Nextflow pipeline exercise

1) If it was possible to execute the pipeline, what would be the workflow of the pipeline (what is the pipeline designed to do)? - You can draw the workflow if you prefer

The pipeline is designed for variant calling starting from raw sequence data (FASTQ -> SAM -> BAM -> VCF). I executed the workflow on a small filtered fastq.gz file downloaded from 1000 genome.

As we see in Figure 1, we can visualize the workflow using the following command:

./nextflow run main.nf -preview -with-dag flowchart.png

The main steps of the workflow can be summarized as follows:

1) We start using a fastq.gz file and the genome reference and index.
2) We align the reads (fastq file) against the genome reference (genomeref), this outputs a sam file using the mimimap tool.
3) We sort and transform the sam file to a bam file using samtool, which outputs a bam file.
4) We apply two variant calling techniques (tools): deepvariant and clair3. Each will produce VCF files (the identified variants).

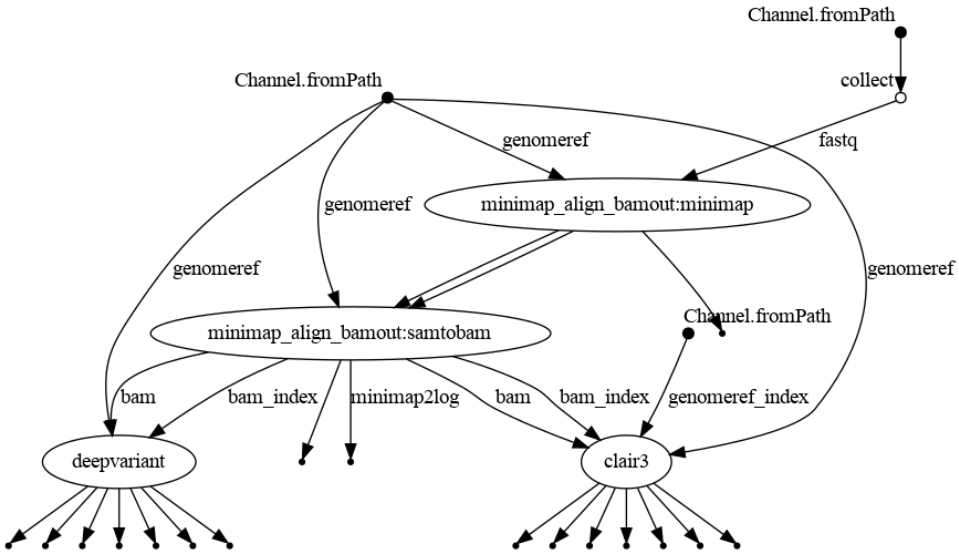

Figure 1: Structure of the workflow

What would be the name of the main output folder? What names would output subfolders, generated inside the main output folder, have? What files would every output subfolders have inside?

In the nextflow.config, we define two parameters sampleid and outdir. Therefore, the main output directory will be named Example_out_ test_run. This will be inside the ./data directory.

```
sampleid            = "test_run"
outdir              = "./data/Example_out_${params.sampleid}"
```

In the samtools process, we defined the publishDir path, which is the directory where the output of the process will be copied to. Therefore, inside the Example_out_ test_run (outdir), I will have a single folder called test_run (sampleid)

```
publishDir path: "${params.outdir}/${params.sampleid}/${task.process.replaceAll(':', '_')}/", mode: 'copy'
```

The output from the workflow will therefore be stored in the test_run directory. We can visualize its content using the tree command, as shown in Figure 2. Also, Figure 3 shows the terminal output.

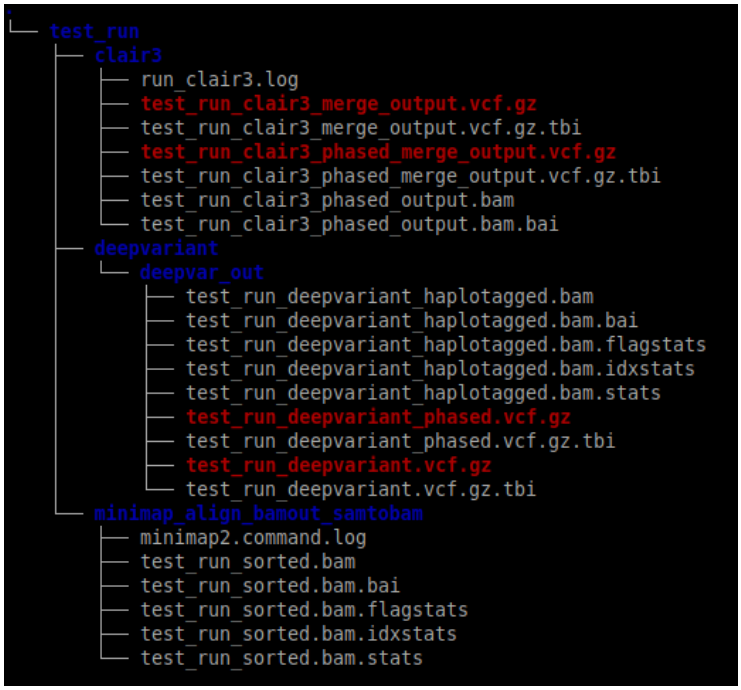

Figure 2: Structure of the output directory of the workflow

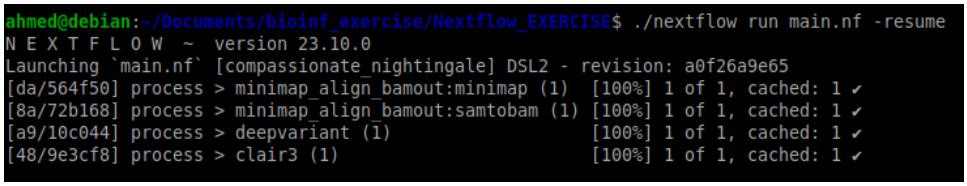

Figure 3: The terminal output of running the workflow