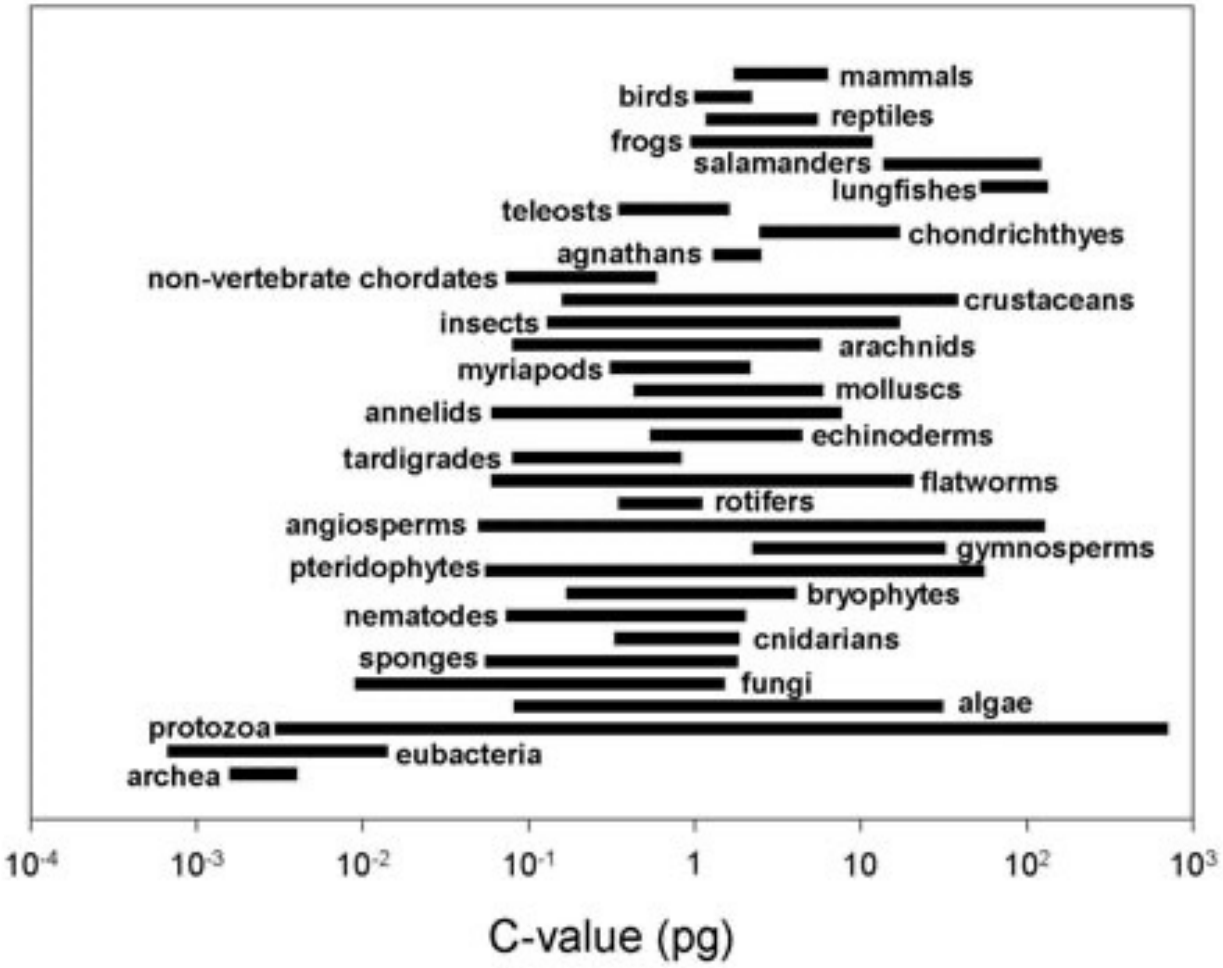
Technology	Typical read length	Accuracy	Gb per run	Cost per Gb	Devices
Illumina	100–300 bp	> 99.9 %	500–8000+	$1–5	NextSeq 2000, NovaSeq X
PacBio	15–25 kb	> 99.9 %	30–480	$10–20	Sequel II, Revio
Nanopore	10–50 kb	95–99.5 %	50–3000+	$5–100	MinION, PromethION

Legend: bp = base pairs, kb = kilo bases (1000 bp), Mb = Mega bases (mill. bp), Gb = Giga bases (bill. bp)

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

Base quality

Phred quality score:

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

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%

Common benchmark:

% bases with $Q \geq 30$

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: CGTAATGGCATATCGCCTAGATTCGAAACG
Read 1:   TAATGGCATATCGCCTAGAT
Read 2:         CATATCGCCTAGATTCGAAA
Read 3:           TATCGCCTAGATTCGAAACG
Depth:  001111112233333333333322222211
```

Assembly metrics

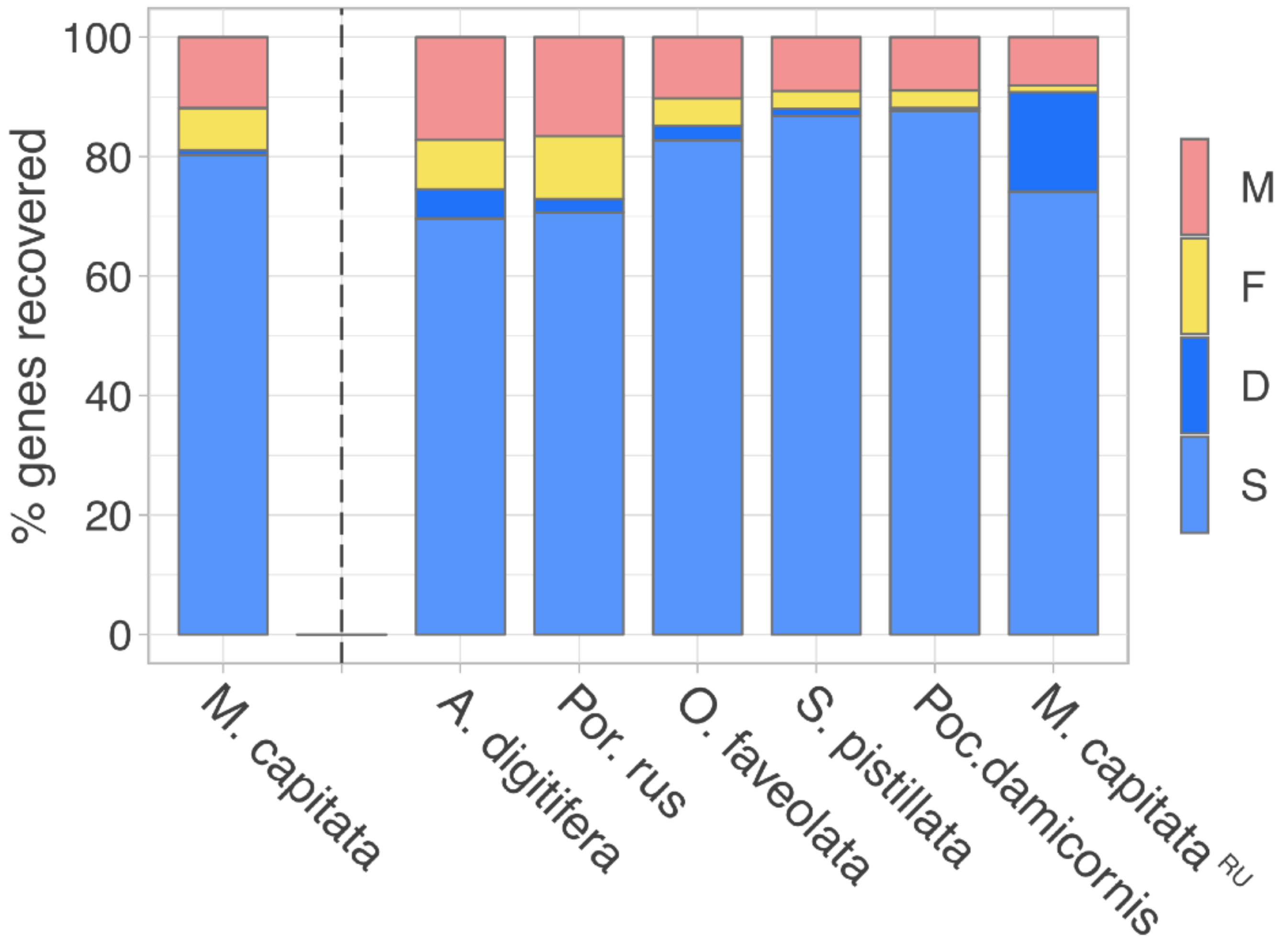
- Total size (compare to expected genome size)
- Number of contigs / scaffolds
- Largest scaffold
- N50: contig / scaffold size where 50% of assembly is found on contigs / scaffolds of equal or larger size (measure for sequence continuity)

```
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

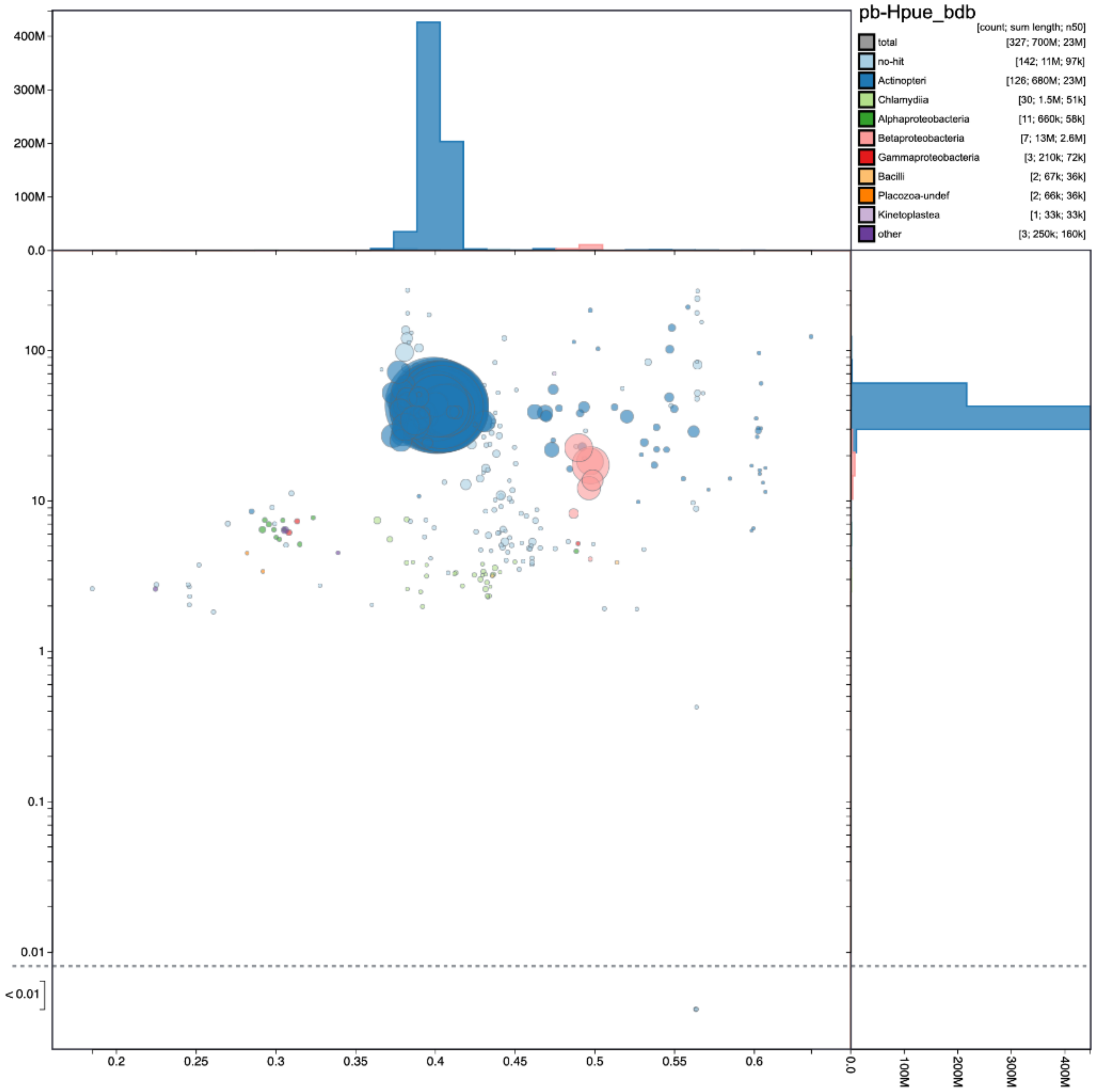
Quantifies assembly completeness based on presence of universal, highly conserved, single-copy genes (e.g. housekeeping genes)



Helmkampf et al. 2019 (Genome Biology and Evolution)

Contamination QC with BlobTools

Coverage



% GC

Color:

Most similar

known taxon

HypPue2.1_pacbio_pctg.fas

Genome size

Class	Date	Topics	Script
01	Apr 11	Introduction, setup	01_intro.R
–	Apr 18	Good Friday	
02	Apr 25	Hardy-Weinberg equilibrium / N_e	
03	May 02	Population structure and gene flow	
04	May 09	Genome assembly and metrics	
05	May 16	Population genomics and SNPs	
06	May 23	Linkage disequilibrium and genetic diversity	
–	May 30	Himmelfahrt break	
07	Jun 06	Population structure II	
–	Jun 13	Selection	
08	Jun 20	Student presentations – no exercises	
09	Jun 27	DNA barcoding	
10	Jul 04	Metabarcoding / eDNA	
11	Jul 11	Introduction to phylogenetics	