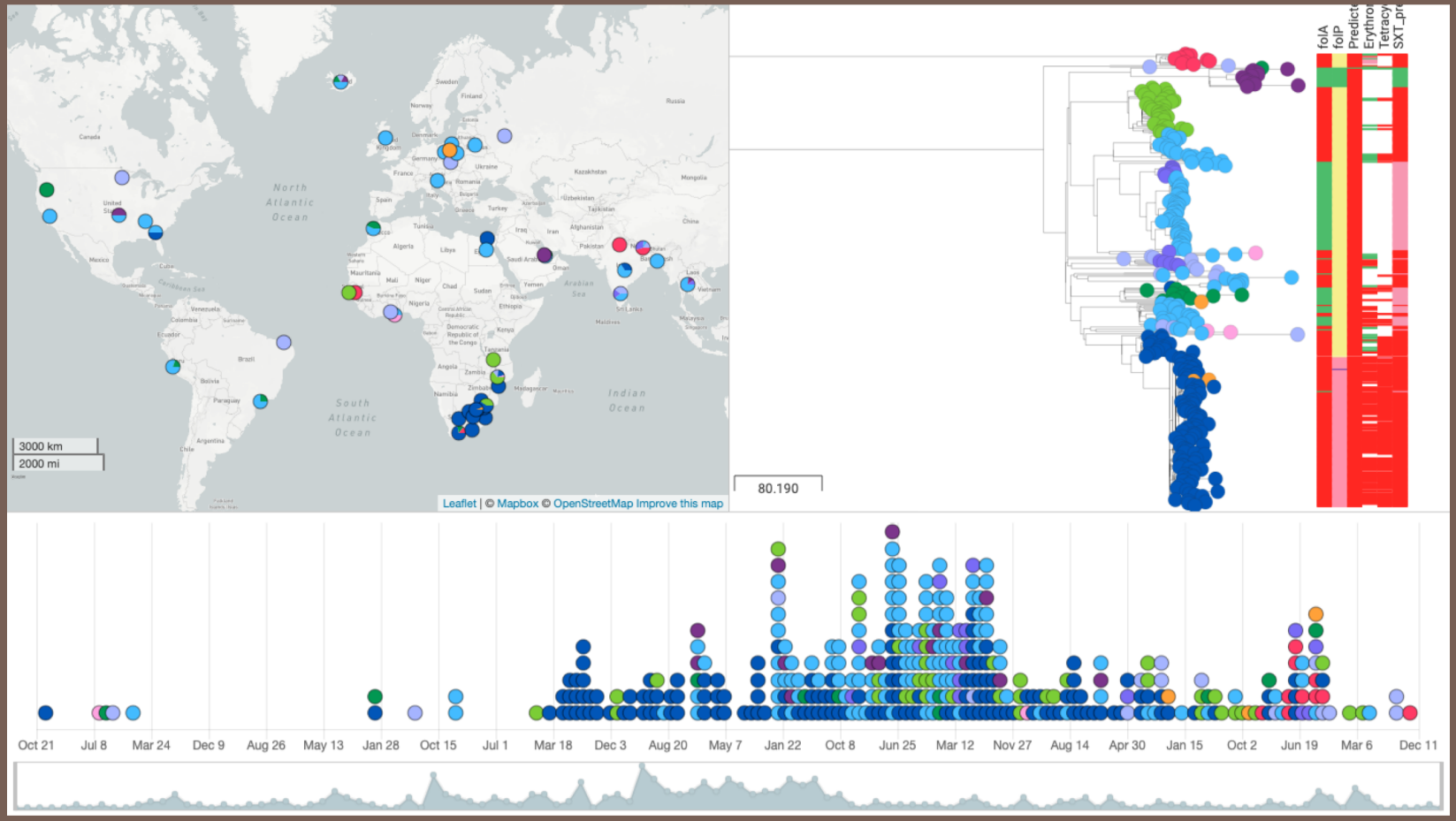


TASK 2: GEO-REFERENCEING

Christine Boinett

Objectives

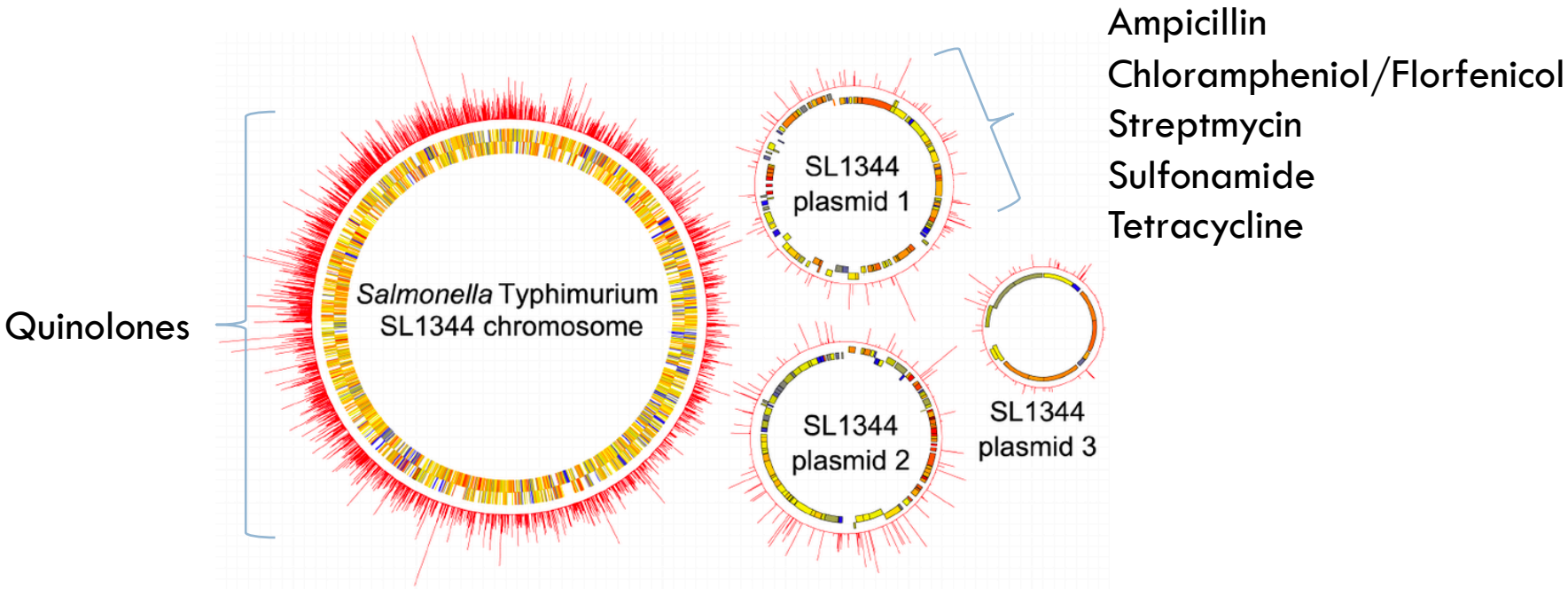
- Experience building and interpreting phylogenetic trees.
- Screen for genetic determinants using curated databases.
- Linking genotypic and phenotypic (metadata) information using geo-referencing tools.
- Gain experiencing presenting sequencing data with biological relevance.

Task 2: Background

- *Salmonella enterica* serovar Typhimurium
  - Enteric pathogen
  - Most common causes of food-borne illness
  - Transmission: faecal-oral route
  - Causes diarrhoea, fever, and stomach cramps
  - Symptoms usually begin six hours to six days after infection and last four to seven days
  - Food-borne pathogens usually emerge in outbreaks
  - Treatment:
    - Most people recover within 4-7 days
    - Antibiotic treatment recommended in severe cases, immunocompromised people, infants <12 months, adults >65
  - More information at https://www.cdc.gov/salmonella/

Salmonella Typhimurium – SL1344

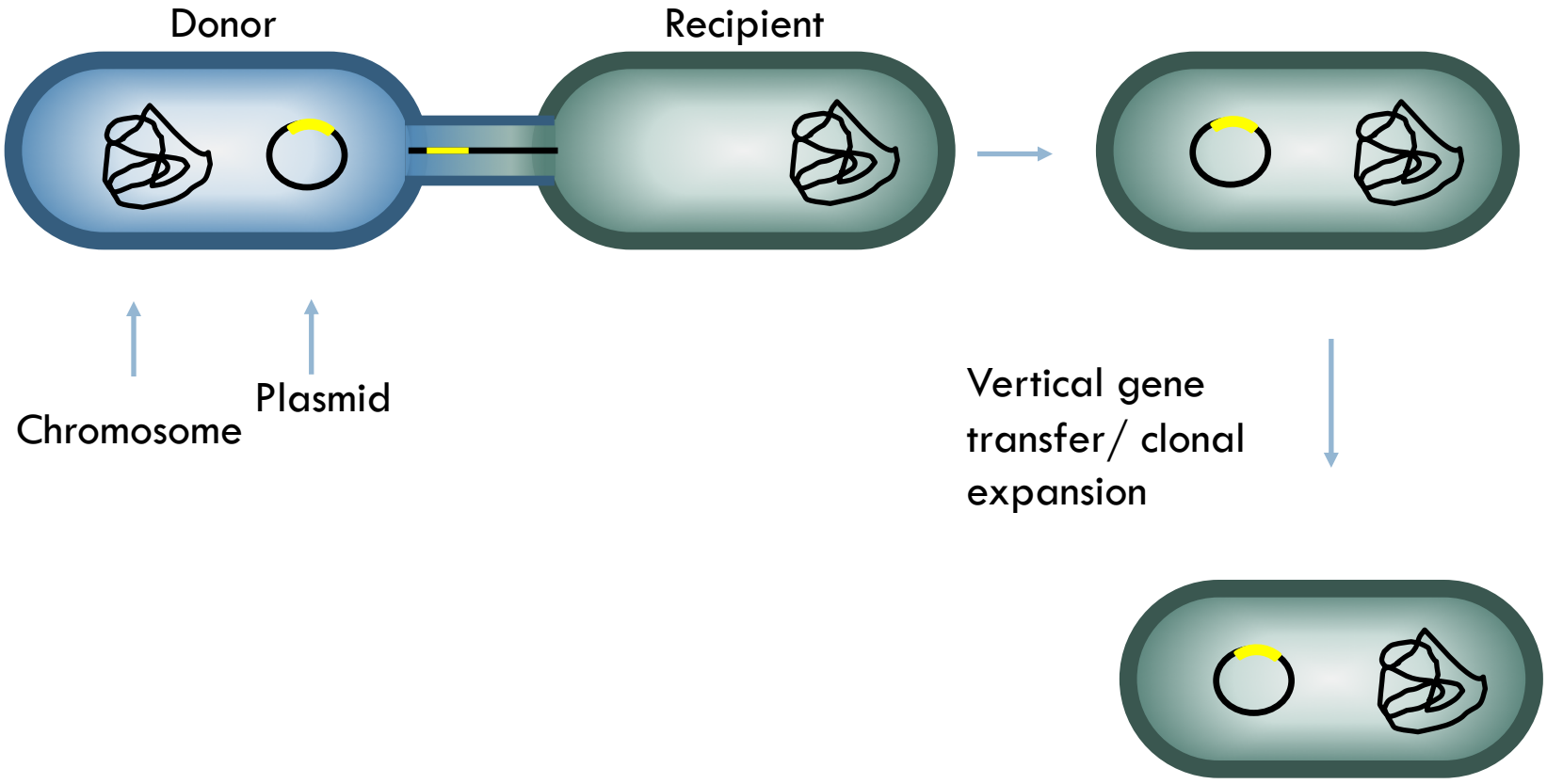
- Antimicrobial (AMR) an virulence resistance genes can be encoded on plasmids and the chromosome



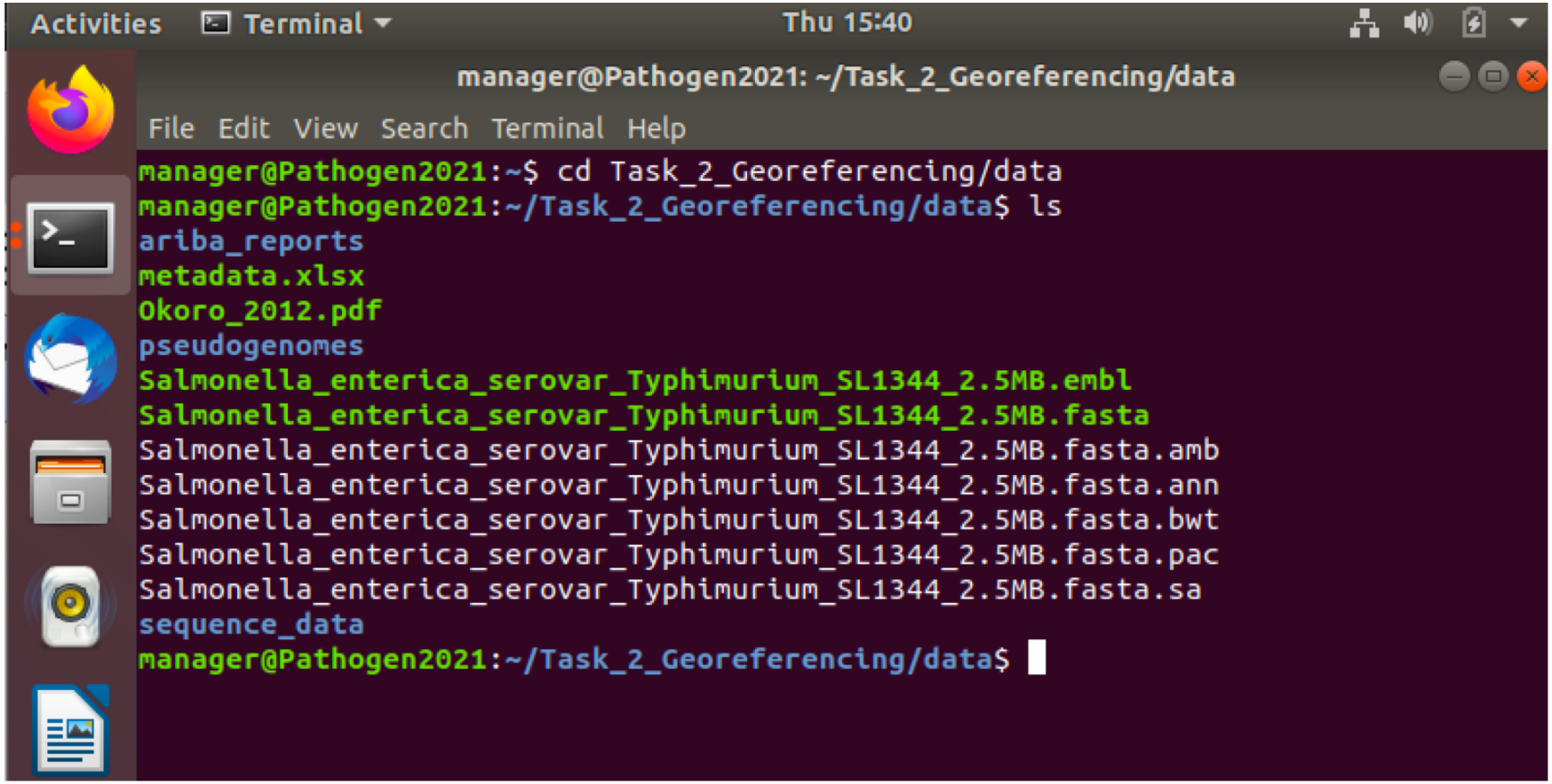
Adapted from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2712085/

Horizontal vs Vertical gene transfer

HGT – Transduction, Transformation, Conjugation



The task!

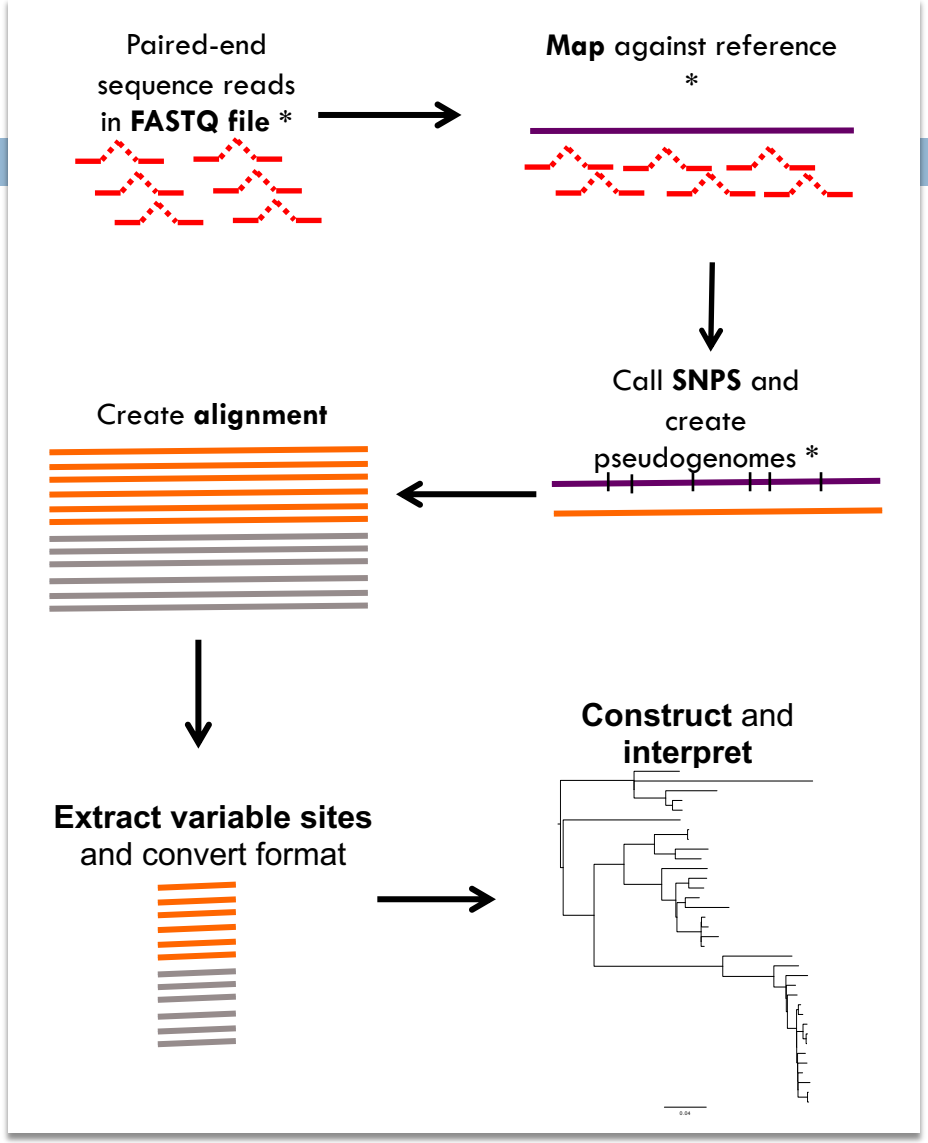

Activities
Terminal
Thu 15:40
manager@Pathogen2021: ~/Task_2_Georeferencing/data
File Edit View Search Terminal Help
manager@Pathogen2021:~$ cd Task_2_Georeferencing/data
manager@Pathogen2021:~/Task_2_Georeferencing/data$ ls
ariba_reports
metadata.xlsx
Okoro_2012.pdf
pseudogenomes
Salmonella_enterica_serovar_Typhimurium_SL1344_2.5MB.embl
Salmonella_enterica_serovar_Typhimurium_SL1344_2.5MB.fasta
Salmonella_enterica_serovar_Typhimurium_SL1344_2.5MB.fasta.amb
Salmonella_enterica_serovar_Typhimurium_SL1344_2.5MB.fasta.ann
Salmonella_enterica_serovar_Typhimurium_SL1344_2.5MB.fasta.bwt
Salmonella_enterica_serovar_Typhimurium_SL1344_2.5MB.fasta.pac
Salmonella_enterica_serovar_Typhimurium_SL1344_2.5MB.fasta.sa
sequence_data
manager@Pathogen2021:~/Task_2_Georeferencing/data$

Steps you will need to do

- Phylogenetic analysis
- AMR typing (new)
- Geo-tagging (new)

Phylogenetic tree

- Steps you will have learnt in modules 3 and 4
- Use RAxML to generate tree.
  - HPC (High performance computing)
  - ML (Statistical model)
  - GTR (Evolutionary model)
  - GAMMA (rate of distribution among sites)



Generating a phylogenetic tree using RAxML

Program optimised for **H**igh **P**erformance **C**omputing (HPC) which uses a **M**aximum **L**ikelihood (ML) statistical model

Output prefix

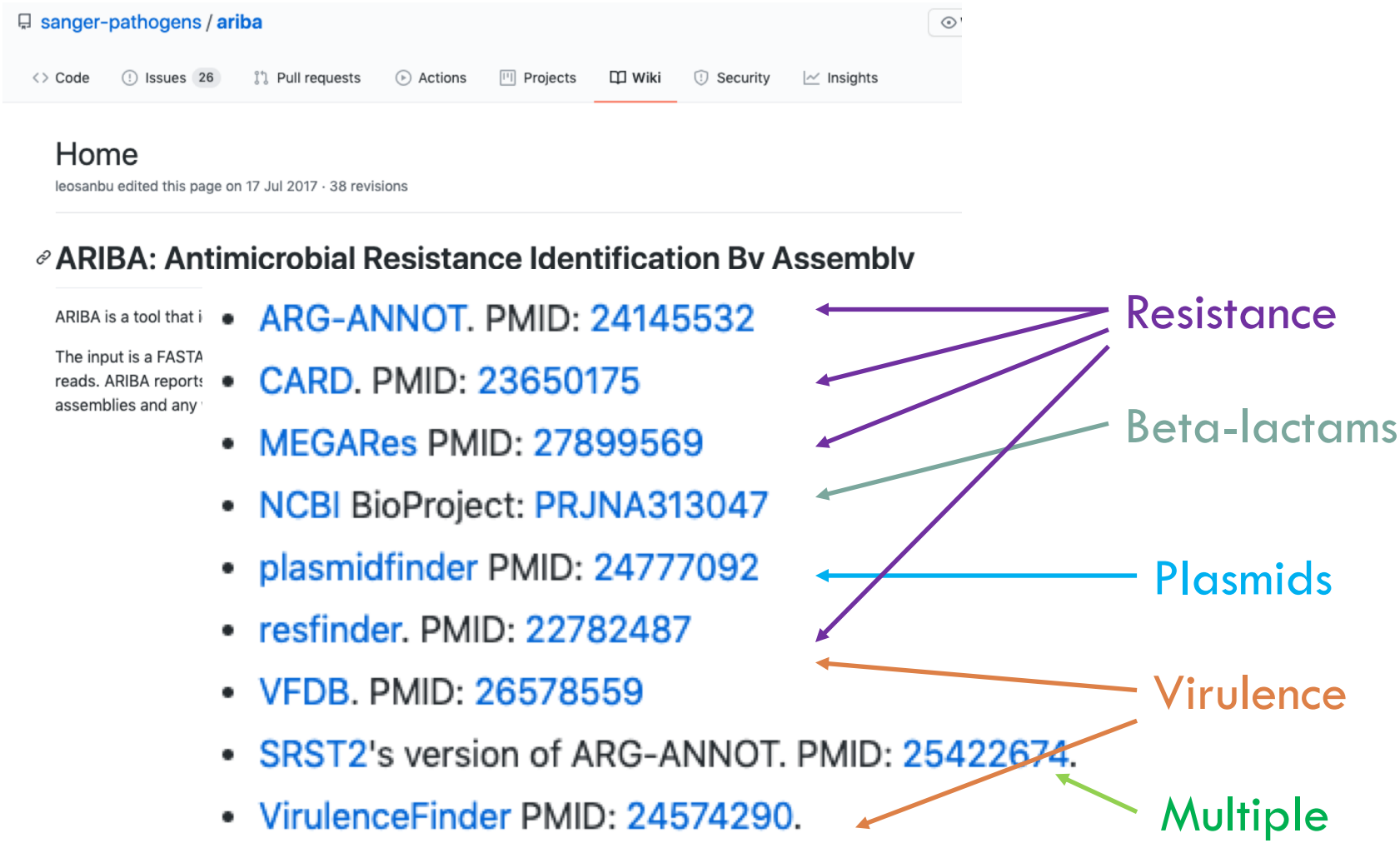
```
raxmlHPC -m GTRGAMMA -p 12345 -n STm -s All_snps.phy
```

GTR (Evolutionary model)
GAMMA (rate of distribution among sites)

Input file

Random seed used for the parsimony inferences

Step 2: Gene detection using ARIBA



ARIBA reference: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5695208/

The database – multista (.fa) file


```
>PDC-4.3002501.FJ666067.0-1194.1619
ATGCGCGATACCAGATTCCCCTGCCTGTGCGGCATCGCCGCTTCCACACTGCTGTTCGCC
ACCACCCCGGCCATTGCCGGCGAGGCCCCGGCGGATCGCCTGAAGGCACTGGTCGACGCC
GCCGTACAACCGGTGATGAAGGCCAATGACATTCCGGGCCTGGCCGTAGCCATCAGCCTG
AAAGGAGAACCGCATTACTTCAGCTATGGGCTGGCCTCGAAAGAGGACGGCCGCCGGGTG
ACGCCGGAGACCCTGTTCGAGATCGGCTCGGTGAGCAAGACCTTCACCGCCACCCTCGCC
GGCTATGCCCTGGCCCAGGACGAGATGCGCCTCGACGACCGCGCCAGCCAGCACTGGCCG
GCACTGCAGGGCAGCCGCTTCGACGGCATCAGCCTGCTCGACCTCGCGACCTATACCGCC
GGCGGCTTGCCGCTGCAGTTCCCCGACTCGGTGCAGAAGGACCAGGCACAGATCCGCGAC
TACTACCGCCAGTGGCAGCCGACCTACGCGCCGGGCAGCCAGCGCCTCTATTCCAACCCG
AGCATCGGCCTGTTCGGCTATCTCGCCGCGCGCAGCCTGGGCCAGCCGTTCGAACGGCTC
ATGGAGCAGCAAGTGTTCCCGGCACTGGGCCTCGAACAGACCCACCTCGACGTGCCCGAG
GCGGCGCTGGCGCAGTACGCCCAGGGCTATGGCAAGGACGACCGCCCGCTACGGGTCGGT
CCCGGCCCGCTGGATGCCGAAGGCTACGGGGTGAAGACCAGCGCGGCCGACCTGCTGCGC
TTCGTCGATGCCAACCTGCATCCGGAGCGCCTGGACAGGCCCTGGGCGCAGGCGCTCGAT
GCCACCCATCGCGGTTACTACAAGGTCGGCGACATGACCCAGGGCCTGGGCTGGGAAGCC
TACGACTGGCCGATCTCCCTGAAGCGCCTGCAGGCCGGCAACTCGACGCCGATGGCGCTG
CAACCGCACAGGATCGCCAGGCTGCCCGCGCCACAGGCGCTGGAGGGCCAGCGCCTGCTG
AACAAGACCGGTTCCACCAACGGCTTCGGCGCCTACGTGGCGTTCGTCCCGGGCCGCGAC
CTGGGCCTGGTGATCCTGGCCAACCGCAACTATCCCAATGCCGAGCGGGTGAAGATCGCC
TACGCCATCCTCAGCGGCCTGGAGCAGCAGGGCAAGGTGCCGCTGAAGCGCTGA
>CblA-1.3002999.GQ343019.132-1023.1188
ATGAAAGCATATTTCATCGCCATACTTACCTTATTCACTTGTATAGCTACCGTCGTCCGG
GCGCAGCAAATGTCTGAACTTGAAAACCGGATTGACAGTCTGCTCAATGGCAAGAAAGCC
ACCGTTGGTATAGCCGTATGGACAGACAAAGGAGACATGCTCCGGTATAACGACCATGTA
CACTTCCCCTTGCTCAGTGTATTCAAATTCCATGTGGCACTGGCCGTACTGGACAAGATG
GATAAGCAAAGCATCAGTCTGGACAGCATTGTTTCCATAAAGGCATCCCAAATGCCGCCC
AATACCTACAGCCCCCTGCGGAAGAAGTTTCCCGACCAGGATTTCACGATTACGCTTAGG
GAACTGATGCAATACAGCATTTCCCAAAGCGACAACAATGCCTGCGACATCTTGATAGAA
TATGCAGGAGGCATCAAACATATCAACGACTATATCCACCGGTTGAGTATCGACTCCTTC
AACCTCTCGGAAACAGAAGACGGCATGCACTCCAGCTTCGAGGCTGTATACCGCAACTGG
AGTACTCCTTCCGCTATGGTCCGACTACTGAGAACGGCTGATGAAAAAGAGTTGTTCTCC
AACAAGGAGCTGAAAGACTTCTTGTGGCAGACCATGATAGATACTGAAACCGGTGCCAAC
AAACTGAAAGGTATGTTGCCAGCCAAAACCGTGGTAGGACACAAGACCGGCTCTTCCGAC
CGCAATGCCGACGGTATGAAAACTGCAGATAATGATGCCGGCCTCGTTATCCTTCCCGAC
GGCCGGAAATACTACATTGCCGCCTTCGTCATGGACTCATACGAGACGGATGAGGACAAT
GCGAACATCATCGCCCGCATATCACGCATGGTATATGATGCGATGAGATGA
>ompF.3003390.U00096.3.985893-986982.5413
ATGATGAAGCGCAATATTCTGGCAGTGATCGTCCCTGCTCTGTTAGTAGCAGGTACTGCA
AACGCTGCAGAAATCTATAACAAAGATGGCAACAAAGTAGATCTGTACGGTAAAGCTGTT
GGTCTGCATTATTTTTCCAAGGGTAACGGTGAAAACAGTTACGGTGGCAATGGCGACATG
ACCTATGCCCGTCTTGGTTTTAAAGGGGAAACTCAAATCAATTCCGATCTGACCGGTTAT
GGTCAGTGGGAATATAACTTCCAGGGTAACAACTCTGAAGGCGCTGACGCTCAAACTGGT
AACAAAACGCGTCTGGCATTCGCGGGTCTTAAATACGCTGACGTTGGTTCTTTCGATTAC
GGCCGTAACTACGGTGTGGTTTATGATGCACTGGGTTACACCGATATGCTGCCAGAATTT
GGTGGTGATACTGCATACAGCGATGACTTCTTCGTTGGTCGTGTTGGCGGCGTTGCTACC
TATCGTAACTCCAACTTCTTTGGTCTGGTTGATGGCCTGAACTTCGCTGTTCAGTACCTG
GGTAAAAACGAGCGTGACACTGCACGCCGTTCTAACGGCGACGGTGTTGGCGGTTCTATC
```

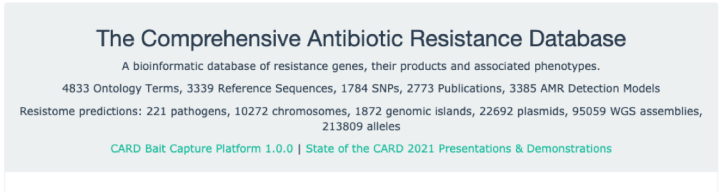
The database - additional information

Ta file with additional information to help with interpretation.

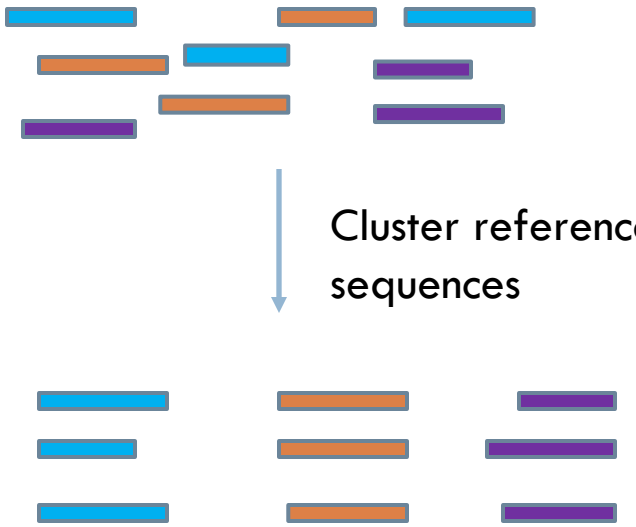
```
PDC-4.3002501.FJ666067.0-1194.1619      1       0       .       .       PDC-4
PDC-4.3002501.FJ666067.0-1194.1619      1       0       .       .       b'PDC-4 is a extended-spectrum beta-lactamase found in Pseudomona
eruginosa.'
CblA-1.3002999.GQ343019.132-1023.1188   1       0       .       .       CblA-1
CblA-1.3002999.GQ343019.132-1023.1188   1       0       .       .       b'CblA-1 beta-lactamase is a class A beta-lactamase found in Bact
ides uniformis that is species-specific.'
ompF.3003390.U00096.3.985893-986982.5413        1       1       .       .       Escherichia coli ompF with mutation conferring resistance
 beta-lactam antibiotics
ompF.3003390.U00096.3.985893-986982.5413        1       1       G141D   .       b'The Escherichia coli ompF (oprF) is a nonspecific porin
annel involved in the membrane translocation of small hydrophilic molecules, including and especially beta-lactam antibiotics. Mutations
ompF can decrease diffusion of antibiotics across the cellular membrane, thereby decreasing overall susceptibility through absence of por
function.'
ompF.3003390.U00096.3.985893-986982.5413        1       1       R154D   .       b'The Escherichia coli ompF (oprF) is a nonspecific porin
annel involved in the membrane translocation of small hydrophilic molecules, including and especially beta-lactam antibiotics. Mutations
ompF can decrease diffusion of antibiotics across the cellular membrane, thereby decreasing overall susceptibility through absence of por
function.'
ompF.3003390.U00096.3.985893-986982.5413        1       1       R154A   .       b'The Escherichia coli ompF (oprF) is a nonspecific porin
annel involved in the membrane translocation of small hydrophilic molecules, including and especially beta-lactam antibiotics. Mutations
ompF can decrease diffusion of antibiotics across the cellular membrane, thereby decreasing overall susceptibility through absence of por
function.'
ompF.3003390.U00096.3.985893-986982.5413        1       1       G141E   .       b'The Escherichia coli ompF (oprF) is a nonspecific porin
annel involved in the membrane translocation of small hydrophilic molecules, including and especially beta-lactam antibiotics. Mutations
ompF can decrease diffusion of antibiotics across the cellular membrane, thereby decreasing overall susceptibility through absence of por
function.'
```

Antimicrobial resistance (AMR) gene detection - ARIBA

ARIBA reference: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5695208/




```
ariba getref
```



Cluster reference sequences

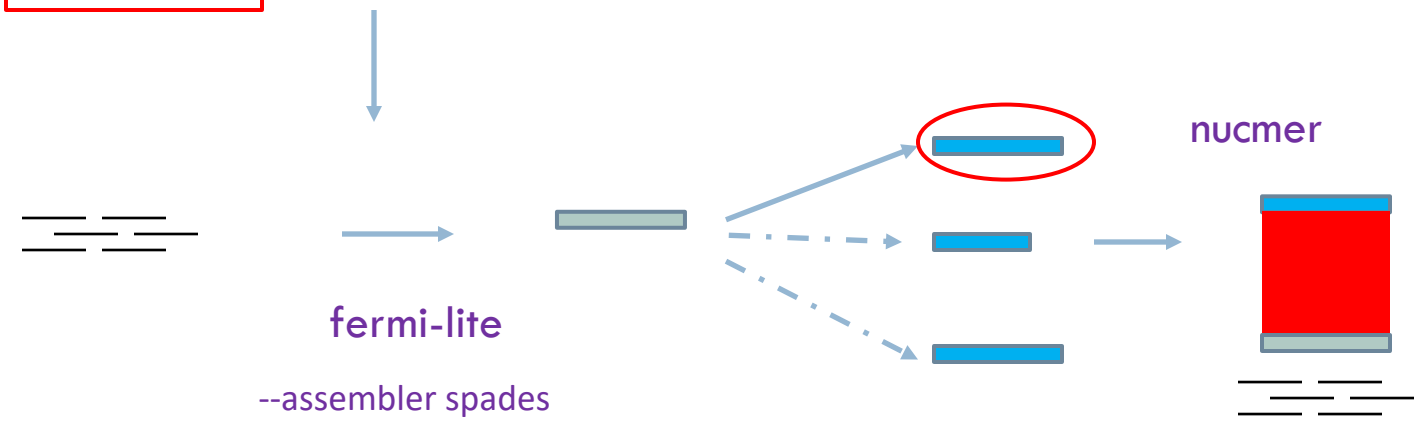
```
ariba prepareref
```

Adapted from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5695208/

Antimicrobial resistance (AMR) gene detection - ARIBA

Map your reads against the database using minimap

```
ariba run
```



Bowtie2 – aligner

Mpileup – snp calling

Adapted from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5695208/

Antimicrobial resistance (AMR) gene detection - ARIBA

Sample_1_report1.tsv

Sample_1_report1.tsv

⋮

Sample_n_report1.tsv

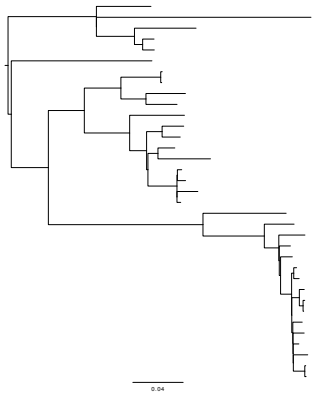
```
ariba summary out report.*.tsv
```

```
out.summary.csv
out.summary.phandango.csv
out.summary.phandango.tre
```

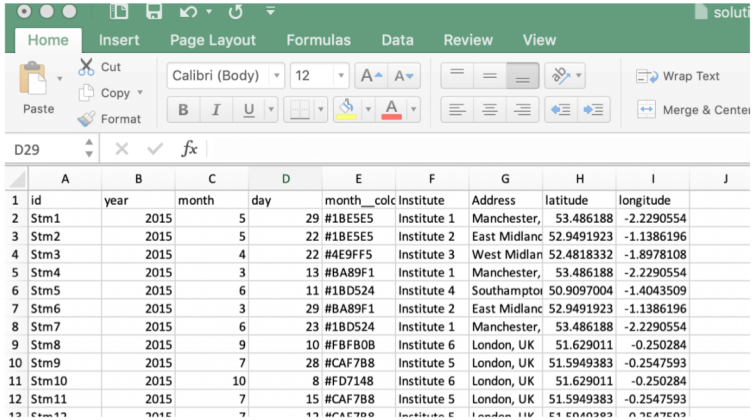
Visualise results in Phandango

Step 3: Geotagging

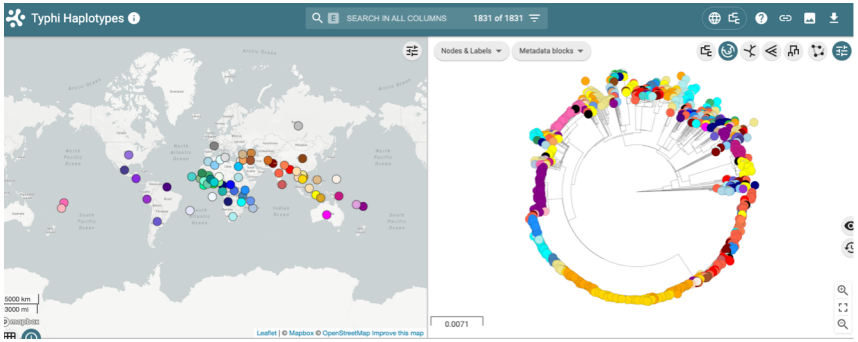
Geotagging and visualisation



+



	A	B	C	D	E	F	G	H	I	J
1	id	year	month	day	month_calc	Institute	Address	latitude	longitude	
2	Stm1	2015	5	29	#1BE5E5	Institute 1	Manchester,	53.486188	-2.2290554	
3	Stm2	2015	5	22	#1BE5E5	Institute 2	East Midlanc	52.9491923	-1.1386196	
4	Stm3	2015	4	22	#4E9FF5	Institute 3	West Midlan	52.4818332	-1.8978108	
5	Stm4	2015	3	13	#BA89F1	Institute 1	Manchester,	53.486188	-2.2290554	
6	Stm5	2015	6	11	#1BD524	Institute 4	Southampto	50.9097004	-1.4043509	
7	Stm6	2015	3	29	#BA89F1	Institute 2	East Midlanc	52.9491923	-1.1386196	
8	Stm7	2015	6	23	#1BD524	Institute 1	Manchester,	53.486188	-2.2290554	
9	Stm8	2015	9	10	#FBFB0B	Institute 6	London, UK	51.629011	-0.250284	
10	Stm9	2015	7	28	#CAF7B8	Institute 5	London, UK	51.5949383	-0.2547593	
11	Stm10	2015	10	8	#FD7148	Institute 6	London, UK	51.629011	-0.250284	
12	Stm11	2015	7	15	#CAF7B8	Institute 5	London, UK	51.5949383	-0.2547593	



Today's task

- **Phylogenetic analysis (1hr)**
  - SNP-calling and phylogenetic inference
- **Antimicrobial Resistance Screening (1hr)**
  - Run ARIBA and visualise with Phandango
- **Geo-referenceing (1hr)**
  - GPS locations and visualise with Microreact
- **Reporting (45min)**
  - Mini group discussions about your findings and consolidating slides for your group question.
- **Discussion with larger group – (45 mins)**

Georeferencing genomic mini groups

Group 1

Group 2

Group 3

Group 4

Q1 → Presentation

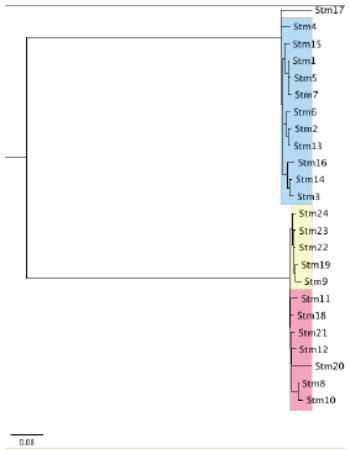
Q2 → Presentation

Q3 → Presentation

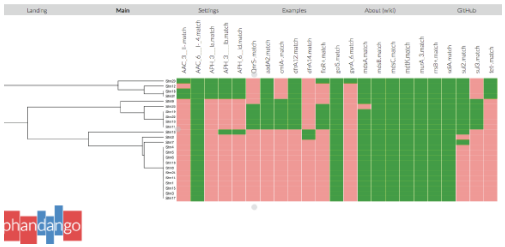
Q4 → Presentation

5 minutes presentation to the larger group

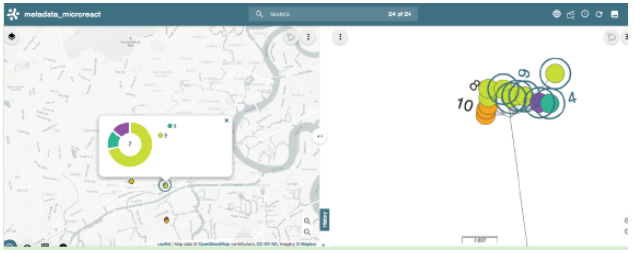
Steps to complete the task:



Build a tree



Resistance detection



Phylogeography

Presentation

Group questions

- **Group 1**: Looking at the phylogenetic relationship of the strains and associated metadata, which strain do you think are part of an outbreak? What measurements can you use to support your hypothesis. Hint: Look at the month of collection.
- **Group 2**: What resistance genes confer resistance to clinically relevant antibiotics for treating Salmonella infections? Hint: read the last page of instructions of the AMR screen.
- **Group 3**: How do you think genes encoded on the plasmids can affect the interpretation of the data? Hint: are any isolates genetically similar to others sensitive to antibiotics whose gene may be encoded on a plasmid.
- **Group 4**: What other data might be useful to have, to improve your hypothesis of which strains may be part of an outbreak? To give you some ideas, we suggest downloading the microreact metadata file from the Lo et al., 2019 study.

Today's task

- **Phylogenetic analysis (1hr)**
  - SNP-calling and phylogenetic inference
- **Antimicrobial Resistance Screening (1hr)**
  - Run ARIBA and visualise with Phandango
- **Geo-referenceing (1hr)**
  - GPS locations and visualise with Microreact
- **Reporting (45min)**
  - Mini group discussions about your findings and consolidating slides for your group question.
- **Discussion with larger group – (45 mins)**

Group questions

- **Group 1**: Looking at the phylogenetic relationship of the strains and associated metadata, which strain do you think are part of an outbreak? What measurements can you use to support your hypothesis. Hint: Look at the month of collection.
  - Samples part of outbreaks closely related.
  - Run beast to estimate the temporal signal
- **Group 2**: What resistance genes confer resistance to clinically relevant antibiotics for treating Salmonella infections? Hint: read the last page of instructions of the AMR screen.
  - Quinolones – 1 strains gyrA and another qnrS
  - Need phenotype information to confirm resistance phenotype
- **Group 3**: How do you think genes encoded on the plasmids can affect the interpretation of the data? Hint: are any isolates genetically similar to others sensitive to antibiotics whose gene may be encoded on a plasmid.
  - Plasmids can be transmitted horizontally therefore acquired by other strains resulting I similar/same AMR profile (microreact: https://microreact.org/project/dfgmM4xLyMZrNgHtZqas5P)
- **Group 4**: What other data might be useful to have, to improve your hypothesis of which strains may be part of an outbreak? To give you some ideas, we suggest downloading the microreact metadata file from the Lo et al., 2019 study.
  - Phenotype (MICs/Etest) information
  - Clinical information

Wrap up Demo: Microreact

https://microreact.org/project/GPS_tetSM