High Throughput Sequencing Data Aggregation and Cell Identification

Hanbo Sun, Hyun-Min Kang

12/15/2017

Publication Forthcoming

1 Introduction

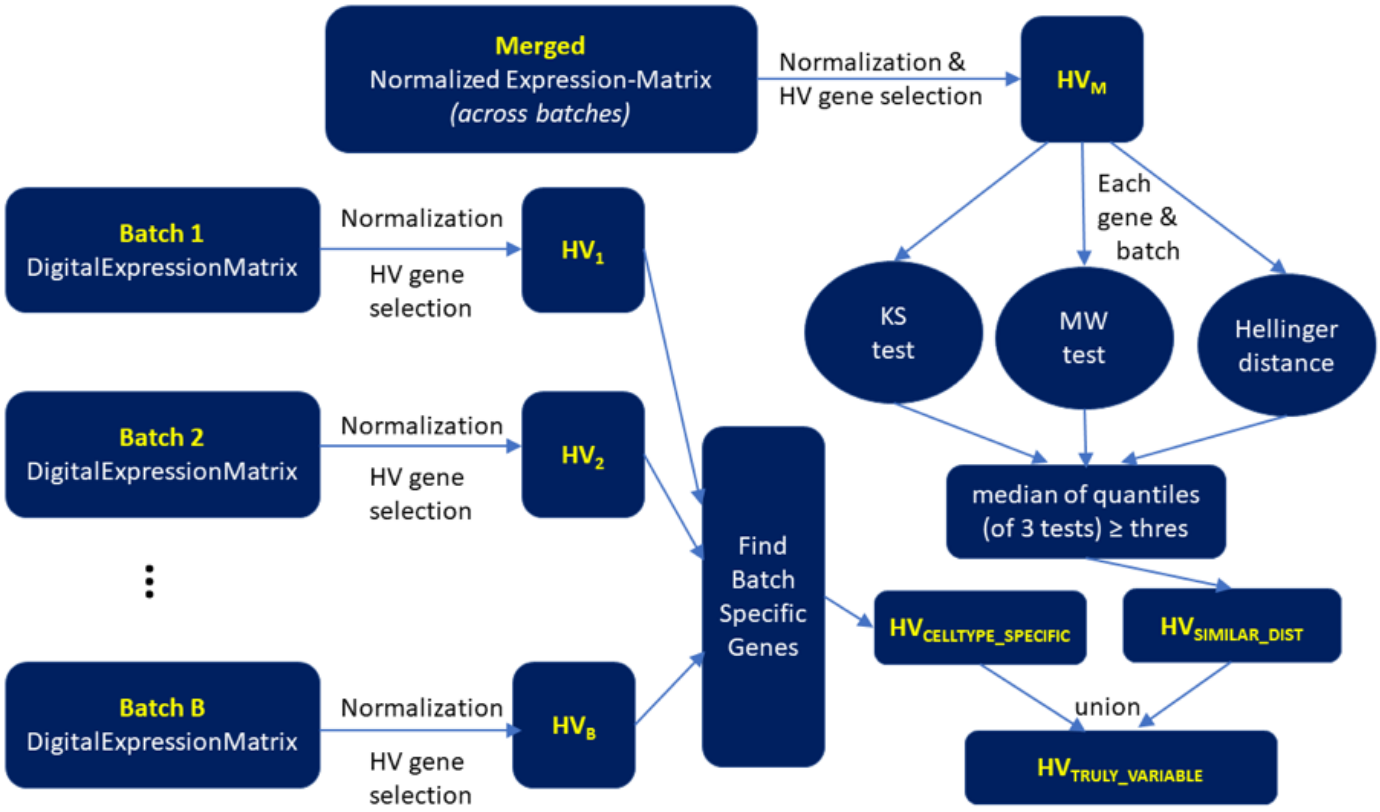
1.1 Data

The data we used include: PBMC6K, generated from Chromium Demonstration (v1 Chemistry) and Cell Ranger 1.1.0; PBMC4K, generated from Chromium Demonstration (v2 Chemistry) and Cell Ranger 2.0.1, as well as PanT3k, generated from Chromium Demonstration (v2 Chemistry) and Cell Ranger 2.0.1. Three batches were collected from different healthy donors. PBMCs include lymphocytes (T cells, B cells, and NK cells), monocytes, dendritic cells, and some rare type such as megakaryocytes. In humans, the frequencies of these populations vary across individuals. For the healthy, lymphocytes are typically in the range of 70 – 90% of PBMCs, monocytes range from 10 – 30% of PBMCs, while dendritic cells and megakaryocytes cells are rare, being only 1 – 2% and 0.5% of PBMCs, respectively. The frequencies of cell types within the lymphocyte population include 70 – 85% CD3+ T cells (i.e., 45 – 70% of PBMC), 5 – 20% B cells (up to 15% of PBMC), and 5 – 20% NK cells (up to 15% of PBMC).

1.2 Methods

The presence of batch effects is a well-known problem in experimental data analysis, and single-cell RNA sequencing (scRNA-seq) data is no exception. Sources of variability in experimentally scRNA-seq data include the biological variation of interest in addition to technical originated from laboratories, sequencing platforms, measuring instrument, experiments design and random measurement error.

The technical variability may fat ally compromise interpretation and mask biology underlying of the data. Building on that, computational batch correction is critical for eliminating uninteresting technical factors and obtaining valid biological conclusions. Here, we present a novel model-based correction method: SCRUB - single cell robust unification batch effect, as part of the preprocessing, to attenuate batch effect. SCRUB determines “truly variable” (TV) genes. Conceptually, the TV genes are a subset of the “highly variable (HV)” genes, which contributes to high variance between cell types, across the batches, excluding the genes that shows signature of batch-specificity. If the distribution of gene expressions significantly differs between batches, the gene is considered to be batch-specific. We determine the batch-specificity of a gene as an omnibus statistics between the three statistics: (1) Kolmogorov-Smirnov (KS) test, (2) Mann-Whitney (MW) test, and (3) Bhattacharya distance. SCRUB can be expanded to multiple batches



2 Results

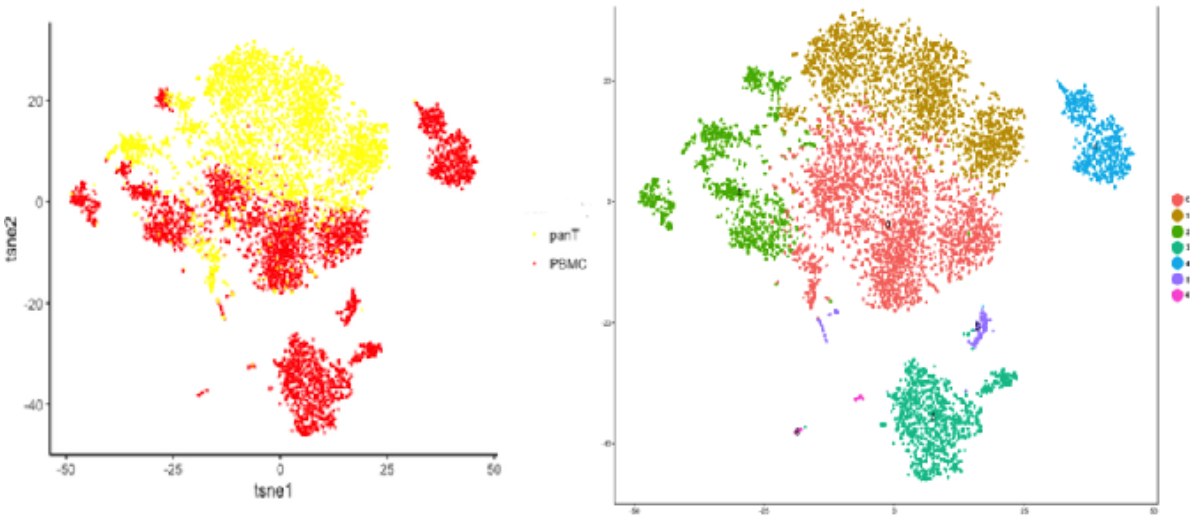
2.1 Two batches: PBMC and Pan T

To demonstrate the applicability of our method on correcting batch effect, we considered PBMC data set and Pan T data set. Two batches cells are collected from different healthy donors and two data sets are generated in different experiments, both based on cell ranger scRNA-seq protocols. PBMC tissue contains 4340 cells and Pan T includes 3555 cells. Denote them as PBMC4K and panT 3k for short. The majority composition of PBMC tissue include T cells, Monocytes cell, B cell, NK cells and some minor cell type (Megakaryocytes, Dendritic Cells for example) To assess performance, we

performed t-SNE dimensionality reduction on the (log-scale and normalized) highly variable genes of the

1. uncorrected data:

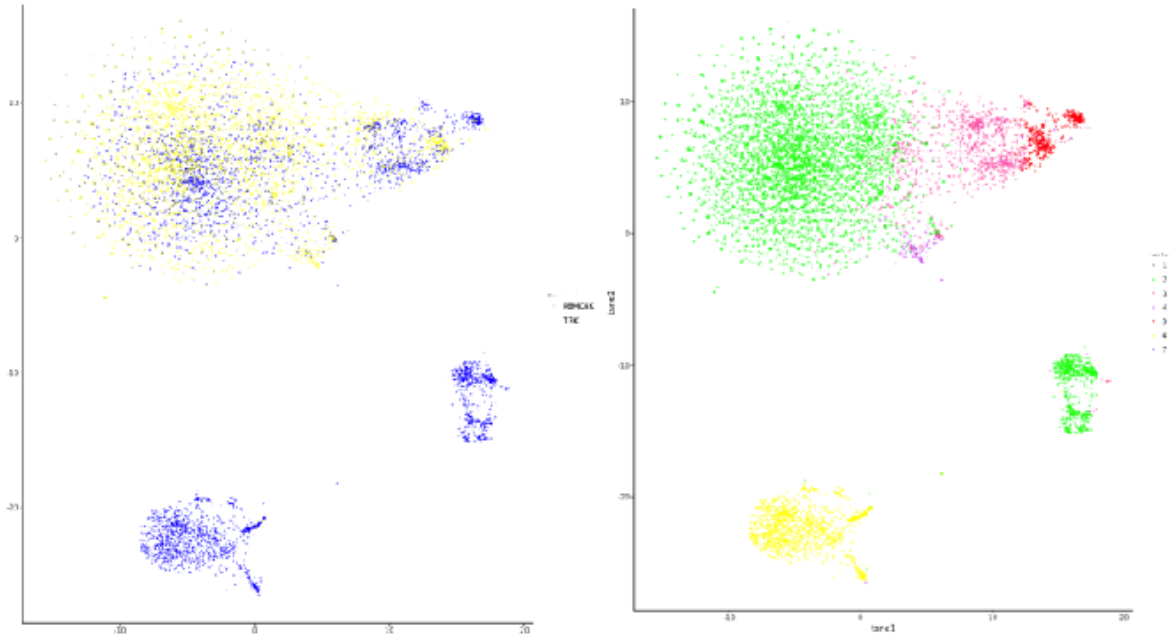
t-SNE was applied to the Seurat normalization uncorrected data. Red group is PBMC4K, yellow group is panT3K. Severe batch effect exists and dominates the clustering. For example, yellow and red clusters right panel, they are clustered based on batches rather than underlying cell differentiation.



2. MNN corrected data:

Apply t-SNE on MNN corrected data. It shows that batch effect is corrected. However, we notice that the performance of clustering is bad. For example, B cells(green cluster in the right panel) cannot be

identified and mixed with T cells. Also, some minor clusters are generated with no biological

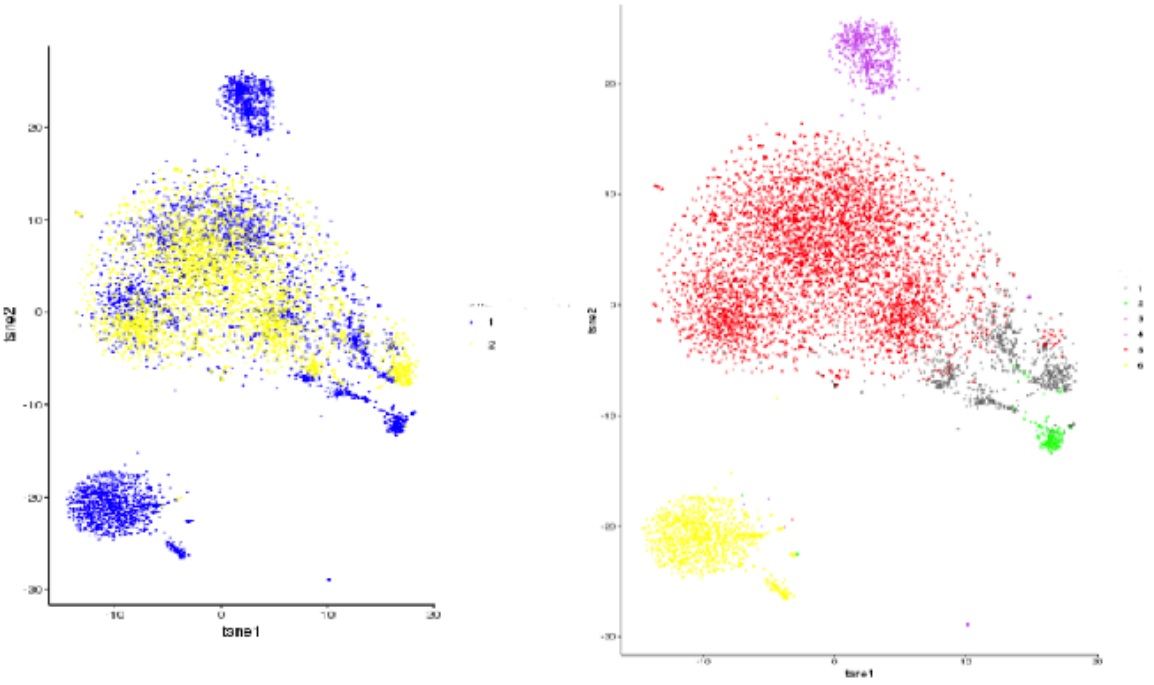


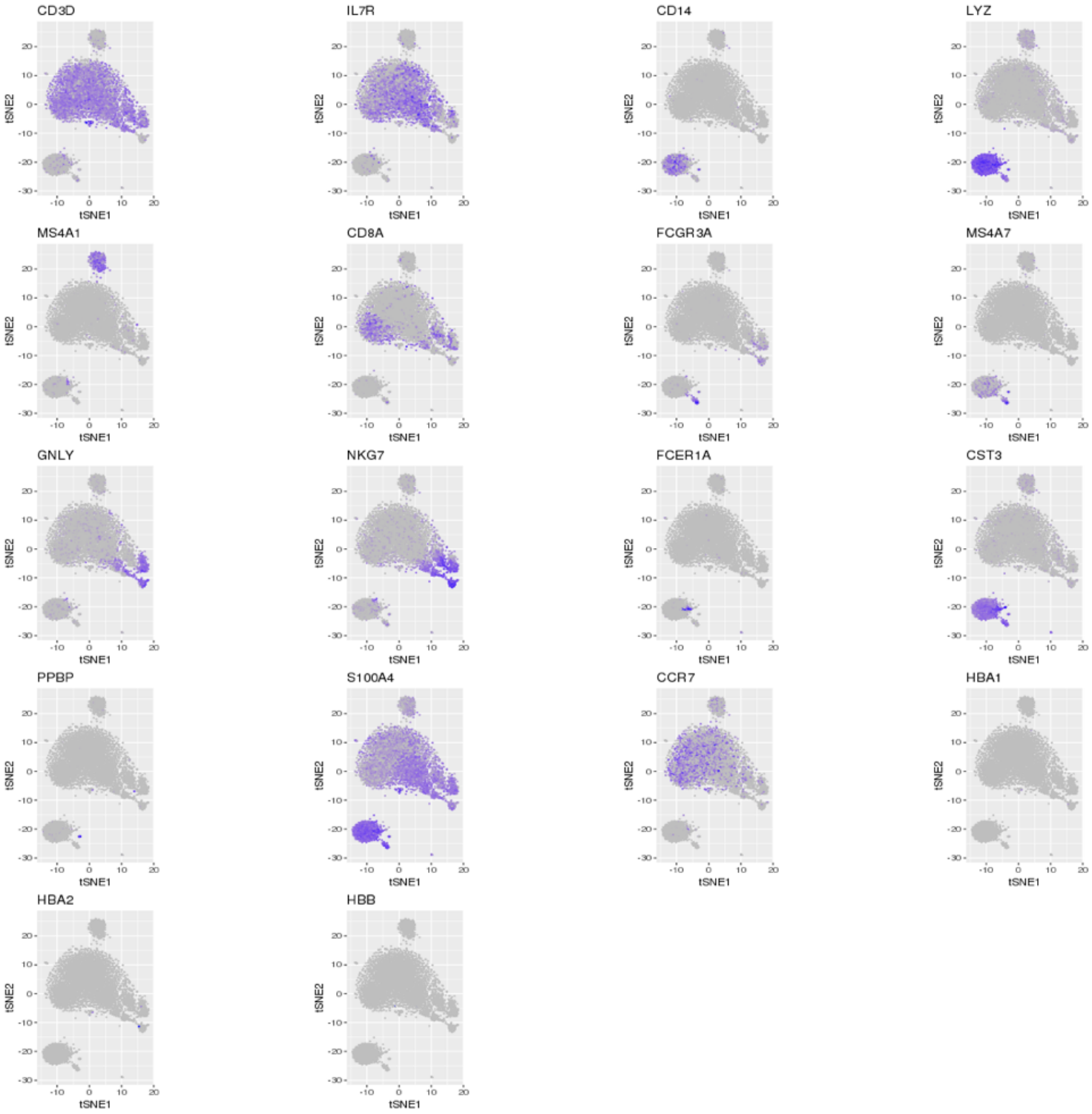
explanation.

(3) SCRUB corrected data:

Again, t-SNE was applied to data corrected by our method - SCRUB. Batch effect is corrected and the clustering can recognize cell type well, which can be revealed from gene mark plots. For example, CD3D+ represent T cell; LYZ or MS4A7 represent Monocytes cell; CD3D-NKG7+ represent NK cell, etc. Thus, we can label clusters in right panel: Red, grey – T cell, Yellow - Monocytes cell, Purple – B cell, Green – NK cell. Complicated T cell clusters can be further paraphrase. Because Grey cluster in right panel is NKG7+, it can be identified as cytotoxic T cell(gene mark: NKG7+CD3D+), while Red cluster is

regular T cell(gene mark: NKG7-CD3D+).





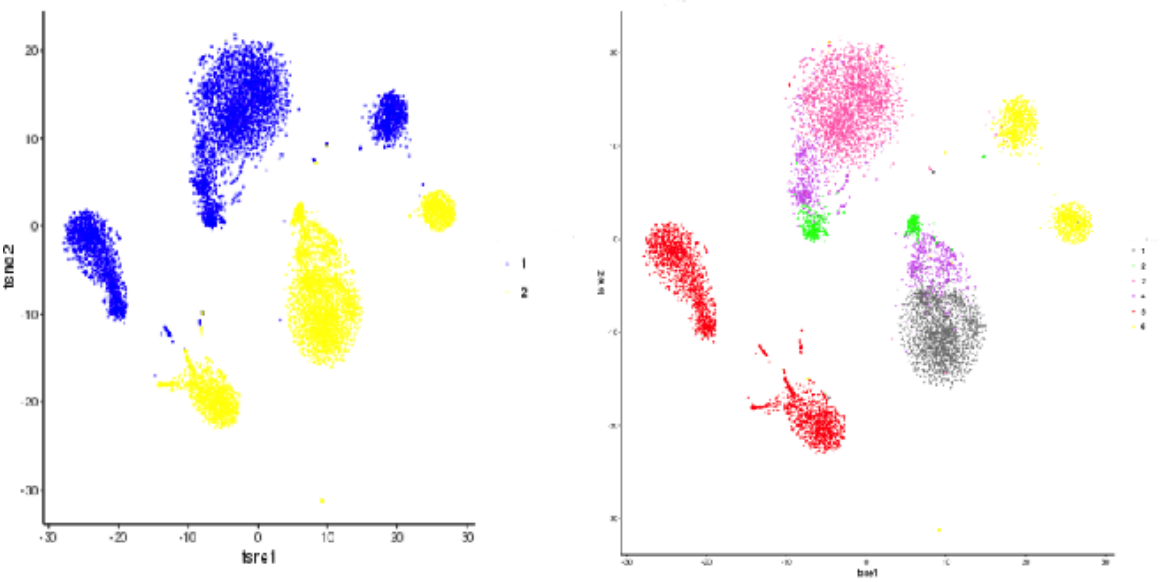
Here is the table of cell identifiers.

Markers	Cell Type
IL7R	CD4 T cells
CD14, LYZ	CD14+ Monocytes
MS4A1	B cells
CD8A	CD8 T cells
FCGR3A, MS4A7	FCGR3A+ Monocytes
GNLY, NKG7	NK cells
FCER1A, CST3	Dendritic Cells
PPBP	Megakaryocytes
S100A4	Naïve T
CCR7	Memory T

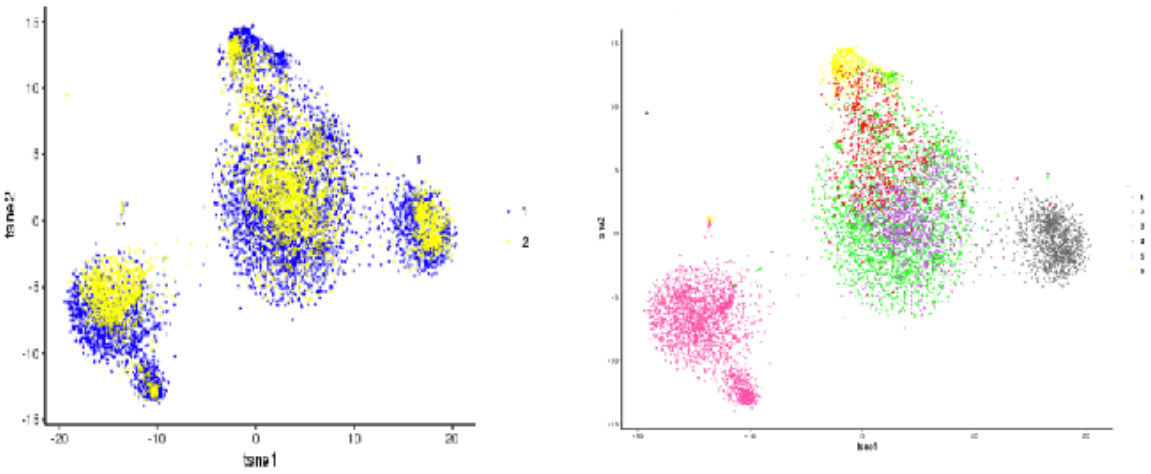
2.2 Two batches: PBMC6K and PBMC4K

Again, compare the performance of (1) no correction, (2) MNN correction and (3) SCRUB correction

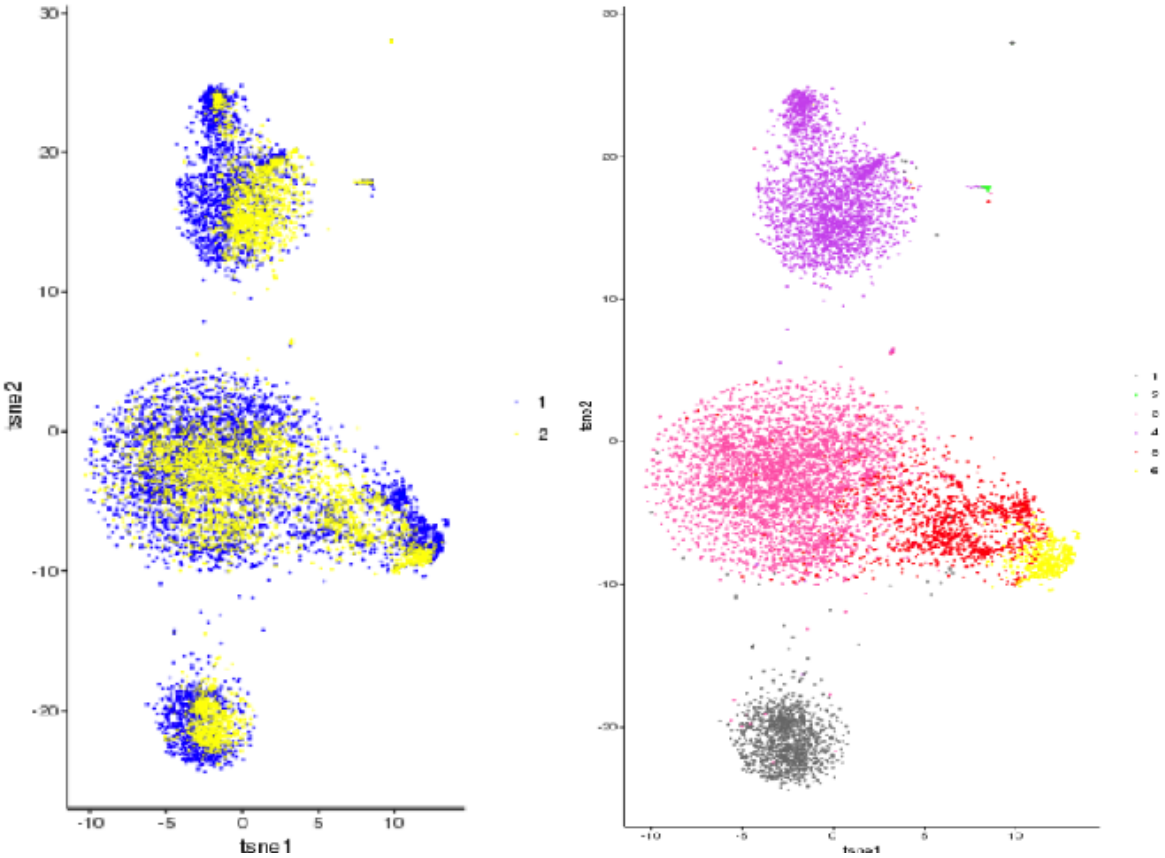
1. uncorrected data:

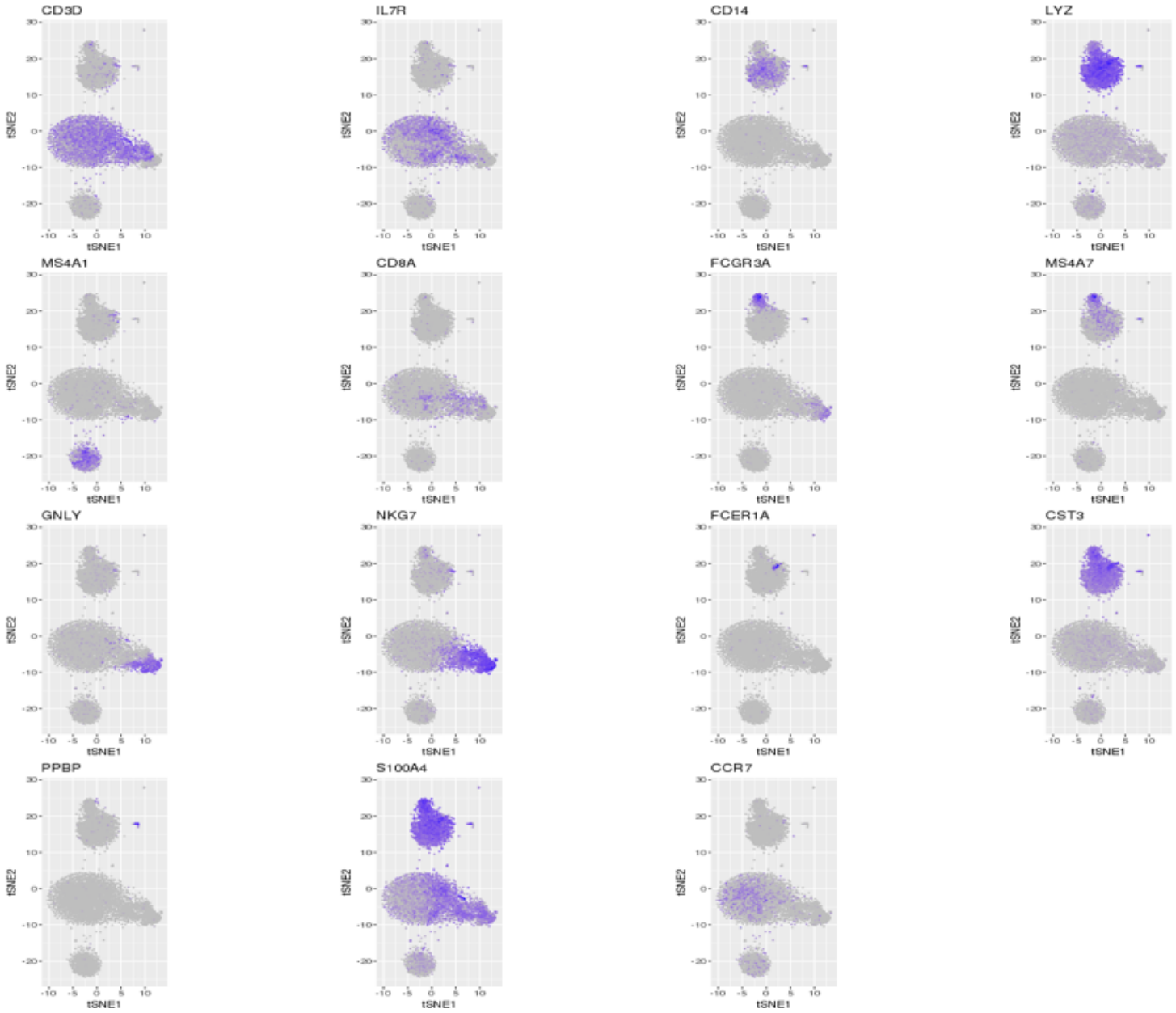


2. MNN corrected data:



(3) SCRUB corrected data:

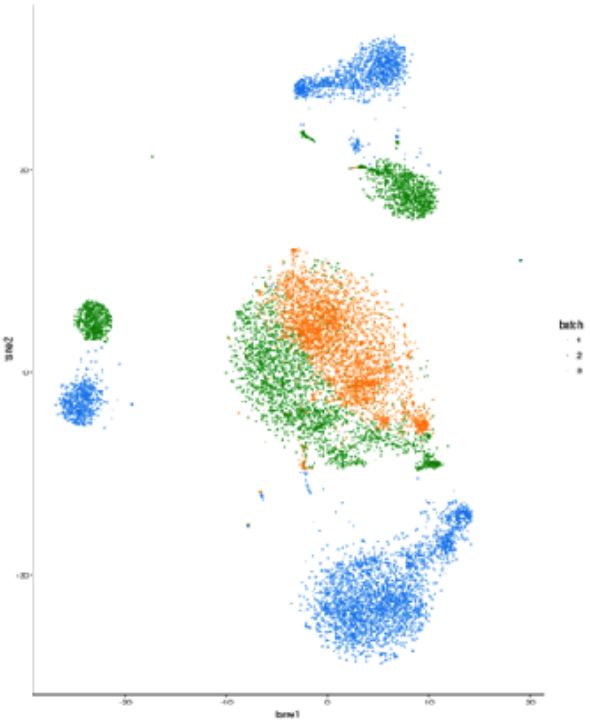


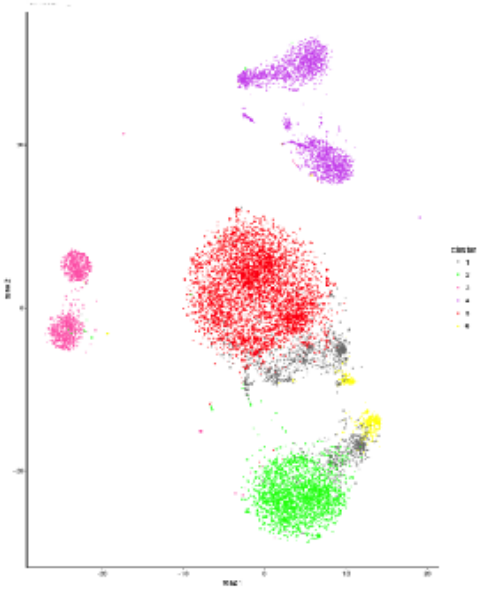


2.3 Multiple batches

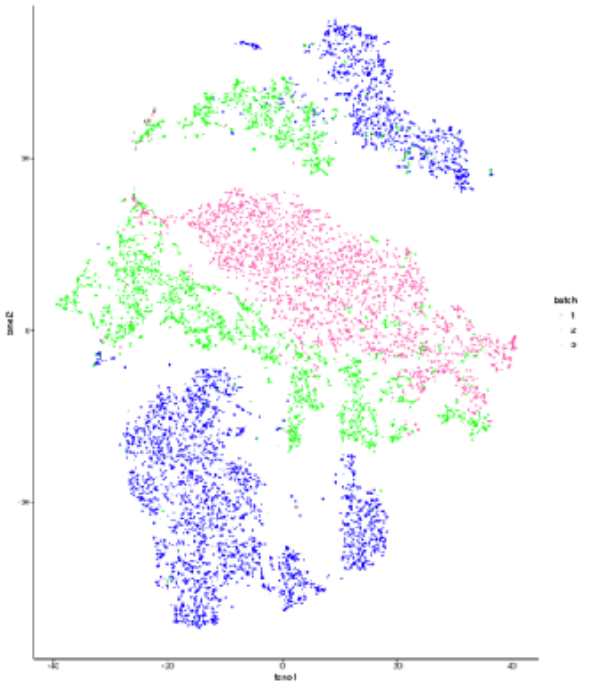
We compared the performance of (1) no correction, (2) MNN correction and (3) SCRUB correction for three batches case: PBMC6K, PBMC4K and Pan T

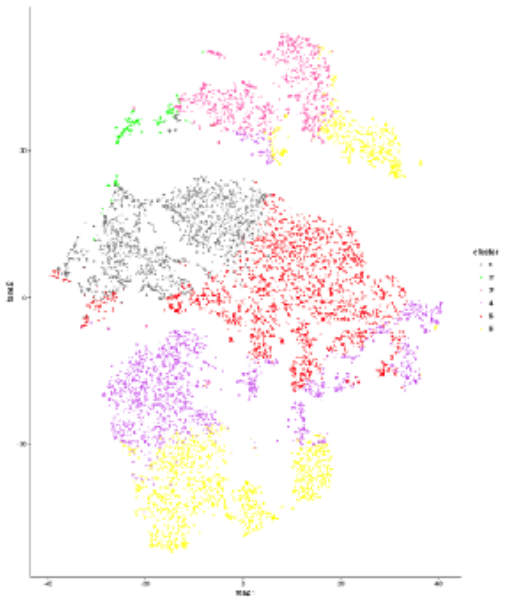
1. uncorrected data:



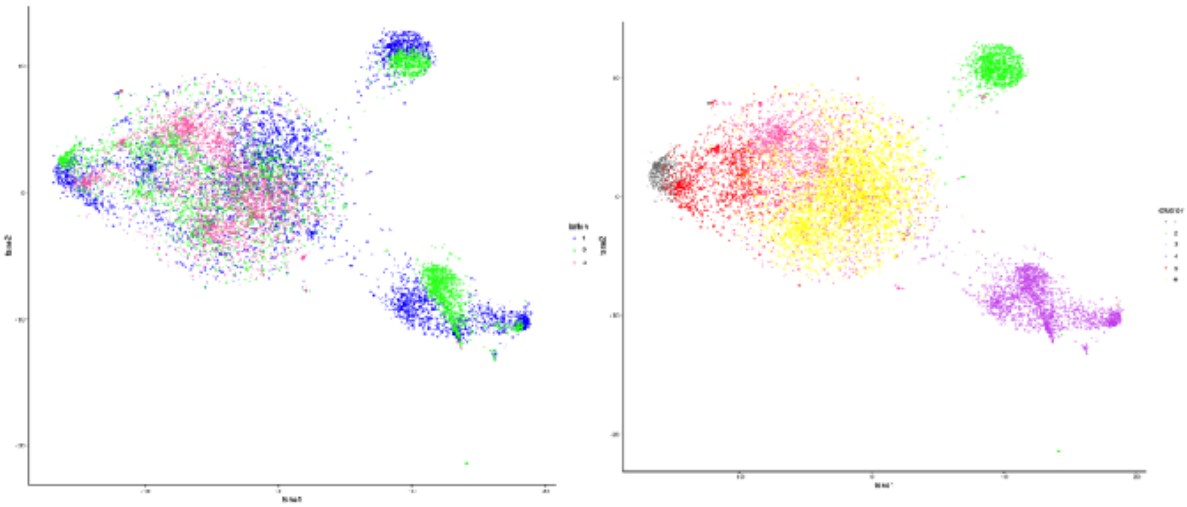


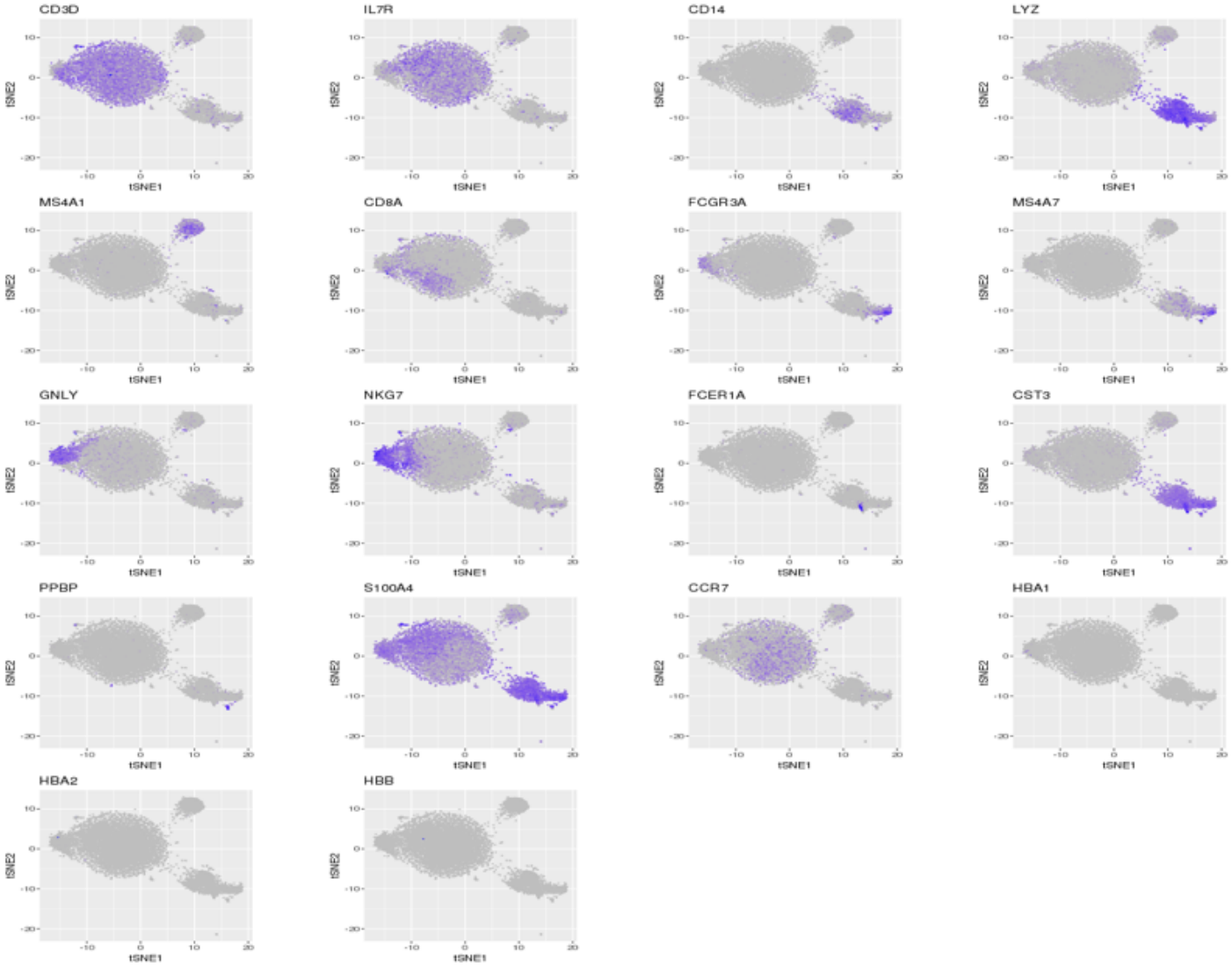
2. MNN corrected data:





(3) SCRUB corrected data:





Conclusion: We demonstrated the superiority of our method: SCRUB over other existed methods on two batches integration and on multiple batches integration. We were notified that cells from multiple batches mixed well. Meanwhile, the clustering pattern was clear and reflected the cell types in t-SNE projection. The clustering pattern were dominated by cells' phenotypes instead of technical variation through SCRUB process. One more noticeable advantage of SCRUB is on the detection of minor cell type, which is resulted from the distribution-based metrics we applied in SCRUB.