GNA5031 Applied Session 3

Case study - Phylogenetic reconstruction

1. Background

The application of wastewater-based epidemiology (WBE) to support the global response to the COVID-19 pandemic has shown encouraging outcomes. The accurate, sensitive, and high-throughput detection of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in municipal wastewater is critical for wastewater-based epidemiology. The orgin the the virus is valuable to inform public health measurements during the global response to the pandemic.

In this case study, we will adopte phylogenetic techniques that was discussed during this week's lectures to trace the origin of SARS-CoV-2 using genomes that were obtained from international flights that landed in Brisbane, Australia in June, 2021. In order to achieve this goal, we will analyze these genomes together with 500 reference genomes that were reported from other parts of the world.

alt

For more information of wastewater-based analysis of SARS-CoV-2, please refer to:

- https://pubs.acs.org/doi/full/10.1021/acs.estlett.1c00408
- https://pubs.acs.org/doi/full/10.1021/acsestwater.2c00083

FOR INSTRUCTORS: students can locate these papers through Monash network.

2. Learning Objectives:

At the conclusion of this session, students are expected to:

- Understand a genome file (with extension ".fasta")
- Align genomes using MAFFT
- Trim genomes using TrimAL
- Infer phylogenetic tree using IQTREE
- Visualize phylogenetic tree using ITOL

3. Prerequisite:

3.1 Virtual machines (VMs)

You have been provided credentials that will allow you to access these virtual machines 24/7. To access any remote server, we use the terminal and the ssh command. If you are on a mac or linux based PC, you will have terminal in-built. If you are on a windows-based PC, please download PuTTY. To sign up for a VM please enter your name next to a line in this Google Doc: https://docs.google.com/spreadsheets/d/1H9k6jPfAR96m_0iKUA_sbLZEUzDmlME8yHlfLmAKTow/edit?usp=sharing

Note: You cannot share a VM, please jot down your virtual machine credentials.

Accessing virtual machines using **PuTTY** (Windows user)

If you are using Windows, you will access your virtual machines using PuTTY. In the hostname (or IP address) box, enter the hostname that you were provided. Ensure the connection type is SSH. Click open. You will be prompted to enter your username and password in the terminal window. Enter your credentials and click enter. Note you will not see the characters as they are typed. You are now in your home directory on the VM.

Accessing virtual machines using the **terminal** app (Mac user)

To access your VM using the terminal app on Mac, search for the terminal app using the search bar in the top right-hand corner of your computer. Type the following command following the convention of `ssh username@hostname` where `username` is your username at Monash, and `hostname` is the hostname provided for this course. Click enter, you will be promted for a password. Enter your Monash password and click enter. Note you will not see the characters as they are typed. You are now in your home directory on the VM.

Example:

```
ssh -x gnii0001@gna5031s1-gnii0001-01.rep.monash.edu
```

FOR INSTRUCTORS: students have successfully logged in.

3.2 The Monash VPN

Due to security constraints, our virtual machines can only be accessed from within the Monash network. If you are offsite, you will need to first connect to the Monash Virtual Private Network (VPN). Detailed instructions can be found on the following website: https://www.monash.edu/esolutions/network/vpn

Install the VPN software and test your connection.

FOR INSTRUCTORS: students can connet through VPN.

3.3 Register at ITOL

Note: Please register for a free account prior to the practical session, at: https://itol.embl.de/

FOR INSTRUCTORS: students can register at ITOL and login with credentials.

4. Hands-on component one

Note : This is on your local PC (i.e. not with virtual machine)

4.1 Prepare local copy of data.

Press **Download ZIP** under **Code** from the following page: https://github.com/GreeningLab/GNA5031_applied3.git Decompress

FOR INSTRUCTORS: students can download and decompress the specified data files

Install **Seaview** for sequence and alignment visualisation: https://doua.prabi.fr/software/seaview

FOR INSTRUCTORS: students can download and install Seaview (Both Windows and Mac versions are available from this website)

Install **Sublime Text 2** for working with text: https://www.sublimetext.com/2 **Note**: Not restricted to a single text editor; if you have BBEdit from session, feel free to use it.

FOR INSTRUCTORS: students have a test editor such sublime text or BBEdit installed)

4.2 Visualize sequence data using Seaview

Under filder data_pregenerated:

Drag and drop `input.fasta` into Seaview Observe nucleotides and colors.

4.3 Visualize aligned data using Seaview

Drag and drop `input.msa.fasta` into Seaview Observe nucleotides and colors.

4.4 Visualize aligned and trimed data using Seaview

Drag and drop `input.msa.trimmed.fasta` into Seaview Observe nucleotides and colors.

FOR INSTRUCTORS: help students to drag and drop the sequence files into Seaview `input.fasta` alt

`input.msa.fasta` alt

`input.msa.trimmed.fasta` alt

4.5 Inspect generated phylogenetic tree file using a text editor

Inspect `input.msa.trimmed.treefile`

FOR INSTRUCTORS: help students to open this treefile with a text editor

`input.msa.trimmed.treefile` alt

5. Hands-on component two (virtual machine except last step)

Note: replace user_name with your own user name in from the following scripts

5.1 access to virutal machine, setting up directory for analysis

```
ssh -x user_name@gna5031s1-user_name-01.rep.monash.edu
```

Make sure you start in your home directory using tilde "~", which indicates home directory in Linux systems

```
cd ~
```

FOR INSTRUCTORS: help students login to virtual machine as instructed

Download course material using `git clone`

```
git clone https://github.com/GreeningLab/GNA5031_applied3.git
```

FOR INSTRUCTORS: help students run the codes to obtain materials for this section as instructed

Install Mamba, and related softwares Mamba is a package manager for the Python programming language that aims to be a faster and more reliable alternative to the popular package manager, conda. Mamba uses the same package and environment specification formats as conda, so it is fully compatible with existing conda environments and packages. In simpler words, it is tool for software installations.

Need to login again

```
ssh -x user_name@gna5031s1-user_name-01.rep.monash.edu
```

```
cd ~/GNA5031_applied3/
mkdir -p tools
cd tools
wget "https://github.com/conda-forge/miniforge/releases/latest/download/Mambaforge-$(uname)-$(uname -m).sh
bash Mambaforge-$(uname)-$(uname -m).sh -p ~/session3/tools/mamgaforge
```

Note: yes to all prompts during mamba installation. Also, need to exit and re-login to take effect

```
exit
ssh -x user_name@gna5031s1-user_name-01.rep.monash.edu
```

FOR INSTRUCTORS: help students to login again and install Mamba as instructed

Create an isolation environment for phylogenetic analysis, install software MAFFT, TrimAL and IQTREE

```
mamba create -y -n phylogenetics
mamba activate phylogenetics
mamba install -c bioconda iqtree # press y and enter to confirm changes
mamba install -c bioconda mafft # press y and enter to confirm changes
mamba install -c bioconda trimal # press y and enter to confirm changes
```

Test whether installation is successful

```
iqtree -h
mafft -h
trimal -h
```

You're ready to go if software information are displayed.

FOR INSTRUCTORS: help students to install the above software using the codes provided, and test their installations

5.2 Align sequences

Sequence alignment is a fundamental step in many bioinformatics applications, including phylogenetic inference and others. It allows us to identify regions of similarity and difference between two or more sequences. By aligning sequences, we can identify conserved regions, mutations, insertions, and deletions, among others. This information is useful for understanding the evolutionary relationships among the sequences, identifying functional domains, and detecting genetic variations that may be associated with disease or other phenotypes.

Common tools for alignment: MAFFT (https://mafft.cbrc.jp/alignment/software/)

Go to data directory

```
cd ~/GNA5031_applied3/
```

```
mafft --preservecase --auto --reorder --thread -1 input.fasta >
input.msa.fasta
# 14m49.445s tested
```

FOR INSTRUCTORS: help students access the data directory and operate mafft using the above codes. Help students to move forward using pre-generated data if didn't work

5.3 Trim alignment

Trimming alignment is a common step in the process of inferring phylogenetic trees from molecular sequence data. The main reason for trimming alignment is to remove any poorly aligned or ambiguous regions in the sequence data that may affect the accuracy of the phylogenetic inference. Common tools for trimming: trimal (http://trimal.cgenomics.org/)

```
trimal -in input.msa.fasta  -out input.msa.trim.fasta -automated1
# 0m8.611s tested
```

FOR INSTRUCTORS: help students operate trimal using the above codes. Help students to move forward using pre-generated data if didn't work

5.4 Phylogenetic inference

Tree inference, also known as phylogenetic inference, is the process of reconstructing the evolutionary relationships among different organisms or groups of organisms based on their molecular or morphological characteristics. The resulting tree structure is called a phylogenetic tree, and it represents the evolutionary history of the group under study. The goal of tree inference is to reconstruct a tree that best explains the observed similarities and differences among the sequences or traits, while minimizing the number of evolutionary changes required to explain the data. A common tool is IQTREE (http://www.iqtree.org/).

```
iqtree -s input.msa.trim.fasta -alrt 1000 -bb 1000 -m TEST -nt 4
# 2m48.720s tested

-s for input
-arlt for Replicates for SH approximate likelihood ratio test. The SH
(Shimodaira–Hasegawa) test is a statistical test used to evaluate the
support for different branches in a phylogenetic tree. Don't mind it for
now.
-B for Replicates for ultrafast bootstrap (1000 is a commonly used
number).
-m for model select. TEST is by testing several models.
```

When all analysis is done, exit analysis environment and exit virtual machine for good practice

```
mamba activate phylogenetics
exit
```

FOR INSTRUCTORS: help students operate iqtree using the above codes, and exit virtual machine as intructed.

5.5 Visualization of phylogenetic trees (exit virtual machine, return to your own computer)

iTOL (Interactive Tree Of Life) is a web-based tool that allows users to visualize and explore phylogenetic trees and other hierarchical data sets. It was developed by Ivica Letunic and Peer Bork at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany, and has become a popular tool in the field of evolutionary biology and comparative genomics.

With iTOL, users can customize the display of their tree using a range of options, such as changing the layout, adding labels and annotations, and highlighting specific clades or branches. The tool also supports the integration of additional data sets, such as functional annotations or environmental information, which can be overlaid onto the tree to provide a more comprehensive view of the relationships between different organisms.

In addition to visualizing phylogenetic trees, iTOL can also be used to explore other types of hierarchical data sets, such as gene ontologies, protein domains, and metabolic pathways. The tool has a user-friendly interface and is freely available for use online, making it accessible to researchers and educators around the world.

Webpage: https://itol.embl.de/

4.5.1 Obtain the generated tree file

Note use pre_generated file if not ready For Mac user:

```
cd ~/Desktop
mkdir tree_file
cd tree_file
scp user_name@gna5031s1-user_name-
01.rep.monash.edu:/GNA5031_applied3/data/input.msa.trim.treefile .
```

For Windows user, use the pre-generated tree file - `input.msa.trim.treefile`

FOR INSTRUCTORS: help students to obtain either pregenerated or self-generated tree file

4.5.2 Upload to ITOL

Go to ITOL webpage (https://itol.embl.de/) press the following:

My Trees => Tree upload (drag and drop) => click on `input.msa.trim.treefile` to enter the tree

FOR INSTRUCTORS: help students to login to ITOL, drag and drop the tree file

4.5.3 Tree exploration

After tree being uploaded, explore:

- circular or rectangular tree format
- Labels and Lable options
- Advanced options
- Export options

Annotate tree lables with pre-generated file `color.strip.txt` Explore the following sections:

- Mandatory settins
- Actual data

Questions

- What is the meaning of `MW240742.1 rgba(104,2,63,0.7) North America`?
- What is the meaning of `RFPL_1 rgba(255,215,0,0.7) RFPL_1`?

FOR INSTRUCTORS: help students open the annotation file in a text editor. FOR INSTRUCTORS: First part (MW240742.1 and RFPL_1) means the tree leave, second part (rgba(104,2,63,0.7)) means color

to be applied; while the last part shows either this is our own sample or reference sequence.

Desired output

alt

4.5.4 Figure export

Export a pdf file for the rendered tree

Further exploration of tree styles at: https://itol.embl.de/gallery.cgi

FOR INSTRUCTORS: help students to explore, annotate and export the tree.