Bachelor project - implementation task

The third phase of the project is the implementation phase. In this phase, you will be doing a minor benchmark of the existing WGBS and ONT tools. The usual approach to detecting 5mC methylations in CpG motifs includes running the WGBS methylation calling pipeline to obtain calls that are considered the ground truth and then analyzing the ONT data with some existing methylation tools tailored for ONT signals. The results obtained with ONT tools and data are evaluated with respect to calls obtained through the WGBS pipeline. The work in this phase will be divided into two subphases, first focusing on the WGBS pipeline and results, second focusing on benchmarking two ONT tools.

Data

WGBS data

1. These reads are obtained by sequencing bisulfite-treated samples with Illumina sequencing technologies.
2. The reads are short (151 bases) and paired (both ends are sequenced simultaneously, therefore producing twice the amount of reads in the same period of time).
3. Types of reads:
    a. Native - some CpG positions are methylated, and some are not
    b. Negative control - no CpG positions are methylated
    c. Positive control - all CpG positions are methylated

ONT data

1. This data consists of raw signals obtained through Nanopore sequencing with r9.4.1 pore.
2. The lengths of reads vary and are usually in thousands of bases. The corresponding signals are therefore in tens of thousands of measurements.

Phase 1 - WGBS methylation calling pipeline

In this phase, you will be working on obtaining the ground truth for further analysis. The pipeline includes several steps and the end result is a list of positions in the reference genome and the corresponding number of methylated/non-methylated reports for each position.

WGBS Pipeline:

The pipeline consists of 4 main steps. In the first step, you perform quality and adapter trimming using **Trim Galore**. In the second step, you map the reads to a reference with **bwa_meth** to obtain .sam file with mappings. The .sam file is converted to .bam file, sorted, and indexed using **samtools.** In the third step, you remove duplicate calls to preserve the true distribution of calls per position with **Picard.** In the final step, you extract the methylation calls to a bedGraph file using **MethylDackel.** An example of the pipeline is listed below.

```
# Trim Galore
trim_galore -j 4 --fastqc --paired --retain_unpaired ${PAIR_ONE} ${PAIR_TWO} -o trim_g
alore

# Bwa-meth alignment
mkdir -p bwa-meth
bwameth.py --threads 64 \
    --reference ${REFERENCE} \
    trim_galore/*_R1_val_1.fq.gz trim_galore/*_R2_val_2.fq.gz > bwa-meth/aln.sam

# Convert to bam and analyze
samtools view -@ 64 -T ${REFERENCE} -bS bwa-meth/aln.sam > bwa-meth/aln.bam
samtools sort -@ 64 bwa-meth/aln.bam > bwa-meth/aln.sorted.bam
mv bwa-meth/aln.sorted.bam bwa-meth/aln.bam
samtools index -@ 64 bwa-meth/aln.bam
samtools flagstat -@ 64 bwa-meth/aln.bam > bwa-meth/aln_flagstat_report.txt

# Deduplication
mkdir -p picard
picard -Xmx128g MarkDuplicates I=bwa-meth/aln.bam O=picard/aln_dedup.bam M=picard/dedu
p_metrics.txt REMOVE_DUPLICATES=TRUE
samtools index -@ 64 picard/aln_dedup.bam

# MethylDackel
mkdir -p methyldackel
MethylDackel extract -@ 64 ${REFERENCE} picard/aln_dedup.bam -o methyldackel/bedGraph
```

You can check all the tools:

1. https://github.com/FelixKrueger/TrimGalore
2. https://github.com/brentp/bwa-meth
3. https://github.com/broadinstitute/picard
4. https://github.com/dpryan79/MethylDackel
5. http://www.htslib.org/

c. You will be running the pipeline on three datasets:

1. Negative control dataset - in this dataset, all methylations are erased through the PCR amplification, the samples are bisulfite treated and all cytosines are converted. The expected outcome of the pipeline for this dataset are reports of non-methylations in all CpG positions.
2. Positive control dataset - in this dataset, all cytosines are methylated by applying an enzyme. The expected outcome of the pipeline are reports of methylations in all CpG positions.
3. Native dataset - this dataset has some positions methylated and some non-methylated. There was no additional preparation of the sample (unlike for positive and negative datasets).

WGBS Analysis

Once you run the entire pipeline for all three datasets, you will additionally analyze the results for positive and negative control to evaluate how well the methylation calls match the expectations for such controls. You expect to obtain a distribution of scores reporting methylation in all CpG positions with high confidence for positive control. On the other hand, in the negative control case, you expect a distribution of scores reporting non-methylation in all CpG positions with high confidence. You will be analyzing control samples for Escherichia Coli and native samples for embryonic stem cell.

1. Controls
    a. Negative control - /mnt/share1_DrOc/sbakic/negative_control/original_reads/
    b. Positive control - /mnt/share1_DrOc/sbakic/positive_control/original_reads
    c. Native data - /mnt/share1_DrOc/sbakic/native_h1esc

    Native samples are sampled in several runs, you can analyze each run separately and combine the calls in post-analysis.