Robust Normalization and Integration of Single-cell Protein Expression across CITE-seq Datasets

Ye Zheng [1], Seong-Hwan Jun [1], Yuan Tian [1], Mair Florian [1], and Raphael Gottardo [1, 2, 3*]

[1] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, USA. [2] Biomedical Data Sciences Center, Lausanne university hospital, Lausanne, Vaud, Switzerland. [3] Biomedical Data Sciences, University of Lausanne, Lausanne, Vaud, Switzerland. * Corresponding author.

Abstract

CITE-seq technology enables the direct measurement of protein expression, known as antibody-derived tags (ADT), in addition to RNA expression. The increase in the copy number of protein molecules leads to a more robust detection of protein features compared to RNA, providing a deep definition of cell types. However, due to added discrepancies of antibodies, such as the different types or concentrations of IgG antibodies, the batch effects of the ADT component of CITE-seq can dominate over biological variations, especially for the across-study integration. We present ADTnorm as a normalization and integration method designed explicitly for the ADT counts of CITE-seq data. Benchmarking with existing scaling and normalization methods, ADTnorm achieves a fast and accurate matching of the negative and positive peaks of the ADT counts across samples, efficiently removing technical variations across batches. Further quantitative evaluations confirm that ADTnorm achieves the best cell-type separation while maintaining the minimal batch effect. Therefore, ADTnorm facilitates the scalable ADT count integration of massive public CITE-seq datasets with distinguished experimental designs, which are essential for creating a corpus of well-annotated single-cell data with deep and standardized annotations.

Contact: raphael.gottardo@chuv.ch

Supplementary information: Supplementary data are available online.

1 Introduction

Cell type annotation in the single-cell analysis is a common problem required for biological interpretation and downstream statistical analyses. While much progress has been made in this area, current state-of-the-art techniques still suffer from major limitations, including lack of ground truth, reliance on limited reference datasets, and lack of standard annotations. With CITE-seq (Stoeckius *et al.*, 2017) and related technologies (Shahi *et al.*, 2017; Peterson *et al.*, 2017; Mimitou *et al.*, 2019), more than a hundred proteins' expression of individual cells can be directly measured in addition to RNA expression or epigenomics (Swanson *et al.*, 2020; Zhang *et al.*, 2022), facilitating robust and deep cell-type annotation. Despite their extraordinary potential, protein expression, measured by the antibody-derived tags (ADT) counts, is often analyzed using tools developed for single-cell RNA-seq even though the characteristics of the data are substantially different. ADT data, however, are much less sparse and have a unique density distribution pattern with a negative peak representing non-specific antibody binding and a positive peak indicating the enrichment of specific cell surface protein. Therefore scRNA-seq normalization tools are not directly applicable.

Apart from the library size and technical noises that other single-cell modalities normalization methods usually account for, ADT counts are also prone to the variability in antibodies, such as the antibody types and concentrations, which changes dramatically across studies and laboratories. Such batch effect hinders the aggregation of ADT data across CITE-seq studies. However, few methods have been developed for this modality. The most commonly used normalization approach is the centered log-ratio (CLR) (Stoeckius *et al.*, 2017; Hao *et al.*, 2021) which mostly can only take care of the library size variations. Mulè *et al.* (2022) introduced the denoised and scaled by background (DSB) normalization method developed specifically for ADT data. DSB can improve the visualization of protein expression within an experiment but fails to properly normalize data across experiments because it only focuses on aligning the negative expression peak. In addition, it requires data from empty droplets to estimate the background ambient noise, but public datasets are usually pre-filtered, and low-quality cells are removed for the sake of data storage and efficient data sharing. Additionally, Harmony has established effectiveness for scRNA-seq batch removal Korsunsky *et al.* (2019); Tran *et al.* (2020) and CytoRUV demonstrates advantages in removing library size batch effects while strengthening biological signals for the flow cytometry data Trussart *et al.* (2020). However, neither is tailored to the high-throughput sequencing protein expression count data. Here, we proposed a functional data analysis normalization method, called ADTnorm, for ADT count data of CITE-seq related assays, building on methods that were originally conceived for cytometry data, including fdaNorm and gaussNorm (Hahne *et al.*, 2010) that try to align landmarks in protein density profiles.

2 Materials and methods

To demonstrate the ADTnorm normalization and integration procedure and benchmark with existing normalization methods designed for scRNA-seq, flow cytometry, and CITE-seq ADT component, 15 public CITE-seq datasets are utilized (Supplementary Table 1). The pipeline for the surface protein expression normalization and integration across CITE-seq datasets starts with CITE-seq sequencing data alignment and quality check to remove low-quality cells, such as empty droplets, doublets, and cells with high mitochondrial gene expression (Fig. 1A). Note that the scRNA-seq component of CITE-seq in this paper is only utilized in the data quality

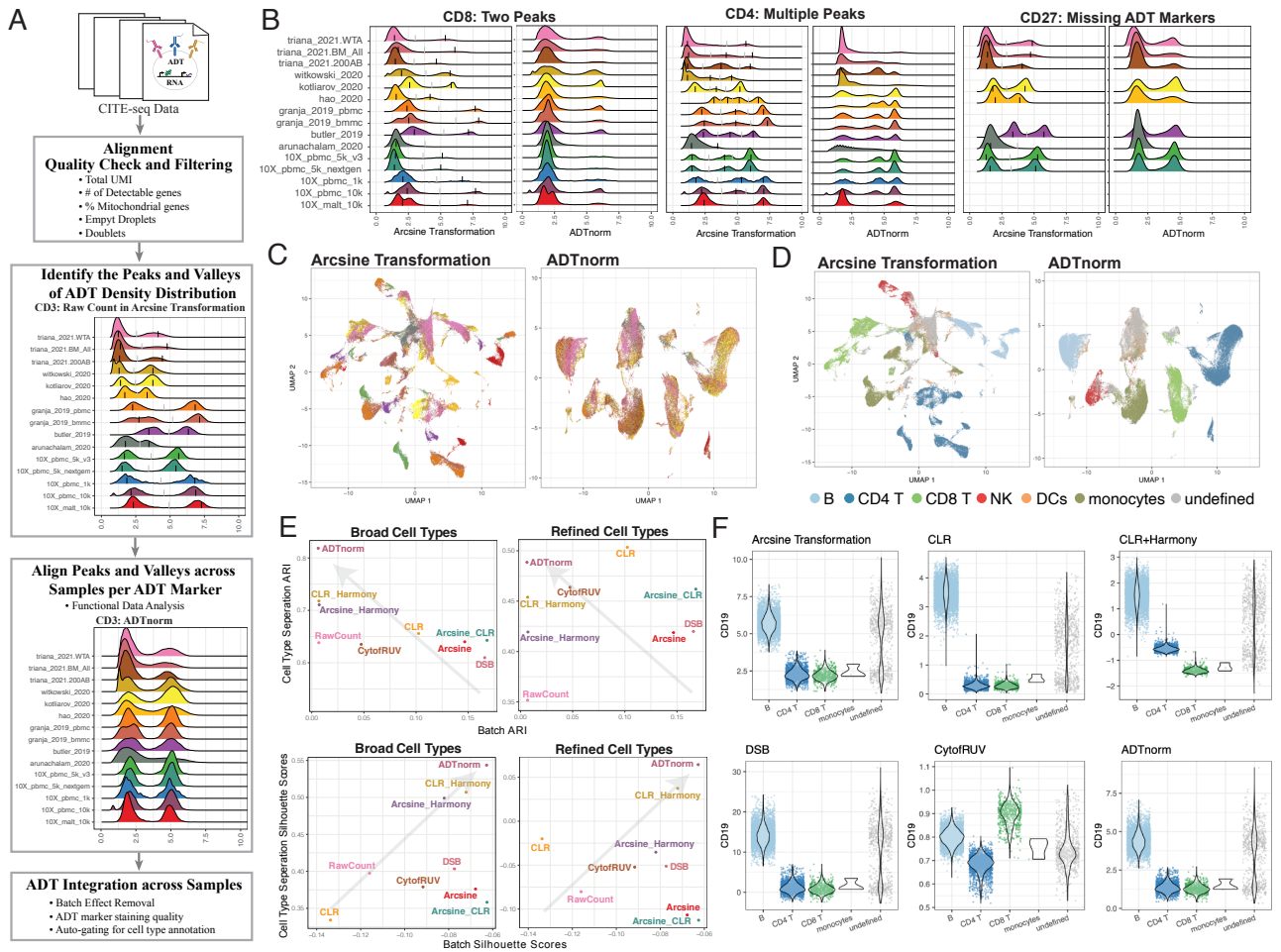

Fig. 1. A. Overview of ADTnorm normalization and integration pipeline. B. Comparison of ADT counts density distribution in raw count scaled by Arcsine transformation and after ADTnorm normalization for different types of surface proteins across CITE-seq datasets. C. Comparison of the low-dimensional embeddings depicting the batch effect with each CITE-seq dataset as one batch. The colors of the plotting symbols in the UMAP visualization represent the datasets with the same datasets matching colors in B. D. Comparison of the low-dimensional embeddings depicting the cell type separations before and after ADTnorm normalization. E. Benchmarking different ADT counts scaling and normalization methods regarding the cell type separation and batch effect elimination, quantified by the Adjust Rand Index (ARI) and Silhouette scores. Grey arrows point to direction where the method performance is better. F. CD19 surface protein expression can be misleading manipulated after normalized by Harmony and CytoRUV in the 10X MALT dataset.

check and necessary cell filtering. The downstream analysis only includes the ADT counts of CITE-seq data. ADTnorm normalizes ADT counts of cells passing quality check by identifying the peaks and valleys of the ADT count density distribution. Subsequently, functional data analysis is implemented to align the negative peak, valley, and positive peak of ADT density distribution across samples from different CITE-seq datasets. The resulting normalized ADT counts are batch effect free and ready for across-study aggregation. Additionally, the landmarks, including peak and valley locations, detected during ADTnorm normalization can also be leveraged to evaluate the ADT markers' staining quality and facilitate the automatic gating for cell-type annotation. Please refer to the Supplementary Note for the ADTnorm model and implementation procedure details.

3 Result

Arcsine transformation is widely used in processing flow cytometry data for better density visualization, which is also leveraged to show the raw count of multiple CITE-seq datasets. We benchmarked ADTnorm with CLR (Stoeckius *et al.*, 2017), Harmony (Korsunsky *et al.*, 2019), DSB (Mulè *et al.*, 2022), CytofRUV (Trussart *et al.*, 2020) in terms of batch effect elimination and cell type separation. Fig. 1B and Supplementary Figs. 1-2 demonstrate the successful removal of variations for negative peaks and positive peaks in each surface protein marker after ADTnorm normalization compared to Arcsine Transformation and CLR scaling. Additionally, ADTnorm can accommodate surface protein markers that are not profiled in all the CITE-seq datasets to be integrated, which have to be discarded by other normalization methods such as Harmony and CytofRUV that cannot handle the missing data (Supplementary Figs. 3-4). UMAP visualization further confirms the removal of the variations across datasets while preserving the biological signals across cell types at both broad and refined cell-type annotation levels (Fig. 1 C-D, Supplementary Fig. 5). In addition to the low-dimensional embeddings visualization, we also leveraged the Adjust Rand Index (ARI) and Silhouette scores to quantitatively evaluate the cell type separation and the remaining batch effect across studies. Fig. 1E illustrates that ADTnorm accomplished the best cell type separation for both evaluation metrics, where the true cell type labels were obtained by an orthogonal manual gating on the surface protein in combination with the cell type annotation provided in the data source paper. At the same time, ADTnorm achieves the minimal batch effect as good, if not better than, as Harmony approach.

Apart from obtaining batch-free low-dimensional embeddings, biologically reasonable normalized ADT counts are also needed for further cell-type annotation and standardization. When scrutinizing the ADT

counts after normalization by different methods, Harmony and CytofRUV tend to manipulate the surface protein counts into an improper value range (Fig. 1F). For instance, in the 10X MALT study, CD19 expression is significantly lower for CD8 T cells than CD4 T after Harmony processing. CytofRUV, on the other hand, even assigns a much higher CD19 value to CD8 T cells compared to B cells. Additionally, CLR reduced the separation between B cells and CD4 T cells regarding the CD19 expression. Similar issues are observed on other protein markers, such as CD3, CD4, CD8, CD25, CD45RA, and CD56, which are all critical lineage markers (Supplementary Figs. 6-7). This re-emphasizes the importance of utilizing the normalization methods tailored to CITE-seq ADT counts.

In summary, ADTnorm provides a fast, accurate and scalable approach to normalizing the ADT counts of CITE-seq for within-study batch removal and across-study integration. The method is also extended to incorporate cytometry data with the CITE-seq data by removing the discrepancies across technologies.

Acknowledgements

We acknowledge the Scientific Computing Infrastructure at Fred Hutch funded by ORIP grant S10OD028685, the J. Orin Edson Foundation, the Translational Data Science Integrated Research Center of the Fred Hutch, and NIH U19AI128914.

References

Hahne, F. *et al.* (2010). Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A: The Journal of the International Society for Advancement of Cytometry*, **77**(2), 121–131.

Hao, Y. *et al.* (2021). Integrated analysis of multimodal single-cell data. *Cell*, **184**(13), 3573–3587.

Korsunsky, I. *et al.* (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, **16**(12), 1289–1296.

Mimitou, E. P. *et al.* (2019). Multiplexed detection of proteins, transcriptomes, clonotypes and crispr perturbations in single cells. *Nature methods*, **16**(5), 409–412.

Mulè, M. P. *et al.* (2022). Normalizing and denoising protein expression data from droplet-based single cell profiling. *Nature Communications*, **13**(1), 1–12.

Peterson, V. M. *et al.* (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nature biotechnology*, **35**(10), 936–939.

Shahi, P. *et al.* (2017). Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. *Scientific reports*, **7**(1), 1–12.

Stoeckius, M. *et al.* (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, **14**(9), 865–868.

Swanson, E. *et al.* (2020). Tea-seq: a trimodal assay for integrated single cell measurement of transcription, epitopes, and chromatin accessibility. *bioRxiv*.

Tran, H. T. N. *et al.* (2020). A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, **21**(1), 1–32.

Trussart, M. *et al.* (2020). Removing unwanted variation with cytofruv to integrate multiple cytof datasets. *Elife*, **9**, e59630.

Zhang, B. *et al.* (2022). Characterizing cellular heterogeneity in chromatin state with sccut&tag-pro. *Nature Biotechnology*, pages 1–11.