[BIOINFO 2024]

Deep Transformer Model for Detecting driver signal in structural variation

Eun-youn Seo[1]

[1]*Department of Biomedical System Informatics, College of Medicine, Yonsei University*

With advances in sequencing technology, whole-genome sequencing (WGS) analysis has become more accessible, significantly enhancing our understanding of non-coding regions. The non-coding regions comprise approximately 98% of the entire genome and contain regulatory elements such as promoters and enhancers that play crucial roles in various biological processes and can influence tumorigenesis. Structural variations (SVs), in particular, frequently occur in non-coding regions and can affect various genetic variations, potentially serving as key cancer drivers. Therefore, discovering structural variations in non-coding regions through WGS analysis may serve as a critical starting point for identifying new cancer drivers and developing therapeutic strategies.

An essential first step in this process is distinguishing driver signals from passenger signals to identify significant structural variation breakpoints. Traditional approaches have primarily used machine learning techniques such as generalized linear models (GLMs) to identify significant variation points through linear and non-linear combinations of genomic features. However, these machine learning models have limitations in capturing the complex non-linear relationships between genomic features that influence the formation of structural variations and in reflecting the broader context of the entire genome sequence.

To overcome these limitations, this study developed an analysis tool based on a deep learning Transformer framework, modeling a background of passenger breakpoints and comparing this background to select significant breakpoints. This approach allows for the analysis of patterns in atypical structural variation breakpoints and the prediction of the likelihood of atypical structural variations, facilitating the identification of driver breakpoints in non-coding regions.

AI models for predicting structural variations are increasingly utilized in various genomic studies and enable the discovery of novel biomarkers with high accuracy. The development of AI models that can be applied to genomic variation, gene expression, and metabolomic analyses simplifies the processing of complex, large-scale data, aids in the identification of biomarkers, and helps elucidate

the mechanisms of diseases. These new biomarkers enhance our understanding of individual disease phenotypes, assist in predicting drug responses, and, in some cases, act as new therapeutic targets that may lead to drug development.

As a result, these advancements are expected to contribute to the progress of precision medicine by providing more accurate treatments and pharmaceuticals for patients, thereby improving therapeutic efficacy and enhancing the quality of life for patients undergoing treatment.