Guidelines to how to get started 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.

Exercise on clonal typing (SNP analysis)

In this exercise you can try to see, how different tools (CSI Phylogeny vs MinTyper) affects the outcome of the SNP analysis. Also, you can try to change some of the settings or exclude 'outlier' isolates to get a better resolution of your analysis. Also, you can try to see the effect of using either a “suitable but not perfect” reference (“Kmerfinder reference”) or a perfect reference (we have fully assembled the index isolate previously) and finally you can try different types of sequencing data (Illumina draft assemblies, Illumina raw data, MinION draft assemblies or MinION raw data). Here please know that CSIPhylogeny does not accept MinION raw data and MinTyper does not accept draft genomes (FASTA files) at the moment. Below are suggestions for analysis schemes to try out and also a link to the results so you don’t have to upload to get the result from the server (valid for this week only). The relevant changes compared to the previous analysis is highlighted with yellow marker. You don’t necessarily have to follow all examples, as these have been included to satisfy the eager students curiosity ☺. It is recommended that you at least examine:

Analysis 2 vs 3: CSI phylogeny analysis (Prune = 100) using draft genomes based on Illumina sequencing and either a KmerFinder reference or the optimal reference.

Analysis 2 vs 4: CSI phylogeny analysis (Prune = 100) using either draft genomes or raw reads based on Illumina sequencing and the KmerFinder reference.

Analysis 4 vs 5: CSI phylogeny analysis (Prune = 100) using raw reads based on Illumina sequencing and the KmerFinder reference and including either all 12 genomes or only the 9 most similar genomes.

Analysis 3 vs 7: CSI phylogeny analysis (Prune = 100) using draft assemblies from either Illumina og ONT data.

Analysis 6 & 8: CSI phylogeny vs MinTyper analysis (Prune = 100) using raw Illumina data the Best reference and only the 9 most similar genomes.

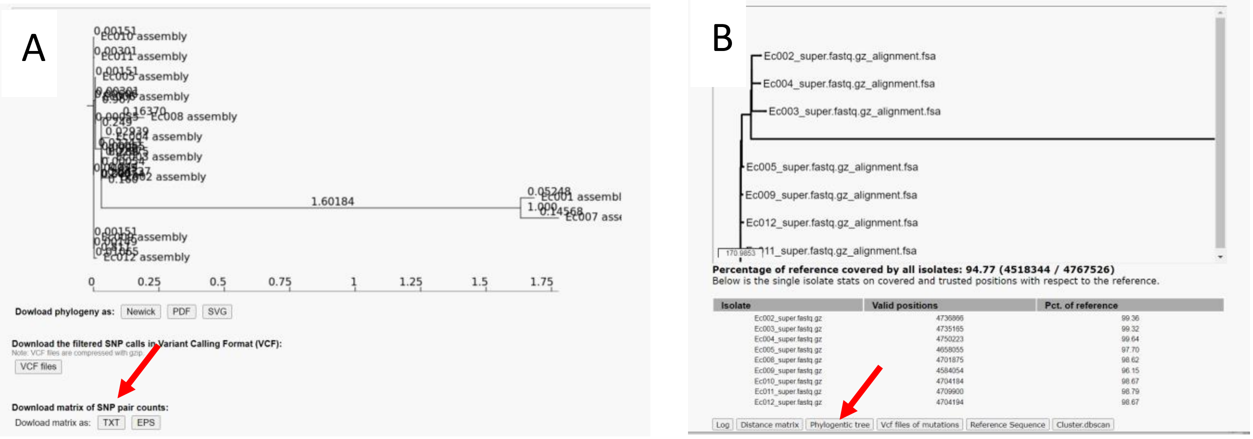
Analysis 8 & 10: MinTyper analysis (Prune = 100) using either raw Illumina or raw MinION (fast basecalling) data with the Best reference and only the 9 most similar genomes.

Analysis 12 & 13: MinTyper analysis (Prune = 100) using both raw Illumina data and raw MinION data basecalled either using the fast basecalling algorithm or the Super accuracy algorithm in Guppy and with the Best reference on all 2 x 12 genomes.

Hint: The fastest way to analyse these 12 genomes is to first perform **Analysis 3** (If a perfect reference is available, otherwise Analysis 2) on the Illumina draft genomes to see if some of the isolates can be omitted and then **Analysis 6** to perform the HQ SNP analysis on a subset of the isolates, which are closest to each other.

Notice: As MinION draft assemblies are of poorer quality, analyzing these is not recommended. Therefore, all 12 genomes in raw format should be included and then a final analysis with the selected subset of genomes can be made.

To do the comparisons, it is recommended that you in the output files of both CSI Phylogeny (Figure 1A) and MinTyper (Figure 1B) focus on the dendrograms to get the overview and then the Distance matrices generated as part of the analysis (find these where red arrows are indicated on the figures below). The format is a raw text format, which will often be difficult to read directly. If so, it is recommended to copy the text into a Spread sheet such as MS Excel.



Dowload phylogeny as:

Download the filtered SNP calls in Variant Calling Format (VCF):
Note: VCF files are compressed with gzip.

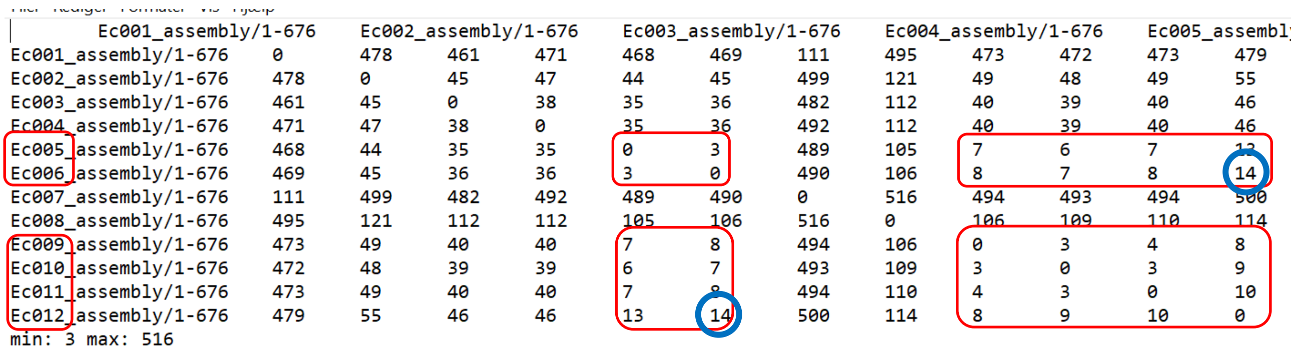
Download matrix of SNP pair counts:
Dowload matrix as:

Percentage of reference covered by all isolates: 94.77 (4518344 / 4767526)
Below is the single isolate stats on covered and trusted positions with respect to the reference.

Isolate	Valid positions	Pct. of reference
Ec002_super.fastq.gz	4736866	99.36
Ec003_super.fastq.gz	4735165	99.32
Ec004_super.fastq.gz	4750223	99.64
Ec005_super.fastq.gz	4658055	97.70
Ec008_super.fastq.gz	4701875	98.62
Ec009_super.fastq.gz	4584054	96.15
Ec010_super.fastq.gz	4704184	98.67
Ec011_super.fastq.gz	4709900	98.79
Ec012_super.fastq.gz	4704194	98.67

Figure 1. Output from CSI Phylogeny (A) and MinTyper (B).

In the Distance matrices, try to see if you are able to group some of the isolates together based on a given SNP distance (e.g. 10 or 15 SNPs) to identify possible outbreak strains. And in the pairwise comparisons suggested above, try to see how the different settings affect the SNP distances as well as the overall topology of the phylogenetic trees. An example of a distance matrix from CSI Phylogeny is given in Figure 2.



```
|         Ec001_assembly/1-676         Ec002_assembly/1-676         Ec003_assembly/1-676         Ec004_assembly/1-676         Ec005_assembl
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
```

Figure 2. Distance matrix from CSI phylogeny of the 12 isolates. Red boxes indicate isolates, which group together with up to 15 SNPs distance and which could constitute outbreak isolates (notice that the matrix from CSI Phylogeny is symmetrical to the diagonal). However, if a cut-off of 10 SNPs is applied, some of the isolates will no longer group together with the rest, as e.g. Ec004 and Ec12 will be 14 SNPs apart (blue circles).

Please notice the “Server run time” information, as this gives you an idea about how fast you can get the information you need, but this may come with the cost of reduced precision in the analysis. However, this “Server run time” does depend on how busy the server is when you submit the job.

CSI Phylogeny

Analysis 1

Tool: CSI Phylogeny

Reference: KmerFinder reference

Prune: 10

Data: Illumina draft genomes (all 12 isolates)

Results: Center for Genomic Epidemology - 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 Epidemology - Results (dtu.dk)

Server run time (approximately): 10 minutes

Analysis 3

Tool: CSI Phylogeny

Reference: Best reference

Prune: 100

Data: Illumina draft genomes (all 12 isolates)

Results: Center for Genomic Epidemology - Results (dtu.dk)

Server run time (approximately): 10 minutes

Analysis 4

Tool: CSI Phylogeny

Reference: KmerFinder reference

Prune: 100

Data: Illumina raw data (all 12 isolates)

Results: Center for Genomic Epidemology - Results (dtu.dk)

Server run time (approximately): 60-90 minutes

Analysis 5

Tool: CSI Phylogeny

Reference: KmerFinder reference

Prune: 100

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

Results: Center for Genomic Epidemology - Results (dtu.dk)

Server run time (approximately): 60-90 minutes

Analysis 6

Tool: CSI Phylogeny

Reference: Best reference

Prune: 100

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

Results: Center for Genomic Epidemology - Results (dtu.dk)

Server run time (approximately): 60-90 minutes

Analysis 7

Tool: CSI Phylogeny

Reference: Best reference

Prune: 100

Data: ONT assembly data (9 closest related isolates only)

Results: Center for Genomic Epidemology - 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

Analysis 9

Tool: MinTyper

Reference: Best reference

Prune: 100

Data: ONT raw data fast basecalling (All 12 isolates)

Results: MINTyper-1.0 - Results (dtu.dk)

Server run time (approximately): 20-30 minutes

Analysis 10

Tool: MinTyper

Reference: Best reference

Prune: 100

Data: ONT raw data fast basecalling (9 closest related isolates only)

Results: MINTyper-1.0 - Results (dtu.dk)

Server run time (approximately): 20-30 minutes

Analysis 11

Tool: MinTyper

Reference: Best reference

Prune: 100

Data: ONT raw data Super basecalling (9 closest related isolates only)

Results: MINTyper-1.0 - Results (dtu.dk)

Server run time (approximately): 20-30 minutes

Analysis 12

Tool: MinTyper

Reference: Best reference

Prune: 100

Data: Illumina and ONT raw data fast basecalling (All 2 x 12 isolates)

Results: MINTyper-1.0 - Results (dtu.dk)

Server run time (approximately): 60-90 minutes

Analysis 13

Tool: MinTyper

Reference: Best reference

Prune: 100

Data: Illumina and ONT raw data super basecalling (All 2 x 12 isolates)

Results: MINTyper-1.0 - Results (dtu.dk)

Server run time (approximately): 60-90 minutes