SCRUB: Correcting Batch Effect in Droplet-based Single Cell RNA-seq

Hanbo Sun[1]; Hyun Min Kang*[2]
[1]Department of Statistics , [2]Department of Biostatistics, University of Michigan

INTRODUCTION

Problem: 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.

Current Solutions: Recently, negative influences of the batch effect in scRNA-seq analysis have been studied (Hicks, et al., 2017 and Tung et al., 2017). The technical variability may fatally 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. Some methods have been proposed to correct batch effect for the microarray or bulk RNA-seq. However, these methods focus primarily on identifying unknown factors of variation and less on power in scRNA-seq. In the 2017, Shaham, et al., Haghverdi, et al., Satija, et al. proposed various integrated methods for scRNA-seq. However, neither of them works for multiple batches or complex archives. Besides, these methods have no biological interpretation and rely on strong assumptions. For example, MNN(Haghverdi, et al.) assumes that two batches have common cells, which is usually not true.

Our Solution: we present a novel model-based correction method: SCRUB to attenuate batch effect. We demonstrate the superiority of data integration and biological variation preservation of SCRUB over existing methods. Also, SCRUB is to eliminate batch effect by excluding only a very small proportion of genes, while remaining all the other genes without any modification. We anticipate it serves not only as a part of preprocess to correct batch effects, but also to general comparisons of scRNA-seq datasets and deeper understandings of critical genes that respond to perturbation, disease, and evolution.

METHOD

SCRUB (Single-Cell-focused Robust Unification of Batch effect) method starts from two or more digital expression matrices that contain observed UMI counts for each cell and gene. The digital expression matrix is first normalized following the procedure (Macosko et al., 2015, Satija, et al., 2015). After the normalization, SCRUB determines "**truly variable**"(TV) genes using multiple steps as described below, and use the TV genes to generate a t-SNE manifold and cluster for each cell. 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. However, we do not require them to have the identical distribution, as one batch may contain only a subset of cell types of the other batches. For these reasons, 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 between the batches. If any of these statistics do not show significant differences between the batches above the pre-defined threshold, we consider the gene as a TV gene.

We also include a gene into the TV gene list if it is HV gene within a specific batch. This step is meant to include genes expressed in cell types that appeared only in certain batches. For example, if a batch of peripheral blood mononuclear cell (PBMC) and another batch of T-cells are merged together, genes specifically expressed in B-cells and monocytes will have different distributions between the batches, and will be identified as batch-specific genes to be removed from the further analysis. These will inevitably result in reduced power to identify and cluster B-cells and monocytes. However, we can include these genes into the TV gene list by leveraging the fact that the genes will be highly variable within one cell type, but not within the other. The overall procedure is illustrated in **Figure 1**.

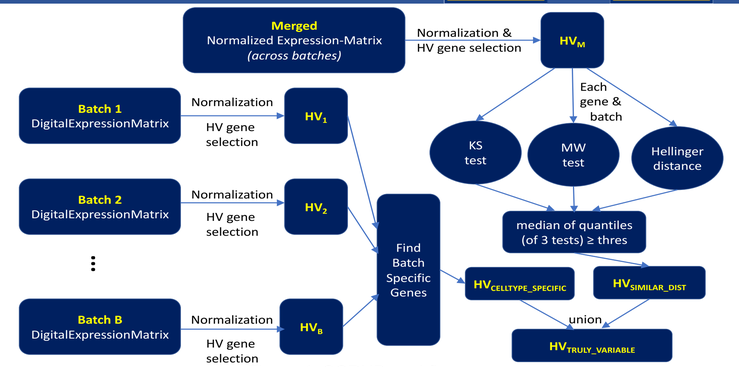

Figure 1. SCRUB workflow

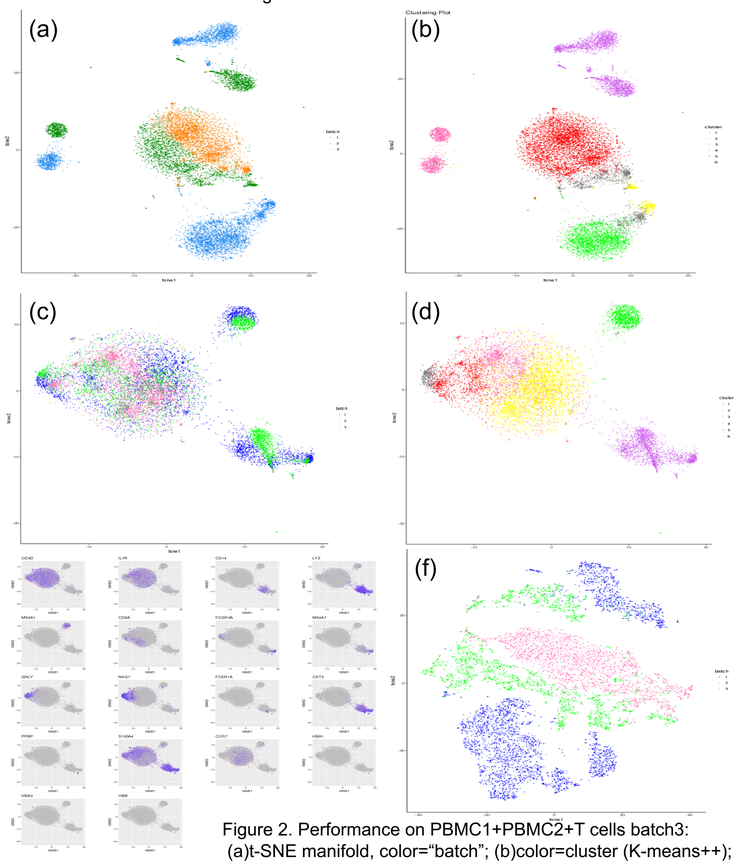

Figure 2. Performance on PBMC1+PBMC2+T cells batch3:
(a)t-SNE manifold, color="batch"; (b)color=cluster (K-means++);
(c)t-SNE manifold of SCRUB processed data; (d)color=cluster;
(e)Gene markers to identify cell types; (f)MNN[1] for comparison.

Contact

Hanbo Sun
Department of Statistics, University of Michigan
Email: hanbosun@umich.edu
Phone: (734) 881-0016

References

[1] Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2017). Correcting batch effects in single-cell RNA sequencing data by matching mutual nearest neighbours. bioRxiv, 1–18.
[2] Macosko, E. Z., Basu, A., Satija, R.,… McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, *161*(5), 1202–1214.
[3] Satija, A. B. (2017). Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. bioRxiv, 37(7), 802–807.
[4] Shaham, U., Stanton, K. P., Zhao, J., & Li, H. (2017). Removal of batch effects using distribution-matching residual networks. Bioinformatics, 33(16), 2539–2546.
[5] Tung, P.-Y., Blischak, J. D., Hsiao, C. J., Knowles, D. A., & Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. Scientific Reports, 7, 39921.
[6] Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. Nature Communications, 8, 14049.