

VIRTUAL MULTIDISCIPLINARY TRAINING WORKSHOP #1

OUTBREAK EXERCISE WITH CRE MINION AND ILLUMINA DATA

September 26th 2022
Jette S. Kjeldgaard

Day 1: Monday Sept 26st 13:00-15:00

❖ Introduction

- purpose of exercise & learning objectives,
- differentiated objectives – there's something for all

❖ Introduction to typing methods

- MLST, serotyping, PlasmidFinder, (Jette)
- cgMLST, SNP analysis (Henrik)

❖ Introduction to tools for SNP analysis

- CSIphylogeny (Jette)
- MinTyper (Henrik)

❖ Information about exercise data and tasks (Jette)

❖ Questions



Day 2: Monday Oct 10th 13:00-15:00

- Presentation of exercise results (cluster analysis)
- Use of different tools
- Additional analyses (cgMLST, plasmids, resistance genes)
- Questions/comments and other results from participants
- Same Teams link – chat should be available between meetings if you have installed Teams

- ❖ To get prepared to start on outbreak investigations – (if you are not already 😊)
 - background information about bacterial subtyping and cluster analysis
 - suggestions for online available analytical tools to get started on bacterial comparison and outbreak detection
- ❖ To learn more about the difference between Minion vs Illumina data and the consequences in analysis
- ❖ See examples of typing and characterisation possibilities by online tools and try working with online tools by yourself



- You will be able to perform a cluster analysis of bacteria to look into possible relatedness in a dataset
- You will learn about other relevant tools for sequence analysis and work with a small data set to characterise
- You will apply the results from the cluster analysis and the additional analyses to elucidate a possible hospital outbreak of CCRE



Data set on carbapenemase 12 producing E. coli (OXA-48-like and/or NDM)

❖ Cluster analysis

- Either on Illumina files, on MinION files or on both types

❖ Further characterisation

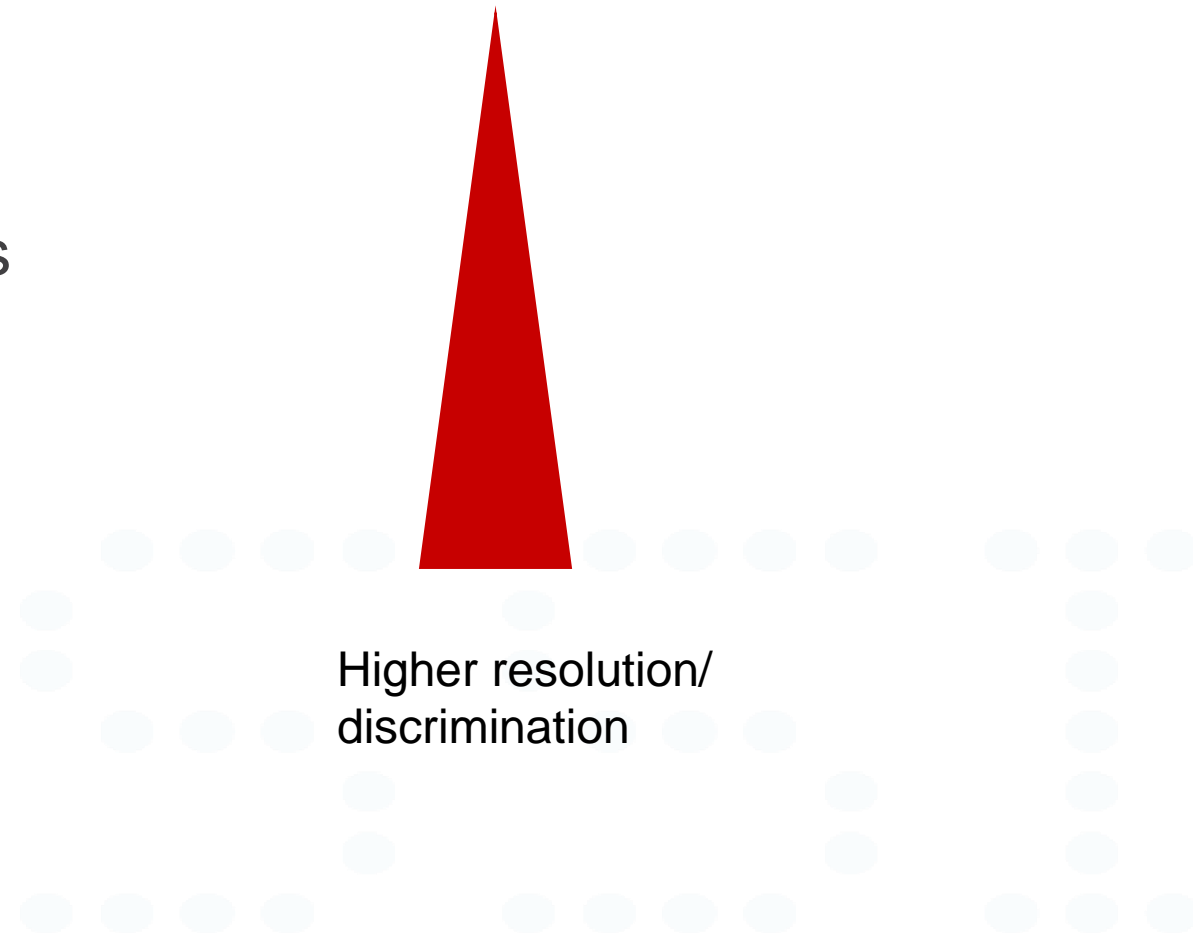
- cgMLST
- ResFinder (AMR prediction)
- PlasmidFinder/MGEFinder (Plasmid prediction)
- (Serotyping by WGS)



- ❖ You will hear more about the different types of sequencing in the course in December
- ❖ Illumina
 - Short reads (ca 150 bp)
 - Low error rate
- ❖ ONT MinION
 - Long reads (up to approx. 200 kbp)
 - Higher error rate
 - Raw reads can be trimmed to remove poor sequence reads
 - Reads contain information on sequence quality (fastq files)
 - Reads are assembled into contigs, which are much longer stretches of DNA
 - Assembly files contain less information and are 'smaller' to work with (fasta files)

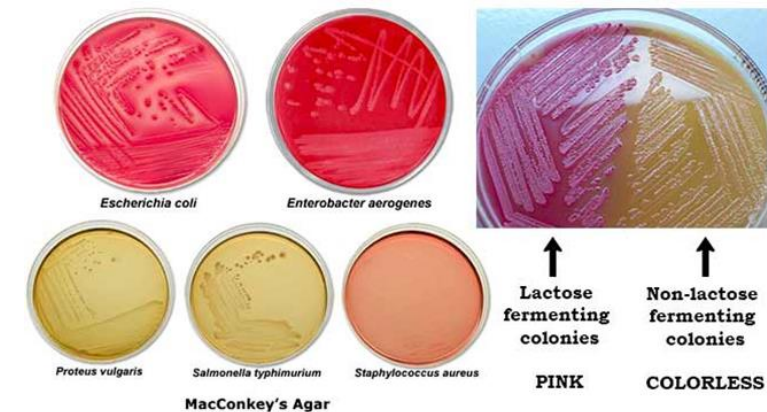
- ✚ Information of bacteria below species level
 - Outbreak detection, clusters, common contamination source, transmission routes..
 - *E. coli*/Salmonella - traditional subtyping
 - serotyping using antisera against the ca. 186 O-antigens and 53 H-flagellar antigens for *E. coli* or 46 O-antigens and 114 H-antigens for *Salmonella* (ca 2600 serovars)
 - Requires anti-sera and trained personnel
 - time consuming and not always accurate
 - Phagotyping
 - Golden standard method for surveillance of *Salmonella* Typhimurium and *S. Enteritidis* – also used for *E. coli* and other bacteria
 - requires a comprehensive panel of different phages, considerable technical expertise

- ❖ Purpose of subtyping?
- ❖ Genus/Species determination
- ❖ Serotyping and MLST
 - Characterisation and grouping of isolates
- ❖ cgMLST and SNP analysis
 - Comparison
- ❖ Resistance patterns
- ❖ pMLST – plasmids
- ❖ Specific genes or combinations

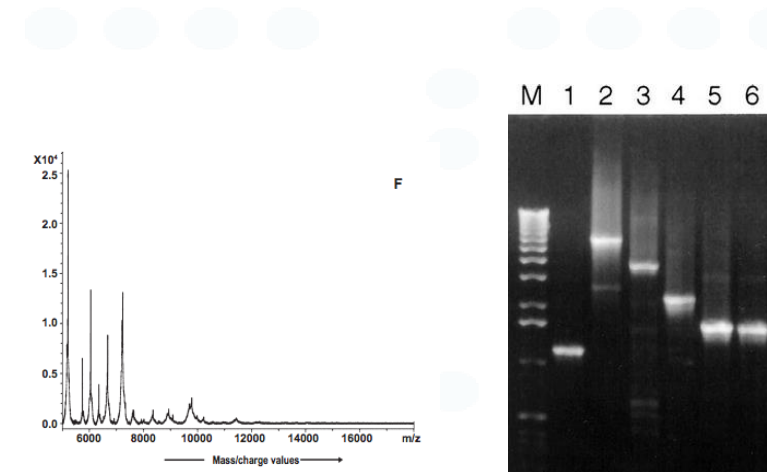


❖ How do you normally do typing in the lab?

- Phenotypic identification
 - Biochemical/metabolic analysis
 - Chromogenic media
 - AST
 - CIM test



- Molecular identification
 - PCR (genus/species/AST)
 - MALDI-TOF MS
 - Microarray (AMR)
 - MLST (PCR/Sequencing)



❖ Species

- KmerFinder (full genome) and/or SpeciesFinder (16s rRNA)

❖ Sub-typing

- Serotyping (*E. coli*, *P. aeruginosa*, *Salmonella*)

❖ Typing

- MLST
- cgMLSTFinder
 - *Campylobacter*, *Clostridium*, *E. coli*, *Listeria*, *Salmonella*, *Yersinia*
- pMLST

❖ Cluster analysis

- CSIPhylogeny & MinTyper

❖ Species

- **KmerFinder** (full genome) and/or SpeciesFinder (16s rRNA)

❖ Sub-typing

- Serotyping (*E. coli*, *P. aeruginosa*, *Salmonella*)

❖ Typing

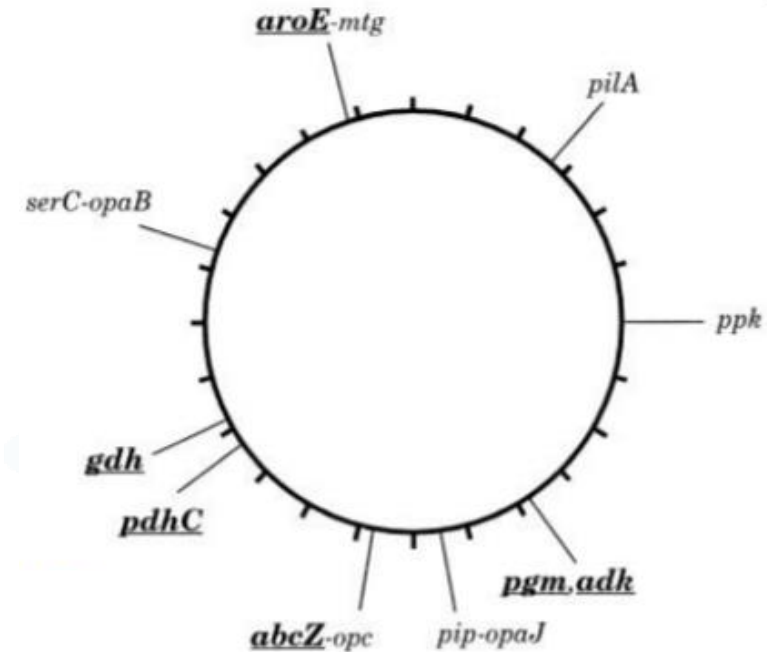
- **MLST**
- **cgMLSTFinder**
 - *Campylobacter*, *Clostridium*, *E. coli*, *Listeria*, *Salmonella*, *Yersinia*
- pMLST

❖ Cluster analysis

- **CSIPhylogeny & MinTyper**

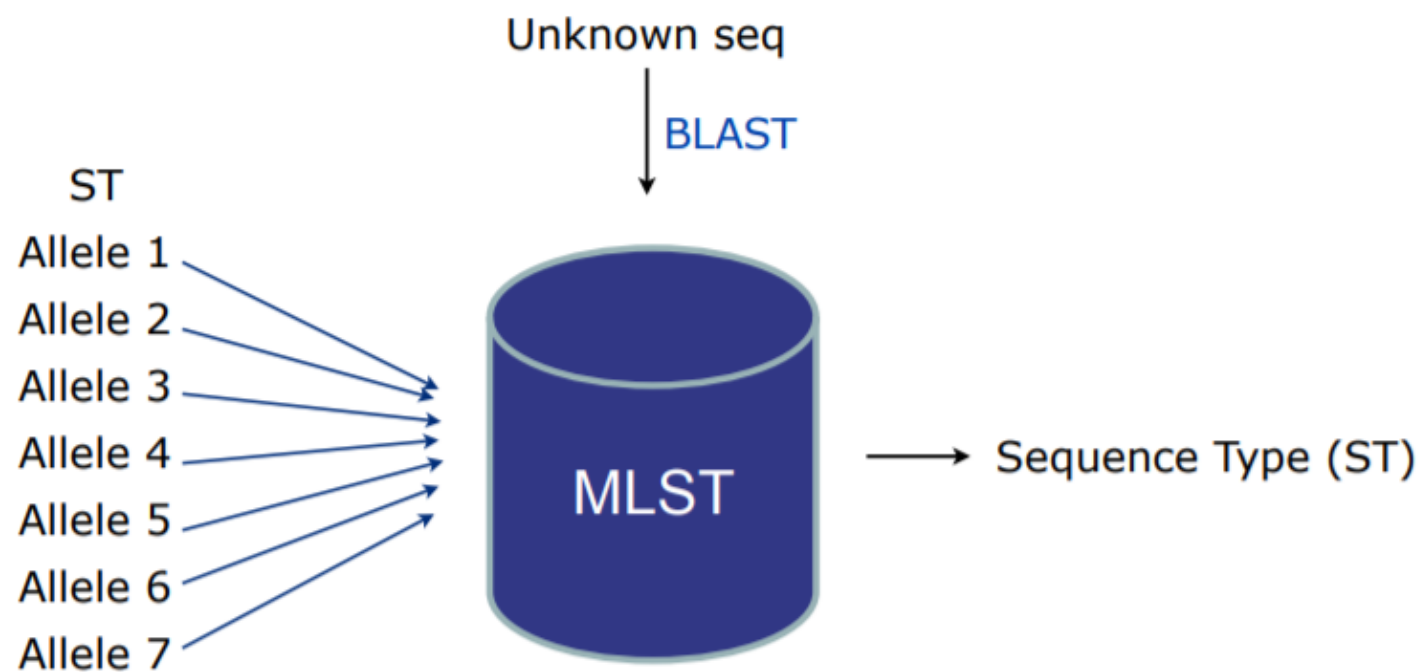
Classical MLST:

- The (new) golden standard for typing
- First developed in 1998 for *Neisseria meningitis* (Maiden et al. PNAS 1998. 95:3140-3145)
- The nucleotide sequence of internal regions of app. 7 housekeeping genes are determined by PCR followed by Sanger sequencing
- Different alleles are each assigned a random number
- The unique combination of alleles is the sequence type (ST)



- ❖ For many bacterial species, MLST is considered the gold standard of typing
 - It is traditionally performed in an expensive and time-consuming way
- ❖ As the cost of WGS continue to decline, it becomes increasingly available to scientists and routine diagnostics laboratories
 - Currently, the WGS cost is typically below that of traditional MLST

7 x PCR and sequencing vs. 1 x WGS



MLST-2.0 Server - Results

mlst Profile: *Imonocytogenes*

Organism: *Listeria monocytogenes*

Sequence Type: 6

Locus	Identity	Coverage	Alignment Length	Allele Length	Gaps	Allele
abcZ	100	100	537	537	0	abcZ_3
bglA	100	100	399	399	0	bglA_9
cat	100	100	486	486	0	cat_9
dapE	100	100	462	462	0	dapE_3
dat	100	100	471	471	0	dat_3
ldh	100	100	453	453	0	ldh_1
lhkA	100	100	480	480	0	lhkA_5

extended output

Input Files: *Lm02.fa*

One limitation: ONE variation in bases of one of the seven genes: new allele number = different ST

Why limit to SEVEN genes when we sequence the whole genome?
-> core genome MLST

- ❖ Single Nucleotide Polymorphism (SNP)
 - Require reference genome

- ❖ Gene-by-gene approach
 - cgMLST – core genome MLST/wgMLST - whole genome MLST
 - No reference genome required
 - Require species specific cgMLST scheme

- ❖ What is phylogeny used for?
 - Classify taxonomy – the classic use
 - Outbreak detection – detection of clones – increasing with WGS data

- ❖ Core genome = genes common for all (almost) within the species
 - *E. coli* has approx. 5000-5500 genes, hereof 2300 are selected for the cgMLST scheme

Locus	Identity	Coverage	Alignment Length	Allele Length	Gaps	Allele
abcZ	100	100	537	537	0	abcZ_3
bglA	100	100	399	399	0	bglA_9
cat	100	100	486	486	0	cat_9
dapE	100	100	462	462	0	dapE_3
dat	100	100	471	471	0	dat_3
ldh	100	100	453	453	0	ldh_1
lhkA	100	100	480	480	0	lhkA_5

Gene08						
Gene09						
Gene10						
Gene11						
Gene12						
Gene13						
Gene14						
Gene15						
Gene16						
Gene17						
Gene18						
Gene19						
Gene20						

Each gene variant has an allele number

Each allele combination has a **cg ST** assigned based on the cgMLST scheme



By cgMLST very closely related genomes are 'lumped' together in a Complex Type (CT)

Can also be used to interpret clusters

••Part 2 – CSIPhylogeny



Table 1 Metadata for the 12 carbapenemase producing *E. coli* isolates

Species	Date	Region of isolation	Travel	MLST	Sequence	Carba genotype (PCR)
<i>E. coli</i>	2015	Copenhagen	Pakistan	ST410	Ec001	OXA-48-like
<i>E. coli</i>	2015	Copenhagen	Thailand	ST410	Ec002	OXA-48-like
<i>E. coli</i>	2015	Jutland - M	India	ST410	Ec003	NDM
<i>E. coli</i>	2015	Copenhagen	Lebanon	ST410	Ec004	OXA-48-like
<i>E. coli</i>	2016	Zealand	No	ST410	Ec005	NDM, OXA-48-like
<i>E. coli</i>	2016	Zealand	No	ST410	Ec006	NDM, OXA-48-like
<i>E. coli</i>	2017	Copenhagen	Pakistan	ST410	Ec007	OXA-48-like
<i>E. coli</i>	2018	Jutland - N	Thailand	ST410	Ec008	NDM
<i>E. coli</i>	2018	Zealand	No	ST410	Ec009	NDM, OXA-48-like
<i>E. coli</i>	2018	Zealand	No	ST410	Ec010	NDM, OXA-48-like
<i>E. coli</i>	2018	Zealand	No	ST410	Ec011	NDM
<i>E. coli</i>	2018	Zealand	No	ST410	Ec012	OXA-48-like

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. Your laboratory has just finalized setting up MinION (Oxford Nanopore; ONT) sequencing, and you wish to use this occasion to work with both types of sequences.

- ❖ Phylogenetic comparisons allow for determining clusters and clonal spread of microorganisms
- ❖ SNP calling – to determine variants in the DNA (Single Nucleotide Polymorphism)
- ❖ Different sequencing technology has a systematic bias making integration of data generated from different platforms difficult.
 - CSIPhylogeny has incorporated two different procedures for identifying variable sites and inferring phylogenies in WGS data across multiple platforms

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.

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

- ❖ 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 ($\geq 30\times$)
- ❖ Consider the quality of your raw data (specifically phred scores)

You will hear more about
sequence quality on the course
in December

❖ 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!**

- ❖ Calling of single nucleotide polymorphism
 - Variants in the DNA – compared to reference

....ATCGAATTCCGGGTTTAAACCGGATCGTACGATCGGGAAAAA..

TTCCAGG

TTCCAGG

TTCCAGG

TTCCAGG

TTCCAGG

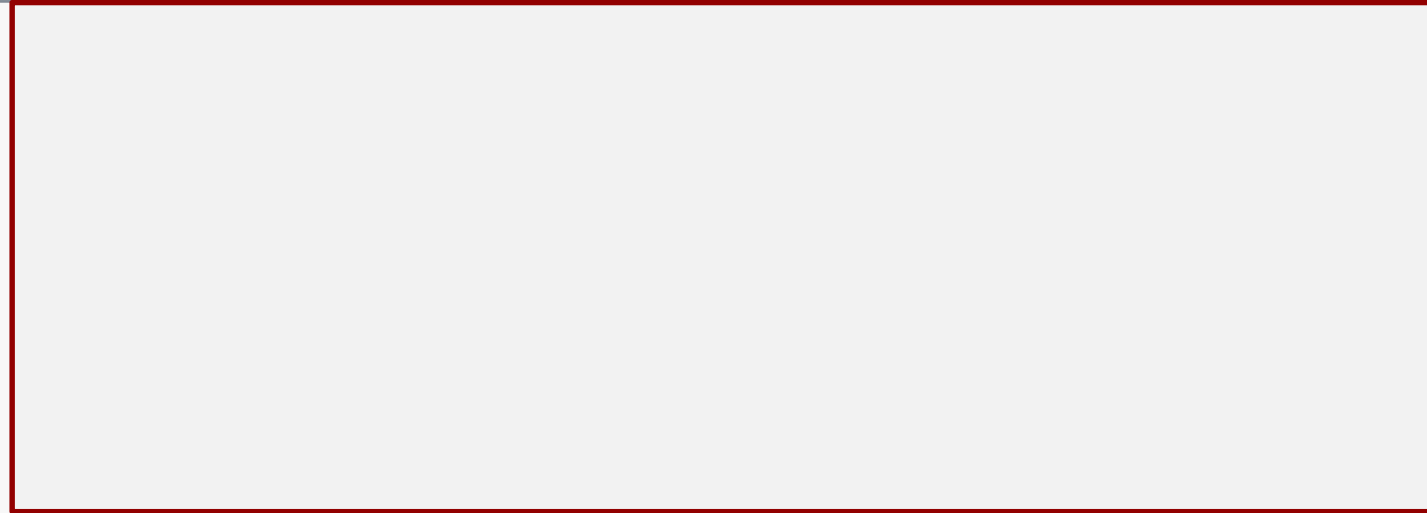
TTCCAGG

SNPs are called on **the nucleotides which all isolates in the analysis share** with the reference.

Higer variation between isolates = higher difference from reference

->

Decreacing amount of nucleotides to call SNPs from
(Valid positions/ percentage of reference covered)



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
1.96

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
1.96

☐ 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

....ATCGAATTCCGGGTTTTTAACCGGATCGTACGATCGGGAAAAA..

TTCCAGGTTTTTTAACCAGATCG

TTCCAGGTTTTTTAACCAGATCG

TTCCAGGTTTTTTAACCAGATCG

TTCCAGGTTTTTTAACCAGATCG

TTCCAGGTTTTTTAACCAGATCG

TTCCAGGTTTTTTAACCAGATCG

11 bp

Input data

Upload reference genome (fasta format)

Note: Reference genome must not be compressed.

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

1.96

☐ 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 j

 Isolate File

❖ Input data:

❖ Reference: Must be fasta format

- Choice of reference impacts the result

Warning!:Uploading too many files can make the job failed...

❖ Additional sequences:

- Can be both fasta and fastq (Illumina)
 - fastq most accurate

OUTPUT: VARIANT CALLING FORMAT (VCF)

- Lists of SNPs called for each sequence, compared to the reference

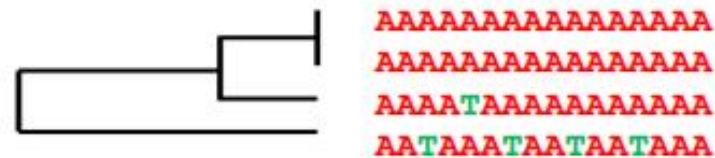
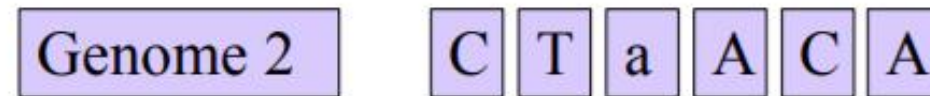
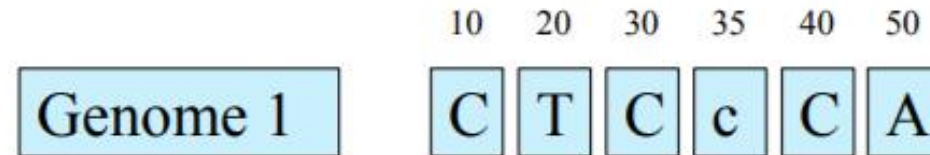
Genome 1	position	ref	change
Ref_genome	10	T	C
Ref_genome	20	C	T
Ref_genome	30	A	C
Ref_genome	40	A	C
Ref_genome	50	G	A

Genome 2	position	ref	change
Ref_genome	10	T	C
Ref_genome	20	C	T
Ref_genome	35	C	A
Ref_genome	40	A	C
Ref_genome	50	G	A

- ❖ Output: Variant calling format (VCF)

Genome 1	position	ref	change
Ref_genome	10	T	C
Ref_genome	20	C	T
Ref_genome	30	A	C
Ref_genome	40	A	C
Ref_genome	50	G	A

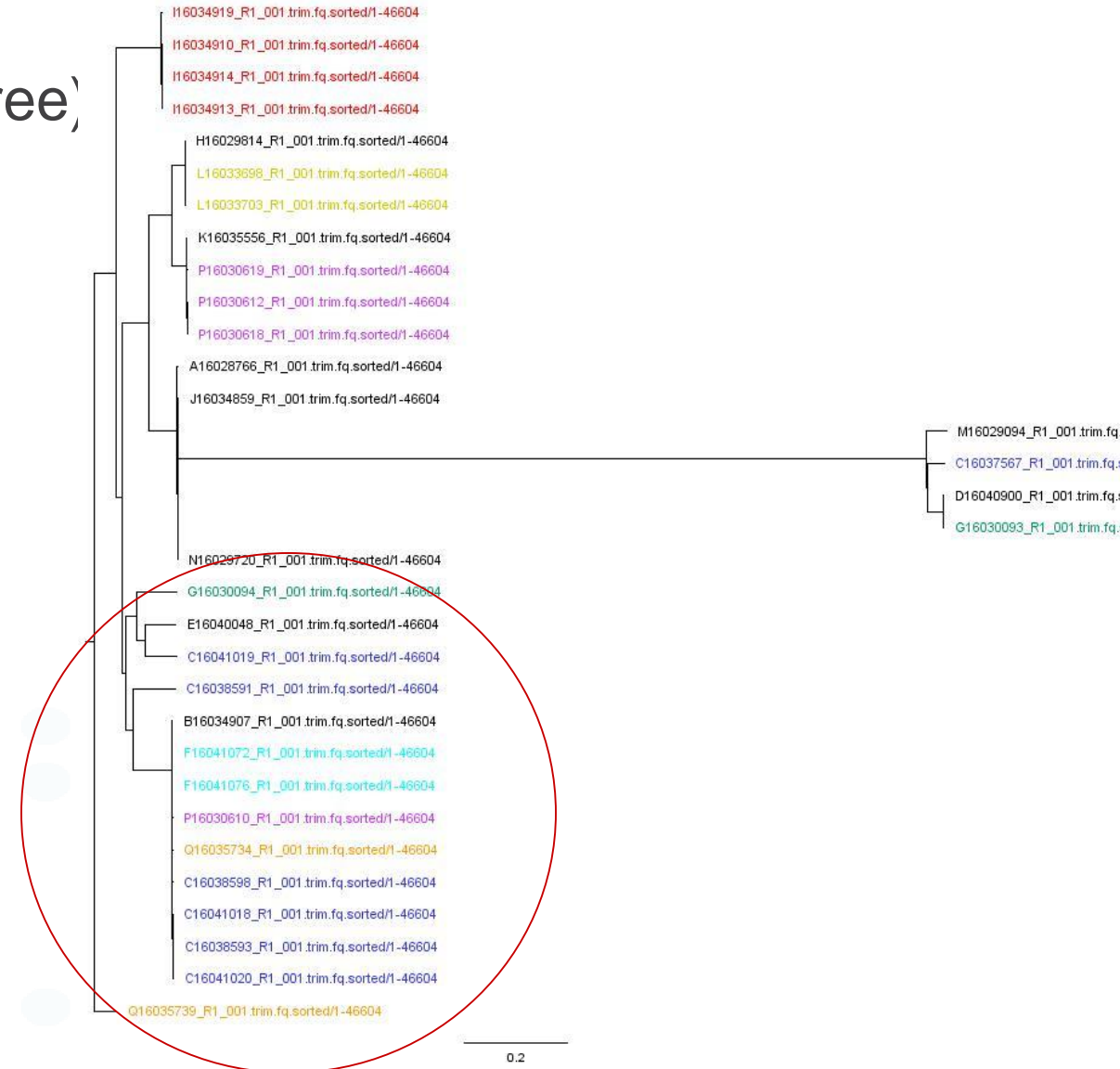
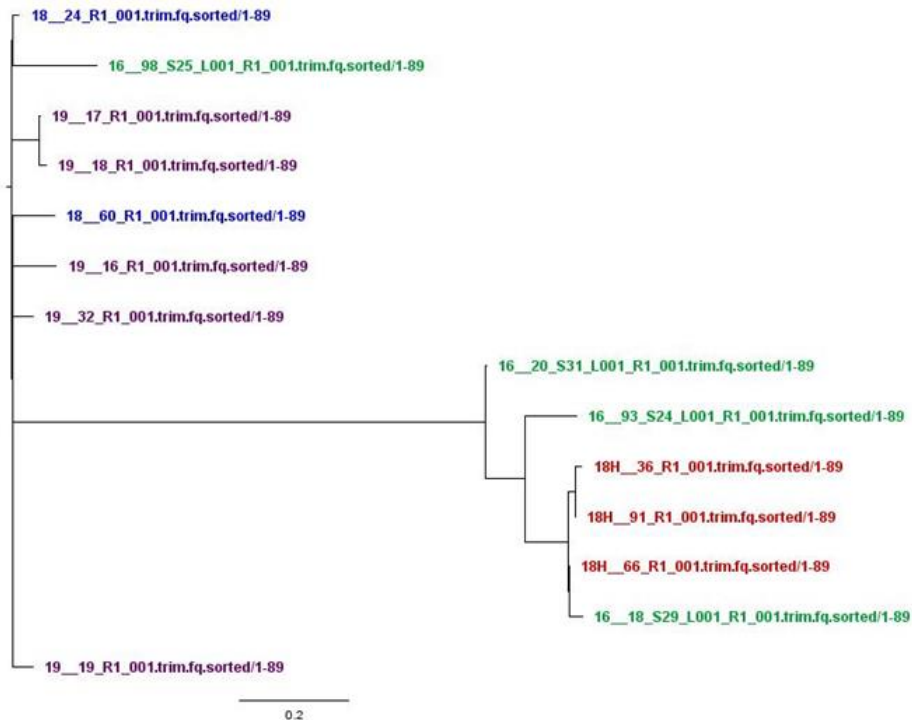
Genome 2	position	ref	change
Ref_genome	10	T	C
Ref_genome	20	C	T
Ref_genome	35	C	A
Ref_genome	40	A	C
Ref_genome	50	G	A



SNP matrix – pairwise comparison of SNPs

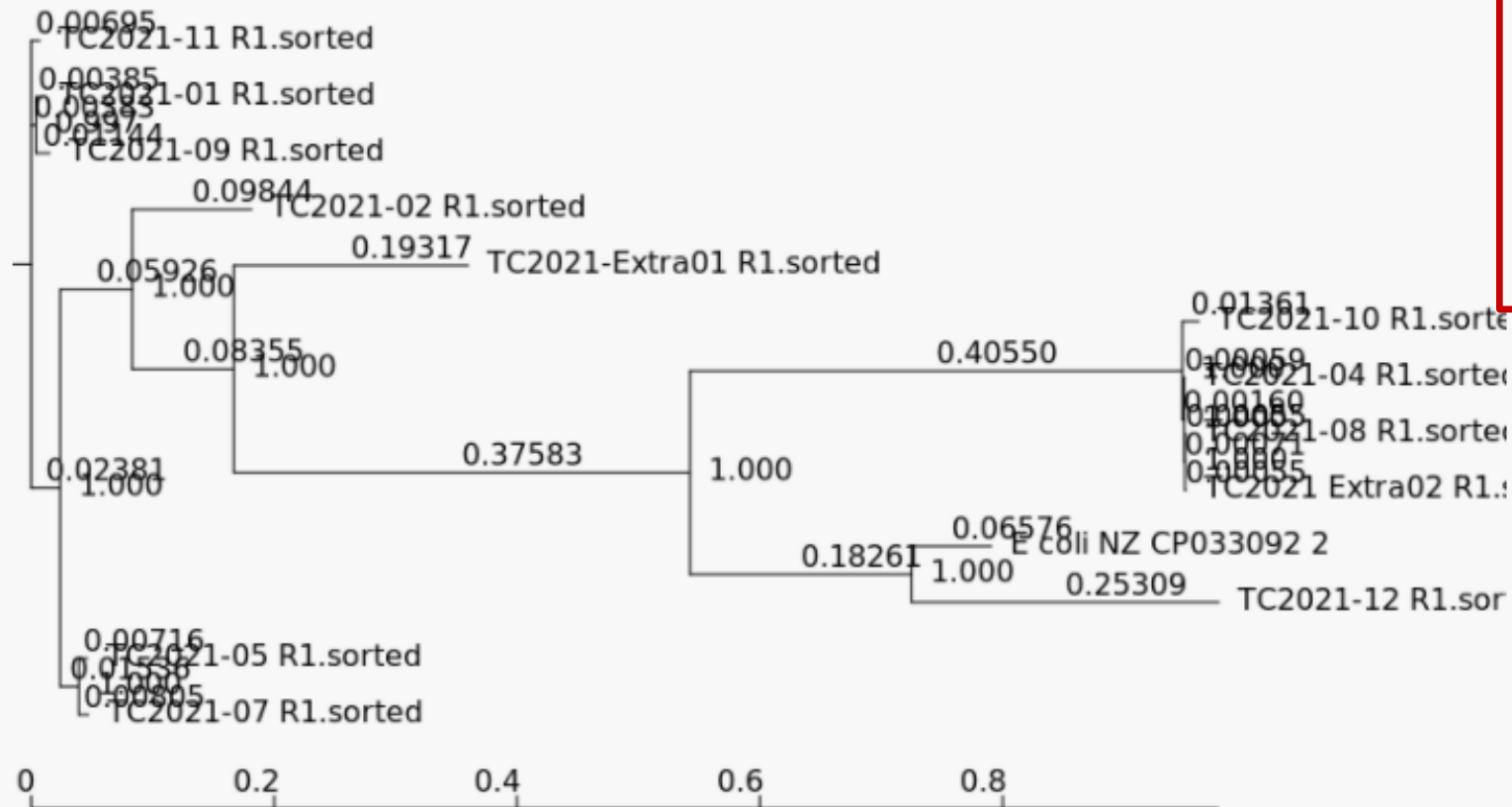
	Strain A	Strain B	Strain C	Strain D	Strain E	Strain F	Strain G	Strain H	
Strain A		0	406	223	388	326	212	324	321
Strain B	406		0	140	51	458	279	459	455
Strain C	223	140		0	12	259	85	259	255
Strain D	388	51	12		0	431	257	432	428
Strain E	326	458	259	431		0	328	6	5
Strain F	212	279	85	257	328		0	329	322
Strain G	324	459	259	432	6	329		0	9
Strain H	321	455	255	428	5	322	9		0

- ❖ Newick file – distance file: phylogeny
 - Visualise by various tools (here: by FigTree)
 - Distance measured on horizontal lines
 - No/short distance = clustering
 - It's a matter of perspective!



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

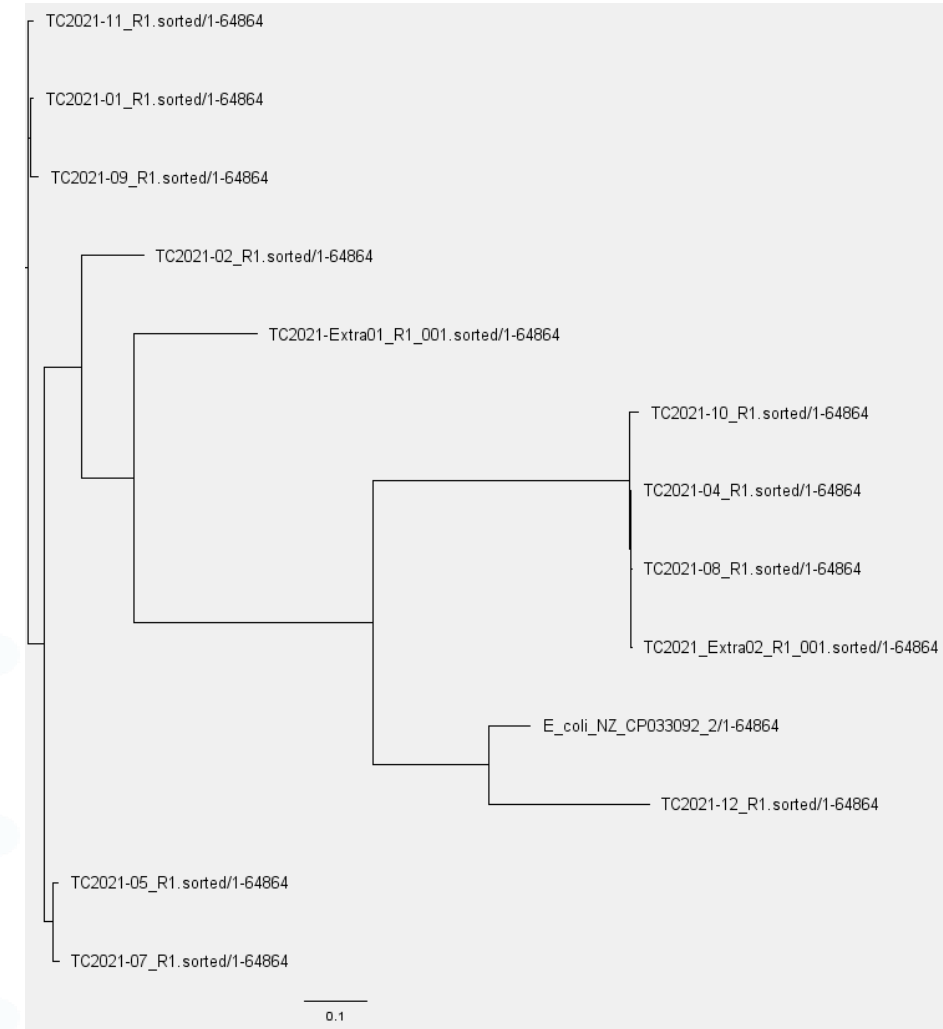
Download matrix of SNP pair counts:

Download matrix as:

Download SNP alignment:

Download phylogeny as:

- ❖ Text file – SNP distances
- ❖ Use various tools to visualise the phylogenetic tree
- ❖ Text file – SNP distances
- ❖ Use various tools to visualise the phylogenetic tree
- ❖ Here: FigTree
- ❖ <https://github.com/rambaut/figtree/releases>
- ❖ CGE tool:
 - TreeViewer
- ❖ Microreact, iTOL...
 - <https://microreact.org/upload>



Percentage of reference genome covered by all isolates: 71.4734023710814
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



❖ Plain text file – open in excel

	E_coli_NZ_C P033092_2	TC2021- 01_	TC2021- 02_	TC2021- 04_	TC2021- 05_	TC2021- 07_	TC2021- 08_	TC2021- 09_	TC2021- 10_	TC2021- 11_	TC2021- 12_	TC2021- Extra01_	TC2021_Extra0 2_
E_coli_NZ_CP03309 2_2	0	29753	30187	26060	29484	29404	26067	29809	26510	29744	15477	30541	26071
TC2021-01_	29753	0	10003	32323	3125	3150	32332	932	32333	862	34921	16898	32336
TC2021-02_	30187	10003	0	32549	9519	9603	32558	10011	32548	10017	35335	17244	32562
TC2021-04_	26060	32323	32549	0	32270	32180	80	32312	962	32425	30575	32712	84
TC2021-05_	29484	3125	9519	32270	0	928	32279	3222	32278	3113	34970	17024	32283
TC2021-07_	29404	3150	9603	32180	928	0	32189	3266	32192	3170	34872	16949	32193
TC2021-08_	26067	32332	32558	80	32279	32189	0	32321	970	32434	30577	32718	4
TC2021-09_	29809	932	10011	32312	3222	3266	32321	0	32322	1309	34977	16753	32325
TC2021-10_	26510	32333	32548	962	32278	32192	970	32322	0	32433	30997	32698	974
TC2021-11_	29744	862	10017	32425	3113	3170	32434	1309	32433	0	34925	16930	32438
TC2021-12_	15477	34921	35335	30575	34970	34872	30577	34977	30997	34925	0	35612	30581
TC2021-Extra01_	30541	16898	17244	32712	17024	16949	32718	16753	32698	16930	35612	0	32722
TC2021_Extra02_	26071	32336	32562	84	32283	32193	4	32325	974	32438	30581	32722	0
min: 4 max: 35612													

SNP MATRIX - EXAMPLE

	E_coli_NZ_C P033092_2	TC2021- 01_	TC2021- 02_	TC2021- 04_	TC2021- 05_	TC2021- 07_	TC2021- 08_	TC2021- 09_	TC2021- 10_	TC2021- 11_	TC2021- 12_	TC2021- Extra01_	TC2021_Extra0 2_
E_coli_NZ_CP03309 2_2	0												
TC2021-01_	29753	0							Below	1000	SNPs		
TC2021-02_	30187	10003	0						Below	100	SNPs		
TC2021-04_	26060	32323	32549	0					Below	10	SNPs		
TC2021-05_	29484	3125	9519	32270	0								
TC2021-07_	29404	3150	9603	32180	928	0							
TC2021-08_	26067	32332	32558	80	32279	32189	0						
TC2021-09_	29809	932	10011	32312	3222	3266	32321	0					
TC2021-10_	26510	32333	32548	962	32278	32192	970	32322	0				
TC2021-11_	29744	862	10017	32425	3113	3170	32434	1309	32433	0			
TC2021-12_	15477	34921	35335	30575	34970	34872	30577	34977	30997	34925	0		
TC2021-Extra01_	30541	16898	17244	32712	17024	16949	32718	16753	32698	16930	35612	0	
TC2021_Extra02_	26071	32336	32562	84	32283	32193	4	32325	974	32438	30581	32722	0
min: 4 max: 35612													

- ❖ 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 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 adress 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><div></div></div>	
Ec001.illumina_R2.trimmed.fastq.gz	96.00 MB	<div><div></div></div>	

Upload

Remove

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

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

← → ↻ 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 ▾ [Send to:](#) ▾

⚠ 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				

For this exercise:

We have uploaded 2
reference sequences on
Sciencedata.dk:
One is the best match found
by KmerFinder
(KmerFinder_ref)

Another is index isolate,
hybrid assembled and
published (optimal_ref)

KmerFinder_ref.fasta






[Optimal_ref.fasta](#)

❖ Back to Henrik and MinTyper



SO – WHAT TO DO?

- ❖ We have distributed the sequence data by sciencedata.dk
- ❖ <https://sciencedata.dk/shared/007e3242ab05e33a01de62cf24ce8eda>

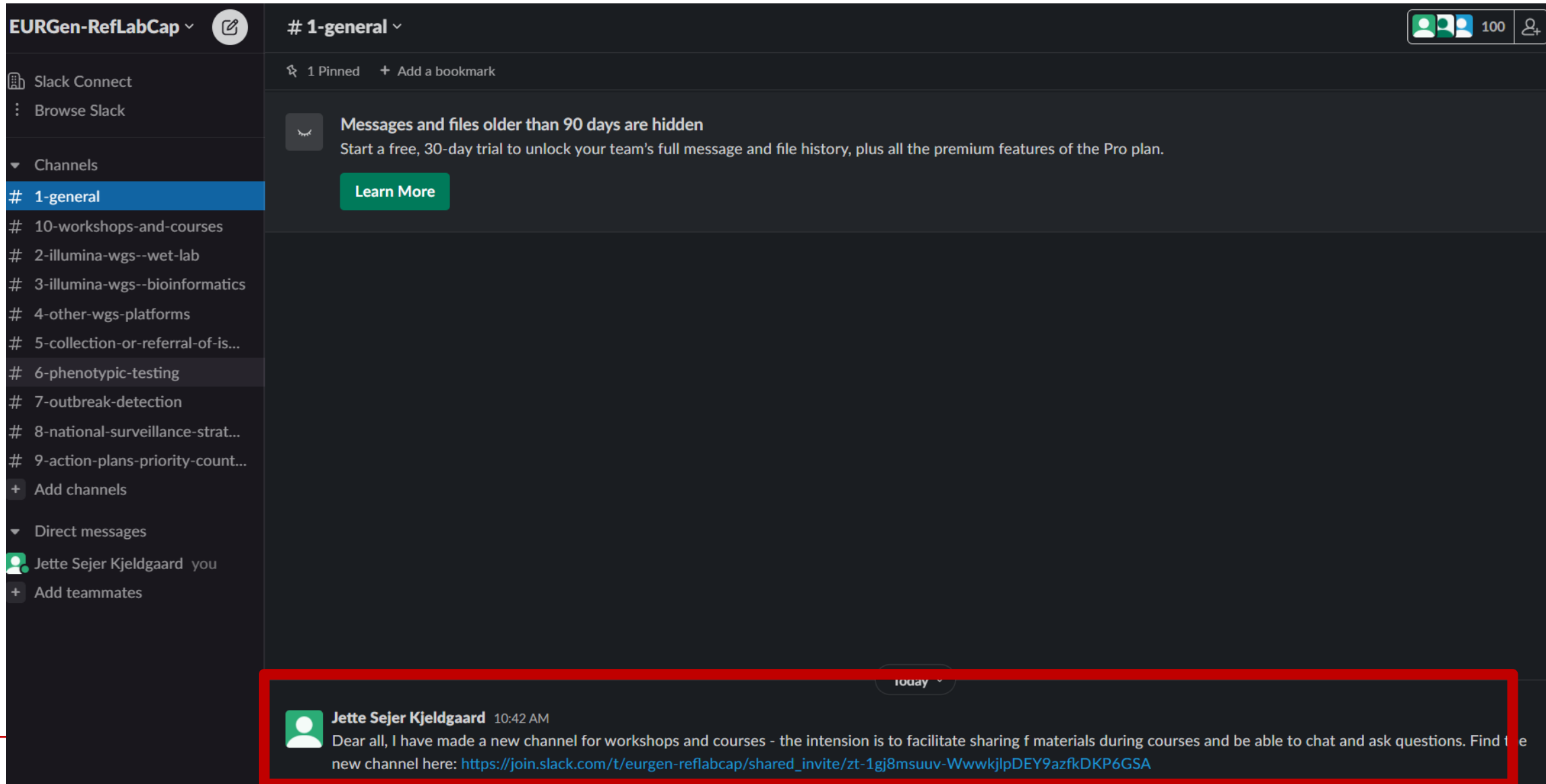
<input type="checkbox"/>	Name ▲	Size
<input type="checkbox"/>	 ONT fastq	5.6 GB
<input type="checkbox"/>	 References	9.3 MB
<input type="checkbox"/>	 Illumina.zip	3.9 GB
<input type="checkbox"/>	 Illumina_Assemblies.zip	14.9 MB
<input type="checkbox"/>	 ONT_assemblies.zip	15.4 MB
2 folders and 3 files		9.6 GB

- ❖ We will post slides and additional material on Slack





SLACK – FILE SHARE AND DISCUSSION FORUM

❖ https://join.slack.com/t/eurgen-reflabcap/shared_invite/zt-1gj8msuuv-WwwkjlPDEY9azfkDKP6GSA



The screenshot shows the Slack interface for the workspace 'EURGen-RefLabCap'. The left sidebar contains a list of channels, with '# 1-general' selected. Below the channels list are options for 'Direct messages' and 'Add teammates'. The main area displays the '# 1-general' channel, which is currently empty except for a pinned message. The pinned message is a system message stating: 'Messages and files older than 90 days are hidden. Start a free, 30-day trial to unlock your team's full message and file history, plus all the premium features of the Pro plan. Learn More'. At the bottom of the screen, a message from 'Jette Sejer Kjeldgaard' is visible, dated '10:42 AM'. The message content is: 'Dear all, I have made a new channel for workshops and courses - the intension is to facilitate sharing f materials during courses and be able to chat and ask questions. Find the new channel here: https://join.slack.com/t/eurgen-reflabcap/shared_invite/zt-1gj8msuuv-WwwkjlPDEY9azfkDKP6GSA'. The message is highlighted with a red border.

EURGen-RefLabCap 

1-general 

1 Pinned + Add a bookmark

Messages and files older than 90 days are hidden
Start a free, 30-day trial to unlock your team's full message and file history, plus all the premium features of the Pro plan.
[Learn More](#)

Channels

- # 1-general
- # 10-workshops-and-courses
- # 2-illumina-wgs--wet-lab
- # 3-illumina-wgs--bioinformatics
- # 4-other-wgs-platforms
- # 5-collection-or-referral-of-is...
- # 6-phenotypic-testing
- # 7-outbreak-detection
- # 8-national-surveillance-strat...
- # 9-action-plans-priority-count...


+ Add channels

Direct messages

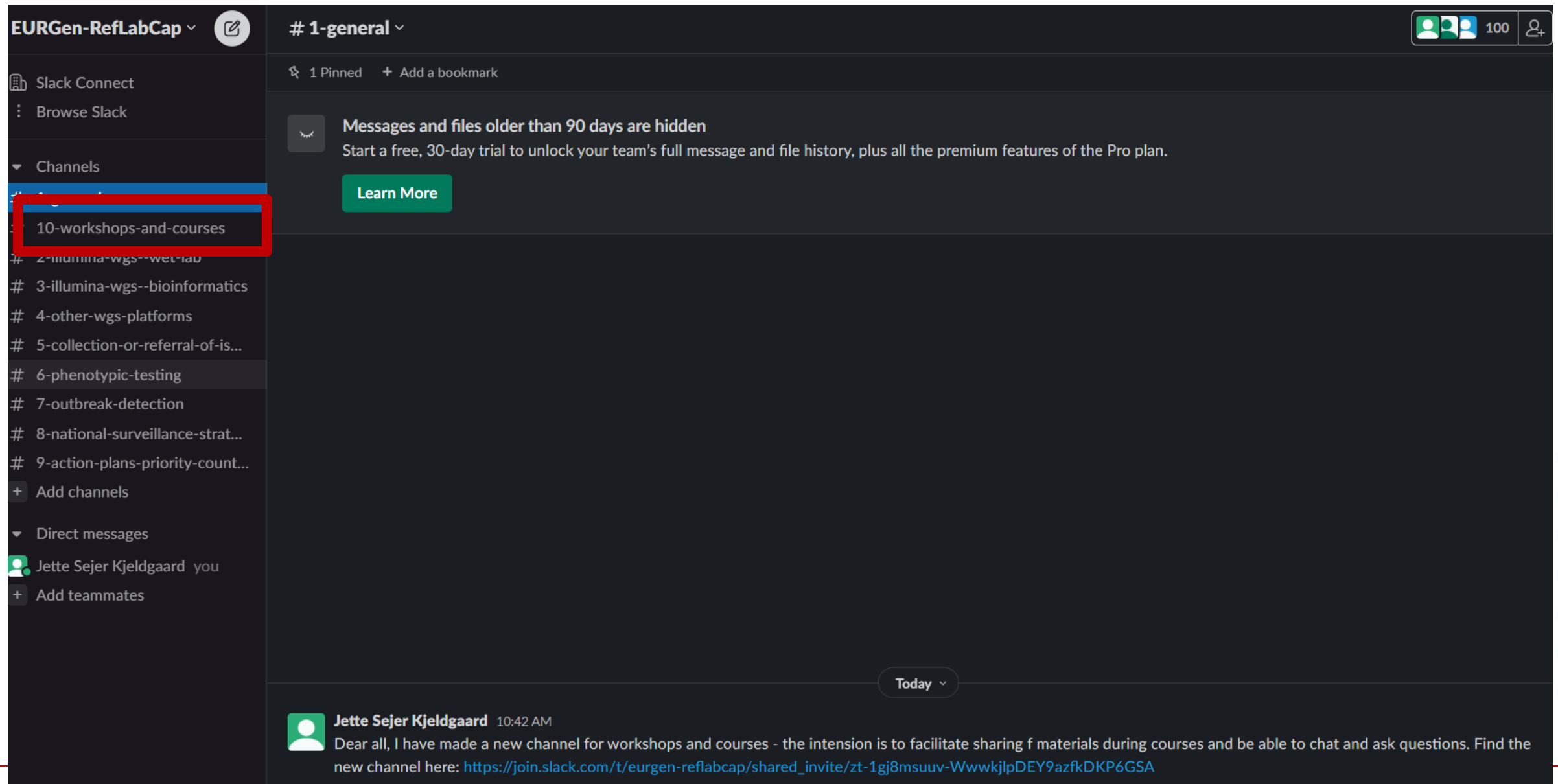
Jette Sejer Kjeldgaard you


+ Add teammates

today

 **Jette Sejer Kjeldgaard** 10:42 AM
Dear all, I have made a new channel for workshops and courses - the intension is to facilitate sharing f materials during courses and be able to chat and ask questions. Find the new channel here: https://join.slack.com/t/eurgen-reflabcap/shared_invite/zt-1gj8msuuv-WwwkjlPDEY9azfkDKP6GSA

SLACK – FILE SHARE AND DISCUSSION FORUM



EURGen-RefLabCap 


Slack Connect
Browse Slack

Channels


- # 1-general
- # 2-illumina-wgs--wet-lab
- # 3-illumina-wgs--bioinformatics
- # 4-other-wgs-platforms
- # 5-collection-or-referral-of-is...
- # 6-phenotypic-testing
- # 7-outbreak-detection
- # 8-national-surveillance-strat...
- # 9-action-plans-priority-count...
- + Add channels


Direct messages


- Jette Sejer Kjeldgaard you
- + Add teammates

1-general 

1 Pinned + Add a bookmark

 Messages and files older than 90 days are hidden
Start a free, 30-day trial to unlock your team's full message and file history, plus all the premium features of the Pro plan.
[Learn More](#)

Today 

 **Jette Sejer Kjeldgaard** 10:42 AM
Dear all, I have made a new channel for workshops and courses - the intension is to facilitate sharing f materials during courses and be able to chat and ask questions. Find the new channel here: https://join.slack.com/t/eurgen-reflabcap/shared_invite/zt-1gj8msuuv-WwwwkjlPDEY9azfkDKP6GSA

Tasks:

The main task in this exercise is to perform cluster analysis of either Illumina data alone or together with MinION data for comparison, and use the results hereof together with the metadata in table 1 to elucidate possible outbreak isolates in the dataset. There is no need to perform the analysis on both fasta and fastq files from the same sequencing technology, so chose the format that is most suitable for your analysis setup and connection. During the first day of the exercise, we will give an introduction to the tools, how to select a reference and how to work with and interpret the results.

Suggested tools, which will be presented during the first day (Sept 26th):

CSiphylogeny (Illumina data only): <https://cge.food.dtu.dk/services/CSIPhylogeny/>

MinTyper (Illumina AND MinION data): <https://cge.food.dtu.dk/services/MINTyper/>

Evaluation:

There will be no hand-in of analysis results from the participants. We will give a walk-through of the cluster analysis results on the second day of the exercise (Oct 10th), including considerations on additional analyses like identification of resistance genes and plasmids, and subtyping by cgMLST. There will be time for questions, and participants are welcome to show their results from e.g. their own pipelines etc.

Guidelines to how to get started when you have a potential outbreak

A SNP analysis is in most cases performed to examine the clonal relations of isolates. The result may then be used to support further epidemiological investigations, but it rarely stands by itself.

Often, the researcher is not completely sure which of the strains are relevant. This can lead to sub-optimal comparisons, as it in essence does not make sense to compare isolates which turns out to be very different. SNP analysis can therefore often be performed where the most distantly related isolates are removed before the next round of analysis is performed. Not to say that all non-cluster isolates should be removed, but it is convenient to have one or more "outgroup" isolates to put the outbreak in context, but genomes with more than approximately 500-1000 SNPs distant from the reference considered to be removed before a rerun of the remaining isolates to utilize the reference data in the analysis.

To save time in the initial analysis, draft genomes can be used to get the overall overview of the chosen isolates for further selection of the relevant genomes for analysis. However, the final analysis should preferably be made on raw sequencing data, which gives the opportunity to only use High-quality SNPs in the analysis...and prevents the risk of spot intra-species contamination of the sequencing reads. This is because

CSI Phylogeny

Analysis 1

Tool: CSI Phylogeny

Reference: KmerFinder reference

Prune: 10

Data: Illumina draft genomes (all 12 isolates)

Results: [Center for Genomic Epidemiology - Results \(dtu.dk\)](#)

Server run time (approximately): 10 minutes

Analysis 2

Tool: CSI Phylogeny

Reference: KmerFinder reference

Prune: 100

Data: Illumina draft genomes (all 12 isolates)

Results: [Center for Genomic Epidemiology - Results \(dtu.dk\)](#)

Server run time (approximately): 10 minutes

MinTyper

Analysis 8

Tool: **MinTyper**

Reference: Best reference

Prune: 100

Data: Illumina raw data (All 12 isolates)

Results: [MINTyper-1.0 - Results \(dtu.dk\)](#)

Server run time (approximately): 20-30 minutes

Analysis 8

Tool: MinTyper

Reference: Best reference

Prune: 100

Data: Illumina raw data (**9 closest related isolates only**)

Results: [MINTyper-1.0 - Results \(dtu.dk\)](#)

Server run time (approximately): 20-30 minutes

ADDITIONAL ANALYSES?

Species	Date	Region	Travel	MLST	Sequence	Carba genotype (PCR)
E. coli	2015	Copenhagen	Pakistan	ST410	Ec001	OXA-48-like
E. coli	2015	Copenhagen	Thailand	ST410	Ec002	OXA-48-like
E. coli	2015	Jutland - M	India	ST410	Ec003	NDM
E. coli	2015	Copenhagen	Lebanon	ST410	Ec004	OXA-48-like
E. coli	2016	Zealand	No	ST410	Ec005	NDM, OXA-48-like
E. coli	2016	Zealand	No	ST410	Ec006	NDM, OXA-48-like
E. coli	2017	Copenhagen	Pakistan	ST410	Ec007	OXA-48-like
E. coli	2018	Jutland - N	Thailand	ST410	Ec008	NDM
E. coli	2018	Zealand	No	ST410	Ec009	NDM, OXA-48-like
E. coli	2018	Zealand	No	ST410	Ec010	NDM, OXA-48-like
E. coli	2018	Zealand	No	ST410	Ec011	NDM
E. coli	2018	Zealand	No	ST410	Ec012	OXA-48-like

Suggestions – but
completely optional

Excel sheet with
data to fill out

Carba genotype WGS	Other resistance genes	Plasmids	Other MGEs	cgMLST	Serotype
	Tool		Illumina	Both Illumina and MinION	
	cgMLST			https://cge.food.dtu.dk/services/cgMLSTFinder/	
	ResFinder			https://cge.food.dtu.dk/services/ResFinder/	
	KmerResistance			https://cge.food.dtu.dk/services/KmerResistance/	
	PlasmidFinder			https://cge.food.dtu.dk/services/plasmidfinder/	
	MobileElementFinder			https://cge.food.dtu.dk/services/MobileElementFinder/	

- ❖ Exercise overview
- ❖ Guidelines and analysis examples
- ❖ Things to be considered... to be uploaded shortly



❖ Time to work 😊

❖ Questions and comments?

❖ Please use Slack (or Teams) for questions – or eventually contact us directly

❖ jetk@food.dtu.dk