Title: Epigenetic landscape and AAV targeting of human neocortical cell classes

Authors: John K. Mich[1*], Erik E. Hess[1$], Lucas T. Graybuck[1$], Saroja Somasundaram[1], Jeremy A. Miller[1], Yi Ding[1], Nadiya V. Shapovalova[1], Olivia Fong[1], Shenqin Yao[1], Marty Mortrud[1], Peter Chong[1], Darren Bertagnolli[1], Jeff Goldy[1], Tamara Casper[1], Matthew Kroll[1], Rebecca D. Hodge[1], Trygve E. Bakken[1], Zizhen Yao[1], Nick Dee[1], Ali Cetin[1], Kimberly A. Smith[1], Ryder P. Gwinn[2], Charles Cobbs[2], Andrew. L. Ko[3], Jeffrey G. Ojemann[4], C. Dirk Keene[5], Daniel. L. Silbergeld[6], Viviana Gradinaru[7], Susan M. Sunkin[1], Hongkui Zeng[1], Ed S. Lein[1], Bosiljka Tasic[1], Jonathan T. Ting[1], Boaz P. Levi[1*]

Affiliations: [1]Allen Institute for Brain Science, Seattle WA USA.

[2]Swedish Neurosci. Institute, Seattle WA USA.

[3]Dept. of Neurolog. Surgery, Univ. of Wash. Sch. of Med., Seattle WA USA; Regional Epilepsy Ctr., Harborview Med. Ctr., Seattle WA USA.

[4]Dept. of Neurological Surgery, Univ. of Wash., Seattle WA USA.

[5]Dept. of Pathology, Univ. of Wash., Seattle WA USA.

[6]Dept. of Neurological Surgery and Alvord Brain Tumor Center, Univ. of Wash., Seattle WA USA.

[7]Div. of Biol. and Biol. Engineering, California Inst. of Tech., Pasadena CA, USA.

[$]Equal contribution.

*Correspondence should be addressed to J.K.M. (johnmi@alleninstitute.org) or B.P.L (boazl@alleninstitute.org).

Abstract: Myriad cell types comprise the human neocortex, but their roles in normal brain function and disease are largely unknown because few tools exist. To find enhancer elements useful for cell type-specific genetic tools, we examined chromatin accessibility in >2,800 high-quality single human neocortical nuclei. Accessible elements frequently are conserved in mouse

(34%), often overlap with hypomethylated sites (27%), and connect cell types with neurological diseases via trait-associated SNPs. Directly testing these elements in viral vectors demonstrates functional enhancer activity with cell type specificity predicted by their chromatin accessibility patterns. In summary we present a catalog of human cell class-specific epigenetic elements, and utilize them for new species-agnostic cell type-specific viral genetic tools, which will illuminate human neuron function and drive gene therapy applications.

One sentence summary: Mapping chromatin accessibility in human brain cell classes links gene regulation to disease and enables human genetic tools.

Main Text: The human neocortex has greatly expanded in size and complexity relative to that of other mammals (*1*, *2*). Along with neocortical expansion are increased abilities such as language and reasoning which are disrupted in human diseases like autism and schizophrenia (*3*, *4*). This structure contains billions of cells grouped into dozens, if not hundreds, of molecularly distinguishable cell types (*5*–*8*), many with unknown roles because specific genetic tools for their study have been lacking.

To find distinguishing neocortical cell enhancers, we generated high-quality chromatin accessibility profiles from multiple fresh neurosurgical specimens (bulk n = 5, single n = 14, Fig. S1), using the assay for transposase-accessible chromatin and high-throughput sequencing (ATAC-seq) (*9*–*11*) on both populations (Fig. S2) and sorted single nuclei (Fig. S3). We prepared 3,660 single nucleus (sn) ATAC-seq libraries (median of 48,542 uniquely mapped reads per nucleus), and used 2,858 quality-filtered nuclei for clustering and mapping to human snRNA-seq data (Figs. 1A, S4, S5, and Materials and Methods). We defined 27 robustly detectable snATAC-seq clusters that mapped to 18 transcriptomically defined types from a previous human temporal lobe snRNA-seq study (Fig. 1B, *8*). These cell types spanned three major classes of brain cell types: excitatory, inhibitory, and non-neuronal, which we subdivided

into eleven subclasses: excitatory layer 2/3 (L23), layer 4 (L4), layer 5/6 intra-telencephalic (L56IT), and deep layer non-intratelencephalic neurons (DL); inhibitory *LAMP5*$^+$, *VIP*$^+$, *SST*$^+$, and *PVALB*$^+$ neurons; and non-neuronal astrocytes, microglia, and oligodendrocytes/OPCs. Cell type subclasses were predominantly recovered from the expected sort strategy (Fig. 1B and Fig. S5G), and all subclasses included nuclei contributed by multiple specimens (Fig. 1C).

To identify putative regulatory elements within each subclass, we aggregated the data for all nuclei within each subclass, and called peaks using Homer (*12*). This analysis revealed peaks proximal to recently identified transcriptomic subclass-specific marker genes (*8*), thereby confirming our clustering and mapping strategy (Fig. 1E-F). We then used chromVAR (*13*) to identify differentially enriched transcription factor (TF) motifs for known neuronal regulators, including DLX1 in inhibitory neurons (Fig. 2A) and NEUROD6 in lower-layer excitatory neurons (Fig. 2B). Across subclasses, these TF motif accessibilities correlated with TF transcript abundances (paired t-tests for correlation; DLX1 $t = 3.0$ $p < 0.01$; NEUROD6 $t = 5.4$ $p < 0.001$; Fig. 2C-D; *8*). As a whole, these observations indicate strong concordance between RNA-seq and ATAC-seq data modalities at the cell subclass level.

To assess the correspondence among accessibility and epigenetic modifications and primary sequence, we first calculated the overlap between subclass snATAC-seq peaks and differentially methylated regions (DMRs) previously identified from human frontal cortex methylcytosine sequencing (Fig. S6, *14*, *15*). For every cell subclass, we observed a greater overlap of snATAC-seq peaks with DMRs than expected by chance (Fig. 2E), revealing thousands of independently validated human neocortical regulatory elements. In total 27 ± 20% (mean ± sd) of all human peaks were also identified as DMRs. Furthermore, we partitioned peaks as transcriptional start site (TSS)-proximal (≤20 kb) or TSS-distal (>20 kb distance to any TSS), and found peak-DMR overlap greater for TSS-distal peaks than TSS-proximal peaks, and greater for inhibitory than excitatory neurons (Fig. 2F; two-way ANOVA with Tukey's post-hoc correction; proximal versus distal peaks $F = 22.5$, $p < 0.001$; inhibitory versus excitatory neurons

F = 10.6, p < 0.01). Looking at primary sequence, peaks from all cell subclasses displayed greater than random conservation (Fig. 2G, *16*); this sequence conservation was significantly stronger in TSS-distal peaks relative to TSS-proximal peaks, and in inhibitory relative to excitatory peaks (Fig. 2H; two-way ANOVA with Tukey's post-hoc correction; proximal versus distal peaks *F* = 12.2, p < 0.01; inhibitory versus excitatory neurons *F* = 8.0, p < 0.01), agreeing with previous observations by Luo et al (*15*). Together, these analyses suggest that snATAC-seq identifies DNA elements that are likely functional, and with differing properties.

To identify regions of chromatin accessibility shared with mouse ("conserved"), as well as those unique to human ("divergent"), we aggregated mouse scATAC-seq peaks (*10*) to match our human dataset, and then computed Jaccard similarity coefficients between human and mouse subclasses by counting peak overlaps (Methods). All mouse subclasses displayed the highest similarity to the orthologous human subclasses, and all but one human subclass matched reciprocally (Fig. 3A). In addition, non-neuronal classes displayed the strongest cross-species similarity, followed by inhibitory neurons, whereas excitatory neurons displayed the weakest correspondence (Fig. 3A). This analysis yielded thousands of conserved and non-conserved peaks for each subclass, and many more conserved peaks than expected by chance alone (Fig. 3B, ** FDR < 0.01 in each subclass). In sum, 34 ± 10% (mean ± sd) of all human peaks were also detected in matching mouse subclasses. Finally, conserved peaks exhibited significantly greater primary sequence conservation than divergent peaks in both species (heteroscedastic t-test; human *t* = 10.3, p < 0.001; mouse *t* = 6.6, p < 0.001; Fig. 3C) and less repetitive element overlap (Fig. S7), suggesting that these elements perform important, evolutionarily conserved functions.

Next we sought to use these neocortical elements as tools to associate cell subclasses with brain diseases. We used Linkage Disequilibrium Score Regression (LDSC, *17*, *18*) to find significant associations between snATAC-seq peaks and SNPs identified in 15 brain disease genome-wide association studies with sufficient power (see Materials and Methods, *19*). As

expected we observed similar association patterns using both snATAC-seq peaks and DMRs (Fig. S8, *14*, *15*), and generally weak associations for the outgroup disease (Crohn's disease) and outgroup peak set (Keratinocytes, Fig. 3D, *20*).

We partitioned human peaks into conserved and divergent sets (above), and found that conserved peaks had generally stronger peak-disease associations than divergent peaks (Fig. 3D). Associations were significant between multiple neuronal (but not non-neuronal) peak sets and educational attainment and schizophrenia (as shown with RNA-seq data , *21–23*), and furthermore conserved regions contain the majority of heritability for these traits (Fig. 3E) despite being numerically fewer (Fig. 3B). Interestingly, we observed the strongest enrichment between divergent microglial peaks and Alzheimer's disease SNPs (similar to previous reports, *21–23*), but this enrichment did not pass significance cutoff, possibly due to low overall total heritability and statistical power in Alzheimer's studies. Regardless, we conclude human-mouse conserved peaks likely regulate physiological human brain function, and their dysfunction may be causative for some neurological diseases.

To determine whether accessible genomic elements might provide useful genetic tools, we cloned several peaks into a reporter adeno-associated virus (AAV) vector packaged with mouse blood-brain barrier-penetrant capsid (PHP.eB) (Fig. 4A, *24*). Peaks were chosen based on proximity to known subclass-specific marker genes from snRNA-seq (*8*) with matching subclass-specific accessibility (Fig. 1E-F). We discovered several enhancers that drive distinct reporter expression patterns consistent with their expected subclass-specific accessibility profiles, including pan-excitatory and pan-inhibitory neurons, and subclass-specific L4, *PVALB*$^{+}$, *SST*$^{+}$, *VIP*$^{+}$, and *LAMP5*$^{+}$ expression patterns, when tested in mouse (Fig. 4B, *25*, *26*). We conclude that snATAC-seq enhancer discovery is a general strategy to identify cell class-specific genetic tools (*10*).

AAV viral tools are functional across species, which allows them to be used in both mouse and human studies. To test the function of an enhancer across species, we selected a

single region containing a conserved regulatory element found near the *LAMP5*$^+$/*VIP*$^+$ cell marker gene *CXCL14* (region eHGT_022, Fig. 4C). AAV vectors containing either the human or mouse ortholog of eHGT_022 (eHGT_022h and eHGT_022m, respectively) were sufficient to drive expression in upper-layer-enriched interneurons in both mouse and human (via live human *ex vivo* slice culture infection, Fig. 4D). When profiled by immunohistochemistry or single-cell RNA-seq, reporter-positive cells in both mouse and human mapped specifically to *LAMP5*$^+$ and *VIP*$^+$ neurons (Fig. 4E-F). These observations and those of the companion manuscript (*10*) suggest that ATAC-seq can identify cell subclass-specific enhancers for genetic tools in human and other species. Furthermore, these new eHGT_022 vectors allow the prospective marking and manipulation of *VIP*$^+$ and *LAMP5*$^+$ cells in human for the first time, and will be useful in combinatorial transduction experiments with future reporter vectors for increasingly specific labeling.

Human brain functions and diseases are often difficult to study because model organisms do not recapitulate human brain circuitry or display clear clinically relevant phenotypes. In particular, functionally relevant cell types are not yet known for many conditions, which leads to undertreatment of many debilitating brain disorders. It is thus critical to understand human brain-specific circuit components and their regulatory apparatus to provide avenues for therapeutic intervention. We have catalogued human neocortical chromatin accessibility with cell type precision, which deepens our knowledge of human brain chromatin structure and uncovers a cell type-specific logic in gene regulation. We expect this knowledge will not only guide models of human cognitive circuitry, but also fuel more precise gene therapy vectors for unmet clinical needs.

References and Notes:

1. H. Zeng *et al.*, *Cell*. **149**, 483–496 (2012).
2. P. Rakic, *Nat Rev Neurosci*. **10**, 724–735 (2009).
3. J. B. King *et al.*, *JAMA Netw Open*. **1**, e184777–e184777 (2018).
4. van den Heuvel MP, Sporns O, Collin G, et al, *JAMA Psychiatry*. **70**, 783–792 (2013).
5. A. Zeisel *et al.*, *Science*. **347**, 1138–1142 (2015).
6. B. Tasic *et al.*, *Nat Neurosci*. **19**, 335–346 (2016).
7. B. Tasic *et al.*, *Nature*. **563**, 72 (2018).
8. R. D. Hodge *et al.*, *bioRxiv*, 384826 (2018).
9. J. D. Buenrostro *et al.*, *Nature*. **523**, 486–490 (2015).
10. L. T. Graybuck *et al.*, *bioRxiv*, 525014 (2019).
11. L. T. Gray *et al.*, *eLife*. **6**, e21883 (2017).
12. S. Heinz *et al.*, *Molecular Cell*. **38**, 576–589 (2010).
13. A. N. Schep, B. Wu, J. D. Buenrostro, W. J. Greenleaf, *Nature Methods*. **14**, 975–978 (2017).
14. R. Lister *et al.*, *Science*. **341**, 1237905 (2013).
15. C. Luo *et al.*, *Science*. **357**, 600–604 (2017).
16. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, *Genome Res.* **20**, 110–121 (2010).
17. B. K. Bulik-Sullivan *et al.*, *Nature Genetics*. **47**, 291–295 (2015).
18. H. K. Finucane *et al.*, *Nat Genet*. **47**, 1228–1235 (2015).
19. See supplementary references for genome-wide association studies.
20. The ENCODE Project Consortium, *Nature*. **489**, 57–74 (2012).
21. N. G. Skene *et al.*, *Nature Genetics*. **50**, 825 (2018).
22. K. Girdhar *et al.*, *Nature Neuroscience*. **21**, 1126–1136 (2018).
23. D. A. Cusanovich *et al.*, *Cell*. **174**, 1309-1324.e18 (2018).
24. K. Y. Chan *et al.*, *Nat Neurosci*. **20**, 1172–1179 (2017).
25. T. Zerucha *et al.*, *J. Neurosci.* **20**, 709–721 (2000).

26. J. Dimidschstein *et al.*, *Nature Neuroscience*. **19**, 1743–1749 (2016).

Acknowledgements: Funding: We thank Allison Beller for assistance procuring and distributing tissue. This work is supported by NIH BRAIN Initiative award #1RF1MH114126-01 from the National Institute of Mental Health to BT, JKM, ZY, ESL, JTT, and BPL, and National Institute on Drug Abuse award #1R01DA036909-01 to LTG, HZ, and BT, and the Nancy and Buster Alvord Endowment to CDK. The content is solely the responsibility of the authors and does not necessarily represent the views of the funding agencies. In addition, we wish to thank the Allen Institute for Brain Science founder, Paul G. Allen, for his vision, encouragement and support. **Author contributions:** JTT, NS, EEH, TC, MK, ND, and JKM assisted on tissue handling and flow cytometry. EEH, BPL, and JKM performed ATAC-seq with assistance from DB and KAS. JKM analyzed ATAC-seq data using techniques developed by LTG and ZY, with assistance from SS, JAM, LTG, JG, and YD. EEH, JKM, and JTT performed molecular biology and design. Single cell RNA-seq was conducted by DB, KAS, and analysis by OF, JG, ZY, and BPL. JKM and JTT tested AAV vectors. RPG, CC, JGO, ALK, CDK, and DLS procured human tissue and consent. BPL, JTT, and JKM conceived of the study design. JKM wrote the manuscript and prepared figures. RDH and TEB provided human RNA-seq data. LTG and BT provided mouse ATAC-seq data. VG provided PHP.eB capsid. SMS provided program and budgetary management. HZ, ESL, BT, JTT, and BPL provided program leadership. **Competing interests:** The authors declare no outside interests. **Data and materials availability:** Consented data will be deposited in a public repository upon publication.

Supplementary Materials:

Materials and Methods

Figures S1-S8

Supplementary References (26-57)

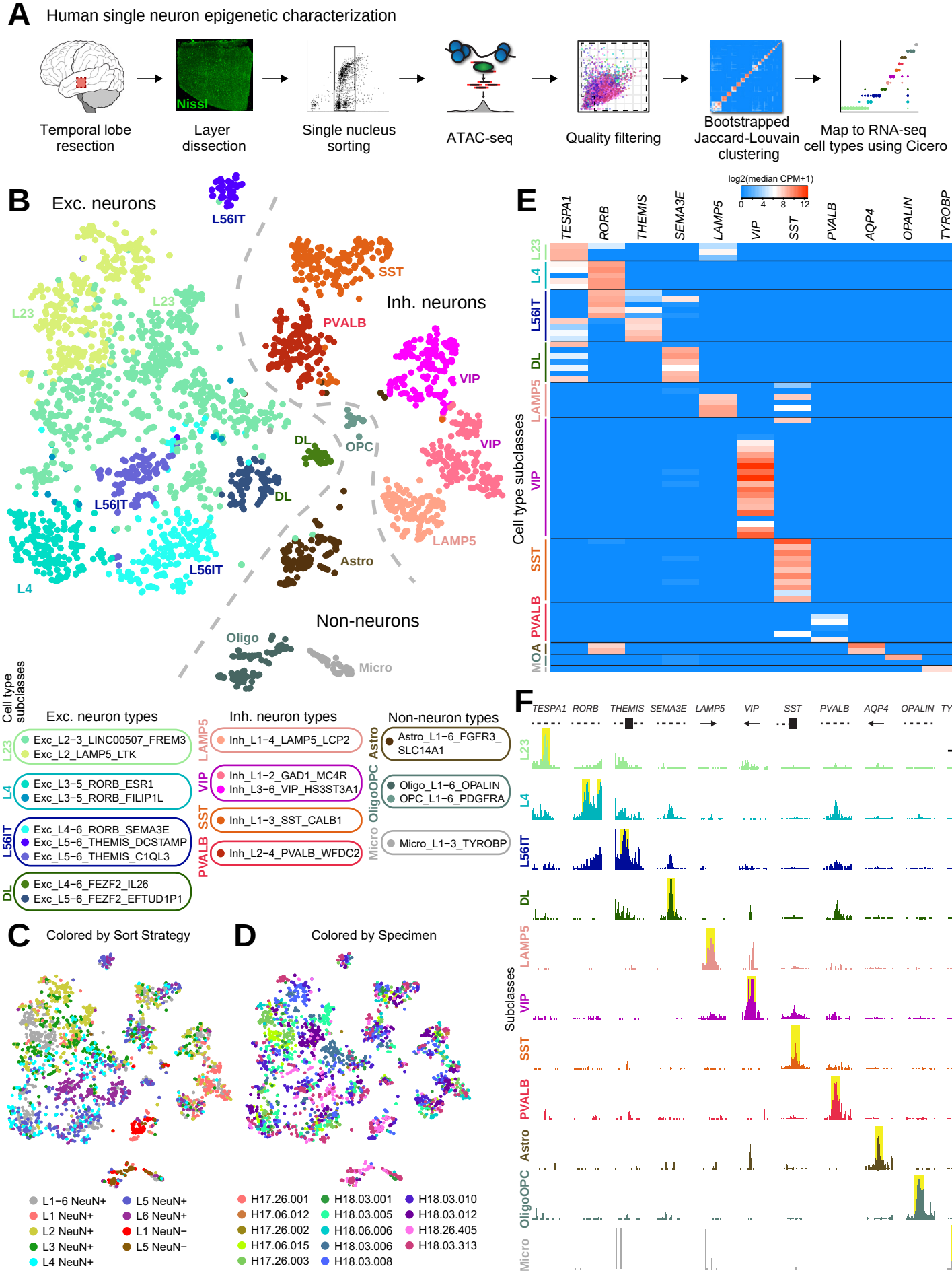

Figure 1: A database of human neocortical cell subclass-specific accessible chromatin elements.

A) Workflow for human neocortical epigenetic characterization.

B-D) High-quality nuclei (2858 from 14 specimens) visualized by *t*-SNE and colored according to mapped transcriptomic cell type (B), sort strategy (C), or specimen (D).

E) Transcriptomic abundances of eleven cell subclass-specific marker genes (median CPM for 75 cell types) identified in human temporal cortex middle temporal gyrus (*8*).

F) These eleven subclass-specific marker genes also demonstrate uniquely accessible chromatin elements in their vicinity (less than 50 kb distance to gene). Pileup heights are scaled proportionally to read number, and yellow bars highlight subclass-specific peaks for visualization. Dashed lines: introns, thick bars: exons, arrows: direction to gene body.

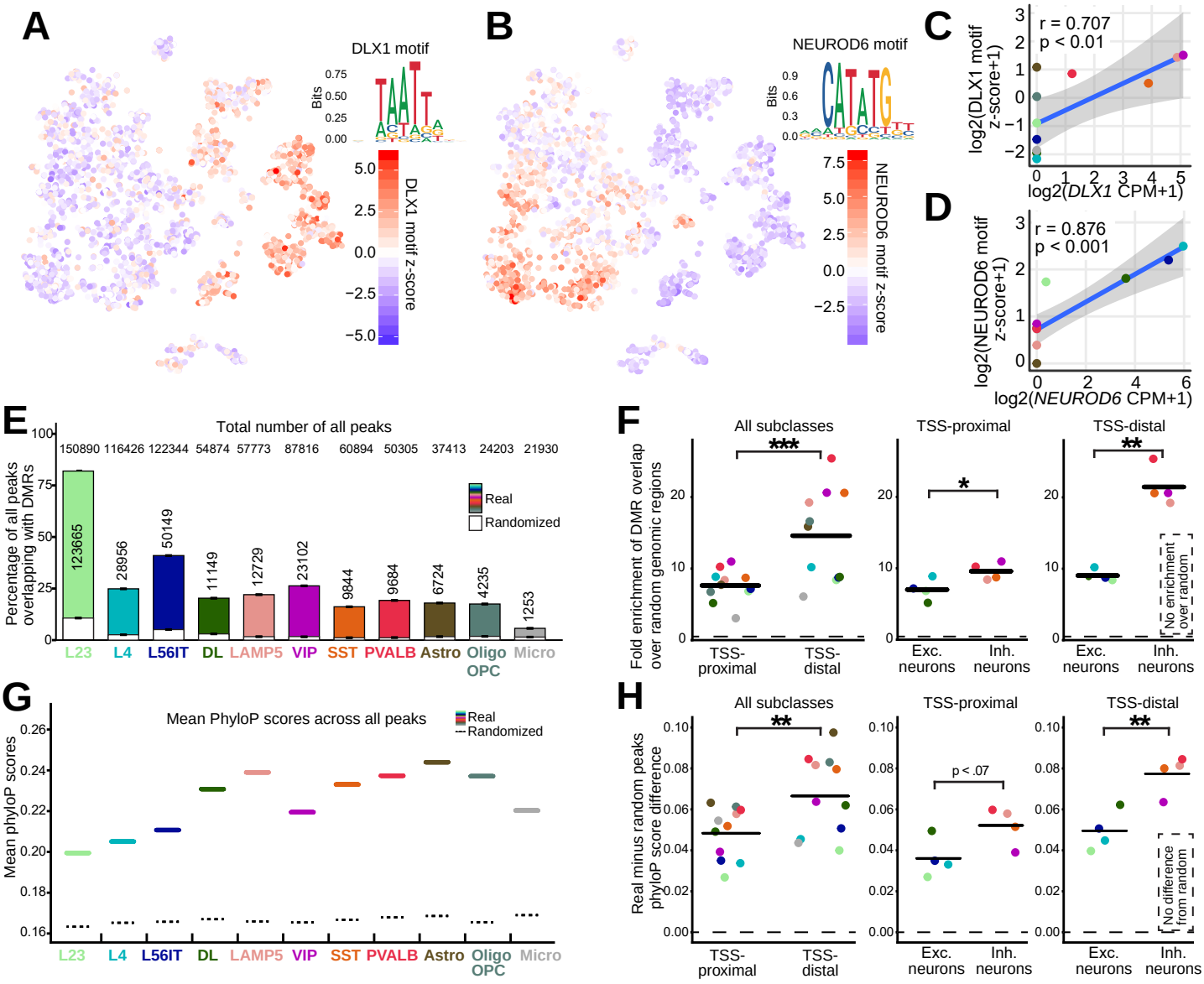

Figure 2: Properties of human neocortical cell subclass-specific accessible genomic elements.

A-B) Nuclei visualized by tSNE and colored by motif accessibilities for A) DLX1 and B) NEUROD6 as calculated by chromVAR (*13*).

C-D) Correlation between motif accessibilities and transcript abundances across cell subclasses for C) DLX1 and D) NEUROD6. 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 (*14*, *15*), comparing real peaks to randomized peak positions. Absolute numbers of detected peaks and peak-DMR overlaps are shown.

F) Peakset fold enrichments for DMR regions, calculated as ratio of real overlap to randomized overlap. Bars represent mean across subclasses, and dashed line indicates no enrichment over random. ***$p < 0.001$ (two-way ANOVA, $F = 22.5$), *$p < 0.05$ (heteroscedastic t-test, $t = 2.6$, df = 5.8), **$p < 0.01$ (heteroscedastic t-test, $t = 8.8$, df = 3.5).

G) Mean phyloP scores across all peaks for cell subclass ATAC-seq peaks, compared to randomized peak positions.

H) Real peak minus randomized peak phyloP score differences. Dashed line indicates no difference between randomized and real peaks. **$p < 0.01$ (two-way ANOVA, $F = 12.2$), $p = 0.07$ (heteroscedastic t-test, $t = 2.2$, df = 6.0), **$p < 0.01$ (heteroscedastic t-test, $t = 4.6$, df = 5.9).

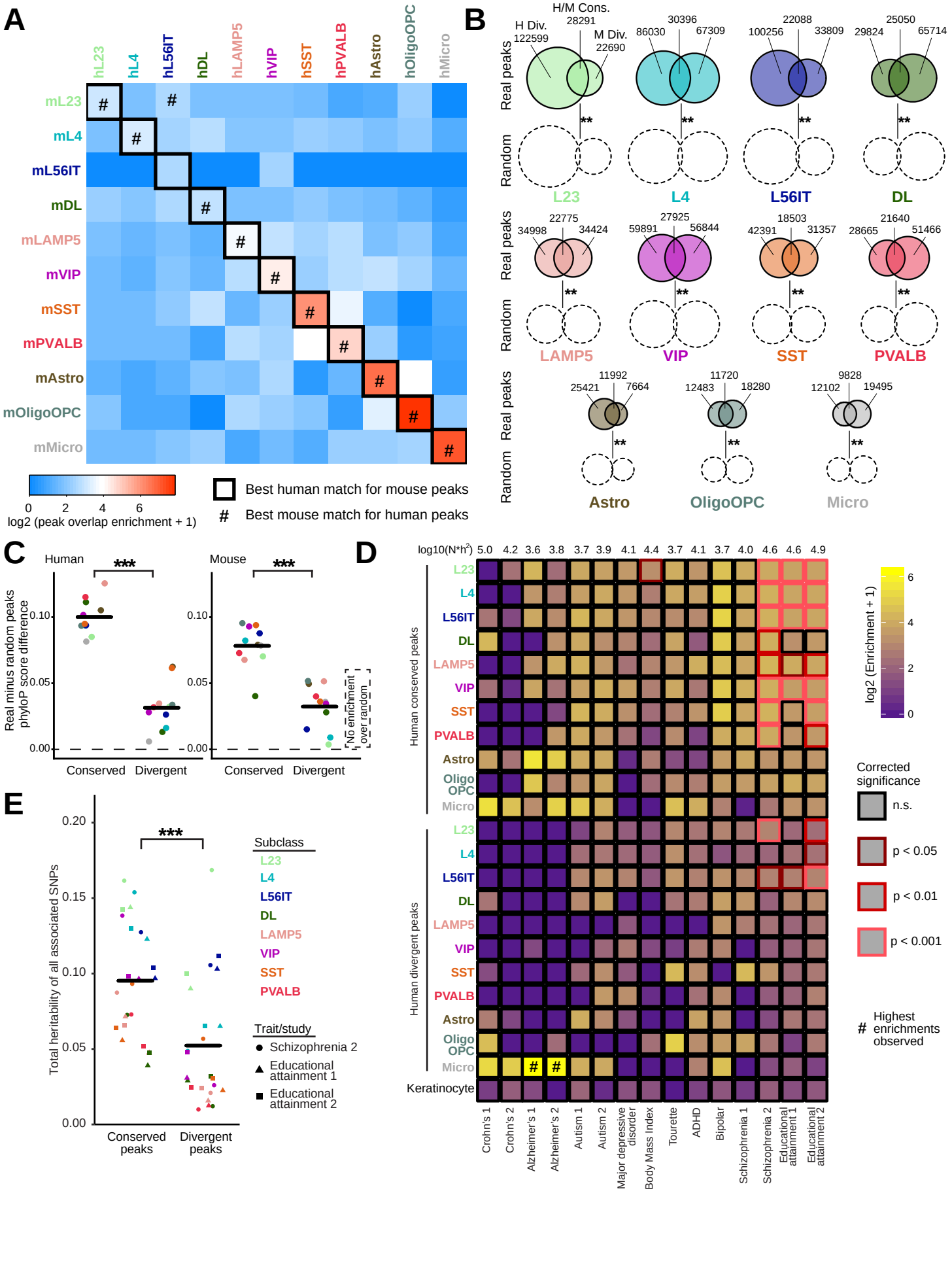

Figure 3: High conservation and disease relevance of human neocortical accessible genomic elements.

A) Jaccard similarity coefficient enrichments (ratio of real to randomized peaks) between human and mouse neocortical cell subclasses. Subclass peaksets almost always best match their orthologous peakset across species.

B) Visualization of conserved (Cons.) and divergent (Div.) peak counts across cell subclass in human and mouse. Conserved peaks are more frequent than expected by chance (** FDR < 0.01).

C) Conservedly accessible peaks display greater primary sequence conservation than divergently accessible peaks in both human and mouse. ***p < 0.001, by heteroscedastic t-test (human t = 10.3, df = 18.5; mouse t = 6.6, df = 19.9). Dashed line indicates no difference between real and randomized peak positions.

D) We observe generally greater enrichment and more significant associations for conserved peaks than for divergent peaks. A notable exception is microglia in Alzheimer's disease, which shows greater enrichment in divergent peaks. Heatmap colors displays log-transformed enrichment for peaks in disease/trait heritabilities, enrichment = Pr(h^2) / Pr(SNPs) as calculated by LDSC (*17*, *18*). Bonferroni-corrected p-values are employed (345 tests performed).

E) Overall, conserved peaks contain significantly more total heritability than divergent peaks, for three studies with multiple significant neuron subclass associations. ***p < 0.01 by heteroscedastic t-test, t = 3.8, df = 45.6.

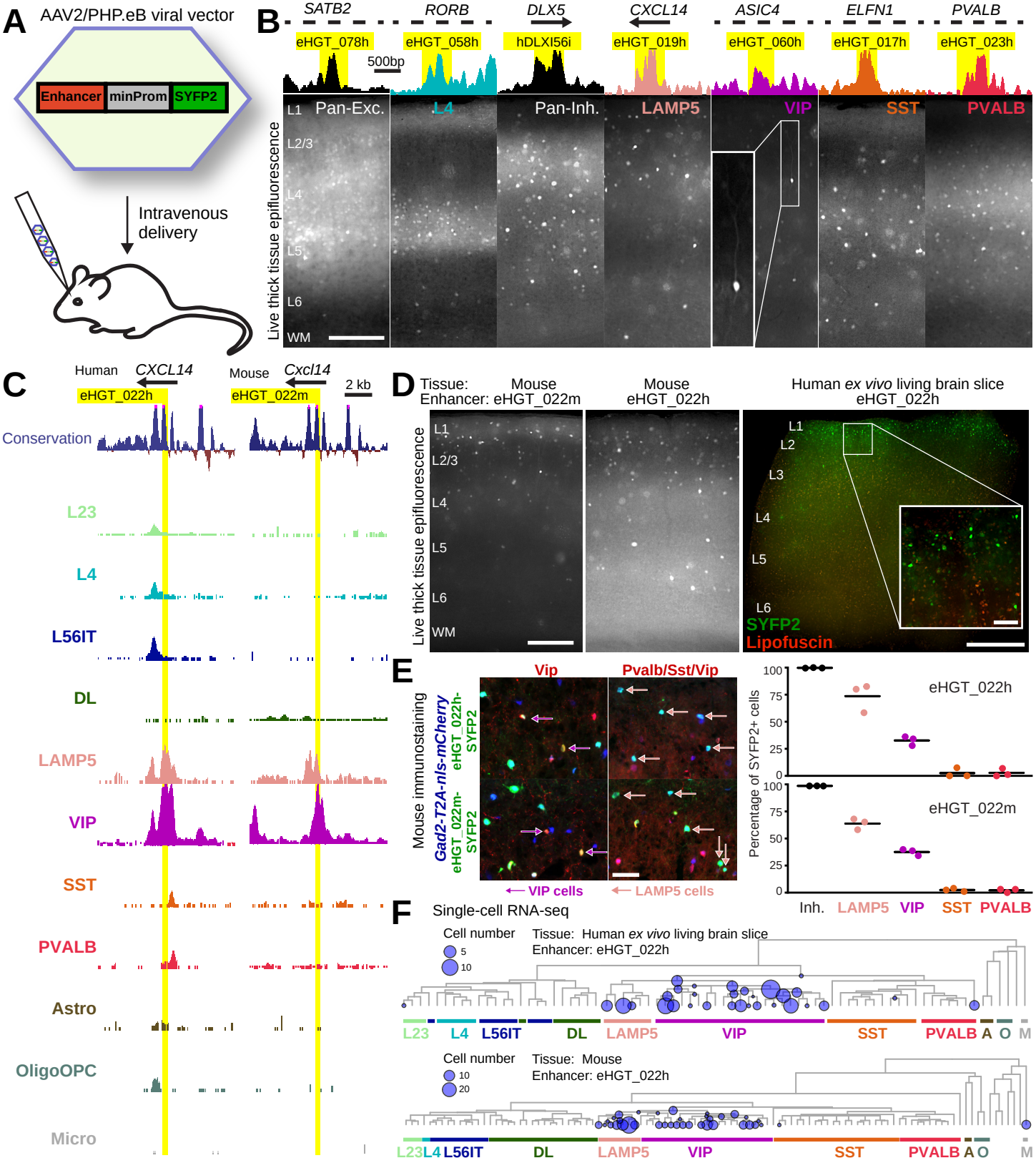

Figure 4: Accessible chromatin elements enable human genetic tools.

(A) AAV2/PHP.eB viral reporter vector design for testing presumptive enhancers.

(B) Multiple enhancer-AAV vectors yield distinct subclass selectivities. For seven enhancer-AAV reporter vectors we show ATAC-seq read pileups and thick tissue expression patterns in mouse V1. Scale 200 μm.

(C) eHGT_022h/m is accessible only in *LAMP5*$^+$ and *VIP*$^+$ neuron subclasses.

(D) eHGT_022h/m drives expression in primarily upper-layer interneurons in mouse V1 and human *ex vivo* neorcortex. Scale 200 μm (mouse), 1.0 mm (human), 100 μm (human inset).

(E) Immunostaining eHGT_022h/m-SYFP2$^+$ V1 mouse cells suggests they are primarily *VIP*$^+$ or *LAMP5*$^+$ (Gad2$^+$Pvalb/Sst/Vip$^-$) neurons. Over 1000 cells from 3 mice counted for each reporter.

(F) RNA-seq profiling of single eHGT_022h-SYFP2$^+$ cells from mouse (142 cells from three experiments) and human (106 cells from one experiment) confirms selective labeling of *VIP*$^+$ and *LAMP5*$^+$ neurons in both species. Dendrogram leaves represent transcriptomic cell types (75 in human and 104 in mouse V1, *7, 8*) and circles represent cell types detected.