Resistant Fit Regression Normalization for Single-cell RNA-seq Data

Da Kuang
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, USA
kuangda@seas.upenn.edu

Junhyong Kim
Department of Biology
University of Pennsylvania
Philadelphia, USA
junhyong@sas.upenn.edu

***Abstract*—All mRNA quantification, including single-cell mRNA sequencing, requires normalization to correct for technical variation and to make measurements of two or more samples comparable. The choice of normalization method impacts the downstream analysis. All common approaches (applying scaling factors, variational inference, and quantile regression) currently focus on removing technical variations but ignore localized variations of biological origin. To address this problem, we propose a new framework to normalize for technical effects while also aligning RNA-seq datasets for a biologically meaningful comparison. We present an iterative optimization method using the notion of a resistant fit regression to isolate localized perturbations. Both simulated data and real data are resistant-fit normalized and compared with popular normalization methods. This comparison shows that the resistant fit works better under localized biological variations.**

***Keywords*-Single-cell, RNA-seq, Normalization, Robust Regression, Resistant Fit**

I. Introduction

A primary goal of single-cell RNA sequencing (scRNA-seq) is to estimate the molecular counts of all the mRNA types in a cell (i.e., the cell's transcriptome). As the essential first step, normalization attempts to correct various factors that affect the mRNA counts such that measurements from two or more experiments can be compared to each other. The factors that affect the measurement variations can be classified into technical variations and biological variations. The technical variations reflect noise, reduced sensitivity, and bias introduced during sample processing steps such as mRNA capture, reverse transcription, PCR amplification, fragmentation, and sequencing. Additionally, Single-cell data compiled from multiple experiments often demonstrate significant variations or "batch effects" because of the difference in capturing times, handling personnel, reagent lots, types of equipment, and technology platforms. [1] The biological variations represent differences in the mRNA numbers that arise due to inherent biological differences such as differentiated cell types, different life cycle stages, responses to genetic or environmental changes.

Recently, unique molecular identifiers (UMI) have been widely used in RNA sequencing, which attaches an artificially designed random sequence to each mRNA molecule. The random sequences are long enough to assume uniqueness, and the UMI-protocol removes biases related to amplification since multiple reads associated with the same UMI are collapsed into a unique count. However, other aforementioned factors during sequencing still contribute to considerable technical variations. Most of the normalization methods are focused on removing the effect of sequencing depth (= sampling depth) per sample or per cell, which in the case of scRNA-seq can vary by orders of magnitude.

Existing normalization methods for sampling depth fall into three broad types. The first type involves estimating a single scalar factor (so-called size factor) for each sample (cell). For example, Seurat [2] and Scanpy [3] use the total sequencing depth as size factors. Scran [4] pools cells with similar library sizes and uses the summed expression values to estimate pool-based size factors. These methods assume that a single scalar quantity is sufficient to correct for most of the technical variation. The second class of normalization approaches applies a probabilistic model to the molecule counts to obtain a model-corrected estimation. For single-cell data, it is common to assume "drop-outs" where zero-count classes arise from catastrophic loss. A common approach involves adding a zero-inflation term coupled to a Negative Binomial distribution (ZINB). For instance, scVI [5] uses the ZINB noise model in a variational inference auto-encoder for normalization and dimensionality reduction. Also, DCA [6] normalizes data based on the ZINB model using a denoising autoencoder. The first two normalization classes involve using the identities of the mRNA molecules (which we will simply call "genes"), assuming that the same gene measured in two samples should be comparable. The third class of approaches involves identifying the quantiles of the transcriptome rather than the genes. SCnrom [7] uses quantile regression to estimate scale factors within each group of genes. SCTransform [8] fits a generalized linear model for each gene and regularizes the model by the genes with similar abundances.

All three classes of methods are focused on technical variations. However, transcriptome analysis requires correcting for more than just technical variations. For example, Li et al. give the following example [9]:

Suppose we have 101 genes in two samples, g1 to g101.

In sample 1, we have the following RNAseq counts: (100, 100,...,100, 0); that is, all 100 except for g101 = 0. In sample 2, we have (80, 80,...,80, 2000); that is, all 80 except for g101 = 2000. Both samples have the same number of total reads so normalization by the inverse of the total reads would yield (0.01, 0.01,...,0) and (0.008, 0.008,... 0.008, 0.2).

Similarly, Robinson and Oshlack [10] also point out problems when one cell has 1,000 non-zero genes while another has 10,000 non-zero genes. A proportional quantity of 0.01 means very different things in these two cases. In the first example, some biological responses might have changed the expression of gene g101; in the second example, cell differentiation might have led to very different numbers of non-zero gene expression. In both cases, biological factors, not technical factors, underlie the variation. Any biologists would "normalize" the first example by correcting the first sample by a 0.8 size factor, making genes g1 to g101 equivalent in both datasets. Thus, the notion of normalization should involve correcting for technical variations and "aligning" across biological variations such that two or more datasets become comparable to each other.

In this manuscript, we present a new framework to normalize for technical effects and align RNA-seq datasets in a biologically meaningful comparison. We first formalize the notion of transfer functions in RNA quantification and define an objective for normalization. Then we introduce the concept of resistant fit regression, which aligns two or more datasets allowing for biological variations.

II. Method

A. Normalization Framework

Suppose there are m cells with g genes in an experiment, for each cell i, there is a vector $v_i \in \mathbb{R}^g$ describing the expression levels of genes. We call the subspace $V_{bio} = \{v_i, i \in [1, N]\}$ the *biology space*. Values in the *biology space* are unknowable to us. We now assume that RNA-seq measurement on a cell i is an application of the measuring function f_i mapping v_i to a *measurement space* consisting of the integer counts of UMIs after sequencing. Ideally we would like to invert the measuring function f_i, but we assume this is impossible (Fig.1a). Instead we wish to construct a *normalized space* by mapping f_i with a normalization function ϕ_i such that

Normalizing implicit rule (NI). *if $x_1 = x_2$ in biology space, then $\phi_1(f_1(x_1)) = \phi_2(f_2(x_2))$.*

The key component of this idea is to find a set of normalization functions that satisfy NI. Considering RNA-seq in more detail, multiple factors affect the shape of a measurement function:

- Experimental details governing the total number of molecules in a sequencing library

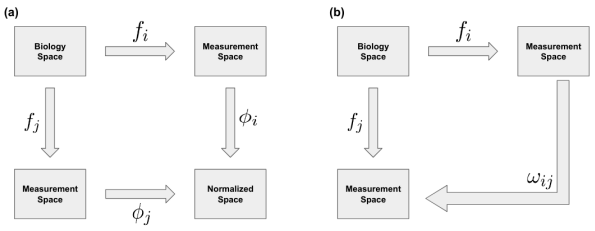

Figure 1. Schematic of the normalization framework. f_i and f_j are two different measuring functions.(a) ϕ_i and ϕ_j are two transfer functions mapping observations to Normalized Space. (b) ω_{ij} is an normalizing function matching one measurement to another where $\omega_{ij} = \phi_i \circ \phi_j^{-1}$.

- Experiments details governing the differential gain/loss of RNA molecules of a particular type
- Characteristics of the RNA molecules
- Composition of cell types in the sequencing library

In the worst case, the above considerations could mean that the measurement of every RNA molecule involves a unique measurement function each time a measurement is made. If this is the case, it would be impossible to construct a normalization function that satisfies NI. In the best-case scenario, the measurement function might be a simple function whose parameters can be obtained from experiment configuration such as the total number of reads. Here we propose that the measurement function is the same for a given cell. Also for simplicity, for different cells, we assume that they only differ by a scalar size factor representing the sequencing depth. For this simple case, it can be shown that the optimal fit (see below) normalization can be obtained by a simple regression to estimate a size factor for a normalization function $\phi_i = \lambda_i f_i$.

So far, the above treatment is the same as the standard methods for estimating a single size factor. Here, we now consider the case when biological variations cause localized significant variations (e.g., large over-expression of 20% of the genes). For such a case, it is clear that we should only use a subset of the genes that did not experience localized large deviation as the biological feature set to estimate the size factor for the normalization function. In the next section, we introduce the notion of a resistant fit for this goal.

B. Normalization for Two Transcriptions

We propose an normalizing function ω_{ij} to match the measurements of any two cells through the normalized space as Fig.1(b). Suppose we have observations of two cell $\boldsymbol{x} = \{x_1, x_2, \cdots, x_g\}$ and $\boldsymbol{y} = \{y_1, y_2, \cdots, y_g\}$, where g is the number of genes in the library, so we have g points, $(x_1, y_1), (x_2, y_2) \cdots (x_g, y_g)$. We find a normalizing function by minimizing the objective function,

$$O(\lambda) = ||\boldsymbol{x} - \omega_\lambda(\boldsymbol{y})||^2$$

where ω_λ is the normalizing function action on the measurement space of $\boldsymbol{y}$ and $||\ ||$ is the Euclidean norm. As mentioned, for simplicity, in this paper we only consider a two parameter linear function for the normalizing function. We use robust regression to resistant fit the normalizing function while excluding some of the extremely distorted data. The regression is based on a subset of genes as the biological feature set S. The size of the set is p, which is a user predefined parameter. By default, we provide the choice of setting $p = 0.7$ or finding the p that corresponds to $1.25\times$(median of the residuals). Even though p is given, the types of genes in S could be different from cell to cell.

Therefore, instead of choosing a fixed piece-wise function as the M-Estimator [11], we use Expectation-maximization (EM) algorithm as a dynamical estimator while minimizing the objective function. In the initial E step, a robust linear function is fitted using repeated medians based on all the points [12]. Then in the subsequent M-step, all the points are ranked by the ascending of residual respect to the linear model from the previous step. Only the top p points are selected for the next linear regression as the E step. Then successive M and E steps are repeated until the sum of squared residuals (SSR) is less than a reasonable tolerance δ.

C. Normalizing Multiple Transcriptomes

The objective function can be extended to multiple pairwise transcriptomes but simultaneous optimization will be computationally complex. Instead, we suggest an iterated normalization where we match all the measurements to a centroid transcriptome.

Let $T_1, T_2, \cdots, T_n$ be n different transcriptomes for normalization and T_c is defined as the centroid transcriptome. Given the natural counting scale of RNAseq data, we suggest using arithmetic average to compute the centroid. The following algorithm is computed for a certain epoch. In our experiment, usually 1 to 3 epoch would be sufficient to normalize a dataset.

Algorithm 1 Normalizing Multiple Transcriptomes

```
for j in epoch do
    T_c = arithmetic average (T_1, T_2, ..., T_n)
    for i in 1:n do
        Normalize T_i to T_c
    end for
end for
```

III. Result

A dataset of 33,148 human peripheral blood mononuclear cells (PBMC) was selected to examine Resistant Fit normalization performance. The dataset can be freely accessed from 10x Genomics and is characteristic of current scRNA-seq experiments [8]. Comparing with other popular methods, such as Log-scale, SCRAN, SCTransform (SCT), and scVI, we show that resistant fit normalization (RF Norm) works with downstream analysis (dimension reduction, clustering, and differentially expression). Moreover, enrichment analysis of DE genes from resistant fit normalized data shows more biological sense.

A. Resistant Fit Normalization Works with Dimension Reduction

The whole 33k PBMC dataset was used to examine the performance of Resistant Fit Normalization. Christoph et al. provide a cell subtype list for this dataset in SCTransform paper [8]. Different normalization methods were used on the full PBMC dataset, then PCA and UMAP were applied to the normalized values for dimension reduction. The embedding representations were colored by biological heterogeneity in Fig 2. Resistant Fit Normalization shows a similar pattern compared with Log-scale, SCT, SCRAN, and scVI.

B. Resistant Fit Normalization Enhances DE Analysis

Next, we use the number of false-positive (FP) differential expressions (DE) to demonstrate that sequencing depth can confound scale-factor based normalization but have less effect on Resistant Fit Normalization. For each sub-cell type in the PBMC dataset, cells were evenly separated into two groups. In one of the groups, cell counts were randomly downsampled to 50% of the original expression level. Therefore, ideally, we expect no DE genes between groups since they are biologically identical. RF Norm, log-scale norm, SCRAN, and SCT, as well as scVI, were applied to each pair of groups.

Then we performed DE analysis on each pair of normalized sub-cell types. For the first four normalization methods, the Wilcoxon rank-sum test [13] was used where the significant threshold was FDR less than 0.05, and log fold changes less than 0.3. For scVI, auto-encoder was trained with the default architecture (1 layer, 128 hidden units, and 10 latent variables), and $|\ln \text{Bayes factor}| > 3$ was the significant threshold as recommended [5].

In Table I, most of the normalization methods have zero FPs when the population of subtypes is notable, while more FPs are discovered with the decreasing cell number. It indicates that technical variations are evenly distributed and can be canceled out between large communities. Nevertheless, the technical variations introduce unwanted heterogeneity to small pairs of subtypes (DC, NK Bright, Mk, and pDC). SCRAN reduces the technical noise and performs better than Log-scale but still has relatively high FPs compared with others. RF Norm and SCT discover significantly fewer FPs revealing the robustness to technical variations. The difference is that RF Norm has a better performance for small groups, while SCT is more accurate at a moderate population size (NK Dim and Mono CD 16). scVI is

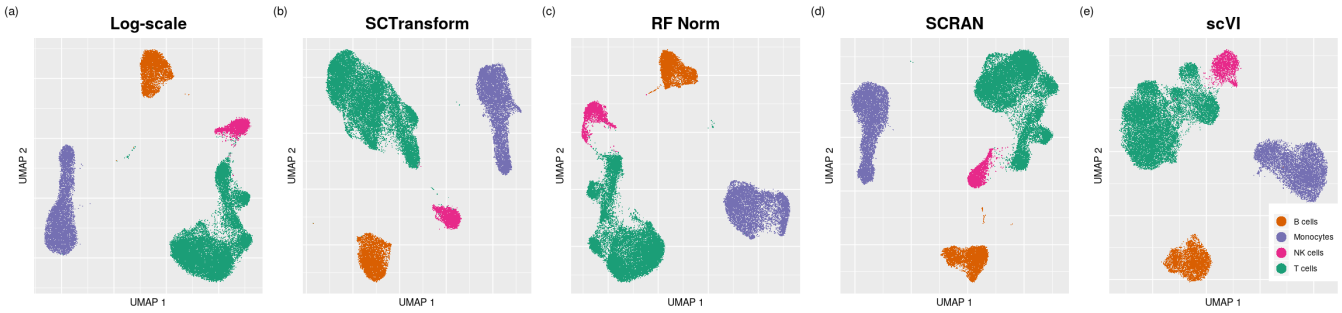

Figure 2. UMAP Embedding of the 33k PBMC dataset using popular normalization methods. All five low-dimensional representations show similar pattern respect to cell type.

Table I

Cell Types	N Cells	Normalization Methods				
		RF Norm	Log-scale	SCRAN	SCT	scVI
CD4 Memory	5961	0	0	0	0	1
CD4 Naive	4431	0	0	0	0	1
Mono CD14	5551	0	0	0	0	0
B Pre	2886	0	0	0	0	1
CD8 Memory	2067	0	0	0	0	1
CD8 Naive	1917	0	0	0	0	2
NK Dim	1655	3	0	0	0	1
Mono CD16	1475	2	0	1	0	1
CD8 Effector	1453	0	0	0	0	4
B Pro	953	0	1	1	0	0
DC	204	6	57	27	1	1
NK Bright	217	4	59	14	2	0
Mk	140	16	119	74	9	0
pDC	96	18	203	116	30	8

Number of false positive genes from differential analysis on the different sub cell types normalized by current popular approaches

not affected by community size and has a good overall performance.

C. Resistant Fit Normalization Gives More Biological Sense

CD4 memory T cells and CD8 Effector T cells were subsetted from the PBMC dataset to further demonstrate the impact of different normalization methods. We did DE analysis on normalized counts with the same significance threshold as described above. Based on Table II, SCTransform (SCT), scVI, and Resistant Fit Normalization (RF Norm) detected fewer DE genes than Log-scale and SCRAN. Moreover, the DE genes detected after Resistant Fit is a subset of that found with Log-scale and SCRAN.

In Table II, the down regulated DE genes are the ones with higher expression in CD8 Effector T cells. After RF Norm, we discovered 32 down-regulated genes. The same DE analysis detected 45 down regulated DE Genes with SCT. It turns out the 32 down-regulated gene from RF Norm is a subset to the 45 down-regulated DE of SCT. Enrichment analysis was conducted using Enrichr [14] to examine the function of DE genes. Pathway Enrichment based on *KEGG 2019* on the 32 RF Normalized down-regulated genes shows that "Antigen processing and presentation" is enriched in the gene list. Ontologies Enrichment based on *Gene Ontology Molecular Function 2018* shows that "MHC class I protein binding" is significantly enriched in those 32 genes. The enrichment results are consistent with the fact that those 32 genes are expressed more in CD8 Effector T cells.

The same analysis was applied to the 45 SCT down-regulated genes. We found "Primary immunodeficiency" and "MHC class II protein binding" as the enriched pathway and molecular function. Compared with RF Norm, the result of SCT down-regulated genes is less consistent with known differential gene levels in CD8 effector T cells. The extra 13 genes' expression level is more likely due to false positives created by quantile regularization by SCTransform.

The up-regulated DE genes are the ones up-regulated in CD4 memory T cells. There are 73 up-regulated genes associated with RF norm and 41 up-regulated genes with SCT. The RF Norm up-regulated genes cover 70% of the SCT up-regulated genes. We conducted enrichment analysis on RF Normalized up-regulated genes, SCT up-regulated genes, and the overlapping genes. However, we did not find any clearly enriched pathway previously related to CD4 cells.

Given that other methods have difficulty revealing the up-regulated genes in a biologically meaningful manner, discovering fewer genes with scVI normalization potentially suggests that it might have higher degree of precision. However, as a deep learning model, the default architectures of scVI may not be able to find an accurate latent representation for this dataset. Besides, the latent representation of scVI confounds genes' abundance and leads to more false positives for down-regulated genes. Only 30% of its discoveries are in common with RF Normalized down-regulated genes related to CD 8. The rest of the down-regulated genes were difficult to interpret.

In sum, Resistant Fit Normalization enhances the DE analysis and helps find a group of 32 genes as the markers of CD8 Effector T cells.

IV. Conclusion

Here, we present a new framework for normalization while making the samples biologically comparable. Resistant fit based on a dynamic biological feature set is used to estimate a robust size factor for each cell. The size factor is estimated excluding large deviations which helps make

Table II

CD4 Memory vs. CD8 Effector			
Method	Total	Down	Up
Log-scale	724	161	563
SCRAN	210	70	140
SCTransform	86	45	41
SCVI	87	60	27
Resistant Fit	105	32	73
Shared with Log-scale	104	24	69
Shared with SCRAN	105	32	73
Shared with SCT	61	32	29
Shared with SCVI	29	19	10

Number of DE genes from differential analysis on CD4 memory T cell and CD 8 memory T cell with different normalization methods.
Up refers to increased expression in CD4 memory T cell; down refers to increased expression in CD 8 Effector T cell.

the measurements biologically comparable. The biological feature set makes the method ignore the distorted data and discover the linearity among the common genes between two or more cells. In addition to the above biological explorations, we carried out two scenarios (SIM I and II) with Splatter [15] to demonstrate that resistant fit normalization works under extreme differential expression and extreme drop-out rate. The simulation results are organized in the supplementary for readers' reference. The supplementary along with the code to reproduce the experiments can be found at our GitHub repository: https://github.com/kimpenn/resistant-fit-norm-bibe-2020.

When analyzing the PBMC dataset, our results reveal that Resistant Fit Normalize has a low false-positive rate, comparable to the best of existing methods. The DE analysis for actual biological contrast, CD4, and CD8 positive T cells, resulted in enriched pathways that were more consistent with known biological differences than SCTransform and scVI.

Finally, Resistant Fit Normalization is a flexible framework that can be generalized to other forms of measurement function models, and the particular vector norm can be adapted to the characteristics of the experiment protocol and the data.

ACKNOWLEDGMENT

This work was funded in part by Health Research Formula Funds from the Commonwealth of Pennsylvania, who played no direct role in the content of the paper.

REFERENCES

[1] X. Zhang, C. Xu, and N. Yosef, "Simulating multiple faceted variability in single cell RNA sequencing," *Nature Communications*, vol. 10, pp. 1–16, June 2019. Number: 1 Publisher: Nature Publishing Group.

[2] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nature Biotechnology*, vol. 36, pp. 411–420, May 2018.

[3] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY : large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, pp. 1–5, Dec. 2018. Number: 1 Publisher: BioMed Central.

[4] A. T. L. Lun, K. Bach, and J. C. Marioni, "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts," *Genome Biology*, vol. 17, pp. 1–14, Dec. 2016. Number: 1 Publisher: BioMed Central.

[5] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature Methods*, vol. 15, pp. 1053–1058, Dec. 2018.

[6] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, "Single-cell RNA-seq denoising using a deep count autoencoder," *Nature Communications*, vol. 10, p. 390, Jan. 2019. Number: 1 Publisher: Nature Publishing Group.

[7] R. Bacher, L.-F. Chu, N. Leng, A. P. Gasch, J. A. Thomson, R. M. Stewart, M. Newton, and C. Kendziorski, "SCnorm: robust normalization of single-cell RNA-seq data," *Nature Methods*, vol. 14, pp. 584–586, June 2017. Number: 6 Publisher: Nature Publishing Group.

[8] C. Hafemeister and R. Satija, "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression," *Genome Biology*, vol. 20, p. 296, Dec. 2019.

[9] J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani, "Normalization, testing, and false discovery rate estimation for RNA-sequencing data," *Biostatistics (Oxford, England)*, vol. 13, pp. 523–538, July 2012.

[10] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biology*, vol. 11, no. 3, p. R25, 2010.

[11] P. Cizek and J. A. Visek, "Least Trimmed Squares," pp. 49–63, 2000.

[12] A. F. Siegel, "Robust regression using repeated medians," *Biometrika*, vol. 69, pp. 242–244, Apr. 1982. Publisher: Oxford Academic.

[13] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. Publisher: [International Biometric Society, Wiley].

[14] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Research*, vol. 44, pp. W90–W97, July 2016.

[15] L. Zappia, B. Phipson, and A. Oshlack, "Splatter: simulation of single-cell RNA sequencing data," *Genome Biology*, vol. 18, p. 174, Sept. 2017.