eLIFE elifesciences.org

RESEARCH ARTICLE

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

Leon A Furchtgott[1,2*†], Samuel Melton[1,3†], Vilas Menon[4,5], Sharad Ramanathan[1,3,4,6,7*]

[1]FAS Center for Systems Biology, Harvard University, Cambridge, United States; [2]Biophysics Program, Harvard University, Cambridge, United States; [3]Harvard Stem Cell Institute, Harvard University, Cambridge, United States; [4]Allen Institute for Brain Science, Seattle, United States; [5]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, United States; [6]Department of Molecular and Cellular Biology, Harvard University, Cambridge, United States; [7]School of Engineering and Applied Sciences, Harvard University, Cambridge, United States

Abstract Computational analysis of gene expression to determine both the sequence of lineage choices made by multipotent cells and to identify the genes influencing these decisions is challenging. Here we discover a pattern in the expression levels of a sparse subset of genes among cell types in B- and T-cell developmental lineages that correlates with developmental topologies. We develop a statistical framework using this pattern to simultaneously infer lineage transitions and the genes that determine these relationships. We use this technique to reconstruct the early hematopoietic and intestinal developmental trees. We extend this framework to analyze single-cell RNA-seq data from early human cortical development, inferring a neocortical-hindbrain split in early progenitor cells and the key genes that could control this lineage decision. Our work allows us to simultaneously infer both the identity and lineage of cell types as well as a small set of key genes whose expression patterns reflect these relationships.

DOI: 10.7554/eLife.20488.001

***For correspondence:** leon.furchtgott@gmail.com (LAF); sharad@cgr.harvard.edu (SR)

†These authors contributed equally to this work

Competing interests: The authors declare that no competing interests exist.

Funding: See page 26

Received: 09 August 2016
Accepted: 31 January 2017
Published: 15 March 2017

Reviewing editor: Nir Yosef, University of California, United States

Introduction

During development, pluripotent cells make a series of lineage decisions to give rise to the different cell types of the body. These lineage decisions are controlled by intra-cellular molecular networks that include transcription factors and signaling molecules. There are two fundamental challenges associated with understanding the differentiation of individual cells. The first is to identify lineage relationships: how cells and their progeny move from pluripotent through intermediate to terminally differentiated cell states. The second is to identify the key molecular drivers that allow cells to make fate decisions along their developmental trajectory.

Reconstructing cell lineages has traditionally involved prospectively tracking cells and their progeny using a variety of imaging or genetic tools (***Buckingham and Meilhac, 2011***; ***Frumkin et al., 2008***; ***Orkin and Zon, 2008***; ***Sulston et al., 1983***). Recent progress in single-cell sequencing techniques (***Grün et al., 2015***; ***Jaitin et al., 2014***; ***Macosko et al., 2015***; ***Patel et al., 2014***; ***Paul et al., 2015***; ***Treutlein et al., 2014***; ***Zeisel et al., 2015***) allows for a complementary view of the transcriptional states of individual cells during the course of development, providing static snapshots of the dynamics of the underlying molecular network. But inferring lineage relationships and the dynamics of the underlying molecular networks has proved difficult using transcriptional data alone, in part because of the high dimensional nature of these data. Overcoming this challenge would be

particularly useful for understanding the development of human organs such as the brain where traditional lineage tracing experiments are more difficult.

High dimensional data analysis techniques such as PCA, ICA, or t-SNE, are useful at reducing dimensionality. However, since the resulting axes represent linear combinations of a large number of features (for example, expression levels of each gene), interpreting the analysis or making experimental predictions is sometimes challenging. Meanwhile, traditional statistical methods such as linear multivariate regression have limited applicability for detecting patterns in high dimensional data (***Advani and Ganguli, 2016***; ***Donoho and Tanner, 2009***). The challenges inherent in high-dimensional data analysis such as identifying discriminatory features are further exacerbated as the fraction of relevant features decreases (***Donoho and Tanner, 2009***). Computational techniques currently in use to cluster single cell data or to infer relationships among cells are built on these approaches, thereby assuming that all high-variance genes are equally relevant for pattern detection (***Marco et al., 2014***; ***Satija et al., 2015***; ***Trapnell et al., 2014***). In contrast, decades of work in developmental biology have revealed that combinations of a few transcription factors can be sufficient to experimentally perturb cell fate and developmental decisions (***Gilbert, 2014***; ***Graf and Enver, 2009***; ***Takahashi and Yamanaka, 2006***) suggesting that the expression patterns of a few genes may be most relevant for making computational inferences. Therefore, there is a need to detect patterns involving a small fraction of all genes. Unfortunately, except in the case of well-studied lineage decisions, we do not know the identity of this fraction.

In statistics, techniques relying on L1 regularization have been successful in contexts where the number of informative variables is known to be small but whose identities are unknown, both for regression problems (***Baraniuk, 2007***; ***Candès et al., 2006***; ***Tibshirani, 1996***; ***Wainwright, 2009***) and for clustering (***Witten and Tibshirani, 2010***). Inspired by these successes in statistics, our aim here is to discover generalizable sparse patterns in gene expression data during development (if they exist), and to exploit these patterns to computationally infer the dynamics of cell state transitions from high-dimensional transcriptional data obtained during the course of development.

In this manuscript we analyze expression patterns among cell types with known lineage relationships in late hematopoiesis and discover a pattern in a sparse subset of genes that correlates with these relationships. We develop a Bayesian framework based on this gene expression pattern to simultaneously infer lineage transitions and the key genes that drive them. We apply this method to reconstruct the lineage tree among a different set of cell types in early hematopoietic development, and in this process identify many known drivers of early hematopoiesis, including *Gata1*, *Cebpa* and *Ebf1*. We further extend our method to analyze single-cell gene expression data, using genes exhibiting the discovered pattern to cluster cells from early brain development and to infer lineage relationships between these clusters. Our analysis reveals a split from early progenitors to putative neocortex and mid/hindbrain cell types, as evidenced by the mutually exclusive expression of region-specific genes such as *FOXG1*, *LHX1*, and *POU3F2* (*BRN2*). This prediction was validated experimentally in a separate work (***Yao et al., 2017***). We finally discuss the advantages of using sparse patterns for making inferences and for modeling the underlying gene regulatory networks.

Results

Discovering sparse patterns correlated with lineage transitions

In order to identify gene expression patterns that are robustly predictive of lineage relationships, we analyzed gene expression data from 41 cell types during B- and T- cell development that have an experimentally established developmental lineage (***Figure 1A***, ***Heng et al., 2008***). We searched for sparse patterns of gene expression amongst groups of three cell types from this collection; subsets of three are the minimal set in which measures of relative similarity can be used infer relative lineage relationships.

We identified 150 triplets of cell types with experimentally verified lineage relationships from B- and T- cell development (***Heng et al., 2008***) (***Figure 1—source data 1***). Three such triplets are shown in ***Figure 1A***. These triplets constituted both cell fate decisions (for example, cell type A gives rise to cell type B and C) and lineage progressions (cell type B gives rise to cell type A which then gives rise to cell type C). For each triplet, we noted which cell type was the progenitor or intermediate cell type ('root' cell type A) and which cell types were not ('leaf' cell types B and C). Note

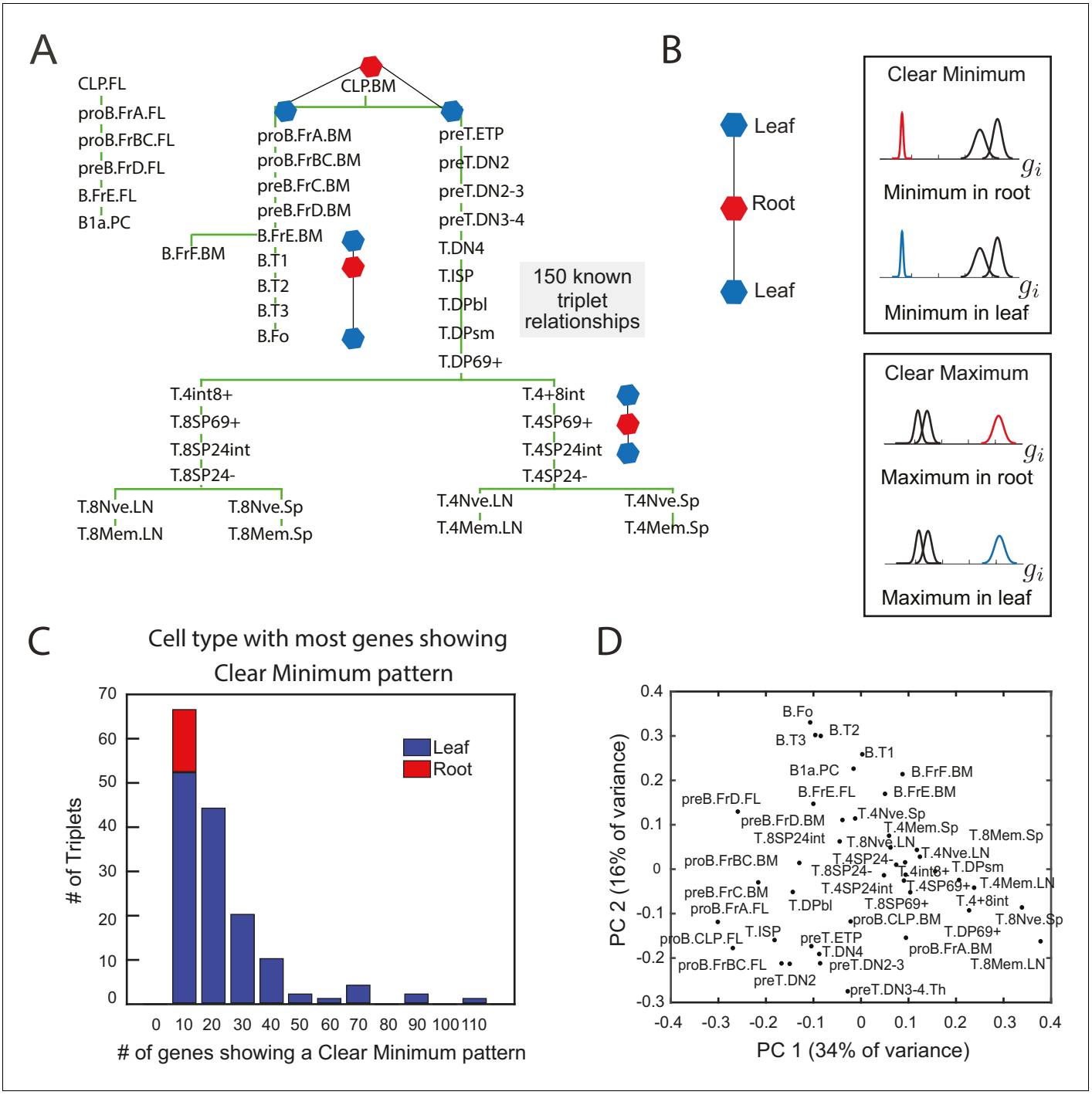


Figure 1. The clear minimum pattern is robustly detected in leaf cell types throughout triplet lineages in B- and T-cell development. (**A**) Developmental lineage tree showing relationships among 41 cell types in B- and T-cell development (***Heng et al., 2008***). Three triplets – the minimal subset 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

Figure 1 continued on next page

Figure 1 continued

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 and figure supplements are 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

Figure supplement 1. Clear minimum and clear maximum patterns.

DOI: 10.7554/eLife.20488.005

Figure supplement 2. The clear minimum pattern is observed across different types of triplets.

DOI: 10.7554/eLife.20488.006

Figure supplement 3. Distinction between related and unrelated triplets.

DOI: 10.7554/eLife.20488.007

that even in the case in which cell type B gives rise to cell type A which gives rise to cell type C, we will refer to A as the 'root' and B and C as 'leaves' for that particular triplet. We analyzed transcription factor gene expression data for these triplets from the Immunological Genome Consortium (***Heng et al., 2008***) since transcription factors are the ultimate drivers of cell fate decisions.

Surprisingly, we found that specific expression patterns involving only single genes correlated well with lineage relationships between three cell types. Genes that are not differentially expressed within a triplet of cell types convey no information about relationships between the cell types. Therefore, we select genes with expression variability among the three cell types. The expression pattern of such genes can belong to one of only two possible patterns: (***Figure 1B***): the *clear minimum pattern*: the gene has a clear minimum level in one of the three cell types, with its distribution of gene expression levels being well-separated from the other two ($p<0.005$ in both two sample *t*-tests between the minimum cell type and the other two cell types; ***Figure 1—figure supplement 1A***); the *clear maximum pattern*: the gene has a clear maximum level in one of the cell types ($p<0.005$ in both sample *t*-tests; ***Figure 1—figure supplement 1B***). Note that both patterns can be satisfied simultaneously if the distribution of expression levels in the three cell types are all well-separated with a clear maximum and minimum. We tested if either of the two patterns correlated with the lineage topologies between the three cell types with known lineage relationships (***Figure 1C***, ***Figure 1—figure supplement 1C***).

Triplets of cell types can be separated into categories based on how many genes exhibit the aforementioned patterns. 56% of the triplets contained more than 10 genes exhibiting the clear minimum pattern, and in 100% of these triplets the majority of genes with expression fitting this pattern reached their minimum expression in one of the leaves (***Figure 1C***) and never in the root of the triplet. The frequency with which the pattern correctly indicated the lineage relationship increased with the number of genes within a triplet exhibiting the pattern, thus suggesting a confidence measure. The genes showing the clear minimum pattern fell into two distinct groups, corresponding to whether the minimum expression level was in one or the other of the two leaves (***Figure 1—figure supplement 1A***; ***Figure 1—figure supplement 2C***). Thus the expression pattern of the total set of clear minimum genes correlated with the topology. Since genes showing the clear minimum pattern correlated with lineage relationships (***Figure 1C***) between cell states both in the case of branches and linear sequences of cell state transitions, we refer to them as *transition genes*.

We further verified that the clear minimum pattern could be observed (a) in the set of all 25,194 genes (***Figure 1—figure supplement 2A***), (b) using FDR-adjusted p-values (***Benjamini and Hochberg, 1995***) (***Figure 1—figure supplement 2B***), (c) in triplets of different lengths along the lineage tree (***Figure 1—figure supplement 2D***), (d) in triplets both containing only internal nodes and including terminal nodes (***Figure 1—figure supplement 2E***), (e) and in both lineage progression and cell fate decision triplets (***Figure 1—figure supplement 2F***).

The clear maximum pattern was a poorer indicator of lineage relationships (***Figure 1—figure supplement 1C***). 83% of the triplets had more than 10 genes exhibiting this pattern, but 10% of those showed the majority of genes with expression fitting this pattern reaching their maximum in the root

while the others did so in the leaves. Crucially, the integrity of the relationship between the clear maximum pattern and lineage topology was not correlated with the number of genes exhibiting the pattern (***Figure 1—figure supplement 1C***). While the clear maximum pattern did not correlate with lineage relationships, genes exhibiting this pattern identify individual cell types, and therefore we will refer to them as *marker genes*.

There are many examples of genes known to be functionally important for lineage decisions whose expression patterns fit the clear minimum pattern. In the case of lateral inhibition commonly used during development, progenitor cells express genes together (for example, *Notch* and *Delta*) which are differentially expressed in the differentiated states (only *Notch* or only *Delta*) (***Perrimon et al., 2012***) reaching the minimum expression level in one of the leaves. The same pattern is also seen in multiple examples of lineage decisions often involving mutual inhibition, where key genes expressed in the progenitor are differentially regulated in the progeny (***Graf and Enver, 2009***; ***Qi et al., 2013***; ***Thomson et al., 2011***; ***Zhang et al., 1999***). In each of these cases, key genes reach minimal expression levels in one of the leaves of the triplets.

The observation that genes exhibiting the clear minimum pattern are correlated with the lineage topology of a triplet of cell types further revealed that (i) only this fraction of transcription factors can be useful for inferring lineage relationships, and (ii) the identity of this fraction depends on which group of three cell types were analyzed. As the subset of genes that are informative varies based on the triplet of cell types being considered, establishing lineage relationship between all cell types at once as opposed to three at a time could be challenging. Indeed, our attempt to reconstruct the lineage relationships between all cell types using methods based on PCA failed (***Figure 1D***).

We further evaluated the clear minimum pattern in 100 triplets in which there was no clear relation between the cell types (***Figure 1—source data 1***). We found that while there were a substantial number of genes exhibiting a clear minimum in one of the three cell types in unrelated triplets, their minima were evenly distributed amongst the cell types. To quantify that the minima were evenly distributed, we counted the fraction of genes f_i which reached a clear minimum in cell type i = *A, B, C* (for A, B, and C unrelated cell types), and for each triplet, we computed the entropy $S = -\sum_{i=A,B,C} f_i \log(f_i)$. We compared the distribution of the entropy and the number of genes showing a minimum in any triplet for unrelated and related triplets (***Figure 1—figure supplement 3A***). The unrelated triplets have higher entropy and typically more genes with a minimum level. This suggested that unrelated triplets show a distinct pattern from the related triplets.

Using patterns to infer lineages

Together, these observations suggested a strategy for inferring the lineage topology between three cell types: each gene showing the clear minimum pattern with a minimum expression level in a particular cell type increases to the probability that this cell type is not the root of the topology (***Figure 2A***). We next developed a statistical machinery to systematically detect this pattern in gene expression data and to use the resulting sparse subset of genes to infer lineage relationships between three cell types at a time. We then used the inferred relationships between all sets of three cell types as constraints to determine the full developmental lineage tree.

The classifications of genes as transition or marker genes in ***Figure 1*** were based on p<0.005 in a two-sample *t*-test. To implement such classifications probabilistically without arbitrary cutoffs we developed a statistical framework to infer the lineage relationships between each set of three cell types A, B, and C and find the key sets of transition genes (those genes that show the clear minimum pattern), given gene expression data $\left\{g_i^{A,B,C}\right\}$ in those cell types. We determined the probability of any possible topological relationship between the cell types $T = \{A,\ B,\ C,\ \emptyset\}$ referring to cell type A, B, or C being the root of the triplet, or $\emptyset$ which corresponds to the case where the data does not suggest any lineage relationship between the three cell types because either no significant pattern could be detected, or multiple genes exhibiting the minimum pattern suggested conflicting topologies. Rather than an absolute classification of genes as showing a pattern or not, we calculated the probability $p\left(\alpha_i = 1 \mid g_i^{A,B,C}\right)$ of each gene i being a marker gene (denoted by $\alpha_i = 1$), i.e. gene i showing the clear maximum pattern, and the probability $p\left(\beta_i = 1 \mid g_i^{A,B,C}\right)$ of it being a transition gene denoted by $\beta_i = 1$, i.e. gene i showing the clear minimum pattern (Materials and methods).

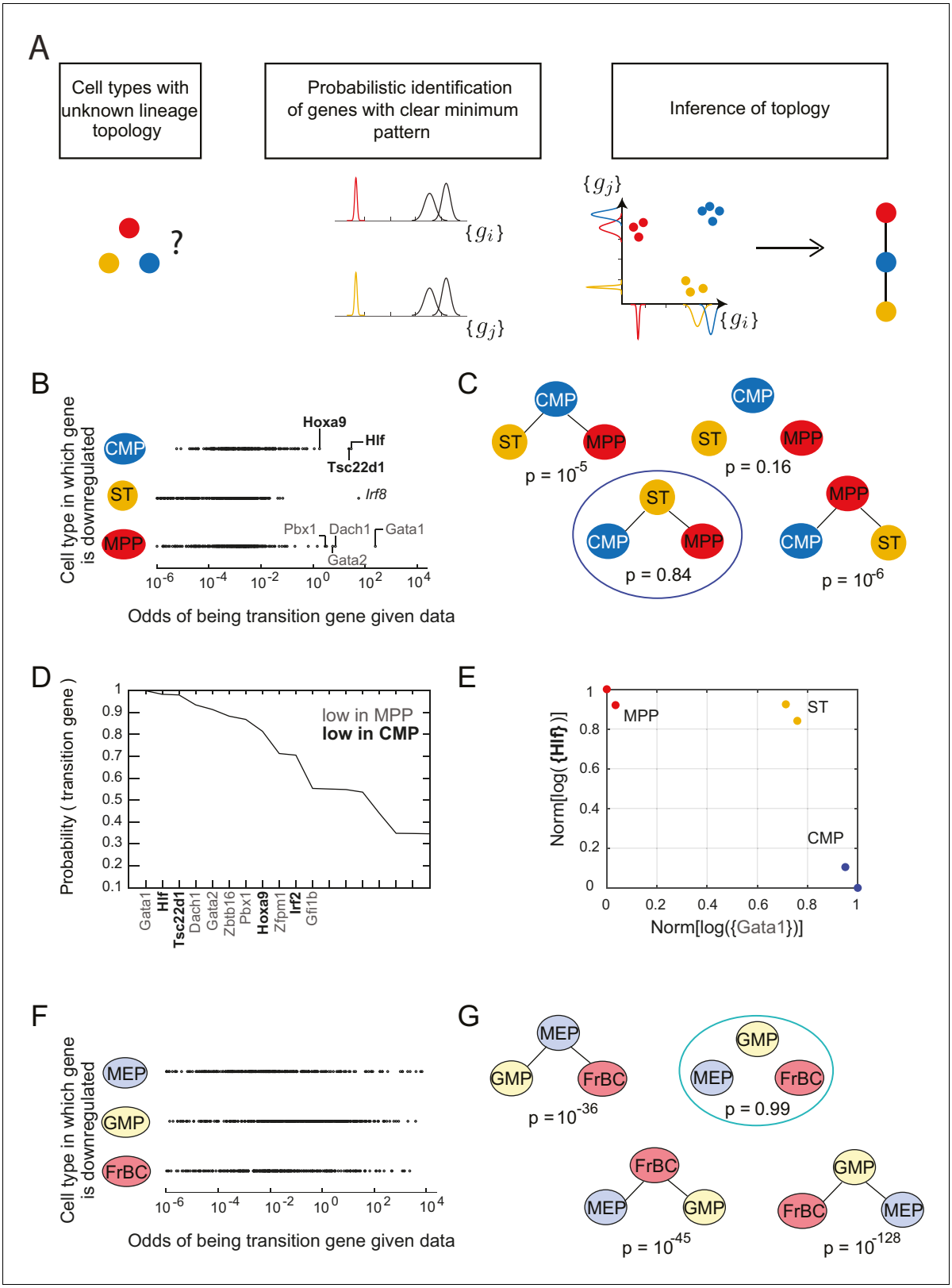


Figure 2. Identification of topology and transition genes (showing clear minimum pattern) for each triplet of cell types. (**A**) Schematic for the statistical inference of lineage topology for 3 cell types. Genes with a clear minimum pattern indicate which cell types that are not the root (see ***Figure 1C***) and 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 $\mathcal{O}_i$ of that gene being a transition gene. Each gene votes

Figure 2 continued on next page

Figure 2 continued

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 $\frac{p(\beta_i=1)}{p(\beta_i=0)}$ = 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 $\mathcal{O}_i$ 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 and figure supplements are 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 supplement 1. Probability of topology depends on prior odds.

DOI: 10.7554/eLife.20488.012

Figure supplement 2. Plots of the length of the triplets distinguishing the traditional model and in the Adolfsson model.

DOI: 10.7554/eLife.20488.013

We derived the probability of a given topology T given the expression data as (Materials and methods):

$$p\left(T \mid \left\{g_i^{A,B,C}\right\}\right) \propto \prod_i \left(1+\frac{3}{2}\mathcal{O}_i\left[1-p\left(\mu_T^i \,\mathrm{is\,min} \mid g_i^{A,B,C}\right)\right]\right), \quad (1)$$

where, $\mathcal{O}_i = p\left(\beta_i = 1 \mid g_i^{A,B,C}\right)/p\left(\beta_i = 0 \mid g_i^{A,B,C}\right)$ is the odds of gene i being a transition gene and thus having a unique minimum. The term $p\left(\mu_T^i \,\mathrm{is\,min} \mid g_i^{A,B,C}\right)$ is the probability that the mean μ_T^i of the distribution of the expression levels of gene i in the root cell type T is less than the mean in the other two cell types. The odds implicitly contains the only free parameter in our analysis, the prior odds $\frac{p(\beta_i=1)}{p(\beta_i=0)}$, which defines the number of genes we expect to show the clear minimum pattern *a priori*, and functions as a sparsity parameter for the inference. Qualitatively, in the above equation, every gene casts a vote $-p\left(\mu_T^i \,\mathrm{is\,min} \mid g_i^{A,B,C}\right)$ against the cell type T in which its mean expression is minimal being the root. Further, this vote is weighted by the odds $\mathcal{O}_i$ of gene i being a transition gene. Thus, genes with a clearer minimum pattern get larger votes in determining which cell type is not the root. In practice, these quantities are computed numerically (Materials and methods).

We note further that if a substantial number of genes cast votes against each of the cell types, then the probability of the null topology ø increases. We computed the probability of obtaining the null topology among the 150 related triplets and 100 unrelated triplets from our training set. The distribution of the probability of obtaining the null topology was considerably different between the related triplets and the unrelated triplets, with an AUC of 0.96 (***Figure 1—figure supplement 3B–C***).

Application to hematopoietic gene expression data

We used our statistical framework to recreate the lineage of early hematopoietic differentiation. We considered 11 early hematopoietic progenitors from the ImmGen Consortium microarray data set (***Heng et al., 2008***) (***Figure 2—source data 1***). These cell types and their associated relationships were not included in the data set used earlier to study the correlations of the two patterns and lineage topologies. Several features of the early hematopoietic lineage tree are debated

(*Adolfsson et al., 2005*; *Iwasaki and Akashi, 2007*) (*Figure 2—figure supplement 2A*). Given only the gene expression data for these different subpopulations of cells, we determined the lineage relationships and the key factors associated with each lineage decision. We calculated the probabilities of topology and marker and transition genes for the $\binom{11}{3} = 165$ possible triplets of cell types using our statistical framework (*Figure 2—source data 2*). To illustrate our method, we first described the analysis of the expression data from two such triplets of cell types: CMP/ST/MPP and MEP/GMP/FrBC (*Figure 2B–G*). We then assembled the triplets to form an undirected lineage tree (*Figure 3*; *Video 1*).

Following *Equation 1*, each gene votes against the topology whose central node has the minimum expression of that gene among the three cell types, and this vote is weighted by the odds that the gene is a transition gene. To illustrate this for the triplet of cell types CMP, ST and MPP, we plotted the topology each gene voted most against, *i.e.* the topology T for which $p\left(\mu_T^i \text{ is min} \mid g_i^{CMP,ST,MPP}\right)$ is the maximum, versus the odds $\mathcal{O}_i$ of that gene being a transition gene (*Figure 2B*).

We find two groups of genes that are much more likely to be transition genes than any of the other genes, with values of $O_i \sim 10^2$ compared to 10^0 at most for other genes (*Figure 2B*, regular and bold fonts). These two groups of genes have a large value for either $p\left(\mu_{CMP}^i \text{ is min} \mid g_i^{CMP,ST,MPP}\right)$ or $p\left(\mu_{MPP}^i \text{ is min} \mid g_i^{CMP,ST,MPP}\right)$ and thus vote against $T = \boldsymbol{CMP}$ (cell type CMP is the intermediate) or against $T = \boldsymbol{MPP}$ (cell type MPP is the intermediate). Together these genes that have a high odds of being transition genes appear to most support topology $T = \boldsymbol{ST} \equiv CMP - \boldsymbol{ST} - MPP$.

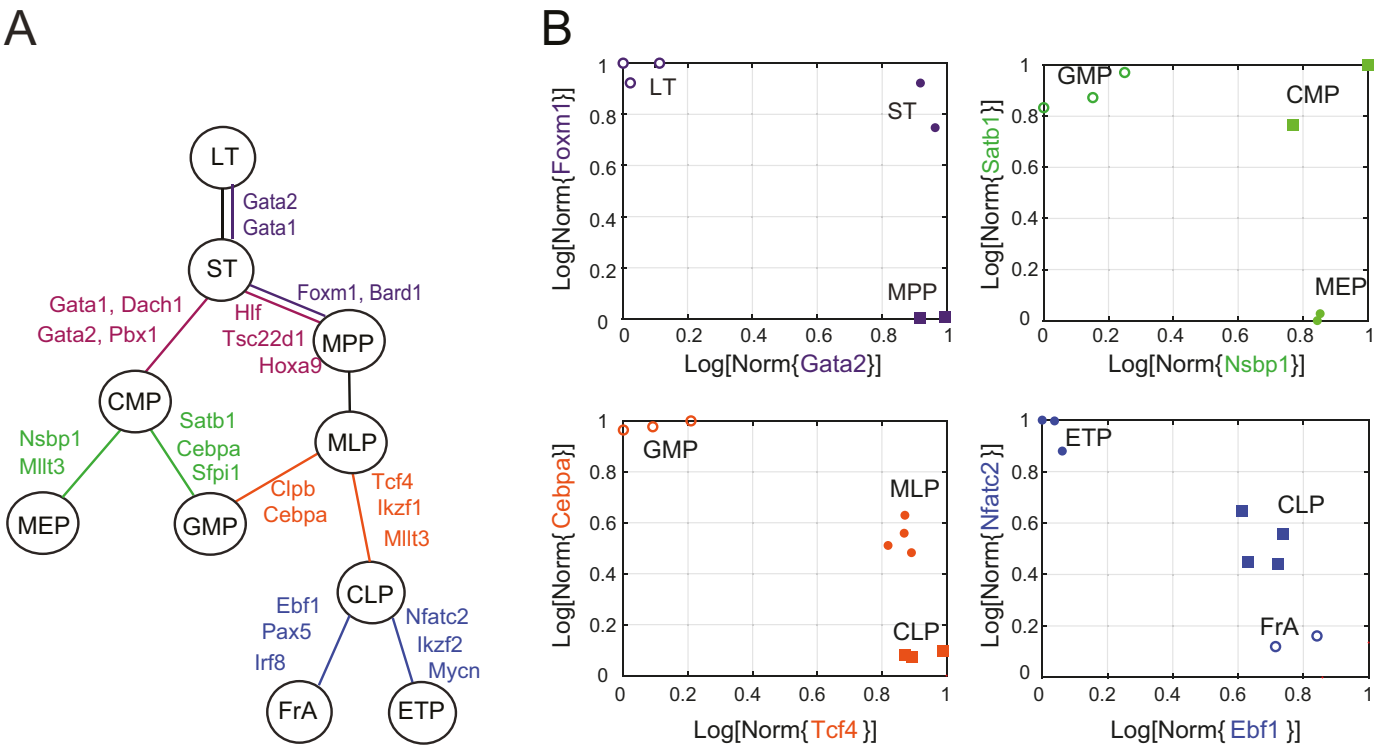
In fact, the intuition in *Figure 2B* is borne out in the calculation of $p\left(T \mid \left\{g_i^{CMP,ST,MPP}\right\}\right)$. Using *Equation 1* above and assuming a sparsity parameter of 0.05, we calculate that there is an 84% chance that the topology is $\boldsymbol{ST}$ (*Figure 2C*; *Figure 2—figure supplement 2B*). Although gene *Irf8* (*Figure 2B*, italic font) is strongly downregulated in ST and is expressed at higher levels in CMP and MPP (*Figure 2C*), we note that its signal is overwhelmed by the large number of genes downregulated in either CMP or MPP, illustrating the statistical nature of the framework.

For each triplet, we evaluated each gene's probability of being a transition or marker gene (*Figure 2—source data 3*). *Figure 2D* shows the names and associated probabilities of the 12 genes most likely to be transition genes for the triplet $\text{CMP} - \mathbf{ST} - \text{MPP}$. The transition genes fall into two groups, corresponding to the two groups in *Figure 2B*. One group, which includes genes *Gata1*, *Dach1*, and *Gata2*, has higher expression in CMP than in MPP; the other group, which includes *Hlf*, *Tsc22d1*, and *Hoxa9*, has higher expression in MPP. Although the values of the probabilities of the genes being transition genes vary with the value of the sparsity parameter, the relative order of different genes does not change. The genes identified include many genes previously identified as being important for lineage specification (*Crispino, 2005*; *Gazit et al., 2013*; *Miyawaki et al., 2015*). The transition genes we discovered thus not only have gene expression patterns that reflect the lineage decision but also include functionally important genes.

In addition to the transition genes, we identified marker genes ($p(\alpha_i = 1 \mid \{g_i\}, T) > 0.8$) present only in ST (including *Mpl* and *Rai14*, consistent with [*Solar et al., 1998*]) and then symmetrically downregulated in both CMP and MPP (*Figure 2—figure supplement 1C*). Marker genes for CMP include *Srf*, *Zeb2*, *Rbpj* and *Irf8* (consistent with [*Goossens et al., 2011*; *Kurotaki et al., 2013*; *Ragu et al., 2010*; *Robert-Moreno et al., 2005*; *Tamura et al., 2000*]); marker genes for MPP include *Satb1*, consistent with (*Satoh et al., 2013*). Although these genes were not used to determine the topology, they are good markers for cell types ST, CMP and MPP.

Plotting the cell types using the mean expression levels of the two transition gene class captures the fork in the gene expression space associated with the cell-fate decision (*Figure 2E*). In contrast with the PCA analysis of the cell types (*Figure 2—figure supplement 1E*), in which MPP appears to be an intermediate between the hematopoietic stem cell types (LT and ST) and CMP, the projection of the cell types onto the transition-gene subspace shows that ST splits into CMP and MPP.

In contrast to the case of the triplet of cell types CMP/ST/MPP, for triplet MEP/GMP/FrBC, the distributions of genes supporting different topologies are similar (*Figure 2F*). Thus the most likely topology calculated using *Equation 1* is the null hypothesis (99%), which is that transition genes, if



C

Cell fate transition	High probability transition genes	References
CMP – **ST** – MPP	*Gata1*, *Dach1*, ***Gata2***, ***Hlf***, *Tsc22d1*, ***Hoxa9***	Crispino, 2005; Gazit et al., 2013; Miyawaki et al., 2015; Riddell et al., 2014; Doulatov et al., 2013; Batta et al., 2014
MEP – **CMP** – GMP	***Gata1***, *Mllt3*, ***Sfpi1***, ***Cebpa***, *Gfi1*	Pina et al., 2008; Zhang et al., 1999; Koschmieder et al., 2005; van der Meer et al., 2010; Radomska et al., 1998; Zhang et al., 1999; Iwasaki and Akashi, 2007
GMP – **MLP** – CLP	*Satb1*, *Ikzf1*, *Notch1*	Rebollo and Schmitt, 2003; Satoh et al., 2013; Stier et al., 2002
ETP – **CLP** – FrA / FrBC	***Ebf1***, *Pax5*, *Irf8*, *Nfatc2*, *Runx2*	Busslinger, 2004; Macian, 2005; Vaillant et al., 2002; Wang et al., 2008; Pongubala et al., 2008

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 and figure supplements are available for figure 3:

Source data 1. Marker genes for early hematopoiesis.

DOI: 10.7554/eLife.20488.015

Figure supplement 1. Reconstruction of lineage tree from individual triplets.

DOI: 10.7554/eLife.20488.016

Figure supplement 2. Inferred lineage tree and transition genes for intestinal development.

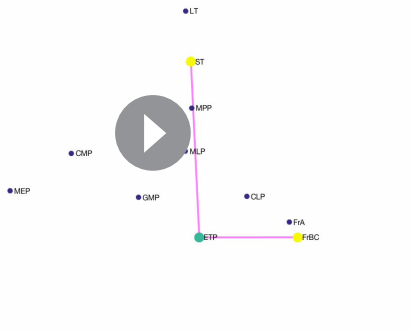
DOI: 10.7554/eLife.20488.017

they exist, do not have patterns that depend on the cellular topology (***Figure 2G***, ***Figure 2—figure supplement 1D***), in which case there is insufficient evidence to classify the triplet according to a particular non-null topology. The distribution of the maximal probability of non-null topologies in the different triplets is heavily concentrated near 1, allowing for clear separation between null and non-null triplets (***Figure 2—figure supplement 1E***). Null topologies were identified by the algorithm for triplets with cells that are from three terminal nodes (for example, the triplet MEP/GMP/FrBC) or from triplets that contain very distantly related triplets (for example, LT/CLP/ETP) (***Figure 2—source data 2***).

A lineage tree for early hematopoiesis

We next reconstructed the early hematopoietic lineage tree and identified transition genes involved in the different cell state transitions using all the non-null triplet relationships as constraints on the lineage relationships between all cell types. Out of the 165 possible triplets of hematopoietic progenitors, 144 showed one single non-null topology with probability greater than 0.6 over a range of the prior odds from 10^{-6} to 10^{2}. We next determined an undirected graph that recapitulates all of the individual triplet topologies (note that we are only inferring triplet topologies and are not inferring directionality). For example, although triplet CMP/LT/MPP has topology CMP – **LT** – MPP (***Figure 3—figure supplement 1A***), we could determine that LT cannot be the *direct* progenitor of CMP or MPP, because ST is an intermediate between LT and both cell types (***Figure 3—figure supplement 1B–C***). We could thus 'prune' this triplet when inferring the full graph (***Figure 3—figure supplement 1D–E***). A visualization of the pruning process is shown in ***Video 1***, where successive triplets are added to the graph, creating new edges and pruning others, leading to the final undirected tree. In practice, though, the pruning process was performed on all triplets simultaneously, not in succession.

The ST/CMP/MPP triplet (***Figure 2B–E***) immediately distinguishes between two competing models regarding the hierarchy of early hematopoietic progenitors. According to the traditional picture (***Iwasaki and Akashi, 2007***), MPP is the progenitor of CMP, and ST is the progenitor of MPP – therefore MPP should be an intermediate between ST and CMP and the topology of triplet ST/MPP/CMP should be ST –



Video 1. Tree-building process for early hematopoietic lineage Animation of the triplet assembly and pruning process for reconstructing the early hematopoietic lineage. For illustrative purposes, triplets (with p>0.6) are successively selected at random (in practice, the assembly and pruning process was performed on all triplets simultaneously; the resulting tree does not depend on the order in which the triplets are selected). The nodes of the current triplet are highlighted in yellow; if a topology is recognized for the triplet, the root is shown in green and the leaves in yellow, and the triplet edges are shown in magenta. If adding the triplet causes another triplet to be pruned, the soon-to-be-pruned (i.e. offending) edge is highlighted in red. The resulting pruned graph is then shown before adding the next triplet. As more triplets are considered, more edges between nodes are added and then pruned, leading to the final tree.

DOI: 10.7554/eLife.20488.018

MPP – CMP (***Figure 2—figure supplement 1A***, left). According to a model suggested by Adolfsson and colleagues (***Adolfsson et al., 2005***), ST splits into CMP and MPP (***Figure 2—figure supplement 1A***, right), and the topology should be CMP – **ST** – MPP. We identify both CMP – **ST** – MPP and CMP – **LT** – MPP as the correct topologies, lending support to the Adolfsson model. The Adolfsson and traditional models differ in the topology of 9 triplets. The inferred expected topologies of 8 out of these nine triplets support the Adolfsson model, which led to the identification of the final tree (***Figure 2—figure supplement 2***).

The lineage tree that we determined is consistent with the Adolfsson model and contains three additional lineage decisions (***Figure 3A–B***). First, CMP gives rise to MEP (megakaryocyte/erythroid progenitor) and GMP (granulocyte/macrophage progenitor). Second, MPP gives rise to MLP (multilineage progenitor), which then splits into the GMP and CLP (common lymphoid progenitor) cell types. In the final lineage decision, CLP gives rise to the ETP (pre-T) and FrA (pre-pro-B) cell types.

For each triplet of cell types along the tree, we identified transition and marker classes of genes. Among the 14 triplets that contained only adjacent cell types, we identified on average 24 marker genes per cell type and 25 transition genes (probability threshold of 0.8, prior odds $p(\beta_i = 1)/p(\beta_i = 0) = 0.05$). Many genes we discovered as belonging with high probability to the transition and marker classes of genes at each lineage decision are known in the literature to be functionally important genes, including classic hematopoietic regulators such as *Cebpa*, *Sfpi1*, *Gata1*, *Satb1*, *Irf8* and *Ebf1* (see full tables with references in ***Figure 3C*** and ***Figure 3—source data 1***). Additionally, the genes identified include many genes successfully used in hematopoietic reprogramming experiments, including *Gata2* and *Pbx1* (***Figure 3C***). Together these observations suggest that the sparse subspace of transition and marker genes identified by our framework not only allows for accurate reconstruction of the lineage hierarchy but also constitutes a set of candidates for relevant biological functions.

As further validation of the inference method, we compared it to the method proposed by Grun et al. on a single-cell intestinal development data set (***Grün et al., 2015***, ***2016***). We inferred lineages between each cell type based on their cluster identifications, excluding clusters with fewer than 10 cells, and constructed an undirected lineage tree by taking triplets with probability > 0.6 and applying the pruning rule (***Figure 3—figure supplement 2A***). The only disagreement between the two methods is the progression from crypt base columnar cells (C_2) to different populations of Goblet cells (C_4 and C_8). Grun et al. hypothesize a $C_2 – C_8 – C_4$ progression, while we infer the triplet $C_8 – C_2 – C_4$ with p>0.99, suggesting that the progenitor C_2 gives rise to both differentiated Goblet subpopulations. Both lineage trees are supported by the literature (***van der Flier and Clevers, 2009***). The high probability transition genes included many factors well known for their roles in tissue homeostasis and development (***Figure 3—figure supplement 2B***), notably *Klf4* (***Yu et al., 2012***), *Atoh1* (***VanDussen and Samuelson, 2010***), *Spdef* (***Noah et al., 2010***), *Foxa1/Foxa2* (***Ye and Kaestner, 2009***), and *Tcf3* (***Merrill et al., 2001***).

Inferred lineage tree for human excitatory neuronal progenitors from in vitro single-cell data over 80 days of differentiation

The ease with which single-cell transcriptomic data can be generated (***Grün et al., 2015***; ***Jaitin et al., 2014***; ***Macosko et al., 2015***; ***Patel et al., 2014***; ***Paul et al., 2015***; ***Treutlein et al., 2014***; ***Zeisel et al., 2015***) presents an opportunity to understand the dynamics of the underlying networks that lead individual cells to their final fate. We studied the differentiation of stem cells both into germ layer progenitors (***Jang et al., 2017***) and into cortical neurons. To study the latter, we analyzed single-cell gene expression data from 2217 cells from an in vitro differentiation protocol for early human neuronal development (***Yao et al., 2017***). Briefly, human embryonic stem cells (hESCs) were subjected to a SMAD inhibition-based cortical induction phase, a progenitor expansion phase, and a neural differentiation phase. Single cells were sorted at 12, 26, 54, and 80 days into differentiation, and their gene expression was profiled using the SMART-Seq2 technique (***Picelli et al., 2013***). In the initial clustering of the single-cell data, dimensionality reduction by PCA (into 15 above-noise components) followed by t-SNE (***Van Der Maaten and Hinton, 2008***; ***Satija et al., 2015***) showed separation by day and *SOX2* expression (***Figure 4A***). However, the number of predicted clusters varied depending on the perplexity parameter (***Figure 4—figure supplement 1A–B***). In addition, no clear lineage or distance relationship among the putative types is immediately apparent from this clustering. Analysis of this data with other recent methods such as Monocle and

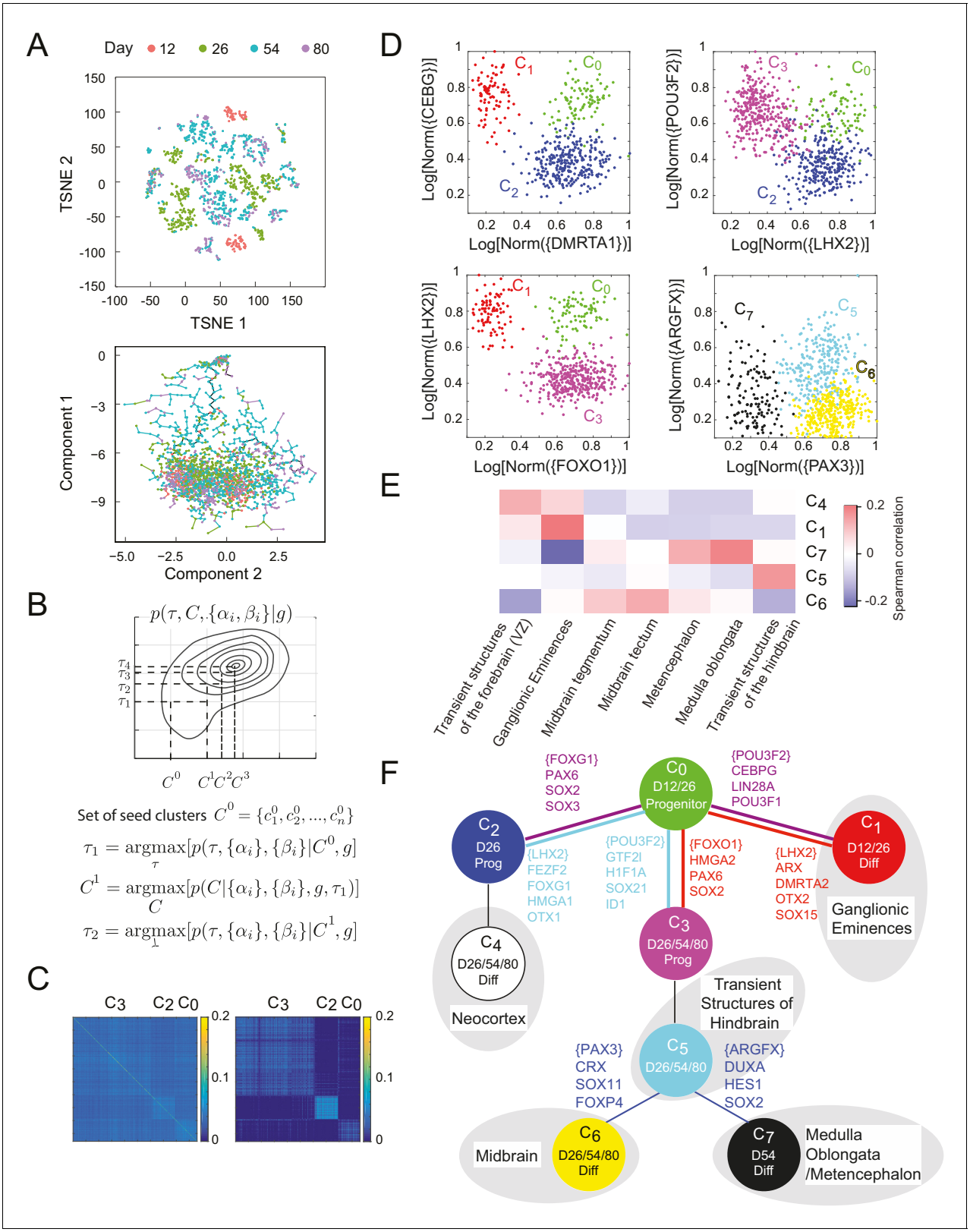


Figure 4. Inference of lineage tree and key transitions genes using single cell expression data from in vitro differentiated developing human brain. (**A**) RNA-seq data from single cells collected at days 12, 26, 54, and 80 from a human brain in vitro differentiation protocol (***Yao et al., 2017***) 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 ***Figure 4 – Figure Supplement 1A-B***) and does not allow for mechanistic understanding. Independent

Figure 4 continued on next page

Figure 4 continued

component analysis of all transcription factors with Monocle (bottom) does not show clear structure and could not inform reconstruction of lineage relationships. (**B**) Maximization algorithm to determine most likely cluster identities $\{C\} \equiv \{c_1, c_2, \ldots, c_n\}$, sets of transitions $\{T\}$, marker genes ($\alpha_i = 1$) and transition genes ($\beta_i = 1$), given single-cell gene expression data $\{g_i\}$. Starting from a seed clustering scheme $\{C_0\}$, iterative maximization of the conditional probabilities $p(\{T\}, \{\alpha_i\}, \{\beta_i\}|\{g_i\}, \{C\})$ and $p(\{C\}|\{g_i\}, \{T\}, \{\alpha_i\}, \{\beta_i\})$ converges to most likely set $(\{C\}, \{T\}, \{\alpha_i\}, \{\beta_i\})$ (**C**) 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). (**D**) Selected high probability triplets of clusters plotted in the axes defined by two sets of transition gene classes for each triplet. $c_1 - c_0 - c_2$ (top right, $p(T = c_0 \mid \{g_i^{c_0,c_1,c_2}\}) > 0.99$), plotted in transition gene class {*CEBPG*} also including *POU3F1, POU3F2, NR2F1, NR2F2, ARX, LIN28A, TOX3, ZBTB20, PROX1* and *SOX15*, and class {*DMRTA1*} also including *HES1, HES5, FOXG1, PAX6, HMGA2, SOX2, SOX3, SOX9, SOX6, SP8, OTX2, TGIF, ID4, TCF7L2*, and *TCFL1*. $c_2 - c_0 - c_3$ (top left, $p(T = c_0 \mid \{g_i^{c_0,c_2,c_3}\}) = 0.96$), plotted in transition gene class {*LHX2*} also including *FEZF2, FOXG1, HMGA1, SP8, OTX1, SOX11, GLI3, SIX3, ETV5*, and class {*POU3F2*} also including *GTF2I, HIF1A, ID1, ID3, PROX1, SALL1, SOX21, TCF12, TRPS1, ZHX2*.$c_1 - c_0 - c_3$ (bottom left, $p(T = c_0 \mid \{g_i^{c_0,c_1,c_3}\}) > 0.99$), plotted in transition gene class {*FOXO1*} also including *HMGA2, PAX6*, and *SOX2*, and class {*LHX2*} also including *DMRTA2, HMGA1, ARX, LIN28A, OTX2, LITAF, NANOG, POU3F1, SOX15*. $c_6 - c_5 - c_7$ (bottom right, $p(T = c_5 \mid \{g_i^{c_5,c_6,c_7}\}) > 0.99$), plotted in transition gene class {*PAX3*} also including *CRX, SOX11, EBF2, FOXP4, ASCL1, FOXO3*, and *SIX3*, and class {*ARGFX*} also including *DUXA, HES1, NFIB, PPARA, SOX2, SOX7*, and *SOX9*. (**E**) 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. (**F**) 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 data. Clusters color-coded similarly to (**D**). 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 (***Figure 4—figure supplement 1C–D***).

DOI: 10.7554/eLife.20488.019

The following source data and figure supplements are available for figure 4:

Source data 1. Final cluster identities of single cells from in vitro cortical differentiation.

DOI: 10.7554/eLife.20488.020

Source data 2. Probabilities of topologies for triplets of single-cell clusters.

DOI: 10.7554/eLife.20488.021

Source data 3. Probabilities of Transition and Marker Genes for the Human Brain Developmental Lineage Tree.

DOI: 10.7554/eLife.20488.022

Source data 4. Human Brain Development SmartSeq2 Census.

DOI: 10.7554/eLife.20488.023

Source data 5. List of Human Transcription Factors.

DOI: 10.7554/eLife.20488.024

Figure supplement 1. Cluster identity and sparse coding in neuronal differentiation.

DOI: 10.7554/eLife.20488.025

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***).

DOI: 10.7554/eLife.20488.026

Monocle2 (***Trapnell et al., 2014***), TSCAN (***Ji and Ji, 2016***), and StemID (***Grün et al., 2016***) did not clearly reconstruct lineage or infer key genes regulating transitions (***Figure 4A*** – Bottom, ***Figure 4—figure supplement 2***). Monocle2 (***Figure 4—figure supplement 2A***) produces a tree with complex branching, but *SOX2*+ progenitors and *DCX*+ differentiated neurons do not clearly separate.

Analyzing data from single-cell profiling presents an additional challenge relative to data from population-level profiling because cell types are not previously known and must be inferred from the data. Computationally, it is necessary define a measure of similarity in the gene expression profiles of individual cells so as to be able to cluster them and define cell states. Here again, it is necessary to identify the correct gene subspace to use for clustering. Clustering and determining lineage have typically been performed sequentially and treated as independent problems (***Satija et al., 2015***; ***Trapnell, 2015***; ***Trapnell et al., 2014***). However, we found that it is informative to solve both these problems simultaneously.

In our framework, the relevant feature set for clustering is the set of marker and transition transcription factors from the triplets with non-null topologies. But determination of these sets of genes and of the transition topologies depends itself on knowledge of the cluster identities. Following previous work on sparse clustering (***Witten and Tibshirani, 2010***), we simultaneously determined

optimal clusters, lineage topology and sets of transition and marker genes by iteratively selecting transition and marker genes and clustering the data using this set of features (**Figure 4B**).

In order to utilize information about developmental distances in clustering, we iteratively maximized a joint distribution, $P(T,\{C\}^m|\{g_i\})$, over the developmental tree, T, and the set of clusters of cells, $\{C\}^m=\{c_1^m,c_2^m,\ldots,c_K^m\}$ (**Figure 4B**). Starting with a clustering $\{C\}^m$, we first inferred the set of genes $\{\alpha_i=1\}$ and $\{\beta_i=1\}$ which were identified in high probability triplets, $p(T|\{g_i\},\ \{C\}^m)>0.6$, and we then re-clustered in this new subspace to obtain the clusters for the next iteration $\{C\}^{m+1}$.

Initial analysis based on the gap statistic (**Tibshirani et al., 2001**) suggested that the single-cell gene expression profiles clustered into 20 clusters of cell types. We chose a seed $\{C\}^0$ for the iterated clustering procedure by intentionally over-clustering the data into 40 clusters using spectral K-medoids. By being overly discriminative in our initial clustering, we ensured that all genes with differential regulation would be classified as either marker genes or transition genes, and would be preserved in later clustering iterations. We iterated the clustering-inference procedure until the dimension of the re-clustering subspace changed by less than 10% of the total transcription factor space. In this resulting subspace of 469 genes, we finally clustered the cells into the final configuration $\{C\}^f=\{c_0,\ c_1,\ldots,c_7\}$ using K-medoids (**Figure 4C**, **Figure 4—source data 1**) where the number of clusters K = 8 was chosen based on the gap statistic (**Tibshirani et al., 2001**).

We inferred 45 high probability triplets ($p(T\mid\{g_i\})>0.6$) between the final cell clusters (**Figure 4—source data 2**). Four such triplets are shown in **Figure 4D**, plotted in axes defined by transition genes for each triplet (**Figure 4—source data 3**). Starting with progenitor cell states (*SOX2+*, **Figure 4—figure supplement 1C**), we manually appended cell clusters to the tree according to their time information and in agreement with inferred topological restrictions. The first transition involves the production of day 12 neuronal cell type c_1 from day 12 progenitor c_0, which is observed in the triplets $c_1-c_0-c_2$ and $c_1-c_0-c_3$. The $c_1-c_0-c_2$ triplet ($p\left(T=c_0\mid\left\{g_i^{c_0,c_1,c_2,}\right\}\right)=\ 0.99$, **Figure 4D** – top left) is mediated by 47 transition genes between c_0 and c_1 and 87 between c_0 and c_2 ($p(\beta_i=1\mid\{g_i\},T)>0.8$). Transition genes expressed highly in the c_1-c_0 branch include *CEBPG, POU3F1, POU3F2, NR2F1, NR2F2, ARX, LIN28A, TOX3, ZBTB20, PROX1*, and *SOX15* which have been previously implicated in proliferation of forebrain progenitors (**Au et al., 2013**; **Borello et al., 2014**; **Cimadamore et al., 2013**; **Dominguez et al., 2013**; **Yang et al., 2015**) and the migratory behaviors of ganglionic eminences (**Kanatani et al., 2008**; **Kessaris et al., 2014**; **Lodato et al., 2014**; **Olivetti and Noebels, 2012**; **Reinchisi et al., 2012**). The c_0-c_2 transition genes include *DMRTA1, HES1, HES5, FOXG1, PAX6, HMGA2, SOX2, SOX3, SOX9, SOX6, SP8, OTX2, TGIF, ID4, SOX3, TCF7L2*, and *TCFL1* which are known to be expressed in forebrain progenitors of the developing neocortex, thalamus, and hypothalamus (**Abraham et al., 2013**; **Pozniak et al., 2010**; **Shimojo et al., 2011**; **Tzeng and de Vellis, 1998**; **Wang et al., 2006**), and are known to establish dorsal forebrain regional identity; (**Azim et al., 2009**; **Bani-Yaghoub et al., 2006**; **Borello et al., 2014**; **Gaston-Massuet et al., 2016**; **Hagey et al., 2014**; **Hutton and Pevny, 2011**; **Johansson et al., 2013**; **Kikkawa et al., 2013**; **Kishi et al., 2012**; **Manuel et al., 2011**; **Miyoshi and Fishell, 2012**; **Ohtsuka et al., 2001**; **Ross et al., 2003**; **Shen and Walsh, 2005**; **Sur and Rubenstein, 2005**; **Yang et al., 2015**; **Zembrzycki et al., 2007**).

The progenitor cell types form the triplet $c_2-c_0-c_3$ $\left(p\left(T=c_0|\left\{g_i^{c_0,c_2,c_3}\right\}\right)=\ 0.96\right.$, **Figure 4D** – top right). The c_2-c_0 branch is mediated by 39 transition genes including *LHX2, FEZF2, FOXG1, HMGA1, SP8, OTX1, SOX11, GLI3, SIX3*, and *ETV5* which suggest that c_2 is comprised of cortical progenitors (**Appolloni et al., 2008**; **Greig et al., 2013**; **Kishi et al., 2012**; **Manuel et al., 2011**; **Raciti et al., 2013**). The c_0-c_3 branch is mediated by 55 transition genes including *GTF2I, HIF1A, ID1, ID3, PROX1, POU3F2, SALL1, SOX21 TCF12, TRPS1* and *ZHX2*, which are associated with mesencephalon and metencephalon regional development, as well as having known involvement with midbrain/hindbrain organizer identity (**Buck et al., 2001**; **Enkhmandakh et al., 2009**; **Inoue et al., 2012**; **Jaegle et al., 2003**; **Kunath et al., 2002**; **Lavado and Oliver, 2007**; **Milosevic et al., 2007**; **Ohba et al., 2004**; **Uittenbogaard and Chiaramello, 2002**; **Yao et al., 2017**). We additionally inferred the triplet $c_2-c_0-c_3$ $\left(p\left(T=c_0\mid\left\{g_i^{c_0,c_1,c_3}\right\}\right)>\ 0.99\right.$, **Figure 4D** – bottom left), which suggests a three way split from early progenitor c_0 into early differentiated neuron c_1, and progenitors c_2 and c_3.

The continuation of the c_2 branch is inferred through triplet $c_0-c_2-c_4$ ($p\left(T=c_2\mid\left\{g_i^{c_0,c_2,c_4}\right\}\right)=\ 0.97$). The c_2-c_4 branch includes transition genes *BCL11A, EMX2, FOXP2*

and *RORB*, which are known to be associated with the neocortex and neuronal identity (**Cánovas et al., 2015**; **Ebisu et al., 2016**; **Greig et al., 2016**; **Jabaudon et al., 2012**; **Wiegreffe et al., 2015**; **Woodworth et al., 2016**; **Zembrzycki et al., 2007**).The triplet $c_0 - c_3 - c_5$ ($p(T = c_3 \mid \{g_i^{c_0,c_3,c_5}\}) > 0.99$) meanwhile is characterized by transition genes in the $c_3 - c_5$ branch including *ASCL1, FOXP2, PAX3, POU3F4, ZIC1, ZIC4, HOXB2*, and *EN2*, which have been shown to regulate fate acquisition in the midbrain/hindbrain (**Agoston et al., 2012**; **Ang, 2006**; **Di Bonito et al., 2013**; **Elsen et al., 2008**; **Hegarty et al., 2013**; **Miller et al., 2011**; **Tan et al., 2014**).

The c_5 cell cluster differentiates into two distinct clusters of post-mitotic neurons – c_6 and c_7 ($p(T = c_5 \mid \{g_i^{c_5,c_6,c_7}\}) > 0.99$, **Figure 4D** – bottom right). The $c_5 - c_6$ branch is inferred from transition genes including *PAX3, CRX, SOX11, EBF2, FOXP4, ASCL1, FOXO3*, and *SIX3* which have strong expression in developing dopaminergic and gabaergic neurons of the midbrain (**Agoston et al., 2012**; **Erickson et al., 2010**; **Pino et al., 2014**; **Yang et al., 2015**; **Yin et al., 2009**; **Zhang et al., 2002**). Transition genes in the $c_5 - c_7$ branch include *ARGFX, DUXA, HES1, NFIB, PPARA, SOX2, SOX7*, and *SOX9*, which are known to be associated with the medulla oblongata region of the hindbrain (**Fawcett and Klymkowsky, 2004**; **Kameda et al., 2011**; **Kumbasar et al., 2009**; **Madissoon et al., 2016**; **Matsui et al., 2000**; **Stolt et al., 2003**).

In addition to interpreting these individual transition genes defining the major branch splits, we correlated the expression over all predicted transition and marker genes in neuronal clusters to in vivo developmental human data (**Miller et al., 2014**). The in vivo data comprise a representative range of microarray data sampled from different parts of the developing brain at post-conception week 15, including forebrain proliferative regions, midbrain, and hindbrain. The differences in data acquisition methods (RNAseq vs. microarray, single-cell vs heterogeneous populations) resulted in relatively low correlations overall, but there are clear associations between individual clusters and specific brain regions (**Figure 4E**). Specifically, c_1, maps to the ganglionic eminences, suggesting an interneuron identity, whereas c_5, c_6 and c_7 show better mapping to mid- and hindbrain structures, and c_4 appears to be more closely related to neocortex. Overall, this global comparison, combined with the identification of genes with known regional expression, suggests that the inferred clusters from the in vitro data capture the diversity of differentiation into the early stages of the major neuronal lineages (**Figure 4F**). These lineage predictions based on our analysis techniques were verified experimentally using viral barcoding in a separate work (**Yao et al., 2017**).

To estimate the sparsity of the underlying network and to find a minimal subset of genes through which lineage could be inferred, we replicated the analysis while only considering a limited set of genes per triplet. We assembled a collection of 20 triplets with maximal leaf-to-leaf distance of 4 nodes, and non-null inferred topology. For each triplet, we ranked genes based on their odds of being a transition gene, $\mathcal{O}_i$, agnostic of the true topology of the triplet. We then replicated the inference process using only the N genes with the greatest odds (**Figure 4—figure supplement 1E**). We found that with as few as 4 genes per triplet, the correct lineage topology could be inferred for all of the triplets. Genes with greatest odds comprise a restricted subset of genes for further experimental investigation. Further, these findings suggest that the dynamics of expression of just four specific genes are sufficient to monitor a particular lineage decision in single cells.

Discussion

Finding an informative subspace of variables for data analysis is a general problem in machine learning, both for regression and clustering (**Tibshirani, 1996**; **Witten and Tibshirani, 2010**); the innovation in this paper is to use a statistical pattern learned from known biology to inform this subspace search. The approach we take here is complementary to methods that project expression variability onto coordinates of PCA, ICA or t-SNE maps (**Marco et al., 2014**; **Satija et al., 2015**; **Trapnell et al., 2014**), which are combinations of all variables. Searching for sparse representation of the dynamics has the advantage of providing interpretability and experimental direction (**McGibbon and Pande, 2017**). Following the dynamics of this small set of high-probability transition genes via fluorescent tagging could allow for the tracking of lineage decisions of individual cells in real time. Further, these genes provide a list of candidates for drivers of fate decisions, and hence a set of experimental hypotheses.

Not all genes give us equal information about the dynamics of differentiation. We discovered that genes showing the clear minimum pattern are most predictive of the sequence of lineage transitions

during development. Although our pattern discovery and subsequent lineage reconstruction does not assume any functional role for the clear minimum pattern, we note that this pattern is shown by genes known to be regulators of development during hematopoiesis. The same pattern observed in many differentiating systems (**Graf and Enver, 2009**; **Qi et al., 2013**; **Thomson et al., 2011**; **Zhang et al., 1999**), and is consistent with mutual inhibition. Mutual inhibition, in turn, is hypothesized to play an important role in maintaining multi-stable systems and in mediating transitions between different stable states of multi-stable systems (**Ferrell, 2012**).

Discovery of sparse representations of the cell states and variability between them demonstrates the efficacy of low dimensional descriptions of the system. Understanding the dynamics and transitions of complex physical systems composed of a large number of variables has been driven by the discovery of low dimensional order parameters (**Anderson, 1978**; **Landau and Lifshitz, 1951**). As opposed to measuring and modeling states as high dimensional objects in their native representation, order parameters provide low dimensional descriptions of the states and dynamics, which has proven crucial in developing both qualitative and quantitative models. Finding small subsets of genes which captures the lineage transitions in cells analogously provides a low dimensional subspace that captures the dynamics in genetic networks and can be useful for modeling (**Jang et al., 2017**). An accompanying paper allows us to exploit this idea to extract mathematical models for the underlying molecular circuits from single cell gene expression data obtained during germ layer differentiation (**Jang et al., 2017**).

Materials and methods

In vitro neuronal differentiation

Single-cell transcriptomic data from the in vitro neural differentiation procedure was obtained as described in **Yao et al. (2017)** (Supplemental information):

hESCs were dissociated with Accutase and plated on Matrigel-coated 24-well plates at 2.5 × 105 cells/cm2 in DMEM/F12 (#11330–032), 1 × N2, 1 × B27 without vitamin A, 2 mM Glutamax, 100 µM non-essential amino acids, 0.5 mg/mL BSA, 1X Pen-Strep, and 100 µM 2-mercaptoethanol (referred to as basal medium; all from Thermo Fisher, Waltham, MA) with 20 ng/mL FGF2 (Thermo Fisher) and 2 µM thiazovivin. Cortical induction was initiated by changing to the basal medium with 5 µM SB431542 (StemRD, Burlingame, CA), 50 nM LDN193189 (Reagents Direct, Encinitas, CA) and 1 µM cyclopamine (Stemgent, Lexington, MA) (referred to as NIM). NIM was changed daily for 11 days. On day 12, cells were dissociated and seeded on Matrigelcoated 24-well plates at 5 × 105/cm2 in basal medium with 20 ng/mL FGF2 and 2 µM thiazovivin. Progenitor expansion was initiated on D13 by changing to serum-free human neural stem cell culture medium (NSCM, #A10509–01 from Thermo Fisher) containing 20 ng/mL FGF2 and 20 ng/mL EGF. NSCM was changed daily for 6 days. Cultures were passaged once more on D19 with Accutase and replated at 5 × 105 cells/cm2. On D26, cells were dissociated with Accutase and seeded on 24-well plates sequentially coated with poly-D-lysine (Millipore, Billerica, MA) and laminin (Thermo Fisher) at 1 × 105 cells/cm2 in basal medium supplemented with 20 ng/mL FGF2 and 2 µM thiazovivin. On D27, medium was changed to a 1:1 mixture of DMEM/F12 and Neurobasal medium (#21103–049) supplemented with 100 µM cAMP (Sigma-Aldrich, St Louis, MO), 10 ng/mL BDNF (R and D Systems, Minneapolis, MN), 10 ng/mL GDNF (R and D Systems) and 10 ng/mL NT-3 (R and D Systems) (referred to as ND). Cells were maintained in ND medium for four weeks until day 54 with half medium change every other day. Quality of differentiations was routinely assessed by immunostaining at D12 (*PAX6* and *DCX*), at D26 (*LHX2, SOX2, EOMES, POU3F2*, and *TBR1*), and at D54 (*MAP2* costained with *TBR1, CTIP2, SATB2*). In addition flow cytometry at D26 (*EOMES*, *SOX2* and *PAX6*) was performed. Typically, *EOMES* at day 26 proved the most valuable quality control metric (~10% of cells by both flow cytometry and immunostaining) and predicted failure at D54. Specifically, when *EOMES* was low (<1% of cells) differentiations failed and were typically dominated by *POU3F2*+ cell types and/or non-neural 'other' cell types. These failed differentiations were eliminated from further analysis, typically ~20% of experiments (5 of 19 experiments in 2016). 50 bp paired-end Smart-Seq2 libraries were prepared from these cells as previously described (**Picelli et al., 2013**) and mapped as described in **Thomsen et al., 2016** and **Yao et al. (2017)**. As each cell was profiled independently (without pooling before amplification, as in methods such as Cel-Seq, STRT, or Drop-Seq), we did not observe the

batch effects present in pooling-based methods. Although cells were profiled in plates (batches), there was no significant plate-related variation - this Fixed single-cell transcriptomic characteriza is most likely due to amplification being carried out separately in each well of the plate, thereby precluding cross-talk among barcodes, or plate-related differences in amplification. This is clear from the mixing of cells from different plates in the clustering. A complete census is provided (***Figure 4—source data 4***).

Gene expression data

Hematopoietic gene expression data were downloaded from the Immunological Genome Project (***Heng et al., 2008***); GEO GSE15907) and log-2 transformed. We restricted the genes considered to 1459 transcription factors. Brain development expression data were log-2 transformed, and 1460 transcription factors were individually normalized by dividing by the 90th-percentile expression value. Lists of mouse and human transcription factors are provided (***Figure 1—source data 2***, ***Figure 4—source data 5***).

Software

Calculations were performed using custom written MATLAB code (The Mathworks) on the Harvard Research Computing Odyssey cluster. Code is available at https://github.com/furchtgott/sibilant. t-SNE was done using the package provided in (***Van Der Maaten, 2009***). Monocole was run in R (***Core Team, 2015***; ***Trapnell et al., 2014***).

Description of algorithm

The algorithm proceeds according to the following steps:

1. Find initial seed clustering configuration $\{C\}^0$ using K-medoids where K is chosen to be larger than the cluster number inferred from the Gap statistic (***Tibshirani et al., 2001***)
2. For all triplets of clusters, find most likely T and $\{\alpha_i\}$ and $\{\beta_i\}$ given $\{C\}^m$:
   a. Compute $p\left(g_i^{A,B,C} \mid T, \alpha_i, \beta_i, \{C\}^m\right)$ by integrating numerically over $p(\mu,\ \sigma)$ (***Equations 6, 9, and 10***).
   b. Compute $p\left(T, \{\alpha_i\}, \{\beta_i\} \mid \left\{g_i^{A,B,C}\right\}, \{C\}^m\right)$ using ***Equations 20 and 29***.
   c. Identify mostly likely topology T and set of $\{\alpha_i\}$ and $\{\beta_i\}$.
3. Recluster $\{g\}$ using K-medoids in the space of all $\{\alpha_i\}$ and $\{\beta_i\}$ for the triplets with probability $p\left(T \mid \left\{g_i^{A,B,C}\right\}, \{C\}^m\right) > 0.6$ of being non-null. Determine new clustering configuration $\{C\}^{m+1}$.
4. Repeat steps 1 and 2 until convergence of $\{C\}$.
5. Determine the most likely tree connecting cell clusters, recapitulating high-probability triplet topologies.

Each of these steps is described in the following section.

Bayesian framework for inferring cluster identities, state transitions, and marker and transition genes simultaneously

Notation; Bayes' rule

Given gene expression data from single cells $\{g_i\}$, we built a probabilistic framework to simultaneously infer cell cluster identities, $\{C\} \equiv \{c_A, c_B, \ldots\}$, the sequence of transitions T between these clusters, the key sets of marker genes $\{\alpha_i\}$ that define each cell cluster, and genes $\{\beta_i\}$ that determine the sequence of transitions between clusters. We maximized the joint probability distribution of these variables given the gene expression data, $p(T, \{c_A, c_B, \ldots\}, \{\alpha_i\}, \{\beta_i\} \mid \{g_i\})$ to determine the maximum likelihood estimates of these parameters.

We first consider how to solve this problem in the case in which there are three cell clusters, and we will later build a tree using all possible combinations of three cell clusters. Let the set of three cell clusters be c_A, c_B and c_C with gene expression data $\left\{g_i^{A,B,C}\right\}$ for all genes $(i = 1\ to\ N)$ and all cells. The term $g_i^{A,B,C}$ denotes the expression data for just gene i in cells in clusters c_A, c_B, and c_C. The topology T of the relationships between cell clusters c_A, c_B and c_C can take on four possible values: $T = \mathcal{A}$: cell cluster c_A is in the middle (either c_A is the progenitor of c_B and c_C, or c_A is an intermediate cell type between c_B and c_C); $T = \mathcal{B}$: cell cluster c_B is in the middle; $T = \mathcal{C}$: cell cluster c_C is

in the middle; or $T = \emptyset$: we cannot determine the topology. Complementarily, for each gene i we define variables α_i and β_i, where $\alpha_i = 1$ and $\beta_i = 0$ if the gene is a marker gene, $\alpha_i = 0$ and $\beta_i = 1$ if the gene is a transition gene, and $\alpha_i = \beta_i = 0$ otherwise. Our task is to determine the probability $p\left(T, \{c_A, c_B, \ldots\}, \{\alpha_i\}, \{\beta_i\} \mid \left\{g_i^{A,B,C}\right\}\right)$ given gene expression data for all genes $\left\{g_i^{A,B,C}\right\}$.

According to Bayes' rule, $p\left(T, \{c_A, c_B, \ldots\}, \{\alpha_i\}, \{\beta_i\} \mid \left\{g_i^{A,B,C}\right\}\right)$ is proportional to the probability of the gene expression data given T, $\{C\}$, $\{\alpha_i\}$ and $\{\beta_i\}$:

$$p\left(T, \{c_A, c_B, \ldots\}, \{\alpha_i\}, \{\beta_i\} \mid \left\{g_i^{A,B,C}\right\}\right) = \frac{p\left(\left\{g_i^{A,B,C}\right\} \mid T, \{C\}, \{\alpha_i\}, \{\beta_i\}\right) p(\{\alpha_i\}, \{\beta_i\} \mid T, \{C\})\, p(T|\{C\})\, p(\{C\})}{p\left(\left\{g_i^{A,B,C}\right\}\right)} \quad (2)$$

The denominator of the right hand side of ***Equation 2*** is a normalization constant. Expressions $p(\{\alpha_i\}, \{\beta_i\} \mid T, \{C\})$, $p(T|\{C\})$, and $p(\{C\})$ are respectively the prior probabilities of $\{\alpha_i\}$ and $\{\beta_i\}$ given T and $\{C\}$, the prior probability of T given $\{C\}$, and the prior probability of $\{C\}$. We assume that in the absence of any expression data, the probability that a gene is a transition or marker gene is independent of the topology and clustering configuration: $p(\{\alpha_i\}, \{\beta_i\} \mid T, \{C\})\, p(T|\{C\})\, p(\{C\}) = p(\{C\}) p(T) \prod_i p(\alpha_i, \beta_i)$, and $p(T) = 1/4$.

Conditional independence

In our model, we assume that knowing the clustering configuration $\{C\}$, the topology T and whether or not a gene is a marker or transition gene is sufficient to determine the probability distribution for its expression levels in each of the cell clusters. Therefore, the gene expression patterns of different genes are conditionally independent given the topology, clustering and gene type:

$$p\left(\left\{g_i^{A,B,C}\right\} \mid T, \{C\}, \{\alpha_i\}, \{\beta_i\}\right) = \prod_i p\left(g_i^{A,B,C} \mid T, \{C\}, \alpha_i, \beta_i\right) \quad (3)$$

Thus, ***Equation 2*** becomes

$$p\left(T, \{c_A, c_B, \ldots\}, \{\alpha_i\}, \{\beta_i\} \mid \left\{g_i^{A,B,C}\right\}\right) = \frac{p(\{C\}) p(T) \prod_i p\left(g_i^{A,B,C} \mid T, \{C\}, \alpha_i, \beta_i\right) p(\alpha_i, \beta_i)}{p\left(\left\{g_i^{A,B,C}\right\}\right)} \quad (4)$$

We maximize the evaluated $p\left(T, \{c_A, c_B, \ldots\}, \{\alpha_i\}, \{\beta_i\} \mid \left\{g_i^{A,B,C}\right\}\right)$ with respect to T, $\{C\}$ and each of the α_i and β_i to obtain the most likely relationships between cell types c_A, c_B and c_C, as well as the genes most likely to be marker and transition genes.

Expression for $p\left(g_i^{A,B,C} \mid T, \{C\}, \alpha_i = 0,\ \beta_i = 1\right)$ (transition genes)

To infer $p\left(T, \{c_A, c_B, \ldots\}, \{\alpha_i\}, \{\beta_i\} \mid \left\{g_i^{A,B,C}\right\}\right)$, we need a model to compute the probability of the gene expression data for each gene, $p\left(g_i^{A,B,C} \mid T, \{C\}, \alpha_i, \beta_i\right)$, given T, $\{C\}$,α_i and β_i, following ***Equation 4***. Our model for the probability distribution of the expression of a single transition gene i in the three cell types $p\left(g_i^{A,B,C} \mid T, \{C\}, \alpha_i = 0,\ \beta_i = 1\right)$ is defined solely by the geometry of the arrangement of the cell types in gene expression space, as described in the main text. For example, for $T = \mathcal{A}$ and $\beta_i = 1$, our model is that the distribution of the expression levels of $g_i^{A,B,C}$ in the three cell types A, B and C has the smallest mean value in either B or C but not in A (***Figure 2A***). If the distribution of the expression of gene i in cell type A is $D_A\left(g_i^A \mid \mu_A^i, \sigma_A^i, \{C\}\right)$ (we assume a log-normal distribution) with a mean μ_A^i and standard deviation σ_A^i, with analogous expressions for cell types B and C, then our model defining $p\left(g_i^{A,B,C} \mid T = \mathcal{A}, \beta_i = 1, \{C\}\right)$ is that either $\mu_B^i < \mu_C^i$ and $\mu_B^i < \mu_A^i$ or $\mu_C^i < \mu_B^i$ and $\mu_C^i < \mu_A^i$, where μ_A^i, μ_B^i and μ_C^i are the mean values of the expression levels of g_i in cell types A, B and C. Thus,

$$p\left(g_i^{A,B,C} \mid T=\mathcal{A},\beta_i=1,\{C\}\right)=\frac{1}{2}\left\{\begin{array}{l} p\left(g_i^{A,B,C} \mid \mu_B^i<\mu_A^i,\mu_B^i<\mu_C^i,\{C\}\right) \\ +p\left(g_i^{A,B,C} \mid \mu_C^i<\mu_A^i,\mu_C^i<\mu_B^i,\{C\}\right)\end{array}\right\} \tag{5}$$

The terms in ***Equation 5*** can be calculated by integrating over the prior probability distribution of the means μ_A^i, μ_B^i and μ_C^i and standard deviations σ_A^i, σ_B^i and σ_C^i, with the conditions on the means constraining the domains of integration:

$$\begin{aligned} &p(g_i^{A,B,C}|T=\mathcal{A},\beta_i=1,\{C\}) \\ &=\tfrac{1}{2}\iiint_{\mu_B^i<\mu_A^i,\mu_B^i<\mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i} D_A(g_i^A|\mu_A^i,\sigma_A^i)D_B(g_i^B|\mu_B^i,\sigma_B^i)D_C(g_i^C|\mu_C^i,\sigma_C^i)p(\mu_A^i,\mu_B^i,\mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i) \\ &+\tfrac{1}{2}\iiint_{\mu_C^i<\mu_A^i,\mu_C^i<\mu_B^i,\sigma_A^i,\sigma_B^i,\sigma_C^i} D_A(g_i^A|\mu_A^i,\sigma_A^i)D_B(g_i^B|\mu_B^i,\sigma_B^i)D_C(g_i^C|\mu_C^i,\sigma_C^i)p(\mu_A^i,\mu_B^i,\mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i) \end{aligned} \tag{6}$$

Probabilities $p\left(g_i^{A,B,C} \mid T=\mathcal{B},\beta_i=1,\{C\}\right)$ and $p\left(g_i^{A,B,C} \mid T=\mathcal{C},\beta_i=1,\{C\}\right)$ are defined similarly.

In addition to topologies $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, we consider a null hypothesis $\emptyset$ in which transition genes have differential expression levels between states, but these levels are not correlated with any particular topology of states. This corresponds to having gene expression levels from cell-types A, B and C coming from three distributions with no restrictions on the relative order of the three means:

$$\begin{aligned} &p\left(g_i^{A,B,C} \mid T=\emptyset,\beta_i=1,\{C\}\right) \\ &=\iiint_{\mu_A^i,\mu_B^i,\ \mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i} D_A\left(g_i^A \mid \mu_A^i,\sigma_A^i\right)D_B\left(g_i^B \mid \mu_B^i,\sigma_B^i\right)D_C\left(g_i^C \mid \mu_C^i,\sigma_C^i\right)p\left(\mu_A^i,\mu_B^i,\mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i\right) \end{aligned} \tag{7}$$

Note that the probability of the data given the null hypothesis is the average of the probabilities of the data given the non-null hypotheses:

$$p(g_i^{A,B,C}|T=\emptyset,\beta_i=1,\{C\})=\frac{1}{3}\sum_{T=A,B,C} p(g_i^{A,B,C}|T,\beta_i=1,\{C\}) \tag{8}$$

Note that $p(g_i^{A,B,C}|T,\{C\},\alpha_i=0,\beta_i=1)$ depends on both T and $\{C\}$.

Expression for $p\left(g_i^{A,B,C} \mid T,\{C\},\alpha_i=1,\beta_i=0\right)$ (marker genes)

Our model for marker genes assumes that the probability distribution for the expression level of such genes, $p\left(g_i^{A,B,C} \mid T,\{C\},\alpha_i=1,\beta_i=0\right)$ to be independent of T and to be generated from distributions with two cell-types having a low value and the third a high value (for example, $D_{AB}\left(g_i^{A,B,C} \mid \mu_{AB}^i,\sigma_{AB}^i\right)$ for cell-types A and B and $D_C\left(g_i^C \mid \mu_C^i,\sigma_c^i\right)$ for cell-type C, with the constraint $\mu_{AB}^i<\mu_C^i$):

$$\begin{aligned} &p(g_i^{A,B,C}|T,\{C\},\alpha_i=1,\beta_i=0) \\ &=\frac{1}{3}\iiint_{\mu_{AB}^i<\mu_C^i,\sigma_{AB}^i,\sigma_C^i} D_{AB}(g_i^{AB}|\mu_{AB}^i,\sigma_{AB}^i)D_C(g_i^C|\mu_C^i,\sigma_C^i)p(\mu_{AB}^i,\mu_C^i,\sigma_{AB}^i,\sigma_C^i) \\ &+\frac{1}{3}\iiint_{\mu_{AC}^i<\mu_B^i,\sigma_{AC}^i,\sigma_B^i} D_{AC}(g_i^{AC}|\mu_{AC}^i,\sigma_{AC}^i)D_B(g_i^B|\mu_B^i,\sigma_B^i)p(\mu_{AC}^i,\mu_B^i,\sigma_{AC}^i,\sigma_B^i) \\ &+\frac{1}{3}\iiint_{\mu_{BC}^i<\mu_A^i,\sigma_{BC}^i,\sigma_A^i} D_{BC}(g_i^{BC}|\mu_{BC}^i,\sigma_{BC}^i)D_B(g_i^A|\mu_A^i,\sigma_A^i)p(\mu_{BC}^i,\mu_A^i,\sigma_{BC}^i,\sigma_A^i) \end{aligned} \tag{9}$$

Note that $p\left(g_i^{A,B,C} \mid T,\{C\},\ \alpha_i=1,\beta_i=0\right) = p\left(g_i^{A,B,C} \mid \{C\},\ \alpha_i=1,\beta_i=0\right)$ does not depend on T but does depend on $\{C\}$.

Expression for $p\left(g_i^{A,B,C} \mid T,\{C\},\ \alpha_i=0,\beta_i=0\right)$ (irrelevant genes)

Our model for genes that are neither marker nor transition genes is that the expression levels of such genes, $p\left(g_i^{A,B,C} \mid T,\{C\},\ \alpha_i=0,\beta_i=0\right)$, is generated from one single distribution $D_{ABC}\left(g_i^{A,B,C} \mid \mu^i,\sigma^i\right)$:

$$p\left(g_i^{A,B,C} \mid T,\{C\},\alpha_i=0,\beta_i=0\right) = \iint\limits_{\mu^i,\sigma^i} D_{ABC}\left(g_i^{A,B,C} \mid \mu^i,\sigma^i\right)p(\mu^i,\sigma^i) \tag{10}$$

Note that $p\left(g_i^{A,B,C} \mid T,\{C\},\ \alpha_i=0,\beta_i=0\right) = p\left(g_i^{A,B,C} \mid \ \alpha_i=0,\beta_i=0\right)$ does not depend on T or $\{C\}$.

Numerical integration

Each of the probabilities on the right hand side of ***Equation 4*** is evaluated numerically as above. We assume the distribution of the expression of gene i in cluster c_A $D_A\left(g_i^A \mid \mu_A^i,\sigma_A^i\right)$ to be log-normal. Given m log2-transformed replicate measurements g_i^A of gene expression of gene i in cells belonging to cluster c_A, the probability of the data assuming mean μ_A^i and standard deviation σ_A^i is:

$$D_A(g_i^A|\mu_A^i,\sigma_A^i) = \left(\frac{1}{\sqrt{2\pi\sigma_A^{i2}}}\right)^m \prod_{g_i^A}^{m} e^{-\frac{(g_i^A-\mu_A^i)^2}{2\sigma_A^{i2}}}. \tag{11}$$

Distributions D_B, D_C, D_{AB}, D_{AC}, D_{BC} and D_{ABC} are defined analogously.

We take the *a priori* probability distribution of μ^i and σ^i, $p(\mu^i,\sigma^i)$ as uniform over a certain range of means and standard deviations estimated from the data. For the log2-transformed hematopoietic gene expression data, we take $2<\mu^i<14$ and $0<\sigma^i<0.75$.

We take the prior probabilities for the distributions in different cell types to be independent: $p\left(\mu_A^i,\mu_B^i,\mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i\right) = p\left(\mu_A^i,\sigma_A^i\right)\,p\left(\mu_B^i,\sigma_B^i\right)\,p\left(\mu_C^i,\sigma_C^i\right)$. The constraints on the order of the means are enforced by the domain of integration, and the prior must be properly normalized over this domain. For example, in ***Equation 6***,

$$\begin{aligned}\frac{1}{2}\iiint\limits_{\mu_B^i<\mu_A^i,\mu_B^i<\mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i} p\left(\mu_A^i,\mu_B^i,\mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i\right)\\ +\frac{1}{2}\iiint\limits_{\mu_C^i<\mu_A^i,\mu_C^i<\mu_B^i,\sigma_A^i,\sigma_B^i,\sigma_C^i} p\left(\mu_A^i,\mu_B^i,\mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i\right) = 1.\end{aligned} \tag{12}$$

Integrals were evaluated numerically in MATLAB using trapezoidal integration with step-sizes $\delta\mu$ = 0.05 and $\delta\sigma$ = 0.01.

The choice of a log-normal distribution could potentially confound the results, particularly for single-cell RNA-Seq data with significant zero-inflation. This is a potential area for improvement in our algorithm, but the method can easily be adapted to model different distributions of RNA expression, such as gamma distributions (***Shahrezaei and Swain, 2008***; ***Wills et al., 2013***) or beta-Poisson distributions (***Delmans and Hemberg, 2016***; ***Vu et al., 2016***). In either case, the probability of the data given different topologies would be computed by numeric integration over the parameters of the distribution, for example, α and β for the Gamma distribution, by replacing the log-normal distributions in ***Equations 7, 9 and 10*** with ones from the appropriate model.

The choice of the right parametric form for single-cell RNA expression is still an area of active research. Our choice of log-normal distributions assumes that higher order moments of the

distributions (beyond standard deviation) ought to have a minimal contribution to the predictions, but we have not tested this extensively.

Although the default prior was the uniform prior, we also implemented an empirical prior $p(\mu,\ \sigma)$ by estimating the empirical distribution over all the Immgen cell types, using the kernel density estimation code provided in (**Botev et al., 2010**). The resulting hematopoietic lineage tree was identical. Using the kernel-density-estimated empirical prior may provide more stability in future analyses.

Probability of topology given gene expression and cluster identities $p\left(T \mid \left\{g_i^{A,B,C}\right\},\{C\}\right)$

We can derive the probability of the topology given the gene expression data and cluster identities $p\left(T \mid \left\{g_i^{A,B,C}\right\},\{C\}\right)$ by summing over all the $\{\alpha_i\}$ and $\{\beta_i\}$ to find the probability of the data given topology $p\left(\left\{g_i^{A,B,C}\right\} \mid T,\{C\}\right)$:

$$\begin{aligned} p\left(\left\{g_i^{A,B,C}\right\} \mid T,\{C\}\right) &= \sum_{\alpha_i,\beta_i} p\left(\left\{g_i^{A,B,C}\right\} \mid T,\{\alpha_i\},\{\beta_i\},\{C\}\right) p(\{\alpha_i\},\{\beta_i\} \mid T,\{C\}) \\ &= \sum_{\beta_1}\sum_{\beta_2}\cdots\sum_{\beta_N}\prod_i p\left(g_i^{A,B,C} \mid T,\alpha_i,\ \beta_i,\{C\}\right) p(\alpha_i,\ \beta_i) \\ &= \prod_i\left(\sum_{\beta_i} p\left(g_i^{A,B,C} \mid T,\alpha_i,\ \beta_i,\{C\}\right) p(\alpha_i,\beta_i)\right) \end{aligned} \tag{13}$$

$$p\left(\left\{g_i^{A,B,C}\right\} \mid T,\{C\}\right) = \prod_i p\left(g_i^{A,B,C} \mid T,\{C\}\right), \tag{14}$$

where the probability of gene expression data for gene i given topology $p\left(g_i^{A,B,C} \mid T,\{C\}\right)$ is obtained by summing $p\left(g_i^{A,B,C} \mid T,\{C\},\ \alpha_i,\beta_i\right)$ over α_i and β_i:

$$\begin{aligned} p\left(g_i^{A,B,C} \mid T,\{C\}\right) = \ & p\left(g_i^{A,B,C} \mid T,\{C\},\alpha_i=0,\beta_i=1\right)p(\beta_i=1) \\ & +p\left(g_i^{A,B,C} \mid \{C\},\alpha_i=1,\beta_i=0\right)p(\alpha_i=1,\beta_i=0) \\ & +p\left(g_i^{A,B,C} \mid \alpha_i=0,\beta_i=0\right)p(\alpha_i=0,\beta_i=0). \end{aligned} \tag{15}$$

We consider the odds of a gene being a transition gene $p(\beta_i=1)/p(\beta_i=0)$ as a sparsity parameter, which we vary. We take $p(\alpha_i=0 \mid \beta_i=0) = p(\alpha_i=1 \mid \beta_i=0) = 1/2$.

The probability of topology T given data is proportional to the probability of the data given topology T (using Bayes' rule):

$$p\left(T \mid \left\{g_i^{A,B,C}\right\},\{C\}\right) = \frac{p\left(\left\{g_i^{A,B,C}\right\} \mid T,\{C\}\right)p(T)}{p\left(\left\{g_i^{A,B,C}\right\} \mid \{C\}\right)}, \tag{16}$$

where

$$p\left(\left\{g_i^{A,B,C}\right\} \mid \{C\}\right) = \sum_T p(T)p\left(\left\{g_i^{A,B,C}\right\} \mid T,\{C\}\right). \tag{17}$$

Therefore, using ***Equation 14***, we obtain the following expression for $p\left(T \mid \left\{g_i^{A,B,C}\right\},\{C\}\right)$:

$$p\left(T \mid \left\{g_i^{A,B,C}\right\},\{C\}\right) = \frac{p(T)\prod_i p\left(g_i^{A,B,C} \mid T,\{C\}\right)}{\sum_T p(T)\prod_i p\left(g_i^{A,B,C} \mid T,\{C\}\right)}. \tag{18}$$

Equation 18 can be written more explicitly by rewriting ***Equation 15*** as follows:

$$p\left(g_i^{A,B,C} \mid T,\{C\}\right) = p\left(g_i^{A,B,C} \mid T,\beta_i=0,\{C\}\right)p(\beta_i=0)+p\left(g_i^{A,B,C} \mid T,\beta_i=1,\{C\}\right)p(\beta_i=1)$$
$$= p\left(g_i^{A,B,C} \mid \beta_i=0,\{C\}\right)\,p(\beta_i=0)\left(1+\frac{p\left(g_i^{A,B,C} \mid T,\beta_i=1,\{C\}\right)p(\beta_i=1)}{p\left(g_i^{A,B,C} \mid \beta_i=0,\{C\}\right)p(\beta_i=0)}\right). \quad (19)$$

Here we have used the fact noted earlier that in our generating model, $p\left(g_i^{A,B,C} \mid T,\alpha_i=1,\ \beta_i=0,\ \{C\}\right)=p\left(g_i^{A,B,C} \mid \alpha_i=1,\ \beta_i=0,\{C\}\right)$ and $p\left(g_i^{A,B,C} \mid T,\alpha_i=0,\ \beta_i=0,\ \{C\}\right)=p\left(g_i^{A,B,C} \mid \alpha_i=0,\ \beta_i=0\right)$ do not depend on T (***Equation 10***). The terms $\prod_i p\left(g_i^{A,B,C} \mid \beta_i=0,\{C\}\right)\,p(\beta_i=0)$ cancel out in the numerator and denominator of ***Equation 18***, and we can write ***Equation 18*** in terms of ratios of the probabilities of the data given transition-gene and non-transition-gene status:

$$p\left(T \mid \left\{g_i^{A,B,C}\right\},\{C\}\right)=\frac{p(T)\prod_i\left(1+\frac{p\left(g_i^{A,B,C} \mid T,\beta_i=1,\{C\}\right)p(\beta_i=1)}{p\left(g_i^{A,B,C} \mid \beta_i=0,\{C\}\right)p(\beta_i=0)}\right)}{\sum_T p(T)\prod_i\left(1+\frac{p\left(g_i^{A,B,C} \mid T,\beta_i=1,\{C\}\right)p(\beta_i=1)}{p\left(g_i^{A,B,C} \mid \beta_i=0,\{C\}\right)p(\beta_i=0)}\right)}. \quad (20)$$

We can rewrite ***Equation 20*** as:

$$p\left(T \mid \left\{g_i^{A,B,C}\right\},\{C\}\right)=\frac{p(T)\prod_i\left(1+\frac{1}{p(T)}\mathcal{O}_i\,p\left(T \mid g_i^{A,B,C},\ \beta_i=1,\{C\}\right)\right)}{\sum_T p(T)\prod_i\left(1+\frac{1}{p(T)}\mathcal{O}_i\,p\left(T \mid g_i^{A,B,C},\ \beta_i=1,\{C\}\right)\right)}, \quad (21)$$

where $\mathcal{O}_i$ is the odds that gene i is a transition gene, given clustering:

$$\mathcal{O}_i=\frac{p\left(\beta_i=1 \mid g_i^{A,B,C},\{C\}\right)}{p\left(\beta_i=0 \mid g_i^{A,B,C},\{C\}\right)}=\frac{p\left(g_i^{A,B,C} \mid \beta_i=1,\{C\}\right)}{p\left(g_i^{A,B,C} \mid \beta_i=0,\{C\}\right)}\frac{p(\beta_i=1)}{p(\beta_i=0)} \quad (22)$$

and $p\left(T \mid g_i^{A,B,C},\ \beta_i=1,\{C\}\right)$ is the probability of T given only gene expression data for gene i, clustering and that gene i is a transition gene:

$$p\left(T \mid g_i^{A,B,C},\ \beta_i=1,\{C\}\right)=\frac{p\left(g_i^{A,B,C} \mid \beta_i=1,\ T,\{C\}\right)p(T)}{p\left(g_i^{A,B,C} \mid \beta_i=1,\{C\}\right)}. \quad (23)$$

Thus each gene's contribution $p\left(T \mid g_i^{A,B,C},\ \beta_i=1,\{C\}\right)$ to the probability of the topology given total gene expression $p\left(T \mid \left\{g_i^{A,B,C}\right\},\{C\}\right)$ is weighted by the odds $\mathcal{O}_i$ that it is transition gene.

Rewriting ***Equation 21*** in terms of negative votes

Let us denote the probability of gene expression data for gene i given that cell cluster ξ has the distribution with minimum mean expression as $p\left(g_i^{A,B,C} \mid \mu_\xi^i \text{ is min},\ \{\text{C}\}\right)$. For example, $p\left(g_i^{A,B,C} \mid \mu_B^i \text{ is min},\ \{\text{C}\}\right)=p\left(g_i^{A,B,C} \mid \mu_B^i<\mu_A^i,\mu_B^i<\mu_C^i,\ \{C\}\right)$. Then, using $p(T)=1/4$ and ***Equations 5 and 8***, we can write:

$$\begin{aligned} p\left(g_i^{A,B,C} \mid \beta_i=1,\{C\}\right) &= \tfrac{1}{4}\left[\sum_{T=\mathcal{A},\mathcal{B},\mathcal{C},\emptyset} p\left(g_i^{A,B,C} \mid T,\beta_i=0,\{C\}\right)\right] \\ &= \tfrac{1}{4}\left[\underbrace{\sum_{\xi=A,B,C} p\left(g_i^{A,B,C} \mid \mu_\xi^i \text{ is min, } \{C\}\right)}_{T=\mathcal{A},\mathcal{B},\mathcal{C}} + \tfrac{1}{3}\underbrace{\sum_{\xi=A,B,C} p\left(g_i^{A,B,C} \mid \mu_\xi^i \text{ is min, } \{C\}\right)}_{T=\emptyset}\right] \\ &= \tfrac{1}{3}\left[\sum_{\xi=A,B,C} p\left(g_i^{A,B,C} \mid \mu_\xi^i \text{ is min, } \{C\}\right)\right] \end{aligned} \tag{24}$$

Therefore, for $T=\mathcal{A},\mathcal{B},\mathcal{C}$, we can rewrite ***Equation 5*** as:

$$p\left(g_i^{A,B,C} \mid T,\beta_i=1,\{C\}\right) = \frac{1}{2}\left[3\,p\left(g_i^{A,B,C} \mid \beta_i=1,\{C\}\right) - p\left(g_i^{A,B,C} \mid \mu_T^i \text{ is min, } \{C\}\right)\right] \tag{25}$$

Combining ***Equations 21 and 25***, we derive, for $T\neq\emptyset$:

$$\begin{aligned} p\left(T \mid \left\{g_i^{A,B,C}\right\},\{C\}\right) &\propto p(T)\prod_i\left(1+\frac{p\left(g_i^{A,B,C}|T,\beta_i=1,\{C\}\right)}{p\left(g_i^{A,B,C}|\beta_i=1,\{C\}\right)}\mathcal{O}_i\right) \\ &\propto p(T)\prod_i\left(1+\frac{\frac{1}{2}\left[3\,p\left(g_i^{A,B,C} \mid \beta_i=1,\{C\}\right)-p\left(g_i^{A,B,C} \mid \mu_T^i \text{ is min, } \{C\}\right)\right]}{p\left(g_i^{A,B,C}|\beta_i=1,\{C\}\right)}\mathcal{O}_i\right) \\ &\propto\ p(T)\prod_i\left(1+\tfrac{3}{2}\mathcal{O}_i\left[1-p\left(\mu_T^i \text{ is min} \mid g_i^{A,B,C},\beta_i=1,\{C\}\right)\right]\right), \end{aligned} \tag{26}$$

where $p\left(\mu_T^i \text{ is min} \mid g_i^{A,B,C},\beta_i=1,\{C\}\right)$ is the probability that cell cluster T (the intermediate cluster in topology T) has the distribution with the minimum mean for gene i:

$$\begin{aligned} p\left(\mu_T^i \text{ is min} \mid g_i^{A,B,C},\beta_i=1,\{C\}\right) &= \frac{p\left(g_i^{A,B,C} \mid \mu_T^i \text{ is min},\{C\}\right)p\left(\mu_T^i \text{ is min}|\beta_i=1,\{C\}\right)}{p\left(g_i^{A,B,C} \mid \beta_i=1\right)} \\ &= \tfrac{1}{3}\frac{p\left(g_i^{A,B,C} \mid \mu_T^i \text{ is min, } \{C\}\right)}{p\left(g_i^{A,B,C} \mid \beta_i=1,\{C\}\right)}. \end{aligned} \tag{27}$$

Every gene can be thought of as casting a vote $-p\left(\mu_T^i \text{ is min} \mid g_i^{A,B,C},\beta_i=1,\{C\}\right)$ against cell type T being the intermediate, and this vote is weighted by the odds $\mathcal{O}_i$ of the gene i being a transition gene and having a unique minimum, given the clustering. This corresponds to ***Equation 1*** in the main text.

Expression for $p\left(T,\{\alpha_i\},\{\beta_i\} \mid \left\{g_i^{A,B,C}\right\},\{C\}\right)$

Once $p\left(T \mid \left\{g_i^{A,B,C}\right\},\{C\}\right)$ is calculated, it is straightforward to find $p\left(T,\{\alpha_i\},\{\beta_i\} \mid \left\{g_i^{A,B,C}\right\},\{C\}\right)$:

$$p\left(T,\{\alpha_i\},\{\beta_i\} \mid \left\{g_i^{A,B,C}\right\},\{C\}\right) = p\left(\{\alpha_i\},\{\beta_i\} \mid \left\{g_i^{A,B,C}\right\},T,\{C\}\right)p\left(T \mid \left\{g_i^{A,B,C}\right\},\{C\}\right), \tag{28}$$

where $p\left(\{\alpha_i\},\{\beta_i\} \mid \left\{g_i^{A,B,C}\right\},T,\{C\}\right)$ is the probability of $\{\alpha_i\}$ and $\{\beta_i\}$ given the particular topology T, clustering $\{C\}$ and gene expression. Because we have assumed that gene expression patterns $p\left(g_i^{A,B,C} \mid T,\{C\},\alpha_i,\beta_i\right)$ are conditionally independent given T, $\{C\}$, α_i and β_i (***Equation 3***), the probabilities of being marker or transition genes α_i or β_i are also conditionally independent given gene expression, clustering and the topology:

$$\begin{aligned} p\left(\{\alpha_i\},\{\beta_i\} \mid \left\{g_i^{A,B,C}\right\},T,\{C\}\right) &= \frac{p\left(\left\{g_i^{A,B,C}\right\} \mid T,\{\alpha_i\},\{\beta_i\},\{C\}\right)p\left(\{\alpha_i\},\{\beta_i\} \mid T,\{C\}\right)}{p\left(\left\{g_i^{A,B,C}\right\} \mid T,\{C\}\right)} \\ &= \prod_i \frac{p\left(g_i^{A,B,C} \mid T,\alpha_i,\beta_i,\{C\}\right)p\left(\alpha_i,\beta_i \mid T,\{C\}\right)}{p\left(g_i^{A,B,C} \mid T,\{C\}\right)} \\ &= \prod_i p\left(\alpha_i,\beta_i \mid T,\{C\},\ g_i^{A,B,C}\right), \end{aligned} \tag{29}$$

where $p\left(\alpha_i,\beta_i \mid T,\{C\},\ g_i^{A,B,C}\right)$ is the probability that gene i is a marker or transition gene given its gene expression, the clustering, and that the topology is T.

Choice of prior odds does not affect the most likely topology

The only free parameter in our calculation above is the prior odds of gene *i* being a transition gene, $p(\beta_i=1)/p(\beta_i=0)$. At one extreme, if $p(\beta_i=1)/p(\beta_i=0)\to 0$, then $p\left(T \mid \left\{g_i^{A,B,C}\right\}\right)\to p(T)$: if we assume that none of the genes are transition genes, then knowing gene expression does not give us any new knowledge of the topology T, since only transition genes are informative about T. At the other extreme, if $p(\beta_i=1)/p(\beta_i=0)\to\infty$ then the null hypothesis dominates: if all genes are transition genes, then there will be negative votes against all topologies. We computed the behavior of $p\left(T \mid \left\{g_i^{A,B,C}\right\}\right)$ between these two limits to determine the sensitivity of our answer to $p(\beta_i=1)/p(\beta_i=0)$.

Figure 2—figure supplement 1 shows the dependence of the probabilities $p\left(T \mid \left\{g_i^{A,B,C}\right\}\right)$ on the prior odds for triplets CMP/ST/MPP and GMP/MEP/FrBC for values of $p(\beta_i=1)/p(\beta_i=0)$ between 10^{-8} and 10^{2}. For triplet CMP/ST/MPP the topology $\boldsymbol{ST}\equiv CMP-\boldsymbol{ST}-MPP$ dominates for $p(\beta_i=1)/p(\beta_i=0)$ between 10^{-2} and 10, whereas for triplet MEP/GMP/FrBC there is no value of the prior odds that strongly favors a non-null topology. For most triplets, the most likely topology does not depend on the choice of prior odds; when building lineage trees, we ignore those triplets where different choices of prior odds lead to different most-likely topologies, i.e. there is more than one non-null topology that reaches probability 0.6 over the range of prior odds.

Determination of lineage tree from triplet topologies

Selection of triplets

In order to build lineage trees from the topologies we determine for each cell type, we select the triplets for which our determination of the topology is most robust. There is one free parameter in our model: the prior odds for a gene to be a transition gene in the absence of gene expression data, $p(\beta_i=1)/p(\beta_i=0)$. For each triplet, we vary this parameter between 10^{-6} and 10^{2} and calculate the probability of the topology given gene expression data $p\left(T \mid \left\{g_i^{A,B,C}\right\}\right)$ as a function of the prior odds.

We want to consider only triplets which showed a single dominant topology. We exclude triplets which show a weak probability for a particular topology or ones which depend on a particular choice of prior odds. We also do not consider triplets which show a strong probability for two different topologies, depending on the choice of prior odds.

There were 14 such triplets among the 165 hematopoietic triplets. Mathematically, these cases come up when the genes that are most likely to show the clear minimum pattern (furthest on the right in a 'dot plot') suggest one topology, but if one used a more permissive value of the sparsity parameter, a different topology wins out. One of the cell types might have a small number of genes with very high odds, but then fewer genes with moderately high odds compared to the other cell types. We did not notice a clear pattern in the identity of the triplets exhibiting this behavior, but 9 of the 14 were triplets of length five or greater in the Adolfsson model. One of the triplets with this behavior was the MLP/CMP/GMP triplet, and the dominant topology was either MLP (at low prior odds) or CMP (at higher prior odds). Interestingly, both cell types are progenitors to GMP in the Adolfsson model.

We consider triplets for which only one non-null topology has probability $p\left(T \mid \left\{g_i^{A,B,C}\right\}\right)$ greater than 0.6. The probabilities of the different topologies for each triplet in the hematopoietic tree are shown in ***Figure 2—source data 1*** and for the triplets in the cortical development tree in ***Figure 4—source data 2***.

Pruning rule

We assemble the triplets with known topology into an undirected graph. Since we determined topologies by considering cell types three at a time, we obtain topological relationships involving both cell types that are nearest neighbors and cell types that are more distantly related. In order to

reconstruct the tree, we must determine which cell types are nearest neighbors and which ones are separated by one or more intermediate cell types.

The set of inferred topologies allows us to determine which cell types are separated by intermediates. For every pair of cell types, we ask whether any of the inferred topologies features an intermediate between the two cell types. If such a topology has been inferred, we consider that the two cell types are not nearest neighbors, and that at least one other cell type is an intermediate. For example, we can ignore triplet CMP – **LT** – MPP because triplet LT – **ST** – CMP testifies that there exists an intermediate between LT and CMP, and triplet LT – **ST** – MPP testifies that there exists an intermediate between LT and MPP (***Figure 3—figure supplement 1***).

Note that this pruning rule does not assume the absence of loops. The lineage tree we infer for the hematopoietic progenitors contains a loop that includes ST to CMP to GMP on one side and ST to MPP to MLP to GMP on the other side. The loop involves triplets CMP – **ST** – MPP, ST – **CMP** – GMP, ST – **MPP** – MLP and MPP – **MLP** – GMP (we cannot determine the topology of triplet CMP/MLP/GMP). None of the triplets shows a topology that would allow us to break up the loop.

Distinguishing between two models of hematopoiesis

The topologies we infer support the model from ***Adolfsson et al. (2005)***, in which CMP splits from ST-HSC. In particular, there are several triplets that can distinguish the Adolfsson model from the traditional picture, and they support the Adolfsson model. These triplets include CMP – **ST** – MPP, CMP – **ST** – MLP, MEP – **ST** – MLP and CMP – **LT** – MPP. These triplets show that, unlike in the traditional picture, cell types CMP and MEP split from ST are not descended from MPP or MLP.

On the other hand, triplets LT – **MLP** – GMP, ST – **MLP** – GMP and MPP – **MLP** – GMP show that MLP is an intermediate between the earliest progenitors and GMP. See also ***Figure 2—figure supplement 2*** and ***Figure 2—source data 2***.

Stability analysis

The inference algorithm depends on several parameters and priors. We performed a stability analysis for both the microarray hematopoietic data and the human brain single cell data to determine the parameter ranges for which the inferred lineage tree was unchanged.

1. Hematopoiesis
   - For the prior probability, given that a gene is not a transition gene, that it is an irrelevant gene, our default value was $p(\alpha_i = 0 \mid \beta_i = 0)$ =0.5, but the tree was unchanged for values of $p(\alpha_i = 0 \mid \beta_i = 0)$ between 0.25 and 0.65.
   - For the prior probabilities of different topologies, our default value was $p(\emptyset) = p(\mathcal{A}) = p(\mathcal{B}) = p(\mathcal{C}) = 0.25$. We varied $p(\emptyset)$ while keeping the prior probabilities of the non-null topologies equal: $p(T \neq \emptyset) = \frac{1}{3}(1 - p(\emptyset))$. The tree was unchanged for values of $p(\emptyset)$ between 0.1 and 0.35.
   - We used a threshold of 0.6 to consider a triplet significant for the tree-building step. The tree was unchanged for thresholds between 0.5 and 0.65.
   - A key input parameter into the algorithm is the expected prior distribution of means and standard deviations $p(\mu, \sigma)$ used in the numeric integration. Our default prior was a uniform prior for both means and standard deviations over reasonable ranges of these parameters. We also implemented an empirical prior $p(\mu, \sigma)$ by estimating the empirical distribution over all the Immgen cell types, using kernel density estimation (***Botev et al., 2010***). The resulting hematopoietic lineage tree was identical. Using the kernel-density-estimated empirical prior may provide more stability in future analyses.
2. Brain Development
   - For the prior probability, given that a gene is not a transition gene, that it is an irrelevant gene, our default value was $p(\alpha_i = 0 \mid \beta_i = 0)$ = 0.5, but the tree was unchanged for values of $p(\alpha_i = 0 \mid \beta_i = 0)$ between 0.02 and 0.97.
   - For the prior probabilities of different topologies, our default value was $p(\emptyset) = p(\mathcal{A}) = p(\mathcal{B}) = p(\mathcal{C}) = 0.25$. We varied $p(\emptyset)$ while keeping the prior probabilities of the non-null topologies equal: $p(T \neq \emptyset) = \frac{1}{3}(1 - p(\emptyset))$. The tree was unchanged for values of $p(\emptyset)$ between 0.06 and 0.96.
   - We used a threshold of 0.6 to consider a triplet significant for the tree-building step. The tree was unchanged for thresholds between 0.5 and 0.8.

- For the single cell data, we also changed the number of initial seed clusters for the iterative algorithm within a range from 30 to 45. In each case, the tree remained the same, and never more than 23% of the cells were clustered into a different cell type.

Acknowledgements

We thank Sandeep Choubey, Alex Schier, Andrew Murray, David Scadden and Toshihiko Oki for scientific discussions. We also thank Christof Koch, Andrew Murray, KC Huang, Jim Valcourt and Dann Huh for detailed comments and feedback on this work. We are grateful to Simona Lodato for helping us interpret our results and in particular help us write the section on the genes we discover to be important during cortical development. We are particularly grateful to the Senior and Reviewing Editors and to the four peer reviewers, including Nir Yosef, for their very thoughtful comments and suggestions. The study was supported by NSF-GRFP and NDSEG Fellowships (LF), an NIH Pioneer Award (SR), and the Allen Institute for Brain Science (SR).

Additional information

Funding

Funder	Author
National Science Foundation	Leon A. Furchtgott
National Institutes of Health	Sharad Ramanathan
Allen Foundation	Vilas Menon Sharad Ramanathan

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

LAF, Conceptualization, Data curation, Software, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing; SM, Conceptualization, Data curation, Software, Formal analysis, Validation, Visualization, Methodology, Writing—original draft, Writing—review and editing; VM, Validation, Visualization, Writing—review and editing; SR, Conceptualization, Resources, Formal analysis, Supervision, Funding acquisition, Investigation, Visualization, Methodology, Writing—original draft, Project administration, Writing—review and editing

Author ORCIDs

Leon A Furchtgott, http://orcid.org/0000-0002-4258-0950

Additional files

Major datasets

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database, license, and accessibility information
Gopalan G	2009	Immunological Genome Project	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15907	Publicly available at the NCBI Gene Expression Omnibus (accession no: GSE15907)
Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A	2015	Single-Cell mRNA Sequencing Reveals Rare Intestinal Cell Types	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62270	Publicly available at the NCBI Gene Expression Omnibus (accession no: GSE62270)
Yao Z, Mich JK, Ku	2016	Region-specific neural stem cell	https://www.ncbi.nlm.	Publicly available at

S, Menon V, Krostag A, Martinez RA, Grimley JS, Wang Y, Ramanathan S, Levi BP	lineages revealed by single-cell rna-seq from human embryonic stem cells [Smart-seq]	nih.gov/geo/query/acc.cgi?acc=GSE86982	the NCBI Gene Expression Omnibus (accession no: GSE86982)

References

Abraham AB, Bronstein R, Chen EI, Koller A, Ronfani L, Maletic-Savatic M, Tsirka SE. 2013. Members of the high mobility group B protein family are dynamically expressed in embryonic neural stem cells. *Proteome Science* **11**:18. doi: 10.1186/1477-5956-11-18, PMID: 23621913

Adolfsson J, Månsson R, Buza-Vidas N, Hultquist A, Liuba K, Jensen CT, Bryder D, Yang L, Borge OJ, Thoren LA, Anderson K, Sitnicka E, Sasaki Y, Sigvardsson M, Jacobsen SE. 2005. Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell* **121**:295–306. doi: 10.1016/j.cell.2005.02.013, PMID: 15851035

Advani M, Ganguli S. 2016. Statistical mechanics of optimal convex inference in high dimensions. *Physical Review X* **6**:031034. doi: 10.1103/PhysRevX.6.031034

Agoston Z, Li N, Haslinger A, Wizenmann A, Schulte D. 2012. Genetic and physical interaction of Meis2, Pax3 and Pax7 during dorsal midbrain development. *BMC Developmental Biology* **12**:10. doi: 10.1186/1471-213X-12-10, PMID: 22390724

Akashi K, Traver D, Miyamoto T, Weissman IL. 2000. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404**:193–197. doi: 10.1038/35004599, PMID: 10724173

Anderson PW. 1978. Local moments and localized states. *Reviews of Modern Physics* **50**:191–201. doi: 10.1103/RevModPhys.50.191

Ang SL. 2006. Transcriptional control of midbrain dopaminergic neuron development. *Development* **133**:3499–3506. doi: 10.1242/dev.02501, PMID: 16899537

Appolloni I, Calzolari F, Corte G, Perris R, Malatesta P. 2008. Six3 controls the neural progenitor status in the murine CNS. *Cerebral Cortex* **18**:553–562. doi: 10.1093/cercor/bhm092, PMID: 17576749

Au E, Ahmed T, Karayannis T, Biswas S, Gan L, Fishell G. 2013. A modular gain-of-function approach to generate cortical interneuron subtypes from ES cells. *Neuron* **80**:1145–1158. doi: 10.1016/j.neuron.2013.09.022, PMID: 24314726

Azim E, Jabaudon D, Fame RM, Macklis JD. 2009. SOX6 controls dorsal progenitor identity and interneuron diversity during neocortical development. *Nature Neuroscience* **12**:1238–1247. doi: 10.1038/nn.2387, PMID: 19657336

Bani-Yaghoub M, Tremblay RG, Lei JX, Zhang D, Zurakowski B, Sandhu JK, Smith B, Ribecco-Lutkiewicz M, Kennedy J, Walker PR, Sikorska M. 2006. Role of Sox2 in the development of the mouse neocortex. *Developmental Biology* **295**:52–66. doi: 10.1016/j.ydbio.2006.03.007, PMID: 16631155

Baraniuk RG. 2007. Compressive sensing. *IEEE signal processing magazine* **118**:12–13.

Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A, Weinberg RA. 2008. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nature Genetics* **40**:499–507. doi: 10.1038/ng.127, PMID: 18443585

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* **57**:289–3000.

Borello U, Madhavan M, Vilinsky I, Faedo A, Pierani A, Rubenstein J, Campbell K. 2014. Sp8 and COUP-TF1 reciprocally regulate patterning and Fgf signaling in cortical progenitors. *Cerebral Cortex* **24**:1409–1421. doi: 10.1093/cercor/bhs412, PMID: 23307639

Botev ZI, Grotowski JF, Kroese DP. 2010. Kernel density estimation via diffusion. *The Annals of Statistics* **38**:2916–2957. doi: 10.1214/10-AOS799

Buck A, Kispert A, Kohlhase J. 2001. Embryonic expression of the murine homologue of SALL1, the gene mutated in Townes–Brocks syndrome. *Mechanisms of Development* **104**:143–146. doi: 10.1016/S0925-4773(01)00364-1, PMID: 11404093

Buckingham ME, Meilhac SM. 2011. Tracing cells for tracking cell lineage and clonal behavior. *Developmental Cell* **21**:394–409. doi: 10.1016/j.devcel.2011.07.019, PMID: 21920310

Candès EJ, Romberg JK, Tao T. 2006. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics* **59**:1207–1223. doi: 10.1002/cpa.20124

Chang CY, Pasolli HA, Giannopoulou EG, Guasch G, Gronostajski RM, Elemento O, Fuchs E. 2013. NFIB is a governor of epithelial-melanocyte stem cell behaviour in a shared niche. *Nature* **495**:98–102. doi: 10.1038/nature11847, PMID: 23389444

Cimadamore F, Amador-Arjona A, Chen C, Huang CT, Terskikh AV. 2013. SOX2-LIN28/let-7 pathway regulates proliferation and neurogenesis in neural precursors. *PNAS* **110**:E3017–3026. doi: 10.1073/pnas.1220176110, PMID: 23884650

Core Team R. 2015. *R: A Language and Environment for Statistical Computing*.

Crispino JD. 2005. GATA1 in normal and malignant hematopoiesis. *Seminars in Cell & Developmental Biology* **16**:137–147. doi: 10.1016/j.semcdb.2004.11.002, PMID: 15659348

Cánovas J, Berndt FA, Sepúlveda H, Aguilar R, Veloso FA, Montecino M, Oliva C, Maass JC, Sierralta J, Kukuljan M. 2015. The specification of cortical subcerebral projection neurons depends on the direct repression of TBR1

by CTIP1/BCL11a. *Journal of Neuroscience* **35**:7552–7564. doi: 10.1523/JNEUROSCI.0169-15.2015, PMID: 25972180

de Santa Barbara P, van den Brink GR, Roberts DJ. 2003. Development and differentiation of the intestinal epithelium. *Cellular and Molecular Life Sciences* **60**:1322–1332. doi: 10.1007/s00018-003-2289-3, PMID: 12943221

Delmans M, Hemberg M. 2016. Discrete distributional differential expression (D3E)–a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics* **17**:110. doi: 10.1186/s12859-016-0944-6, PMID: 26927822

Di Bonito M, Narita Y, Avallone B, Sequino L, Mancuso M, Andolfi G, Franzè AM, Puelles L, Rijli FM, Studer M. 2013. Assembly of the auditory circuitry by a Hox genetic network in the mouse brainstem. *PLoS Genetics* **9**: e1003249. doi: 10.1371/journal.pgen.1003249, PMID: 23408898

Dominguez MH, Ayoub AE, Rakic P. 2013. POU-III transcription factors (Brn1, Brn2, and Oct6) influence neurogenesis, molecular identity, and migratory destination of upper-layer cells of the cerebral cortex. *Cerebral Cortex* **23**:2632–2643. doi: 10.1093/cercor/bhs252, PMID: 22892427

Donoho D, Tanner J. 2009. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**:4273–4293. doi: 10.1098/rsta.2009.0152, PMID: 19805445

Duggan SP, Behan FM, Kirca M, Zaheer A, McGarrigle SA, Reynolds JV, Vaz GM, Senge MO, Kelleher D. 2016. The characterization of an intestine-like genomic signature maintained during Barrett's-associated adenocarcinogenesis reveals an NR5A2-mediated promotion of cancer cell survival. *Scientific Reports* **6**:32638. doi: 10.1038/srep32638, PMID: 27586588

Ebisu H, Iwai-Takekoshi L, Fujita-Jimbo E, Momoi T, Kawasaki H. 2016. Foxp2 regulates identities and projection patterns of thalamic nuclei during development. *Cerebral Cortex*:bhw187. doi: 10.1093/cercor/bhw187

Elsen GE, Choi LY, Millen KJ, Grinblat Y, Prince VE. 2008. Zic1 and Zic4 regulate zebrafish roof plate specification and hindbrain ventricle morphogenesis. *Developmental Biology* **314**:376–392. doi: 10.1016/j.ydbio.2007.12.006, PMID: 18191121

Enkhmandakh B, Makeyev AV, Erdenechimeg L, Ruddle FH, Chimge NO, Tussie-Luna MI, Roy AL, Bayarsaihan D. 2009. Essential functions of the Williams-Beuren syndrome-associated TFII-I genes in embryonic development. *PNAS* **106**:181–186. doi: 10.1073/pnas.0811531106, PMID: 19109438

Erickson T, French CR, Waskiewicz AJ. 2010. Meis1 specifies positional information in the retina and tectum to organize the zebrafish visual system. *Neural Development* **5**:22. doi: 10.1186/1749-8104-5-22, PMID: 20809932

Fawcett SR, Klymkowsky MW. 2004. Embryonic expression of xenopus laevis SOX7. *Gene Expression Patterns* **4**: 29–33. doi: 10.1016/j.modgep.2003.08.003

Ferrell JE. 2012. Bistability, bifurcations, and Waddington's epigenetic landscape. *Current Biology* **22**:R458–R466. doi: 10.1016/j.cub.2012.03.045, PMID: 22677291

Fre S, Huyghe M, Mourikis P, Robine S, Louvard D, Artavanis-Tsakonas S. 2005. Notch signals control the fate of immature progenitor cells in the intestine. *Nature* **435**:964–968. doi: 10.1038/nature03589, PMID: 15959516

Frumkin D, Wasserstrom A, Itzkovitz S, Stern T, Harmelin A, Eilam R, Rechavi G, Shapiro E. 2008. Cell lineage analysis of a mouse tumor. *Cancer Research* **68**:5924–5931. doi: 10.1158/0008-5472.CAN-07-6216, PMID: 18632647

Garcia H, Fleyshman D, Kolesnikova K, Safina A, Commane M, Paszkiewicz G, Omelian A, Morrison C, Gurova K. 2011. Expression of FACT in mammalian tissues suggests its role in maintaining of undifferentiated state of cells. *Oncotarget* **2**:783–796. doi: 10.18632/oncotarget.340, PMID: 21998152

Gaston-Massuet C, McCabe MJ, Scagliotti V, Young RM, Carreno G, Gregory LC, Jayakody SA, Pozzi S, Gualtieri A, Basu B, Koniordou M, Wu CI, Bancalari RE, Rahikkala E, Veijola R, Lopponen T, Graziola F, Turton J, Signore M, Mousavy Gharavy SN, et al. 2016. Transcription factor 7-like 1 is involved in hypothalamo-pituitary axis development in mice and humans. *PNAS* **113**:E548–557. doi: 10.1073/pnas.1503346113, PMID: 26764381

Gazit R, Garrison BS, Rao TN, Shay T, Costello J, Ericson J, Kim F, Collins JJ, Regev A, Wagers AJ, Rossi DJ, Immunological Genome Project Consortium. 2013. Transcriptome analysis identifies regulators of hematopoietic stem and progenitor cells. *Stem Cell Reports* **1**:266–280. doi: 10.1016/j.stemcr.2013.07.004, PMID: 24319662

Gilbert SF. 2014. *Developmental biology*. Sinauer.

Goossens S, Janzen V, Bartunkova S, Yokomizo T, Drogat B, Crisan M, Haigh K, Seuntjens E, Umans L, Riedt T, Bogaert P, Haenebalcke L, Berx G, Dzierzak E, Huylebroeck D, Haigh JJ. 2011. The EMT regulator Zeb2/Sip1 is essential for murine embryonic hematopoietic stem/progenitor cell differentiation and mobilization. *Blood* **117**: 5620–5630. doi: 10.1182/blood-2010-08-300236, PMID: 21355089

Graf T, Enver T. 2009. Forcing cells to change lineages. *Nature* **462**:587–594. doi: 10.1038/nature08533, PMID: 19956253

Greig LC, Woodworth MB, Galazo MJ, Padmanabhan H, Macklis JD. 2013. Molecular logic of neocortical projection neuron specification, development and diversity. *Nature Reviews Neuroscience* **14**:755–769. doi: 10.1038/nrn3586, PMID: 24105342

Greig LC, Woodworth MB, Greppi C, Macklis JD, Abdel-Majid RM, Leong WL, Schalkwyk LC, Smallman DS, Wong ST, Storm DR. 2016. Ctip1 controls acquisition of sensory area identity and establishment of sensory input fields in the developing neocortex. *Neuron* **90**:261–277. doi: 10.1016/j.neuron.2016.03.008, PMID: 27100196

Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. 2015. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**:251–255. doi: 10.1038/nature14966, PMID: 26287467

Grün D, Muraro MJ, Boisset JC, Wiebrands K, Lyubimova A, Dharmadhikari G, van den Born M, van Es J, Jansen E, Clevers H, de Koning EJ, van Oudenaarden A. 2016. De novo prediction of stem cell identity using Single-Cell transcriptome data. *Cell Stem Cell* **19**:266–277. doi: 10.1016/j.stem.2016.05.010, PMID: 27345837

Hagey DW, Muhr J. 2014. Sox2 acts in a dose-dependent fashion to regulate proliferation of cortical progenitors. *Cell Reports* **9**:1908–1920. doi: 10.1016/j.celrep.2014.11.013, PMID: 25482558

Hegarty SV, Sullivan AM, O'Keeffe GW. 2013. Midbrain dopaminergic neurons: a review of the molecular circuitry that regulates their development. *Developmental Biology* **379**:123–138. doi: 10.1016/j.ydbio.2013.04.014, PMID: 23603197

Heng TS, Painter MW, Immunological Genome Project Consortium. 2008. The immunological genome project: networks of gene expression in immune cells. *Nature Immunology* **9**:1091–1094. doi: 10.1038/ni1008-1091, PMID: 18800157

Hutton SR, Pevny LH. 2011. SOX2 expression levels distinguish between neural progenitor populations of the developing dorsal telencephalon. *Developmental Biology* **352**:40–47. doi: 10.1016/j.ydbio.2011.01.015, PMID: 21256837

Inoue F, Kurokawa D, Takahashi M, Aizawa S. 2012. Gbx2 directly restricts Otx2 expression to forebrain and midbrain, competing with class III POU factors. *Molecular and Cellular Biology* **32**:2618–2627. doi: 10.1128/MCB.00083-12, PMID: 22566684

Iwasaki H, Akashi K. 2007. Myeloid lineage commitment from the hematopoietic stem cell. *Immunity* **26**:726–740. doi: 10.1016/j.immuni.2007.06.004, PMID: 17582345

Jabaudon D, Shnider SJ, Tischfield DJ, Galazo MJ, Macklis JD. 2012. Rorβ induces barrel-like neuronal clusters in the developing neocortex. *Cerebral Cortex* **22**:996–1006. doi: 10.1093/cercor/bhr182, PMID: 21799210

Jaegle M, Ghazvini M, Mandemakers W, Piirsoo M, Driegen S, Levavasseur F, Raghoenath S, Grosveld F, Meijer D. 2003. The POU proteins Brn-2 and Oct-6 share important functions in schwann cell development. *Genes & development* **17**:1380–1391. doi: 10.1101/gad.258203, PMID: 12782656

Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I. 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**:776–779. doi: 10.1126/science.1247651, PMID: 24531970

Jang S, Furchtgott L, Choubey S, Zou L-N, Doyle A, Menon V, Loew E, Krostag A-R, Martinez RA, Madisen L, Levi BP, Ramanathan S. 2017. Probabilistic model of gene networks controlling embryonic stem cell differentiation inferred from single-cell transcriptomics. *eLife* **6**:e20487. doi: 10.7554/eLife.20487

Jenny M, Uhl C, Roche C, Duluc I, Guillermin V, Guillemot F, Jensen J, Kedinger M, Gradwohl G. 2002. Neurogenin3 is differentially required for endocrine cell fate specification in the intestinal and gastric epithelium. *The EMBO Journal* **21**:6338–6347. doi: 10.1093/emboj/cdf649, PMID: 12456641

Jensen J, Pedersen EE, Galante P, Hald J, Heller RS, Ishibashi M, Kageyama R, Guillemot F, Serup P, Madsen OD. 2000. Control of endodermal endocrine development by Hes-1. *Nature Genetics* **24**:36–44. doi: 10.1038/71657, PMID: 10615124

Ji Z, Ji H. 2016. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research* **44**:e117. doi: 10.1093/nar/gkw430, PMID: 27179027

Johansson PA, Irmler M, Acampora D, Beckers J, Simeone A, Götz M. 2013. The transcription factor Otx2 regulates choroid plexus development and function. *Development* **140**:1055–1066. doi: 10.1242/dev.090860, PMID: 23364326

Kameda Y, Saitoh T, Fujimura T. 2011. Hes1 regulates the number and anterior-posterior patterning of mesencephalic dopaminergic neurons at the mid/hindbrain boundary (isthmus). *Developmental Biology* **358**:91–101. doi: 10.1016/j.ydbio.2011.07.016, PMID: 21798254

Kanatani S, Yozu M, Tabata H, Nakajima K. 2008. COUP-TFII is preferentially expressed in the caudal ganglionic eminence and is involved in the caudal migratory stream. *Journal of Neuroscience* **28**:13582–13591. doi: 10.1523/JNEUROSCI.2132-08.2008, PMID: 19074032

Katz JP, Perreault N, Goldstein BG, Lee CS, Labosky PA, Yang VW, Kaestner KH. 2002. The zinc-finger transcription factor Klf4 is required for terminal differentiation of goblet cells in the colon. *Development* **129**:2619–2628. PMID: 12015290

Kessaris N, Magno L, Rubin AN, Oliveira MG. 2014. Genetic programs controlling cortical Interneuron fate. *Current Opinion in Neurobiology* **26**:79–87. doi: 10.1016/j.conb.2013.12.012, PMID: 24440413

Khan WI, Blennerhasset P, Ma C, Matthaei KI, Collins SM. 2001. Stat6 dependent goblet cell hyperplasia during intestinal nematode infection. *Parasite Immunology* **23**:39–42. doi: 10.1046/j.1365-3024.2001.00353.x, PMID: 11136476

Kikkawa T, Obayashi T, Takahashi M, Fukuzaki-Dohi U, Numayama-Tsuruta K, Osumi N. 2013. Dmrta1 regulates proneural gene expression downstream of Pax6 in the mammalian telencephalon. *Genes to Cells* **18**:636–649. doi: 10.1111/gtc.12061, PMID: 23679989

Kishi Y, Fujii Y, Hirabayashi Y, Gotoh Y. 2012. HMGA regulates the global chromatin state and neurogenic potential in neocortical precursor cells. *Nature Neuroscience* **15**:1127–1133. doi: 10.1038/nn.3165, PMID: 22797695

Knight JM, Davidson LA, Herman D, Martin CR, Goldsby JS, Ivanov IV, Donovan SM, Chapkin RS. 2014. Non-invasive analysis of intestinal development in preterm and term infants using RNA-Sequencing. *Scientific Reports* **4**:159–173. doi: 10.1038/srep05453

Kondo M, Weissman IL, Akashi K. 1997. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell* **91**:661–672. doi: 10.1016/S0092-8674(00)80453-5, PMID: 9393859

Kumbasar A, Plachez C, Gronostajski RM, Richards LJ, Litwack ED. 2009. Absence of the transcription factor *Nfib* delays the formation of the basilar pontine and other mossy fiber nuclei. *The Journal of Comparative Neurology* **513**:98–112. doi: 10.1002/cne.21943, PMID: 19107796

Kunath M, Lüdecke HJ, Vortkamp A. 2002. Expression of Trps1 during mouse embryonic development. *Mechanisms of Development* **119 Suppl 1**:S117–120 . doi: 10.1016/S0925-4773(03)00103-5, PMID: 14516672

Kurotaki D, Osato N, Nishiyama A, Yamamoto M, Ban T, Sato H, Nakabayashi J, Umehara M, Miyake N, Matsumoto N, Nakazawa M, Ozato K, Tamura T. 2013. Essential role of the IRF8-KLF4 transcription factor cascade in murine monocyte differentiation. *Blood* **121**:1839–1849. doi: 10.1182/blood-2012-06-437863, PMID: 23319570

Landau LD, Lifshitz EM. 1951. *Statistical Physics*. Elsevier. **5**.https://books.google.com/books?id=VzgJN-XPTRsC&pgis=1

Lavado A, Oliver G. 2007. Prox1 expression patterns in the developing and adult murine brain. *Developmental Dynamics* **236**:518–524. doi: 10.1002/dvdy.21024, PMID: 17117441

Li X, Udager AM, Hu C, Qiao XT, Richards N, Gumucio DL. 2009. Dynamic patterning at the pylorus: formation of an epithelial intestine-stomach boundary in late fetal life. *Developmental Dynamics* **238**:3205–3217. doi: 10.1002/dvdy.22134, PMID: 19877272

Lodato S, Molyneaux BJ, Zuccaro E, Goff LA, Chen HH, Yuan W, Meleski A, Takahashi E, Mahony S, Rinn JL, Gifford DK, Arlotta P. 2014. Gene co-regulation by Fezf2 selects neurotransmitter identity and connectivity of corticospinal neurons. *Nature Neuroscience* **17**:1046–1054. doi: 10.1038/nn.3757, PMID: 24997765

Lorente-Trigos A, Varnat F, Melotti A, Ruiz i Altaba A. 2010. BMP signaling promotes the growth of primary human colon carcinomas in vivo. *Journal of Molecular Cell Biology* **2**:318–332. doi: 10.1093/jmcb/mjq035, PMID: 21098050

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. 2015. Highly parallel Genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**:1202–1214. doi: 10.1016/j.cell.2015.05.002, PMID: 26000488

Madissoon E, Jouhilahti EM, Vesterlund L, Töhönen V, Krjutškov K, Petropoulous S, Einarsdottir E, Linnarsson S, Lanner F, Månsson R, Hovatta O, Bürglin TR, Katayama S, Kere J. 2016. Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. *Scientific Reports* **6**:28995. doi: 10.1038/srep28995, PMID: 27412763

Manuel MN, Martynoga B, Molinek MD, Quinn JC, Kroemmer C, Mason JO, Price DJ. 2011. The transcription factor Foxg1 regulates telencephalic progenitor proliferation cell autonomously, in part by controlling Pax6 expression levels. *Neural Development* **6**:9. doi: 10.1186/1749-8104-6-9, PMID: 21418559

Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, Yuan GC. 2014. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *PNAS* **111**:E5643–E5650. doi: 10.1073/pnas.1408993111, PMID: 25512504

Marjoram L, Alvers A, Deerhake ME, Bagwell J, Mankiewicz J, Cocchiaro JL, Beerman RW, Willer J, Sumigray KD, Katsanis N, Tobin DM, Rawls JF, Goll MG, Bagnat M. 2015. Epigenetic control of intestinal barrier function and inflammation in zebrafish. *PNAS* **112**:2770–2775. doi: 10.1073/pnas.1424089112, PMID: 25730872

Matsui H, Kimura A, Yamashiki N, Moriyama A, Kaya M, Yoshida I, Takagi N, Takahashi T. 2000. Molecular and biochemical characterization of a serine proteinase predominantly expressed in the medulla oblongata and cerebellar white matter of mouse brain. *Journal of Biological Chemistry* **275**:11050–11057. doi: 10.1074/jbc.275.15.11050, PMID: 10753908

McGibbon RT, Husic BE, Pande VS. 2017. Identification of simple reaction coordinates from complex dynamics. *The Journal of Chemical Physics* **146**:044109. doi: 10.1063/1.4974306

Mellor P, Deibert L, Calvert B, Bonham K, Carlsen SA, Anderson DH. 2013. CREB3L1 is a metastasis suppressor that represses expression of genes regulating metastasis, invasion, and angiogenesis. *Molecular and Cellular Biology* **33**:4985–4995. doi: 10.1128/MCB.00959-13, PMID: 24126059

Merrill BJ, Gat U, DasGupta R, Fuchs E. 2001. Tcf3 and Lef1 regulate lineage differentiation of multipotent stem cells in skin. *Genes & Development* **15**:1688–1705. doi: 10.1101/gad.891401, PMID: 11445543

Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall JJ, Aiona K, Arnold JM, Bennet C, Bertagnolli D, Brouner K, Butler S, Caldejon S, Carey A, Cuhaciyan C, Dalley RA, Dee N, et al. 2014. Transcriptional landscape of the prenatal human brain. *Nature* **508**:199–206. doi: 10.1038/nature13185, PMID: 24695229

Miller RL, Stein MK, Loewy AD. 2011. Serotonergic inputs to FoxP2 neurons of the pre-locus coeruleus and parabrachial nuclei that project to the ventral tegmental area. *Neuroscience* **193**:229–240. doi: 10.1016/j.neuroscience.2011.07.008, PMID: 21784133

Milosevic J, Maisel M, Wegner F, Leuchtenberger J, Wenger RH, Gerlach M, Storch A, Schwarz J. 2007. Lack of hypoxia-inducible factor-1 alpha impairs midbrain neural precursor cells involving vascular endothelial growth factor signaling. *Journal of Neuroscience* **27**:412–421. doi: 10.1523/JNEUROSCI.2482-06.2007, PMID: 17215402

Miyawaki K, Arinobu Y, Iwasaki H, Kohno K, Tsuzuki H, Iino T, Shima T, Kikushige Y, Takenaka K, Miyamoto T, Akashi K. 2015. CD41 marks the initial myelo-erythroid lineage specification in adult mouse hematopoiesis: redefinition of murine common myeloid progenitor. *Stem Cells* **33**:976–987. doi: 10.1002/stem.1906, PMID: 25446279

Miyoshi G, Fishell G. 2012. Dynamic FoxG1 expression coordinates the integration of Multipolar pyramidal neuron precursors into the cortical plate. *Neuron* **74**:1045–1058. doi: 10.1016/j.neuron.2012.04.025, PMID: 22726835

Muncan V, Sansom OJ, Tertoolen L, Phesse TJ, Begthel H, Sancho E, Cole AM, Gregorieff A, de Alboran IM, Clevers H, Clarke AR. 2006. Rapid loss of intestinal crypts upon conditional deletion of the Wnt/Tcf-4 target gene c-Myc. *Molecular and Cellular Biology* **26**:8418–8426. doi: 10.1128/MCB.00821-06, PMID: 16954380

Negishi H, Miki S, Sarashina H, Taguchi-Atarashi N, Nakajima A, Matsuki K, Endo N, Yanai H, Nishio J, Honda K, Taniguchi T. 2012. Essential contribution of IRF3 to intestinal homeostasis and microbiota-mediated Tslp gene induction. *PNAS* **109**:21016–21021. doi: 10.1073/pnas.1219482110, PMID: 23213237

Noah TK, Kazanjian A, Whitsett J, Shroyer NF. 2010. SAM pointed domain ETS factor (SPDEF) regulates terminal differentiation and maturation of intestinal goblet cells. *Experimental Cell Research* **316**:452–465. doi: 10.1016/j.yexcr.2009.09.020, PMID: 19786015

Ohba H, Chiyoda T, Endo E, Yano M, Hayakawa Y, Sakaguchi M, Darnell RB, Okano HJ, Okano H. 2004. Sox21 is a repressor of neuronal differentiation and is antagonized by YB-1. *Neuroscience Letters* **358**:157–160. doi: 10.1016/j.neulet.2004.01.026, PMID: 15039105

Ohtsuka T, Sakamoto M, Guillemot F, Kageyama R. 2001. Roles of the basic helix-loop-helix genes Hes1 and Hes5 in expansion of neural stem cells of the developing brain. *Journal of Biological Chemistry* **276**:30467–30474. doi: 10.1074/jbc.M102420200, PMID: 11399758

Olivetti PR, Noebels JL. 2012. Interneuron, interrupted: molecular pathogenesis of ARX mutations and X-linked infantile spasms. *Current Opinion in Neurobiology* **22**:859–865. doi: 10.1016/j.conb.2012.04.006, PMID: 22565167

Orkin SH, Zon LI. 2008. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**:631–644. doi: 10.1016/j.cell.2008.01.025, PMID: 18295580

Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, Louis DN, Rozenblatt-Rosen O, Suvà ML, Regev A, Bernstein BE. 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**:1396–1401. doi: 10.1126/science.1254257, PMID: 24925914

Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, Winter D, Lara-Astiaso D, Gury M, Weiner A, David E, Cohen N, Lauridsen FK, Haas S, Schlitzer A, Mildner A, Ginhoux F, Jung S, Trumpp A, Porse BT, et al. 2015. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**:1663–1677. doi: 10.1016/j.cell.2015.11.013, PMID: 26627738

Perrimon N, Pitsouli C, Shilo BZ. 2012. Signaling mechanisms controlling cell fate and embryonic patterning. *Cold Spring Harbor Perspectives in Biology* **4**:a005975. doi: 10.1101/cshperspect.a005975, PMID: 22855721

Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10**:1096–1098. doi: 10.1038/nmeth.2639, PMID: 24056875

Pino E, Amamoto R, Zheng L, Cacquevel M, Sarria JC, Knott GW, Schneider BL. 2014. FOXO3 determines the accumulation of α-synuclein and controls the fate of dopaminergic neurons in the substantia nigra. *Human Molecular Genetics* **23**:1435–1452. doi: 10.1093/hmg/ddt530, PMID: 24158851

Pozniak CD, Langseth AJ, Dijkgraaf GJ, Choe Y, Werb Z, Pleasure SJ. 2010. Sox10 directs neural stem cells toward the oligodendrocyte lineage by decreasing suppressor of fused expression. *PNAS* **107**:21795–21800. doi: 10.1073/pnas.1016485107, PMID: 21098272

Qi X, Hong J, Chaves L, Zhuang Y, Chen Y, Wang D, Chabon J, Graham B, Ohmori K, Li Y, Huang H. 2013. Antagonistic regulation by the transcription factors C/EBPα and MITF specifies basophil and mast cell fates. *Immunity* **39**:97–110. doi: 10.1016/j.immuni.2013.06.012, PMID: 23871207

Raciti M, Granzotto M, Duc MD, Fimiani C, Cellot G, Cherubini E, Mallamaci A. 2013. Reprogramming fibroblasts to neural-precursor-like cells by structured overexpression of pallial patterning genes. *Molecular and Cellular Neuroscience* **57**:42–53. doi: 10.1016/j.mcn.2013.10.004, PMID: 24128663

Ragu C, Boukour S, Elain G, Wagner-Ballon O, Raslova H, Debili N, Olson EN, Daegelen D, Vainchenker W, Bernard OA, Penard-Lacronique V. 2010. The serum response factor (SRF)/megakaryocytic acute leukemia (MAL) network participates in megakaryocyte development. *Leukemia* **24**:1227–1230. doi: 10.1038/leu.2010.80, PMID: 20428204

Reinchisi G, Ijichi K, Glidden N, Jakovcevski I, Zecevic N. 2012. COUP-TFII expressing interneurons in human fetal forebrain. *Cerebral Cortex* **22**:2820–2830. doi: 10.1093/cercor/bhr359, PMID: 22178710

Reya T, Morrison SJ, Clarke MF, Weissman IL. 2001. Stem cells, cancer, and cancer stem cells. *Nature* **414**:105–111. doi: 10.1038/35102167, PMID: 11689955

Riccio O, van Gijn ME, Bezdek AC, Pellegrinet L, van Es JH, Zimber-Strobl U, Strobl LJ, Honjo T, Clevers H, Radtke F. 2008. Loss of intestinal crypt progenitor cells owing to inactivation of both Notch1 and Notch2 is accompanied by derepression of CDK inhibitors p27Kip1 and p57Kip2. *EMBO reports* **9**:377–383. doi: 10.1038/embor.2008.7, PMID: 18274550

Robert-Moreno A, Espinosa L, de la Pompa JL, Bigas A. 2005. RBPjkappa-dependent notch function regulates Gata2 and is essential for the formation of intra-embryonic hematopoietic cells. *Development* **132**:1117–1126. doi: 10.1242/dev.01660, PMID: 15689374

Ross SE, Greenberg ME, Stiles CD. 2003. Basic helix-loop-helix factors in cortical development. *Neuron* **39**:13–25. doi: 10.1016/S0896-6273(03)00365-9, PMID: 12848929

Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**:495–502. doi: 10.1038/nbt.3192, PMID: 25867923

Satoh Y, Yokota T, Sudo T, Kondo M, Lai A, Kincade PW, Kouro T, Iida R, Kokame K, Miyata T, Habuchi Y, Matsui K, Tanaka H, Matsumura I, Oritani K, Kohwi-Shigematsu T, Kanakura Y. 2013. The Satb1 protein directs hematopoietic stem cell differentiation toward lymphoid lineages. *Immunity* **38**:1105–1115. doi: 10.1016/j.immuni.2013.05.014, PMID: 23791645

Shahrezaei V, Swain PS. 2008. Analytical distributions for stochastic gene expression. *PNAS* **105**:17256–17261. doi: 10.1073/pnas.0803850105, PMID: 18988743

Shen J, Walsh CA. 2005. Targeted disruption of Tgif, the mouse ortholog of a human holoprosencephaly gene, does not result in holoprosencephaly in mice. *Molecular and Cellular Biology* **25**:3639–3647. doi: 10.1128/MCB.25.9.3639-3647.2005, PMID: 15831469

Shimojo H, Ohtsuka T, Kageyama R. 2011. Dynamic expression of notch signaling genes in neural stem/progenitor cells. *Frontiers in Neuroscience* **5**:78. doi: 10.3389/fnins.2011.00078, PMID: 21716644

Sillars-Hardebol AH, Carvalho B, Beliën JAM, de Wit M, Delis-van Diemen PM, Tijssen M, van de Wiel MA, Pontén F, Meijer GA, Fijneman RJA. 2012. CSE1L, DIDO1 and RBM39 in colorectal adenoma to carcinoma progression. *Cellular Oncology* **35**:293–300. doi: 10.1007/s13402-012-0088-2

Solar GP, Kerr WG, Zeigler FC, Hess D, Donahue C, de Sauvage FJ, Eaton DL. 1998. Role of c-mpl in early hematopoiesis. *Blood* **92**:4–10. PMID: 9639492

Spehlmann ME, Manthey CF, Dann SM, Hanson E, Sandhu SS, Liu LY, Abdelmalak FK, Diamanti MA, Retzlaff K, Scheller J, Rose-John S, Greten FR, Wang JYJ, Eckmann L. 2013. Trp53 deficiency protects against acute intestinal inflammation. *The Journal of Immunology* **191**:837–847. doi: 10.4049/jimmunol.1201716

Stolt CC, Lommes P, Sock E, Chaboissier MC, Schedl A, Wegner M. 2003. The Sox9 transcription factor determines glial fate choice in the developing spinal cord. *Genes & Development* **17**:1677–1689. doi: 10.1101/gad.259003, PMID: 12842915

Sulston JE, Schierenberg E, White JG, Thomson JN. 1983. The embryonic cell lineage of the nematode caenorhabditis elegans. *Developmental Biology* **100**:64–119. doi: 10.1016/0012-1606(83)90201-4, PMID: 6684600

Sur M, Rubenstein JL. 2005. Patterning and plasticity of the cerebral cortex. *Science* **310**:805–810. doi: 10.1126/science.1112070, PMID: 16272112

Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**:663–676. doi: 10.1016/j.cell.2006.07.024, PMID: 16904174

Tamura T, Nagamura-Inoue T, Shmeltzer Z, Kuwata T, Ozato K. 2000. ICSBP directs bipotential myeloid progenitor cells to differentiate into mature macrophages. *Immunity* **13**:155–165. doi: 10.1016/S1074-7613(00)00016-9, PMID: 10981959

Tan X, Zhang L, Zhu H, Qin J, Tian M, Dong C, Li H, Jin G. 2014. Brn4 and TH synergistically promote the differentiation of neural stem cells into dopaminergic neurons. *Neuroscience Letters* **571**:23–28. doi: 10.1016/j.neulet.2014.04.019, PMID: 24769320

Thomsen ER, Mich JK, Yao Z, Hodge RD, Doyle AM, Jang S, Shehata SI, Nelson AM, Shapovalova NV, Levi BP, Ramanathan S. 2016. Fixed single-cell transcriptomic characterization of human radial glial diversity. *Nature Methods* **13**:87–93. doi: 10.1038/nmeth.3629, PMID: 26524239

Thomson M, Liu SJ, Zou LN, Smith Z, Meissner A, Ramanathan S. 2011. Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. *Cell* **145**:875–889. doi: 10.1016/j.cell.2011.05.017, PMID: 21663792

Tibshirani R, Walther G, Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B* **63**:411–423. doi: 10.1111/1467-9868.00293

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* **58**:267–288 . doi: 10.1111/j.1467-9868.2011.00771.x

Tou L, Liu Q, Shivdasani RA. 2004. Regulation of mammalian epithelial differentiation and intestine development by class I histone deacetylases. *Molecular and Cellular Biology* **24**:3132–3139. doi: 10.1128/MCB.24.8.3132-3139.2004, PMID: 15060137

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32**:381–386. doi: 10.1038/nbt.2859, PMID: 24658644

Trapnell C. 2015. Defining cell types and states with single-cell genomics. *Genome research* **25**:1491–1498. doi: 10.1101/gr.190595.115, PMID: 26430159

Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR. 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**:371–375. doi: 10.1038/nature13173, PMID: 24739965

Tzeng SF, de Vellis J. 1998. Id1, Id2, and Id3 gene expression in neural cells during development. *Glia* **24**:372–381. doi: 10.1002/(SICI)1098-1136(199812)24:4<372::AID-GLIA2>3.0.CO;2-B, PMID: 9814817

Uittenbogaard M, Chiaramello A. 2002. Expression of the bHLH transcription factor Tcf12 (ME1) gene is linked to the expansion of precursor cell populations during neurogenesis. *Gene Expression Patterns* **1**:115–121. doi: 10.1016/S1567-133X(01)00022-9, PMID: 15018808

van der Flier LG, Clevers H. 2009. Stem cells, self-renewal, and differentiation in the intestinal epithelium. *Annual Review of Physiology* **71**:241–260. doi: 10.1146/annurev.physiol.010908.163145, PMID: 18808327

Van Der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**:2579–2605.

Van Der Maaten L. 2009. Learning a parametric embedding by preserving local structure. Twelfth International Conference on Artificial Intelligence and Statistics (AI-STATS) 384–391.

VanDussen KL, Samuelson LC. 2010. Mouse atonal homolog 1 directs intestinal progenitors to secretory cell rather than absorptive cell fate. *Developmental Biology* **346**:215–223. doi: 10.1016/j.ydbio.2010.07.026, PMID: 20691176

Varelas X. 2014. The hippo pathway effectors TAZ and YAP in development, homeostasis and disease. *Development* **141**. doi: 10.1242/dev.102376, PMID: 24715453

Vu TN, Wills QF, Kalari KR, Niu N, Wang L, Rantalainen M, Pawitan Y. 2016. Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32**:2128–2135. doi: 10.1093/bioinformatics/btw202, PMID: 27153638

Wainwright MJ. 2009. Sharp thresholds for High-Dimensional and noisy sparsity recovery using $\ell _{1}$ - Constrained Quadratic Programming (Lasso). *IEEE Transactions on Information Theory* **55**:2183–2202.

Wang M, Gu D, Du M, Xu Z, Zhang S, Zhu L, Lu J, Zhang R, Xing J, Miao X, Chu H, Hu Z, Yang L, Tang C, Pan L, Du H, Zhao J, Du J, Tong N, Sun J, et al. 2016. Common genetic variation in ETV6 is associated with colorectal cancer susceptibility. *Nature Communications* **7**:11478. doi: 10.1038/ncomms11478, PMID: 27145994

Wang TW, Stromberg GP, Whitney JT, Brower NW, Klymkowsky MW, Parent JM. 2006. Sox3 expression identifies neural progenitors in persistent neonatal and adult mouse forebrain germinative zones. *The Journal of Comparative Neurology* **497**:88–100. doi: 10.1002/cne.20984, PMID: 16680766

Wiegreffe C, Simon R, Peschkes K, Kling C, Strehle M, Cheng J, Srivatsa S, Liu P, Jenkins NA, Copeland NG, Tarabykin V, Britsch S. 2015. Bcl11a (Ctip1) controls migration of cortical projection neurons through regulation of sema3c. *Neuron* **87**:311–325. doi: 10.1016/j.neuron.2015.06.023, PMID: 26182416

Wills QF, Livak KJ, Tipping AJ, Enver T, Goldson AJ, Sexton DW, Holmes C. 2013. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature Biotechnology* **31**:748–752. doi: 10.1038/nbt.2642, PMID: 23873083

Witten DM, Tibshirani R. 2010. A framework for feature selection in clustering. *Journal of the American Statistical Association* **105**:713–726. doi: 10.1198/jasa.2010.tm09415, PMID: 20811510

Woodworth MB, Greig LC, Liu KX, Ippolito GC, Tucker HO, Macklis JD. 2016. Ctip1 regulates the balance between specification of distinct projection neuron subtypes in deep cortical layers. *Cell Reports* **15**:999–1012. doi: 10.1016/j.celrep.2016.03.064, PMID: 27117402

Wullaert A, Bonnet MC, Pasparakis M. 2011. NF-κB in the regulation of epithelial homeostasis and inflammation. *Cell Research* **21**:146–158. doi: 10.1038/cr.2010.175, PMID: 21151201

Yang Q, Liu S, Yin M, Yin Y, Zhou G, Zhou J. 2015. Ebf2 is required for development of dopamine neurons in the midbrain periaqueductal gray matter of mouse. *Developmental Neurobiology* **75**:1282–1294. doi: 10.1002/dneu.22284, PMID: 25762221

Yao Z, Mich JK, Ku S, Menon V, Krostag AR, Martinez RA, Furchtgott L, Mulholland H, Bort S, Fuqua MA, Gregor BW, Hodge RD, Jayabalu A, May RC, Melton S, Nelson AM, Ngo NK, Shapovalova NV, Shehata SI, Smith MW, et al. 2017. A Single-Cell roadmap of lineage bifurcation in human ESC models of embryonic brain development. *Cell Stem Cell* **20**:120–134. doi: 10.1016/j.stem.2016.09.011, PMID: 28094016

Ye DZ, Kaestner KH. 2009. Foxa1 and Foxa2 control the differentiation of goblet and enteroendocrine L- and D-cells in mice. *Gastroenterology* **137**:2052–2062. doi: 10.1053/j.gastro.2009.08.059, PMID: 19737569

Yin M, Liu S, Yin Y, Li S, Li Z, Wu X, Zhang B, Ang SL, Ding Y, Zhou J. 2009. Ventral mesencephalon-enriched genes that regulate the development of dopaminergic neurons in vivo. *Journal of Neuroscience* **29**:5170–5182. doi: 10.1523/JNEUROSCI.5569-08.2009, PMID: 19386913

Yu T, Chen X, Zhang W, Li J, Xu R, Wang TC, Ai W, Liu C, Ghaleb A, McConnell B. 2012. Krüppel-like factor 4 regulates intestinal epithelial cell morphology and polarity. *PLoS One* **7**:e32492. doi: 10.1371/journal.pone.0032492, PMID: 22384261

Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C, Rolny C, Castelo-Branco G, Hjerling-Leffler J, Linnarsson S. 2015. Brain structure. cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**:1138–1142. doi: 10.1126/science.aaa1934, PMID: 25700174

Zembrzycki A, Griesel G, Stoykova A, Mansouri A. 2007. Genetic interplay between the transcription factors Sp8 and Emx2 in the patterning of the forebrain. *Neural Development* **2**:8. doi: 10.1186/1749-8104-2-8, PMID: 17470284

Zhang P, Behre G, Pan J, Iwama A, Wara-Aswapati N, Radomska HS, Auron PE, Tenen DG, Sun Z. 1999. Negative cross-talk between hematopoietic regulators: GATA proteins repress PU.1. *PNAS* **96**:8705–8710. doi: 10.1073/pnas.96.15.8705, PMID: 10411939

Zhang Y, Miki T, Iwanaga T, Koseki Y, Okuno M, Sunaga Y, Ozaki N, Yano H, Koseki H, Seino S. 2002. Identification, tissue expression, and functional characterization of Otx3, a novel member of the Otx family. *Journal of Biological Chemistry* **277**:28065–28069. doi: 10.1074/jbc.C100767200, PMID: 12055180