**[BIOINFO 2025]**

# CluVar: Clustering of variants using Autoencoder for inferring the phylogeny of cancer subclones in single cell RNA sequencing data

Chae Won Kim[1*], Heewon Park[2*], Dohyeon Kim[1,3], Yuchang Seong[1], Minhae Kwon[2,4†], Junil Kim[1,5†]

1 Department of Bioinformatics and Life Science, Soongsil University, Seoul 06978, Republic of Korea

2 Department of Intelligent Semiconductors, Soongsil University, Seoul 06978, Republic of Korea

3 Center for Natural Product Systems Biology, Korea Institute of Science and Technology, Gangneung 25451, Republic of Korea

4 School of Electronic Engineering, Soongsil University, Seoul 06978, Republic of Korea

5 School of Systems Biomedical Science, Soongsil University, Seoul 06978, Republic of Korea

[*] These authors are contributed equally.

[†] Corresponding authors: MK (minhae@ssu.ac.kr) and JK (junilkim@ssu.ac.kr)

## Abstract

Tumor tissues are composed of malignant subclones with diverse genetic profiles. Reconstructing the evolutionary trajectory of these subclones is crucial for understanding how tumors acquire malignant traits. However, current approaches to subclonal tree reconstruction are limited either by their reliance on single-cell DNA sequencing (scDNA-seq) which involve a small number of cells and thus yield low-resolution results, or using single-cell RNA sequencing (scRNA-seq) data, which despite including larger cell populations, remain susceptible to bias from high dropout rates and technical noise. Here, we introduce CluVar, an autoencoder-based framework for inferring the phylogeny of cancer subclones from scRNA-seq data using mutation profile analysis. To address the extensive missing variant information inherent in scRNA-seq datasets, CluVar incorporates a customized loss function and multiple hidden layers optimized for clustering. CluVar demonstrated superior performance in reconstructing phylogenetic trees of cancer subclones under a range of erroneous conditions. When applied to cancer scRNA-seq data, the phylogenetic tree predicted using CluVar aligned well with the transcriptomic profiles. These findings highlight its utility for tracing evolutionary trajectories and identifying novel variants associated with cancer progression.