Figures and figure supplements

Discovering sparse transcription factor codes for cell states and state transitions during development

Leon A Furchtgott et al
Figure 1. The clear minimum pattern is robustly detected in leaf cell types throughout triplet lineages in B- and T-cell development. (A) Developmental lineage tree showing relationships among 41 cell types in B- and T-cell development (Heng et al., 2008). Three triplets – the minimal subset of the tree from which relative distances can be studied – are denoted, each including an intermediate root cell type (red) and terminal leaf cell types (blue). 150 triplets among all sets of three cell types within five steps on the lineage tree were extracted for pattern-detection. (B) For each triplet of cell types (left), each gene’s expression level can have the clear minimum pattern in either the root or the leaves (right, top box) where the distribution of gene expression levels in one cell type is well separated from the other two (left, p<0.005 in a two sample t-test); clear maximum pattern (left, bottom) in the root or the leaves: the gene has a clear maximum in one of the cell types (right, p<0.005 in a two sample t-test). (C) A histogram of the number of genes showing the clear minimum pattern among the 150 triplets with known developmental topology. Triplets in which the root has the most genes showing the pattern shown in red; triplets in which one of the leaves has the most genes showing the pattern shown in blue. None of the triplets with more than 10 genes showing the pattern have the most genes with a clear minimum in the root (no red in any histogram bar except for the

Furchtgott et al. eLife 2017;6:e20488. DOI: 10.7554/eLife.20488
left most)(D) Principal component analysis of microarray data from the cell types in B- and T-cell development does not reflect known lineage relationships in (A).

DOI: 10.7554/eLife.20488.002

The following source data is available for figure 1:

Source data 1. Triplets used for pattern detection.
DOI: 10.7554/eLife.20488.003

Source data 2. List of mouse transcription factors.
DOI: 10.7554/eLife.20488.004
We used a two-distribution t-test to quantify the likelihood of a gene (each denoted by a single dot) showing either the clear minimum (A) or clear maximum pattern (B). We show the p-value from this t-test and the cell type in which the gene shows differential expression. The p-value cutoff is used to identify genes which show these patterns with high significance (dotted line). In this particular triplet, there are fewer genes with a clear minimum and with a clear maximum in the root of the triplet (B.T1) than in either leaf, as seen by fewer genes in the first row to the right of the cutoff. (C) A histogram of the number of genes showing the clear maximum pattern among the 150 triplets with known developmental topology. Triplets in which the root as the most genes showing the pattern are shown in blue. The clear maximum pattern does not show strong predictive power over the lineage relationships between triplets, and even when a large number of genes exhibit the pattern, the majority often reach their maximum in a leaf.

Figure 1—figure supplement 1. Clear minimum and clear maximum patterns. (A, B) We used a two-distribution t-test to quantify the likelihood of a gene (each denoted by a single dot) showing either the clear minimum (A) or clear maximum pattern (B). We show the p-value from this t-test and the cell type in which the gene shows differential expression. The p-value cutoff is used to identify genes which show these patterns with high significance (dotted line). In this particular triplet, there are fewer genes with a clear minimum and with a clear maximum in the root of the triplet (B.T1) than in either leaf, as seen by fewer genes in the first row to the right of the cutoff. (C) A histogram of the number of genes showing the clear maximum pattern among the 150 triplets with known developmental topology. Triplets in which the root as the most genes showing the pattern are shown in blue. The clear maximum pattern does not show strong predictive power over the lineage relationships between triplets, and even when a large number of genes exhibit the pattern, the majority often reach their maximum in a leaf.

DOI: 10.7554/eLife.20488.005
Figure 1—figure supplement 2. The clear minimum pattern is observed across different types of triplets. (A) Histogram of the number of genes showing the clear minimum pattern evaluated over all 25,194 genes among the 150 triplets with known developmental topology. Triplets in which the
root has the most genes showing the pattern shown in red; triplets in which one of the leaves has the most genes showing the pattern shown in blue. (B) Histogram of the number of genes showing the clear minimum pattern evaluated over transcription factors, using FDR-adjusted p-values, among the 150 triplets with known developmental topology. Triplets in which the root has the most genes showing the pattern shown in red; triplets in which one of the leaves has the most genes showing the pattern shown in blue. (C) Number of genes showing the clear minimum pattern in the three cell types of each triplet. The number of genes in the root cell type of each triplet is indicated in blue; the number of genes in the two leaves are represented in yellow and red. The triplets are ordered according to the total number of genes showing the pattern in the triplet. (D–F) A histogram of the number of genes showing the clear minimum pattern. As in Figure 1C, triplets in which the root has the most genes showing the pattern shown in red; triplets in which one of the leaves has the most genes showing the pattern shown in blue. Triples are shaded according to different characteristics: length of triplet (D), presence of terminal node in triplet (E), cell-fate decision or lineage progression (F). DOI: 10.7554/eLife.20488.006
Figure 1—figure supplement 3. Distinction between related and unrelated triplets. (A) Plot, for triplets of related and unrelated cell types, of the entropy $S$ of the fraction of genes downregulated in the 3 cell types (x-axis) and...
Figure 1—figure supplement 3 continued

the total number of genes showing a minimum in any triplet for 100 unrelated and 150 related triplets (Figure 1—source data 1). Related triplets are shown in blue or red, depending on whether the cell type with the most genes showing the downregulated pattern is the root (red) or a leaf (blue), as in Figure 1C. Unrelated triplets are shown in green. For each triplet, we computed the entropy $S = - \sum_{i=A,B,C} f_i \log(f_i)$, where $f_i$ is the fraction of the genes which reach a clear minimum in cell type $i = A, B, C$. Jitter added to triplets with $S = 0$ for clarity. (B) Histogram of the maximum inferred probability of a non-null topology in the Bayesian algorithm for the set of related triplets (top) and the set of unrelated triplets (bottom). (C) Receiver operating characteristic (ROC) curve for the probability of a non-null topology (B) for binary classification of the related and unrelated triplets. The area under the curve (AUC) is 0.96.

DOI: 10.7554/eLife.20488.007
Figure 2. Identification of topology and transition genes (showing clear minimum pattern) for each triplet of cell types. (A) Schematic for the statistical inference of lineage topology for 3 cell types. Genes with a clear minimum pattern indicate which cell types that are not the root (see Figure 1C) and Figure 2 continued on next page.
Figure 2 continued

hence allow inference of the topological relationship. (B) Dot plot (each dot representing a gene) of the cell type that is most likely to have the minimum mean expression of each gene among CMP, ST and MPP as a function of the odds $O_i$ of that gene being a transition gene. Each gene votes against the topology whose root has the minimum mean among the three cell types, and this vote is weighted by the odds that the gene is a transition gene (Equation 1). Two groups of genes, labeled by their names, have high odds of being transition genes and thus cast a strong vote against CMP or MPP being the root. (C) The computed probability of the topology given gene expression data indicates 0.84 probability that ST is the intermediate type. (D) The plot of the probabilities of genes being transition genes for triplet ST/MPP/CMP, given gene expression data and that the topology is MPP – ST – CMP. The names of the 10 genes with the highest probability of being transition genes are shown. Probabilities are calculated assuming the prior odds $p(b_i = 1) = 0.05$ (see main text). There are two classes of transition genes: one for which gene expression in CMP is greater than expression in MPP (regular font), and another for which gene expression in MPP is greater than expression in CMP (bold font). (E) Plot of the replicates of ST, MPP and CMP in the gene-expression space of the two classes of transition genes (with probability > 0.8). Plotted on each axis is the mean normalized log expression level of the transition genes in the class, each class is denoted in curly brackets by the name of the transition gene with the highest probability. (F) Dot plot for triplet MEP/GMP/FrBC of the cell type that is most likely to have the minimum mean expression as a function of the odds $O_i$ of that gene being a transition gene. (G) The computed probability of the topology given gene expression data is the null hypothesis ($p=0.99$).

DOI: 10.7554/eLife.20488.008

The following source data is available for figure 2:

Source data 1. Early hematopoietic cell types considered.
DOI: 10.7554/eLife.20488.009

Source data 2. Probabilities of topologies for triplets of hematopoietic cell types.
DOI: 10.7554/eLife.20488.010

Source data 3. Probabilities of transition and marker genes for the hematopoietic lineage tree.
DOI: 10.7554/eLife.20488.011
Figure 2—figure supplement 1. Probability of topology depends on prior odds. (A) Two models of the hierarchy of early hematopoietic progenitors, both built based on prospective isolation of lineage restricted progenitors, include (left) the traditional model, in which the first split strictly separates...
myeloid and lymphoid lineages (Akashi et al., 2000; Kondo et al., 1997, Reya et al., 2001) and (right) an alternative hierarchy proposed by Adolfsson and colleagues, in which lymphoid progenitors subsequent to the first split retain some myeloid potential (Adolfsson et al., 2005). (B) The inferred probability of different topological configurations as a function of the prior odds of genes being transition genes. At low values of the prior odds, very few genes are given substantial votes in the inference procedure. The topology MPP-ST-CMP has high probability for a range of prior odds. (C) Genes with high probability of being marker genes for MPP (bold), CMP (plaintext), or ST (italics). (D) The triplet MEP, GMP, FrBC shows no significant probability of a non-null topology for a wide range of prior odds. In this case, our inference method makes no claim as to the topological configuration based on the gene expression evidence provided. (E) Histogram of the maximum inferred probability of a non-null topology for the 165 triplets of hematopoietic progenitors. (F) Projections of the early hematopoietic progenitors along first two principal components. (PC1: 30%; PC2: 17%). Each point represents a different replicate. Note the proximity between ETP and MPP samples (red circle), which does not reflect either of the lineage models shown in (A). DOI: 10.7554/eLife.20488.012
Figure 2—figure supplement 2. Plots of the length of the triplets distinguishing the traditional model and in the Adolfsson model. Separate plots are shown for triplets whose inferred topology is consistent with only the Adolfsson model (A) and the traditional model (B). Triplets whose inferred topology is null are shown in turquoise; those whose inferred topology is non-null are shown in black. Jitter was added to the points for clarity.

DOI: 10.7554/eLife.20488.013
Figure 3. Reconstruction of lineage tree and key transition gene for early hematopoiesis. (A) Final lineage tree, recapitulating the inferred triplet topologies, with top inferred transition genes indicated along cell fate decisions. (B) Plot of the replicates of different cell types in the gene-expression space of the transition gene classes (probability > 0.8) for 4 cell-fate transitions along the inferred lineage tree in (A). Plotted on each axis is the mean normalized log expression level of the transition genes in the class. The axis labels and data points are color-coded according to the colors in (A). (C) Table with selected transition genes for early hematopoietic cell-fate transitions, along with references to published validations of their functional role.

Genes known to be effective for reprogramming are shown in bold.

DOI: 10.7554/eLife.20488.014

Figure 3 continued on next page
Figure 3 continued

The following source data is available for figure 3:

**Source data 1.** Marker genes for early hematopoiesis.

DOI: 10.7554/eLife.20488.015
Figure 3—figure supplement 1. Reconstruction of lineage tree from individual triplets. (A–D) Probability of all topological configurations for a triplet as a function of prior odds for all triplets between 4 cell types: MPP, ST, LT, and CMP. (A) The LT-ST-MPP triplet, with \( p = 0.99 \). (B) The CMP-ST-MPP triplet, with \( p = 0.95 \). (C) The LT-ST-CMP triplet, with \( p = 0.99 \). (D) The MPP-LT-CMP triplet, with \( p = 0.65 \). Note that the next highest topology here is the null topology. (E) These four triplets can be combined to infer a simple relationship between all four cell types that is consistent with the topological restrictions inferred from each high probability triplet.

DOI: 10.7554/eLife.20488.016
Cell fate transition | High probability transition genes | References
--- | --- | ---
$C_2$-$C_7$-$C_3$ | *Rbm39, Ssrp1, Nfib, Tcf3, Nfkbie, Creb3l3, Irf3, Klf4* | Sillars-Hardebol et al., 2012; Garcia et al., 2011; Chang et al., 2013; Merrill et al., 2001; Wullaert et al., 2011; Li et al., 2009; Negishi et al., 2012; Yu et al., 2012; Katz et al., 2002; Mellor et al., 2013
$C_4$-$C_1$-$C_8$ | *Id3, Stat6, Esrra, Tcf3, Notch1, Atoh1, Spdef, Klf4, Neurog3, Foxa1* | De Santa Barbara et al., 2008; Khan et al., 2001; Knight et al., 2014; Merrill et al., 2001; Fre et al., 2005; VanDussen et al., 2010; Noah et al., 2010; Yu et al., 2012; Jenny et al., 2002; Ye and Kaestner, 2009
$C_4$-$C_2$-$C_8$ | *Id3, Stat6, Esrra, Hes1, Trp53, Uhrf1, Mycl1* | De Santa Barbara et al., 2008; Khan et al., 2001; Knight et al., 2014; Jensen et al., 2000; Spehlmann et al., 2013; Marjoram et al., 2015; Muncan et al., 2006
$C_4$-$C_2$-$C_1$ | *Atoh1, Klf4, Spdef, Nfkbie, Foxa2, Hes1, Trp53, Uhrf1* | VanDussen et al., 2010; Yu et al., 2012; Noah et al., 2010; Wullaert et al., 2011; Ye and Kaestner, 2009
$C_4$-$C_2$-$C_5$ | *Etv6, Spdef, Nfkbie, Hes1, Tcf3, Mycl1* | Wang et al., 2016; Noah et al., 2010; Wullaert et al., 2011; Jensen et al., 2000; Merrill et al., 2001; Muncan et al., 2006
$C_1$-$C_2$-$C_5$ | *Etv6, Cdk9, Hdac2, Atoh1, Wwtr1, Klf4* | Wang et al., 2016; Riccio et al., 2008; Tou et al., 2004; VanDussen et al., 2010; Varelas, 2014; Yu et al., 2012
$C_8$-$C_2$-$C_5$ | *Cdk9, Spdef, Nr5a2, Id3, Stat6, Tcf3, Wwtr1* | Riccio et al., 2008; Noah et al., 2010; Duggan et al., 2016; De Santa Barbara et al., 2008; Merrill et al., 2001; Varelas, 2014

**Figure 3—figure supplement 2.** Inferred lineage tree and transition genes for intestinal development. (A) Inferred lineage tree from intestinal single cell data taken from (Grun et al., 2015). Cluster identifications were taken from RaceID2 (Grun et al., 2016), and a similar tree was inferred from Figure 3—figure supplement 2 continued on next page
StemID ([Grun et al., 2016], not shown). Using 1369 transcription factors, we inferred these lineage relationships between all cell clusters with more than 10 cells per cluster. The tree differs from that in (Grun et al., 2016) only in the progression from C2 – C8 – C4, while we infer the triplet C4 – C2 – C8 with p>0.99, both of which are consistent with literature (van der Flier and Clevers, 2009), and await experimental validation. (B) High-probability transition genes with known relevance in intestinal homeostasis or differentiation taken from the top 20 transition genes for each lineage progression. DOI: 10.7554/eLife.20488.017
Figure 4. Inference of lineage tree and key transitions genes using single cell expression data from in vitro differentiated developing human brain. (A) RNA-seq data from single cells collected at days 12, 26, 54, and 80 from a human brain in vitro differentiation protocol (Yao et al., 2017) were analyzed.
using a variety of existing methods. Partitioning single-cells into cell types through non-linear dimensionality reduction using t-SNE (top) depends on the perplexity parameter (set here to 5, see (37x345)). Maximization algorithm to determine most likely cluster identities \( \{C\} \equiv \{c_1, c_2, \ldots, c_n\} \), sets of transitions \( \{T\} \), marker genes \( \{a_i\} = 1 \) and transition genes \( \{b_i\} = 1 \), given single-cell gene expression data \( \{g_i\} \). Starting from a seed clustering scheme \( \{C_0\} \), iterative maximization of the conditional probabilities \( p(T|\{a_i\}, \{b_i\}|\{g_i\}, \{C\}) \) and \( p(C|\{g_i\}, \{T\}, \{a_i\}, \{b_i\}) \) converges to most likely set \( (\{C\}, \{T\}, \{a_i\}, \{b_i\}) \). Cell-cell covariance matrix between cells using only the associated high probability marker and transition genes show the final cluster assignments \( c_0, c_2 \) and \( c_3 \) (right) in contrast to using all transcription factors (left). (37x345) Correlations between differentiated cell clusters (Figure 4—Figure Supplement 4D) and bulk population samples from brain regions (in vivo developmental human data) (Miller et al., 2014). Neuronal cell types can be identified with specific spatial regions of the brain to interpret the topology of the lineage tree. Expression signatures of SOX2+ cell types \( c_0, c_2 \) and \( c_3 \) were dominated by pluripotency factors, and are not shown. (37x377) Inferred lineage tree for brain development. Genes associated with neocortical development, and mid-/hind-brain progenitors, and specific neuronal cell types are identified as high probability transition genes and are corroborated by mapping information from in vivo clusters. Colors coded similarly to (37x337). D12/26/54/80 labels indicate time of collection of cells within each cell type. Prog refers to SOX2+ cells, Diff refers to SOX2-DCX+ cells (37x527). Source data is available for figure 4:

Source data 1. Final cluster identities of single cells from in vitro cortical differentiation.
DOI: 10.7554/eLife.20488.019
Source data 2. Probabilities of topologies for triplets of single-cell clusters.
DOI: 10.7554/eLife.20488.020
Source data 3. Probabilities of Transition and Marker Genes for the Human Brain Developmental Lineage Tree.
DOI: 10.7554/eLife.20488.021
Source data 4. Human Brain Development SmartSeq2 Census.
DOI: 10.7554/eLife.20488.022
Source data 5. List of Human Transcription Factors.
DOI: 10.7554/eLife.20488.023
Figure 4—figure supplement 1. Cluster identity and sparse coding in neuronal differentiation. (A–B) t-SNE of developing neuronal cells displays seemingly different number of cell types depending on the perplexity parameter. (C) SOX2 expression clearly falls into bimodal distribution. Cell types with high levels of SOX2 (blue, c₀, c₂, and c₃) are labeled progenitor cell populations. (D) DCX expression in final clusters is also bimodal, and cell types with high expression (green, c₁, c₄, c₆, and c₇) are labeled differentiated (post-mitotic) neurons. (E) We assembled 20 triplets that were inferred to be non-null, and had a maximal distance between leaf nodes of less than 5. Amongst these triplets, we attempted to infer the correct topology using a minimal subset of genes. Restricting the inference to the N genes per triplet with the greatest odds ratio of being transition genes, the tree can be reconstructed accurately for any N ≥ 4.

DOI: 10.7554/eLife.20488.025
Figure 4—figure supplement 2. A selection of recent lineage-determination methods for single cell transcriptomic analysis applied to an in vitro neuronal differentiation data set (Yao et al., 2017). Each of these methods was run with multiple sets of parameters in an attempt to optimize the results. Figure 4—figure supplement 2 continued on next page.
lineage inference; however, it is possible that an unexplored parameter regime might yield more interpretable results, although it is unclear which parameters these might be. (A) Monocle2 (Trapnell et al., 2014) shows a complex tree with clear progression and multiple branches, but does not separate progenitor cells (DCX- cells) from neurons (DCX+ cells): progenitor cells are known to give rise to neurons, whereas neurons are post-mitotic and do not give rise to progenitors, in contradiction with certain portions of the tree. In addition, portions of the tree are in conflict with the general time point information specifying when cells were collected (Yao et al., 2017). (B) TSCAN (Ji and Ji, 2016) shows separation of DCX- and DCX+ cells along PC dimension 2 (note that cells are not labeled by DCX expression in this plot), but fails to capture structure among the differentiated neurons, particularly a fore-/hind-brain split. (C) StemID (Grun et al., 2016) produces a complex lineage that is not directly interpretable, and no clear separation of progenitors from neurons or forebrain-like from mid/hindbrain-like neurons is apparent.

DOI: 10.7554/eLife.20488.026