Task 2 Georeferencing genomic data

Introduction

Phylogenetic trees based on whole genome data tell us the about the relationships of bacterial isolates to each other on a very fine scale. When we combine that high resolution information about the evolutionary relationships of isolates with geographical data it can inform our understanding of the current distribution of the pathogen and allow us to **infer the epidemiological processes** that have acted on the pathogen over time. The simplest example of this would be if a phylogeny showed that a pathogen was **geographically constrained** (e.g. isolates from the same region always cluster together). This might indicate that the pathogen is not highly mobile, whereas a pathogen with a phylogeny that shows isolates from distant regions are equally likely to be related to each other as isolates from nearby is likely to be highly mobile across regional borders. Geographical referencing of genomic data can also be **combined with temporal information** to study the movement of pathogens in space and time in real time for use in outbreak detection and monitoring.

For this task, you will be split into teams. Using the skills you learned in the structured modules, your team will use the all the **mapped sequences (.fa)** produced in Module 7 and to identify **single nucleotide polymorphism (SNP)** sites based on the reference sequence strain SL1344 and subsequently construct a **phylogenetic tree.** You will use the software **FigTree** to view and interpret your phylogeny and geo-reference your data using **Microreact** to 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.

The aims of this exercise are:

- 1) Introduce the biology & workflow
- 2) Gain experience in building and interpreting phylogenetic trees
- 3) Introduce concepts and tools for geo-referencing metadata
- 4) Show how **Next Generation Sequencing data** can be used to describe the evolution and distribution of a pathogen across a geographical area
- 5) Combine phylogenetic analysis and comparative genomic techniques
- 6) Demonstrate the value of shared data resources
- 7) Gain experience in presenting the results of Next Generation Sequencing Data analysis

Background

Biology

To learn about phylogenetic reconstruction and geo-referencing for epidemiological inference, we will work with some software that has already been introduced as well as some new software introduced in this module. We will work with real data from *Salmonella enetrica* serovar Typhimurium sampled from regional labs in England and Wales, United Kingdom in 2015.

Salmonella enetrica serovar Typhimurium

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 servoars 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 (Okoro *et al. Nature Genetics*, 2012).

Whole genome sequence analysis of this organism of provides some insight into the shortterm microevolution of *S*. Typhimurium. Understanding the level of diversity in this timeperiod is crucial in attempting to identify if this is an outbreak or sporadic infection.

Your task

The Global Health Authority (GHA) has asked you to provide an overview of *Salmonella enterica* serovar Typhimurium*in* England and Wales, using retrospective samples. In teams you will develop a whole-genome sequencing based tree from all 24 sequences and correlate this to the geography of the city. You will also look into the distribution of antimicrobial resistance and investigate the genetic basis for the resistance phenotype you identified in the laboratory. At the end of the task each group will present their findings.

The five teams are the will have been assigned in the previous day.

The following division of responsibilities in your teams is recommended. If there are extra people, then they should help with the tree builder and both should work on the georefencing task once you have a tree.

- Tree Builder SNP-calling and phylogenetic inference
- Antimicrobial Resistance Investigator ARIBA and Phandango
- Geo-referencer and Reporter geo-referencing with Microreact
- Presentation All group members

GENERAL INFORMATION

Data provided

As a team you will create a phylogenetic tree of the *S*. Typhimurium isolates from England and Wales. You will geo-reference this information, as well as antimicrobial resistance data, against the address of the isolates to form ideas about the distribution and epidemiology of the pathogen. Additionally, the genetic basis for the antimicrobial resistance will be explored.

To achieve this, each team is provided with the following files in the Task folder:

- A metadata table (**metadata.xls**) which contains information on the isolates including the date and address of collection.
- Your sequence data folder in each of your groups folders, which contain symlinked or symbolic linked sequenced data, fastq.gz (unix command: ln -s). These act like 'hyperlink' data. Symlinked data is often used to save space when you do not want to copy large files.
- An **ariba_reports** folder that contains a summary the resistance reports from ARIBA to save time. You are encouraged to run the ARIBA analysis on the samples allocated to your group.
- S. Typhimurium fasta and embl files, which you will use as a reverence.
- A pseudogenomes folder that contains an .fa file
- And a **PDF** of the literature reference cited on page 2.

group_1 group_8 group_10 group_9 group_2 metadata.xls group_3 Okoro_2012.pdf group_4 pseudogenomes group_5 Salmonella_enterica_serovar_Typhimurium_SL group_6 salmonella_enterica_serovar_Typhimurium_SL wt@Pathogens:~/Group_task_Georeferencing\$ ssigned group sm_1 ssigned group sm_2 ssigned group sm_3 ssigned group sm_2 ssigned group sm_3 ssigned group sm_3 ssigned group sm_3 ssigned group sm_3 ssigned group sgroup ssigned group <th>L1344_2.5MB. L1344_2.5MB. Ben Nylex Ward, Guiden Sch Minh Cary Ben Nylex Ward, Guiden Sch Minh Cary Ben Nylex Ward, Guiden 1</th> <th>embl fasta ude latitude</th> <th></th>	L1344_2.5MB. L1344_2.5MB. Ben Nylex Ward, Guiden Sch Minh Cary Ben Nylex Ward, Guiden Sch Minh Cary Ben Nylex Ward, Guiden 1	embl fasta ude latitude	
group_10 group_9 group_21 metadata.xls group_3 Okoro_2012.pdf group_4 pseudogenomes group_6 Salmonella_enterica_serovar_Typhimurium_SL group_6 Salmonella_enterica_serovar_Typhimurium_SL wt@Pathogens:~/Group_task_Georeferencing\$ axigned group MST smig. 19	L1344_2.5MB. L1344_2.5MB. Ben Nghe Word, Quận 1 Số Chỉ Minh Cây Ben Nghe Word, Quận 1	embl fasta	
group_10 group_9 group_2 metadata.xls group_3 Okoro_2012.pdf group_4 pseudogenomes group_5 Salmonella_enterica_serovar_Typhimurium_SL group_6 Salmonella_enterica_serovar_Typhimurium_SL wt@Pathogens:~/Group_task_Georeferencing\$ 1 Sm_1 2 Sm_3 3 Sm_5 2 Sm_5 3 Sm_3 3 Sm_1	L1344_2.5MB. L1344_2.5MB.	embl fasta ude latitude	
group_2 metadata.xls group_3 Okoro_2012.pdf group_4 pseudogenomes group_5 Salmonella_enterica_serovar_Typhimurium_SL group_6 Salmonella_enterica_serovar_Typhimurium_SL wt@Pathogens:~/Group_task_Georeferencing\$ amigned group MIST sm_3 19 sm_3 19 sm_6 19 sm_7 19 sm_8 19 sm_7 19 sm_8 19 sm_7 19 sm_7 19 sm_8 19 sm_7 19 sm_8 10 sm_8 <th>L1344_2.5MB. L1344_2.5MB. Ben Nghe Word, Quận 1 50 Chi Minh Cây Ben Nghe Word, Quận 1</th> <th>embl fasta ude latitude</th> <th></th>	L1344_2.5MB. L1344_2.5MB. Ben Nghe Word, Quận 1 50 Chi Minh Cây Ben Nghe Word, Quận 1	embl fasta ude latitude	
stopp_1 Okcoro_2012.pdf group_3 Okcoro_2012.pdf group_4 pseudogenomes group_5 Salmonella_enterica_serovar_Typhimurium_SL group_6 Salmonella_enterica_serovar_Typhimurium_SL wt@Pathogens:~/Group_task_Georeferencing\$ Mark staiped group smm_1 1 2015 staiped group smm_2 19 2015 Mark Behv Win dight, 82 Cool, pubme 4, Cub R Mark Mark Street, Behv Street, Behv Win dight, 82 Cool, pubme 4, Cub R staiped group smm_1 19 2015 Mark Behv Win dight, 82 Cool, pubme 4, Cub R Behv Win dight, 82 Cool, Dubme 4, Cub R smm_5 19 2015 Mark Behv Win dight, 82 Cool, Dubme 4, Cub R Behv Win dight, 82 Cool, Dubme 4, Cub R smm_5 19 2015 Lune Behv Win dight, 84 Cool, Dubme 4, Cub R Behv Win dight, 84 Cool, Dubme 4, Cub R smm_5 19 2015 Lune Behv Win dight, 84 Cool, Dubme 4, Cub R Behv Win dight, 84 Cool, Dubme 3, Cub R smm_6 19 2015 Lune Behv Win dight, 84 Cool, Dubme 3, Cub R Behv Win dight, 84 Cool, Dubme 3, Cub R smm_11 19 2015 <th< th=""><th>L1344_2.5MB. L1344_2.5MB. Ben Nghe Ward, Quận 1 tố Chi Minh City Ben Nghe Ward, Quận 1</th><th>embl fasta</th><th></th></th<>	L1344_2.5MB. L1344_2.5MB. Ben Nghe Ward, Quận 1 tố Chi Minh City Ben Nghe Ward, Quận 1	embl fasta	
group_3 Okoro_2012.pdf group_4 pseudogenomes group_5 Salmonella_enterica_serovar_Typhimurium_SL group_6 Salmonella_enterica_serovar_Typhimurium_SL wt@Pathogens:~/Group_task_Georeferencing\$ migned group Mist year neath singerd group Mist year neath Addres singerd group Smm_3 19 2015 Mark Helm Win cube R. 82 Co. 0, public R, 40 de 8 singerd group Smm_3 19 2015 Mark Helm Win cube R. 82 Co. 0, public R, 40 de 8 smm_4 19 2015 Arrill Helm Win cube R. 82 Co. 0, public R, 40 de 8 smm_3 19 2015 Arrill Helm Win cube R, 82 Co. 0, public R, 40 de 9 smm_4 19 2015 Arrill Helm Win cube R, 40 de 0, 40 de 8 smm_4 19 2015 Line Helm Win cube R, 40 de 0, 40 de 8 smm_5 19 2015 Line Helm Win cube R, 40 de 0, 40 de 8 smm_10 19 2015 Line Helm Win cube R, 40 de 0, 40 de 8 smm_111 19 2015 Line	L1344_2.5MB. L1344_2.5MB. Ben Nghe Word, Quide 1 fo Chi Minh City Ben Nghe Word, Quide 1	embl fasta de latitude	
group_4 pseudogenomes group_5 Salmonella_enterica_serovar_Typhimurium_SL group_6 salmonella_enterica_serovar_Typhimurium_SL wt@Pathogens:~/Group_task_Georeferencing\$ migged group 5 salmonella_enterica_serovar_Typhimurium_SL wt@Pathogens:~/Group_task_Georeferencing\$ migged group 5 salmonella_enterica_serovar_Typhimurium_SL wt@Pathogens:~/Group_task_Georeferencing\$ migged group 5 salmonella_enterica_serovar_Typhimurium_SL 2 ssigned group 5 5 5	L1344_2.5MB. L1344_2.5MB. Ben Nghe Ward, Quận 1 to Chi Minh City Ben Nghe Ward, Quận 1	embl fasta	
group_5 Salmonella_enterica_serovar_Typhimurium_SL group_6 group_6 Salmonella_enterica_serovar_Typhimurium_SL wt@Pathogens:~/Group_task_Georeferencing\$ msigned group Sm_1 19 2015 sm_3 19 2015 May Hahn Mink Mail Street, Be Beh Wen quint R. Stock O, brubing 4, Oule 8 2 Sm_4 19 2015 May Hahn Mink Mail Street, Be Beh Wen Mink and 115, 227 Vien Henh, 12th Wesd, Oule 10, Oule 0, O	L1344_2.5MB. L1344_2.5MB. Ben Nghe Ward, Quận 1 to Chi Minh Ctty Ben Nghe Ward, Quận 1	embl fasta	
group_5 Salmonella_enterica_serovar_Typhimurium_SL group_6 Salmonella_enterica_serovar_Typhimurium_SL wt@Pathogens:~/Group_task_Georeferencing\$ #ssigned group smm_1 2005 May Beh wing dgn.8, 82 Co.0, briving 4, Outh 8 2 Sm_2 19 2005 May Beh wing dgn.8, 82 Co.0, briving 4, Outh 8 2 Sm_2 19 2005 May Beh wing dgn.8, 82 Co.0, briving 4, Outh 8 2 Sm_3 19 2005 May Beh wing dgn.8, 82 Co.0, briving 4, Outh 8 2 Sm_3 19 2005 May Beh wing dgn.8, 82 Co.0, briving 4, Outh 8 3 Sm_3 19 2005 May Beh wing dgn.8, 82 Co.0, briving 4, Outh 8 3 Sm_3 19 2005 May Beh wing dgn.8, 82 Co.0, briving 4, Outh 8 3 Sm_3 19 2005 May Beh wing dgn.8, 82 Co.0, briving 4, Outh 8 4 Sm_3 19 2005 June Beh wing dgn.8, 82 Co.0, briving 4, Outh 8 4 Sm_11 2005 July Beh wing dgn.8, 82 Co.0, briving 4, Outh 8 4<	L1344_2.SMB. L1344_2.SMB. Ben Nghe Ward, Quận 1 tố Chi Minh City Ben Nghe Ward, Quận 1	embl fasta Ide latitude	
group_6 Salmonella_enterica_serovar Typhimurium_SL wt@Pathogens:~//Group_task_Georeferencing\$	L1344_2.5MB . longit: Ben Nghe Ward, Quận 1 16 Chi Minh City Ben Nghe Ward, Quận 1	fasta ude latitude	
Solution <th colspan="</th> <th>longitu Ben Nghe Ward, Quận 1 to Chi Minh City to Chi Minh City Ben Nghe Ward, Quận 1</th> <th>ude latitude</th> <th></th>	longitu Ben Nghe Ward, Quận 1 to Chi Minh City to Chi Minh City Ben Nghe Ward, Quận 1	ude latitude	
wt@Pathogens:~/Group_task_Georeferencing\$ assigned group Smm_1 19 2015 May Medress 1 Smm_2 19 2015 May Heah Win oudin 8, 82 Cool Ch phubme 4, Oudin 8 2 Smm_2 19 2015 May Heah Win hind and 115, 527 SU Vin Heah , 124 Wins Chall 9, 004 2 Smm_4 19 2015 May Heah Win hind and 115, 527 SU Vin Heah , 124 Wins Chall 9, 004 2 Smm_5 19 2015 March Heah Win hind and 115, 527 SU Vin Heah , 124 Wins Chall 9, 004 2 Smm_5 19 2015 March Heah Win hind and 115, 527 SU Vin Heah , 124 Wins Chall 9, 104 W	Ben Nghe Ward, Quận 1 to Chi Minh City to Chi Minh City Ben Nghe Ward, Quận 1	Jde latitude	
assigned group isomple Smm_1 VLST year month Address 1 Smm_1 1 2015 May Behn Wein right R, 82 Clo LS, phut/ng 4, Club R 8, 1 May Behn Wein right R, 82 Clo LS, phut/ng 4, Club R 8, 1 May Behn Wein right R, 82 Clo LS, phut/ng 4, Club R 8, 1 Mark R 8, 22 Clo LS, phut/ng 4, Club R 8, 1 Mark R 8, 22 Clo LS, phut/ng 4, Club R 8, 1 Mark R 8, 22 Clo LS, phut/ng 4, Club R 8, 1 Mark R 8, 22 Clo LS, phut/ng 4, Club R 10, 1 Mark R 8, 22 Clo LS, phut/ng 4, Club R 10, 1 Mark R 8, 22 Clo LS, phut/ng 4, Club R 10, 1 Mark R 10, 1 </th <th>longit Ben Nghe Ward, Quận 1 Ho Chi Minh City Ho Chi Minh City Ben Nghe Ward, Quận 1</th> <th>ude latitude</th> <th></th>	longit Ben Nghe Ward, Quận 1 Ho Chi Minh City Ho Chi Minh City Ben Nghe Ward, Quận 1	ude latitude	
assigned group sample Smm_1 MLST year month Address 1 Smm_2 19 2015 May Hash Pulkin (and/in 8, 82 Cao LA), phut/ong 4, Oud/in 8 1 Smm_2 19 2015 May Hash Pulkin (International Hospital)?/, Rayven Thi Minh Khall Street, Be Smm_4 2 Smm_3 19 2015 Ann Behn Vehn Rind and 115, 527 2V veh Hann, Jizh Wez, Quieh 10, to Smm_4 Street, Be Smm_4 2015 Ann Behn Vehn Rind and 115, 527 2V veh Hann, Jizh Wez, Quieh 10, to Smm_5 19 2015 Ann Behn Vehn Rind and 115, 527 2V veh Hann, Jizh Wez, Quieh 10, to Smm_5 10 2015 Ann Behn Vehn agin 8, 82 Cao LA, phut/org 4, Quieh 8, 64 Smm_1 20 10 30	longiti Ben Nghe Ward, Quận 1 Ho Chi Minh City Ho Chi Minh City Ben Nghe Ward, Quận 1	ude latitude	
assigned group Isomplex MLST year month Address 1 Sm_1 2015 May Behn Velon qidin 8, 82 Cao L0, phương 4, Quán 8 1 Sm_2 19 2015 May Hanh Phuc International Hospital 37, Negurem 1Minh Nhu Street, Behn Velon Nhinh cell 11, 55, 275 Ven Heim, 12, 124 Ven Heim, 124 Ven Logel, 124 Ven Heim, 124 Ven Logel, 124 Ven Heim, 124 Ven Logel, 1	Ben Nghe Ward, Quận 1 Ho Chi Minh City Ho Chi Minh City Ben Nghe Ward, Quận 1	ude latitude	
Sm1 19 2015 May Behn Velon galn 8, 82 Clo LG, burbing 4, Qualt 8 1 Stm2 19 2015 May Hanh Photo, International Hospital 37, Reyner Th Mith Nika Street, Be Stm3 19 2015 April Behn Velon Nihan Galn 15, 527 SU Ven Heinh, 12th Verar, Quie 10, Io Stm4 10 Stm. Annot Behn Velon Nihan Galn 15, 527 SU Ven Heinh, 12th Verar, Quie 10, Io Stm5 10 2015 Annot Behn Velon Galn A, Ba Clo LG, Dirbürg 4, Quie 10, Io Stm5 10 2015 Annot Behn Velon Galn A, Ba Clo LG, Dirbürg 4, Quie 10, Io Stm5 10 2015 Annot Heinh Photo, Galn A, Ba Clo LG, Dirbürg 4, Quie 10, Io Stm5 10 2015 Annot Heinh Photo, Galn A, Ba Clo LG, Dirbürg 4, Quie 10, Io Stm5 10 2015 Stm5 2015 Stm. Street, Be Stm5 10 2015 Stm. 10 2015 Stm. Annot Heinh Nich Ad Street, Be Stm10 10 2015 Stm4 10 10 10 2015 Stm4 10 1	Ben Nghe Ward, Quận 1 Ho Chi Minh City Ho Chi Minh City Ben Nghe Ward, Quận 1		
1 Stm_2 19 2015 May Hanh Phace International Hospital (7), Nayuem H Mith Rink H Steets (18) 56m_3 10 2015 Ann Behn Vehn Stm 11 15, 25, 25, 45, 45, 45, 45, 45, 45, 45, 45, 45, 4	Ben Nghe Ward, Quận 1 Ho Chi Minh City Ho Chi Minh City Ben Nghe Ward, Quận 1		
Sm.,1 30 2015 April Berth wide Namia and 115, 527 59 Van Henden, 114 Waard, Lagel Bull, Out Sm.,4 3015 Akm Berth wide Namia and 115, 527 59 Van Henden, 116, 528 50, 519 59 Van Henden, 116, 528 50, 519 59 Van Henden, 116, 528 500 Van Henden,	Ho Chi Minh City Ho Chi Minh City Ben Nghe Ward, Quận 1		
Stm S 13 2013 Num BigHt Wath Indian, Big 2015, XU, Wath Indian, Big 2015, X	Ben Nghe Ward, Quận 1		
Sm. 5 19 2015 March Hain PRpuic Intervational Hospital 57, Nayore 11, Mitor Mais Street, Be Sm. 7 3 Sm. 7 2015 June Beh vidy and print, 82 cold cold structures 4, Outpin 4, Sm. 9 100 mitor 40, 074 Vidy and kpin 4, 200 mitor 4,	Ben Nghe Ward, Quận 1		
Sm7 19 2015 June Behn Vehn guln 8, 22 Cos LD, pruiving 4, Outor 8 3 Sm8 1 2015 Spember Behn Vehn guln 8, 22 Cos LD, pruiving 4, Outor 8 5 Sm8 1 2015 Spember Behn Vehn guln 8, 24 Cos LD, pruiving 4, Outor 8 5 Sm8 1 2015 July As Binh Inegatis L4A cells m, 7m Verun Guln 9, 15 Spember 4 Sm11 1 2015 July Behn Vehn Behn Nihef 60, 754 Vehn Keip, pruiving 1, 204h 5 Spember Spember Behn Vehn Behn Nihef 60, 754 Vehn Keip, pruiving 1, 204h 5 Spember Spember Behn Vehn Behn Nihef 60, 754 Vehn Keip, pruiving 1, 204h 5 Spember Spember Behn Vehn Behn Nihef 60, 754 Vehn Keip, pruiving 1, 204h 5 Spember Spember Behn Vehn Behn Nihef 60, 754 Vehn Keip, pruiving 1, 204h 5 Spember Spember Behn Vehn Behn Nihef 60, 754 Vehn Keip, pruiving 1, 204h 5 Spember 1, 204h 5 Spember 1, 204h 5 Spember 1, 204h 5 Spember 2, 204h 5			
3 Stm_8 19 2015 September Behn Wein Behn Nihet 6x0, 764 Vol Van Kick, prudvog LQuén 5 Stm_9 10 2015 Juhy An Behn Nihet 6x0, 764 Vol Van Kick, prudvog LQuén 5 Stm_10 19 2015 October Behn Ville Behn Nihet 6x0, 764 Vol Van Kick, prudvog LQuén 5 Stm_11 19 2015 Unit Behn Nihet 6x0, 764 Vol Van Kick, prudvog LQuén 5 Stm_12 19 2015 Juhy Behn Ville Behn Nihet 6x0, 764 Vol Van Kick, prudvog LQuén 5 Stm_12 19 2015 Juhy Behn Ville 6x0, 764 Vol Van Kick, prudvog LQuén 5 Stm_12 19 2015 Juhy Behn Ville 6x0, 764 Vol Van Kick, prudvog LQuén 5			
Sm. 19 2015 July An Binh Hoopital (34 An Binh), Thi Ward (24 Binh Jinh) An Binh Hoopital (37 Vin Wark (45 p. hu/ong 1,04 ho 4 Sm. 11 19 2015 July Bighn Wing dot, 19 Vin Wark (40 p. hu/ong 1,04 ho 5 4 Sm. 11 19 2015 July Bighn Wing dot, 19 Vin Wark (40 y. 64 Vin Wark (40 p. 104 hog) 2,04 hog 5 Sm. 12 19 2015 July Bighn Wing Bighn Nhight (40 Yi Vin Wark (40 p. 104 hog) 1,04 hog 1,04 hog 1,04 hog 1,04 hog 5 Sm. 1 10 2015 July Bighn Wing Bighn Nhight (40 Yin Vin Wark (40 p. 104 hog 1, 104 hog 1,04 hog <td></td> <td></td> <td></td>			
Sm_11 1 2015 October Behn Weit Behn Nihet 601, 764 Vol Win Kills, pruhoma, Zuden 5 4 Sm_11 1 2015 July Behn Weit Behn Nihet 601, 764 Vol Win Kills, pruhoma, Zuden 5 5m_112 13 2015 July Behn Weit 601, 764 Vol Win Kills, pruhoma, Zuden 5 5m_112 13 2015 May Hanh Phut 601, 764 Vol Win Kills, pruhoma, Zuden 5 5 Sm_12 13 2015 May Hanh Phut 601, 764 Vol Win Kills, pruhoma, Zuden 5			
 Stm_11 19 2015 Juny 19 depin vigin gerin riving rout, 764 Vo Yan Kiệt, phương 1, Quán S Stm_12 19 2015 Juny 19 Behn kiệt don, 764 Vo Yan Kiệt, phương 1, Quán S Stm_13 19 2015 May Hanh Phuc International Hospital, 97, Nguyen Thi Minh Khai Street, Be 			
5 Stm_13 19 2015 May Hanh Phuc International Hospital,97, Nguyen Thi Minh Khai Street, Be			
5	Ben Nghe Ward, Quân 1		
5tm 14 19 2015 April Běnh viên Nhân dân 115, 527 Sự Van Hanh, 12th Ward, Quân 10, Ho	Ho Chi Minh City		
5 Stm_15 19 2015 March Bênh viện quận 8, 82 Cao Lỗ, phường 4, Quận 8			
Stm_16 19 2015 April Bệnh viện Nhân dân 115, 527 Sư Vạn Hạnh, 12th Ward, Quận 10, Ho	Ho Chi Minh City		
7 Stm_17 19 2015 January Bệnh viện Nhân dân 115, 527 Sư Vạn Hạnh, 12th Ward, Quận 10, Ho	Ho Chi Minh City		
Stm_18 19 2015 July Benn Wen Benn Nniet acti, /s64 Vo Van Kiet, prucing 1, Quan 5			
8 Stm 20 19 2015 July Ránh viện Ránh Niệt đới 754 Vă Văn Kiết nhưởng 1 Quận 5			
Stm 21 19 2015 July Bénh viên Bénh Nhiệt đới, 764 võ Văn Kiệt, phương LQuận 5			
9 Stm_22 19 2015 June An Binh Hospital,146 An Binh, 7th Ward, Quận 5			
10 Stm_23 19 2015 June An Binh Hospital,146 An Binh, 7th Ward, Quận 5			

GENERAL INFORMATION

Your isolate names

The isolate names you can see in the subfolders of your group folder. There are two files for every isolate _1.fastq.gz and _2.fastq.gz. These represent the forward and reverse reads of paired end sequencing for that isolate.

When you work with your own sequencing data after the course, other naming conventions will be used. As in the example above, it is likely this formats will include helpful pieces of information, so find out what your own sequencing data names mean when the time comes!

How to use this module

As in some previous modules, you will be provided with many of the commands you will need to perform the analysis. As you will be distributing the tasks between people in your group, each role has their own set of guiding pages and focuses on different skills. You will learn about the other roles while integrating the results of your individual analyses and have the opportunity to work through the other sections in your own time.

GENERAL INFORMATION

Team Presentation

At the end of the task compile your findings and interpretations into a 5 minute presentation. All team members should contribute to making the presentation. Examples of some of the exciting key images you might produce are below, but don't be limited by these ideas - please be as creative as you like!



General Information

In this role you are responsible for the construction of a tree from whole genome sequencing data for your isolates. During this section you and your fellow tree maker will map the sequence data of your isolates to the *Salmonella enterica* Typhimurium reference genome SL1344. To save on time, you will only be mapping the fastq files to 2.5 million base pairs of the genome. Although this will take a long time, keep in mind that this step would ordinarily take many more hours of computation time. If there is more than one of you working on the mapping portion, you can work on half the samples in step 1. When you get to Step 2, combine your data and work on one computer together.

Bioinformatic processing of data into biologically-meaningful outputs involves the **conversion of data** into many different forms. Just like working in the laboratory, it's useful to to break this process down into individual steps and have a plan.

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

Step 1. Map and call SNPs for each isolate using commands introduced earlier in the course

Step 2. Create a whole genome sequence alignment

Step 3. Build a phylogenetic tree from the SNP data in your alignment

Step 4. Interpret your phylogeny and report the lineages to the geo-referencer

Step 1: Map and call SNPs for each isolate

Your reference sequence for this is *Salmonella enterica* serovar Typhimurium strain SL1344, called Salmonella_enterica_serovar_Typhimurium_SL1344_2.5MB.fasta in the task folder. You may want to create a local copy in your working directory by using the **cp** command.

Map the sequencing data for each isolate to the reference genome and obtain a pseudogenome (incorporating the isolate SNPs into the reference sequence). The required commands were covered in the **mapping and phylogeny** modules. If you struggle with the commands, ask the instructors for a command cheat sheet.

NOTE: Before you continue onto the next step, you must do some housekeeping. Refer to the mapping and phylogeny module, for which files you should remove.

Schematic of task workflow



*do for each isolate

Step 2: Create a whole genome sequence alignment for your data

Now you have created pseudogenomes (.fasta **NOT** .fastq) for each of your samples, you can use this data to **create a sequence alignment** to build a phylogenetic tree. Using this mapping based approach we are able to avoid the computational power required to align millions of base pairs of DNA that would be needed with e.g. CLUSTAL or MUSCLE. Here, because all of the isolates were mapped to the same reference genome, they are already the same length, so they can just be pasted together to form an alignment. Then, you can combine them with **information from global reference isolates** that were created for you in the same way.

Due to time constrains we have mapped all the samples to the same reference. The file can be found in the **pseudogenome** folder. The pseudogenomes of all 24 strains were combined together using the **cat** command as below.

cat *_pseudogenome.fasta > All_pseudogenomes.fa

This produces a multifasta file 'All_pseudogenomes.fa' that contains all 24 sequences. You can check all 24 sequences are present by opening it in seaview.

Here, the * acts as a wildcard symbol and a single file containing all of the pseudogenome sequences pasted one after the other is created. Both .fa and .fasta files are sequence files, but the extension is useful for distinguishing files with single (.fasta) and multiple (.fa) sequences

You should now have a file containing 24 taxa each 2.5MB long. Most of the sites in this alignment will be conserved and not provide useful information for phylogenetic inference, so we will shorten the alignment by **extracting the variable sites** using the program **snp-sites**

snp-sites -o All_snps.aln All.aln

Step 3: Build a phylogenetic tree from the SNP data in your genome alignment

Now you will build a phylogenetic tree from the SNP alignment that you created in the last step. There are a lot of programs for building phylogenetic trees, and here we are going to use one called RAxML which evaluates trees based on maximum likelihood.

The reference is:

A. Stamatakis: "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models". In *Bioinformatics*, 2006

Like all programs, RAxML has requirements for the format of input files. Your **All_snps.aln** file is multifasta format and RAxML requires **phylip format**, so open the file in **seaview** and save it as phylip format under the name **All_snps.phy** by typing

seaview All_snps.aln then doing File > Save As > Format > Phylip(*.phy)

😣 🗐 🗉 🛛 All_snps.aln		🛛 🖨 🔲 All_snps.aln
File ∇ Edit ∇ Align ∇ Prop	s ∇ Sites ∇ Species ∇ Footers ∇ Search: Goto:	File \(\nabla\) Edit \(\nabla\) Align \(\nabla\) Props \(\nabla\) Sites \(\nabla\) Species \(\nabla\) Footers \(\nabla\) Search:
Open Ctrl+O Open Mase Open Phylip Open Clustal Open MSF Open NEXUS Import from DBs	Seq:1 Pos:1 1 [5008_7#11 Lineage 1] Trefacerprocessor and considering and considering and considered and a second and a s	Sel=0 1 Seq:22 Pos:41 41 [5236 7#10 Lineage 3] S008 7#11 Lineage 1 GCCCAGCCATCGCCCATCGCCCCATCGCCCCCCCCCCCC
Save Ctrl+S Save as Save selection Save prot alignmt	סימום שליין ארב אבבע הבריקט אבדי האינדי האידי האידי האידי האידי האידי האידי האידי אידי	Argentina/ CCC Chile/ CCC Colombia/ CCC Costa_Rica/ CCC Peru/ CCC

Then, back at the command line, run **RAxML** by typing the following:

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

Recall that with a single iteration of a maximum likelihood method you risk recovering a tree from a local maximum, which means it might not be the best one. This can be avoided by running multiple iterations with different starting points (we can't do that now because of time). The addition of multiple runs is done by adding the following flag to the command.

-N 20 would run the program with 20 different starting trees (which is typically enough to find a problem if one exists)

Step 4: Interpret your phylogenetic tree

Open you final tree file (**RAxML_result.STm**) in FigTree and midpoint root it by selecting **Tree > Midpoint Root**.

You will now need to give this file that you have saved in FigTree to your georeferencers in your group. Go to **File > Export trees >** select **Newick** file format. Remember to save your file with a **.nwk** suffix so that you know what type of file it is.

Interpret your phylogenetic tree by first taking some time to make some general observations:

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

Then, using the relationships with the known lineages, define each of your isolates as belonging to lineage 1, 2, 3, 4 or Other and pass the information on to your georeferencer. A picture of your tree as well as your general observations about it should go into your team presentation. Take some time to make figure(s) you are happy with and create a pdf picture file by selecting **File > export PDF**



General Information

In this role, you are responsible for the investigation of antimicrobial resistance in isolates from your isolates. You will correlate the phenotypic metadata with the genetic information contained in the isolates using a local assembly approach with **ARIBA**. ARIBA, Antimicrobial Resistance Identifier by Assembly, is a freely available tool that can be installed from the ARIBA github repository. This tool required a FASTA input of reference sequences, which can be a multi-fasta file or database of antibiotic resistance genes or non-coding sequences. This **database** will serve as one of your inputs and the other is **paired sequence reads**. ARIBA reports which of the reference sequences were found, plus detailed information on the quality of the assemblies and any variants between the sequencing reads and the reference sequences.

We have installed ARIBA in the virtual machine. You will download the CARD database (<u>https://card.mcmaster.ca/home</u>) for resistance detection for your samples, however other databases can be installed.

Further information and installation instructions are detailed in the github wiki page: <u>https://github.com/sanger-pathogens/ariba/wiki</u>. The data can then be visualised using Phandango, an interactive also freely available tool to visualise your outputs <u>http://jameshadfield.github.io/phandango/</u>.

Step 1. Run ARIBA

Step 2. Visualise outputs (phandango.csv and .phandango.tre) in Phandango

Step 3. Compare resistance gene present with metadata

Step 4. Summarise your findings in text and screen shots for the presentation

Step 1: Run ARIBA

On the command line, navigate to your group folder in the **Group_task_Georeferencing** folder. To run ARIBA you will need to download and format the database. Type:

ariba getref card out.card

Next you will need to format the reference database for ARIBA. Type:

ariba prepareref -f out.card.fa -m out.card.tsv out.card.prepareref

Next you will need to run local assemblies and call variants, type:

ariba run out.card.prepareref reads_1.fastq.gz reads_2.fastq.gz out.run

The command should take about 5 minutes per sample. Be patient and wait for the command prompt (denoted by a \$ sign).

Next you will need to summarise the data from several runs. These are included in newly generated folders. You will combine the data in **report.tsv** files.

ariba summary out.summary out.run1/report1.tsv out.run2/report2.tsv out.run3/report3.tsv

Three files will be generated, a .csv file with the summary of all the runs and two .phandango files. You will need to drag and drop the out.summary.phandango.tre and out.summary.phandango.csv into the Phandago window.

To understand the whole picture you will need to run ARIBA for ALL 24 samples. To save time, we have already done this for all 24 samples. Please use the files in the ariba_reports folder for subsequent steps. We suggest you run the analysis for a few samples to get an idea of the results.





On the left hand side is a dendogram of the phylogenetic relationship of the resistance data and the strains. 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 (<u>https://card.mcmaster.ca/home</u>) 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 in a file names '**01.filter.check_metadata.tsv**' produced when you **prepared** your database (p. 11). Consult the report.tsv of the particular sample of interest for the gene names. You can open both .tsv files in excel.

ARD					
e or Download Copyright & Disclaimer		Search			
The Comprehensive Antibiotic	Resistance Database				
A bioinformatic database of resist	ance genes, their products and associated p	henotypes.			
3598 Ontology Terms, 2346 Refe	rence Sequences, 867 SNPs, 2160 Publication	ons, 2272 AMR D	etection Models		
Browse	Analyze	Dowr	nload		
The CARD is a rigorously	The CARD includes tools	CARD	data can be		
curated collection of	for analysis of molecular sequences including	downlo of form	paded in a numbe	er e	
determinants and	BLAST and the	is avai	lable as a	c	
associated antibiotics,	Resistance Gene	comma	and-line tool. Vie	N	
organized by the	Identifier (RGI) software	change	es in CARD		
Antibiotic Resistance	for prediction of	curatio	on.		
gene detection models.	homology and SNP				
gene detection models.	models				

Some general points to consider are:

- Does the presence of the gene correlate well with the phenotypic results?
- Is it the same in multiple isolates that share the resistance?
- Do you think it is vertically or horizontally transmitted?

Step 4: Summarise your findings for the presentation

Coordinate with your other team members to investigate the relationship of your resistance with where the isolates lie in the phylogenetic tree (that the Tree builders produced) and in the region of England and Wales (Georeferencer).

Consolidate your findings into some slides for the presentation and ensure the georeferencer produces a map of the distribution of resistance to complement your work.

General Information

Geo-referencing information from pathogens can provide insight into the processes that drive their epidemiology. This can be used to infer whether single introductions of a pathogen have occurred followed by local evolution (as in the *S. sonnei* in Vietnam story described in the introductory talk and Holt *et al*, *Nature Genetics*, 2013) or whether it transmits frequently across borders. It can also indicate regions affected by antimicrobial resistance.

In this role, you are going to use the metadata provided and the tools **spatialepidemiology.net** and **Microreact.**

In this role, you will complete the following steps:

Step 1. Identify the global positioning coordinates (longitude and latitude) of the addresses where the isolates were collected

Step 2. Create a map of the metadata of the isolates



To start, open the metadata file for the strains is in the inbuilt spread sheet program on the virtual machine and note the address column, as well as the two empty global positioning columns.

In **Step 1** you will be locating isolates based on their **Address** and filling out the information in the Latitude and Longitude columns.

Eventually you will obtain phylogenetic relationships for your samples from the Tree Builders in your team, which you will use as the input for **Microreact**. For now **get started with Steps 1 and 2**.

Step 1: Identify the longitude and latitude of the isolate addresses

Open a browser window and navigate to <u>www.spatialepidemiology.net</u> to obtain latitude and longitude coordinates for your addresses.

Click on the Create User Maps option on the right hand column (red arrow)

Then click on the 2nd tab **Batch Geocode addresses** and copy and paste the address column from your metadata file into this field. Click **Start geocoding**.



spatialepidemiology.net
Create User maps
Map Latitude/Longitude Lookup Batch Geocode addresses Simple Map Creation Advanced Map Creation
Batch Geocoding - Here you can enter a list of addresses and each is geocoded and latitude / longitude values returned.
Please enter your list of addresses one per line - PLEASE READ INSTRUCTIONS BEFORE DOING SO
North West, UK West Midlands, UK Wales, UK London, UK South East, UK London, UK
London, UK South East, UK
South East, UK
Geocoding results - Results are displayed in TAB delimited format allowing you to copy and paste directly into Excel. Six columns are given details of which can be found here. Address I Latitude I Longitude I Accuracy I Number of Addresses Returned I Address or error code.
North West, UK 41.5545253 - 67.33737550UU0001 9 2 2135 W 35th AVE, Gary, IN 46408, USA
London, UK 51.5073509 -0.1277582999998223 4 1 London, UK
London, UK 51.5073509 -0.12775829999998223 4 1 London, UK
London, UK 51.5073509 -0.12775829999998223 4 1 London, UK
London, UK 51.5073509 -0.12775829999998223 4 1 London, UK
West Midlands, UK 52.4750743 -1.829833000000078 3 1 West Midlands, UK
West Midlands, UK 52.4750743 -1.829833000000078 3 1 West Midlands, UK
North West, UK 48.6655294 -123.40820329999997 9 2 2212 Harbour Rd, Sidney, BC V8L 2P6, Canada
North West, UK 41.5545253 -87.3373/550000001 9 2 2135 W 35th Ave, Gary, IN 46408, USA
West Midianids, UK 52.4/30/43 -1.8296330000000/76 3 I West Midiands, UK
Wates, UK 52.1506007 -3.75371170000020 2 1 Wates, UK
South East, UK 0 0 0 Unknown Address: No corresponding geographic location could be found for the specified address.

As described on the website, this returns the address geocoding in six columns. The first three: Address, Latitude and Longitude are what we are after to update our metadata file.

You are also given a measure of address accuracy (for how specific the address is) and multiple addresses where a single one could not be specified.

Copy and Paste this information directly from the field into the spread sheet program and **manually curate** (i.e. decided between the options for each one) the address until you have a single latitude and longitude for each isolate.

Then, update the latitude and longitude columns in your metadata file.

Step 2: Create a map of the metadata of the isolates

Although <u>www.spatialepidemiology.net</u> is a complete geo-referencing tool, we are going to use some of the added functionality available in **Microreact** to visualize and explore your trees and metadata.

Microreact enables you to visualize phylogenetic relationships of isolates linked to geographic locations. Dynamic visualization of the data with interactive map, tree and metadata windows. <u>http://microreact.org</u>

C https://microreact.org/showcase ~ *Microreact = Microreact allows you to link, visualise and explore your data using trees, maps and timelines. Streptococcus pneumoniae Salmonella Typhi Y-chromosome Human PMFN2 Phylogeny Croucher NJ et al. 2014. Wong V et al. 2015. Hallast P et al. 2015. ographical analysis of the vnamics during tree hursts into lea

To prepare for the next step, save your updated metadata file with GPS locations as a .csv by doing **File > Save as > metadata.csv**

Read the instructions on how to set format your metadata file to visualise in microreact. This is vital for the next steps.

Note: You can obtain more HTML colour codes at http://htmlcolorcodes.com/

You will need a **NEWICK (.nwk)** file from the tree builders and the **.csv** metadata file you have just saved for the next step.

.csv file URL .nwk file URL One data file (.csv or .tsv) is required an CON .csv file?	nd a tree file (.nwk or .newick) is optional. TINUE
.nwk file URL One data file (.csv or .tsv) is required an CON .csv file?	nd a tree file (.nwk or .newick) is optional.
One data file (.csv or .tsv) is required at CON	nd a tree file (.nwk or .newick) is optional.
.csv file?	TINUE
.csv file?	nucle file 2
	.riwk lile?
an id column with a valid identifier for every row, which n full stops or commas:	This is your tree file which must be in valid Newick format. Every leaf label must correspond to an identifier that is specified in the <u>id</u> column of your dat
ititude and longitude columns. You can find the cation using <u>this service</u> .	file The number of labels in the $\pi_{\rm NW}$ file must match the number of identifiers within the $\pm {\rm d}$
ate a new project	
a_microreact	
your project (briefly):	<i>,</i>
	<i>h</i>
/our email address:	
	full stops or commas: tituted and longitude columns. You can find the cation using this service. ear, month and day columns. Attended and the project name of your project: * a_microreact your project (briefly):

You can include a different name for your projects and a brief description if you would like.

Leave 'project website' section blank and your email is optional. Then 'create project'.



Look at the distribution of your isolates across England and Wales when coloured by:

- **Clade** how are the isolates distributed? Are there any patterns you can see to the distribution? What factors might be driving the distribution?
- Antimicrobial resistances coordinate with your team antimicrobial resistance investigator for this are there patterns to any of the resistances? And is this related to clades?

Take snap shots of the images and report your findings in the group presentation. Some examples are below.

Other geo-tagging resources

Pathogens do not respect borders and global travel is increasingly frequent. For this reason the effective tracking and tracing of pathogens internationally is more important then ever. The analysis of your *S*. Typhimurium isolates tells us about how the pathogen behaves on a city-wide scale. To see if the epidemiology and resistance patterns you observed in your HCMC translate to the global scale, we need effective collaboration. For the geo-tagging recourses mentioned below, you can use either your own geographical data or data from the course for practice.

Other free geo-tagging resources.

WGSA: Is a web application for the processing, clustering and exploration of microbial genome assemblies. You can upload your assemblies and accompanying metadata to view assembly stats and view other metadata. <u>https://www.wgsa.net/</u>

EpiCollect: Is a freely available web and mobile app tool that is used for data collection (questionnaires), using multiple mobile phones and the data can be centrally viewed using Google maps/ tables and charts. <u>www.epicollect.net</u>

Phylocanvas: Metadata in binary format can be displayed next to the tree leaves by uploading a .csv file together with the tree file. <u>http://phylocanvas.net</u>

CartoDB: Much like the Google maps exercise you completed, CartoDB allows the user to map and analyze location data . This tool can take multiple file formats as input e.g. XLS, CSV and SQL amongst others. <u>https://cartodb.com/</u>

DISCLAIMER: All the locations and dates of the Salmonella isolates are fictitious and solely for educational purposes. No data was collected from Public Health England.

End of module...

ANY QUESTIONS? Please feel free to ask at any time!