Final task

NGS data analysis/pipelining module

by Fynn Freyer
(but shamelessly stolen from W. Dabrowski)

2025-07-11

Background

You are a bioinformatician in a workgroup that works with Hepatitis delta, a small human-pathogenic virus. One of your collegues analyses samples from molecular surveillance of Hepatitis delta and thus needs to compare genomes of viruses from those samples to a reference genome.

Fortunately, the NGS lab at your workplace is able to provide the complete (short) genomes from the sequencing. However, custom analysis is not part of the service.

Required analysis

For the comparison, your colleague requires the following:

- An alignment of all genomes from a sequencing run to a reference genome
- Cleanup of the alignment (removing positions from the alignment that have a low quality)
- Simple visualization of the alignment

Fortunately, you have already been able to go through the analysis manually with you colleague and you already agreed on which tools and which parameters to use. Now, you just need to automate the task in order to be able to run it quickly and reproducibly whenever new sequencing results come in.

Pipeline development task

The nextflow pipeline you develop should:

- Take an NCBI GenBank accession as commandline parameter (`--accession`) and downloads the FASTA file for that accession. The default accession (if none is given on the commandline) should be M21012
- Take a directory of genomes in FASTA format from your colleague[1]
- Combine all of those single FASTA files into one FASTA file[2] – this is necessary for the alignment tool used in the next step
- Run the mafft aligner on the combined FASTA file
- Clean-up the resulting aligned FASTA file using trimal, using the `-automated1`[3] operation
- Publish both the cleaned up alignment and the HTML report generated by trimal into an output directory

[1] Do this by taking a directory containing the FASTA files with the genomes. You can use the FASTA files in the supplied `hepatitis/` folder as input.

[2] Using Nextflow. Not by manually copying them.

[3] There are different options, your colleague and you looked at the outputs all of these generate, and your colleague decided that that's the one they want.

You can publish more files, but the two files that your colleague is interested in are just the resulting FASTA file and the HTML report (which includes a visualization of the alignment) from trimal. You can find an example visualization[4] as `alignment_trimmed.html` next to this file.

Important:

- Implement the whole pipeline in nextflow
- Use conda packages for your dependencies
  - `mafft=7.525`
  - `trimal=1.5.0`
  - `entrez-direct=24.0`
- Submit your solution through GitHub:
  - Create a new repository
  - Commit and push your code (everything necessary to run the pipeline) (**at least** `main.nf` and `nextflow.config`)
  - Either make the repository public and send me a link, or make it private and invite me (**FynnFreyer?**) to it

Tips

Downloading data

Downloading a FASTA file for an accession from GenBank can be done using the following commands:

```
esearch -db nucleotide -query "M21012" | efetch -format fasta > "M21012.fasta"
```

In this example, the accession M21012 is used – please adapt this to your needs, and please note that in this command, the accession is given twice: Once in the URL (defining which accession to download), and once as output filename in the redirect.

Combining files

You can combine multiple files on the commandline using cat. For instance, to combine “file1.txt” and “file2.txt” into “both.txt”, you could run:

```
cat file1.txt file2.txt > both.txt
```

or, using wildcards (assuming “file1.txt” and “file2.txt” are the only text files in the directory):

```
cat *.txt > both.txt
```

or, using operators. The choice is yours.

[4] This visualization is generated from the output generated from the accession M21012 and the provided example data.