BIO-312: Practical on Environmental DNA Analysis

Session 2: Kākāpō eDNA Part II

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. 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.

In the last practical session, you aligned DNA sequences from the soil samples to the kākāpō reference genome and visualized the alignment.

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 tools and software.

Open up the Terminal on the virtual machine and try out the commands listed below in black to familiarize yourself with Linux. 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 'return' to
scroll and '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 tools/software on the command line generally follows the rough format
# tool_name command -arg1 [value] -arg2 [value] input_file.txt output_file.txt
```

Part 1: Variant Calling

Variant calling is the process of identifying genetic variants by comparing genome sequencing data to a reference genome. Medaka is a tool developed by Oxford Nanopore Technologies to create consensus sequences and variant calls from nanopore sequencing data. We will use this tool to call DNA sequence variants from each of our soil sample alignments.

We will pick up working with our mapped and sorted reads generated last week. In the directory `/opt/data/Bio312/`, you'll find the `SRR*_mapped.bam` files for each sample. These were prepared as shown in the previous practical.

First, we perform variant calling with medaka. It can take some time, so be patient:

```
# activate the medaka virtual environment (prepare the machine to run medaka)
$ . /opt/medaka/bin/activate

# copy over the reference genome and reference genome index medaka will use
# '.' indicates the current directory (should be the Desktop)
$ cp /opt/data/Bio312/kakapo_ref_genome.fna .

# medaka_variant is the software and command, -i input_file, -r reference_genome,
-o output_directory
$ medaka_variant -i /opt/data/Bio312/SRR18208523_mapped.bam -r
kakapo_ref_genome.fna -o SRR18208523_medaka

# take a look at what files are in the output directory
$ ls SRR18208523_medaka
```

Let's take a look at the output files. BCFtools is a common package used for working with DNA sequence variants in the Variant Called Format (VCF) files and their binary compressed binary version (BCF):

```
# head -n N to print the first N lines instead of default 10
$ bcftools view SRR18208523_medaka/medaka.vcf | head -n 60

# grep -v prints all lines except those containing the string "#"
$ bcftools view SRR18208523_medaka/medaka.vcf | grep -v "#" | head

# wc -l counts how many lines are printed
$ bcftools view SRR18208523_medaka/medaka.vcf | grep -v "#" | wc -l
```

1. *What information is contained in the header of the VCF file? (lines starting with '#')*
2. *What information does each column of a VCF file contain?*
3. *Where is the quality score located in the VCF? How is variant quality calculated?*
4. *What is the total number of variants compared to the reference kākāpō genome contained in this alignment?*

Let's look at a nice summary of the variants:

```
$ bcftools stats SRR18208523_medaka/medaka.vcf | head -n 40
```

1. *How many of each different type of variant is there?*
2. *What is the difference between SNPs, MNPs, and indels?*

Repeat the previous steps sorting, converting, and variant calling the other two samples SRR18208524 and SRR18208525. The variant calling will be faster because the reference genome index file has already been created. (If you are running out of time, the SRR1820852*.medaka.vcf files are also provided in `/opt/data/Bio312` which you may copy to the Desktop and use in the next steps.)

Next, let's use **bcftools** to intersect the variants between samples. We first need to compress (zip) and index our VCF files:

```
# bgzip is a general command line program for compressing files
$ bgzip SRR18208523_medaka/medaka.vcf
$ bgzip SRR18208524_medaka/medaka.vcf
$ bgzip SRR18208525_medaka/medaka.vcf

$ ls SRR18208523_medaka

# bcftools index creates an index for bgzip compressed VCF files for random
indexing
$ bcftools index SRR18208523_medaka/medaka.vcf.gz
$ bcftools index SRR18208524_medaka/medaka.vcf.gz
$ bcftools index SRR18208525_medaka/medaka.vcf.gz

$ ls SRR18208523_medaka
```

1. *How were the file names changed when zipping the files?*
2. *What files were created in the second step?*

```
# bcftools isec creates intersection of the variants from the 3 input files
$ bcftools isec SRR18208523_medaka/medaka.vcf.gz SRR18208524_medaka/medaka.vcf.gz
SRR18208525_medaka/medaka.vcf.gz > variant_intersection.txt

$ head variant_intersection.txt
```

3. *What does the 5th column in variant_intersection.txt encode?*
4. *How many variants are present in only the second sample? Hint: use grep and wc -l.*
5. *How many variants are shared between all three samples?*

Part 2: Variant Visualization

Let's look at the alignments and variants of the samples on IGV. To open IGV open another Terminal window and run.

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

We need to sort and index the alignment files to view them on IGV:

```
$ samtools sort -O BAM -o SRR18208523_mapped_sorted.bam
        /opt/data/Bio312/SRR18208523_mapped.bam
$ samtools sort -O BAM -o SRR18208524_mapped_sorted.bam
        /opt/data/Bio312/SRR18208524_mapped.bam
$ samtools sort -O BAM -o SRR18208525_mapped_sorted.bam
        /opt/data/Bio312/SRR18208525_mapped.bam

$ samtools index SRR18208523_mapped_sorted.bam
$ samtools index SRR18208524_mapped_sorted.bam
$ samtools index SRR18208525_mapped_sorted.bam
```

Select *Genomes > Load Genome from File > Desktop/kakapo_ref_genome.fna*
Select *File > Load from File > Desktop/SRR18208523_medaka/medaka.vcf.gz*
Select *File > Load from File > Desktop/SRR18208524_medaka/medaka.vcf.gz*
Select *File > Load from File > Desktop/SRR18208525_medaka/medaka.vcf.gz*
Select *File > Load from File > Desktop/SRR18208523_mapped_sorted.bam*
Select *File > Load from File > Desktop/SRR18208524_mapped_sorted.bam*
Select *File > Load from File > Desktop/SRR18208525_mapped_sorted.bam*

Navigate to `NC044278.2:105372700`.

1. *Take a look at the alternative alleles in the .bam file compared to the called variants .vcf file. Is there support in the alignment file for the called variant?*
2. *Notice where there are alternative alleles in the .bam file, but no called variant at the position – does this make sense?*

Part 3: Taxonomic Classification

Last week, we saw that less than 1% of the soil sequencing reads aligned to the kākāpō reference genome. What could be the source of the rest of the DNA?

In this part, we will look at taxonomic classification using czid.org. It is an open-source cloud-based metagenomics pipeline.

Open up **czid.org** and make an account using your EPFL email address. Join the kākāpō project by giving your account email address to Katherine and she'll invite you to our shared project.

Looking at the homepage, notice that samples SRR18208523 and SRR18208525 were split into multiple parts due to their large size (the compressed files of unmapped reads are 7.9G and 6.0G respectively).

Take a look at the home page and answer the following questions:

1. *What does czid.org do to reduce computational burden of large sample files? (Compare the Total Reads to the Passed Filters counts.)*
2. *Which column provides an overview of the sequence quality?*

Click on a sample, in the top right *Download > View Pipeline Visualization*. Click on some of the pipeline steps to see a more in-depth description.

Slightly above the column names, you can switch the view to a taxonomical tree. Take some time to explore the taxonomical tree.

1. *Are there species that make sense given the sampling location?*
2. *Are there species that do not make sense given the sampling location?*
3. *Is there anything missing from the taxonomical tree you expected to be there?*