Bioinformatics tools for WGS analysis of infectious diseases

Pimlapas Leekitcharoenphon (Shinny)
Researcher
Research Group for Genomic Epidemiology
DTU Food

Whole-genome sequence based species ID using KmerFinder

KmerFinder - bitbucket

https://bitbucket.org/genomicepidemiology/kmerfinder/src/master/

Installation

Setting up KmerFinder program

```
# Go to wanted location for resfinder
cd /path/to/some/dir
# Clone and enter the KmerFinder directory
git clone https://bitbucket.org/genomicepidemiology/kmerfinder.git
cd kmerfinder
```

Build Docker image from Dockerfile

```
# Build container
docker build -t kmerfinder .
# Run test
docker run --rm -it \
       --entrypoint=/test/test.sh kmerfinder
```

KmerFinder database - bitbucket

https://bitbucket.org/genomicepidemiology/kmerfinder_db/src/master/

Installation - Clone repository

```
# Go to wanted location for KmerFinder db
cd /path/to/some/dir
# Clone and enter the KmerFinder db directory
git clone https://bitbucket.org/genomicepidemiology/kmerfinder_db.git
cd kmerfinder_db
```

K-mer ?

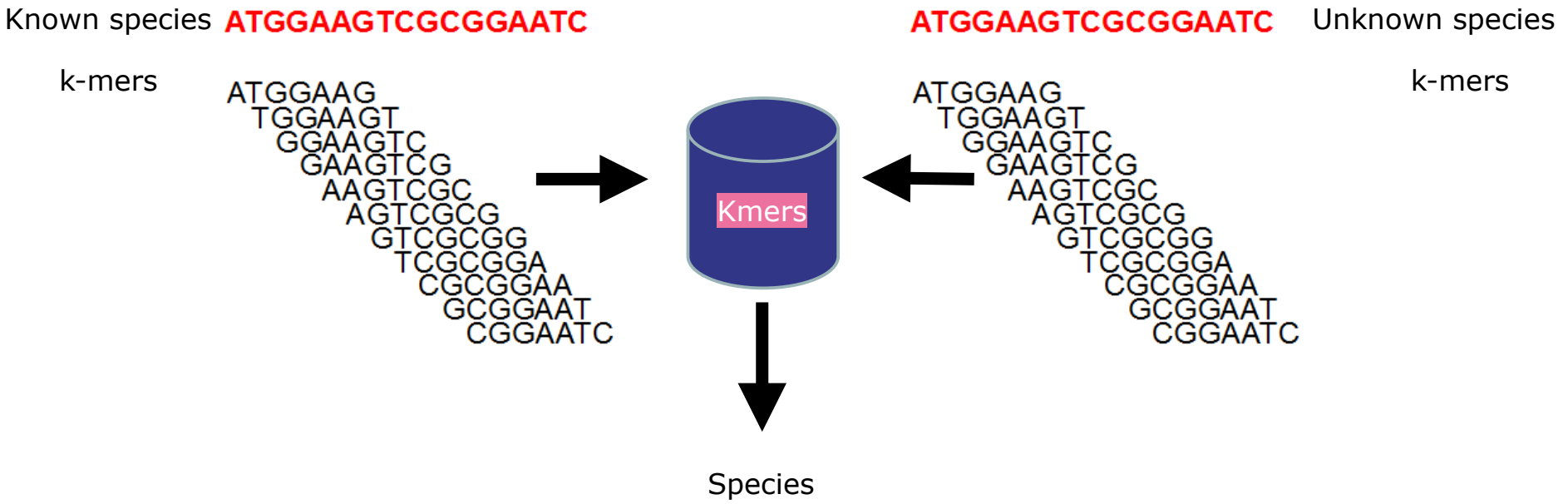
- A k-mer is a contiguous sequence of k bases
- k is any positive integer
- Sequences with high similarity must share k-mers

sequence **ATGGAAGTCGCGGAATC**

7 mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

Species identification by K-mer

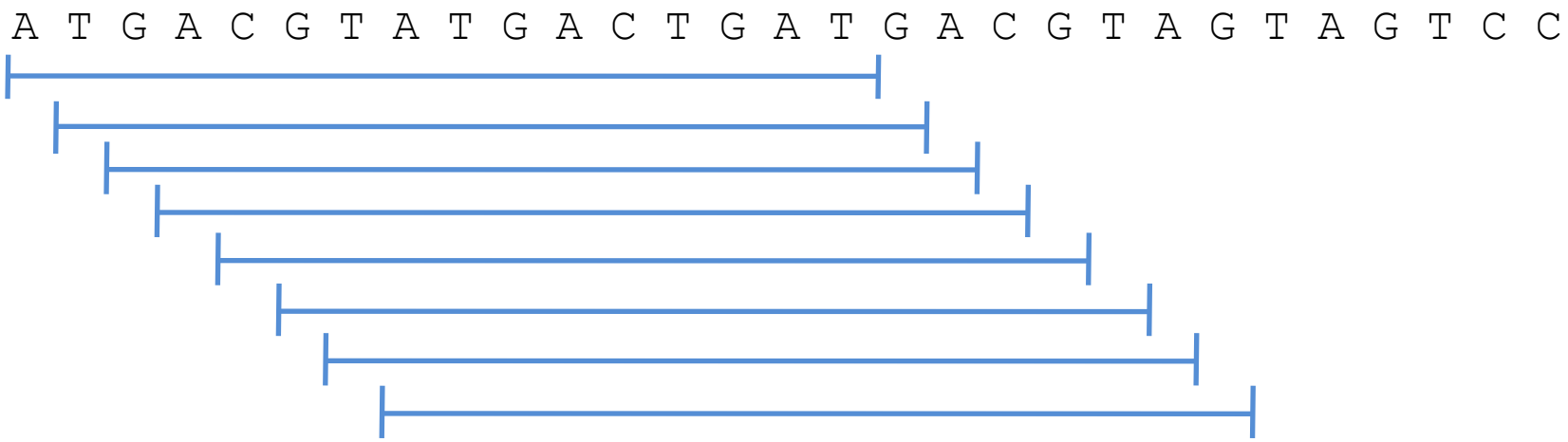


KmerFinder

Using all information in the WGS data

almost

- Genomes is split into 16mers:



- It is necessary to reduce the total amount of 16mers:
  - Only 16mers with particular prefix are (ATGAC) kept

Query bacteria of unknown species

Reference db bacteria of known species (template)

Bact1->*E. coli*

Bact2->*S. enterica*

KMA (k-mer alignment)

Bact3->*K. pneumoniae*

Bact4->S. aureus

Prediction: Query bacteria is a *S. aureus*

16mer database

CP001921 (*Acinetobacter baumanii*)
CP000521 (*Acinetobacter baumanii*)
CP002522 (*Acinetobacter baumanii*)

CP001921 (*Acinetobacter baumanii*)
CP002301 (*Buchnera aphidicola*)

Unknown bacteria

16 mers:

ATGAATGTGTGAGTGA

ATGACTGTGCCCCTGA

ATGAAAAAAAAAAAA

Species	Match	No. of Kmer hits
Acinetobacter baumannii	CP001921	2
Acinetobacter baumannii	CP000521	1
Acinetobacter baumannii	CP002521	1
Buchnera aphidicola	CP002301	1

KmerFinder output – standard scoring method

Center for Genomic Epidemiology

Home | Services | Instructions | Output

KmerFinder-3.1 Server - Results

KmerFinder 3.1 results:

Template	Num	Score	Expected	Template_length	Query_Coverage	Template_Coverage	Depth
NC_016854.1 Salmonella enterica subsp. enterica serovar Typhimurium str. D23580 complete genome	7004	6094006	30	157485	96.86	99.99	38.70

EXTENDED OUTPUT

Input Files: *Salmonella-spp-02-03-002_R1_001.trim.fq* *Salmonella-spp-02-03-002_R2_001.trim.fq*

RESULTS as text (tab separated)

Other KmerFinder statistics

$$\text{Query coverage} = \frac{\text{Score (total number of kmers in query sequence that match kmers in template sequence)}}{\text{Total number of kmers in query sequence}}$$

$$\text{Template coverage} = \frac{\text{Score (total number of kmers in template sequence that match kmers in query sequence)}}{\text{Total number of kmers in template sequence (database seqeunce)}}$$

$$\text{Depth} = \frac{\text{Score (total number of kmers in query sequence that match kmers in template sequence)}}{\text{Total number of kmers in template sequence (database seqeunce)}}$$

Explanation of the columns in standard and extended output

The following contains a briefly explanition of all columns of the output including the columns in the extended output

Template: shows the accession numbers or name of the template sequences
Assembly: RefSeq assembly accession ID
Num: is the sequence number of accession entry in the KmerFinder database
Score: is the total number of matching Kmers between the query and the template
Expected: is the expected score, i.e.the expected total number of matching Kmers between query and template (randomly selected).
Template length: is the number of Kmers in the template
query_coverage [%]: is the percentage of input query Kmers that match the template.
Coverage [%]: is the template coverage.
Depth: is the number of matched kmers in the query sequence divided by the total number of Kmers in the template. For read files this estimates the sequencing depth.
tot_query_coverage [%]: is calculated based on the ratio of the score and the number of kmers in the query sequence, where the score includes kmers matched before.
tot_coverage [%]: is calculated based on ratio of the score and the number of unique kmers in the template sequence, where the score includes kmers matched before.
tot_depth: depth value based on all query kmers that can be found in the template sequence .
q_value: is the quantile in a standard Pearson Chi-square test, to test whether the current template is a significant hit.
p_value: is the p-value corresponding to the obtained q_value.
Accession number: accession number of entry ID in fasta file.
Description: additional descriptions available in fasta file, or in the case of organism databases the identifier lines of fasta files.
TAXID: NCBI's TaxID number of the hit
Taxonomy: complete taxonomy of the hit
TAXID Species: NCBI's species TaxID number of the hit (sometimes bacterial strain or substrain TaxIDs can be given above)
Species: Species name

Dataset: 5 unknown genomes

`/home/projects/course_23262/course/week03/assembly`

`/home/projects/course_23262/course/week03/trimmed_reads`

Guildline on how to run WGS tool:
DTU Learn > Course Content > Content > Week 3 > Exercises

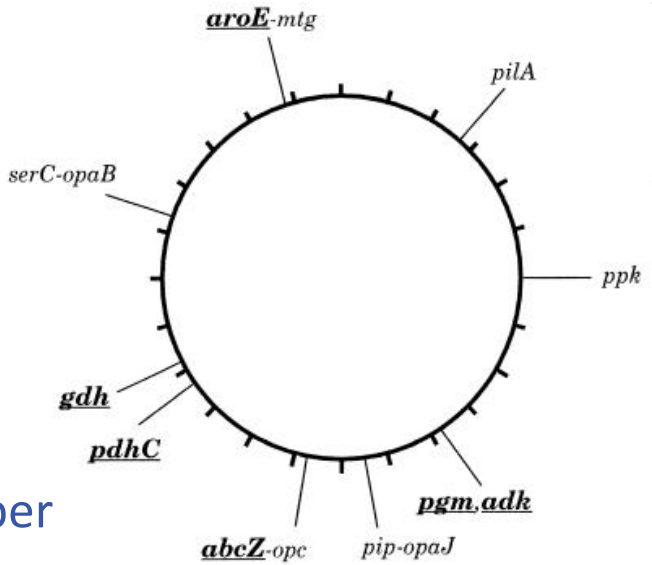
Output table

https://docs.google.com/spreadsheets/d/1WmmtTCFTVp7uOe9vYlRyDilyYHYbR-liPiEyoPkbZo4/edit?usp=sharing

MLST typing

Multilocus Sequence Typing (MLST) - the golden standard for typing

First developed in 1998 for *Neisseria meningitis*
(Maiden et al. PNAS 1998. 95:3140-3145)



❖ The nucleotide sequence of internal regions of app. 7 housekeeping genes are determined by PCR followed by Sanger sequencing

❖ Different alleles are each assigned a random number

❖ The unique combination of alleles is the sequence type (**ST**)

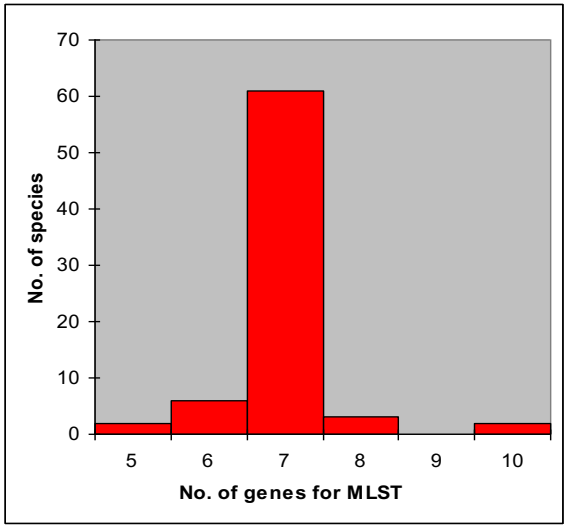
The first 4 ST profiles of *Neisseria meningitis*

	Loci						
Sequence type (ST)	*abcZ*	*adk*	*aroE*	*fumC*	*gdh*	*pdhC*	*pgm*
1	1	3	1	1	1	1	3
2	1	3	4	7	1	1	3
3	1	3	1	1	1	23	13
4	1	3	3	1	4	2	3

A total of 446,824 different sequences (alleles) and 11,113 different Sequence Types exist (13. Feb 2015)

MLST schemes for ~100 species have now been developed and incorporated in the MLST database (called MLST configurations)

For most species, seven genes are used



Original Article

Genomics of an emerging clone of *Salmonella* serovar Typhimurium ST313 from Nigeria and the Democratic Republic of Congo

Pimlapas Leekitcharoenphon[1,2], Carsten Friis[1], Ea Zankari[1], Christina A. Svendsen[1], Lance B. Price[3], Maral Rahmani[1], Ana Herrero-Fresno[1], Kayode Fashae[4], Olivier Vandenberg[5, 6], Frank M. Aarestrup[1], Rene S. Hendriksen[1]

[1]*Division of Bacterial Genomics and Epidemiology, WHO Collaborating Center for Antimicrobial Resistance in Food borne Pathogens and European Union Reference Laboratory for Antimicrobial Resistance, National Food Institute, Technical University of Denmark, Kemitorvet, Denmark*
[2]*Center for Biological Sequence Analysis, Department of System Biology, Technical University of Denmark, Kemitorvet, Denmark*
[3]*Division of Pathogen Genomics, Translational Genomics Research Institute (TGen), Flagstaff, USA*
[4]*Department of Microbiology, University of Ibadan, Ibadan, Nigeria.*
[5]*Infectious Diseases Epidemiological Unit, Public Health School, Université Libre de Bruxelles, Brussels, Belgium.*
[6]*Department of Microbiology, Saint-Pierre University Hospital, Brussels, Belgium.*

Abstract

Introduction: *Salmonella enterica* serovar Typhimurium ST313 is an invasive and phylogenetically distinct lineage present in sub-Saharan Africa. We report the presence of *S.* Typhimurium ST313 from patients in the Democratic Republic of Congo and Nigeria.

Methodology: Eighteen *S.* Typhimurium ST313 isolates were characterized by antimicrobial susceptibility testing, pulsed-field gel electrophoresis (PFGE), and multilocus sequence typing (MLST). Additionally, six of the isolates were characterized by whole genome sequence typing (WGST). The presence of a putative virulence determinant was examined in 177 *Salmonella* isolates belonging to 57 different serovars.

Results: All *S.* Typhimurium ST313 isolates harbored resistant genes encoded by bla_{TEM1b}, *cat*A1, *str*A/B, *sul*1, and *dfr*A1. Additionally, *aac(6')1aa* gene was detected. Phylogenetic analyses revealed close genetic relationships among Congolese and Nigerian isolates from both blood and stool. Comparative genomic analyses identified a putative virulence fragment (ST313-TD) unique to *S.* Typhimurium ST313 and *S.* Dublin.

Conclusion: We showed in a limited number of isolates that *S.* Typhimurium ST313 is a prevalent sequence-type causing gastrointestinal diseases and septicemia in patients from Nigeria and DRC. We found three distinct phylogenetic clusters based on the origin of isolation suggesting some spatial evolution. Comparative genomics showed an interesting putative virulence fragment (ST313-TD) unique to *S.* Typhimurium ST313 and invasive *S.* Dublin.

Implementation of MLST using WGS data

- Automatic, local download of MLST database once a week
- For each species, we get an ST profile table file and an allele file

ST profile table file

ST	leuA	petC	malF	cysG	holC	nuoL	gltT
1	1	1	1	1	1	1	1
2	1	1	4	1	1	1	1
3	1	1	1	20	1	1	1
4	1	1	1	4	1	1	1
5	2	2	2	2	2	2	2
6	3	3	3	3	3	3	3
7	3	3	3	7	3	3	3
9	3	3	5	5	4	3	4
10	5	4	3	3	6	3	5
11	7	7	7	9	10	8	8
13	7	6	7	9	10	7	8
14	8	8	8	11	12	9	9
16	7	6	8	10	11	8	8
17	1	1	10	12	18	10	1
18	9	1	9	13	14	5	10
19	10	1	10	14	15	11	1
20	1	1	10	12	17	11	11
21	10	1	10	14	15	11	12
26	5	3	3	3	6	3	5
33	11	9	14	15	19	13	10
39	3	3	5	19	4	3	7

Allele file

```
>leuA_1
TACTATCAATGGCATTGGTGAGCGCGCTGGTAACTGCGCGCTGGAAGAACTCACTATGGT
GTTGAAAGTACGCAACGCGTTTTACAACATTGATACTTCAATCCACACATCACGTATCGT
CTCCACCTCCCAGTTACTGCAACGATTGGTTGGCATGCCCGTGCAACGTAACAAGGCAGT
AGTAGGTGCCAATGCCTTTGCACATGAATCGGGTATCCACCAGCACGGTATGCTGCGCCA
TCGCGGCACCTACGAAATCATGCGTCCACAAGAAGTCGGTTGGGTATGTTCGCATATGGT
GCTCGGCCGCCATAGCGGCCGTGCTGCGGTCGAACAGCGTCTACGCGCACTGGGCTACTT
GCTGGAGGAAGAAGATCTAAAACTGGTATTTGAAGAATTCAAACAGCTTTGTGAGAAACA
GCGTTTGGTCACCGATGTCGACCTGCAAGTACTGATGCAAGATACAACAGTACAGCATGG
CTATCGCTTGGCCTCAATGACAATCAGTGATGTTGGTAACCGGGCCAACGCACTGGTGGA
ATTGTCCGATCCCCAAGGTCAGCGTGTTGCCGAAACTGCGCAAGGCAACGGCCCAGTAGA
TGCACTGTTCGGAGCGCTGGCTGCAGCAACTGGGGTCAAACTGGAGTTGGACAGCTATCA
GGTACACAGTGTTGGTATCGGGGCCGATGCACGTGGTGAAGCAAACCT
>leuA_2
TACTATCAATGGCATTGGTGAGCGCGCTGGTAACTGCGCGCTGGAAGAACTCACTATGGT
GTTGAAAGTACGCAACGCGTTTTACAACATTGATACTTCGATCCACACATCACGTATCGT
CTCCACCTCCCAGTTACTGCAACGATTGGTTGGCATGCCCGTGCAACGTAACAAGGCAGT
AGTAGGTGCCAATGCCTTTGCACATGAATCGGGTATCCATCAGCACGGTATGCTGCGCCA
TCGCGGCACCTACGAAATCATGCGTCCACAAGAAGTCGGTTGGGTATGTTCGCACATGGT
ACTCGGCCGCCATAGCGGCCGTGCTGCGGTCGAACAGCGTCTACGCGCACTGGGCTACTT
GCTGGAGGAAGAAGATCTAAAACTGGTATTTGAAGAATTCAAACAGCTTTGTGAGAAACA
GCGTTTGGTCACCGATGTCGACCTGCAAGTACTGATGCAAGATACAACAGTACAGCATGG
CTATCGCTTGGCCTCAATGACAATCAGTGATGTTGGTAACCGGGCCAATGCACTGGTGGA
```

Implementation of MLST using WGS data

Method

- The genome is converted to a blast database

- For the specified species, all alleles for all genes are aligned to the database (using blast for assembled genomes/ using KmerFinder (KMA) for raw reads)

- For each gene, the perfectly matching allele is picked (all nucleotides must match across the whole length of the allele)

- If there is no perfectly matching allele, the closest matching allele is outputtet along with warnings

- The ST is determined from the combination of alleles

DTU

MLST in CGE bitbucket

https://bitbucket.org/genomicepidemiology/mlst/src/master/

Installation

Setting up MLST program

```
# Go to wanted location for mlst
cd /path/to/some/dir
# Clone and enter the mlst directory
git clone https://bitbucket.org/genomicepidemiology/mlst.git
cd mlst
```

Build Docker container

```
# Build container
docker build -t mlst .
# Run test
docker run --rm -it \
       --entrypoint=/test/test.sh mlst
```

#Download and install MLST database

```
# Go to the directory where you want to store the mlst database
cd /path/to/some/dir
# Clone database from git repository (develop branch)
git clone https://bitbucket.org/genomicepidemiology/mlst_db.git
cd mlst_db
MLST_DB=$(pwd)
# Install MLST database with executable kma_index program
python3 INSTALL.py kma_index
```

If kma_index has not bin install please install kma_index from the kma repository: https://bitbucket.org/genomicepidemiology/kma

MLST in CGE bitbucket

https://bitbucket.org/genomicepidemiology/mlst/src/master/

Usage

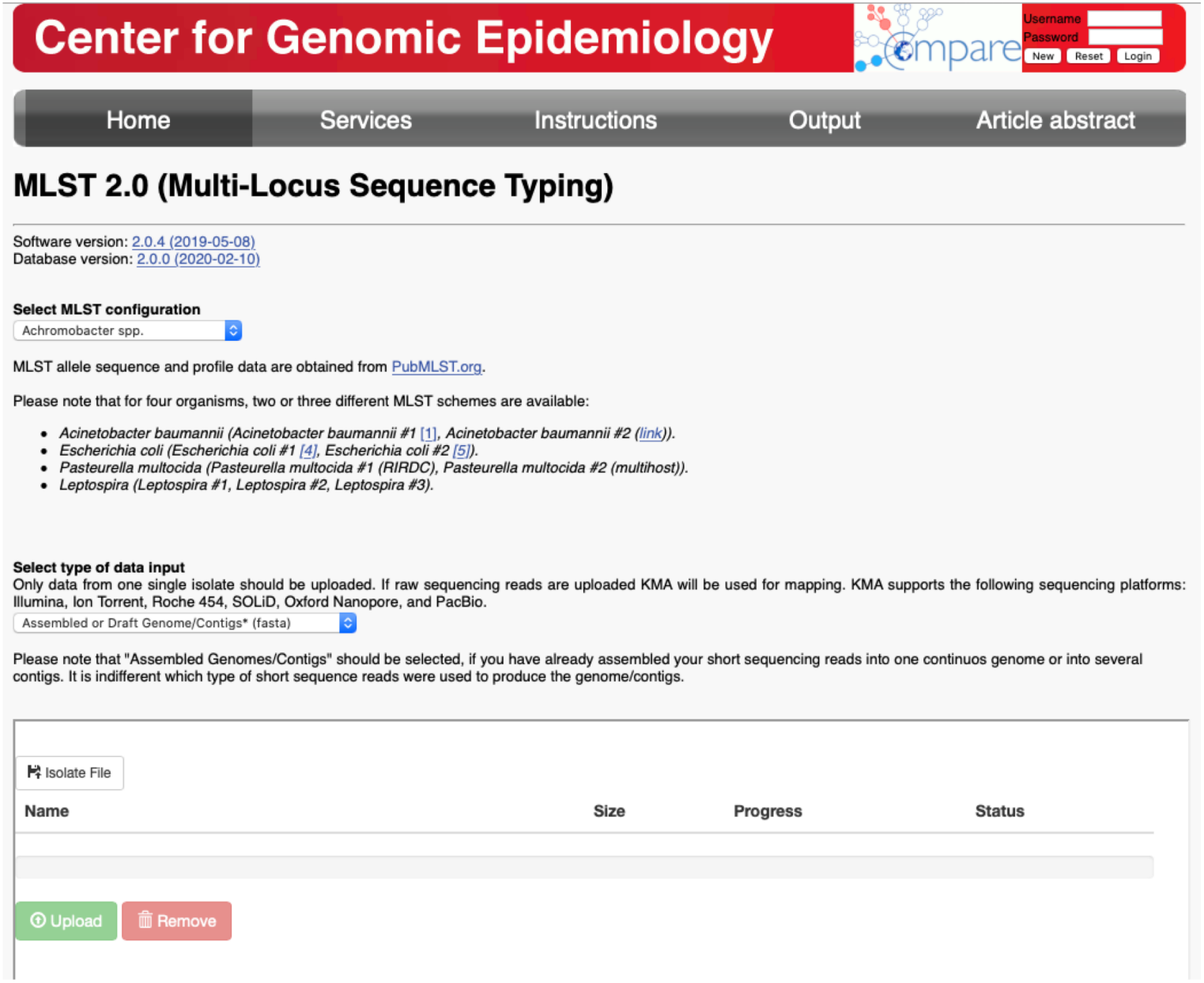
The program can be invoked with the -h option to get help and more information of the service. Run Docker container

```
# Run mlst container
docker run --rm -it \
       -v $MLST_DB:/database \
       -v $(pwd):/workdir \
       mlst -i [INPUTFILE] -o . -s [SPECIES] [-x]
```

When running the docker file you have to mount 2 directory: 1. mlst_db (MLST database) downloaded from bitbucket 2. An output/input folder from where the input file can be reached and an output files can be saved. Here we mount the current working directory (using $pwd) and use this as the output directory, the input file should be reachable from this directory as well.

-i INPUTFILE input file (fasta or fastq) relative to pwd -s SPECIES species origin of input file -o OUTDIR outpur directory relative to pwd -x extended output. Will create an extented output

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



Center for Genomic Epidemiology

Home | Services | Instructions | Output

MLST-1.7 Server - Typing Results

Sequence Type: *Unknown ST*

Locus	% Identity	HSP Length	Allele Length	Gaps	Allele
acs	99.74	390	390	0	*acs_28*
aro	100.00	349	498	0	*aro_122*
gua	100.00	373	373	0	*gua_11*
mut	100.00	442	442	0	*mut_11*
nuo	100.00	366	366	0	*nuo_4*
pps	100.00	370	370	0	*pps_12*
trp	100.00	443	443	0	*trp_3*

Please note that one or more loci do not match perfectly to any previously registered MLST allele.
We recommend verifying the results by traditional methods for MLST!

extended output

MLST Profile: *paeruginosa*

Organism: *Pseudomonas aeruginosa*

Input Files: *18_tag18.contigs.fa*

Extended output
Imperfect match, *acs* loci

acs: WARNING, Identity: 99.74%, HSP/Length: 390/390, Gaps: 0, Best Match: *acs_28*

```
MLST allele seq:    ggcccgttggccaacggcgccaccaccattctgttcgagggcgtgccgaactaccccgac
Hit in genome:      ggcccgttggccaacggcgccaccaccattctgttcgagggcgtgccgaactaccccgac

MLST allele seq:    gtgacccgcgtggcgaaaatcatcgacaagcacaaggtcaacatcctctacaccgcgccg
Hit in genome:      gtgacccgcgtggcgaaaatcatcgacaagcacaaggtcaacatcctctacaccgcgccg

MLST allele seq:    accgcgatccgcgcgatgatggccgaaggcaaggcggcggtggccggtgccgacggttcc
Hit in genome:      accgcgatccgcgcgatgatggccgaaggcaaggcggcggtggccggtgccgacggttcc

MLST allele seq:    agcctgcgtctgctcggttcggtgggcgagccgatcaacccggaagcctggcagtggtac
Hit in genome:      agcctgcgtctgctcggttctgtgggcgagccgatcaacccggaagcctggcagtggtac

MLST allele seq:    tacgagaccgtcggccagtcgcgctgcccgatcgtcgacacctggtggcagaccgagacc
Hit in genome:      tacgagaccgtcggccagtcgcgctgcccgatcgtcgacacctggtggcagaccgagacc

MLST allele seq:    ggcgcctgcctgatgaccccgctgccgggcgcccacgcgatgaagccgggctctgcagcc
Hit in genome:      ggcgcctgcctgatgaccccgctgccgggcgcccacgcgatgaagccgggctctgcagcc

MLST allele seq:    aagccgttcttcggcgtggtaccggcactg
Hit in genome:      aagccgttcttcggcgtggtaccggcactg
```

Extended output
Imperfect match, *aro* loci

aro*: WARNING, Identity: 100%, HSP/Length: 349/498, Gaps: 0, *aro_122* is the best match for *aro

```
MLST allele seq: atgtcaccgtgccgttcaaggaagaggcctatcgtctggtggacgaattgagcgagcggg
Hit in genome:

MLST allele seq: ccacccgggccggggcggtgaacaccctgatccgcctcgccgacggtcgcctgcgcggcg
Hit in genome:

MLST allele seq: acaacaccgacggcgcgggcttgctgcgggacctgacggcgaacgccggggtcgagctgc
Hit in genome:                                gacctgacggcgaacgccggggtcgagctgc

MLST allele seq: gcggcaagcgggttctcctgctcggcgccggcggtgcggtgcgtggggtgctcgaaccct
Hit in genome:   gcggcaagcgggttctcctgctcggcgccggcggtgcggtgcgtggggtgctcgaaccct

MLST allele seq: tcctcggcgagtgcccggcggagttgctgatcgccaaccgcacggcgcggaaggccgtgg
Hit in genome:   tcctcggcgagtgcccggcggagttgctgatcgccaaccgcacggcgcggaaggccgtgg

MLST allele seq: acctggccgagcggttcgccgacctcggcgcggtgcacggctgcggtttcgccgaggtcg
Hit in genome:   acctggccgagcggttcgccgacctcggcgcggtgcacggctgcggtttcgccgaggtcg

MLST allele seq: aagggcctttcgacctgatcgtcaacggcacctcggccagtcttgccggcgacgtgccgc
Hit in genome:   aagggcctttcgacctgatcgtcaacggcacctcggccagtcttgccggcgacgtgccgc

MLST allele seq: cgctggcgcagagcgtgatcgagcccggccgtaccgtctgctacgacatgatgtatgcca
Hit in genome:   cgctggcgcagagcgtgatcgagcccggccgtaccgtctgctacgacatgatgtatgcca

MLST allele seq: aggaaccgactgccttca
Hit in genome:   aggaaccgactgccttca
```

https://pubmlst.org

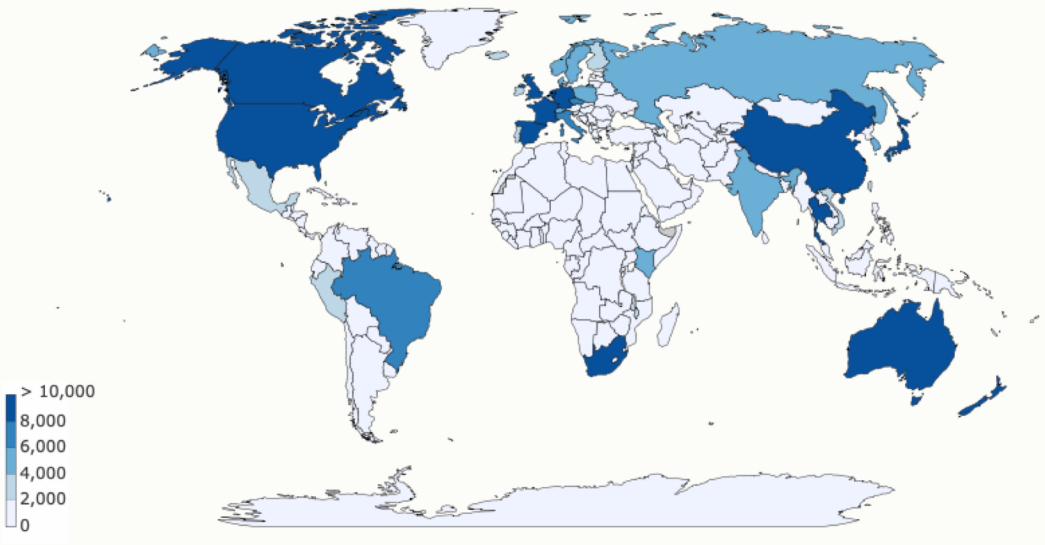
PubMLST Databases Downloads BIGSdb Contact Account

PubMLST

Welcome to PubMLST - Public databases for molecular typing and microbial genome diversity.

- Log in or register for a PubMLST account
- Databases
- Species identification by ribosomal MLST

Source of isolates submitted to PubMLST



More details

isolates: 740,594
genomes: 510,111
alleles: 20,933,675

This site is hosted at The Department of Zoology, University of Oxford, UK and is funded by The Wellcome Trust.

Is super MLST, e.g. rMLST, the next step?

Microbiology. 2012 Jan 27. [Epub ahead of print]

Ribosomal Multi-Locus Sequence Typing: universal characterisation of bacteria from domain to strain.

Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony CM, Colles FM, Wimalarathna HM, Harrison OB, Sheppard SK, Cody AJ, Maiden MC.

Abstract

No single genealogical reconstruction or typing method currently encompasses all levels of bacterial diversity, from domain to strain. We propose **Ribosomal Multi Locus Sequence Typing (rMLST)**, an approach which **indexes variation of the 53 genes encoding the bacterial ribosome protein subunits (rps genes)**, as a means of integrating microbial genealogy and typing. As with MLST, rMLST employs curated reference sequences to identify gene variants efficiently and rapidly. **The rps loci are ideal targets for a universal characterization scheme as they are: (i) present in all bacteria**; (ii) distributed around the chromosome; and (iii) encode proteins which are under stabilising selection for functional conservation. Collectively, the rps loci exhibit variation that resolves bacteria in to groups at all taxonomic and most typing levels providing significantly more resolution than 16S small subunit rRNA gene phylogenies. A web-accessible expandable database, comprising whole genome data from more than 1900 bacterial isolates, including 28 draft genomes assembled de novo from the EBI

Ribosomal Multilocus Sequence Typing (rMLST)

Ribosomal Multilocus Sequence Typing (rMLST) is an approach that indexes variation of the 53 genes encoding the bacterial ribosome protein subunits (*rps* genes) as a means of integrating microbial taxonomy and typing.

The *rps* gene variation catalogued in this database permits rapid and computationally non-intensive identification of the phylogenetic position of any bacterial sequence at the domain, phylum, class, order, family, genus, species and strain levels.

rMLST is described in Jolley *et al.* 2012 *Microbiology* **158:**1005-15.

Identify species

[No login required]

genomes: 343,524
alleles: 2,145,412
profiles: 133,582

- Databases
  - rMLST sequence definitions
  - Bacterial genomes

C. jejuni/C. coli core genome MLST (cgMLST) scheme

A C. jejuni / C. coli core genome MLST (cgMLST) scheme v1.0 has been defined from 2,742 Oxfordshire Human Surveillance genomes, for use in the analysis of human disease isolates.

Loci appearing in 95% or more of genomes were detected, using the Genome Comparator function of BIGSdb, from the 1,643 loci defined by the re-annotation of the NCTC11168 genome (Gundogdu et al., 2007). Potential paralogues were identified in 5 subsets of 10 isolates by running Genome Comparator and excluding paralogous loci from distance matrix calculations; firstly, when loci were paralogous in all isolates (default BLAST settings) and secondly when loci were paralogous in any isolates (min 90% sequence similarity BLAST setting). Sequence similarities to loci thus detected were identified in further loci using GeneDB. This resulted in the identification of 22 potential paralogous loci which were removed from the original 95% core list to provide a cgMLST scheme of 1,343 loci.

This scheme is available for selection within the sequence definition and isolate databases. It can also be accessed via the RESTful API.

Resistant gene detection: ResFinder

Source	Accessibility	Approach	Type	Year	Status
Resfinder	Web and standalone	Assembly-based and read based tool	Tool and database	2012	Active update regularly
ARG-ANNOT	Standalone	Assembly-based tools	Tool and database	2014	Archived last update in May 2018
ResfinderFG	Web	Assembly-based tool	Tool and database	2016	Active last update in November 2016
RGI	Web and/or standalone	Assembly-based tools	Tool	2015	Active
ARGs-OAP (v2)	Web and/or standalone	Assembly-based tools	Tool	2016	Active
ARIBA	Standalone	Assembly-based tools	Tool	2017	Active
NCBI-AMRFinder	Standalone	Assembly-based tools	Tool	2018	Active
SRST2	Standalone	Read-based tools	Tool	2014	Active
SEAR	Web and/or standalone (archived)	Read-based tools	Tool	2015	Archived
ShortBRED	Standalone	Read-based tools	Tool	2015	Active
PATRIC	Web	Read-based tools	Tool	2016	Active
SSTAR	Standalone	Read-based tools	Tool	2016	Active
KmerResistance	Web	Read-based tools	Tool	2016	Active
GROOT	Standalone	Read-based tools	Tool	2018	Active
DeepArgs	Web	Read-based tools	Tool	2018	Active

Boolchandani M, D'Souza AW, Dantas G. Nat Rev Genet. 2019

Source	Accessibility	Type	Year	Status
CARD	Web	Database	2013	Active updated monthly
Resfams	Web	Database	2015	Active last update in January 2015
ARDB	Web	Database	2009	Archived last updated in 2009
MEGARes	Web	Database	2016	Active; last update in December 2016
NDARO	Web	Database	2016	Active; started in 2016
Mustard	Web	Database	2018	Active; last update in November 2018
FARME database	Web	Database	2017	Active; last update in 2017
SARG (v2)	Web	Database		Active
Lahey list of β-lactamases	Web	Database	2015	Archived; last update in 2015
BLDB	Web	Database	2018	Active; last update in November 2018
LacED	Web	Database		TEMLacED active: last update in 2017; SHVED archived: last update in April 2010
CBMAR	Web	Database		Last update in September 2014
MUBII-TB-DB	Web	Database		Last update in December 2013
u-CARE	Web	Database		Last update in 2016

Boolchandani M, D'Souza AW, Dantas G. Nat Rev Genet. 2019

ResFinder

- ResFinder is based on curated database, public databases as well as on scientific papers
- The ResFinder is a web-friendly interface and freely accessible tool (It is also a stand-alone tool)
- ResFinder will detect the presence of whole resistance genes, AND chromosomal point mutations causing resistance in the whole genome sequence data (raw reads or assembled genomes)
- ResFinder 4.0 provides in silico antibiograms as reliable as those obtained by phenotypic antimicrobial susceptibility testing
- High concordance (>95%) between phenotypic and predicted antimicrobial susceptibility was observed. Discrepancies were mainly linked to criteria for interpretation of phenotypic tests and suboptimal sequence quality, and not to ResFinder 4.0 performance.

- Bortolaia V. et al. ResFinder 4.0 for predictions of phenotypes from genotypes. J. Antimicrob Chemother. 2020 Aug 11;dkaa345

Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agersø Y, Lund O, Larsen MV,. Aarestrup FM. Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. J. Antimicrob Chemother. 2013 Apr;68(4):771-7.

ResFinder command line in CGE bitbucket

https://bitbucket.org/genomicepidemiology/resfinder/src/master/

ResFinder tool

Setting up ResFinder script and database

```
# Go to wanted location for resfinder
cd /path/to/some/dir

# Clone the latest version and enter the resfinder directory
git clone https://git@bitbucket.org/genomicepidemiology/resfinder.git
cd resfinder
```

Dependencies:

Depending on how you plan to run ResFinder BLAST and KMA can be optional. BLAST is used to analyse assemblies (ie. FASTA files). KMA is used to analyse read data (ie. FASTQ files).

Python modules: Tabulate, BioPython, CGECore and Python-Git

To install the needed python modules you can use pip

```
pip3 install tabulate biopython cgecore gitpython python-dateutil
```

For more information visit the respective website

```
https://bitbucket.org/astanin/python-tabulate
https://biopython.org
https://bitbucket.org/genomicepidemiology/cge_core_module
https://gitpython.readthedocs.io/en/stable/index.html
```

BLAST (optional)

If you don't want to specify the path of blastn every time you run ResFinder, make sure that blastn is in you PATH.

Blastn can be obtained from:

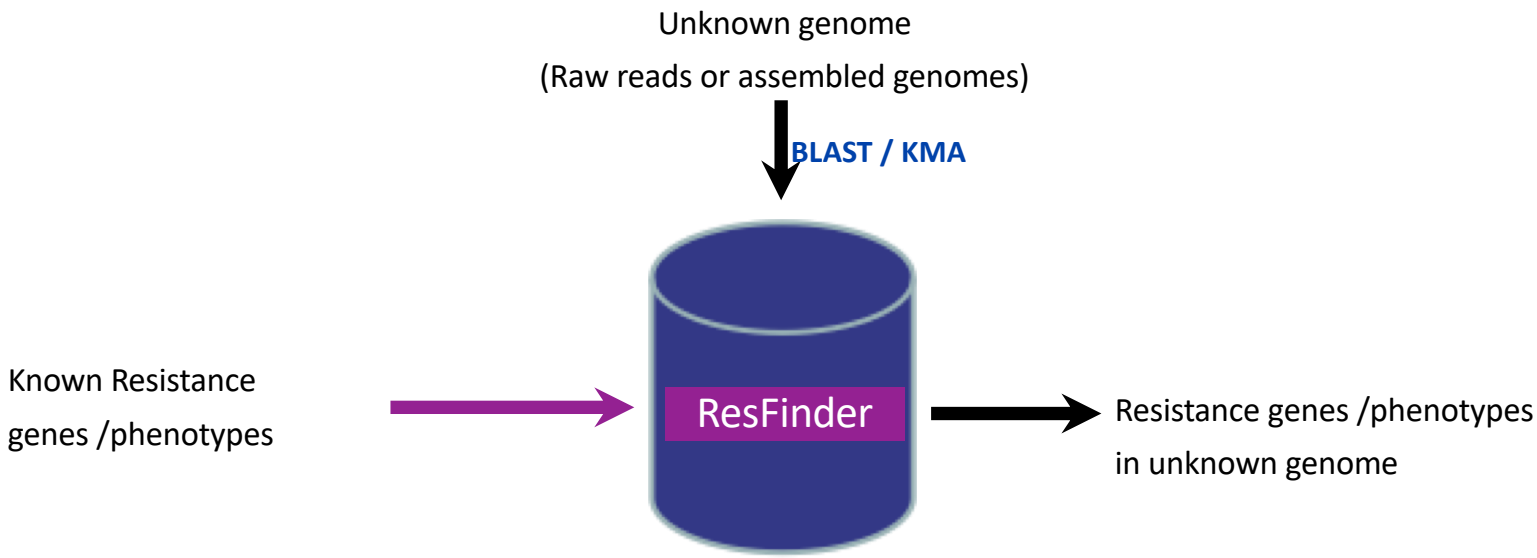
```
ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
```

ResFinder 4.0 database

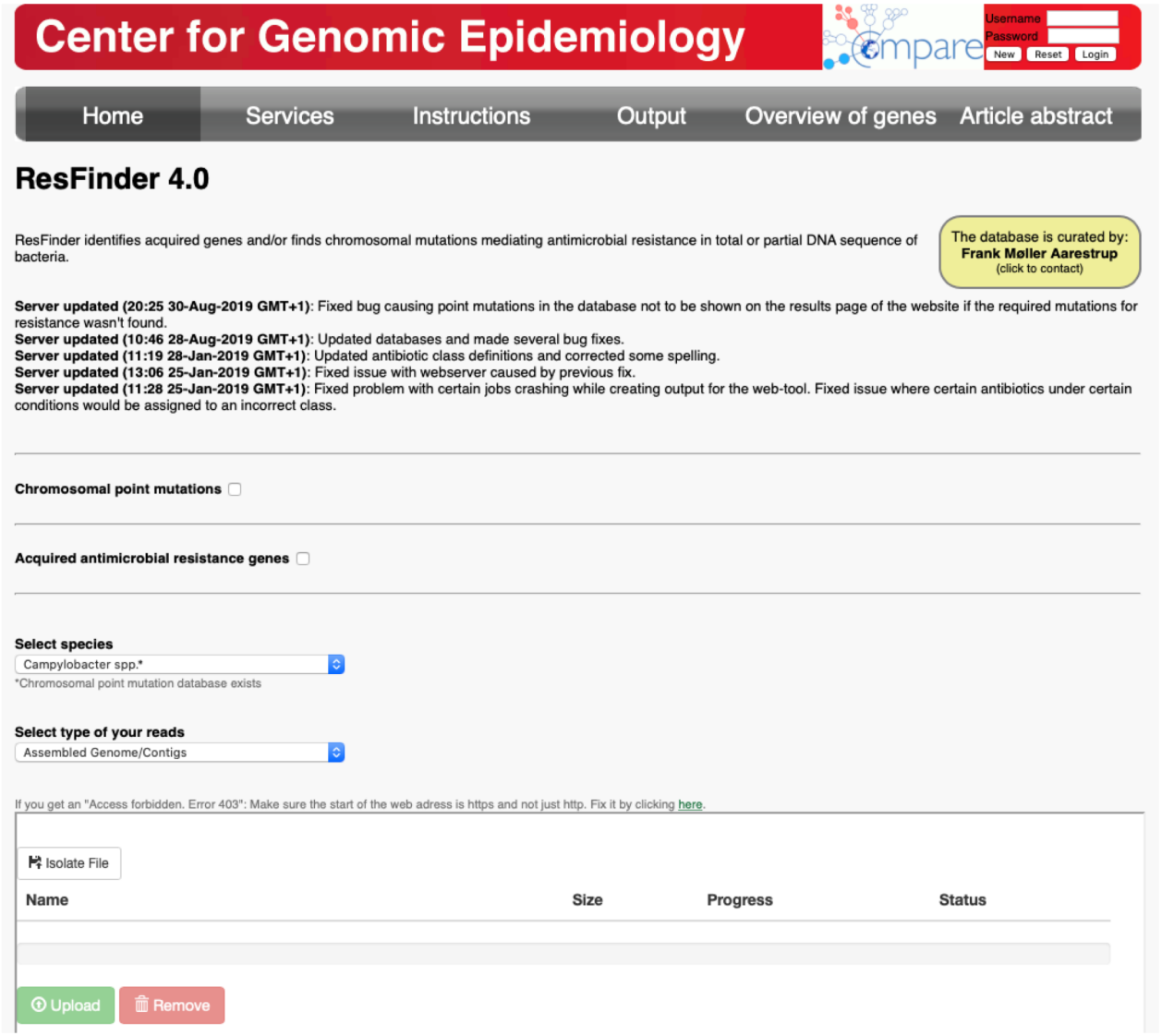
ResFinder 4.0 contains 4 databases

- AMR genes (ResFinder) **https://bitbucket.org/genomicepidemiology/resfinder_db/src/master/**
- Chromosomal point mutation (PointFinder) **https://bitbucket.org/genomicepidemiology/pointfinder_db/src/master/**
- Translation of genotypes into phenotypes
- Species-specific panels for in silicon antibiograms

ResFinder tool



http://cge.cbs.dtu.dk/services/ResFinder/



ResFinder-4.0 Server - Results

Input Files: *strain_1.fasta*

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	Resistant	blaCTX-M-15 (blaCTX-M-15_AY044436)
chloramphenicol	phenicol	No resistance	
piperacillin+tazobactam	beta-lactam	No resistance	
cefoxitin	beta-lactam	No resistance	
ampicillin	beta-lactam	Resistant	blaCTX-M-15 (blaCTX-M-15_AY044436), blaTEM-1B (blaTEM-1B_AY458016)
ampicillin+clavulanic acid	beta-lactam	No resistance	
cefotaxime	beta-lactam	Resistant	blaCTX-M-15 (blaCTX-M-15_AY044436)
ciprofloxacin	fluoroquinolone	Resistant	gyrA (p.S83A)
colistin	polymyxin	No resistance	
sulfamethoxazole	folate pathway antagonist	Resistant	sul2 (sul2_HQ840942), sul1 (sul1_AY115475), sul1 (sul1_AY522923), sul1 (sul1_DQ914960), sul1 (sul1_U12338)
imipenem	beta-lactam	No resistance	
trimethoprim	folate pathway antagonist	Resistant	dfrA7 (dfrA7_AB161450)
nalidixic acid	fluoroquinolone	Resistant	gyrA (p.S83A)
ertapenem	beta-lactam	No resistance	
tetracycline	tetracycline	Resistant	tet(A) (tet(A)_AJ517790)
fosfomycin	fosfomycin	No resistance	
ceftazidime	beta-lactam	Resistant	blaCTX-M-15 (blaCTX-M-15_AY044436)
temocillin	beta-lactam	No resistance	
gentamicin	aminoglycoside	No resistance	
meropenem	beta-lactam	No resistance	
azithromycin	macrolide	No resistance	

Download phenotype table (txt) | Download species specific phenotype table (txt)

Fluoroquinolone					
Mutation	Nucleotide change	Amino acid change	Phenotype	PMID	Notes
gyrA:p.S83A	tcg -> gcg	s -> a	nalidixic acid,ciprofloxacin	8891148, 2168148, 12654733, 12654733	

Aminoglycoside									
Resistance gene	Identity	Alignment Length/Gene Length	Position in reference	Contig or Depth	Position in contig	Phenotype	PMID	Accession no.	Notes
aph(6)-Id	100.0	837/837	1..837	strain_1_contig_11	4362..5198	streptomycin	2653965	M28829	Alternative name strB
aph(3")-Ib	100.0	804/804	1..804	strain_1_contig_11	3559..4362	streptomycin	12029529	AF321551	Alternative name strA

Tetracycline									
Resistance gene	Identity	Alignment Length/Gene Length	Position in reference	Contig or Depth	Position in contig	Phenotype	PMID	Accession no.	Notes
tet(A)	100.0	1200/1200	1..1200	strain_1_contig_11	15402..16601	doxycycline,tetracycline	12654659	AJ517790	

Beta-lactam									
Resistance gene	Identity	Alignment Length/Gene Length	Position in reference	Contig or Depth	Position in contig	Phenotype	PMID	Accession no.	Notes
blaTEM-1B	100.0	861/861	1..861	strain_1_contig_14	84807..85667	amoxicillin,ampicillin,cephalothin,piperacillin,ticarcillin	15388431	AY458016	Class A
blaCTX-M-15	100.0	876/876	1..876	strain_1_contig_14	81110..81985	amoxicillin,ampicillin,aztreonam,cefepime,cefotaxime,ceftazidime,ceftriaxone,piperacillin,ticarcillin	11470367, 26169409	AY044436	Class A

Folate pathway antagonist									
Resistance gene	Identity	Alignment Length/Gene Length	Position in reference	Contig or Depth	Position in contig	Phenotype	PMID	Accession no.	Notes
sul1	100.0	761/840	1..761	strain_1_contig_9	442478..443238	sulfamethoxazole	unpublished	U12338	

Output

The grey colour

indicates a warning due to a non-perfect match, alignment length is shorter than resistance gene length, % ID <= 100%.

The light green colour

indicates a warning due to a non-perfect match, % ID < 100%, alignment length = resistance gene length.

PlasmidFinder

Background

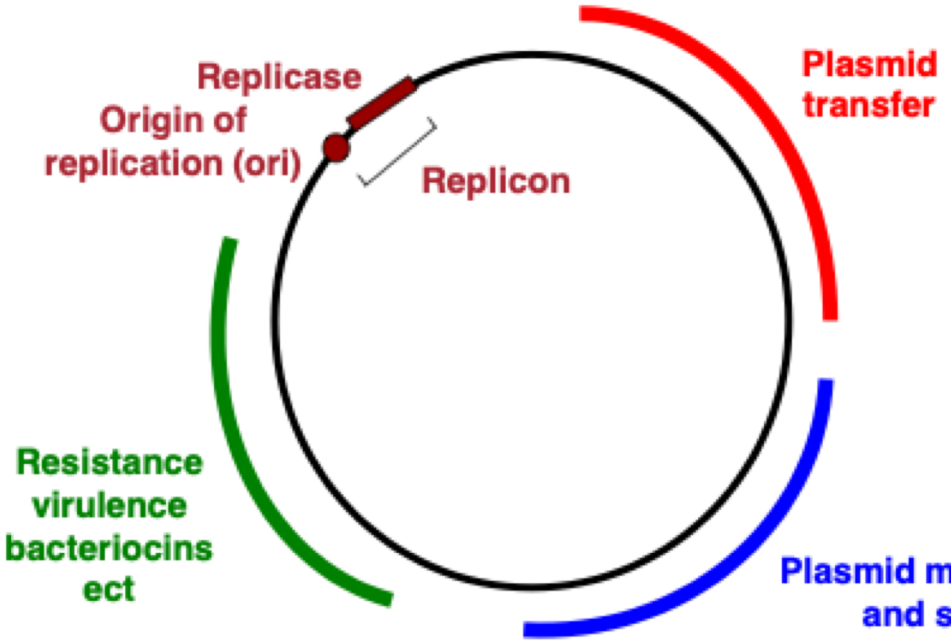
- Plasmids are double-stranded circular or linear DNA molecules. They can replicate and transfer between different bacterial species and clones

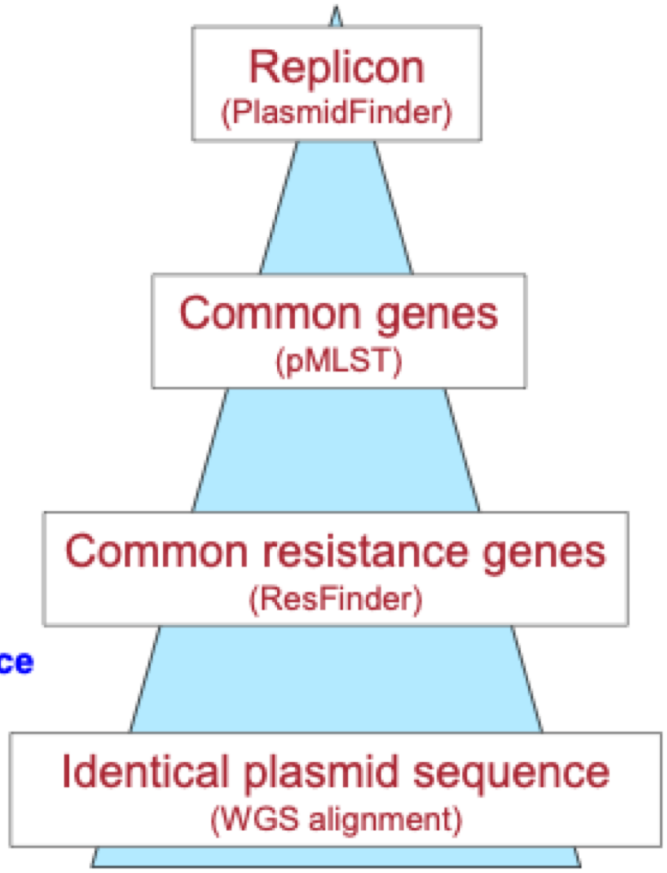
- Most of the known plasmids have been identified because they confer phenotypes that are subject to positive selection on bacterial host such as the presence of antimicrobial resistance or virulence genes

- It is important not only to study the molecular epidemiology of different bacterial clones but also to study and understand the molecular epidemiology of transferable plasmids

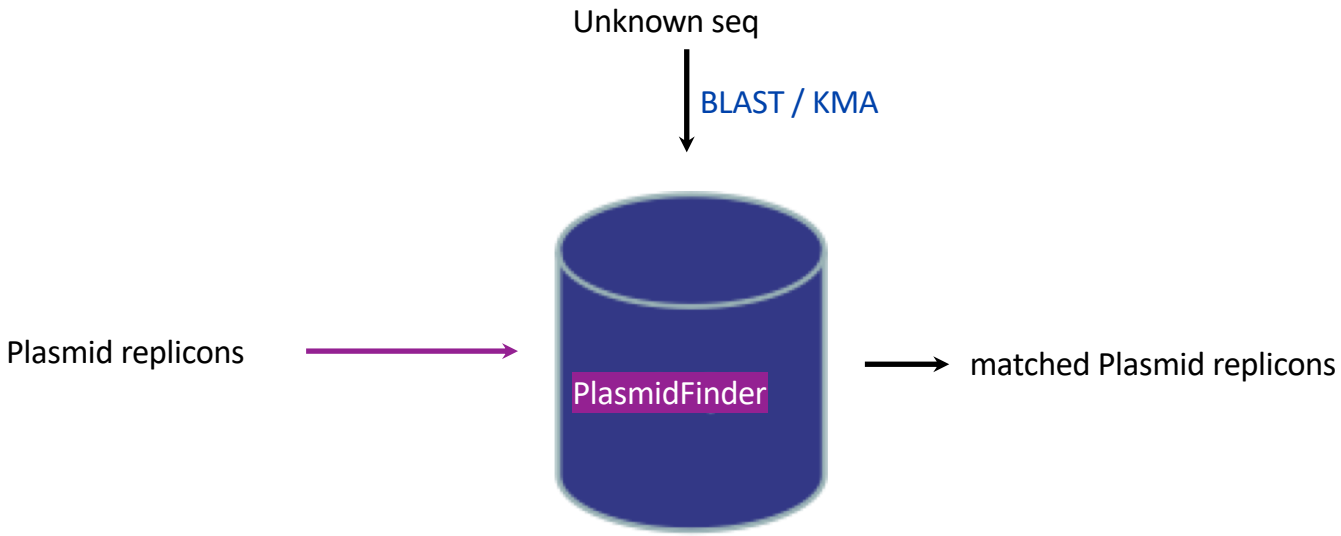
- For this specific purpose, plasmid typing systems are needed

Hieratical typing strategy

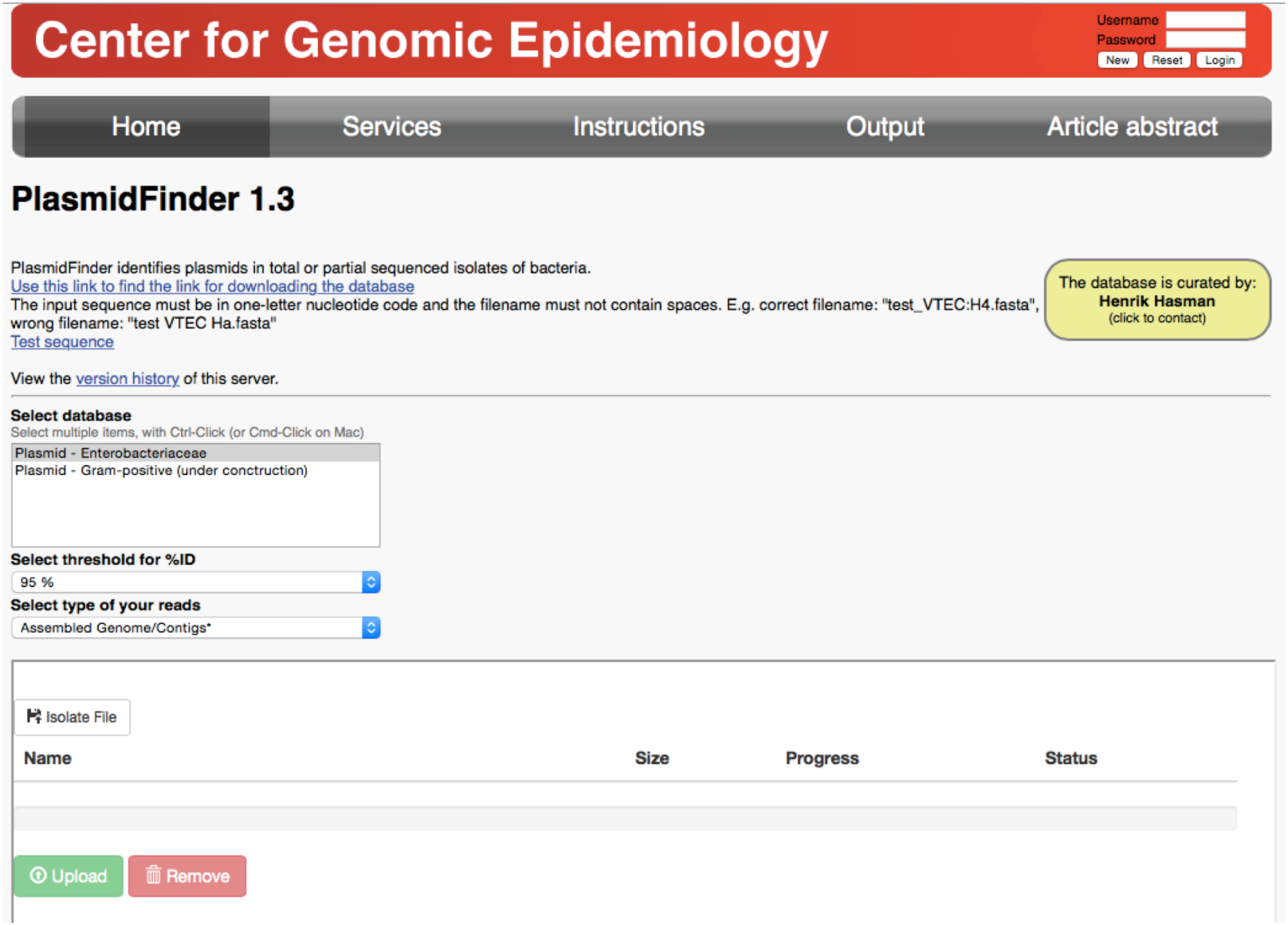




PlasmidFinder



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



Center for Genomic Epidemiology

Home | Services | Instructions | Output

PlasmidFinder-1.3 Server - Results

PlasmidFinder Results

SETTINGS:
Selected %ID threshold: *95.00*

PlasmidFinder - Enterobacteriaceae						
Plasmid	**%Identity**	**Query/HSP length**	**Contig**	**Position in contig**	**Note**	**Accession number**
IncP	99.44	535 / 534	strain_1_contig_11	13583..14117	alpha	L27758
IncFII(pRSB107)	97.70	261 / 261	strain_1_contig_18	23064..23324	pRSB107	AJ851089
IncFIB(AP001918)	96.92	682 / 682	strain_1_contig_18	27994..28675		AP001918
IncI1	97.89	142 / 142	strain_1_contig_14	72210..72351	Alpha	AP005147
IncQ1	100.00	450 / 450	strain_1_contig_11	895..1344		HE654726

extended output

Results as text | Results tab separated | Hit in genome sequences | Plasmid sequences

Input Files: *strain_1.fasta.txt*

VirulenceFinder

Bacterial pathogenecity and virulence

- **Pathogenicity.** This is the potential capacity of certain species of microbes to cause an infectious process.
- **Virulence.** signifies the degree of pathogenicity of the given strain. Virulence, therefore, is an index of the qualitative individual nature of the pathogenic microorganism.

E. coli is a **pathogenic** bacterium. However, not all *E. coli* are equally **virulent**. Some have specific virulence factors, which make them more **virulent** than others.

Credited: Henrik Hasman

VirulenceFinder in CGE bitbucket

https://bitbucket.org/genomicepidemiology/virulencefinder/src

Installation

Setting up VirulenceFinder program

Warning: Due to bugs in BioPython 1.74, if you are not using the Docker container, do not use that version if not using Python 3.7.

```
# Go to wanted location for virulencefinder
cd /path/to/some/dir
# Clone and enter the virulencefinder directory
git clone https://bitbucket.org/genomicepidemiology/virulencefinder.git
cd virulencefinder
```

Build Docker container

```
# Build container
docker build -t virulencefinder .
# Run test
docker run --rm -it \
       --entrypoint=/test/test.sh virulencefinder
```

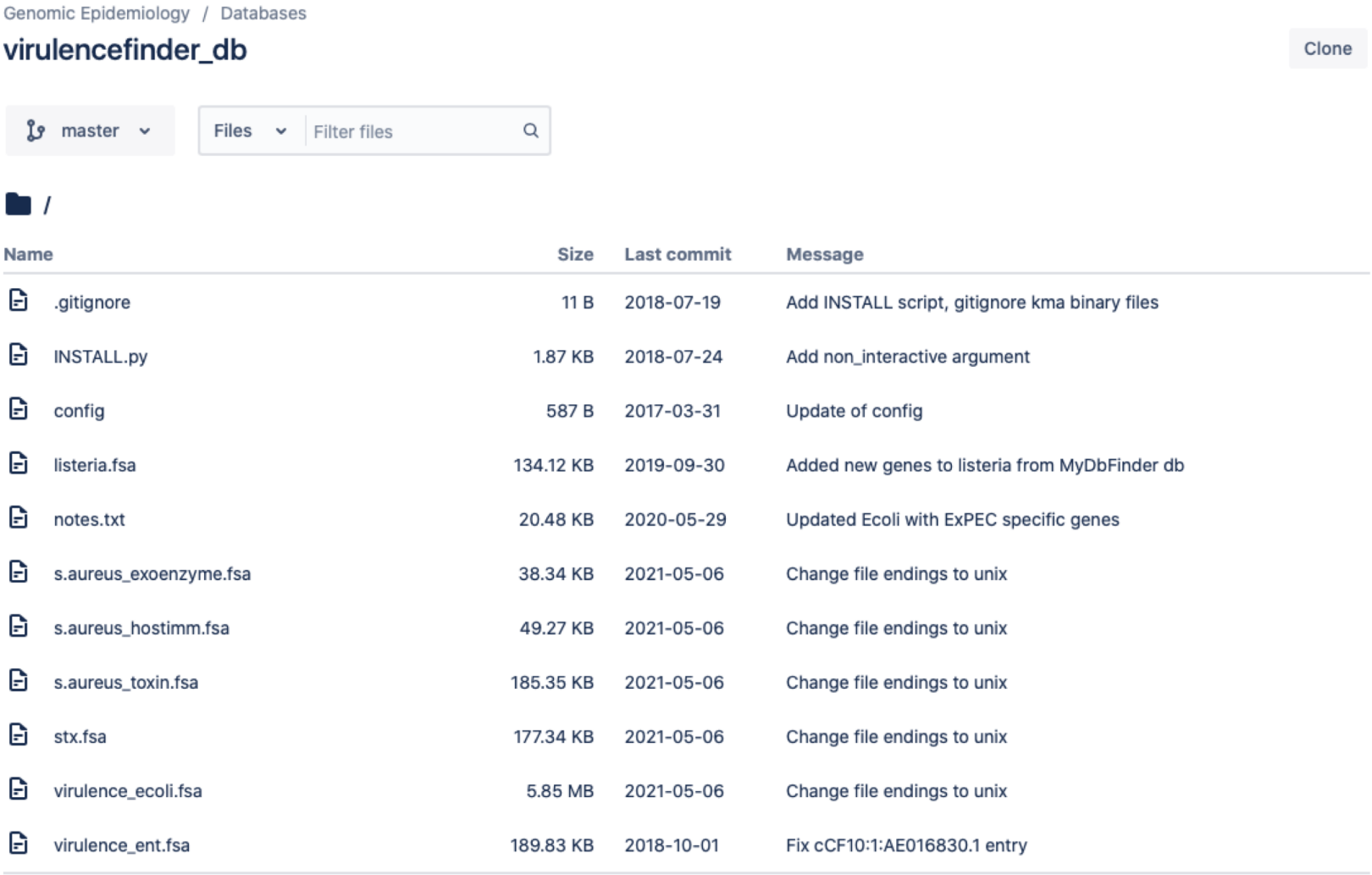
#Download and install VirulenceFinder database

```
# Go to the directory where you want to store the virulencefinder database
cd /path/to/some/dir
# Clone database from git repository (develop branch)
git clone https://bitbucket.org/genomicepidemiology/virulencefinder_db.git
cd virulencefinder_db
VIRULENCE_DB=$(pwd)
# Install VirulenceFinder database with executable kma_index program
python3 INSTALL.py kma_index
```

If kma_index has no bin install please install kma_index from the kma repository: https://bitbucket.org/genomicepidemiology/kma

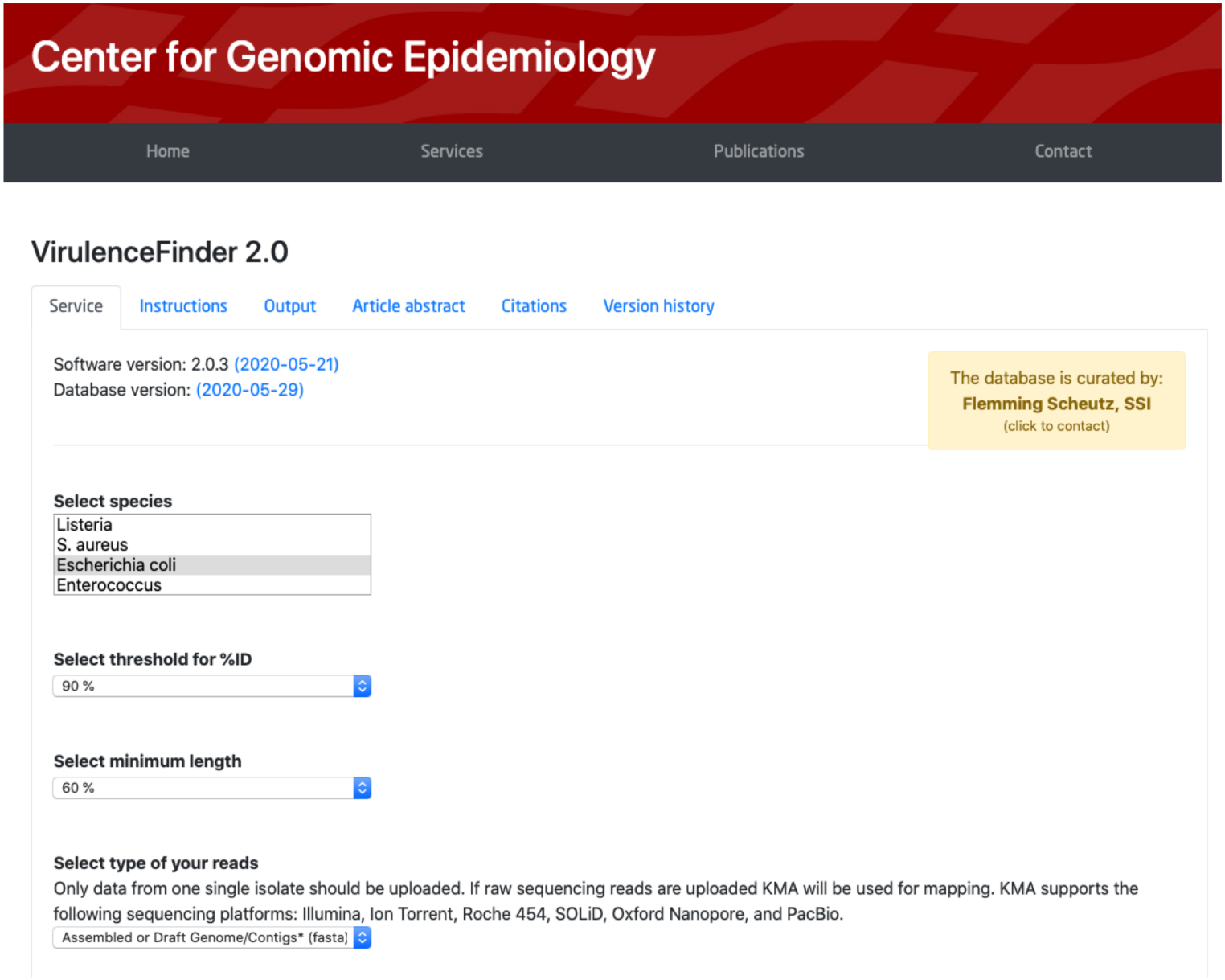
VirulenceFinder database

https://bitbucket.org/genomicepidemiology/virulencefinder_db/src/master/



Name	Size	Last commit	Message
.gitignore	11 B	2018-07-19	Add INSTALL script, gitignore kma binary files
INSTALL.py	1.87 KB	2018-07-24	Add non_interactive argument
config	587 B	2017-03-31	Update of config
listeria.fsa	134.12 KB	2019-09-30	Added new genes to listeria from MyDbFinder db
notes.txt	20.48 KB	2020-05-29	Updated Ecoli with ExPEC specific genes
s.aureus_exoenzyme.fsa	38.34 KB	2021-05-06	Change file endings to unix
s.aureus_hostimm.fsa	49.27 KB	2021-05-06	Change file endings to unix
s.aureus_toxin.fsa	185.35 KB	2021-05-06	Change file endings to unix
stx.fsa	177.34 KB	2021-05-06	Change file endings to unix
virulence_ecoli.fsa	5.85 MB	2021-05-06	Change file endings to unix
virulence_ent.fsa	189.83 KB	2018-10-01	Fix cCF10:1:AE016830.1 entry

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



VirulenceFinder-1.2 Server - Results

SETTINGS:
Selected %ID threshold: *98.00*

Virulence - E. coli						
Virulence factor	**%Identity**	**Query/HSP length**	**Contig**	**Position in contig**	**Protein function**	**Accession number**
mcmA	99.64	279 / 279	NODE_17_length_48340_cov_62.616714	40909..41187	Microcin M part of colicin H	AJ515251
lpfA	100.00	573 / 573	NODE_4_length_115337_cov_62.053581	84857..85429	Long polar fimbriae	KC207123
iss	99.71	342 / 342	NODE_195_length_89121_cov_54.610832	87701..88042	Increased serum survival	CU928160
prfB	100.00	882 / 882	NODE_75_length_157387_cov_57.585850	94324..95205	P-related fimbriae regulatory gene	CP002970

extended output

Results as text | Results tab separated | Hit in genome sequences | Virulence gene sequences

Input Files: *EC19_2011_70_34_3-illumina_pe_velvet1.1.04_kmer63_cov57_cut0.fna*

- **Non-pathogenic:** Few or no obvious virulence factors (you already have information regarding which this could be).

- **UPEC:** Adhesion factors needed to avoid being flushed away from the urinary tract. Also siderophore proteins such as the one encoded by the iroN gene for iron chelation in urine can be relevant due to the iron-limiting environment in the bladder. Presence Pap (P) fimbriae (papG adhesion) can be a sign of increased virulence, as these fimbriae are associated with progression of a urinary tract infection into pyelonephritis (Dan: Nyrebetændelse).

- **VTEC, STEC (EHEC):** A hall-mark virulence factor is the Shiga toxin. This is encoded by the stx2A and the stx2B genes.

- **EAEC:** Many different virulence factors (especially aggregative adherence fimbriae (AAFs) located on a 100-kb pAA plasmid, mycolycins such as those encoded by the pic gene and toxins such as those encoded by the pet and the astA gene) are believed to be responsible for the EAEC phenotype. However, it has recently been found that the regulator encoded by the aggR gene also located on the pAA plasmid is coordinating the virulence factors. Therefore, detection of the aggR gene is a good marker for EAEC.

- **ETEC:** This pathotype is known for its production of Heat-Stabile Toxin (ST) or Heat-Labile Toxin (LT). The former can be encoded by the sta1 or stb genes and the latter by the elt or ltcA genes.

EPEC	Enteropathogenic E.coli	pEAF	ehlY					
ETEC	Enterotoxigenic E.coli	elt	ltcA	sta1	stb	cfa	elt	est
EAEC	Enteroaggregative E.coli	aaf	aggR	aaiC	aatA			
VTEC, STEC (EHEC)	Shiga toxin-producing E.coli	stx2A	stx2B					
EIEC	Enteroinvasive E.coli	pINV						

SalmonellaTypeFinder

Accept only FASTQ

SalmonellaTypeFinder in CGE bitbucket

https://bitbucket.org/genomicepidemiology/salmonellatypefinder/src/master/

Installation using Docker

Build time: Docker image takes approximately 1 hour to build.

Setting up SalmonellaTypeFinder

```
# Go to wanted location for SalmonellaTypeFinder
cd /path/to/some/dir
# Clone and enter the salmonellatypefinder directory
git clone https://bitbucket.org/genomicepidemiology/salmonellatypefinder.git
cd salmonellatypefinder
```

The installation can either be with MLST typing capability or without. Choose without, if you intend to provide SalmonellaTypeFinder with an MLST type each time you run it, or if you have the CGE MLST tool installed, and intend to provide the path to it when invoking SalmonellaTypeFinder. Choose with, if you sometimes or always need SalmonellaTypeFinder to find the MLST type.

Build Docker container without MLST typing capability

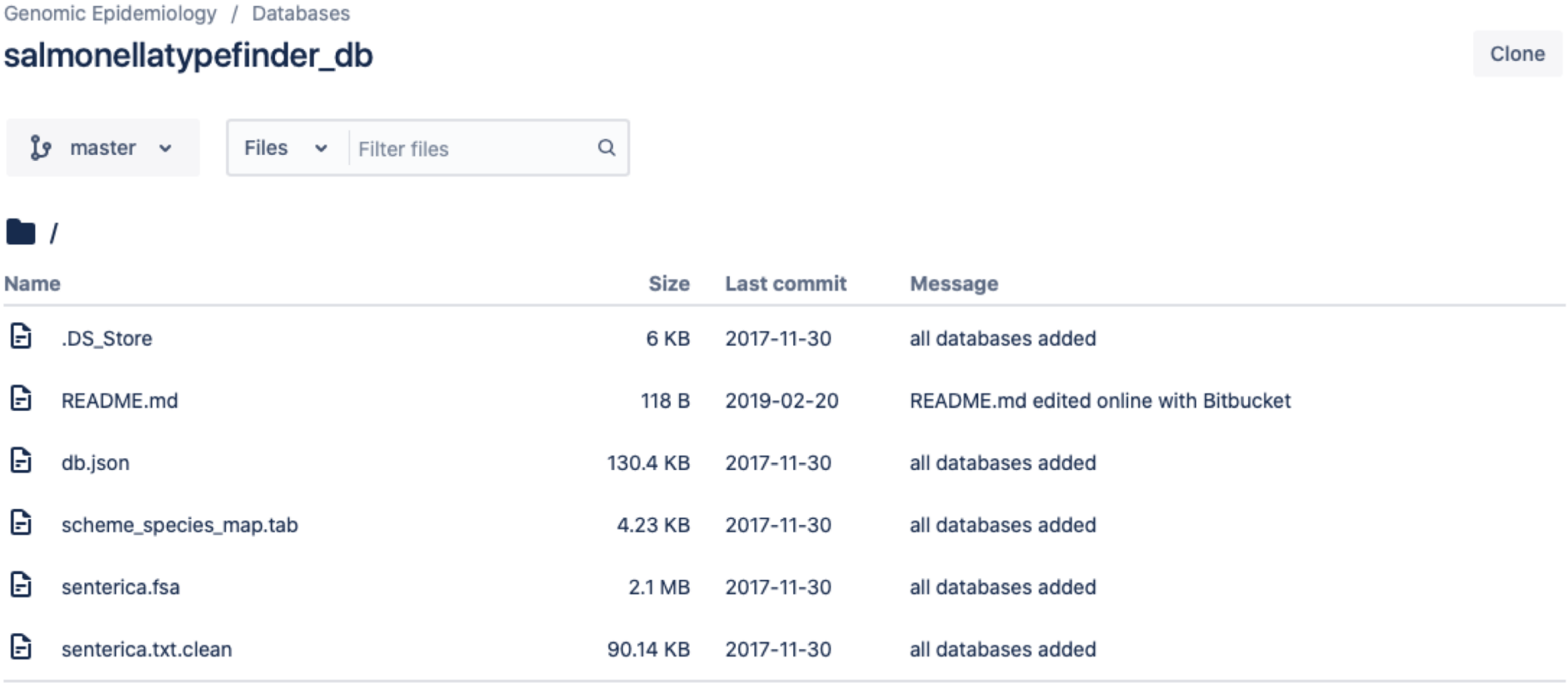
```
# Rename Docker files
mv Dockerfile Dockerfile_complete
mv Dockerfile_light Dockerfile
# Build container
docker build -t salmonellatypefinder .
```

Build Docker container with MLST typing capability Dependencies: KMA. If kma and kma_index has not bin installed please install kma_index from the kma repository: https://bitbucket.org/genomicepidemiology/kma

```
# Build container
docker build -t salmonellatypefinder .
# Go to wanted location for MLST database
# Note: The location needs to be able to be attached to your docker container.
cd /path/to/db_dir
# Clone database from git repository
git clone https://bitbucket.org/genomicepidemiology/mlst_db.git
# Install MLST database with executable kma_index program
cd mlst_db
python3 INSTALL.py kma_index non_interactive
```

SalmonellaTypeFinder database

https://bitbucket.org/genomicepidemiology/salmonellatypefinder_db/src/master/



SerotypeFinder

For serotyping of *E.coli*

SerotypeFinder in CGE bitbucket

https://bitbucket.org/genomicepidemiology/serotypefinder/src/master/

Installation

Setting up SerotypeFinder program

Warning: Due to bugs in the BioPython 1.74, if not using the Docker container, do not use that version if not using Python 3.7.

```
# Go to wanted location for serotypefinder
cd /path/to/some/dir
# Clone and enter the serotypefinder directory
git clone https://bitbucket.org/genomicepidemiology/serotypefinder.git
cd serotypefinder
```

Build Docker container

```
# Build container
docker build -t serotypefinder .
```

#Download and install SerotypeFinder database

```
# Go to the directory where you want to store the serotypefinder database
cd /path/to/some/dir
# Clone database from git repository (develop branch)
git clone https://bitbucket.org/genomicepidemiology/serotypefinder_db.git
cd serotypefinder_db
STFinder_DB=$(pwd)
# Install SerotypeFinder database with executable kma_index program
python3 INSTALL.py kma_index
```

If kma_index has no bin install please install kma_index from the kma repository: https://bitbucket.org/genomicepidemiology/kma

SerotypeFinder database

https://bitbucket.org/genomicepidemiology/serotypefinder_db/src/master/