Introduction to online tools: Kmers, MLST and Serotyping

SEQAFRICA Module 1

The
Fleming Fund
Regional Grants

19 February 2021

Stanford Kwenda
Lead Data Analyst

National Institute for Communicable Diseases
South Africa

Bioinformatics web-based tools

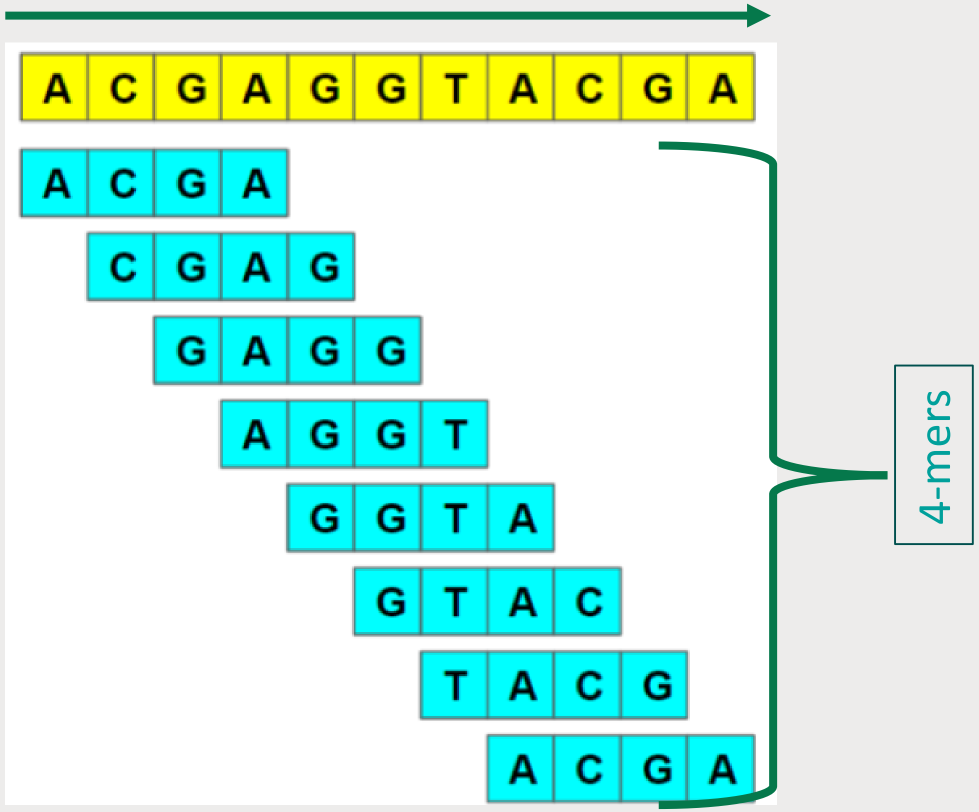
- Variety of methods and tools are available to analyze bacterial pathogens
- Most bioinformatics tools are implemented in Unix environments
  - Require at least some bioinformatics expertise for usage
- Web-based bioinformatics tools
  - Often free for use
  - Do not require computational power from the user
  - Limited bioinformatics knowledge
  - In some cases demand that that users deposit the analyzed data in public repositories
- Always make an effort to browse through the documentation of web based platforms
  - Helps with choice of parameters and interpretation of results

Uelze *et al*. One Health Outlook (2020)

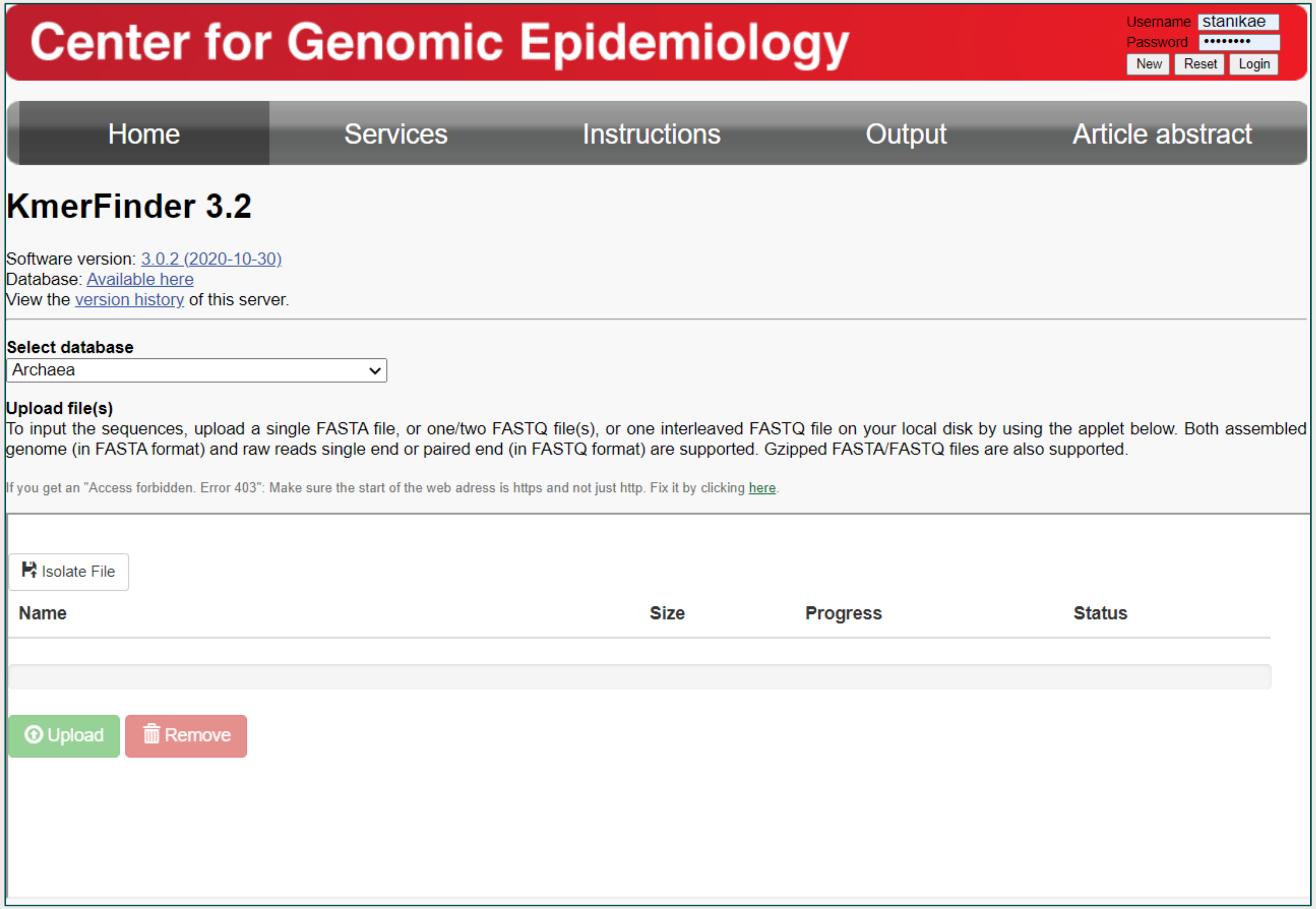
The Fleming Fund
Regional Grants

Introduction to Kmers

- A k-mer typically refers to all the possible substrings of length k that are contained in a string
- E.g. For sequence with length N
  - No. of K-mers = N-k+1
- K-mer counting
  - Total count, distinct count and unique count
- Computing k-mers is much faster than producing alignments
- Applications using k-mers:
  - Error correction
    - Rare k-mers are more likely due to sequence errors
  - Classification
    - Certain k-mers may uniquely identify genomes
  - Pseudo-alignment
    - New pseudo-aligners can match reads to locations based solely on the presence of common k-mers



Web-based species identification using Kmers



KmerFinder 3.2

Software version: 3.0.2 (2020-10-30)
Database: Available here
View the version history of this server.

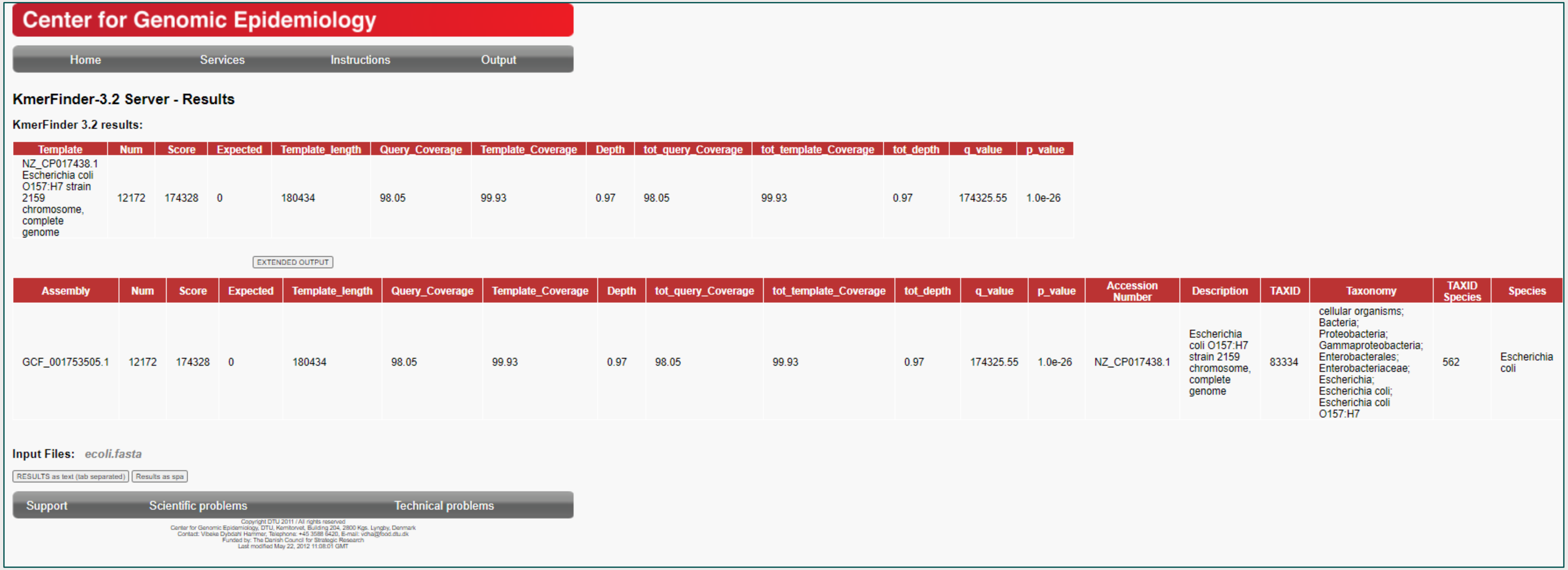
Select database
Archaea

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.

https://cge.cbs.dtu.dk/services/KmerFinder/

Web-based species identification using Kmers



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_Coverage	tot_depth	q_value	p_value
NZ_CP017438.1 Escherichia coli O157:H7 strain 2159 chromosome, complete genome	12172	174328	0	180434	98.05	99.93	0.97	98.05	99.93	0.97	174325.55	1.0e-26

Assembly	Num	Score	Expected	Template_length	Query_Coverage	Template_Coverage	Depth	tot_query_Coverage	tot_template_Coverage	tot_depth	q_value	p_value	Accession Number	Description	TAXID	Taxonomy	TAXID Species	Species
GCF_001753505.1	12172	174328	0	180434	98.05	99.93	0.97	98.05	99.93	0.97	174325.55	1.0e-26	NZ_CP017438.1	Escherichia coli O157:H7 strain 2159 chromosome, complete genome	83334	cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia; Escherichia coli; Escherichia coli O157:H7	562	Escherichia coli

Input Files: *ecoli.fasta*

Center for Genomic Epidemiology, DTU, Kemitorvet, Building 204, 2800 Kgs. Lyngby, Denmark
Contact: Vibeke Dybdahl Hammer, Telephone: +45 3588 6420, E-mail: vdha@food.dtu.dk
Funded by: The Danish Council for Strategic Research
Last modified May 22, 2012 11:08:01 GMT

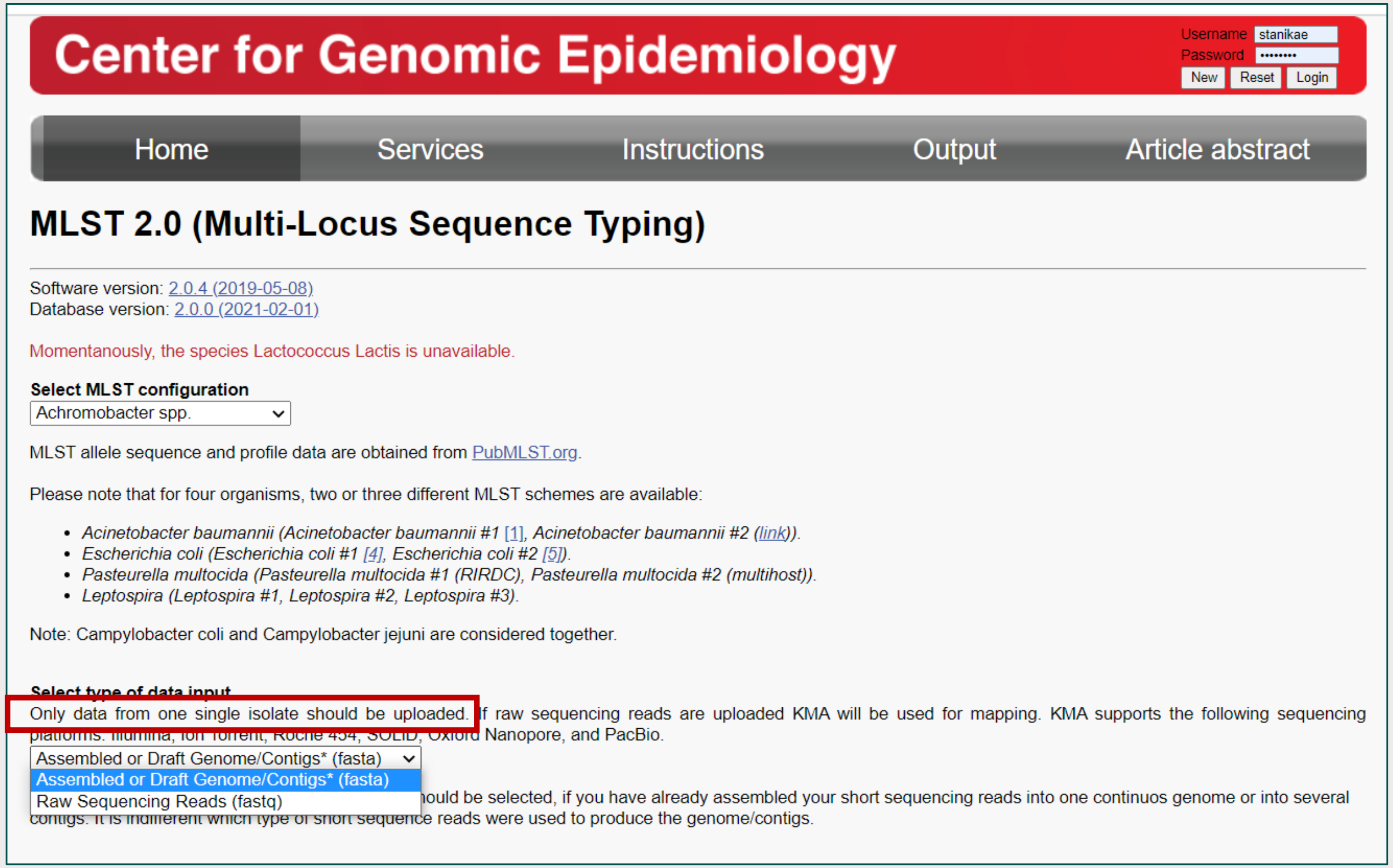
https://cge.cbs.dtu.dk/services/KmerFinder/

Multi-Locus Sequence Typing (MLST)

- A universal, portable, and precise means of typing bacteria
  - Traditionally involved PCR amplification and DNA sequencing of PCR fragments
- Indexes sequences (~500 bp) of representative housekeeping genes
  - Usually from seven loci
  - Each unique allele is assigned an arbitrary integer identifier
  - Sequence type (ST) – unique combinations of the alleles at each locus
- NGS sequence data can be matched to a database of allelic profiles
  - A single nucleotide variation at any of these loci defines a different allele and informs the ST
- MLST often not discriminative enough for outbreak detection or distinguishing highly related strains

Belen et al. Methods Mol Bio (2009)
https://pubmlst.org/multilocus-sequence-typing
Jolley and Maiden. BMC Bioinformatics (2010)

Web-based MLST typing tools: MLST2.0



MLST 2.0 (Multi-Locus Sequence Typing)

Software version: 2.0.4 (2019-05-08)
Database version: 2.0.0 (2021-02-01)

Momentanously, the species Lactococcus Lactis is unavailable.

Select MLST configuration
Achromobacter spp.

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

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

- *Acinetobacter baumannii (Acinetobacter baumannii #1* [1], *Acinetobacter baumannii #2 (link)).*
- *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).*

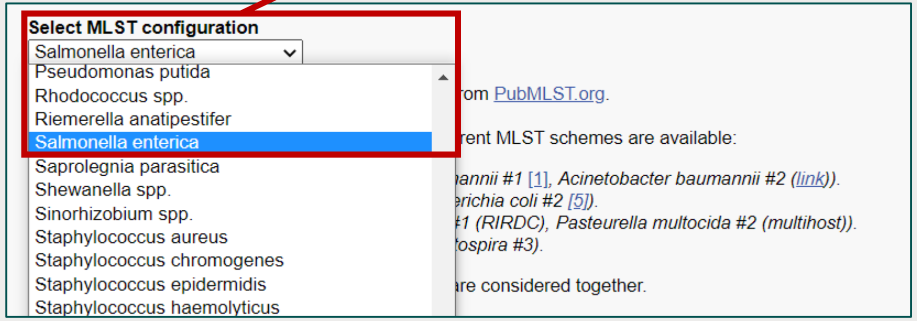
Note: Campylobacter coli and Campylobacter jejuni are considered together.

Select 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* (fasta)
Assembled or Draft Genome/Contigs* (fasta)
Raw Sequencing Reads (fastq)

hould 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.

Must select MLST scheme prior to launching the tool



Select MLST configuration
Salmonella enterica

Pseudomonas putida
Rhodococcus spp.
Riemerella anatipestifer
Salmonella enterica
Saprolegnia parasitica
Shewanella spp.
Sinorhizobium spp.
Staphylococcus aureus
Staphylococcus chromogenes
Staphylococcus epidermidis
Staphylococcus haemolyticus

rom PubMLST.org.

rent MLST schemes are available:

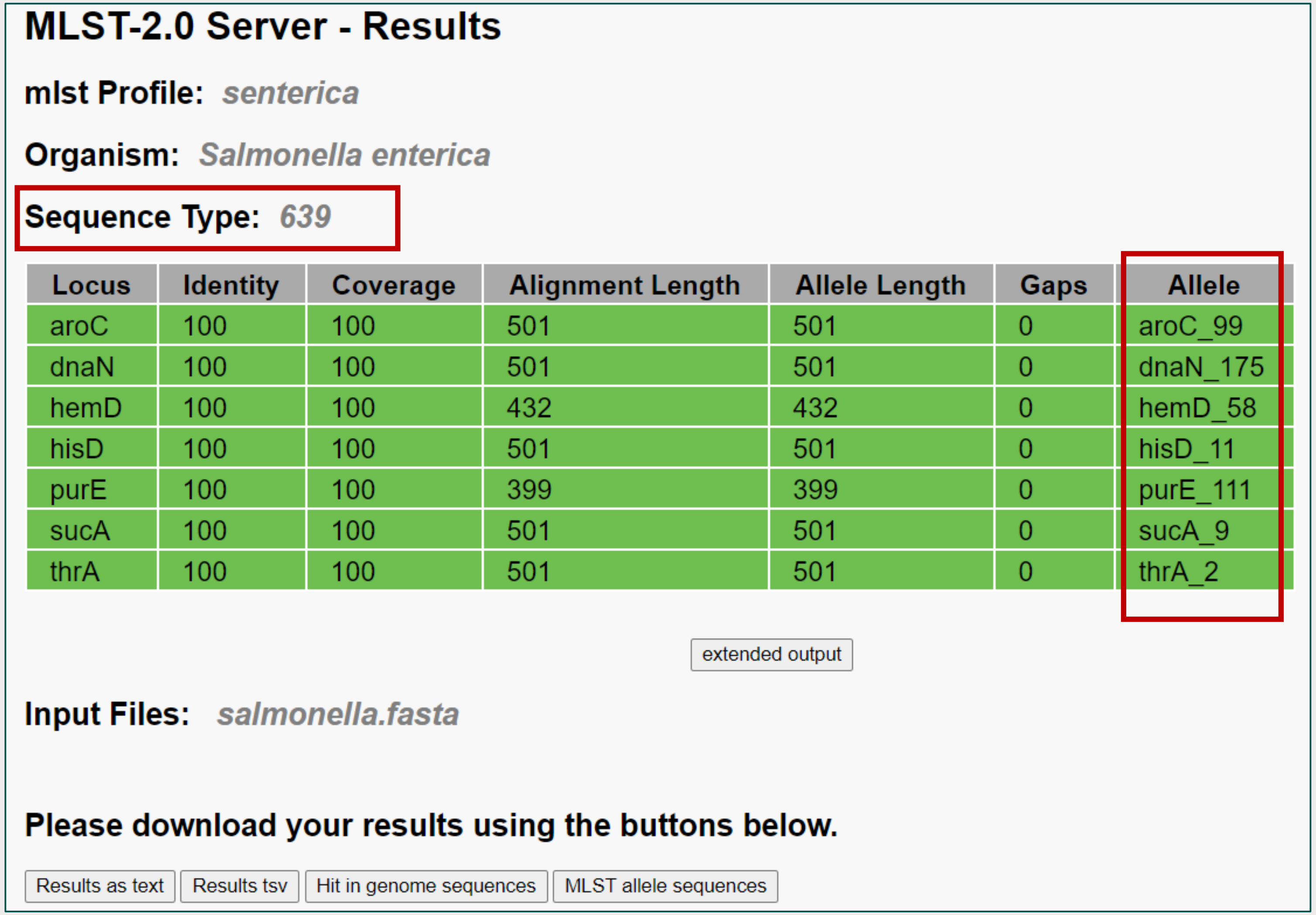
annii #1 [1], *Acinetobacter baumannii #2 (link)).*
richia coli #2 [5]).
1 (RIRDC), Pasteurella multocida #2 (multihost)).
ospira #3).

re considered together.

- Input: raw reads or assembled genomes
- Can only take data for a single isolate at a time
- Some species will have more than one MLST scheme

https://cge.cbs.dtu.dk/services/MLST/

Web-based MLST typing tools: MLST2.0



MLST-2.0 Server - Results

mlst Profile: *senterica*

Organism: *Salmonella enterica*

Sequence Type: *639*

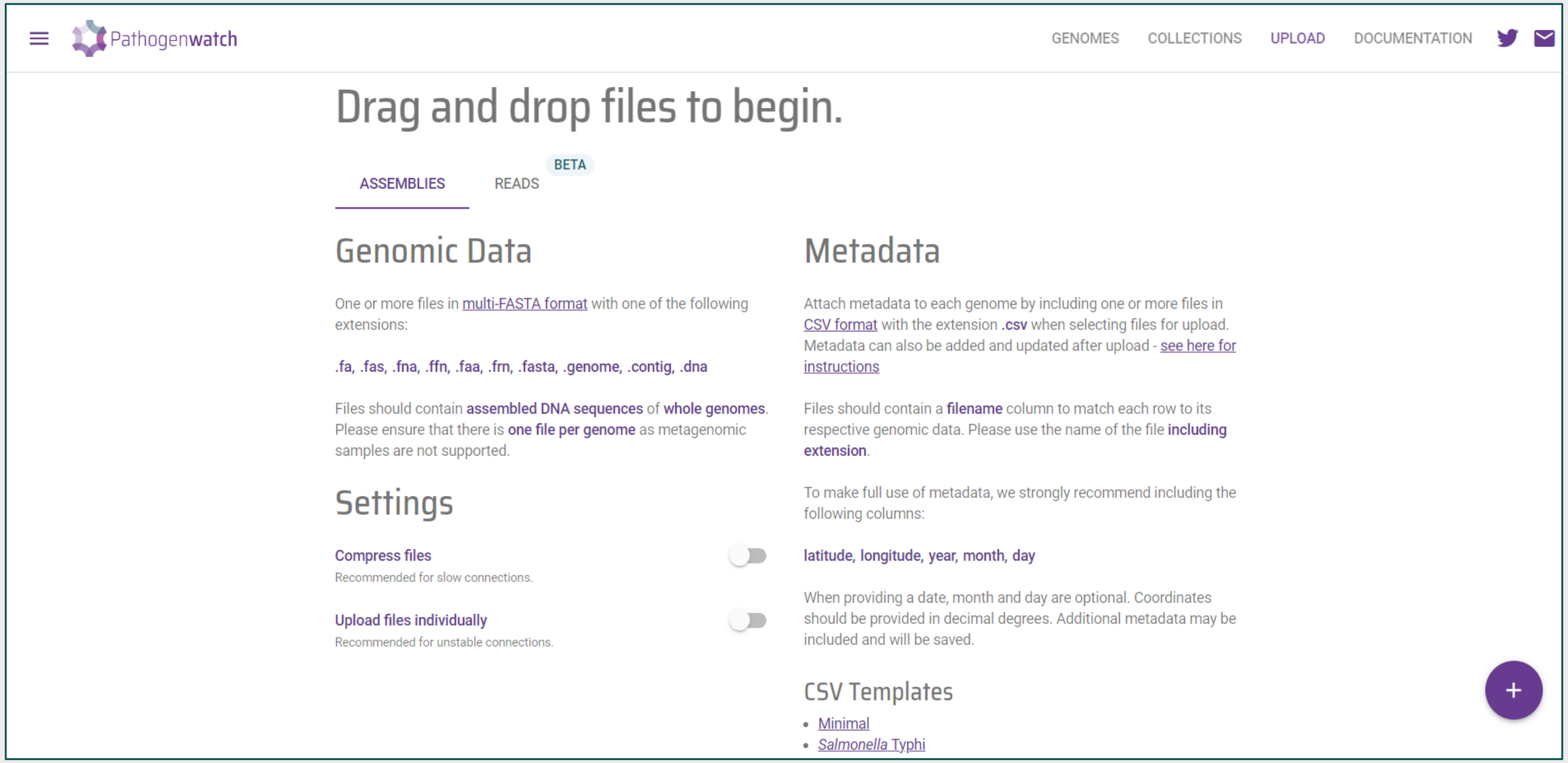
Locus	Identity	Coverage	Alignment Length	Allele Length	Gaps	Allele
aroC	100	100	501	501	0	aroC_99
dnaN	100	100	501	501	0	dnaN_175
hemD	100	100	432	432	0	hemD_58
hisD	100	100	501	501	0	hisD_11
purE	100	100	399	399	0	purE_111
sucA	100	100	501	501	0	sucA_9
thrA	100	100	501	501	0	thrA_2

Input Files: *salmonella.fasta*

Please download your results using the buttons below.

https://cge.cbs.dtu.dk/services/MLST/

Web-based MLST typing tools: PathogenWatch



Drag and drop files to begin.

ASSEMBLIES READS BETA

Genomic Data

One or more files in multi-FASTA format with one of the following extensions:

.fa, .fas, .fna, .ffn, .faa, .frn, .fasta, .genome, .contig, .dna

Files should contain **assembled DNA sequences** of **whole genomes**. Please ensure that there is **one file per genome** as metagenomic samples are not supported.

Settings

Compress files
Recommended for slow connections.

Upload files individually
Recommended for unstable connections.

Metadata

Attach metadata to each genome by including one or more files in CSV format with the extension **.csv** when selecting files for upload. Metadata can also be added and updated after upload - see here for instructions

Files should contain a **filename** column to match each row to its respective genomic data. Please use the name of the file **including extension**.

To make full use of metadata, we strongly recommend including the following columns:

latitude, longitude, year, month, day

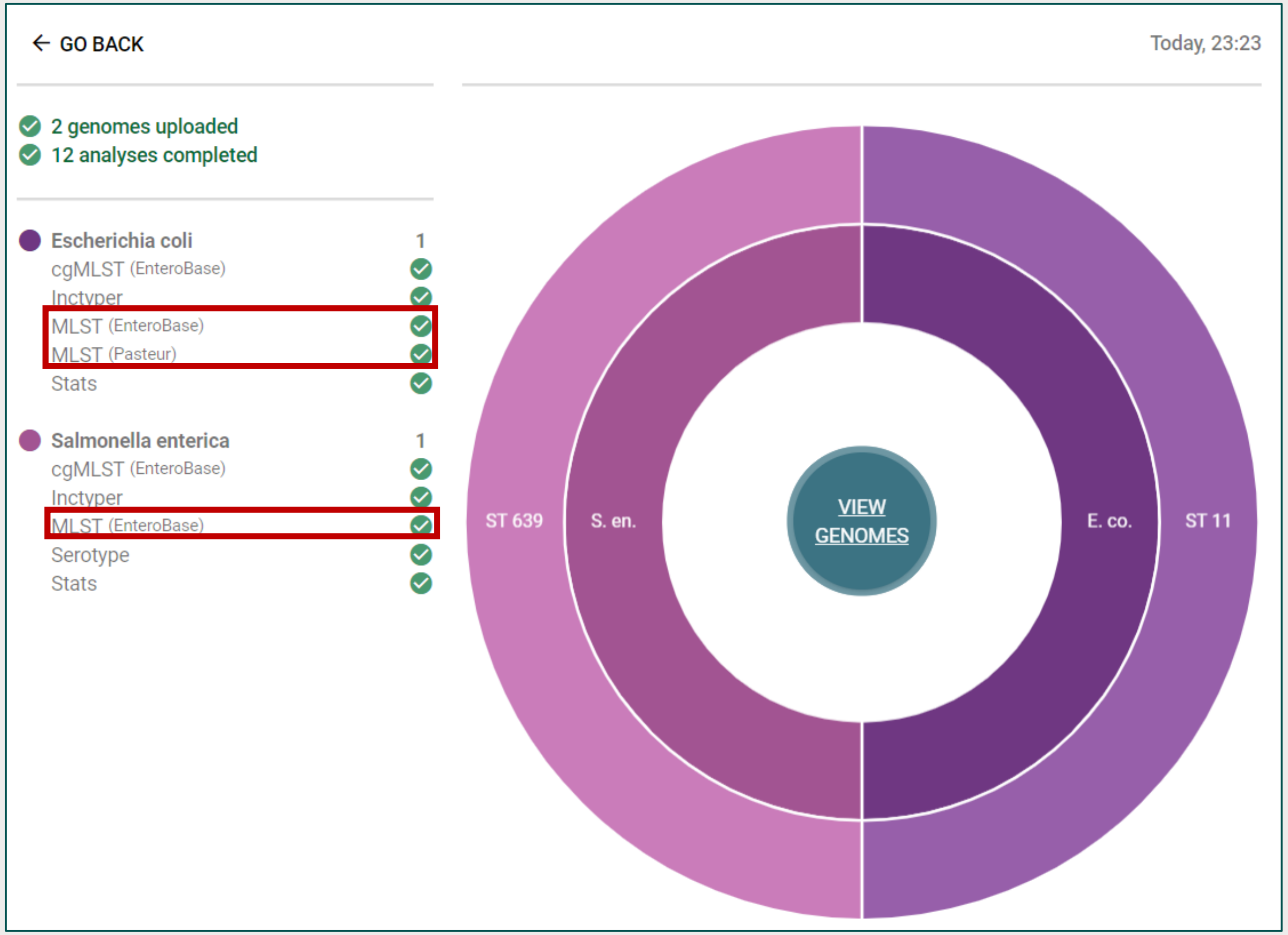
When providing a date, month and day are optional. Coordinates should be provided in decimal degrees. Additional metadata may be included and will be saved.

CSV Templates

- Minimal
- *Salmonella* Typhi

https://pathogen.watch/

Web-based MLST typing tools: PathogenWatch

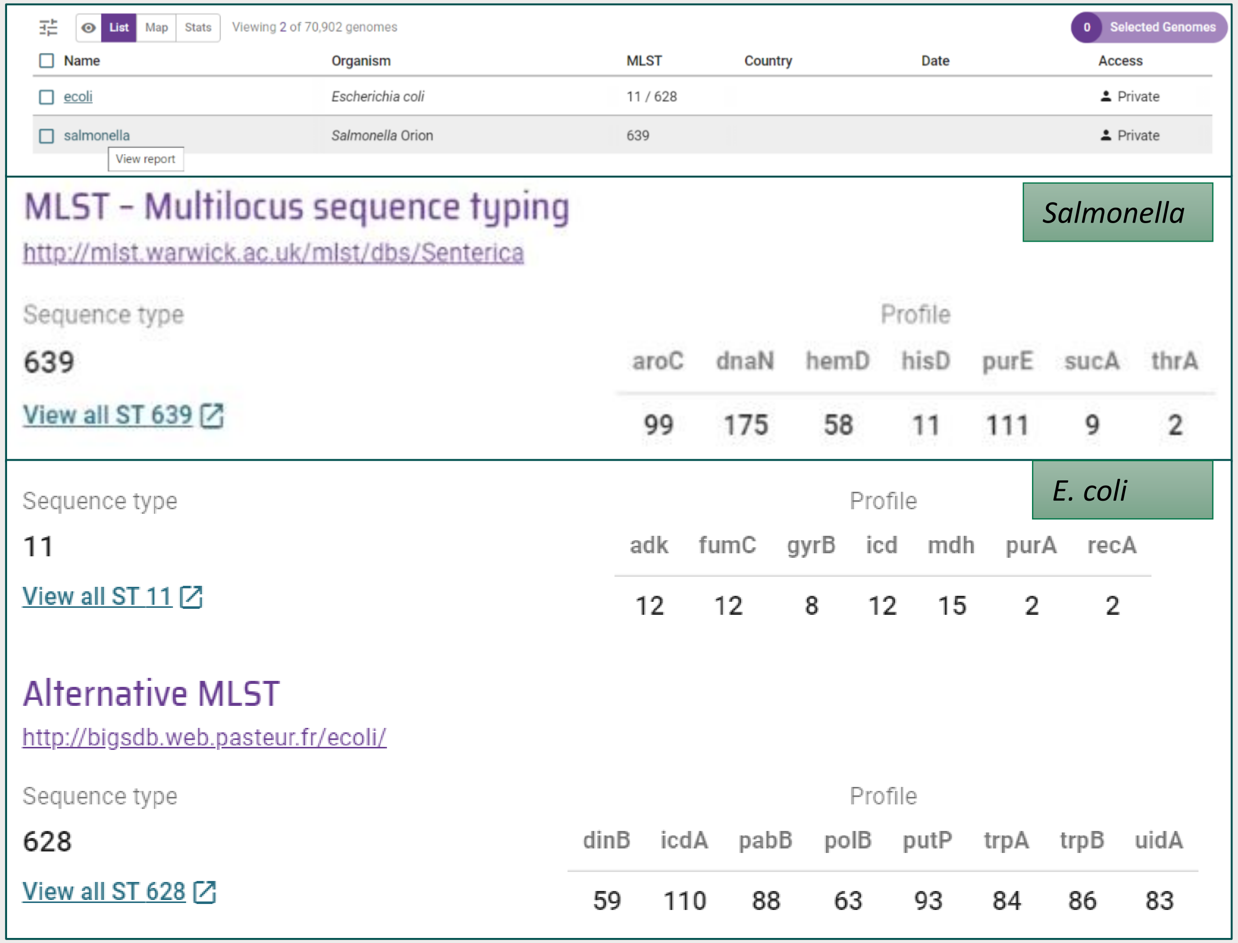


← GO BACK

Today, 23:23

✓ 2 genomes uploaded
✓ 12 analyses completed

● Escherichia coli	1
cgMLST (EnteroBase)	✓
Inctyper	✓
MLST (EnteroBase)	✓
MLST (Pasteur)	✓
Stats	✓
● Salmonella enterica	1
cgMLST (EnteroBase)	✓
Inctyper	✓
MLST (EnteroBase)	✓
Serotype	✓
Stats	✓



List | Map | Stats — Viewing 2 of 70,902 genomes — 0 Selected Genomes

□ Name	Organism	MLST	Country	Date	Access
□ ecoli	*Escherichia coli*	11 / 628			Private
□ salmonella	*Salmonella* Orion	639			Private

View report

MLST – Multilocus sequence typing

Salmonella

http://mlst.warwick.ac.uk/mlst/dbs/Senterica

Sequence type	aroC	dnaN	hemD	hisD	purE	sucA	thrA
639	99	175	58	11	111	9	2

View all ST 639

E. coli

Sequence type	adk	fumC	gyrB	icd	mdh	purA	recA
11	12	12	8	12	15	2	2

View all ST 11

Alternative MLST

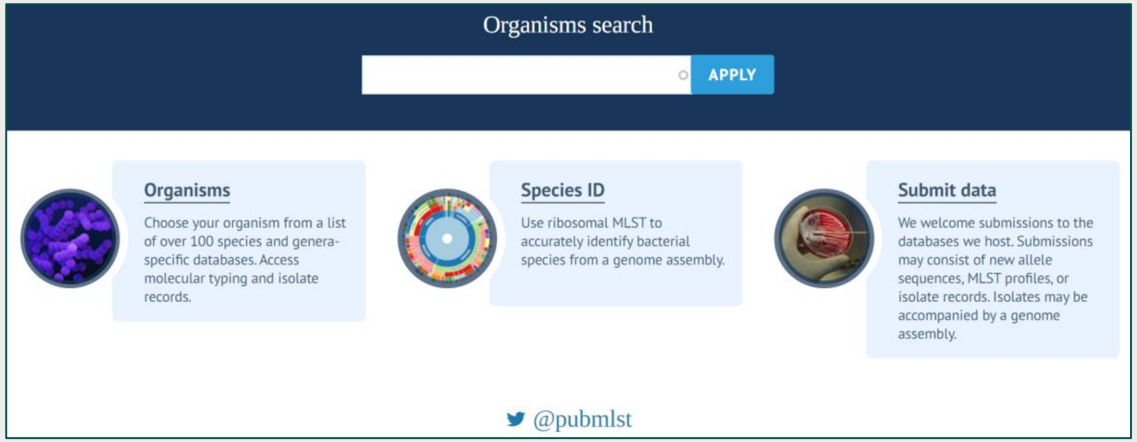
http://bigsdb.web.pasteur.fr/ecoli/

Sequence type	dinB	icdA	pabB	polB	putP	trpA	trpB	uidA
628	59	110	88	63	93	84	86	83

View all ST 628

https://pathogen.watch/

Web-based MLST typing tools: MLST2.0



Organisms

Choose your organism from a list of over 100 species and genera-specific databases. Access molecular typing and isolate records.

Species ID

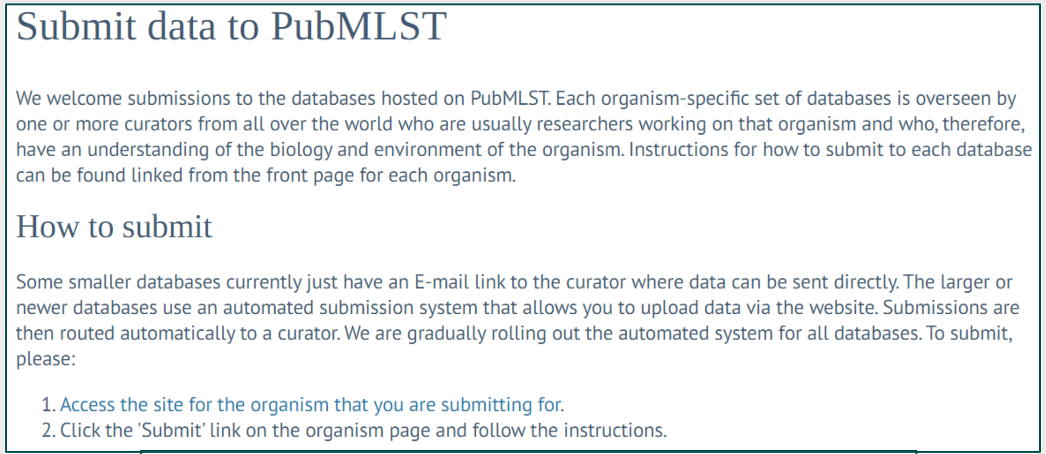
Use ribosomal MLST to accurately identify bacterial species from a genome assembly.

Submit data

We welcome submissions to the databases we host. Submissions may consist of new allele sequences, MLST profiles, or isolate records. Isolates may be accompanied by a genome assembly.

@pubmlst

https://pubmlst.org/



Submit data to PubMLST

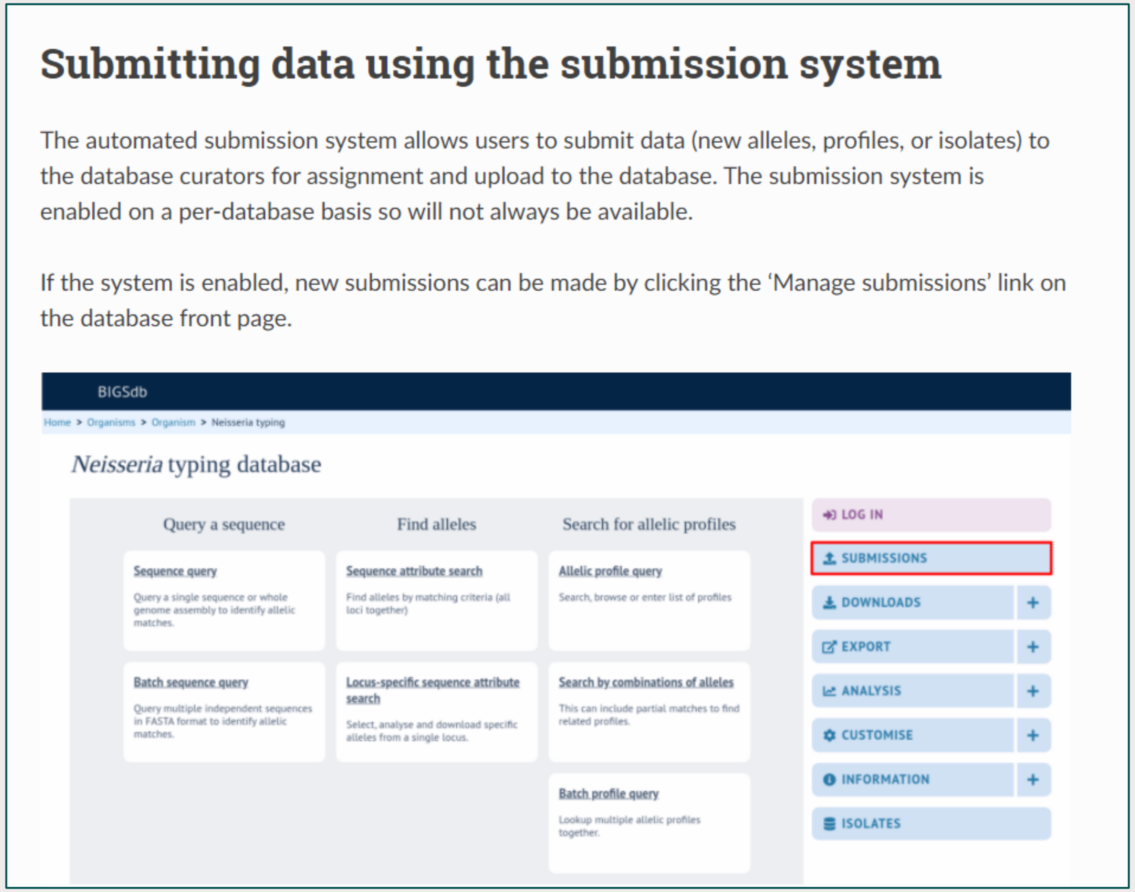
We welcome submissions to the databases hosted on PubMLST. Each organism-specific set of databases is overseen by one or more curators from all over the world who are usually researchers working on that organism and who, therefore, have an understanding of the biology and environment of the organism. Instructions for how to submit to each database can be found linked from the front page for each organism.

How to submit

Some smaller databases currently just have an E-mail link to the curator where data can be sent directly. The larger or newer databases use an automated submission system that allows you to upload data via the website. Submissions are then routed automatically to a curator. We are gradually rolling out the automated system for all databases. To submit, please:

1. Access the site for the organism that you are submitting for.
2. Click the 'Submit' link on the organism page and follow the instructions.

https://pubmlst.org/submit-data



Submitting data using the submission system

The automated submission system allows users to submit data (new alleles, profiles, or isolates) to the database curators for assignment and upload to the database. The submission system is enabled on a per-database basis so will not always be available.

If the system is enabled, new submissions can be made by clicking the 'Manage submissions' link on the database front page.

BIGSdb
Home > Organisms > Organism > Neisseria typing
Neisseria typing database
Query a sequence
Find alleles
Search for allelic profiles
Sequence query
Query a single sequence or whole genome assembly to identify allelic matches.
Sequence attribute search
Find alleles by matching criteria (all loci together)
Allelic profile query
Search, browse or enter list of profiles
Batch sequence query
Query multiple independent sequences in FASTA format to identify allelic matches.
Locus-specific sequence attribute search
Select, analyse and download specific alleles from a single locus.
Search by combinations of alleles
This can include partial matches to find related profiles.
Batch profile query
Lookup multiple allelic profiles together.
LOG IN
SUBMISSIONS
DOWNLOADS +
EXPORT +
ANALYSIS +
CUSTOMISE +
INFORMATION +
ISOLATES

https://bigsdb.readthedocs.io/en/latest/submissions.html

Bacterial serotyping

- Serotypes (serovars) – groups within a single species of microorganisms, (e.g. bacteria) which share distinctive surface structures e.g. surface antigens
  - Serotyping – classifying species at a sub-species level based on e.g. antigen properties into serogroups
- Strain – single isolates from pure cultures of a given species with distinct phenotypic/genotypic traits
- Virulence and pathogenicity tend to correlate well with subtype assignments

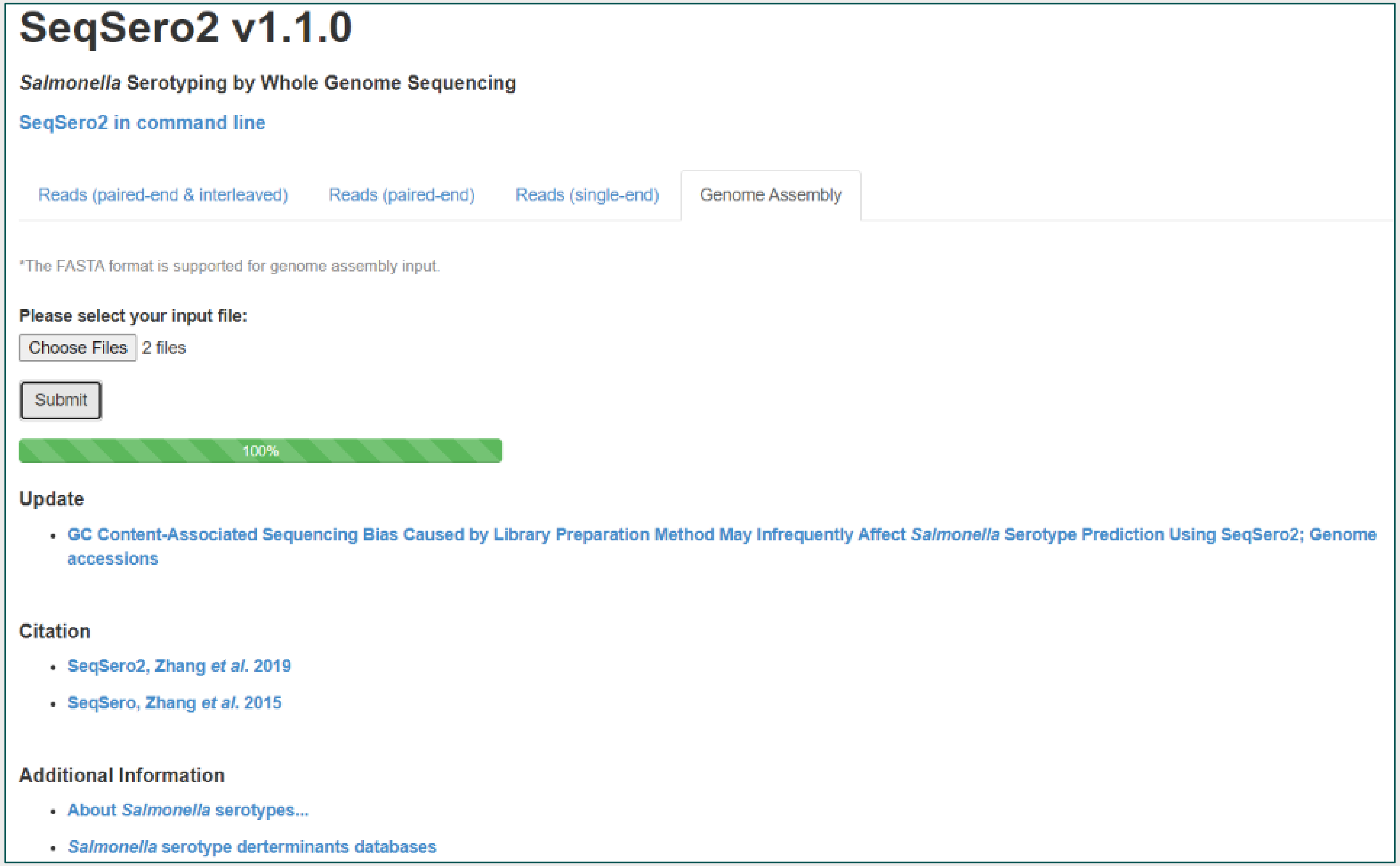
https://www.microscopemaster.com/serotype.html
Uelze *et al.* One Health Outlook (2020)
https://www.cdc.gov/salmonella/reportspubs/salmonella-atlas/serotyping-importance.html

Serotyping in *Salmonella*

- *Salmonella* can be separated into many serotypes based on:
    - O antigen – outermost portion of the bacteria’s surface covering
    - H antigen – slender threadlike structure that is part of the flagella
- Serotypes determined based on the distinct combination of O and H antigens
    - >2500 serotypes described for *Salmonella*
    - <100 serotypes account for most human infections
- Each known antigen is assigned a number and letter code
    - Combined into a seroformula
    - For example the White-Kauffmann-Le Minor scheme for *Salmonella*

https://www.cdc.gov/salmonella/reportspubs/salmonella-atlas/serotyping-importance.html
Uelze *et al.* One Health Outlook (2020)

Web-based tools for Salmonella serotyping



SeqSero2 v1.1.0

Salmonella Serotyping by Whole Genome Sequencing

SeqSero2 in command line

*The FASTA format is supported for genome assembly input.

Please select your input file:

Update

- GC Content-Associated Sequencing Bias Caused by Library Preparation Method May Infrequently Affect *Salmonella* Serotype Prediction Using SeqSero2; Genome accessions

Citation

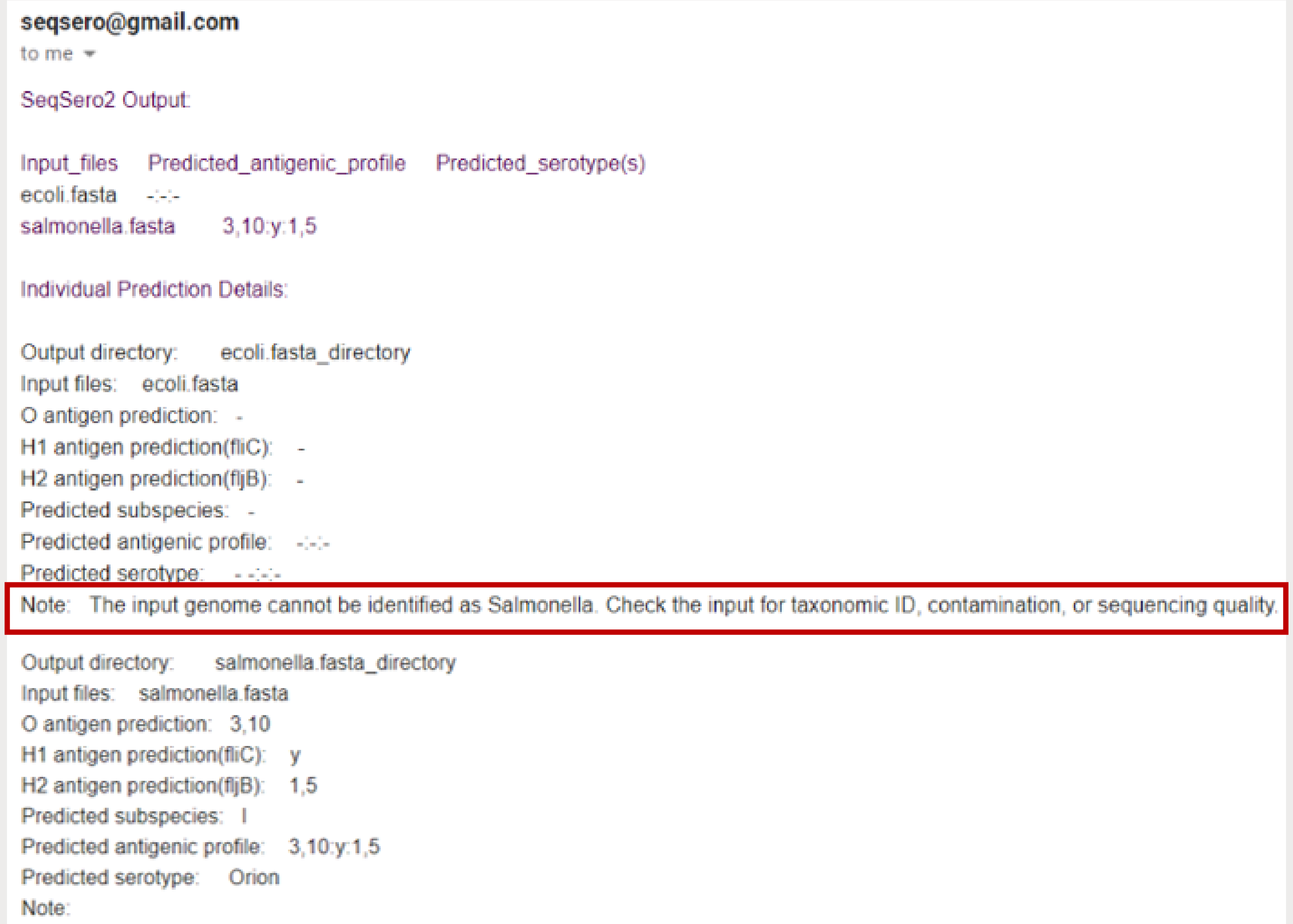
- SeqSero2, Zhang *et al.* 2019
- SeqSero, Zhang *et al.* 2015

Additional Information

- About *Salmonella* serotypes...
- *Salmonella* serotype derterminants databases

http://www.denglab.info/SeqSero2

Web-based tools for *Salmonella* serotyping

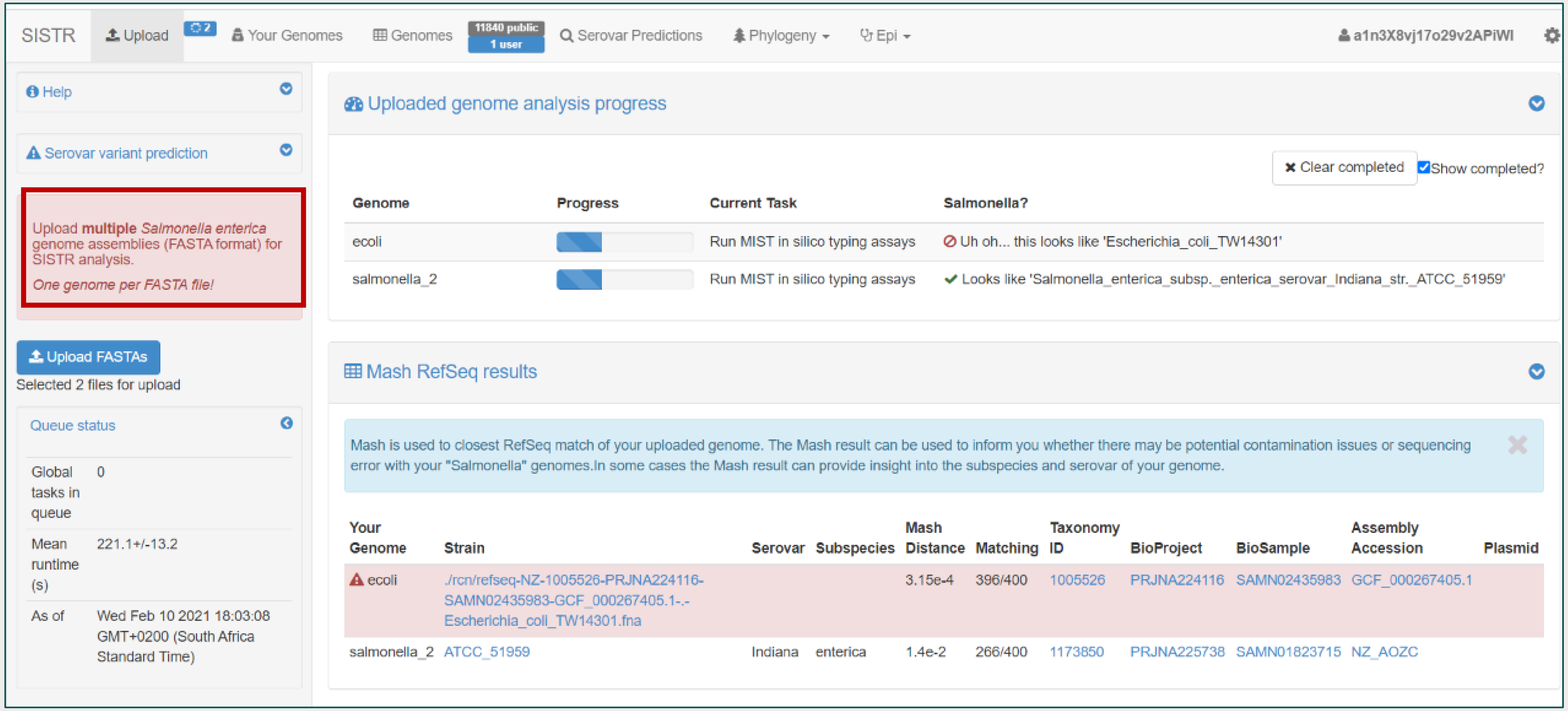


seqsero@gmail.com
to me

```
SeqSero2 Output:

Input_files     Predicted_antigenic_profile     Predicted_serotype(s)
ecoli.fasta     -:-:-
salmonella.fasta       3,10:y:1,5

Individual Prediction Details:

Output directory:      ecoli.fasta_directory
Input files:    ecoli.fasta
O antigen prediction:   -
H1 antigen prediction(fliC):    -
H2 antigen prediction(fljB):    -
Predicted subspecies:   -
Predicted antigenic profile:    -:-:-
Predicted serotype:     - -:-:-
Note:   The input genome cannot be identified as Salmonella. Check the input for taxonomic ID, contamination, or sequencing quality.

Output directory:      salmonella.fasta_directory
Input files:    salmonella.fasta
O antigen prediction:   3,10
H1 antigen prediction(fliC):    y
H2 antigen prediction(fljB):    1,5
Predicted subspecies:   I
Predicted antigenic profile:    3,10:y:1,5
Predicted serotype:     Orion
Note:
```

http://www.denglab.info/SeqSero2

Web-based tools for *Salmonella* serotyping



Upload **multiple** *Salmonella enterica* genome assemblies (FASTA format) for SISTR analysis.

One genome per FASTA file!

Upload FASTAs

Selected 2 files for upload

Queue status

Global tasks in queue	0
Mean runtime (s)	221.1+/-13.2
As of	Wed Feb 10 2021 18:03:08 GMT+0200 (South Africa Standard Time)

Uploaded genome analysis progress

✖ Clear completed ☑ Show completed?

Genome	Progress	Current Task	Salmonella?
ecoli		Run MIST in silico typing assays	⊘ Uh oh... this looks like 'Escherichia_coli_TW14301'
salmonella_2		Run MIST in silico typing assays	✓ Looks like 'Salmonella_enterica_subsp._enterica_serovar_Indiana_str._ATCC_51959'

Mash RefSeq results

Mash is used to closest RefSeq match of your uploaded genome. The Mash result can be used to inform you whether there may be potential contamination issues or sequencing error with your "Salmonella" genomes.In some cases the Mash result can provide insight into the subspecies and serovar of your genome.

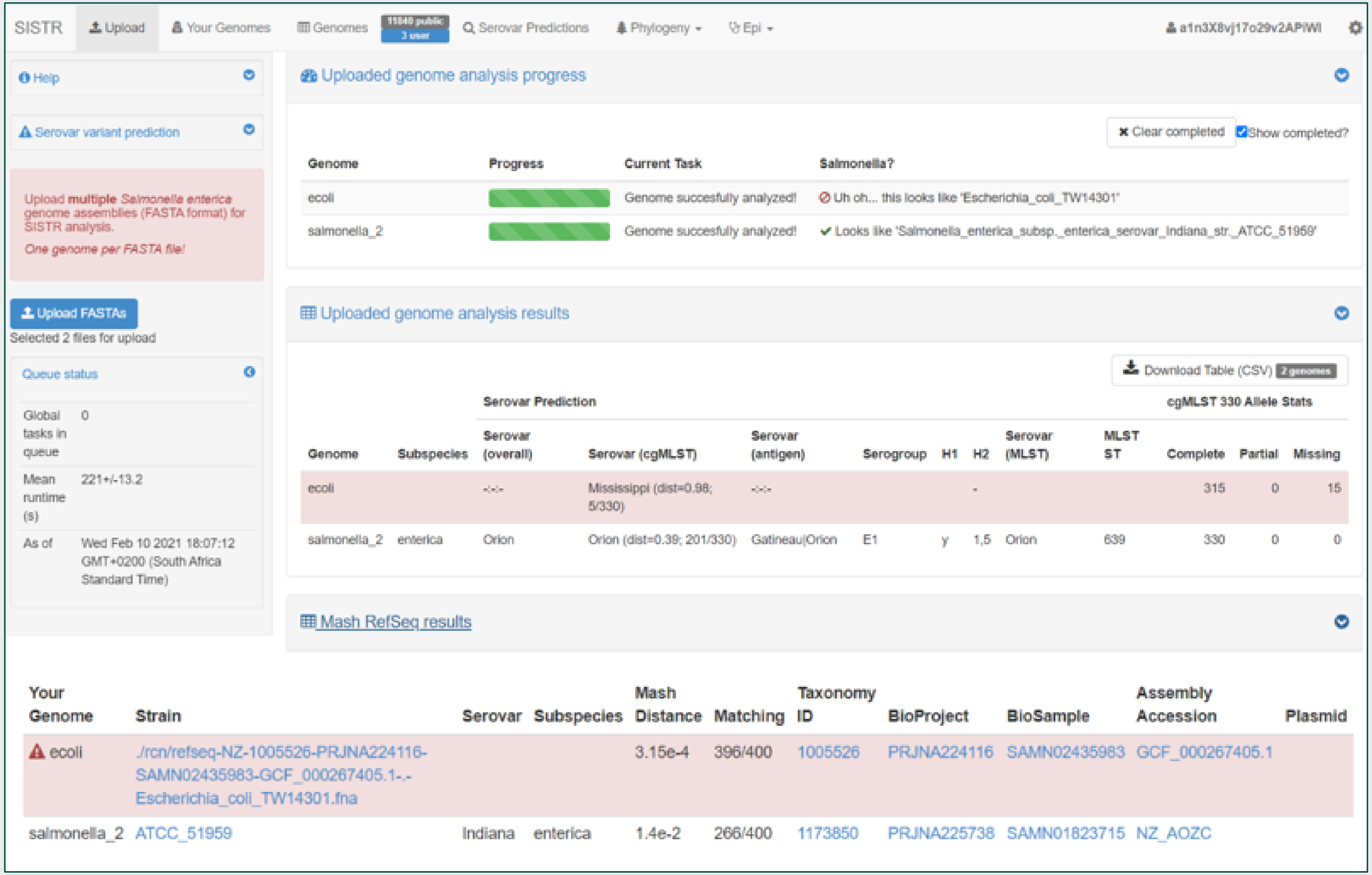
Your Genome	Strain	Serovar	Subspecies	Mash Distance	Matching	Taxonomy ID	BioProject	BioSample	Assembly Accession	Plasmid
ecoli	./rcn/refseq-NZ-1005526-PRJNA224116-SAMN02435983-GCF_000267405.1-.-Escherichia_coli_TW14301.fna			3.15e-4	396/400	1005526	PRJNA224116	SAMN02435983	GCF_000267405.1	
salmonella_2	ATCC_51959	Indiana	enterica	1.4e-2	266/400	1173850	PRJNA225738	SAMN01823715	NZ_AOZC	

- Input: assembled genome
- Can accept multiple assemblies

https://lfz.corefacility.ca/sistr-app/

The Fleming Fund
Regional Grants

Web-based tools for *Salmonella* serotyping



Upload **multiple** *Salmonella enterica* genome assemblies (FASTA format) for SISTR analysis.

One genome per FASTA file!

Selected 2 files for upload

Queue status

Global tasks in queue	0
Mean runtime (s)	221+/-13.2
As of	Wed Feb 10 2021 18:07:12 GMT+0200 (South Africa Standard Time)

Uploaded genome analysis progress

Genome	Progress	Current Task	Salmonella?
ecoli	🟩	Genome succesfully analyzed!	Uh oh... this looks like 'Escherichia_coli_TW14301'
salmonella_2	🟩	Genome succesfully analyzed!	✓ Looks like 'Salmonella_enterica_subsp._enterica_serovar_Indiana_str._ATCC_51959'

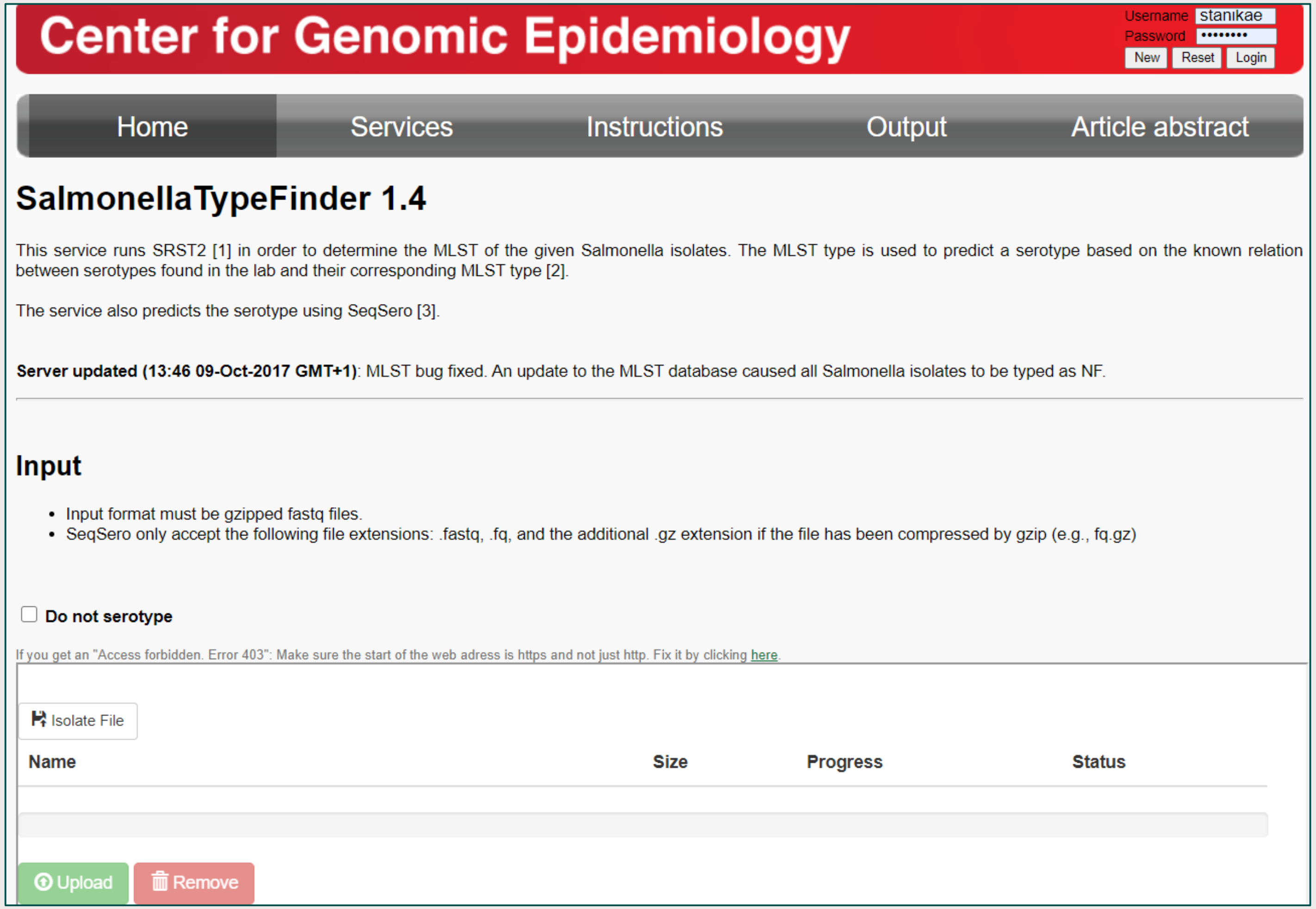
Uploaded genome analysis results

		Serovar Prediction								cgMLST 330 Allele Stats		
Genome	Subspecies	Serovar (overall)	Serovar (cgMLST)	Serovar (antigen)	Serogroup	H1	H2	Serovar (MLST)	MLST ST	Complete	Partial	Missing
ecoli		-:-:-	Mississippi (dist=0.98; 5/330)	-:-:-			-			315	0	15
salmonella_2	enterica	Orion	Orion (dist=0.39; 201/330)	Gatineau|Orion	E1	y	1,5	Orion	639	330	0	0

Mash RefSeq results

Your Genome	Strain	Serovar	Subspecies	Mash Distance	Matching	Taxonomy ID	BioProject	BioSample	Assembly Accession	Plasmid
ecoli	./rcn/refseq-NZ-1005526-PRJNA224116-SAMN02435983-GCF_000267405.1-.-Escherichia_coli_TW14301.fna			3.15e-4	396/400	1005526	PRJNA224116	SAMN02435983	GCF_000267405.1	
salmonella_2	ATCC_51959	Indiana	enterica	1.4e-2	266/400	1173850	PRJNA225738	SAMN01823715	NZ_AOZC	

Web-based tools for *Salmonella* serotyping



Username stanikae
Password ••••••••

SalmonellaTypeFinder 1.4

This service runs SRST2 [1] in order to determine the MLST of the given Salmonella isolates. The MLST type is used to predict a serotype based on the known relation between serotypes found in the lab and their corresponding MLST type [2].

The service also predicts the serotype using SeqSero [3].

Server updated (13:46 09-Oct-2017 GMT+1): MLST bug fixed. An update to the MLST database caused all Salmonella isolates to be typed as NF.

Input

- Input format must be gzipped fastq files.
- SeqSero only accept the following file extensions: .fastq, .fq, and the additional .gz extension if the file has been compressed by gzip (e.g., fq.gz)

□ **Do not serotype**

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

https://cge.cbs.dtu.dk/services/SalmonellaTypeFinder-1.4/

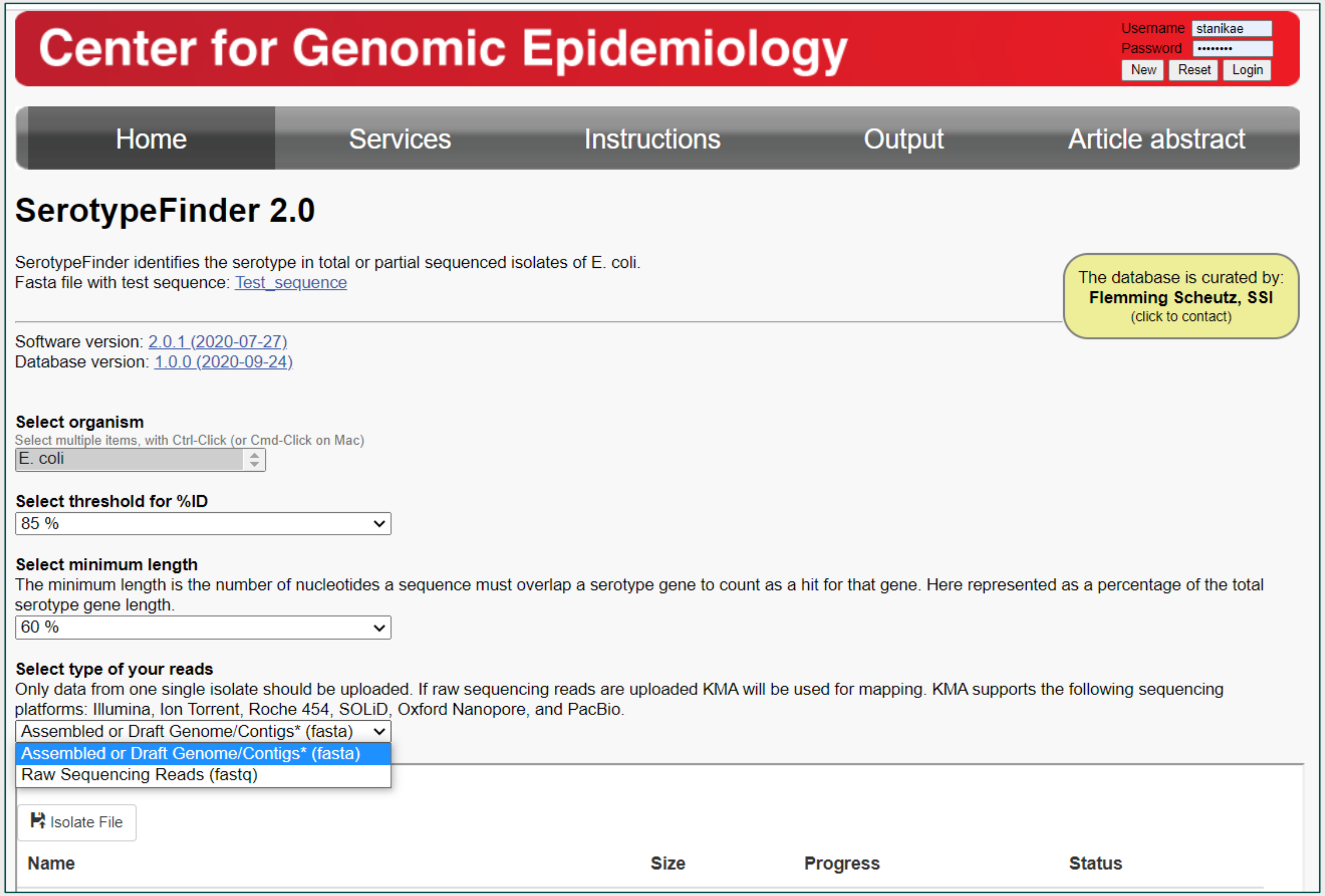
Serotyping in *E. coli*

- WGS of *E. coli* is replacing established subtyping methods such as PFGE
  - Traditional methods such as use of antibodies to test for surface antigens e.g. O, H, and K antigens

- Approx. 190 known *E. coli* serovars
  - Currently ~186 different O-groups and 53 H-types
  - Serotyping is highly complex

Fratamico et al. Front Microbiol (2016)
Uelze *et al.* One Health Outlook (2020)

Web-based tools for *E. coli* serotyping



Username: stanikae
Password: ••••••••

SerotypeFinder 2.0

SerotypeFinder identifies the serotype in total or partial sequenced isolates of E. coli.
Fasta file with test sequence: Test_sequence

The database is curated by:
Flemming Scheutz, SSI
(click to contact)

Software version: 2.0.1 (2020-07-27)
Database version: 1.0.0 (2020-09-24)

Select organism
Select multiple items, with Ctrl-Click (or Cmd-Click on Mac)
E. coli

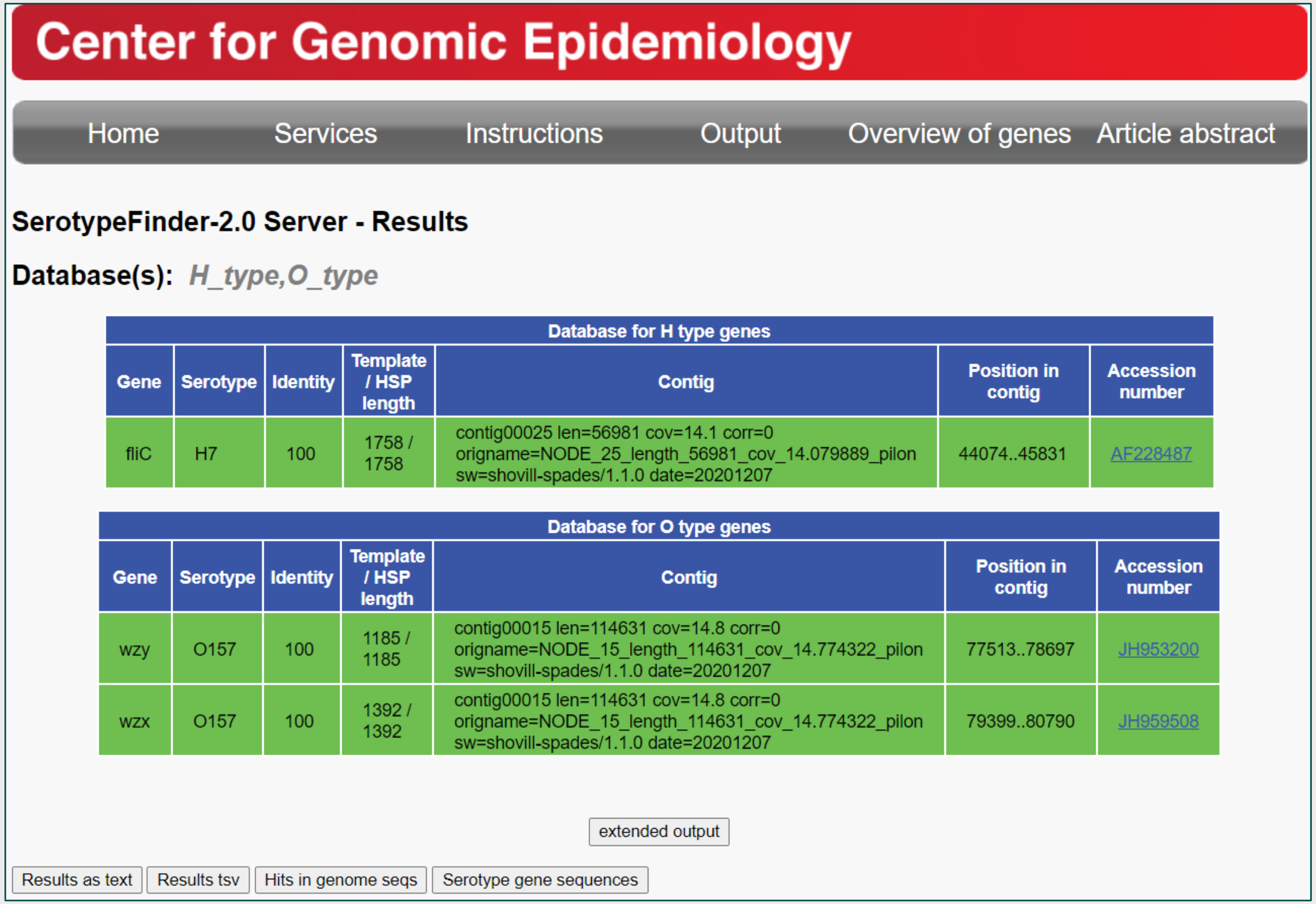
Select threshold for %ID
85 %

Select minimum length
The minimum length is the number of nucleotides a sequence must overlap a serotype gene to count as a hit for that gene. Here represented as a percentage of the total serotype gene length.
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* (fasta)
- Assembled or Draft Genome/Contigs* (fasta)
- Raw Sequencing Reads (fastq)

https://cge.cbs.dtu.dk/services/SerotypeFinder/

Web-based tools for *E. coli* serotyping



SerotypeFinder-2.0 Server - Results

Database(s): *H_type,O_type*

Database for H type genes

Gene	Serotype	Identity	Template / HSP length	Contig	Position in contig	Accession number
fliC	H7	100	1758 / 1758	contig00025 len=56981 cov=14.1 corr=0 origname=NODE_25_length_56981_cov_14.079889_pilon sw=shovill-spades/1.1.0 date=20201207	44074..45831	AF228487

Database for O type genes

Gene	Serotype	Identity	Template / HSP length	Contig	Position in contig	Accession number
wzy	O157	100	1185 / 1185	contig00015 len=114631 cov=14.8 corr=0 origname=NODE_15_length_114631_cov_14.774322_pilon sw=shovill-spades/1.1.0 date=20201207	77513..78697	JH953200
wzx	O157	100	1392 / 1392	contig00015 len=114631 cov=14.8 corr=0 origname=NODE_15_length_114631_cov_14.774322_pilon sw=shovill-spades/1.1.0 date=20201207	79399..80790	JH959508

https://cge.cbs.dtu.dk/services/SerotypeFinder/

The Fleming Fund
Regional Grants

Thank you

NOGUCHI MEMORIAL INSTITUTE FOR MEDICAL RESEARCH
UNIVERSITY OF GHANA , LEGON

UNIVERSITY OF IBADAN
RECTE SAPERE FONS

NATIONAL INSTITUTE FOR COMMUNICABLE DISEASES
Division of the National Health Laboratory Service

KCRI Kilimanjaro Clinical Research Institute
Academic Centre for Evidence Based Health Interventions
An Institution of the Good Samaritan Foundation

UKaid
from the British people

MOTT MACDONALD

SEQAFRICA

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.