# Phylogenetic Analysis

Phylogenetics can be used in evolutionary biology to look at how closely species are related. As pathogens generate genetic variation rapidly, in pathogen genomics it can be used to investigate a pathogen outbreak including other pathogens that it is related to and to narrow down the location which is the original source of the outbreak. This can determine if the pathogen is novel in infecting humans and whether there are treatment strategies currently available to deal with the outbreak.

## Identify a closely related microbe to a sequence

A sample from a person with severe stomach pain and diarrhoea has been sequenced and the reads assembled in a pathogen laboratory. This sequence is called TY2482.fasta. It is in the directory /Users/trainee/Documents/ecoli\_outbreak/ on the Mac computer

We would like to identify if the pathogen from this sample could be related to any pathogens known to infect humans.

* Use microbial blast to identify possible related pathogens
  1. What is the top hit?
     1. How good is this match?
  2. Within the top 10 hits are there any other different pathogens? If so, what are they?
     1. Why might there be hits from this pathogen that match more than one other pathogen?

## Identify possible genes in the pathogen sample

Glimmer software can be used to predict genes in microbial DNA. You would like to see what its predictions are for your sample.

Log into the trainee account on the Mac computer.

* Open the terminal and run glimmer using the script g3-from-scratch.csh
  + Navigate to the directory /Users/trainee/Documents/ecoli\_outbreak/
  + Type csh /Users/trainee/Downloads/glimmer3.02/scripts/g3-from-scratch.csh TY2482\_assembly.fasta TY2482\_genes
    - This will generate predicted genes in the file TY2482\_genes.predict
      1. Take a screenshot of the top of this file once it is generated
  + To convert the file to GFF3 format run the bash script make\_gff.sh located in the current directory
    - Type bash make\_gff.sh > TY2482\_assembly.gff
* Navigate to the directory /Users/trainee/Documents/ecoli\_outbreak/biopython/
* Activate a virtual environment to use biopython
  + Type source activate biopython\_env
  + Retrieve the sequences associated with the predicted genes using bedtools
    - Type bedtools getfasta -fi ../TY2482\_assembly.fasta -bed ../TY2482\_assembly.gff -name -s > TY2482\_assembly\_cds.fasta
      1. Genes on the minus strand will be reverse complemented
      2. Open the generated file (TY2482\_assembly\_cds.fasta) and take a screenshot of the top of the file
  + Translate the sequences into predicted protein sequences using the script make\_protein.py (this script uses the helpful BioPython package and takes two command line arguments)
    - Type (all on one line) python make\_protein.py “/Users/trainee/Documents/ecoli\_outbreak/biopython/TY2482\_assembly\_cds.fasta” “/Users/trainee/Documents/ecoli\_outbreak/biopython/TY2482\_assembly\_protein.fasta”
      1. Open the generated file (TY2482\_assembly\_protein.fasta) and take a screenshot of the top of the file
  + Close the virtual environment by typing source deactivate
  + Use the predicted protein sequences to predict possible functions of the predicted genes in your sample based on sequence homology (similarity) with other sequences in a database. We could do this for all of our predicted genes locally, but as a demonstration we will select a couple of them and use a website
    - Open a web browser and navigate to <https://orthomcl.org/orthomcl/>
      1. Open New Search>Sequences>BLAST
      2. Copy the protein sequence from two of the predicted genes from the file TY2482\_assembly\_protein.fasta (one by one) (use the default settings)
         * Take a screenshot of the output

1. Explain (briefly) what the output might indicate
2. Explain (briefly) why it might be useful to find potential genes in a new pathogen isolate

## Explore relationships between isolates

Log into the trainee account on the Mac computer.

### *E.coli*

The software Mauve aligns regions that have homology between different samples. This is useful because during recombination, the genome can be rearranged meaning that sequences can end up in a different order or the other way around from the ancestral sequence. The software lines up blocks of the genome that have homology and colour codes them.

* **A.** You receive assembled sequences (contigs) from samples from other laboratories around the world with local outbreaks of a similar disease. You are interested in whether these outbreaks might be related to one another and to your original isolate. These files are named 1.fasta, 2.fasta and 3.fasta and are in the directory ecoli\_outbreak in the documents directory (/Users/trainee/Documents/ecoli\_outbreak/).
* On the Mac, take the files and use Mauve to make a multiple sequence alignment of these sequences received from other laboratories and your sequence (TY2482\_assembly.fasta). Use the default settings.
  + Choose the progressiveMauve alignment for the sequences
  + Save the output (use the output box before running the software) as MauveAln
  + Export the image created as Mauve.jpeg and screenshot it
  + The lines between samples link the similar sequences between samples. Click on a coloured block on one sample. The software will line it up with similar blocks in the other samples. This illustrates the relationship between the sequences in the samples. Sometimes these blocks can be in very different positions in the sequences perhaps showing a genome rearrangement event.

**B.** You have assembled your original sample’s genome into contigs. Mauve can also be used to refine a draft assembly if we have a reference genome that we know that it is related to.

* Download the reference genome for the top hit from your BLAST search performed earlier
* Load the reference genome into Mauve first and your assembly (TY2482\_assembly.fasta) into Mauve as the second sample. Use progressiveMauve to align
  + Take a screenshot
* Select the Move Contigs option in Mauve and load in your reference genome as the first sample and your draft assembly (TY2482\_assembly.fasta) as the second sample.
  + You need to provide an output directory- use the directory reordering within the e\_coli outbreak directory
  + Mauve will move the contigs in the draft assembly to attempt to better match the reference genome
  + Take a screenshot to compare to the original alignment

### *S. enteritidis*

You have a number of isolates from a different outbreak from different laboratories. You are interested in how closely related the different samples may be to one another. You know that the isolates are from salmonella (*S. enteritidis*) and you would like to use the reference genome for this organism as a baseline for comparison with your samples.

To facilitate this analysis, you have set up an instance of SnapperDB (created by Public Health England), a postgres database with associated scripts (<https://github.com/phe-bioinformatics/snapperdb>).

You have already downloaded the relevant reference genome AM933172 from NCBI and loaded it into your database. You have also populated your database with your fastq files (reads), mapped them to this reference genome using BWA and called variants using the PHEnix pipeline from Public Health England for seven out of your eight isolates.

As a prerequisite to completing your analysis you need to upload, map and variant call your eighth sample.

* On the Mac computer, open Postgres.app and make sure that the PostgreSQL 11 server is running
* Open the terminal and navigate to the directory /Users/trainee/Documents/salmonella\_outbreak/
  + If you prefer, you can also open this directory in the finder window to make it easier to inspect by eye
  + In this directory you will see a file called snapper.txt and three sub directories, configs, fastqs and reference genomes.
    - The snapper.txt file contains environment variables you will need to run the scripts that process the data for the snapperdb.
      * Load the variables into your open terminal by typing
        + . <space> snapper.txt
    - The configuration file for your database is within the configs folder. It is called ebg4\_config.txt
      * This file will be passed to the snapperdb scripts, to pass the name and connectivity paramaters of the postgres database, the reference genome identifier and parameters for the mapping and variant calling parts of the process
    - The fastqs containing the reads from your isolates are in the fastqs directory. Nested within this directory is a directory called snpdb which contains the vcf files output by the PHEnix pipeline. There are currently seven of these vcf files. After uploading the final sample, there will be eight.
    - The ref\_genomes directory contains the reference genome (AM933172) and associated files.
* Upload the final sample fastqs and process them using the PHEnix pipeline by typing (all on one line)
  + ../snapperdb/run\_snapperdb.py fastq\_to\_db -c /Users/trainee/Documents/salmonella\_outbreak/configs/ebp4\_config.txt /Users/trainee/Documents/salmonella\_outbreak/fastqs/SRR6131972\_1.fastq /Users/trainee/Documents/salmonella\_outbreak/fastqs/SRR6131972\_2.fastq

Once the above step has completed, we can look at the relationships between our samples which are now mapped to our reference genome. One way we can achieve this is to look at the SNP address of the samples. The reference genome is given the SNP address 1.1.1.1.1.1.1 (These thresholds are 250, 100, 50, 25, 10, 5 and 0 SNPs).

* To use a clustering algorithm to generate the SNP addresses for all of the samples, in the terminal type
  + ../snapperdb/run\_snapperdb.py update\_distance\_matrix -c /Users/trainee/Documents/salmonella\_outbreak/configs/ebp4\_config.txt
  + The output shows how closely the samples are related to one another in terms of numbers of SNPs that they have in common
    - Samples that have the same address at a position have fewer than that number of SNPs in common with at least one other sample within that cluster
      * Example: a sample with the SNP address 1.1.1.1.2.1.1 has less than 10 different SNPs from a sample with the SNP address 1.1.1.1.2.2.3
    - Take a screenshot of the output

1. Which sample is the most different from the others?
2. How do you know this?
3. Could it being older or from a different region geographically be why it is the most different? Why (briefly)?

The difference between samples can be visualised with a phylogenetic tree. As a starting point, a multiple sequence alignment is required. We need to input a list of the sample names to include

* In the current directory, make a list of the samples to include in the alignment and save it as a text file called sample\_list.txt.
  + Our sample names are
    - ERR2200244\_1
    - SRR5055288\_1
    - SRR5194193\_1
    - SRR5583186\_1
    - SRR5815674\_1
    - SRR5850014\_1
    - SRR5864444\_1
    - SRR6131972\_1
* To generate a SNP alignment, type the following into the terminal
  + ../snapperdb/run\_snapperdb.py get\_the\_snps -c /Users/trainee/Documents/salmonella\_outbreak/configs/ebp4\_config.txt -l /Users/trainee/Documents/salmonella\_outbreak/sample\_list.txt -o /Users/trainee/Documents/salmonella\_outbreak/salmonella\_alignment
  + This will generate a file called salmonella\_alignment.fa

We can use SplitsTree4 to compute the phylogenetic tree.

* Open the SplitsTree app on the Mac
* Load in the alignment file generated above (salmonella\_alignment.fa) and use the default settings
  1. Which sample looks like it is the most different?
     1. Is this the same file which appeared the most different in the SNP address analysis?
  2. Which samples are the most related?
* Take a screenshot of your phylogenetic network

Have a look at the different options available to compute a phylogenetic network (under Distances, Trees, Networks). There are a number of different methods to compute a phylogenetic network and definitions of phylogenetic network.

To draw a phylogenetic tree using the Neighbour-Joining tree method:

* Go to the Trees tab and select NJ and apply
* Go to the Draw tab and select Phylogram and apply
* Take a screenshot of your phylogenetic tree
  1. Which sample appears the most different?

## Supplementary Questions

1. For what reasons (give a couple) might a pathogen genomics laboratory be asked to sequence a pathogen or pathogen samples?
2. How might a patient benefit directly from having a pathogen sample that they have provided sequenced?

## Supplementary Information

Take a look at the map on the website <https://www.contagionlive.com/outbreak-monitor>. Here you can see reported outbreaks live marked by geographical location.

Combining this information with the relatedness of the strains could enable us to potentially track the spread of a disease from its original source as the pathogen evolves and changes over time. An example of this for HIV can be seen in this paper <https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(18)30647-9/fulltext>

<https://marlin-prod.literatumonline.com/cms/attachment/69bc8c9b-f726-4de6-a5b8-afea023c7d2d/gr3.jpg>

<https://marlin-prod.literatumonline.com/cms/attachment/dc3ff9c9-7977-4084-ac68-8b2f54a28864/gr4.jpg>

## Resources

<https://www.nature.com/articles/s41564-018-0296-2>

<https://cmr.asm.org/content/28/3/541>

<https://www.nejm.org/doi/full/10.1056/NEJMoa1107643>

<https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=MicrobialGenomes>

<http://ccb.jhu.edu/software/glimmer/index.shtml>

<https://ccb.jhu.edu/papers/glimmer3.pdf>

<https://www.ncbi.nlm.nih.gov/pubmed/15231754>

<https://github.com/phe-bioinformatics/snapperdb>

<https://www.ncbi.nlm.nih.gov/pubmed/29659710>

<http://www.splitstree.org/>

<https://academic.oup.com/mbe/article/23/2/254/1118872>

<https://orthomcl.org/orthomcl/>

<https://genome.cshlp.org/content/13/9/2178.full>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3196566/>

<https://www.contagionlive.com/outbreak-monitor>

<https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(18)30647-9/fulltext>

### Files

TY2482.fasta

TY2482\_assembly.fasta

1.fasta

2.fasta

3.fasta

SRR6131972\_1.fastq

SRR6131972\_2.fastq