Bioinformatics Practical Part 2

Mapping Oxford Nanopore reads to the reference genome

MP235

SS 2022

Introductory information.

Question 1

What is Oxford Nanopore Sequencing Technology? How does it work? What read length distribution do you expect? What error rate do you expect?

What materials you need:

1. Reference genome sequence

   Question 2

   What is the reference genome file format?

   Question 3

   How are reference genome sequences obtained?

2. File with unaligned reads

   Question 4

   What is the file format unaligned reads are stored in?

3. Mapping software

   Question 5

   Give three examples of software that can be used to map Oxford Nanopore reads

You will be working in a conda environment. There is no need to understand all the details, but you need to run the command below to activate it

```
source /mnt/practical2022/miniconda3/bin/activate pgb_env
```

This will ensure you can access all the necessary software for the practical.
You can find the reference genome in `/vol/volume/practical2022/part2/files`

```
$ cd /mnt/practical2022/part2/files
$ ll
```

The file name is Express617_v1.fa.
Let's count the sequences in the file:

```
$ grep "^>" Express617_v1.fa | wc -l
```

Question 6

What does the ^ mean?

Question 7

How many sequences are there in the file?

Question 8

How many chromosomes does *B. napus* have?

Let's have a look at chromosome names in the file

```
$ grep "^>" Express617_v1.fa | head -n 20
```

Question 9

Can you see all the chromosomes?

Question 10

Why are there more sequences than chromosomes?

Question 11

How would you display just the top five sequences?

Now let's obtain the raw sequencing data. Sequencing reads associated with existing publications are stored in the NCBI Sequence Read Archive.

Question 12

What is NCBI? How big is the NCBI SRA database?

Every read dataset has an associated entry and a graphical summary. Go to the following page:
https://www.ncbi.nlm.nih.gov/sra/?term=SRR15731030

Let's explore the webpage.

Question 13

What can you learn about the reads? How many are there? What is the mean length?

You can find download links under: 'Data access'. You can download the files directly using wget.

There are other options to download using the SRA toolkit (https://github.com/ncbi/sra-tools).

Please DO NOT execute the following command! The file is big and will take a long time to download.

```
$ wget https://sra-pub-run-odp.s3.amazonaws.com/sra/SRR15731030/SRR15731030
```

The files are downloaded in .sra format, which can be converted to the fastq format we need, using the SRA toolkit `fasterq-dump` tool (https://github.com/ncbi/sra-tools/wiki/HowTo:-fasterq-dump). We will not be doing that. A ready fastq file is available in `/mnt/practical2022/part2/files`. The file name is: `SRR15731030.5x.fq`

Please create your working directory

```
$ mkdir /mnt/practical2022/part2/analysis
```

and move the reference sequence and sequencing reads to your working directory.

```
$ mv /mnt/practical2022/part2/files/*.fa /mnt/practical2022/part2/analysis
$ mv /mnt/practical2022/part2/files/*.fq /mnt/practical2022/part2/analysis
```

Question 14

What does the star stand for?

Let's have a look at the `SRR15731030.5x.fq` file. Let's count the number of lines.

```
$ cd /mnt/practical2022/part2/analysis
$ wc -l SRR15731030.5x.fq
```

Question 15

With the knowledge of the fastq format can you infer how many reads are there in the file?

We will use the `n50` command (https://github.com/quadram-institute-bioscience/seqfu/wiki/n50) to get some statistics for the fastq file.

```
$ n50 --format tsv SRR15731030.5x.fq
```

Question 16

What does the output look like?

Question 17

What is the N50 value of our sequences?

Question 18

What is an N50 value?

Now we will map the Oxford Nanopore Sequencing reads to the reference genome.

Question 19

Explain what read mapping is?

We will use minimap2 for this purpose: https://github.com/lh3/minimap2
Many bioinformatics software tools are now available on github, where often you can also find tool descriptions, sample usage etc. Go ahead and explore the minimap2 github page.

Question 20

What kind of information can you get from the github page?

Often it is possible to get information about software usage just by typing its name into the terminal.

Type minimap2 into the terminal.

Question 21

What do you see? What is the software version installed? Is it the newest version available on github? Why is it important to use up-to-date software versions?

Look at the minimap2 sample usage displayed.

Question 22

What is a target, what is a query?

Question 23

Try to design a mapping command by substituting our reference and read file names for target and query.

Question 24

Is this a complete command?

We are using Oxford Nanopore reads so we need to supply minimap2 with additional information. Type the software name again and look at all the parameters (options).

Question 25

Which parameters do you think may be relevant? Note that we want our mappings in SAM format. Try to design a complete command.

```
$ minimap2 -ax map-ont --MD -t 8 -o SRR15731030.minimap2.5x.sam \
> Express617_v1.fa SRR15731030.5x.fq
```

Note that the backslash (\) indicates that the command is not finished and will be continued on the next line. The ">" at the beginning of the next line is the prompt given by the shell, expecting a continuation of the previous line. Do not confuse this with the same symbol being used to redirect standard output into a file. Breaking up the command like this is not necessary in the terminal and is only used so that the command fits on this page. This command will take a while to run. We have an already prepared sample SAM file (SRR15731030.minimap2.5x.chrA01.sam) in the /mnt/practical2022/part2/files folder. To stop a command press ctrl+c (strg+c).

To reduce size this file has been produced by running the above command and then extracting only reads mapping to chrA01. Have a look at the SAM file using the head command.

```
$ mv /mnt/practical2022/part2/files/SRR15731030.minimap2.5x.chrA01.sam .
$ head SRR15731030.minimap2.5x.chrA01.sam
$ head -n 2000 SRR15731030.minimap2.5x.chrA01.sam
```

The SAM file format is one of the most important file formats in bioinformatics. Have a look at the specification: https://samtools.github.io/hts-specs/SAMv1.pdf

Question 26

What is a header in sam files? How many sequences are listed in the header of SRR15731030.5x.chrA01.sam? What do they correspond to?

Question 27

How many columns are there in the alignment section? What information do these columns store?

Now let's have a look at the alignment section of the file (skip the header)

```
grep -v "^@" SRR15731030.minimap2.5x.chrA01.sam | head
```

SAM files are in textual (human readable) format. For processing they can be converted to binary BAM files, which take up much less space and can be sorted and indexed for faster access.
Sam files are operated on using samtools. Have a look at the webpage: http://www.htslib.org/doc/samtools.html

There are quite a lot of commands. Probably the most popular are view (allows conversion, subsetting etc), sort and index.

Let's convert our SAM file to a sorted, indexed BAM file.

```
$ samtools view -bh  SRR15731030.minimap2.5x.chrA01.sam > \
> SRR15731030.minimap2.5x.chrA01.bam
$ samtools sort -o SRR15731030.minimap2.5x.chrA01.sorted.bam \
> SRR15731030.minimap2.5x.chrA01.bam
$ samtools index SRR15731030.minimap2.5x.chrA01.sorted.bam
```

Type `samtools view` into your terminal.

Question 28

What do the options `-b` and `-h` mean in the command above?

It is possible to extract only a portion of a bam file, for example if we want to have a look in the viewer on our local machine, but don't want to copy over the entire file.
Type the following to see reads in human readable format:

```
$ samtools view SRR15731030.minimap2.5x.chrA01.sorted.bam \
> chrA01:100000-200000 | head
```

Type the following to get reads in machine readable format.

```
$ samtools view -bh  SRR15731030.minimap2.5x.chrA01.sorted.bam \
> chrA01:1000000-2000000 > SRR15731030.minimap2.5x.chrA01.sorted.1Mb.bam
$ samtools index SRR15731030.minimap2.5x.chrA01.sorted.1Mb.bam
```

Let's have a look at this small section of bam file in a viewer.
The most popular viewer for genomes/annotations/bam files is IGV https://igv.org/. Today we will be using a web version. https://igv.org/app/

Demonstration on how to use IGV.

Often you want to have a look at mappings close to genes of interest. Gene annotations are stored in gff3 format. Have a look at the specification here: https://www.ensembl.org/info/website/upload/gff3.html