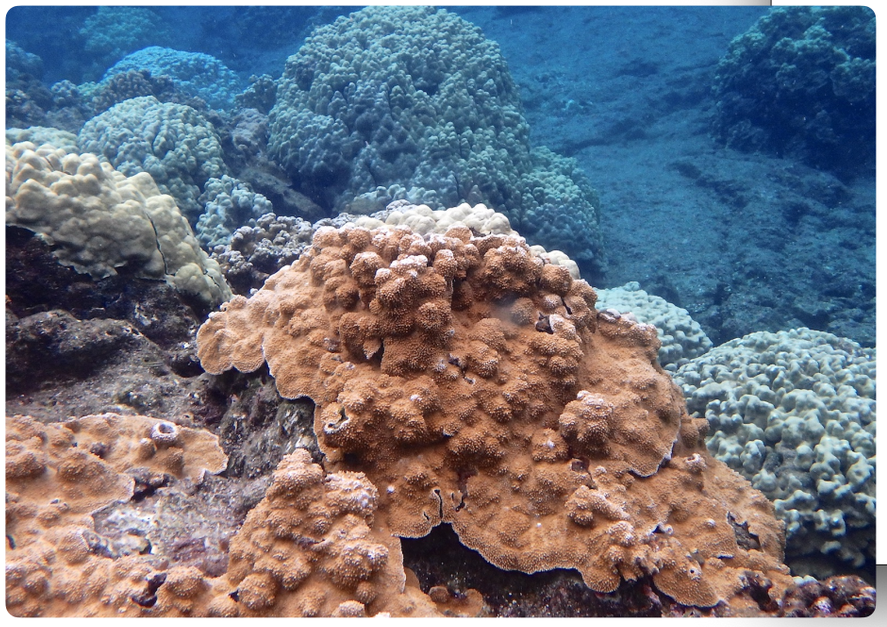
Genomics

An introduction to genome sequencing, data processing, and analysis

#DatAlumni course series @ ZMT, July 6/7, 2021

Martin Helmkampf



Workshop plan

Time	Block	Topics
Jul 06, 14:00–14:30	1. Background and introduction	– Sequencing techniques and applications – Workflow overview – Planning a genome project
Jul 06, 14:30–16:00	2. Data pre-processing and QA/QC	– Introduction to bioinformatics (exercise) – Raw sequencing data, trimming & filtering – Quality assurance / control
Jul 07, 14:00–15:00	3. Assembly and annotation	– Genome assembly (exercise) – Genome re-sequencing – Genome annotation
Jul 06, 15:00–16:00	4. Data management	– Documentation – Data organization – Publishing / sharing genome data

Objective: Convey basic knowledge of the steps required to sequence a genome, from project planning to data publication

Exercises

Required software

- git
- bash

Windows

- install git from here: https://git-scm.com/download (use all default settings)

  follow instructions at: https://www.computerhope.com/issues/ch001927.htm
- bash comes with git, type “git bash” in Start menu to launch

macOS

- both should come preinstalled
- open Terminal (/Applications/Utilities/Terminal.app), type “git --version” to check
- If you receive an error message, install from here: https://sourceforge.net/projects/git-osx-installer/

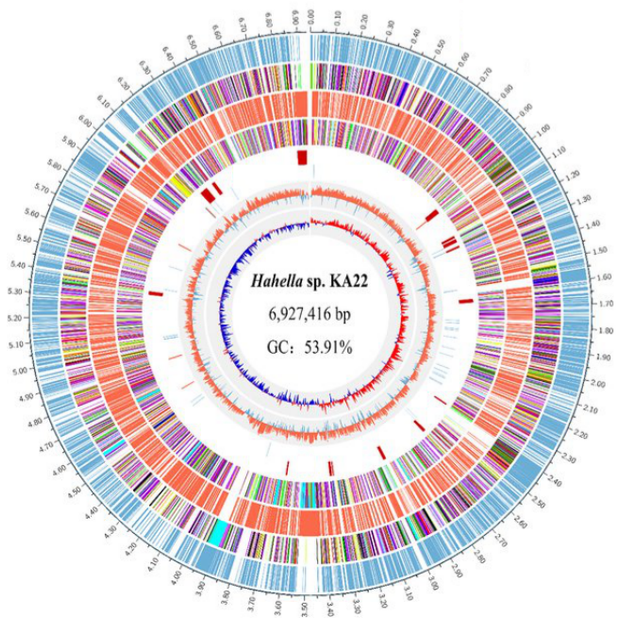
1. Background and introduction

- What is genomics?
- Sequencing technologies and applications
- Basic workflow
- Project planning

What is genomics and what can we do with it?

Understanding genomes

- Structure
- Function (e.g. gene prediction and interactions)
- Evolution (> comparative genomics)

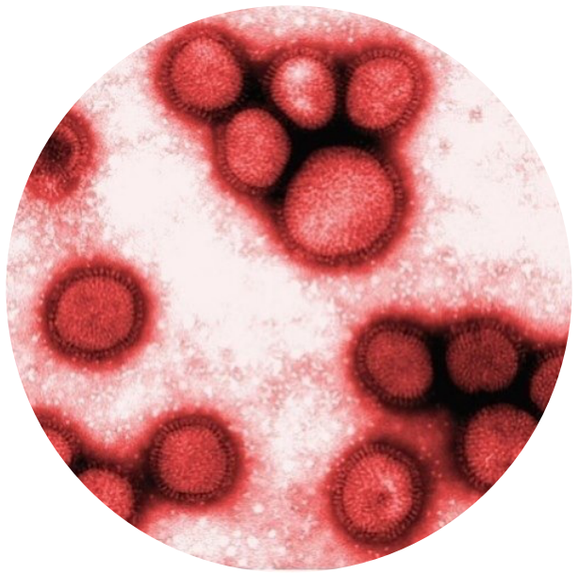


adapted from Feng et al. 2019, Marine Genomics

What is genomics and what can we do with it?

Studying phylogenetic relationships

- Reconstruct evolutionary history of organisms or genes
- Identify unknown organisms (e.g. new pathogens, metagenomics)
- How do species evolve?



What is genomics and what can we do with it?

Understanding the genetic basis of phenotypic traits

(e.g. GWAS, Genome-wide association studies)

- Diseases and risk factors
- Yield and resistance in domestic plants and animals

What is genomics and what can we do with it?

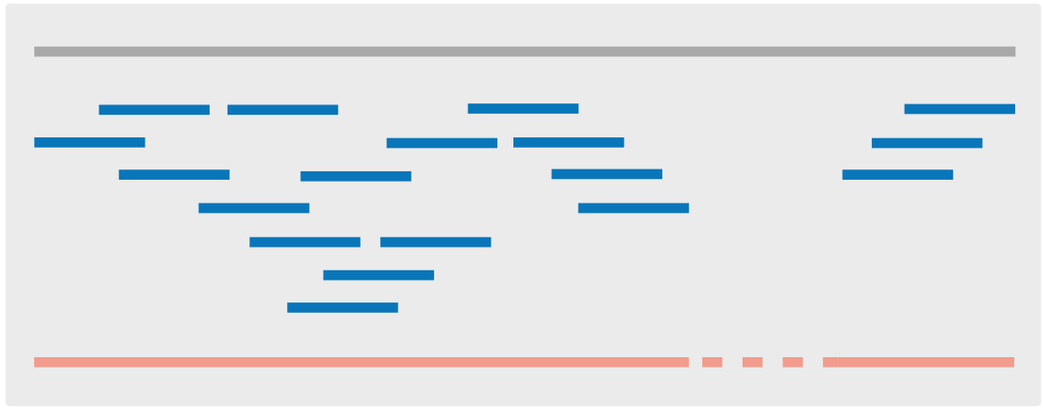
Population and conservation genomics

- Estimate population history and size
- Assess genetic diversity / identify adaptive loci
- Detect hybrids / effects of hybridization

Genome sequencing basics

Whole genome sequencing (WGS) strategies

- *de novo*
- re-sequencing (mapping)

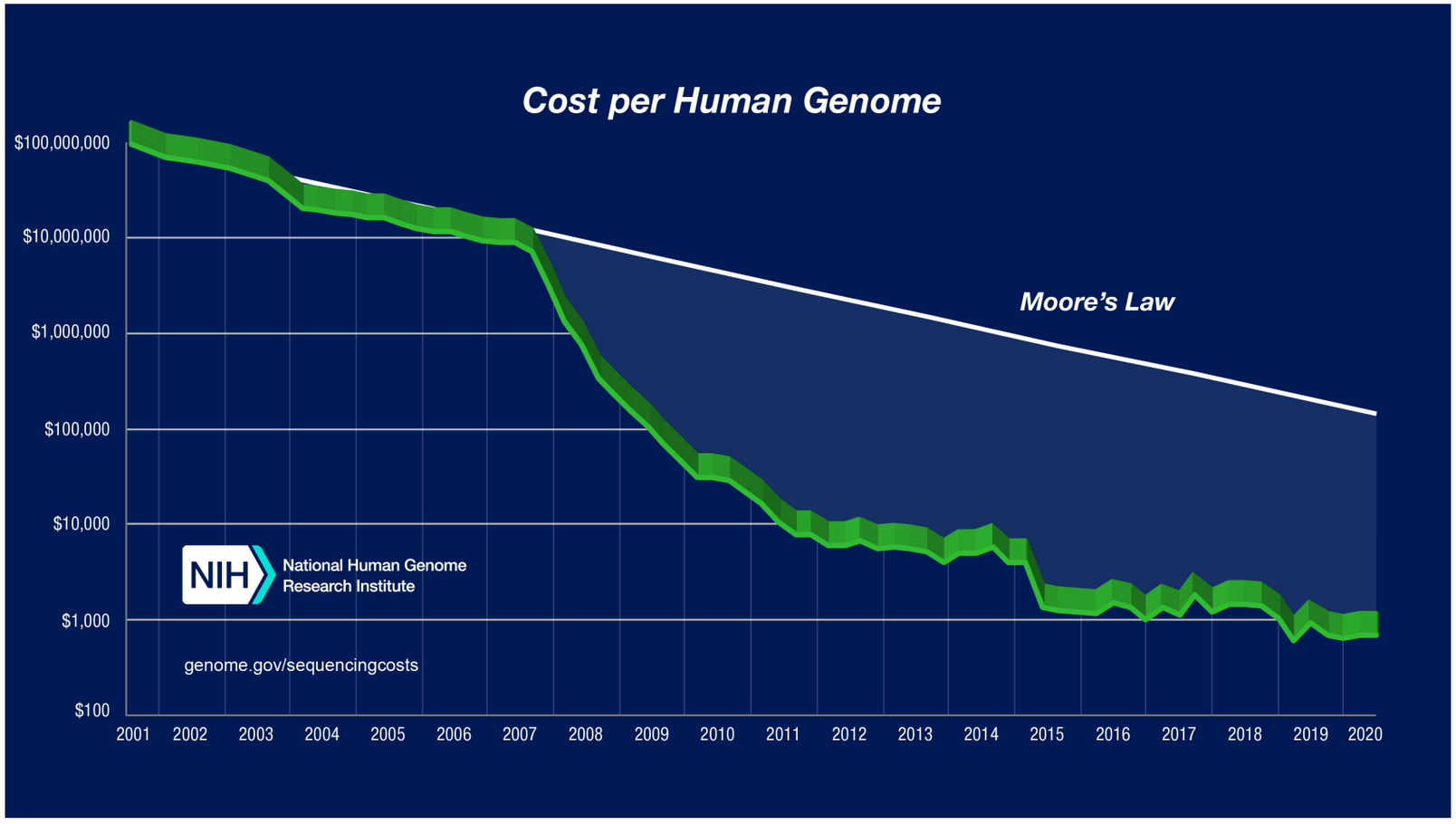


Shotgun principle /
Massive parallel sequencing

Many competing high-throughput methods

- short read sequencing: Illumina (next-gen)
- long read sequencing: PacBio, Nanopore (third gen)
- linked read sequencing: e.g. 10x Genomics

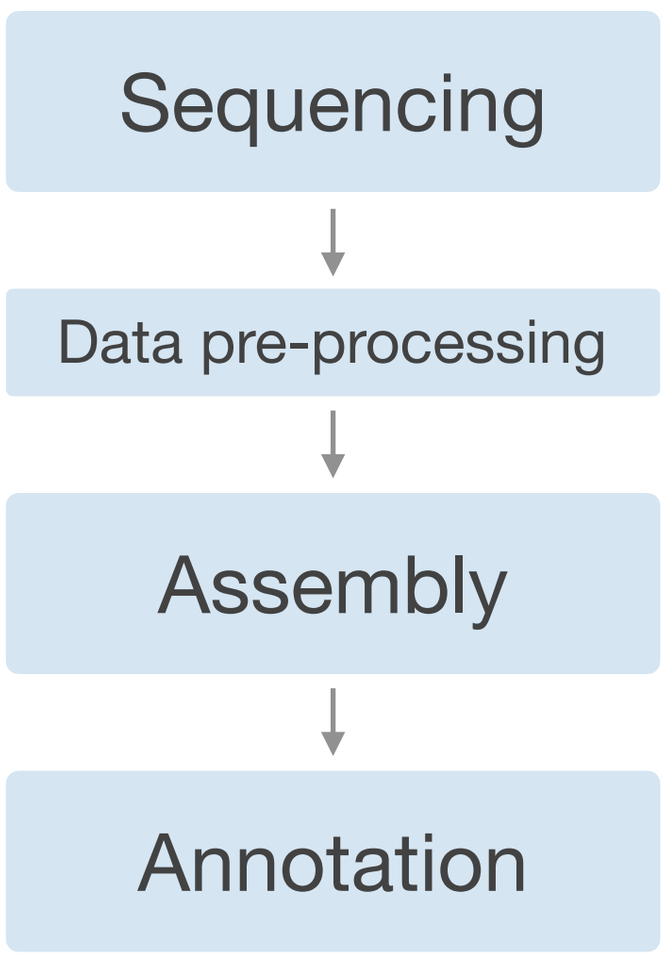
Genome sequencing costs



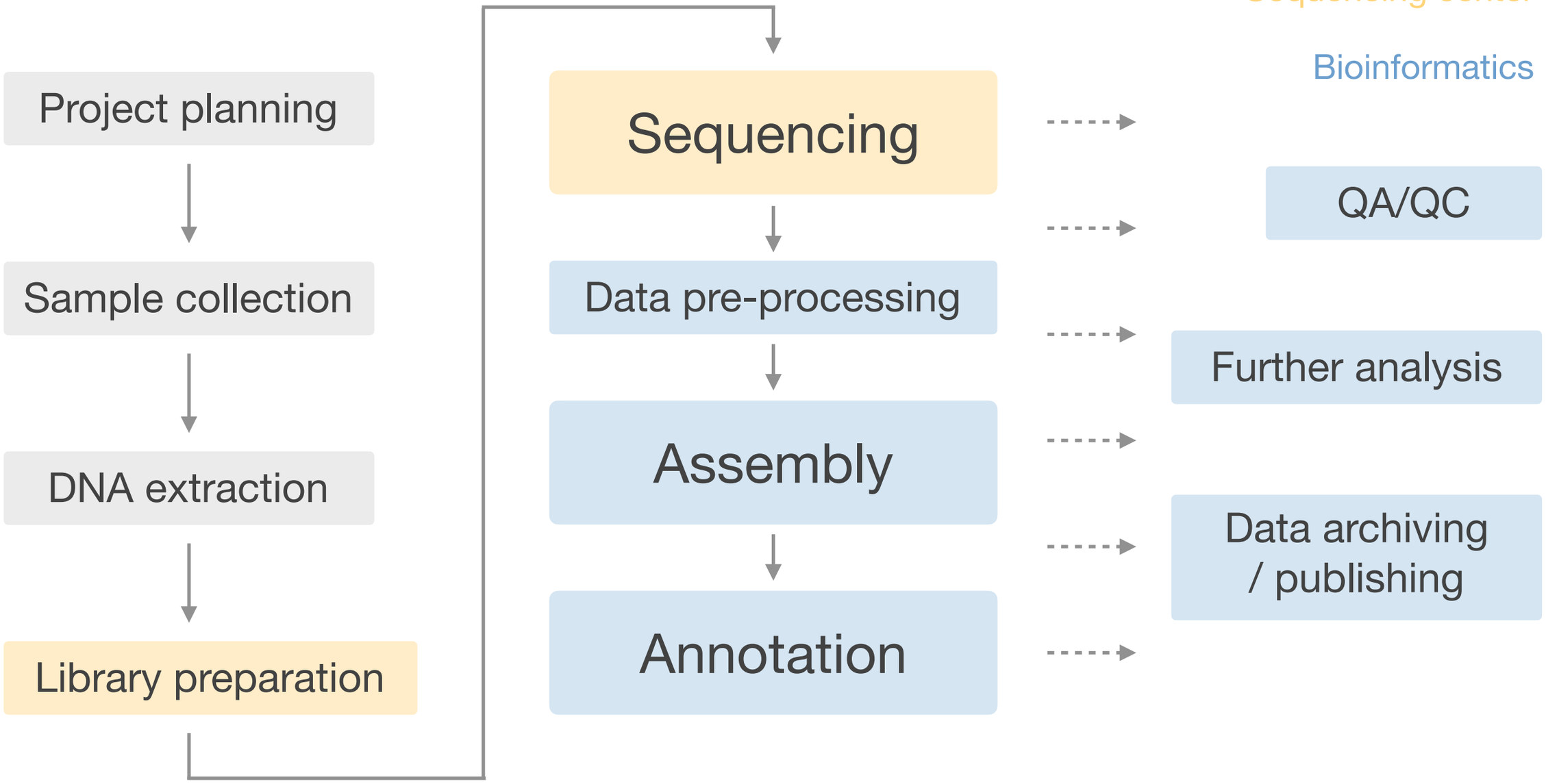
The bottleneck is not cost anymore, but knowing how to process the data

www.genome.gov/sequencingcostsdata, accessed Jul 5 2021

De novo genome sequencing workflow

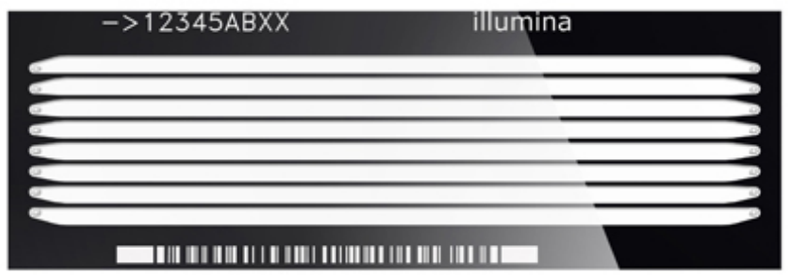


De novo genome sequencing workflow

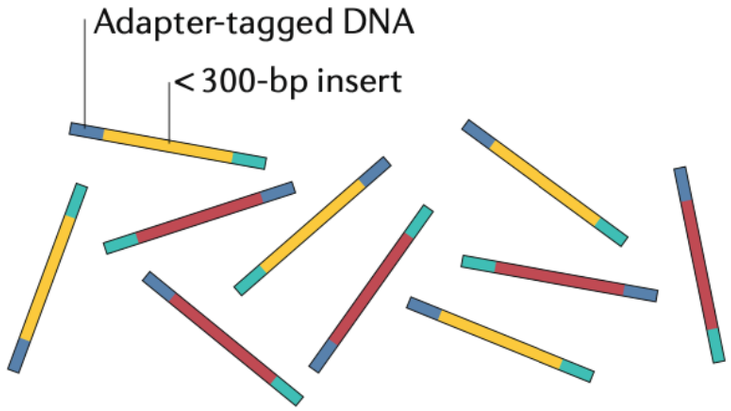


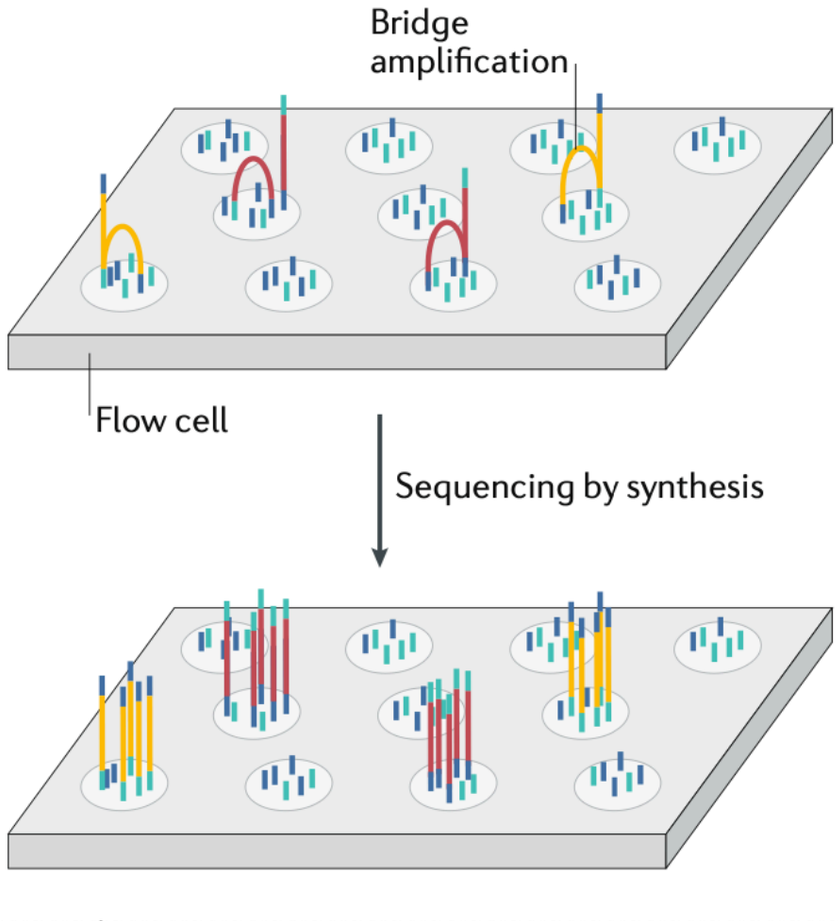
Illumina sequencing (short reads)

1. DNA fragmentation
2. Adapter ligation
3. Attachment to flow cell
4. Cluster generation (bridge PCR)
5. Sequencing by synthesis



Flow cell

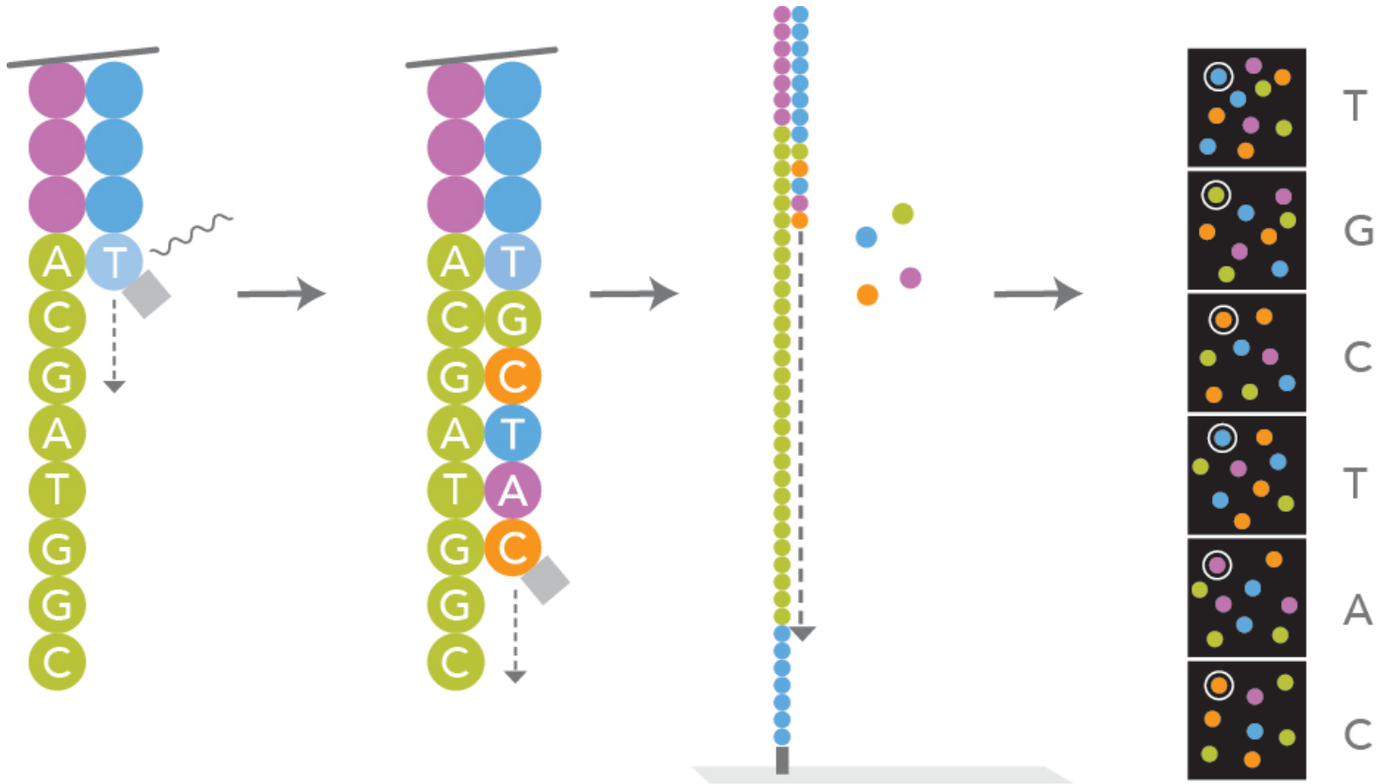




Logsdon et al. 2020, Nature Reviews

Illumina sequencing (short reads)

Sequencing by synthesis

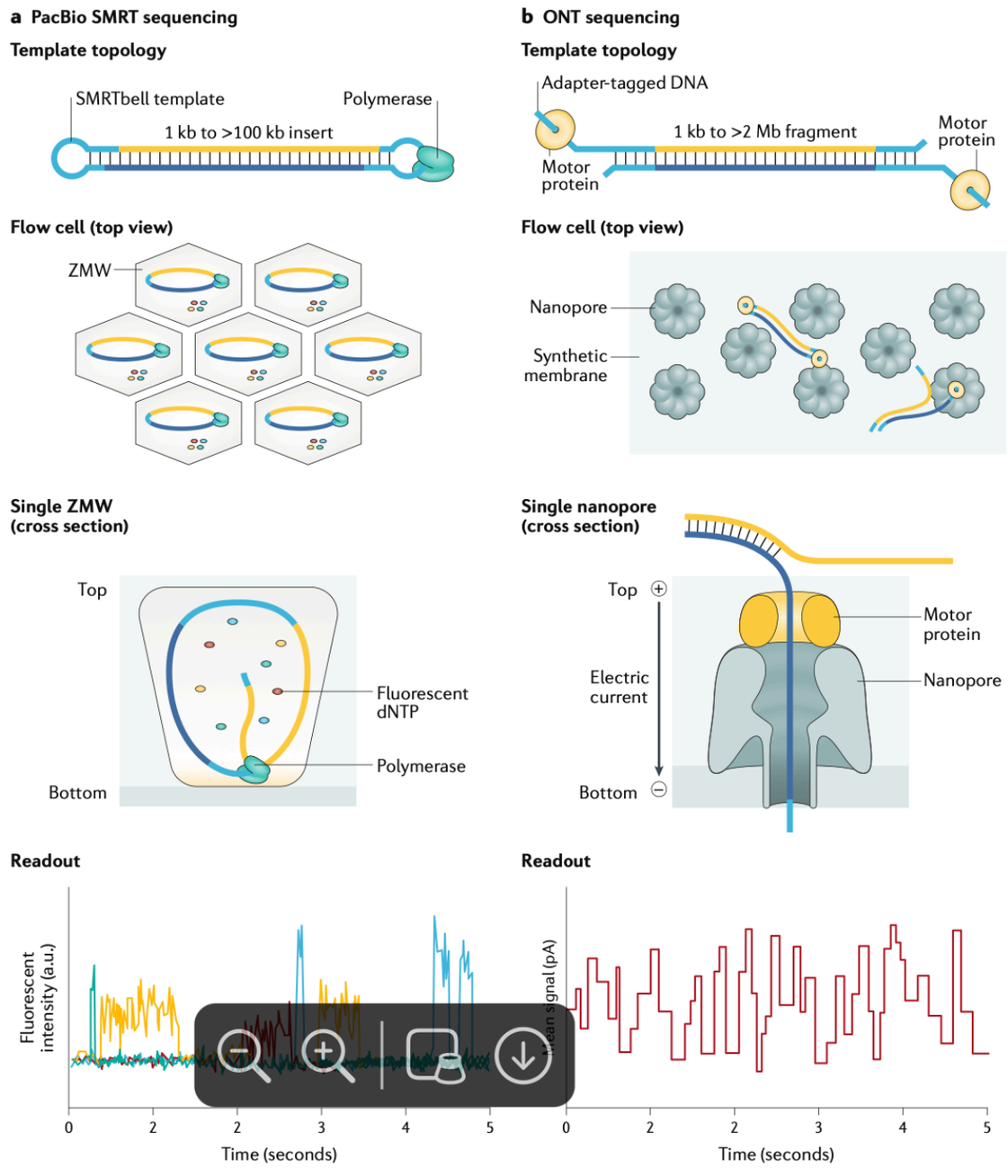


Illumina, Inc.

Long read sequencing

PacBio

SMRT



Oxford

Nanopore

Technologies

Logsdon et al. 2020, Nature Reviews

Sequencing technologies compared

Technology	Typical read length	Accuracy	Gb per run	Cost per Gb	Devices
Illumina	50–250 bp	>99.9 %	15–3000	$10–60	HiSeq, MiSeq, NovaSeq
PacBio	10–60 kb	87–99%	5–100	$20–200	Sequel, Sequel II
Nanopore	10–200 kb	87–98%	1–100	$20–2000	PromethION, MinION

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

Logsdon et al. 2020, Nature Reviews

Sequencing technologies compared

Illumina HiSeq 4000

Illumina, Inc.

Nanopore MinION

Oxford Nanopore Technologies

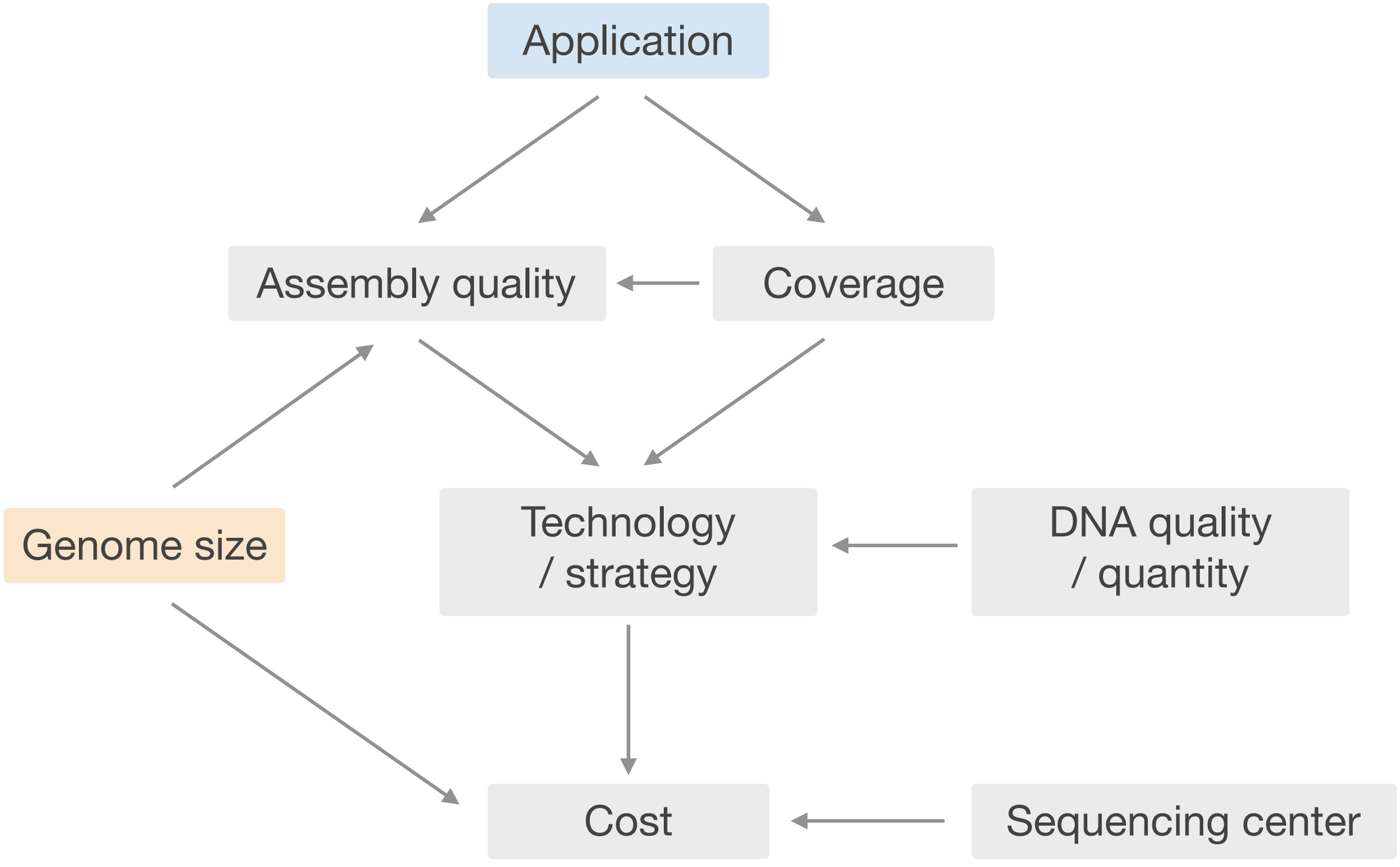
Coverage

- 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:  00111111223333333333322222211
```

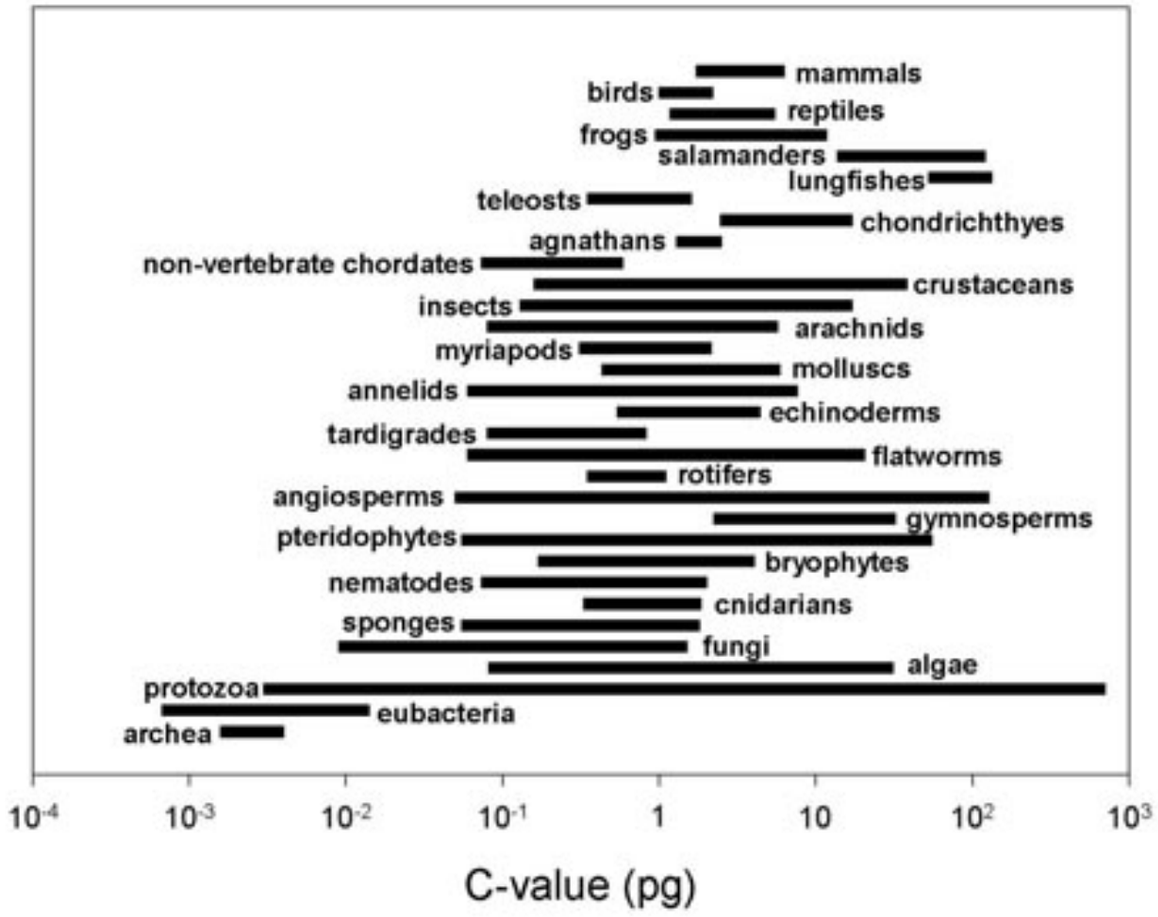
Planning your genome project



Genome size

How large is your genome?

Search Animal Genome Size Database: www.genomesize.com



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

Gregory 2021, Animal Genome Size Database

Sample collection

- Permits required?
- Suitable tissue / cell types vary by organism

  (e.g. whole body, coral cores, blood, gills, fin clips)
- Store in 100% ethanol or stabilizing solution (e.g. DNA Shield), ideally cold
- Do not freeze and thaw repeatedly
- Collect metadata

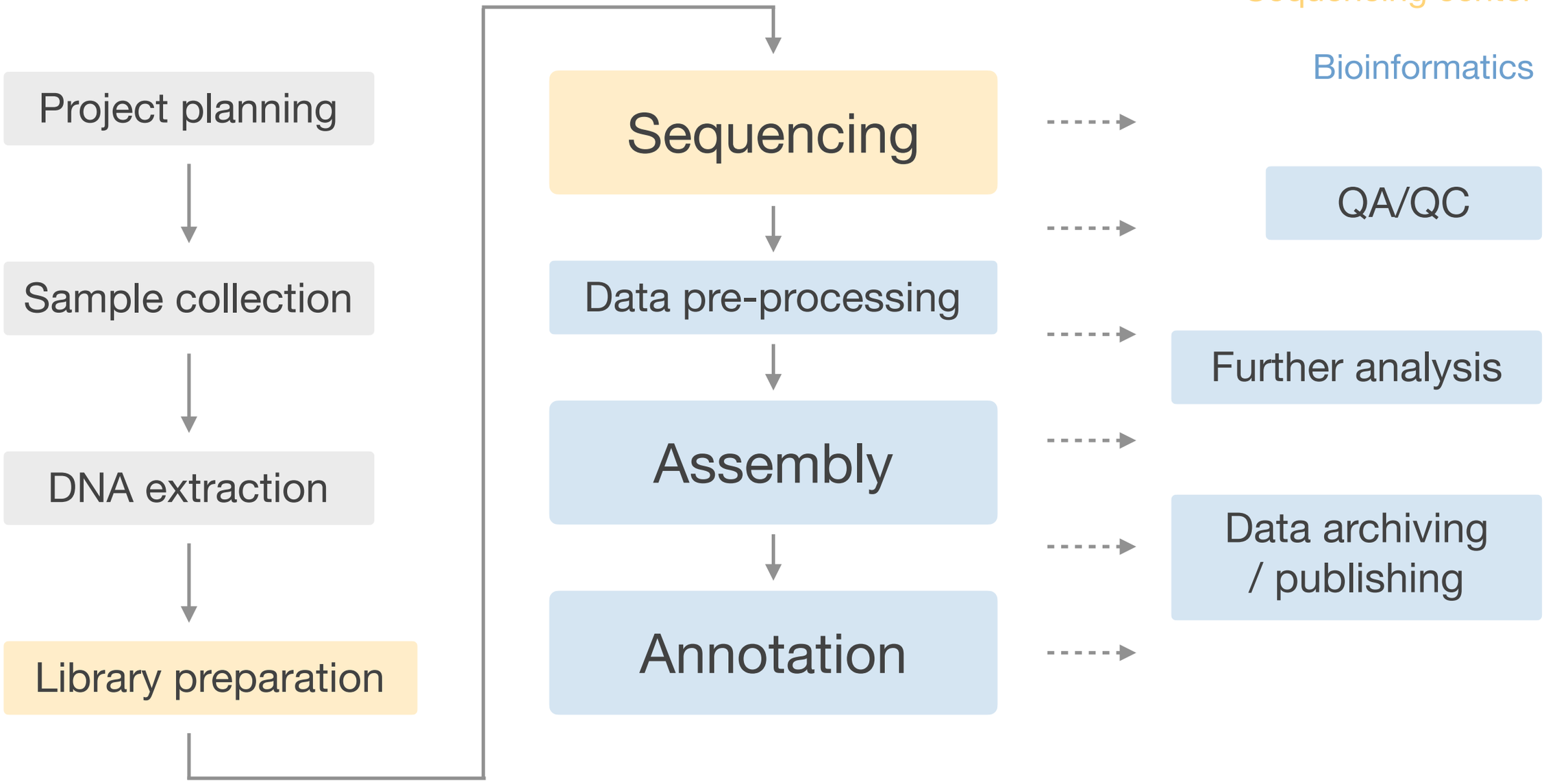
Metadata

- Species / population
- Date and time of day
- Geographic coordinates (possibly depth, site name)
- Tissue type (possibly developmental stage)
- Person collecting
- Possibly ecological context (e.g. coral reef)
- Sampling method (e.g. snorkeling, core sample)
- Assign unique sample IDs if collecting population samples

DNA extraction

- Protocol depends on tissue and desired quality (perform test runs)
- Common methods include commercial kits and phenol:chloroform extraction
- Document DNA quality and concentration
- Store and ship DNA cold if possible
- Do not freeze and thaw repeatedly

De novo genome sequencing workflow



Library preparation and sequencing

- Typically outsourced to academic or commercial sequencing center
- Academic providers usually offer discounts to affiliates
- Services, DNA requirements often found on website, but not prices

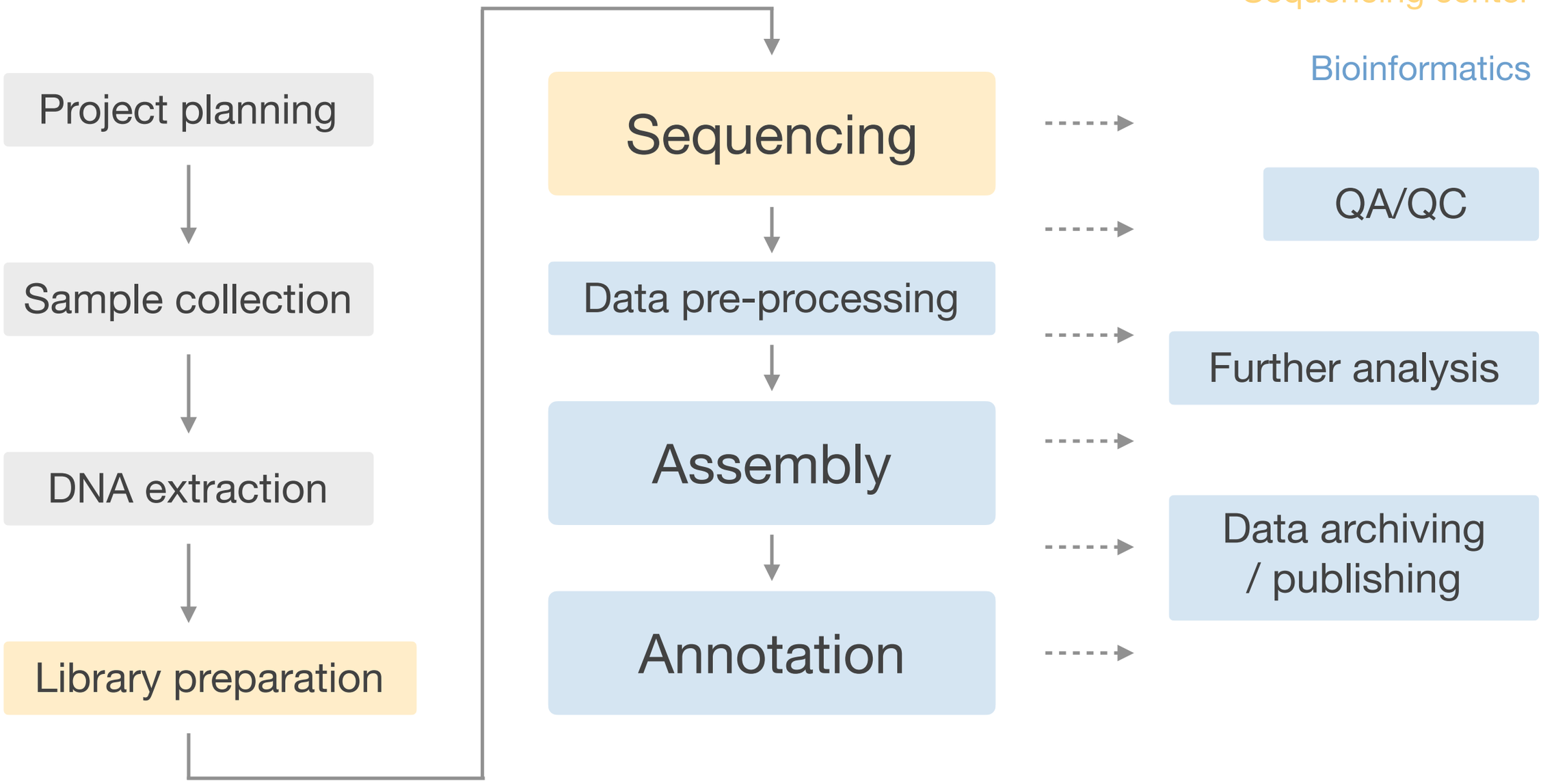
  (request quotes, possibly try www.scienceexchange.com)
- Contact center well in advance for scheduling, shipping and labeling

  preferences

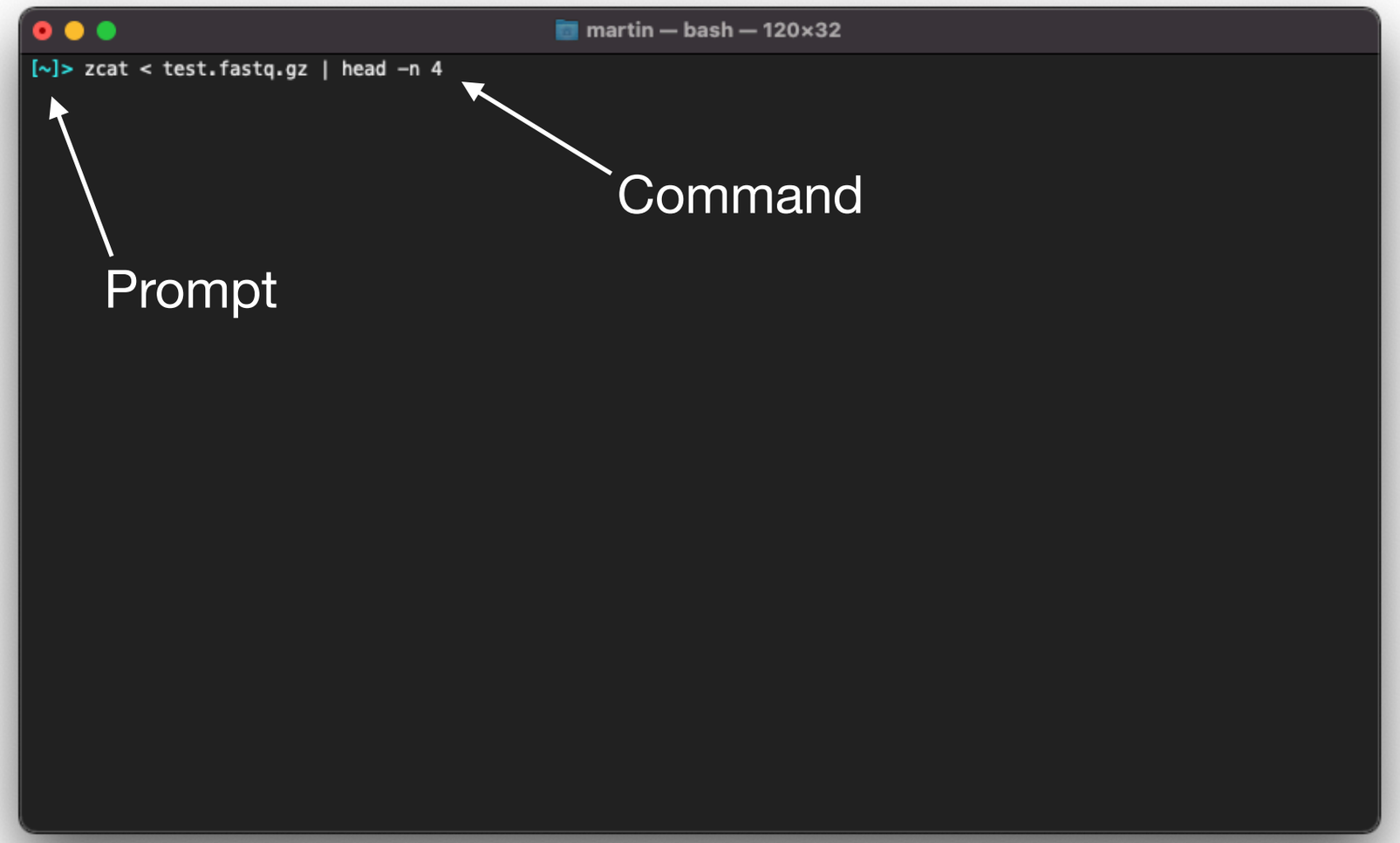
2. Data pre-processing and QA/QC

- Introduction to bioinformatics and the command line
- Raw sequencing data, trimming & filtering
- Read quality assurance / control

De novo genome sequencing workflow



The command line



Shell window

(terminal / console)

Command line interface + command language = shell

Most common shell: bash

The command line

Advantages

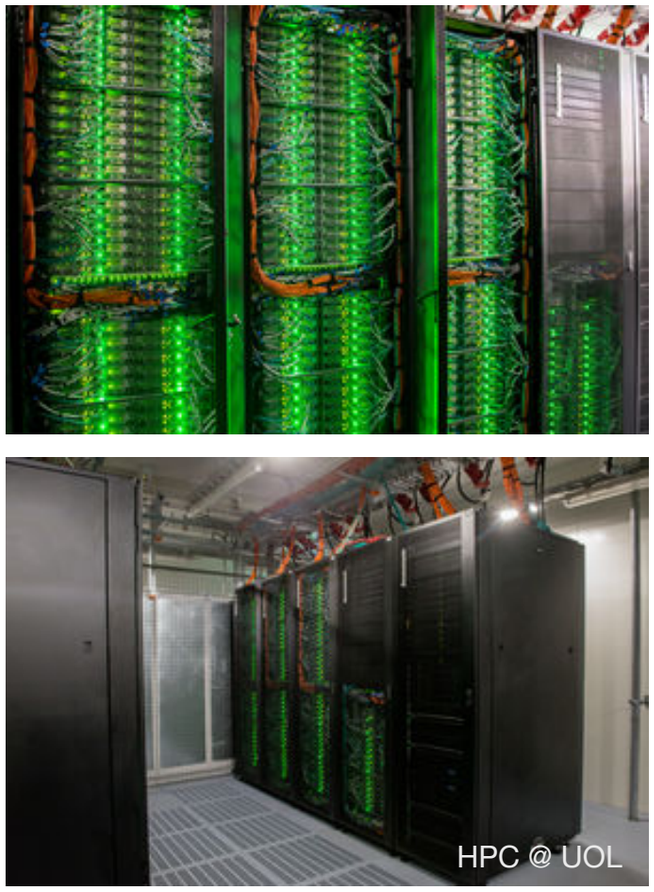
- High flexibility
- Automation / pipelines > complex workflows
- High reproducibility / ease of documentation
- Use of computing clusters

Many tools are only available on the command line

A note on high performance computing

HPC cluster advantages

- Parallelization
- Memory capacity (storage and RAM)
- Rich, actively managed software environment (Linux based)
- Job submission system


HPC @ UOL

Possible alternative: cloud computing (e.g. Amazon Web Services)

How to access and use bash

macOS

- open Terminal: /Applications/Utilities/Terminal.app
- (type “bash”)

Windows

- shortcut for this workshop: install git, type “git bash” from Start menu
- for serious use, run Linux subsystem on Windows
- Alternatively, install Cygwin

Text editors

Large text-based files are ubiquitous in genomics and programming

Recommended for all OS: Atom (https://atom.io)

Built-in:

- TextEdit (macOS)
- Notepad (Windows)

Exercise instructions

```
Command (type this and press enter to execute)      # comments will be ignored
                                                    # replace < > with appropriate content
```

```
Output (this should appear as a result to your commands)
```

```
Additional information and tips
```

Exercise instructions

Download example data using git

```
git clone https://github.com/mhelmkampf/workshop_zmt.git          # download example data
```

```
Cloning into ‘workshop_zmt’ …
…
…
…
Resolving deltas: 100% (7/7), done.
```

Bash basics

Typical usage (syntax)

```
command [-options] [file]
```

Navigating the file system

```
ls                                          # list directory contents
ls -l                                       # list in long format

cd <dir>                                    # change directory
cd ..                                       # move back one level (to parent directory)

pwd                                         # display path of current directory

man <command>                               # help manual (not available in git bash)
```

Bash basics

Path: location of file or directory in system's file structure

`workshop_zmt/project/1_rawdata`

Path and keyboard shortcuts

```
..                                          # = parent directory
.                                           # = current directory
tab                                         # autocomplete file and directory names
up arrow                                    # display previous command
```

Bash basics

Organizing files

```
mkdir <dir>                                    # create new directory

mv <file / dir>                                # move or rename file or directory
cp <file> <new path / name>                    # make copy of file under new name
rm <file>                                      # delete file
rm -r <dir>                                    # delete directory

wget <URL>                                     # download file
gzip <file>                                    # compress file
gzip -d <file>                                 # uncompress file
```

Bash basics

Displaying and manipulating files

```
cat <file>                                  # display text file
head <file>                                 # display the first 10 lines of file
tail <file>                                 # display the last 10 lines of file

nano <file>                                 # edit file using the Nano text editor

grep ‘pattern’ <file>                       # search for ‘pattern’ in file
wc –l <file>                                # count number of lines in file
```

Bash basics

Operators

```
<command> > <file>                    # write output to file
<command1> | <command2>               # execute second command on output of first (pipe)
```

Looping over multiple files

```
for FILE in <list>                    # list: e.g. *.fastq
do
  <command> ${FILE}
  …
done
```

Raw data

Receiving data: FTP server link, `wget` command

```
cd workshop_zmt/projects/1_rawdata
gzip -d Hpue_raw300_F.fastq.gz                                  # decompress file
ls -l                                                           # check if successful
```

```
total 68936
-rw-r--r--  1 martin  staff    26M May 21 13:13 Hpue_raw300_F.fastq
-rw-r--r--  1 martin  staff   7.7M May 21 13:16 Hpue_raw300_R.fastq.gz
drwxr-xr-x  4 martin  staff   128B Jun 28 20:42 md5
```

Raw data

The FASTQ format

```
head -n 4 Hpue_raw300_F.fastq                    # display first 4 lines of file
```

```
@HWI-ST1293:199:HA9JHADXX:1:1101:1044:1603 1:N:0:CGATGT
NCCCTGTTAAAGGATCATCTCTGACCTATCATTGTGGTGTAATCACATTTAACACAATCACGATGTGCTTTACCTGCAGCATCTTTACAG
CAGGGCTGGGAGATATGACCAAAAACAGTTATGATAATATGTTTATTTCTATTGAAAATCA
+
#1=DDFFFHHHHHHJJJJJJJJJJJJJJJJJJJJJJGHIJJJJJIIIJJJJHJJJJJJJJJJJJIJIJJJJJJJHHHHHFFFFFFEEEE
EEDDDDDDDDD@DDDEDEDDDBDDDDDDCDCEDDEEDDEEEDEECDDEDEDEEDDDDDDDC
```

Raw data

The FASTQ format

```
head -n 4 Hpue_raw300_F.fastq                    # display first 4 lines of file
```

```
@HWI-ST1293:199:HA9JHADXX:1:1101:1044:1603 1:N:0:CGATGT
NCCCTGTTAAAGGATCATCTCTGACCTATCATTGTGGTGTAATCACATTTAACACAATCACGATGTGCTTTACCTGCAGCATCTTTACAG
CAGGGCTGGGAGATATGACCAAAAACAGTTATGATAATATGTTTATTTCTATTGAAAATCA
+
#1=DDFFFHHHHHHJJJJJJJJJJJJJJJJJJJJJGHIJJJJJIIIJJJJHJJJJJJJJJJJJIJIJJJJJJJHHHHHFFFFFFEEEE
EEDDDDDDDDD@DDDEDEDDDBDDDDDDCDCEDDEEDDEEEDEECDDEDEDEEDDDDDDDC
```

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

Raw data

The FASTQ format

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
```

Raw read quality check

Using the FastQC software

```
cd ../../apps                                   # change into apps folder
unzip fastqc_v0.11.9.zip                        # decompress software
FastQC/fastqc -h                                # confirm software works, show help

cd ../project/1_rawdata                         # change into raw data folder
../../apps/FastQC/fastqc Hpue_raw300_F.fastq    # run FastQC on file
```

```
Started analysis of Hpue_raw300_F.fastq
Approx 5% complete for Hpue_raw300_F.fastq
Approx 10% complete for Hpue_raw300_F.fastq
…
Analysis complete for Hpue_raw300_F.fastq
```

Results can also be found in `workshop_zmt/project/2_qc`

Raw read quality check

Using the FastQC software

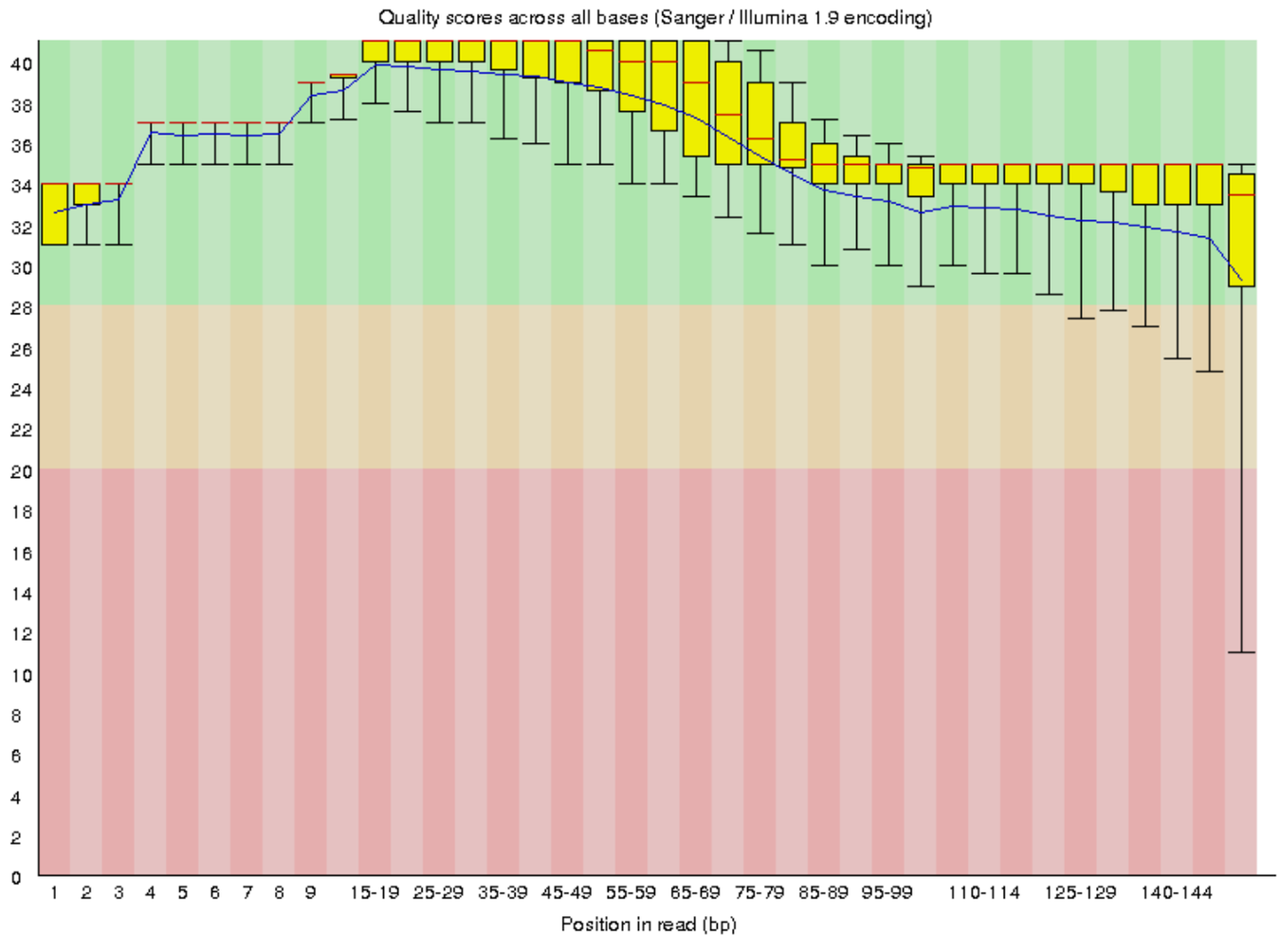
Basic Statistics

Measure	Value
Filename	Hpue_raw300_F.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	75000
Sequences flagged as poor quality	0
Sequence length	151
%GC	40

Raw read quality check

Using the FastQC software

Per base sequence quality



Raw read quality check

Using the FastQC software

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	640	0.8533333333333334	TruSeq Adapter, Index 2 (100% over 50bp)

Example for bad Illumina data:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Read trimming & filtering

Using cutadapt for adapter removal and quality trimming

```
../../apps/cutadapt-3.4.exe -help          # confirm app works, show help (from 1_rawdata)

../../apps/cutadapt-3.4.exe -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -q 10 -o
Hpue_raw300_F_trim.fastq Hpue_raw300_F.fastq
# run FastQC (-q 10: quality-filter last 10 bases, -a: trim Illumina adapter)
```

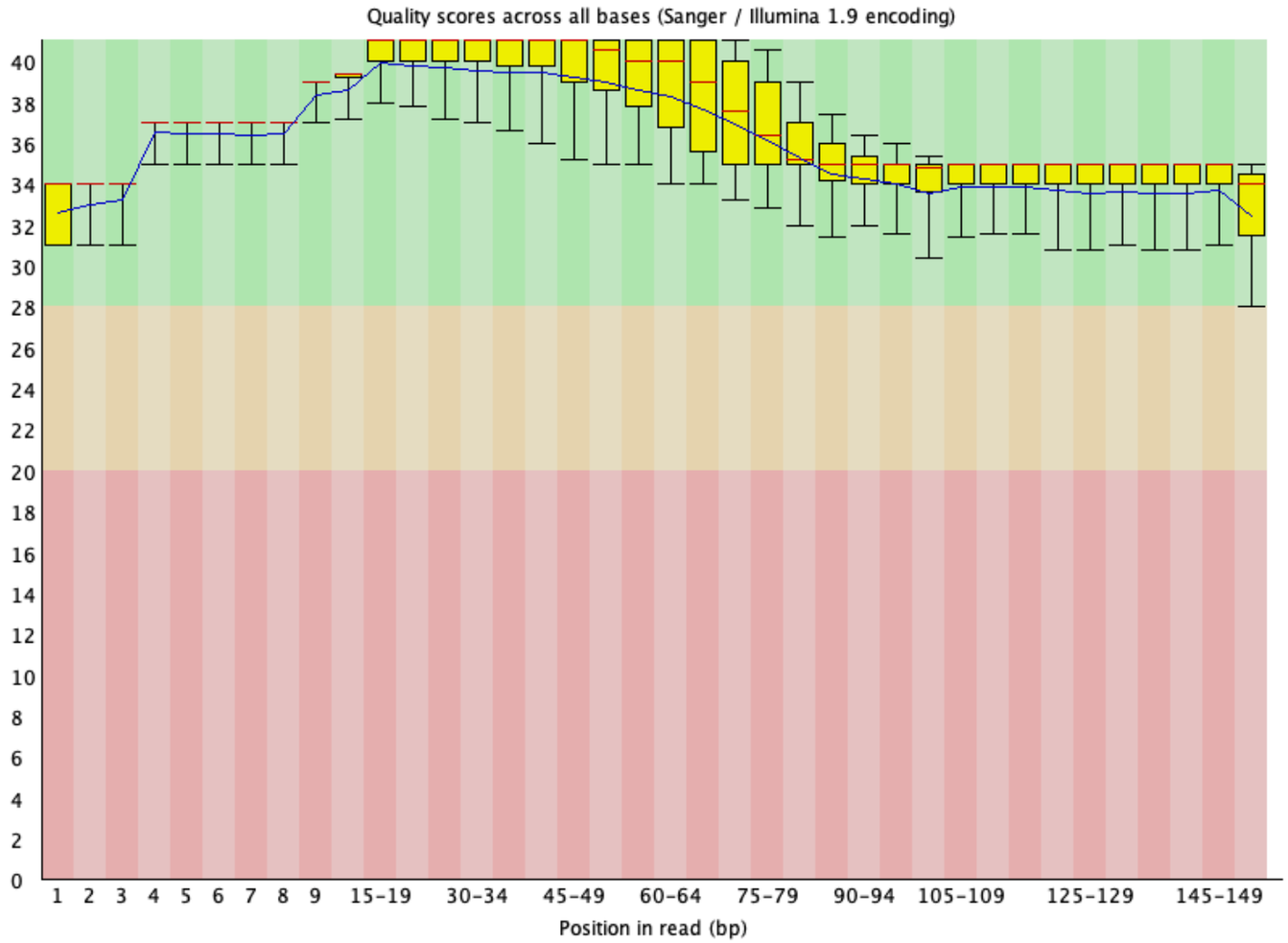
```
This is cutadapt 3.4 with Python 3.9.2
Command line parameters: …
Processsing reads …
```

Rerun FastQC

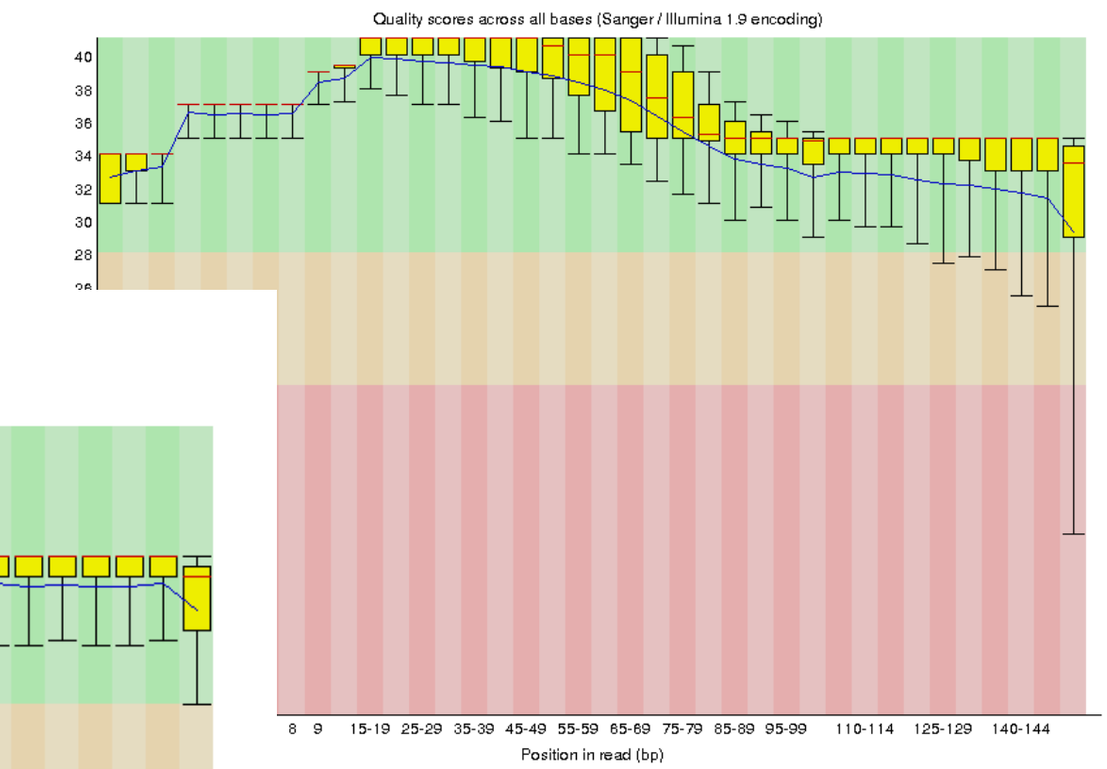
Filtered read quality check

Using the FastQC software

Per base sequence quality



Per base sequence quality



Raw reads

Filtered reads

Data pre-processing summary

- Raw read quality assessment
- Adapter removal
- Quality filtering (trim low-quality bases / reads)
- Check for contamination (reads of viral, bacterial, human origin)
- Filtered read quality assessment

De novo genome sequencing workflow