MSA construction & SNP distance computation

Bharkbhoom Jamesia & Pakorn Aiewsakun

Pornchai Matangkasombut Center for Microbial Genomics

Department of Microbiology, Faculty of Science, Mahidol University

# Introduction

From the previous session, you should now have a set of GVCF files for your samples. In this practical, we will use ‘*GATK GenotypeGVCFs*’ to combine all individual GVCF files into one per-dataset/multi-sample joint-calling file in Variant Call Format (VCF). This approach to sample processing enables us to conveniently incorporate additional new sequence data to the pre-existing dataset.

After joint-calling, the output can then be easily processed into a simple single-nucleotide polymorphism (SNP) matrix, which can subsequently be converted into a multiple sequence alignment (MSA) for further analyses that may help you answer your specific questions.

We will do all of these in this session.

# Per-cohort variant calling (or joint-calling)

To joint-call samples, we will first merge all (pre-called) GVCF files generated by ‘*gatk HaplotypeCaller*’ into a single GVCF file using ‘*gatk CombineGVCFs*’, and then perform joint-calling by using ‘*gatk GenotypeGVCFs*’. For a very large dataset (n > 1000), GATK developer recommends another tool called ‘*gatk GenomicsDBImport*’ for this task. To do this, change your current directory to

*‘****/home/[user]/mtb\_wgs\_analysis\_workshop/’***

and run the following commands.

*gatk CombineGVCFs \*

*-V /per-sample\_gvcf/sample\_01.g.vcf \*

*-V /per-sample\_gvcf/sample\_02.g.vcf \*

*-V /per-sample\_gvcf/sample\_03.g.vcf \*

*-V /per-sample\_gvcf/ …g.vcf \*

*-V /per-sample\_gvcf/ …g.vcf \*

*-V /per-sample\_gvcf/your\_last\_sample.g.vcf \*

*-O /per-cohort\_gvcf/cohort.g.vcf.gz*

*gatk GenotypeGVCFs \*

*-V /per-cohort\_gvcf/cohort.g.vcf.gz \*

*-O /per-cohort\_vcf/cohort.vcf.gz*

With these two commands, we should now have a per-cohort variant data in VCF format ‘*cohort.vcf.gz*’.

# Variant site filtering

Similar to what we did in the sequence data quality control practical yesterday, as part of the data cleaning process, we will now do site filtering to remove low quality variant sites from our VCF file, which could bias our downstream analyses.

Settings for variant filtering should be adjusted according to the dataset and questions that we are investigating. When mapping short sequence reads of MTB samples to the reference genome, we typically exclude repetitive regions (such as PE/PPE genes, phages, and drug resistant genes), which are prone to read mapping errors. These regions account for ~10 percent of the H37Rv genome. For today, we are going to filter our variant data using two scores:

*‘QualByDepth (QD)’* score, which is the variant confidence score (*‘QUAL’*) normalized (or divided) by unfiltered allele depth (*‘AD’*), and

‘*RMSMappingQuality (MQ)’* score*,* which is the root mean square mapping quality of reads across all samples at the site.

We will use two gatk tools for this. First, ‘*gatk VariantFiltration*’, which will flag, or annotate, low-quality sites in the VCF file based on QD and MQ scores by modifying their ‘*FILTER*’ field. We then will filter out the annotated low-quality sites (as well as and non-SNP sites, e.g. INDEL or MIXED sites) by using ‘*gatk SelectVariants*’. To do this, change you current working directory to

*‘****/home/[user]/mtb\_wgs\_analysis\_workshop/per-cohort\_vcf/***

and run the following commands.

*gatk VariantFiltration \*

*-V joint-called.vcf.gz \*

*-O joint-called\_flaggedQD2MQ40.vcf.gz \*

*--filter-expression "QD < 2" --filter-name "QD2" \*

*--filter-expression "MQ < 40" --filter-name "MQ40"*

*gatk SelectVariants \*

*-V joint-called\_flaggedQD2MQ40.vcf.gz \*

*-O clean.vcf.gz \*

*--select-type-to-include SNP \*

*--exclude-filtered*

The first command will annotate sites with QD scores >=2 with *'PASS'*, and those with QD < 2 with *'QD2'*. Similarly, it will annotate sites MQ scores >=40 with *'PASS'*, and those with MQ < 40 with *'MQ40'.* The second command drop all ‘non-PASS’ (and non-SNP) sites from the file.

With these analyses, we should have a clean high-quality joint-calling VCF file

# Generating a multiple sequence alignment

Multiple sequence alignment is a starting point of many molecular sequence analyses. Next, we will create a SNP alignment from your clean high-quality joint-calling VCF file, using the ‘pipeline’ below. What it does is, basically, first make a new directory (with the program ‘*mkdir’*) for storing the outputs, generating a SNP matrix from your multi-sample VCF file (using the program ‘*gatk VariantsToTable’*), and then converting it into a standard format for sequence data storage, FASTA format (using a combination of several programs).

To do this, change your current working directory to

***‘/home/[user]/mtb\_wgs\_analysis\_workshop/per-cohort\_vcf/****’*

and run the commands below.

*mkdir -p ../msa*

*gatk VariantsToTable -V clean.vcf.gz -O ../msa/mat -F POS -GF GT*

*cd ../msa*

*datamash transpose --output-delimiter=, < mat > tmp*

*sed -i '1d' tmp*

*sed $'s/^/>/;s/.GT,/\\\n/g;s/,//g;s/[\.\\*]/-/g' tmp > aln.fasta*

Now we have a SNP alignment that can be opened with any text editors (Notepad, Sublime, or MS Word) or alignment visualisers (such as Mega, Aliview, and BioEdit).

# Computing pairwise SNP distances

We can do many things with whole genome sequence data, such as identifying the lineage / sub-lineage of your samples, predicting drug resistance phenotypes, inferring evolutionary history, and investigating epidemiology and transmission dynamics the epidemic strains, for example. For today’s workshop, we will perform a quick and simple analysis – computing pairwise SNP distances from our SNP alignment.

Change your current directory to the directory ‘*/msa*’ and use the program ‘*snp-dists*’ to compute pairwise SNP distances. We will run it by using the default settings, which count only differences between bases A, C, G, and T, while ignoring ambiguous bases and gaps. The command is simple:

snp-dists aln.fasta > pairwise\_snp\_dist.tsv

The output ‘*pairwise\_snp\_dist.tsv*’ is a matrix table, which can be opened with any text editor program like Notepad.