Bioinformatics Summer School Report

Name: Eyad Abdullah Alqaysi
Internship Organization: Broad Institute of MIT and Harvard
Supervisor's Name: Dr. Ricky Magner
Project title: Understanding Mistakes in Variant Calling

1. Introduction

During my summer internship at the Broad Institute, I worked on a project titled "Understanding Mistakes in Variant Calling." The primary objective was to develop a pipeline that models the impact of reference context on mistakes in variant calling. This project was inspired by Stratomod [1], a framework designed to statistically model genomic data, and it aimed to understand why certain variants are difficult to call accurately. I explored the use of Explainable Boosting Machines (EBMs) to rank the importance of different features that contribute to these mistakes. My work involved both the development of computational tools and the application of machine learning models to genomic data, ultimately helping to improve the accuracy of variant calling protocols.

2. Materials and Methodology

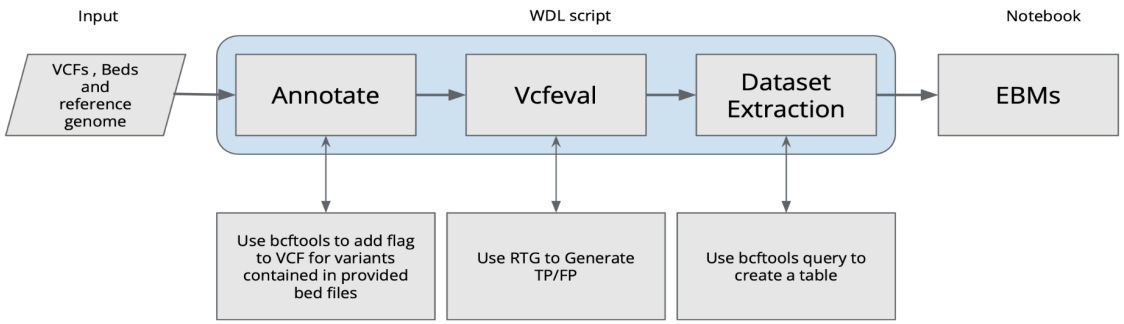


Figure 1.pipeline

To achieve the objectives of the project, I engineered a robust pipeline by integrating various software tools and custom scripts. The pipeline's architecture was meticulously designed with the following key components:

Data Annotation:
I utilized bcftools to enrich Variant Call Format (VCF) files with annotations derived from BED files, identifying regions with specific reference genome attributes, such as low mappability and high GC content. These annotations were critical in assessing the contextual factors contributing to variant calling errors.

Benchmarking:
The vcfeval tool from RTG Tools was employed to benchmark the VCF files against a reference truth dataset curated by NIST [2], classifying the variants into True Positives (TP), False Positives (FP), and False Negatives (FN). This benchmarking step was pivotal in establishing a ground truth for model training.
Feature Extraction:
I developed scripts to query the annotated VCF files and extract relevant features, which served as the input for the machine learning models. These features included both the annotated genomic contexts and other relevant variant characteristics.

Dataset: We used the dataset extracted using the WDL script from the HG002 VCF and region-specific BED files.

Model Training:
An Explainable Boosting Machine (EBM) model was implemented to rank the importance of various features in predicting TP/FP/FN labels [3]. The interpretability of the EBM model was a significant advantage, enabling us to understand which genomics contexts played a larger role in producing variant calling errors.

The entire process was encapsulated in a Workflow Description Language (WDL) script, ensuring seamless automation of the pipeline and cloud integration (Figure 1). This script was designed with flexibility in mind, allowing it to be executed across datasets from different sequencing platforms, including Illumina, Ultima, and PacBio.

3. Results and Discussion

The pipeline effectively processed data from various sequencing platforms, training EBMs tailored to each dataset. The models consistently identified regions with low mappability and segmental duplication content as prone to variant calling errors. Cross-platform analysis revealed differences in error-prone regions, providing insights into the strengths and limitations of different sequencing technologies (Figure 2).

These findings indicate that while the pipeline is robust across platforms, certain sequencing technologies are more prone to specific types of errors. For example, regions with low mappability were more problematic for Illumina datasets compared to PacBio (Figure 3), which is expected due to Illumina's short-read technology, whereas PacBio's long-read technology handles these regions better. These results underscore the importance of tailoring variant calling strategies to the specific characteristics of the sequencing platform being used.

Furthermore, the modular design of the pipeline allows for the integration of additional features and application to new datasets, ensuring its continued utility in future variant calling research. Future work should explore the potential of incorporating additional contextual features, such as local sequence homology or repeat regions, to further improve the accuracy of variant calling models.

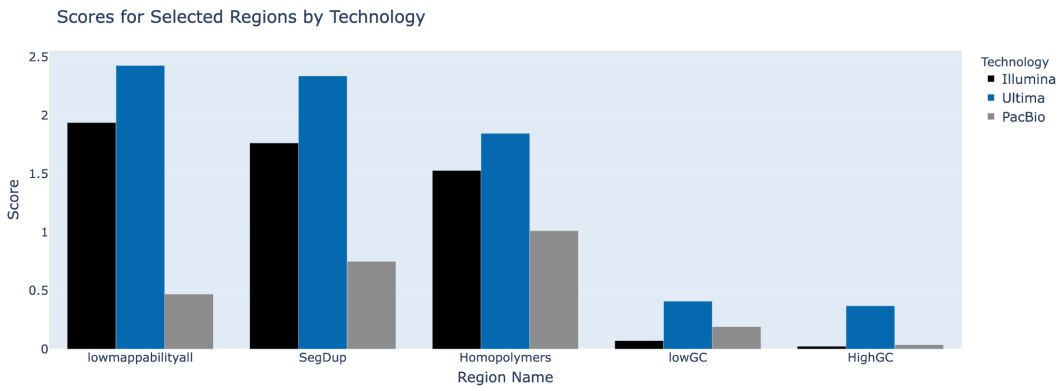


Figure 2. Comparison of error-prone regions across Illumina, Ultima, and PacBio platforms.

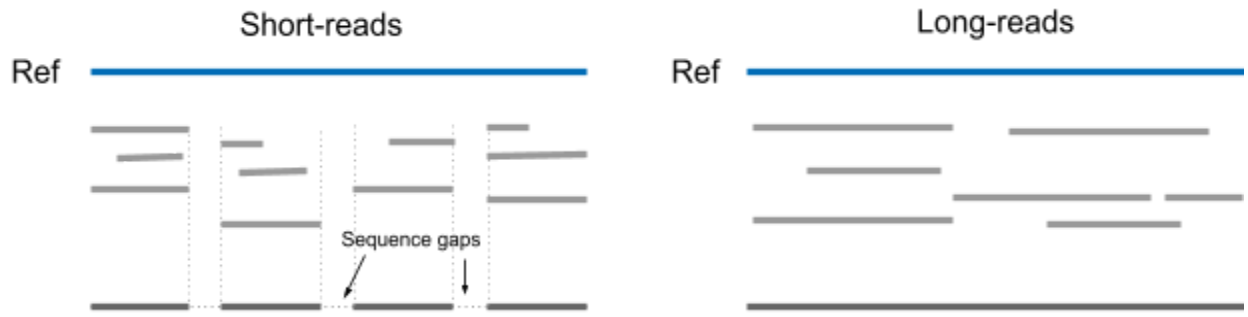


Figure 3. long-read vs short-read sequencing in mappability

4. Skills and Knowledge Gained

Technical Skills:

- **Data Manipulation:** Proficient in using bcftools and vcfeval for genomic data processing and annotation.
- **Machine Learning:** Gained hands-on experience with Explainable Boosting Machines (EBMs).
- **Automation:** Built and executed WDL pipelines on the Terra platform for large-scale genomic analysis.

Soft Skills:

- **Problem-Solving:** Improved my ability to troubleshoot and optimize complex pipelines.
- **Communication:** Enhanced my skills in conveying technical concepts clearly in both presentations and discussions.
- **Collaboration:** Worked effectively in a team, incorporating feedback to refine project outcomes.
- **Time Management:** Improved task prioritization and time management to meet project deadlines.
- **Adaptability:** Quickly learned and applied new tools and techniques to meet project needs.

5. Challenges and Solutions

One of the primary challenges I encountered was the handling of large and complex genomic datasets. Efficient data processing and storage were critical to the success of the project. To address these

challenges, I optimized my scripts for performance and leveraged cloud computing resources to manage the computational demands effectively.

Ensuring the interpretability of the machine learning model was another challenge. Although EBMs are designed for interpretability, selecting the most informative features and understanding their interactions required careful consideration. Iterative testing, along with guidance from my mentor and extensive literature review, was key to overcoming this obstacle.

6. Conclusion

My internship at the Broad Institute was a highly rewarding experience that significantly influenced my career goals. I gained a deeper understanding of bioinformatics and the complexities of variant calling, and I am now more motivated to pursue a career at the intersection of genomics and machine learning. The skills and knowledge I acquired will be invaluable as I continue to explore the field of computational biology. Additionally, the pipeline I developed during this internship will be utilized in future research efforts at the Broad Institute, contributing to ongoing advancements in variant calling accuracy.

7. Acknowledgments

I extend my heartfelt gratitude to Dr. Ricky Magner for his mentorship and guidance throughout this project. I would also like to thank Danielle Pace, Shadi Zaheri, Valentina Dsouza, Michael Gatzen, Paul Koch (InterpretML), Raghed Alkharouby, and the Hydro.gen team members for their invaluable support and contributions to my learning experience. Finally, I express my appreciation to the KAUST Academy team for making this opportunity possible.

8. References

1. **Stratomod by NIST:**
   Zook, J. M., Wagner, J., & Phillips, D. (2023). Stratomod: A tool for benchmarking genomic variant calling accuracy based on genomic reference sequences. bioRxiv. DOI: 10.1101/2023.01.20.524401v1.
2. **Extensive sequencing of seven human genomes to characterize benchmark reference materials by NIST:**
   Zook, J. M., McDaniel, J., Olson, N. D., Wagner, J., Parikh, H., Heaton, H., ... & Salit, M. (2019). Extensive sequencing of seven human genomes to characterize benchmark reference materials. Scientific Data, *3*, 160025. https://www.nature.com/articles/sdata201625
3. **Accurate Intelligible Models with Pairwise Interactions by Microsoft Research :**
   Yin Lou, Rich Caruana, Johannes Gehrke, Giles Hooker. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13). Available at: https://www.cs.cornell.edu/~yinlou/papers/lou-kdd13.pdf.