Genome Analysis

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

Yuchen Yang[1*], Gang Li[2*], Huijun Qian[2], Kirk C. Wilhelmsen[1,3], Yin Shen[4,5], and Yun Li[1,6,7,†]

[1]Department of Genetics, [2]Statistics and Operations Research, [6]Biostatistics, [7]Computer Science, [3]Renaissance Computing Institute, University of North Carolina, Chapel Hill, North Carolina. 27599, USA. [4]Institute for Human Genetics, [5]Department of Neurology, University of California, San Francisco, San Francisco, CA 94143, USA

* Equal contribution

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: 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, MNN is unsupervised in that it ignores cluster label information of single cells, which can further improve effectiveness of batch effect correction, particularly under realistic scenarios where true biological differences are not orthogonal to batch effect.

Results: In this work, we propose SMNN for batch effect correction of scRNA-seq data via supervised mutual nearest neighbor detection. SMNN either takes cluster/cell-type label information as input or infers cell types using scRNA-seq clustering in the absence of such information. It then detects mutual nearest neighbors within matched cell types and corrects batch effect accordingly. Compared to MNN, SMNN provides improved merging within the corresponding cell types across batches and retains more cell type-specific features after correction.

Availability and implementation: SMNN is implemented in R, and freely available at https://yunliweb.its.unc.edu/SMNN/ and https://github.com/yycunc/SMNNcorrect.

Contact: yunli@med.unc.edu

INTRODUCTION

An ever-increasing amount of single cell RNA-sequencing (scRNA-seq) data has been generated as scRNA-seq technologies mature and sequencing costs continue dropping. However, large scale scRNA-seq data, for example, those profiling tens of thousands to millions of cells (such as the Human Cell Atlas Project (Rozenblatt-Rosen *et al.*, 2017)), almost inevitably involve multiple batches across time points, laboratories or experimental protocols. The presence of batch effect renders joint analysis across batches challenging (Stegle *et al.*, 2015). Batch effect, or systematic differences in gene expression profiles across batches, cannot only obscure the true underlying biology, but also lead to spurious findings. Thus, batch effect correction, which aims to mitigate the discrepancies across batches, is crucial and deemed indispensable for the analysis of scRNA-seq data across batches (Stuart and Satija, 2019).

A recent study proposed MNN for batch effect correction, using information of mutual nearest neighbors across batches. MNN has demonstrated superior performance over alternative methods (Haghverdi *et al.*, 2018). MNN makes a critical assumption that true biological differences are orthogonal to batch effect, requiring that at the minimum that variation from batch effect is much smaller than that from biological effect. Under this assumption, MNN finds, across batches, mutual nearest neighboring cells for each cell to be corrected, and then computes batch-effect correction vectors under a Gaussian kernel. However, the orthogonality assumption might not hold in real data, particularly given that different batches may easily differ in many aspects, including samples used, single cell capture method, or library preparation approach. Under non-orthogonal scenarios, MNN will not be optimal using its global (ignoring cell type information) nearest neighbor search strategy, leading to undesired correction results. For example, under the scenario depicted in Fig. 1a, MNN leads to cluster 1 (C1) and cluster 2 (C2) mis-corrected due to mismatching single cells in the two clusters/cell-types across batches. To address this issue, here we present SMNN, a supervised version of MNN by incorporating cell type information. SMNN performs nearest neighbor searching within the same cell type, instead of global searching regardless of cell type (Fig. 1b). Cell type information, when unknown *a priori*, can be inferred via clustering methods (Kiselev *et al.*, 2019; Duò *et al.*, 2018).

IMPLEMENTATION

SMNN is implemented in R, encompassing two major steps: one optional clustering step and the other batch effect correction step (Supplementary Fig. S1). In the first step, SMNN takes the expression matrix as input, and performs clustering using Seurat v. 3.0 (Butler *et al.*, 2018), where dimension reduction is first conducted via principal component analysis (PCA) to the default of 20 PCs, and then graph-based clustering follows on the dimension-reduced data with *resolution* parameter of 0.9 (Yang *et al.*, 2019). Corresponding clusters/cell types are then matched across batches based on marker genes specified by users. This entire clustering step can be bypassed by feeding SMNN cell cluster labels.

With cell cluster label information, SMNN searches mutual nearest neighbors within each cell type and performs batch effect correction via the *SMNNcorrect* function, which calculates a cell-specific batch-correction vector by exploiting a Gaussian kernel to obtain a weighted average based on its mutual nearest neighbors, following MNN.

APPLICATION EXAMPLES

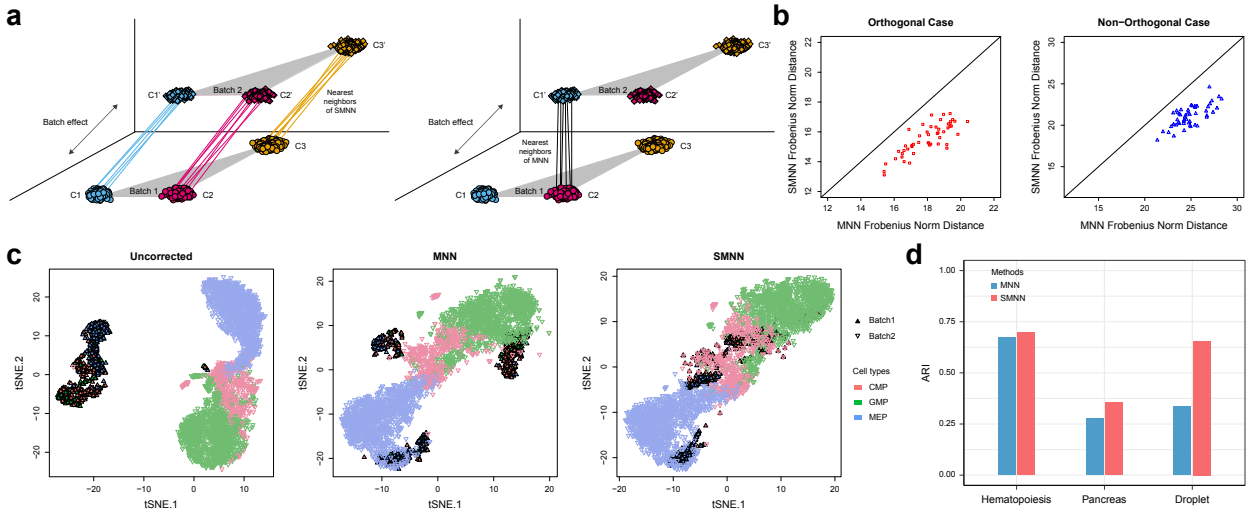

Fig. 1. Schematics of mutual nearest neighbor detection in SMNN and MNN correction and the benchmarking of SMNN correction in both simulation and real data. (**a**) Schematics of detecting mutual nearest neighbors between two batches in SMNN and MNN in non-orthogonal scenario. (**b**) Frobenius norm distance between two batches after SMNN and MNN correction in simulation data under orthogonal and non-orthogonal scenarios, respectively. (**c**) t-SNE plot for two hematopoietic datasets before and after correction with SMNN and MNN, respectively. Solid and inverted triangle represents the first and second batch, respectively, and different cell types are specified by different colors. (**d**) Accuracy evaluation of the ability of retaining the cell type specific features in SMNN and MNN correction in three datasets. Adjusted Rand Index (ARI) is employed to measure the similarity between the clustering before and after batch effect correction.

Since MNN has been shown to excel alternative methods (Haghverdi *et al.*, 2018; Stuart and Satija, 2019), we here focus on comparing our SMNN with MNN. In our simulations (see Supplementary method for details), SMNN demonstrates superior performance over MNN, as manifested by smaller Frobenius norm distance, under both orthogonal and non-orthogonal scenarios (Fig. 1b; Supplementary Fig. S2-5).

To assess the performance of SMNN in real data, we applied it to two hematopoietic scRNA-seq datasets, generated using different sequencing platforms, MARs-seq and SMART-seq2 (Supplementary Table 1) (Paul *et al.*, 2015; Nestorowa *et al.*, 2016). Three cell types shared between the two datasets were extracted for our analysis. Fig. 1c shows the t-SNE plot before and after correction with SMNN and MNN. Notably, both SMNN and MNN substantially mitigated discrepancy between the two datasets. Comparatively, SMNN better mixed cells of the same cell type across batches (here the two datasets) (Supplementary Fig. S6), especially for common myeloid progenitor (CMP) cells and megakaryocyte-erythrocyte progenitor (MEP) cells, which were wrongly corrected by MNN due to sub-optimal nearest neighbor search ignoring cell type information (Supplementary Fig. S7). Supplement Fig. S8-9 further show that cells from the same cell type across the two batches are closer (2.8 - 8.2% reduction when measured by Euclidean distance, detailed in Supplementary results) after SMNN correction than MNN.

We also examined the ability to differentiate cell types after SMNN and MNN correction in three datasets (Supplementary Table 1). In all three real datasets, clustering after SMNN correction shows improved (by 7.6 - 42.3%) Adjusted Rand Index (ARI) than MNN (Fig. 1d), suggesting that SMNN correction more effectively recovers cell-type specific features.

CONCLUSION

In summary, we present SMNN for batch effect correction of scRNA-seq data. SMNN adopts a supervised mutual nearest neighbor matching, effectively mitigates batch effect, and better recovers cell-type specific features. We believe that SMNN is valuable for the analysis of scRNA-seq data.

FUNDING

This research was supported by the National Institute of Health grant R01 HL129132 (awarded to YL).

Conflict of interest: none declared.

REFERENCES

Butler,A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411.

Duò,A. *et al.* (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, **7**.

Haghverdi,L. *et al.* (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, **36**, 421.

Kiselev,V.Y. *et al.* (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, 1.

Nestorowa,S. *et al.* (2016) A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, **128**, e20–e31.

Paul,F. *et al.* (2015) Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, **163**, 1663–1677.

Rozenblatt-Rosen,O. *et al.* (2017) The Human Cell Atlas: from vision to reality. *Nature*.

Stegle,O. *et al.* (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, **16**, 133.

Stuart,T. and Satija,R. (2019) Integrative single-cell analysis. *Nat. Rev. Genet.*, 1.

Yang,Y. *et al.* (2019) SAFE-clustering: Single-cell aggregated (from ensemble) clustering for single-cell RNA-seq data. *Bioinformatics*, **35**, 1269–1277.