**[BIOINFO 2024]**

# Development of a deep learning model for deconvolution of bulk *in situ* Hi-C by referencing pseudo-bulk single-cell epigenome profiles

Kyukwang Kim,[1] and Inkyung Jung[1,*]

[1]*Department of Biological Sciences, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea*

**Corresponding author: ijung@kaist.ac.kr*

The advancement of single-cell sequencing technology enables the dissection of complex cell-type specific gene regulation in heterogeneous tissues. Despite such advancements, most single-cell studies are based on single-cell RNA and ATAC-seq, with standardized high-throughput experiment methods. Even though the 3D genome plays a key role in cell type-specific gene regulation, the data is lacking due to the low throughput and unstandardized protocols of single-cell Hi-C. To address this issue, we developed a multimodal deep learning model that utilizes the scATAC-seq identified pseudo-bulk and cell type-specific epigenome profile differences to define cell type-specific 3D genome organization from the bulk *in situ* Hi-C contact map. We used a variational autoencoder (VAE) to project the Hi-C contact maps into a 1D latent vector, and the latent vector was adjusted by a transformer-based model to produce a deconvolution target output by referring to the input epigenome profile. For the model training, synthetic mixed data was generated using ATAC-seq and Hi-C data of the cardiomyocyte differentiating RUES2 stem cell line. The developed model reproduced the 128x128 bin Hi-C contact map target with an average mean squared error (MSE) loss of 0.0002 or less. Also, it operates at a resolution 2.5 times higher than existing scRNA-seq-based deconvolution models. Due to the limitation of the convolution window, only deconvolution of a ~2.5Mb distance region from the diagonal is available, rather than the entire Hi-C contact map. However, even with this limitation, cell type-specific 3D genome structures such as TADs and loops can be identified, and deconvolution of such targets is expected to help understand the 3D genome of complex tissues in situations where single-cell Hi-C data is lacking.