SeqAsia

Mini exercise on interpreting outbreak investigation data including SNP phylogeny

Aug/Sep 2025

Participant Instructions

Faisal Ahmad Khan

Praissy Zefi Jeyakumar

Contents

1 BACKGROUND AND OBJECTIVE OF THE EXERCISE

The aim of this exercise is to practice the various bioinformatics analyses for outbreak investigation that were covered in SeqAsia training sessions using sequencing data provided by the SeqAsia team. The exercise will focus on MLST, detection of AMR genes, and cluster analysis using SNP based phylogeny. The exercise will provide hands-on experience for the participants who have limited or no experience of working with sequencing data and cluster analyses.

Specific objectives:

1. Train in SNP-based cluster analyses of *E. coli* isolates using CSIPhylogeny
2. Train in performing MLST for subtyping (CGE MLST tool)
3. Train in annotation of AMR (ResFinder)

2 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). MLST was performed using PCR (12 *E. coli* isolates, Achtman Scheme) and isolates have been transported to your laboratory for further analysis.

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

3 PRE-ANALYSIS PERFORMED

Analysis 1: CSIPhylogeny using Illumina assemblies (.fasta files)

Data used: Illumina assemblies (.fasta) of 12 genomes

Reference: KmerFinder Reference

CSIPhylogeny parameters: default (pruning=10)

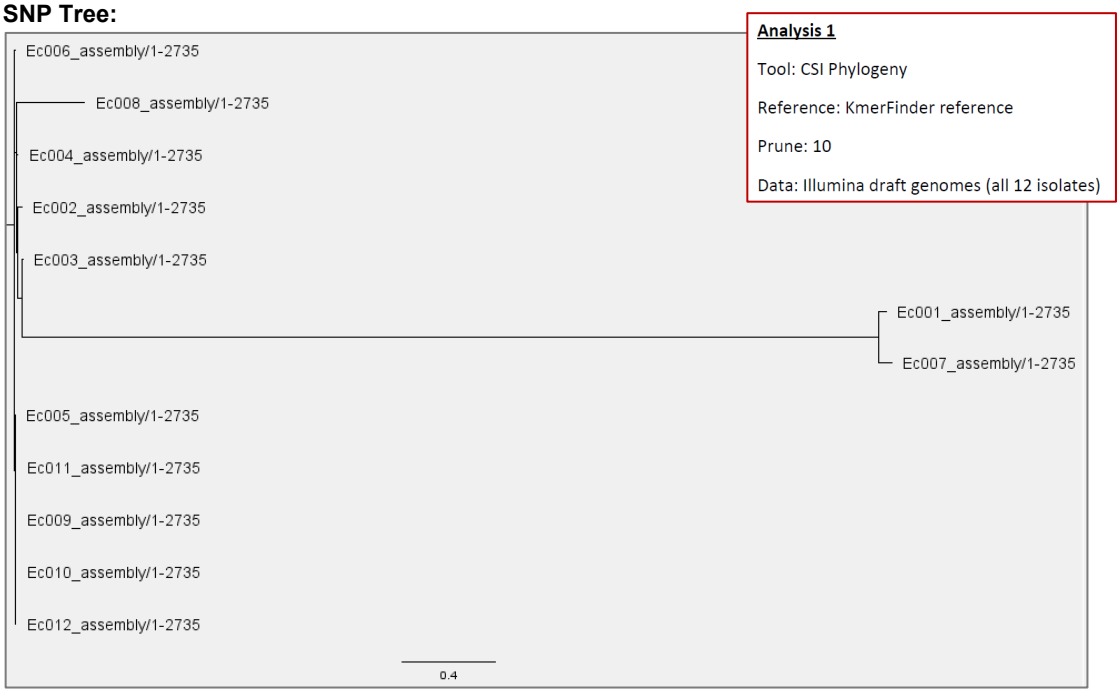
Results of Analysis 1:

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

SNP Tree:



SNP Matrix:

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: CSIPhylogeny using Illumina assemblies (.fasta files)

Data used: Illumina assemblies (.fasta) of 12 genomes

Reference: KmerFinder Reference

CSIPhylogeny parameters: default except pruning=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

SNP Tree:

SNP Matrix:

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
min: 3 max: 516												

Analysis 3: CSIPhylogeny using raw reads data (.fastq files)

Data used: Illumina raw reads (.fastq) of 12 genomes

Reference: KmerFinder Reference

CSIPhylogeny parameters: default except pruning=100

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

SNP Tree:

SNP Matrix

	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												

Analysis 4: CSIPhylogeny using raw reads data (.fastq files) after removing distant isolates

Data used: Illumina raw reads (.fastq) of 9 genomes (excluding more distant isolates)

Reference: KmerFinder Reference

CSIPhylogeny parameters: default except pruning=100

Results:

We have run the analysis for you. You can access the results page and using the results, answer the questions in the questionnaire!

Link to the analysis 4: https://cge.food.dtu.dk//cgi-bin/webface.fcgi?jobid=68C01C1B000022CABD5232D7

Please note that the analysis page will **expire** at 2025-09-16 14:22:51 (CET).

4 DATA ACCESS

Sequencing and exercise data can be accessed at:

https://sciencedata.dk/shared/033382cdb569d3331b4e8bd5fdfef34a

5 QUESTIONNAIRE

Link to the survey: https://ec.europa.eu/eusurvey/runner/01f12a5d-9831-4f2c-86f6-d31882f00b17

Deadline: 21 September 2025, 23:59 CET

Goodluck!

6 ANNEXES

Annex 1: Guidelines to how to get started with Phylogenetic analysis when you have a potential outbreak

A SNP analysis is in most cases performed to examine the clonal relationship between two or more isolates. The result may then be used to support further epidemiological investigations but can rarely stand by itself.

Often, the researcher is not completely sure which of the strains are relevant to compare, and this can lead to sub-optimal comparisons, as it in essence does not make sense to compare things, which turns out to be very different. SNP analysis can therefore often be an iterative process 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, though. Sometimes it is convenient to have one or more “outgroup” isolates to put the outbreak genomes into the right context, but genomes with more than approximately 500-1000 SNPs distance should be considered to be removed before a re-run of the remaining isolates to utilize as much as possible of the reference data in the analysis.

To save time in the initial analysis, draft genomes can be used to get the overall phylogenetic overview of the chosen isolates for further selection of the relevant genomic data before the final analysis. However, the final analysis should preferably be made on raw sequencing reads, as this gives the opportunity to only use High-quality SNPs in the analysis… and potentially also being able to spot intra-species contamination of the sequencing reads.

Most SNPs analysis tools (such as CSI Phylogeny at CGE) can only work with short reads such as those generated by Illumina sequencers because the DNA aligners (such as BWA and Bowtie) can only handle short reads. Long reads from PacBIO or Oxford Nanopore Technology (ONT) are too long to be handled and will cause the SNP tool to crash. Therefore, alternative SNP mapping analysis tools have been developed. One example is MinTyper (also at CGE). MinTyper relies on a different DNA aligner called KMA, which splits the reads into short kmers, but also still take the sequencing signal next to a given kmer hit into account, thus giving more weight to kmer hits adjacent to each other in a read, if they also are located adjacent to each other in the reference. This method is also applicable with short reads, so both short and long reads can be analyzed with MinTyper, and in principle together. However, because especially ONT long reads have systematic errors (often generated by DNA modifications such as methylation), the two data types are not always directly compatible and may group according to sequencing method rather than true phylogeny. This issue is most pronounced in older versions of the ONT Guppy basecaller and if a fast-basecalling algorithm is used. Workarounds such as masking (removing) specific methylation sites (e.g. Dcm methylation signals in *E. coli* references = CC(A/T)GG) may decrease this problem, but other bacterial species may have other DNA modification signals, which may be difficult to identify and therefore difficult to mask.