[BIOINFO 2024]

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

Chaewon Kim[1], Dohyeon Kim[1], Yuchang Seong[1] and Junil Kim1,2,*

[1]*Department of Bioinformatics, Soongsil University*

[2]*School of Systems Biomedical Science, Soongsil University*

**Corresponding author: junilkim@ssu.ac.kr*

Tumor tissues consist of subclones of malignant cells with various genetic profiles. Reconstructing the evolutionary trajectory of tumor cells allows us to understand how tumors acquire malignant traits. Here, we introduce CluVar (Clustering of Variants using autoencoder), a framework for reconstructing the phylogeny of cancer subclones. CluVar utilizes deep autoencoder model to address the high missing values of variants in single cell RNA sequencing data. CluVar clusters distinct subclones and reconstructs phylogenetic trees using a four-steps approach: 1) Learning autoencoder model for dimension reduction, 2) Bayesian Gaussian Mixture Models for clustering, 3) Gibbs sampling for genotype estimation, and 4) the neighbor-joining algorithm for tree reconstruction. This approach demonstrates robust performance to reconstruct phylogenetic trees of cancer subclones with various error conditions. This performance evaluation highlights its potential application in identifying novel variants for cancer progression using single cell RNAseq data.