[BIOINFO 2024]

SVDP: Somatic Variation Detecting Pipeline Using PCA and Machine Learning

Minsu Kim[1], KeumSeok Kang[1] , and Giltae Song [1,2,3,*]

[1] *Division of Artificial Intelligence, Department of Information Convergence Engineering, Pusan National University, Busan, 46241, South Korea*

[2] *Center for Artificial Intelligence Research, Pusan National University, Busan, 46241, South Korea*

[3] *School of Computer Science and Engineering, Pusan National University, Busan, 46241, South Korea*

**Corresponding author: gsong@pusan.ac.kr*

Structural variation (SV) refers to any genomic alteration exceeding a single nucleotide, encompassing insertions, deletions, inversions, duplications, and translocations. SV can be divided into germline and somatic mutations. Germline mutations, inherited from the parents through sperm and egg cells, result in identical mutations across all cells of an organism. In contrast, somatic mutations occur in localized regions within the organism. These mutations often contribute to the development of somatic tumors, and their rapid proliferation into cancer cells highlights the critical importance of early detection. To address this, we propose a Somatic Variation Detection Pipeline (SVDP) that utilizes Matrix PCA and machine learning models for early identification of such mutations. Our study focuses on three key types of structural variation: deletions, duplications, and complex combinations. To create a robust training dataset, we leveraged somatic SV data from various sources, including human genome sequences like NA12878 and NA12877, as well as somatic genome sequences from tissues such as the heart and brain. We used NA12878 as a reference to build a simulation dataset representing a virtual human subject by sampling other structural variations at specific frequencies. This approach allowed us to systematically train the model to accurately detect and analyze somatic mutations in a controlled environment. We enhanced SV feature extraction using ARC-SV calling, applied PCA to further refine these features, and trained an XGBoost machine learning model for each of the three SV types.