Variant Calling and Phylogenomics

Developed by: A/Prof. Torsten Seemann, Dr Kristy Horan, Dr Mun Hua Tan
Institution: University of Melbourne, Victoria, Australia
Date: 2025-10-06
Contact: munhua.tan@unimelb.edu.au

Learning Objectives

By the end of this session, you should be able to:

- Identify variants (SNPs, insertions, deletions)
- Generate a core genome SNP alignment
- Build a phylogenetic tree
- Visualise the tree.

The Toolkit

In this tutorial, we will be using:

- Snippy
- FastTree
- Microreact

Snippy

Used for rapid variant calling from raw sequencing data (FASTQ) of haploid genomes and for generating core genome SNP alignment. Snippy uses bwa-mem for read mapping and FreeBayes for variant calling.

FastTree

Used to build maximum-likelihood phylogenetic trees from aligned DNA or protein sequences. It computes local support values with the Shimodaira-Hasegawa test and produces trees in Newick format.

Microreact

A web-based tool for interactive phylogenetic tree visualisation. It allows you to upload data files (e.g., trees, metadata), explore relationships, and share/export visualisations.

Tutorial

1. Introduction

Variant calling and phylogenetic analysis are essential tools for understanding the genetic relationships among microbial or viral isolates.

Variant calling identifies single nucleotide polymorphisms (SNPs) and other genetic differences (e.g., insertions and deletions) relative to a reference genome. These differences are then used to construct phylogenetic trees to illustrate how closely related different isolates are. These methods are valuable for outbreak investigation, transmission tracking, and evolutionary studies.

---

2. Setup - Install Toolkit

> ⚠ **Important:** Take care at this point to make sure you are aware of which directory you are currently in and where your data is stored.

2.1. Check your current directory

```
pwd
```

Expected output:

```
/home/jovyan
```

If it does not look like the expected output, run:

```
cd ~
```

Your terminal prompt should now look like this (look for the ~ after your username):

```
[yourusername:~] $
```

2.2. Install the environment

You will need to get the new environment that we will be using in this session. To do this, run the installation script:

> **Note:** This is one long command.

```
bash
/nbt_main/k8sfs/pvs/manual/workbench/share/Environments/install_module_21.
sh
```

Activate the environment:

```
conda activate snps_phylo-env
```

Confirm that the installation has worked as expected:

```
snippy --help
FastTree
gotree --help
snp-dists -h
```

---

3. Setup - Get the Data

3.1. Check the data

In addition to installing the environment and tools, the previous step also retrieved and extracted the data for you.

For this tutorial, you have been provided with:

- paired-end sequence files for 5 samples (`*.fastq.gz`)
- a reference genome file (`reference.gbk`)

Check that the files are present:

```
ls module_21/*
```

You should see the following files:

```
module_21/reference.gbk

module_21/Strain_001:
Strain_001_nextseq_R1.fastq.gz  Strain_001_nextseq_R2.fastq.gz

module_21/Strain_002:
Strain_002_nextseq_R1.fastq.gz  Strain_002_nextseq_R2.fastq.gz

module_21/Strain_003:
Strain_003_nextseq_R1.fastq.gz  Strain_003_nextseq_R2.fastq.gz

module_21/Strain_004:
Strain_004_R1.fastq.gz  Strain_004_R2.fastq.gz

module_21/Strain_005:
Strain_005_nextseq_R1.fastq.gz  Strain_005_nextseq_R2.fastq.gz
```

3.2. Set up a results directory

You should have a directory were you can keep all of your analysis results. Below is a suggested format for doing this. Please note if you take a different approach the commands listed below will need to be modified accordingly.

Make a new folder to keep all the results for this module:

```
mkdir -p module_21/results
```

---

4. Run Analysis

This section will walk you through the process of identifying SNPs, generating a core genome SNP alignment, building an approximate maximum-likelihood tree for 5 samples, and visualising the tree.

4.1. Identify samples

Inspect your data and determine the names of the samples based on the FASTQ file names. Sample names are typically labeled as part of filename prefixes before the read-pair identifiers (e.g., _R1, _R2).

For example, for the files `Strain_001_nextseq_R1.fastq.gz` and `Strain_001_nextseq_R2.fastq.gz`, the sample name could be `Strain_001`.

4.2. Run Snippy (variant calling)

Snippy requires:

- A reference genome (FASTA or GENBANK format)
- Single- or paired-end sequencing reads from one or more isolates
- An output folder to put results in

Run `snippy` for each sample. Example command for one sample:

```
snippy --cpus 8 --outdir module_21/results/Strain_001 --ref
module_21/reference.gbk --R1
module_21/Strain_001/Strain_001_nextseq_R1.fastq.gz --R2
module_21/Strain_001/Strain_001_nextseq_R2.fastq.gz
```

Snippy generates many files reporting the identified variants, including table of variants (e.g., `.tab`) and consensus genome sequences (e.g., `.consensus.fa`). Visit the Snippy GitHub repository for detailed information on the various output files and formats.

To look at the SNP table (Note, press '`q`' to exit the view by `less`):

```
csvtk pretty -t module_21/results/Strain_001/snps.tab | less -S
```

To look at the consensus sequence (Note, press '`q`' to exit the view by `less`):

```
less -S module_21/results/Strain_001/snps.consensus.fa
```

> ➡️ **Task:** Repeat Snippy for the remaining 4 samples.

4.3. Run Snippy (core genome SNP alignment)

Run `snippy-core` to generate a core genome SNP alignment files across all samples:

> **Note:** This is one long command.

```
snippy-core --prefix module_21/results/core --ref module_21/reference.gbk
module_21/results/Strain_001 module_21/results/Strain_002
module_21/results/Strain_003 module_21/results/Strain_004
module_21/results/Strain_005
```

This will generate files with `core.*` prefix. To view a list of the files:

```
ls module_21/results/core*
```

You should see the following files:

```
module_21/results/core.aln
module_21/results/core.full.aln
module_21/results/core.ref.fa
module_21/results/core.tab
module_21/results/core.txt
module_21/results/core.vcf
```

Notable files include:

- `core.aln`: Core genome SNP alignment
- `core.tab`: Tab-separated table of core SNPs with alleles
- `core.full.aln`: Whole genome SNP alignment, including invariant sites

Run `snp-dists` to generate a SNP distance matrix for the core genome SNP alignment:

```
snp-dists module_21/results/core.aln > module_21/results/core.snpdist.tab
```

To look at the SNP table (Note, press 'q' to exit the view by less):

```
csvtk pretty -t module_21/results/core.snpdist.tab | less -S
```

4.4. Run FastTree (build phylogenomic tree)

Run FastTree to create a tree from the core SNP alignment:

```
FastTree -nt module_21/results/core.aln > module_21/results/core.nwk
```

This creates a phylogenomic tree core.nwk in Newick format, which you can visualise using tree visualisation tools.

Additionally, you can root the tree by specifiying an outgroup (e.g., the "Reference" lineage). Run gotree to reroot the tree, specifying "Reference" as outgroup:

```
gotree reroot outgroup -i module_21/results/core.nwk -o
module_21/results/core.reroot.nwk Reference
```

(Optional) Run gotree to draw the re-rooted tree in terminal (Note, press 'q' to exit the view by less):

```
gotree draw text -i module_21/results/core.reroot.nwk | less -S
```

4.5. Visualise the tree with Microreact

In this tutorial, we will use Microreact to visualise the tree. Note that other visualisation tools are also available such as FigTree and iTOL.

- Download the core.reroot.nwk file from JupyterLab to your local machine. This can be found in your module_21/results folder.
- Open a web browser and go to the Microreact website.
- Upload the core.reroot.nwk file.
- Click Continue to visualise and explore the tree.
- Explore the various features. For example:
    - display leaf labels
    - adjust alignment of leaf labels
    - zoom, pan, and collapse branches to explore the tree

- export high-resolution images
- share your `.microreact` file to allow others to view or build on the same project

> ⚠ **Caution:** Your Microreact file includes all underlying metadata, whether or not it is displayed in the tree. So, ensure that you remove or anonymise any sensitive data before sharing, unless you are okay for the person you're sharing it with to also have access.

❓ Questions to consider:

> *Feel free to contact us at the email above for feedback.*

1. Using your SNP distance matrix, report the number of core genome SNP differences between:
    - Strain_001 and the Reference
    - Strain_005 and the Reference
    - Strain_002 and Strain_003
    - Strain_002 and Strain_004
2. Compare your phylogenetic tree with the SNP distance matrix:
    - Do the tree relationships match the pairwise SNP distances?
    - Explain why you think they correspond (or do not).
3. Name two interactive features you explored in `Microreact` that you found helpful but could not achieve with the `gotree draw text` output.