

Introduction to Mini SIMEX

(mini simulated exercises)

Presented by: Praissy Zefi J

pzeje@dtu.dk

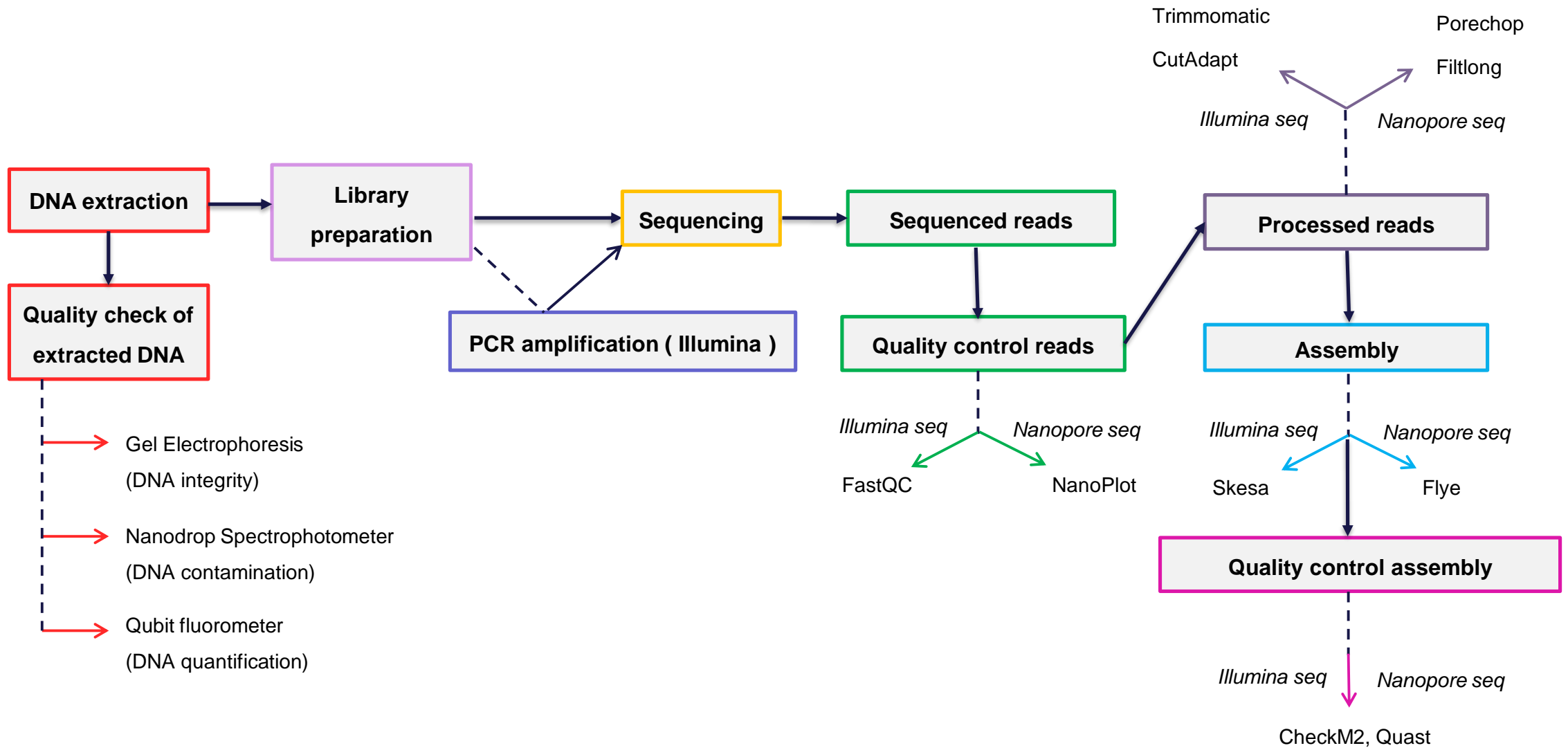
Bioinformatician, DTU

Presented by: Faisal Ahmad Khan

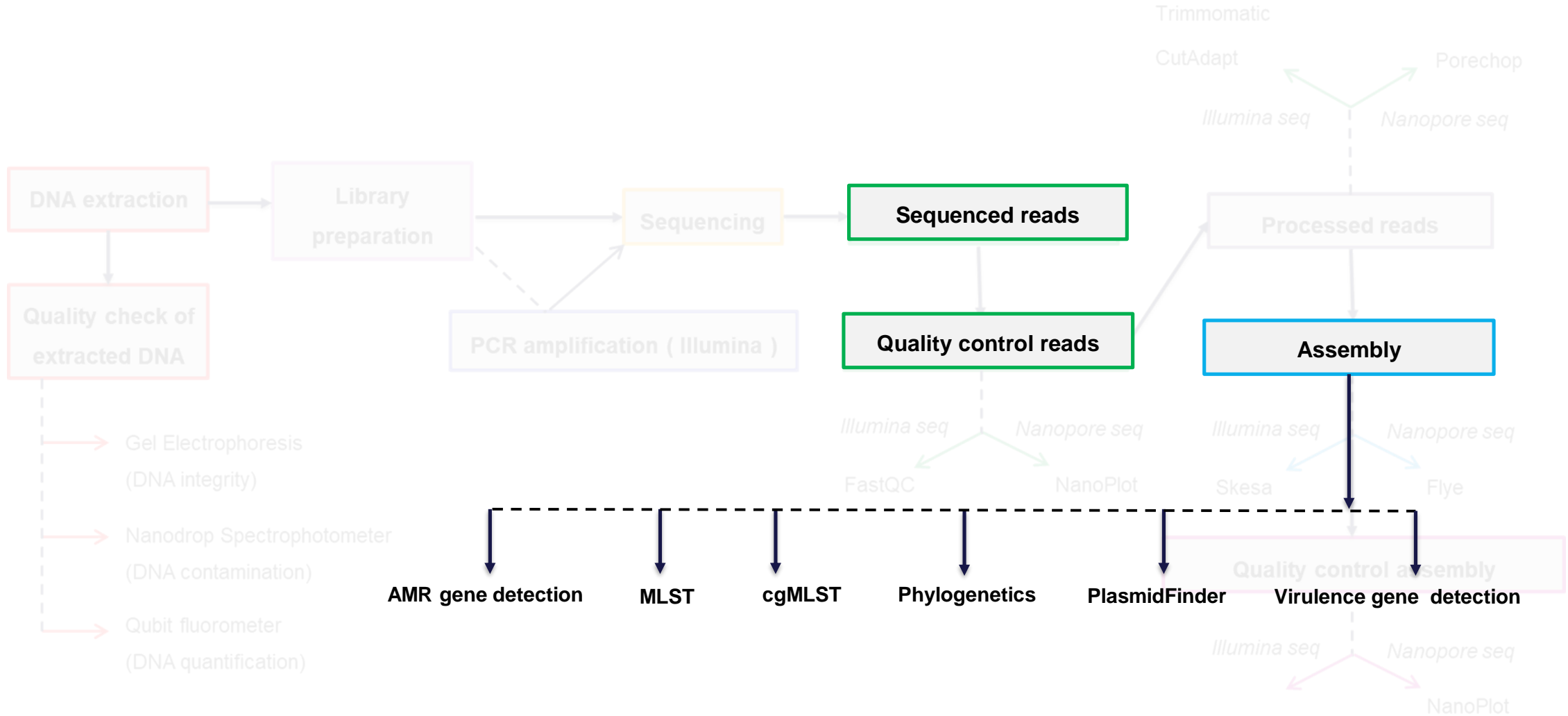
fakh@dtu.dk

Microbiologist Post-doc, DTU

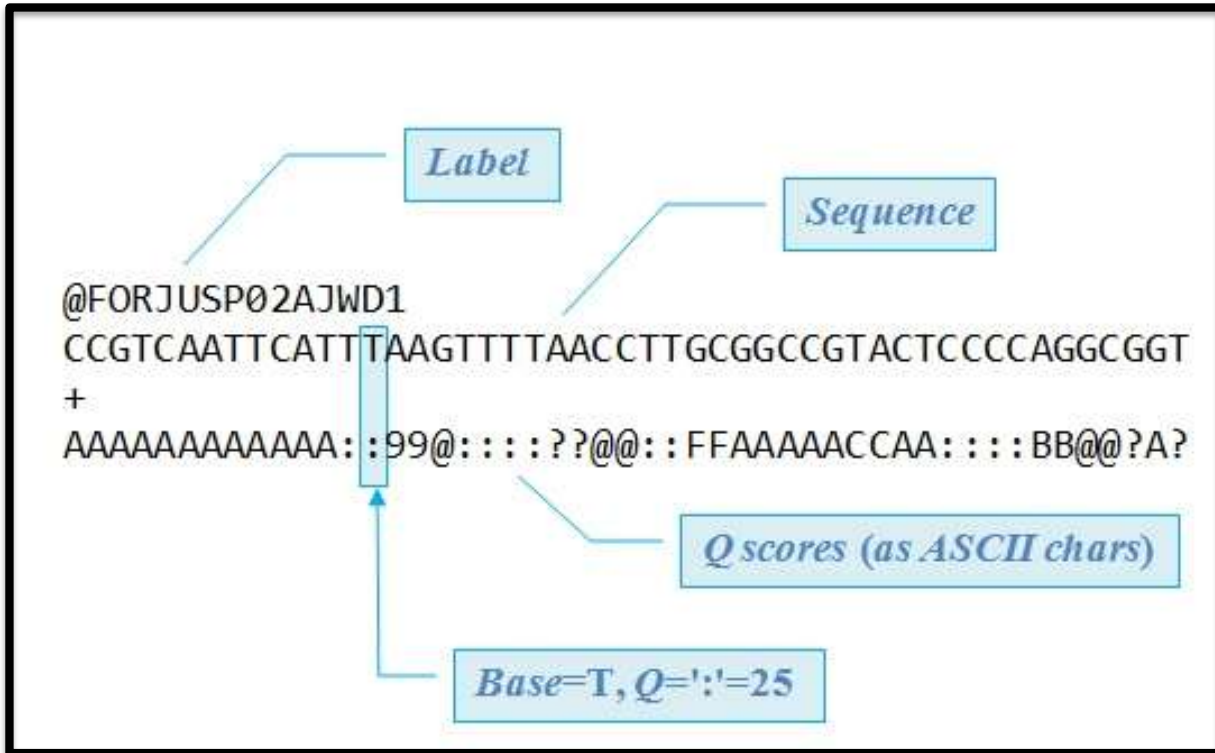
BACTERIAL WHOLE GENOME SEQUENCING



BACTERIAL WHOLE GENOME SEQUENCING



Illumina sequencing → FASTQ file

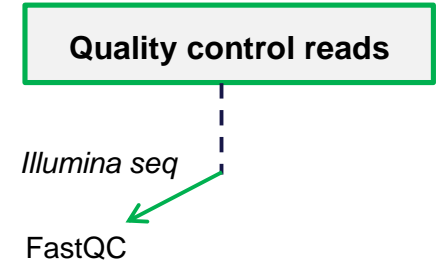


Nanopore sequencing → FASTA file

```

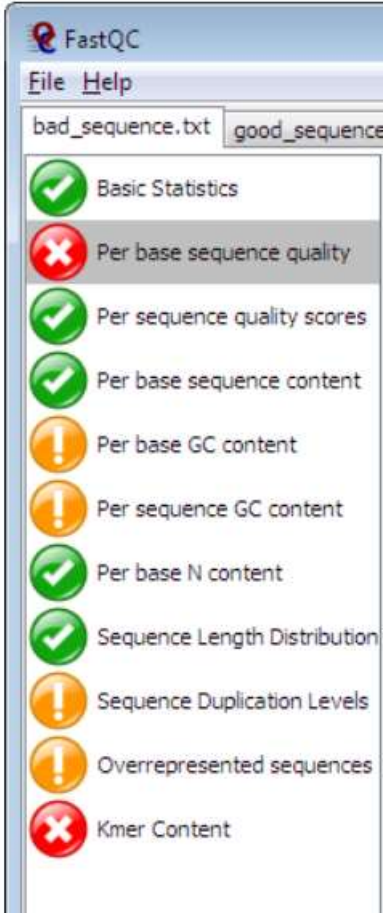
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACC CGGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCCGGCGAGGTGCAGCAGAAGTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
  
```

**Q30 → 1 in 1000 chance that a base is called incorrectly
 => 1/1000 = 0.001 error rate**



Illumina - 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 |

ILLUMINA SEQUENCING – QUALITY CONTROL

1)



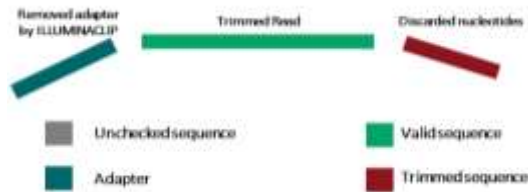
```
@SEQ_ID_1
GATAAAGCAGTATCGAT
+
!"*%%%).1***+**))**
@SEQ_ID_2
TTTGGGGTTCAAAT
+
%%%).1***CCF>>>>>CC
@SEQ_ID_3
GATCAAAGCAGTATCGAT
+
!"*(1***+**))**55CCF>>>>>
```

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 |

cutadapt

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

ERROR TOLERANCE

Allowed errors : matches, mismatches, deletions, insertions

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

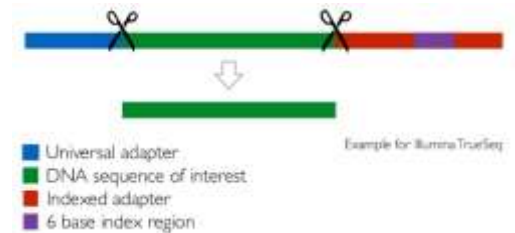
Eg 1: adapter match of length 8 containing 1 error →

error rate = $1 / 8 = 0.125$

Default maximum error rate = 0.1

Will this be detected or not? No!

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



ONT SEQUENCING - NANO PLOT

Quality control reads

Nanopore seq

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 Online

You can create NanoPlot reports online using this simple tool.



FAQ

How do I use the online tool ?

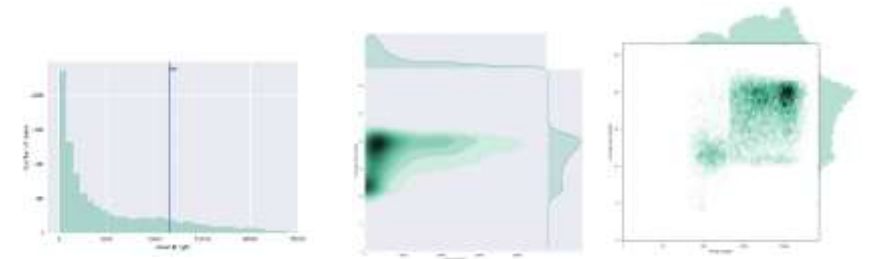
Albacore/Guppy creates a summary: `sequencing_summary.txt`. This summary is used by nanoplot to generate its plots. You can upload the data using the upload-field above. Once the run is completed, a report will be generated.

How long is my data stored ?

The uploaded summary is stored until the run is finished and is then removed from our server. The resulting report is only removed upon user request.

Can I upload bam/sam/fastq files ?

No, this tool is aimed at exploring NanoPlot. If you wish to use more features, install it locally: (see github.com/walbrun/nanoPlot)



```
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
```

Trimming tools – Nanopore sequencing

PORECHOP



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:

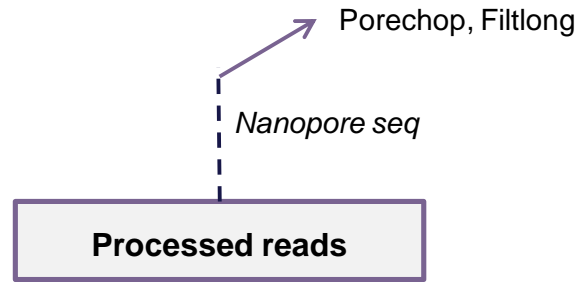
```
porechop -i input_reads.fasta -o output_reads.fasta
```

More verbose output:

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

Got a big server?

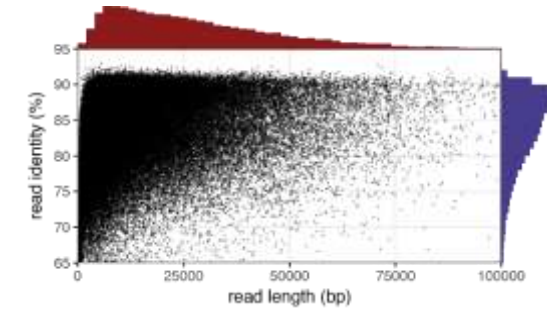
```
porechop -i input_reads.fastq.gz -o output_reads.fastq.gz --threads 40
```



FILTLONG

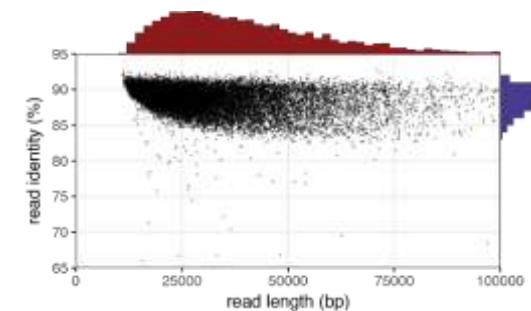


Figure A)



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

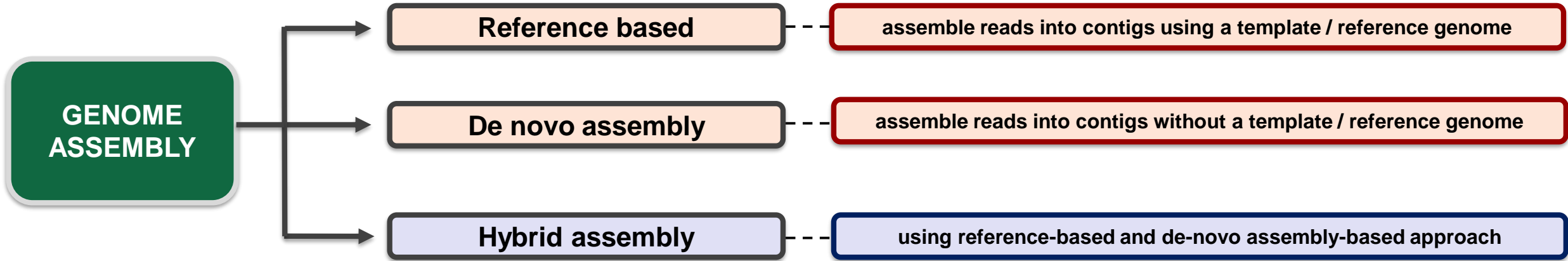
Figure B)



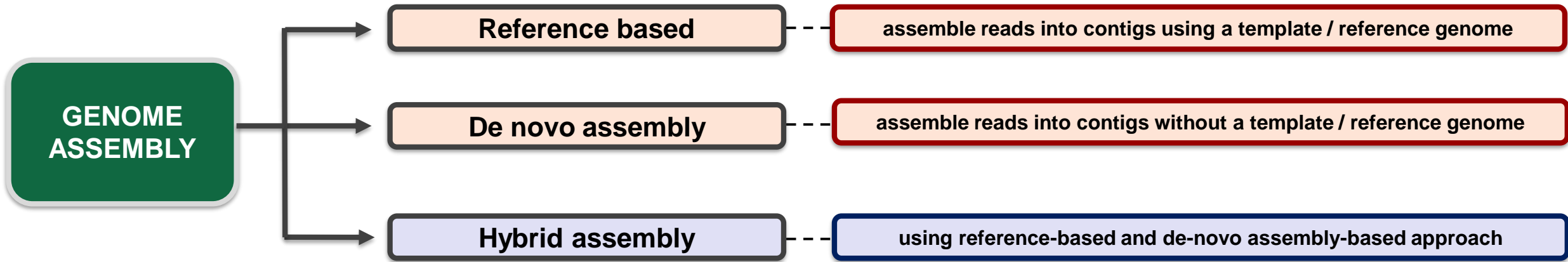
```
filtlong -1 illumina_1.fastq.gz -2 illumina_2.fastq.gz --min_length 1000 --keep_percent 90 --target_bases 500000000 --trim --split 500 input.fastq.gz | gzip > output.fastq.gz
```

GENOME ASSEMBLY

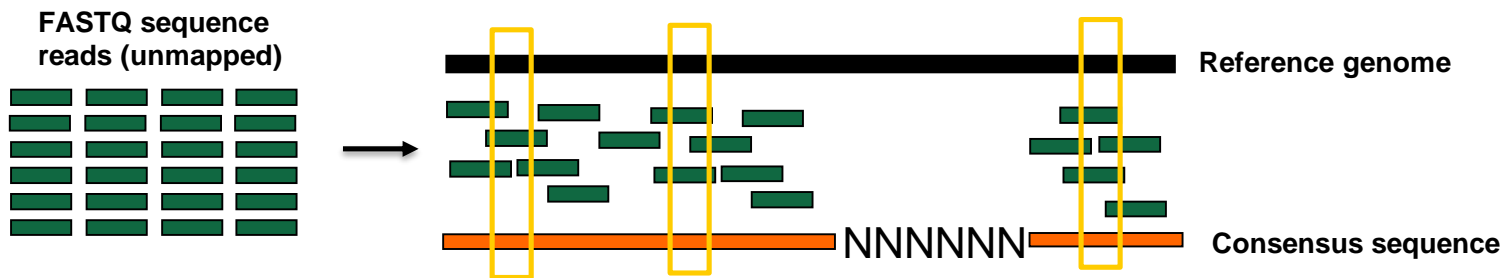
Assembly



GENOME ASSEMBLY



REFERENCE BASED ASSEMBLY



TOOLS USED

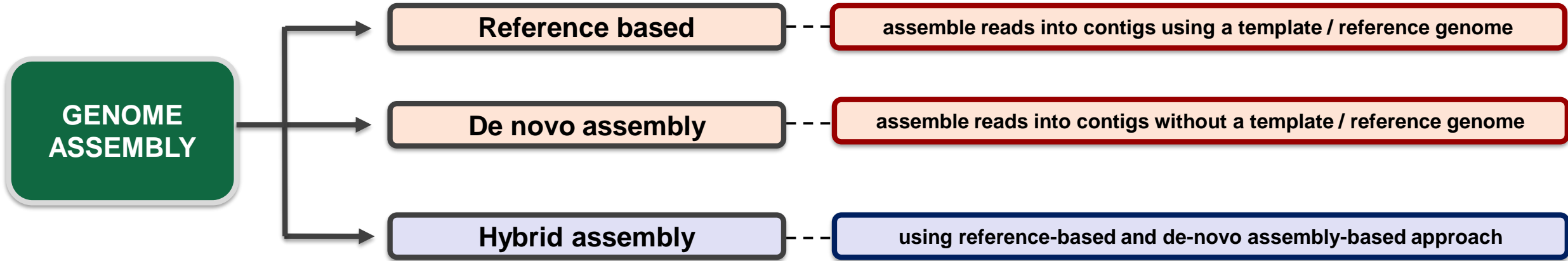
1) Burrows-Wheeler aligner (short reads)

<https://bio-bwa.sourceforge.net/>

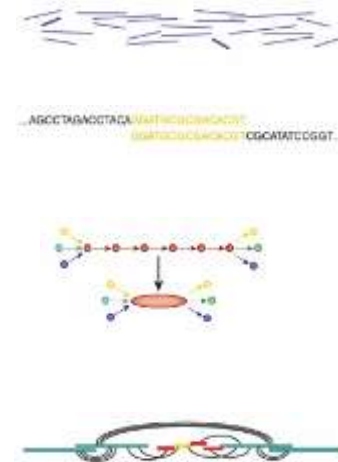
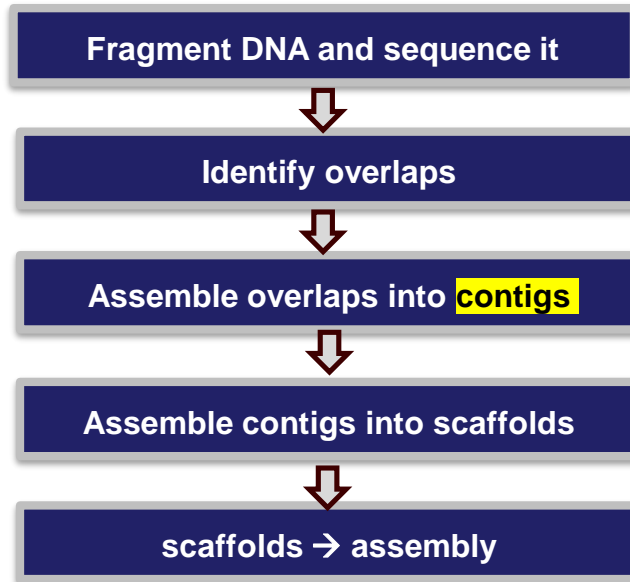
2) MiniMap2 (long reads)

<https://github.com/lh3/minimap2>

GENOME ASSEMBLY



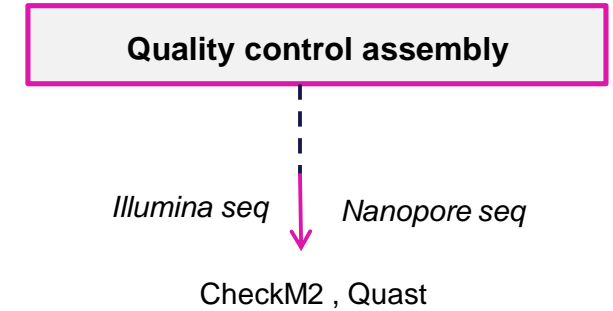
De novo ASSEMBLY



Assembly tools

- Short read assembly**
 - SPAdes
 - SKESA
- Long read assembly**
 - Flye (OLC-style repeat graph)
 - Canu

ASSEMBLY QUALITY ASSESSMENT



CheckM2 - [GitHub](#)

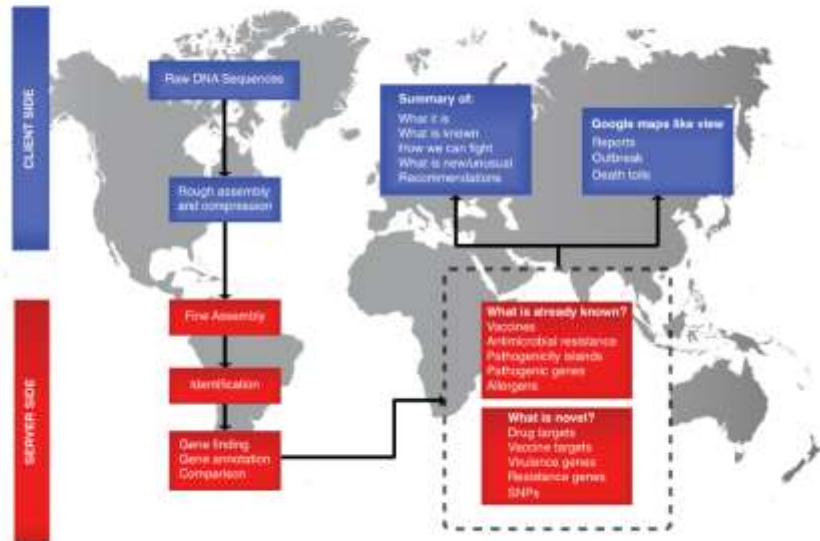
- Assess the quality of genome assembly – specifically **completeness** and **contamination**
- >90% completeness ; <5% contamination

QUAST - [manual](#)

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

Let's look at how it works and the output!

MLST and cgMLST



Welcome to the Center for Genomic Epidemiology

News

Scaling neighbor joining to one million taxa with dynamic and heuristic neighbor joining
January 2022
[Link to article...](#)

Sourcefinder: a Machine-Learning-Based Tool for Identification of Chromosomal, Plasmid, and Bacteriophage Sequences from Assemblies
November 2022
[Link to article...](#)

PlasmidHostFinder: prediction of plasmid hosts using random forest
April 2022
[Link to article...](#)

ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes
January 2012
[Link to article...](#)

MINITyper: an outbreak-detection method for accurate and rapid SNP typing of clonal clusters with noisy long reads
April 2021
[Link to article...](#)

Automated download and clean-up of family-specific databases for enter-based virus identification
October 2020

Center for Genomic Epidemiology

Home Services Publications Contact

MLST 2.0

Service [Instructions](#) [Output](#) [Article abstract](#) [Citations](#)

Software version: 2.0.0 (2022-05-11)
Database version: (2021-08-04)
MLST allele sequence and profile data is obtained from PubMLST.org.

Warning: the species *Lactococcus Lactis* is unavailable.

Select MLST configuration

Please note that for four organisms, two or three different MLST schemes are available:

- *Acinetobacter baumannii* (*Acinetobacter baumannii* #1 [1], *Acinetobacter baumannii* #2 [2]).
- *Escherichia coli* (*Escherichia coli* #1 [4], *Escherichia coli* #2 [3]).
- *Pasteurella multocida* (*Pasteurella multocida* #1 (MDC), *Pasteurella multocida* #2 (multihost)).
- *Leptospira* (*Leptospira* #1, *Leptospira* #2, *Leptospira* #3).

Note: *Campylobacter coli* and *Campylobacter jejuni* are considered together.

Select min. depth for an allele

Select type of data input

Only data from one single isolate should be uploaded. If raw sequencing reads are uploaded KMA will be used for mapping. KMA supports the following sequencing platforms: Illumina, Ion Torrent, Roche 454, SQUID, Oxford Nanopore, and PacBio.

Please note that "Assembled Genomes/Contigs" should be selected. If you have already assembled your short sequencing reads into one continuous genome or into several contigs, it is indifferent which type of short sequence reads were used to produce the genome/contigs.

<https://cge.food.dtu.dk/services/MLST/>



Scenario

| Species | Date | Region | Travel | MLST | Sequence | Carba genotype (PCR) |
|---------|------|-------------|----------|---------|----------|----------------------|
| E. coli | 2015 | Copenhagen | Pakistan | ST410 | Ec001 | OXA |
| E. coli | 2015 | Copenhagen | Thailand | ST410 | Ec002 | OXA |
| E. coli | 2015 | Jutland - M | India | ST410 | Ec003 | NDM |
| E. coli | 2015 | Copenhagen | Lebanon | Missing | Ec004 | OXA |
| E. coli | 2016 | Zealand | No | Missing | Ec005 | NDM, OXA |
| E. coli | 2016 | Zealand | No | ST410 | Ec006 | NDM, OXA |
| E. coli | 2017 | Copenhagen | Pakistan | ST410 | Ec007 | OXA |
| E. coli | 2018 | Jutland - N | Thailand | ST410 | Ec008 | NDM |
| E. coli | 2018 | Zealand | No | ST410 | Ec009 | NDM, OXA |
| E. coli | 2018 | Zealand | No | ST410 | Ec010 | NDM, OXA |
| E. coli | 2018 | Zealand | No | ST410 | Ec011 | NDM |
| E. coli | 2018 | Zealand | No | ST410 | Ec012 | OXA |

Scenario

A recent rise in cases of carbapenemase producing *E. coli* in several regional hospitals indicate one or more ongoing outbreaks, and it has been suggested that the NRL could give assistance by performing outbreak investigation by WGS. Patients involve both domestic and travel-related cases and a batch of samples has already been sequenced using Illumina sequencing platform (NextSeq). From these sequences, subtyping by MLST was performed and a selection (12 *E. coli* isolates) of the most predominant MLST (ST410) isolates has been transported to your laboratory for further analysis.

Data Quality and SNP Calling

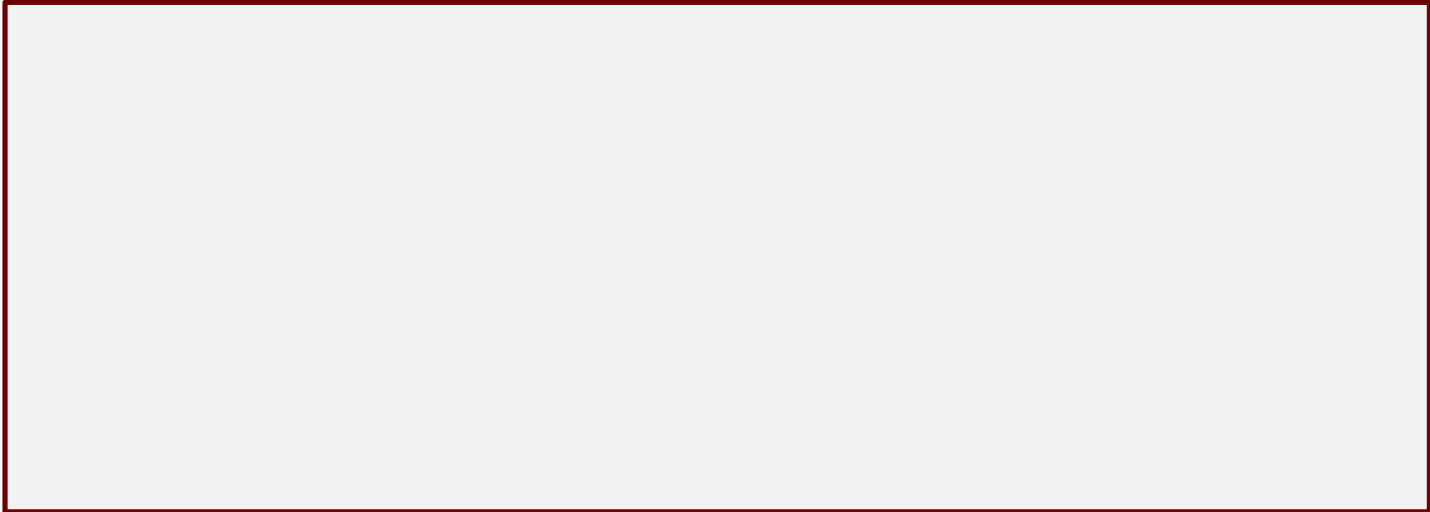
- Good data quality ensures reliability of your analysis
 - Poor quality sequences can rarely be used for SNP analysis
- For assembled contigs - good coverage is essential ($\geq 30x$)
- Consider the quality of your raw data (specifically phred scores)

- CSI Phylogeny SNP filtering criteria:
 - SNP quality: ≥ 30 (Phred score, base call accuracy: 99.9%)
 - SNPs with a sequence depth of < 10 are removed
 - A SNP is removed if it is < 10 bps from the nearest SNP (Pruning)
(recombination do not reflect naturally evolved SNPs)

Preferably analyse raw reads for better resolution!

CSI Phylogeny 1.4 (Call SNPs & Infer Phylogeny)

CSI Phylogeny calls SNPs, filters the SNPs, does site validation and infers a phylogeny based on the concatenated alignment of the high quality* SNPs.
The service is having some issues with files compressed. Please submit all files uncompressed.



Input data

Upload reference genome (fasta format)

Note: Reference genome must not be compressed.

Vælg fil Der er ikke valgt nogen fil
 Include reference in final phylogeny.

Select min. depth at SNP positions
10x

Select min. relative depth at SNP positions
10 %

Select minimum distance between SNPs (prune)
10 bp

Select min. SNP quality
30

Select min. read mapping quality
25

Select min. Z-score
1.98

Ignore heterozygous SNPs

Comment (to yourself)

This comment will appear unaltered on your output page. It has no effect on the analysis

Use altered FastTree (more accurate)

Note: Read more [here](#)

Upload read files and/or assembled genomes (fasta or

Note: Read files must be compressed with gzip (compressed files often e
If you get an "Access forbidden. Error 403": Make sure the start of the we

Isolate File

Name

Upload

Remove

Select min. depth at SNP positions

10x

Select min. relative depth at SNP positions

10 %

Select minimum distance between SNPs (prune)

10 bp

Select min. SNP quality

30

Select min. read mapping quality

25

-score

...ATCGAATTCGGGTTTTTAACCGGATCGTACGATCGGGAAAAA..

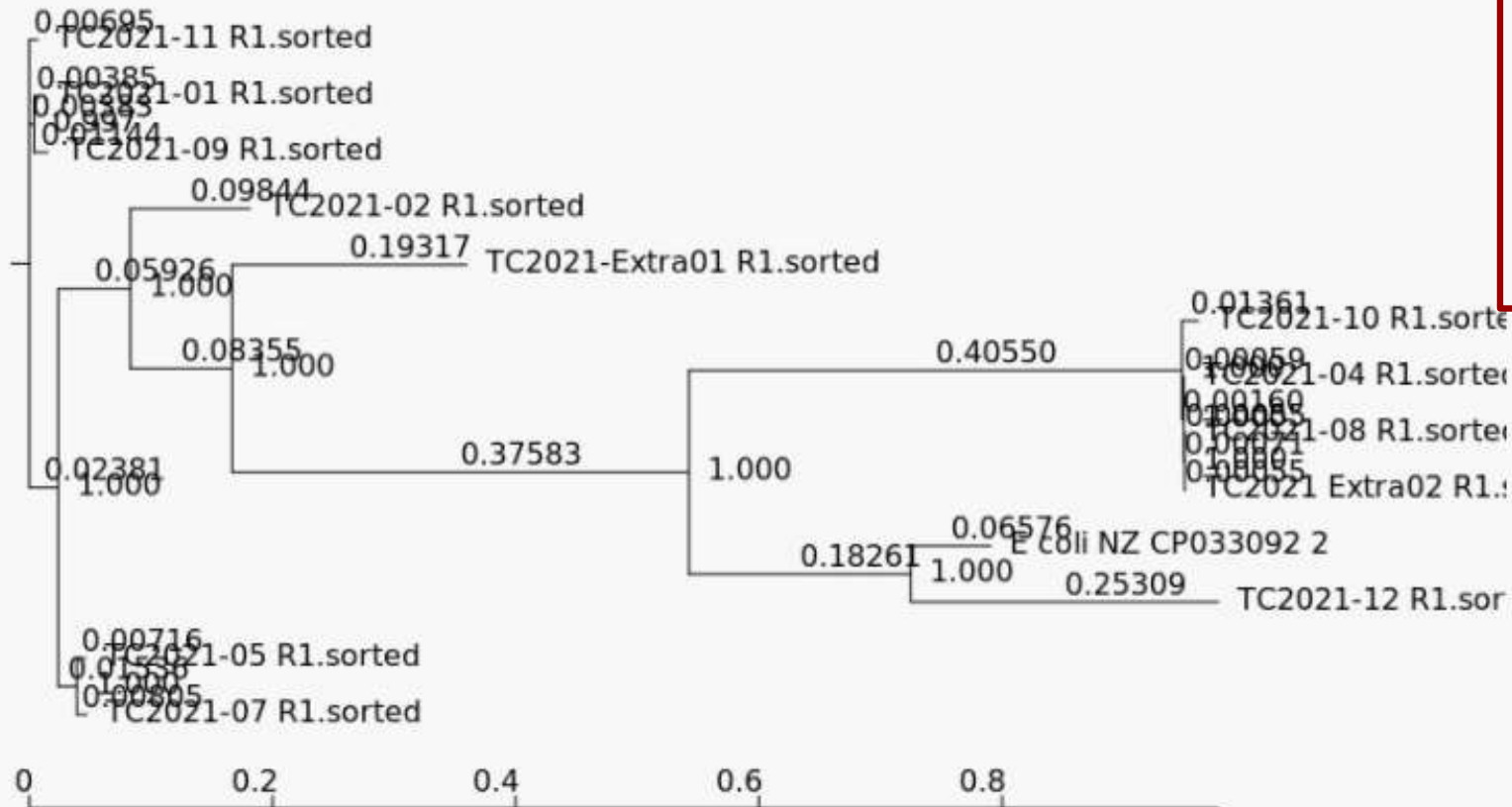
TTCCAGGTTTTTAACCAGATCG
TTCCAGGTTTTTAACCAGATCG
TTCCAGGTTTTTAACCAGATCG
TTCCAGGTTTTTAACCAGATCG
TTCCAGGTTTTTAACCAGATCG
TTCCAGGTTTTTAACCAGATCG

11 bp

CSI output – web interface

CSIPhylogeny Results

The tree presented in the picture below is only meant as a preview. If the tree is meant to be shared or published, we strongly recommend that the 'Newick' file is downloaded and processed using software created for this purpose. We suggest ([FigTree](#)).



Download the filtered SNP calls in Variant Calling Format (VCF):

Note: VCF files are compressed with gzip.

VCF files

Download matrix of SNP pair counts:

Download matrix as:

Download SNP alignment:

Download phylogeny as:

CSIPhylogeny output page

Percentage of reference genome covered by all isolates: 71.473402371081
 3504699 positions was found in all analyzed genomes.
 Size of reference genome: 4903501

Below is listed the number of positions that are shared and trusted between each isolate and the reference genome.

| File | Valid positions | Pct. of reference |
|-----------------------------------|-----------------|-------------------|
| TC2021-05_R1.ignored_snps | 3978591 | 81.137762590443 |
| TC2021-12_R1.ignored_snps | 4307863 | 87.852801498358 |
| TC2021-02_R1.ignored_snps | 4039549 | 82.3809151869246 |
| TC2021-01_R1.ignored_snps | 4048331 | 82.5600117140794 |
| TC2021-09_R1.ignored_snps | 4003614 | 81.6480714493583 |
| TC2021-08_R1.ignored_snps | 4101898 | 83.6524352702284 |
| TC2021-10_R1.ignored_snps | 4117054 | 83.9615205543957 |
| TC2021-Extra01_R1.ignored_snps | 3985371 | 81.2760311459098 |
| TC2021-07_R1.ignored_snps | 4048219 | 82.5577276317472 |
| E_coli_NZ_CP033092_2.ignored_snps | 4903501 | 100 |
| TC2021-11_R1.ignored_snps | 3986463 | 81.2983009486487 |
| TC2021-04_R1.ignored_snps | 4142652 | 84.4835557288558 |
| TC2021_Extra02_R1.ignored_snps | 4067475 | 82.9504266441467 |

How to choose a reference genome

- The reference should be somewhat similar to the isolates you test
 - You can use an internal reference in your collection
- Better described (annotated strain)
 - Search for something similar in kmerFinder
- The more distant your reference is from the dataset you analyse, the less bases you will build the SNP analysis on
 - -> false lower number of SNPs if you choose a bad reference

KmerFinder –species ID and contamination

KmerFinder 3.2

Service
Instructions
Output
Article abstract
Citations

Software version: 3.0.2 (2020-10-30)
 Database version: (2022-07-11)
 The database can be downloaded [here](#)

Select database

Bacteria organisms

Upload file(s)
 To input the sequences, upload a single FASTA file, or one/two FASTQ file(s), or one interleaved FASTQ file on your local disk by using the applet below. Both assembled genome (in FASTA format) and raw reads single end or paired end (in FASTQ format) are supported. Gzipped FASTA/FASTQ files are also supported.

If you get an "Access forbidden. Error 403". Make sure the start of the web address is https and not just http. Fix it by clicking [here](#).

Choose File(s)

| Name | Size | Progress | Status |
|------------------------------------|-----------|--|--------|
| Ec001.illumina_R1.trimmed.fastq.gz | 113.15 MB | <div style="width: 20%; height: 10px; background-color: #007bff;"></div> | |
| Ec001.illumina_R2.trimmed.fastq.gz | 96.00 MB | <div style="width: 0%; height: 10px; background-color: #ccc;"></div> | |

Upload
Remove

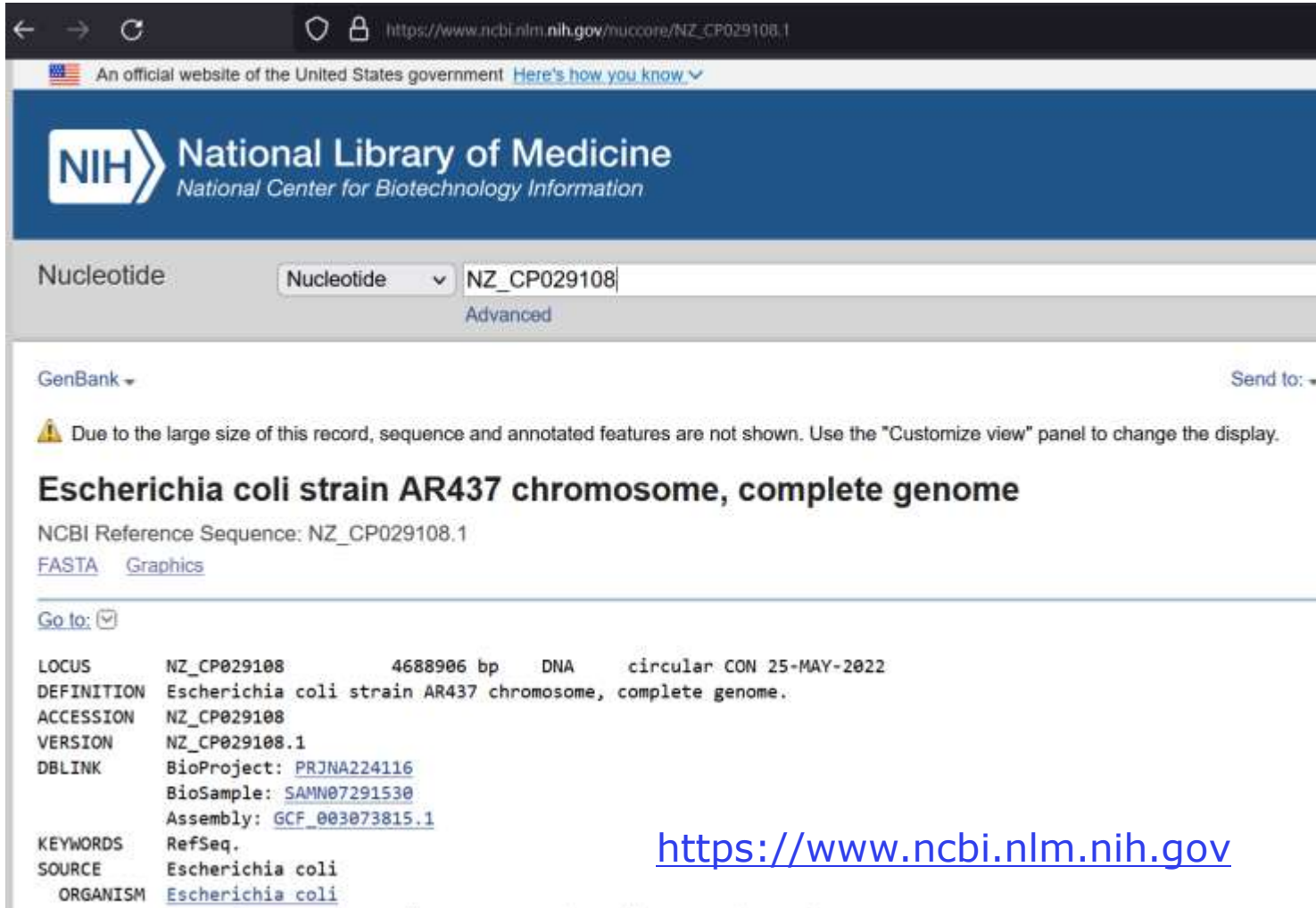
KmerFinder – find a reference

KmerFinder-3.2 Server - Results


KmerFinder 3.2 results:

| Template | Num | Score | Expected | Template_length | Query_Coverage | Template_Coverage | Depth | tot_query_Coverage | tot_template |
|---|-------|---------|----------|-----------------|----------------|-------------------|-------|--------------------|--------------|
| NZ_CP029108.1 Escherichia coli strain AR437 chromosome, complete genome | 14538 | 7191229 | 231 | 154903 | 82.45 | 99.04 | 46.42 | 82.45 | 99.04 |
| NZ_CP018991.1 Escherichia coli strain Ecol_AZ146 chromosome, complete genome | 18701 | 168049 | 2651 | 181206 | 1.93 | 3.19 | 0.93 | 49.86 | 51.43 |
| NZ_CP083869.1 Escherichia coli strain NDM6 chromosome, complete genome | 24430 | 68824 | 2318 | 156510 | 0.79 | 1.20 | 0.44 | 64.63 | 76.67 |
| NZ_CP080139.1 Escherichia coli strain PK8241 chromosome, complete genome | 2178 | 32981 | 2655 | 184405 | 0.38 | 1.21 | 0.18 | 65.23 | 68.71 |
| NZ_CP031653.1 Escherichia coli strain UK_Dog_Liverpool chromosome, complete genome | 9127 | 27836 | 2406 | 161066 | 0.32 | 1.00 | 0.17 | 81.94 | 95.45 |
| NC_011586.2 Acinetobacter baumannii AB0057, complete genome | 18517 | 6592 | 2266 | 152543 | 0.08 | 1.98 | 0.04 | 0.54 | 2.13 |

Download reference




https://www.ncbi.nlm.nih.gov/nucleotide/NZ_CP029108.1
 An official website of the United States government [Here's how you know.](#)


National Library of Medicine
National Center for Biotechnology Information

Nucleotide
 Advanced

GenBank

 Due to the large size of this record, sequence and annotated features are not shown. Use the "Customize view" panel to change the display.

Escherichia coli strain AR437 chromosome, complete genome

NCBI Reference Sequence: NZ_CP029108.1
[FASTA](#) [Graphics](#)

[Go to:](#)

| | | | | | |
|------------|--|------------|-----|----------|-----------------|
| LOCUS | NZ_CP029108 | 4688906 bp | DNA | circular | CON 25-MAY-2022 |
| DEFINITION | Escherichia coli strain AR437 chromosome, complete genome. | | | | |
| ACCESSION | NZ_CP029108 | | | | |
| VERSION | NZ_CP029108.1 | | | | |
| DBLINK | BioProject: PRJNA224116 | | | | |
| | BioSample: SAMN07291530 | | | | |
| | Assembly: GCF_003073815.1 | | | | |
| KEYWORDS | RefSeq. | | | | |
| SOURCE | Escherichia coli | | | | |
| ORGANISM | Escherichia coli | | | | |

<https://www.ncbi.nlm.nih.gov>

Guidelines – how to get started

- Iterative process – what should be included?
 - Inclusion criteria
 - Time/origin/other epi-data
 - Resistance phenotype or gene
 - Sub-typing
 - Pre-analysis - Fast check for overview
 - Include all isolates
 - Consider choice of reference & settings of tool
 - Fasta files
 - Refine analysis
 - Exclude most distant isolates
 - Run analysis on fastq files

Pre-analysis

- All strains
- Fasta files – faster analysis
- Test of pruning 10 vs 100

Input data

Upload reference genome (fasta format)
Note: Reference genome must not be compressed.

Ingen fil valgt.

Include reference in final phylogeny.

Select min. depth at SNP positions

Select min. relative depth at SNP positions

Select minimum distance between SNPs (prune)

Select min. SNP quality

Select min. read mapping quality

Select min. Z-score

Ignore heterozygous SNPs

Comment (to yourself)
This comment will appear unaltered on your output page. It has no effect on the analysis.

Use altered FastTree (more accurate)
Note: Read more [here](#)

Upload read files and/or assembled genomes (fasta or fastq format)
Note: Read files must be compressed with gzip (compressed files often ends with .gz).
 If you get an "Access forbidden. Error 403": Make sure the start of the web address is https and not just http. Fix it by clicking [here](#).

| Name | Size | Progress |
|------|------|----------|
| | | |

Analysis 1 output: Illumina assembly

CSIPhylogeny Results

The tree presented in the picture below is only meant as a preview. If the tree is meant to be shared or published and processed using software created for this purpose. We suggest ([FigTree](#)).

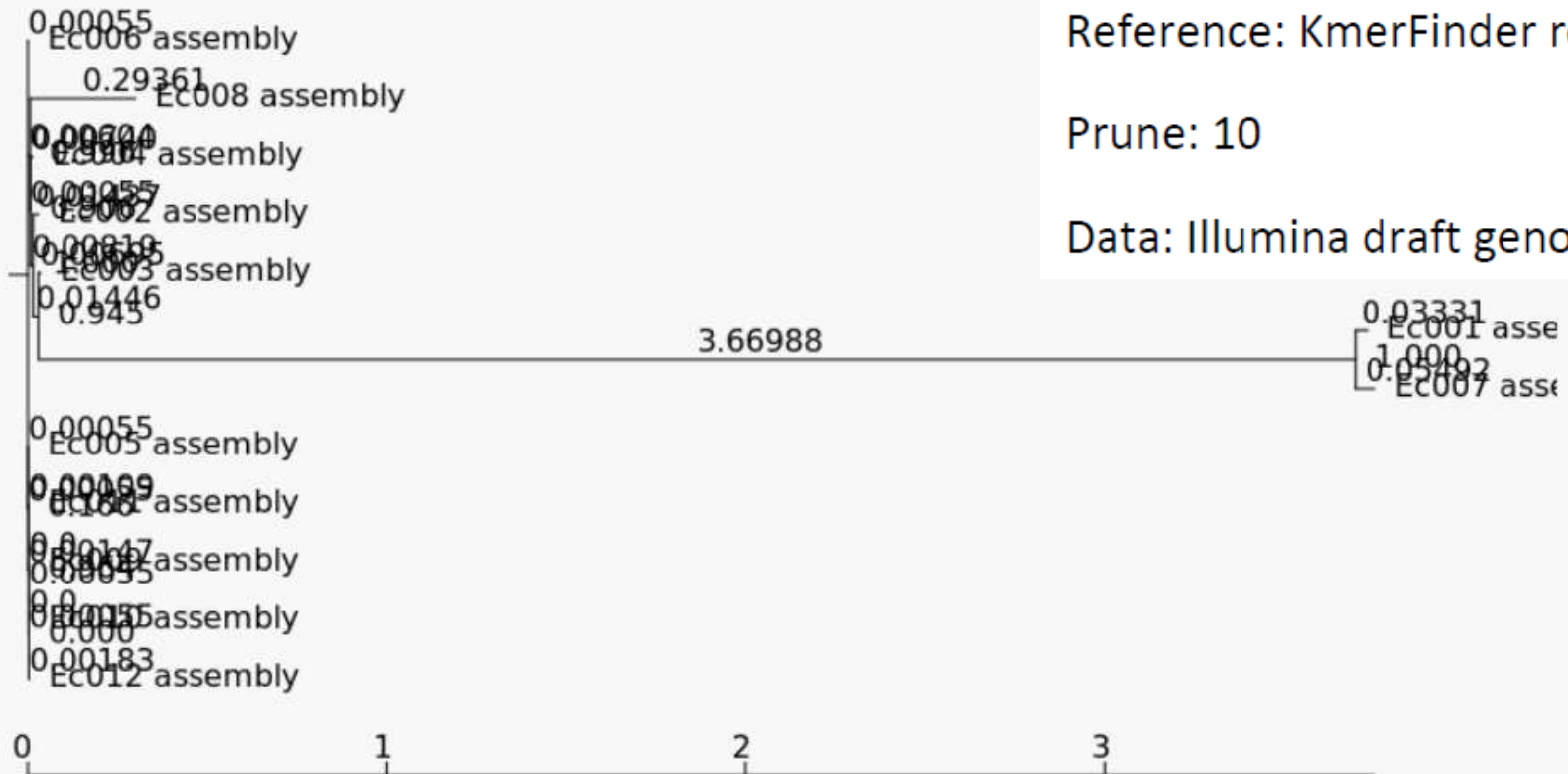
Analysis 1

Tool: CSI Phylogeny

Reference: KmerFinder reference

Prune: 10

Data: Illumina draft genomes (all 12 isolates)



Download phylogeny as:

Analysis 1 output: Illumina assembly

Percentage of reference genome covered by all isolates: 90.018466029857

4383433 positions was found in all analyzed genomes.

Size of reference genome: 4869482

Below is listed the number of positions that are shared and trusted between each isolate and the reference genome.

| File | Valid positions | Pct. of reference |
|-----------------------------|-----------------|-------------------|
| Ec002_assembly.ignored_snps | 4582041 | 94.0970928735336 |
| Ec003_assembly.ignored_snps | 4593110 | 94.324406579591 |
| Ec001_assembly.ignored_snps | 4612609 | 94.7248393155576 |
| Ec009_assembly.ignored_snps | 4544249 | 93.3209938962707 |
| Ec005_assembly.ignored_snps | 4607895 | 94.6280323040521 |
| Ec004_assembly.ignored_snps | 4607871 | 94.6275394384865 |
| Ec008_assembly.ignored_snps | 4578451 | 94.0233683993493 |
| Ec007_assembly.ignored_snps | 4622929 | 94.936771508756 |
| Ec011_assembly.ignored_snps | 4567232 | 93.792974283507 |
| Ec012_assembly.ignored_snps | 4586595 | 94.1906141146019 |
| Ec006_assembly.ignored_snps | 4610605 | 94.6836850408319 |
| Ec010_assembly.ignored_snps | 4570946 | 93.8692452297801 |

Visualisation of tree (newick file)



We cannot determine clustering based on the visual tree only

SNP matrix – pairwise comparison of SNPs

A1 prune 10

| | Ec001_assembly/1-2735 | Ec002_assembly/1-2735 | Ec003_assembly/1-2735 | Ec004_assembly/1-2735 | Ec005_assembly/1-2735 | Ec006_assembly/1-2735 | Ec007_assembly/1-2735 | Ec008_assembly/1-2735 | Ec009_assembly/1-2735 | Ec010_assembly/1-2735 | Ec011_assembly/1-2735 | Ec012_assembly/1-2735 |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Ec001_assembly/1-2735 | 0 | 2176 | 2122 | 2180 | 2176 | 2176 | 216 | 2280 | 2179 | 2179 | 2182 | 2184 |
| Ec002_assembly/1-2735 | 2176 | 0 | 94 | 80 | 78 | 78 | 2212 | 644 | 81 | 81 | 84 | 86 |
| Ec003_assembly/1-2735 | 2122 | 94 | 0 | 98 | 96 | 96 | 2170 | 662 | 99 | 99 | 102 | 104 |
| Ec004_assembly/1-2735 | 2180 | 80 | 98 | 0 | 38 | 38 | 2222 | 604 | 41 | 41 | 44 | 46 |
| Ec005_assembly/1-2735 | 2176 | 78 | 96 | 38 | 0 | 2 | 2218 | 598 | 5 | 5 | 8 | 10 |
| Ec006_assembly/1-2735 | 2176 | 78 | 96 | 38 | 2 | 0 | 2218 | 598 | 5 | 5 | 8 | 10 |
| Ec007_assembly/1-2735 | 216 | 2212 | 2170 | 2222 | 2218 | 2218 | 0 | 2322 | 2221 | 2221 | 2224 | 2226 |
| Ec008_assembly/1-2735 | 2280 | 644 | 662 | 604 | 598 | 598 | 2322 | 0 | 601 | 601 | 604 | 606 |
| Ec009_assembly/1-2735 | 2179 | 81 | 99 | 41 | 5 | 5 | 2221 | 601 | 0 | 0 | 3 | 5 |
| Ec010_assembly/1-2735 | 2179 | 81 | 99 | 41 | 5 | 5 | 2221 | 601 | 0 | 0 | 3 | 5 |
| Ec011_assembly/1-2735 | 2182 | 84 | 102 | 44 | 8 | 8 | 2224 | 604 | 3 | 3 | 0 | 8 |
| Ec012_assembly/1-2735 | 2184 | 86 | 104 | 46 | 10 | 10 | 2226 | 606 | 5 | 5 | 8 | 0 |

min: 0 max: 2322

Analysis 2: Pruning 10 vs 100

Percentage of reference genome covered by all isolates: 90.018466029857

4383433 positions was found in all analyzed genomes.

Size of reference genome: 4869482

Below is listed the number of positions that are shared and trusted between each isolate and the reference genome.

| File | Valid positions | Pct. of reference |
|-----------------------------|-----------------|-------------------|
| Ec008_assembly.ignored_snps | 4578451 | 94.0233683993493 |
| Ec006_assembly.ignored_snps | 4610605 | 94.6836850408319 |
| Ec007_assembly.ignored_snps | 4622929 | 94.936771508756 |
| Ec012_assembly.ignored_snps | 4586595 | 94.1906141146019 |
| Ec010_assembly.ignored_snps | 4570946 | 93.8692452297801 |
| Ec001_assembly.ignored_snps | 4612609 | 94.7248393155576 |
| Ec011_assembly.ignored_snps | 4567232 | 93.792974283507 |
| Ec004_assembly.ignored_snps | 4607871 | 94.6275394384865 |
| Ec009_assembly.ignored_snps | 4544249 | 93.3209938962707 |
| Ec005_assembly.ignored_snps | 4607895 | 94.6280323040521 |
| Ec003_assembly.ignored_snps | 4593110 | 94.324406579591 |
| Ec002_assembly.ignored_snps | 4582041 | 94.0970928735336 |

Different pruning (100) -> different number of SNPs

Pruning 10: min: 0 max: 2322 vs. Pruning 100: min:3 max: 516

| A2 prune 100 | Ec001_assembly/1-676 | Ec002_assembly/1-676 | Ec003_assembly/1-676 | Ec004_assembly/1-676 | Ec005_assembly/1-676 | Ec006_assembly/1-676 | Ec007_assembly/1-676 | Ec008_assembly/1-676 | Ec009_assembly/1-676 | Ec010_assembly/1-676 | Ec011_assembly/1-676 | Ec012_assembly/1-676 |
|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| Ec001_assembly/1-676 | 0 | 478 | 461 | 471 | 468 | 469 | 111 | 495 | 473 | 472 | 473 | 479 |
| Ec002_assembly/1-676 | 478 | 0 | 45 | 47 | 44 | 45 | 499 | 121 | 49 | 48 | 49 | 55 |
| Ec003_assembly/1-676 | 461 | 45 | 0 | 38 | 35 | 36 | 482 | 112 | 40 | 39 | 40 | 46 |
| Ec004_assembly/1-676 | 471 | 47 | 38 | 0 | 35 | 36 | 492 | 112 | 40 | 39 | 40 | 46 |
| Ec005_assembly/1-676 | 468 | 44 | 35 | 35 | 0 | 3 | 489 | 105 | 7 | 6 | 7 | 13 |
| Ec006_assembly/1-676 | 469 | 45 | 36 | 36 | 3 | 0 | 490 | 106 | 8 | 7 | 8 | 14 |
| Ec007_assembly/1-676 | 111 | 499 | 482 | 492 | 489 | 490 | 0 | 516 | 494 | 493 | 494 | 500 |
| Ec008_assembly/1-676 | 495 | 121 | 112 | 112 | 105 | 106 | 516 | 0 | 106 | 109 | 110 | 114 |
| Ec009_assembly/1-676 | 473 | 49 | 40 | 40 | 7 | 8 | 494 | 106 | 0 | 3 | 4 | 8 |
| Ec010_assembly/1-676 | 472 | 48 | 39 | 39 | 6 | 7 | 493 | 109 | 3 | 0 | 3 | 9 |
| Ec011_assembly/1-676 | 473 | 49 | 40 | 40 | 7 | 8 | 494 | 110 | 4 | 3 | 0 | 10 |
| Ec012_assembly/1-676 | 479 | 55 | 46 | 46 | 13 | 14 | 500 | 114 | 8 | 9 | 10 | 0 |

Analysis 3: fasta vs fastq

Percentage of reference genome covered by all isolates: 91.0057168298394
4431507 positions was found in all analyzed genomes.
Size of reference genome: 4869482

| File | Valid positions | Pct. of reference |
|--|-----------------|-------------------|
| Ec011.illumina_R1.trimmed.ignored_snps | 4602885 | 94.5251466172377 |
| Ec006.illumina_R1.trimmed.ignored_snps | 4649224 | 95.4767673440419 |
| Ec001.illumina_R1.trimmed.ignored_snps | 4650803 | 95.5091937910439 |
| Ec009.illumina_R1.trimmed.ignored_snps | 4557370 | 93.590447608185 |
| Ec003.illumina_R1.trimmed.ignored_snps | 4652967 | 95.5536338362068 |
| Ec005.illumina_R1.trimmed.ignored_snps | 4650928 | 95.5117607991979 |
| Ec004.illumina_R1.trimmed.ignored_snps | 4623318 | 94.9447600381314 |
| Ec012.illumina_R1.trimmed.ignored_snps | 4602711 | 94.5215733418873 |
| Ec008.illumina_R1.trimmed.ignored_snps | 4620255 | 94.8818580703245 |
| Ec010.illumina_R1.trimmed.ignored_snps | 4598637 | 94.4379094121305 |
| Ec002.illumina_R1.trimmed.ignored_snps | 4639182 | 95.2705441769782 |
| Ec007.illumina_R1.trimmed.ignored_snps | 4673824 | 95.981954548759 |

Fasta vs fastq -> different number of SNPs

Fasta: min:3 max: 516 vs. fastq: min: 0 max 413

Remove 3 most distant strains 1,7,8

A4

| | Ec001.illumina_R1.trimmed | Ec002.illumina_R1.trimmed | Ec003.illumina_R1.trimmed | Ec004.illumina_R1.trimmed | Ec005.illumina_R1.trimmed | Ec006.illumina_R1.trimmed | Ec007.illumina_R1.trimmed | Ec008.illumina_R1.trimmed | Ec009.illumina_R1.trimmed | Ec010.illumina_R1.trimmed | Ec011.illumina_R1.trimmed | Ec012.illumina_R1.trimmed |
|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|
| Ec001.illumina_R1.trimmed | 0 | 374 | 361 | 369 | 361 | 362 | 99 | 390 | 365 | 365 | 367 | 370 |
| Ec002.illumina_R1.trimmed | 374 | 0 | 41 | 47 | 39 | 40 | 397 | 88 | 43 | 43 | 45 | 48 |
| Ec003.illumina_R1.trimmed | 361 | 41 | 0 | 36 | 28 | 29 | 384 | 77 | 32 | 32 | 34 | 37 |
| Ec004.illumina_R1.trimmed | 369 | 47 | 36 | 0 | 34 | 35 | 392 | 83 | 38 | 38 | 40 | 43 |
| Ec005.illumina_R1.trimmed | 361 | 39 | 28 | 34 | 0 | 1 | 384 | 71 | 4 | 4 | 6 | 9 |
| Ec006.illumina_R1.trimmed | 362 | 40 | 29 | 35 | 1 | 0 | 385 | 72 | 5 | 5 | 7 | 10 |
| Ec007.illumina_R1.trimmed | 99 | 397 | 384 | 392 | 384 | 385 | 0 | 413 | 388 | 388 | 390 | 393 |
| Ec008.illumina_R1.trimmed | 390 | 88 | 77 | 83 | 71 | 72 | 413 | 0 | 75 | 75 | 77 | 80 |
| Ec009.illumina_R1.trimmed | 365 | 43 | 32 | 38 | 4 | 5 | 388 | 75 | 0 | 0 | 2 | 5 |
| Ec010.illumina_R1.trimmed | 365 | 43 | 32 | 38 | 4 | 5 | 388 | 75 | 0 | 0 | 2 | 5 |
| Ec011.illumina_R1.trimmed | 367 | 45 | 34 | 40 | 6 | 7 | 390 | 77 | 2 | 2 | 0 | 7 |
| Ec012.illumina_R1.trimmed | 370 | 48 | 37 | 43 | 9 | 10 | 393 | 80 | 5 | 5 | 7 | 0 |
| min: 0 max: 413 | | | | | | | | | | | | |

Sum-up of analysis

- Improved analysis result by:
 - Pruning 100 instead of 10
 - Using fastq instead of fasta

Exercise

- Instructions will be sent out today!
- Following data will be provided
 - Raw sequencing data
 - Assembled data
 - Results of the SNP analysis performed (No requirement to run CSIPhylogeny but you will be given a link to the CSIPhylogeny output page)
- Survey with questions will be sent out and must be completed by Sep 12, 2026 (23:59 CET)