Genome Biology and Genomics

Tom Williams, School of Biological Sciences
(tom.a.williams@bristol.ac.uk)

Session 5: Evolutionary inference from genomes

Monitoring of SARS-CoV-2 epidemiology in UK & NI (Wellcome Trust)

Overview: In the earlier practicals, you learned how to sequence, assemble and annotate genomes, and to map reads against them to study gene expression. In this practical, we'll look at comparing genomes to learn about population-level variation and the processes of evolution. We'll return to one of the lifeforms featured in Session 2 - the SARS-Cov-2 virus - and use variant calling and some light phylogenetics to identify and track new coronavirus variants.

Learning objectives: By the end of this practical, you will be able to map reads to a reference genome, evaluate the evidence for single nucleotide polymorphisms compared to a reference, and interpret phylogenetic trees in an epidemiological context.

Running jobs on BluePebble

If you plan to work through the practical on BluePebble, you should first log in in the usual way, then request an interactive job:

```
srun --mpi=pmi2 --account=bisc033844 --partition=teach_cpu --nodes=1 --ntasks-per-node=1
--time=01:00:00 --pty bash -i
```

This will launch an interactive bash session on one of the compute nodes, so you can run commands directly. **Whatever you do, do not run jobs directly on the head (login) node!**

This practical makes use of the following software, all of which is already installed on BluePebble:

bwa: module load bwa/0.7.17
samtools: module load samtools/1.19.2
bcftools: module load bcftools/1.19-openblas
snpEff: module load apps/snpeff/5.2f

Software documentation

NB: If you cannot find something, try Googling for it. The Github pages of software packages usually point to some documentation. Sources like Stack Overflow and Biostars can also be very useful.

Bwa: https://github.com/lh3/bwa
Samtools/bcftools: https://www.htslib.org/
snpEff: https://pcingola.github.io/SnpEff/

Further (optional) reading

Bwa paper: https://academic.oup.com/bioinformatics/article/25/14/1754/225615
Samtools paper: https://pubmed.ncbi.nlm.nih.gov/19505943/
SNP calling statistics: https://pubmed.ncbi.nlm.nih.gov/21903627/

Part 1. Identify single nucleotide polymorphisms in a new SARS-CoV-2 variant

1. **Index the SARS-CoV-2 reference genome.** You can use the assembly you produced in Session 2, or you can use the NCBI reference genome, which is provided in the "session4" folder on the course repository. To index the reference sequence using the bwa aligner, run

```
bwa index covid19_wuhan_reference.fasta
```

2. **Map the reads** from the first variant to the reference genome using bwa.

```
bwa mem covid19_wuhan_reference.fasta variant1_R1.fastq variant1_R2.fastq -o
variant1.sam
```

This command writes the alignment of reads to the reference to a Sequence Alignment Map (SAM) format file, the standard for storing this kind of information in genomics. A compressed version of this kind of file, a BAM (binary alignment map) format file, is used for variant calling analysis.

3. Convert the SAM file to a compressed BAM file, then sort it (needed for variant calling):

```
samtools view -S -b variant1.sam > variant1.bam

samtools sort variant1.bam > variant1_sorted.bam
```

4. **Visualise/inspect the alignment of reads from the variant against the reference sequence using the samtools tview option.** You will need to index the sorted BAM file first (`samtools index variant1_sorted.bam`). Then:

```
samtools tview variant1_sorted.bam covid19_wuhan_reference.fasta
```

5. **Call variants present in the new sample, but not the reference.** We will use bcftools mpileup and bcftools call for variant calling. Note the use of the unix pipe (|) to pass information from one method to the next.

```
bcftools mpileup -f covid19_wuhan_reference.fasta variant1_sorted.bam |
bcftools call -mv -Ob --ploidy 1 -o calls.bcf
```

6. Output a readable version of the binary calls file:

```
bcftools view calls.bcf > calls.vcf
```

7. Predict the functional/coding-level consequences of the inferred variants using snpEff:

```
java -Xmx2g -jar /software/local/apps/snpEff/snpEff.jar
 -v NC_045512.2 calls.vcf > annotation.vcf
```

snpEff produces a HTML summary, an annotated VCF, and a summary of variants called per gene.

8. How many high-quality SNPs do you detect? What are the amino acid-level changes that are caused by these SNPs? In which genes do they occur?

9. Repeat the variant calling procedure for the second variant.

10. The two variants are not closely related, but both are estimated to be more transmissible than the original SARS-CoV-2 virus. Can you propose a hypothesis as to which SNPs may cause this increased transmissibility? (http://sars2.cvr.gla.ac.uk/cog-uk/)

Part 2. Phylogenetic origin of SARS-CoV-2 variants

In Part 2 of the practical, you will assemble the genomes of the two SARS-CoV-2 variants and determine their geographical origin using phylogenetics. This part of the practical is rendered fairly straightforward thanks to the Pangolin ("Phylogenetic Assessment of Named Global Outbreak Lineages", https://pangolin.cog-uk.io/) infrastructure for COVID epidemiology/phylogenetic placement.

1. Assemble a genome for each of the two variants.
2. Use the Pangolin webserver (https://pangolin.cog-uk.io/) to determine the place of origin for each.
3. To which lineage do these variants belong?
4. How accurate were your inferences about the causative substitutions in Part 1?