Stem Cell Research 15 (2015) 427–434

Contents lists available at ScienceDirect

Stem Cell Research

journal homepage: www.elsevier.com/locate/scr

Short report

A computational strategy for predicting lineage specifiers in stem cell subpopulations

Satoshi Okawa, Antonio del Sol *

Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 7, Avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg

ARTICLE INFO

Article history:
Received 10 June 2015
Received in revised form 17 July 2015
Accepted 16 August 2015
Available online 2 September 2015

Keywords:
Stem cell differentiation
Lineage specifier
Single-cell gene expression
Transcriptional regulatory network
Seesaw model of differentiation

ABSTRACT

Stem cell differentiation is a complex biological event. Our understanding of this process is partly hampered by the co-existence of different cell subpopulations within a given population, which are characterized by different gene expression states driven by different underlying transcriptional regulatory networks (TRNs). Such cellular heterogeneity has been recently explored with the modern single-cell gene expression profiling technologies, such as single-cell RT-PCR and RNA-seq. However, the identification of cell subpopulation-specific TRNs and genes determining specific lineage commitment (i.e., lineage specifiers) remains a challenge due to the slower development of appropriate computational and experimental workflows. Here, we propose a computational method for predicting lineage specifiers for different cell subpopulations in binary-fate differentiation events. Our method first reconstructs subpopulation-specific TRNs, which is more realistic than reconstructing a single TRN representing multiple cell subpopulations. Then, it predicts lineage specifiers based on a model that assumes that each parental stem cell subpopulation is in a stable state maintained by its specific TRN stability core. In addition, this stable state is maintained in the parental cell subpopulation by the balanced gene expression pattern of pairs of opposing lineage specifiers for mutually exclusive different daughter cell subpopulations. To this end, we devised a statistical metric for identifying opposing lineage specifier pairs that show a significant ratio change upon differentiation. Application of this computational method to three different stem cell systems predicted known and putative novel lineage specifiers, which could be experimentally tested. Our method does not require pre-selection of putative candidate genes, and can be applied to any binary-fate differentiation system for which single-cell gene expression data are available. Furthermore, this method is compatible with both single-cell RT-PCR and single-cell RNA-seq data. Given the increasing importance of single-cell gene expression data in stem cell biology and regenerative medicine, approaches like ours would be useful for the identification of lineage specifiers and their associated TRN stability cores.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Stem cell differentiation is a complex process that involves a multitude of regulatory mechanisms at different organizational levels. Despite accumulating experimental evidence, identification of lineage specifiers and understanding of the regulatory mechanisms of cell-fate commitments are partially hampered by the heterogeneity in stem cell populations. Indeed, stem cells in tissues and culture exist as a heterogeneous population consisting of different subpopulations, which are characterized by different gene expression states driven by different underlying TRNs. Different TRNs in turn determine different propensities for cell fate decision. Hence, conventional bulk gene expression profiling and ChIP-seq approaches generated from a heterogeneous population of cells appear to be suboptimal for studying stem cell differentiation (Moignard et al., 2013). Recent development of modern technologies for single-cell gene expression studies, such as single-cell RT-PCR and RNA-seq, have made possible gene expression profiling of hundreds of cells. They have been successfully used for elucidating heterogeneity in different stem cell systems, including the early embryonic development (Guo et al., 2010; Tang et al., 2010), hematopoiesis (Moignard et al., 2013; Guo et al., 2013), induced pluripotent stem cells (Buganim et al., 2012) and lung alveolar development (Treutlein et al., 2014). Nevertheless, a remaining challenge is the development of computational methods for elucidating complex molecular interaction networks and predicting lineage specifiers within a heterogeneous cell population. A couple of studies has proposed computational workflows for predicting cell lineage specifiers by reconstructing a single TRN that represents multiple cell types (Xu et al., 2014; Moignard et al., 2015). However, it has been revealed that cell subpopulation-specific TRNs showed significant rewiring during differentiation (Moignard et al., 2013). Hence, TRNs that are differentially reconstructed for different cell subpopulations provide a more realistic picture of underlying transcriptional regulatory mechanisms.

* Corresponding author.
E-mail address: antonio.delsol@uni.lu (A. del Sol).

http://dx.doi.org/10.1016/j.scr.2015.08.006
1873-5061/© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Here, we introduce a general method for predicting lineage specifiers in binary-fate differentiation events based on the reconstruction of cell subpopulation-specific TRNs using single-cell gene expression data. Our method is based on a model, in which each stem cell subpopulation is considered to be in a stable state maintained by a TRN stability motif. We particularly focused on a set of circuits known as strongly connected components (SCCs) that we previously used for the prediction of reprogramming determinants (Crespo and Del Sol, 2013). The model further assumes that the stability of a parental stem cell subpopulation, which differentiates into two mutually exclusive daughter cell subpopulations, is maintained by a balance between the two opposing differentiation forces exerted by lineage specifiers for each of the two daughter cell subpopulations. Indeed, this "seesaw model" of stem cell differentiation has been observed during mesendodermal and ectodermal specification of embryonic stem cells (ESCs) (Montserrat et al., 2013; Shu et al., 2013). In this case, the balanced expression of a mesendodermal specifier, Pou5f1, and an ectodermal specifier, Sox2, which mutually activate each other, maintains the pluripotent state. Hence, the method searches for opposing lineage specifier pairs that reside in the TRN stability core of the parental cell subpopulation, and exhibit a significantly unbalanced expression ratio in the daughter cell sub-populations with respect to the parental cell subpopulation.

To assess the applicability of our method, we selected three binary-fate stem cell differentiation systems for which high-quality single-cell gene expression data are available. These examples include the differentiation of inner cell mass (ICM) into either primitive endoderm (PE) or epiblast (EPI) (Guo et al., 2010), the differentiation of different progenitor cells in the hematopoietic system (hematopoietic stem cell (HSC) into either multipotent progenitor (MPP) or megakaryocyte–erythroid progenitor (MEP), MPP into common myeloid progenitor (CMP) or common lymphoid progenitor (CLP), and CMP into either MEP or granulocyte–macrophage progenitor (GMP)) (Guo et al., 2013), and the differentiation of lung alveolar bipotential progenitor (BP) into either alveolar type 1 (AT1) or alveolar type 2 (AT2) (Treutlein et al., 2014). In the first example *Gata6* for PE and *Klf2* for EPI were predicted, which is in full agreement with previously reported experimental observations (Fujikura et al., 2002; Yeo et al., 2014; Gillich et al., 2012). In addition, many well-known lineage specifiers in the hematopoietic system, such as *Cebpa* (Radomska et al., 1998), *Gata1* (Pevny et al., 1991), *Gfi1* (Li et al., 2010) and *Spi1* (PU.1) (Voso et al., 1994) were correctly predicted for appropriate subpopulations, demonstrating the validity of our approach. Finally, our predictions in the relatively understudied lung BP developmental system provided novel candidate lineage specifiers with prior associations with lung development, including *Hes1* (Ito et al., 2000) and *Pou6f1* (Sandbo et al., 2009).

To our knowledge, this is the first computational method that systematically predicts cell lineage specifiers based on cell subpopulation-specific TRNs. Our method does not require pre-selection of candidate genes, and can be applied to any binary-fate differentiation event for which single-cell gene expression data are available. Finally, this method is compatible with both single-cell RT-PCR and single-cell RNA-seq data. Given the increasing importance of single-cell gene expression data in stem cell biology, we believe that approaches like ours would be useful for the identification of lineage specifiers. This should aid in understanding stem cell lineage specification and the development of strategies for regenerative medicine (Li and Kirschner, 2014).

2. Materials and methods

2.1. Formulation of binary-fate stem cell differentiation model

Our model assumes that each stem cell subpopulation is in a stable state – i.e., an attractor – in the gene expression landscape determined by their TRNs. Within TRNs, SCCs, which consist of a set of circuits and confer autonomous stability to TRNs, have been previously used for identifying cell fate determinants (Crespo and Del Sol, 2013; Ertaylan et al., 2014). The model further assumes that such stability is maintained by the balanced expression pattern between opposing lineage specifiers, as was previously demonstrated in the ESC system (Montserrat et al., 2013; Shu et al., 2013). Therefore, we propose that genes involved in lineage specification belong to the SCC of the parental cell subpopulation, and that they exhibit a significantly unbalanced gene expression pattern in the daughter cell subpopulations in comparison to the parental cell subpopulation. Finally, we assume that lineage specifiers for one daughter cell subpopulation should be differentially active in comparison to the other daughter cell subpopulation.

2.2. Single-cell gene expression data processing

The single-cell gene expression datasets for mouse ICM differentiation (Guo et al., 2010), HSC differentiation (Guo et al., 2013) and lung BP differentiation (Treutlein et al., 2014) were obtained from Gene Expression Omnibus (GEO). Transcription factors/regulators (TFs) annotated at (http://www.bioguo.org/AnimalTFDB/) (Zhang et al., 2012) were extracted from these datasets, resulting in around 26, 55 and 900 total TFs, respectively. In the first two RT-PCR datasets the normalized C_T values were converted into gene expression values by applying a base 2 exponential transformation as described in (Schmittgen and Livak, 2008). For the third dataset, the FPKM values were used and the missing values were imputed with the lowest expression value. We used the same single-cell sample classes as in the respective datasets. The ICM, PE and EPI subpopulations were unbiasedly classified by principle component analysis (PCA) (Guo et al., 2010), the HSC, MPP, CMP, MEP, GMP and CLP subpopulations were classified by combinations of surface markers (Guo et al., 2013), and the BP, AT1 and AT2 subpopulations were classified by PCA (Treutlein et al., 2014).

2.3. Gene expression booleanization

For Booleanization of the gene expression data, we compared the significance of the expression of each gene in each subpopulation against the background distribution formed by the union of the expression values of all cell subpopulations that co-exist at a given moment. For example, the ICM and trophoectoderm (TE) cell subpopulations co-exist in the 32-cell stage cells and therefore the expression of ICM genes was compared against the background expression formed by both ICM and TE cells. Similarly, the Booleanization of the gene expression of PE and EPI was performed against the background expression formed by all 64-cell stage cells (i.e., PE, EPI and TE (64C)). The six subpopulations of the HSC dataset co-exist in the mouse bone marrow, therefore the background expression was formed by combining all the six subpopulations. The BP, AT1 and AT2 cell subpopulations also co-exist at embryonic day 18.5 and the background expression was formed by combining all these three subpopulations. Since the gene expression values did not follow a normal distribution, the significance p-value of a gene against the background expression was non-parametrically computed using the one-sided Mann–Whitney–Wilcoxon test. The cutoff of p-value ≤ 0.4 was set, below which the expression of a gene was considered differentially active "1", and otherwise "0" (i.e., not significantly differentially active) in a Boolean manner. This significance threshold was empirically determined based on several marker genes whose expression states are well-known to be active in certain subpopulations. The Booleanized expression data are available in Tables S1–S3.

2.4. TRN reconstruction

1. Network inference from literature knowledge: The information about experimentally validated interactions among TFs was retrieved from the MetaCore™ server (Nikolsky et al., 2005). The interaction types "Transcriptional regulation" and "Binding" were selected. These data include the information on the directionality of the interactions and its mode of action (i.e., activation or inhibition, or unspecified

otherwise). This set of interactions included distal element-mediated interactions.

2. Network inference by TF-DNA binding-site prediction: The prediction of TF-DNA binding-site was carried out using the MATCH tool (Kel et al., 2003). The information regarding the transcription start sites (TSSs) was obtained from the RefSeq database (Pruitt et al., 2014). Promoter sequences comprising 2000 base pairs upstream and 1000 base pairs downstream from TSSs were obtained using twobitToFa utility and 0.2 bit genome sequence files (hg19, mm10) from UCSC (http://hgdownload.soe.ucsc.edu/downloads.html).
3. Network inference from single-cell gene expression data: Single-cell gene expression data allow us to infer more realistic co-expression relationships between genes, which can significantly increase the reliability of network inference (Moignard and Gottgens, 2014; Luo et al., 2015). Since the gene expression patterns between gene pairs were not following a normal distribution, mutual information was used as a statistical metric since it makes no assumption about the underlying statistical distribution. For this purpose, we used MRNET (Meyer et al., 2007) implemented in R (Meyer et al., 2008), which employs the maximum relevance–minimum redundancy. Next, we filtered out weakly inferred interactions by computing the null distribution of interaction strengths. To do this, a randomized expression matrix of a given cell subpopulation was made by randomly shuffling the gene expression values of each single-cell sample and the same MRNET inference was performed on this randomized expression matrix for inferring the interactions corresponding to the null distribution. This procedure was repeated 10,000 times and a significance p-value for each interaction was computed against the interaction strength inferred from these randomized expression matrices. The p-values were then ranked and the interactions among the top half of the total interactions were considered as putative interactions. Note, we did not set a strict cutoff at this stage, as these interactions will be further filtered in the subsequent procedures.

These three different sources of interactions were combined to reconstruct raw TRNs. To this end, we took the union of the intersection between 1 and 3, and the intersection between 2 and 3 (Fig. S1). The rationale behind this approach is that: i) the interactions in 1 and 2 are not cell-type specific and mainly come from bulk data that may contain a heterogeneous population of cells, therefore 3 can add cell-subpopulation specificity to each TRN, and ii) the interactions in 1 are already reported interactions, therefore adding the intersection between 2 and 3 could add novel, proximal element (promoter)-mediated transcriptional interactions that were supported by two different information sources. Although we could not add novel interactions mediated by distal elements, to our knowledge, no comprehensive approach for accurately linking distal elements with regulated genes is available.

Next, raw TRNs were contextualized (i.e., pruned) to the Booleanized gene expression profiles of each cell subpopulation using a method previously developed in our group (Crespo et al., 2013), which was re-implemented and modified in MATLAB using a genetic algorithm (GA) function. This algorithm assumes that each cellular phenotype is a stable steady state attractor of a Boolean network, and removes edges that are inconsistent with the Booleanized gene expression data. The Boolean simulation was carried out using the pbn-matlab-toolbox (http://code.google.com/p/pbn-matlab-toolbox/downloads/list) with a synchronous updating scheme. A logic rule was defined for genes receiving multiple interactions, such that the number of activating and inhibiting edges acting on a gene were compared and the one with a higher number dominates (i.e., the threshold rule with all node/edge weights being 1). If the numbers of activations and inhibitions are equal, the state of the target gene is set to remain in its current state. During this process, "unassigned" interactions – i.e., interactions for which the effect of activation or inhibition is unknown – were randomly assigned "activation" or "inhibition" and the GA was used to find the best solutions for explaining the gene expression profile. We did not incorporate specific, experimentally validated interaction logic rules into our workflow in order to keep the method applicable to systems without any prior logic knowledge. The optimization function was designed to minimize mismatches between the simulated Boolean attractor and gene expression data. When a gene has more than 30 incoming edges, the number was reduced to 29 by randomly removing the incoming edges to reduce the computational load. For the BP system, the contextualization was performed on the SCC of the entire raw TRN, as the genetic algorithm could not converge when the entire TRN was used due to its size. The contextualized network was visualized in Cytoscape (version 2.7.0) (Shannon et al., 2003).

2.5. Prediction of opposing lineage specifier pairs

As mentioned above, our model of stem cell differentiation assumes that lineage specifiers reside in autonomous network stability cores, namely SCCs, and exhibit a significantly unbalanced gene expression pattern in the daughter cell subpopulation in comparison to the parental cell subpopulation. This change in the expression ratios between pairs of TFs during differentiation was quantified by the following metric:

$$\frac{\left(\frac{E_{TF1}^{\text{Parent}}}{E_{TF2}^{\text{Parent}}}-\frac{E_{TF1}^{\text{Daughter}}}{E_{TF2}^{\text{Daughter}}}\right)}{\left(\frac{E_{TF1}^{\text{Parent}}}{E_{TF2}^{\text{Parent}}}\right)}$$

where E_{TF1}^{Parent}, E_{TF2}^{Parent}, $E_{TF1}^{Daughter}$ and $E_{TF2}^{Daughter}$ are the expression values of TF1 and TF2 in the parental cell subpopulation and a daughter cell subpopulation, respectively. The formula says that for a given TF pair, the expression ratio difference between the parental and daughter cell subpopulations (numerator) is normalized by the same ratio for the parental subpopulation (denominator). The normalization is necessary since here we aim to quantify the ratio change upon differentiation from the parental- to daughter cell subpopulations with respect to the initial parental ratio. This value was calculated for all pairs of TFs in each combination of a randomly selected parental single cell and a daughter single cell. This analysis was conducted for both differentiation paths in each binary-fate differentiation event. For each pair of TFs, the median value across all combinations (no replacement) of a parental single cell and a daughter single cell was computed. Then, the difference in the median values between the two opposing lineages was calculated. We call this final value the "median expression ratio unbalance index" throughout this study.

Next, we filtered out opposing lineage specifier pairs whose median expression ratio unbalance index was not significantly high. To this end, we computed the null distribution of median expression ratio unbalance indices from randomly shuffled values with 1500 iterations. The significance p-value for each TF pair was computed against this null distribution. The p-value was then corrected for multiple testing using the Benjamini–Hochberg method. The false discovery rate cutoff was set to 0.05. In addition, all TF pairs with the median expression ratio unbalance index below 1 were discarded. Among the remaining TF pairs, we only kept those, whose Booleanized states of TF1 and TF2 were "1" and "0" in one daughter cell subpopulation but "0" and "1" (i.e., the opposite state) in the other daughter cell subpopulation, respectively, with more than 2-fold change in the expression value. If both TFs of these selected opposing lineage specifier pairs were present in the SCC of the parental cell subpopulation and if they were directly connected to each other (i.e., likely exerting their influence on themselves), they were considered the final predicted opposing lineage specifier pairs.

3. Results

Based on the proposed model above, we implemented a computational method for predicting opposing lineage specifier pairs of stem cell differentiation. The schematic view of the method is shown in (Fig. 1). Briefly, we first generated a raw TRN for each parental cell subpopulation by combining literature-based interactions, predicted TF-DNA binding interactions, and single-cell gene co-expression-based interactions (Fig. S1). We then performed network contextualization of raw TRNs using an improved version of the method developed in our group (Crespo et al., 2013), which removes interactions that are inconsistent with Booleanized gene expression profiles. Once contextualized TRNs were reconstructed, candidate opposing lineage specifier pairs were predicted based on their expression values, and their presence in the SCC of the parental cell subpopulation. We propose that the change in expression ratio between parental and daughter cell subpopulations is biologically more relevant than the expression ratio itself within each cell subpopulation, since the basal/effective level of

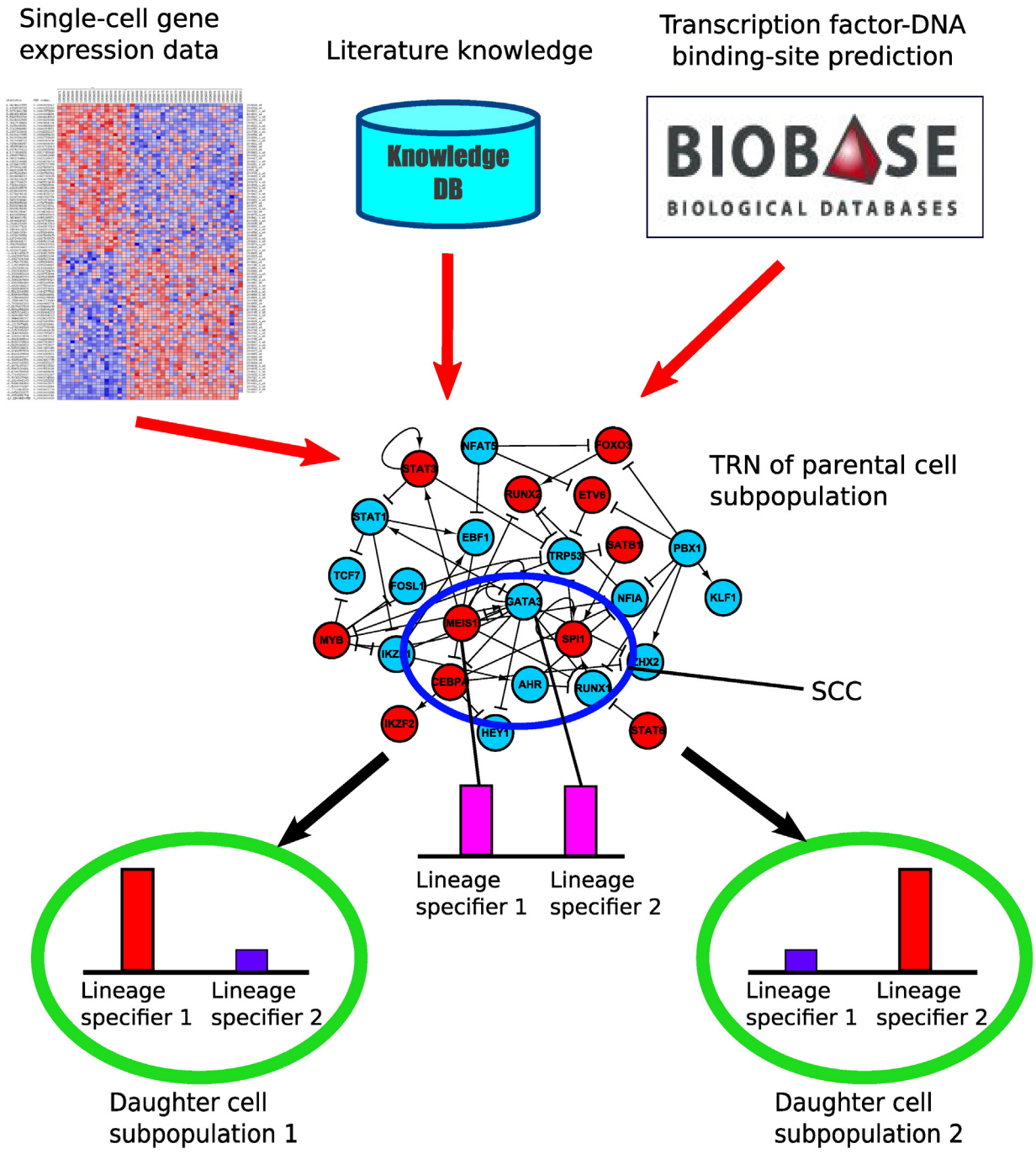

Fig. 1. Schematic view of the proposed method for predicting opposing lineage specifier pairs of stem cell differentiation using single-cell gene expression data. Candidate lineage specifiers are identified in three steps. First, a transcriptional regulatory network (TRN) is inferred using single-cell gene expression data, literature knowledge and transcription factor (TF)-DNA binding-site prediction. This TRN is then contextualized by removing edges that are inconsistent with Booleanized gene expression data. Then, the strongly connected component (SCC) is identified in the TRN of parental cell subpopulation (see Materials and methods). In parallel, pairs of TFs that showed a significant change in expression ratio upon differentiation to two daughter cell subpopulations are identified. These TF pairs that are present in the parental SCC and whose genes are differentially active in daughter cell subpopulations are considered candidate opposing lineage specifier pairs.

expression differs among TFs due to several factors, including the binding strength and number of target genes (see Materials and methods section for complete explanation).

3.1. Application of the method for predicting opposing lineage specifier pairs

To test the applicability of our method, we selected three binary-fate stem cell differentiation systems, for which high-quality single-cell gene expression data were available. The first example (Guo et al., 2010) is the differentiation of ICM into either PE or EPI (Fig. 3A). The second example is the hematopoietic system (Guo et al., 2013), in which HSC differentiates into either MPP or MEP, MPP differentiates into CMP or CLP, and CMP differentiates into either MEP or GMP (Fig. 2). The third example is the differentiation of lung BP into either AT1 or AT2 (Treutlein et al., 2014) (Fig. 3B). The predicted opposing lineage specifier pairs and their median expression ratio unbalance index are shown in Table 1. In the first two examples our method accurately predicted experimentally validated lineage specifiers. Therefore, the method was further applied to the lung BP differentiation system, which has been studied in less detail, resulting in novel candidate lineage specifiers that could be experimentally tested. The predicted lineage specifiers in each differentiation event are discussed in detail in the following sections.

3.1.1. ICM into PE and EPI

One opposing lineage specifier pair, *Gata6–Klf2*, was predicted in this event, which is in line with the previous evidence that *Gata6* is a lineage specifier for PE (Fujikura et al., 2002) and *Klf2* is strongly implicated in the establishment of EPI (Yeo et al., 2014; Gillich et al., 2012). Therefore, our prediction recapitulated known lineage specifiers of this system without predicting false positives. However, we also missed *Gata4* and *Nanog* for the PE and EPI lineage specifications, respectively (Fujikura et al., 2002; Mitsui et al., 2003),

3.1.2. HSC into MPP or MEP

In this differentiation event five opposing lineage specifier pairs were predicted. For the MPP specification, a lymphoid lineage specifier *Gata3* (Ting et al., 1996) and a myeloid lineage specifiers *Spi1* (PU.1) (Voso et al., 1994) were predicted. The prediction of both of these lineage specifiers is reasonable since MPP has the potential for both lineages. On the other hand, a well-known lineage specifier for the megakaryocyte-erythroid lineage, *Gata1*, (Pevny et al., 1991) was predicted for MEP specification, which counteracts the lymphoid and myeloid lineages. In addition, our prediction of the *Meis1–Gata1* pair is in good agreement with the previous study that showed that *Meis1* regulates the *Gata1* activity, and controls the lineage specification between myelopoiesis and erythropoiesis (Pillay et al., 2010). Although *Stat3* is not definitively shown to specify myelolymphoid lineage, it is directly regulated by *Meis1*, and plays a key role in many cell fate decisions downstream in the myelolymphoid lineage, such as T helper cell and B cell commitment, as well as dendritic cell development and maturation (Hankey, 2009).

3.1.3. MPP into CMP or CLP

In the MPP to CMP transition, *Gata2* was predicted, which is known to be expressed in committed progenitor cells and plays a role in megakaryopoiesis (Mouthon et al., 1993; Orlic et al., 1995). In addition, the other predicted CMP lineage specifier, *Stat1*, has been shown to mediate the cell fate decision between erythropoiesis and megakaryopoiesis (Duek et al., 2014). In the MPP to CLP transition a known lymphoid lineage specifier, *Satb1*, (Satoh et al., 2013) was predicted to counteract *Gata2*. Finally, a T-lineage gene, *Est1* (Liu et al., 2010), was also predicted to specify the CLP lineage. Although the direct transition from the MPP to the CLP subpopulations was previously accepted (Akashi et al., 2000), it is now known that LMPP is a more appropriate intermediate subpopulation that gives rise to GMP and CLP. However, we could not use the LMPP subpopulation in the current study, as this subpopulation was not profiled in the study from which we obtained the dataset (Guo et al., 2013). Therefore, the relatively small number of predicted opposing lineage specifier pairs (for example, *Ikzf1*, important for lymphoid development, was missed for the CLP specification) in this differentiation system might be due to the fact that we did not consider LMPP instead of CLP.

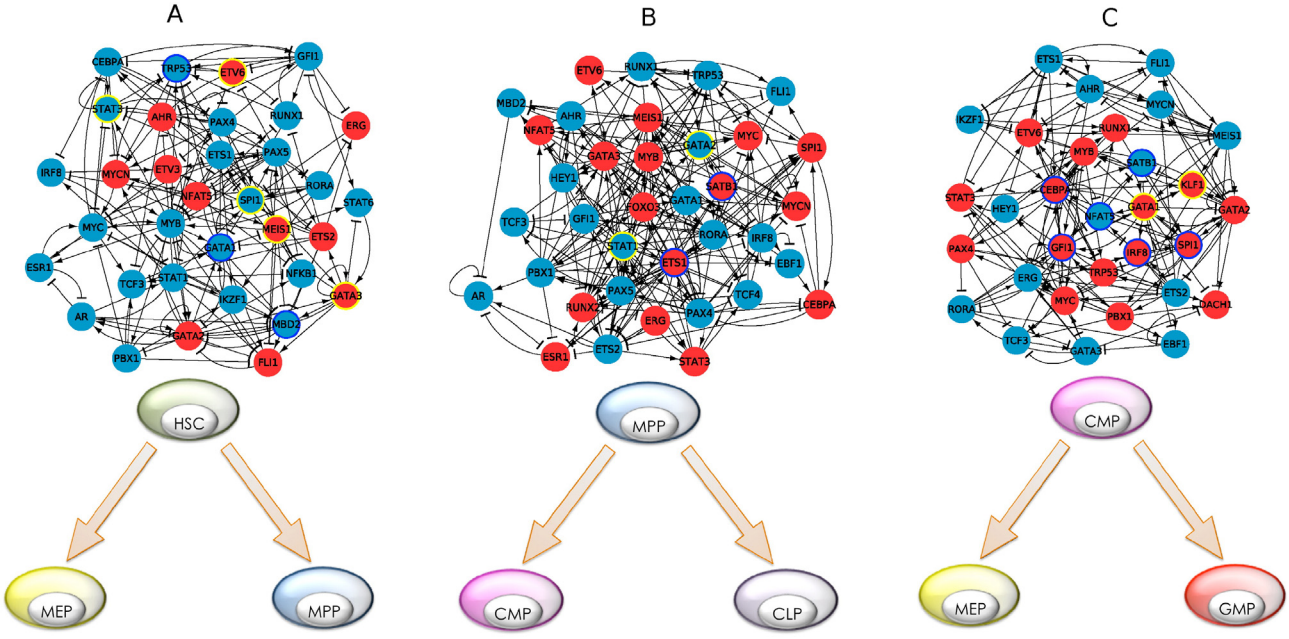

Fig. 2. SCCs of TRNs of parental cell subpopulations for three binary bifurcations during hematopoiesis. A, Differentiation of HSC into either MEP or MPP. B, Differentiation of MPP into either CMP to CLP. C, Differentiation of CMP into either MEP or GMP. Red and blue nodes indicate not over-expressed and over-expressed genes, respectively. Pointed arrows indicate activation and blunted arrows indicate inhibition.

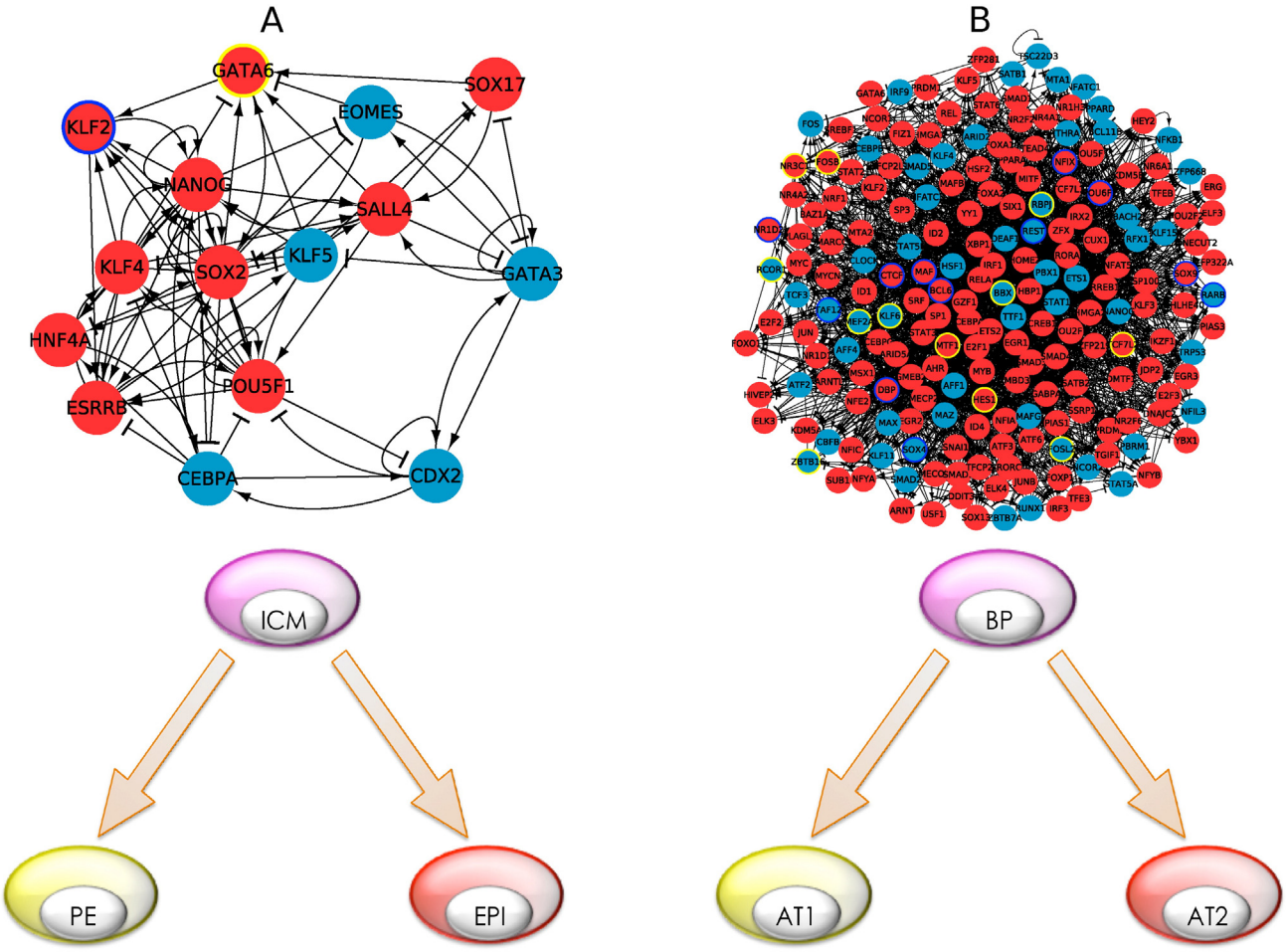

Fig. 3. SCCs of TRNs of parental cell subpopulations for binary bifurcations during early embryonic development and lung BP development. A, Differentiation of ICM into either to PE or EPI. B, Differentiation of BP into either AT1 or AT2. The graphical properties are described in Fig. 1.

3.1.4. CMP into MEP or GMP

In this bifurcation event six opposing linage specifier pairs were predicted, five of which included *Gata1* for the MEP specification, underscoring the importance of this gene for MEP specification. Furthermore, *Klf1* was also predicted as an MEP specifier, which is in accordance with the previous evidence that it is strongly involved in the establishment and maintenance of the erythroid lineage (Miller and Bieker, 1993; Siatecka et al., 2007). Some of the other lineage specifiers predicted for the HSC-to-MEP transition, such as *Mbd2* and *Trp53*, were not predicted in this event, suggesting that lineage specifiers for a same cell subpopulation vary with the initial cellular subpopulation. In the CMP to GMP transition our method predicted several known myeloid specifiers, *Cebpa* (Radomska et al., 1998), *Gfi1* (Li et al., 2010), *Irf8* (Becker et al., 2012) and *Spi1*, as lineage specifiers. In addition, lymphoid lineage specifiers, *Satb1* and *Nfat5* were predicted as GMP lineage specifiers, suggesting that these genes might also play a role in the specification of myeloid lineage from CMP.

3.1.5. BP into AT1 and AT2

Since our method recapitulated many well-known lineage specifiers in the last two examples, we applied it to the less studied, lung BP differentiation system, which resulted in 17 opposing lineage specifier pairs. As expected, not so much is known about the lineage specification of BP into AT1 or AT2, however, for AT1 specification, *Fosb* has been implicated to play a role in correct alveolar sac development (Millien et al., 2006). In addition, Notch targets *Rbpj* and *Hes1* have been implicated in mouse lung development (Ito et al., 2000; Dou et al., 2008). For AT2 specification, *Pou6f1* has been shown to be associated with lung developmental pathway (Hu et al., 2012) and D-site binding protein, *Dbp*, belongs to the bZIP protein family, and has been shown to bind to the promoter region of pulmonary surfactant *Sftpb*, (Bein et al., 2011) which is formed by AT2. Thus, several of the predicted lineage specifiers exhibited, to varying degrees, prior associations with lung development.

4. Discussion

Understanding lineage specification has been partly hampered by the co-existence of different cell subpopulations within a heterogeneous stem cell population. In the current study we have proposed a model of binary-fate stem cell differentiation, in which each parental stem cell subpopulation is in a stable state maintained by its specific TRN stability core. Furthermore, this stability core is maintained by the balanced expression pattern of opposing lineage specifiers for different daughter cell subpopulations. Dysregulation of this balanced expression pattern induces differentiation. Based on this model, we have developed a computational method for predicting opposing lineage specifier pairs for a binary-fate differentiation event. Single-cell gene expression data enabled us to reconstruct TRNs and to identify their stability cores specific for different parental cell subpopulations. Indeed, subpopulation-specific TRNs exhibited significant network rewiring, as was previously reported in Moignard et al. (2013). Using these subpopulation-specific TRNs, our method was, albeit a few false negatives such as *Gata4* and *Nanog* for PE and EPI lineage specifications, able to predict many known lineage specifiers in the two well-studied examples (Guo et al., 2010, 2013). This method was further applied to a less-studied example,

Table 1
Predicted opposing lineage specifier pairs in each binary-fate differentiation event. Each binary-fate differentiation event is indicated with a combination of parental cell subpopulation and daughter cell subpopulations. Genes in bold are known/strongly implicated lineage specifiers for that cell subpopulation.

Gene1	Gene2	Median expression ratio unbalance index
ICM differentiation (Guo et al., 2010)		
PE	ICM	
Gata6	**Klf2**	1.78
HSC differentiation (Guo et al., 2013)		
MPP	**MEP**	
Gata3	Mbd2	26.07
Meis1	**Gata1**	18.56
Etv6	Trp53	13.31
Spi1	**Gata1**	12.92
Stat3	Trp53	7.53
MPP differentiation (Guo et al., 2013)		
CMP	**CLP**	
Stat1	Ets1	5.46
Gata2	**Satb1**	2.14
CMP differentiation (Guo et al., 2013)		
MEP	**GMP**	
Gata1	Nfat5	14.06
Klf1	**Cebpa**	11.95
Gata1	**Gfi1**	8.21
Gata1	**Irf8**	8.01
Gata1	**Spi1**	6.42
Gata1	Satb1	4.18
BP differentiation (Treutlein et al., 2014)		
AT1	**AT2**	
Fosl2	Sox4	12.28
Mef2a	Nr1d2	8.26
Mtf1	Pou6f1	8.26
Mef2a	Nfix	5.66
Zbtb16	Ctcf	3.39
Klf6	Dbp	2.95
Bbx	Rest	2.76
Rcor1	Rest	2.61
Rbpj	Sox9	2.35
Nr3c1	Bcl6	2.08
Bbx	Rarb	2.02
Bbx	Maf	1.81
Tcf7l2	Sox9	1.68
Bbx	Rarb	1.31
Bbx	Maf	1.21
Fosb	Ctcf	1.05
Hes1	Taf12	1.02

the lung BP differentiation system (Treutlein et al., 2014), and predicted novel candidate lineage specifiers, several of which have been previously shown to have some association with lung development, and could be experimentally validated in future.

Importantly, our method does not require pre-selection of putative candidate genes, and can be applicable to any binary-fate differentiation system for which single-cell gene expression data are available. Furthermore, this method is compatible with both single-cell RT-PCR and single-cell RNA-seq data. Given the increasing importance of single-cell gene expression data in stem cell biology, we believe that approaches like ours would be useful for the identification of lineage specifiers and their associated network stability cores in TRNs. This should help us design cell differentiation protocols with higher efficiency and fidelity, and therefore constitute an important aid in regenerative medicine.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.scr.2015.08.006.

Disclosure of potential conflict of interest

The authors state that there are no conflicts of interest.

Acknowledgments

SO is supported by an FNR AFR Postdoctoral Grant.

References

Akashi, K., et al., 2000. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. Nature 404 (6774), 193–197.
Becker, A.M., et al., 2012. IRF-8 extinguishes neutrophil production and promotes dendritic cell lineage commitment in both myeloid and lymphoid mouse progenitors. Blood 119 (9), 2003–2012.
Bein, K., Leight, H., Leikauf, G.D., 2011. JUN-CCAAT/enhancer-binding protein complexes inhibit surfactant-associated protein B promoter activity. Am. J. Respir. Cell Mol. Biol. 45 (2), 436–444.
Buganim, Y., et al., 2012. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell 150 (6), 1209–1222.
Crespo, I., Del Sol, A., 2013. A general strategy for cellular reprogramming: the importance of transcription factor cross-repression. Stem Cells 31 (10), 2127–2135.
Crespo, I., et al., 2013. Predicting missing expression values in gene regulatory networks using a discrete logic modeling optimization guided by network stable states. Nucleic Acids Res. 41 (1), e8.
Dou, G.R., et al., 2008. RBP-J, the transcription factor downstream of Notch receptors, is essential for the maintenance of vascular homeostasis in adult mice. FASEB J. 22 (5), 1606–1617.
Duek, A., et al., 2014. Loss of Stat1 decreases megakaryopoiesis and favors erythropoiesis in a JAK2-V617F-driven mouse model of MPNs. Blood 123 (25), 3943–3950.
Ertaylan, G., et al., 2014. Gene regulatory network analysis reveals differences in site-specific cell fate determination in mammalian brain. Front. Cell. Neurosci. 8, 437.
Fujikura, J., et al., 2002. Differentiation of embryonic stem cells is induced by GATA factors. Genes Dev. 16 (7), 784–789.
Gillich, A., et al., 2012. Epiblast stem cell-based system reveals reprogramming synergy of germline factors. Cell Stem Cell 10 (4), 425–439.
Guo, G., et al., 2010. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev. Cell 18 (4), 675–685.
Guo, G., et al., 2013. Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. Cell Stem Cell 13 (4), 492–505.
Hankey, P.A., 2009. Regulation of hematopoietic cell development and function by Stat3. Front. Biosci. (Landmark Ed.) 14, 5273–5290.
Hu, P., et al., 2012. Microarray meta-analysis identifies acute lung injury biomarkers in donor lungs that predict development of primary graft failure in recipients. PLoS One 7 (10), e45506.
Ito, T., et al., 2000. Basic helix–loop–helix transcription factors regulate the neuroendocrine differentiation of fetal mouse pulmonary epithelium. Development 127 (18), 3913–3921.
Kel, A.E., et al., 2003. MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. 31 (13), 3576–3579.
Li, V.C., Kirschner, M.W., 2014. Molecular ties between the cell cycle and differentiation in embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. 111 (26), 9503–9508.
Li, H., et al., 2010. Repression of Id2 expression by Gfi-1 is required for B-cell and myeloid development. Blood 116 (7), 1060–1069.
Liu, P., Li, P., Burke, S., 2010. Critical roles of Bcl11b in T-cell development and maintenance of T-cell identity. Immunol. Rev. 238 (1), 138–149.
Luo, Y., et al., 2015. Single-cell transcriptome analyses reveal signals to activate dormant neural stem cells. Cell 161 (5), 1175–1186.
Meyer, P.E., et al., 2007. Information-theoretic inference of large transcriptional regulatory networks. EURASIP J. Bioinform. Syst. Biol. 79879.
Meyer, P.E., Lafitte, F., Bontempi, G., 2008. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinf. 9, 461.
Miller, I.J., Bieker, J.J., 1993. A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Kruppel family of nuclear proteins. Mol. Cell. Biol. 13 (5), 2776–2786.
Millien, G., et al., 2006. Alterations in gene expression in T1 alpha null lung: a model of deficient alveolar sac development. BMC Dev. Biol. 6, 35.
Mitsui, K., et al., 2003. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell 113 (5), 631–642.
Moignard, V., Gottgens, B., 2014. Transcriptional mechanisms of cell fate decisions revealed by single cell expression profiling. Bioessays 36 (4), 419–426.
Moignard, V., et al., 2013. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. Nat. Cell Biol. 15 (4), 363–372.
Moignard, V., et al., 2015. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat. Biotechnol. 33 (3), 269–276.
Montserrat, N., et al., 2013. Reprogramming of human fibroblasts to pluripotency with lineage specifiers. Cell Stem Cell 13 (3), 341–350.
Mouthon, M.A., et al., 1993. Expression of tal-1 and GATA-binding proteins during human hematopoiesis. Blood 81 (3), 647–655.
Nikolsky, Y., et al., 2005. A novel method for generation of signature networks as biomarkers from complex high throughput data. Toxicol. Lett. 158 (1), 20–29.
Orlic, D., et al., 1995. Pluripotent hematopoietic stem cells contain high levels of mRNA for c-kit, GATA-2, p45 NF-E2, and c-myb and low levels or no mRNA for c-fms and the receptors for granulocyte colony-stimulating factor and interleukins 5 and 7. Proc. Natl. Acad. Sci. U. S. A. 92 (10), 4601–4605.
Pevny, L., et al., 1991. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. Nature 349 (6306), 257–260.

Pillay, L.M., et al., 2010. The Hox cofactors Meis1 and Pbx act upstream of gata1 to regulate primitive hematopoiesis. Dev. Biol. 340 (2), 306–317.
Pruitt, K.D., et al., 2014. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 42 (Database issue), D756–D763.
Radomska, H.S., et al., 1998. CCAAT/enhancer binding protein alpha is a regulatory switch sufficient for induction of granulocytic development from bipotential myeloid progenitors. Mol. Cell. Biol. 18 (7), 4301–4314.
Sandbo, N., et al., 2009. Critical role of serum response factor in pulmonary myofibroblast differentiation induced by TGF-beta. Am. J. Respir. Cell Mol. Biol. 41 (3), 332–338.
Satoh, Y., et al., 2013. The Satb1 protein directs hematopoietic stem cell differentiation toward lymphoid lineages. Immunity 38 (6), 1105–1115.
Schmittgen, T.D., Livak, K.J., 2008. Analyzing real-time PCR data by the comparative C(T) method. Nat. Protoc. 3 (6), 1101–1108.
Shannon, P., et al., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13 (11), 2498–2504.
Shu, J., et al., 2013. Induction of pluripotency in mouse somatic cells with lineage specifiers. Cell 153 (5), 963–975.
Siatecka, M., Xue, L., Bieker, J.J., 2007. Sumoylation of EKLF promotes transcriptional repression and is involved in inhibition of megakaryopoiesis. Mol. Cell. Biol. 27 (24), 8547–8560.
Tang, F., et al., 2010. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell 6 (5), 468–478.
Ting, C.N., et al., 1996. Transcription factor GATA-3 is required for development of the T-cell lineage. Nature 384 (6608), 474–478.
Treutlein, B., et al., 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509 (7500), 371–375.
Voso, M.T., et al., 1994. Inhibition of hematopoiesis by competitive binding of transcription factor PU.1. Proc. Natl. Acad. Sci. U. S. A. 91 (17), 7932–7936.
Xu, H., et al., 2014. Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. PLoS Comput. Biol. 10 (8), e1003777.
Yeo, J.C., et al., 2014. Klf2 is an essential factor that sustains ground state pluripotency. Cell Stem Cell 14 (6), 864–872.
Zhang, H.M., et al., 2012. AnimalTFDB: a comprehensive animal transcription factor database. Nucleic Acids Res. 40 (Database issue), D144–D149.