Supplementary Information

Gene pair signatures in cell type transcriptomes reveal lineage control

Merja Heinäniemi, Matti Nykter, Roger Kramer, Anke Wienecke-Baldacchino, Lasse Sinkkonen, Joseph Xu Zhou, Richard Kreisberg, Stuart A. Kauffman, Sui Huang and Ilya Shmulevich.

correspondence to: ishmulevich@systemsbiology.org

Other supplementary material for this manuscript:

Supplementary Tables 1-3, 5, 7-17
Online webpage at http://trel.systemsbiology.net/

Supplementary Figures and Legends

Supplementary Tables and Legends

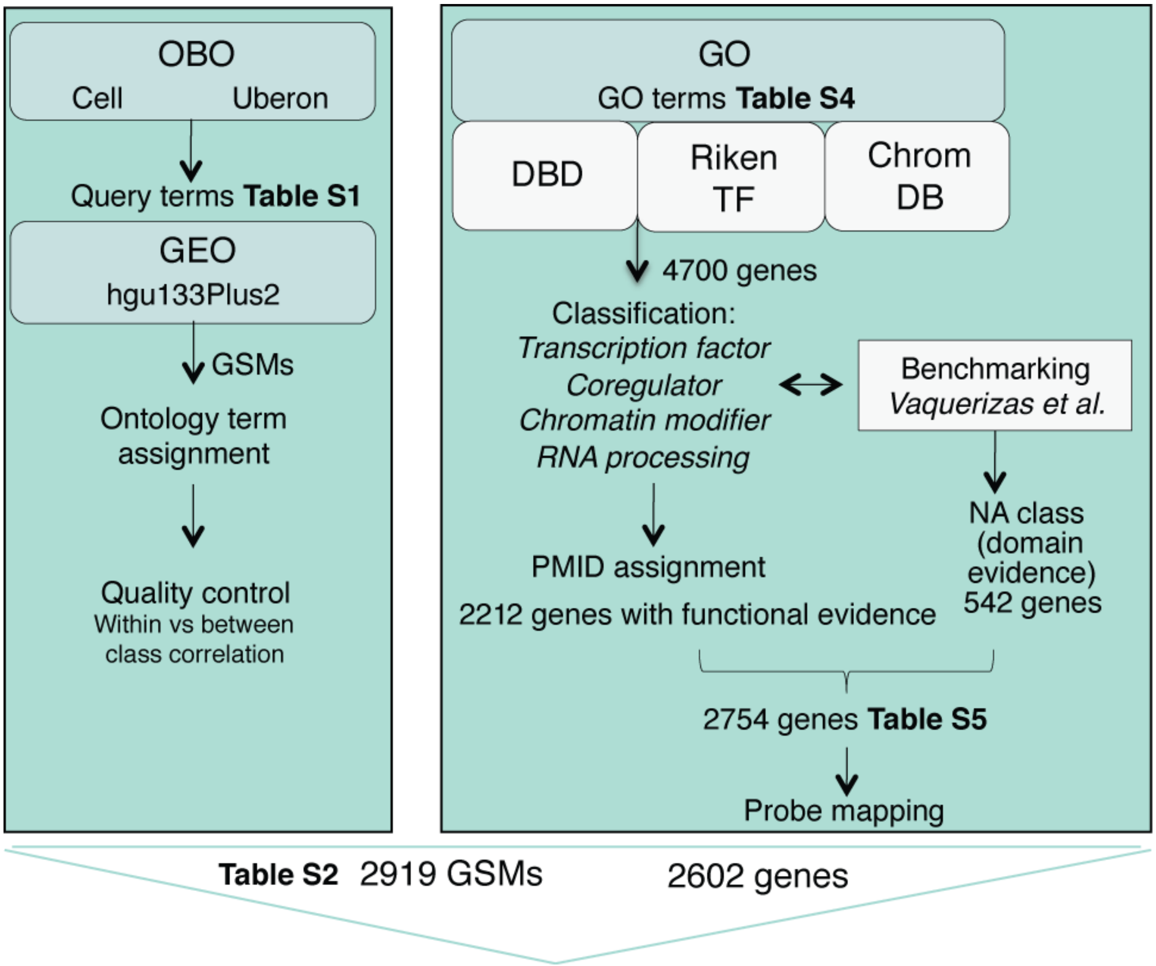
Supplementary Results

Captions for online Supplementary Tables

Table S3. The order of cell types as it appears in heat maps presented *(xls file available online for download)*
Table S5. Functional evidence for role in transcription regulation found in the gene set curation *(xls file available online for download)*
Table S7. Identification of candidate toggle pairs *(xls file available online for download)*
Table S8. Rank-based differential expression analysis using RCoS *(xls file available online for download)*
Table S9. Rank-based differential expression analysis using RDAM *(xls file available online for download)*
Table S10. Public ChIPseq datasets used *(xls file available online for download)*
Table S11. Genomic region enrichment results for GATA1 ChIP-seq data *(xls file available online for download)*
Table S12. Genomic region enrichment results for TAL1 ChIP-seq data *(xls file available online for download)*
Table S13. Genomic region enrichment results for SPI1 ChIP-seq data *(xls file available online for download)*
Table S14. Genomic region enrichment results for EBF1 ChIP-seq data *(xls file available online for download)*
Table S15. Genomic region enrichment results for GATA3 ChIP-seq data *(xls file available online for download)*
Table S16. Mouse knockout phenotypes of *Gata1*, *Tal1*, *Sfpi1*, *Ebf1* and *Gata3 (xls file available online for download)*
Table S17. Additional microarray data used for validation *(xls file available online for download)*

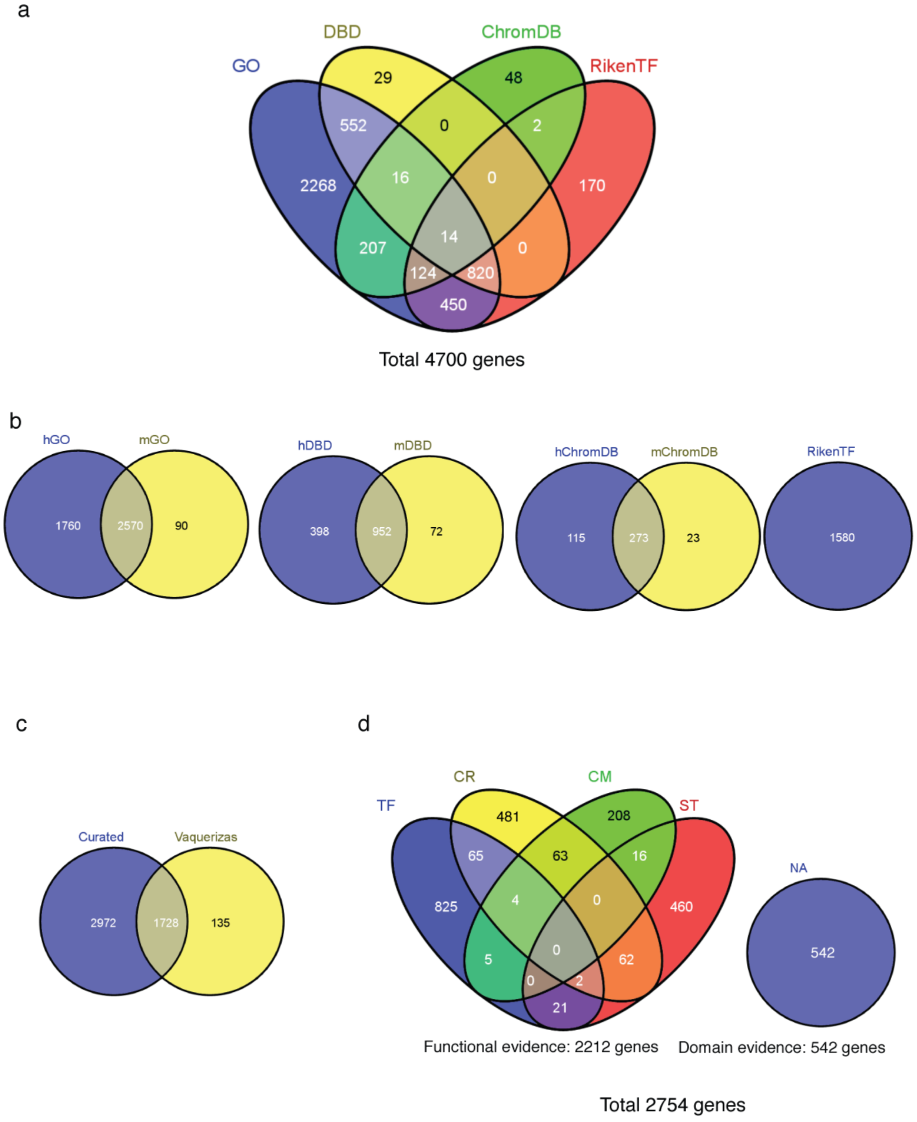
References

Supplementary Fig. 1. Data set assembly to capture changes in the transcriptional network between different cell types.



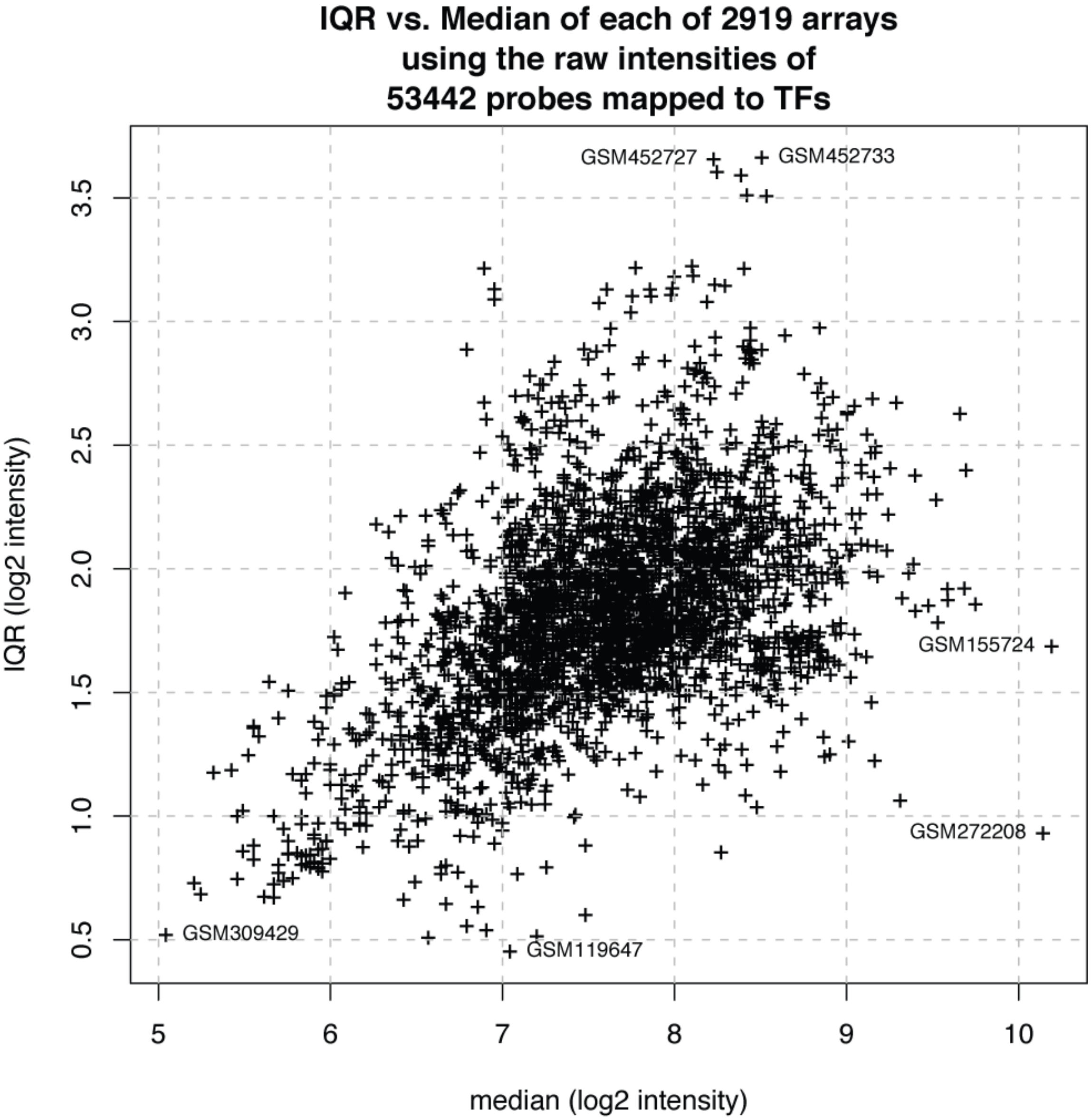
The different steps to assemble the microarray data and the gene data are illustrated. The cell[46] (cell.obo) and anatomical site[47] (uberon.obo) ontologies from the OBO database (http://www.obofoundry.org/) were used to generate a query over the GEO microarray sample annotations. The retrieved samples were assigned a corresponding ontology term by manual curation and this was validated using Spearman's rho rank correlation. Samples with higher correlation to other cell type classes than their annotated origin were discarded resulting in 2919 microarrays. The transcription regulating gene set was assembled from GO[48], DBD[49], Riken TF[50] and ChromDB[51] databases. The dataset was compared against an independent data set[52] and went through manual curation to assign Pubmed ID indicating a function in four different classes. Functional evidence was attributed to 2212 genes. This set was complemented by a domain annotation –based[52] class resulting in 2754 genes. After probe mapping 2602 genes remained that were quantified from the hgu133Plus2 microarrays.

Supplementary Fig. 2. Data source overlaps for the gene set collection.



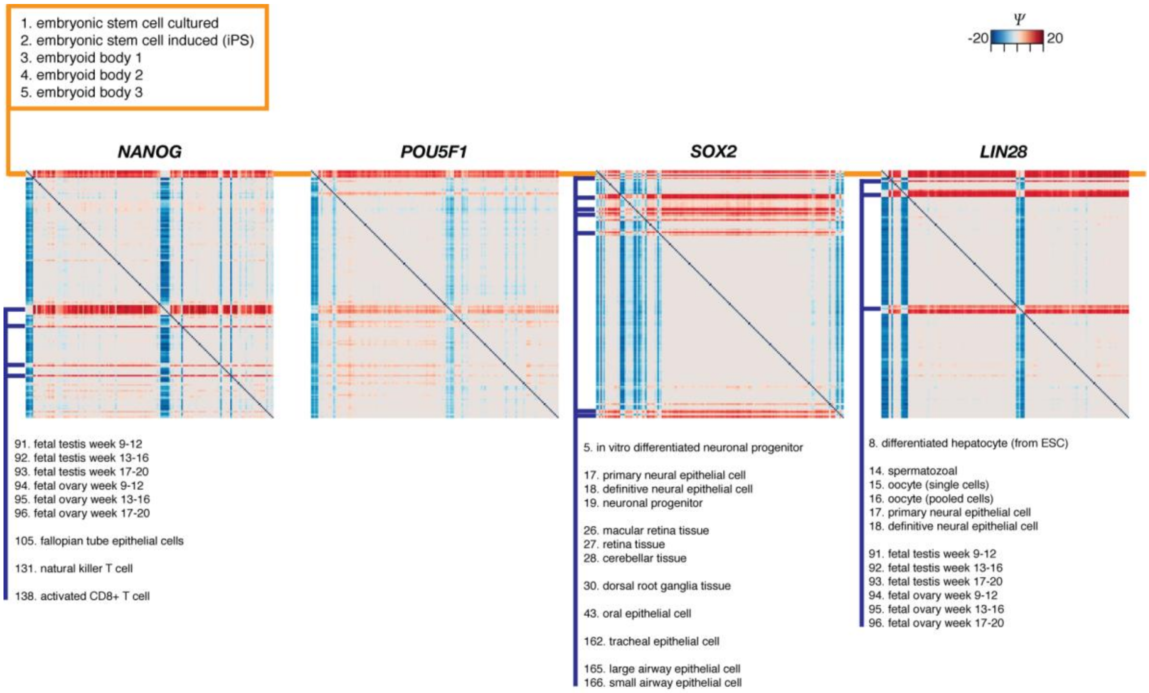
The overlap of genes (in total 4700) that were extracted from the four indicated data sources is shown in (**a**). When available, data was extracted for both human and mouse genes. The overlap between species shows (**b**) that this retrieved non-redundant entries. Comparison of the collected gene set with the census for transcription factors compiled using domain-based evidence[52] is shown in (**c**). (**d**) Division of genes to different functional classes was done by manual curation of evidence available for a function in transcription. TF = transcription factor, CR = coregulator, CM = chromatin modifier, ST = RNA processing (splicing, transcription, processing). This resulted in a curated set of 2212 genes with associated Pubmed ID references. Additionally, genes with strong domain-based evidence[52] were included as the NA category as no classification for function is available from literature.

Supplementary Fig. 3. Distribution of probe signals across all microarrays.



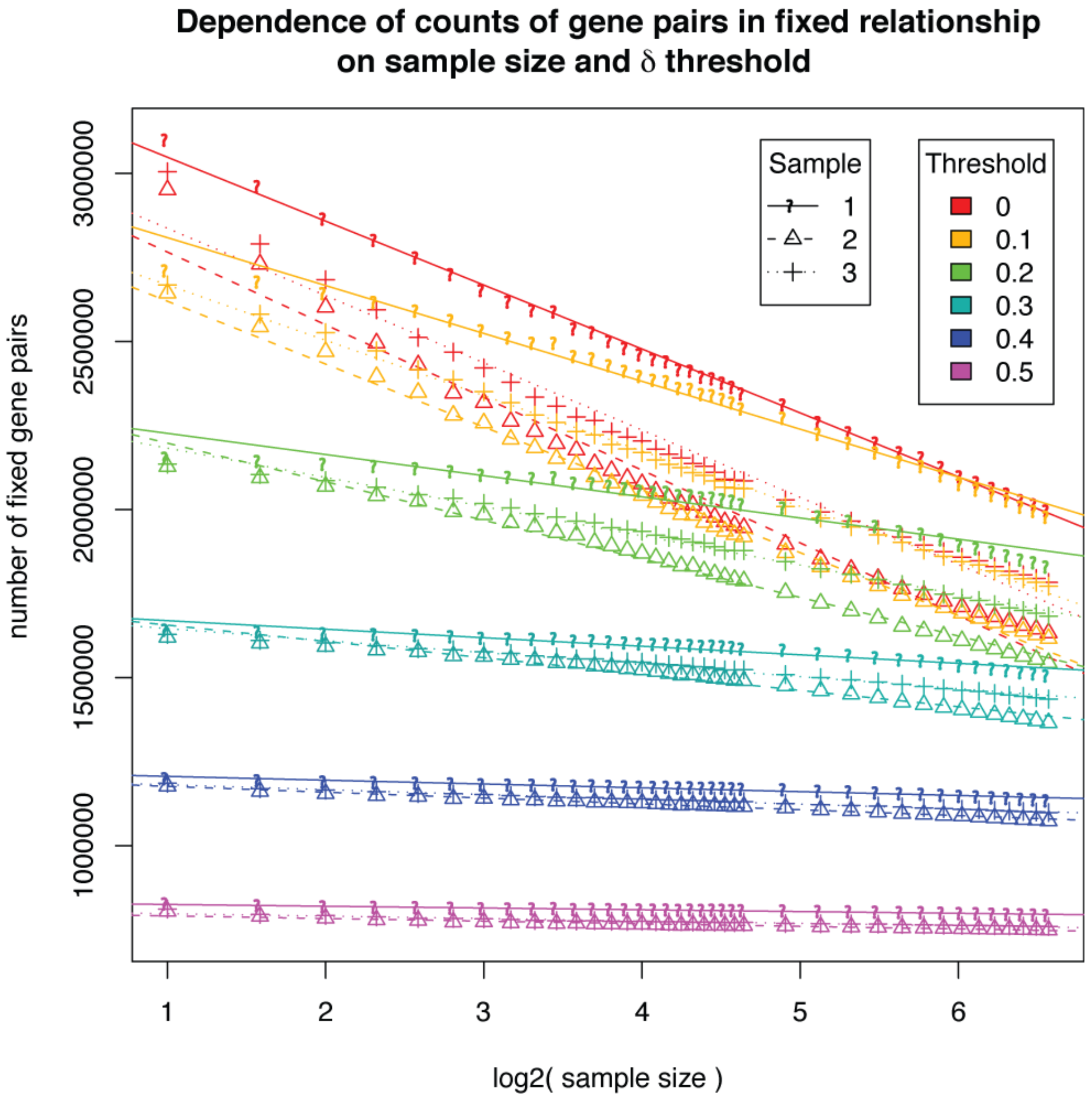
The raw intensities of 53442 probes mapped to TFs from 2919 arrays were extracted. The ratio of the 3rd quartile to the 1st quartile is plotted against the median of the probe signals to characterize the distribution of expression values non-parametrically. Each point in the plot represents one microarray.

Supplementary Fig. 4. Reversal participation of the pluripotency-inducing gene set is highest in pluripotent cells and samples of early fetal origin.



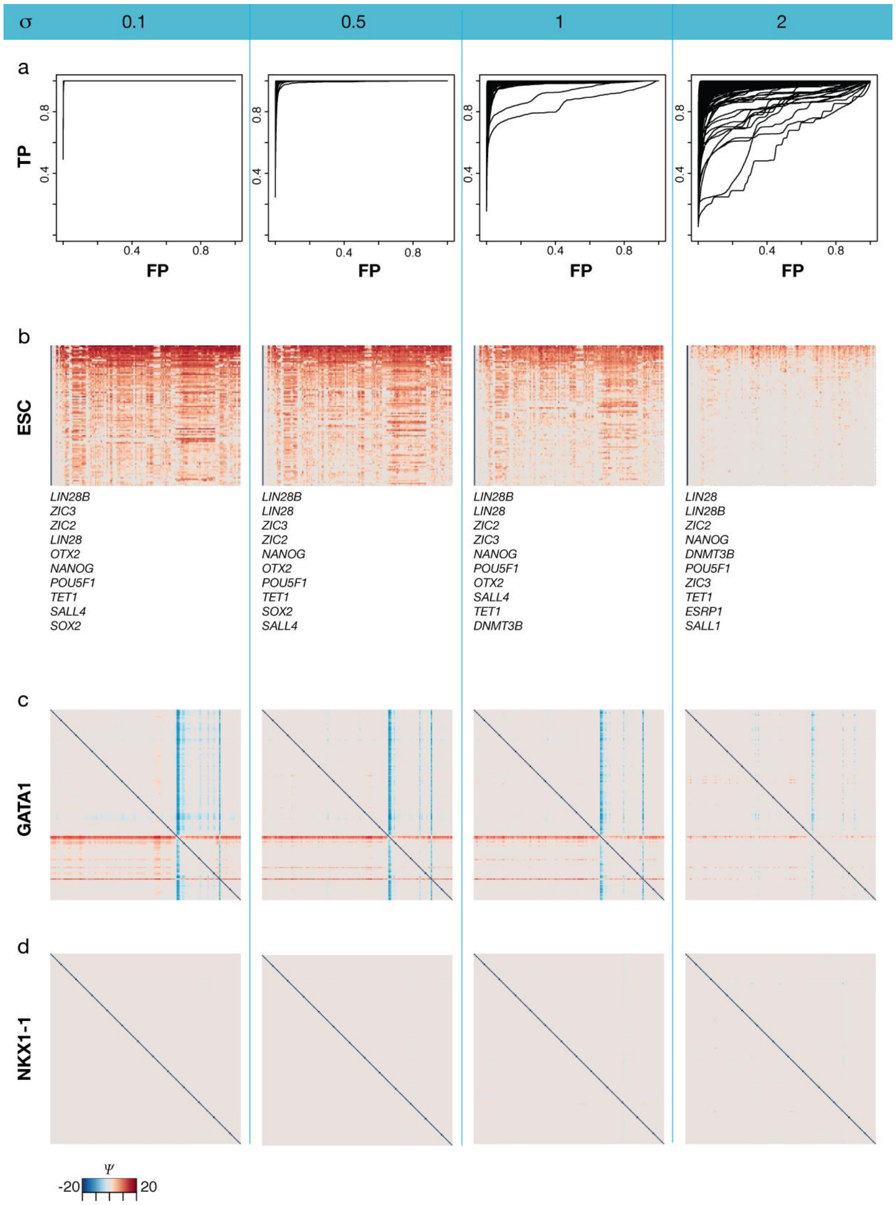
Reversal participation Ψ gene portraits. The Ψ value for gene *g* reflects the number of reversal pairs that involve *g* and their maximum Δ value. As a consequence, specific gene pair configuration (in pairs with *G)* will be reflected by a high score (dark red or blue) and this highlights restricted high or low expression of *g* in a particular cell type (in a row). Reversal participation Ψ gene portraits of the pluripotency-inducing[11] genes *NANOG*, *POU5F1*, *SOX2* and *LIN28* are shown. The cell types with highest row sums reflect early-development restricted expression.

Supplementary Fig. 5. The pair reversal score behaves robustly even in low-sample number groups.



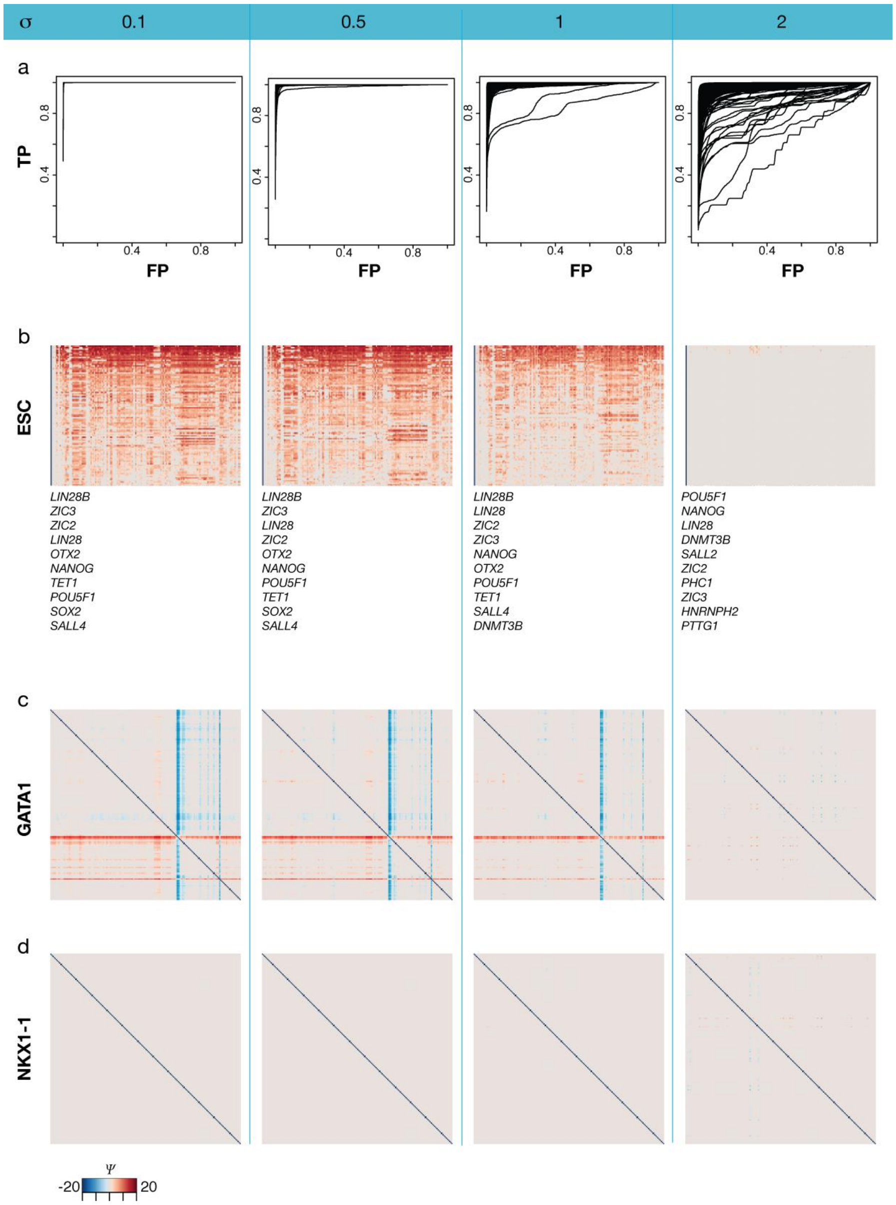
The effect of decreasing sample size on the fixed pair configuration at different δ value cut-offs was tested using random sampling from three cell types with highest number of samples in the original dataset (ductal breast epithelial cells, skeletal muscle tissue, monocytes). The pair states of 100 random array sets generated from each cell type are plotted against number of samples (notice the logarithmic scale on the x-axis).

Supplementary Fig. 6. Simulation of the effect of random fluctuations in gene pair ranks on the reversal participation results using a zero mean Gaussian noise injection model.



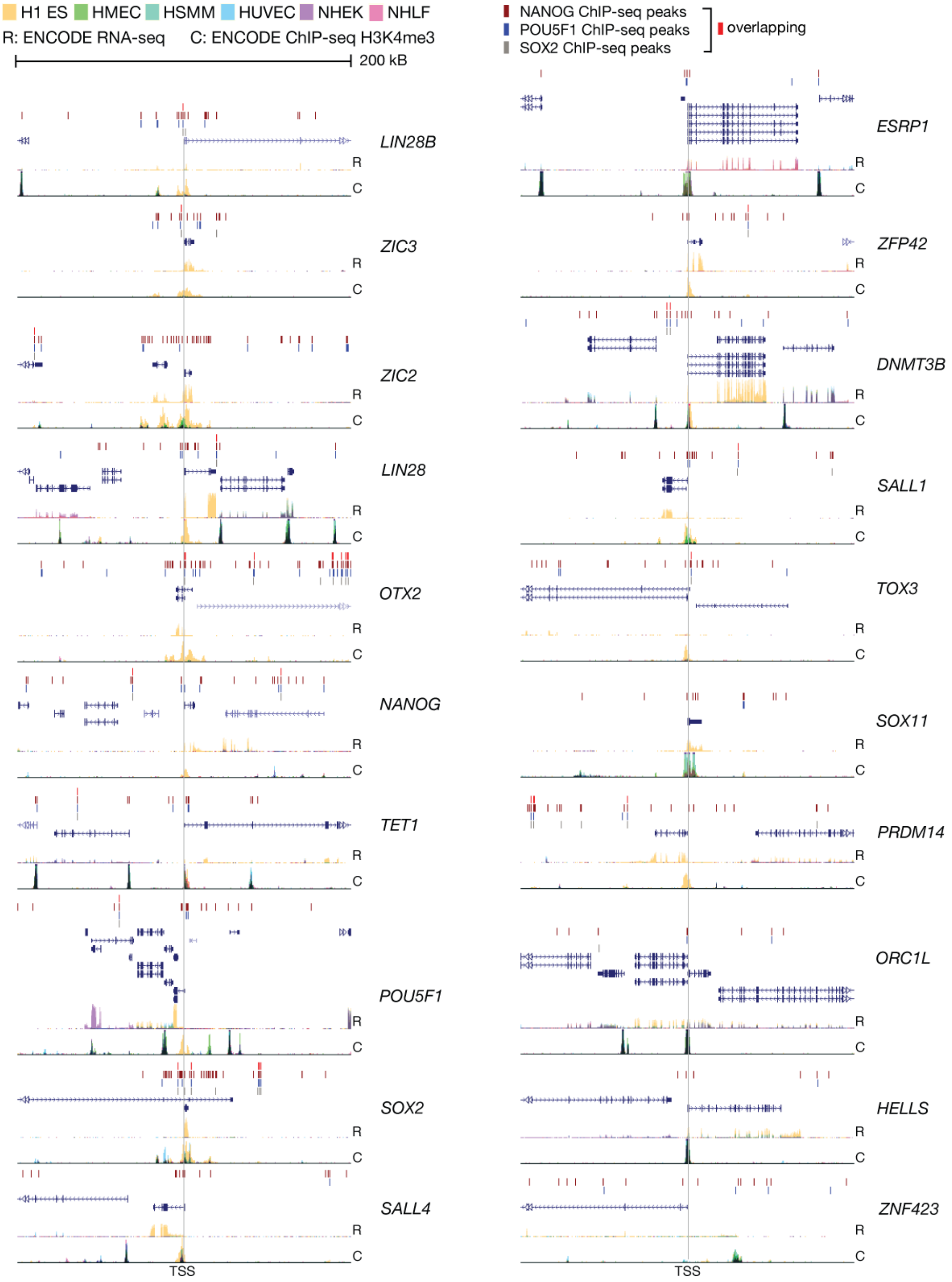
The cell portrait ranking of top 20 most lineage specific genes was taken to represent the gold standard for each 166 cell types. Simulations (*n* = 100) at different noise levels (standard deviation values are indicated in the figure) were performed and displayed as ROC curves (a separate line for each cell type) in (**a**). Representative examples are shown in **b-d** to demonstrate that even high noise levels do not easily cause loss of signal i.e. false negatives (the lineage-specific signal persists for *GATA1* in (**c**) or false positive signal (no random pattern emerges) for a gene that initially lacks signal shown in panel (**d**).

Supplementary Fig. 7. Simulation of the effect of random fluctuations in gene pair ranks on the reversal participation results using a zero mean Lapplacian noise injection model.



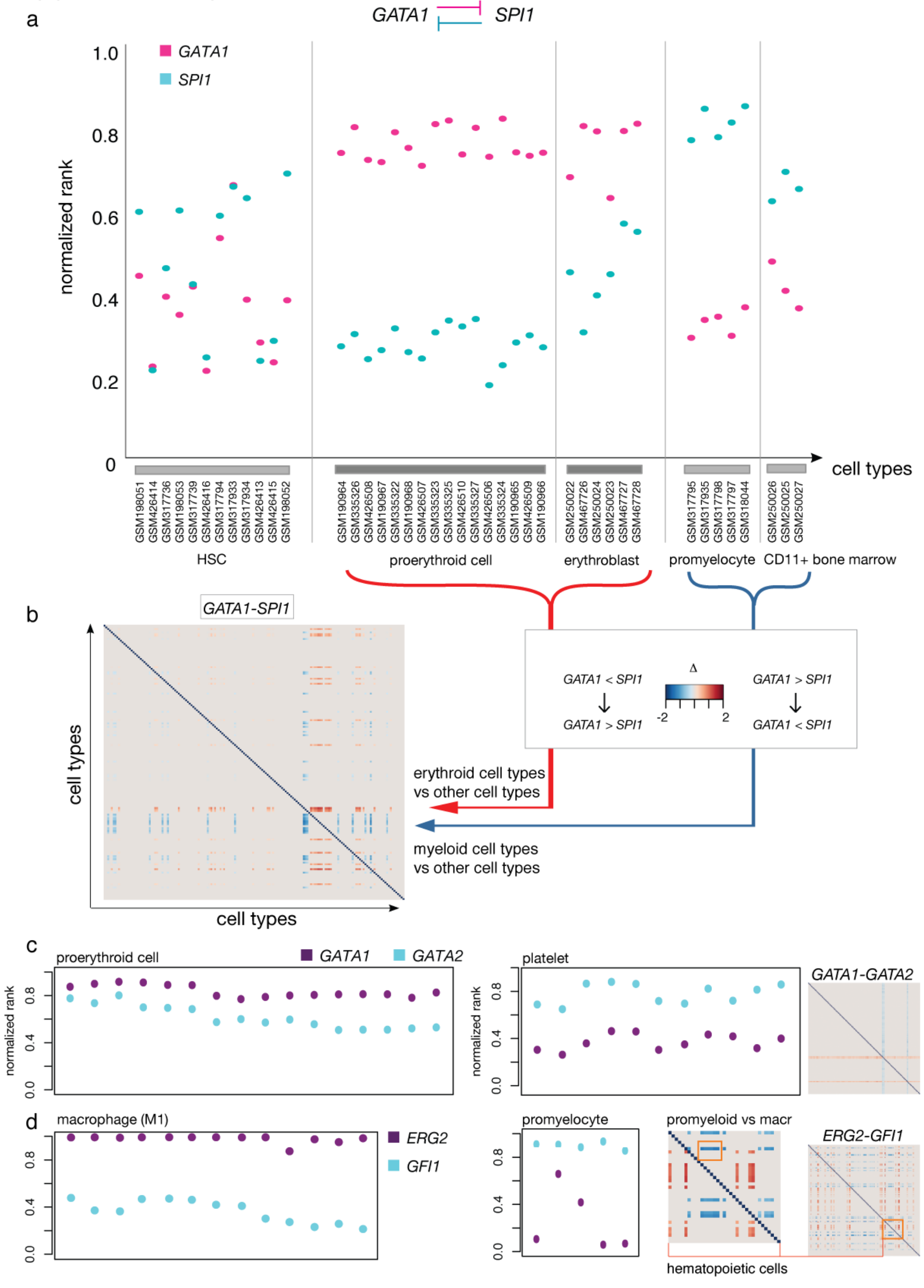
The cell portrait ranking of top 20 most lineage specific genes was taken to represent the gold standard for each 166 cell types. Simulations (*n* = 100) at different noise levels (standard deviation values are indicated in the figure) were performed and displayed as ROC curves (a separate line for each cell type) in (**a**). Representative examples are shown in **b-d** to demonstrate that even high noise levels do not easily cause loss of signal i.e. false negatives (the lineage-specific signal persists for *GATA1* in (**c**) or false positive signal (no random pattern emerges) for a gene that initially lacks signal shown in panel (**d**).

Supplementary Fig. 8. NANOG, POU5F1 and SOX2 occupancy at the ESC restricted gene TSS regions.



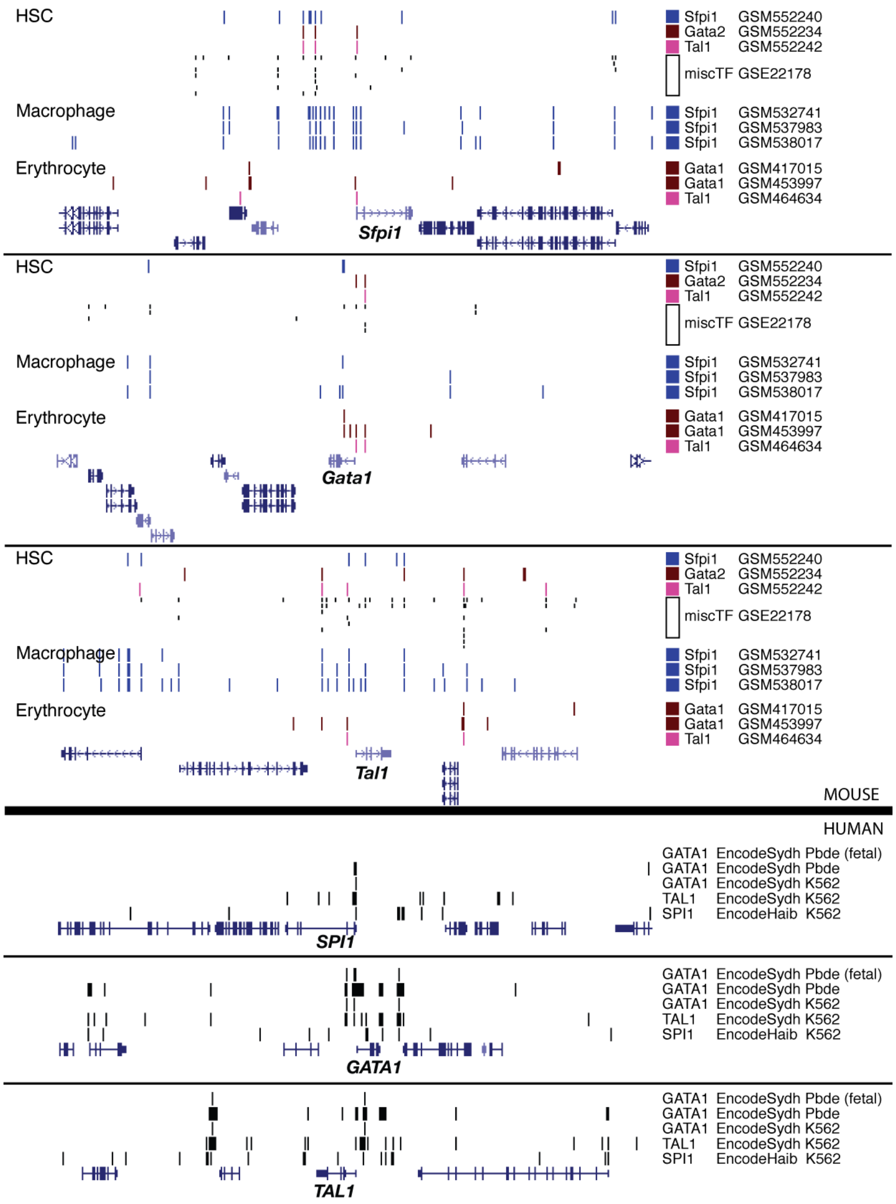
The peak lists from different ChIP-seq experiments that measured genome-wide occupancy of the key ES TFs from human ESCs were combined[53,54]. Occupancy of each TF and overlapping binding sites for all three TFs are shown from a 200 kB region centered at the respective TSS. The ENCODE[12] active promoter marker (H3K4me3) ChIPseq and RNAseq results for this extended region are also displayed. The six normal ENCODE cell types shown are H1 ES: human embryonic stem cell line H1, HMEC: breast epithelial cell, HSMM: skeletal muscle myoblast, HUVEC: umbilical vein endothelial cell, NHEK: epithelial keratinocyte, NHLF: lung fibroblast. RNA-seq data is available from H1 ES, HUVEC and NHEK cells. The presence of overlapping binding sites is statistically significant (*p*-value of 2.205x10^{-4} calculated using hypergeometric distribution).

Supplementary Fig. 9. Gene pair expression reversal exemplified by known toggle switch gene pairs.



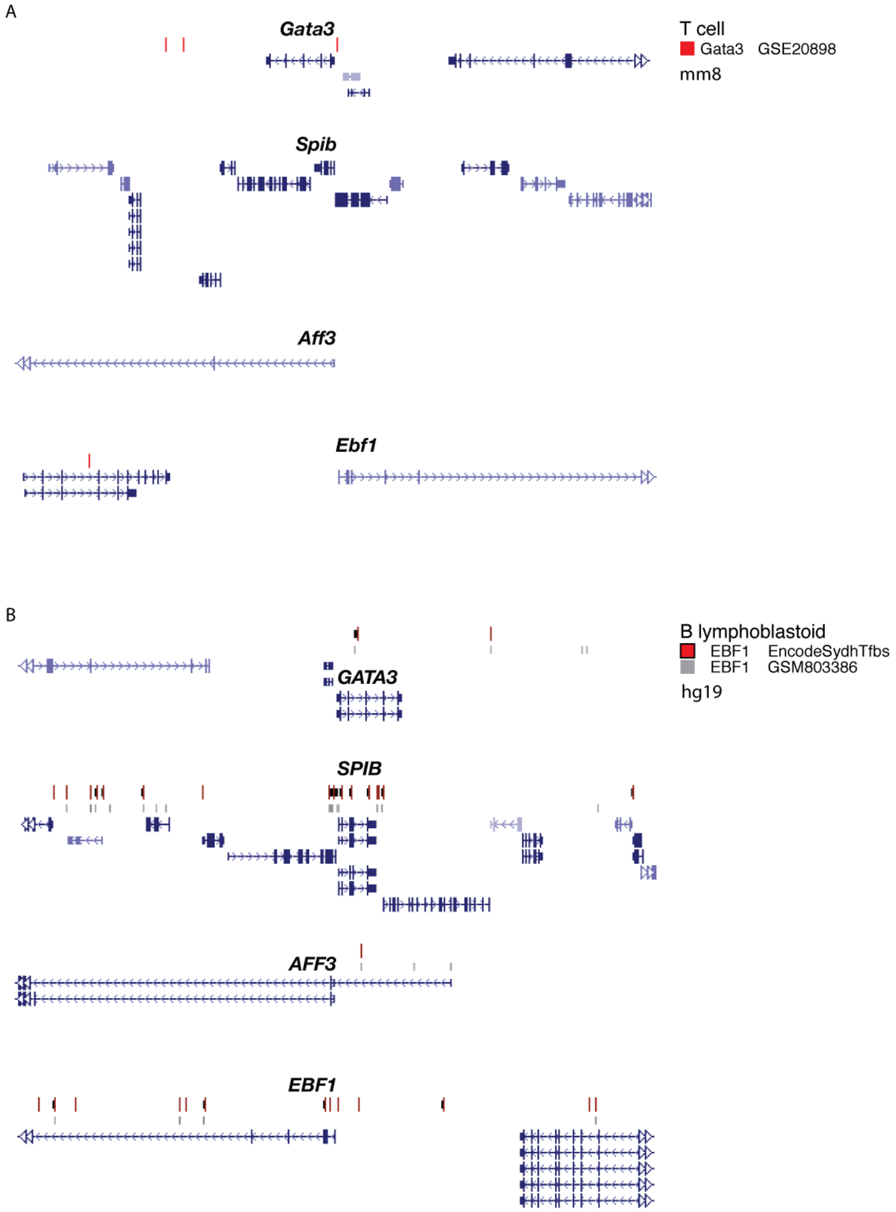
The ranks of *GATA1* and *SPI1* are plotted in (**a**) from each microarray sample that corresponds to the HSC, proerythroid, erythroblast and promyeloid samples. Higher rank corresponds to higher expression in the given cell type. (**b**) Gene pair reversal plot. The reversal behavior of the GATA1-SPI1 gene pair quantified for all pair-wise comparisons of *n* = 166 cell types is shown as an *n* x *n* symmetric matrix. The *Δ* value, indicating the extent of reversal behavior is represented by the color in the heat map. Red tones indicate that the pair configuration changes from *GATA1* >> *SPI1* in the first cell type of a comparison pair ("row-to-column comparison") to *GATA1* << *SPI1* in the second cell type. A reversal of the gene pair configuration in the opposite direction in cell type comparisons are indicated in blue shades. Unlike the familiar heat maps representing similarity (e.g. correlation) between cell type transcriptomes that exhibit diagonal symmetry note here the characteristic asymmetry of colors (but symmetry of shades) indicating gene expression reversal of the gene pair {*TF1*, *TF2*}. For order of cell types refer to Supplementary Table 3 online. Cell types belonging to the erythroid and myeloid lineages are indicated by arrows. Similarly as in (**a**), the ranks of *GATA1*, *GATA2*, *EGR2* and *GFI1* are plotted from arrays belonging to the cell types indicated in (**c**) and (**d**) and the reversal behaviour across all cell type comparisons (as in (**b**)) is shown in the gene pair plot.

Supplementary Fig. 10. Putative cross-regulatory and auto-regulatory interactions of the erythroid-myeloid toggle candidates identified from ChIP-seq datasets.



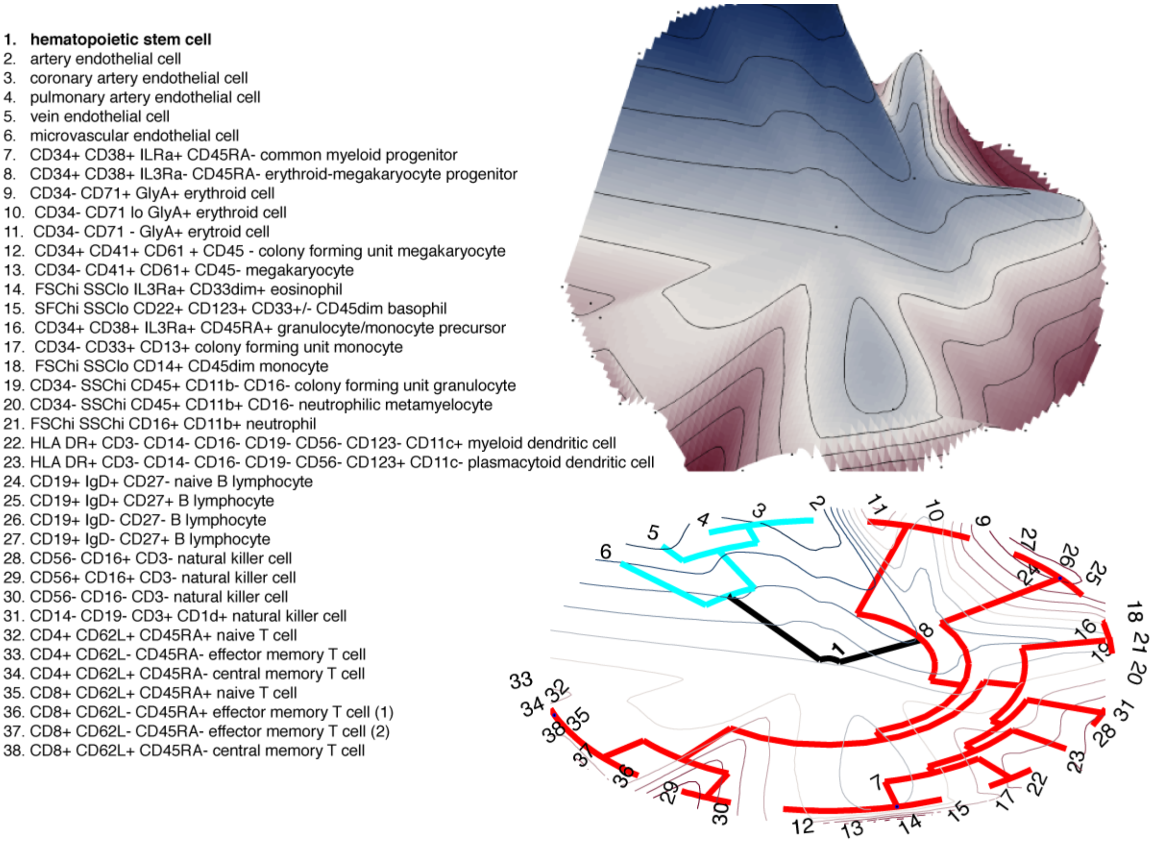
The peak lists from different ChIP-seq experiments that measured genome-wide occupancy of the candidate toggle TFs from mouse HSC, erythroid or myeloid cells were combined (see Supplementary Table 10). Binding sites for the candidate toggle switch circuit pairs are shown from a 200 kb region centered at the respective TSS.

Supplementary Fig. 11. Putative cross-regulatory and auto-regulatory interactions of the lymphoid toggle candidates identified from ChIP-seq datasets.



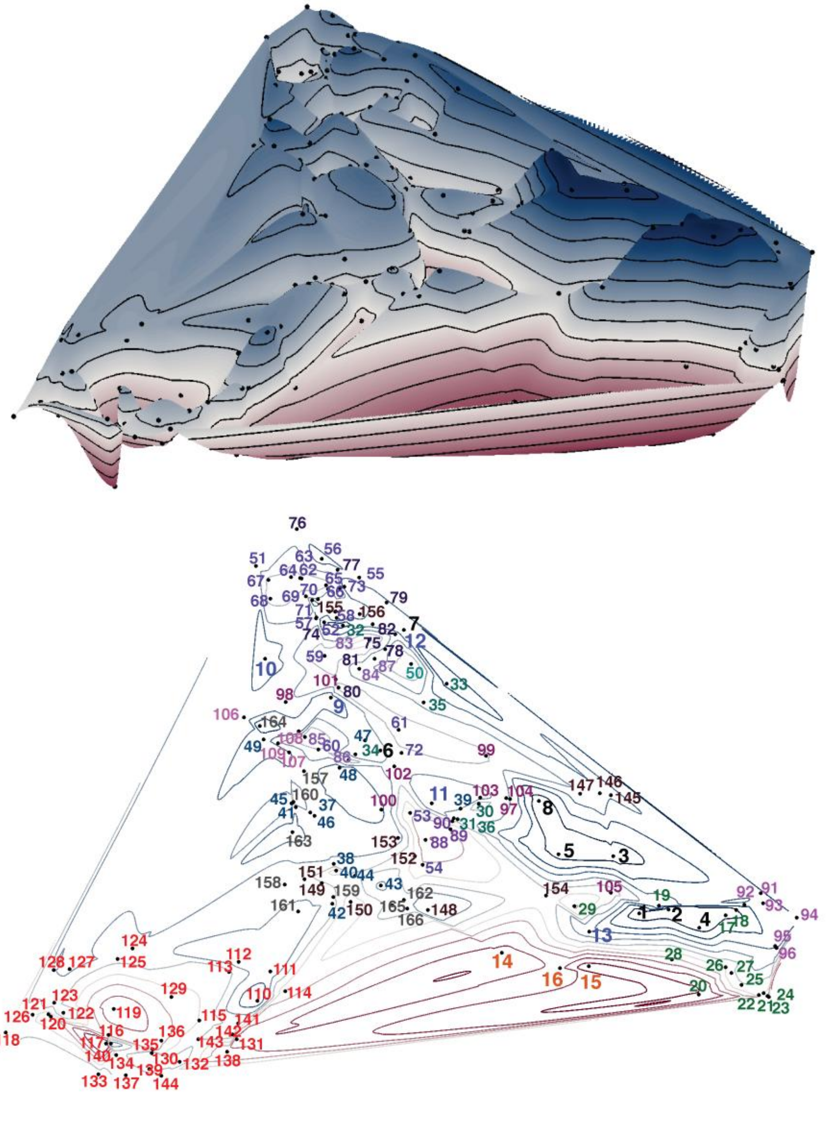
The peak lists from different ChIP-seq experiments that measured genome-wide occupancy of the candidate toggle TFs from mouse T cells or human lymphoblastoid cells were combined (see Supplementary Table 10). Binding sites for the candidate toggle switch circuit pairs are shown from a 200 kb region centered at the respective TSS.

Supplementary Fig. 12. Lineage relationships among hematopoietic and endothelial cell types reproduced from independent microarray data.


1. hematopoietic stem cell
2. artery endothelial cell
3. coronary artery endothelial cell
4. pulmonary artery endothelial cell
5. vein endothelial cell
6. microvascular endothelial cell
7. CD34+ CD38+ ILRa+ CD45RA- common myeloid progenitor
8. CD34+ CD38+ IL3Ra- CD45RA- erythroid-megakaryocyte progenitor
9. CD34- CD71+ GlyA+ erythroid cell
10. CD34- CD71 lo GlyA+ erythroid cell
11. CD34- CD71 - GlyA+ erytroid cell
12. CD34+ CD41+ CD61 + CD45 - colony forming unit megakaryocyte
13. CD34- CD41+ CD61+ CD45- megakaryocyte
14. FSChi SSClo IL3Ra+ CD33dim+ eosinophil
15. SFChi SSClo CD22+ CD123+ CD33+/- CD45dim basophil
16. CD34+ CD38+ IL3Ra+ CD45RA+ granulocyte/monocyte precursor
17. CD34- CD33+ CD13+ colony forming unit monocyte
18. FSChi SSClo CD14+ CD45dim monocyte
19. CD34- SSChi CD45+ CD11b- CD16- colony forming unit granulocyte
20. CD34- SSChi CD45+ CD11b+ CD16- neutrophilic metamyelocyte
21. FSChi SSChi CD16+ CD11b+ neutrophil
22. HLA DR+ CD3- CD14- CD16- CD19- CD56- CD123- CD11c+ myeloid dendritic cell
23. HLA DR+ CD3- CD14- CD16- CD19- CD56- CD123+ CD11c- plasmacytoid dendritic cell
24. CD19+ IgD+ CD27- naive B lymphocyte
25. CD19+ IgD+ CD27+ B lymphocyte
26. CD19+ IgD- CD27- B lymphocyte
27. CD19+ IgD- CD27+ B lymphocyte
28. CD56- CD16+ CD3- natural killer cell
29. CD56+ CD16+ CD3- natural killer cell
30. CD56- CD16- CD3- natural killer cell
31. CD14- CD19- CD3+ CD1d+ natural killer cell
32. CD4+ CD62L+ CD45RA+ naive T cell
33. CD4+ CD62L- CD45RA- effector memory T cell
34. CD4+ CD62L+ CD45RA- central memory T cell
35. CD8+ CD62L+ CD45RA+ naive T cell
36. CD8+ CD62L- CD45RA+ effector memory T cell (1)
37. CD8+ CD62L- CD45RA- effector memory T cell (2)
38. CD8+ CD62L+ CD45RA- central memory T cell
6
5
4
3
2
11
10
9
27
26
25
24
8
1
16
19
18
21
20
31
28
23
22
17
15
14
13
12
7
30
29
36
37
38
35
34
32
33

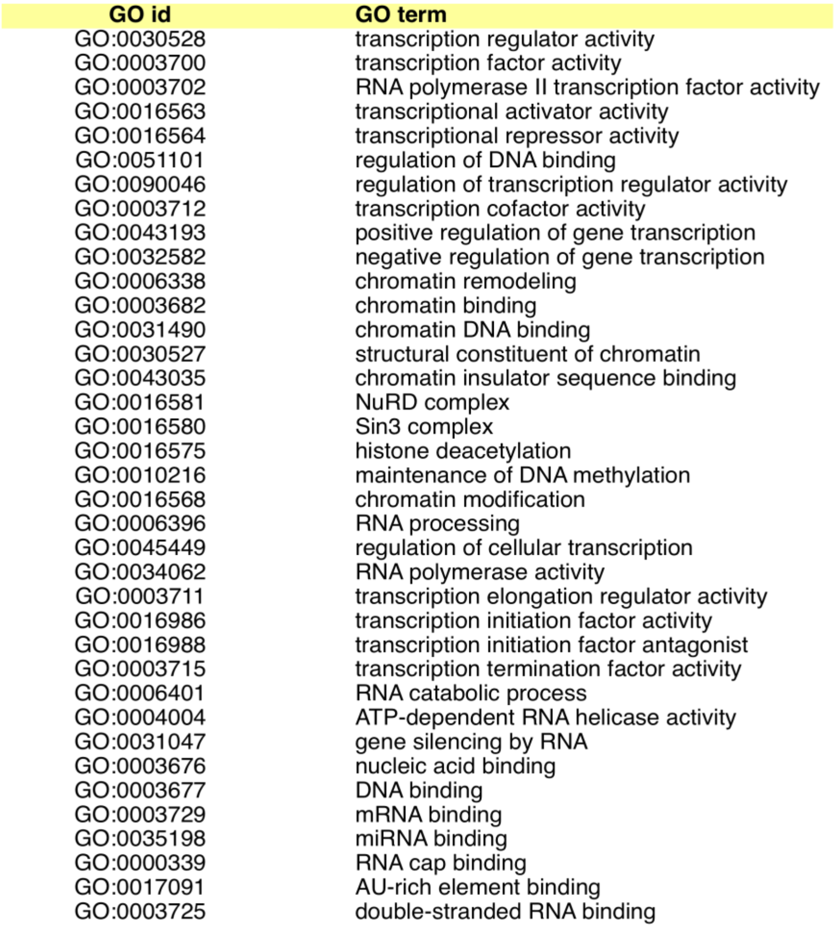
Hierarchical clustering of differentiated cell types from the independent array set (see Supplementary Table 17) was performed as for Fig. 5. Placement of precursor cell types that were selected to match closely those in our dataset and mapping of the tree to a landscape is shown. The landscape elevation (z-dimension) represents the similarity Φ to the ESC where blue color and high altitude on the landscape corresponds to large similarity to the pluripotent cells.

Supplementary Fig. 13. Multidimensional scaling representation of all cell type dissimilarities.



Two-dimensional multidimensional scaling was used to visualize the cell type reversal similarity matrix. The landscape elevation (z-dimension) represents the similarity Φ to the ESC, similar to Fig. 2. Blue color and high altitude on the landscape correspond to large similarity to the pluripotent cells. Numbering refers to order of cell types given in Supplementary Table 3.

Supplementary Table 4. Gene ontology terms used to query the GO database for transcription regulating genes.



GO id	GO term
GO:0030528	transcription regulator activity
GO:0003700	transcription factor activity
GO:0003702	RNA polymerase II transcription factor activity
GO:0016563	transcriptional activator activity
GO:0016564	transcriptional repressor activity
GO:0051101	regulation of DNA binding
GO:0090046	regulation of transcription regulator activity
GO:0003712	transcription cofactor activity
GO:0043193	positive regulation of gene transcription
GO:0032582	negative regulation of gene transcription
GO:0006338	chromatin remodeling
GO:0003682	chromatin binding
GO:0031490	chromatin DNA binding
GO:0030527	structural constituent of chromatin
GO:0043035	chromatin insulator sequence binding
GO:0016581	NuRD complex
GO:0016580	Sin3 complex
GO:0016575	histone deacetylation
GO:0010216	maintenance of DNA methylation
GO:0016568	chromatin modification
GO:0006396	RNA processing
GO:0045449	regulation of cellular transcription
GO:0034062	RNA polymerase activity
GO:0003711	transcription elongation regulator activity
GO:0016986	transcription initiation factor activity
GO:0016988	transcription initiation factor antagonist
GO:0003715	transcription termination factor activity
GO:0006401	RNA catabolic process
GO:0004004	ATP-dependent RNA helicase activity
GO:0031047	gene silencing by RNA
GO:0003676	nucleic acid binding
GO:0003677	DNA binding
GO:0003729	mRNA binding
GO:0035198	miRNA binding
GO:0000339	RNA cap binding
GO:0017091	AU-rich element binding
GO:0003725	double-stranded RNA binding

Supplementary Table 6. Publication references for the genes with ESC-restricted Ψ scores. Pubmed was queried for literature references showing functional evidence in ESCs for the top 20 genes identified from the ESC cell portrait.

Gene	EntrezID	Class	Function in stem cells	Reference
LIN28B	389421	ST	miRNA processing	Hagan, J.P., Piskounova, E. & Gregory, R.I. Lin28 recruits the TUTase Zcchc11 to inhibit let-7 maturation in mouse embryonic stem cells. Nat Struct Mol Biol 16, 1021-5 (2009).
ZIC3	7547	TF	TF network	Shushan Lim, L. et al. The Pluripotency Regulator Zic3 is a Direct Activator of the Nanog Promoter in Embryonic Stem Cells. Stem cells 28, 1961-9 (2010).
ZIC2	7546	TF		Salero, E. & Hatten, M.E. Differentiation of ES cells into cerebellar neurons. Proc. Natl. Acad. Sci. U.S.A. 104, 2997-3002 (2007).
LIN28	79727	ST	miRNA processing	Yu, J. et al. Induced pluripotent stem cell lines derived from human somatic cells. Science 318, 1917-20 (2007).
OTX2	5015	TF		Ahn, J.-I. et al. Comprehensive transcriptome analysis of differentiation of embryonic stem cells into midbrain and hindbrain neurons. Dev Biol 265, 491-501 (2004).
NANOG	79923	TF	TF network	Yu, J. et al. Induced pluripotent stem cell lines derived from human somatic cells. Science 318, 1917-20 (2007).
TET1	80312	CM	DNA methylation	Ito, S. et al. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature 466, 1129-33 (2010).
POU5F1	5460	TF	TF network	Yu, J. et al. Induced pluripotent stem cell lines derived from human somatic cells. Science 318, 1917-20(2007).
SOX2	6657	TF	TF network	Yu, J. et al. Induced pluripotent stem cell lines derived from human somatic cells. Science 318, 1917-20 (2007).
SALL4	57167	TF	TF network	Yang, J. et al. A novel SALL4/OCT4 transcriptional feedback network for pluripotency of embryonic stem cells. PloS One 5, e10766 (2010).
ESRP1	54845	ST		Warzecha, C.C. et al. An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. EMBO J 29, 3286-300 (2010).
ZFP42	132625	TF	TF network	Scotland, K.B. et al. Analysis of Rex1 (zfp42) function in embryonic stem cell differentiation. Dev Dyn 238, 1863-77 (2009).
DNMT3B	1789	CM	DNA methylation	Li, J.-Y. et al. Synergistic function of DNA methyltransferases Dnmt3a and Dnmt3b in the methylation of Oct4 and Nanog. Mol Cell Biol 27, 8748-59 (2007).
SALL1	6299	TF	TF network	Yang, J. et al. A novel SALL4/OCT4 transcriptional feedback network for pluripotency of embryonic stem cells. PloS One 5, e10766 (2010).
TOX3	27324	NA		Dittmer, S. et al. TOX3 is a neuronal survival factor that induces transcription depending on the presence of CITED1 or phosphorylated CREB in the transcriptionally active complex. J Cell Sci 124, 252-60 (2011).
SOX11	6664	TF	TF network	Thomas, S. et al. Human neural crest cells display molecular and phenotypic hallmarks of stem cells. Human Mol Genet 17, 3411-25 (2008).
PRDM14	63978	TF, CM	histone methylation	Tsuneyoshi, N. et al. PRDM14 suppresses expression of differentiation marker genes in human embryonic stem cells. Biochem Biophys Res Commun. 367, 899-905 (2008).
ORC1L	4998	CR	TF regulation	Sun, Y. et al. Evolutionarily conserved transcriptional co-expression guiding embryonic stem cell differentiation. PloS One 3, e3406 (2008).
HELLS	3070	CM	DNA methylation	Xi, S. et al. Lsh participates in DNA methylation and silencing of stem cell genes. Stem cells 27, 2691-702 (2009).
ZNF423	23090	TF, CR		Huang, S. et al. ZNF423 is critically required for retinoic acid-induced differentiation and is a marker of neuroblastoma outcome. Cancer cell 15, 328-40 (2009).

Supplementary Results

Motivation for method development

We describe a data-driven method motivated by a two-gene circuit motif known to control binary developmental decisions[2] that contains a pair of mutually-repressive TFs and effectively constitutes a toggle switch. Their role in lineage determination is reviewed in[55,56].

A critical issue when collecting data from different experiments is whether these samples can be combined into one analysis in such a way that the differences in sample handling, technical variability in signal detection, and data value distributions will not dominate over biologically salient differences between the samples. The huge untapped potential for new knowledge inherent in the vast diversity of published datasets motivates an analysis method that bypasses these challenges to perform large-scale comparative analysis across cell types.

Statistical methodology development over the past decade has mitigated many of the problems outlined above in gene expression analysis[41,57-59]. Commonly, such methodology is referred to as data normalization, although normalization itself can consist of several very different steps. One common approach for microarrays is to apply normalization across arrays in order to generate a dataset with similar value range and distribution[41,57]. To do so, some assumptions as to how this variation is distributed are made and some information (e.g. absolute scale) that one deems to be artifactual or not useful is sacrificed. RNA-seq technology has emerged as an alternative to microarrays and initial reports claimed it was devoid of the main nonlinear distortions present in microarrays, namely chemical saturation in hybridization and optical saturation due to scanner limitations. However, RNA-seq data has other sources of nonlinear distortions that create unwanted and obscuring variability that still requires normalization[58,59].

Beyond the technical issue of data distributions, our central goal is to extract readily interpretable information on cell lineage decisions. Separation of groups of samples is a task that can generally be tackled using clustering or classification. Indeed, one can computationally identify features (genes) that are able to statistically distinguish two groups of samples, such as cell types. Such feature sets are typically not unique (in terms of their estimated classification performance) nor do they reflect any prior knowledge of mechanisms or ontogenic relationships among cell types.

In this work, we place emphasis on preserving the biologically intuitive placement of stem cells and precursor cells onto branch points of tree dendrograms that can be drawn to visualize distances as a cell type lineage tree. Finally, experimental evidence of cell type plasticity shows that transitions between multiple cell types can be induced, provided with sufficient knowledge of key regulatory factors (mainly TFs and other transcription regulating genes). To address lineage switching for each possible cell type pair (out of 13695 such pairs), pair-wise cell type to cell type comparisons summarized at the appropriate biological feature level, namely genes, are required. To our knowledge, no methods exist for this purpose.

A concept of ‘relative expression reversals’, was first introduced in the context of cancer sample classification[7-9] where its performance as a classification method was demonstrated to be highly accurate. The original method generates simple and accurate decision rules for a two-sample classification task (with some extension to a multi-class case). In this work, we develop a method that uses relative expression reversals in a large-scale setting to produce intuitive and practically relevant gene- and cell type-level datasets, and provide a developmentally grounded interpretation that directly connects the results to cell lineage specification. By building on the concept of relative expression, the method is invariant to normalization across samples, as are all rank-based methods.

Microarray dataset assembly

Gene expression patterns reflect the dynamics of regulatory circuits that govern lineage specification. We present a biologically motivated analysis method to reveal lineage determination from gene expression signatures. To achieve this goal, microarray and gene datasets were assembled as outlined in **Supplementary Fig. 1**.

Selection of microarrays from the GEO public microarray repository

The Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) is a public functional genomics data repository that can be queried for published gene expression datasets from different platforms. Our aim was to select a human expression data platform that contained a large number of different experiments, with a maximal genome coverage. We queried GEO for expression profiles of normal (i.e. not patient, cancer or long-time treatment) cell or tissue samples.

Choice of array platform and probe mapping

To obtain maximal genome coverage and a large number of available datasets, we compared different expression data platforms. The Affymetrix hgu133 Plus2 (GPL570) microarrays represented by far the most widely used platform with over 67000 samples available. The array platform chosen represents the so called 3’ array generation where probes were mainly designed to the 3’ end of transcripts and in general are not sensitive to different gene isoforms (compared to the newer generation exon arrays). We mapped the probes to the Refseq mRNA sequences to check what proportion of genes were differentially probed in case of multiple transcript variants. Only 1613/20072 genes (8%) were differentially probed. Therefore, we chose to map all probes to genes, discarding transcript information.

1) The human genome (a local copy of "GRCh37 primary reference assembly" from NCBI) and the RefSeq database (human.rna.fna dated 13 Dec 2010) were independently scanned for exact matches to the HG-U133_Plus_2 Affymetrix PM probes defined in HG-U133_Plus_2.probe_fasta.

2) The RefSeq sequences were aligned to the genome using the downloadable NCBI alignment tool Splign. The alignments were filtered to include those strictly contained within the genomic windows (+/– 1 bp) listed in the gene2refseq annotation available from NCBI. This step was included to validate existing transcript alignments against the local copy of the genome that was scanned for probe sequences ensuring all downstream coordinates reference a single physical copy of the genome. This step does not yield new alignments.

3) The NCBI gene2refseq file associates one or more transcripts with each Entrez gene identifier. For the purpose of downstream computations, a gene's location is defined as the smallest genomic window containing all the alignments validated in Step 2 of all its transcripts.

4) An Affymetrix PM probe is associated with a gene locus if either:

a) It exactly matched a 25 bp segment strictly contained within a gene's locus (as defined in 3), or

b) It exactly matched any of the gene's transcripts (as defined by gene2refseq)

This definition is the key point distinguishing our mapping from others.

5) Probes were selected that were associated (as defined in Step 4) with exactly one locus by default. All multi-locus probes were curated for inclusion which could result from a situation when i) the probes target a pseudogene locus in addition to the actual locus; ii) a read-through transcript that generates an identical protein (and no other protein) overlaps the actual transcript; iii) the probes target a gene that is present in multiple copies in the genome, yet codes for the same protein; iv) the probes target a bicistronic transcript. (The Entrez IDs of the 14 genes which pass as multi-locus hits are 4207, 5460, 5940, 6606, 6607, 6638, 22947, 84321, 86614, 90316, 136319, 159119, 253175 and 100271849.)

6) A custom CDF was created using a Python script that integrated a template CDF from Affy (HG-U133_Plus_2.cdf) with the results of the preceding steps.

6a) Control probe data was copied verbatim from the template CDF

6b) All non-control PM probes in the resulting CDF were taken from the preceding steps.

6c) The respective mismatch probes were inferred from the template CDF and included without further validation.

6d) Finally, only transcription regulating genes (described later) were included in the CDF, and we required a minimum of three probes per gene, resulting in 2602 probed genes (844 in the high confidence TF only dataset).

The goal was, as always, to achieve a balance between sensitivity and specificity. The mapping may include probes that, while only associated with one locus by the above definition, nonetheless match arbitrarily many locations in the genome. However, the RNA sample processing should have removed genomic DNA contamination. Note that, because of splicing, a probe may match a gene's transcript but nowhere in the gene's genomic locus.

Annotating microarray experiments with cell type or tissue ontology terms

We focus on normal human cells and tissues and group the array samples by cell type (/tissue). We used the open ontology (cell.obo and uberon.obo) definitions of cell types[46] and tissues[47] (available from http://www.obofoundry.org/) to generate SQL queries over the GEOmetadb database of GEO annotation information with names of cell types and their synonyms as search terms (**Supplementary Table 1**). The experiments were next hand-curated for inclusion and associated cell type term. The following distinctions were made:

1) Tissue samples were not discarded at this point although they are known to contain multiple cell types, however these were kept as separate groups from samples representing cell cultures or primary cell isolations.

2) It is known that cells may change their phenotype during long-term culture, therefore a separation was also made between freshly isolated primary cultures and long-term/immortalized cultures and cell lines.

3) Some experiments studied the differentiation process of cells. These samples were separated as their own group to allow for comparison with precursor state and fully differentiated samples.

Samples representing cancer cells, cell lines derived from patients and > 8 h exposures to natural or chemical compounds were discarded.

Choice of microarray preprocessing method

To address the possibility to apply any of the commonly used normalization methods to the dataset presented here, we have plotted from each array the ratio of the probe value upper quartile to the lower quartile (Q3/Q1) against the array median (**Supplementary Fig. 3**). These plots were inspected for systematic effects within and between cell types and across different array generations. Overall, there was a very large spread in the values indicating that co-normalization would pose a problem. Inspections of systematic trends revealed that arrays representing samples from amplified RNA typically had low Q3/Q1 values. These effects would be very difficult to reconcile by any kind of normalization, which could even be considered data destructive in this setting.

We tested different background correction methods to select one which would work best to rank the expression levels of genes within an array. In order to perform well, the method should consider probe affinity effects and correct for background signal. The standard RMA method operates on PM match probes only, which does not allow a clear separation between low expressed genes and non-expressed ones. The GC-RMA[41] full model performed better in this setting. The GC-RMA calculated expression matrix was converted to a rank matrix and used in the downstream analysis presented. In summary, we chose a strategy where the probe signal values were background corrected using the GC-RMA method, no normalization step was applied, and the probe values were combined to one expression value per Entrez Gene ID using median polish.

Quality control of the collected microarray samples

In order to eliminate possible misclassification of arrays to their cell type term, a rank correlation matrix (Spearman's rho) was calculated using all samples. Arrays that belong to the same cell type needed to fulfill the following criteria for their inclusion in the subsequent analysis:

1) Max within-group correlation > 0.9
2) Min within-group correlation > 0.8
3) Max between-group correlation < Max within-group correlation

The oocyte and spermatogonial cell samples represent an exception where the maximum correlation within the group of arrays was much lower than in other cell types, attributable to RNA isolation from single cells or a few hundred cells, respectively. This resulted in a matrix with 2919 microarrays representing 166 cell or tissue type terms (**Supplementary Table 2**). The cell types and number of samples per cell type in the final data set are listed in **Supplementary Table 3**.

Collection of transcription regulating genes

Transcriptional regulation encompasses many levels, not just the TFs that recognize specific regulatory sites, including the modification of chromatin state, the regulation of TF activity and RNA processing. The genes involved at different levels are potentially tightly integrated into the core regulatory networks. In order to identify genes with a shared function, we used the Gene Ontology database (http://www.geneontology.org/). This was the main data source for our gene selection, which was then compared to and complemented with additional dedicated data sources that host collections of genes functioning in transcription regulation. The GO Online SQL Environment query and BioMart Perl commands are available upon request.

The data sources used and their overlap for gene set assembly

GO database

The Gene Ontology project[48] provides gene product annotation data from the GO consortium that can be queried online using the GO Online SQL Environment (http://berkeleybop.org/goose). Each gene is associated with ontology terms that are structured as cellular component, molecular function and biological process ontologies. GO terms related to transcription regulation process were selected (**Supplementary Table 4**) and an SQL query was formed to find genes that have a matching annotation in either human or mouse (Feb 2010, AmiGO v. 1.6.0.0). The list of genes from human annotations (symbols and swissprot accession terms) was mapped to respective EntrezIDs using the Ensembl Biomart database[60]. The list of genes from mouse annotations (MGI IDs) was mapped similarly to the respective human EntrezIDs using the BioMart annotations and gene homology information.

DBD

The DBD database[49] (http://dbd.mrc-lmb.cam.ac.uk/DBD/index.cgi?Home) hosts genome-wide TF predictions from multiple completely sequenced genomes, including human and mouse. The predicted TFs contain assignments to sequence specific DNA-binding domain families based on hidden Markov model libraries. The genes obtained from this source (release 2.0) for human and mouse (Ensembl protein IDs) were mapped to respective human EntrezGene IDs using the Ensembl BioMart database.

ChromDB

The Chromatin Database[51] (http://www.chromdb.org/index.html) hosts sequence information in two broad functional classes: chromatin-associated proteins and RNA interference associated proteins. The list of genes representing all protein groups in the database was obtained (gene symbols) from human and mouse (Feb 2010 release) and mapped to respective human EntrezGene IDs using the Ensembl BioMart database.

Riken mouse TF database

The Mouse Transcription Factor database[50] (http://genome.gsc.riken.jp/TFdb/) is a database containing mouse TF genes and their related genes. The list of genes obtained (EntrezGeneIDs) was converted to respective human EntrezGene IDs using the Ensembl BioMart database.

The overlap of gene IDs from the different sources (**Supplementary Fig. 2a**), and between human and mouse annotations (in (**b**)) were compared. The two dedicated TF databases had an overlap of 834 genes, and all of these overlapping genes were also identified using the GO term search (representing 81% of total number of genes from these dedicated data sources). The DBD database provides 29 genes not found via the other sources, and the Riken TF database, 170. Similarly, the GO search identified 375/425 genes (88%) of the genes functioning in chromatin modification or RNA silencing according to ChromDB. In total, combining the list from these sources resulted in 4700 genes that were selected for curation.

Curation of transcription regulating genes

The list of genes obtained from the various data sources represents a heterogeneous collection of genes, both on the level of function in transcription regulation and on the level of evidence supporting each function. An initial check for possible false positive hits indicated that a number of genes with only indirect effects on gene transcription (membrane receptors, secreted proteins, signalling cascade components) were included. The GO evidence level annotations were also variable, and did not serve as a good filter (data not shown).

Automated text-retrieval from NCBI databases and word pair -based text search

The Entrez GeneID list of potential transcription regulating genes was used as an input to a python implemented text retrieval query. The Biopython package Entrez was used to extract the following information from the NCBI Entrez database: the gene name, a short summary provided by Refseq, Refseq status, possible alias, if existing, the MIM (Omim identifier), GO annotations (function, process, component, evidence codes) and PubMed IDs. Accordingly the information per gene in a complete dataset included the following information in the format of a python dictionary:

gene[EntrezId] = {name:(string), summary:(string), status:(string), alias:(string), omim: (omimId, omimtext), pubmed:(string), function:[GO-function evidenceCode, GO-function2 evidenceCode,...],process:[GO-process1 evidenceCode, GO-process2 evidenceCode,...], component:[GO-component1 evidenceCode, GO-component2 evidenceCode,...]}

In case references to OMIM or PubMed were available, further information (gene function and biochemical features from OMIM, abstract from PubMed) was extracted in a subsequent step.

A gene set representing approximately 10% of all candidate genes was curated and classified into functional categories (see below for Gene classification). The curated text for this gene set was then used to automate the process for the remaining genes by identifying informative word pairs. The text data for these genes were processed to extract all possible word pairs occurring after filtering out symbols and numbers using a regular expression search. Word pairs related to an assignment of function in transcription regulation were manually selected. These word pairs were then used to highlight key parts in all subsequent text retrieval steps. The automated steps extract the respective text region when a word pair is detected.

For those genes that failed the Entrez-based curation, an additional step was included where all PubMed abstracts referencing the official gene name (or its synonyms) together with the word ‘transcription’ were retrieved (max 500) and passed through the word-pair extraction step and subsequent manual curation and classification.

Gene classification

All genes were classified to four main sub-categories: TF, co-regulator, chromatin modifier and mRNA transcript synthesis/processing. The descriptions for each category are listed below, and for each gene, at least one PubMed ID reference indicating evidence for that function was associated with the gene. The association to multiple categories was allowed since many genes can carry out multiple functions.

Transcription factor (TF): The encoded protein binds to DNA at regulatory regions of its target genes and through its binding affects (positively or negatively) the transcription of the target genes.

Co-regulator (CR): The encoded protein (or protein complex whose component the protein is) binds directly to a TF to regulate its function: Activation/inactivation of TFs by protein modification / Degradation of TFs by targeting to proteasome / Recruitment of chromatin modifying enzyme complexes to the TF bound chromatin regions / Bridging the TF to the basal transcription machinery.

Chromatin modifier (CM): The encoded protein (or protein complex whose component the protein is) possesses enzymatic activity to modify DNA or histones to alter the local chromatin status.

mRNA transcript synthesis/processing (ST): The encoded protein (or protein complex whose component the protein is) is involved in RNA polymerase II mediated mRNA synthesis / mRNA processing / mRNA transport/ mRNA degradation: Component of the basal transcriptional machinery / Recognition of the TSS / mRNA elongation / mRNA splicing / mRNA stability / microRNA-mediated silencing / mRNA degradation.

The following categories of genes were not included: DNA repair, DNA replication, histone proteins, histone RNA synthesis and processing, mRNA translation. Direct evidence available for multiple family members was considered sufficient for highly likely similar function. If the gene could not be classified to one or more of the subclasses listed, it was discarded due to lack of evidence for direct involvement in transcription regulation. The initial list was complemented by 55 genes, representing missed genes from gene families, during curation. The curation of TFs was extended to contain evidence level (1=clear evidence, 2=indirect evidence, 3=weak evidence).

The resulting list of genes represents genes functioning in transcription regulation that are supported by experimental evidence. It contains 2212 genes (**Supplementary Fig. 2d**) of which 922 genes are TFs (Evidence level 1: 816 genes, 2: 61 genes, 3: 45 genes), 677 are co-regulators, 296 are chromatin modifiers and 561 function in mRNA transcript synthesis and processing (notice that some genes were annotated to multiple categories).

Comparison of the TF list to known resources

To assess the quality of the curated list, the genes annotated as TFs were selected. A census of human TFs that was based on domain search and manual curations for evidence[52] was selected for the comparison. The full list of potential TFs provided in that paper were mapped to EntrezGene IDs using Ensembl BioMart, resulting in 1863 genes. Of those, 1728 (93%) are included in our curated gene set, leaving 135 genes that were not included (**Supplementary Fig. 2c**).

Our dataset includes functional evidence for a role in transcription regulation for 1040 genes (56%) listed in[52], of which 837 genes were annotated as TFs. For the remaining genes, our curation rarely found evidence against direct function in transcriptional regulation, only for 16 genes. The low number of genes that were rejected from this list (false positive rate = 0.9%) indicates a high likelihood that the genes with unknown function do in fact represent transcription regulating genes. Therefore, we included 542 genes that were annotated in the census[52] dataset (with evidence levels a-c given in the publication) for the follow-up analysis as the NA category, to potentially generate new hypothesis concerning their role in transcriptional regulation from their expression profiles.

The 135 genes that had not been curated were passed through our curation process, but no genes with evidence for function as TFs were identified: evidence against direct function in transcription regulation was found for 83 genes, for 41 genes no functional information was available, one gene represented a pseudogene, one gene was identified as functioning as a chromatin modifier, 8 functioning as co-regulators and one functioning in mRNA transcript synthesis/processing (these 10 genes with evidence were included). The low number of genes missed by our approach (10/1863, false negative rate = 0.5%) indicates high representativeness of transcription regulating genes. Of note, our dataset had 84 additional TFs that were missing from the census[52] dataset (including one TF that was added during curation).

The final list of genes (2754) with their associated classification terms (TF/CR/CM/ST/NA) and PubMed references can be found in **Supplementary Table 5**.

Gene pair analysis

The concept of expression reversals in a pair of conditions is illustrated by the example of the hypothetical {gene *g*, gene *g'*} gene pair (**Fig. 1**) and explained in more detail here. To find two genes that exhibit a high reversal property, the expression levels of all genes are first converted to ranks in each array and normalized by the number of genes to the range (0,1). In this case, the ranks of the two genes across a set of microarrays (samples 1-9) encompassing three hypothetical cell types, are plotted to show the gene pair configuration within each array (**Fig. 1a**). The first cell type expresses variable amounts of transcripts of both TF genes, whereas a pronounced reversal of their relative mRNA expression levels (ranks) is observed between the second and third cell types: {gene g >> gene *g'*} in samples 4-6 while {gene g << gene *g'*} in samples 7-9.

Such reversal of a gene pair for pairs of cell types can be evaluated across a set of *N* cell types, each represented by at least two microarray samples. Each gene pair {gene *g*, gene *g'*} can be assigned an *N*x*N* matrix displayed as a color heat map (**Fig. 1b**). In such plots, referred to as gene pair reversal plots, rows and columns represent skew-symmetrically all possible pairs of the *N* cell types being compared with respect to a gene pair, with each matrix element representing a quantity that reflects the extent of reversal of the relative ranking of gene *g* and gene *g'*, in the given pair of cell types being compared. This quantity, which we denote as $\varDelta$, is the change in the normalized mean rank difference δ of gene *g* and gene *g'* (shown in **Fig. 1a** for cell type 2 and 3). If the pair configuration is not fixed in one or both cell types being compared (as in cell type 1, **Fig. 1a**), $\varDelta$ is assigned a value 0 (see Methods for details). $\varDelta$ is thus a property of a TF pair that is considered in the context of two cell type profiles. The $\varDelta$ score is a key element of our analysis.

To explore cell type specificity across large scale data sets, we introduce the reversal participation score, $\varPsi$, that examines a gene's reversal behavior in a large number of gene reversal pairs in a large set of cell type comparison pairs (see Methods Eq. 4). Thus, while $\varDelta$ is a property of a gene pair, $\varPsi$ is a property of a gene. Each of the 2602 genes investigated in this study is a member of 2601 nominal gene pairs, for each of which a unique reversal pattern can be computed. The results are visualized in "reversal participation $\varPsi$ gene/cell type portraits" that encompass all cell type comparisons (**Fig. 1c**). Rows in such matrices correspond to a particular cell type compared to all other cell types (32 hypothetical cell types are compared in **Fig. 1c**). These results can be calculated for all genes and sorted based on signal strength to get a ranked gene list for a particular cell type (shown for hypothetical cell type 12). In addition, color heat maps can be associated with an individual TF to form the cell type comparison matrix (here shown for the top ranked gene of cell type 12). These results are to be distinguished from the previously described gene pair reversal plots that quantify the gene pair score Δ. The Ψ matrix values reflect how often a gene *g* was involved in pairs that reversed for a given cell type comparison (irrespective of which other gene was paired with *g*) and the relative magnitude and direction of the change.

Effect of the number of samples on the gene pair reversal results

To quantify the effect of the number of arrays used to derive the δ values, we utilize the strict inequality requirement and determine the number of gene pairs in a fixed configuration (a requirement for a non-zero δ value) at different δ value cut-offs from random sampling. The three largest cell type groups (ductal breast epithelial cells, skeletal muscle tissue, monocyte cells) were each sampled 100x taking each time an increasing amount of arrays to represent the cell type (see **Supplementary Fig. 5**). The method is very robust in the δ value range (0.4,2) where practically no dependence on sample size is observed.

The effect of fluctuations in gene pair ranks on the reversal participation results

To quantify the effect of random fluctuations in gene pair rankings that may propagate to downstream results, we designed two noise injection simulations and evaluated the effect on the reversal participation results. First, we quantified the normal variance observed in the expression data. This was calculated by taking an average over all genes, where for each gene the variance was determined within each cell type. We then additively injected zero mean Gaussian or Laplacian noise to the original expression matrix and ran repeated simulations (n = 100) with several noise levels. We used ROC analysis to show that the ranking of genes in the cell type portraits is hardly affected and only shows a decline in performance at the highest noise levels, unlikely to be observed in real expression data. The top 20 genes from the original data were taken to represent the true positive set. At each noise level, the performance is evaluated by the cell portrait ranking with respect to these 20 genes averaged over 100 simulations. The results for each cell type displayed as a separate curve are displayed in **Supplementary Figs. 6 and 7**. To make the interpretation more intuitive, the individual cell portraits and gene portraits shown represent results obtained from a single representative run (**Supplementary Figs. 6 and 7**) and agree with what can be seen in the ROC curve: lineage-restricted signal persists even at increased noise levels and for genes that initially show no such signal, a false signal is not present even at the highest noise levels. Such robustness, particularly to heavy tailed noise distributions, is attributable to rank order statistics.

Comparison to rank-based differential expression analysis methods

A number of rank-based analysis methods have been proposed that are considered to be invariant to normalization across arrays[61-63]. Thus, rank-based approaches are well suited for comprehensive comparative genomic studies. One of the published methods, RDAM[61], first replaces raw expression values by their ranks. RDAM always makes comparisons between two samples by considering variation (rank differences) between genes. Thus, the method does not require averaging the samples from a given class, but rather, different samples are merged at the level of the significance analysis by the product of p-values. Another method, RCoS[62], looks for genes with consistently high ranks, calculated from fold-changes of paired samples, across different classes of samples by computing a score using the observed ranks of each gene across all samples. It also incorporates flexibility to outliers by introducing a rank consistency score, which is based on a robust mean of ranks. Finally, the statistical significance is evaluated by comparing the observed rank consistency scores to randomly drawn data from the null model.

Our method also uses rank values instead of raw intensity values. We transform gene ranks to gene pair scores as an initial step, which is a key difference to the above mentioned differential expression based methods. In our approach, we combine samples from a given cell type by requiring that the *pair* ranking remains consistent throughout, as for example RCoS[62] evaluates consistency using the rank consistency score. The above rank based methods are motivated by their ability to extract reliable gene lists even from a small number of samples and running them becomes computationally expensive on large sample collections. Our approach directly benefits from larger amount of data (but works well with small number of samples, **Supplementary Fig. 5**) and will produce more accurate results as the number of samples increases with approximately linear increase in computational cost.

To show the benefit of using the gene pair analysis over standard analysis of differential expression, we performed the parallel analysis using published rank-based differential expression analysis methods, RDAM[61] and RCoS[62] (see **Supplementary Tables 8 and 9**), to discover mutually antagonistic gene pairs for lineage separation between erythroid and myeloid cells. With both of the above mentioned algorithms we test differential expression between two randomly paired arrays from cell types *X* and *Y*. The number of array pairs tested is set to 5 for RDAM (we settled on this number after testing with higher number of arrays that, either due to the increase in computational complexity inherent in this method, or related to the implementation that was not previously tested in a large-sample setting, was deemed infeasible; the results from independent runs were comparable with 5 arrays) and all arrays from both cell types are used for RCoS. When there are less samples in cell type *Y* than in *X*, the same arrays are paired multiple times. To make the comparison more robust, we apply circular permutation as proposed[61]. Circular shifting of samples is used to produce permutations: arrays of cell type *Y* are randomly paired with arrays from cell type *X* multiple times. Circular shifting is repeated a maximum of 100 times or the number of arrays if it is less than 100. For RDAM, the random pairing of array samples is used as an input. For RCoS, we reduce each cell type comparison to an average logratio across all array pairings as proposed by the authors. To make the results more robust, we run the algorithm 100 times and average the obtained FDR values. Differentially expressed genes are called by an average FDR < 0.05 (RDAM) or Bonferroni corrected p-value (RCoS).

Using both methods, we can distinguish both up- and downregulated genes from each lineage comparison that correspond to those tested in **Fig. 4** (see **Supplementary Tables 8 and 9**). To propose a list of candidate toggle gene pairs, we pair up the genes with antagonistic expression profiles and then check whether the individual genes, or the gene pairs formed, are specific to a particular lineage comparison (the numbers are summarized on the first sheet and individual lists are included on the following data sheets). The two tools used (RDAM and RCoS) differ in their outputs and one limitation in comparing to RCoS was that this tool does not report a list of genes that passed only few comparisons when the number of cell type comparisons increased (useful to filter pairs that make it into the list in any of the comparisons to progenitor or lymphoid cells). The gene (and pair) lists for the other lineage comparisons would actually be longer if the filtering would be fully in accordance with the way our pair analysis treats the non-relevant lineage comparisons (i.e., not allowing any reversals in non-relevant cell type comparisons, non-relevant referring here to comparisons outside the lineage split of interest). We report here the more conservative lists of genes passing all/all comparisons between cell types for the progenitor-erythroid, progenitor-myeloid, lymphoid-erythroid and lymphoid-myeloid comparisons (we always require all/all comparisons to be passed for the erythroid-myeloid comparison consistent with our analysis). With RDAM, it was possible to report pairs that pass all erythroid-myeloid cell type comparisons, but do not occur in any of the non-relevant lineage comparisons (we also perform the filtering as we did using RCoS to allow comparison).

Utilizing anti-correlating differential expression between the two lineages as an additional criterion, these approaches resulted in implausibly many candidates (198 pairs using RDAM[61] with FDR q<0.05, 4352 pairs using RCoS[62] with Bonferroni p<0.05, **Supplementary Tables 8 and 9**), suggesting that most must be false positives. The additional lineage comparisons (between progenitor and lymphoid cell types) reduced the RCoS list to 3214 pairs (including the {*GATA1, SPI1*} pair), while the RDAM list held only 9 pairs. However, the lack of the {*GATA1, SPI1*} pair on the RDAM list, and the inclusion of TF pairs that either had stabilized an opposite configuration already in the progenitor cells or did not maintain a mutually antagonistic expression within cell types that committed to a particular lineage, questions the utility of these candidate TF pairs.

Our method directly evaluates mutual repression by considering the gene pair consistency for a given lineage split, and provides a statistical evaluation of the specificity towards that lineage that is a key distinction that allows both narrowing down the list of candidates and ranking them. Moreover, it is clear from this comparison that the differential expression reveals lineage-relevant genes, but fails to associate the relevant genes to a lineage-split, a feature that is demonstrated only at the level of gene pairing.

Additional data used in validation

ChIP-seq datasets The GEO and SRA accession numbers for the ChIP-seq data are indicated in **Supplementary Table 10**.

Hematopoietic system: To further evaluate the gene network regulated by the two SPI1-containing gene pairs that were most specific to the erythroid-myeloid split, we integrated results from published mouse ChIP-seq datasets[64-69]: four studies measuring the genome-wide binding profile of Sfpi1 (also known as Pu.1); two of Gata1, one of Gata2 (known to occupy mostly the same loci as Gata1[70]); two of Tal1. In addition, we display data from the ENCODE consortium for the erythroleukemia K562 cells for all three TFs and from peripheral blood derived erythrocytes for GATA1. For the GATA3 pair with EBF1 data was available from two Ebf1 mouse ChIP-seq datasets[32,71], two from human ENCODE consortium[12] (displayed in **Supplementary Fig. 11**), and a study across different mouse T cell populations for Gata3[29] (data from double negative, double positive, naïve CD4+ and naïve CD8+ cells was combined for display). These represent the interaction of SPI1-mediated regulation with that of the two highest ranked SPI1-pairs identified and the interactions between Ebf1 and Gata3. Possible competitive/inhibitory interactions at genomic sites are revealed by overlapping peak regions and/or target genes for these toggle switch circuit candidates.

ESCs: Similarly, as for the hematopoietic data, we assembled published ChIP-seq data for POU5F1, NANOG and SOX2 (in human)[53,54]. Peak lists were combined and overlapping binding sites were identified for these TFs.

ENCODE datasets: Data from the publicly released ENCODE[12] datasets were displayed for the gene regions of interest. Data from normal cell types was selected for display from the ENCODE Regulation Supertrack (hg18).

Genomic region enrichment analysis. ChIP-seq analysis provides lists of binding sites (peaks) identified across the genome. However, to understand the impact of this binding, that is, how it manifests as gene regulation, it is necessary to associate the peaks to gene regulatory domains. It has become evident from ChIP-seq studies that most TFs bind to distal regulatory elements, spanning up to several hundred kb upstream and downstream of gene TSSs. We adopt here the default gene regulatory domain definition used by the GREAT tool[45]: each gene is assigned a basal regulatory domain and this domain is extended upstream and downstream to the next basal regulatory domain encountered (however not more than 1 Mb away from the gene TSS).

With the gene regulatory domains assigned to each gene, we can now associate binding observed with potential function by matching the ChIP-seq peak coordinates to the gene regulatory domain coordinates. The next question to consider is whether this binding is concentrated nearby genes from particular functional categories as defined by Gene Ontology or pathway databases. This is an important question, since to carry out a specific function in the cell type a TF is expected to regulate a defined set of genes associated with that function. In our case, we are interested to discover whether the binding of the TFs studied is concentrated nearby genes that function in cell differentiation or key pathways required for carrying out the specific biological role of the given cell type, i.e., that there is evidence that the TF is able to establish the given cell phenotype. For this purpose, we employ the binomial genomic region enrichment test, which for each ontology term tests for significance of the fraction of the genome spanned by TF-bound regulatory domains from that category vs. the total fraction of the genome spanned by regulatory domains from that category. For more information about the test performed[45], please refer to http://great.stanford.edu/help/display/GREAT/Home. We used each ChIP-seq dataset independently to perform the genomic region enrichment analysis. Peak data as published by the data providers was used and the binomial enrichment test was performed to select significant ontology terms using the tool GREAT[45]. We collected data from GO Biological function, MGI Mouse Phenotype, Pathway commons and MSigPert perturbation experiments for each ontology category at an FDR level of 1%. These are shown in **Supplementary Tables 11-15** (notice the multiple sheets in these xls files; each corresponds to a separate dataset).

We found that the results were very informative and highlighted the specific blood cell type functions without additional filtering. However, to summarize some key results across these lists, we considered the overlap with the MGI Mouse Phenotype terms and report separately those that were supported by at least two datasets and had at least 1% genomic region coverage. This ontology category was chosen because these terms are related to mouse knockout phenotypes for which the annotations are assigned to genes in a very consistent fashion. Moreover, this allows us to compare to the TF knockout phenotypes themselves (the phenotype of a TF is in fact that caused by lack of expression / misregulation of its key target genes) shown in **Supplementary Table 16**.

Additional microarray datasets To show consistency of results beyond the microarray data set collected from the hgu133Plus2 arrays, we included data from another Affymetrix array type (ht-hgu133a) to support several results obtained. These arrays do not contain probes for all genes, but otherwise represent a comparable independent microarray set (see **Supplementary Table 17** for a list of GSM ids).

Captions

Supplementary Table 1 (separate file)

Cell type and tissue ontology terms. The ontologies for cell type[46] and anatomical site[47] were obtained from OBO. The terms list was extended when encountering subtypes or sample types not directly corresponding to an existing term. *(xls file online)*

Supplementary Table 2 (separate file)

Microarray samples mapped to ontology terms. The samples included in the final dataset are listed with the associated cell or tissue type ontology ID. *(xls file online)*

Supplementary Table 3 (separate file)

The order of cell types as it appears in heat maps presented. The names of cell types corresponding to the 166 rows and columns in the heat map representations of analysis results are listed with the number of microarrays belonging to each group indicated. *(xls file online)*

Supplementary Table 5 (separate file)

Functional evidence for role in transcription regulation found in the gene set curation. The Entrez Gene IDs, assigned class and PMID to a publication that present evidence for the indicated function are listed for the gene set that passed the curation. TF=transcription factor, CR=coregulator, CM=chromatin modifier, ST=RNA splicing, transcription and processing. For the TF set evidence level for the function assignment is provided 1=direct experimental evidence, 2=indirect evidence based on likely shared function of related gene family members, 3=weak evidence. *(xls file online)*

Supplementary Table 7 (separate file)

Candidate toggle pair search results using less stringent cut-offs. *(xls file online)*

Supplementary Table 8 (separate file)

Rank-based differential expression analysis using RCoS. *(xls file online)*

Supplementary Table 9 (separate file)

Rank-based differential expression analysis using RDAM. *(xls file online)*

Supplementary Table 10 (separate file)

The GEO/SRA/ENCODE identifiers for the ChIP-seq datasets used for Supplementary Figs. 8,10 and 11. The peak lists were either available via GEO or were extracted from the respective publications. *(xls file online)*

Supplementary Tables 11-15 (separate files)

Genomic region enrichment results from the GO Biological Function, MGI Mouse Phenotype, Pathway commons and MSigPerturbation data sources for GATA1, TAL1, SPI1, EBF1 and GATA3 respectively. Each file holds multiple sheets that display results of independent ChIP-seq experiments. The first sheet summarizes the MGI Mouse Phenotype results. *(xls files online)*

Supplementary Table 16 (separate file)

Knockout phenotypes for *Gata1*, *Tal1*, *Sfpi1*, *Ebf1* and *Gata3* from MGI. *(xls file online)*

Supplementary Table 17 (separate file)

Datasets used from the affymetrix ht-hgu133a platform. *(xls file online)*

Supplementary Online Resources for reversal participation and landscape results

The reversal participation gene portraits are generated to serve a gene-centric or a cell type-centric analysis. Lineage-restricted reversals make a given gene *g* a potential candidate for a lineage-determining gene. We provide all gene reversal participation portraits and cell type portraits on our online webpage http://trel.systemsbiology.net/. The heatmaps are interactive enabling the user to choose

1) from a gene portrait: cell types with specific high/low expression of the gene to display the cell type portraits of the selection; or

2) from a cell type portrait: genes with the highest/lowest Ψ row sum for the given cell type. The genes of the selection can then be displayed as gene portraits to check how restricted their reversal pattern is to the given cell type.

A workflow is outlined in our online user manual for the use of these results to generate candidate lists for use in re-programming experiments.

The landscape with all cell types can be used to identify (i) neighboring cell types to design direct conversion experiments (closely related cell types require less factors to induce conversion[56]), or (ii) to examine how distant (in terms of transcription regulating gene pair state) the starting and target cell type of interest is from the ESC for experiments where the first step is to induce pluripotency and subsequently to convert to a cell type of interest.

The ranked cell type portraits (top 100 genes) can be used to select candidate genes that when overexpressed may induce lineage conversion. As shown in **Fig. 2** this analysis efficiently identified the genes to induce pluripotency[11] (*NANOG, POU5F1, SOX2, LIN28*) and genes that enhance the efficiency (such as *PRDM14*[72]). The candidate genes can be further examined to select for maximally lineage-restricted genes by visualizing their gene portraits. As shown in **Fig. 4**, genes with neuronal cell type restricted pattern correspond to those that are most efficient to convert fibroblasts to neurons. Additionally, a link to the UCSC Genome Browser is provided from each gene. Similarly to **Fig. 2**, the lineage-restriction can be confirmed from public (e.g. ENCODE[12]) RNA-seq or histone marker data by navigating to the gene area. UCSC Genome Browser also accepts custom tracks enabling users to view own or public data of interest, e.g. from TFs known to be master-regulators for the cell type of interest, to examine how the selected genes may be embedded into regulatory networks (as demonstrated in **Supplementary Figs. 4** and **10**).

References

46. Bard, J., Rhee, S.Y. & Ashburner, M. An ontology for cell types. *Genome Biol.* **6**, R21 (2005).
47. Mungall, C.J. et al. Integrating phenotype ontologies across multiple species. *Genome Biol.* **11**, R2 (2010).
48. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* **25**, 25-9 (2000).
49. Kummerfeld, S.K. & Teichmann, S.A. DBD: a transcription factor prediction database. *Nucleic Acids Res.* **34**, D74-81 (2006).
50. Kanamori, M. et al. A genome-wide and nonredundant mouse transcription factor database. *Biochem Biophys Res Commun.* **322**, 787-93 (2004).
51. Gendler, K., Paulsen, T. & Napoli, C. ChromDB: the chromatin database. *Nucleic Acids Res.* **36**, D298-302 (2008).
52. Vaquerizas, J.M. et al. A census of human transcription factors: function, expression and evolution. *Nature Rev Genet.* **10**, 252-63 (2009).
53. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315-22(2009).
54. Kunarso, G. et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet 42, 631-4(2010).
55. Huang, S. Reprogramming cell fates: reconciling rarity with robustness. BioEssays 31, 546-60 (2009).
56. Graf, T. & Enver, T. Forcing cells to change lineages. Nature 462, 587-94 (2009).
57. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). *Biostatistics*. **11**, 242-53. (2010)
58. Hansen, K.D., Irizarry, R.A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics.* **13**, 204-16 (2012).
59. Loven J., et al. Revisiting global gene expression analysis. *Cell*. **151**, 476-482. (2012)
60. Kasprzyk, A. et al. EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.* **14**, 160-9 (2004).
61. Martin, D.E., Demougin, P., Hall, M.N. & Bellis, M. Rank Difference Analysis of Microarrays (RDAM), a novel approach to statistical analysis of microarray expression profiling data. BMC Bioinformatics 5, 148 (2004).
62. Navon, R. et al. Novel rank-based statistical methods reveal microRNAs with differential expression in multiple cancer types. PloS One 4, e8003 (2009).
63. Aerts, S. *et al.* Gene prioritization through genomic data fusion. *Nat Biotechnol.* **24**, 537-544 (2006).
64. Wilson, N.K. et al. Combinatorial Transcriptional Control In Blood Stem/Progenitor Cells: Genome-wide Analysis of Ten Major Transcriptional Regulators. Cell Stem Cell 7, 532-44 (2010).
65. Yu, M. et al. Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol Cell* **36**, 682-95 (2009).
66. Cheng, Y. et al. Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res.* **19**, 2172-84 (2009).
67. Kassouf, M.T. et al. Genome-wide identification of TAL1's functional targets: Insights into its mechanisms of action in primary erythroid cells. *Genome Res.* **20**, 1064-83 (2010).
68. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-89 (2010).
69. Lefterova, M.I. et al. Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages. *Mol. Cell. Biol.* **30**, 2078-89

(2010).

70. Fujiwara, T. et al. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell* **36**, 667-81 (2009).
71. Lin, Y.C. et al. A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. Nat Immunol. 11, 635-43 (2010).
72. Chia, N.-Y. et al. A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature* **468**, 316-320 (2010).