Introduction to Web-based Analyses for Bacterial Typing

Erkison Ewomazino Odih

SEQAFRICA

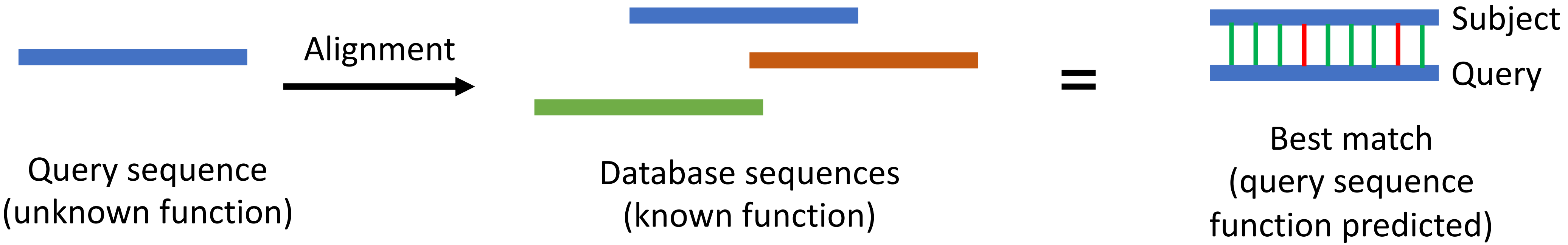
Outline

- General concepts
- Species identification
- Bacterial subtyping (MLST, serotyping)
- Practical

General concepts

Sequence Alignment (Applications)

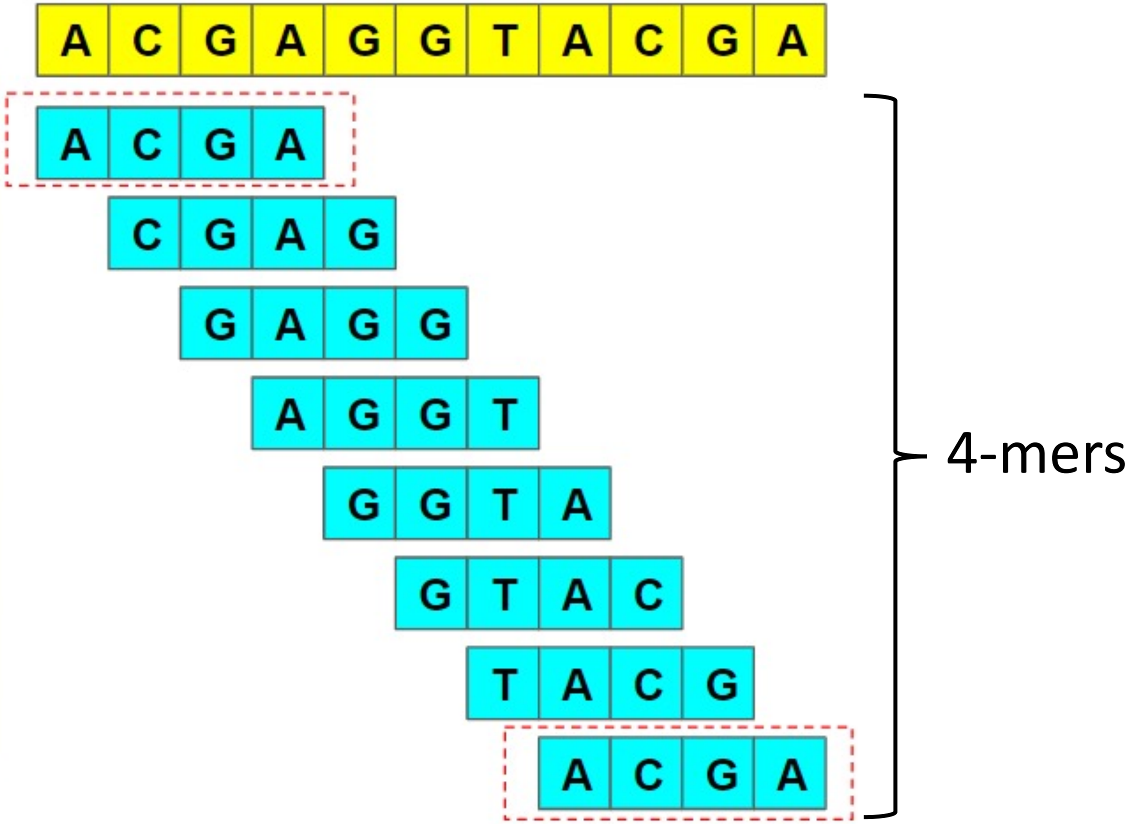
- Discover functional information (annotation)
- Predict molecular structure
- Predict shared ancestry
- Phylogenetic analysis
- Sequence typing and identification (speciation, MLST prediction etc)


Alignment
=
Subject
Query
Query sequence (unknown function)
Database sequences (known function)
Best match (query sequence function predicted)

Kmers

- Contiguous substrings of a given sequence, each having length k (positive integer)

- Similar sequences expected to share more kmers in common



A C G A G G T A C G A

A C G A

C G A G

G A G G

A G G T

G G T A

G T A C

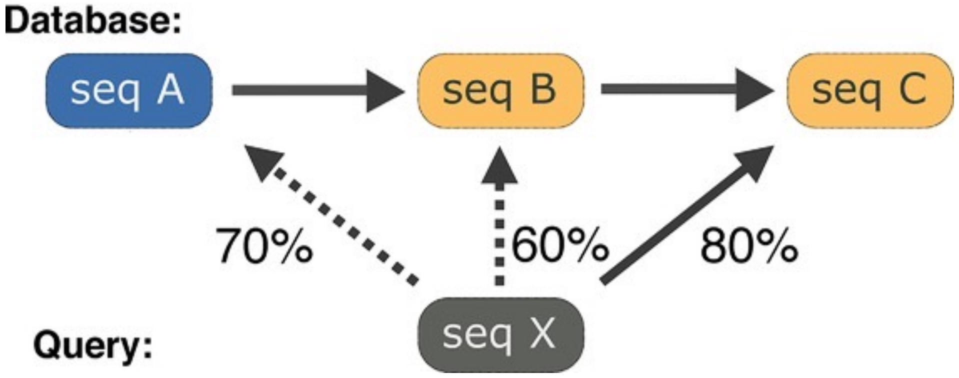
T A C G

A C G A

4-mers

Databases

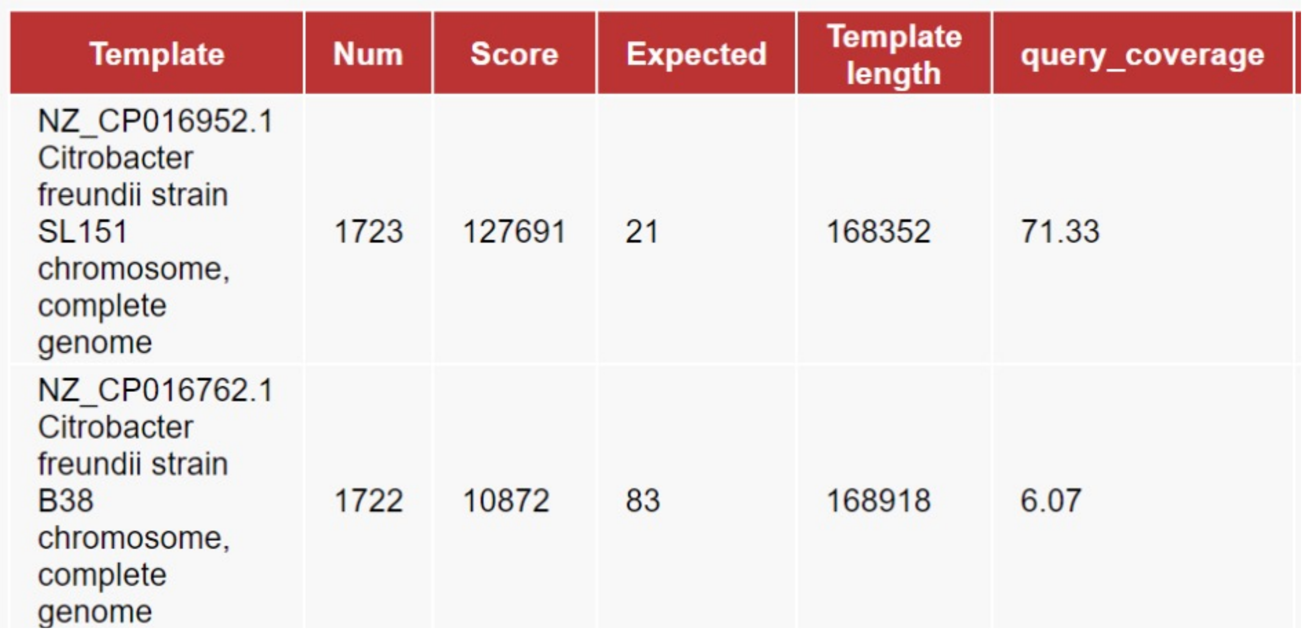
- Publicly accessible repository of [annotated] sequences
- Any prediction tool is only as good as its queried database
- Curated vs non-curated databases
- Find best matches for a query sequence in the database of sequences



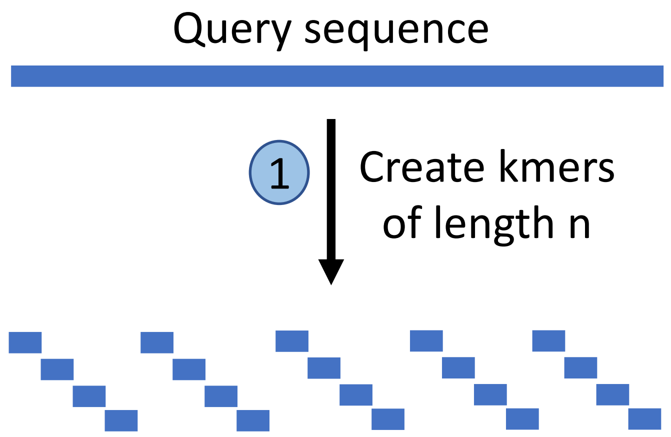
Species identification

KmerFinder Tool

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

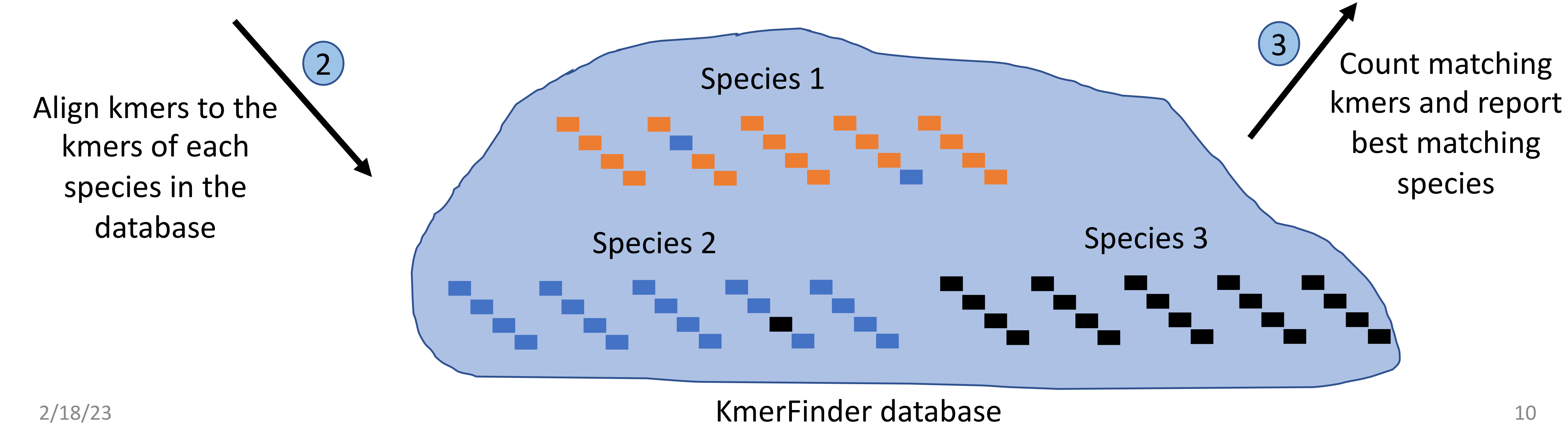


Template	Num	Score	Expected	Template length	query_coverage
NZ_CP016952.1 Citrobacter freundii strain SL151 chromosome, complete genome	1723	127691	21	168352	71.33
NZ_CP016762.1 Citrobacter freundii strain B38 chromosome, complete genome	1722	10872	83	168918	6.07



Query sequence

1 Create kmers of length n

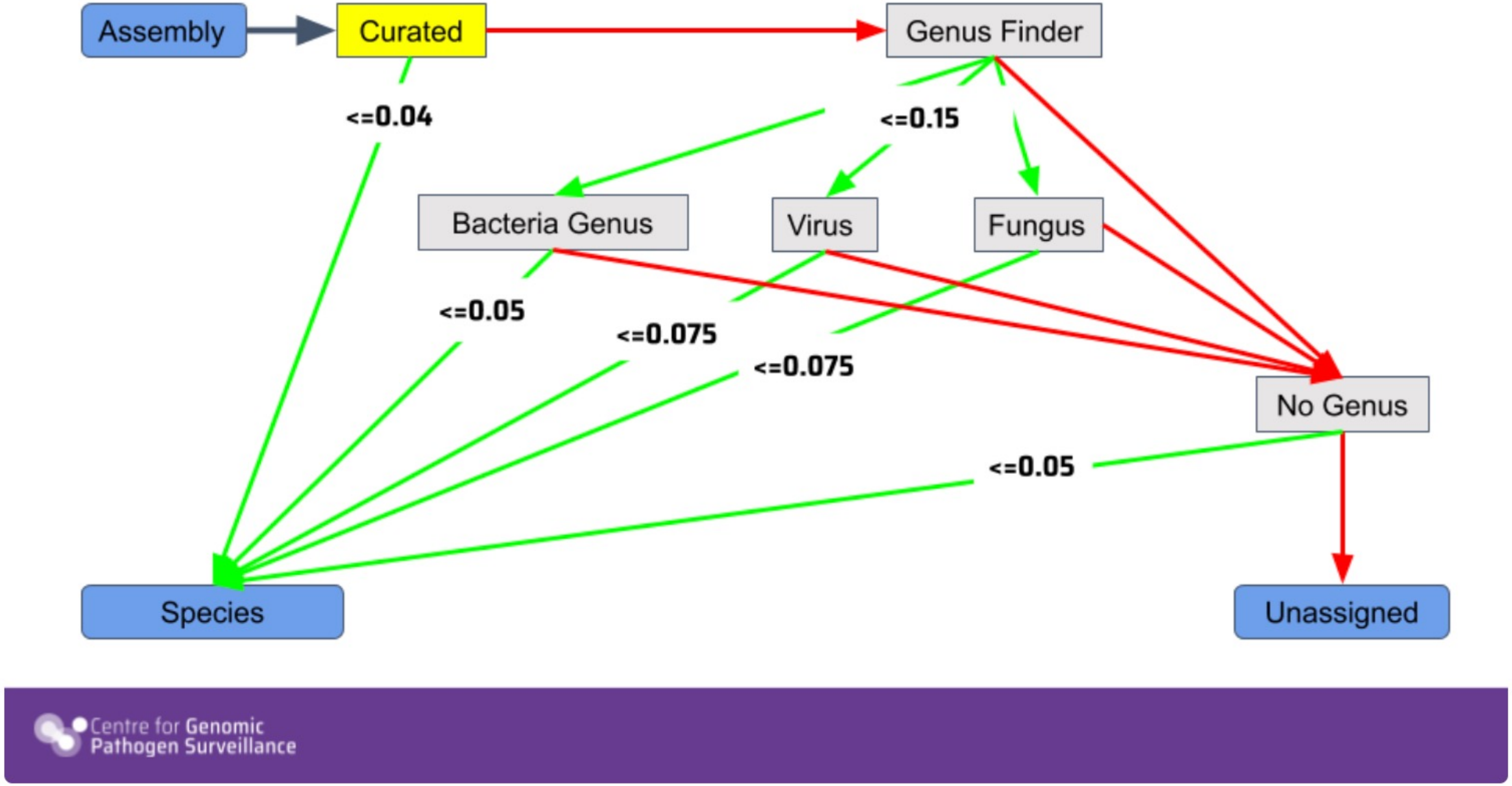


2 Align kmers to the kmers of each species in the database

3 Count matching kmers and report best matching species

Pathogenwatch (Speciator)

https://pathogen.watch



Bacterial subtyping

Why subtype?

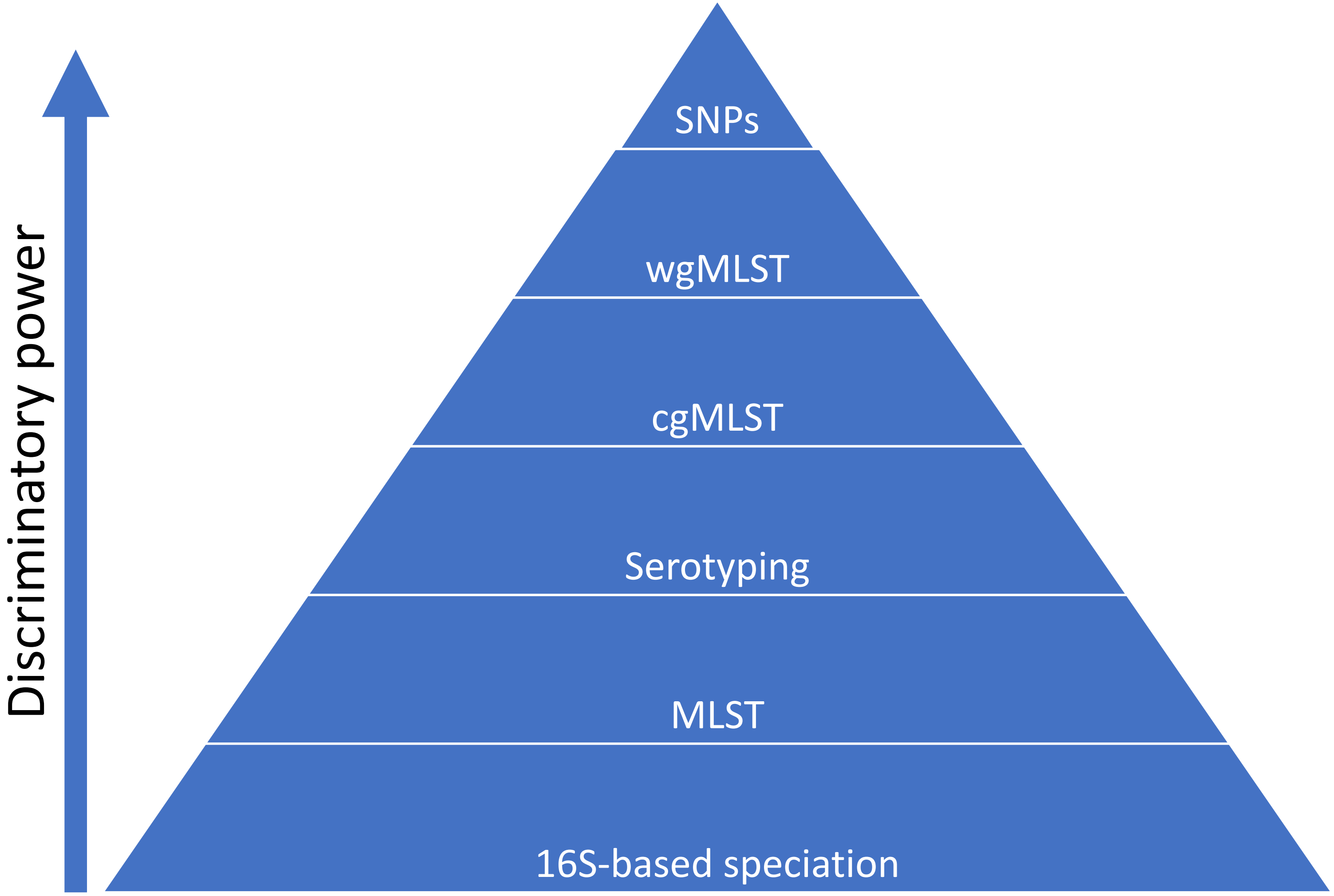
Sub-classification of separate bacterial strains within the same species

- Delineate virulent subtypes (e.g., *E. coli* O157:H7)
- Epidemiological surveillance
- Outbreak investigation
- Identify emerging pathogenic strains (e.g., hybrid pathotypes)

Sequence-based subtyping

Merits:

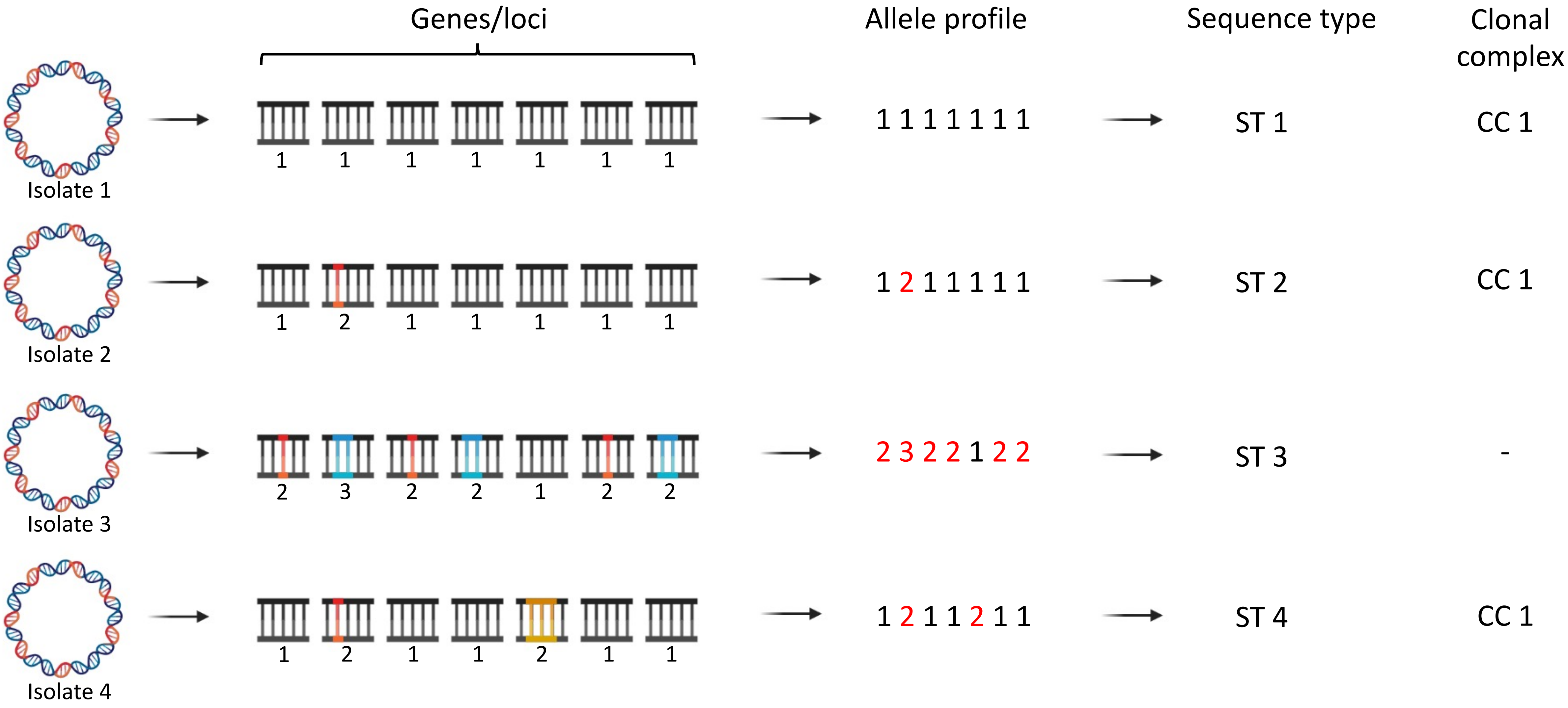
- Not all genetic changes translate to morphological change
- Morphological convergence
- More “characters” compared
- Increased WGS access



SNPs – single nucleotide polymorphisms; wg – whole genome; cg – core genome

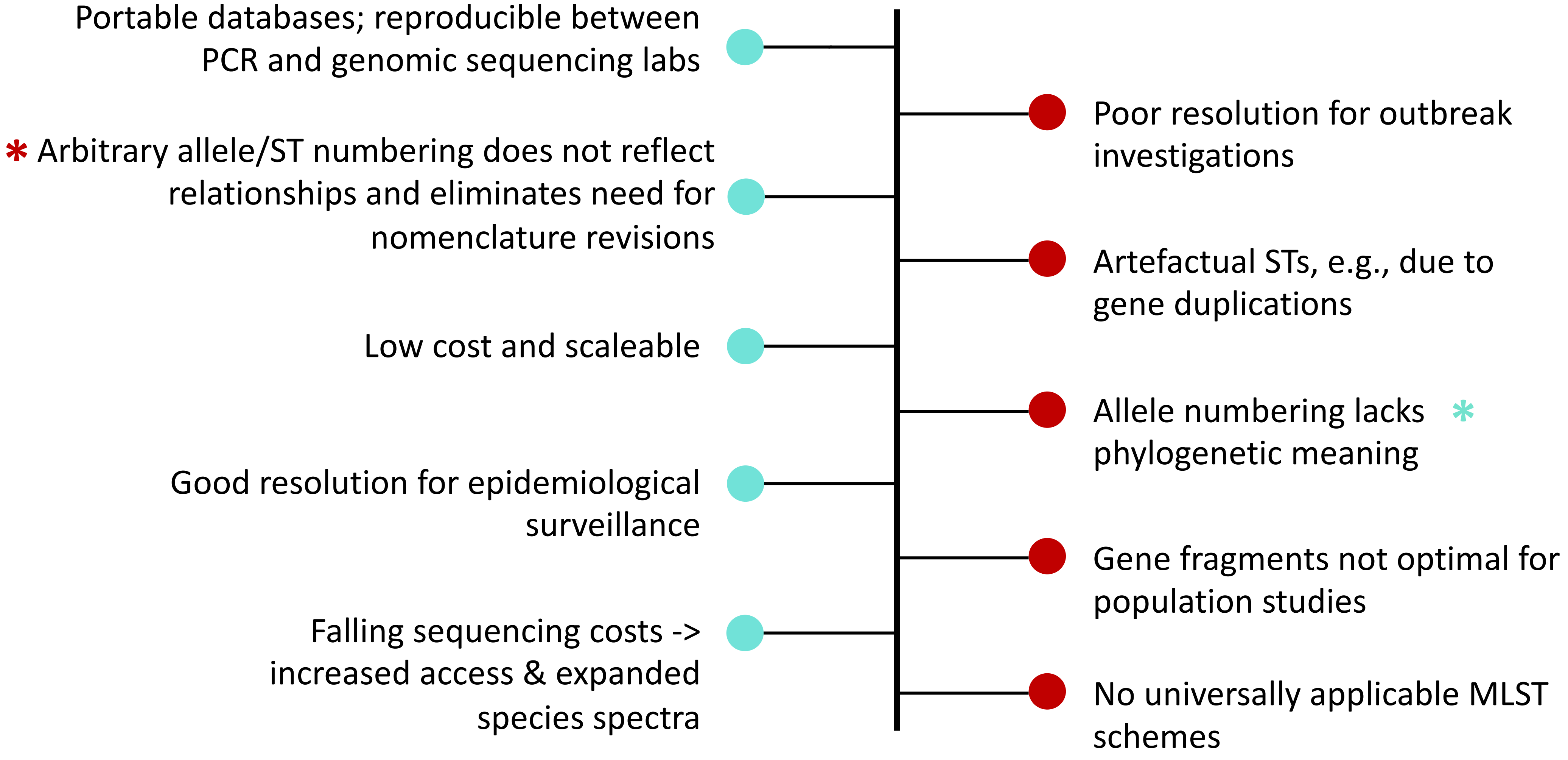
Multi-locus sequence typing (MLST)

- Whole genome- or PCR-based sequencing of selected (n) housekeeping genes
- Genetic changes in these genes are constrained but discriminatory
- Multiple loci (usually 7) studied to address lack of congruence in bacteria
- Typically compares sequences of gene fragments (300 – 550 bp)
- Allele numbers for each loci assigned in order of discovery
- All allele numbers form an allelic profile
- Each ST corresponds to a unique allelic profile; STs are also assigned in order of discovery; database updated with new alleles and STs
- Clonal complexes typically defined as clusters with 1 or 2 varying loci



	Genes/loci	Allele profile	Sequence type	Clonal complex
Isolate 1	1 1 1 1 1 1 1	1 1 1 1 1 1 1	ST 1	CC 1
Isolate 2	1 2 1 1 1 1 1	1 2 1 1 1 1 1	ST 2	CC 1
Isolate 3	2 3 2 2 1 2 2	2 3 2 2 1 2 2	ST 3	-
Isolate 4	1 2 1 1 2 1 1	1 2 1 1 2 1 1	ST 4	CC 1

MLST: Pros and cons



Portable databases; reproducible between PCR and genomic sequencing labs

* Arbitrary allele/ST numbering does not reflect relationships and eliminates need for nomenclature revisions

Low cost and scaleable

Good resolution for epidemiological surveillance

Falling sequencing costs -> increased access & expanded species spectra

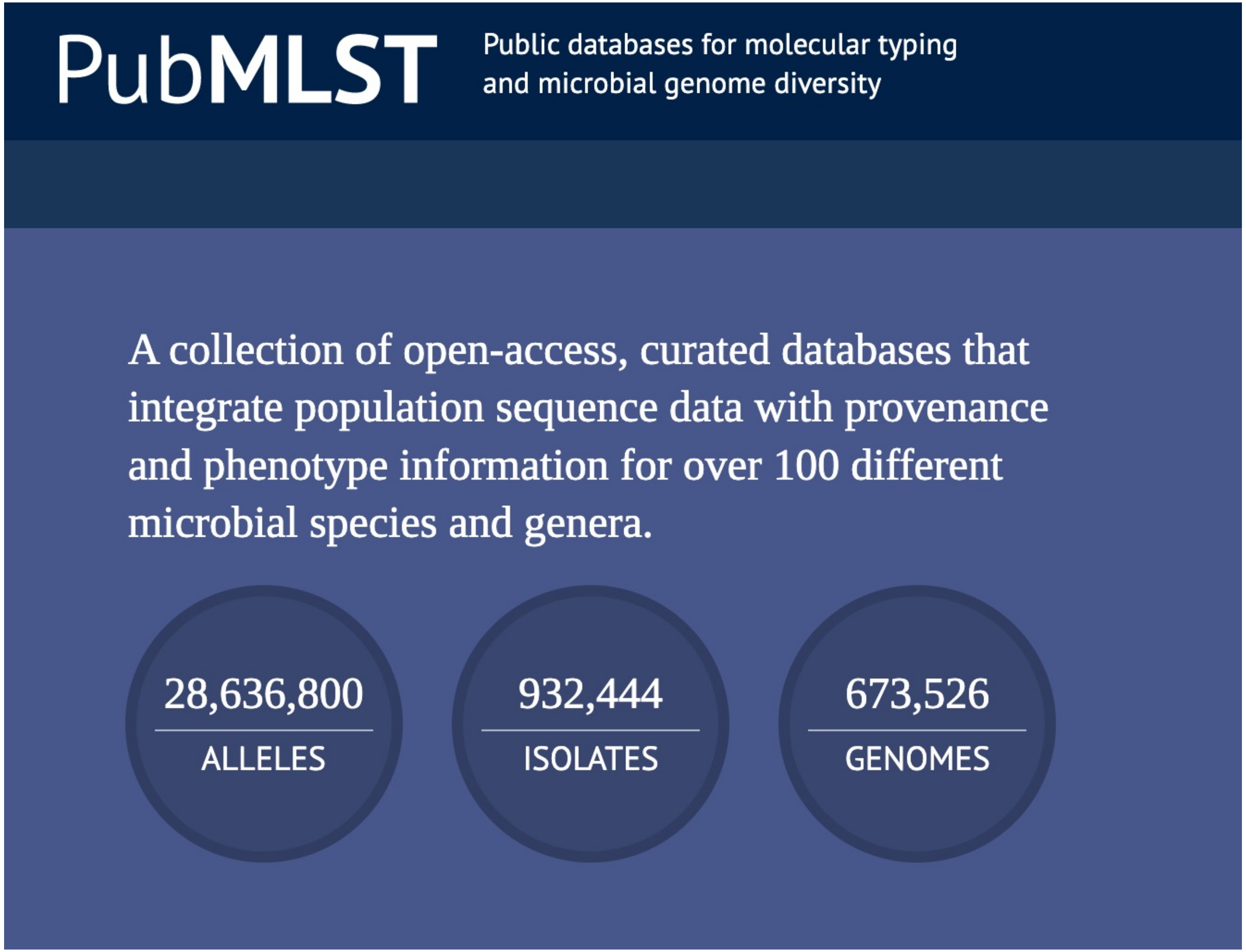
Poor resolution for outbreak investigations

Artefactual STs, e.g., due to gene duplications

Allele numbering lacks phylogenetic meaning *

Gene fragments not optimal for population studies

No universally applicable MLST schemes



PubMLST

Public databases for molecular typing and microbial genome diversity

A collection of open-access, curated databases that integrate population sequence data with provenance and phenotype information for over 100 different microbial species and genera.

28,636,800 ALLELES

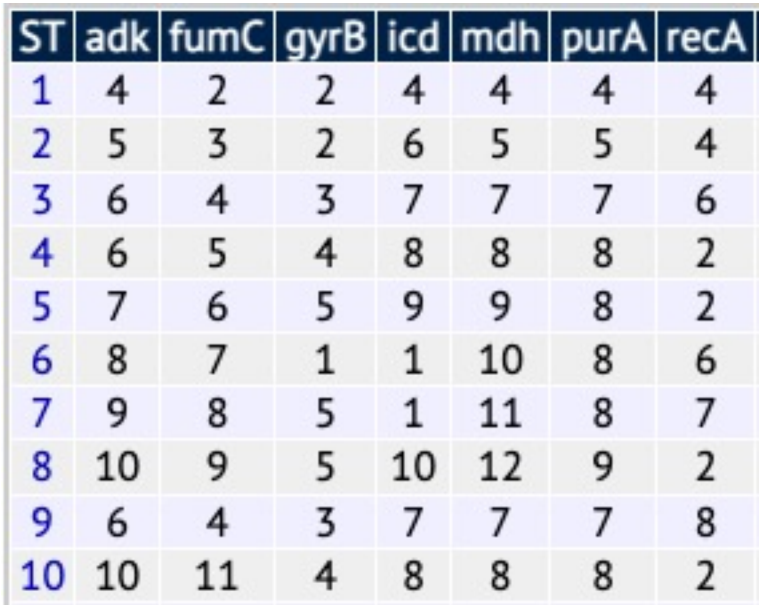
932,444 ISOLATES

673,526 GENOMES

pubmlst.org

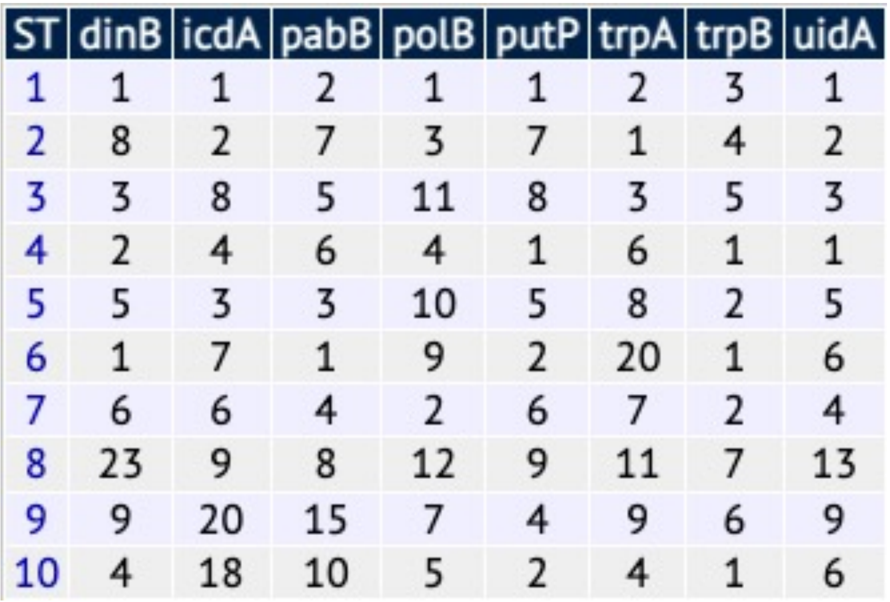
Escherichia coli MLST schemes (pubmlst.org)

Achtman scheme



ST	adk	fumC	gyrB	icd	mdh	purA	recA
1	4	2	2	4	4	4	4
2	5	3	2	6	5	5	4
3	6	4	3	7	7	7	6
4	6	5	4	8	8	8	2
5	7	6	5	9	9	8	2
6	8	7	1	1	10	8	6
7	9	8	5	1	11	8	7
8	10	9	5	10	12	9	2
9	6	4	3	7	7	7	8
10	10	11	4	8	8	8	2

Pasteur scheme



ST	dinB	icdA	pabB	polB	putP	trpA	trpB	uidA
1	1	1	2	1	1	2	3	1
2	8	2	7	3	7	1	4	2
3	3	8	5	11	8	3	5	3
4	2	4	6	4	1	6	1	1
5	5	3	3	10	5	8	2	5
6	1	7	1	9	2	20	1	6
7	6	6	4	2	6	7	2	4
8	23	9	8	12	9	11	7	13
9	9	20	15	7	4	9	6	9
10	4	18	10	5	2	4	1	6

https://pubmlst.org/bigsdb?db=pubmlst_escherichia_seqdef
https://enterobase.readthedocs.io/en/latest/mlst/mlst-legacy-info-ecoli.html

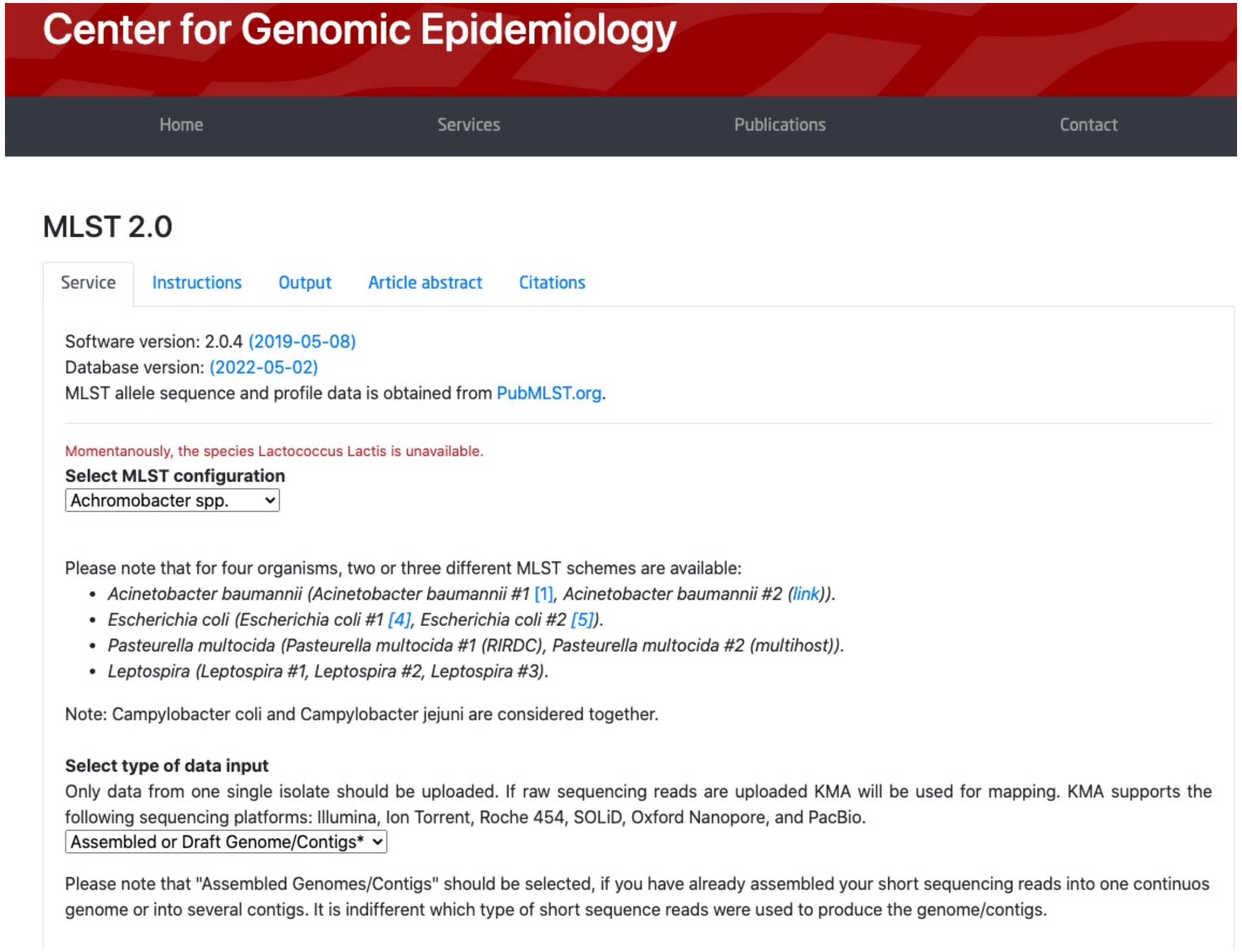
MLST 2.0 (CGE)

Platform: Web-based

Input: Sequence reads (.fastq) OR assemblies (.fasta)

URL: https://cge.food.dtu.dk/services/MLST/

Comments: Available for many clinically important species; regularly updated DB



MLST 2.0

Software version: 2.0.4 (2019-05-08)
Database version: (2022-05-02)
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 (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*

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.

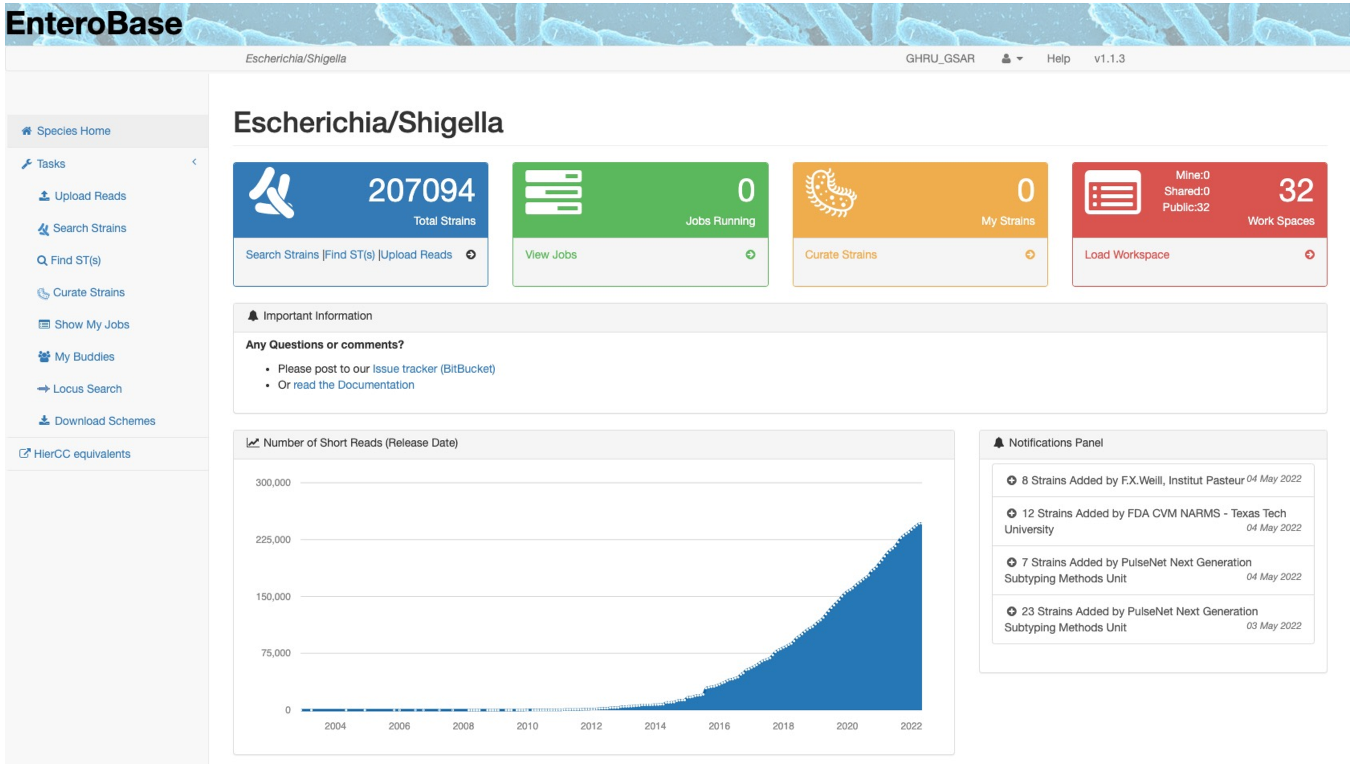
Enterobase

Platform: Web-based

Input: Sequence reads (fastq)

URL: https://enterobase.warwick.ac.uk/

Comments: Specific for *E. coli*, *Salmonella*, etc. Also generates cgMLST, serotyping results, etc.



Any Questions or comments?

- Please post to our Issue tracker (BitBucket)
- Or read the Documentation

8 Strains Added by F.X.Weill, Institut Pasteur 04 May 2022

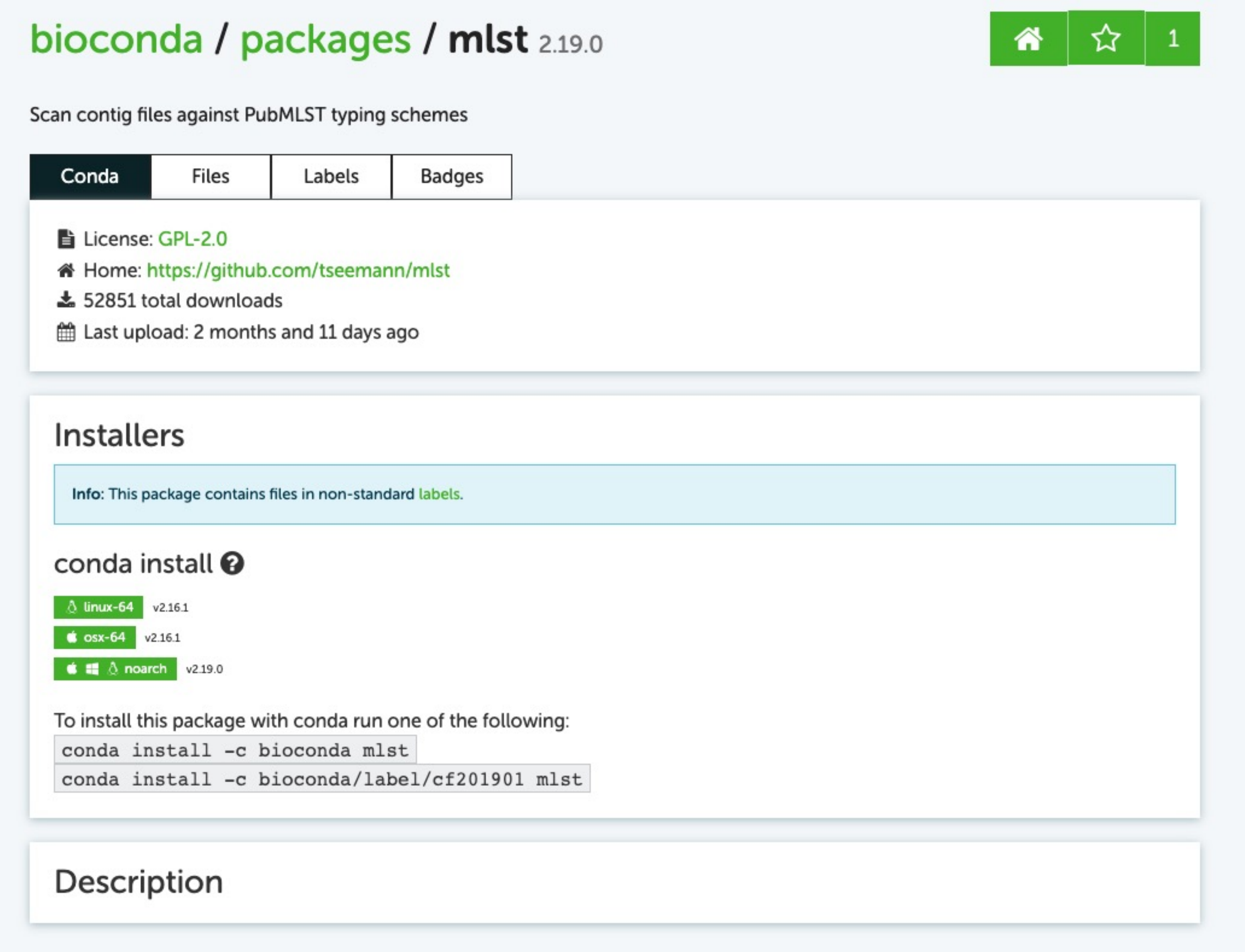
12 Strains Added by FDA CVM NARMS - Texas Tech University 04 May 2022

7 Strains Added by PulseNet Next Generation Subtyping Methods Unit 04 May 2022

23 Strains Added by PulseNet Next Generation Subtyping Methods Unit 03 May 2022

MLST

Platform:	CLI (Linux/OS X)
Input:	Assembled and/or annotated genomes (fasta/GenBank/EMBL)
URL:	https://github.com/tseemann/mlst
Comments:	Preferred for large datasets; manually update database once installed



bioconda / packages / mlst 2.19.0

Scan contig files against PubMLST typing schemes

License: GPL-2.0
Home: https://github.com/tseemann/mlst
52851 total downloads
Last upload: 2 months and 11 days ago

Installers

Info: This package contains files in non-standard labels.

conda install

linux-64	v2.16.1
osx-64	v2.16.1
noarch	v2.19.0

To install this package with conda run one of the following:

```
conda install -c bioconda mlst
conda install -c bioconda/label/cf201901 mlst
```

Description

E. coli serotyping

- Serological typing based on differences in the lipopolysaccharide O antigen, capsular K antigen, and flagellar H antigen.
- Serotype information very useful epidemiologically because it is directly linked to antigenic response
- Poor phylogenetic correlation due to high propensity for recombination in these genes
- WGS serotyping based on sequence similarity
- O:H typing (standard serotyping) from WGS data:
  - O group:
    - O-antigen processing genes: *wzx, wzy, wzm*, and *wzt*
  - H group:
    - Flagellin-encoding genes – *fliC, flkA, fllA, flmA* and *flnA*

Salmonella serotyping

- Phenotypic serotyping
  - Labor-intensive and expensive
  - Requires procurement and storage of multiple antisera
- Serovars designated by names or antigenic formula in the format: **O:H1:H2**
- Over 2500 serotypes in White-Kauffmann-Le Minor scheme
- Specific combinations of O and H antigenic types represent serotypes/serovars
- Genome-based serotyping targets same antigens as phenotypic assays:
  - Somatic (O) group antigen – *rfb* gene cluster (*wzx, wzy,* others)
  - Flagellar antigens – *fliC* and *fljB*

Zhang et al (2015) https://doi.org/10.1128/JCM.00323-15

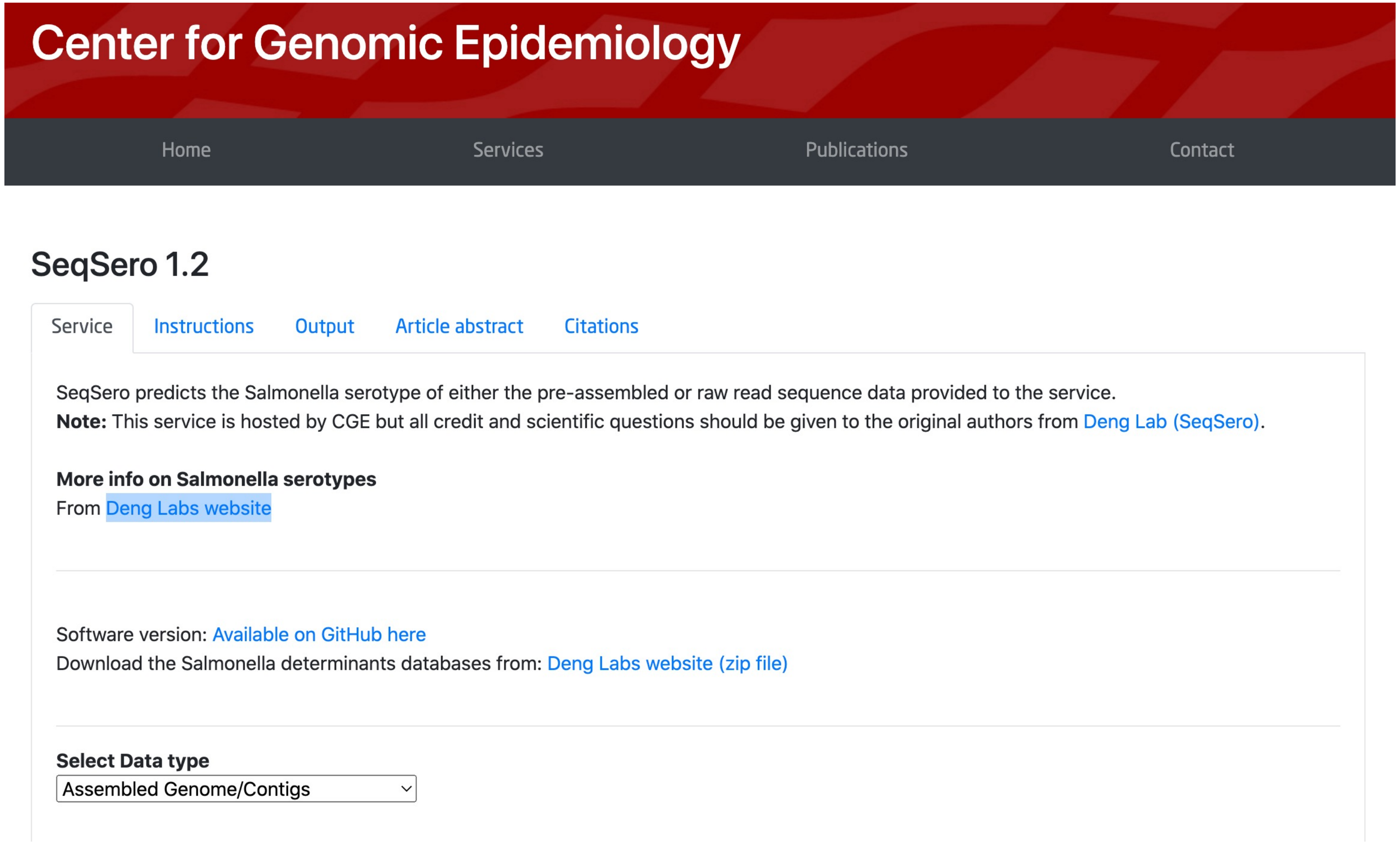
SeqSero (CGE)

Platform: Web-based

Input: Sequence reads (.fastq) OR assemblies (.fasta)

URL: https://cge.food.dtu.dk/services/SeqSero/

Comments: *Salmonella* serotyping



SeqSero 1.2

SeqSero predicts the Salmonella serotype of either the pre-assembled or raw read sequence data provided to the service.
Note: This service is hosted by CGE but all credit and scientific questions should be given to the original authors from Deng Lab (SeqSero).

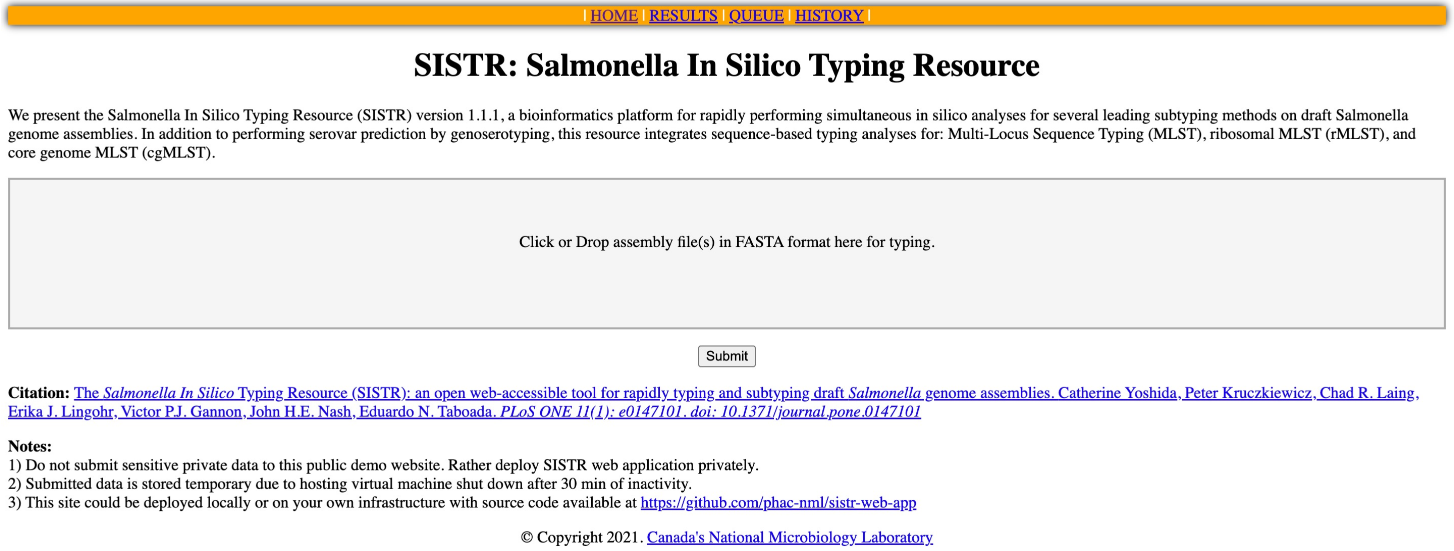
More info on Salmonella serotypes
From Deng Labs website

Software version: Available on GitHub here
Download the Salmonella determinants databases from: Deng Labs website (zip file)

Select Data type
Assembled Genome/Contigs

SISTR

Platform:	Web and CLI (Linux / OS X)
Input:	Assemblies (fasta)
URL:	https://github.com/phac-nml/sistr_cmd https://sistr-app.herokuapp.com/
Comments:	*Salmonella* typing: serovar and serogroup prediction, cgMLST, etc.



SISTR: Salmonella In Silico Typing Resource

We present the Salmonella In Silico Typing Resource (SISTR) version 1.1.1, a bioinformatics platform for rapidly performing simultaneous in silico analyses for several leading subtyping methods on draft Salmonella genome assemblies. In addition to performing serovar prediction by genoserotyping, this resource integrates sequence-based typing analyses for: Multi-Locus Sequence Typing (MLST), ribosomal MLST (rMLST), and core genome MLST (cgMLST).

Click or Drop assembly file(s) in FASTA format here for typing.

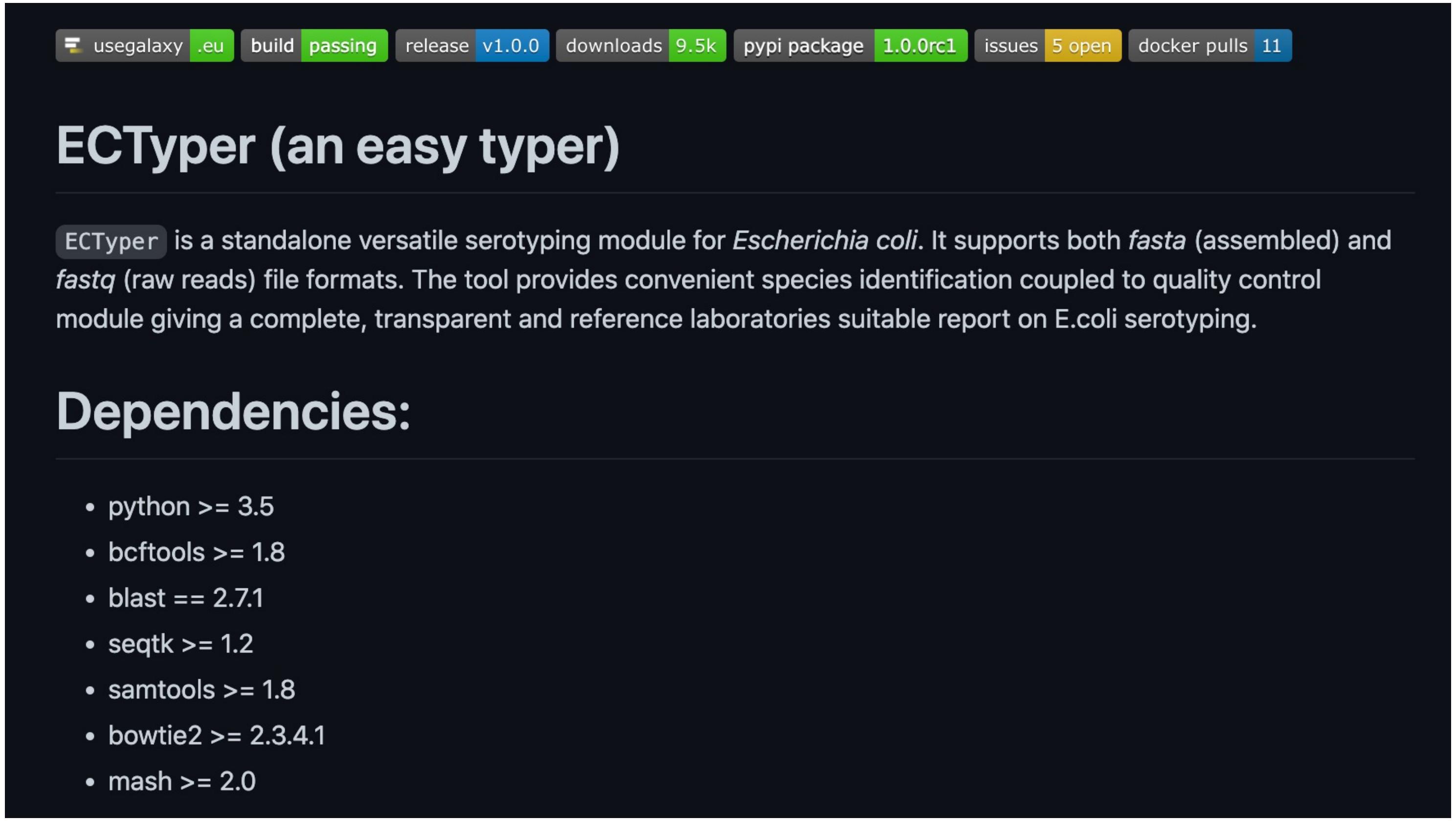
Citation: The *Salmonella In Silico* Typing Resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. Catherine Yoshida, Peter Kruczkiewicz, Chad R. Laing, Erika J. Lingohr, Victor P.J. Gannon, John H.E. Nash, Eduardo N. Taboada. *PLoS ONE 11(1): e0147101. doi: 10.1371/journal.pone.0147101*

Notes:
1) Do not submit sensitive private data to this public demo website. Rather deploy SISTR web application privately.
2) Submitted data is stored temporary due to hosting virtual machine shut down after 30 min of inactivity.
3) This site could be deployed locally or on your own infrastructure with source code available at https://github.com/phac-nml/sistr-web-app

© Copyright 2021. Canada's National Microbiology Laboratory

ECTyper

Platform:	Web and CLI
Input:	Assemblies (fasta) or raw reads (fastq)
URL:	https://github.com/phac-nml/ecoli_serotyping https://usegalaxy.eu/root?tool_id=ectyper
Comments:	*E. coli* serotyping



ECTyper (an easy typer)

`ECTyper` is a standalone versatile serotyping module for *Escherichia coli*. It supports both *fasta* (assembled) and *fastq* (raw reads) file formats. The tool provides convenient species identification coupled to quality control module giving a complete, transparent and reference laboratories suitable report on E.coli serotyping.

Dependencies:

- python >= 3.5
- bcftools >= 1.8
- blast == 2.7.1
- seqtk >= 1.2
- samtools >= 1.8
- bowtie2 >= 2.3.4.1
- mash >= 2.0

Questions

Practical

- CGE tools: https://cge.food.dtu.dk/services/
    - Species identification
    - MLST determination
    - Serotyping (if applicable)
- Multi-analyses tools:
    - Pathogenwatch [and Kleborate] - https://pathogen.watch/
    - Enterobase - https://enterobase.warwick.ac.uk/