A non-linear reverse-engineering method for inferring genetic regulatory networks

Siyuan Wu[1], Tiangang Cui[1], Xinan Zhang[2] and Tianhai Tian[1]

[1] School of Mathematics, Monash University, Clayton, VIC, Australia
[2] School of Mathematics and Statistics, Central China Normal University, Wuhan, PR China

ABSTRACT

Hematopoiesis is a highly complex developmental process that produces various types of blood cells. This process is regulated by different genetic networks that control the proliferation, differentiation, and maturation of hematopoietic stem cells (HSCs). Although substantial progress has been made for understanding hematopoiesis, the detailed regulatory mechanisms for the fate determination of HSCs are still unraveled. In this study, we propose a novel approach to infer the detailed regulatory mechanisms. This work is designed to develop a mathematical framework that is able to realize nonlinear gene expression dynamics accurately. In particular, we intended to investigate the effect of possible protein heterodimers and/or synergistic effect in genetic regulation. This approach includes the Extended Forward Search Algorithm to infer network structure (top-down approach) and a non-linear mathematical model to infer dynamical property (bottom-up approach). Based on the published experimental data, we study two regulatory networks of 11 genes for regulating the erythrocyte differentiation pathway and the neutrophil differentiation pathway. The proposed algorithm is first applied to predict the network topologies among 11 genes and 55 non-linear terms which may be for heterodimers and/or synergistic effect. Then, the unknown model parameters are estimated by fitting simulations to the expression data of two different differentiation pathways. In addition, the edge deletion test is conducted to remove possible insignificant regulations from the inferred networks. Furthermore, the robustness property of the mathematical model is employed as an additional criterion to choose better network reconstruction results. Our simulation results successfully realized experimental data for two different differentiation pathways, which suggests that the proposed approach is an effective method to infer the topological structure and dynamic property of genetic regulations.

INTRODUCTION

Hematopoiesis is a highly complex process that controls the proliferation, differentiation and maturation of hematopoietic stem cells (HSCs) (*Ng & Alexander, 2017*). It has been widely accepted that genetic regulatory networks control the developmental processes of various types of blood cells (*Cedar & Bergman, 2011*). Although the regulatory mechanisms have been studied over a century, there are still many challenging questions

regarding the cell fate determination in hematopoiesis (*Ottersbach et al., 2010*). Thus, it is imperative to unravel the regulatory mechanisms for the study of hematopoiesis.

In adult mammals, hematopoiesis occurs mostly in the bone marrow (*Birbrair & Frenette, 2016*). HSCs have the feature of self-renewal and multipotent as well as the ability to differentiate into multipotent progenitors (MPPs). Then, MPPs will differentiate into two main lineages of blood cells, namely the myeloid lineage which starts at common myeloid progenitors (CMPs) and the lymphoid lineage which starts at common lymphoid progenitors (CLPs). In addition, the myeloid lineage has two distinct progenitors, namely megakaryocyte-erythroid progenitors (MEPs) and granulocyte-macrophage progenitors (GMPs). MEPs can differentiate into megakaryocytes and erythrocytes, and GMPs can give rise to mast cells, macrophages and granulocytes. Lymphoid lineage cells include T lymphocytes (T-cells), B lymphocytes (B-cells) and natural killer cells (NK-cells) (*Orkin & Zon, 2008*). In this work, we focus on the fate determination of HSCs in the myeloid lineage for the choice between erythrocytes and neutrophils.

During the developmental process, a number of transcriptional factors (TFs) act as regulators to control the fate determination of HSCs (*Aggarwal et al., 2012*). Among them, the genetic complex Gata1-Gata2-PU.1 is a very important module for the cell-fate choice of CMPs between erythrocytes or neutrophils (*Friedman, 2007*; *Liew et al., 2006*; *Ling et al., 2004*). In particular, the Gata1-PU.1 complex forms a double negative feedback module, in which each gene inhibits the expression of the other (*Friedman, 2007*). Recently it has been elucidated that the fate determination of HSCs was defined not only by the ratio of Gata1 and PU.1 (*Hoppe et al., 2016*), but also by a third party during the regulation. For example, *FOG-1* is a significant third party to regulate the Gata1-PU.1 module (*Chang et al., 2002*; *Mancini et al., 2012*). Erythropoietin receptor (EpoR) signaling also acts the essential role in regulating the Gata1-PU.1 Module (*Zhao et al., 2006*). Although the regulatory mechanisms of the Gata1-Gata2-PU.1 complex in hematopoiesis are relatively well studied, the connection of this triad complex with other significant genes as well as the role of these genes in hematopoiesis are mostly unclear (*Liew et al., 2006*).

Mathematical modeling is an important method for inferring the detailed regulatory mechanisms. In 1910, Archibald Hill proposed a classical non-linear ordinary differential equation (ODE or Hill equation) model to describe the sigmoidal oxygen binding curve of hemoglobin (*Hill, 1910*). Since then, the Hill equation has been applied to explore the mechanisms in a wide range of genetic regulatory networks and biological systems. For example, the genetic toggle switching was achieved by the models with Hill equations (*Gardner, Cantor & Collins, 2000*). In addition, the Hill equation was employed to formalize the mechanisms of cell fate determination (*Xiong & Ferrell, 2003*; *Laslo et al., 2006*; *Huang et al., 2007*). Recently, the Hill equation was also used to discover a regulatory network of 52 genes with the uniform activation and repression strengthes (*Li & Wang, 2013*). Another widely used approach is the Shea–Ackers formalism for studying the thermodynamics of regulatory networks (*Shea & Ackers, 1985*). We developed a mathematical model based on the Shea–Ackers formalism to study the regulations of the Gata1-Gata2-PU.1 complex (*Tian & Smith-Miles, 2014*). A stochastic model was also

proposed to explore the function of noise in regulating the fate determination of HSCs. Simulations suggested that fluctuations of protein numbers may lead the HSC to different developmental pathways. In recent years substantial process has been made to design various types of mathematical models for describing the regulatory mechanisms of gene networks, including stochastic differential equations, stochastic kinetic systems, qualitative differential equations, Michaelis–Menten formalism, S-system and power-law formalism (*de Jong, 2002*; *Liu & Wang, 2008*; *Wang & Tian, 2010*; *Maetschke et al., 2013*; *Woods et al., 2016*; *Olariu & Peterson, 2018*; *Yang et al., 2018*; *Yang & Bao, 2019*). In particular, a number of mathematical models have been designed to realize the stable states of gene expression levels in the differentiation of HSCs (*Chang et al., 2006*; *Huang et al., 2007*; *Chickarmane, Enver & Peterson, 2009*; *Olariu & Peterson, 2018*). However, the majority of these models only considered the functions of each gene independently, namely variable x_i for the expression level of gene i in the model is in the form of $\sum_i a_i x_i$. Nonetheless, this type of models fails to represent the co-operation functions of genes together. There is a lack of investigations for the effect of possible protein heterodimers and/or synergistic effect in genetic regulations, namely variables x_i in the model are also in the form of $\sum_{i,j} b_{ij} x_i x_j$. Most recently, single-cell studies have been conducted to explore the hematopoietic system. Compared with the analysis of bulk cells, the advantage of single-cell analysis is the ability to understand the heterogeneity within the cell population (*Guo et al., 2010*; *Ye, Huang & Guo, 2017*). With the development of single-cell analysis, researchers have raised more novel computational and statistical methods to explore the regulatory mechanism of hematopoiesis. For example, the partial correlation method, Boolean model and ODE model were employed to construct the genetic regulatory networks from the single-cell expression profiles (*Hamey et al., 2017*; *Wei et al., 2017*). In addition, a deep learning method was applied to unravel the fate decision in hematopoiesis (*Athanasiadis et al., 2017*).

Recently, we proposed a general approach that combines both top-down and bottom-up approaches to reconstruct the genetic regulatory networks of the fate choice between erythrocytes and neutrophils (*Wu, Cui & Tian, 2018*). The key issue in this work includes a large number of unknown parameters and a high computational cost to add potential regulations. For the issue of parameter number, a linear ODE model may have the least number of unknown parameters among the models for all possible regulations between genes. However, since the linear model is limited to describe the linear relationship, it is not appropriate to use the linear model to study systems with complex non-linear dynamics. Although the non-linearity has been addressed by the reverse-engineering methods with the cost of more unknown parameters (*Chickarmane & Peterson, 2008*; *Crombach et al., 2012*; *Meister et al., 2013*; *Li & Wang, 2013*; *Wang et al., 2016*), the issue of protein heterodimers and/or synergistic effect between genes has not been discussed in the majority of literature at all. This work is designed to address these issues by proposing a novel approach for reconstructing genetic regulatory networks. The first innovation of this approach is the new non-linear ODE model as the bottom-up approach to study the effect of protein heterodimers and/or synergistic effect explicitly. The second innovation of this work is the proposed Extended Forward Search Algorithm as the

top-down approach to infer the structure of networks in our newly proposed non-linear model. The proposed approach thus is able to not only reduce the complex structure of genetic regulatory networks but also improve the inference efficiency substantially because the number of parameters in the mathematical model is decreased. We examined the capability of our proposed method by studying the genetic regulatory networks for the fate determination of HSCs.

METHODS

Experimental data

In this work, we used the sub-series GSE49987 as the experimental data from the published microarray dataset GSE49991 (*May et al., 2013*). This dataset contains the expression profiles collected by experiments using the cell line FDCPmix. This dataset was generated with the probe name version of Agilent Whole Mouse Genome Microarray 4 × 44 K (*May et al., 2013*). It provides microarray gene expression profiles of hematopoietic stem cells (HSCs) differentiating into erythrocytes and neutrophils. This microarray dataset is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49991. To convert all microarray probe IDs to gene names, we pre-processed this dataset based on the Ensembl BioMart and GO Enrichment Analysis (*The Gene Ontology Consortium, 2017*). From a previous study, the regulatory network of 18 core genes during the hematopoiesis has been curated (*Moignard et al., 2013*). Moreover, the same research team studied the regulatory interaction of 26 core genes during the hematopoiesis (*Moignard et al., 2015*). The total number of distinct genes in these two studies is 30. Thus, in our work we considered 30 genes whose names are listed in Table S1. There are three repeated experiments for each developmental process, each of which contains the expression levels of 30 genes from HSCs to differentiated cells at 30 time points spanning over 1 week. The observation time points are those starting from the HSCs/progenitors stage (1 point), then every 2 h over the first day (12 points), every 3 h over the second day (8 points), every 4 h over the third day (6 points), every 24 h until the fifth day (2 points), and the seventh day (1 point). In this study, we used the average data of these three repeated tests as the experimental data for each time point. The time points and expression data of four genes can be found in the following Figs. 1 and 2.

Selection of candidate genes

Based on our research experience (*Wang et al., 2016*), it is challenging to study a dynamic network with 30 genes. Thus, we conducted an extensive literature review for selecting a smaller number of important genes based on their relationship with the three genes *Gata1, Gata2* and *PU.1*. These candidate genes should be essential for the cell-fate choice in hematopoiesis, or they significantly interact with these three genes. For example, gene *Scl/Tal1* interacts with *Gata1, Eto2/Cbfa2t3* and *Ldb1* (*Goardon et al., 2006*), and is a regulator in the differentiation of hematopoietic stem cells (HSCs) (*Shivdasani, Mayer & Orkin, 1995*; *Zhang et al., 2005*; *Porcher et al., 1996*; *Real et al., 2012*). In addition,
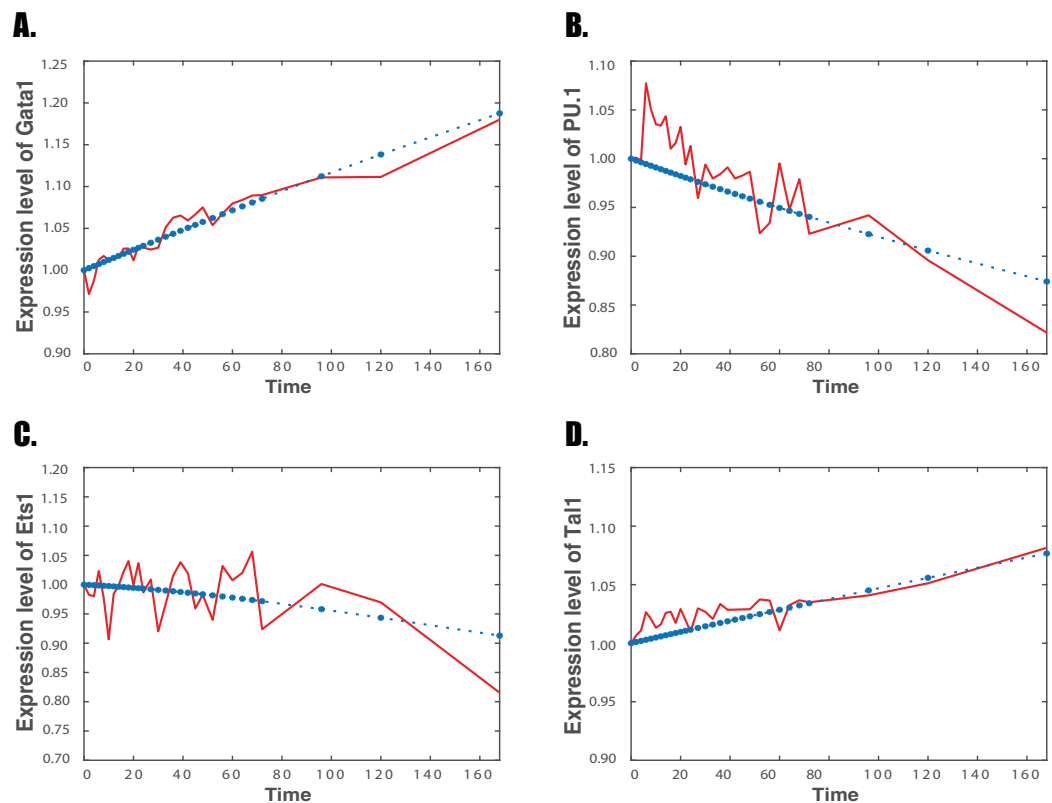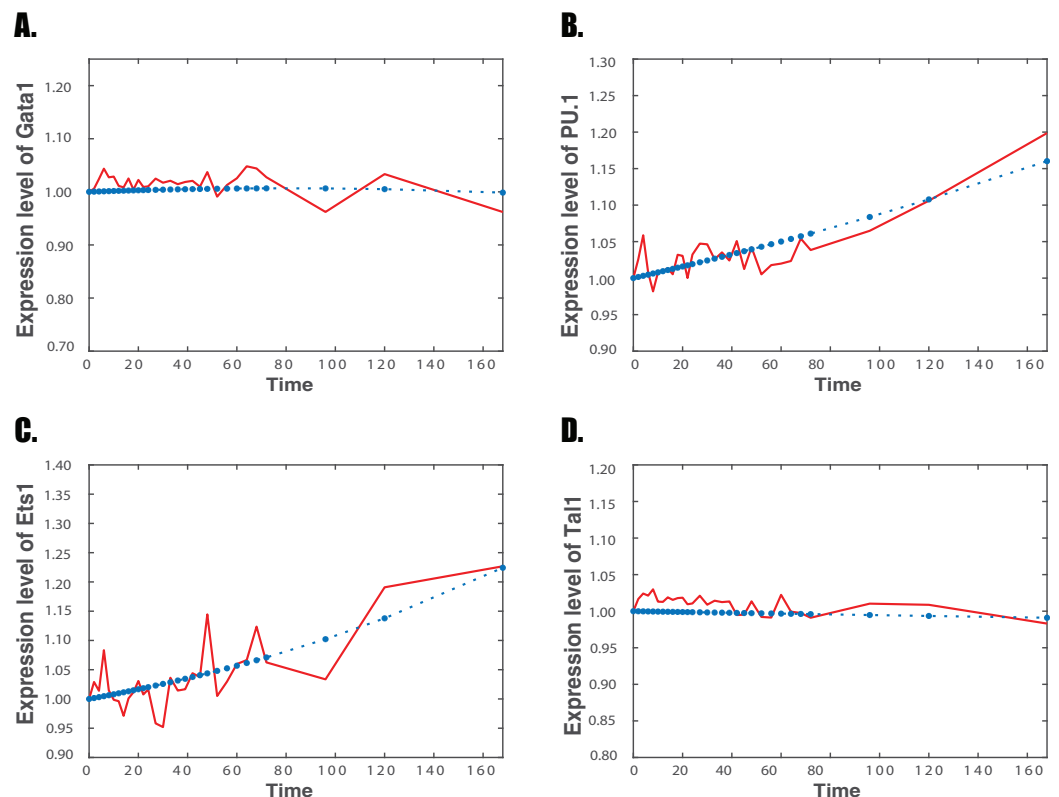
Figure 1 Simulation results and experimental data of the regulatory network for erythrocyte differentiation Red solid line: experimental microarray data; Blue star dash line: simulation of the regulatory network. (A) Gene *Gata1*; (B) gene *PU.1*; (C) gene *Ets1*; (D) gene *Tal1*.

Full-size 🖼 DOI: 10.7717/peerj.9065/fig-1

Eto2/Cbfa2t3 regulates the differentiation of HSCs by repressing the expression of target gene *Scl/Tal1* (*Goardon et al., 2006*). Moreover, *Ldb1* is a significant transcriptional factor (TF) for the differentiation of erythroid lineage (*Soler et al., 2010*). According to the ChIPSeq analysis, *Ldb1* is necessary for HSCs to control their maintenance since it binds to the majority of enhancer elements in hematopoiesis (*Li et al., 2011*).

We also included a number of genes with potential regulatory relationship with the three genes *Gata1, Gata2* and *PU.1*. For example, it was indicated that there might be unclear regulations between *Gata2* and *Gfi1* (*Moignard et al., 2013*). *Gfi1* is an important TF in the regulation of HSCs differentiation (*van der Meer, Jansen & van der Reijden, 2010*; *Lancrin et al., 2012*). *Gfi1* is required for the differentiation of common lymphoid progenitors (CLPs) and common myeloid progenitors (CMPs) from HSCs and exists in the majority of HSCs, CLPs and CMPs. Similar to gene *Gfi1*, gene *Runx1* is also expressed in most HSCs and progenitor cells as well. Then, *Gfi1* and/or *Runx1* are expressed continually in most cells which differentiate into the granulocyte lineage (*North et al., 2004*). *Lmo2* is a master regulator of hematopoiesis (*Inouea et al., 2013*). However, its specific role in regulation is still unclear. Experimental studies suggested that the knockdown of *Lmo2* does not affect the expression of *Gata1* and *Scl/Tal1* (*Inouea et al., 2013*). However, the overexpression of *Lmo2* gene also inhibited erythroid

Figure 2 Simulation results and experimental data of the regulatory network for neutrophil differentiation Red solid line: experimental microarray data; Blue star dash line: simulation of the regulatory network. (A) Gene *Gata1*; (B) gene *PU.1*; (C) gene *Ets1*; (D) gene *Tal1*.

Full-size 🖼 DOI: 10.7717/peerj.9065/fig-2

differentiation (*Visvader et al., 1997*). In addition, gene *Ets1* is a suppressor in the erythrocyte differentiation. It is downregulated in erythrocyte differentiation by binding to and activating the Gata2 promoter (*Lulli et al., 2006*). The last candidate gene is *Notch1* that inhibits the differentiation of granulocyte lineage by maintaining the expression of gene *Gata2*. It also enhances the HSCs differentiate to CLPs (*Kumano et al., 2001*; *Stier et al., 2002*). Therefore, in this study we considered the regulatory networks with the following 11 genes: *Gata1, Gata2, PU.1/Sfpi1, Runx1, Eto2/Cbfa2t3, Ets1, Notch1, Scl/Tal1, Ldb1, Gfi1* and *Lmo2*. The detailed information of the references for these 11 genes is also given in Table S2 in Supplemental Information.

Top-down approach: extended forward search algorithm

To reduce the number of unknown parameters in our proposed mathematical model, we used the probabilistic graphical models as the top-down approach to infer the topological structure of gene regulatory networks. Probabilistic graphical model is a useful tool for inferring the network structure (*Noor et al., 2013*). One type of probabilistic graphical models is the Gaussian graphical model (GGM), which provides a simple and effective method to characterize the regulatory relationship between genes. The GGM is based on the calculation of the conditional dependencies among genes using the gene

expression data. The edge connecting two genes in the model is neglected if they are conditionally independent given all other genes (*Krämer, Schäfer & Boulesteix, 2009*). In this work it is assumed that a system includes genes $\{G_1, \ldots, G_m\}$ with expression levels x_{ij} for gene G_i at time point j. Compared with the existing methods that study networks with genes only, this work will study gene networks that include not only genes in the form of monomers $\{G_1, \ldots, G_m\}$, which are represented by the linear terms in the model, but also protein heterodimers and/or synergistic effect $\{G_k–G_l\}$ ($k, l = 1,\ldots,m$), which are represented by the non-linear terms (NLTs) in the model. There are two reasons for using the NLTs $\{G_k–G_l\}$. Firstly, we can use the product of two variables to represent the synergistic effect of these two genes. Secondly, if the NLT represents the protein heterodimer, we assumed that the binding and disassociation reactions for the heterodimer $\{G_k–G_l\}$ reach an equilibrium state quickly. Thus the level of the heterodimer $\{G_k–G_l\}$ can be written as $C_{kl} \times G_k \times G_l$, where C_{kl} is the equilibrium constant. We can consider this constant C_{kl} as a coefficient in our mathematical model. In both cases, we only need to consider the product of the expression levels of these two genes, namely $y_{klj} = x_{kj}x_{lj}$, as the level of NLT $\{G_k–G_l\}$ at time t_j for our algorithm computation. Since the number of possible regulations from NLTs to genes is much larger than that of possible regulations among genes (i.e. 726 vs 110), the regulations from NLTs to genes will dominate the whole genetic regulatory system with high probability. However, the regulations among genes should be the core mechanisms rather than the regulations from NLTs to genes. To avoid the dominance of NLTs regulations, we assume that the number of regulations from NLTs to genes does not exceed that between genes.

According to the GGM (*Wang, Myklebost & Hovig, 2003*; *Wang et al., 2016*), we proposed a new algorithm, named Extended Forward Search Algorithm (EFSA), to infer the topological structure of regulatory networks that includes both genes and NLTs. Let $\mathbf{X} = (x_1, x_2, \ldots, x_N)$ be a vector that consists of m genes and n NLTs ($N = m + n$). The following three matrices are constructed, namely a $m \times m$ covariance matrix $\mathbf{A}$ of m genes, a $m \times n$ covariance matrix $\mathbf{B}$ to measure the covariance between m genes and n NLTs, and a $n \times n$ covariance matrix $\mathbf{C}$ of n NLTs. The N-dimensional matrix $\mathbf{M}$ is defined by

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{C} \end{bmatrix}, \tag{1}$$

where $\mathbf{B}'$ is the transpose of $\mathbf{B}$. An initial empty graph $\mathbf{G}$ is built by the N-dimensional identity matrix. This initial graph $\mathbf{G}$ consists of four matrices $\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3$ and $\mathbf{G}_4$ which have the same dimensions as $\mathbf{A}, \mathbf{B}, \mathbf{B}'$ and $\mathbf{C}$, respectively, namely

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{G}_2 \\ \mathbf{G}_3 & \mathbf{G}_4 \end{bmatrix}, \tag{2}$$

where $\mathbf{G}_1$ and $\mathbf{G}_4$ are identity matrix with dimensions m and n, respectively, and $\mathbf{G}_2$ and $\mathbf{G}_3$ are $m \times n$ and $n \times m$ zero matrices, respectively.

The proposed algorithm is given below.

Algorithm 1: extended forward search algorithm

1. Let $\mathbf{X} = (x_1, x_2, \ldots, x_N)$ be a vector with N elements, and N be the number of components consist of m genes and n NLTs. An initial empty graph $\mathbf{G}$ is built by the N-dimensional identity matrix, which is defined by Eq. (2).

2. Substitute all covariance values from the diagonal positions of sub-matrix $\mathbf{A}$ into the corresponding positions of sub-matrix $\mathbf{G}_1$, and then based on the updated $\mathbf{G}_1$, use the Iterative Maximum Likelihood Estimates Algorithm (IMLEA) to compute the new covariance matrix (*Dempster, Laird & Rubin, 1977*).

3. Add an undirected edge E_{ij}^1 $((i, j) \in [1, m]^2)$ into $\mathbf{G}_1$, namely add the symmetrical covariance value between the ith gene and jth gene from the positions $\mathbf{A}(i, j)$ and $\mathbf{A}(j, i)$ into the positions $\mathbf{G}_1(i, j)$ and $\mathbf{G}_1(j, i)$, respectively. Then compute a new covariance matrix by the IMLEA. Based on the deviance difference between the new covariance matrix and that before addition, test the significance of the added edge E_{ij}^1 by using the Chi-square distribution with one degree of freedom. The p-value of the Chi-square test is used in the next step as the edge selection criterion. Record the p-value of this tested edge and remove it from $\mathbf{G}_1$.

4. Add a new undirected edge into $\mathbf{G}_1$. Then, repeat the computation in Step 3. After all possible undirected edges have been tested, sort all tested edges in ascending order by their p-values. If the smallest p-value is lower than the predefined cut-off value, add the edge with the smallest p-value into the sub-graph $\mathbf{G}_1$ permanently.

5. Go back to step 3, add the second edge in the updated sub-graph $\mathbf{G}_1$. Repeat the computation in steps 3 and 4 until the smallest p-value of an added edge is larger than the cutoff p-value.

6. Based on the last updated undirected graph $\mathbf{G}_1$, the graph orientation rules are applied to transform the undirected graph into a directed acyclic graph (DAG) (*Meek, 1995*). The inferred DAG with m_1 directed edges, denoted as $\mathbf{A_s}$, represents the predicted regulatory network among m genes.

7. Test the possible edges between m genes and n NLTs. Based on the latest matrix $\mathbf{G}$, add an undirected edge E_{ij}^2 between the ith gene and the jth NLT. That is, add the symmetrical covariance value between the ith gene and jth NLT from the positions $\mathbf{B}(i, j)$ and $\mathbf{B}'(j, i)$ into the positions $\mathbf{G}_2(i, j)$ and $\mathbf{G}_3(j, i)$, respectively. Then, compute a new covariance matrix by the IMLEA. Based on the deviance difference between the new covariance matrix and that before addition, test the significance of the added edge E_{ij}^2 by using the Chi-square distribution with one degree of freedom. The p-value of the Chi-square test is used as the edge selection criterion. Record the p-value of this tested edge E_{ij}^2 and remove it from $\mathbf{G}$.

8. Repeat the computation in steps 7 for the regulation between genes and NLTs. The last updated sub-graph $\mathbf{G}_3$ with n_1 edges, denoted as $\mathbf{B'_s}$, is the predicted directed regulatory

network from n NLTs to m genes. Since we only consider regulations among genes and those from NLTs to genes, the result matrix is given as follows:

$$\mathbf{G_s} = \begin{bmatrix} \mathbf{A_s} \\ \mathbf{B_s'} \end{bmatrix}. \tag{3}$$

The output network includes m_1 directed edges among m gene and n_1 directed edges from n NLTs to m genes.

Note that we have initially applied the GGM in our previous work to the whole matrix $\mathbf{M}$ directly (*Wang et al., 2016*). However, since the number of NLTs is much larger than that of genes, numerical results showed that the majority of selected edges connect NLTs, but few edges are selected to connect genes. This result is not appropriate because the regulations between genes should be the primary mechanisms of the network. Then we conducted another test, in which we did not consider the regulations between NLTs by changing matrix $\mathbf{C}$ into an identity matrix $\mathbf{I_m}$. Matrix $\mathbf{M}$ now is

$$\mathbf{M_1} = \begin{bmatrix} \mathbf{A_s} & \mathbf{B} \\ \mathbf{B'} & \mathbf{I_m} \end{bmatrix}. \tag{4}$$

However, when we applied the GGM to $\mathbf{M_1}$ directly, the singular problem arose during the computation of IMLEA. To satisfy our intention and make the algorithm stable, we proposed EFSA which is executed in two steps. The first step selects regulations between genes and the second step finds regulations from NLTs to genes. The EFSA can be used to predict the gene-gene interactions and the effect from NLTs to genes based on the time-course experimental data.

Bottom-up approach: mathematical model

For a regulatory network with m genes, the expression levels of the i-th gene at time t is denoted as $x_i(t)$. We used the following ordinary differential equation (ODE) model to describe the dynamics of the network (*de Jong, 2002*)

$$\frac{d\mathbf{x}}{dt} = F(t, \mathbf{x}), \tag{5}$$

where $\mathbf{x} = (x_1,\dots,x_m)$ is a vector representing the expression levels of m genes. A number of mathematical formalisms have been proposed to describe the dynamical interactions between different genes in the network, such as the models with linear functions (*de Jong, 2002*)

$$F_i(t, \mathbf{x}) = \sum_{j=1, j \neq i}^{n} a_{ij}x_j - k_i x_i \tag{6}$$

or the models with non-linear functions (*Olariu & Peterson, 2018*)

$$F_i(t, \mathbf{x}) = \frac{\sum_{j=1}^{n} a_{ij}x_j}{1 + \sum_{j=1}^{n} b_{ij}x_j} - k_i x_i \tag{7}$$

The advantage of the model (Eq. 5) with the linear functions (Eq. 6) is that it has a much smaller number of unknown parameters than the non-linear functions (Eq. 7). However, the non-linear model is able to describe the non-linear dynamics more precisely. Therefore, we proposed a method that combines the feature of additive terms in the linear model and the advantages of non-linear model. We applied the second truncated Taylor series approach to approximate the non-linear function (Eq. 7). Here the Taylor series is a mathematical formula to approximate a function by using a polynomial function (*Stewart, 2018*). Thus, we proposed an ODE model (Eq. 5) with the following functions

$$F_i(t, \mathbf{x}) = \sum_{j=1, j\neq i}^{m} \alpha_{ij}x_j + \sum_{1\leq j < k \leq n} \beta_{ijk}x_jx_k - k_ix_i \qquad (8)$$

where k_i is the degradation rate of x_i. This proposed model (Eq. 5) with the non-linear function (Eq. 8) is based on the following assumptions:

1. The regulations from different genes to a particular gene are additive. Similarly, the regulations from non-linear terms (NLTs) to a particular gene are also additive.

2. The regulations from gene j to gene i is represented by $\alpha_{ij}x_j$, where α_{ij} is the coefficient of regulation strength.

3. The regulation of NLT x_jx_k to gene i is represented by $\beta_{ijk}x_jx_k$, where β_{ijk} consists of the regulation strength and equilibrium constant C_{ij}, as we discussed in the sub-section Top-down Approach.

4. The auto-regulation is not considered, namely $\alpha_{ii} = 0$, to avoid confusion between auto-regulation term $\alpha_{ii}x_i$ and degradation term k_ix_i. Note that the issue of auto-regulation may be addressed using a model with non-linear function (Eq. 7). In addition, we just consider the effect of NLTs x_jx_k for $j \neq k$ since the expression levels of x_j may be highly correlated to that of x_j^2. Therefore, we assume that $\beta_{ijj} = 0$.

5. If the value of α_{ij} is positive (negative or zero), it means that gene x_j activates (represses or has no regulation to) the expression of gene x_i. Similar assumption is applied to the value of β_{ijk}.

We emphasize that the proposed method in this work is substantially different from our previous work (*Wang et al., 2016*). The first difference is that the proposed non-linear model (Eq. 8) is different from the non-linear model in *Wang et al. (2016)*. This new model not only can study the regulations from genes to genes, as we considered in our previously proposed model (*Wang et al., 2016*), but also can investigate the effects of heterodimers and/or synergistic effect in genetic regulation. This new model also leads to the second difference compared with our previous top-down approach, namely the proposed Extended Forward Search Algorithm (EFSA) not only includes the probabilistic graphical model in our previous work (*Wang et al., 2016*) but also can predict the possible regulations from NLTs to genes. In addition, in this work, we will infer a medium-sized network first by using EFSA and then reduce the network size by removing

regulations from the network in the Results section, rather than inferring a core network first and then adding regulations to the core network in our previous approach (*Wang et al., 2016*).

Parameter inference

When considering the full connected graph among m genes and n non-linear terms (NLTs), we have an ordinary differential equation (ODE) system with m differential equations. The total number of all unknown coefficients is $m(m + n)$. After applying the Extended Forward Search Algorithm (EFSA), we have an inferred regulatory network which contains only m_1 edges among genes and n_1 edges from NLTs to genes. Thus, the numbers of coefficients α_{ij} and β_{ijk} are reduced from $m(m - 1)$ to m_1 and from mn to n_1, respectively. It is easier to estimate the parameters for the inferred network than for the fully connected network.

In this work, we used a MATLAB toolbox of Genetic Algorithm to estimate the parameters in the proposed mathematical model (*Chipperfield, Fleming & Fonseca, 1994*). The algorithm begins by generating a population of initial parameter values, for example, 100 values. Each initial value is called an individual and the whole population is called one generation. Then it calculates the fitness value for each individual of current generation. Based on the fitness values, the algorithm next creates new values for each individual and thus forms a population of the next generation. This process is repeated until a pre-defined number of generations have been calculated. In this work, we used the following functions, namely function *crtbp* to generate initially binary populations, function *reins* to effect fitness-based reinsertion, function *select* to give a convenient interface to the selection routines, function *recombine* to conduct crossover operators, and function *mut* to conduct binary and integer mutations. The detailed information of these functions and their alternatives can be found in the relevant reference (*Chipperfield, Fleming & Fonseca, 1994*).

To ensure the accuracy of estimates, we set the number of generations as 1,000 and the number of individuals for each generation as 300. For the parameter vector $(\alpha_{ij}, \beta_{ijk}, k_i)$, we used the uniform distribution over the interval $(W_{\min}, W_{\max})$ to generate the initial estimates. Here $W_{\min}$ and $W_{\max}$ are the minimal value and maximal value, respectively, for choosing the samples of the parameters. The values of $W_{\min}$ and $W_{\max}$ are adjusted by computation. For example, if the majority of estimated parameters all are close to $W_{\min}$, then we will further decrease the value of $W_{\min}$. However, if the majority of estimated values are well above $W_{\min}$, then we need to increase the value of $W_{\min}$ accordingly. The similar consideration is applied to $W_{\max}$. In this study, for the erythroid lineage pathway, numerical results suggest that the values of $W_{\min}$ and $W_{\max}$ for $(\alpha_{ij}, \beta_{ijk}, k_i)$ are $(-3, -3, 0)$ and $(3, 3, 1)$, respectively. In addition, for the neutrophil lineage pathway, numerical results suggest that the values of $W_{\min}$ and $W_{\max}$ for $(\alpha_{ij}, \beta_{ijk}, k_i)$ are $(-2.5, -2.5, 0)$ and $(2.5, 2.5, 1)$, respectively. We run the algorithm using an initial random number to generate an initial set of model parameters, which leads to a set of estimated parameters. For each model, we used 200 different initial random numbers, which lead to 200 different sets of estimated model parameters. Denote $x_i(t_j)$ and $x^*_i(t_j)$ as

the observation data and numerical simulations at time point t_j for $j = \{1,2,\ldots,M\}$, respectively. The simulation error is calculated by

$$E = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{M} (x_i(t_j) - x_i^*(t_j))^2}. \tag{9}$$

We selected the top ten sets with the minimal estimated errors out of 200 estimates for further analysis and comparison.

Robustness analysis

We noted that, if a model with the estimated parameters is not robust, a perturbation to the parameters might lead to substantial variations of the model output. Thus, we next used the robustness property of the model to select the inferred model parameter sets from the Genetic Algorithm. This property was designed to examine the robustness of the inferred model to the perturbations of model parameters (*Kitano, 2004*). Robustness property is also an important method for understanding the variations in genetic regulatory networks mathematically (*Masel & Siegal, 2009*). Note that our perturbation test is a mathematical technique. It is different from the perturbation of biological experiments, which may be conducted by the over-expression/knock-down tests. Although we will conduct removal tests by removing edges from the developed model, these tests are designed to remove the unnecessary (or unimportant) regulations in the network.

In this perturbation test, based on the inferred parameter k_i that is assumed to be the unperturbed one, the perturbed parameter is generated by

$$\overline{k_i} = k_i \times (1 + \mu \times \varepsilon) \tag{10}$$

where ε is a sample generated from either the normal distribution or the uniform distribution. In this work, we used the standard Gaussian random variable $\mathbf{N}(0,1)$ to generate samples. In addition, μ is a parameter to determine the values of perturbation (*Wang et al., 2016*). The value of parameter μ determines the variations of simulations. Numerical results suggest that when the value of μ is small, perturbation has small effect on the system dynamics, and it is difficult to distinguish the robustness properties of the model with different parameter sets. However, if the value of μ is large, perturbation will make the model output substantially different, and it will be difficult to measure the robustness property. To make the variations of simulations appropriately for robustness analysis, $\mu = 0.4$ was employed in this study.

For each of the top ten sets of parameters determined in the previous sub-section, we firstly obtained N ($= 5,000$) sets of perturbed model parameters by using (Eq. 10) and then used these parameter sets to obtain N corresponding simulations. We used $x_{ij}^{(k)}(p)$ and $x_{ij}^{(k)}$ to denote the simulation of variable x_i at time point t_j obtained by the k-th perturbed and unperturbed model parameters, respectively. Then, we defined

$$E^{(k)} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{M} (x_{ij}^{(k)}(p) - x_{ij}^{(k)})^2} \tag{11}$$

as the measure for the robustness property of the model with the k-th perturbed parameter set. Afterwards, we defined the robust average for the given parameter set as

$$RA = \frac{1}{N}\sum_{k=1}^{N} E^{(k)}, \tag{12}$$

and robust standard deviation as

$$RSTD = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(E^{(k)} - RA\right)^2} \tag{13}$$

over N perturbation tests. Smaller values of RA and RSTD mean that the model with the given parameter set is more robust.

RESULTS

Inference of regulatory network

To reduce the complexity of regulatory networks, we first used the Extended Forward Search Algorithm (EFSA) to predict the topological structure of genetic networks. The algorithm controls the number of edges by adjusting a pre-defined cut-off value. This value is equivalent to the significant value in statistics. If the threshold is too low, we may miss some significant regulations. However, if the threshold is relatively high, it is quite possible to select insignificant regulations. This work considers the networks including 11 genes and 55 non-linear terms (NLTs). For the sub-network of 11 genes only (i.e. matrix $\mathbf{A_s}$ in Eq. 3), to ensure the statistically significant, we set a specific threshold as 0.1 for both the erythroid regulatory network and neutrophil regulatory network. The selection of this threshold value (i.e. 0.1) is based on the balance between neither selecting much insignificant regulations nor choosing a small number of candidate regulations. Then we had 46 and 40 directed edges for the erythroid regulatory network and neutrophil network, respectively.

For the regulations from NLTs to genes (i.e. matrix $\mathbf{B'_s}$ in Eq. 3), the size of matrix $\mathbf{B'_s}$ is much larger than that of $\mathbf{A_s}$. To avoid the dominance of the regulations from NLTs to genes, we also set the cut-off value as 0.1 for the two networks, or take the first 46 and 40 directed edges from NLTs to genes for the erythrocyte and neutrophil differentiation, respectively, if more edges are selected when using the cut-off value 0.1. The reason we still applied threshold 0.1 here is that the number of selected edges that satisfy this value is much larger than the required number (i.e. 46 for the erythroid regulatory network and 40 for the neutrophil regulatory network). Since the edges are selected and ranked by their significance, we can simply select the top 46 edges and 40 edges for the erythroid and neutrophil pathway, respectively, without conducting any further numerical tests.

Figures S1 and S2 in Supplemental Information present the inferred regulatory networks for the erythroid and neutrophil networks, respectively. Note that there are 11 and 17 isolated NLTs in the erythroid and neutrophil networks, respectively, since no significant edges have been selected from these NLTs by our algorithm. All these isolated

NLTs are listed in the "Isolated NLTs Table". Moreover, all arrows in these figures only represent the direction of regulations, rather than the types of regulations (i.e. positive or negative regulation). We will study the detailed regulatory mechanisms in the next subsection. We found that the targeted gene of the protein heterodimer is a component of that heterodimer in all situations. The possible explanation of this observation is that the expression levels of a heterodimer are the product of the expression levels of the two corresponding genes (namely $x_i x_j$ for genes i and j with expression levels x_i and x_j, respectively). Thus, the expression data of the NLTs $\{x_i x_j\}$ may be highly correlated to those of the component genes, namely $\{x_i\}$ or $\{x_j\}$.

Inference of mathematical model

After the success of constructing regulatory networks in the previous sub-section, we next study the detailed dynamics of genetic networks in fate determination of hematopoietic stem cells (HSCs) by using our proposed mathematical model. The major step is to infer the values of unknown parameters in the model (Eq. 8). If we consider the fully connected model, there should be 11 × (11 + 55) = 726 parameters. However, after the application of EFSA, the number of unknown parameters is reduced to 103 (including 46 directed edges between genes, 46 directed edges from non-linear terms (NLTs) to genes and 11 self-degradation rate constants) for the differentiation of erythrocytes and 91 (including 40 directed edges between genes, 40 directed edges from NLTs to genes and 11 self-degradation rate constants) for the differentiation of neutrophils. We next applied the Genetic-Algorithm to estimate these unknown parameters for two networks.

We used 200 different random numbers to obtain different initial values of rate constants (α_{ij}, β_{ijk}, k_i) over the defined range ($W_{\min}$, $W_{\max}$), which was discussed in the Methods section. This leads to 200 different sets of estimated parameters. Then, we chose the top ten sets of estimated results for each differentiated lineage with the smallest estimation errors for further robustness analysis. According to the definition of estimation error (Eq. 9), the optimal inferred network for the erythrocyte differentiation in our tests has estimation error 0.9902. In addition, the robust average (Eq. 12) and robust standard deviation (Eq. 13) are 0.3977 and 0.1066, respectively. For the neutrophil differentiation, the optimal inferred network has estimation error 0.8726, robust average 0.3983 and robust standard deviation 0.1275.

Figures 1 and 2 present the simulation results based on the optimal estimated parameters for the expression levels of four genes, namely genes *Gata1, PU.1, Ets1* and *Tal1*, for the differentiation of erythrocyte and neutrophil, respectively. The expression levels of *Gata1* increase continuously in both simulated and experimental data during the erythrocyte differentiation. However, during the neutrophil differentiation, experimental data of *Gata1* keep fluctuations and then turn to slightly decreasing at the end of differentiation, which is matched by our simulation. For gene *PU. 1*, both microarray and simulated data decline in the differentiation of erythrocyte but climb during the differentiation of neutrophil. Similarly, the expression levels of *Ets1* in microarray data increase during erythrocyte differentiation but decrease during neutrophil differentiation. Simulation results also fit the trends for both differentiation pathways.

Table 1 Edge deletion test for erythrocyte differentiation. RR, Removed regulation; SE, Simulation error, defined by Eq. (9); RA, Robust average, defined by Eq. (12); RSTD, Robust standard deviation, defined by Eq. (13).

Model	RR	SE	RA	RSTD
OES	N/A	0.9902	0.3977	0.1066
DEL1	Gata2-Notch1 → Notch1 Tal1-Gfi1 → Gfi1 Cbfa2t3-Gfi1 → Gfi1	0.9826	0.4594	0.1259
DEL2	Ldb1 → Lmo2	0.9955	0.3938	0.1124
DEL3	Notch1 → Lmo2	0.9861	0.4506	0.1263
DEL4	Cbfa2t3 → Lmo2	1.0451	0.3820	0.0962
DEL5	Runx1 → Lmo2	1.0298	0.3471	0.0904

Note:
Description of different models: OES, The original model without any deletion; DEL1, Model based on OES by removing regulations from NLTs to genes; DEL2, Model based on DEL1 by removing a regulation among genes; DEL3, Model based on DEL2 by removing a regulation among genes; DEL4, Model based on DEL3 by removing a regulation among genes; DEL5, Model based on DEL4 by removing a regulation among genes.

The experimental data of *Tal1* increase with fluctuations during the first 60 h of erythrocyte differentiation, but then rises rapidly after the first 60 h. Our simulated results are consistent with the expression levels of *Tal1* with the same trend in expression levels. Thus, our simulation results fit the trend of expression levels of these genes very well during two developmental processes. Figures S3 and S4 in Supplemental Information give the simulation results of the other six genes for the differentiation of erythrocyte and neutrophil, respectively.

Reduction of network model—edge deletion

We have obtained two regulatory networks with 92 directed edges and 80 directed edges for erythroid and neutrophil differentiation, respectively. Next we tested the possibility to delete the potential insignificant edges from our predicted regulatory networks. In the first step, we tested the deletion of regulations from non-linear terms (NLTs) to genes. We removed one edge in each test to form a temporary system model, and then examined the simulation error and robustness property of the new model. Afterwards, we removed one specific edge permanently if the corresponding new system has the minimal change in simulation error and robustness property, and then formed an updated model. This test is repeated until both the simulation error and robustness property of the updated model are much worse than the original network without any removal. In the second step, we evaluated the regulatory interactions between 11 genes using the same method in the first step.

For the erythrocyte differentiation, Table 1 suggests that after removing 3 regulations from NLTs to genes, the estimation error (Eq. 9) is improved (shown in DEL1). Then, we tested the regulation reduction from gene to gene. The final result suggests that, after we deleted (*Ldb1 → Lmo2*), (*Notch1 → Lmo2*), (*Cbfa2t3 → Lmo2*) and (*Runx1 → Lmo2*) edges, the estimation error (Eq. 9) is slightly increased. However, the robustness property is better than that of the DEL1 model since the robust average (Eq. 12) is decreased. Thus, for the erythroid differentiation, numerical tests recommended to remove total

Table 2 Edge deletion test for neutrophil differentiation. RR, Removed regulation; SE, Simulation error, defined by Eq. (9); RA, Robust average, defined by Eq. (12); RSTD, Robust standard deviation, defined by Eq. (13).

Model	RR	SE	RA	RSTD
OES	N/A	0.8726	0.3983	0.1275
DEL1	No Suggestion	N/A	N/A	N/A
DEL2	Gata2 → Ldb1	0.8726	0.3943	0.1273
DEL3	Runx1 → Cbfa2t3	0.8726	0.3928	0.1265
DEL4	Ldb1 → Lmo2	0.8748	0.4183	0.1333
DEL5	Tal1 → Lmo2	0.8809	0.3925	0.1237

Note:
Description of different models: OES, The original model without any deletion; DEL1, Model based on OES by removing regulations from NLTs to genes; DEL2, Model based on DEL1 by removing a regulation among genes; DEL3, Model based on DEL2 by removing a regulation among genes; DEL4, Model based on DEL3 by removing a regulation among genes; DEL5, Model based on DEL4 by removing a regulation among genes.

seven edges from our predicted regulatory network. We stopped the deletion test after obtaining the DEL5 model. If we proceed further deletion, both simulation error and robustness property of the temporary network are much worse than the original network without removal.

Table 2 shows that, for the neutrophil differentiation, there are no insignificant regulations from NLTs to genes, because the removal of any edge from NLTs to genes will increase the simulation error (Eq. 9) substantially and/or decrease the robustness property by increasing the values of robust average (Eq. 12) and robust standard deviation (Eq. 13). For the regulations between genes, we have removed the following four regulations, namely (*Gata2 → Ldb1*), (*Runx1 → Cbfa2t3*), (*Ldb1 → Lmo2*) and (*Tal1 → Lmo2*), and formed an updated system. Table 2 shows that the simulation error and robustness property of the updated system are close to those of the original system without any removal of edges. Thus, for the neutrophil differentiation, numerical tests recommended to remove only four edges from our predicted regulatory network. Coincidentally, we stopped the deletion test after obtaining the DEL5 model because of the same reason for the erythrocyte differentiation.

Figures 3 and 4 present the inferred regulatory networks after edge deletion test for erythroid and neutrophil differentiation, respectively. Initially, we have 92 directed edges for the erythrocyte pathway and 80 directed edges for the neutrophil pathway. After the edges deletion, seven and four directed edges have been taken away from the erythrocyte network and neutrophil network, respectively, since the removal of these edges has not much negative influence on simulation error (Eq. 9), robust average (Eq. 12) and robust standard deviation (Eq. 13). Thus, there are 85 and 76 directed edges left for the erythrocyte and neutrophil pathways, respectively.

DISCUSSION

This work was designed to develop a mathematical framework that was able to realize nonlinear gene expression dynamics accurately. In particular, we intended to investigate the effect of possible protein heterodimers and/or synergistic effect in genetic regulation.
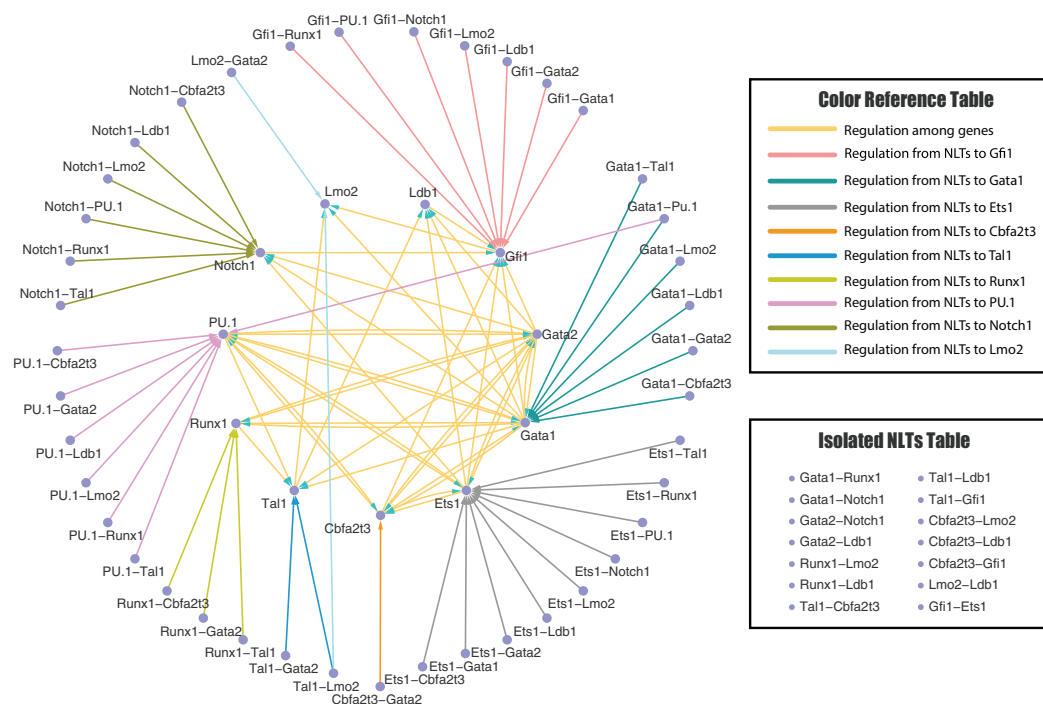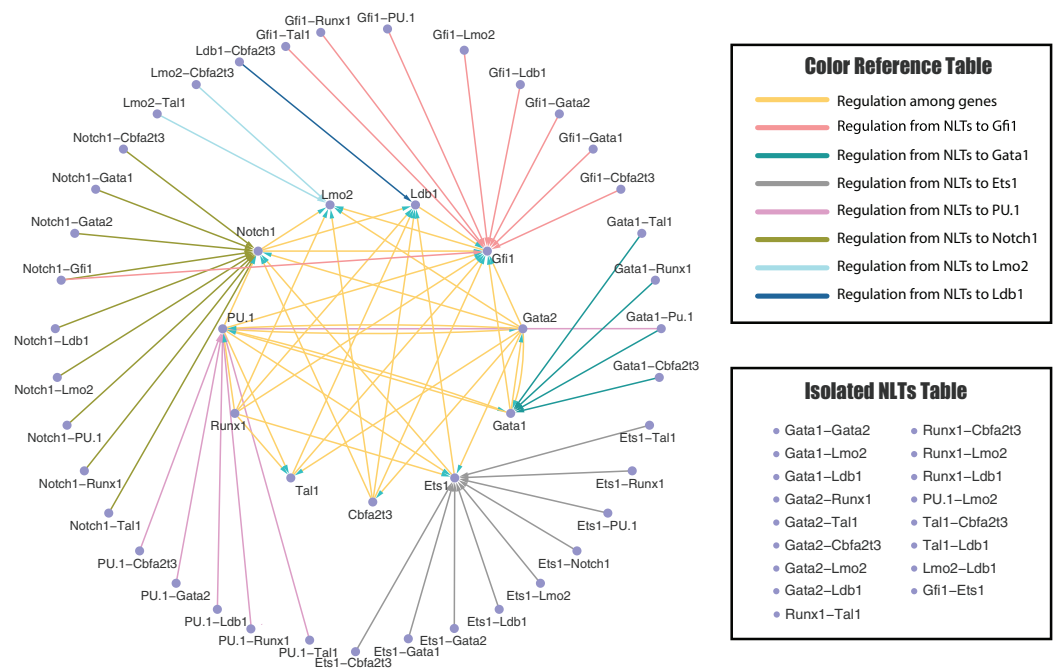
Figure 3 Predicted genetic regulatory network of erythrocyte pathway. The genetic regulatory network predicted by the Extended Forward Search Algorithm with 11 genes and 41 non-linear terms (NLTs) (14 isolated NLTs excluded) after edges deletion test, which is related to the fate determination of erythrocyte pathway: regulatory network for hematopoietic stem cells differentiate to megakaryocyte-erythroid progenitors. The network is visualized by Cytoscape software.

Full-size 🖼 DOI: 10.7717/peerj.9065/fig-3

In this study, we designed the Extended Forward Search Algorithm (EFSA) to predict the topology of regulatory networks connecting genes and heterodimers. We also proposed a new mathematical model for inferring dynamic mechanisms of regulatory networks. Using the EFSA, we derived two regulatory networks of 11 genes for erythrocyte and neutrophil differentiation pathways. According to the predicted networks and experimental data, we estimated parameters in our proposed mathematical model based on the criteria of simulation error and robustness property. By removing regulations with less importance based on simulation error and robustness property, we developed two gene networks that regulate erythrocyte and neutrophil differentiation pathways. Numerical results suggested that our proposed method is capable of reconstructing genetic regulatory networks effectively and accurately.

To infer the regulatory mechanisms of heterodimers, we combined both the top-down approach (i.e. probabilistic graphical model) and the bottom-up approach (i.e. mathematical model). We used the top-down approach first to simplify the network topology and reduced the number of unknown parameters in the mathematical model. Then the Genetic-Algorithm was used to estimate the unknown parameters. The combination of these two approaches reduced the errors in simulation and also improved the robustness property of the mathematical model. In this work, we considered

Figure 4 Predicted genetic regulatory network of neutrophil pathway. The genetic regulatory networks predicted by the Extended Forward Search Algorithm with 11 genes and 38 non-linear terms (NLTs) (17 isolated NLTs excluded) after edges deletion test, which is related to the fate determination of neutrophil pathway: regulatory network for hematopoietic stem cells differentiate to granulocyte-macrophage progenitors. The network is visualized by Cytoscape software.

Full-size ⬜ DOI: 10.7717/peerj.9065/fig-4

the network with medium-sized complexity initially. We then reduced the network complexity by removing edges from the network, rather than studying the core network and then adding the edges to the network in our previous study (*Wang et al., 2016*). The reason for changing the method from "adding edge" to "removing edge" in this work is mainly due to the high computational cost in the "adding edge" tests since the number of candidate edges in the "removing edge test" is much smaller than that in the "adding edge test". Thus, in this work, we used the EFSA to obtain more candidate edges and then used the dynamic model to remove unimportant edges. If the number of potential regulations derived from the probabilistic graphical model is relatively large, the removal of one single regulation from the potential network may not have any changes in simulation error. Numerical results suggested that a couple of regulations should be removed simultaneously in order to achieve changes in simulation error.

The inferred regulatory networks from our proposed models are partially supported by experimental observations. For example, the regulation of *Gata1-Gata2-PU.1* complex in our inferred networks agrees with the experimental results (*May et al., 2013*). The *Gata1-PU.1* heterodimer plays an important role in regulating the hematopoiesis (*Zhang et al., 2000*), which is also included in our inferred model. In addition, the *Ldb1-Lmo2* dimer is activated with significantly expression profiles during the erythroid

differentiation process (*Xu et al., 2003*), which is consistent with our prediction. Moreover, there are evidences to show the existence of synergistic effect of *Tal1, Lmo2* and *Gata1* (*Mead et al., 2001*), which has been inferred in our regulatory networks as well. However, not all of the predictions can be confirmed by the existing experimental observations. The first explanation is that the non-linear terms in our mathematical model are introduced by mathematical operation (i.e. the Taylor series). Some of these non-linear terms may be needed for realizing the nonlinear dynamics accurately, but not supported by biological mechanisms. Note that another inference method, called semi-supervised method, can include the validated regulations first and then infer the invalidated regulations (*Maetschke et al., 2013*). Secondly, our inferred regulatory network may predict some potential possible regulations between genes and from non-linear terms to genes, which may be confirmed by future experimental studies. Thus, the inferred regulations in this work may provide testable prediction for further experimental studies to explore the detailed mechanism of hematopoiesis.

This work also raised a number of important issues in the study of genetic regulations. One question is that our non-linear model still cannot fit all the expression data very well due to noise in the data. Figures 1 and 2 show that the noise in expression data may increase the simulation error of our proposed model. If the noise ratio in expression data is large, it is a challenging issue in mathematical modeling. Large variations in the data may lead to incorrect inference results. In that case, stochastic modeling may be a more appropriate approach to describe the noise in gene expression data (*Samad et al., 2005*; *Tian, 2010*; *Chowdhury, Chetty & Evans, 2015*). In addition, the Gaussian graphical model is based on the covariance matrix. However, the correlation coefficient is suitable to measure the linear correlation relationship. Currently, other approaches, such as mutual information and conditional mutual information, have been used to measures both linear and non-linear correlation relationships between the gene expression data (*Zhang et al., 2012*, *2015*; *Zhao et al., 2016*). Finally, this research determines the regulatory mechanisms based on numerical simulation and robustness property. More information from experimental studies will be important to improve the accuracy of the model and make more reasonable predictions. In addition, we may use other key criteria to select mathematical models, such as Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and Bayesian factor (*Kadane & Lazar, 2004*). All these issues will be the interesting topics of our future research.

CONCLUSION

In conclusion, this study proposes a new method to construct the network topology from genes and heterodimers by a new top-down approach and then develops a non-linear ordinary differential equation model to infer the dynamic mechanisms of regulatory networks. The derived two networks may provide insights regarding the genetic regulations in the cell fate determination of hematopoietic stem cells. The proposed method can also be applied to model other regulatory pathways and biological systems.

ADDITIONAL INFORMATION AND DECLARATIONS

Competing Interests
The authors declare that they have no competing interests.

Author Contributions
- Siyuan Wu performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Tiangang Cui analyzed the data, prepared figures and/or tables, and approved the final draft.
- Xinan Zhang analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Tianhai Tian conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability
The following information was supplied regarding data availability:
The data is available at NCBI GEO: GSE49991. The code is available at GitHub: https://github.com/ThaddeusWu/PeerJ_Submission.

Supplemental Information
Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.9065#supplemental-information.

REFERENCES

Aggarwal R, Lu J, Pompili VJ, Das H. 2012. Hematopoietic stem cells: transcriptional regulation, ex vivo expansion and clinical application. *Current Molecular Medicine* **12(1)**:34–49 DOI 10.2174/156652412798376125.

Athanasiadis EI, Botthof JG, Andres H, Ferreira L, Lio P, Cvejic A. 2017. Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *Nature Commmunication* **8(1)**:631 DOI 10.1038/s41467-017-02305-6.

Birbrair A, Frenette PS. 2016. Niche heterogeneity in the bone marrow. *Annals of the New York Academy of Science* **1370(1)**:82–96 DOI 10.1111/nyas.13016.

Cedar H, Bergman Y. 2011. Epigenetics of haematopoietic cell development. *Nature Reviews Immunology* **11(7)**:478–488 DOI 10.1038/nri2991.

Chang AN, Cantor AB, Fujiwara Y, Lodish MB, Droho S, Crispino JD, Orkin SH. 2002. GATA-factor dependence of the multitype zinc-finger protein FOG-1 for its essential role in megakaryopoiesis. *Proceedings of the National Academy of Sciences of the United States of America* **99(14)**:9237–9242 DOI 10.1073/pnas.142302099.

Chang HH, Oh PY, Ingber DE, Huang S. 2006. Multistable and multistep dynamics in neutrophil differentiation. *BMC Cell Biology* **7(1)**:11 DOI 10.1186/1471-2121-7-11.

Chickarmane V, Enver T, Peterson C. 2009. Computational modeling of the hematopoietic erythroid-myeloid switch reveals insights into cooperativity, priming, and irreversibility. *PLOS Computational Biology* **5(1)**:e1000268 DOI 10.1371/journal.pcbi.1000268.

Chickarmane V, Peterson C. 2008. A computational model for understanding stem cell, trophectoderm and endoderm lineage determination. *PLOS ONE* **3(10)**:e3478 DOI 10.1371/journal.pone.0003478.

Chipperfield AJ, Fleming PJ, Fonseca CM. 1994. Genetic algorithm tools for control systems engineering. In: *Proceedings of Adaptive Computing in Engineering Design and Control*. 128–133.

Chowdhury AR, Chetty M, Evans R. 2015. Stochastic S-system modeling of gene regulatory network. *Cognitive Neurodynamics* **9(5)**:535–547 DOI 10.1007/s11571-015-9346-0.

Crombach A, Wotton KR, Cicin-Sain D, Ashyraliyev M, Jaeger J. 2012. Efficient reverse-engineering of a developmental gene regulatory network. *PLOS Computational Biology* **8(7)**:e1002589 DOI 10.1371/journal.pcbi.1002589.

de Jong H. 2002. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology* **9(1)**:67–103 DOI 10.1089/10665270252833208.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39(1)**:1–22 DOI 10.1111/j.2517-6161.1977.tb01600.x.

Friedman AD. 2007. Transcriptional control of granulocyte and monocyte development. *Oncogene* **26(47)**:6816–6828 DOI 10.1038/sj.onc.1210764.

Gardner TS, Cantor CR, Collins JJ. 2000. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403(6767)**:339–342 DOI 10.1038/35002131.

Goardon N, Lambert JA, Rodriguez P, Nissaire P, Herblot S, Thibault P, Dumenil D, Strouboulis J, Romeo P-H, Hoang T. 2006. ETO2 coordinates cellular proliferation and differentiation during erythropoiesis. *EMBO Journal* **25(2)**:357–366 DOI 10.1038/sj.emboj.7600934.

Guo G, Huss M, Tong GQ, Wang C, Sun LL, Clarke ND, Robson P. 2010. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental Cell* **18(4)**:675–685 DOI 10.1016/j.devcel.2010.02.012.

Hamey FK, Nestorowa S, Kinston SJ, Kent DG, Wilson NK, Göttgens B. 2017. Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proceedings of the National Academy of Sciences of the United States of America* **114(23)**:5822–5829 DOI 10.1073/pnas.1610609114.

Hill AV. 1910. The possible effects of the aggregation of the molecules of hæmoglobin on its dissociation curves. *Journal of Physiology* **40**:iv–vii.

Hoppe PS, Schwarzfischer M, Loeffler D, Kokkaliaris KD, Hilsenbeck O, Moritz N, Endele M, Filipczyk A, Gambardella A, Ahmed N, Etzrodt M, Coutu DL, Rieger MA, Marr C, Strasser MK, Schauberger B, Burtscher I, Ermakova O, Bürger A, Lickert H, Nerlov C,

Theis FJ, Schroeder T. 2016. Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature* **535(7611)**:299–302 DOI 10.1038/nature18320.

Huang S, Guo Y, May G, Enver T. 2007. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental Biology* **305(2)**:695–713 DOI 10.1016/j.ydbio.2007.02.036.

Inouea A, Fujiwaraa T, Okitsua Y, Katsuokaa Y, Fukuharaa N, Onishia Y, Ishizawaa K, Harigaea H. 2013. Elucidation of the role of LMO2 in human erythroid cells. *Experimental Hematology* **41(12)**:1062–1076 DOI 10.1016/j.exphem.2013.09.003.

Kadane JB, Lazar NA. 2004. Methods and criteria for model selection. *Journal of the American Statistical Association* **99(465)**:279–290 DOI 10.1198/016214504000000269.

Kitano H. 2004. Biological robustness. *Nature Reviews Genetics* **5(11)**:826–837 DOI 10.1038/nrg1471.

Krämer N, Schäfer J, Boulesteix A-L. 2009. Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC Bioinformatics* **10(1)**:384 DOI 10.1186/1471-2105-10-384.

Kumano K, Chiba S, Shimizu K, Yamagata T, Hosoya N, Saito T, Takahashi T, Hamada Y, Hirai H. 2001. Notch1 inhibits differentiation of hematopoietic cells by sustaining GATA-2 expression. *Blood* **98(12)**:3283–3289 DOI 10.1182/blood.V98.12.3283.

Lancrin C, Mazan M, Stefanska M, Patel R, Lichtinger M, Costa G, Vargel Ö, Wilson NK, Möröy T, Bonifer C, Göttgens B, Kouskoff V, Lacaud G. 2012. GFI1 and GFI1B control the loss of endothelial identity of hemogenic endothelium during hematopoietic commitment. *Blood* **120(2)**:314–322 DOI 10.1182/blood-2011-10-386094.

Laslo P, Spooner CJ, Warmflash A, Lancki DW, Lee H-J, Sciammas R, Gantner BN, Dinner AR, Singh H. 2006. Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell* **126(4)**:755–766 DOI 10.1016/j.cell.2006.06.052.

Li C, Wang J. 2013. Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: landscape and biological paths. *PLOS Computational Biology* **9(8)**:e1003165 DOI 10.1371/journal.pcbi.1003165.

Li L, Jothi R, Cui K, Lee JY, Cohen T, Gorivodsky M, Tzchori I, Zhao Y, Hayes SM, Bresnick EH, Zhao K, Westphal H, Love PE. 2011. Nuclear adaptor Ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. *Nature Immunology* **12(2)**:129–136 DOI 10.1038/ni.1978.

Liew CW, Rand KD, Simpson RJY, Yung WW, Mansfield RE, Crossley M, Proetorius-Ibba M, Nerlov C, Poulsen FM, Mackay JP. 2006. Molecular analysis of the interaction between the hematopoietic master transcription factors GATA-1 and PU.1. *Journal of Biological Chemistry* **281(38)**:28296–28306 DOI 10.1074/jbc.M602830200.

Ling KW, Ottersbach K, Van Hamburg JP, Oziemlak A, Tsai FY, Orkin SH, Ploemacher R, Hendriks RW, Dzierzak E. 2004. GATA-2 plays two functionally distinct roles during the ontogeny of hematopoietic stem cells. *Journal of Experimental Medicine* **200(7)**:871–872 DOI 10.1084/jem.20031556.

Liu P, Wang F. 2008. Inference of biochemical network models in S-system using multi-objective optimization approach. *Bioinformatics* **24(8)**:1085–1092 DOI 10.1093/bioinformatics/btn075.

Lulli V, Romania P, Morsilli O, Gabbianelli M, Pagliuca A, Mazzeo S, Testa U, Peschle C, Marziali G. 2006. Overexpression of Ets-1 in human hematopoietic progenitor cells blocks erythroid and promotes megakaryocytic differentiation. *Cell Death and Differentiation* **13(7)**:1064–1074 DOI 10.1038/sj.cdd.4401811.

Maetschke SR, Madhamshettiwar PB, Davis MJ, Ragan MA. 2013. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in Bioinformatics* **15(2)**:195–211 DOI 10.1093/bib/bbt034.

Mancini E, Sanjuan-Pla A, Luciani L, Moore S, Grover A, Zay A, Rasmussen KD, Luc S, Bilbao D, O'Carroll D, Jacobsen SE, Nerlov C. 2012. FOG-1 and GATA-1 act sequentially to specify definitive megakaryocytic and erythroid progenitors. *EMBO Journal* **31(2)**:351–365 DOI 10.1038/emboj.2011.390.

Masel J, Siegal ML. 2009. Robustness: mechanisms and consequences. *Trends in Genetics* **25(9)**:395–403 DOI 10.1016/j.tig.2009.07.005.

May G, Soneji S, Tipping AJ, Teles J, McGowan SJ, Wu M, Guo Y, Fugazza C, Brown J, Karlsson G, Pina C, Olariu V, Taylor S, Tenen DG, Peterson C, Enver T. 2013. Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Cell Stem Cell* **13(6)**:754–768 DOI 10.1016/j.stem.2013.09.003.

Mead P, Deconinck A, Huber T, Orkin S, Zon L. 2001. Primitive erythropoiesis in the Xenopus embryo: the synergistic role of LMO-2, SCL and GATA-binding proteins. *Development* **128(12)**:2301–2308.

Meek C. 1995. Causal inference and causal explanation with background knowledge. *Uncertainty in Artificial Intelligence* **11**:403–410.

Meister A, Li YH, Choi B, Wong WH. 2013. Learning a nonlinear dynamical system model of gene regulation: a perturbed steady-state approach. *Annals of Applied Statistics* **7(3)**:1311–1333 DOI 10.1214/13-AOAS645.

Moignard V, Macaulay IC, Swiers G, Buettner F, Schütte J, Calero-Nieto FJ, Kinston S, Joshi A, Hannah R, Theis FJ, Jacobsen SE, De Bruijn M, Göttgens B. 2013. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nature Cell Biology* **15(4)**:363–372 DOI 10.1038/ncb2709.

Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E, Nishikawa S-I, Piterman N, Kouskoff V, Theis FJ, Fisher J, Göttgens B. 2015. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology* **33(3)**:269–279 DOI 10.1038/nbt.3154.

Ng AP, Alexander WS. 2017. Haematopoietic stem cells: past, present and future. *Cell Death & Disease* **3(17002)**:371–380.

Noor A, Serpedin E, Nounou M, Nounou H, Mohamed N, Chouchane L. 2013. An overview of the statistical methods used for inferring gene regulatory networks and protein–protein interaction networks. *Advances in Bioinformatics* **2013(6912)**:1–12 DOI 10.1155/2013/953814.

North TE, Stacy T, Matheny CJ, Speck NA, De Bruijn MF. 2004. Runx1 is expressed in adult mouse hematopoietic stem cells and differentiating myeloid and lymphoid cells, but not in maturing erythroid cells. *Stem Cells* **22(2)**:158–168 DOI 10.1634/stemcells.22-2-158.

Olariu V, Peterson C. 2018. Kinetic models of hematopoietic differentiation. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **11(1)**:e1424 DOI 10.1002/wsbm.1424.

Orkin SH, Zon LI. 2008. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132(4)**:631–644 DOI 10.1016/j.cell.2008.01.025.

Ottersbach K, Smith A, Wood A, Göttgens B. 2010. Ontogeny of haematopoiesis: recent advances and open questions. *British Journal of Haematology* **148(3)**:345–355 DOI 10.1111/j.1365-2141.2009.07953.x.

Porcher C, Swat W, Rockwell K, Fujiwara Y, Alt F, Orkin SH. 1996. The T cell leukemia oncoprotein SCL/tal-1 Is essential for development of all hematopoietic lineages. *Cell* **86(1)**:47–57 DOI 10.1016/S0092-8674(00)80076-8.

Real PJ, Ligero G, Ayllon V, Ramos-Mejia V, Bueno C, Gutierrez-Aranda I, Navarro-Montero O, Lako M, Menendez P. 2012. SCL/TAL1 regulates hematopoietic specification from human embryonic stem cells. *Molecular Therapy* **20(7)**:1443–1453 DOI 10.1038/mt.2012.49.

Samad HE, Khammash M, Petzold L, Gillespie D. 2005. Stochastic modelling of gene regulatory networks. *International Journal of Robust and Nonlinear Control* **15(15)**:691–711 DOI 10.1002/rnc.1018.

Shea MA, Ackers GK. 1985. The OR control system of bacteriophage lambda a physical-chemical model for gene regulation. *Journal of Molecular Biology* **181(2)**:211–230 DOI 10.1016/0022-2836(85)90086-5.

Shivdasani RA, Mayer EL, Orkin SH. 1995. Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein Tal1/SCL. *Nature* **373(6513)**:432–434 DOI 10.1038/373432a0.

Soler E, Andrieu-Soler C, De Boer E, Bryne JC, Thongjuea S, Stadhouders R, Palstra1 R-J, Stevens M, Kockx C, Van IJcken W, Hou J, Steinhoff C, Rijkers E, Lenhard B, Grosveld F. 2010. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes and Development* **24(3)**:277–289 DOI 10.1101/gad.551810.

Stewart J. 2018. *Calculus*, chapter infinite sequences and series. Cengage learning. *Available at* https://www.stewartcalculus.com/.

Stier S, Cheng T, Dombkowski D, Carlesso N, Scadden DT. 2002. Notch1 activation increases hematopoietic stem cell self-renewal in vivo and favors lymphoid over myeloid lineage outcome. *Blood* **99(7)**:2369–2378 DOI 10.1182/blood.V99.7.2369.

The Gene Ontology Consortium. 2017. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Research* **45(D1)**:D331–D338.

Tian T. 2010. Stochastic models for inferring genetic regulation from microarray gene expression data. *Biosystems* **99(3)**:192–200 DOI 10.1016/j.biosystems.2009.11.002.

Tian T, Smith-Miles K. 2014. Mathematical modelling of GATA-switching for regulating the differentiation of hematopoietic stem cell. *BMC Systems Biology* **8(Suppl. 1)**:S8 DOI 10.1186/1752-0509-8-S1-S8.

van der Meer LT, Jansen JH, van der Reijden BA. 2010. Gfi1 and Gfi1b: key regulators of hematopoiesis. *Leukemia* **24(11)**:1834–1843 DOI 10.1038/leu.2010.195.

Visvader JE, Mao X, Fujiwara Y, Hahm K, Orkin SH. 1997. The LIM-domain binding protein Ldb1 and its partner LMO2 act as negative regulators of erythroid differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **94(25)**:13707–13712 DOI 10.1073/pnas.94.25.13707.

Wang J, Myklebost O, Hovig E. 2003. MGraph: graphical models for microarray data analysis. *Bioinformatics* **19(17)**:2210–2211 DOI 10.1093/bioinformatics/btg298.

Wang J, Tian T. 2010. Quantitative model for inferring dynamic regulation of the tumour suppressor gene P53. *BMC Bioinformatics* **11(36)**:45 DOI 10.1186/1471-2105-11-36.

Wang J, Wu Q, Hu XT, Tian T. 2016. An integrated approach to infer dynamic protein–gene interactions, a case study of the human P53 protein. *Methods* **110**:3–13 DOI 10.1016/j.ymeth.2016.08.001.

Wei J, Hu X, Zou X, Tian T. 2017. Reverse-engineering of gene networks for regulating early blood development from single-cell measurements. *BMC Medical Genomics* **10(S5)**:72 DOI 10.1186/s12920-017-0312-z.

Woods ML, Leon M, Perez-Carrasco R, Barnes CP. 2016. A statistical approach reveals designs for the most robust stochastic gene oscillators. *ACS Synthetic Biology* **5(6)**:459–470 DOI 10.1021/acssynbio.5b00179.

Wu S, Cui T, Tian T. 2018. Mathematical modelling of genetic network for regulating the fate determination of hematopoietic stem cells. In: *2018 International Conference on Bioinformatics and Biomedicine (BIBM)*. 2167–2173.

Xiong W, Ferrell JE. 2003. A positive-feedback-based bistable 'memory module' that governs a cell fate decision. *Nature* **426(6965)**:460–465 DOI 10.1038/nature02089.

Xu Z, Huang S, Chang L-S, Agulnick A, Brandt S. 2003. Identification of a Tal1 target gene reveals a positive role for the LIM domain-binding protein Ldb1 in erythroid gene expression and differentiation. *Molecular and Cellular Biology* **23(21)**:7585–7599 DOI 10.1128/MCB.23.21.7585-7599.2003.

Yang B, Bao W. 2019. RNDEtree: regulatory network with differential equation based on flexible neural tree with novel criterion function. *IEEE Access* **7**:58255–58263 DOI 10.1109/ACCESS.2019.2913084.

Yang B, Bao W, Huang D-S, Chen Y. 2018. Inference of large-scale time-delayed gene regulatory network with parallel mapreduce cloud platform. *Scientific Reports* **8(1)**:17787 DOI 10.1038/s41598-018-36180-y.

Ye F, Huang W, Guo G. 2017. Studying hematopoiesis using single-cell technologies. *Journal of Hematology & Oncology* **10(1)**:27 DOI 10.1186/s13045-017-0401-7.

Zhang P, Zhang X, Iwama A, Yu C, Smith KA, Mueller BU, Narravula S, Torbett BE, Orkin SH, Tenen DG. 2000. PU.1 inhibits GATA-1 function and erythroid differentiation by blocking GATA-1 DNA binding. *Blood* **96(8)**:2641–2648 DOI 10.1182/blood.V96.8.2641.

Zhang X, Zhao J, Hao J-K, Zhao X-M, Chen L. 2015. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic Acids Research* **43(5)**:e31 DOI 10.1093/nar/gku1315.

Zhang X, Zhao X-M, He K, Lu L, Cao Y, Liu J, Hao J-K, Liu Z-P, Chen L. 2012. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* **28(1)**:98–104 DOI 10.1093/bioinformatics/btr626.

Zhang Y, Payne KJ, Zhu Y, Price MA, Parrish YK, Zielinska E, Barsky LW, Crooks GM. 2005. SCL expression at critical points in human hematopoietic lineage commitment. *Stem Cells* **23(6)**:852–860 DOI 10.1634/stemcells.2004-0260.

Zhao J, Zhou Y, Zhang X, Chen L. 2016. Part mutual information for quantifying direct associations in networks. *Proceedings of the National Academy of Sciences of the United States of America* **113(18)**:5130–5135 DOI 10.1073/pnas.1522586113.

Zhao W, Kitidis C, Fleming MD, Lodish HF, Ghaffari S. 2006. Erythropoietin stimulates phosphorylation and activation of GATA-1 via the PI3-kinase/AKT signaling pathway. *Blood* **107(3)**:907–915 DOI 10.1182/blood-2005-06-2516.