## **Stage 2**

## **Project 3: Run a simple NGS analysis pipeline**

In this section, you will implement a simple NGS analysis on a simple dataset

**Starting Datasets:**

* [Forward Strand](https://zenodo.org/records/10426436/files/ERR8774458_1.fastq.gz?download=1)
* [Reverse Strand](https://zenodo.org/records/10426436/files/ERR8774458_2.fastq.gz?download=1)
* [Reference](https://zenodo.org/records/10886725/files/Reference.fasta?download=1) Genome

**Proposed Pipeline**:

Download dataset (**wget**) => Quality Control (**FastQC**) => Trimming (**FastP**) => Genome Mapping (**bwa**) => Variant Calling (**bcftools/freebayes**)

Feel free to add software as you prefer.

**Let’s get bigger:**

Use your pipeline to analyze more datasets

**Reference**: <https://raw.githubusercontent.com/josoga2/yt-dataset/main/dataset/raw_reads/reference.fasta>

**ACBarrie**  
<https://github.com/josoga2/yt-dataset/raw/main/dataset/raw_reads/ACBarrie_R1.fastq.gz><https://github.com/josoga2/yt-dataset/raw/main/dataset/raw_reads/ACBarrie_R2.fastq.gz>

**Alsen**  
<https://github.com/josoga2/yt-dataset/raw/main/dataset/raw_reads/Alsen_R1.fastq.gz><https://github.com/josoga2/yt-dataset/raw/main/dataset/raw_reads/Alsen_R2.fastq.gz>

**Baxter**  
<https://github.com/josoga2/yt-dataset/raw/main/dataset/raw_reads/Baxter_R1.fastq.gz><https://github.com/josoga2/yt-dataset/raw/main/dataset/raw_reads/Baxter_R2.fastq.gz>

**Chara**  
<https://github.com/josoga2/yt-dataset/raw/main/dataset/raw_reads/Chara_R1.fastq.gz><https://github.com/josoga2/yt-dataset/raw/main/dataset/raw_reads/Chara_R2.fastq.gz>

**Drysdale**  
<https://github.com/josoga2/yt-dataset/raw/main/dataset/raw_reads/Drysdale_R1.fastq.gz><https://github.com/josoga2/yt-dataset/raw/main/dataset/raw_reads/Drysdale_R2.fastq.gz>

**Submission**:

* We look forward to receiving your final pipeline **script.sh** (you can use bash, snakemake, nextflow or any pipeline tool you know how to use).
* Alongside, create a **setup.sh** file that anyone can use to install all the tools needed for making the pipeline work.
* Make a **requirement.txt** file that simply lists all the tools you used
* Upload the 3 files to your team’s github repo. Each team member should have a folder and their folder should contain their 3 scripts.
* Copy the link to the team’s repo and paste it on HackBio Submission platform
* Finally, be ready to discuss your pipeline with everyone

**Resources**

* [Introduction to Whole Genome Sequencing and Variant Calling](https://www.youtube.com/watch?v=NxRECdxKP40)
* [Raw Sequence to Variant Calling Pipeline with FreeBayes](https://www.youtube.com/watch?v=gmJ6LteXAq0) (Hands-On)
* [Galaxy Tutorial for Variant Calling (with Code)](https://training.galaxyproject.org/training-material/topics/data-science/tutorials/bash-variant-calling/tutorial.html)
* [Using For loops in BASh](https://www.youtube.com/watch?v=T7hVOiTsSUU)
* [HackBio Video for loops for multiple datasets using FastP](https://youtu.be/HNE0VPZK8yM)