WARSAW UNIVERSITY OF TECHNOLOGY

FACULTY OF MATHEMATICS AND
INFORMATION SCIENCE

FIRST SEMESTER 2021/2022

Computational Genomics

PROJECT 1

BY:

Andra Umoru (324334)

SUBMITTED TO:

Tojek Zofia (DOKT)

1. Table of Contents

2. List of Figures

3. Project Description

Project 1 is to create a program that will accept standard short-read data alignment files against the GrCH38 reference genome (SAM / BAM standard), while the output will be a list of sorted structural variants in the standard VCF format. The student may use any approach of detecting variants.

3.1 Requirements:

- Working algorithm for detecting structural variants as per specification.
- Unit tests (e.g., unit tests may be on data for other organisms that have smaller genomes; however, there must be at least one test for human data; may be shallow-sequenced data).
- Comparison with another tool / tools (published / state-of-the-art).
- Final report
- Presentation of the program operation and results during the last / second to last classes.

4. Brief Overview on genomics

4.1 Reference Genome

A reference genome is a digital nucleic acid sequence database assembled by scientists to reflect the set of genes in a single idealized individual organism of a species [5]. Reference genomes do not correctly represent the set of genes of any single particular organism because they are generated through the sequencing of DNA from a number of individual contributors. A reference, on the other hand, gives a haploid mosaic of various DNA sequences from each donor [5]. Multiple species of viruses, bacteria, fungi, plants, and mammals have reference genomes. The Genome Reference Consortium's human reference genome, GRCh38, for example, was derived from thirteen anonymous volunteers [2].

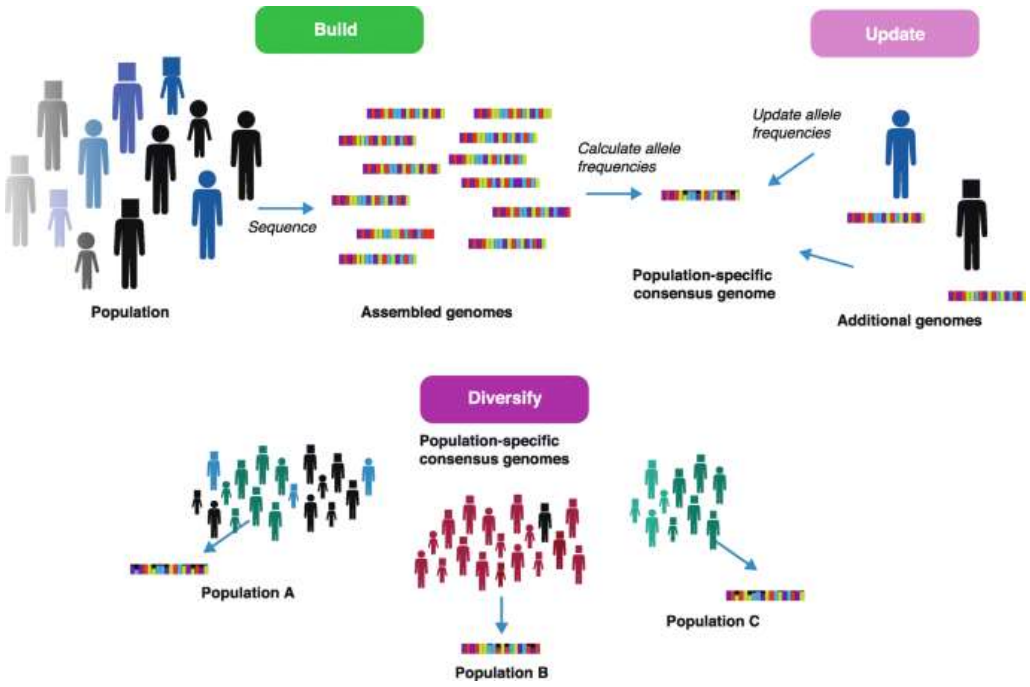

Figure 1: A brief of genomics

Source of figure: [7]

4.2 Structural variations

Large variations in chromosomal structure can be defined as these. Deletions, duplications, insertions, inversions, and translocations are examples of these changes. Many studies have recently demonstrated the significance of structural variations in genetic diversity and disease risk. Structural Variants are DNA areas that have changed by at least 50 base pairs compared to the reference genome. The majority of structural variations are discovered utilizing paired-end or mate-pair data from next-generation sequencing. Structural variant detection approaches for NGS data have limited sensitivity and precision, owing to short read lengths [1], [3]. Long-read sequencing technologies create lengthy reads that can span huge genomic regions and be mapped with high accuracy, which is especially important for insertions. Structural variants (SVs) are thought to play a key role in genetic variability and phenotypic control. Recently, many computer techniques for detecting SVs have been created. Single nucleotide variants only affect a small fraction of the human genome, whereas structural variants affect the entire genome [2].

4.3. Variants detection techniques

There are some few outlined techniques for detecting structural variants. Some of these techniques are: read-depth, read-pair, and split-read etc. In this implementation, I am going to be implementing the **read-depth** detection technique.

4.4 Read Depth Technique for SV Detection

Read depth calculates the number of reads mapped in a region: if a genomic location has fewer mapped reads than the average, it signals the presence of deletion in that locus. Furthermore, we may be witnessing an insertion if the number of mapped reads in an area is higher than typical. The method's high-reliability and ability to detect only deletions and insertions are both directly derived from it [2], [6].

5. Algorithm for Read Depth

i. Start
ii. Load the reference Genome **RG (GrCH38 – BAM)** standard.
iii. Divide the genome into bins of equal size
iv. Perform a calculation of the read depth signals as the number of reads.
v. Align the non-overlapping bins and calculate the average.
vi. Select, sort and merge the read depth signal,
vii. Return list the structural variants.
viii. Write the list of sorted structural variants and save it into VCF file.
ix. End

6. Tests

The test mechanism for this implementation is the reference Genome **RG (GrCH38 – BAM).** The results obtained from the implementation of the program are given below.

7. Results

The following are the results obtained from the program:

7.1 Reading the Chromosomes

After successfully loading the reference genome, the figure below shows the chromosomes read obtained from the BAM file.

```
GCCGTCCCGGGACTCCGCTCACCTTTATTATCTTGTCCTTTTGTTCTCTGCACCGCCGCAGGCCGGGACGTGGGGT
GGACTCCGCTCACCTTTATTATCTTGTCCTTTTGTTCTCTGCACCGCCGCAGGCCGGGACGTGGGGTACACACAAC
ATTATCTTGTCCTTTTGTTCTCTGCACCGCCGCAGGCCGGGACGTGGGGTACACACAACCCGGGGGAAGAGGGAAA
GGCTCTGCGTTCCAGCCCCAGGACCTCAACCCAGACCCCGCGCCTCGGCCCCGGCCGCGGCCCCTGTAACCCGGCC
AGGTCTGGTTGGCGGGAACAGGGGGCCCTAAAGTGAATTGAATGCAGGGGGTGTGGCACCAGCAGGGGGTGCTGAG
CAGGGGGTGCTGAGGTCCCCAGGCCAGCTCTGGGGGGAAAGCCTTTATCTTGGCCAGGCCTGAGGACCTGGGAGGG
ACCCGCCGTCCACCCACCTCGTCCCTCTCCAGGGAGGGCATGGGCTCCAGCCTCATCCAGTGAGTGTCCTGGCCCC
CACCTCGTCCCTCTCCAGGGAGGGCATGGGCTCCAGCCTCATCCAGTGAGTGTCCTGGCCCCAACACACATGCACA
GAGGTGTGGTCCAGACATCTGCCGTGGCACTGAGGAATGCTGACCATCCAGACTGAGACCCATGGTCGGGTCAGCA
CACTGAGGAATGCTGACCATCCAGACTGAGACCCATGGTCGGGTCAGCATCCAGCCCAGCCCAGTCCAGCAGGTGG
```

Figure 2: A Screenshot of the chromosome read

7.2 Sequence Length

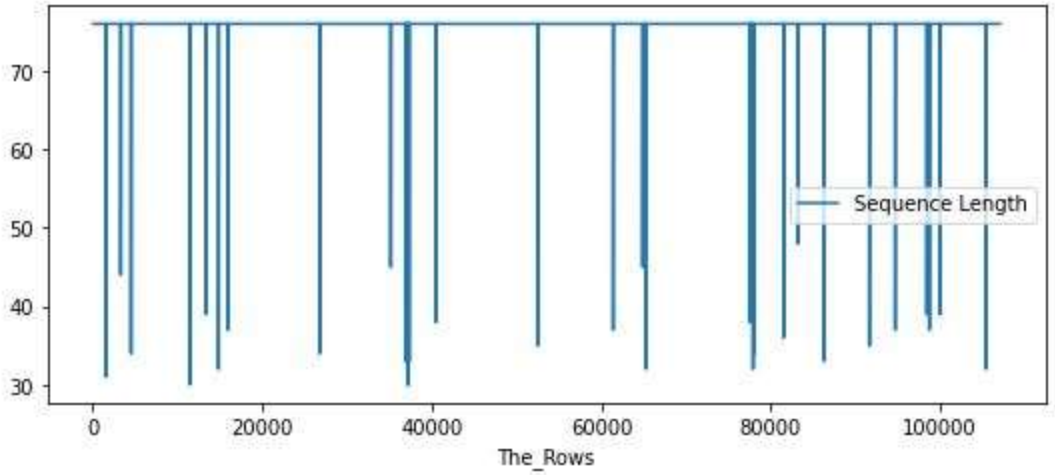

Figure 3: Plot showing the length of Sequence

7.3 Histogram (Sequences per Row)

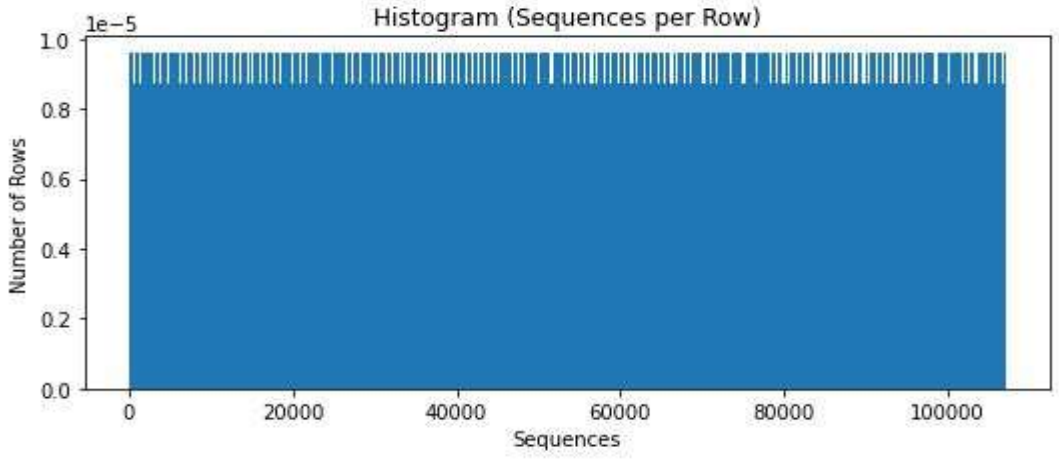

Figure 4: A Histogram showing the number of Sequences per Row

7.4 Plots (Histogram and Bar Chart) showing Bin Divisions

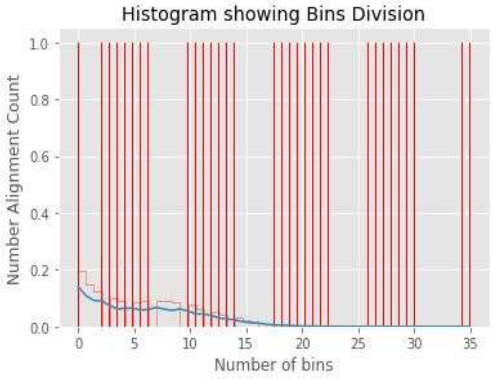

Figure 6: Histogram for Bins Division

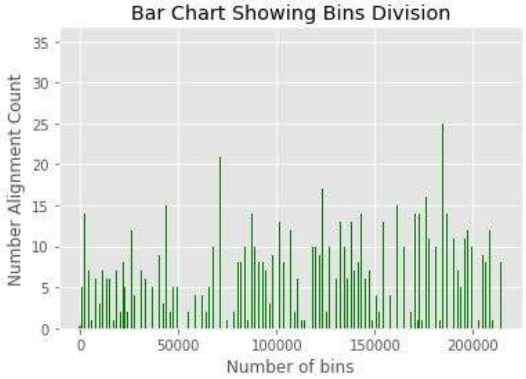

Figure 5: Bar Chart for Bins Division

7.5 Screenshot showing Number of Alignments

	Start_Position	End_Position	Number_of_Alignments
0	60280	60317	6
1	60258	60295	4
2	60295	60332	12
3	60298	60335	6
4	60296	60259	5
..	...	...	...
995	279267	279230	13
996	279234	279271	6
997	279220	279257	1
998	279236	279273	10
999	279266	279303	3

Figure 7: Screenshot Showing Alignments Counts

7.6 Histogram for Numbers Alignments Counts

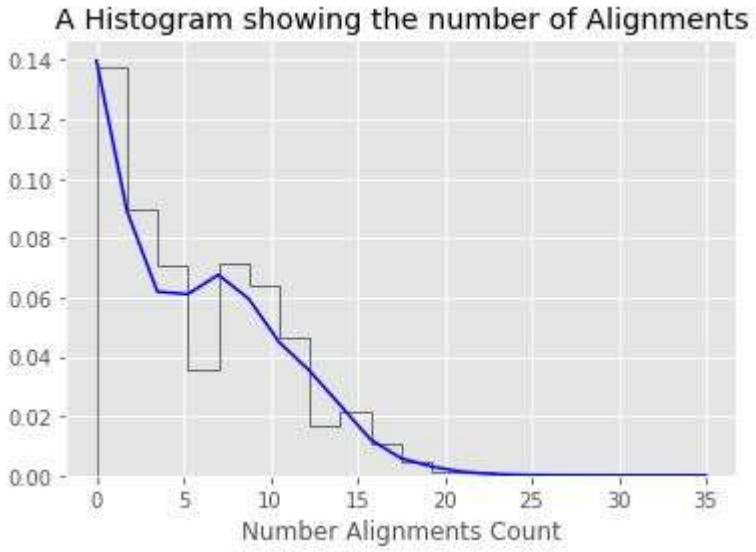

Figure 8: Histogram showing Alignments Count

8. Existing solution

There are many existing solutions that are already deployed into use in sorting and listing structural variants. Some of these includes: Delly, breakdance, IMR/DENOM, Platypus, Lumpy, MetaSV, Sniffles, and SVIM etc. Since my implementation is not completely done, it is may not be possible to compare it with existing state-of-the-art solution. The program I have here is far less than the existing ones, but with more work fine-tuning and error correction, it will can stand shoulder to shoulder with the existing solutions.

9. Conclusion

It is a known fact that there are other state-of-the-art solutions or tools that are used in detecting a structural variant. My attempt to produce a program similar to the existing was not 100% successful. The errors I encountered proved difficult for me to debug, this is due to novelty of the field to me, and inadequate of technical-know-how.

10. References

1. Rebecca E., Hsin-Ta W., Benjamin J. R. (2018). Identifying structural variants using linked-read sequencing data. *Bioinformatics*, Volume 34, Issue 2, Pages 353-360, https://doi.org/10.1093/bioinformatics/btx712

2. Liu, Y., Zhang, M., Sun, J. *et al.* Comparison of multiple algorithms to reliably detect structural variants in pears. *BMC Genomics* **21,** 61 (2020). https://doi.org/10.1186/s12864-020-6455-x
3. Hui-Jou C. (2017). A Project on: An algorithm for structural variant detection with third generation sequencing.
4. Yueyuan L., et. al. (2020). Comparison of multiple algorithms to reliably detect structural variants in pears.
5. https://en.wikipedia.org/wiki/Reference_genome
6. Computational Genomics Lab Presentation slides
7. https://www.google.com/url?sa=i&url=https%3A%2F%2Fgenomebiology.biomedcentral.com%2Farticles%2F10.1186%2Fs13059-019-1774-4&psig=AOvVaw2fH2mif_cV60-amSsPNpMi&ust=1650921508107000&source=images&cd=vfe&ved=2ahUKEwjHpM6B0K33AhUFx4sKHdqyD8kQr4kDegUIARDXAQ