EQAsia – with Lauge Holm Sørensen

Exercises in WGS analysis and the CGE tools

Contents

- Whole genome sequencing
  - What is whole genome sequencing
  - sequencing technologies
  - strengths and weaknesses
- Next generation sequencing – Illumina platforms
  - Library preparation
  - Read processing
  - Assembly
  - Quality control
- Genomic analysis
  - Species verification and typing
  - Antimicrobial resistance
  - Plasmids
  - Phylogeny
- Introduction to exercises

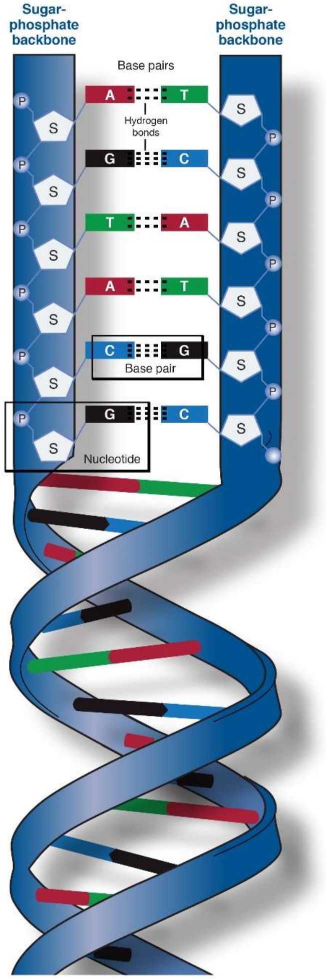
DNA Sequencing

- The DNA encodes all genetic information needed for a cell to survive and prosper

- DNA consists of two strands of sugar-phosphate backbones, each residue (called a nucleotide) containing one of four bases
  - Adenine (A)
  - Tyrosine (T)
  - Guanine (G)
  - Cytosine (C)

- The two strands are complementary, with each nucleotide base pairing with a specific complementary base o nthe opposite strand, A with T, G with C

- Sequencing is the process of reading a stretch of DNA, reproducing the ordered combination of its constituent residues



Public domain image, courtesy: National Human Genome Research Institute, [National Human Genome Research Institute Home | NHGRI]

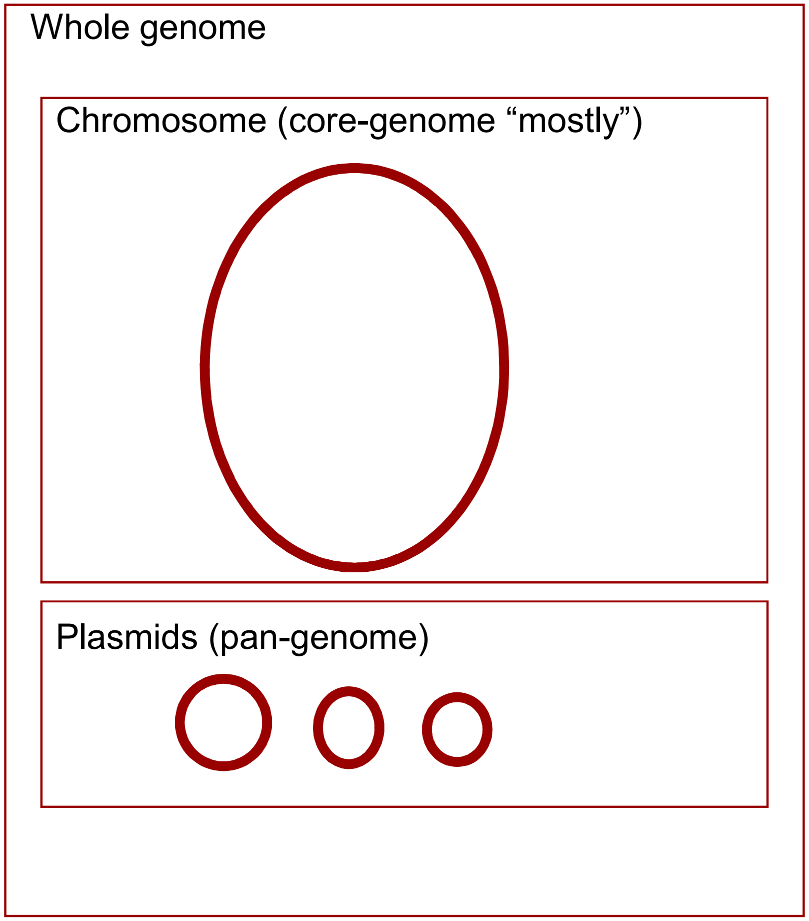
Whole Genome Sequencing

- In bacteria DNA is ordered into circular molecules
  - Large DNA molecules are classified as chromosomes
  - Smaller DNA molecules are classified as plasmids

- Most bacteria contains a single chromosome, which encodes all the most necessary genes for survival, these are referred to as “core-genes” or “housekeeping genes”

- The cell also contains DNA coding for genes not necessary for survival, these are called “pan-genes” and can be found in the chromosome or plasmids

- The genome of a bacteria refers to all chromosomes + all plasmids



Whole genome

Chromosome (core-genome “mostly”)

Plasmids (pan-genome)

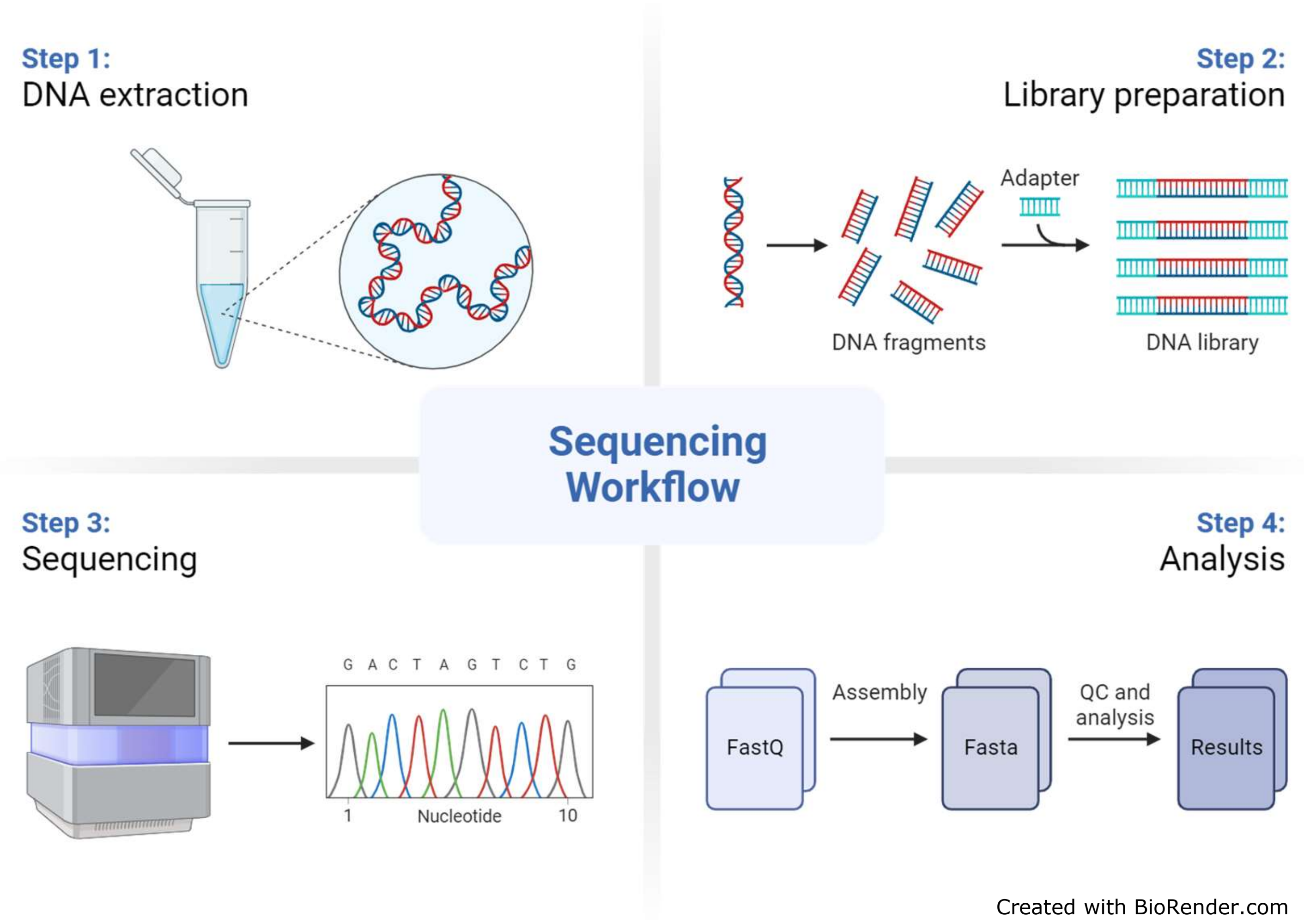
Overview

1) DNA is extracted from a pure culture of a bacterial isolate

2) DNA is fragmented to smaller pieces and adapters are attached

3) DNA library is loaded to sequencing platform and the sequence of nucleotides in each fragment determined

4) The machine outputs results as a fastQ file and analysis is conducted



Step 1: DNA extraction

Step 2: Library preparation

Sequencing Workflow

Step 3: Sequencing

Step 4: Analysis

Created with BioRender.com

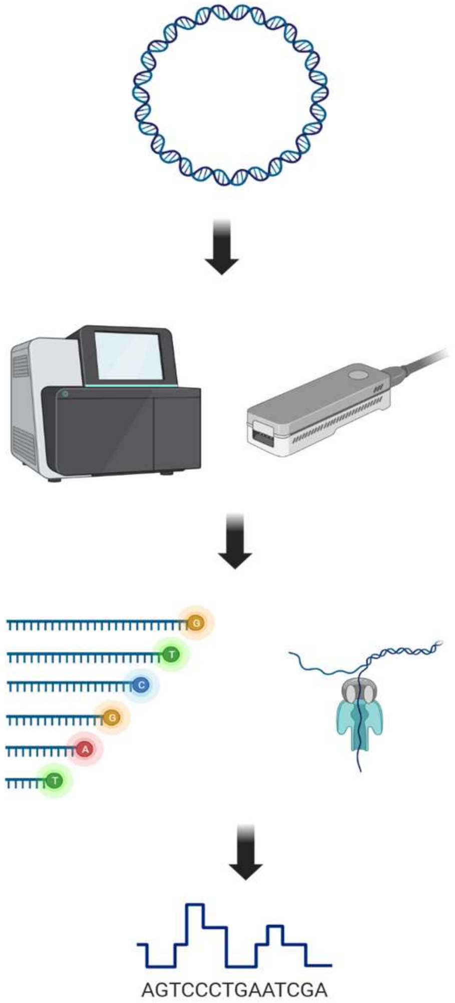
Sequencing technologies

- Different technologies have been developed for genome sequencing, currently the Illumina next generation sequencing platforms are the most used in surveillance (Segerman, 2020)

- 3rd generation sequencing platforms are seeing wider usage, mainly due to the Oxford Nanopore MinIon sequencers smaller size and affordability.

- 3rd generation sequencers (Nanopore, PacBio) are able to read longer stretches of DNA, but are generally more prone to error and costly compared to the 2nd generation

- In particular, mobile genetic elements and structural variation is simpler to find with 3rd generation sequencing



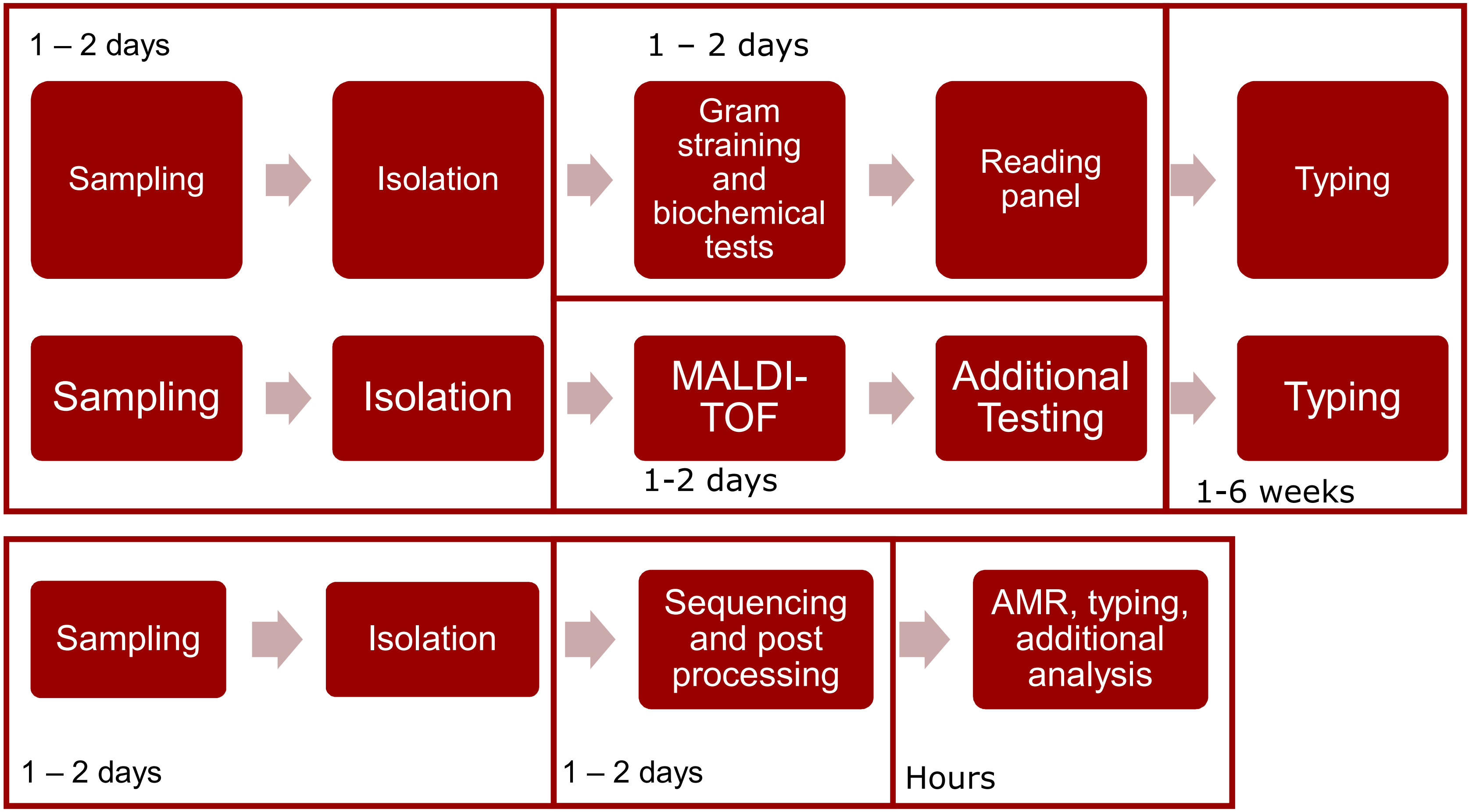
Created with BioRender.com

Segerman B. The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases. Front Cell Infect Microbiol. 2020 Oct 19;10:527102. doi: 10.3389/fcimb.2020.527102. PMID: 33194784; PMCID: PMC7604302.

Overview timeframe

Rapid biochemical methods

Whole genome sequencing



1 – 2 days

Sampling

Isolation

1 – 2 days

Gram straining and biochemical tests

Reading panel

Typing

Sampling

Isolation

MALDI-TOF

Additional Testing

1-2 days

Typing

1-6 weeks

Sampling

Isolation

1 – 2 days

Sequencing and post processing

1 – 2 days

AMR, typing, additional analysis

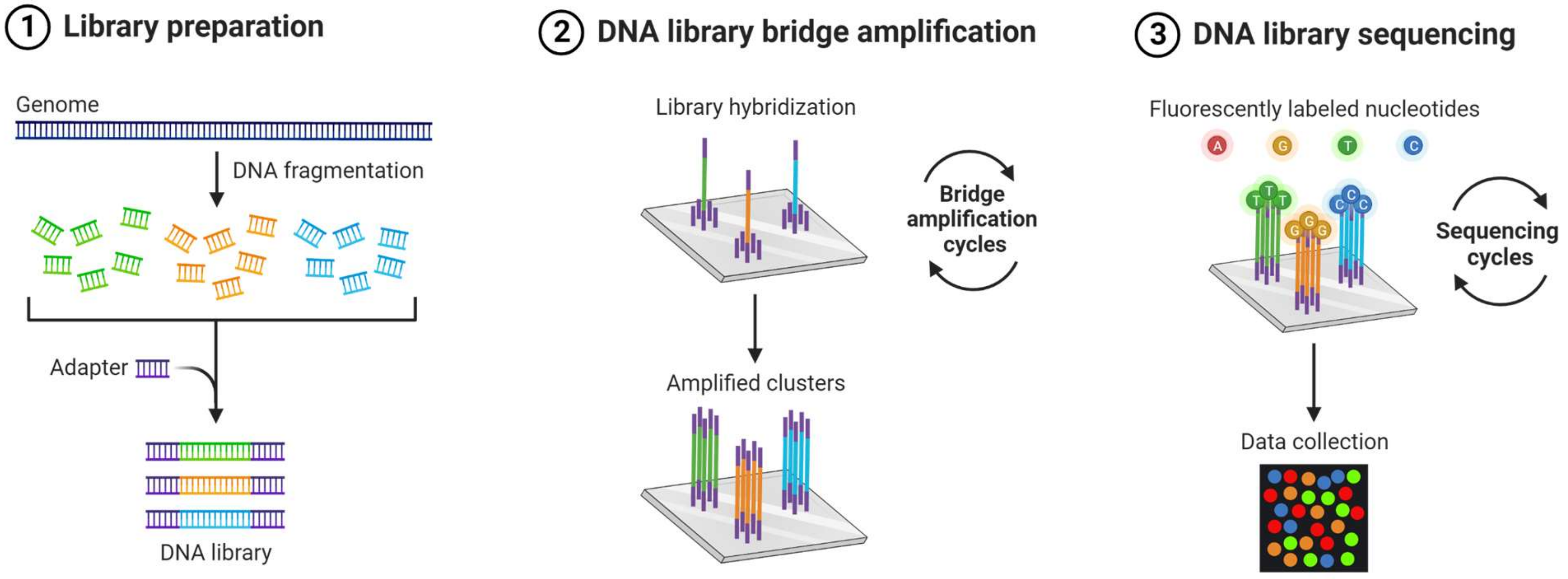
Hours

Strengths and weaknesses

Pros	Cons
Captures a lot of information: We aim to capture all the genetic information of the isolate	Storage: large amounts of data requires large hard drives
Additional analysis is easy to conduct, including in future research	CPU power: Programs demand computing power
High resolution: We can estimate the phylogenetic relationship between strains at a very in-depth level	Costs: machines are expensive and so are reagents (possible less so with new long reads sequencing)
Relatively fast, Ferrer et al. 2014 found a 1% increase in mortality per hour treatment was delayed after sepsis	Previous knowledge: databases need a solid foundation of knowledge to be precise
Scalable: good if surveillance needs to be expanded	

Ferrer R, Martin-Loeches I, Phillips G, Osborn TM, Townsend S, Dellinger RP, Artigas A, Schorr C, Levy MM. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. Crit Care Med. 2014 Aug;42(8):1749-55. doi: 10.1097/CCM.0000000000000330. PMID: 24717459.

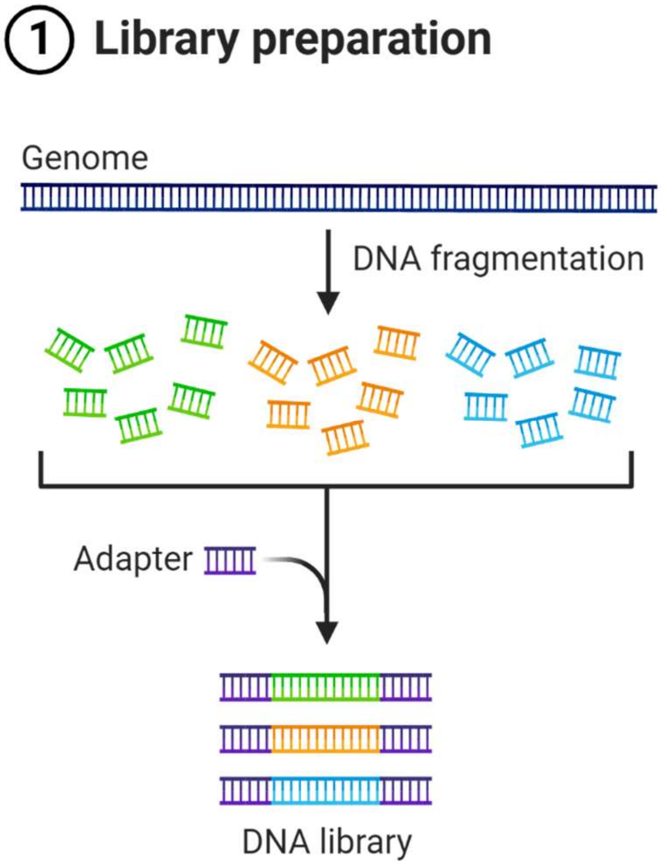
Overview of Illumina sequencing



Created with BioRender.com

Library prep

- After the pure culture have been grown, the cells are pelleted and the DNA extracted.
- The DNA is fragmented to produce smaller pieces suitable for NGS and adapters are ligated to fragments.
- Fragments are then selected by size to achieve a more homogenous library size.
- The adapters make the fragment able to bind to the flow cell in the subsequent sequencing.
- It also contains indexes for multiplexing libraries, making it possible to run multiple isolates at the same time.



Created with BioRender.com

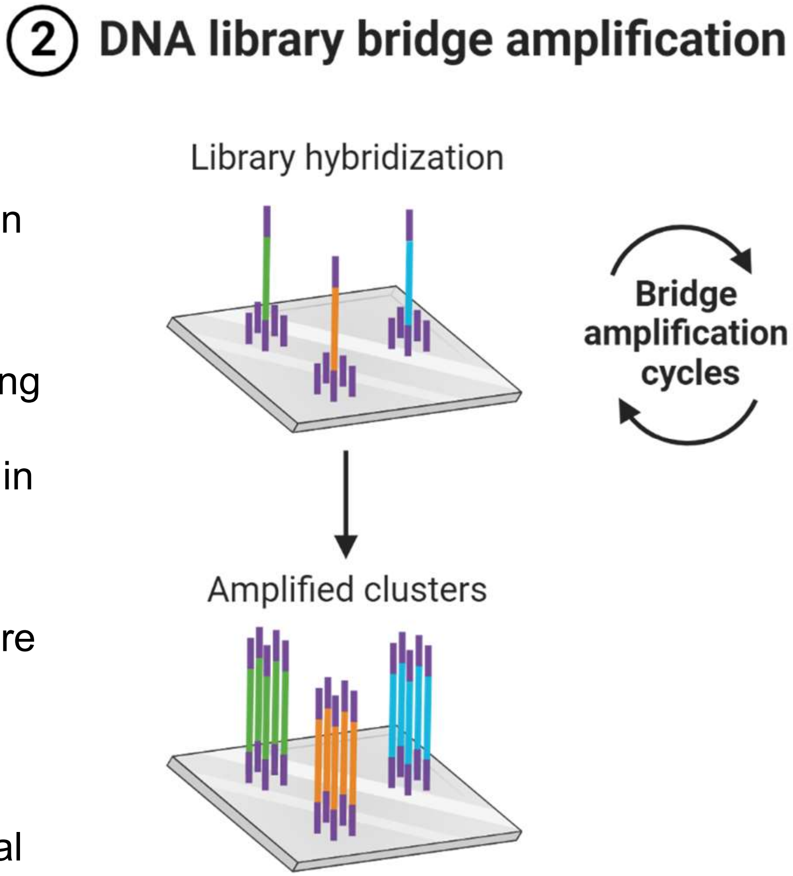
Initial amplification

- The library is loaded to the flow cell to be sequenced on the sequencing platform.

- The adapter adheres to a surface in the flow cell, binding the fragment. The concentration of the loaded DNA is important to leave sufficient space between fragments in this step.

- Each fragment is amplified, meaning identical copies are made in close proximity to original fragment, forming a cluster.

- This step is needed to amplify the signal from the actual sequencing.


2 DNA library bridge amplification
Library hybridization
Bridge amplification cycles
Amplified clusters

Created with BioRender.com

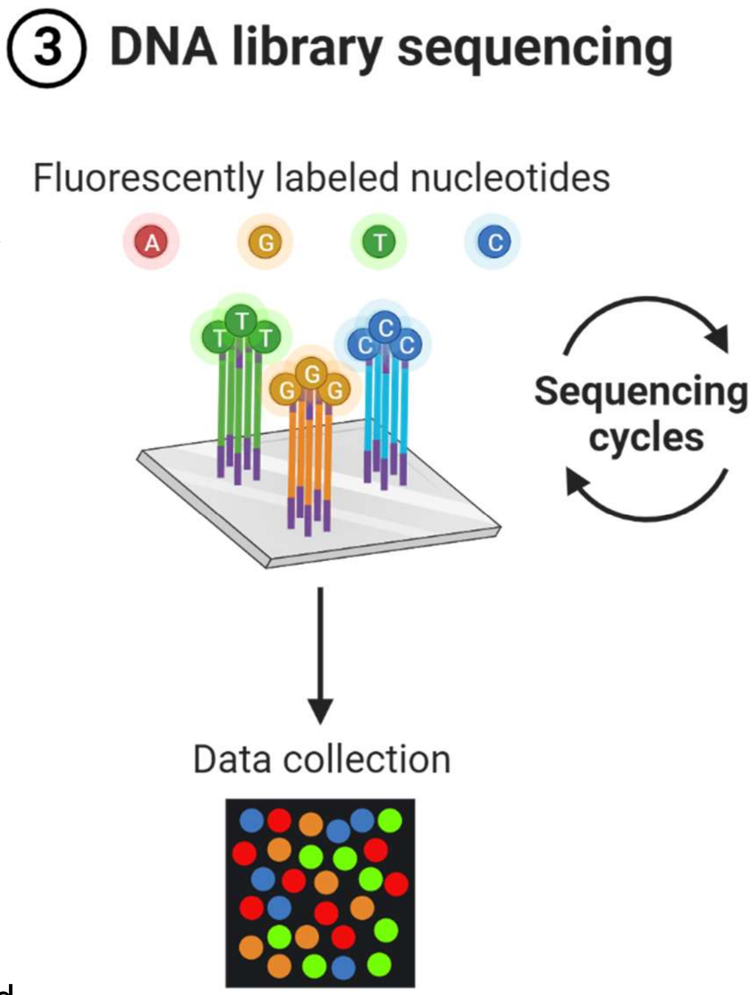
Sequencing

- The sequencing now begins, each fragment is copied in a stepwise manner, allowing only a single nucleotide to be added.

- Nucleotides are modified with fluorescent dyes which makes the reaction stop after the addition of a single nucleotide. Each nucleotide type (A,T,G,C) is label with a different fluorescent dye.

- After the addition of every modified nucleotide, the fluorescent dye is exited, which makes it emit a light of a color dependent on the nucleotide. The sequencing machine thus interprets the light as a specific nucleotide.

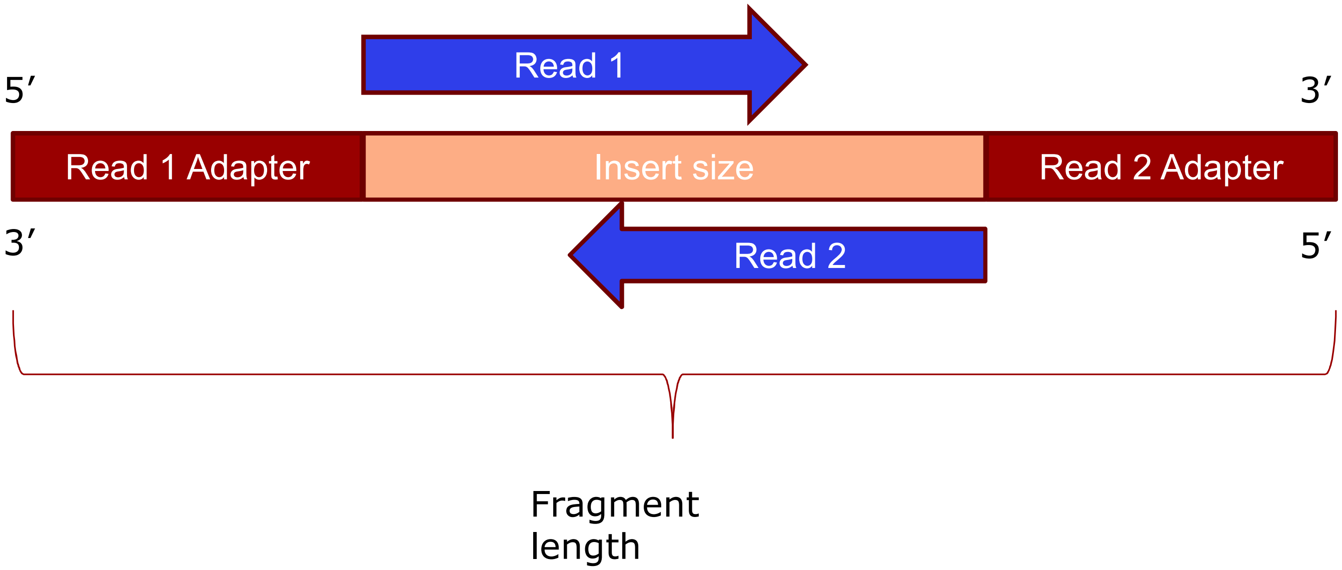
- The dye is then chemically cleaved from the modified nucleotide, which allows a new modified nucleotide to bind and a new round of sequencing can begin.



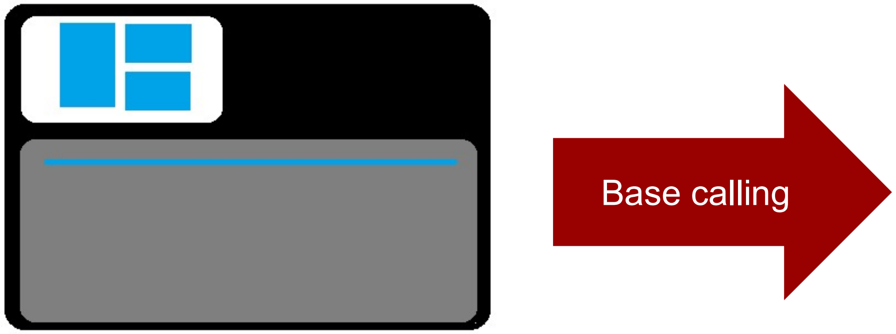
Created with BioRender.com

Paired-end libraries of DNA fragments

- When conducting paired-end libraries, adapters will be attached in pairs
- Insert size is the distance between adapters
- A read pair is produced by reading the insert from opposite ends
- This give positional information for the downstream analysis

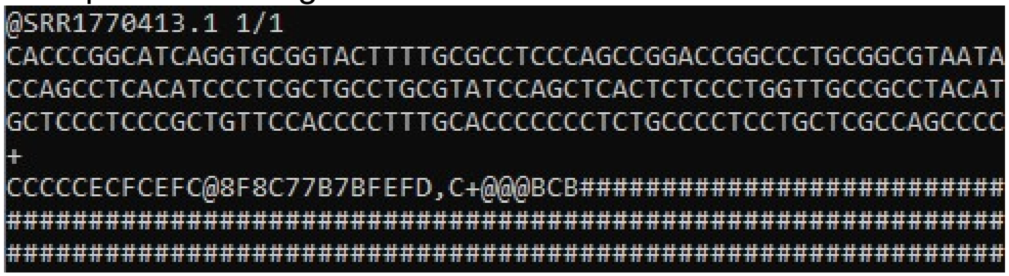


Next generation sequencing data processing



Base calling

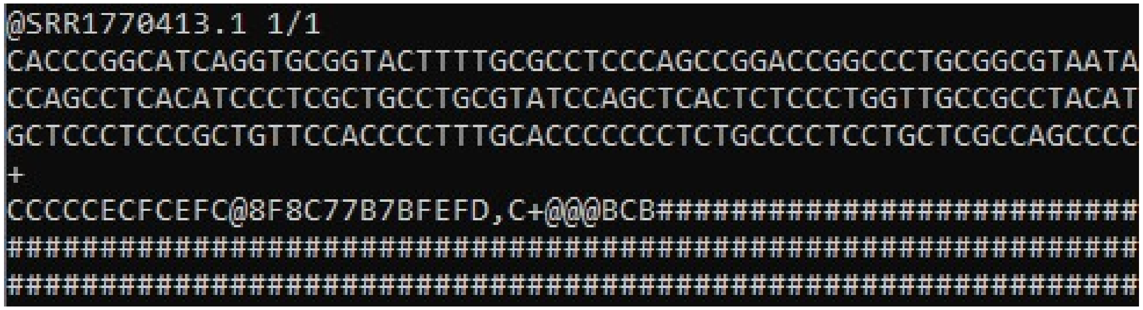
Fastq file containing millions of reads



```
@SRR1770413.1 1/1
CACCCGGCATCAGGTGCGGTACTTTTGCGCCTCCCAGCCGGACCGGCCCTGCGGCGTAATA
CCAGCCTCACATCCCTCGCTGCCTGCGTATCCAGCTCACTCTCCCTGGTTGCCGCCTACAT
GCTCCCTCCCGCTGTTCCACCCCTTTGCACCCCCCCTCTGCCCCTCCTGCTCGCCAGCCCC
+
CCCCCECFCEFC@8F8C77B7BFEFD,C+@@@BCB##########################
#############################################################
#############################################################
```

What is fastq?

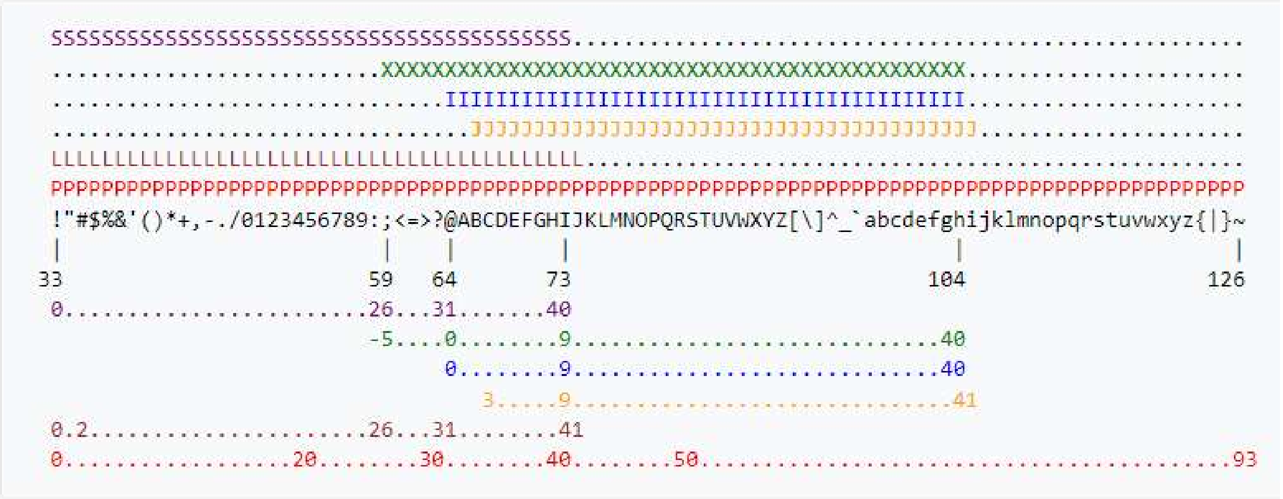
- Fastq are the the read files produced by sequencing machines, after base-calling.
- It has a particular format:
  - Header
    - Contains info on the run, depends on machine
    - Unique ID
  - Called bases
    - Sequence
  - Spacer line
    - Spacing
  - Base quality scores
    - Phred-score giving the probability that the base call is incorrect.



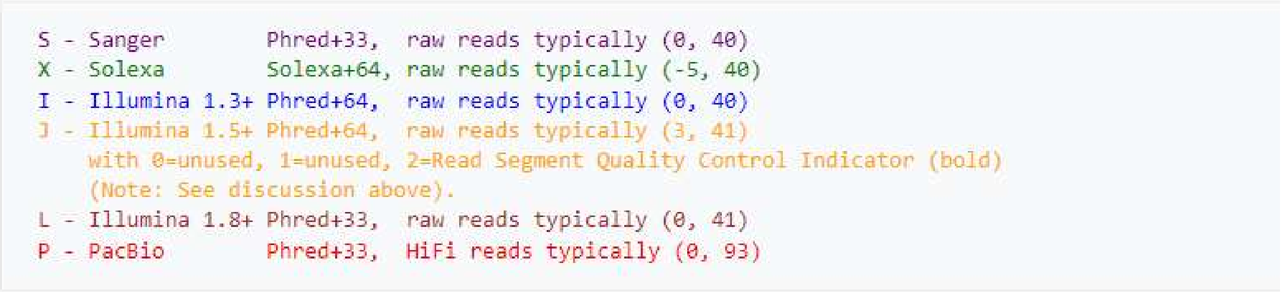
```
@SRR1770413.1 1/1
CACCCGGCATCAGGTGCGGTACTTTTGCGCCTCCCAGCCGGACCGGCCCTGCGGCGTAATA
CCAGCCTCACATCCCTCGCTGCCTGCGTATCCAGCTCACTCTCCCTGGTTGCCGCCTACAT
GCTCCCTCCCGCTGTTCCACCCCTTTGCACCCCCCCTCTGCCCCTCCTGCTCGCCAGCCCC
+
CCCCCECFCEFC@8F8C77B7BFEFD,C+@@@BCB##########################
#############################################################
#############################################################
```

Phred scores?

- The Phred quality score given as one of the 127 standard ASCII characters
- The scale is off-set, with different sequencing machines use different scales
- New Illumina machines use the sanger scale
- The base quality score is important in correctly calling Single Nucleotide Polymorphisms (SNP), used in phylogeny and outbreak detection



```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ...................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126
  0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9..............................41
  0.2......................26...31........41
  0..................20........30........40........50..........................................93
```



```
 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
 P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

Phred scales used in different machines, from the FASTQ format entry on wikipedia: FASTQ format - Wikipedia

The probability of error

- The Phred quality score is a logarithmic score based on the probability that the base call (nucleotide) is incorrect

- Q10 = 1/10 risk of incorrect base
- Q20 = 1/100 risk of incorrect base
- Q30 = 1/1000 risk of incorrect base

- This means that in a sequence of 100 bp at Q20, there will most likely be at least 1 bp called incorrectly

$$Q = -10 \cdot \log_{10}(P)$$

or in terms of probability

$$P = 10^{-\frac{Q}{10}}$$

Where

P = probability of incorrect base call

Q = Phred quality score

Phred quality score	Probability of incorrect base call	Probability of being correct
10	0.1	90%
20	0.01	99%
30	0.001	99.9%

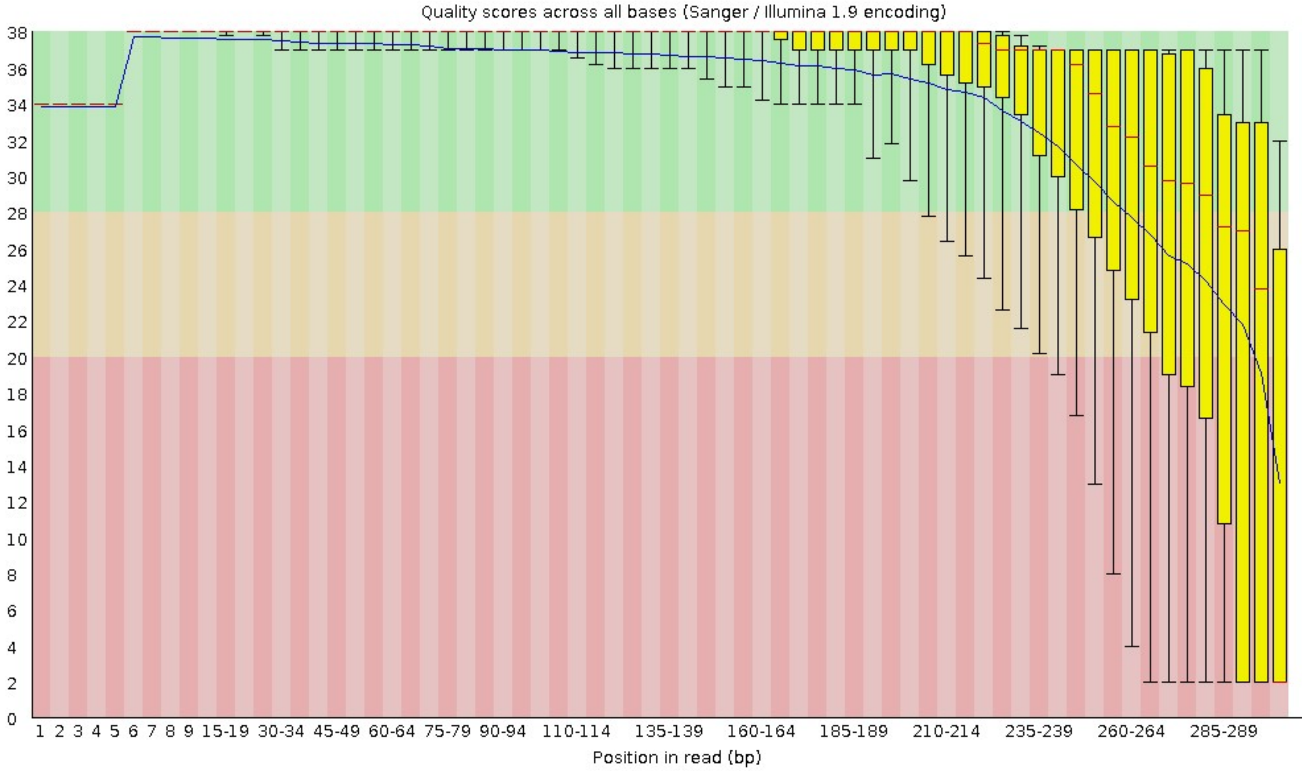
Why does errors occur?

- As multiple rounds of sequencing are conducted, the probability of erroneous base calls increases

- Every time a new base is called an error may occur, meaning the signal for the correct base gets weaker

- Degradation of enzymes used in the reaction may introduce more errors

- This means sequencing with shorter fragments improves base call accuracy

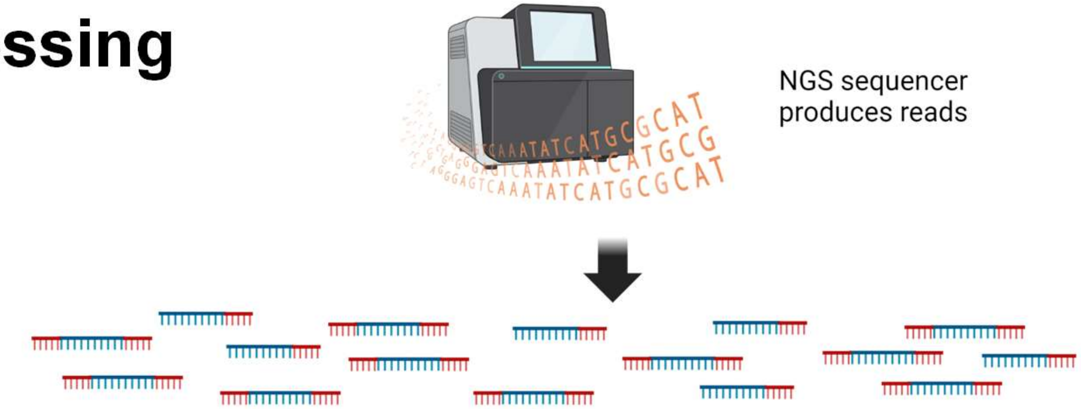


Section of output from running fastqc

Created with BioRender.com

NGS data processing

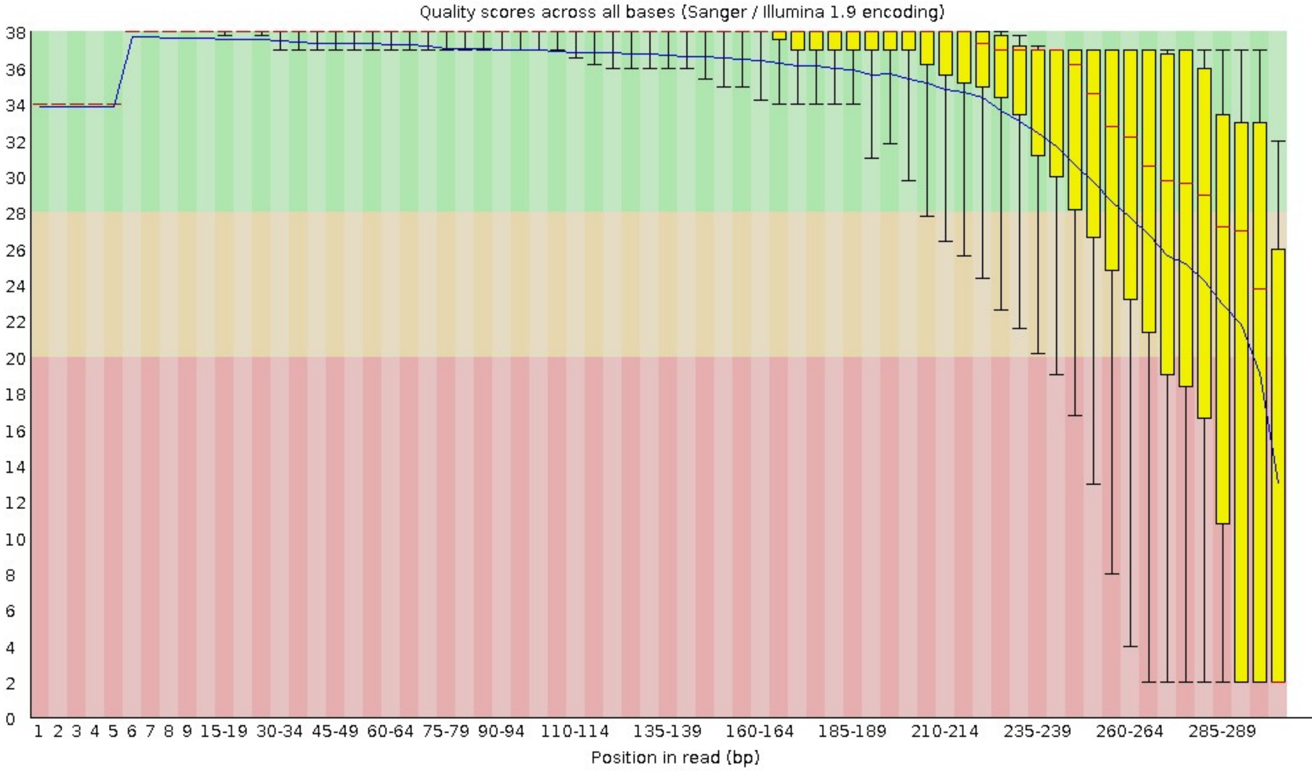
- The raw reads are produced by the sequencing platform


NGS sequencer produces reads

Created with BioRender.com

Trimming

- On Illumina platforms, adapter sequences are not sequenced at the 5’ end of the read, however we can sequence through the entire fragment and start sequencing the adapter at the 3’ end
- We base call at the end of the read may also be of too poor quality for analysis.
- Wrong base calls can impact phylogenetic analysis and gene annotation.



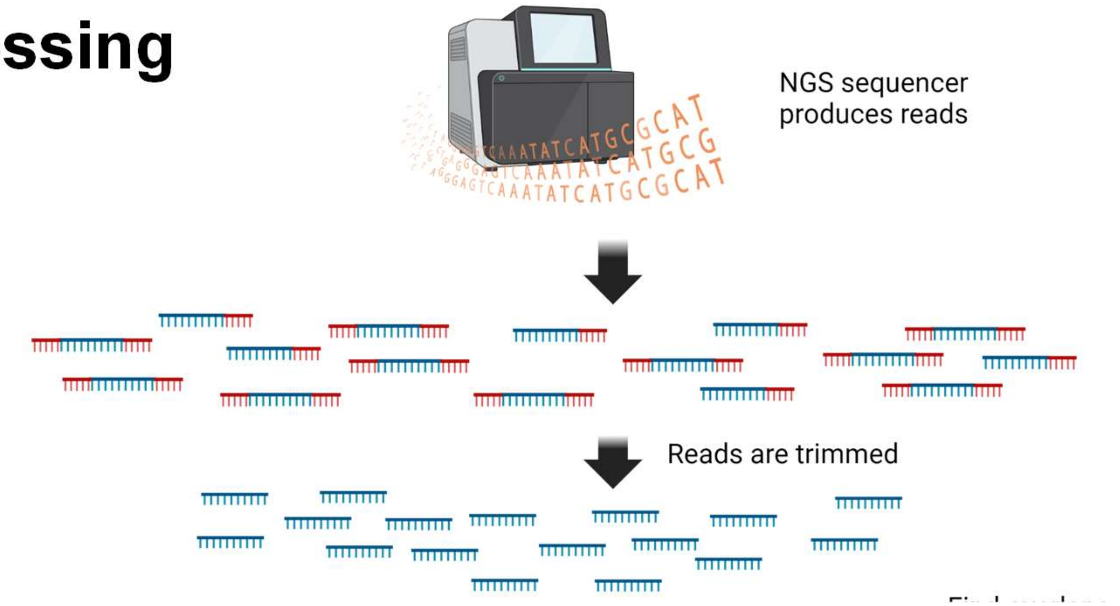
Section of output from running fastqc

NGS data processing

- The raw reads are produced by the sequencing platform

- Poor sequences are trimmed of the raw reads, leaving high confidence DNA stretches (trimmed reads)

Created with BioRender.com



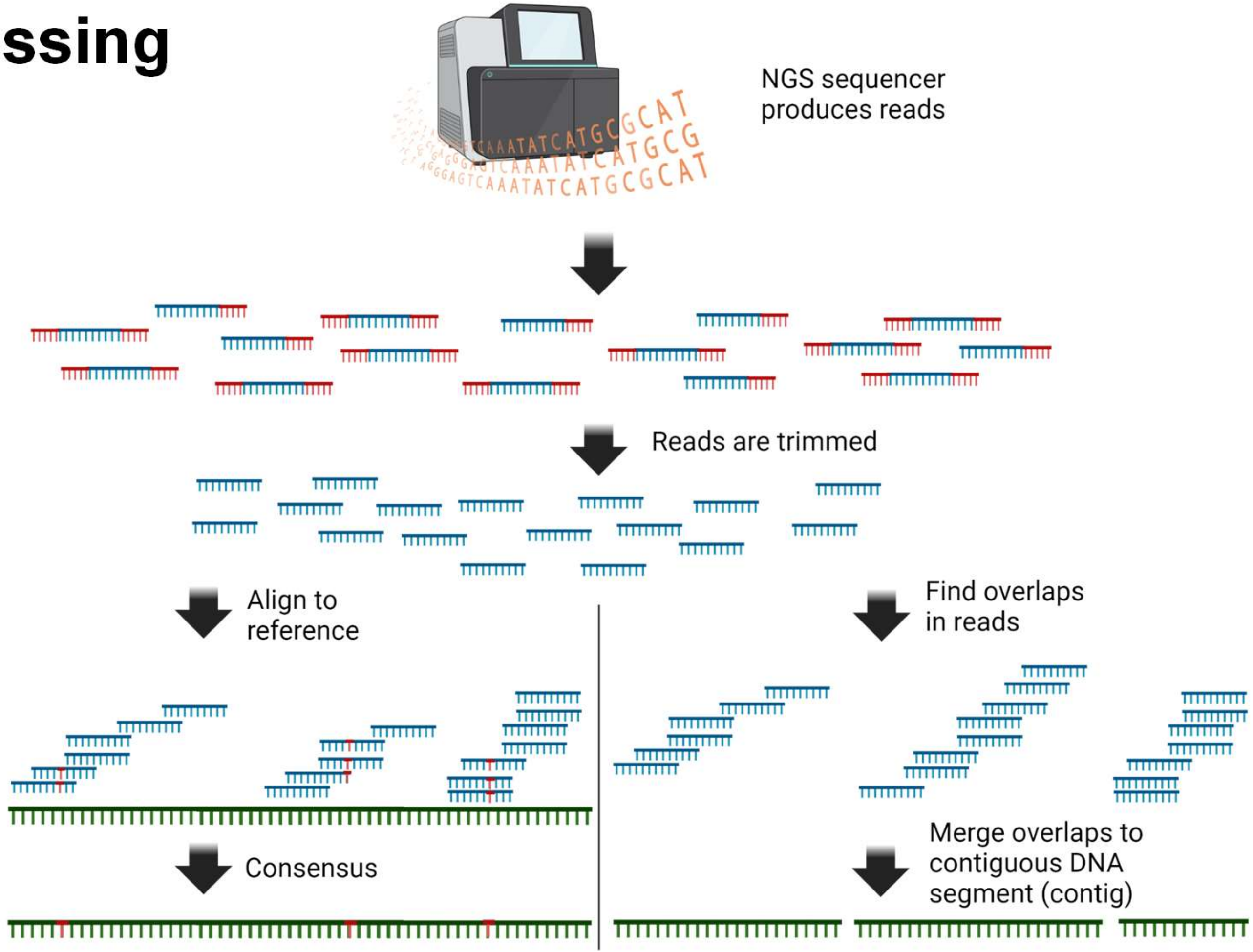
NGS data processing

- The raw reads are produced by the sequencing platform

- Poor sequences are trimmed of the raw reads, leaving high confidence DNA stretches (trimmed reads)

- We can then apply two standard approaches:
  - Mapping: Is we are sequencing a known pathogen (e.g. from a outbreak) we can align reads to a previously constructed assembly (a reference genome)
  - De novo assembly: We can infer the genome of the pathogen by constructing an assembly

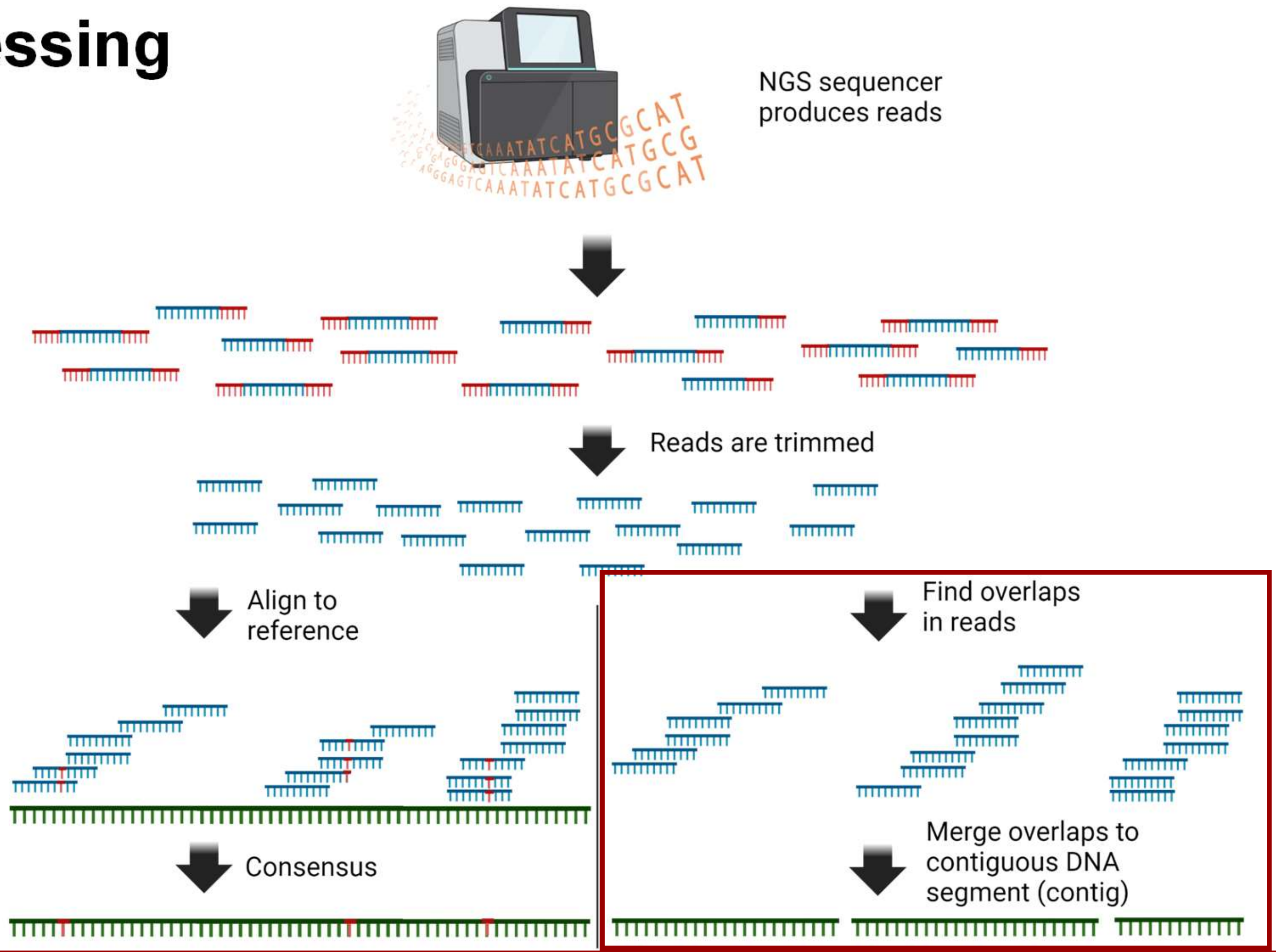
Created with BioRender.com



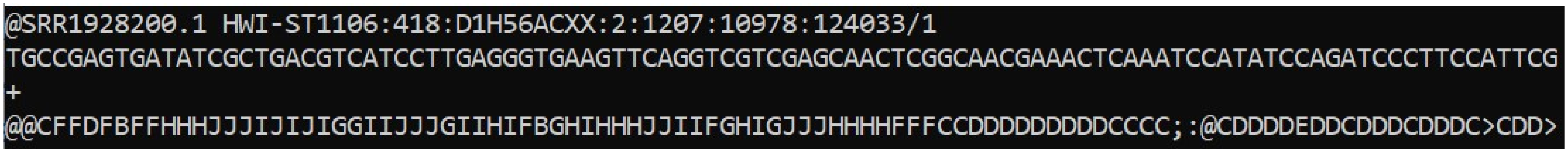
NGS data processing

- The raw reads are produced by the sequencing platform
- Poor sequences are trimmed of the raw reads, leaving high confidence DNA stretches (trimmed reads)
- We can then apply two standard approaches:
  - Mapping: Is we are sequencing a known pathogen (e.g. from a outbreak) we can align reads to a previously constructed assembly (a reference genome)
  - De novo assembly: We can infer the genome of the pathogen by constructing an assembly

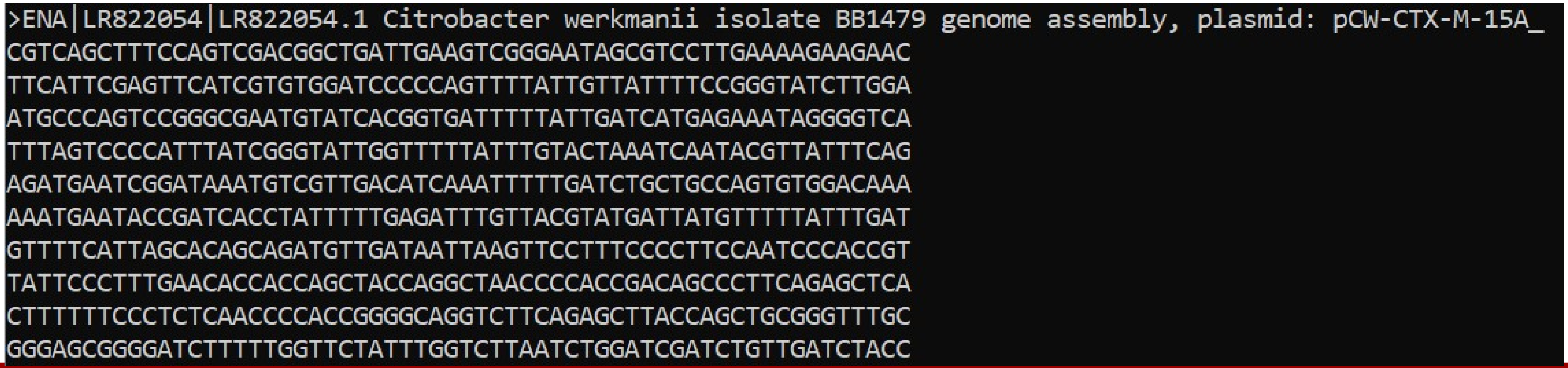
Created with BioRender.com



From fastq to fasta



```
@SRR1928200.1 HWI-ST1106:418:D1H56ACXX:2:1207:10978:124033/1
TGCCGAGTGATATCGCTGACGTCATCCTTGAGGGTGAAGTTCAGGTCGTCGAGCAACTCGGCAACGAAACTCAAATCCATATCCAGATCCCTTCCATTCG
+
@@CFFDFBFFHHHJJJIJIJIGGIIJJJGIIHIFBGHIHHHJJIIFGHIGJJJHHHHFFFCCDDDDDDDDDCCCC;:@CDDDDEDDCDDDCDDDC>CDD>
```

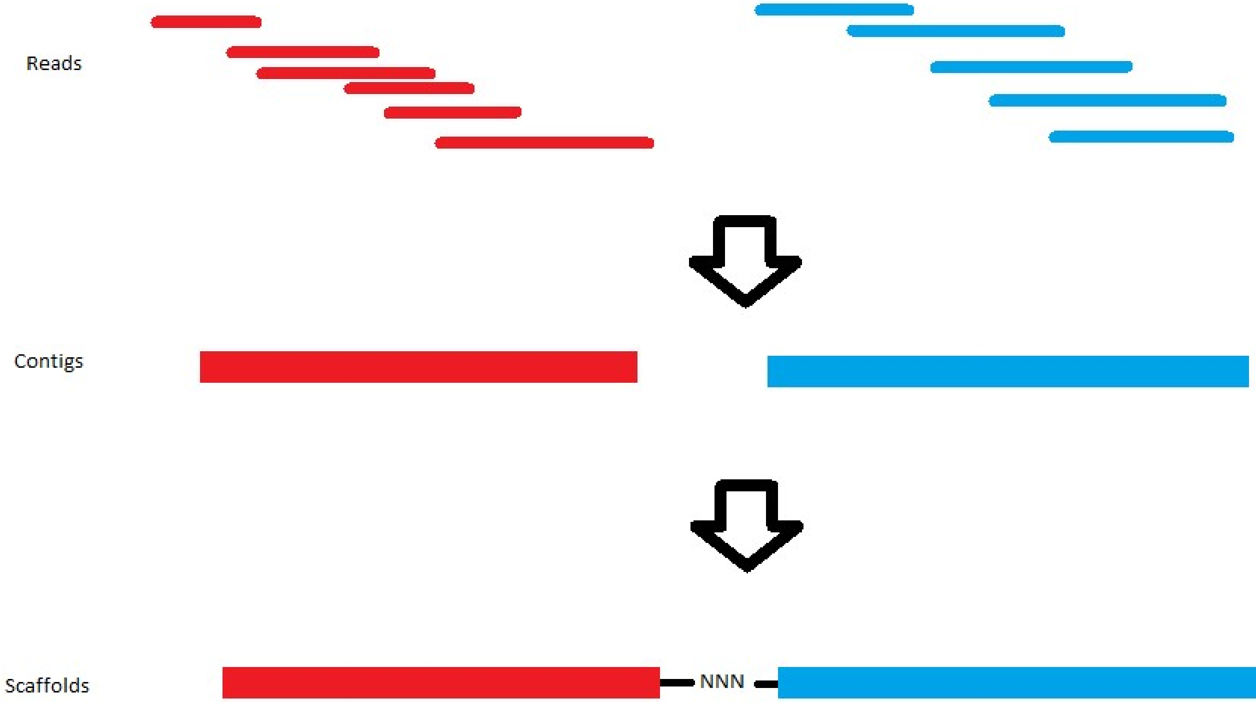


```
>ENA|LR822054|LR822054.1 Citrobacter werkmanii isolate BB1479 genome assembly, plasmid: pCW-CTX-M-15A_
CGTCAGCTTTCCAGTCGACGGCTGATTGAAGTCGGGAATAGCGTCCTTGAAAAGAAGAAC
TTCATTCGAGTTCATCGTGTGGATCCCCCAGTTTTATTGTTATTTTCCGGGTATCTTGGA
ATGCCCAGTCCGGGCGAATGTATCACGGTGATTTTTATTGATCATGAGAAATAGGGGTCA
TTTAGTCCCCATTTATCGGGTATTGGTTTTTATTTGTACTAAATCAATACGTTATTTCAG
AGATGAATCGGATAAATGTCGTTGACATCAAATTTTTGATCTGCTGCCAGTGTGGACAAA
AAATGAATACCGATCACCTATTTTTGAGATTTGTTACGTATGATTATGTTTTTATTTGAT
GTTTTCATTAGCACAGCAGATGTTGATAATTAAGTTCCTTTCCCCTTCCAATCCCACCGT
TATTCCCTTTGAACACCACCAGCTACCAGGCTAACCCCACCGACAGCCCTTCAGAGCTCA
CTTTTTTCCCTCTCAACCCCACCGGGGCAGGTCTTCAGAGCTTACCAGCTGCGGGTTTGC
GGGAGCGGGGATCTTTTTGGTTCTATTTGGTCTTAATCTGGATCGATCTGTTGATCTACC
```

De novo assembly

- Many programs can do assembly, they differentiate by how precisely they can construct the assembly, how fast and how computationally heavy their workload
  - SPAdes
  - SOAPdenovo2
  - MEGAHIT
  - Velvet
  - “shovill"

- The assembly should not contain unknown bases (N), e.g. we usually work with the contigs, and not the scaffolds



Sequencing Quality Control

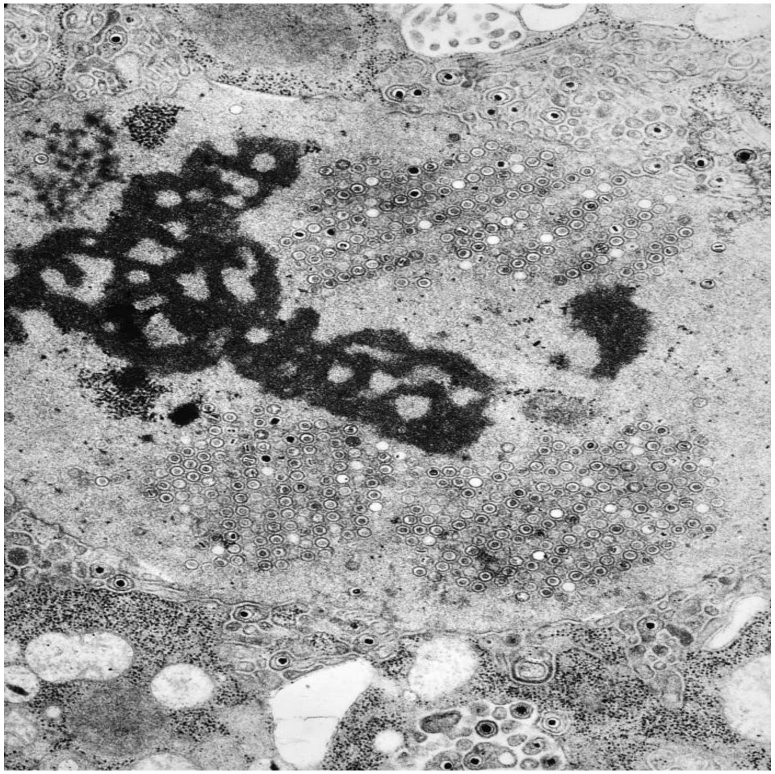
- Many different parameters are used for evaluation of the sequencing
  - Total size of assembly
  - N50
  - Number of contigs (>200 bp)
  - Sequence depth/coverage
  - Genomic coverage

- Another possible option is checking the GC% content which is expected to be in a very narrow range for a species.

- It is important to know how successful the sequencing was both for internal purposes and to evaluate data used from else (e.g. online sequence repositories)

Assembly statistics – total base pairs

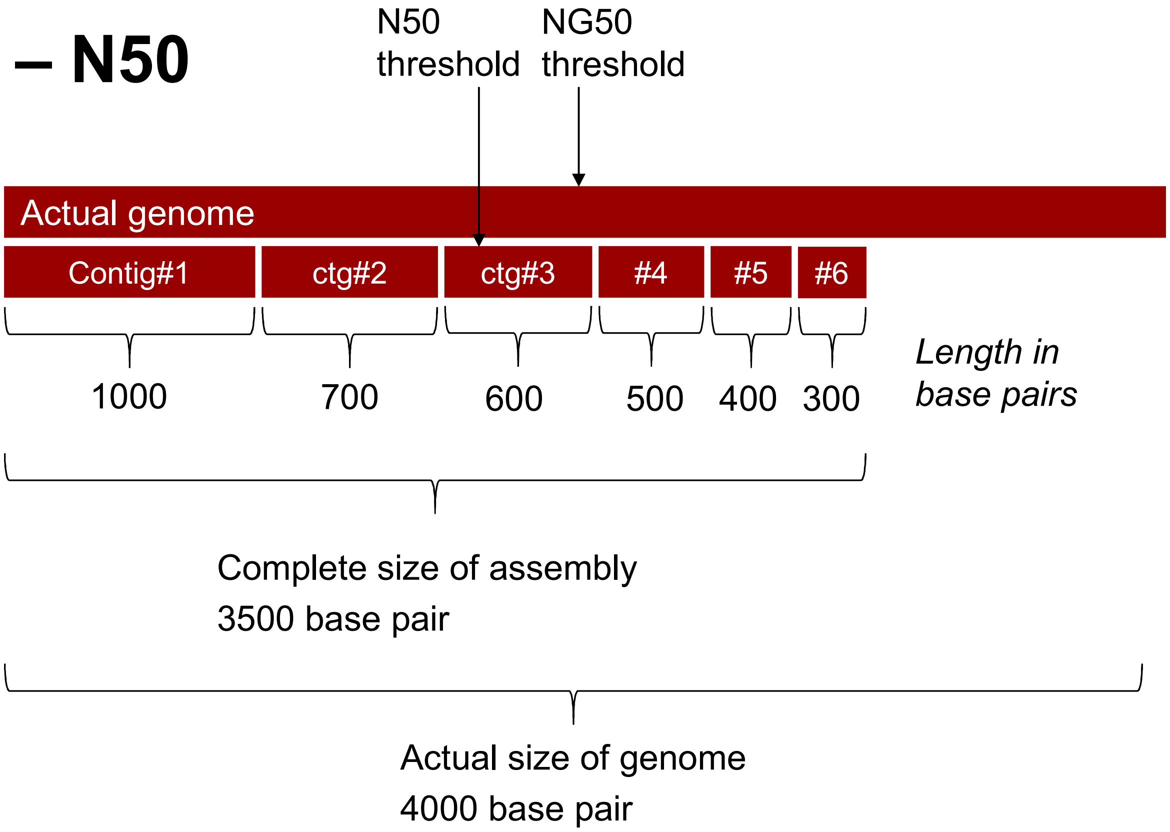
- Total base pairs are the total length of all contigs in your assembly
- For whole genome sequencing we expect it to be close to the actual size of the genome
- Comparing the total base pairs of an assembly with a reference of the same expected sp. can reveal contamination or misidentification
- E.g. *Salmonella enterica* is expected to be 4.4-5.0 Mbp, if assembly contains 8 Mb, it is like due to contamination



Source: CDC/ Dr. Fred Murphy; Sylvia Whitfield

Assembly statistics – N50

- N50 is found by:
  - Sorting all contigs in assembly from longest to shortest, starting with the longest
  - Adding together the length of the longest contigs until half the assembly is included
  - The length of the last added contig to reach 50% of the assembly is the N50
- N50 gives a measure for how much of the assembly is captured in as few contigs as possible
- The higher the N50, the better the assembly, the better the sequencing



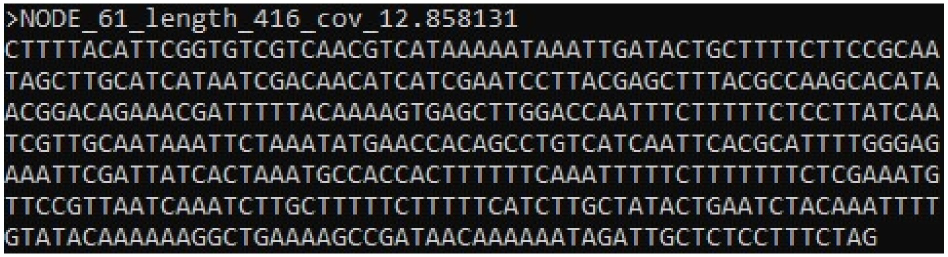
Assembly statistics – number of contigs

- When we assembly we never expect to be able to produce a closed genome (at least not using short read sequencing)

- This is due to several factors including repeated sequences

- We want the lowest number of contigs possible, as this makes e.g. gene identification and annotation more feasible

- Often, contigs below 200 bp are not counted



```
>NODE_61_length_416_cov_12.858131
CTTTTACATTCGGTGTCGTCAACGTCATAAAAATAAATTGATACTGCTTTTCTTCCGCAA
TAGCTTGCATCATAATCGACAACATCATCGAATCCTTACGAGCTTTACGCCAAGCACATA
ACGGACAGAAACGATTTTTACAAAAGTGAGCTTGGACCAATTTCTTTTTCTCCTTATCAA
TCGTTGCAATAAATTCTAAATATGAACCACAGCCTGTCATCAATTCACGCATTTTGGGAG
AAATTCGATTATCACTAAATGCCACCACTTTTTTCAAATTTTTCTTTTTTTCTCGAAATG
TTCCGTTAATCAAATCTTGCTTTTTCTTTTTCATCTTGCTATACTGAATCTACAAATTTT
GTATACAAAAAAGGCTGAAAAGCCGATAACAAAAAATAGATTGCTCTCCTTTCTAG
```

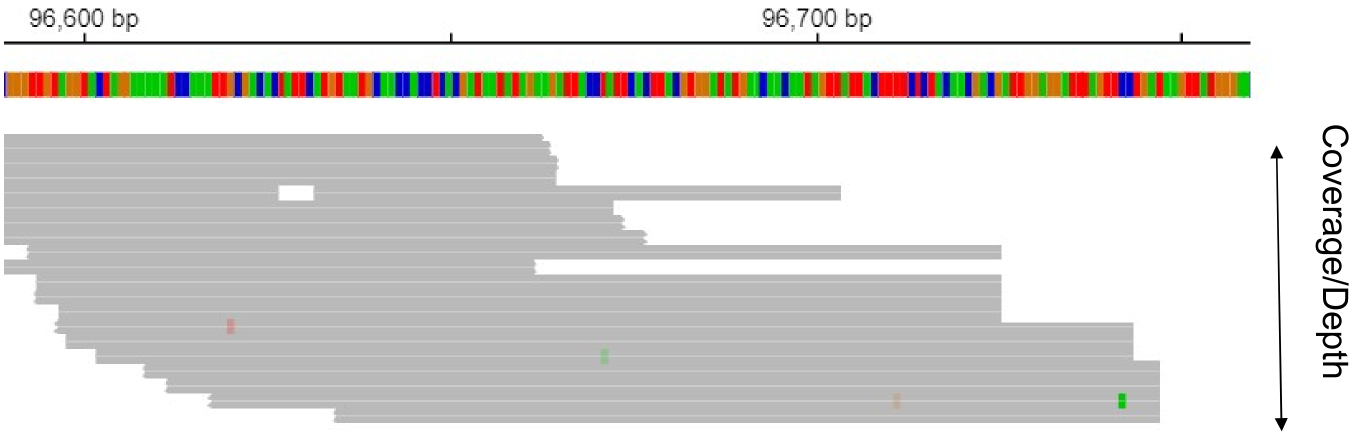
Assembly statistics – Depth (Sequence coverage)

- The number times we cover a part of the assembled genome is called sequencing depth

- Often also called coverage

- The deeper we sequence a part of the genome, the more sure we are about the called bases

- Average coverage would be:

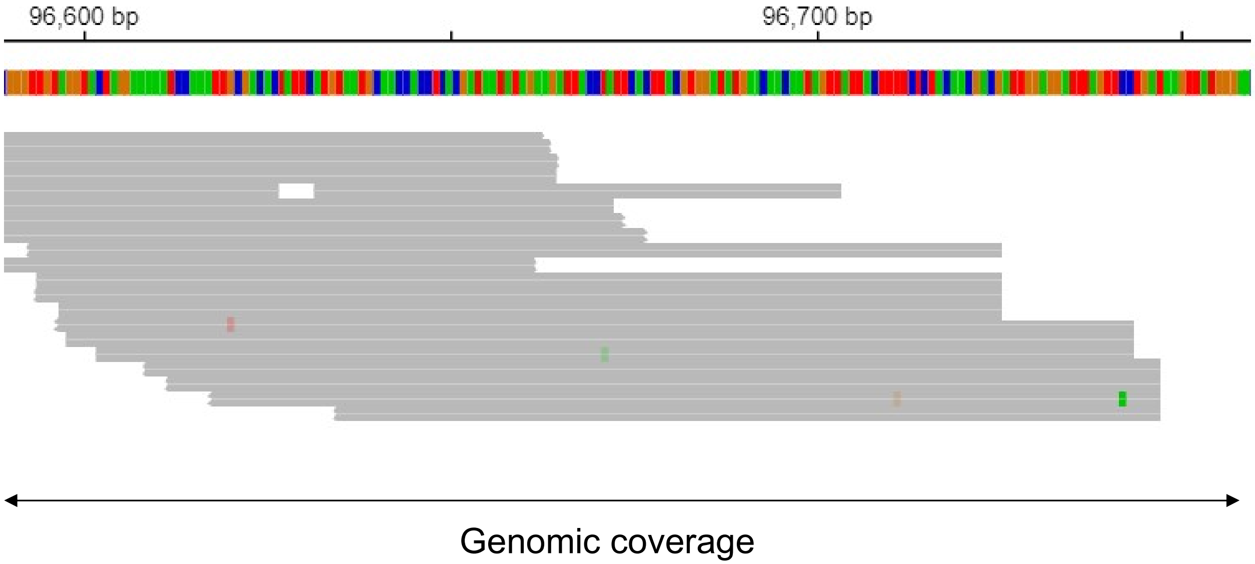


$$sequence\ coverage = \frac{number\ of\ reads\ *average\ read\ length}{Total\ genome\ size}$$

$$sequence\ coverage = \frac{9\ *100bp}{800bp} = 1.125x$$

Assembly statistics – Physical coverage

- If a closed reference genome is available the physical coverage can likewise be calculated
- The physical coverage is the percentage of the assembly covered by reads
- The percentage should be as high as possible



Suggestions for thresholds

- There is no universal thresholds for the quality metrics described and they can be expected to vary depending on the specific species and strain. The table below are suggestion based on experience and available literature

Species	Size of assembly (Mbp)	N50	Number of contigs
E. Coli	~4.5 - 5.9	>50,000	<500
Campylobacter	~1.5 - 1.9	>100,000	<250
Klebsiella	~5.0 - 6.2	>50,000	<500
Salmonella	~4.3 - 5.2	>50,000	<300

Further reading: Vornhagen, J. *et al.* (2022). Timme, R.E. *et al* (2020). Kristensen, T. *et al* (2023). Ellington, M.J. *et al* (2016) [see next slide]

- Vornhagen, J., Roberts, E.K., Unverdorben, L. *et al.* Combined comparative genomics and clinical modeling reveals plasmid-encoded genes are independently associated with *Klebsiella* infection. *Nat Commun* **13**, 4459 (2022). https://doi.org/10.1038/s41467-022-31990-1

- Timme RE, Wolfgang WJ, Balkey M, Venkata SLG, Randolph R, Allard M, Strain E. Optimizing open data to support one health: best practices to ensure interoperability of genomic data from bacterial pathogens. One Health Outlook. 2020;2(1):20. doi: 10.1186/s42522-020-00026-3. Epub 2020 Oct 19. PMID: 33103064; PMCID: PMC7568946.

- Kristensen T, Sørensen LH, Pedersen SK, Jensen JD, Mordhorst H, Lacy-Roberts N, Lukjancenko O, Luo Y, Hoffmann M, Hendriksen RS. Results of the 2020 Genomic Proficiency Test for the network of European Union Reference Laboratory for Antimicrobial Resistance assessing whole-genome-sequencing capacities. Microb Genom. 2023 Aug;9(8):mgen001076. doi: 10.1099/mgen.0.001076. PMID: 37526643; PMCID: PMC10483428.

- Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, Grundman H, Hasman H, Holden MTG, Hopkins KL, Iredell J, Kahlmeter G, Köser CU, MacGowan A, Mevius D, Mulvey M, Naas T, Peto T, Rolain JM, Samuelsen Ø, Woodford N. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. Clin Microbiol Infect. 2017 Jan;23(1):2-22. doi: 10.1016/j.cmi.2016.11.012. Epub 2016 Nov 23. PMID: 27890457.

Detection of specific resistance mechanisms – ESBL and CRE

The Fleming Fund
Regional Grants

UKaid from the British people

MOTT MACDONALD

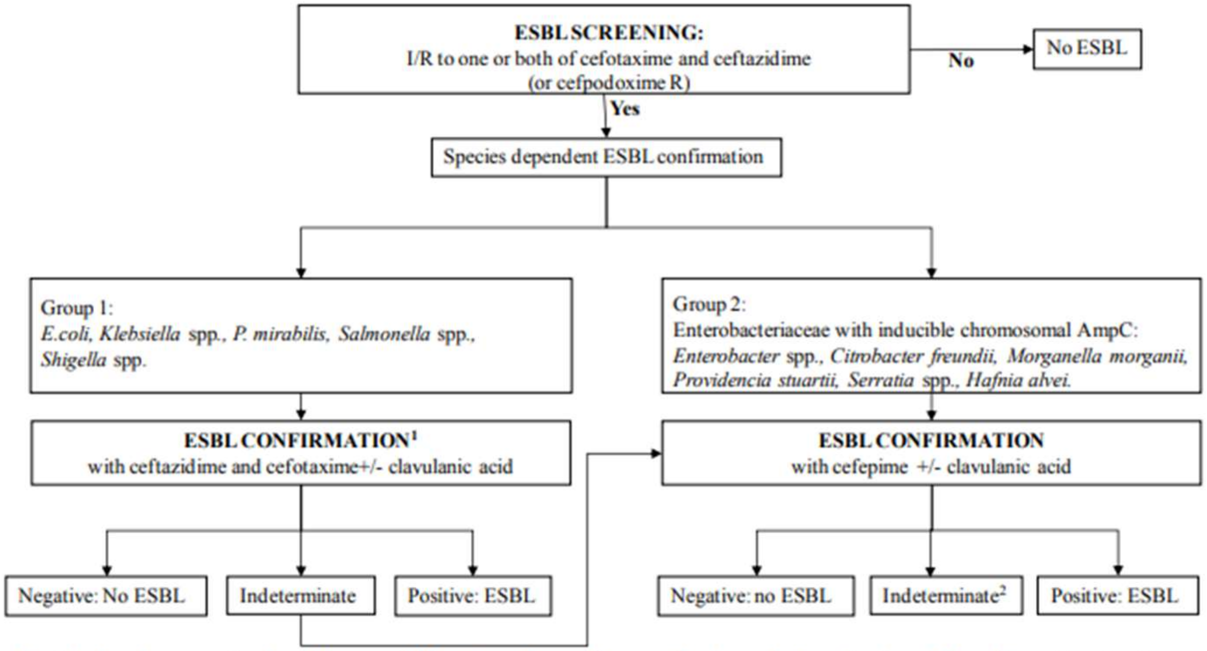
International Vaccine Institute

Faculty of Veterinary Science • Chulalongkorn University •

2023
DTU

ESBL detection

Figure 1. Algorithm for phenotypic detection of ESBLs



ESBL SCREENING:
I/R to one or both of cefotaxime and ceftazidime
(or cefpodoxime R)

No → No ESBL

Yes → Species dependent ESBL confirmation

Group 1:
E.coli, *Klebsiella* spp., *P. mirabilis*, *Salmonella* spp., *Shigella* spp.

ESBL CONFIRMATION[1]
with ceftazidime and cefotaxime+/- clavulanic acid

Negative: No ESBL

Indeterminate

Positive: ESBL

Group 2:
Enterobacteriaceae with inducible chromosomal AmpC:
Enterobacter spp., *Citrobacter freundii*, *Morganella morganii*, *Providencia stuartii*, *Serratia* spp., *Hafnia alvei*.

ESBL CONFIRMATION
with cefepime +/- clavulanic acid

Negative: no ESBL

Indeterminate[2]

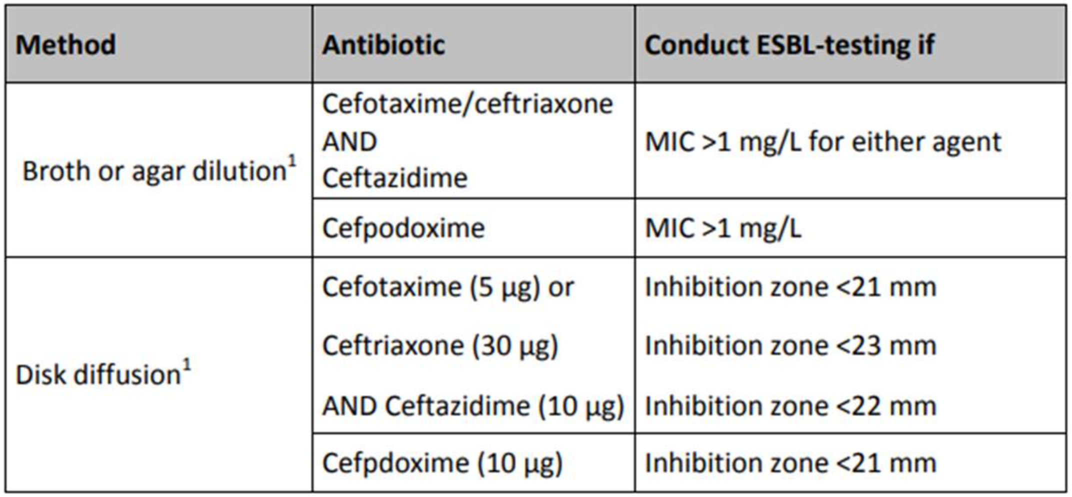
Positive: ESBL

[1] If cefoxitin has been tested and has an MIC >8 mg/L, perform cefepime+/- clavulanic acid confirmation test

[2] Cannot be determined as either positive or negative (e.g. if a gradient diffusion strip cannot be read due to growth beyond the MIC range of the strip or there is no clear synergy in combination-disk and double-disk synergy tests). In confirmation with cefepime +/- clavulanic acid is still indeterminate, genotypic testing is required.

ESBL screening methods

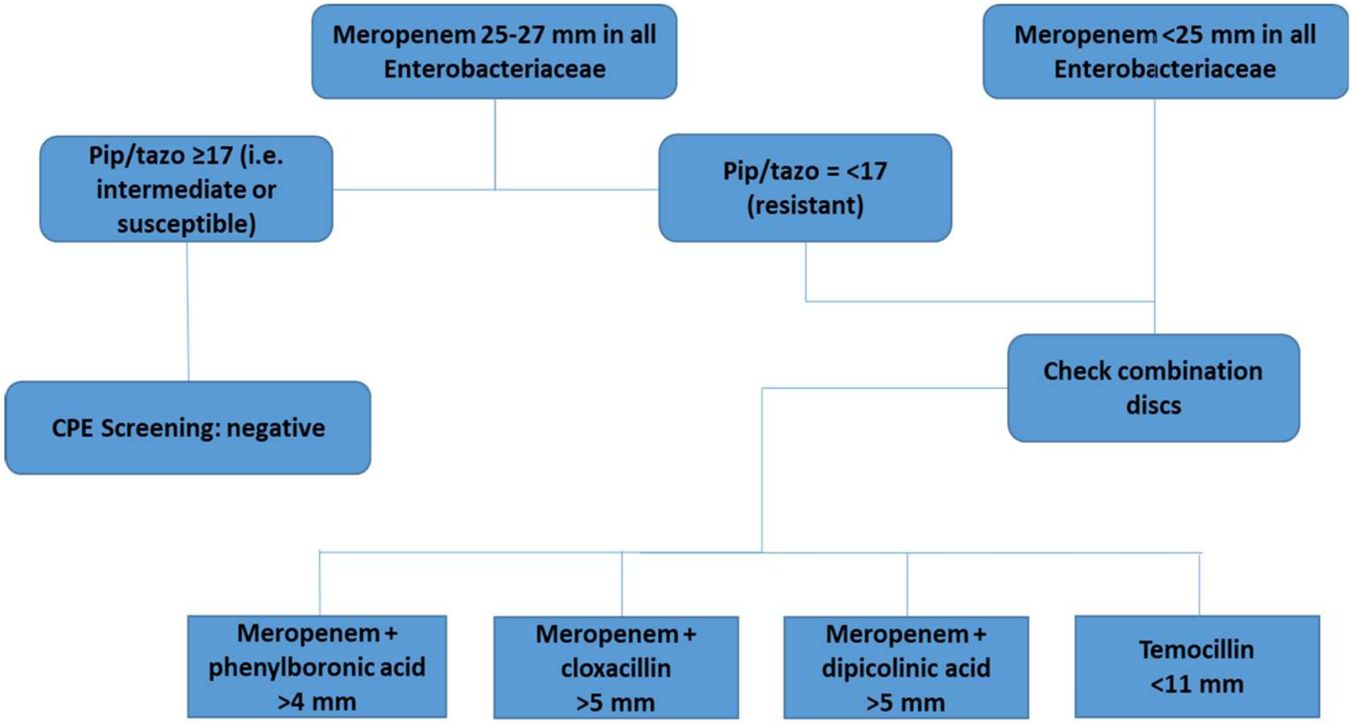
Table 1. ESBL screening methods for Enterobacteriaceae (13-19).



Method	Antibiotic	Conduct ESBL-testing if
Broth or agar dilution[1]	Cefotaxime/ceftriaxone AND Ceftazidime	MIC >1 mg/L for either agent
	Cefpodoxime	MIC >1 mg/L
Disk diffusion[1]	Cefotaxime (5 µg) or Ceftriaxone (30 µg) AND Ceftazidime (10 µg)	Inhibition zone <21 mm Inhibition zone <23 mm Inhibition zone <22 mm
	Cefpdoxime (10 µg)	Inhibition zone <21 mm

EUCAST guidelines for detection of resistance mechanisms and specific resistances of clinical and/or epidemiological importance. Version 2,0,

CRE screening



Meropenem 25-27 mm in all Enterobacteriaceae

Meropenem <25 mm in all Enterobacteriaceae

Pip/tazo ≥17 (i.e. intermediate or susceptible)

Pip/tazo = <17 (resistant)

CPE Screening: negative

Check combination discs

Meropenem + phenylboronic acid >4 mm

Meropenem + cloxacillin >5 mm

Meropenem + dipicolinic acid >5 mm

Temocillin <11 mm

EUCAST guidelines for detection of resistance mechanisms and specific resistances of clinical and/or epidemiological importance. Version 2,0, July 2017

Combination disk method

- meropenem (10µg) +/- various inhibitors

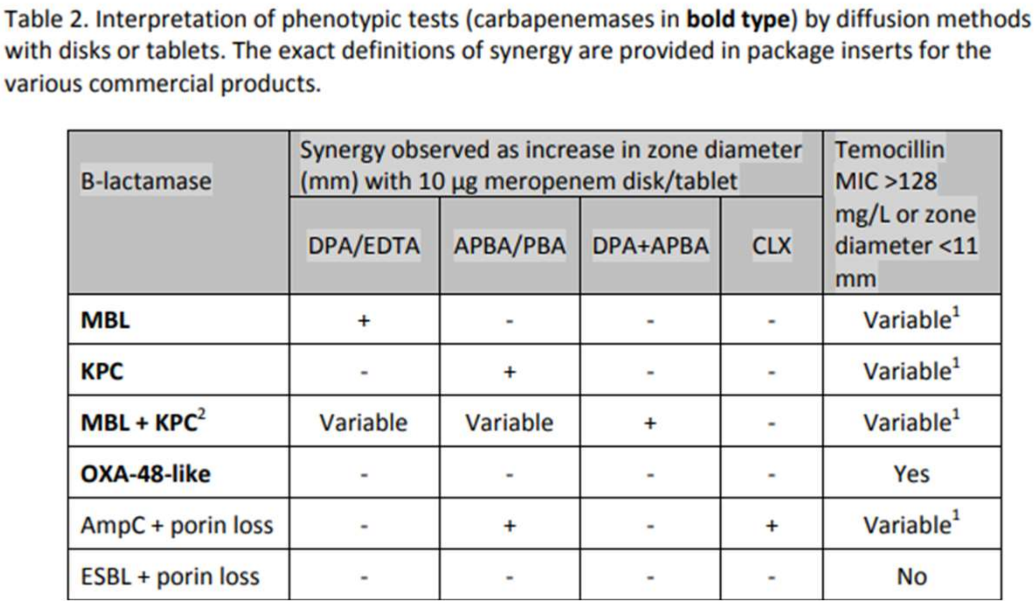
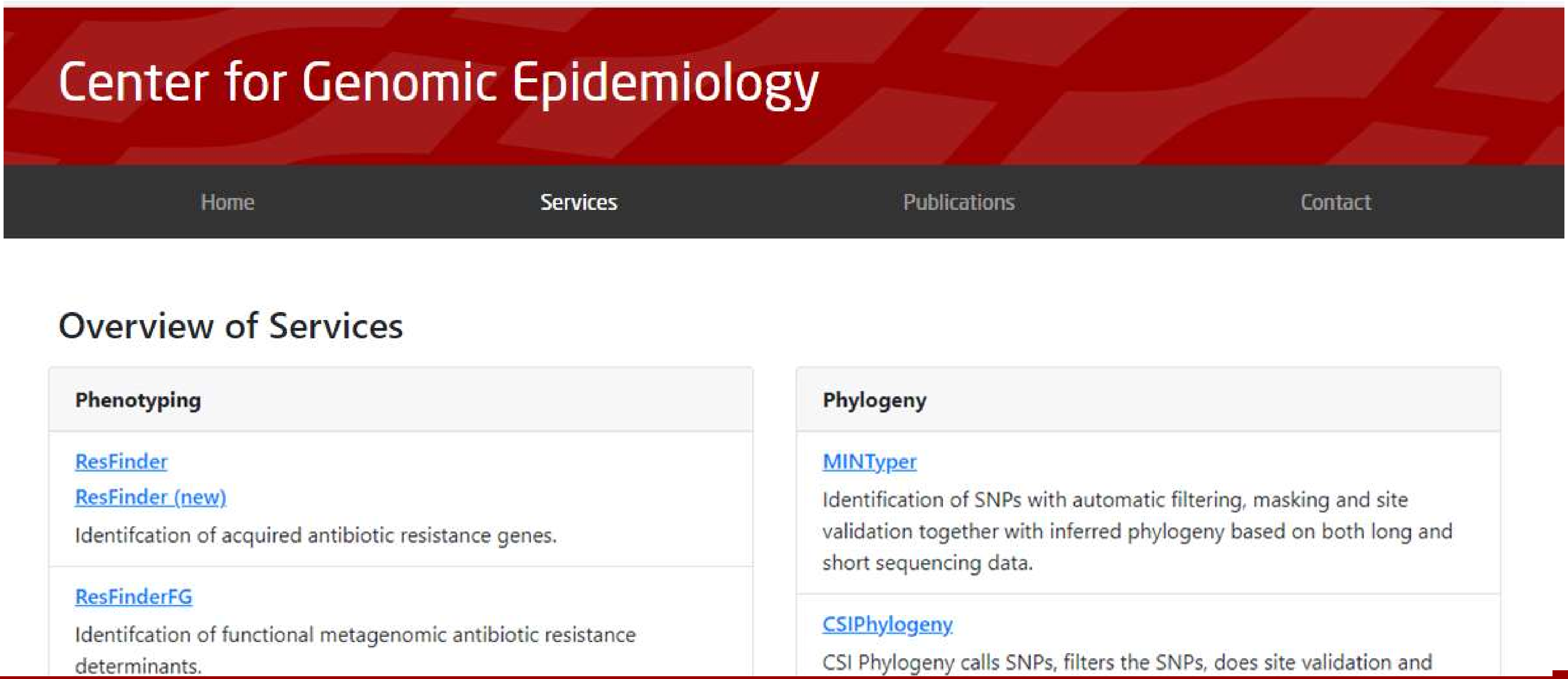


Table 2. Interpretation of phenotypic tests (carbapenemases in **bold type**) by diffusion methods with disks or tablets. The exact definitions of synergy are provided in package inserts for the various commercial products.

B-lactamase	Synergy observed as increase in zone diameter (mm) with 10 µg meropenem disk/tablet				Temocillin MIC >128 mg/L or zone diameter <11 mm
	DPA/EDTA	APBA/PBA	DPA+APBA	CLX	
MBL	+	-	-	-	Variable[1]
KPC	-	+	-	-	Variable[1]
MBL + KPC[2]	Variable	Variable	+	-	Variable[1]
OXA-48-like	-	-	-	-	Yes
AmpC + porin loss	-	+	-	+	Variable[1]
ESBL + porin loss	-	-	-	-	No

EUCAST guidelines for detection of resistance mechanisms and specific resistances of clinical and/or epidemiological importance. Version 2,0, July 2017

Genomic analysis – Using the CGE tools



Overview of Services

Phenotyping

ResFinder
ResFinder (new)
Identifcation of acquired antibiotic resistance genes.

ResFinderFG
Identifcation of functional metagenomic antibiotic resistance determinants.

Phylogeny

MINTyper
Identification of SNPs with automatic filtering, masking and site validation together with inferred phylogeny based on both long and short sequencing data.

CSIPhylogeny
CSI Phylogeny calls SNPs, filters the SNPs, does site validation and

Genomic analysis – Using the CGE tools

- Available at: https://www.genomicepidemiology.org/services/

- We will talk about genomic analysis and look at the associated tools:
  - Kmerfinder (for species verification)
  - MLST (for typing)
  - Resfinder (for detection of AMR genes and mutations)
  - Plasmidfinder (for identification of plasmid replicons)
  - CSIphylogeny (for SNP-based characterization)

Genomic analysis – species verification

- The term bacterial species is widely used, but poorly defined

- In general bacterial species are defined by phenotypic and genotypic differences, meaning bacteria showing high genomic similarity and phenotypic traits are clustered into a single species

- Ribosomal 16S gene have been used to identify species and is still used in metagenomics - but does not provide enough discriminatory power between closely related species (*Shigella* spp – *Escherichia coli*)

- Multiple approaches have been used, we will look into a kmer-based method

What is a kmer?

- A kmer is a substring within a stretch of DNA of length “k”

- When dividing a DNA sequence into kmers, you start with the first k basepairs and then proceed by moving one nucleotide at a time

- E.g. let us look at the sequence to the right and divide it into kmers of length 4 (into 4mers)

ATGCATATTG

What is a kmer?

- A kmer is a substring within a stretch of DNA of length “k”
- When dividing a DNA sequence into kmers, you start with the first k basepairs and then procede by moving one nucleotide at a time
- E.g. let us look at the sequence to the right and divide it into kmers of length 4 (into 4mers)
- The first 4mer consist of the first 4 bases

ATGCATATTG

ATGC

What is a kmer?

- A kmer is a substring within a stretch of DNA of length “k”
- When dividing a DNA sequence into kmers, you start with the first k basepairs and then procede by moving one nucleotide at a time
- E.g. let us look at the sequence to the right and divide it into kmers of length 4 (into 4mers)
- The first 4mer consist of the first 4 bases
- We then move one space to the right to identify the next 4mer

```
ATGCATATTG
ATGC
 TGCA
```

What is a kmer?

- A kmer is a substring within a stretch of DNA of length “k”
- When dividing a DNA sequence into kmers, you start with the first k basepairs and then procede by moving one nucleotide at a time
- E.g. let us look at the sequence to the right and divide it into kmers of length 4 (into 4mers)
- The first 4mer consist of the first 4 bases
- We then move one space to the right to identify the next 4mer
- We end up with 7 unique 4mers

```
ATGCATATTG
ATGC
 TGCA
  GCAT
   CATA
    ATAT
     TATT
      ATTG
```

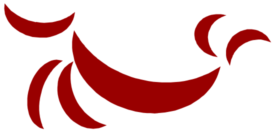
But why?

- Kmers are used in multiple settings to make dealing with sequence data more manageable
  - In search functions like blast
  - In assembly (de brujn graphs)
  - DNA profiling

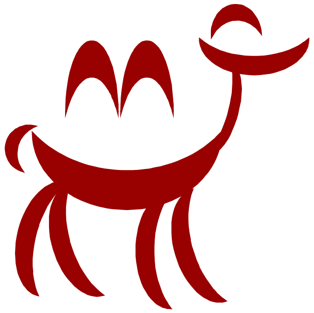
- The longer kmers we use, the more unique their signature

- Kmerfinder uses 16mers to align submitted sequences against a database constructed from the overlapping 16kmers starting with ATGAC

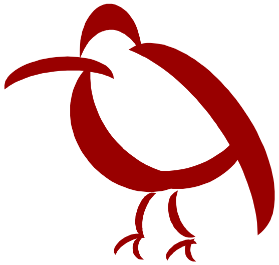
ATGGCCAATTATAGCCCGTCT



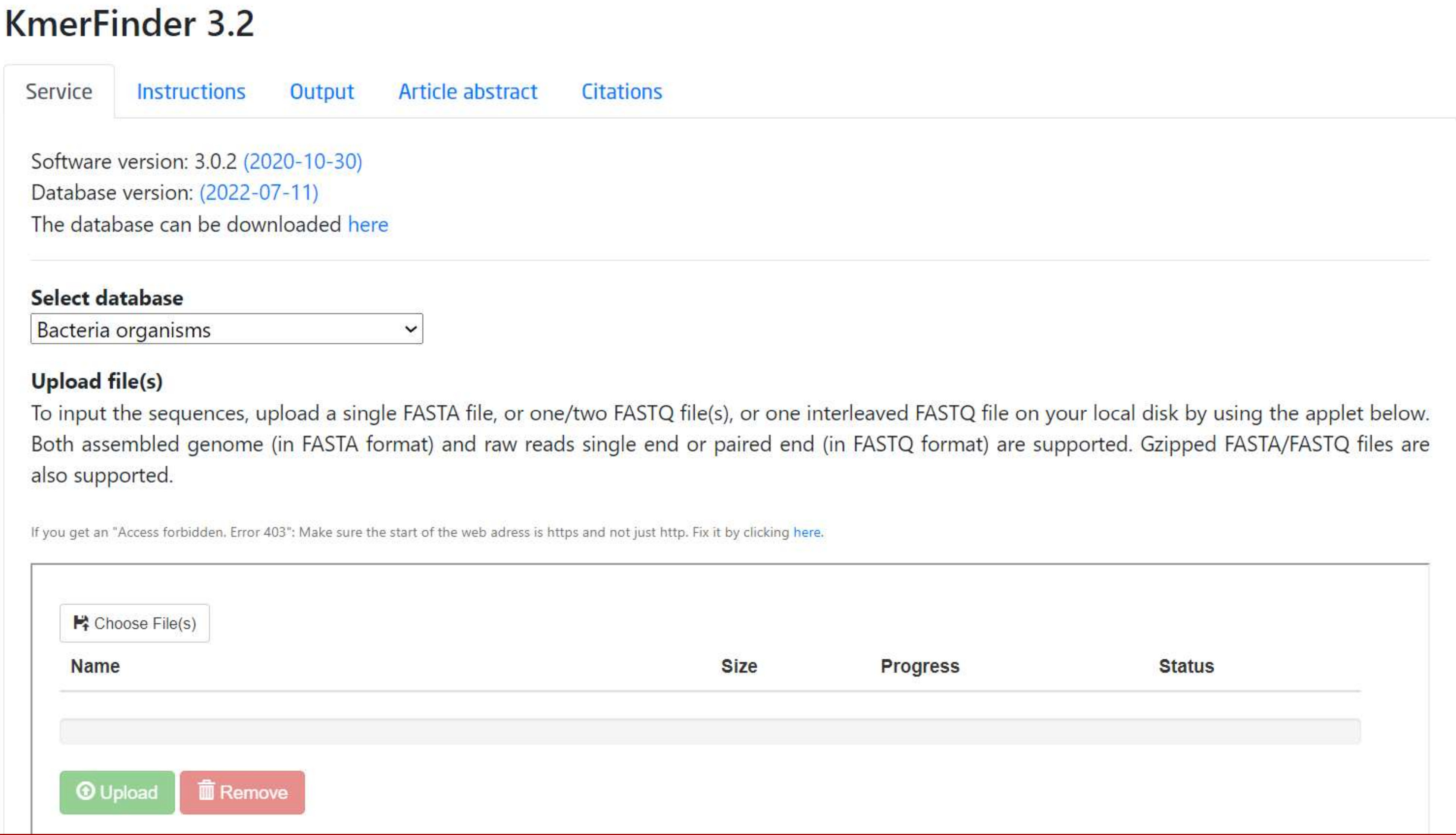
TTAATGGCCAATTATAGCCCG



AGCTGGCCAATTATAGCCC



GATGGCCAATTATAGCTCC



KmerFinder 3.2

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

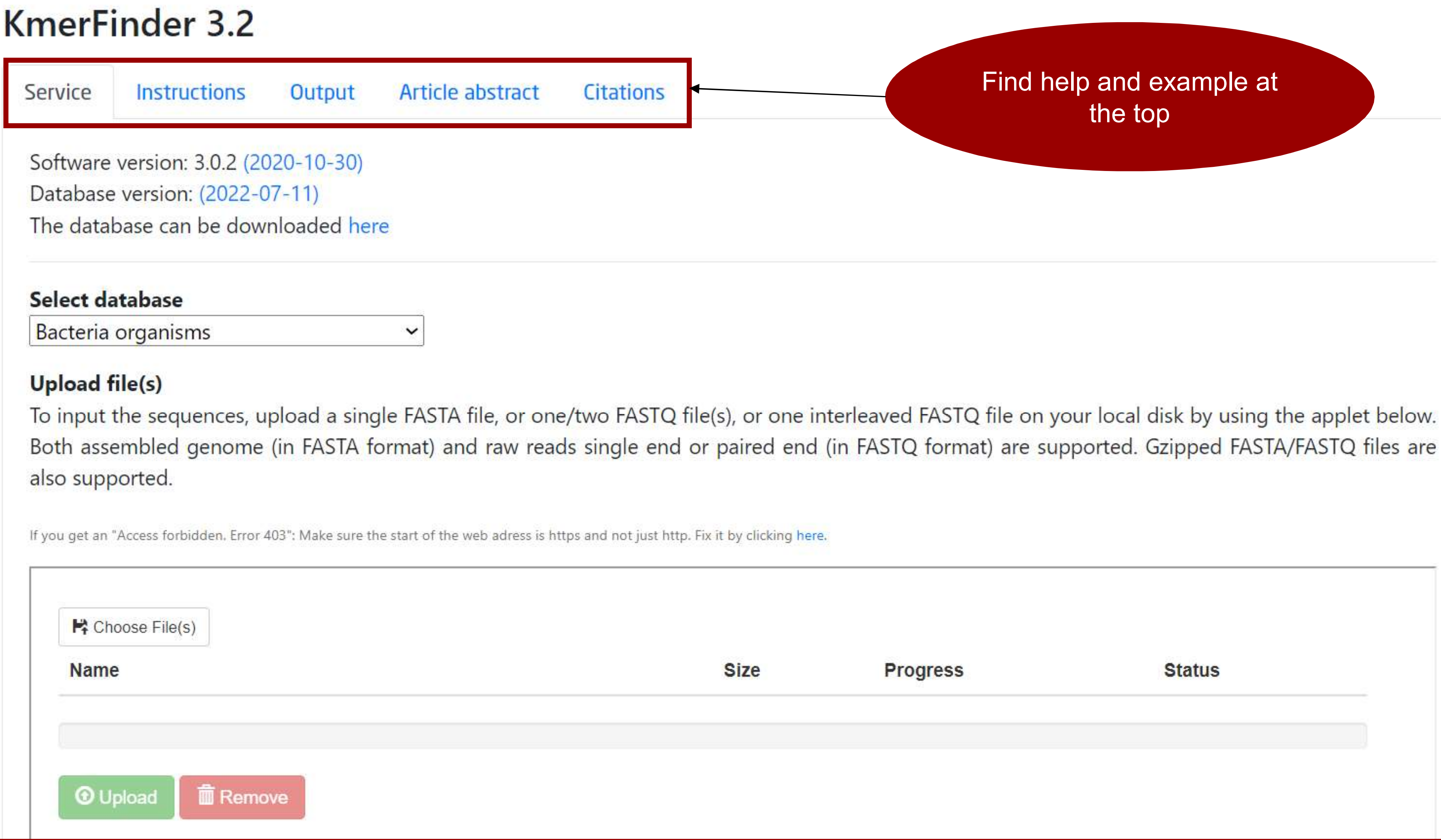
Select database
Bacteria organisms

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



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

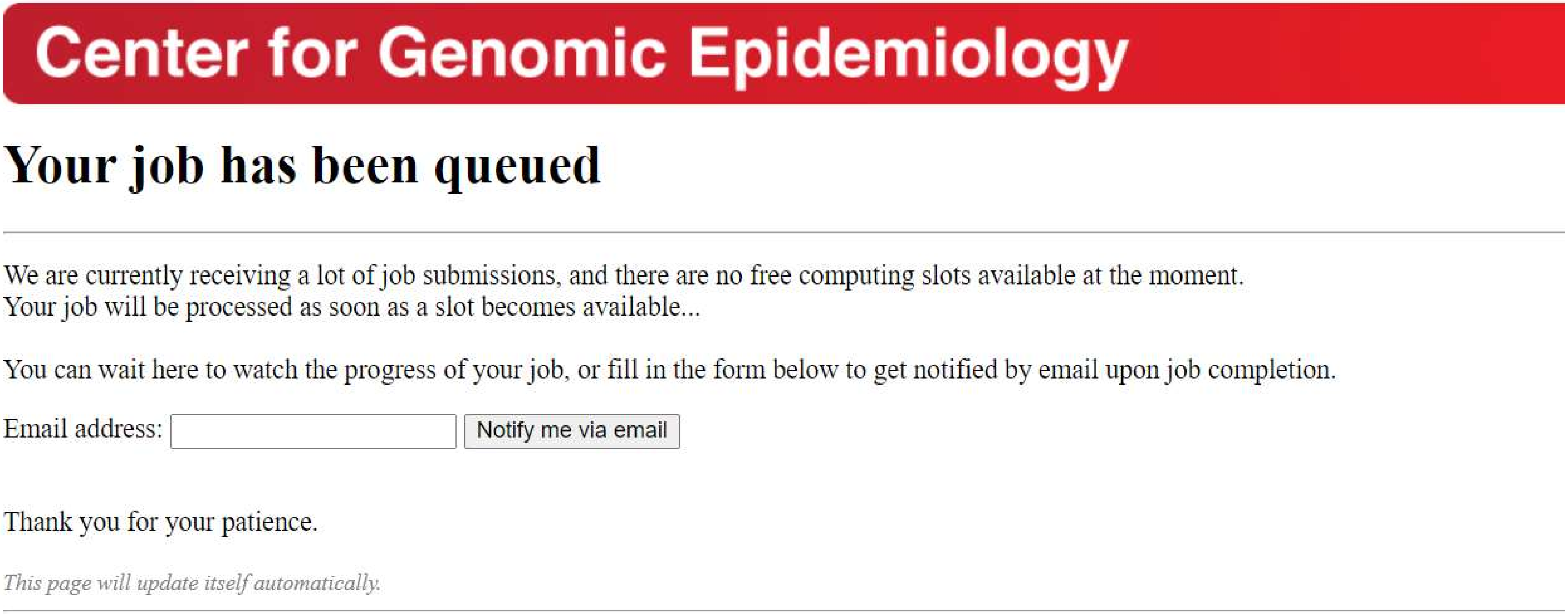
Upload file(s)

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

Find help and example at the top



Center for Genomic Epidemiology

Your job has been queued

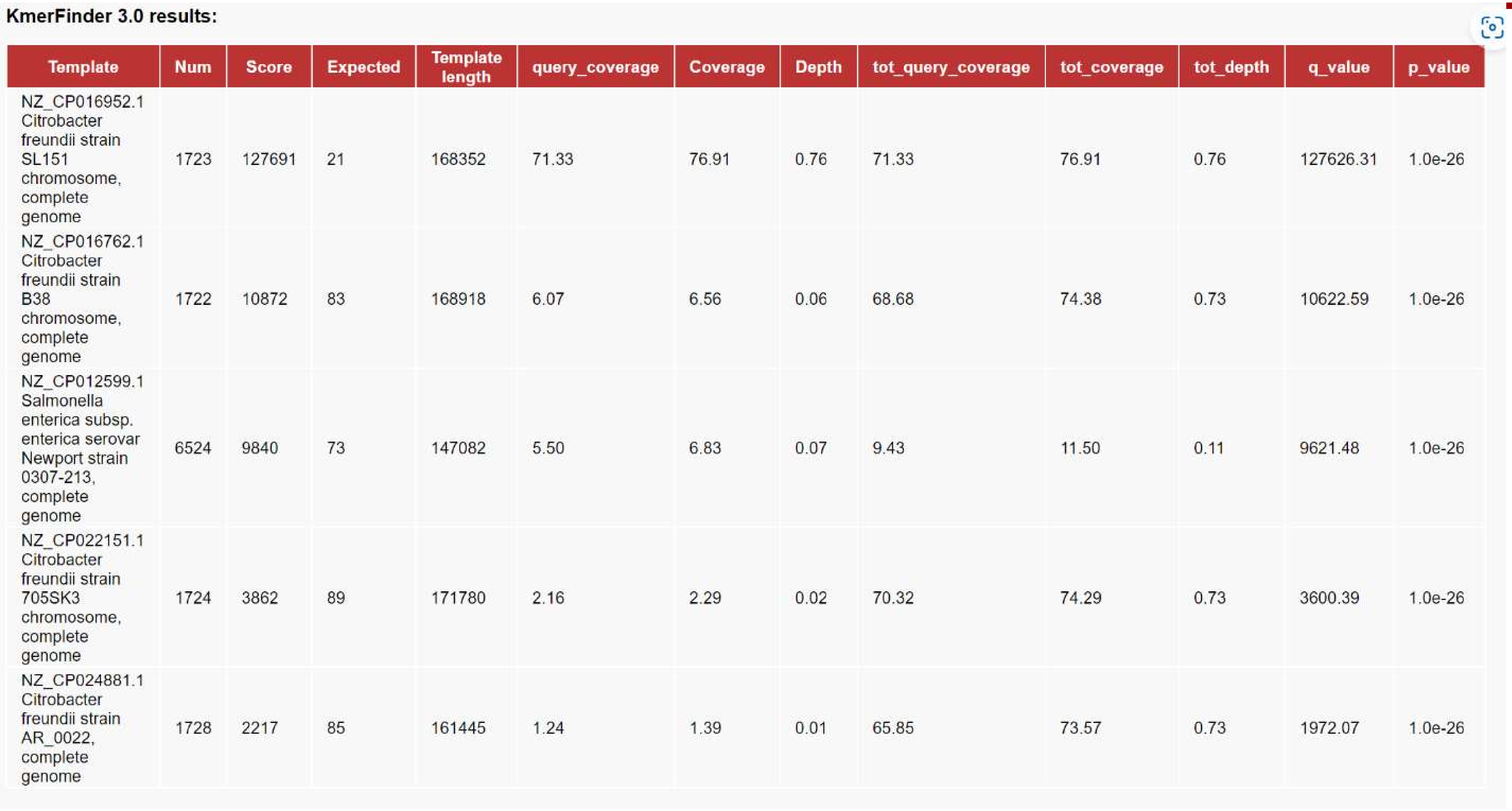
We are currently receiving a lot of job submissions, and there are no free computing slots available at the moment.
Your job will be processed as soon as a slot becomes available...

You can wait here to watch the progress of your job, or fill in the form below to get notified by email upon job completion.

Email address:

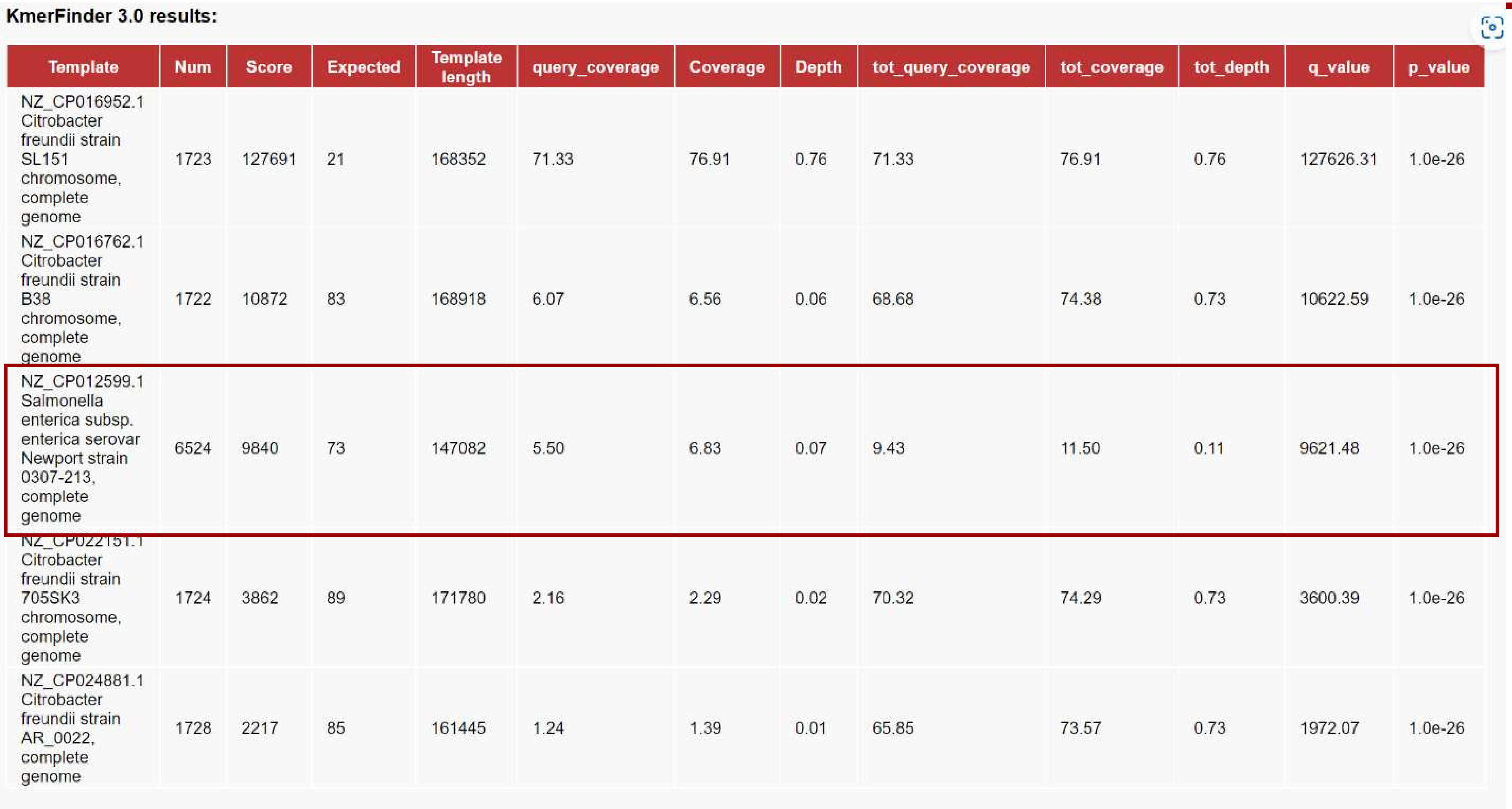
Thank you for your patience.

This page will update itself automatically.



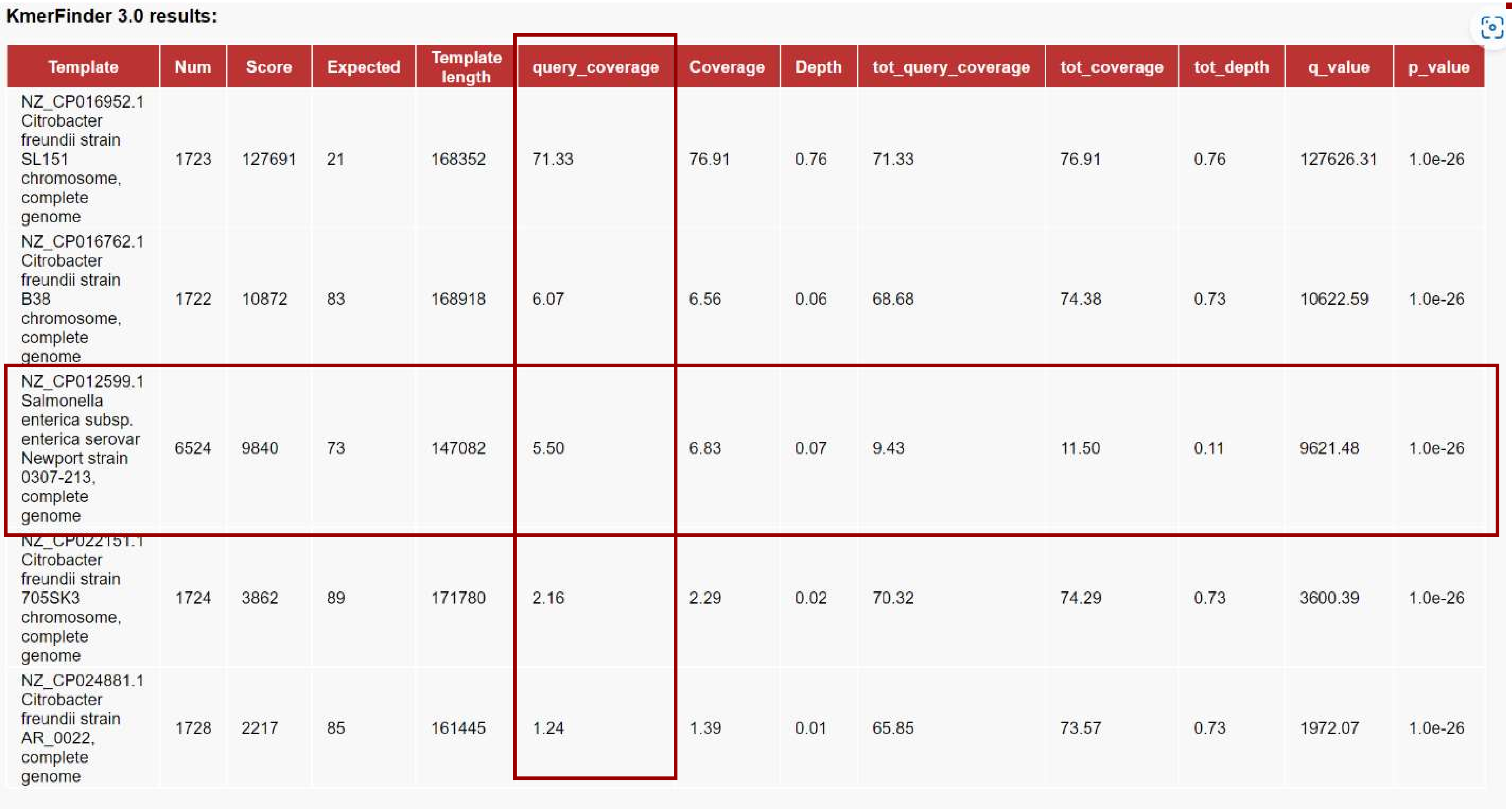
KmerFinder 3.0 results:

Template	Num	Score	Expected	Template length	query_coverage	Coverage	Depth	tot_query_coverage	tot_coverage	tot_depth	q_value	p_value
NZ_CP016952.1 Citrobacter freundii strain SL151 chromosome, complete genome	1723	127691	21	168352	71.33	76.91	0.76	71.33	76.91	0.76	127626.31	1.0e-26
NZ_CP016762.1 Citrobacter freundii strain B38 chromosome, complete genome	1722	10872	83	168918	6.07	6.56	0.06	68.68	74.38	0.73	10622.59	1.0e-26
NZ_CP012599.1 Salmonella enterica subsp. enterica serovar Newport strain 0307-213, complete genome	6524	9840	73	147082	5.50	6.83	0.07	9.43	11.50	0.11	9621.48	1.0e-26
NZ_CP022151.1 Citrobacter freundii strain 705SK3 chromosome, complete genome	1724	3862	89	171780	2.16	2.29	0.02	70.32	74.29	0.73	3600.39	1.0e-26
NZ_CP024881.1 Citrobacter freundii strain AR_0022, complete genome	1728	2217	85	161445	1.24	1.39	0.01	65.85	73.57	0.73	1972.07	1.0e-26



KmerFinder 3.0 results:

Template	Num	Score	Expected	Template length	query_coverage	Coverage	Depth	tot_query_coverage	tot_coverage	tot_depth	q_value	p_value
NZ_CP016952.1 Citrobacter freundii strain SL151 chromosome, complete genome	1723	127691	21	168352	71.33	76.91	0.76	71.33	76.91	0.76	127626.31	1.0e-26
NZ_CP016762.1 Citrobacter freundii strain B38 chromosome, complete genome	1722	10872	83	168918	6.07	6.56	0.06	68.68	74.38	0.73	10622.59	1.0e-26
NZ_CP012599.1 Salmonella enterica subsp. enterica serovar Newport strain 0307-213, complete genome	6524	9840	73	147082	5.50	6.83	0.07	9.43	11.50	0.11	9621.48	1.0e-26
NZ_CP022151.1 Citrobacter freundii strain 705SK3 chromosome, complete genome	1724	3862	89	171780	2.16	2.29	0.02	70.32	74.29	0.73	3600.39	1.0e-26
NZ_CP024881.1 Citrobacter freundii strain AR_0022, complete genome	1728	2217	85	161445	1.24	1.39	0.01	65.85	73.57	0.73	1972.07	1.0e-26



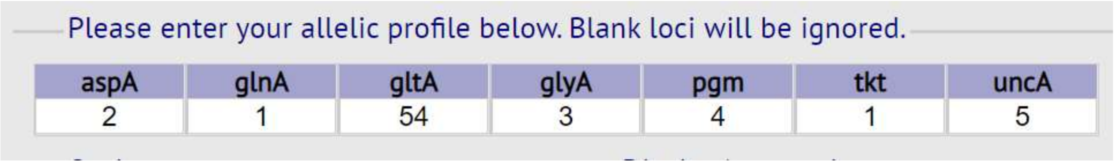
KmerFinder 3.0 results:

Template	Num	Score	Expected	Template length	query_coverage	Coverage	Depth	tot_query_coverage	tot_coverage	tot_depth	q_value	p_value
NZ_CP016952.1 Citrobacter freundii strain SL151 chromosome, complete genome	1723	127691	21	168352	71.33	76.91	0.76	71.33	76.91	0.76	127626.31	1.0e-26
NZ_CP016762.1 Citrobacter freundii strain B38 chromosome, complete genome	1722	10872	83	168918	6.07	6.56	0.06	68.68	74.38	0.73	10622.59	1.0e-26
NZ_CP012599.1 Salmonella enterica subsp. enterica serovar Newport strain 0307-213, complete genome	6524	9840	73	147082	5.50	6.83	0.07	9.43	11.50	0.11	9621.48	1.0e-26
NZ_CP022151.1 Citrobacter freundii strain 705SK3 chromosome, complete genome	1724	3862	89	171780	2.16	2.29	0.02	70.32	74.29	0.73	3600.39	1.0e-26
NZ_CP024881.1 Citrobacter freundii strain AR_0022, complete genome	1728	2217	85	161445	1.24	1.39	0.01	65.85	73.57	0.73	1972.07	1.0e-26

MLST

- MultiLocus Sequence Typing (MLST), is a scheme of 7 genes specific for a species
- The Unique Allele (DNA sequence) for each of these 7 genes are given a number
- Any time a new allele is discovered, its sequence is given a new number and added to the database
- Each unique combination of alleles are given a number, this is the sequence type
- Useful for tracking highly pathogenic lineages, some sequence types are known to cause more severe infections e.g. L. monocytogenes ST6 (Koopmans, 2013)

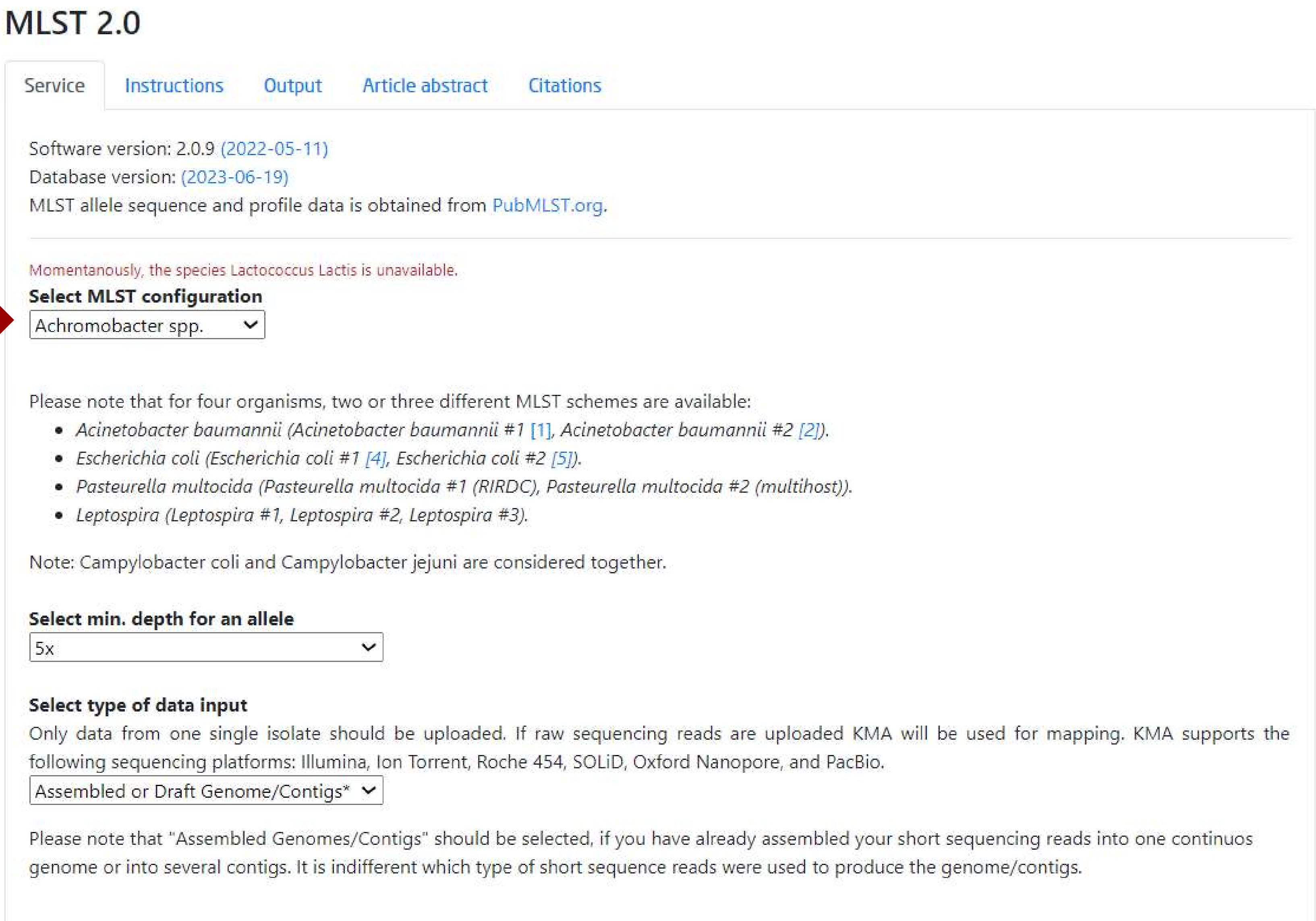
Allele profile for sequence type (ST) 1 in campylobacter jejuni/coli, source: Pubmlst Search by locus combinations (pubmlst.org)



Please enter your allelic profile below. Blank loci will be ignored.

aspA	glnA	gltA	glyA	pgm	tkt	uncA
2	1	54	3	4	1	5

Koopmans MM, Brouwer MC, Bijlsma MW, Bovenkerk S, Keijzers W, van der Ende A, van de Beek D. Listeria monocytogenes sequence type 6 and increased rate of unfavorable outcome in meningitis: epidemiologic cohort study. Clin Infect Dis. 2013 Jul;57(2):247-53. doi: 10.1093/cid/cit250. Epub 2013 Ap 16. PMID: 23592828.



MLST 2.0

Software version: 2.0.9 (2022-05-11)
Database version: (2023-06-19)
MLST allele sequence and profile data is obtained from PubMLST.org.

Momentanously, the species Lactococcus Lactis is unavailable.

Select MLST configuration
Achromobacter spp.

Please note that for four organisms, two or three different MLST schemes are available:

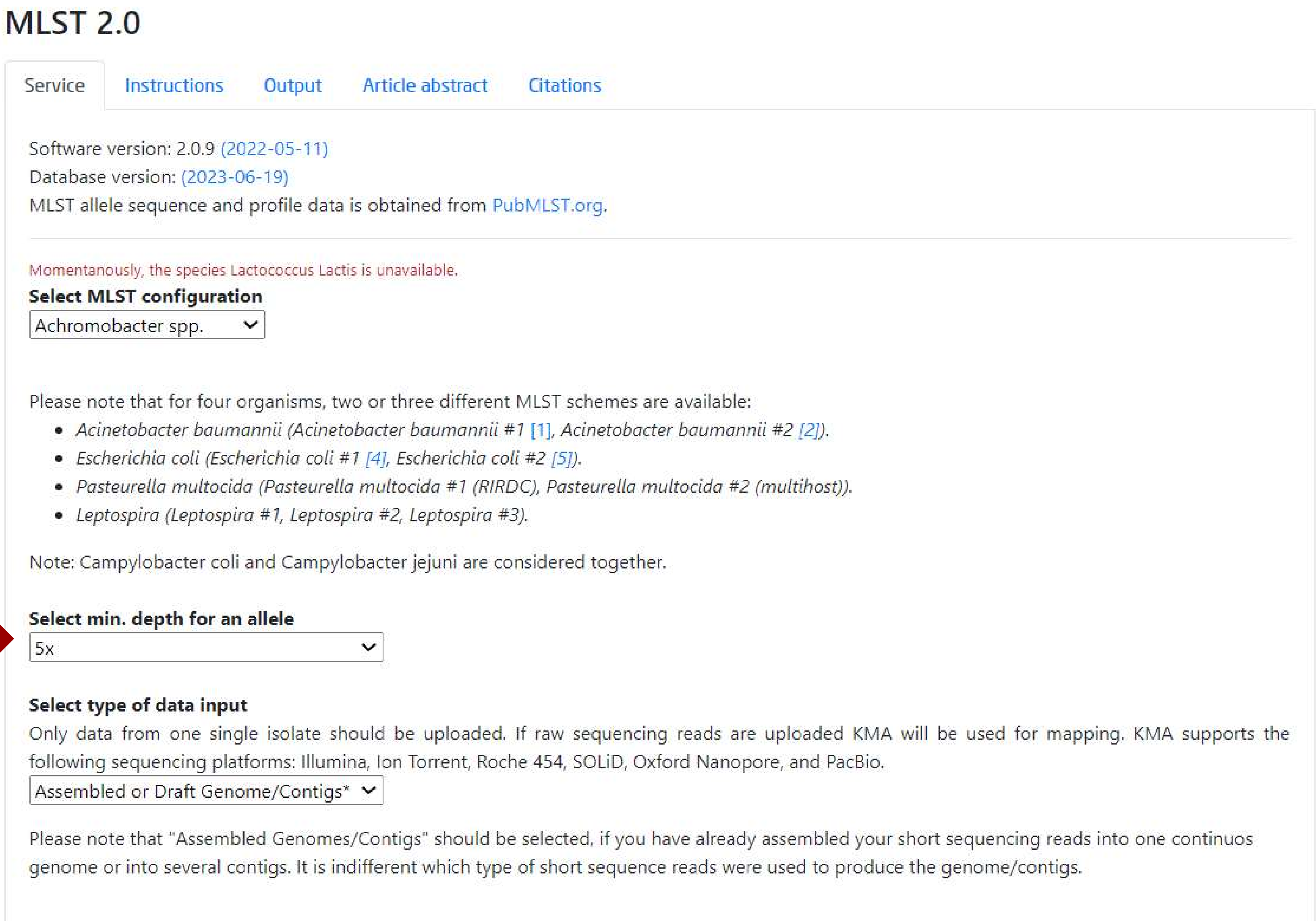
- *Acinetobacter baumannii (Acinetobacter baumannii #1* [1], *Acinetobacter baumannii #2* [2]*).*
- *Escherichia coli (Escherichia coli #1* [4], *Escherichia coli #2* [5]*).*
- *Pasteurella multocida (Pasteurella multocida #1 (RIRDC), Pasteurella multocida #2 (multihost)).*
- *Leptospira (Leptospira #1, Leptospira #2, Leptospira #3).*

Note: Campylobacter coli and Campylobacter jejuni are considered together.

Select min. depth for an allele
5x

Select type of data input
Only data from one single isolate should be uploaded. If raw sequencing reads are uploaded KMA will be used for mapping. KMA supports the following sequencing platforms: Illumina, Ion Torrent, Roche 454, SOLiD, Oxford Nanopore, and PacBio.
Assembled or Draft Genome/Contigs*

Please note that "Assembled Genomes/Contigs" should be selected, if you have already assembled your short sequencing reads into one continuos genome or into several contigs. It is indifferent which type of short sequence reads were used to produce the genome/contigs.



MLST 2.0

Software version: 2.0.9 (2022-05-11)
Database version: (2023-06-19)
MLST allele sequence and profile data is obtained from PubMLST.org.

Momentanously, the species Lactococcus Lactis is unavailable.

Select MLST configuration

Achromobacter spp.

Please note that for four organisms, two or three different MLST schemes are available:

- *Acinetobacter baumannii (Acinetobacter baumannii #1 [1], Acinetobacter baumannii #2 [2]).*
- *Escherichia coli (Escherichia coli #1 [4], Escherichia coli #2 [5]).*
- *Pasteurella multocida (Pasteurella multocida #1 (RIRDC), Pasteurella multocida #2 (multihost)).*
- *Leptospira (Leptospira #1, Leptospira #2, Leptospira #3).*

Note: Campylobacter coli and Campylobacter jejuni are considered together.

Select min. depth for an allele

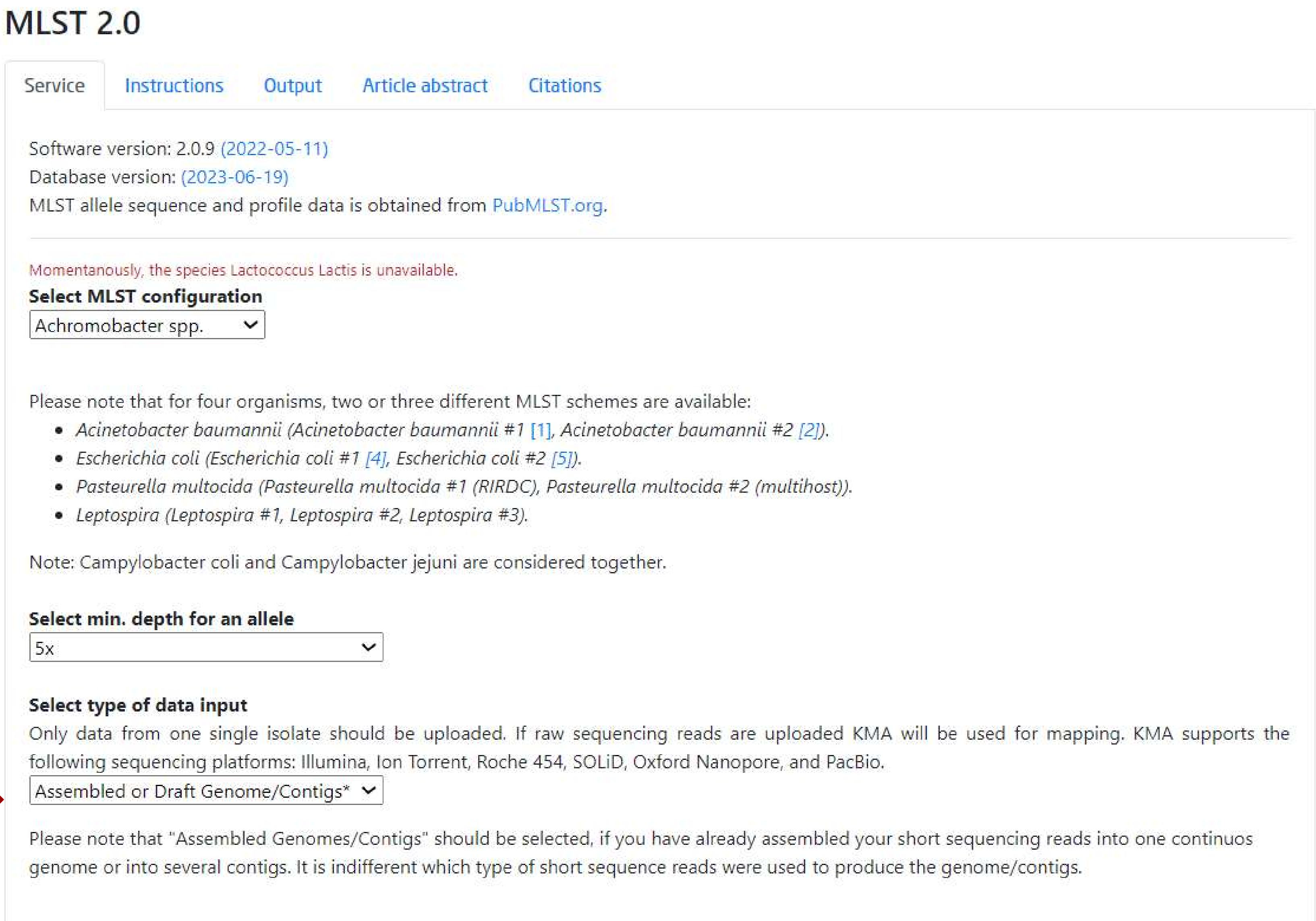
5x

Select type of data input

Only data from one single isolate should be uploaded. If raw sequencing reads are uploaded KMA will be used for mapping. KMA supports the following sequencing platforms: Illumina, Ion Torrent, Roche 454, SOLiD, Oxford Nanopore, and PacBio.

Assembled or Draft Genome/Contigs*

Please note that "Assembled Genomes/Contigs" should be selected, if you have already assembled your short sequencing reads into one continuos genome or into several contigs. It is indifferent which type of short sequence reads were used to produce the genome/contigs.



MLST 2.0

Software version: 2.0.9 (2022-05-11)

Database version: (2023-06-19)

MLST allele sequence and profile data is obtained from PubMLST.org.

Momentanously, the species Lactococcus Lactis is unavailable.

Select MLST configuration

Achromobacter spp.

Please note that for four organisms, two or three different MLST schemes are available:

- *Acinetobacter baumannii (Acinetobacter baumannii #1 [1], Acinetobacter baumannii #2 [2]).*
- *Escherichia coli (Escherichia coli #1 [4], Escherichia coli #2 [5]).*
- *Pasteurella multocida (Pasteurella multocida #1 (RIRDC), Pasteurella multocida #2 (multihost)).*
- *Leptospira (Leptospira #1, Leptospira #2, Leptospira #3).*

Note: Campylobacter coli and Campylobacter jejuni are considered together.

Select min. depth for an allele

5x

Select type of data input

Only data from one single isolate should be uploaded. If raw sequencing reads are uploaded KMA will be used for mapping. KMA supports the following sequencing platforms: Illumina, Ion Torrent, Roche 454, SOLiD, Oxford Nanopore, and PacBio.

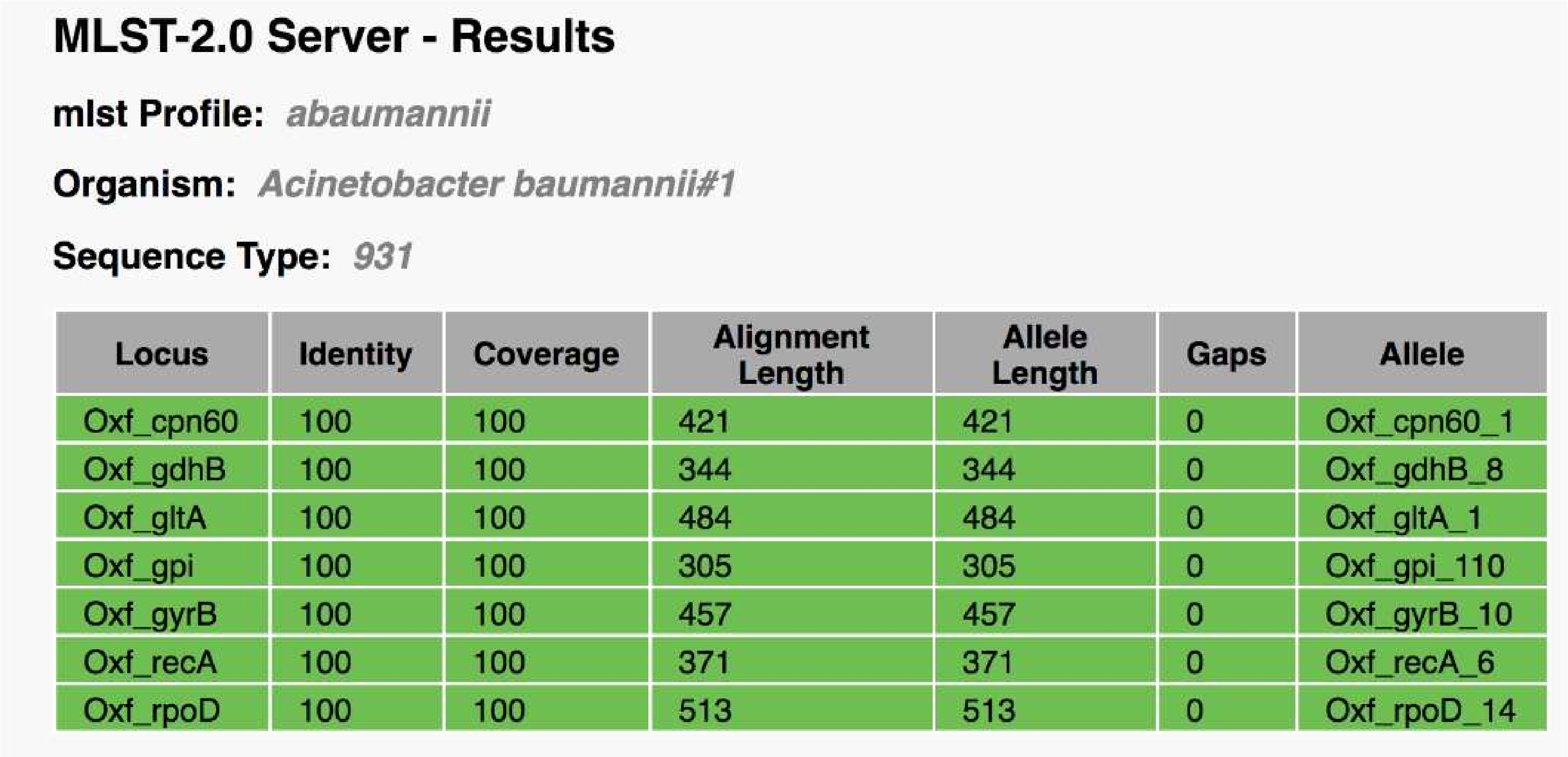
Assembled or Draft Genome/Contigs*

Please note that "Assembled Genomes/Contigs" should be selected, if you have already assembled your short sequencing reads into one continuos genome or into several contigs. It is indifferent which type of short sequence reads were used to produce the genome/contigs.

A matching Sequence type means all alleles had perfect matches in the database

Database is sourced from pubMLST

If any allele does not have a perfect match or is missing the sequence type cannot be determined or marked with a “*” or “!” to indicate an issue



MLST-2.0 Server - Results

mlst Profile: *abaumannii*

Organism: *Acinetobacter baumannii#1*

Sequence Type: *931*

Locus	Identity	Coverage	Alignment Length	Allele Length	Gaps	Allele
Oxf_cpn60	100	100	421	421	0	Oxf_cpn60_1
Oxf_gdhB	100	100	344	344	0	Oxf_gdhB_8
Oxf_gltA	100	100	484	484	0	Oxf_gltA_1
Oxf_gpi	100	100	305	305	0	Oxf_gpi_110
Oxf_gyrB	100	100	457	457	0	Oxf_gyrB_10
Oxf_recA	100	100	371	371	0	Oxf_recA_6
Oxf_rpoD	100	100	513	513	0	Oxf_rpoD_14

Annotation in general

- Attaching biological, chemical or otherwise functional information to a DNA sequence

- Often you are only interested in a limited set of genes, we will look further into antimicrobial resistance (AMR)

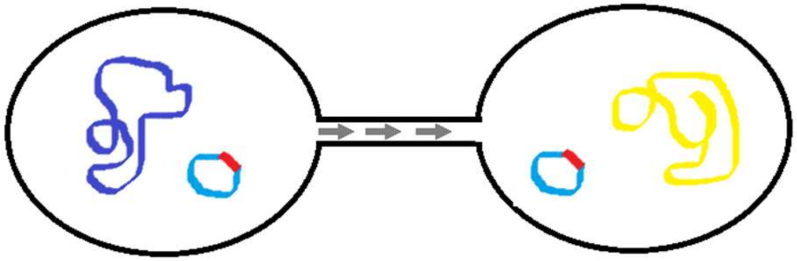
- AMR is a large threat to public health
  - Carried on mobile genetic elements (MGE) -> horizontal gene transfer
  - Estimated 1.27 million people died due to AMR in 2019 and estimated up to 10 million deaths by 2050 (Murray et al., 2019)
  - Development of new drugs is slow (Norrby et al., 2005)

- Murray, Christopher J. L., et al. “Global Burden of Bacterial Antimicrobial Resistance in 2019: a Systematic Analysis.” Lancet, vol. 399, no. 10325, Elsevier B.V., 2022, pp. 629–55, doi:10.1016/S0140-6736(21)02724-0.
- Norrby, S. Ragnar, et al. “Lack of Development of New Antimicrobial Drugs: A Potential Serious Threat to Public Health.” Lancet Infectious Diseases, vol. 5, no. 2, Lancet Publishing Group, 2005, pp. 115–19, doi:10.1016/S1473-3099(05)70086-4.

Genetic basis of Antimicrobial resistance

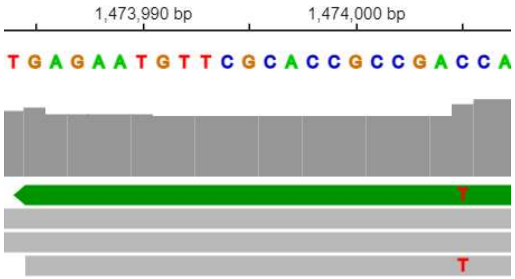
- AMR is conferred by different mechanisms:
  - Acquired resistance genes
  - Mutation
  - (Copy numbers)

- Mobile Genetic Elements (MGE) can transfer resistance genes between isolates closely or distantly related
- Resistance genes tend to aggregate, meaning MGEs often confer resistance to multiple classes
- May integrate into host chromosome



Transfer of plasmid with resistance gene

- Point mutations can confer resistance by various mechanisms:
  - Change the target of a drug, making the strain resistant
  - Upregulate the expression of a gene
  - Downregulate the expression of a gene
  - Change target specificity of protein
  - Usually species specific



Possible point mutation

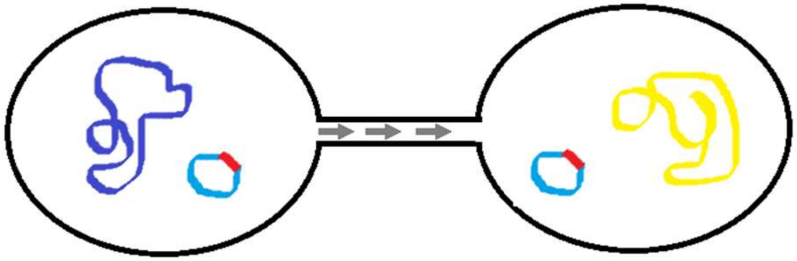
Genetic basis of AMR

- AMR is conferred by different mechanisms:
  - Acquired resistance genes
  - Mutation
  - (Copy numbers)
- MGEs can transfer resistance genes between isolates closely or distantly related
- Resistance genes tend to aggregate, meaning MGEs often confer resistance to multiple classes
- May integrate into host chromosome

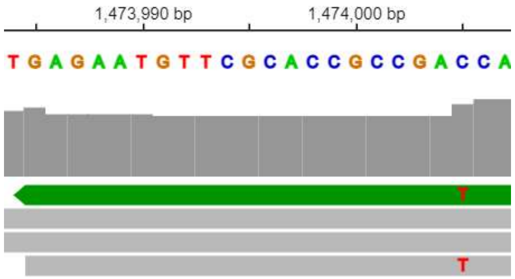
Note!

We also have intrinsic resistance in certain species, e.g. *Mycobacterium tuberculosis* inherently possess erm(37) protecting against macrolides, lincosamide and streptogramin

- Point mutations can confer resistance by various mechanisms:
  - Change the target of a drug, making the strain resistant
  - Upregulate the expression of a gene
  - Downregulate the expression of a gene
  - Change target specificity of protein
  - Usually species specific



Transfer of plasmid with resistance gene



Possible point mutation

AMR tools and databases

- There are multiple tools which all utilize their own and/or each others databases for predicting antimicrobial resistance
  - Resfinder (ResFinder 4.1 (dtu.dk)), AMRfinderplus (Releases · ncbi/amr (github.com)) , CARD (https://card.mcmaster.ca/home), KmerResistance, ARIBA
  - Differences exists due to
    - How the database is created and curated
    - How the tool conducts its search
  - The correct tool/database will likely depend on the type of analysis or workflow you are using
  - Approach results from tools with a critical mindset!

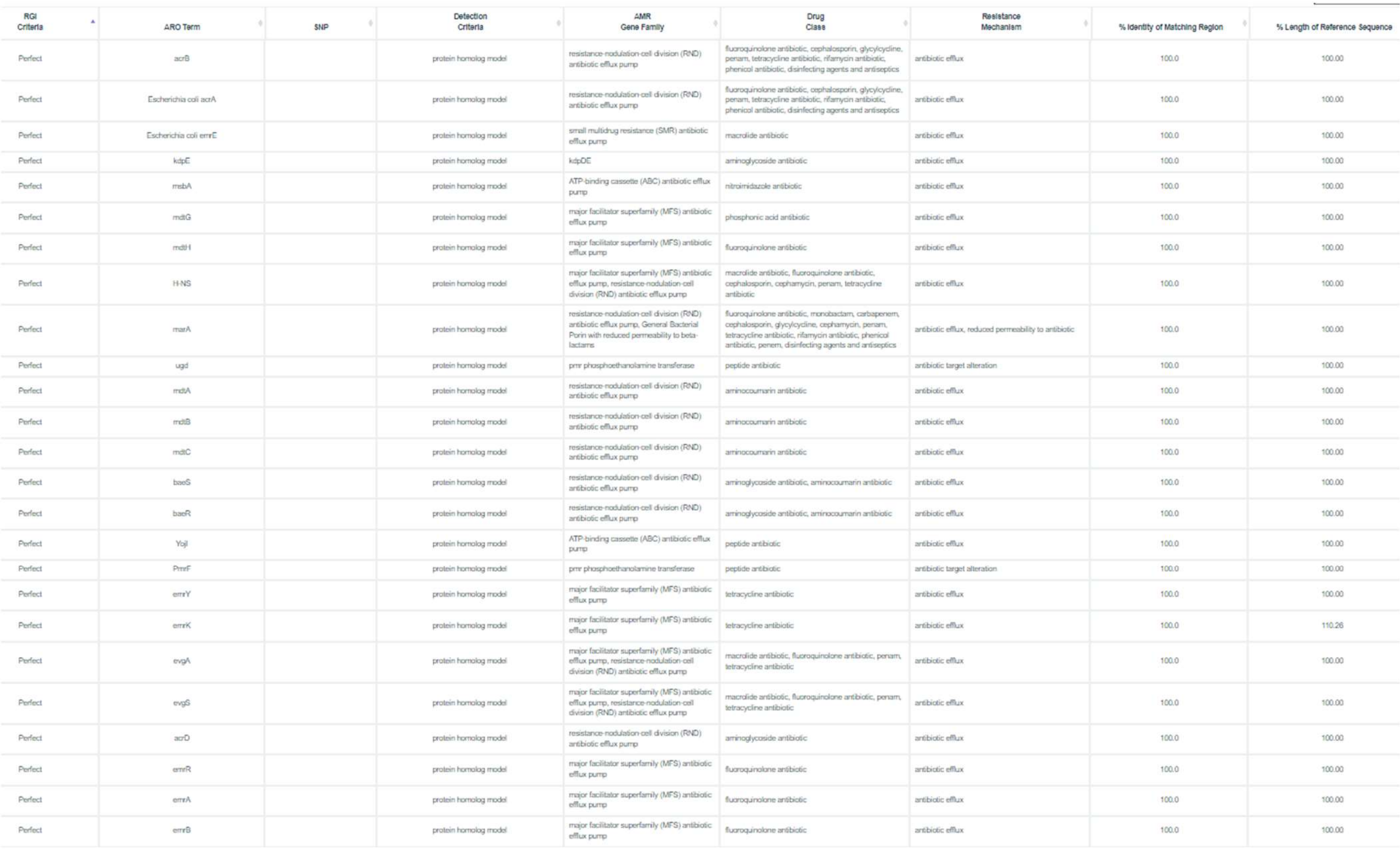
EXAMPLE

CARD output:

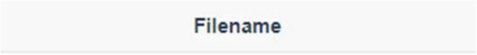
Data was complete genome of E. Coli strain

44 hits in total!

Let us take a closer look



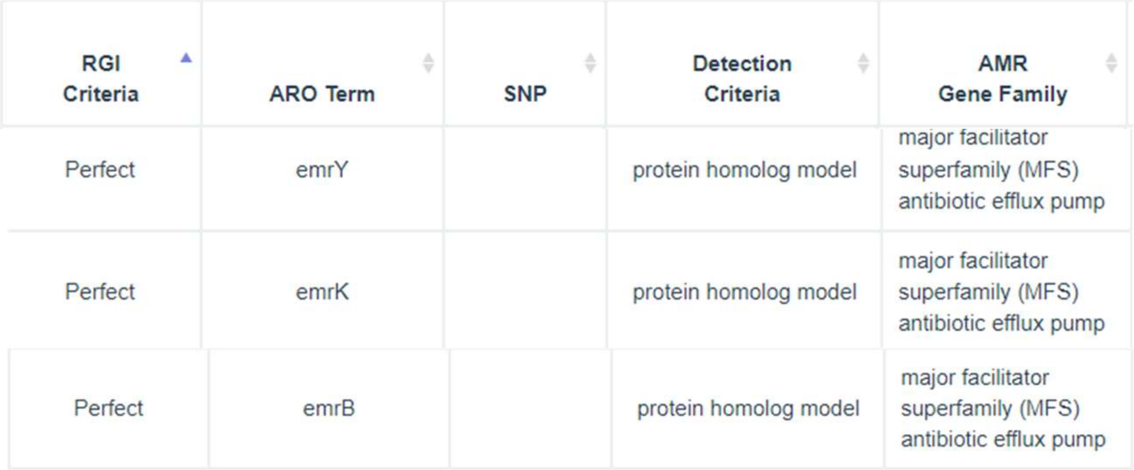
RGI Criteria	ARO Term	SNP	Detection Criteria	AMR Gene Family	Drug Class	Resistance Mechanism	% Identity of Matching Region	% Length of Reference Sequence
Perfect	acrB		protein homolog model	resistance-nodulation-cell division (RND) antibiotic efflux pump	fluoroquinolone antibiotic, cephalosporin, glycylcycline, penam, tetracycline antibiotic, rifamycin antibiotic, phenicol antibiotic, disinfecting agents and antiseptics	antibiotic efflux	100.0	100.00
Perfect	Escherichia coli acrA		protein homolog model	resistance-nodulation-cell division (RND) antibiotic efflux pump	fluoroquinolone antibiotic, cephalosporin, glycylcycline, penam, tetracycline antibiotic, rifamycin antibiotic, phenicol antibiotic, disinfecting agents and antiseptics	antibiotic efflux	100.0	100.00
Perfect	Escherichia coli emrE		protein homolog model	small multidrug resistance (SMR) antibiotic efflux pump	macrolide antibiotic	antibiotic efflux	100.0	100.00
Perfect	kdpE		protein homolog model	kdpDE	aminoglycoside antibiotic	antibiotic efflux	100.0	100.00
Perfect	msbA		protein homolog model	ATP-binding cassette (ABC) antibiotic efflux pump	nitroimidazole antibiotic	antibiotic efflux	100.0	100.00
Perfect	mdtG		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump	phosphonic acid antibiotic	antibiotic efflux	100.0	100.00
Perfect	mdtH		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump	fluoroquinolone antibiotic	antibiotic efflux	100.0	100.00
Perfect	H-NS		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump, resistance-nodulation-cell division (RND) antibiotic efflux pump	macrolide antibiotic, fluoroquinolone antibiotic, cephalosporin, cephamycin, penam, tetracycline antibiotic	antibiotic efflux	100.0	100.00
Perfect	marA		protein homolog model	resistance-nodulation-cell division (RND) antibiotic efflux pump, General Bacterial Porin with reduced permeability to beta-lactams	fluoroquinolone antibiotic, monobactam, carbapenem, cephalosporin, glycylcycline, cephamycin, penam, tetracycline antibiotic, rifamycin antibiotic, phenicol antibiotic, penem, disinfecting agents and antiseptics	antibiotic efflux, reduced permeability to antibiotic	100.0	100.00
Perfect	ugd		protein homolog model	pmr phosphoethanolamine transferase	peptide antibiotic	antibiotic target alteration	100.0	100.00
Perfect	mdtA		protein homolog model	resistance-nodulation-cell division (RND) antibiotic efflux pump	aminocoumarin antibiotic	antibiotic efflux	100.0	100.00
Perfect	mdtB		protein homolog model	resistance-nodulation-cell division (RND) antibiotic efflux pump	aminocoumarin antibiotic	antibiotic efflux	100.0	100.00
Perfect	mdtC		protein homolog model	resistance-nodulation-cell division (RND) antibiotic efflux pump	aminocoumarin antibiotic	antibiotic efflux	100.0	100.00
Perfect	baeS		protein homolog model	resistance-nodulation-cell division (RND) antibiotic efflux pump	aminoglycoside antibiotic, aminocoumarin antibiotic	antibiotic efflux	100.0	100.00
Perfect	baeR		protein homolog model	resistance-nodulation-cell division (RND) antibiotic efflux pump	aminoglycoside antibiotic, aminocoumarin antibiotic	antibiotic efflux	100.0	100.00
Perfect	YojI		protein homolog model	ATP-binding cassette (ABC) antibiotic efflux pump	peptide antibiotic	antibiotic efflux	100.0	100.00
Perfect	PmrF		protein homolog model	pmr phosphoethanolamine transferase	peptide antibiotic	antibiotic target alteration	100.0	100.00
Perfect	emrY		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump	tetracycline antibiotic	antibiotic efflux	100.0	100.00
Perfect	emrK		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump	tetracycline antibiotic	antibiotic efflux	100.0	110.26
Perfect	evgA		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump, resistance-nodulation-cell division (RND) antibiotic efflux pump	macrolide antibiotic, fluoroquinolone antibiotic, penam, tetracycline antibiotic	antibiotic efflux	100.0	100.00
Perfect	evgS		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump, resistance-nodulation-cell division (RND) antibiotic efflux pump	macrolide antibiotic, fluoroquinolone antibiotic, penam, tetracycline antibiotic	antibiotic efflux	100.0	100.00
Perfect	acrD		protein homolog model	resistance-nodulation-cell division (RND) antibiotic efflux pump	aminoglycoside antibiotic	antibiotic efflux	100.0	100.00
Perfect	emrR		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump	fluoroquinolone antibiotic	antibiotic efflux	100.0	100.00
Perfect	emrA		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump	fluoroquinolone antibiotic	antibiotic efflux	100.0	100.00
Perfect	emrB		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump	fluoroquinolone antibiotic	antibiotic efflux	100.0	100.00



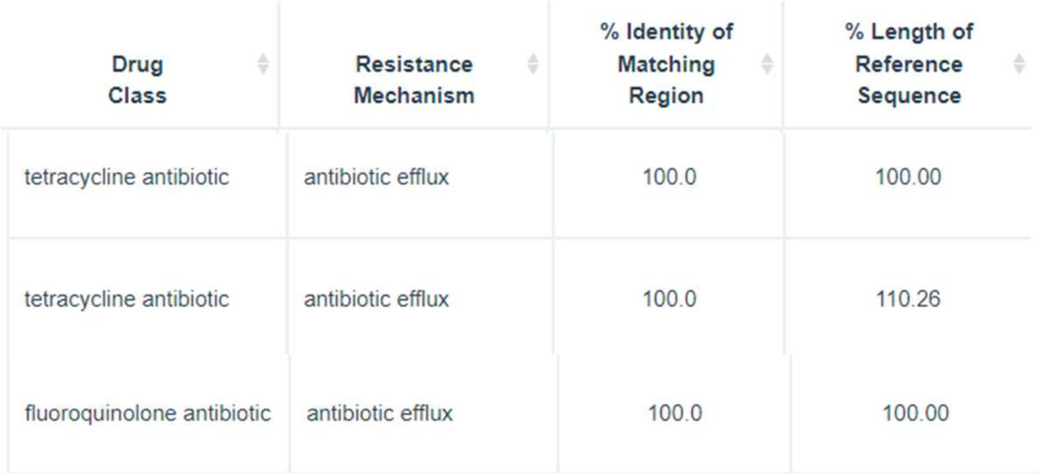
Filename
GCF_000005845.2_ASM584v2_genomic

EXAMPLE CARD output:

- EmrY, emrK and emrB
- Perfect hits!
    - Expect for emrK, ID and COV are 100%
- Should we expect resistance to tetracycline and fluoroquinolones in this isolate?



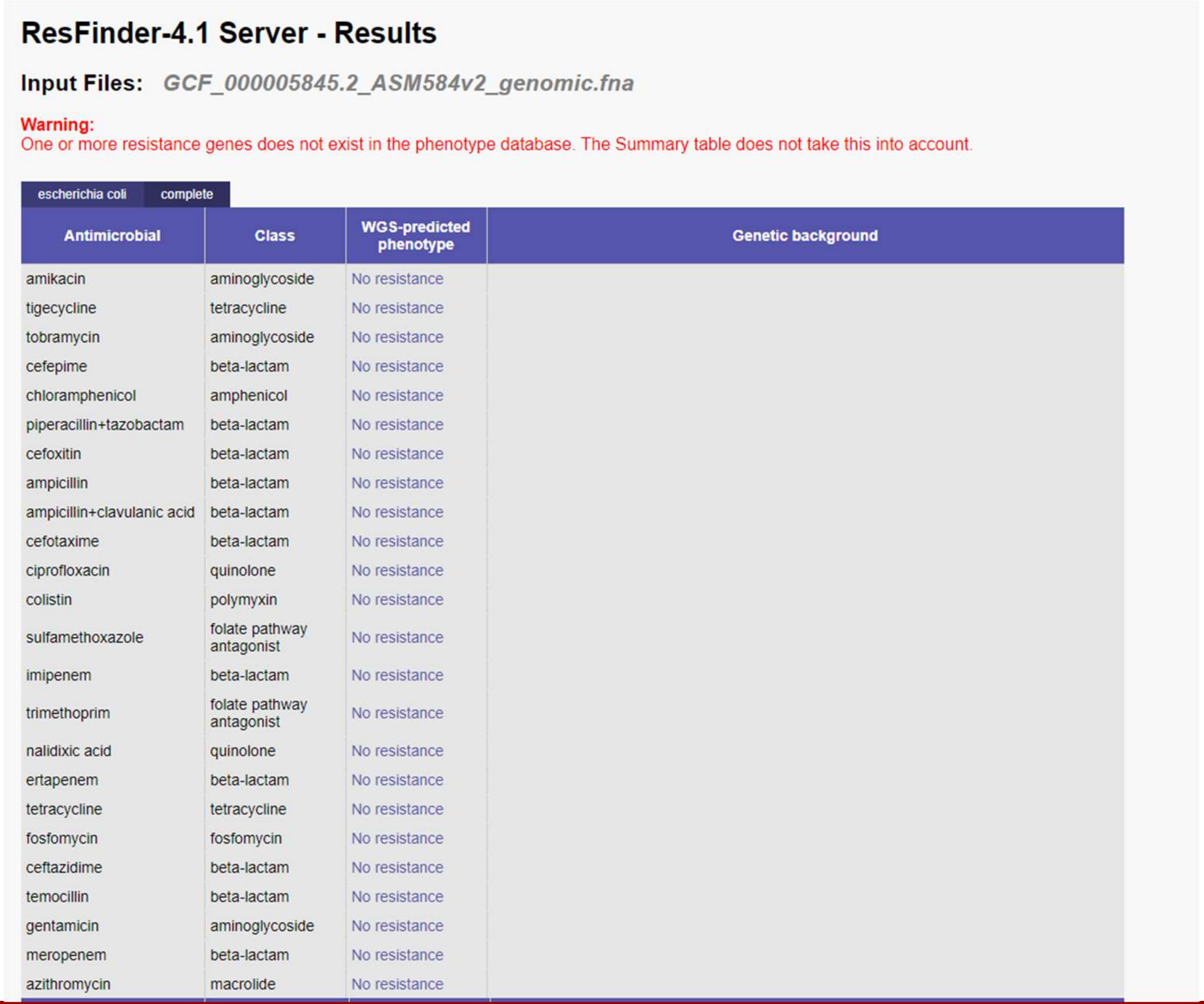
RGI Criteria	ARO Term	SNP	Detection Criteria	AMR Gene Family
Perfect	emrY		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump
Perfect	emrK		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump
Perfect	emrB		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump



Drug Class	Resistance Mechanism	% Identity of Matching Region	% Length of Reference Sequence
tetracycline antibiotic	antibiotic efflux	100.0	100.00
tetracycline antibiotic	antibiotic efflux	100.0	110.26
fluoroquinolone antibiotic	antibiotic efflux	100.0	100.00

Lets try a different tool for the strain: ResFinder

- No resistance at all?



ResFinder-4.1 Server - Results

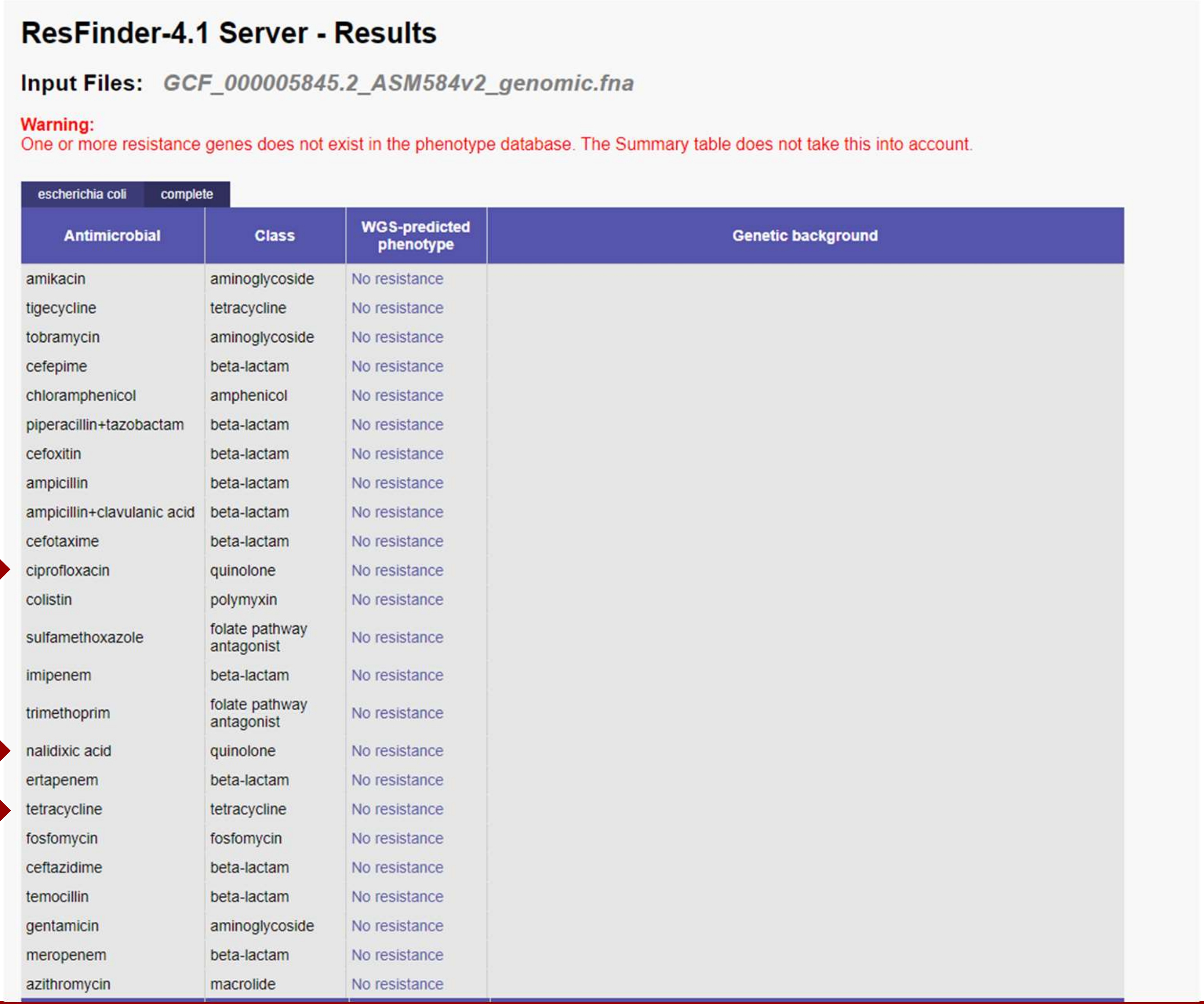
Input Files: *GCF_000005845.2_ASM584v2_genomic.fna*

Warning:
One or more resistance genes does not exist in the phenotype database. The Summary table does not take this into account.

Antimicrobial	Class	WGS-predicted phenotype	Genetic background
amikacin	aminoglycoside	No resistance	
tigecycline	tetracycline	No resistance	
tobramycin	aminoglycoside	No resistance	
cefepime	beta-lactam	No resistance	
chloramphenicol	amphenicol	No resistance	
piperacillin+tazobactam	beta-lactam	No resistance	
cefoxitin	beta-lactam	No resistance	
ampicillin	beta-lactam	No resistance	
ampicillin+clavulanic acid	beta-lactam	No resistance	
cefotaxime	beta-lactam	No resistance	
ciprofloxacin	quinolone	No resistance	
colistin	polymyxin	No resistance	
sulfamethoxazole	folate pathway antagonist	No resistance	
imipenem	beta-lactam	No resistance	
trimethoprim	folate pathway antagonist	No resistance	
nalidixic acid	quinolone	No resistance	
ertapenem	beta-lactam	No resistance	
tetracycline	tetracycline	No resistance	
fosfomycin	fosfomycin	No resistance	
ceftazidime	beta-lactam	No resistance	
temocillin	beta-lactam	No resistance	
gentamicin	aminoglycoside	No resistance	
meropenem	beta-lactam	No resistance	
azithromycin	macrolide	No resistance	

Lets try a different tool for the strain: ResFinder

- No resistance at all?
- No resistance to tetracycline or quinolones?



ResFinder-4.1 Server - Results

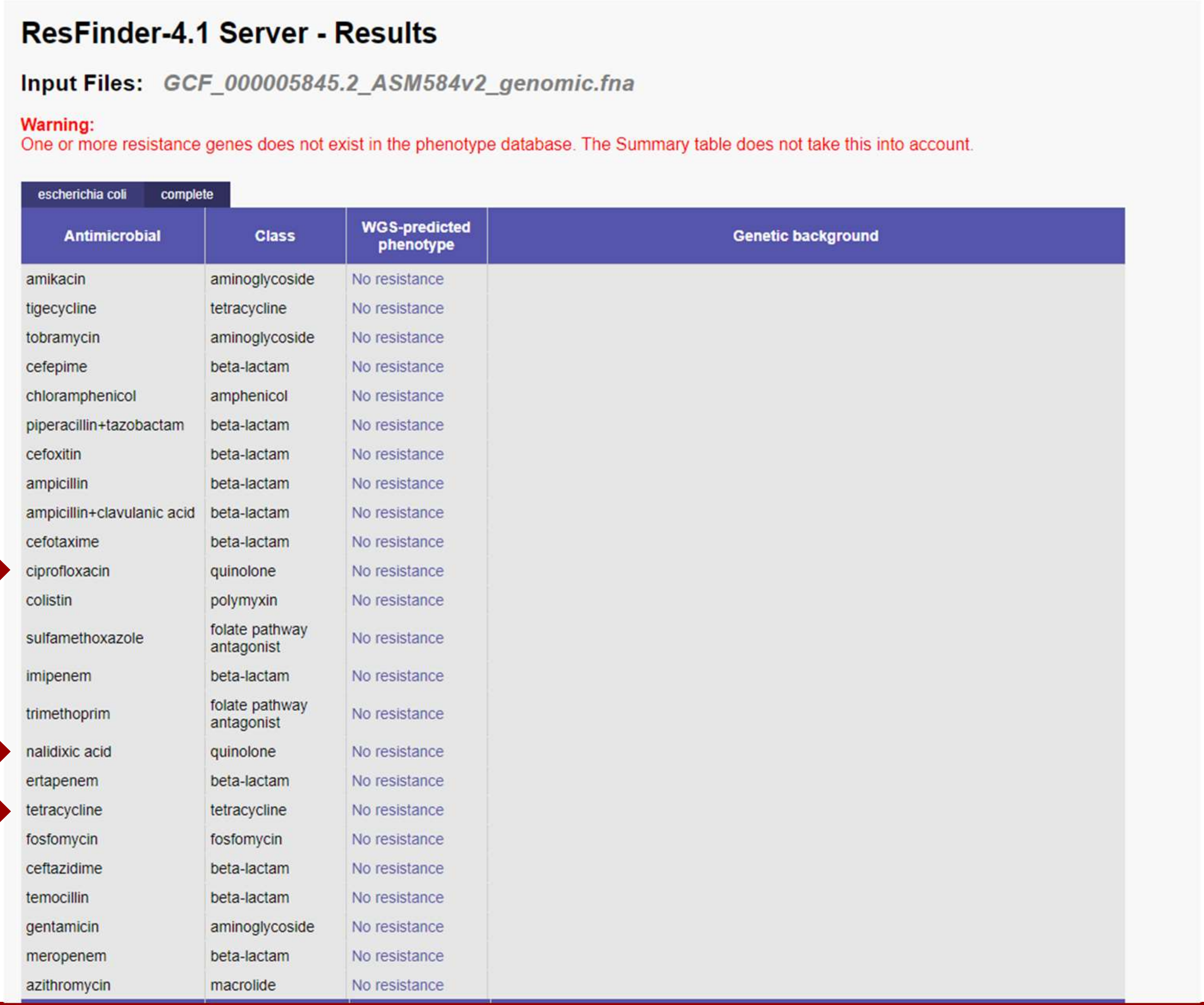
Input Files: *GCF_000005845.2_ASM584v2_genomic.fna*

Warning:
One or more resistance genes does not exist in the phenotype database. The Summary table does not take this into account.

Antimicrobial	Class	WGS-predicted phenotype	Genetic background
amikacin	aminoglycoside	No resistance	
tigecycline	tetracycline	No resistance	
tobramycin	aminoglycoside	No resistance	
cefepime	beta-lactam	No resistance	
chloramphenicol	amphenicol	No resistance	
piperacillin+tazobactam	beta-lactam	No resistance	
cefoxitin	beta-lactam	No resistance	
ampicillin	beta-lactam	No resistance	
ampicillin+clavulanic acid	beta-lactam	No resistance	
cefotaxime	beta-lactam	No resistance	
ciprofloxacin	quinolone	No resistance	
colistin	polymyxin	No resistance	
sulfamethoxazole	folate pathway antagonist	No resistance	
imipenem	beta-lactam	No resistance	
trimethoprim	folate pathway antagonist	No resistance	
nalidixic acid	quinolone	No resistance	
ertapenem	beta-lactam	No resistance	
tetracycline	tetracycline	No resistance	
fosfomycin	fosfomycin	No resistance	
ceftazidime	beta-lactam	No resistance	
temocillin	beta-lactam	No resistance	
gentamicin	aminoglycoside	No resistance	
meropenem	beta-lactam	No resistance	
azithromycin	macrolide	No resistance	

Lets try a different tool for the strain: ResFinder

- No resistance at all?
- No resistance to tetracycline or quinolones?
- One tool gives 44 hits, another gives 0 what is the truth?



ResFinder-4.1 Server - Results

Input Files: *GCF_000005845.2_ASM584v2_genomic.fna*

Warning:
One or more resistance genes does not exist in the phenotype database. The Summary table does not take this into account.

escherichia coli | complete

Antimicrobial	Class	WGS-predicted phenotype	Genetic background
amikacin	aminoglycoside	No resistance	
tigecycline	tetracycline	No resistance	
tobramycin	aminoglycoside	No resistance	
cefepime	beta-lactam	No resistance	
chloramphenicol	amphenicol	No resistance	
piperacillin+tazobactam	beta-lactam	No resistance	
cefoxitin	beta-lactam	No resistance	
ampicillin	beta-lactam	No resistance	
ampicillin+clavulanic acid	beta-lactam	No resistance	
cefotaxime	beta-lactam	No resistance	
ciprofloxacin	quinolone	No resistance	
colistin	polymyxin	No resistance	
sulfamethoxazole	folate pathway antagonist	No resistance	
imipenem	beta-lactam	No resistance	
trimethoprim	folate pathway antagonist	No resistance	
nalidixic acid	quinolone	No resistance	
ertapenem	beta-lactam	No resistance	
tetracycline	tetracycline	No resistance	
fosfomycin	fosfomycin	No resistance	
ceftazidime	beta-lactam	No resistance	
temocillin	beta-lactam	No resistance	
gentamicin	aminoglycoside	No resistance	
meropenem	beta-lactam	No resistance	
azithromycin	macrolide	No resistance	

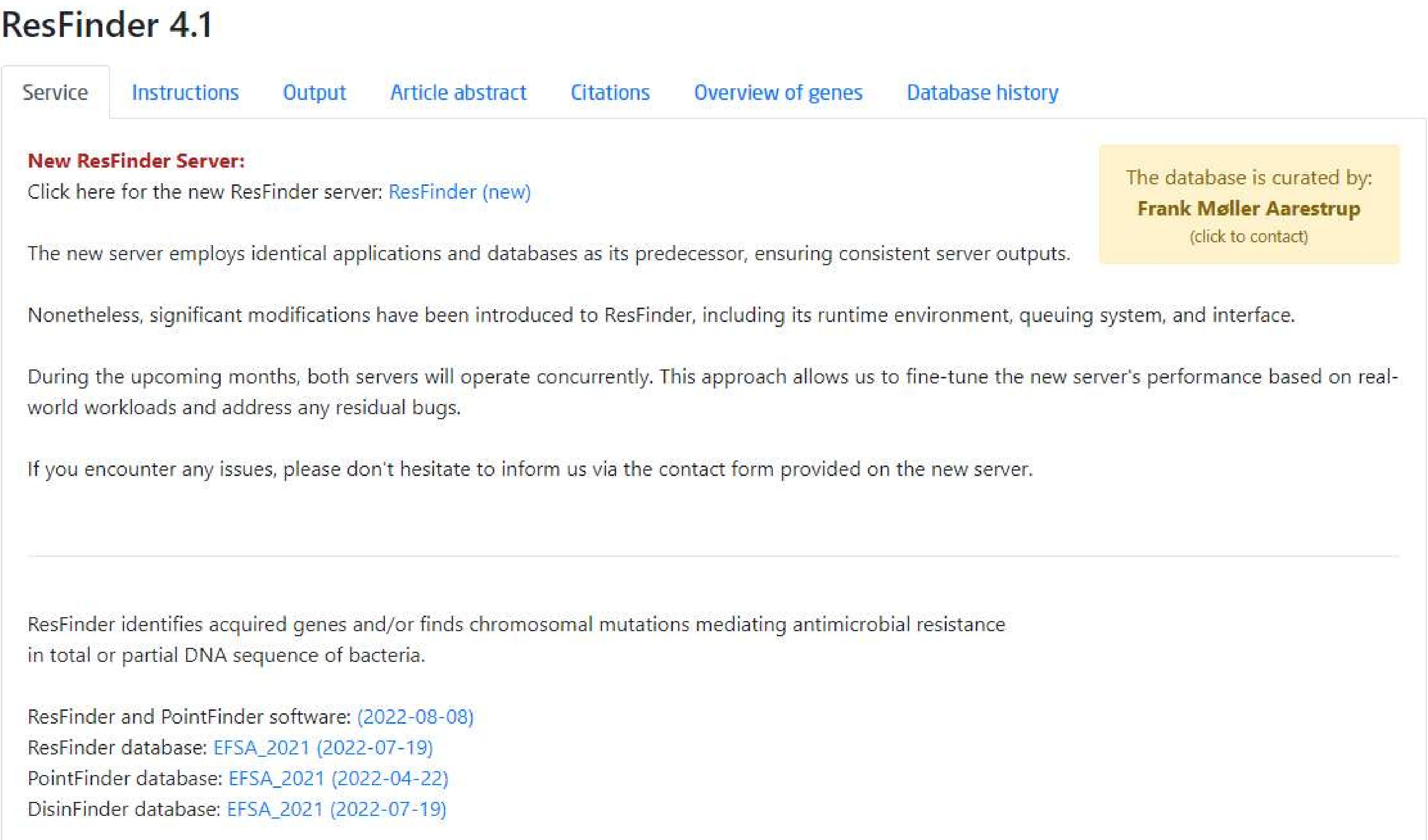
Differences in output example

- The strain run in this example is a standard laboratory strain E. coli K-12 substrain MG1655

- It is not expected to have any phenotypic resistance to tetracycline (Zhang et al., 2022)
  - Not actually expected to have any particular phenotypic resistance different from wild-type e. coli

- If run on AMRfinderplus, no resistance genes are found either.

- Approach databases with care and select based on your scope
  - How does results translate to the laboratory, genotypic =/= phenotypic
  - How much expertise is demanded to utilize findings
  - What is the aim of your analysis



ResFinder 4.1

New ResFinder Server:
Click here for the new ResFinder server: ResFinder (new)

The new server employs identical applications and databases as its predecessor, ensuring consistent server outputs.

The database is curated by:
Frank Møller Aarestrup
(click to contact)

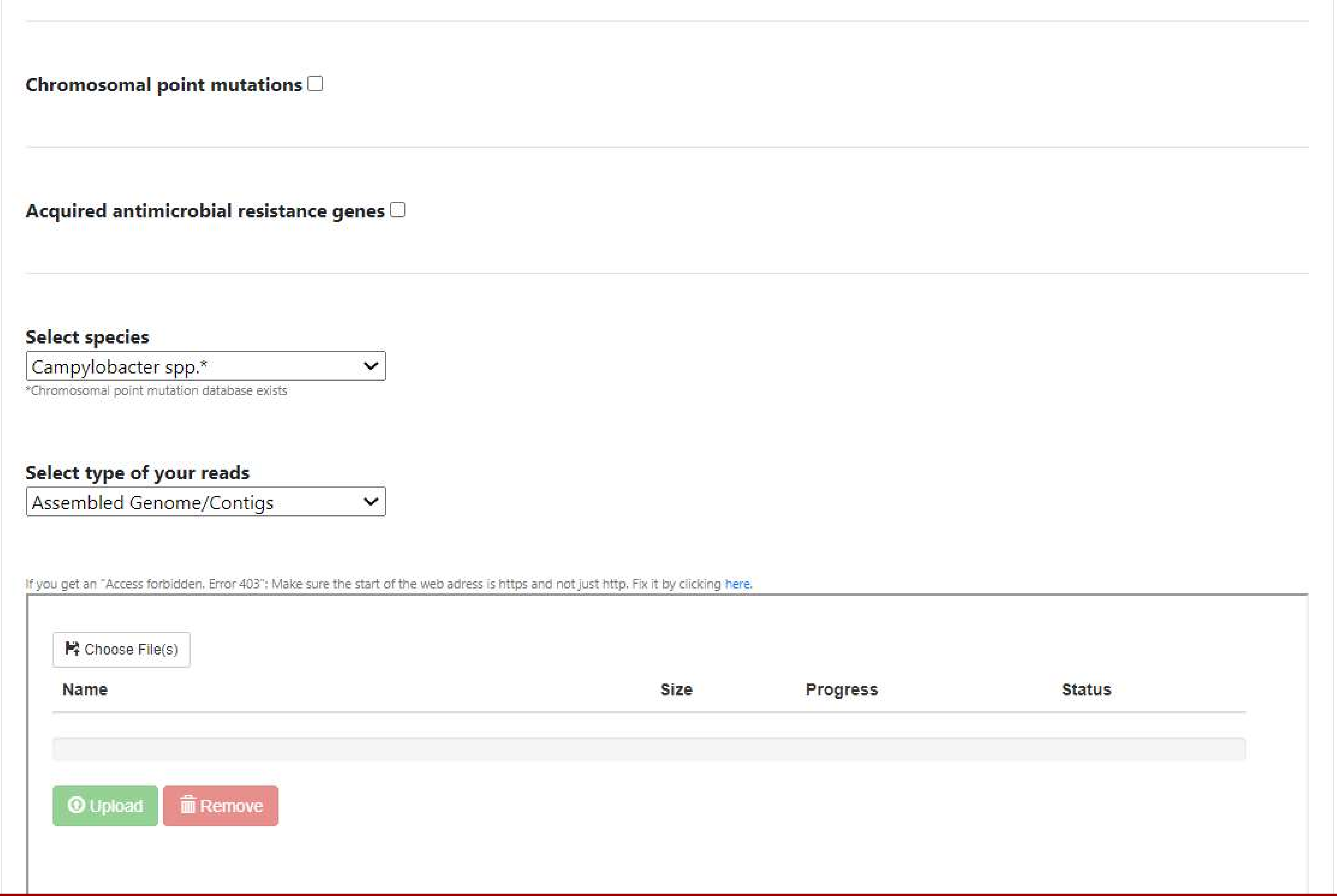
Nonetheless, significant modifications have been introduced to ResFinder, including its runtime environment, queuing system, and interface.

During the upcoming months, both servers will operate concurrently. This approach allows us to fine-tune the new server's performance based on real-world workloads and address any residual bugs.

If you encounter any issues, please don't hesitate to inform us via the contact form provided on the new server.

ResFinder identifies acquired genes and/or finds chromosomal mutations mediating antimicrobial resistance
in total or partial DNA sequence of bacteria.

ResFinder and PointFinder software: (2022-08-08)
ResFinder database: EFSA_2021 (2022-07-19)
PointFinder database: EFSA_2021 (2022-04-22)
DisinFinder database: EFSA_2021 (2022-07-19)



Chromosomal point mutations ☐

Acquired antimicrobial resistance genes ☐

Select species

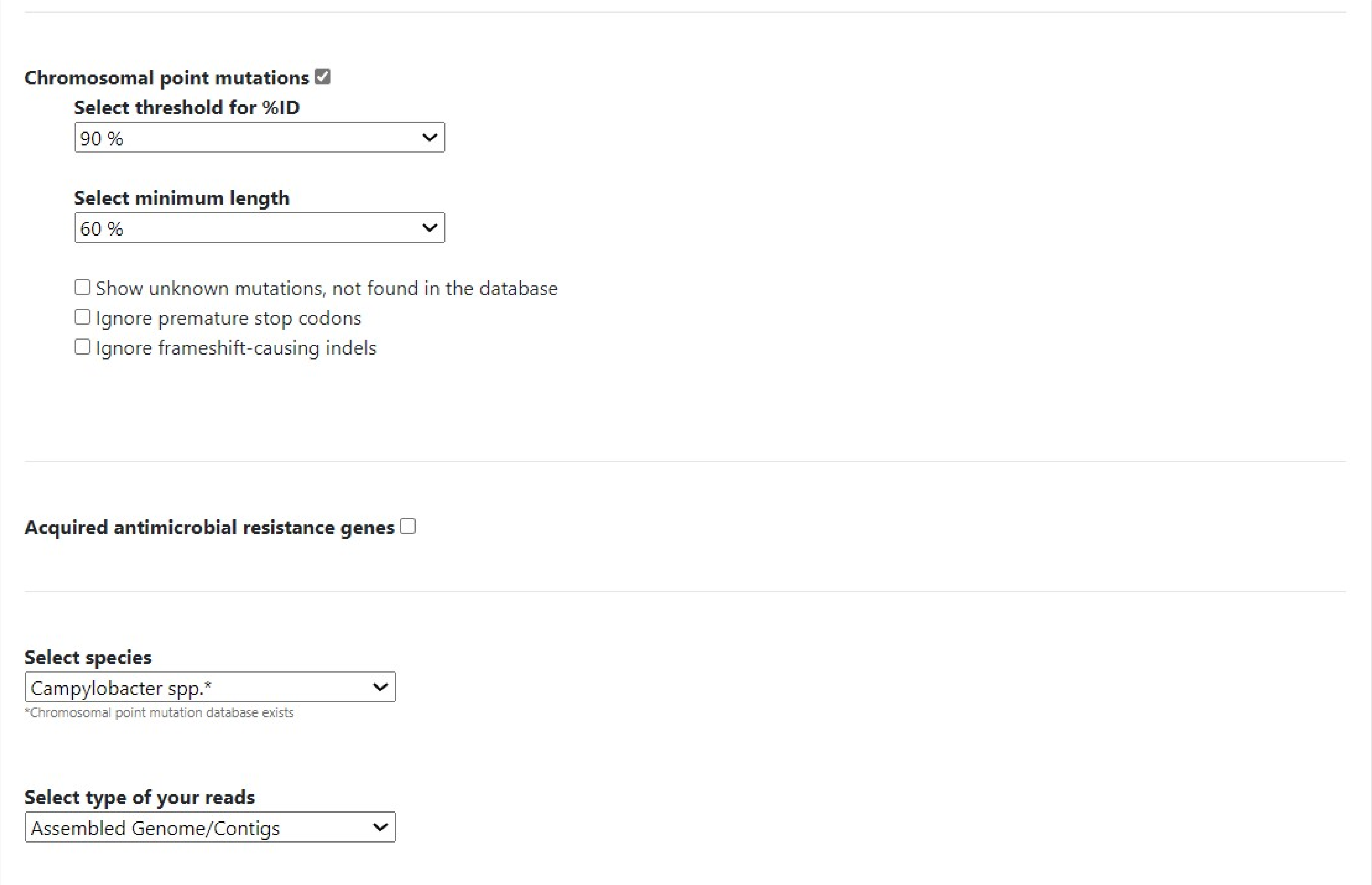
Campylobacter spp.*

*Chromosomal point mutation database exists

Select type of your reads

Assembled Genome/Contigs

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.



Chromosomal point mutations ☑

Select threshold for %ID

90 %

Select minimum length

60 %

☐ Show unknown mutations, not found in the database

☐ Ignore premature stop codons

☐ Ignore frameshift-causing indels

Acquired antimicrobial resistance genes ☐

Select species

Campylobacter spp.*

*Chromosomal point mutation database exists

Select type of your reads

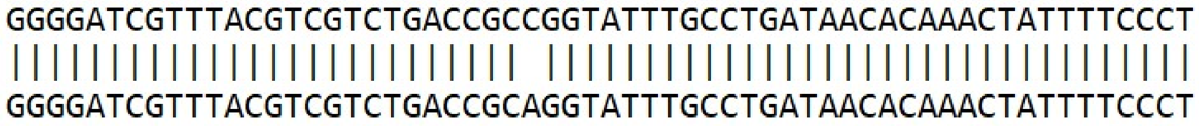
Assembled Genome/Contigs

Sequence identity

- Another term we encounter in the cge tools is % identity (ID)

- The identity describes how many bases of the aligned sequences are identical

- Given the alignment:



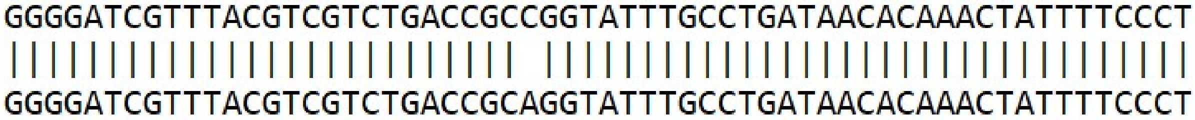
```
GGGGATCGTTTACGTCGTCTGACCGCCGGTATTTGCCTGATAACACAAACTATTTTCCCT
|||||||||||||||||||||||||| |||||||||||||||||||||||||||||||||
GGGGATCGTTTACGTCGTCTGACCGCAGGTATTTGCCTGATAACACAAACTATTTTCCCT
```

Sequence identity

- Another term we encounter in the cge tools is % identity (ID)

- The identity describes how many bases of the aligned sequences are identical

- Given the alignment:

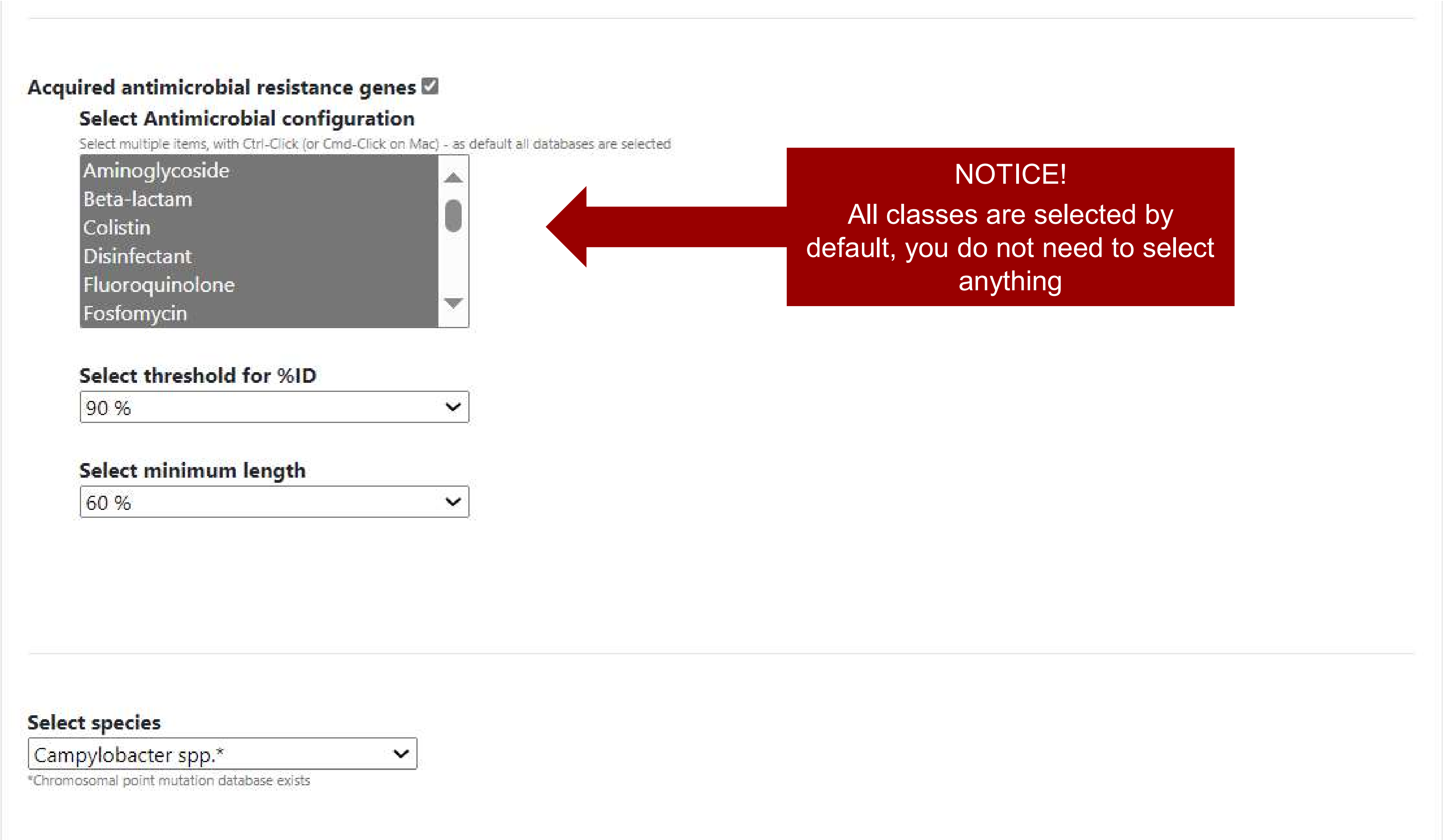


```
GGGGATCGTTTACGTCGTCTGACCGCCGGTATTTGCCTGATAACACAAACTATTTTCCCT
||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||||
GGGGATCGTTTACGTCGTCTGACCGCAGGTATTTGCCTGATAACACAAACTATTTTCCCT
```

- Sequence length 60

- Matches 59

- %ID = 59/60*100% = 98.3%



Acquired antimicrobial resistance genes ☑

Select Antimicrobial configuration

Select multiple items, with Ctrl-Click (or Cmd-Click on Mac) - as default all databases are selected

- Aminoglycoside
- Beta-lactam
- Colistin
- Disinfectant
- Fluoroquinolone
- Fosfomycin

Select threshold for %ID: 90 %

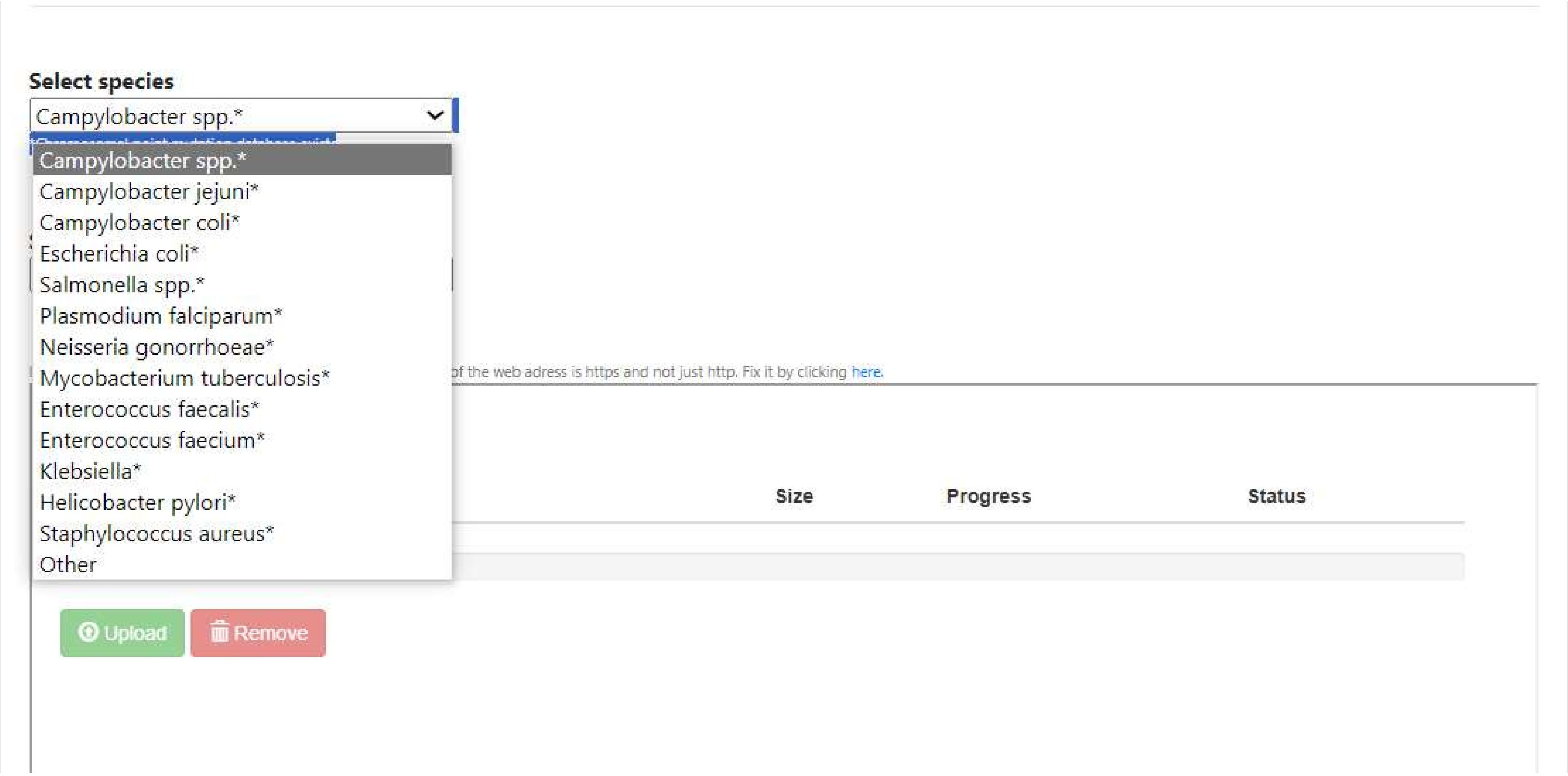
Select minimum length: 60 %

Select species: Campylobacter spp.*

*Chromosomal point mutation database exists

NOTICE!

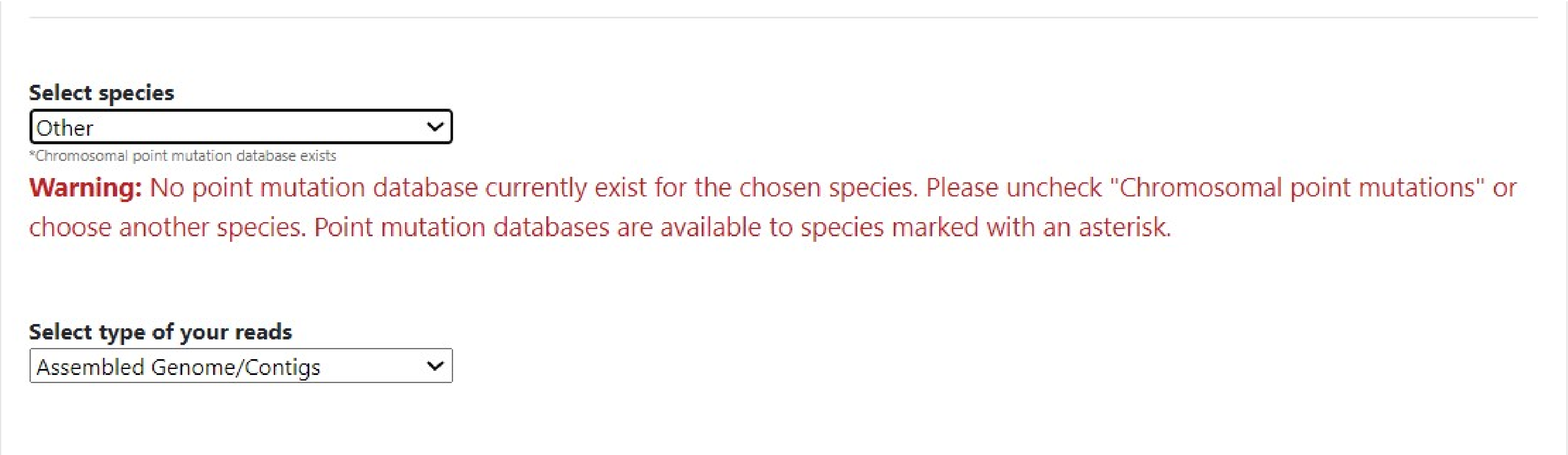
All classes are selected by default, you do not need to select anything



Select species

Campylobacter spp.*

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



Select species

Other

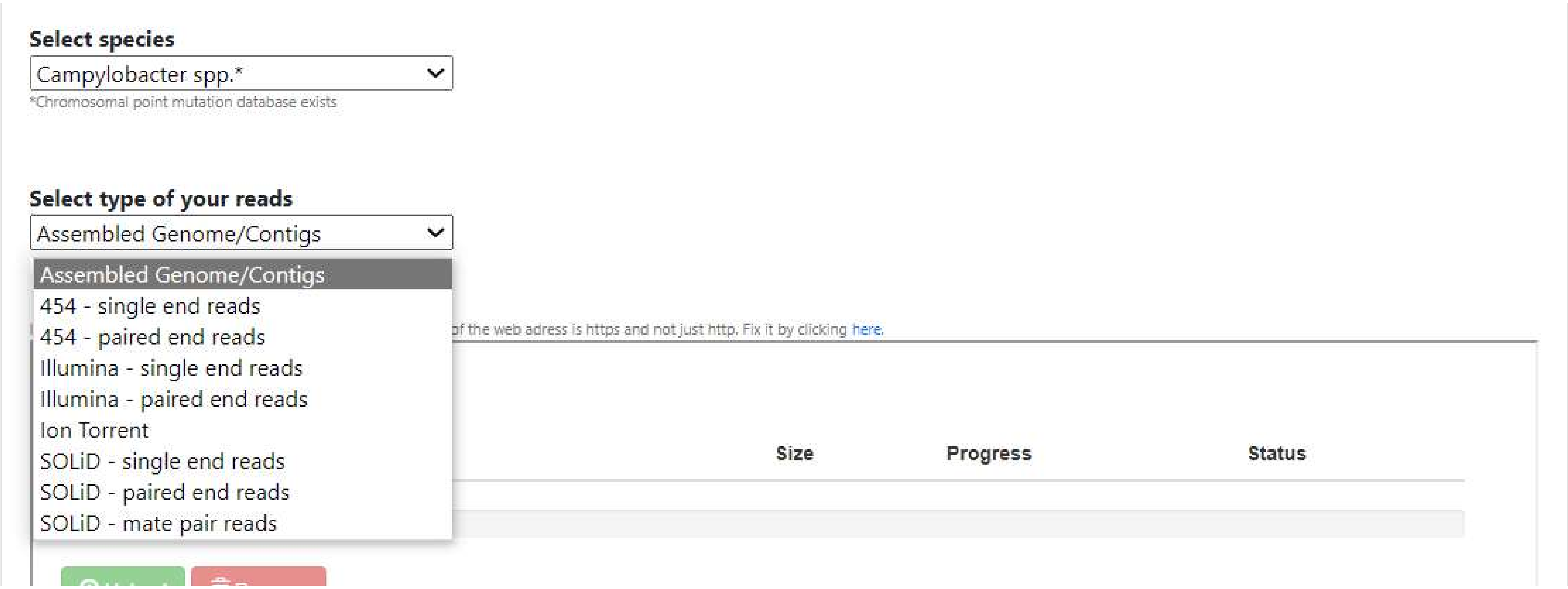
*Chromosomal point mutation database exists

Warning: No point mutation database currently exist for the chosen species. Please uncheck "Chromosomal point mutations" or choose another species. Point mutation databases are available to species marked with an asterisk.

Select type of your reads

Assembled Genome/Contigs

If you cannot find a suitable option among species you can chose “other”, but chromosomal mutations cannot be selected if running with other selected



Select species

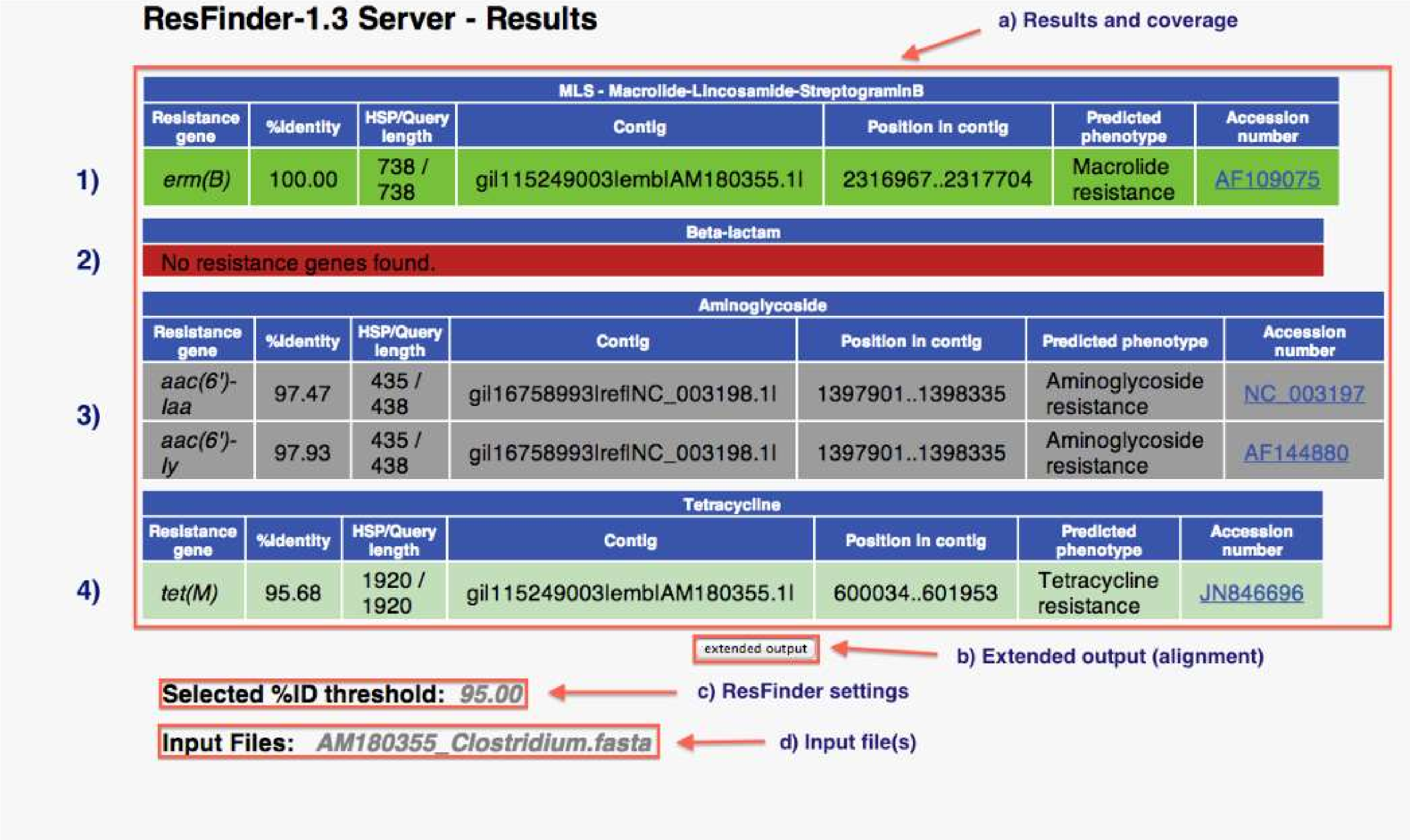
Campylobacter spp.*

*Chromosomal point mutation database exists

Select type of your reads

Assembled Genome/Contigs

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



ResFinder-1.3 Server - Results

a) Results and coverage

1)

MLS - Macrolide-Lincosamide-StreptograminB						
Resistance gene	%Identity	HSP/Query length	Contig	Position in contig	Predicted phenotype	Accession number
erm(B)	100.00	738 / 738	gil115249003lemblAM180355.1l	2316967..2317704	Macrolide resistance	AF109075

2)

Beta-lactam
No resistance genes found.

3)

Aminoglycoside						
Resistance gene	%Identity	HSP/Query length	Contig	Position in contig	Predicted phenotype	Accession number
aac(6')-Iaa	97.47	435 / 438	gil16758993lreflNC_003198.1l	1397901..1398335	Aminoglycoside resistance	NC_003197
aac(6')-Iy	97.93	435 / 438	gil16758993lreflNC_003198.1l	1397901..1398335	Aminoglycoside resistance	AF144880

4)

Tetracycline						
Resistance gene	%Identity	HSP/Query length	Contig	Position in contig	Predicted phenotype	Accession number
tet(M)	95.68	1920 / 1920	gil115249003lemblAM180355.1l	600034..601953	Tetracycline resistance	JN846696

extended output

b) Extended output (alignment)

Selected %ID threshold: *95.00*

c) ResFinder settings

Input Files: *AM180355_Clostridium.fasta*

d) Input file(s)

or

A case of tet(M)

1_______________________________1920 tet(M)_FN433596

1____________801 Hit 2

793____________________1920 Hit 1

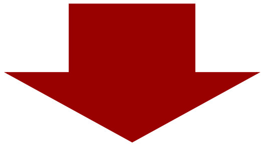
1___________801___________(5000 bp inserted)________793________1920

A case of tet(M)

1______________________________1920 tet(M)_FN433596

1____________801 Hit 2

793____________________1920 Hit 1



1___________801___________(mobile element)________793________1920

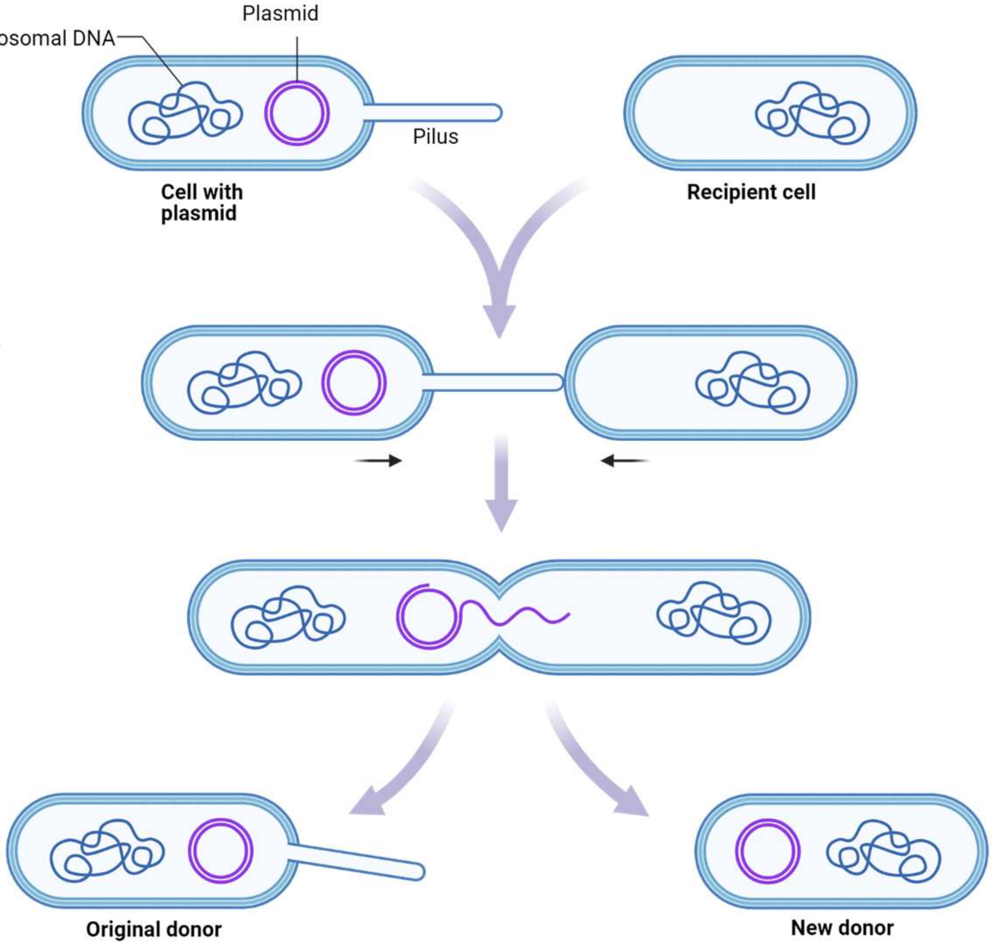
Plasmids

- Plasmids can in some cases be transferred between strains

- This makes them important for AMR surveillance as they can transfer resistance genes between lineages of bacteria (Horizontal gene transfer)

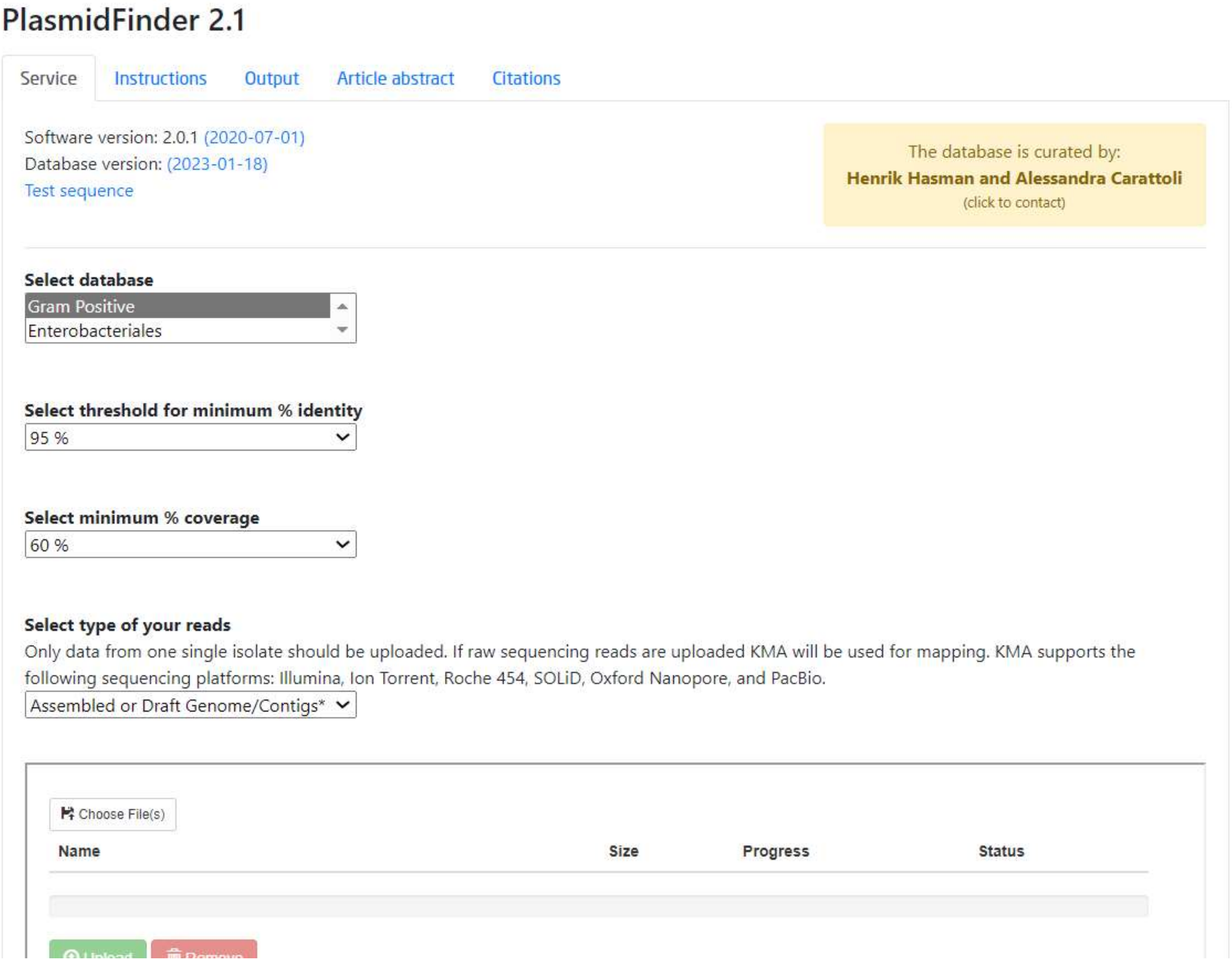
- Plasmids can be typed by the mechanism of replication, which differ from both the chromosomal DNA replication and among plasmids

- We will try to identify the “replicon”, the genes that conduct the replication and the origin site



Created with BioRender.com

We will mainly be looking into enterobacteriales



PlasmidFinder 2.1

Software version: 2.0.1 (2020-07-01)

Database version: (2023-01-18)

Test sequence

The database is curated by:
Henrik Hasman and Alessandra Carattoli
(click to contact)

Select database

Gram Positive
Enterobacteriales

Select threshold for minimum % identity

95 %

Select minimum % coverage

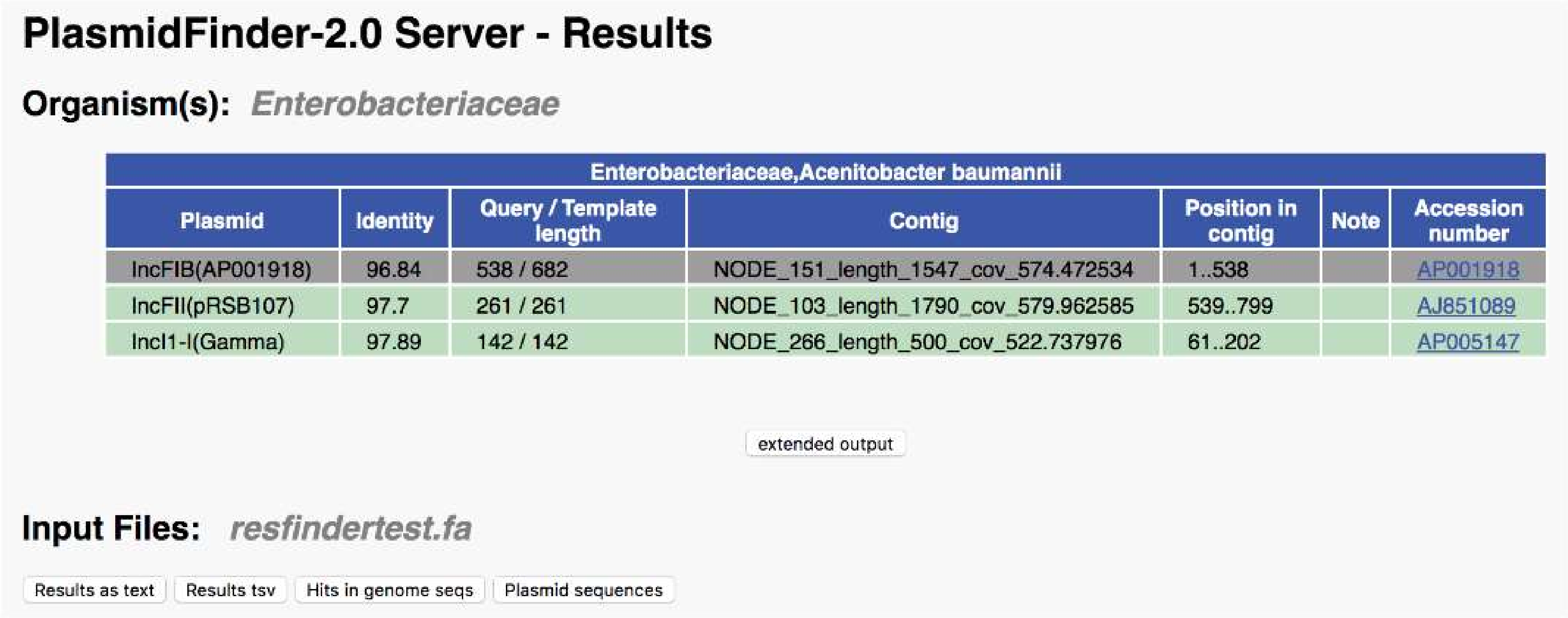
60 %

Select type of your reads

Only data from one single isolate should be uploaded. If raw sequencing reads are uploaded KMA will be used for mapping. KMA supports the following sequencing platforms: Illumina, Ion Torrent, Roche 454, SOLiD, Oxford Nanopore, and PacBio.

Assembled or Draft Genome/Contigs*

Output



PlasmidFinder-2.0 Server - Results

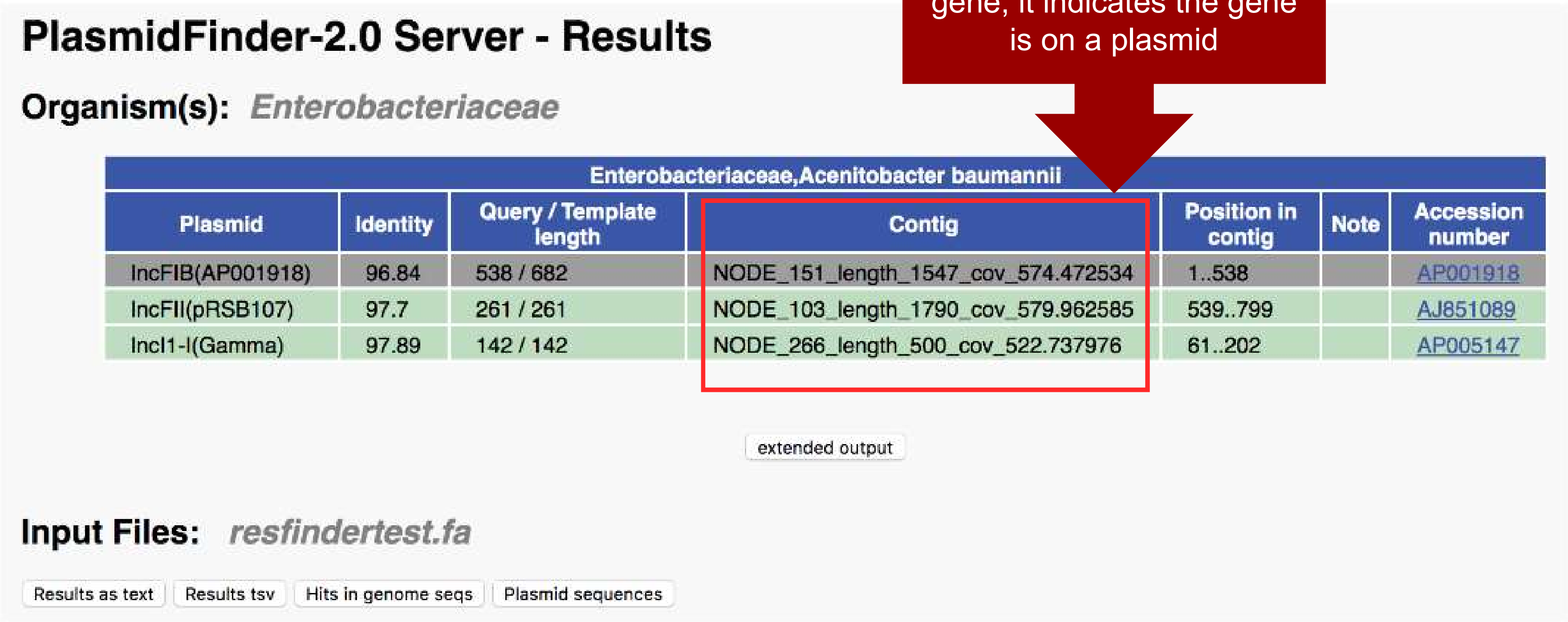
Organism(s): *Enterobacteriaceae*

Enterobacteriaceae,Acenitobacter baumannii						
Plasmid	**Identity**	**Query / Template length**	**Contig**	**Position in contig**	**Note**	**Accession number**
IncFIB(AP001918)	96.84	538 / 682	NODE_151_length_1547_cov_574.472534	1..538		AP001918
IncFII(pRSB107)	97.7	261 / 261	NODE_103_length_1790_cov_579.962585	539..799		AJ851089
IncI1-I(Gamma)	97.89	142 / 142	NODE_266_length_500_cov_522.737976	61..202		AP005147

Input Files: *resfindertest.fa*

Output

If the replicon is found on the same contig as a AMR gene, it indicates the gene is on a plasmid



PlasmidFinder-2.0 Server - Results

Organism(s): *Enterobacteriaceae*

Enterobacteriaceae,Acenitobacter baumannii						
Plasmid	**Identity**	**Query / Template length**	**Contig**	**Position in contig**	**Note**	**Accession number**
IncFIB(AP001918)	96.84	538 / 682	NODE_151_length_1547_cov_574.472534	1..538		AP001918
IncFII(pRSB107)	97.7	261 / 261	NODE_103_length_1790_cov_579.962585	539..799		AJ851089
IncI1-I(Gamma)	97.89	142 / 142	NODE_266_length_500_cov_522.737976	61..202		AP005147

Input Files: *resfindertest.fa*

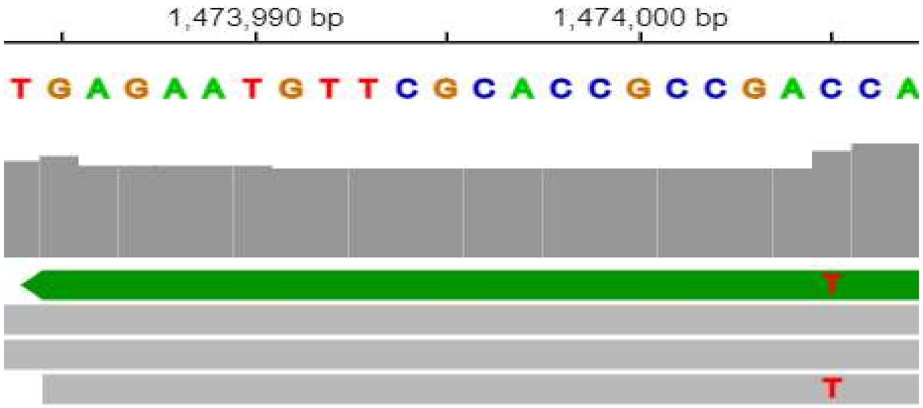
Single nucleotide polymorphism (SNP)

- 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

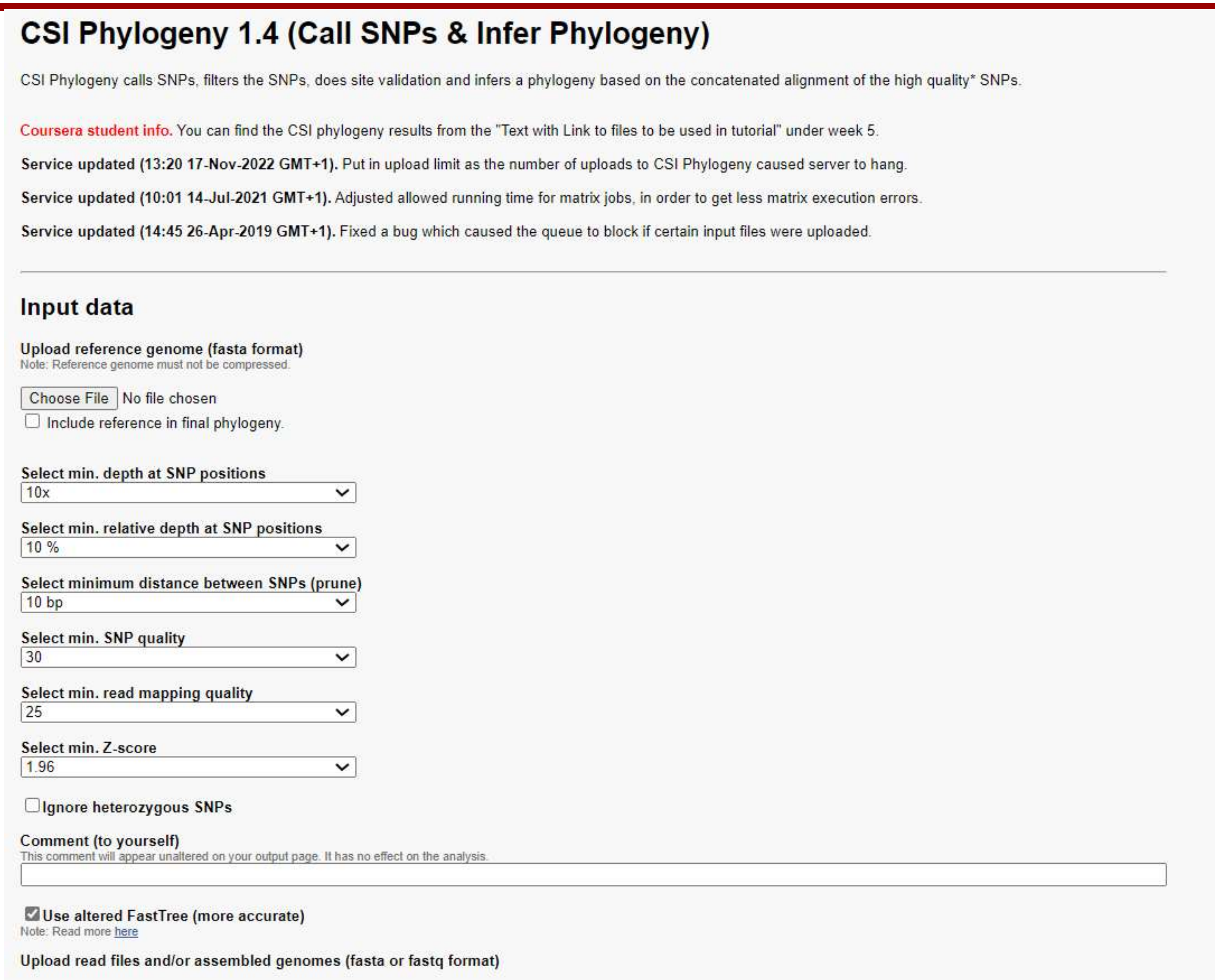
- When we know the amounts of SNP differences we can infer the phylogenic relationship between strains

- High resolution



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

Chose a reference, this is the sequence all other isolates will be compared to



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.

No file chosen

□ 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

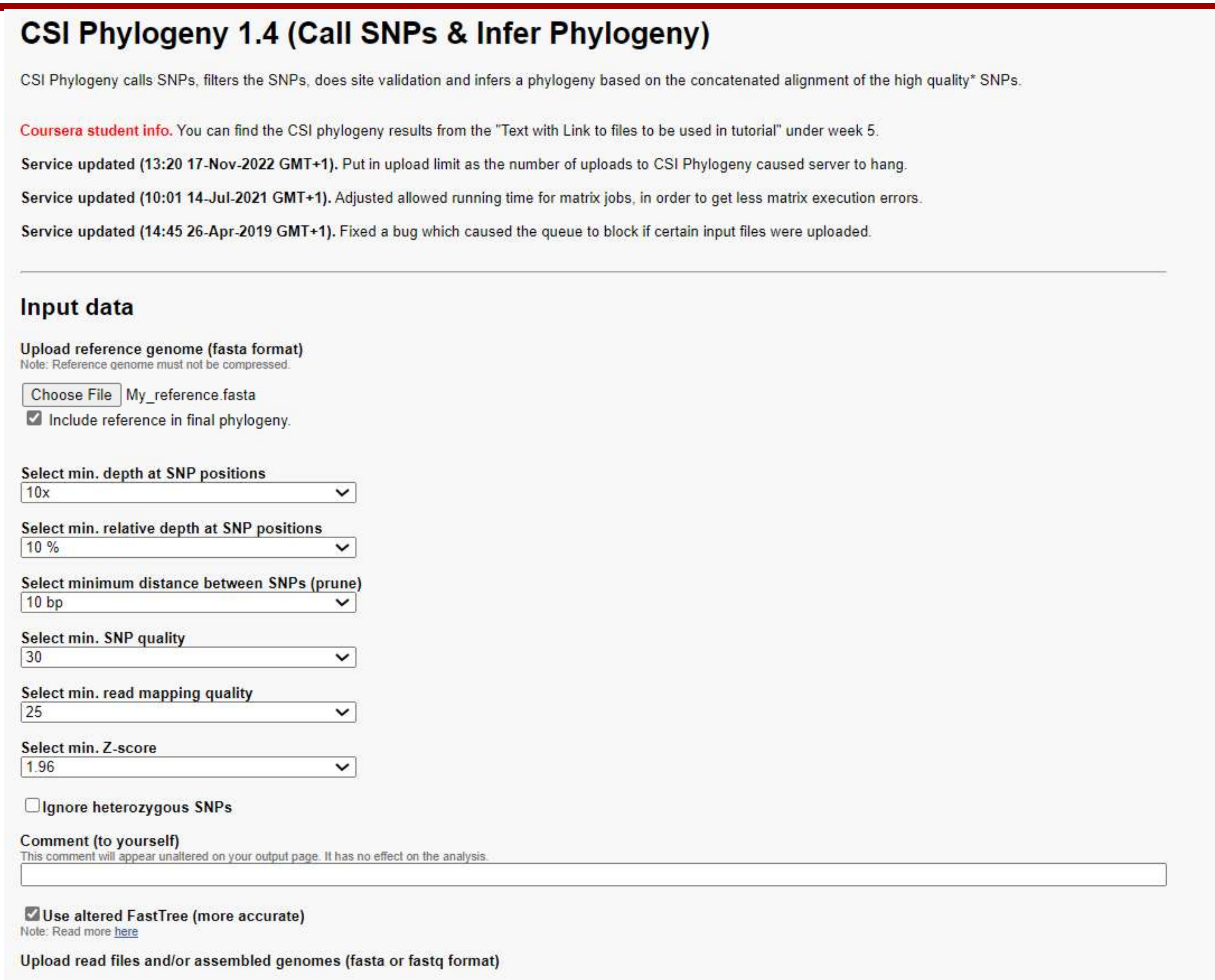
□ **Ignore heterozygous SNPs**

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)

Chose a reference, this is the sequence all other isolates will be compared to



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.

My_reference.fasta

☑ 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

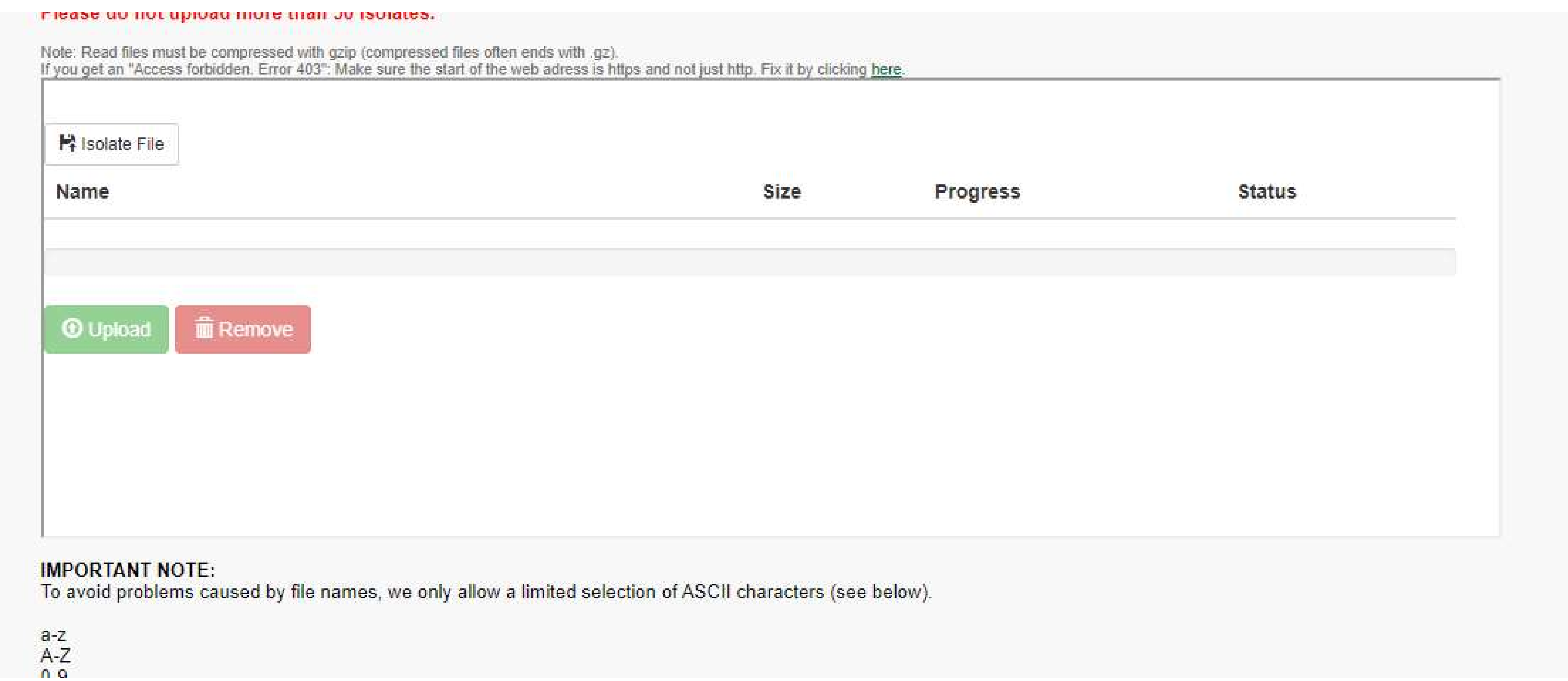
☐ **Ignore heterozygous SNPs**

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)

Ready to upload!



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.

IMPORTANT NOTE:
To avoid problems caused by file names, we only allow a limited selection of ASCII characters (see below).

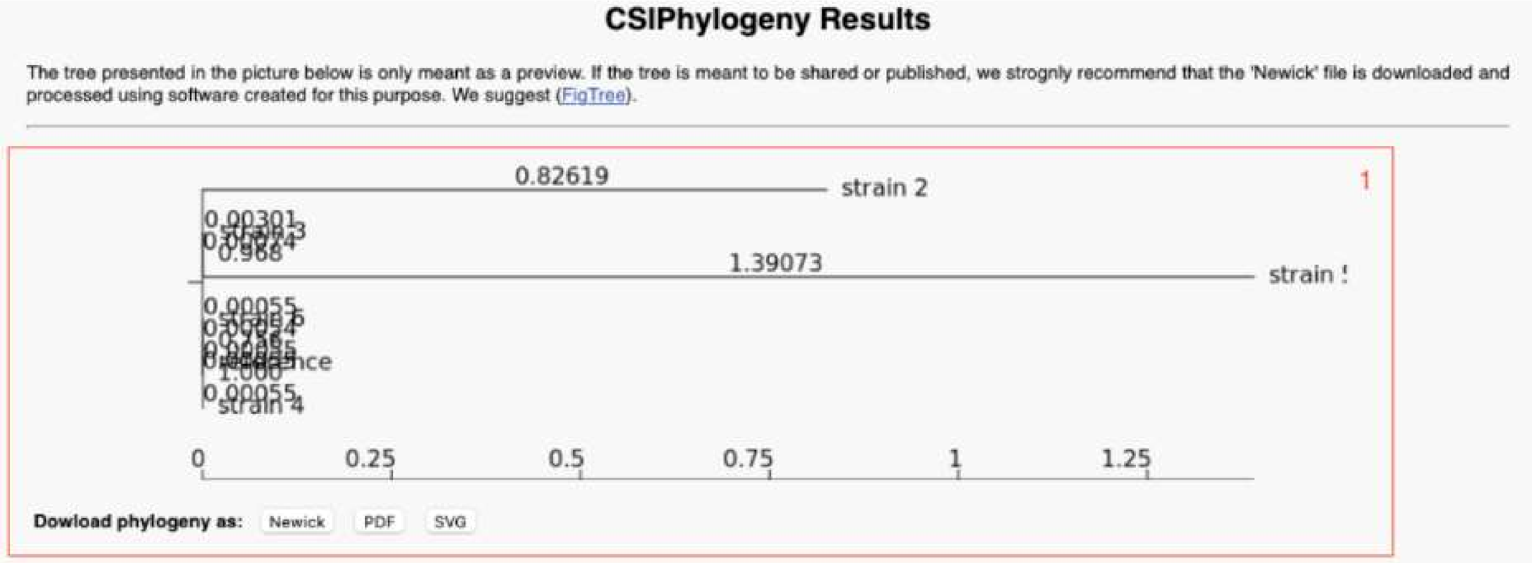
a-z
A-Z
0-9

Interpretation

- Some of the important outputs from CSIphylogeny are:
- The newick file, containing the phylogenetic tree
- The SNP matrix, which contains the number of SNPs between isolates

- In our exercise we will try to identify isolates belonging to an outbreak

- Isolates that cluster with our outbreak reference are presumably part of the outbreak



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).

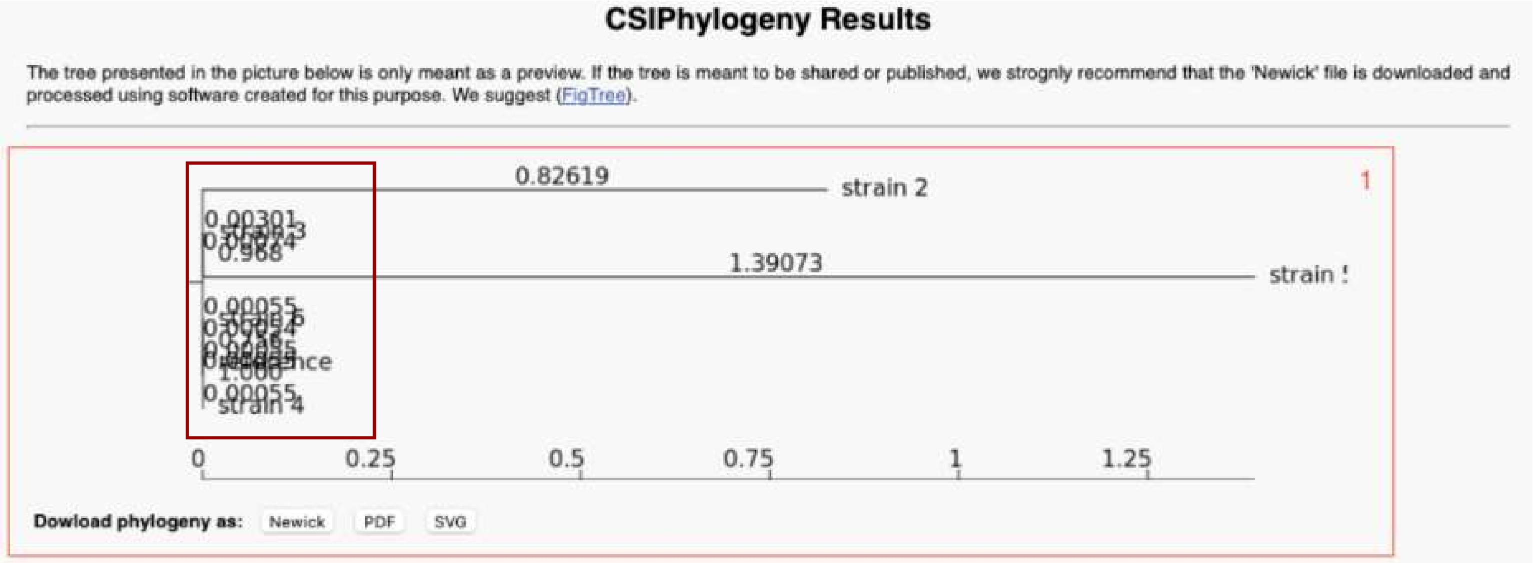
Dowload phylogeny as: Newick PDF SVG

Interpretation

- Some of the important outputs from CSIphylogeny are:
- The newick file, containing the phylogenetic tree
- The SNP matrix, which contains the number of SNPs between isolates

- In our exercise we will try to identify isolates belonging to an outbreak

- Isolates that cluster with our outbreak reference are presumably part of the outbreak



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).

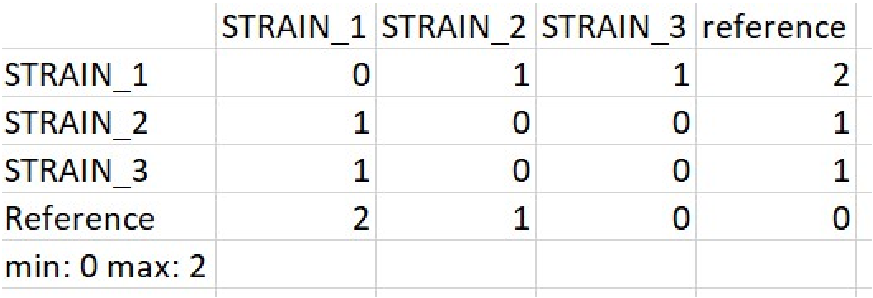
Interpretation

- The SNP matrix shows the distance between isolates

- In the table we can see that for strain_1:
  - 0 SNP differences to strain_1
  - 1 SNP difference to strain_2
  - 1 SNP difference to strain_3
  - 2 SNP differences to the reference

- The number of difference to determine whether a isolate is part of a cluster will depend on the setting, such as time interval between sampling and rate of mutation for the strain/species

- We often expect less than 5-10 SNP differences in an outbreak with this tool, but this is not a rule



	STRAIN_1	STRAIN_2	STRAIN_3	reference
STRAIN_1	0	1	1	2
STRAIN_2	1	0	0	1
STRAIN_3	1	0	0	1
Reference	2	1	0	0
min: 0 max: 2				

The Exercises

- Exercises will be sent as Excel file, with 5 sheets along with 18 fasta files (15th December)
- Exercise 1: Quality control of WGS
- Exercise 2: Phenotypic classification
- Exercise 3: Genotypic profiling (AMR)
- Exercise 4: Outbreak investigation
- Exercise 5: Your previous experiences
- **Please return your answers before January 21st** by sending them to lahoso@food.dtu.dk, we will be holding a **Summary session with correct results February 1st.** More on time to follow.

- These exercises can be completed using the webtools discussed above.

- If you are new to genomic analysis, do not worry, this is a learning experience. Fill out as much as possible, we do not expect you to get everything correct.

- If you have experience in genomic analysis, some of the included isolates are purposefully a bit irregular to engourage interpretation of the results, I hope they will be interesting.

- You are always welcome to send questions to me or the EQAsia team