TENET: Gene network reconstruction using single cell transcriptomic data reveals key factors for embryonic stem cell differentiation

Anonymous Authors[1]

Abstract

We propose a novel approach called TENET (https://github.com/neocaleb/TENET) to reconstruct the gene regulatory networks (GRNs) from single cell RNA sequencing (scRNAseq) data by calculating causal relationships between genes using transfer entropy (TE). We show that known target genes have significantly higher TE values. Predicted genes with a higher TE value were more influenced by the perturbation of their regulator. Comprehensive benchmarking showed that TENET outperformed other GRN prediction algorithms. Uniquely, TENET is capable of capturing condition specific key regulators as the hubs of the GRNs including markers for pluripotency in ESCs and culture condition specific stem cell factors.

1. Introduction

Gene expression data has been widely used to infer GRNs. Recent scRNAseq data, containing the expression information of the individual cells (or status), are highly useful in blindly reconstructing regulatory mechanisms.

When dealing with causal relationships, time is often involved. To utilize time to identify the cause (the regulator) and the effect (the target genes), a series of expression data across multiple time points would be useful. scRNAseq can provide sequential expression data from the cells aligned along the pseudo-time. It is based on an assumption that the expression profile of a potential regulator is proceeded by the expression pattern of a target gene along the pseudo-time. However, systematic approaches to quantify potential causal relationships between genes and reconstruct GRNs are still highly required.

We hypothesized that we can quantify the strength of causality between genes by using an information theory, called transfer entropy (TE). TE measures the amount of directed information transfer between two variables while considering the past events by quantifying the contribution of the past events of a variable to the other variable in reducing uncertainty. By adopting TE, we developed an approach called TENET, an algorithm to reconstruct GRNs from scRNAseq data. TENET shows that target genes with higher TE values were affected more profoundly by the perturbation analysis, suggesting that TE measures the dependency of a target gene to its regulator. By benchmarking tests, we show that TENET outperforms previous GRN constructors in identifying target genes using various scRNAseq data. More importantly, unique to TENET is the ability to represent regulators of the key biological processes with the hub nodes in the reconstructed GRNs.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review at the ICML 2020 Workshop on Computational Biology (WCB). Do not distribute.

2. RESULTS

2.1. TENET quantifies causal relationships between genes from scRNAseq data aligned along the pseudo-time

TENET measures TE for all pairs of genes to reconstruct a GRN. To involve time into the gene expression, TENET aligns cells along the pseudo-time. Gene expression levels of the gene pairs along the aligned cells (Fig. 1a) are used to calculate TE between them. Given the pseudo-time ordered expression profiles, TE applied to scRNAseq quantifies the causal relationships of a gene X to a gene Y by considering the past events of the two genes. TE represents the level of the information in X that contributes to the prediction of the current event Yt (Fig. 1a). We obtained the significantly high relationships between genes after modeling all possible relationships with normal distribution (Benjamini-Hochberg's false discovery rate (FDR) ¡0.01). Potential indirect relationships were removed by applying data processing inequality (Fig. 1a).

2.2. The TF target genes showed significantly higher TE values than randomly selected genes

We applied TENET to our scRNAseq data during mESC differentiation into NPCs (**?**). We profile mESCs cultured

in 2iL (serum-free media with MEK and GSK3 inhibitors and cytokine LIF) and SL (serum media and cytokine LIF), and differentiated them to neural progenitor cells. Visualization of the scRNAseq data during mESC differentiation using tSNE showed the differentiation trajectory from naïve ground state pluripotency (2iL) to differentiation-permissive (SL) to NPCs (Fig. 1b). First, we evaluated the TE values of the target genes supported by ChIP-seq at the promoter proximal (+/-2kbps) region. We chose c-Myc (**?**) as their occupancy is often observed at the promoter region of their target genes. The TE values of the c-Myc targets were compared with those of the same number of randomly selected genes. Repeating it for 1,000 times, we observed that the 541 c-Myc target genes showed significantly higher TE values than the randomly selected genes (Fig. 1c).

2.3. TE values reflect the degree of dependency to the regulator

Gene perturbation followed by bulk expression data has been widely used to determine potential target genes. We further examined the TE values of the potential TF target genes identified by overexpression of naïve pluripotency markers including Tbx3 (**??**). We divided the genes based on their TE values and investigated the fold change upon the perturbation of the corresponding TF. As expected, we observed that the expression levels of the genes with low TE values (¡0.05) were not influenced by the perturbation. However, the expression levels of the genes with high TE values increased upon overexpression of Tbx3 (Fig. 1d). These indicate that TE values reflects the degree of dependency of the target genes to the expression of their regulator.

2.4. TENET outperforms other GRN reconstruction algorithms

To assess how TENET compares to other GRN reconstruction algorithms, we used Beeline (**?**), a benchmarking software for GRN inference algorithms. We evaluated the performance using the mESC scRNAseq dataset. The receiver operating characteristic (ROC) curves (Fig. 1e) showed that TENET, GENIE3 and LEAP outperformed other predictors in predicting targets of Nanog, Pou5f1, Esrrb, and Tbx3. SCRIBE, which was designed based on TE as well, showed worse performance than TENET.

2.5. TENET can predict key regulators from scRNAseq data

The performance evaluation by counting the number of correct or false prediction does not reflect the importance of the inferred network. It is still required to evaluate if the inferred networks reflect the key underlying biological processes. We, therefore, evaluated if the key regulators were well represented in the networks by investigating hub nodes. From the reconstructed GRNs, we further evaluated if the key regulators (based on number of outgoing edges) in the GRNs are associated with the stem cell or neural cell biology. We investigated whether the hubs in the networks are associated with "pluripotency" or "neural differentiation" using the list of the genes obtained from gene ontology (GO) database. Collectively, TENET identified far exceeding number of genes related with these key GO terms compared to other methods (Fig. 1f).

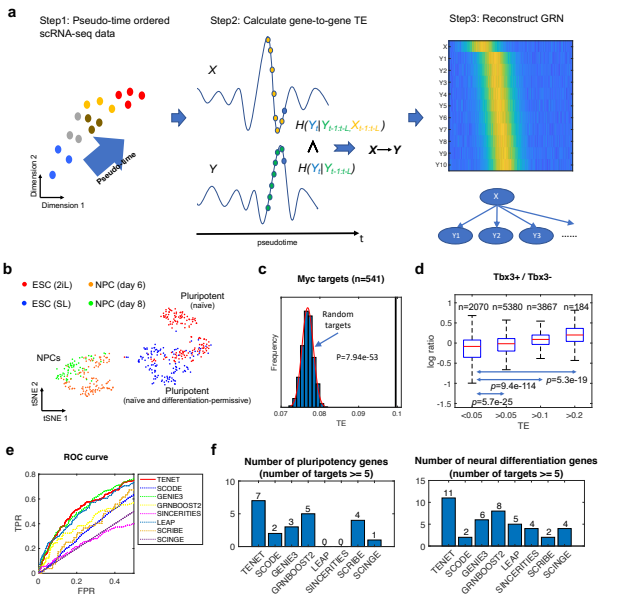

Figure 1. a. TENET reconstructs GRNs from a pseudo-time ordered single cell transcriptome data using TE. b. A tSNE plot of the mESCs (2iL and SL) and NPCs shows distinct expression. c. The c-Myc target genes have higher TE values than the randomly selected 541 genes (repeated 1000 times). d. The expression ratio of predicted Tbx3 target genes (Tbx3 overexpression (Tbx3+) against control (Tbx3-)). e. ROC curves for the mESC GRN by TENET and seven different algorithms. f. The number of hub genes (number of targets larger than 5) related with pluripotency genes and neural differentiation genes