Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse

Trygve E. Bakken, Nikolas L. Jorstad, Qiwen Hu, Blue B. Lake, Wei Tian, Brian E. Kalmbach, Megan Crow, Rebecca D. Hodge, Fenna M. Krienen, Staci A. Sorensen, Jeroen Eggermont, Zizhen Yao, Brian D. Aevermann, Andrew I. Aldridge, Anna Bartlett, Darren Bertagnolli, Tamara Casper, Rosa G. Castanon, Kirsten Crichton, Tanya L. Daigle, Rachel Dalley, Nick Dee, Nikolai Dembrow, Dinh Diep, Song-Lin Ding, Weixiu Dong, Rongxin Fang, Stephan Fischer, Melissa Goldman, Jeff Goldy, Lucas T. Graybuck, Brian R. Herb, Xiaomeng Hou, Jayaram Kancherla, Matthew Kroll, Kanan Lathia, Baldur van Lew, Yang Eric Li, Christine S. Liu, Hanqing Liu, Anup Mahurkar, Delissa McMillen, Jeremy A. Miller, Marmar Moussa, Joseph R. Nery, Joshua Orvis, Scott Owen, Carter R. Palmer, Thanh Pham, Nongluk Plongthongkum, Olivier Poirion, Nora M. Reed, Christine Rimorin, Angeline Rivkin, William J. Romanow, Adriana E. Sedeño-Cortés, Kimberly Siletti, Saroja Somasundaram, Josef Sulc, Michael Tieu, Amy Torkelson, Herman Tung, Xinxin Wang, Fangming Xie, Anna Marie Yanny, Renee Zhang, Seth A. Ament, Hector Corrada Bravo, Jerold Chun, Alexander Dobin, Jesse Gillis, Ronna Hertzano, Patrick R. Hof, Thomas Höllt, Gregory D. Horwitz, C. Dirk Keene, Peter V. Kharchenko, Andrew L. Ko, Boudewijn P. Lelieveldt, Chongyuan Luo, Eran A. Mukamel, Sebastian Preissl, Aviv Regev, Bing Ren, Richard H. Scheuermann, Kimberly Smith, William J. Spain, Owen R. White, Christof Koch, Michael Hawrylycz, Bosiljka Tasic, Evan Z. Macosko, Steven A. McCarroll, Jonathan T. Ting, Hongkui Zeng, Kun Zhang, Guoping Feng, Joseph R. Ecker, Sten Linnarsson, Ed S. Lein

Correspondence: Ed S. Lein (edl@alleninstitute.org), Trygve E. Bakken (trygveb@alleninstitute.org)

Abstract

The primary motor cortex (M1) is essential for voluntary fine motor control and is functionally conserved across mammals. Using high-throughput transcriptomic and epigenomic profiling of over 450,000 single nuclei in human, marmoset monkey, and mouse, we demonstrate a broadly conserved cellular makeup of this region, whose similarity mirrors evolutionary distance and is consistent between the

transcriptome and epigenome. The core conserved molecular identity of neuronal and non-neuronal types allowed the generation of a cross-species consensus cell type classification and inference of conserved cell type properties across species. Despite overall conservation, many species specializations were apparent, including differences in cell type proportions, gene expression, DNA methylation, and chromatin state. Few cell type marker genes were conserved across species, providing a short list of candidate genes and regulatory mechanisms responsible for conserved features of homologous cell types, such as the GABAergic chandelier cells. This consensus transcriptomic classification allowed the Patch-seq identification of layer 5 (L5) corticospinal Betz cells in non-human primate and human and characterization of their highly specialized physiology and anatomy. These findings highlight the robust molecular underpinnings of cell type diversity in M1 across mammals and point to the genes and regulatory pathways responsible for the functional identity of cell types and their species-specific adaptations.

Introduction

Single-cell transcriptomic and epigenomic methods provide a powerful lens on understanding the cellular makeup of highly complex brain tissues based on distinct patterns of gene expression and underlying regulatory mechanisms [1–7]. Applied to mouse and human neocortex, single-cell or single-nucleus transcriptomic analysis has yielded a complex but finite classification of cell types with approximately 100 discriminable neuronal and non-neuronal types in any given neocortical region [1,2,6,8]. Similar analyses using epigenomic methods have shown that many cortical cell types can be distinguished on the basis of regions of open chromatin or DNA methylation [5,9,10]. Furthermore, several recent studies have shown that transcriptomically-defined cell types can be aligned across species [2,11–13], indicating that these methods provide a path to quantitatively study evolutionary conservation and divergence at the level of cell types. However, application of these methods has been highly fragmented to date. Human and mouse comparisons have been performed in different cortical regions, using single-cell (with biases in cell proportions) versus single-nucleus (with biases in transcript

makeup) analysis, and most single-cell transcriptomic and epigenomic studies have been performed independently.

The primary motor cortex (MOp in mouse, M1 in human and non-human primates, all referred to as M1 herein) provides an ideal cortical region to address questions about cellular evolution in rodents and primates by integrating these approaches. Unlike the primary visual cortex (V1), which is highly specialized in primates, or frontal and temporal association areas, whose homologues in rodents remain poorly defined, M1 is essential for fine motor control and is functionally conserved across placental mammals. M1 is an agranular cortex, lacking a defined L4, although neurons with L4-like properties have been described [14]. L5 of carnivore and primate M1 contains exceptionally large "giganto-cellular" corticospinal neurons (Betz cells in primates [15,16] that contribute to the pyramidal tract and are highly specialized for their unusually large size with distinctive "taproot"-style dendrites [17,18]. Extracellular recordings from macaque corticospinal neurons reveal distinctive action potential properties supportive of a high conduction velocity and similar, unique properties have been reported during intracellular recordings from giganto-cellular neurons in cats[19–21]. Additionally, some primate Betz cells directly synapse onto alpha motor neurons, whereas in cats and rodents these neurons synapse instead onto spinal interneurons [22,23]. These observations suggest that Betz cells possess specialized intrinsic mechanisms to support rapid communication, some of which are primate specific.

Conservation of cellular features across species is strong evidence for evolutionary constraints on important cellular function. To explore evolutionary conservation and divergence of the M1 cellular makeup and its underlying molecular and gene regulatory mechanisms, we combined saturation coverage single-nucleus transcriptome analysis, DNA methylation, and combined open chromatin and transcriptome analysis of mouse, marmoset, and human M1 and transcriptomic profiling of macaque M1 L5. We describe a robust classification of neuronal and non-neuronal cell types in each species that is highly consistent between the transcriptome and epigenome. Cell type alignment accuracy and similarity varied as a function of evolutionary distance, with human more similar to non-human primate

than to mouse. We derived a consensus mammalian classification with globally similar cellular diversity, varying proportions, and species specializations in gene expression between conserved cell classes. Few genes had conserved cell type-specific expression across species and likely contribute to other conserved cellular properties, such as the unique morphology of chandelier GABAergic neurons. Conversely, these data also allow a targeted search for genes responsible for species specializations such as the distinctive anatomy, physiology and axonal projections of Betz cells, large corticospinal neurons in primates that are responsible for voluntary fine motor control. Together these findings highlight the strength of a comparative approach to understand cortical cellular diversity and identify conserved and specialized gene and gene regulatory mechanisms underlying cellular identity and function.

We made all primary and analyzed data publicly available. Raw sequence data are available for download from the Neuroscience Multi-omics Archive (nemoarchive.org) and the Brain Cell Data Center (biccn.org/data). Visualization and analysis tools are available at NeMO Analytics (nemoanalytics.org) and Cytosplore Viewer (viewer.cytosplore.org). These tools allow users to compare cross-species datasets and consensus clusters via genome and cell browsers and calculate differential expression within and among species. A semantic representation of the cell types defined through these studies is available in the provisional Cell Ontology (https://bioportal.bioontology.org/ontologies/PCL; Supplementary Table 1).

Results

Multi-omic cell type taxonomies

To characterize the molecular diversity of M1 neurons and non-neuronal cells, we applied multiple single-nucleus transcriptomic (plate-based SMART-seq v4, SSv4, and droplet-based Chromium v3, Cv3, RNA-sequencing) and epigenomic (single-nucleus methylcytosine sequencing 2, snmC-seq2; single-nucleus chromatin accessibility and mRNA expression sequencing, SNARE-seq2) assays on

107 isolated M1 samples from human (Extended Data Fig. 1a), marmoset, and mouse brain. Cellular
108 diversity was also profiled selectively in M1 L5 from macaque monkeys using Cv3 (Fig. 1b) to allow
109 Patch-seq mapping in physiology experiments. M1 was identified in each species based on its
110 stereotyped location in the caudal portion of frontal cortex and histological features such as the
111 presence of exceptionally large pyramidal neurons in L5 of M1, classically known as Betz cells in
112 human, other primates, and carnivores (Fig. 1a; [17]). Single nuclei were dissociated, sorted, and
113 transcripts were quantified with Cv3 for deep sampling in all four species, and additionally using SSv4
114 in human and mouse for full-length transcript information. For human and a subset of mouse nuclei,
115 individual layers of M1 were profiled independently using SSv4. Whole-genome DNA methylation, and
116 open chromatin combined with transcriptome measurements, were quantified in single nuclei from a
117 subset of species (Fig. 1b). Mouse datasets are also reported in a companion paper [6]. Median neuronal
118 gene detection was higher in human using SSv4 (7296 genes) than Cv3 (5657), partially due to 20-fold
119 greater read depth, and detection was lower in marmoset (4211) and mouse (5046) using Cv3
120 (Extended Data Fig. 1b-i).
121
122 For each species, a diverse set of neuronal and non-neuronal cell type clusters were defined based on
123 unsupervised clustering of snRNA-seq datasets (cluster metadata in Supplementary Table 2). Human
124 SSv4 and Cv3 data were integrated based on shared co-expression using Seurat [24], and 127 clusters
125 were identified that included nuclei from both RNA-seq platforms (Extended Data Fig. j-l). Marmoset
126 clusters (94) were determined based on independent clustering of Cv3 data using a similar analysis
127 pipeline. Mouse clusters (116) were defined in a companion paper [6] using seven integrated
128 transcriptomics datasets. These differences in the number of clusters are likely due to a combination of
129 statistical methodological differences as well as sampling and data quality differences rather than true
130 biological differences in cell diversity. For example, more non-neuronal nuclei were sampled in mouse
131 (58,098) and marmoset (21,189) compared to human (4,005), resulting in greater non-neuronal
132 resolution in those species. t-SNE visualizations of transcriptomic similarities across nuclei revealed

133 well-separated clusters in all species and mixing among donors, with some donor-specific technical
134 effects in marmoset (Extended Data Fig. 1m,n).
135
136 Post-clustering, cell types were organized into taxonomies based on transcriptomic similarities and
137 given a standardized nomenclature (Supplementary Table 3). As described previously for a different
138 cortical region [2], taxonomies were broadly conserved across species and reflected different
139 developmental origins of major non-neuronal and neuronal classes (e.g. GABAergic neurons from
140 ganglionic eminences (GEs) versus glutamatergic neurons from the cortical plate) and subclasses (e.g.
141 GABAergic CGE-derived *Lamp5/Sncg* and *Vip* versus MGE-derived *Pvalb* and *Sst*), allowing
142 identification and naming of these subclasses across species. Consequently, cardinal cell subclass
143 properties can be inferred, such as intratelencephalic (IT) projection patterns. Greater species variation
144 was seen at the highest level of resolution (cell types) that are named based on transcription data in
145 each species including the layer (if available), major class, subclass marker gene, and most specific
146 marker gene (e.g. L3 Exc *RORB OTOGL* in human; additional markers in Supplementary Tables 4-6).
147 GABAergic types were uniformly rare (< 4.5% of neurons), whereas more variable frequencies were
148 found for glutamatergic types (0.01 to 18.4% of neurons) and non-neuronal types (0.15% to 56.2% of
149 non-neuronal cells).
150
151 Laminar dissections in human M1 further allowed the estimation of laminar distributions of cell types
152 based on the proportions of nuclei dissected from each layer (Fig. 1c). As expected and previously
153 reported in middle temporal gyrus (MTG) of human neocortex [2], glutamatergic neuron types were
154 specific to layers. A subset of CGE-derived *Lamp5/Sncg* GABAergic neurons were restricted to L1, and
155 MGE-derived GABAergic types (*Sst* and *Pvalb*) displayed laminar patterning, with transcriptomically
156 similar types showing proximal laminar distributions, whereas *Vip* GABAergic neuron types displayed
157 the least laminar specificity. Three astrocyte subtypes had frequencies and layer distributions that
158 correlated with known morphologically-defined astrocyte types [25], including a common type in all layers
159 (protoplasmic), a rare type in L1 (interlaminar) [26], and a rare type in L6 (fibrous).

160

161 Single-nucleus sampling provides a relatively unbiased survey of cellular diversity [2,27] and enables
162 comparison of cell subclass proportions across species. For each donor, we estimated the proportion of
163 GABAergic and glutamatergic cells among all neurons and compared the proportions across species.
164 Consistent with previously reported differences in GABAergic neuron frequencies in primate versus
165 rodent somato-motor cortex based on histological measurements (reviewed in [28]), we found twice as
166 many GABAergic neurons in human (33%) compared to mouse M1 (16%) an intermediate proportion
167 (23%) in marmoset (Fig. 1f). Despite these differences, the relative proportions of GABAergic neuron
168 subclasses were similar. Exceptions to this included an increased proportion of *Vip* and *Sncg* cells and
169 decreased proportion of *Pvalb* cells in marmoset. Among glutamatergic neurons, there were
170 significantly more L2 and L3 IT neurons in human than marmoset and mouse (Fig. 1f), consistent with
171 the dramatic expansion of supragranular cortical layers in human (Fig. 1a) [29]. The L5
172 extratelencephalic-projecting (ET) types (also known as pyramidal tract, PT, or subcerebral types),
173 including corticospinal neurons and Betz cells in primate M1, comprised a significantly smaller
174 proportion of glutamatergic neurons in primates than mouse. This species difference was also reported
175 in MTG [2], possibly reflecting the spatial dilution of these cells with the expansion of neocortex in
176 primates. Similarly, the L6 cortico-thalamic (CT) neuron populations were less than half as frequent in
177 primates compared to mouse, whereas the L6 *Car3* type was rare in all species and relatively more
178 abundant in marmoset.

179

180 Individual nuclei were isolated from M1 of the same donors for each species and molecular profiles
181 were derived for DNA methylation (snmC-seq2) and open chromatin combined with mRNA (SNARE-
182 seq2). Independent unsupervised clustering of epigenomic data also resulted in diverse clusters (see
183 below, Figs. 4 and 5) that were mapped back to RNA clusters based on shared (directly measured or
184 inferred) marker gene expression. Cell epigenomes were highly correlated with transcriptomes, and all
185 epigenomic clusters mapped to one or more transcriptomic clusters. The epigenome data generally had
186 lower cell type resolution (Fig. 1c-e), although this may be due to sampling fewer cells or sparse

187 genomic coverage. Interestingly, snmC-seq2 and SNARE-seq2 resolved different granularities of cell
188 types. For example, more GABAergic *Vip* neuron types were identified in human M1 based on DNA-
189 methylation than open chromatin, despite profiling only 5% as many nuclei with snmC-seq2 (Fig. 1c).
190

191 **Consensus cellular M1 taxonomy across species**

192 A consensus cell type classification identifies conserved molecular makeup and allows direct cross-
193 species comparisons. The snRNA-seq Cv3 datasets were integrated using Seurat [24] that aligns nuclei
194 across species based on shared co-expression of a subset of orthologous genes with variable
195 expression. We repeated this analysis for three cell classes: GABAergic neurons (Fig. 2), glutamatergic
196 neurons (Extended Data Fig. 3) and non-neuronal cells (Extended Data Fig. 4). As represented in a
197 reduced dimension UMAP space (Fig.2a), GABAergic neuronal nuclei were well-mixed across species.
198 Eight well-defined populations formed distinct islands populated by cells from all three species,
199 including CGE-derived (*Lamp5*, *Sncg*, *Vip*) and MGE-derived (*Pvalb*, *Sst*, *Sst Chodl*) subclasses, and
200 *Lamp5 Lhx6* and chandelier cell (ChC) types. To identify conserved molecular expression for each
201 subclass across species, we first identified genes that were enriched in each subclass ("markers")
202 compared to all GABAergic neurons in each species (ROC test; AUC > 0.7). Then, we looked for
203 overlap among these genes across species. Each subclass had a core set of conserved markers (Fig.
204 2b, markers listed in Supplementary Table 7), and many subclass markers were species-specific. The
205 contrast between a minority of conserved and majority of species-specific marker genes enriched in
206 subclasses is particularly clear in the heatmap in Figure 2c (Supplementary Table 8). As expected
207 based on their closer evolutionary distance, human and marmoset shared more subclass markers with
208 each other than with mouse (Fig. 2b).
209

210 Cell types remained distinct within species and aligned across species in the integrated transcriptomic
211 space (Fig. 2d). To establish a consensus taxonomy of cross-species clusters, we used unsupervised
212 clustering to split the integrated space into more than 500 small clusters ('metacells') and built a
213 dendrogram and quantified branch stability by subsampling metacells and reclustering (Extended Data

214 Fig. 2a). Metacells were merged with nearest neighbors until all branches were stable and included
215 nuclei from the three species (see Methods). We used cluster overlap heatmaps to visualize the
216 alignment of cell types across species based on membership in merged metacells (Fig. 2e). 24
217 GABAergic consensus clusters displayed consistent overlap of clusters among the three species and
218 are highlighted as blue boxes in the heatmaps (Fig. 2e).
219
220 We next constructed a consensus taxonomy by pruning the metacell dendrogram (Extended Data Fig.
221 2a), and demonstrated that all types were well mixed across species (Fig. 2f, grey branches). The
222 robustness of consensus types was bolstered by a conserved set of marker genes (Extended Data Fig.
223 2d) and high classification accuracy of subclasses (Extended Data Fig. 2e, data in Supplementary
224 Table 9) and types (Extended Data Fig. 2f, data in Supplementary Table 9) compared to nearest
225 neighbors within and among species using a MetaNeighbor analysis [30]. Distinct consensus types (ChC,
226 *Sst Chodl*) were the most robust (mean AUROC = 0.99 within-species and 0.88 cross-species), while
227 *Sncg* and *Sst* types could not be as reliably differentiated from closely related types (mean AUROC =
228 0.84 within-species and 0.50 cross-species). Most consensus GABAergic types were enriched in the
229 same layers in human and mouse (Fig. 2f), although there were also notable species differences. For
230 example, ChCs were enriched in L2/3 in mouse and distributed across all layers in human as was seen
231 in temporal cortex (MTG) based on RNA ISH [2]. *Sst Chodl* was restricted to L6 in mouse and also found
232 in L1 and L2 in human, consistent with previous observations of sparse expression of *SST* in L1 in
233 human not mouse cortex [31].
234
235 More consensus clusters could be resolved by pairwise alignment of human and marmoset than
236 primates and mouse, particularly *Vip* subtypes (Fig. 2g, Extended Data Fig. 2b). Higher resolution
237 integration of cell types was also apparent in cluster overlap plots between human and marmoset
238 clusters (Fig. 2e, Extended Data Fig. 2c). We quantified the expression conservation of functionally
239 annotated sets of genes by testing the ability of gene sets to discriminate GABAergic consensus types.
240 This analysis was framed as a supervised learning task, both within- and between-species [30]. Within-

241 species, gene sets related to neuronal connectivity and signaling were most informative for cell type
242 identity (Extended Data Fig. 2g), as reported in human and mouse cortex [2,32]. All gene sets had
243 remarkably similar consensus type classification performance across species (r > 0.95; Fig. 2h),
244 pointing to strong evolutionary constraints on the cell type specificity of gene expression central to
245 neuronal function. Gene set classification performance was systematically reduced when training and
246 testing between primates (44% reduction) and between primates and mouse (65% reduction; Fig. 2h).
247 Therefore, many cell type marker genes were expressed in different consensus types between species.
248 Future comparative work can compare reductions in classification performance to evolutionary
249 distances between species to estimate rates of expression change across phylogenies.
250
251 Cross-species consensus types were defined for glutamatergic neurons using an identical approach as
252 for GABAergic neurons (Extended Data Fig. 3). In general, glutamatergic subclasses aligned well
253 across species and had a core set of conserved markers as well as many species-specific markers
254 (Extended Data Fig. 3a-c, genes listed in Supplementary Tables 10-11). 13 consensus types were
255 defined across species. Glutamatergic types had fewer conserved markers than GABAergic types
256 (Extended Data Fig. 3d-f,j), although subclasses and types were similarly robust (mean within-species
257 AUROC = 0.86 for GABAergic types and 0.85 for glutamatergic types) based on classification
258 performance (Extended Data Fig. 3k,l and Supplementary Table 9). Human and marmoset had
259 consistently more conserved marker genes than primates and mouse (Extended Data Fig. 3i) and could
260 be aligned at somewhat higher resolution (Extended Data Fig. 3g,h) for L5/6 NP and L5 IT subclasses.
261
262 Integration of non-neuronal cells was performed similarly to neurons (Extended Data Fig. 4a).
263 Consensus clusters (blue boxes in Extended Data Fig. 4c) that shared many marker genes were
264 identified across species (Extended Data Fig. 4d), and there was also evidence for the evolutionary
265 divergence of gene expression in consensus types. For example, the Astro_1 type had 560 DEGs
266 (Wilcox test; FDR < 0.01, log-fold change > 2) between human and mouse and only 221 DEGs
267 between human and marmoset (Extended Data Fig. 4e). The human cortex contains several

268 morphologically distinct astrocyte types [33]: interlaminar (ILA) in L1, protoplasmic in all layers, varicose
269 projection in deep layers, and fibrous in white matter (WM). We previously reported two transcriptomic
270 clusters in human MTG that corresponded to protoplasmic astrocytes and ILAs [2], and we validated
271 these types in M1 (Extended Data Fig. 4g,h). We identified a third type, Astro L1-6 *FGFR3 AQP1*, that
272 expresses *APQ4* and *TNC* and corresponds to fibrous astrocytes in WM (Extended Data Fig. 4g, left
273 ISH). A putative varicose projection astrocyte did not express human astrocyte markers (Extended Data
274 Fig. 4g, middle and right ISH), and this rare type may not have been sampled or is not
275 transcriptomically distinct.
276
277 Species comparison of non-neuronal cell types was more challenging than for neurons due to variable
278 sampling across species and more immature non-neuronal cells in mouse. 5- to 15-fold lower sampling
279 of non-neuronal cells in human impacted detection of rare types. For example, pericytes, smooth
280 muscle cells (SMCs), and some subtypes of vascular and leptomeningeal cells (VLMCs) were present
281 in marmoset and mouse and not human datasets (Extended Data Fig. 4c, right plot, blue arrows),
282 although these cells are clearly present in human cortex (for example, see [34]). A maturation lineage
283 between oligodendrocyte precursor cells (OPCs) and oligodendrocytes based on reported marker
284 genes [35] that was present in mouse and not primates (Extended Data Fig. 4b) likely represents the
285 younger age of mouse tissues used. Mitotic astrocytes (Astro_*Top2a*) were also only present in mouse
286 (Extended Data Fig. 4a,c) and represented 0.1% of non-neuronal cells. Primates had a unique
287 oligodendrocyte population (Oligo *SLC1A3 LOC103793418* in marmoset and Oligo L2-6 *OPALIN*
288 *MAP6D1* in human) that was not a distinct cluster in mouse (Extended Data Fig. 4c, left plot, blue
289 arrow). Surprisingly this oligodendrocyte clustered with glutamatergic types (Fig. 1c,d) and was
290 associated with neuronal transcripts such as *NPTX1*, *OLFM3*, and *GRIA1* (Extended Data Fig. 4i). This
291 was not an artifact, as FISH for markers of this type (*SOX10*, *ST18*) co-localized with neuronal markers
292 in the nuclei of cells that were sparsely distributed across all layers of human and marmoset M1
293 (Extended Data Fig. 4j). This may represent an oligodendrocyte type that expresses neuronal genes or

294 could represent phagocytosis of parts of neurons and accompanying transcripts that are sequestered in
295 phagolysosomes adjacent to nuclei.
296
297 To assess differential isoform usage between human and mouse, we used SSv4 data with full transcript
298 coverage and estimated isoform abundance in cell subclasses. Remarkably, 25% of moderately
299 expressed (> 10 transcripts per million) isoforms showed a large change (>9-fold) in usage between
300 species, and isoform switching was 30-60% more common in non-neuronal than neuronal cells
301 (Extended Data Fig. 2h,i, Supplementary Table 12). For example, β2-Chimaerin (*CHN2*), a gene shown
302 to mediate axonal pruning in the hippocampus [36], was highly expressed in human and mouse L5/6 NP
303 neurons. In mouse, the short isoform was almost exclusively expressed, while in human, longer
304 isoforms were also expressed (Extended Data Fig. 2j).
305
306 **Open chromatin profiling reveals distinct cell type gene regulation**
307 To directly match accessible chromatin profiles to RNA-defined cell populations, we used SNARE-Seq
308 [37], now modified for highly multiplexed combinatorial barcoding (SNARE-Seq2) [38]. We generated
309 84,178 and 9,946 dual-omic single-nucleus RNA and accessible chromatin (AC) datasets from human
310 (n = 2) and marmoset (n = 2) M1, respectively (Extended Data Fig. 5a-b, Supplementary Table 13). On
311 average, 2,242 genes (5,764 unique transcripts) were detected per nucleus for human and 3,858 genes
312 (12,400 unique transcripts) per nucleus for marmoset, due to more than 4-fold greater sequencing
313 depth for marmoset (average 17,576 reads per nucleus for human and 77,816 reads per nucleus for
314 marmoset).
315
316 To define consensus clusters, SNARE-seq2 single-nucleus RNA expression data were mapped to
317 human and marmoset transcriptomic clusters (Fig. 1c,d) based on correlated expression of cell type
318 marker genes. SNARE-seq2 transcriptomes were also independently clustered, with both approaches
319 giving consistent results (Extended Data Fig. 5c-f). Consensus clusters were more highly resolved in
320 transcriptomic compared to AC data (Extended Data Fig. 5g), and so an integrative approach was used

321 to achieve best matched AC-level cluster annotations (Extended Data Fig. 5h-k). AC peak calling at
322 multiple levels of cellular identity (for RNA consensus clusters, resolved AC clusters, subclasses and
323 classes) yielded a combined total of 273,103 (human) and 134,769 (marmoset) accessible sites, with
324 an average of 1527 or 1322 unique accessible peak fragment counts per nucleus, respectively. Gene
325 activity estimates based on cis-regulatory interactions predicted from co-accessible promoter and distal
326 peak regions using Cicero [39] were highly correlated with RNA expression values. This highlights the
327 ability of SNARE-Seq2 to meaningfully characterize AC at RNA-defined cellular resolution that cannot
328 be achieved using only AC data (Extended Data Fig. 6a-b). The AC-level clusters (Fig. 3a,b) that
329 showed similar coverage across individual samples (Extended Data Fig. 6c-f) revealed regions of open
330 chromatin that are extremely cell type specific (Fig. 3c). These regulatory regions were relatively more
331 abundant in glutamatergic compared to GABAergic neuron subpopulations (Fig. 3c-d, Supplementary
332 Table 14).
333
334 To better understand the interplay of gene regulation and expression, we compared transcript counts
335 and open chromatin measured in the same nuclei. For example, the GABAergic neuron marker *GAD2*
336 and the L2/3 glutamatergic neuron marker *CUX2* showed cell-type specific chromatin profiles for co-
337 accessible sites that were consistent with their corresponding transcript abundances (Fig. 3e-g).
338 Transcription factor binding site (TFBS) activities were calculated using chromVAR [40], permitting
339 discovery of differentially active TFBSs between cell types. To investigate the regulatory factors that
340 may contribute to marker gene expression, we evaluated active TFBSs for their enrichment within
341 marker gene co-accessible sites. This permitted direct cell type mapping of gene expression and
342 activity levels with the expression and activity of associated regulatory factors (Fig. 3g). Using this
343 strategy, we identified TFBS activities associated with subclass (Fig. 3h-i) and AC-cluster level
344 differentially expressed genes (DEGs) in human and marmoset (Supplementary Table 15). DEG
345 transcript levels and AC-inferred gene activity scores showed high correspondence (Fig. 3h). While
346 most subclasses also showed distinct TFBS activities, correspondence between human and marmoset
347 was higher for glutamatergic rather than GABAergic neurons (Fig. 3h,j). For GABAergic neuron

348 subclasses, gene expression profiles were more conserved than TFBS activities, consistent with fewer
349 differences between GABAergic subpopulations based on AC sites (Fig. 3a,b). This observation is also
350 consistent with fewer distinct TFBS activities among some inhibitory neuron subclasses (*Lamp5*, *Sncg*)
351 in human compared to marmoset (Fig. 3h), despite these cell types having a similar number of AC peak
352 counts (Extended Data Fig. 6d-f). Interestingly, glutamatergic neurons in L5 and L6 showed higher
353 correspondence between primates based on TFBS activities compared to average expression,
354 suggesting that gene regulatory processes are more highly conserved in these subclasses than target
355 gene expression.
356
357 **Methylomic profiling reveals conserved gene regulation**
358 We used snmC-seq2 [41] to profile the DNA methylome from individual cells in M1. Single-nuclei were
359 labeled with an anti-NeuN antibody and isolated by fluorescence-activated cell sorting (FACS), and
360 neurons were enriched (90% NeuN+ nuclei) to increase detection of rare types. Using snmC-seq2, we
361 generated single-nucleus methylcytosine datasets from M1 of human (n = 2 donors, 6,095 nuclei),
362 marmoset (n = 2, 6,090), and mouse (9,876) (Liu et al. companion paper) (Supplementary Table 16).
363 On average, 5.5 ± 2.7% (mean ± s.d.) of human, 5.6 ± 2.9% of marmoset and 6.2 ± 2.6% of mouse
364 genomes were covered by stringently filtered reads per cell, with 3.4×10^4 (56%), 1.8×10^4 (62%) and
365 4.5×10^4 (81%) genes detected per cell in the three species, respectively. Based on the DNA
366 methylome profiles in both CpG sites (CG methylation or mCG) and non-CpG sites (CH methylation or
367 mCH), we clustered nuclei (Methods) to group cell populations into 31 cell types in human, 36 cell types
368 in marmoset, and 42 cell types in mouse (Fig. 4a and Extended Data Fig. 7a,b). For each species, cell
369 type clusters could be robustly discriminated using a supervised classification model and had distinct
370 marker genes based on DNA methylation signatures for neurons (mCH) or non-neuronal cells (mCG)
371 (Methods). Differentially methylated regions (DMR) were determined for each cell type versus all other
372 cell types and yielded 9.8×10^5 DMRs in human, 1.0×10^6 in marmoset, and 1.8×10^6 in mouse.
373

374 We determined a consensus molecular classification of cell types in each species by integrating single-
375 nucleus methylomic data with the Cv3 transcriptomic data described above using measurements of
376 gene body differential methylation (CH-DMG) to approximate expression levels. Nuclei from the two
377 data modalities mixed well as visualized in ensemble UMAPs (Fig. 4b,c). Methylation clusters have
378 one-to-one, one-to-many, or many-to-many mapping relation to transcriptomic clusters (Fig. 1c-e and
379 Extended Data Fig. 7d-f). DMRs were quantified for each subclass versus all other subclasses (Fig.
380 4d), and glutamatergic neurons had more hypo-methylated DMRs compared to GABAergic neurons.
381 Methylome tracks at subclass level can be found at http://neomorph.salk.edu/aj2/pages/cross-species-
382 M1/. To identify enriched transcription factor binding sites (TFBS) in each species and subclass, we
383 performed motif enrichment analysis with hypo-methylated DMRs from one subclass against other
384 DMRs of the same species, and identified 102 ± 57 (mean ± s.d.) TFBS in each subclass (Extended
385 Data Fig. 8 and Supplementary Table 17). We repeated the enrichment analysis using TFBS motif
386 clusters [42] and found similarly distinct subclass signatures (Supplementary Table 18). Although
387 subclasses had unique marker genes (Fig. 2c, genes listed in Supplementary Table 8) and CH-DMG
388 across species, they had remarkably conserved TFBS motif enrichment (Fig. 4e,f and Extended Data
389 Fig. 8). For example, *TCF4* is robustly expressed in L5 IT neurons across species and shows
390 significant TFBS enrichment in hypo-methylated DMRs and AC sites. DMRs and AC sites provide
391 independent epigenomic information (Extended Data Fig. 7f,g) and can identify different TFBS
392 enrichment, such as for *ZNF148* in L5 IT neurons. These results are consistent with previous
393 observations of conserved TF network architectures in neural cell types between human and mouse
394 (Stergachis et al. 2014). Conserved sets of TFs have the potential to determine conserved and
395 divergent expression in consensus types based on shared or altered genomic locations of TFBS motifs
396 across species.
397
398 **Layer 4-like neurons in human M1**
399 M1 lacks a L4 defined by a thin band of densely packed “granular” neurons that is present in other
400 cortical areas, such as MTG (Fig. 5a), although prior studies have identified neurons with L4-like

401 synaptic properties in mice [14] and expression of *RORB*, a L4 marker, in non-human primate M1 [43]. To
402 address the potential existence of L4-like neurons in human M1 from a transcriptomic perspective, we
403 integrated snRNA-seq data from M1 and the granular MTG, where we previously described multiple L4
404 glutamatergic neuron types [2]. This alignment revealed a broadly conserved cellular architecture
405 between M1 and MTG (Fig. 5b,c, Extended Data Fig. 9) including M1 neuron types Exc L3 *RORB*
406 *OTOGL* (here, *OTOGL*) and Exc L3-5 *RORB LINC01202* (here, *LINC01202*) that map closely to MTG
407 neurons in deep L3 to L4 (Fig. 5c, red outlines). Interestingly, four MTG L2/3 IT types (*LTK*, *GLP2R*,
408 *FREM3*, and *CARM1P1*) whose distinct physiology and morphology are reported in a companion paper
409 [44] had less clear homology in M1 than other types (Extended Data Fig. 9a-c), pointing to more
410 variability across cortical areas of superficial as compared to deep glutamatergic neurons. To compare
411 laminar positioning in M1 and MTG, the relative cortical depth from pia for each neuron was estimated
412 based on the layer dissection and average layer thickness [45]. Transcriptomically similar cell types were
413 found at similar cortical depths in M1 and MTG, and the *OTOGL* and *LINC01202* types were located in
414 deep L3 and superficial L5 in M1 (Fig. 5d).
415
416 MTG contains three main transcriptomically-defined L4 glutamatergic neuron types (*FILIP1L*, *TWIST2*
417 and *ESR1*) and a deep L3 type (*COL22A1*) that is found on the border of L3 and L4 (Fig. 5e-g). The M1
418 types *OTOGL* and *LINC01202* matched one-to-one with MTG *COL22A1* and *ESR1*, whereas there
419 were no matches for the other two MTG L4 types (Fig. 5f). Based on snRNA-seq proportions, the L4-
420 like *OTOGL* type was much sparser in M1 than the *ESR1* type in MTG (Fig. 5e). Multiplex fluorescent in
421 situ hybridization (mFISH) with probes to cell type marker genes confirmed these findings. The MTG
422 *ESR1* type was highly enriched in L4, [2], and the homologous M1 *LINC01202* type was sparser and
423 more widely distributed across L3 and L5 (Fig. 5g). The MTG *COL22A1* type was tightly restricted to
424 the L3/4 border [2], and the M1 *OTOGL* type was similarly found at the L3/5 border. Quantification of
425 labeled cells as a fraction of DAPI+ cells in L3-5 showed similar frequencies of M1 *OTOGL* and MTG
426 *COL22A1* types and 4-fold sparser M1 *LINC01202* type versus MTG *ESR1* type (Fig. 5h). These data
427 indicate a conservation of deep L3 glutamatergic types and proportions across human cortical areas,

428 but with reduced diversity and sparsification of L4-like neurons to a single (ESR1) type in M1,
429 distributed more broadly where L4 would be if tightly aggregated.
430
431 **Chandelier cells share a core molecular identity across species**
432 Conserved transcriptomic and epigenomic features of consensus types likely contribute to cell function
433 and generate hypotheses about the gene regulatory mechanisms underlying cell type identity. Focused
434 analysis of *Pvalb*-expressing GABAergic neurons illustrates the power of these data to predict such
435 gene-function relationships. Cortical *Pvalb*-expressing neurons comprise two major types — basket
436 cells (BCs) and ChCs — that have fast-spiking electrical properties and distinctive cellular
437 morphologies. BCs selectively synapse onto the perisomatic region of glutamatergic pyramidal
438 neurons. ChCs, also called axo-axonic cells [46], selectively innervate the axon initial segment (AIS) of
439 pyramidal cells and have unique synaptic specializations called axon cartridges. These cartridges run
440 perpendicular to their post-synaptic target axon, giving a characteristic morphological appearance of
441 candlesticks on a chandelier. This highly conserved feature is shown with biocytin-filled cells from
442 mouse, rhesus macaque, and human (Fig. 6a). To reveal evolutionarily conserved transcriptomic
443 hallmarks of ChCs, we identified DEGs in ChCs versus BCs in each species using an ROC test. 357
444 DEGs were identified in at least one species, and marmoset ChCs shared more DEGs with human (61
445 genes) than mouse (29; Fig. 6b, Supplementary Table 19). Remarkably, only 25 DEGs were conserved
446 across all three species. One conserved gene, *UNC5B* (Fig. 6c), is a netrin receptor involved in axon
447 guidance and may help target ChC to pyramidal neuron AIS. Three transcription factors (*RORA*,
448 *TRPS1*, and *NFIB*) were conserved markers and may contribute to gene regulatory networks that
449 determine the unique attributes of ChCs.
450
451 To determine if ChCs had enriched epigenomic signatures for *RORA* and *NFIB* (*TRPS1* lacked motif
452 data), we compared DMRs between ChCs and BCs. In all species, *RORA* and *NFIB* had significant
453 CH-DMGs in ChCs not BCs (Fig. 6d), consistent with differential expression. To discern if these TFs
454 may preferentially bind to DNA in ChCs, we tested for TF motif enrichment in hypo-methylated (mCG)

455 DMRs and AC sites genome-wide. We found that the RORA motif was significantly enriched in DMRs
456 in primates (Fig. 6d) and in AC sites of ChCs in all species (Fig. 6e, Supplementary Table 14). The
457 NFIB binding motif was only significantly enriched in AC sites of mouse ChCs, possibly because
458 enrichment was transient during development or NFIB specificity is due to expression alone. Three
459 independent genomic assays converge to implicate *RORA* as a ChC-specific TF among *Pvalb*-
460 expression neurons. Notably, 60 of 357 DEGs contained a ROR-motif in DMRs and AC regions in at
461 least one species, further implicating *RORA* in defining ChC identity.
462
463 **Primate Betz cell specialization**
464 In mouse cortex, L5 glutamatergic neurons have distinct long-range projection targets (ET versus IT)
465 and transcriptomes [1]. L5 ET and IT neuron subclasses clearly align between human and mouse using
466 snRNA-seq in M1 (Extended Data Fig. 3) and in temporal [2] and fronto-insular cortex [12]. Betz cells in L5
467 of primate M1 connect to spinal motor-neurons via the pyramidal tracts and are predicted to be L5 ET
468 neurons. The species aligned transcriptomic types allow for the identification of genes whose
469 expression may contribute to conserved ET versus IT features and primate-specific physiology,
470 anatomy, and connectivity. Furthermore, Patch-seq methods that jointly measure the transcriptome,
471 physiological properties and morphology of cells, allow the direct identification and characterization of
472 L5 ET and IT neurons across mouse, non-human primate, and human. As primate physiology
473 experiments are largely restricted to macaque, we also profiled L5 of macaque M1 with snRNA-seq
474 (Cv3) to allow accurate Patch-seq mapping.
475
476 L5 ET neurons had many DEGs compared to L5 IT neurons in all 4 species. Approximately 50 DEGs
477 were conserved across all species and similarity to human varied as a function of evolutionary distance
478 (Fig. 7a, Supplementary Table 20). Several genes encoding ion channel subunits were enriched in ET
479 versus IT neurons in all species, potentially mediating conserved ET physiological properties (Fig. 7b).
480 A number of additional potassium and calcium channels were primate-enriched (Fig. 7c), potentially
481 underlying primate-specific ET or Betz cell physiology. Interestingly, many of these primate-specific ET-

482 enriched genes showed gradually increasing ET specificity in species more closely related to human.
483 To explore this idea of gradual evolutionary change further, we identified genes with increasing L5 ET
484 versus IT specificity as a function of evolutionary distance from human (Fig. 7d, Supplementary Table
485 21). Interestingly, this gene set was highly enriched for genes associated with axon guidance including
486 members of the Robo, Slit and Ephrin gene families. These genes are potential candidates for
487 regulating the cortico-motoneuronal connections associated with increasingly dexterous fine motor
488 control across these species [23].
489
490 To investigate if transcriptomically defined L5 ET types contain anatomically-defined Betz cells, FISH
491 for L5 ET neurons was combined with immunolabeling against SMI-32, a protein enriched in Betz cells
492 and other long-range projecting neurons in macaque [47–49] (Fig. 7e). Cells consistent with the size and
493 shape of Betz cells were identified in two L5 ET clusters (Exc L5 *FEZF2 ASGR2* and Exc L5 *FEZF2*
494 *CSN1S1*). Similar to previous reports on von Economo neurons in the insular cortex [12], ET clusters in
495 M1 also included neurons with non-Betz morphologies.
496
497 To facilitate cross-species comparisons of Betz cells and mouse ET neurons we made patch clamp
498 recordings from L5 neurons in acute and cultured slice preparations of mouse and macaque M1. For a
499 subset of recordings, Patch-seq analysis was applied for transcriptomic cell type identification
500 (Extended Data Fig. 10h). To permit visualization of cells in heavily myelinated macaque M1, we used
501 AAV viruses to drive fluorophore expression in glutamatergic neurons in macaque slice culture
502 (Extended Data Fig. 10g). As shown in Figure 7f, Patch-seq neurons mapping to the macaque Betz/ET
503 cluster (Exc L5 *FEZF2 LOC114676463*) had large somata (diameter > 65 μm) and long “tap root” basal
504 dendrites, canonical hallmarks of Betz cell morphology [17,50]. A unique opportunity to record from
505 neurosurgical tissue excised from human premotor cortex (near the confluence of the precentral and
506 superior frontal gyri) during an epilepsy treatment surgery using the same methods as for macaque
507 yielded multiple neurons that mapped transcriptomically to one of the Betz-containing cell types and
508 had canonical Betz cell morphology (Fig. 7g). Macaque and human ET neurons were grouped for

509 physiological analysis because intrinsic properties were not significantly different, and many
510 corticospinal axons originate from premotor cortex[23].
511
512 Shared transcriptomic profiles of mouse, primate, and human L5 ET neurons predicted conservation of
513 some physiological properties of rodent and primate neurons. Transcriptomically-defined ET neurons
514 across species expressed high levels of genes encoding an HCN channel-subunit and a regulatory
515 protein (*HCN1* and *PEX5L;* Fig. 7b). We hypothesized that HCN-dependent membrane properties,
516 which are used to distinguish rodent ET from IT neurons [51], would similarly separate cell types in
517 primates. Some primate L5 neurons possessed distinctive HCN-related properties such as a lower
518 input resistance (R_N) and a peak resonance (f_R) in voltage response around 3-9 Hz (Fig. 7h,i), similar to
519 rodent ET neurons. To determine whether HCN-related physiology is a conserved feature of L5
520 neurons, we grouped all neurons into physiologically defined ET and non-ET neurons based on their R_N
521 and f_R. We asked whether these physiologically-defined neurons corresponded to genetically-defined
522 ET/Betz or non-ET neurons using Patch-seq and cell-type specific mouse lines. For mouse M1, the ET-
523 specific *Thy1*-YFP[21,52] and IT specific *Etv1*-EGFP[53] mouse lines preferentially labeled physiologically
524 defined ET and non-ET neurons, respectively (Fig. 7j). For primates, transcriptomically-defined Betz
525 cells were physiologically defined ET neurons, whereas transcriptomically defined non-ET neurons
526 were physiologically defined non-ET neurons (Fig. 7k). Thus, there was broad correspondence
527 between physiologically-defined and genetically-defined ET neurons in both mouse and primate M1.
528 There were notable differences in physiology between mouse and primate ET neurons, however. A
529 greater fraction of primate ET neurons exhibited an exceptionally low R_N compared to mouse (Fig. 7l).
530 Additional differences in action potential properties across cell types and species may be explained in
531 part by differences in the expression of ion channel-related genes (Fig. 7c, Extended Data Fig. 10).
532
533 Most strikingly, primate Betz/ET neurons displayed a distinctive biphasic-firing pattern during long spike
534 trains. The firing rate of both primate and mouse non-ET neurons decreased to a steady state within
535 the first second of a 10 second depolarizing current injection, whereas the firing rate of mouse ET

536 neurons increased moderately over the same time period (Fig. 7m,n; Extended Data Fig. 10m,n). The
537 acceleration in rodent ET neurons has been attributed to the expression of Kv1-containing voltage-
538 gated K^+ channels that are encoded by genes like the conserved ET gene *KCNA1*. In macaque and
539 human ET/Betz neurons, a distinctive biphasic pattern was characterized by an early cessation of firing
540 followed by a sustained and dramatic increase in firing later in the current injection. Thus, while ET
541 neurons in both primate and rodent M1 displayed spike frequency acceleration, the temporal dynamics
542 and magnitude of this acceleration appears to be a unique feature of primate ET/Betz neurons. These
543 data emphasize how transcriptomic data from this specialized neuron type can be linked to shared and
544 unique physiological properties across species.
545

546 Discussion

547 Comparative analysis is a powerful strategy to understand brain structure and function. Species
548 conservation is strong evidence for functional relevance under evolutionary constraints that can help
549 identify critical molecular and regulatory mechanisms [54,55]. Conversely, divergence indicates adaptive
550 specialization, which may be essential to understand the mechanistic underpinnings of human brain
551 function and susceptibility to human-specific diseases. In the current study, we applied a comparative
552 approach to understand conserved and species-specific features of M1 at the level of cell types using
553 single-nucleus RNA-seq (Cv3 and SSv4), open chromatin (SNARE-seq2 and ATAC-seq) and DNA-
554 methylation (snmC-seq2) technologies. Integrated analysis of over 450,000 nuclei in human, non-
555 human primates (marmoset, a New World monkey, and to a lesser degree macaque, an Old World
556 monkey that is evolutionarily more closely related to humans), and mouse (see also companion paper
557 [6]) yielded a high-resolution, multimodal classification of cell types in each species, and a coarser
558 consensus classification conserved between rodent and primate lineages. Robust species conservation
559 strongly argues for the functional relevance of this consensus cellular architecture. Species
560 specializations are also apparent, both in the additional granularity in cell types within species and
561 differences between conserved cell types. A comparative evolutionary approach provides an anchor
562 point to define the cellular architecture of any tissue and to discover species-specific adaptations.

563

564 A key result of the current study is the identification of a consensus classification of cell types across
565 species that allows the comparison of relative similarities in human compared to common mammalian
566 model organisms in biomedical research. Prior studies have demonstrated that high resolution cellular
567 taxonomies can be generated in mouse, non-human primate and human cortex, and that there is
568 generally good concordance across species [2,11]. However, inconsistencies in the methods and
569 sampling depths used made strong conclusions difficult, compounded by the analysis of different
570 cortical regions in different species. The current study overcame these challenges by focusing on M1, a
571 functionally and anatomically conserved cortical region across mammals, and comparing a variety of
572 methods on similarly isolated tissues (and the same specimens from human and marmoset). Several
573 important points emerged from these integrated analyses. First, with deeper sampling and the same
574 methodology (snRNA-seq with Cv3), a similar cellular complexity on the order of 100 cell types was
575 seen in all three species. The highest resolution molecular classification was seen with RNA-seq
576 compared to epigenomic methods, and among RNA-seq methods with those that allow the most cells
577 to be analyzed. Strikingly, the molecular classifications were well aligned across all methods tested,
578 albeit at different levels of resolution as a function of the information content of the assay and the
579 number of cells profiled. All methods were consistent at the level of subclasses as defined above, both
580 across methods and species; significantly better alignment was achieved among species based on
581 transcriptomics, and with epigenomic methods in some subclasses. Mismatches in cellular sampling
582 affect the ability to compare across species; for example, higher non-neuronal sampling in mouse and
583 marmoset increased detection of rare cell types compared to human. One important comparison was
584 between plate-based (SSv4) and droplet-based (Cv3) RNA-seq of human nuclei, where we compared
585 results between approximately 10,000 SSv4 and 100,000 Cv3 nuclei. On average, SSv4 detected 30%
586 more genes per nucleus and enabled comparisons of isoform usage between cell types, albeit with 20-
587 fold greater sequencing depth. However, SSv4 cost 10 times as much as Cv3 and did not allow
588 detection of additional cell types.

589

590 The snmC-seq2 clustering aligned closely with the transcriptomic classification, although with
591 significantly lower resolution in rarer subclasses. Hypo-methylated sites correlated with gene
592 expression and specific transcription factor binding motifs were enriched in cell type specific sites.
593 Multi-omic SNARE-seq2 measured RNA profiles of nuclei that allowed high confidence assignment to
594 transcriptomic clusters. Examining accessible chromatin (AC) regions within the same nuclei led to
595 strong correlations between cell subclass or type gene expression and active regulatory regions of
596 open chromatin. Using this strategy, gene regulatory activities could be identified within RNA-defined
597 cell populations (including RNA consensus clusters) that could not be resolved from AC data alone
598 (Extended Data Fig. 6a, Supplementary Table 15). By joint consideration of these epigenomic
599 modalities, glutamatergic neurons were found to have more hypo-methylated DMRs and differentially
600 accessible chromatin, consistent with having larger somata and expressing more genes. Within-
601 species, cell types have many more unique AC sites than uniquely expressed marker genes. At the
602 same time, there is striking conservation across species of subclass TFBS motif enrichment within AC
603 and hypo-methylated DMRs. Most subclasses have distinct motifs, although L2/3 and L6 IT and *Lamp5*
604 and *Sncg* subclasses share many motifs and are more clearly distinguished based on gene expression.
605 Taken together, these results show a robust cell type classification that is consistent at the level of
606 subclasses both across transcriptomic and chromatin measures and across species, with additional cell
607 type-level granularity identified with transcriptomics.
608
609 Alignment across species allowed a comparison of relative similarities and differences between
610 species. A common (and expected) theme was that more closely related species are more similar to
611 one another. This was true at the level of gene expression and epigenome patterning across cell types,
612 and in the precision with which transcriptomically-defined cell types could be aligned across species.
613 For example, human and marmoset GABAergic types could be aligned at higher resolution than human
614 and mouse. Human was more similar to macaque than to marmoset. This indicates that cell type
615 similarity increases as a function of evolutionary distance to our closest common ancestors with mouse
616 (~70 mya), marmoset (~40 mya), and macaque (~25 mya). Interestingly, many gene expression

differences may change gradually over evolution. This is apparent in the graded changes in expression levels of genes enriched in L5 ET versus L5 IT neurons and in the reduced performance of cell type classification based on marker gene expression that is correlated with evolutionary distance between species.

Several prominent species differences in cell type proportions were observed. First, the ratio of glutamatergic excitatory projection neurons compared to GABAergic inhibitory interneurons was 2:1 in human compared to 3:1 in marmoset and 5:1 in mouse and leads to a profound shift in the overall excitation-inhibition balance of the cortex. A similar species difference has been described based on histological measures (reviewed in [28]), indicating that snRNA-seq gives a reasonably accurate measurement of cell type proportions. Surprisingly, the relative proportions of GABAergic subclasses and types were similar across species. These results suggest a developmental shift in the size of the GABAergic progenitor pool in the ganglionic eminences or an extended period of neurogenesis and migration. A decreased proportion of the subcortically targeting L5 ET neurons in human was also seen, as previously shown in temporal [2] and frontoinsular [12] cortex. This shift likely reflects the evolutionary increase in cortical neurons relative to their subcortical targets [56] and was less prominent in M1, suggesting regional variation in the proportion of L5 ET neurons. Finally, a large increase in the proportion of L2 and L3 IT neurons was seen in human compared to mouse and marmoset. This increase parallels the disproportionate expansion of human cortical area and supragranular layers that contain neurons projecting to other parts of the cortex, presumably to facilitate greater corticocortical communication. Interestingly, L2 and L3 IT neurons appear to be particularly highly variable across cortical areas and species, and also are more diverse and specialized in human compared to mouse (see companion paper [44]).

A striking and somewhat paradoxical observation is the high degree of species specialization of consensus types. The majority of DEGs between cell types were consistently species-specific. This result suggests that the conserved cellular features of a cell type are largely due to a minority of DEGs

644 with conserved expression patterns. The current study demonstrates this point for one of the most
645 distinctive brain cell types, the cortical *Pvalb*-expressing GABAergic ChC. ChCs in mouse, non-human
646 primate, and human have 100-150 genes with highly enriched expression compared to other *Pvalb*-
647 expressing interneurons (BCs); however, only 25 of these ChC-enriched genes are shared across
648 species. This small overlapping gene set includes several transcription factors and a member of the
649 netrin family (*UNC5B*) that could be responsible for AIS targeting. Binding sites for these TFs are
650 enriched in ChC cluster regions of open chromatin and in hypo-methylated regions around ChC-
651 enriched genes. While these associations between genes and cellular phenotypes for conserved and
652 divergent features remain to be tested, a comparative strategy can identify these core conserved genes
653 and make strong predictions about the TF code for cell types and the genes responsible for their
654 evolutionarily constrained functions.
655
656 M1 is an agranular cortex lacking a L4, although a recent study demonstrated that there are neurons
657 with L4-like properties in mouse [14]. Here we confirm and extend this finding in human M1. We find a L4-
658 like neuron type in M1 that aligns to a L4 type in human MTG and is scattered between the deep part of
659 L3 and the superficial part of L5 where L4 would be if aggregated into a layer. However, MTG
660 contained several additional L4 types not found in M1, and with a much higher frequency. The human
661 M1 L4-like type is part of the L5 IT_1 consensus cluster that includes several IT types in all species,
662 including two L4-like types in mouse (L4/5 IT_1 and L4/5 IT_2) that also express the canonical L4
663 marker *Rorb* (see companion paper [6]). Therefore, it appears that M1 has L4-like cells from a
664 transcriptomic perspective, but only a subset of the types compared to granular cortical areas, at much
665 lower density, and scattered rather than aggregated into a tight layer.
666
667 The most distinctive cellular hallmark of M1 in primates and cats is the enormous Betz cell, which
668 contributes to direct corticospinal connections to spinal motoneurons in primates that participate in fine
669 motor control [15,16,57–59]. Intracellular recordings from cats have shown highly distinctive characteristics
670 including HCN channel-related membrane properties, spike frequency acceleration, and extremely fast

671 maintained firing rates [19,20]. However, they have never been recorded in primates using patch clamp
672 physiology due to the high degree of myelination in M1 that prevents their visualization, and the inability
673 to obtain motor cortex tissue from neurosurgical procedures which are careful to be function-sparing. A
674 goal of the current project was to identify the transcriptomic cluster corresponding to Betz cells and use
675 this to understand gene expression that may underlie their distinctive properties and species
676 specializations. We have recently taken a similar approach to study von Economo neurons in the
677 fronto-insular cortex, showing they are found within a transcriptomic class consisting of ET neurons [12].
678 Betz cells are classical ET neurons that, together with the axons of smaller corticospinal neurons, make
679 up part of the pyramidal tract from the cortex to the spinal cord [16,60]. We show that neurons with Betz
680 cell morphology label with markers for the M1 ET clusters. Like von Economo neurons, there does not
681 appear to be an exclusively Betz transcriptomic type. Rather, M1 ET clusters are not exclusive for
682 neurons with Betz morphology, and we find more than one ET cluster contains neurons with Betz
683 morphology.
684
685 Although comparative transcriptomic alignments provide strong evidence for functional similarity, the
686 distinctions between corticospinal neurons across species or even between L5 ET and IT neuron types
687 in primates or humans has not been demonstrated physiologically. We recently developed a suite of
688 methodologies for studying specific neuron types in human and non-human tissues, including triple
689 modality Patch-seq to combine physiology, morphology and transcriptome analysis, acute and cultured
690 slice physiology in adult human neurosurgical resections and macaque brain, and AAV-based neuronal
691 labeling to allow targeting of neurons in highly myelinated tissues (companion paper [44]; [61] . Specifically,
692 these tools allow the targeting of L5 neurons in mouse and non-human primate and the assignment of
693 neurons to their transcriptomic types using Patch-seq, which we facilitated by generating and aligning a
694 L5 transcriptomic classification in macaque where such analyses could be performed. We show here
695 that several of the characteristic features of L5 ET versus IT neurons are conserved, and can be
696 reliably resolved from one another in mouse and non-human primate. Furthermore, macaque neurons
697 with Betz-like morphologies mapped to the Betz-containing clusters. However, as predicted by

differences in ion channel-related gene expression, not all physiological features were conserved between macaque and mouse ET neurons. Betz/ET neurons had the distinctive pauses, bursting and spike-frequency acceleration described previously in cats but not seen in rodents [19,20]. Finally, we had access to an extremely rare human neurosurgical case where a region of premotor cortex was resected. Similar to macaque M1, this premotor region contained large neurons with characteristic Betz-like morphology that mapped transcriptomically to the Betz-containing clusters. Together these results highlight the predictive power of transcriptomic mapping and cross-species inference of cell types for L5 pyramidal neurons including the Betz cells. Furthermore, these data are consistent with observations that Betz cells may not in fact be completely restricted to M1 but distribute across other proximal motor-related areas that contribute to the pyramidal tract [62]. Finally, a number of ion channels that may contribute to conserved ET versus IT features as well as species specializations of Betz cell function were identified that provide candidate genes to explore gene-function relationships. For example, axon guidance-associated genes are enriched in Betz-containing ET neuron types in primates, possibly explaining why Betz cells in primates directly contact spinal motor neurons rather than spinal interneurons as in rodents. Thus, as the comparative approach is helpful in identifying core conserved molecular programs, it may be equally valuable to understand what is different in human or can be well modeled in closer non-human primate relatives. This is particularly relevant in the context of Betz cells and other ET neuron types that are selectively vulnerable in amyotrophic lateral sclerosis, some forms of frontotemporal dementia, and other neurodegenerative conditions.

718 ## References

719 1. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**,
720 72–78 (2018).
721 2. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex.
722 *Nature* **573**, 61–68 (2019).
723 3. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat.*
724 *Neurosci.* **19**, 335–346 (2016).
725 4. Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of
726 the human brain. *Science* **352**, 1586–1590 (2016).
727 5. Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in
728 mammalian cortex. *Science* **357**, 600–604 (2017).
729 6. Yao, Z. *et al.* An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell
730 types. *bioRxiv* 2020.02.29.970558 (2020) doi:10.1101/2020.02.29.970558.
731 7. Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the
732 human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
733 8. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by
734 single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
735 9. Gray, L. T. *et al.* Layer-specific chromatin accessibility landscapes reveal regulatory networks in
736 adult mouse visual cortex. *Elife* **6**, (2017).
737 10. Lee, D.-S. *et al.* Simultaneous profiling of 3D genome structure and DNA methylation in single
738 human cells. *Nat. Methods* **16**, 999–1006 (2019).
739 11. Krienen, F. M., Goldman, M., Zhang, Q. & del Rosario, R. Innovations in primate interneuron
740 repertoire. *bioRxiv* (2019).
741 12. Hodge, R. D. *et al.* Transcriptomic evidence that von Economo neurons are regionally specialized
742 extratelencephalic-projecting excitatory neurons. *Nat. Commun.* **11**, 1172 (2020).
743 13. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem

744 Cells. *Cell* **167**, 566–580.e19 (2016).
745 14. Yamawaki, N., Borges, K., Suter, B. A., Harris, K. D. & Shepherd, G. M. G. A genuine layer 4 in
746 motor cortex with prototypical synaptic circuit connectivity. *Elife* **3**, e05422 (2014).
747 15. Betz, W. Anatomischer Nachweis zweier Gehirncentra. *Zentralbl Med Wiss* **12**, (1874).
748 16. Lassek, A. M. The Human Pyramidal Tract II. A Numerical Investigation of the Betz Cells of the
749 Motor Area. *J. Nerv. Ment. Dis.* **94**, 225–226 (1941).
750 17. Jacobs, B. *et al.* Comparative morphology of gigantopyramidal neurons in primary motor cortex
751 across mammals. *J. Comp. Neurol.* **526**, 496–536 (2018).
752 18. Kaiserman-Abramof, I. R. & Peters, A. Some aspects of the morphology of Betz cells in the
753 cerebral cortex of the cat. *Brain Res.* **43**, 527–546 (1972).
754 19. Spain, W. J., Schwindt, P. C. & Crill, W. E. Post-inhibitory excitation and inhibition in layer V
755 pyramidal neurones from cat sensorimotor cortex. *The Journal of Physiology* vol. 434 609–626
756 (1991).
757 20. Chen, W., Zhang, J. J., Hu, G. Y. & Wu, C. P. Electrophysiological and morphological properties of
758 pyramidal and nonpyramidal neurons in the cat motor cortex in vitro. *Neuroscience* **73**, 39–55
759 (1996).
760 21. Miller, M. N., Okaty, B. W. & Nelson, S. B. Region-Specific Spike-Frequency Acceleration in Layer
761 5 Pyramidal Neurons Mediated by Kv1 Subunits. *Journal of Neuroscience* vol. 28 13716–13726
762 (2008).
763 22. Gu, Z. *et al.* Control of species-dependent cortico-motoneuronal connections underlying manual
764 dexterity. *Science* **357**, 400–404 (2017).
765 23. Lemon, R. N. Descending pathways in motor control. *Annu. Rev. Neurosci.* **31**, 195–218 (2008).
766 24. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
767 25. Oberheim, N. A. *et al.* Uniquely Hominid Features of Adult Human Astrocytes. *Journal of*
768 *Neuroscience* vol. 29 3276–3287 (2009).
769 26. Colombo, J. A. The interlaminar glia: from serendipity to hypothesis. *Brain Struct. Funct.* **222**,
770 1109–1129 (2017).

771 27. Bakken, T. E. *et al.* Single-nucleus and single-cell transcriptomes compared in matched cortical
772 cell types. *PLoS One* **13**, e0209648 (2018).
773 28. Džaja, D., Hladnik, A., Bičanić, I., Baković, M. & Petanjek, Z. Neocortical calretinin neurons in
774 primates: increase in proportion and microcircuitry structure. *Front. Neuroanat.* **8**, 103 (2014).
775 29. DeFelipe, J., Alonso-Nanclares, L. & Arellano, J. I. Microstructure of the neocortex: comparative
776 aspects. *J. Neurocytol.* **31**, 299–316 (2002).
777 30. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types
778 defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9**, 884 (2018).
779 31. Boldog, E. *et al.* Transcriptomic and morphophysiological evidence for a specialized human cortical
780 GABAergic cell type. *Nat. Neurosci.* **21**, 1185–1195 (2018).
781 32. Paul, A. *et al.* Transcriptional Architecture of Synaptic Communication Delineates GABAergic
782 Neuron Identity. *Cell* **171**, 522–539.e20 (2017).
783 33. Verkhratsky, A. & Nedergaard, M. The homeostatic astroglia emerges from evolutionary
784 specialization of neural cells. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **371**, (2016).
785 34. Nortley, R. *et al.* Amyloid β oligomers constrict human capillaries in Alzheimer's disease via
786 signaling to pericytes. *Science* vol. 365 eaav9518 (2019).
787 35. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
788 36. Riccomagno, M. M. *et al.* The RacGAP β2-Chimaerin selectively mediates axonal pruning in the
789 hippocampus. *Cell* **149**, 1594–1606 (2012).
790 37. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin
791 accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
792 38. Nongluk Plongthongkum, Dinh Diep, Song Chen, Blue B. Lake, Kun Zhang. Scalable Dual-omic
793 Profiling with Single-nucleus Chromatin Accessibility and mRNA Expression Sequencing 2
794 (SNARE-Seq2). (2020).
795 39. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin
796 Accessibility Data. *Mol. Cell* **71**, 858–871.e8 (2018).
797 40. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-

798 factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
799 41. Luo, C. *et al.* Robust single-cell DNA methylome profiling with snmC-seq2. *Nat. Commun.* **9**, 3824
800 (2018).
801 42. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding
802 profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).
803 43. Bernard, A. *et al.* Transcriptional architecture of the primate neocortex. *Neuron* **73**, 1083–1099
804 (2012).
805 44. Berg, J. *et al.* Human cortical expansion involves diversification and specialization of supragranular
806 intratelencephalic-projecting neurons. (2020).
807 45. von Economo, C. & Koskinas, G. N. *Die Cytoarchitektonik der Hirnrinde des Erwachsenen*
808 *Menschen*. (J. Springer, 1925).
809 46. Somogyi, P., Freund, T. F. & Cowey, A. The axo-axonic interneuron in the cerebral cortex of the
810 rat, cat and monkey. *Neuroscience* **7**, 2577–2607 (1982).
811 47. Hof, P. R., Nimchinsky, E. A. & Morrison, J. H. Neurochemical phenotype of corticocortical
812 connections in the macaque monkey: quantitative analysis of a subset of neurofilament protein-
813 immunoreactive projection neurons in frontal, parietal, temporal, and cingulate cortices. *J. Comp.*
814 *Neurol.* **362**, 109–133 (1995).
815 48. Tsang, Y. M., Chiong, F., Kuznetsov, D., Kasarskis, E. & Geula, C. Motor neurons are rich in non-
816 phosphorylated neurofilaments: cross-species comparison and alterations in ALS. *Brain Res.* **861**,
817 45–58 (2000).
818 49. Preuss, T. M., Stepniewska, I., Jain, N. & Kaas, J. H. Multiple divisions of macaque precentral
819 motor cortex identified with neurofilament antibody SMI-32. *Brain Research* vol. 767 148–153
820 (1997).
821 50. Scheibel, M. E., Davies, T. L., Lindsay, R. D. & Scheibel, A. B. Basilar dendrite bundles of giant
822 pyramidal cells. *Exp. Neurol.* **42**, 307–319 (1974).
823 51. Baker, A. *et al.* Specialized Subpopulations of Deep-Layer Pyramidal Neurons in the Neocortex:
824 Bridging Cellular Properties to Functional Consequences. *J. Neurosci.* **38**, 5441–5455 (2018).

825 52. Feng, G. *et al.* Imaging neuronal subsets in transgenic mice expressing multiple spectral variants
826 of GFP. *Neuron* **28**, 41–51 (2000).
827 53. Groh, A. *et al.* Cell-Type Specific Properties of Pyramidal Neurons in Neocortex Underlying a
828 Layout that Is Modifiable Depending on the Cortical Area. *Cerebral Cortex* vol. 20 826–836 (2010).
829 54. Tosches, M. A. *et al.* Evolution of pallium, hippocampus, and cortical cell types revealed by single-
830 cell transcriptomics in reptiles. *Science* **360**, 881–888 (2018).
831 55. Arendt, D. *et al.* The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 (2016).
832 56. Herculano-Houzel, S., Catania, K., Manger, P. R. & Kaas, J. H. Mammalian Brains Are Made of
833 These: A Dataset of the Numbers and Densities of Neuronal and Nonneuronal Cells in the Brain of
834 Glires, Primates, Scandentia, Eulipotyphlans, Afrotherians and Artiodactyls, and Their Relationship
835 with Body Mass. *Brain Behav. Evol.* **86**, 145–163 (2015).
836 57. Rivara, C.-B., Sherwood, C. C., Bouras, C. & Hof, P. R. Stereologic characterization and spatial
837 distribution patterns of Betz cells in the human primary motor cortex. *Anat. Rec. A Discov. Mol.*
838 *Cell. Evol. Biol.* **270**, 137–151 (2003).
839 58. Evarts, E. V. Representation of movements and muscles by pyramidal tract neurons of the
840 precentral motor cortex. in *Neurophysiological basis of normal and abnormal motor activities* 215–
841 253 (Raven Press New York, 1967).
842 59. Evarts, E. V. RELATION OF DISCHARGE FREQUENCY TO CONDUCTION VELOCITY IN
843 PYRAMIDAL TRACT NEURONS. *J. Neurophysiol.* **28**, 216–228 (1965).
844 60. Lassek, A. M. THE PYRAMIDAL TRACT: A STUDY OF RETROGRADE DEGENERATION IN THE
845 MONKEY. *Arch NeurPsych* **48**, 561–567 (1942).
846 61. Ting, J. T. *et al.* A robust ex vivo experimental platform for molecular-genetic dissection of adult
847 human neocortical cell types and circuits. *Sci. Rep.* **8**, 8407 (2018).
848 62. Vigneswaran, G., Kraskov, A. & Lemon, R. N. Large Identified Pyramidal Cells in Macaque Motor
849 and Premotor Cortex Exhibit 'Thin Spikes': Implications for Cell Type Classification. *Journal of*
850 *Neuroscience* vol. 31 14235–14242 (2011).
851 63. Krimer, L. S. *et al.* Cluster Analysis–Based Physiological Classification and Morphological

852 Properties of Inhibitory Neurons in Layers 2–3 of Monkey Dorsolateral Prefrontal Cortex. *J.*
853 *Neurophysiol.* **94**, 3009–3022 (2005).
854 64. Rotaru, D. C. *et al.* Functional properties of GABA synaptic inputs onto GABA neurons in monkey
855 prefrontal cortex. *J. Neurophysiol.* **113**, 1850–1861 (2015).
856 65. Fetz, E. E., Cheney, P. D., Mewes, K. & Palmer, S. Control of forelimb muscle activity by
857 populations of corticomotoneuronal and rubromotoneuronal cells. *Prog. Brain Res.* **80**, 437–49;
858 discussion 427–30 (1989).
859 66. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more
860 genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic*
861 *Acids Res.* **47**, D419–D426 (2019).
862 67. Koopmans, F. *et al.* SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the
863 Synapse. *Neuron* **103**, 217–234.e4 (2019).
864 68. Luo, C. *et al.* Single nucleus multi-omics links human cortical cell regulatory genome diversity to
865 disease risk variants. *bioRxiv* 2019.12.11.873398 (2019) doi:10.1101/2019.12.11.873398.
866 69. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic
867 Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).
868 70. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data
869 analysis. *Genome Biol.* **19**, 15 (2018).
870 71. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected
871 communities. *Sci. Rep.* **9**, 5233 (2019).
872 72. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using
873 Support Vector Machines. *Mach. Learn.* **46**, 389–422 (2002).
874 73. Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The Balanced Accuracy and Its
875 Posterior Distribution. in *2010 20th International Conference on Pattern Recognition* 3121–3124
876 (2010).
877 74. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A Python Toolbox to Tackle the
878 Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **18**, 1–5 (2017).

879 75. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes
880 using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
881 76. He, Y. *et al.* Spatiotemporal DNA Methylome Dynamics of the Developing Mammalian Fetus.
882 doi:10.1101/166744.
883 77. McLeay, R. C. & Bailey, T. L. Motif Enrichment Analysis: a unified framework and an evaluation on
884 ChIP data. *BMC Bioinformatics* **11**, 165 (2010).
885 78. Palmer, C., Liu, C. & Chun, J. Nuclei Isolation for SNARE-seq2 v1 (protocols.io.8tvhwn6).
886 doi:10.17504/protocols.io.8tvhwn6.
887 79. Lun, A. T. L. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-
888 cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
889 80. Gayoso, A. & Shor, J. *GitHub: DoubletDetection*. (2019). doi:10.5281/zenodo.2678042.
890 81. Hahne, F. & Ivanek, R. Visualizing Genomic Data Using Gviz and Bioconductor. *Methods Mol. Biol.*
891 **1418**, 335–351 (2016).
892 82. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation.
893 *Nature* **523**, 486–490 (2015).
894 83. Cusanovich, D. A. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial
895 cellular indexing. *Science* vol. 348 910–914 (2015).
896 84. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**,
897 355–364 (2014).
898 85. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that
899 Correlate with Prognosis. *Cell* **162**, 184–197 (2015).
900 86. Graybuck, L. T. *et al.* Prospective, brain-wide labeling of neuronal subclasses with enhancer-driven
901 AAVs. (2019) doi:10.1101/525014.
902 87. Ting, J. T., Daigle, T. L., Chen, Q. & Feng, G. Acute brain slice methods for adult and aging
903 animals: application of targeted patch clamp analysis and optogenetics. *Methods Mol. Biol.* **1183**,
904 221–242 (2014).
905 88. Chan, K. Y. *et al.* Engineered AAVs for efficient noninvasive gene delivery to the central and

906 peripheral nervous systems. *Nat. Neurosci.* **20**, 1172–1179 (2017).

907 89. Rudy, B. & McBain, C. J. Kv3 channels: voltage-gated K+ channels designed for high-frequency

908 repetitive firing. *Trends Neurosci.* **24**, 517–526 (2001).

909

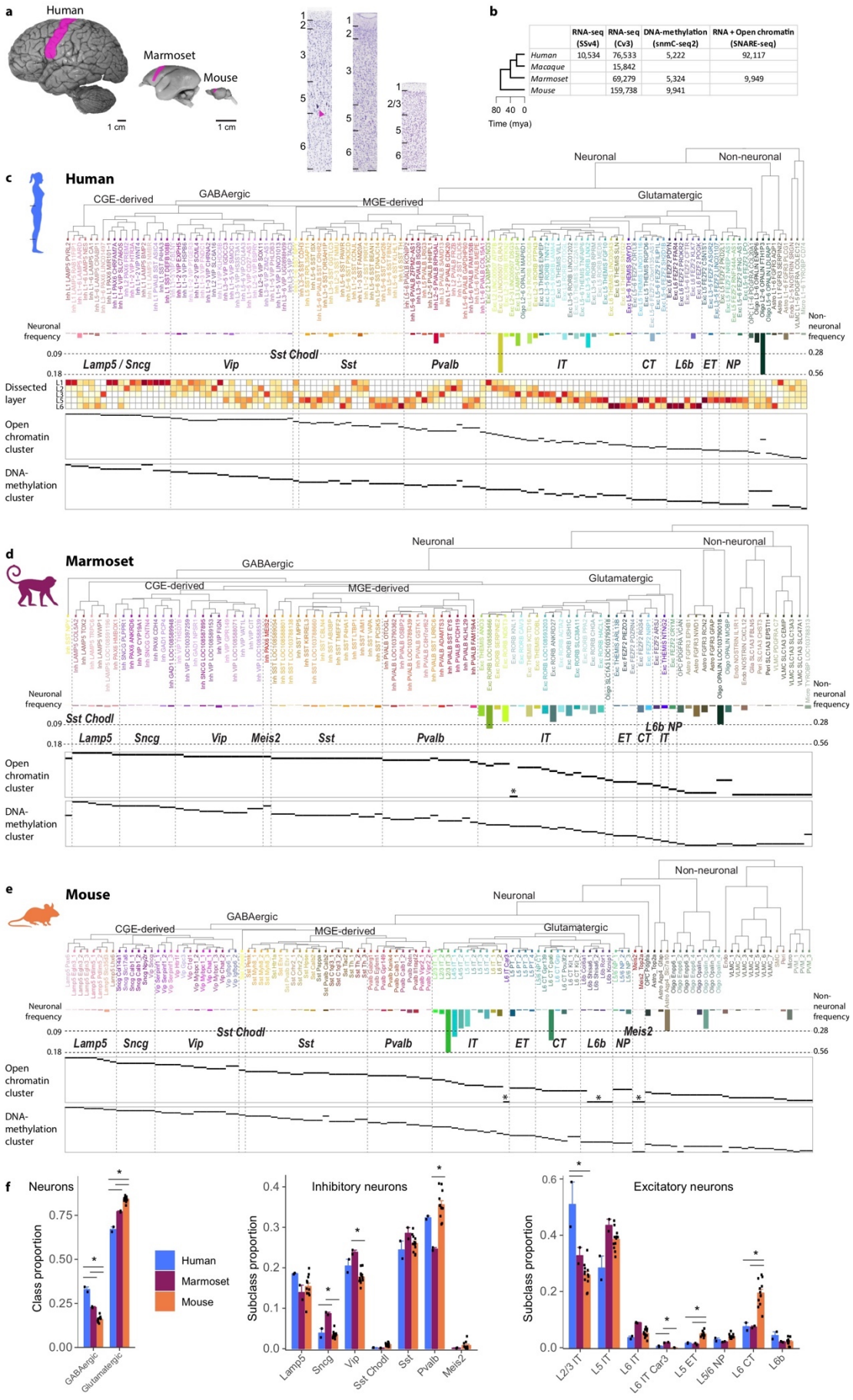

a
Human
Marmoset
Mouse
1 cm
1 cm
b
RNA-seq (SSv4)
RNA-seq (Cv3)
DNA-methylation (snmC-seq2)
RNA + Open chromatin (SNARE-seq)
Human 10,534 76,533 5,222 92,117
Macaque 15,842
Marmoset 69,279 5,324 9,949
Mouse 159,738 9,941
80 40 0
Time (mya)
c
Human
Neuronal
Non-neuronal
GABAergic
Glutamatergic
CGE-derived
MGE-derived
Neuronal frequency
Non-neuronal frequency
Lamp5 / Sncg
Vip
Sst Chodl
Sst
Pvalb
IT
CT
L6b
ET
NP
Dissected layer
Open chromatin cluster
DNA-methylation cluster
d
Marmoset
Neuronal
Non-neuronal
GABAergic
Glutamatergic
CGE-derived
MGE-derived
Neuronal frequency
Non-neuronal frequency
Sst Chodl
Lamp5
Sncg
Vip
Meis2
Sst
Pvalb
IT
ET
CT
L6b
NP
Open chromatin cluster
DNA-methylation cluster
e
Mouse
Non-neuronal
Neuronal
GABAergic
Glutamatergic
CGE-derived
MGE-derived
Neuronal frequency
Non-neuronal frequency
Sst Chodl
Meis2
Lamp5
Sncg
Vip
Sst
Pvalb
IT
ET
CT
L6b
NP
Open chromatin cluster
DNA-methylation cluster
f
Neurons
Class proportion
Human
Marmoset
Mouse
GABAergic
Glutamatergic
Inhibitory neurons
Subclass proportion
Lamp5
Sncg
Vip
Sst Chodl
Sst
Pvalb
Meis2
Excitatory neurons
Subclass proportion
L2/3 IT
L5 IT
L6 IT
L6 IT Car3
L5 ET
L5/6 NP
L6 CT
L6b

Figure 1. Molecular taxonomy of cell types in M1 of human, marmoset, and mouse. a, M1 highlighted in lateral views of neocortex across species. Nissl-stained sections of M1 annotated with layers and showing the relative expansion of cortical thickness, particularly L2 and L3 in primates, and large pyramidal neurons or 'Betz' cells in human L5 (arrowhead). Scale bars, 100 μm. **b**, Phylogeny of species and number of nuclei included in analysis for each molecular assay. All assays used nuclei isolated from the same donors for human and marmoset. SSv4, SMART-Seq v4; Cv3, Chromium v3; mya, millions of years ago. **c-e**, Dendrograms of cell types defined by RNA-seq (Cv3) for human (**c**), marmoset (**d**), and mouse (**e**) and annotated with cluster frequency and dissected layer (human only). Epigenomic clusters (in rows) aligned to RNA-seq clusters as indicated by horizontal black bars. Asterisks denote RNA clusters that lack corresponding epigenomic clusters. **f**, Relative proportions of cells in several classes and subclasses were significantly different between species based on an ANOVA followed by Tukey's HSD tests (asterisk, adjusted $P < 0.05$).

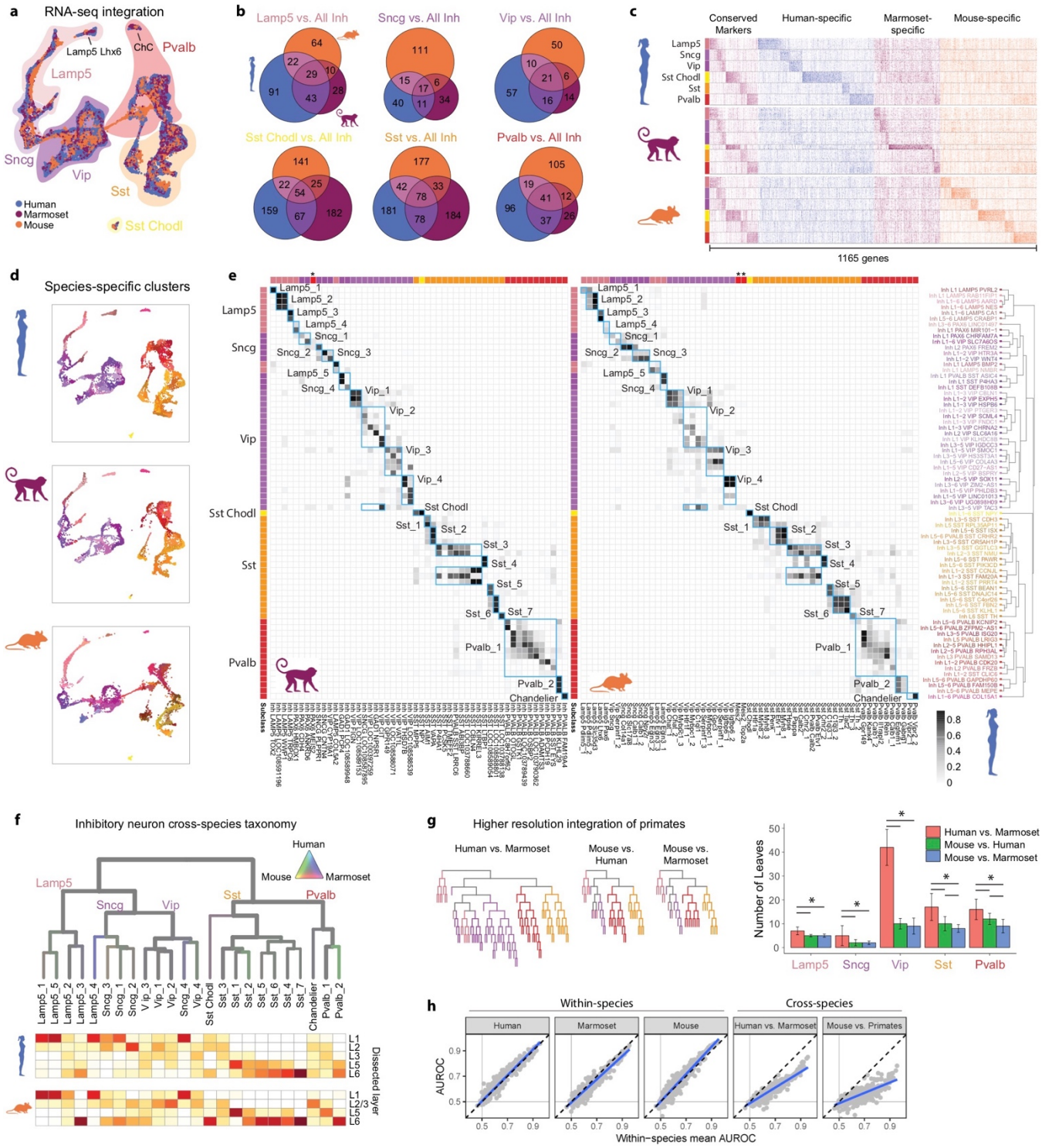


Figure 2. Evolution of GABAergic neuron types across species. a, UMAP projection of integrated snRNA-seq data from human, marmoset, and mouse GABAergic neurons. Filled outlines indicate cell subclasses. **b,** Venn diagrams indicating the number of shared DEGs across species by subclass. DEGs were determined by ROC tests of each subclass versus all other GABAergic subclasses within a

species. **c,** Heatmap of all DEGs from **b** ordered by subclass and species enrichment. Heatmap shows gene expression scaled by column for up to 50 randomly sampled nuclei from each subclass for each species. **d,** UMAP projection from **a**, separated by species, and colored by within-species clusters. **e,** Cluster overlap heatmap showing the proportion of nuclei in each pair of species clusters that are mixed in the cross-species integrated space. Cross-species consensus clusters are indicated by labeled blue boxes. Human clusters (rows) are ordered by the dendrogram reproduced from **Figure 1c**. Marmoset (left columns) and mouse (right columns) clusters are ordered to align with human clusters. Color bars at top and left indicate subclasses of within-species clusters. Asterisks indicate marmoset and mouse Meis2 subclasses, which were not present in human. **f,** Dendrogram of GABAergic neuron consensus clusters with edges colored by species mixture (grey, well mixed). Below: Estimated spatial distributions of clusters based on layer dissections in human (top) and mouse (bottom). **g,** Dendrograms of pairwise species integrations, colored by subclass. Bar plots quantifying well-mixed leaf nodes. Significant differences (adjusted $P < 0.05$, Tukey's HSD test) between species are indicated for each subclass. **h,** Scatter plots of MetaNeighbor analysis showing the performance (AUROC) of gene-sets to classify GABAergic neurons within and between species. Blue lines, linear regression fits; black lines, mean within species performance; grey lines, performance equivalent to chance.

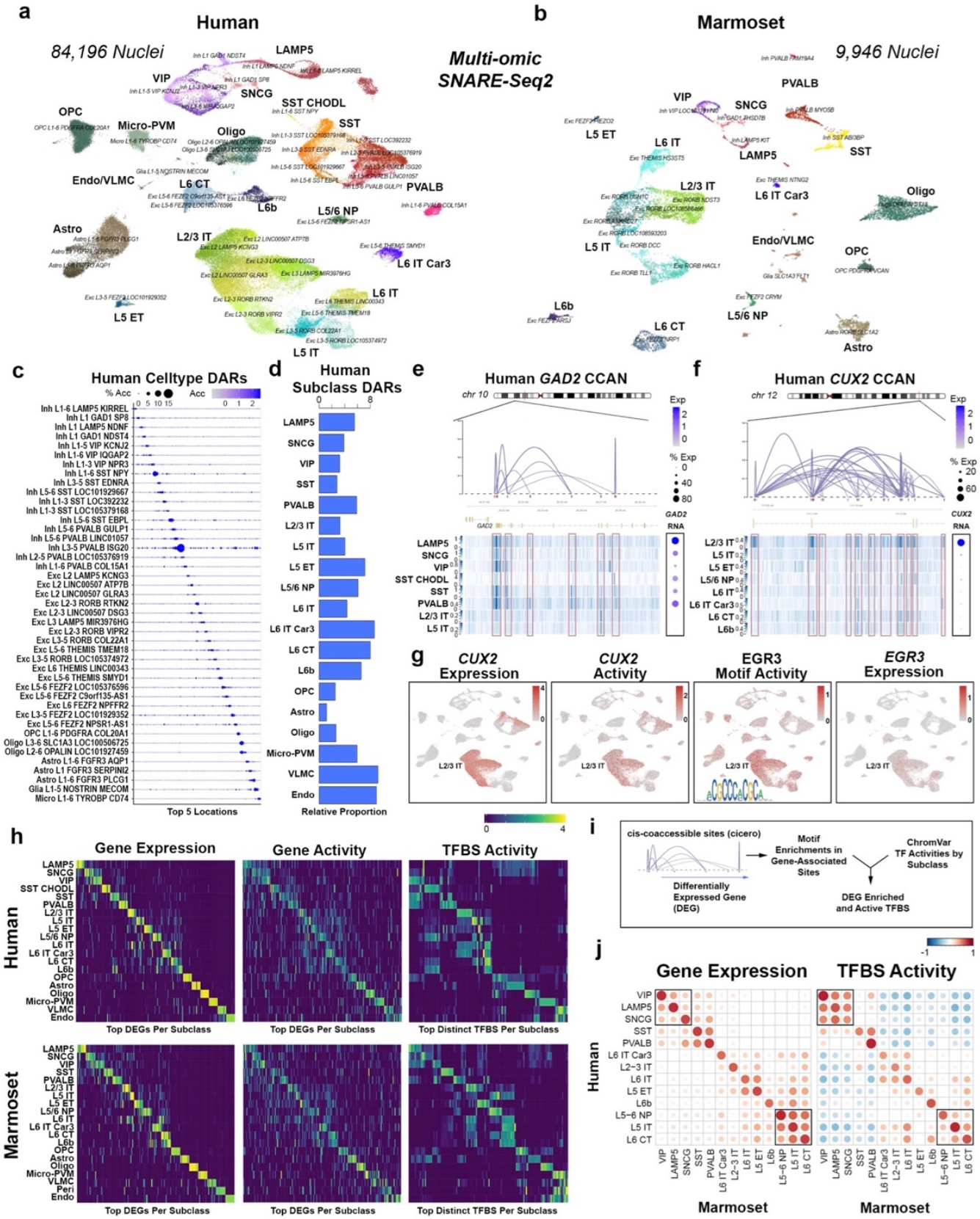


Figure 3. Dual-omic expression and chromatin accessibility reveals regulatory processes defining M1 cell types. a-b UMAP visualizations of human (**a**) and marmoset (**b**) M1 SNARE-Seq2

951 data (2 individuals per species) indicating both subclass and accessibility-level cluster identities. **c**, Dot
952 plot showing proportion and scaled average accessibility of differentially accessible regions (DARs)
953 identified between human AC clusters (adjusted P < 0.001, log-fold change > 1, top 5 distinct sites per
954 cluster). **d**, Proportion of total human or marmoset DARs identified between subclasses (adjusted P <
955 0.001, log-fold change > 1) after normalization to cluster sizes. **e**-**f**, Connection plots for cis-co-
956 accessible network (CCAN) sites associated with the human *GAD2* (**e**) and *CUX2* (**f**) genes.
957 Corresponding AC read pile-up tracts for GABAergic and select glutamatergic subclasses are shown.
958 Right panels are dot plots showing the percentage of expressing nuclei and average gene expression
959 values (log scale) for *GAD2* or *CUX2* within each of the clusters indicated. **g**, UMAP plots as in Figure
960 5a (human) showing (scaled from low—gray to high—red) *CUX2* gene expression (RNA) and activity
961 level predicted from AC data. UMAP plots for activity level of the EGR3-binding motif, identified using
962 chromVAR and found to be enriched within *CUX2* co-accessible sites, and the corresponding
963 expression (RNA) of the *EGR3* gene are shown. **h**, Heatmaps for human (top) and marmoset (bottom)
964 showing TFBS enrichments, according to the scheme outlined in (**i**), within genes differentially
965 expressed between subclasses and having at least two cis-co-accessible sites. Left panels show
966 scaled average (log scale) gene expression values (RNA) for the top DEGs (adjusted P < 0.05, log-fold
967 change > 1, top 10 distinct sites per cluster visualized), middle panels show the corresponding scaled
968 average cicero gene activity scores and the right panels show scaled values for the corresponding top
969 distinct chromVAR TFBS activities (adjusted P < 0.05, log-fold change > 0.5, top 10 distinct sites per
970 cluster visualized). **j**, Correlation plots comparing scaled average gene expression profiles (left panel)
971 or chromVAR TFBS activity scores (right panel) between human and marmoset matched subclasses.
972
973

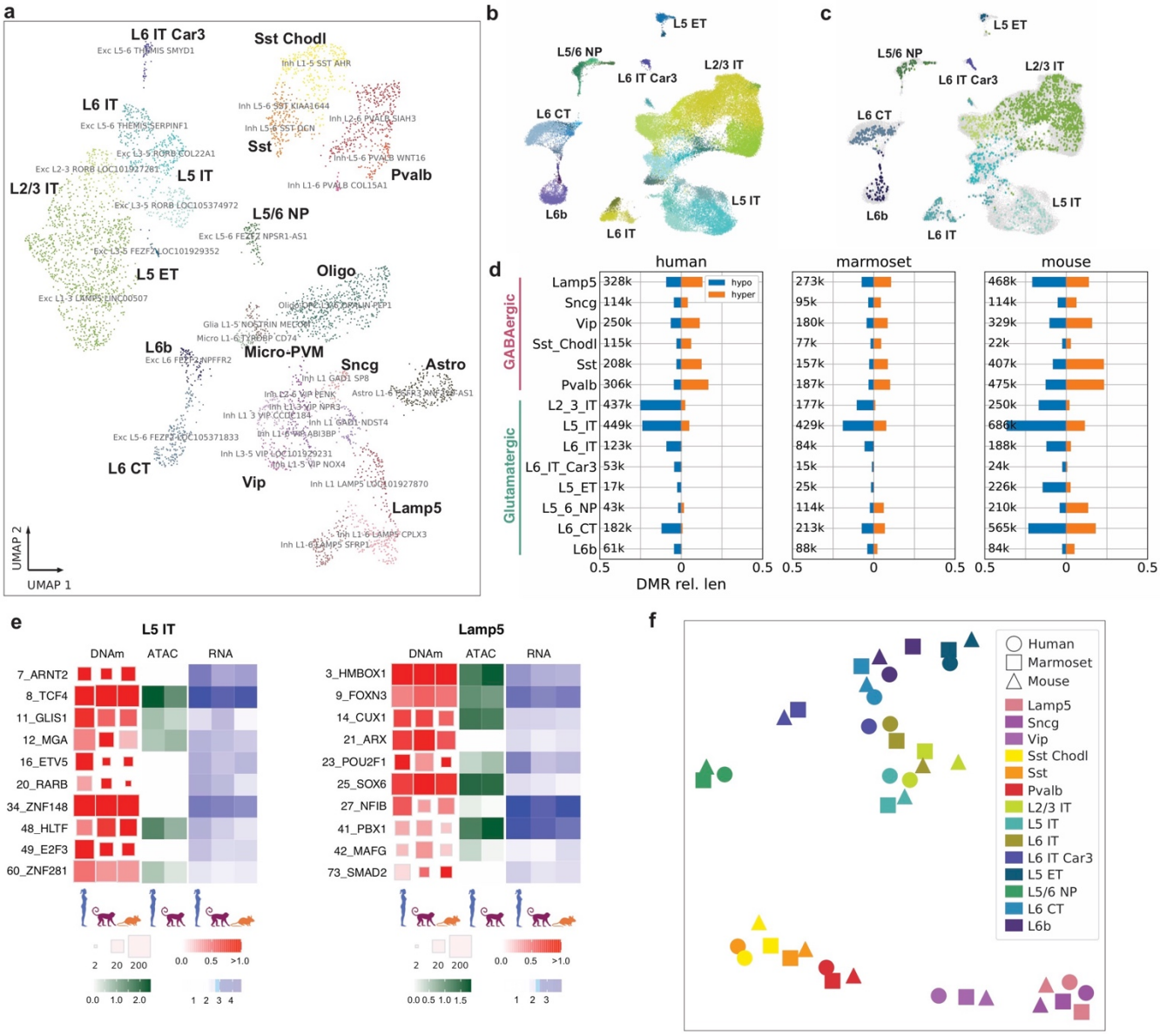


Figure 4. DNA methylation difference across clusters and species. a, UMAP visualization of human M1 DNAm-seq (snmC-seq2) data indicating both subclass and DNAm cluster identities. **b**,**c**, UMAP visualization of integration between DNAm-seq and RNA-seq of human glutamatergic neurons colored by cell subclass for all nuclei (**b**) or only nuclei profiled with DNAm-seq (**c**). **d**, Barplots of the relative lengths of hypo- and hyper-methylated DMRs among cell subclasses across three species normalized by cytosine coverage genome-wide (Methods). Total number of DMRs for each subclass are listed (k, thousands). **e**, Distinct TF motif enrichment for L5 IT and *Lamp5* subclasses across species. **f**, t-SNE visualization of subclass TF motif enrichment that is conserved across species.

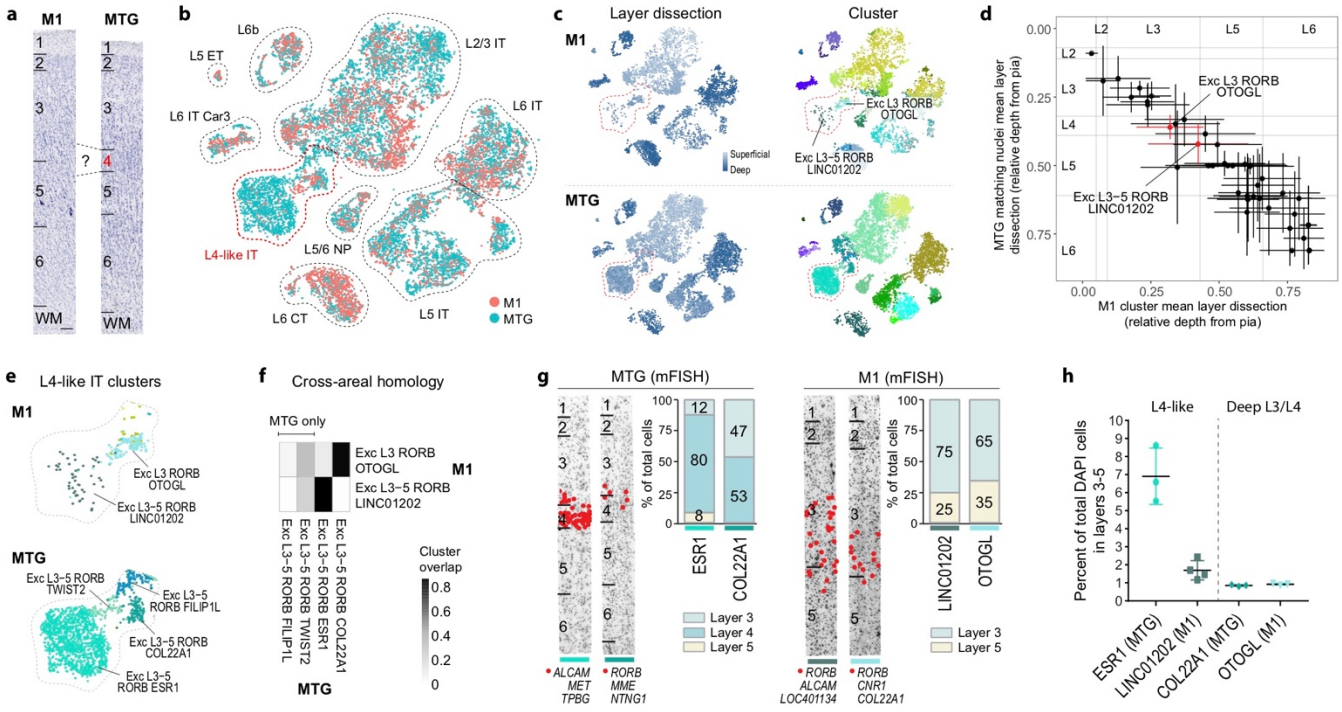


Figure 5. L4-like neurons identified in M1 based on cross-areal cell type homology. a, t-SNE projection of glutamatergic neuronal nuclei from M1 and MTG based on similarity of integrated expression levels. Nuclei are intermixed within all cell subclasses. **b**, Nuclei annotated based on the relative depth of the dissected layer and within-area cluster. A subset of clusters from superficial layers are highlighted. **c**, Proportion of nuclei in each cluster that overlap between areas. MTG clusters *COL22A1* and *ESR1* map almost one-to-one with M1 clusters *OTOGL* and *LINC01202*, respectively. **d**, Estimated relative depth from pia of M1 glutamatergic clusters and closest matching MTG neurons based on similarity of integrated expression. Mean (points) and standard deviation (bars) of the dissected layer are shown for each cluster and approximate layer boundaries are indicated for M1 and MTG. **e**, Magnified view of L4-like clusters in M1 and MTG. Note that MTG clusters *FILIP1L* and *TWIST2* have little overlap with any M1 clusters. **f**, Overlap of M1 and MTG clusters in integrated space identifies two one-to-one cell type homologies and two MTG-specific clusters. **g**, ISH labeling of MTG and M1 clusters quantifies differences in layer distributions for homologous types between cortical areas. Cells (red dots) in each cluster were labeled using the markers listed below each representative inverted image of a DAPI-stained cortical column. **h**, ISH estimated frequencies of homologous clusters shows M1 has a 4-fold sparser L4-like type and similarly rare deep L3 type.

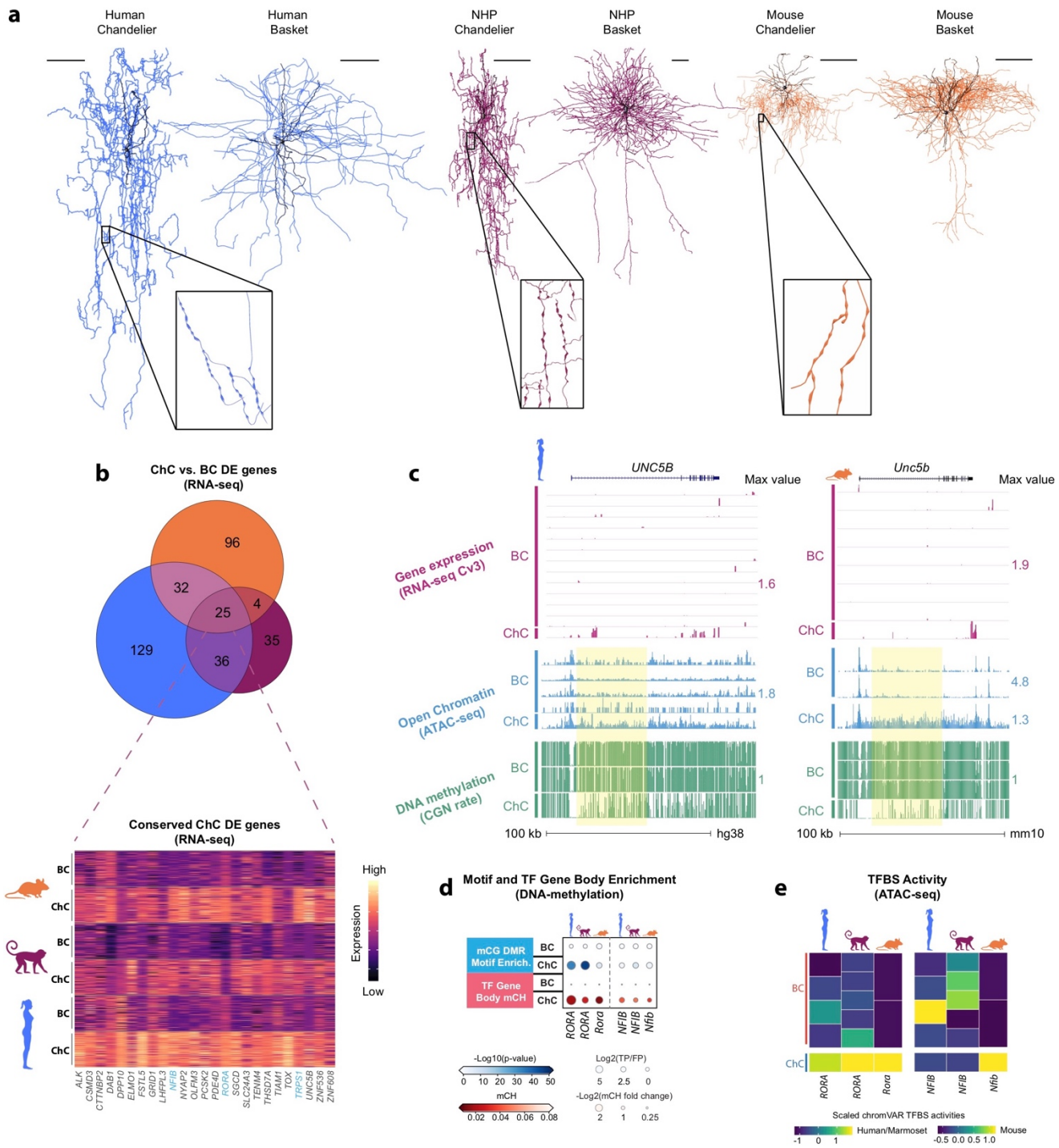


Figure 6. Chandelier neurons have conserved molecular features that may contribute to similar morphology across species. a, Representative ultrastructure reconstructions of a ChC and BC from human (left), macaque (middle), and mouse (right). Insets show higher magnification of ChC axon cartridges. Macaque reconstructions were from source data available in Neuromorpho [63,64]. **b,** Venn diagram indicating the number of shared ChC-enriched genes across species (top). DEGs were

007 determined by a ROC test of ChCs against BCs within a species. Heatmap showing scaled expression
008 of the 25 conserved DEGs in 100 randomly selected ChC and BC nuclei for each species (bottom);
009 transcription factors are colored in blue. **c**, Genome browser tracks showing *UNC5B* locus in human
010 (left) and mouse (right) ChCs and BCs. Tracks show aligned transcripts, regions of accessible
011 chromatin, CGN methylation rate, and CHN methylation rate. Yellow highlights mark examples of ChC-
012 enriched regions of accessible chromatin with hypo-methylated CGN. **d,** Heatmaps of TF gene body
013 hypo-methylation (mCH) state (bottom half, red) and genome-wide enrichment of TF motif across mCG
014 DMRs in ChCs and BCs (top half, blue). **e,** Scaled TFBS activities identified from SNARE-seq2 for
015 human and marmoset according to the scheme in Figure 4i and from mouse snATAC-seq data, using
016 genes enriched in ChC versus BC (Supplementary Table 19). Rows correspond to BC and ChC
017 clusters identified in snATAC-seq and SNARE-seq2 datasets.
018

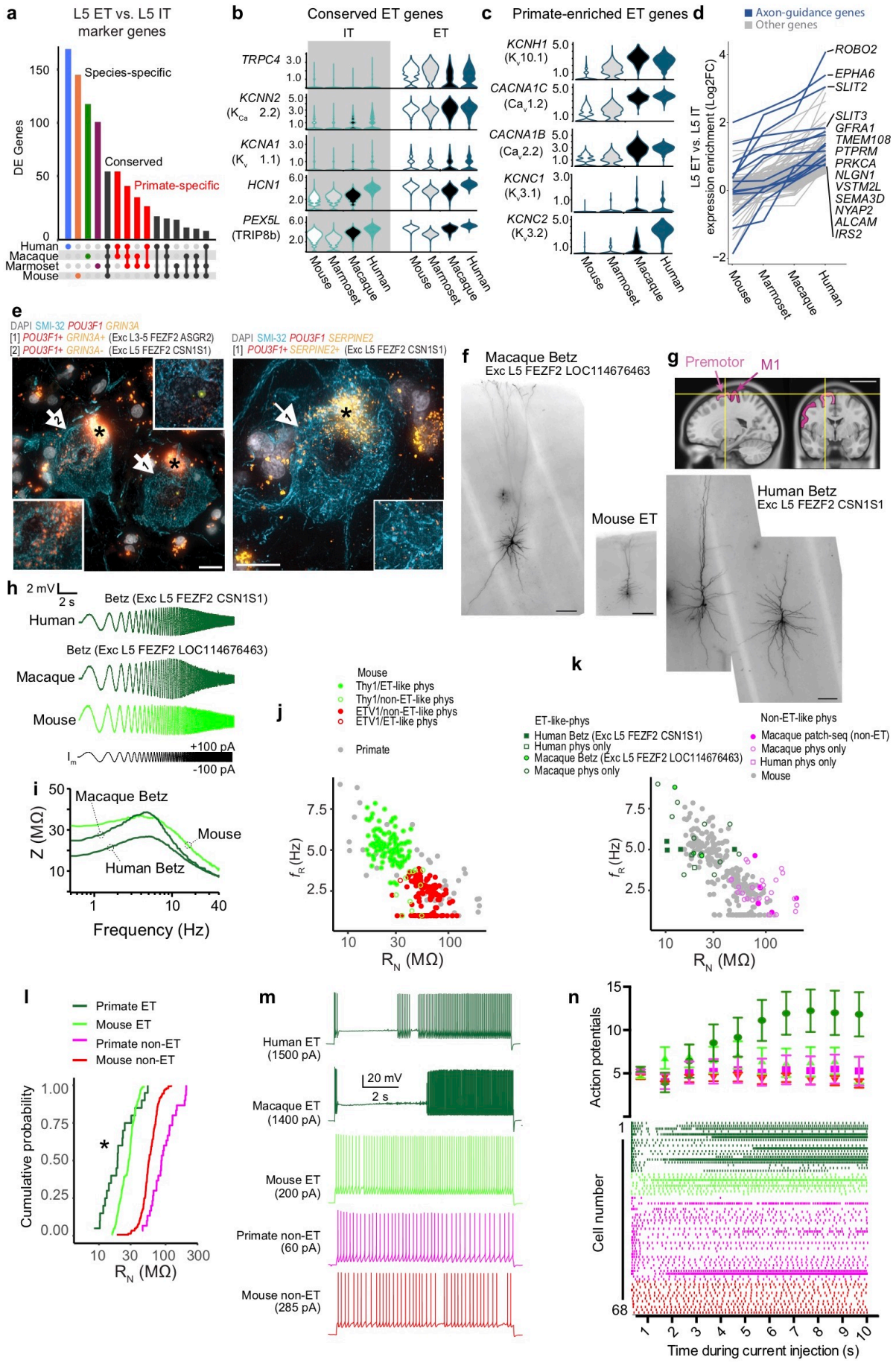

a
L5 ET vs. L5 IT marker genes
DE Genes
Species-specific
Conserved
Primate-specific
Human
Macaque
Marmoset
Mouse
b
Conserved ET genes
IT
ET
TRPC4
KCNN2 (K_Ca 2.2)
KCNA1 (K_v 1.1)
HCN1
PEX5L (TRIP8b)
Mouse
Marmoset
Macaque
Human
c
Primate-enriched ET genes
KCNH1 (K_v10.1)
CACNA1C (Ca_v1.2)
CACNA1B (Ca_v2.2)
KCNC1 (K_v3.1)
KCNC2 (K_v3.2)
d
Axon-guidance genes
Other genes
L5 ET vs. L5 IT expression enrichment (Log2FC)
ROBO2
EPHA6
SLIT2
SLIT3
GFRA1
TMEM108
PTPRM
PRKCA
NLGN1
VSTM2L
SEMA3D
NYAP2
ALCAM
IRS2
e
DAPI SMI-32 POU3F1 GRIN3A
[1] POU3F1+ GRIN3A+ (Exc L3-5 FEZF2 ASGR2)
[2] POU3F1+ GRIN3A- (Exc L5 FEZF2 CSN1S1)
DAPI SMI-32 POU3F1 SERPINE2
[1] POU3F1+ SERPINE2+ (Exc L5 FEZF2 CSN1S1)
f
Macaque Betz
Exc L5 FEZF2 LOC114676463
Mouse ET
g
Premotor
M1
Human Betz
Exc L5 FEZF2 CSN1S1
h
2 mV
2 s
Betz (Exc L5 FEZF2 CSN1S1)
Human
Betz (Exc L5 FEZF2 LOC114676463)
Macaque
Mouse
+100 pA
-100 pA
i
Macaque Betz
Mouse
Human Betz
Z (MΩ)
Frequency (Hz)
j
Mouse
Thy1/ET-like phys
Thy1/non-ET-like phys
ETV1/non-ET-like phys
ETV1/ET-like phys
Primate
f_R (Hz)
R_N (MΩ)
k
ET-like-phys
Human Betz (Exc L5 FEZF2 CSN1S1)
Human phys only
Macaque Betz (Exc L5 FEZF2 LOC114676463)
Macaque phys only
Non-ET-like phys
Macaque patch-seq (non-ET)
Macaque phys only
Human phys only
Mouse
f_R (Hz)
R_N (MΩ)
l
Primate ET
Mouse ET
Primate non-ET
Mouse non-ET
Cumulative probability
R_N (MΩ)
m
Human ET (1500 pA)
20 mV
2 s
Macaque ET (1400 pA)
Mouse ET (200 pA)
Primate non-ET (60 pA)
Mouse non-ET (285 pA)
n
Action potentials
Cell number
Time during current injection (s)

Figure 7. Betz cells have specialized molecular and physiological properties. a, Upset plot showing conserved and divergent L5 ET glutamatergic neuron marker genes. DEGs were determined by performing a ROC test between L5 ET and L5 IT within each species. **b, c,** Violin plots of ion channel-related gene expression for genes that are enriched in (**b**) ET versus IT neurons and in (**c**) primate versus mouse ET neurons. Protein names are in parentheses. **d,** Line graph of 131 genes with expression enrichment in L5 ET versus IT neurons in human (>0.5 $\log_2$ fold-change) that decreases with evolutionary distance from human. **e,** Two example photomicrographs of ISH labeled, SMI-32 IF stained Betz cells in L5 human M1. Cells corresponding to two L5 ET clusters are labeled based on two sets of marker genes. Insets show higher magnification of ISH in corresponding cells. Asterisks mark lipofuscin; scale bar, 20 µm. **f, g,** Exemplar biocytin fills obtained from patch-seq experiments in human, macaque and mouse brain slices. The example human and macaque neurons mapped to a Betz cell transcriptomic cell type. Scale bars, 200 µm. **g,** MRI image in sagittal and coronal planes and approximate location of excised premotor cortex tissue (yellow lines) and adjacent M1. **h,** Voltage responses to a chirp stimulus for the neurons shown in **f** and **g** (left human neuron). **i,** Corresponding ZAP profiles. All neurons were clustered into putative ET and non-IT neurons based upon their resonant frequency and input resistance. **j,** For mouse L5 neurons (*Thy1*-YFP line H, n=117; *Etv1*-EGFP line, n=123; unlabeled, n=21) 99.2 % of neurons in the *Etv1*-EGFP line possessed non-ET-like physiology, whereas, 91.4% of neurons in the *Thy1*-YFP line H had ET-like physiology. **k,** For primate L5 neurons (human, n=8, macaque n=42), all transcriptomically-defined Betz cells (human, n=4, macaque n=3) had ET-like physiology (human n=6, macaque, n=14), whereas all transcriptomically-defined non-ET neurons (human n=2, macaque n=3) had non-ET like physiology (human n=2, macaque n=28). **l,** Cumulative probability distribution of L5 neuron input resistance for primate versus mouse. * $p = 0.0064$, Kolmogorov-Smirnov test between mouse and primate ET neurons. **m,** Example voltage responses to 10s step current injections for monkey, mouse and human ET and non-ET neurons. The amplitude of the current injection was adjusted to produce ~5 spikes during the first second. **n**, Raster plot (below) and average firing rate (above) during 1 s epochs during the 10s DC current injection. Primate ET neurons (pooled data from human and macaque) displayed a distinctive

047 decrease followed by a pronounced increase in firing rate over the course of the current injection.
048 Notably, a similar biphasic-firing pattern is observed in macaque corticospinal neurons *in vivo* during
049 prolonged motor movements [65], suggesting that the firing pattern of these neurons during behavior is
050 intimately tied to their intrinsic membrane properties.
051

052 Methods

053 ***Ethical compliance***

054 Postmortem adult human brain tissue was collected after obtaining permission from decedent next-of-
055 kin. Postmortem tissue collection was performed in accordance with the provisions of the United States
056 Uniform Anatomical Gift Act of 2006 described in the California Health and Safety Code section 7150
057 (effective 1/1/2008) and other applicable state and federal laws and regulations. The Western
058 Institutional Review Board reviewed tissue collection processes and determined that they did not
059 constitute human subjects research requiring institutional review board (IRB) review.
060

061 ***Postmortem human tissue specimens***

062 Male and female donors 18–68 years of age with no known history of neuropsychiatric or neurological
063 conditions ('control' cases) were considered for inclusion in the study (Extended Data Table 1). Routine
064 serological screening for infectious disease (HIV, Hepatitis B, and Hepatitis C) was conducted using
065 donor blood samples and only donors negative for all three tests were considered for inclusion in the
066 study. Only specimens with RNA integrity (RIN) values ≥7.0 were considered for inclusion in the study.
067 Postmortem brain specimens were processed as previously described [2]. Briefly, coronal brain slabs
068 were cut at 1cm intervals and frozen for storage at -80°C until the time of further use. Putative hand and
069 trunk-lower limb regions of the primary motor cortex were identified, removed from slabs of interest, and
070 subdivided into smaller blocks. One block from each donor was processed for cryosectioning and

071 fluorescent Nissl staining (Neurotrace 500/525, ThermoFisher Scientific). Stained sections were
072 screened for histological hallmarks of primary motor cortex. After verifying that regions of interest
073 contained M1, blocks were processed for nucleus isolation as described below.
074
075 ***Human RNA-sequencing, QC and clustering***
076 SMART-seq v4 nucleus isolation and sorting. Vibratome sections were stained with fluorescent Nissl
077 permitting microdissection of individual cortical layers (dx.doi.org/10.17504/protocols.io.7aehibe).
078 Nucleus isolation was performed as previously described (dx.doi.org/10.17504/protocols.io.ztqf6mw).
079 NeuN staining was carried out using mouse anti-NeuN conjugated to PE (FCMAB317PE, EMD
080 Millipore) at a dilution of 1:500. Control samples were incubated with mouse IgG1k-PE Isotype control
081 (BD Biosciences, 555749). DAPI (4',6-diamidino-2-phenylindole dihydrochloride, ThermoFisher
082 Scientific, D1306) was applied to nuclei samples at a concentration of 0.1µg/ml. Single-nucleus sorting
083 was carried out on either a BD FACSAria II SORP or BD FACSAria Fusion instrument (BD
084 Biosciences) using a 130 µm nozzle. A standard gating strategy based on DAPI and NeuN staining was
085 applied to all samples as previously described [2]. Doublet discrimination gates were used to exclude
086 nuclei aggregates.
087
088 SMART-seq v4 RNA-sequencing. The SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (Takara
089 #634894) was used per the manufacturer's instructions. Standard controls were processed with each
090 batch of experimental samples as previously described. After reverse transcription, cDNA was amplified
091 with 21 PCR cycles. The NexteraXT DNA Library Preparation (Illumina FC-131-1096) kit with
092 NexteraXT Index Kit V2 Sets A-D (FC-131-2001, 2002, 2003, or 2004) was used for sequencing library
093 preparation. Libraries were sequenced on an Illumina HiSeq 2500 instrument using Illumina High
094 Output V4 chemistry.
095
096 SMART-seq v4 gene expression quantification. Raw read (fastq) files were aligned to the GRCh38
097 human genome sequence (Genome Reference Consortium, 2011) with the RefSeq transcriptome

098 version GRCh38.p2 (current as of 4/13/2015) and updated by removing duplicate Entrez gene entries
099 from the gtf reference file for STAR processing. For alignment, Illumina sequencing adapters were
100 clipped from the reads using the fastqMCF program. After clipping, the paired-end reads were mapped
101 using Spliced Transcripts Alignment to a Reference (STAR) using default settings. Reads that did not
102 map to the genome were then aligned to synthetic construct (i.e. ERCC) sequences and the *E. coli*
103 genome (version ASM584v2). Quantification was performed using summerizeOverlaps from the R
104 package GenomicAlignments. Expression levels were calculated as counts per million (CPM) of exonic
105 plus intronic reads.
106
107 10x Chromium RNA-sequencing. Nucleus isolation for 10x Chromium RNA-sequencing was conducted
108 as described (dx.doi.org/10.17504/protocols.io.y6rfzd6). After sorting, single-nucleus suspensions were
109 frozen in a solution of 1X PBS, 1% BSA, 10% DMSO, and 0.5% RNAsin Plus RNase inhibitor
110 (Promega, N2611) and stored at -80°C. At the time of use, frozen nuclei were thawed at 37°C and
111 processed for loading on the 10x Chromium instrument as described
112 (dx.doi.org/10.17504/protocols.io.nx3dfqn). Samples were processed using the 10x Chromium Single-
113 Cell 3' Reagent Kit v3. 10x chip loading and sample processing was done according to the
114 manufacturer's protocol. Gene expression was quantified using the default 10x Cell Ranger v3 pipeline
115 except substituting the curated genome annotation used for SMART-seq v4 quantification. Introns were
116 annotated as "mRNA", and intronic reads were included in expression quantification.
117
118 Quality control of RNA-seq data. Nuclei were included for analysis if they passed all QC criteria.
119 SMART-seq v4 criteria:
120 > 30% cDNA longer than 400 base pairs
121 > 500,000 reads aligned to exonic or intronic sequence
122 > 40% of total reads aligned
123 > 50% unique reads
124 > 0.7 TA nucleotide ratio

Cv3 criteria:

> 500 (non-neuronal nuclei) or > 1000 (neuronal nuclei) genes detected

< 0.3 doublet score

Clustering RNA-seq data. Nuclei passing QC criteria were grouped into transcriptomic cell types using a previously reported iterative clustering procedure (Tasic et al. 2018; Hodge, Bakken et al., 2019). Briefly, intronic and exonic read counts were summed, and $\log_2$-transformed expression was centered and scaled across nuclei. X- and Y-chromosomes and mitochondrial genes were excluded to avoid nuclei clustering based on sex or nuclei quality. DEGs were selected, principal components analysis (PCA) reduced dimensionality, and a nearest neighbor graph was built using up to 20 principal components. Clusters were identified with Louvain community detection (or Ward's hierarchical clustering if N < 3000 nuclei), and pairs of clusters were merged if either cluster lacked marker genes. Clustering was applied iteratively to each subcluster until clusters could not be further split.

Cluster robustness was assessed by repeating iterative clustering 100 times for random subsets of 80% of nuclei. A co-clustering matrix was generated that represented the proportion of clustering iterations that each pair of nuclei were assigned to the same cluster. We defined consensus clusters by iteratively splitting the co-clustering matrix as described (Tasic et al. 2018; Hodge, Bakken et al., 2019). The clustering pipeline is implemented in the R package “scrattch.hicat”, and the clustering method is provided by the “run_consensus_clust” function (https://github.com/AllenInstitute/scrattch.hicat).

Clusters were curated based on QC criteria or cell class marker expression (*GAD1*, *SLC17A7*, *SNAP25*). Clusters were identified as donor-specific if they included fewer nuclei sampled from donors than expected by chance. To confirm exclusion, clusters automatically flagged as outliers or donor-specific were manually inspected for expression of broad cell class marker genes, mitochondrial genes related to quality, and known activity-dependent genes.

152 ***Marmoset sample processing and nuclei isolation***

153 Marmoset experiments were approved by and in accordance with Massachusetts Institute of
154 Technology IACUC protocol number 051705020. Two adult marmosets (2.3 and 3.1 years old; one
155 male, one female; Extended Data Table 2) were deeply sedated by intramuscular injection of ketamine
156 (20-40 mg/kg) or alfaxalone (5-10 mg/kg), followed by intravenous injection of sodium pentobarbital
157 (10–30 mg/kg). When pedal withdrawal reflex was eliminated and/or respiratory rate was diminished,
158 animals were transcardially perfused with ice-cold sucrose-HEPES buffer. Whole brains were rapidly
159 extracted into fresh buffer on ice. Sixteen 2-mm coronal blocking cuts were rapidly made using a
160 custom-designed marmoset brain matrix. Coronal slabs were snap-frozen in liquid nitrogen and stored
161 at -80°C until use.

162

163 As with human samples, M1 was isolated from thawed slabs using fluorescent Nissl staining
164 (Neurotrace 500/525, ThermoFisher Scientific). Stained sections were screened for histological
165 hallmarks of primary motor cortex. Nuclei were isolated from the dissected regions following this
166 protocol: https://www.protocols.io/view/extraction-of-nuclei-from-brain-tissue-2srged6 and processed
167 using the 10x Chromium Single-Cell 3' Reagent Kit v3. 10x chip loading and sample processing was
168 done according to the manufacturer's protocol.

169

170 ***Marmoset RNA-sequencing, QC and clustering***

171 RNA-sequencing. Libraries were sequenced on NovaSeq S2 instruments (Illumina). Raw sequencing
172 reads were aligned to calJac3. Mitochondrial sequence was added into the published reference
173 assembly. Human sequences of RNR1 and RNR2 (mitochondrial) and RNA5S (ribosomal), were
174 aligned using gmap to the marmoset genome and added to the calJac3 annotation. Reads that mapped
175 to exons or introns of each assembly were assigned to annotated genes. Libraries were sequenced to a
176 median read depth of 5.95 reads per unique molecular index (UMI). The alignment pipeline can be
177 found at https://github.com/broadinstitute/Drop-seq.

178

179 Cell filtering. Cell barcodes were filtered to distinguish true nuclei barcodes from empty beads and PCR
180 artifacts by assessing proportions of ribosomal and mitochondrial reads, ratio of intronic/exonic reads (>
181 50% of intronic reads), library size (> 1000 UMIs) and sequencing efficiency (true cell barcodes have
182 higher reads/UMI). The resulting digital gene expression matrix (DGE) from each library was carried
183 forward for clustering.
184
185 Clustering. Clustering analysis proceeded as in Krienen et al (2019, bioRxiv). Briefly, independent
186 component analysis. (ICA, using the fastICA package in R) was performed jointly on all marmoset
187 DGEs after normalization and variable gene selection as in (Saunders et al 2018, *Cell*). The first-round
188 clustering resulted in 15 clusters corresponding to major cell classes (neurons, glia, endothelial). Each
189 cluster was curated as in (Saunders et al 2018, *Cell*) to remove doublets and outliers. Independent
190 components (ICs) were partitioned into those reflecting artifactual signals (e.g. those for which cell
191 loading indicated replicate or batch effects). Remaining ICs were used to determine clustering (Louvain
192 community detection algorithm igraph package in R); for each cluster nearest neighbor and resolution
193 parameters were set to optimize 1:1 mapping between each IC and a cluster.
194
195 ***Mouse snRNA-seq data generation and analysis***
196 Single-nuclei were isolated from mouse primary motor cortex, gene expression was quantified using
197 Cv3 and SSv4 RNA-sequencing, and transcriptomic cell types and dendrogram were defined as
198 described in a companion paper [6].
199
200 ***Integrating and clustering human Cv3 and SSv4 snRNA-seq datasets***
201 To establish a set of human consensus cell types, we performed a separate integration of snRNA-seq
202 technologies on the major cell classes (Glutamatergic, GABAergic, and Non-neuronal). Broadly, this
203 integration is comprised of 6 steps: (1) subsetting the major cell class from each technology (e.g. Cv3
204 GABAergic and SSv4 GABAergic); (2) finding marker genes for all clusters within each technology; (3)
205 integrating both datasets with Seurat's standard workflow using marker genes to guide integration

206 (Seurat 3.1.1); (4) overclustering the data to a greater number of clusters than were originally identified
207 within a given individual dataset; (5) finding marker genes for all integrated clusters; and (6) merging
208 similar integrated clusters together based on marker genes until all merging criteria were sufficed,
209 resulting in the final human consensus taxonomy.
210
211 More specifically, each expression matrix was $\log_2$(CPM + 1) transformed then placed into a Seurat
212 object with accompanying metadata. Variable genes were determined by downsampling each
213 expression matrix to a maximum of 300 nuclei per scrattch.hicat-defined cluster (from a previous step;
214 see scrattch.hicat clustering) and running select_markers (scrattch.io 0.1.0) with n set to 20, to
215 generate a list of up to 20 marker genes per cluster. The union of the Cv3 and SSv4 gene lists were
216 then used as input for anchor finding, dimensionality reduction, and Louvain clustering of the full
217 expression matrices. We used 100 dimensions for steps in the workflow, and 100 random starts during
218 clustering. Louvain clustering was performed to overcluster the dataset to identify more integrated
219 clusters than the number of scrattch.hicat-defined clusters. For example, GABAergic neurons had 79
220 and 37 scrattch.hicat-defined clusters, 225 overclustered integrated clusters, and 72 final human
221 consensus clusters after merging for Cv3 and SSv4 datasets, respectively. To merge the overclustered
222 integrated clusters, up to 20 marker genes were found for each cluster to establish the neighborhoods
223 of the integrated dataset. Clusters were then merged with their nearest neighbor if there were not a
224 minimum of ten Cv3 and two SSv4 nuclei in a cluster, and a minimum of 4 DEGs that distinguished the
225 query cluster from the nearest neighbor (note: these were the same parameters used to perform the
226 initial scrattch.hicat clustering of each dataset).
227
228 ***Integrating and clustering MTG and M1 SSv4 snRNA-seq datasets***
229 To compare cell types between our M1 human cell type taxonomy and our previously described human
230 MTG taxonomy [2], we used Seurat's standard integration workflow to perform a supervised integration
231 of the M1 and MTG SSv4 datasets. Intronic and exonic reads were summed into a single expression
232 matrix for each dataset, CPM normalized, and placed into a Seurat object with accompanying

metadata. All nuclei from each major cell class were integrated and clustered separately. Up to 100 marker genes for each cluster within each dataset were identified, and the union of these two gene lists was used as input to guide alignment of the two datasets during integration, dimensionality reduction, and clustering steps. We used 100 dimensions for all steps in the workflow.

Integrating Cv3 snRNA-seq datasets across species

To identify homologous cell types across species, we used Seurat's SCTransform workflow to perform a separate supervised integration on each cell class across species. Raw expression matrices were reduced to only include genes with one-to-one orthologs defined in the three species (14,870 genes; downloaded from NCBI Homologene in November, 2019) and placed into Seurat objects with accompanying metadata. To avoid having one species dominate the integrated space and to account for potential differences in each species' clustering resolution, we downsampled the number of nuclei to have similar numbers across species at the subclass level (e.g. *Lamp5*, *Sst*, L2/3 IT, L6b, etc.). The species with the largest number of clusters under a given subclass was allowed a maximum of 200 nuclei per cluster. The remaining species then split this theoretical maximum (200 nuclei times the max number of clusters under subclass) evenly across their clusters. For example, the L2/3 IT subclass had 8, 4, and 3 clusters for human, marmoset, and mouse, respectively. All species were allowed a maximum of 1600 L2/3 IT nuclei total; or a maximum of 200 human, 400 marmoset, and 533 mouse nuclei per cluster. To integrate across species, all Seurat objects were merged and normalized using Seurat's SCTransform function. To better guide the alignment of cell types from each species, we found up to 100 marker genes for each cluster within a given species. We used the union of these gene lists as input for integration and dimensionality reduction steps, with 30 dimensions used for integration and 100 for dimensionality reduction and clustering. Clustering the human-marmoset-mouse integrated space provided an additional quality control mechanism, revealing numerous small, species-specific integrated clusters that contained only low-quality nuclei (low UMIs and genes detected). We excluded 4836 nuclei from the marmoset dataset that constituted low-quality integrated neuronal clusters.

To identify which clusters in our three species taxonomy aligned with macaque clusters from our L5 dissected Cv3 dataset, we performed an identical integration workflow on Glutamatergic neurons as was used for the three species integration. Macaque clusters were assigned subclass labels based on their corresponding alignment with subclasses from the other species. The annotated L5 dissected macaque Cv3 dataset was then used as a reference for mapping macaque patch-seq nuclei (see section below).

Estimation of cell type homology

To identify homologous groups from different species, we applied a tree-based method (https://github.com/AllenInstitute/BICCN_M1_Evo and package: https://github.com/huqiwen0313/speciesTree). In brief, the approach consists of 4 steps: 1) metacell clustering, 2) hierarchical reconstruction of a metacell tree, 3) measurements of species mixing and stability of splits and 4) dynamic pruning of the hierarchical tree.

Firstly, to reduce noise in single-cell datasets and to remove species-specific batch effects, we clustered cells into small highly similar groups based on the integrated matrix generated by Seurat, as described in the previous section. These cells were further aggregated into metacells and the expression values of the metacells were calculated by averaging the gene expression of individual cells that belong to each metacell. Correlation was calculated based on the metacell gene expression matrix to infer the similarity of each metacell cluster. Then hierarchical clustering was performed based on the metacell gene expression matrix using Ward's method. For each node or corresponding branch in the hierarchical tree, we calculated 3 measurements, and the hierarchical tree was visualized based on these measurements: 1) Cluster size visualized as the thickness of branches in the tree; 2) Species mixing calculated based on entropy of the normalized cell distribution and visualized as the color of each node and branch; 3) Stability of each node. The entropy of cells was calculated as: $H = -\sum_i p_i log p_i$, where p_i is the probability of cells from one species appearing among all the cells in a node. We assessed the node stability by evaluating the agreement between the original hierarchical

287 tree and a result on a subsampled dataset calculated based on the optimal subtree in the subsampled
288 hierarchical trees derived from subsampling 95% of cells in the original dataset. The entire subsampling
289 process was repeated 100 times and the mean stability score for every node in the original tree was
290 calculated. Finally, we recursively searched each node in the tree. If the heuristic criteria (see below)
291 were not met for any node below the upper node, the entire subtree below the upper node was pruned
292 and all the cells belonging to this subtree were merged into one homologous group.

293 To identify robust homologous groups, we applied criteria in two steps to dynamically search the cross-
294 species tree. Firstly, for each node in the tree, we computed the mixing of cells from 3 species based
295 on entropy and set it as a tuning parameter. For each integrated tree, we tuned the entropy parameter
296 to make sure the tree method generated the highest resolution of homologous clusters without losing
297 the ability to identify potential species-specific clusters. Nodes with entropy larger than 2.9 (for inhibitory
298 neurons) or 2.75 (for excitatory neurons) were considered as well-mixed nodes. For example, an
299 entropy of 2.9 corresponded to a mixture of human, marmoset, and mouse equal to (0.43, 0.37, 0.2) or
300 (0.38, 0.30, 0.32). We recursively searched all the nodes in the tree until we found the node nearest the
301 leaves of the tree that was well-mixed among species, and this node was defined as a well-mixed
302 upper node. Secondly, we further checked the within-species cell composition for the subtrees below
303 the well-mixed upper node to determine if further splits were needed. For the cells on the subtrees
304 below the well mixed upper node, we measured the purity of within-species cell composition by
305 calculating the percentage of cells that fall into a specific sub-group in each individual species. If the
306 purity for any species was larger than 0.8, we went one step further below the well mixed upper node
307 so that its children were selected. Any branches below these nodes (or well-mixed upper node if the
308 within-species cell composition criteria was not met) were pruned and cells from these nodes were
309 merged into the same homologous groups, then the final integrated tree was generated.

310 As a final curation step, the homologous groups generated by the tree method were merged to be
311 consistent with within-species clusters. We defined consensus types by comparing the overlap of
312 within-species clusters between human and marmoset and human and mouse, as previously described

313 [2]. For each pair of human and mouse clusters and human and marmoset clusters, the overlap was
314 defined as the sum of the minimum proportion of nuclei in each cluster that overlapped within each leaf
315 of the pruned tree. This approach identified pairs of clusters that consistently co-clustered within one or
316 more leaves. Cluster overlaps varied from 0 to 1 and were visualized as a heatmap with human M1
317 clusters in rows and mouse or marmoset M1 clusters in columns. Cell type homologies were identified
318 as one-to-one, one-to-many, or many-to many so that they were consistent in all three species. For
319 example, the Vip_2 consensus type could be resolved into multiple homologous types between human
320 and marmoset but not human and mouse, and the coarser homology was retained. Consensus type
321 names were assigned based on the annotations of member clusters from human and mouse and
322 avoided specific marker gene names due to the variability of marker expression across species.
323
324 To quantify cell type alignment between pairs of species, we pruned the hierarchical tree described
325 above based on the stability and mixing of two species. We performed this analysis for human-
326 marmoset, human-mouse, and marmoset-mouse and compared the alignment resolution of each
327 subclass. The pruning criteria were tuned to fit the two-species comparison and to remove bias, and we
328 set the same criteria for all comparisons (entropy cutoff 3.0). Specifically, for each subclass and
329 pairwise species comparison, we calculated the number of leaves in the pruned tree. We repeated this
330 analysis on the 100 subsampled datasets and calculated the mean and standard deviation of the
331 number of leaves in the pruned trees. For each subclass, we tested for significant differences in the
332 average number of leaves across pairs of species using an ANOVA test followed by post-hoc Tukey
333 HSD tests.
334
335 ***Marker determination for cell type clusters by NS-Forest v2.1***
336 NS-Forest v2.1 was used to determine the minimum set of marker genes whose combined expression
337 identified cells of a given type with maximum classification accuracy (T. Bakken et al. 2017; Aevermann
338 et al. 2018). (https://github.com/JCVenterInstitute/NSForest/releases). Briefly, for each cluster NS-
339 Forest produces a Random Forest (RF) model using a one vs all binary classification approach. The

340 top ranked genes from RF are then filtered by expression level to retain genes that are expressed in at
341 least 50% of the cells within the target cluster. The selected genes are then reranked by Binary Score
342 calculated by first finding median cluster expression values for a given gene and dividing by the target
343 median cluster expression value. Next, one minus this scaled value is calculated resulting in 0 for the
344 target cluster and 1 for clusters that have no expression, while negative scaled values are set to 0.
345 These values are then summed and normalized by dividing by the total number of clusters. In the ideal
346 case, where all off-target clusters have no expression, the binary score is 1. Finally, for the top 6 binary
347 genes optimal expression level cutoffs are determined and all permutations of genes are evaluated by
348 f-beta score, where the beta is weighted to favor precision. This f-beta score indicates the power of
349 discrimination for a cluster and a given set of marker genes. The gene combination giving the highest f-
350 beta score is selected as the optimal marker gene combination. Marker gene sets for human, mouse
351 and marmoset primary motor cortex are listed in Supplementary Tables 4, 5, and 6, respectively, and
352 were used to construct the semantic cell type definitions provided in Supplementary Table 1.
353
354 ***Calculating differentially expressed genes (DEGs)***
355 To identify subclass level DEGs that are conserved and divergent across species, we used the
356 integrated Seurat objects from the species integration step. Seurat objects for each major cell class
357 were downsampled to have up to 200 cells per species cell type. Positive DEGs were then found using
358 Seurat's `FindAllMarkers` function using the ROC test with default parameters. We compared each
359 subclass within species to all remaining nuclei in that class and used the SCT normalized counts to test
360 for differential expression. For example, human *Sst* nuclei were compared to all other GABAergic
361 human neurons using the ROC test. Venn diagrams were generated using the eulerr package (6.0.0) to
362 visualize the relationship of DEGs across species for a given subclass. Heatmaps of DEGs for all
363 subclasses under a given class were generated by downsampling each subclass to 50 random nuclei
364 per species. SCT normalized counts were then scaled and visualized with Seurat's `DoHeatmap`
365 function.
366

367 To identify ChC DEGs that are enriched over BCs, we used the integrated Seurat objects from the
368 species integration step. The *Pvalb* subclass was subset and species cell types were then designated
369 as either ChCs or BCs. Positive DEGs were then found using Seurat's `FindAllMarkers` function
370 using the ROC test to compare ChCs and BCs for each species. Venn diagrams were generated using
371 the eulerr package (6.0.0) to visualize the relationship of ChC-enriched DEGs across species.
372 Heatmaps of conserved DEGs were generated by downsampling the dataset to have 100 randomly
373 selected BCs and ChCs from each species. SCT normalized counts were then scaled and visualized
374 with Seurat's `DoHeatmap` function.
375
376 We used the four species (human, macaque, marmoset, and mouse) integrated Glutamatergic Seurat
377 object from the species integration step for all L5 ET DEG figures. L5 ET and L5 IT subclasses were
378 downsampled to 200 randomly selected nuclei per species. A ROC test was then performed using
379 Seurat's `FindAllMarkers` function between the two subclasses for each species to identify L5 ET-
380 specific marker genes. We then used the UpSetR (1.4.0) package to visualize the intersections of the
381 marker genes across all four species as an upset plot. To determine genes that decrease in expression
382 across evolutionary distance in L5 ET neurons, we found the log-fold change between L5 ET and L5 IT
383 for each species across all genes. We then filtered the gene lists to only include genes that had a trend
384 of decreasing log-fold change (human > macaque > marmoset > mouse). Lastly, we excluded any gene
385 that did not have a log-fold change of 0.5 or greater in the human comparison. These 131 genes were
386 then used as input for GO analysis with the PANTHER Classification System [66] for the biological
387 process category, with organism set to Homo sapiens. All significant GO terms for this gene list were
388 associated with cell-cell adhesion and axon-guidance, and are colored blue in the line graph of their
389 expression enrichment.
390
391 ***Estimating differential isoform usage between human and mouse***
392 To assess changes of isoform usage between mouse and human, we used SSv4 data with full
393 transcript coverage and estimated isoform abundance in each cell subclasses. To mitigate low read

394 depth in each cell, we aggregated reads from all cells in each subclass. We estimated the relative
395 isoform usage in each subclass by calculating its genic proportion (P), defined as the ratio (R) of
396 isoform expression to the gene expression, where $R = (P_{human} - P_{mouse}) / (P_{human} + P_{mouse})$. For a common
397 set of transcripts for mouse and human, we used UCSC browser TransMapV5 set of human transcripts
398 (hg38 assembly, Gencode v31 annotations) mapped to the mouse genome (mm10 assembly)
399 http://hgdownload.soe.ucsc.edu/gbdb/mm10/transMap/V5/mm10.ensembl.transMapV5.bigPsl. We
400 considered only medium to highly expressed isoforms, which have abundance > 10 TPM (Transcripts
401 per Million) and P > 0.2 in either mouse or human and gene expression > 10 TPM in both mouse and
402 human.
403
404 Calculating isoform abundance in each cell subclass:
405 1) Aggregated reads from each subclass
406 2) Mapped reads to the mouse or human reference genome with STAR 2.7.3a using default
407 parameters
408 3) Transformed genomic coordinates into transcriptomic coordinates using STAR parameter: --
409 quantMode TranscriptomeSAM
410 4) Quantified isoform and gene expression using RSEM 1.3.3 parameters: --bam --seed 12345 --
411 paired-end --forward-prob 0.5 --single-cell-prior --calc-ci
412
413 Estimating statistical significance:
414 1) Calculated the standard deviation of isoform genic proportion (P_{human} and P_{mouse}) from the
415 RSEM's 95% confidence intervals of isoform expression
416 2) Calculated the P-value using normal distribution for the ($P_{human} - P_{mouse}$) and the summed
417 (mouse + human) variance
418 3) Bonferroni-adjusted P-values by multiplying nominal P-values by the number of medium to
419 highly expressed isoforms in each subclass
420

421 ***Species cluster dendrograms***

422 DEGs for a given species were identified using Seurat's `FindAllMarkers` function with a Wilcox test
423 and comparing each cluster to every other cluster under the same subclass, with logfc.threshold set to
424 0.7 and min.pct set to 0.5. The union of up to 100 genes per cluster with the highest avg_logFC were
425 used.The average $\log_2$ expression of the DEGs were then used as input for the `build_dend` function
426 from scrattch.hicat to create the dendrograms. This was performed on both human and marmoset
427 datasets. For mouse dendrogram methods, see the companion paper [6].

428

429 ***Multiplex fluorescent in situ hybridization (FISH)***

430 Fresh-frozen human postmortem brain tissues were sectioned at 14-16 µm onto Superfrost Plus glass
431 slides (Fisher Scientific). Sections were dried for 20 minutes at -20°C and then vacuum sealed and
432 stored at -80°C until use. The RNAscope multiplex fluorescent v1 kit was used per the manufacturer's
433 instructions for fresh-frozen tissue sections (ACD Bio), except that fixation was performed for 60
434 minutes in 4% paraformaldehyde in 1X PBS at 4°C and protease treatment was shortened to 5minutes.
435 Primary antibodies were applied to tissues after completion of mFISH staining. Primary antibodies used
436 were mouse anti-GFAP (EMD Millipore, MAB360, 1:250 dilution) and mouse anti-Neurofilament H
437 (SMI-32, Biolegend, 801701). Secondary antibodies were goat anti-mouse IgG (H+L) Alexa Fluor
438 conjugates (594, 647). Sections were imaged using a 60X oil immersion lens on a Nikon TiE
439 fluorescence microscope equipped with NIS-Elements Advanced Research imaging software (version
440 4.20). For all RNAscope mFISH experiments, positive cells were called by manually counting RNA
441 spots for each gene. Cells were called positive for a gene if they contained ≥ 3 RNA spots for that

442 gene. Lipofuscin autofluorescence was distinguished from RNA spot signal based on the larger size of

443 lipofuscin granules and broad fluorescence spectrum of lipofuscin.

444

445 ***Gene family conservation***

446 To investigate the conservation and divergence of gene family coexpression between primates and

447 mouse, MetaNeighbor analysis [30] was performed using gene groups curated by the HUGO Gene

448 Nomenclature Committee (HGNC) at the European Bioinformatics Institute

449 (https://www.genenames.org; downloaded January 2020) and by the Synaptic Gene Ontology (SynGO)

450 [67] (downloaded February 2020). HGNC annotations were propagated via the provided group hierarchy

451 to ensure the comprehensiveness of parent annotations. Only groups containing five or more genes

452 were included in the analysis.

453

454 After splitting data by class, MetaNeighbor was used to compare data at the cluster level using labels

455 from cross-species integration with Seurat. Cross-species comparisons were performed at two levels of

456 the phylogeny: 1) between the two primate species, marmoset and human; and 2) between mouse and

457 primates. In the first case, the data from the two species were each used as the testing and training set

458 across two folds of cross-validation, reporting the average performance (AUROC) across folds. In the

459 second case, the primate data were used as an aggregate training set, and performance in mouse was

460 reported. Results were compared to average within-species performance.

461

462 ***Replicability of clusters***

463 MetaNeighbor was used to provide a measure of neuronal subclass and cluster replicability within and

464 across species. For this application, we tested all pairs of species (human-marmoset, marmoset-

465 mouse, human-mouse) as well as testing within each species. After splitting the data by class, highly

466 variable genes were identified using the get_variable_genes function from MetaNeighbor, yielding 928

467 genes for GABAergic and 763 genes for Glutamatergic neuron classes, respectively. These were used

468 as input for the MetaNeighborUS function, which was run using the fast_version and one_vs_best
469 parameters set to TRUE. Using the one_vs_best parameter means that only the two closest
470 neighboring clusters are tested for their similarity to the training cluster, with results reported as the
471 AUROC for the closest neighbor over the second closest. AUROCs are plotted in heatmaps in
472 Extended Data Figures 2 and 3. Data to reproduce these figures can be found in Supplementary Table
473 9, and scripts are on GitHub (http://github.com/gillislab/MetaNeighbor).
474
475 ***Single-cell methylome data (snmC-seq2): Sequencing and quantification***
476 Library preparation and Illumina sequencing. Detailed methods for bisulfite conversion and library
477 preparation were previously described for snmC-seq2[5,41]. The snmC-seq2 libraries generated from
478 mouse brain tissues were sequenced using an Illumina Novaseq 6000 instrument with S4 flowcells and
479 150 bp paired-end mode.
480
481 Mapping and feature count pipeline. We implemented a versatile mapping pipeline (http://cemba-
482 data.rtfd.io) for all the single-cell methylome based technologies developed by our group [5,41,68]. The
483 main steps of this pipeline included: 1) demultiplexing FASTQ files into single-cell; 2) reads level QC; 3)
484 mapping; 4) BAM file processing and QC; and 5) final molecular profile generation. The details of the
485 five steps for snmC-seq2 were described previously [41]. We mapped all the reads from the three
486 corresponding species onto the human hg19 genome, the marmoset ASM275486v1 genome, and the
487 mouse mm10 genome. After mapping, we calculated the methyl-cytosine counts and total cytosine
488 counts for two sets of genome regions in each cell: the non-overlapping chromosome 100-kb bins of
489 each genome, the methylation levels of which were used for clustering analysis, and the gene body
490 regions, the methylation levels of which were used for cluster annotation and integration with RNA
491 expression data.
492
493 ***snmC-seq2: Quality control and preprocessing***

494 Cell filtering. We filtered the cells based on these main mapping metrics: 1) mCCC rate < 0.03. mCCC
495 rate reliably estimates the upper bound of bisulfite non-conversion rate [5]; 2) overall mCG rate > 0.5; 3)
496 overall mCH rate < 0.2; 4) total final reads > 500,000; and 5) bismark mapping rate > 0.5. Other metrics
497 such as genome coverage, PCR duplicates rate, and index ratio were also generated and evaluated
498 during filtering. However, after removing outliers with the main metrics 1-5, few additional outliers can
499 be found.
500
501 Feature filtering. 100kb genomic bin features were filtered by removing bins with mean total cytosine
502 base calls < 250 or > 3000. Regions overlap with the ENCODE blacklist [69] were also excluded from
503 further analysis.
504
505 Computation and normalization of the methylation rate. For CG and CH methylation, the computation of
506 methylation rate from the methyl-cytosine and total cytosine matrices contains two steps: 1) prior
507 estimation for the beta-binomial distribution and 2) posterior rate calculation and normalization per cell.
508 Step 1. For each cell we calculated the sample mean, m, and variance, v, of the raw mc rate *(mc / cov)*
509 for each sequence context (CG, CH). The shape parameters (α, β) of the beta distribution were then
510 estimated using the method of moments:

511 $$\alpha = m(m(1-m)/v - 1)$$

512 $$\beta = (1-m)(m(1-m)/v - 1)$$

513 This approach used different priors for different methylation types for each cell and used weaker prior to
514 cells with more information (higher raw variance).
515
516 Step 2. We then calculated the posterior: $\widehat{mc} = \frac{\alpha + mc}{\alpha + \beta + cov}$., We normalized this rate by the cell's global
517 mean methylation, $m = \alpha/(\alpha + \beta)$. Thus, all the posterior $\widehat{mc}$ with 0 *cov* will be constant 1 after
518 normalization. The resulting normalized *mc* rate matrix contains no NA (not available) value, and
519 features with less *cov* tend to have a mean value close to 1.
520

521 Selection of highly variable features. Highly variable methylation features were selected based on a
522 modified approach using the scanpy package *scanpy.pp.highly_variable_genes* function [70]. In brief, the
523 *scanpy.pp.highly_variable_genes* function normalized the dispersion of a gene by scaling with the
524 mean and standard deviation of the dispersions for genes falling into a given bin for mean expression of
525 genes. In our modified approach, we reasoned that both the mean methylation level and the mean *cov*
526 of a feature (100kb bin or gene) could impact *mc* rate dispersion. We grouped features that fall into a
527 combined bin of mean and *cov*, and then normalized the dispersion within each *mean-cov* group. After
528 dispersion normalization, we selected the top 3000 features based on normalized dispersion for
529 clustering analysis.
530
531 Dimension reduction and combination of different mC types. For each selected feature, *mc* rates were
532 scaled to unit variance, and zero mean. PCA was then performed on the scaled *mc* rate matrix. The
533 number of significant PCs was selected by inspecting the variance ratio of each PC using the elbow
534 method. The CH and CG PCs were then concatenated together for further analysis in clustering and
535 manifold learning.
536
537 ***snmC-seq2: Data analysis***
538 Consensus clustering on concatenated PCs. We used a consensus clustering approach based on
539 multiple Leiden-clustering [71] over K-Nearest Neighbor (KNN) graph to account for the randomness of
540 the Leiden clustering algorithms. After selecting dominant PCs from PCA in both mCH and mCG
541 matrix, we concatenated the PCs together to construct a KNN graph using *scanpy.pp.neighbors* with
542 Euclidean distance. Given fixed resolution parameters, we repeated the Leiden clustering 300 times on
543 the KNN graph with different random starts and combined these cluster assignments as a new feature
544 matrix, where each single Leiden result is a feature. We then used the outlier-aware DBSCAN
545 algorithm from the scikit-learn package to perform consensus clustering over the Leiden feature matrix
546 using the hamming distance. Different epsilon parameters of DBSCAN are traversed to generate
547 consensus cluster versions with the number of clusters that range from minimum to the maximum

548 number of clusters observed in the multiple Leiden runs. Each version contained a few outliers that
549 usually fall into three categories: 1) cells located between two clusters that had gradient differences
550 instead of clear borders; 2) cells with a low number of reads that potentially lack information in essential
551 features to determine the specific cluster; and 3) cells with a high number of reads that were potential
552 doublets. The amount of type 1 and 2 outliers depends on the resolution parameter and is discussed in
553 the choice of the resolution parameter section. The type 3 outliers were very rare after cell filtering. The
554 supervised model evaluation then determined the final consensus cluster version.
555
556 Supervised model evaluation on the clustering assignment. For each consensus clustering version, we
557 performed a Recursive Feature Elimination with Cross-Validation (RFECV) [72] process from the scikit-
558 learn package to evaluate clustering reproducibility. We first removed the outliers from this process,
559 and then we held out 10% of the cells as the final testing dataset. For the remaining 90% of the cells,
560 we used tenfold cross-validation to train a multiclass prediction model using the input PCs as features
561 and *sklearn.metrics.balanced_accuracy_score* [73] as an evaluation score. The multiclass prediction
562 model is based on *BalancedRandomForestClassifier* from the imblearn package that accounts for
563 imbalanced classification problems[74]. After training, we used the 10% testing dataset to test the model
564 performance using the *balanced_accuracy_score* score. We kept the best model and corresponding
565 clustering assignments as the final clustering version. Finally, we used this prediction model to predict
566 outliers' cluster assignments, we rescued the outlier with prediction probability > 0.3, otherwise labeling
567 them as outliers.
568
569 Choice of resolution parameter. Choosing the resolution parameter of the Leiden algorithm is critical for
570 determining the final number of clusters. We selected the resolution parameter by three criteria: 1. The
571 portion of outliers < 0.05 in the final consensus clustering version. 2. The ultimate prediction model
572 accuracy > 0.95. 3. The average cell per cluster ≥ 30, which controls the cluster size to reach the

573 minimum coverage required for further epigenome analysis such as DMR calls. All three criteria

574 prevented the over-splitting of the clusters; thus, we selected the maximum resolution parameter under

575 meeting the criteria using a grid search.

576

577 Three-level of iterative clustering analysis. We used an iterative approach to cluster the data into three

578 levels of categories with the consensus clustering procedure described above. In the first level termed

579 CellClass, clustering analysis is done on all cells. The resulting clusters are then manually merged into

580 three canonical classes, glutamatergic neurons, GABAergic neurons, and non-neurons, based on

581 marker genes. The same clustering procedure was then conducted within each CellClass to get

582 clusters as the MajorType level. Within each MajorType, we got the final clusters as the SubTypes in

583 the same way.

584

585 Integrating cell clusters identified from snmC-seq2 and from Cv3. We identified gene markers based on

586 gene body mCH hypo-methylation for each level of clustering of snmC-seq2 data using our in-house

587 analysis utilities (https://github.com/lhqing/cemba_data), and identified gene markers for cell class from

588 Cv3 analysis using scanpy [70]. We then used Scanorama [75] to integrate the two modalities.

589

590 Calling CG differentially methylated regions (DMRs). We identified CG DMRs using methylpy

591 (https://github.com/yupenghe/methylpy) as previously described [76]. Briefly, we first called CG

592 differentially methylated sites and then merged them into blocks if they both showed similar sample-

593 specific methylation patterns and were within 250bp. Normalized relative lengths of DMRs (Figure 4d)

594 were calculated by summation of lengths of DMRs and 250bp around divided by numbers of cytosine

595 covered in sequencing.

596

597 TFBS motif enrichment analysis. For each cell subclass (cluster), we performed TFBS motif enrichment
598 analysis for its hypo-methylated DMRs against the hypo-methylated DMRs from other cell subclasses
599 (clusters) using software AME [77]. DMRs and 250bp regions around were used in the analysis.
600
601 ***SNARE-Seq2: Sample preparation***
602 Human and marmoset primary motor cortex nuclei were isolated for SNARE-seq2 according to the
603 following protocol: https://www.protocols.io/view/nuclei-isolation-for-snare-seq2-8tvhwn6 [7,78].
604 Fluorescence-activated nuclei sorting (FANS) was then performed on a FACSAria Fusion (BD
605 Biosciences, Franklin Lakes, NJ) gating out debris from FSC and SSC plots and selecting $DAPI^+$
606 singlets (Extended Data Fig. 5a). Samples were kept on ice until sorting was complete and were used
607 immediately for SNARE-seq2.
608
609 ***SNARE-Seq2: Library preparation and sequencing***
610 A detailed step-by-step protocol for SNARE-Seq2 has been outlined in a companion paper [38]. The
611 resulting AC libraries were sequenced on MiSeq (Illumina) (R1: 75 cycles for the 1st end of AC DNA
612 read, R2: 94 cycles for cell barcodes and UMI read, R3: 8 cycles for i5 read, R4: 75 cycles for the 2nd
613 end of AC DNA read) for library validation, then on NovaSeq6000 (Illumina) using 300 cycles reagent
614 kit for data generation. RNA libraries were combined at equimolar ratio and sequenced on MiSeq
615 (Illumina) (Read 1: 70 cycles for the cDNA read, Index 1: 6 cycles for i7 read, Read 2: 94 cycles for cell
616 barcodes and UMI read) for library validation, then on NovaSeq6000 (Illumina) using 200 cycles
617 reagent kit for data generation.
618
619 ***SNARE-Seq2: Data processing***
620 A detailed step-by-step SNARE-seq2 data processing pipeline has been provided in a companion
621 paper [38]. For RNA data, this has involved the use of dropEst to extract cell barcodes and STAR
622 (v2.5.2b) to align tagged reads to the genome (GRCh38 version 3.0.0 for human; GCF 000004665.1
623 Callithrix jacchus-3.2, marmoset). For AC data, this involved snaptools for alignment to the genome

624 (cellranger-atac-GRCh38-1.1.0 for human, GCF 000004665.1 Callithrix jacchus-3.2, marmoset) and to
625 generate snap objects for processing using the R package snapATAC.
626
627 ***SNARE-Seq2: Data analysis***
628 RNA quality filtering. For SNARE-Seq2 data, quality filtering of cell barcodes and clustering analysis
629 were first performed on transcriptomic (RNA) counts and used to inform on subsequent accessible
630 chromatin quality filtering and analysis. Each cell barcode was tagged by an associated library batch ID
631 (for example MOP1, MOP2… etc.), RNA read counts associated with dT and n6 adaptor primers were
632 merged, libraries were combined for each sample within each experiment and empty barcodes
633 removed using the emptyDrops() function of DropletUtils [79], mitochondrial transcripts were removed,
634 doublets were identified using the DoubletDetection software [80] and removed. All samples were
635 combined across experiments within species and cell barcodes having greater than 200 and less than
636 7500 genes detected were kept for downstream analyses. To further remove low quality datasets, a
637 gene UMI ratio filter (gene.vs.molecule.cell.filter) was applied using Pagoda2 (https://github.com/hms-
638 dbmi/pagoda2).
639
640 RNA data clustering. For human SNARE-seq2 RNA data, clustering analysis was first performed using
641 Pagoda2 where counts were normalized to the total number per nucleus and batch variations were
642 corrected by scaling expression of each gene to the dataset-wide average. After variance
643 normalization, the top 6000 over-dispersed genes were used for principal component analysis.
644 Clustering was performed using an approximate k-nearest neighbor graph (k values between 50 – 500)
645 based on the top 75 principal components and cluster identities were determined using the infomap
646 community detection algorithm. Major cell types were identified using a common set of broad cell type
647 marker genes: *GAD1/GAD2* (GABAergic neurons), *SLC17A7/SATB2* (glutamatergic neurons),
648 *PDGFRA* (oligodendrocyte progenitor cells), *AQP4* (astrocytes), *PLP1/MOBP* (oligodendrocytes),
649 *MRC1* (perivascular macrophages), *PTPRC* (T cells), *PDGFRB* (vascular smooth muscle cells), *FLT1*
650 (vascular endothelial cells), *DCN* (vascular fibroblasts), *APBB1IP* (microglia) (Extended Data Fig. 5c).

651 Low quality clusters that showed very low gene/UMI detection rates, low marker gene detection and/or
652 mixed cell type marker profiles were removed. Oligodendrocytes were over-represented (54,080 total),
653 possibly reflecting a deeper subcortical sampling then intended, therefore, to ensure a more balanced
654 distribution of cell types, we capped the number of oligodendrocytes at 5000 total and repeated the
655 PAGODA2 clustering as above. To achieve optimal clustering of the different cell types, different k
656 values were used to identify cluster subpopulations for different cell types (L2/3 glutamatergic neurons,
657 k = 500; all other glutamatergic neurons, astrocytes, oligodendrocytes, OPCs, k = 100; GABAergic
658 neurons, vascular cells, microglia/perivascular macrophages, k = 50). To assess the appropriateness of
659 the chosen k values, clusters were compared against SMARTer clustering of data generated on human
660 M1 through correlation of cluster-averaged scaled gene expression values using the corrplot package
661 (https://github.com/taiyun/corrplot) (Extended Data Fig. 5d). For cluster visualization, uniform manifold
662 approximation and projection (UMAP) dimensional reduction was performed in Seurat (version 3.1.0)
663 using the top 75 principal components identified using Pagoda2. For marmoset, clustering was initially
664 performed using Seurat, where the top 2000 variable features were selected from the mean variance
665 plot using the 'vst' method and used for principal component analysis. UMAP embeddings were
666 generated using the top 75 principal components. To harmonize cellular populations across platforms
667 and modalities, snRNA-seq within-species cluster identities were then predicted from both human and
668 marmoset data. We used an iterative nearest centroid classifier algorithm (Methods, 'Mapping of
669 samples to reference taxonomies') to generate probability scores for each SNARE-seq2 nuclei mapping
670 to their respective species' snRNA-seq reference cluster (Cv3 for marmoset and SMART-Seqv4 for
671 human). Comparing the predicted RNA cluster assignment of each nuclei with Pagoda2-identified
672 clusters showed highly consistent cluster membership using Jaccard similarity index (Extended Data
673 Fig. 5e), confirming the robustness of these cell identities discovered using different analysis platforms.
674
675 AC quality filtering and peak calling. Initial analysis of corresponding SNARE-Seq2 chromatin
676 accessibility data was performed using SnapATAC software (version 2)
677 (https://github.com/r3fang/SnapATAC) (https://doi.org/10.1101/615179). Snap objects were generated

678 by combining individual snap files across libraries within each species. Cell barcodes were included for
679 downstream analyses only if cell barcodes passed RNA quality filtering (above) and showed greater
680 than 1000 read fragments and 500 UMI. Read fragments were then binned to 5000 bp windows of the
681 genome and only cell barcodes showing the fraction of binned reads within promoters greater than 10%
682 (15% for marmoset) and less than 80% were kept for downstream analysis. Peak regions were called
683 independently for RNA cluster, subclass and class groupings using MACS2 software
684 (https://github.com/taoliu/MACS) using the following options "--nomodel --shift 100 --ext 200 --qval 5e-2
685 -B --SPMR". Peak regions were combined across peak callings and used to generate a single peak
686 count matrix (cell barcodes by chromosomal peak locations) using the "createPmat" function of
687 SnapATAC.
688
689 AC data clustering. The peak count matrices were filtered to keep only locations from chromosomes 1-
690 22, x or y, and processed using Seurat (version 3.1.0) and Signac (version 0.1.4) software [24]
691 (https://satijalab.org). All peaks having at least 100 counts (20 for marmoset) across cells were used for
692 dimensionality reduction using latent semantic indexing ("RunLSI" function) and visualized by UMAP
693 using the first 50 dimensions (40 for marmoset).
694
695 Calculating gene activity scores. For a gene activity matrix from accessibility data, cis-co-accessible
696 sites and gene activity scores were calculated using Cicero software (v1.2.0) [39] (https://cole-trapnell-
697 lab.github.io/cicero-release/). The binary peak matrix was used as input with expression family variable
698 set to "binomialff" to make the aggregated input Cicero CDS object using the AC peak-derived UMAP
699 coordinates and setting 50 cells to aggregate per bin. Co-accessible sites were then identified using the
700 "run_cicero" function using default settings and modules of cis-co-accessible sites identified using the
701 "generate_ccans" function. Co-accessible sites were annotated to a gene if they fell within a region
702 spanning 10,000 bp upstream and downstream of the gene's transcription start site (TSS). The Cicero
703 gene activity matrix was then calculated using the "build_gene_activity_matrix" function using a co-
704 accessibility cutoff of 0.25 and added to a separate assay of the Seurat object.

705

706 Integrating RNA/AC data modalities. For reconciliation of differing resolutions achievable from RNA and
707 accessible chromatin (Extended Data Fig. 5f-k), integrative analysis was performed using Seurat.
708 Transfer anchors were identified between the activity and RNA matrices using the
709 "FindTransferAnchors" function. For human, transfer anchors were generated using an intersected list
710 of variable genes identified from Pagoda2 analysis of RNA clusters (top 2000 genes) and marker genes
711 for clusters identified from SSv4 data (2492 genes having β-scores > 0.4), and canonical correlation
712 analysis (CCA) for dimension reduction. For marmoset, transfer anchors were generated using an
713 intersected list of variable genes identified using Seurat (top 2000 genes) and DEGs identified between
714 marmoset consensus clusters (Cv3 snRNA-seq data, $P < 0.05$, top 100 markers per cluster). Imputed
715 RNA expression values were then calculated using the "TransferData" function from the Cicero gene
716 activity matrix using normalized RNA expression values for reference and LSI for dimension reduction.
717 RNA and imputed expression data were merged, a UMAP co-embedding and shared nearest neighbor
718 (SNN) graph generated using the top 50 principal components (40 for marmoset) and clusters identified
719 ("FindClusters") using a resolution of 4. Resulting integrated clusters were compared against
720 consensus RNA clusters by calculating jaccard similarity scores using scratch.hicat software. Cell
721 populations identified as T-cells from Pagoda2 analysis (human only) and those representing low
722 quality integrated clusters, showing a mixture of disparate cell types, were removed from these
723 analyses. RNA clusters were assigned to co-embedded clusters based on the highest jaccard similarity
724 score and frequency and then merged to generate the best matched co-embedded clusters, taking in
725 account cell type and subclass to ensure more accurate merging of ambiguous populations. This
726 enabled AC-level clusters that directly matched the RNA-defined populations (Extended Data Fig. 5k).
727 For consensus cluster and subclass level predictions (Extended Data Fig. 5g) the Seurat
728 "TransferData" function was used to transfer RNA consensus cluster or subclass labels to AC data
729 using the pre-computed transfer anchors and LSI dimensionality reduction.
730

731 Final AC peak and gene activity matrices. A final combined list of peak regions was then generated
732 using MACS2 as detailed above for all cell populations corresponding to RNA consensus (> 100
733 nuclei), accessibility-level, subclass (> 50 nuclei) and class level barcode groupings. The corresponding
734 peak by cell barcode matrix generated by SnapATAC was used to establish a Seurat object as outlined
735 above, with peak counts, Cicero gene activity scores and RNA expression values for matched cell
736 barcodes contained within different assay slots.
737
738 Transcription factor motif analyses. Jaspar motifs (JASPAR2020, all vertebrate) were used to generate
739 a motif matrix and motif object that was added to the Seurat object using Signac ("CreateMotifMatrix",
740 "CreateMotifObject", "AddMotifObject") and GC content, region lengths and dinucleotide base
741 frequencies calculated using the "RegionStats" function. Motif enrichments within specific chromosomal
742 sites were calculated using the FindMotifs function. For motif activity scores, chromVAR
743 (https://greenleaflab.github.io/chromVAR) was performed according to default parameters (marmoset)
744 or using Signac "RunChromVAR" function on the peak count matrix (human). The chromVAR deviation
745 score matrix was then added to a separate assay slot of the Seurat object and differential activity of
746 TFBS between different populations were assessed using the "Find[All]Markers" function through
747 logistic regression and using the number of peak counts as a latent variable.
748
749 Differentially accessible regions (DARs) between cell populations (Fig. 4b) were identified using the
750 "find_all_diff" function (https://github.com/yanwu2014/chromfunks) and p-values calculated using a
751 hypergeometric test. For visualization, the top DARs (q value < 0.001 and log-fold change > 1) were
752 selected and the top distinct sites visualized by dot plot in Seurat. For motif enrichment analyses, peak
753 counts associated with the clusters selected for comparison (all subclasses, all AC-level clusters,
754 *PVALB*-positive for ChC analyses) were used to identify cis-co-accessible site networks or CCANs
755 using cicero as indicated above. Peak locations were annotated to the nearest gene (10,000 bases
756 upstream and downstream of the TSS) and only genes identified from SNARE-seq2 RNA data as being
757 differentially expressed (Seurat, Wilcoxon Rank Sum test) within the clusters of interest (adjusted P <

758 0.05, average log-fold change > 0.5) were used. Genes having more than one co-accessible site were
759 assessed for motif enrichments within all overlapping sites using the “FindMotifs” function in Signac
760 (using peaks for all cell barcodes for subclass and AC-level, or only peaks for ChC or L5 ET cells).
761 Motifs were then trimmed to only those showing significant differential activity (chromVAR) between the
762 clusters of interest (P < 0.05) as assessed using the “FindMarkers” function on the chromVAR assay
763 slot using Seurat and using the number of total peaks as a latent variable. The top distinct genes
764 (subclass, AC-level) or all genes (ChC, Betz) used for motif enrichment analysis were visualized for
765 scaled average RNA expression levels and scaled average cicero gene activities using the ggHeat
766 plotting function (SWNE package, https://github.com/yanwu2014/swne). Top chromVAR TFBS activities
767 were also visualized using ggHeat.
768
769 Correlation plots. For correlation of RNA expression and associated AC activities for consensus and
770 AC-level clusters (Extended Data Fig. 6a-b), average scaled expression values were generated and
771 pairwise correlations performed for marker genes identified from an intersected list of variable genes
772 identified from Pagoda2 analysis of RNA clusters (top 2000 genes) and marker genes for clusters
773 identified from SSv4 data (2492 genes having β-scores > 0.4). For correlation across species,
774 expression values for genes used to integrate human and marmoset GABAergic and glutamatergic
775 clusters (Cv3 scRNA-seq data), or chromVAR TFBS activity scores for all Jaspar motifs were averaged
776 by subclass, scaled (trimming values to a minimum of 0 and a maximum of 4) for each species
777 separately, then correlated and visualized using corrplot.
778
779 Plots and figures. All UMAP, feature, dot, and violin plots were generated using Seurat. Connection
780 plots were generated using cicero and peak track gradient heatmaps were generated using Gviz [81] from
781 bedGraph files generated during peak calling using SnapATAC. Correlation plots were generated using
782 the corrplot package.
783
784 ***Mouse chandelier cell ATAC-Seq: Data acquisition and analysis***

785 Chandelier cells are rare in mouse cortex and were enriched by isolating individual neurons from
786 transgenically-labelled mouse primary visual cortex (VISp). Many of the transgenic mouse lines have
787 previously been characterized by single-cell RNA-seq [1]. Single-cell suspensions of cortical neurons
788 were generated as described previously [1] and subjected to tagmentation (ATAC-seq) [82,83]. Mixed
789 libraries, containing 60 to 96 samples were sequenced on an Illumina MiSeq. In total, 4,275 single-cells
790 were collected from 36 driver-reporter combinations in 67 mice. After sequencing, raw FASTQ files
791 were aligned to the GRCm38 (mm10) mouse genome using Bowtie v1.1.0 as previously described [9].
792 Following alignment, duplicate reads were removed using samtools rmdup, which yielded only single
793 copies of uniquely mapped paired reads in BAM format. Quality control filtering was applied to select
794 samples with >10,000 uniquely mapped paired-end fragments, >10% of which were longer than 250
795 base pairs and with >25% of their fragments overlapping high-depth cortical DNase-seq peaks from
796 ENCODE [84]. The resulting dataset contained a total of 2,799 samples.
797
798 To increase the cell-type resolution of chromatin accessibility profiles beyond that provided by driver
799 lines, a feature-free method for computation of pairwise distances (Jaccard) was used. Using Jaccard
800 distances, principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE)
801 were performed, followed by Phenograph clustering [85]. This clustering method grouped cells from
802 class-specific driver lines together, but also segregated them into multiple clusters. Phenograph-defined
803 neighborhoods were assigned to cell subclasses and clusters by comparison of accessibility near
804 transcription start site (TSS ± 20 kb) to median expression values of scRNA-seq clusters at the cell type
805 and at the subclass level from mouse primary visual cortex [86]. From this analysis, a total of 226
806 samples were assigned to *Pvalb* and 124 samples to *Pvalb Vipr2* (ChC) clusters. The sequence data
807 for these samples were grouped together and further processed through the Snap-ATAC pipeline.
808
809 Mouse scATAC-seq peak counts for *Pvalb* and ChC were used to generate a Seurat object as outlined
810 for human and marmoset SNARE-Seq2 AC data. Cicero cis-co-accessible sites were identified, gene
811 activity scores calculated, and motif enrichment analyses performed as outlined above. Genes used for

812 motif enrichment were ChC markers identified from differential expression analysis between *PVALB-*
813 *positive* clusters in mouse Cv3 scRNA-seq data (adjusted $P < 0.05$).
814
815 ***Patch-seq neuronal physiology, morphology, and transcriptomics***
816 Subjects. The human neurosurgical specimen was obtained from a 61-year old female patient that
817 underwent deep tumor resection (glioblastoma) from the frontal lobe at a local hospital (Harborview
818 Medical Center). The patient provided informed consent and experimental procedures were approved
819 by the hospital institute review board before commencing the study. Post-hoc analysis revealed that the
820 neocortical tissue obtained from this patient was from a premotor region near the confluence of the
821 superior frontal gyrus and the precentral gyrus (Fig. 7g). All procedures involving macaques and mice
822 were approved by the Institutional Animal Care and Use Committee at either the University of
823 Washington or the Allen Institute for Brain Science. Macaque M1 tissue was obtained from male (n=4)
824 and female (n=5) animals (mean age= 10 ± 2.21 years) designated for euthanasia via the Washington
825 National Primate Research Center's Tissue Distribution Program. Mouse M1 tissue was obtained from
826 4-12 week old male and female mice from the following transgenic lines: *Thy1h*-eyfp (B6.Cg-Tg(*Thy1*-
827 YFP)-HJrs/J: JAX Stock No. 003782), *Etv1*-egfp Tg(*Etv1*-EGFP)BZ192Gsat/Mmucd (etv1) mice
828 maintained with the outbred Charles River Swiss Webster background (Crl:CFW(SW) CR Stock No.
829 024), and C57BL/6-Tg(*Pvalb*-tdTomato)15Gfng/J: JAX stock No. 027395.
830
831 Brain slice preparation. Brain slice preparation was similar for *Pvalb*-TdTomato mice, macaque and
832 human samples. Upon resection, human neurosurgical tissue was immediately placed in a chilled and
833 oxygenated solution formulated to prevent excitotoxicity and preserve neural function [87]. This artificial
834 cerebral spinal fluid (NMDG aCSF) consisted of (in mM): 92 with N-methyl-D-glucamine (NMDG), 2.5
835 KCl, 1.25 NaH_2PO_4, 30 $NaHCO_3$, 20 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 25
836 glucose, 2 thiourea, 5 Na-ascorbate, 3 Na-pyruvate, 0.5 $CaCl_2 \cdot 4H_2O$ and 10 $MgSO_4 \cdot 7H_2O$. The pH of
837 the NMDG aCSF was titrated to pH 7.3–7.4 with concentrated hydrochloric acid and the osmolality was
838 300-305 mOsmoles/Kg. The solution was pre-chilled to 2-4°C and thoroughly bubbled with carbogen

839 (95% O_2/5% CO_2) prior to collection. Macaques were anesthetized with sevoflurane gas during which
840 the entire cerebrum was extracted and placed in the same protective solution described above. After
841 extraction, macaques were euthanized with sodium-pentobarbital. We dissected the trunk/limb area of
842 the primary motor cortex for brain slice preparation. *Pvalb*-TdTomato mice were deeply anesthetized by
843 intraperitoneal administration of Advertin (20mg/kg IP) and were perfused through the heart with NMDG
844 aCSF (bubbled with carbogen).
845
846 Brains were sliced at 300-micron thickness on a vibratome using the NMDG protective recovery
847 method and a zirconium ceramic blade [61,87]. Mouse brains were sectioned coronally, and human and
848 macaque brains were sectioned such that the angle of slicing was perpendicular to the pial surface.
849 After sections were obtained, slices were transferred to a warmed (32-34° C) initial recovery chamber
850 filled with NMDG aCSF under constant carbogenation. After 12 minutes, slices were transferred to a
851 chamber containing an aCSF solution consisting of (in mM): 92 NaCl, 2.5 KCl, 1.25 NaH_2PO_4, 30
852 $NaHCO_3$, 20 HEPES, 25 glucose, 2 thiourea, 5 Na-ascorbate, 3 Na-pyruvate, 2 $CaCl_2 \cdot 4H_2O$ and 2
853 $MgSO_4 \cdot 7H_2O$ continuously bubbled with 95% O_2/5% CO_2. Slices were held in this chamber for use in
854 acute recordings or transferred to a 6-well plate for long-term culture and viral transduction. Cultured
855 slices were placed on membrane inserts and wells were filled with culture medium consisting of 8.4 g/L
856 MEM Eagle medium, 20% heat-inactivated horse serum, 30 mM HEPES, 13 mM D-glucose, 15 mM
857 $NaHCO_3$, 1 mM ascorbic acid, 2 mM $MgSO_4 \cdot 7H_2O$, 1 mM $CaCl_2.4H_2O$, 0.5 mM GlutaMAX-I, and 1 mg/L
858 insulin (Ting et al 2018). The slice culture medium was carefully adjusted to pH 7.2-7.3, osmolality of
859 300-310 mOsmoles/Kg by addition of pure H_2O, sterile-filtered and stored at 4°C for up to two weeks.
860 Culture plates were placed in a humidified 5% CO_2 incubator at 35°C and the slice culture medium was
861 replaced every 2-3 days until end point analysis. 1-3 hours after brain slices were plated on cell culture
862 inserts, brain slices were infected by direct application of concentrated AAV viral particles over the slice
863 surface (Ting et al 2018).
864

865 Thy1 and Etv1 mice were deeply anesthetized by IP administration of ketamine (130 mg/kg) and
866 xylazine (8.8 mg/kg) mix and were perfused through the heart with chilled (2-4°C) sodium-free aCSF
867 consisting of (in mM): 210 Sucrose, 7 D-glucose, 25 $NaHCO_3$, 2.5 KCl, 1.25 NaH_2PO_4, 7 $MgCl_2$, 0.5
868 $CaCl_2$, 1.3 Na-ascorbate, 3 Na-pyruvate bubbled with carbogen (95% O_2/5% CO_2). Near coronal slices
869 300 microns thick were generated using a Leica vibratome (VT1200) in the same sodium-free aCSF
870 and were transferred to warmed (35°C) holding solution (in mM): 125 NaCl, 2.5 KCl, 1.25 NaH_2PO_4, 26
871 $NaHCO_3$, 2 $CaCl_2$, 2 $MgCl_2$, 17 dextrose, and 1.3 sodium pyruvate bubbled with carbogen (95% O_2/5%
872 CO_2). After 30 minutes of recovery, the chamber holding slices was allowed to cool to room
873 temperature.
874
875 Patch clamp electrophysiology. Macaque, human and *Pvalb*-TdTomato mouse brain slices were placed
876 in a submerged, heated (32-34°C) recording chamber that was continually perfused (3-4 mL/min) with
877 aCSF under constant carbogenation and containing (in mM) 1): 119 NaCl, 2.5 KCl, 1.25 NaH_2PO_4, 24
878 $NaHCO_3$, 12.5 glucose, 2 $CaCl_2 \cdot 4H_2O$ and 2 $MgSO_4 \cdot 7H_2O$ (pH 7.3-7.4). Slices were viewed with an
879 Olympus BX51WI microscope and infrared differential interference contrast (IR-DIC) optics and a 40x
880 water immersion objective. The infragranular layers of macaque primary motor cortex and human
881 premotor cortex are heavily myelinated, which makes visualization of neurons under IR-DIC virtually
882 impossible. To overcome this challenge, we labeled neurons using various viral constructs in
883 organotypic slice cultures (Extended Data Fig. 10g).
884 Patch pipettes (2-6 MΩ) were filled with an internal solution containing (in mM): 110.0 K-gluconate,
885 10.0 HEPES, 0.2 EGTA, 4 KCl, 0.3 Na2-GTP, 10 phosphocreatine disodium salt hydrate, 1 Mg-ATP, 20
886 µg/ml glycogen, 0.5U/µL RNAse inhibitor (Takara, 2313A) and 0.5% biocytin (Sigma B4261), pH 7.3.
887 Fluorescently labeled neurons from *Thy1* or *Etv1* mice were visualized through a 40x objective using
888 either Dodt contrast with a CCD camera (Hamamatsu) and/or a 2-photon imaging/ uncaging system
889 from Prairie (Bruker) Technologies. Recordings were made in aCSF: (in mM): 125 NaCl, 3.0 KCl, 1.25
890 NaH_2PO_4, 26 $NaHCO_3$, 2 $CaCl_2$, 1 MgCl2, 17 dextrose, and 1.3 sodium pyruvate bubbled with
891 carbogen (95% O_2/5% CO_2) at 32-35°, with synaptic inhibition blocked using 100 µM picrotoxin.

892 Sylgard-coated patch pipettes (3-6 MΩ) were filled with an internal solution containing (in mM): 135 K-
893 gluconate, 12 KCl, 11 HEPES, 4 MgATP, 0.3 NaGTP, 7 K_2-phosphocreatine, 4 Na_2-phophocreatine (pH
894 7.42 with KOH) with neurobiotin (0.1-0.2%), Alexa 594 (40 μM) and Oregon Green BAPTA 6F (100
895 μM).
896
897 Whole cell somatic recordings were acquired using either a Multiclamp 700B amplifier, or an AxoClamp
898 2B amplifier (Molecular Devices) and were digitized using an ITC-18 (HEKA). Data acquisition software
899 was either MIES (https://github.com/AllenInstitute/MIES/) or custom software written in Igor Pro.
900 Electrical signals were digitized at 20-50 kHz and filtered at 2-10 kHz. Upon attaining whole-cell current
901 clamp mode, the pipette capacitance was compensated and the bridge was balanced. Access
902 resistance was monitored throughout the recording and was 8-25 MΩ.
903
904 Data analysis. Data were analyzed using custom analysis software written in Igor Pro. All
905 measurements were made at resting membrane potential. Input resistance (R_N) was measured from a
906 series of 1 s hyperpolarizing steps from -150 pA to +50 pA in +20 pA increments. For neurons with low
907 input resistance (e.g. the Betz cells) this current injection series was scaled by upwards of 4x. Input
908 resistance (R_N) was calculated from the linear portion of the current−steady state voltage relationship
909 generated in response to these current injections. Resonance (f_R) was determined from the voltage
910 response to a constant amplitude sinusoidal current injection (Chirp stimulus). The chirp stimulus
911 increased in frequency either linearly from 1-20 Hz over 20 s or logarithmically from 0.2-40 Hz over 20s.
912 The amplitude of the Chirp was adjusted in each cell to produce a peak-to-peak voltage deflection of
913 ~10 mV. The impedance amplitude profile (ZAP) was constructed from the ratio of the fast Fourier
914 transform of the voltage response to the fast Fourier transform of the current injection. ZAPs were
915 produced by averaging at least three presentations of the Chirp and were smoothed using a running
916 median smoothing function. The frequency corresponding to the peak impedance (Z_{max}) was defined as
917 the resonant frequency. Spike input/output curves were constructed in response to 1 s step current

injections (50 pA-500 pA in 50 pA steps). For a subset of experiments, this current injection series was extended to 3A in 600 pA steps to probe the full dynamic range of low R_N neurons. Spike frequency acceleration analysis was performed for current injections producing ~10 spikes during the 1 s step. Acceleration ratio was defined as the ratio of the second to the last interspike interval. To examine the dynamics of spike timing over longer periods, we also measured spiking in response to 10 s step current injections in which the amplitude of the current was adjusted to produce ~5 spikes in the first second. Action potential properties were measured for currents near rheobase. Action potential threshold was defined as the voltage at which the first derivative of the voltage response exceeded 20 V/s. AP width was measured at half the amplitude between threshold and the peak voltage. Fast AHP was defined relative to threshold. We clustered mouse, macaque and human pyramidal neurons into two broad groups based on their R_N and f_R using Ward's algorithm.

<u>Viral vector production and transduction.</u>

Recombinant AAV vectors were produced by triple-transfection of ITR-containing enhancer plasmids along with AAV helper and rep/cap plasmids using the AAV293 cell line, followed by harvest, purification and concentration of the viral particles. The AAV293 packaging cell line and plasmid supplying the helper function are available from a commercial source (Cell Biolabs). The PHP.eB capsid variant was generated by Dr. Viviana Gradinaru at the California Institute of Technology [88] and the DNA plasmid for AAV packaging is available from Addgene (plasmid#103005). Quality control of the packaged AAV was determined by viral titering to determine an adequate concentration was achieved (>$5E^{12}$ viral genomes per mL), and by sequencing the AAV genome to confirm the identity of the viral vector that was packaged. Human and NHP L5 ET neurons including Betz cells were targeted in cultured slices by transducing the slices with viral vectors that either generically label neurons (AAV-hSyn1-tdTomato), or that enrich for L5 ET neurons by expressing reporter transgene under the control of the msCRE4 enhancer [86].

944 Processing of Patch-seq samples. For a subset of experiments, the nucleus was extracted at the end of
945 the recording and processed for RNA-sequencing. Prior to data collection for these experiments, all
946 surfaces were thoroughly cleaned with RNAse Zap. The contents of the pipette were expelled into a
947 PCR tube containing lysis buffer (Takara, 634894). cDNA libraries were produced using the SMART-
948 Seq v4 Ultra Low Input RNA Kit for Sequencing according to the manufacturer's instructions. We
949 performed reverse transcription and cDNA amplification for X PCR cycles. Sample proceeded through
950 Nextera NT DNA Library Preparation using Nextera XT Index Kit V2 Set A(FC-131-2001).
951
952 Mapping of samples to reference taxonomies. To identify which cell type a given patch-seq nuclei
953 mapped to, we used our previously described nearest centroid classifier [1]. Briefly, a centroid classifier
954 was constructed for Glutamatergic reference data (human SSv4 or macaque Cv3) using marker genes
955 for each cluster. Patch-seq nuclei were then mapped to the appropriate species reference 100 times,
956 using 80% of randomly sampled marker genes during each iteration. Probabilities for each nuclei
957 mapping to each cluster were computed over the 100 iterations, resulting in a confidence score ranging
958 from 0 to 100. We identified four human patch-seq nuclei that mapped with > 85% confidence and four
959 macaque nuclei that mapped with > 93% confidence to a cluster in the L5 ET subclass.
960

961 ## Data availability

962 Raw sequence data are available for download from the Neuroscience Multi-omics Archive
963 (https://nemoarchive.org/) and the Brain Cell Data Center (https://biccn.org/data). Visualization and
964 analysis tools are available at NeMO Analytics (Individual species:
965 https://nemoanalytics.org//index.html?layout_id=ac9863bf; Integrated species:
966 https://nemoanalytics.org//index.html?layout_id=34603c2b) and Cytosplore Viewer
967 (https://viewer.cytosplore.org/). These tools allow users to compare cross-species datasets and
968 consensus clusters via genome and cell browsers and calculate differential expression within and
969 among species. A semantic representation of the cell types defined through these studies is available in

970 the provisional Cell Ontology (https://bioportal.bioontology.org/ontologies/PCL; Supplementary Table
971 1).
972

973 Code availability

974 Code to reproduce figures will be available for download from
975 https://github.com/AllenInstitute/BICCN_M1_Evo.
976

977 Acknowledgements

978 We thank the Tissue Procurement, Tissue Processing and Facilities teams at the Allen Institute for
979 Brain Science for assistance with the transport and processing of postmortem and neurosurgical brain
980 specimens; the Technology team at the Allen Institute for assistance with data management; M.
981 Vawter, J. Davis and the San Diego Medical Examiner's Office for assistance with postmortem tissue
982 donations. We thank Ximena Opitz-Araya and Allen Institute for Brain Science Viral Technology team
983 for AAV packaging. We thank Lindsay Ng, Dijon Hill and Ram Rajanbabu for patching the human and
984 mouse cells in the figure describing chandelier neurons and Sara Kebede, Alice Mukora, Grace
985 Willams for reconstructing them. This work was funded by the Allen Institute for Brain Science and by
986 US National Institutes of Health grant U01 MH114812-02 to E.S.L. Support for the development of NS-
987 Forest v.2 and the provisional cell ontology was provided by the Chan–Zuckerberg Initiative DAF, an
988 advised fund of the Silicon Valley Community Foundation (2018-182730). G.Q. is supported by NSF
989 CAREER award 1846559. This work was partially supported by an NWO Gravitation project:
990 BRAINSCAPES: A Roadmap from Neurogenetics to Neurobiology (NWO: 024.004.012) and NWO
991 TTW project 3DOMICS (NWO: 17126). This project was supported in part by NIH grants
992 P51OD010425 from the Office of Research Infrastructure Programs (ORIP) and UL1TR000423 from
993 the National Center for Advancing Translational Sciences (NCATS). Its contents are solely the
994 responsibility of the authors and do not necessarily represent the official view of NIH, ORIP, NCATS,
995 the Institute of Translational Health Sciences or the University of Washington National Primate

Research Center. This work is supported in part by NIH BRAIN Initiative award RF1MH114126 from the National Institute of Mental Health to E.S.L., J.T.T., and B.P.L., NIH BRAIN Initiative award U19MH121282 to J.R.E., the National Institute on Drug Abuse award R01DA036909 to B.T., National Institute of Neurological Disorders and Stroke award R01NS044163 to W.J.S. and the California Institute for Regenerative Medicine (GC1R-06673-B) and the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation (2018–182730) to R.H.S. J.R.E. is an Investigator of the Howard Hughes Medical Institute. The authors thank the Allen Institute founder, Paul G. Allen, for his vision, encouragement and support.

Author contributions

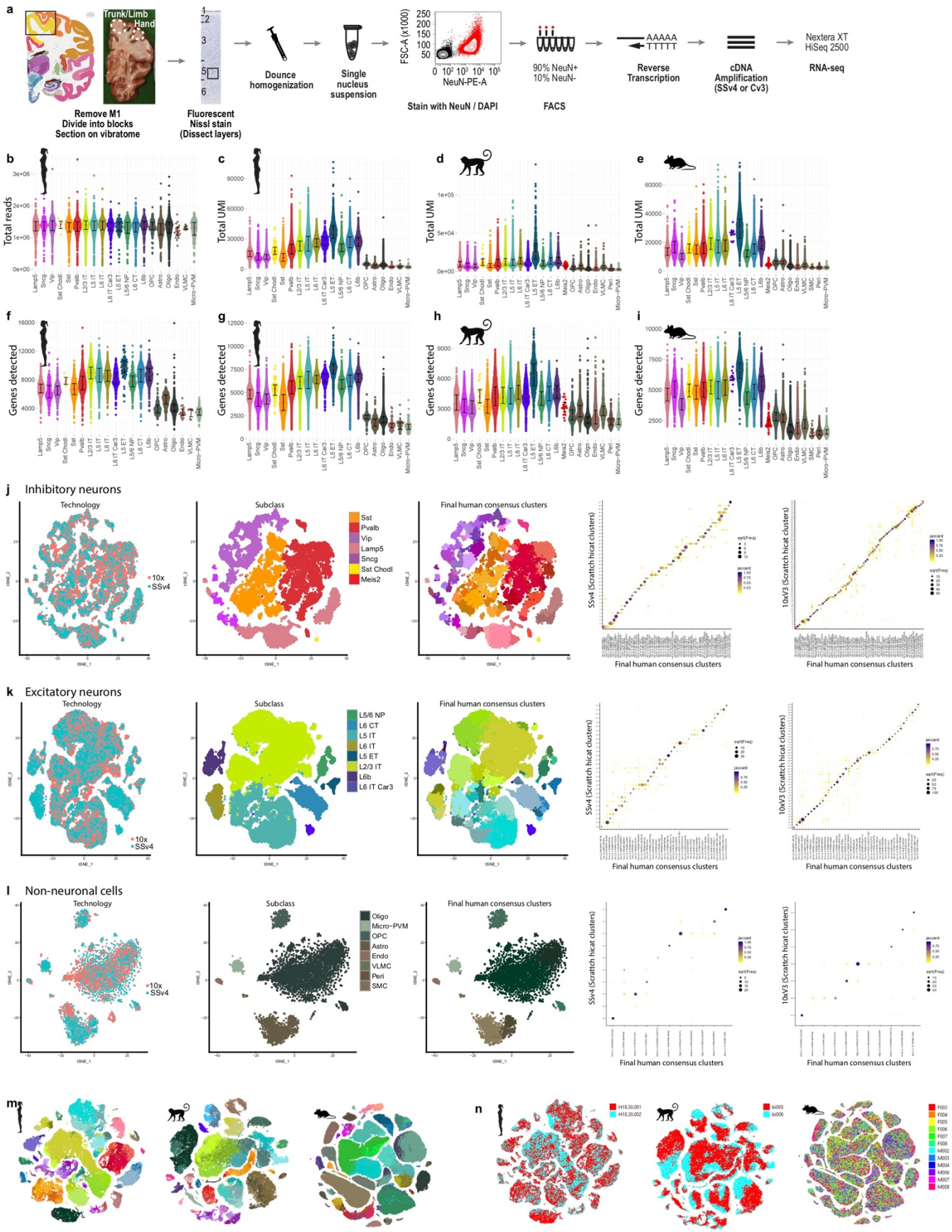
RNA data generation: A.M.Y., A.R., A.T., B.B.L., B.T., C.D.K., C.R., C.R.P., C.S.L., D.B., D.D., D.M., E.S.L., E.Z.M., F.M.K., G.F., H.T., H.Z., J.C., J.G., J.S., K.C., K.L., K.S., K.S., K.Z., M.G., M.K., M.T., N.D., N.M.R., N.P., R.D.H., S.A.M., S.D., S.L., T.C., T.E.B., T.P., W.J.R. mC data generation: A.B., A.I.A., A.R., C.L., H.L., J.R.E., J.R.N., R.G.C. ATAC data generation: A.E.S., B.B.L., B.R., B.T., C.R.P., C.S.L., D.D., J.C., K.Z., L.T.G., N.P., S.P., W.J.R., X.H., X.W. Electrophysiology, morphology, and Patch-seq data generation: A.L.K., B.E.K., D.M., E.S.L., G.D.H., J.G., J.T.T., K.S., M.T., N.D., S.A.S., S.O., T.L.D., T.P., W.J.S. Data archive and infrastructure: A.E.S., A.M., B.R.H., H.C.B., J.A.M., J.G., J.K., J.O., M.M., O.R.W., R.H., S.A.A., S.S., Z.Y. Cytosplore Viewer software: B.P.L., B.V.L., J.E., T.H. Data analysis: A.D., B.B.L., B.D.A., B.E.K., B.P.L., B.V.L., D.D., E.A.M., E.S.L., F.M.K., F.X., H.L., J.E., J.G., J.G., J.R.E., J.T.T., K.S., M.C., N.D., N.L.J., O.P., P.V.K., Q.H., R.F., R.H.S., R.Z., S.F., S.O., T.E.B., T.H., W.D., W.T., Y.E.L., Z.Y. Data interpretation: A.D., A.R., B.B.L., B.E.K., B.T., C.K., C.L., E.S.L., F.X., H.L., H.Z., J.G., J.G., J.R.E., J.T.T., M.C., M.H., N.D., N.L.J., P.R.H., P.V.K., Q.H., R.D.H., R.H.S., R.Z., S.D., S.O., T.E.B., W.T., Y.E.L., Z.Y. Writing manuscript: A.D., B.B.L., B.E.K., C.K., E.S.L., F.M.K., M.C., N.D., N.L.J., P.R.H., Q.H., R.H.S., T.E.B., W.J.S., W.T.

Competing interests

A.R. is an equity holder and founder of Celsius Therapeutics, a founder of Immunitas, and an SAB member in Syros Pharmaceuticals, Neogene Therapeutics, Asimov, and Thermo Fisher Scientific. B.R. is a shareholder of Arima Genomics, Inc. K.Z. is a co-founder, equity holder and serves on the Scientific Advisor Board of Singlera Genomics. P.V.K. serves on the Scientific Advisory Board to Celsius Therapeutics Inc.

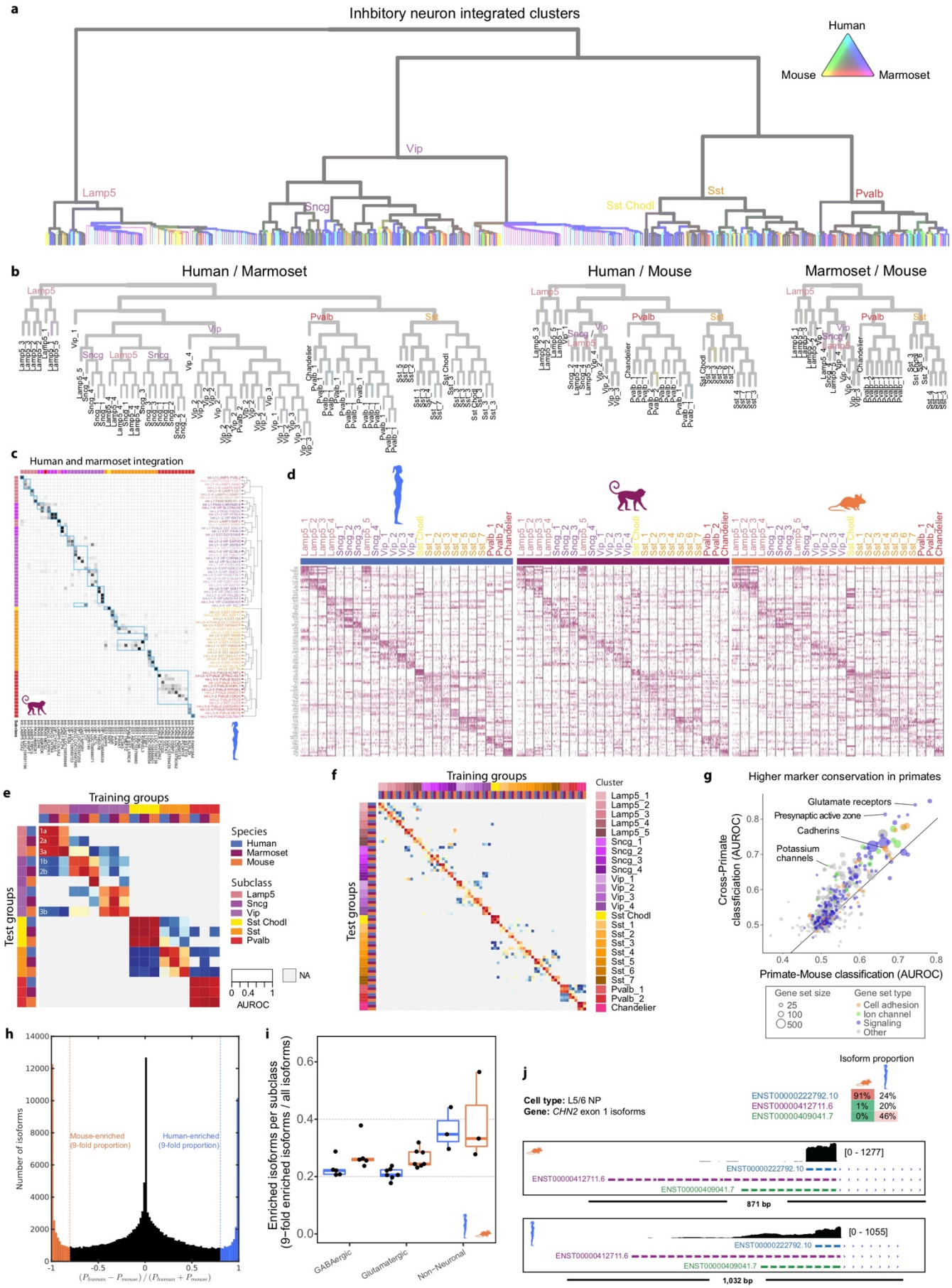
Materials & Correspondence

Correspondence and requests for materials should be addressed to E.S.L. and T.E.B.


a
Trunk/Limb
Hand
Remove M1
Divide into blocks
Section on vibratome
Fluorescent
Nissl stain
(Dissect layers)
Dounce
homogenization
Single
nucleus
suspension
FSC-A (x1000)
NeuN-PE-A
Stain with NeuN / DAPI
90% NeuN+
10% NeuN-
FACS
AAAAA
TTTTT
Reverse
Transcription
cDNA
Amplification
(SSv4 or Cv3)
Nextera XT
HiSeq 2500
RNA-seq
b
Total reads
c
Total UMI
d
Total UMI
e
Total UMI
f
Genes detected
g
Genes detected
h
Genes detected
i
Genes detected
j Inhibitory neurons
Technology
10x
SSv4
Subclass
Sst
Pvalb
Vip
Lamp5
Sncg
Sst Chodl
Meis2
Final human consensus clusters
SSv4 (Scratch hicat clusters)
10xV3 (Scratch hicat clusters)
k Excitatory neurons
L5/6 NP
L6 CT
L5 IT
L6 IT
L5 ET
L2/3 IT
L6b
L6 IT Car3
l Non-neuronal cells
Oligo
Micro-PVM
OPC
Astro
Endo
VLMC
Peri
SMC
m
n

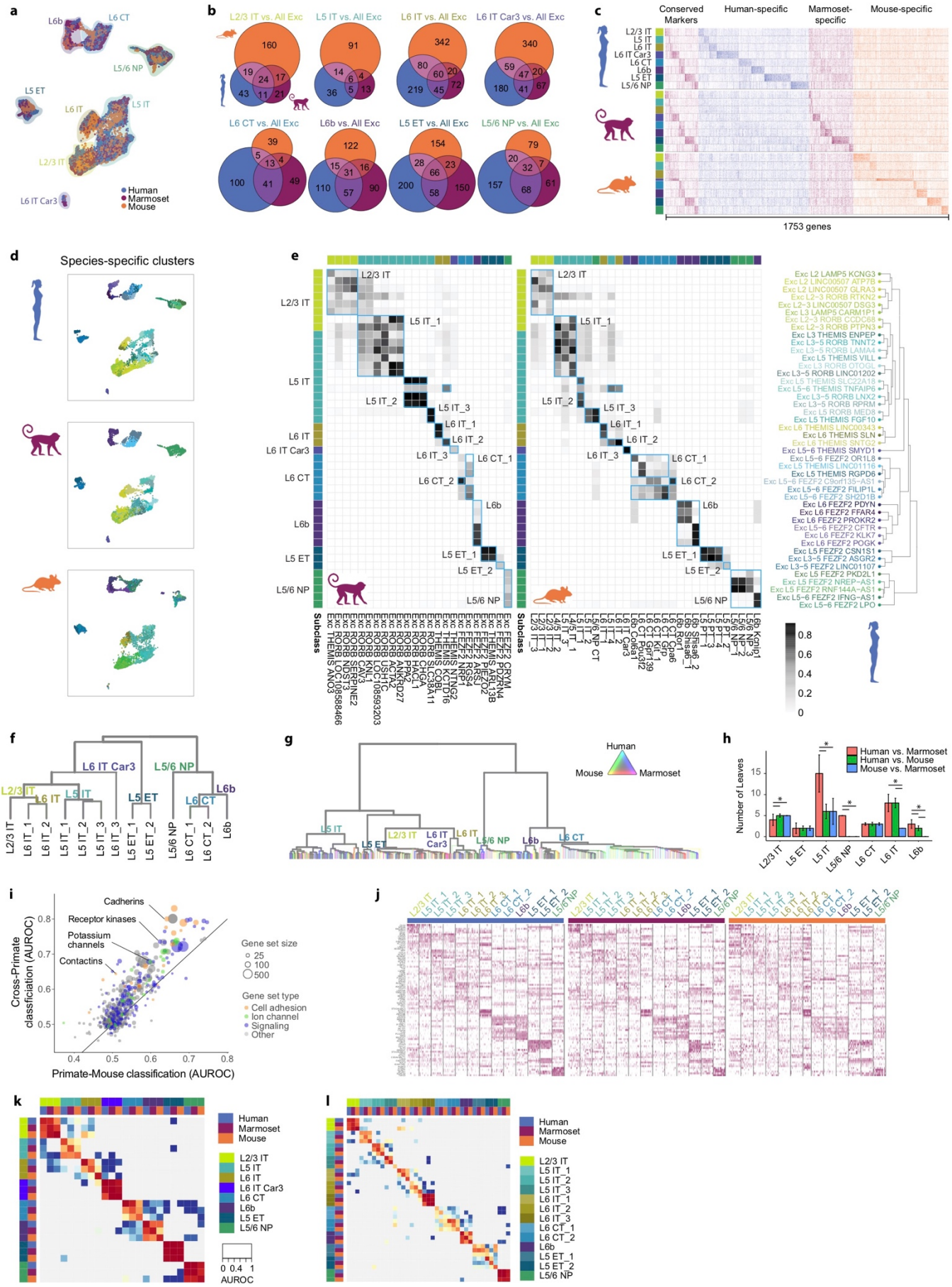
031

Extended Data Figure 1. RNA-seq quality metrics and integration of human datasets. a, Schematic of single-nucleus isolation from M1 of post-mortem human brain and profiling with RNA-seq. Box in the Nissl image highlights a cluster of Betz cells in L5. **b**, Using SSv4, > 1 million total reads were sequenced across all subclasses in human. **c-e**, Using Cv3, total unique molecular identifiers (UMI) varies between subclasses, and these differences are shared across species. **f-i**, Gene detection (expression > 0) is highest in human using SSv4 (**e**) and lowest for marmoset using Cv3 (**h**). Note that the average read depth used for SSv4 was approximately 20-fold greater than for Cv3 (target 60,000 reads per nucleus). **j-k**, tSNE projections of single nuclei based on expression of several thousand genes with variable gene expression and colored by cluster label (**j**) or donor (**k**). **l-n**, Integration of SSv4 and Cv3 RNA-seq datasets from human single nuclei isolated from GABAergic (**l**) and glutamatergic (**m**) neurons and non-neuronal cells (**n**). Left: UMAP visualizations colored by RNA-seq technology, cell subclass, and unsupervised consensus clusters. Right: Confusion matrices show membership of SSv4 and Cv3 nuclei within integrated consensus clusters.


a
Inhbitory neuron integrated clusters
Human
Mouse
Marmoset
Lamp5
Sncg
Vip
Sst Chodl
Sst
Pvalb
b
Human / Marmoset
Human / Mouse
Marmoset / Mouse
c
Human and marmoset integration
d
e
Training groups
Test groups
Species
Human
Marmoset
Mouse
Subclass
Lamp5
Sncg
Vip
Sst Chodl
Sst
Pvalb
NA
AUROC
f
Training groups
Test groups
Cluster
g
Higher marker conservation in primates
Glutamate receptors
Presynaptic active zone
Cadherins
Potassium channels
Cross-Primate classficiation (AUROC)
Primate-Mouse classification (AUROC)
Gene set size
25
100
500
Gene set type
Cell adhesion
Ion channel
Signaling
Other
h
Number of isoforms
Mouse-enriched (9-fold proportion)
Human-enriched (9-fold proportion)
i
Enriched isoforms per subclass (9-fold enriched isoforms / all isoforms)
GABAergic
Glutamatergic
Non-Neuronal
j
Isoform proportion
Cell type: L5/6 NP
Gene: CHN2 exon 1 isoforms
ENST00000222792.10
ENST00000412711.6
ENST00000409041.7
91%
24%
1%
20%
0%
46%
[0 - 1277]
871 bp
[0 - 1055]
1,032 bp

Extended Data Figure 2. RNA-seq integration of GABAergic neurons across species. a, Dendrogram of GABAergic neuron clusters from unsupervised clustering of integrated RNA-seq data from human, marmoset and mouse. Edge thickness indicates the relative number of nuclei, and edge color indicates species mixing (grey is well mixed). Major branches are labeled by subclass. Dendrogram shown in **Figure 2f** is derived from this tree based on pruning species-specific branches. **b,** Dendrograms of pairwise species integrations from **Figure 2g** with leaves labeled by cross-species clusters and edges colored by species mixing. **c,** Cluster overlap heatmap from human-marmoset pairwise Seurat integration showing the proportion of within-species clusters that coalesce within integrated clusters. Columns and rows are ordered as in **Figure 2e** with cross-species consensus clusters indicated by blue boxes. Top and left color bars indicate subclasses of within-species clusters. **d,** Heatmaps showing scaled expression of the top 5 marker genes for each GABAergic cross-species cluster, and 5 marker genes for *Lamp5* and *Sst*. Initial genes were identified by performing a Wilcox test of every integrated cluster against every other GABAergic nuclei. Additional DEGs were identified for *Lamp5* and *Sst* cross-species clusters, by comparing one of the cross-species clusters to all other related nuclei (e.g. *Sst*_1 against all other *Sst*). **e-f,** Heatmap of 1-vs-best MetaNeighbor scores for GABAergic subclasses (**e**) and clusters (**f**). Each column shows the performance for a single training group across the three test datasets. AUROCs are computed between the two closest neighbors in the test dataset, where the closer neighbor will have the higher score, and all others are shown in gray (NA). For example, in **e** the first column contains results of training on human *Lamp5*, labeled with numbers to indicate test datasets, where 1 is human, 2 is marmoset and 3 is mouse, and letters to indicate closest (**a**) and second-closest (**b**) neighboring groups. Dark red 3x3 blocks along the diagonal indicate high transcriptomic similarity across all three species. **g**, Scatter plot of MetaNeighbor analysis showing the performance (AUROC) of gene sets to classify GABAergic neuron consensus types by training with human or marmoset data and testing with the other species (Cross-Primate, y-axis) or training with primate data and testing with mouse (Primate-Mouse, x-axis). Gene set size and type are indicated by point size and color, respectively. **h**, Histogram of the relative difference in isoform genic proportion (P) between human and mouse for all subclass comparisons. All moderately to highly

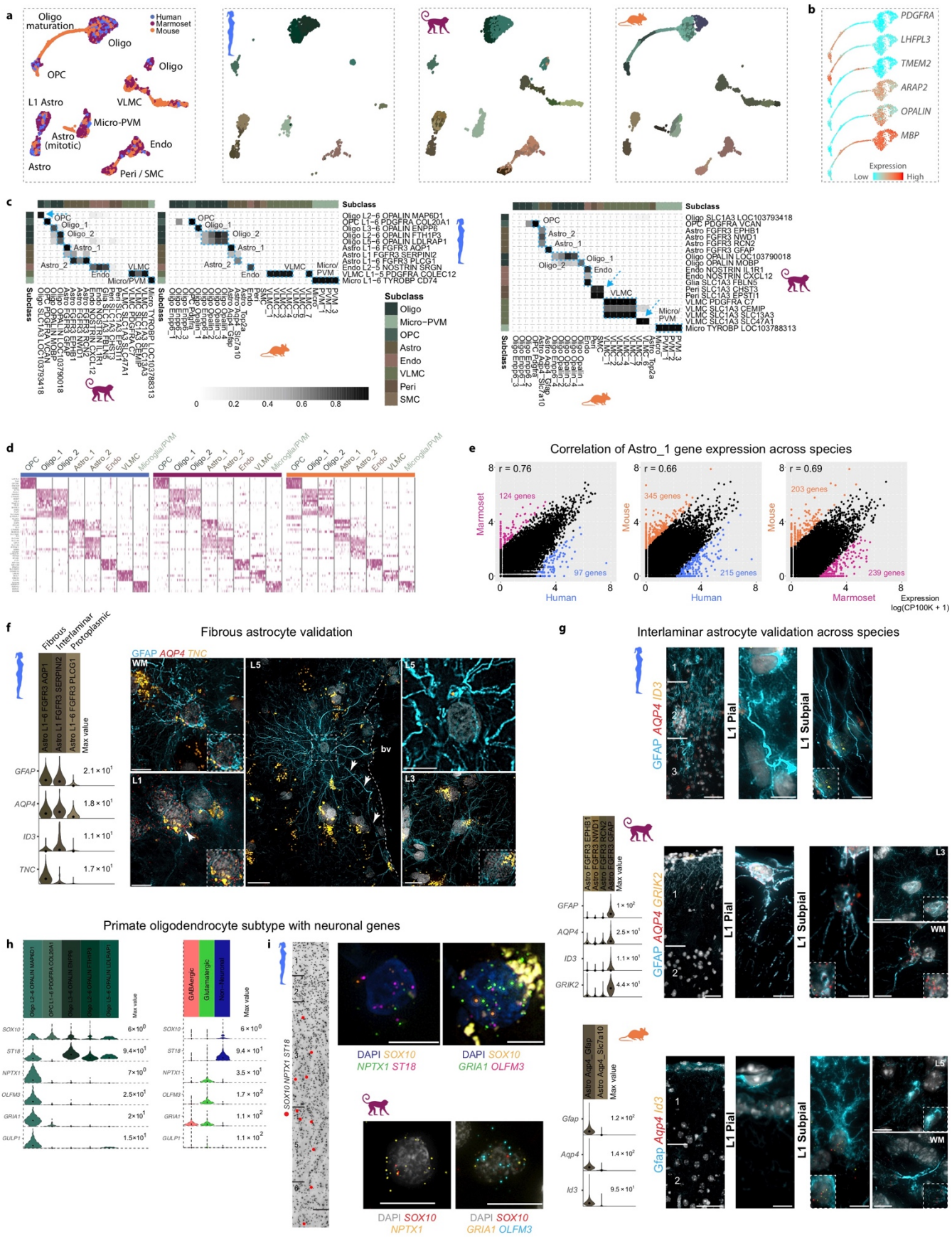
074 expressed isoforms were included (gene TPM > 10 in both species; isoform TPM > 10 and proportion >
075 0.2 in either species). Vertical lines indicate >9-fold change in mouse or human. **i**, Proportion of all
076 isoforms in **h** that switch between species (FDR P < 0.05; >9-fold change in P) summarized by
077 subclass and grouped by cell class. **j**, Comparison between species of isoform genic proportions for the
078 top three most common isoforms of Chimerin 2 (*CHN2*) expressed in the L5/6 NP subclass. Genome
079 browser tracks of RNA-seq (SSv4) reads in human and mouse at the *CHN2* locus.
080


a
L6b
L6 CT
L5/6 NP
L5 ET
L6 IT
L5 IT
L2/3 IT
L6 IT Car3
Human
Marmoset
Mouse
b
L2/3 IT vs. All Exc
L5 IT vs. All Exc
L6 IT vs. All Exc
L6 IT Car3 vs. All Exc
L6 CT vs. All Exc
L6b vs. All Exc
L5 ET vs. All Exc
L5/6 NP vs. All Exc
c
Conserved Markers
Human-specific
Marmoset-specific
Mouse-specific
1753 genes
d
Species-specific clusters
e
Subclass
f
g
Human
Mouse
Marmoset
h
Human vs. Marmoset
Human vs. Mouse
Mouse vs. Marmoset
Number of Leaves
i
Cadherins
Receptor kinases
Potassium channels
Contactins
Cross-Primate classficiation (AUROC)
Primate-Mouse classification (AUROC)
Gene set size
Gene set type
Cell adhesion
Ion channel
Signaling
Other
j
k
l
AUROC

081

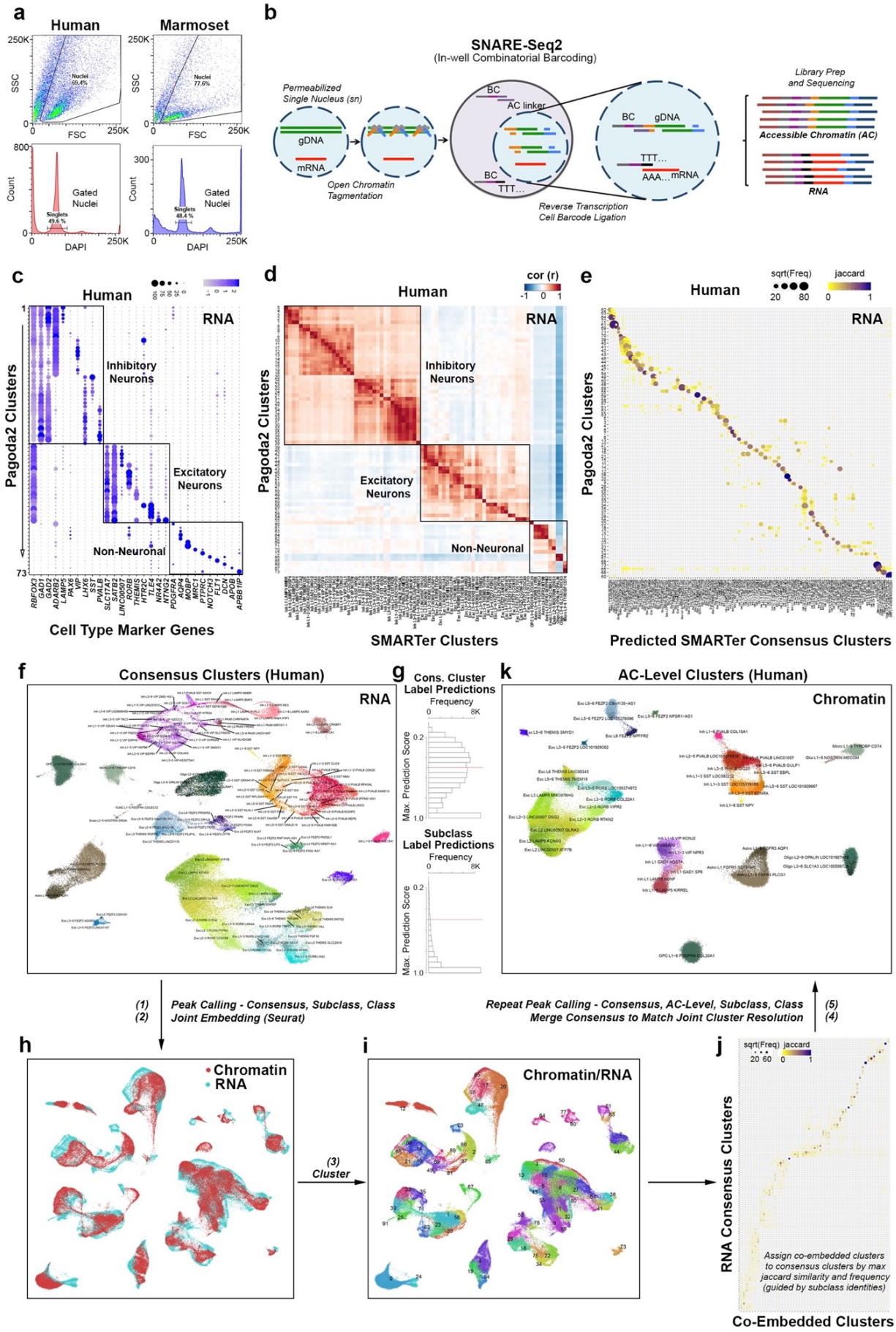
082 **Extended Data Figure 3. Glutamatergic neuron cell type homology across species. a,** UMAP
083 visualization of integrated snRNA-seq data from human, marmoset, and mouse glutamatergic neurons.
084 Highlighted colors indicate subclass. **b,** Venn diagrams indicating number of shared DEGs across
085 species by subclass. DEGs determined by ROC test of subclass against all other glutamatergic
086 subclasses within a species. **c,** Heatmap of all DEGs from **b** ordered by subclass and species
087 enrichment. Heatmap shows expression scaled by column for up to 50 randomly sampled nuclei from
088 each subclass for each species. **d,** UMAP visualization of integrated snRNA-seq data with projected
089 nuclei split by species. Colors indicate different within-species clusters. **e,** Cluster overlap heatmap
090 showing the proportion of within-species clusters that coalesce with a given integrated cross-species
091 cluster. Cross-species clusters are labelled and indicated by blue boxes with human-marmoset overlap
092 shown to the left and human-mouse overlap shown to the right. Top and left axes indicate the subclass
093 of a given within-species cluster by color. Bottom axis indicates marmoset (left) and mouse (right)
094 within species clusters. Right axis shows the glutamatergic branch of the human dendrogram from
095 **Figure 1c**. **f,** Dendrogram of glutamatergic neuron cross-species clusters. **g,** Unpruned dendrogram of
096 glutamatergic neuron clusters from unsupervised clustering of integrated RNA-seq data. Edge
097 thickness indicates the relative number of nuclei, and edge color indicates species mixing. Major
098 branches are labeled by subclass. **h**, Bar plots quantifying the number of well-mixed clusters from
099 unsupervised clustering of pairwise species integrations. Significant differences (adjusted $P < 0.05$,
100 Tukey's HSD test) between species are indicated for each subclass. **i,** Scatter plot of MetaNeighbor
101 analysis showing the performance (AUROC) of gene sets to classify glutamatergic neuron consensus
102 types by training with human or marmoset data and testing with the other species (Cross-Primate, y-
103 axis) or training with primate data and testing with mouse (Primate-Mouse, x-axis). Gene set size and
104 type are indicated by point size and color, respectively. **j,** Heatmaps showing scaled expression of
105 marker genes for each glutamatergic cross-species cluster. The top 5 marker genes for each cross-
106 species cluster are shown, with an additional 5 genes for L5 ET, L5 IT, and L6 IT. Initial genes were
107 identified by performing a Wilcox test of every integrated cluster against every other glutamatergic
108 nuclei. Additional DEGs were identified for L5 ET, L5 IT, and L6 IT cross-species clusters, by

comparing one of the cross-species clusters to all other related nuclei (e.g. L5 IT_1 against all other L5 IT). **k**, **l,** Heatmap of 1-vs-best MetaNeighbor scores for glutamatergic subclasses (**k**) and clusters (**l**). Results are displayed as in **Extended Data Fig. 2e,f**.

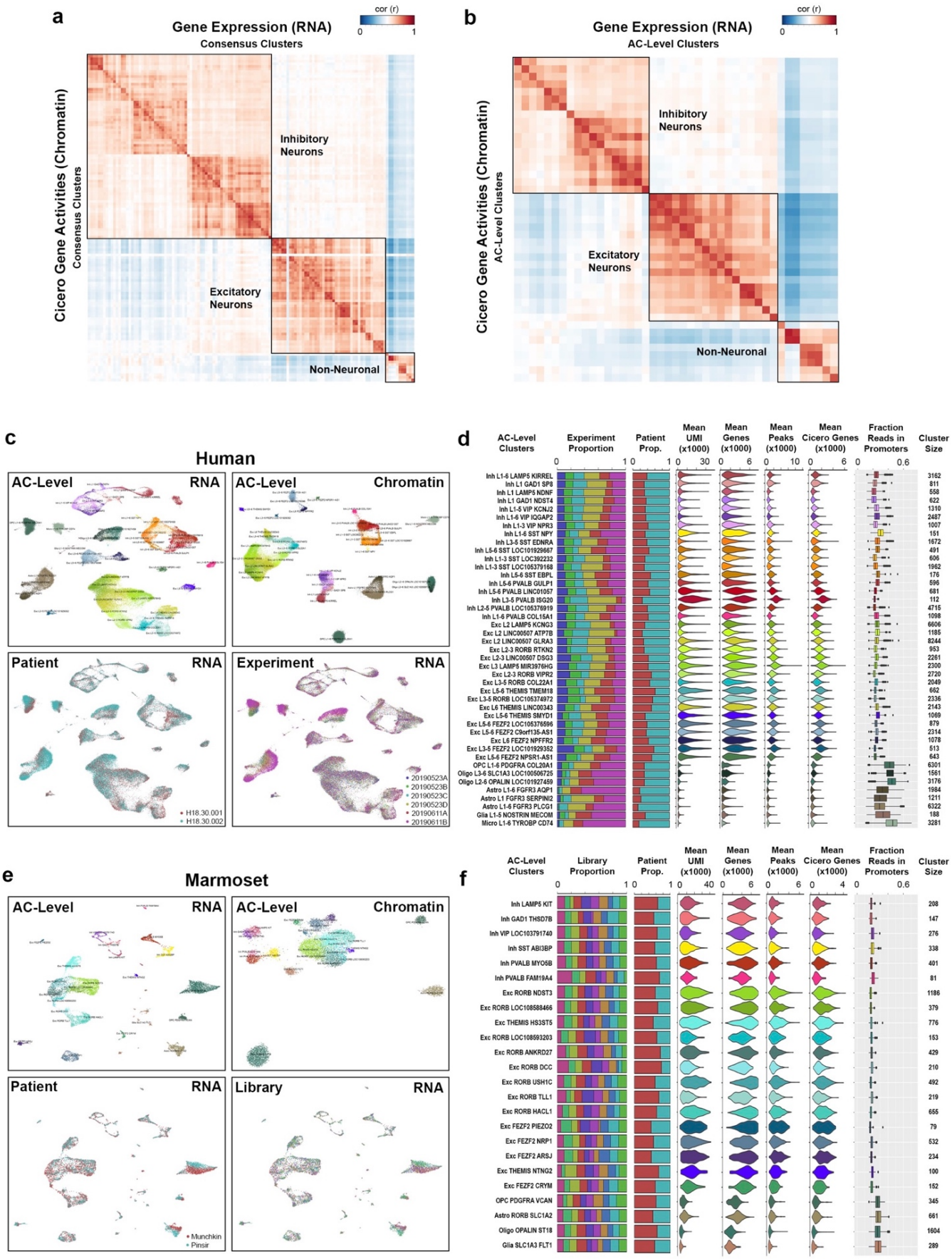

a
Oligo maturation
Human
Marmoset
Mouse
Oligo
OPC
L1 Astro
VLMC
Micro-PVM
Astro (mitotic)
Endo
Astro
Peri / SMC
b
PDGFRA
LHFPL3
TMEM2
ARAP2
OPALIN
MBP
Expression
Low
High
c
Subclass
Oligo
Micro−PVM
OPC
Astro
Endo
VLMC
Peri
SMC
d
e
Correlation of Astro_1 gene expression across species
r = 0.76
124 genes
97 genes
r = 0.66
345 genes
215 genes
r = 0.69
203 genes
239 genes
Marmoset
Mouse
Human
Expression
log(CP100K + 1)
f
Fibrous astrocyte validation
Fibrous
Interlaminar
Protoplasmic
GFAP AQP4 TNC
WM
L5
L1
L3
bv
Max value
GFAP
AQP4
ID3
TNC
g
Interlaminar astrocyte validation across species
GFAP AQP4 ID3
L1 Pial
L1 Subpial
GFAP AQP4 GRIK2
Gfap Aqp4 Id3
h
Primate oligodendrocyte subtype with neuronal genes
SOX10
ST18
NPTX1
OLFM3
GRIA1
GULP1
GABAergic
Glutamatergic
i
SOX10 NPTX1 ST18
DAPI SOX10 NPTX1 ST18
DAPI SOX10 GRIA1 OLFM3
DAPI SOX10 NPTX1
DAPI SOX10 GRIA1 OLFM3

114 **Extended Data Figure 4. Non-neuronal cell type homology across species. a**, UMAP plots of
115 integrated RNA-seq data for non-neuronal nuclei, colored by species and within-species clusters. Note
116 that some cell types are present in only one or two species. **b**, UMAP of mouse oligodendrocyte
117 precursors and mature cells showing expression levels of marker genes for different stages of cell
118 maturation. **c**, Heatmaps of the proportion of nuclei in each species-specific cluster that overlap in the
119 integrated RNA-seq analysis. Blue boxes define homologous cell types that can be resolved across all
120 three species. Arrows highlight clusters that overlap between two species and are not detected in the
121 third species, due to differences in sampling depth of non-neuronal cells, relative abundances of cell
122 types between species, or evolutionary divergence. **d**, Conserved marker genes for homologous cell
123 types across species. **e**, Pairwise comparisons between species of log-transformed gene expression of
124 the Astro_1 type. Colored points correspond to significantly differentially expressed (DE) genes (FDR <
125 0.01, log-fold change > 2). r, Spearman correlation. **f**, Fibrous astrocyte in situ validation. Violin plots of
126 marker genes of human astrocyte clusters that correspond to fibrous, interlaminar, and protoplasmic
127 types based on in situ labeling of types. Left ISH: Fibrous astrocytes located in the white matter (WM,
128 top) and a subset of L1 (bottom) astrocytes express the Astro L1-6 *FGFR3 AQP1* marker gene *TNC*.
129 Middle ISH: Image of putative varicose projection astrocyte located in cortical L5 adjacent to a blood
130 vessel (bv) and extending long GFAP-labeled processes (white arrows) does not express the marker
131 gene *TNC*. The white dashed box indicates the area shown at higher magnification in the top right
132 panel. Likewise, the L3 protoplasmic astrocyte shown in the bottom right panel does not express *TNC*.
133 **g**, Combined GFAP immunohistochemistry and RNAscope FISH for markers of L1 astrocytes in
134 human, mouse, and marmoset. In human (top), pial and subpial interlaminar astrocytes are labeled with
135 *AQP4* and *ID3* and extend long processes from L1 down to L3. In marmoset (middle), both pial and
136 subpial L1 astrocytes express *AQP4* and *GRIK2* and extend GFAP-labeled processes through L1 that
137 terminate before reaching L2. An image of a marmoset protoplasmic astrocyte located in L3 shows that
138 this astrocyte type does not express the marker gene *GRIK2*. A subset of marmoset fibrous astrocytes
139 located in the white matter (WM) express *GRIK2*, suggesting that fibrous and L1 astrocytes have a
140 shared gene expression signature as shown in human [2]. L1 astrocytes in mouse (bottom) consist of pial

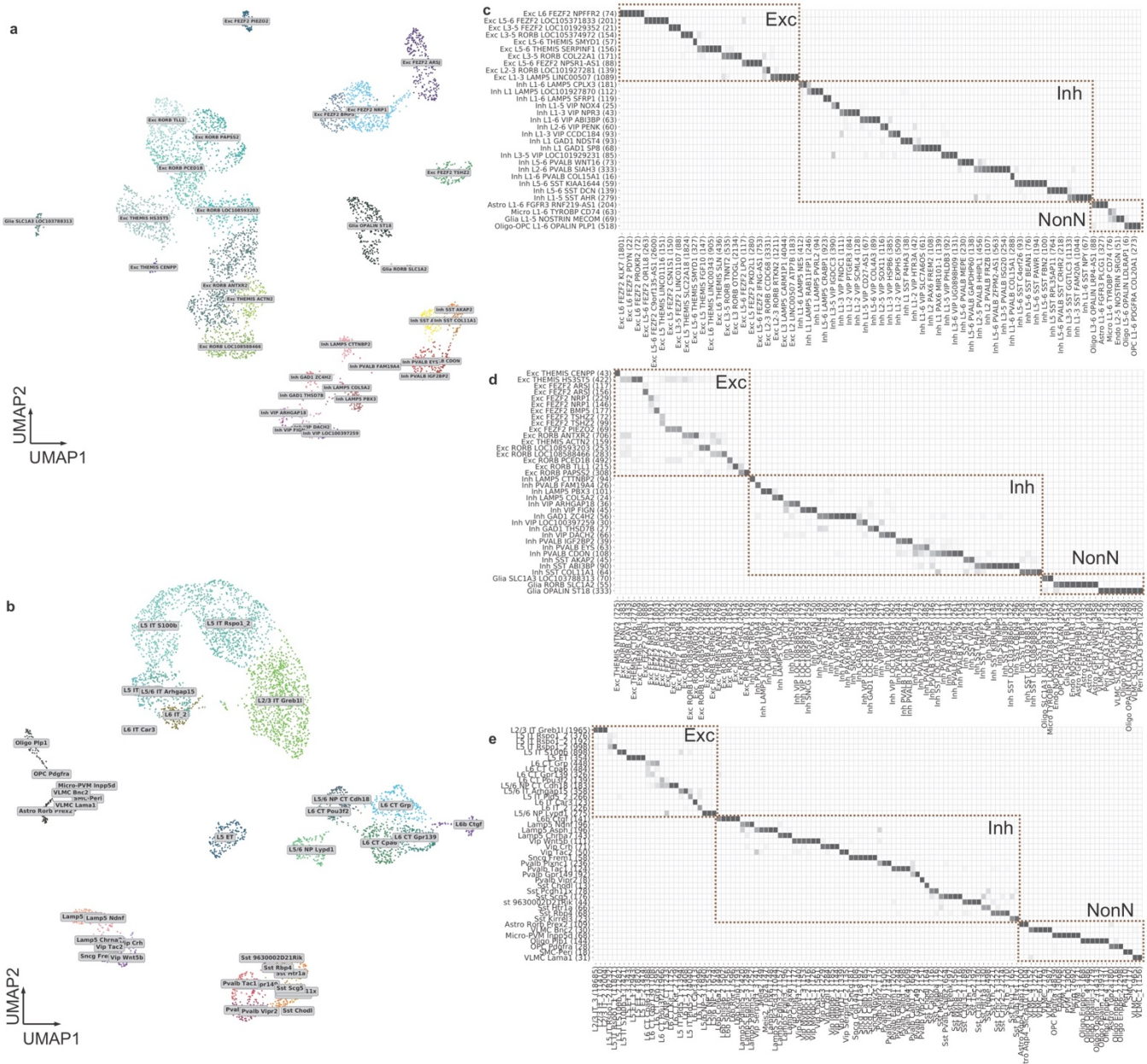
141 and subpial types that differ morphologically but are characterized by their expression of the genes
142 *Aqp4* and *Id3*. Pial astrocytes in mouse extend short Gfap-labeled processes that terminate within L1
143 whereas mouse subpial astrocytes appear to extend processes predominantly toward the pial surface.
144 Protoplasmic astrocytes (example shown in L5) do not express *Id3*, whereas fibrous astrocytes in
145 mouse share expression of *Id3* with L1 astrocyte types. Inset images outlined with white dashed boxes
146 illustrate cells in each of the accompanying images at higher magnification to show RNAscope spots for
147 each gene labeled. Scale bars, 20 μm. **h**, Violin plots of marker genes of oligodendrocyte lineage
148 clusters in human. Transcripts detected in the Oligo L2−6 *OPALIN MAP6D1* cluster include genes
149 expressed almost exclusively in neuronal cells. Scale bars, 20 μm. **i**, Left: Inverted DAPI image
150 showing a column of cortex labeled with markers of the human Oligo L2-6 *OPALIN MAP6D1* type. Red
151 dots show cells triple labeled with *SOX10*, *NPTX1*, and *ST18*. Top right: Examples of cells labeled with
152 marker gene combinations specific for the human Oligo L2-6 *OPALIN MAP6D1* type. Bottom right:
153 Example of a marmoset cell labeled with the marker genes *OLIG2* and *NRXN3*. Scale bars, 20 μm.
154

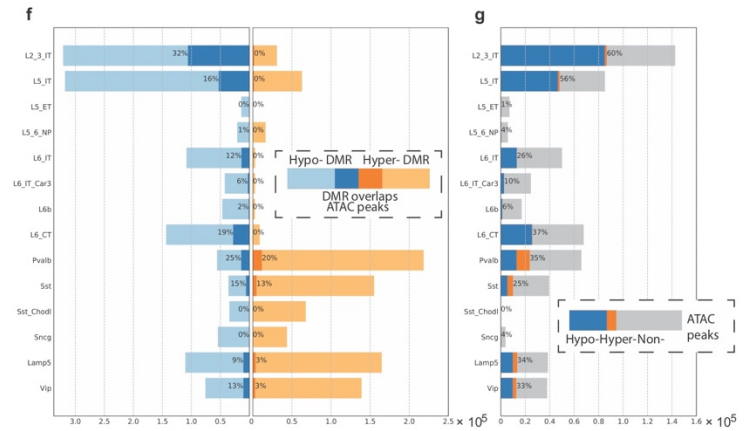

a
Human
Marmoset
SSC
FSC
Nuclei 69.4%
Nuclei 77.6%
Count
Gated Nuclei
Singlets 69.6 %
Singlets 48.4 %
DAPI
b
SNARE-Seq2
(In-well Combinatorial Barcoding)
Permeabilized Single Nucleus (sn)
gDNA
mRNA
Open Chromatin Tagmentation
BC
AC linker
TTT...
Reverse Transcription Cell Barcode Ligation
AAA...mRNA
Library Prep and Sequencing
Accessible Chromatin (AC)
RNA
c
Human
RNA
Pagoda2 Clusters
Inhibitory Neurons
Excitatory Neurons
Non-Neuronal
Cell Type Marker Genes
d
cor (r)
SMARTer Clusters
e
sqrt(Freq)
jaccard
Predicted SMARTer Consensus Clusters
f
Consensus Clusters (Human)
g
Cons. Cluster Label Predictions
Frequency
Max. Prediction Score
Subclass Label Predictions
k
AC-Level Clusters (Human)
Chromatin
(1) Peak Calling - Consensus, Subclass, Class
(2) Joint Embedding (Seurat)
Repeat Peak Calling - Consensus, AC-Level, Subclass, Class
Merge Consensus to Match Joint Cluster Resolution
(5)
(4)
h
Chromatin
RNA
(3) Cluster
i
Chromatin/RNA
j
RNA Consensus Clusters
Assign co-embedded clusters to consensus clusters by max jaccard similarity and frequency (guided by subclass identities)
Co-Embedded Clusters

156 **Extended Data Figure 5. SNARE-seq2 transcriptomic profiling resolves M1 cell types**. **a-b**, FACS
157 gating parameters used for sorting human and marmoset single nuclei (**a**) that were used for SNARE-
158 seq2 as outlined in (**b**), to generate both RNA and accessible chromatin (AC) libraries having the same
159 cell barcodes. **c**, Dot plot showing averaged marker gene expression values (log scale) and proportion
160 expressed for clusters identified in a preliminary analysis of SNARE-seq2 RNA using Pagoda2. **d**,
161 Correlation heatmap of averaged scaled gene expression values for Pagoda2 clusters against SSv4
162 clusters from the same M1 region. **e**, Jaccard similarity plot for cell barcodes grouped according to
163 Pagoda2 clustering compared against the predicted SSv4 consensus clustering. **f-k**, Overview of AC-
164 level cluster assignment using RNA-defined clusters indicating the five main steps of the process. **f**,
165 Consensus clusters visualized by UMAP on RNA expression data and that were used to independently
166 call peaks from AC data. **g**, Histograms showing maximum prediction scores for consensus cluster
167 (top) and subclass (bottom) labels from RNA data to corresponding accessibility data (cicero gene
168 activities). **h**, Consensus cluster peaks, as well as those identified from subclass and class level
169 barcode groupings, were combined and the corresponding peak by cell barcode matrix was used to
170 predict gene activity scores using Cicero for integrative RNA/AC analyses. UMAP shows joint
171 embedding of RNA and imputed AC expression values using Seurat/Signac. **i**, UMAP showing clusters
172 identified from the joint embedding (**h**). **j**, Jaccard similarity plot comparing cell barcodes either grouped
173 according to RNA consensus clustering or joint RNA/AC clustering (**i**). RNA consensus clusters were
174 merged to best match the cluster resolution achieved from co-embedded clusters. Chromatin peak
175 counts generated from peak calling independently on consensus, AC-level, subclass, and class
176 barcode groupings were used to generate a final peak by cell barcode matrix. **k**, Final AC-level clusters
177 visualized using UMAP.
178

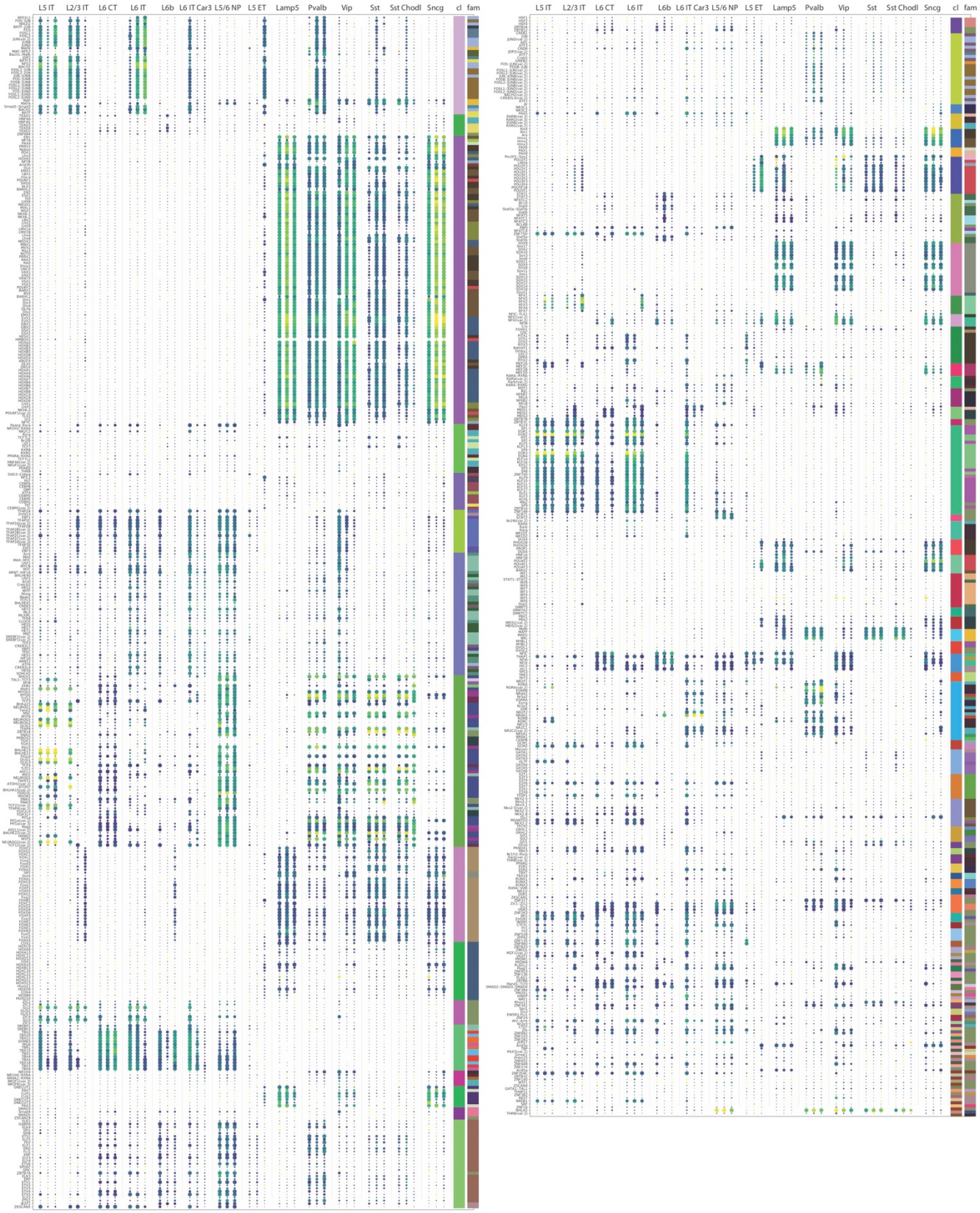

a
Gene Expression (RNA)
Consensus Clusters
cor (r)
0
1
Cicero Gene Activities (Chromatin)
Consensus Clusters
Inhibitory Neurons
Excitatory Neurons
Non-Neuronal
b
Gene Expression (RNA)
AC-Level Clusters
cor (r)
0
1
Cicero Gene Activities (Chromatin)
AC-Level Clusters
Inhibitory Neurons
Excitatory Neurons
Non-Neuronal
c
Human
AC-Level
RNA
AC-Level
Chromatin
Patient
RNA
Experiment
RNA
H18.30.001
H18.30.002
20190523A
20190523B
20190523C
20190523D
20190611A
20190611B
d
AC-Level Clusters
Experiment Proportion
Patient Prop.
Mean UMI (x1000)
Mean Genes (x1000)
Mean Peaks (x1000)
Mean Cicero Genes (x1000)
Fraction Reads in Promoters
Cluster Size
Inh L1-6 LAMP5 KIRREL 3162
Inh L1 GAD1 SP9 811
Inh L1 LAMP5 NDNF 558
Inh L1 GAD1 NDST4 622
Inh L1-5 VIP KCNJ2 1310
Inh L1-6 VIP IQGAP2 2487
Inh L1-3 VIP NPR3 1007
Inh L1-6 SST NPY 151
Inh L3-5 SST EDNRA 1672
Inh L5-6 SST LOC101929667 491
Inh L1-3 SST LOC392232 606
Inh L1-3 SST LOC105379168 1962
Inh L5-6 SST EBPL 176
Inh L5-6 PVALB GULP1 596
Inh L5-6 PVALB LINC01057 681
Inh L3-5 PVALB ISG20 112
Inh L2-5 PVALB LOC105376919 4715
Inh L1-6 PVALB COL15A1 1098
Exc L2 LAMP5 KCNG3 6606
Exc L2 LINC00507 ATP7B 1185
Exc L2 LINC00507 GLRA3 8244
Exc L2-3 RORB RTKN2 953
Exc L2-3 LINC00507 DSG3 2261
Exc L3 LAMP5 MIR3976HG 2300
Exc L2-3 RORB VIPR2 2720
Exc L3-5 RORB COL22A1 2049
Exc L5-6 THEMIS TMEM18 662
Exc L3-5 RORB LOC105374972 2336
Exc L6 THEMIS LINC00343 2143
Exc L5-6 THEMIS SMYD1 1069
Exc L5-6 FEZF2 LOC105376596 879
Exc L5-6 FEZF2 C9orf135-AS1 2314
Exc L6 FEZF2 NPFFR2 1078
Exc L3-5 FEZF2 LOC101929352 513
Exc L5-6 FEZF2 NPSR1-AS1 643
OPC L1-6 PDGFRA COL20A1 6301
Oligo L3-6 SLC1A3 LOC100506725 1561
Oligo L2-6 OPALIN LOC101927459 3176
Astro L1-6 FGFR3 AQP1 1984
Astro L1 FGFR3 SERPINI2 1211
Astro L1-6 FGFR3 PLCG1 6322
Glia L1-5 NOSTRIN MECOM 188
Micro L1-6 TYROBP CD74 3281
e
Marmoset
AC-Level
RNA
AC-Level
Chromatin
Patient
RNA
Library
RNA
Munchkin
Pinsir
f
AC-Level Clusters
Library Proportion
Patient Prop.
Mean UMI (x1000)
Mean Genes (x1000)
Mean Peaks (x1000)
Mean Cicero Genes (x1000)
Fraction Reads in Promoters
Cluster Size
Inh LAMP5 KIT 208
Inh GAD1 THSD7B 147
Inh VIP LOC103791740 276
Inh SST ABI3BP 338
Inh PVALB MYO5B 401
Inh PVALB FAM19A4 81
Exc RORB NDST3 1186
Exc RORB LOC108588466 379
Exc THEMIS HS3ST5 776
Exc RORB LOC108593203 153
Exc RORB ANKRD27 429
Exc RORB DCC 210
Exc RORB USH1C 492
Exc RORB TLL1 219
Exc RORB HACL1 655
Exc FEZF2 PIEZO2 79
Exc FEZF2 NRP1 532
Exc FEZF2 ARSJ 234
Exc THEMIS NTNG2 100
Exc FEZF2 CRYM 152
OPC PDGFRA VCAN 345
Astro RORB SLC1A2 661
Oligo OPALIN ST18 1604
Glia SLC1A3 FLT1 289

Extended Data Figure 6. SNARE-Seq2 quality statistics. a-b, Correlation heatmaps of average scaled gene expression values against average scaled Cicero gene activity values for consensus clusters (**a**) and AC-level clusters (**b**). **c**, UMAP plots showing human AC-level clusters for both RNA and chromatin data, as well as the corresponding patient and experiment identities for the RNA embeddings. **d**, Bar, violin and box plots for human AC-level clusters showing proportion contributed by each experiment or patient, mean UMI and genes detected from the RNA data, the mean peaks and cicero active genes detected from AC data, the fraction of reads found in promoters for AC data, and the number of nuclei making up each of the clusters. **e**, UMAP plots showing marmoset AC-level clusters for both RNA and chromatin data, as well as the corresponding patient and library identities for the RNA embeddings. **f**, Bar, violin and box plots for marmoset AC-level clusters showing proportion contributed by each library or patient, mean UMI and genes detected from the RNA data, the mean peaks and cicero active genes detected from AC data, the fraction of reads found in promoters for AC data, and the number of nuclei making up each of the clusters.

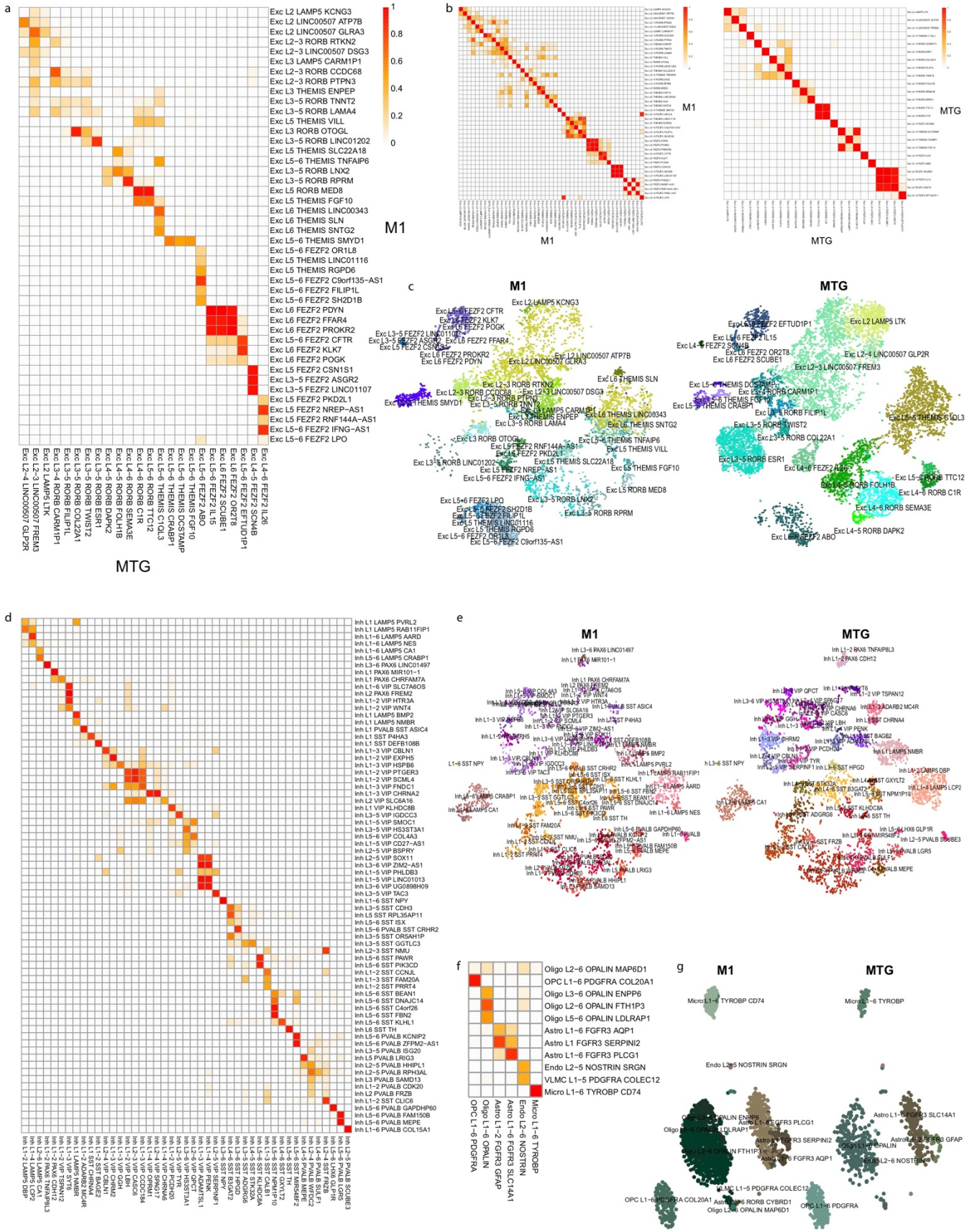

a
UMAP2
UMAP1
b
UMAP2
UMAP1
c
Exc
Inh
NonN
d
Exc
Inh
NonN
e
Exc
Inh
NonN


f
L2_3_IT
L5_IT
L5_ET
L5_6_NP
L6_IT
L6_IT_Car3
L6b
L6_CT
Pvalb
Sst
Sst_Chodl
Sncg
Lamp5
Vip
Hypo- DMR
Hyper- DMR
DMR overlaps ATAC peaks
g
Hypo-
Hyper-
Non-
ATAC peaks

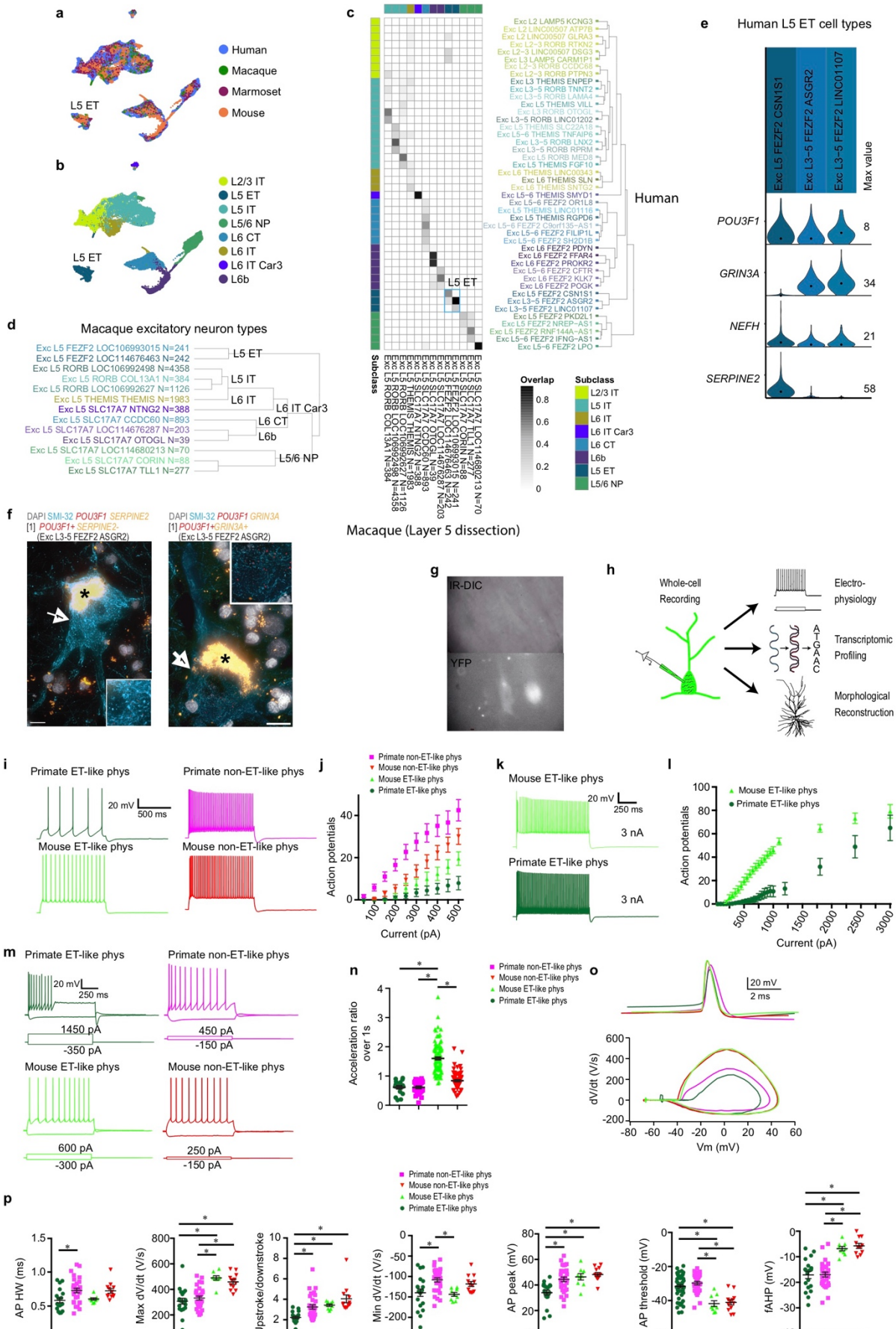
Extended Data Figure 7. DNA-methylation cell type and integration with RNA-seq data. a-b, UMAP visualization of marmoset M1 and mouse MOp DNA methylation (snmC-seq2) data and cell clusters. **c-e,** Mapping between DNAm-seq and RNA-seq clusters from human (**c**), marmoset (**d**), and mouse (**e**). Number of nuclei in each cluster are listed in parentheses. **f,** Numbers of hypo- and hyper-methylated DMRs and overlap with chromatin accessible peaks in each subclass of human. **g,** Numbers of chromatin accessible peaks and overlap with DMRs in each subclass of human.


L5 IT
L2/3 IT
L6 CT
L6 IT
L6b
L6 IT Car3
L5/6 NP
L5 ET
Lamp5
Pvalb
Vip
Sst
Sst Chodl
Sncg
cl
fam

204 **Extended Data Figure 8. TFBS enrichment analysis on hypo-methylated DMRs at subclass level**
205 **show conservativity of gene regulation across species.** Motif enrichment analysis of TFBS were
206 conducted using JASPAR's non-redundant core vertebrata TF motifs for neuronal subclasses in each
207 species. Each subclass tri-column shows the results of human, marmoset and mouse, respectively
208 from left to right. The size of a dot denotes the p-value of the corresponding motif, while the color
209 denotes the fold change. The rightmost two columns show TF clusters (cl) identified from motif profiles
210 and TF family (fam) identified from TF structures.
211


a
M1
MTG
b
M1
MTG
c
M1
MTG
d
e
M1
MTG
f
g
M1
MTG

Extended Data Figure 9. Cell type homologies between human cortical areas based on RNA-seq integration. a, Heatmap of glutamatergic neuron cluster overlap between M1 and MTG. **b**, Heatmaps of glutamatergic neuron cluster overlap for M1 and MTG test datasets. Clusters were split in half and two datasets were integrated using the same analysis pipeline as the M1 and MTG integration. Most clusters mapped correctly (along the diagonal) with some loss in resolution between closely related clusters (red blocks). **c**, tSNE plots of integrated glutamatergic neurons labeled with M1 and MTG clusters. **d-g**, Cluster overlap heatmaps and tSNE plots of integrations of GABAergic neurons (**d**, **e**) and non-neuronal cells (**f**, **g**), as described for glutamatergic neurons.


a
Human
Macaque
Marmoset
Mouse
L5 ET
b
L2/3 IT
L5 ET
L5 IT
L5/6 NP
L6 CT
L6 IT
L6 IT Car3
L6b
L5 ET
c
Human
L5 ET
Overlap
Subclass
L2/3 IT
L5 IT
L6 IT
L6 IT Car3
L6 CT
L6b
L5 ET
L5/6 NP
Macaque (Layer 5 dissection)
d
Macaque excitatory neuron types
Exc L5 FEZF2 LOC106993015 N=241
Exc L5 FEZF2 LOC114676463 N=242
Exc L5 RORB LOC106992498 N=4358
Exc L5 RORB COL13A1 N=384
Exc L5 RORB LOC106992627 N=1126
Exc L5 THEMIS THEMIS N=1983
Exc L5 SLC17A7 NTNG2 N=388
Exc L5 SLC17A7 CCDC60 N=893
Exc L5 SLC17A7 LOC114676287 N=203
Exc L5 SLC17A7 OTOGL N=39
Exc L5 SLC17A7 LOC114680213 N=70
Exc L5 SLC17A7 CORIN N=88
Exc L5 SLC17A7 TLL1 N=277
L5 ET
L5 IT
L6 IT
L6 IT Car3
L6 CT
L6b
L5/6 NP
e
Human L5 ET cell types
Exc L5 FEZF2 CSN1S1
Exc L3–5 FEZF2 ASGR2
Exc L3–5 FEZF2 LINC01107
Max value
POU3F1
8
GRIN3A
34
NEFH
21
SERPINE2
58
f
DAPI SMI-32 POU3F1 SERPINE2
[1] POU3F1+ SERPINE2-
(Exc L3-5 FEZF2 ASGR2)
DAPI SMI-32 POU3F1 GRIN3A
[1] POU3F1+GRIN3A+
(Exc L3-5 FEZF2 ASGR2)
g
IR-DIC
YFP
h
Whole-cell Recording
Electro-physiology
Transcriptomic Profiling
Morphological Reconstruction
i
Primate ET-like phys
Primate non-ET-like phys
Mouse ET-like phys
Mouse non-ET-like phys
20 mV
500 ms
j
Primate non-ET-like phys
Mouse non-ET-like phys
Mouse ET-like phys
Primate ET-like phys
Action potentials
Current (pA)
k
Mouse ET-like phys
20 mV
250 ms
3 nA
Primate ET-like phys
3 nA
l
Mouse ET-like phys
Primate ET-like phys
Action potentials
Current (pA)
m
Primate ET-like phys
Primate non-ET-like phys
20 mV
250 ms
1450 pA
-350 pA
450 pA
-150 pA
Mouse ET-like phys
Mouse non-ET-like phys
600 pA
-300 pA
250 pA
-150 pA
n
Primate non-ET-like phys
Mouse non-ET-like phys
Mouse ET-like phys
Primate ET-like phys
Acceleration ratio over 1s
o
20 mV
2 ms
dV/dt (V/s)
Vm (mV)
p
Primate non-ET-like phys
Mouse non-ET-like phys
Mouse ET-like phys
Primate ET-like phys
AP HW (ms)
Max dV/dt (V/s)
Upstroke/downstroke
Min dV/dt (V/s)
AP peak (mV)
AP threshold (mV)
fAHP (mV)

Extended Data Figure 10. Cross-species alignment of glutamatergic neurons and differences in L5 neuron spike trains and single spike properties. a, b, UMAP visualizations of cross-species integration of snRNA-seq data for glutamatergic neurons isolated from human, macaque (L5 dissection only), marmoset, and mouse. Colors indicate species (**a**) or cell subclass (**b**). **c**, Cluster overlap heatmap showing the proportion of nuclei from within-species clusters that are mixed within the same integrated clusters. Human clusters (rows) are ordered by the dendrogram reproduced from **Figure 1c**. Macaque clusters (columns) are ordered to align with human clusters. Color bars at top and left indicate subclasses of within-species clusters. Blue box denotes the L5 ET subclass. **d,** Dendrogram showing all macaque clusters from L5 dissection with subclasses denoted to the right. **e,** Violin plot showing expression of marker genes for human L5 ET neuron subtypes. **f,** Two examples of ISH labeled, SMI-32 IF stained Betz cells in L5 of human M1 that correspond to the L5 ET cluster Exc L3-5 *FEZF2 ASGR2*. Insets show higher magnification of ISH-labeled transcripts in corresponding cells. Scale bars, 20 µm. Asterisks mark lipofuscin. **g,** Example IR-DIC (top) and fluorescent (bottom) images obtained from a macaque organotypic slice culture. Note the inability to visualize the fluorescently labeled neurons in IR-DIC because of dense myelination. **h**, patch-seq involves the collection of morphological, physiological and transcriptomic data from the same neuron. Following electrophysiological recording and cell filling with biocytin via whole cell patch clamp, the contents of the cell are aspirated and processed for RNA-sequencing. This permits a transcriptomic cell type to be pinned on the physiologically-probed neuron. **i**, Example voltage responses to a 1 s, 500 pA step current injection. **j.** Action potentials as a function of current injection amplitude. Primate ET neurons display shallowest action potential-current injection relationship, perhaps partially because of their exceptionally low input resistance. **k**, Voltage responses to a 1 s, 3 nA step current injection. **l**, Action potentials as a function of current injection for a subset of experiments in which current injection amplitude was increased incrementally to 3 nA. While both mouse and primate ET neurons could sustain high firing rates, primate neurons required 3 nA of current over 1s to reach similar average firing rates as mouse ET neurons. **m**, Example voltage responses to 1 s depolarizing step current injections. The amplitude of the current injection was adjusted to produce ~10 spikes. Also shown are voltage responses to a

250 hyperpolarizing current injection. **n**, The firing rate of primate ET and IT neurons decreased during the 1
251 s step current injection, whereas, the firing rate of mouse ET neurons increased. Acceleration
252 ratio=2nd/last interspike interval. **o**, Example single action potentials (above) and phase plane plots
253 (below). **p**, Various action potential features are plotted as a function of cell type. Notably, action
254 potentials in primate ET neurons were reminiscent of fast spiking interneurons in that they were shorter
255 and more symmetrical compared with action potentials in other neuron types/species. Intriguingly, K^+
256 channel subunits $K_v3.1$ and $K_v3.2$ that are implicated in fast spiking physiology[89] are encoded by highly
257 expressed genes (*KCNC1* and *KCNC2*) in primate ET neurons (Fig. 7c) * $p < 0.05$, Bonferroni
258 corrected t-test.
259

Specimen ID	Age	Sex	Race	Cause of Death	PMI (hr)	Tissue RIN	Hemisphere Sampled	Data Type
H200.1023	43	F	Iranian descent	Mitral valve prolapse	18.5	7.4 ± 0.7	L	SSv4
H200.1025	50	M	Caucasian	Cardiovascular	24.5	7.6 ± 1.0	L	SSv4
H200.1030	54	M	Caucasian	Cardiovascular	25	7.7 ± 0.8	L	SSv4
H18.30.001	60	F	Unknown	Car accident	18	7.9 ± 2.5	R	SSv4, Cv3, SNARE-seq2, sn-methlyome
H18.30.002	50	M	Unknown	Cardiovascular	10	8.2 ± 0.4	R	SSv4, Cv3, SNARE-seq2, snmC-seq2

260 **Extended Data Table 1.** Summary of human tissue donors. RIN, RNA integrity number. Data type:
261 SMART-Seqv4 (SSv4), 10x Genomics Single Cell 3' Kit v3 (Cv3), Single-Nucleus Chromatin
262 Accessibility and mRNA Expression sequencing (SNARE-seq2), Single-nucleus methylcytosine
263 sequencing (snmC-seq2).
264

Specimen ID	Age (years)	Sex	Data Type
bi005	2.3	M	Cv3
bi006	3.1	F	Cv3
bi003	1.9	M	FISH

265 **Extended Data Table 2.** Summary of marmoset specimens. Data type: 10x Genomics Chromium
266 Single Cell 3' Kit v3 (Cv3). ACD Bio multiplex fluorescent in situ hybridization (FISH).
267

Supplementary Table legends

Supplementary Table 1. Provisional cell ontology (pCL) terms for human, mouse, and marmoset primary motor cortex cell types. Column headers are described as follows: pCL_id is a unique alphanumeric identifier assigned to each provisional cell type. CL_id is the Cell Ontology (CL) identifier for those parent cell type classes already represented in CL. pCL_name and Transcriptome data cluster are labels given according to each species naming convention that combines information about cortical layer enrichment and genes expressed in data cluster transcriptomes. TDC_id is a unique identifier assigned to the transcriptome data cluster. The part_of (uberon_id) and part_of (uberon_name) columns contain unique identifiers and names for tissue anatomic regions from which the experiment specimen was derived, in this case primary motor cortex. The is_a (CL or pCL_id) and is_a (CL or pCL_name) columns contain parent cell type or provisional cell type identifiers and names, respectively. Cluster_size indicates the number of single-nucleus or cell transcriptomes that were assigned membership to the transcriptome data cluster. Marker_gene_evidence indicates the number of marker genes that are necessary and sufficient to define the transcriptome cell type data cluster with maximal classification accuracy based on the NS-Forest v2.1 algorithm (see Supplementary Tables 4-6). F-measure_evidence is the f-beta score of classification accuracy from the NS-Forest v2.1 algorithm using the marker genes listed. The selectively_expresses column lists the minimum set of marker genes necessary and sufficient to define the transcriptome cell type data cluster. The definition brings together features to form a data driven ontological representation for each cell type cluster. The pCL annotations are available at https://github.com/mkeshk2018/Provisional_Cell_Ontology and https://bioportal.bioontology.org/ontologies/PCL.

Supplementary Table 2. Cluster annotations for human, marmoset, and mouse in separate worksheets. Cluster_label column identifies the RNA-seq cluster within each species. Cluster_size column denotes the number of nuclei that reside within each cluster (cluster_label). Class column identifies which cell class each cluster belongs to. Subclass column identifies which cell subclass each cluster belongs to. Cross-species cluster column indicates the cross-species consensus cluster

taxonomy. DNAm_cluster_label column identifies the transcriptomic cluster (cluster_label) that is aligned to DNAm-determined clusters. ATAC_cluster label column identifies the transcriptomic cluster (cluster_label) that is aligned to ATAC-determined clusters.

Supplementary Table 3. Application of Allen Institute nomenclature schema to mouse, marmoset, and human M1 taxonomies. The “taxonomy_ids” tab lists ids and descriptions for the 11 taxonomies included and which tab those taxonomies are shown on. The “preferred_aliases” tab shows a list of preferred aliases for linking between taxonomies, as well as descriptions for these. The next five tabs show nomenclatures for each of the taxonomies and have the following column headers: “tree_order” is the order shown in the tree (if any); “cell_set_alias”, “cell_set_label”, and “cell_set_accession” are unique identifiers, as described in the Allen Institute nomenclature page (https://portal.brain-map.org/explore/classes/nomenclature), with “cell_set_alias” including the names used in this manuscript; “cell_set_preferred_alias” indicates which clusters correspond to the “preferred_alias”es from the previous tab, if any; “cell_set_alias_integrated” shows linkages between single species transcriptomics taxonomies and the integrated taxonomy; “cell_set_labels_CS191213#" columns indicate linkages between cell sets in the transcriptomics and other modalities within a single species; “cell_set_descriptor” shows the type of cell set (or level of ontology); and “taxonomy_id” links to the “taxonomy_id” tab. Finally, the “Cell class hierarchy” tab shows the ordered class, level2, and subclass hierarchy and associated colors used as cell sets in previous tabs.

Supplementary Table 4. NS-Forest v2.1 was used to determine cell type cluster marker genes for all annotated levels of the human primary motor cortex cell type taxonomy defined by RNA-seq (Cv3). “clusterName” corresponds to the annotation label, either a cell type cluster name or a parent cell type class in the taxonomy. “markerCount” gives the optimal number of marker genes in the set that best discriminates the label. The “f-measure” column gives the f-beta score for classification using the set of markers. The next four columns “True Negative”, “False Positive”, “False Negative”, “True Positive” give

321 the confusion matrix for the label given the set of markers. Finally, “Marker 1-5” lists the gene symbols
322 corresponding to the optimal set of markers.
323
324 **Supplementary Table 5.** NS-Forest v2.1 was used to determine cell type cluster marker genes for all
325 annotated levels of the mouse primary motor cortex cell type taxonomy defined by RNA-seq (Cv3).
326 “clusterName” corresponds to the annotation label, either a cell type cluster name or a parent cell type
327 class in the taxonomy. “markerCount” gives the optimal number of marker genes in the set that best
328 discriminates the label. The “f-measure” column gives the f-beta score for classification using the set of
329 markers. The next four columns “True Negative”, “False Positive”, “False Negative”, “True Positive” give
330 the confusion matrix for the label given the set of markers. Finally, “Marker 1-5” lists the gene symbols
331 corresponding to the optimal set of markers.
332
333 **Supplementary Table 6.** NS-Forest v2.1 was used to determine cell type cluster marker genes for all
334 annotated levels of the marmoset primary motor cortex cell type taxonomy defined by RNA-seq (Cv3).
335 “clusterName” corresponds to the annotation label, either a cell type cluster name or a parent cell type
336 class in the taxonomy. “markerCount” gives the optimal number of marker genes in the set that best
337 discriminates the label. The “f-measure” column gives the f-beta score for classification using the set of
338 markers. The next four columns “True Negative”, “False Positive”, “False Negative”, “True Positive” give
339 the confusion matrix for the label given the set of markers. Finally, “Marker 1-5” lists the gene symbols
340 corresponding to the optimal set of markers.
341
342 **Supplementary Table 7.** DEGs determined by ROC test between each GABAergic neuron subclass
343 and all other GABAergic nuclei within each species. Columns are labeled myAUC, which contains AUC
344 scores > 0.7; avg_diff, which contains difference in expression between target subclass and all other
345 GABAergic neurons; power; pct.1, which indicates the percent of nuclei that express the gene in the
346 target cluster; pct.2, which indicates the percent of non-target nuclei that express the gene; cluster,

347 which denotes the target cluster; gene, indicating the gene that was identified as DE; and species,
348 which indicates the species the test was performed in.
349
350 **Supplementary Table 8.** List of DEGs (from Supplementary Table 7) that is sorted according to the
351 order the genes appear within the heatmap.
352
353 **Supplementary Table 9.** Supervised MetaNeighbor results, within- and across-species. Each row
354 corresponds to a unique entry for a given gene set and a given cell class, either Glutamatergic or
355 GABAergic. The first five columns provide information about the gene sets, namely their provenance
356 (SynGO or HGNC); numerical IDs; descriptive labels; manual classifications for plotting and
357 interpretation; and finally the number of genes included in the analysis (after subsetting to genes with 1-
358 1 orthologs across all three species). The sixth column indicates cell class. The remaining columns
359 contain MetaNeighbor AUROCs for various analyses: within_species_meanROC (column 7) provides
360 the mean of within-mouse (column 8), within-marmoset (column 9) and within-human (column 10)
361 performance. For each species, tests were run with random 3-fold cross-validation, and the average
362 across folds is reported. Columns 11 and 12 contain results from cross-species analyses, detailed in
363 the methods. Results are sorted by their AUROC across primates (column 12).
364
365 **Supplementary Table 10.** DEGs determined by ROC test between each glutamatergic neuron
366 subclass and all other glutamatergic nuclei within each species. Columns are labeled myAUC, which
367 contains AUC scores > 0.7; avg_diff, which contains difference in expression between target subclass
368 and all other glutamatergic neurons; power; pct.1, which indicates the percent of nuclei that express the
369 gene in the target cluster; pct.2, which indicates the percent of non-target nuclei that express the gene;
370 cluster, which denotes the target cluster; gene, indicating the gene that was identified as DE; and
371 species, which indicates the species the test was performed in.
372

373 **Supplementary Table 11.** List of DEGs (from Supplementary Table 10) that is sorted according to the
374 order the genes appear within the heatmap.
375
376 **Supplementary Table 12**. Average expression of isoforms in human and mouse subclasses and
377 estimates of isoform genic proportions (P) based on the ratio of isoform to gene expression. Isoforms
378 were included if they had at least moderate expression (TPM > 10) and P > 0.2 in either human or
379 mouse and at least moderate gene expression (TPM > 10) in both species.
380
381 **Supplementary Table 13.** SNARE-Seq2 metadata, cluster annotations and quality statistics. Tab 14a
382 indicates SNARE-Seq2 experiment level metadata (experiment name, library, patient, species,
383 purification, age, sex) and mapping statistics for RNA (mean UMI detected, mean genes detected) and
384 AC (mean fraction of reads in promoters or FRiP, mean uniquely mapped fragments grouped by 5000
385 base pair chromosomal bins, mean unique fragment counts per final peak locations, total number of
386 final nuclei). Tab 14b indicates the SNARE-Seq2 local RNA clusters for human M1 generated using
387 Pagoda2 (local cluster, annotated cluster name, broad cell type and abbreviation, k value used for
388 Pagoda2 clustering, broad cell type markers, level 1 and level2 classes and associated markers,
389 unique cluster markers). Tabs 14c-d indicates SNARE-Seq2 consensus or harmonized RNA and AC-
390 Level cluster annotations for human and marmoset M1, respectively, including annotated cluster name,
391 cluster order, associated subclass and class, and the number of datasets making up the clusters. Tabs
392 14e-f lists all metadata outlined in tabs 14a-d for all SNARE-Seq2 cell barcodes from human and
393 marmoset M1 samples, respectively.
394
395 **Supplementary Table 14.** SNARE-Seq2 differentially accessible regions for human and marmoset M1.
396 Tabs 15a and 15b show SNARE-Seq2 differentially accessible regions (DARs, q value < 0.001, log-fold
397 change > 1) identified by AC-Level clusters (15a) or subclass level (15b) for human M1, indicating for
398 each chromosomal location the p value (hypergeometric test), q value (Benjamini-Hochberg adjusted p
399 value), log-fold change and associated cluster or subclass. Tab 15b shows subclass DARs (q value <

400 0.001, log-fold change > 1) for marmoset subclasses as in tab 15b. Tab 15d shows a summarization of
401 human and marmoset DARs detected by matched subclasses, indicating actual number of DARs
402 detected (tabs 15b and 15c) and the values normalized to cluster size and total number of DARs
403 detected per species.
404
405 **Supplementary Table 15.** Cis-co-accessible sites, TF motif enrichments and differential TFBS
406 activities for human and marmoset M1. Tab 16a (human M1) and 16b (marmoset M1) show cis-
407 coaccessible network (CCAN) sites for subclass distinct markers genes (Wilcoxon Rank Sum test,
408 adjusted P < 0.05, average log-fold change > 0.5). pct.1 indicates the percent of nuclei that express the
409 gene in the target cluster, pct.2 indicates the percent of non-target nuclei that express the gene. For
410 each cluster and marker gene, corresponding motif enrichment values (hypergeometric test) for gene-
411 associated CCAN sites are shown ("observed" indicates number of features containing the motif,
412 "background" indicates the total number of features from a random selection of 40000 features that
413 contain the motif), and the motif associated differential chromVAR activity values identified using
414 logistic regression. The full list of chromVAR differentially active TFBS activities are also provided. Tab
415 16c summarizes the number of CCAN-associated marker genes, associated TFBSs enriched and or
416 active by subclass for both human and marmoset M1. Tabs 16d and 16e show cis-co-accessible sites,
417 TFBS enrichments and differential activities by AC-level clusters for human and marmoset M1,
418 respectively, similar to that provided in tabs 16a and 16b. Tab 16f shows chromVAR differentially active
419 TFBS activities by consensus or harmonized cluster using logistic regression. Tabs 16g, 16h, and 16i
420 show cis-co-accessible sites, TF motif enrichments and differential TFBS activities for human,
421 marmoset and mouse M1 ChCs compared against BCs.
422
423 **Supplementary Table 16**. snmC-seq2 metadata. The table shows experiment level metadata,
424 including species, sample name, gender, purification information, experiment nuclei numbers and pass-
425 QC nuclei numbers.
426

427 **Supplementary Table 17**. Subclass TFBS enrichment results. TFBS enrichment analysis was done
428 with AME [77] using JASPAR2020 motifs . Within a species, hypo-methylated DMRs in each subclass
429 were tested against hypo-methylated DMRs of all the other subclasses (background). DMRs and 250bp
430 around regions were used in the analysis. This table includes p-values and effect sizes (log2(TP/FP)) of
431 the analysis results.
432
433 **Supplementary Table 18**. Subclass TFBS enrichment at TF cluster level. TFs in SI Tab 18 were
434 grouped using clusters defined in Ref [42]. The table lists the most significant p-values and the largest
435 effect size of each TF cluster group.
436
437 **Supplementary Table 19.** DEGs determined by ROC test between chandelier cells and basket cells
438 within each species. Columns are labeled as species, with true/false values indicating if a gene was
439 enriched in chandelier cells for that species.
440
441 **Supplementary Table 20.** DEGs determined by ROC test between L5 ET subclass and L5 IT subclass
442 within each species. Columns are labeled as species, with values of 1 indicating a gene was enriched
443 in the L5 ET subclass for that species. A value of 0 indicates that the gene was not enriched in the L5
444 ET subclass for that species.
445
446 **Supplementary Table 21.** Genes with expression enrichment in L5 ET versus L5 IT that decreases
447 with evolutionary distance from human (human > macaque > marmoset > mouse). Columns are labeled
448 by species, and values indicate the log-fold change between L5 ET and L5 IT for that species. Genes
449 were included if they had a minimum log-fold change equal to 0.5 in human.