Bioinformatics Advance Access published February 24, 2015

Application Note

Normalization and noise reduction for single cell RNA-seq experiments

Bo Ding[1,#], Lina Zheng[1,#], Yun Zhu[1], Nan Li[1], Haiyang Jia[1,2], Rizi Ai[1], Andre Wildberg[1] and Wei Wang[1,3*]

[1]Department of Chemistry and Biochemistry, University of California, La Jolla, CA 92093, USA,

[2] College of Computer Science and Technology, Jilin University, Changchun 130012, China.

[3]Department of Cellular and Molecular Medicine, University of California, La Jolla, CA 92093, USA,

[#]Equal contribution

Associate Editor: Dr. Ziv Bar-Joseph

ABSTRACT

A major roadblock towards accurate interpretation of single cell RNA-seq data is large technical noise resulted from small amount of input materials. The existing methods mainly aim to find differentially expressed genes rather than directly de-noise the single cell data. We present here a powerful but simple method to remove technical noise and explicitly compute the true gene expression levels based on spike-in ERCC molecules.

Availability and implementation: The software is implemented by R and the download version is available at http://wanglab.ucsd.edu/star/GRM.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Single-cell RNA-seq is a promising technology with broad applications and discerning biological noise from technical noise is critical for correctly interpreting the data (Jaitin, et al., 2014). Recently, statistical methods are developed to model the technical noise from spike-in ERCC molecules, whose concentrations are presumably same across the samples, and then identify differentially expressed genes, whose variations across samples are significantly larger than technical noise(Brennecke, et al., 2013). A limit for such an approach is that the true gene expression level is not explicitly calculated, which is needed for many analyses based on quantification of transcriptions.

Here we propose a novel strategy to normalize and de-noise single cell RNA-seq data. This method calculates RNA concentrations from the sequencing reads, which is opposite to the other published methods that model sequencing reads from RNA concentrations; it is much simpler than the existing methods but importantly it allows to remove technical noise and explicitly compute gene expression. Specifically, we fit a gamma regression model (GRM) between the sequencing reads (RPKM, FPKM or TPM) and the concentration of spike-in ERCC molecules. The trained model is then used to estimate the de-noised molecular concentration of the genes from the reads. GRM shows great power of reducing technical noise and superior performance compared to several popular normalization methods such as FPKM(Tu, et al., 2012), TMM(Robinson and Oshlack, 2010) and FQ(Bullard, et al., 2010) in analyzing single cell RNA-seq data.

2 RESULTS

2.1 Fit a gamma regression model from read counts to RNA concentrations

Spike-in ERCCs can be added equally to each sample during the library preparation to calibrate measurements of single cell RNA-seq. A natural approach is to train a model to compute read counts such as FPKM from the concentrations of ERCC (FPKM=$function$(concentration)). This model is then used to calculate the expression level or molecular concentration of each gene from its FPKM using the reversed relationship (concentration=$function$(FPKM)). However, substantial technical noise in single cell RNA-seq makes it non-trivial to construct such a model(Grun, et al., 2014). In addition, it can be challenging to analytically or numerically solve the reverse model. We therefore propose to fit the "reverse" model directly (concentration=$function$(FPKM)) using ERCCs. This way, gene expression levels can be directly computed from FPKM. Such a strategy is novel and much simpler than the published methods that model noise distribution without explicitly computing the de-noised gene expression levels.

We choose to use gamma distribution to model the distribution of molecular concentrations because of its flexibility to fit diverse shapes. As the values of molecular concentration (10^{-2}-10^{4}) and FPKM (always 0-$10^{4\sim5}$) vary in a large range, we first perform log transformation of these data, $x = log$(FPKM) (log-R) and $y = log$(concentration) (log-C). Instead of fitting a gamma regression model between x and y directly, we model the non-linearity of single cell signals using a polynomial function $\mu(x) = \sum_{i=0}^{n} \beta_i x^i$. The model is the following:

$$y \sim Gamma(y; \mu(x), \varphi)$$

with the probability density function:

$$f(y) = \frac{1}{y\Gamma(\varphi)} \left(\frac{\varphi y}{\mu(x)}\right)^{\varphi} exp\left(-\frac{\varphi y}{\mu(x)}\right)$$

The parameters are determined using maximum likelihood estimation (MLE). The optimal value of n is determined by an empirical search: we train multiple models with n=1 to n=4 and select n with smallest average technical noise of ERCCs.

Using the regression model trained from spike-in ERCCs in one single cell sample, we compute the true expression levels of genes

*Correspondence: wei-wang@ucsd.edu

from their FPKMs by calculating the expectation of a given FPKM, $\hat{y}_{gene} = E(y_{gene}) = \hat{\mu}(x_{gene}) = \sum_{i=0}^{n} \hat{\beta}_i x_{gene}^i$.

2.2 Successful de-noise of single cell RNA-seq data

To demonstrate the effectiveness of our strategy, we need a data set that has a gold standard of cataloguing single cells and such a data set is still rare. Treutlein et al. performed single cell experiments at four different developmental stages (GEO GSE52583)(Treutlein, et al., 2014). This dataset includes 198 individual mouse lung cells derived from 4 different developmental stages: E14.5 (45 samples), E16.5 (27 samples), E18.5 (80 samples) and adult (46 samples).

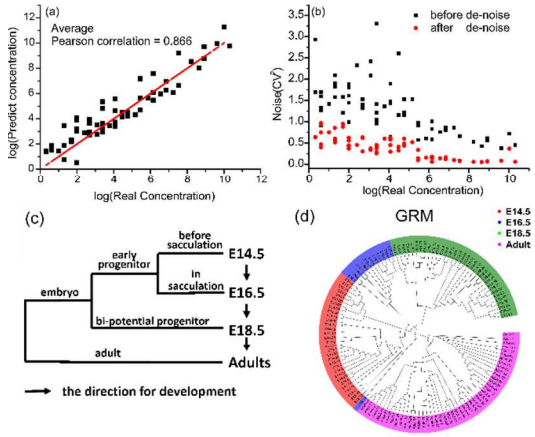

Figure 1. (a) The correlation between the true and predicted molecular concentrations for the spike-in ERCCs in one sample. (b) Noise represented by CV^2 before and after de-noise using GRM. (c) The dendrogram of the hierarchical clustering of de-noised data correctly correspond to the four developmental stages. (d) Hierarchical clustering on 124 Sftpc$^+$ single cells before and after de-noise.

We first perform gamma regression on all the spike-in ERCCs. Our model achieves an average Pearson correlation between the predicted and true concentrations of 0.866 over all the 198 samples (Figure 1(a)), significantly higher than a value of 0.119 if using a linear regression model fit between log-R and log-C. We measure the noise for the spike-in ERCCs across all the samples using CV^2, which is defined as the variance divided by the square of mean(Brennecke, et al., 2013) (Figure 1(b), Table S1). After de-noise, on average the CV^2 value is reduced 70%, from 1.301 to 0.408, with the largest reduction of 90% (from 0.584 to 0.056). These results suggest that our strategy can drastically reduce technical noise.

We next apply the model fit in each single cell to remove technical noise of the genes in the same single cell. Treutlein et al. selected 124 Sftpc-positive cells to monitor the mature process of alveolar type 2 (AT2) cells at the four developmental stages (Sftpc is the marker of AT2 cells). Four distinct groups of cells are expected to be found corresponding to the four developmental stages. Treutlein et al. selected 10,946 genes that were observed in more than two samples and had a variance of transcript level (log2(FPKM)) across all the sample larger than 0.5. If all the cells are clustered using these genes by hierarchical clustering, single cells from different stages are mixed (Figure S1(a)). After de-noise, distinct clusters corresponding to the four developmental stages are observed (Figure 1(d)). Furthermore, the dendrogram of the hierarchical clustering correctly represent the developmental distance between the single cells. Namely, the adult cells are most distant from embryonic ones of E14.5, E16.5 and E18.5. E14.5 and E16.5 are most similar to each other and form the early progenitor branch, which is connected to E18.5. This hierarchy is consistent with the development of AT2 cells (Figure 1(c)). Similarly, when PCA is applied to all the data, the de-noised data clearly show significantly better separation between the single cells derived from different developmental stages (Figure S2(a)-(e)).

Taken together, the significant lower noise in ERCCs and achievement of more biologically meaningful clusters in the AT2 cell development indicate that our de-noise strategy can successfully remove technical noise in single cell RNA-seq.

CONCLUSION

We present here a simple but powerful method for removing technical noise of single cell RNA-seq data. This method is distinct from the existing approaches as it derives the relationship between RNA concentrations and sequencing read counts from ERCC molecules and then applies this relationship to calculate gene expression from read counts. We demonstrated the success of normalization and noise reduction of single cell RNA-seq data by showing significantly improved clustering of single cells after denoise. Furthermore, this method is general and also applicable to bulk RNA-seq data with spike-in ERCCs.

ACKNOWLEDGEMENTS

This study was supported by NIH (U01 MH098977)

REFERENCES

Brennecke, P., *et al.* (2013) Accounting for technical noise in single-cell RNA-seq experiments, *Nat Methods*, **10**, 1093-1095.

Bullard, J.H., *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, *BMC Bioinformatics*, **11**, 94.

Jaitin, D.A., *et al.* (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types, *Science*, **343**, 776-779.

Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data, *Genome Biol*, **11**, R25.

Treutlein, B., *et al.* (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq, *Nature*, **509**, 371-375.

Tu, Q., *et al.* (2012) Gene structure in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis, *Genome Res*, **22**, 2079-20