STATENS SERUM INSTITUT

MINTYPER TO ANALYSE ONT (MINION) DATA

Senior scientist
HENRIK HASMAN
Statens Serum Institut (SSI)
Denmark

MINION – THE NEW(ISH) KID ON THE BLOCK

STATENS SERUM INSTITUT

6-15 days

6-48 hours

Relatively..

- low price per isolat
- well-proven tec gy
- high precisi ow error rate)
- Slow (d ding on the setup)

..but n ds in real-time

Tools for outbreak detection validated

Relatively..

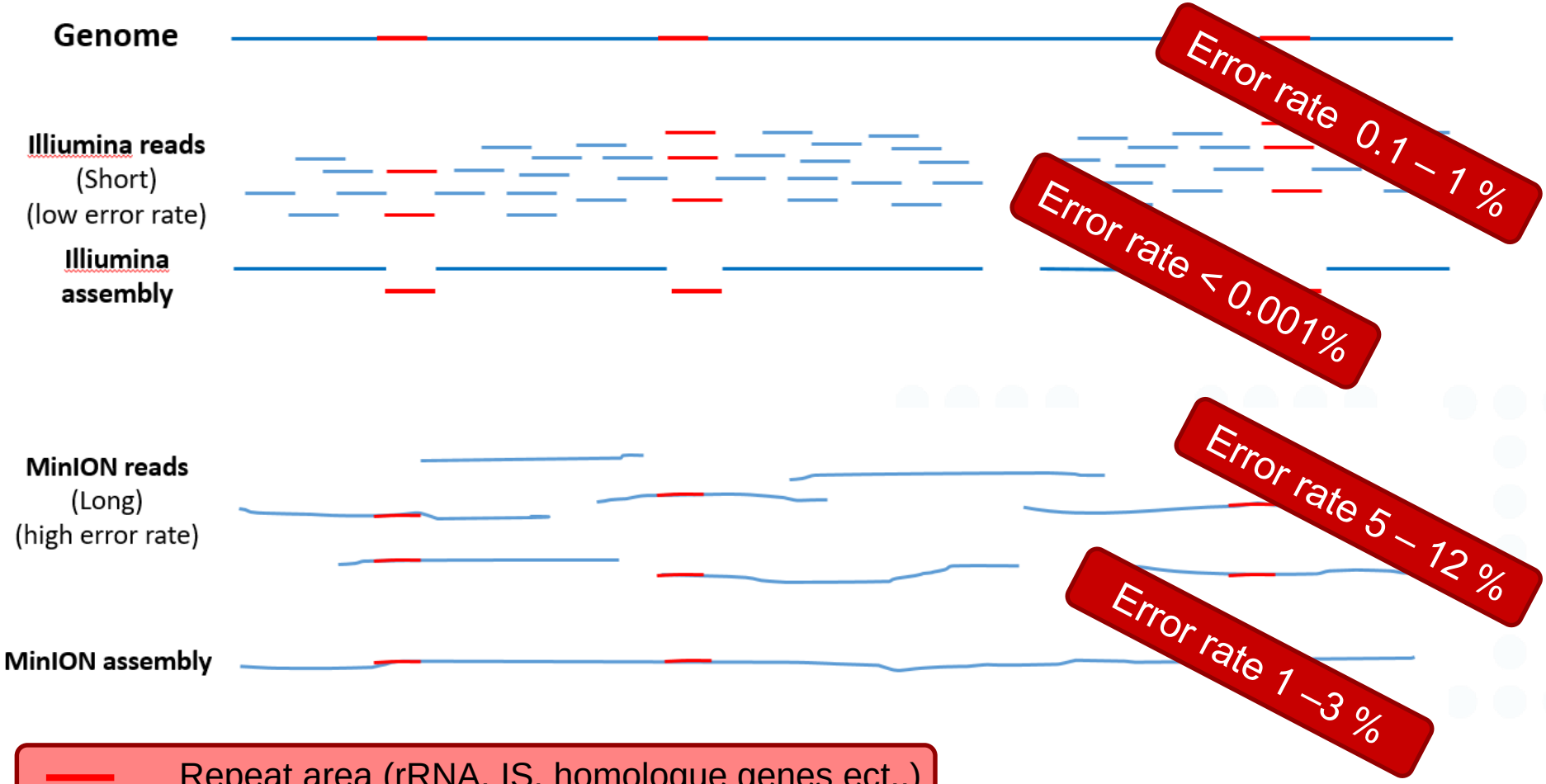
- Low-to medium per isolate
- experiment nology
- low pre (high error rate)
- fast

..a ads available in real-time

Tools for outbreak detection emerging

ILLUMINA VS MINION (R9.4.1) DATA

STATENS SERUM INSTITUT



Repeat area (rRNA, IS, homologue genes ect..)

ILLUMINA VS MINION DATA

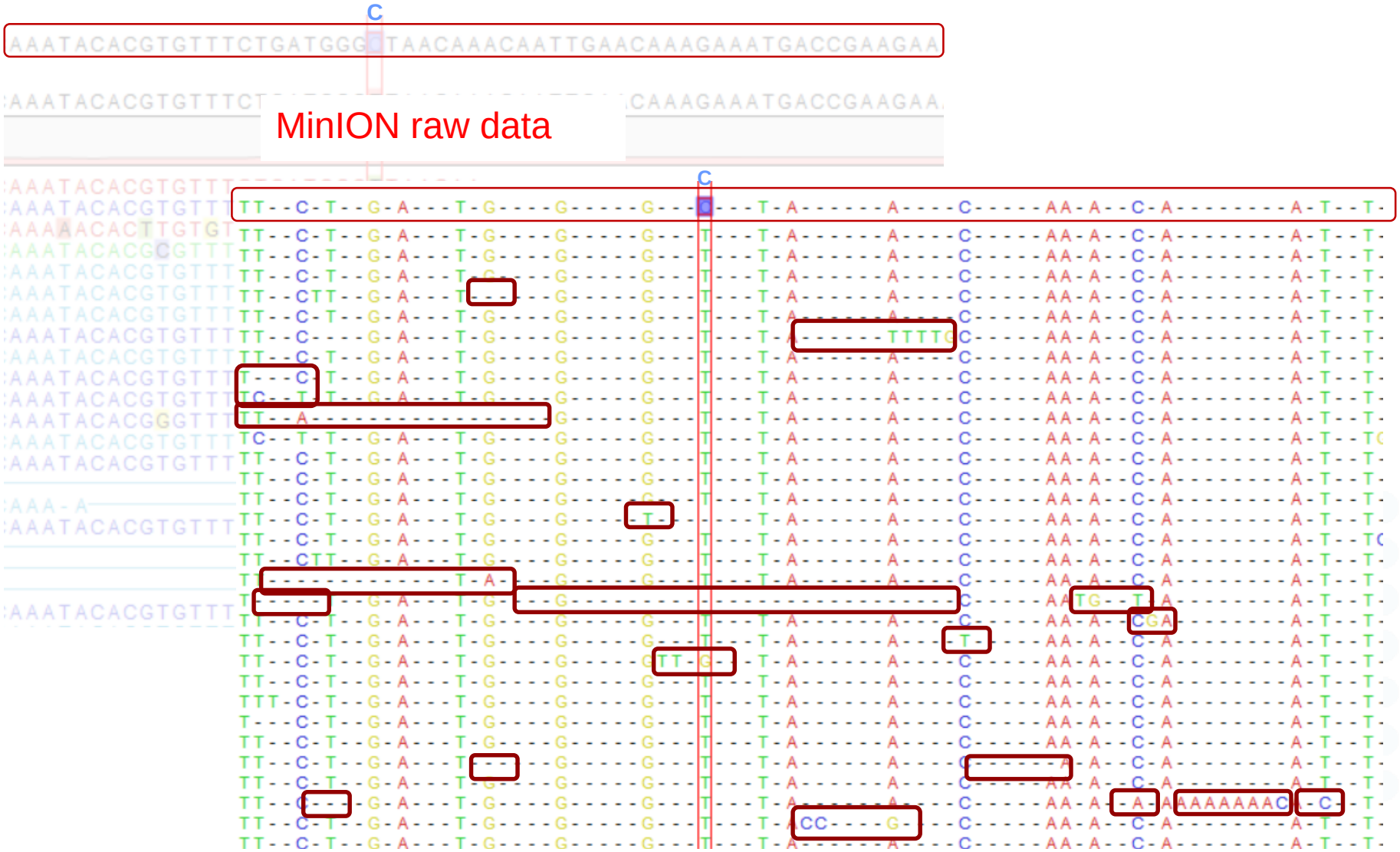
Illumina raw data

```
AAATACACGTGTTTCTGATGGGCTAACAAACAATTGAACAAAGAAATGACCGAAGAA

AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA

AAATACACGTGTTTCTGATGGGTTAACAA
AAATACACGTGTTTCTGATGKGTTAACAAACAATTGAACAAAGAAATGACCRAAGAA
AAAAACACTTGTGTCTGATGGGTTAACAAACAAGTGAACAAAGAAATGACCGAAGAA
AAATACACGCGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGCA
AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA
AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA
AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA
AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA
AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA
AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA
AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA
AAATACACGGGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA
AAATACACGTGTTTCTGATGGGTTAACAAAC
AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA

AAA-A
AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA


                TGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA
AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA
```

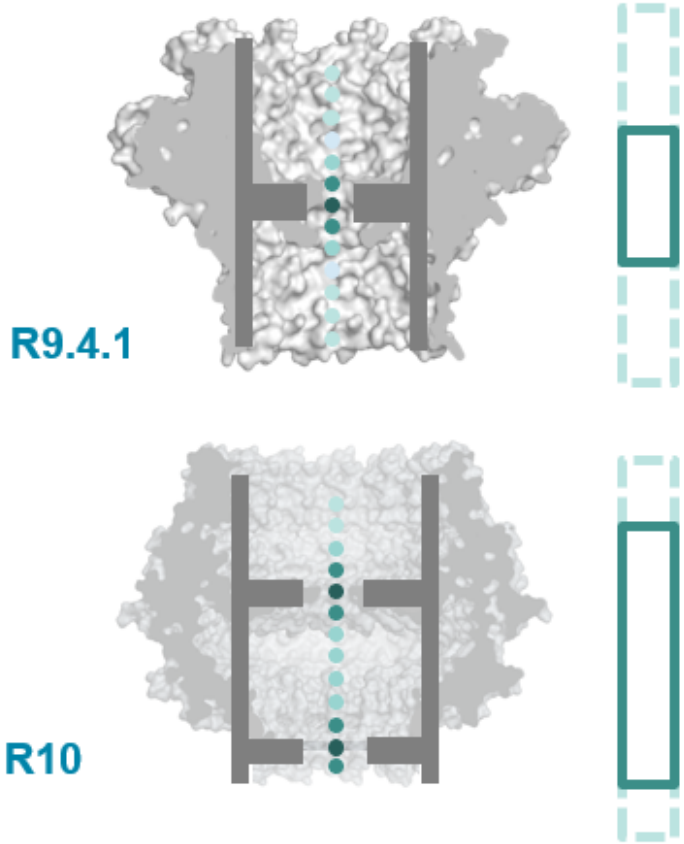
ILLUMINA VS MINION DATA

STATENS SERUM INSTITUT

Illumina raw data



R9.4.1 VS R10.4.1 PORE



CHOICE OF FLOWCELL / PORE

STATENS SERUM INSTITUT

Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

Oxford Nanopore R10.4 long-read sequencing enables near-perfect bacterial genomes from pure cultures and metagenomes without short-read or reference polishing

Mantas Sereika[a*], Rasmus Hansen Kirkegaard[a,b*], Søren Michael Karst[a], Thomas Yssing Michaelsen[a], Emil Aarre Sørensen[a], Rasmus Dam Wollenberg[c] and Mads Albertsen[a**]

[a]Center for microbial communities, Aalborg University, Denmark

[b]Joint Microbiome Facility, University of Vienna, Austria

[c]DNASense ApS, Denmark

*These authors contributed equally to the paper

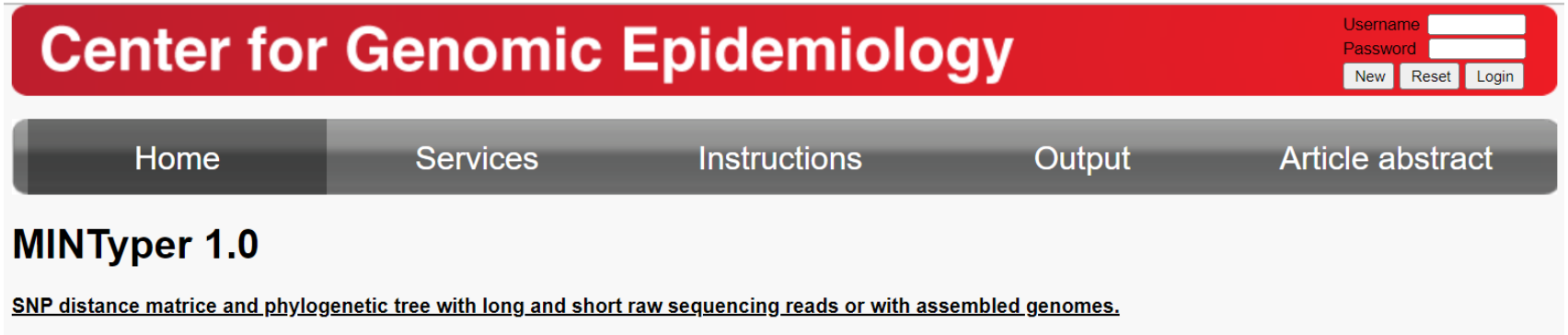
**Corresponding author ma@bio.aau.dk

20	0.01000
19	0.01259
18	0.01585
17	0.01995
16	0.02512
15	0.03162
14	0.03981
13	0.05012
12	0.06310
11	0.07943
10	0.10000
9	0.12589
8	0.15849
7	0.19953
6	0.25119
5	0.31623
4	0.39811
3	0.50119
2	0.63096
1	0.79433

https://www.biorxiv.org/content/10.1101/2021.10.27.466057v2

THE MINTYPER TOOL AT CGE

STATENS SERUM INSTITUT



- Will only accept raw data (Illumina and ONT)
- Will fail if not all input data (strains) cover at least 50% of the reference
- Allow for the use to give her own reference genome (fasta format)
- Allow to filter out Dcm methylation signals, which may give issues with the fast basecaller (at least in old versions of Guppy).
- Exists as a command-line tool (genomicepidemiology / mintyper — Bitbucket).

MINTYPER V1.0

STATENS SERUM INSTITUT

Center for Genomic Epidemiology

Username
Password
New Reset Login

Home | Services | Instructions | Output | Article abstract

MINTyper 1.0

SNP distance matrice and phylogenetic tree with long and short raw sequencing reads or with assembled genomes.

*** For large datasets (>100 isolates), consider running the analysis locally, as uploading large quantities of data to the webserver may be troublesome. For a local installation of MINTyper, please see https://bitbucket.org/genomicepidemiology/mintyper**

View the version history of this server.

Single reference of your choosing
Note: If you would like to choose a Vælg fil Der er ingen fil valgt

Select the host database
Bacteria organisms (KmerFinder DB)

Motif masking
No masking

Prune significance
Significant calls only

Pruning length:
The pruning length should be non-negative - the default is 10
10

Cluster length:
Maximum SNP distance to determine if two isolates belongs to the same cluster.
10

Input files: fastq and fasta formats are supported, fastq are recommended. Assemblies are not handled yet. Note: 2 or more samples are required as input!

- MinTyper can search (a now a bit outdated version of) the NCBI RefSeq genome database (KmerFinder DB) for the best reference.
- You can also upload your own reference (e.g. a draft genome of what you think is your index isolate).

MINTYPER V1.0

STATENS SERUM INSTITUT

Center for Genomic Epidemiology

Username
Password
New Reset Login

Home | Services | Instructions | Output | Article abstract

MINTyper 1.0

SNP distance matrice and phylogenetic tree with long and short raw sequencing reads or with assembled genomes.

*** For large datasets (>100 isolates), consider running the analysis locally, as uploading large quantities of data to the webserver may be troublesome. For a local installation of MINTyper, please see https://bitbucket.org/genomicepidemiology/mintyper**

View the version history of this server.

Single reference of your choosing
Note: If you would like to choose a Vælg fil Der er ingen fil valgt

Select the host database
Bacteria organisms (KmerFinder DB)

Motif masking
No masking

Prune significance
Significant calls only

Pruning length:
The pruning length should be non-negative - the default is 10
10

Cluster length:
Maximum SNP distance to determine if two isolates belongs to the same cluster.
10

Input files: fastq and fasta formats are supported, fastq are recommended. Assemblies are not handled yet. Note: 2 or more samples are required as input!

- Choose no masking if you have Illumina data and/or MinION data, which has been basecalled to correct for Dcm methylation.
- If your Illumina data and MinION data of the same strain does not allign in the analysis, try to apply the ”DCM masking option”

MINTYPER V1.0

STATENS SERUM INSTITUT

Center for Genomic Epidemiology

Username
Password
New Reset Login

Home | Services | Instructions | Output | Article abstract

MINTyper 1.0

SNP distance matrice and phylogenetic tree with long and short raw sequencing reads or with assembled genomes.

*** For large datasets (>100 isolates), consider running the analysis locally, as uploading large quantities of data to the webserver may be troublesome. For a local installation of MINTyper, please see https://bitbucket.org/genomicepidemiology/mintyper**

View the version history of this server.

Single reference of your choosing
Note: If you would like to choose a Vælg fil Der er ingen fil valgt

Select the host database
Bacteria organisms (KmerFinder DB)

Motif masking
No masking

Prune significance
Significant calls only

- Significant calls are HQ SNPs
- Insignificant calls include more ambiguous calls (not advised).

Pruning length:
The pruning length should be non-negative - the default is 10
10

Cluster length:
Maximum SNP distance to determine if two isolates belongs to the same cluster.
10

Input files: fastq and fasta formats are supported, fastq are recommended. Assemblies are not handled yet. Note: 2 or more samples are required as input!

MINTYPER V1.0

STATENS SERUM INSTITUT

Center for Genomic Epidemiology

Username
Password
New Reset Login

Home | Services | Instructions | Output | Article abstract

MINTyper 1.0

SNP distance matrice and phylogenetic tree with long and short raw sequencing reads or with assembled genomes.

*** For large datasets (>100 isolates), consider running the analysis locally, as uploading large quantities of data to the webserver may be troublesome. For a local installation of MINTyper, please see https://bitbucket.org/genomicepidemiology/mintyper**

View the version history of this server.

Single reference of your choosing
Note: If you would like to choose a Vælg fil Der er ingen fil valgt

Select the host database
Bacteria organisms (KmerFinder DB)

Motif masking
No masking

Prune significance
Significant calls only

Pruning length:
The pruning length should be non-negative - the default is 10
10

- Select pruning distance.
- Use default or perhaps 100 bp.

Cluster length:
Maximum SNP distance to determine if two isolates belongs to the same cluster.
10

Input files: fastq and fasta formats are supported, fastq are recommended. Assemblies are not handled yet. Note: 2 or more samples are required as input!

MINTYPER V1.0

STATENS SERUM INSTITUT

Center for Genomic Epidemiology

Username
Password
New Reset Login

Home | Services | Instructions | Output | Article abstract

MINTyper 1.0

SNP distance matrice and phylogenetic tree with long and short raw sequencing reads or with assembled genomes.

*** For large datasets (>100 isolates), consider running the analysis locally, as uploading large quantities of data to the webserver may be troublesome. For a local installation of MINTyper, please see https://bitbucket.org/genomicepidemiology/mintyper**

View the version history of this server.

Single reference of your choosing
Note: If you would like to choose a Vælg fil Der er ingen fil valgt

Select the host database
Bacteria organisms (KmerFinder DB)

Motif masking
No masking

Prune significance
Significant calls only

Pruning length:
The pruning length should be non-negative - the default is 10
10

Cluster length:
Maximum SNP distance to determine if two isolates belongs to the same cluster.
10

Input files: fastq and fasta formats are supported, fastq are recommended. Assemblies are not handled yet. Note: 2 or more samples are required as input!

- Define a SNP distance for clusters
- Often between 10 and 20 (but depends on the length and nature of the outbreak).

UPLOADING DATA

STATENS SERUM INSTITUT

Input files: fastq and fasta formats are supported, fastq are recommended. Assemblies are not handled yet. Note: 2 or more samples are required as input!

Choose File(s)

Name

Status

Upload

Remove

- Click here to find your data
- Raw data only!
- Can not exceed around 1 GB per file

- Click and run the analysis

REFERENCES

1. Clausen PTLC, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics **2018**; 19:307.

Center for Genomic Epidemiology

Your job is being processed

Wait here to watch the progress of your job, or fill in the form below to get an email message upon completion.

To get notified by email: henh@ssi.dk Notify me via email

This page will update itself automatically.

- Insert your email address

Center for Genomic Epidemiology

Your job is being processed

Wait here to watch the progress of your job, or fill in the form below to get an email message upon completion.
henh@ssi.dk

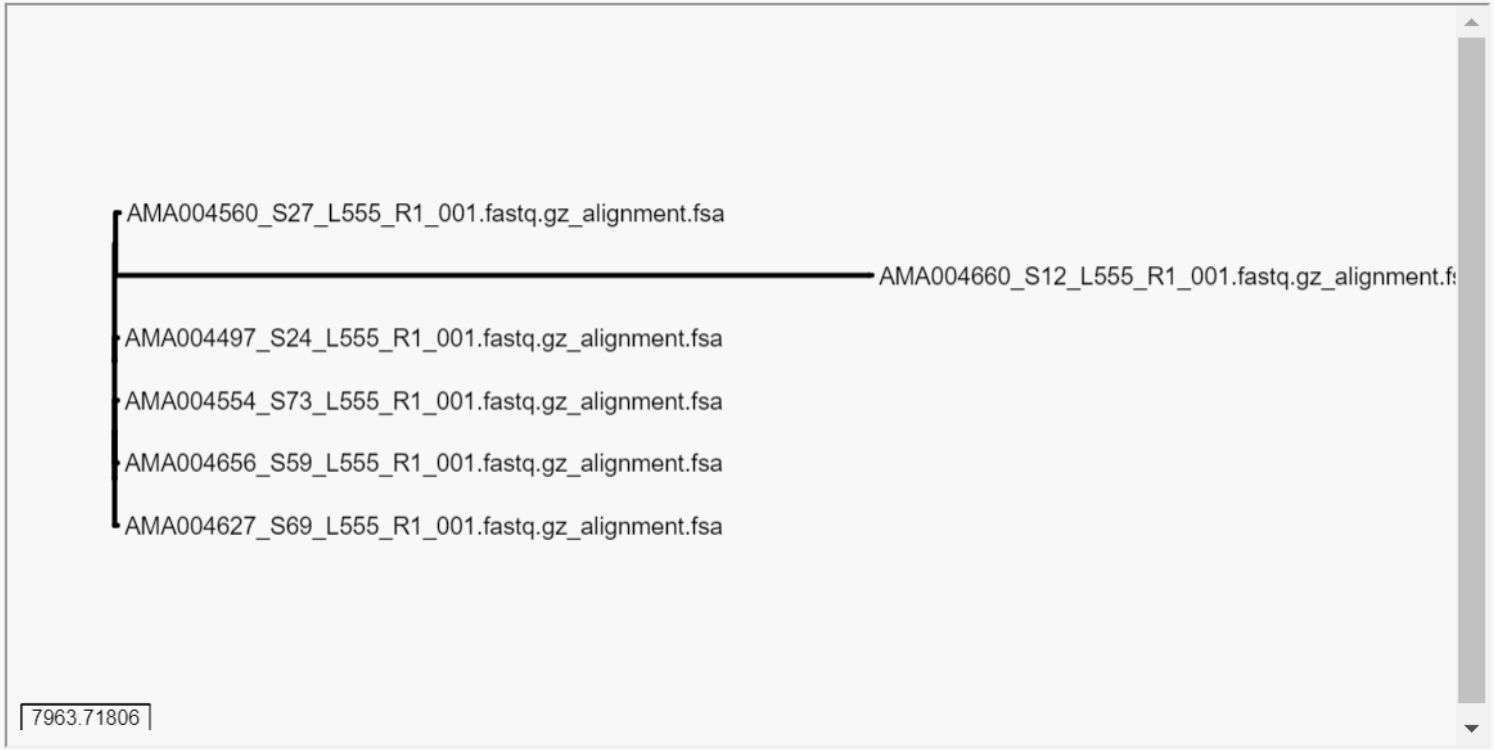
To get notified by email: Notify me via

This page will update itself automatically.

- Then wait for the result (if you start many different analysis, it is advised to make a log of what you have started and with what settings…and perhaps also the hypothesis).

MINTYPER OUTPUT

STATENS SERUM INSTITUT



Percentage of reference covered by all isolates: 84.71 (4149824 / 4899014)

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

Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67
AMA004660_S12_L555_R1_001.fastq.gz	4327141	88.33

Log | Distance matrix | Phylogentic tree | Vcf files of mutations | Reference Sequence | Cluster.dbscan

MINTYPER OUTPUT

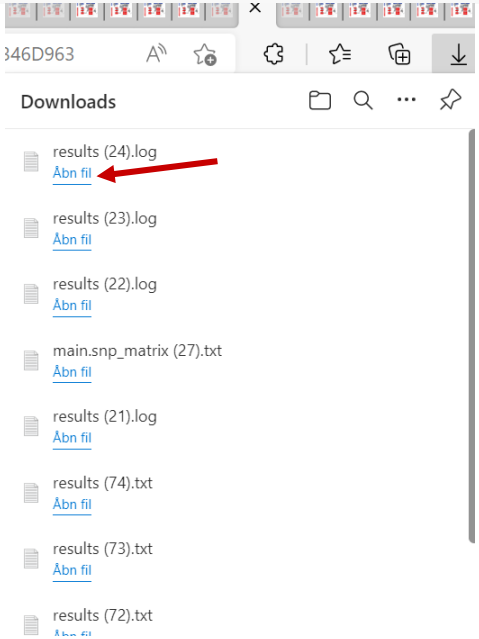
STATENS SERUM INSTITUT

Percentage of reference covered by all isolates: 84.71 (4149824 / 4899014)

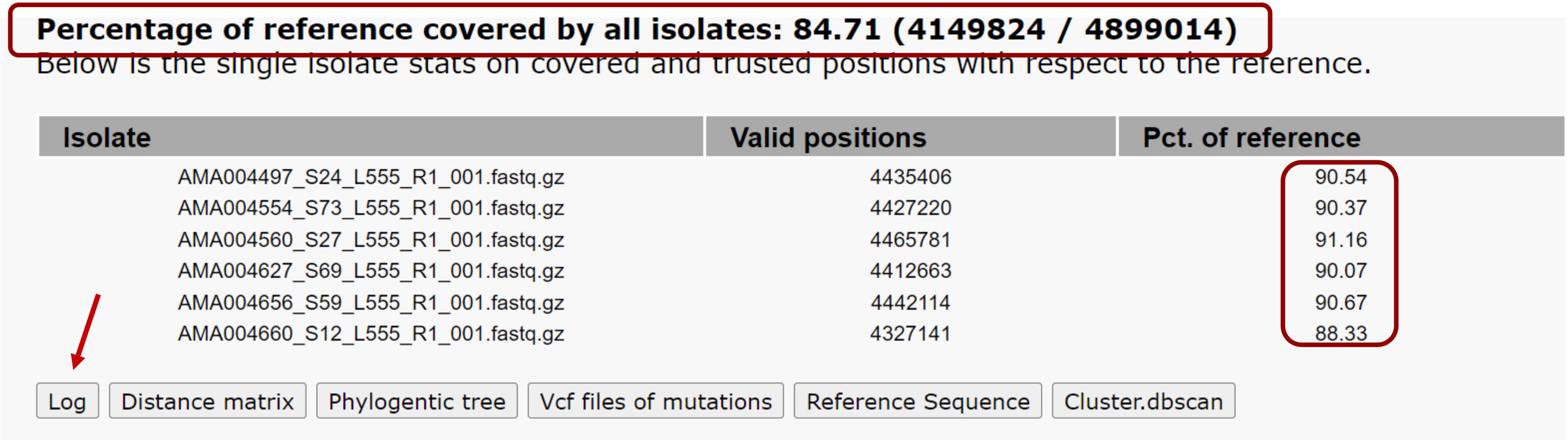
Below is the single isolate stats on covered and trusted positions with respect to the reference.

Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67
AMA004660_S12_L555_R1_001.fastq.gz	4327141	88.33

Log | Distance matrix | Phylogentic tree | Vcf files of mutations | Reference Sequence | Cluster.dbscan



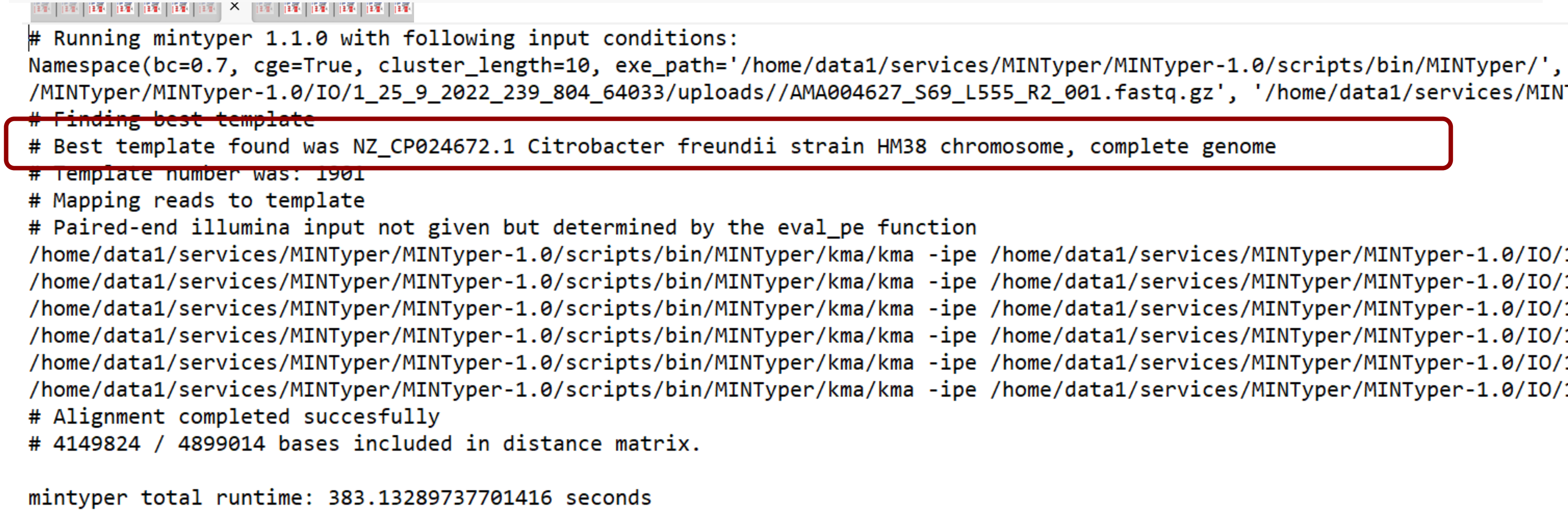
MINTYPER OUTPUT



Percentage of reference covered by all isolates: 84.71 (4149824 / 4899014)

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

Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67
AMA004660_S12_L555_R1_001.fastq.gz	4327141	88.33



```
# Running mintyper 1.1.0 with following input conditions:
Namespace(bc=0.7, cge=True, cluster_length=10, exe_path='/home/data1/services/MINTyper/MINTyper-1.0/scripts/bin/MINTyper/',
/MINTyper/MINTyper-1.0/IO/1_25_9_2022_239_804_64033/uploads//AMA004627_S69_L555_R2_001.fastq.gz', '/home/data1/services/MINT
# Finding best template
# Best template found was NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome
# Template number was: 1901
# Mapping reads to template
# Paired-end illumina input not given but determined by the eval_pe function
/home/data1/services/MINTyper/MINTyper-1.0/scripts/bin/MINTyper/kma/kma -ipe /home/data1/services/MINTyper/MINTyper-1.0/IO/
/home/data1/services/MINTyper/MINTyper-1.0/scripts/bin/MINTyper/kma/kma -ipe /home/data1/services/MINTyper/MINTyper-1.0/IO/
/home/data1/services/MINTyper/MINTyper-1.0/scripts/bin/MINTyper/kma/kma -ipe /home/data1/services/MINTyper/MINTyper-1.0/IO/
/home/data1/services/MINTyper/MINTyper-1.0/scripts/bin/MINTyper/kma/kma -ipe /home/data1/services/MINTyper/MINTyper-1.0/IO/
/home/data1/services/MINTyper/MINTyper-1.0/scripts/bin/MINTyper/kma/kma -ipe /home/data1/services/MINTyper/MINTyper-1.0/IO/
/home/data1/services/MINTyper/MINTyper-1.0/scripts/bin/MINTyper/kma/kma -ipe /home/data1/services/MINTyper/MINTyper-1.0/IO/
# Alignment completed succesfully
# 4149824 / 4899014 bases included in distance matrix.

mintyper total runtime: 383.13289737701416 seconds
```

MINTYPER OUTPUT

STATENS SERUM INSTITUT

Percentage of reference covered by all isolates: 84.71 (4149824 / 4899014)

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

Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67
AMA004660_S12_L555_R1_001.fastq.gz	4327141	88.33

ST18

ST91

Log | Distance matrix | Phylogentic tree | Vcf files of mutations | Reference Sequence | Cluster.dbscan

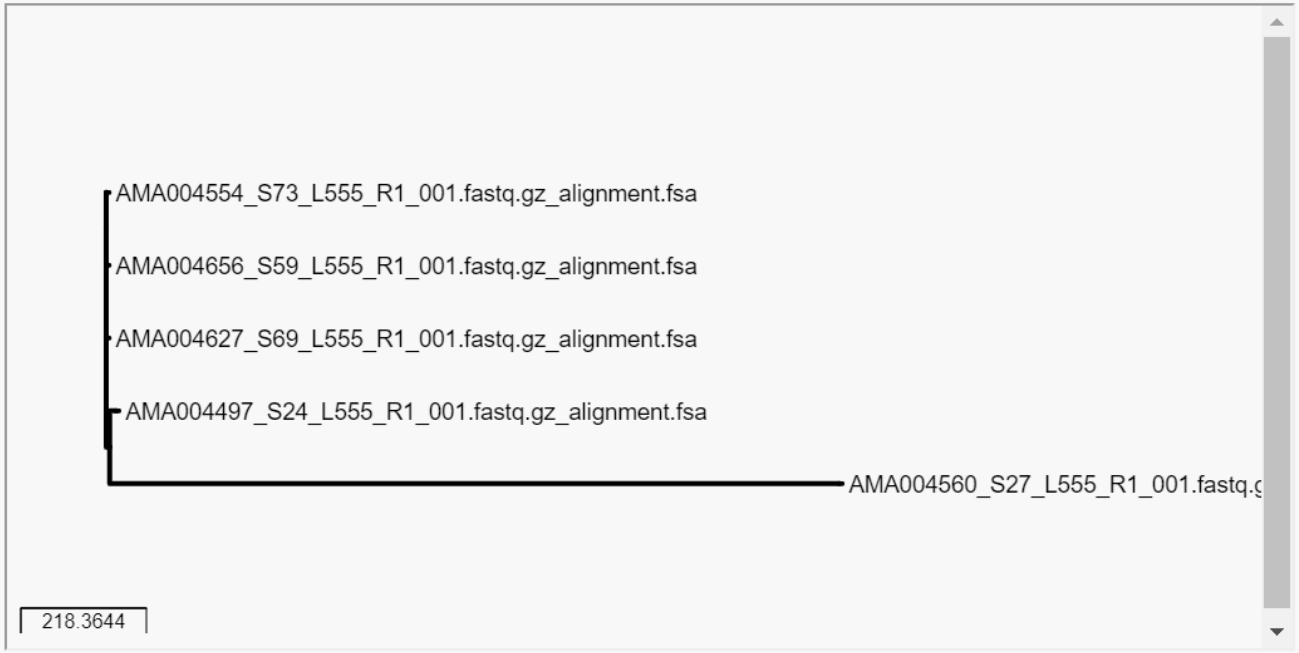
		1	2	3	4	5	6
	6						
1	AMA004497_S24_L555_R1_001.fastq.gz_alignment.fsa	0					
2	AMA004554_S73_L555_R1_001.fastq.gz_alignment.fsa	15	0				
3	AMA004560_S27_L555_R1_001.fastq.gz_alignment.fsa	133	130	0			
4	AMA004627_S69_L555_R1_001.fastq.gz_alignment.fsa	15	0	130	0		
5	AMA004656_S59_L555_R1_001.fastq.gz_alignment.fsa	15	0	130	0	0	
6	AMA004660_S12_L555_R1_001.fastq.gz_alignment.fsa	46761	46758	46758	46758	46758	0

RERUN WITHOUT AMA004660

STATENS SERUM INSTITUT

Center for Genomic Epidemiology

Home | Services | Instructions | Output



Percentage of reference covered by all isolates: 89.18 (4368832 / 4899014)

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

Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67

Log | Distance matrix | Phylogentic tree | Vcf files of mutations | Reference Sequence | Cluster.dbscan

RERUN WITHOUT AMA004660

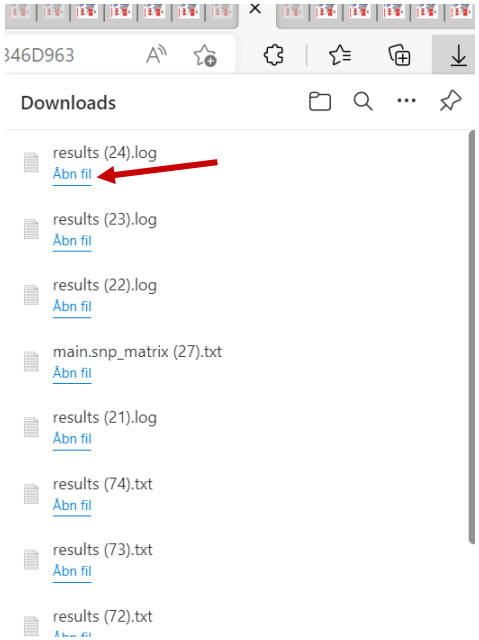
STATENS SERUM INSTITUT

Percentage of reference covered by all isolates: 89.18 (4368832 / 4899014)

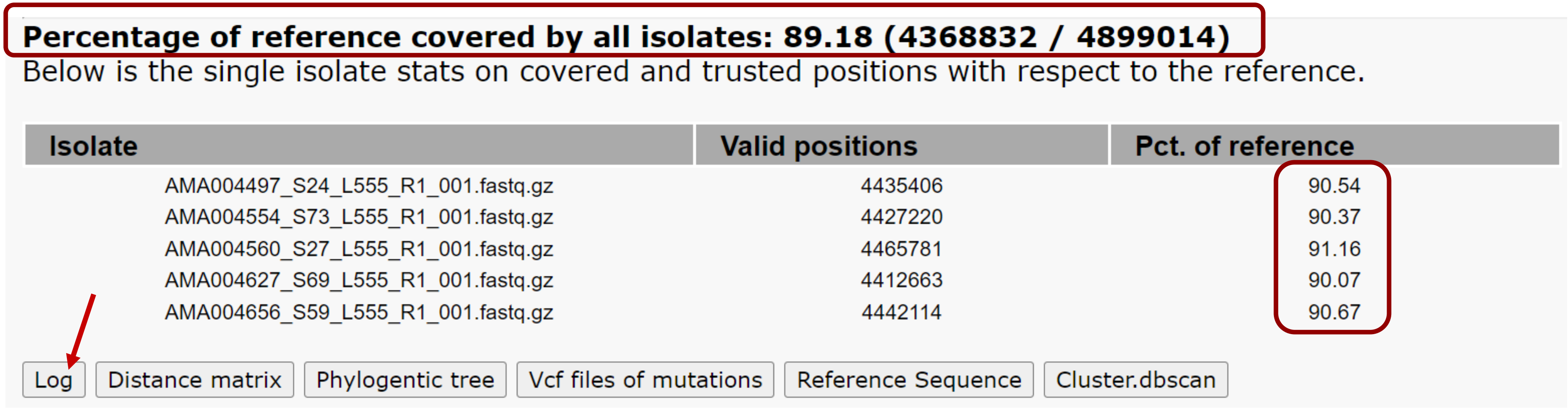
Below is the single isolate stats on covered and trusted positions with respect to the reference.

Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67

Log | Distance matrix | Phylogentic tree | Vcf files of mutations | Reference Sequence | Cluster.dbscan



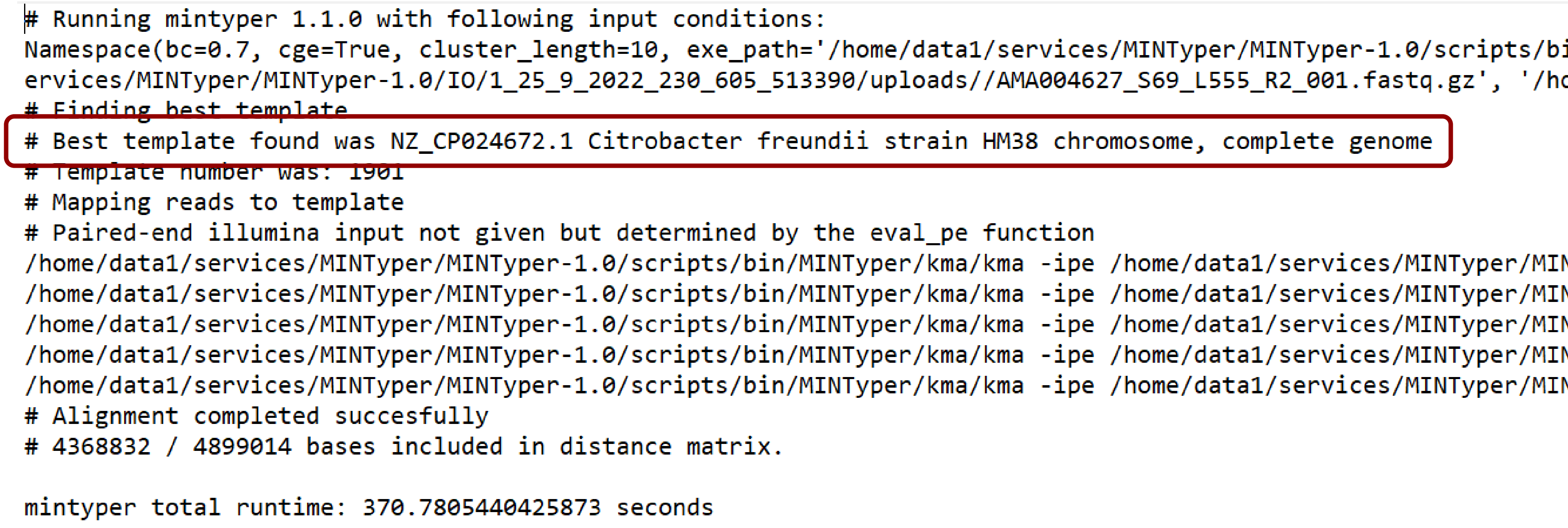
RERUN WITHOUT AMA004660



Percentage of reference covered by all isolates: 89.18 (4368832 / 4899014)

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

Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67



```
# Running mintyper 1.1.0 with following input conditions:
Namespace(bc=0.7, cge=True, cluster_length=10, exe_path='/home/data1/services/MINTyper/MINTyper-1.0/scripts/bi
ervices/MINTyper/MINTyper-1.0/IO/1_25_9_2022_230_605_513390/uploads//AMA004627_S69_L555_R2_001.fastq.gz', '/h
# Finding best template
# Best template found was NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome
# Template number was: 1901
# Mapping reads to template
# Paired-end illumina input not given but determined by the eval_pe function
/home/data1/services/MINTyper/MINTyper-1.0/scripts/bin/MINTyper/kma/kma -ipe /home/data1/services/MINTyper/MIN
/home/data1/services/MINTyper/MINTyper-1.0/scripts/bin/MINTyper/kma/kma -ipe /home/data1/services/MINTyper/MIN
/home/data1/services/MINTyper/MINTyper-1.0/scripts/bin/MINTyper/kma/kma -ipe /home/data1/services/MINTyper/MIN
/home/data1/services/MINTyper/MINTyper-1.0/scripts/bin/MINTyper/kma/kma -ipe /home/data1/services/MINTyper/MIN
/home/data1/services/MINTyper/MINTyper-1.0/scripts/bin/MINTyper/kma/kma -ipe /home/data1/services/MINTyper/MIN
# Alignment completed succesfully
# 4368832 / 4899014 bases included in distance matrix.

mintyper total runtime: 370.7805440425873 seconds
```

RERUN WITHOUT AMA004660

STATENS SERUM INSTITUT

Percentage of reference covered by all isolates: 89.18 (4368832 / 4899014)

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

Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67

Log | Distance matrix | Phylogentic tree | Vcf files of mutations | Reference Sequence | Cluster.dbscan

		1	2	3	4	5
1	AMA004497_S24_L555_R1_001.fastq.gz_alignment.fsa					
2	AMA004554_S73_L555_R1_001.fastq.gz_alignment.fsa	17	0			
3	AMA004560_S27_L555_R1_001.fastq.gz_alignment.fsa	1280	1275	0		
4	AMA004627_S69_L555_R1_001.fastq.gz_alignment.fsa	17	0	1275	0	
5	AMA004656_S59_L555_R1_001.fastq.gz_alignment.fsa	17	0	1275	0	0

RERUN WITHOUT AMA004660

STATENS SERUM INSTITUT

Percentage of reference covered by all isolates: 89.18 (4368832 / 4899014)

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

Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67

Log | Distance matrix | Phylogentic tree | Vcf files of mutations | Reference Sequence | Cluster.dbscan

		1	2	3	4	5
1	AMA004497_S24_L555_R1_001.fastq.gz_alignment.fsa	0				
2	AMA004554_S73_L555_R1_001.fastq.gz_alignment.fsa	17	0			
3	AMA004560_S27_L555_R1_001.fastq.gz_alignment.fsa	1280	1275	0		
4	AMA004627_S69_L555_R1_001.fastq.gz_alignment.fsa	17	0	1275	0	
5	AMA004656_S59_L555_R1_001.fastq.gz_alignment.fsa	17	0	1275	0	0

MINTYPER OUTPUT - VISUALIZATIONS

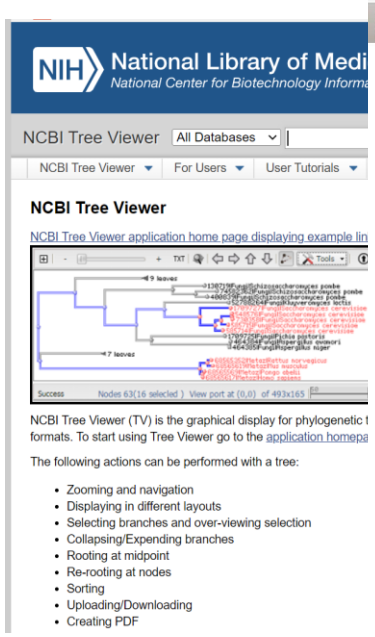
STATENS SERUM INSTITUT

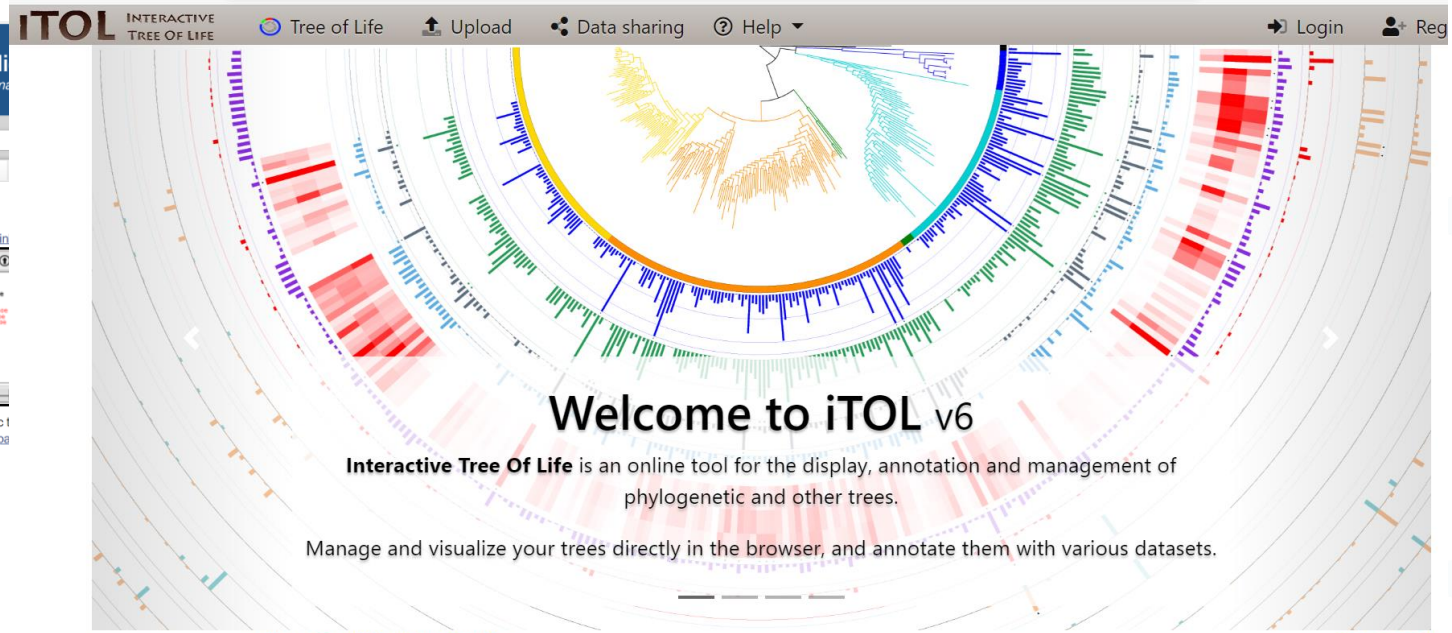
Percentage of reference covered by all isolates: 89.18 (4368832 / 4899014)

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

Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67

Log | Distance matrix | Phylogentic tree | Vcf files of mutations | Reference Sequence | Cluster.dbscan





Note: See the details on iTOL access modes and subscriptions

Current changelog: version 6.5.8

MINTYPER OUTPUT – VCF DATA

STATENS SERUM INSTITUT

Percentage of reference covered by all isolates: 89.18 (4368832 / 4899014)

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

Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67

Log | Distance matrix | Phylogentic tree | Vcf files of mutations | Reference Sequence | Cluster.dbscan

AMA004497_S24_L555_R1_001.fastq.gz_alignment.vcf - Notesblok

Filer Rediger Formater Vis Hjælp

```
##fileformat=VCFv4.2
##kmaVersion=1.4.2
##FILTER=<ID=LowQual,Description="Low quality">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AD,Number=1,Type=Integer,Description="Allele Depth">
##INFO=<ID=AF,Number=1,Type=Float,Description="Allele Fraction">
##INFO=<ID=RAF,Number=1,Type=Float,Description="Revised Allele Fraction">
##INFO=<ID=DEL,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=AD6,Number=6,Type=Integer,Description="Count of all alternative alleles: A,C,G,T,N,-">
##FORMAT=<ID=Q,Number=1,Type=Float,Description="McNemar quantile">
##FORMAT=<ID=P,Number=1,Type=Float,Description="McNemar p-value">
##FORMAT=<ID=FT,Number=1,Type=String,Description="Filter">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  bacteria.ATG
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    338     .       A       a       277     .       DP=76;AD=65;AF=0.86;RAF=0.86
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    471     .       A       G       367     .       DP=61;AD=61;AF=1.00;RAF=1.00
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    489     .       C       T       325     .       DP=54;AD=54;AF=1.00;RAF=1.00
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    492     .       G       T       314     .       DP=56;AD=55;AF=0.98;RAF=0.98
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    508     .       T       C       264     .       DP=44;AD=44;AF=1.00;RAF=1.00
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    672     .       C       T       273     .       DP=49;AD=48;AF=0.98;RAF=0.98
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    756     .       A       a       200     .       DP=50;AD=44;AF=0.88;RAF=0.88
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    760     .       A       a       194     .       DP=49;AD=43;AF=0.88;RAF=0.88
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    894     .       T       C       270     .       DP=45;AD=45;AF=1.00;RAF=1.00
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1251    .       C       T       338     .       DP=60;AD=59;AF=0.98;RAF=0.98
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1548    .       T       G       559     .       DP=97;AD=96;AF=0.99;RAF=0.99
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1549    .       T       t       361     .       DP=94;AD=82;AF=0.87;RAF=0.87
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1568    .       C       c       355     .       DP=88;AD=78;AF=0.89;RAF=0.89
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1569    .       A       G       529     .       DP=88;AD=88;AF=1.00;RAF=1.00
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1594    .       A       a       336     .       DP=87;AD=76;AF=0.87;RAF=0.87
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1597    .       A       a       324     .       DP=87;AD=75;AF=0.86;RAF=0.86
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1604    .       T       t       361     .       DP=89;AD=79;AF=0.89;RAF=0.89
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1612    .       A       a       304     .       DP=81;AD=70;AF=0.86;RAF=0.86
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1743    .       G       T       385     .       DP=64;AD=64;AF=1.00;RAF=1.00
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1753    .       T       G       379     .       DP=63;AD=63;AF=1.00;RAF=1.00
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1764    .       C       T       385     .       DP=64;AD=64;AF=1.00;RAF=1.00
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1773    .       C       T       391     .       DP=65;AD=65;AF=1.00;RAF=1.00
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1777    .       T       C       379     .       DP=63;AD=63;AF=1.00;RAF=1.00
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    1816    .       G       T       392     .       DP=69;AD=68;AF=0.99;RAF=0.99
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    2047    .       A       C       270     .       DP=45;AD=45;AF=1.00;RAF=1.00
NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome    2100    .       A       G       344     .       DP=61;AD=60;AF=0.98;RAF=0.98
```

MINTYPER OUTPUT – REFERENCE

STATENS SERUM INSTITUT

Percentage of reference covered by all isolates: 89.18 (4368832 / 4899014)

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

Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67

Log | Distance matrix | Phylogentic tree | Vcf files of mutations | Reference Sequence | Cluster.dbscan

template_sequence (2) - Notesblok

Filer Rediger Formater Vis Hjælp

```
>NZ_CP024672.1 Citrobacter freundii strain HM38 chromosome, complete genome
TCGGAATGTGATCAATTTAAAAATTTATTGACTTAGGTCAGCAGATACTTTAACCCTAAAAGAATACAAGACAGACAGATAATAATTACAGAGCACACAACATCATGAAACGCATCAGCATCACTATTACCACAACCATCACCATTACCACAGC
ACTTCGGCGCCAAAGTGCTGCACCCGCGCACCATTACCCCTATTGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATACCGGCAATCCACAAGCGCCTGGCACGTTGATTGGCGCCAGCAGTGATGACGACGATTTGCCGGTAAAAGGTATTTC
GCGTAAATTCCTCTACGACACCAACGTGGGCGCAGGCTTGCCGGTAATTGAAAACCTGCAAAACCTGCTCAGCGCAGGTGATGAATTGCAGCGTTTCTCCGGTATTCTTTCCGGCTCGCTGTCGTTTATTTTCGGCAAGCTGGATGAAGGCATC
TACGCGCCTGTTGGCGATAATGGGCGAGCTGGAAGGGCGTATCTCCGGCAGTATTCATTACGATAATGTTGCCCCGTGCTTTTTGGGCGGCATGCAGTTAATGATCGAAGAAAACGGCATTATCAGCCAGCAGGTGCCAGGCTTTGATGAGTGC
GTGCGCGTGGTGATCCTCTACCCGAACGGCAAGATCAGCCCATTACAGGAAAAACTGTTCTGTACGCTGGGCGGTAACATTGAAACTGTAGCGATTGATGGCGATTTTGATGCCTGCCAGGCGCTGGTGAAACAGGCGTTTGATGACGAAGAGC
AAGCGGTTCTCAATGATGTAGCGGCTCATCAGGCCGCGCGCTTTCTTGGCATAAAAGCTGATGACTTTGAACTTGCCGTTTTTCTCGTCTAGGAAGACTGGTTTAATCAGCTCGGCATTGAGCTTTTTCGGCTTCACGGATTTAAAATACTCAT
GATGCGCATATTGTCCAGTTTCAGAAAGACGAGATTGTTCTCCGCATAGATGTAGTTGGCGACGATCGAGCTAAAGGCAAACAGAATCACAATAAAAGCCACGAATCCGGCTCCCCATTCTCCCGTCAGCGTTACCATTGCTTTTTGAAGTAAC
CCAAACTGGCGAATGTAGCGGAACTGCACGAACCCTGTACGCCAGGTGAACCAACATCCTGCGCCAAGCAGGAGGTAGATCATCACCGACCCCCAGAGTATTTCATTAATAAATGAGAAAAAATCAGGCATTAACATCCCTCTTGTTGATGCCC
CGGAAATCTACGAGTACTACAAACAGCACGGTTACGAAACCGTGGTGATGGGTGCGAGCTTCCGTAACATCGGCGAAATCATTGAGCTGGCAGGCTGCGATCGTCTGACGATTGCTCCGGCACTGCTGAAAGAGCTGGCGGAAAGCGAAGGCGC
AACAGCTGAAAGTAGCCCGATCGTGACATAAGTGCCAATCACAGTGGAGTATTGAATTGTTTACGATATAGTAATTTTTGCTTTACATATATCGACAGATAGATTCAATAACTCCACTGACTGATGGTTTGCTATGTCACAACACACTCGACCC
GGATTTATTGTCGACCGGGTCGGCGCCAGCAAAACCTTTATTGTCGGCAGCCTGCTGCTGGCCTGTTCGAGCTGGTTTTTCTATCACCTGACCGGCAGCCATCCTCAGCACTTGTTCCTGTTATACGGGCTGGTGGGATTATGCGTTGGCGTGC
CAAAATCATCATTGATTGAATGGTGAAATAGTTTCCCCAAATAACGATCACTGTCTTCGGGGCGCGGCATAATAATCAGGGGAGAGGGCACTGTCTATGATCTAACGAAGGGAAAACGAATTATTTTCCCTGTGATGGGCATCACGCTTGTGCC
TTGATGGAGGAAGGCACGTTCAGACTCTGGCTGGACATGCGGGCAGCTTTGAAATAAACCGATGCACCACTGAGCTGTAAATCACCATGATCGGCCGTAAGTTGAATGCGTTTCACCACGCGGCAAACGGGAAGTTTCAGCGTCAGATCGTTGC
AAAACGGCTGAAGATTACCTGGGTGAGCCGGTAACTGAAGCAGTTATCACCGTACCTGCATACTTTAACGATGCTCAGCGTCAGGCAACTAAAGACGCCGGTCGTATCGCAGGTCTGGAAGTAAAACGTATCATCAACGAACCAACCGCAGCTC
CGCGCGGTATGCCGCAGATCGAAGTTACTTTTGACATCGATGCCGACGGTATCCTGCACGTTTCCGCGAAAGACAAAAACAGCGGTAAAGAGCAGAAGATCACCATCAAGGCTTCTTCTGGTCTGAACGAAGAAGAAATTCAGAAAATGGTTCC
GTATTCCGACTCTGGAAGAGTGTGACGTCTGCCACGGTAGCGGCGCGAAGGCGGGCACTCAGCCGCAGACCTGTCCAACCTGTCACGGTTCTGGTCAGGTACAGATGCGTCAGGGCTTTTTTGCCGTACAGCAGACCTGTCCACACTGTCAGGC
ACACTACCAGTAATAACGATTCACCTTCCTTAATGTGAGATTTACATCACTATCCTGTCATTCTATTTCATTATTATTTGTCACCGCGGTAACACAATATTTATTGATGGATTAATTATAATTATGGGGAATACATTCAATGCCTCGGATAAAA
TGATTATTATGGCTGATGATTTGGGCTATGGCGATCTCGCGACTTATGGCCATCAAATCGTCAAAACGCCTAATATCGACAAGCTGGCACAGGAAGGGGTGAAATTTACCGACTACTATGCGCCCGCGCCGCTGTGCTCTCCTTCTCGTGCGGC
CATGATGAATTTTAAACTGCCAACAGATCGTACCTACGATGGGCAGTCTCTGGTTCCATTACTTGAACAGAAAACGTTAGCACGTCAGAAACCACTCATCTTTGGCATTGATATGCCGTTCCAGGATGATCCTACTGACGAATGGGCGATCCGT
GAAAAGCATTGGCAGTCGCTATATGCAGTTTATTCCGCTCATCGAAAGAAGCGCTGAAGAGGGCGGTATTGGTCTGGCACATCCTGAAGGCGGACAAACAAGAATGATGCCGTGGTCGGTTGATTCACTGCAATTTGGCCATTTTCTCAACGCT
```

MINTYPER OUTPUT – CLUSTER ANALYSIS

STATENS SERUM INSTITUT

Percentage of reference covered by all isolates: 89.18 (4368832 / 4899014)
Below is the single isolate stats on covered and trusted positions with respect to the reference.

Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67

Log | Distance matrix | Phylogentic tree | Vcf files of mutations | Reference Sequence | Cluster.dbscan

other Cluster # isolates

```
# 5      3        10.000000        1
"AMA004497_S24_L555_R1_001.fastq.gz_alignment.fsa"      0        0
"AMA004554_S73_L555_R1_001.fastq.gz_alignment.fsa"      2        1
"AMA004560_S27_L555_R1_001.fastq.gz_alignment.fsa"      0        2
"AMA004627_S69_L555_R1_001.fastq.gz_alignment.fsa"      2        1
"AMA004656_S59_L555_R1_001.fastq.gz_alignment.fsa"      2        1
```

```
         5
AMA004497_S24_L555_R1_001.fastq.gz_alignment.fsa
AMA004554_S73_L555_R1_001.fastq.gz_alignment.fsa        17
AMA004560_S27_L555_R1_001.fastq.gz_alignment.fsa        1280     1275
AMA004627_S69_L555_R1_001.fastq.gz_alignment.fsa        17       0        1275
AMA004656_S59_L555_R1_001.fastq.gz_alignment.fsa        17       0        1275     0
```

THE EXCERCISE - GUIDELINES

STATENS SERUM INSTITUT

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 rerun 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 choosen 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.This is because

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

THE EXCERCISE - COMPARISONS

STATENS SERUM INSTITUT

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 Gubby 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.

THE EXCERCISE – ANALYSIS EXAMPLES

STATENS SERUM INSTITUT

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

THE END

STATENS
SERUM
INSTITUT