

Whole Genome Sequencing and Bioinformatics SeqAfrica Training

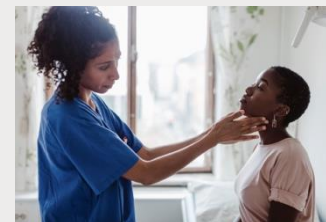
2-5 June 2025
Lusaka

Marco van Zwetselaar
Niamh Lacy-Roberts

Day 4 – Simulation Exercise



Scenario



Scenario:

- A recent rise in cases of carbapenemase producing *E. coli* in several regional hospitals indicate one or more ongoing outbreaks
- Suggested that the NRL could give assistance by performing outbreak investigation by WGS.
- Patients include both domestic and travel-related cases and a batch of samples has already been sequenced using Illumina sequencing (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.

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

Simulated exercise results

Pre-analysis

- All strains
- Fasta files – faster analysis
- Test of pruning 10 vs 100
- Test of reference kmerfinder vs optimal
- CSIPhylogeny

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
 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 just http. Fix it by clicking [here](#).

Name	Size	Progress

Depth & quality => fastq files

Analysis 1 output

Analysis 1

Tool: CSI Phylogeny

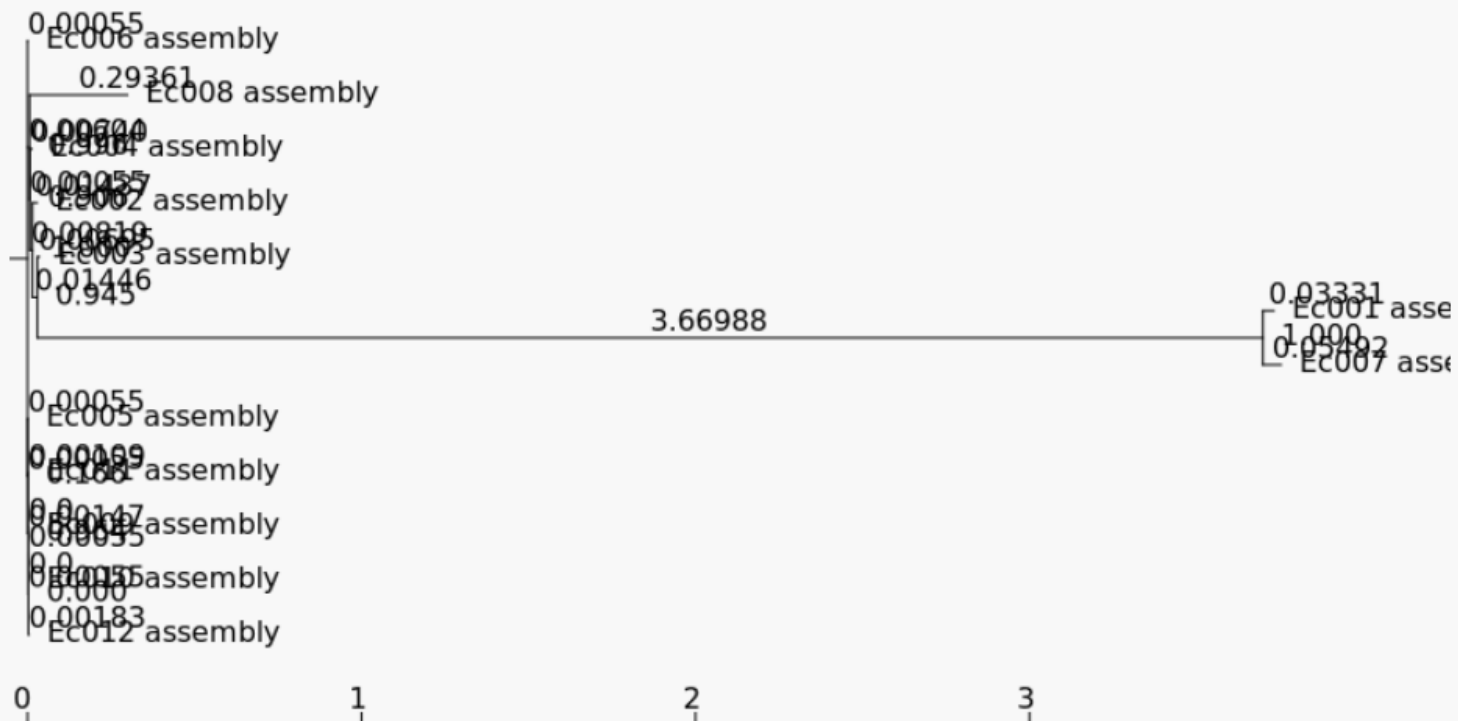
Reference: KmerFinder reference

Prune: 10

Data: Illumina draft genomes (all 12 isolates)

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](#)).



Download phylogeny as:

Newick

PDF

SVG

Analysis 1 output

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

Analysis 1 output

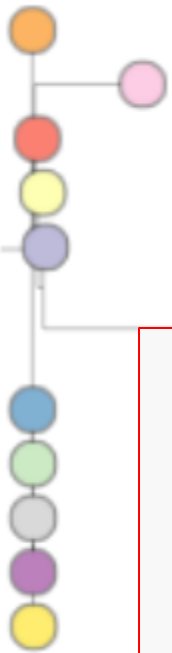
Below is listed the number of positions, covered by at least one read, in each isolate that did not meet the minimum thresholds.

Applies only for reads (fastq)

Index	Isolate	Ignored pos.
1	Ec004_assembly	0
2	Ec006_assembly	0
3	Ec005_assembly	0
4	Ec012_assembly	0
5	Ec009_assembly	0
6	Ec008_assembly	0
7	Ec010_assembly	0
8	Ec002_assembly	0
9	Ec007_assembly	0
10	Ec003_assembly	0
11	Ec001_assembly	0
12	Ec011_assembly	0

Visualisation of tree (newick file)

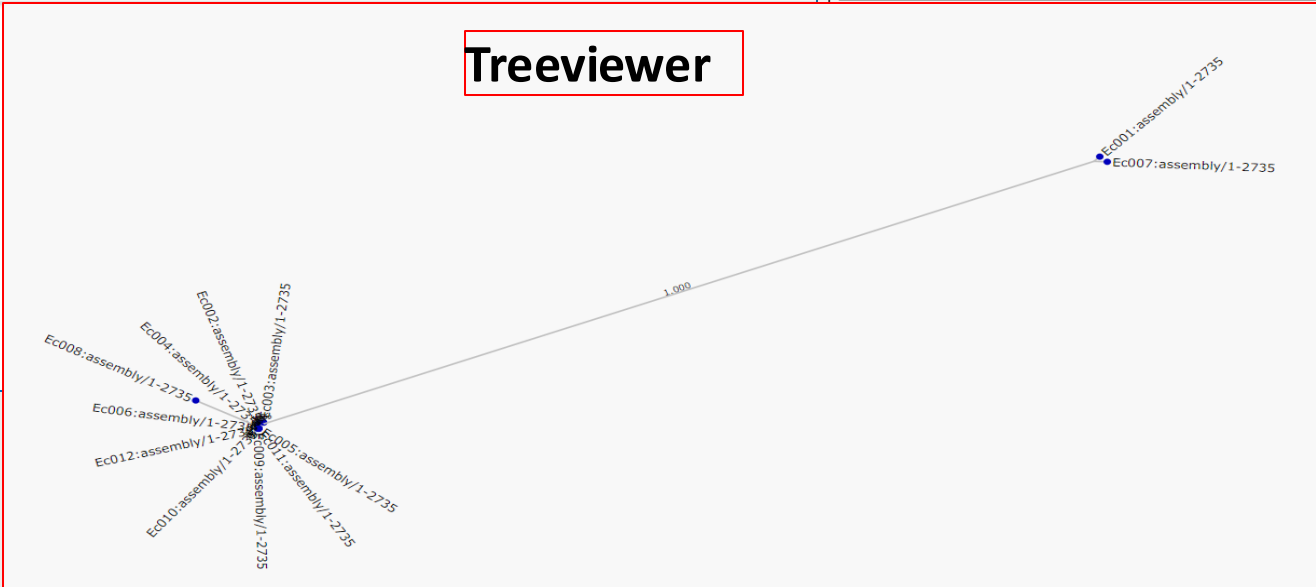
Microreact



FigTree



Treeviewer



We cannot determine clustering based on the visual tree only

SNP matrix – pairwise comparison of SNPs

A1 prune 10

	Ec001_assembl ly/1-2735	Ec002_assembl y/1-2735	Ec003_asse mbly/1-2735	Ec004_asse mbly/1-2735	Ec005_assembl y/1-2735	Ec006_assembl y/1-2735	Ec007_assembl y/1-2735	Ec008_assembl y/1-2735	Ec009_assembl y/1-2735	Ec010_assembl y/1-2735	Ec011_assembl y/1-2735	Ec012_assembl y/1-2735
Ec001_assembl ly/1-2735	0	2176	2122	2180	2176	2176	216	2280	2179	2179	2182	2184
Ec002_assembl ly/1-2735	2176	0	94	80	78	78	2212	644	81	81	84	86
Ec003_assembl ly/1-2735	2122	94	0	98	96	96	2170	662	99	99	102	104
Ec004_assembl ly/1-2735	2180	80	98	0	38	38	2222	604	41	41	44	46
Ec005_assembl ly/1-2735	2176	78	96	38	0	2	2218	598	5	5	8	10
Ec006_assembl ly/1-2735	2176	78	96	38	2	0	2218	598	5	5	8	10
Ec007_assembl ly/1-2735	216	2212	2170	2222	2218	2218	0	2322	2221	2221	2224	2226
Ec008_assembl ly/1-2735	2280	644	662	604	598	598	2322	0	601	601	604	606
Ec009_assembl ly/1-2735	2179	81	99	41	5	5	2221	601	0	0	3	5
Ec010_assembl ly/1-2735	2179	81	99	41	5	5	2221	601	0	0	3	5
Ec011_assembl ly/1-2735	2182	84	102	44	8	8	2224	604	3	3	0	8
Ec012_assembl ly/1-2735	2184	86	104	46	10	10	2226	606	5	5	8	0

min: 0 max: 2322

Analysis 2 output

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_asse mbly/1-676	Ec002_asse mbly/1-676	Ec003_asse mbly/1-676	Ec004_asse mbly/1-676	Ec005_asse mbly/1-676	Ec006_asse mbly/1-676	Ec007_asse mbly/1-676	Ec008_asse mbly/1-676	Ec009_asse mbly/1-676	Ec010_asse mbly/1-676	Ec011_asse mbly/1-676	Ec012_asse mbly/1-676
Ec001_asse mbly/1-676	0	478	461	471	468	469	111	495	473	472	473	479
Ec002_asse mbly/1-676	478	0	45	47	44	45	499	121	49	48	49	55
Ec003_asse mbly/1-676	461	45	0	38	35	36	482	112	40	39	40	46
Ec004_asse mbly/1-676	471	47	38	0	35	36	492	112	40	39	40	46
Ec005_asse mbly/1-676	468	44	35	35	0	3	489	105	7	6	7	13
Ec006_asse mbly/1-676	469	45	36	36	3	0	490	106	8	7	8	14
Ec007_asse mbly/1-676	111	499	482	492	489	490	0	516	494	493	494	500
Ec008_asse mbly/1-676	495	121	112	112	105	106	516	0	106	109	110	114
Ec009_asse mbly/1-676	473	49	40	40	7	8	494	106	0	3	4	8
Ec010_asse mbly/1-676	472	48	39	39	6	7	493	109	3	0	3	9
Ec011_asse mbly/1-676	473	49	40	40	7	8	494	110	4	3	0	10
Ec012_asse mbly/1-676	479	55	46	46	13	14	500	114	8	9	10	0
min: 3 max: 516												

Analysis 3 best reference

More bases included in analysis
with better reference

Percentage of reference genome covered by all isolates **91.8256974372033**

4377814 positions was found in all analyzed genomes.

Size of reference genome: 4767526

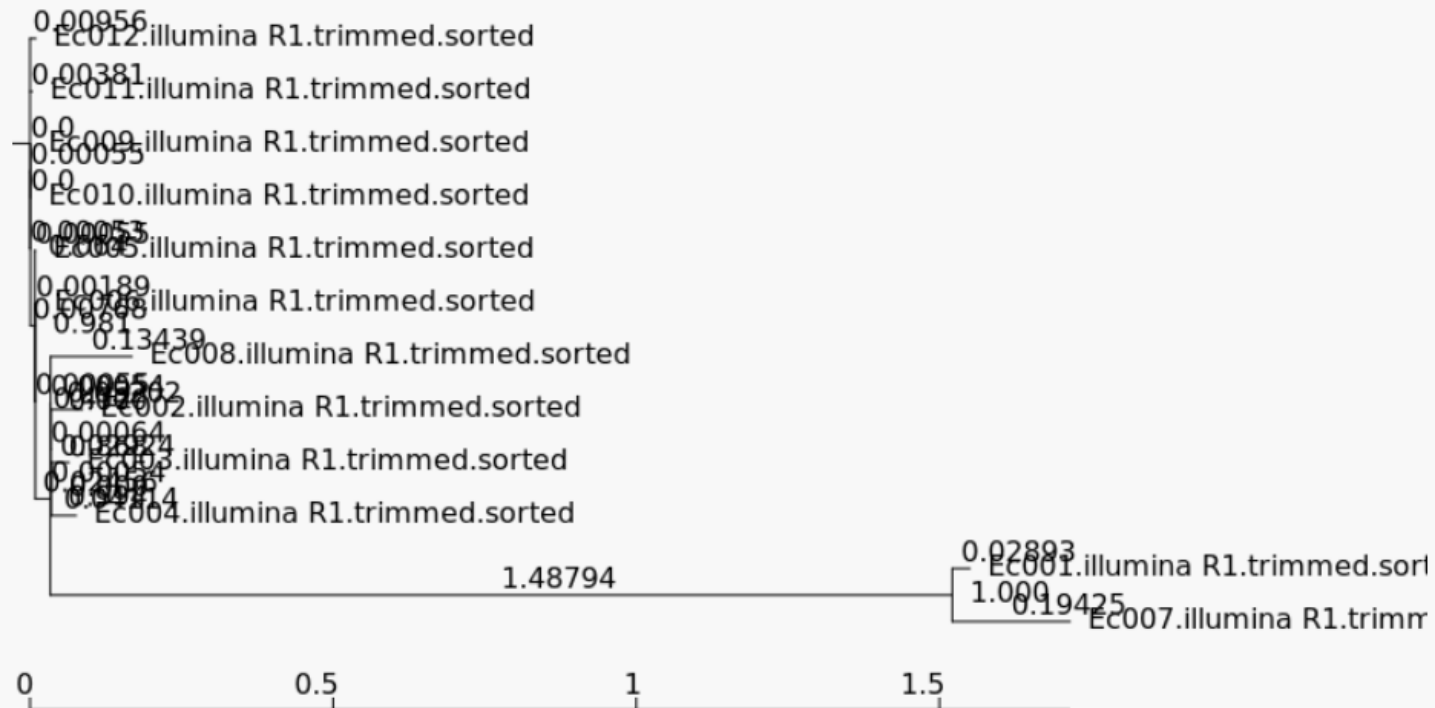
Below is listed the number of positions that are shared and trusted between each isolate and the reference genome.

File	Valid positions	Pct. of reference
Ec010_assembly.ignored_snps	4611790	96.7334000905291
Ec002_assembly.ignored_snps	4706382	98.7174899518115
Ec012_assembly.ignored_snps	4636647	97.2547816204883
Ec009_assembly.ignored_snps	4586479	96.2024958018058
Ec007_assembly.ignored_snps	4493555	94.253392640124
Ec006_assembly.ignored_snps	4677620	98.1142001113366
Ec004_assembly.ignored_snps	4698122	98.544234472974
Ec005_assembly.ignored_snps	4677036	98.1019505714285
Ec008_assembly.ignored_snps	4610963	96.7160535674058
Ec003_assembly.ignored_snps	4700677	98.5978262100721
Ec011_assembly.ignored_snps	4627145	97.0554748941065
Ec001_assembly.ignored_snps	4509616	94.5902759628369

Let's compare with raw data (fastq) analysis (A4)

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](#)).



Analysis 4 output

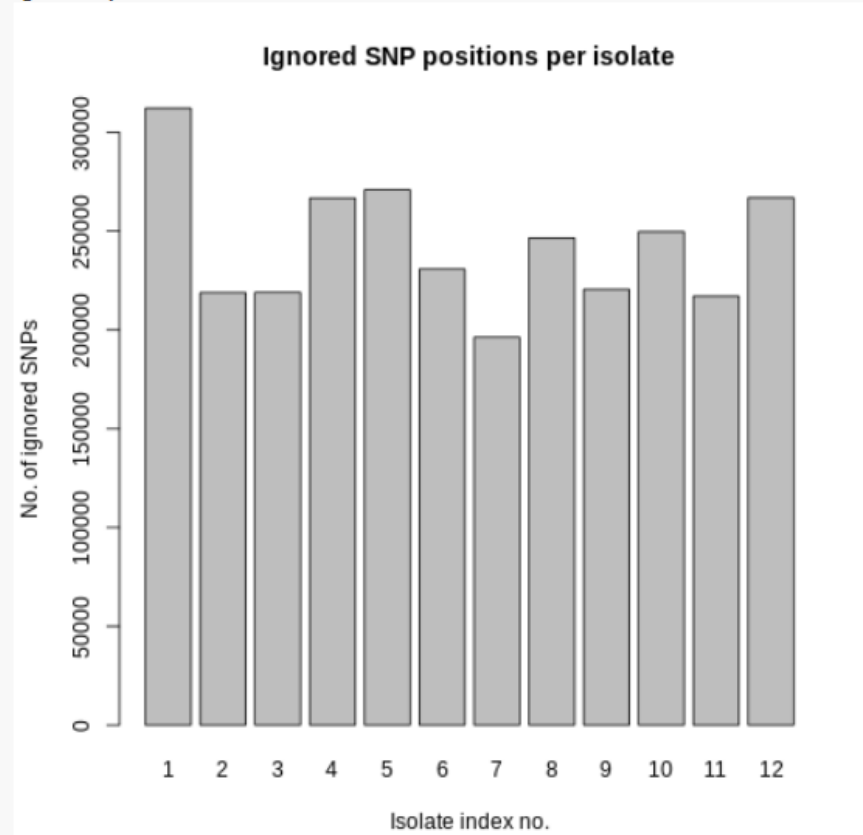
Percentage of reference genome covered by all isolates: 91.0057168298394
 4431507 positions was found in all analyzed genomes.
 Size of reference genome: 4869482

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
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

Fastq = quality parameters (A4)

Ignored positions in each isolate



Index	Isolate	Ignored pos.
1	Ec009.illumina_R1.trimmed	312295
2	Ec005.illumina_R1.trimmed	218893
3	Ec001.illumina_R1.trimmed	219045
4	Ec011.illumina_R1.trimmed	266752
5	Ec010.illumina_R1.trimmed	270983
6	Ec002.illumina_R1.trimmed	230834
7	Ec007.illumina_R1.trimmed	196339
8	Ec004.illumina_R1.trimmed	246623
9	Ec006.illumina_R1.trimmed	220583
10	Ec008.illumina_R1.trimmed	249733
11	Ec003.illumina_R1.trimmed	217114
12	Ec012.illumina_R1.trimmed	266987



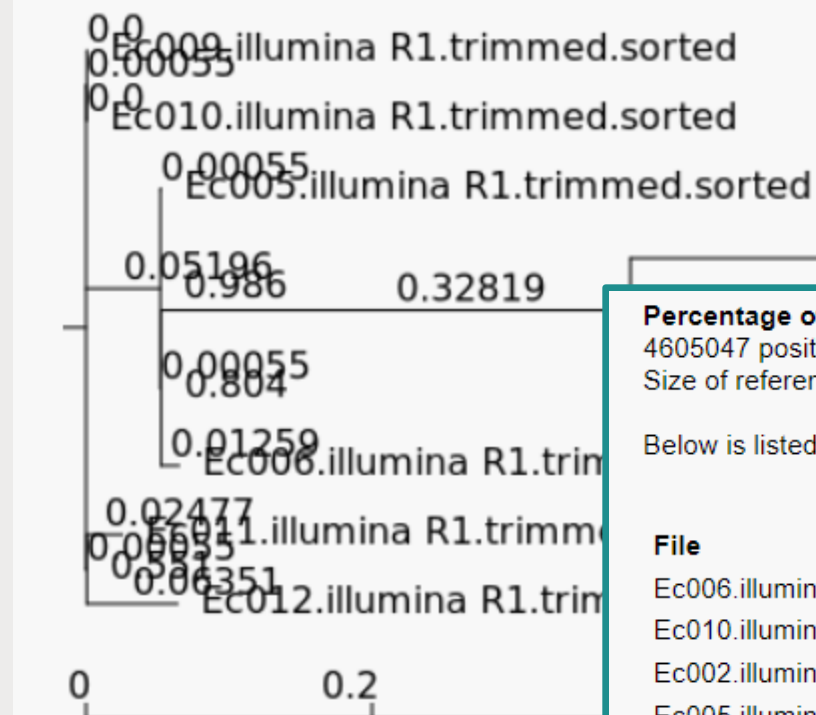
Fasta vs fastq -> different number of SNPs

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												

Analysis 6 output



Newick

SVG

SVG

Index	Isolate	Ignored pos.
1	Ec009.illumina_R1.trimmed	143356
2	Ec002.illumina_R1.trimmed	15087
3	Ec005.illumina_R1.trimmed	25080
4	Ec004.illumina_R1.trimmed	23813
5	Ec006.illumina_R1.trimmed	29039
6	Ec011.illumina_R1.trimmed	97355
7	Ec003.illumina_R1.trimmed	16678
8	Ec010.illumina_R1.trimmed	105374
9	Ec012.illumina_R1.trimmed	87795

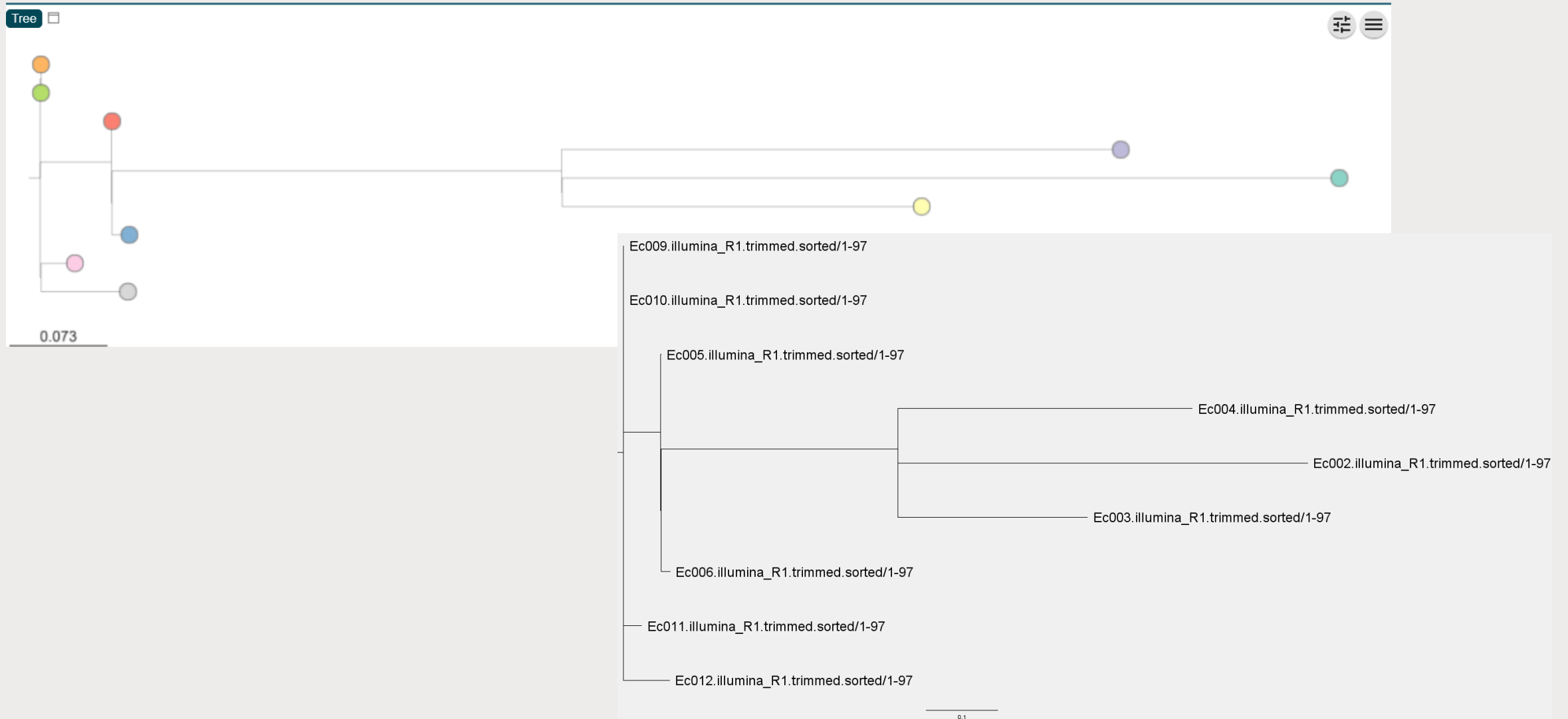
Percentage of reference genome covered by all isolates: 96.5919640501174

4605047 positions was found in all analyzed genomes.
Size of reference genome: 4767526

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

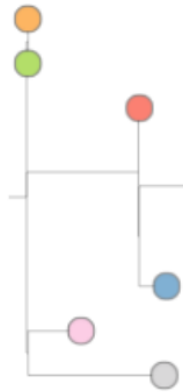
File	Valid positions	Pct. of reference
Ec006.illumina_R1.trimmed.ignored_snps	4738518	99.3915502505912
Ec010.illumina_R1.trimmed.ignored_snps	4662180	97.7903424123959
Ec002.illumina_R1.trimmed.ignored_snps	4752591	99.6867347970415
Ec005.illumina_R1.trimmed.ignored_snps	4742469	99.4744234221271
Ec012.illumina_R1.trimmed.ignored_snps	4679761	98.1591080992532
Ec009.illumina_R1.trimmed.ignored_snps	4624202	96.9937447640558
Ec003.illumina_R1.trimmed.ignored_snps	4750902	99.65130761741
Ec004.illumina_R1.trimmed.ignored_snps	4743766	99.5016283078477
Ec011.illumina_R1.trimmed.ignored_snps	4670190	97.9583540813411

Analysis 6 output



Analysis output 6

Tree ☐



	Ec005.illumina_R1.trimmed.sorted/1-545	Ec006.illumina_R1.trimmed.sorted/1-545	Ec009.illumina_R1.trimmed.sorted/1-545	Ec010.illumina_R1.trimmed.sorted/1-545	Ec011.illumina_R1.trimmed.sorted/1-545	Ec012.illumina_R1.trimmed.sorted/1-545
Ec005.illumina_R1	0	1	4	4	6	9
Ec006.illumina_R1	1	0	5	5	7	10
Ec009.illumina_R1	4	5	0	0	2	5
Ec010.illumina_R1	4	5	0	0	2	5
Ec011.illumina_R1	6	7	2	2	0	7
Ec012.illumina_R1	9	10	5	5	7	0

Region	Travel	MLST	Sequence
Copenhagen	Pakistan	ST410	Ec001
Copenhagen	Thailand	ST410	Ec002
Jutland - M	India	ST410	Ec003
Copenhagen	Lebanon	ST410	Ec004
Zealand	No	ST410	Ec005
Zealand	No	ST410	Ec006
Copenhagen	Pakistan	ST410	Ec007
Jutland - N	Thailand	ST410	Ec008
Zealand	No	ST410	Ec009
Zealand	No	ST410	Ec010
Zealand	No	ST410	Ec011
Zealand	No	ST410	Ec012

Ec009.illumina_R1.trimmed.sorted/1-97

Ec010.illumina_R1.trimmed.sorted/1-97

Ec005.illumina_R1.trimmed.sorted/1-97

Ec004.illumina_R1.trimmed.sorted/1-97

Ec002.illumina_R1.trimmed.sorted/1-97

Ec003.illumina_R1.trimmed.sorted/1-97

Ec006.illumina_R1.trimmed.sorted/1-97

Ec011.illumina_R1.trimmed.sorted/1-97

Ec012.illumina_R1.trimmed.sorted/1-97

Apparent outbreak of 6 non-travel related isolates

Sum-up of analysis of Illumina data

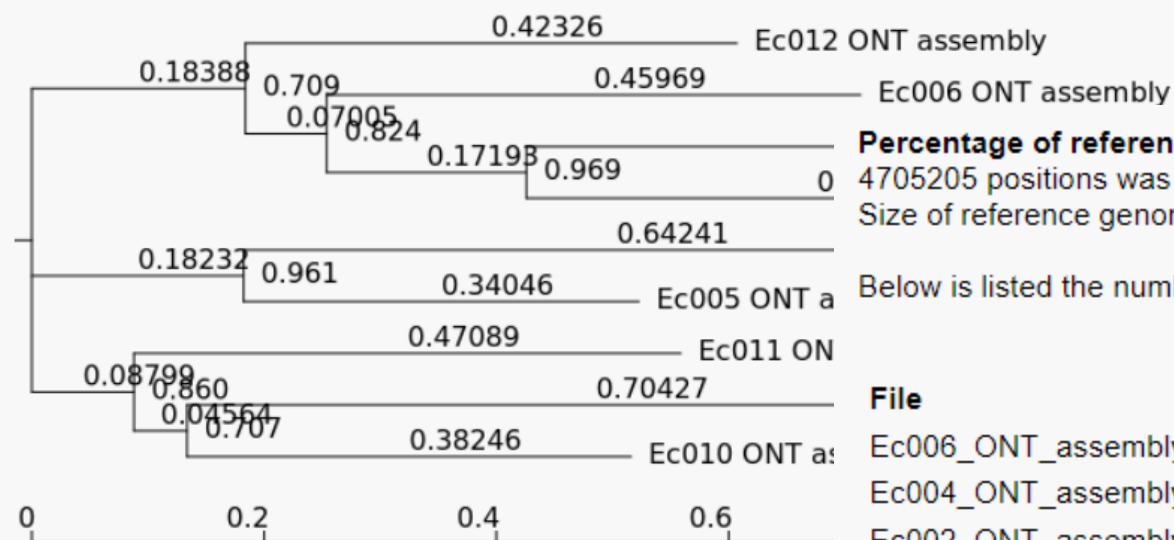
- Improved analysis result by:
 - Pruning 100 instead of 10
 - Better reference (index isolate)
 - Using fastq instead of fasta
- Determined 6 isolates (5,6,9,10,11,12) with ≤ 10 SNPs as an outbreak

Analysis 7 on ONT assembly

No apparent clustering
– more scattered

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 phylogeny as: [Newick](#) [PDF](#) [SVG](#)

Percentage of reference genome covered by all isolates: 98.6928020948391
4705205 positions was found in all analyzed genomes.
Size of reference genome: 4767526

Below is listed the number of positions that are shared and trusted between each

File	Valid positions	Pct. of reference
Ec006_ONT_assembly.ignored_snps	4740160	99.425991593963
Ec004_ONT_assembly.ignored_snps	4762644	99.8975988804256
Ec002_ONT_assembly.ignored_snps	4750672	99.6464833123092
Ec009_ONT_assembly.ignored_snps	4708623	98.7644954636849
Ec011_ONT_assembly.ignored_snps	4811173	100.915506281455
Ec010_ONT_assembly.ignored_snps	4756051	99.7593091259492
Ec003_ONT_assembly.ignored_snps	4748341	99.5975900288745
Ec012_ONT_assembly.ignored_snps	4708629	98.7646213151223
Ec005_ONT_assembly.ignored_snps	4740160	99.425991593963

Analysis output 7 ONT assembly

Running CSIPhylogeny on ONT assemblies (fasta) is not recommended

SNPs: min 276 max 451

	Ec002_ONT_asse mbly/1-1134	Ec003_ONT_asse mbly/1-1134	Ec004_ONT_asse mbly/1-1134	Ec005_ONT_asse mbly/1-1134	Ec006_ONT_asse mbly/1-1134	Ec009_ONT_asse mbly/1-1134	Ec010_ONT_asse mbly/1-1134	Ec011_ONT_asse mbly/1-1134	Ec012_ONT_asse mbly/1-1134
Ec002_ONT_assembly/1-1134	0	352	351	362	351	443	376	376	351
Ec003_ONT_assembly/1-1134	352	0	378	277	357	371	336	354	356
Ec004_ONT_assembly/1-1134	351	378	0	374	345	451	394	377	340
Ec005_ONT_assembly/1-1134	362	277	374	0	326	303	279	310	309
Ec006_ONT_assembly/1-1134	351	357	345	326	0	386	350	345	319
Ec009_ONT_assembly/1-1134	443	371	451	303	386	0	304	361	377
Ec010_ONT_assembly/1-1134	376	336	394	279	350	304	0	276	323
Ec011_ONT_assembly/1-1134	376	354	377	310	345	361	276	0	320
Ec012_ONT_assembly/1-1134	351	356	340	309	319	377	323	320	0
min: 276 max: 451									

Analysis 8b (9 isolates) MinTyper

Percentage of reference covered by all isolates: **95.41 (4548752 / 4767526)**
Below is the single isolate stats on covered and trusted positions with respect to the reference.

Isolate	Valid positions	Pct. of reference
Ec002.illumina_R1.trimmed.fastq.gz	4688571	98.34
Ec003.illumina_R1.trimmed.fastq.gz	4696326	98.51
Ec004.illumina_R1.trimmed.fastq.gz	4696130	98.50
Ec005.illumina_R1.trimmed.fastq.gz	4651076	97.56
Ec006.illumina_R1.trimmed.fastq.gz	4671229	97.98
Ec009.illumina_R1.trimmed.fastq.gz	4648702	97.51
Ec010.illumina_R1.trimmed.fastq.gz	4667494	97.90
Ec011.illumina_R1.trimmed.fastq.gz	4666373	97.88
Ec012.illumina_R1.trimmed.fastq.gz	4669504	97.94



Ec012.illumina_R1.trimmed.fastq.gz_alignment.fsa

Ec002.illumina_R1.trimmed.fastq.gz_alignment.fsa

Ec009.illumina_R1.trimmed.fastq.gz_alignment.fsa

Ec004.illumina_R1.trimmed.fastq.gz_alignment.fsa

Ec005.illumina_R1.trimmed.fastq.gz_alignment.fsa

Ec006.illumina_R1.trimmed.fastq.gz_alignment.fsa

Ec009.illumina_R1.trimmed.fastq.gz_alignment.fsa

Ec010.illumina_R1.trimmed.fastq.gz_alignment.fsa

Ec011.illumina_R1.trimmed.fastq.gz_alignment.fsa

Ec012.illumina_R1.trimmed.fastq.gz_alignment.fsa

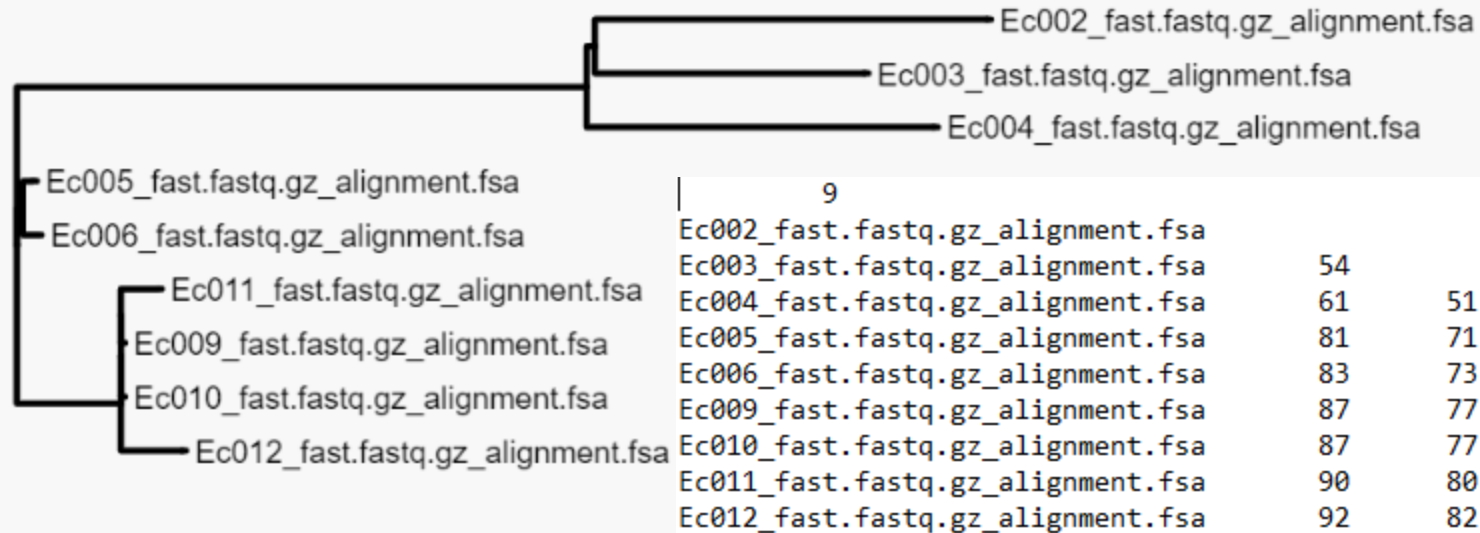
Ec012.illumina_R1.trimmed.fastq.gz_alignment.fsa

Running MinTyper with
Illumina reads (fastq)
works fine

Ec002.illumina_R1.trimmed.fastq.gz_alignment.fsa									
Ec003.illumina_R1.trimmed.fastq.gz_alignment.fsa	57								
Ec004.illumina_R1.trimmed.fastq.gz_alignment.fsa	68	43							
Ec005.illumina_R1.trimmed.fastq.gz_alignment.fsa	63	56	57						
Ec006.illumina_R1.trimmed.fastq.gz_alignment.fsa	64	57	58	1					
Ec009.illumina_R1.trimmed.fastq.gz_alignment.fsa	66	59	70	13	14				
Ec010.illumina_R1.trimmed.fastq.gz_alignment.fsa	66	59	70	13	14	0			
Ec011.illumina_R1.trimmed.fastq.gz_alignment.fsa	68	61	72	15	16	2	2		
Ec012.illumina_R1.trimmed.fastq.gz_alignment.fsa	71	64	75	18	19	5	5		

Analysis output 10 ONT reads (MinTyper)

Running MinTyper with
ONT reads shows the
same clustering.
2-15 SNPs detected



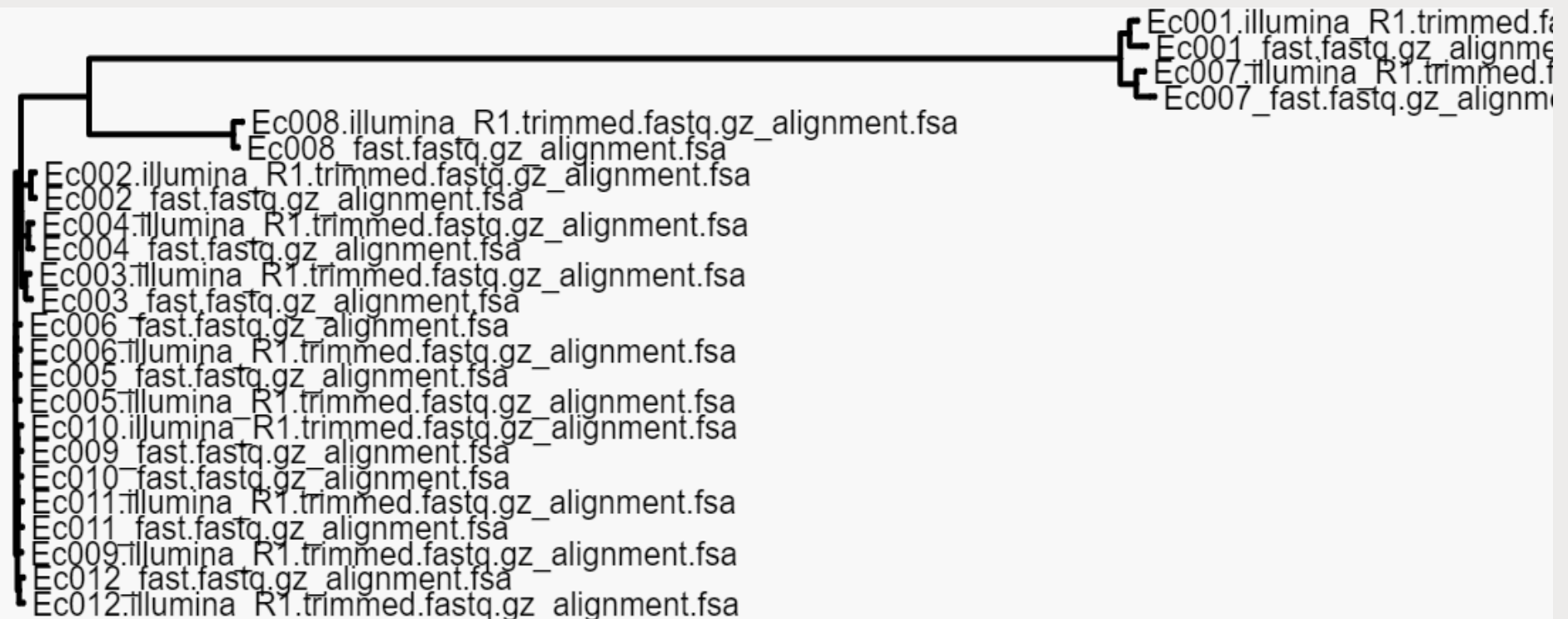
Percentage of reference covered by all isolates: 97.55 (4650614 / 4767526)

Below is the single isolate stats on covered and trusted positions with respect to the reference.

Analysis 11 output MinTyper

Illumina and ONT raw data (All 2 x 12 isolates)

- Illumina outbreak isolates 0-11 SNPs
 - ONT outbreak isolates 0-11 SNPs
 - Illumina vs ONT isolates 2-11 SNPs
- Comparable results
between platforms



MinTyper - Basecalling

Percentage of reference covered by all isolates: 89.50 (4267149 / 4767526)

Below is the single isolate stats on covered and trusted positions with respect to the reference.

ONT raw data was
processed with 'fast'
basecalling

Isolate	Valid positions	Pct. of reference
Ec001_fast.fastq.gz	4563512	95.72
Ec001.illumina_R1.trimmed.fastq.gz	4501213	94.41
Ec002_fast.fastq.gz	4719352	98.99
Ec002.illumina_R1.trimmed.fastq.gz	4688571	98.34
Ec003_fast.fastq.gz	4717284	98.95
Ec003.illumina_R1.trimmed.fastq.gz	4696326	98.51
Ec004_fast.fastq.gz	4750223	99.64
Ec004.illumina_R1.trimmed.fastq.gz	4696130	98.50
Ec005_fast.fastq.gz	4713769	98.87
Ec005.illumina_R1.trimmed.fastq.gz	4651076	97.56
Ec006_fast.fastq.gz	4737522	99.37
Ec006.illumina_R1.trimmed.fastq.gz	4671229	97.98
Ec007_fast.fastq.gz	4537645	95.18
Ec007.illumina_R1.trimmed.fastq.gz	4523451	94.88
Ec008_fast.fastq.gz	4677416	98.11
Ec008.illumina_R1.trimmed.fastq.gz	4675681	98.07
Ec009_fast.fastq.gz	4679011	98.14
Ec009.illumina_R1.trimmed.fastq.gz	4648702	97.51
Ec010_fast.fastq.gz	4679842	98.16
Ec010.illumina_R1.trimmed.fastq.gz	4667494	97.90
Ec011_fast.fastq.gz	4683906	98.25
Ec011.illumina_R1.trimmed.fastq.gz	4666373	97.88
Ec012_fast.fastq.gz	4682650	98.22
Ec012.illumina_R1.trimmed.fastq.gz	4669504	97.94

MinTyper - Basecalling

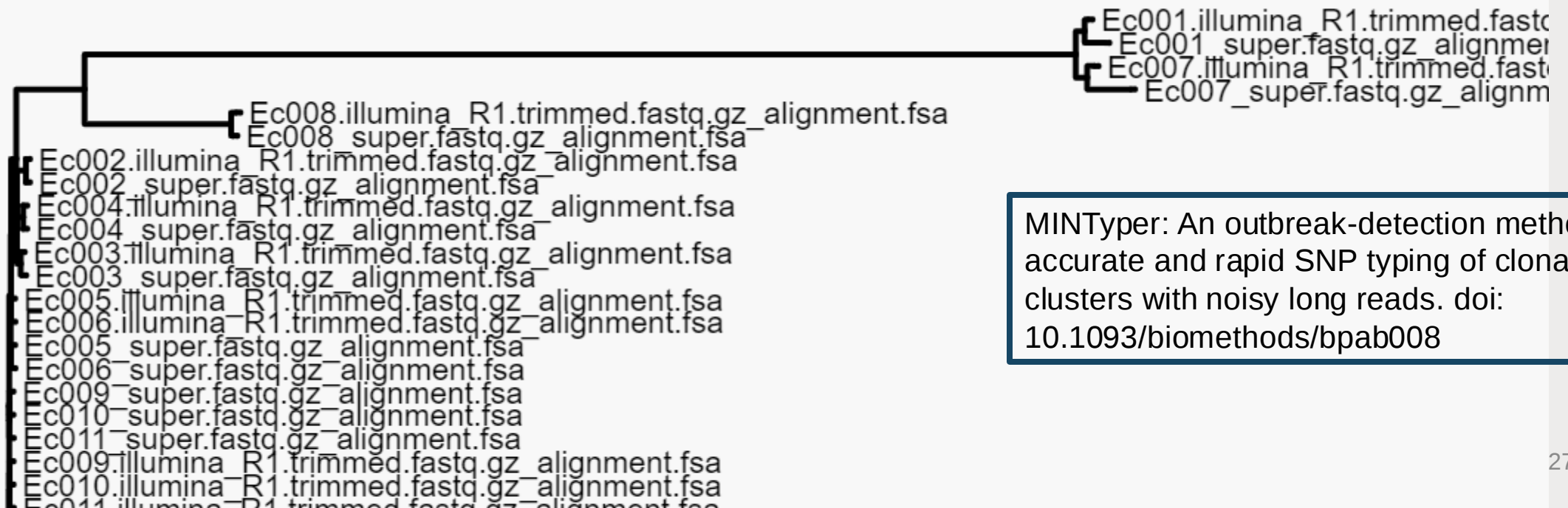
Illumina and ONT raw data (All 2 x 12 isolates – Super basecalling)

- Illumina outbreak isolates 2-12 SNPs
- ONT outbreak isolates 1-12 SNPs
- Illumina vs ONT isolates 3-15 SNPs

ONT raw data was
processed with
'super' basecalling

Percentage of reference covered by all isolates: 90.26 (4302960 / 4767526)

Below is the single isolate stats on covered and trusted positions with respect to the reference.



MINTyper: An outbreak-detection method for accurate and rapid SNP typing of clonal clusters with noisy long reads. doi: 10.1093/biomethods/bpab008

Sum-up of analysis of ONT (+Illumina) data

- Result from MinTyper with ONT data
 - Outbreak cluster
- ONT assembly data is not suitable for SNP analysis!
- Result from Mintyper with both ONT and Illumina data
 - Comparable results between data from different platforms
 - Determined 6 isolates (5,6,9,10,11,12) with ≤ 15 SNPs as an outbreak
- Slight improvement in analysis with improved basecalling
 - Additional note: the MinION basecaller is trained on *E. coli* data, so it could be less accurate with other bacterial species

Thank you



This programme is being funded by the UK Department of Health and Social Care.
The views expressed do not necessarily reflect the UK Government's official policies.