1 Genomic analysis of Salmonella Typhimurium

For this analysis project, you will use the all the skills acquired from earlier in the week to map the sequnce data of 25 isolates of Salmonella Typhimurium to the reference sequence strain construct a phylogenetic tree. You will use the software FigTree to view and interpret your phylogeny and also use Microreact to visualise location data and develop your own hypotheses about the pathogen distribution. In addition, you will use the software ARIBA (Antibiotic Resistance Identification By Assembly) to investigate resistance genes in these strains and a free visualisation tool, Phandango to compare with the pathogen distribution observed in your tree. You will also prepare a short 5 mins presentation describing the analysis approach and your key findings.

2 Learning outcomes:

On completion of the assignment, you can expect to have acquired the skills to:

- Build and interpret phylogenetic trees from Whole Genome Sequence (WGS) data of bacterial samples
- Show how WGS can be used to describe the evolution and distribution of a pathogen across a geographical area
- Combine phylogenetic analysis and AMR results
- Present the results of WGS data analysis

3 Introducing the dataset

Salmonella enterica is a diverse bacterial species that can cause disease in both human and animals. Human infections caused by Salmonella can be divided into two, typhoidal Salmonella or non-typhoidal Salmonella (NTS). The former include Typhi and Paratyphi serovars that cause typhoid. NTS comprises of multiple serovars that cause self-limiting gastroenteritis in humans and is normally associated with zoonotic Salmonella reservoirs, typically domesticated animals, with little or no sustained human-to-human transmission.

Salmonella enterica serovar Typhimurium (S. Typhimurium), unlike the classical views of NTS, can cause an invasive form of NTS (iNTS), with distinct clinical representations to typhoid and gastroenteritis and normally characterized by a nonspecific fever that can be indistinguishable from malaria and in rare cases is accompanied by diarrhoea.

You have been provided with sequence data for 25 S. Typhimurium isolates collected from regional labs in England and Wales throughout 2022. Whole genome sequence analysis of this organism can provide some insight into the short-term microevolution of S. Typhimurium. Understanding the level of diversity in this time-period is crucial in attempting to identify if this is an outbreak or sporadic infection.

To download the data that you need for this asignment:

[ ]:
```
cd ~/course_data/group_projects/data/project1
git pull
```

This will download the following files:

- A reference genome (FASTA) for S. Typhimurium SL1344 found in the `ref` directory.
- Sequence data (FASTQ files) for 6 of the 25 samples. These are in the `fastq` folder.
- A `pseudogenomes` directory containing pseudogenomes for 19 of the 26 samples, you will need to generate pseudogenomes for the remaining isolates.
- A metadata table (`metadata.xls`) which contains information on the isolates including the date and location of collection and to save time we have prepared a `metadata_microreact.csv` file compatible with Microreact.
- An `ariba_results` folder that contains the resistance reports from ARIBA for 19 of the 26 samples. You will need to run ARIBA on the remaining samples and combine the reports into a summary report.
- A summary of the phenotypic data (`phenotype_data.csv`) from the lab.

The download may take some time so read ahead and make a plan for your analysis.

4 Tasks

You will need to generate a whole-genome sequencing based tree for the isolates and correlate this to the geography of the country. You will also need to look into the distribution of antimicrobial resistance and investigate the genetic basis for the resistance phenotype identified in the laboratory.

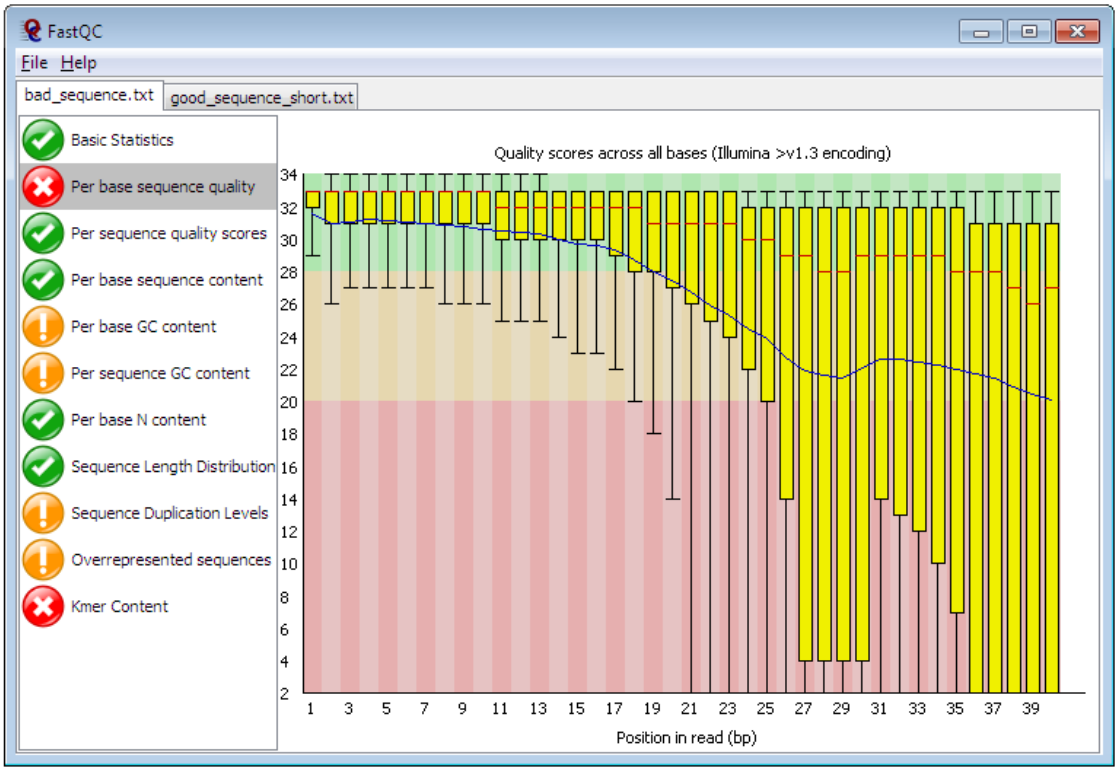
Therfore, you will need to complete the following tasks:

- QC - QC your sequence data (30 mins)
- Phylogenetic analysis - SNP-calling and phylogenetic inference (1hr)
- Antimicrobial resistance screening - ARIBA and Phandango (1hr)
- Data visualisation with Microreact (30 mins)
- Preparation of presentation (30 mins)
- Group presentations (30 mins)

There is a lot to do here, so you may want to divide up the tasks among the group. For example, you could split into smaller groups and assign subsets of the data to each sub-group or assign different tasks to each subgroup. The choice is yours, just ensure that each of the sub-groups continues to communicate with each other during the task!

4.1 Task 1: Sequence QC

In this task you will QC the sequence data for all 25 *Salmonella typhimurium* isolates.



4.1.1 Analysis Steps

Bioinformatic processing of data into biologically-meaningful outputs is often a multi stage process. Just like working in the laboratory, it's useful to break this process down into individual steps and have a plan before you start the analysis.

- **Step 1**. Set up a directory for the QC analysis
- **Step 2**. Use FastQC to QC the sequence data for each isolate
- **Step 3**. Gather together all the FastQC reports into a single QC report
- **Step 4**. Run bactinspector on all the isolates to determine the species
- **Step 5**. Assess the QC results and remove any isolates that do not pass your QC threshold from further analysis
- **Step 6**. Summarise your findings (text and screenshots) for inclusion in your presentation

4.1.2 Software Tools

The recommended software to use in this task are:

- FastQC
- MultiQC
- Bactinspector

There is a conda environment called `qc` that contains these tools. To activate it use:

```
conda activate qc
```

4.1.3 Questions

The questions you want to ask of the data are:

- What is the yield and sequencing depth for each isolate and is it adequate?
- Is the sequencing quality adequate for each of the isolates?
- Are all the isolates the species that you expect?

4.2 Task 2: Phylogenetic analysis

In this task you will map the sequence data to the *Salmonella enterica Typhimurium SL1344* reference genome, call SNPs and build a phylogenetic tree from the SNP data.

4.2.1 Analysis Steps

A rough guide of the steps involved in this task is below. Check that you understand the principles of each one and then get started:

- **Step 1**. Set up a directory for the phylogenetic analysis
- **Step 2**. Map, call SNPs and create pseudogenomes for each isolate
- **Step 3**. Create a whole genome sequence alignment fusing all the pseudogenomes
- **Step 4**. Identify the SNPs in the pseudogenome alignment
- **Step 5**. Build a phylogenetic tree
- **Step 6**. Root the tree
- **Step 7**. Interpret your phylogeny
- **Step 8**. Summarise your findings (text and screenshots) for inclusion in your presentation

4.2.2 Software Tools

The recommended software to use in this task are:

- Bactmap nextflow pipeline
- assembly-stats
- snp-sites
- iqtree
- FigTree

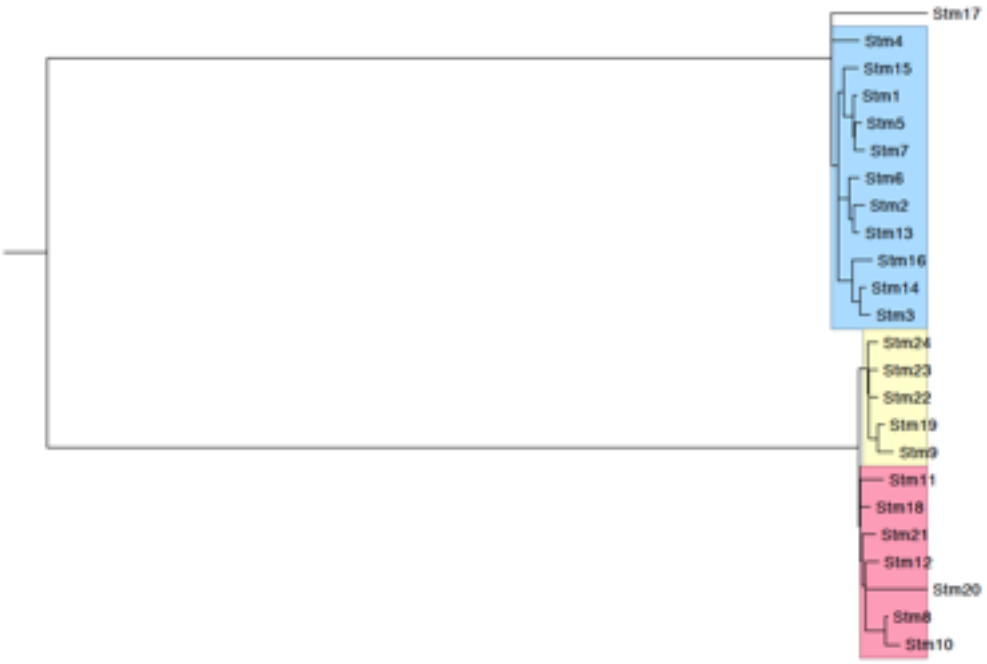
The conda environments that contain the required software are `nextflow-pipelines` and `snp-phylogeny`.

4.2.3 Questions

Look at your phylogenetic tree in Figtree, take some time to make some general observations:

- Are there distinct clades present in the isolates?
- Are there isolates that do not cluster with other isolates?

A picture of your tree as well as your general observations about it should go into your group presentation. Take some time to make figure(s) you are happy with and create a PDF picture file by selecting `File > export PDF`.



4.3 Task 3: Antimicrobial resistance screening

In this task you will screen your isolates for antimicrobial resistance (AMR) genes using ARIBA and the CARD database.

4.3.1 Analysis Steps

- **Step 1.** Set up a directory for the AMR analysis
- **Step 2.** Download and prepare the CARD database for use with ARIBA
- **Step 3.** Run ARIBA on your isolates
- **Step 4.** Summarise all the ARIBA results into one report
- **Step 5.** Visualise the outputs in Phandango
- **Step 6.** Compare resistance genes found with your phenotypic data
- **Step 7.** Summarise your findings (text and screenshots) for inclusion in your presentation

4.3.2 Software Tools

The recommended software to use in this task are:

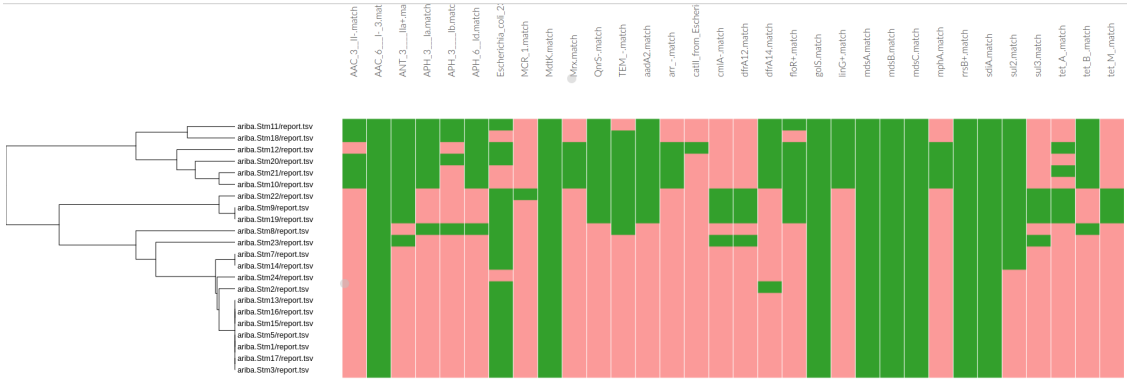
- CARD database
- ARIBA
- Phandango

The conda environment that contain the required software is `ariba-2.14.6`

4.3.3 Questions

Open the `all_results.summary.phandango.tre` and `all_results.summary.phandango.csv` produced by ARIBA in Phandago. On the left hand side is a dendogram of the phylogenetic relationship

of the resistance data and the isolates. On the top panel are the matching resistance genes found. The green colour indicates positive match and salmon pink is a negative match.



Consult the CARD database for the resistance phenotype of the genes detected. Note that underscores (_) in the output data denotes prime (‘) or bracket, therefore AAC_3_-II is AAC(3)-II. The codes for these are in a file named `01.filter.check_metadata.tsv` produced when you prepared your database. Consult the `report.tsv` of the particular sample of interest for the gene names. You can open both .tsv files in excel.

Some general points to consider when summarizing your finding for antimicrobial resistance (AMR) screen.

- Does the presence of the gene correlate well with the phenotypic results?
- Is it the same in multiple isolates that share the resistance?

To help you in your discussions, summarized in the table below are the most frequently observed antimicrobial resistance patterns of S. Typhimurium strains between 1996 and 2016 taken from Wang et al., 2019 (PMID: 311340204).

Below are genes found in your* isolates that confer resistance to some of the antibiotics below. Can you match isolates to any of the patterns listed? Ampicillin (A): aac, aph. Chloramphenicol/Florfenicol (C): cmlA, floR. Streptmycin (S): aadA2. Sulfonamide (Su): sul2 and sul3. Tetracycline (T): tet.

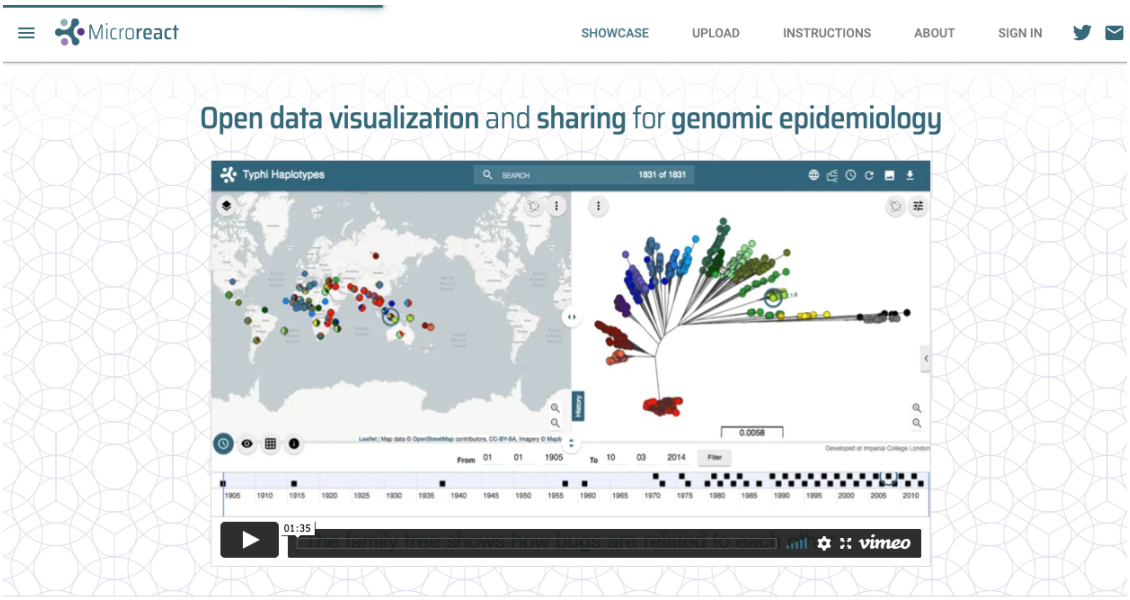
*Multiple genes can confer resistance to the same class of antimicrobials.

Resistance Patterns	Resistance phenotype
ASSuT	Ampicillin, Streptomycin, Sulfonamides, and Tetracycline
ACSSuT	Ampicillin, Streptomycin, Sulfonamides, and Tetracycline
ACSSuTAmc	Ampicillin, Chloramphenicol, Streptomycin, Sulfonamides, Tetracycline, and Amoxicillin-clavulanic acid
ACSSuTAmcAxoTio	Ampicillin, Chloramphenicol, Streptomycin, Sulfonamides, Tetracycline, Amoxicillin-clavulanic acid, Ceftriaxone, and Ceftiofur

Letter abbreviations and resistance class:
A: Ampicillin, C: Chloramphenicol, S: Streptomycin, Su: Sulfonamides, T: Tetracycline, Amc: Amoxicillin-clavulanic acid, Axo: Ceftriaxone, Tio: Ceftiofur.

4.4 Task 4: Data visualisation with Microreact

Use Microreact to visualize and explore your tree and metadata. Microreact enables you to visualize phylogenetic relationships of isolates linked to geographic locations. You can also display other information you find useful and to do so, you need only format your metadata table. To save time we have prepared a metadata table compatible with microreact and it is called `microreact_metadata.csv`. It contains latitude and longitude values for location, date of collection and some selected amr results from the ARIBA analysis. The different metadata fields have also been assigned a colour code that will be interpreted by Microreact. Take a look at this file.



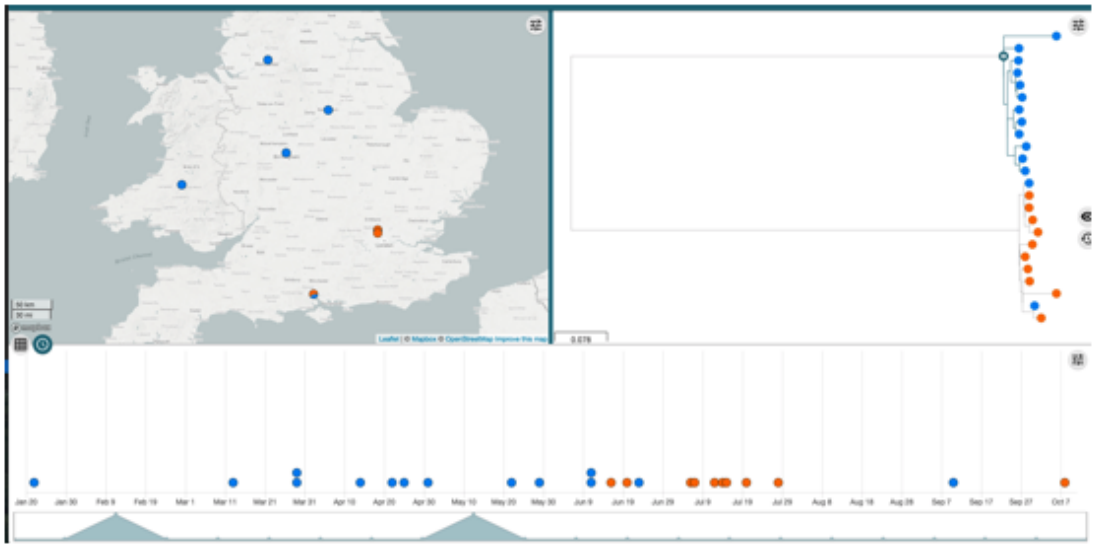
4.4.1 Location information

There are plenty of ways to generate latitude and longitude coordinates (e.g. using a simple search on google) for locations, however with multiple samples, it is easier to submit this in batches. You

can use a toll called Data-flo, which allows you to paste the name your locations, returns a list of geographic coordinates. To save time, we have already generated these coordinates for the locations and included them in the metadata file.

4.4.2 Creating a project in Microreact

You will need the NEWICK (.nwk) file from your phylogenetic analysis and the .csv metadata file mentioned above. Create a new project in Microreact by dragging and dropping them into the browser on the Microreact page. Create a map by selecting the edit icon in the right hand corner and selecting 'Create New Map'.



The resulting map and tree enables you to query your data. Notice we have assigned colour to the samples resistant or sensitive to aminoglycosides encoded by aadA2, but you can choose to group the samples however you want. For example, you can include other antimicrobial resistance information. All you will have to do is add the resistance gene data from the ariba results.

4.5 Task 5: Summarise your findings for the presentation

Create a short 5 mins presentation summarising your findings. A suggested outline for the presemtations is:

- Title slide (include your group name and individual names)
- Aims of the project
- Summary of the data
- Methods for each task (how you analysed your data)
- Results from each task (use lots of screenshots!)
- Findings and conclusions

Some things to consider when you are interpreting your results, look at the distribution of your isolates across the country when coloured by:

4.5.1 Clade

- How many clades are there?
- How are the isolates distributed?
- Are there any patterns you can see to the distribution?
- What factors might be driving the distribution?

4.5.2 Antimicrobial resistances

- Are there patterns to any of the resistances?
- And is this related to clades?
- Are there any genes that confer resistance to antimicrobials used to treat Salmonella cases?

Decide among the group who will present each slide, if possible we would like to see everyone from the group present at least one slide each.

DISCLAIMER: All the locations and dates of the Salmonella isolates are fictitious and solely for educational purposes.