Neuron types in the developing mouse CNS can be divided into several epigenomic and transcriptomic classes

Sami Kilpinen[1], Heidi Heliölä, Kaia Achim[1]

Molecular and Integrative Biosciences Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Finland

[1]Corresponding author

Summary

We hypothesize that the differential activity of transcription factors (TFs) regulating patterning and cell fate determination manifests in a differential genome activation in different cell lineages and cell types. Here, we have mapped mRNA expression and accessible chromatin regions in single cells derived from the developing mouse CNS. We first defined a reference set of features for scATAC-seq read quantitation across samples, allowing comparison of chromatin accessibility between brain regions and cell types directly. Second, we integrated the scRNA and scATAC data to form a unified resource of transcriptome and chromatin accessibility landscape for the cell types in diencephalon to anterior hindbrain in E14.5 mouse embryo. Importantly, we implemented methods to achieve high level of statistically supported resolution of clustering for single-cell -omics. We show high level of concordance between the cell clustering based on the chromatin accessibility and transcriptome in analyzed neuronal lineages, indicating that both genome and transcriptome can be used for cell type definition. Hierarchical clustering by the similarity in genome activation reveals complex interrelationships between the cell types derived from different anteroposterior domains of the E14.5 CNS. Further work is required to describe the fraction of genome activation resulting from the activity of patterning genes and the fraction activated by the cell fate selection regulators.

Introduction

Clustering the single cells by their chromatin accessibility allows to identify cell types and measure molecular distance between the cell types [1,26].The topology of the cell type tree can reveal developmental and evolutionary relatedness or depict developmental lineages in the sample. Here, we use an approach of single-cell data integration [2], where we integrate the open chromatin profile (scATAC-seq reads, later scATAC) and mRNA expression profile (scRNA-seq reads, later scRNA) of cells derived from the diencephalon, midbrain and rhombomere 1 regions of E14.5 mouse brain. In our data analysis, we use novel joined feature space created here by adapting method first described by Cusanovich *et al.*[3] where we optimize feature detection by performing cellular clade level peak calling across different brain regions. Cell type detection by clustering is prone to bias due to overclustering [6]. We mitigated this risk in our chromatin accessibility-based cell cluster detection with iterative method to find optimal resolution parameter, maximizing the statistical confidence of clustering. Finally, we integrated scATAC and scRNA modalities together with canonical correlation analysis (CCA) [4]. This allows dataset to be used to study transcriptional and epigenetic variation among cell types.

In the integrated dataset created here, we studied the distribution and relatedness of cells in correlation with clustering, brain region and selector gene expression. The expression of previously known regional markers as well as cell fate specific selector transcription factors aligned with the scATAC-based clustering of single cells and could be correlated with novel regulatory DNA elements and RNA expression. This creates a necessary starting point for studying the genetic regulation of GABAergic neuron development, considering the cell type specific open chromatin features, the expression and DNA-binding of regulatory transcription factors, as well as the target gene expression levels simultaneously.

In prospect, our multimodal single-cell data from developing brain can be used to better understand the logic of the transcriptional regulation of cell type-specific gene expression during neurogenesis and neuron differentiation.

Materials and methods

Data and code availability

The E12.5 mouse rhombomere 1 scRNA-seq and scATAC-seq and E14.5 rhombomere 1, midbrain and diencephalon scRNA-seq and scATAC-seq reads generated in our study will be deposited in SRA upon publication. All original code will be deposited in GitHub at https://github.com/ComputationalNeurogenetics/NeuronalFeatureSpace
Any additional information required to reanalyze the data reported in this paper is available upon request.

Single-cell RNA-sequencing

Single cells were isolated from *Nkx2-2*$^{Cre/+}$; *R26R*$^{TdTomato/+}$ mouse embryos [5]. *R26R*^{TdTomato} allele was obtained from Jackson laboratories, catalog number 007909). Embryonic day 0.5 (E0.5) was defined as noon of the day of the vaginal plug. All animal experiments were approved by the Laboratory Animal Center, University of Helsinki, and the National Animal Experiment Board in Finland.
For the scRNA, brains were isolated from E14.5 mouse embryos, and the ventral forebrain, ventral midbrain and ventral R1 pieces were separated. Cells were dissociated using the Papain cell dissociation kit (Worthington, LK003150). Chromium 3' single-cell RNA-seq kit (10xGenomics, 1000146) was used for cell capture and cDNA synthesis, with the cell yield set at 5000 cells per sample. Per each brain region, single cell collection was performed from two individual embryos (see table at Figure 1B for the number of biological replicates). The samples' cDNA libraries were sequenced on Illumina NovaSeq 6000 system using read lengths: 28bp (Read 1), 8bp (i7 Index), 0 bp (i5 Index) and 89bp (Read 2). The reads were quality filtered and mapped against the Mus musculus genome *mm10*, using CellRanger v3.0.1 pipeline for scRNA.
The exact cell number, read yields and read quality for each sample are found in (https://github.com/ComputationalNeurogenetics/NeuronalFeatureSpace/seq_quality_reports).

Processing of scRNA data

scRNA data processing per brain region is described in detail in R code in (https://github.com/ComputationalNeurogenetics/NeuronalFeatureSpace) with main steps outlined below.

scRNA data was read into *Seurat* objects from 10x *filtered_feature_bc* matrix by using Ensembl gene IDs as features. Replicates per each brain region, or all replicates in the later analysis steps were merged as *Seurat* objects. After merging, quality control (QC) was performed by calculating *mt-*

percentage. Cells were filtered requiring *nFeature_RNA* > 2800 and *nFeature_RNA* < 5800 and *percent.mt* < 15. Data was then log-normalized, top 3000 most variable features were detected, data was scaled and *percent.mt*, *nFeature_RNA*, *nCount_RNA* were set to be regressed out.

Data was then run through PCA using variable features, and *Seurat RunUMAP* with PCA dimensions 1-50 was used to generate Uniform Manifold Approximation and Projection (UMAP) for further analysis. *Seurat* functions of *FindNeighbors*(dims=1:50) and *FindClusters*(resolution=3) were applied for neighborhood and community detection-based clustering. Clustering resolution was optimized with *chooseR* [6] as described below. Neurotransmitter type was estimated by using *ScType* [7] as described below.

Single-cell ATAC-seq

Embryonic brains were isolated from E12.5 and E14.5 wildtype mouse embryos, and the ventral forebrain, ventral midbrain (E14.5) and ventral R1 pieces (E14.5, E12.5) were separated. Nuclei were isolated using the recommended protocol for scATAC-seq (Assay for Transposase Accessible Chromatin, 10xGenomics protocol CG000212). At least 2 replicates were collected per each brain region and stage. Chromium Single Cell Next GEM ATAC Solution (10X Genomics, 1000162) was used for the capture of nuclei and the accessible chromatin capture. In chip loading, the nuclei yield was set at 5000 nuclei per sample. Chromium Single Cell Next GEM ATAC Reagent chemistry v1.1 was used for the DNA library preparation. Sample libraries were sequenced on Illumina NovaSeq 6000 system using read lengths: 50bp (Read 1N), 8bp (i7 Index), 16bp (i5 Index) and 50bp (Read 2N). The reads were quality filtered and mapped against the Mus musculus genome *mm10* using Cell Ranger ATAC v1.2.0 pipeline.

Definition of feature space of scATAC-seq data

In order to define joined feature space that could be used in the processing of scATAC samples from different brain regions, we adapted the approach first described by Cusanovich *et al.* [3].

As a first phase, scATAC data from the replicates of the same brain region were processed through pipeline described in (https://github.com/ComputationalNeurogenetics/NeuronalFeatureSpace) by using *Seurat* and *Signac* packages in R [2,8], obtaining read count tables per cell per 5kbps bin over the entire mm10 genome (excluding X- and Y-chromosomes).

In second phase the read count tables were merged into one *Seurat* object and additional QC was performed based on nucleosome signal, TSS enrichment, blacklisted regions and reads in peaks fractions, as well as setting min and max limits for number of cells having count per feature and number of features having count per cell.

We then binarized the data and used Latent Semantic Analysis (LSA) [3,8] approach to reduce the dimensionality. The first singular component was omitted as it significantly correlated with read depth. Remaining 49 components were used to construct hierarchical tree of cellular clades with using cosine distance and ward.D2 as linkage methods.

Next, we computed Dunn2 and Silhouette statistics [9] at various cut heights of the cellular clade tree (see Supplementary Figure 1A) by using R package *fpc*. Based on these statistics we selected cutting height (5.5) which resulted in in 93 cellular clades, with 72-1816 cells per clade (Supplementary Figure 1B). First, the reads from each clade were processed separately, using *Macs2* for the peak detection, thus each clade had equal possibility to provide its own peaks to merged collection [3]. Subsequently, peaks from all clades were merged [7] to form joined feature space [2].

Feature comparisons and characterization

In the comparisons of sample-specific features in VENN diagrams, the features were considered overlapping when they shared an overlap of at least of 100 bp. Features of the joined feature space were further characterized by using *ChIPseeker* [10] and visualized with *circlize* [11] R packages (Figure 2, Supplementary Figure 1B).

Processing of scATAC data with common feature space

Jointly defined features were then used to process the scATAC data as previously described, but instead of fixed 5 kbps windows, the features of the joined feature space were used. Data was subjected to QC, binarization and LSA, steps described in more detail in (https://github.com/ComputationalNeurogenetics/NeuronalFeatureSpace) largely following steps described by Stuart et al. [4]. scATAC data was integrated with scRNA using an approach described by Hao *et al.* [8].

UMAP projection based on scATAC data was done with *Seurat* by using SVD components 2-59 for cells with reliable score from scRNA integration. First component was omitted due to the association to read depth. Neighborhood and community detection were made with Seurat functions of *FindNeighbors*(dims=2:59) and *FindClusters*(resolution=8). Clustering resolution was optimized with *chooseR* [6] as described below.

Identification of neurotransmitter type

Neurotransmitter types for each cell were identified by using *ScType* [7] from which we returned cell level typing, instead of more typical cluster level type label, and decided to add an 'unknown' label on cell level based on not having score for any defined cell type higher than mean(scores) + (1.3*sd(scores)). *ScType* was applied on integrated RNA data in scATAC object and separately to RNA data in scRNA object. The neurotransmitter phenotype of the cell types was identified based on the expression of positive and negative marker genes describing each neurotransmitter type (Supplementary Table 1). In addition to the different neurotransmitter types, we created 'progenitor' label using artificial composite genes called *g2m* and *s*, which describe the cells in G2/M or S-phase of the cell cycle, respectively. We adjusted a *gene_set_prepare* function in *ScType* to accept our added artificial composite genes (code in https://github.com/ComputationalNeurogenetics/NeuronalFeatureSpace). Otherwise, the code was run with default parameters. Marker genes for G2/M and S-phase were collected from *Seurat* cell cycle gene collection. As some of these genes are not entirely cell cycle specific in the context of developing mouse CNS, we applied additional filtering by excluding genes showing expression in clearly postmitotic neuronal cells in our dataset (Supplementary table 1).

Clustering resolution optimization

We used *ChooseR* [6] package as wrapper to test clustering resolution parameter values between 1-20 for clustering of both scATAC and scRNA data before visualizing with UMAP. We modified *ChooseR* package code to accommodate scATAC data and Leiden clustering by adding an option to skip first SVD component in scATAC data and by changing clustering algorithm from Louvain to Leiden. The optimal resolution parameter was chosen based on silhouette values (Supplementary Table 2, Supplementary Figure 3A).

Chromatin accessibility-based similarity of cell clusters and transcriptome correspondence

Similarity of scATAC based cell clusters was further analyzed by hierarchical clustering (cosine distance and ward.D2 linkage) of their average accessibility profile across all features. Entire collection of 130 identified clusters were visualized with *circlize* (Figure 5) with additional information of brain region of origin and neurotransmitter types visualized in stacked barplots.

In analysis focusing to GABAergic cell clusters we selected all cell clusters where >0.75 proportion of cells received 'GABAergic' label from ScType, and additionally two clusters recognized to be GABAergic progenitors based on marker expression. These clusters were subjected to separate hierarchical clustering (cosine distance and ward.D2 linkage) based on their average accessibility profile across all features and visualized with *circlize* (Figure 6A). Proportion of brain region of origin was additionally visualized as stacked barplot track.

We used generalized trajectory inference method VIA [12] to examine trajectories and pseudotime among prominently GABAergic cell clusters based on chromatin accessibility (Figure 6B). Additional two clusters (21 and 90) were added to the analysis to provide prominently progenitor cells for the lineages. Details and parameters of VIA analysis are available from (https://github.com/ComputationalNeurogenetics/ NeuronalFeatureSpace), with main steps outlined here. Data subset was exported from R and subjected to VIA analysis in Python (Jupyter) based workflow where VIA was run with *pyVIA* library. Progenitor clusters were set as roots, and VIA was run both in coarse and fine mode on top of coarse results.

Correspondence between transcriptome and chromatin accessibility was analyzed by selecting all individual cells from four prominently GABAergic cell clusters from diencephalon (Figure 6A) and subjecting these cells to separate clustering based on their top 3000 most variable expressed genes (imputed RNA). Clustered heatmap in Figure 6C was drawn using euclidean distance and ward.D2 linkage.

Characterisation of differentially accessible chromatin

Pairwise differentially accessible chromatin features between all pairs of clusters were identified with *Seurat FindMarkers* function applying *min.pct=0.05*, logistic regression and latent variable set to *peak_region_fragments*. This was run in wrapper across all cluster combinations and differential accessibility (DA) was defined when avg_log2FC > 0.75 or avg_log2FC < -0.75 and p_val_adj < 0.05.
The closest gene for each DA feature was identified with *Signac ClosestFeature* with threshold of avg_log2FC > 0.7 for DA feature. Enrichment for the genes in Gene Ontology (GO) biological process categories was calculated with go_*enrich* function from R package *GOfuncR* [13].

NT purity and Gini index

To assess the biological validity of clusters, we calculated purity value [14] across clusters, as well as Gini index [15] of expression of housekeeping genes and neuronal markers genes. Both indexes were calculated based on imputed RNA. For both indices, the maximum value of 1 would indicate perfect segregation of NT-types/marker gene expression among the clusters [16]

ENCODE feature comparison

Comparisons of our features to previously known mouse genome features were calculated with *findOverlapsOfPeaks* function from *ChIPpeakAnno* R package [17], by using parameters (connectedPeaks = "keepAll", minoverlap = 1). The previously known feature definitions were obtained from ENCODE database [18] from SCREEN service for E14.5 Di, MB and Hindbrain, as well as E12.5 Hindbrain.

scATAC vs scRNA cell cluster composition

In order to assess similarities in scATAC and scRNA cell cluster composition, for every cell, the cluster ID vectors of scATAC clusters (n=130) and label transferred scRNA clusters (n=80 originally) were registered. The resulting contingency table (Supplementary table 3) was visualized in heatmap with *ComplexHeatmap* package [19]. Ordering of the rows and columns in the heatmap was calculated with R package *seriation* [16] by using BEA_TSP as method.

Results

Definition of open chromatin features in E12.5 - E14.5 neuronal cells

Here, we aimed to study the cell types among the single cells derived from the neuroepithelium of the ventral forebrain, hypothalamus, midbrain and rhombomere 1 of a developing mouse (Figure 1A). To map the open chromatin areas and the mRNA molecules expressed in the cells, we performed scATAC-seq or scRNA-seq on stage- and region matched samples (Figure 1B, Methods).

The cell type complements between the sampled brain regions are partially overlapping: all regions contain neuronal progenitors, GABAergic neurons and glutamatergic neurons. In addition, there are cell types unique to brain regions. To compare the gene expression and regulation in the cell types side-by-side, we decided to pool the cells and perform a combined analysis. While the scRNA features are already measured in a common reference frame, this is not the case for the scATAC-sequencing data. Instead, in ATAC-seq data analysis, features are formed either i) based on fixed genomic windows or ii) based on locations of aggregated peaks of reads along the genome. The latter analysis approach has clear benefit in increased resolution, but the results vary between samples. Analysis of the set of pooled cells originating from different scATAC-seq experiments thus requires some extra steps of defining the joined feature space applicable across the samples. We addressed this by merging the reads from E14.5 R1, E14.5 MB, E14.5 DI and E12.5 R1 scATAC-seq experiments, and then defining features stepwise: we first used the reads from cells that formed high-level clades separately. To determine optimal cutting level for clade detection we used Dunn2 and Silhouette statistics [20] (Figure 1C, Supplementary Figure 1A). After clade definition, we performed separate peak detection in each of the clades. Clade level peak definition was followed by peak merging (Figure 1C; Supplementary Figure 1, see Methods for details). This allows rarer cell types to contribute to peak detection more equally. Finally, we combined the peak definitions, resulting in a final set of n= 482 157 features (Figure 1C).

Chromatin features are enriched around genes and are accessible in a temporally and spatially differential manner

We further characterized the defined features in terms of feature length, density, and localization to genomic context. Defined features were found to be mostly under 1000 bp long (Supplementary figure 2A) and covered entire mouse genome (mm10) with numerous hotspots throughout the visible genome (Supplementary figure 2B). Altogether, the defined features cover approximately 7.1% of the mouse genome sequence (193 825 312 bp out of total of 2 730 855 475 bp). Features were found to be centered around TSS (transcription start site) of genes (Figure 2A). Nearly 30% of the features are +/- 10kb around TSS, which altogether is relatively narrow part of the genome. Approximately 80% of the features are +/-100kb around TSS, already containing intergenic regions. Average distribution profile of the feature locations over gene models from TSS to TTS (transcript termination site) clearly peaks at the TSS but also shows that in some cases, regions downstream of TTS are highly accessible (Figure 2B). This highlights the fact that while TSS regions are in general the most dynamically regulated areas of chromatin, there are also significant stretches of intergenic chromatin that is under dynamic regulation of accessibility. We also annotated the features by the genomic region (Figure 2C-D). In total, 45% of the features are found in intergenic regions while 55% are located within gene structures (Figure 2C-D).

To assess how brain region specific features are distributed in joined feature space, we also defined open chromatin features in each brain region separately. Most of the features present in the separately analyzed samples were included in the jointly defined feature space (Supplementary Figure 1C for E14DI). However, the joined feature space contains significantly more features than any brain

region alone (Supplementary figure 1D-E). Likely, combining reads from the samples enhances the signal from weakly detected, but otherwise commonly used features. We calculated the overlaps between the features detected in the brain regions and embryonic stages (Supplementary Figure 1F-G). The majority of the features were detected in all analyzed samples, and 20-40% of features in two or three samples.

Next, we compared our open chromatin feature set to the functional DNA elements in ENCODE [18]. As the chromatin accessibility varies in cell types and developmental stages, we considered the ENCODE features that are detected in the E14.5 mouse hindbrain, midbrain and diencephalon, and E12.5 hindbrain, collectively. 82.45% of the ENCODE elements are found in our feature set (64634/78390 elements). However, our feature set contains 417 509 additional features, making the ENCODE elements a minority (15.48%) of the defined features. Aside from the noise inherent to single-cell data, our feature set may be enriched for rare features that are expressed transiently and not to a significant extent in the cell type level. To ask whether the significantly expressed features in our data better match ENCODE elements, we calculated the overlap between the ENCODE elements from E14.5 diencephalon, midbrain and hindbrain and the 213 175 features used in clustering of E14.5 integrated data (see below, Figure 1D). 38.69% (59463/213175) of those features overlap with ENCODE elements (Figure 2E-F). The features overlapping with ENCODE are enriched in promoters (Figure 2G-H).

Cell type identification and comparison between brain compartments

Having validated the jointly defined feature space for all studied brain regions, we next used this feature space to process the scATAC-seq data from E14.5 DI, MB and R1 in a combined fashion (Figure 1D). We then integrated the scATAC-seq data with scRNA-seq data as described by Hao *et al*. [2] and Stuart *et al.* [8]. The integration and further analyses were here focused on the E14.5 stage and therefore exclude the data from E12.5 R1 in downstream analysis steps. Including the E12.5 R1 samples in the feature space definition steps keeps the feature space compatible with future studies. Using joined feature space to process the combined E14.5 scATAC-seq data resulted in the dataset of 213 175 expressed features over 12 903 cells, after filtering features based on detected accessibility in at least 1% of the cells and in maximum of 97.5% of the cells, and cells based on having at least 2.5% of features accessible and maximum of 97.5% of features accessible.

Following the sample merging and data integration, we proceeded to cluster the E14.5 cells based on their scATAC profile (Figure 3A). The resolution parameter for cluster detection was estimated by iterating cluster detection over range of resolution values and selecting the highest statistically supported resolution using *chooseR* [6] (Supplementary Figure 3A-B). To characterize the cells in each cluster based on their neurotransmitter type or progenitor cell identity, we established *ScType* [7] with a customized reference database (Supplementary Table 1, Figure 3A-C). We observed high level of concordance between cluster identity and neurotransmitter phenotype (NT type), with NT type purity value [14] being 0.85. To analyze the enrichment of known neuronal marker genes or housekeeping genes among the differentially expressed genes in the clusters, we calculated Gini index [15]. Mean Gini index was 0.80 for the neuronal marker genes and 0.12 for the housekeeping genes, indicating that housekeeping gene expression is relatively equal across the clusters, whereas expression of neuronal marker genes is highly variable across the clusters. Consistent with the NT type score, the markers for the cell types present in the diencephalon, mid- and hindbrain at E14.5 were differentially expressed in clusters (Figure 4). The identified cell types include the GABAergic and glutamatergic neurons derived from ventrolateral and lateral neural tube, dopaminergic and serotonergic neurons lineages derived from floor plate and basal plate (Figure 4), and glial lineages (astrocytes and oligodendrocytes) derived from lateral and ventral neural tube (data not shown).

Next, we aimed to further explore the relationships between cell clusters and cell types. For that, we calculated average chromatin accessibility profile for each cell cluster and subjected those to hierarchical clustering (Figure 5). To further understand the scale of difference in chromatin accessibility between these cell clusters, we performed pairwise tests of differentially accessible chromatin between all clusters. Surprisingly, in overwhelming majority of cases significant accessibility differences between clusters comprised less than 100 features out of over 213 175 (Supplementary figure 4A, Methods). The total length of chromatin (sum of kb in the DA features) found to be differentially accessible between clusters ranged from 0.7-1350 kbp, with the mean of 176 kbp (Supplementary figure 4B).

Hierarchical clustering by chromatin accessibility revealed some segregation of clusters by both brain region of origin as well as by neurotransmitter type (Figure 5). However, there seems to be no clear preference to group by either characteristic. It would be interesting to further analyze the different branches or the cluster tree that comprise the similar neurotransmitter type, for example GABAergic neurons. Such branches represent cell types with different chromatin accessibility configuration that yet can reach and maintain similar cellular phenotype.

Qualitative differences between the transcriptomic and epigenetic features in defining cellular state

The accessible chromatin areas affect gene expression and thus largely define RNA expression in the cells. To explore the correlation between the cell clusters formed based on accessible chromatin regions or on mRNA expression in our data, we first clustered the cells based on the scRNA, independently of the scATAC (Supplementary Figure 5A). The clustering based on scRNA resulted in 80 clusters at the optimal resolution (Supplementary Figure 5A). We next calculated the cluster-to-cluster matches between the scATAC and scRNA cell clusters (Methods, Supplementary table 3). We observed high level of concordance between the cell clustering. We also compared the distribution of cells with each neurotransmitter phenotype between the clusters (Supplementary Figure 5C-D). In the matching clusters, the cell types and their ratio were mostly similar (Supplementary Figure 5D, note for example the adrenergic neurons, Adr; and progenitor cell groups, Prog; dopaminergic neuron groups D). Nevertheless, cells could also be grouped differently based on the scRNA and scATAC modalities (Supplementary Figure 5D, mixed clusters S+D vs S+Unk, indicated with #). However, there are some differences between the modality specific clustering results. Consistent with earlier reports (Ma et al, 2020), many clusters were defined at higher resolution in either modality. As the scATAC data contains more clusters, the cells forming scRNA based cluster often match cells from several scATAC clusters (Supplementary Figure 5D, indicated GABAergic (GB) and Glutamatergic (GLU) cell clusters). Nevertheless, opposite one-to-many cluster matches also appear. Notably, in cluster integration step, ca 30% of the cells scRNA samples could not be matched to the cells in scATAC samples, probably due to the low cell number or low sequencing signal in those cell types. This indicates that the sparsity of data might limit the efficiency of cluster matching. However, the efficiency and accuracy of matching is not systematically tested or quantified here.

Alternatively, the differential grouping of cells between modalities can be explained by differential gene expression. Indeed, the cells forming the R1 dorsal/dorsolateral GABAergic neuron clusters distribute differently in the scATAC or scRNA based clustering and show clear differences and even within-cluster substructure in mRNA expression profiles (Supplementary Figure 6). It is possible that the optimal resolution reached here does not allow definition of all the cell types or cellular stages present in the E14.5 mouse brain. This might be improved by sampling more cells.

GABAergic neuron precursors in E14.5 mouse brain cluster based on the selector gene expression and differentiation stage

We next focused our analysis on the GABAergic neuron subclusters. The predominantly GABAergic clusters were defined as clusters containing >75% cells with NT-type=GABAergic with cluster mean silhouette reliability score > 0.5. We selected these clusters (n=36) and studied their relatedness in terms of chromatin accessibility profiles in isolation. Hierarchical clustering of the GABAergic neurons resulted in a complex tree structure in terms of brain region of origin (Figure 6A). During development, GABAergic neuron precursors express proneural TF *Ascl1* and a combination of selector transcription factors *Dlx1/2, Gata2/3* and *Tal1/2* that are required for the acquisition of GABAergic phenotype. We asked how these known GABAergic fate regulators are expressed in the clusters (Figure 6A). We observe clear differential expression of *Ascl1*, *Tal1, Gata2, Gata3, Dlx1* and *Dlx2* correlating with the clustering (Figure 6A). GABAergic neuron marker genes *Gad1* and *Gad2* were expressed in all the identified GABAergic neuron clusters. Consistent with previous studies, the cells originating from diencephalon (E14.5 DI) expressed *Dlx1* and *Dlx2*, whereas the cells originating in midbrain and rhombomere 1 expressed *Tal1*, *Gata2* and *Gata3* in various combinations. In E14.5 DI, we distinguish five different cell clusters of predominantly GABAergic phenotype. The four *Dlx1/2*+ clusters (18, 19, 60, 122) form one distinct branch (branch b) in the DA feature -based hierarchical clustering tree as well as pseudotime trajectory (Figure 6B). Some E14.5 DI cells are found in cluster 34 that expresses *Gata3* and could thus contain the derivatives of the P2R or P1 region of the diencephalon [21].

Interestingly, while some DI and R1 GABAergic clusters form distinct groups, the R1 and MB-derived GABAergic cells also form mixed branches. The branch c can be subdivided in groups of differentiating neurons or mixed origin (c1, expressing *MAPT, Tubb3*), early midbrain-derived GABAergic neuron precursors (c2, *Ascl1*) and intermediate, Tal1+ R1 precursors (c3). Branch d also contains *Tal1/2, Gata2/3* positive, intermediate GABAergic neuron precursors of mixed origin. Some of the mixing might be explained by contamination by the cells from the neighboring regions and/or cell migration. Branch a and e are more uniform in regional identity and gene expression, containing *Ascl1, Tal1, Gata2, Gata3* positive, ventrolateral R1 precursors (branch a) and previously characterized dorsal and dorsolateral rhombomere 1 GABAergic cell lineages (branch e) [22]. In conclusion, the grouping of midbrain-R1 GABAergic neuron clusters shows a loose correlation with the region of origin, differentiation stage and selector gene expression. While early-stage precursors group according to the positional identity, at the late progenitor/early precursor stage, the cells of distinct lineages may display a transient variation in genome accessibility, which leads to similar priming of GABAergic differentiation genes. This phenomenon has earlier been characterized as *phenotypic convergence* [23].

To further study the variation of genetic landscape associated with the transition from an early precursor to late precursor and differentiating neuron stage, we focused on the four closely related diencephalic GABAergic clusters. These cell clusters are found at subsequent positions in the pseudotime trajectory (Figure 6B, arrows). Two distinct lineages could be identified: the anterior, ventral forebrain (MGE/POA, positive for *Lhx6, Lhx8, Nkx2-1, Foxg1*) and more posterior, diencephalic prosomere 3 (P3, *Nkx2-1, Six3*, *Lhx1, Sst, Npy*) lineage. We next compared the gene expression differences between the cluster constituent cells and the differential feature accessibility analysis in each step of the pseudotime (Figure 6C-D). Surprisingly, the total length of variable chromatin between the clusters is small, ranging from 250-800 kbps. The DA chromatin length correlates with the proximity of the cell clusters in the pseudotime trajectory, where the closest placed clusters show the least amount of DA features. GO term enrichment analysis showed that the genes nearest to the DA features are greatly enriched to genes involved in neuronal function or fate (*modulation of chemical synaptic transmission, growth cone, neuronal cell body, dendrite*) or development (*neuron projection morphogenesis, dendrite development, axon guidance, cell-cell adhesion, cell*

morphogenesis involved in neuron differentiation) (Figure 6 E-H). We observed a bias towards having more DA features in the more differentiated cells when compared to less differentiated along the pseudotime axis, however, this has not been further quantified or validated across all cluster pairs yet.

Discussion

The advantages and limitations of the study

As open chromatin features are stage- and tissue specific to a considerable extent, our study provides a pertinent method to combine data from different experiments. Nevertheless, due to the high context specificity, still additional experiments may be required to thoroughly understand the regulation of cell differentiation. The chromatin features defined in this study cover 7.1% of mouse genome and can be used for the analysis regulatory landscape and its dynamics in the E14.5 mouse CNS. However, using predefined areas for scoring chromatin openness has several limitations. First, dynamic use of chromatin outside of predefined features is not considered.

Second, we lack information of the way feature is used in genome regulation (steric or regulative interactions), and for gene regulatory features, feature-to-target gene links can be challenging to define. Defining the target genes for the genomic features of regulatory TF binding events is not trivial neither, especially for the features that locate far away (>100 Mb) from genes. Enhancers can associate with the target gene promoter by chromatin looping and their target genes can be either up-or downstream of the feature. Several genes have intergenic enhancers further confounding the model of regulation. Third, we were able to characterize the cell clusters using the RNA expression inferred from a stage- and region- matched dataset. Nevertheless, rather than the inferred RNA expression level, true multimodal -omics data would allow to better assess the match between the RNA expression-based clustering vs clustering by open chromatin features. On the other hand, the mRNA expression response from chromatin accessibility change might require some distance in time [24,25].

Finally, mRNA expression and open chromatin features describe only a fraction of the cell. Post-transcriptional aspects such as the alternative splicing of RNA could not be addressed in this study. scRNA also lacks information of the regulatory RNA molecules and non-polyadenylated RNA, such as long non-coding RNA. Providing a true snapshot of the genetic regulation network (GRN) in a cell would also require in-parallel information of protein expression in the cells. Currently available methods limit the proteomics and protein-protein interaction studies as well as studies of post-translational regulation of protein expression and activity in a comparable cell type resolution.

Detailed characterization of the regulative genomic features in mouse CNS development

Regardless of the limitations, our results and data comprising cells of numerous types and developmental stages, will be essential addition to the studies of neurogenesis and early fate specification of neurons. These earlier developmental stages are under-represented in the current brain cell atlases yet feature high level of regulatory interactions.

In this study, we have demonstrated the use of open chromatin features for clustering single cells, reaching a high resolution where ‘cell types’ are separated by only hundreds to thousands of base pairs difference in open chromatin. It is yet unclear whether the observed difference is regulative of cell fate decisions. Cell fate is regulated by inter- and intracellular signalling that is, to a large extent, mediated by transcription factors. Transcription factor binding profile in differentially accessible

chromatin features could therefore be indicative of developmental history. The TF and RNA polymerase binding at promoters and enhancers is accompanied by the modification of histones. Histone modification profile could thus also be utilized as supporting data during GRN reconstruction. Importantly, we have here integrated the chromatin feature usage and the mRNA expression data in single cells. mRNA expression adds multiple layers of value, first, it provides readout of gene transcription, and second, information about the molecular function of the cell or cell type.

Our comparison of the differentially expressed genes and the genes near DA features in the respective clusters shows that the achieved resolution allows extraction of regulatory events associated with changes of cell state or progression along differentiation steps. In the future, it would be interesting to dissect the fractions of chromatin specifically associated with distinct but interlinked developmental events: such as the cell fate determination and regional patterning.

Cell type and state

Which modality, chromatin accessibility or the mRNA expression, is a better predictor of cell identity? Due to both technical and biological variability among single cells, cell type should not be defined as a transcriptionally and epigenetically fixed entity, but the signature should allow certain level of variation. Our cross-modality cluster comparisons indicate that we can detect different transcriptional and epigenetic states within cell types. Compared to the number of genes, the higher number of open chromatin features allows to reach higher clustering resolution. Maximum resolution was achieved here by combining two essential steps: first, forming cell clades before feature definition allows to detect and consider rare and cell-type specific features, and second, using the highest clustering resolution achievable for given features and number of cells. Indeed, the number of clusters detected in the scATAC-seq experiment is double compared to the number of clusters detected in the scRNA-seq data at optimal resolution. This has earlier been interpreted as segregation of different cell states [26]. Alternatively, the multiple matching effect could be due to the use of alternative gene regulatory programs. In this case, rather than cell state, the apparent 'resolution' reflects parallel regulatory or co-regulatory events, leading to similar mRNA expression profile. Careful analysis of marker gene combination is required to address this issue.

Outlook

Here, we have demonstrated an approach to study divergent gene expression regulation events in cell types within and across regional boundaries. The study can easily be extended to various questions in developmental biology and gene expression regulation. We demonstrate how specific points of the "cell type tree" can be analyzed in detail for the specific changes in enhancer activation. For the gene regulatory network reconstruction further on, the enhancer usage should be associated with the target gene expression and correlated with TF binding events.

Author contributions

SK and AK conceptualized the study, planned the experiments, wrote the manuscript and compiled the figures. SK contributed to the formal analysis, implementation of methods, as well as manuscript writing and editing. HH established the ScTyping method. KA contributed to single cell RNA-seq and single cell ATAC-seq experiments, interpretation of results, manuscript writing and editing, supervising the project and acquisition of funding.

Acknowledgements

We thank FIMM single-cell genomics core facility, Jenni Lahtela and Bishwa Ghimire for the advice and assistance in sample processing and data analysis.

This research was supported by the Academy of Finland, decision numbers 319906 and 345737.

Figure legends

Figure 1. Graphical overview of the presented workflow.
A. Outline of mouse embryo with studies brain regions labeled. B. Number of samples collected per brain region and single-cell sequencing method (scRNA-seq or scATAC-seq) used. Each sample consists of ca 5000 cells or nuclei from a single embryo. C. Outline of the process for defining joined feature space across studied brain regions. D. Outline of the clustering based on the chromatin accessibility (scATAC-seq data scored along the previously defined features) and the integration of scATAC-seq and scRNA-seq data. DI, diencephalon; MB, midbrain; R1, rhombomere 1; GE, ganglionic eminences of ventral forebrain (subpallium); poa, preoptic area; LSA, Latent Semantic Analysis ; QC, Quality Control; PCA, Principal Component Analysis; scATAC, scATACseq, single-cell ATAC-sequencing; scRNA, scRNAseq, single-cell mRNA-sequencing.

Figure 2. Characterisation of the joined feature space.
A. Distribution of features relative to the TSS of genes in mouse genome. X-axis shows cumulative percentage of features locating upstream or downstream of the TSS of genes. Colors indicate distance intervals from TSS in kb. B. Count frequency of feature peaks location across average profile of gene from TSS to TTS (5´-3´direction). C. Pie chart of features annotated for the type of genomic region they are overlapping with. D. Venn-pie diagram showing amount of overlapping annotation labels as features can overlap more than one type of genomic region. E-F. overlap with the features expressed in the E14.5 FB, MB and R1 and E12.5 R1 according to ENCODE resource and according to our scATACseq experiments. G-H. Genomic regions of the features overlapping with the ENCODE features (H) and the features not overlapping with the ENCODE features (G).

Figure 3. Clustering of the E14.5 Di, MB and R1 cells based on the scATAC-seq feature expression.
A. UMAP projection of the clusters of E14.5 cells. Clustering is based on the scATAC-seq feature expression. Community detection-based clusters are colored and numbered. B. UMAP of the clusters showing the neurotransmitter identity of the cells in colours. C. UMAP of the clusters showing the brain region of origin for each cell.

Figure 4. Neurotransmitter phenotypes and the related marker gene expression in the E14.5 mouse brain.
A. Expression of known neuronal marker genes over the UMAP projection of cell clusters. Level of gene expression is indicated with the intensity of blue color. The cell type is indicated for each marker gene set. The shown marker genes are used in the NT-typing method (see Supplementary Table 1 for the reference database).

Figure 5. Analysis of the cluster tree by the region of origin and neurotransmitter phenotype of the cells.
A. Circos plot showing the dendrogram of E14.5 cell clusters, the neurotransmitter type per cell as well as brain region of origin. Outermost track: cluster numbers corresponding to the numbering in UMAP (Figure 3). Stacked barplot in second track shows the proportion of the cells derived from different brain regions in the cluster. Third track: the proportion of cells of different neurotransmitter identity per cluster.

Figure 6. The differential chromatin accessibility associated with the GABAergic neuron differentiation in different brain regions.
A. Circos plot showing the dendrogram of dominantly GABAergic clusters in E14.5 mouse brain and the expression of genes known to regulate GABAergic fate acquisition. The tracks show proportion of brain region origins (inner track), and the average expression of the key proneural genes (Ascl1, Helt) and selector genes (Tal1, Gata2, Gata3, Dlx1, Dlx2) known to drive GABAergic neuron

differentiation per cluster. B. Pseudotime analysis of the GABAergic cell clusters and respective progenitor cell clusters. The lineages representing the diencephalic prosomere 3 (P3) and preoptic area/medial ganglionic eminence (PA/MGE) GABAergic neurons are indicated with arrows. C. Analysis of the transcriptomes of the Dlx1/2 expressing GABAergic neuron clusters based on top 1000 variable genes. Both axes have been clustered with Euclidean distance and ward.D2 linkage. scATAC cluster of origin of each cell is shown on the right. D. Dlx1/2 expressing clusters clustered by the similarity in the differentially accessible (DA) DNA features. Heatmap shows the total length of the DA chromatin. E-F. Position of the DA features and the nearest gene to each feature. The number of DA features near a gene is indicated in brackets. The numbers of the compared clusters are indicated in the middle. Red and blue colour indicates the cluster where a DA feature is upregulated. G-H. Enrichment analysis across Gene Ontology database categories. The numbers of the compared clusters are indicated.

Supplementary Figure legends

Supplementary Figure 1. Distribution of cells into clades and the comparison of the features defined on individual scATAC-seq samples and the joined feature space.
A. Dunn2 statistic (green) and cluster average Silhouette statistic (red) for each configuration of clusters acquired by cutting cellular tree from heights 4 to 10, with 0.5 steps. The method seeks for maximum of both values, and minimal cutting height where distinct cell populations are separated. B. Dendrogram of cellular clades after applying the cutting height 5.5. The number of cells per clade and the proportion of cells derived from each brain region per clade. Outermost track contains clade numbers and the number of cells per clade. Stacked barplot in second track show proportions of cells originating from various brain regions. C. Histogram of accessible features in the E14.5 DI sample that overlap with the features in joined feature space, stratified by the proportion of overlap. D. Histogram of jointly defined feature space features that overlap with the features defined using only E14.5 DI (E14.5 DI features), stratified by the proportion of overlap. E. Venn diagram of feature space correspondences between all four samples and the jointly defined features.F. Venn diagram between joined feature space (joint.space) and the features detected in E14.5 DI and E14.5 MB samples. Features were considered overlapping when they contained at least 100bp overlap in sequence. G. Venn diagram between joined feature space and the features detected in E12.5 R1 and E14.5 R1 samples.

Supplementary Figure 2. The length and genomic position of the scATAC features.
A. Histogram of the length of the features (bp) included in the joined feature space. B. Density plot of the features aligned to the mouse mm10 genome.

Supplementary Figure 3. Clustering resolution, cluster reliability and the similarity of labels after the scRNA and scATAC based clustering.
A. Silhouette statistic scores per clustering resolution calculated using chooseR. Optimal resolution is indicated with red line. B. Silhouette scores per cluster overlayed on the UMAP projection of clusters. Higher values (max 1) indicate clusters with high reliability. Note that the clusters 65 and 74 get clearly low reliability score. In further analysis, the representativeness of actual separate cell types should be carefully considered for these clusters.

Supplementary Figure 4. The number and overall length of differentially accessible DNA features in the E14.5 Di, MB and R1 cells.
A. Histogram of the number of differentially accessible features between the cell clusters. B. Clustered heatmap of total kb of differentially accessible chromatin between all pairwise combinations of cell clusters.

Supplementary Figure 5. Clustering and NT-type annotation of the combined scRNA data from E14.5 Di, MB and R1.
A. UMAP of the scRNA clusters. The cluster resolution is chosen according to chooseR optimization. B. The brain region labels overlaid on the cluster UMAP. C. NT type labels overlaid on the scRNA cluster UMAP. D. Heatmap of cluster-to-cluster matches between the integrated and scRNA-seq based clusters. The stacked barplots next to the cluster numbers show the proportion of cells of each neurotansmitter type in the respective cluster. The matching clusters mostly received the same NT-type label.

Supplementary Figure 6. Heatmap of the RNA expression in the D/DL R1 GABAergic cell clusters.
Cells (y-axis) were reclustered based on top 1000 variable genes (x-axis). Both axes are clustered applying Euclidean distance and ward.D2 linkage. scATAC cluster ID is shown on the right.

Supplementary Table 1.
Tab 1: ScType database for assigning the NT type labels. Tab 2: CellCycleScore genes expressed in the E14.5 mouse neuronal progenitors and list of genes used in g2m and s artificial gene for ScType.

Supplementary Table 2.
Results of chooseR iteration for optimal clustering resolution.

Supplementary table 3.
Contingency table of cell placement in scATAC clusters and scRNA clusters.

References

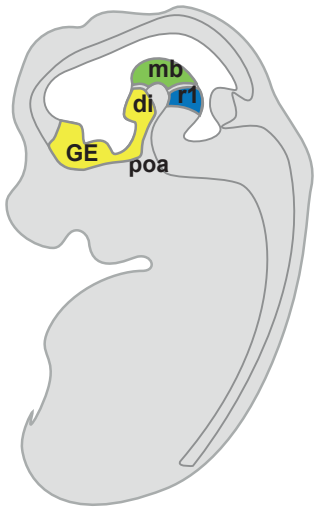
1. Shema, E., Bernstein, B.E., and Buenrostro, J.D. (2019). Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. Nat Genet *51*, 19–25. 10.1038/S41588-018-0290-X.
2. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck 3rd, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., *et al.* (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587 e29. 10.1016/j.cell.2021.04.048.
3. Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., *et al.* (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell *174*, 1309-1324.e18. 10.1016/j.cell.2018.06.052.
4. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902.e21. 10.1016/j.cell.2019.05.031.
5. Balderes, D.A., Magnuson, M.A., and Sussel, L. (2013). Nkx2.2:Cre knock-in mouse line: A novel tool for pancreas- and CNS-specific gene deletion. genesis *51*, 844–851. 10.1002/DVG.22715.
6. Patterson-Cross, R.B., Levine, A.J., and Menon, V. (2021). Selecting single cell clustering parameter values using subsampling-based robustness metrics. BMC Bioinformatics *22*, 1–13. 10.1186/S12859-021-03957-4/FIGURES/3.
7. Ianevski, A., Giri, A.K., and Aittokallio, T. (2022). Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun *13*. 10.1038/s41467-022-28803-w.

8. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902.e21. 10.1016/j.cell.2019.05.031.
9. Akhanli, S.E., and Hennig, C. (2020). Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. Statistics and Computing 2020 30:5 *30*, 1523–1544. 10.1007/S11222-020-09958-2.
10. Yu, G., Wang, L.G., and He, Q.Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics *31*, 2382–2383. 10.1093/BIOINFORMATICS/BTV145.
11. Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize implements and enhances circular visualization in R. Bioinformatics *30*, 2811–2812. 10.1093/BIOINFORMATICS/BTU393.
12. Stassen, S. v., Yip, G.G.K., Wong, K.K.Y., Ho, J.W.K., and Tsia, K.K. (2021). Generalized and scalable trajectory inference in single-cell omics data with VIA. Nat Commun *12*. 10.1038/S41467-021-25773-3.
13. Prüfer, K., Muetzel, B., Do, H.H., Weiss, G., Khaitovich, P., Rahm, E., Pääbo, S., Lachmann, M., and Enard, W. (2007). FUNC: A package for detecting significant associations between gene sets and ontological annotations. BMC Bioinformatics *8*. 10.1186/1471-2105-8-41.
14. Manning, C.D., Raghavan, Prabhakar., and Schütze, Hinrich. (2008). Introduction to information retrieval (Cambridge University Press).
15. Farris, F.A. (2010). The gini index and measures of inequality. American Mathematical Monthly *117*, 851–864. 10.4169/000298910X523344.
16. Hahsler, M. (2017). An experimental comparison of seriation methods for one-mode two-way data. Eur J Oper Res *257*. 10.1016/j.ejor.2016.08.066.
17. Zhu, L.J. (2013). Integrative analysis of ChIP-chip and ChIP-seq dataset. Methods in Molecular Biology *1067*. 10.1007/978-1-62703-607-8_8.
18. Consortium, E.P., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., *et al.* (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. 10.1038/s41586-020-2493-4.
19. Gu, Z., and Hübschmann, D. (2022). Make Interactive Complex Heatmaps in R. Bioinformatics *38*. 10.1093/bioinformatics/btab806.
20. Hennig, C. (2019). Cluster Validation by Measurement of Clustering Characteristics Relevant to the User. Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining, 1–24. 10.1002/9781119597568.CH1.
21. Virolainen, S.-M.M., Achim, K., Peltopuro, P., Salminen, M., and Partanen, J. (2012). Transcriptional regulatory mechanisms underlying the GABAergic neuron fate in different diencephalic prosomeres. Development *139*, 3795–3805. 10.1242/dev.075192.
22. Morello, F., Borshagovski, D., Survila, M., Tikker, L., Sadik-Ogli, S., Kirjavainen, A., Estartús, N., Knaapi, L., Lahti, L., Törönen, P., *et al.* (2020). Molecular Fingerprint and Developmental Regulation of the Tegmental GABAergic and Glutamatergic Neurons Derived from the Anterior Hindbrain. Cell Rep *33*. 10.1016/j.celrep.2020.108268.
23. Konstantinides, N., Kapuralin, K., Fadil, C., Barboza, L., Satija, R., and Desplan, C. (2018). Phenotypic Convergence: Distinct Transcription Factors Regulate Common Terminal Features. Cell *174*, 622-635.e13. 10.1016/j.cell.2018.05.021.
24. Allaway, K.C., Gabitto, M.I., Wapinski, O., Saldi, G., Wang, C.Y., Bandler, R.C., Wu, S.J., Bonneau, R., and Fishell, G. (2021). Genetic and epigenetic coordination of cortical interneuron development. Nature *597*, 693–697. 10.1038/s41586-021-03933-1.
25. di Bella, D.J., Habibi, E., Stickels, R.R., Scalia, G., Brown, J., Yadollahpour, P., Yang, S.M., Abbate, C., Biancalani, T., Macosko, E.Z., *et al.* (2021). Molecular logic of cellular diversification in the mouse cerebral cortex. Nature *595*, 554–559. 10.1038/s41586-021-03670-5.

26. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., *et al.* (2020). Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. Cell *183*, 1103-1116.e20. 10.1016/j.cell.2020.09.056.

Figure 1

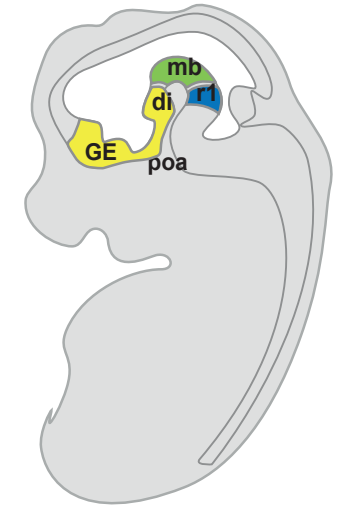
A



B

	scRNAseq	scATACseq
E14.5 Di	2	3
E14.5 MB	2	3
E14.5 R1	2	3
E12.5 R1		2

1 replicate = cells or nuclei from one embryo



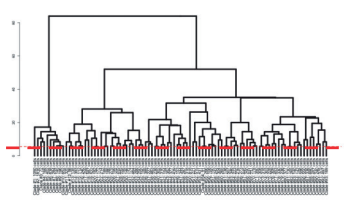
C

scATAC samples
E14.5 DI x3
E14.5 MB x3
E14.5 R1 x3
E12.5 R1 x2

↓ Replicates merged
Regions merged

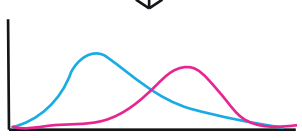
scATAC data read counts binned to 5kbp windows

↓ LSA



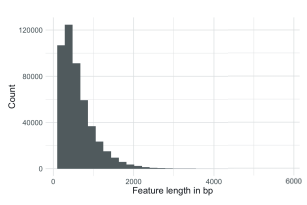
Optimal clade detection threshold

↓



Peak detection per clade

↓ Merge peak profiles from clades



Joined final feature space
n(features)= 482 157

D

scATAC samples
E14.5 DI x3
E14.5 MB x3
E14.5 R1 x3

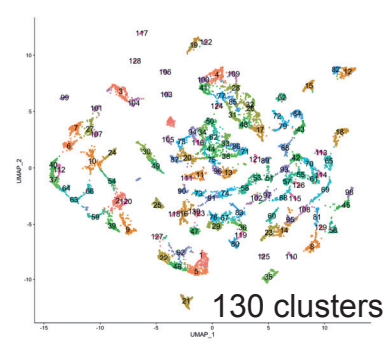
↓ Replicates merged
Regions merged

scATAC data read counts binned to joined feature space features

↓ QC + LSI

Iterative search for optimal clustering resolution

↓



↓

↓

Downstream analysis

scRNA samples
E14.5 DI x2
E14.5 MB x2
E14.5 R1 x2

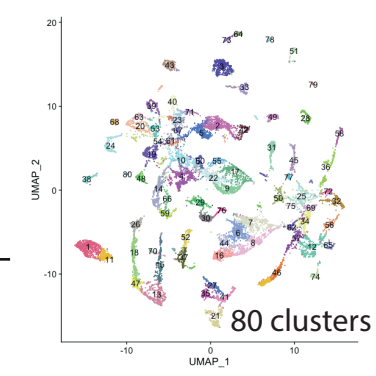
↓ Replicates merged
QC + normalization
+ PCA

scRNA data on gene level across brain regions

↓

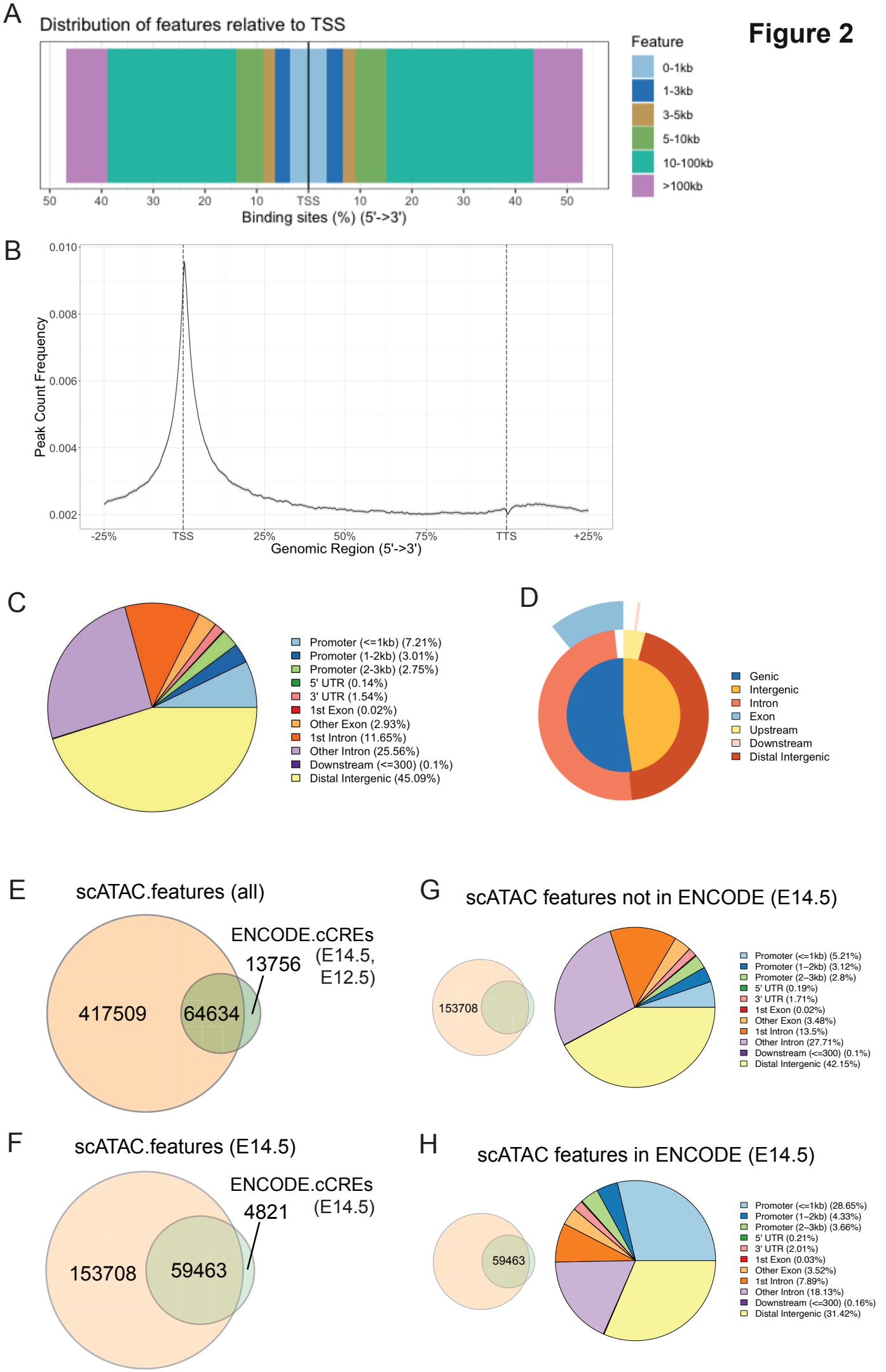
Iterative search for optimal clustering resolution

↓



← RNA level integration to scATAC data objects ←

Figure 2


A
Distribution of features relative to TSS
Feature
0-1kb
1-3kb
3-5kb
5-10kb
10-100kb
>100kb
50
40
30
20
10
TSS
10
20
30
40
50
Binding sites (%) (5'->3')
B
Peak Count Frequency
0.010
0.008
0.006
0.004
0.002
-25%
TSS
25%
50%
75%
TTS
+25%
Genomic Region (5'->3')
C
Promoter (<=1kb) (7.21%)
Promoter (1-2kb) (3.01%)
Promoter (2-3kb) (2.75%)
5' UTR (0.14%)
3' UTR (1.54%)
1st Exon (0.02%)
Other Exon (2.93%)
1st Intron (11.65%)
Other Intron (25.56%)
Downstream (<=300) (0.1%)
Distal Intergenic (45.09%)
D
Genic
Intergenic
Intron
Exon
Upstream
Downstream
Distal Intergenic
E
scATAC.features (all)
ENCODE.cCREs
(E14.5,
E12.5)
13756
417509
64634
G
scATAC features not in ENCODE (E14.5)
153708
Promoter (<=1kb) (5.21%)
Promoter (1–2kb) (3.12%)
Promoter (2–3kb) (2.8%)
5' UTR (0.19%)
3' UTR (1.71%)
1st Exon (0.02%)
Other Exon (3.48%)
1st Intron (13.5%)
Other Intron (27.71%)
Downstream (<=300) (0.1%)
Distal Intergenic (42.15%)
F
scATAC.features (E14.5)
ENCODE.cCREs
(E14.5)
4821
153708
59463
H
scATAC features in ENCODE (E14.5)
59463
Promoter (<=1kb) (28.65%)
Promoter (1–2kb) (4.33%)
Promoter (2–3kb) (3.66%)
5' UTR (0.21%)
3' UTR (2.01%)
1st Exon (0.03%)
Other Exon (3.52%)
1st Intron (7.89%)
Other Intron (18.13%)
Downstream (<=300) (0.16%)
Distal Intergenic (31.42%)

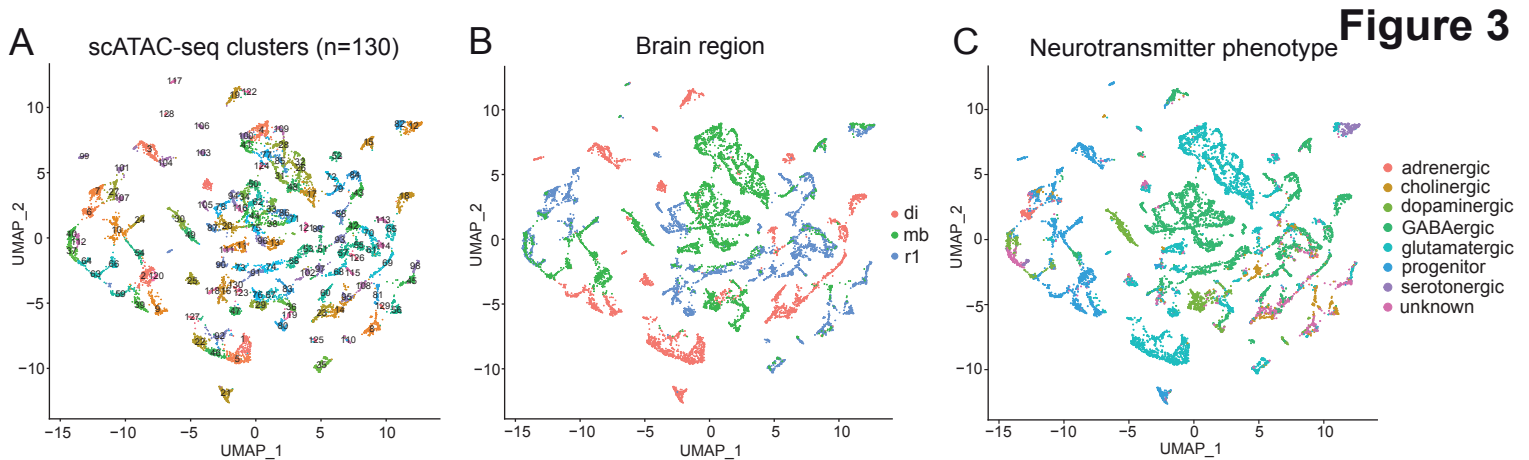
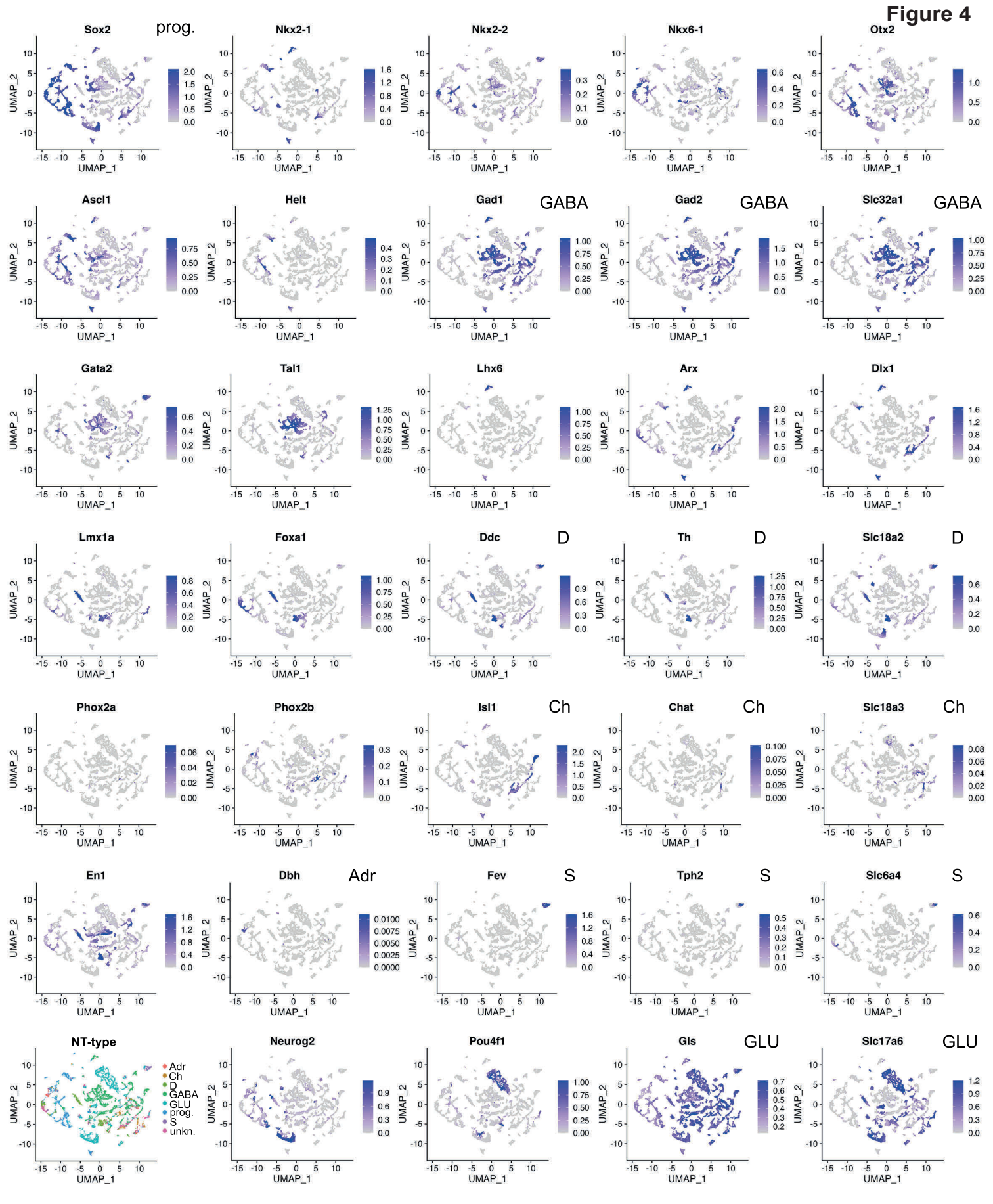


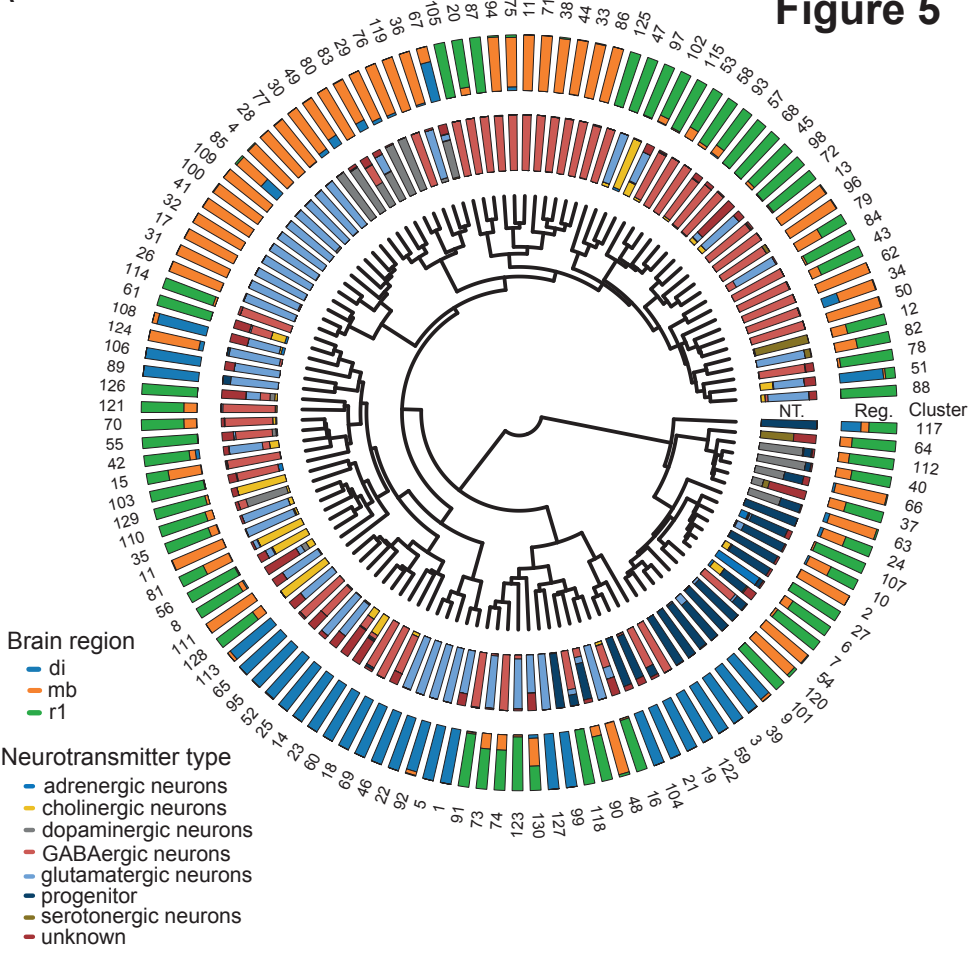
Figure 3
A
scATAC-seq clusters (n=130)
B
Brain region
C
Neurotransmitter phenotype
UMAP_1
UMAP_2
di
mb
r1
adrenergic
cholinergic
dopaminergic
GABAergic
glutamatergic
progenitor
serotonergic
unknown

Figure 4


Sox2
prog.
Nkx2-1
Nkx2-2
Nkx6-1
Otx2
Ascl1
Helt
Gad1
GABA
Gad2
GABA
Slc32a1
GABA
Gata2
Tal1
Lhx6
Arx
Dlx1
Lmx1a
Foxa1
Ddc
D
Th
D
Slc18a2
D
Phox2a
Phox2b
Isl1
Ch
Chat
Ch
Slc18a3
Ch
En1
Dbh
Adr
Fev
S
Tph2
S
Slc6a4
S
NT-type
Adr
Ch
D
GABA
GLU
prog.
S
unkn.
Neurog2
Pou4f1
Gls
GLU
Slc17a6
GLU
UMAP_1
UMAP_2

A

Figure 5


NT.
Reg.
Cluster
Brain region
di
mb
r1
Neurotransmitter type
adrenergic neurons
cholinergic neurons
dopaminergic neurons
GABAergic neurons
glutamatergic neurons
progenitor
serotonergic neurons
unknown

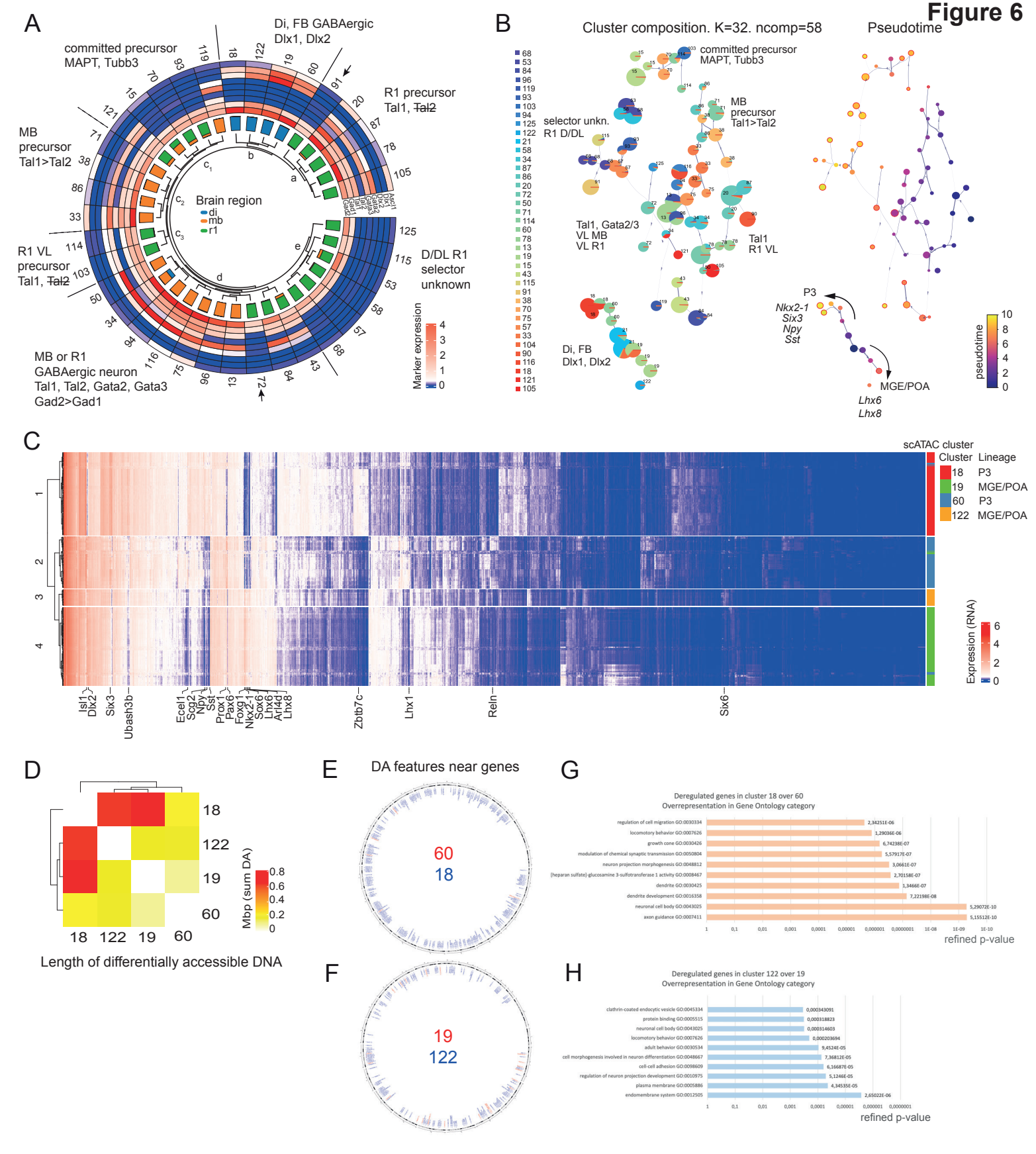
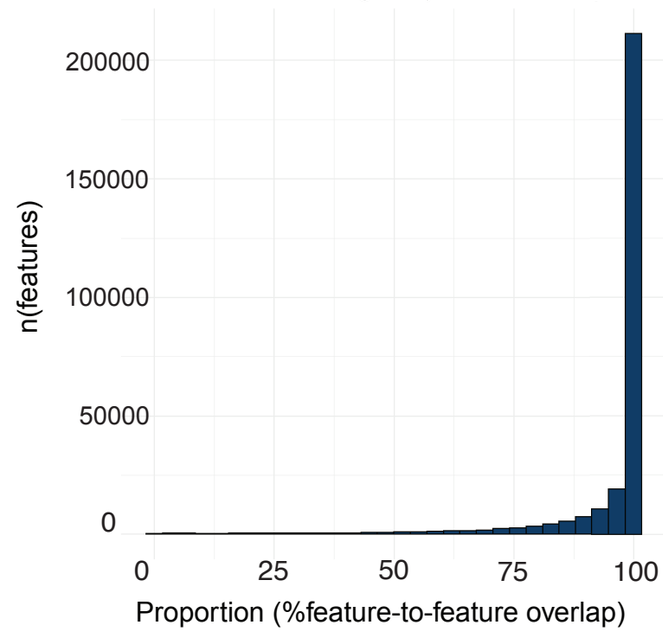


Figure 6
A
Di, FB GABAergic
Dlx1, Dlx2
committed precursor
MAPT, Tubb3
R1 precursor
Tal1, Tal2
MB
precursor
Tal1>Tal2
Brain region
di
mb
r1
R1 VL
precursor
Tal1, Tal2
D/DL R1
selector
unknown
MB or R1
GABAergic neuron
Tal1, Tal2, Gata2, Gata3
Gad2>Gad1
Marker expression
B
Cluster composition. K=32. ncomp=58
Pseudotime
committed precursor
MAPT, Tubb3
MB
precursor
Tal1>Tal2
selector unkn.
R1 D/DL
Tal1, Gata2/3
VL MB
VL R1
Tal1
R1 VL
Di, FB
Dlx1, Dlx2
P3
Nkx2-1
Six3
Npy
Sst
MGE/POA
Lhx6
Lhx8
pseudotime
C
scATAC cluster
Cluster Lineage
18 P3
19 MGE/POA
60 P3
122 MGE/POA
Expression (RNA)
Isl1
Dlx2
Six3
Ubash3b
Ecel1
Scg2
Npy
Sst
Prox1
Pax6
Foxg1
Nkx2-1
Sox6
Lhx6
Arl4d
Lhx8
Zbtb7c
Lhx1
Reln
Six6
D
Mbp (sum DA)
Length of differentially accessible DNA
E
DA features near genes
60
18
F
19
122
G
Deregulated genes in cluster 18 over 60
Overrepresentation in Gene Ontology category
regulation of cell migration GO:0030334
locomotory behavior GO:0007626
growth cone GO:0030426
modulation of chemical synaptic transmission GO:0050804
neuron projection morphogenesis GO:0048812
[heparan sulfate]-glucosamine 3-sulfotransferase 1 activity GO:0008467
dendrite GO:0030425
dendrite development GO:0016358
neuronal cell body GO:0043025
axon guidance GO:0007411
refined p-value
H
Deregulated genes in cluster 122 over 19
Overrepresentation in Gene Ontology category
clathrin-coated endocytic vesicle GO:0045334
protein binding GO:0005515
neuronal cell body GO:0043025
locomotory behavior GO:0007626
adult behavior GO:0030534
cell morphogenesis involved in neuron differentiation GO:0048667
cell-cell adhesion GO:0098609
regulation of neuron projection development GO:0010975
plasma membrane GO:0005886
endomembrane system GO:0012505
refined p-value

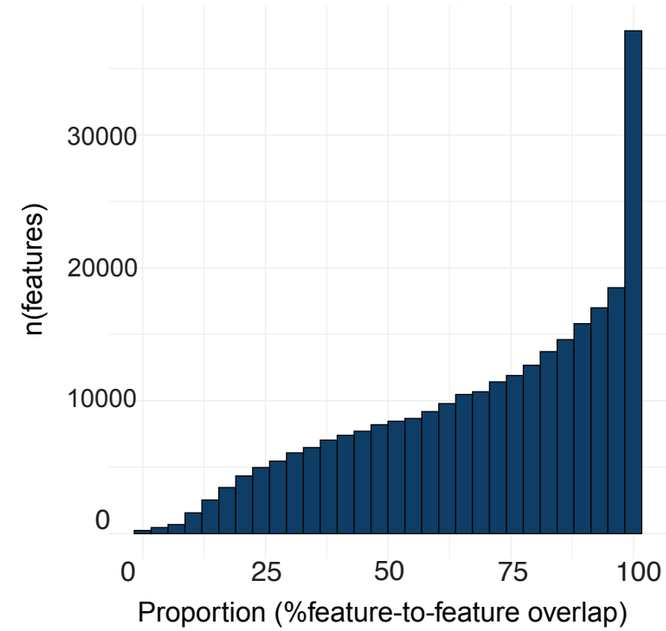
Suppl. Fig. 1

A

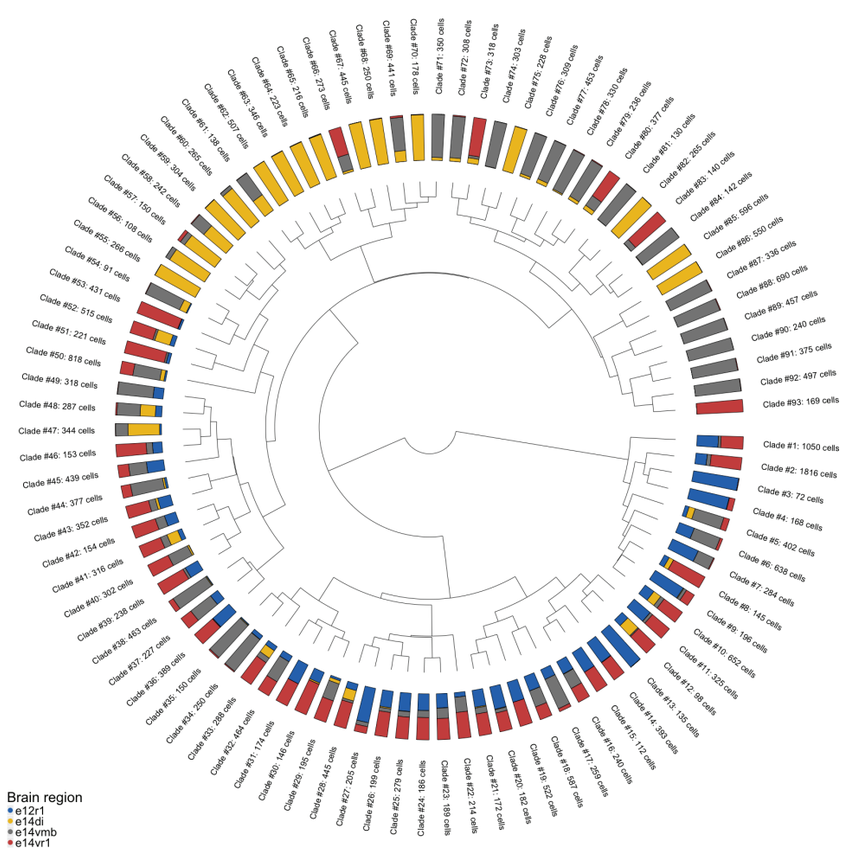
C



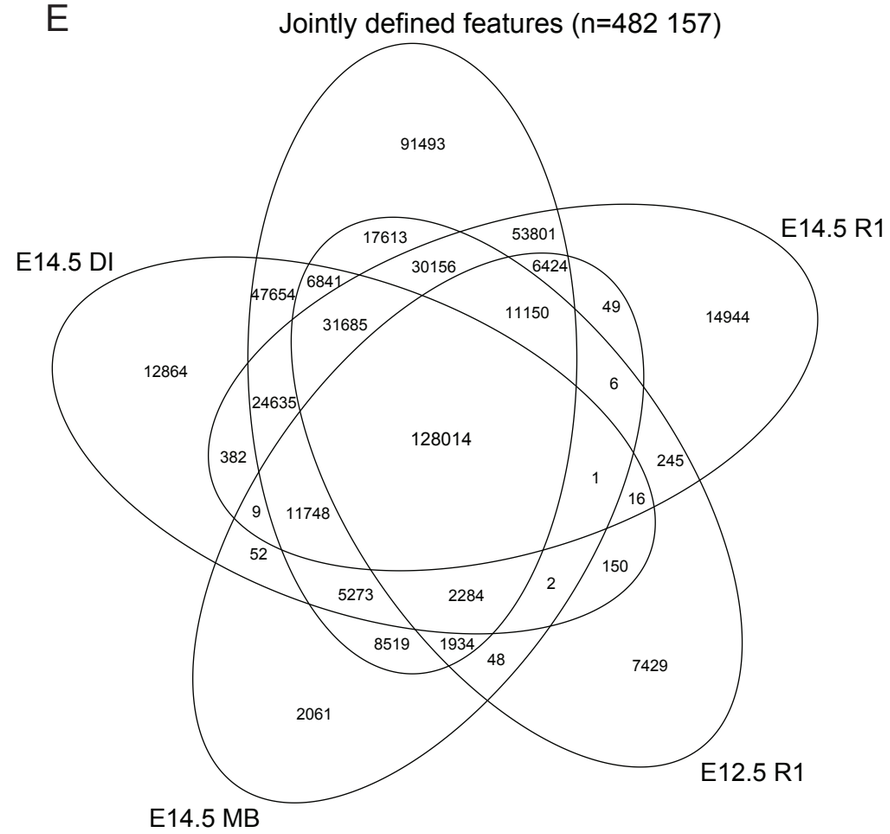
D



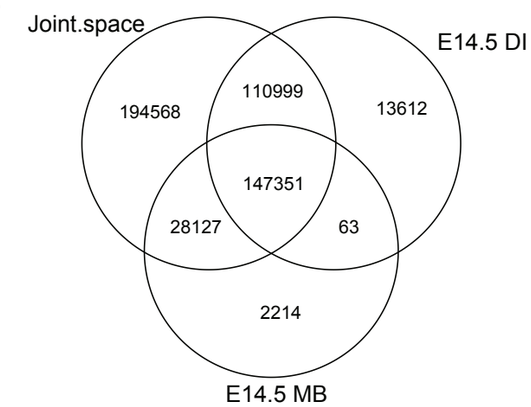
B



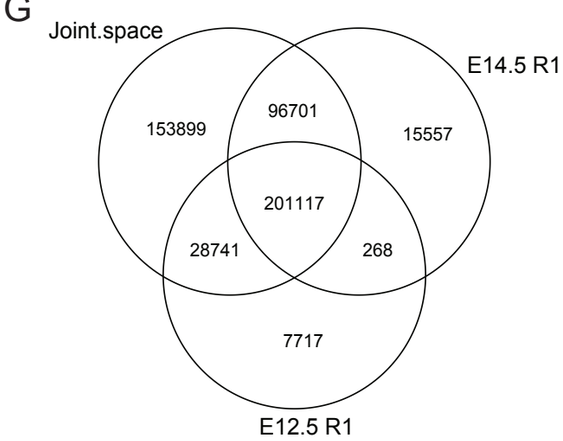
E



F

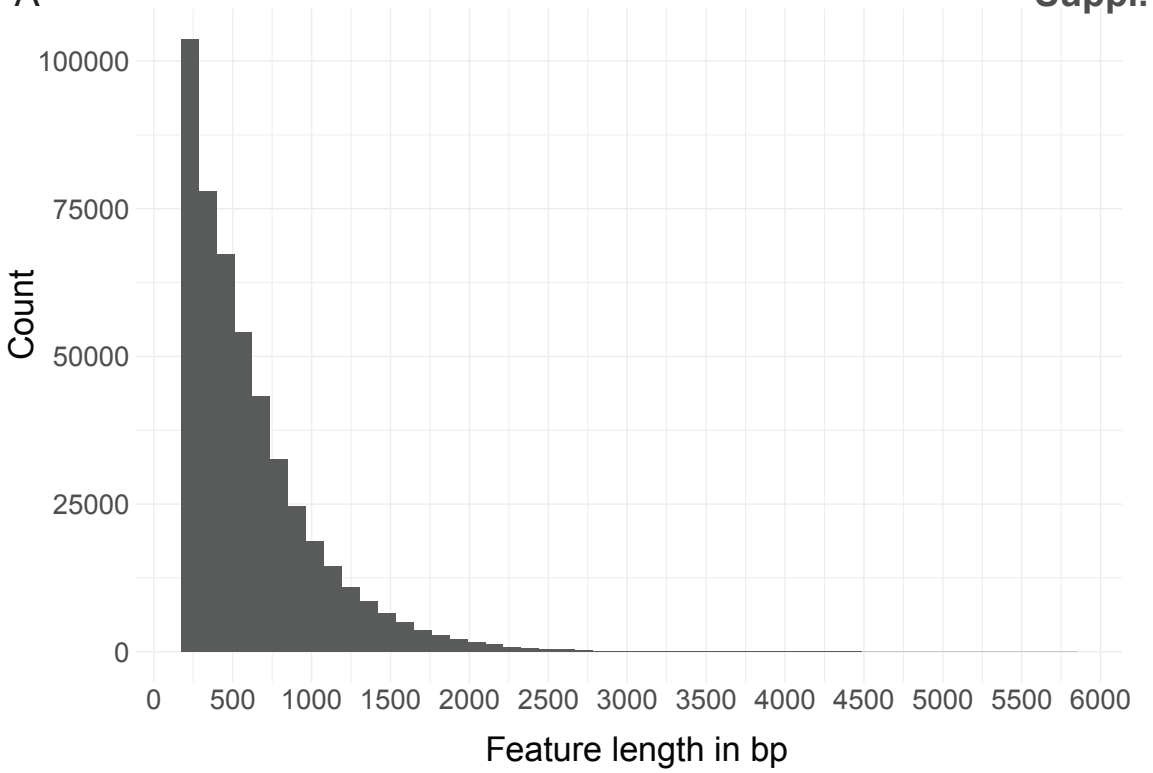


G

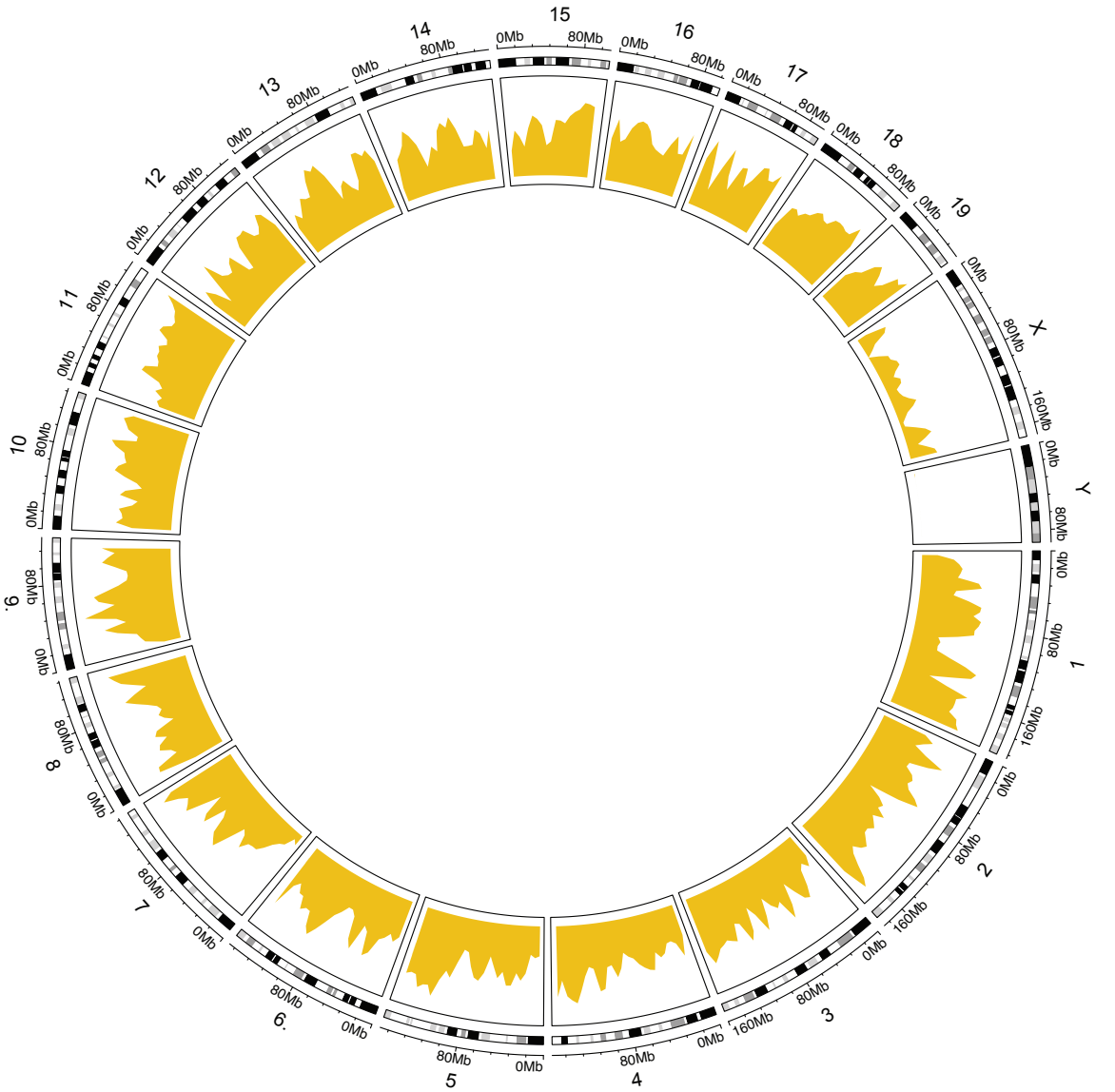


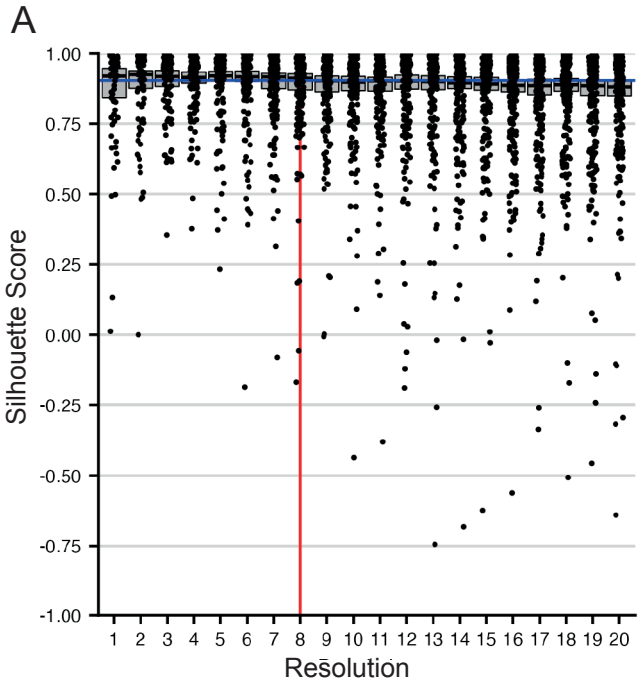
A

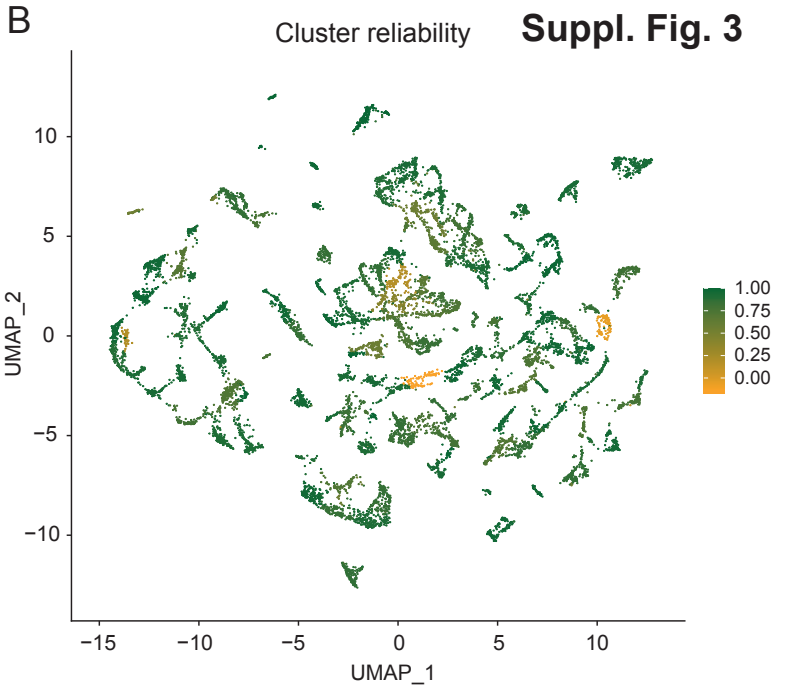
Suppl. Fig. 2



B



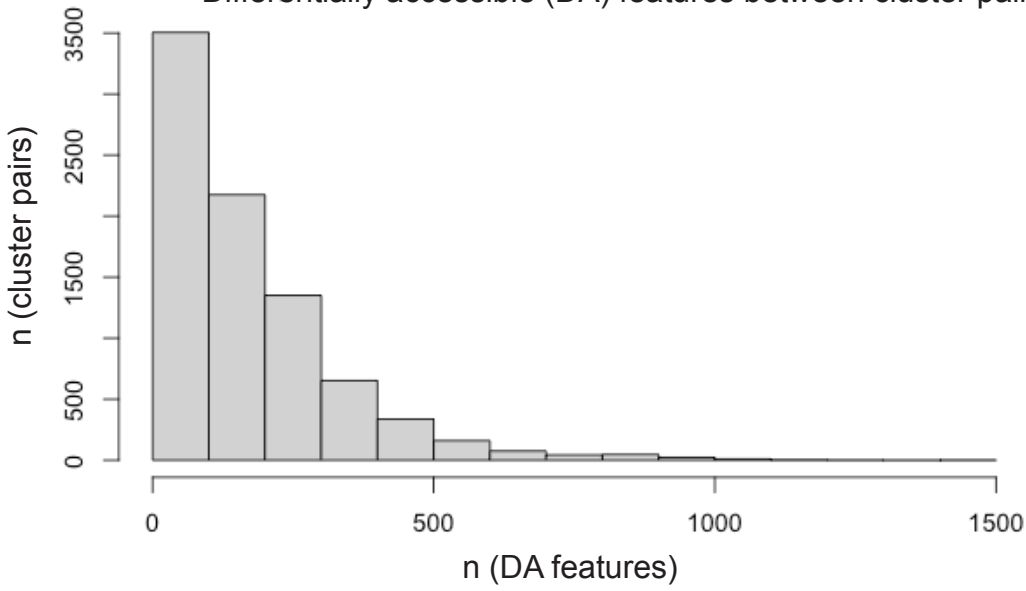

A
Silhouette Score
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Resolution


B
Cluster reliability
Suppl. Fig. 3
UMAP_2
10
5
0
−5
−10
−15 −10 −5 0 5 10
UMAP_1
1.00
0.75
0.50
0.25
0.00

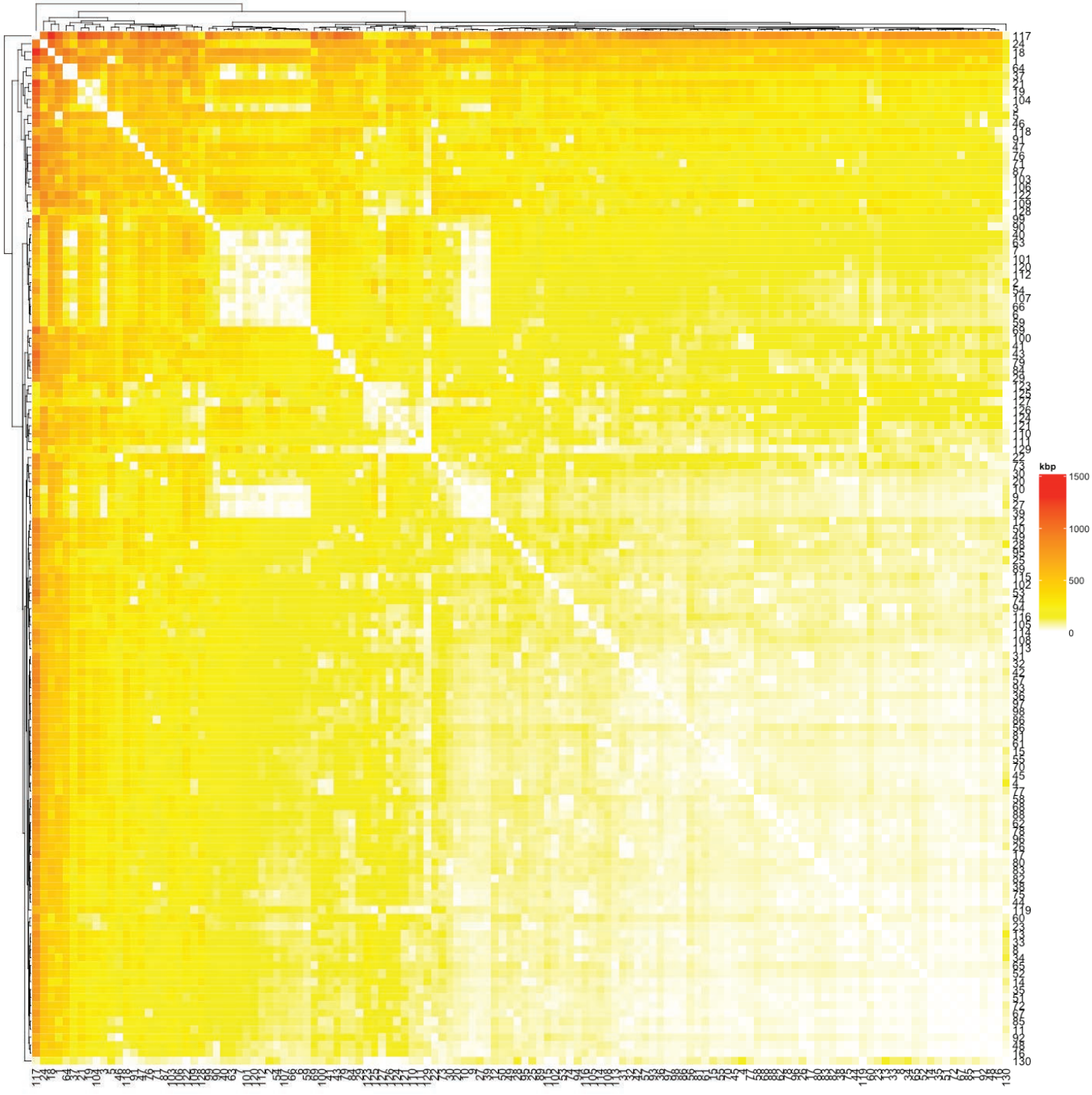
A

Differentially accessible (DA) features between cluster pairs

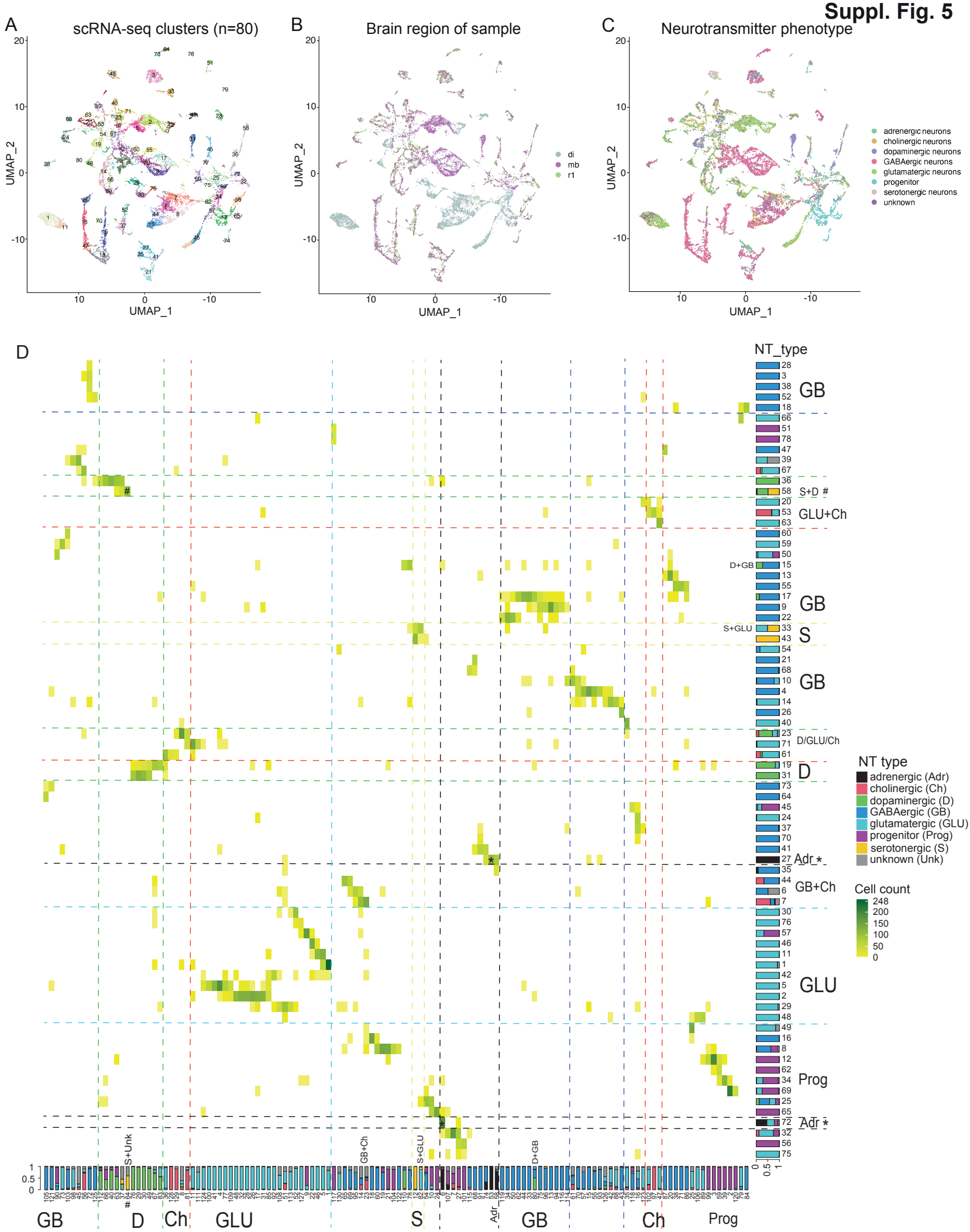
Suppl. Fig. 4



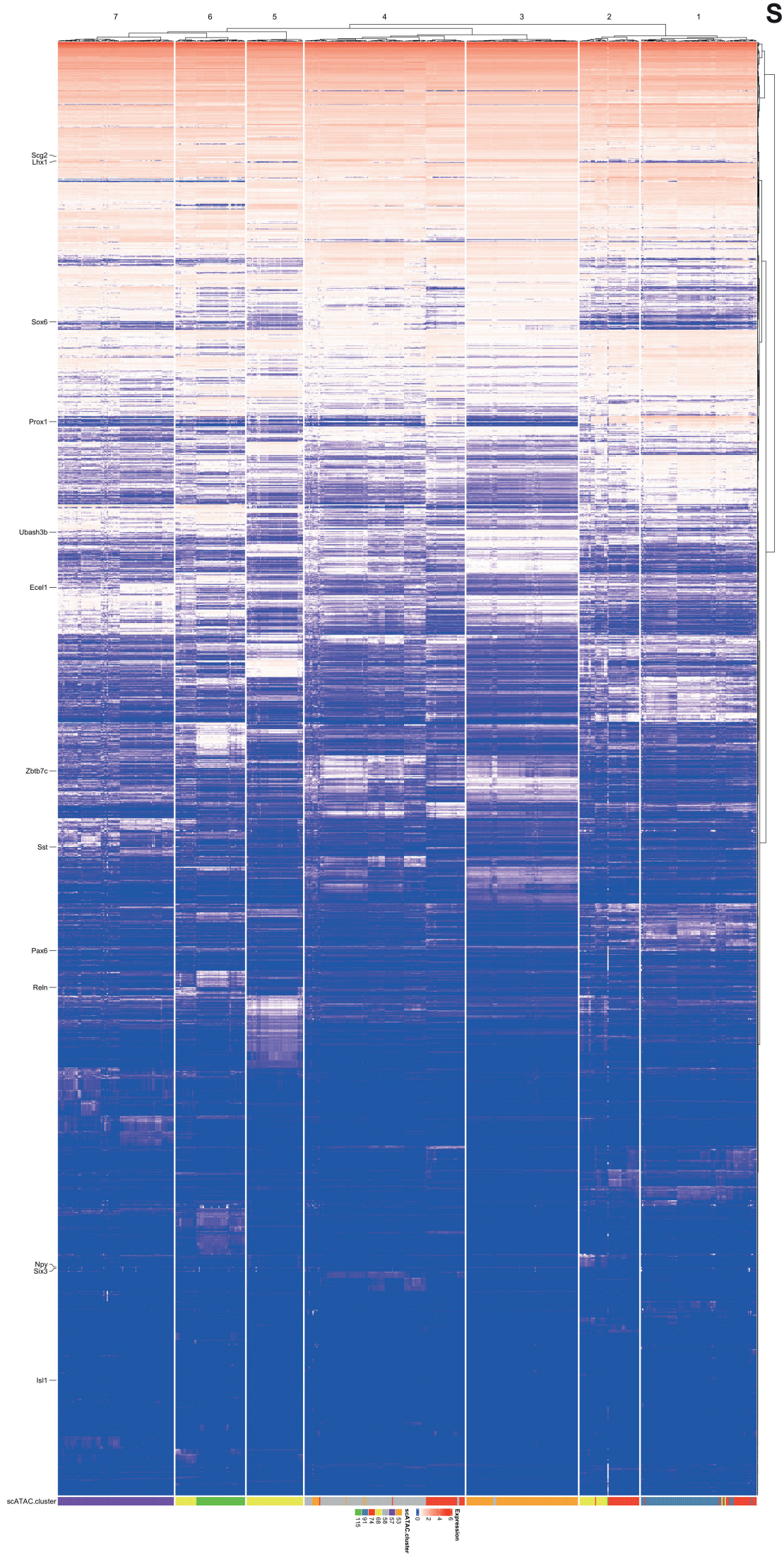
B



Suppl. Fig. 5


A
scRNA-seq clusters (n=80)
B
Brain region of sample
di
mb
r1
C
Neurotransmitter phenotype
adrenergic neurons
cholinergic neurons
dopaminergic neurons
GABAergic neurons
glutamatergic neurons
progenitor
serotonergic neurons
unknown
UMAP_1
UMAP_2
D
NT_type
GB
S+D #
GLU+Ch
D+GB
S+GLU
S
D/GLU/Ch
D
Adr *
GB+Ch
GLU
Prog
NT type
adrenergic (Adr)
cholinergic (Ch)
dopaminergic (D)
GABAergic (GB)
glutamatergic (GLU)
progenitor (Prog)
serotonergic (S)
unknown (Unk)
Cell count
248
200
150
100
50
0
S+Unk
GB+Ch
S+GLU
D+GB
GB
D
Ch
GLU
S
Adr
GB
Ch
Prog

Suppl. Fig. 6


7
6
5
4
3
2
1
Scg2
Lhx1
Sox6
Prox1
Ubash3b
Ecel1
Zbtb7c
Sst
Pax6
Reln
Npy
Six3
Isl1
scATAC.cluster
Expression
6
4
2
0
scATAC.cluster
53
57
58
68
74
97
115