Bacterial species identification using sequencing data

Introduction to Bioinformatics - Module 1

Public Health Alliance for Genomic Epidemiology

Ifeoluwa J. Akintayo

Bacteria Identification

The process of determining the taxon a bacteria isolate belongs.

It is important as a bacteriologist to identify pathogens in order to give the appropriate treatment options.

Bacteria Identification Methods

- Phenotypic Method (Macroscopy & Microscopy)
    - ✓ Appearance
    - ✓ Gram stains etc

- Biochemical test Method
    - ✓ Acid production
    - ✓ Gas production
    - ✓ Enzyme production
    - ✓ Sugar fermentation etc

- **Genotypic Method –** Genetic material of the organism
    - ✓ Pattern-Fingerprint based technology
    - ✓ **Sequence based technology**

Sequence-based technique

The use of DNA sequences of the unknown species, compared to a comprehensive sequence database (e.g *National Center for Biotechnology Information* - NCBI) of known species -*Many organisms have similarities to a known species.*

with the use of

Sequence alignment tool– *arrangement of DNA nucleotides of an unknown species to a set of known species to identify regions of similarities*.

Methods for species identification using sequencing data

- Basic Local Alignment Search Tool – BLAST
- SpeciesFinder
- K-merFinder
- Pathogenwatch

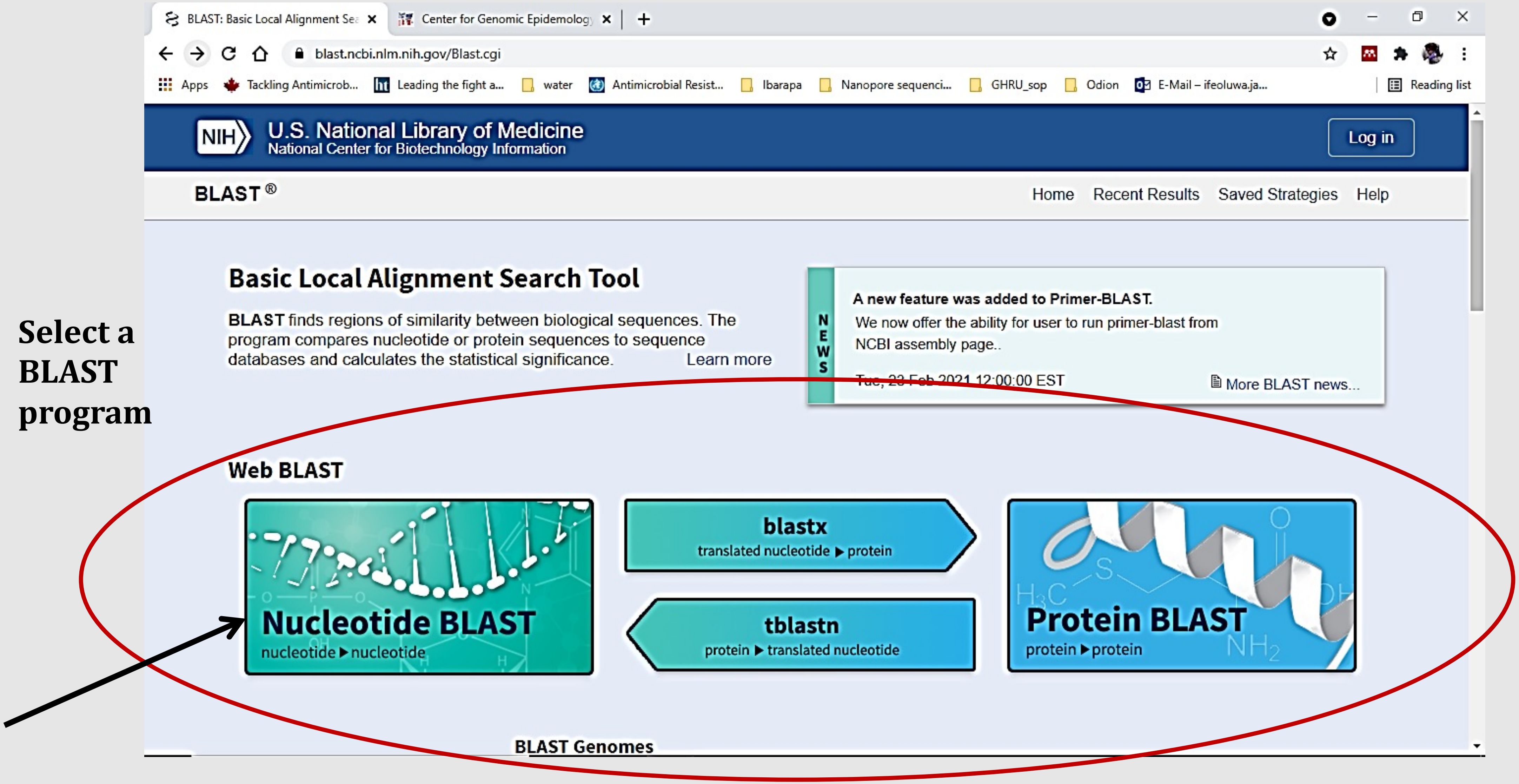
Basic Local Alignment Search Tool - BLAST

- Matches unknown sequence to known published sequences
- Compares genes and protein sequences against public database
- Search database for maximum alignment
- Returns most similar sequences from the database

How does BLAST work?

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch

Select a BLAST program



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

A new feature was added to Primer-BLAST.

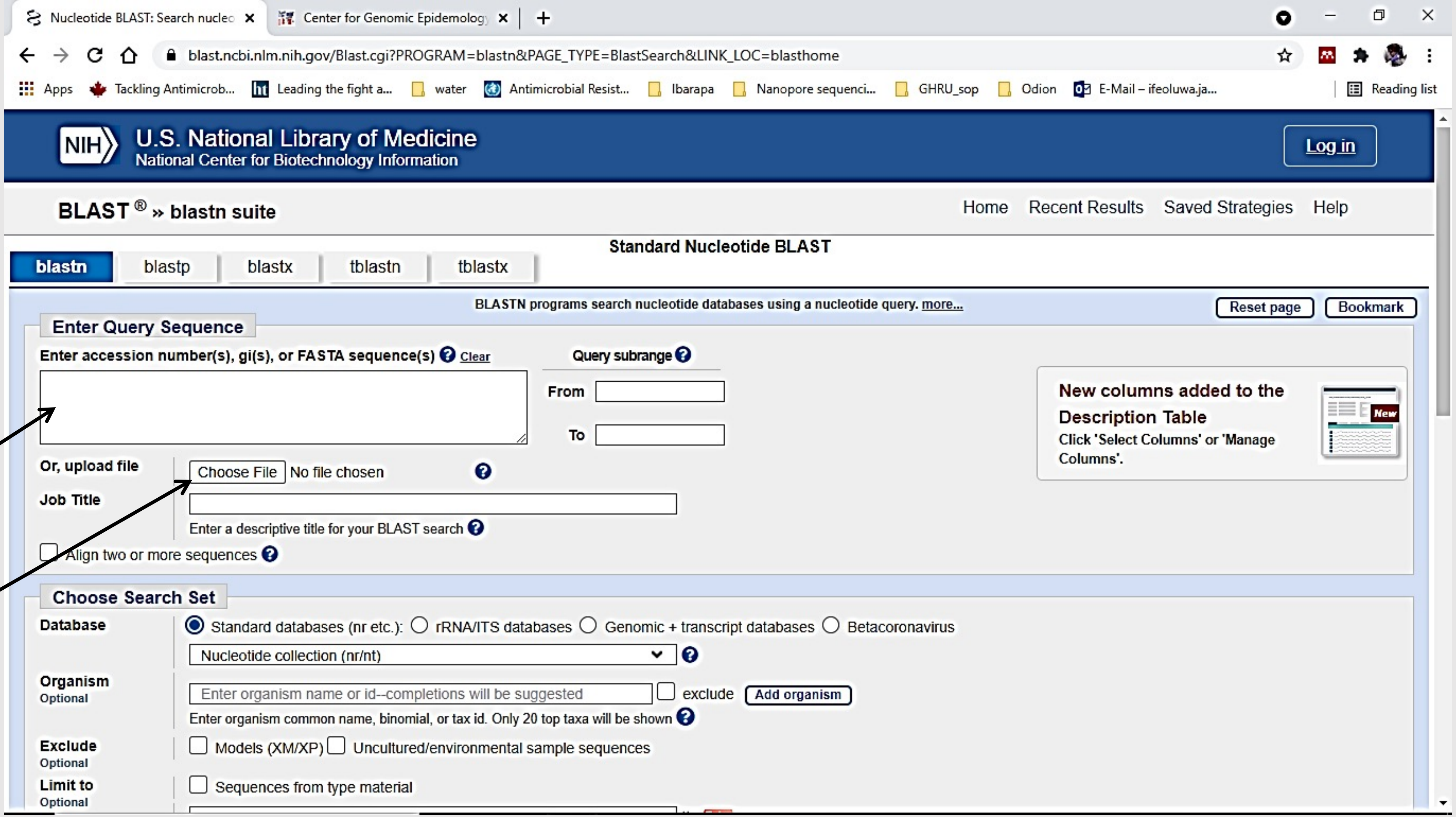
We now offer the ability for user to run primer-blast from NCBI assembly page..

Tue, 23 Feb 2021 12:00:00 EST

More BLAST news...

Web BLAST

Submit your Sequence



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Query subrange

From

To

Or, upload file — No file chosen

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

New columns added to the Description Table

Click 'Select Columns' or 'Manage Columns'.

Choose Search Set

Database: ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus

Nucleotide collection (nr/nt)

Organism
Optional

Enter organism name or id--completions will be suggested □ exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

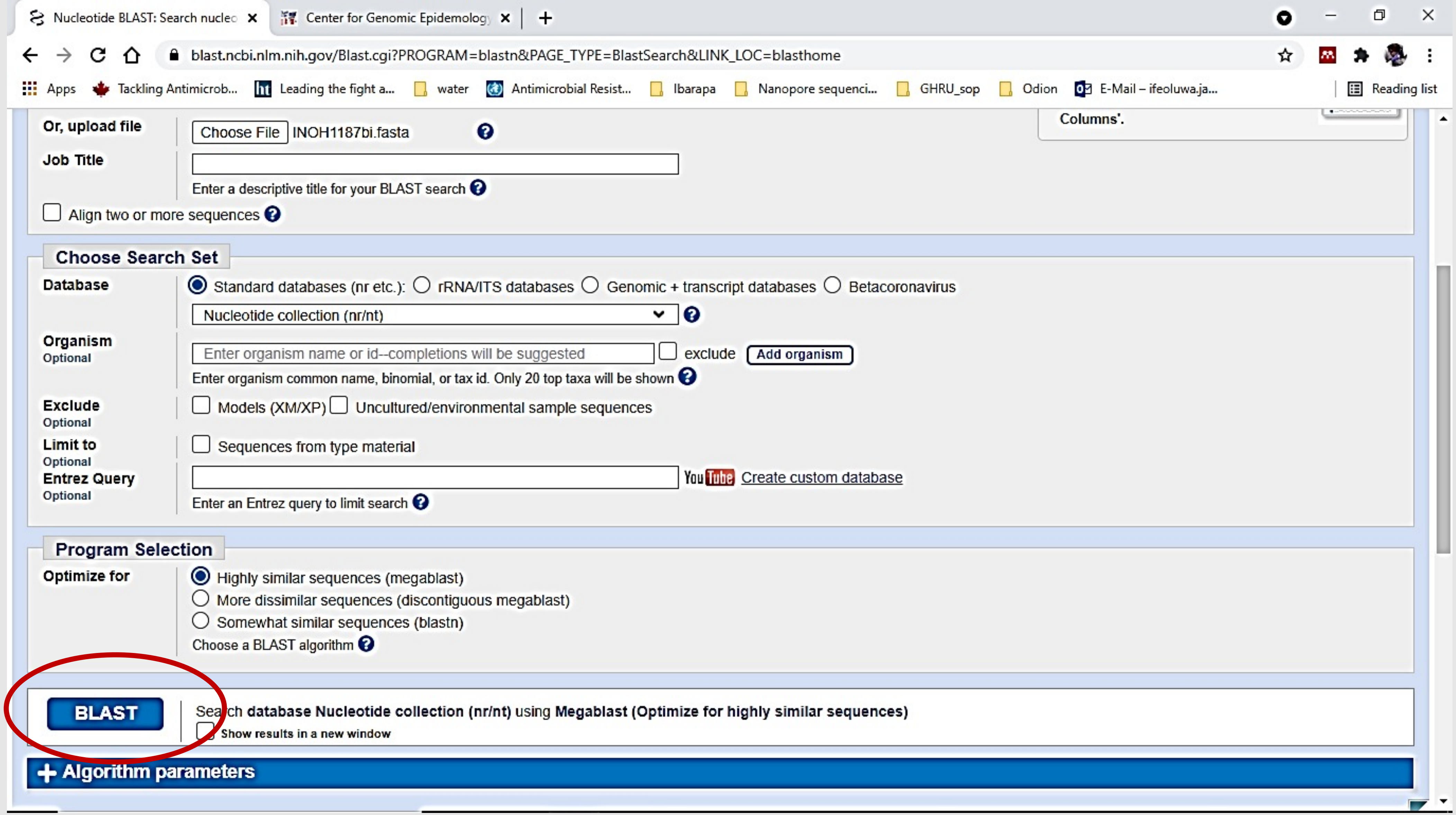
Exclude
Optional

□ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to
Optional

□ Sequences from type material

Public Health Alliance for Genomic Epidemiology



Or, upload file: Choose File INOH1187bi.fasta

Job Title:

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus

Nucleotide collection (nr/nt)

Organism
Optional: Enter organism name or id--completions will be suggested □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude
Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to
Optional: □ Sequences from type material

Entrez Query
Optional: YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for:
● Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

+ Algorithm parameters

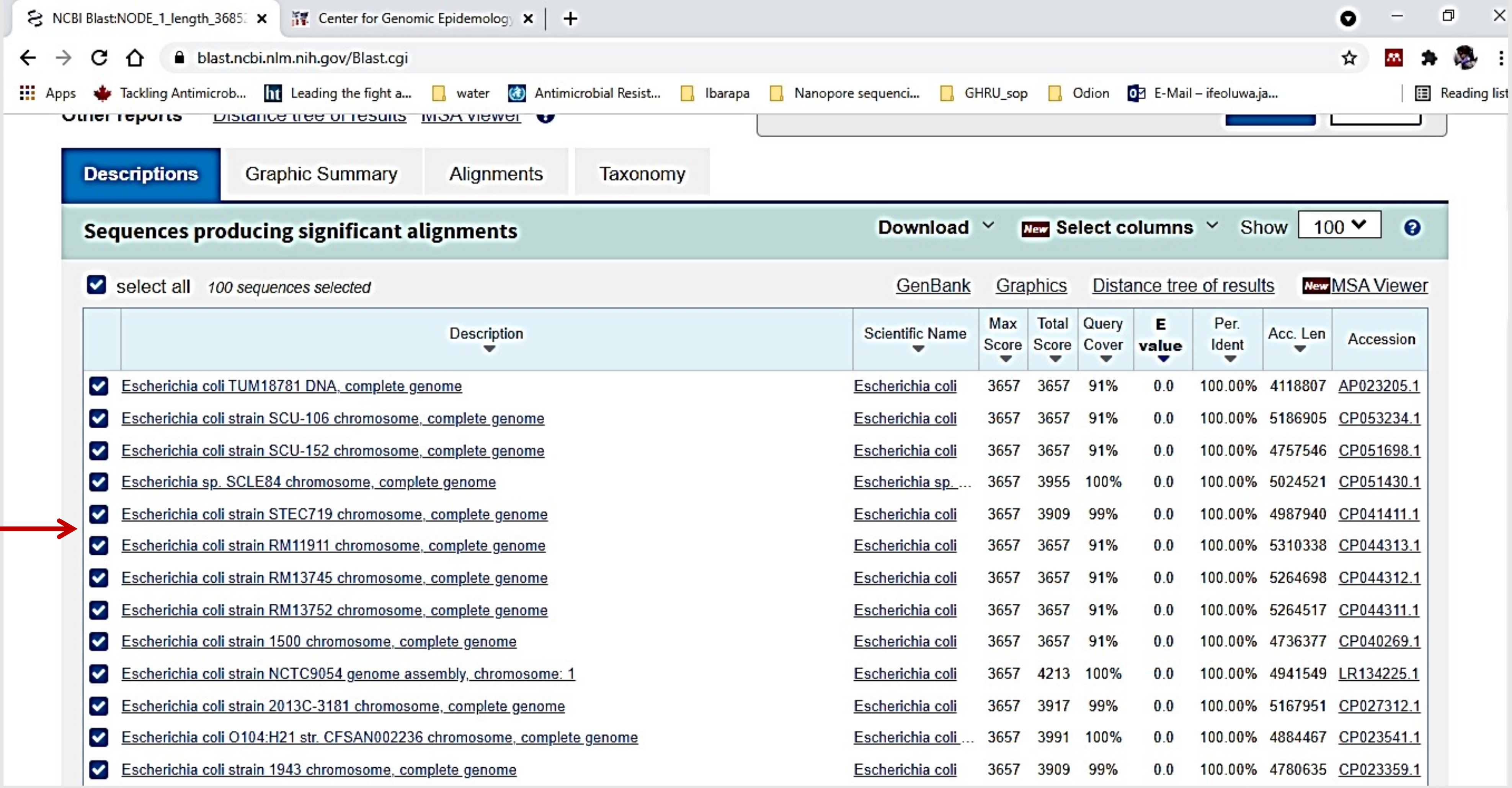
Click on BLAST

BLAST OUTPUT

List of organisms in the database containing segments where there is high similarities with your query

Per. Identity: how similar the query is to the sequences

Query cover: how much of the query is covered by the sequence



Sequences producing significant alignments

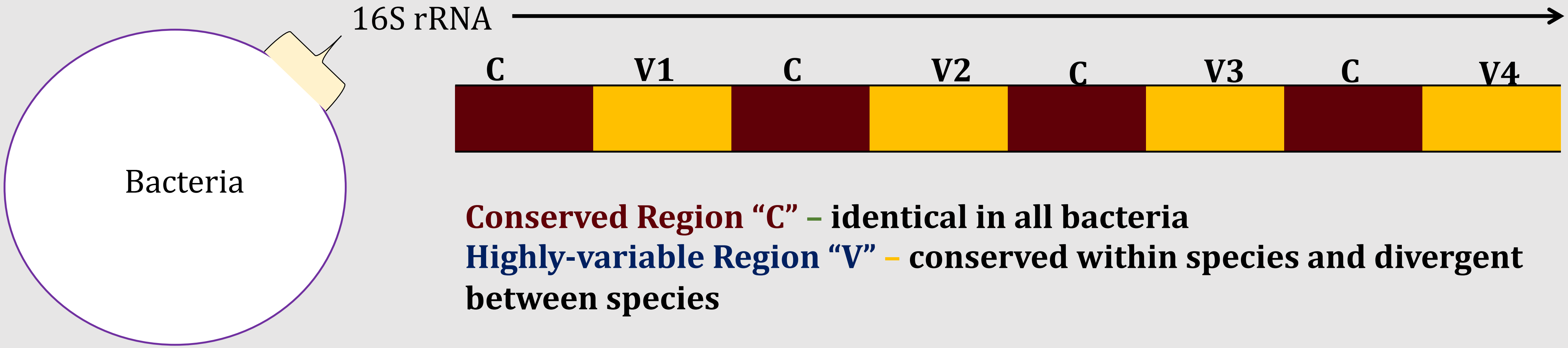
☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Escherichia coli TUM18781 DNA, complete genome	Escherichia coli	3657	3657	91%	0.0	100.00%	4118807	AP023205.1
☑	Escherichia coli strain SCU-106 chromosome, complete genome	Escherichia coli	3657	3657	91%	0.0	100.00%	5186905	CP053234.1
☑	Escherichia coli strain SCU-152 chromosome, complete genome	Escherichia coli	3657	3657	91%	0.0	100.00%	4757546	CP051698.1
☑	Escherichia sp. SCLE84 chromosome, complete genome	Escherichia sp. ...	3657	3955	100%	0.0	100.00%	5024521	CP051430.1
☑	Escherichia coli strain STEC719 chromosome, complete genome	Escherichia coli	3657	3909	99%	0.0	100.00%	4987940	CP041411.1
☑	Escherichia coli strain RM11911 chromosome, complete genome	Escherichia coli	3657	3657	91%	0.0	100.00%	5310338	CP044313.1
☑	Escherichia coli strain RM13745 chromosome, complete genome	Escherichia coli	3657	3657	91%	0.0	100.00%	5264698	CP044312.1
☑	Escherichia coli strain RM13752 chromosome, complete genome	Escherichia coli	3657	3657	91%	0.0	100.00%	5264517	CP044311.1
☑	Escherichia coli strain 1500 chromosome, complete genome	Escherichia coli	3657	3657	91%	0.0	100.00%	4736377	CP040269.1
☑	Escherichia coli strain NCTC9054 genome assembly, chromosome: 1	Escherichia coli	3657	4213	100%	0.0	100.00%	4941549	LR134225.1
☑	Escherichia coli strain 2013C-3181 chromosome, complete genome	Escherichia coli	3657	3917	99%	0.0	100.00%	5167951	CP027312.1
☑	Escherichia coli O104:H21 str. CFSAN002236 chromosome, complete genome	Escherichia coli ...	3657	3991	100%	0.0	100.00%	4884467	CP023541.1
☑	Escherichia coli strain 1943 chromosome, complete genome	Escherichia coli	3657	3909	99%	0.0	100.00%	4780635	CP023359.1

SpeciesFinder

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

Predicts species by using 16S rRNA genes- *a gene present in all bacteria species*



Conserved Region "C" – identical in all bacteria

Highly-variable Region "V" – conserved within species and divergent between species

- The gene is predicted from the sequence, predicted sequence is aligned against the 16S database
- The species associated with the best hit is the final prediction.

K-merFinder

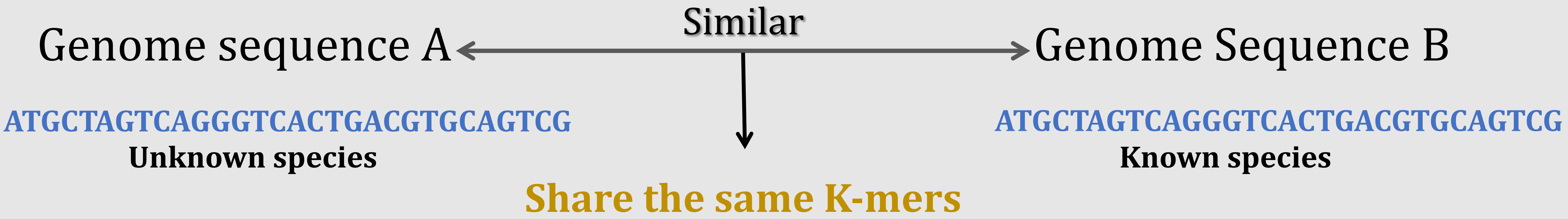
http://cge.cbs.dtu.dk/services/KmerFinder/

- Predicts species by examining the number of overlapping 16-mers

K is number of bases =16

- KmerFinder tool cuts the unknown species genome into 16mers, this is compared with 16mers of the known species in the database and gives the best matching species.

Sequence	**ATGCTAGTCAGGGTCACTGACGTGCAGTCG**
16-mers	**ATGCTAGTCAGGGTCA**
	TGCTAGTCAGGGTCAC
	GCTAGTCAGGGTCACT



Similar

Genome sequence A

Genome Sequence B

ATGCTAGTCAGGGTCACTGACGTGCAGTCG
Unknown species

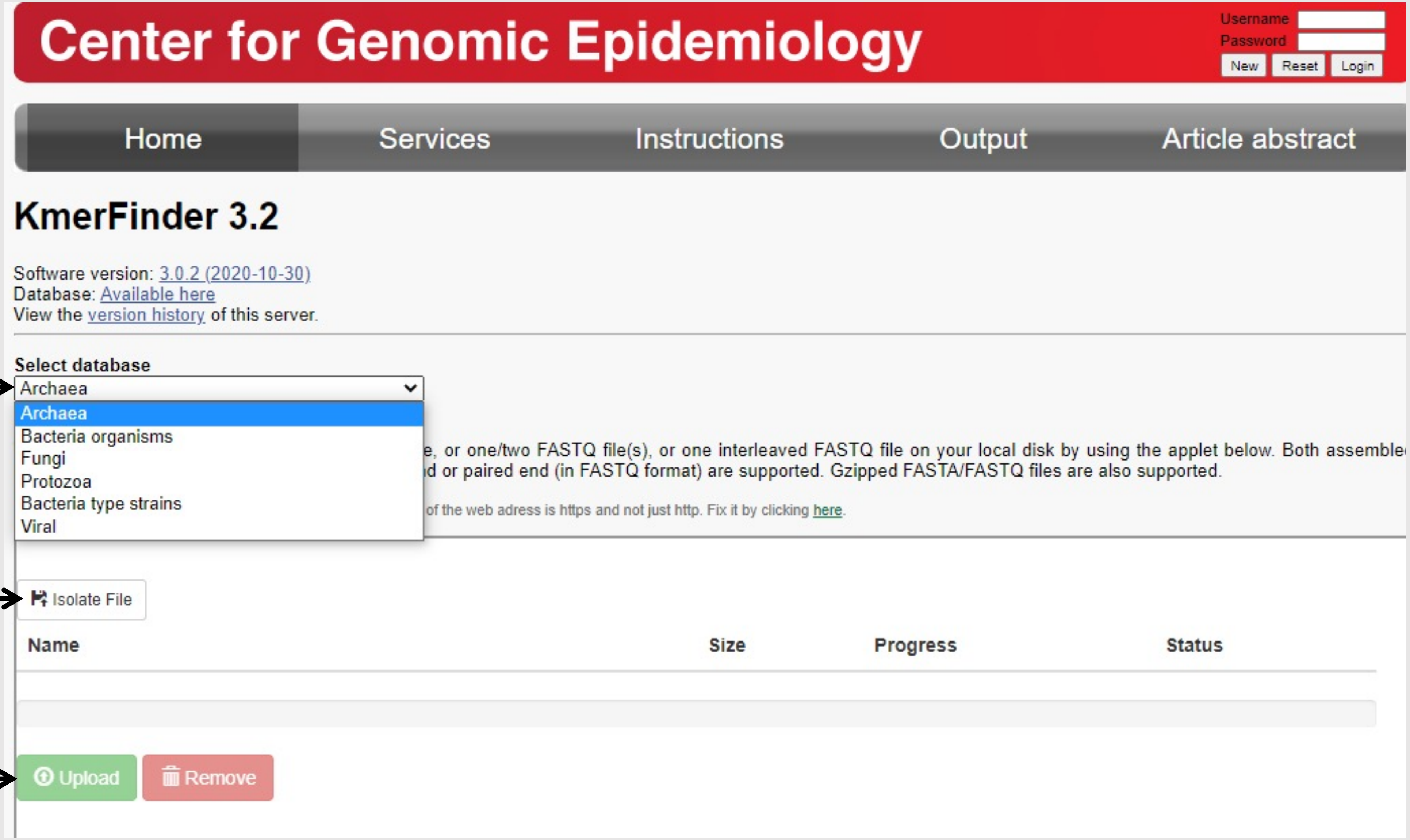
ATGCTAGTCAGGGTCACTGACGTGCAGTCG
Known species

Share the same K-mers

Database

Select file

Upload



KmerFinder 3.2

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

Select database

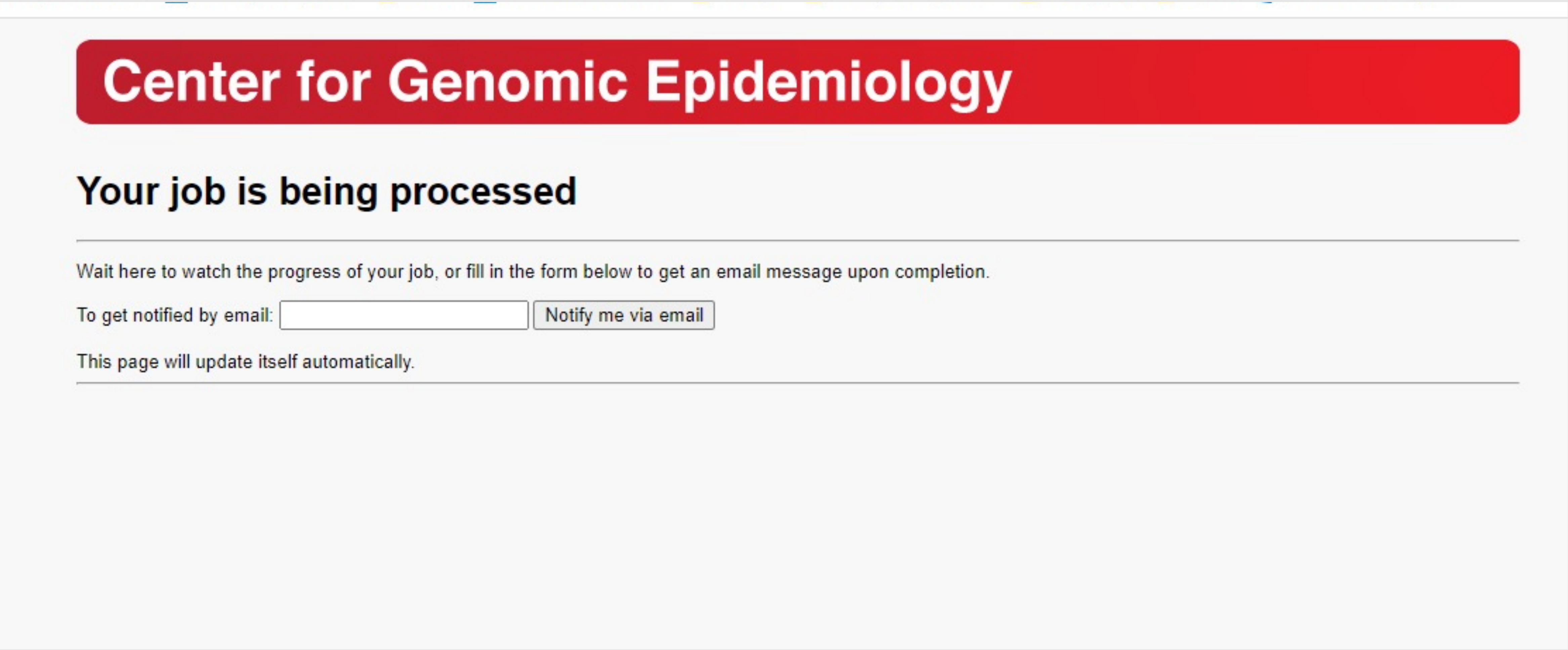
Archaea

- Archaea
- Bacteria organisms
- Fungi
- Protozoa
- Bacteria type strains
- Viral

e, or one/two FASTQ file(s), or one interleaved FASTQ file on your local disk by using the applet below. Both assemble
d or paired end (in FASTQ format) are supported. Gzipped FASTA/FASTQ files are also supported.

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

Name	Size	Progress	Status



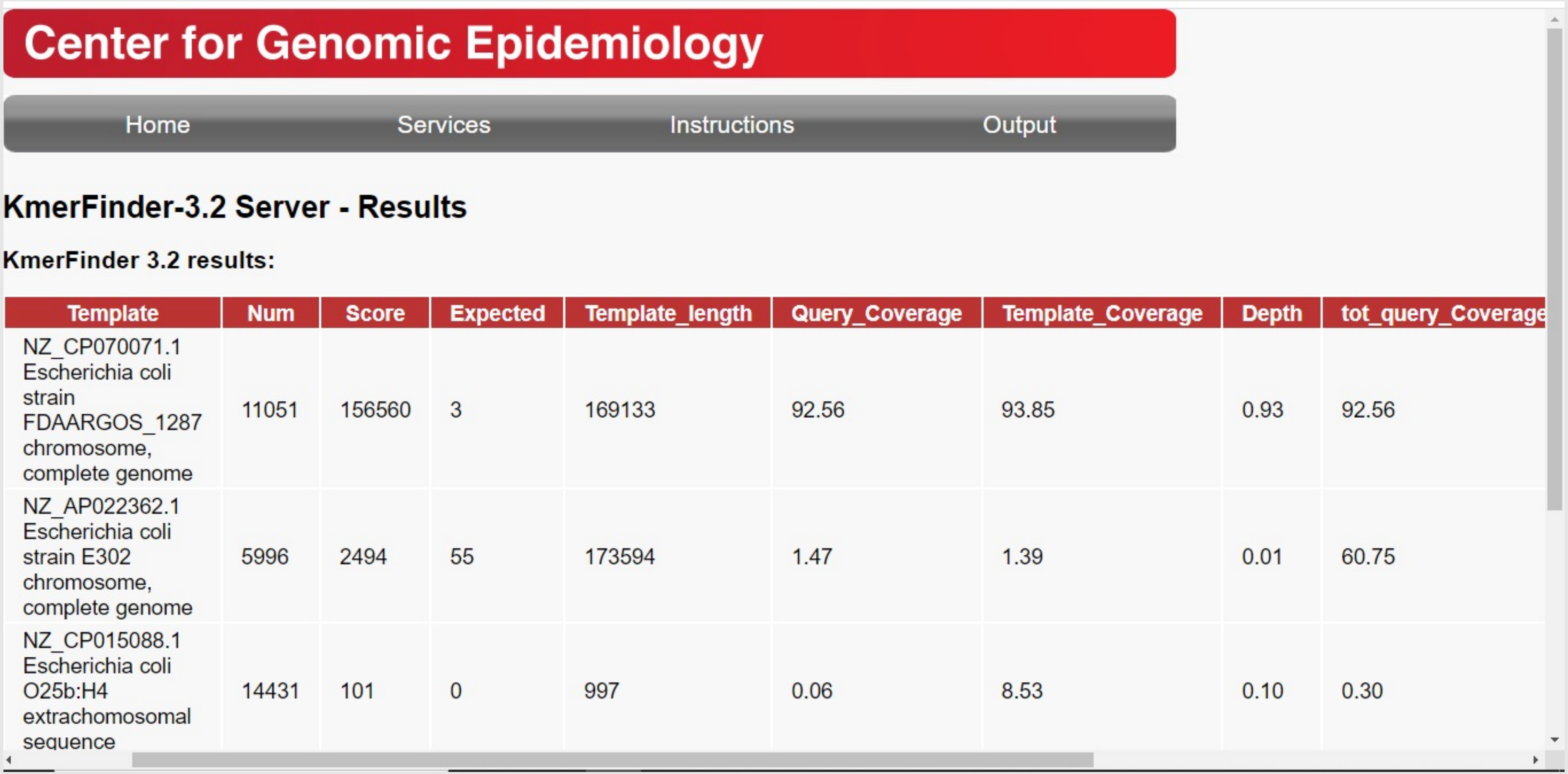
Center for Genomic Epidemiology

Your job is being processed

Wait here to watch the progress of your job, or fill in the form below to get an email message upon completion.

To get notified by email: [ ] Notify me via email

This page will update itself automatically.



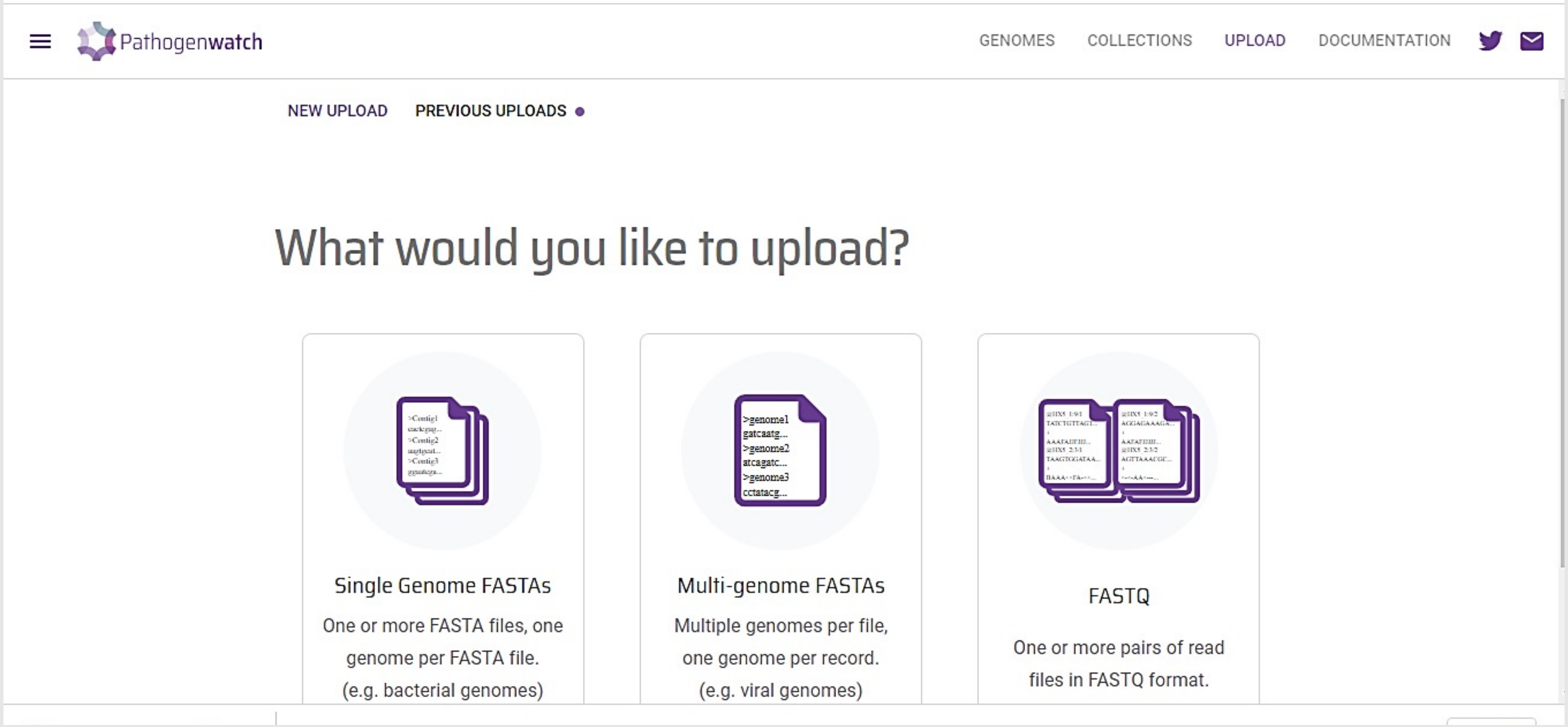
KmerFinder-3.2 Server - Results

KmerFinder 3.2 results:

Template	Num	Score	Expected	Template_length	Query_Coverage	Template_Coverage	Depth	tot_query_Coverage
NZ_CP070071.1 Escherichia coli strain FDAARGOS_1287 chromosome, complete genome	11051	156560	3	169133	92.56	93.85	0.93	92.56
NZ_AP022362.1 Escherichia coli strain E302 chromosome, complete genome	5996	2494	55	173594	1.47	1.39	0.01	60.75
NZ_CP015088.1 Escherichia coli O25b:H4 extrachomosomal sequence	14431	101	0	997	0.06	8.53	0.10	0.30

Public Health Alliance for Genomic Epidemiology

Pathogenwatch



What would you like to upload?

Single Genome FASTAs

One or more FASTA files, one genome per FASTA file. (e.g. bacterial genomes)

Multi-genome FASTAs

Multiple genomes per file, one genome per record. (e.g. viral genomes)

FASTQ

One or more pairs of read files in FASTQ format.

- SpeciesFinder
- KmerFinder
- Pathogenwatch

Public Health Alliance for
Genomic Epidemiology

Thank you

Global Health
Research Unit
Genomic Surveillance of AMR

The
Fleming
Fund

RECTE SAPERE FONS