# Untangling biological factors influencing trajectory inference from single cell data

Charrout, Mohammed; Reinders, Marcel J.T.; Mahfouz, Ahmed

**DOI**
10.1093/nargab/lqaa053

**Publication date**
2020

**Document Version**
Final published version

**Published in**
NAR Genomics and Bioinformatics

*Published online 22 July 2020*

*NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 3* **1**
*doi: 10.1093/nargab/lqaa053*

# Untangling biological factors influencing trajectory inference from single cell data

**Mohammed Charrout [1,2], Marcel J.T. Reinders[1,2] and Ahmed Mahfouz [1,2,3,*]**

[1]Delft Bioinformatics Lab, Delft University of Technology, Delft 2628 XE, The Netherlands, [2]Leiden Computational Biology Center, Leiden University Medical Center, Leiden 2333 ZC, The Netherlands and [3]Department of Human Genetics, Leiden University Medical Center, Leiden 2333 ZC, The Netherlands

Received April 14, 2020; Revised June 17, 2020; Editorial Decision June 30, 2020; Accepted July 07, 2020

## ABSTRACT

**Advances in single-cell RNA sequencing over the past decade has shifted the discussion of cell identity toward the transcriptional state of the cell. While the incredible resolution provided by single-cell RNA sequencing has led to great advances in unraveling tissue heterogeneity and inferring cell differentiation dynamics, it raises the question of which sources of variation are important for determining cellular identity. Here we show that confounding biological sources of variation, most notably the cell cycle, can distort the inference of differentiation trajectories. We show that by factorizing single cell data into distinct sources of variation, we can select a relevant set of factors that constitute the core regulators for trajectory inference, while filtering out confounding sources of variation (e.g. cell cycle) which can perturb the inferred trajectory. Script are available publicly on https://github.com/mochar/cell_variation.**

## INTRODUCTION

Single cell RNA-sequencing enables quantitative gene expression profiling of individual cells. From an RNA viewpoint, these cells live in a high-dimensional space defined by the expression of their genes. A critical step when analyzing such data is the identification of cells in order to find and label the cell types present in the data. This is often achieved by grouping together cells with similar expression profiles by applying a clustering method. The resulting cell clusters are thus separated from one another by a set of genes that vary in expression between the clusters. These so-called marker genes can then be used for identification by cross-referencing with known marker genes or marker genes found in other studies. This clustering-based approach for cell identification relies on the general presumption that the measured expression levels are reflective of the cell's identity, which may be violated due to shared transcriptional programs between two or more types. Large variations within cell type clusters due to many exclusive programs may also pose a problem as it can become hard to discern between cell types and cell states (1). More generally, sources of variation that contribute significantly to the cell–cell distances in gene space, yet do not reflect the cell type, can be detrimental to the identification task. These can vary from small transient changes e.g. cell communication, up to complex shifts in the cell's regulatory state such as the cell cycle, which has been reported to contribute a substantial portion of the gene expression variance (2). Moreover, cell identification is often preceded by a gene filtering step whereby genes with low variance are discarded to ease the computational burden in downstream analysis. Gene filtering can lead to a lower dimensional space that further amplifies unwanted variabilities, depending on the normalization and filtering criteria that is used. It becomes clear then that identifying and filtering out unwanted biological sources of variation can serve an important step in cell identification.

The immediate question to ask is how to identify what genes are necessary and what genes need to be left out when carrying out such analysis, which leads into a broader discussion of the definition of a cell type. Classically, cells were characterized using a combination of morphology, lineage, location and overall cell function. (1,3) However it has long been demonstrated that terminally differentiated cells can convert into other selected cell types by overexpression of key regulators (4–6). It has therefore been argued that cell types can be identified by the expression of a unique combination of transcription factors that make up the core regulatory complex, which is preserved along all states of the cell (1,7). Successfully identifying these core regulators allows one to differentiate between clusters of cell types and clusters of cell states, as the latter would share core regulators. Furthermore, focusing only on this stable set of differentially expressed genes relieves us from determining what transcriptional programs relate to the identity of a cell. In developmental systems, before complete cell maturity is reached, developing cells have been shown to undergo a series of discrete metastable states here referred to as dif-

*To whom correspondence should be addressed. Tel: +31 71 52 69513; Fax: +31 71 52 68285; Email: a.mahfouz@lumc.nl

ferentiation checkpoints (7–9). By relaxing the aforementioned definition of regulatory complexes to also include differentiation checkpoints, an analogous approach can be used to identify types when dealing with continuous cell transitions.

We focus on the problem of pseudotime inference where the aim is to order developing cells along a 'pseudotime' axis based on their transcriptional similarities. These similarities should therefore strictly reflect differences between cell types as they progress through the differentiation trajectory. The majority of pseudotime inference tools rely on the existence of a continuous manifold that reflects this trajectory such that a one-dimensional curve or graph can be fit (10). Confounding biological sources of variation (such as the cell cycle) can therefore perturb the inferred trajectory. We therefore hypothesize that by factorizing the matrix into distinct sources of variation, a relevant set of factors that constitute the core regulatory complexes can be selected for improving trajectory analysis.

## MATERIALS AND METHODS

### Data and preprocessing

Two single cell RNA-seq datasets with continuous cell type transitions were acquired from La Manno *et al*. (11): developing glutamatergic neurons in the human forebrain, which has a well-defined linear manifold and a complex branching dataset of the developing mouse hippocampus (Gene Expression Omnibus accession code GSE104323). Both datasets were available as raw unspliced and spliced count matrices with the corresponding gene and cell annotations. Normalization was done with Seurat 3.0's scTransform (12). In short, a regularized negative binomial model is fit on the data to model the counts as a function of total cell size. The Pearson residuals yielded by this fit can then be treated as normalized expression values, where a positive value indicates a higher count than expected, and *vice versa*. Seurat uses these residuals to return a count matrix that is unbiased by cell size, which was used for pseudotime inference. However, scHPF models the genes directly as negative binomial and was therefore passed the unnormalized spliced matrix instead. Filtering of low quality genes and cells was done in correspondence to the Jupyter notebooks provided by the authors (https://github.com/velocyto-team/velocyto-notebooks): In the human forebrain dataset, genes in the unspliced matrix with a total count $<25$ or a minimum cell count $<20$ were removed, and in the spliced matrix the thresholds were set at 30 and 20, respectively. The 1720 cells in the dataset were not filtered. In the mouse hippocampus dataset, the same gene threshold was set for the unspliced matrix while the thresholds were increased to 40 and 30, respectively for the spliced matrix. Cells with a total gene count lower than the 0.4 percentile were also filtered out. In addition, neuronal clusters identified by the source (Subiculum, CA1, CA3, CA2/4, Granule) were removed, leaving a total of 6673 out of 18 213 cells. Variant genes in both datasets were selected on the basis of residual variance provided by the model fit rather than the mean-variance association.

### Cell-cycle effect analysis

The branching mouse hippocampus dataset was truncated by removing the neuronal specification branches identified by the original paper, which included everything beyond the neuroblast clusters. A UMAP embedding was created with parameters $n\_neighbor = 30$, $min\_dist = 0.1$, and $metric = correlation$. scHPF was run with $K = 60$ (number of factors) and the resulting factors were annotated by Gene Ontology (GO) enrichment of the top scoring factor genes using the Enrichr web service (13). Factors with enriched GO terms within the cell cycle hierarchy were recognized as cell-cycle factors (Figure 2E). Two new UMAP embeddings were then calculated with the top scoring factor genes, one including and one excluding the cell-cycle factors. To quantify the agreement of both embeddings with the original, the distribution of Jaccard distances between the 500 nearest neighbors of each cell were calculated (Figure 2D). Results were validated using RNA velocity, a method by which the RNA splicing dynamics are used to extrapolate the expression profile at a future time point (11). The similarity of the extrapolated profile of a cell and that of the measured profiles of all other cells can be restated as a transition probability (Supplementary Note 1 of (11)). In summary, the transition probability between two cells $i$ and $j$ is found by calculating the Pearson correlation of the difference between the two cell profiles and the RNA velocity vector of cell $i$, which is then passed through an exponential kernel. Repeating this procedure for each pair of cells yields a matrix of similarity values which is then transformed into probability-like values by normalizing the rows to sum up to 1. These cell-to-cell transition probabilities were used to showcase the effect of cell cycle on the transition.

### Pseudotime linear dataset

The human glutamatergic neurogenesis dataset was factorized into 15 scHPF factors. Cell type and differentiation checkpoint factors were selected on the basis of known marker genes and enriched GO terms linked to neuronal differentiation such as *dendrite extension* and *cell morphogenesis in differentiation*. The top 10 factor genes were selected and subsequently passed to Ouija for pseudotime inference. However, kNN-smoothed expression values were passed instead of the log-transformed counts as this leads to a better model fit (Supplementary Figure S5). A principal curve was also fitted on the first four principal components following the aforementioned notebook provided by the data source. Pseudotimes of the two methods were then compared by simply calculating the differences in pseudotime assignment per cell. This revealed a disagreement in pseudotime assignment between the two methods. This disagreement was evaluated by correlating the downregulated genes with the pseudotimes under the basic presumption that these should be negatively correlated, i.e. downregulated genes decreases monotonically in expression as a function of pseudotime.

### Data Availability

Both datasets are acquired from La Manno *et al*. (11). The developing mouse hippocampus dataset is available

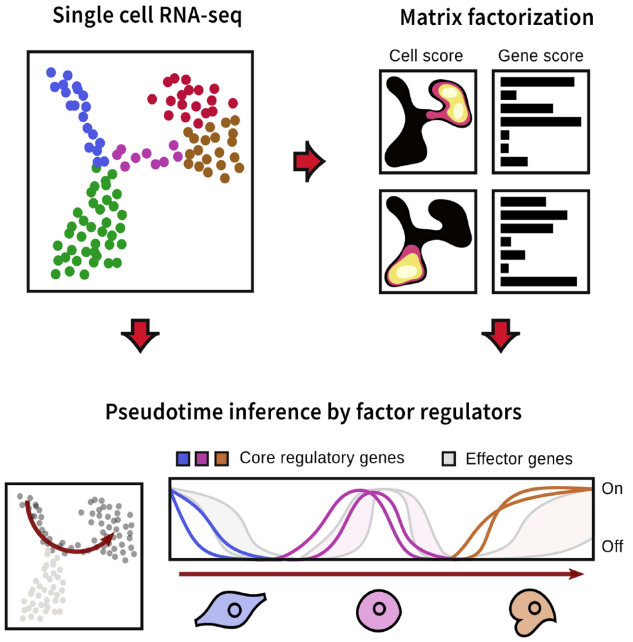


**Figure 1.** A graphical overview of the methodology. A single cell RNA-seq dataset of developing cells is first decomposed into a set of factors using matrix factorization. Each factor captures a different source of variation, a subset of them of which constitute the core regulatory genes driving the differentiation. A gene-based modeling approach is then used to find the unobstructed trajectory using only the regulatory genes of the differentiation factors.

in the Gene Expression Omnibus with accession code GSE104323. Analysis scripts can be found as a Snakemake pipeline in https://github.com/mochar/cell_variation.

## RESULTS

### Overview

A summary of our approach is visualized in Figure 1. First, the count matrix is factorized into a set amount of factors, each representing a transcriptional program. For this, scHPF (14) was used, which assigns for each factor a score per gene that quantifies the contribution of that gene to the factor. Similarly, cells are assigned a score based on how active the factor is in the cell. Scores concentrated on a select subpopulation of cells and a small subset of genes indicate specialized processes, whilst factors with a more uniform score distribution indicate broader processes active in many cells. The number of factors is selected based on the presumed number of major and intermediate cell types. However, by increasing this number of factors, the resolution can be increased from broad subpopulations to highly specific cell states. Next, the factors representing the cell types and differentiation checkpoints are selected. This is a manual process done by a combination of known marker genes, GO annotations, the number of highly variant genes and preservation of the factors across different runs. Next, the top 10 factor genes are passed to Ouija (9), a pseudotime inference method that models gene expression directly as either switch-like, where the gene is activated or repressed at some point during differentiation, or transient, when expression is only active for a short period of time. The direct modeling of a small set of marker genes allows for more interpretable inference and is therefore more in line with our biologically motivated approach. However, Ouija is limited to linear, non-branching data but can be used nonetheless by repeating the process for each sequence of progenitor to mature cell type factors.

### Cycling progenitor cells lead to spurious embeddings

As a first demonstration of our methodology, we study radial glia cells that are known progenitors of glia cells, the astrocytes and oligodendrocytes, as well as granule and pyramidal neurons in the developing hippocampus (15). When activated from their quiescent state, radial glia differentiate into neuronal intermediate progenitor cells (nIPC) and undergo continuous cell division during the development period, which is reflected strongly in the transcriptomic profile of these cells. To show to what extent this cell state can affect the analysis of the continuous embedding, the cells of the developing mouse hippocampus from La Manno *et al.* (11) were reanalyzed, focusing on the subset affected by the cell cycle by excluding the neuron specification branch. Figure 2A shows the UMAP embedding based on the top 3000 most variable genes, annotated by cell types as identified by the data source. The effect of the cell cycle was analyzed by first identifying a set of three factors associated with the cell cycle (Figure 2E), and subsequently reconstructing the embedding twice, with and without the cell-cycle factors (Figure 2B and C, respectively). A significant difference in the resulting embedding can be observed, which is further quantified by the jaccard distances of the cell neighbors visualized for both embeddings in Figure 2D. In all embeddings, the radial glia progenitor cells, as well as the astrocytes, neuroblasts and oligodendrocytes precursor cells (OPC) are clearly separated, with developing cells forming a bridge between all four clusters. Of note are the nIPC cells, which in the first two embeddings allude to being a differentiation checkpoint for the OPCs and neuroblasts. However, this observation disappears once the cell-cycle genes are removed, where instead the nIPC cells have a transcriptional profile that agrees with developing cells in both glia and neuronal lineages. This reveals that the nIPC cells do not form an intermediate cell type *per se*, but rather cluster together due to the significant transcriptional change attributed to the rapid cell division during development. Instead, a separate factor active at the right-hand site of the astrocyte cluster (Figure 2 C) suggests that there exists an intermediate checkpoint between the astrocytes and OPCs. This factor is characterized by the highly variable *Fabp7*, a regulator of both astrocytes as well as OPCs (16,17). Further effect of the cell cycle can be observed in a subset of the OPC 2 cells active in factors 34 and 48 (Figure 2E, top two embeddings). Indeed, oligodendrocytes in the developing brain have previously been shown to enter the cell cycle after reaching a more mature state (18). We note that the common approach of regressing out cell cycle genes can similarly remove the nIPC cluster, however at the cost of a less coherent embedding due to correlated signals (Supplementary Figure S1).

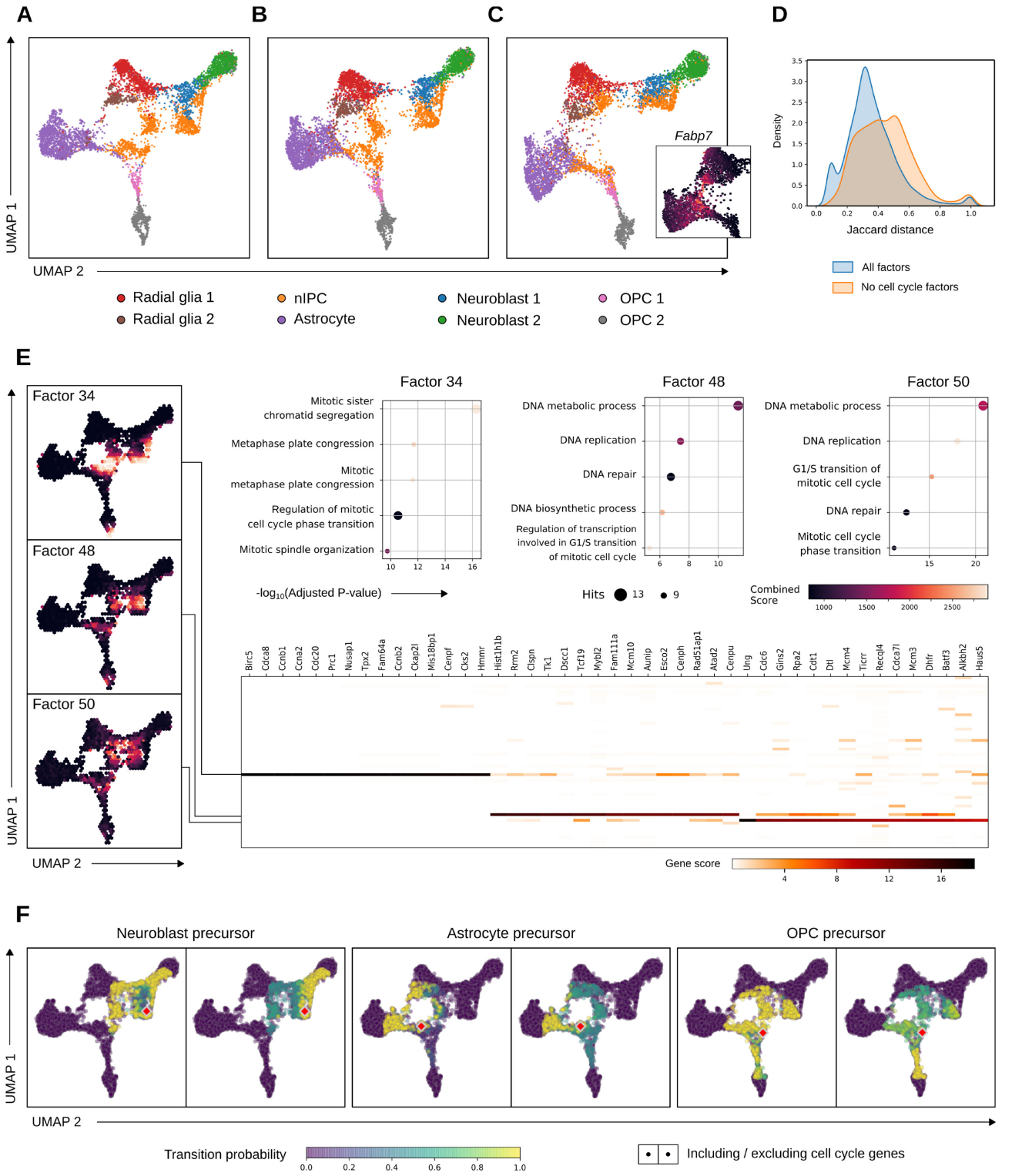


**Figure 2.** Cell cycle as a biological confounder in the developing mouse hippocampus. (**A**) UMAP embedding of developing mouse hippocampus cells annotated by the cell types identified by La Manno *et al.* (11). OPC, oligodendrocyte precursor cell; nIPC, neuronal intermediate progenitor cell. (**B** and **C**) The reconstructed embeddings using the top 50 genes of each factor including and excluding the cell-cycle factors identified by scHPF. (**D**) The distribution of Jaccard distances between the nodes of the embedding's knn-graph and that of the embedding in A. (**E**) Cell-cycle factors (34, 48, 50) found by scHPF identified by their GO annotations. The cell scores are shown in the embeddings on the left. The heatmap shows the gene scores of the top 10 genes of these factors. GO annotations are shown on top. (**F**) Transition probabilities including (left) and excluding (right) the cell-cycle genes for three selected cells highlighted in red. The latter shows a clearer transition affinity to one of three cell types (rows).

### RNA velocity is influenced by confounding factors

A scRNA-seq dataset is a static snapshot of the transcriptional state of a cell and therefore does not reveal the regulation of a gene, i.e. if it is currently up- or downregulated or in a steady state of transcription. However by comparing the fraction of spliced to unspliced counts, RNA velocity allows the quantification of a gene's regulatory state, effectively extracting a time component from the static snapshot (11). This information was used to further explore the effects of the cell-cycle dynamics by calculating the cell-to-cell transition probabilities with and without the previously identified cell-cycle genes. Figure 2F shows three examples of nIPC cells that show a strong transition probability to other nIPC cells when cell-cycle genes are included in the calculation, with no clear commitment to any one cell fate. This implies that the transition between the stages of the cell-cycle dominate, or at least contribute significantly to the aggregated velocities. However, when excluding the velocities of the cell cycle genes in the calculation of the cell transition probabilities, a much clearer picture can be established of the differentiation, as the probabilities becomes more concentrated at the different cell type clusters. This shows that the intermediate nIPC cells already have a clear commitment to a certain cell type, an observation that is missed when selecting genes exclusively on dominant signal rather than biological contribution as showcased before in Figure 2A. Previous results from the data source (11) further exemplify this, as no transition from radial glia to the nIPC cluster can be observed in the velocity field.

### Matrix factorization captures regulators of cell differentiation

The second dataset from La Manno *et al*. is of developing glutamatergic neurons in the human forebrain. It follows a linear path from the radial glia progenitor cells to the mature neurons through a sequence of differentiation checkpoints. These checkpoints were captured in six out of 15 factors generated with scHPF and are shown sequentially in Figure 3. For each of these factors, the cell scores on the embedding are shown in addition to the expression dynamics of the top 10 most contributing genes recovered by pseudotime inference using Ouija. Factor genes are largely preserved even with variable numbers of K (see Supplementary Figure S3). The first factor, factor 8, captures the radial glia cells identified by the highest scoring gene, the homeobox transcription factor *HOPX*, a well-known marker gene for radial glia (11,19). Factor 6 follows immediately after and is marked by an early deactivation of the *KLF5* gene which belongs to the Kruppel-like family of transcription factors. *KLF5* and other KLF genes are known repressors of neurite growth, and their downregulation are linked to cell-cycle arrest and neuronal development (20,21). The later suppression of Vimentin (*VIM*) in the same factor, which is a highly variable gene (Supplementary Figure S2) and known marker of gliogenesis, indicates that this might be a commitment point of the progenitor cells to either neuronal or astrocytic cell fates. Factor 7 has *EOMES* as the highest scoring gene which is a well-known transcription factor in early neuroblasts that regulates neurogenesis (11). Though no direct link can be found between *EPB41L4A* and neurogenesis, the early activated *SDC2* gene has a known role in regulating axon morphology in developing neurons in mice (22). This function is continued in factor 3 where GO terms such as *regulation of dendritic spine morphogenesis* and *regulation of cell morphogenesis involved in differentiation* are highly enriched. Of note in this factor is *LAMC1*, a target gene of miR-124, a microRNA that has an important regulatory function in differentiation and maturation of neuroblasts (23,24). The excitatory effects of glutamatergic neurons are already specified by the *GPR39* gene in factor 3 and is further regulated by the two synaptotagmins *SYT4* and *SYT1* (Supplementary Figure S2) in factor 1. The activity of synaptotagmins are indicative of synaptic integration, the final stage of neurogenesis (23,23,25–26). The highly variable *LMO3* gene in the final factor is a co-factor that physically interacts with other regulatory proteins to form transcriptional regulators in the developing brain (27,28). Together these results show that genes captured with matrix factorization have important regulatory roles in the timing of neuronal differentiation.

### Pseudotime curve fitting ignored biological confounds

We compared the pseudotime assignment obtained using Ouija on the factor genes (shown in Figure 3), with the pseudotimes identified by the data source, where a principle curve was fit on the first four principal components. Figure 4A shows the fitted principal curve while Figure 4B shows the alignment of the cells based on the two pseudotime assigned by the two methods. A disagreement is visible in the first half of the trajectory with a delay in the radial glia cluster (blue) by Ouija. This delay is propagated until the midpoint is reached which corresponds to the start of factor 3. We hypothesize the cause of the delay to be due to confounding variability within the radial glia cluster, most likely related to cell-cycle effects. Indeed, GO enrichment on factor 13, which is active in the delayed first half, captures cell cycle-related activities (Figure 4 and Supplementary Figure S4). As the true labels are unknown, we resort to working on the basic assumption that early downregulated genes (e.g. those found in factors 8 and 6) are to be negatively correlated with pseudotime. A higher correlation value would therefore indicate a better ordering of cells along the pseudotime axis. Figure 4C shows the Pearson correlations that support the idea of a necessary delay within the radial glia cluster as captured by our approach. Scatterplot of the downregulated genes as a function of both pseudotime assignments are shown in Supplementary Figure S7.

### Matrix factorization decouples cell type and cell states

Cells within each cell type cluster can be further divided into cell states. One such example are the radial glia, which differentiate to produce both neurons and astrocytes, but have been shown to switch their preference from the former to the latter at a later stage (29–31). Another example of different cell states within the same cell type cluster are the astrocytes themselves. Among other differences, astrocytes are long known to specialize as GFAP-positive protoplasmic and GFAP-negative fibrous astrocytes (29,32–33). The

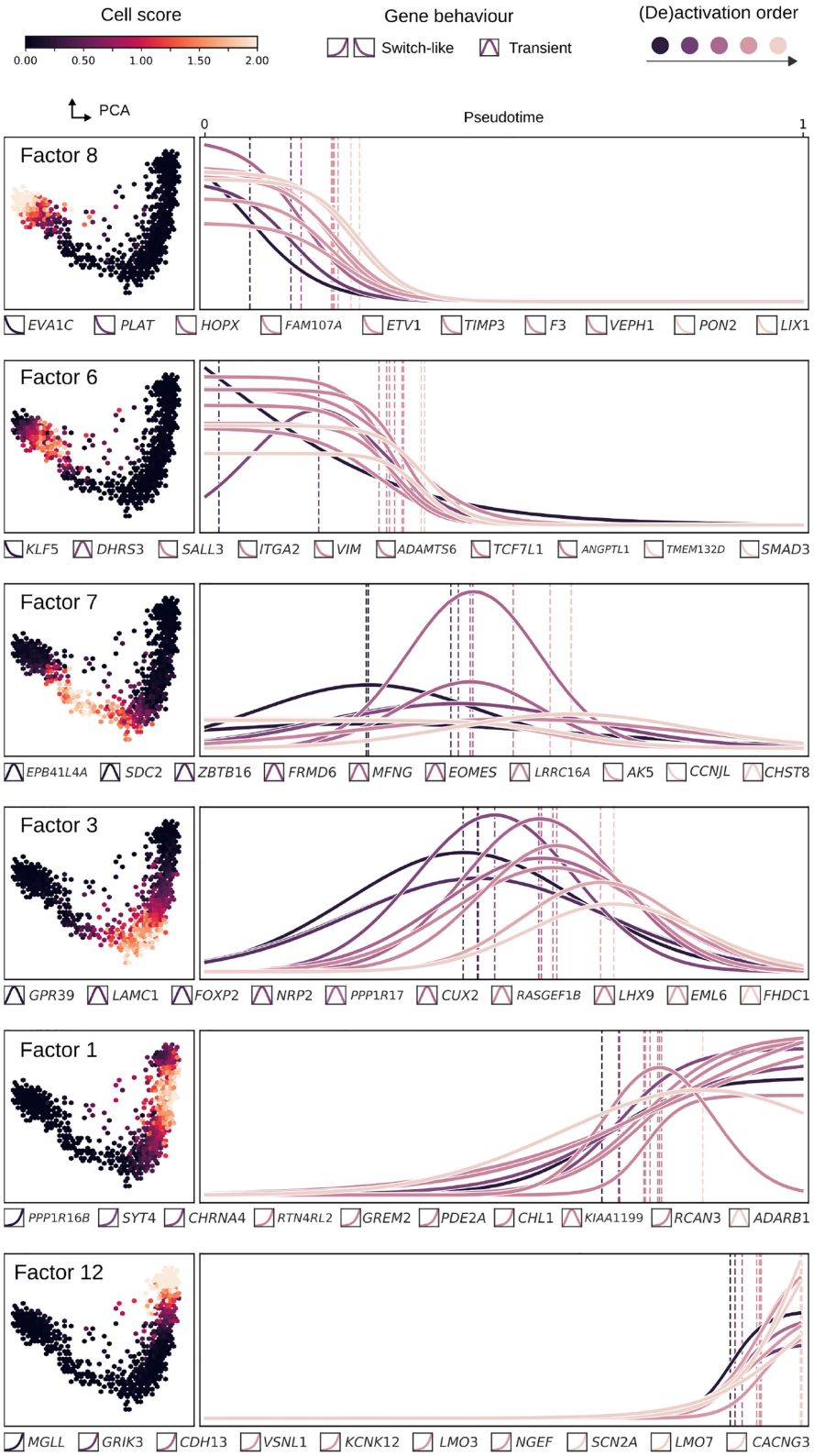


**Figure 3.** Gene expression dynamics of cell type and differentiation checkpoint factors. Each row shows for each factor the cell scores plotted on the first two principal components. On the right are the Ouija-fitted curves of the top 10 factor genes. Each vertical line corresponds to the activation/deactivation time for genes with switch-like behavior, and peak time for transiently activated genes. Curve and line color indicate order of activation and deactivation. Gene names with corresponding expression behavior type is shown below each plot.

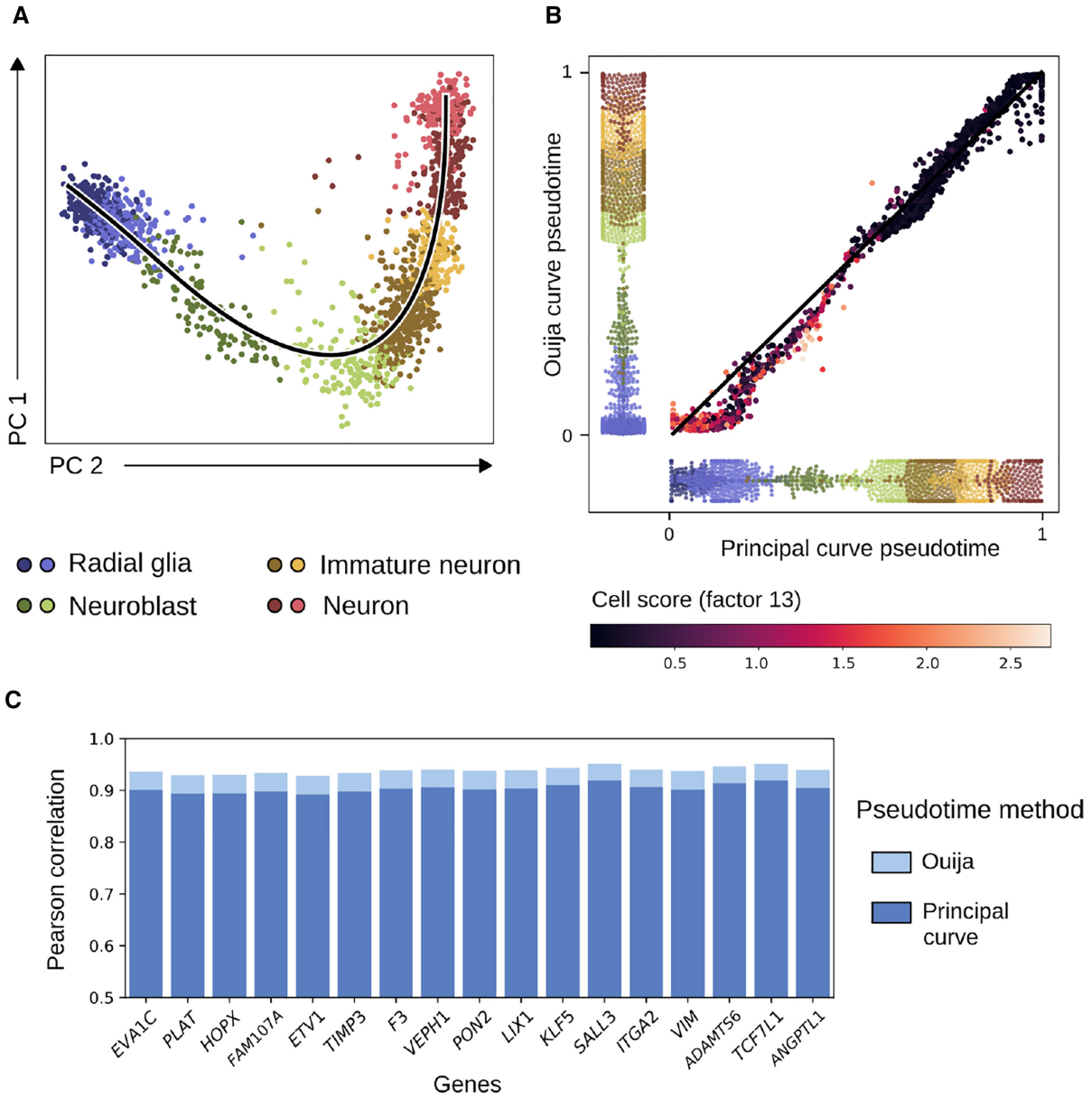


**Figure 4.** (**A**) Cells of developing human forebrain projected on the first two principal components. Black line indicates the fitted principal curve. Colors indicate clusters. (**B**) Cells plotted based on the pseudotime of the principal curve (*x*-axis) and Ouija (*y*-axis). Cells are colored by the cell scores of the cell cycle factor (factor 13, Supplementary Figures S2 and 4). Shown behind each axis are the distribution of the cells on the basis of the pseudotime assignments, colored according to their cluster. (**C**) Pearson correlation between the normalized expression values and the pseudotimes of the downregulated genes of factors 8 and 6 (see Figure 3).

factor shown in Figure 5 identified such a subpopulation within the astrocyte cluster with a high specificity for GFAP, which might reflect the protoplasmic astrocyte state. More examples are found in the second dataset where multiple factors are found that overlap with the cell type and checkpoint factors identified in Figure 3 (Supplementary Figure S2).

## DISCUSSION

We have made an attempt at understanding the complex process of cell differentiation by directly modeling regulators of selected transcriptional programs presumed to be cell type specific. This approach tackles the question of what biological sources of variation are relevant to developing cells. We followed a 'bottom-up' approach where a small subset of core marker genes is utilized to represent cell identity. Alternatively, one can follow a top down approach such that all biological processes active in the cells are annotated and subsequently stripped away based on their perceived relevance. An example of this is of Buettner *et al.* (2), where the same problem formulation of confounding cell states was tackled. There, a latent variable model was developed that factorizes the expression matrix into a set of

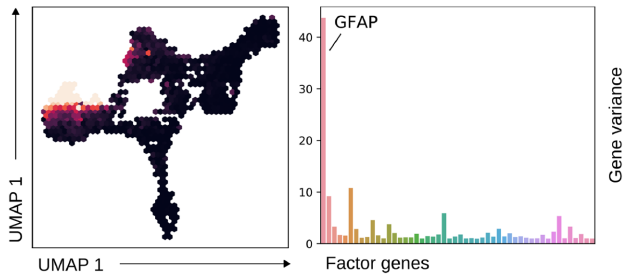


**Figure 5.** Factor active in a subset of astrocyte cells (left). The variance of the top genes contributing to the factor are shown on the right. The high specificity of GFAP in the factor alludes to it being the protoplasmic subtype.

factors determined by a database of pathways and another set of unannotated factors. However, their analysis was limited to removing technical rather than biological confounding sources of variation. Nonetheless we argue that while both approaches require a fair amount of domain knowledge and manual interpretation of gene sets, a bottom-up approach alleviates this by a large margin while still adhering to a concrete definition of cell identity.

### Dimensionality reduction in pseudotime inference

The end goal of our methodology is to improve pseudotime inference rather than to develop a new algorithm, and to shift the discussion toward the use of biology motivated tooling and interpretation. This is in response to most tools developed for reconstructing trajectories from single cell data, namely that they adhere to the pipeline of dimensionality reduction followed by trajectory modeling, by either fitting a curve in the resulting embedding or by finding a path in the embedding's neighbor graph (10,34). While it is known that dimensionality reduction leads to a loss of interpretability, other relevant concerns can be raised as well. First, the issue of the frequently used principal component analysis (PCA) method in count based data such as scRNA-seq was recently addressed by (among others) Townes *et al.* (35), where the authors show that there exists an implicit assumption of normality of the features. The authors show that this results in distorted components as scRNA-seq data violates this assumption. Second, non-linear dimensionality reduction methods such as t-SNE and UMAP can exaggerate the distances between cell clusters with large transcriptional differences, which can lead to disconnected embeddings (36). This problem can be amplified with stringent variance-based gene filtering as developing cells are identified by more subtle differences in gene expression. Finally, it has been shown that complex high-dimensional structures found in scRNA-seq data cannot be fully preserved in a small number of dimensions, and can therefore miss important variation or lead to distorted embeddings (37). Our methodology is unaffected by these issues as dimensionality reduction is not a prerequisite for matrix factorization and gene expression modeling. However it must be noted that the use of gene expression smoothing used in the RNA velocity pipeline, which was subsequently passed to Ouija for gene modeling, does rely on a PCA step, and may therefore be affected by possible distortions mentioned before. This is no shortcoming of Ouija as no smoothing is required per se, however we did find the model fit to improve with smoothing (Supplementary Figures S5 and 6).

### Cell type and cell state

Of specific interest in this study is the definition of a cell type as a summation of its core regulators and its different states. Trajectory modeling is affected by this as cell distances can reflect cell state instead of cell types. We argue that reducing the genes to a small set of core regulators that preserve the identity of the cells, we are able to circumvent the problem of confounding cell states. Another consequence of decoupling cell type from state is that a shared state between different cell types can lead to high similarity between cells of different identities. This is exemplified in the first analysis shown in Figure 2 where it was found that the existence of the nIPC cluster is a direct effect of the cell cycle on differentiating cells. A more subtle example is regarding the substantial amount of functional similarities between radial glia and astrocytes, hence their combined classification as neuroglia (33). The expression of the GFAP protein in both cell types is an example of their similarities, which can explain the close proximity observed in Figure 2B between the radial glia and the aforementioned protoplasmic astrocyte subpopulation.

### Regulator genes for cell type identification

We have motivated and shown that the use of cell type regulators for pseudotime inference can be a useful alternative to curve-based fitting due to their biological relevancy. However not all transcription factors are annotated or even identified (38). This problem is circumvented in this study by taking the top factor genes, and although many have important regulatory functions as shown in the results, it is not obvious if they are directly coupled to the cell types and differentiation checkpoints. Further validation must be needed to confirm such associations. For example, knockout experiments of the putative transcription factors can be used to determine their necessity as important regulators in cell differentiation. Similarly, one can overexpress the putative regulator in pluripotent stem cells and observe any structural or regulatory similarities with the cell type under study. One might argue that as long as the genes used for pseudotime inference are faithful proxies of the differentiation process, the use of effector genes rather than regulatory genes might not influence the resulting pseudotimes significantly. A downside to this however is that the genes cannot be validated or utilized in other similar studies which hurts the interoperability and reusability of the study. Furthermore, the genes may no longer provide insight into what processes regulate cell fate decisions. Effector genes are also much larger in quantity than their regulators (1), which can lead to an uneven distribution of signals across the different cell types and checkpoints and result in distorted pseudotimes. Even when restricting the number of genes to circumvent this, one must also keep in mind that there exists a delay in activity between transcription factors and effector genes. Namely, when a pathway

used during differentiation is activated in developing cells, there exists a delay between when its regulators respond to the activation signals, and the actual transcription and activity of the effector genes. This means that the snapshot provided by scRNA-seq of the transcript counts fluctate in time, which has led to the utilization of time alignment algorithms such as Dynamic Time Warping, used predominantly in the field of metabolomics (39). Current computational solutions might be the use of regulatory network inference algorithms that have emerged in quantity due to the granularity provided by scRNA-seq data (40). Another approach is provided by algorithms that predict physical interactions between proteins, as increasing evidence shows that core regulators form physical interactions (1), exemplified by the co-factor LMO3 found in factor 12 in the developing mouse forebrain. Supplementary Figure S8 in the supplementary shows how the STRING service (41) is used to find many interactions between top factor genes in the last stages of glutamatergic neurogenesis.

The incredible resolution provided by single cell RNA-sequencing data raises the question of what sources of variation within it are important for the study at hand. Here we have focused on the problem of cell identification, more specifically in developmental systems using pseudotime inference. We have argued and shown that confounding sources of variation, most notably the cell cycle, can distort inference of the differentiation trajectory. We have then shown that this problem can be circumvented by limiting the scope to a select subset of genes assumed to play a regulatory role in cell development and directly modeling their expression.

## SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

## FUNDING

This project has received funding from: (1) The European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 861190 (PAVE); (2) The collaboration project TIMID (LSHM18057-SGF), itself financed by the PPP allowance, made available by Top Sector Life Sciences & Health (LSH) to Samenwekende Gezondheidsfondsen (SGF) to stimulate public-private partnerships and co-financing by health foundations that are part of the SGF; (3) The NWO Gravity program BRAINSCAPES; (4) The European Commission of a H2020 MSCA award under proposal number [675743] (ISPIC).

*Conflict of interest statement.* None declared.

## REFERENCES

1. Arendt,D., Musser,J.M., Baker,C.V.H., Bergman,A., Cepko,C., Erwin,D.H., Pavlicev,M., Schlosser,G., Widder,S., Laubichler,M.D. *et al.* (2016) The origin and evolution of cell types. *Nat. Rev. Genet.*, **17**, 744–757.
2. Buettner,F., Natarajan,K.N., Casale,F.P., Proserpio,V., Scialdone,A., Theis,F.J., Teichmann,S.A., Marioni,J.C. and Stegle,O. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155–160.
3. Nguyen,D.H., Jaszczak,R.G. and Laird,D.J. (2019) Heterogeneity of primordial germ cells. In: Lehmann,R. (ed). *The Immortal Germline*. Academic Press, Cambridge, Massachusetts. Vol. **135**, pp. 155–201.
4. Davis,R.L., Weintraub,H. and Lassar,A.B. (1987) Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell*, **51**, 987–1000.
5. Graf,T. and Enver,T. (2009) Forcing cells to change lineages. *Nature*, **462**, 587–594.
6. Takahashi,K. and Yamanaka,S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**, 663–676.
7. Xia,B. and Yanai,I. (2019) A periodic table of cell types. *Development*, **146**, dev169854.
8. McFaline-Figueroa,J.L., Hill,A.J., Qiu,X., Jackson,D., Shendure,J. and Trapnell,C. (2019) A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nat. Genet.*, **51**, 1389–1398.
9. Campbell,K.R. and Yau,C. (2019) A descriptive marker gene approach to single-cell pseudotime inference. *Bioinformatics*, **35**, 28–35.
10. Saelens,W., Cannoodt,R., Todorov,H. and Saeys,Y.(2019). A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *Nature Biotechnoly*,**37**, 547–554.
11. La Manno,G., Soldatov,R., Zeisel,A., Braun,E., Hochgerner,H., Petukhov,V., Lidschreiber,K., Kastriti,M.E., Lönnerberg,P., Furlan,A. *et al.* (2018) RNA velocity of single cells. *Nature*, **560**, 494–498.
12. Hafemeister,C. and Satija,R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, **20**, 296.
13. Kuleshov,M.V., Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z., Koplev,S., Jenkins,S.L., Jagodnik,K.M., Lachmann,A. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.
14. Levitin,H.M., Yuan,J., Cheng,Y.L., Ruiz,F.J., Bush,E.C., Bruce,J.N., Canoll,P., Iavarone,A., Lasorella,A., Blei,D.M. *et al.* (2019) De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization. *Mol. Syst. Biol.*, **15**, e8557.
15. Malatesta,P., Hack,M.A., Hartfuss,E., Kettenmann,H., Klinkert,W., Kirchhoff,F. and Gotz,M. (2003) Neuronal or Glial Progeny: regional differences in radial glia fate. *Neuron*, **37**, 751–764.
16. Sharifi,K., Morihiro,Y., Maekawa,M., Yasumoto,Y., Hoshi,H., Adachi,Y., Sawada,T., Tokuda,N., Kondo,H., Yoshikawa,T. *et al.* (2011) FABP7 expression in normal and stab-injured brain cortex and its role in astrocyte proliferation. *Histochem. Cell Biol.*, **136**, 501–513.
17. Sharifi,K., Ebrahimi,M., Kagawa,Y., Islam,A., Tuerxun,T., Yasumoto,Y., Hara,T., Yamamoto,Y., Miyazaki,H., Tokuda,N. *et al.* (2013) Differential expression and regulatory roles of FABP5 and FABP7 in oligodendrocyte lineage cells. *Cell Tissue Res.*, **354**, 683–695.
18. Goldman,S.A. and Kuypers,N.J. (2015) How to make an oligodendrocyte. *Development*, **142**, 3983–3995.
19. Hochgerner,H., Zeisel,A., Lönnerberg,P. and Linnarsson,S. (2018) Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nat. Neurosci.*, **21**, 290–299.
20. Qin,S. and Zhang,C.-L. (2012) Role of Krüppel-like factor 4 in neurogenesis and radial neuronal migration in the developing cerebral cortex. *Mol. Cell. Biol.*, **32**, 4297–4305.
21. Moore,D.L., Apara,A. and Goldberg,J.L. (2011) Kruppel-like transcription factors in the nervous system: novel players in neurite outgrowth and axon regeneration. *Mol. Cell. Neurosci.*, **47**, 233–243.
22. Barak,B., Feldman,N. and Okun,E. (2014) Toll-like receptors as developmental tools that regulate neurogenesis during development: an update. *Front. Neurosci.*, **8**, 272.
23. Bielefeld,P., Mooney,C., Henshall,D.C. and Fitzsimons,C.P. (2017) miRNA-mediated regulation of adult hippocampal neurogenesis; implications for epilepsy. *Brain Plast.*, **3**, 43–59.
24. Lang,M.-F. and Shi,Y. (2012) Dynamic roles of microRNAs in neurogenesis. *Front. Neurosci.*, **6**, 71.
25. Tocco,G., Bi,X., Vician,L., Lim,I.K., Herschman,H. and Baudry,M. (1996) Two synaptotagmin genes, Syt1 and Syt4, are differentially

regulated in adult brain and during postnatal development following kainic acid-induced seizures. *Mol. Brain Res.*, **40**, 229–239.
26. Ullrich,B., Li,C., Zhang,J.Z., McMahon,H., Anderson,R. G.W., Geppert,M. and Südhof,T.C. (1994) Functional properties of multiple synaptotagmins in brain. *Neuron*, **13**, 1281–1291.
27. Abellán,A., Desfilis,E. and Medina,L. (2014) Combinatorial expression of Lef1, Lhx2, Lhx5, Lhx9, Lmo3, Lmo4, and Prox1 helps to identify comparable subdivisions in the developing hippocampal formation of mouse and chicken. *Front. Neuroanat.*, **8**, 59.
28. Sang,M., Ma,L., Sang,M., Zhou,X., Gao,W. and Geng,C. (2014) LIM-domain-only proteins: multifunctional nuclear transcription coregulators that interacts with diverse proteins. *Mol. Biol. Rep.*, **41**, 1067–1073.
29. Tabata,H. (2015) Diverse subtypes of astrocytes and their development during corticogenesis. *Front. Neurosci.-Switz.*, **9**, 114.
30. Molofsky,A.V. and Deneen,B. (2015) Astrocyte development: a guide for the perplexed. *Glia*, **63**, 1320–1329.
31. Beattie,R. and Hippenmeyer,S. (2017) Mechanisms of radial glia progenitor cell lineage progression. *FEBS Lett.*, **591**, 3993–4008.
32. Bayraktar,O.A., Fuentealba,L.C., Alvarez-Buylla,A. and Rowitch,D.H. (2015) Astrocyte development and heterogeneity. *CSH Perspect. Biol.*, **7**, a020362.
33. Mori,T., Buffo,A. and Götz,M. (2005). The novel roles of glial cells revisited: The contribution of radial glia and astrocytes to neurogenesis. In: Schatten,G.P. (ed). *Current Topics in Developmental Biology*. Amsterdam, Elsevier. Vol. **69**, pp. 67–99.
34. Cannoodt,R., Saelens,W. and Saeys,Y. (2016) Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.*, **46**, 2496–2506.
35. Townes,F.W., Hicks,S.C., Aryee,M.J. and Irizarry,R.A. (2019) Feature Selection and Dimension Reduction for Single Cell RNA-Seq based on a Multinomial Model. *Genome Biology*, **20**, 295.
36. Konstorum,A., Jekel,N., Vidal,E. and Laubenbacher,R. (2018) Comparative Analysis of Linear and Nonlinear Dimension Reduction Techniques on Mass Cytometry Data. bioRxiv doi: https://doi.org/10.1101/273862, 01 March 2018, preprint: not peer reviewed.
37. Cooley,S.M., Hamilton,T., Deeds,E.J. and Ray,J. C.J. (2019) A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-Seq data. bioRxiv doi: https://doi.org/10.1101/689851, 02 July, 2019, preprint: not peer reviewed.
38. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
39. Christin,C., Hoefsloot,H. C.J., Smilde,A.K., Suits,F., Bischoff,R. and Horvatovich,P.L. (2010) Time alignment algorithms based on selected mass traces for complex LC-MS data. *J. Proteome Res.*, **9**, 1483–1495.
40. Todorov,H., Cannoodt,R., Saelens,W. and Saeys,Y. (2019) Network inference from single-cell transcriptomic data. *Methods Mol. Biol.*, **1883**, 235–249.
41. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.