BIO-312: Practical on Environmental DNA Analysis

Session 1: Kākāpō eDNA

By Katherine Delevaux and Prof. Sebastian Waszak
Laboratory of Computational Neuro-Oncology, Swiss Institute of Experimental Cancer Research, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Switzerland
https://www.epfl.ch/labs/upwaszak/

Background

In the week 6 lecture, the kākāpō, a ground-dwelling parrot endemic to New Zealand, was mentioned as an endangered species that has been studied using eDNA from soil samples. In this session, you will work with the actual eDNA data published as part of Non-invasive real-time genomic monitoring of the critically endangered kākāpō *eLife* (2023). (Also available on Moodle.)

To summarize, the authors collected about 5-10 grams of soil from the display sites of two kākāpō individuals and another soil sample from the feeding station of another kākāpō individual. From these three soil samples, DNA was extracted and prepared for nanopore sequencing. The data was made publicly available through NCBI.

Part 0: Connect to virtual machine

We will use a virtual machine to access software and datafiles. Launch VMware Horizon Client if you already have it or download from here:
https://vdi.epfl.ch/?includeNativeClientLaunch=true

Log into the virtual machine called SVUbuntuUpwaszak2025.
Throughout the session, run the code through Terminal on the virtual machine.

Your personal EPFL storage is mounted on the virtual machine *Desktop/MyFiles*. For files to persist after closing the virtual machine, make sure to place any files here. Files left on the Desktop will be destroyed upon logging out.

Please make sure to log out of the virtual machine when you are done.

Part 0.5: Linux Crash Course

Bioinformatics and computational biology require the use of Linux command line for running various software and viewing files.

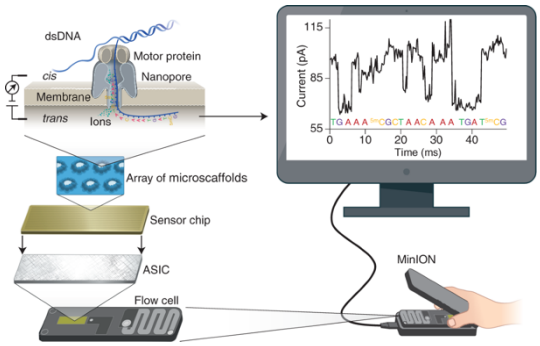
Open up the Terminal and try out the commands listed below in black to familiarize yourself. Lines starting with '#' and in green are comments and are not run.

```
# list the files and directories in the current working directory
# (directory = folder)
$ ls

# change working directory to the specified location
# cd /where/you/want/to/go
$ cd Desktop

# '..' stands for the parent directory and '.' for the current working directory
$ cd ..
$ cd Desktop

# copy a file to a new location and/or filename
# cp old_directory/original_file.txt new_directory/new_file.txt
$ cp /opt/data/Bio312/kakapo_ref_genome.fna ./kakapo_ref_genome.fna

# print out the entire contents of a file
$ cat kakapo_ref_genome.fna
# ctrl-c to stop

# print out the first 10 lines of a file
$ head kakapo_ref_genome.fna

# '|' allows you to link two commands together
$ cat kakapo_ref_genome.fna | head

# view the entire contents of the file in a scrollable manor (press 'q' to close)
$ cat kakapo_ref_genome.fna | more

# instead of printing the output to Terminal use '>' to save it to a file
$ head kakapo_ref_genome.fna > kakapo_ref_genome_first_10_lines.fna

$ cat kakapo_ref_genome_first_10_lines.fna

# running software on the command line generally follows the rough format
# software command -arg1 [value] -arg2 [value] input_file.txt output_file.txt
```

Part 1: Processing raw Nanopore DNA sequencing data

Nanopore sequencing measures changes in current through a nanopore as a nucleotide strand passes through.

Dorado is a package developed by Oxford Nanopore Technologies that uses a machine learning model to deconvolve the electrical signal into a strand of base pairs.



Wang et al. *Nature* 2021

Nanopore sequencing data is stored in POD5 file format. Let's use the Dorado basecaller to call nucleotide bases from nanopore signal. For this example, we use a small subset of human genome sequencing.

```
$ dorado basecaller fast /opt/data/Bio312/NA12878_DNA.pod5 > NA12878_DNA.bam

# head prints only the first 10 lines of output
$ samtools fastq NA12878_DNA.bam | more
```

1. *What is the format of the FASTQ file? How many lines per sequence?*
2. *How to FASTQ files differ from FASTA files?*

Part 2: Aligning sequencing reads

Now, we'll start with the kākāpō eDNA data. In this dataset, the base pairs were called from the POD5 files and the FASTQ files were made available. We will now align the FASTQ reads to the kākāpō reference genome. Minimap2 is a sequence alignment program developed for long-read sequencing.

In the directory `/opt/data/Bio312/`, `kakapo_ref_genome.fna` is the kākāpō reference genome downloaded from the NCBI. `SRR18208523.fastq.gz`, `SRR18208524.fastq.gz`, and `SRR18208525.fastq.gz` are the sequences from the three kākāpō sites sampled.

Let's start with `SRR18208524.fastq.gz` and use minimap2 to generate a minimap2 index of the reference genome then align our sequencing reads using dorado.

```
$ minimap2 -d ref.mmi /opt/data/Bio312/kakapo_ref_genome.fna
$ dorado aligner ref.mmi /opt/data/Bio312/SRR18208524.fastq.gz > SRR18208524.bam
```

Samtools is a common library for working with Sequence Alignment/Map (SAM) files and the compressed binary version (BAM). Let's use samtools to take a look at the alignment. Run the following code and use the samtools documentation and this site

https://broadinstitute.github.io/picard/explain-flags.html to understand what each line is doing.

```
$ samtools view SRR1820524.bam | head
$ samtools view -b -F 4 -q 7 SRR18208524.bam > SRR18208524_mapped.bam
$ samtools view -b -f 4 –threads 8 SRR18208524.bam > SRR18208524_unmapped.bam
$ samtools sort SRR18208524_mapped.bam > SRR18208524_mapped_sorted.bam
$ samtools index SRR18208524_mapped_sorted.bam

# wc -l counts how many lines are printed
$ samtools view SRR18208524_mapped_sorted.bam | wc -l
$ samtools view SRR18208524_unmapped.bam | wc -l
```

1. *What is the format of the SAM and BAM files? What information does each column contain?*
2. *What do the samtools flags -F and -q do?*
3. *How many reads were mapped to the kākāpō reference genome and how many did not map? Is this surprising?*

Now let's do some quality control on the alignments.

```
$ samtools view SRR18208524_mapped_sorted.bam | awk '{print length($10)}' | sort -
n > mapped_read_lengths.txt
$ samtools view SRR18208524_unmapped.bam " | awk '{print length($10)}' >
unmapped_read_lengths.txt

$ samtools view SRR18208524_mapped_sorted.bam | awk '{print $5}' > mapped_mapq.txt
```

1. *What is the range of mapping qualities?*
2. *What are the read lengths of mapped vs unmapped?*

Let's use R visualize the sequence lengths.

```
$ R
> read_lengths = read.table("mapped_read_lengths.txt")
> png("mapped_read_lengths_hist.png")
> hist(log(read_lengths$V1), main="mapped read lengths")
> dev.off()
> q(save= 'no')
```

Similarly, make bar plots for the unmapped read lengths and mapping quality.

Bonus: make a scatter plot of read length vs mapping quality for the mapped reads.

Part 3: Alignment Visualization

There's a nicer way to visualize the alignments with the kākāpō reference genome using IGV. To open IGV open another Terminal window and run:

```
$ /usr/local/bin/IGV_Linux_2.19.4/igv.sh
```

Select *Genomes > Load Genome from File > Desktop/Files/kakapo_ref_genome.fna*
Select *File > Load from File > SRR18208524_mapped.bam*

The alignment is sparse; therefore, you probably won't see many aligned reads. So, how can we see where in the reference genome the sequences aligned to?

In a new Terminal window run:

```
$ samtools coverage SRR18208524_mapped_sorted.bam
```

Based on the output above we see that some reads aligned to the contig NC_044277.2. Let's look at the exact location for reads that map to NC_044277.2.

```
$ samtools view SRR18208524_mapped_sorted.bam | head

# grep prints only the lines containing the specified pattern
# cut prints only selected columns
$ samtools view SRR18208524_mapped_sorted.bam | grep "NC_044277.2" | cut -f1-5
```

Notice the start position of the reads (4th column). Choose a base pair location where multiple reads start.

Next to the top left drop down that shows the reference genome, go to the next drop down and select the contig NC_044277.2. Then, using the next box to the right select the exact base pair location as follow NC_044277.2:base_pair.

On IGV you can click and drag to move along the chromosome.

1. *How are alternate alleles ("SNPs") marked?*
2. *How are DNA sequence deletions marked?*
3. *How are DNA sequence insertions marked?*

Repeat the same process and take a look at another section of alignments from another contig.

Part 4: Rinse and Repeat

Congrats! You have now processed environmental DNA of the SRR18208524 soil sample from a kākāpō nesting site.

Repeat parts 2 and 3 to process the other two soil samples: `SRR18208523.fastq.gz` & `SRR18208525.fastq.gz`. Note that these two files are significantly larger, so each of the steps will take longer.