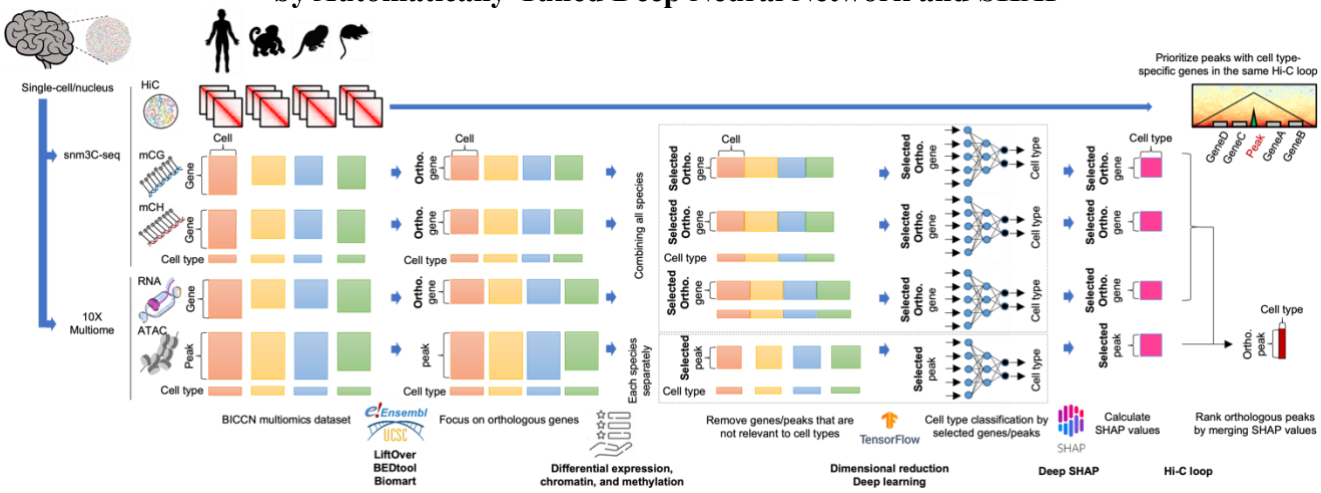
Yoshiaki Tanaka lab, Maisonneuve-Rosemont Hospital Research Centre, University of Montreal

cisMultiDeep: Identifying Cell Type-Specific Cis-Regulatory Regions by Automatically-Tuned Deep Neural Network and SHAP



Background: Brain comprises a complex variety of cell types of cells, such as neurons and astrocytes, which are intimately interacted for proper function and homeostasis. The cell type is determined by expression of a specific set of genes that are governed by cis-regulatory elements (CREs). Recent studies have demonstrated that the cell type-specific CREs are not fully conserved (Villar et al., 2015), whereas the expression patterns of the cell type-specific genes are evolutionarily conserved (Shay et al., 2012). Here, we aim to identify the set of both 'conserved' and 'non-conserved' cell type-specific CREs that controls the expression of 'conserved' cell type-specific genes. We developed cisMultiDeep (https://github.com/ytanaka-bio/cisMultiDeep (not public yet)) that employs automatically-tuned deep learning with SHAP feature importance assessment in cross-species single-cell multiomics data.

Identify cell type-specific 'conserved' genes: Given high conservation of the cell type-specific genes, we first obtained orthologous gene list from *Biomart*. Then, we sorted the orthologous genes by cell type specificity, which was assessed by Wilcoxon's test in their RNA expression or mCG/mCH DNA methylation profiles. In each cell type, top and bottom 600 genes were selected. Then, feature-barcode matrices from four species were combined by these selected cell type-specific 'conserved' genes.

Identify cell type-specific peaks: Unlike the RNA and methylation profiles, the differential chromatin accessibility was analyzed in all 'conserved' and 'non-conserved' peaks. *LiftOver* and *Bedtools* were used to define the peak conservation. Top and bottom 600 peaks were selected from the mixture of conserved and non-conserved peaks. The feature-barcode matrix was saved in each species, separately.

Deep learning: To ask if the selected genes/peaks are sufficient to define the cell types, we employed automatically-tuned deep neural network that was designed by *Keras Tuner*. Using RNA expression, chromatin accessibility, or DNA methylation of the selected genes/peaks as input and cell types as output, we trained the deep neural network models. Notably, the models trained by RNA expression and methylation profiles classified individual cells into each cell type with more than 90% accuracy, whereas the classification by ATAC displayed relatively low accuracy, maybe due to its sparsity.

Feature importance: SHAP value is a widely used measurement to estimate how much each feature contributes to the model's prediction. *DeepSHAP* is a technique that can handle the complex and non-linear interactions across features and is optimized to calculate SHAP value for deep neural network. Using the trained deep neural network models, we calculated mean absolute SHAP value that represents the impact of each gene or peak on the cell type classification.

Peak prioritization: Chromatin looping enables distal CREs to contact their target genes. Here, we hypothesize that 'functional' cell type-specific CREs are physically contacted with various cell type-specific genes, and ranked the peaks by the sum of SHAP values of the contacted genes by Hi-C loop. If the peak is conserved, we added up the sum of SHAP values of the contacted genes in other species.

BICCN Challenge: Predicting Cell Type-Specific Enhancers from Cross-Species Multiomics