Analysis of SplitThreader: A Weeb tool for exploration and analysis of rearrangements in cancer genomes

Vicente Machaca Arceda

Universidad Nacional de San Agustín, Arequipa-Perú,
vmachacaa@unsa.edu.pe

Abstract. Phylogenetics analysis is a very important task in Bioinformatics, we could learn about evolution, the relation between specimens. Nevertheless, a phylogenetics tree depends on the similarity analysis performed before. This similarity analysis is based on sequence alignment methods like BLAST and CLUSTALW, but they are too slow and we need other algorithms to process similarity between sequences. In this work, we present an analysis of four alignment-free algorithms based on the image texture computed from a sequence. We compared first-order statistics, gray level co-occurrence matrix, local binary patterns, and multi-resolution local binary patterns. Moreover, we used several mapping functions for each base. Then, we compared which of these algorithms were more similar to CLUSTALW. Finally, we got that first-order statistics is the method that is more likely to CLUSTALW with the advantage of having a low computational cost.

Keywords: Similarity analysis, Phylogentics trees, alignment-free methods, image textures.

1 Introduction

Genomics data has growth up exponentially, for example in 2009, they reached about 0.8 ZB, moreover in 2020 they reached about 40 ZB [1]. Furthermore, cancer related data are generated from: gene expression data (Microarray), NGS data, protein–protein interaction (PPI), pathway annotation data y gene ontology (GO). These data are important for research in cancer diagnosis and treatment. Big data resources allow researchers to observe large retrospective, and heterogeneous data of cancer patients [2].

For instance, the human genome is made of approximately 3.2 billions bp of DNA [3]. The HIV-1 genome is made of 20k bp of DNA, meanwhile the COVID-19 is made of 32k bp [4]. Additionally, there are approximately 19000 to 25000 genes (no one knows for sure) [3]. Finally, human genes have dozens of introns, each of which can be tens of thousands of nucleotides. Distinguishing exons from

introns and other forms of non-coding DNA is challenging [3]. This lack of information, makes difficult the research in cancer genomics.

Moreover, genomic instability is one of the hallmarks of cancer [5,6], resulting in a widespread copy number changes, structural variants and chromosome-scale rearrangements [7]. Furthermore, copy number variants and gene fusions are common drivers in cancer [8,9]. In this context, it is very important to detect these structural variants, but the available algorithms for identifying gene fusions do not have perfect specificity (false positive rate) and they require a joint analysis of genomic and transcriptomic data. Moreover, rearrangements variants are difficult to study because of the sheer complexity of rearrangements, which often include adjacencies between distant regions of a chromosome or even between unrelated chromosomes [7].

In this work, we reviewed and replicated the tool SplitThreader [7]. It is an open source interactive web application for analysis and visualization of genomics rearrangements and copy number variation in cancer genomes.

2 Concepts

In this section, we present the most relevant concepts related to bioinformatics and cancer genomics.

2.1 Genomics data

"DNA is abbreviation of deoxyribonucleic acid, organic chemical of complex molecular structure that is found in all prokaryotic and eukaryotic cells and in many viruses. DNA codes genetic information for the transmission of inherited traits" [10]. Moreover, the term genomics is used to refer the sum total of DNA in cells [3]. For instance, in Figure 1, we present a piece of COVID-19 DNA (genome) in FASTA format. DNA data is just a string of four characters: A, C, G, T that represent the nitrogen bases Adenine, Cytosine, Guanine and Thymine respectively. Unfortunately, this data is not 100% accurate and for a single human genome, it could reaches about 4 GB (3.2 billion bases).

```
>gb:MN988668|Organism:Wuhan seafood market pneumonia virus|Strain
Name:2019-nCoV WHU01|Segment:null|Host:Human
TTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAAC
GAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACT
AATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGT
TGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCC
CTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACG
TGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGGC
TTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGATG
CTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTCG
TAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTT
CTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAG
GCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGT
TACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGC
CCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTGT
CCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTGC
TTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAAA
TTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAAC
CAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACC
```

Fig. 1: A piece of COVID-19 DNA.

2.2 Sequence alignment

In Bio-informatics, sequence alignment could be defined as a way to arranging DNA, RNA and amino-acids sequences in order to find similarities [11]. For example, in Figure 2, we present two alignments, the top alignment (no alignment) seems to denote that there is not identity or similarity regions between two sequences, meanwhile, the bottom alignment shows that both sequences are similar.

```
No alignment

 CGATGCTAGCGTATCGTAGTCTATCGTAC
              |     ||
 ACGATGCTAGCGTTTCGTATCATCGTA

Alignned

 -CGATGCTAGCGTATCGTAGTCTATCGTAC
  |||||||||||| ||||||||||||||
 ACGATGCTAGCGTTTCGTA-TC-ATCGTA-
```

Fig. 2: Example of sequence alignment adding gaps.

Concordance reads .- This type of alignment refers when the reads have span size within the range of expected fragment size and consistent orientation of read pairs with respect to reference [12].

Discordance reads .- In this case, the reads have unexpected span size or inconsistent orientation of read pairs. It is important to identify this type of read in order to analyze genome alteration events [12].

Split reads .- When one portion of an read, maps to several locations of the same read map. Then, these are reads that have two or more alignments to the reference from unique region of the read. [12].

2.3 Structural variants

According to the National Center for Biotechnology Information (NCBI): "Structural variation (SV) is generally defined as a region of DNA approximately 1 kb and larger in size and can include inversions and balanced translocations or genomic imbalances (insertions and deletions), commonly referred to as copy number variants (CNVs)" [13]. In other words, this variations represent mutation in DNA, this mutations could be: insertions, deletions, inversions and translocations. In Figure 3, we present some examples.

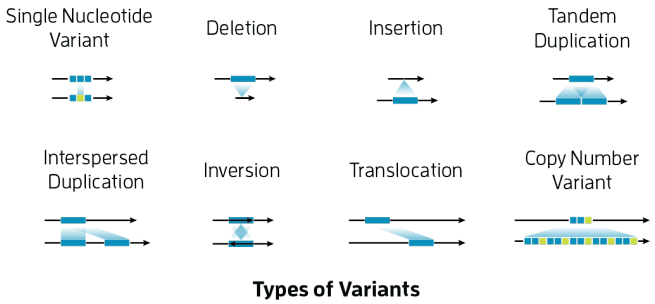

Fig. 3: Example of structural variants. Source: [14]

Copy number variants .- According to the National Human Genome (NIH): "A copy number variation (CNV) is when the number of copies of a particular gene varies from one individual to the next" [15]. For example in Figure 4, we present some examples of CNV, we could see how the number of genes varies individual 2 to 6. Additionally, it is recognized that some cancer diseases are associated to CNV [7–9, 15].

Gene fusions .- Gene fusion is a gene made by two or more genes [16]. For example, the first gene fusion discovered in cancer was BCR/ABL (related to leukemia), it is resulted from a fusion of chromosomes [17].

3 Related work

Detection of structural variants are the key point in cancer rearrangement analysis. For example, Lumpy [18], stands as a probabilistic framework for structural variant discovery. This framework uses three alignments inputs: concordant alignment, discordant alignment and split reads. Then, it uses a probabilistic method to detect structural variants like breakpoints (a pair of bases that are adjacent in an sequence samṕle but not in the reference genome).

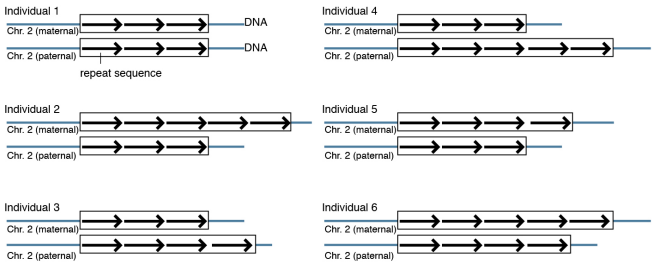

Fig. 4: Example of copy number variation. Source: [15]

Furthermore, there are studies about the perspective and challenges of structural variant detection for precision oncology [19]. Some techniques are used for short-reads [20, 21]. For long-read sequencing, technologies like PacBio and Oxford Nanopore Technologies (ONT) are valuable for structural variant detection [22]. Moreover, some algorithms have been develop to improve the quality of alignments and structural variant detection [23, 24].

Despite, there are many techniques to detect structural variants, they does not have 100% accuracy. Furthermore, it is difficult to evaluate and see the structural variants from Variant Call Files (VCF), these are files that represent the structural variants detected in genomes. In this context, there are some visualization tools, that stand for the analysis and discovery of structural variants. SplitThreader [7], for example, is used to graphically see copy number variations. Additionally, this tool used an algorithm to detect gene fusions.

MoMI-G is other tool that used a graph-based approach to represent structural variants [25]. This tool, used the same methodology of SplitThreader, but it is designed specifically for long-reads. Finally, there are several tools to detect and analyze structural variants, some authors reviewed and analyzed them [26].

4 Proposal

In this work, we replicated the results of SplitThreader [7]. This is a open source web application that stand for analysis and visualization of genomic rearrangements and copy number variation in cancer genomes. It constructs a graph of genomic rearrangements and uses a priority queue breadth-first search algorithm to search for novel interactions.

SplitThreader follows the pipeline in Figure 5. First, we aligns the sample genome with a reference genome using SpeedSeq (we could use another tool

that generated bam files). After alignment, three files are generated: concordance alignment, discordance alignment and split-reads. These files are taken by Lumpy and Copycat in order to detect structural variants. This step, generates two files: a variant calling file and a copy number file, these files are taken by SplitThreader in order to detect gene fusions, then SplitThreader plot the rearrangements using circle plots.

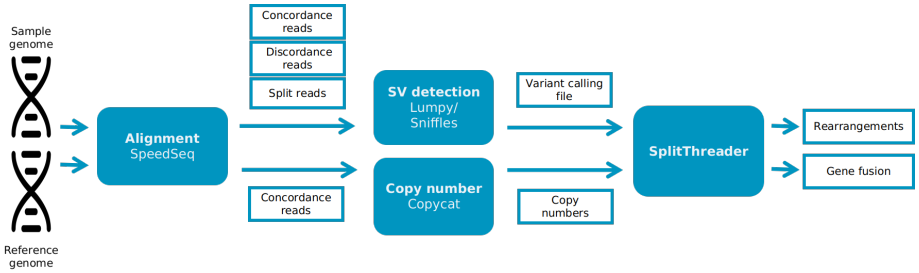

Fig. 5: SplitThreader pipeline. Two sequences are aligned using SpeedSeq, then output files are taken by Lumpy and Copycat in order to detect structural variants, finally SplitThreader takes its outputs and plot the rearrangements and gene fusions.

Formally, SplitThreader uses a graph, where a node represent sequences from DNA (reference genome) spanning between rearrangements breakpoints. Moreover, edges are used to represent rearrangements variants and no-rearranged reference-spanning connections. First, the graph is constructed where each node represents a chromosome, then each node is split each time we detect breakpoints. In Figure 7, we present this graph-based approach.

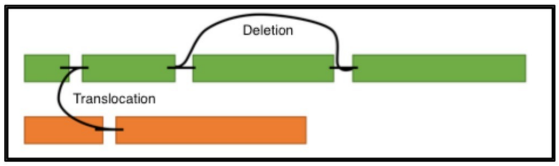

Fig. 6: Representation of rearrangements using a graph-based approach (SplitThreader proposal). Source: [7]

4.1 Gene fusion detection

Normally, gene fusion could be detected using RNA-seq and PCR validation to confirm transcriptome link between fusion genes [27]. However, there are scenarios where this analysis could no detect all gene fusions. For instance, when there is not a single variant that intersect both gene, when cancer genomes have "two hop" gene fusions and when the corresponding genomic region are not directly fused to each other but instead, it require passing through a third or more genomic region.

SplitThreader detect gene fusions searching for the shortest and lowest variant count path that connects the fusion genes. SplitThread uses a priority queue breath-first search to detect the shortest path in base pairs connecting two genes.

4.2 Variant neighborhood and copy number concordance

SplitThreader, can detect rearrangements responsible for copy number changes. The Web tool, categorizes each rearrangement variant by its copy number concordance. These concordances could be: matching, partial, non-matching or neutral. Finally, SplitThreader categorize each the rearrangements like reciprocal, simple, solo or crowded.

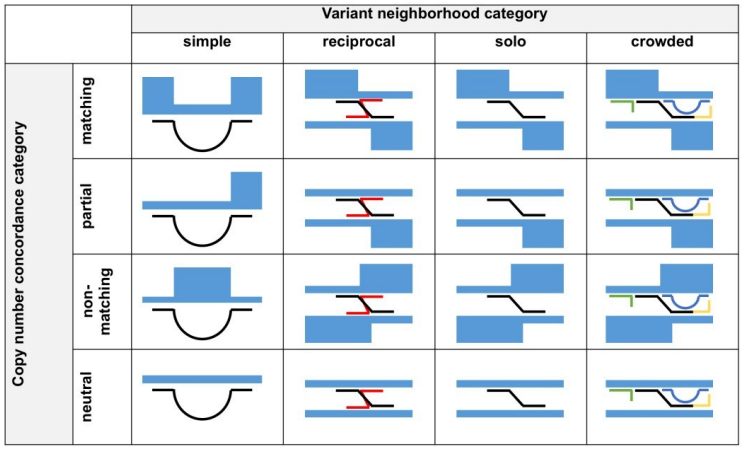

Fig. 7: Variant rearrangement categorization propose by SplitThreader based on copy number concordance. Source: [7]

5 Results

In this work we deployed SplitThreader locally. This Web tool is hosted on Github. In Figure 8, we show SplitThreader deployed locally. We performed some changes in order to have a more simple interface. SplitThreader uses two source files for the analysis: a Variant Calling File (VCF) and a Copy number profile, in Tables 1 and 2, we show examples of these files.

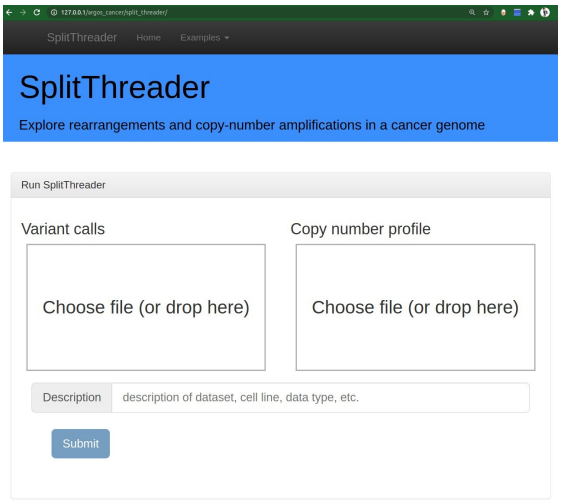

Fig. 8: SplitThreader locally hosted.

Table 1: Example of a Variant Calling File (VCF). Some columns and several rows were deleted. It was built using Lumpy.

chrom1	start1	stop1	chrom2	start2	stop2	name	type	split
1	17051740	17051740	1	234912188	234912188	35665	BND	71
1	47659735	47659735	8	105739138	105739138	573599	BND	6
1	87069066	87069066	2	82854932	82854932	571553	BND	6
1	109650635	109650635	22	30163373	30163373	575755	BND	36
1	150593722	150593722	5	55447995	55447995	572639	BND	19
1	153306043	153306043	16	76228788	76228788	575219	BND	6
1	168186186	168186186	1	182274316	182274316	20968	BND	11
1	201288206	201288206	10	52642286	52642286	574013	BND	9
1	208992122	208992122	3	87327147	87327147	572038	BND	
...								

Table 2: Example of a copy number profile. It was built using Copycat.

chromosome	start	end	coverage
1	0	10000	0
1	10000	20000	0.9605
1	20000	30000	0
1	30000	40000	0
1	40000	50000	0.0059
1	50000	60000	0.775
1	60000	70000	0.6154
1	100000	110000	0.3666
...			

After analysis, SplitThreader plot graphs (Figure 9). It uses a circle plot in order to show genome rearrangements. Moreover, we could see copy numbers and gene fusion for a pair of chromosomes. For instance, in Figure 9, we see chromosome 1 and chromosome 2 (histograms), then the lines between them, represent copy regions.

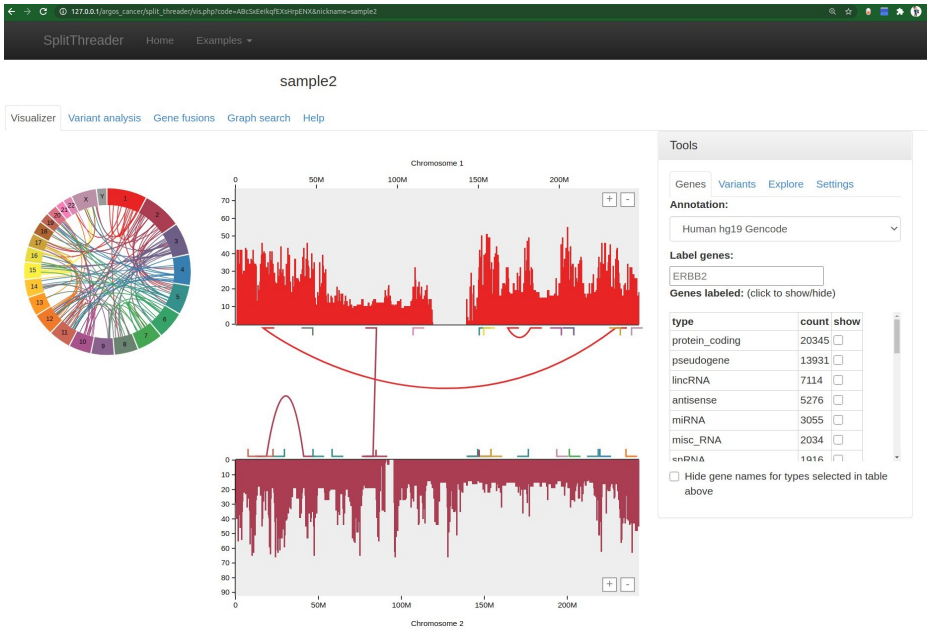

Fig. 9: Plots generated by SplitThreader.

6 Conclusions

In this work, we analyzed and replicated the results of SplitThreader. It is a Web tool for exploration and analysis of rearrangements in cancer genomes.

SplitThreader have a simple architecture but it is robust for genome analysis. It basically uses PHP to perform API Restfull services and Javascript to plot the graphs. Additionally, it uses R and Python for processing the data.

References

1. Archana Prabahar and Subashini Swaminathan. Perspectives of machine learning techniques in big data mining of cancer. In *Big Data Analytics in Genomics*, pages 317–336. Springer, 2016.
2. Jules J Berman. *Principles of big data: preparing, sharing, and analyzing complex information*. Newnes, 2013.
3. John M Archibald. *Genomics: A Very Short Introduction*, volume 559. Oxford University Press, 2018.
4. Gurjit S Randhawa, Maximillian PM Soltysiak, Hadi El Roz, Camila PE de Souza, Kathleen A Hill, and Lila Kari. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study. *bioRxiv*, 2020.
5. Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.
6. Philip J Hastings, James R Lupski, Susan M Rosenberg, and Grzegorz Ira. Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551–564, 2009.
7. Maria Nattestad, Marley C Alford, Fritz J Sedlazeck, and Michael C Schatz. Splitthreader: Exploration and analysis of rearrangements in cancer genomes. *bioRxiv*, page 087981, 2016.
8. Adam Shlien and David Malkin. Copy number variations and cancer. *Genome medicine*, 1(6):1–9, 2009.
9. Felix Mitelman, Bertil Johansson, and Fredrik Mertens. The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*, 7(4):233–245, 2007.
10. Britannica definitions. Dna. https://www.britannica.com/science/DNA, 2021. Accessed: 2021-05-07.
11. Jin Xiong. *Essential bioinformatics*. Cambridge University Press, 2006.
12. Bioinfomratics handbook. Concordance, discordance and split reads. https://www.biostars.org/p/278412/, 2021. Accessed: 2021-05-07.
13. NCBI. Overview of structural variation. https://www.ncbi.nlm.nih.gov/dbvar/content/overview/, 2021. Accessed: 2021-05-07.
14. PacBio. Two review articles assess structural variation in human genomes. https://www.pacb.com/blog/two-review-articles-assess-structural-variation-in-human-genomes/, 2021. Accessed: 2021-05-07.
15. National Human Genome. Copy number variation (cnv). https://www.genome.gov/genetics-glossary/Copy-Number-Variation, 2021. Accessed: 2021-05-07.

16. National Cancer Institute. Gene fusion definition. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/fusion-gene, 2021. Accessed: 2021-05-07.
17. Kees Stam, Nora Heisterkamp, Gerard Grosveld, Annelies de Klein, Ram S Verma, Morton Coleman, Harvey Dosik, and John Groffen. Evidence of a new chimeric bcr/c-abl mrna in patients with chronic myelocytic leukemia and the philadelphia chromosome. *New England Journal of Medicine*, 313(23):1429–1433, 1985.
18. Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. Lumpy: a probabilistic framework for structural variant discovery. *Genome biology*, 15(6):1–19, 2014.
19. Ianthe AEM van Belzen, Alexander Schönhuth, Patrick Kemmeren, and Jayne Y Hehir-Kwa. Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. *NPJ Precision Oncology*, 5(1):1–11, 2021.
20. Matthew Ruffalo, Thomas LaFramboise, and Mehmet Koyutürk. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27(20):2790–2796, 2011.
21. Biao Liu, Jeffrey M Conroy, Carl D Morrison, Adekunle O Odunsi, Maochun Qin, Lei Wei, Donald L Trump, Candace S Johnson, Song Liu, and Jianmin Wang. Structural variation discovery in the cancer genome using next generation sequencing: computational solutions and perspectives. *Oncotarget*, 6(8):5477, 2015.
22. Wouter De Coster and Christine Van Broeckhoven. Newest methods for detecting structural variations. *Trends in biotechnology*, 37(9):973–982, 2019.
23. Franka J Rang, Wigard P Kloosterman, and Jeroen de Ridder. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome biology*, 19(1):1–11, 2018.
24. Kevin J Travers, Chen-Shan Chin, David R Rank, John S Eid, and Stephen W Turner. A flexible and efficient template format for circular consensus sequencing and snp detection. *Nucleic acids research*, 38(15):e159–e159, 2010.
25. Toshiyuki T Yokoyama, Yoshitaka Sakamoto, Masahide Seki, Yutaka Suzuki, and Masahiro Kasahara. Momi-g: modular multi-scale integrated genome graph browser. *BMC bioinformatics*, 20(1):1–14, 2019.
26. Toshiyuki T Yokoyama and Masahiro Kasahara. Visualization tools for human structural variations identified by whole-genome sequencing. *Journal of human genetics*, 65(1):49–60, 2020.
27. Henrik Edgren, Astrid Murumagi, Sara Kangaspeska, Daniel Nicorici, Vesa Hongisto, Kristine Kleivi, Inga H Rye, Sandra Nyberg, Maija Wolf, Anne-Lise Borresen-Dale, et al. Identification of fusion genes in breast cancer by paired-end rna-sequencing. *Genome biology*, 12(1):1–13, 2011.