Integration of Regulatory Networks and Expression Profiles (IRNEP) to Infer Master Regulators' Activities in scRNA-seq Data

Haoquan Zhao[1] and Xianzhe Peng[2]

[1]Department of Biomedical Informatics, Columbia University, New York, New York, United States

[2]Department of Computer Science, Columbia University, New York, New York, United States

Abstract

Motivation: Dropouts and mixed cell/tissue types are the main issues in the noisy scRNA-seq data analysis. Many models and algorithms were developed to reduce the noise in the gene expression level. Our approach tried to solve it from the perspective of master regulators' (MR) activities. However, the challenges lie in (1) how to identify the MRs and their targets? (2) how to quantify the MR activities?

Results: We developed a new pipeline called IRNEP to systematically infer the MR activities in scRNA-seq data. Compared to previous methods on co-expression network construction, ChIP-X and motif data were introduced for better inference of regulatory networks. We showed that our method can remove artifacts like batch effects, eliminate the negative effects of drop-outs, but keep the original cell-type heterogeneity of the data.

Availability: Publicly available on Github: https://github.com/pengxianzhe/COMS-4761-IRNEP

Contact: hz2441@columbia.edu, xp2155@columbia.edu

1 Introduction

Single cell RNA sequencing (scRNA-seq) technique allows us to measure gene expressions at single cell resolution and becomes critical in cell type identification and heterogeneity studies[1]. However, scRNA-seq data has a larger sources of noise compared to bulk RNA-seq, one of which is the dropout issue[2] that means a gene is observed at moderate expression level in one cell but is not detected in another cell at all, even if the real expression level is non-zero. Traditional statistical models for bulk RNA-seq don't address this issue well and there are many new probabilistic like zero-inflated negative binomial models[3] or generative models like variational autoencoder[4] (VAE) specifically developed for scRNA-seq without consistently good performances for different datasets. In addition to that, scRNA-seq data typically have huge number of cell types and tissue types from different time points, which makes the task even harder.

One observation is that dropout noise can be reduced if multiple gene expression levels are represented by master regulators' activities. Master regulators are key genes that are at the very top of a regulatory hierarchy and are basically not under the regulatory influence of any other genes[5]. Virtual inference of protein activity by enriched regulon analysis (VIPER) is an new algorithm for inferring master regulators' activities for bulk RNA-seq data[6]. The regulons, used by VIPER algorithm, are defined as the expression of targets of transcription factors. VIPER algorithm is a enrichment analysis based method which takes gene expression profiles and regulons as inputs and infers the activities of master regulators using normalized enrichment scores (NES). However, VIPER algorithm has certain flaws for scRNA-seq which is often not cell type specific, thus the problem of identifying good regulons is still unsolved.

In this report, we proposed a novel pipeline to infer master regulators' activity profiles from gene expression profiles for scRNA-seq tasks. Our pipeline is developed to solve three fundamental questions: 1. How to identify master regulators and regulons; 2. How to infer master regulators' activities. and 3. How to evaluate master regulator activity profiles compared to the original expression profile. Our pipeline was implemented in R and the program is publicly available on Github.

2 Methods

The IRNEP pipeline we proposed contains four sequential modules, each with unique functionalities and purposes. The first module is the data processing module whose purpose is to prepare for formatted, high-quality data for later steps. This module involves quality control, cell and gene filtering, and normalization of the gene expression data. For the raw read-count expression data, we removed cells with low read-count and extremely high read-count across all the genes according to the read count distribution, cells with few detected genes (the number of genes that have non-zero read counts) and cells with more than 10% Mitochondria genes count. In addition, genes with <~1% non-zero counts across all the cells will not be used for downstream analysis. All the cleaned datasets are normalized to $\log_2$(TPM+1) where TPM is transcripts per million. For some analysis like principle component analysis (PCA), the datasets also need to be scaled and centered according to a negative binomial model.

Master regulators and their regulons are extracted in the second module of our pipeline. In this part, gene regulatory networks are constructed based on input scRNA-seq data and there are two separate steps. First, GENIE3[7] , a tree-based method is run over the gene expression data to generate an adjacency weight matrix among genes. Higher weight corresponds to more likely regulatory links so that we applied the algorithm to obtain co-expression modules for transcription factors (TFs) by pruning the gene-gene pairs with low weight in expression profiles. Additionally, only TF modules with more than 20 target genes are kept. However, these modules are only drafts of regulons because they might contain numerous indirect targets. In the second step of this module, indirect targets are removed using RcisTarget, which is based on enriched cis-regulatory motif analysis on each of the TF regulons, and the corresponding regulons are continually pruned by CHEA dataset[8] which includes the common regulators and their corresponding targets according to multiple ChIP-X studies.

The third module infers the activity profiles of master regulators using the normalized gene expression profile and corresponding regulons generated from the second module via VIPER algorithm. In addition, another option for this module is meta-VIPER algorithm, which integrates regulons inferred from multiple gene expression profiles instead of one.

The accuracy and robustness of master regulators' activity profiles is evaluated against gene expression profiles in the last module. It is evaluated based on the results from common dimensionality reduction algorithms (PCA and t-SNE) and differential expression analysis conducted by Seurat package.

3 Main Results

IRNEP was tested on the human brain embryonic cells dataset from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76381. For comparison, t-SNE was first applied on normalized and scaled gene matrix, using the same number of variable genes as the number of MRs in the MR matrix for PCA and using the first twenty PCs to perform the t-SNE with perplexity 30 (same for other t-SNE analysis). It turned out that different brain cell-types marked with different colors were separated in the plot (Fig.1A). When we used MR matrix generated from single regulon with target pruning for t-SNE, we found that our algorithms can still capture the cell-type heterogeneity pretty well (Fig.1C). We argued that the key to achieve this is that enriched motifs and ChIP-X study evidence were applied to select targets for each TF regulons after we compared the differences between Fig.1C and Fig.1B, the t-SNE plot generated without target pruning. That's partly because if too many targets especially indirect targets are involved in the inference step, the noise as well as the cell-type signal will be greatly impaired. Interestingly, the t-SNE result with multi-regulons inference (Fig.1D) appears to do a worse job in cell-type capture compare to single regulon inference, which might due to more MRs are introduced.

Furthermore, from the elbow plot of PCA which describe the percentage of variance explained by each principle component (PC), the power of IRNEP algorithms are demonstrated clearly. Firstly, much fewer variances remain after the first ten PCs for Fig.2B, Fig.2C and Fig.2D which MR matrices were used; Secondly, compared to Fig.2A, the first two components explain most of the variances, which suggests that our method is good for low dimensional representation. To sum up, compared to using the gene matrix, using MR matrix can greatly reduce the noise. However, at the same time, it might also reduce the signal (first two PCs) and MR matrices with target pruning have a good balance between the keeping the cell-type diversity signal and reducing the unnecessary noise.

4 Conclusions

We developed a new pipeline, IRNEP to systematically infer the MR activities in scRNA-seq data. Compared to previous methods on co-expression network construction, ChIP-X and motif data were introduced for better inference of regulatory networks. We showed that our method can remove artifacts like batch effects, eliminate the negative effects of drop-outs, but keep the original cell-type heterogeneity of the data. The trade-off here is, the more we are trying to reduce the non-biological noise, the more difficult we are trying to capture the cell-type heterogeneity, and better algorithms have to be developed on precise inference of regulons.

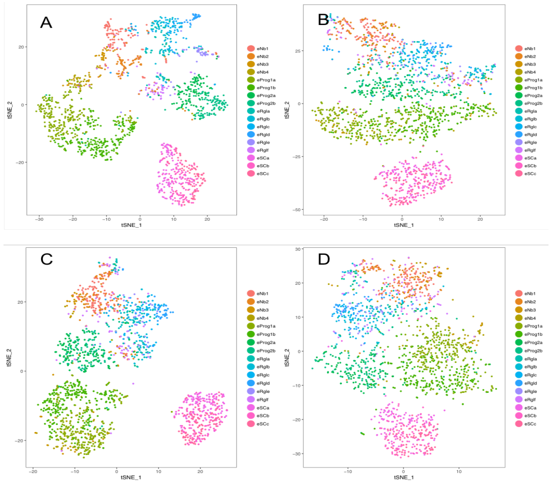

Fig. 1. t-SNE for gene matrix and three MR activity matrices.

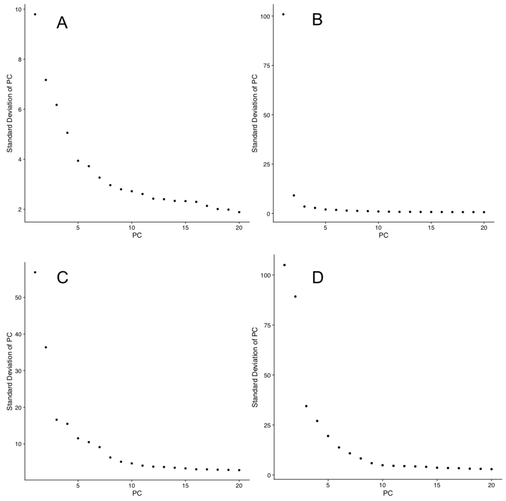

Fig. 2. Elbow plot for gene matrix and three MR activity matrices.

References

[1] Tang, Fuchou, et al. "mRNA-Seq whole-transcriptome analysis of a single cell." Nature methods 6.5 (2009): 377.

[2] Stegle, Oliver, Sarah A. Teichmann, and John C. Marioni. 2015. “Computational and Analytical Challenges in Single-Cell Transcriptomics.” Nat Rev Genet 16 (3). Springer Nature: 133–45.

[3] Risso, Davide, et al. "ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data." bioRxiv (2017): 125112.

[4] Wang, Dongfang, and Jin Gu. "VASC: dimension reduction and visualization of single cell RNA sequencing data by deep variational autoencoder." bioRxiv (2017): 199315.

[5] Ohno S, *Major sex-determining genes*. Berlin, Germany: Springer-Verlag, 1979.

[6] M. J. Alvarez, Y. Shen, F. M. Giorgi, A. Lachmann, B. B. Ding, B. H. Ye, A. Califano, “Functional characterization of somatic mutations in cancer using network-based inference of protein activity”, Nat. Genet. 48, 838–847 (2016)

[7] S. Aibar, C. Bravo Gonzalez-Blas, T. Moerman, J. Wouters, VA. Huynh-Thu, H. Imrichnova, Z. Kalender Atak, G. Hulselmans, M. Deweale, F. Rambow, P. Geurts, J. Aerts, C. Marine, J. van den Oord, S. Aerts, “SCENIC: Single-cell Regulatory Network Inference and Clustering”, Nature Methods 14, 1083–1086 (2017). doi: 10.1038/nmeth.4463

[8] Lachmann, Alexander, et al. "ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments." Bioinformatics 26.19 (2010): 2438-2444.