Bonus Exercises

Bioinformática Aplicada às Ciências da Vida

Exercises, 05/12/25

Let's first look at yesterday's mapped data from SARS-COV2 Delta variant.

1) Copy the reference file, the reference index file, the gff3 file, the mapped bam file and the bam index to your local disk, and open them with IGV. In windows use Filezilla, in Linux open a new terminal on your local machine, navigate to to where you want to save the results, and copy with `scp`.

```
cd ~/
mkdir sars_cov_results
cd sars_cov_results
scp -r bnevado@motoo:/home/bnevado/bacv/sars_map/sars-cov-2-reference.fa ./
scp -r bnevado@motoo:/home/bnevado/bacv/sars_map/sars-cov-2-reference.fa.fai ./
scp -r bnevado@motoo:/home/bnevado/bacv/sars_map/SARS-COV-2.gff3 ./
scp -r bnevado@motoo:/home/bnevado/bacv/sars_map/delta_mapping/delta_map.bam ./
scp -r bnevado@motoo:/home/bnevado/bacv/sars_map/delta_mapping/delta_map.bam.bai ./
```

2) Find the position of the Spike protein gene in the genome, using the GFF file provided. If you are running Windows, do this on the server, otherwise you can use the local copy.

```
grep spike SARS-COV-2.gff3
```

Output: our Spike protein gene starts on position 21563 and ends on position 25384 of chromosome NC_045512.2

```
NC_045512.2 RefSeq  gene    21563   25384   .   +   .   ID=gene-GU280_gp02;Dbxref=GeneID:43740568;Name=
NC_045512.2 RefSeq  CDS 21563   25384   .   +   0   ID=cds-YP_009724390.1;Parent=gene-GU280_gp02;Dbxref
```

3) Now navigate to the Spike gene location in IGV, and report the positions where the Delta variant is different from the reference.

SNP CALLING FROM MAPPED DATA

4) Can we find these variants in a better way? This is bioinformatics after all.. Let's create a VCF file from the bam file, covering only the Spike protein gene. Get back on Motoo, and execute this command inside the delta_mapping folder.

```
bcftools mpileup -Ob -f sars-cov-2-reference.fa -r NC_045512.2:21563-25384 delta_map.bam | bcftools cal
```

5) Have a look at the resulting VCF file *delta.vcf*. Does it match what you see on IGV?

Now let's look at the Omicron variant, which caused so many troubles.

1) Perform the same things you did in yesterday's class, but on the *omicron* dataset (look for it in the `/home/data/sars` directory of *Motoo*). Copy the files locally and don't forget to verify the files' integrity (sha256sum) after copying them. Do the quality control, trimming, re-do quality control and create the bam file and the VCF file.

>> The main difference here is that the *omicron dataset* was prepared using a different protocol (Nextera), so you may have to perform some trial and error to get the trimming right.

2) Why was this variant so scary (tip: check out variants in the Spike protein gene)?

Data manipulation in Linux

1) For this task we will perform a PCA in R from the environmental data contained in this file on *Motoo* : /home/data/exercises/data.tar. Create a folder for this analysis in your home folder, copy the file there and uncompress it.

2) A kind researcher has already prepared a R script that will perform the PCA analysis, but you need to prepare the input file so it matches the expected format. First, copy the R script in /home/data/exercises/PCA.R to the folder you just created. Open it with a text editor (e.g. nano) and check what it does.This PCA script expects as input a file named input.txt, and will write an image into a file called output.png.

3) You need to prepare the input file so it has the expected characteristics, namely:
* The input file must be named input.txt
* The first column must contain the grouping variable (in this case, the Habitat type) and the name of this column must be "group"
* All the other columns must contain only environmental data
* Fields in this file must be separated by tabulations

4) Once you have your input file ready, you can run the script with `Rscript PCA.R` in the folder with the *PCA.R* and the *input.txt* files.

5) Assuming all went well, you should have a image (output.png) in your folder. Copy it to your local computer and have a look. Do the environmental variables help separate the different types of habitat?

6) Now do a new PCA, but this time using only data points from Forests and Swamps.

>> Hint1: There are many ways to solve this problem. The following commands would be useful: *cut*, *paste*, *grep* and *sed*.
>> Hint2: For *sed*, you need to use the 'global' option to replace all occurrences of a given character in the line (otherwise, just the first occurrence is replaced). Also, the symbol `\t` can be used in sed to represent a taulations. Example: `sed 's/\t/,/g'` would replace all tabs in each line with commas.

Quality control and mapping of sequencing reads Pt2

1) In this exercise, you are tasked with performing quality control on paired-end Illumina sequencing data for *E. coli* strain "O118/O151:H2" and mapping the clean data using Bowtie2 against the *E. coli* K-12 reference genome. Subsequently, your goal is to explore genomic differences, specifically focusing on genes associated with fluoroquinolone resistance. You will find the required data under `/home/data/mapping`. The FASTA file contains the K-12 strain reference genome, and the `FASTQ` files contain the sequencing data of the O118/O151:H2 strain. There is also a K-12 GFF3 file in the directory. Don't forget to verify your copy's integrity against the known checksums!

>> Control and Mapping:

>> Perform quality control on the paired-end Illumina sequencing data, ensuring that it is clean and ready for analysis.
>> Map the clean data using Bowtie2 against the *E. coli* K-12 reference genome.

>> Identification of Fluoroquinolone Resistance Genes:

>> Explore the GFF3 file of the E. coli K-12 genome to identify fluoroquinolone resistance genes.
>> Focus on the gyrA gene, which is known to be associated with resistance to fluoroquinolone antibiotics.

Analysis and Comparison: * Analyze the mapped data to determine the variations, if any, in the gyrA genes between the E. coli "O118/O151:H2" strain and the K-12 reference strain.

Is this too much to search using *IGV*? Try generating a *VCF file* instead:

```
bcftools mpileup -Ob -f reference.fa -r scaffold:start-end input.bam | bcftools call -vmO v > output.vc
```

>> *Hint*: The VCF format may look overwhelming at first, but keep in ming that after the "header" lines (starting with "#"), it is easy to find what you are looking for. The first column is the chromosome name (only one in this case), and the second column is the position in the chromosome where a variant has been identified. The rest you will have to figure out from the header description and the documentation linked above (have fun!).

Ohhh, another bonus exercise!

Data manipulation in Linux Pt2

We are now interested in comparing the number of genomic variants present **along the genome** in the Delta and Omicron strains, compared to the original virus strain.
To help doing this, another kind researcher (or perhaps it was the same?) prepared a R script called `GenomeScan.R`, which you can copy from `/home/data/sars/fullvcfs/GenomeScan.R`.
To run this script, you will need to create two input files named `delta.positions` and `omicron.positions`. Inside these files you should have all the variable positions on the genomes of the two strains. You can get this information from the VCF files, but make sure you prepare new VCF files that span the entire genome, not just the spike protein region as done before, and then use grep/sed etc... to prepare your input files.

Was the spike protein gene in omicron particularly variable (compared to the rest of the genome)?