Methods

Evgenii O. Tretiakov

2023-06-12

# Presentation  
library("glue")  
library("knitr")  
library("reticulate")  
  
# JSON  
library("jsonlite")  
  
# Tidyverse  
library("tidyverse")

if (!file.exists(here::here("output/methods/packages.bib"))) {  
 dir.create(here::here("output", DOCNAME), showWarnings = FALSE)  
  
 write\_bib(  
 c(  
 "base", "Seurat", "SeuratWrappers", "SeuratDisk", "sctransform",  
 "glmGamPoi", "patchwork", "scCustomize", "Nebulosa", "clustree",  
 "mrtree", "gprofiler2", "cowplot", "UpSetR", "ggstatsplot",  
 "gridExtra", "tidyverse", "dplyr", "tidyr", "magrittr", "stringr",  
 "skimr", "future", "purrr", "here", "workflowr", "zeallot", "knitr",  
 "kableExtra", "rmarkdown", "reticulate"  
 ),  
 file = here::here("output/methods/packages.bib")  
 )  
}

versions <- list(  
 biomaRt = packageVersion("biomaRt"),  
 cellbender = "docker://etretiakov/cellbender:v0.0.1",  
 cellranger = "7.1.0",  
 cellranger\_ref = "mm10\_optimized\_v.1.0",  
 clustree = packageVersion("clustree"),  
 cowplot = packageVersion("cowplot"),  
 dplyr = packageVersion("dplyr"),  
 future = packageVersion("future"),  
 ggplot2 = packageVersion("ggplot2"),  
 ggstatsplot = packageVersion("ggstatsplot"),  
 glmGamPoi = packageVersion("glmGamPoi"),  
 gprofiler2 = packageVersion("gprofiler2"),  
 gridExtra = packageVersion("gridExtra"),  
 here = packageVersion("here"),  
 quarto = packageVersion("quarto"),  
 kableExtra = packageVersion("kableExtra"),  
 knitr = packageVersion("knitr"),  
 leidenalg = "0.9.1@pypi\_0",  
 magrittr = packageVersion("magrittr"),  
 mrtree = packageVersion("mrtree"),  
 Nebulosa = packageVersion("Nebulosa"),  
 pandoc = rmarkdown::pandoc\_version(),  
 patchwork = packageVersion("patchwork"),  
 purrr = packageVersion("purrr"),  
 pyroe = "docker://combinelab/usefulaf:0.9.3",  
 python = "3.8.8",  
 R = str\_extract(R.version.string, "[0-9\\.]+"),  
 reticulate = packageVersion("reticulate"),  
 rmarkdown = packageVersion("rmarkdown"),  
 scCustomize = packageVersion("scCustomize"),  
 sctransform = packageVersion("sctransform"),  
 Seurat = packageVersion("Seurat"),  
 simpleaf = "docker://etretiakov/usefulaf:0.9.0",  
 skimr = packageVersion("skimr"),  
 stringr = packageVersion("stringr"),  
 Snakemake = "7.21.0",  
 scrublet = "0.2.3@pyh5e36f6f\_1",  
 tidyr = packageVersion("tidyr"),  
 tidyverse = packageVersion("tidyverse"),  
 UpSetR = packageVersion("UpSetR"),  
 viridis = packageVersion("viridis"),  
 workflowr = packageVersion("workflowr"),  
 zeallot = packageVersion("zeallot")  
)

# Mouse strains (WIP SECTION)

All mice were housed in groups in clear plastic cages on a 12h–12h light–dark cycle (lights on at 08:00 h) and in a temperature (22 ± 2 °C) and humidity (50 ± 10%) controlled environment. Food and water were available ad libitum. Embryos and tissues were obtained from timed matings with the day of vaginal plug considered as embryonic day (E) 0.5. The day of birth was always registered as postnatal day (P)0. Postnatal animals were weaned on P21. Commercial mouse lines were: C57Bl/6J wild-type (RRID:IMSR\_JAX:000664), Ai14 (RRID:IMSR\_JAX:007914), Ascl1-creERT2 (RRID:IMSR\_JAX:012882), Th–Gfp (RRID:IMSR\_RBRC03162), (BAC)GAD65–eGFP (RRID:MMRRC\_011849-UCD), GAD67 (RRID:IMSR\_RBRC03674), Pomc–Gfp (RRID:IMSR\_JAX:009593), Slc6a3-Ires-cre (RRID:IMSR\_JAX:006660), Nfia−/− (RRID:MMRRC\_010318-UNC), Robo1−/− (RRID:IMSR\_APB:5320), Slit1−/− (RRID:MMRRC\_030404-MU), Slit2−/− (RRID:MMRRC\_030405-MU), Isl1-cre (RRID:IMSR\_JAX:024242) and OxtrVenus/+ (MGI:3838764) (REF ). Ascl1-creERT2 knock-in mice were used as heterozygotes when performing lineage tracing and as homozygotes to study developmental consequences of the lack of Ascl1 since both copies of the gene were replaced by the Cre coding region (referred to as Ascl1 ko). Tracing experiments for all other Cre lines were performed using heterozygotes. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

# Tissue collection and fixation

Dissected brains were collected and fixed in 4% paraformaldehyde (PFA) in phosphate-buffered saline (PBS, 0.05 M, pH 7.4) at 4 °C for 24 h. Animals were transcardially perfused with 4% PFA in 0.1 M phosphate buffer (PB; pH 7.4) and dissected brains post-fixed overnight. Samples were then washed in PB and cryoprotected by incubating in 30% sucrose in distilled water at 4 °C overnight.

# Ethical approval of animal studies

Experiments on live animals conformed to the 2010/63/EU European Communities Council Directive and regulated by applicable local laws (Tierversuchsgesetz 2012, BGBI, Nr. 114/2012 (Austria)) and were approved by the Austrian Ministry of Science and Research (66.009/0145-WF/II/3b/2014, and 66.009/0277-WF/V/3b/2017). Particular effort was directed towards minimizing the number of animals used and their suffering during experiments.

# Tamoxifen injection and tissue processing (WIP SECTION)

Acutely stressed and control *“TRAP”* mice (formalin stress induced by injection of 4% PFA into the left paw; REF Alán Alpár EMBOJ) experimental groups were injected with tamoxifen (150 mg/kg) to induce Cre-mediated recombination. We used a 6h time point because global changes in mRNA peak ∼5–9 h after experimental manipulation. After 72h to accumulate *ZsGreen* expression in “stress-responder” astrocytes, animals were transcardially perfused with a 50–100 ml of fixative containing 4% PFA in 0.1 M phosphate buffer (PB, pH 7.4), then brains were collected and immersion fixed in 4% PFA in PB (pH 7.4) for 12–24h before being immersed into 30% sucrose for cryoprotection (48 h). Brains were then cut on a cryostat as 50-μm-thick serial free-floating coronal sections. In contrast for the circadian activity analysis we used “*TRAP2*” model and 4OH-tamoxifen that allowed to collect more narrow time window of circadian cFos activation.

# Tissue preparation and immunohistochemistry (WIP SECTION)

After rinsing in 0.1 M PB, specimens were exposed to a blocking solution composed of 0.1 M PB, 10% normal donkey serum, 5% BSA and 0.3% TX-100 for 3 h followed by 48 h incubation with select combinations of primary antibodies: rabbit anti-TH (1:500; Millipore AB152, lot 2593900, 3199177), sheep anti-TH (1:1,000, Novus Biologicals, #NB300-110, lot ajo1217p), sheep anti-ONECUT2 (1:250; R&D Systems, AF6294, lot CDKS0116081), guinea pig anti-ONECUT3 (1:5,000)95, rabbit anti-VGLUT2 (1:800; a gift from M. Watanabe)96, goat anti-GFP (1:1,000; Abcam, #ab6662, lot GR311622-15, GR311622-7), chicken anti-GFP (1:500, Aves Labs Inc., #GFP-1020, lot GFP697986), rabbit anti-SOX2 (1:500, Abcam, #ab97959, lot GR3244885-1), chicken anti-mCherry (1:1,000; EnCor Biotech, #CPCA-mCHERRY, lot 7670-4), mouse anti-MASH1 (1:100, BD Pharmingen, 556604, clone: 24B72D11.1), guinea pig anti-GFAP (1:500, Synaptic Systems, 173004, lot 2-15, 2-17), rabbit anti-phospho-histone H3 (1:500; Cell Signaling Technology, 9701, lot 7), chicken anti-NeuN (1:500, Merck Millipore, ABN91, lot 3132967), mouse anti-FLAG-tag (1:1,000; Sigma, F1804, lot SLBR7936V), mouse anti-HA-tag (1:600; Cell Signaling Technology, mAb2367, lot 1). Secondary antibodies were from Jackson ImmunoResearch, including Alexa Fluor 488-AffiniPure donkey anti-goat (705-545-147, lot 131669), Alexa Fluor 488 donkey anti-mouse (715-545-151, lot 127820), Alexa Fluor 488-AffiniPure donkey anti-guinea pig (706-545-148, lot 138058), Alexa Fluor 647-AffiniPure donkey anti-guinea pig (706-605-148, lot 135631), Alexa Fluor 647-AffiniPure donkey anti-rabbit (711-605-152, lot 127614), carbocyanine (Cy)2-AffiniPure donkey anti-rabbit (711-225-152, lot 139999), Cy3-AffiniPure donkey anti-chicken (703-165-155, lot 142225), Cy3-AffiniPure donkey anti-goat (705-165-147, lot 134527), Cy3-AffiniPure donkey anti-guinea pig (706-165-148, lot 134844), Cy3-AffiniPure donkey anti-mouse (715-165-150, lot 116881), and Cy3-AffiniPure donkey anti-rabbit (711-165-152, lot 141941) and applied at a dilution of 1:300 in 0.1 M PB supplemented with 2% BSA (20-22 °C, 2 h). Nuclei were routinely counterstained with Hoechst 33,342 (1:10,000; Sigma). Tissues were photographed on a Zeiss LSM880 laser-scanning microscope. Images were acquired in the ZEN2010 software package. Multi-panel images were assembled in CorelDraw X7 (Corel Corp.).

# Fluorescent in situ hybridization (HCR 3.0) (WIP SECTION)

Staining was performed on fresh-frozen tissue sectioned at 16 μm following the HCR v3.0 protocol for ‘generic sample on the slide’ (Molecular Instruments)99. The pre-treatment of tissue sections included fixation with 4% PFA for 15 min, two washing steps with PBS and dehydration using an ascending EtOH gradient (25%, 50%, 75% and 100%, each step for 5 min with subsequent drying for 15 min). The tissue used for these experiments was obtained from 8-12 weeks old mice. The probes used (Apoe NM\_009696.4; Gfap NM\_001131020; Slit2 NM\_001291227.2; Aldh1a1 NM\_001361503.1; Tafa1 NM\_182808.3; Plcb1 NM\_001145830.1; Sgcd NM\_011891.5; Slc38a1 NM\_001166456.1; Fos NM\_010234.3; Gja1 NM\_010288.3; Snap25 NM\_011428.3; Olig1 NM\_016968.4; Npy2r; Otp) were designed and purchased from Molecular Instruments.

# Assessment of Arc-astrocytic mitochondria impact on High Fat Diet phenotype

## **Animal care**

All experimental procedures were performed in accordance with the Yale Animal Resources Center and Institutional Animal Care and Use Committee policies. Mice were housed in groups of 3–5 at 22 C°–24 C° using a 12-h light/12-h dark cycle. Animals had *ad libitum* access to water and the prescribed diet at all times. Animals were fed a regular chow diet containing 57 % calories from carbohydrates, 34 % calories from protein, and 9 % calories from fat, or an HFD containing 35% calories from carbohydrates, 20% calories from protein, and 45 % protein from fat (Research Diets, US). All experiments were performed in adult male mice at the age of 6-25 weeks. Food intake and body weight were assessed weekly. Representative food intake shown in the figure was measured on the last week of HFD exposure.

## **Stereotaxic virus injection**

Bilateral virus injections were made into the ARC of anesthetized 6/7-week-old male *mfn2flox-flox*mice placed into a stereotaxic apparatus (model 902; David Kopf instruments).  AAV8/GFAP-eGFP (for control mice) or AAV8/GFAP-GFp-Cre (for mfn2GFAP mice) (Virus Vector Core, UNC, US) were applied into each hemisphere (300nl) (coordinates: bregma, anterior–posterior: −1.2 mm, dorsal–ventral: −5.8 mm, lateral: ±0.3 mm) by using an air pressure system (injection time 5 min). After surgery, mice were allowed to recover for 10 days before exposure to HFD. Accurate virus injection into the ARC was verified by analyzing local GFP fluorescence. Mice with 'missed' or 'partial' hits were excluded. Specific AAV expression in ARC was signified by double fluorescence labeling for GFP and GFAP.

## **Body composition**

Lean and fat mass were analyzed with EchoMRI (EchoMRI LLC, US) (Varela, Kim, et al. 2021)

## **Electron Microscopy**

Mice (at least 4 per group) were anesthetized and transcardially perfused with freshly prepared 4% PFA and 0.1% glutaraldehyde, as previously reported (Varela et al. 2017; Varela, Stutz, et al. 2021). After post-fixation overnight, vibratome sections (50 μm) containing the ARC were immunostained with primary antibody anti-POMC (dilution 1:7500, H-029-30, Phoenix Pharmaceuticals,) or anti-GFAP (dilution 1:4500, Sigma). After overnight incubation at room temperature, sections were washed with PB, incubated with biotin-conjugated donkey anti-rabbit IgG or donkey anti-mouse IgG secondary antibody respectively (dilution 1:250, Jackson ImmunoResearch Laboratories) for 2 h, washed again, put in avidin–biotin complex (ABC; Vector Laboratories), and developed with 3,3-diaminobenzidine (DAB). Sections were then osmicated (15 min in 1% osmium tetroxide) and dehydrated in increasing ethanol concentrations. During the dehydration, 1% uranyl acetate was added to the 70% ethanol to enhance ultrastructural membrane contrast. Flat embedding in Durcupan followed dehydration. Ultrathin sections were cut on a Leica ultra-microtome, collected on Formvar-coated single-slot grids, and analyzed with a Tecnai 12 Biotwin electron microscope (FEI) with an AMT XR-16 camera (Varela et al. 2017; Varela, Stutz, et al. 2021).

## **Mitochondria quantification and glia coverage/synaptic inputs analyses**

Hypothalamic sections containing POMC or GFAP immunoreactive cells with a visible nucleus were analyzed by electron microscopy. Mitochondrial cross-sectional area was calculated. For the glia coverage and the synaptic inputs, a blinded investigator scored the number of synapses and the percentage of glia per POMC cell in high-magnification images (>4.800x) (Varela, Kim, et al. 2021; Varela, Stutz, et al. 2021). p ≤ 0.05 was considered statistically significant.

# Pre-processing of single-cell RNA sequencing data

## Downloading data

Raw sequencing data files were obtained from the NCBI Gene Expression Onimbus mirror of European Bioinformatics Institute using IBM Aspera tool-kit (example command - ascp -QT -l 800m -k 1 --overwrite=diff -P 33001 -i ~/asperaweb\_id\_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:/path/to/file/on/server .) or directly from NCBI mirror using ncbi-toolkit based on BioProject ID (example command: esearch -db sra -query PRJNA\*\*\*\*\*\* | efetch -format runinfo | cut -d ',' -f 1 | grep SRR | xargs -n 1 -P 20 prefetch --max-size u && esearch -db sra -query PRJNA\*\*\*\*\*\* | efetch -format runinfo | cut -d ',' -f 1 | grep SRR | xargs -n 1 -P 20 fasterq-dump -p -x --threads 10 --mem 20000M --outdir fastq --split-files --include-technical && pigz -p 20 fastq/SRR\*.fastq). If an original bam-file from 10X cellranger count pipeline was deposited it needs to be further converted to fastq (example command: ~/src/cellranger-7.1.0/bin/cellranger bamtofastq --nthreads=12 /data/PRJNA\*\*\*\*\*\*/bam/SRR\*\*\*\*\*\*\*.bam /data/PRJNA\*\*\*\*\*\*/fastq/SRR\*\*\*\*\*\*\*/). Information regarding publicly available data deposition included in **ED Table 1**.

## Deriving initial UMI-count matrices

The Cell Ranger pipeline (v7.1.0) (**zhengMassivelyParallelDigital2017?**) was used to perform sample demultiplexing, barcode processing and single-nuclei gene counting. Reads containing sequence information were aligned using the optimised mouse genome reference (vmm10\_optimized\_v.1.0) provided by Pool’s lab based on the default Cell Ranger mm10 genome version 2020-A that was cleared from gene overlaps, poorly annotated exons and 3’-UTRs and intergenic fragments (**poolEnhancedRecoverySinglecell2022?**). PCR duplicates were removed by selecting unique combinations of cell barcodes, unique molecular identifiers (UMIs) and gene ids with the final results being a gene expression matrix that was used for further analysis. We aligned reads using --include-introns quantification mode.

## RNA Velocity analysis enabling count matices

To obtain multiple count matrices with regard of read source (related to exon/introne genes structure) we build piscem index (v) (He, Soneson, and Patro, n.d.) of the refernce genome for spliceu version of simpleaf quantification pipeline (vdocker://etretiakov/usefulaf:0.9.0) based on alevin-fry (v) (He et al. 2022). This approach allows high quality resolution of ambiguoty of the read source which is especially important in case of single nuclei RNA-seq (He, Soneson, and Patro, n.d.; Kuo, Hansen, and Hicks, n.d.; **eldjarnhjorleifssonAccurateQuantificationSinglenucleus2022?**), which is strongly affect results of RNA Velocity estimates (Soneson et al. 2021; Gorin et al. 2022).

# Preparation of individual single-cell RNA sequencing datasets for analysis

## Quality control

### Droplet selection

The droplet selection method of Cell Ranger is based on EmptyDrops method (**lunEmptyDropsDistinguishingCells2019?**) incorporated into cellranger count pipeline. Information regarding number of detected cells/nuclei included in Supplementary Table 1.

### Ambient RNA removal

Using those values as the expected number of cells, we applied a neural network-based approach called CellBender (docker://etretiakov/cellbender:v0.0.1) (**fleming2022?**). To establish additional information regarding degree of regionally determined contamination and quality of particular samples, we set a false positive rate threshold at different levels (0.1, 0.01 and 0.001) and set the neural network to learn over 150 epochs with a total numbers of droplets included based on knee plots (please see Online supplementary Cell Ranger reports). Thus, analysis described below were performed using all three resolutions and not corrected data variants. Information regarding size of set of droplets used in estimation is included in **ED Table 1**.

### Doublets detection

For each sample separately we quantified probability to be a doublet for every cell based on expected doublets-rate reference table provided by 10X Genomics and number of cells/nuclei in samples (Wolock, Lopez, and Klein 2019) ([v0.2.3@pyh5e36f6f](mailto:v0.2.3@pyh5e36f6f)\_1) Information regarding expected doublet rate used in estimation is included in Supplementary Table 1. (also see GetDoubletRate in function.R and scrublet\_cb.py files of code directory).

### Further filtering

Gene annotation information was added using the gprofiler2 package (v0.2.1) (**reimandProfileraWebServer2016?**); thus we filter cells based on high content of mitochondrial, ribosomal or hemoglobin proteins genes, specific thresholds were chosen individually for each dataset (please see Online GitHub Exploratory Analysis reports and params.json files); additionally pseudogenes and poorly annotated genes were also deleted from count matrix. Moreover, cells of low complexity were filtered out as (). Therefore, cells were assigned cell cycle scores using the CellCycleScoring function in the Seurat package (v4.3.0) (**satijaSpatialReconstructionSinglecell2015?**; **stuartIntegrativeSinglecellAnalysis2019?**). Additionally, in case of correlation analysis of arcuate nucleus astrocytes (REF deng) we filtered out cells based on expression of 13 mitochondria-coded genes using more elaborated and strict approach to avoid bias caused by sequencing technical variability when we explored correlation with the mitofusin 2 (Mfn2). So we applied individual filtering models fitted for each sample of the dataset separately (REF packages splines, flexmix, miQC): 536-1\_chow-diet, 536-3\_chow-diet, 537-5\_538-2\_high-fat-diet - mixtureModel with posterior cutoff 0.95; 536-5\_chow-diet - spline mixtureModel with posterior cutoff 0.7, 537-1\_537-3\_high-fat-diet - spline mixtureModel with posterior cutoff 0.95, 536-2\_537-4\_high-fat-diet - spline mixtureModel with posterior cutoff 0.999999.

## Astrocytes selection

Analysis described in next three sections were done twice were the first iteration explore all cell types and second iteration focuses on particular glial cell types: Astrocytes, Tanycytes, Ependymal cells and NG2-glia, oligodendrocytes precursor cells.

### Gene selection

We used the selection method in the Seurat package (v4.3.0) (**satijaSpatialReconstructionSinglecell2015?**; **stuartIntegrativeSinglecellAnalysis2019?**), which uses a modern variance stabilising transformation statistical technic that utilises scaling to person residuals (**hafemeister2019?**). That way, we selected 3000 highly variable genes per dataset and regressed out complexity and cell-cycle variability prior to the final scaling of filtered matrixes.

### Graph-based and multi-level reconcile tree clustering

We performed Leiden algorithm graph-based clustering. PCA was performed using the selected genes and the jackknife tested (**chungStatisticalSignificanceVariables2015?**) principal components (we tested the significance of feature for randomly picked 100 samples by 2% of data each over 1000 iterations; see PCScore function in functions.R script of code directory) were used to construct a shared nearest neighbour graph using the overlap between the 15 nearest neighbours of each cell. Leiden modularity optimisation (**traag2019?**) was used to partition this graph with an array of resolution parameters where 30 modularity events were sampled between 0.2 and 2.5. Clustering tree visualisations (**zappiaClusteringTreesVisualization2018?**) were produced using the clustree package (v0.5.0) showing the resolution of previously identified clusters. By inspecting these resolutions reconcile tree produced by mrtree package (v0.0.0.9000) (**pengCellTypeHierarchy2021?**) and calculating adjusted multi-resolution Rand index chosen as maximum value if there is no higher modularity within 0.05 AMRI difference (see SelectResolution in function.R file of code directory).

### Marker genes

Marker genes for each cluster were identified using logreg test (**ntranos2019?**) implemented in Seurat framework (v) (**stuartIntegrativeSinglecellAnalysis2019?**). Genes were considered significant markers for a cluster if they had an FDR less than 0.001. Identities were assigned to each cluster by comparing the detected genes to previously published markers and our own validation experiments; we assigned the astrocyte lablel to a cell only if it contained at least seven marker genes. The list of 22 marker genes that we used as inclusion critereas is available as **ED Table 2** (also see class\_cello.py files of code repository).

### Classification of cell types based on gene-scoring with publicly available database information

Additionally, we further explore gene signatures using enrichment with Over Representation Analysis (Badia-i-Mompel et al. 2022). For that we used canonical markers from PanglaoDB with particular focus on Astrocytes, Ependymocytes, Tanycytes and OPCs markers (also see class\_cello.py file of code repository).

### Filtering criterea for astrocytes

As described above, we filtered out astrocytes with cell quality markers enlisted in the code repository as individual params.json files for each dataset. Regardless of the dataset and method of astrocytes seclection additionally, we filtered out cells using manually curated list of 59 marker genes for each major subtype of cells (glutamatergic, GABAergic and peptidergic neurons, oligodendrocytes on different developmental stages from oligodendrocyte precursor cells (OPCs) to mature myelinesing oligodendrocytes, pericytes, vascular muscle cells, macrophage and microglia immune cells, and most importantly more close on developmental trajectory – tanycytes and ependymal cells). Therefore, any astrocyte to pass our filter might contain maximum two genes from the whole exclusion critera list as strict trade-off taking into account the effect of ambient RNA described above. Information regarding number of astrocytes included in analysis is included in **ED Table 1**. The list of Marker genes that we used as exclusion critereas is available as **ED Table 2** (also see class\_cello.py file of code repository).

## Prepare train and test data split to enable hyperparameters optimisation of supervised machine learning models by classification performance metrics

As we had varible number of input cells in subregional datasets, we defined the probability of a cell to be in the training set as 90% or if it was smaller - the ratio of 1000 cells to the total number of cells in the dataset. We used this probability to randomly assign cells to the training set. We used this process to obtain more balanced training sets regardless of initial number of astrocytes. We used the same training sets for all models to ensure that the results are comparable. We used the remaining cells as the test set to enable evaluation of the model performance on unseen data. We used the same test sets for all models to ensure that the results are comparable. The code for this procedure is available in class\_cello.py file of code repository.

As result of previous steps we derived subsets of astrocytes from each sample of 12 individual subregional dataset splitted into train/test sets and four whole hypothalamus dataset (**ED Figure 4A**: Step 0).

# Identification of astrocytic signatures and diversity description across subregions

## RNA velocity analysis of individual datasets

## Full conventional integration

First, to compare astrocytes from different subregions of hypothalamus we performed integration using well-established anchor-based integration using Seurat package (v4.3.0) (**satijaSpatialReconstructionSinglecell2015?**; **stuartIntegrativeSinglecellAnalysis2019?**). We used both SCTransform-based and “normal” GLM-scaled versions of the default pipeline trying to explore heterogeneity of astrocytes across the hypothalamus with vary number of highly variable genes selected (from 2500 to 7500 initial hvg param). Particular problem was caused by small fraction of variability shared across all analysed subregions (~ maximum 426 genes) which represented mostly commonly known general markers of astrocytes across CNS.

## Iterative Paired integrations with shared astrocytes signature masking

To answer the question if there is any level of subregional heterogeneity behind identified astrocyte types transitions between functional states, we decided to change the strategy and elaborated the approach where we diminished the shared signatures of astrocytes. To avoid observed obstackles with rapidly reduced number of anchor genes that allow integration of datasets, we performed pairwised integrations of each subregional dataset with each whole hypothalamus dataset using python implementation of Harmony algorithm (REF). Next to highlight subregional differences intead of functional states idenified above, we extracted anchor genes (subset of original highly variable features corrected for concatanated datasets to maximise explained variance as evaluated by optimised variance stabilising transform procedure (REF Kobak)) of earlier performed pairwise integrations resulting with 48 sets of 2000 genes in total or 12 sets (REF Arc: deng, lutomska; LHA: mickelson2019; MnPO: poolEnhancedRecoverySinglecell2022; POA: moffit; PVN: lopez; SCN: wen, morrison; VMH: liu; VPH: mickelson2020; MBH: rupp) per whole hypothalamus reference (REF zeisel, romanov, kim, hajdorovic; also see get\_full\_pair\_mtx.py file of the code repository and **ED Figure 3,75** to see example of PaCMAP embeddings of integrated datasets before ambient RNA removal; **ED Figure 4A**: Step 1).

Thus, as we had genes enlisted in the order of descending variance on the same plane of the whole hypothalamus reference, we adapted the robust rank aggregation algorithm (REF) to identify broadly expressed genes of functional states in two steps: first, we derived scores per whole hypothalamus reference taking into account complete number of resulted genes with used version of the mouse genome (45163 genes in initial matrixes; also see get\_shared\_signature.R file of the code repository for the first step; **ED Figure 4A**: Step 2), and second, we estimated robust rank aggregation score of four derived lists (see all lists in **ED Table 3**; also see get\_aggregated\_shared\_signature.R file of the code repository for the second step; ; **ED Figure 4A**: Step 3).

Consequently, we derived integrated manifolds between hajdorovic and 12 subregional datasets again after removal of shared signature using 100, 250 and 500 genes thresholds (we also removed redundunt correlated genes with absolute Pearson’s correlation > 0.5 corresponding to each threshold (REF nancorrmp; **ED Figure 4A**: Steps 4-5)) in order to understand impact of functional genes (see get\_substr\_pair\_mtx.py file and particularly get\_top\_abs\_correlations() and remove\_irrelevant\_top\_abs\_correlations() functions in the code repository; **ED Figure 4A**: Steps 6-7 and **ED Figure 9,75** to see PaCMAP embeddings of integrated datasets at 0,001, 0,01 FDR levels and without (*nc*) ambient RNA removal). These steps were used to unmask potentially hidden spatial marks in astrocytes from different hypothalamic nuclei cleaning genes-axis of the expression matrices and their respective manifolds dimensionally reduced by the integration.

Downstream analysis of astrocytes subregional heterogeneity was reported here using datasets after removal of ambient RNA at 0.001 FDR levels and removal of shared signature at 100 genes threshold with corresponding correlated genes as described above and using k=10 for nearest neighbours (NN) graphs constraciton (and ks=k+10=20 for for expanded mutual kNN-graph with path-connectivity model) as our limited testing (FDR=[0.001, 0.01, "nc"]; signature\_substruction=[100, 250, 500, "full"], neighbors\_k=[5, 10, 25, 50], connectivity\_mst=["full\_tree", "min\_tree"]) demonstrated best performance with such parameters and we had to reduce complexity of the analysis (see Snakefile in the root of the main GitHub repository). To explore those parameters combinations, we used simplified approach: 1) we performed overclustering of train part of each integrated pair using clustering from leidenalg ([v0.9.1@pypi](mailto:v0.9.1@pypi)\_0; REF leiden and original paper), 2) we subset each of 12 matrixes to the same reference dataset (REF hajdarovic), and 3) derived fuzzy simplical set using UMAP-based path-connectivity model (REF UMAP original and connectivity) then 4) we aggregated results (microclusters and graphs) using partitioning based graph abstruction (PAGA REF) 5) that PAGA graphs were utilised as layers of multiplexed graph to optimised multiplex partitions (REF leiden, multiplex science) and 6) we used those partitions to train support vector machine (SVM) classifier (SVC) and to project those partitions on test set astrocytes, 7) finally, we embedded full reference dataset using derived labels for supervised densMAP (see microclustering\_limited\_test.ipynb file of the analysis repository; see example of unfiltered (FDR=“nc”, signature\_substruction=“full”, k=10, ks=20, connectivity\_mst=“min\_tree”) data embedding **Figure 1 K, M**).

# Classification of astrocytes subregion with *Astro Trap* approach

To support performance of classifier model we used two steps of feature selection:

## Filter genes that are allow to discriminate between subregions/nuclei

We prefiltered features using the ensemble of methods to capture genes of known subregional origin (**ED Figure 4A**: Step 11). For that we used a union of gene sets selected by chi2 statistics, ANOVA F-value and mutual information score. We defined sets of features by those three different feature selection methods as it improves the accuracy and stability of the selected features, leading to better performance in machine learning models. We used GridSearchCV to optimise parameters of pipelines that consisted of three steps: scaling - MinMaxScaler for consistency (as negative values are not allowed for all statistics we calculated), reduce dimentionallity - SelectKBest taking input one of chi2, f\_classif or mutual\_info\_classif, and classify - SVC estimator with scaled gamma and maximum number of iteration fixed to 1000. The parameter grid included SVM fit strenght (C\_OPTIONS = [1, 10, 100, 1000]) and number of features K to select with (2000, 1000, 500, 250) genes thresholds. We used RepeatedStratifiedKFold(n\_splits=10, n\_repeats=10) cross validation to select classifiers with the highest weighted F-measure scores. Finally, we extracted features from each best\_estimator and merged them into one set of genes (results are available as **ED Table 4** ; **ED Figure 4A**: Step 11). The union of features sets was used to filter input to the downstream pipeline.

## Find the best support vector machine classifier model with *ad hoc* feature selection for the region-label assignment using Grid-Search groupped k-fold cross validation

To build an effective classifier model that reconcile features of subregional origin regardless of functional states (that we captured by the shared signature) we used initially selected astrocytes from subregional datasets (**ED Figure 4A**: Step 0) and prefiltered genes set (**ED Figure 4A**: Step 11). However, to learn robust features that consistently represent this factor of regional heterogeneity we must recognise standalone functional roles of nuclei/areas of the hypothalamus. The astrocytic functional states occur across the whole hypothalamus and are promped by various environmental factors; thus we expect that they are not specific to any subregion in longterm and reflect continuous complex trajectory transitions. To resolve this issue we clustered this shared structure into tiny groups of cells that being examined separately; hence we suppose diminished intraregional roles versus allowed exploration of interregional importances of genes. After having reduced number of genes and meaningful groups for stratification we could apply more complex model for feature selection. Pipelines consisted from scaling - RobustScalerthat smooth outliers, feature selection and classification by SVC. We implemented two types of feature selection: first is simplier interpretable LogisticRegression that utilised L1-regularisation to fit coefficients for multiclass problem as one versus rest (for the sake of computational performance in contrast to multinomial estimation); second is the more complex XGBoost classifier model. To fit these models we encoded region classes as sorted proportions of each class. To select best model we performed GridSearchCV with StratifiedGroupKFold(n\_splits=10) cross validation on the training set scored by Matthews correlation coefficient (MCC) as subregion classes proportions were unbiased.

### Deriving the microclusters of astrocytic subfunctional states

Therefore, in order to further consolidate the signal we learned graph representations using Minimal Spanning Tree with path-connectivity model implementation of the UMAP algorithm (REF UMAP original and connectivity papers). First, we reduced 12 integration manifolds of training data (REF hajdorovic) using the UMAP algorithm (REF UMAP original) with following parameters: n\_components=6, n\_neighbors=20, n\_epochs=1000, metric="cosine", init="spectral", learning\_rate=0.1, min\_dist=1, spread=2, repulsion\_strength=2.0, negative\_sample\_rate=10, angular\_rp\_forest=True, densmap=False. Second, we used UMAP algorithm to derive the 10-NN graph from this matrix in euclidean space with pynndescent optimisation. Third, we used 20-NN graph and distances matrix to bulid 20-mutual-NN graph using minimal tree path-connectivity model (see functions min\_spanning\_tree(), create\_connected\_graph(), find\_new\_nn(), mutual\_nn\_nearest() in microclusters\_to\_groups\_cv\_feature\_selection\_svc.ipynb; REF UMAP connectivity). Fourth, we derived fuzzy simplical set for the new graph using the UMAP algorithm (REF UMAP). Those graphs information were saved in anndata container being transformed from\_scipy\_sparse\_array to iGraph, and through igraph.Graph.get\_adjacency\_sparse() to adata.obsp["connectivities"] object, also we derived required adata.obsp["distances"] using get\_sparse\_matrix\_from\_indices\_distances\_umap() function (see microclusters\_to\_groups\_cv\_feature\_selection\_svc.ipynb file in the main GitHub analysis repository; REF anndata, igraph, networkx; **ED Figure 4A**: Step 8). Next, we used Optimiser class from leidenalg package to explore the resolution profile of the Constant Pots Model clustering on the weighted graph at range from 0 to 2 till full convergence (REF leiden, optimiser and original paper). Thus we partitioned the weighetd graph using CPMVertexPartition function with resolution\_parameter=0.0112 and applied optimise\_partition function of optimiser on top till convergence (REF leiden, optimiser and original paper). The partitioning membership was saved as categorical vector to the PRJNA779749\_init.obs["subfunct\_groups"] (see **ED Figure 4A**: Step 9). Therefore, we trained 12 support vector machine classifiers (linear kernel, svm fit strenght - C=100) from sklearn package to fit those labels (PRJNA779749\_init.obs["subfunct\_groups"]) to the subsets of 12 integrated matrices; hence we projected learned groups labels to the corresponding subregional parts of training sets (**ED Figure 4A**: Step 10).

### Optimisation of hyperparameters of a simple logit-SVM classifier model

We trained Pipeline with SelectFromModel(LogisticRegression(solver="saga", multi\_class="ovr", penalty="l1", max\_iter=10000, n\_jobs=-1,), threshold=-np.inf) feature selection and SVC estimator with scaled gamma and maximum number of iteration fixed to 10000. The parameter grid included SVM fit strenght (C\_OPTIONS = [1, 10, 100, 200]) and maximal number of features to select with N\_FEATURES\_OPTIONS = [round(n\_features / 3), round(n\_features / 4), round(n\_features / 8), round(n\_features / 10), round(n\_features / 20)] (where n\_features equals to number of genes in the union of prefiltering sets). Also, we extracted features from best\_estimator (results are available as **ED Table 4**). Finally, we trained the best model on the whole training set and dumped it for further evaluation (**ED Figure 4A**: Step 12).

### Optimisation of hyperparameters of the more complex XGBoost-SVM classifier model

We trained Pipeline with SelectFromModel(XGBClassifier(tree\_method="gpu\_hist"), threshold=-np.inf) feature selection and SVC estimator with scaled gamma and maximum number of iteration fixed to 10000. The parameter grid included SVM fit strenght (C\_OPTIONS = [1, 10, 100, 200]); at the feature selection step we optimised such parameters of the estimator as learning\_rate": [0.1, 0.3, 0.5], boosting trees depth and number of rounds max\_depth: [3, 5, 7], n\_estimators": [100, 250, 500], as well as the maximal number of features to select with N\_FEATURES\_OPTIONS (see logit-SVC section). Also, we extracted features from best\_estimator (results are available as **ED Table 4**). Finally, we trained the best model on the whole training set and dumped it for further evaluation (**ED Figure 4A**: Step 12).

## Evaluation of the best classifier model on the test set and additional validations of whole hypothalamus predictions

As the result of such optimisation we built two SVM-based classificaton pipelines:

1. LogisticRegression (logit)-based features selection with L1-regularisation + SVC
   1. pick 1406 genes during feature selection stage (**ED Table 4**)
   2. performed with 0.4408352445523123 MCC and 0.6273951088822771 weighted F-measure on the evaluation set (during cross-validataion mean test MCC score was 0.619153; see all results **ED Table 5**)
2. XGBoost-based feature selection + SVC
   1. suggested 211 genes, learning\_rate=0.3, boosting trees depth max\_depth=3 and number of rounds n\_estimators=100 for the feature selection and strong (C=200) SVM classifier (**ED Table 4**)
   2. outperformed simplier logit-SVC model with 0.5776601427855099 MCC and 0.723413495875078 weighted F-measure on the evaluation set (**Figure 2A; ED Figure 4A**; during cross-validataion mean test MCC score was 0.870621; see all results **ED Table 5**)

Next, these models were used to project learned subregional classes onto whole hypothalamus datasets (REF zeisel, romanov, kim, hajdarovic). For that, we used the original model selected directly from the whole set of all available genes from the full subregional datasets matrixes. Thus, the model should be refitted to the subset of genes available for prediction in a particular dataset resulting with four different models per pipeline.

Therefore, we also come up with *ad hoc* approach that can safe time and interpretation resulting with a shared model across each predicted datasets. For that before applying AstroTrap approach, we selected subset genes that: 1) were expressed in more than 10 cells in each whole hypothalamus dataset and than 2) took intersection of these genes across all four datasets. This resulted with two best-models selected by cross-validation: 1) fit strenght C=10 and 1378 genes (**ED Table 4**) that were used for the logit-SVC model, which performed with 0.714431 mean test MCC score, and 2) fit strenght C=200, learning\_rate=0.5, boosting trees depth max\_depth=3 and number of rounds n\_estimators=100, 207 genes (**ED Table 4**) for the XGBoost-SVC model, which performed with 0.769805 mean test MCC score. The results are available in **ED Table 5**. We used these models to predict subregional classes for evaluation set and in each whole hypothalamus dataset. On the evaluation set results of the *ad hoc* version models were 1) 0.5391718659990032 MCC and 0.70261086217399 weighted F-measure for logit-SVC model, and 2) 0.5183861279482299 MCC and 0.6863289741984487 weighted F-measure for XGBoost-SVC model.

Subregional classes predicted by direct version of *Astro Trap* (**ED Figure 4A**: Step 13) were used for further functional biological analysis.

# Functional annotation and validation of astrocytic subregional heterogeneity

## Gene-regulatory networks (regulons) analysis with SCENIC pipeline

Like in our previous analysis using SCENIC, here we could confirm pilot observation of regulons-driven heterogeneity in astrocytes and magnified the analysis of the same dataset to astroglia assigned to different hypothalamic subregions (see **Figure 2 B, C**; also **ED Table 6**). For that we used normalised expression matrixes of astrocytic subsets from the whole hypothalamus reference datasets and applied SCENIC pipeline (REF SCENIC; docker://aertslab/pyscenic:0.12.1) to infer regulons and their activity scores. For aucell procedure we built regulons only using enriched motifs with a NES of 3.0 or higher; we took only directly annotated TFs or TF annotated for an orthologous gene into account; and we only kept regulons with at least 3 genes. We used aucell matrix to visualise dotplot of regulons activity and to build dendrogramm (using Wald clustering of regulons (REF)). We found that regulons are highly specific to the subregional classes of astrocytes (**Figure 2B; ED Figure 5**).

## Ligand-receptor expression pairs between Latexin positive vs. negative astroglia with Arcuate nuclei neurons governing feeding behavior (WIP SECTION)

To validate the assumption of the subregional specificity of Lxn+ astocytes, we analysed selected astrocytes signalling data estimated using Liana-py and Rank Aggregation method (including CellPhoneDB, CellChat, ICELLNET, connectomeDB2020, and CellTalkDB). (Fig. 4E,F; ED Fig. 7A-E)

# Sample sizes, statistics and reproducibility (WIP SECTION)

1. Differential gene expression analysis of full astrocytic integration cluster analysis SCTransform corrected UMI-count matrices were statistically tested to obtain DEGs using log-normalized values with pseudocount = 1 for 12 identified clusters as previously described (REF Mayer, REF Hafemeister) using MAST test (REF Finak). Results of the DGE tests are specified at **ED Table 3**.
2. Comparisons of models performance for hyperparameters optimisation during groupped 10-fold cross-validation trained and evaluated using GridSearchCV. Perfomance of models was evaluated using matthews\_corrcoef (MCC). The output of GridSearchCV does not provide information on the certainty of the differences between the models. To evaluate this, we need to conduct a statistical test on MCC sesults. Several variance-corrected statistical tests have been developed for these cases. We used the one proven to obtain the highest replicability scores (which rate how similar the performance of a model is when evaluating it on different random partitions of the same dataset) while maintaining a low rate of false positives and false negatives: the Nadeau and Bengio’s corrected t-test ((Nadeau and Bengio 1999)). We used implementations of Nadeau and Bengio’s corrected t-test under two different statistical frameworks: frequentist and Bayesian. Therefore, approach assumed a corrected right-tailed paired t-test to evaluate if the performance of the first model is significantly better than that of the second model; where, our null hypothesis was that the second model performs at least as good as the first model. Moreover, we used Bayesian estimation to calculate the probability that the first model is better than the second. Bayesian estimation output a distribution followed by the mean μ of the differences in the performance of two models. To obtain the posterior distribution we need to define a prior that models our beliefs of how the mean is distributed before looking at the data, and multiply it by a likelihood function that computes how likely our observed differences are, given the values that the mean of differences could take. One way of defining our posterior using a closed-form expression is to select a prior conjugate to the likelihood function. Benavoli and colleagues ((Benavoli et al. 2017)) show that when comparing the performance of two classifiers we can model the prior as a Normal-Gamma distribution (with both mean and variance unknown) conjugate to a normal likelihood, to thus express the posterior as a normal distribution. Marginalizing out the variance from this normal posterior, we can define the posterior of the mean parameter as a Student’s t-distribution. Overall, using the Bayesian approach we computed the probability that a model performs better, worse or practically equivalent to another.
3. Differential gene expression analysis of projected subregional classes of astroglia onto whole hypothalmus datasets and UMIs counts matrices were statistically tested to obtain DEGs for eight available subregional classes using MAST test (REF Finak). We used the Logreg test (REF Ntranos) to define differential regulons expression across cells labeled by projected subregional classes of astroglia onto whole hypothalmus datasets. Finally, we applied robust rank aggregation of differentially expressed genes across different datasets to derive individual lists of markers for each of nuclei (REF RRA); we targeted RRA on the p-value to determine ***Specificity Score*** **and on the** log2 Fold change to determine ***Magnitude Score*** (**ED Figure 4A**: Step 14; **Figure 2A-F**; **ED Table 6**).
4. Data analysis with Bootstrap-coupled Estimation (DABEST) of cell-to-cell interactions between vs.  astrocytes and neurons (multiple groups)
5. Circadian activity analysis estimated using Fos expression in the SCN nucleus with TRAP2 model was analysed across neurons and astrocytes (multiple groups) between two conditions (CT: ); sample size was XXX animals per condition; two-tailed t-test with Welch’s correction was used for statistical analysis; bar plots whiskers indicate standard error.
6. Acute stress condition analysis estimated using Fos expression in the PVN nucleus with TRAP model was analysed in astrocytes (multiple groups) compared to control; sample size was 7 animals per condition; two-tailed t-test with Welch’s correction was used for statistical analysis; bar plots whiskers indicate standard error.
7. HFD condition analysed using electron microscopy measured morphological parameters of astrocytes and immuno-histochemistry imaging within the Arc nucleus and phenotypical parameters monitoring with Mfn2-KO model compared to control; sample size was XXX animals per condition; two-tailed t-test with Welch’s correction was used for statistical analysis; bar plots whiskers indicate standard error.

# Other packages

Visualisations and figures were primarily created using the ggplot2 (v3.4.2), cowplot (v1.1.1) (Wilke 2020), patchwork (v1.1.2.9000) and scCustomize (v1.1.1) packages using the viridis colour palettes (v0.6.2) for continuous data. UpSet plots (**conwayUpSetRPackageVisualization2017?**) were produced using the UpSetR package (v1.4.0) (Gehlenborg 2019) with help from the gridExtra package (v2.3) (Auguie 2017). Data manipulation was performed using other packages in the tidyverse (v2.0.0.9000) (Wickham 2023) particularly dplyr (v1.1.2) (Wickham et al. 2023), tidyr (v1.1.2) (Wickham, Vaughan, and Girlich 2023) and purrr (v1.0.1) (Wickham and Henry 2023). The analysis project was managed using the Snakemake system (v ) (**molderSustainableDataAnalysis2021?**) and the workflowr (v1.7.0) (Blischak, Carbonetto, and Stephens 2021) package which was also used to produce the publicly available website displaying the analysis code, results and output. Reproducible reports were produced using Quarto (v1.2), knitr (v1.42) (Xie 2023) and R Markdown (v2.21) (Allaire et al. 2023) and converted to HTML using Pandoc (v2.19.2).

# Data availability

All scRNA-seq datasets had been deposited in GEO previously and reanalysed here (accession numbers and metadata are available in **ED Table 1**). All data presented (for example, imaging) will be made available by T. Harkany ([tibor.harkany@ki.se](mailto:tibor.harkany@ki.se) or [tibor.harkany@meduniwien.ac.at](mailto:tibor.harkany@meduniwien.ac.at)) upon reasonable request.

# Code availability

The code used is available at <https://doi.org/10.6084/m9.figshare.11867889>.

# Summary

## Output files

versions <- purrr::map(versions, as.character)  
versions <- jsonlite::toJSON(versions, pretty = TRUE)  
readr::write\_lines(  
 versions,  
 here::here("output", DOCNAME, "package-versions.json")  
)

# References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2023. *Rmarkdown: Dynamic Documents for r*.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*.

Badia-i-Mompel, Pau, Jesús Vélez Santiago, Jana Braunger, Celina Geiss, Daniel Dimitrov, Sophia Müller-Dott, Petr Taus, et al. 2022. “decoupleR: Ensemble of Computational Methods to Infer Biological Activities from Omics Data.” Edited by Marieke Lydia Kuijjer. *Bioinformatics Advances* 2 (1): vbac016. <https://doi.org/10.1093/bioadv/vbac016>.

Benavoli, Alessio, Giorgio Corani, Janez Demšar, and Marco Zaffalon. 2017. “Time for a Change: A Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis.” *The Journal of Machine Learning Research* 18 (1): 2653–88.

Blischak, John, Peter Carbonetto, and Matthew Stephens. 2021. *Workflowr: A Framework for Reproducible and Collaborative Data Science*. <https://github.com/workflowr/workflowr>.

Gehlenborg, Nils. 2019. *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. <http://github.com/hms-dbmi/UpSetR>.

Gorin, Gennady, Meichen Fang, Tara Chari, and Lior Pachter. 2022. “RNA velocity unraveled.” *PLoS computational biology* 18 (9): e1010492. <https://doi.org/10.1371/journal.pcbi.1010492>.

He, Dongze, Charlotte Soneson, and Rob Patro. n.d. “Understanding and Evaluating Ambiguity in Single-Cell and Single-Nucleus RNA-Sequencing.” <https://doi.org/10.1101/2023.01.04.522742>.

He, Dongze, Mohsen Zakeri, Hirak Sarkar, Charlotte Soneson, Avi Srivastava, and Rob Patro. 2022. “Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data.” *Nature Methods* 19 (3): 316–22. <https://doi.org/10.1038/s41592-022-01408-3>.

Kuo, Albert, Kasper D. Hansen, and Stephanie C. Hicks. n.d. “Quantification and Statistical Modeling of Chromium-Based Single-Nucleus RNA-Sequencing Data.” <https://doi.org/10.1101/2022.05.20.492835>.

Nadeau, Claude, and Yoshua Bengio. 1999. “Inference for the Generalization Error.” In *Advances in Neural Information Processing Systems*. Vol. 12. MIT Press. <https://proceedings.neurips.cc/paper/1999/hash/7d12b66d3df6af8d429c1a357d8b9e1a-Abstract.html>.

Soneson, Charlotte, Avi Srivastava, Rob Patro, and Michael B. Stadler. 2021. “Preprocessing choices affect RNA velocity results for droplet scRNA-seq data.” *PLoS computational biology* 17 (1): e1008585. <https://doi.org/10.1371/journal.pcbi.1008585>.

Varela, Luis, Jae Geun Kim, Pablo Fernández-Tussy, Binod Aryal, Zhong Wu Liu, Carlos Fernández-Hernando, and Tamas L. Horvath. 2021. “Astrocytic Lipid Metabolism Determines Susceptibility to Diet-Induced Obesity.” *Science Advances* 7 (50): eabj2814. <https://doi.org/10.1126/sciadv.abj2814>.

Varela, Luis, Bernardo Stutz, Jae Eun Song, Jae Geun Kim, Zhong-Wu Liu, Xiao-Bing Gao, and Tamas L. Horvath. 2021. “Hunger-Promoting AgRP Neurons Trigger an Astrocyte-Mediated Feed-Forward Autoactivation Loop in Mice.” *The Journal of Clinical Investigation* 131 (10). <https://doi.org/10.1172/jci144239>.

Varela, Luis, Shigetomo Suyama, Yan Huang, Marya Shanabrough, Matthias H. Tschöp, Xiao-Bing Gao, Frank J. Giordano, and Tamas L. Horvath. 2017. “Endothelial HIF-1α Enables Hypothalamic Glucose Uptake to Drive POMC Neurons.” *Diabetes* 66 (6): 1511–20. <https://doi.org/10.2337/db16-1106>.

Wickham, Hadley. 2023. *Tidyverse: Easily Install and Load the Tidyverse*.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*.

Wickham, Hadley, and Lionel Henry. 2023. *Purrr: Functional Programming Tools*.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2023. *Tidyr: Tidy Messy Data*.

Wilke, Claus O. 2020. *Cowplot: Streamlined Plot Theme and Plot Annotations for Ggplot2*. <https://wilkelab.org/cowplot/>.

Wolock, Samuel L., Romain Lopez, and Allon M. Klein. 2019. “Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data.” *Cell systems* 8 (4): 281–291.e9. <https://doi.org/10.1016/j.cels.2018.11.005>.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. <https://yihui.org/knitr/>.