

Computational practical 7

Online tools for assembly and antimicrobial resistance prediction

7.1 Introduction

Whole genome sequencing is rapidly being used for understanding evolution and spread of antimicrobial resistance. This has fostered the development of various bioinformatics tools that are more user-friendly and require minimal bioinformatics expertise. Through global efforts a number of antimicrobial resistance databases and tools have been developed that can help identify determinants of resistance from whole genome sequences.

In this chapter, we will be downloading publicly available sequences (genome assemblies and raw reads) from the ENA (European Nucleotide Archive) database. Afterwards de novo assembly and detection of genetic determinants of resistance using web-based tools will be performed using freely accessible web-based tools. We begin by learning how to access and download the assembled genome sequences and raw sequence reads from ENA. Next step will be to assemble the downloaded reads using web-based tools to generate contigs (long contiguous stretch of nucleotides), which will then be used to detect genetic determinants of resistance.

7.2 Downloading the assemblies/genome sequence

Step1: Open the European Nucleotide Archive (ENA) website (<https://www.ebi.ac.uk/ena>) in your web-browser.

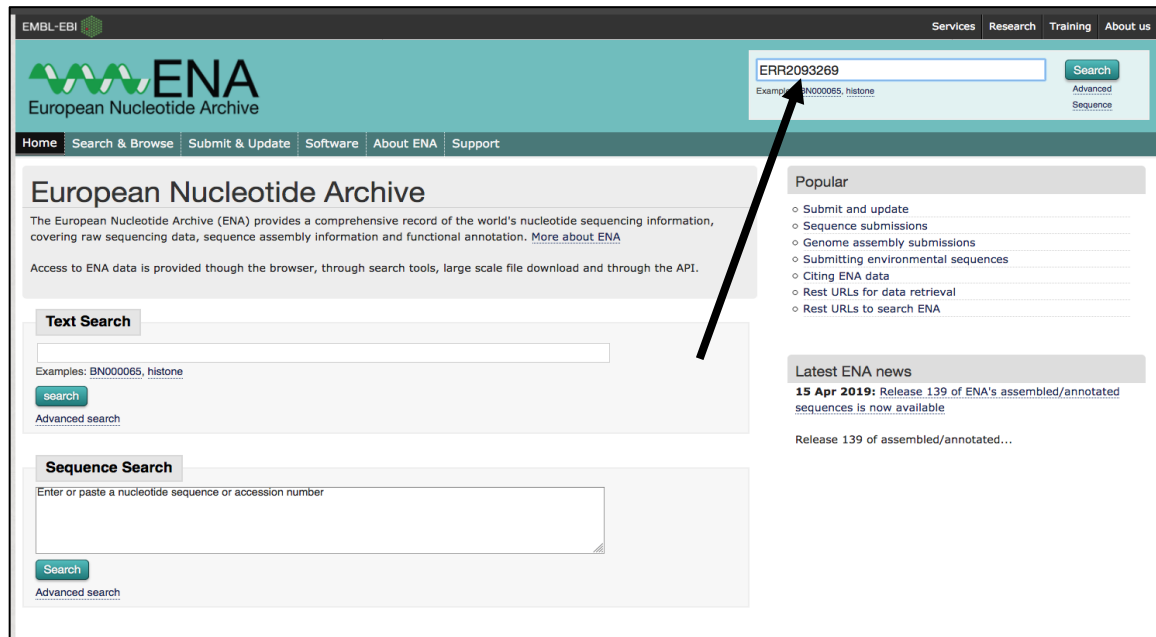


Figure 1: European Nucleotide Archive page

Enter the accession number given to you (BX571856) in the search box indicated by the arrow in figure 1 and click on the “search” button to initiate the process.

Step 2: The search will return a page with details of the experiment and run associated with the accession ID as shown below.

EMBL-EBI ENA European Nucleotide Archive

Services Research Training About us

Home Search & Browse Submit & Update Software About ENA Support

Search results for **BX571856** [Show more data from EMBL-EBI](#)

Sequence
Sequence (Release) (140)

Coding
Coding (Update) (477)
Coding (Release) (787)

Non-coding
Non-coding (Release) (16)

Sequence (Release) (140 results found)
BX571856 Staphylococcus aureus subsp. aureus strain MRSA252, complete genome
[View all 140 results](#)

Coding (Update) (477 results found)
VGU68421 Streptococcus pyogenes Ortholog of S. aureus MRSA252 (BX571856) SAR1694
[View all 477 results](#)

Coding (Release) (787 results found)
CZT40499 Streptococcus agalactiae Ortholog of S. aureus MRSA252 (BX571856)
[View all 787 results](#)

Non-coding (Release) (16 results found)
BX571856.1:519224..519338:rRNA Staphylococcus aureus subsp. aureus MRSA252 ribosomal RNA
[View all 16 results](#)

Figure 2: Accession search page

You can see the information associated with the submission details along with other information such as coding and non-coding regions with release dates. Click on the accession ID that you just entered as shown by the arrow in the figure 2.

Step 3: The window now shows the submission details including the Organism, molecule type that was sequenced, sequence length etc. Click on the “sequence” tab as shown by the **arrow1** in figure 3. Now, you can see the genome sequence in fasta format with the first line starting with the symbol “>” called the header followed by the sequence of nucleotides. Right click on the “**Show full sequence**” link as pointed by the arrow2 in and select “**save link as**” (shown by arrow3) to download the genome assembly and save it in the folder **cp7**.

Sequence: BX571856.1

Staphylococcus aureus subsp. aureus strain MRSA252, complete genome

View: TEXT FASTA XML

Download: XML FASTA TEXT

Organism	Molecule type	Topology	Data class	Taxonomic Division
Staphylococcus aureus subsp. aureus MRSA252	genomic DNA	circular	STD	PRO
Sequence length	Sequence Version	First public	Last updated	Show Version History
2,902,619	1	23-JUN-2004	29-JAN-2016	BX571856

Keywords

complete genome.

Lineage

Bacteria, Firmicutes, Bacilli, Bacillales, Staphylococcaceae, Staphylococcus

Navigation

Overview

Source Feature(s)

Sequence

Publications

Submission Details

Aligned Reads

Other Feature(s)

Showing first 1 - 1000 of 2902619

Find similar sequences

>ENA[BX571856][BX571856.1 Staphylococcus aureus subsp. aureus strain MRSA252, complete genome : Location:1..1000

CGATTAAAGATAGAAATACAGATCGAGCAATCAATTTTCATAACATCACCATGAGTTT

GATCCAAGCATGAGTGTTCACATGTTTGAATACCTTTATACAGTCTTATACATACTTT

ATAAATTATTTCCCAAGCTGTTTGTATACACACACTAACAGATACTTATAGAAGGAAA

GTTATCCACTTATCCACACTTATACTTTTGAAGATTGTGGATATAGAAATACACACA

AAGTTATACTATTTTAGCAACATATTCACAGGTATTGACATATAGAGAACTGAAAAG

TATAATTGTGGATAAGTCCTCCAACTCATGTTTATAAGGATTATTATTGATATT

TACATAAAATACCTGTGCATAACTAATAGCAGGATAAAGTTATCCACCGATTGTTATA

ACTTGTGGATAATATTAAACATGTTGTTTAGAAGTTATCCACGGTGTATTATTGTTG

TATAACTTAAAAATTTAAGAGATGAGTAAATTTATCTCGAAGAAAGAAATTTGGGAA

AAAGTCGCTTAAATGCTCAAGAAATATATCAAGCTGTAAGTACTCAACTTCCFAAA

GATCTGAGCTTTACACGATCAAGATGCTGAAGCTATCTATATCGAGTATTCCTTT

AAAGCAATTTGTTAAATCAACATATGCTGAATATCAAGCAATCTTATTGATGTT

GTAGGCTATGAGTAAAGCTCACTTTATTAATCTGAAGAAATAGCAAAATATAGTAAT

AAAGCAATCTGCTACTCCAAAGAGCAACAAACCTTCTACTGAACCAACTGAG

CATGTGCTGTGTGAGAGAGCAATTCAGGCTCATACACATTTGACACTTTGTGTA

CCTGTATA

TACAATCC

...

Open Link in New Tab

Open Link in New Window

Open Link in New Private Window

Bookmark This Link

Save Link As...

Save Link to Pocket

Copy Link Location

Search Google for "Show full sequ..."

Send Link to Device

Inspect Element

Show full sequen

Figure 3: genome assemblies/sequence search

7.3 Downloading sequence reads

Step 1: Open the ENA website (<https://www.ebi.ac.uk/ena>) in the browser and repeat step 1 from the above exercise with the accession number given to you (**ERR2093269**). This will open a window as shown in figure 4 containing the information associated with the accession ID. Click on the “Run” as pointed by the arrow in figure 4.

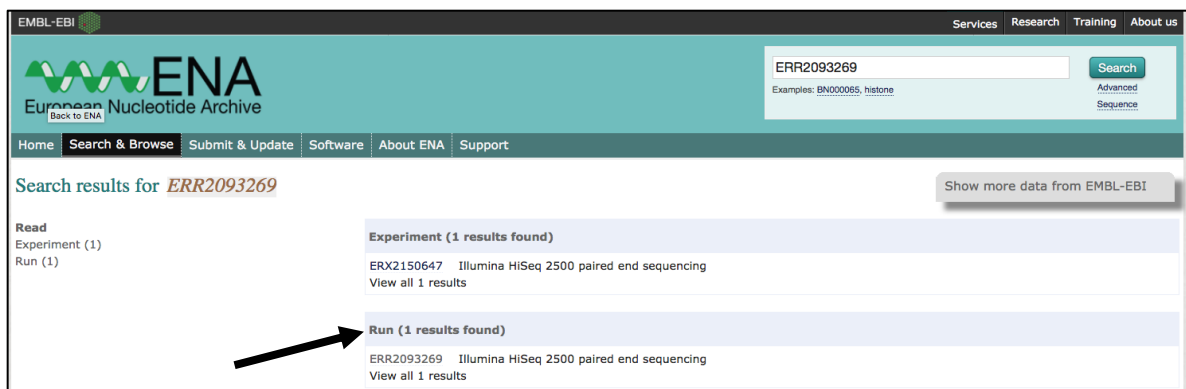


Figure 4: The accession page

Step2: Right click on “**File1**” as pointed by the arrow in figure 5 and select “**save link as**” figure 5. Save the compressed FASTQ file in the folder **cp7**. Repeat the same steps for “**File 2**”. Since the read files contain millions of fragments of the genome it is larger in size which is the reason why these are usually stored as compressed files which are easier to handle computationally.

Run: ERR2093269

Illumina HiSeq 2500 paired end sequencing

View: XML

Download: XML

Navigation: Show

Read Files: Hide

Organism: [Salmonella enterica subsp. enterica serovar Typhi](#)

Sample Accession: SAMEA103981538

Instrument Platform: ILLUMINA

Instrument Model: Illumina HiSeq 2500

Read Count: 2718556

Base Count: 679639000

Center Name: SC

Library Layout: PAIRED

Library Strategy: WGS

Library Source: GENOMIC

Show More

Read Files

Show Column Selection

Download report: JSON TSV

Download Files as ZIP

Download selected files

Download All

Study Accession	Sample Accession	Experiment Accession	Run Accession	Tax Id	Scientific Name	Generated FASTQ files: FTP	Su
PRJEB20363	SAMEA103981538	ERX2150647	ERR2093269	90370	Salmonella enterica subsp. enterica serovar Typhi	<input type="checkbox"/> ERR2093269_1.fastq.gz <input type="checkbox"/> ERR2093269_2.fastq.gz	2

Figure 5: Sequencing information page

Now we have learned about accessing the ENA database, identifying specific strain data using accession Ids and downloading the assemblies and sequence reads. In practice, when we perform sequencing two sequence files called reads are generated. These are mostly the called paired end reads. In order to detect the resistance determinants a series of steps are performed which we are going to understand in the next section.

7.4 Performing the assembly of sequence reads

The raw sequence reads contain the genome information in form of millions of short reads and therefore needs to be assembled into a larger set of contigs. There are a number of freely available computational tools such as velvet, SPAdes etc. These are command-line tools therefore require some basic computational knowledge to be able to use them.

Pathogenwatch is one web tool that can perform assembly in addition to many other tasks such as mlst typing, serotyping and antimicrobial resistance.

Step1: Open the website (<https://pathogen.watch/>) in your web-browser. Click on the upload tab on the top right corner (figure6), it will take you to a sign-in page. After you have signed in you can drag and drop the fastq files. The upload process will start automatically followed by assembly into contigs and other analyses.

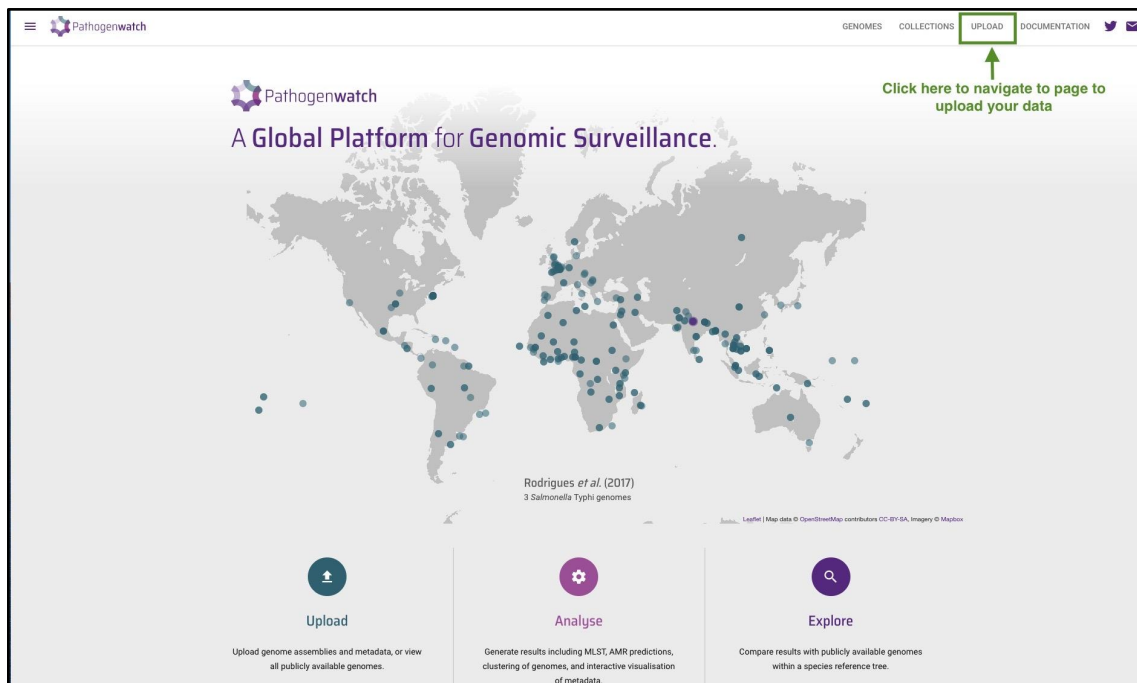


Figure 6: Pathogenwatch

Step2: Once the analysis is complete you will see the page display the information shown below (figure7). Click on the view genomes link to see the page with all the information about the uploaded genome.

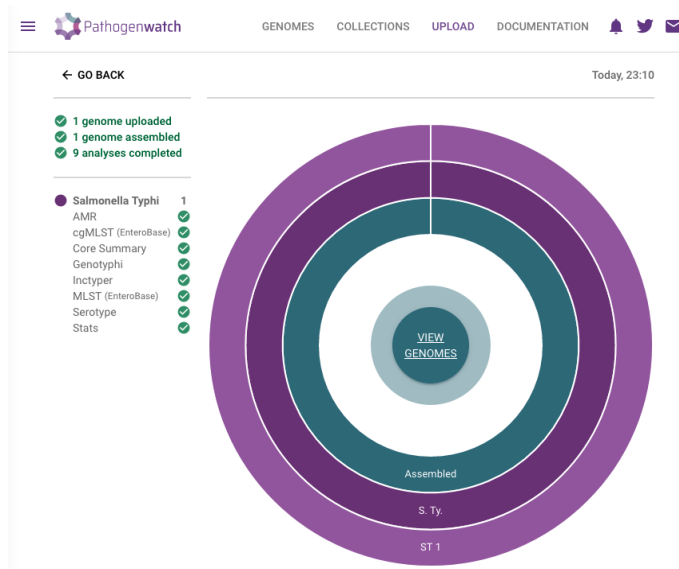


Figure 7: uploading the fastq files

Once we see the webpage as shown in figure 8, click on the tab “download data”. From the options presented select “fasta files” option to download the assembly file.

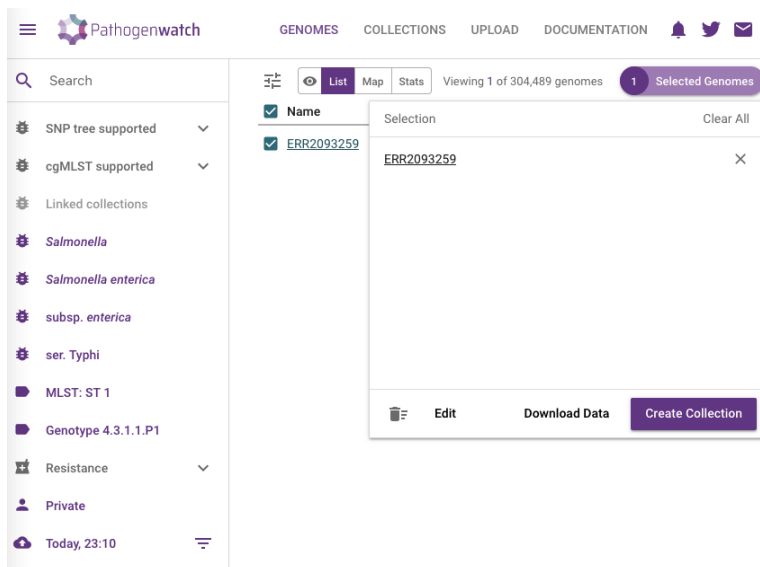


Figure 8: downloading the assembly file from pathogenwatch

7.5 Detecting the genetic determinants of resistance

The genome sequence of an organism constitutes the information about the genes that are translated into proteins. Over the years, a considerable number of genes and mutations have been found to mediate resistance to particular antibiotics. Bacteria can either acquire these genes horizontally or can evolve mutations in the genes that mediate resistance. There are several databases such as Comprehensive Antimicrobial Resistance Database (CARD) and ResFinder that contain information about the genes and mutations that confer resistance.

In this section we are going to use three different web-based tools (Pathogenwatch, ResFinder and CARD) to identify genetic determinants in the whole genome assemblies that we just created above.

7.5.1 Detection with Pathogenwatch

Pathogenwatch (<https://pathogen.watch/>) is one of the simplest web-based platforms developed by the Centre for genomic Epidemiology group that can be used to detect resistance in the genomes in many bacterial pathogens (but not all). The assemblies that we generated/ downloaded can be directly uploaded as input for this tool. Once uploaded the tool performs strains identification, MLST determination and resistance prediction in an automated manner. Recently, the website has been upgraded with an option to directly upload the raw reads but the analysis takes more time than usual so we will be using assemblies that we have already downloaded.

Step 1: Open the website (<https://pathogen.watch/>) in the Firefox web-browser. Click on the “upload” button on the top right corner as indicated by the arrow in figure 9 and select “single genome fasta”.

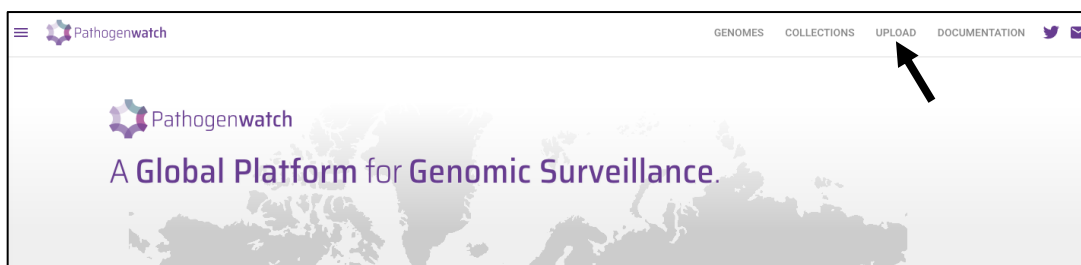


Figure 9: Pathogenwatch website

Step 2: Go to the folder and select the assembled sequence files. Drag the selected files into the web-browser where the above site is open (figure10). The files are then automatically uploaded and analysed.

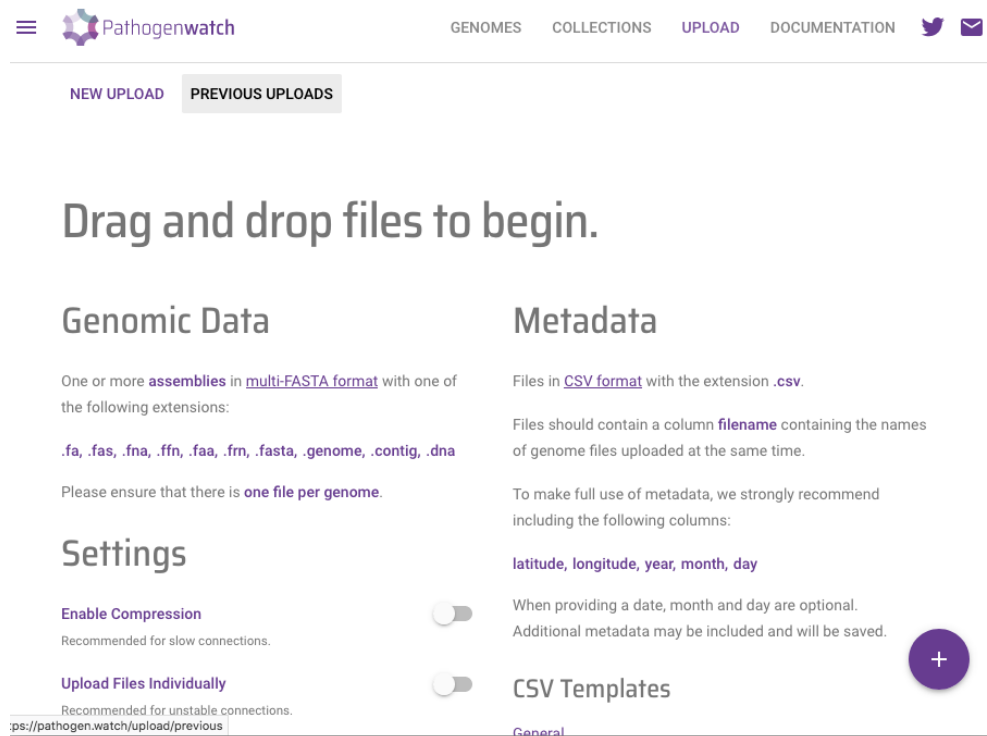


Figure 10: Uploading the assemblies to Pathogenwatch

Step 4: Once completed select “view genomes” which will open a tabular window with details on the strain (figure 11). Click on the filename you just uploaded to see the prediction results.

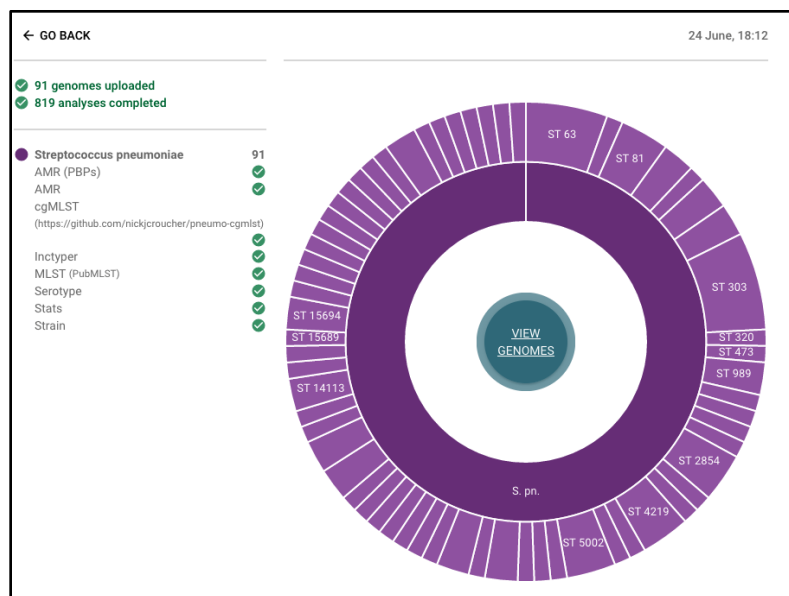


Figure 11: The status of the analysis by Pathogenwatch

Step 5: A sample result of prediction analysis performed by Pathogenwatch is shown in figure 12. As we can see the tool has performed MLST analysis in addition to identifying the resistance determinants. The results page clearly shows the information about the genome. In addition to predicting the resistance, it can simultaneously detect the sequence type information as well. You should record the SNPs/genes detected for your strain/genome.

<

Figure 12: The tabular view of the analysed genomes by Pathogenwatch

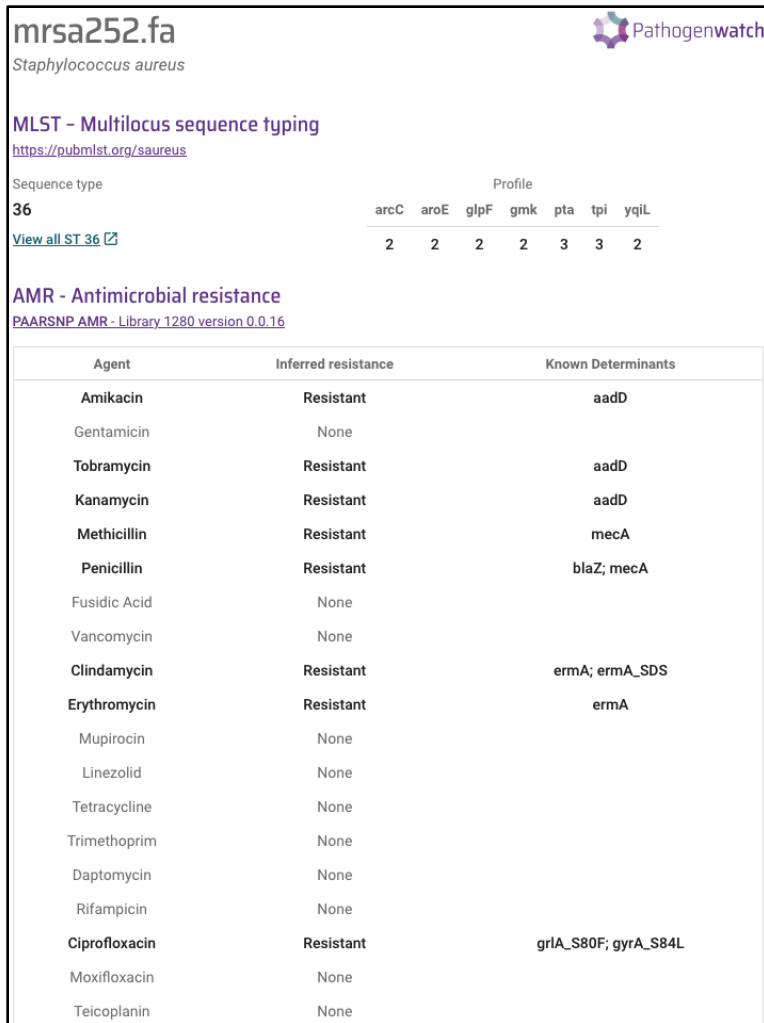


Figure 13: The results of the analysis by Pathogenwatch

Here, we can see the strain name “mrsa252”, detected sequence type “36” and alleles together with resistance determinants detected in the form of a table. For each drug the corresponding genetic determinant identified for example the gene “mecA” was identified as shown in figure 18 which confers resistance to methicillin (and most other Beta-lactam antibiotics). For SNPs the mutations identified are mentioned along with the gene name for example as shown in figure 18 the SNPs griA_S80F (non-synonymous mutation causing change of serine to phenylalanine at the 80th codon of the GriA) and gyrA_S84L (non-synonymous mutation causing change of serine to leucine at the 84th codon of GyrA) were identified to be present and which confer resistance to the fluoroquinolone antibiotic ciprofloxacin.

Step 6: The results for the prediction should be stored in a tabular form as shown below. In the first column write the strain name, in the second column write the genes and SNPs detected as observed in the previous step and third column is for recording the corresponding drug names.

An example for mrsa252 is shown below:

IsolateID	Drugs	Genes/SNPs detected (Pathogenwatch)
mrsa252	Amikacin,Tobramycin, Kanamycin	<i>aad</i>
	Methicillin	<i>mecA</i>
	Erythromycin, Clindamycin	<i>ermA</i> , <i>ermA</i> _SDS
	Ciprofloxacin	<i>griA</i> _S80F, <i>gyrA</i> _S84L

Pathogenwatch tool is the simplest to use without much requirement of bioinformatics expertise. One has to be careful when using the tool as the resistance prediction is done only for a limited number of bacterial species. The information about the bacterial species the tool works for at the moment can be found on the website. Therefore, you need to be careful while deciding on the tools for analysis. Prediction of resistance is generally based on a database of previously known genetic determinants which is maintained and updated by the developers of the specific tools. Therefore, it might be possible that a certain database might not contain newly identified or novel genetic determinants. Hence it would be important to confirm the predictions by comparing with other tools which we will be doing in the next section.

While WGSA is a web-based tool which is easy to use, the essential step is to assemble the raw reads into contigs. The results as you have observed resistance prediction is not limited to drug classes/mechanisms instead it provides output with specific drugs a particular strain is resistant to. This makes it really easy to understand the results and compare it with the phenotypic drug sensitivity results.

7.5.2 Using ResFinder webtool.

ResFinder is developed by researchers at the Centre for Genomic Epidemiology at DTU in Denmark. It's another web-based tool designed for automated detection of resistance conferring genes and mutations. The tool uses a database of previously determined resistance conferring genetic determinants and uses two different bioinformatic tools in an automated manner to detect genes and mutations respectively. Its usage requires minimum bioinformatics expertise and is freely available to users worldwide but can be slow at times as the jobs are run on the basis of queues.

Again, we will be using the “mrssa252.fa” file.

Step 1: Open the website (<https://cge.food.dtu.dk/services/ResFinder/>) in your web-browser. Select the chromosomal mutations (figure 14, this will inform the pipeline to look for SNPs as well in the genome which aren't detected by default) and select the appropriate organism. Select the option for acquired resistance genes indicated by Arrow 2. also then select the species of the genome in our case it is “*Staphylococcus aureus*”. Click on the “isolate” tab and select the file “mrssa252.fa” and click open. Click on the “upload” button to upload the data and initiate the analysis. Once the analysis is finished the window will appear as shown in figure 15.

Center for Genomic Epidemiology

Home

Services

Publications

Contact

ResFinder 4.1

Service

Instructions

Output

Article abstract

Citations

Overview of genes

Database history

Unstable services.

Dear user of the CGE services. As you may have noticed, our services have been suffering from several periods of down time lately.

Coming soon!

We have been working on an entirely new platform for CGE. This includes completely new servers and a completely new infrastructure, which will make our platform much more stable.

We will start moving services after New Year.

We are very sorry for the inconvenience these down times are causing, and we thank you for your patience. We are very excited about the new infrastructure, and we are working as hard as we can to get it online.

The database is curated by:

Frank Møller Aarestrup

(click to contact)

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

ResFinder and PointFinder software: (2022-08-08)

ResFinder database: (2022-05-24)

PointFinder database: (2022-04-22)

For analysis part of EFSA, go to [ResFinder-EFSA](#)

Chromosomal point mutations

Select threshold for %ID

90 %

Select minimum length

60 %

Show unknown mutations, not found in the database

Acquired antimicrobial resistance genes

Select Antimicrobial configuration

Aminoglycoside

Beta-lactam

Colistin

Disinfectant

Fluoroquinolone

Fosfomycin

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

Select threshold for %ID

90 %

Select minimum length

60 %

Select species

Salmonella spp.*

*Chromosomal point mutation database entry

Select type of your reads

Assembled Genome/Contigs

If you get an "Access forbidden. Error 403". Make sure the start of the web address is https and not just http. Fix it by clicking [here](#).

Choose File(s)

Name

Size

Progress

Status

Upload

Remove

Confidentiality:

The sequences are kept confidential and will be deleted after 48 hours.

Figure 14: ResFinder web server

Step 2: Once the analysis is complete the prediction results appear in a tabular form which contains the columns like detected resistance genes, identity etc. As you can see, the prediction is made according to the drug class and not individual drugs and if the gene conferring resistance to a particular class is identified is detailed in the table. Here. The results for the strain MRSA 252 showed the presence of genes such as *blaZ*, *mecA*, *aadD*, *ermA* and *ant(4)-Ib*. In addition, two SNPs: S80F in *griA* and S84L in *gyrA* have been identified.

Step 3: Now that we have seen how to use the tool ResFinder, you should record the in the table that you created previously. Add the results of the findings from ResFinder both genes and mutations identified as shown below. Sometimes the same gene can be known with different names and therefore we should be careful when comparing the results from different tools. For example, *aad* gene found in the results from both Pathogenwatch and ResFinder can also be known as *ant(4)-Ib* and *aadD2*. This information is also shown in the column “notes” of the ResFinder results. Remember to add these alternate names as well in the table which will be useful when comparing results of different prediction tools.

Strain	Drugs	Genes/SNPs detected (Pathogenwatch)	Genes/SNPs detected (ResFinder)
mrsa252	Amikacin, Tobramycin, Kanamycin	aadD / ant(4)-Ib / aadD2	present
	Methicillin	mecA	present
	Penicillin	blaZ	present
	Erythromycin, Clindamycin	ermA, ermA_SDS	present
	Ciprofloxacin	griA_S80F, gyrA_S84L	present

7.5.3 Using CARD (Comprehensive Antimicrobial Resistance Database).

CARD database (<https://card.mcmaster.ca/home>) is a collection of curated reference sequences of the genes and mutations that confer resistance to various drugs. The database is developed and maintained by laboratories of Drs. Gerry Wright and Andrew G McArthur of McMaster University's Department of Biochemistry & Biomedical Sciences (Hamilton, Ontario, Canada). This is a freely available online tool that can be used to investigate the presence of resistance in the genomes.

In this section we will again be using the sequence file “mrsa252.fa” to detect resistance determinants.

Step 1: Open the website (<https://card.mcmaster.ca/analyze>) in your web-browser and click “Analyze” on the top-right corner. You will see the screen as shown in figure 15 and click on “RGI”.

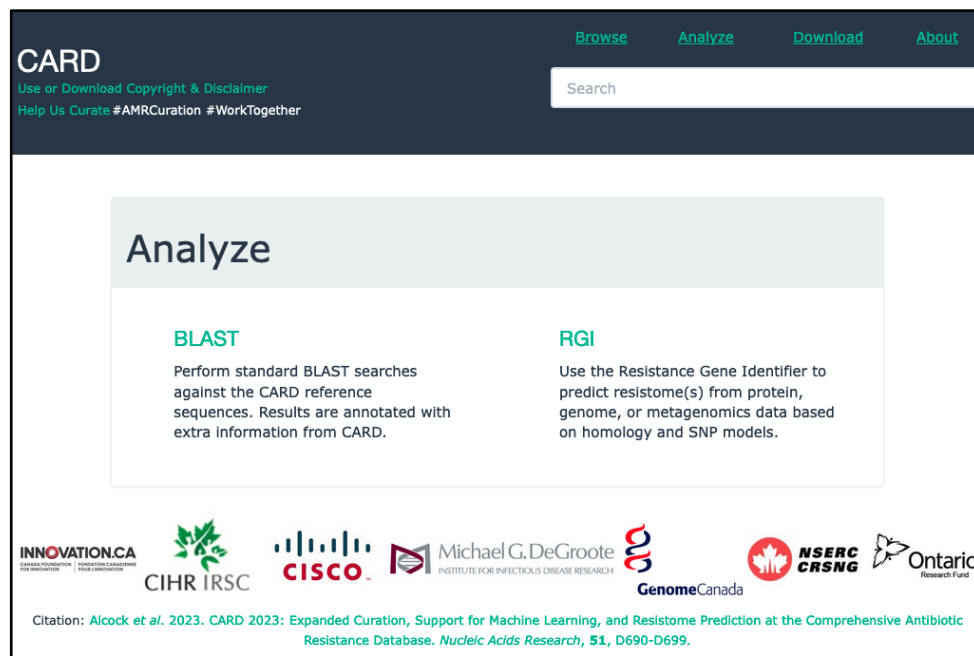


Figure 15: CARD database web server

Step 2: Click on the “Choose” button as indicated in figure 16, select the file “mrsa252.fa”. Click “open” to upload the sequence. Upload the file and click “submit” as indicated to initiate the analysis.

CARD
Use or Download Copyright & Disclaimer
Help Us Curate #AMRCuration #WorkTogether

[Browse](#) [Analyze](#) [Download](#) [About](#)

Search

RGI Resistance Gene Identifier

RGI can be used to predict resistomes from protein or nucleotide data based on homology and SNP models. Analyses can be performed via this web portal (20 Mb limit), via the command line, or via use of a [Galaxy wrapper](#). The command line version can be obtained from the [Download section of the CARD website](#). You can additionally install RGI from Conda or run RGI from Docker.

This web portal supports analysis of genomes, genome assemblies, metagenomic contigs, or proteomes. The command line tool additionally supports analysis of metagenomic reads and k-mer prediction of pathogen-of-origin for AMR genes.

Web portal - RGI 5.0.0, CARD 3.0.3: Open Reading Frame (ORF) prediction using [Prodigal](#), homolog detection using [DIAMOND](#), and Strict significance based on CARD curated bitscore cut-offs. Options included for percent identity filtering, low quality/coverage assemblies, merged metagenomic reads, small plasmids or assembly contigs (<20,000 bp).


Online RGI results cached for 7 days. As the CARD curation evolves, the results of the RGI evolve. RGI targets, reference sequences, and significance cut-offs are under constant curation. Full documentation for the RGI can be found at [GitHub](#).

Use RGI:

Enter a GenBank accession(s):
Enter accessions separated by commas
Nucleotide sequences will undergo ORF calling to generate predicted protein sequences. Examples: JN420336.1, AY123251.1, HQ451074.1, AL123456

Upload FASTA sequence file(s):
Choose Files no files selected
Upload a **plain text** file containing DNA or protein sequence(s) in FASTA format (20 Mb limit). The file can contain more than one FASTA formatted sequence, such as assembly contigs or multiple proteins. Each file will be treated as a single sample.

Sequence Quality:
☒ High quality/coverage¹
☐ Low quality/coverage²

☐ I'm not a robot 

Select Data Type:
☒ DNA sequence
☐ Protein sequence

Select Criteria:
☒ Perfect and Strict hits only
☐ Perfect, Strict and Loose hits

Nudge ≥95% identity Loose hits to Strict:
☒ Exclude nudge
☐ Include nudge

Submit

Figure 16: Uploading the assemblies to CARD database

studies. The CARD results only mentions the name *parC* rather than *grlA* so one had to be careful when recording these results.

Strain	Drugs	Genes/SNPs detected (Pathogenwatch)	Genes/SNPs detected (ResFinder)	Genes/SNPs detected (CARD)
mrsa252	Amikacin, Tobramycin, Kanamycin	aadD / ant(4)-Ib / aadD2	present	present
	Methicillin	mecA	present	present
	Penicillin	blaZ	present	present
	Erythromycin, Clindamycin	ermA, ermA_SDS	present	present
	Ciprofloxacin	grlA_S80F, gyrA_S84L	present	present

Now, we have put together the prediction results of 3 different tools for the genome of MRSA 252. We first identified the genes and the drugs to which the strain was predicted to be resistant using the tool pathogen watch. Then we analyzed the same genome with two other tools ResFinder and CARD and confirmed the prediction made by Pathogenwatch and therefore the resulting table is the drug resistance profile for the strain MRSA 252. This increases our confidence in the predictions made using the genomic data.