Whole Genome Sequencing and Bioinformatics SeqAfrica Training

2-5 June 2025
Lusaka

Marco van Zwetselaar

Niamh Lacy-Roberts

Day 4

The Fleming Fund Regional Grants

UKaid from the British people

DTU

SeqAfrica

The Fleming Fund
Regional Grants

Typing and Phylogenetic Analysis

Multi-Locus Sequence Typing (MLST)

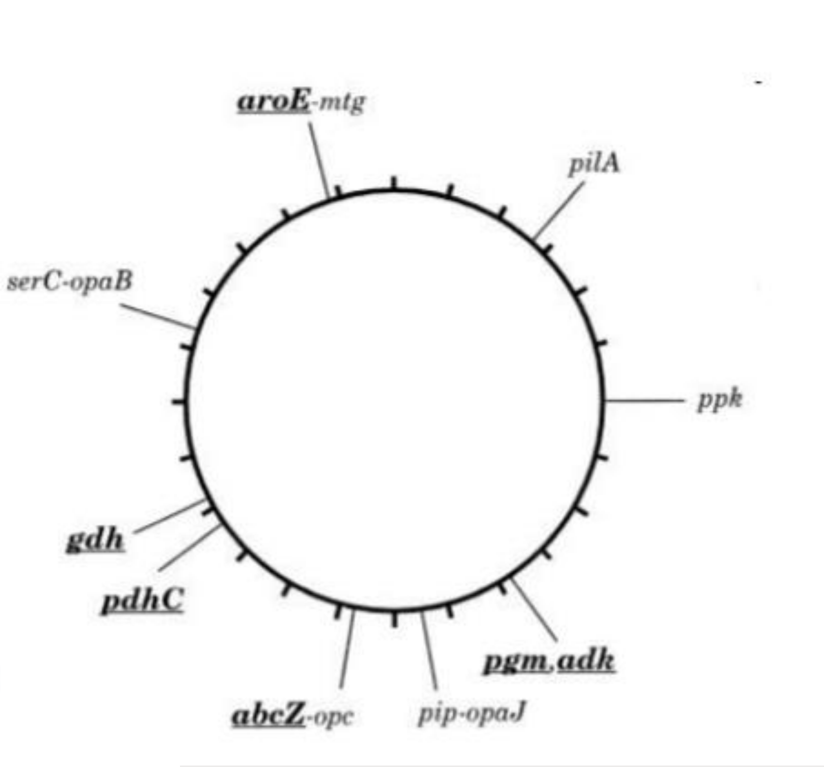
Classical MLST:

- The (old) gold 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).



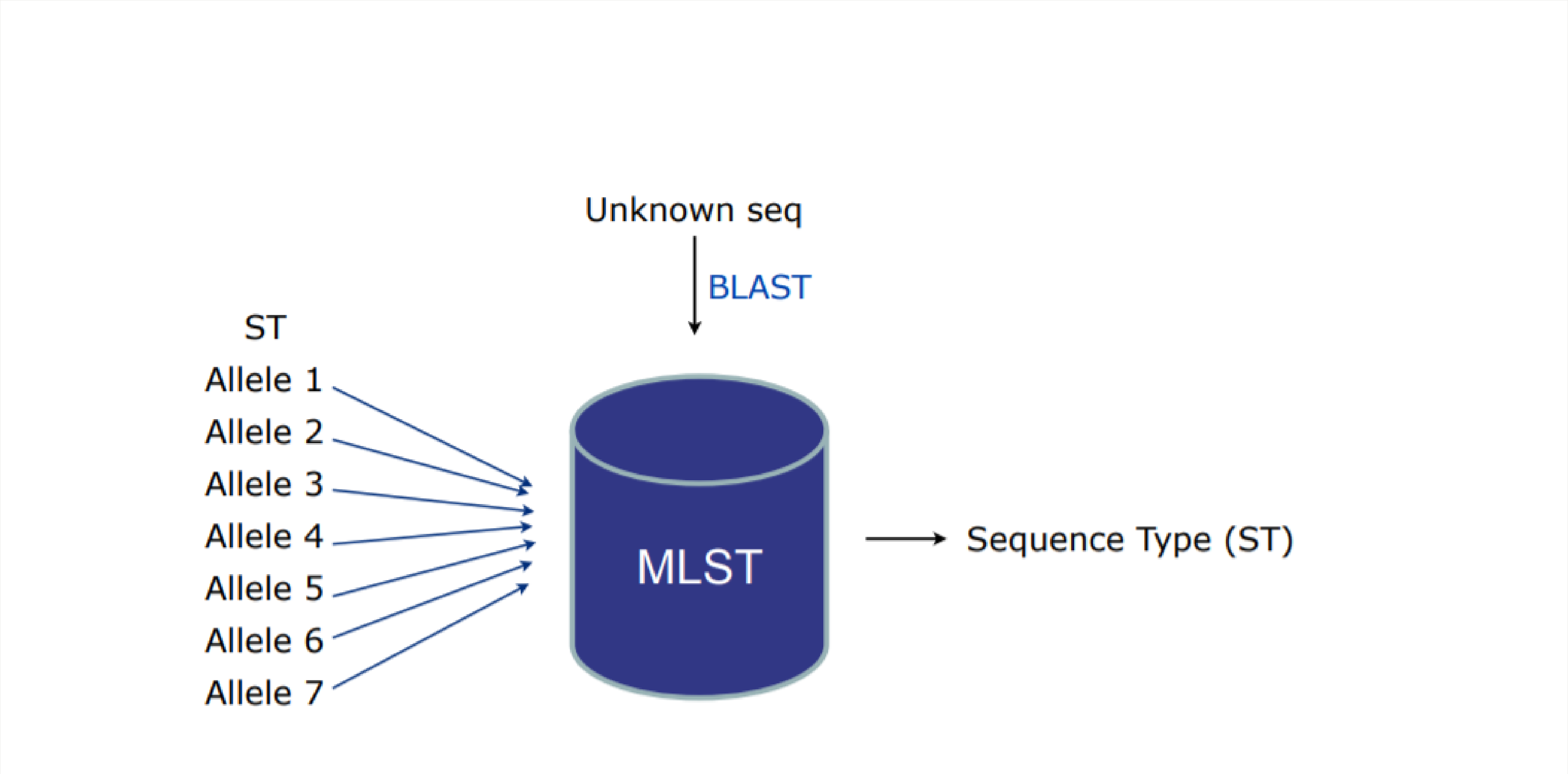
MLST now

- 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 continues to decline, it becomes increasingly available to scientists and routine diagnostics laboratories.
  - Currently, the WGS cost is typically below that of traditional MLST.

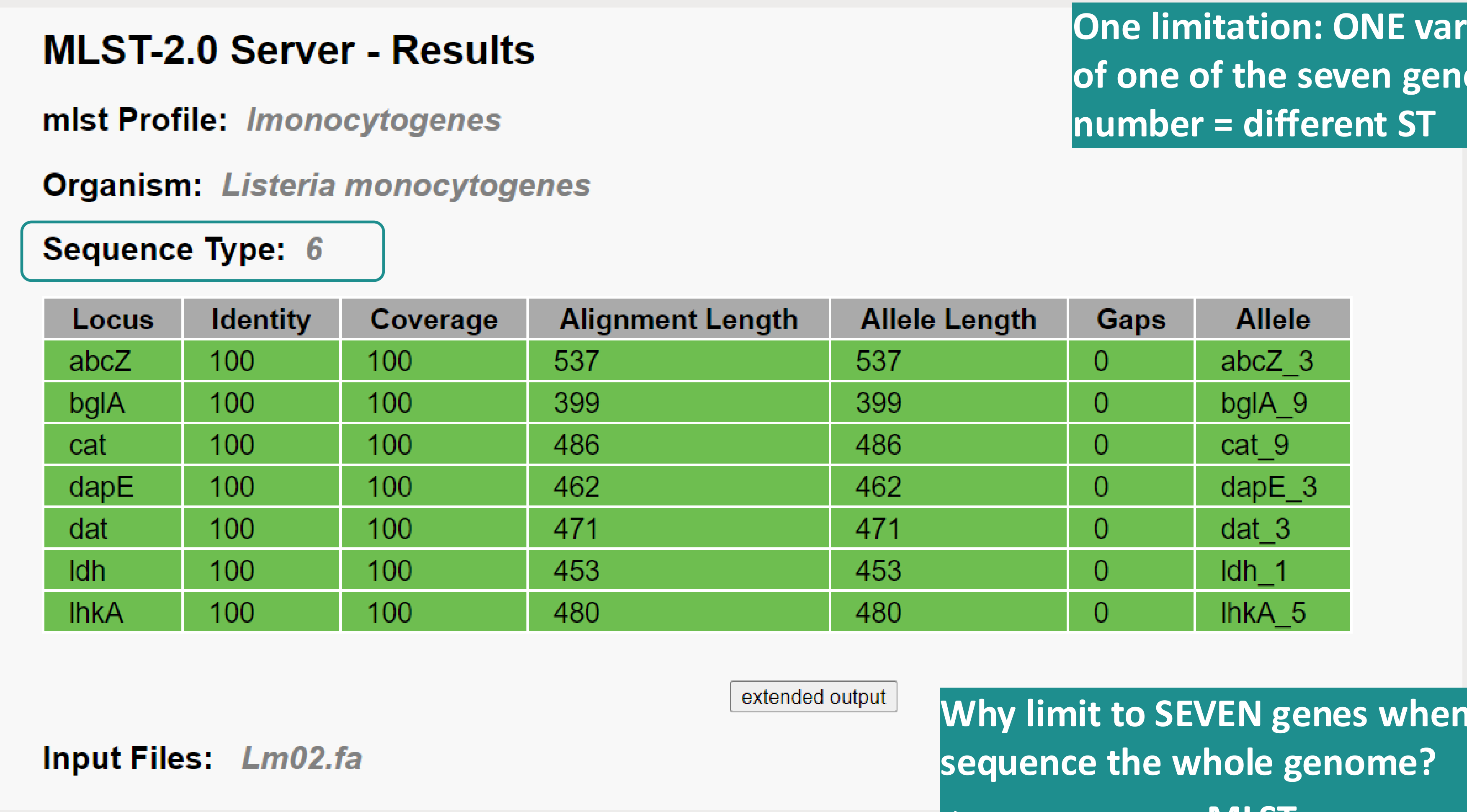
7 x PCR and sequencing vs. 1 x WGS

The Fleming Fund
Regional Grants

MLST Typing by WGS



MLST result output



MLST-2.0 Server - Results

mlst Profile: *lmonocytogenes*

Organism: *Listeria monocytogenes*

Sequence Type: *6*

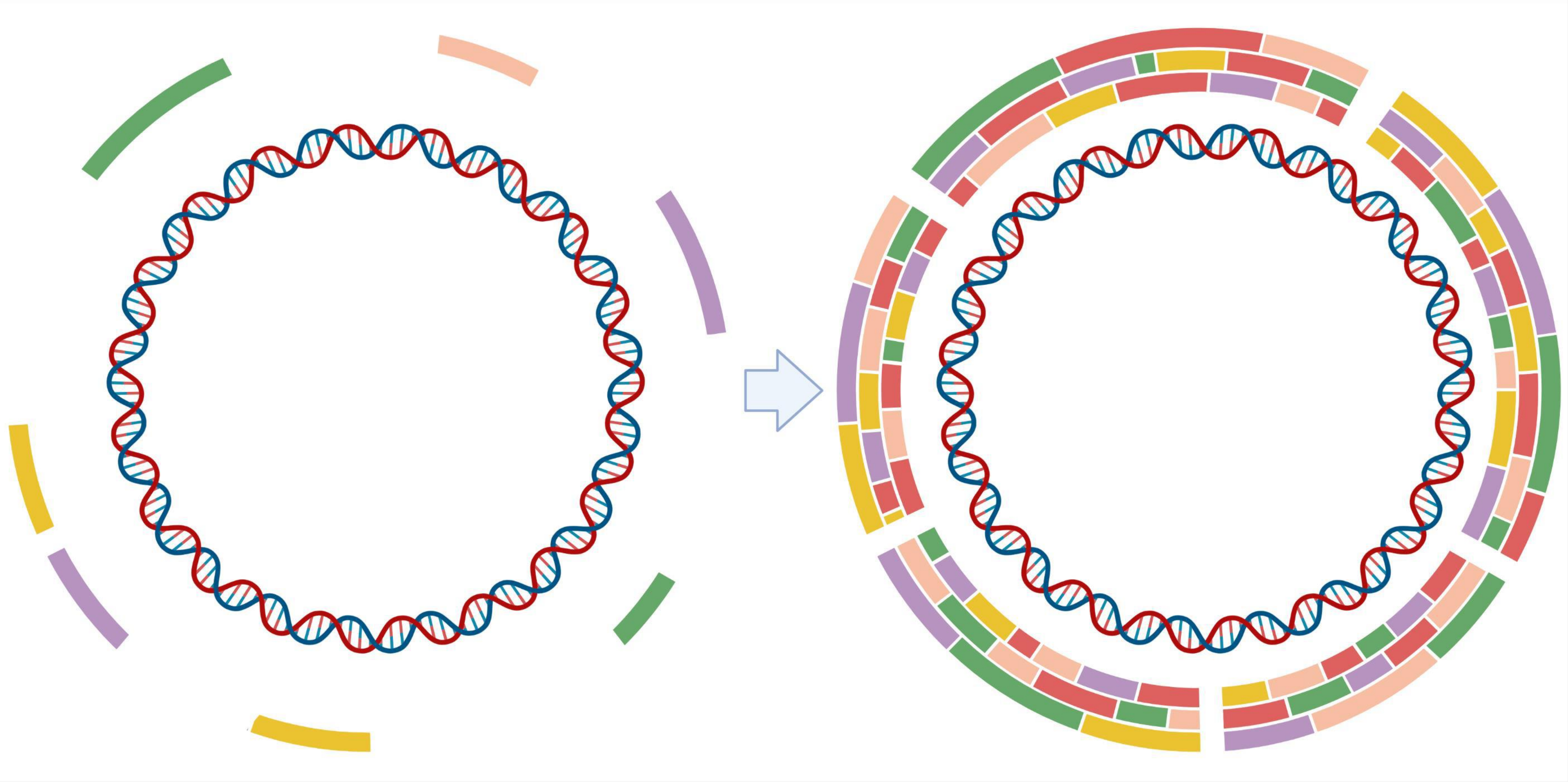
Locus	Identity	Coverage	Alignment Length	Allele Length	Gaps	Allele
abcZ	100	100	537	537	0	abcZ_3
bglA	100	100	399	399	0	bglA_9
cat	100	100	486	486	0	cat_9
dapE	100	100	462	462	0	dapE_3
dat	100	100	471	471	0	dat_3
ldh	100	100	453	453	0	ldh_1
lhkA	100	100	480	480	0	lhkA_5

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?

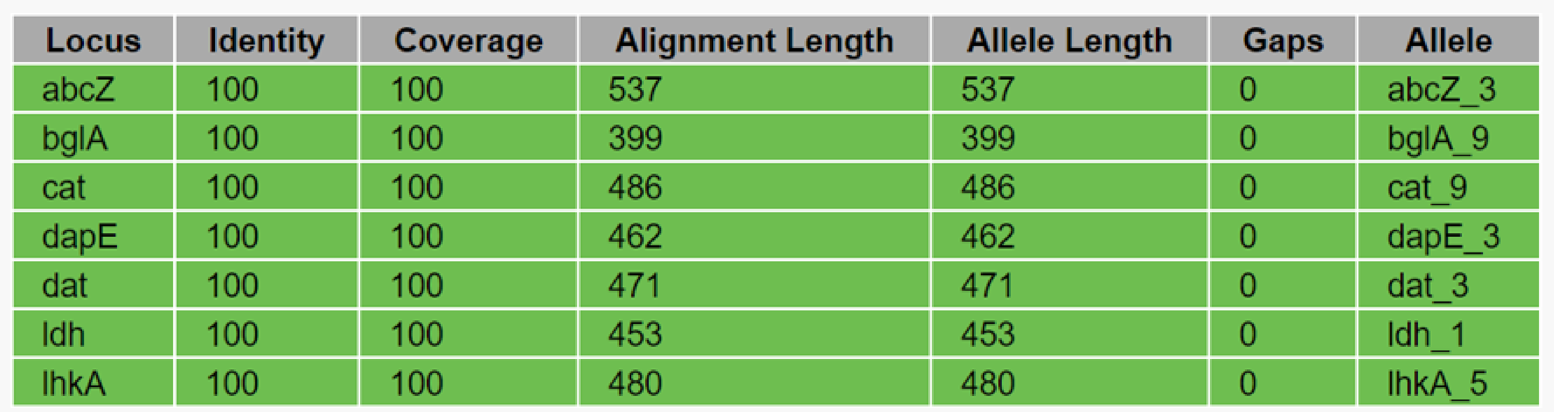
-> core genome MLST



Created in BioRender.com

cgMLST – core genome

- Core genome = genes common for (almost) all 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						

Each gene variant has an allele number

Each allele combination has a **cg ST** assigend 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

Whole genome based phylogeny

- 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

Sequence-based typing

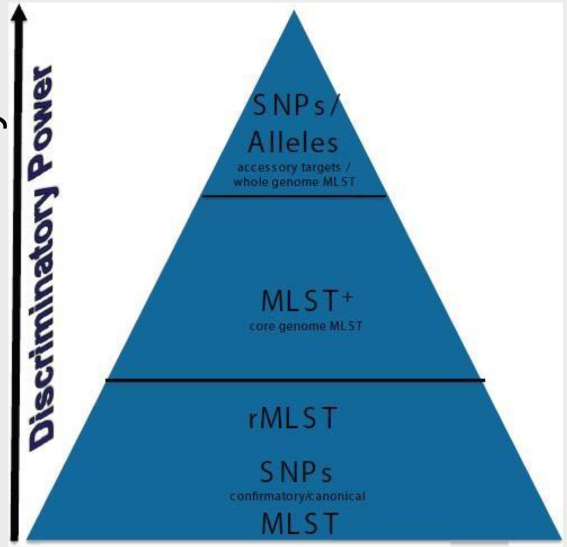
- MLST

- cgMLST / SNP (Core/Whole Genome Comparison)

- Presence/absence of genes and mobile elements

…..often a combination of the above is used to study outbreaks.

Fraction of the genome used for the analysis



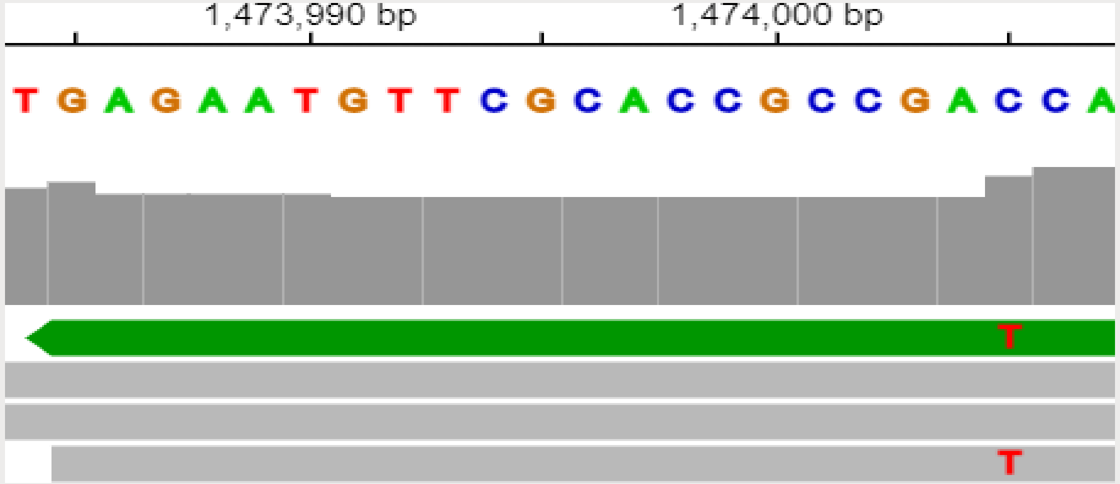
Single nucleotide polymorphism (SNP)

- A SNP is a mutation within a subpopulations of individuals, essentially it is a point mutation which distinguishes two “closely” related strains of the same species

- To separate sequencing error from true SNPs, we need to have:
  - Proper sequencing depth at the position
  - High Q-score

- When we know the amounts of SNP differences we can infer the phylogenic relationship between strains

- High resolution



Section of reads mapped to reference, visualized using integrative genomics viewer, [IGV: Integrative Genomics Viewer]

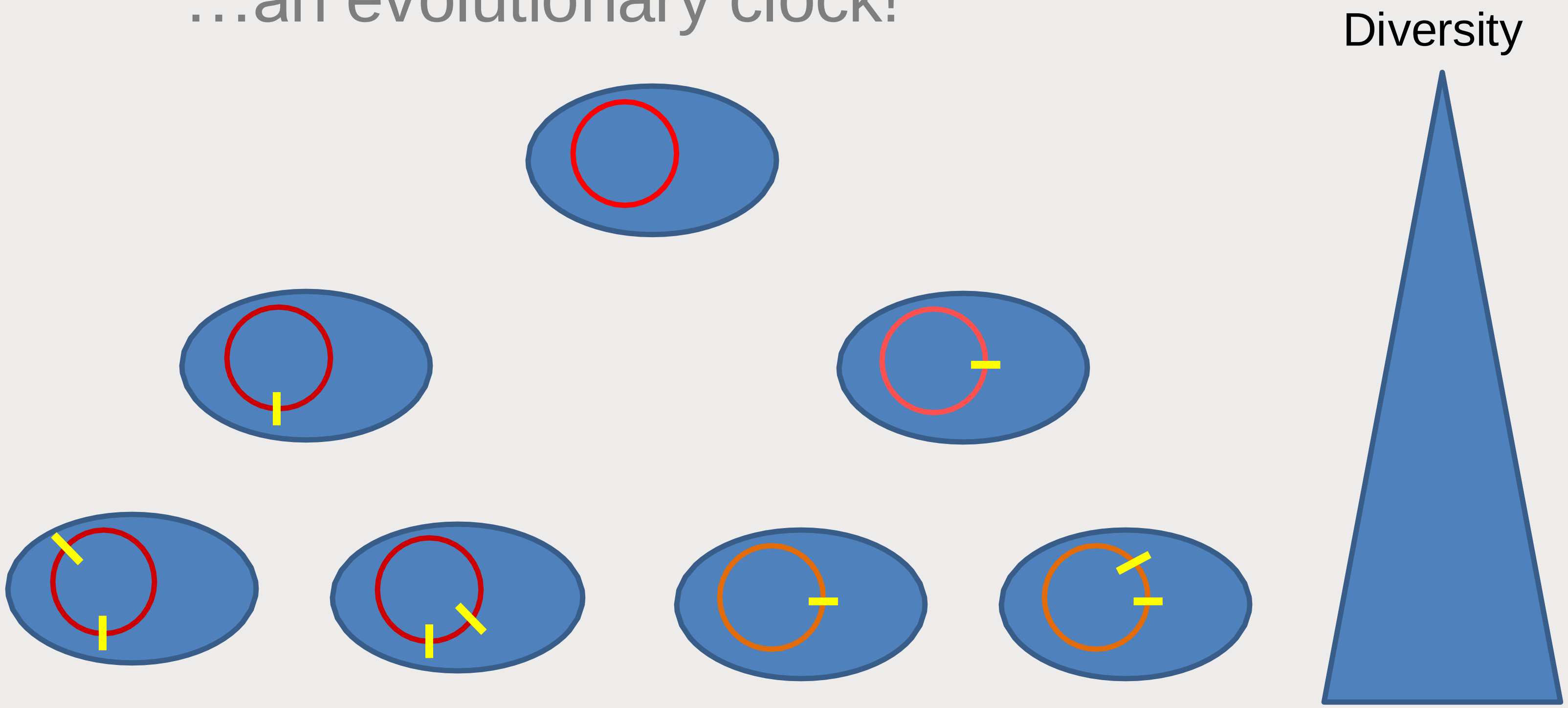
Clone theory 101

- Textbook – A clone is:
  - ”a group of genotypic identical isolates descending from a common ancestor as part of a direct chain of replication”

- A more realistic definition:
  - ”the word clone will be used to denote bacterial cultures isolated independently from different sources, in different locations, and perhaps at different times, but showing so many identical phenotypic and genotypic traits that the most likely explanation for this identity is a common origin”
- (*Ørskov & Ørskov, 1983*)

The Chromosomograph

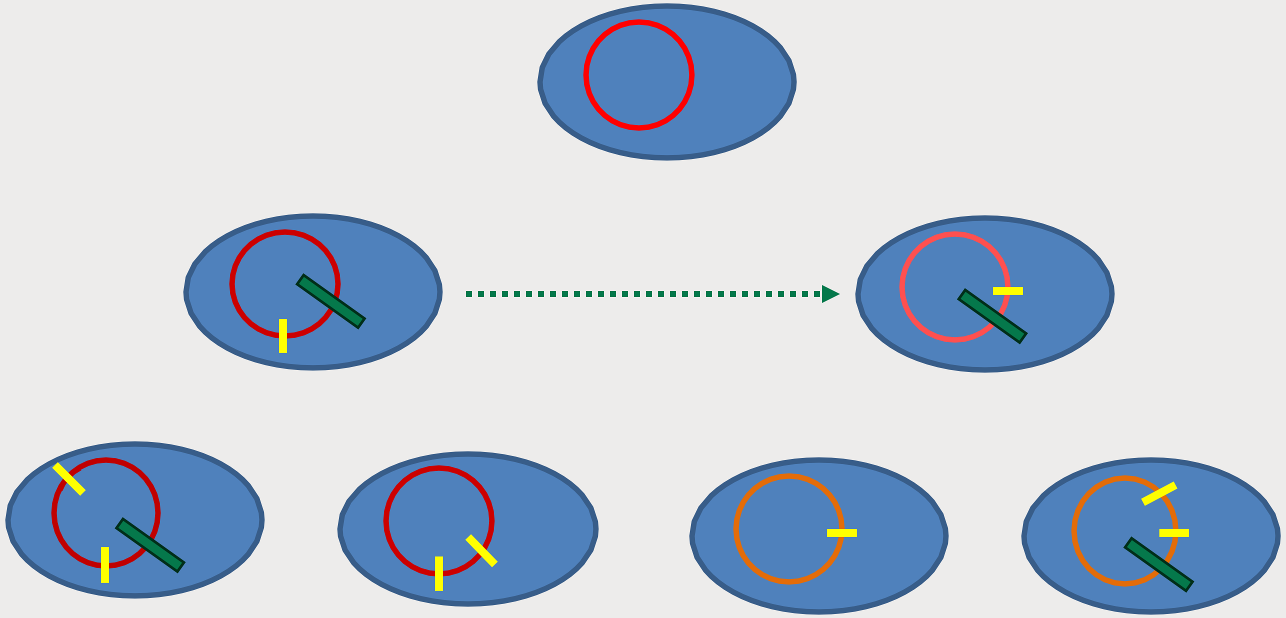
…an evolutionary clock!



- Randomly generated across the chromosome over time (“The mutation rate”)
- …but influenced by external factors…

Horizontal gene transfer

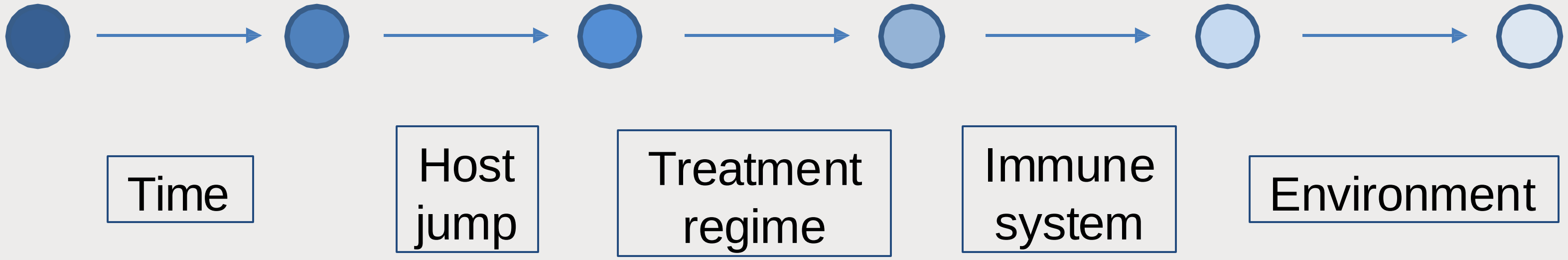
The Chromosomograph's evil nemesis



- Horizontal gene transfer circumvents the linearity of the evolutionary clock
- …and needs to be addressed in any whole genome analysis such as SNPs…

Advanced clone theory

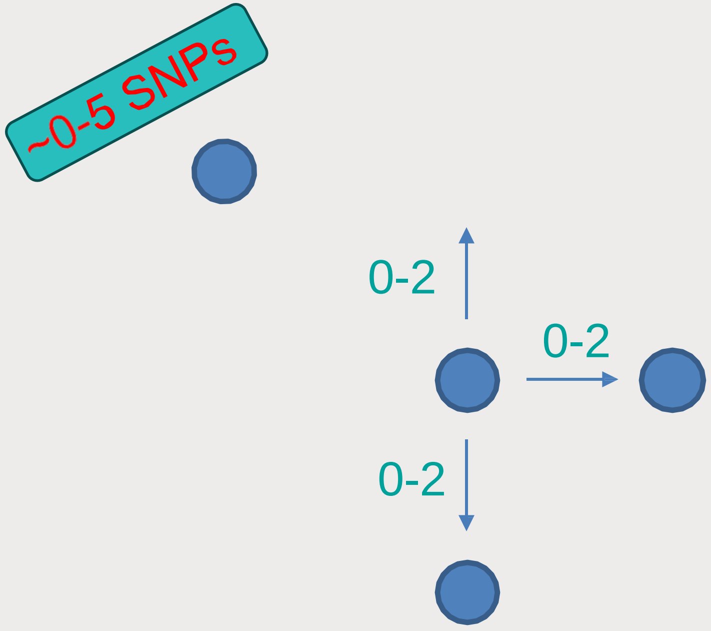
Clonal drift



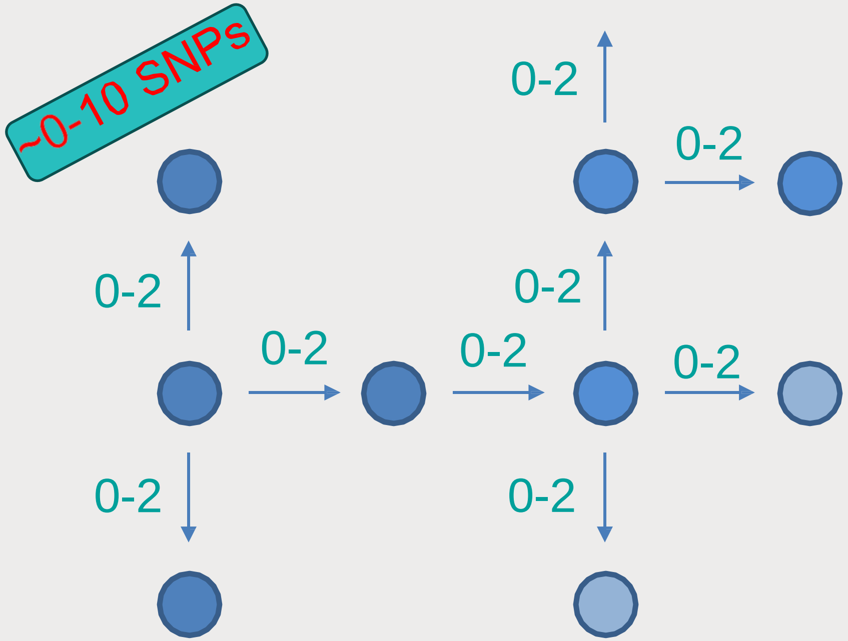
- The more discriminatory a typing method is, the more difficult it will be for it to accommodate *biological variation* caused by clonal drift over time (stability issues).

- On top of this, all typing methods will add *methodological variation* (repeatability and reproducibility issues) thus blurring the picture even more.

Single source outbreaks

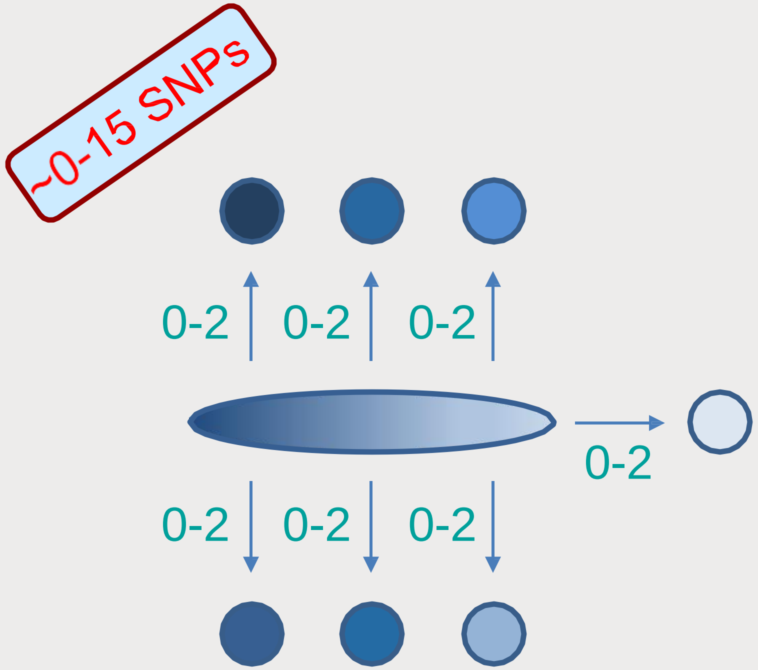


Single source Short time span
“Contaminated dish”
“Single infected patient”



Single source – local spread
Long time span
“Hospital or regional outbreak”

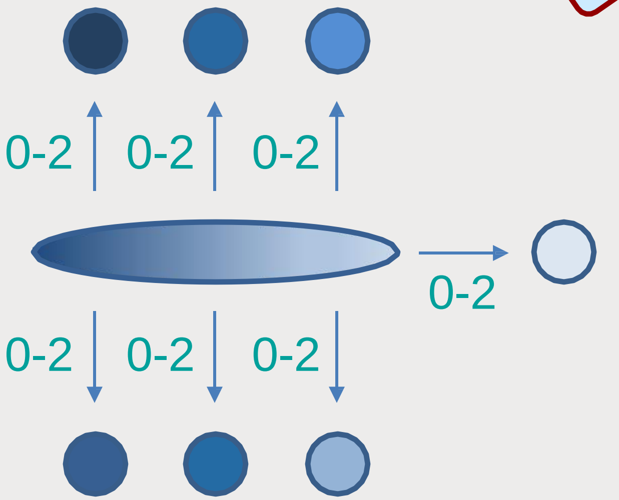
Complicated outbreaks



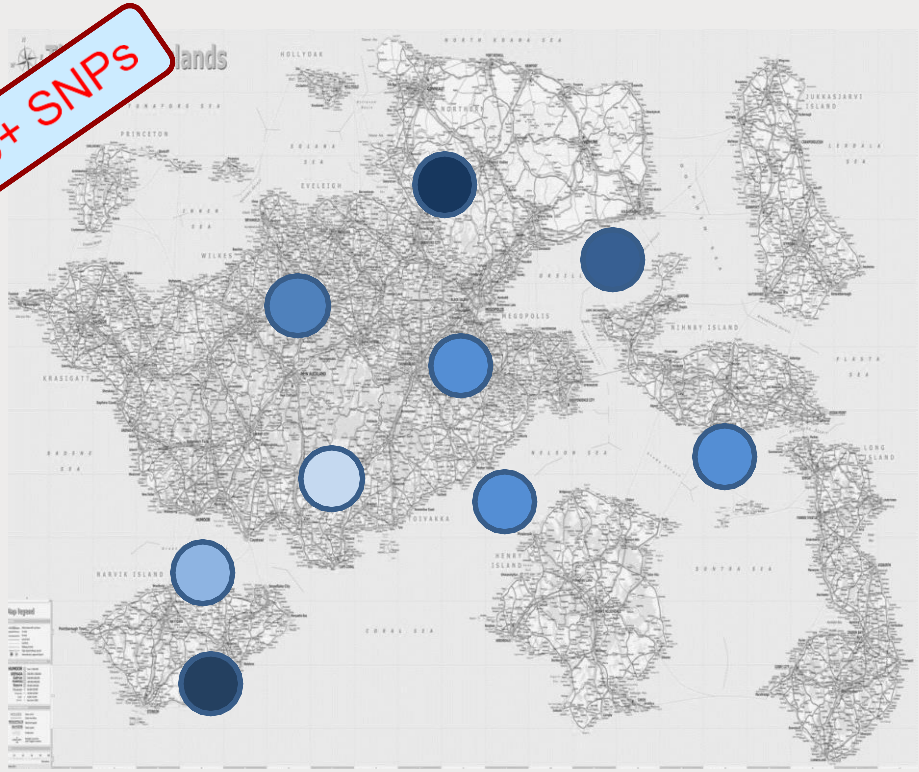
Single source
Long time span
"Contaminated processing plant / industry"
"Long-term colonized patient / healthcare worker"

Complicated outbreaks

~0-15 SNPs



~10-50+ SNPs

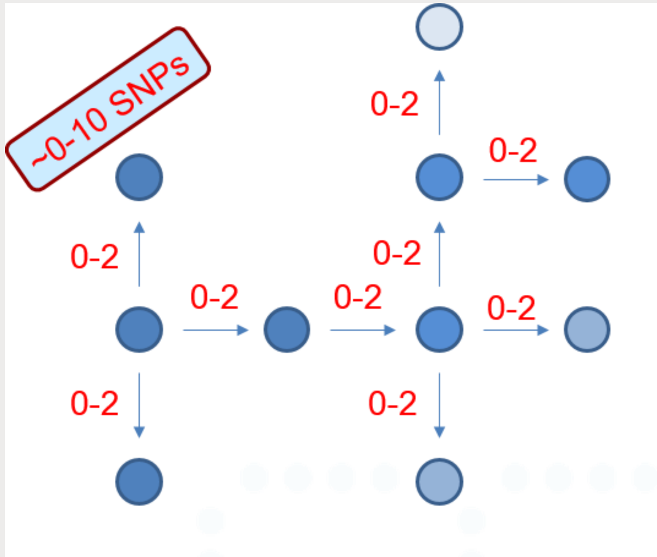


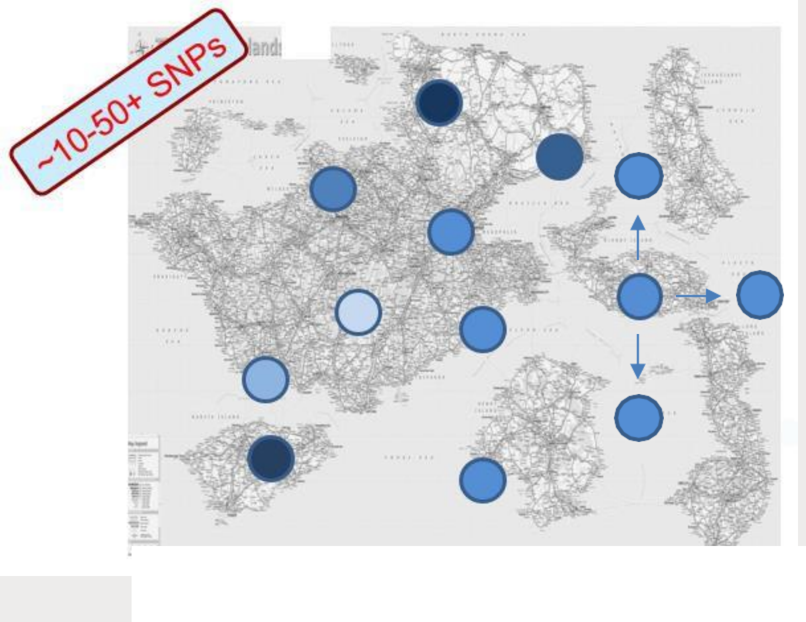
Single source
Long time span
“Contaminated processing plant / industry”
“Long-term colonized patient / healthcare worker”

International s
Long tim n
“Importe source”
“Trav ted outbreak”

International clones

PO = Possible outbreaks(E. COLI)



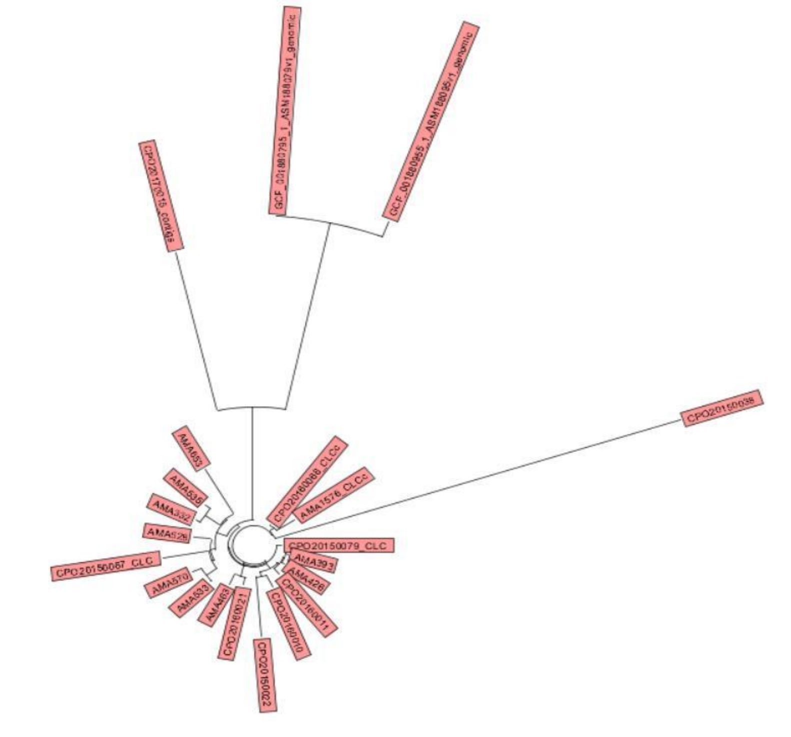


Tentative definition of possible outbreak (PO)

If two isolates have a SNP distance ≤ 10 (termed PO_{10}), they are considered to be so genetically related that they may be part of the same outbreak.

Phylogenetic analysis

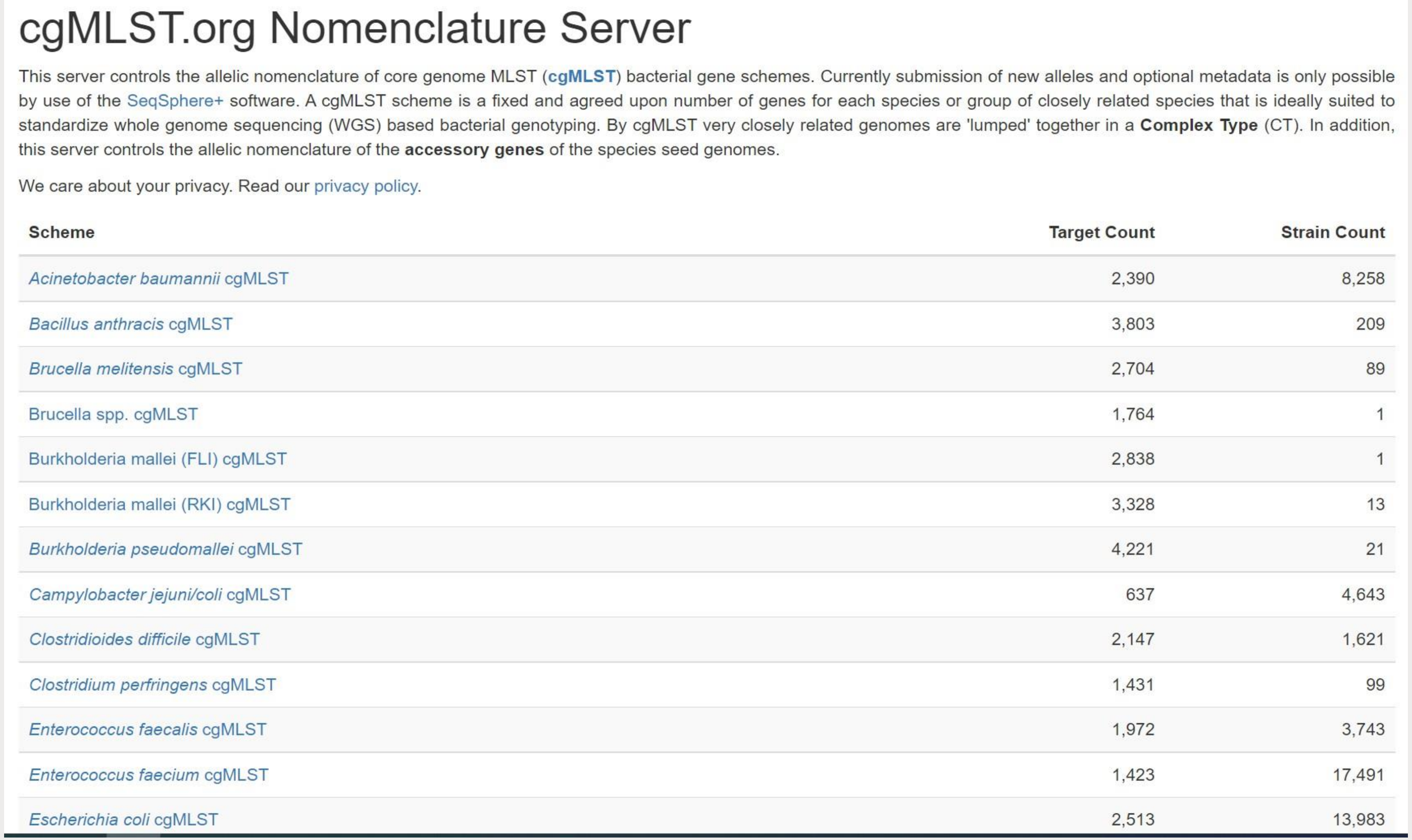
Core genome MLST (cgMLST) vs
Single Nucleotide Polymorphism (SNP)



Core genome MLST (cgMLST)

- Reference based gene-by-gene comparison
- “Super MLST”
- Increased number of genes → Increased discriminatory power requires curated and validated schemes
- Requires software to remove gene homologues if you want to build your own scheme.

cgMLST.org



cgMLST.org Nomenclature Server

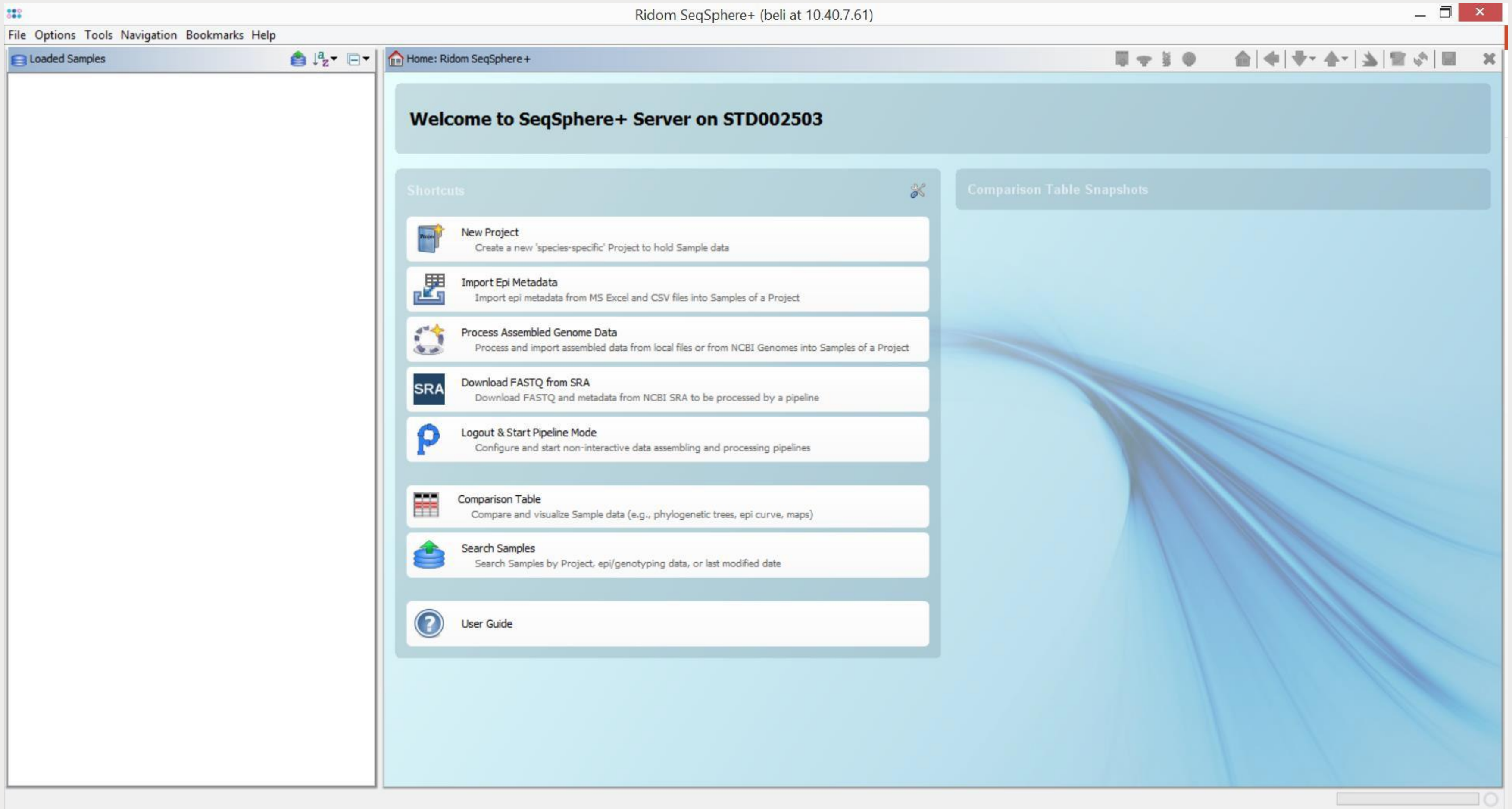
This server controls the allelic nomenclature of core genome MLST (**cgMLST**) bacterial gene schemes. Currently submission of new alleles and optional metadata is only possible by use of the SeqSphere+ software. A cgMLST scheme is a fixed and agreed upon number of genes for each species or group of closely related species that is ideally suited to standardize whole genome sequencing (WGS) based bacterial genotyping. By cgMLST very closely related genomes are 'lumped' together in a **Complex Type** (CT). In addition, this server controls the allelic nomenclature of the **accessory genes** of the species seed genomes.

We care about your privacy. Read our privacy policy.

Scheme	Target Count	Strain Count
Acinetobacter baumannii cgMLST	2,390	8,258
Bacillus anthracis cgMLST	3,803	209
Brucella melitensis cgMLST	2,704	89
Brucella spp. cgMLST	1,764	1
Burkholderia mallei (FLI) cgMLST	2,838	1
Burkholderia mallei (RKI) cgMLST	3,328	13
Burkholderia pseudomallei cgMLST	4,221	21
Campylobacter jejuni/coli cgMLST	637	4,643
Clostridioides difficile cgMLST	2,147	1,621
Clostridium perfringens cgMLST	1,431	99
Enterococcus faecalis cgMLST	1,972	3,743
Enterococcus faecium cgMLST	1,423	17,491
Escherichia coli cgMLST	2,513	13,983

The Fleming Fund
Regional Grants

SeqSphere+ Software


Ridom SeqSphere+ (beli at 10.40.7.61)
File Options Tools Navigation Bookmarks Help
Loaded Samples
Home: Ridom SeqSphere+
Shortcuts
Comparison Table Snapshots

Welcome to SeqSphere+ Server on STD002503

New Project
Create a new 'species-specific' Project to hold Sample data

Import Epi Metadata
Import epi metadata from MS Excel and CSV files into Samples of a Project

Process Assembled Genome Data
Process and import assembled data from local files or from NCBI Genomes into Samples of a Project

Download FASTQ from SRA
Download FASTQ and metadata from NCBI SRA to be processed by a pipeline

Logout & Start Pipeline Mode
Configure and start non-interactive data assembling and processing pipelines

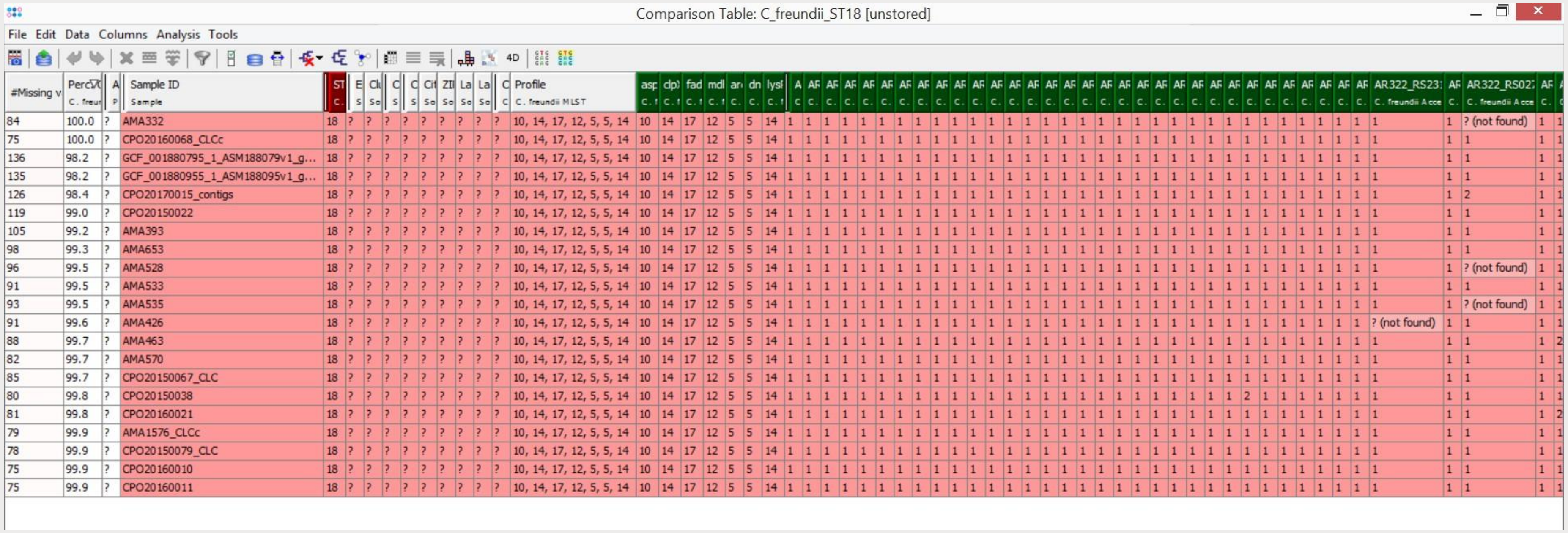
Comparison Table
Compare and visualize Sample data (e.g., phylogenetic trees, epi curve, maps)

Search Samples
Search Samples by Project, epi/genotyping data, or last modified date

User Guide

Available schemes: *S. aureus* – *E. coli* – *E. faecium* – *A. baumannii* – *K. pneumoniae* ... and more

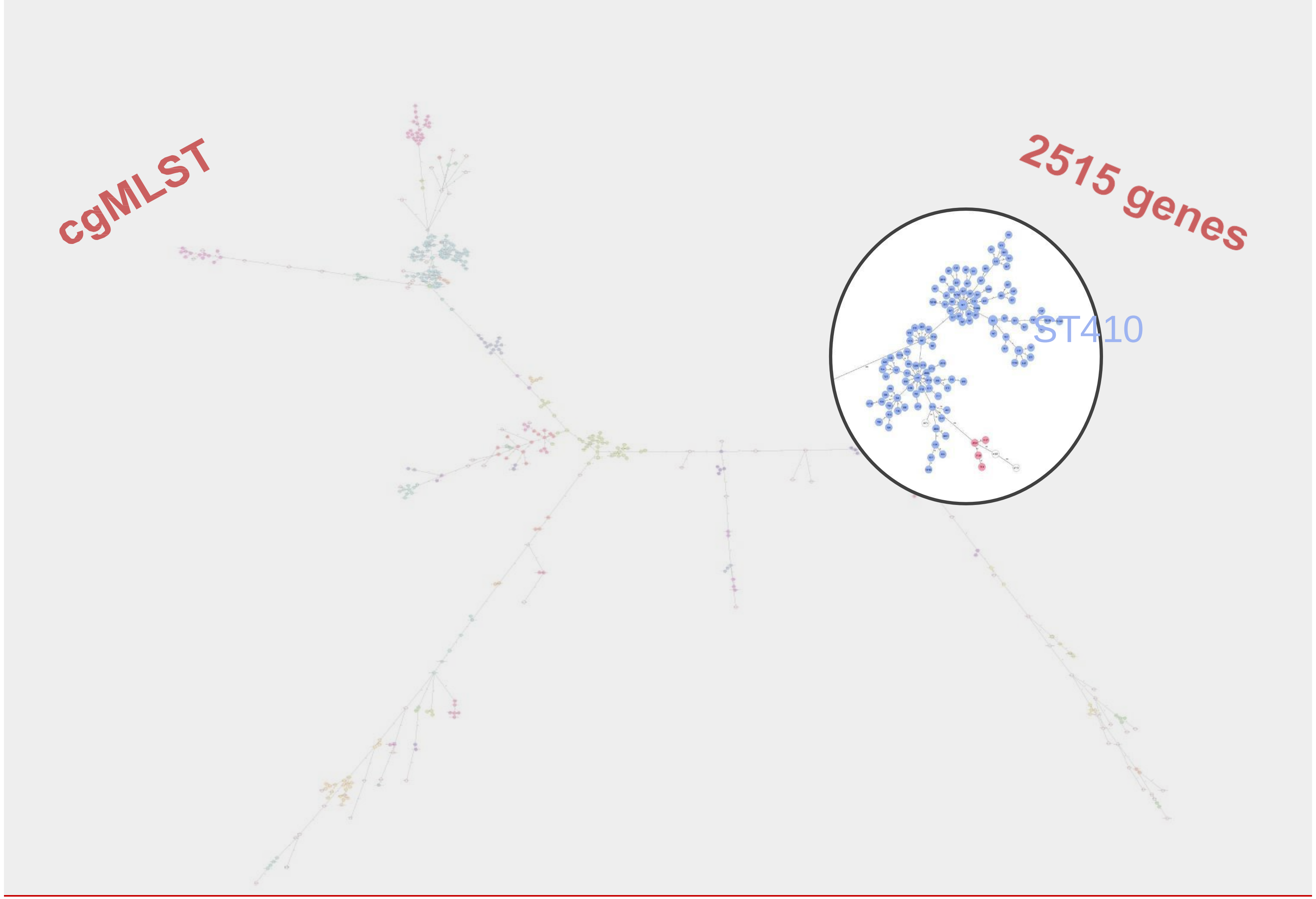
Core genome MLST (cgMLST)



Comparison Table: C_freundii_ST18 [unstored]

#Missing v	Perc▽(	A	Sample ID	ST	E	Clu	C	C	Cit	ZII	La	La	C	Profile	asp	clp	fad	mdl	ar	dn	lysP	A	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF	AR322_RS23:	AF	AR322_RS02:	AF	A
	C. freu	P	Sample	C.	S	So	S	S	So	So	So	So	C	C. freundii MLST	C. f	C. f	C. f	C. f	C.	C.	C. f	C	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C.	C. freundii A cce	C.	C. freundii A cce	C.	C
84	100.0	?	AMA332	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	? (not found)	1	1
75	100.0	?	CPO20160068_CLCc	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
136	98.2	?	GCF_001880795_1_ASM188079v1_g...	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
135	98.2	?	GCF_001880955_1_ASM188095v1_g...	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
126	98.4	?	CPO20170015_contigs	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1
119	99.0	?	CPO20150022	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
105	99.2	?	AMA393	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
98	99.3	?	AMA653	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
96	99.5	?	AMA528	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	? (not found)	1	1
91	99.5	?	AMA533	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
93	99.5	?	AMA535	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	? (not found)	1	1
91	99.6	?	AMA426	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	? (not found)	1	1	1	1
88	99.7	?	AMA463	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
82	99.7	?	AMA570	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
85	99.7	?	CPO20150067_CLC	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
80	99.8	?	CPO20150038	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1
81	99.8	?	CPO20160021	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
79	99.9	?	AMA1576_CLCc	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
78	99.9	?	CPO20150079_CLC	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
75	99.9	?	CPO20160010	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
75	99.9	?	CPO20160011	18	?	?	?	?	?	?	?	?	?	10, 14, 17, 12, 5, 5, 14	10	14	17	12	5	5	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

- All isolates are assigned to specific Complex Types (CTs)
- Different cgMLST schemes use different cut-off values for new CTs

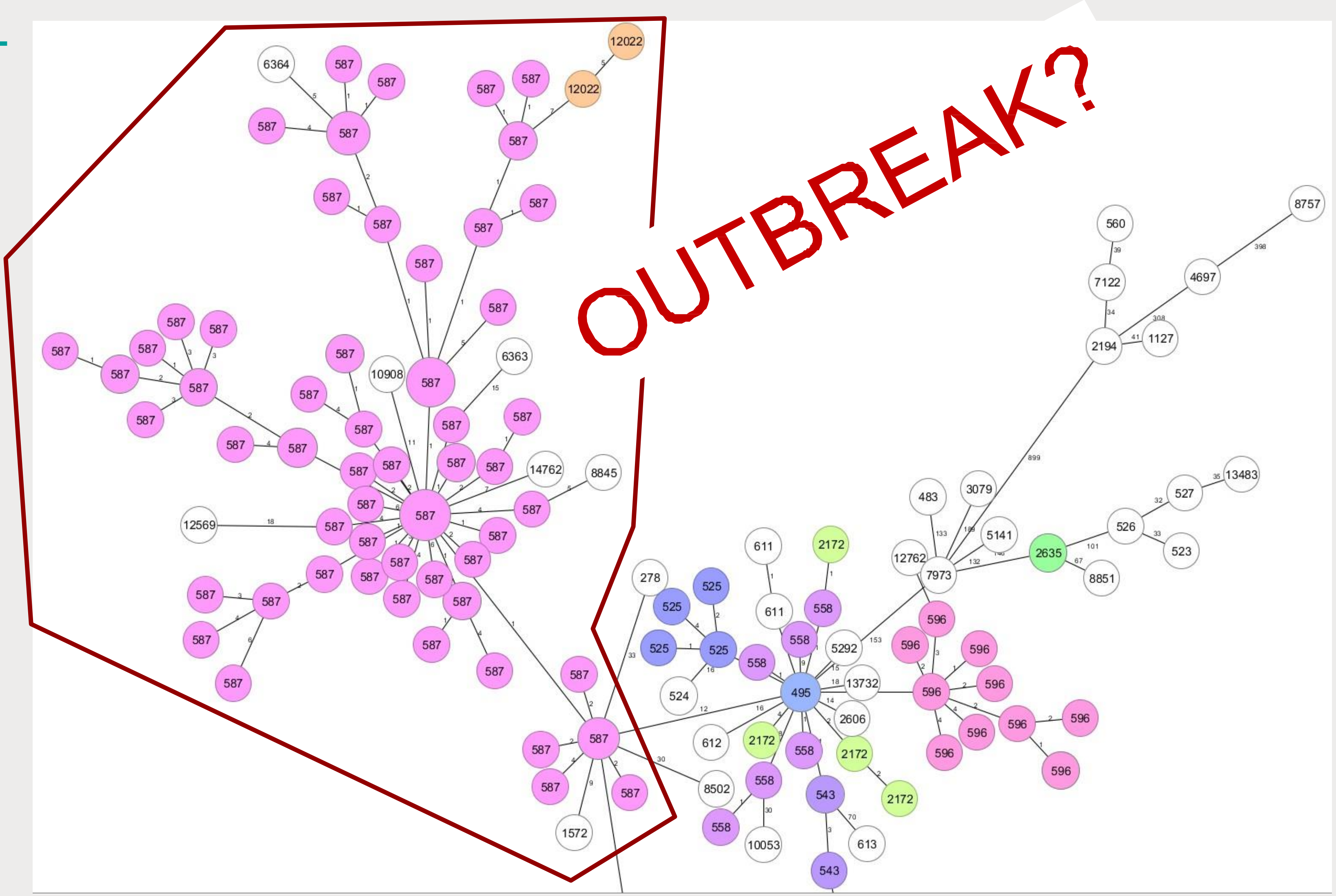


Ridom – SeqSphere+

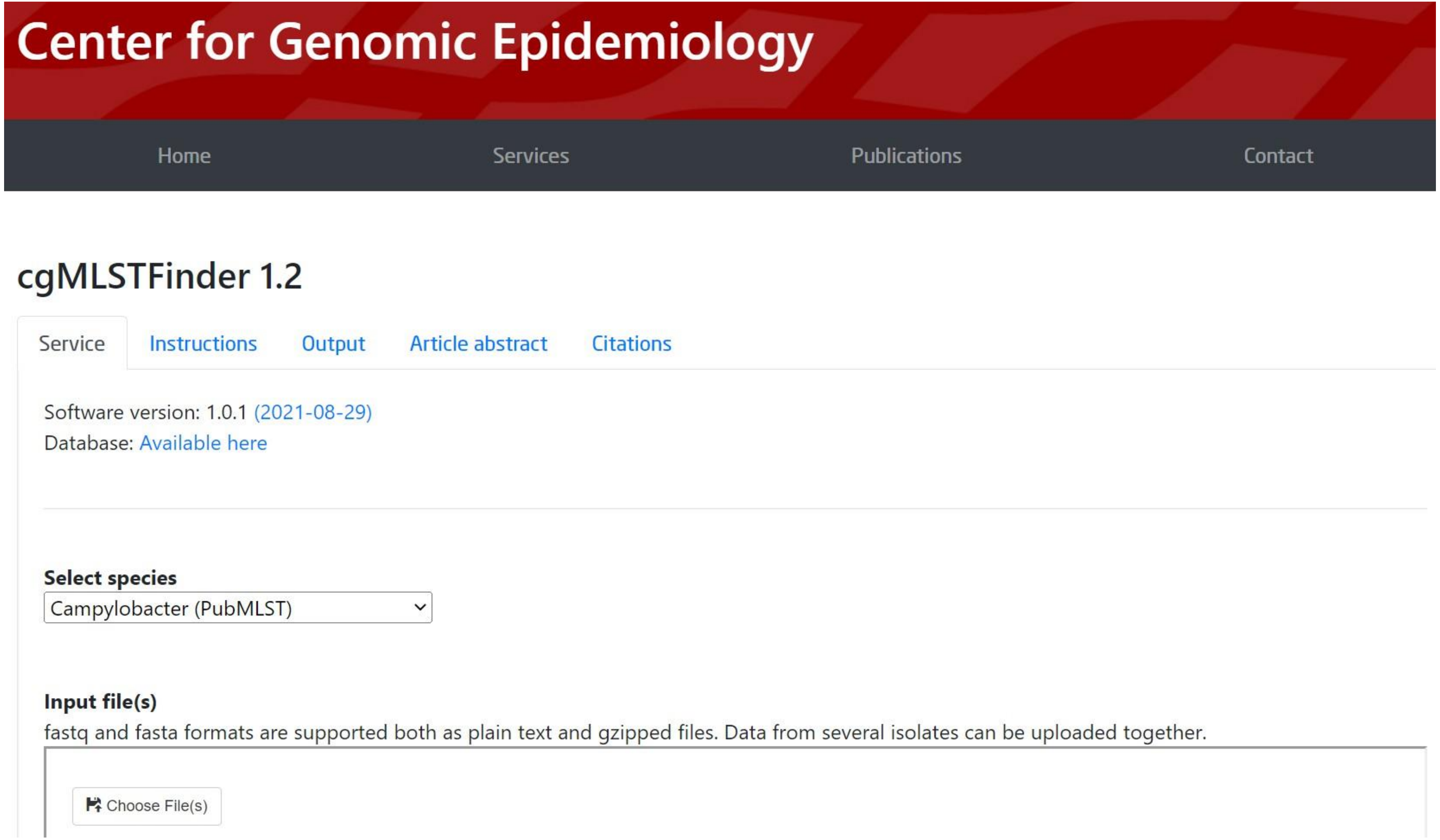
The Fleming Fund
Regional Grants

CPO in Denmark – *E. coli* ST410

cgMLST (2505 genes)



Core genome MLST (cgMLST)



cgMLSTFinder 1.2

Software version: 1.0.1 (2021-08-29)

Database: Available here

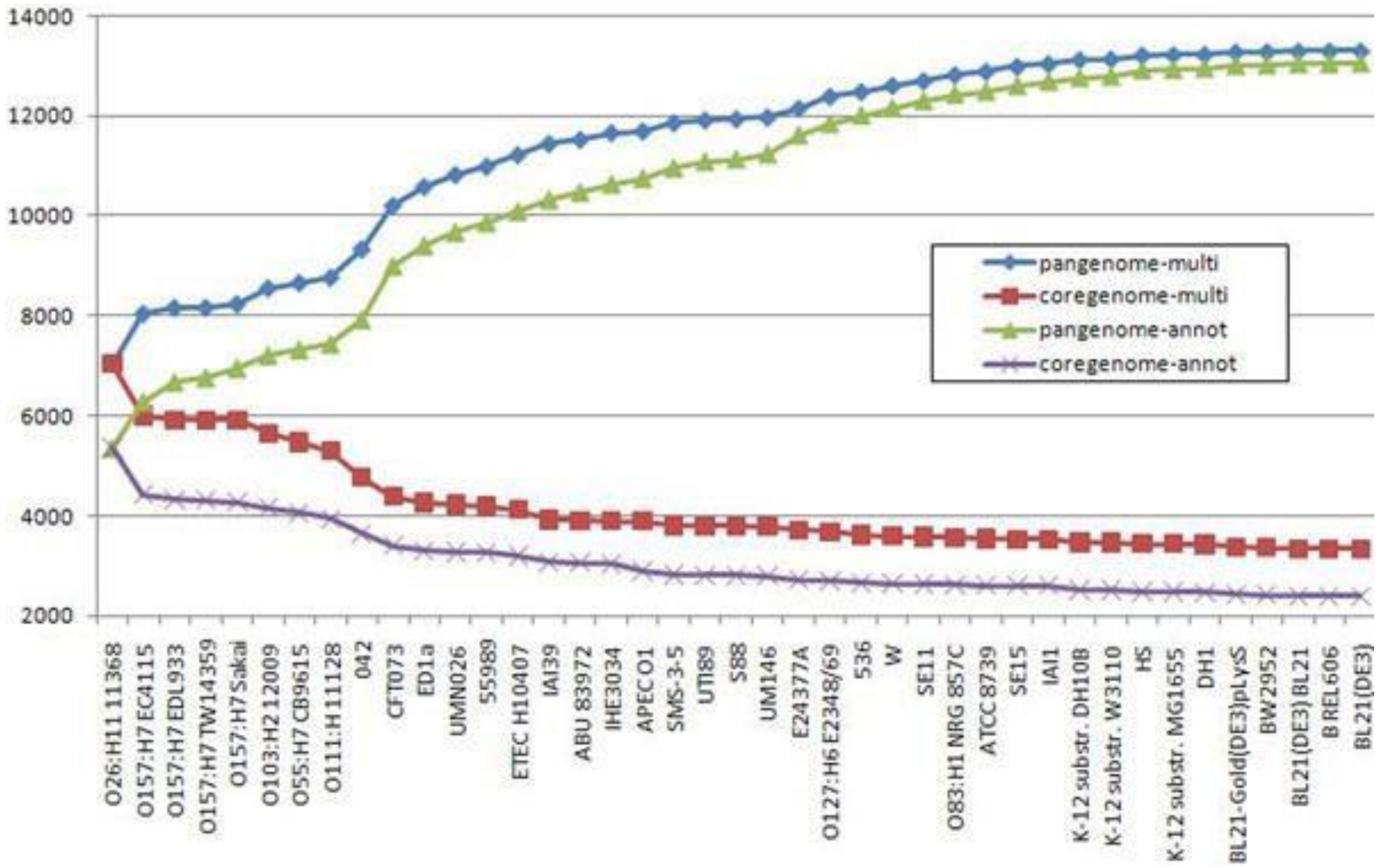
Select species

Campylobacter (PubMLST)

Input file(s)

fastq and fasta formats are supported both as plain text and gzipped files. Data from several isolates can be uploaded together.

Core Genome MLST (cgMLST)



Core Genome MLST (cgMLST)

Main advantages

- Common nomenclature (Cluster types)
- Fixed set of reference genes
- Recombination has been filtered out
- Curated database
- Fast, as it runs on draft assemblies

Main disadvantages

- Requires a validated cgMLST scheme
- May be sensitive to assembly method
- Requires a curator to manage the database
- The discriminatory power may be a bit lower than for SNP analysis
- Have a tendency to drift over time – especially in long-lasting outbreaks

The Fleming Fund
Regional Grants

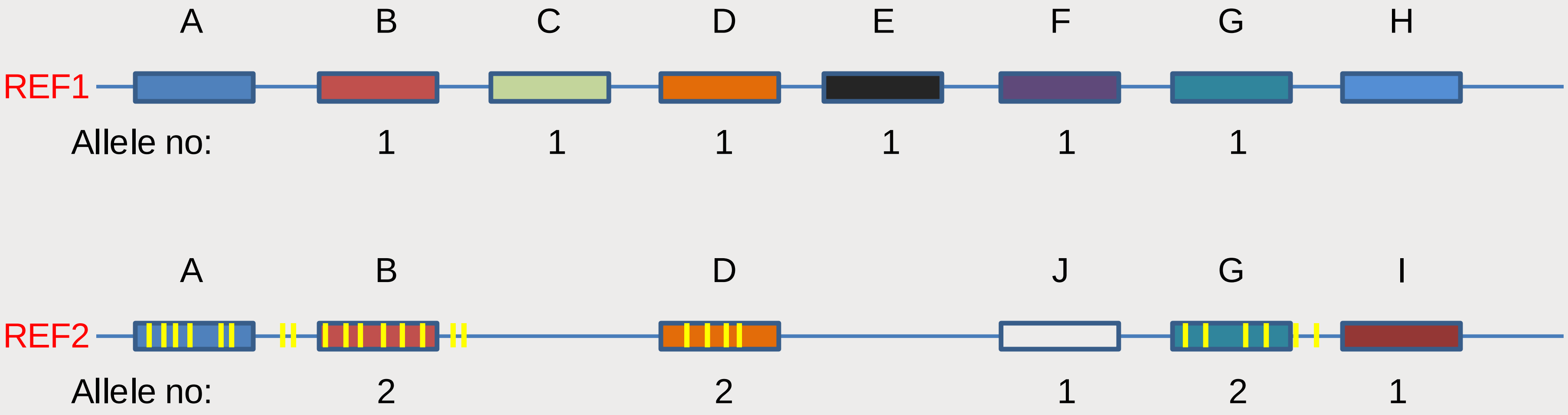
SNP analysis

practical considerations

- Choosing the best reference
- Global SNP vs HQ SNP analysis
- Detecting contamination
- Recombination events

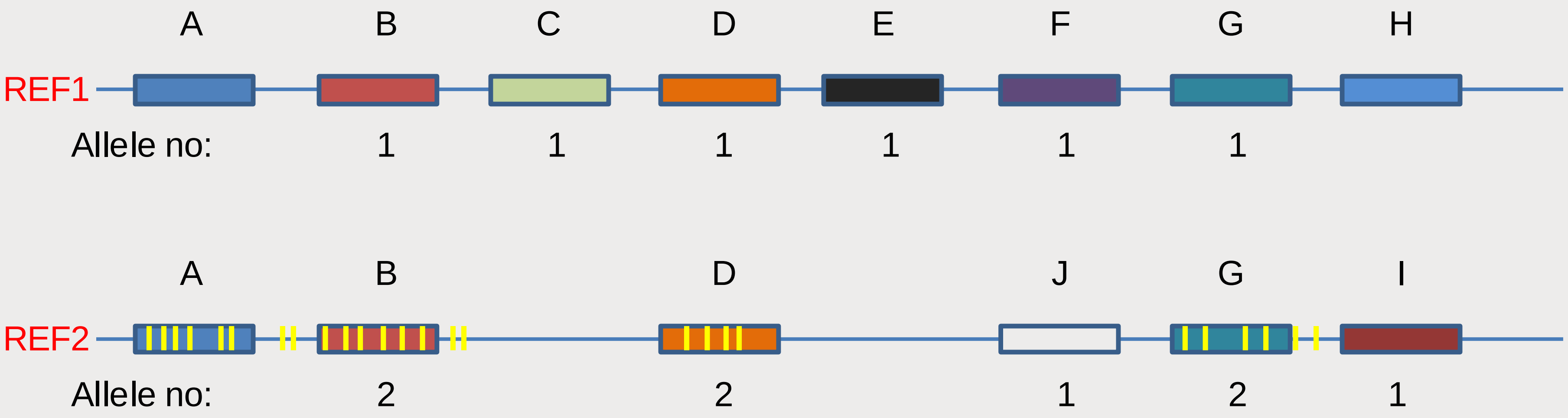
The Fleming Fund
Regional Grants

Choosing the best reference



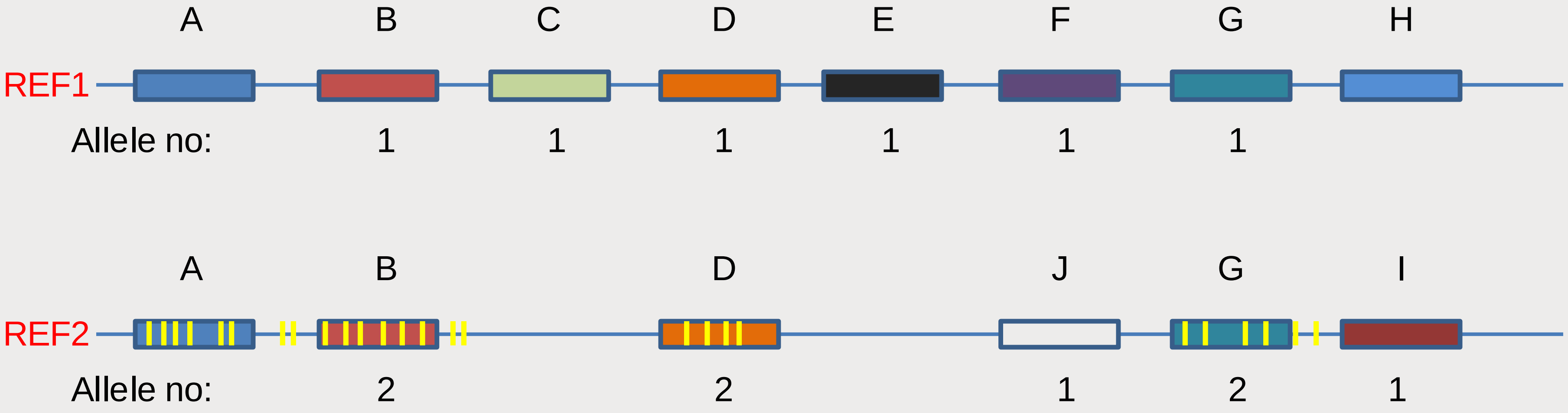
- In general, a closely related reference is desired.
- A best match in NCBI RefSeq can be searched using KmerFinder.
- Complete genomes can also be searched at NCBI (but is not easy to use).
- A draft genome of the index isolate can be considered for use.
- Or you can make your own complete genome by using MinION or PacBIO.

Choosing the best reference



- In general, a closely related reference is desired.
- A best match in NCBI RefSeq can be searched using KmerFinder.
- Complete genomes can also be searched at NCBI (but is not easy to use).
- A draft genome of the index isolate can be considered for use.
- Or you can make your own complete genome by using MinION or PacBIO.

Choosing the best reference

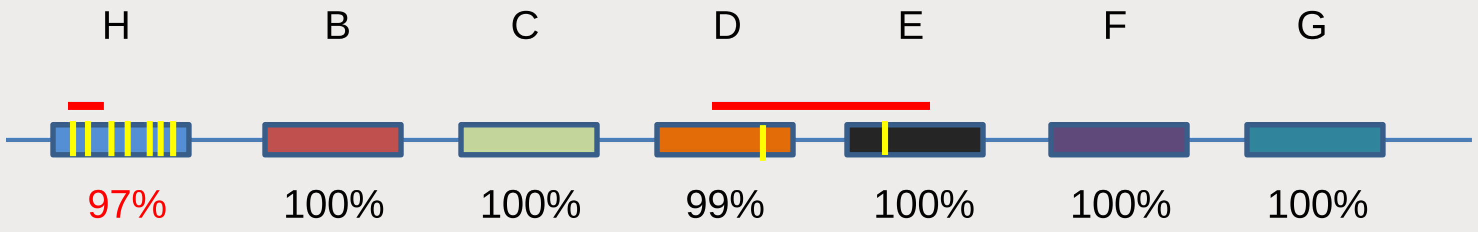


- In general, a closely related reference is desired.
- A best match in NCBI RefSeq can be searched using KmerFinder.
- Complete genomes can also be searched at NCBI (but is not easy to use).
- A draft genome of the index isolate can be considered for use.
- Or you can make your own complete genome by using MinION or PacBIO.

Recombination events

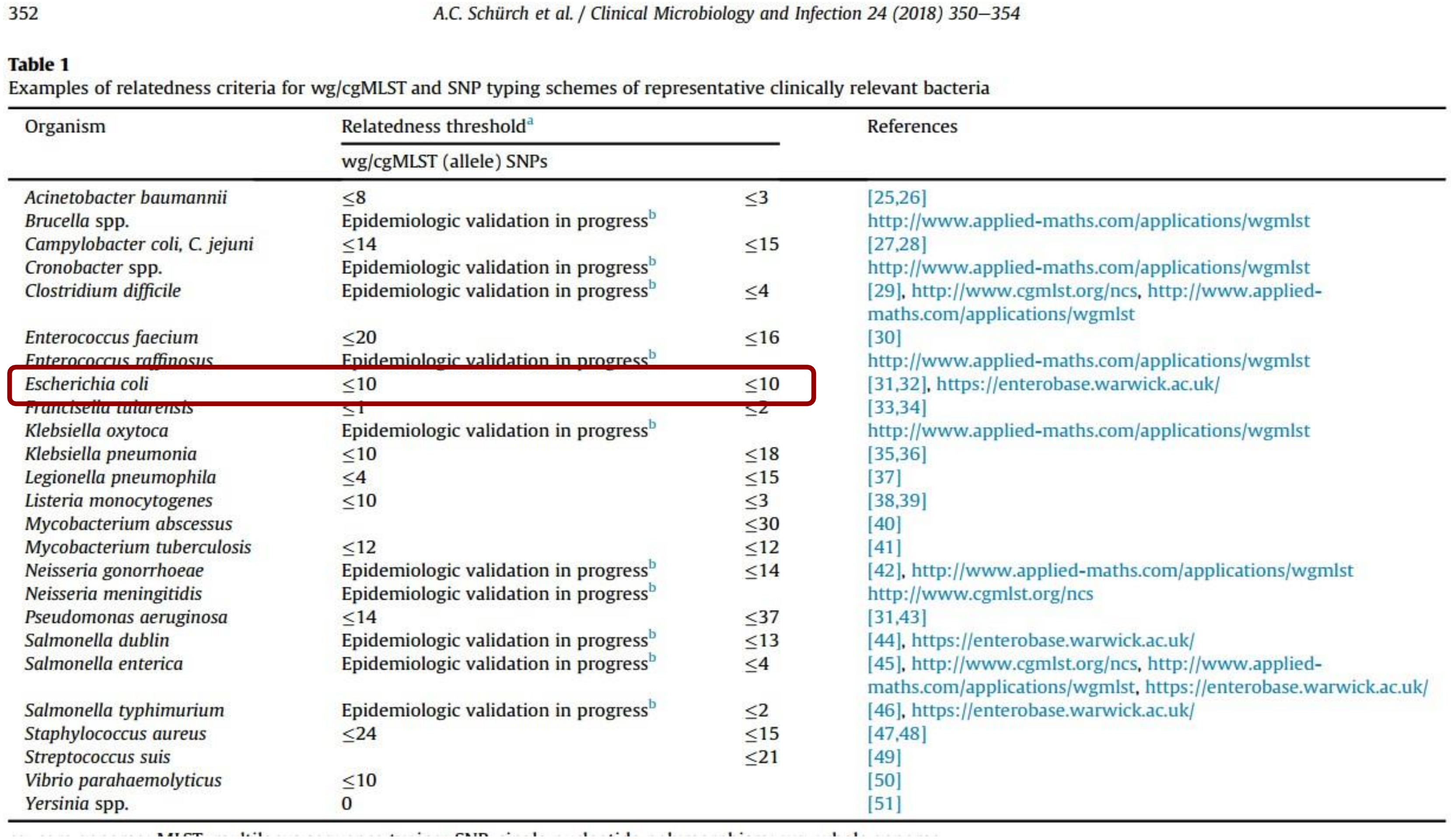
- Horizontal gene transfer
- Repetitive elements (IS-elements, AMR genes ect..)
- Gene duplication and diversification

Can to some extent be removed by using bioinformatic tools such as GUBBINS or by ignoring SNPs that are “close” to each other (called *pruning*).



No pruning	Kept	Kept	Kept
Prune 100	Ignored	Kept	Kept
Prune 1000	Ignored	Ignored	Ignored

What’s in a SNP?

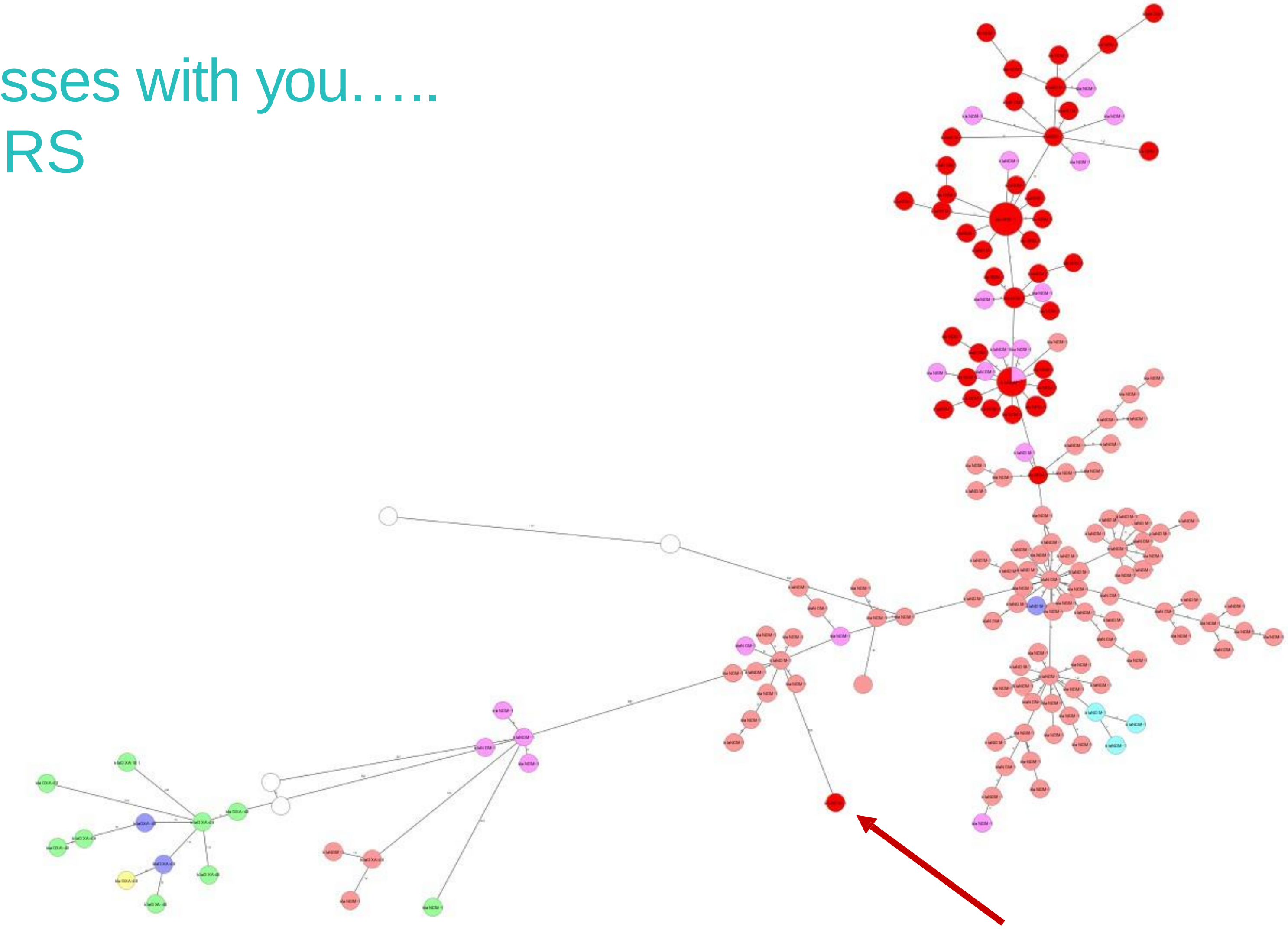


352 A.C. Schürch et al. / Clinical Microbiology and Infection 24 (2018) 350–354

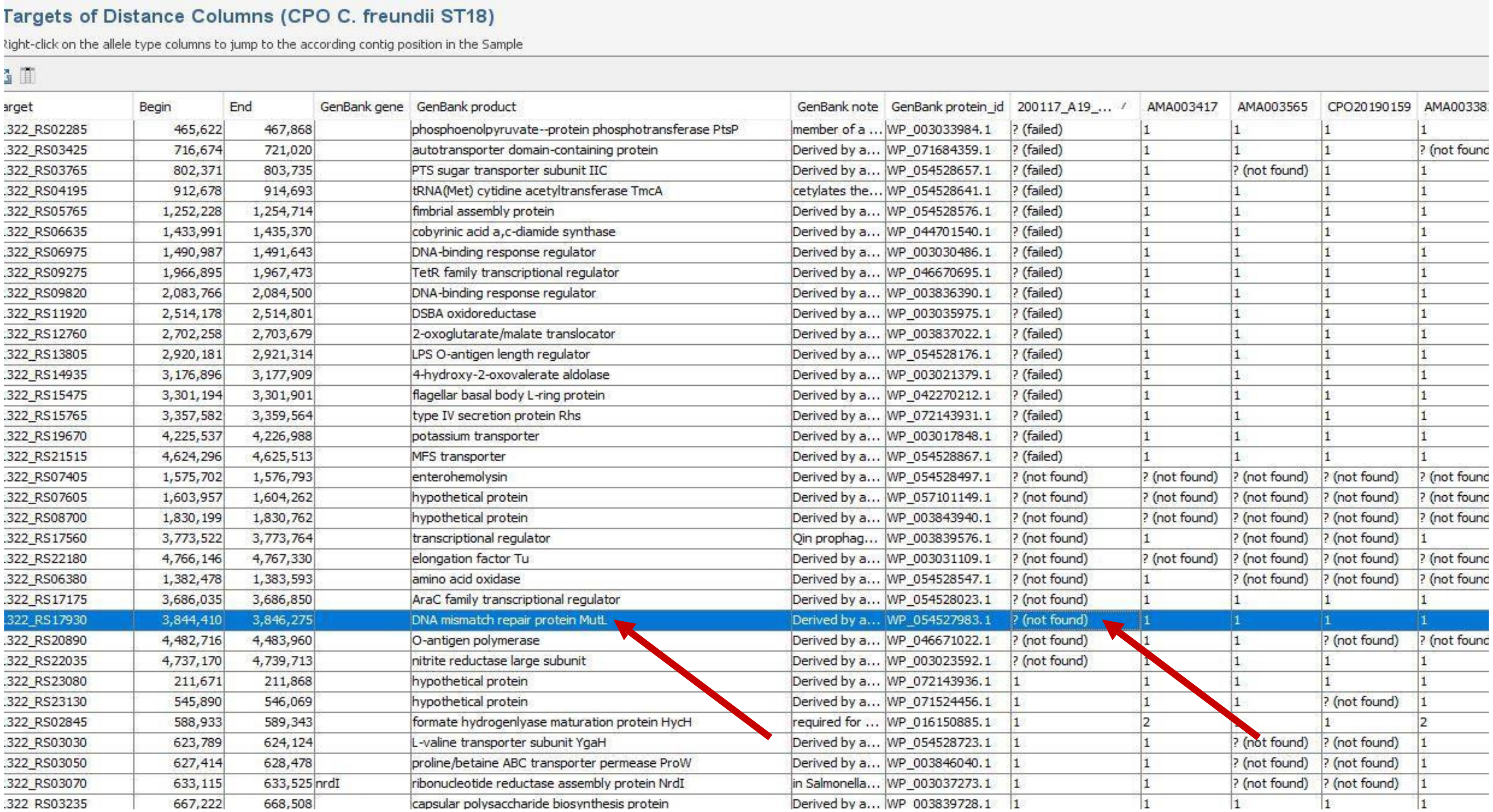
Table 1
Examples of relatedness criteria for wg/cgMLST and SNP typing schemes of representative clinically relevant bacteria

Organism	Relatedness threshold[a]		References
	wg/cgMLST (allele)	SNPs	
Acinetobacter baumannii	≤8	≤3	[25,26]
Brucella spp.	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Campylobacter coli, C. jejuni	≤14	≤15	[27,28]
Cronobacter spp.	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Clostridium difficile	Epidemiologic validation in progress[b]	≤4	[29], http://www.cgmlst.org/ncs, http://www.applied-maths.com/applications/wgmlst
Enterococcus faecium	≤20	≤16	[30]
Enterococcus raffinosus	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Escherichia coli	≤10	≤10	[31,32], https://enterobase.warwick.ac.uk/
Francisella tularensis	≤1	≤2	[33,34]
Klebsiella oxytoca	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Klebsiella pneumonia	≤10	≤18	[35,36]
Legionella pneumophila	≤4	≤15	[37]
Listeria monocytogenes	≤10	≤3	[38,39]
Mycobacterium abscessus		≤30	[40]
Mycobacterium tuberculosis	≤12	≤12	[41]
Neisseria gonorrhoeae	Epidemiologic validation in progress[b]	≤14	[42], http://www.applied-maths.com/applications/wgmlst
Neisseria meningitidis	Epidemiologic validation in progress[b]		http://www.cgmlst.org/ncs
Pseudomonas aeruginosa	≤14	≤37	[31,43]
Salmonella dublin	Epidemiologic validation in progress[b]	≤13	[44], https://enterobase.warwick.ac.uk/
Salmonella enterica	Epidemiologic validation in progress[b]	≤4	[45], http://www.cgmlst.org/ncs, http://www.applied-maths.com/applications/wgmlst, https://enterobase.warwick.ac.uk/
Salmonella typhimurium	Epidemiologic validation in progress[b]	≤2	[46], https://enterobase.warwick.ac.uk/
Staphylococcus aureus	≤24	≤15	[47,48]
Streptococcus suis		≤21	[49]
Vibrio parahaemolyticus	≤10		[50]
Yersinia spp.	0		[51]

When nature messes with you…..
HYPERMUTATORS



When nature messes with you. HYPERMUTATORS

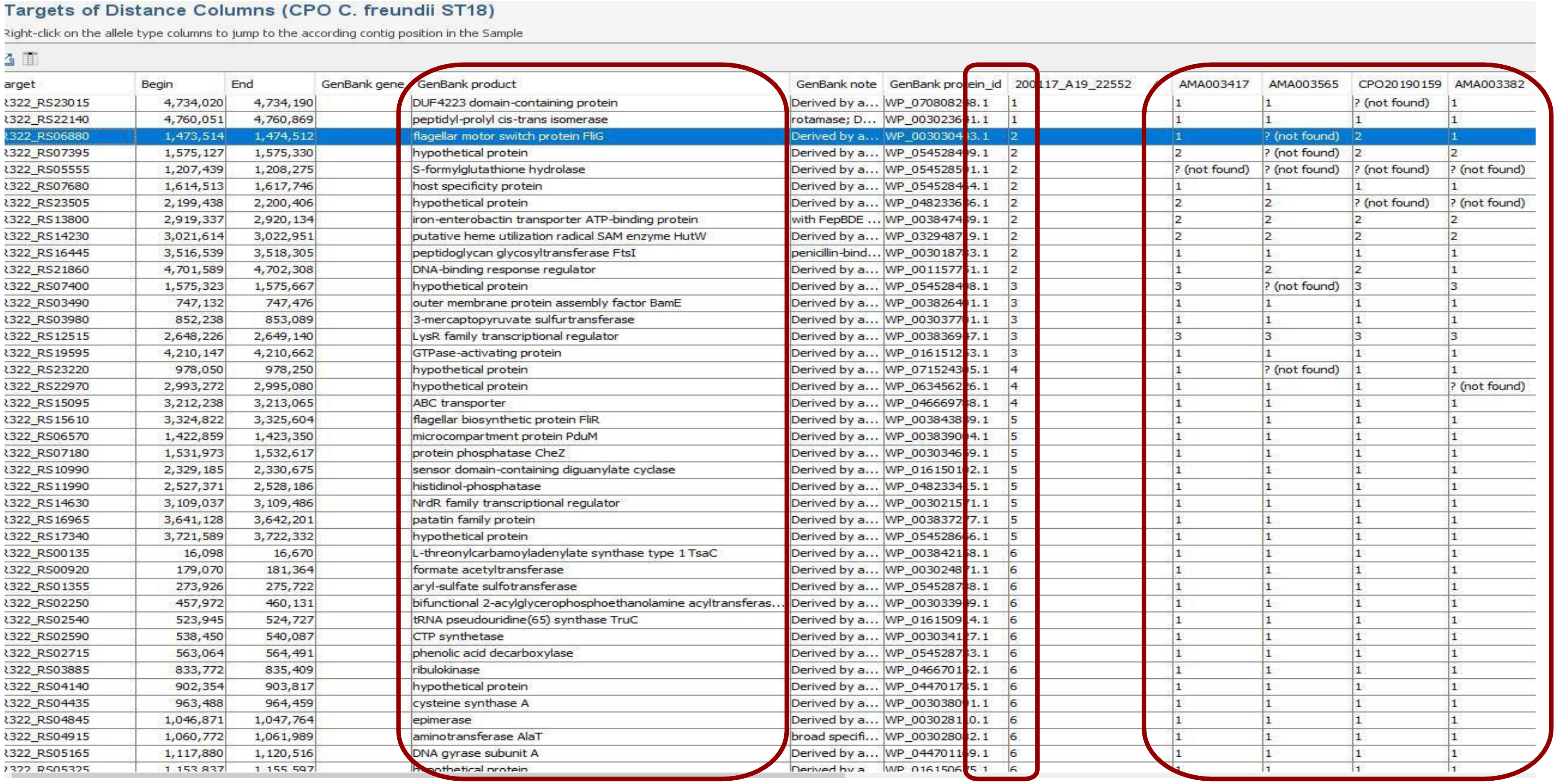


Targets of Distance Columns (CPO C. freundii ST18)

Right-click on the allele type columns to jump to the according contig position in the Sample

arget	Begin	End	GenBank gene	GenBank product	GenBank note	GenBank protein_id	200117_A19_...	AMA003417	AMA003565	CPO20190159	AMA00338
.322_RS02285	465,622	467,868		phosphoenolpyruvate--protein phosphotransferase PtsP	member of a ...	WP_003033984.1	? (failed)	1	1	1	1
.322_RS03425	716,674	721,020		autotransporter domain-containing protein	Derived by a...	WP_071684359.1	? (failed)	1	1	1	? (not found
.322_RS03765	802,371	803,735		PTS sugar transporter subunit IIC	Derived by a...	WP_054528657.1	? (failed)	1	? (not found)	1	1
.322_RS04195	912,678	914,693		tRNA(Met) cytidine acetyltransferase TmcA	cetylates the...	WP_054528641.1	? (failed)	1	1	1	1
.322_RS05765	1,252,228	1,254,714		fimbrial assembly protein	Derived by a...	WP_054528576.1	? (failed)	1	1	1	1
.322_RS06635	1,433,991	1,435,370		cobyrinic acid a,c-diamide synthase	Derived by a...	WP_044701540.1	? (failed)	1	1	1	1
.322_RS06975	1,490,987	1,491,643		DNA-binding response regulator	Derived by a...	WP_003030486.1	? (failed)	1	1	1	1
.322_RS09275	1,966,895	1,967,473		TetR family transcriptional regulator	Derived by a...	WP_046670695.1	? (failed)	1	1	1	1
.322_RS09820	2,083,766	2,084,500		DNA-binding response regulator	Derived by a...	WP_003836390.1	? (failed)	1	1	1	1
.322_RS11920	2,514,178	2,514,801		DSBA oxidoreductase	Derived by a...	WP_003035975.1	? (failed)	1	1	1	1
.322_RS12760	2,702,258	2,703,679		2-oxoglutarate/malate translocator	Derived by a...	WP_003837022.1	? (failed)	1	1	1	1
.322_RS13805	2,920,181	2,921,314		LPS O-antigen length regulator	Derived by a...	WP_054528176.1	? (failed)	1	1	1	1
.322_RS14935	3,176,896	3,177,909		4-hydroxy-2-oxovalerate aldolase	Derived by a...	WP_003021379.1	? (failed)	1	1	1	1
.322_RS15475	3,301,194	3,301,901		flagellar basal body L-ring protein	Derived by a...	WP_042270212.1	? (failed)	1	1	1	1
.322_RS15765	3,357,582	3,359,564		type IV secretion protein Rhs	Derived by a...	WP_072143931.1	? (failed)	1	1	1	1
.322_RS19670	4,225,537	4,226,988		potassium transporter	Derived by a...	WP_003017848.1	? (failed)	1	1	1	1
.322_RS21515	4,624,296	4,625,513		MFS transporter	Derived by a...	WP_054528867.1	? (failed)	1	1	1	1
.322_RS07405	1,575,702	1,576,793		enterohemolysin	Derived by a...	WP_054528497.1	? (not found)	? (not found)	? (not found)	? (not found)	? (not found
.322_RS07605	1,603,957	1,604,262		hypothetical protein	Derived by a...	WP_057101149.1	? (not found)	? (not found)	? (not found)	? (not found)	? (not found
.322_RS08700	1,830,199	1,830,762		hypothetical protein	Derived by a...	WP_003843940.1	? (not found)	? (not found)	? (not found)	? (not found)	? (not found
.322_RS17560	3,773,522	3,773,764		transcriptional regulator	Qin prophag...	WP_003839576.1	? (not found)	1	? (not found)	? (not found)	1
.322_RS22180	4,766,146	4,767,330		elongation factor Tu	Derived by a...	WP_003031109.1	? (not found)	? (not found)	? (not found)	? (not found)	? (not found
.322_RS06380	1,382,478	1,383,593		amino acid oxidase	Derived by a...	WP_054528547.1	? (not found)	1	? (not found)	? (not found)	? (not found
.322_RS17175	3,686,035	3,686,850		AraC family transcriptional regulator	Derived by a...	WP_054528023.1	? (not found)	1	1	1	1
.322_RS17930	3,844,410	3,846,275		DNA mismatch repair protein MutL	Derived by a...	WP_054527983.1	? (not found)	1	1	1	1
.322_RS20890	4,482,716	4,483,960		O-antigen polymerase	Derived by a...	WP_046671022.1	? (not found)	1	1	? (not found)	? (not found
.322_RS22035	4,737,170	4,739,713		nitrite reductase large subunit	Derived by a...	WP_003023592.1	? (not found)	1	1	1	1
.322_RS23080	211,671	211,868		hypothetical protein	Derived by a...	WP_072143936.1	1	1	1	1	1
.322_RS23130	545,890	546,069		hypothetical protein	Derived by a...	WP_071524456.1	1	1	1	? (not found)	1
.322_RS02845	588,933	589,343		formate hydrogenlyase maturation protein HycH	required for ...	WP_016150885.1	1	2	1	1	2
.322_RS03030	623,789	624,124		L-valine transporter subunit YgaH	Derived by a...	WP_054528723.1	1	1	? (not found)	? (not found)	1
.322_RS03050	627,414	628,478		proline/betaine ABC transporter permease ProW	Derived by a...	WP_003846040.1	1	1	? (not found)	? (not found)	1
.322_RS03070	633,115	633,525	nrdI	ribonucleotide reductase assembly protein NrdI	in Salmonella...	WP_003037273.1	1	1	? (not found)	? (not found)	1
.322_RS03235	667,222	668,508		capsular polysaccharide biosynthesis protein	Derived by a...	WP_003839728.1	1	1	1	1	1

When nature messes with you. HYPERMUTATORS



Targets of Distance Columns (CPO C. freundii ST18)

Right-click on the allele type columns to jump to the according contig position in the Sample

arget	Begin	End	GenBank gene	GenBank product	GenBank note	GenBank pro[illegible]ein_id	200[illegible]17_A19_22552	AMA003417	AMA003565	CPO20190159	AMA003382
322_RS23015	4,734,020	4,734,190		DUF4223 domain-containing protein	Derived by a...	WP_0708082[illegible]8.1	1	1	1	? (not found)	1
322_RS22140	4,760,051	4,760,869		peptidyl-prolyl cis-trans isomerase	rotamase; D...	WP_0030236[illegible]1.1	1	1	1	1	1
322_RS06880	1,473,514	1,474,512		flagellar motor switch protein FliG	Derived by a...	WP_0030304[illegible]3.1	2	1	? (not found)	2	1
322_RS07395	1,575,127	1,575,330		hypothetical protein	Derived by a...	WP_0545284[illegible]9.1	2	2	? (not found)	2	2
322_RS05555	1,207,439	1,208,275		S-formylglutathione hydrolase	Derived by a...	WP_0545285[illegible]1.1	2	? (not found)	? (not found)	? (not found)	? (not found)
322_RS07680	1,614,513	1,617,746		host specificity protein	Derived by a...	WP_0545284[illegible]4.1	2	1	1	1	1
322_RS23505	2,199,438	2,200,406		hypothetical protein	Derived by a...	WP_0482336[illegible]6.1	2	2	2	? (not found)	? (not found)
322_RS13800	2,919,337	2,920,134		iron-enterobactin transporter ATP-binding protein	with FepBDE ...	WP_0038474[illegible]9.1	2	2	2	2	2
322_RS14230	3,021,614	3,022,951		putative heme utilization radical SAM enzyme HutW	Derived by a...	WP_0329487[illegible]9.1	2	2	2	2	2
322_RS16445	3,516,539	3,518,305		peptidoglycan glycosyltransferase FtsI	penicillin-bind...	WP_0030187[illegible]3.1	2	1	1	1	1
322_RS21860	4,701,589	4,702,308		DNA-binding response regulator	Derived by a...	WP_0011577[illegible]1.1	2	1	2	2	1
322_RS07400	1,575,323	1,575,667		hypothetical protein	Derived by a...	WP_0545284[illegible]8.1	3	3	? (not found)	3	3
322_RS03490	747,132	747,476		outer membrane protein assembly factor BamE	Derived by a...	WP_0038264[illegible]1.1	3	1	1	1	1
322_RS03980	852,238	853,089		3-mercaptopyruvate sulfurtransferase	Derived by a...	WP_0030377[illegible]1.1	3	1	1	1	1
322_RS12515	2,648,226	2,649,140		LysR family transcriptional regulator	Derived by a...	WP_0038369[illegible]7.1	3	3	3	3	3
322_RS19595	4,210,147	4,210,662		GTPase-activating protein	Derived by a...	WP_0161512[illegible]3.1	3	1	1	1	1
322_RS23220	978,050	978,250		hypothetical protein	Derived by a...	WP_0715243[illegible]5.1	4	1	? (not found)	1	1
322_RS22970	2,993,272	2,995,080		hypothetical protein	Derived by a...	WP_0634562[illegible]6.1	4	1	1	1	? (not found)
322_RS15095	3,212,238	3,213,065		ABC transporter	Derived by a...	WP_0466697[illegible]8.1	4	1	1	1	1
322_RS15610	3,324,822	3,325,604		flagellar biosynthetic protein FliR	Derived by a...	WP_0038438[illegible]9.1	5	1	1	1	1
322_RS06570	1,422,859	1,423,350		microcompartment protein PduM	Derived by a...	WP_0038390[illegible]4.1	5	1	1	1	1
322_RS07180	1,531,973	1,532,617		protein phosphatase CheZ	Derived by a...	WP_0030346[illegible]9.1	5	1	1	1	1
322_RS10990	2,329,185	2,330,675		sensor domain-containing diguanylate cyclase	Derived by a...	WP_0161501[illegible]2.1	5	1	1	1	1
322_RS11990	2,527,371	2,528,186		histidinol-phosphatase	Derived by a...	WP_0482334[illegible]5.1	5	1	1	1	1
322_RS14630	3,109,037	3,109,486		NrdR family transcriptional regulator	Derived by a...	WP_0030215[illegible]1.1	5	1	1	1	1
322_RS16965	3,641,128	3,642,201		patatin family protein	Derived by a...	WP_0038372[illegible]7.1	5	1	1	1	1
322_RS17340	3,721,589	3,722,332		hypothetical protein	Derived by a...	WP_0545286[illegible]6.1	5	1	1	1	1
322_RS00135	16,098	16,670		L-threonylcarbamoyladenylate synthase type 1 TsaC	Derived by a...	WP_0038421[illegible]8.1	6	1	1	1	1
322_RS00920	179,070	181,364		formate acetyltransferase	Derived by a...	WP_0030248[illegible]1.1	6	1	1	1	1
322_RS01355	273,926	275,722		aryl-sulfate sulfotransferase	Derived by a...	WP_0545287[illegible]8.1	6	1	1	1	1
322_RS02250	457,972	460,131		bifunctional 2-acylglycerophosphoethanolamine acyltransferas...	Derived by a...	WP_0030339[illegible]9.1	6	1	1	1	1
322_RS02540	523,945	524,727		tRNA pseudouridine(65) synthase TruC	Derived by a...	WP_0161509[illegible]4.1	6	1	1	1	1
322_RS02590	538,450	540,087		CTP synthetase	Derived by a...	WP_0030341[illegible]7.1	6	1	1	1	1
322_RS02715	563,064	564,491		phenolic acid decarboxylase	Derived by a...	WP_0545287[illegible]3.1	6	1	1	1	1
322_RS03885	833,772	835,409		ribulokinase	Derived by a...	WP_0466701[illegible]2.1	6	1	1	1	1
322_RS04140	902,354	903,817		hypothetical protein	Derived by a...	WP_0447017[illegible]5.1	6	1	1	1	1
322_RS04435	963,488	964,459		cysteine synthase A	Derived by a...	WP_0030380[illegible]1.1	6	1	1	1	1
322_RS04845	1,046,871	1,047,764		epimerase	Derived by a...	WP_0030281[illegible]0.1	6	1	1	1	1
322_RS04915	1,060,772	1,061,989		aminotransferase AlaT	broad specifi...	WP_0030280[illegible]2.1	6	1	1	1	1
322_RS05165	1,117,880	1,120,516		DNA gyrase subunit A	Derived by a...	WP_0447011[illegible]9.1	6	1	1	1	1
322_RS05325	1,153,837	1,155,597		hypothetical protein	Derived by a...	WP_0161506[illegible]5.1	6	1	1	1	1

The Fleming Fund
Regional Grants

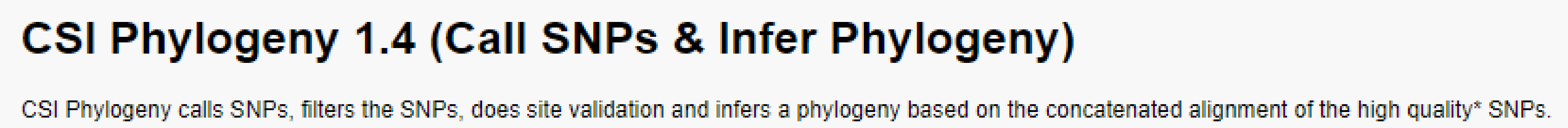
Let’s take a break ☺

The Fleming Fund
Regional Grants

CSI Phylogeny

Focus on (CSI) phylogeny

- Phylogenetic comparisons allow for determining clusters and clonal spread of microorganisms.
- SNP calling – to determine variants in the DNA (Single Nucleotide Polymorphism)
- Different sequencing technologies have systematic biases, 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/

Data quality and SNP calling

- 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 (≥30x).
- Consider the quality of your raw data (specifically phred scores).

- CSI Phylogeny SNP filtering criteria:

- SNP quality: ≥30 (Phred score, base call accuracy: 99.9%)

- SNPs with a sequence depth of <10 are removed.

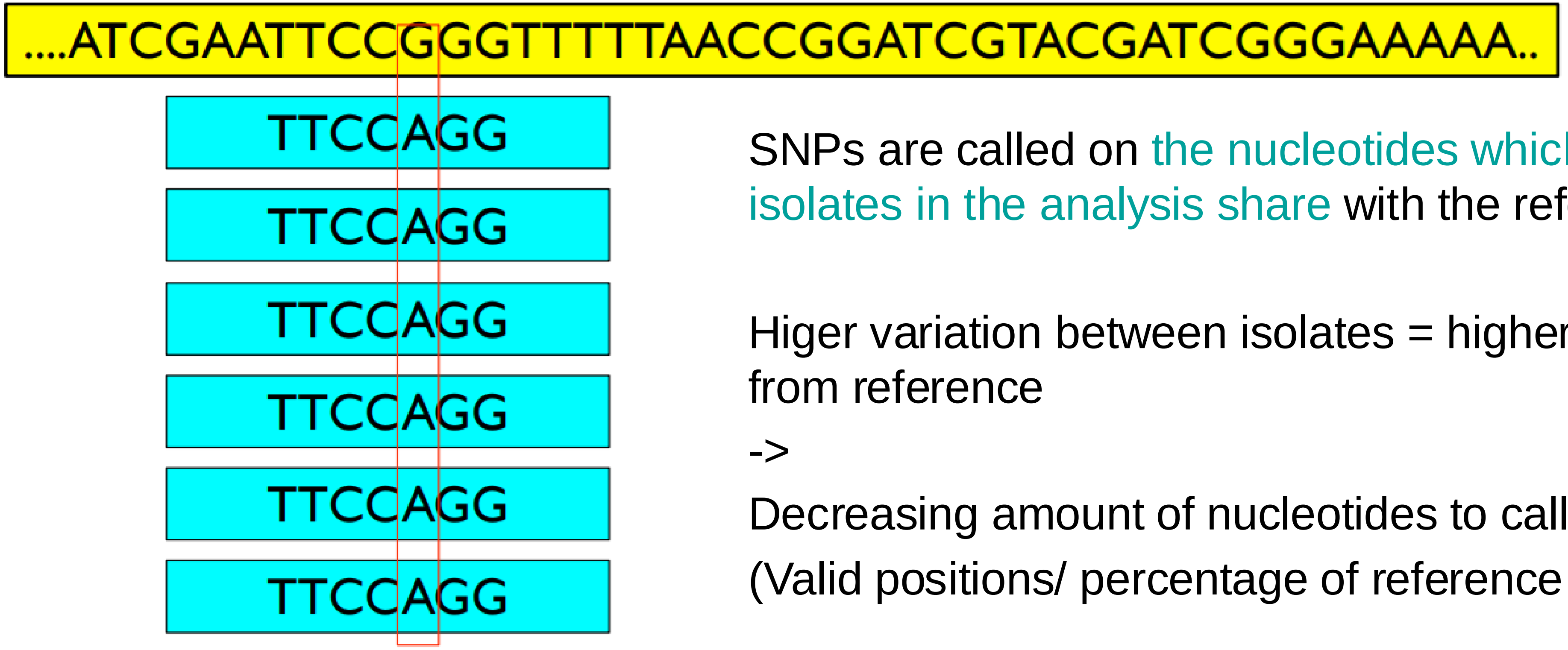
- A SNP is removed if it is <10 bps from the nearest SNP (Pruning) (recombination do not reflect naturally evolved SNPs).

Preferably analyse raw reads for better resolution!

SNPs detection (CSIPhylogeny)

Calling of single nucleotide polymorphism

- Variants in the DNA – compared to reference



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

Higer variation between isolates = higher difference from reference

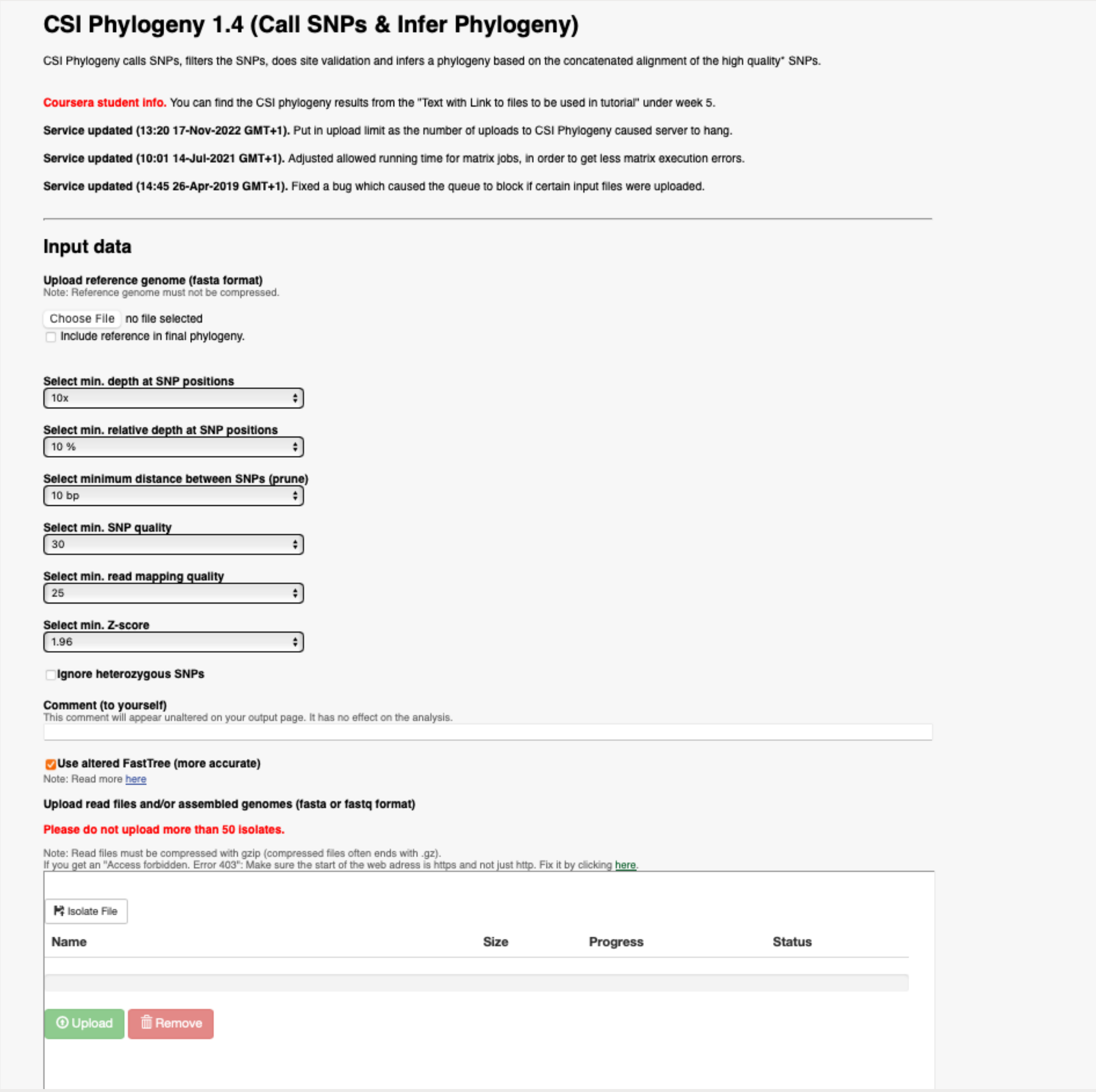
->

Decreasing amount of nucleotides to call SNPs from

(Valid positions/ percentage of reference covered)

The Fleming Fund
Regional Grants

CSIphylogeny - webtool



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.

Coursera student info. You can find the CSI phylogeny results from the "Text with Link to files to be used in tutorial" under week 5.

Service updated (13:20 17-Nov-2022 GMT+1). Put in upload limit as the number of uploads to CSI Phylogeny caused server to hang.

Service updated (10:01 14-Jul-2021 GMT+1). Adjusted allowed running time for matrix jobs, in order to get less matrix execution errors.

Service updated (14:45 26-Apr-2019 GMT+1). Fixed a bug which caused the queue to block if certain input files were uploaded.

Input data

Upload reference genome (fasta format)
Note: Reference genome must not be compressed.

Choose File no file selected
□ 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**

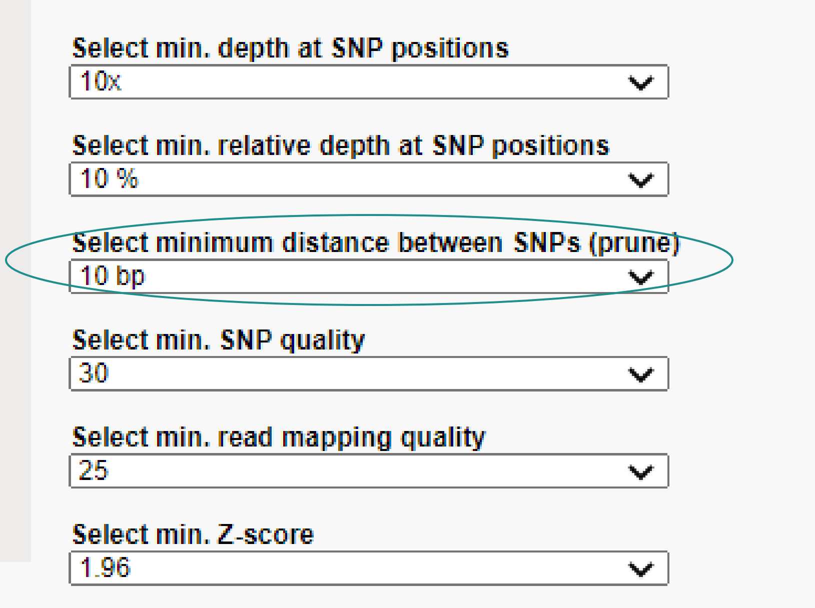
Comment (to yourself)
This comment will appear unaltered on your output page. It has no effect on the analysis.

☑ **Use altered FastTree (more accurate)**
Note: Read more here

Upload read files and/or assembled genomes (fasta or fastq format)

Please do not upload more than 50 isolates.

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 adress is https and not just http. Fix it by clicking here.



Select min. depth at SNP positions
10x

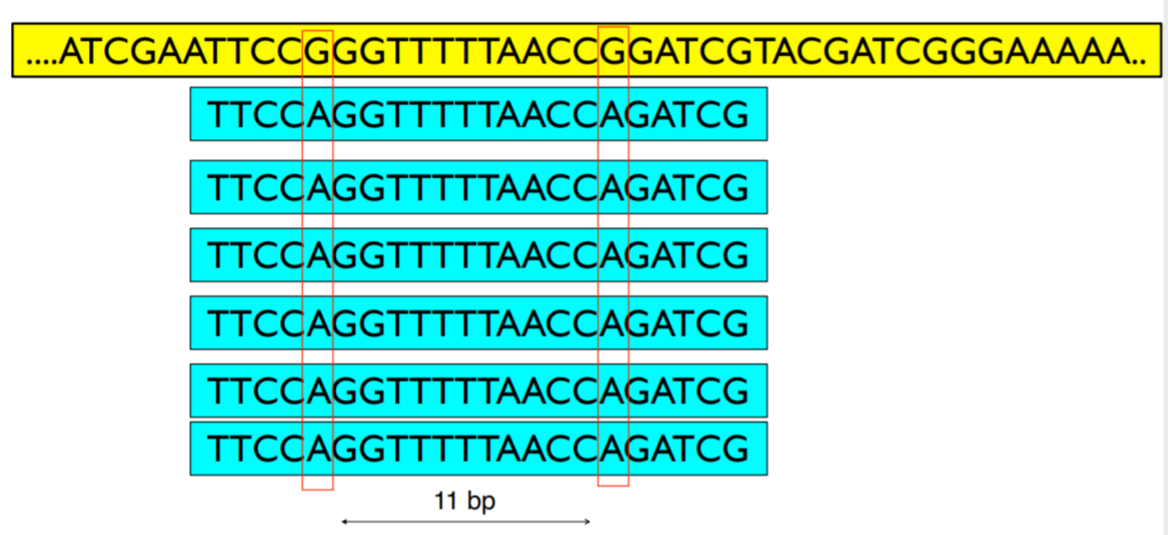
Select min. relative depth at SNP positions
10 %

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

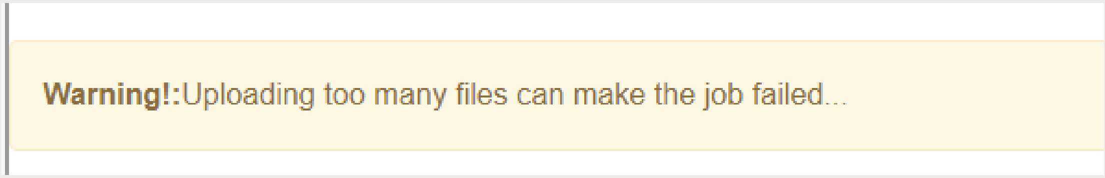


The Fleming Fund
Regional Grants

CSIphylogeny - webtool

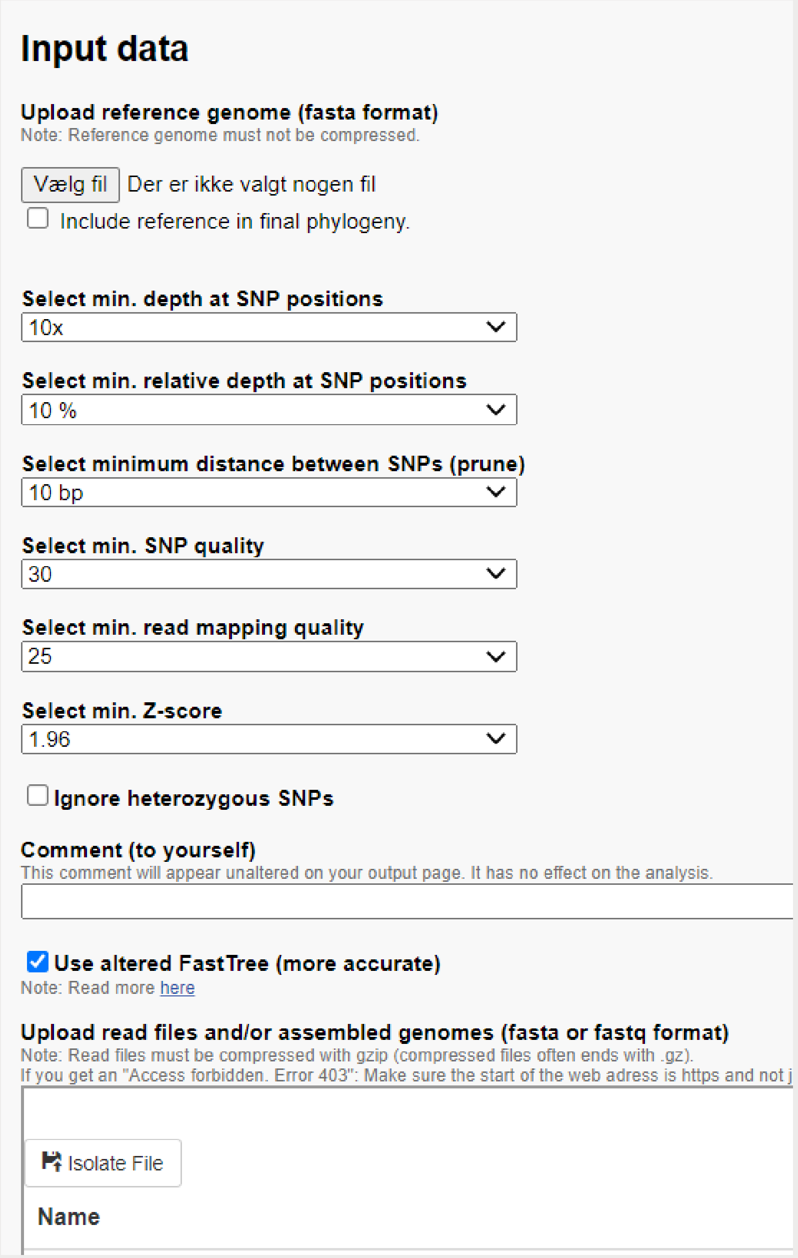
- Input data:

- 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



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**

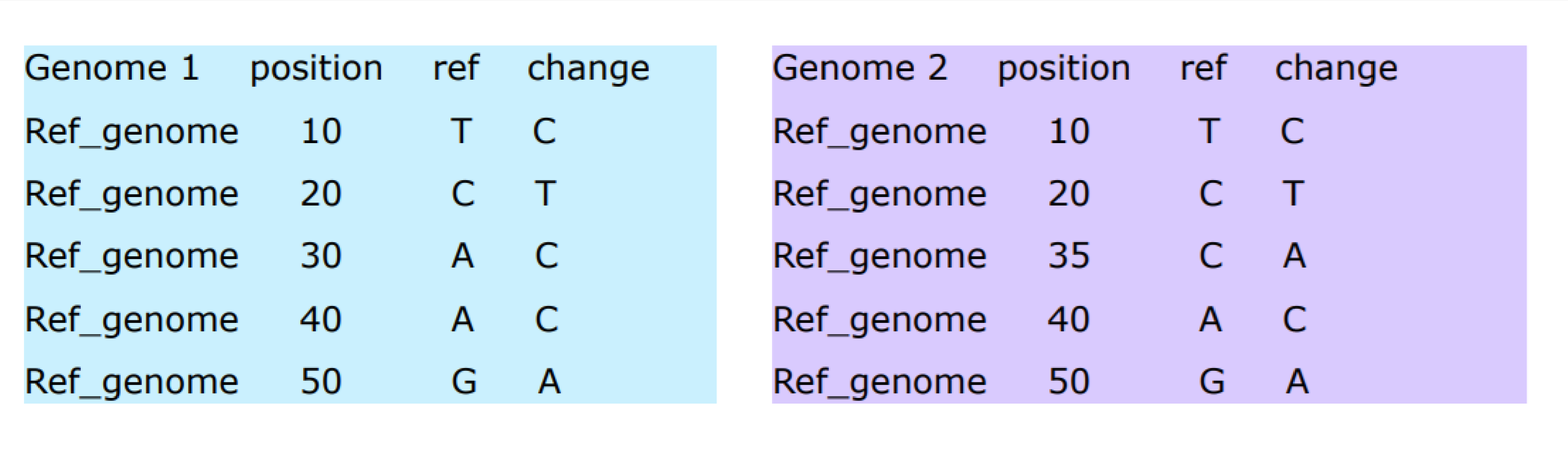
Comment (to yourself)
This comment will appear unaltered on your output page. It has no effect on the analysis.

☑ **Use altered FastTree (more accurate)**
Note: Read more here

Upload read files and/or assembled genomes (fasta or 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 adress is https and not j

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: SNP matrix

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

SNP Matrix - example

- Plain text file – open in Excel

	E_coli_NZ_CP033092_2	TC2021-01_	TC2021-02_	TC2021-04_	TC2021-05_	TC2021-07_	TC2021-08_	TC2021-09_	TC2021-10_	TC2021-11_	TC2021-12_	TC2021-Extra01_	TC2021_Extra02_
E_coli_NZ_CP033092_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

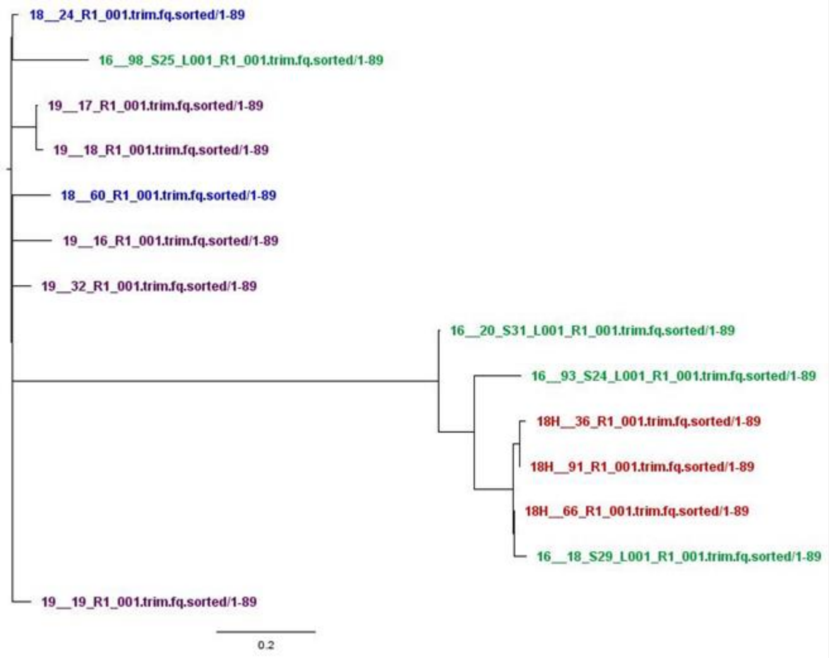
	E_coli_NZ_CP033092_2	TC2021-01_	TC2021-02_	TC2021-04_	TC2021-05_	TC2021-07_	TC2021-08_	TC2021-09_	TC2021-10_	TC2021-11_	TC2021-12_	TC2021-Extra01_	TC2021_Extra02_
E_coli_NZ_CP033092_2	0												
TC2021-01_	29753	0											
TC2021-02_	30187	10003	0										
TC2021-04_	26060	32323	32549	0									
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

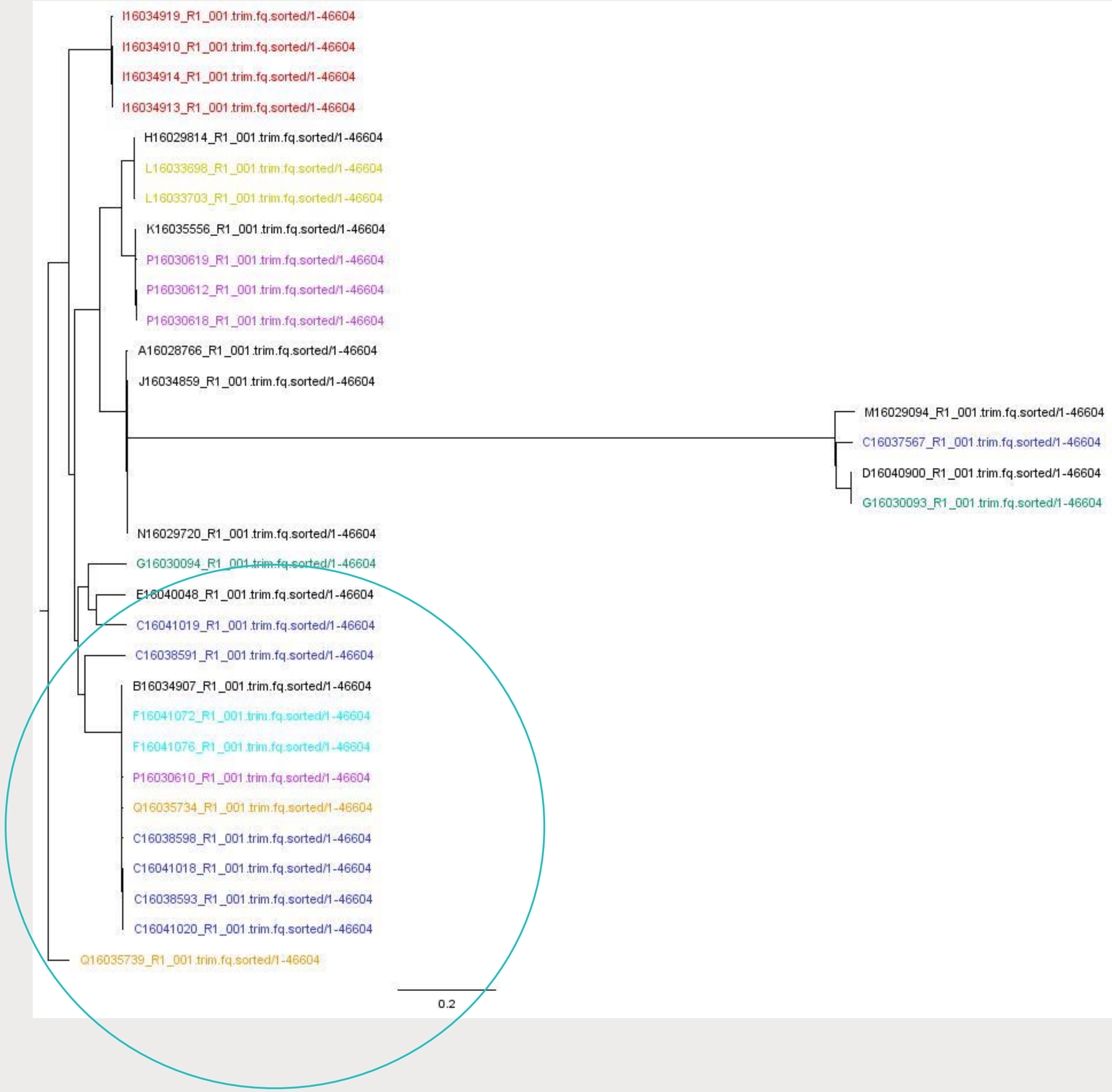
Below	1000	SNPs
Below	100	SNPs
Below	10	SNPs

Outputs from SNP analysis: Newick file

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

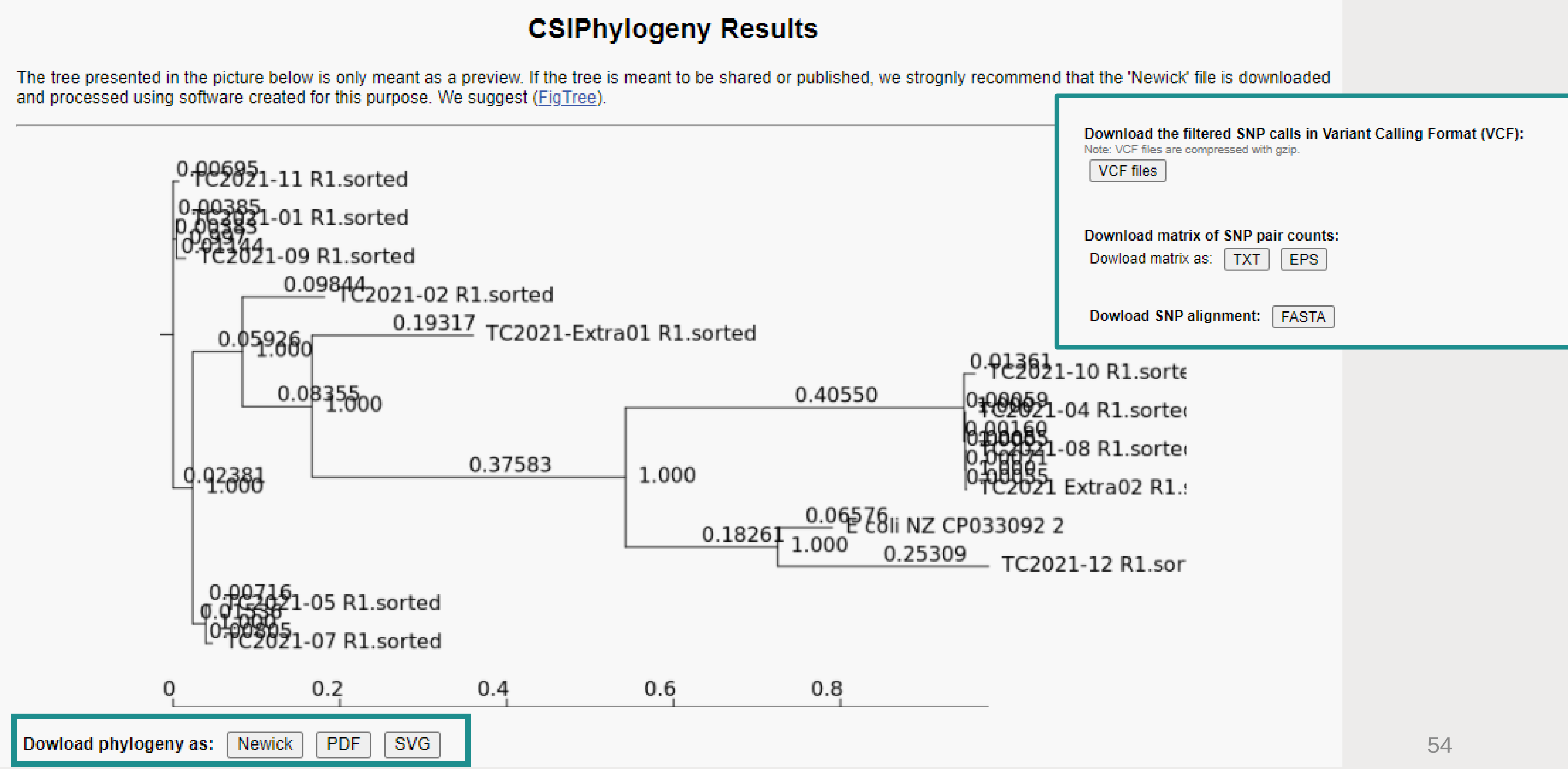


18__24_R1_001.trim.fq.sorted/1-89
16__98_S25_L001_R1_001.trim.fq.sorted/1-89
19__17_R1_001.trim.fq.sorted/1-89
19__18_R1_001.trim.fq.sorted/1-89
18__60_R1_001.trim.fq.sorted/1-89
19__16_R1_001.trim.fq.sorted/1-89
19__32_R1_001.trim.fq.sorted/1-89
16__20_S31_L001_R1_001.trim.fq.sorted/1-89
16__93_S24_L001_R1_001.trim.fq.sorted/1-89
18H__36_R1_001.trim.fq.sorted/1-89
18H__91_R1_001.trim.fq.sorted/1-89
18H__66_R1_001.trim.fq.sorted/1-89
16__18_S29_L001_R1_001.trim.fq.sorted/1-89
19__19_R1_001.trim.fq.sorted/1-89
0.2



I16034919_R1_001.trim.fq.sorted/1-46604
I16034910_R1_001.trim.fq.sorted/1-46604
I16034914_R1_001.trim.fq.sorted/1-46604
I16034913_R1_001.trim.fq.sorted/1-46604
H16029814_R1_001.trim.fq.sorted/1-46604
L16033698_R1_001.trim.fq.sorted/1-46604
L16033703_R1_001.trim.fq.sorted/1-46604
K16035556_R1_001.trim.fq.sorted/1-46604
P16030619_R1_001.trim.fq.sorted/1-46604
P16030612_R1_001.trim.fq.sorted/1-46604
P16030618_R1_001.trim.fq.sorted/1-46604
A16028766_R1_001.trim.fq.sorted/1-46604
J16034859_R1_001.trim.fq.sorted/1-46604
M16029094_R1_001.trim.fq.sorted/1-46604
C16037567_R1_001.trim.fq.sorted/1-46604
D16040900_R1_001.trim.fq.sorted/1-46604
G16030093_R1_001.trim.fq.sorted/1-46604
N16029720_R1_001.trim.fq.sorted/1-46604
G16030094_R1_001.trim.fq.sorted/1-46604
E16040048_R1_001.trim.fq.sorted/1-46604
C16041019_R1_001.trim.fq.sorted/1-46604
C16038591_R1_001.trim.fq.sorted/1-46604
B16034907_R1_001.trim.fq.sorted/1-46604
F16041072_R1_001.trim.fq.sorted/1-46604
F16041076_R1_001.trim.fq.sorted/1-46604
P16030610_R1_001.trim.fq.sorted/1-46604
Q16035734_R1_001.trim.fq.sorted/1-46604
C16038598_R1_001.trim.fq.sorted/1-46604
C16041018_R1_001.trim.fq.sorted/1-46604
C16038593_R1_001.trim.fq.sorted/1-46604
C16041020_R1_001.trim.fq.sorted/1-46604
Q16035739_R1_001.trim.fq.sorted/1-46604
0.2

CSI output – web interface



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 strognly 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:
Dowload matrix as: TXT EPS

Dowload SNP alignment: FASTA

Dowload phylogeny as: Newick PDF SVG

The Fleming Fund
Regional Grants

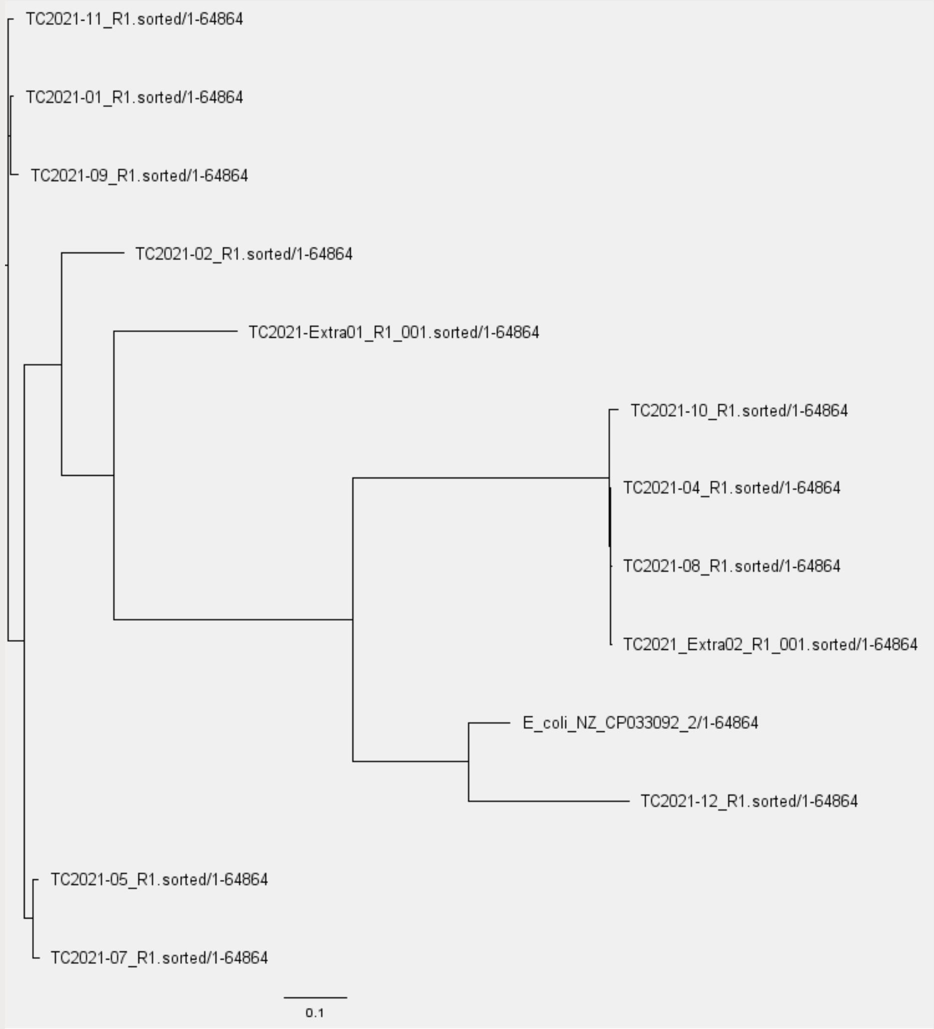
Newick file

- Text file – SNP distances
- Use various tools to visualise the phylogenetic tree

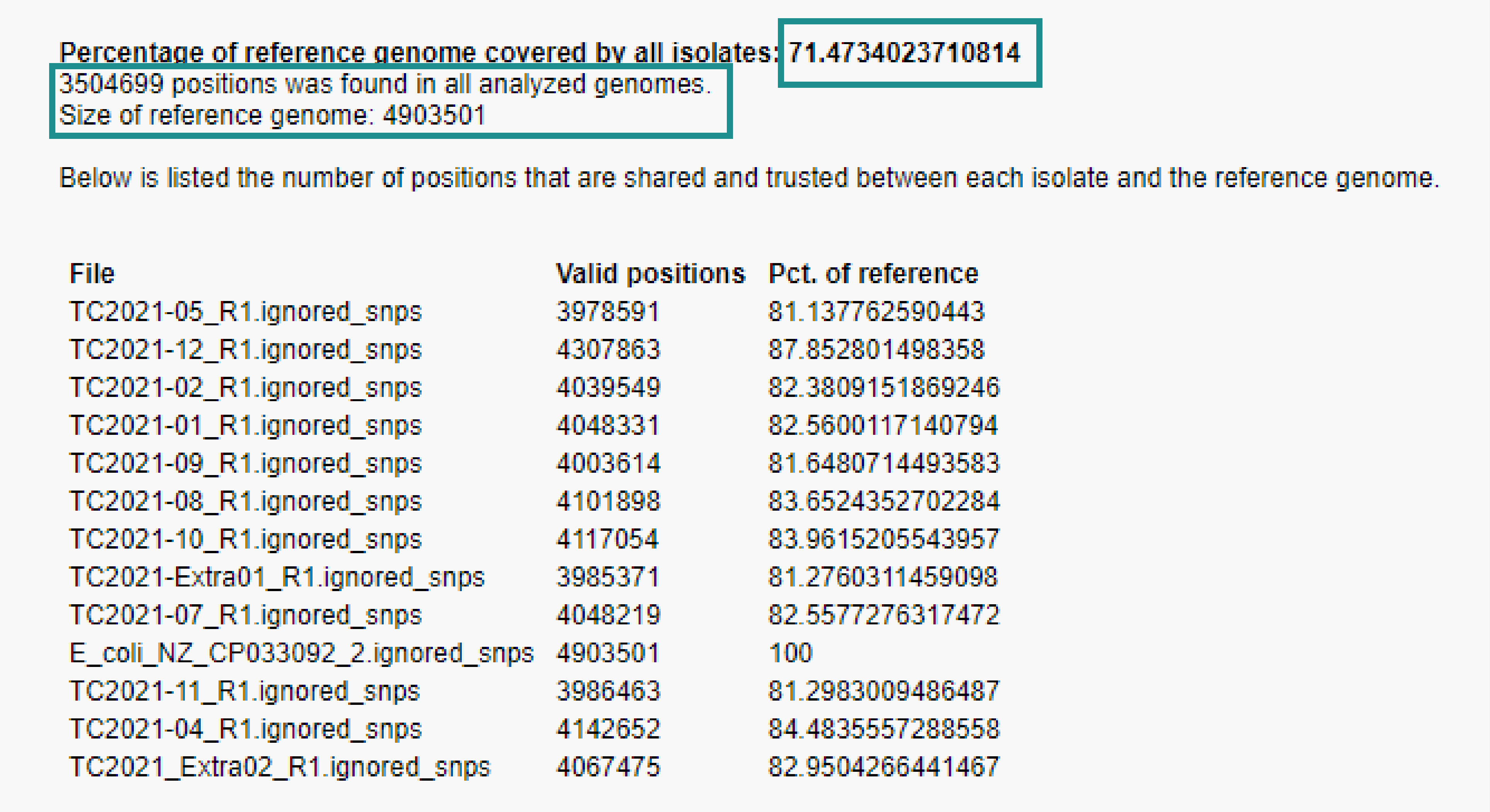
- Here: FigTree
- https://github.com/rambaut/figtree/releases

- CGE tool:
  - TreeViewer

- Microreact, iTOL…
  - https://microreact.org/upload


TC2021-11_R1.sorted/1-64864
TC2021-01_R1.sorted/1-64864
TC2021-09_R1.sorted/1-64864
TC2021-02_R1.sorted/1-64864
TC2021-Extra01_R1_001.sorted/1-64864
TC2021-10_R1.sorted/1-64864
TC2021-04_R1.sorted/1-64864
TC2021-08_R1.sorted/1-64864
TC2021_Extra02_R1_001.sorted/1-64864
E_coli_NZ_CP033092_2/1-64864
TC2021-12_R1.sorted/1-64864
TC2021-05_R1.sorted/1-64864
TC2021-07_R1.sorted/1-64864
0.1

CSI outputs



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

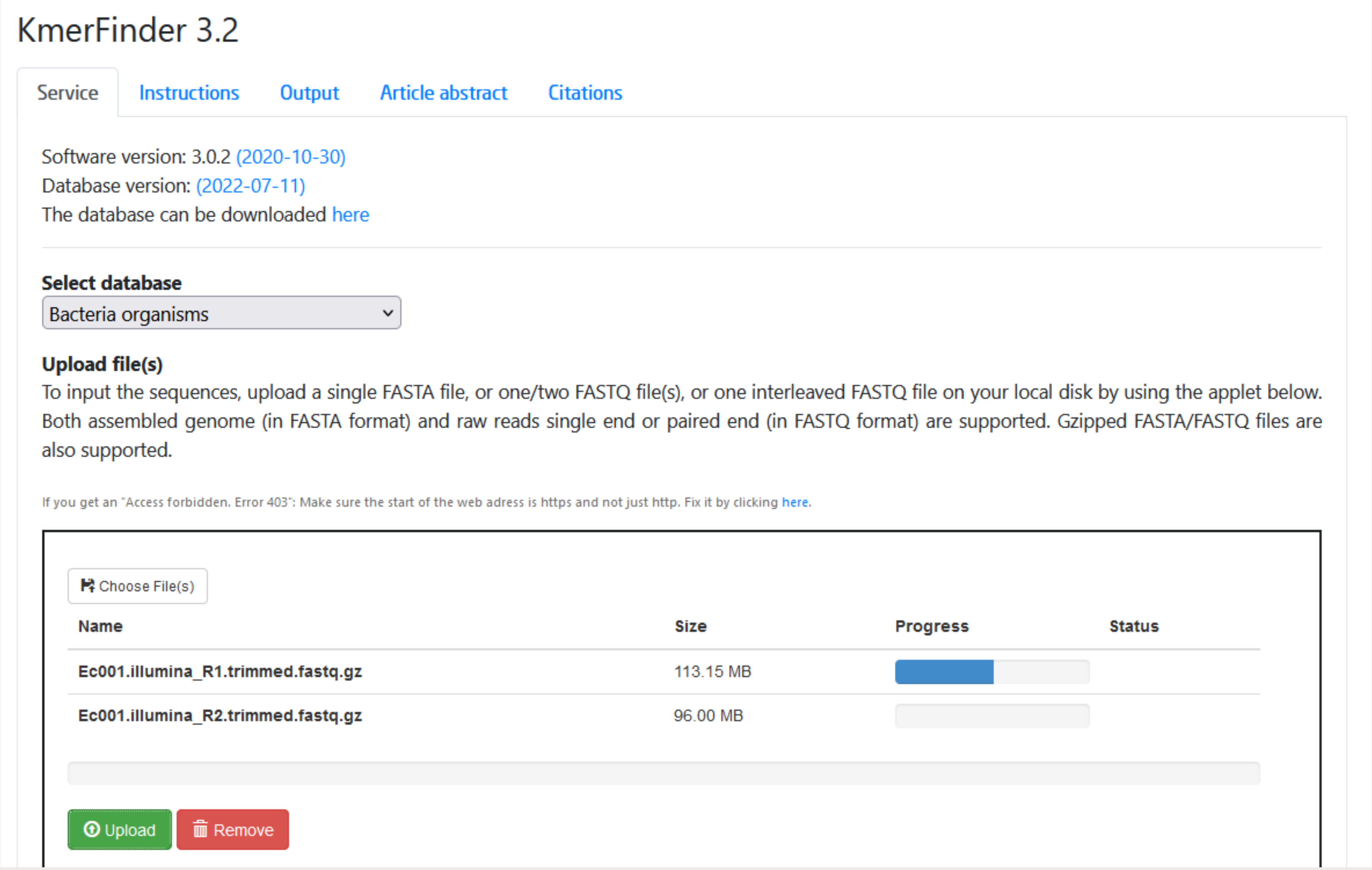
How to choose a reference

- 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

Kmer-finder –species ID and contamination



KmerFinder 3.2

Software version: 3.0.2 (2020-10-30)
Database version: (2022-07-11)
The database can be downloaded here

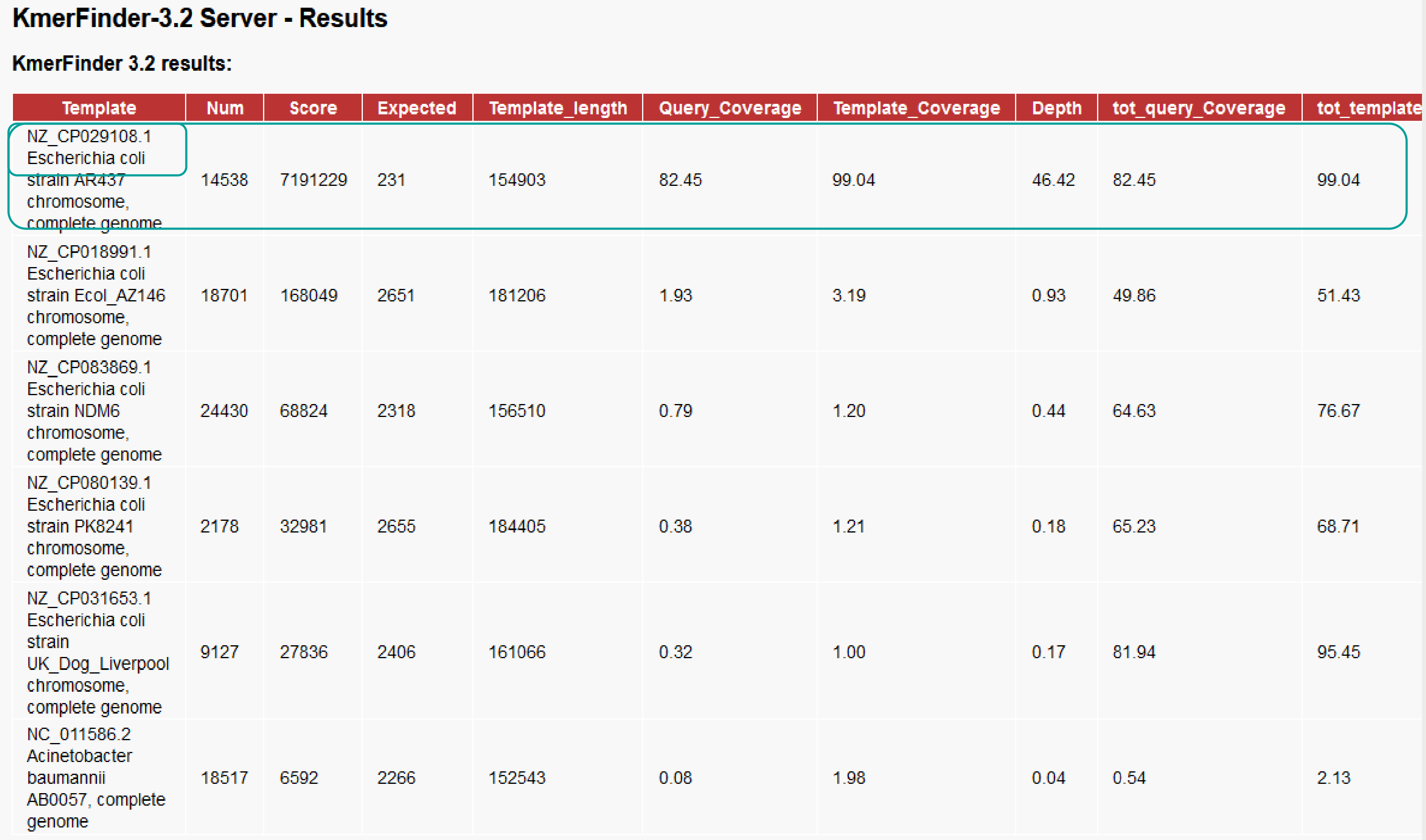
Select database
Bacteria organisms

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.

Name	Size	Progress	Status
Ec001.illumina_R1.trimmed.fastq.gz	113.15 MB		
Ec001.illumina_R2.trimmed.fastq.gz	96.00 MB		

Kmer-finder – find a reference



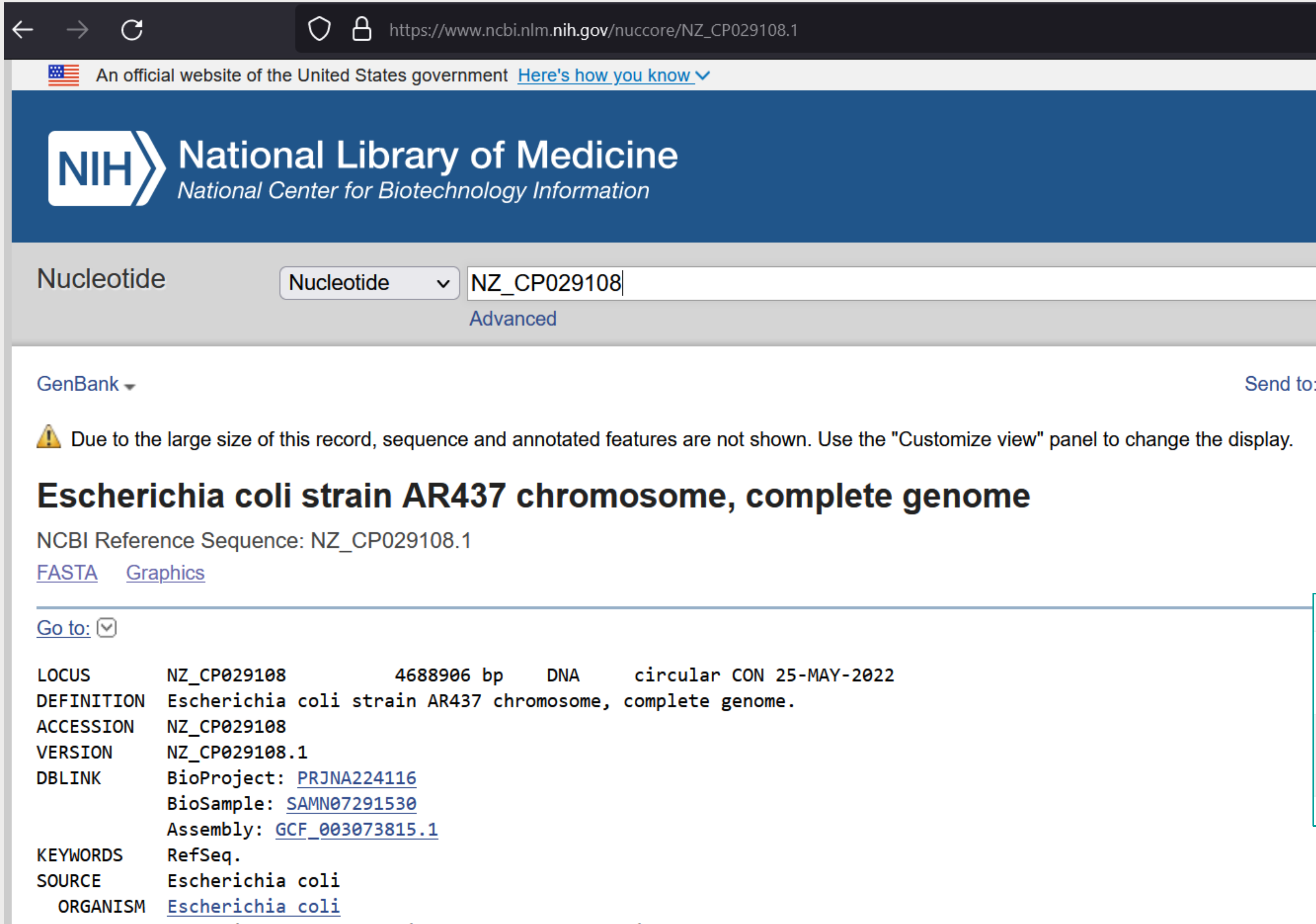
KmerFinder-3.2 Server - Results

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

The Fleming Fund
Regional Grants

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


https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP029108.1

An official website of the United States government Here's how you know

NIH National Library of Medicine
National Center for Biotechnology Information

Nucleotide | Nucleotide | NZ_CP029108
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

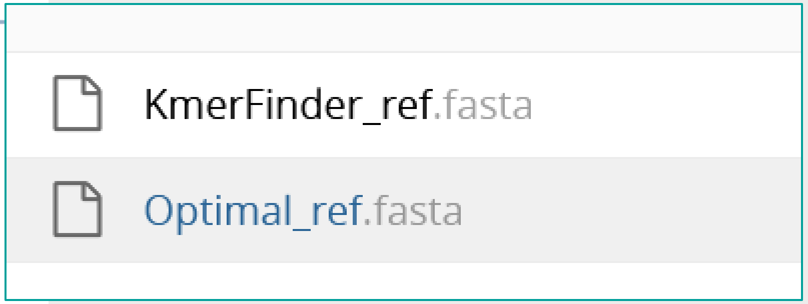
FASTA Graphics

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

Mintyper

MinION – the new(ish) kid on the block

6-15 days

6-48 hours

Relatively..

- low price per isolate
- well-proven technology
- high precision (low error rate)
- Slow (depending on the setup)

..but no reads in real-time

Tools for outbreak detection validated

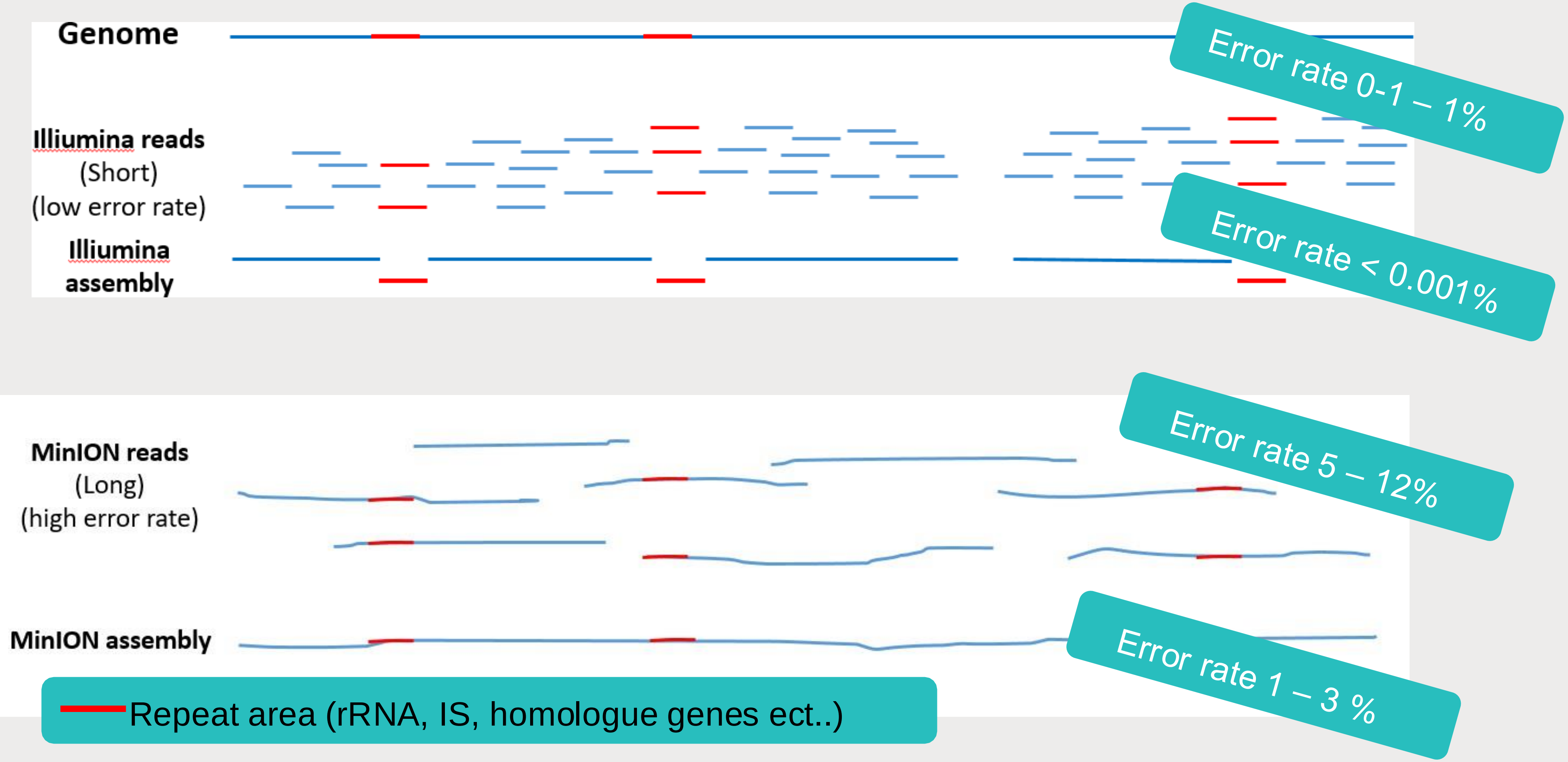
Relatively..

- Low-to medium price per isolate
- experimental technology
- low precision (high error rate)?
- fast

..and reads available in real-time

Tools for outbreak detection emerging

Illumina vs. MinION (R9.4.1) data



Error rate 0-1 – 1%

Error rate < 0.001%

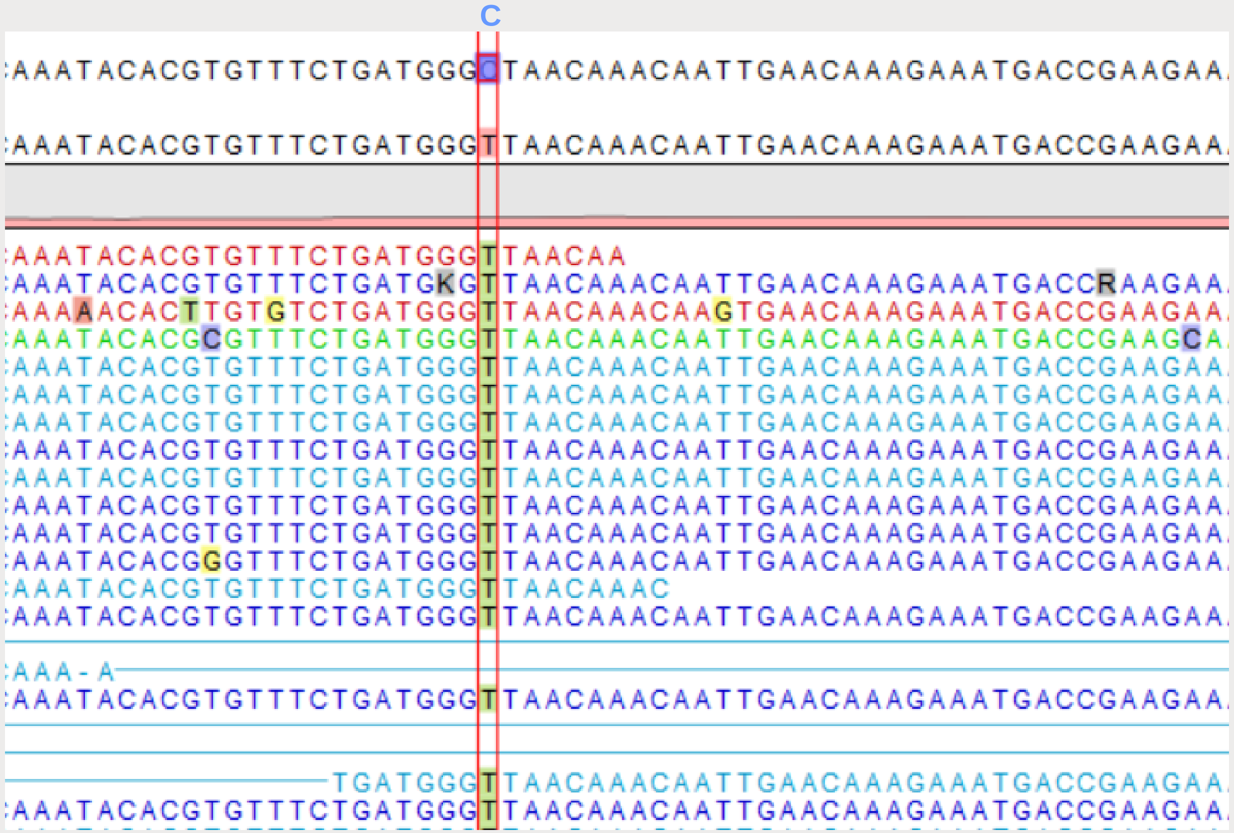
Error rate 5 – 12%

Error rate 1 – 3 %

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

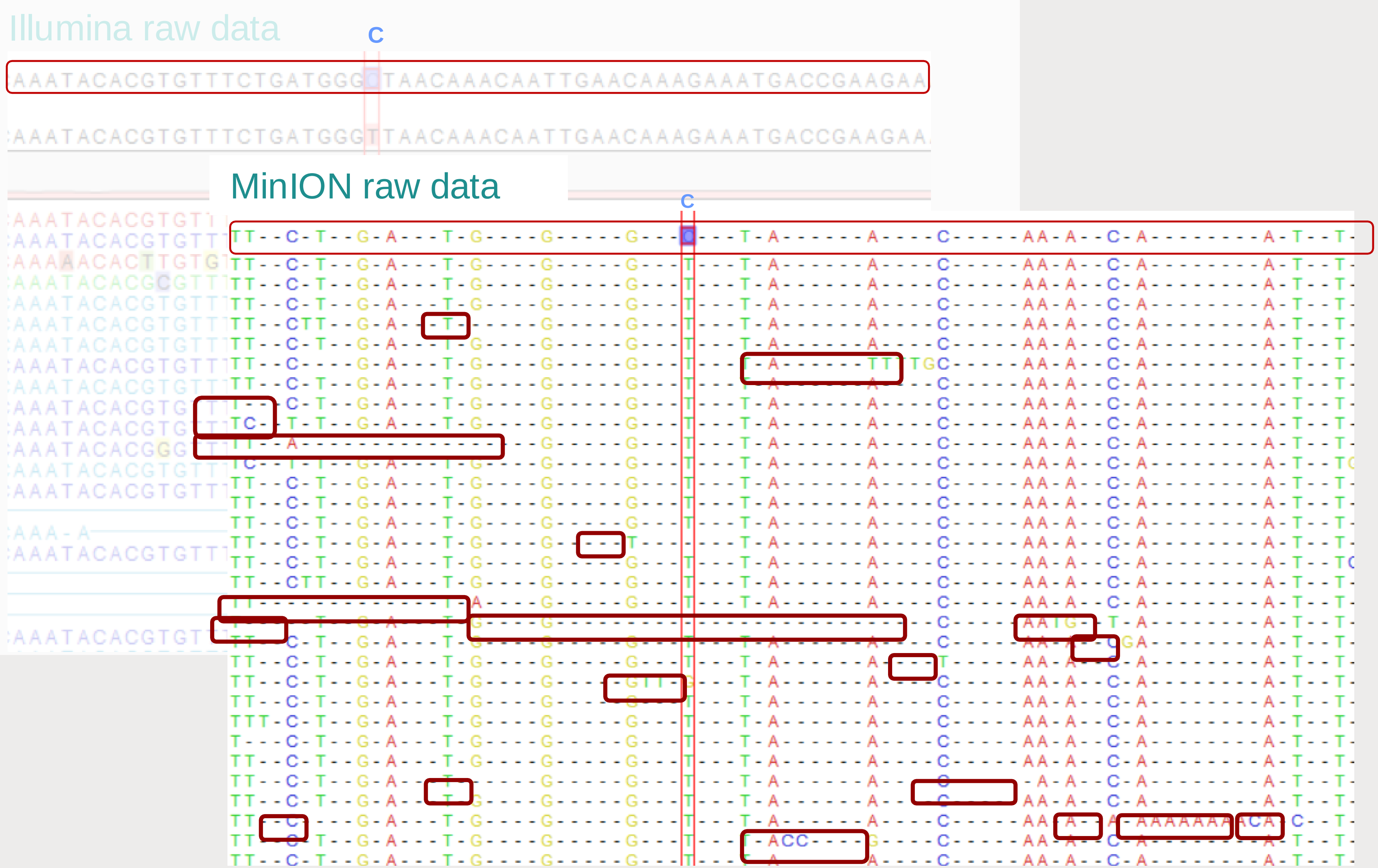
Illumina vs. MinION data

Illumina raw data

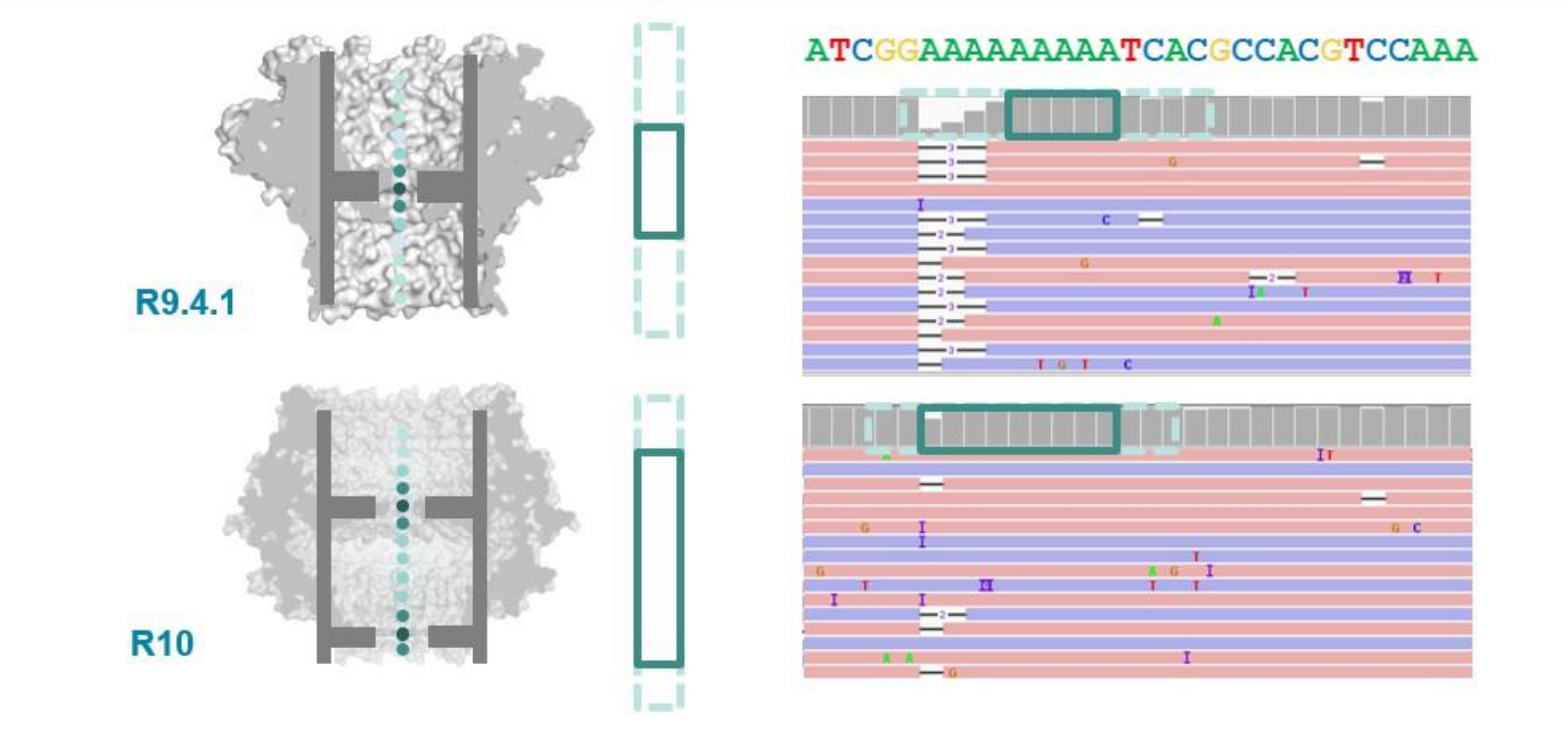


```
AAATACACGTGTTTCTGATGGGCTAACAAACAATTGAACAAAGAAATGACCGAAGAA

AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA

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

AAA-A
AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA

               TGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA
AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA
```

Illumina vs. MinION data


Illumina raw data
C
AAATACACGTGTTTCTGATGGGCTAACAAACAATTGAACAAAGAAATGACCGAAGAA
AAATACACGTGTTTCTGATGGGTTAACAAACAATTGAACAAAGAAATGACCGAAGAA
MinION raw data
C

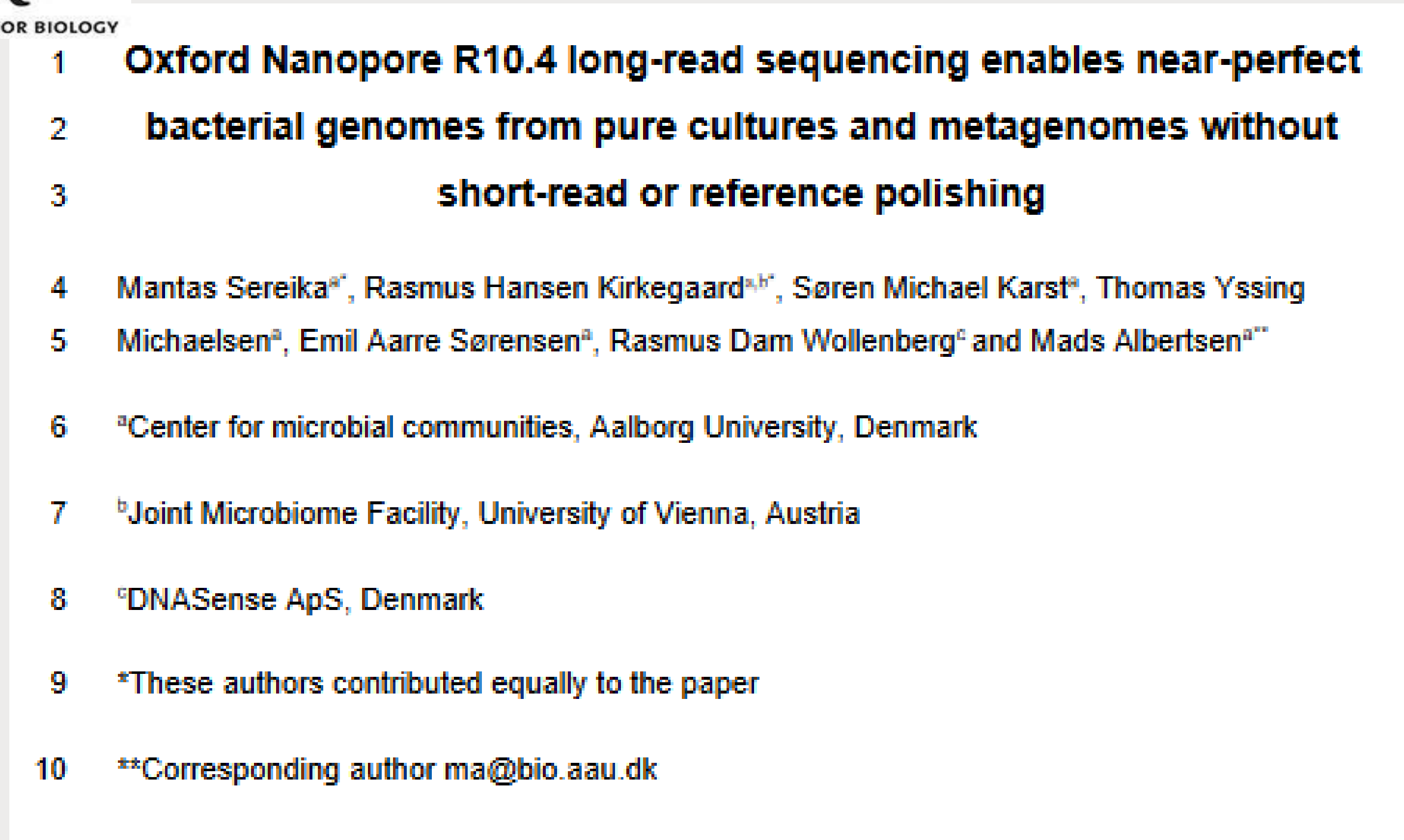
R9.4.1 vs. R10.4.1 pore



Choice of flowcell/pore

CSH Cold Spring Harbor Laboratory

bioRxiv
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**]

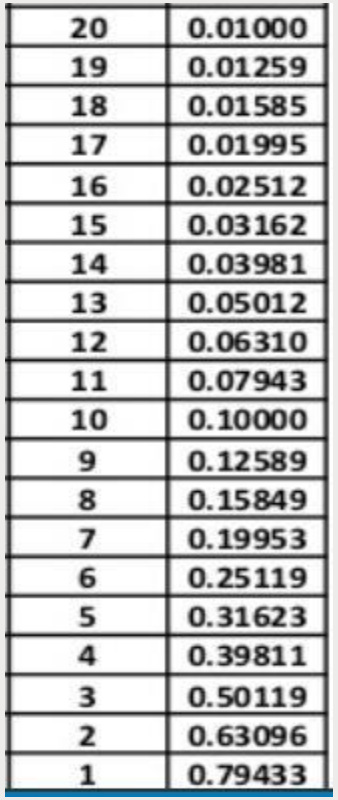
[a]Center for microbial communities, Aalborg University, Denmark

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

[c]DNASense ApS, Denmark

*These authors contributed equally to the paper

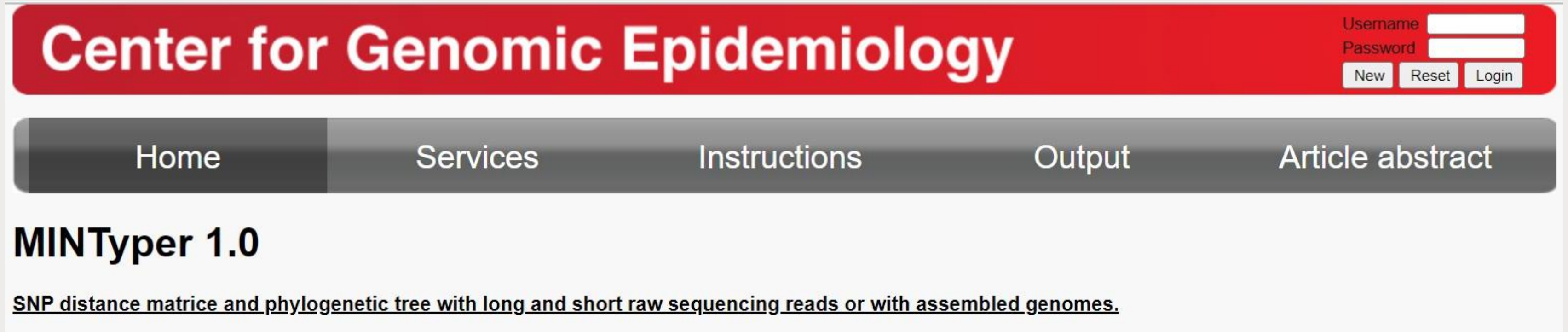
**Corresponding author ma@bio.aau.dk



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



Center for Genomic Epidemiology

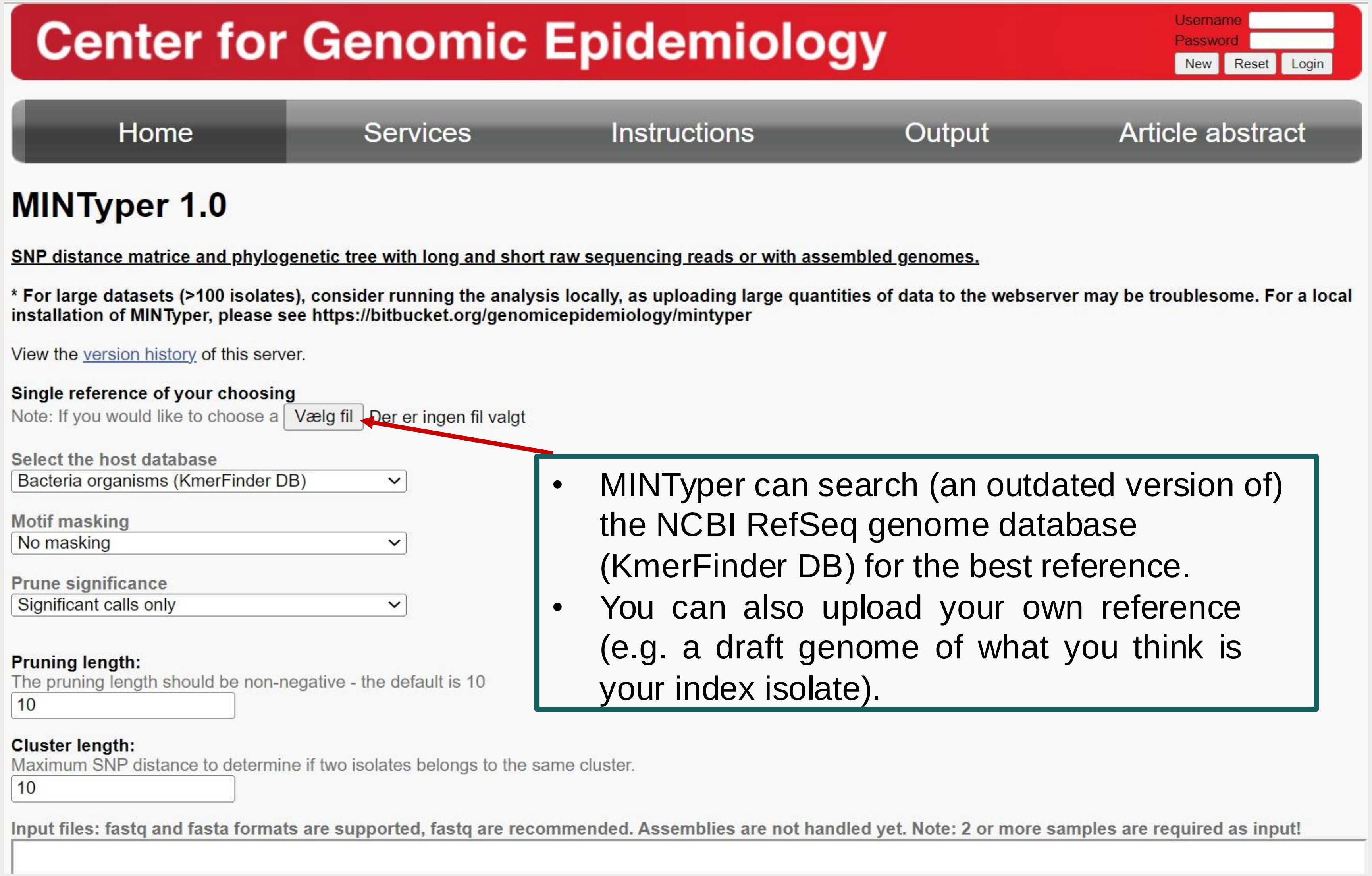
MINTyper 1.0

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

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

The Fleming Fund
Regional Grants

MINTyper V1.0



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

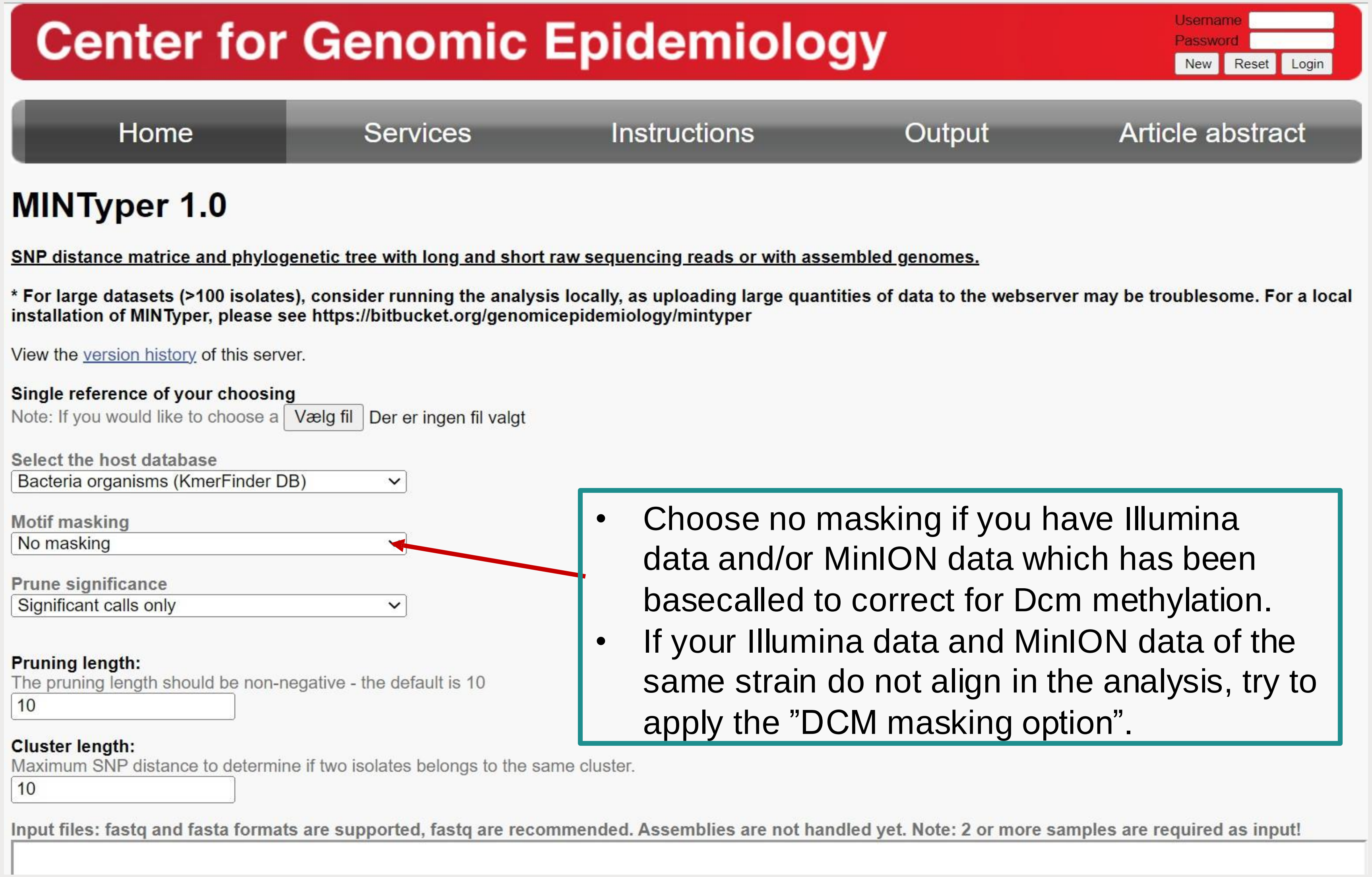
Pruning length:
The pruning length should be non-negative - the default is 10
10

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 (an 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



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

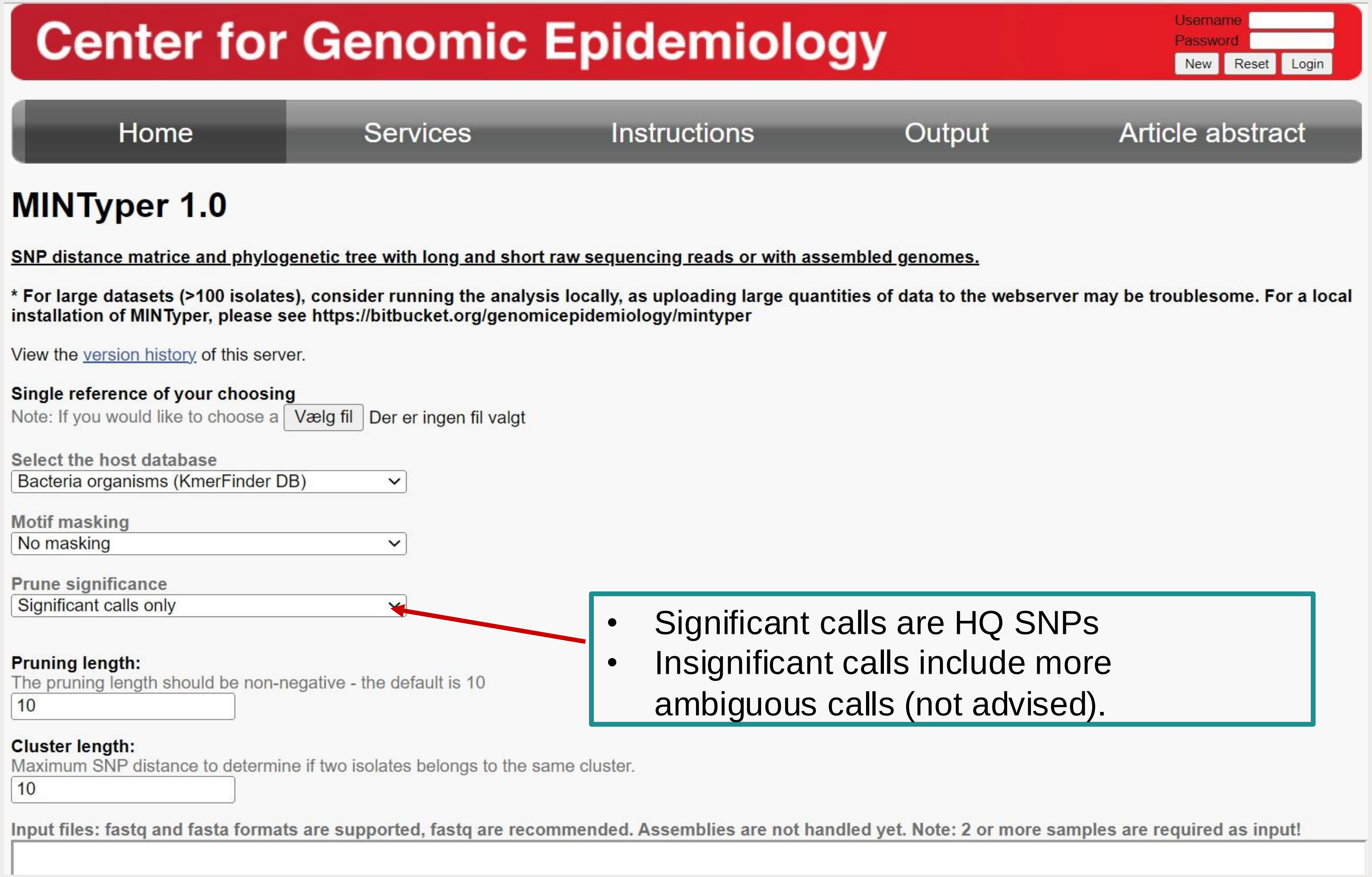
Pruning length:
The pruning length should be non-negative - the default is 10
10

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 do not align in the analysis, try to apply the ”DCM masking option”.

MINTyper V1.0



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

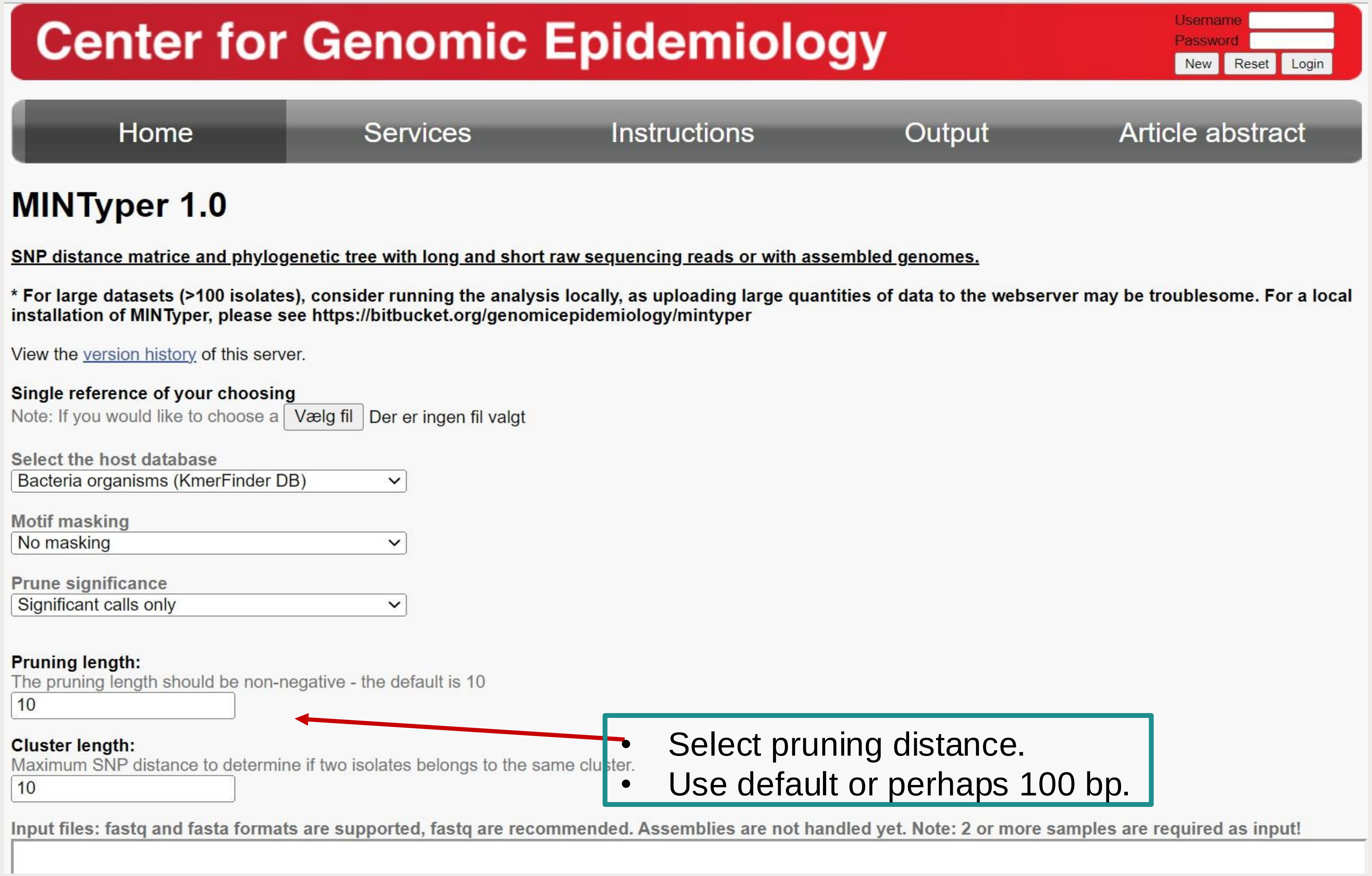
Pruning length:
The pruning length should be non-negative - the default is 10
10

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!

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

MINTyper V1.0



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

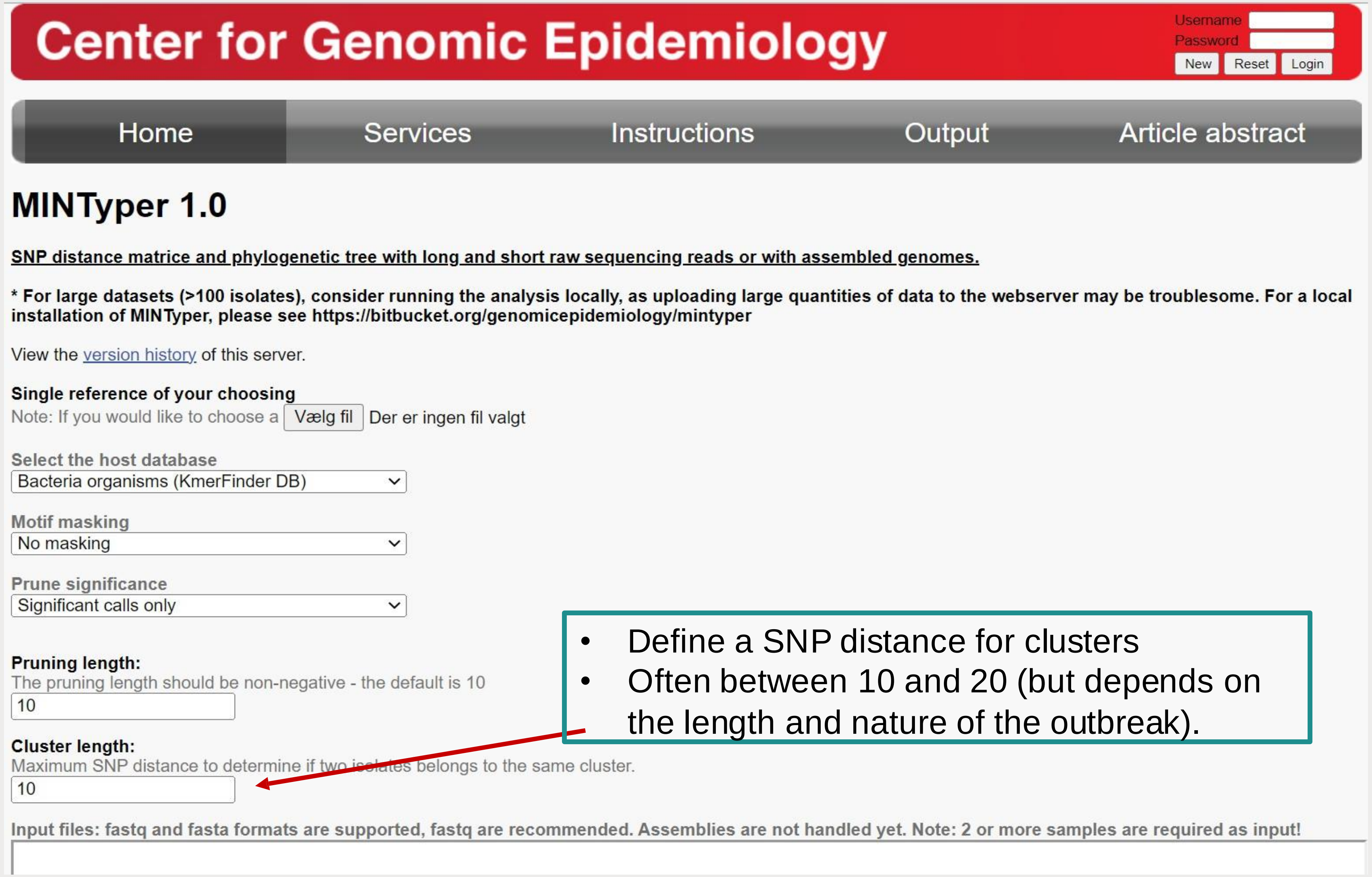
Pruning length:
The pruning length should be non-negative - the default is 10
10

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!

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

MINTyper V1.0



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

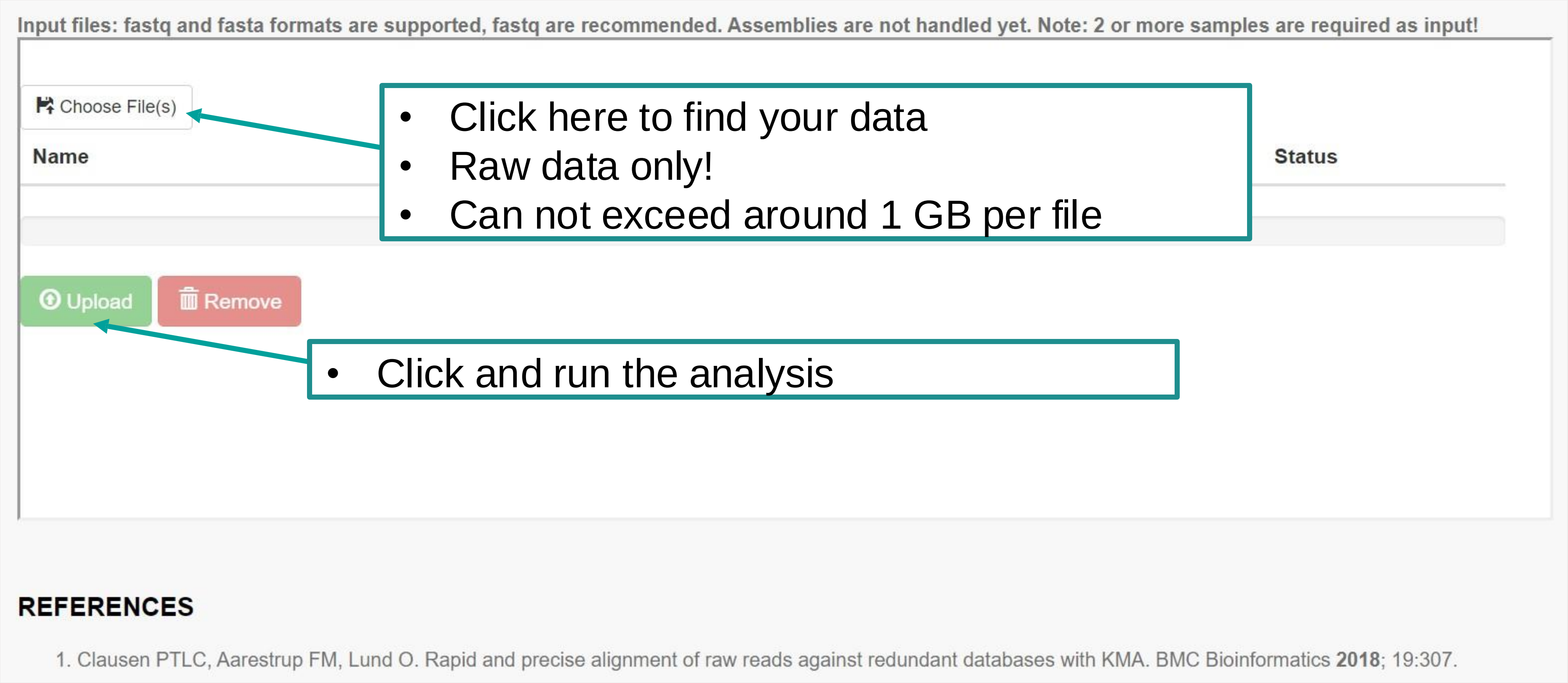
Pruning length:
The pruning length should be non-negative - the default is 10
10

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



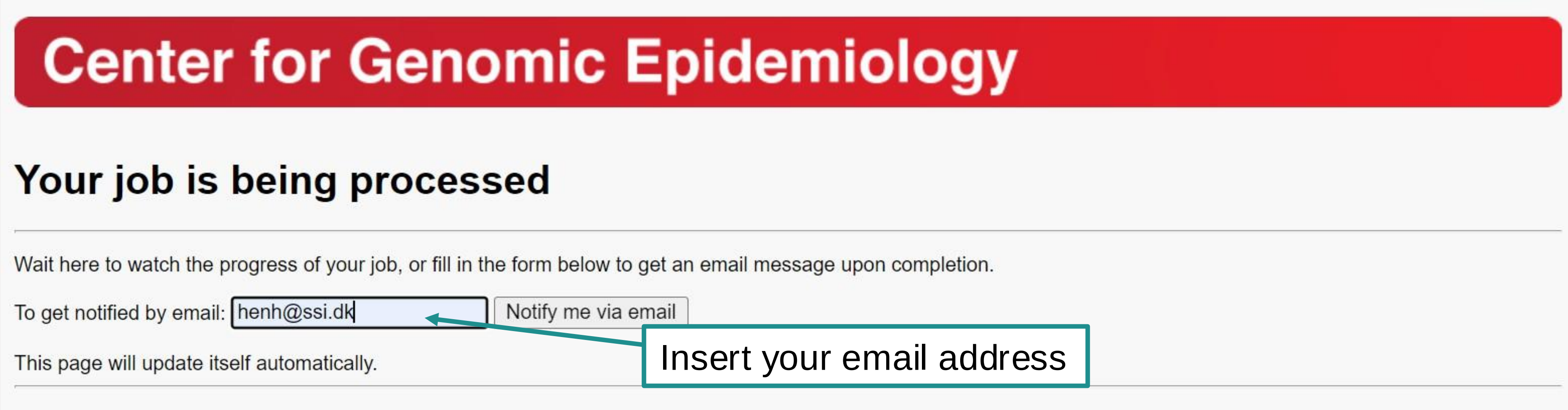
Input files: fastq and fasta formats are supported, fastq are recommended. Assemblies are not handled yet. Note: 2 or more samples are required as input!

- 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

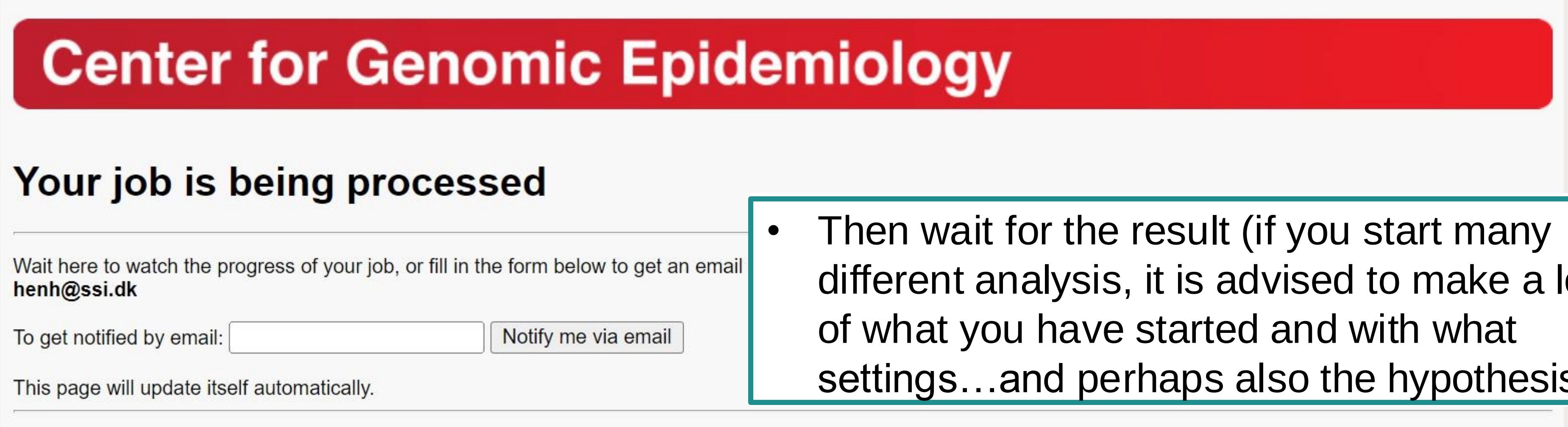
Your job is being processed

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

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
henh@ssi.dk

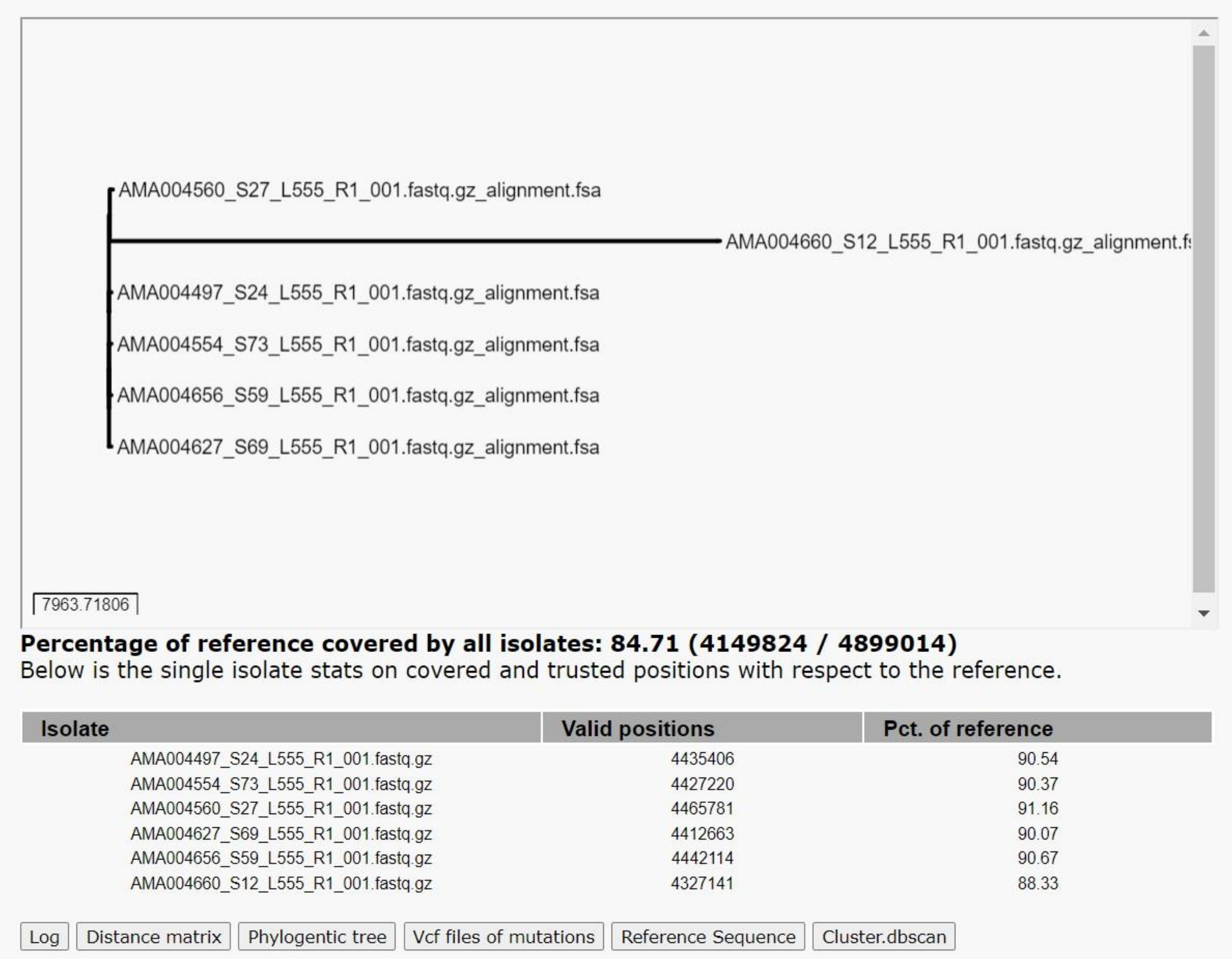
To get notified by email:

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).

The Fleming Fund
Regional Grants

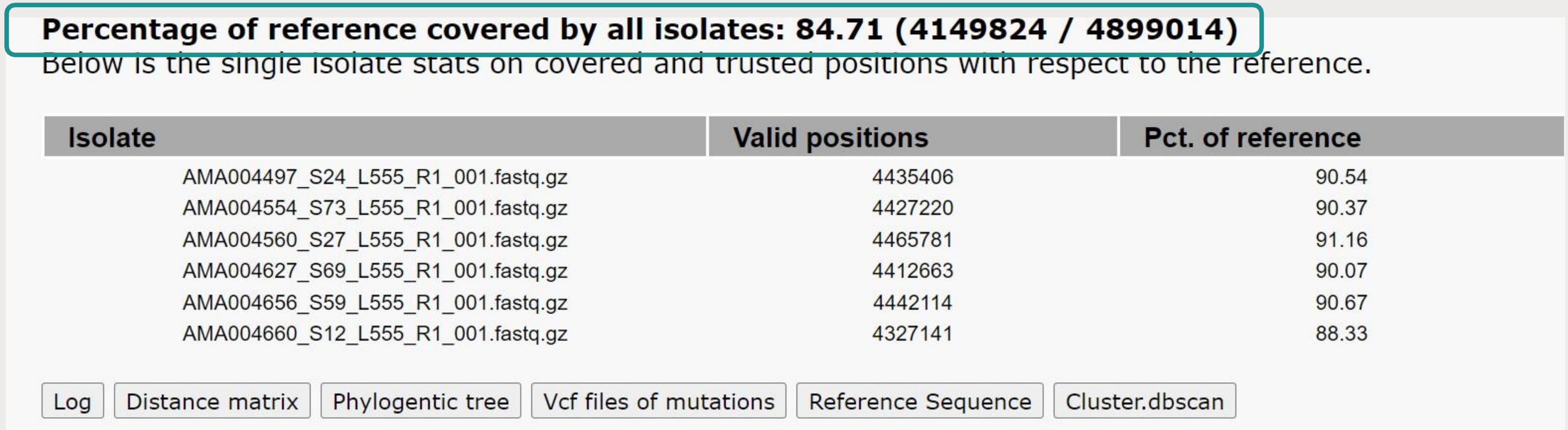
MINTyper output


AMA004560_S27_L555_R1_001.fastq.gz_alignment.fsa
AMA004660_S12_L555_R1_001.fastq.gz_alignment.f
AMA004497_S24_L555_R1_001.fastq.gz_alignment.fsa
AMA004554_S73_L555_R1_001.fastq.gz_alignment.fsa
AMA004656_S59_L555_R1_001.fastq.gz_alignment.fsa
AMA004627_S69_L555_R1_001.fastq.gz_alignment.fsa
7963.71806

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

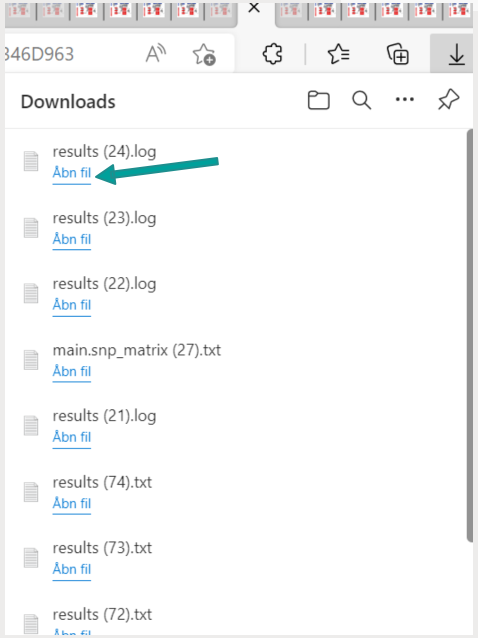
MINTyper output



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



results (24).log
Åbn fil

results (23).log
Åbn fil

results (22).log
Åbn fil

main.snp_matrix (27).txt
Åbn fil

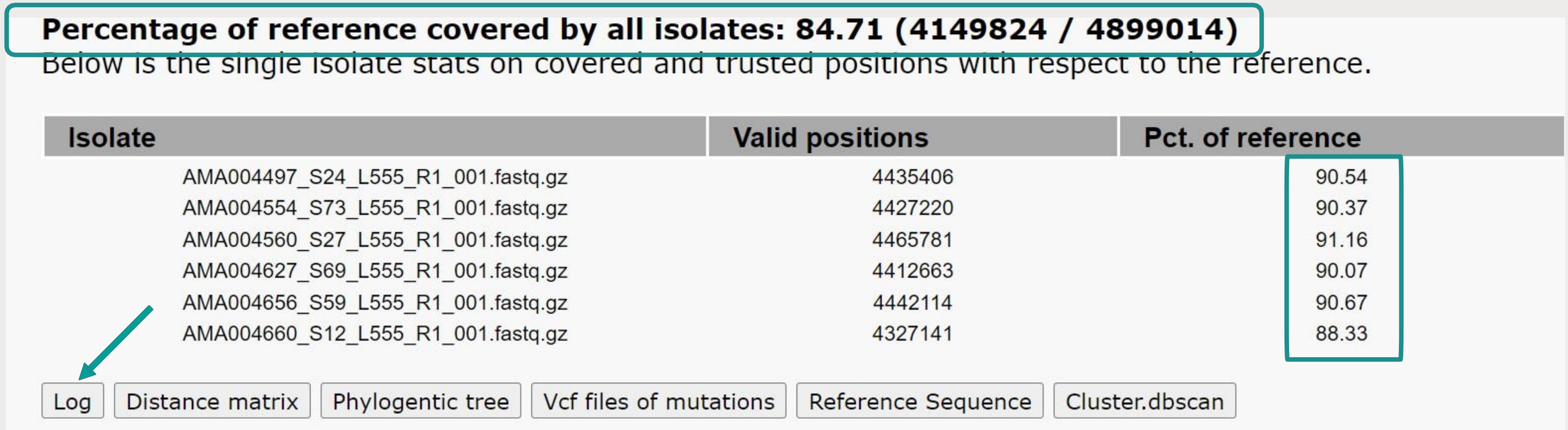
results (21).log
Åbn fil

results (74).txt
Åbn fil

results (73).txt
Åbn fil

results (72).txt

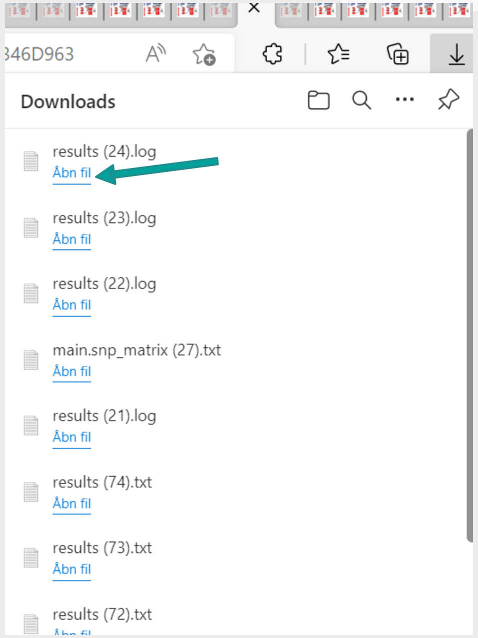
MINTyper output



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



results (24).log
Åbn fil

results (23).log
Åbn fil

results (22).log
Åbn fil

main.snp_matrix (27).txt
Åbn fil

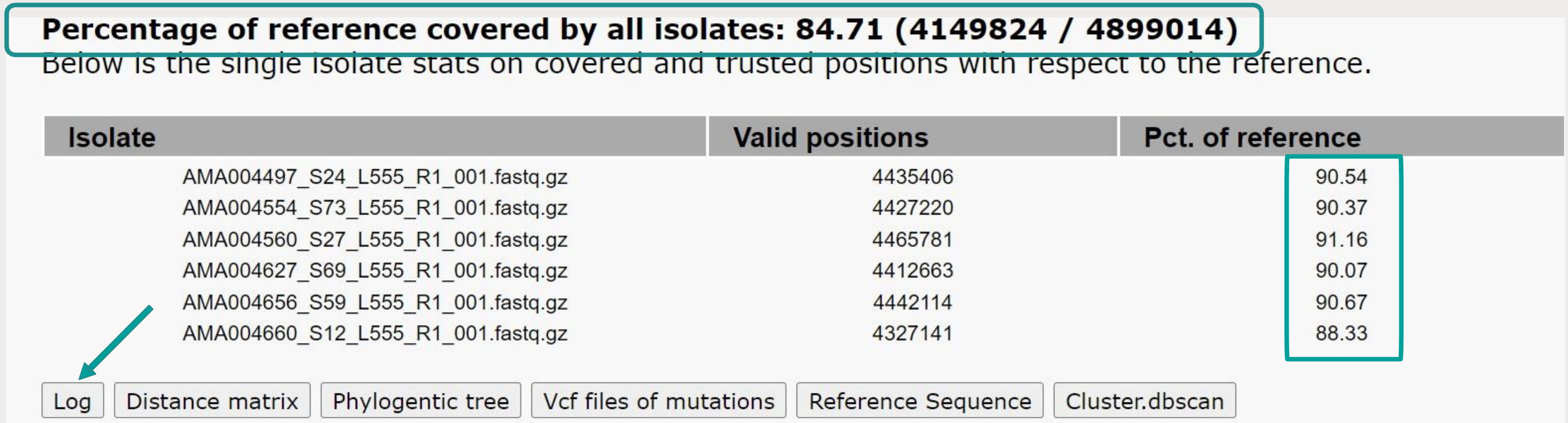
results (21).log
Åbn fil

results (74).txt
Åbn fil

results (73).txt
Åbn fil

results (72).txt

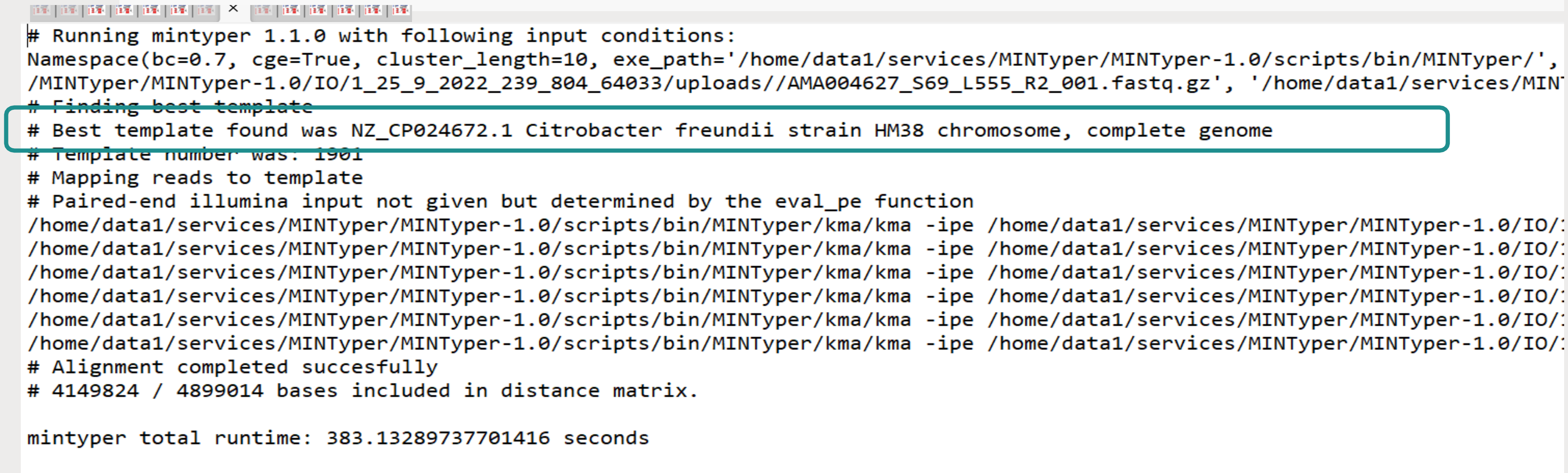
MINTyper output



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

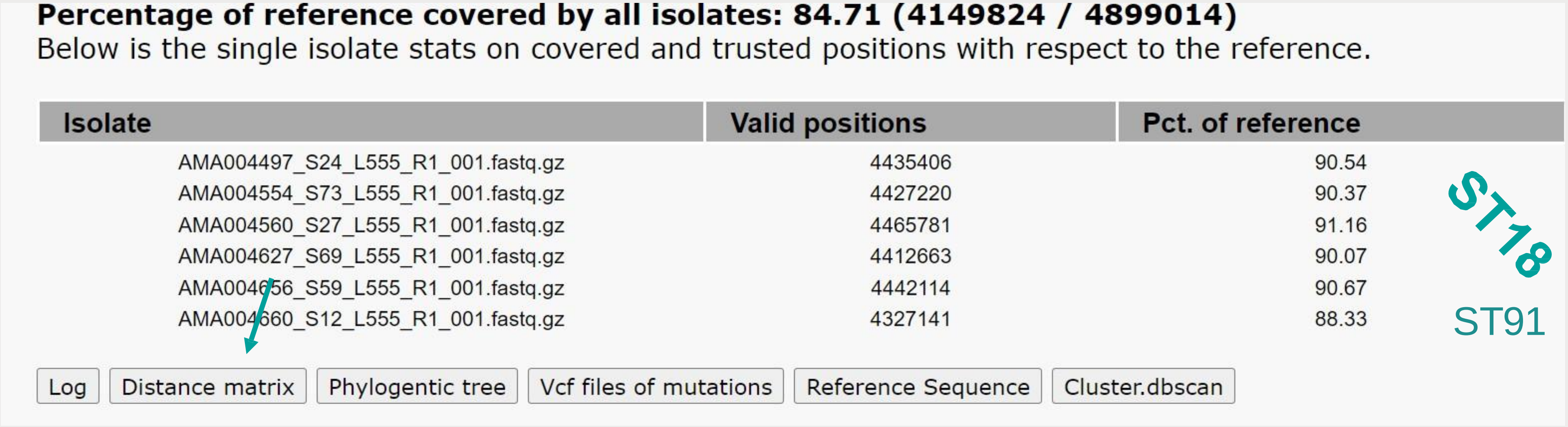
Below is the single isolate stats on covered and trusted positions with respect to the reference.

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

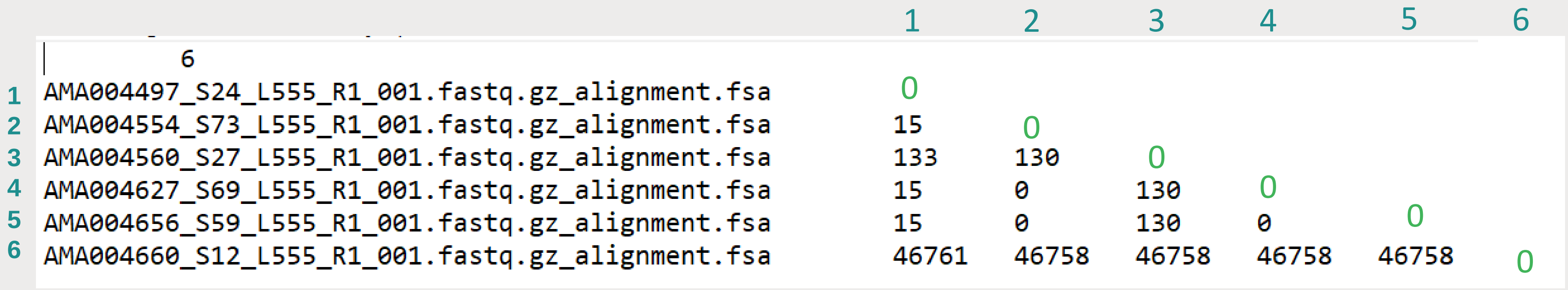


Percentage of reference covered by all isolates: 84.71 (4149824 / 4899014)
Below is the single isolate stats on covered and trusted positions with respect to the reference.

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

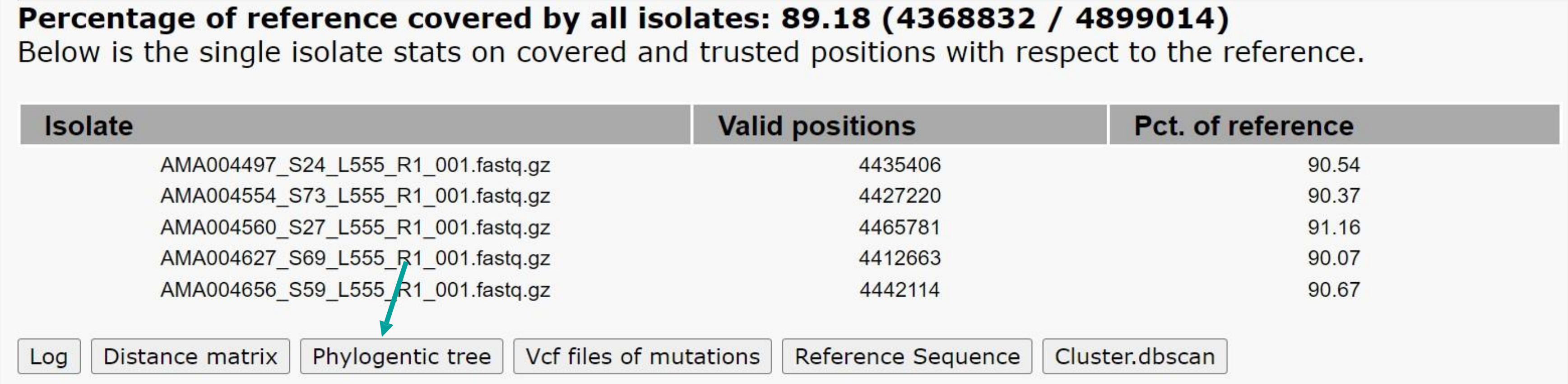
ST18

ST91



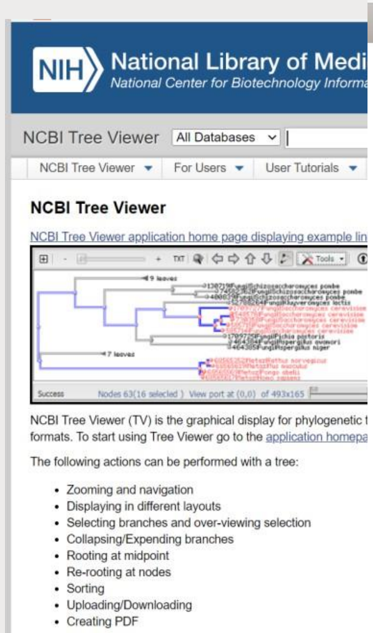
```
                                                      1      2      3      4      5      6
        6
1 AMA004497_S24_L555_R1_001.fastq.gz_alignment.fsa    0
2 AMA004554_S73_L555_R1_001.fastq.gz_alignment.fsa    15     0
3 AMA004560_S27_L555_R1_001.fastq.gz_alignment.fsa    133    130    0
4 AMA004627_S69_L555_R1_001.fastq.gz_alignment.fsa    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
```

MINTyper output - visualizations



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



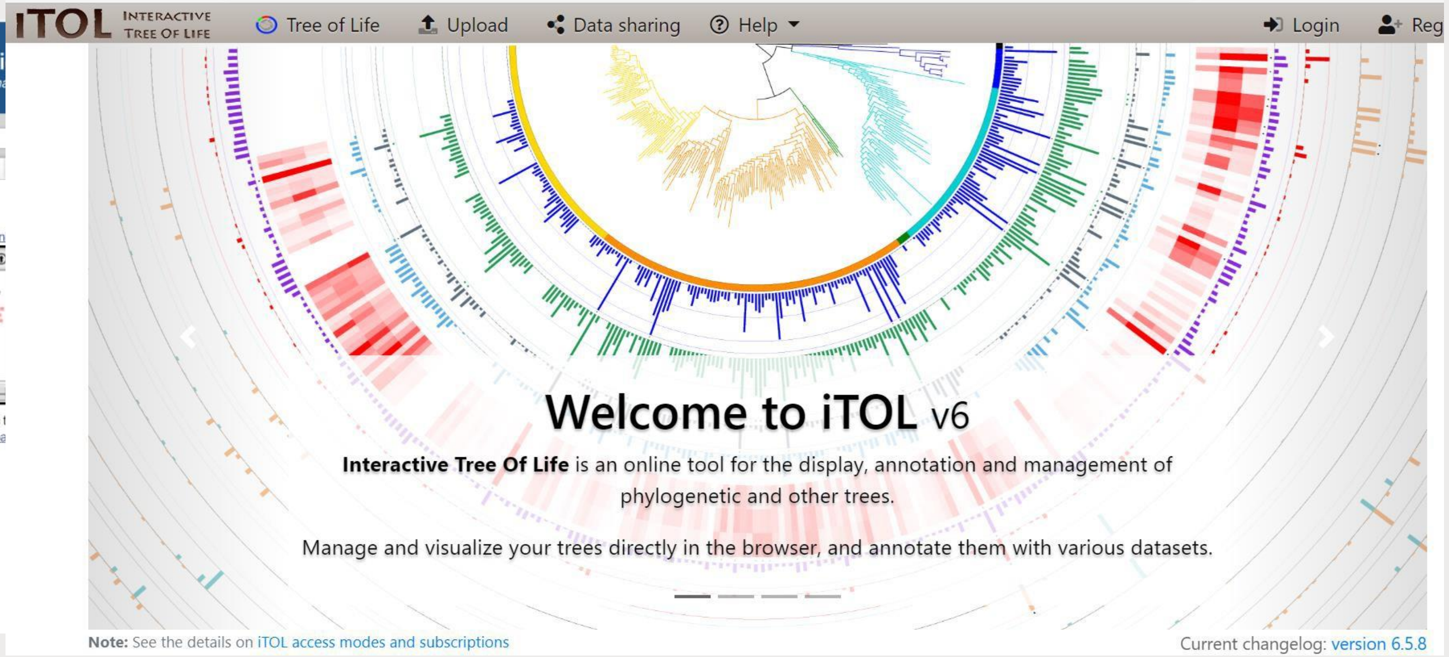
NCBI Tree Viewer

NCBI Tree Viewer application home page displaying example lin

NCBI Tree Viewer (TV) is the graphical display for phylogenetic t formats. To start using Tree Viewer go to the application homepa

The following actions can be performed with a tree:

- Zooming and navigation
- Displaying in different layouts
- Selecting branches and over-viewing selection
- Collapsing/Expending branches
- Rooting at midpoint
- Re-rooting at nodes
- Sorting
- Uploading/Downloading
- Creating PDF



Welcome to iTOL v6

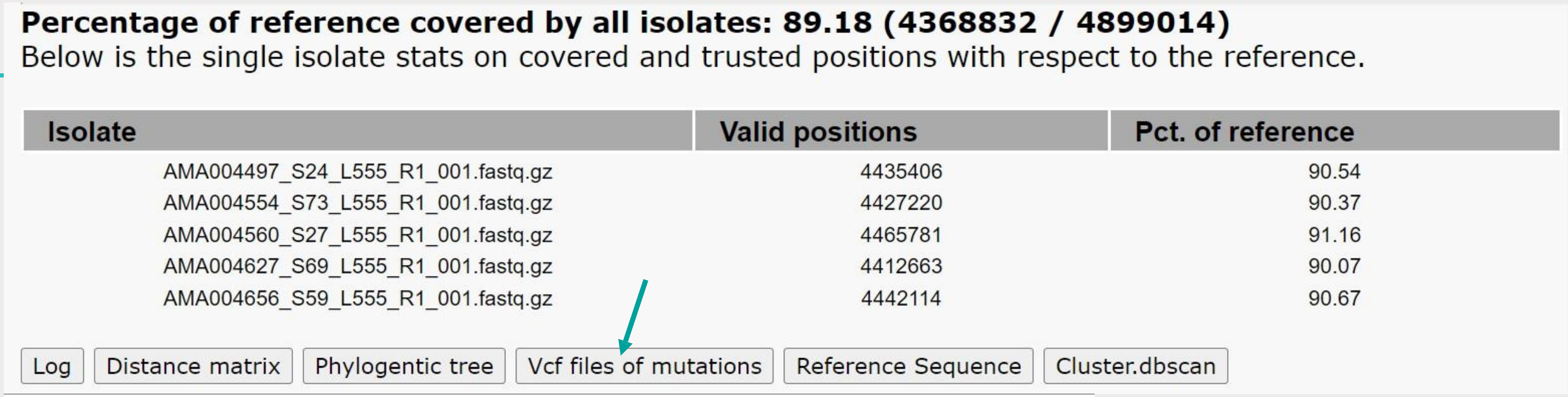
Interactive Tree Of Life is an online tool for the display, annotation and management of phylogenetic and other trees.

Manage and visualize your trees directly in the browser, and annotate them with various datasets.

Note: See the details on iTOL access modes and subscriptions

Current changelog: version 6.5.8

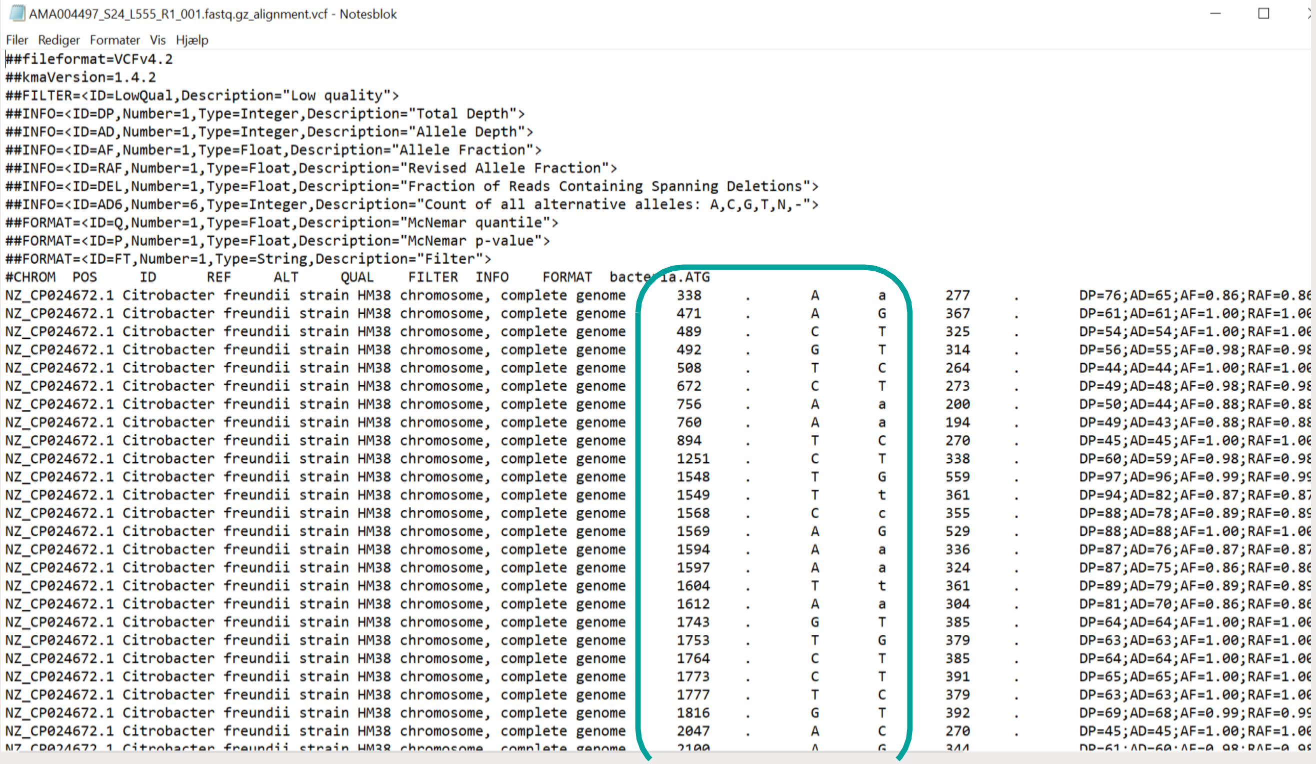
MINTyper output– VCF data



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.

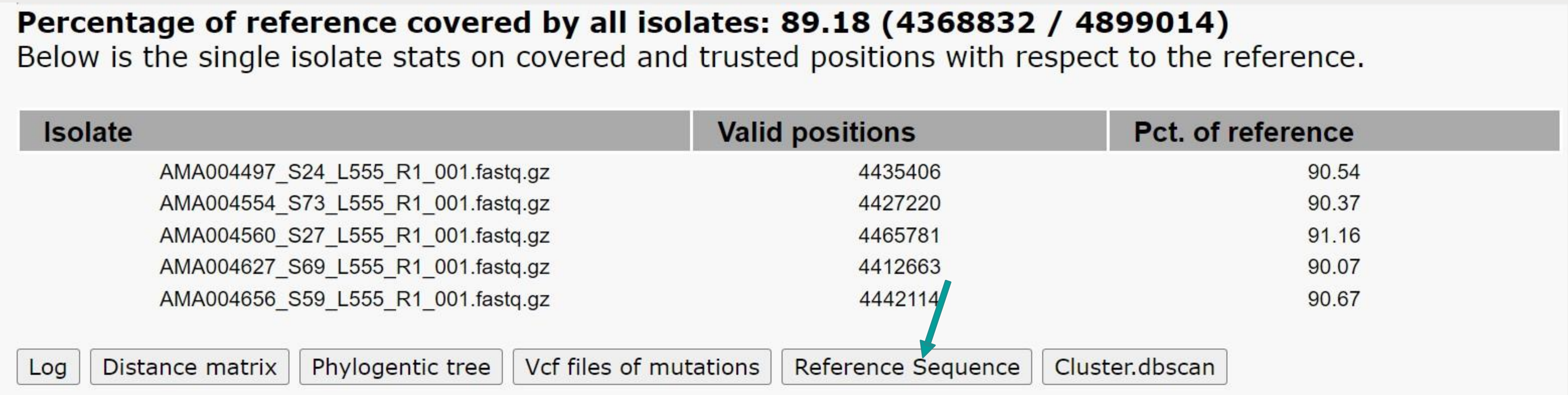
Isolate	Valid positions	Pct. of reference
AMA004497_S24_L555_R1_001.fastq.gz	4435406	90.54
AMA004554_S73_L555_R1_001.fastq.gz	4427220	90.37
AMA004560_S27_L555_R1_001.fastq.gz	4465781	91.16
AMA004627_S69_L555_R1_001.fastq.gz	4412663	90.07
AMA004656_S59_L555_R1_001.fastq.gz	4442114	90.67



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
```

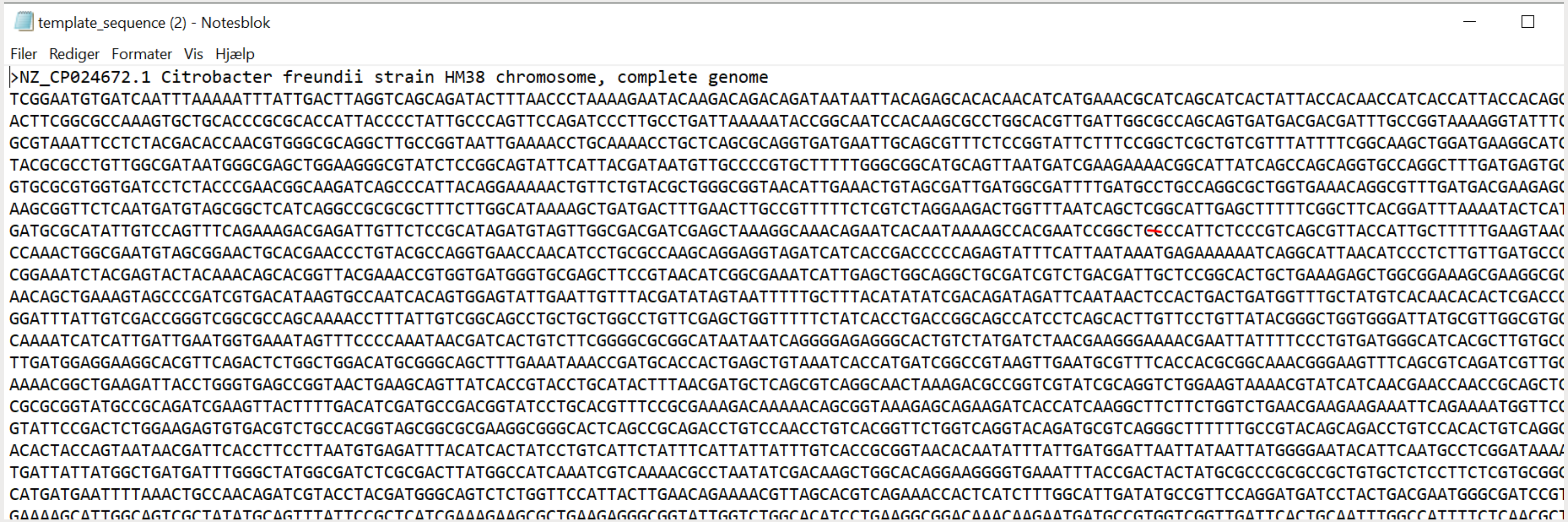
MINTyper output– reference



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



template_sequence (2) - Notesblok
Filer Rediger Formater Vis Hjælp

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

The Fleming Fund
Regional Grants

Let’s take a break ☺

The Fleming Fund
Regional Grants

Scenario

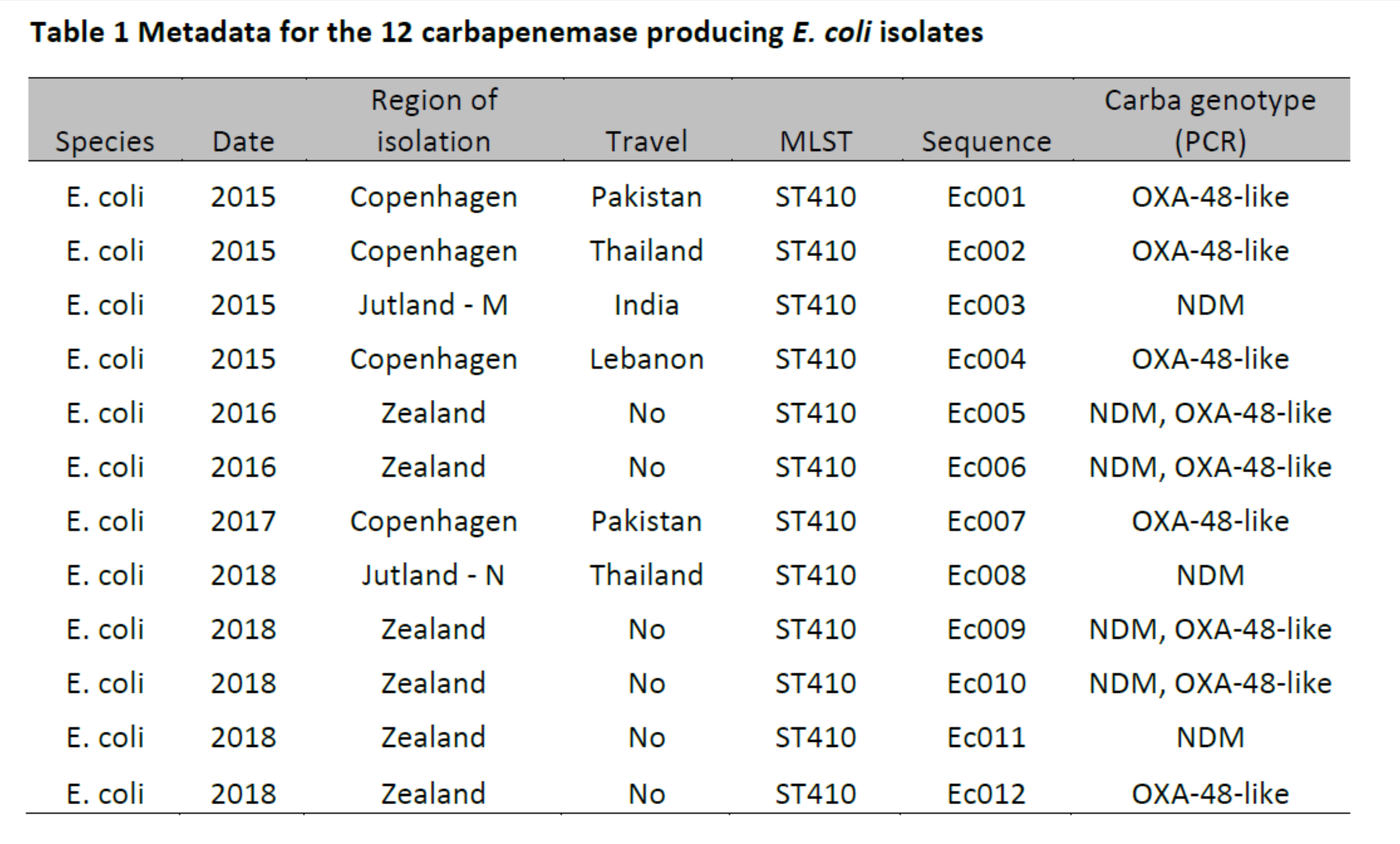


Table 1 Metadata for the 12 carbapenemase producing *E. coli* isolates

Species	Date	Region of isolation	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

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.

The Fleming Fund
Regional Grants

Thank you

UKaid from the British people

MOTT MACDONALD

DTU

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.