Computational Practical 7: Online tools for assembly and antimicrobial resistance prediction

Module Developers: Dr Narender Kumar and Dr Anthony Smith

Table Of Contents

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.

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

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.

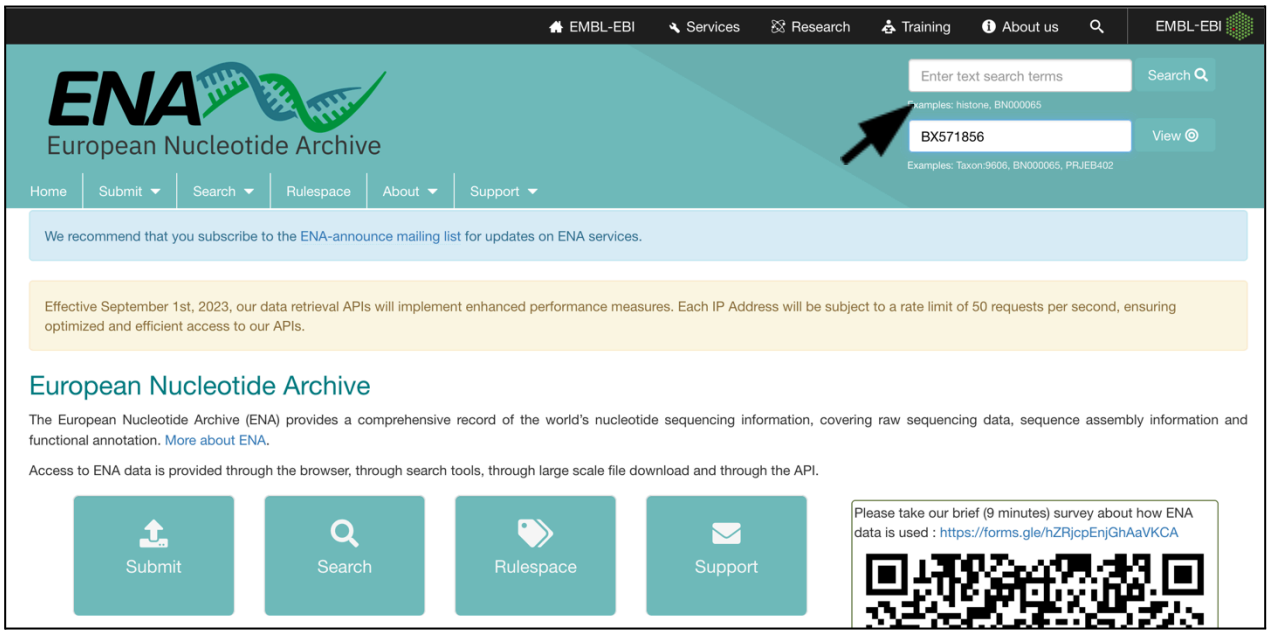


Figure 1: European Nucleotide Archive page

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 associated with the accession ID as shown below.

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

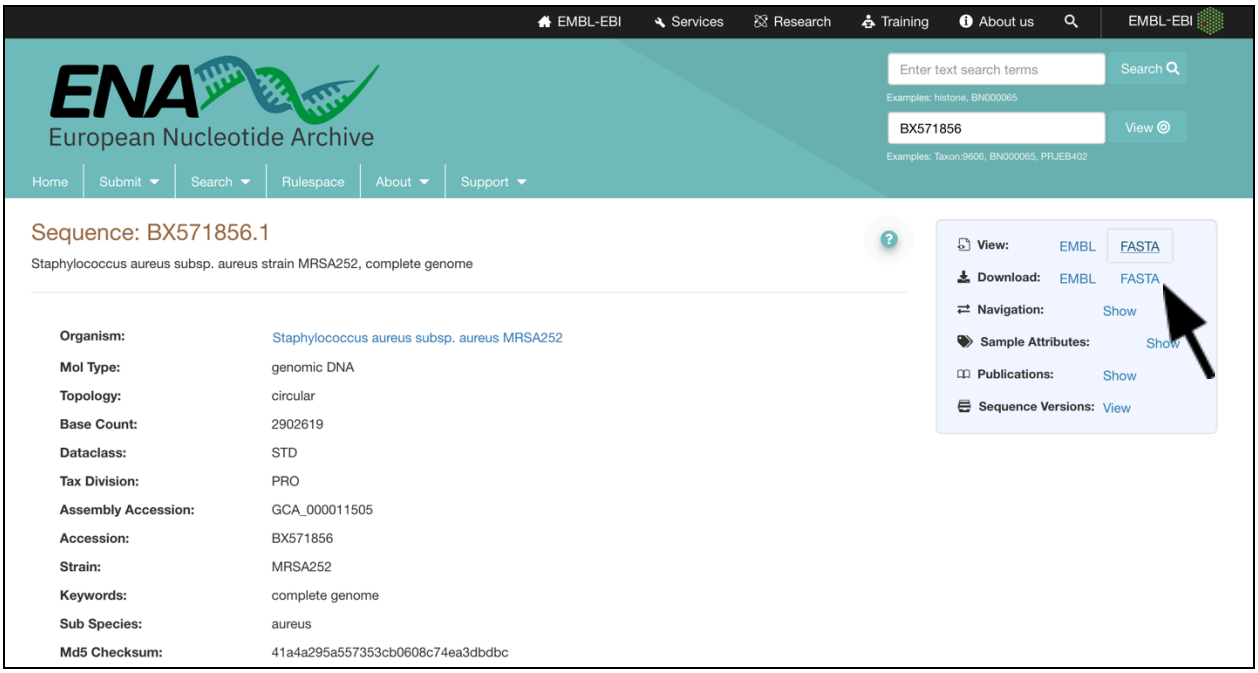


Figure 2: Accession search page

You can see the information associated with the submission details along with other information such as Base Count (genome size), Strain name etc. Right click on the "FASTA" as pointed by the arrow in the figure above and save the file to the **cp7** folder on your computer.

You can open this file to see the genome sequence in fasta format with the first line starting with the symbol ">" called the header followed by the sequence of nucleotides as shown below (**figure 3**). Here, the header line contains the information: accession number, species name, strain name and the indication that this is a complete genome sequence.

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

Figure 3: A snapshot of the downloaded fasta file

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

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 this accession ID.

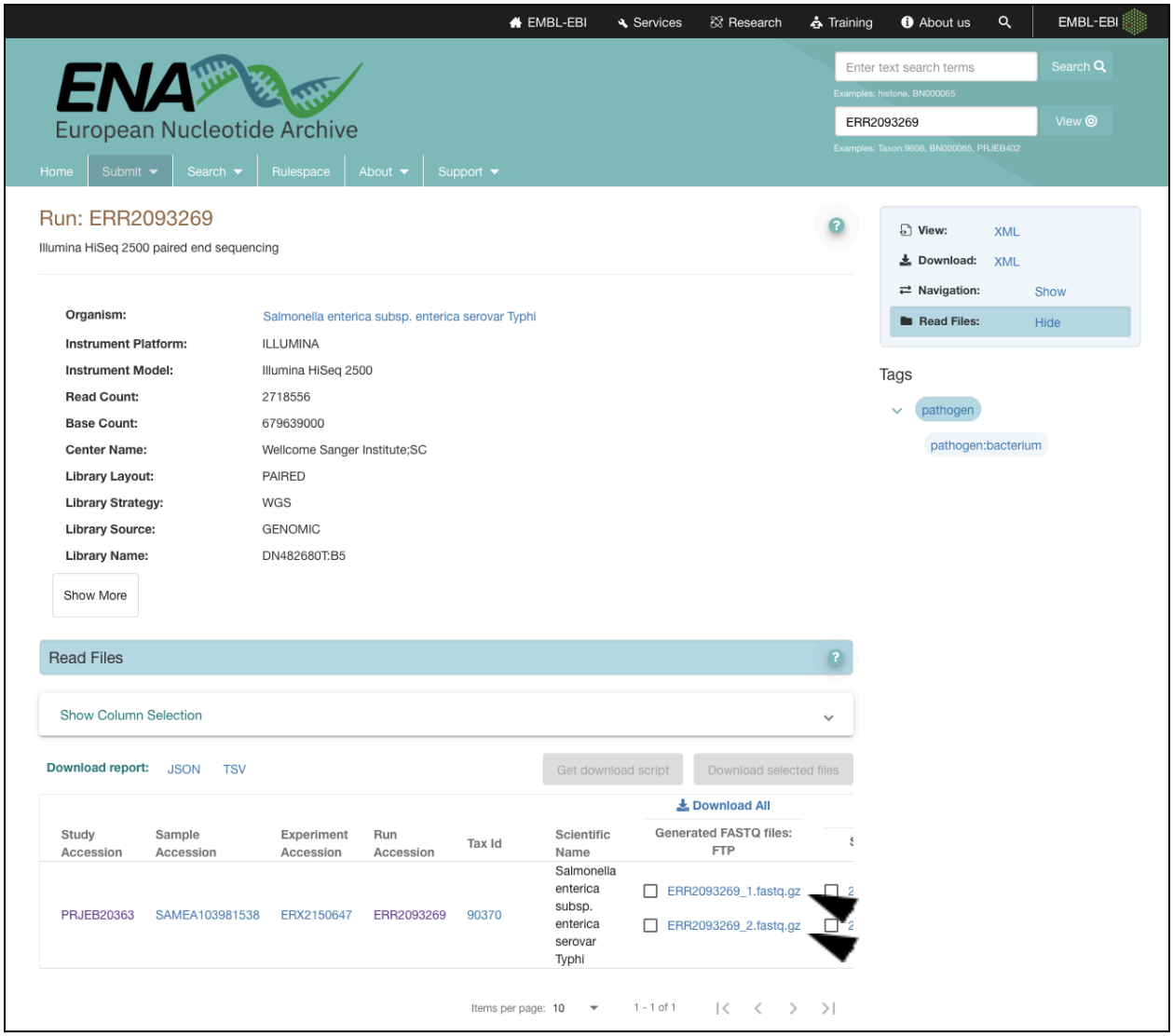


Figure 4: The information page for accession ID: ERR2093269

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

Step2: Right click on "**File1**" as pointed by the arrows in figure 4, select "**save link as**" and save the compressed FASTQ file in the folder **cp7**. Repeat the same steps for "**File 2**". Since the read files contain millions of reads and the corresponding quality scores of each base within the reads, it is larger in size. This is the reason why these are usually stored as compressed files.

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 paired-end sequencing two sequence files are generated. These are primarily called paired-end read files. In order to detect the resistance determinants a series of steps are performed which we are going to understand in the next sections.

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 (figure 5), 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.

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

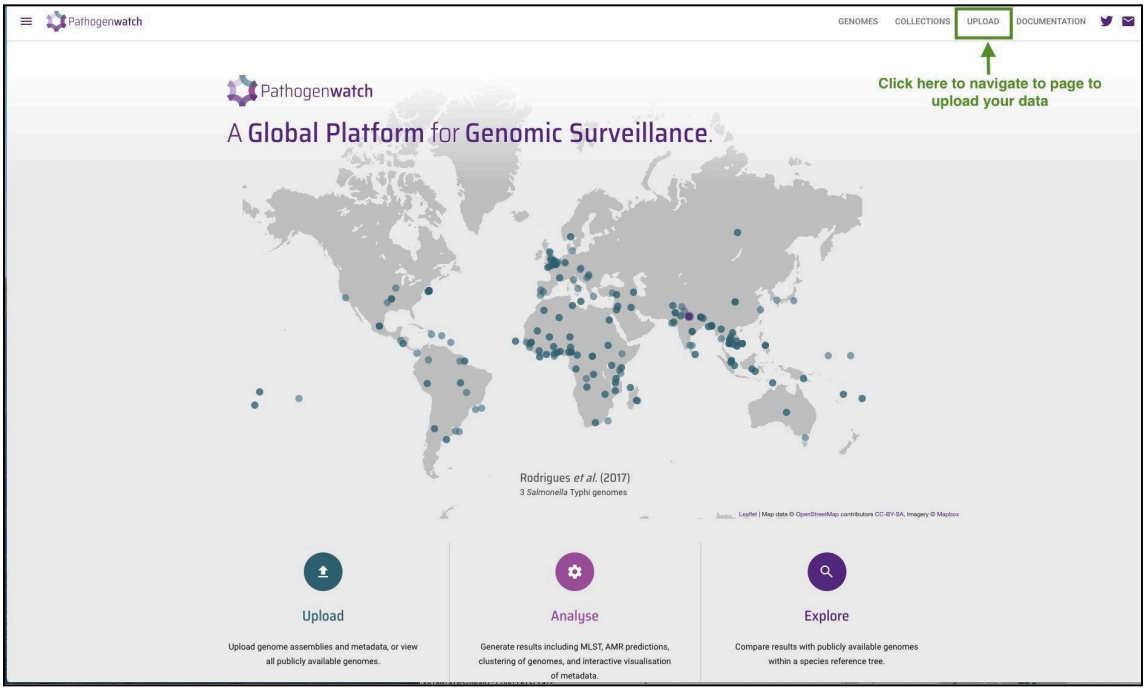


Figure 5: Pathogenwatch website

Step2: Once the upload and the analysis are completed, you will see the page display the information shown below (figure 6). Click on the "**view genomes**" link to see the page with all the information about the uploaded genome.

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

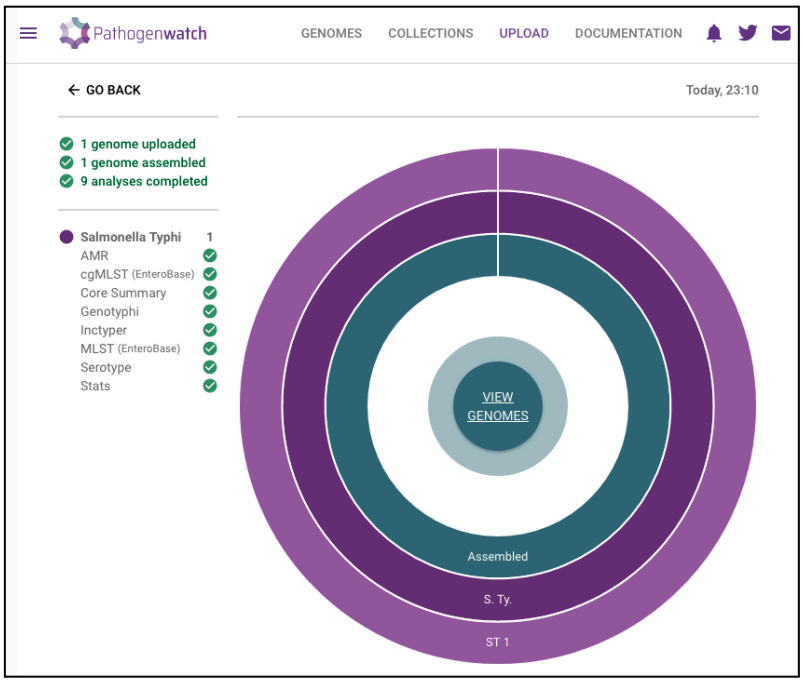


Figure 6: Pathogenwatch view after completing genome upload and processing

Step3: The resulting webpage will display information about the uploaded genome displayed in a tabular format (figure 7). To download the genome, you can select the genome (**arrow 1**), click on the selected genomes (**arrow 2**) and select "**download data**". The webpage would now resemble the figure 7 and you can download the assemblies (contigs) by clicking on "fasta files" (**arrow 3**). Other features displayed below can be downloaded for the genome in a similar manner.

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

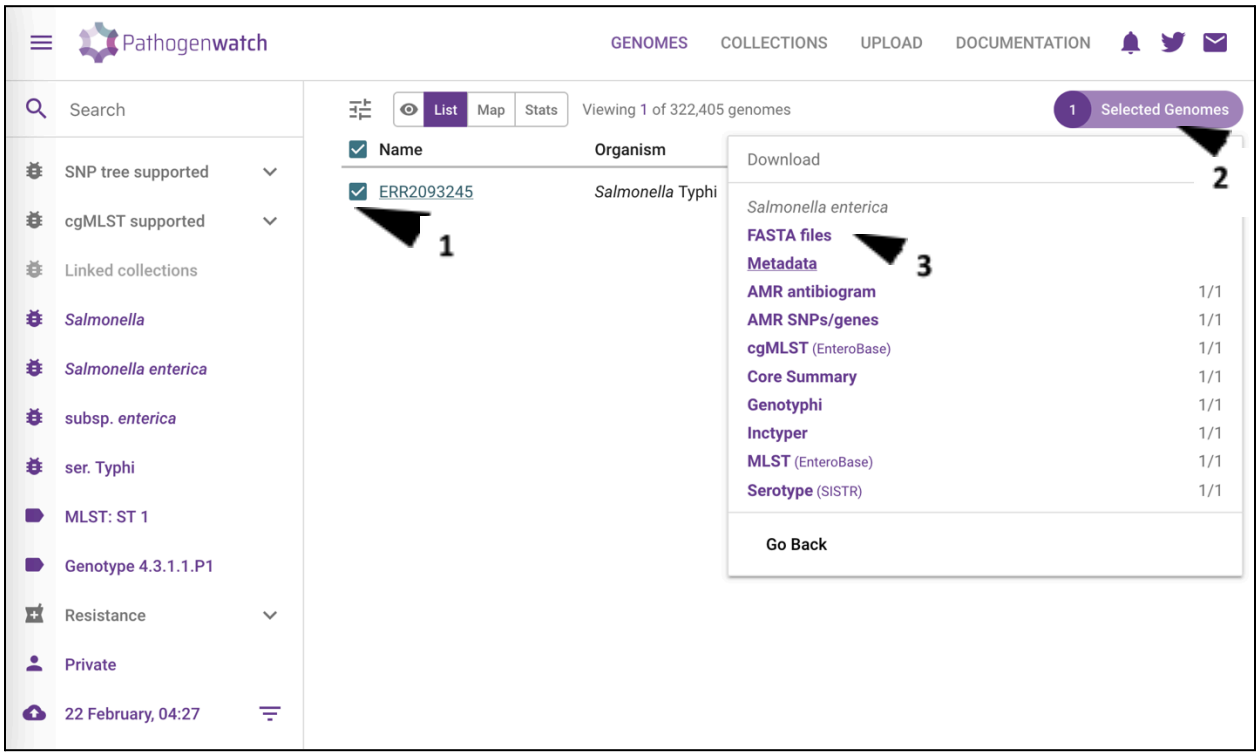


Figure 7: Downloading genome features from Pathogenwatch

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

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 8 and select “single genome fasta”.

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care’s Fleming Fund using UK aid.

Figure 8: Pathogewatch 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 (figure 9). The files are then automatically uploaded and analysed.

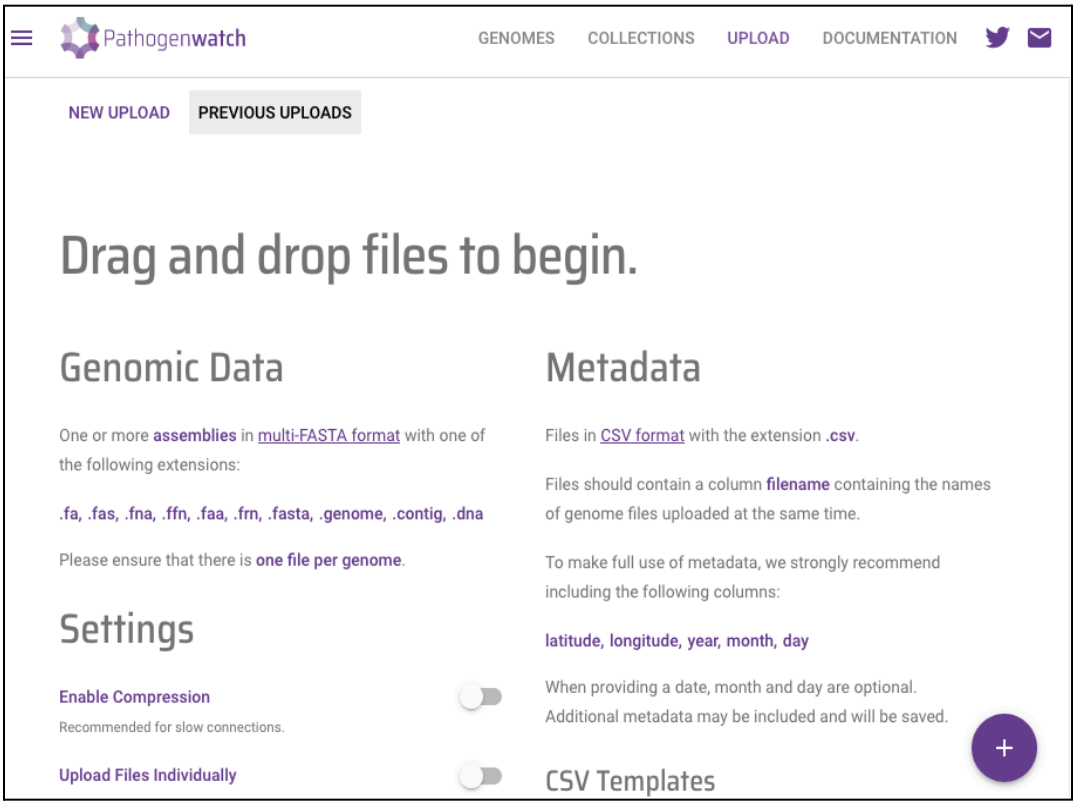


Figure 9: Uploading the assemblies to Pathogenwatch

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

Step 3: Once completed click "**view genomes**" which will open a tabular window with details on the strain (figure 10).

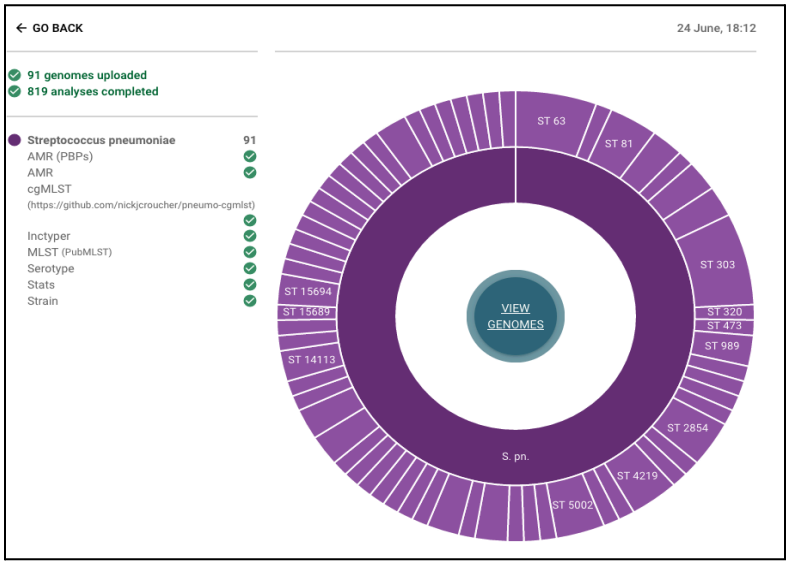


Figure 10: The status of the analysis by Pathogenwatch

Step 4: Now you will be able to see a list of all the genomes/assemblies uploaded along with associated information about the genomes. In order to see all the detailed information of the characterisation carried out on the assemblies, please click on the name/ID of that particular genome (figure 11).

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

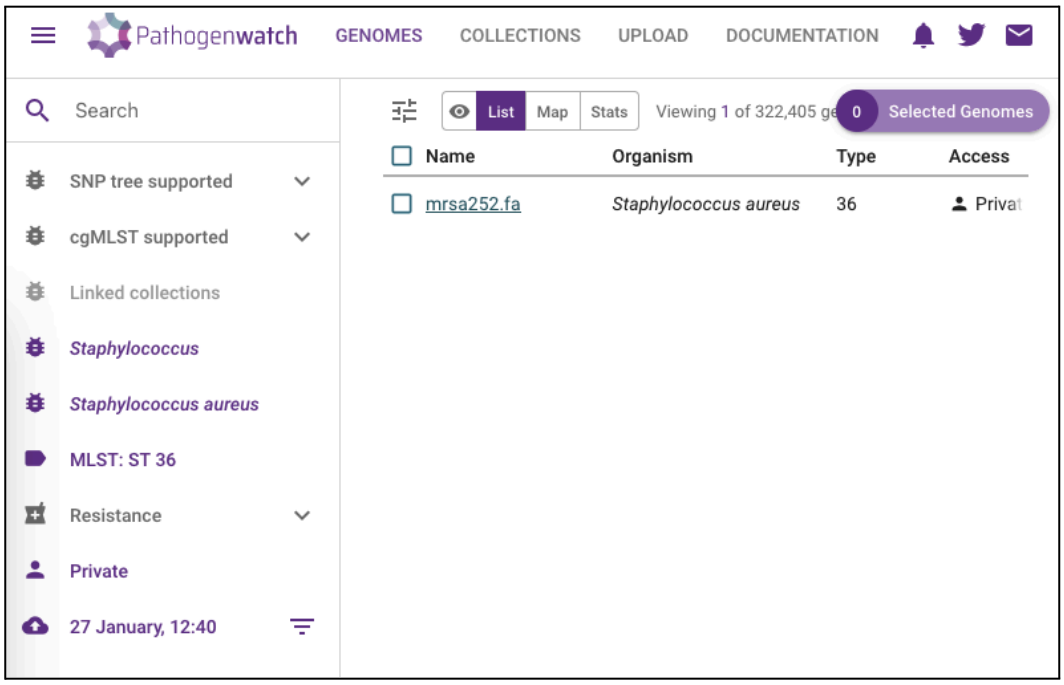
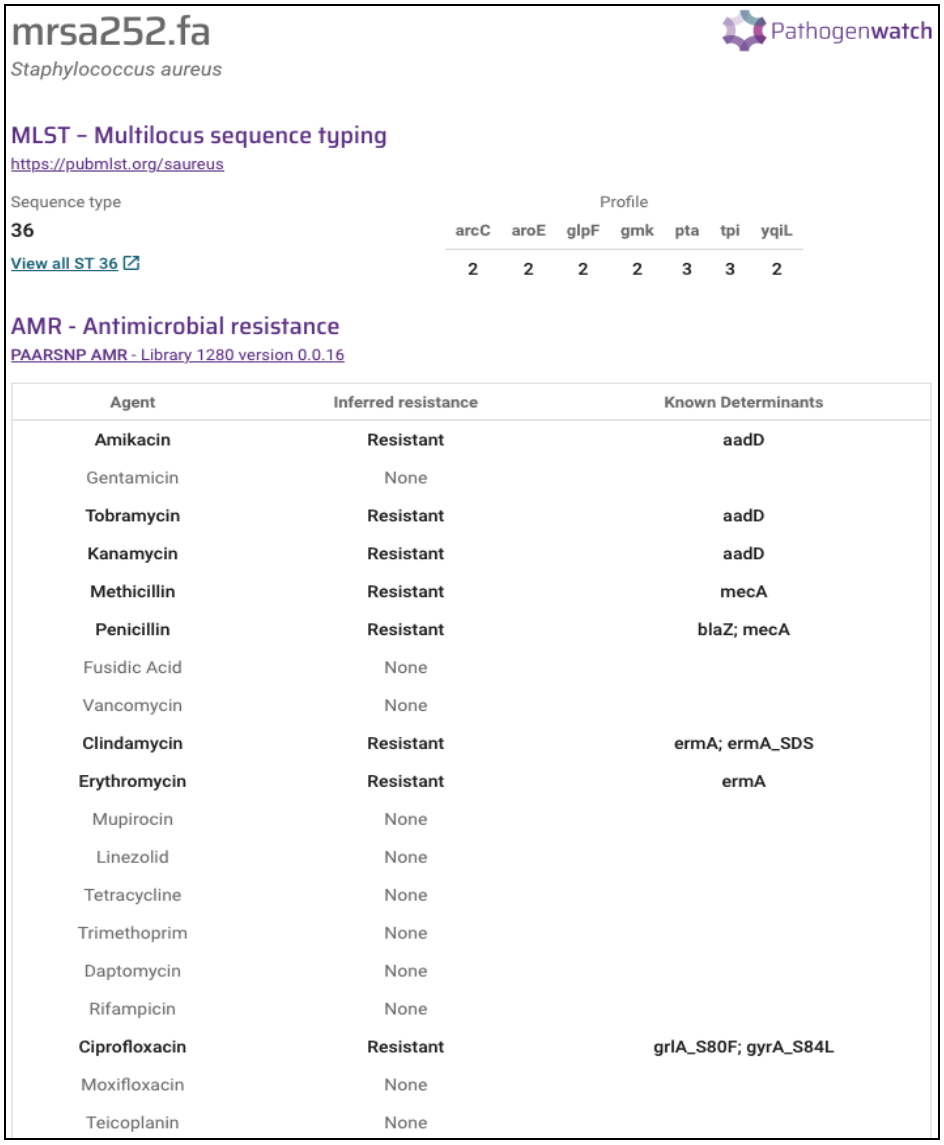


Figure 11: The tabular view of the analysed genomes by Pathogenwatch

Step 5: A sample result of prediction analysis performed by Pathogenwatch is shown in figure 12. We can see that Pathogenwatch carries out a lot of different analysis automatically to identify the species, perform MLST and detect AMR for the uploaded genome.

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.



mrsa252.fa

Staphylococcus aureus

Pathogenwatch

MLST – Multilocus sequence typing

https://pubmlst.org/saureus

Sequence type

36

View all ST 36

Profile

arcC	aroE	glpF	gmk	pta	tpi	yqiL
2	2	2	2	3	3	2

AMR - Antimicrobial resistance

PAARSNP AMR - Library 1280 version 0.0.16

Agent	Inferred resistance	Known Determinants
Amikacin	**Resistant**	**aadD**
Gentamicin	None	
Tobramycin	**Resistant**	**aadD**
Kanamycin	**Resistant**	**aadD**
Methicillin	**Resistant**	**mecA**
Penicillin	**Resistant**	**blaZ; mecA**
Fusidic Acid	None	
Vancomycin	None	
Clindamycin	**Resistant**	**ermA; ermA_SDS**
Erythromycin	**Resistant**	**ermA**
Mupirocin	None	
Linezolid	None	
Tetracycline	None	
Trimethoprim	None	
Daptomycin	None	
Rifampicin	None	
Ciprofloxacin	**Resistant**	**grlA_S80F; gyrA_S84L**
Moxifloxacin	None	
Teicoplanin	None	

Figure 12: The results of the analysis by Pathogenwatch

Here, we can see the strain name “mrsa252”, detected sequence type (ST) “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 grlA_S80F (non-synonymous mutation causing change of serine to phenylalanine at the 80th codon of the GrlA) and

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care’s Fleming Fund using UK aid.

gyrA_S84L (non-synonymous mutation causing change of serine to leucine at the84th 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	*aaD*
	Methicillin	*mecA*
	Erythromycin, Clindamycin	*ermA, ermA*_SDS
	Ciprofloxacin	*grlA*_S80F, *gyrA*_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.

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

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 "mrsa252.fa" file.

Step 1: Open the website (https://cge.food.dtu.dk/services/ResFinder/) in your web-browser. Select the chromosomal mutations and acquired antimicrobial resistance genes (figure 13, this will inform the pipeline to look for SNPs and genes in the assemblies which aren't detected by default). Then, choose the right species in our case it is "*Staphylococcus aureus*". Click on the "isolate" tab and select the file "mrsa252.fa" and click open. Click on the "upload" button to upload the data and initiate the analysis.

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

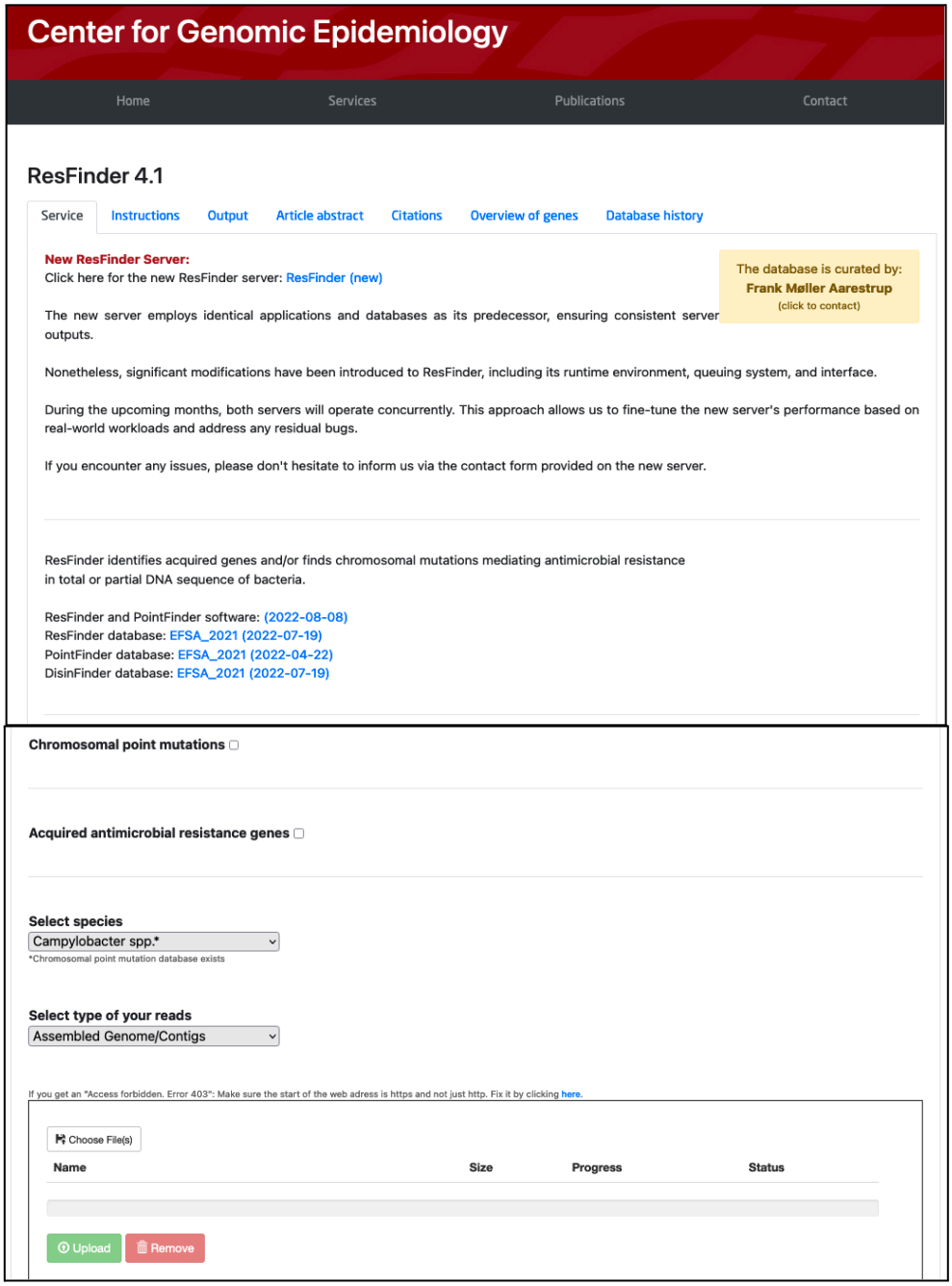


Figure 13: ResFinder web-page

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. 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

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

the strain MRSA 252 showed the presence of genes such as *blaZ, mecA, aaD, ermA* and *ant(4)Ia*. In addition, two SNPs: S80F in *grl*A and S84L in *gyr*A have been identified.

Step 3: 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	grlA_S80F, gyrA_S84L	present

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

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 14 and click on “RGI”.

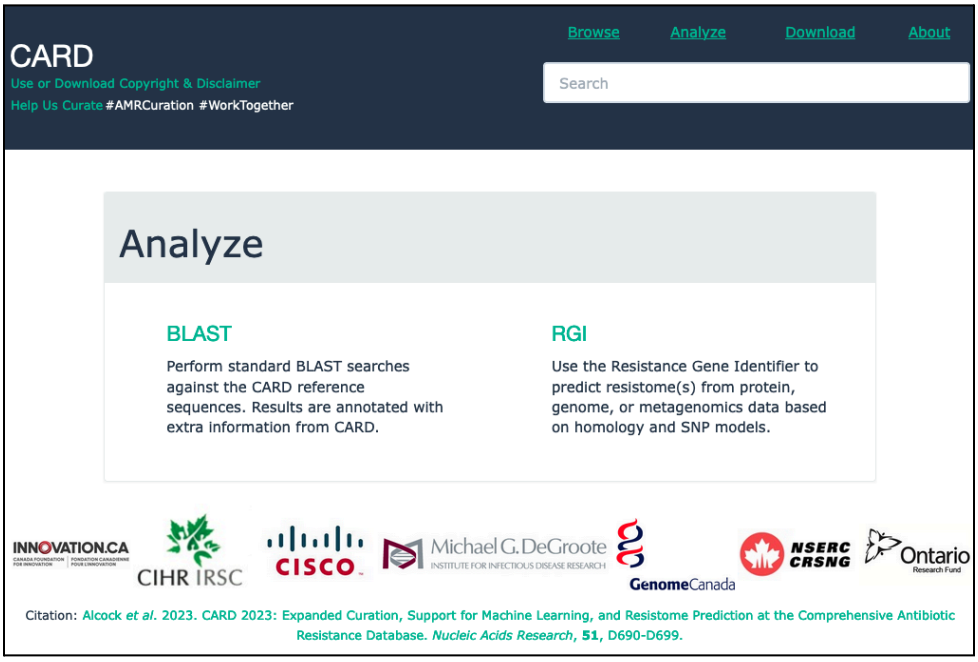


Figure 14: CARD database web server

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care’s Fleming Fund using UK aid.

Step 2: Click on the "Choose" button as indicated in figure 15, select the file "mrsa252.fa". Click "open" to upload the sequence. Upload the file and click "submit" as indicated to initiate the analysis.

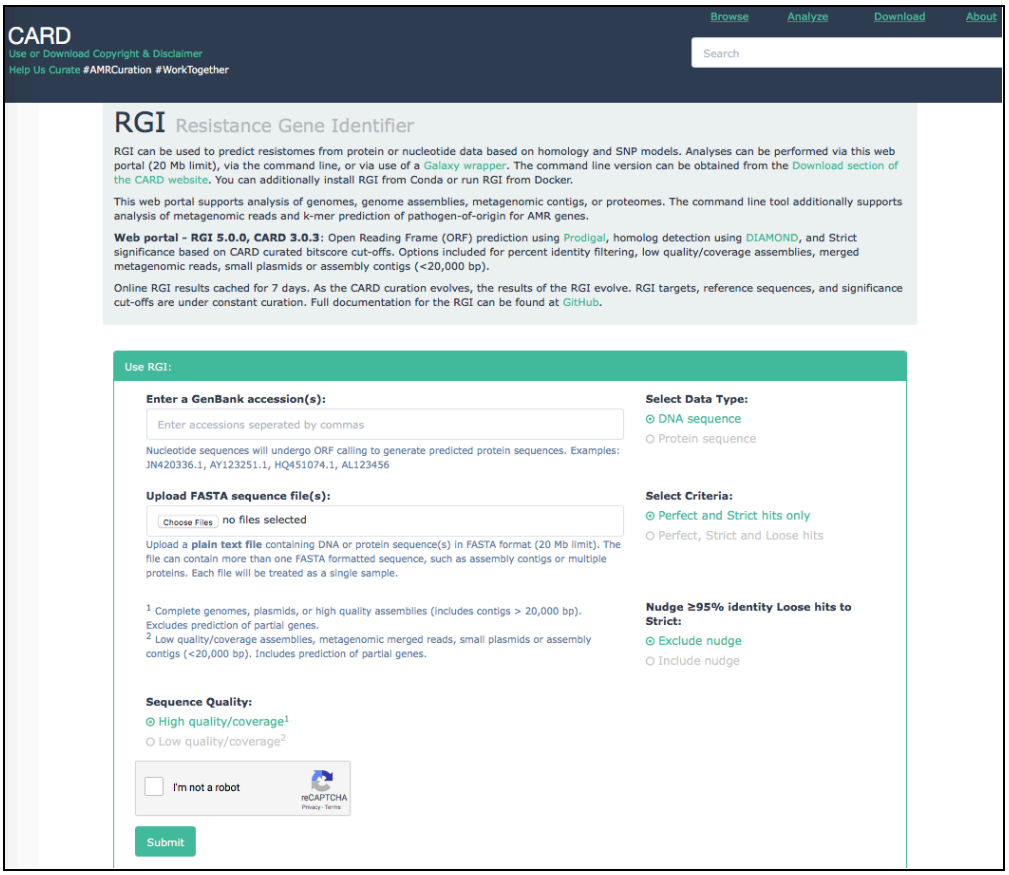


Figure 15: Uploading the assemblies to CARD database

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

Step 3: Once finished the window will appear as shown in figure 16. Since the CARD database has a large collection of genetic determinants from various bacterial species all the genetic determinants are used to mine the submitted genome without selecting for any species. Therefore, the results generated are usually a long list of determinants identified. The genes and the mutations detected are represented in the columns “ARO type” and “SNPs” respectively.

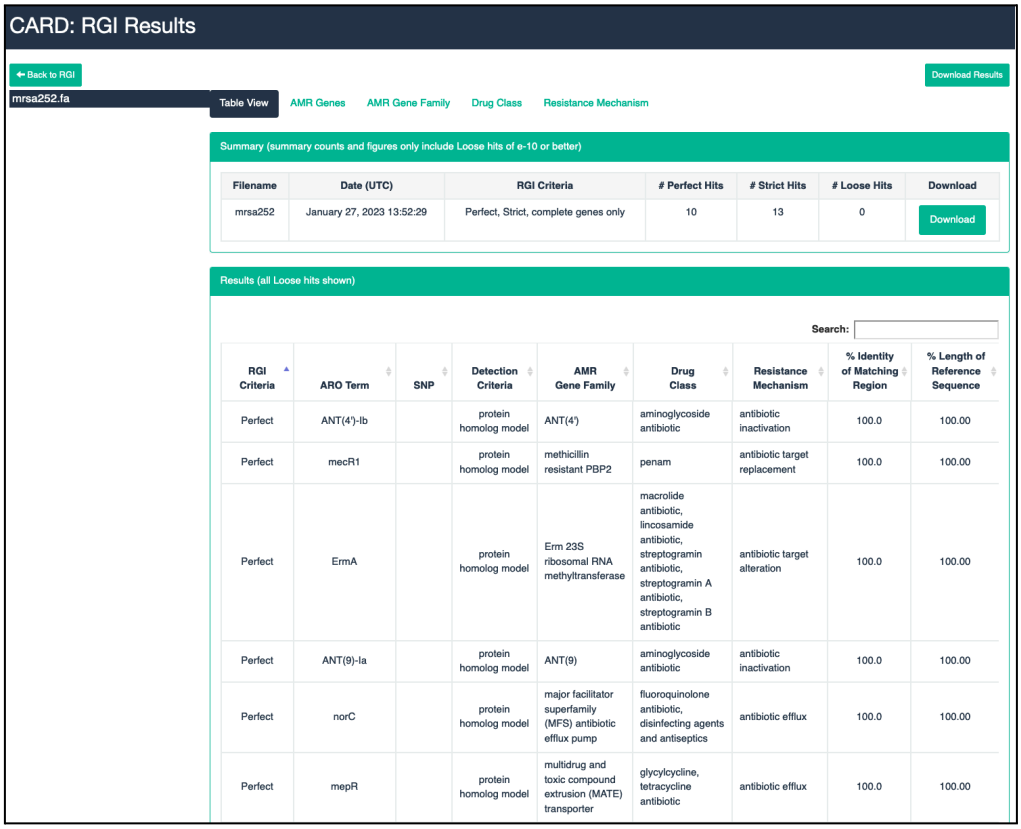


Figure 16: The prediction results of the CARD database

Step 4: Recording the prediction results:

Open the tabular file that we have already created. Record the results of prediction from CARD in a separate column as shown below. Please note that the results here don’t show the gene *aadD* instead the alternate gene name *ant(4)-Ib* is used here. This demonstrates the need for prior knowledge on the resistance mechanisms at the alternate names if possible. Another example is the gene *grlA* which is also referred to

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care’s Fleming Fund using UK aid.

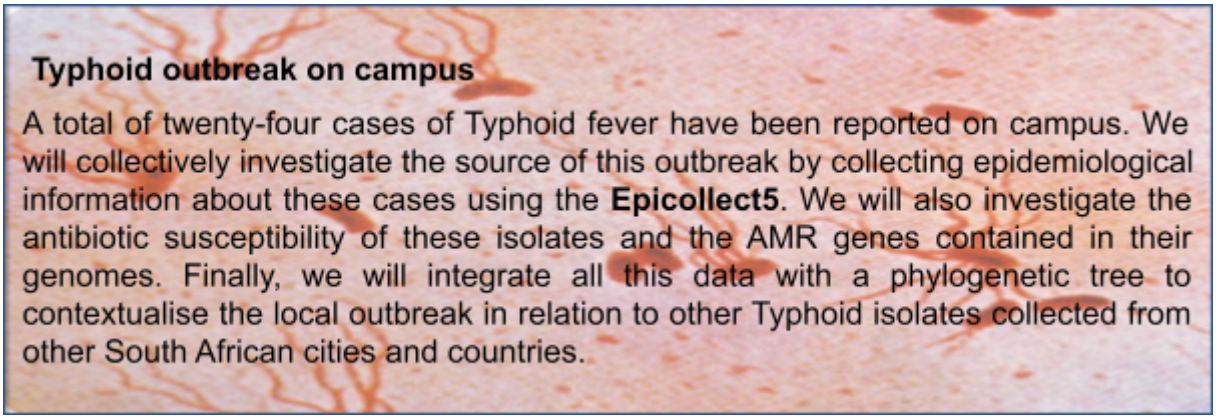
as parC in some studies. The CARD results only motions the name *parC* rather than *grlA* so one had to be careful when recording these results.

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.

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

Adding *in silico* antibiogram of *S. typhi* outbreak strains into EpiCollect



In this practical, identify the genome assembly of your assigned S. typhi isolate (e.g., ERR2093236.assembly.fa) and submit your assembly to the Online AMR detection tools used in this practical. After submitting the genomes of your assigned *S. typhi* sample, make sure the AMR genetic determinants identified by the AMR detection tools are included into Epicollect5.

Following on from the EpiCollect practical in which you collected epidemiological information about four assigned Typhoid case reported on campus, you will now need to input the AMR genes found in your *S. typhi* genome into EpiCollect. Remember that you have already done this based on the results of the command-line tools used in Computational Practical 6. Now, compare the previous results with those obtained by the Online tools used in this practical and amend the introduced results if necessary.

First, use the same mobile phone you used in the previous EpiCollect practical. Make sure WIFI is turned on and connect to the WIFI network using the credentials provided.

1. Open the Epicollect5 app on your phone and select the project named ‘Typhoid XDR outbreak’ (Figure 1)

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care’s Fleming Fund using UK aid.

Figure 1: Click on the 'Typhoid XDR outbreak' project

2. Then click on the entry you want to edit (Figure 2).

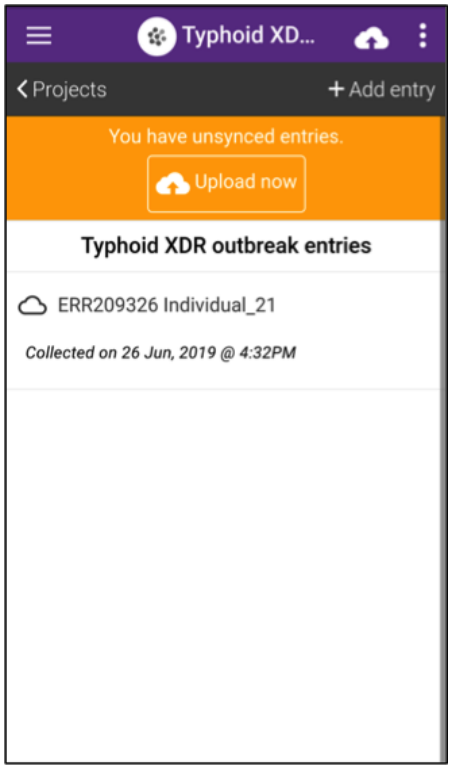


Figure 2: Click on an existing entry

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

3. Scroll down to the field 'Antibiotic resistance genes' and select the AMR genes/mutations detected in your Typhoid isolates (Figure 3 and Figure 4).

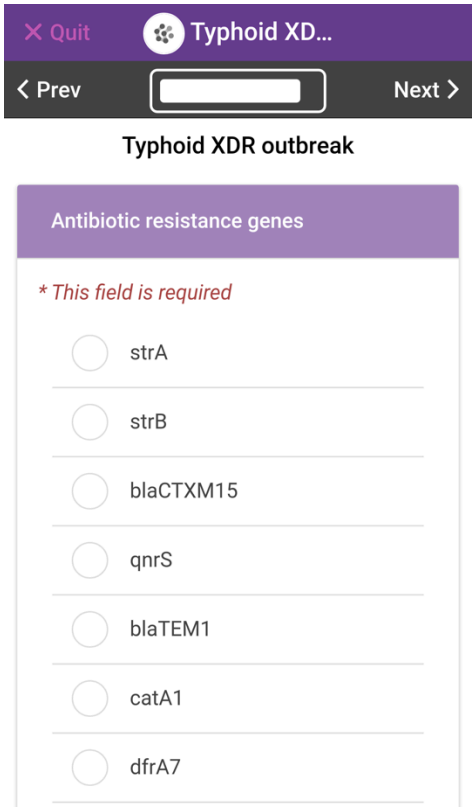


Figure 3

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.

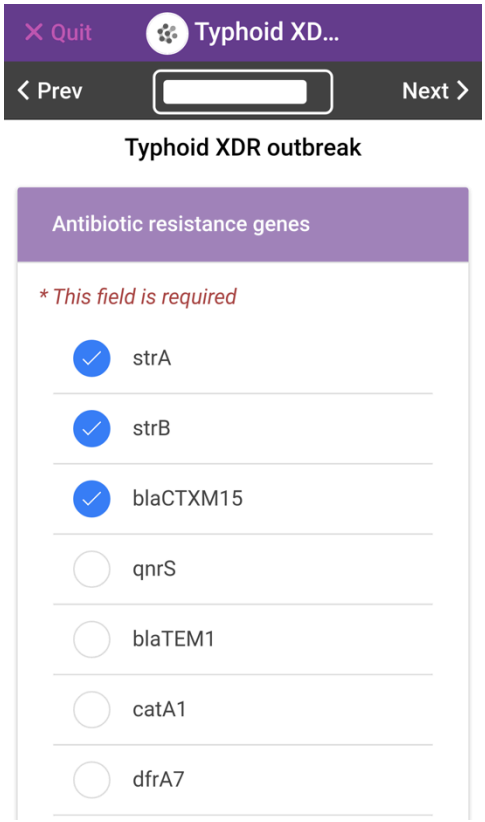


Figure 4

4. Do not forget to save changes in your entry by clicking on ‘Save entry’ at the end of the form (Figure 5).

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care’s Fleming Fund using UK aid.

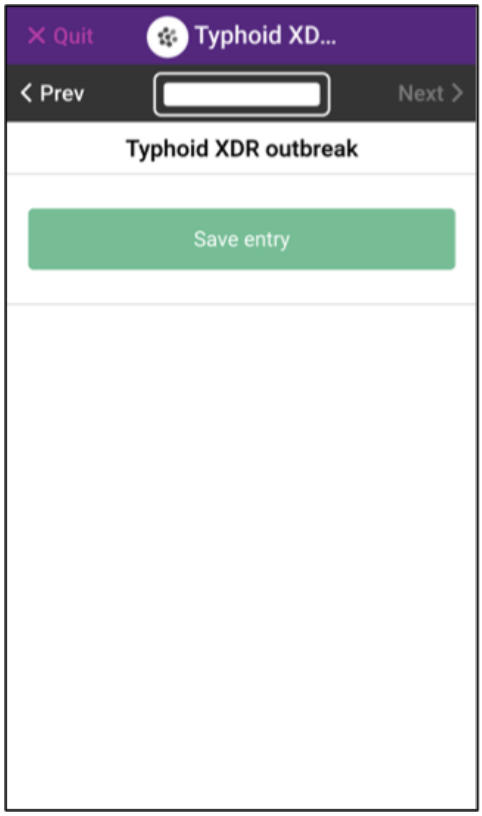


Figure 5: Save changes by clicking on 'Save entry' at the end of the form

Genomic DNA samples (*Salmonella* species isolated in South Africa) and genomic data (*Vibrio cholerae* O1 isolated in South Africa) used in the training course were made possible by support from the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid.