Published in final edited form as:
Nature. 2018 August ; 560(7719): 494–498. doi:10.1038/s41586-018-0414-6.

RNA velocity of single cells

Gioele La Manno[1,2], **Ruslan Soldatov**[3], **Amit Zeisel**[1,2], **Emelie Braun**[1,2], **Hannah Hochgerner**[1,2], **Viktor Petukhov**[3,4], **Katja Lidschreiber**[5], **Maria E. Kastriti**[6], **Peter Lönnerberg**[1,2], **Alessandro Furlan**[1], **Jean Fan**[3], **Lars E. Borm**[1,2], **Zehua Liu**[3], **David van Bruggen**[1], **Jimin Guo**[3], **Xiaoling He**[7], **Roger Barker**[7], **Erik Sundström**[8], **Gonçalo Castelo-Branco**[1], **Patrick Cramer**[5,9], **Igor Adameyko**[6], **Sten Linnarsson**[1,2,†], and **Peter V. Kharchenko**[3,10,†]

[1]Laboratory of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet,171 77 Stockholm, Sweden

[2]Science for Life Laboratory, 171 21 Solna, Sweden

[3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02446, USA

[4]Department of Applied Mathematics, Peter The Great St. Petersburg Polytechnic University, St. Petersburg, Russia

[5]Department of Biosciences and Nutrition, Karolinska Institutet, 171 77 Stockholm, Sweden

[6]Department of Physiology and Pharmacology, Karolinska Institutet, 171 77, Stockholm, Sweden

[†]Correspondence and requests for materials should be addressed to peter_kharchenko@hms.harvard.edu and sten.linnarsson@ki.se.

Ethical compliance. All experimental procedures followed the guidelines and recommendations of Swedish animal protection legislation and were approved by the local ethical committee for experiments on laboratory animals (N68/14, Stockholms Norra Djurförsöksetiska nämnd, Sweden).

Data availability. All sequencing data are deposited in the Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA). Human postconception week 10 forebrain dataset is available on SRA under accession code SRP129388. The data from Mouse P0 and P5 hippocampus is extracted from Dataset C of 22 (GEO; GSE104323). The Oligodendrocyte differentiation dataset was obtained from a recently published survey of the mouse nervous system20 (SRA; SRP135960). Metabolic labelling data is available in GEO under the accession code (GEO; GSE115813). The bulk RNA-seq data on the mouse liver circadian variation was taken from12 (SRA; SRA025656). The SMART-seq2 data on the chromaffin cell differentiation was taken from13 (GEO; GSE99933). Data on the mouse bone marrow dataset is described in 19 (GEO; GSE109989). The Visual cortex inDrop datatset is described in 21 (GEO; GSE102827.). The Intestinal epithelium dataset is described in 23 (GEO; GSE92332). All other data are available from the corresponding author upon reasonable request.

Code availability. The software described in this paper, in the form of a pipeline called Velocyto (velox, quick; κύτος, cell) is available at http://velocyto.org. This includes complete analysis libraries in R and Python, as well as R and Python notebooks.

Author Contributions

S.L. conceived of the concept of RNA velocity and P.V.K. showed that RNA velocity could be detected through analysis of unspliced transcripts in single cells. P.V.K. and S.L. designed and supervised the study. P.V.K., S.L., G.L.M. and R.S. developed the analytical framework, analyzed data, made figures and drafted the manuscript, with contributions from all co-authors. P.V.K., G.L.M., R.S. and P.L. implemented the software, with assistance from V.P. and J.F. Z.L. examined RNA degradation signals. A.Z. and H.H. performed the mouse hippocampus experiment. P.C. supervised and K.L. and H.H. performed metabolic labeling. M.E.K. and I.A. have performed validations of chromaffin differentiation rate. E.B. and L.E.B performed and analyzed the fluorescent in situ hybridization experiment on tissue dissected by X.H. E.S. and R.B. provided human embryonic brain tissue. D.B. performed the human forebrain single-cell RNA-seq experiment, under supervision of G.C.B. J.G. assisted with measurement and interpretation of mouse bone marrow. The paper was read and approved by all co-authors.

Author Information

Reprints and permissions information is available at www.nature.com/reprints

The authors declare no competing interests.

[7]John van Geest Centre for Brain Repair, Department of Clinical Neurosciences, University of Cambridge, Cambridge CB2 0PY, UK

[8]Division of Neurodegeneration, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, 171 77 Stockholm, Sweden

[9]Max Planck Institute for Biophysical Chemistry, Department of Molecular Biology, Am Fassberg 11, 37077 Göttingen, Germany

[10]Harvard Stem Cell Institute, Cambridge, MA 02315, USA

Abstract

RNA abundance is a powerful indicator of the state of individual cells. Single-cell RNA sequencing can reveal RNA abundance with high quantitative accuracy, sensitivity and throughput1. However, this approach captures only a static snapshot at a point in time, posing a challenge for the analysis of time-resolved phenomena, such as embryogenesis or tissue regeneration. Here we show that RNA velocity—the time derivative of the gene expression state—can be directly estimated by distinguishing unspliced and spliced mRNAs in common single-cell RNA sequencing protocols. RNA velocity is a high-dimensional vector that predicts the future state of individual cells on a timescale of hours. We validate its accuracy in the neural crest lineage, demonstrate its use on multiple published datasets and technical platforms, reveal the branching lineage tree of the developing mouse hippocampus, and examine the kinetics of transcription in human embryonic brain. We expect RNA velocity to greatly aid the analysis of developmental lineages and cellular dynamics, particularly in humans.

During development, differentiation occurs on a time scale of hours to days, which is comparable to the typical half-life of mRNA. The relative abundance of nascent (unspliced) and mature (spliced) mRNA can be exploited to estimate the rates of gene splicing and degradation, without the need for metabolic labelling, as previously shown in bulk2–4. We reasoned similar signals may be detectable in single-cell RNA-seq data, and could reveal the rate and direction of change of the entire transcriptome during dynamic processes.

All common single-cell RNA-seq protocols rely on oligo-dT primers to enrich for polyadenylated mRNA molecules. Nevertheless, examining single-cell RNA-seq datasets based on the SMART-seq2, STRT/C1, inDrop, and 10x Chromium protocols5–8, we found that 15-25% of reads contained unspliced intronic sequences (Fig. 1a), in agreement with previous observations in bulk4 (14.6%) and single-cell5 (~20%) RNA sequencing. Most such reads originated from secondary priming positions within the intronic regions (Extended Data Fig. 1). In 10x Genomics Chromium libraries, we also found abundant discordant priming from the more commonly occurring intronic polyT sequences (Extended Data Fig. 1), which may have been generated during PCR amplification by priming on the first-strand cDNA. The substantial number of intronic molecules and their correlation with the exonic counts suggest that these molecules represent unspliced precursor mRNAs. This was confirmed by metabolic labeling of newly transcribed RNA9 followed by RNA sequencing using oligo-dT-primed STRT10 (Extended Data Fig. 2); 83% of all genes showed expression time courses consistent with simple first-order kinetics, as expected if unspliced reads represented nascent mRNA.

To quantify the time-dependent relationship between precursor and mature mRNA abundances, we assumed a simple model for transcriptional dynamics2, where the first time derivative of the spliced mRNA abundance (RNA velocity) is determined by the balance between production of spliced mRNA from unspliced mRNA, and the mRNA degradation (Fig. 1b, Supplementary Note 1). Under such a model, steady states are approached asymptotically when the rate of transcription α is constant, with the steady-state abundances of spliced (s) and unspliced (u) molecules determined by α, and constrained to a fixed-slope relationship where $u = \gamma s$ (Supplementary Note 2 Section 1). The equilibrium slope γ combines degradation and splicing rates, capturing gene-specific regulatory properties, the ratio of intronic and exonic lengths, and the number of internal priming sites. Examining a recently published compendium of mouse tissues11, steady-state behavior of most genes across a wide range of cell types was consistent with a single fixed slope γ (Extended Data Fig. 3a-c). However, 11% of genes showed distinct slopes in different subsets of tissues (Extended Data Fig. 3d-e), suggesting tissue-specific alternative splicing (Extended Data Fig. 3f) or degradation rates.

During a dynamic process, an increase in the transcription rate α results in a rapid increase of unspliced mRNA, followed by a subsequent increase of spliced mRNA (Fig. 1c and Supplementary Note 2 Section 1) until a new steady state is reached. Conversely, a drop in the rate of transcription first leads to a rapid drop in unspliced mRNA, followed by reduction of spliced mRNAs. During induction of gene expression, unspliced mRNAs are present in excess of the expectation based on the equilibrium rate γ, whereas the opposite is true during repression (Fig. 1d). The balance of unspliced and spliced mRNA abundance is, therefore, an indicator of the future state of mature mRNA abundance, and thus the future state of the cell.

To illustrate that such a simple model can be used to extrapolate the mature mRNA abundance into the future, we examined a timecourse of bulk RNA-seq measurements of the mouse liver circadian cycle12. Unspliced mRNA levels at each time point were consistently more similar to the spliced mRNA of the subsequent time (Fig. 1e), and many circadian-associated genes showed the expected excess of unspliced mRNA relative to slope γ during up-regulation, and a corresponding deficit during down-regulation (Fig. 1f-g). Solving the proposed differential equations for each gene allowed us to extrapolate each measurement throughout the circadian cycle, accurately capturing the expected direction of progression of the circadian cycle (Fig. 1h).

Next, to demonstrate ability to predict transcriptional dynamics in single-cell measurements, we analyzed recently-published single-cell mRNA-seq data on mouse chromaffin cells13, obtained using SMART-seq25 (Fig. 2). During development, a substantial proportion of chromaffin cells, which are neuroendocrine cells of the adrenal medulla, arise from Schwann cell precursors, providing a convenient test case in which the direction of differentiation can be validated by lineage tracing. Phase portraits of many genes showed the expected deviations from the predicted steady-state relationship (Fig. 2b-c). RNA velocity estimates of the individual cells accurately recapitulated the transcriptional dynamics within this dataset, including general movement of the differentiating cells towards chromaffin fate (Fig. 2d), as well as movement towards and away from the intermediate differentiation state. The velocity also captured cell cycle dynamics involved in the chromaffin differentiation, both in PCA projection and in a focused analysis of cell-cycle associated genes (Supplementary Note 2 Section 5).

Our velocity estimation procedure incorporates several features to accommodate the complexity of splicing biology (Supplementary Note 1). The estimation of the gene-specific equilibrium coefficient γ is performed using regression on the extreme expression quantiles, ensuring robust estimation even when most of the observed cells are outside of the steady state (Supplementary Note 2 Section 2). To accommodate genes observed far outside of their steady state, we also developed an alternative fit based on gene structure (Extended Data Fig. 4). A variety of techniques can be used to visualize the velocity estimates in low dimensions. The observed and extrapolated cell states can be jointly embedded in a common low-dimensional space (*e.g.* PCA in Fig. 2d). Alternatively, velocities can be projected onto existing low-dimensional embeddings such as t-SNE based on the similarity of the extrapolated state to other cells in the local neighborhood (Fig. 2h, see Supplementary Note 1). In large datasets, it is easier to visualize the prevalent pattern of cell velocities with locally averaged vector fields (Fig. 2i). Since cells can have RNA velocities along multiple independent components simultaneously, such as differentiation, maturation and proliferation, care must be taken when interpreting low-dimensional representations, as cells that lack apparent velocity in one particular embedding can nevertheless have substantial velocity in some subspace that is not visualized.

Cell-specific RNA velocity estimates provide a natural basis for quantitative modeling of cell fates. Metabolic labelling showed that for most genes, changes in the spliced/unspliced ratio would be detectable after 10 - 100 minutes (Extended Data Fig. 2). The effective timescale of extrapolation, on the other hand, depends on the biological process being analyzed. Based on the pulse labeling of chromaffin progenitor cells by EdU (Supplementary Note 2 Section 6), we estimate that we were able to extrapolate 2.5 - 3.8 hours into the future (Fig. 2f,g), which is also consistent with the ability to resolve cell-cycle events. Given the linear nature of the extrapolation, however, this predictive time-scale will depend on the shape of the gene expression trajectory (*i.e.* the curvature of the expression manifold). Cell fates can be predicted over longer time scales by tracing a sequence of small extrapolation steps on the observed expression manifold (Supplementary Note 2 Section 7).

To demonstrate the generality of our approach we analyzed data generated using additional single-cell RNA-seq protocols. We observed the transcriptional dynamics of neutrophil maturation in mouse bone marrow, and of light-induced neuronal activation in mouse cortex measured using the inDrop protocol (Extended Data Fig. 5), and of the intestinal epithelium (Extended Data Fig. 6), oligodendrocyte differentiation (Extended Data Fig. 7), and hippocampus development (see below), measured using 10x Genomics Chromium7. Estimates of RNA velocity were robust to variation of model parameters, gene and cell subsampling, with the most sensitive parameter being the size of the neighborhood used in visualization of velocity in pre-defined embeddings (Supplementary Note 2 Sections 10,11).

Most genes showed a positive correlation between velocity estimates and empirically observed expression derivatives (Extended Data Fig. 8), confirming that velocity vectors are informative. Failures in specific cases were due to several apparent causes, including genes observed exclusively far from equilibrium, uneven contribution of non-coding transcripts, and alternative splicing leading to multiple γ rates across the measured populations (Supplementary Note 2 Section 4).

We next applied RNA velocity to the branching lineage of the developing mouse hippocampus14. After removing vascular and immune cells, and GABAergic and Cajal-Retzius neurons (which originate from outside the hippocampus), t-SNE embedding revealed a complex manifold with multiple branches (Fig. 3a). We used known markers to identify the tips of the branches as corresponding to astrocytes, oligodendrocyte precursors (OPCs), dentate gyrus granule neurons, and pyramidal neurons of the five fields of the hippocampus: the subiculum, CA1, CA2, CA3, and hilus (Extended Data Fig. 9). Phase portraits of individual genes showed specific induction and repression of gene expression along the manifold (Fig. 3b, Extended Data Fig. 10). For example, *Pdgfra* (a marker of OPCs) was induced in pre-OPCs and maintained in OPCs; it showed corresponding positive velocity in the pre-OPC state, but neutral in the OPCs. Similarly, *Igfbpl1* was expressed specifically in neuroblasts, and showed positive velocity from radial glia to neuroblasts, but negative velocity going from neuroblasts to the two main neuronal branches.

RNA velocity showed a strong directional flow towards each of the main branches (Fig. 3c, Extended Data Fig. 10), originating in a small group of cells arranged in a band (Fig. 3c inset, dashed line). We identified these cells as radial glia based on the expression of markers including the Notch target *Hes1* and the homeobox transcription factor *Hopx* (Extended Data Fig. 9). Indeed, fate mapping has previously shown radial glia to be the true origin of the lineage tree of the hippocampus15. Using a Markov random walk model on the velocity field, the terminal and root states could be automatically identified (Fig. 3c), demonstrating the power of RNA velocity to orient the lineage tree without prior knowledge about the developmental process. On one side, velocity pointed towards astrocytes (expressing *Aqp4*) without intervening cell division, or alternatively to a pre-OPC state, leading through a narrow passage to proliferating OPCs. We speculated that the narrow passage represented the moment of commitment to the oligodendrocyte lineage. At this microstate level, fate choice is likely a non-deterministic process involving the tilting of gene expression in favor of one or the other fate, followed by a lock-in of the final fate once transcription factor feedback loops are established16. Comparing the probability distribution of future states for a cell starting among the pre-OPCs, versus one starting in the narrow passage leading to OPCs, revealed a clear difference, where the latter cell was overwhelmingly likely to end up as a fully formed OPC, whereas the former was as likely to remain in the pre-OPC state (Fig. 3d).

Some cycling progenitor cells (Extended Data Figs. 9b) expressed neurogenic transcription factors (*e.g. Neurod2*, *Neurod4*, *Eomes*) and those cells showed velocity pointing toward the immature neuroblast state, leading towards the three main neuronal branches in the upper part of the manifold. Granule neurons of the dentate gyrus first split from the hippocampus proper, and a second split divided the hippocampal cells into subiculum/CA1 and CA2-4, respectively (Extended Data Figs. 9, 10), in agreement with the major functional and anatomical subdivisions of the hippocampus. The detailed, single-cell view of a branching lineage allowed us to ask questions about fate choice. Examining two adjacent neuroblasts, just at the entrance to the branching point between CA and granule fates (Fig. 3e), we found that although their current states were neighbors (in gene expression space), their futures were already tilted towards different fates, distinguished by activation of *Prox1* (Fig. 3c, insert). Consistent with these findings, it has been shown that *Prox1* is required for the formation of granule neurons, and that when *Prox1* is deleted, neuroblasts instead adopt a pyramidal neuron fate17.

To demonstrate that RNA velocity is detectable in the human embryo, we performed droplet-based single-cell mRNA-seq of the developing human forebrain at ten weeks post-conception, focusing on the glutamatergic neuronal lineage (Fig. 4a). We found a strong velocity pattern originating from a proliferating progenitor state (radial glia), and proceeding through a sequence of intermediate neuroblast stages to a more mature differentiated glutamatergic neuron expressing *SL17A7* (the vesicular glutamate transporter used in forebrain excitatory neurons). We validated the expression of known and novel markers of cortical neuron development by multiplexed *in situ* hybridization (Fig. 4b-c), confirming the predicted expression of *CLU* and *FBXO32* in the ventricular zone (radial glia; marked by *SOX2*), *UNC5D* in the intermediate zone (neuroblasts; marked by *EOMES*) and *SEZ6* and *RBFOX1* in the cortical plate (neurons; marked by *SLC17A7*, also known as VGLUT1). The layered expression of these genes in the tissue (Fig. 4c) corresponded closely to the pseudotemporal distribution of their expression in the single-cell RNA-seq data (Fig. 4b).

We used principal curve analysis to order the cells according to a differentiation pseudotime, and examined the temporal progression of transcription in human primary cells. We confirmed that unspliced mRNAs consistently preceded spliced mRNAs during both up- and down-regulation (Fig. 4d).We observed both fast and slow kinetics. For example, *RNASEH2B* showed fast kinetics, with little difference between unspliced and spliced RNAs. In contrast, genes such as *DCX, ELAVL4* and *STMN2* showed evidence of an initial burst of rapid transcription, followed by sustained transcription at a reduced level (as evidenced by the shape of the unspliced RNA curve, Fig. 4d), with spliced transcripts following a noticeably delayed trajectory. Such dynamic induction with overshooting has been proposed to help quickly induce genes whose degradation kinetics are slow2, but have not been possible to study in human embryos.

The fact that RNA velocity is grounded in real transcription kinetics promises to bring a more solid quantitative foundation to our understanding of the dynamics of cells in gene expression space during differentiation. We envision future manifold learning algorithms that simultaneously fit a manifold and the kinetics on that manifold, based on RNA velocity. RNA velocity has already enabled the detailed study of dynamic processes in whole organisms18, and will greatly facilitate lineage analysis particularly in the human embryo.

Methods

Theoretical description of RNA velocity

Based on the model of transcription shown in Fig. 1, we developed a computational framework for robust inference of RNA velocity. A detailed description of the theory and computational methods is available as Supplementary Note 1.

Analysis pipeline, parameters and implementations details

We implemented the procedures above as two complete pipelines, one in R and one in Python, called velocyto.R and velocyto.py, respectively. These were used to generate all the analyses in the paper, with detailed settings as described in the following sections.

Annotation of spliced and unspliced reads

Read annotation for all protocols was performed using velocyto.py command-line tools. The velocyto.py annotation starts with bam file(s). For the 10x genomics platform datasets, the bam file was processed using default parameters of the *cellranger* software (10x Genomics). For the inDrop dataset, the reads were demultiplexed using dropEst pipeline19, using '-F -L eiEIBA' options to produce an annotated bam file analogous to *cellranger* output. For SMART-seq2 data, demultiplexed cell-specific bam files were fed into velocyto.py directly. The genome annotations GRCm38.84 and GRCh37.82 from the *cellranger* pre-built packages were used to count molecules while separating them into three categories: "spliced", "unspliced" or "ambiguous".

The annotation process considered only reads that could be mapped uniquely. Reads with multiple mappings and reads mapped inside repeat-masked (based on the UCSC genome browser repeat masker output) regions were discarded. For UMI-based protocols, the counting was performed on the level of molecules, taking into account annotation (spliced, unspliced, etc.) of all reads associated with that molecule (supporting read sets) into consideration. The supporting read sets for each molecule were determined by a combination of cell barcode and UMI sequence. For inDrops datasets, where UMI barcode does not have sufficient complexity to uniquely identify a molecule in the dataset, the reads were grouped based on the cell barcode, UMI and the region of the genome where it mapped (chromosomes, binned in 10Mbase regions). For each molecule, all annotated transcripts that were compatible with the given set of read mappings were considered, and cases where the set of reads associated with a given molecule was not compatible with any annotated transcript model were discarded. Cases where a set of supporting read mappings was compatible with transcript models of two or more different genes were also discarded.

The following set of rules was applied to annotate a set of read supporting a given molecule as spliced, unspliced or ambiguous:

1. A molecule was annotated as spliced if all of the reads in the set supporting a given molecule map only to the exonic regions of the compatible transcripts.
2. A molecule was annotated as unspliced if all of the compatible transcript models had at least one read among the supporting set of reads for this molecule

mapping that i) spanned exon-intron boundary, or ii) mapped to the intron of that transcript.

Molecules for which some of the compatible transcript models had exonic-only mappings, while others included intronic mappings were annotated as ambiguous and not used in the downstream analyses.

Similar logic was applied in annotating and counting reads for the SMART-seq2 dataset, with the following notable differences: 1) as reads lacked UMI, each read was considered to be an independent molecule; 2) as the protocol does not distinguish strands, transcript annotations on both strands were considered when annotating each read.

Chromaffin datasets processing (Fig. 2)

Chromaffin E12.5 and E13.5 datasets were processed using velocyto.R pipeline. The γ coefficients and velocity estimates were calculated for genes meeting a number of filtering criteria: $\gamma \geq 0.1$; Spearman rank correlation between s and $u \geq 0.1$; average s counts for a gene ≥ 5 for at least one cell subpopulation (cluster); average u counts for a gene ≥ 1 for at least one cell subpopulation; for the datasets where spanning reads were annotated (E12.5, E13.5), average spanning read counts were required to be ≥ 0.5 in at least one subpopulation. For SMART-seq2 datasets, the abundance of reads spanning intron and exon boundaries is sufficiently high to enable estimation of the unspliced offset o. The offset was estimated using a linear regression.

Mouse hippocampus dataset analysis (Fig. 3)

A total of 18,213 cells were analyzed (postnatal day 0: 8,113 cells; postnatal day 5: 10,100 cells). The embedding was computed on the correlation similarity matrix using pagoda2 (https://github.com/hms-dbmi/pagoda2). Briefly, gene variance normalization was performed by fitting a generalized additive model of variance on expression magnitude, and rescaling the gene variance by matching the tail probabilities of log residuals from the F distribution to the chi squared distribution with the degrees of freedom corresponding to the total number of cells. Cell distances were determined as $1 - r_{ij}$, where r_{ij} is Pearson linear correlation of the cell i and j scores on the first 100 principal components of the top 3000 variable genes in the dataset. Clustering was performed using the Louvain community detection algorithm on the nearest neighbor cell graph (k=30, pagoda2 implementation). For the velocity analysis lowly expressed (spliced) genes were excluded (requiring 40 minimum expressed counts and detected over 30 cells) and the top 3000 high variable genes were selected on the basis of a non-parametric fit of coefficient of variation (CV) *vs.* mean (using support vector regression). Only 1706 genes that had unspliced molecule counts above a detection threshold (25 minimum expressed counts and detected over 20 cells) were kept for the analysis. To normalize for the cell size, the counts were divided by the total number of molecules in each cell, and multiplied by the mean number of molecules across all cells. Spliced and unspliced counts were normalized separately. To reduce dimensionality, PCA was performed and the top 19 variable components were kept on the basis of the explained variance ratio profile. Euclidean distance in this reduced PCA space was used to construct a k-nearest neighbor graph (k=500), using a greedy balanced k-NN algorithm that limits each node to have no more than 4*k incoming edges. This graph was used to perfrom k-NN pooling. Velocity-based extrapolation was performed using Model I assumptions.

Human glutamatergic neurogenesis analysis (Fig. 4)

Pseudotime analysis was performed by fitting principal curve in the space of the top four principal components (using the R package *princurve*). The cell positions were projected onto the curve and the length of the arc between projections was used as pseudotime coordinates. The direction of the pseudotime was determined using the velocity field. Clusters were determined using Louvain community detection algorithm on the nearest neighbor graph in the same subspace. For the velocity analysis lowly expressed (spliced) genes were excluded (requiring 30 minimum expressed counts and detected over 20 cells). The top 2000 most variable genes were selected on the basis of a non-parametric fit of CV *vs.* mean (using support vector regression). A total of 987 genes that had unspliced molecules above a detection threshold (requiring 25 minimum expressed counts and detected over 20 cells; average spliced counts for a gene 0.06 in a subpopulation and average unspliced counts for a gene 0.007 in a subpopulation) were kept for the analysis. To normalize for the cell size, the counts were divided by the total number of molecules in each cell, and multiplied by the median number of molecules across all cells. For cell k-NN pooling, a k-nearest neighbor graph (k=550) was constructed based on Euclidean distance in the space of the top six principal components, as described above. The gamma coefficients were fit using the extreme quantile fit with diagonal quantiles, as described above.

For the visualizations in Figure 4b, the following maxprojection procedure was used to color the cells according to expression of the pre-defined gene set. First, the (cell-size normalized) expression of each gene included in the set was rescaled, dividing it by the 98th percentile magnitude. After rescaling, each cell was colored with the color corresponding to the gene that was expressed at highest level compared to other genes, and the saturation of the color was chosen to be proportional to the level of expression in the cell. The rescaled expression of the gene was required to exceed 0.45 in order for the cell to be colored.

Genes whose expression peaks at different stages of neurogenesis were selected using a heuristic gene enrichment score: $\frac{\mu_{cluster} * f_{cluster}}{\mu_{all} * f_{all}}$ where μ indicates the average molecule count of a gene and f is the fraction of cells in which the gene is detected. Figure 4d shows a selection of top-enriched genes, spliced and unspliced molecules were brought to a comparable scale by multiplying spliced molecular counts by the estimated γ.

Analysis of Mouse Oligodendrocytes lineage (Extended Data Fig. 7)

We analyzed a dataset of oligodendrocyte differentiation from murine pons extracted from a recently published cellular atlas20. We restricted the analysis to the trajectory of differentiation from oligodendrocyte precursor cells (OPCs) to mature oligodendrocytes by selecting cells that were labeled in the atlas as OPCs, COPs. NFOLs or MFOLs, for a total of 6307 cells.

As an initial step, for the Supp. Figure 7d-f, we performed a straightforward feature selection, first removing genes expressed lower than 15 spliced molecules, or lower than 8 unspliced molecules, requiring a minimal average spliced expression of 0.075 and minimal unspliced expression of 0.03 in the highest expressing cluster. A CV-mean fit was used to select the 606 most variable genes.

As the simple procedure above retained significant sex-driven batch effect (shown in Supp. Figure 7e), we then used a different approach aimed at minimizing batch effects by focusing on the genes that were uniquely relevant to the observed oligeodendrocytes. Specifically, a list of genes enriched in the oligodendrocyte lineage when compared to all other cell types was used to analyze the dataset. For each cell cluster we used the top 190 genes as sorted by enrichment (differential upregulation) scores, calculated as described in 20. The resulting set of genes was subjected to further filtering where lowly detected genes where excluded, requiring at least 5 spliced and 3 unspliced mRNA molecules detected in the whole dataset, resulting in 606 genes. We then normalized the cell total molecule counts using the initial molecule count as normalization factor. For cell k-NN pooling we built a k-nearest neighbor graph (k=90) based on Euclidean distance in the top nine principal components. Data was clustered using Louvain community detection algorithm on the nearest neighbor graph and colored according a pseudotime computed by a principal curve. Finally, we calculated gammas, velocity and extrapolation as described above; transition probabilities were computed using n_sight=300 and log transform.

Analysis of visual cortex response to light simulation (Extended Data Fig. 5)

For the pre-processing of the inDrops light stimulated mouse visual cortex dataset21 we used the dropEst pipeline (https://github.com/hms-dbmi/dropEst). First the *droptag* command was run on each fastq file using 10 as the minimum quality parameter. Then, mapping was performed using the STAR aligner. Finally, the *dropest* command was run to perform UMI and cell barcode correction, and the following flags were passed "`–m –V –b –L eiEIBA`” to produce a *cellranger*-like bam file. velocyto.py “`run_dropest`” command was used to annotate and count molecules.

Cell annotations from the original publication were used to extract ExcL23_1 (the largest and most homogeneous cell population described as responsive to stimulus in the original publication). We excluded cells whose total spliced RNA abundance was below 15th percentile (as low quality cells) and above the 99.5th percentile (as possible doublets). The dataset was further balanced by equalizing the number of cells representing each stimulation condition (unstimulated, 1h stimulation, 4h stimulation), randomly down-sampling subpopulations to match the number of cells in the less abundant condition. Genes whose total spliced molecule count was less than 250, or the number of expression cells was less than 150 were removed. Similarly, we removed genes whose total unspliced molecule count was less than 18, or number of expression cells was less than 15. To focus our analysis on the stimulation process and to avoid capturing orthogonal variation we performed a model-based feature selection. Briefly, we considered a negative binomial generalized linear model with predictors: size (as estimated by total number of molecules), the stimulation time (categorical and interaction with size) and no offset (i.e. correspondent to the R formula:

`expression ~ size + size:stimulation - 1`). We then performed likelihood ratio test comparing our model against the alternative model that does not take the stimulation predictor in account. Only statistically significant genes ($p < 0.001$ for spliced and $p < 0.03$ for unspliced molecules) were considered for the downstream analysis. After this step we further eliminated the cells ranking in bottom 10% of total molecular counts over all of the selected genes. For the cell k-NN pooling, we built a k-nearest neighbor graph (k=70) based on the Euclidean distance. Importantly, in this case, we reasoned that it was not correct to average across different independent stimulation conditions (*e.g.* non-stimulated and 1h-stimulation), therefore pooling was only allowed between cells of the same stimulation condition. Model 2 was used for velocity-based extrapolation, with *t* set to 15. For the transition probability calculation, the *n_sight* parameter was set to 200, and square root was used as a variance stabilizing transformation. Early and late response genes illustrated in Extended Data Figure 6 were extracted from the Supplementary Table 3 of the original publication, containing a list of significantly induced genes in different cell types21.

Analysis of gammas over different cell types using Tabula Muris (Extended Data Fig. 3)

The Tabula Muris dataset (including only the samples generated using droplet-based 10x Genomics Chromium protocol) was analyzed using velocyto.py, using the bam files and the valid barcodes list provided by the authors. All of the experiments were merged into a single dataset. The average of spliced and unspliced raw molecule counts over the different annotated cell types were calculated, and Pearson's correlation coefficient was computed. To reduce bias associated with variation in cell coverage, we removed from the analysis the clusters with less than 120 cells as well as several outlier clusters that had more than 3000 cells. Erythrocytes were also excluded, as they lack nuclei. To avoid inflating our correlations with trivial cases where a gene is expressed by just one or two cell types we applied the following filters: A gene was taken into consideration only if its expression levels met all of the following conditions: (1) at least 5 cell types with average of at least 0.04 spliced molecules; (2) at least 4 cell types with average of at least 0.02 unspliced molecules; (3) the highest expressing cell type expressed the gene at an average of at least 0.15 spliced molecules; (4) at least 2 other cell types express the gene at least 15% the level of the maximum expressing cell type. Furthermore, to avoid that inflation of correlation estimates by zeros, correlation of each gene was calculated considering only the cell types that expressed the gene at minimum 10^{-5} spliced and 5×10^{-6} unspliced levels. The estimates of gammas provided in Extended Data Fig. 3 were obtained as the slope of RANSAC regression without intercept. Double gammas were estimated using a mixture of generalized linear regression models fitted by expectation maximization, as implemented in the R package *flexmix*. The fraction of genes that are better explained by two or more values of gammas than by a single gamma was estimated by comparing the double gamma model fit with a single-gamma generalized linear model fit. Specifically, a log likelihood ratio test was used with the difference in degrees of freedom between the single- and double-gamma models taken to be the number of cell types + 1. Bonferroni correction was applied, and genes with $p<0.05$ were reported as being significantly better explained by two gammas.

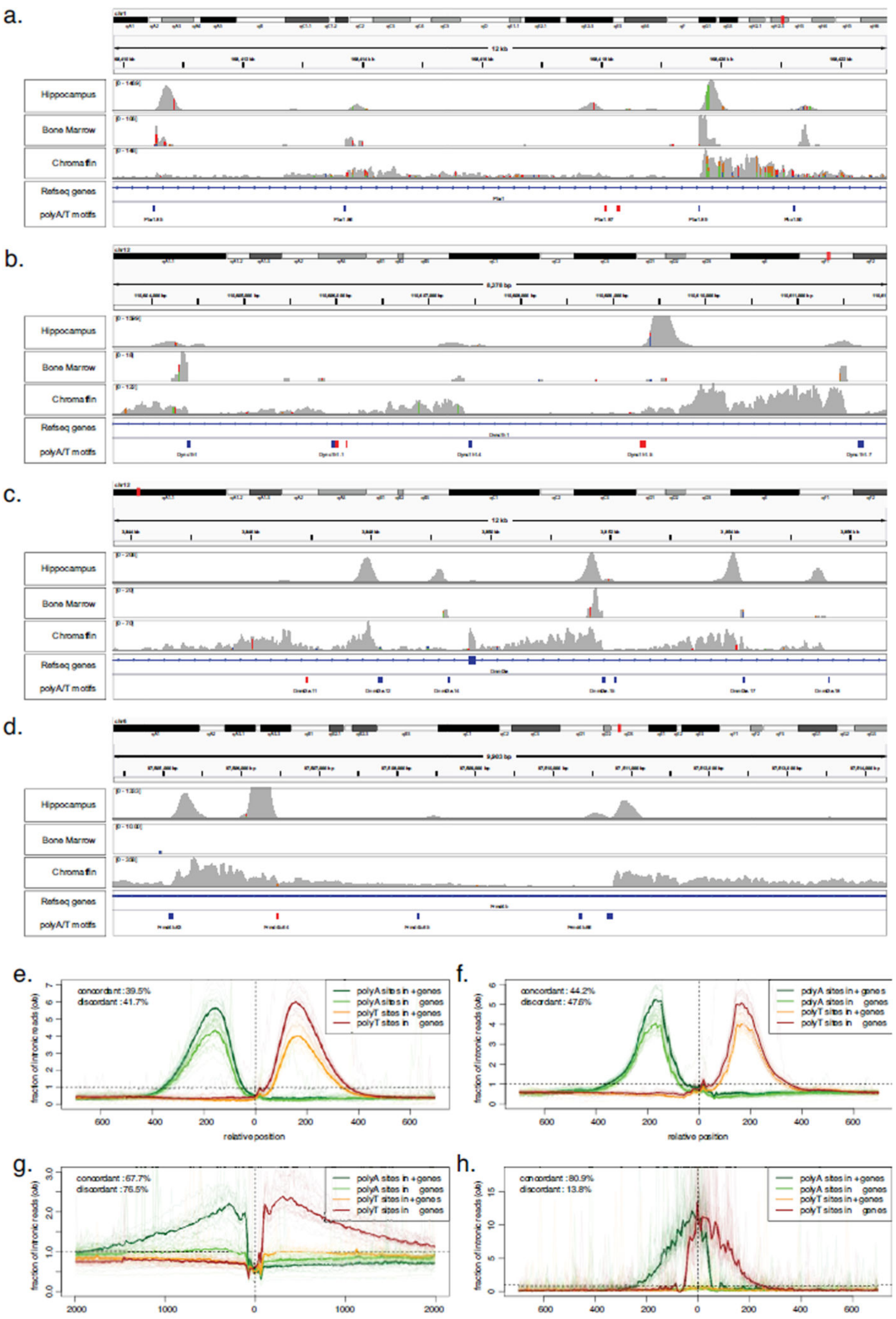
Analysis of the Intestinal epithelium (Extended Data Fig. 6)

velocyto.py, was run on the bam files and the valid barcode list provided by the authors. Cells with low levels of spliced (< 2000 molecules) and unspliced (< 300 molecules) were filtered out. Cell cycle genes, as defined by gene ontology annotation (using *Goatools*) were removed from the analysis. Genes with at least 30 spliced molecules and 20 unspliced molecules in the dataset were used in the downstream analysis. No clustering was performed, instead the cell type cell type annotation from the original publication was used. Feature selection was performed using these clusters. Specifically, the top 110 genes differentially upregulated in each cluster were selected. Genes whose minimum average expression in the highest expressing cluster was low were removed (unspliced <0.008, and spliced <0.08). Principal component analysis was performed on the cell-size-normalized data, and the first nine principal components were retained and used to calculate the t-SNE embedding (*cytograph* implementation, Euclidean distance). We calculated cell kNN pooling using the 70 nearest neighbors, as determined by the Euclidean distance in the same nine dimensional PCA space. Gammas were fitted, velocities computed using default parameters, and extrapolation carried on using Model II with $t = 4$. Transition probability was computed using *n_sight* of 30, using square root variance stabilizing transformation.

Human tissue and single-cell RNA sequencing (Fig. 4)

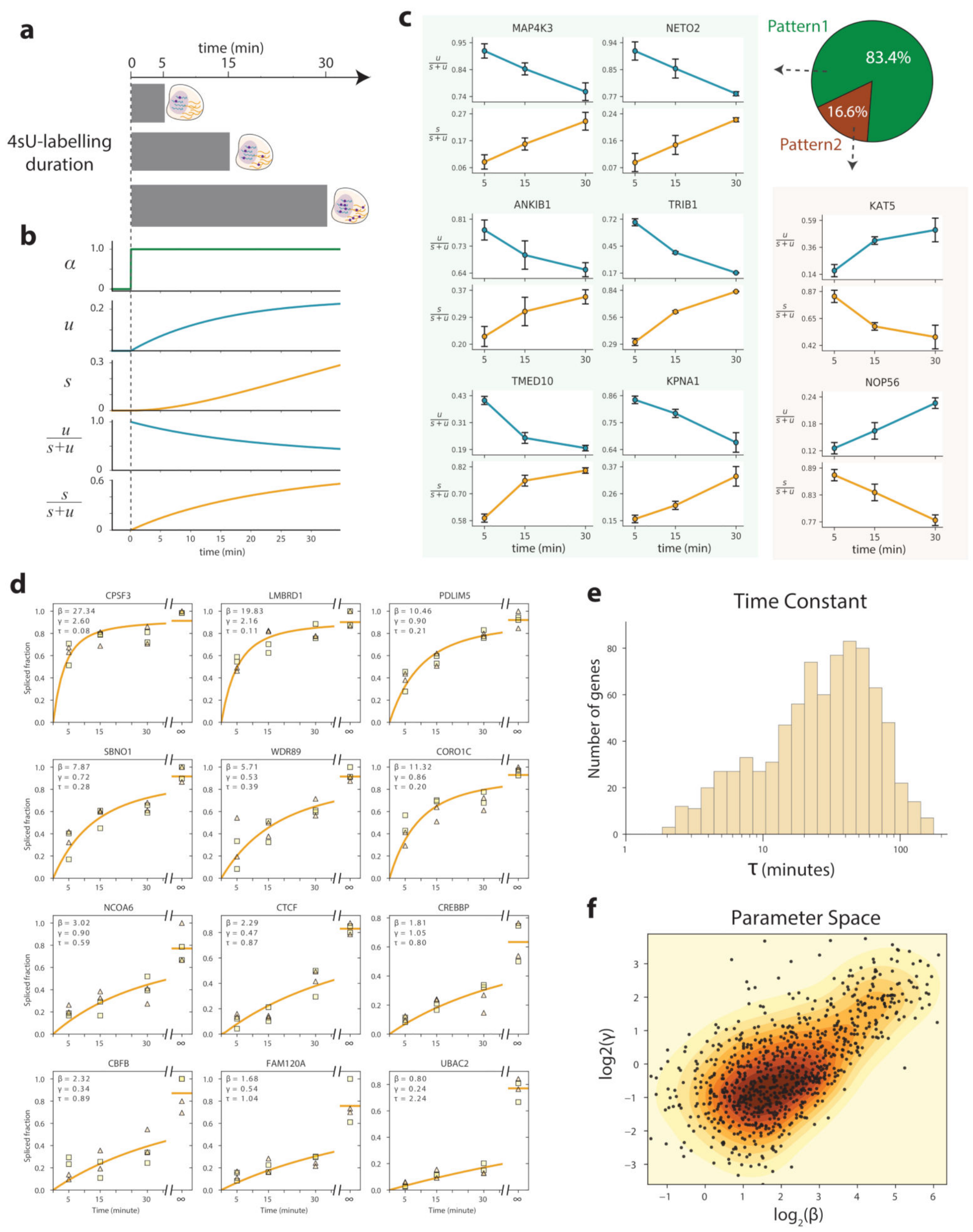
Human first trimester subcortical forebrain tissue was obtained from elective routine abortions (10 weeks postconception) with the written informed consent of the pregnant woman and in accordance with the ethical permit given by the Regional Ethics Vetting Board (Stockholm, Sweden). Human fetal forebrain tissue was collected and stored in hibernation media with addition of GlutaMAX and B-27 supplements according to the manufacture’s instructions (overnight, 4oC, Hibernate-A, Thermo-Fisher). The tissue was then cut into small cubic pieces of approximately 1-2mm length. Tissue was dissociated using a dissociation kit (Miltenyi, Neural Tissue Dissociation Kit (P)) according to manufacture’s instructions. In short, tissue was prepared in the kit buffer containing 0.067mM beta-mercaptoethanol. After addition of enzyme mix 1 and 2, the tissue was mechanically dissociated using three increasingly smaller gauges of fire polished Pasteur pipettes, pipetted 20, 15 and 10 times up and down respectively. Ultimately, collected cells were stored on ice in PBS containing 1% BSA and immediately prepared for single cell library preparation. Single-cell RNA sequencing was performed using the 10X Genomics Chromium V2 kit, following the manufacturer’s protocol, and sequenced on an Illumina Hiseq 2500.

Extended Data



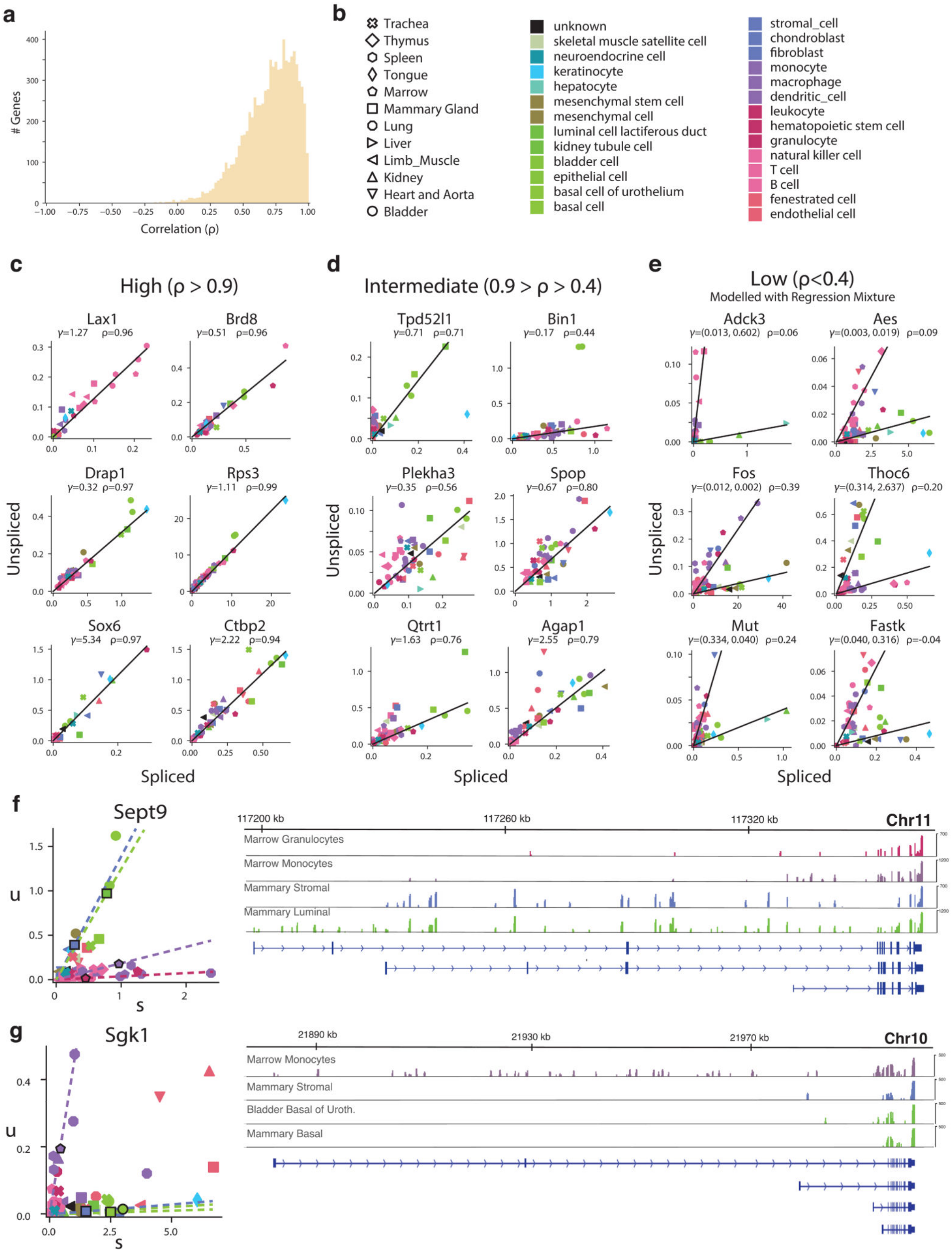
Extended Data Figure 1. Most of the intronic reads arise due to internal priming from stable positions.

a-d. Examples of read density around intronic polyA and polyT sequences. The browser screenshots show density of reads from the 10x Chromium mouse hippocampus dataset (top track of each panel), mouse bone marrow inDrop dataset (second track from the top), and chromaffin differentiation assessed using SMART-seq2 (third track). The bottom two tracks show gene annotation, and positions of polyA or polyT sequences (of length at least 15bp with one allowed mismatch). The polyA/polyT boxes are colored blue if the stretch is in a concordant orientation to the transcription of the underlying gene (i.e. would result in a polyA sequence in the nascent RNA molecule being transcribed), or red if they are oriented in the discordant position (i.e. would result in a polyT sequence in the RNA). The 3'-end based 10x Chromium and inDrop protocols show discrete peaks downstream of the polyA priming sites, with the 10x dataset also showing peaks upstream of the polyT sites. The SMART-seq2 protocol shows much more diffused peaks, expected from the full-length purification procedure used by the protocol. **e-h.** Average read density profiles around concordant and discordant internal priming sites. The plots show observed/expected intronic read density around $(A)_{15}$ or $(T)_{15}$ sequences (with 1 allowed mismatch) within the intronic regions. The x axis shows position relative to the motif position (in basepairs), in a genomic reference orientation. The bold lines show genome-wide average (trimmed of two extreme values among chromosomes for each position). The averages of individual chromosomes are shown semi-transparent lines. (e.) shows the profiles of mouse hippocampus 10x Chromium dataset (n=18,213), (f.) shows profile for human forebrain 10x data (n=1720), (g.) shows profile for the chromaffin differentiation data measured using SMART-seq2 (n=385), and (h.) shows profile for the mouse bone marrow data measured using inDrop (n=3018). The top left corner of each plot shows the number of all intronic reads (i.e. falling within the gene, but not touching an exon) that falls within the 250bp around internal priming sites (1500bp was used for the SMART-seq2 dataset). In 10x data, while concordant internal priming sites produce stronger signal, their frequency within the genome is lower than those of discordant sites, so that overall discordant sites account for slightly higher fraction of intronic signals. By contrast, the inDrop dataset appears to have very limited discordant priming.



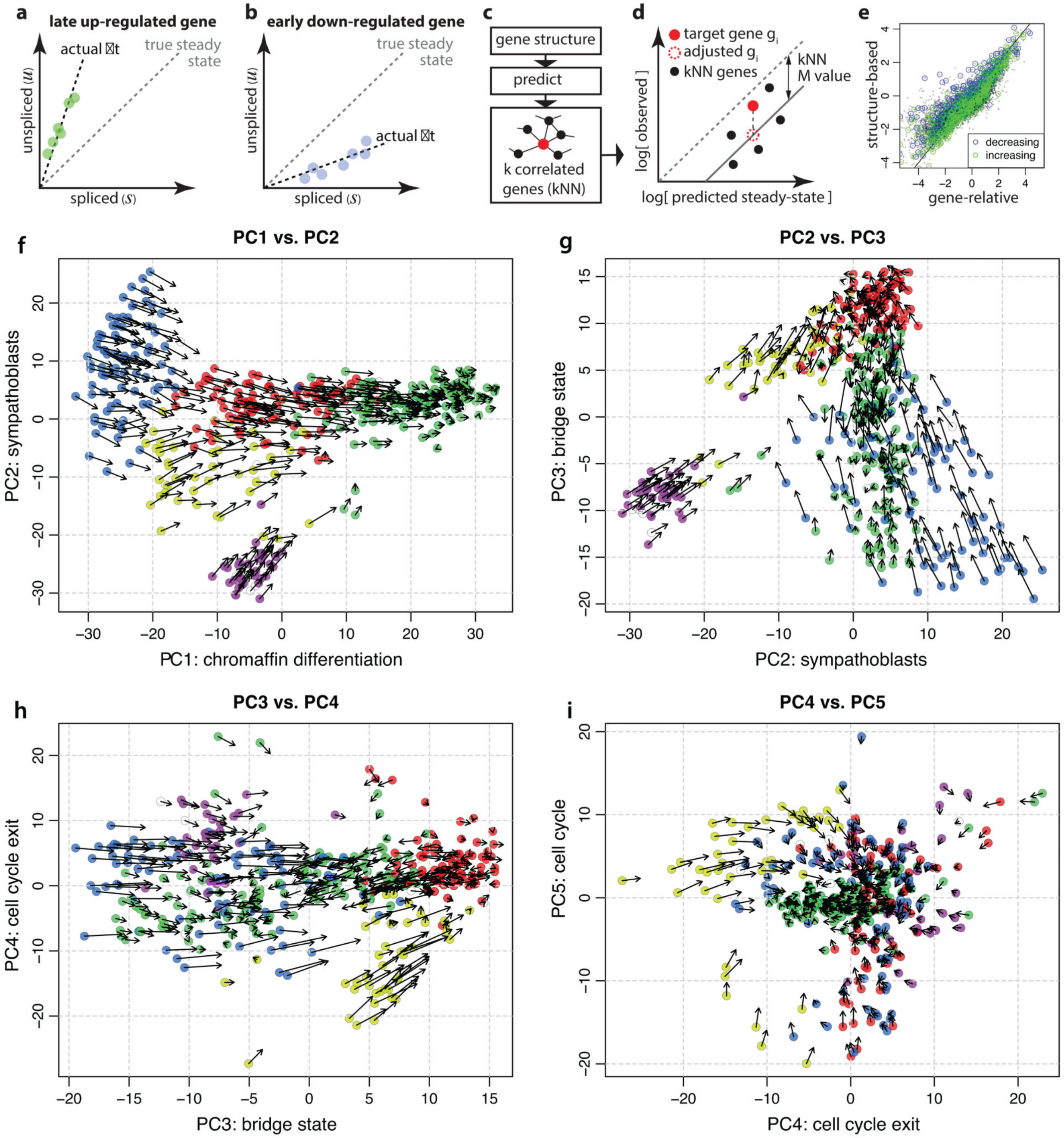
Extended Data Figure 2. Estimation of the characteristic time of RNA metabolism in human cells.

a. Design of the metabolic labeling experiment in human cells. HEK 293 cells were exposed to 4sU for 5, 15 or 30 min, the labeled fraction was isolated and analyzed. A no pull-down control was also analyzed, and represents the equilibrium state (indicated by ∞). **b.** Expected profiles of the abundance and fraction of labeled spliced and unspliced RNA molecules. **c.** The observed dynamic profiles of genes were clustered, yielding two groups: the majority (83.4%) were concordant with the expectation of increasing labeling; and a smaller fraction (16.6%) of discordant genes. Bars indicate SEM. n_{genes}=998, $N_{technical}$=2 $N_{biological}$=2. **d.** Curves showing maximum likelihood fit to the data, based on the analytical solution for a step increase of the transcription rate. The fit yields values of β and γ, and of the characteristic time constant τ, defined as the time required to reach 1 – 1 /e ≈ 63.2 % of the asymptotic value. **e.** The distribution of τ values and **f.** The joint distribution of the fit β and γ parameters (n=832).



Extended Data Figure 3. Degradation rates are conserved over a wide range of terminally differentiated cell types.

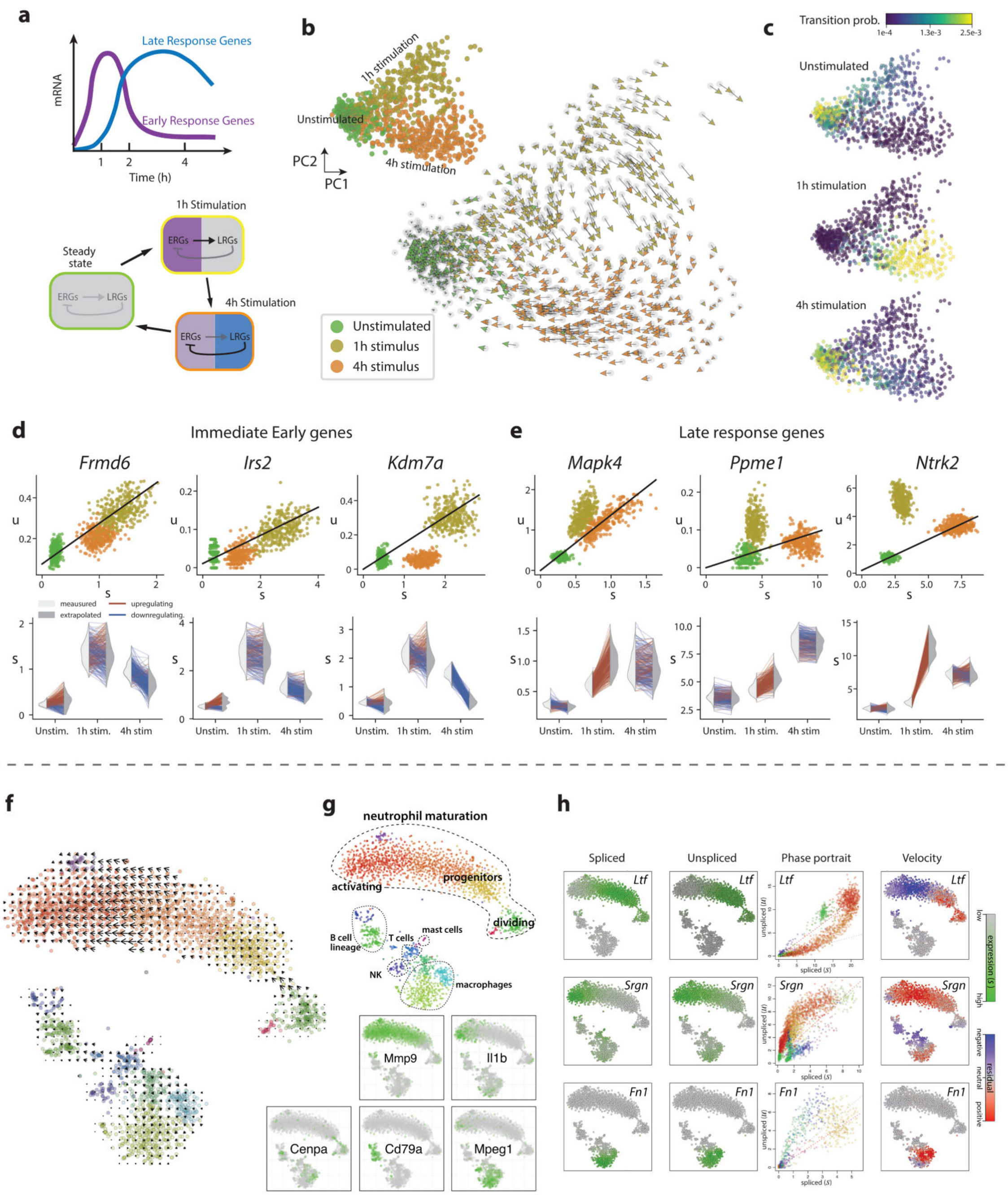
Conservation of the RNA degradation rate over a wide range of different cell types in the adult mouse (Tabula Muris dataset). **a.** The distribution over the genes of the correlation of spliced and unspliced molecule counts across all the cell types (n_{genes}=8,385). **b.** Legend enumerating the tissues and cell classes annotated by the Tabula Muris consortium (n=48). Functionally, developmentally or phenotypically related are colored with similar colors to aid the interpretation of the plots below. **c.** A representative selection of genes with high correlation ($\rho > 0.9$) and **d.** typical correlation ($0.9 > \rho > 0.4$). γ was estimated by robust linear regression (RANSAC) **e.** Plots show a selection of genes displaying two clearly distinct degradation rates (such genes with double γ amounted to 10.8% of the total). The values of the two different γ were estimated by regression mixture modeling. **f,g.** Two examples of genes where multiple gammas are explained by alternative splicing in different cell types.



Extended Data Figure 4. Structure-based velocity estimation.

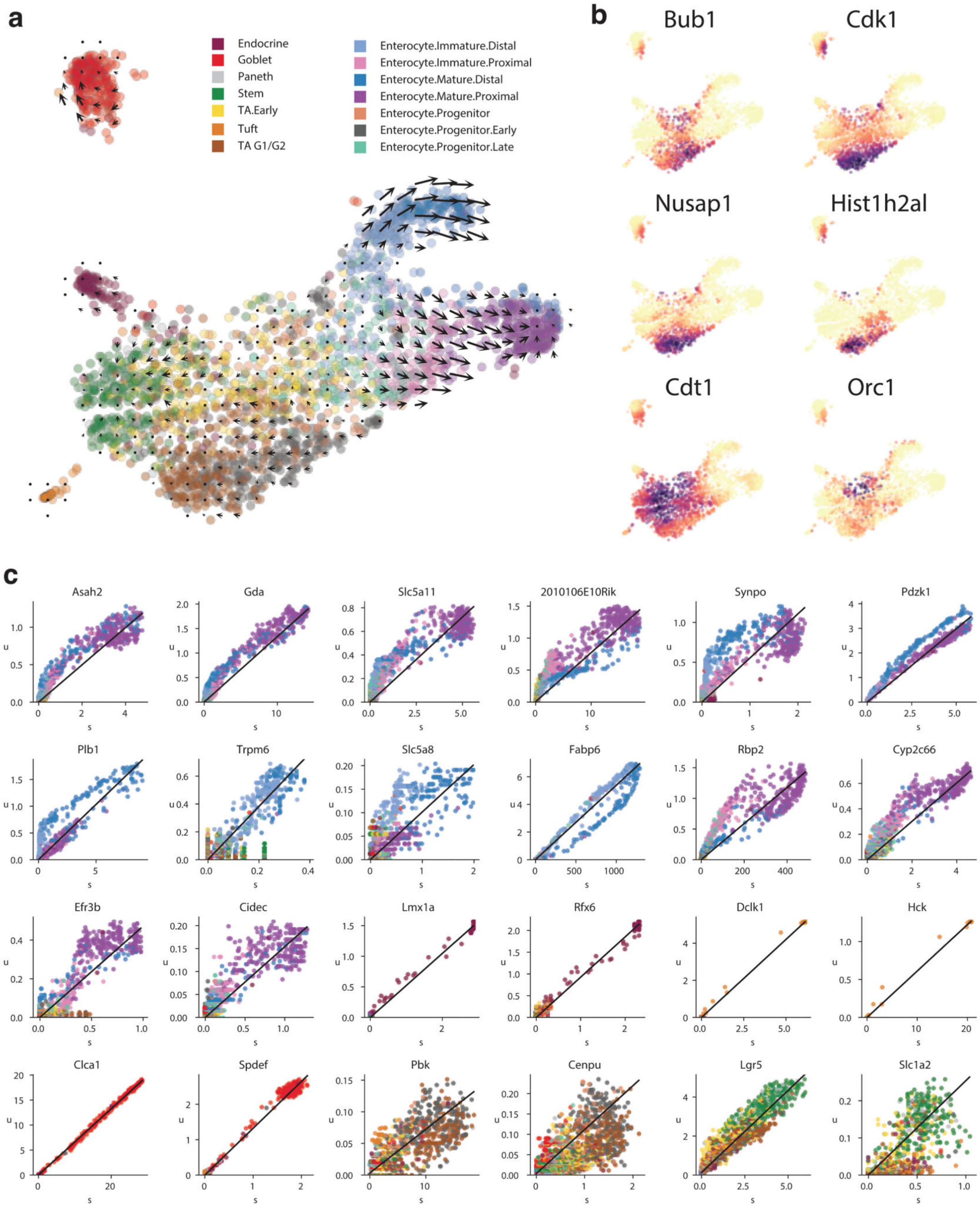
a,b. For genes that are observed only outside of the steady state, such as genes upregulated late in the chromaffin differentiation (a) or down-regulated early in the Schwann cell precursors (b), gene-relative γ fit will likely deviate from its steady-state value. **c,d.** To correct for such effects, a structure-based γ fit will first predict γ for every gene based on its structural parameters, and then use k most correlated genes in the dataset to adjust M-value ($M = log_2[u_o/u_{ss}]$, where u_{ss} is the unspliced counts predicted from spliced counts under steady-state, and u_o is the observed unspliced count) using robust mean, and re-estimate γ.

e. Scatter-plot comparing gene-relative and structure-based γ estimates, with colored circles highlighting γ adjustments for genes down-regulated early in SCPs (blue) and up-regulated late in chromaffin cells (green). The values are shown on a natural log scale. **f-i.** Cell expression velocity in the chromaffin E12.5 dataset, based on the structure-based γ estimates, shown on the first five PCs.



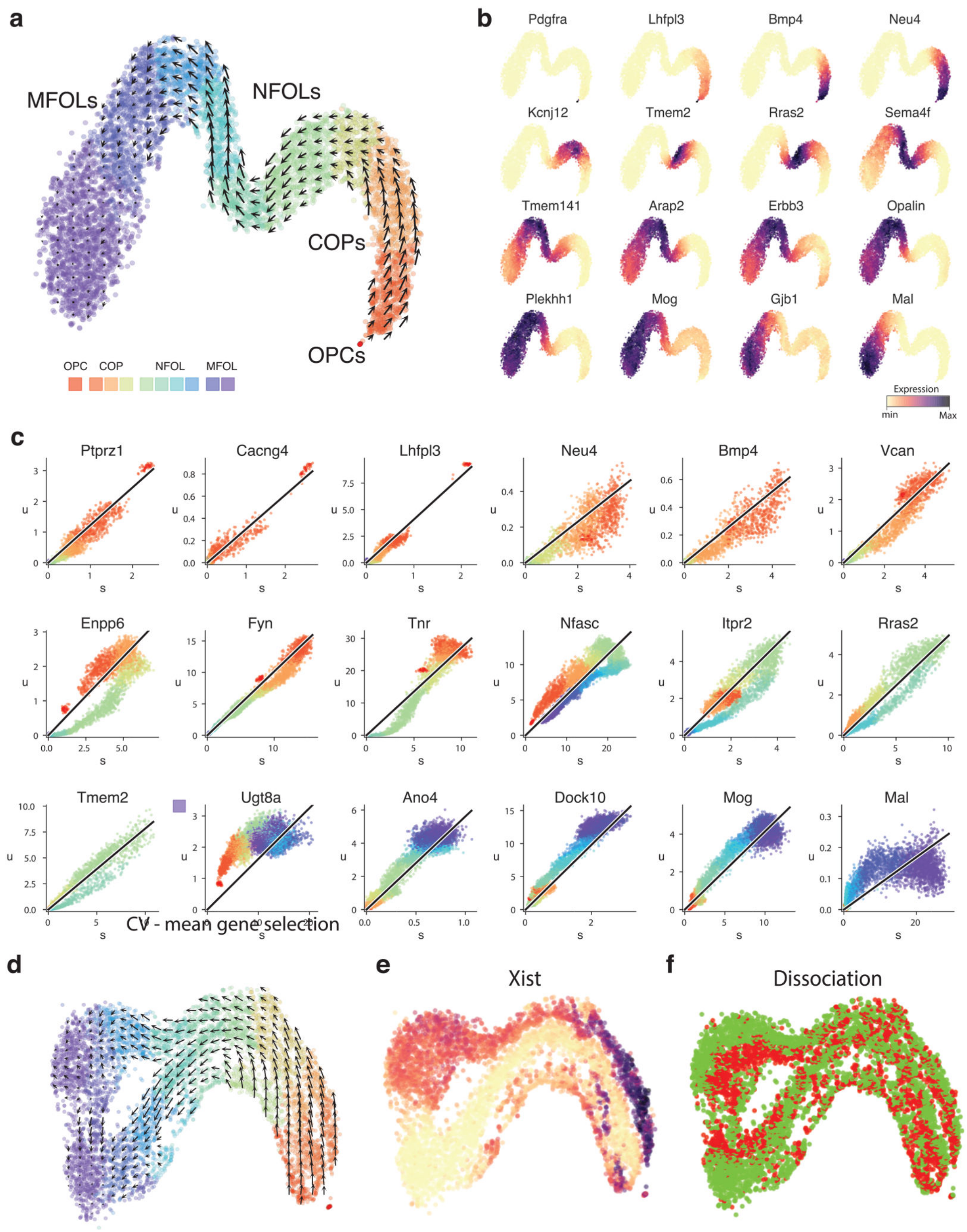
Extended Data Figure 5. RNA velocity analysis of inDrop datasets: visual stimulus response of cortical pyramidal neurons and neutrophil differentiation.

a. Simplified illustration of a model of activation of pyramidal neurons of the visual cortex after exposure to a light stimulus. **b.** Velocity estimates projected onto a two-dimensional PCA embedding of the dataset (n=952) **c.** Average transition probability of unstimulated cells (top), cells stimulated for 1h (middle) and cell stimulated for 4h (bottom). The unstimulated cells mostly were stationary and only few cells show tendency of activating early response genes (likely as a result of the dissociation procedure). Cells stimulated for 1h were characterized by expressing immediate early genes and high velocity in late response genes, and they were therefore transitioning to a state more similar to the one observed 4h activation time point. After 4h of stimulations cells appeared to be reverting to a state comparable to the unstimulated sample (bottom). **d,e.** Above, phase portraits of early (d) and late (e) response genes. Below, Violin plots show expression distribution over the cell population at each time point (left half of the violin) and extrapolation in to the future using velocity (right half of the violin). In the plot, transitions of single cells are indicated by lines connecting the two halves of the violins and colored by the sign of the velocity of each gene. **f.** Grid visualization shows cell expression velocity estimates for the inDrop mouse bone marrow dataset on a t-SNE embedding (n=3018). **g.** Major cell populations are labeled based on manual annotation. The velocity flow in (a) captures neutrophil maturation, starting from the dividing cells on the right, all the way to Il1b activation on the left. Expression profiles for five marker genes are shown below. **h.** The plots illustrate gene-relative model fits for several example genes. For each gene, the first column shows spliced molecular counts in different cells. The second column shows unspliced molecular counts. Third column shows phase portrait of a gene (unspliced *vs.* spliced dependency) and the resulting γ fit (dashed red line), as determined using extreme quantile method. Each point corresponds to a cell, colored according to cluster labels shown in (g). The last column shows unspliced count signal residual based on the estimated γ fit, with positive residuals indicating expected upregulation, and negative residuals indicating expected downregulation of a gene.



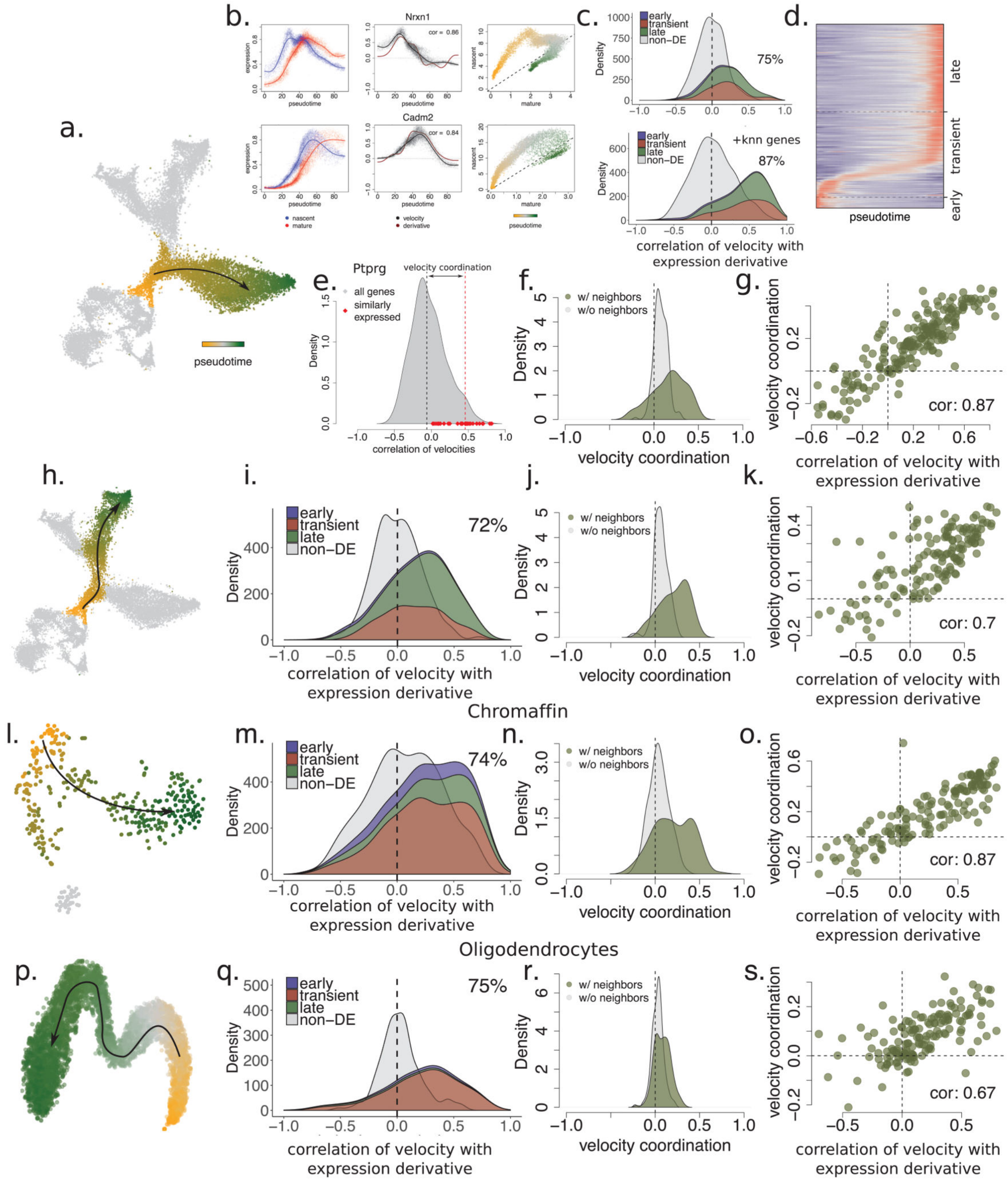
Extended Data Figure 6. Dynamics of maturation of enterocytes during intestinal homeostasis.

a. Velocity field projected on a 2D t-SNE embedding. The clusters are labeled and colored as in the original publication to facilitate comparison (n=2683). Velocity analysis revealed a transition related to the maturation of distal and proximal enterocytes. No consistent velocity was observed in the part of the manifold occupied by stem cells and transit amplifying (TA) cells, suggesting that stem cell dynamics is more difficult to capture either for its slower rate or a more stochastic nature. The small velocities of transit amplifying cells were likely driven by cell cycle process. **b.** A selection of the cell cycle genes that were removed in the analysis are plotted on the t-SNE. Despite the removal of the genes annotated as cell cycle genes we still observed important segregation by cell cycle, illustrating the difficulty of disentangling cell cycle phase from the cell state. **c.** A selection of phase portraits that show genes underlying the observed velocity field. Markers of Endocrine, Goblet and Tuft cells displayed no detectable velocity. Velocity towards and from stem cell states was detectable for limited set of genes (like the stem cell marker Lgr5), however on the genome-wide level the exact dynamics of this process was likely confounded by the high correlation with cell cycle.



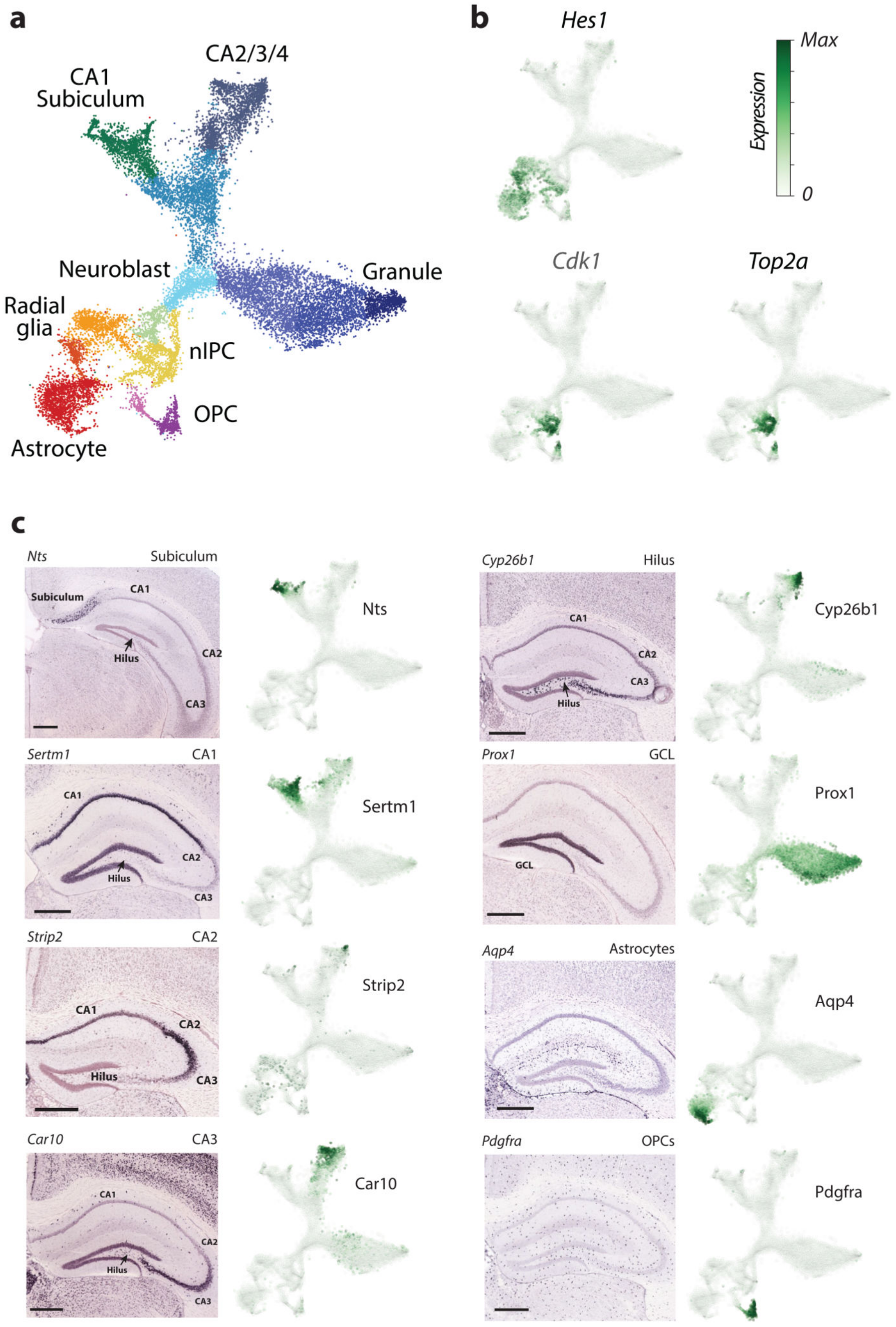
Extended Data Figure 7. RNA velocity unveils the dynamics of differentiation and myelination of oligodendrocytes.

a. t-SNE projection shows the landscape of oligodendrocyte lineage differentiation and myelination process in the hindbrain (pons) of adolescent (P20) mice (n=6307). The velocity field reflects the dynamics of expression of both the initial differentiation wave and the following expression changes associated to the myelination process. The cell clusters are colored by pseudotime as in (c) to facilitate interpretation. **b.** Expression patterns of landmark genes of the differentiation process. *Pdgfra* is the canonical marker of oligodendrocyte precursors (OPCs), *Neu4* marks committed oligodendrocyte precursors (COPs), *Tmem2* is enriched in newly formed oligodendrocyte (NFOLs) and the expression of *Mog* is upregulated at the beginning of the myelination process in myelin forming oligodendrocytes (MFOLs). **c.** A selection of phase portraits underlying the velocity field showed in (a). **d.** t-SNE projections and velocity vector field of the same dataset, but analyzed using a more naïve feature selection that has retained other axes of variation on top of the oligodendrocyte maturation (sex and day of dissection). Notice that despite separation of populations into *Xist+* and *Xist-* tracks, the velocity field correctly captures progression from progenitors to newly formed oligodendrocytes in the two parallel tracks. **e.** Level of expression of *Xist* showing that most of the extra variation is driven by the sex of the animal. **f.** Cells colored by the day the experiment was performed in.



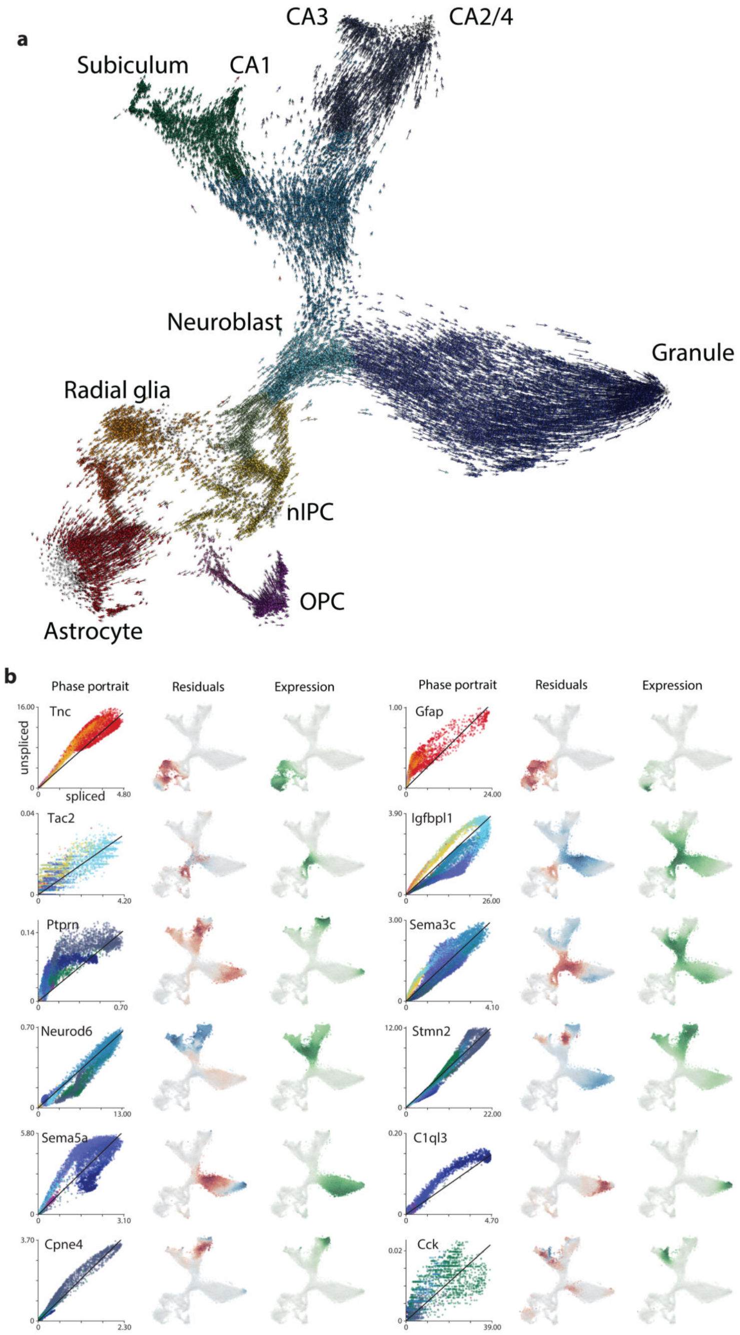
Extended Data Figure 8. Agreement of velocity predictions with the observed expression derivatives.

a. Maturation progression of granule neurons in the mouse hippocampus dataset is approximated by pseudotime (estimated with a principal curve). **b.** For a pair of example genes (rows), the plots show unspliced and spliced gene expression profiles along the pseudotime (left panels), empirically-estimated smoothed pseudotime derivative of the observed gene expression and the estimated RNA velocity (middle panels), as well as the relationship between spliced and unspliced expression (right panel). The velocity estimates for the two chosen genes are highly correlated with the empirically-observed derivative, indicating accurate velocity estimation. **c.** The majority (75%) of the genes that were differentially regulated along the pseudotime trajectory showed positive correlation with the empirical expression derivative. The distribution of such genes is split according to three classes of trajectory-associated genes as shown in d. By contrast, velocity estimates for genes that were not differentially expressed along the pseudotime trajectory did not show such correlation (grey). Incorporating information about co-regulated genes into velocity estimation using gene kNN clustering (see Supplementary Note 1) can significantly boost the accuracy of the velocity predictions (lower panel). **d.** Trajectory-associated genes were classified as early, transient and late, according to their peak expression time. x-axis: cells ordered by pseudotime, y-axis: genes ordered by their peak expression time. **e.** The genes that were well-correlated in terms of their spliced expression patterns with *Ptprg*, also showed high correlation of their velocity estimates with *Ptprg*. To assess the degree consistency of the velocities of co-regulated genes, we introduced a measure of velocity coordination for a given gene, as a difference between the mean correlations of the velocity estimates of the co-regulated genes and the velocity estimates of all genes. The two quantities being compared are shown for *Ptprg* with dotted vertical lines: grey – mean velocity correlation with all genes, red – mean velocity correlation with top co-regulated genes. Velocity coordination provides an unbiased measure of quality of velocity estimates. **f.** Velocities of co-regulated genes were correlated. Distribution of gene velocity coordination values is shown for genes that had co-regulated genes (*i.e.* the genes that had well-correlated gene neighbors in terms of their spliced expression pattern, green), as well as for the genes that did not have enough co-regulated genes (without neighbors, grey). **g.** Co-regulated genes that had high velocity coordination tended to have high correlation with the empirical derivatives. Spearman correlation coefficient is shown. **h-k.** Velocity performance during maturation of pyramidal neurons (h). Genes differentially expressed during maturation had high correlation of velocity with empirical derivative (i), co-regulated genes tended to have correlated velocity estimates (j) and the degree of velocity coordination was associated with its correlation with empirical derivative (k). **l-m.** Velocity performance during chromaffin differentiation. **p-s.** Velocity performance during maturation of oligodendrocytes. Number of top co-regulated genes analyzed for velocity correlation: (g): 200 genes, (k,o,s): 150 genes.



Extended Data Figure 9. Branching developmental trajectories of developing hippocampus.

a. t-SNE embedding of the developmental dentate gyrus dataset. Cells are colored by cluster identities, with labels shown for the major cell types. **b.** Expression of radial glia (and astrocyte) marker *Hes1*, and cell cycle genes *Top2a* and *Cdk1* shown on the t-SNE embedding. **c.** Marker genes of different regions of the hippocampus (*in situ* hybridization images from Allen Brain Atlas) show prominent expression signals at different extremities of the branching embedding. Scale bars, 0.5 mm.



Extended Data Figure 10. Single cell velocity estimates for individual cells in the embryonic hippocampus dataset.

a. Arrows indicate the extrapolated state projected onto the t-SNE embedding of the manifold. **b.** Selected phase portraits and fits of the equilibrium slope (γ) for the developing cells in the embryonic hippocampus dataset. For each gene, the first column shows spliced-unspliced phase portrait. The dashed line shows the γ fit. The second column illustrates the magnitude of the residuals (*i.e.* difference between observed and expected unspliced abundance, which closely tracks with velocity) for several genes involved in the development of different neural lineages. The third column shows the observed expression profile for spliced molecules.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The work reported here was supported by the Swedish Foundation for Strategic Research (RIF14-0057 and SB16-0065), the Knut and Alice Wallenberg Foundation (2015.0041), the Erling Persson Foundation (HumDevCellAtlas) and the Wellcome Trust (108726/Z/15/Z) to S.L.; Center for Innovative Medicine (CIMED) to K.L. and P.C.; Swedish Research Council, Marie Curie Integration Grant EPIOPC,333713, European Research Council EPIScOPE, 681893, Swedish Brain Foundation, Ming Wai Lau Centre for Reparative Medicine, Cancerfonden and Karolinska Institutet to G.C.B.; PVK was supported by NIH R01HL131768 from NHLBI and CAREER (NSF-14-532) award from NSF.

References

1. Linnarsson S, Teichmann SA. Single-cell genomics: coming of age. Genome Biol. 2016; 17:97. [PubMed: 27160975]
2. Zeisel A, et al. Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. Mol Syst Biol. 2011; 7:529. [PubMed: 21915116]
3. Gray JM, et al. SnapShot-Seq: A method for extracting genome-wide, in Vivo mRNA dynamics from a single total RNA sample. PLoS One. 2014; 9
4. Gaidatzis D, Burger L, Florescu M, Stadler MB. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. Nat Biotechnol. 2015; 33:722–729. [PubMed: 26098447]
5. Picelli S, et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013; 10:1096–8. [PubMed: 24056875]
6. Islam S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2013; 11:163–166. [PubMed: 24363023]
7. Klein AMM, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015; 161:1187–1201. [PubMed: 26000487]
8. Zheng GXY, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017; 8:14049. [PubMed: 28091601]
9. Schwalb B, et al. TT-seq maps the human transient transcriptome. Science (80-.). 2016; 352:1225–1228.
10. Islam S, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011; 21:1160–1167. [PubMed: 21543516]
11. Quake SR, Wyss-Coray T, Darmanis S, Consortium, TM. et al. Single-cell transcriptomic characterization of 20 organs and tissues from individual mice creates a Tabula Muris. bioRxiv. 2018 237446.
12. Vollmers C, et al. Circadian oscillations of protein-coding and regulatory RNAs in a highly dynamic mammalian liver epigenome. Cell Metab. 2012; 16:833–845. [PubMed: 23217262]
13. Furlan A, et al. Multipotent peripheral glial cells generate neuroendocrine cells of the adrenal medulla. Science (80-.). 2017; 357 eaal3753.
14. Kriegstein A, Alvarez-Buylla A. The Glial Nature of Embryonic and Adult Neural Stem Cells. Annu Rev Neurosci. 2009; 32:149–184. [PubMed: 19555289]
15. Malatesta P, et al. Neuronal or glial progeny: Regional differences in radial glia fate. Neuron. 2003; 37:751–764. [PubMed: 12628166]
16. Johnston RJ, Desplan C. Stochastic mechanisms of cell fate specification that yield random or robust outcomes. Annu Rev Cell Dev Biol. 2010; 26:689–719. [PubMed: 20590453]

17. Iwano T, Masuda A, Kiyonari H, Enomoto H, Matsuzaki F. Prox1 postmitotically defines dentate gyrus cells by specifying granule cell identity over CA3 pyramidal cell fate in the hippocampus. Development. 2012; 139:3051–3062. [PubMed: 22791897]

18. Plass M, et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science (80-.). 2018; 1723 eaaq1723.

19. Petukhov V, et al. Accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. Genome Res. 2018; doi: 10.1186/s13059-018-1449-6

20. Zeisel A, et al. Molecular architecture of the mouse nervous system. 2018

21. Hrvatin S, et al. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. Nat Neurosci. 2018; 21:120–129. [PubMed: 29230054]

22. Hochgerner H, Zeisel A, Lönnerberg P, Linnarsson S. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. Nat Neurosci. 2018; 21:290–299. [PubMed: 29335606]

23. Haber AL, et al. A single-cell survey of the small intestinal epithelium. Nature. 2017; 551:333–339. [PubMed: 29144463]

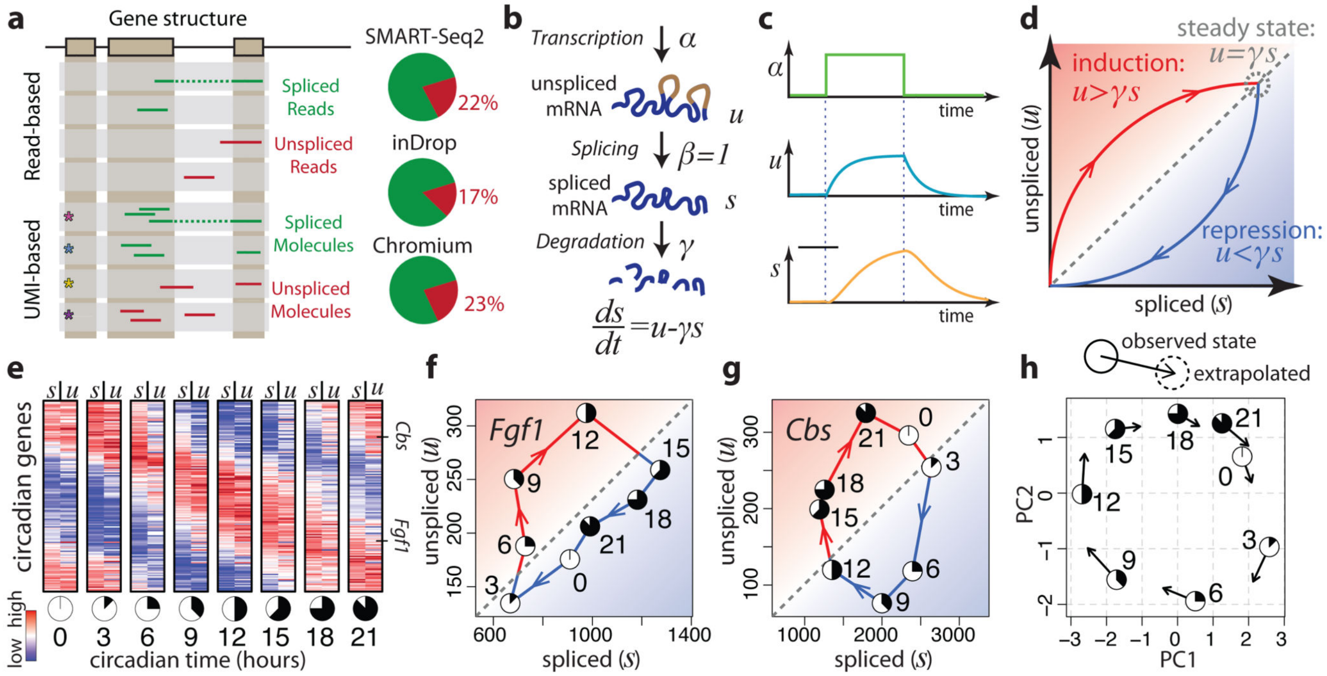


Figure 1. Balance between unspliced and spliced mRNAs is predictive of cellular state progression.

a. Spliced and unspliced counts are estimated by separately counting reads that incorporate intronic sequence. Multiple reads associated with a given molecule are grouped (* boxes) for UMI-based protocols. Pie charts show typical fractions of unspliced molecules.

b. Model of transcriptional dynamics, capturing transcription (α), splicing (*β*), and degradation (γ) rates involved in production of unspliced (*u*) and spliced (*s*) mRNA products.

c. Solution of the model in (b) as a function of time, showing unspliced and spliced mRNA dynamics in response to step changes in α.

d. Phase portrait showing the same solution (solid curves). Steady states for different values of transcription rates α fall on the diagonal given by slope γ (dashed line). Levels of unspliced mRNA above or below that proportion indicate increasing (red shading) or decreasing (blue shading) expression of a gene, respectively.

e. Abundance of spliced (*s*) and unspliced (*u*) mRNAs for circadian-associated genes in a 24h time course of mouse liver12. The unspliced mRNAs are predictive of spliced mRNA at the next time point.

f,g. Phase portraits observed for a pair of circadian-driven genes: *Fgf1* (f) and *Cbs* (g). The circadian time of each point is shown using a clock symbol (see bottom of Fig. 1e). The dashed diagonal line shows steady-state relationship, as predicted by γ fit.

h. Change in expression state at a future time *t*, as predicted by the model, is shown in the space of the first two principal components (PCs), recapitulating the progression along the circadian cycle. Each circle shows the observed expression state, with the arrow pointing at the position of the future state, extrapolated from velocity estimates.

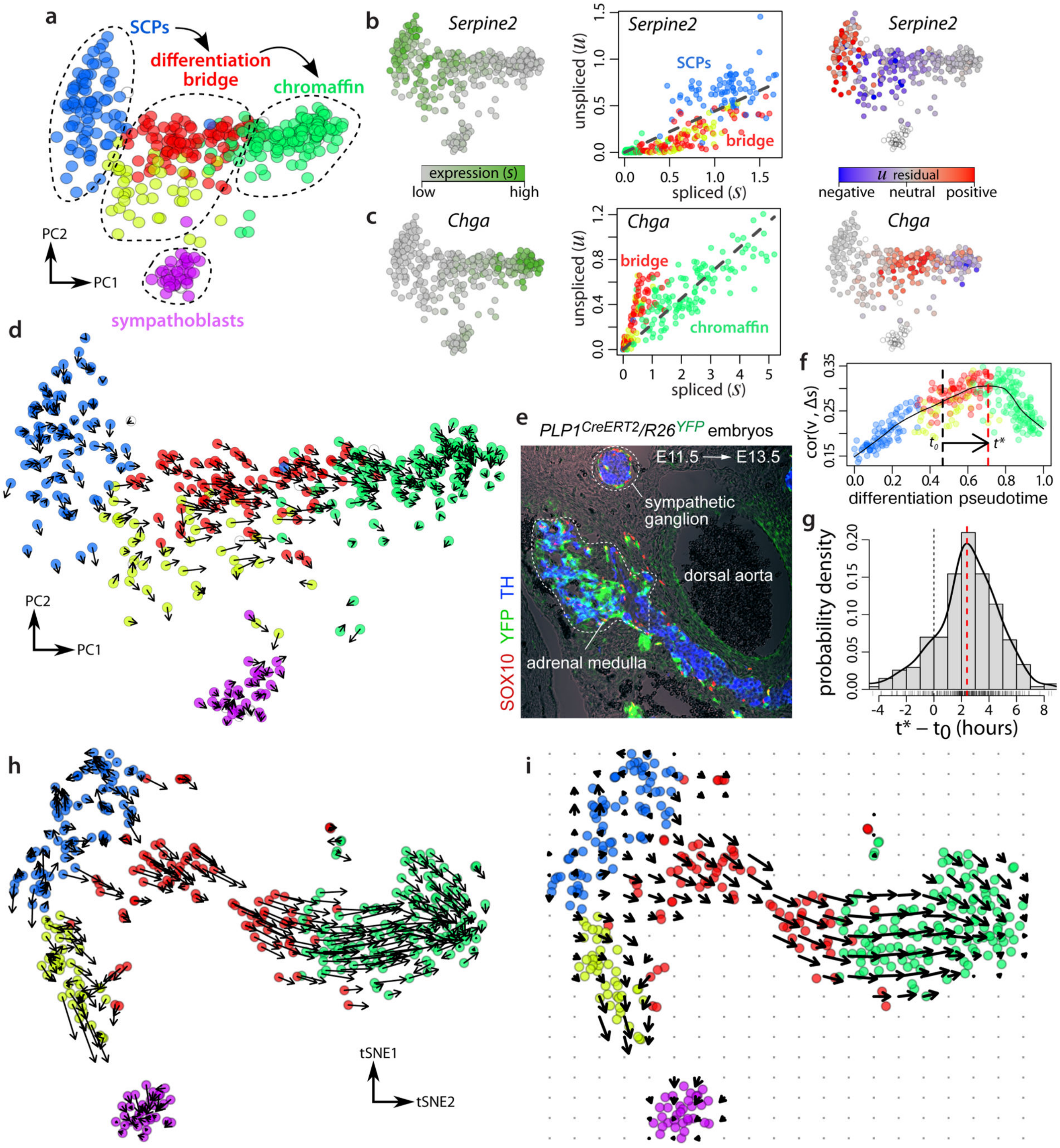


Figure 2. RNA velocity recapitulates dynamics of chromaffin cell differentiation.

a. PCA projection showing major subpopulations of Schwann cell precursors (SCPs) differentiating into chromaffin cells in E12.5 mouse (n=385 cells).

b,c. Expression pattern (left), unspliced/spliced phase portraits (center, cells colored according to a), and *u* residuals (right) are shown for the repressed *Serpine2* (b) and induced *Chga* (c) genes. Read counts were pooled across $k = 5$ nearest cell neighbors.

d. The observed and the extrapolated future (arrows) states are shown on the first two PCs. RNA velocity was estimated without cell or gene pooling.

e. SCP-to-chromaffin cell transition as evidenced by lineage tracing with SCP-specific PLP1-CreERT2 line. A cross-section through the developing adrenal medulla is shown. Note high proportion of TH+/YFP+ cells in the developing medulla and the absence of such double-positive cells in the sympathetic ganglion (N=3 replicates).

f. Extrapolation distance along the chromaffin differentiation trajectory is estimated for a single cell at pseudotime t_0, based on the correlation (y axis) between the velocity v and cell expression difference. Red line shows optimal extrapolation time (t^*) (see Supplementary Note 2 Section 6).

g. Distribution of optimal extrapolation times ($t^*- t_0$) for the chromaffin differentiation timecrouse. Red line marks the distribution mode (2.1 hours).

h. The velocities are visualized on the pre-defined t-SNE embedding from the original publication13. Velocity estimates based on nearest-cell pooling ($k = 5$) were used.

i. Same velocity field as (h) visualized using Gaussian smoothing on a regular grid.

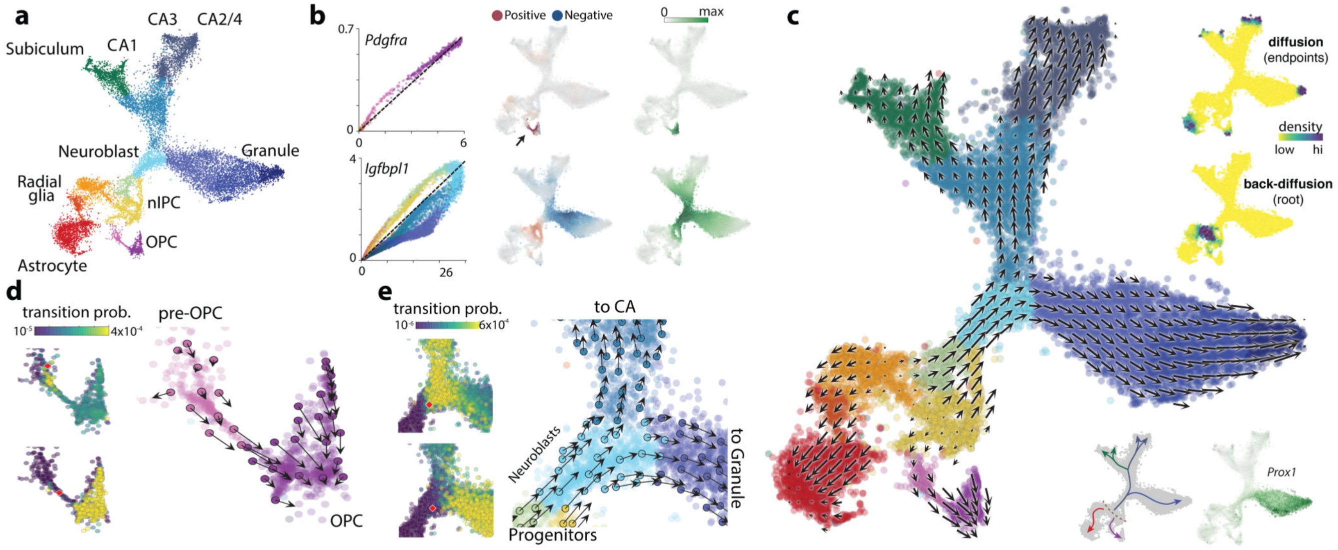


Figure 3. RNA velocity field describes fate decisions of major neural lineages in the hippocampus.

a. A t-SNE embedding of the developing mouse hippocampus cells (n=18,213 cells), showing major transient and mature subpopulations.

b. Phase portraits (left, colored as in a), unspliced residuals (middle), and spliced expression (right) are shown for two regulated genes. kNN cell pooling was used.

c. Velocity field projected onto the t-SNE embedding. Arrows show the local average velocity evaluated on a regular grid. Upper right insert: differentiation endpoints as high density regions on the manifold after forward Markov process with velocity-based transition probabilities; the root of the branching tree is identified simulating the process in the reverse direction. Lower right insert: Summary schematic of the RNA velocity field, and expression of the transcription factor *Prox1.*

d. Commitment to oligodendrocyte fate. Left, visualization of single-step transition probabilities from two starting cells (red) to neighbouring cells. Right, velocities of a sampled subset of cells shown on the t-SNE embedding in (c).

e. Fate decision of neuroblasts. Left, visualization of single-step transition probabilities from two starting cells (red) to neighbouring cells. Right, velocities of a sampled subset of cells shown on the t-SNE embedding in (c).

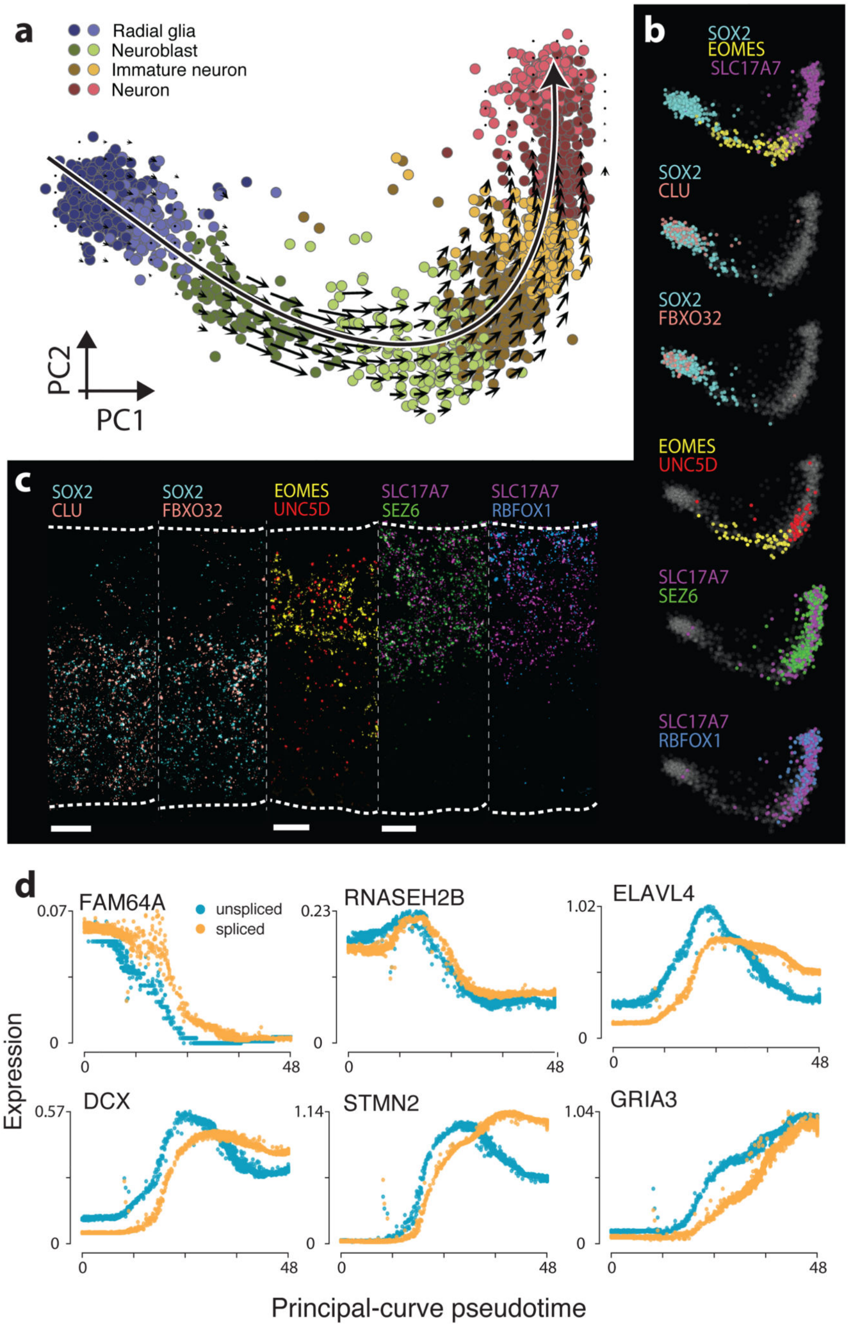


Figure 4. Kinetics of transcription during human embryonic glutamatergic neurogenesis.

a. PCA projection of human glutamatergic neuron differentiation (n=1,720 cells) at post-conception week 10, shown with velocity field. Colors indicate cell types and intermediate states. A corresponding principal curve is shown in bold.

b. Gene expression for known markers of radial glia (*SOX2*), neuroblasts (*EOMES*) and neurons (*SLC17A7*) and for novel markers are visualized on the PCA projection as in indicated genes in pseudocolor.

c. Fluorescent *in situ* hybridization (RNAscope) for the same genes as in (b) on a cross-section of human developing cortex, oriented with the ventricular zone towards the bottom and the cortical surface towards the top (N=1). Scale bars, 25 μm.

d. Pseudotime expression profiles for six example genes regulated in glutamatergic neuron maturation. Unspliced abundance was divided by γ to match the scale of spliced abundance.