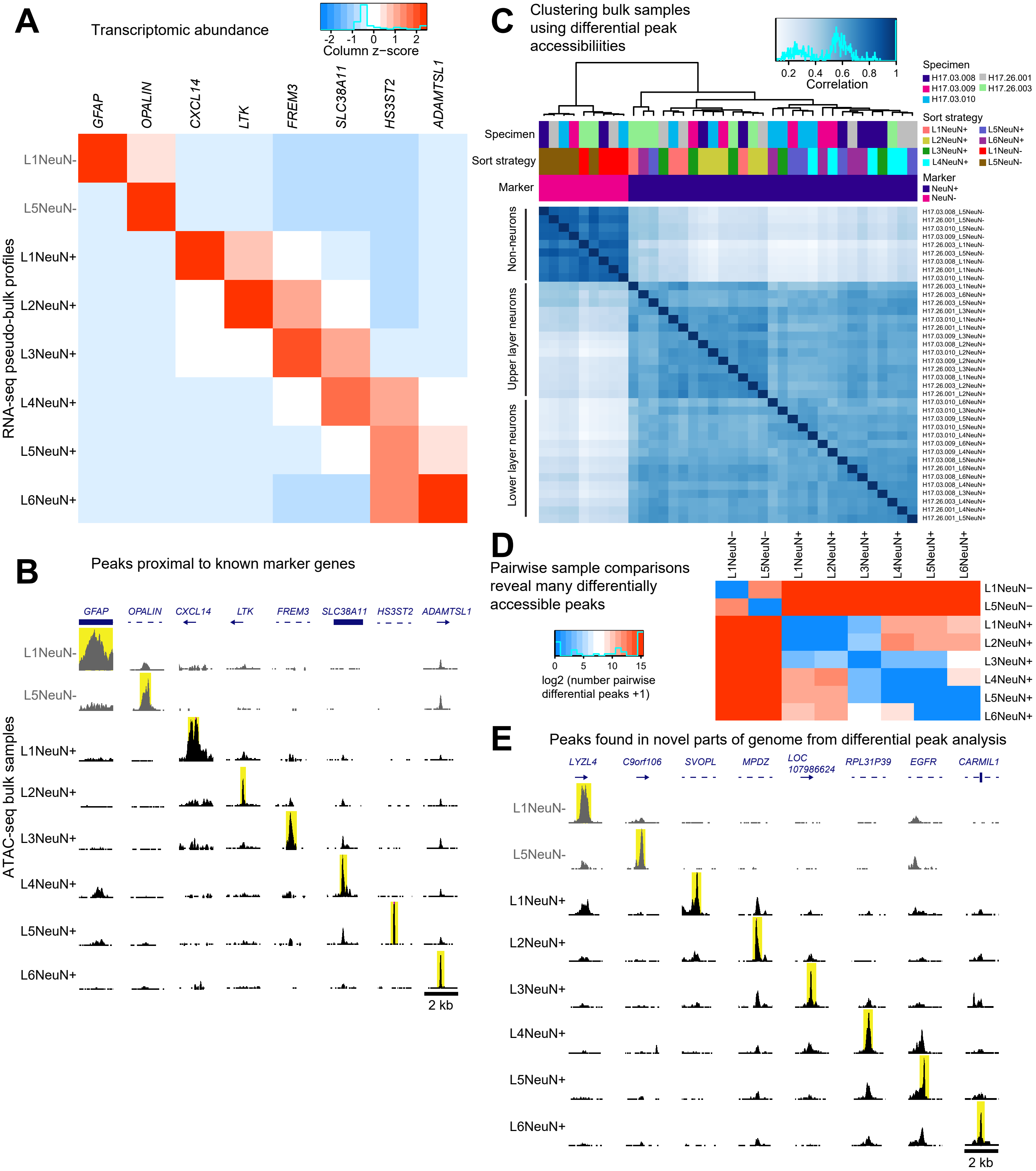
Cell Reports, Volume *34*

Supplemental information

Functional enhancer elements drive subclass-selective expression from mouse to primate neocortex

John K. Mich, Lucas T. Graybuck, Erik E. Hess, Joseph T. Mahoney, Yoshiko Kojima, Yi Ding, Saroja Somasundaram, Jeremy A. Miller, Brian E. Kalmbach, Cristina Radaelli, Bryan B. Gore, Natalie Weed, Victoria Omstead, Yemeserach Bishaw, Nadiya V. Shapovalova, Refugio A. Martinez, Olivia Fong, Shenqin Yao, Marty Mortrud, Peter Chong, Luke Loftus, Darren Bertagnolli, Jeff Goldy, Tamara Casper, Nick Dee, Ximena Opitz-Araya, Ali Cetin, Kimberly A. Smith, Ryder P. Gwinn, Charles Cobbs, Andrew L. Ko, Jeffrey G. Ojemann, C. Dirk Keene, Daniel L. Silbergeld, Susan M. Sunkin, Viviana Gradinaru, Gregory D. Horwitz, Hongkui Zeng, Bosiljka Tasic, Ed S. Lein, Jonathan T. Ting, and Boaz P. Levi


A
Transcriptomic abundance
-2 -1 0 1 2
Column z−score
GFAP
OPALIN
CXCL14
LTK
FREM3
SLC38A11
HS3ST2
ADAMTSL1
RNA-seq pseudo-bulk profiles
L1NeuN-
L5NeuN-
L1NeuN+
L2NeuN+
L3NeuN+
L4NeuN+
L5NeuN+
L6NeuN+
B
Peaks proximal to known marker genes
GFAP
OPALIN
CXCL14
LTK
FREM3
SLC38A11
HS3ST2
ADAMTSL1
ATAC-seq bulk samples
L1NeuN-
L5NeuN-
L1NeuN+
L2NeuN+
L3NeuN+
L4NeuN+
L5NeuN+
L6NeuN+
2 kb
C
Clustering bulk samples using differential peak accessibiliities
0.2 0.4 0.6 0.8 1
Correlation
Specimen
Sort strategy
Marker
Specimen
H17.03.008
H17.03.009
H17.03.010
H17.26.001
H17.26.003
Sort strategy
L1NeuN+
L2NeuN+
L3NeuN+
L4NeuN+
L5NeuN+
L6NeuN+
L1NeuN-
L5NeuN-
Marker
NeuN+
NeuN-
Non-neurons
Upper layer neurons
Lower layer neurons
H17.03.008_L5NeuN-
H17.26.001_L5NeuN-
H17.03.010_L5NeuN-
H17.03.009_L5NeuN-
H17.26.003_L1NeuN-
H17.26.003_L5NeuN-
H17.03.008_L1NeuN-
H17.26.001_L1NeuN-
H17.03.010_L1NeuN-
H17.26.003_L1NeuN+
H17.26.003_L6NeuN+
H17.26.003_L5NeuN+
H17.26.001_L3NeuN+
H17.03.010_L1NeuN+
H17.26.001_L1NeuN+
H17.03.009_L3NeuN+
H17.03.008_L2NeuN+
H17.03.010_L2NeuN+
H17.03.009_L2NeuN+
H17.26.003_L3NeuN+
H17.03.008_L1NeuN+
H17.26.003_L2NeuN+
H17.26.001_L2NeuN+
H17.03.010_L6NeuN+
H17.03.010_L3NeuN+
H17.03.009_L5NeuN+
H17.03.010_L5NeuN+
H17.03.010_L4NeuN+
H17.03.009_L6NeuN+
H17.03.009_L4NeuN+
H17.03.008_L5NeuN+
H17.26.001_L6NeuN+
H17.03.008_L6NeuN+
H17.03.008_L4NeuN+
H17.03.008_L3NeuN+
H17.26.003_L4NeuN+
H17.26.001_L4NeuN+
H17.26.001_L5NeuN+
D
Pairwise sample comparisons reveal many differentially accessible peaks
0 5 10 15
log2 (number pairwise differential peaks +1)
L1NeuN−
L5NeuN−
L1NeuN+
L2NeuN+
L3NeuN+
L4NeuN+
L5NeuN+
L6NeuN+
L1NeuN−
L5NeuN−
L1NeuN+
L2NeuN+
L3NeuN+
L4NeuN+
L5NeuN+
L6NeuN+
E
Peaks found in novel parts of genome from differential peak analysis
LYZL4
C9orf106
SVOPL
MPDZ
LOC 107986624
RPL31P39
EGFR
CARMIL1
L1NeuN-
L5NeuN-
L1NeuN+
L2NeuN+
L3NeuN+
L4NeuN+
L5NeuN+
L6NeuN+
2 kb

Supplementary Figure 1 (Related to Figure 1): Bulk ATAC-seq data demonstrates differentially accessible chromatin elements around known marker genes, and in novel genomic regions.

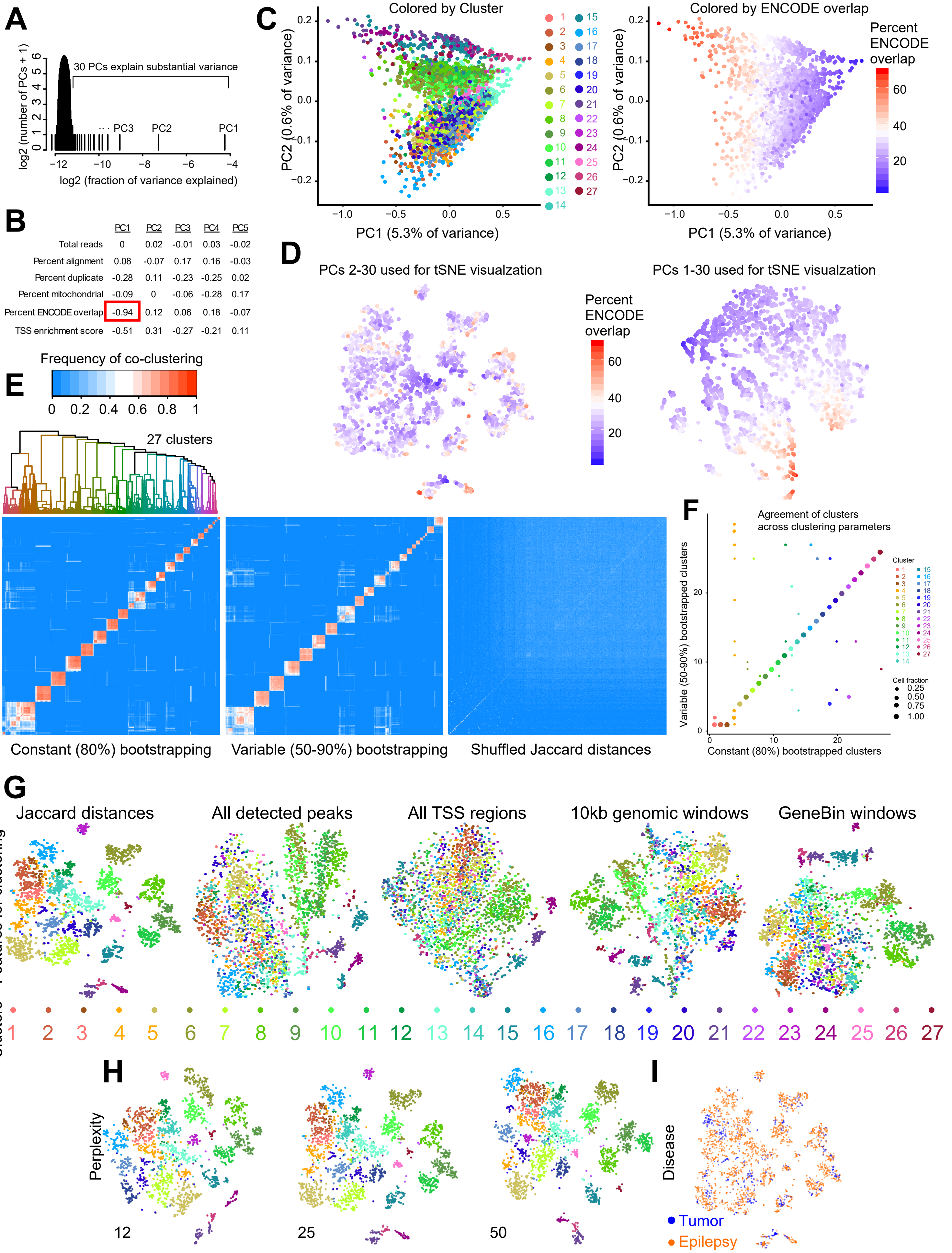
A) snRNA-seq data (Hodge et al., 2019), aggregated into pseudo-bulk profiles by weighted averages of gene CPM medians for 75 transcriptomic clusters. Weights were assigned by their frequencies within the eight sort strategies, and the heatmap is scaled by z-score within each column (gene). Relative expressions of eight sort strategy-specific marker genes are displayed.

B) Example sort strategy-specific peaks proximal to (<50kb distance to gene body) the eight sort strategy-specific transcriptomic marker genes. Pileups indicate aggregated data within a 2 kb genomic window across five independent experiments. In B and E, dashed lines indicate introns, thick lines indicate exons, and arrows indicate direction towards proximal marker gene. Yellow highlights demarcate sort strategy-specific chromatin accessibility peaks.

C) DiffBind (Ross-Innes et al., 2012) identification of 72,218 peaks that were differentially accessible among any pairwise comparison of sort strategies (FDR 0.01). Read counts within those 72,218 differentially accessible peaks then clustered samples using a correlation distance matrix, which revealed separate groupings of non-neuronal samples, and upper- and lower-layer neuronal samples. One sample was omitted from this analysis (H17.03.009 L1 NeuN+) because this sample appeared intermediate between NeuN+ and NeuN- cells, suggesting a failed sort.

D) Number of peaks differentiating each pairwise sample contrast.

E) Example sort strategy-specific peaks resulting from pairwise DiffBind differential peak analysis. These peaks were found in novel genomic regions (not proximal to known marker genes), and closest genes are shown.



	PC1	PC2	PC3	PC4	PC5
Total reads	0	0.02	-0.01	0.03	-0.02
Percent alignment	0.08	-0.07	0.17	0.16	-0.03
Percent duplicate	-0.28	0.11	-0.23	-0.25	0.02
Percent mitochondrial	-0.09	0	-0.06	-0.28	0.17
Percent ENCODE overlap	-0.94	0.12	0.06	0.18	-0.07
TSS enrichment score	-0.51	0.31	-0.27	-0.21	0.11

Supplementary Figure 2 (Related to Figure 1): High confidence clustering for single nucleus ATAC-seq data.

A) Histogram showing the percentages of variance explained by each principal component of the Jaccard single cell distance matrix. The first 30 principal components explain substantial variance within the dataset.

B) Correlation of the first five principal components with quality metrics. Principal component 1 was omitted from further analysis due to strong negative correlation with ENCODE overlap.

C) Single nuclei evaluated by principal component analysis, with nuclei colored by cluster membership (*left*). Three major groups of nuclei were separated by PC2. Single nuclei were also colored by ENCODE overlap percentage, which is strongly negatively correlated with PC1 (*right*).

D) tSNE plot to visualize either principal components 2 to 30 (*left*) or 1 to 30 (*right*). Note, PCs 2 to 30 permit clear groupings with no ENCODE overlap gradient, whereas PCs 1 to 30 result in blurred cluster separations with a gradient of ENCODE spanning the clusters.

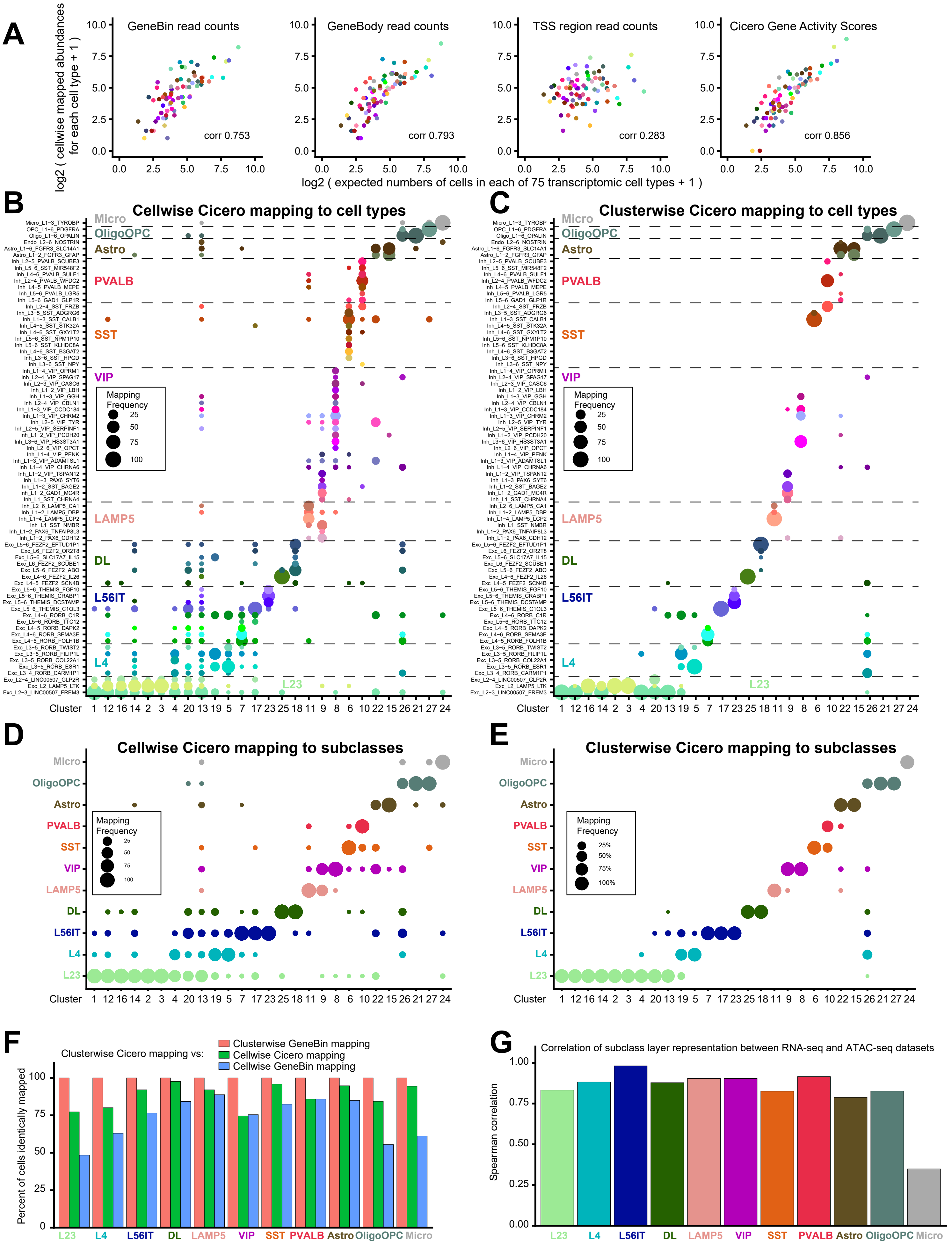
E) Bootstrapped iterative clustering to identify reproducible nuclear clusters. From the 2858 x 29 matrix of nuclei x principal component scores, we subsampled to either a constant 80% of nuclei (*left*) or a variable 50-90% of nuclei (*middle*), and calculated clusters using Jaccard-Louvain clustering (Tasic et al., 2018), which was repeated 200 times. Shuffled Jaccard distance matrix as input to PCA is shown (*right)*. Heatmaps display the frequency of co-clustering among nuclei. The constant 80% bootstrapping co-clustering matrix was used as input into Euclidean distance clustering, which yielded the final 27 clusters by cutting the tree to the major blocks of co-clustering nuclei. Nucleus order is not matched across the three plots.

F) Agreement between cluster memberships resulting from constant 80% bootstrapping and variable 50-90% bootstrapping, for most nuclei.

G) Visualization of nucleus groupings using five different feature sets (see Methods) using tSNE. Jaccard distances yielded clearest cell groupings. Cluster colors are applied in both (G) and (H).

H) Different perplexity parameters for tSNE visualization of cell groupings. Nuclear cluster groupings are evident at a wide range of perplexity values.

I) Visualization of disease status (tumor or epilepsy) for nuclei. Note that nuclei largely intermix regardless of disease status.

A
log2 (cellwise mapped abundances for each cell type + 1)
GeneBin read counts
corr 0.753
GeneBody read counts
corr 0.793
TSS region read counts
corr 0.283
Cicero Gene Activity Scores
corr 0.856
0.0
2.5
5.0
7.5
10.0
log2 (expected numbers of cells in each of 75 transcriptomic cell types + 1)
B
Cellwise Cicero mapping to cell types
Micro
OligoOPC
Astro
PVALB
SST
VIP
LAMP5
DL
L56IT
L4
L23
Micro_L1−3_TYROBP
OPC_L1−6_PDGFRA
Oligo_L1−6_OPALIN
Endo_L2−6_NOSTRIN
Astro_L1−6_FGFR3_SLC14A1
Astro_L1−2_FGFR3_GFAP
Inh_L2−5_PVALB_SCUBE3
Inh_L5−6_SST_MIR548F2
Inh_L4−6_PVALB_SULF1
Inh_L2−4_PVALB_WFDC2
Inh_L4−5_PVALB_MEPE
Inh_L5−6_PVALB_LGR5
Inh_L5−6_GAD1_GLP1R
Inh_L2−4_SST_FRZB
Inh_L3−5_SST_ADGRG6
Inh_L1−3_SST_CALB1
Inh_L4−5_SST_STK32A
Inh_L4−6_SST_GXYLT2
Inh_L5−6_SST_NPM1P10
Inh_L5−6_SST_KLHDC8A
Inh_L4−6_SST_B3GAT2
Inh_L3−6_SST_HPGD
Inh_L3−6_SST_NPY
Inh_L1−4_VIP_OPRM1
Inh_L2−4_VIP_SPAG17
Inh_L2−3_VIP_CASC6
Inh_L1−2_VIP_LBH
Inh_L1−3_VIP_GGH
Inh_L2−4_VIP_CBLN1
Inh_L1−3_VIP_CCDC184
Inh_L1−3_VIP_CHRM2
Inh_L2−5_VIP_TYR
Inh_L2−5_VIP_SERPINF1
Inh_L1−2_VIP_PCDH20
Inh_L3−6_VIP_HS3ST3A1
Inh_L2−6_VIP_QPCT
Inh_L1−4_VIP_PENK
Inh_L1−3_VIP_ADAMTSL1
Inh_L1−4_VIP_CHRNA6
Inh_L1−2_VIP_TSPAN12
Inh_L1−3_PAX6_SYT6
Inh_L1−2_SST_BAGE2
Inh_L1−2_GAD1_MC4R
Inh_L1_SST_CHRNA4
Inh_L2−6_LAMP5_CA1
Inh_L1−2_LAMP5_DBP
Inh_L1−4_LAMP5_LCP2
Inh_L1_SST_NMBR
Inh_L1−2_PAX6_TNFAIP8L3
Inh_L1−2_PAX6_CDH12
Exc_L5−6_FEZF2_EFTUD1P1
Exc_L6_FEZF2_OR2T8
Exc_L5−6_SLC17A7_IL15
Exc_L6_FEZF2_SCUBE1
Exc_L5−6_FEZF2_ABO
Exc_L4−6_FEZF2_IL26
Exc_L4−5_FEZF2_SCN4B
Exc_L5−6_THEMIS_FGF10
Exc_L5−6_THEMIS_CRABP1
Exc_L5−6_THEMIS_DCSTAMP
Exc_L5−6_THEMIS_C1QL3
Exc_L4−6_RORB_C1R
Exc_L5−6_RORB_TTC12
Exc_L4−5_RORB_DAPK2
Exc_L4−6_RORB_SEMA3E
Exc_L4−5_RORB_FOLH1B
Exc_L3−5_RORB_TWIST2
Exc_L3−5_RORB_FILIP1L
Exc_L3−5_RORB_COL22A1
Exc_L3−5_RORB_ESR1
Exc_L3−4_RORB_CARM1P1
Exc_L2−4_LINC00507_GLP2R
Exc_L2_LAMP5_LTK
Exc_L2−3_LINC00507_FREM3
Mapping Frequency
25
50
75
100
Cluster
1 12 16 14 2 3 4 20 13 19 5 7 17 23 25 18 11 9 8 6 10 22 15 26 21 27 24
C
Clusterwise Cicero mapping to cell types
Micro
OligoOPC
Astro
PVALB
SST
VIP
LAMP5
DL
L56IT
L4
L23
Micro_L1−3_TYROBP
OPC_L1−6_PDGFRA
Oligo_L1−6_OPALIN
Endo_L2−6_NOSTRIN
Astro_L1−6_FGFR3_SLC14A1
Astro_L1−2_FGFR3_GFAP
Inh_L2−5_PVALB_SCUBE3
Inh_L5−6_SST_MIR548F2
Inh_L4−6_PVALB_SULF1
Inh_L2−4_PVALB_WFDC2
Inh_L4−5_PVALB_MEPE
Inh_L5−6_PVALB_LGR5
Inh_L5−6_GAD1_GLP1R
Inh_L2−4_SST_FRZB
Inh_L3−5_SST_ADGRG6
Inh_L1−3_SST_CALB1
Inh_L4−5_SST_STK32A
Inh_L4−6_SST_GXYLT2
Inh_L5−6_SST_NPM1P10
Inh_L5−6_SST_KLHDC8A
Inh_L4−6_SST_B3GAT2
Inh_L3−6_SST_HPGD
Inh_L3−6_SST_NPY
Inh_L1−4_VIP_OPRM1
Inh_L2−4_VIP_SPAG17
Inh_L2−3_VIP_CASC6
Inh_L1−2_VIP_LBH
Inh_L1−3_VIP_GGH
Inh_L2−4_VIP_CBLN1
Inh_L1−3_VIP_CCDC184
Inh_L1−3_VIP_CHRM2
Inh_L2−5_VIP_TYR
Inh_L2−5_VIP_SERPINF1
Inh_L1−2_VIP_PCDH20
Inh_L3−6_VIP_HS3ST3A1
Inh_L2−6_VIP_QPCT
Inh_L1−4_VIP_PENK
Inh_L1−3_VIP_ADAMTSL1
Inh_L1−4_VIP_CHRNA6
Inh_L1−2_VIP_TSPAN12
Inh_L1−3_PAX6_SYT6
Inh_L1−2_SST_BAGE2
Inh_L1−2_GAD1_MC4R
Inh_L1_SST_CHRNA4
Inh_L2−6_LAMP5_CA1
Inh_L1−2_LAMP5_DBP
Inh_L1−4_LAMP5_LCP2
Inh_L1_SST_NMBR
Inh_L1−2_PAX6_TNFAIP8L3
Inh_L1−2_PAX6_CDH12
Exc_L5−6_FEZF2_EFTUD1P1
Exc_L6_FEZF2_OR2T8
Exc_L5−6_SLC17A7_IL15
Exc_L6_FEZF2_SCUBE1
Exc_L5−6_FEZF2_ABO
Exc_L4−6_FEZF2_IL26
Exc_L4−5_FEZF2_SCN4B
Exc_L5−6_THEMIS_FGF10
Exc_L5−6_THEMIS_CRABP1
Exc_L5−6_THEMIS_DCSTAMP
Exc_L5−6_THEMIS_C1QL3
Exc_L4−6_RORB_C1R
Exc_L5−6_RORB_TTC12
Exc_L4−5_RORB_DAPK2
Exc_L4−6_RORB_SEMA3E
Exc_L4−5_RORB_FOLH1B
Exc_L3−5_RORB_TWIST2
Exc_L3−5_RORB_FILIP1L
Exc_L3−5_RORB_COL22A1
Exc_L3−5_RORB_ESR1
Exc_L3−4_RORB_CARM1P1
Exc_L2−4_LINC00507_GLP2R
Exc_L2_LAMP5_LTK
Exc_L2−3_LINC00507_FREM3
Mapping Frequency
25
50
75
100
Cluster
1 12 16 14 2 3 4 20 13 19 5 7 17 23 25 18 11 9 8 6 10 22 15 26 21 27 24
D
Cellwise Cicero mapping to subclasses
Micro
OligoOPC
Astro
PVALB
SST
VIP
LAMP5
DL
L56IT
L4
L23
Mapping Frequency
25
50
75
100
Cluster
1 12 16 14 2 3 4 20 13 19 5 7 17 23 25 18 11 9 8 6 10 22 15 26 21 27 24
E
Clusterwise Cicero mapping to subclasses
Micro
OligoOPC
Astro
PVALB
SST
VIP
LAMP5
DL
L56IT
L4
L23
Mapping Frequency
25%
50%
75%
100%
Cluster
1 12 16 14 2 3 4 20 13 19 5 7 17 23 25 18 11 9 8 6 10 22 15 26 21 27 24
F
Clusterwise Cicero mapping vs:
Clusterwise GeneBin mapping
Cellwise Cicero mapping
Cellwise GeneBin mapping
Percent of cells identically mapped
0
25
50
75
100
L23
L4
L56IT
DL
LAMP5
VIP
SST
PVALB
Astro
OligoOPC
Micro
G
Correlation of subclass layer representation between RNA-seq and ATAC-seq datasets
Spearman correlation
0.00
0.25
0.50
0.75
1.00
L23
L4
L56IT
DL
LAMP5
VIP
SST
PVALB
Astro
OligoOPC
Micro

Supplementary Figure 3 (Related to Figure 1): Mapping ATAC-seq clusters to RNA-seq cell types and subclasses.

A) Expected abundances of each of the 75 transcriptomic cell types (Hodge et al., 2019), correlated with observed abundances of those cell types, using four different methods for computing gene-level information from each nucleus: *far left*: read counts in gene bins, *middle left*: read counts in gene bodies, *middle right:* read counts in 10kb-extended TSS regions, and *far right:* Cicero gene activity scores. Correlation values are Pearson correlation statistics between log-transformed expected and observed abundances plus one, for each of the 75 transcriptomic cell types. Computing gene-level information using cicero gene activity scores results in the greatest correlation between expectation and observation for cell type abundances.

B) Bootstrapped mapping of single nuclei ("cellwise") to 75 transcriptomic cell types. Dot sizes indicate the frequencies of cell type mappings within each of the 27 ATAC-seq clusters.

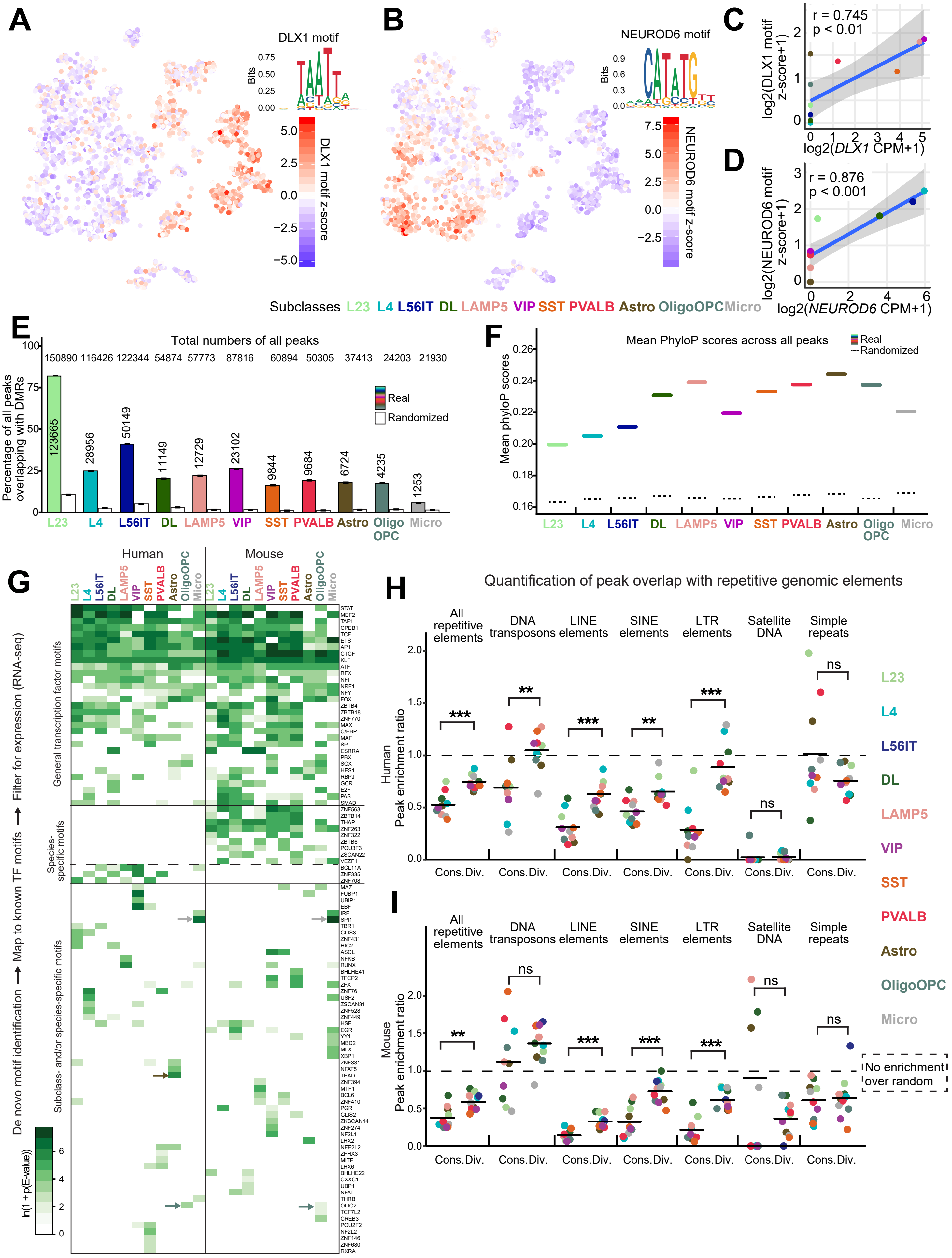
C) Bootstrapped mapping of clusters ("clusterwise") to 75 transcriptomic cell types. Dot sizes indicate the frequency each cluster maps to each transcriptomic cell type.

D) Bootstrapped mapping of single nuclei ("cellwise") to 11 transcriptomic cell type subclasses. Dot sizes indicate the frequencies of subclass mappings within each of the 27 ATAC-seq clusters.

E) Bootstrapped mapping of clusters ("clusterwise") to 11 transcriptomic cell type subclasses. Dot sizes indicate the frequency each cluster maps to each subclass. This plot represents the final mapped subclass assigned as the most frequent mapping for each cluster, which are used throughout the text.

F) Correlation of subclass mappings for all cells using four different mapping techniques. Overall, most cells are identically mapped to the same subclass with most of the techniques, with especially good agreement between both clusterwise mapping techniques.

G) Correlation between RNA-seq and ATAC-seq dataset layerwise distributions for the 11 subclasses. Most of the subclasses are observed in similar layer distributions in both datasets.


A
DLX1 motif
Bits
0.75
0.50
0.25
0.00
5.0
2.5
0.0
−2.5
−5.0
DLX1 motif z-score
B
NEUROD6 motif
Bits
0.9
0.6
0.3
0.0
7.5
5.0
2.5
0.0
−2.5
NEUROD6 motif z-score
C
r = 0.745
p < 0.01
log2(DLX1 motif z-score+1)
log2(DLX1 CPM+1)
0 1 2 3 4 5
D
r = 0.876
p < 0.001
log2(NEUROD6 motif z-score+1)
log2(NEUROD6 CPM+1)
0 2 4 6
Subclasses L23 L4 L56IT DL LAMP5 VIP SST PVALB Astro OligoOPC Micro
E
Total numbers of all peaks
150890 116426 122344 54874 57773 87816 60894 50305 37413 24203 21930
Percentage of all peaks overlapping with DMRs
123665
28956
50149
11149
12729
23102
9844
9684
6724
4235
1253
Real
Randomized
L23 L4 L56IT DL LAMP5 VIP SST PVALB Astro Oligo OPC Micro
F
Mean PhyloP scores across all peaks
Real
--- Randomized
Mean phyloP scores
0.26
0.24
0.22
0.20
0.18
0.16
L23 L4 L56IT DL LAMP5 VIP SST PVALB Astro Oligo OPC Micro
G
Human
Mouse
L23 L4 L56IT DL LAMP5 VIP SST PVALB Astro OligoOPC Micro
De novo motif identification → Map to known TF motifs → Filter for expression (RNA-seq)
General transcription factor motifs
Species-specific motifs
Subclass- and/or species-specific motifs
ln(1 + p(E-value))
0 2 4 6
STAT
MEF2
TAF1
CPEB1
TCF
ETS
AP1
CTCF
KLF
ATF
RFX
NFI
NRF1
NFY
FOX
ZBTB4
ZBTB18
ZNF770
MAX
C/EBP
MAF
SP
ESRRA
PBX
SOX
HES1
RBPJ
GCR
E2F
PAS
SMAD
ZNF563
ZBTB14
THAP
ZNF263
ZNF322
ZBTB6
POU3F3
ZSCAN22
VEZF1
BCL11A
ZNF335
ZNF708
MAZ
FUBP1
UBIP1
EBF
IRF
SPI1
TBR1
GLIS3
ZNF431
HIC2
ASCL
NFKB
RUNX
BHLHE41
TFCP2
ZFX
ZNF76
USF2
ZSCAN31
ZNF528
ZNF449
HSF
EGR
YY1
MBD2
MLX
XBP1
ZNF331
NFAT5
TEAD
ZNF394
MTF1
BCL6
ZNF410
PGR
GLIS2
ZKSCAN14
ZNF274
NF2L1
LHX2
NFE2L2
ZFHX3
MITF
LHX6
BHLHE22
CXXC1
UBP1
NFAT
THRB
OLIG2
TCF7L2
CREB3
POU2F2
NF2L2
ZNF146
ZNF680
RXRA
H
Quantification of peak overlap with repetitive genomic elements
All repetitive elements
DNA transposons
LINE elements
SINE elements
LTR elements
Satellite DNA
Simple repeats
Human Peak enrichment ratio
2.0
1.5
1.0
0.5
0.0
***
**
***
**
***
ns
ns
Cons. Div.
I
Mouse Peak enrichment ratio
**
ns
***
***
***
ns
ns
No enrichment over random
L23
L4
L56IT
DL
LAMP5
VIP
SST
PVALB
Astro
OligoOPC
Micro

Supplementary Figure 4 (Related to Figure 2): Properties of human neocortical cell subclass-specific accessible genomic elements.

A-B) Nuclei visualized by tSNE and colored by motif accessibilities for A) DLX1 and B) NEUROD6 as calculated by chromVAR (Schep et al., 2017). *DLX1* transcripts are specifically detected in inhibitory neurons (Hodge et al., 2019).

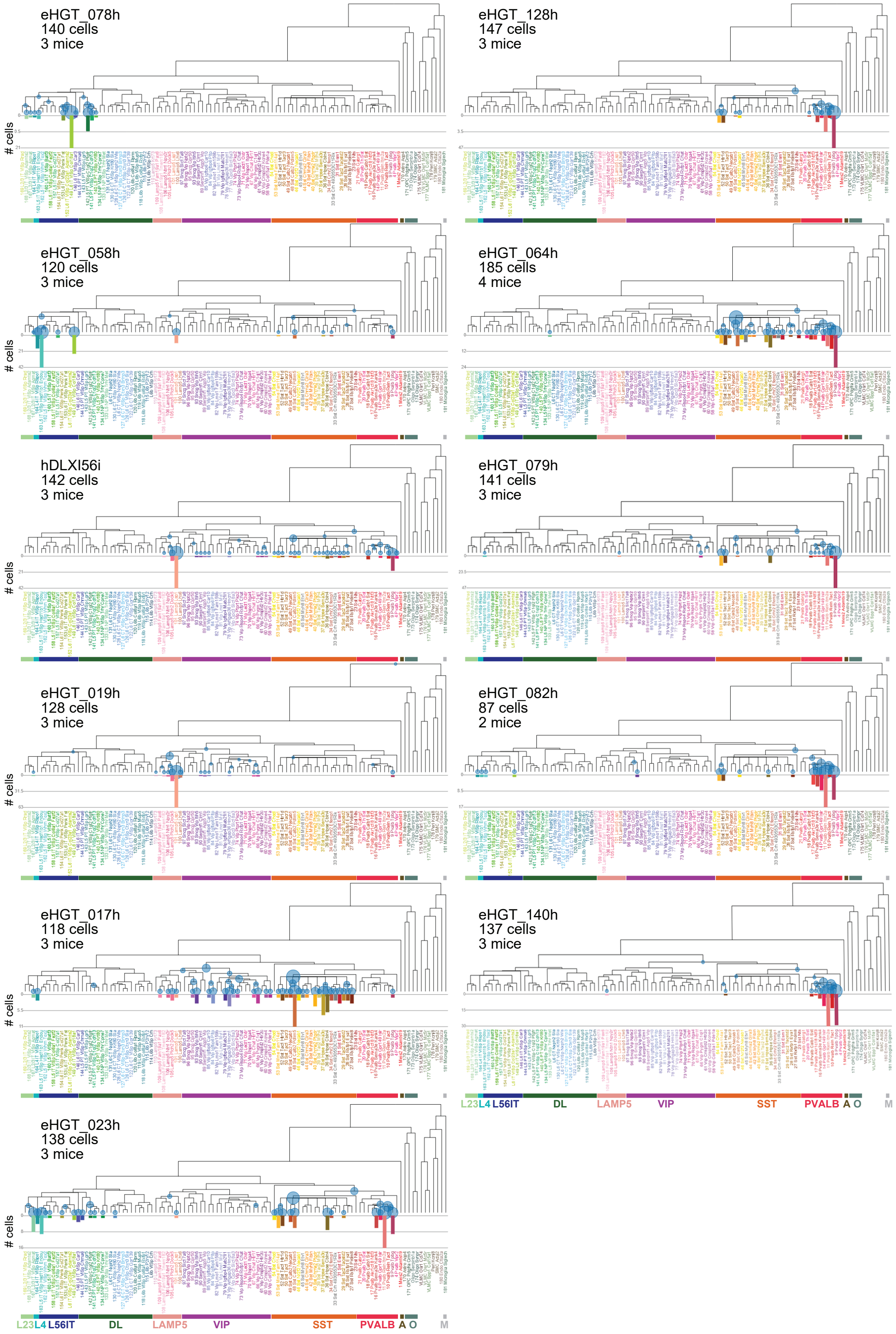
C-D) Correlation between motif accessibilities and transcript abundances across cell subclasses for C) DLX1 and D) NEUROD6 (grouping by average for motif accessibility, and by sum for transcript abundances). r, Pearson correlation coefficient. Two-tailed paired t-tests for significant correlation: DLX1 t = 3.0 df = 9 p < 0.01; NEUROD6 t = 5.4 df = 9 p < 0.001.

E) Percent overlap of ATAC-seq peaks with previously identified DMRs (Lister et al., 2013; Luo et al., 2017), comparing real peaks to randomized peak positions. Absolute numbers of detected peaks and peak-DMR overlaps are shown. Error bars represent standard deviation across 100 bootstrapped iterations using a subsampling rate of 80%.

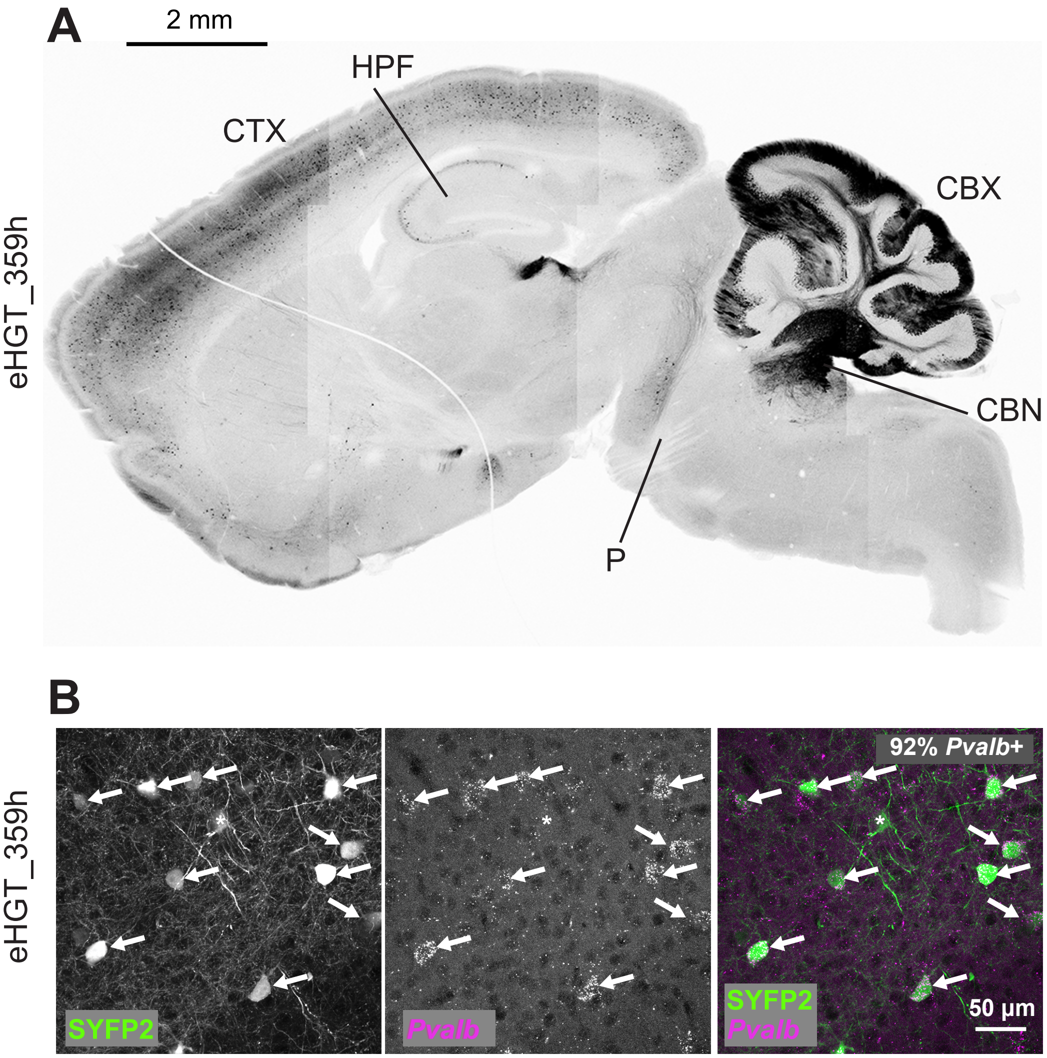
F) Mean phyloP scores across all peaks for cell subclass ATAC-seq peaks (colored line), compared to randomized peak positions (broken line).

G) Active transcriptional regulators in human and mouse brain cell subclasses, revealed by motifs in ATAC-seq peaks and gene expression by transcriptomics (Tasic et al., 2018; Hodge et al., 2019). E-value indicates the p-value from Fisher's exact test, corrected for multiple testing as calculated by MEME-CHIP. Arrows indicate strong and specific microglial enrichments for SPI1/PU.1 (gray) and for TEAD in human astrocytes (brown) and for OLIG2 in oligodendrocytes/OPCs (cadet blue).

H-I) Overlap of conserved or divergent peaks by cell subclass with multiple classes of repetitive genomic elements in both human (H) and mouse (I) sn/scATACseq datasets. An enrichment value of 1.0 corresponds to no fold change between real and random peak enrichment. Black bars represent the mean across the eleven neocortical cell subclasses. Heteroscedastic t-tests: *** p < 0.001, ** p < 0.01, ns not significant. Human all elements t = 5.2, df = 14.5; human DNA transposons t = 3.4, df = 14.8; human LINE t = 5.3, df = 18.3; human SINE t = 3.8, df = 18.8; human LTR t = 6.1, df = 18.3; human satellite t = 0.2, df = 12.4; human simple t = 1.6, df = 10.1; mouse all elements t = 3.7, df = 16.9; mouse DNA transposons t = 1.3, df = 12.7; mouse LINE t = 5.2, df = 18.5; mouse SINE t = 5.2, df = 18.0; mouse LTR t = 6.1, df = 17.5; mouse satellite t = 1.6, df = 9.7; mouse simple t = 0.3, df = 19.0.


eHGT_078h
140 cells
3 mice
eHGT_128h
147 cells
3 mice
eHGT_058h
120 cells
3 mice
eHGT_064h
185 cells
4 mice
hDLXI56i
142 cells
3 mice
eHGT_079h
141 cells
3 mice
eHGT_019h
128 cells
3 mice
eHGT_082h
87 cells
2 mice
eHGT_017h
118 cells
3 mice
eHGT_140h
137 cells
3 mice
eHGT_023h
138 cells
3 mice
# cells
L23
L4
L56IT
DL
LAMP5
VIP
SST
PVALB
A
O
M

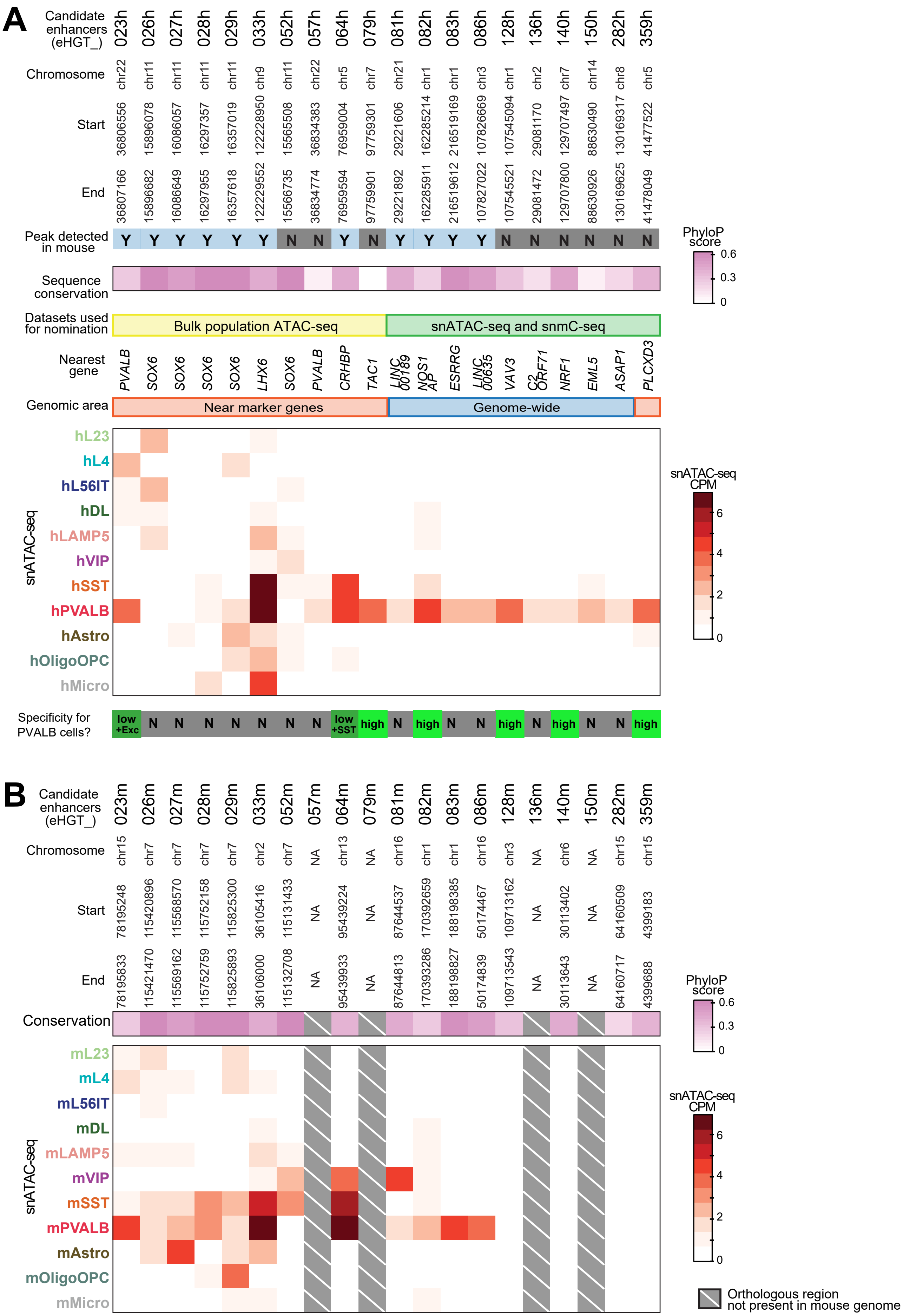
Supplementary Figure 5 (Related to Figure 3): Cell type validation of enhancer-AAV-labeled cells via scRNA-seq. Numbers of sorted labeled cells with each enhancer-AAV vector shown in Figures 3 and 4, mapped to the cell type transcriptomic taxonomy of mouse VISp (Tasic et al., 2018). Dendrogram leaves represent 111 transcriptomic cell types. Circles on the dendrogram represent the number of cells that could be mapped to that point in the dendrogram (starting from the root) and bar plots below the leaves represent the number of each cell type recovered that mapped to that final leaf.


A
2 mm
HPF
CTX
CBX
CBN
P
eHGT_359h
B
eHGT_359h
SYFP2
Pvalb
92% Pvalb+
SYFP2
Pvalb
50 µm

Supplementary Figure 6 (Related to Figure 4): eHGT_359h labels neocortical PVALB cells with high specificity, and also subcortical zones of *Pvalb* expression.

A) Whole-brain labeling pattern by eHGT_359h. Abbreviations: CTX cerebral cortex, HPF hippocampal formation, P pons, CBX cerebellar cortex, CBN cerebellar nuclei.

B) mFISH in L2/3 of VISp demonstrating positive labeling of *Pvalb*$^+$ cells (arrows) by eHGT_359h. An asterisk denotes an off-target *Pvalb*$^-$ labeled cell. Data represents n = 1 experiment.



A

Candidate enhancers (eHGT_)	Chromosome	Start	End	Peak detected in mouse	Datasets used for nomination	Nearest gene	Genomic area	Specificity for PVALB cells?
023h	chr22	36806556	36807166	Y	Bulk population ATAC-seq	*PVALB*	Near marker genes	low +Exc
026h	chr11	15896078	15896682	Y		*SOX6*		N
027h	chr11	16086057	16086649	Y		*SOX6*		N
028h	chr11	16297357	16297955	Y		*SOX6*		N
029h	chr11	16357019	16357618	Y		*SOX6*		N
033h	chr9	122228950	122229552	Y		*LHX6*		N
052h	chr11	15565508	15566735	N		*SOX6*		N
057h	chr22	36834383	36834774	N		*PVALB*		N
064h	chr5	76959004	76959594	Y		*CRHBP*		low +SST
079h	chr7	97759301	97759901	N		*TAC1*		high
081h	chr21	29221606	29221892	Y	snATAC-seq and snmC-seq	*LINC 00189*	Genome-wide	N
082h	chr1	162285214	162285911	Y		*NOS1 AP*		high
083h	chr1	216519169	216519612	Y		*ESRRG*		N
086h	chr3	107826669	107827022	Y		*LINC 00635*		N
128h	chr1	107545094	107545521	N		*VAV3*		high
136h	chr2	29081170	29081472	N		*C2 ORF71*		N
140h	chr7	129707497	129707800	N		*NRF1*		high
150h	chr14	88630490	88630926	N		*EML5*		N
282h	chr8	130169317	130169625	N		*ASAP1*		N
359h	chr5	41477522	41478049	N		*PLCXD3*		high

B

Candidate enhancers (eHGT_)	Chromosome	Start	End
023m	chr15	78195248	78195833
026m	chr7	115420896	115421470
027m	chr7	115568570	115569162
028m	chr7	115752158	115752759
029m	chr7	115825300	115825893
033m	chr2	36105416	36106000
052m	chr7	115131433	115132708
057m	NA	NA	NA
064m	chr13	95439224	95439933
079m	NA	NA	NA
081m	chr16	87644537	87644813
082m	chr1	170392659	170393286
083m	chr1	188198385	188198827
086m	chr16	50174467	50174839
128m	chr3	109713162	109713543
136m	NA	NA	NA
140m	chr6	30113402	30113643
150m	NA	NA	NA
282m	chr15	64160509	64160717
359m	chr15	4399183	4399688

Sequence conservation

Conservation

Supplementary Figure 7 (Related to Figure 4): Accessibility of candidate PVALB enhancers in human and mouse neocortical subclasses.

A) Twenty candidate PVALB enhancers from human epigenetic data characterized by conservation, method of identification, genomic location, and ATAC-seq profiles across cell subclasses. Coordinates correspond to hg38 genome in (A).

B) Orthologous mouse regions characterized by the same metrics. No mouse regions orthologous to eHGT_057h, 079h, 136h, and 150h could be identified, using liftOver with minMatch parameter set to either 0.6 or 0.5. CPM counts per million. Coordinates correspond to mm10 genome in (B).