

# SMNN: Batch Effect Correction for Single-cell RNA-seq data via Supervised Mutual Nearest Neighbor Detection

Yun Li

*University of North Carolina  
E-mail: yunli@med.unc.edu*

**Abstract:** An ever-increasing deluge of single-cell RNA-sequencing (scRNA-seq) data has been generated, often involving different time points, laboratories or sequencing protocols. Batch effect correction has been recognized to be indispensable when integrating scRNA-seq data from multiple batches. A recent study proposed an effective correction method based on mutual nearest neighbors (MNN) across batches. However, the proposed MNN method is unsupervised in that it ignores cluster label information of single cells. Such cluster or cell type label information can further improve effectiveness of batch effect correction, particularly under realistic scenarios where true biological differences are not orthogonal to batch effect. Under this motivation, we propose SMNN which performs supervised mutual nearest neighbor detection for batch effect correction of scRNA-seq data. Our SMNN either takes cluster/cell-type label information as input, or, in the absence of such information, infers cell types by performing clustering of scRNA-seq data. It then detects mutual nearest neighbors within matched cell types and corrects batch effect accordingly. Our extensive evaluations in simulated and real datasets show that SMNN provides improved merging within the corresponding cell types across batches, leading to reduced differentiation across batches over MNN. Furthermore, SMNN retains more cell type-specific features after correction. Differentially expressed genes (DEGs) identified between cell types after SMNN correction are biologically more relevant, and the DEG true positive rates improve by up to 841%. SMNN is implemented in R, and freely available at <https://yunliweb.its.unc.edu/SMNN/> and <https://github.com/yycunc/SMNNcorrect>.