Bioinformatics Summer School Report

Name: Kumail Alawa
Internship Organization: Broad Institute of MIT and Harvard
Supervisor's Name: Fabio Cunial & Ryan Lorig-Roach
Project title: Evaluating Long-Read Breakend Calls
Code: https://github.com/Kumail00Alawa/sv_merge

1. Introduction

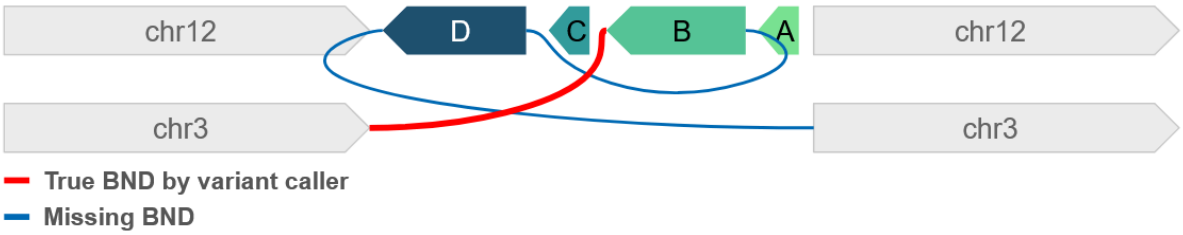
Structural Variants (SV) are any variation (e.g. insertion, deletion, inversion, duplication) of length ≥50 base pairs that is found when comparing a sample genome to a reference genome. SVs are important because they contribute significantly to human genetic diversity and disease, and they are one of the main goals of long-read sequencing projects. A Breakend (BNDs) is defined as a novel adjacency: two positions that are not consecutive in the reference become consecutive in a sample. Every SV can be represented as a set of BNDs in standard VCF format. In particular, BNDs are most useful for representing complex SVs. For example, an insertion where multiple segments of several chromosomes are permuted and copied into a target chromosome would not be informative if represented as a single INS record.

2. Materials and Methodology

The All of Us Long Reads Project involves ~1000 PacBio CCS samples sequenced at 8x and uses multiple SV callers: Sniffles, PBSV, and PAV. Using HPRC data processed in the same conditions, we found that there are ~50 distinct BND calls per sample, which are typically discarded because they are thought to be unreliable. We extend an existing pipeline that classifies true and false SVs to classify BNDs. We add new features that include the direction of a BND (i.e. whether it is in the forward direction or in the reverse complemented direction) and features about the connection of the BNDs (i.e. whether they are inter-chromosomal, intra-chromosomal, or single-ended). We train our model on HPRC data, where we can leverage high-quality assemblies as a source of truth.

3. Results and Discussion

We found that our machine learning model could classify the true BNDs from the false ones (AUC=0.9). Also, we found that including the features slightly improved the performance of the model. Furthermore, we examined several true BNDs, and we investigated the structure of one LINE1 event in detail, using both PacBio reads and high-quality assemblies from HPRC. We found that state-of-the-art variant callers emit only some of the BNDs that are necessary to fully describe a complex event as shown in the figure below.



4. Skills and Knowledge Gained

This internship helped me gain lots of information in a short time, which include:

- Understanding how short- and long-read sequencing work, after being taken on a tour to see the sequencing process at the Broad Genomics Platform.

- Understanding structural variants and complex events.
- Learning about various bioinformatics tools and standard formats that were built by people in the same organization and are currently being used by the wider bioinformatics community.
- Learning about Windows CLI and Linux terminal commands.

Alongside the technical knowledge, I had the opportunity to improve my soft skills, including:

- Delivering multiple presentations, significantly enhancing my public speaking and presentation skills.
- Communicating, collaborating, and contributing effectively with team members, accelerating our project's progress.
- Engaging in deep thinking about complex problems, thereby enhancing my problem-solving abilities.

5. Challenges and Solutions

I faced the following challenges, and I am glad that I was able to overcome them:

- Discovering that there were a few Python libraries that do not support Windows. To overcome this challenge, I had to use Windows Subsystem for Linux (WSL) as a terminal to be able to run Python scripts.
- Working with the VCF data format that I wasn't used to. To overcome this challenge, I attended an introductory session about dealing with the VCF data format, where I learned about various terminal commands to navigate through the VCF files.
- Learning about using the Integrative Genomics Viewer (IGV) tool to analyze various variants. To overcome this challenge, I had to meet my team members to learn how IGV works to contribute to the analysis we did.
- Obtaining various BND types from the VCF file. To overcome this issue, I had to use specific Python attributes to obtain the contents of the data that would help me obtain the various BND types.

6. Conclusion

In conclusion, I had the opportunity to contribute to the bioinformatics sector during this internship, where I worked on a project that aims to evaluate long-read BND calls. I used various bioinformatics-specific tools and contributed in a teamwork setting to advance the understanding of these BND calls that are emitted by state-of-the-art variant callers. We found that state-of-the-art variant callers emit only some of the BNDs that are necessary to fully describe a complex event. I am grateful for this opportunity because it helped me learn new skills and enhance my technical and soft skills. In the future, I look forward to creating a Saudi company that would lead the health sector, advancing the research and development of disease and treatment.

7. Acknowledgments

I am grateful to the Long Reads Lab for creating a welcoming environment:

- Kiran Garimella
- Shadi Zaheri
- Jonn Smith
- Hang Su
- Beri Shifaw
- Bridget Knight
- Maya Mau
- Ryan Lorig-Roach
- James Emery
- Steve Huang
- Lucas van Dijk
- Samuel Lee
- Willard Ford
- Madisyn Koza
- Fabio Cunial
- Wanying Dou
- Kevin Lydon

Special thanks to my KAUST Academy Supervisors for making this opportunity possible:

- Sultan Albarakati
- Naeemullah Khan
- Oroba Alnasr

Thanks to the KAUST Academy interns for making this internship memorable:

- Eyad Alqaysi
- Ebtisam Alshehri
- Omar Rawas
- Raghad Alkharouby
- Abdulrahman Alhomayany

8. References

Collins et al. "A structural variation reference for medical and population genetics." Nature 581.7809 (2020): 444-451.

Liao et al. "A draft human pangenome reference." Nature 617.7960 (2023): 312-324.

Mahmoud, Medhat, et al. "Structural variant calling: the long and the short of it." Genome biology 20 (2019): 1-14.

Schloissnig et al. "Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 Genomes Project." bioRxiv (2024): 2024-04.

The Variant Call Format Specification. VCFv4.5 and BCFv2.2. 28 Jun 2024. https://samtools.github.io/hts-specs/VCFv4.5.pdf