bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TENET: Gene network reconstruction using transfer entropy reveals key regulatory factors from single cell transcriptomic data

Junil Kim[1,2], Simon Toftholm Jakobsen[3], Kedar Nath Natarajan[3,4*], and Kyoung Jae Won[1,2*]

[1]Biotech Research and Innovation Centre (BRIC), University of Copenhagen, 2200 Copenhagen N, Denmark,[2]Novo Nordisk Foundation Center for Stem Cell Biology, DanStem, Faculty of Health and Medical Sciences, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark, [3]Functional Genomics and Metabolism Unit, Department of Biochemistry and Molecular Biology, University of Southern Denmark, Denmark, [4]Danish Institute of Advanced Study (D-IAS), University of Southern Denmark, Denmark

* To whom correspondence should be addressed to KJW (Tel: +45-353-31419 ; Fax: +45-726-20285 Email: kyoung.won@bric.ku.dk). Correspondence may also be addressed to KNN (Tel: +45-655-08929; Email: knn@bmb.sdu.dk).

ABSTRACT

Accurate prediction of gene regulatory rules is important towards understanding of cellular processes. Existing computational algorithms devised for bulk transcriptomics typically require a large number of time points to infer gene regulatory networks (GRNs), are applicable for a small number of genes, and fail to detect potential causal relationships effectively. Here, we propose a novel approach 'TENET' to reconstruct GRNs from single cell RNA sequencing (scRNAseq) datasets. Employing transfer entropy (TE) to measure the amount of causal relationships between genes, TENET predicts large-scale gene regulatory cascades/relationships from scRNAseq data. TENET showed better performance than other GRN reconstructors, in identifying key regulators from public datasets. Specifically from scRNAseq, TENET identified key transcriptional factors in embryonic stem cells (ESCs) and during direct

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cardiomyocytes reprogramming, where other predictors failed. We further demonstrate that known target genes have significantly higher TE values, and TENET predicted higher TE genes were more influenced by the perturbation of their regulator. Using TENET, we identified and validated that Nme2 is a culture condition specific stem cell factor. These results indicate that TENET is uniquely capable of identifying key regulators from scRNAseq data.

Keyword: Gene regulatory network, single cell RNA sequencing, causal relationship, transfer entropy, regulator, hub, embryonic stem cells, transcription factors

Key Points

- TENET measures putative causal relationships between genes using transfer entropy.
- TENET shows outstanding performance in identifying key regulators compared to existing methods.
- TENET can reveal previously uncharacterized regulators.

INTRODUCTION

Regulatory mechanisms are key to understanding cellular processes. The cell-type specific functions and responses to external cues is governed by complex gene regulatory networks (GRNs) (Davidson and Levine 2008; Møller and Natarajan 2020; Kim et al. 2012). Various approaches including genome-wide location analysis using chromatin immunoprecipitation followed by genome-wide sequencing (ChIP-seq) (Gerstein et al. 2012; Chen et al. 2008) and perturbation analysis were designed to explain the putative causal relationships between genes (Loh et al. 2006; Hormoz et al. 2016). However, protein binding information is limited by the availability of antibodies and identification of target genes is difficult when bound at intergenic regions. Moreover, using perturbation experiments, it is hard to measure the strength of the putative causal relationships with the target genes. Systems biology approaches have been suggested to predict regulators and their target genes, prior to experimental wet-lab validation to reduce the cost and time (Hartemink 2005; Zou and Conzen 2005; Margolin et al. 2006; Møller and Natarajan 2020; Cho et al. 2007). However, previous attempts to infer GRNs have been limited to a small number of genes (Moignard et al. 2015; Li et al. 2004; Sanchez-Castillo et al. 2018) and/or cannot detect putative causal relationships effectively (Aibar et al. 2017; Chan et al. 2017).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

When dealing with causal relationships, time is often involved, i.e. an effect cannot occur before its cause. In order to utilize time to identify the cause (the regulator) and the effect (the target genes), a time series analysis of gene expression data would be useful. Single cell RNA sequencing (scRNAseq) provides sequential static snapshots of expression data from cells aligned along the virtual time also known as pseudo-time (Setty et al. 2016; Trapnell et al. 2014; Haghverdi et al. 2016). Indeed, gene expression patterns and peak levels across pseudo-time have been used to infer potential regulatory relationships between genes previously (Trapnell et al. 2014; van Dijk et al. 2018). It is based on an assumption that the expression profile of a potential regulator precedes the expression pattern of a target gene along the pseudo-time. Moreover, current approaches rely on visual and manual inspection and the gene expression dependencies are not extensively investigated. Systematic approaches that quantify potential causal relationships between genes and reconstruct GRNs are still highly required.

We aim to quantify the strength of causality between genes by using a concept originating from information theory, called transfer entropy (TE). TE measures the amount of directed information transfer between two variables (Schreiber 2000; Hlaváčková-Schindler et al. 2007). Leveraging the power to measure potential causality, TE has been successfully applied to estimating functional connectivity of neurons (Orlandi et al. 2014; Wollstadt et al. 2014; Spinney et al. 2017) and social influence in social networks (Kim et al. 2016). Based on TE, we developed TENET (https://github.com/neocaleb/TENET) to reconstruct GRNs from scRNAseq data. Using single-cell gene expression profile along the pseudo-time, TENET calculates TE values between each set of gene pairs.

We found that TE values of the known critical regulators (i.e. target genes) were significantly higher than that of randomly selected targets. Interestingly, target genes with higher TE values were influenced more profoundly by the perturbation analysis. We also show that TENET outperforms previous GRN constructors in identifying target genes.

Pseudo-time has been used in a number of GRN reconstructors (Matsumoto et al. 2017; Specht and Li 2017; Papili Gao et al. 2018; Qiu et al. 2020; Deshpande et al. 2019). Unique to TENET is the ability to represent key regulators with the hub nodes in the reconstructed GRNs. For instance, TENET identified pluripotency factors from scRNAseq during mouse embryonic stem cell (mESC) differentiation (Tuck et al. 2018) and the key programming factors from scRNAseq for the direct reprogramming toward cardiomyocytes (Liu et al. 2017b), where existing methods either failed to identify or capture their importance for the regulatory network. Interestingly, the factors that TENET identified were more negatively correlated with the number of final states (or attractors) in the Boolean

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

networks (Moignard et al. 2015), which confirms the importance of the identified hub nodes. An alternative method SCENIC also infers GRNs and their target genes using co-expression and the motif information (Aibar et al. 2017). Compared with SCENIC, TENET determines the regulatory relationships using the expression profiles alone along the pseudo-time. Therefore, TENET can be used to search for any type of regulators regardless of their binding to DNA.

Applying TENET to scRNAseq data for mESC differentiation into neural progenitor cells (NPCs), we found that Nme2 is a key transcription factor (TF) that regulates pluripotency genes in a culture condition dependent manner. Inhibition of Nme2 in mESC cultured in ground state (2-inhibitors and LIF; '**2iL**') conditions leads to dramatically reduced cell proliferation and reduced expression of its direct targets, compared to differentiation in permissive conditions (Serum and LIF; '**SL**'). In summary, TENET has a potential to elucidate previously uncharacterized regulatory mechanisms by reprocessing scRNAseq data.

METHODS

The TENET algorithms

TENET measures the amount of putative causal relationships using the scRNAseq data aligned along pseudo-time. From pseudo-time ordered scRNAseq data (Figure 1a), TENET calculates bidirectional pairwise TE values for selected genes using JAVA Information Dynamics Toolkit (JIDT) (Lizier 2014) (Figure 1b). We calculated TE values by estimating the joint probability density functions (PDFs) for mutual information (MI) using a non-linear non-parametric estimator "kernel estimator"(Schreiber 2000). The joint PDF of two genes *x* and *y* can be calculated as follows:

$$\hat{p}_r(x_n, y_n) = \frac{1}{N}\sum_{n'=1}^{N} \Theta\left(\left|\begin{pmatrix} x_n - x_{n'} \\ y_n - y_{n'} \end{pmatrix}\right| - r\right), \qquad (1)$$

where Θ is a kernel function and *N* is the number of cells. We used step kernel (Θ(*x*>0)=1, Θ(*x*≤0)=0) with kernel width *r*=0.5 as default. The TE from *X* to *Y* is defined as follows:

$$TE_{X\to Y} = H(Y_t|Y_{t-1:t-L}) - H(Y_t|Y_{t-1:t-L}, X_{t-1:t-L}), \qquad (2)$$

where *H*(*X*) is Shannon's entropy of *X* and *L* denotes the length of the past events considered for calculating TE. It calculates the amount of uncertainty of Y_t reduced by considering $X_{t\text{-}1:t\text{-}L}$. We reconstructed the GRNs by integrating all TE values for gene pairs (Figure 1c). To remove potential indirect relationships, we applied the data processing inequality (Margolin et al. 2006), i.e. iteratively eliminating feed-forward loops. The feed-forward loop is defined by a network motif composed of three genes, where gene X regulates gene Y and both gene X and Y regulate gene Z. We trimmed the

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

link from gene X to gene Z if $TE_{X \to Z}$ is less than the minimum value of $TE_{X \to Y}$ and $TE_{Y \to Z}$. Finally, we reconstructed a GRN consisting of the significant links using Benjamini-Hochberg's false discovery rate (FDR) (Benjamini and Hochberg 1995) after performing the one-sided z-test while considering the all trimmed TE values as a normal distribution. The hub node is identified by calculating the number of targets (outgoing links).

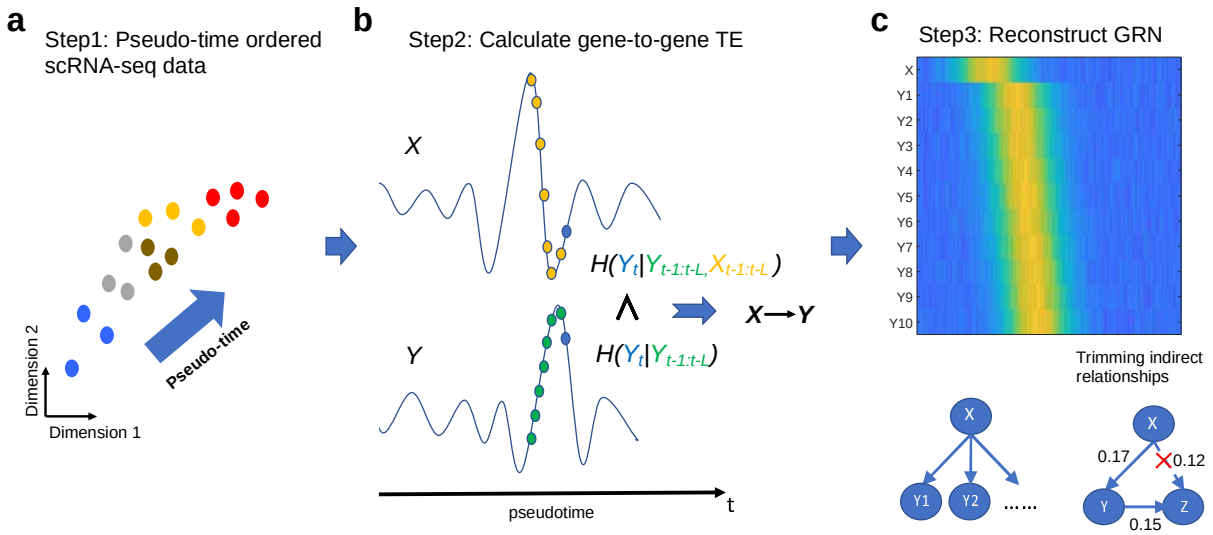


Figure 1. **TENET reconstructs GRNs from pseudo-time ordered single cell transcriptome data using TE**. **a**. Step 1: Pseudo-time ordered scRNAseq data are used as the input for TENET. **b**. Step 2: TENET calculates gene-to-gene pairwise TE while considering the past events of X and Y. **c**. Step 3: A reconstructed GRN is composed of putative but significant causal relationships followed by trimming indirect relationships. The heatmap shows the gene expression levels for a regulator (X) and its target genes.

Statistical analysis

A two-sided one-sample z-test was performed to evaluate the mean of TE values for the targets of key factors (c-Myc, n-Myc, E2f1, Zfx, Nme2) in mESCs and Gata4 in mouse cardiomyocytes. This was accomplished by generating a fitted z-distribution of the TE values using the same number of randomly selected genes (1,000 times). A two-sided two-sample Student t-test was performed to evaluate the relative gene expression changes after knocking-in of Tbx3 and Esrrb and knocking-down of Pou5f1 and Nanog for the specified TE values, respectively.

Data processing of scRNAseq data

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To test TENET, we used the scRNAseq dataset obtained from mESCs (Tuck et al. 2018) and mouse cardiomyocytes (Liu et al. 2017b). Wishbone (Setty et al. 2016) was used for pseudo-time analysis. As an input gene list for the benchmarking of mESC dataset, we used 3,277 highly variable genes with log2(count)>1 for more than 10% of all cells and a coefficient of variation > 1.5. For the scRNAseq data during the reprogramming into cardiomyocytes, we used 8,640 highly variable genes with log2(count)>1 for more than 10% of all cells and a coefficient of variation >1. To reconstruct the GRN, we used a regulator gene list which includes genes with a GO term "regulation of transcription (GO:0006355)" for the mESC. We generated all the network figures using Cytoscape 3.6.1 (Shannon et al. 2003).

Gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for the functional gene group

All enriched GO terms and KEGG pathways were obtained using Enrichr (Kuleshov et al. 2016). The "pluripotency gene" and the "neural differentiation gene" were obtained from the genes with a GO term "stem cell population maintenance (GO:0019827)" and "neuron differentiation (GO:0030182)", respectively. We used GO terms "cardiac muscle cell differentiation (GO:0055007)", "cardiac muscle contraction (GO:0060048)" for cardiomyocyte gene.

Gene expression and ChIP-seq data for validation

We downloaded an RNAseq dataset in mESCs with three different combinations of double knock-in for Esrrb and Tbx3 (Esrrb-/Tbx3-, Esrrb+/Tbx3-, Esrrb+/Tbx3+) (Hormoz et al. 2016). The gold standard target genes of Esrrb and Tbx3 was obtained by comparing Esrrb-/Tbx3- versus Esrrb+/Tbx3- samples and Esrrb+/Tbx3- versus Esrrb+/Tbx3+ samples with 2-fold change criterion, respectively. The target genes of Nanog and Pou5f1 were identified by using microarray data in mESC with Nanog and Pou5f1 knockdown (Loh et al. 2006). To identify target genes of these two TFs, we used a 2-fold change and a p-value < 0.01 as described in the original data analysis.

ChIP-seq data for Pou5f1, Esrrb, Nanog in mESCs were reanalyzed for peak calling (Chen et al. 2008). After removing the adapter sequence using CutAdapt (Martin 2011) implemented in TrimGalore-0.4.5, we aligned the ChIP-seq reads to the mm10 genome using Bowtie2 (Langmead et al. 2013). ChIP-seq peak was called against GFP control using the "findPeaks" command in the Homer package (Heinz et al. 2010).

Robustness of the performance of TENET

In order to evaluate the robustness of TENET, we ran the Wishbone 57 times with different options on the Boolean expression data of single-cells obtained from early blood development experiments (Moignard et al. 2015). 57 Wishbone trajectories were obtained by running Wishbone with 19 different

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

initial states provided in the reference paper (Moignard et al. 2015) and three different choices of cells based on the branches (total cells, trunk + first branch, trunk + second branch). The other GRN reconstructors from Beeline were also run based on these 57 Wishbone trajectories.

Condition specific targets

To identify condition specific targets, we reconstructed GRNs using the pseudo-time ordered expression data of 2iL+NPCs and SL+NPCs using TENET. Subsequently, the condition specific targets of the top 20 factors in the common GRN were obtained by selecting targets in the culture condition specific GRNs. For example, the 35 target genes of Nme2 were included in the 2iL-specific but not in the SL-specific GRN whereas the 14 target genes were included in the SL-specific but not in the 2iL-specific GRN.

ESC culture

E14 mESC were cultured on plastic plates coated with 0.1% gelatin (Sigma #G1393) in either DMEM knockout (Gibco #10829), 15% FBS (Gibco #10270), 1xPen-Strep-Glutamine (Gibco #10378), 1xMEM (Gibco #11140), 1xB-ME (Gibco #21985) and 1000U/mL LIF (Merck #ESG1107) ("Serum") or in NDiff 227 (Takara #Y40002), 3uM CHIR99021 (Tocris #4423), 1uM PD0325901 (Tocris #4192) and 1000U/mL ("2iL"). For Nme2 experiments, mESCs were treated with either vehicle DMSO (Sigma #02660) or 0.5uM STP (StemCell technologies #72652) for 24 or 48hr.

Alkaline phosphatase staining

For AP staining, 1000 mESCs were seeded in a 12-well plate and cultured for 24 or 48hr. The cells were washed in PBS, fixed in 1% formaldehyde and stained with AP following manufacturers instruction (Merck #SCR004). For quantification of positive stained colonies, four randomly selected areas of each well were imaged (10x magnification; Nikon Eclipse TS2) and manually counted. Colonies were marked as pluripotent or primed based on morphology and intensity of AP staining. The mean and standard error of mean (SEM) were calculated over four independent replicates.

Cell proliferation assay

For proliferation assay, 100,000 mESCs were seeded in a 6-well plate in both 2iL and SL conditions. Cells were initially allowed to attach for 24hrs before treatment with either DMSO or STP. After either 24hr or 48hr of DMSO or STP treatment, cells were detached from the plate using Accutase and counted using the TC-20 automated cell counter (BioRad). Data are mean + SEM from 4 biological replicates.

RNA extraction and qPCR

Total RNA was harvested using Trizol (Ambion #15596026), lock-gel columns (5prime #733-2478) and precipitated in chloroform/isopropanol using with glycogen. Reverse transcription was performed

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with 1ug of RNA using high capacity cDNA kit (Applied Biosystem #4368814). Quantitative-PCR was performed using SYBR-green with LightCycler480. To obtain relative gene expression levels, expression levels were normalized to Gapdh as a control.

RESULTS

TENET quantifies the strength of putative 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 assign time to the cells, TENET aligns cells along the pseudo-time. The paired gene expression levels along the pseudo-time are used to calculate TE (Figure 1a). Given the pseudo-time ordered expression profiles (Figure 1a), TE quantifies the strength of putative causal relationships of a gene *X* to a gene *Y* (Figure 1b) by considering the past events of the two genes. TE represents the level of information in gene *X* that contributes to the prediction of the current event Y_t. The highly significant relationships between genes are obtained by modeling all possible relationships with normal distribution (Benjamini-Hochberg's FDR (Benjamini and Hochberg 1995)<0.01). The potential indirect relationships are removed by applying data processing inequality measure (Margolin et al. 2006) (Figure 1c) (see Methods). TENET can be run on various sets of cell type regulators including either known set of genes or the set of all TFs, or even on the entire set of genes depending on the network of interest. After feature selection, the network analysis is applied to understand key regulators and relationships within the networks. In sum, TENET is useful in identifying target genes of a regulator and predicting key regulators.

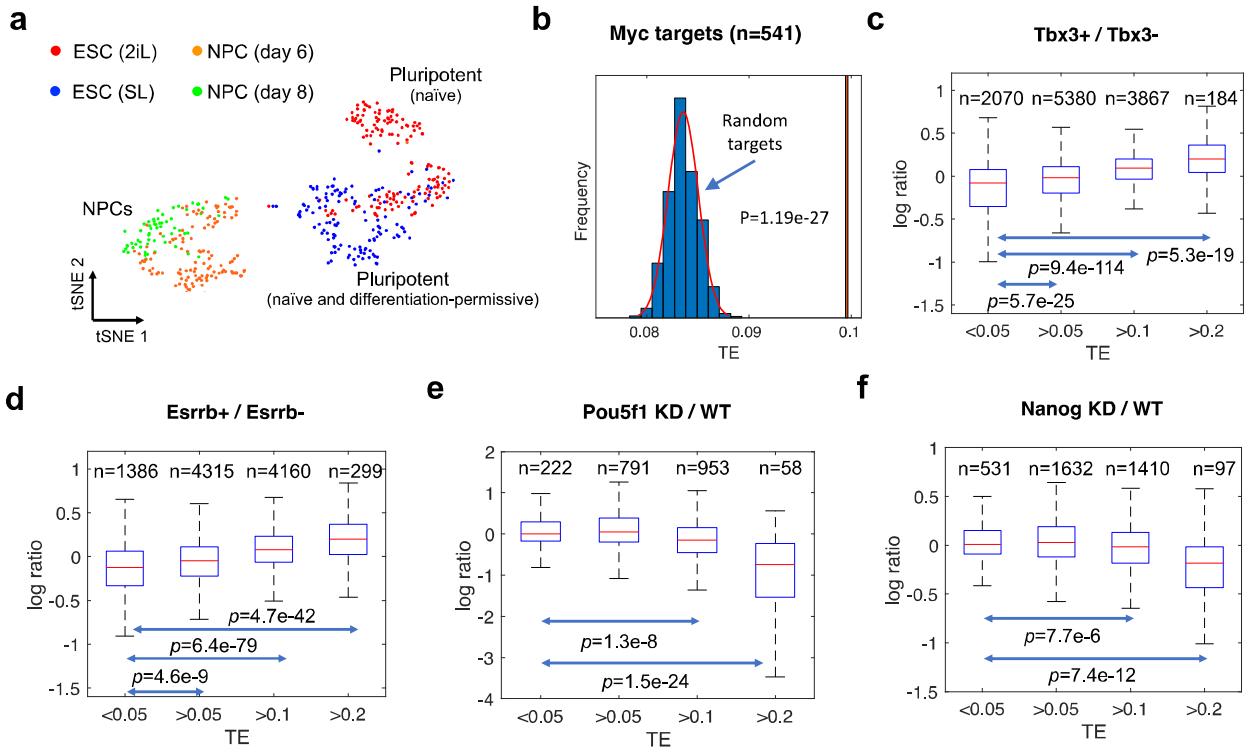
The TF target genes showed significantly higher TE values than randomly selected genes

We applied TENET to the scRNAseq data during mESC differentiation into NPCs (Tuck et al. 2018). We profiled mESCs cultured in 2iL (serum-free media with MEK and GSK3 inhibitors and cytokine LIF) and SL (serum media and cytokine LIF), and induced differentiation into neural progenitor cells (Bibel et al. 2007). The 2iL cultured mESCs (termed 'ground state') homogeneously express naïve pluripotency markers mimicking mouse epiblast, while SL cultures contain a heterogeneous mix of undifferentiated and differentiating ESCs (Alexandrova et al. 2016; Kalkan et al. 2017). Another motivation for choosing the mESC experimental system was that a number of ChIP-seq and RNAseq datasets are publicly available for validation (Chen et al. 2008; Loh et al. 2006; Hormoz et al. 2016). Visualization of the scRNAseq data during mESC differentiation using tSNE showed the differentiation trajectory from naïve ground state pluripotency (2iL) to differentiation-permissive (SL)

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to NPCs (Figure 2a). Consistent with the differentiation time course, general and naïve pluripotency markers including Pou5f1 (or Oct4), Sox2 and Nanog were highly expressed in the mESC population whereas NPC markers such as Pax6 and Slc1a3 were highly expressed in the NPCs (Figure S1). Then, we evaluated the TE values of the target genes supported by ChIP-seq at the promoter proximal (+/-2kbps) region. We chose c-Myc, n-Myc, E2f1 and Zfx (Chen et al. 2008) as their occupancy is often observed at the promoter region of their target genes. Applying peak calling using Homer (Heinz et al. 2010), we found 541 c-Myc promoter proximal peaks. The TE values of the c-Myc targets were compared to the randomly selected genes (as control) with the same sample size, similar GC contents and expression levels. Repeating the process 1,000 times, we observed that the 541 c-Myc target genes showed significantly higher TE values (p-value=1.19e-27) than the randomly selected genes (Figure 2b). We also confirmed that ChIP-seq binding targets for other promoter binding TFs such as n-Myc, E2f1 and Zfx also have significantly higher TE values compared with the random targets (Figure S2a-c).

Additionally, we performed evaluation of TE values using the scRNAseq dataset for the reprogramming of mouse fibroblasts into induced cardiomyocytes (Liu et al. 2017b). Investigation using Gata4 ChIP-seq in cardiomyocytes (Luna-Zurita et al. 2016) confirmed that the 331 potential target genes with Gata4 promoter occupancy also possess significantly higher TE values compared with random targets (Figure S2d).



bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2. **Validation of TENET-inferred GRNs for the mouse embryonic stem cell (mESC) pluripotency**. **a**. A tSNE plot of the mESCs (2iL and SL) and NPCs shows distinct expression. **b**. The c-Myc target genes have higher TE values than the randomly selected 541 genes (repeated 1000 times). The expression ratio of predicted Tbx3 (**c**) or Esrrb (**d**) target genes (Tbx3 or Esrrb overexpression (Tbx3+ or Esrrb+) against control (Tbx3- or Esrrb-)). The expression ratio of predicted Pou5f1 (**e**) or Nanog (**f**) target genes (knockdown versus wild-type).

TE values reflect the degree of dependency to the regulator

Gene perturbation followed by gene expression measurement by bulk RNAseq 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 Esrrb and Tbx3 as well as knockdown of Pou5f1 and Nanog (Loh et al. 2006; Hormoz et al. 2016). We divided the genes based on their TE values and investigated the fold changes upon the perturbation of the corresponding TF. Interestingly, the expression levels of the genes with low TE values (<0.05) had little or no influence upon perturbation. However, the expression levels of the genes with high TE values were markedly increased upon overexpression of Esrrb and Tbx3 and consistently, decreased upon knockdown of Pou5f1 and Nanog. These changes were particularly more significant for genes with higher TE values (>0.2) (Figure 2c-f). These results indicate that TE values reflects the degree of dependency of the target genes to the expression of their regulator.

TENET can predict key regulators from scRNAseq data

To determine whether the TENET captures the key biological processes, we investigated hub nodes and evaluated if key regulators were well represented. From the reconstructed GRNs from mESC to neural cells, we assessed if key regulators (based on the number of outgoing edges) in the GRNs are associated with stem cell or neural cell biology. The gene ontology (GO) terms and KEGG pathways enrichment tests showed that the hub regulators (number of outgoing edges >= 5) are mostly associated with pluripotent stem cells and cellular differentiation functions (Figure S3a). We also benchmarked and compared TENET's performance to other methods including SCODE (Matsumoto et al. 2017), GENIE3 (Huynh-Thu et al. 2010), GRNBOOST2 (Moerman et al. 2019), SINCERITIES (Papili Gao et al. 2018), LEAP (Specht and Li 2017), SCRIBE (Qiu et al. 2020), and SCINGE (Deshpande et al. 2019). For unbiased comparison, we ran each GRN method on the same set of 3,277 highly variable genes (see Methods).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The top 4 ranked regulators determined by TENET were markers for pluripotency (Pou5f1, Nanog, Esrrb, and Tbx3) (Figure 3a). Compared to TENET, most methods failed to identify these key genes in the hub list except for SCRIBE. For instance, GENIE3 and GRNBOOST2 only found Nanog as the 14th and 5th of the top regulators, respectively; but they did not detect Pou5f1. SCRIBE, another TE-based GRN predictor identified Nanog, Pou5f1, Esrrb, Tbx3 as the top regulators, suggesting the algorithmic advantages of TE especially for scRNAseq data (Figure S4). However, both SCRIBE and SCODE the numbers of target genes in the hub node were drastically reduced beyond the 5th regulator, which highlights that these methods emphasize on a few potential regulators during network reconstruction.

Intrigued by this, we investigated whether the hubs in the networks are associated with "pluripotency" or "neural differentiation" using the list of the genes obtained from GO database (see Methods). We investigated both receiver operating characteristic (ROC) curves and precision-recall curves (PRCs) while regarding genes in the GO database as true. The ROC curves for the pluripotency and neural differentiation demonstrates TENET's far exceeding capability in predicting key regulatory factors related with these key GO terms compared to other methods (Figure 3b and S5a-b). The area under precision-recall curve (AUPRC) further confirmed the increased performance of TENET in capturing key regulators (Figure 3c). As TE values rely on pseudo-time, we also investigated whether TENET results were sensitive to other pseudo-time inference methods. Computing pseudo-time using PAGA (Wolf et al. 2019) and Slingshot (Street et al. 2018) showed that TENET is robust to the choice of pseudo-time inference and outperformed other GRN reconstructors (Figure S6).

To further test if TENET can suggest key regulatory factors in various biological systems, we reconstructed a GRN based on the scRNAseq data for direct reprogramming of mouse fibroblast into cardiomyocyte by overexpressing Mef2c, Tbx5 and Gata4 (Liu et al. 2017b). We first examined if these overexpressed factors were well predicted in the inferred GRNs. Consistently, TENET identified those three major reprogramming factors (Mef2c, Tbx5 and Gata4) as well as other genes associated with cardiomyocytes as top ranked regulators (Figure 3d and Figure S3b). Not surprisingly, these factors were not well observed in the GRNs inferred by other reconstruction methods (Figure 3e-f and S5c-d) with exception of SCRIBE that only found Mef2c (Figure S7). Additionally, while GRNBOOST2 showed relatively better performance in detecting pluripotency and neuronal differentiation factors, it failed in detecting the key factors during cardiomyocyte reprogramming. Collectively, our results show that TENET can robustly capture key regulatory genes for biological processes.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

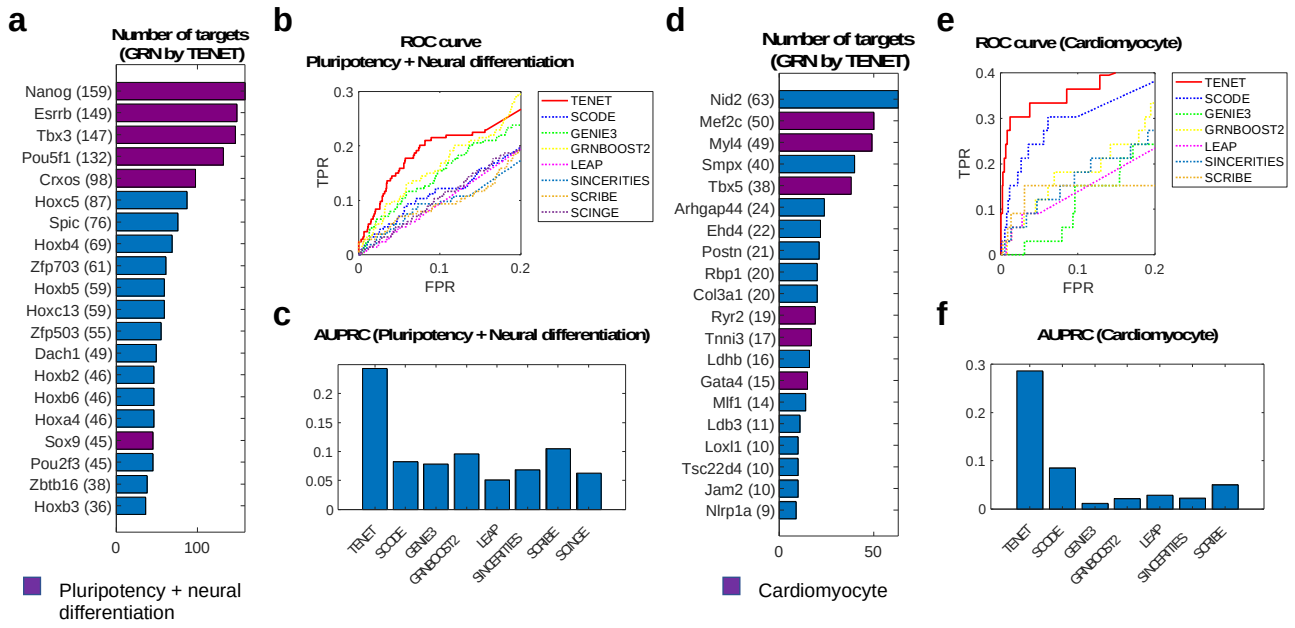


Figure 3. **TENET outperformed other tools when predicting key regulatory factors for mESC pluripotency and direct reprogramming from mouse fibroblast into cardiomyocyte**. **a**. Key regulatory factors for mESC pluripotency and neural differentiation predicted by TENET. The purple bar denotes pluripotency and neural differentiation genes. **b**. ROC curves and **c**. Area Under Precision-Recall Curves (AUPRCs) for the prediction of key regulatory factors of pluripotency and neural differentiation. **d**. Key regulatory factors for direct reprogramming into cardiomyocyte in the TENET-inferred GRN. Three major reprogramming factors Mef2c, Tbx5, Gata4 have a large number of targets. **e**. ROC curves and **f**. AUPRCs for the prediction of key regulatory factors of cardiomyocyte.

TENET's hub nodes were associated with the controllability of Boolean network dynamics

To further investigate the characteristics of TENET in finding key regulators, we compared the reconstructed networks with Boolean networks (BNs) (Moignard et al. 2015). BNs consider all possible binary status of its members (genes) and have been widely used to model biological systems (Li et al. 2004; Choi et al. 2012; Wang et al. 2010). BNs can simulate overexpression or knock-out of a gene and its consequences from the inferred networks. Therefore, BNs can be used to evaluate how much a member (i.e. gene) can influence the steady-state dynamics of the networks, and in combination with other members (called "controllability") (Kim et al. 2013). Previously, a BN based GRN using 20 TFs was built using single cell qRT-PCR during mouse early blood development (Moignard et al. 2015) (Figure S8). Using the BN-inferred GRN as a surrogate for the gold standard, we first evaluate if the networks from GRN reconstructors accurately mimic the BN-inferred GRN. The comparison showed

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

that TENET and GRNBOOST2 outperforms other approaches in both directed and undirected networks in this example (Figure S9a-b).

In the BNs, the number of final stable states (known as attractors) can be calculated while simulating all possible states of the members except one member of interest (a gene with perturbation). A critical member usually has a small number of attractors. Therefore, the predicted hub genes in the GRN will negatively correlate with the number of attractors if the hub genes are the key genes. In a series of experiments, TENET showed an ability to find key regulators. We further tested if the predicted key regulators are negatively correlated with the attractors found in the BNs. Our simulation showed that the TENET-inferred network has the strongest negative correlation with the number of attractors compared followed by SCRIBE, SCODE and GRNBOOST2 (Figure S9c), while other methods showed either no or positive correlation. This further demonstrated that TENET has the capability to identify key regulators.

TENET outperforms other GRN reconstruction algorithms in identifying target genes

To further assess TENET, we used Beeline (Pratapa et al. 2020), a benchmarking software for GRN inference algorithms. Among them, we performed benchmarking only for those algorithms that can implement large scale GRN reconstruction including SCODE (Matsumoto et al. 2017), GENIE3 (Huynh-Thu et al. 2010), GRNBOOST2 (Moerman et al. 2019), SINCERITIES (Papili Gao et al. 2018), LEAP (Specht and Li 2017), SCRIBE (Qiu et al. 2020), and SCINGE (Deshpande et al. 2019), using the mESC scRNAseq dataset (Tuck et al. 2018). To prepare stringent datasets for evaluation, we regarded a target as true if the expression of the target gene is changed significantly by the perturbation study (Loh et al. 2006; Hormoz et al. 2016) and the binding occupancy of the corresponding regulator is observed nearby (+/-50kbps to transcription start sites (TSSs)) (Chen et al. 2008) (see Figure S10a and Methods).

We benchmarked all methods running Beeline on 3,277 highly variable genes (see Methods). Beeline (Pratapa et al. 2020) provided comprehensive results after running all GRN reconstructors. The ROC curves showed that TENET, GENIE3 and LEAP outperformed other reconstructors in predicting targets of Nanog, Pou5f1, Esrrb, and Tbx3 (Figure S10b-c). Interestingly, SCRIBE showed worse performance than TENET, while GENIE3 and LEAP failed to find key regulatory genes but showed good performance in this benchmarking test (even with a small number of regulators).

TENET identifies culture condition specific regulators

To search for potential regulators besides the known TFs during stem cell differentiation (Tuck et al. 2018), we extended the GRN by considering 13,694 highly variable genes as well as target genes (see Methods). In addition to several known pluripotency (Nanog, Sox2, Pou5f1, Tfcp2l1 etc.,) and

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

neural regulators (Meis1, Tbx3 etc.,), we were intrigued to find Fgf4 and Nme2 as the top regulators (ranked by number of targets, Figure S11b and Table S1). Fgf4 is known to be dispensable for embryonic stem cells, but is critical for exit from self-renewal and differentiation (Almousailleakh et al. 2007; Lanner and Rossant 2010), while Nme2 (Zhu et al. 2009) has been implicated in stem cell pluripotency (Figure S11).

As we profiled mESCs in both ground-state 2iL and heterogeneous SL conditions, we assessed whether TENET could further distinguish them and identify culture-condition specific GRNs. We reconstructed GRNs for 2iL and SL condition separately and compared the regulators as well as their specific targets (Figure S12, see Methods). We found several naïve pluripotency markers specifically enriched in 2iL condition including Nanog (Wray et al. 2010), Esrrb (Martello et al. 2012), and Tfcp2l1 (Martello et al. 2013; Qiu et al. 2015), whereas heterogeneous and hypermethylated SL condition (Habibi et al. 2013; Leitch et al. 2013) regulators included Tet1 (Pantier et al. 2019; Ito et al. 2010) and Dnmt3l (Ficz et al. 2013) and Zfp57 (Riso et al. 2016) (Figure S12). Interestingly, Nme2 was a top mESC regulator for the 2iL GRNs. Intrigued by this, we sought to investigate the effect of Nme2 perturbation using an small molecule inhibitor Stauprimide (STP) that blocks the nuclear localization (Zhu et al. 2009). We cultured mESCs in 2iL and SL conditions and treated cells with 0.5μM STP (Figure 4a). The cellular proliferation and division were significantly inhibited in both culture conditions, but were drastic in 2iL (Figure 4b). Upon 24hr STP treatment in 2iL, we could visually observe high levels of apoptotic cells, detached colonies and few viable cells at 48hrs. We quantified the STP effect on pluripotency by alkaline phosphatase staining (AP; Figure 4c) and scored cells either as *undifferentiated* (high AP staining, rounded colonies; naïve mESCs) or as *mixed* (low/no AP staining, flattened colonies; differentiation-like/apoptotic cells) (Kalkan et al. 2017; Liu et al. 2017a). The STP treatment in 2iL led to a drastic decrease in undifferentiated colonies (45% ± 5.9% colonies) and the remaining mixed cells were mostly composed of apoptotic cells.

Previously, c-Myc has been reported as a target gene of Nme2 (Zhu et al. 2009), consistent with TENET prediction. We confirmed that c-Myc expression was significantly downregulated upon STP treatment in both culture conditions (Figure 4d). TENET also predicted several TFs including Nanog and Ctnnb1 at the target of Nme2. We found that both Nanog and Ctnnb1 transcripts were highly upregulated upon STP treatment in both culture conditions but more significant in 2iL condition, indicating condition specific regulation of Nme2 as predicted by TENET (Figure 4d).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

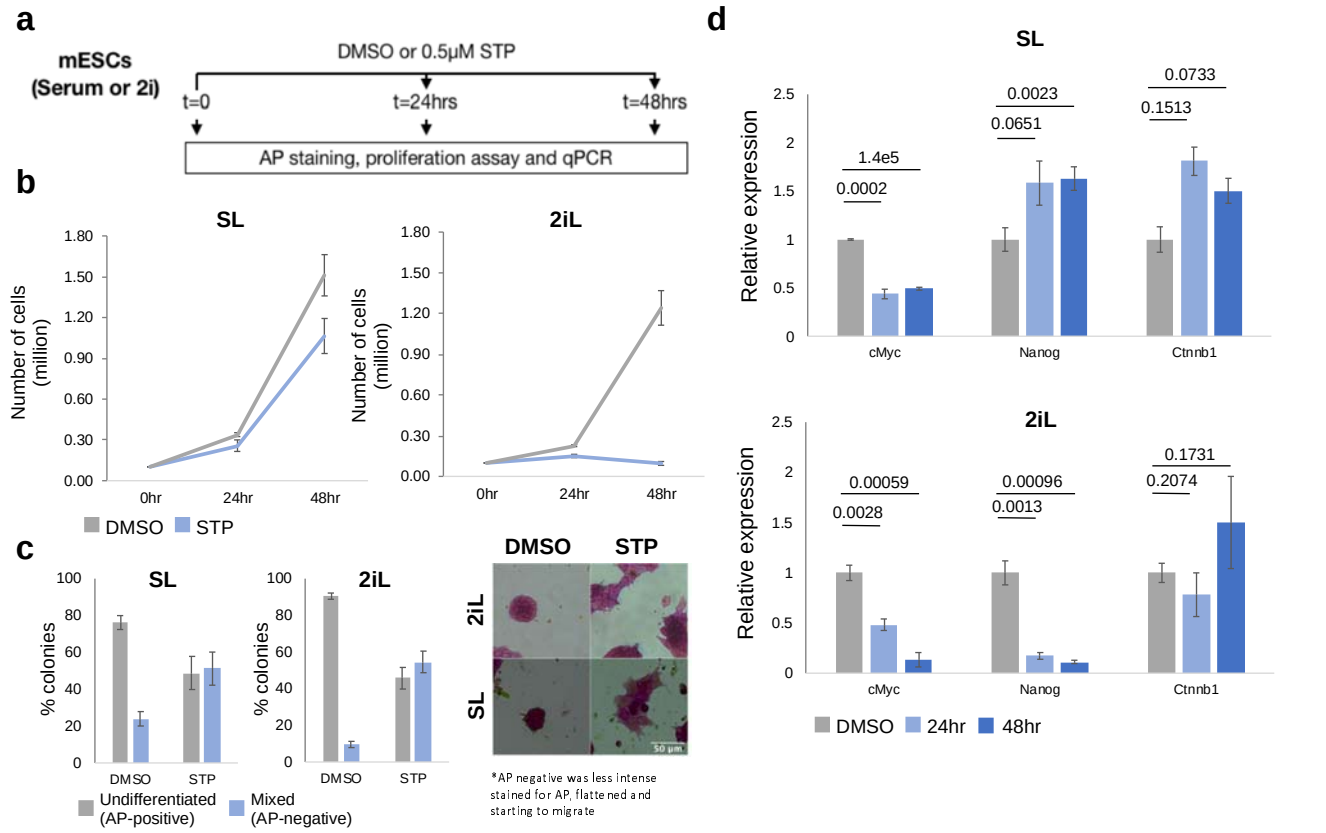


Figure 4. Nme2 inhibition blocks proliferation of mESC in 2iL condition. **a**. Experimental design: The mESCs were seeded in either SL or 2iL culture conditions and were treated with either DMSO (control) or 0.5μM STP for 24hr and 48hr. The 6 samples were assayed for proliferation rates, relative transcript expression and for pluripotency using alkaline phosphatase (AP). **b**. Cell proliferation assay for mESCs cultured in SL and 2iL conditions with either DMSO (Control) or 0.5μM STP. The mean and SEM were calculated over four independent replicates. **c**. STP treatment leads reduction in undifferentiated colonies and increase in mixed differentiation-like colonies across both SL (48hrs) and 2iL (24hrs) conditions, based on AP staining. AP staining intensity and colony morphology are used to classify undifferentiated (high AP staining, rounded colonies) and mixed (low/no AP staining, flattened colonies; differentiation-like; apoptotic cells) populations. Representative images of undifferentiated and mixed colonies in control and STP treated colonies across both culture conditions. The data in the barplots describe the mean ± SEM from 2 biologically independent replicates. **d**. The c-Myc transcript levels are down regulated both in 2iL and SL upon STP treatment, owing to impaired Nme2 nuclear localization. The Nme2 target genes in TENET (Nanog and Ctnnb1) are selectively regulated between culture conditions. Significance (p-value) are highlighted above barplots. The data in barplots describe the mean ± SEM from 3 biologically independent replicates.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DISCUSSION

Systems biology approaches to infer GRNs can provide a hypothesis for further experimental validation. Existing methods for bulk transcriptomics datasets are limited because they cannot capture the continuous cellular dynamics and/or require cell synchronization to avoid "average out" expression. scRNAseq has emerged as an alternative because of its power to provide the transcriptomic snapshots of hundreds, thousands of cells on a massive scale, from same population. Subsequently, computational approaches used scRNAseq for GRN reconstruction (Moignard et al. 2015; Sanchez-Castillo et al. 2018; Matsumoto et al. 2017; Moerman et al. 2019; Papili Gao et al. 2018; Specht and Li 2017; Deshpande et al. 2019; Qiu et al. 2020; Møller and Natarajan 2020; Aibar et al. 2017).

Many GRN reconstruction algorithms including TENET use the temporal gene expression changes, after ordering cells across pseudo-time. For example, GENIE3 (Huynh-Thu et al. 2010) and GRNBOOST2 (Moerman et al. 2019) were originally applied the ensembles of regression trees to temporal bulk expression data. LEAP (Specht and Li 2017) calculates possible maximum time-lagged correlations. SINCERITIES (Papili Gao et al. 2018) and SCINGE (Deshpande et al. 2019) used Granger causality from pseudo-time ordered data. SCODE (Matsumoto et al. 2017) uses a mechanistical model of ordinary differential equations on the pseudo-time aligned scRNAseq data. Compared with current methods, TENET makes use of the power of information theory by adopting TE on gene expression along the pseudo-time. Therefore, the performance of these predicted regulators could be dependent on the performance of the pseudo-time inference. However, we found that TENET is robust to the multiple pseudo-time inference approaches in comparison with other GRN reconstructors (Figure S6).

We showed that TE values of the known target genes were significantly higher than randomly selected genes (Figure 2b and S2). The target genes with higher TE values were more significantly perturbed by either overexpression or knockdown of the corresponding regulators (Figure 2c-f). We also performed comprehensive benchmarking of TENET and several GRN reconstructors using Beeline (Pratapa et al. 2020) and its automated pipeline. TENET was consistently one of the top performing GRN reconstructors in these tests.

The evaluation of the performances of GRN reconstructors by counting the number of true or false prediction does not fully reflect the importance of the inferred network. We observe that TENET consistently predicts and identifies key regulators. This is important because upstream regulators for a biological process are often of interest to explain the underlying mechanisms. It is still required to evaluate if the inferred networks reflect the key underlying biological processes. Applying TENET to a

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

series of scRNAseq datasets including 1) mESC differentiation and 2) reprogramming to cardiomyocytes, we find that TENET identified key factors as the top scoring hubs. For mESC differentiation, TENET ranked Nanog, Pou5f1, Esrrb and Tbx3 as the top 4 regulators, while existing methods failed to identify these key factors. In an additional test using GO terms, TENET identified gene relationships associated with pluripotency and neural differentiation (Figure 3b-c). Interestingly, existing methods including LEAP and SINCERITIES did not find any genes related to pluripotency in their networks (Figure S5b). Analyzing the reprogramming to cardiomyocytes scRNAseq data, only TENET identified the reprogramming factors (Mef2c, Tbx5 and Gata4) (Liu et al. 2017b) (Figure 3d-f and Figure S7). These results suggest that while other approaches successful in finding some regulatory rules, they cannot make networks focusing on the key biological process.

We further questioned if TENET is capable of identifying key regulators using BNs. While BNs may not be a perfect model of biological system, they can still provide a comprehensive systematic overview by visiting all potential states. In BN, the key nodes usually have small number of attractors as they drive the networks into more determined status. In our analysis using BNs, TENET-inferred networks were negatively correlated with the number of attractors (Figure S9), indicating the key ability to capture biological processes.

A number of studies showed distinct expression patterns in the pseudo-space (Halpern et al. 2018; Nowotschin et al. 2019). Since pseudo-time inference can lead to multiple branched trajectories, we also applied TENET to individual branches. These expression changes for some genes may be attributed to association along the spatial axis. However, the associating potential causal relationships for them may not be relevant.

With the power to predict key regulators, we applied TENET to identify mESC culture-condition specific regulators. TENET predicted several TFs (Nanog, Esrrb, and Nme2) as specific for 2iL compared to SL culture conditions (Figure S12). Although Nme2 is expressed both in 2iL and SL, perturbing Nme2 leads to more dramatic effects (reduced proliferation, AP staining and apoptosis) in the 2iL condition; consistent with our prediction. In sum, TENET is a useful approach to predict previously uncharacterized regulatory mechanisms from scRNAseq.

DATA AVAILABILITY

A source code for TENET and input files for the benchmarking datasets are available at https://github.com/neocaleb/TENET.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

FUNDING

The Novo Nordisk Foundation Center for Stem Cell Biology is supported by a Novo Nordisk Foundation grant No. NNF17CC0027852. KJW is supported by the Lundbeck Ascending Investigator (R313-2019-421), the Independent Research Fund Denmark (0135-00243B) and a National Institute of Health (National Institute of Diabetes and Digestive and Kidney Diseases) grant (R01 DK106027). The research in KNN lab is supported by Villum Young Investigator grant (#00025397), Danish Institute of Advanced Study (D-IAS) and Novo Nordisk grants (#NNF18OC0052874, #NNF19OC0056962). We thank Dr. Sen Li for contributing software development.

ACKNOLEDGEMENT

We are grateful to Patrick Martin for proofreading of the manuscript and Sen Li for updating TENET.

AUTHOR CONTRIBUTIONS

K.J.W. conceived and designed TENET. J.K., S.T.J. and K.N.N. performed the experiments and analyzed the data. J.K., K.J.W. and K.N.N. wrote the paper.

CONFLICT OF INTEREST

The authors declare no competing interests.

REFERENCES

Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine JC, Geurts P, Aerts J, et al. 2017. SCENIC: Single-cell regulatory network inference and clustering. *Nat Methods* **14**: 1083–1086.

Alexandrova S, Kalkan T, Humphreys P, Riddell A, Scognamiglio R, Trumpp A, Nichols J. 2016. Selection and dynamics of embryonic stem cell integration into early mouse embryos. *Dev* **143**: 24–34.

Almousailleakh M, Saba-El-Leil MK, Wray J, Smith A, Meloche S, Kunath T. 2007. FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. *Development* **134**: 2895–902.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* **57**: 289–300.

Bibel M, Richter J, Lacroix E, Barde YA. 2007. Generation of a defined and uniform population of

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CNS progenitors and neurons from mouse embryonic stem cells. *Nat Protoc* **2**: 1034–43.

Chan TE, Stumpf MPH, Babtie AC. 2017. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst* **5**: 251–267.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, et al. 2008. Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell* **133**: 1106–17.

Cho KH, Choo SM, Jung SH, Kim JR, Choi HS, Kim J. 2007. Reverse engineering of gene regulatory networks. *IET Syst Biol* **1**: 149–163.

Choi M, Shi J, Jung SH, Chen X, Cho KH. 2012. Attractor landscape analysis reveals feedback loops in the p53 network that control the cellular response to DNA damage. *Sci Signal* **5**: ra83.

Davidson EH, Levine MS. 2008. Properties of developmental gene regulatory networks. *Proc Natl Acad Sci U S A* **105**: 20063–20066.

Deshpande A, Chu L-F, Stewart R, Gitter A. 2019. Network Inference with Granger Causality Ensembles on Single-Cell Transcriptomic Data. *bioRxiv*.

Ficz G, Hore TA, Santos F, Lee HJ, Dean W, Arand J, Krueger F, Oxley D, Paul YL, Walter J, et al. 2013. FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell* **13**: 351–9.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, et al. 2012. Architecture of the human regulatory network derived from ENCODE data Supplementary Information. *Nature* **489**: 91–100.

Habibi E, Brinkman AB, Arand J, Kroeze LI, Kerstens HHD, Matarese F, Lepikhov K, Gut M, Brun-Heath I, Hubner NC, et al. 2013. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* **13**: 360–369.

Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. 2016. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13**: 845–8.

Halpern KB, Shenhav R, Massalha H, Toth B, Egozi A, Massasa EE, Medgalia C, David E, Giladi A, Moor AE, et al. 2018. Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. *Nat Biotechnol* **36**: 962–970.

Hartemink AJ. 2005. Reverse engineering gene regulatory networks. *Nat Biotechnol* **23**: 554–555.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38**: 576–589.

Hlaváčková-Schindler K, Paluš M, Vejmelka M, Bhattacharya J. 2007. Causality detection based on

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

information-theoretic approaches in time series analysis. *Phys Rep* **441**: 1–46.

Hormoz S, Singer ZS, Linton JM, Antebi YE, Shraiman BI, Elowitz MB. 2016. Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. *Cell Syst* **3**: 419–433.

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. 2010. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **5**.

Ito S, Dalessio AC, Taranova O V., Hong K, Sowers LC, Zhang Y. 2010. Role of tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**: 1129–33.

Kalkan T, Olova N, Roode M, Mulas C, Lee HJ, Nett I, Marks H, Walker R, Stunnenberg HG, Lilley KS, et al. 2017. Tracking the embryonic stem cell transition from ground state pluripotency. *Dev* **144**: 1221–1234.

Kim J, Choi M, Kim J-R, Jin H, Kim VN, Cho K-H. 2012. The co-regulation mechanism of transcription factors in the human gene regulatory network. *Nucleic Acids Res* **40**: 8849–61. http://www.ncbi.nlm.nih.gov/pubmed/22798495%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3467061.

Kim J, Park S-M, Cho K-H. 2013. Discovery of a kernel for controlling biomolecular regulatory networks. *Sci Rep* **3**: 2223.

Kim M, Newth D, Christen P. 2016. Macro-level information transfer in social media: Reflections of crowd phenomena. *Neurocomputing* **172**: 84–99.

Kuleshov M V., Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**: W90–W97.

Langmead B, Salzberg SL, Langmead. 2013. Bowtie2. *Nat Methods* **9**: 357–359.

Lanner F, Rossant J. 2010. The role of FGF/Erk signaling in pluripotent cells. *Development* **137**: 3351–60.

Leitch HG, Mcewen KR, Turp A, Encheva V, Carroll T, Grabole N, Mansfield W, Nashun B, Knezovich JG, Smith A, et al. 2013. Naive pluripotency is associated with global DNA hypomethylation. *Nat Struct Mol Biol* **20**: 311–316.

Li F, Long T, Lu Y, Ouyang Q, Tang C. 2004. The yeast cell-cycle network is robustly designed. *Proc Natl Acad Sci* **101**: 4781–4786.

Liu L, Michowski W, Inuzuka H, Shimizu K, Nihira NT, Chick JM, Li N, Geng Y, Meng AY, Ordureau A, et al. 2017a. G1 cyclins link proliferation, pluripotency and differentiation of embryonic stem cells. *Nat Cell Biol* **19**: 177–188.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Liu Z, Wang L, Welch JD, Ma H, Zhou Y, Vaseghi HR, Yu S, Wall JB, Alimohamadi S, Zheng M, et al. 2017b. Single-cell transcriptomics reconstructs fate conversion from fibroblast to cardiomyocyte. *Nature* **551**: 100–104.

Lizier JT. 2014. JIDT: An Information-Theoretic Toolkit for Studying the Dynamics of Complex Systems. *Front Robot AI* **1**: 11.

Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, et al. 2006. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* **38**: 431–40.

Luna-Zurita L, Stirnimann CU, Glatt S, Kaynak BL, Thomas S, Baudin F, Samee MAH, He D, Small EM, Mileikovsky M, et al. 2016. Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. *Cell* **164**: 999–1014.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A. 2006. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**: S7.

Martello G, Bertone P, Smith A. 2013. Identification of the missing pluripotency mediator downstream of leukaemia inhibitory factor. *EMBO J* **32**: 2561–74.

Martello G, Sugimoto T, Diamanti E, Joshi A, Hannah R, Ohtsuka S, Göttgens B, Niwa H, Smith A. 2012. Esrrb is a pivotal target of the Gsk3/Tcf3 axis regulating embryonic stem cell self-renewal. *Cell Stem Cell* **11**: 491–504.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**.

Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, Hayashi T, Nikaido I. 2017. SCODE: An efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33**: 2314–2321.

Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, Aerts S. 2019. GRNBoost2 and Arboreto: Efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**: 2159–2161.

Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, MacAulay IC, Jawaid W, Diamanti E, et al. 2015. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol* **33**: 269–276.

Møller AF, Natarajan KN. 2020. Predicting gene regulatory networks from cell atlases. *Life Sci Alliance* **3**: e202000658.

Nowotschin S, Setty M, Kuo YY, Liu V, Garg V, Sharma R, Simon CS, Saiz N, Gardner R, Boutet SC,

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

et al. 2019. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* **569**: 361–367.

Orlandi JG, Stetter O, Soriano J, Geisel T, Battaglia D. 2014. Transfer entropy reconstruction and labeling of neuronal connections from simulated calcium imaging. *PLoS One*.

Pantier R, Tatar T, Colby D, Chambers I. 2019. Endogenous epitope-tagging of Tet1, Tet2 and Tet3 identifies TET2 as a naïve pluripotency marker. *Life Sci Alliance* **2**: e201900516.

Papili Gao N, Ud-Dean SMM, Gandrillon O, Gunawan R. 2018. SINCERITIES: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* **34**: 258–266.

Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. 2020. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods*.

Qiu D, Ye S, Ruiz B, Zhou X, Liu D, Zhang Q, Ying QL. 2015. Klf2 and Tfcp2l1, Two Wnt/β-Catenin Targets, Act Synergistically to Induce and Maintain Naive Pluripotency. *Stem Cell Reports* **5**: 314–22.

Qiu X, Rahimzamani A, Wang L, Ren B, Mao Q, Durham T, McFaline-Figueroa JL, Saunders L, Trapnell C, Kannan S. 2020. Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. *Cell Syst* **10**: 265–274.

Riso V, Cammisa M, Kukreja H, Anvar Z, Verde G, Sparago A, Acurzio B, Lad S, Lonardo E, Sankar A, et al. 2016. ZFP57 maintains the parent-of-origin-specific expression of the imprinted genes and differentially affects non-imprinted targets in mouse embryonic stem cells. *Nucleic Acids Res* **44**: 8165–78.

Sanchez-Castillo M, Blanco D, Tienda-Luna IM, Carrion MC, Huang Y. 2018. A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics* **34**: 964–970.

Schreiber T. 2000. Measuring information transfer. *Phys Rev Lett* **85**: 461–464.

Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N, Pe'Er D. 2016. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* **34**: 637–645.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–504.

Specht AT, Li J. 2017. LEAP: Constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics* **33**: 764–766.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884163; this version posted April 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Spinney RE, Prokopenko M, Lizier JT. 2017. Transfer entropy in continuous time, with applications to jump and neural spiking processes. *Phys Rev E*.

Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S. 2018. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**: 477.

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**: 381–6.

Tuck AC, Natarajan KN, Rice GM, Borawski J, Mohn F, Rankova A, Flemr M, Wenger A, Nutiu R, Teichmann S, et al. 2018. Distinctive features of lincRNA gene expression suggest widespread RNA-independent functions. *Life Sci Alliance* **1**: e201800124.

van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, et al. 2018. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**: 1–14.

Wang G, Du C, Chen H, Simha R, Rong Y, Xiao Y, Zeng C. 2010. Process-based network decomposition reveals backbone motif structure. *Proc Natl Acad Sci U S A* **107**: 10478–10483.

Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, Rajewsky N, Simon L, Theis FJ. 2019. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* **20**: 59.

Wollstadt P, Martínez-Zarzuela M, Vicente R, Díaz-Pernas FJ, Wibral M. 2014. Efficient transfer entropy analysis of non-stationary neural time series. *PLoS One*.

Wray J, Kalkan T, Smith AG. 2010. The ground state of pluripotency. *Biochem Soc Trans* **38**: 1027–32.

Zhu S, Wurdak H, Wang J, Lyssiotis CA, Peters EC, Cho CY, Wu X, Schultz PG. 2009. A Small Molecule Primes Embryonic Stem Cells for Differentiation. *Cell Stem Cell* **4**: 416–426.

Zou M, Conzen SD. 2005. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* **21**: 71–79.