
Supplementary information

Molecular design of hypothalamus development

In the format provided by the authors and unedited

Roman A. Romanov, Evgenii O. Tretiakov, Maria Eleni Kastriti, Maja Zupancic, Martin Häring, Solomiia Korchynska, Konstantin Popadin, Marco Benevento, Patrick Rebernig, Francois Lallemend, Katsuhiko Nishimori, Frédéric Clotman, William D. Andrews, John G. Parnavelas, Matthias Farlik, Christoph Bock, Igor Adameyko, Tomas Hökfelt, Erik Keimpema & Tibor Harkany✉

Supplementary information

Molecular design of hypothalamus development

In the format provided by the authors and unedited

Roman A. Romanov, Evgenii O. Tretiakov, Maria Eleni Kastriti, Maja Zupancic, Martin Häring, Solomiia Korchynska, Konstantin Popadin, Marco Benevento, Patrick Rebernig, Francois Lallemend, Katsuhiko Nishimori, Frédéric Clotman, William D. Andrews, John G. Parnavelas, Matthias Farlik, Christoph Bock, Igor Adameyko, Tomas Hökfelt, Erik Keimpema & Tibor Harkany✉

1. Integral data analysis: quality control, data filtration, integration and clustering

We addressed the specificity of cellular differentiation programs that generate hypothalamic cell pools by parallel scRNA-seq (10x Genomics) on dissociated and then methanol-fixed cells at embryonic days (E)15.5, E17.5, at birth, and postnatal day (P)2, P10 and P23 (**Figure S1**). In sum, a total of 51,199 cells of ectodermal origin (that is, excluding microglia, endothelial cells, pericytes and smooth muscle cells) were analysed at these points as follows: 8,290/E15.5; 11,213/E17.5; 7,492/P0; 12,824/P2; 8,965/P10 and 2,415/P23. Data were quality controlled and filtered as shown in **Figures S2-S5**.

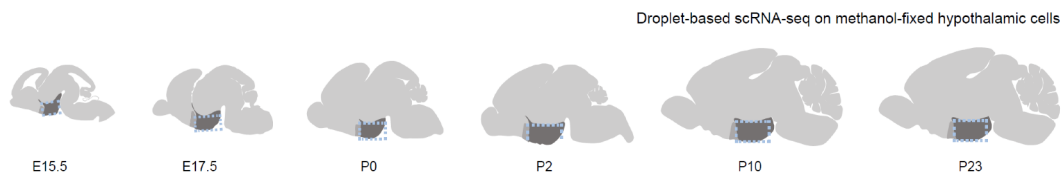


Figure S1. scRNA-seq of dissociated hypothalamic cells. Droplet-based scRNA-seq was performed at consecutive stages of hypothalamus development (E, embryonic; P, postnatal). Grey rectangle and shading demarcates the hypothalamus; the site of cell collection for methanol fixation and subsequent processing.

The *dropEst* work-flow was used for quality control and filtering based on its superior resolution relative to other algorithms (**Figure S5**).

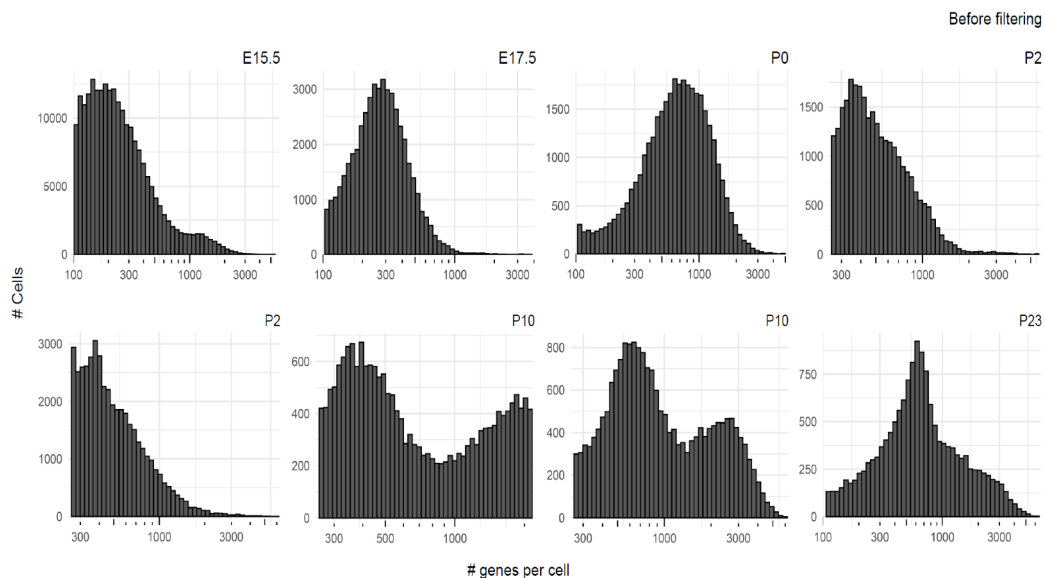


Figure S2. Distribution of the number of genes per cell before filtering of 10x scRNA-seq data. Histograms correspond to the developmental stages specified (E, embryonic; P, postnatal days).

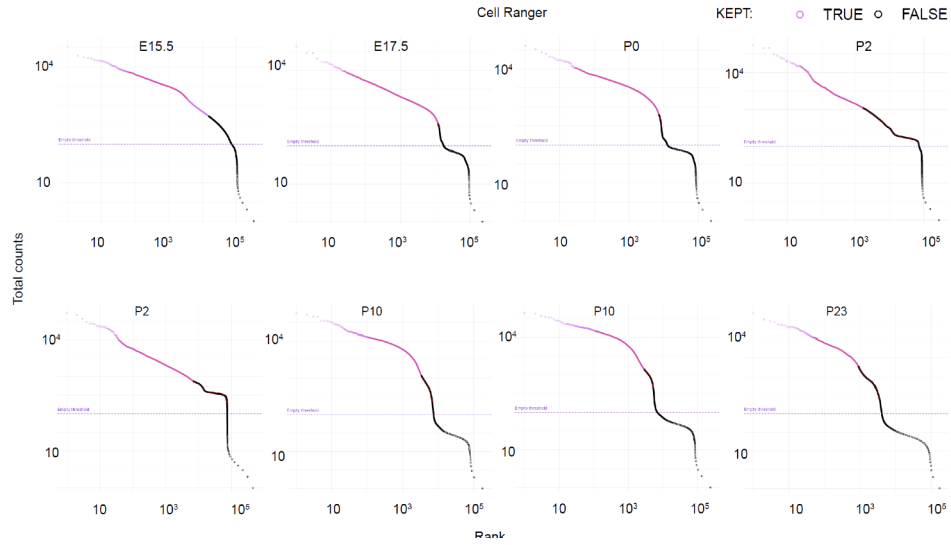


Figure S3. Default Cell Ranger filtering procedure for 10x Genomics scRNA-seq samples.

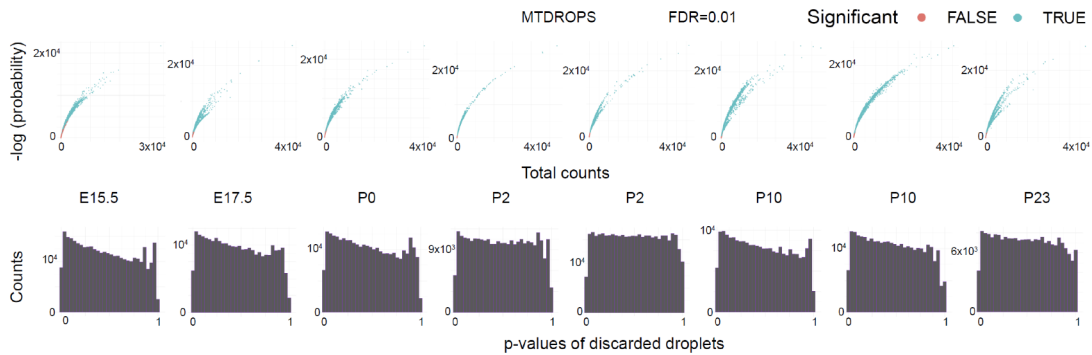


Figure S4. *emptyDrops* method. *Top:* This approach was used for the iterative estimation of probabilities to discard cells. *Bottom:* The uniform distribution of p values for discarded cells is shown. To filter droplets that did not contain cells, we analyzed all cellular barcodes identified in non-zero droplets: E15.5 – 421,673; E17.5 – 297,896; P0 – 316,449; P2 – 367,646 and 532,658; P10 – 241,249 and 312,670; P23 – 217,430. The *EmptyDrops* method that utilizes Monte Carlo simulations was used to calculate exact p values. We repeated simulation for 200,000 iterations. We corrected for multiple testing by controlling false discovery rate (FDR, set to a threshold of 0.01) using the Benjamini-Hochberg method.

Sequencing data on all cells were integrated into a single matrix to assign lineage proto-groups using *Seurat3.1* alignment (based on canonical correlation analysis (CCA)) along both fetal and postnatal developmental stages (**Figures S6 and S7**). We have also validated the greater robustness of this approach by coincidentally resolving a mixture of temporal and cell type segregation criteria by comparison to other algorithms: balanced batch k-nearest neighbours (*BBKNN*), mutual nearest neighbour (*MNN*), *Conos*, *Liger*, *Harmony* and *Scanorama* (**Figure S6**).

DROPEST - used in our pipeline for cell quality estimation and filtering

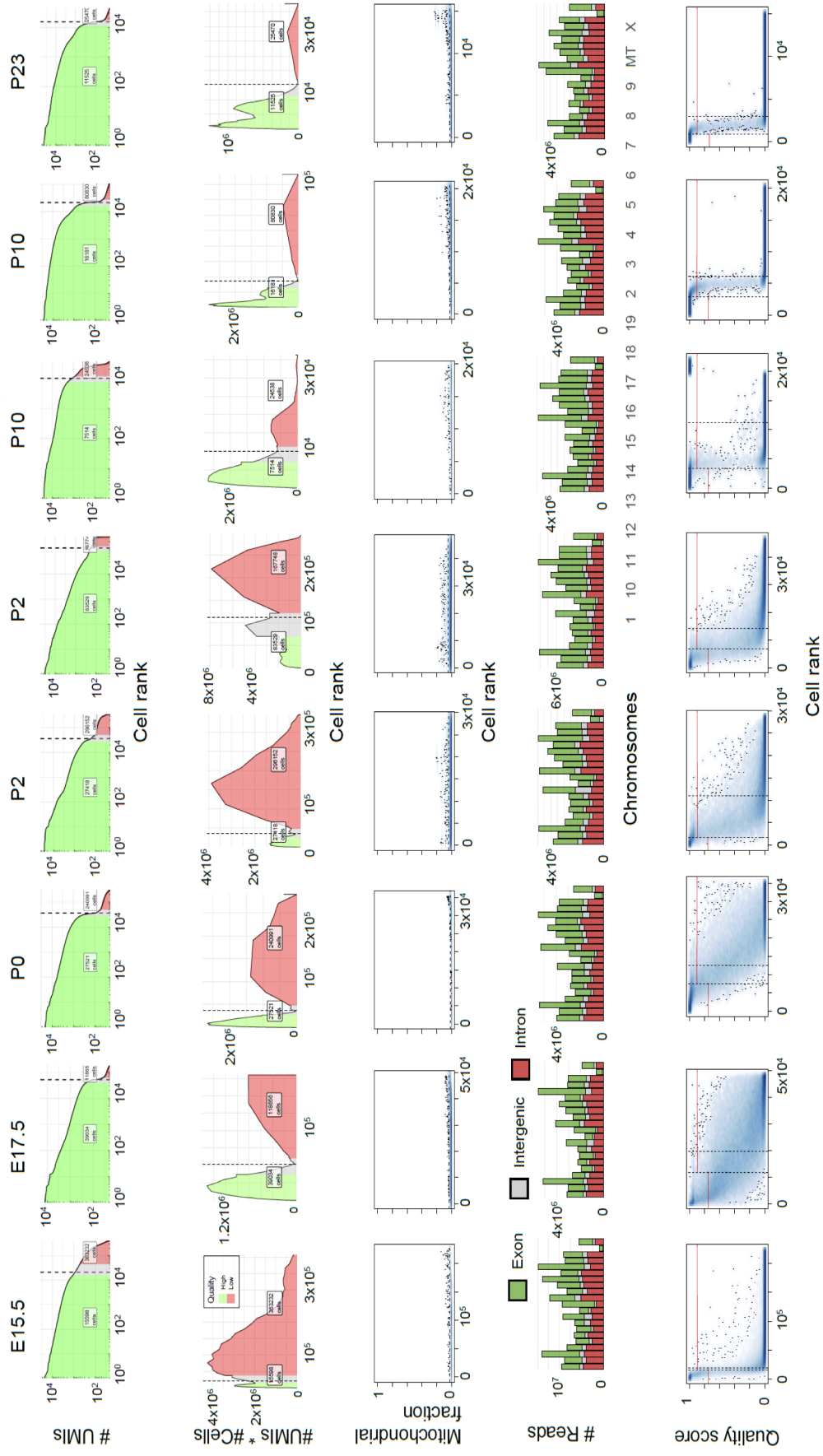


Figure S5 (previous page). Reported output of the *dropEst* pipeline. *dropEst* utilizes the Bayesian correction of cellular barcodes and UMIs, also taking into account Hamming distance on multiple metadata sources. These include sequencing quality of a nucleotide in position (*Phred* score) and estimate damaged and low-quality cells using the kernel density estimate (*KDE*) classifier. Plots show default results of the algorithm on each sample. High-quality cells are in green while low-quality cells are marked red. Low level of mitochondrial fraction (3rd row) and distribution of reads on chromosomes (4th row) confirmed satisfactory data quality and the absence of contamination. On “*Quality score*” vs. “*Cell rank*” plots, red lines are *dropEst* thresholds, dashed lines to the left and right are the *CellRanger* threshold and manually counted number of cells used to prepare 10x libraries.

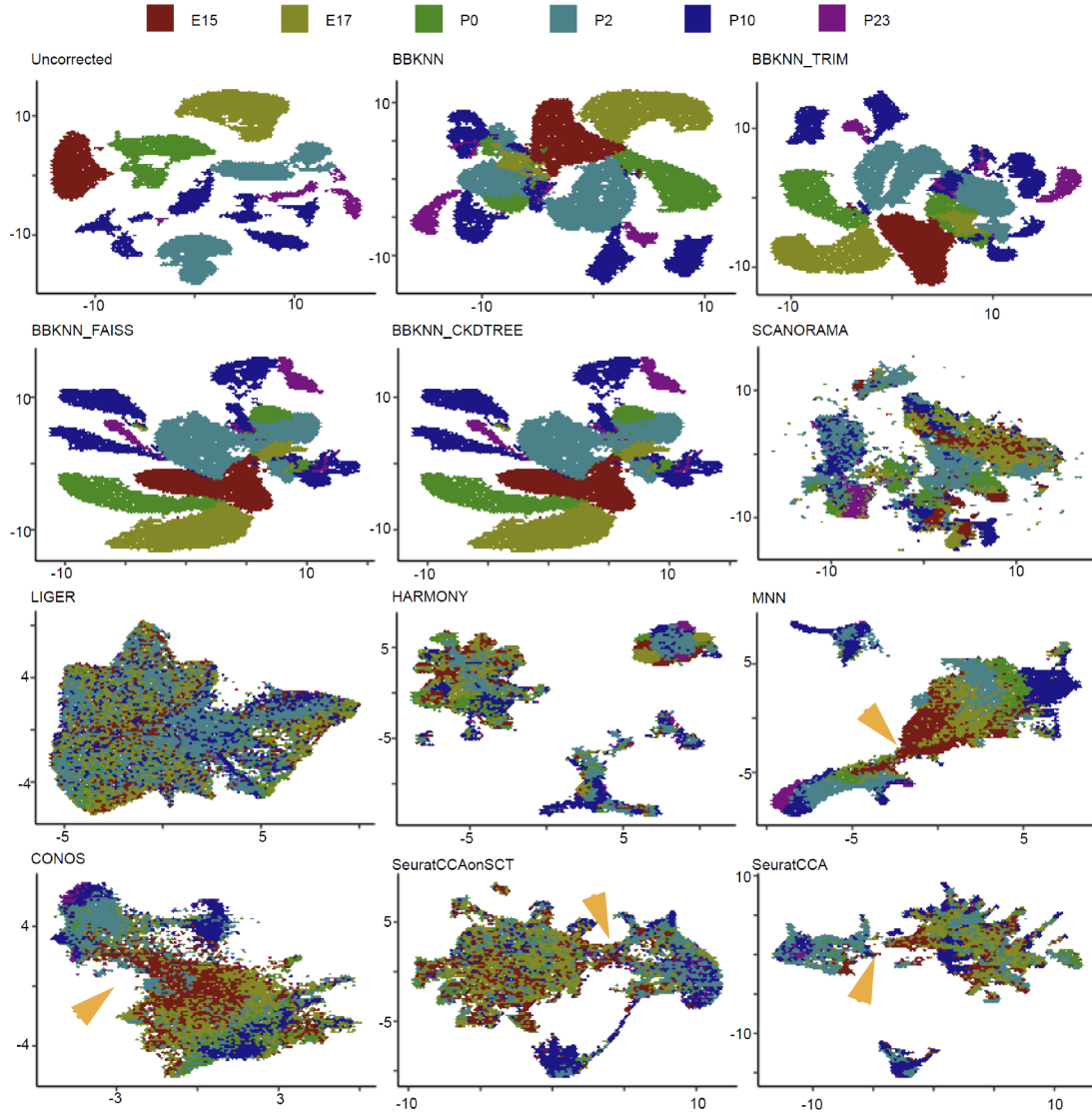


Figure S6. Comparison of presently available algorithms for data integration and batch correction. UMAP embeddings on manifolds integrated across developmental stages were color-coded to define stages of hypothalamus development. Orange arrowheads indicate the “cell bridge”, which we were able to clearly delineate using multiple approaches, such as CCA alignment, *MNN* and *Conos* integration, *RNA-velocity* at discrete time-points and *PAGA* on a multidimensional integrated dataset. We used the same dataset of $n = 51,199$ ectodermal cells for each alignment algorithm.

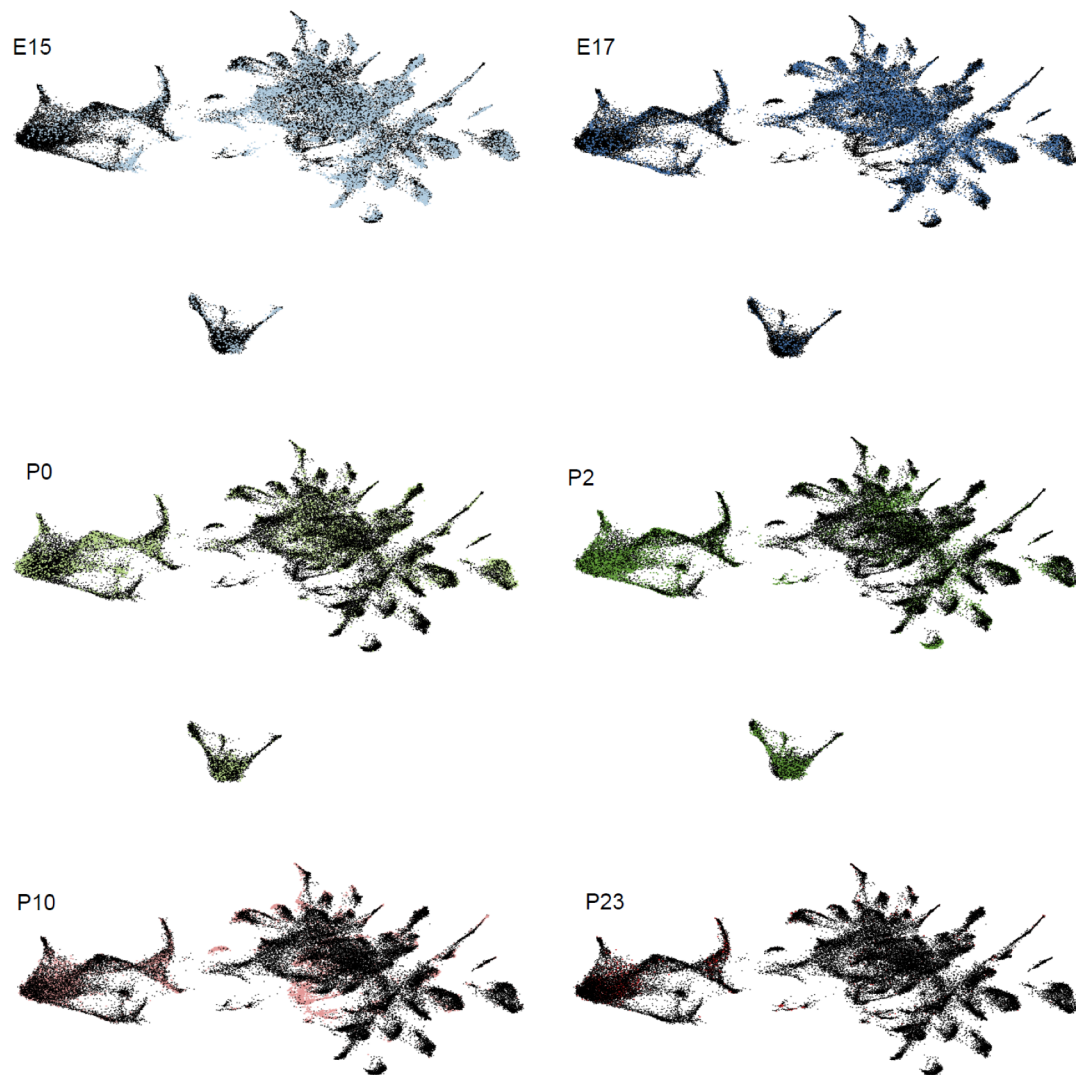


Figure S7. Seurat 3 CCA alignment. This procedure in UMAP space was selected because this algorithm resolved developmental stages and, simultaneously, kept their local structures. Separated prenatal and postnatal stages are shown with subsets of data colored to improve clarity of the distribution of age-specific data subsets. The sequential increase in data complexity as a factor of age substantiated the correctness of CCA matrix-based clustering. Accordingly, each proto-group was defined by unique sets of differentially-expressed genes, notably TFs.

Next, we used *MetaNeighbor* to compare our results with cellular annotations that exist for adult hypothalamic scRNA-seq data (**Figure S8**) even though we found an unexpected lack of consensus in published literature.



Figure S8. Cellular annotations in the adult hypothalamus. Cross-comparison of the classification power of published datasets. Colors correspond to studies as specified in the legend.

Therefore, we have selected a recent dataset (Zeisel *et al.* “Molecular Architecture of the Mouse Nervous System” *Cell* 2018, 174:999-1014.) as reference because of its completeness of anatomical sampling. Integration of these and our developmental data

revealed a close intersection at the diversity of terminally-differentiated progenies (**Figure S9a**).

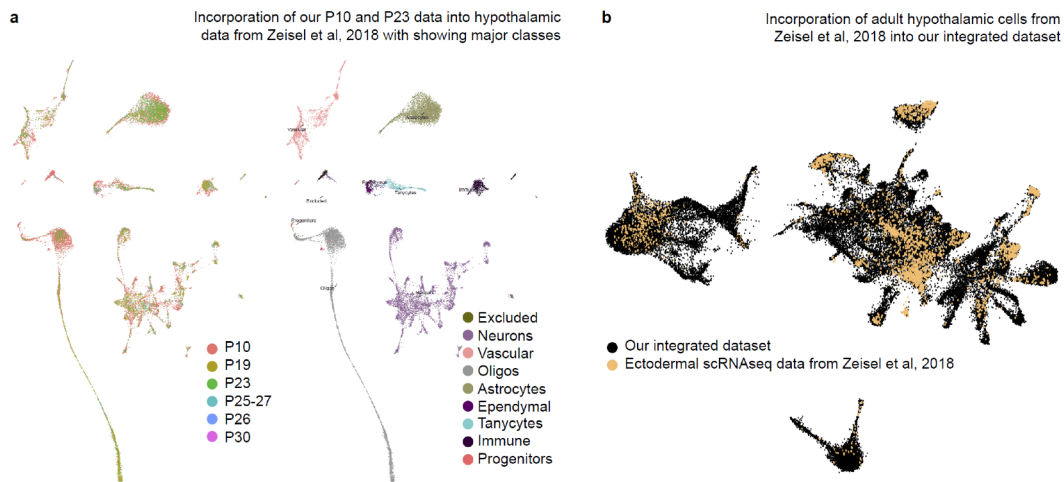


Figure S9. Integrating developmental biology data with an adult reference dataset. (a) CCA-based integration of cells at the juvenile stage into an adult hypothalamic dataset by Zeisel *et al.* **(b)** CCA-based integration of the same adult dataset into our developmental manifold.

At the same time, we showed uncovered gaps for progenitors and transient (~47% of all cells) cell states (**Figure S9b**), which necessitated the introduction of an unbiased strategy of annotation for which different algorithms were compared (**Figure S10**). Here, we used an approach initially developed for paired embryonic/adult analysis (*SCTransform*), which combines consistent properties of cell lineages derived from a single integrated CCA matrix through the *Walktrap* algorithm but along sequential time points and multiple cell cohorts.

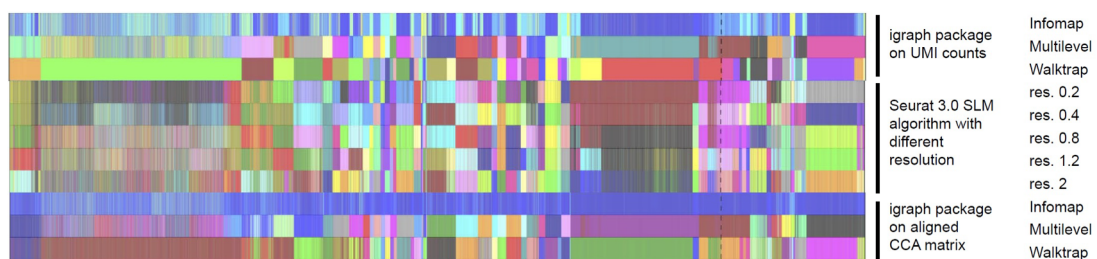


Figure S10. Comparison of clustering precision of the integrated manifold by different published algorithms.

To define the relationship between precursors and terminally-differentiated neuronal proto-groups, we have re-partitioned the CCA matrix with a Leiden algorithm (Traag *et al.* “From Louvain to Leiden: guaranteeing well-connected communities.” *Scientific reports* 2019, 9) and deployed a partition-based graph abstraction (*PAGA*) method (Wolf *et al.* “PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells.” *Genome biology* 2019, 20:59) to define

branching nodes (**Figure S11**). Our attempt to re-partition the CCA matrix with a Leiden algorithm confirmed that neuroblasts in cluster #19 ('bridge cells') remain invariably separated as a distinct cell group.

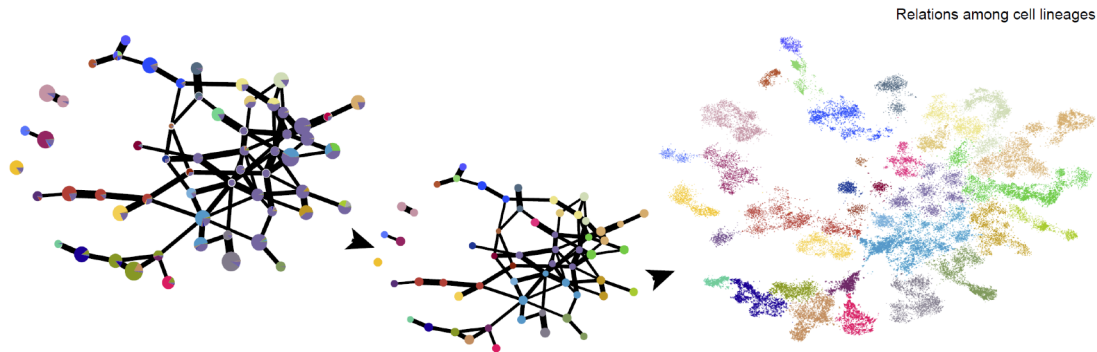


Figure S11. The topological structure of cells produces a guide to repartition immature cells into adult groups by PAGA and Leiden algorithms. A Leiden algorithm on the CCA matrix was used to assign intermediate and immature cells to terminally-differentiated cell groups, which were color-coded as in **Figure 1a**. Of note, 'bridge cells' (light blue) that enter a neuronal differentiation program appear as a distinct group. Immature neurons (purple) were mostly recruited to final groups, and ependymal cells readily split off from astrocytes.

Additionally, we applied *RNA-velocity* (**Figure ED2b**) and the probabilistic *Palantir* algorithm (Setty *et al.* "Characterization of cell fate probabilities in single-cell data with Palantir." *Nature Biotechnology* 2019, **37**:451-460; **Figure S12**) to construct pseudotime scales to recapitulate developmental trajectories and transition states, and to estimate gene expression-based developmental dynamics. Lastly, the *Scenic* work-flow allowed us to map the diversity and complexity of regulatory gene networks ('regulons') with particular focus on their relationship to phenotypic perturbations when correlating their regulatory constituents with genome-wide association studies (UK Biobank). In particular, we have validated 'master regulators' that have the power to steer cell fate decisions towards astroglial *vs.* neuronal specificities and neuronal subgroups. Thus, our methodological repertoire not only encompasses the most advanced analysis tools available to date but also provides multiple levels of quality control both by bioinformatics and experimental neurobiology.

2. Identification of dopamine cell groups and reconstruction of their developmental trajectories

We explored if hypothalamic dopamine neurons share, and if so to what extent, a developmental trajectory. Because the ability to produce, accumulate (in synaptic vesicles) and release dopamine are pivotal to assign any neuron as dopaminergic, we relied on the co-expression of *Th*, dopa decarboxylase/*Ddc* and vesicular amine transporter 2/*Slc18a2* as classifier. Otherwise, neurons with partial and permuted subsets of these genes far outnumber neurons that can be considered as capable of initiating and terminating dopamine neurotransmission. Firstly, we embedded the RNA-velocity vectors of all tyrosine hydroxylase (*Th*)⁺ cells, which unequivocally identified 10 molecularly far-placed *Th*⁺ neuronal groups. Herein, we recapitulated published literature by showing that *Avp*⁺/*Oxt*⁺/*Pou3f2*⁺ magnocellular command neurons (#10) were also *Th*⁺ (Meister *et al.* “Localization of chemical messengers in magnocellular neurons of the hypothalamic supraoptic and paraventricular nuclei: an immunohistochemical study using experimental manipulations.” *Neuroscience* 1990, 37:603-633). Nevertheless, and based on their molecular, positional and functional differences from the other 9 parvocellular *Th*⁺ groups, we used *Avp*⁺/*Oxt*⁺/*Pou3f2*⁺ neurons as positive controls throughout. We then subjected *Th*⁺ neuronal groups to the *Palantir* probabilistic workflow to construct their pseudotime scales (**Figure S12**) for the determination of their differentiation potential (if any, **Figure S12**) and to reconstruct their individual developmental trajectories.

Direct translation of pseudotime data into age and real-time (**Figure S12**) pinpointed substantial differences in the differentiation potential of dopamine neurons: *Th*⁺ neurons in groups #4 (*Ghrh*), #7 (*Six6/Prhr*), and #8 (*Lhx6/8*) were stationary without differentiation potential throughout our window of analysis, suggesting that they became terminally differentiated prior to E15.5 (**Figure S12,13**). Nevertheless, they retained the ability of age-dependent gene enrichment for, e.g., neurotransmitters and neuropeptides (**Figure S13**). In contrast, groups #1 (*Ascl1/Dll1/Gal*^{low}), #2 (*Npy/Sst*), #3 (*Gal*), #5 (*Arx/Pax6/Onecut2*), #6 (*Meis2*) and #9 (*Onecut3*) showed high differentiation potential at E15-P0 (**Figure S12**) and demonstrated age-dependent expressional dynamics for TFs (**Figure S13**). These findings suggest the lack of terminal differentiation until the end of the embryonic period. Particularly, and even if we identified heterogeneity within each group, *Th*, *Ddc* and *Slc18a2* exhibited dynamic enrichment in all groups (except #5). These conclusions were also supported by negative findings through the probabilistic analysis of branching (**Figure S12**): groups #2, #3, #6 and #8 failed to further sub-diversify, thus supporting their unequivocal commitment towards a dopaminergic phenotype.

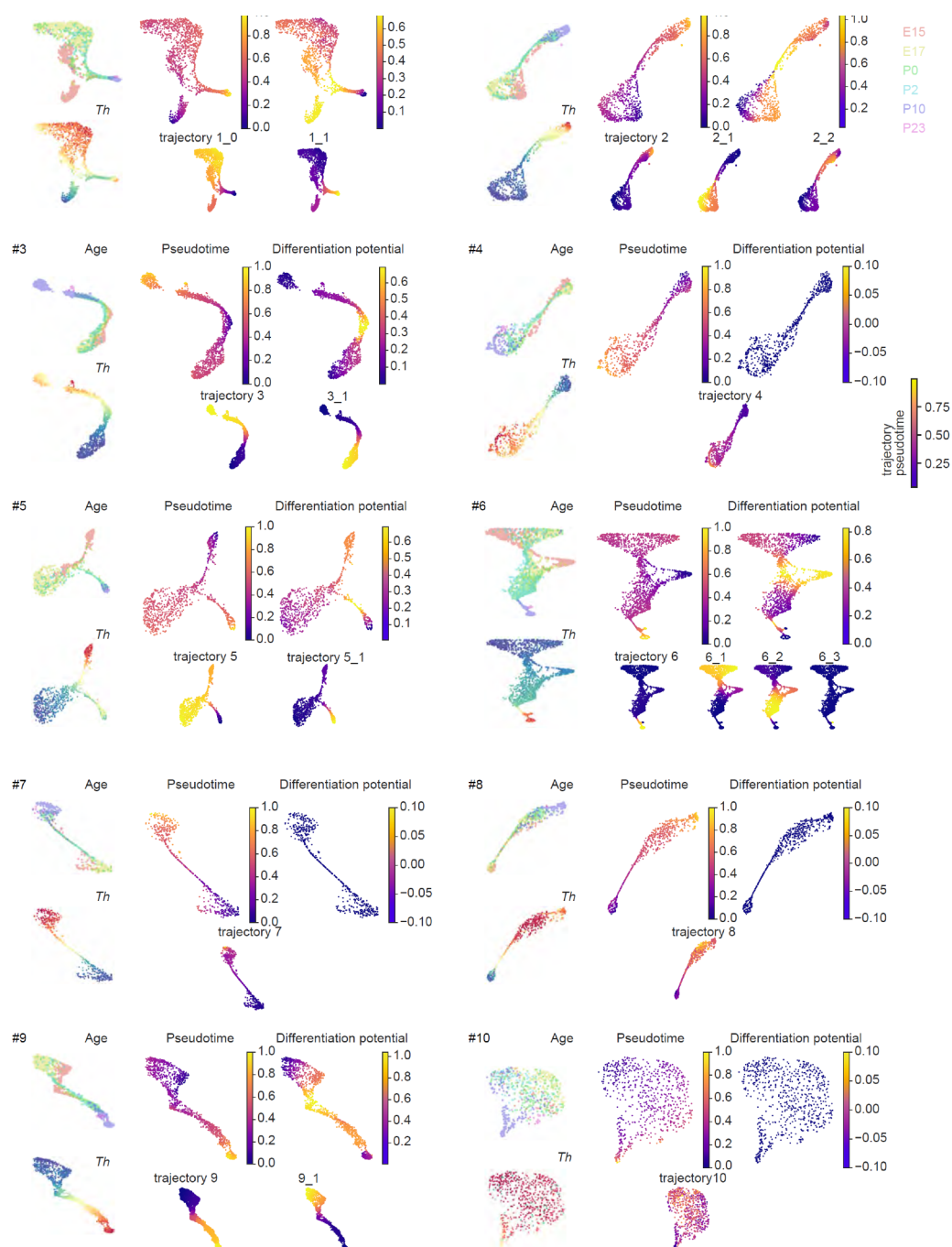


Figure S12. Pseudotime analysis and identification of trajectories for dopamine neurons (in *Palantir*) in tyrosine hydroxylase (*Th*)-containing neuronal groups. Data show estimated pseudotime, differentiation potential, *Th* expression, age contribution and developmental trajectories for 10 *Th*-containing lineages. Minimal-to-maximal mRNA expression for *Th* was normalized separately in each group investigated. Note that groups #4 (*Ghrh*), 7 (*Six6/Prhr*), 8 (*Lhx6/8*) and 10 (*Oxt/Avp/Pou3f2*) lack differentiation potential suggesting their existence prior to E15.5.

Thirdly, we have interrogated the expression of *Slc6a3*, coding for the plasmalemmal dopamine (reuptake) transporter for the removal of extracellular dopamine: *Slc6a3* was invariably expressed by groups #7 (*Six6/Prhr*, A12) and #9 (*Onecut3*, A14), temporarily

in #6 (*Meis2*, A13) and sporadically in #3 (*Gal*, likely A12/14; **Figure S13**). These data were confirmed by life-time genetic tracing in *Slc6a3*-Ires-Cre:R26^{TdTomato} mice, including fluctuating *Slc6a3*-driven *tdTomato* signal in A13 neurons. Thus, we have dissected both general and specific genetic programs generating molecular diversity among hypothalamic dopamine neurons.

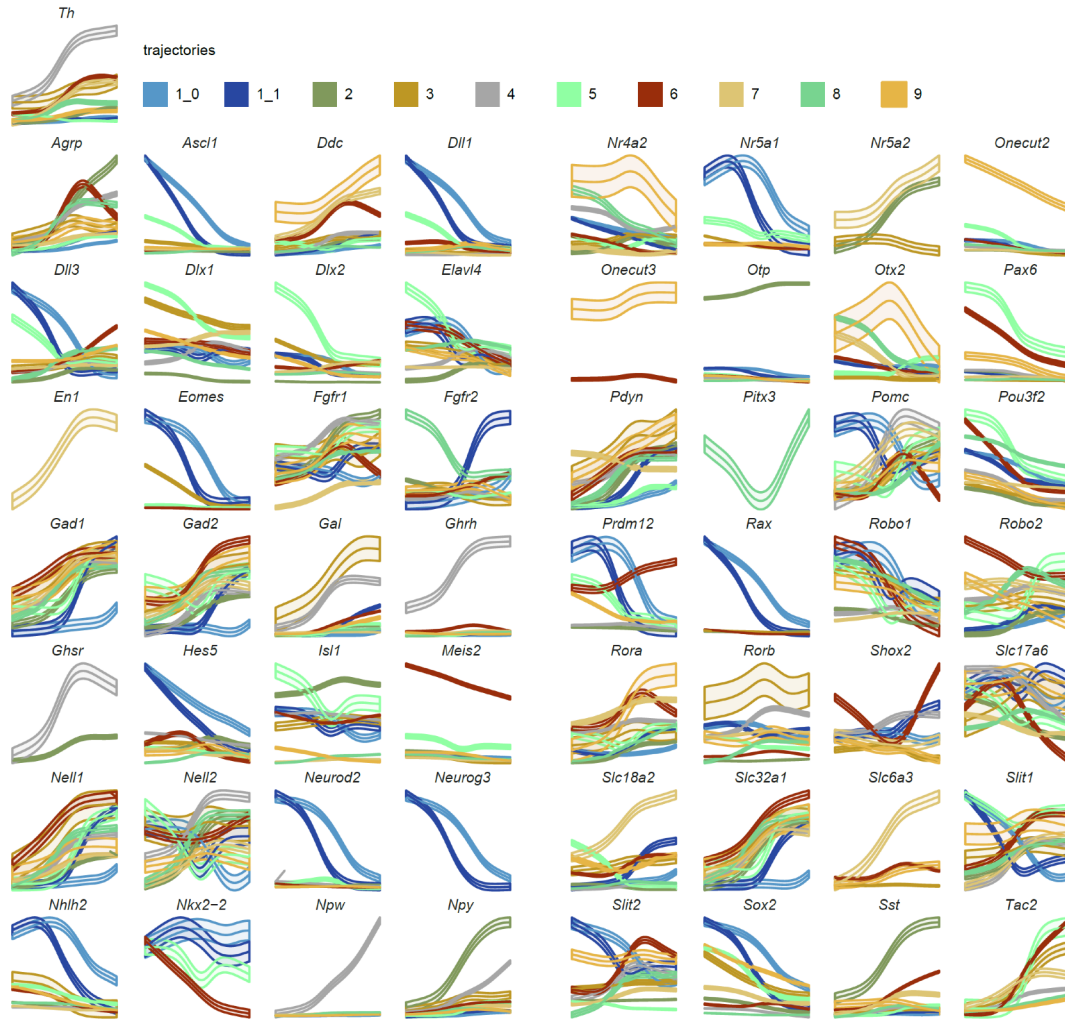


Figure S13. Pseudotime gene expression for dopamine lineages. Fifty-three genes had been manually selected and include markers of neurotransmission, neuropeptides and TFs. Dynamic transcriptional trends of differentiation are exemplified for trajectories in groups 1-9. Amplitudes are shown in log₁₀ scale, and line shades correspond to means $\pm$ s.e.m. To compute gene trends, we independently sampled differentiation trajectories containing *Th*⁺ positive neurons (1 – 1,506, 2 – 997, 3 – 1,453, 4 – 696, 5 – 948, 6 – 1,779, 7 – 538, 8 – 649, 9 – 1,181) to 500 bins. For each bin means $\pm$ s.e.m was estimated by generalized additive models.

3. Additional remarks to source data on Figshare:

For Supplementary File 5 (DOI:10.6084/m9.figshare.11867889), we used the whole dataset of $n = 51,199$ ectodermal cells testing each cell group against background. Adjusted p values were based on Bonferroni correction using all genes in the dataset.

For Supplementary File 6 (DOI:10.6084/m9.figshare.11867889), $n = 33,893$ cells in total were used. Two-tailed Wilcoxon rank sum test was used.