Structural Variant Calling in Genomes Using Deep Learning

Anwarul Bashir Shuaib
Department of CSE, BUET
1805010@ugrad.cse.buet.ac.bd

Abu Humayed Azim Fahmid
Department of CSE, BUET
1805036@ugrad.cse.buet.ac.bd

Supervisor — Dr. Atif Hasan Rahman
Department of CSE, BUET
atif@cse.buet.ac.bd

***Abstract*—The exploration of structural variations (SVs) in the genome is pivotal for understanding genetic diversity and disease mechanisms. However, the inherent complexity of structural variants poses considerable challenges for accurate detection, particularly with traditional short-read sequencing technologies. Long-read sequencing has emerged as a promising alternative, offering enhanced resolution and the ability to span larger genomic regions. In this study, we introduce a novel deep learning-enhanced methodology that builds upon the PEPPER-Margin-DeepVariant framework, specifically tailored for the detection of structural variants using long-read sequencing data. Our approach represents the first-of-its-kind integration of deep learning for the detection of structural variants. By clustering similar structural variants based on normalized indel similarity score and analyzing soft clips within alignment files, we enhance the detection signals for structural variants that are often missed by conventional methods. We evaluate our methodology against state-of-the-art SV calling methods in the Challenging Medically Relevant Genes (CMRG) and HG002 Tier1 Benchmark datasets, demonstrating superior performance with an F1 score of 98.87% on the CMRG dataset, and a highly competitive F1 score of 96.66% on the HG002 dataset.**

***Index Terms*—structural variants, deep learning, long-read sequencing, variant detection, genomics, levenshtein distance**

I. Introduction

The human genome, comprising approximately 3 billion base pairs, encodes between 20,000 and 25,000 genes and harbors millions of genetic variations [1]. These variations profoundly affect gene expression and function. While single nucleotide variations (SNVs) and small indels are prevalent, with over 100 million identified to date, structural variations (SVs) impact a much larger portion of the genome than SNVs and small indels combined [2].

Structural variants are also linked to several fatal diseases, including amyotrophic lateral sclerosis (ALS), Alzheimer's, chronic myeloid leukemia (CML), and various forms of cancer [3]. However, detecting SVs remains a formidable challenge. Many SVs are located in poorly mapped regions of the genome, and larger SVs often span multiple sequencing reads. Additionally, in highly repetitive regions, SVs tend to disperse, weakening detection signals and often causing many variants to go undetected. Traditional short-read sequencing technologies often fail to map these complex regions uniquely, but recent advances in long-read sequencing methods like Pacific Biosciences and Oxford Nanopore Technologies (ONT) are showing promise in this area.

Aiming to address these challenges, we present PEPPER-SV, an extension of the PEPPER-Margin-DeepVariant [5] framework for detecting structural variants. Some of the key contributions include:

1) Our approach represents the first-of-its-kind integration of deep learning techniques specifically designed for SV detection.
2) We consolidate similar variants within repetitive regions using a normalized indel similarity score (1).
3) We utilize soft-clipped reads to extract SV signals, which are often overlooked in traditional analyses.

PEPPER-SV utilizes long-read sequencing data and merges these sophisticated techniques to provide a robust and dependable tool for SV identification.

II. Related Works

In the field of genomics, considerable progress has been made in developing tools for detecting genetic variants. For SNVs and small indels, deep learning-based tools like DeepVariant [4] and PEPPER-Margin-DeepVariant [5] have shown exceptional performance, particularly with long-read sequencing data, outperforming traditional methods. However, the detection of SVs has largely relied on algorithmic approaches such as CuteSV [6], Delly [7], and Sniffles [8]. Only recently has the landscape begun to change with the introduction of Dysgu [9], a tool that enhances SV detection by analyzing alignment gaps, discordant and supplementary mappings, and utilizing machine learning for classifying variants.

Despite these advancements, the specific application of deep learning to SV calling has remained largely unexplored. Motivated by the success of PEPPER-Margin-DeepVariant, we adapt its framework to develop a novel deep learning-based pipeline tailored for robust and efficient detection of structural variants. Our method aims to bridge the gap in the current toolkit, aiming to overcome the existing limitations and enhance the SV detection in long-read sequencing data.

III. Methodology

A. Variant Breakpoint Detection

Our pipeline begins by processing an input alignment file, which it divides into multiple smaller, manageable windows for concurrent analysis. Within each window, we examine potential SV breakpoints by interpreting the signatures present in the alignment information, represented by a string. We specifically focus on insertions and deletions, which can appear either with their own signatures or with soft-clipped reads, and track the total variant support count across all reads base by base. If the support count exceeds a pre-defined threshold, that position is identified as a candidate variant breakpoint.

B. Consolidation of Similar Variants

In regions where SVs are sparsely distributed, the support counts may fall below the necessary threshold for reliable detection. To counteract this, we consolidate similar candidate SVs using a metric known as the normalized indel similarity score (1). This is calculated by taking the Levenshtein distance between two candidate variants, normalizing it, and then taking the complement to get a similarity score between 0 and 1. Here, we use $\alpha = 2, \beta = 1, \gamma = 1$ as mismatch, insertion and deletion penalty consecutively. L_1, L_2 represents the length of the two variants under consideration.

$$S_{\text{norm}} = 1 - \frac{\alpha \cdot n_{\text{mis}} + \beta \cdot n_{\text{ins}} + \gamma \cdot n_{\text{del}}}{L_1 + L_2} \qquad (1)$$

This scoring system ensures that dissimilar variants are kept separate while similar variants are merged together.

C. Matrix Generation

Following the identification and consolidation of these breakpoints, our pipeline constructs a pileup summary matrix for each candidate variant. We position each breakpoint centrally within the matrix, accompanied by a 64-base wide context window on either side, resulting in a 128-base window around the breakpoint. This matrix integrates initial features from the existing pipeline, which includes: variant support counts, variant length, reference base encoding, read base encoding and read depth, both for forward and reverse strands. Additionally, we enrich the feature set with two novel features that help capture the soft-clip signatures within each read, totaling a set of 28 features.

D. Variant Category Prediction

To classify and predict the characteristics of these variants, we train two LSTM models: one focused on determining the variant category and the other on predicting the genotype from the pileup matrices. These models leverage the deep learning capabilities to refine the accuracy of variant categorization. Finally, the results are compiled and output in the VCF (Variant Call Format) file format, providing a detailed and actionable list of identified variants.

TABLE I
IMPACT OF MODEL ON PERFORMANCE OUTCOMES.

Benchmark	**Analysis Type**	**Precision**	**Recall**	**F1 Score**
HG002 Tier 1	Without Model	89.74%	97.47%	93.44%
	With Model	94.67%	98.73%	96.66%
CMRG	Without Model	90.91%	97.04%	93.88%
	With Model	98.24%	99.51%	98.87%

TABLE II
PERFORMANCE ON HG002 TIER 1 BENCHMARK DATASET.

Method	**Precision**	**Recall**	**F1 Score**
Dysgu	96.04%	98.05%	97.04%
CuteSV	95.81%	98.20%	96.99%
PEPPER-SV	94.67%	98.73%	96.66%
Delly	96.26%	96.83%	96.55%
Sniffles	95.97%	96.59%	96.28%

TABLE III
PERFORMANCE ON CMRG SV BENCHMARK DATASET.

Method	**Precision**	**Recall**	**F1 Score**
PEPPER-SV	98.24%	99.51%	98.87%
Dysgu	98.91%	98.52%	98.72%
Sniffles	98.41%	99.01%	98.71%
Delly	98.90%	98.03%	98.46%
CuteSV	98.40%	97.54%	97.96%

IV. MAIN RESULTS

We evaluate our results on HG002 PacBio HiFi sequence data aligned against GRCh37 reference genome. As demonstrated in Table I, our model profoundly enhances both precision and recall. Performance benchmarks sorted in non-decreasing order of F1 score are presented on Table II and Table III. Specifically, on the HG002 Tier 1 SV benchmark dataset, our approach achieved an F1 score of 96.66%, outperforming Delly and Sniffles. Furthermore, on the CMRG SV benchmark dataset, our model achieved an F1 score of 98.87%, outperforming all competing methodologies. These results highlight the effectiveness of our approach, establishing a new paradigm in structural variant calling using deep learning.

REFERENCES

[1] S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, et al., “The complete sequence of a human genome,” *Science*, vol. 376, no. 6588, pp. 44–53, 2022.

[2] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, et al., ”Origins and functional impact of copy number variation in the human genome,” *Nature*, vol. 464, no. 7289, pp. 704–712, 2010.

[3] P. Stankiewicz and J. R. Lupski, “Structural variation in the human genome and its role in disease,” *Annual Review of Medicine*, vol. 61, pp. 437–455, 2010.

[4] R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, et al., ”A universal SNP and small-indel variant caller using deep neural networks,” *Nature biotechnology*, vol. 36, no. 10, pp. 983–987, 2018.

[5] K. Shafin, T. Pesout, P.-C. Chang, M. Nattestad, A. Kolesnikov, S. Goel, et al., ”Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads,” *Nature methods*, vol. 18, no. 11, pp. 1322–1332, 2021.

[6] T. Jiang, Y. Liu, Y. Jiang, J. Li, Y. Gao, Z. Cui, Y. Liu, B. Liu, and Y. Wang, ”Long-read-based human genomic structural variation detection with cuteSV,” *Genome biology*, vol. 21, pp. 1–24, 2020.

[7] T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel, ”DELLY: structural variant discovery by integrated paired-end and split-read analysis,” *Bioinformatics*, vol. 28, no. 18, pp. i333–i339, 2012.

[8] F. J. Sedlazeck, P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. Von Haeseler, and M. C. Schatz, ”Accurate detection of complex structural variations using single-molecule sequencing,” *Nature methods*, vol. 15, no. 6, pp. 461–468, 2018.

[9] K. Cleal, and D. M. Baird, ”Dysgu: efficient structural variant calling using short or long reads,” *Nucleic acids research*, vol. 50, no. 9, pp. e53–e53, 2022.