# ASSESSMENT REPORT

## INTRODUCTION

In this bioinformatics assessment, we delve into a set of simulated amplicon sequencing data obtained under two distinct experimental conditions. The data generated using R9.4.1 flow cells. My exploration includes an examination of true positive variants, an evaluation of the base caller model employed, and a comprehensive analysis of various quality metrics and coverage. The goal is to make an informed recommendation for the preferred condition to be utilized in future experiments. This report encapsulates our findings, reasoning, and evidence-based conclusions derived from the exploration of the provided data.

## METHODS

**Data Retrieval:** Simulated amplicon sequencing data under two conditions were obtained from the provided tar ball. The data includes fastq files for each sample, associated reference genome information and a file of true variants information.
**Data Exploration:** The contents of the fastq files, true positive variants file (true_positives.vcf), reference genome file (reference.fasta), and reference genome index file (reference.fasta.fai) were examined using Python, BioPython, and standard file reading techniques.
**Quality Assessment with FastQC:** The FastQC tool was employed to assess the quality of the sequencing data. For each condition, the first five (condition1) and last five (condition2) samples were subjected to FastQC analysis, and the generated reports were saved for further comparison. Later, these reports are used to make a comparative analysis in multiqc.
**Alignment and Read Metrics Calculation**: Read metrics were calculated using the bwa mem aligner and samtools. A temporary BAM file was generated for each sample, and metrics were obtained using samtools stats. The resulting read metrics reports were saved for each condition.
**Coverage Calculation**: Coverage was calculated using bedtools and the BAM files generated during the read metrics calculation. The bedtools genomecov command was applied to determine coverage, and the results were saved for each condition.
**Comparative Analysis**: Comparative analysis was conducted between conditions for both read metrics and coverage. Differences in the reports were identified using the diff command. Reads VCF file with true positive variants, filters them for two conditions ("condition1" and "condition2"), and calculates precision, recall, and F1 scores. It then displays these metrics for both conditions. “calculate_true_false_positives_negatives_for_condition” function was used to calculate true positives, false positives, and false negatives based on "PASS" and "FAIL" labels in the content of a VCF file and prints them to the console. **Epi2me workflow** analyzation was also used.

## RESULTS:

In the **analysis of true positive variants**, no significant differences were observed between the two conditions, as evidenced by consistent **precision** (1.0000), **recall** (0.0769), **F1 scores** (0.1429), 1 **true positive**, 0 **false positives**, and 12 **false negatives**. In the **FastQC results analysis, no overrepresented sequences** were identified in samples from both conditions. Regarding **sequence duplication levels,** in the first condition, the percentages were 0.8, 9.75, 9.95, 1.20, and 0.95, respectively. In the second condition, percentages were 1.20, 1.57, 1.53, 1.50, and 1.80. Both conditions exhibited failures in **per-sequence** and **per-base quality scores.** The analysis revealed notable differences in various **metrics**, as summarized in Table 1 below.

| Metric | Condition1 | Condition2 |
|---|---|---|
| Raw Total Sequences | 2000 | 3000 |
| Reads Mapped | 1999 | 3000 |
| Reads Unmapped | 1 | 0 |
| Bases Mapped | 495,362 | 1,683,132 |
| Bases Mapped (CIGAR) | 491,559 | 1,677,464 |
| Mismatches | 22,159 | 82,642 |

| | | |
|---|---|---|
| Error Rate | 4.507902e-02 | 4.926604e-02 |
| Average Read Length | 247 | 561 |
| Average 1st Frag Length | 248 | 561 |
| Maximum Read Length | 400 | 1297 |

Table1 Comparison of Metrics Between Condition 1 and Condition 2

## DISCUSSION/CONCLUSION

The thorough examination of **simulated amplicon sequencing data** generated under two conditions, utilizing **R9.4.1 flow cells** and the "**dna_r9.4.1_450bps_hac**" base caller model, has yielded valuable insights into the performance of each condition and the associated base caller model. The performance of the "dna_r9.4.1_450bps_hac" base caller model appears robust, with no significant discrepancies observed between conditions. The "dna_r9.4.1_450bps_hac" base caller model significantly influences sequencing data accuracy. Despite not being explicitly addressed in the provided data, considerations of its pros and cons are essential. On the positive side, the model consistently demonstrates high accuracy, particularly in precise variant calling, and is specifically tailored for R9.4.1 flow cells, ensuring compatibility and optimized performance. However, potential drawbacks include high computational intensity, dependency on available resources, and challenges in accurately calling bases in homopolymer regions, which may introduce errors.

Firstly, each of the five barcode reads of both the conditions were from sequencing of a Type B/ Victoria flu virus. When analyzed in Epi2me workflow, it was found that all the samples were from rapid barkoding kit and condition 2 is better in all the points we have considered such as coverage, matrics, read lenghts, and with a lower min q score threshold. Additionally, when created phylogenetic tree it was found that all the samples are replicates of the same biological sample.

The analysis of **true positive variants** revealed consistent **precision, recall, and F1 scores** for both conditions, also analyzed **True Positives (1), False Positives (0), False Negatives (12)** and found no differences suggesting comparable accuracy in variant calling. The uniformity of **failures in per-sequence and per-base quality scores**, coupled with the **absence of overrepresented sequences**, suggests that the base caller model effectively maintains sequencing data quality across both conditions. The model's performance in preserving sequence characteristics contributes positively to the overall **data reliability**. A detailed comparison of **read metrics** between conditions highlighted the strengths of Condition 2, which exhibited **superior mapping characteristics**, **longer read lengths,** and a **larger dataset**. These attributes, crucial for downstream analyses, contribute to the overall advantages of Condition 2. The **inconsistency in sequence duplication levels** in the first condition, with three samples showing extremely low percentages and two around 10%, raises **concerns about data homogeneity**. This variability suggests **potential technical artifacts** or **uneven library preparation, impacting data reproducibility**. In contrast, the second condition maintains more consistent and **lower sequence duplication levels** (1.20% to 1.80%), indicating a more **uniform distribution of sequences** and a dataset likely to be more representative of the **true genomic content**. Considering this, the condition2 is recommended due to its stable and **consistently low sequence duplication levels,** crucial for ensuring **data integrity in downstream analyses**.

**In conclusion**, the meticulous analysis of simulated amplicon sequencing data under two conditions using R9.4.1 flow cells and the "dna_r9.4.1_450bps_hac" base caller model demonstrates comparable accuracy in variant calling for both conditions. The robust performance of the base caller model, evident in consistent precision, recall, and F1 scores, contributes to the overall reliability of sequencing data. Despite the model's high accuracy, considerations of potential drawbacks such as computational intensity and challenges in homopolymer regions are crucial. The recommendation favors Condition 2, supported by its superior mapping characteristics, longer read lengths, more consistent and lower sequence duplication levels, ensuring data integrity for downstream analyses.