

Course 1: High Throughput Sequencing: Introduction to Genome Sequencing & Quality Control

Course Provider:

Dr. Tam Tran, USTH

Mathis Arnal, IRD, Chiang Mai University

Bioinformatics for Infectious Disease Ecology Research (BInDER)

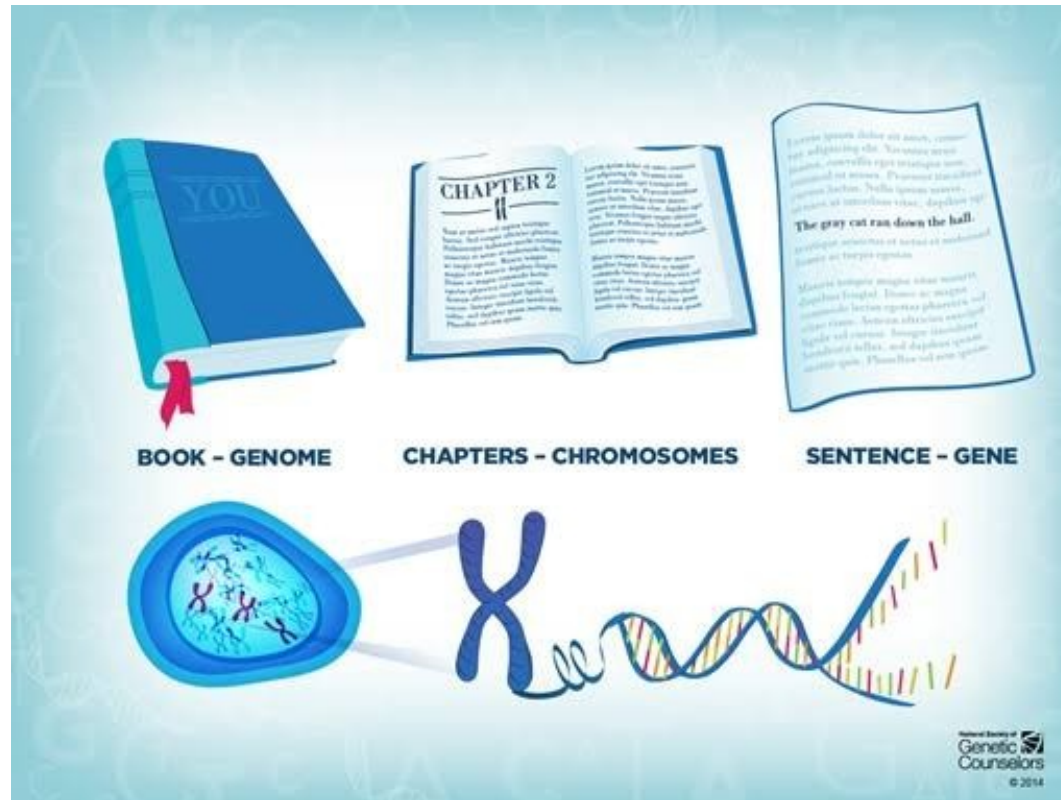
Outline

1. Introduction to genomics
2. How is genome sequenced?
3. Sequence File Formats
4. Upload Data to UseGalaxy
5. Quality check
6. Quality filtering (*Option*)

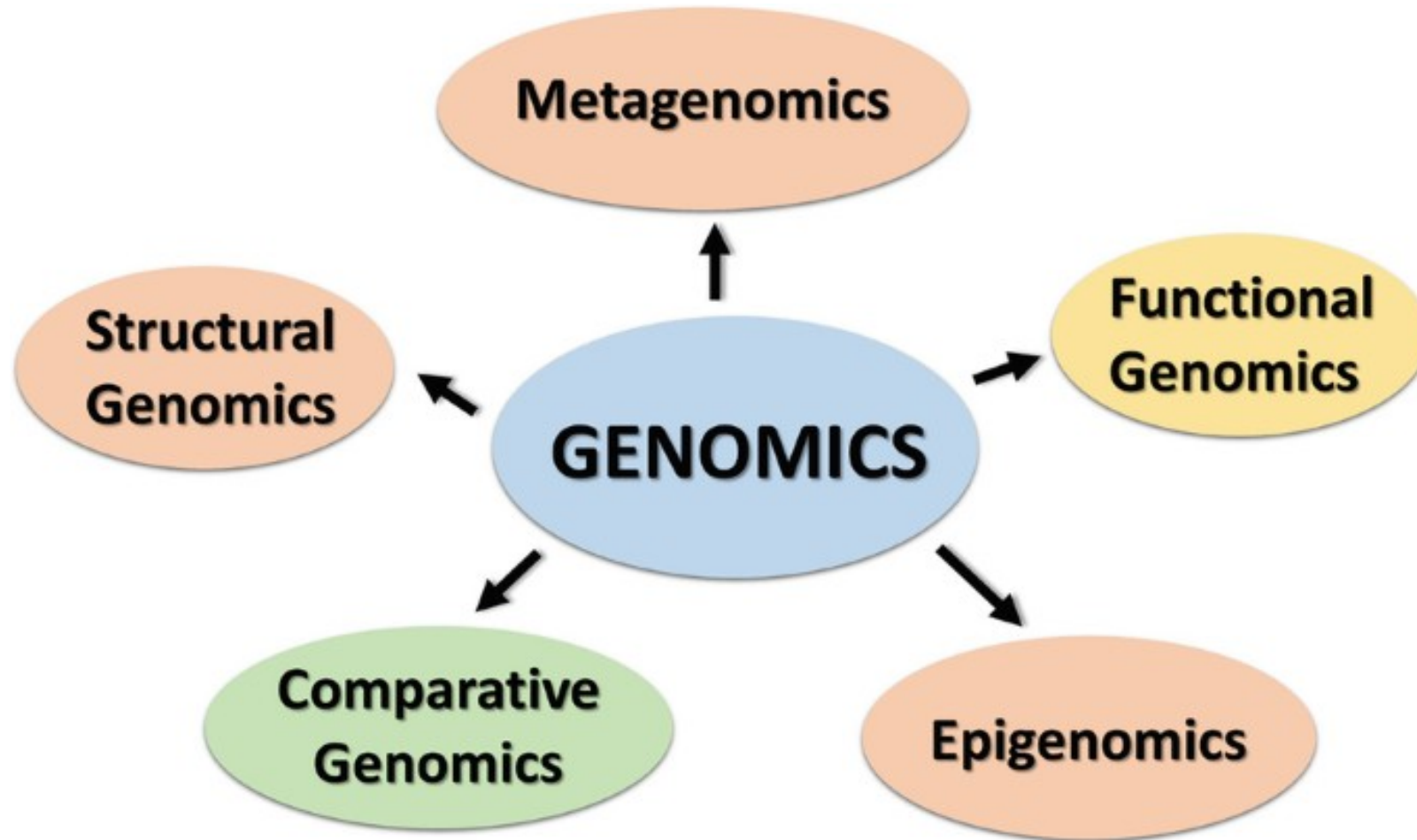


1. Introduction to genomics

- **Genome:** The entire DNA content (and thus all the genes) of an organism. It contains all information necessary for reproduction, development and physiology of an organism
- **Genomics** is the study of all genes present in an organism



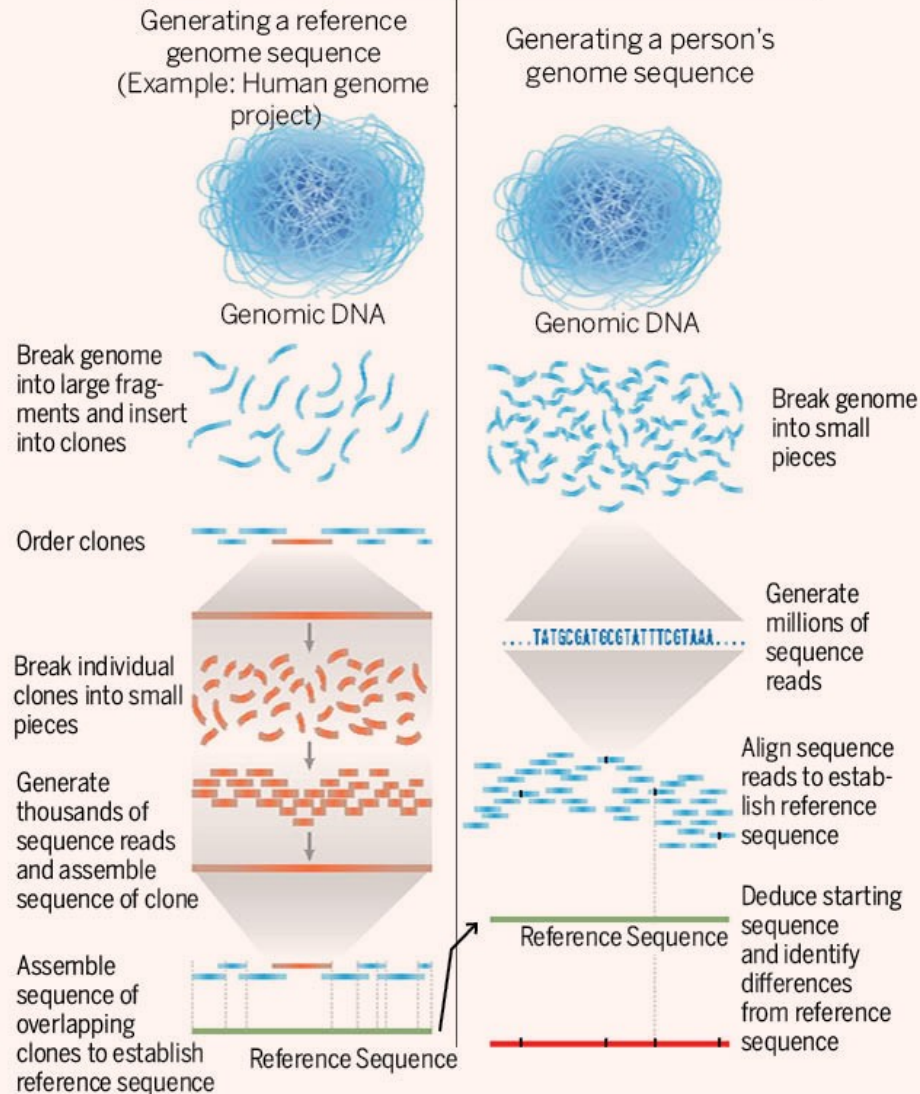
Types of the Genomics



2. How is genome sequenced?

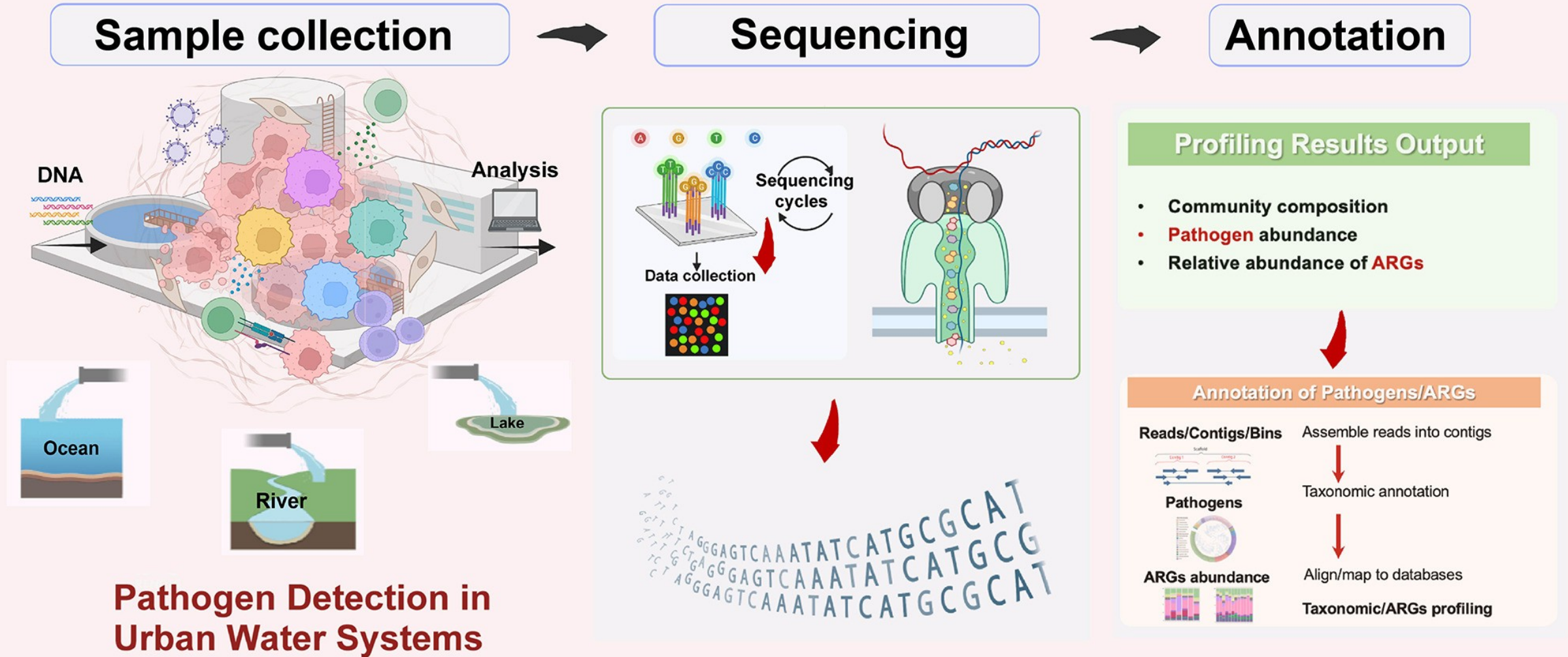
1st draft completed in 2001
13-year-long, \$2.7 billion

Human Genome Sequencing



Today cost:
\$1000/genome

Genomics research workflow



An example of high-throughput sequencing for pathogen detection in urban water systems

(DOI: 10.1016/j.scitotenv.2023.165867)

The evolution of sequencing technology

First Generation



Sanger Sequencing
Maxam and Gilbert
Sanger Chain-termination

- Infer nucleotide identity using dNTPs then visualize with electrophoresis
- 500-1000 bp fragments

Second Generation (Next Generation Sequencing)



454, Solexa,
Ion Torrent
Illumina

- High throughput from the parallelization of sequencing reactions
- ~50-500 bp fragments

Third Generation



PacBio
Oxford Nanopore

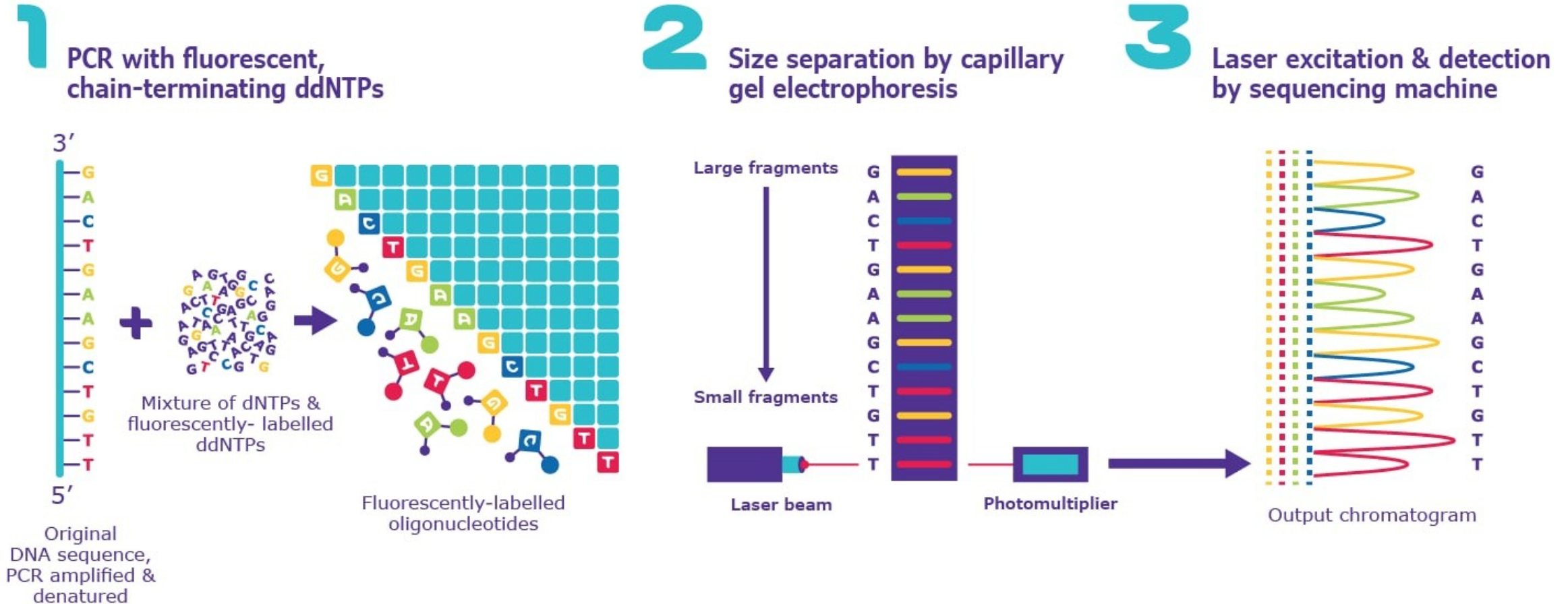
- Sequence native DNA in real time with single-molecule resolution
- Tens of kb fragments, on average

Short-read sequencing

Long-read sequencing

Sanger sequencing

<https://www.youtube.com/watch?v=dVRB4CaLizc>



Three Basic Steps of Automated Sanger Sequencing (<https://www.sigmaaldrich.com/>).

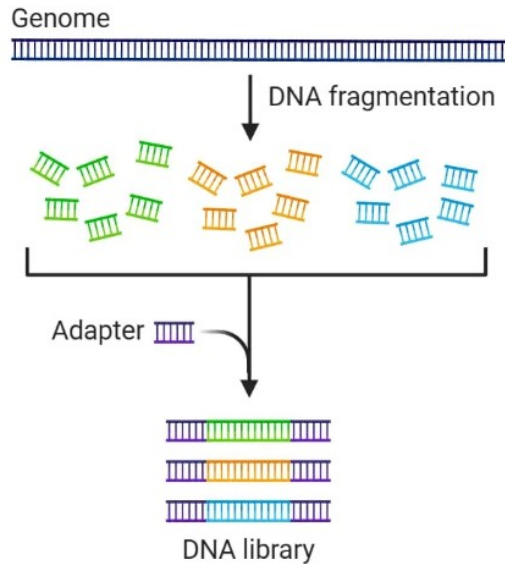
Next Generation Sequencing (NGS)

- *Short Reads Become Fast and Efficient*

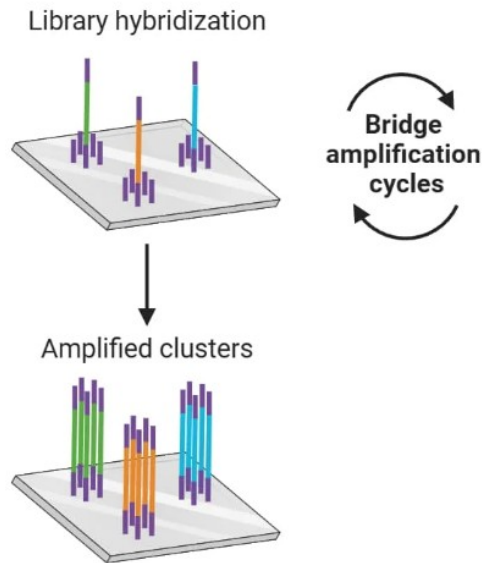
Illumina sequencing: <https://www.youtube.com/watch?v=womKfikWlxM>

- Illumina are used more commonly
- Based on the **Sequencing By Synthesis** principle
- Advantage of technology: Cost-effective
- Disadvantage of technology: Short reads & Long run time

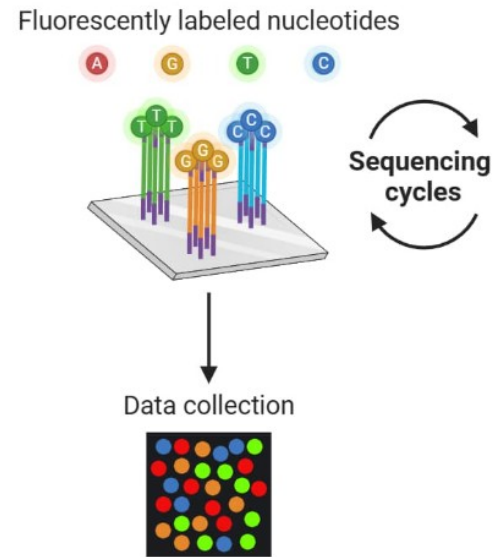
① Library preparation



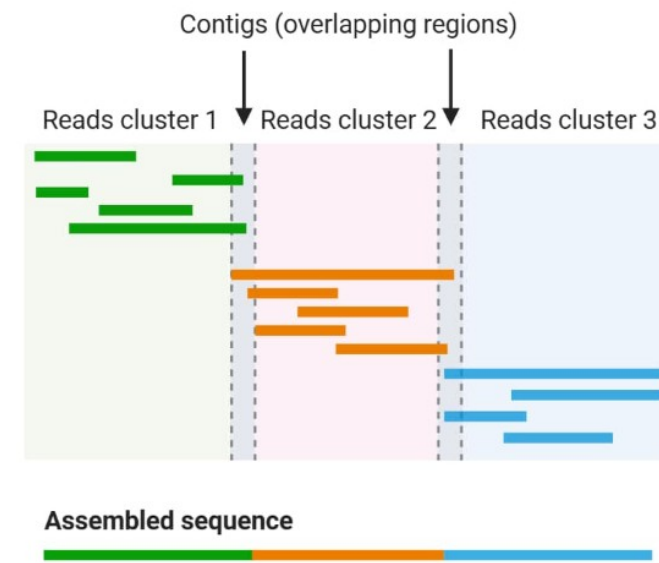
② DNA library bridge amplification



③ DNA library sequencing

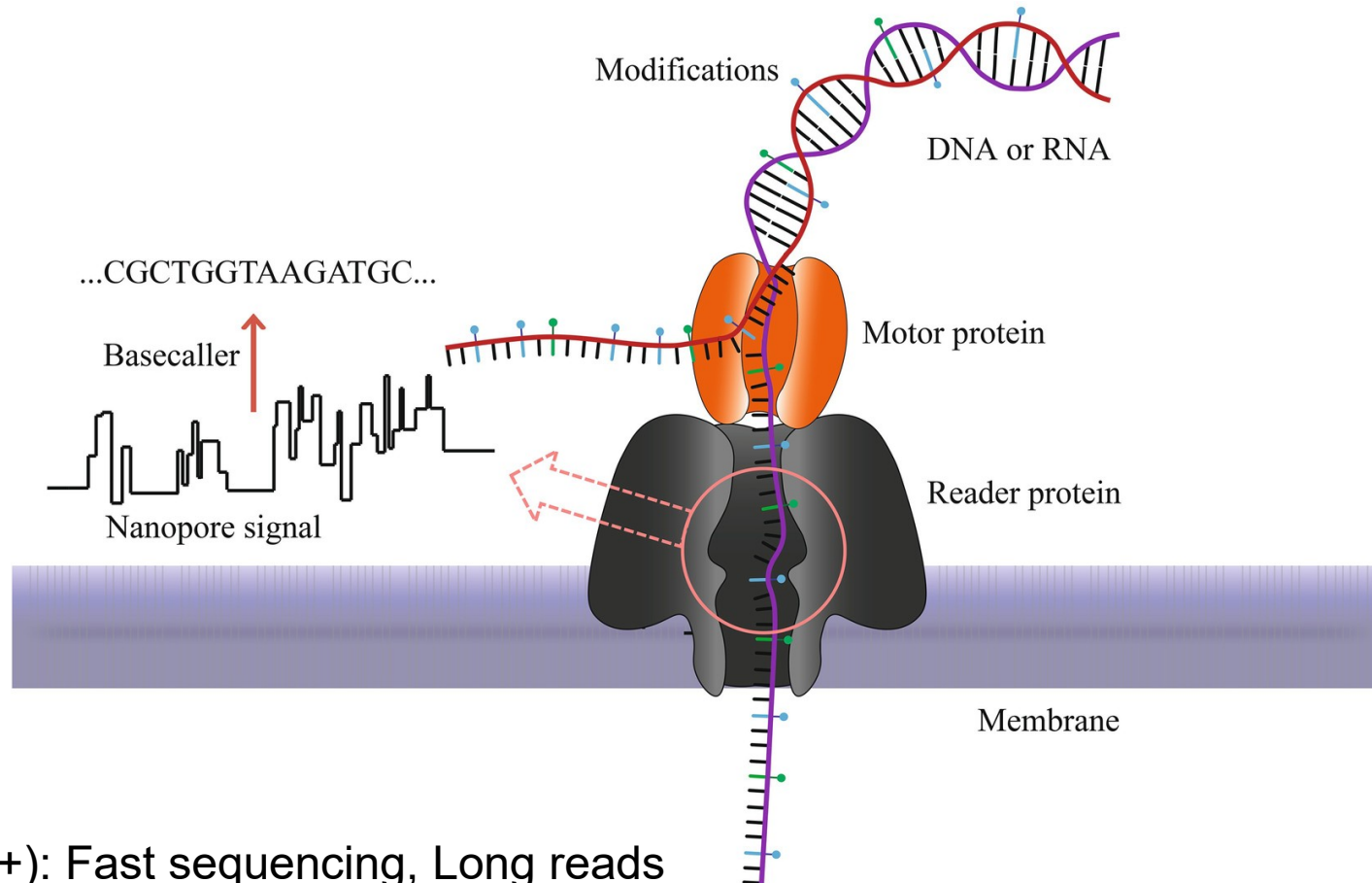


④ Alignment and data analysis



Oxford Nanopore sequencing

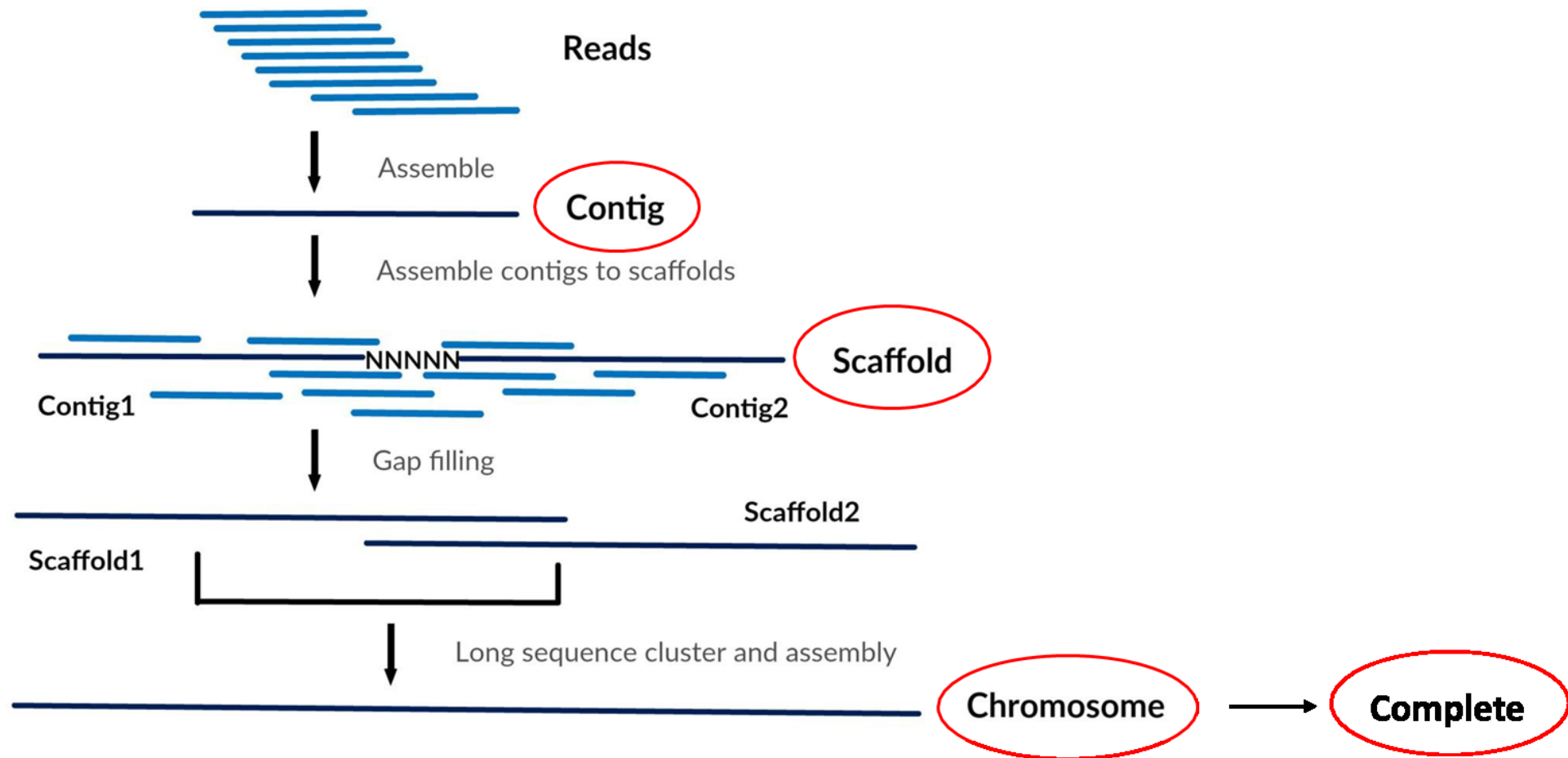
<https://www.youtube.com/watch?v=E9-Rm5AoZGw>



(+): Fast sequencing, Long reads

(-): High error rates

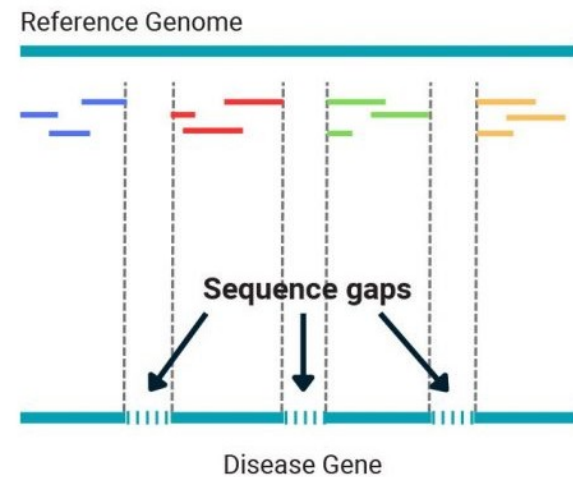
Completeness of the genome sequence



- **Contigs:** longer continuous sequences formed by merging overlapping reads
- **Scaffolds:** Ordered and oriented collections of contigs connected by gaps (a run of N's)

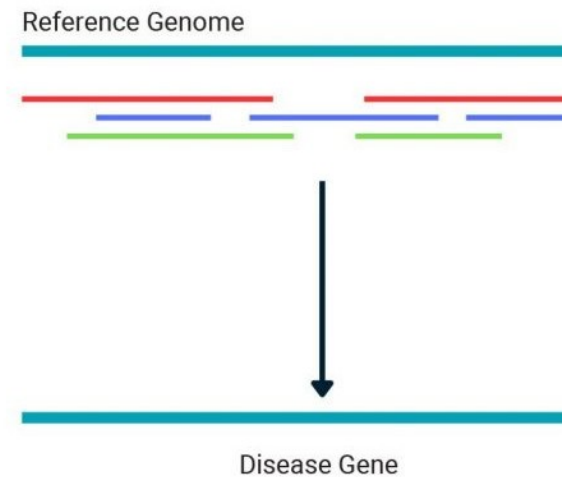
Challenges of Short-Read & Long-Read Sequencing

① SHORT READS



Short Read analysis struggles to cover harder to sequence areas, resulting in sequence gaps preventing accurate analysis

② LONG READS

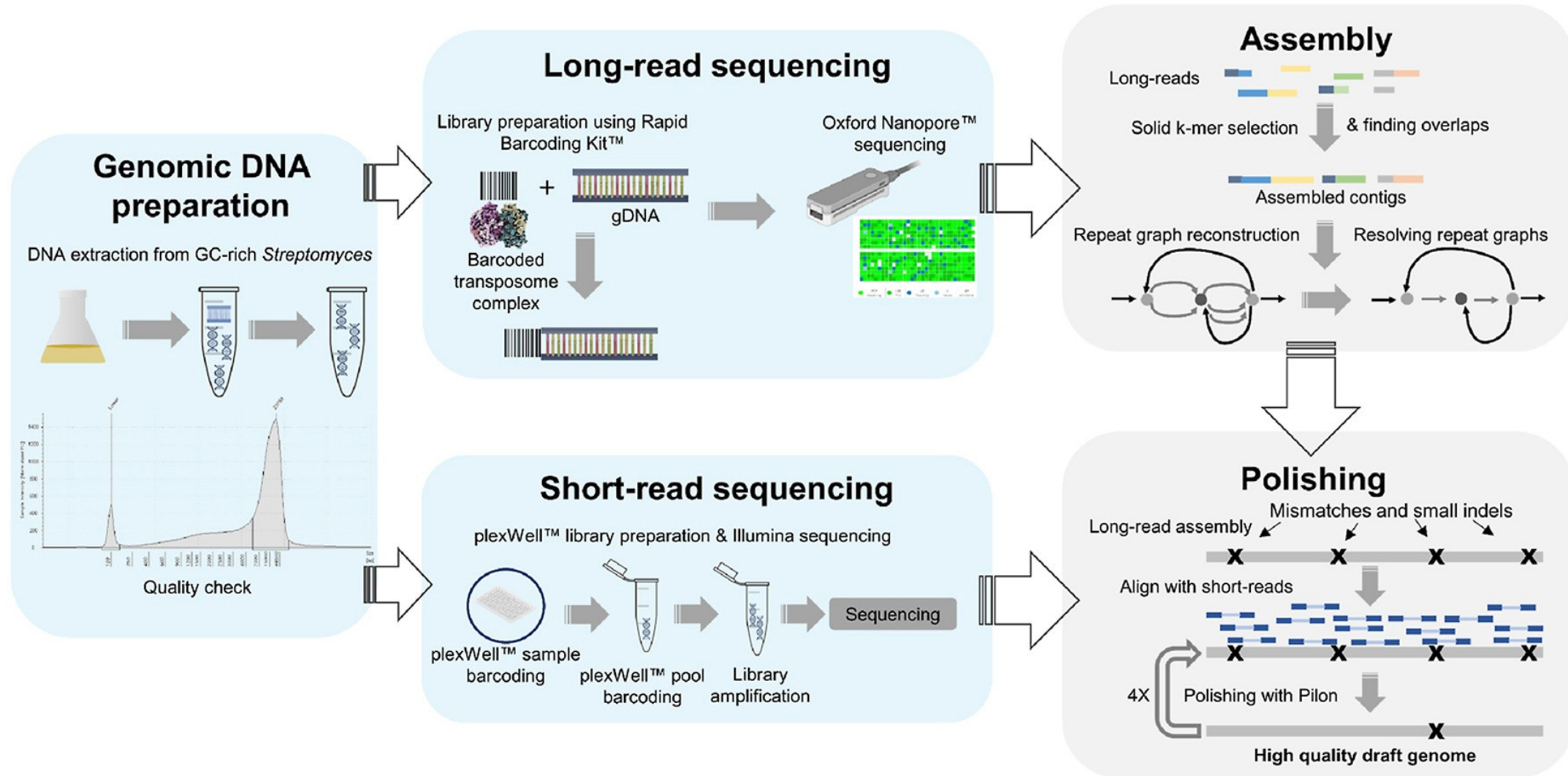


Long Reads are able to span these gaps for a more comprehensive analysis of the genome

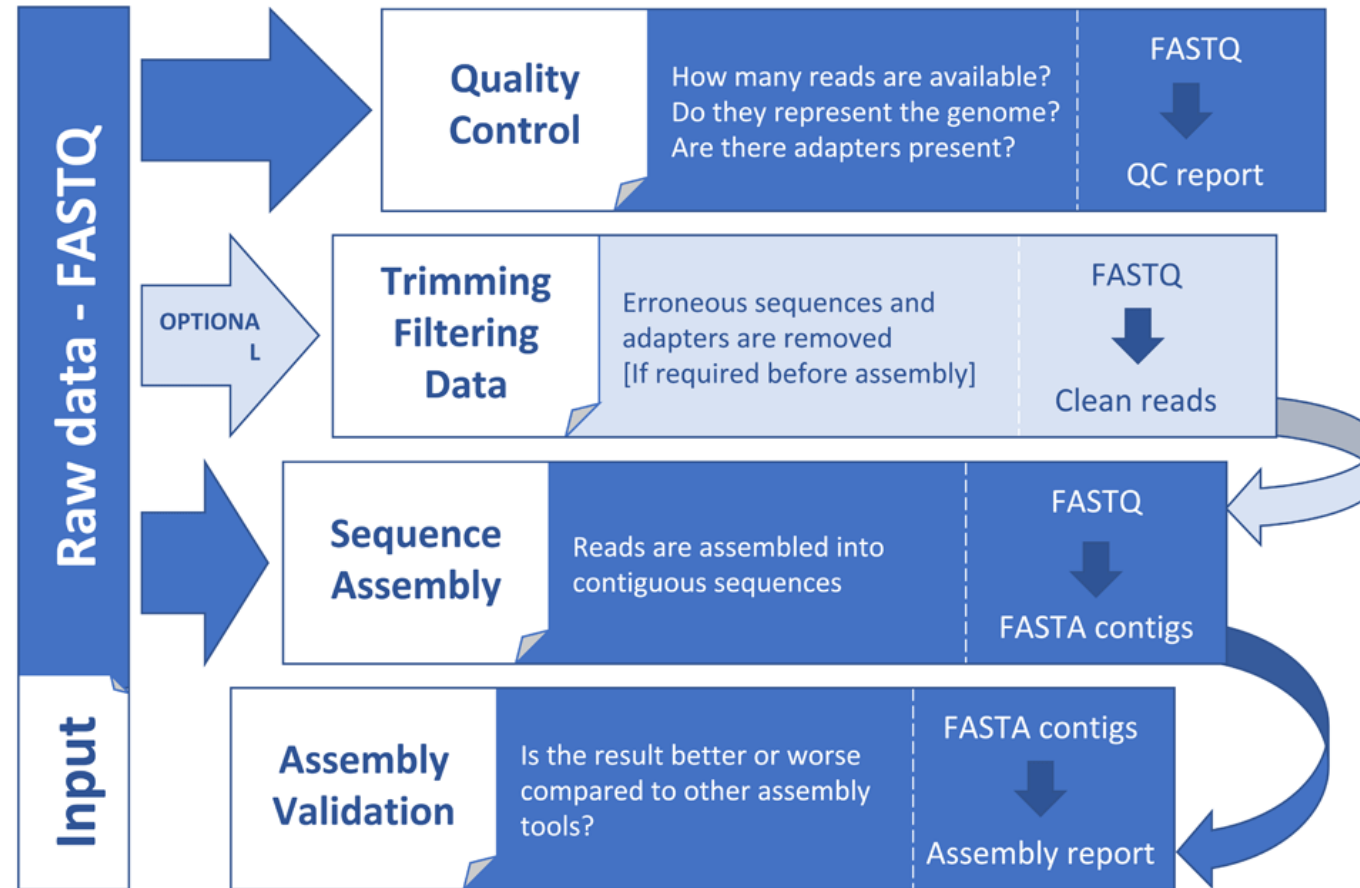
- Difficulty resolving repetitive sequences
- Gaps and fragmented genome assemblies

- Error rates
- Cost
- Computational resources

Hybrid Genome Assembly: Integrating Short- & Long-Read Sequencing



General steps in a genome assembly workflow



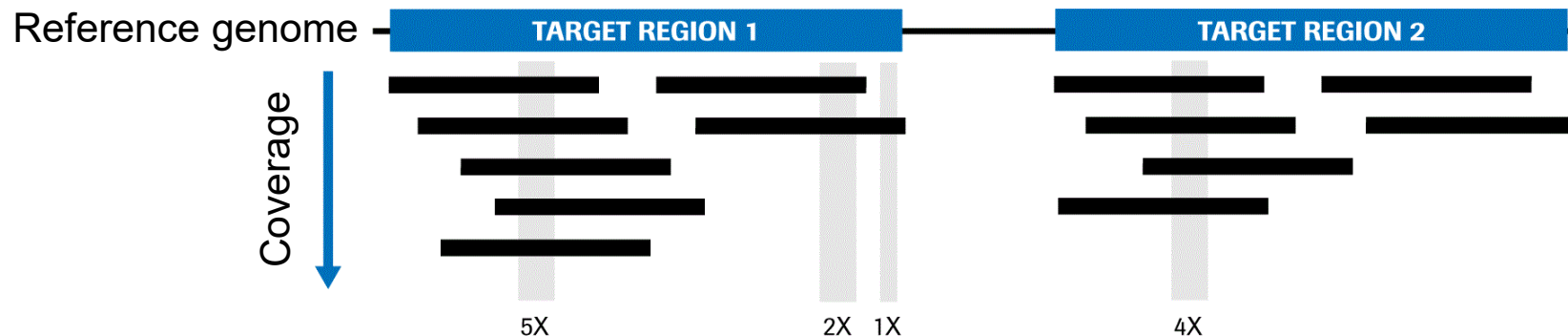
Coverage/Sequencing depth

Coverage/Sequencing depth: the average number of sequencing reads that align to each base position in the reference genome.

1) Overall genome coverage



2) Coverage of each target region



Example: Estimating Genome Coverage

$$\text{Converage} = \frac{\text{Total number of bases obtained from reads}}{\text{Reference genome size}}$$

CTAGGCCCTCAATTTT
CTCTAGGCCCTCAATTTT
GGCTCTAGGCCCTCATTTTT
CTCGGCTCTAGCCCCTCATTTT
TATCTCGACTCTAGGCCCTCA
TATCTCGACTCTAGGCC
TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCG
GGCGTCGATATCT
GGCGTCTATATCT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTT

177 nucleotides
35 nucleotides

Average coverage = $177 / 35 \approx 7x$

Question:

Assume that the *E. coli* genome sequencing data have the following characteristics:

- Total number of reads obtained: 1,000,000 reads
- Length of each read: 150 bp (paired-end sequencing)
- Genome size of *Escherichia coli* str. K-12: approximately 4.6 Mb

Calculate the average coverage of the *E. coli* genome.

3. Sequence File Formats



3.1. Example from Sanger sequences

- File format: **ab1** file
- A view of sequence from Sanger sequencing



FASTA Format

FASTA is a text format used to store biological sequences, such as DNA, RNA, or protein sequences.

FASTA Definition Line (begin with ">")

Nucleotide
sequence or
amino acid
sequence

[illegible]

3.2. Nanopore sequencing data

Raw data files:

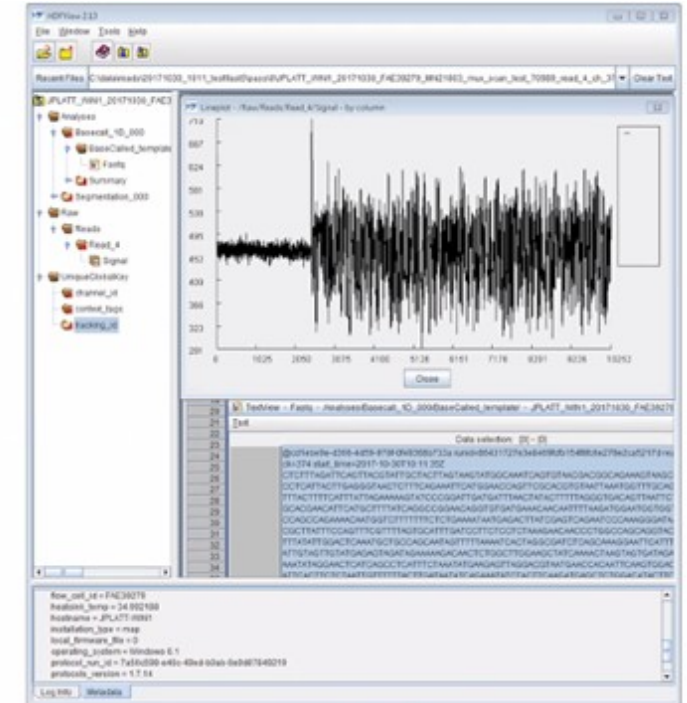
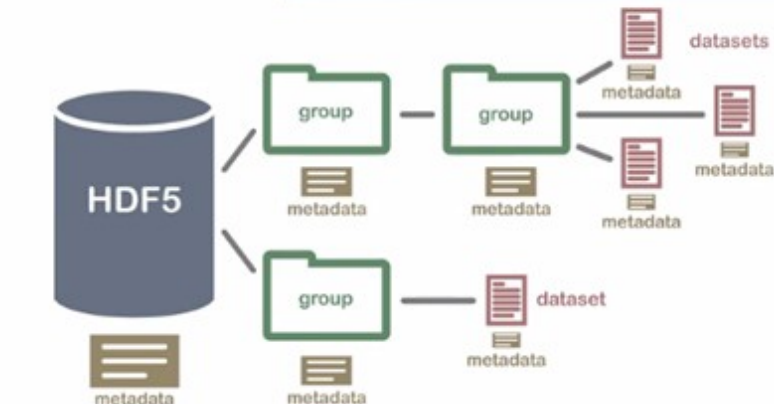
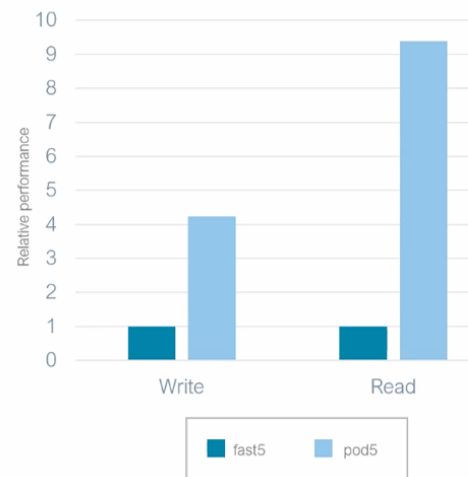
FAST5 (Hierarchical Data Format):

- A hierarchical file format for storing large datasets.
- Contains raw signals generated during sequencing.
- Superseded by POD5 starting with MinKNOW v23.04.05.

POD5 (Pore Open Data):

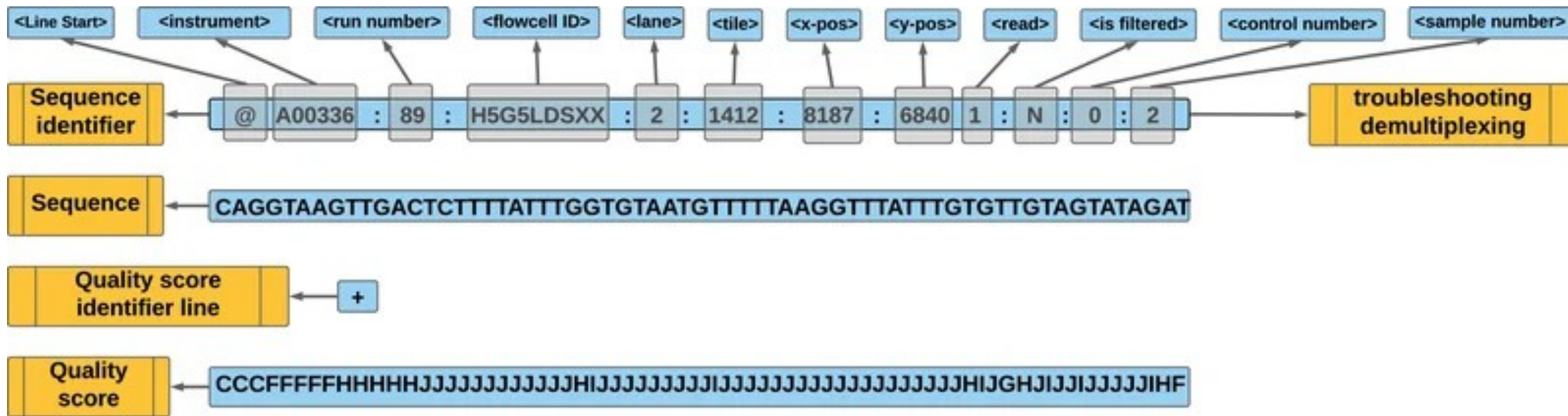
- Smaller file size than FAST5.
- Improved file read/write performance.

Base calling 
FASTQ files



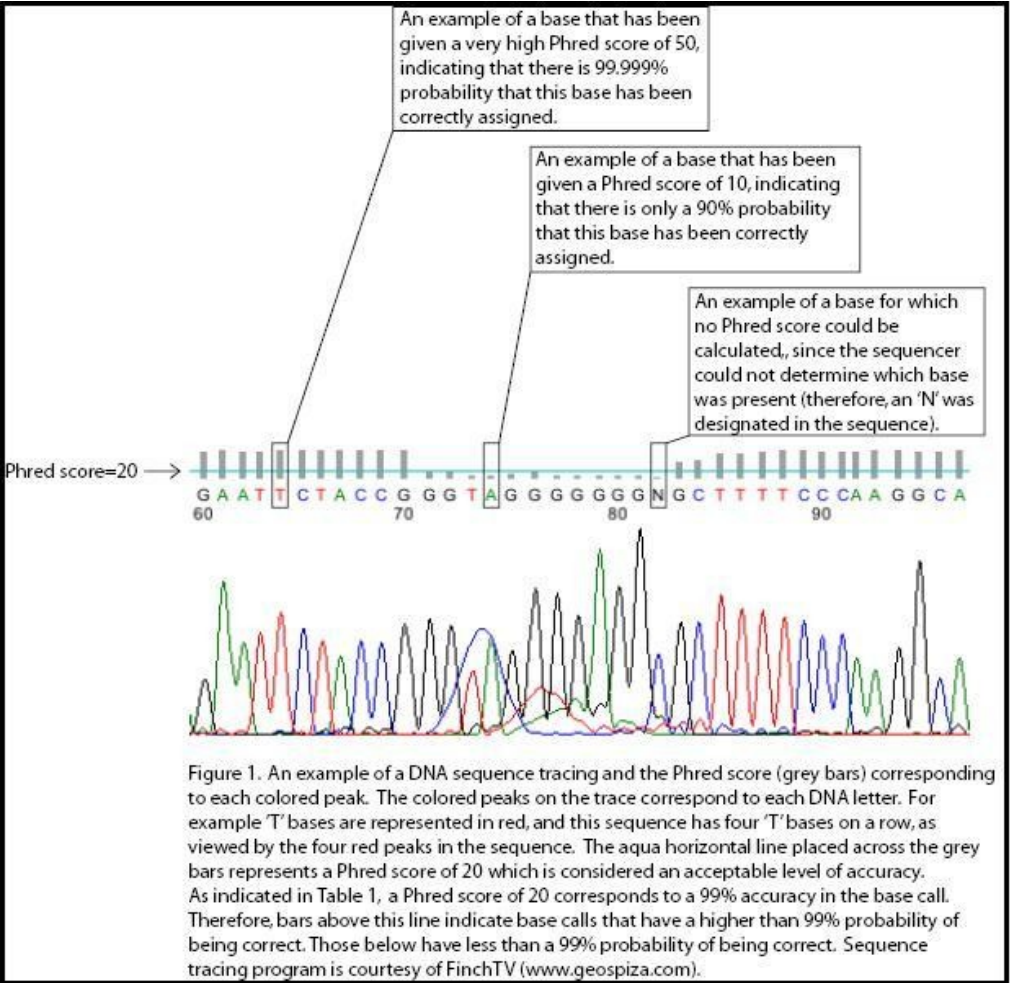
Sequence data file: FASTQ

- Contains nucleotide sequences and per-base quality scores.
- Each read consists of four lines: identifier, sequence, “+” separator, and quality scores.



(DOI:[10.1002/smt.202101251](https://doi.org/10.1002/smt.202101251))

How to read quality score?



Phred quality score

Phred quality scores Q are defined as a property which is logarithmically related to the base-calling error probabilities P

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

or

$$P = 10^{\frac{-Q}{10}}$$

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of 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 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

If Phred assigns a quality score of 30 to a base, the chances that this base is called incorrectly are 1 in 1000.

How to read quality score?

ASCII encoded Phred score

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 \$	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

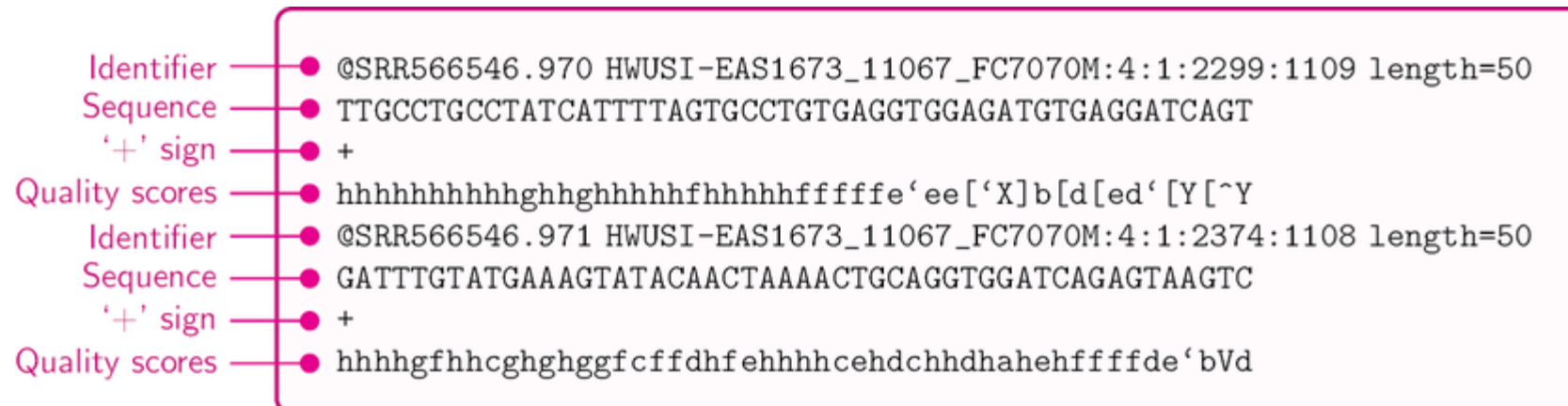
ASCII encoding table used to convert numerical quality scores into single characters

3.3. Sequence File Formats from Illumina sequencing

Each sample has a **pair of raw fastq files** (forward file & reverse file).

e.g. a fastq file

```
@A01056:177:HH2T7DSX2:3:1101:1090:1000 1:N:0:ATCCACTG+ACGCACCT
TTAGGGTGGATAGCTTAGACGATGTCGGTGTCAAGCGCGTCGACTAACCGCGCCGTTAAGGTAGGCTGGCGGGCGCGACACTGGAATGCAGATGGAGAATTCTCTGCAAATTAAAACCTTA
TATAGGAAGAGAACACGTCAGAAAAACAGT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:F,FFFFFFFFFFFFFFFF,
,,::,:::FFF,:F:FFFF,,:F,:F:,F
@A01056:177:HH2T7DSX2:3:1101:3694:1000 1:N:0:ATCCACTG+ACGCACCT
GTTATGCGAGGGGTATTGTCACGACTCCTTTCCGAGGTATCGTGGTGGTACTGAGGCACATGGTAACATGATGTGGGGTTGTGTCTTGTGGGTACAGTGGTACACCTCTGATCAGAGTTTA
ATCTATTCTGAATAGCCGTACCCGCGGTTAT
+
F,:FFFFFFFFF,F:FFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFF:FFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFF,:FFFFFFF:F:
```



4. Upload Data to UseGalaxy



Practical Data

Original Article

Hybrid sequencing of chromosome and plasmids from multidrug-resistant *Escherichia coli* isolated in Cambodia: Are megaplasms vectors of antibiotic resistance genes?

Ella Marcy^{a,b,*}, Sivhour Chiek^c, Mallorie Hidé^{a,b}, Savatey Hak^{b,d}, Chiva Ma^{b,d},
Meymey Lem^c, Gauthier Delvallez^d, Anne-Laure Bañuls^{a,b}, Sokleaph Cheng^{b,d,†},
Juliette Hayer^{a,b,†}

^a MIVEGEC, Montpellier University, IRD, CNRS, Centre IRD, Montpellier, France

^b LMI Drug Resistance in South East Asia (DRISA), Phnom Penh, Cambodia – Montpellier, France

^c Battambang Provincial Referral Hospital, Battambang, Cambodia

^d Laboratoire de Biologie Médicale, Institut Pasteur du Cambodge, Phnom Penh, Cambodia

Long- and short-read sequencing data from **6** clinical *Escherichia coli* isolates in Battambang, Cambodia.

Strain ID	Oxford Nanopore (ONT)	Illumina WGS (Paired-end reads)
327	ERR13363302	ERR13363635
368	ERR13363634	ERR13363306
372	ERR13363630	ERR13363307
366	ERR13363633	ERR13363304
364	ERR13363632	ERR13363303
323	ERR13363629	ERR13363281

(Link to sequencing data: <https://www.ebi.ac.uk/ena/browser/view/PRJEB77560>)

Option 1: Retrieve selected sequencing data from the SRA

- Create a new history and give it a name e.g. *WGS_Ecoli_327*
- Search tool: **Download and Extract Reads in FASTQ** format from NCBI SRA
- Enter SRR accession number e.g. *ERR13363302*, *ERR13363635* (for *E.coli* 327)

The screenshot displays the Galaxy Australia web interface. The top navigation bar includes the Galaxy logo, the text 'Australia', and a status bar showing 'Using 15% of 107.4 GB' and the user 'tam-tran'. The left sidebar contains navigation links: Activities, Upload, Tools (highlighted with an orange arrow), Workflows, Workflow Invocations, Interactive Tools, Visualization, Notifications, and More. The main panel shows the 'Tools' section with a search bar containing 'SRA'. The tool 'Download and Extract Reads in FASTQ format from NCBI SRA' is selected and highlighted with an orange arrow. The tool parameters are displayed: 'select input type' is set to 'SRR accession', 'Accession *' is 'ERR13363302, ERR13363635' (highlighted with an orange arrow), and 'Select output format *' has 'gzip compressed fastq' selected. A note explains that compression reduces space and is accepted by downstream applications. The 'Advanced Options' section is collapsed. The right sidebar shows the 'History' section with a search bar and a message: 'This history is empty. You can load your own data or get data from an external source.' An orange arrow points to the history name 'WGS_Ecoli_327'.

Option 2: Upload sequencing data files to Galaxy

- Create a new history and give it a name.
- Upload data: Choose **Upload** → **Choose local file** → **Start** → **Close**

The screenshot shows the Galaxy Australia web interface. A modal window titled "Upload from Disk or Web to MTB analysis" is open. The modal has tabs for "Regular", "Composite", "Collection", and "Rule-based". Below the tabs, it says "Please wait...5 out of 5 remaining...". The modal displays a list of files being uploaded:

File Name	Size	Type	Search	Reference	Progress	Actions
barcode01_concat.fa	119.6 MB	Auto-detect	Q	unspecified (?)	23%	⚙️ ⏸️
SRR26937092_1.fastq	108.4 MB	Auto-detect	Q	unspecified (?)	0%	⚙️ ⏸️
SRR26937092_2.fastq	110.4 MB	Auto-detect	Q	unspecified (?)	0%	⚙️ ⏸️
VH4-C3_S3_L001_R1.fastq	61.1 MB	Auto-detect	Q	unspecified (?)	0%	⚙️ ⏸️
VH4-C3_S3_L001_R2.fastq	61.5 MB	Auto-detect	Q	unspecified (?)	0%	⚙️ ⏸️

At the bottom of the modal, there are buttons for "New upload Beta", "Choose local file", "Choose from repository", "Paste/Fetch data", "Start", "Pause", "Reset", and "Close".

Three orange arrows with numbers 1, 2, and 3 point to the "Upload" button in the left sidebar, the "Choose local file" button, and the "Start" button respectively.

Create datasets for paired-end illumina data in Galaxy

Chose Illumina fastq files/Advanced Build List → List of Paired Datasets → Builder

The screenshot shows the Galaxy web interface for creating a paired-end dataset list. The interface is divided into three main sections: dataset selection, configuration, and building the list.

Dataset Selection: The left sidebar shows a list of datasets. Two datasets are selected: "3: SRR26937092_1.fastq.gz" and "2: SRR26937092_2.fastq.gz". A red arrow points to the "2 of 5 selected" button.

Configuration: The main panel shows the "Configure auto-pairing" step. It includes a progress bar with three steps: "What are you building?", "Auto Pairing", and "Builder". The "Auto Pairing" step is currently active. Below the progress bar, there are text boxes for filters: "_1" and "_2". The text below the filters explains that these are used to differentiate forward and reverse pairs. It also shows the "Auto-matched Pairs" section, which lists the selected datasets and their corresponding reverse reads. The "No Un-matched Datasets" section indicates that all datasets were successfully paired.

Building the List: The right panel shows the "Builder" step. It includes a progress bar with three steps: "What are you building?", "Auto Pairing", and "Builder". The "Builder" step is currently active. Below the progress bar, there is a section titled "Assemble, label, and sort your list of pairs." which explains that the interface allows building a new Galaxy list of pairs. It also shows a table of auto-matched pairs. At the bottom, there is a text box labeled "Dataset(s)" with the value "Illumina_WGS" entered. A red arrow points to this text box with the text "Điền tên set mẫu e.g. Illumina_WGS".

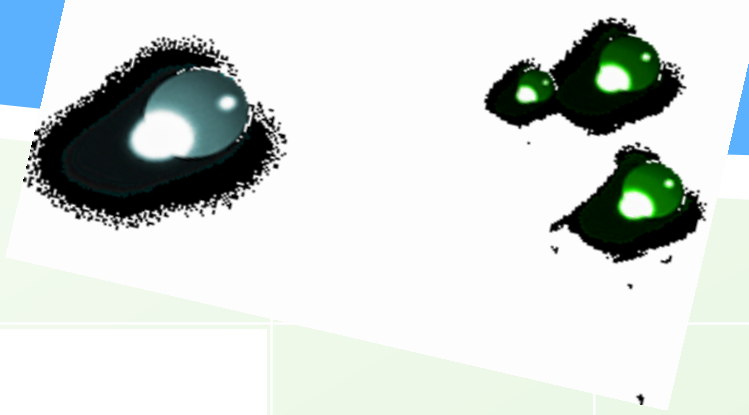
Table of Auto-matched Pairs:

Dataset(s)	List Identifier	Disc...	Status
FORWARD 2: SRR26937092_1.fastq.gz REVERSE 3: SRR26937092_2.fastq.gz	SRR26937092	X	✓

Table of Dataset(s) and List Identifier:

Dataset(s)	List Identifier
Illumina_WGS	

5. Quality check



5.1. Assess Illumina sequencing quality using FastQC

Tool: *FastQC*

Tool Parameters:

- **Raw read data from your current history:** Select **Dataset Collection**, then choose the datasets containing the Illumina sequencing data e.g. Paired-end data (fastq-dump) - ERR13363302
- Select **Run Tool**

The screenshot displays the Galaxy web interface for running the FastQC tool. The top navigation bar shows the user is logged in as 'tam-tran' and is using 16% of 107.4 GB of storage. The left sidebar lists various tool categories, with 'FastQC Read Quality reports' highlighted. The central panel shows the 'FastQC Read Quality reports (Galaxy Version 0.74+galaxy1)' tool configuration. The 'Tool Parameters' section includes three input fields: 'Raw read data from your current history', 'Contaminant list', and 'Adapter list'. The 'Raw read data' field is set to 'Dataset Collection' and shows a selection of '1: Paired-end data (fastq-dump)'. The 'Run Tool' button is visible in the top right of the tool panel. The right sidebar shows the history of datasets, including 'WGS_Ecoli_327' and '2: Single-end data (fastq-dump)'.

FastQC results for sequencing data quality assessment

 Started tool **FastQC** and successfully added 2 jobs to the queue.

It produces 4 outputs:

- **8: FastQC on dataset 3: Webpage**
- **9: FastQC on dataset 3: RawData**
- **10: FastQC on dataset 4: Webpage**
- **11: FastQC on dataset 4: RawData**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Here is a link to each job: [a6e389a98c2d1678fdf1944db3237c97](#) 
[a6e389a98c2d1678706115328e5c9c4f](#) 

search datasets  

WGS_Ecoli_327 

 Default Storage

 399 MB   4  7 

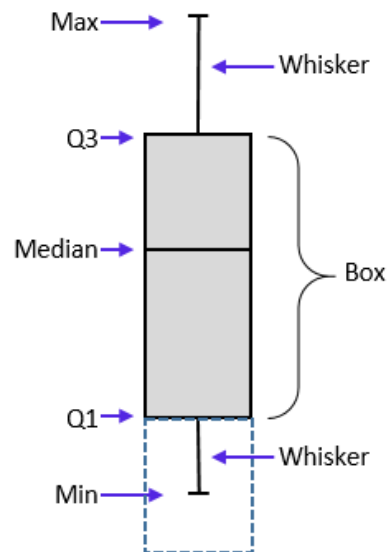
  

7: FastQC on collection 1: RawData  
ta
a list with 1 **txt** pair

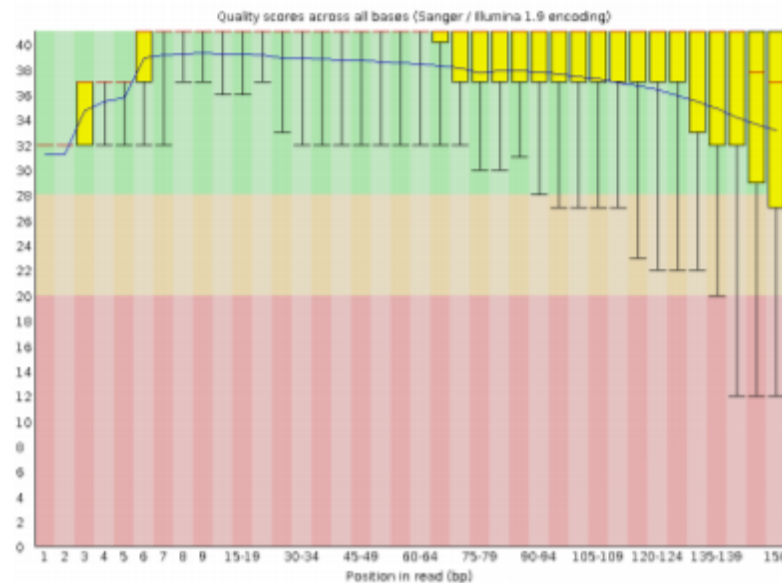
 **6: FastQC on collection 1: Webpage**  
ge
a list with 1 **html** pair

Read the results from the FastQC HTML report (webpage)

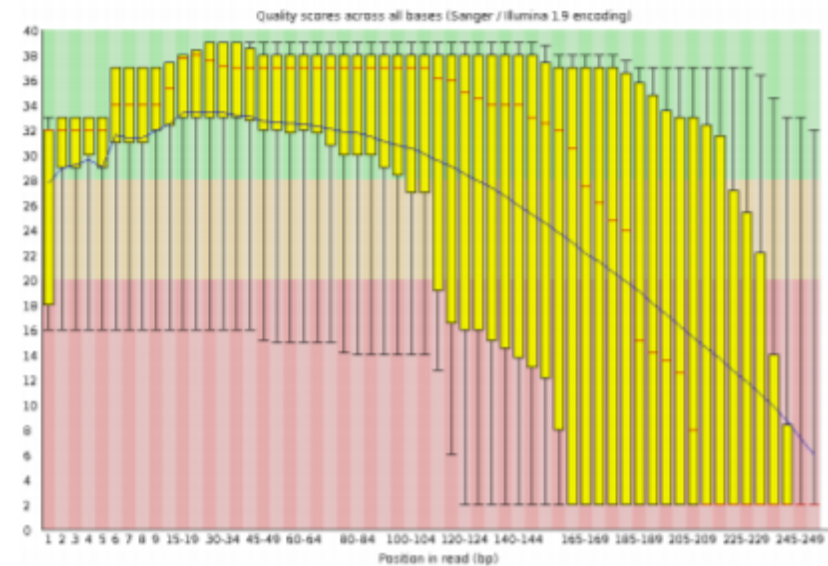
FastQC Report - Examples



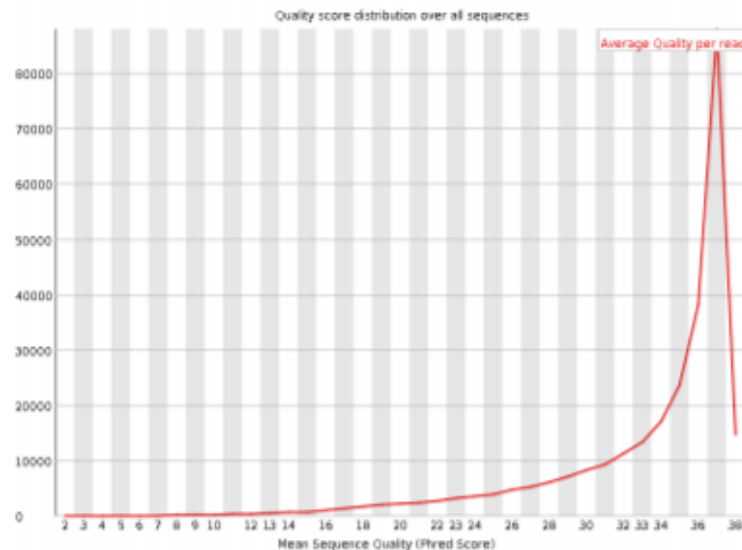
A good per base quality graph



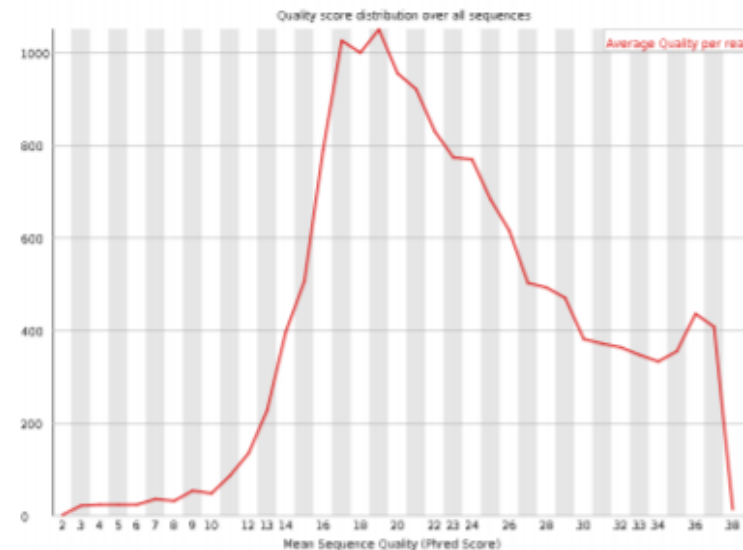
A bad per base quality graph



A good per sequence quality graph

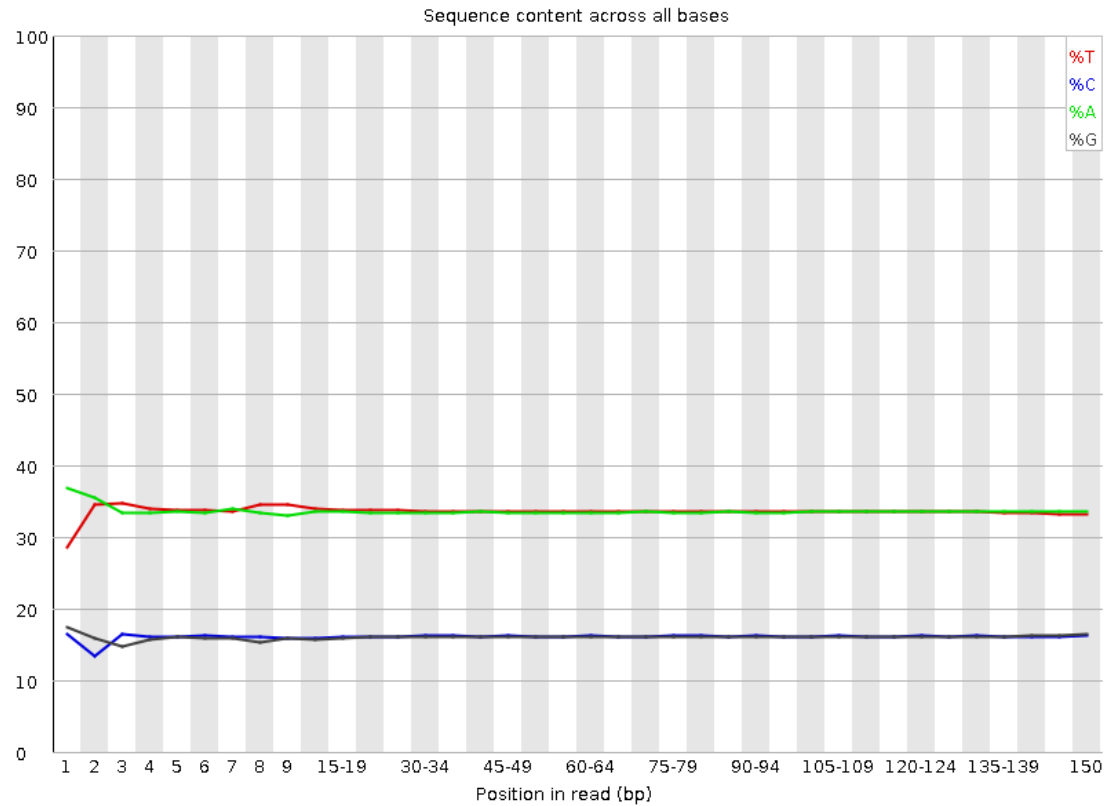


A bad per sequence quality graph

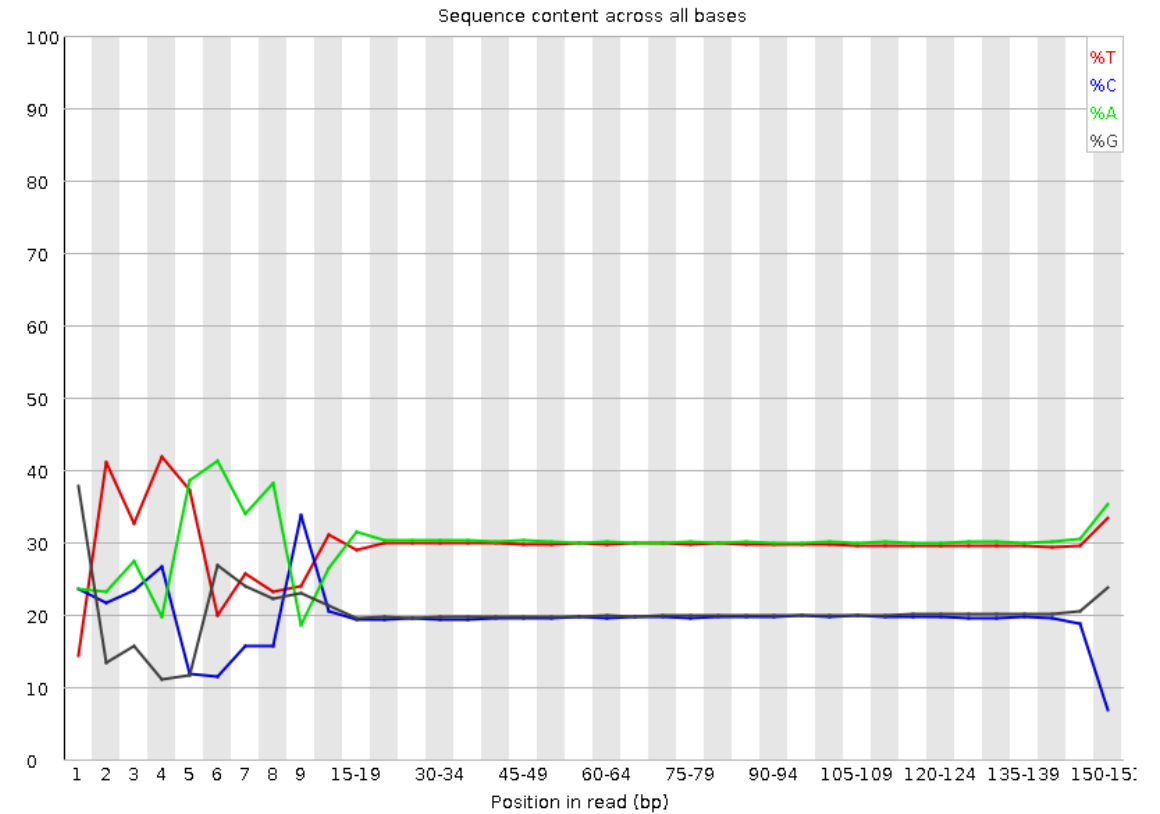


FastQC Report - Examples

✔ Per base sequence content



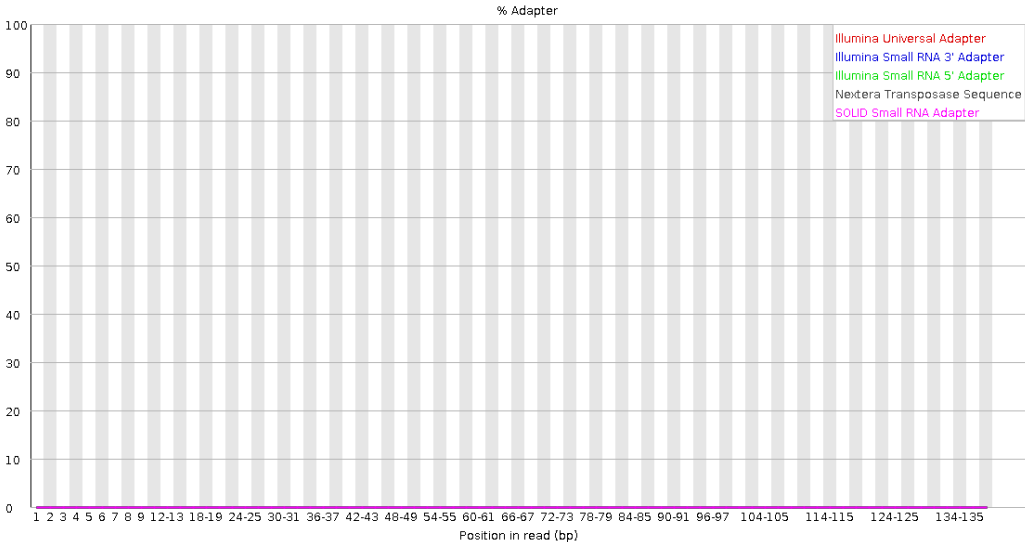
✘ Per base sequence content



FastQC Report - Examples

✔ Overrepresented sequences
No overrepresented sequences

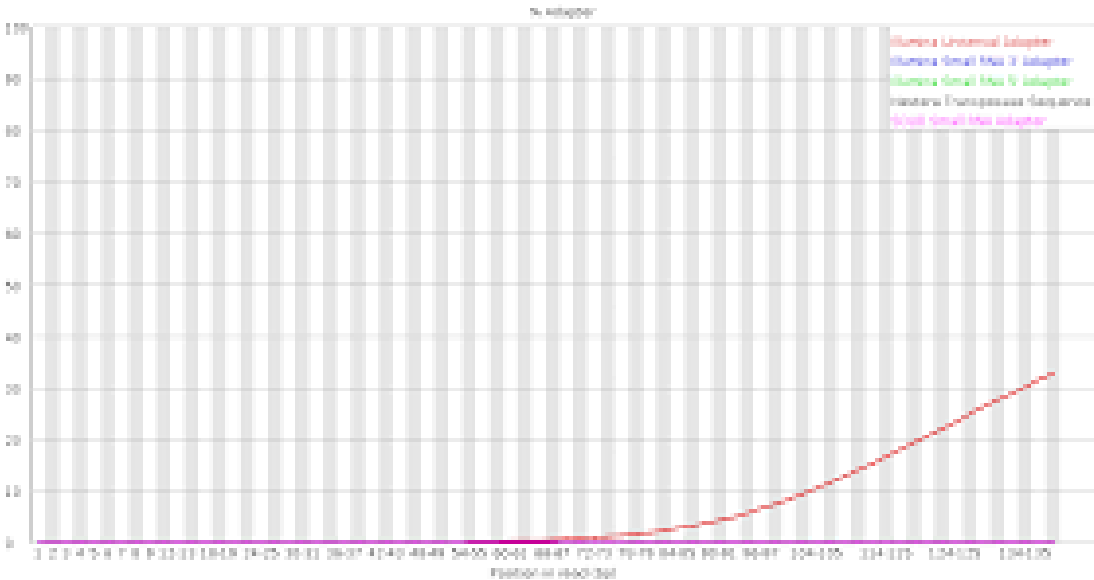
✔ Adapter Content



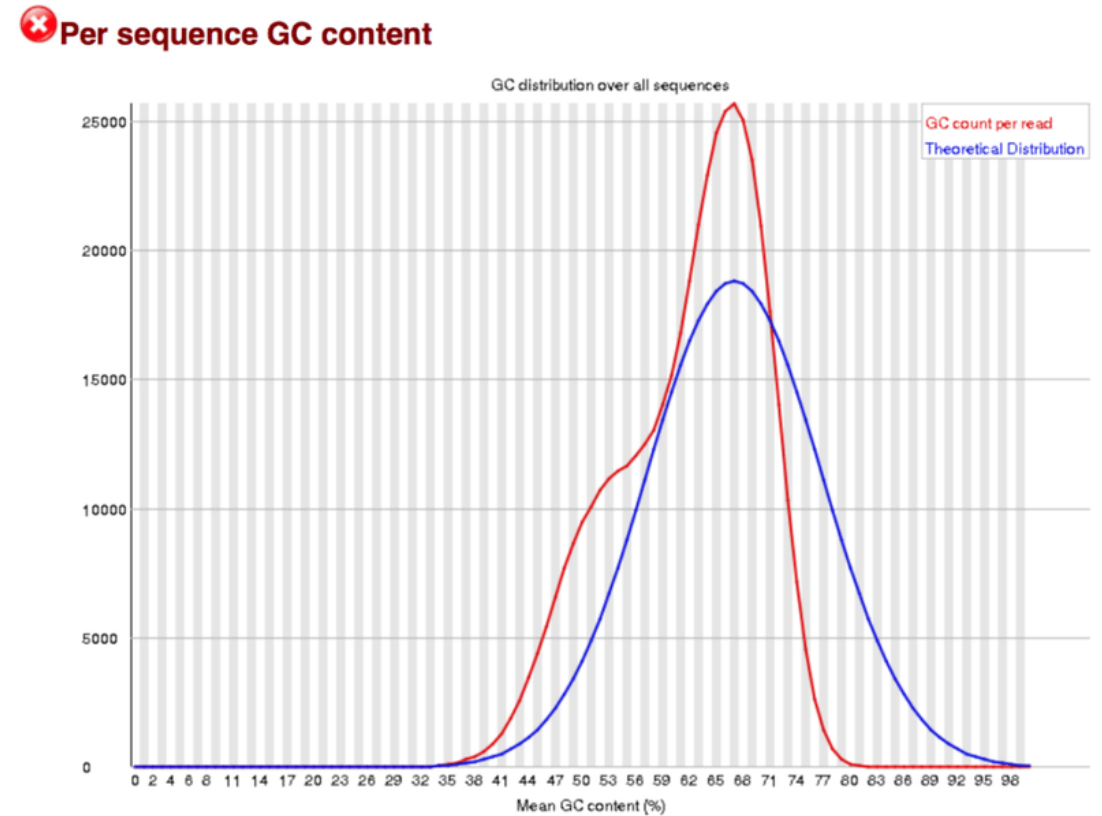
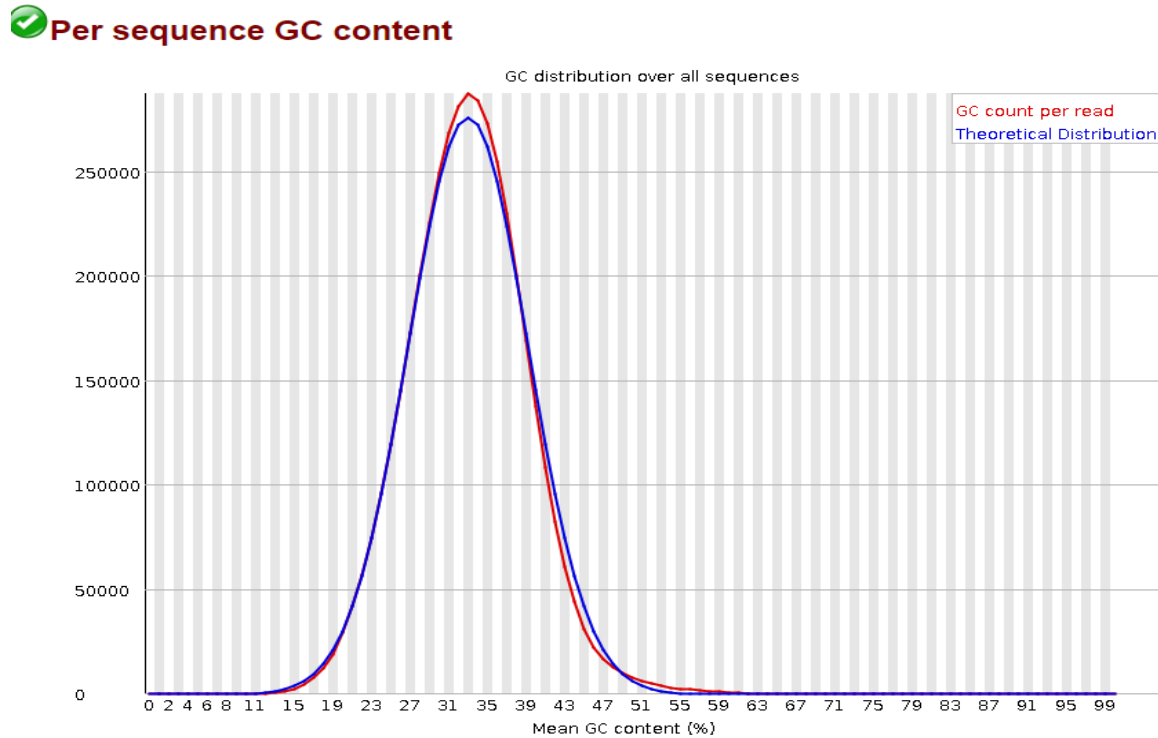
! Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTGCTATGGCCACCAGACTCTCAGGCTCCATGCAGTGGCCAGCCTCATCG	2554	0.8349133703824779	No Hit
CAGCGGTCTAGTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAG	2463	0.8051650866296176	No Hit
GTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTCGATC	1920	0.6276560967636483	No Hit
CCACAGGGTCCCAGGTTCATGGGTACCGAGTCCAGGTTCATAGTCCGGATG	1219	0.39849624060150374	No Hit
GAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTCGATCTTCA	1186	0.3877084014383786	No Hit
GGCAGGTGGACCCGAGCCGTGACAGAGGAGGTGAGCCCTGAGTTGGA	1111	0.3631905851585486	No Hit
CACAGGGTCCCAGGTTCATGGGTACCGAGTCCAGGTTCATAGTCCGGATG	1079	0.35272965021248776	No Hit
GTGCTATGGCCACCAGACTCTCAGGCTCCATGCAGTGGCCAGCCTCATCG	1036	0.3386727688787185	No Hit

! Adapter Content



FastQC Report



GC content of *Escherichia coli*: ~50% - 51%

Question: Assessing the quality of Illumina sequencing data

Based on the FastQC results for Read 1 and Read 2 of the Illumina sequencing dataset for *Escherichia coli* 327, evaluate and report the sequencing data quality according to the following criteria:

- 1. Raw data statistics:** Total number of reads, read length, and GC content (%).
- 2. Read quality assessment:** Analyze the distribution of quality scores across read positions (*Per base sequence quality*) and comment on the overall data quality.
- 3. Assessment of other quality metrics:** Evaluate *Per sequence GC content*, *Sequence Length Distribution*, *Overrepresented sequences*, and *Adapter Content*, among other relevant indicators.
- 4. Comparison of Read 1 and Read 2:** Identify similarities and differences in quality between the two reads, and determine which read has better overall quality.

5.2. Assessing ONT sequencing quality using NanoPlot

- Tool: **NanoPlot**
- Select **Batch** (if processing one file)/ **Type of file(s) to work on: fastq/ files:** select the uploaded ONT sequencing file.
- Under **Specify the bivariate format of the plots**, select **dots (dot)**→ **Run Tool**

The screenshot displays the NanoPlot tool interface in the Galaxy environment. On the left, the 'Tools' sidebar shows 'NanoPlot' selected. The main panel is divided into sections for tool parameters and options.

Tool Parameters

- Select multifile mode:** A dropdown menu is set to 'batch'. Red text annotations explain: 'Batch: Process each file separately.' and 'Combine: Process multiple input files together'.
- Type of the file(s) to work on:** A dropdown menu is set to 'fastq'.
- files *:** A file selection area shows '2: Single-end data (fastq-dump)' selected. An arrow points to the folder icon in the selection area.

Options for customizing the plots created

- Specify a color for the plots. - optional:** A dropdown menu shows 'Nothing selected'.
- Specify the bivariate format of the plots. - optional:** A section with checkboxes. 'dots (dot)' is selected, indicated by a red checkmark and an arrow.
- Show the N50 mark in the read length histogram.:** A toggle switch is set to 'No'.

At the bottom, a warning message states: 'The supplied input will be mapped over this tool. (--fastq)'. A 'Run Tool' button is visible in the top right corner of the tool panel.

NanoPlot report results

WGS_Ecoli_327

Default Storage

400 MB 7 10

14: NanoPlot on collection 2: Nan
oStats post filtering
a list with 1 tabular dataset

13: NanoPlot on collection 2: Nan
oStats
a list with 1 tabular dataset

12: NanoPlot on collection 2: HT
ML report
a list with 1 html dataset

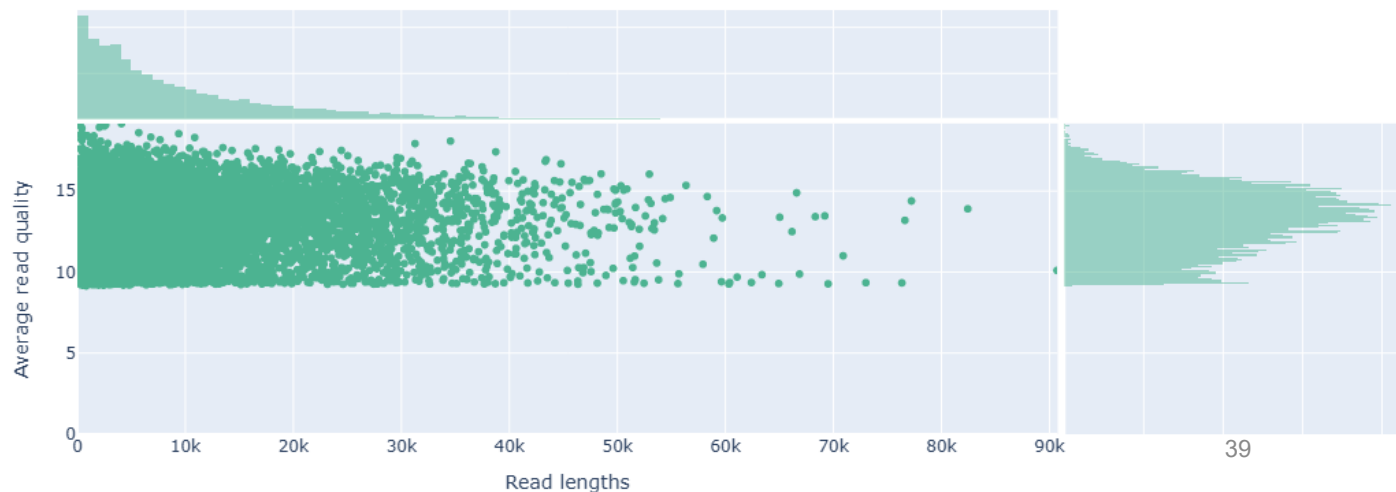
View the results file:
HTML report



Summary statistics

	dataset
Metrics	
number_of_reads	35963
number_of_bases	350203044.0
median_read_length	6309.0
mean_read_length	9737.9
read_length_stdev	10007.5
n50	17401.0
mean_qual	12.7
median_qual	13.3
longest_read_(with_Q):1	96266 (9.3)
longest_read_(with_Q):2	94650 (9.5)
longest_read_(with_Q):3	92830 (14.9)
longest_read_(with_Q):4	90756 (10.1)
longest_read_(with_Q):5	82484 (13.9)
highest_Q_read_(with_length):1	20.2 (6633)
highest_Q_read_(with_length):2	20.1 (2911)
highest_Q_read_(with_length):3	19.9 (14460)
highest_Q_read_(with_length):4	19.6 (1565)
highest_Q_read_(with_length):5	19.6 (489)
Reads >Q10:	33338 (92.7%) 322.1Mb
Reads >Q15:	6953 (19.3%) 66.1Mb
Reads >Q20:	2 (0.0%) 0.0Mb
Reads >Q25:	0 (0.0%) 0.0Mb
Reads >Q30:	0 (0.0%) 0.0Mb

Read lengths vs Average read quality dot plot (downsampled)



Question:

Based on the NanoPlot results for the ONT sequencing dataset of *Escherichia coli* 327:

1. Sequencing quality assessment: Is the overall data quality good or poor? Evaluate read length, average quality score, and quality variation along each read.
2. Are the reads long enough and of high enough quality for our downstream analyses? Is quality filtering or trimming necessary to improve the results?

6. Quality filtering (*Option*)

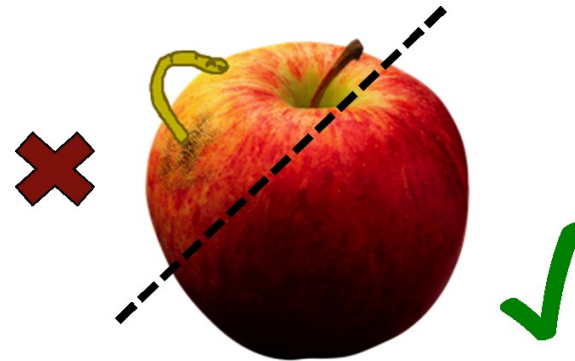


Quality filtering using fastp (*Option*)

Tool: **fastp**- fast all-in-one preprocessing for FASTQ files

It can trim adapter sequences and filter out low-quality reads.

```
@SRR638895.6046 6046 length=76
GTGAAAGACTCTCGTAGCAAACGAAACGTCAAGTCGGTGAGGCCAACTCTTGTCGTAGCCGCGTCCATTGCGCCCT
+SRR638895.6046 6046 length=76
#####EEEEEEEEEEEEEEEEEEEEEEEEEEEE#####
                2                36                2
```



Remove low-quality nucleotides (typically those with a quality score below 20).

Run *fastp* on paired-end Illumina sequencing data in Galaxy

- Tool: **fastp**
- Tool Parameters:
 - ✓ *Single-end or paired-end reads: Paired Collection*
 - ✓ *Select FASTQ dataset collection with R1/R2 pair: Select the raw data from the dataset collection (e.g. Paired-end data (fastq-dump) - ERR13363302).*
 - ✓ *Qualified quality phred: 20*
 - ✓ *Unqualified percent limit: 20*
 - ✓ *Length filtering options*
-> Length required: e.g 50bp
- Click **Run Tool**

Filter Options

Quality filtering options

Disable quality filtering
☐ No
Quality filtering is enabled by default. If this option is specified, quality filtering is disabled. (-Q)

Qualified quality phred - optional
20
Trim bases with a Phred quality score ≤ 20
The quality value that a base is qualified. Default 15 means phred quality $\geq Q15$ is qualified. (-q)

Unqualified percent limit - optional
20
At least 80% of bases must meet the Q20 quality threshold.
How many percents of bases are allowed to be unqualified (0~100). Default 40 means 40%. (-u)

Length filtering options

Disable length filtering
☐ No
Length filtering is enabled by default. If this option is specified, length filtering is disabled. (-L)

Length required - optional
50
Retain only reads that meet the minimum length requirement (e.g., 50 bp)
Reads shorter than this value will be discarded. Default is 15. (-l)

fastp fast all-in-one preprocessing for FASTQ files
(Galaxy Version 1.3.6+galaxy0)

Tool Parameters

Single-end or paired reads

Paired Collection

Select paired collection(s) *

8: Illumina_WGS

accepted formats

Merge forward and reverse reads?

No

(--merge)

Adapter Trimming Options If adapter sequences remain.

Disable adapter trimming
☐ No
Adapter trimming is enabled by default. If this option is specified, adapter trimming is disabled. (-A)

Adapter sequence for input 1 - optional

The adapter for read1. For SE data, if not specified, the adapter will be auto-detected. For PE data, this is used if R1/R2 are found not overlapped. (--adapter_sequence1)

Adapter sequence for input 2 - optional

The adapter for read2. For SE data, if not specified, the adapter will be auto-detected. For PE data, this is used if R1/R2 are found not overlapped. (--adapter_sequence2)

Adapter sequence auto-detection for paired-end
☐ No
(--detect_adapter_for_pe)

fastp report results

View the results file: HTML report
ERR13363302.fastqsanger.gz

Summary

General

fastp version:	1.4.0 (https://github.com/OpenGene/fastp)
sequencing:	paired end (150 cycles + 150 cycles)
mean length before filtering:	150bp, 150bp
mean length after filtering:	148bp, 148bp
Insert size peak:	268

Before filtering

total reads:	848.304000 K
total bases:	127.245600 M
Q20 bases:	123.913718 M (97.381535%)
Q30 bases:	118.210173 M (92.899222%)
Q40 bases:	0 (0.000000%)
GC content:	51.068785%

After filtering

total reads:	809.204000 K
total bases:	120.518872 M
Q20 bases:	118.444667 M (98.278938%)
Q30 bases:	113.610981 M (94.268208%)
Q40 bases:	0 (0.000000%)
GC content:	51.097392%

Filtering result

reads passed filters:	809.204000 K (95.390803%)
reads with low quality:	37.510000 K (4.421764%)
reads with too many N:	84 (0.009902%)
reads too short:	1.506000 K (0.177531%)
reads with adapter dimer:	0 (0.000000%)

✓ Started tool **fastp** and successfully added 1 job to the queue.

It produces 4 outputs:

- 21: fastp
- 22: fastp
- 23: fastp on : HTML report
- 24: fastp on : JSON report

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Here is a link to the job: **a6e389a98c2d1678ffe3106dcd153991**

Tool recommendation

You have used fastp tool. For further analysis, you could try using the following/recommended tools. The recommended tools are shown in the decreasing order of their scores predicted using machine learning analysis on workflows. Therefore, tools at the top may be more useful than the ones at the bottom. Please click on one of the following/recommended tools to open its definition.

Map with BWA-MEM

> History

search datasets

WGS_Ecoli_327

Default Storage

459 MB

10

14

20: fastp on collection 1: JSON report

a list with 1 json dataset

19: fastp on collection 1: HTML report

a list with 1 html dataset

18: fastp on collection 1: Paired-end output

a list with 1 fastqsanger.gz pair

Question: Quality-based trimming using fastp

Use fastp to clean short-read whole-genome sequencing data for sample SRR26937092. Apply the following criteria:

- Remove bases with Phred quality scores ≤ 20 , retaining only reads in which at least 80% of bases meet the Q20 quality threshold.
- Retain reads with a length of > 50 bp after trimming.
- Rerun FastQC for both Read 1 and Read 2 after trimming, and compare the FastQC results before and after quality filtering.

Question:

Report the quality metrics before and after trimming for both Read 1 and Read 2:

- Basic Statistics: number of sequences (reads), sequence length, and GC content (%GC).
- Percentage (%) of reads retained after trimming.
- Evaluate the changes in Per Base Sequence Quality.

Summary



1. Sequencing data: Nanopore and Illumina

- File format: FASTQ (sequence reads and quality scores)

2. Sequencing quality assessment

- *FastQC*: Quality assessment of short reads (Illumina)
- *Nanoplot*: Quality assessment of long reads (Nanopore)

3. Quality filtering

- *fastp*: trim adapter sequences and filter out low-quality reads.

