

Analysing AMR by submitting Data to Pathogenwatch

Learning Objectives

In this tutorial you will learn

1. How to format a CSV file for submission to Pathogenwatch
2. How to upload assemblies and metadata to Pathogenwatch
3. The basics of using Pathogenwatch
4. How to assemble a collection of genomes, predict AMR using the abricate software from [Torsten Seemann](#), and compare it to the analysis in Pathogenwatch.

Tutorial

How to format a CSV file for submission to Pathogenwatch

The CSV file is used to link metadata to assemblies. For example it can contain a column that describes the source of each sample and many columns describing the phenotypic AST result for each sample. The generic format of the CSV file can be found at this [link](#).

The essential columns are filename and displayname. In these columns enter the **exact** name of the fasta format assembly file and the name you want to use as the label for each sample respectively.

If possible it is ideal to fill in the columns called

- latitude
- longitude
- year
- month
- day

However if you do not have this information leave them blank.

After these columns add extra column heading and data depending on what metadata you would like to enter. An example of what this may look like is as below. Save this as a CSV file from your spreadsheet program.

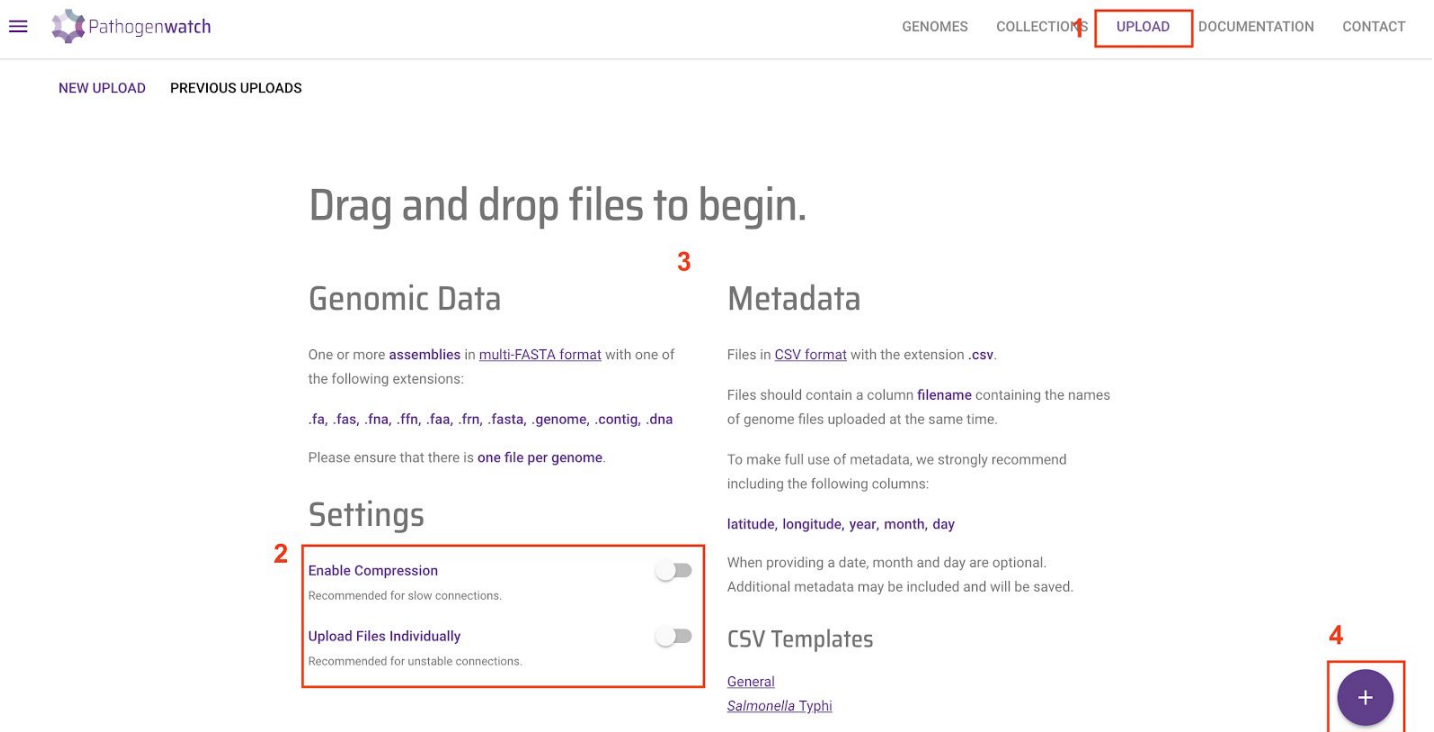
	A	B	C	D	E	F	G	H	I	J
1	filename	displayname	latitude	longitude	year	month	day	ISOLATION SOURCE	MRSA	SCCMEC TYPE
2	ERR064898_scaffolds.fasta	HSA11	41.1579438	-8.6291053	1992			bronchial secret	1	III variant
3	ERR064902_scaffolds.fasta	S38	15.2286861	104.856422	2006	11	21	blood	1	III
4	ERR064903_scaffolds.fasta	S97	15.2286861	104.856422	2007	2	26	blood	1	IIIB
5	ERR064904_scaffolds.fasta	URU110	-34.901113	-56.164531	1998			wound	1	III
6	ERR064906_scaffolds.fasta	URU34	-34.901113	-56.164531	1997			wound	1	IIIA
7	ERR064907_scaffolds.fasta	HSJ216	38.7222524	-9.1393366	1997			bronchial secret	1	IIIA

Uploading assembly files and metadata to Pathogenwatch

Once you have assembly files for each of your samples and a metadata CSV file. These can be uploaded to Pathogenwatch for analysis and visualisation.

Login into Pathogenwatch. Set up an account using Google, Twitter, Facebook or email if you do not have an account already.

1. Click on upload
2. If you have a slow or unstable internet connection turn on the relevant options
3. Drag and Drop all the assembly fasta files and the single metadata csv file onto the browser window **at the same time**. This will upload the assemblies and metadata together so that they can be paired up.
4. **Alternatively** click on the + icon and select the files from the dialog window that appears



The screenshot shows the Pathogenwatch website interface. At the top, there is a navigation bar with 'NEW UPLOAD' and 'PREVIOUS UPLOADS' on the left, and 'GENOMES', 'COLLECTIONS', 'UPLOAD', 'DOCUMENTATION', and 'CONTACT' on the right. The 'UPLOAD' button is highlighted with a red box and a red '3' above it. Below the navigation bar, the main content area is titled 'Drag and drop files to begin.' and is divided into two columns: 'Genomic Data' and 'Metadata'. The 'Genomic Data' section includes instructions on file formats (.fa, .fas, .fna, .ffn, .faa, .frn, .fasta, .genome, .contig, .dna) and a note to ensure one file per genome. The 'Metadata' section explains that files should be in CSV format with a 'filename' column and lists recommended columns: latitude, longitude, year, month, day. Below these sections is a 'Settings' area with two toggle switches: 'Enable Compression' (recommended for slow connections) and 'Upload Files Individually' (recommended for unstable connections). A red box highlights these two settings with a red '2' to its left. To the right of the 'Settings' area is a 'CSV Templates' section with a '+' icon in a blue circle, highlighted with a red box and a red '4' above it. The 'General' template for 'Salmonella Typhi' is visible below the icon.

The basics of using Pathogenwatch

The most complete functionality for Pathogenwatch is available for those organisms that are covered by the AMR and collection functions. These are currently *Neisseria gonorrhoeae*, *Staphylococcus aureus*, and *Salmonella Typhi*. More Enterobacteriaceae to follow soon.

Let's take some *Staphylococcus aureus* samples as an example.

After upload the sample analysis will proceed. Once this has finished click on View Genomes

The screenshot shows the Pathogenwatch interface. At the top, there are navigation links: GENOMES, COLLECTIONS, UPLOAD, DOCUMENTATION, CONTACT. Below the navigation, there are tabs for 'NEW UPLOAD' and 'PREVIOUS UPLOADS', with a timestamp 'Uploaded: 19/10/2018, 12:07:26'. The main content area is divided into 'Progress' and 'Analysis'. Under 'Progress', it says 'Analysis complete'. Under 'Organisms', it lists 'Staphylococcus aureus: 7' with sub-items: AMR ✓, Core ✓, cgMLST ✓, Metrics ✓, MLST ✓. To the right is a donut chart representing the analysis. The chart has two concentric rings. The outer ring is purple and labeled 'ST 130'. The inner ring is a lighter purple and labeled 'S. au.'. In the center of the chart is a circular button labeled 'VIEW GENOMES'.

Select all the samples using the checkbox in the top left and then on the selected genomes button and finally on the create collection button

The screenshot shows the Pathogenwatch interface with a list of 7 genomes. The top navigation links are GENOMES, COLLECTIONS, UPLOAD, DOCUMENTATION, CONTACT. On the left is a search and filter sidebar. The main area shows 'Viewing 7 of 20823 genomes'. A table lists the genomes with checkboxes in the 'Name' column selected. A '7 Selected Genomes' button is highlighted in the top right. A 'CREATE COLLECTION' button is highlighted in the bottom right.

Name	Organism	ST
<input checked="" type="checkbox"/> Cow_A	<i>Staphylococcus aureus</i>	130
<input checked="" type="checkbox"/> Patient_A2	<i>Staphylococcus aureus</i>	130
<input checked="" type="checkbox"/> Patient_B	<i>Staphylococcus aureus</i>	130
<input checked="" type="checkbox"/> Sheep_B2	<i>Staphylococcus aureus</i>	130
<input checked="" type="checkbox"/> Sheep_B1	<i>Staphylococcus aureus</i>	130
<input checked="" type="checkbox"/> Sheep_B3	<i>Staphylococcus aureus</i>	130
<input checked="" type="checkbox"/> Patient_A1	<i>Staphylococcus aureus</i>	130

Give your collection a name and click on create collection

Create Collection

7 Genomes *Staphylococcus aureus*

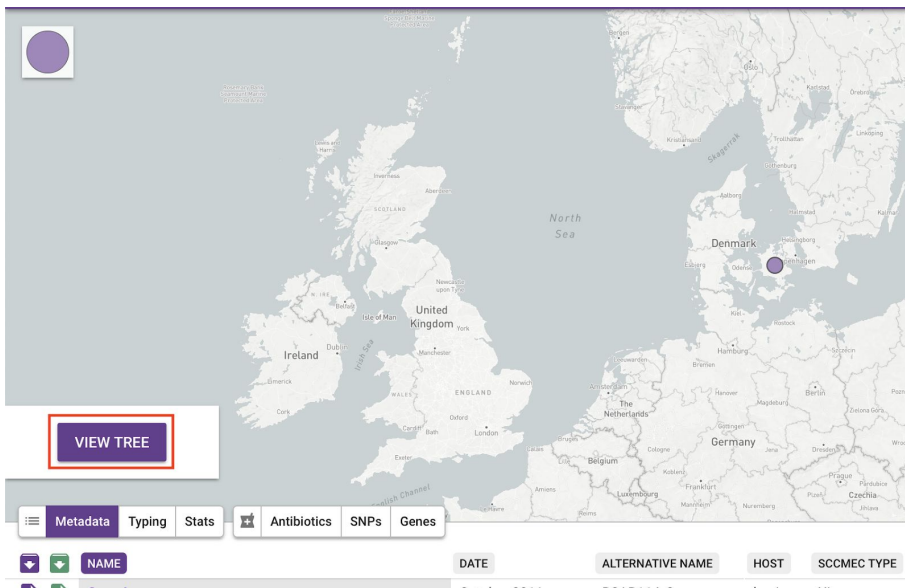
Title
My Collection

Description

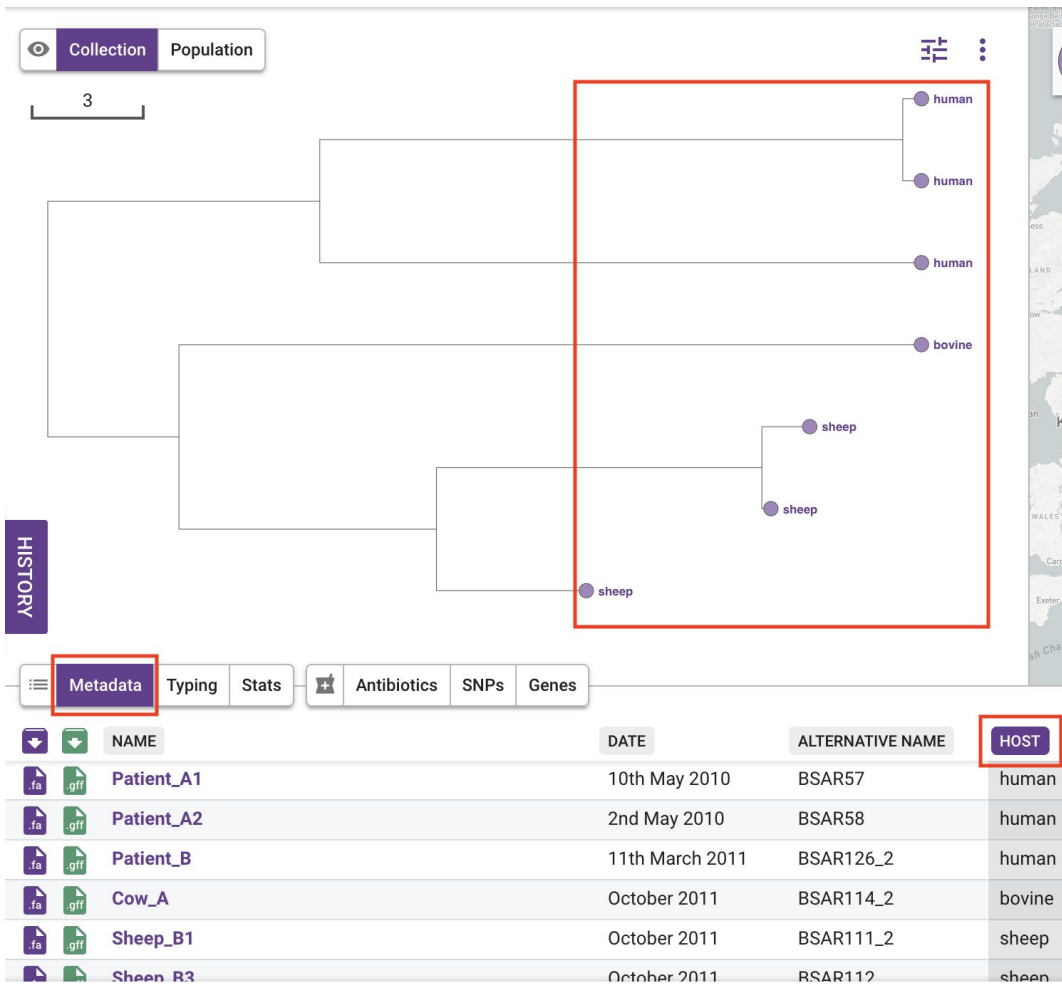
PMID

GO BACK CREATE NOW

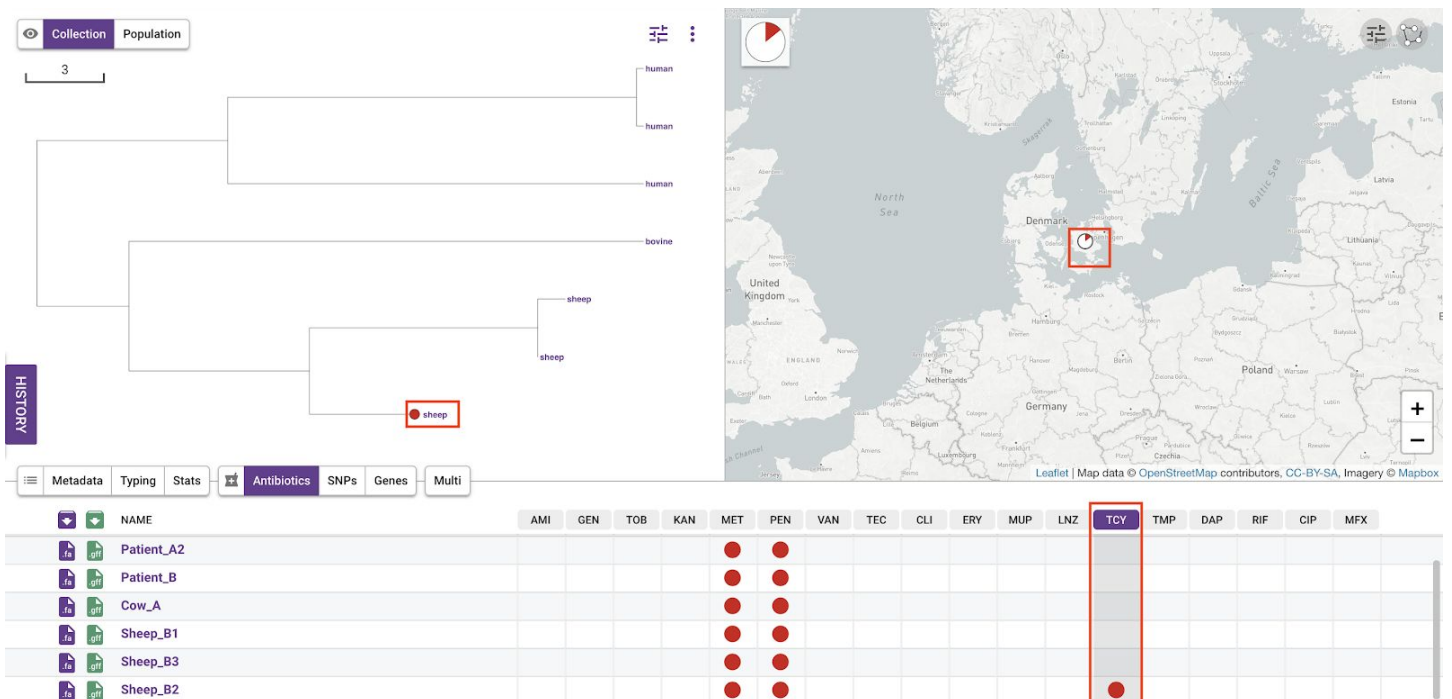
A tree will be built, taking a maximum of a minute or two. Click on the View Tree button once finished



The tree labels can be changed by clicking on a column on the metadata table, having selected Metadata, Typing or Stats on the left of the bottom panel

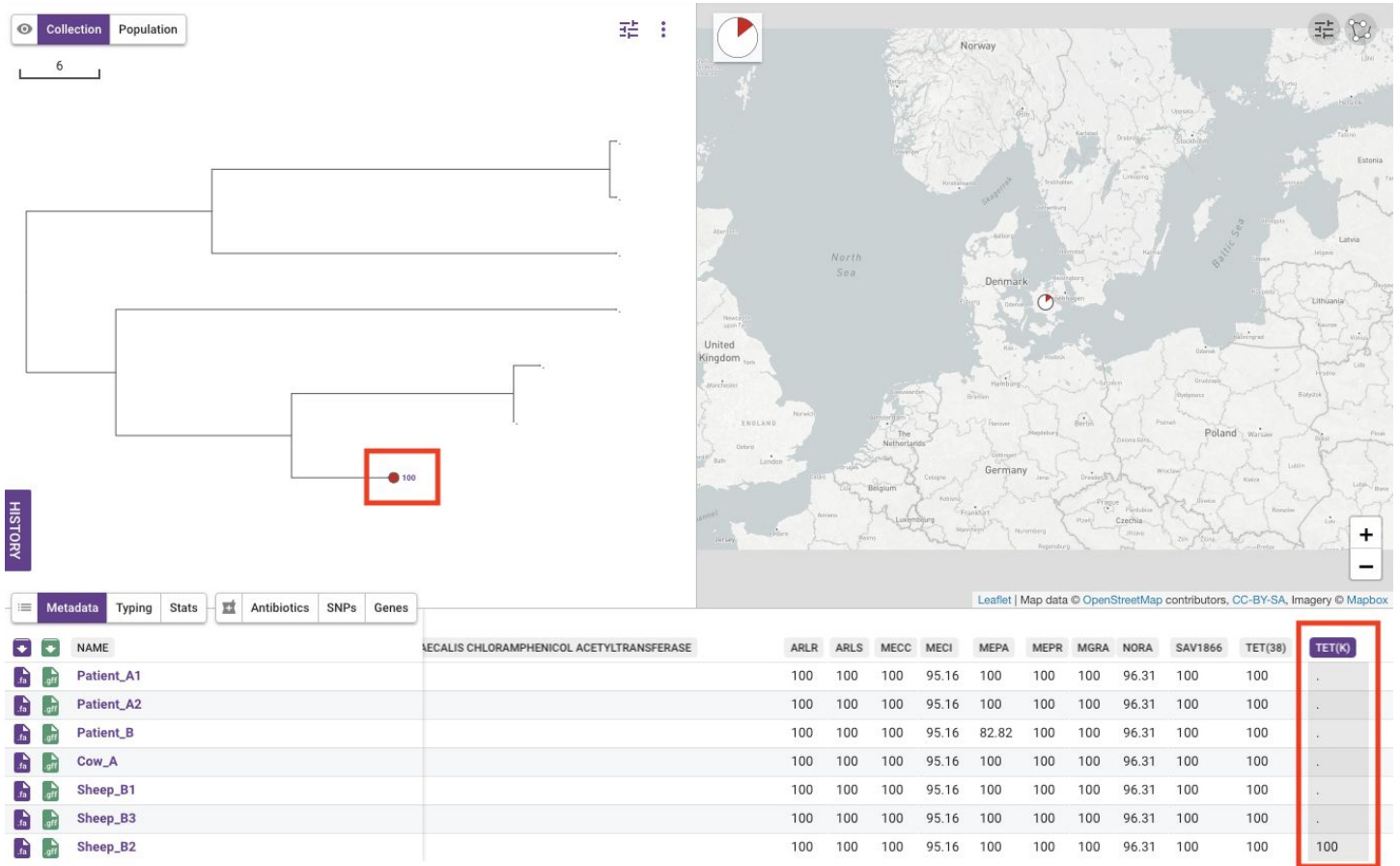


The tree shape colours can be changed by clicking on a column in Antibiotics, SNPs or Genes on the right of the bottom panel



These can be combined to explore the data. For example in the screenshot below TCY was selected in

Antibiotics and a column called TET(K) was selected in Metadata that was the result of an abricate run on the data



This only really touches on the basics of Pathogenwatch, for a much more comprehensive documentation visit the [help pages](#) at the Pathogenwatch website

Assembling genomes, AMR prediction and pathogen watch analysis

In order to follow the process through from start to finish, the next section of the tutorial will guide you through assembling genomes on two *Staphylococcus aureus* genome sets and prediction of AMR using abricate, followed by uploading to Pathogenwatch.

Torok et al 2014

The first data set is from this paper: <https://www.ncbi.nlm.nih.gov/pubmed/24788657>

To obtain the fastq files download from this [google drive folder](#) and then upload them to a directory you have made for this purpose on your server. This will test how feasible it is to upload data from a sequencing run to an online cloud server.

If you find that the upload is too slow then there is an option to download these files directly as part of the assembly process.

Assembly

To run assembly on these genomes use the Nextflow assembly pipeline where the general format of the command is

```
nextflow run /path/to/assembly.nf --input_dir /path/to/fastq_data
--fastq_pattern '*_{1,2}.fastq.gz' --adapter_file /path/to/adapters.fas
--output_dir /path/to/output_dir -with-docker bioinformant/ghru-assembly:1.1
--resume
```

As part of the training exercise to gauge if you have an understanding of Unix file paths, fill in the appropriate file paths.

If you have been unable to upload the fastq files due to a slow/unstable internet connection adjust the command as follows

```
nextflow run /path/to/assembly.nf --accession_number_file
/path/to/accessions.txt --adapter_file /path/to/adapters.fas --output_dir
/path/to/output_dir -with-docker bioinformant/ghru-assembly:1.1 --resume
```

The accessions.txt file can be found [here](#). It is a simple text file with a SRA/ENA accession for each sample on a new line.

AMR prediction using abricate

Once assembly has finished run the [abricate software](#) to predict which AMR-related genes are present in the genomes.

First you will need to ensure that you have the nextflow workflow file and the software dependencies for abricate.

The workflow files are easiest downloaded from the GHRU software repository by making a special location for them on the server and downloading them as follows

```
wget https://gitlab.com/cgps/ghru/pipelines/abricate/-/archive/master/abricate-master.zip
unzip abricate-master.zip
```

To ensure that you have the docker image required to supply the software dependencies you will need to type the command

```
docker pull <DOCKER IMAGE NAME>
```

In the case of the abricate docker image this would be

```
docker pull bioinformant/ghru-abricate:1.0
```

Now to run abricate, use the following command

```
nextflow run /path/to/abricate.nf --input_dir /path/to/assembled_scaffolds
--output_dir /path/to/output_dir --fasta_pattern *.fasta --database card
--with-docker bioinformant/ghru-abricate:1.0
```

As before substitute the paths specific to your set up. The assembled scaffolds directory will an output from the previous assembly nextflow.

Abricate can use one of several databases that store genes related to AMR. In the example above the card database has been specified. Try this and another e.g ncbi

The output will be written as a csv file in the format

```
abricate_summary_<DATABASE NAME>.tsv
```

Once you have this file, combine it with metadata for the samples. This can be found [here](#)

Following the guidelines in the section above on the CSV file for pathogen watch take the data in this file, combine it with the assembly filenames and add additional columns for the AMR genes found in the abricate output. This will require careful sorting of spreadsheets. I suggest using the filenames and ENA RUN headings. At the end of this exercise you should have columns for filename, displayname, location, date and additional columns including those found in the metadata file and in the abricate output.

Once you have the final combined metadata file upload this and the assembled scaffold files to Pathogenwatch. Once uploaded explore the data and see what conclusions you would draw from the data. Read the paper and explore the data on Pathogenwatch to see how they match.

Harris et al 2010

A second data set is taken from a subset of the samples from this paper

<http://www.ncbi.nlm.nih.gov/pubmed/20093474>

The fastqs can be found at this [link](#)

If required an accessions file can be found at this [link](#)

The metadata file to modify and combine with the abricate output can be found at this [link](#)

Follow the same procedure as above to assemble these file and predict AMR genes using abricate. Combine the CSV files and upload to Pathogenwatch along with the assembly files.

Again explore the data in conjunction with the paper.

Congratulations

By completing this tutorial you will have demonstrated the capability to assemble a large batch of fastqs, predict AMR and upload these to Pathogenwatch.

Next steps

If you have local data sets from any of the following species (*Neisseria gonorrhoeae*, *Staphylococcus aureus*, and *Salmonella Typhi*) you will can follow the same procedure to obtain full Pathogenwatch functionality.

If you have datasets from other bacterial datasets, assemble them and predict AMR using abricate and upload to Pathogen and observe the functionality that is available to you and provide feedback on what features are missing.

It is likely that you may require additional functionality for some datasets to be able to display genetic relatedness. In that case [Microreact](#) may provide the best tool whilst more species are added to Pathogenwatch. For Microreact a phylogenetic tree is required in addition to a metadata. Creating this will be the subject of the next tutorial!

Version Control Table

Title	Analysing AMR by submitting Data to Pathogenwatch			
Description	A document describing how to upload assemblies to Pathogenwatch and analyse the resulting data			
Created By	Anthony Underwood			
Date Created	19th October 2018			
Maintained By	Anthony Underwood			
Version Number	Modified By	Modifications Made	Date Modified	Status
1.0	Anthony Underwood	Draft version	19th October 2018	Draft Version