Stem Cell Reports, Volume *7*

Supplemental Information

A Generalized Gene-Regulatory Network Model of Stem Cell Differentiation for Predicting Lineage Specifiers

Satoshi Okawa, Sarah Nicklas, Sascha Zickenrott, Jens C. Schwamborn, and Antonio del Sol

Figure S1

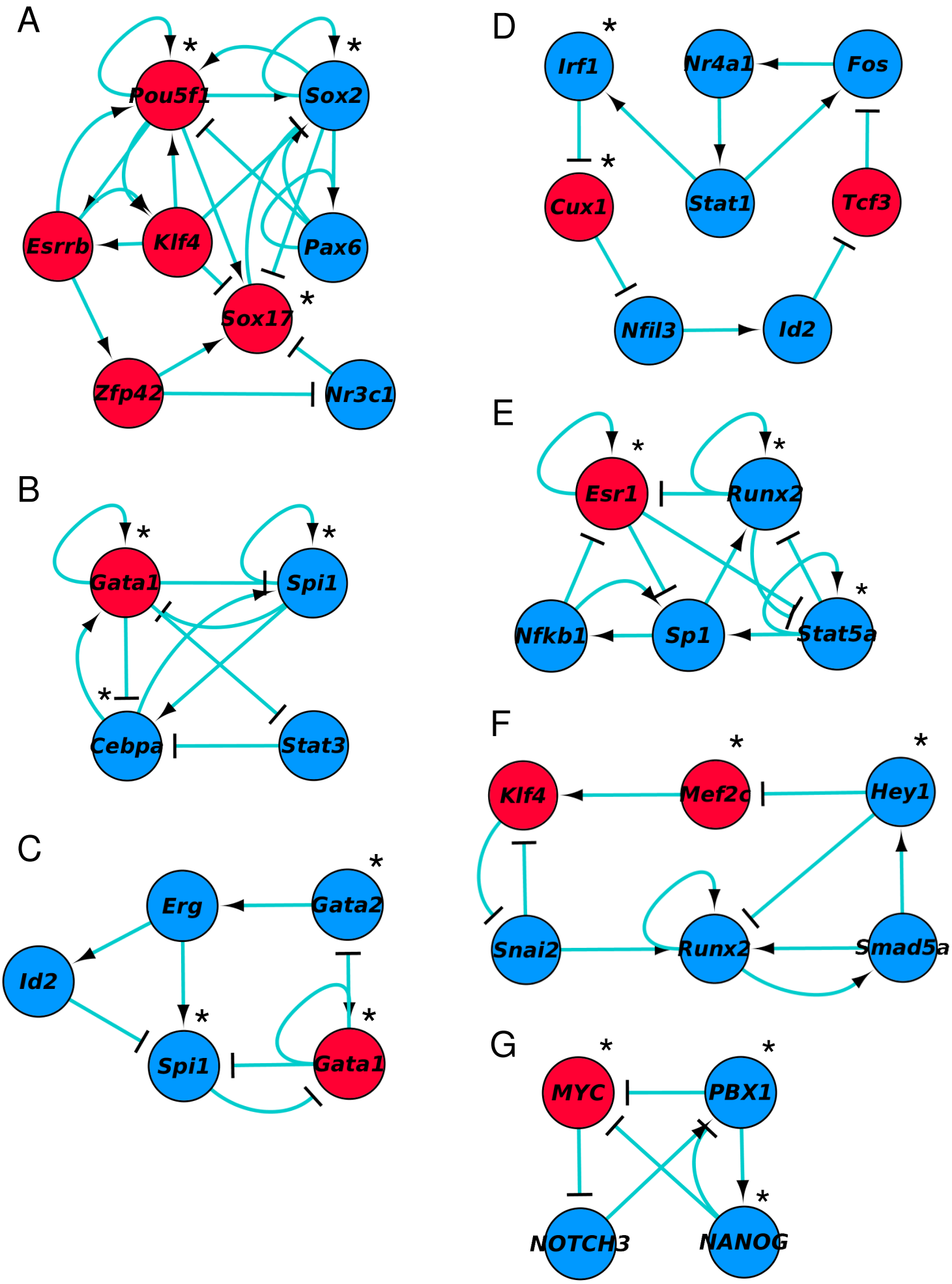

A
Pou5f1
Sox2
Esrrb
Klf4
Pax6
Sox17
Zfp42
Nr3c1
B
Gata1
Spi1
Cebpa
Stat3
C
Erg
Gata2
Id2
Spi1
Gata1
D
Irf1
Nr4a1
Fos
Cux1
Stat1
Tcf3
Nfil3
Id2
E
Esr1
Runx2
Nfkb1
Sp1
Stat5a
F
Klf4
Mef2c
Hey1
Snai2
Runx2
Smad5a
G
MYC
PBX1
NOTCH3
NANOG

Figure S2

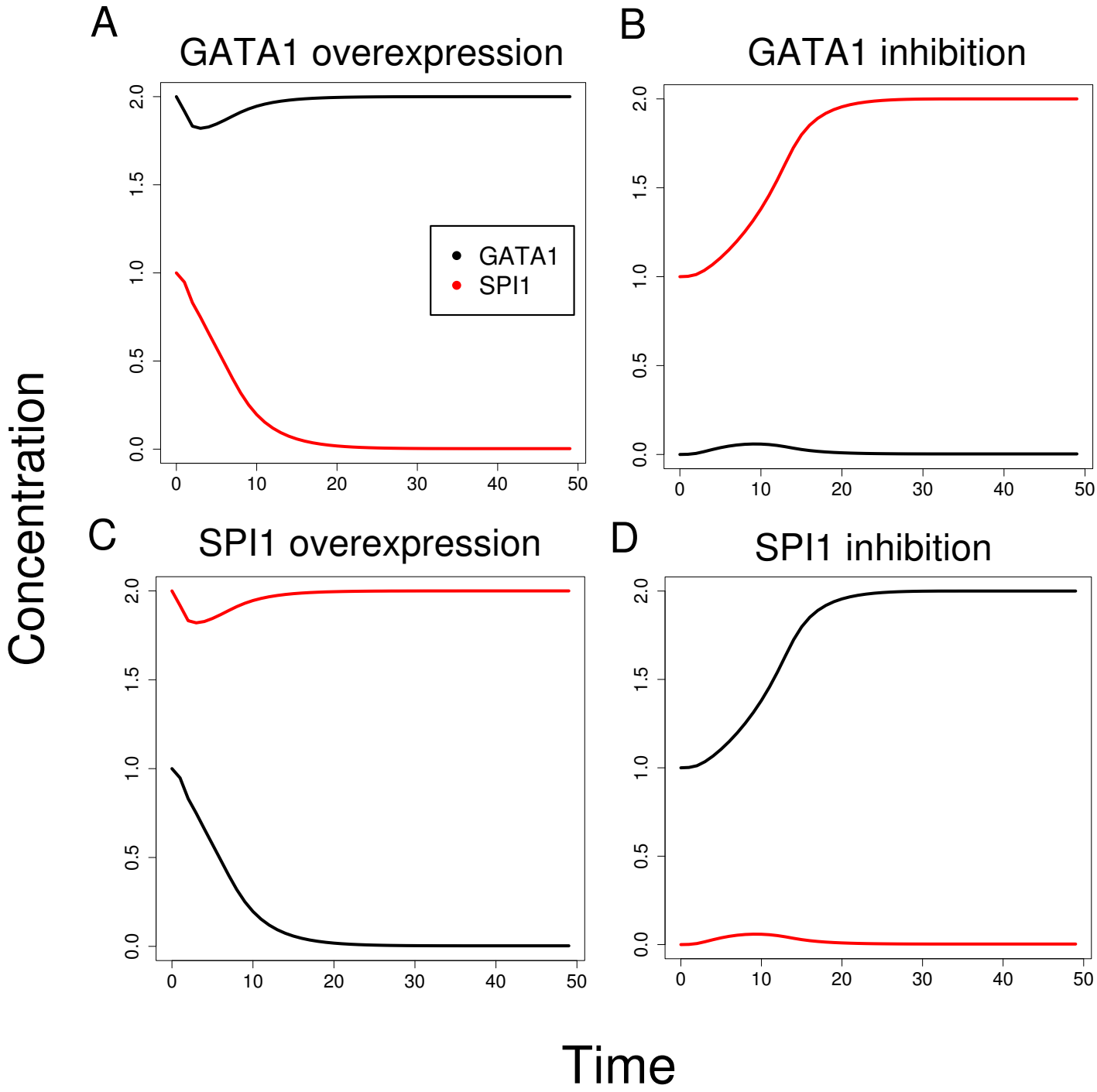

A
GATA1 overexpression
B
GATA1 inhibition
C
SPI1 overexpression
D
SPI1 inhibition
GATA1
SPI1
Concentration
Time
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
50

Figure S3

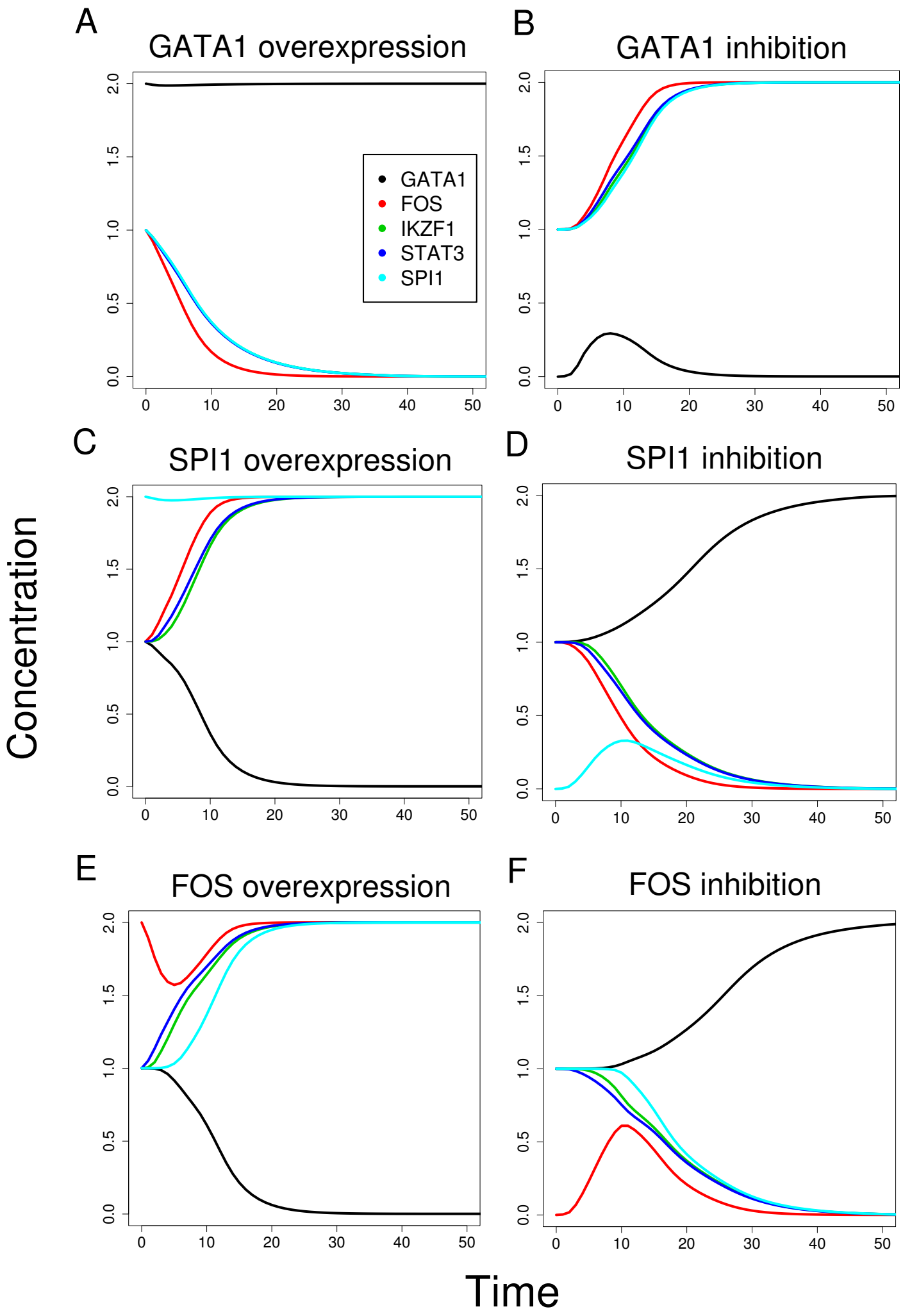
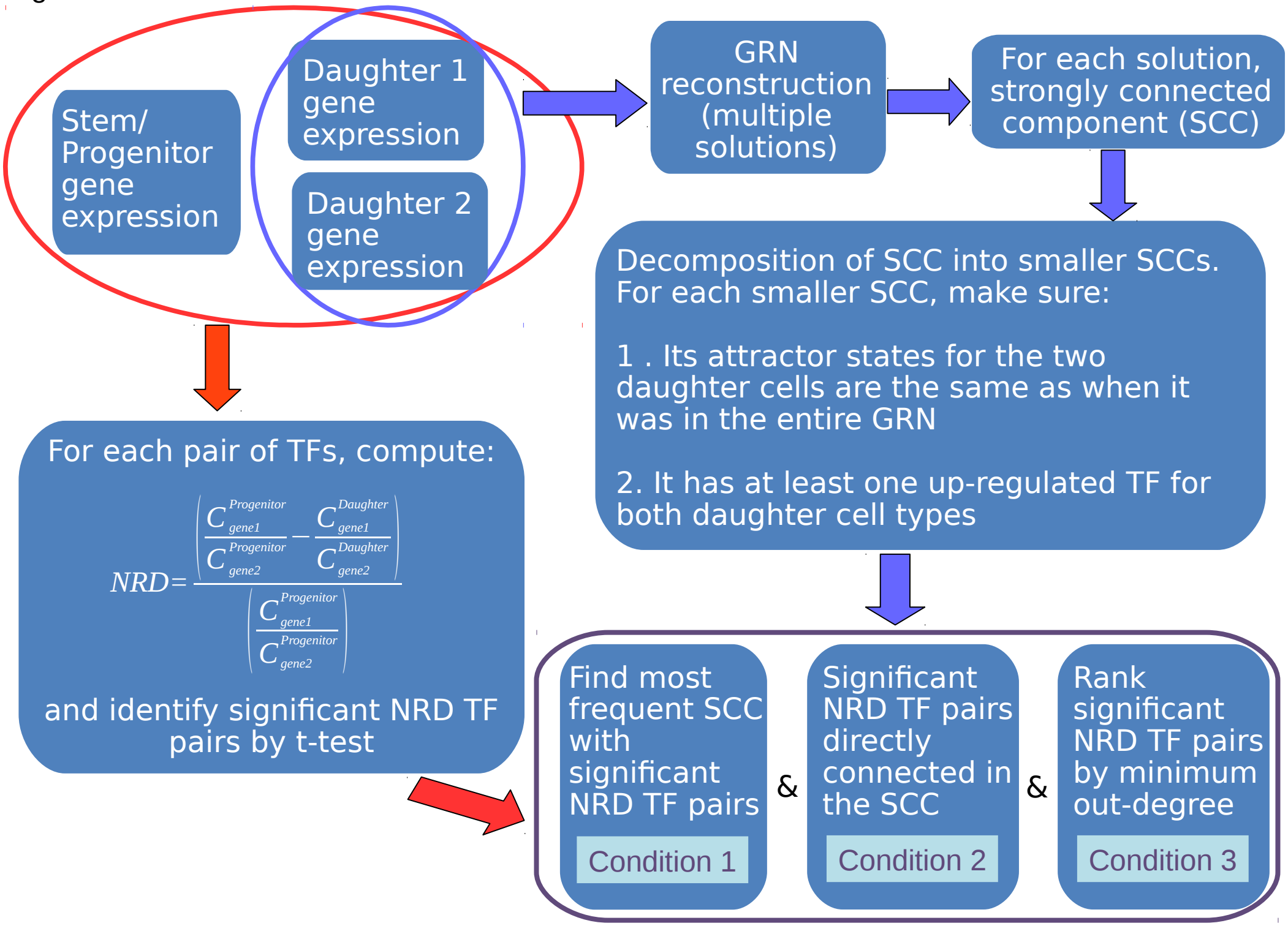

A
GATA1 overexpression
B
GATA1 inhibition
C
SPI1 overexpression
D
SPI1 inhibition
E
FOS overexpression
F
FOS inhibition
GATA1
FOS
IKZF1
STAT3
SPI1
Concentration
Time
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
50

Figure S4


Stem/ Progenitor gene expression
Daughter 1 gene expression
Daughter 2 gene expression
GRN reconstruction (multiple solutions)
For each solution, strongly connected component (SCC)
Decomposition of SCC into smaller SCCs. For each smaller SCC, make sure:
1 . Its attractor states for the two daughter cells are the same as when it was in the entire GRN
2. It has at least one up-regulated TF for both daughter cell types
For each pair of TFs, compute:
$$NRD = \frac{\left(\frac{C_{gene1}^{Progenitor}}{C_{gene2}^{Progenitor}} - \frac{C_{gene1}^{Daughter}}{C_{gene2}^{Daughter}} \right)}{\left(\frac{C_{gene1}^{Progenitor}}{C_{gene2}^{Progenitor}} \right)}$$
and identify significant NRD TF pairs by t-test
Find most frequent SCC with significant NRD TF pairs
Condition 1
&
Significant NRD TF pairs directly connected in the SCC
Condition 2
&
Rank significant NRD TF pairs by minimum out-degree
Condition 3

Supplementary Table and Figure legends

Figure S1. Predicted opposing cell fate determinant pairs and their strongly connected components in mESC, mHSC, mNSC, mMSC and hCPC systems.

Red nodes are genes upregulated in daughter 1 (mesoderm, erythroid, neuron, osteoblast and MESP1+ CPC, respectively). Blue nodes are genes upregulated in daughter 2 (ectoderm, myeloid, astrocyte, adipocyte and MESP1- CPC, respectively). Pointed arrows indicate activation, blunted arrows indicate inhibition. Network motifs of (**A**) *Sox17—Sox2* pair in mESCs, (**B**) *Gata1—Cebpa* pair in mHSCs, (**C**) *Gata1—Gata2* pair in mMSCs, (**D**) *Cux1—Irf1* pair in mHSCs, (**E**) *Esr1—Stat5a* pairs in mNSCs, and (**F**) *Hey1—Mef2c* pair in mNSCs, and (**G**) *MYC—NANOG* and *MYC—PBX1* pairs in hCPCs. * indicates predicted cell fate determinants.

Figure S2. Time trajectories of deterministic continuous simulation of *Gata1—Spi1* toggle switch (Figure 3B) upon perturbation.

Expression of GATA1 and SPI1 at stable steady states were defined as [1, 1], [2, 0] and [0, 2] in arbitrary unit for mHSC progenitors, erythroids and myeloids, respectively (see Methods). Initial conditions are (**A**) GATA1 overexpression, (**B**) GATA1 inhibition, (**C**) SPI1 overexpression, and (**D**) SPI1 inhibition. (**A**) and (**D**) reached erythroid stable steady state, whereas (**B**) and (**C**) reached myeloid stable steady state. ODEs and estimated parameters are shown in (Supplementary methods).

Figure S3. Time trajectories of continuous model for 5-gene motif containing *Gata1*, *Fos*, *Ikzf1*, *Stat3* and *Spi1* shown in Figure 3E.

Expression of GATA1, FOS, IKZF1, STAT3 and SPI1 at stable steady states were defined as [1, 1, 1, 1, 1], [2, 0, 0, 0, 0] and [0, 2, 2, 2, 2] in arbitrary unit for HSC progenitors, erythroids and myeloids, respectively (see Methods). Initial conditions are (**A**) GATA1 overexpression, (**B**) GATA1 inhibition, (**C**) SPI1 overexpression, (**D**) SPI1 inhibition, (**E**) FOS overexpression, and (**F**) FOS inhibition. (**A**), (**D**) and (**F**) reached erythroid stable steady state, whereas (**B**), (**C**) and (**E**) reached myeloid stable steady state. The progenitor state remained stable at [1, 1, 1, 1, 1] (not shown). ODEs and estimated parameters are shown in (Supplemental methods).

Figure S4. Flowchart of the method.

Three transcriptome data for stem/progenitor cell type and two daughter cell types are used for computing significant NRD TF pairs. In parallel, GRNs are reconstructed using transcriptome data for two daufhter cell types. Each GRN is then decomposed into a strongly connected component (SCC), which is then further decomposed into smaller SCCs. Finally, SCCs with significant NRD TF pairs, which satisfy the presented criteria are considered thd final predictions.

Supplemental Experimental Procedures

Microarray data processing and analysis

Micorarray data of five stem cell systems (mESC, mHSC, mNSC, mMSC and hCPC) were obtained from the following sources. For the mESC system, mESCs (GSM720404, GSM720405, GSM720406, GSM720407, GSM720408, GSM720409, GSM720410, GSM720412), ectoderms (GSM338146, GSM338150, GSM338152, GSM338156), and mesoderms (GSM747049, GSM747050, GSM747051). The data for the mHSC system was taken from (May et al., 2013), including mFDCPs (GSM1211192, GSM1211193, GSM1211194), erythroids (GSM1211279, GSM1211280, GSM1211281), and myeloids (GSM1211366, GSM1211367, GSM1211368). The data for the mNSC system consists of mNSCs (Palm et al., 2013), neurons (GSM241896, GSM241897, GSM241899, GSM241901, GSM241903, GSM241904 and GSM241922), and astrocytes (GSM241905, GSM241906, GSM241907, GSM241909, GSM241910, GSM241911, GSM241913, GSM241923 and GSM241924) (Cahoy et al., 2008). The data for mMSCs were obtained from GSM1180589, GSM1180590 and GSM1180591, osteoblasts from GSM234794, GSM234795, GSM234796, and GSM234797 (Schroeder et al., 2007) and adipocytes from GSM1254880, GSM1254881, GSM1254882 and GSM1254883 (Ralston et al., 2014). The data for the differentiation of hESCs (Day 0) into MESP1+ CPCs and MESP1- CPCs (Day 3) were taken from (Den Hartogh et al., 2015).

Raw intensity values were normalized by variance stabilising normalization using the vsn R package (Huber et al., 2002). Quantile normalization was performed when the platforms within each stem cell system were different (i.e., mESC, mNSC and mMSC systems). The differential expression analysis was performed by a moderated t-test using the limma R package (Smyth, 2004) between the ectoderm and mesoderm, erythroids and myeloids, neurons and astrocytes, osteaoblasts and adipocytes, and MESP1+ CPCs and MESP1- CPCs. Genes were binned into 30 bins by intensity and the moderated t-test was applied to each bin. The Benjamini-Hochberg multiple test correction was applied with the false discovery rate (FDR) cutoff 0.05. In all the cases genes with mean log2 fold-change less than 1 were discarded. When a gene had more than one microarray probe, the one with the highest variance across samples was used for subsequent analysis.

GRN reconstruction

Direct gene interactions between the two daughter cells in each differentiation system were retrieved from MetaCore (GeneGo Inc. (Nikolsky et al., 2005)) using differentially expressed TFs. The dates of download were between March and August of 2014. The interaction types "Transcriptional regulation" and "Binding" identified in both mouse and human were kept for the subsequent analyses. In addition, genes with node degree less than seven were discarded to focus on genes with high degrees, since these genes were not forming strongly connected

components in the initial GRN and therefore would not be in the final GRNs. The network edge (interaction) pruning was performed using the modified version of the method proposed by (Crespo et al., 2013) re-implemented in MATLAB using the genetic algorithm (ga) function. Briefly, this algorithm assumes that each cellular phenotype is a Boolean stable steady state attractor of a given network, and removes edges that are inconsistent with the Booleanized mRNA expression data. This pruning was conducted between the two daughter cell types, which resulted in GRNs whose Boolean attractor states correspond to the gene expression states of both daughter cell types. The genetic algorithm was run between 1000-1500 populations and 100 iterations. Although the current version of our algorithm does not regularize potential overfitting, we alleviate this issue by considering all best GRN solutions for subsequent analyses, although this might not fully resolve potential overfitting. The Boolean simulation was carried out using the pbn-matlab-toolbox (http://code.google.com/p/pbn-matlab-toolbox/downloads/list) using the synchronous updating scheme. The node weights were set to all 1. The logic rule was defined, so that the number of activating edges and inhibiting edges acting on a gene were compared and the one with a higher number dominates (i.e., the threshold rule). If both numbers are the same, the state was set to 0 (i.e., inhibition dominant). During this process, "unassigned" interactions (i.e., interactions without knowledge of activation or inhibition) were randomly assigned "activation" or "inhibition" and the one that yielded a better result was taken for the next generation. GRN motifs were visualized in Cytoscape (version 2.7.0) (Shannon et al., 2003).

Deterministic continuous simulation

The dynamics of relative protein abundance was modelled using the ODEs with the "OR" logic described in (Huang et al., 2007), which draw on the Michaelis-Menten formalism with Hill coefficients. The microarray expression value of each gene in the motif was ordered among the three different cell types (stem/progenitor and two daughter cells) and assigned three integers, 0, 1 or 2, for the low, intermediate and high expression values. For example, if a gene has log2 gene expression values 6, 8 and 10 for the daughter cell 1, progenitor, and daughter cell 2, then the integers 0, 1 and 2 were assigned to each stable steady state, respectively. Then the parameters were estimated by equating the ODEs to 0 at these three stable steady states. Note, we did not impose any constraint on the dynamics of the system, since the purpose of this simulation study is to illustrate that the predicted GRN motif can, upon perturbation of its genes, exhibit the expected binary differentiation dynamics from the steady state corresponding to the stem/progenitor cell type, Since this problem is intractable and it is infeasible to explore the entire parameter space, we used MATLAB's "fmincon" function (interior-point method), which was combined with the "GlobalSearch" function. The initial parameters were all set to 1 and the parameter boundary was set between 0.01 and 20. The solutions to ODEs were approximated by the 1st Taylor series using the MATLAB "taylor" function. The simulation was carried out using the Systems Biology Toolbox for MATLAB (Schmidt and Jirstrand, 2006). The "ode23s" function was used for solving ODEs.

The set of ODEs for the *Gata1-Spi1* toggle switch (Figure 3B) is,

$$\frac{d[GATA1]}{dt}=\frac{a_{11}[GATA1]^n}{K_1^n+[GATA1]^n}+\frac{a_{12}K_2^n}{K_2^n+[SPI1]}-\gamma_1[GATA1]$$

$$\frac{d[SPI1]}{dt}=\frac{a_{22}[SPI1]^n}{K_2^n+[SPI1]^n}+\frac{a_{21}K_1^n}{K_1^n+[GATA1]}-\gamma_2[SPI1]$$

where n = 7.4207, a_{11} = 4.824, a_{12} = 4.8712, a_{22} = 4.824, a_{21} = 4.8712, K_1 = 0.91729, K_1 = 0.91729, γ_1 = 4.8403, γ_2 = 4.8403. The set of ODEs for the motif shown in (Figure 3E) is,

$$\frac{d[GATA1]}{dt}=\frac{a_{11}[GATA1]^n}{K_{11}^n+[GATA1]^n}+\frac{a_{13}K_{13}^n}{K_{13}^n+[IKZF1]}+\frac{a_{15}K_{15}^n}{K_{15}^n+[SPI1]}-\gamma_1[GATA1]$$

$$\frac{d[FOS]}{dt}=\frac{a_{21}K_{21}^n}{K_{21}^n+[GATA1]^n}+\frac{a_{24}[STAT3]^n}{K_{24}^n+[STAT3]^n}+\frac{a_{25}[SPI1]^n}{K_{25}^n+[SPI1]^n}-\gamma_2[FOS]$$

$$\frac{d[IKZF1]}{dt}=\frac{a_{31}K_{31}^n}{K_{31}^n+[GATA1]^n}+\frac{a_{34}[STAT3]^n}{K_{34}^n+[STAT3]^n}-\gamma_3[IKZF1]$$

$$\frac{d[STAT3]}{dt}=\frac{a_{41}K_{41}^n}{K_{41}^n+[GATA1]^n}+\frac{a_{42}[FOS]^n}{K_{42}^n+[FOS]^n}-\gamma_4[STAT3]$$

$$\frac{d[SPI1]}{dt}=\frac{a_{51}K_{51}^n}{K_{51}^n+[GATA1]^n}+\frac{a_{55}[SPI1]^n}{K_{55}^n+[SPI1]^n}-\gamma_5[SPI1]$$

where n = 12.311, a_{11} = 4.5282, a_{13} = 2.8374, a_{15} = 2.8374, a_{21} = 7.1558, a_{24} = 6.5507, a_{25} = 6.5507, a_{31} = 4.2685, a_{34} = 6.8057, a_{41} = 4.2684, a_{42} = 6.8064, a_{51} = 4.2683, a_{55} = 6.805, K_{11} = 1.2849, K_{13} = 1.1608, K_{15} = 1.1608, K_{21} = 1.082, K_{24} = 1.0417, K_{25} = 1.0417, K_{31} = 1.1068, K_{34} = 1.0608, K_{41} = 1.1046, K_{42} = 1.0597, K_{51} = 1.1027, K_{55} = 1.0588, γ_1 = 5.0918, γ_2 = 10.126, γ_3 = 5.5357, γ_4 = 5.5361, γ_5 = 5.5353. All the models described above remained in the stable steady states corresponding to the three cell types. It is worth noting that there may exist different other sets of parameters, apart from the ones we showed here, that also reproduce the three stable states and the correct dynamics. Thus, our model provides only a qualitative analysis of the state-space of the motifs.

Pseudo-code for identification of key GRN motifs

1. Decompose strongly connected components (SCCs) into smaller SCCs

Best GRN solutions are the result from the previous step of GRN pruning, which gives rise to

```
# multiple equally best solutions
get all best GRN solutions;

loop through each best GRN solution
    find largest SCC;

    loop through each node in the SCC
        find n shortest paths starting from the node and coming back to it (i.e., feedback loops);
        from the GRN solution, extract all edges among the nodes in each shortest path (i.e.,
        decomposed SCC;

        save the decomposed SCCs;

    end loop

    # Discard duplicated topologically identical, decomposed SCCs
    get unique decomposed SCCs;

    loop through each decomposed SCC
        if (the SCC contains at least one node whose Booleanized gene expression state is up for
        both daughter cell types):

            compute Boolean attractors of the SCC with the initial states being the Booleanized gene
            expression states of the two daughter cell types;

            if (the computed two attractor states of the nodes in the SCC are identical to the
            computed two attractor states of the entire GRN solution):
                if (the computed two attractor states of the nodes in the SCC are identical to the
                Booleanized gene expression states of the two daughter cell types):

                    keep the SCC;

                end if
            end if

        end if
    end loop

end loop


## 2. Compute the frequency of SCCs containing each significant NRD TF pair ##

get all significant NRD TF pairs;

loop through each significant NRD TF pair
    loop through decomposed SCCs from 1.
        if (the SCC contains both TFs):

            compute the frequency of the SCC among all the best GRN solutions;

        end if
```

end loop
end loop

Reagents and plasmids

For immunolabelling the antibodies anti-TUJ1 (BioLegend, #801201) and anti-GFAP (Millipore, #MAB3402) were used. Alexa-fluorophore-conjugated antibodies (Invitrogen, #A11031) were used as secondary antibodies. DNA was counterstained using Hoechst 33258 (Invitrogen, #62249). The following plasmids were used: pCMV-VSV-G, psPAX2 (lentiviral packaging plasmids) (Addgene), pLenti-Runx2-C-mGFP, pLenti-Esr1-C-mGFP (Origene) and pGIPZ pMB049 (Marc Buehler).

Cell culture

Primary NSCs were isolated from C57BL/6N mouse brains at embryonic day 12.5-14.5 and cultivated as described previously (Conti and Cattaneo, 2010; Conti et al., 2005). Briefly, primary NSCs were kept on poly-D-Lysine (Sigma)-coated 10-cm polystyrene tissue culture dishes in DMEM/Ham's F12 medium (PAA) supplemented with 10 ng/mL EGF (Peprotech), 10 ng/mL bFGF-2 (Peprotech), 1 x N2 (Invitrogen), L-Glutamine (PAA), and Penicillin/Streptomycin (PAA). HEK293T cells were cultivated on uncoated 10-cm polystyrene tissue culture dishes in DMEM (Sigma) supplemented with 10% heat-inactivated FCS (PAA), L-Glutamine (PAA) and Penicillin/Streptomycin (PAA).

Lentivirus production

Lentiviruses were produced using a three-plasmid transfection protocol. One day prior to transfection, HEK293T cells were seeded in 10-cm polystyrene tissue culture dishes. The next day, the lentiviral packaging plasmids pCMV-VSV-G and psPAX2 were mixed with either pGIPZ pMB049, pLenti-Runx2-C-mGFP or pLenti-Esr1-C-mGFP and the HEK293T cells were transfected with these plasmids using Fugene6 (Promega) according to manufacturer`s instructions. Three days post transfection, the supernatants were harvested and cleared from remaining cells by centrifuging for 10 min at 3,000 x g and 4 °C. The supernatant was mixed with 1/5 volume of 40% PEG and incubated overnight at 4 °C. The next day, the lentivirus was concentrated by centrifugation for 30 min at 1,500 x g and 4 °C. After removal of the supernatant, the pellet was centrifuged again for 5 min at 1,500 x g and 4 °C. The remaining supernatant was removed and the pellet was resuspended in an appropriate amount of DMEM without supplements and stored at -80 °C.

Viral transductions

For viral transduction, primary NSCs were seeded onto poly-D-Lysine-coated coverslips at a density of 25,000 cells / well. One day after seeding the virus was diluted in growth medium and added to the cells. At two and four days post transduction, half of the growth medium was

exchanged by fresh growth medium.

Immunocytochemistry

For immunocytochemical staining, cells were fixed with 4% paraformaldehyde in 120 mM PBS, pH 7.4 (4% PFA/PBS) followed by permeabilisation for 3 min at 4 °C using 0.05% Triton X-100 in PBS. Next, cells were blocked with 10% FCS in PBS for 1 h at RT and subjected to immunofluorescence staining with primary and secondary antibodies diluted in blocking solution. Images were collected with a Zeiss epifluorescence microscope and image analysis was conducted using ZEN lite (Zeiss) and Adobe Photoshop softwares.

Supplemental References

Cahoy, J.D., Emery, B., Kaushal, A., Foo, L.C., Zamanian, J.L., Christopherson, K.S., Xing, Y., Lubischer, J.L., Krieg, P.A., Krupenko, S.A.*, et al.* (2008). A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. The Journal of neuroscience : the official journal of the Society for Neuroscience *28*, 264-278.

Crespo, I., Krishna, A., Le Bechec, A., and del Sol, A. (2013) Predicting missing expression values in gene regulatory networks using a discrete logic modeling optimization guided by network stable states. *Nucleic acids research 41*, e8.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics *18 Suppl 1*, S96-104.

May, G., Soneji, S., Tipping, A.J., Teles, J., McGowan, S.J., Wu, M., Guo, Y., Fugazza, C., Brown, J., Karlsson, G.*, et al.* (2013). Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. Cell stem cell *13*, 754-768.

Nikolsky, Y., Ekins, S., Nikolskaya, T., and Bugrim, A. (2005) A novel method for generation of signature networks as biomarkers from complex high throughput data. *Toxicology letters 158*, 20-29.

Palm, T., Hemmer, K., Winter, J., Fricke, I.B., Tarbashevich, K., Sadeghi Shakib, F., Rudolph, I.M., Hillje, A.L., De Luca, P., Bahnassawy, L.*, et al.* (2013). A systemic transcriptome analysis reveals the regulation of neural stem cell maintenance by an E2F1-miRNA feedback loop. Nucleic acids research *41*, 3699-3712.

Ralston, J.C., Badoud, F., Cattrysse, B., McNicholas, P.D., and Mutch, D.M. (2014). Inhibition of stearoyl-CoA desaturase-1 in differentiating 3T3-L1 preadipocytes upregulates elongase 6 and downregulates genes affecting triacylglycerol synthesis. Int J Obes (Lond) *38*, 1449-1456.

Schmidt, H., and Jirstrand, M. (2006). Systems Biology Toolbox for MATLAB: a computational platform for research in systems biology. Bioinformatics *22*, 514-515.

Schroeder, T.M., Nair, A.K., Staggs, R., Lamblin, A.F., and Westendorf, J.J. (2007). Gene profile analysis of osteoblast genes differentially regulated by histone deacetylase inhibitors. BMC genomics *8*, 362.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research 13*, 2498-2504.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol *3*, Article3.