Reconstructing gene regulatory networks that control hematopoietic commitment

Fiona K. Hamey & Berthold Göttgens

Affiliation: University of Cambridge, Department of Haematology & Wellcome and MRC Cambridge Stem Cell Institute

Corresponding author: B. Göttgens, e-mail: bg200@cam.ac.uk

i. Summary

Haematopoietic stem cells (HSCs) reside at the apex of the haematopoietic hierarchy, possessing the ability to self-renew and differentiate towards all mature blood lineages. Along with more specialised progenitor cells, HSCs have an essential role in maintaining a healthy blood system. Incorrect regulation of cell fate decisions in stem/progenitor cells can lead to an imbalance of mature blood cell populations – a situation seen in diseases such as leukaemia. Transcription factors, acting as part of complex regulatory networks, are known to play an important role in regulating haematopoietic cell fate decisions. Yet discovering the interactions present in these networks remains a big challenge. Here, we discuss a computational method that uses single cell gene expression data to reconstruct Boolean gene regulatory network models, and show how this technique can be applied to enhance our understanding of transcriptional regulation in haematopoiesis.

ii. Key Words

Gene regulatory network; single cell; pseudotime; transcriptional regulation; haematopoiesis; haematopoietic stem/progenitor cell; Boolean network

1. Introduction

Due to their high turnover, haematopoietic cells require constant replacement in order to sustain the blood system throughout adult life. Extensive research over the past 60 years has revealed that haematopoietic stem and progenitor cells (HSPCs) at different stages of commitment can be found in the mammalian bone marrow, with cells from specific populations specified towards one or more of the more than 10 different lineages of mature blood cells. This led to the popular view of a haematopoietic hierarchy, where haematopoietic stem cells (HSCs) can differentiate into increasingly specialised progenitor cell populations (Fig. 1A). Due to the accessibility of material and the existence of cell surface marker-based sorting strategies, these haematopoietic populations can be readily isolated and studied, and so are well-characterised in comparison to stem and progenitor cells in many other adult systems. The haematopoietic system, therefore, presents attractive opportunities for studying stem cell differentiation.

Transcription factors as key regulators of cell fate decisions

The haematopoietic system must maintain an appropriate balance of all mature cell types, as dysregulation of haematopoietic cell fate decisions is linked to potentially fatal blood disorders such as acute myeloid leukaemia. It is therefore vital to understand how blood stem cells regulate their decision to differentiate towards alternative fates. Across many biological systems it is well established that expression of genes encoding transcription factors plays an important role in determining cell fate decisions [1]. In particular, transcriptional regulation plays a central role in determining cell fate decisions and executing lineage differentiation in the blood [2].

Discovering gene regulatory networks

In their role of driving specific lineage choices, transcription factors act as components of complex regulatory relations known as gene regulatory networks. Whilst experimental approaches have revealed numerous gene interactions with crucial roles in haematopoiesis, relying purely on such time-consuming and expensive investigation to discover regulatory relationships is not feasible. Instead, the increased availability of high quality gene expression data sets has led many researchers to use such data as a starting point for computationally discovering regulatory relationships across a variety of biological systems [3].

Modelling gene regulatory networks

As well as providing a means for discovery, gene regulatory network models enable networks to be simulated in an attempt to better understand the role of gene regulation in a system. One approach used for modelling transcriptional regulatory networks is Boolean abstraction, which has proved popular in studies of the blood system. Here, gene expression levels are converted to ON/OFF expression, and regulation between genes is described as logical rules. Literature curation has been used to build Boolean regulatory network models for common myeloid progenitors [4], lymphoid progenitors [5] and in blood stem cells [6]. Boolean models present a powerful tool for network modelling as they easily scale to large numbers of components, the effect of network perturbations can be easily simulated in silico and network simulation relates to experimental approaches that readily translate into laboratory-based experiments such as gene knock-outs.

The importance of looking at the single cell level

The majority of existing models of blood transcriptional regulatory networks are based on either literature curation, or experimental approaches using bulk expression data. However, investigation at the single cell level is essential, as considering only bulk expression data averages over often heterogeneous populations, where heterogeneity in gene expression levels is linked to cell fate choices [7]. The power of single-cell data in uncovering regulatory relationships is increasingly being recognised, for example correlation analysis of single-cell qRT-PCR data has been successfully used to identify novel regulatory relationships between transcription factors [8, 9]. In systems other than adult haematopoiesis, single-cell expression data have been used to infer Boolean network models in embryonic stem cells [10] and embryonic blood development [11].

In this chapter, we discuss one approach for reconstructing Boolean gene regulatory network models of a haematopoietic differentiation process using snapshot single cell gene expression data [12].

2. Materials

Single cell gene expression data

The input for this network inference method is a set of single cell gene expression values for cells at different stages of a differentiation process of interest. As the method involves calculation of correlations between genes it is beneficial to have a high dynamic range for the majority of genes measured. Additionally, when modelling the Boolean network genes must ultimately be discretised into “ON” or “OFF” states. This requires data with high sensitivity for detecting whether a gene is expressed in a given cell. High quality single cell quantitative real-time PCR (qRT-PCR) data fit both of these requirements. In our case study we used data originally published in [7] and supplemented with additional populations in [12]. These data sampled cells from haematopoietic stem and progenitor cells using 12 different sorting strategies to sample cells at different stages of commitment towards mature blood lineages (Fig. 1B).

Ordering cells through differentiation

Pseudotime describes the concept of ordering single cell molecular profiles based on similarities in the expression states of individual cells. The motivation behind this is the idea that cells close together in terms of differentiation will have smaller changes in gene expression than those further apart (Fig. 1C). This then allows properties such as changes in gene expression levels to be visualised across differentiation (Fig. 1D). Monocle [13] and Wanderlust [14] were the first algorithms designed to order single cell data into trajectories. Since then, improved versions of these algorithms, as well as entirely new approaches, have been published, able to cope with complex branching trajectories [15–17]. When analysis for our case study was carried out, these later algorithms were not available, hence we applied the Wanderlust algorithm [14] to our data.

SAT Solver

The network inference method we describe here involves a step of searching a space of possible Boolean functions to find those with the highest score measuring agreement with the data. With even a relatively small number of genes the space of Boolean functions becomes incredibly large, so in order to make this computationally tractable we encode the problem as a Boolean satisfiability problem to find the rules on a per gene basis. We implemented this using the Python Z3 solver (https://github.com/Z3Prover/z3/), which provides an efficient search for the highest scoring functions.

3. Methods

It has long been recognised that gene expression data provides a powerful way to computationally discover transcriptional regulatory relationships, for example by considering correlations between genes. More recently, large single cell gene expression data sets have proved an effective starting point for building transcriptional networks. Here we describe a method for constructing a Boolean network model from single cell data.

Data pre-processing

Single cell gene expression measurements from cells at different stages of some differentiation process provide the input for the network inference algorithm (Fig. 2A). Single-cell qRT-PCR gene expression measurements are well suited to this analysis. Data can either be acquired pre-normalised from published studies or Ct values can be transformed to ΔCt values by normalisation against housekeeping genes on a per cell basis. Any housekeeping genes, as well as genes that fail quality control due to technical issues, should then be excluded from downstream analysis.

Identifying possible gene regulation

The first step in our network inference method is to identify potential regulatory relationships between pairs of transcription factor encoding genes. This is done by calculating pairwise partial correlation coefficients across the whole dataset. Correlation coefficients are filtered to retain pairs with significant interaction, for example using a threshold of p-value < 0.01. Links between the gene pairs are then ranked by the magnitude of the correlation coefficients, and the strongest correlations retained as edges in a gene correlation network. Positive correlation between gene A and gene B is then treated as possible activation of gene A by gene B, or of gene B by gene A. Negative correlation is treated as potential repression acting in either direction. Potential activating or repressing relationships are combined with self-activation for each gene, and combinations of these high correlation edges describe a set of possible Boolean functions governing the expression of each gene, with each rule featuring one or more regulators (Fig. 2B). To reduce the search time of our algorithm, we restrict the search to functions of the form $F = F_1 \wedge \neg F_2$ with each F_i a Boolean function made from AND and OR gates with at most two inputs per gate. F_1 represents the activating part of the function, consisting of at most four activating transcription factors for a gene, and F_2 the repressing part of the function, formed from at most two repressing transcription factors.

Identification of differentiation trajectories

Whilst high correlation between a pair of genes can be an indication of a regulatory relationship, additional information is required to determine if the regulation is direct, establish the direction of regulation between genes, and understand if this part of a regulatory event that requires the involvement of multiple transcription factors. This can be done using expression data from different time points or perturbed experimental conditions, yet in many cases these data are not available and would be costly and time-consuming to generate. Instead, we exploit the concept of pseudotime [13, 14], by using this computational ordering of cells as the basis for a scoring mechanism that is applied to the set of possible Boolean functions from the correlation network (Fig. 2C). For this we need to order cells from the most immature population in the data through to the mature cells along one or more trajectories. A separate ordering is required for each ‘end point’ corresponding to a different mature population. The first goal is to identify the cells representing the ‘start’ and ‘end’ cells for each trajectory. Here, we enlist the help of diffusion maps, a visualisation tool which has been successfully applied to single-cell gene expression data to capture structure related to differentiation within the data [18]. If cells have been isolated from different time points, or are known to be different types (e.g. were captured using different surface marker sorting strategies in the blood) then this can be used to guide the selection of start and end cells for each trajectory. Alternatively, several pseudotime inference algorithms now provide methods for inferring the tip positions of trajectories automatically. In the case of data sets where cells can branch to multiple cells fates, it is also necessary to identify which cells lie on the differentiation trajectory towards a specific end cell. In our case study below, we followed a previously described method [19] to assign cells to branches. Again, some pseudotime algorithms now provide an inbuilt method for this.

Selecting activating or repressing edges in the network model

So far the single cell gene expression data has provided both a set of possible Boolean functions describing the regulation of each gene, and an ordering of cells from the start to the end of a differentiation trajectory. We next describe how to use the gene changes along the pseudotime ordering to score the functions on a per gene basis, to identify functions that best fit the data. Gene expression in each cell along the trajectory is first discretised into “ON” or “OFF” expression states, by setting any detected values of gene expression to 1, and any undetected values to 0. Each pair of cells positioned k steps apart in the pseudotime ordering is then treated as an input-output pair $P_i = (I_i, O_i)$ for a Boolean function, where $[I_i]_g$ indicates the binary expression of gene g in input cell I_i. Each function F for a gene g is given a score

$S(F) = \sum_i s_i(F)$ where $s_i = \begin{cases} 1, \ if \ [F(I_i)]_g \ = \ [O_i]_g \\ 0, otherwise \end{cases}$ for the pseudotime pairs $P_i = (I_i, O_i)$. This calculates the number of times the value of the gene g as predicted by F applied to I_i equals the value of g in the corresponding output cell O_i. The top scoring functions for each gene can then be considered the 'best' functions for that gene. To identify the top scoring functions, we encoded the problem as a Boolean satisfiability problem, and provide Python code that can be run to identify rules for each gene (https://github.com/fionahamey/Pseudotime-network-inference).

Limitations

As with all computational inference methods, the output from this approach can only be considered as a tool for hypothesis generation, either for identifying potential novel regulations, or utilising the resulting Boolean network to simulate the outcome of scenarios such as overexpression and/or knockdown of specific genes in the network. Consequently, some form of experimental validation is nearly always necessary when applying such methods.

One major limitation of an approach based on qRT-PCR data is that the network model will always be restricted to the genes that have been chosen for profiling in the experiment and passed quality control measures. As a result, especially when using previously published datasets perhaps not originally designed for a network inference study, important regulators may in fact be missing from the network. Nevertheless, our approach provides a valuable way for identifying ways in which known regulators act as part of a network to control cell fate decisions.

Case study on haematopoietic stem cell differentiation

To demonstrate our network inference method, we decided to investigate transcriptional regulation of cell fate decisions in murine haematopoietic stem and progenitor cells. We used previously published single cell qRT-PCR data from 12 different haematopoietic cell-sorting strategies that profiled long-term HSCs, finite self-renewal HSCs, multipotent progenitors, pre-megakaryocyte-erythroid progenitors, megakaryocyte-erythroid progenitors (MEP), granulocyte-monocyte progenitors, lymphoid-primed multipotent progenitors (LMPP) and common myeloid progenitors [7, 12]. These data measured the expression of 42 genes including 32 transcription factors in each cell. For network inference we chose to focus on two trajectories, one from HSCs to MEPs and the other from HSCs to LMPPs. Start and end cells for trajectories were selected based on the diffusion map analysis of these data (calculated using the *destiny* R package [20]). We followed a previously described method [19] to select cells on two branches from HSCs to MEP and LMPP cells. This was done by constructing a $k = 30$ nearest-neighbour graph on coordinates of the single cell profiles in the first four diffusion components. Each branch was then identified by taking the 100 nearest neighbours of all cells lying on the shortest path between the start and end cells. The Wanderlust algorithm [14] was used to order cells along this path, assigning a pseudotime value to each individual cell. A partial correlation network was calculated on the transcription factor encoding genes with edges between the 100 strongest correlating pairs. This correlation network and the pseudotime ordering were used as inputs to our network inference algorithm, which was applied separately to the two trajectories. A simplified network can be viewed in Fig. 3, which indicates the activation or repression between transcription factors identified in our model.

4. Notes

Here we discuss some points for applying the algorithm for network inference to a single cell data set.

Pseudotime method

Although we used the Wanderlust algorithm [14] for ordering cells in pseudotime, many more pseudotime inference algorithms have recently been published, and would also be suitable for this analysis. In particular, users may find algorithms such as diffusion pseudotime [15], which can automatically detect trajectory tips within a data set, useful in cases where the start or end of trajectories is not as well understood as in haematopoiesis. The Python code provided for network inference takes an ordered list of cells along a trajectory as input, hence is compatible with a pseudotime ordering from any algorithm.

Input for Python code of network inference algorithm

Code for our network inference algorithm can be downloaded from github (https://github.com/fionahamey/Pseudotime-network-inference) and requires 3 input files: the matrix of binary gene expression, the pseudotime ordering of cells on a differentiation trajectory, and a list of the possible activators and repressors of each gene. Instructions of the format of these files and the parameters for the algorithm are provided on the github page.

Speeding up the analysis

If a gene has many activators or repressors it can vastly increase the search time to identify the high scoring functions. To reduce this time some parameters can be altered when running the algorithm. Firstly, the maximum numbers of permitted activators or repressors can be reduced in order to search for simpler rules. Secondly, the threshold and threshold step size can be changed. The algorithm works by first searching for any rules that have agreement with a user-defined threshold, for example 90%, of the pseudotime input-output pairs. If no rules are found, the percentage is lowered and the algorithm iterates. If there are a very large number of rules above the starting threshold the algorithm will search for a long time to find all of these, but we are only interested in the highest scoring rule of this set. Therefore it would reduce the search time to start with a much higher threshold. The threshold lowering step size can also be altered in an attempt to limit the number of rules above the newly lowered threshold.

Simplifying the resulting rule set

In many cases, particularly when a gene had a high number of regulators, the algorithm returns several rules with equal scores. We first remove any rule with self-activation that does not score higher than the same rule disregarding self-activation. Groups of rules are then simplified to reduce the overall number of rules, discarding those contained within other rules in the set to give the smallest and simplest set of rules. For example $\{A \rightarrow C, B \rightarrow C, A \vee B \rightarrow C\}$ would be simplified to $A \vee B \rightarrow C$. Whereas $\{A \rightarrow C, A \wedge B \rightarrow C\}$ would be reduced to only $A \rightarrow C$.

Choice of gene sets for pseudotime ordering and network inference

In the above case study in mouse bone marrow 42 genes were measured by single-cell qRT-PCR, 32 of which were transcription factors. For transcriptional regulatory network model inference we were only interested in the expression of transcription factor encoding genes, and so limited the analysis to the set of transcription factors. However, for diffusion map visualisation and pseudotime ordering the continuous expression levels of all genes except housekeeping genes were used, as this gene set demonstrated improved separation of different cells types.

5. References

1. Davidson EH, Peter IS (2015) Genomic Control Process, 2nd ed. Academic Press, Oxford
2. Göttgens B (2015) Regulatory network control of blood stem cells. Blood 125:2614–2620 . doi: 10.1182/blood-2014-08-570226
3. Marbach D, Costello JC, Küffner R, et al (2012) Wisdom of crowds for robust gene network inference. Nat Methods 9:796
4. Krumsiek J, Marr C, Schroeder T, Theis FJ (2011) Hierarchical Differentiation of Myeloid Progenitors Is Encoded in the Transcription Factor Network. PLoS One 6:e22649
5. Collombet S, van Oevelen C, Sardina Ortega JL, et al (2017) Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. Proc Natl Acad Sci 114:5792–5799 . doi: 10.1073/pnas.1610622114
6. Bonzanni N, Garg A, Feenstra KA, et al (2013) Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. Bioinformatics 29:i80-8 . doi: 10.1093/bioinformatics/btt243
7. Wilson NK, Kent DG, Buettner F, et al (2015) Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. Cell Stem Cell 16:712–24 . doi: 10.1016/j.stem.2015.04.004
8. Moignard V, Macaulay IC, Swiers G, et al (2013) Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. Nat Cell Biol 15:363–372 . doi: 10.1038/ncb2709

9. Pina C, Teles J, Fugazza C, et al (2015) Single-Cell Network Analysis Identifies DDIT3 as a Nodal Lineage Regulator in Hematopoiesis. Cell Rep 11:1503–1510 . doi: 10.1016/j.celrep.2015.05.016

10. Xu H, Ang Y-S, Sevilla A, et al (2014) Construction and Validation of a Regulatory Network for Pluripotency and Self-Renewal of Mouse Embryonic Stem Cells. PLOS Comput Biol 10:e1003777

11. Moignard V, Woodhouse S, Haghverdi L, et al (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol 33:269–76 . doi: 10.1038/nbt.3154

12. Hamey FK, Nestorowa S, Kinston SJ, et al (2017) Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. Proc Natl Acad Sci 114:5822–5829 . doi: 10.1073/pnas.1610609114

13. Trapnell C, Cacchiarelli D, Grimsby J, et al (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 32:381–386 . doi: 10.1038/nbt.2859

14. Bendall SC, Davis KL, Amir E-AD, et al (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell 157:714–725 . doi: 10.1016/j.cell.2014.04.005

15. Haghverdi L, Büttner M, Wolf FA, et al (2016) Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods 13:845–848 . doi: 10.1038/nmeth.3971

16. Setty M, Tadmor MD, Reich-Zeliger S, et al (2016) Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol advance on: . doi: 10.1038/nbt.3569

17. Qiu X, Mao Q, Tang Y, et al (2017) Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 14:979–982 . doi:

10.1038/nmeth.4402

18. Haghverdi L, Buettner F, Theis FJ (2015) Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics 1–10

19. Ocone A, Haghverdi L, Mueller NS, Theis FJ (2015) Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. Bioinformatics 31:i89-96 . doi: 10.1093/bioinformatics/btv257

20. Angerer P, Haghverdi L, Büttner M, et al (2016) destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics 32:1241–1243

Acknowledgements

Work in the author's laboratory is supported by grants from the Wellcome, Bloodwise, Cancer Research UK, NIH-NIDDK and core support grants by the Wellcome to the Cambridge Institute for Medical Research and Wellcome & MRC Cambridge Stem Cell Institute. F.K.H. is a recipient of a Medical Research Council PhD Studentship.

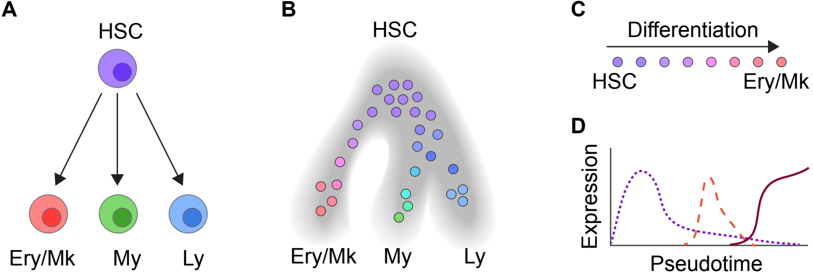
Figure Captions

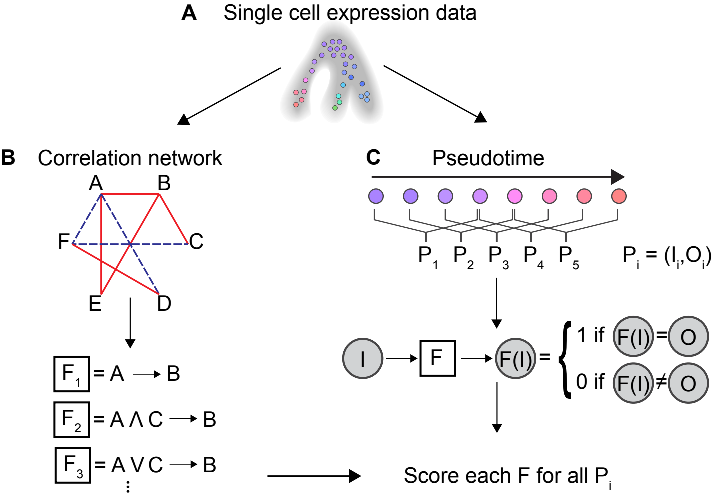
Fig. 1. Single cell snapshot gene expression data can be used to reconstruct the transcriptional landscape of haematopoiesis. (A) Haematopoietic stem cells (HSCs) reside at the apex of the haematopoietic hierarchy. These cells can differentiate towards all of the different blood lineages. Ery, erythroid; Mk, megakaryocytic; My, myeloid; Ly, lymphoid. (B) By sampling single cells from the bone marrow and profiling their gene expression it is possible to capture cells at different stages of differentiation and build up a picture of the underlying transcriptional landscape. (C) Cells can be computationally ordered based on

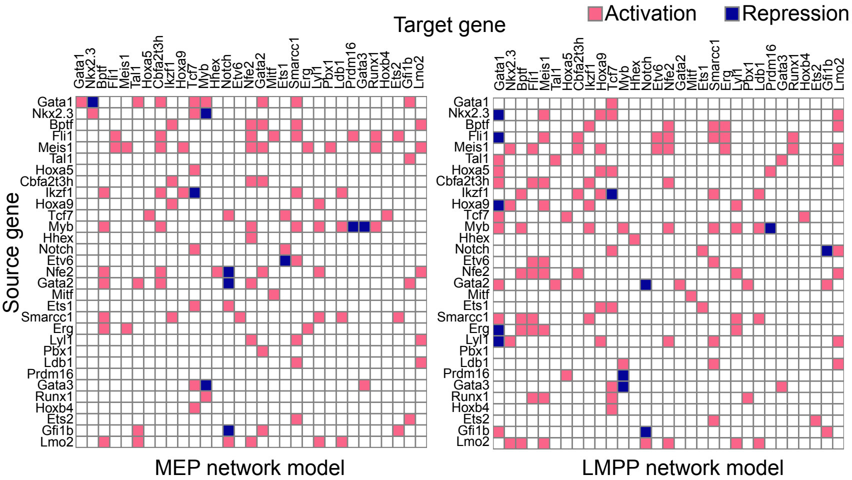
similarities in their transcriptional profile. This ordering, often described as ‘pseudotime’, aims to recapitulate molecular changes during differentiation. (D) Pseudotime orderings can be used to investigate the ordering of expression changes during differentiation for different markers.

Fig. 2. Gene regulatory network models can be inferred from single cell gene expression data. (A) The network inference method is based on single cell gene expression profiles of cells at different stages of differentiation. (B) A set of possible Boolean logic functions describing the regulatory rules of each gene are found by first identifying strong positive and negative correlations between pairs of genes. This then generates a list of possible functions for each gene. Here A-F represent a set of genes. (C) Cells are ordered in pseudotime, and pairs $P_i = (I_i, O_i)$ constructed by taking cells a fixed distance apart in the pseudotime ordering. These cells act as input and output to a Boolean function F. The functions generated by the correlation network are scored against all of the pseudotime pairs.

Fig 3. Alternative gene regulatory network models for differentiation towards megakaryocyte/erythroid progenitors (MEPs) and lymphoid-primed multipotent progenitors (LMPPs) can be identified using single-cell snapshot data from mouse bone marrow. Heatmaps indicate interactions present in simplified network models, showing activation or repression from source gene (rows) to target gene (columns).


A
HSC
Ery/Mk
My
Ly
B
HSC
Ery/Mk
My
Ly
C
Differentiation
HSC
Ery/Mk
D
Expression
Pseudotime


A
Single cell expression data
B
Correlation network
A
B
F
C
E
D
F1 = A → B
F2 = A ∧ C → B
F3 = A ∨ C → B
⋮
C
Pseudotime
P1
P2
P3
P4
P5
Pi = (Ii,Oi)
I
F
F(I) =
1 if F(I) = O
0 if F(I) ≠ O
Score each F for all Pi


Activation
Repression
Target gene
Source gene
Gata1
Nkx2.3
Bptf
Fli1
Meis1
Tal1
Hoxa5
Cbfa2t3h
Ikzf1
Hoxa9
Tcf7
Myb
Hhex
Notch
Etv6
Nfe2
Gata2
Mitf
Ets1
Smarcc1
Erg
Lyl1
Pbx1
Ldb1
Prdm16
Gata3
Runx1
Hoxb4
Ets2
Gfi1b
Lmo2
MEP network model
LMPP network model