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

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 KNN (Email: knn@bmb.sdu.dk) and KJW (Tel: +45-3533-1419 ; Email: kyoung.won@bric.ku.dk).

ABSTRACT

Gene expression data has been widely used to infer gene regulatory networks (GRNs). Recent single-cell RNA sequencing (scRNAseq) data, containing the expression information of the individual cells (or status), are highly useful in blindly reconstructing regulatory mechanisms. However, it is still not easy to understand transcriptional cascade from large amount of expression data. Besides, the reconstructed networks may not capture the major regulatory rules.

Here, we propose a novel approach called TENET to reconstruct the GRNs from scRNAseq data by calculating causal relationships between genes using transfer entropy (TE). We show that known target genes have significantly higher TE values. Genes with higher TE values were more affected by various perturbations. Comprehensive benchmarking showed that TENET outperformed other GRN prediction algorithms. More importantly, TENET is uniquely capable of identifying key regulators. Applying TENET to scRNAseq during embryonic stem cell differentiation to neural cells, we show that Nme2 is a critical factor for 2i condition specific stem cell self-renewal.

Keyword: Gene regulatory network, single cell RNA sequencing, causal relationship, stem cell pluripotency, Nme2

INTRODUCTION

Understanding regulatory mechanisms is a key question in biology to understand cellular processes. Various approaches including genome-wide location analysis using chromatin immunoprecipitation followed by genome-wide sequencing (ChIP-seq)[1,2] and perturbation analysis were designed to explain the causal relationships between genes[3,4]. However, protein binding information is limited by the availability of antibody and identifying target genes is difficult when bound in the intergenic region. Moreover, perturbation analysis is hard to measure the strength of the causal relationships with the target genes. Systems approaches to predict regulators and their target genes have been suggested prior to testing in the wet-lab experiments to reduce the cost and time[5–8]. However, previous attempts were applied to inferring gene regulator networks (GRNs) only for a limited number of genes[9–11] and/or cannot detect causal relationships effectively[12,13].

When dealing with causal relationships, time is involved, i.e. an effect cannot occur before its cause. To utilize time to identify the cause (the regulator) and the effect (the target genes), a series of expression data across multiple time points is useful. Single cell RNA sequencing (scRNAseq) can provide sequential expression data from the cells aligned along the virtual time, called pseudo-time[14–16]. Previously, indeed, the peak locations across pseudo-time have been naively used in predicting potential regulators[15,17]. It is based on the assumption that the expression profile of a potential regulator is proceeded by the expression pattern of a target gene along the pseudo-time. Systematic approaches to quantify potential causal relationships and reconstruct GRNs are still highly required to understand biological processes underlying in the data.

We hypothesize 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 (cause) to the other variable (effect) in reducing uncertainty[18,19]. TE has been successfully applied to the estimation of functional connectivity of neurons[20–22] and social influence in social networks[23]. By adopting TE, we developed an approach called TENET (https://github.com/neocaleb/TENET), an algorithm to reconstruct GRNs from scRNAseq data. Using TENET, we identified the relationships between regulators and the target genes. TE values of the known target genes were significantly higher than random targets. Interestingly, target genes with higher TE values were affected more by the perturbation analysis, suggesting that TE score measure the dependency of a target gene to its regulator.

We performed benchmarking tests for the existing GRN reconstructors. 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 identify regulator of the key biological processes. TENET identified pluripotency factors as the key regulator while other competitors could not find them or with worse prediction performance. Expanding GRNs we newly found that Nme2 is a 2i condition specific key transcription factor (TF) that regulates majority of pluripotency genes. Inhibition of Nme2 in mouse embryonic stem cells (mESC) dramatically prevented cell proliferation in a 2i condition specific manner. In summary, we show that TENET has a potential to elucidate previously uncharacterized regulatory mechanisms by reprocessing scRNAseq data.

RESULTS

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. The TE from X to Y is defined as follows:

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

where $H(X)$ is Shannon entropy of X and L denotes the length of the past events considered for calculating TE. Given the pseudo-time ordered expression profiles (Fig. 1a), TE applied to scRNAseq quantifies the causal relationships of a gene X to a gene Y (Fig. 1b) by considering the past events of the two genes. TE can represent the level of the information in X that contributes to the prediction of the current event Y_t. We obtained the significantly high relationships between genes after modeling all possible relationships with normal distribution (Benjamini-Hochberg's false discovery rate (FDR)[24]<0.01). Potential indirect relationships were removed by applying data processing inequality[8] (Fig. 1c) (see Methods). TENET can run various set of potential regulators. In can run only for known set of gene, entire transcription factors (TFs), or even for entire genes to find new regulators. Network analysis can be followed to understand key regulators in the networks.

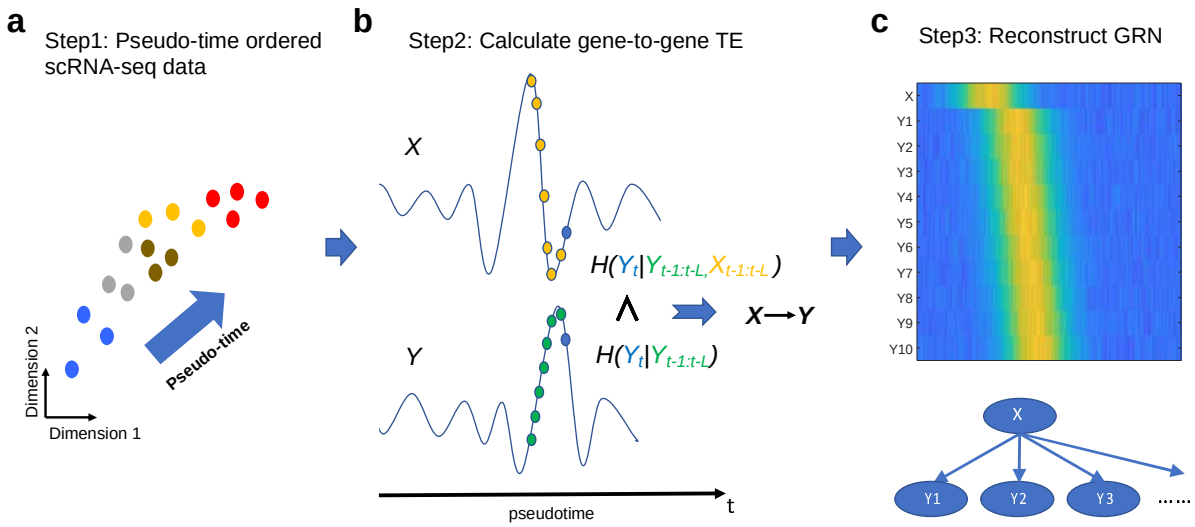

Figure 1. **TENET reconstructs GRNs from a pseudo-time ordered single cell transcriptome data using TE**. **a**. Step 1: Pseudo-time ordered scRNAseq data are used as the input for TENET. **b**. Step 2: TENET calculates gene-to-gene pairwise TE. **c**. Step 3: A reconstructed GRN is composed of strong causal relationships.

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

We applied TENET to the scRNAseq data during mESC differentiation into neural progenitor cells (NPCs)[25]. We chose this experimental system as several ChIP-seq and RNAseq datasets are publicly available for validation[2]. Visualization of the scRNAseq data using tSNE showed potential differentiation trajectory (Fig. 2a). Consistent with differentiation time course, pluripotency markers including Pou5f1 (or Oct4), Sox2 and Nanog were highly expressed in the mESC population whereas NPC markers such as Pax6 and Slc1a3 was highly expressed in the NPCs (Supplementary Fig. 1). First, 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[2] as their occupancy is often observed at the promoter region of their target genes. Peak calling using Homer[26] found 541 c-Myc promoter proximal peaks. 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 (p-value-7.94e-53) than the randomly selected genes (Fig. 2b). We also confirmed that ChIP-seq binding targets for n-Myc, E2f1 and Zfx have significantly higher TE values compared with the random targets (Supplementary Fig. 2a-c).

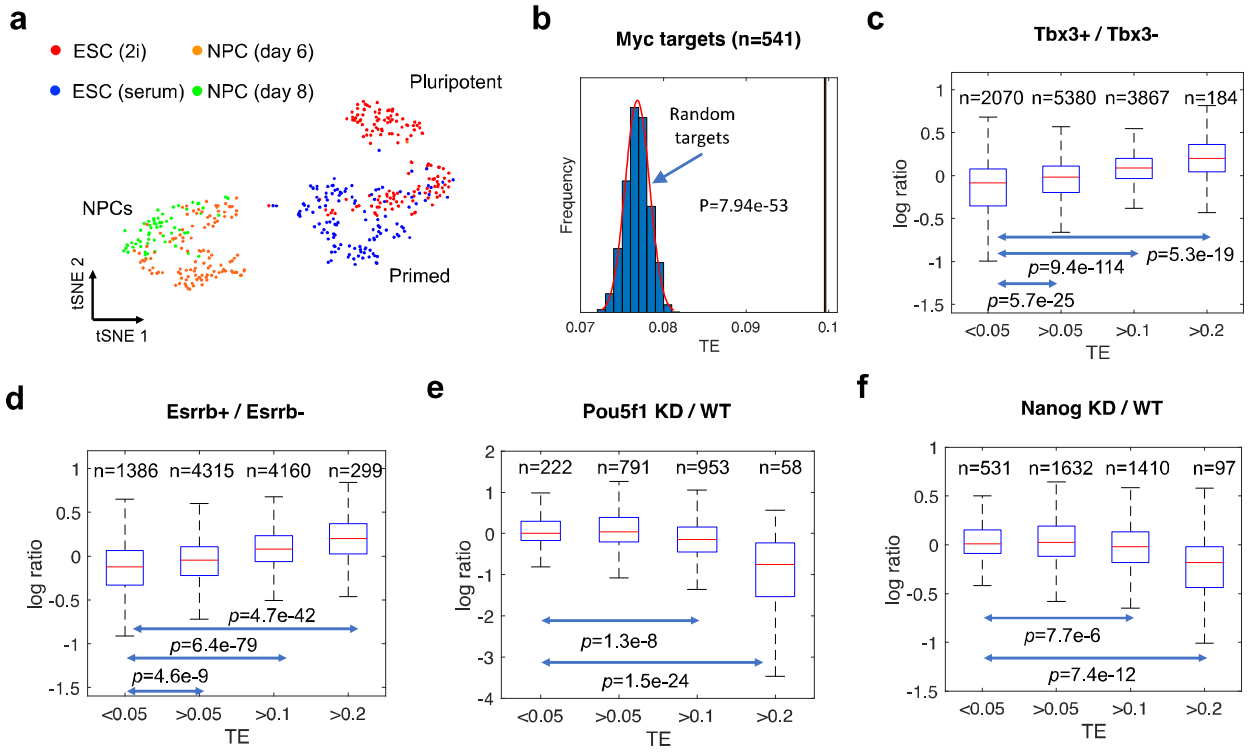

Figure 2. Validation of TENET-inferred GRNs for the mouse embryonic stem cell (mESC) pluripotency. **a**. A tSNE plot of the mESCs (2i and serum) and NPCs shows distinct expression. **b**. The c-Myc target genes have higher TE values than the randomly selected 541 genes (repeated 1000 times). The expression 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).

As an additional test, we performed evaluation using the scRNAseq dataset for the reprogramming of mouse fibroblasts into induced cardiomyocytes[27]. Investigation using Gata4 ChIP-seq in cardiomyocytes[28] further confirmed that the 331 Gata4 target genes also have significantly higher TE values compared with random targets (Supplementary Fig. 2d).

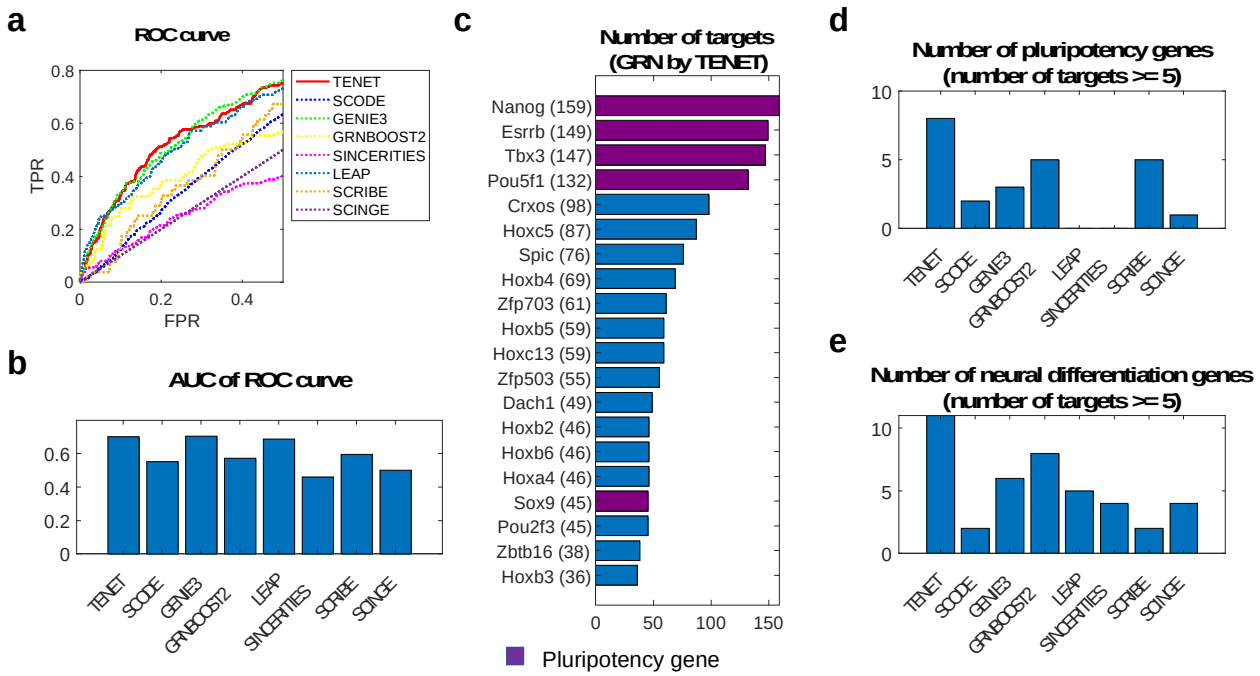

Figure 3. **TENET outperformed in terms of reconstruction of GRN and predicting key regulatory factors for mESC pluripotency**. **a**. Receiver operating characteristic (ROC) curves for the mESC GRN by TENET and seven different algorithms. **b**. Area under curve (AUC) of the ROC curves. **c**. Key regulatory factors for mESC pluripotency predicted by TENET. The purple bar denotes pluripotency gene. Comparison of capability of predicting key genes (pluripotency genes (**d**) and neural differentiation genes (**e**)) using hub genes (number of targets >= 5) in the reconstructed GRNs.

TE values reflect the degree of dependency to the regulator

We further examined the TE values of the potential TF target genes identified by knock-in of Esrrb and Tbx3 and knockdown of Pou5f1 and Nanog[3,4]. 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 genes with low TE values (<0.05) did not change their expression levels upon the perturbation. However, the expression levels of the genes with high TE values increased upon knock-in of Esrrb and Tbx3 and decreased upon knockdown of Pou5f1 and Nanog. The changes were more distinct for the genes with higher TE values (>0.2) (Fig. 2c-g). These indicate that TE values reflects the degree of dependency of the target genes to the expression of their regulator.

TENET outperforms other GRN reconstruction algorithms

To evaluate overall performance of reconstructed GRN, we used Beeline[29], a benchmarking software for GRN inference algorithms for the mESC scRNAseq dataset[25]. Among them, we performed

benchmarking only for those algorithms that can implement large scale GRN reconstruction including SCODE[30], GENIE3[31], GRNBOOST2[32], SINCERITIES[33], LEAP[34], SCRIBE[35], and SCINGE[36]. To prepare stringent datasets for evaluation, we only considered as targets if their expression levels were significantly changed upon Nanog, Pou5f1, Esrrb, and Tbx3 perturbations[3,4] and the binding occupancy from ChIP-seq (+/-50kbps)[2] (see Supplementary Fig. 3a and Methods).

Benchmarking was performed for the 3,280 highly variable genes computed from the mESC scRNAseq data (see Methods). Beeline[29] provided the comprehensive results after running all GRN reconstructors. The receiver operating characteristic (ROC) curves (Fig. 3a) showed that TENET, GENIE3 and LEAP outperformed other predictors in predicting targets of Nanog, Pou5f1, Esrrb, and Tbx3 (Fig. 3a-b and Supplementary Fig. 3b). SCRIBE, which is designed based on TE as well, showed worse performance than TENET. Both GENIE3 and GRNBOOST2 are based on the same tree-based ensemble algorithm. But, we observed different performance in this test.

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 sorted the regulators based on the number of target genes for each GRNs. The top 4 regulators by TENET were the markers for pluripotency (Pou5f1, Nanog, Esrrb, and Tbx3) (Fig. 3c). Compared with it, majority of other predictors did not identify these key genes in the hub list. For instance, GENIE3 which showed comparable performance with TENET found only Nanog as the 14th of the top regulators. LEAP, another competitor, did not find any pluripotency markers. On the other hand, SCRIBE, a TE-based GRN predictor, identified Nanog, Pou5f1, Esrrb, Tbx3 as the top regulators, suggesting the algorithmic advantages of TE (Supplementary Fig. 4).

Intrigued by this, we investigated whether the identified hubs are associated with “pluripotency” or “neural differentiation” using the list of the genes obtained from gene ontology (GO) database (see Methods). Collectively, TENET identified far exceeding number of genes related with the relevant GO terms compared to other methods (Fig. 3d-e). The performance of LEAP, which used time-lagged co-expression along the genes assigned pseudo-time did not perform well in this experiment, suggesting the algorithmic advantages of TENET.

To further test if TENET can suggest key regulatory factors in various biological systems, we reconstructed a GRN based on scRNAseq data for direct reprogramming from mouse fibroblast into cardiomyocyte by overexpressing Mef2c, Tbx5 and Gata4[27] (Supplementary Fig. 5a). We tested if these overexpressed factors were well predicted in TENET. We confirmed that TENET identified those three major reprogramming factors (Mef2c, Tbx5 and Gata4) as well as other genes associated with cardiomyocytes as the key regulators (Supplementary Fig. 5b). In comparison, other predictors found low numbers of genes associated with cardiomyocytes (Supplementary Fig. 5c). Indeed, the networks by other GRN reconstructors failed to identify the 3 reprogramming factors and found only a few regulators related with cardiomyocytes. Interestingly, TE-based SCRIBE also only found Mef2c among the key factors for this experiment (Supplementary Fig. 6). Collectively, our results show that TENET can capture key regulatory genes for the biological processes.

TENET mimics the controllability (key regulators) of Boolean network dynamics

To further investigate the characteristics of TENET in finding key regulators, we compared the reconstructed networks with the Boolean networks (BNs)[9]. Considering all possible binary status of the members, BNs have been widely used to model biological systems[10,37,38]. BNs can simulate overexpression or knock-out of a gene and its consequences from the inferred networks. Therefore, BNs can evaluate how much a member can influence the steady-state dynamics of the networks in combination with other members (called "controllability")[39]. Besides, BNs can represent causal relationships among the members. We found a BN-inferred GRN for 20 TFs inferred from the scRNAseq data for mouse early blood development[9] (Supplementary Fig. 7). Using the BN-inferred GRN as the surrogate of the gold standard, we first evaluate if the networks from GRN reconstructors well mimic the BN-inferred GRN. The comparison showed that TENET and GRNBOOST2 outperforms other approaches in both directed and undirected networks (Supplementary Fig. 8a-b)

In a series of experiment, TENET showed a function to find key regulators. We evaluated this using the simulation of the BNs. The number of final stable states (known as attractors) was calculated when a member is perturbed. For this the BNs simulates all possible states of the members. If the member of interest is a critical one, the number of attractors is small. Therefore, we calculated correlations between the out-degree of a gene in the inferred networks and the number of attractors when perturbing the gene. The simulation showed that the TENET-inferred network has the lowest correlation with the number of attractors compared with other methods (Supplementary Fig. 8c). This further demonstrated that TENET has a capability to identify key regulators.

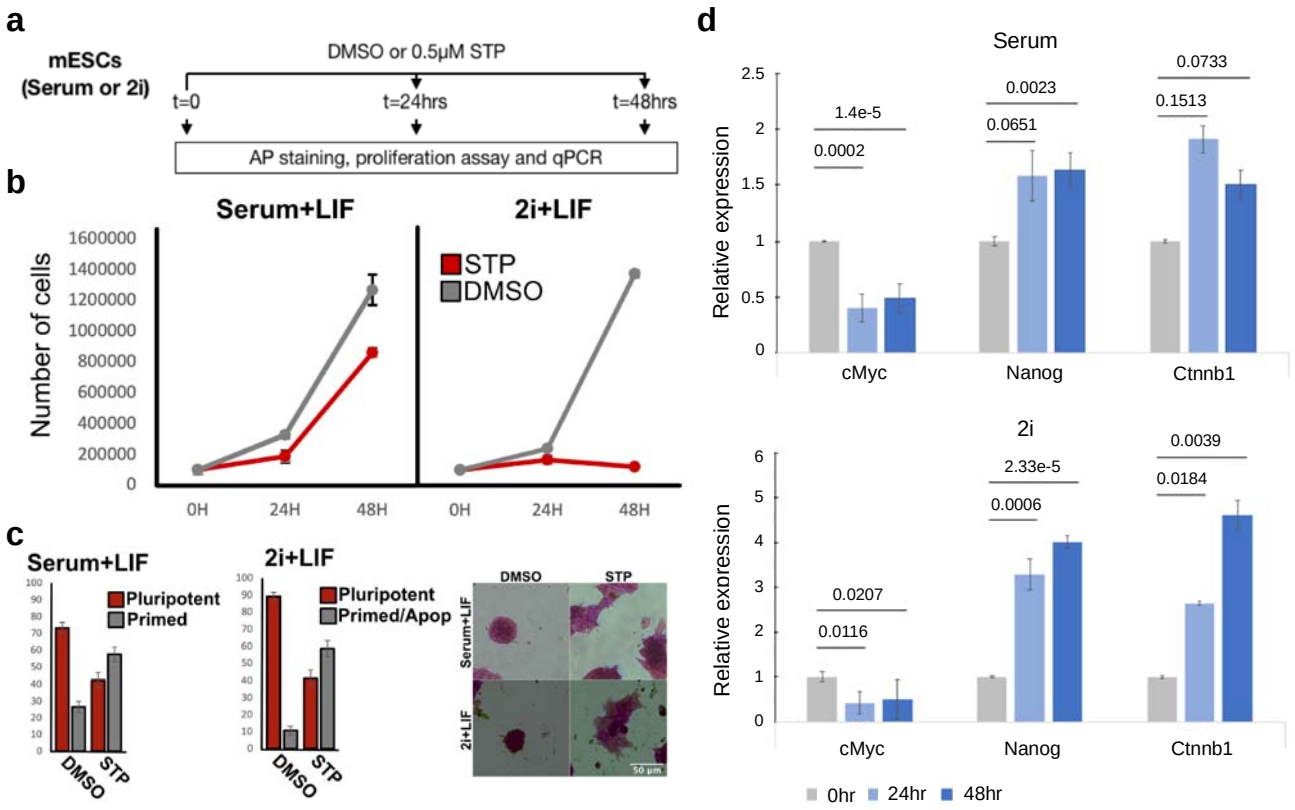

Figure 4. Nme2 inhibition completely blocked proliferation of mESC in 2i condition. **a**. Experimental design of the experiment. The mESC were seeded in either Serum or 2i culture conditions and treated with either DMSO (control) Stauprimide (STP) for 24hr and 48hr. The 6 samples were assayed for proliferation rates, relative transcript expression and for pluripotency using Alkaline Phosphatase (AP). **b**. Cell proliferation assay for mESCs cultured in Serum and 2i conditions with either DMSO (Control) or 0.5uM STP. The data are mean ± SEM from 2 biologically independent replicates. **c**. STP treatment leads reduction in AP positive colonies both in Serum (after 48 hours) and 2i (after 24 hours) condition highlighting differentiation. Representative image of AP positive pluripotent and primed colonies in both culture conditions. The data are mean ± SEM from 3 biologically independent replicates. **d**. The c-Myc transcript levels are down-regulated both in 2i and serum upon STP treatment, owing to impaired Nme2 nuclear localization. The NME2 target genes in TENET (Nanog and Ctnnb1) are selectively upregulated.

TENET identifies condition specific regulator

To search for new regulators besides the known TFs, we extended the GRN considering 13,694 highly variable genes as regulators as well as target genes (see Methods). Therefore, regulators are not limited to the TFs in this setting. We were interested to find Nme2 and Fgf4 as the top regulators. We

confirmed that these are supported by a previous report that these genes have been implicated in either pluripotency or differentiation[40,41] (Supplementary Fig. 9).

As we analyzed mESCs from two culture conditions (namely 2i and serum), we further questioned if TENET can further distinguish them and identify condition specific GRNs. We reconstructed GRNs for 2i and serum condition separately and compared the number of condition specific target genes of the key regulators (see Methods). Interestingly, we found that Nme2 as the top regulator in the 2i-specific GRNs. Besides it, we found that pluripotent factors such as Fgf4, Pou5f1, and Nanog are 2i-specific whereas DNA methylation factors such as Tet1 and Dnmt3l are serum-specific (Supplementary Fig. 10), which is consistent with a previous report about higher DNA methylation in serum than 2i[42,43].

Intrigued by this, we questioned condition specific role for Nme2 in the 2i and serum condition. After culturing mESCs in 2i and serum, we treated them with Stauprimide (STP) which blocks nuclear localization of Nme2 (Fig. 4a). The number of cells did not increase in the 2i condition, showing that the effect of the STP treatment is more profound in the 2i condition (Fig. 4b). The serum cultured mESCs grown in 0.5uM STP (for 24 and 48hrs) had reduced proliferation and increased heterogeneity, as assessed by Alkaline phosphatase (AP) stained colonies (Pluripotent and Primed colonies; Fig. 4b-c). In contrast, 0.5uM STP treatment in 2i cultured cells led to inhibited proliferation at 24 hours, and very few viable cells at 48 hours (Fig. 4b-c). Previously, c-Myc has been reported as the target gene of Nme2[40]. TENET predicted c-Myc as a target of Nme2. We confirmed that c-Myc expression was significantly downregulated upon STP treatment in both culture conditions (Fig. 4d). Additionally, we measured expression levels of several TFs including Nanog and Ctnnb1 targeted by Nme2 in the TENET-inferred GRN. We found that both Nanog and Ctnnb1 transcripts are highly upregulated upon STP treatment in both culture conditions but more significant in 2i condition, indicating condition specific regulation of Nme2 as predicted by TENET (Fig. 4d).

DISCUSSION

Systems approaches to infer GRNs can provide the hypothesis for transcriptional regulation under biological process. Previous approaches using bulk cells were limited because they cannot capture the continuous cellular dynamics because the cells must be synchronized in order to avoid obtaining “average out” expression. scRNAseq has been emerged as an alternative because each cell provides the transcriptomic snapshot in a massive scale. Subsequently, computational approaches have been developed to use of scRNAseq for GRN reconstruction[9,11,12,30–36].

These algorithms including TENET used temporal gene expression to infer GRNs. For example, GENEI3[31] and GRNBOOST2[32] were originally developed for temporal bulk expression data using ensembles of regression trees. LEAP[34] calculates possible maximum time-lagged correlations. SINCERITIES[33] and SCINGE[36] use Granger causality a statistical test for predicting one time series data using another time series data which have been used in economics. SCODE[30] use a mechanistical model ordinary differential equation for the pseudo-time aligned scRNAseq data. Compared with them, TENET makes use of the power of information theory by adopting TE on the gene expression along the pseudo-time.

We showed that TE values of the known target genes were significantly higher than randomly selected genes (Fig. 2b and Supplementary Fig. 2). The higher TE values were associated with gene expression changes once the associated regulator is perturbed (Fig. 2c-f). These show that TE well quantifies the causal relationships between genes. We performed comprehensive benchmarking using the Beeline[29] which provides the comparison of diverse GRN reconstructor in its automated pipeline. The test using the scRNAseq for mESC pluripotency and differentiation showed that TENET is one of the top performing GRN reconstructor along with GENIE3 and LEAP.

We found that TENET is working far better than other approaches in identifying key regulators. This is important because the reconstructed GRNs only find some relationships which are not that important in interpreting the biological processes of the given scRNAseq datasets. For instance, GENIE3 and LEAP, which showed similar performance with TENET could not capture the key known regulators in the networks (Supplementary Fig. 4). TENET, on the other hand, ranked Nanog, Pou5f1, Esrrb and Tbx3 as the top 4 regulators and identified genes related with stem cell or neural development. We also confirmed that TENET (but not others) identified 3 cardiomyocyte reprogramming factors in the inferred network. We further investigate the function of TENET using the BN. TENET best matched with the BN among other predictors. Besides, the regulators that TENET identified also showed to be more important in the independently obtained BN. Even though BN is not a perfect model of the system, it provides the comprehensive overview of the system by visiting all potential status. Our results suggest the algorithmic merits of TENET. In line with it, TE-based SCRIBE also identified key pluripotency factors.

With the power to predict key regulators, TENET has been applied to identify condition specific regulators. Prediction of condition specific regulators is of great value to provide hypothesis about the mechanism that can be subsequently validated. TENET predicted that Nme2 is a key regulator for 2i-condition-specific mESC pluripotency. In consistent with our prediction, perturbing

Nme2 STP lead to reduced proliferation more profoundly in the 2i condition. These results suggest that TENET is a useful approach to predict gene regulatory mechanisms from scRNAseq.

METHODS

Algorithm for TENET

TENET measures the amount of causal relationships using the scRNAseq data aligned along pseudo-time. Given pseudo-time ordered scRNAseq data, TENET calculates bidirectional pairwise TE values for selected genes using JAVA Information Dynamics Toolkit (JIDT)[44]. 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"[18]. The joint PDF of two genes x and y can be calculated as follows:

$$\hat{p}_r(x_n, y_n) = \frac{1}{N}\sum_{n'=1}^{N} \Theta\left(\left|\begin{pmatrix} x_n - x_{n'} \\ y_n - y_{n'} \end{pmatrix}\right| - r\right), \qquad (2)$$

where Θ is a kernel function and N is the number of cells. We used step kernel ($\Theta(x>0)=1$, $\Theta(x\leq 0)=0$) and kernel width r=0.5 as default. The causal relationships using TE are calculated using the Eq (1). We reconstructed the GRNs by integrating all TE values for gene pairs. To remove potential indirect relationships, we applied the data processing inequality[8], i.e. iteratively eliminating feed-forward loops. The feed-forward loop is defined by a network motif composed of three genes, where when gene X regulates gene Y and both gene X and Y regulate gene Z. We trimmed the link from gene X to gene Z if $TE_{X\to Z}$ is less than the minimum value of $TE_{X\to Y}$ and $TE_{Y\to Z}$ while allowing a threshold. Finally, we reconstructed a GRN consisting in the links with significant Benjamini-Hochberg's FDR[24] by performing one-sided z-test considering the all trimmed TE values as a normal distribution. The hub node is identified by calculating the number of targets (outgoing links).

Data processing of scRNAseq data

To test TENET, we downloaded two scRNAseq datasets obtained from mESCs[25] and mouse cardiomyocytes[27]. Wishbone a pseudo-time analysis tool[14] was used on the two datasets. As an input gene list for benchmarking of mESC dataset, we used 3,280 highly variable genes which have expression criterion log2count>1 in more than 10% of the whole cells and coefficient of variation > 1.5. To extend the GRN inferred by TENET, we used 13,694 highly variable genes which have expression criterion log2count>1 in more than 10% of the whole cells. For the mouse cardiomyocytes, we used 8,640 highly variable genes which have expression criterion log2count>1 in more than 10% of the whole cells and coefficient of variation > 1. To reconstruct the GRN, we used a regulator gene list

which includes genes with a gene ontology term "regulation of transcription (GO:0006355)" only for the mESC dataset. We generated all the network figures (Supplementary Fig. 3a, Supplementary Fig. 5a, Supplementary Fig. 7 and Supplementary Fig. 9a) using Cytoscape 3.6.1[45].

Gene ontology for the functional gene group

The "pluripotency gene" in Fig. 3c-d and the "neural differentiation gene" in Fig. 3e was obtained from the genes with a gene ontology term "stem cell population maintenance (GO:0019827)" and "neuron differentiation (GO:0030182)", respectively. We used gene ontology terms "cardiac muscle cell differentiation (GO:0055007)" and "cardiac muscle contraction (GO:0060048)" for "cardiomyocyte gene" in Supplementary Fig. 5 and 6.

Public data of gene expression and ChIP-seq

We downloaded an RNA-seq data obtained from mESCs with three different combinations of double knock-in for Esrrb and Tbx3 (Esrrb-/Tbx3-, Esrrb+/Tbx3-, Esrrb+/Tbx3+)[4]. 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 criterion, respectively. The target genes of Nanog and Pou5f1 were identified by a downloaded microarray data obtained from mESC with Nanog and Pou5f1 knockdown[3]. To identify target genes of these two TFs, we used 2-fold and p-value < 0.01 provided in the original data analysis.

ChIP-seq data for Pou5f1, Esrrb, Nanog in mESCs were reanalyzed for peak calling[2]. After removing the adapter sequence using CutAdapt[46] implemented in TrimGalore-0.4.5, we aligned the ChIP-seq reads to the mm10 genome using Bowtie2[47]. ChIP-seq peak calling for each TF was performed by comparing each ChIP sample with GFP control using the findPeaks command in the Homer package[26].

Statistical analysis

A two-sided, one -sample z-test was performed to evaluate the mean of TE values of targets of key factors (c-Myc, n-Myc, E2f1, Zfx, Nme2) in mESCs and a key factor Gata4 in mouse cardiomyocytes by generating a fitted z-distribution based on an empirical distribution of the thousand means of the TE values of the same number of randomly selected genes (Fig. 2b and Supplementary Fig. 2). A two-sided, two-sample Student t-test was used to evaluate the relative expression values by knocking-in of Tbx3 and Esrrb and knocking-down of Pou5f1 and Nanog along with TE values, respectively (Fig. 2c-f).

Condition specific targets

To identify condition specific targets, we reconstructed GRNs on the pseudo-time ordered expression data of 2i+NPCs and serum+NPCs using TENET. Subsequently, the condition specific targets of the top 20 factors in the common GRN (Supplementary Fig. 9) was obtained by the targets in the common GRN exclusively included in the condition specific GRNs. For example, the 35 target genes of Nme2 was included in the 2i-specific but not in the serum-specific GRN whereas the 14 target genes were included in the serum-specific but not in the 2i-specific GRN (Supplementary Fig. 10).

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, 1uM PD0325901 and 1000U/mL (“2i”). For Nme2 experiments, mESCs were treated with either vehicle DMSO (Sigma #02660) or 0.5uM Stauprimide (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 random selected areas of each well were imaged (10x magnification; MICROSCOPE DETAILS) and manually counted. Colonies were marked as pluripotent or primed based on morphology and intensity of AP staining. The results are presented in percentages of positive stained colonies from two biological replicates.

Cell proliferation assay

For proliferation assay, 140.000 mESCs were seeded in a 6-well plate in both 2i and Serum condition. 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 two 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 with 1ug of RNA using high capacity cDNA kit (Applied Biosystem #4368814). Quantitative-PCR was performed using SYBR-green with LightCycler480. Relative gene expression levels were presented as expression level normalized to Gadph as loading/amplification control.

Robustness of the performance of TENET

In order to evaluate the robustness of TENET, we run the Wishbone 57 times with different options on the Boolean expression data[9] of single-cell obtained from early blood development experiments. 57 Wishbone trajectories were obtained by running the Wishbone with 19 different initial states provided in the reference paper[9] and three different choices of cells based on the branches (total cells, trunk + first branch, trunk + second branch).

AVAILABILITY

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

FUNDING

The Novo Nordisk Foundation Center for Stem Cell Biology is supported by a Novo Nordisk Foundation grant No. NNF17CC0027852 and Lundbeck Foundation's Ascending Investigator grant R313-2019-421.

Competing interests

The authors declare no competing interests.

REFERENCES

1. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data Supplementary Information. *Nature* **489**, 91–100 (2012).
2. Chen, X. *et al.* Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell* **133**, 1106–17 (2008).

3. Loh, Y. H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**, 431–40 (2006).
4. Hormoz, S. *et al.* Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. *Cell Syst.* **3**, 419–433 (2016).
5. Hartemink, A. J. Reverse engineering gene regulatory networks. *Nat. Biotechnol.* **23**, 554–555 (2005).
6. Zou, M. & Conzen, S. D. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* **21**, 71–79 (2005).
7. Kim, S., Kim, J. & Cho, K.-H. Inferring gene regulatory networks from temporal expression profiles under time-delay and noise. *Comput. Biol. Chem.* **31**, 239–245 (2007).
8. Margolin, A. A. *et al.* ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**, S7 (2006).
9. Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* **33**, 269–276 (2015).
10. Li, F., Long, T., Lu, Y., Ouyang, Q. & Tang, C. The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci.* **101**, 4781–4786 (2004).
11. Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I. M., Carrion, M. C. & Huang, Y. A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics* **34**, 964–970 (2018).
12. Aibar, S. *et al.* SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
13. Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst.* **5**, 251–267 (2017).
14. Setty, M. *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).
15. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–6 (2014).
16. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–8 (2016).
17. van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 1–14 (2018).
18. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **85**, 461–464 (2000).
19. Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M. & Bhattacharya, J. Causality detection

based on information-theoretic approaches in time series analysis. *Physics Reports* **441**, 1–46 (2007).

20. Orlandi, J. G., Stetter, O., Soriano, J., Geisel, T. & Battaglia, D. Transfer entropy reconstruction and labeling of neuronal connections from simulated calcium imaging. *PLoS One* (2014).
21. Wollstadt, P., Martínez-Zarzuela, M., Vicente, R., Díaz-Pernas, F. J. & Wibral, M. Efficient transfer entropy analysis of non-stationary neural time series. *PLoS One* (2014).
22. Spinney, R. E., Prokopenko, M. & Lizier, J. T. Transfer entropy in continuous time, with applications to jump and neural spiking processes. *Phys. Rev. E* (2017).
23. Kim, M., Newth, D. & Christen, P. Macro-level information transfer in social media: Reflections of crowd phenomena. *Neurocomputing* **172**, 84–99 (2016).
24. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289–300 (1995).
25. Tuck, A. C. *et al.* Distinctive features of lincRNA gene expression suggest widespread RNA-independent functions. *Life Sci. Alliance* **1**, e201800124 (2018).
26. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).
27. Liu, Z. *et al.* Single-cell transcriptomics reconstructs fate conversion from fibroblast to cardiomyocyte. *Nature* **551**, 100–104 (2017).
28. Luna-Zurita, L. *et al.* Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. *Cell* **164**, 999–1014 (2016).
29. Aditya Pratapa, Amogh Jalihal, Jeffrey Law, Aditya Bharadwaj, and T. M. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *bioRxiv* (2019).
30. Matsumoto, H. *et al.* SCODE: An efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33**, 2314–2321 (2017).
31. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **5**, (2010).
32. Moerman, T. *et al.* GRNBoost2 and Arboreto: Efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2019).
33. Papili Gao, N., Ud-Dean, S. M. M., Gandrillon, O. & Gunawan, R. SINCERITIES: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* **34**, 258–266 (2018).

34. Specht, A. T. & Li, J. LEAP: Constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics* **33**, 764–766 (2017).
35. Qiu, X. *et al.* Towards inferring causal gene regulatory networks from single cell expression Measurements. *bioRxiv* (2018). doi:10.1101/426981
36. Deshpande, A., Chu, L.-F., Stewart, R. & Gitter, A. Network Inference with Granger Causality Ensembles on Single-Cell Transcriptomic Data. *bioRxiv* (2019). doi:10.1101/534834
37. Choi, M., Shi, J., Jung, S. H., Chen, X. & Cho, K. H. Attractor landscape analysis reveals feedback loops in the p53 network that control the cellular response to DNA damage. *Sci. Signal.* **5**, ra83 (2012).
38. Wang, G. *et al.* Process-based network decomposition reveals backbone motif structure. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 10478–10483 (2010).
39. Kim, J., Park, S.-M. & Cho, K.-H. Discovery of a kernel for controlling biomolecular regulatory networks. *Sci. Rep.* **3**, 2223 (2013).
40. Zhu, S. *et al.* A Small Molecule Primes Embryonic Stem Cells for Differentiation. *Cell Stem Cell* **4**, 416–426 (2009).
41. Almousailleakh, M. *et al.* 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 (2007).
42. Habibi, E. *et al.* Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* **13**, 360–369 (2013).
43. Leitch, H. G. *et al.* Naive pluripotency is associated with global DNA hypomethylation. *Nat. Struct. Mol. Biol.* **20**, 311–316 (2013).
44. Lizier, J. T. JIDT: An Information-Theoretic Toolkit for Studying the Dynamics of Complex Systems. *Front. Robot. AI* **1**, 11 (2014).
45. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–504 (2003).
46. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, (2011).
47. Langmead, B., Salzberg, S. L. & Langmead. Bowtie2. *Nat. Methods* **9**, 357–359 (2013).