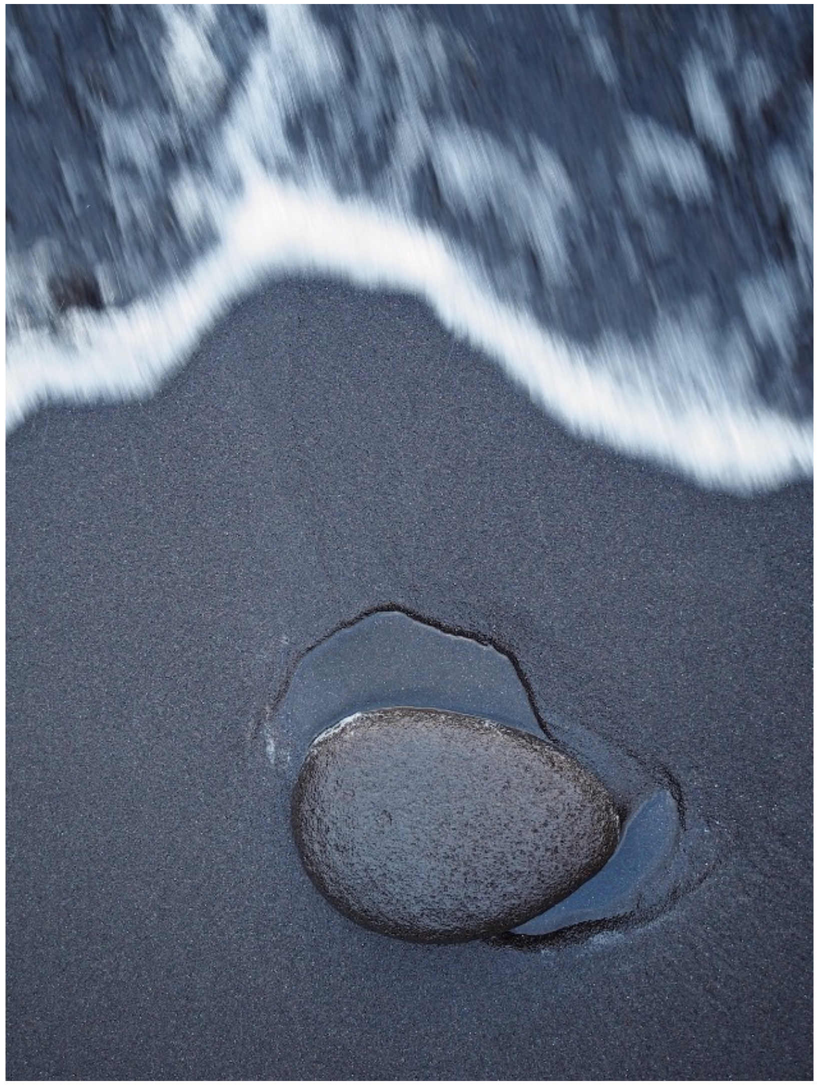
Exercises in Marine Ecological Genetics

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

Course outline

Updated, but may be subject to change

Class	Date	Topics	Script
01	Apr 11	Introduction, setup	01_intro.R
–	Apr 18	Good Friday	
02	Apr 25	Hardy-Weinberg equilibrium	02_hwe.R
03	May 02	Population structure I	03_popst.R
04	May 09	Genome sequencing and assembly	04_asm.sh
05	May 16	Variant calling and SNPs	
06	May 23	Population genomics 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	

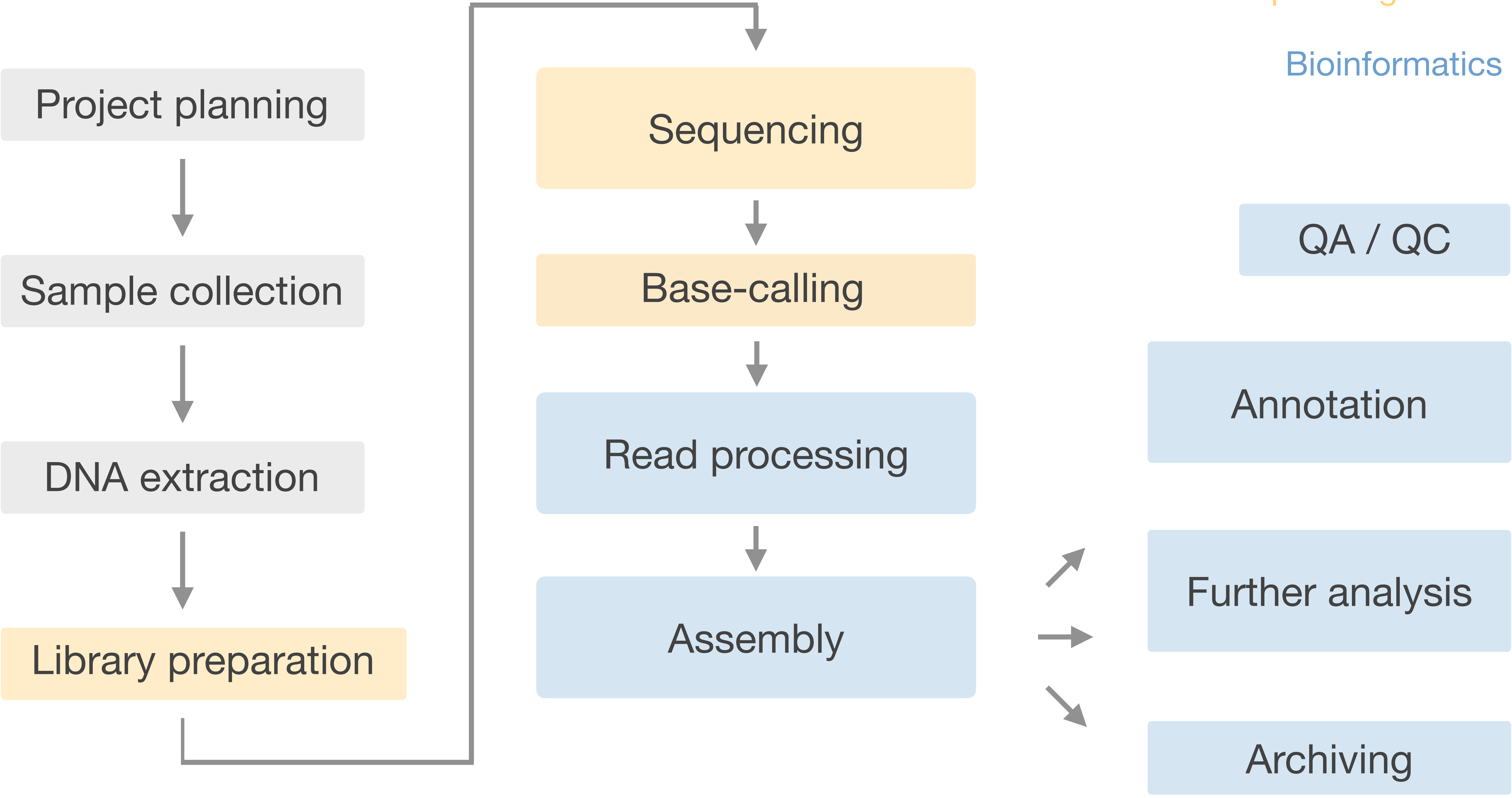
De novo genome sequencing workflow

Legend

Preparation

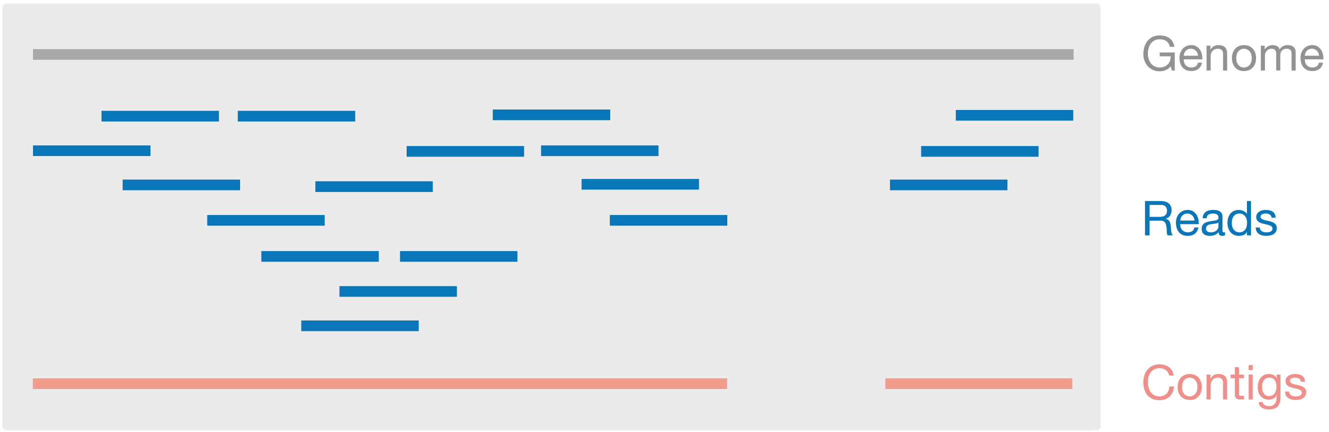
Sequencing center

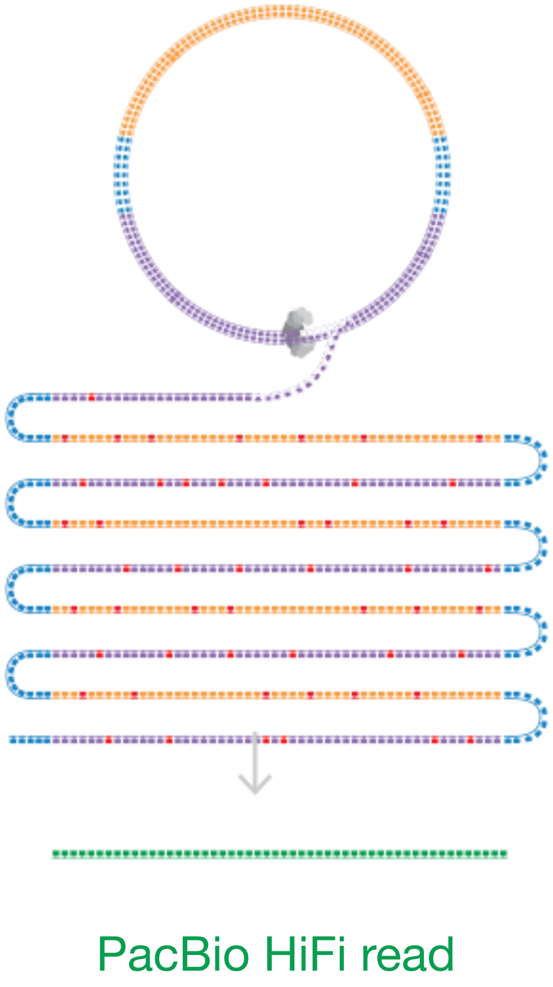
Bioinformatics



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





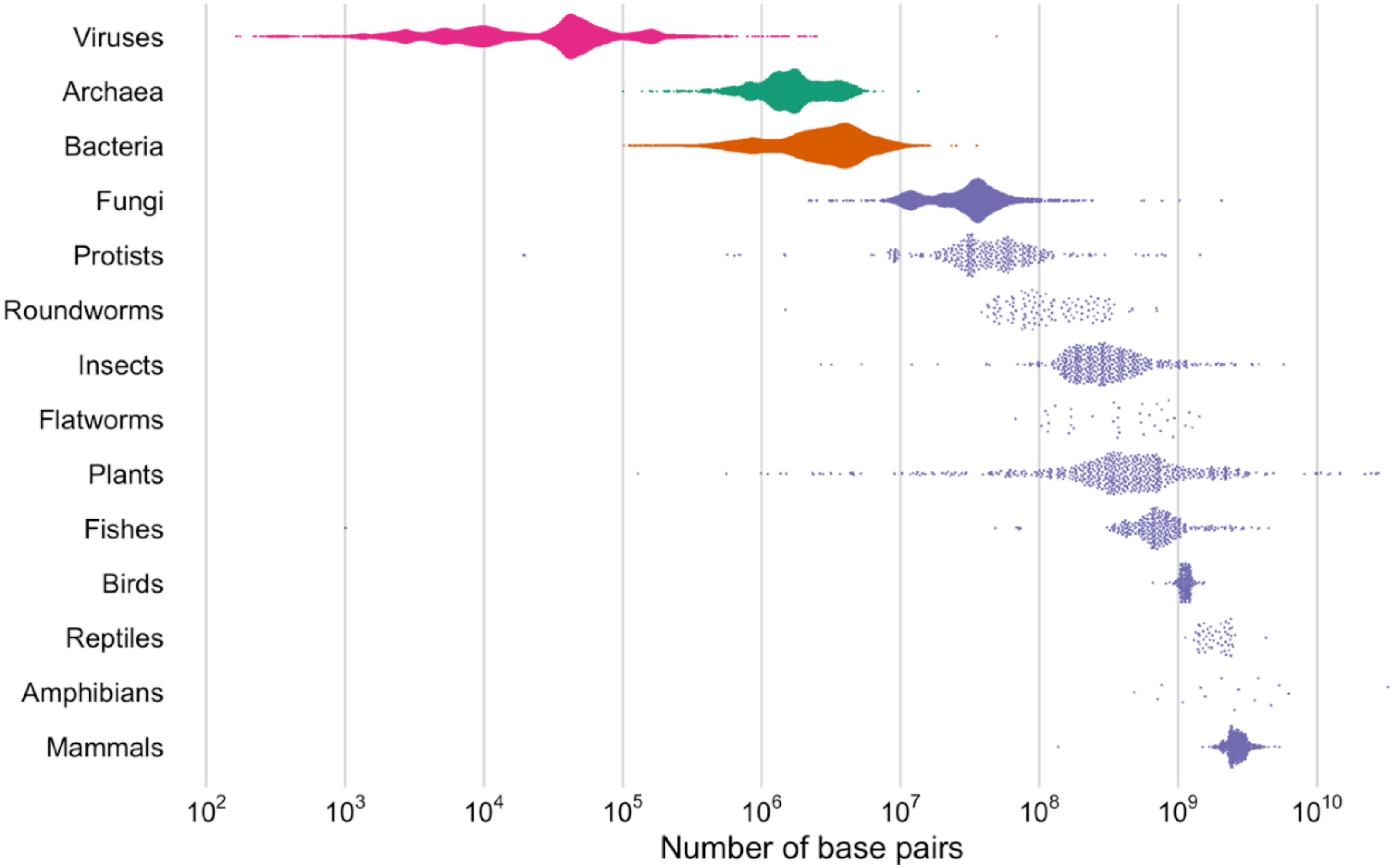
Sequencing technologies compared

Technology	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

As of 2025; read length typical not optimal

bp = base pairs, kb = kilo bases (1000 bp), Mb = Mega bases (million bp), Gb = Giga bases (billion bp)

Genome size



- Nine orders of magnitude
- Genome size correlated with repetitive DNA, which is hard to sequence and assemble

Based on 50,000 organisms with genome information in NCBI

See also www.genomesize.com

Tom E. White, Datavision 2020

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
NCCCTGTTAAAGGATCATCTCTGACCTATCATTGTGGTGTAATCACATTTAACACAATCACGATGTGCTTTACCTGCAGC
ATCTTTACAGCAGGGCTGGGAGATATGACCAAAAACAGTTATGATAATATGTTTATTTCTATTGAAAATCA
+
#1=DDFFFHHHHHHJJJJJJJJJJJJJJJJJJJJJJJGHIJJJJJIIIJJJJHJJJJJJJJJJJJIJIJJJJJJJHHHHH
FFFFFFEEEEEEDDDDDDDDD@DDDEDEDDDBDDDDDDCDCEDDEEDDEEEDEECDDEDEDEEDDDDDDDC
```

```
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:

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

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 ≥ 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

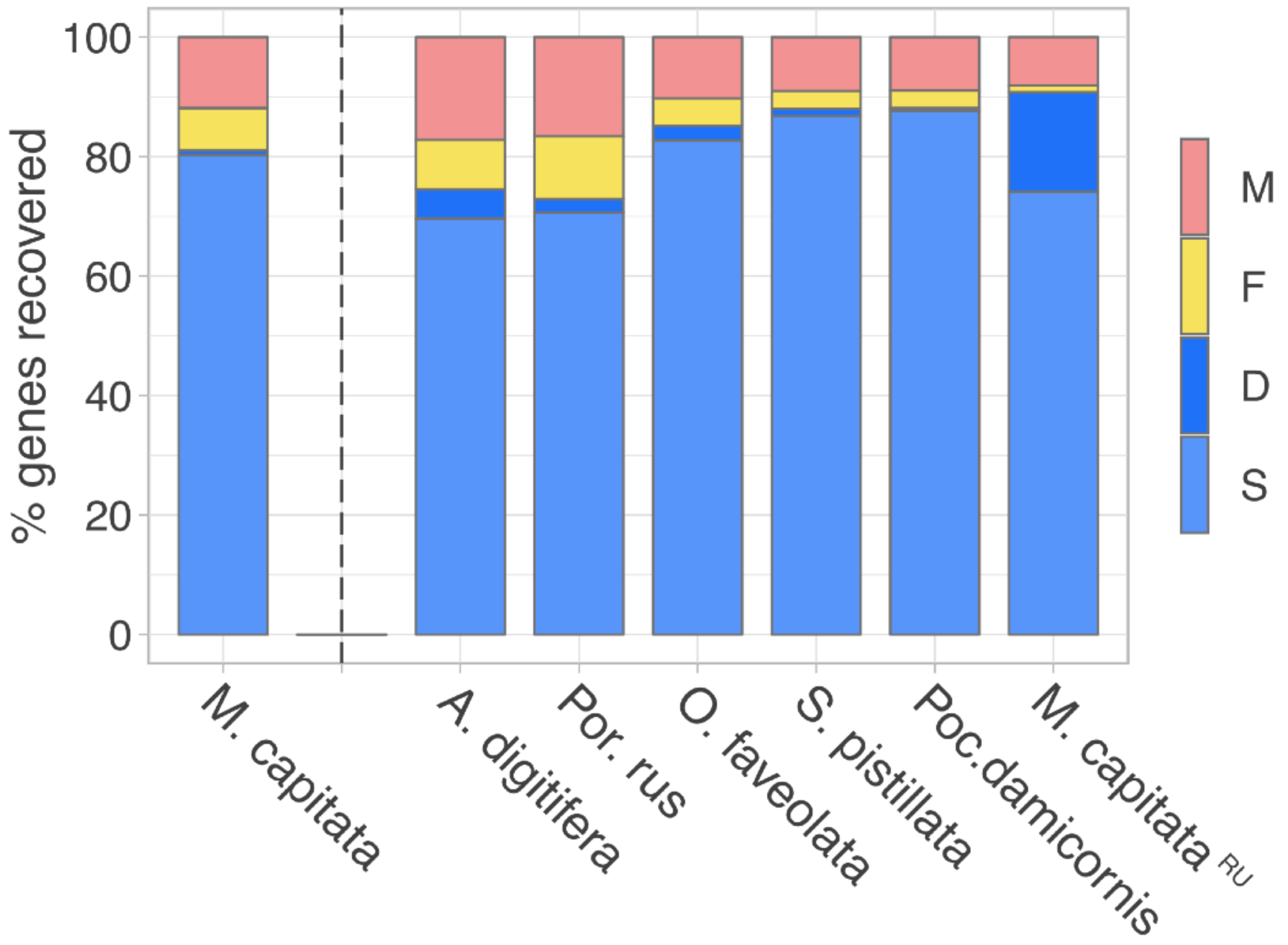
- Total size (compare to expected genome size)
- Number of contigs / scaffolds
- Largest scaffold
- N50: contig / scaffold size where 50% of assembly is found on contigs / scaffolds of equal or larger size (measure for fragmentation / sequence contiguity)

```
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

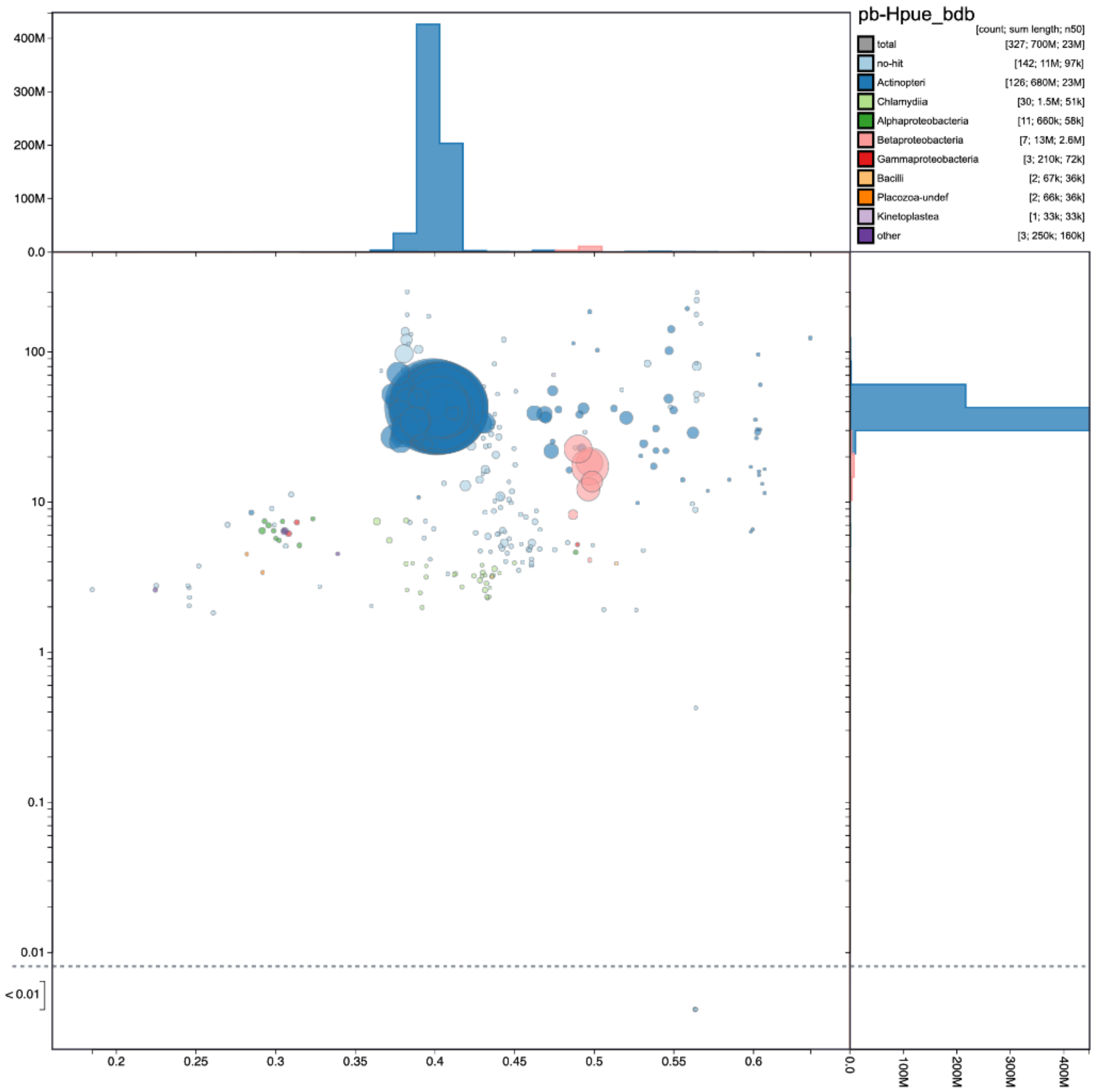
https://busco.ezlab.org

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

Coverage

% GC

HypPue2.1_pacbio_pctg.fas

Genome sequencing and assembly

Recap

```
fastqc -o <output_dir> <read_file>.fastq                   # Assess read quality

hifiasm -o <output_dir> --primary <read_file>.fastq      # Assemble PacBio reads

assembly_stats <assembly_file>.fas    # Calculate assembly metrics (Python tool)
```

- Short- and long-read sequencing technologies offer different but complementary strengths
- Assembly quality may be evaluated using metrics like coverage, contiguity (e.g. N50) and completeness (e.g. BUSCO)
- Assemblies are drafts – often fragmented and with errors – but serve as the foundation for further analyses, including in population and conservation genomics