MODULE 5

SEQUENCE QC AND GENOME ASSEMBLY

PRESENTED BY: Praissy Zefi J
Bioinformatician , DTU
pzeje@dtu.dk

QUALITY CONTROL (QC)

- Post sequencing QC ensures integrity and quality of data before downstream analyses and interpretation
- Varies based on sequencing technology → sometimes requires additional steps
- Advancement in technology → variety of freely available and downloadable tools
- Illumina sequencing → FastQC
- ONT sequencing → Nanoplot, FastQC

ILLUMINA SEQUENCING

- Changes in fluorescence intensities
- FASTQ files
- Short read technology

ONT SEQUENCING

- Changes in electric signals
- Pod5 files / fast5files + fastq files
- Long-read technology

bioconda / packages / nanoplot 1.44.1

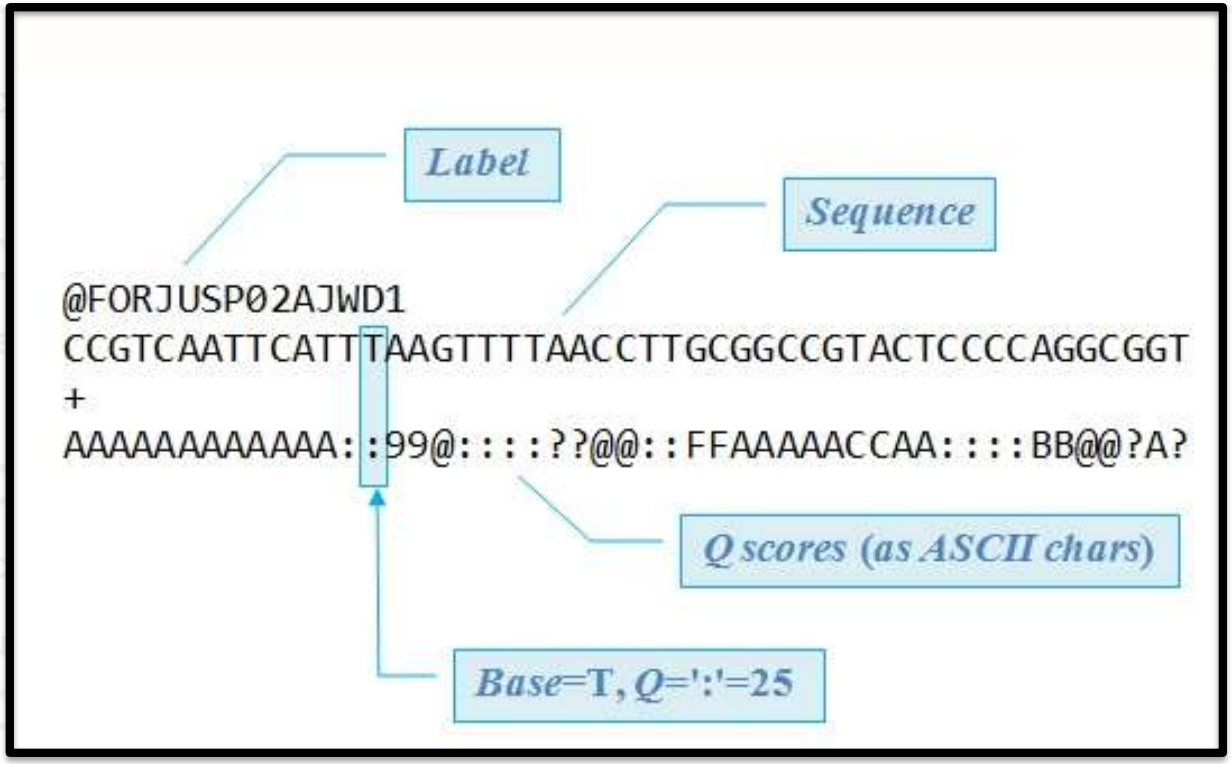
Porechop

Filtlong

QUALITY METRICS

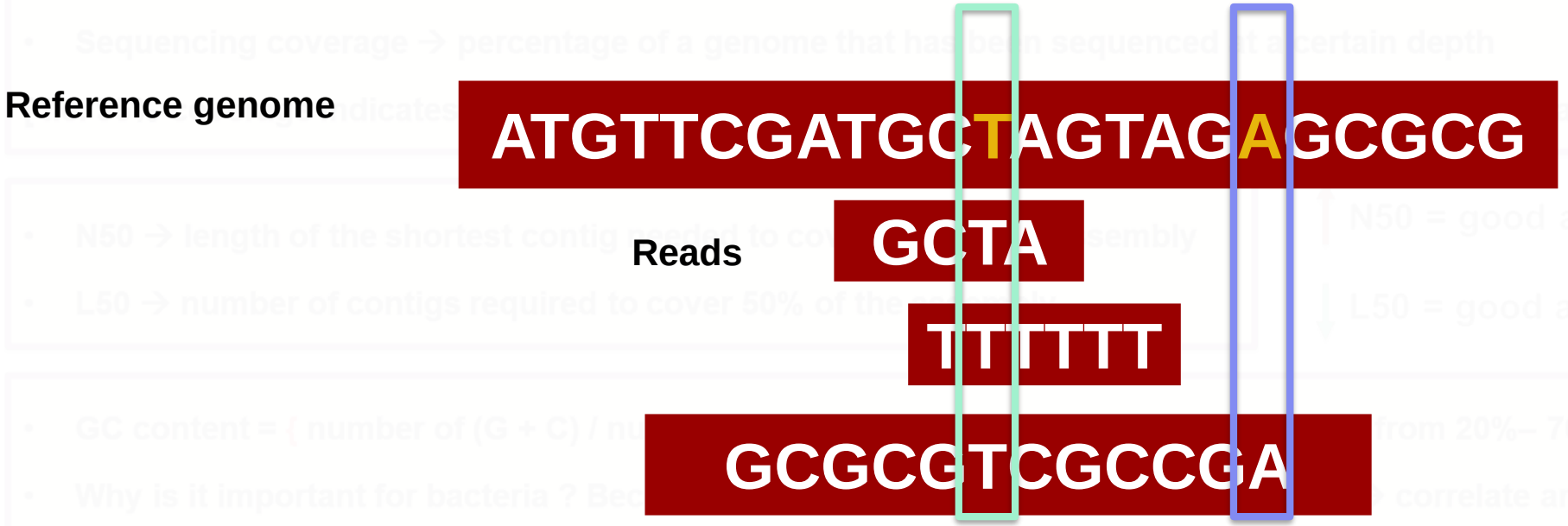
- Q scores → bottom line of fastQ files → assess accuracy and reliability of sequencing data

$Q = -10 \log_{10}(P)$; P → probability that a base is called incorrectly

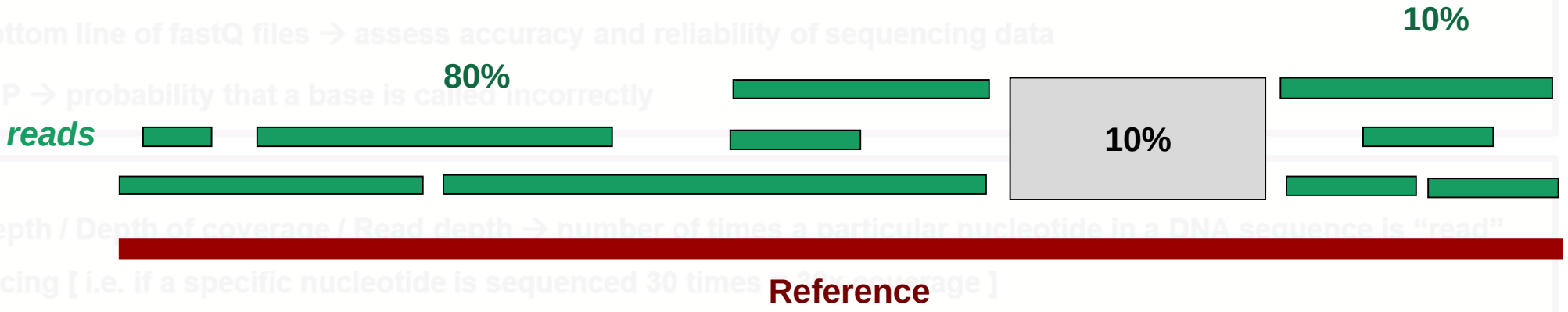


QUALITY METRICS

- **Sequencing depth / Depth of coverage / Read depth → number of times a particular nucleotide in a DNA sequence is "read" during sequencing [i.e. if a specific nucleotide is sequenced 30 times = 30x coverage]**



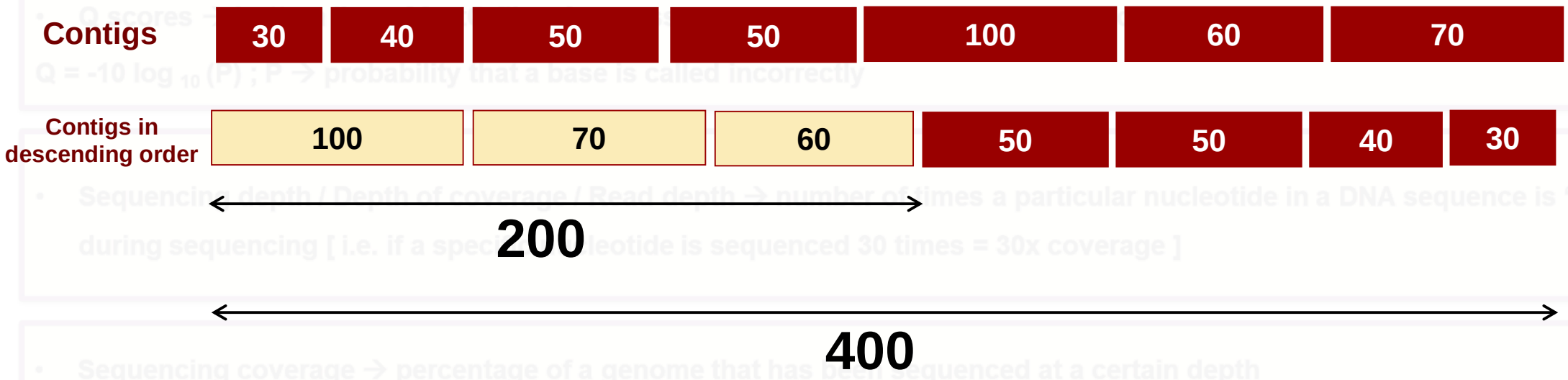
QUALITY METRICS



- **Sequencing coverage → percentage of a genome that has been sequenced at a certain depth**

[i.e. 90% coverage indicates that 90% of the bases in the total DNA sequence have been sequenced atleast once]

QUALITY METRICS – N50

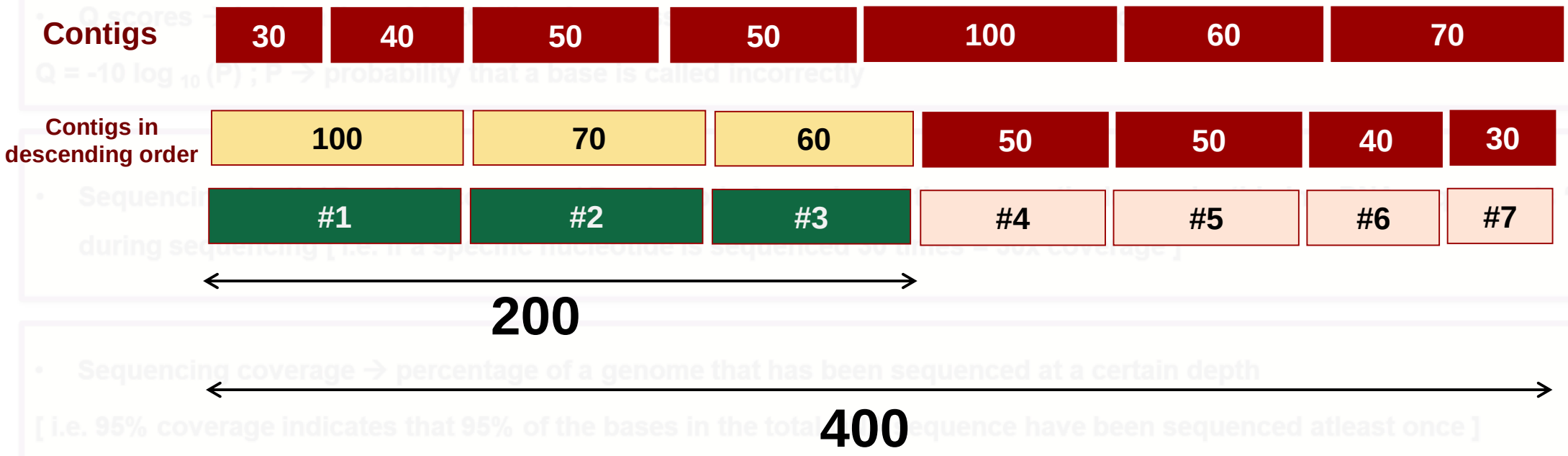


- N50 → length of the shortest contig needed to cover 50% of the assembly
- L50 → number of contigs required to cover 50% of the assembly

N50 = good assembly

L50 = good assembly

QUALITY METRICS – L50

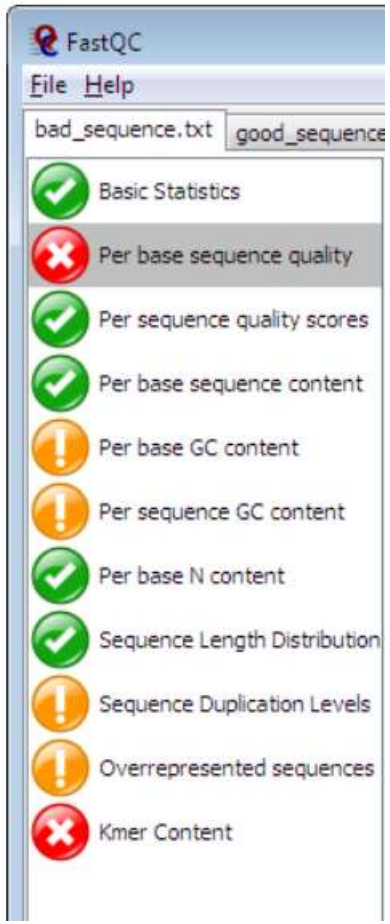


- N50 → length of the shortest contig needed to cover 50% of the assembly
- L50 → number of contigs required to cover 50% of the assembly

↑ N50 = good assembly

↓ L50 = good assembly

FASTQC



FastQC aims to provide a QC report which can spot problems which originate either in the sequencer or in the starting library material.

PASS

WARNING

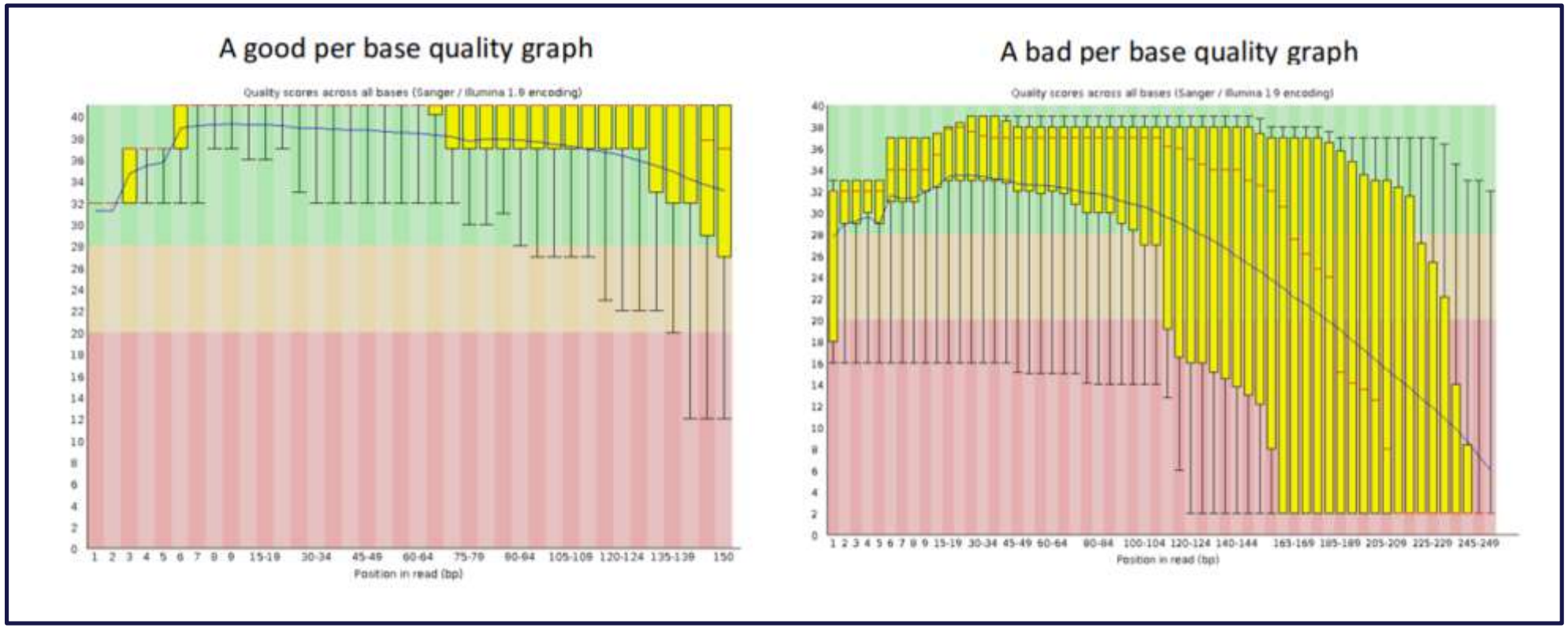
FAIL

Basic Statistics

Measure	Value
Filename	Mov10_oe_1.subset.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	305900
Sequences flagged as poor quality	0
Sequence length	100
%GC	47

FASTQC

1)

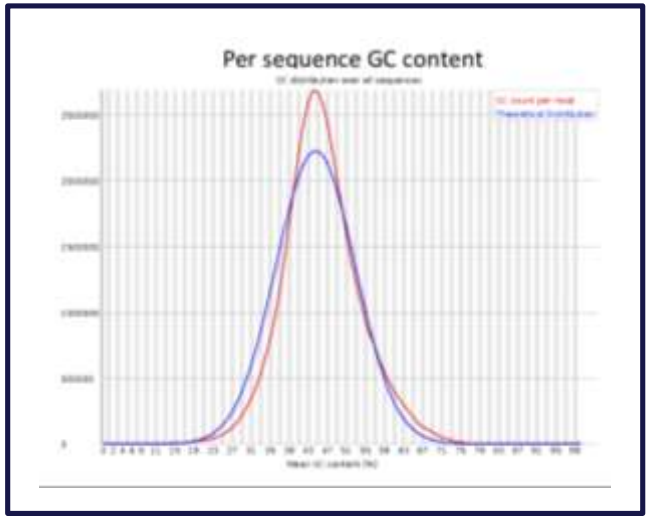



```
@SEQ_ID_1
GATAAAGCAGTATCGAT
+
!''*%%%).1***+*''))**
@SEQ_ID_2
TTTGGGGTTCAAAT
+
%%%).1***CCF>>>>>>CC
@SEQ_ID_3
GATCAAAGCAGTATCGAT
+
!''*(1***+*''))**55CCF>>>>>>
```

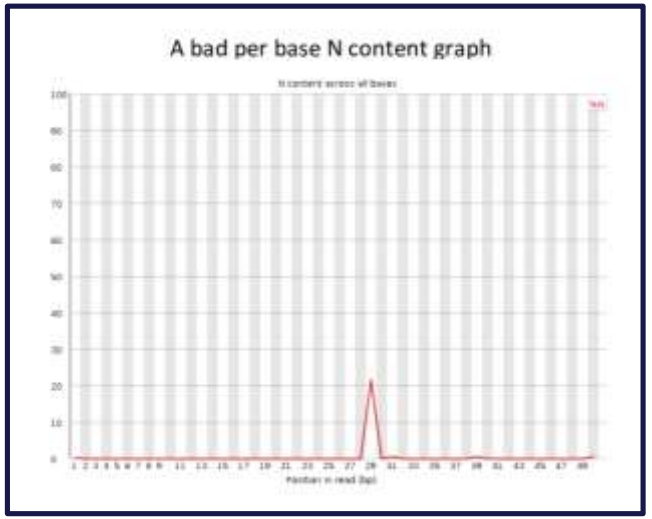
1) Analyses quality score at each position within a sequence to identify potential degradation or bias

- X-axis → base position in read ; Y-axis → quality scores

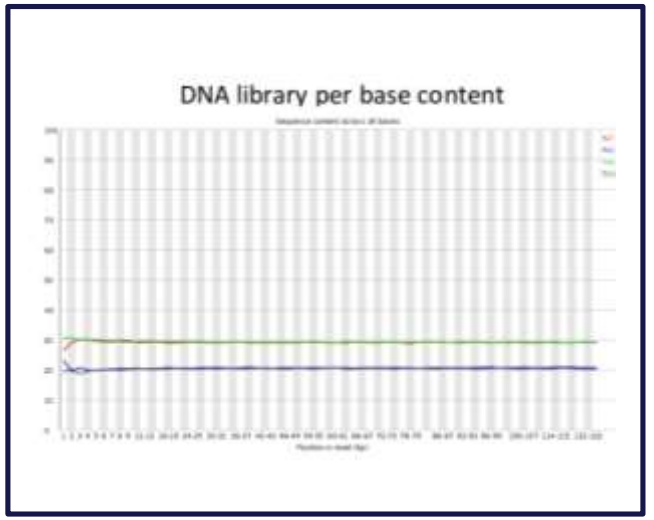
2)



3)



4)



2) Plot of the number of reads vs. GC% per read.

3) Percent of bases at each position or bin with no base call, i.e. ‘N’.

4) Plot of the number of reads vs. GC% per read. The displayed Theoretical Distribution assumes a uniform GC content for all reads

TRIMMING AND REMOVING ADAPTERS

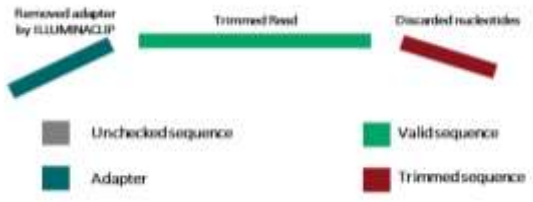
trimmomatic

Command line tool → trim and crop FASTQ data

Single end data: 1 INPUT FILE → 1 OUTPUT FILE

Paired end data: 2 INPUT FILES → 4 OUTPUT FILES ; forward paired, forward unpaired, reverse paired , reverse unpaired

Trimmomatic Manual



COMMAND	action
ILLUMINACLIP	Cut adapter and other illumina-specific sequences from the read
SLIDINGWINDOW	starts scanning at the 5" end and clips the read once the average quality within the window falls below a threshold
LEADING	Cut bases off the start of a read, if below a threshold quality
TRAILING	Cut bases off the end of a read, if below a threshold quality
MINLEN	Drop the read if it is below a specified length

cutadapt

Finds and removes adapter sequences, primers, poly-A tails

Allowed to modify and filter reads

ERROR TOLERANCE

Allowed errors : matches, mismatches, deletions, insertions

--no-indels → only mismatches detected

Maximum error rate = 0.1 (10%) default

Actual error rate = number of errors in the match / length of matching part of adapter

Adapter occurrence is found only if actual error rate of the match does NOT exceed maximum error rate

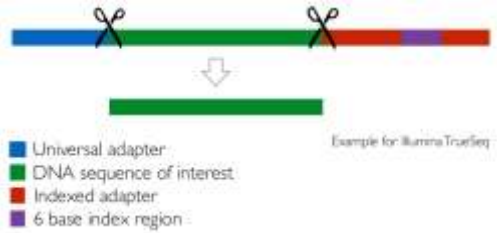
Eg 1: adapter match of length 8 containing 1 error →

error rate = 1 / 8 = 0.125

Default maximum error rate = 0.1

Will this be detected or not?

- e → allows to change maximum error rate between 0 - 1



ONT SEQUENCING - NANOPLOT

Plotting tool for long read sequencing data and alignments

Also available as a web service → submit the summary .txt file

INSTALLATION: **`pip install NanoPlot`**

Upgrade to a newer version using:

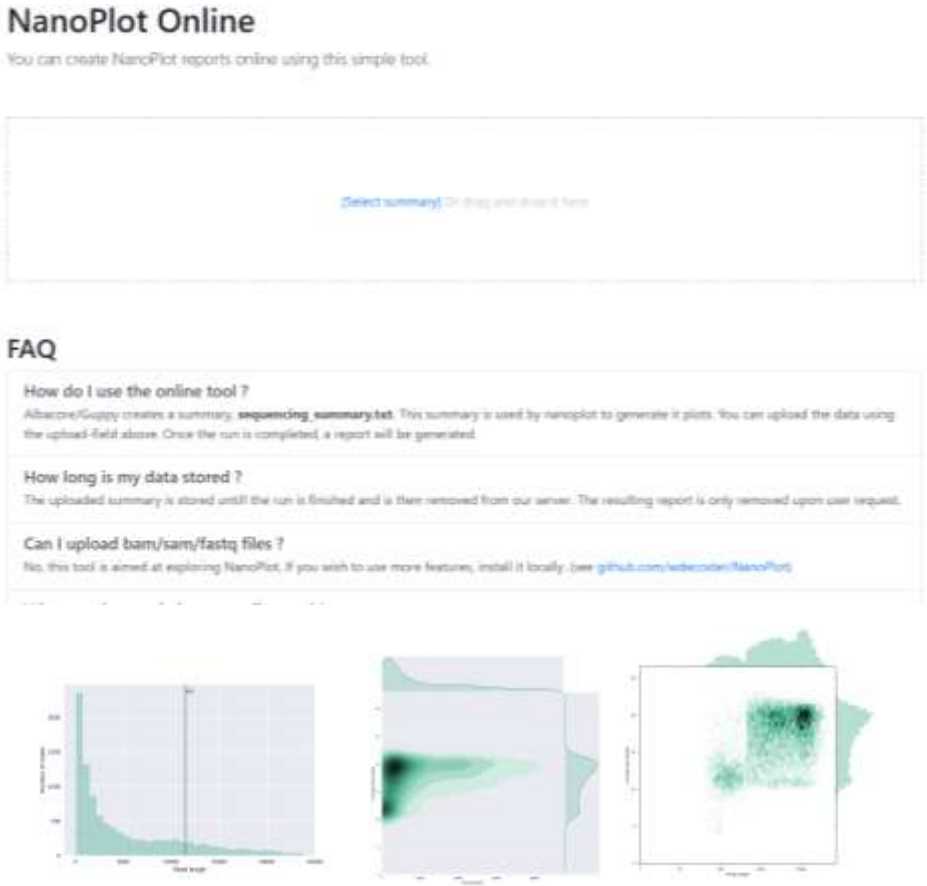
```
pip install NanoPlot --upgrade
```

(or)

```
conda badge
```

```
conda install -c bioconda nanoplot
```

NanoPlot creates → a statistical summary, a number of plots, a html summary file




```
NanoPlot --summary sequencing_summary.txt --loglength -o summary-plots-log-transformed
NanoPlot -t 2 --fastq reads1.fastq.gz reads2.fastq.gz --maxlength 40000 --plots dot --legacy hex
NanoPlot -t 12 --color yellow --bam alignment1.bam alignment2.bam alignment3.bam --downsample 10000 -o bamplots_downsampled
```

PORECHOP

Porechop is a tool for finding and removing adapters from Oxford Nanopore reads

performs thorough alignments to effectively find adapters, even at low sequence identity

Adapters on the ends of reads are trimmed off

If read has an adapter in its middle → treated as chimeric and chopped into separate reads

Porechop also supports demultiplexing of Nanopore reads that were barcoded with the Native Barcoding Kit, PCR Barcoding Kit or Rapid Barcoding Kit.

WORKING:

i. **Find matching adaptors**

ii. **Trim adapter from read ends**

iii. **Split reads with internal adaptors**

iv. **Discards reads with internal adaptors**

v. **Barcode demultiplexing (-b option)**

Basic adapter trimming:

```
porechop -i input_reads.fastq.gz -o output_reads.fastq.gz
```

Trimmed reads to stdout, if you prefer:

```
porechop -i input_reads.fastq.gz > output_reads.fastq
```

Demultiplex barcoded reads:

```
porechop -i input_reads.fastq.gz -b output_dir
```

Demultiplex barcoded reads, straight from Albacore output directory:

```
porechop -i albacore_dir -b output_dir
```

Also works with FASTA:

```
porechop -i input_reads.fasta -o output_reads.fasta
```

More verbose output:

```
porechop -i input_reads.fastq.gz -o output_reads.fastq.gz --verbosity 2
```

Got a big server?

```
porechop -i input_reads.fastq.gz -o output_reads.fastq.gz --threads 40
```

PORECHOP

LIMITATIONS

number of kits/barcodes has increased → adapter-search became increasingly slow

does adapter search on a subset of reads → which means there can be problems with non-randomly ordered read sets (e.g. all barcode 1 reads at the start of a file, followed by barcode 2 reads, etc).

Many ONT adapters share common sequence with each other, making false positive adapter finds possible.

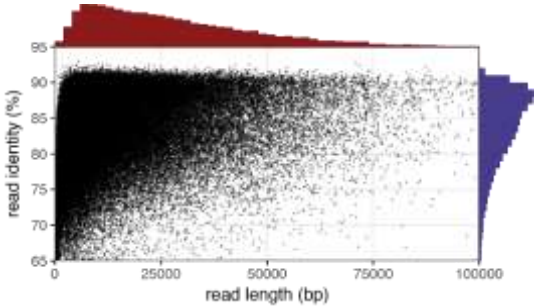
FILTLONG

Filter out low-quality reads from Long read sequencing data (ONT)

Read length + quality metrics data to filter out

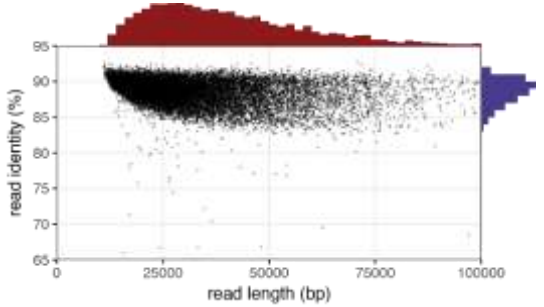
Features: Length filtering, Quality filtering , Read trimming

Figure A)




```
filtlong --min_length 1000 --keep_percent 90 --target_bases 500000000 input.fastq.gz | gzip > output.fastq.gz
```

Figure B)




```
filtlong -1 illumina_1.fastq.gz -2 illumina_2.fastq.gz --min_length 1000 --keep_percent 90 --target_bases 500000000 --trim --split 500 input.fastq.gz | gzip > output.fastq.gz
```

Dot = a single read

Length distribution (top) / → x axis

Identity distribution (right) → y axis

Figure A) Reads before Filtlong → 1.3 Gbp data

Length N50 = 24,077 bp

(i.e. half the bases are in a read 24,077 bp long or longer).

Identity N50 = 85.60%

(i.e. half the bases are in a read with 85.60% or higher identity).

Figure B) Reads after Filtlong - without an external reference

Filtlong has cut the original 1.3 Gbp of reads down to a much better 500 Mbp subset.

Short reads and low identity reads have been mostly removed.

Length N50 = 36,827 bp

Identity N50 = 88.53%

GENOME ASSEMBLY

Aligning and merging fragments from a longer DNA sequence in order to reconstruct the original sequence → whole genomes

ASSEMBLER: Computer program that takes the small sequences and reassembles the original sequence

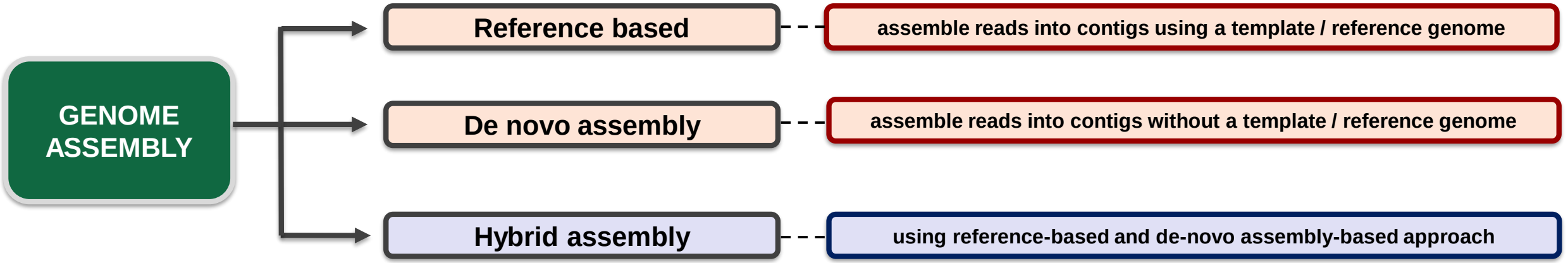
Requirement:

a) Current NGS technologies generate short DNA read lengths

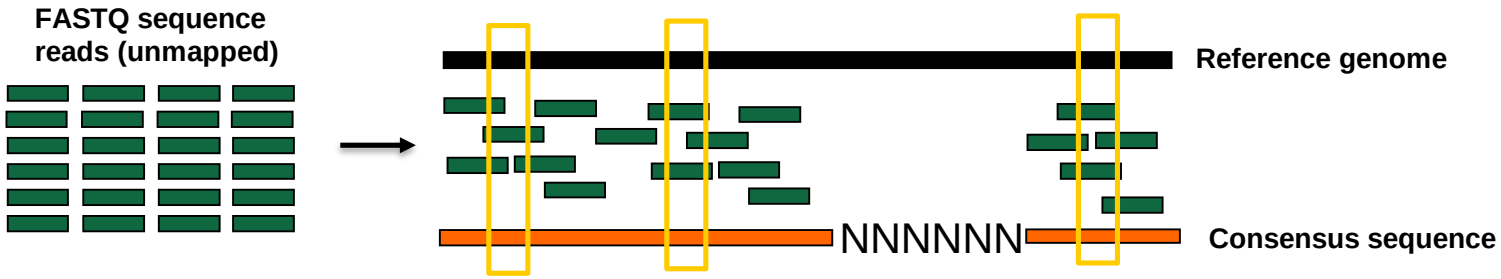
b) Difficult to handle terabytes of sequencing data

c) Resolve repetitive regions

d) Rectify errors generated during sequencing



REFERENCE BASED ASSEMBLY



Sequence reads are mapped to reference genome

i. SNPs → regions that differ between the reference genome and sequence reads

ii. Alignment gaps → large portions of genome present in reference genome but not in sequence reads

iii. Unmapped reads → large regions present in sequence reads but not in reference genome

ADVANTAGES

- Good to identify SNV and small indels
- Less computationally intensive → quick!
- More accurate for identifying regions that are well-represented in the reference

DISADVANTAGES

- High quality genome required
- Read length influences feature detection

TOOLS USED

1) Burrows-Wheeler aligner (short reads)
   https://bio-bwa.sourceforge.net/
2) MiniMap2 (long reads)
   https://github.com/lh3/minimap2

De novo ASSEMBLY

Build the genome using the raw reads only

TWO common approaches :

1. **de Bruijn Graph (DBG) and**
2. **Overlap-Layout-Consensus (OLC)**

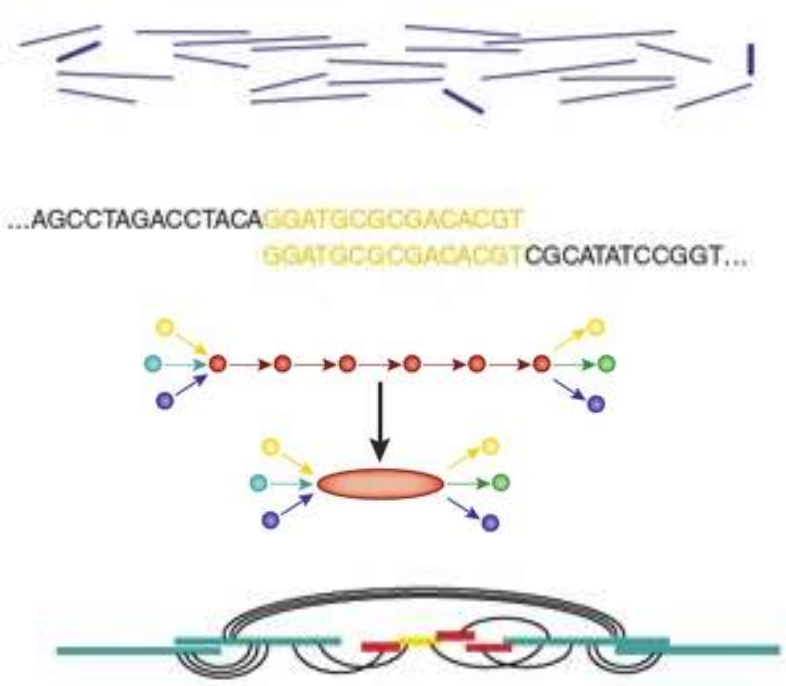
Fragment DNA and sequence it

Identify overlaps

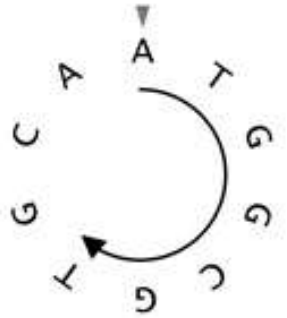
Assemble overlaps into contigs

Assemble contigs into scaffolds

scaffolds → assembly

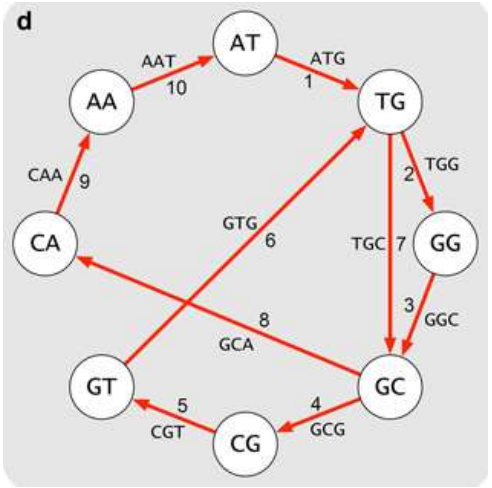


De novo ASSEMBLY: de Bruijn Graph (DBG)

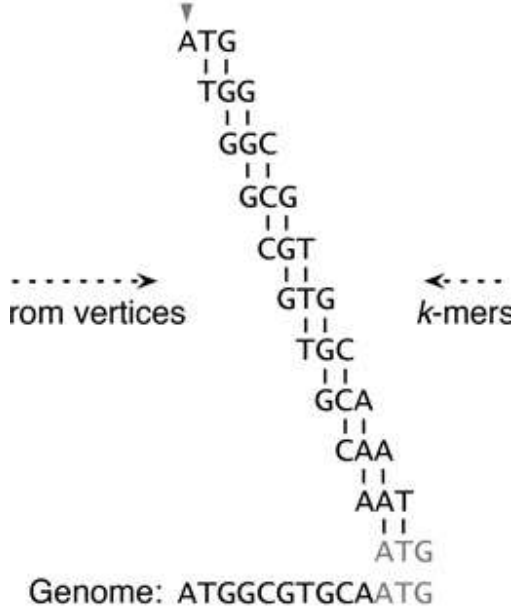


k=2

Split into k-mers of length=2



Looking for the overlaps between each



ADVANTAGES

- **Compact and efficient way to represent large number of k-mers from SHORT reads**
- **Good at handling repeats**
- **Tools incorporate error correction**
- **Yields a high-quality assembly**

DISADVANTAGES

- **Computationally intensive**
- **Choice of k-mer size impacts quality of assembly**
- **Longer complex repeats → difficult to resolve!**

TOOLS FOR SHORT-READ ASSEMBLY

Assembly tools

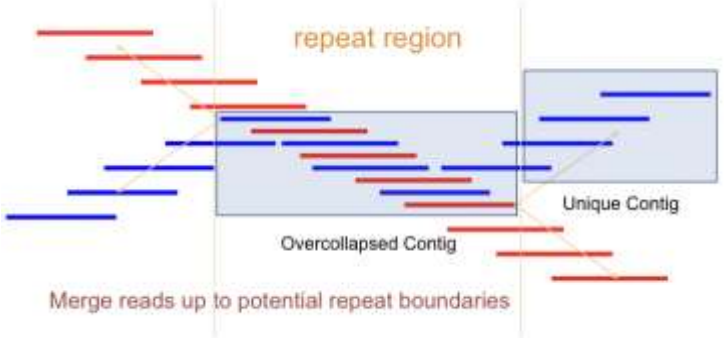
- SPAdes ★
- SKESA ★
- Velvet
- Abyss
- SOAPdenovo2
- MEGAHIT

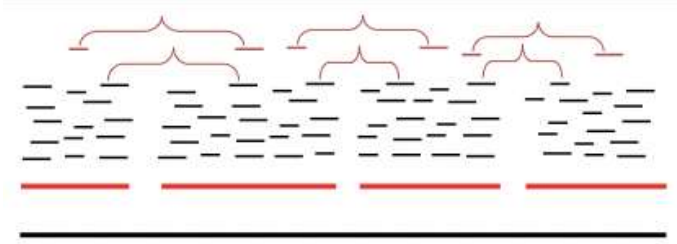
Polishing tools

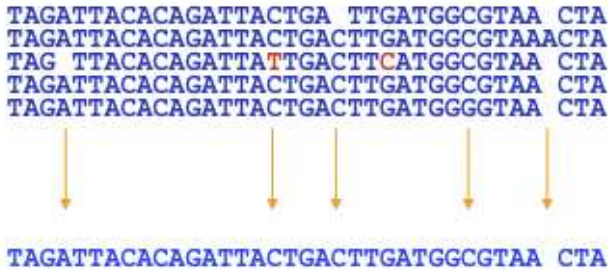
- NextPolish (supports short- and long-read data) ★
- Polypolish ★
- Pypolca (Python-based implementation of POLCA)
- Pilon
- HyPo

Slide inspired by Joana Murao's webinars

De novo ASSEMBLY: Overlap Layout Consensus (OLC)







Overlap → reads obtained from DNA sequencing compared to find overlapping regions

Layout → builds graph based on overlapping reads to arrange reads into contigs

Consensus → Builds final sequence from Layout

ADVANTAGES

- **Handles long reads and complex genomes**
- **Can produce high-quality assemblies**
- **Effective for genome annotation**

DISADVANTAGES

- **Computationally intensive**
- **Can be affected by errors**
- **Requires high coverage**

TOOLS FOR LONG-READ ASSEMBLY

Assembly tools

- Flye (OLC-style repeat graph) ★
- Canu ★
- Dragonflye
- Raven
- SMARTdenovo
- Miniasm
- Autocycler (consensus tool)

Polishing tools

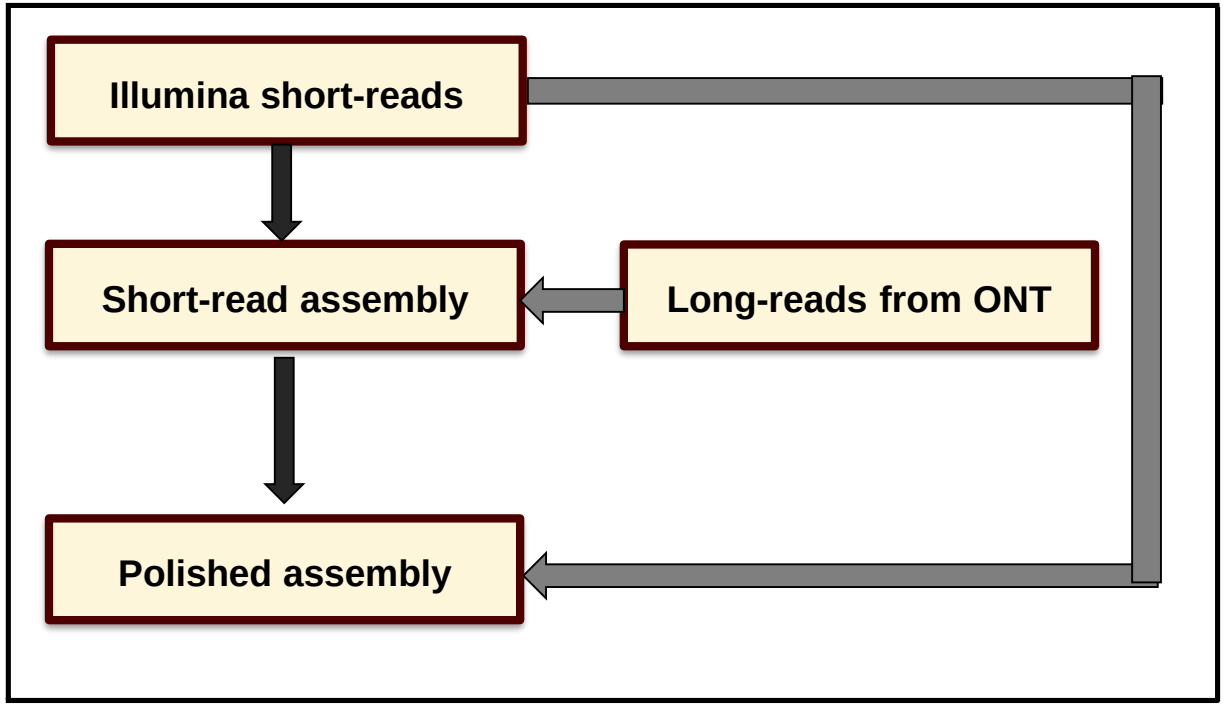
- Medaka (uses a Machine Learning
- model trained on ONT data) ★
- NextPolish ★
- Racon (consensus polishing)
- FMLRC2 (long-read error correction

using high-quality short-read data)

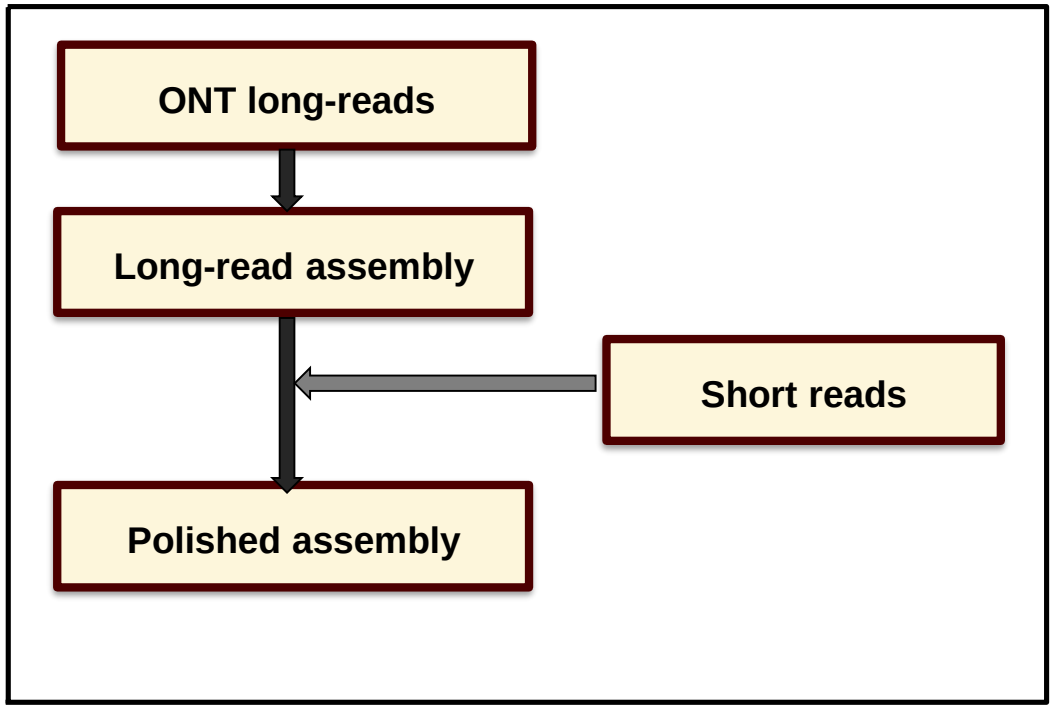
Slide inspired by Joana Murao's webinars

HYBRID ASSEMBLY

channel to strengths of both short-read (highly accurate) and long-read assembly (long fragments that improve accuracy) to produce a highly accurate and complete genome assembly



Tools used: UNICYCLER



Tools used : AUTOCYCLER

channel to strengths of both short-read (highly accurate) and long-read assembly (long fragments that improve accuracy) to produce a highly accurate and complete genome assembly

Illumina short-

Short-read ass

Polished assembly

Tools used: UNICYCLER

ort reads

Polished assembly

Tools used : AUTOCYCLER

ADVANTAGES

- Improved assembly quality
- Better accuracy while resolving complex regions
- Detection of genomic features
- Increased reference coverage

DISADVANTAGES

- Highly computationally intensive
- High probability for error

ASSEMBLY QUALITY ASSESSMENT

CheckM2 - GitHub

- Assess the quality of genome assembly – specifically completeness and contamination
- >90% completeness ; <5% contamination

QUAST - manual

- Provides comprehensive metrics and visual reports to check assembly quality
- Suitable for *de novo* short-read, long-read and hybrid assembly

CheckM2 output:

Genome	GTDB taxonomy	Expected CheckM2 Completeness (v.1.1.0)	Expected CheckM2 Contamination (v.1.1.0)	CheckM1 Completeness	CheckM1 Contamination
TEST1	d__Bacteria;	100	0.76	99.97	0.04
TEST2	p__Patescibacteria;	98.96	0.21	79.86	0.00
TEST3	f__Nanosalinaceae;	98.77	0.50	87.77	0.00

ASSEMBLY QUALITY : QUAST

#contigs → total number of contigs of length x

Less number of contigs = better contiguity

Aligned to "BCep_ref" | 8 605 945 bp | 4 fragments | 66.61% G+C
7705 genomic features

Worst Median Best — Show heatmap

Genome statistics	spades_default	spades_kmers_careful	megahit_default	megahit_min_count_3
Genome fraction (%)	98.14	98.421	98.6	98.603
Duplication ratio	1	1	1.001	1.001
# genomic features	7539 + 75 part	7563 + 62 part	7540 + 105 part	7540 + 104 part
Largest alignment	455 950	505 898	350 746	350 746
Total aligned length	8 436 553	8 470 789	8 492 473	8 493 177
NGA50	143 431	198 969	144 083	125 159
LGA50	18	13	20	21
Misassemblies				
# misassemblies	6	6	9	8
Misassembled contigs length	1 469 048	1 719 134	1 200 775	1 050 989
Mismatches				
# mismatches per 100 kbp	3.85	2.72	2.38	2.3
# indels per 100 kbp	0.67	0.45	0.32	0.28
# N's per 100 kbp	0	0	0	0
Statistics without reference				
# contigs	132	91	156	158
Largest contig	754 490	961 949	539 126	539 126
Total length	8 447 218	8 472 540	8 492 975	8 493 797
Total length (>= 1000 bp)	8 447 218	8 472 540	8 492 975	8 493 797
Total length (>= 10000 bp)	8 324 069	8 384 754	8 296 360	8 308 919
Total length (>= 50000 bp)	7 438 644	7 723 080	6 917 725	6 910 273

Slide inspired by Joana Murao's webinars

ASSEMBLY QUALITY : QUAST

Genomic features → number of annotated elements found in assembly fully/partially from reference genome

Aligned to "BCep_ref" | 8 605 945 bp | 4 fragments | 66.61% G+C
7705 genomic features

Worst Median Best — Show heatmap

Genome statistics	spades_default	spades_kmers_careful	megahit_default	megahit_min_count_3
Genome fraction (%)	98.14	98.421	98.6	98.603
Duplication ratio	1	1	1.001	1.001
# genomic features	7539 + 75 part	7563 + 62 part	7540 + 105 part	7540 + 104 part
Largest alignment	455 950	505 898	350 746	350 746
Total aligned length	8 436 553	8 470 789	8 492 473	8 493 177
NGA50	143 431	198 969	144 083	125 159
LGA50	18	13	20	21
Misassemblies				
# misassemblies	6	6	9	8
Misassembled contigs length	1 469 048	1 719 134	1 200 775	1 050 989
Mismatches				
# mismatches per 100 kbp	3.85	2.72	2.38	2.3
# indels per 100 kbp	0.67	0.45	0.32	0.28
# N's per 100 kbp	0	0	0	0
Statistics without reference				
# contigs	132	91	156	158
Largest contig	754 490	961 949	539 126	539 126
Total length	8 447 218	8 472 540	8 492 975	8 493 797
Total length (>= 1000 bp)	8 447 218	8 472 540	8 492 975	8 493 797
Total length (>= 10000 bp)	8 324 069	8 384 754	8 296 360	8 308 919
Total length (>= 50000 bp)	7 438 644	7 723 080	6 917 725	6 910 273

Slide inspired by Joana Murao's webinars

ASSEMBLY QUALITY : QUAST

Mismatches and indels → average number of base mismatches / indels per 100,000 bp

Low number = high accuracy

Aligned to "BCep_ref" | 8 605 945 bp | 4 fragments | 66.61% G+C
7705 genomic features

Worst Median Best — Show heatmap

Genome statistics	spades_default	spades_kmers_careful	megahit_default	megahit_min_count_3
Genome fraction (%)	98.14	98.421	98.6	98.603
Duplication ratio	1	1	1.001	1.001
# genomic features	7539 + 75 part	7563 + 62 part	7540 + 105 part	7540 + 104 part
Largest alignment	455 950	505 898	350 746	350 746
Total aligned length	8 436 553	8 470 789	8 492 473	8 493 177
NGA50	143 431	198 969	144 083	125 159
LGA50	18	13	20	21
Misassemblies				
# misassemblies	6	6	9	8
Misassembled contigs length	1 469 048	1 719 134	1 200 775	1 050 989
Mismatches				
# mismatches per 100 kbp	3.85	2.72	2.38	2.3
# indels per 100 kbp	0.67	0.45	0.32	0.28
# N's per 100 kbp	0	0	0	0
Statistics without reference				
# contigs	132	91	156	158
Largest contig	754 490	961 949	539 126	539 126
Total length	8 447 218	8 472 540	8 492 975	8 493 797
Total length (>= 1000 bp)	8 447 218	8 472 540	8 492 975	8 493 797
Total length (>= 10000 bp)	8 324 069	8 384 754	8 296 360	8 308 919
Total length (>= 50000 bp)	7 438 644	7 723 080	6 917 725	6 910 273

Slide inspired by Joana Murao's webinars

ASSEMBLY QUALITY : QUAST

Genome fraction →

percentage of aligned bases

in the reference genome

Aligned to "BCep_ref" | 8 605 945 bp | 4 fragments | 66.61% G+C
7705 genomic features

Worst Median Best — Show heatmap

Genome statistics	spades_default	spades_kmers_careful	megahit_default	megahit_min_count_3
Genome fraction (%)	98.14	98.421	98.6	98.603
Duplication ratio	1	1	1.001	1.001
# genomic features	7539 + 75 part	7563 + 62 part	7540 + 105 part	7540 + 104 part
Largest alignment	455 950	505 898	350 746	350 746
Total aligned length	8 436 553	8 470 789	8 492 473	8 493 177
NGA50	143 431	198 969	144 083	125 159
LGA50	18	13	20	21
Misassemblies				
# misassemblies	6	6	9	8
Misassembled contigs length	1 469 048	1 719 134	1 200 775	1 050 989
Mismatches				
# mismatches per 100 kbp	3.85	2.72	2.38	2.3
# indels per 100 kbp	0.67	0.45	0.32	0.28
# N's per 100 kbp	0	0	0	0
Statistics without reference				
# contigs	132	91	156	158
Largest contig	754 490	961 949	539 126	539 126
Total length	8 447 218	8 472 540	8 492 975	8 493 797
Total length (>= 1000 bp)	8 447 218	8 472 540	8 492 975	8 493 797
Total length (>= 10000 bp)	8 324 069	8 384 754	8 296 360	8 308 919
Total length (>= 50000 bp)	7 438 644	7 723 080	6 917 725	6 910 273

Slide inspired by Joana Murao's webinars

THANK YOU FOR LISTENING, ANY QUESTIONS ?

Kindly ensure to drop your email address on the chat to receive the certificates