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.

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

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 $p\text{-value} \leq 0.4$ 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