Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse — **Source link** ↗

Trygve E. Bakken, Nikolas L. Jorstad, Qiwen Hu, Wei Tian ...+99 more authors

Institutions: Allen Institute for Brain Science, Harvard University, Salk Institute for Biological Studies, Cold Spring Harbor Laboratory ...+16 more institutions

Related papers:

- Shared and distinct transcriptomic cell types across neocortical areas

- Conserved cell types with divergent features in human versus mouse cortex.

- Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq

- Adult mouse cortical cell taxonomy revealed by single cell transcriptomics

- Comprehensive Integration of Single-Cell Data.

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

2

3 Trygve E. Bakken, Nikolas L. Jorstad, Qiwen Hu, Blue B. Lake, Wei Tian, Brian E. Kalmbach, Megan

4 Crow, Rebecca D. Hodge, Fenna M. Krienen, Staci A. Sorensen, Jeroen Eggermont, Zizhen Yao, Brian

5 D. Aevermann, Andrew I. Aldridge, Anna Bartlett, Darren Bertagnolli, Tamara Casper, Rosa G.

6 Castanon, Kirsten Crichton, Tanya L. Daigle, Rachel Dalley, Nick Dee, Nikolai Dembrow, Dinh Diep,

7 Song-Lin Ding, Weixiu Dong, Rongxin Fang, Stephan Fischer, Melissa Goldman, Jeff Goldy, Lucas T.

8 Graybuck, Brian R. Herb, Xiaomeng Hou, Jayaram Kancherla, Matthew Kroll, Kanan Lathia, Baldur van

9 Lew, Yang Eric Li, Christine S. Liu, Hanqing Liu, Jacinta D. Lucero, Anup Mahurkar, Delissa McMillen,

10 Jeremy A. Miller, Marmar Moussa, Joseph R. Nery, Philip R. Nicovich, Joshua Orvis, Julia K. Osteen,

11 Scott Owen, Carter R. Palmer, Thanh Pham, Nongluk Plongthongkum, Olivier Poirion, Nora M. Reed,

12 Christine Rimorin, Angeline Rivkin, William J. Romanow, Adriana E. Sedeño-Cortés, Kimberly Siletti,

13 Saroja Somasundaram, Josef Sulc, Michael Tieu, Amy Torkelson, Herman Tung, Xinxin Wang,

14 Fangming Xie, Anna Marie Yanny, Renee Zhang, Seth A. Ament, M. Margarita Behrens, Hector

15 Corrada Bravo, Jerold Chun, Alexander Dobin, Jesse Gillis, Ronna Hertzano, Patrick R. Hof, Thomas

16 Höllt, Gregory D. Horwitz, C. Dirk Keene, Peter V. Kharchenko, Andrew L. Ko, Boudewijn P. Lelieveldt,

17 Chongyuan Luo, Eran A. Mukamel, Sebastian Preissl, Aviv Regev, Bing Ren, Richard H.

18 Scheuermann, Kimberly Smith, William J. Spain, Owen R. White, Christof Koch, Michael Hawrylycz,

19 Bosiljka Tasic, Evan Z. Macosko, Steven A. McCarroll, Jonathan T. Ting, Hongkui Zeng, Kun Zhang,

20 Guoping Feng, Joseph R. Ecker, Sten Linnarsson, Ed S. Lein

21

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

23

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

49   distinguished on the basis of regions of open chromatin or DNA methylation [5,9,10]. Furthermore, several

50   recent studies have shown that transcriptomically-defined cell types can be aligned across species [2,11–

51   13], indicating that these methods provide a path to quantitatively study evolutionary conservation and

52   divergence at the level of cell types. However, application of these methods has been highly

53   fragmented to date. Human and mouse comparisons have been performed in different cortical regions,

54   using single-cell (with biases in cell proportions) versus single-nucleus (with biases in transcript

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

56   independently.

57

58   The primary motor cortex (MOp in mouse, M1 in human and non-human primates, all referred to as M1

59   herein) provides an ideal cortical region to address questions about cellular evolution in rodents and

60   primates by integrating these approaches. Unlike the primary visual cortex (V1), which is highly

61   specialized in primates, or frontal and temporal association areas, whose homologues in rodents

62   remain poorly defined, M1 is essential for fine motor control and is functionally conserved across

63   placental mammals. M1 is an agranular cortex, lacking a defined L4, although neurons with L4-like

64   properties have been described [14]. L5 of carnivore and primate M1 contains exceptionally large

65   "giganto-cellular" corticospinal neurons (Betz cells in primates [15,16] that contribute to the pyramidal tract

66   and are highly specialized for their unusually large size with distinctive "taproot"-style dendrites [17,18].

67   Extracellular recordings from macaque corticospinal neurons reveal distinctive action potential

68   properties supportive of a high conduction velocity and similar, unique properties have been reported

69   during intracellular recordings from giganto-cellular neurons in cats[19–21]. Additionally, some primate Betz

70   cells directly synapse onto alpha motor neurons, whereas in cats and rodents these neurons synapse

71   instead onto spinal interneurons [22,23]. These observations suggest that Betz cells possess specialized

72   intrinsic mechanisms to support rapid communication, some of which are primate specific.

73

74 Conservation of cellular features across species is strong evidence for evolutionary constraints on

75 important cellular function. To explore evolutionary conservation and divergence of the M1 cellular

76 makeup and its underlying molecular and gene regulatory mechanisms, we combined saturation

77 coverage single-nucleus transcriptome analysis, DNA methylation, and combined open chromatin and

78 transcriptome analysis of mouse, marmoset, and human M1 and transcriptomic profiling of macaque

79 M1 L5. We describe a robust classification of neuronal and non-neuronal cell types in each species that

80 is highly consistent between the transcriptome and epigenome. Cell type alignment accuracy and

81 similarity varied as a function of evolutionary distance, with human more similar to non-human primate

82 than to mouse. We derived a consensus mammalian classification with globally similar cellular diversity,

83 varying proportions, and species specializations in gene expression between conserved cell classes.

84 Few genes had conserved cell type-specific expression across species and likely contribute to other

85 conserved cellular properties, such as the unique morphology of chandelier GABAergic neurons.

86 Conversely, these data also allow a targeted search for genes responsible for species specializations

87 such as the distinctive anatomy, physiology and axonal projections of Betz cells, large corticospinal

88 neurons in primates that are responsible for voluntary fine motor control. Together these findings

89 highlight the strength of a comparative approach to understand cortical cellular diversity and identify

90 conserved and specialized gene and gene regulatory mechanisms underlying cellular identity and

91 function.

92

93 We made all primary and analyzed data publicly available. Raw sequence data are available for

94 download from the Neuroscience Multi-omics Archive (nemoarchive.org) and the Brain Cell Data

95 Center (biccn.org/data). Visualization and analysis tools are available at NeMO Analytics

96 (nemoanalytics.org) and Cytosplore Viewer (viewer.cytosplore.org). These tools allow users to compare

97 cross-species datasets and consensus clusters via genome and cell browsers and calculate differential

98 expression within and among species. A semantic representation of the cell types defined through

99   these studies is available in the provisional Cell Ontology

100   (https://bioportal.bioontology.org/ontologies/PCL; Supplementary Table 1).

101

102   ## Results

103   **Multi-omic cell type taxonomies**

104   To characterize the molecular diversity of M1 neurons and non-neuronal cells, we applied multiple

105   single-nucleus transcriptomic (plate-based SMART-seq v4, SSv4, and droplet-based Chromium v3,

106   Cv3, RNA-sequencing) and epigenomic (single-nucleus methylcytosine sequencing 2, snmC-seq2;

107   single-nucleus chromatin accessibility and mRNA expression sequencing, SNARE-seq2) assays on

108   isolated M1 samples from human (Extended Data Fig. 1a), marmoset, and mouse brain. Cellular

109   diversity was also profiled selectively in M1 L5 from macaque monkeys using Cv3 (Fig. 1b) to allow

110   Patch-seq mapping in physiology experiments. M1 was identified in each species based on its

111   stereotyped location in the caudal portion of frontal cortex and histological features such as the

112   presence of exceptionally large pyramidal neurons in L5 of M1, classically known as Betz cells in

113   human, other primates, and carnivores (Fig. 1a; [17]). Single nuclei were dissociated, sorted, and

114   transcripts were quantified with Cv3 for deep sampling in all four species, and additionally using SSv4

115   in human and mouse for full-length transcript information. For human and a subset of mouse nuclei,

116   individual layers of M1 were profiled independently using SSv4. Whole-genome DNA methylation, and

117   open chromatin combined with transcriptome measurements, were quantified in single nuclei from a

118   subset of species (Fig. 1b). Mouse datasets are also reported in a companion paper [6]. Median neuronal

119   gene detection was higher in human using SSv4 (7296 genes) than Cv3 (5657), partially due to 20-fold

120   greater read depth, and detection was lower in marmoset (4211) and mouse (5046) using Cv3

121   (Extended Data Fig. 1b-i).

122

123 For each species, a diverse set of neuronal and non-neuronal cell type clusters were defined based on

124 unsupervised clustering of snRNA-seq datasets (cluster metadata in Supplementary Table 2). Human

125 SSv4 and Cv3 data were integrated based on shared co-expression using Seurat [24], and 127 clusters

126 were identified that included nuclei from both RNA-seq platforms (Extended Data Fig. j-l). Marmoset

127 clusters (94) were determined based on independent clustering of Cv3 data using a similar analysis

128 pipeline. Mouse clusters (116) were defined in a companion paper [6] using seven integrated

129 transcriptomics datasets. These differences in the number of clusters are likely due to a combination of

130 statistical methodological differences as well as sampling and data quality differences rather than true

131 biological differences in cell diversity. For example, more non-neuronal nuclei were sampled in mouse

132 (58,098) and marmoset (21,189) compared to human (4,005), resulting in greater non-neuronal

133 resolution in those species. t-SNE visualizations of transcriptomic similarities across nuclei revealed

134 well-separated clusters in all species and mixing among donors, with some donor-specific technical

135 effects in marmoset (Extended Data Fig. 1m,n).

136

137 Post-clustering, cell types were organized into taxonomies based on transcriptomic similarities and

138 given a standardized nomenclature (Supplementary Table 3). As described previously for a different

139 cortical region [2], taxonomies were broadly conserved across species and reflected different

140 developmental origins of major non-neuronal and neuronal classes (e.g. GABAergic neurons from

141 ganglionic eminences (GEs) versus glutamatergic neurons from the cortical plate) and subclasses (e.g.

142 GABAergic CGE-derived *Lamp5/Sncg* and *Vip* versus MGE-derived *Pvalb* and *Sst*), allowing

143 identification and naming of these subclasses across species. Consequently, cardinal cell subclass

144 properties can be inferred, such as intratelencephalic (IT) projection patterns. Greater species variation

145 was seen at the highest level of resolution (cell types) that are named based on transcription data in

146 each species including the layer (if available), major class, subclass marker gene, and most specific

147 marker gene (e.g. L3 Exc *RORB OTOGL* in human; additional markers in Supplementary Tables 4-6).

148 GABAergic types were uniformly rare (< 4.5% of neurons), whereas more variable frequencies were

149 found for glutamatergic types (0.01 to 18.4% of neurons) and non-neuronal types (0.15% to 56.2% of

150 non-neuronal cells).

151

152 Laminar dissections in human M1 further allowed the estimation of laminar distributions of cell types

153 based on the proportions of nuclei dissected from each layer (Fig. 1c). As expected and previously

154 reported in middle temporal gyrus (MTG) of human neocortex [2], glutamatergic neuron types were

155 specific to layers. A subset of CGE-derived *Lamp5/Sncg* GABAergic neurons were restricted to L1, and

156 MGE-derived GABAergic types (*Sst* and *Pvalb*) displayed laminar patterning, with transcriptomically

157 similar types showing proximal laminar distributions, whereas *Vip* GABAergic neuron types displayed

158 the least laminar specificity. Three astrocyte subtypes had frequencies and layer distributions that

159 correlated with known morphologically-defined astrocyte types [25], including a common type in all layers

160 (protoplasmic), a rare type in L1 (interlaminar) [26], and a rare type in L6 (fibrous).

161

162 Single-nucleus sampling provides a relatively unbiased survey of cellular diversity [2,27] and enables

163 comparison of cell subclass proportions across species. For each donor, we estimated the proportion of

164 GABAergic and glutamatergic cells among all neurons and compared the proportions across species.

165 Consistent with previously reported differences in GABAergic neuron frequencies in primate versus

166 rodent somato-motor cortex based on histological measurements (reviewed in [28]), we found twice as

167 many GABAergic neurons in human (33%) compared to mouse M1 (16%) an intermediate proportion

168 (23%) in marmoset (Fig. 1f). Despite these differences, the relative proportions of GABAergic neuron

169 subclasses were similar. Exceptions to this included an increased proportion of *Vip* and *Sncg* cells and

170 decreased proportion of *Pvalb* cells in marmoset. Among glutamatergic neurons, there were

171 significantly more L2 and L3 IT neurons in human than marmoset and mouse (Fig. 1f), consistent with

172 the dramatic expansion of supragranular cortical layers in human (Fig. 1a) [29]. The L5

173 extratelencephalic-projecting (ET) types (also known as pyramidal tract, PT, or subcerebral types),

174 including corticospinal neurons and Betz cells in primate M1, comprised a significantly smaller

175   proportion of glutamatergic neurons in primates than mouse. This species difference was also reported

176   in MTG [2], possibly reflecting the spatial dilution of these cells with the expansion of neocortex in

177   primates. Similarly, the L6 cortico-thalamic (CT) neuron populations were less than half as frequent in

178   primates compared to mouse, whereas the L6 *Car3* type was rare in all species and relatively more

179   abundant in marmoset.

180

181   Individual nuclei were isolated from M1 of the same donors for each species and molecular profiles

182   were derived for DNA methylation (snmC-seq2) and open chromatin combined with mRNA (SNARE-

183   seq2). Independent unsupervised clustering of epigenomic data also resulted in diverse clusters (see

184   below, Figs. 4 and 5) that were mapped back to RNA clusters based on shared (directly measured or

185   inferred) marker gene expression. Cell epigenomes were highly correlated with transcriptomes, and all

186   epigenomic clusters mapped to one or more transcriptomic clusters. The epigenome data generally had

187   lower cell type resolution (Fig. 1c-e), although this may be due to sampling fewer cells or sparse

188   genomic coverage. Interestingly, snmC-seq2 and SNARE-seq2 resolved different granularities of cell

189   types. For example, more GABAergic *Vip* neuron types were identified in human M1 based on DNA-

190   methylation than open chromatin, despite profiling only 5% as many nuclei with snmC-seq2 (Fig. 1c).

191

192   **Consensus cellular M1 taxonomy across species**

193   A consensus cell type classification identifies conserved molecular makeup and allows direct cross-

194   species comparisons. The snRNA-seq Cv3 datasets were integrated using Seurat [24] that aligns nuclei

195   across species based on shared co-expression of a subset of orthologous genes with variable

196   expression. We repeated this analysis for three cell classes: GABAergic neurons (Fig. 2), glutamatergic

197   neurons (Extended Data Fig. 3) and non-neuronal cells (Extended Data Fig. 4). As represented in a

198   reduced dimension UMAP space (Fig.2a), GABAergic neuronal nuclei were well-mixed across species.

199   Eight well-defined populations formed distinct islands populated by cells from all three species,

200   including CGE-derived (*Lamp5*, *Sncg*, *Vip*) and MGE-derived (*Pvalb*, *Sst*, *Sst Chodl*) subclasses, and

201 *Lamp5 Lhx6* and chandelier cell (ChC) types. To identify conserved molecular expression for each

202 subclass across species, we first identified genes that were enriched in each subclass ("markers")

203 compared to all GABAergic neurons in each species (ROC test; AUC > 0.7). Then, we looked for

204 overlap among these genes across species. Each subclass had a core set of conserved markers (Fig.

205 2b, markers listed in Supplementary Table 7), and many subclass markers were species-specific. The

206 contrast between a minority of conserved and majority of species-specific marker genes enriched in

207 subclasses is particularly clear in the heatmap in Figure 2c (Supplementary Table 8). As expected

208 based on their closer evolutionary distance, human and marmoset shared more subclass markers with

209 each other than with mouse (Fig. 2b).

210

211 Cell types remained distinct within species and aligned across species in the integrated transcriptomic

212 space (Fig. 2d). To establish a consensus taxonomy of cross-species clusters, we used unsupervised

213 clustering to split the integrated space into more than 500 small clusters ('metacells') and built a

214 dendrogram and quantified branch stability by subsampling metacells and reclustering (Extended Data

215 Fig. 2a). Metacells were merged with nearest neighbors until all branches were stable and included

216 nuclei from the three species (see Methods). We used cluster overlap heatmaps to visualize the

217 alignment of cell types across species based on membership in merged metacells (Fig. 2e). 24

218 GABAergic consensus clusters displayed consistent overlap of clusters among the three species and

219 are highlighted as blue boxes in the heatmaps (Fig. 2e).

220

221 We next constructed a consensus taxonomy by pruning the metacell dendrogram (Extended Data Fig.

222 2a), and demonstrated that all types were well mixed across species (Fig. 2f, grey branches). The

223 robustness of consensus types was bolstered by a conserved set of marker genes (Extended Data Fig.

224 2d) and high classification accuracy of subclasses (Extended Data Fig. 2e, data in Supplementary

225 Table 9) and types (Extended Data Fig. 2f, data in Supplementary Table 9) compared to nearest

226 neighbors within and among species using a MetaNeighbor analysis [30]. Distinct consensus types (ChC,

227 *Sst Chodl*) were the most robust (mean AUROC = 0.99 within-species and 0.88 cross-species), while

228 *Sncg* and *Sst* types could not be as reliably differentiated from closely related types (mean AUROC =

229 0.84 within-species and 0.50 cross-species). Most consensus GABAergic types were enriched in the

230 same layers in human and mouse (Fig. 2f), although there were also notable species differences. For

231 example, ChCs were enriched in L2/3 in mouse and distributed across all layers in human as was seen

232 in temporal cortex (MTG) based on RNA ISH [2]. *Sst Chodl* was restricted to L6 in mouse and also found

233 in L1 and L2 in human, consistent with previous observations of sparse expression of *SST* in L1 in

234 human not mouse cortex [31].

235

236 More consensus clusters could be resolved by pairwise alignment of human and marmoset than

237 primates and mouse, particularly *Vip* subtypes (Fig. 2g, Extended Data Fig. 2b). Higher resolution

238 integration of cell types was also apparent in cluster overlap plots between human and marmoset

239 clusters (Fig. 2e, Extended Data Fig. 2c). We quantified the expression conservation of functionally

240 annotated sets of genes by testing the ability of gene sets to discriminate GABAergic consensus types.

241 This analysis was framed as a supervised learning task, both within- and between-species [30]. Within-

242 species, gene sets related to neuronal connectivity and signaling were most informative for cell type

243 identity (Extended Data Fig. 2g), as reported in human and mouse cortex [2,32]. All gene sets had

244 remarkably similar consensus type classification performance across species (r > 0.95; Fig. 2h),

245 pointing to strong evolutionary constraints on the cell type specificity of gene expression central to

246 neuronal function. Gene set classification performance was systematically reduced when training and

247 testing between primates (44% reduction) and between primates and mouse (65% reduction; Fig. 2h).

248 Therefore, many cell type marker genes were expressed in different consensus types between species.

249 Future comparative work can compare reductions in classification performance to evolutionary

250 distances between species to estimate rates of expression change across phylogenies.

251

252 Cross-species consensus types were defined for glutamatergic neurons using an identical approach as

253 for GABAergic neurons (Extended Data Fig. 3). In general, glutamatergic subclasses aligned well

254 across species and had a core set of conserved markers as well as many species-specific markers

255 (Extended Data Fig. 3a-c, genes listed in Supplementary Tables 10-11). 13 consensus types were

256 defined across species. Glutamatergic types had fewer conserved markers than GABAergic types

257 (Extended Data Fig. 3d-f,j), although subclasses and types were similarly robust (mean within-species

258 AUROC = 0.86 for GABAergic types and 0.85 for glutamatergic types) based on classification

259 performance (Extended Data Fig. 3k,l and Supplementary Table 9). Human and marmoset had

260 consistently more conserved marker genes than primates and mouse (Extended Data Fig. 3i) and could

261 be aligned at somewhat higher resolution (Extended Data Fig. 3g,h) for L5/6 NP and L5 IT subclasses.

262

263 Integration of non-neuronal cells was performed similarly to neurons (Extended Data Fig. 4a).

264 Consensus clusters (blue boxes in Extended Data Fig. 4c) that shared many marker genes were

265 identified across species (Extended Data Fig. 4d), and there was also evidence for the evolutionary

266 divergence of gene expression in consensus types. For example, the Astro_1 type had 560 DEGs

267 (Wilcox test; FDR < 0.01, log-fold change > 2) between human and mouse and only 221 DEGs

268 between human and marmoset (Extended Data Fig. 4e). The human cortex contains several

269 morphologically distinct astrocyte types [33]: interlaminar (ILA) in L1, protoplasmic in all layers, varicose

270 projection in deep layers, and fibrous in white matter (WM). We previously reported two transcriptomic

271 clusters in human MTG that corresponded to protoplasmic astrocytes and ILAs [2], and we validated

272 these types in M1 (Extended Data Fig. 4g,h). We identified a third type, Astro L1-6 *FGFR3 AQP1*, that

273 expresses *APQ4* and *TNC* and corresponds to fibrous astrocytes in WM (Extended Data Fig. 4g, left

274 ISH). A putative varicose projection astrocyte did not express human astrocyte markers (Extended Data

275 Fig. 4g, middle and right ISH), and this rare type may not have been sampled or is not

276 transcriptomically distinct.

277

278 Species comparison of non-neuronal cell types was more challenging than for neurons due to variable

279 sampling across species and more immature non-neuronal cells in mouse. 5- to 15-fold lower sampling

280 of non-neuronal cells in human impacted detection of rare types. For example, pericytes, smooth

281 muscle cells (SMCs), and some subtypes of vascular and leptomeningeal cells (VLMCs) were present

282 in marmoset and mouse and not human datasets (Extended Data Fig. 4c, right plot, blue arrows),

283 although these cells are clearly present in human cortex (for example, see [34]). A maturation lineage

284 between oligodendrocyte precursor cells (OPCs) and oligodendrocytes based on reported marker

285 genes [35] that was present in mouse and not primates (Extended Data Fig. 4b) likely represents the

286 younger age of mouse tissues used. Mitotic astrocytes (Astro_*Top2a*) were also only present in mouse

287 (Extended Data Fig. 4a,c) and represented 0.1% of non-neuronal cells. Primates had a unique

288 oligodendrocyte population (Oligo *SLC1A3 LOC103793418* in marmoset and Oligo L2-6 *OPALIN*

289 *MAP6D1* in human) that was not a distinct cluster in mouse (Extended Data Fig. 4c, left plot, blue

290 arrow). Surprisingly this oligodendrocyte clustered with glutamatergic types (Fig. 1c,d) and was

291 associated with neuronal transcripts such as *NPTX1*, *OLFM3*, and *GRIA1* (Extended Data Fig. 4i). This

292 was not an artifact, as FISH for markers of this type (*SOX10*, *ST18*) co-localized with neuronal markers

293 in the nuclei of cells that were sparsely distributed across all layers of human and marmoset M1

294 (Extended Data Fig. 4j). This may represent an oligodendrocyte type that expresses neuronal genes or

295 could represent phagocytosis of parts of neurons and accompanying transcripts that are sequestered in

296 phagolysosomes adjacent to nuclei.

297

298 To assess differential isoform usage between human and mouse, we used SSv4 data with full transcript

299 coverage and estimated isoform abundance in cell subclasses. Remarkably, 25% of moderately

300 expressed (> 10 transcripts per million) isoforms showed a large change (>9-fold) in usage between

301 species, and isoform switching was 30-60% more common in non-neuronal than neuronal cells

302 (Extended Data Fig. 2h,i, Supplementary Table 12). For example, β2-Chimaerin (*CHN2*), a gene shown

303 to mediate axonal pruning in the hippocampus [36], was highly expressed in human and mouse L5/6 NP

304   neurons. In mouse, the short isoform was almost exclusively expressed, while in human, longer

305   isoforms were also expressed (Extended Data Fig. 2j).

306

307   **Open chromatin profiling reveals distinct cell type gene regulation**

308   To directly match accessible chromatin profiles to RNA-defined cell populations, we used SNARE-Seq

309   [37], now modified for highly multiplexed combinatorial barcoding (SNARE-Seq2) [38]. We generated

310   84,178 and 9,946 dual-omic single-nucleus RNA and accessible chromatin (AC) datasets from human

311   (n = 2) and marmoset (n = 2) M1, respectively (Extended Data Fig. 5a-b, Supplementary Table 13). On

312   average, 2,242 genes (5,764 unique transcripts) were detected per nucleus for human and 3,858 genes

313   (12,400 unique transcripts) per nucleus for marmoset, due to more than 4-fold greater sequencing

314   depth for marmoset (average 17,576 reads per nucleus for human and 77,816 reads per nucleus for

315   marmoset).

316

317   To define consensus clusters, SNARE-seq2 single-nucleus RNA expression data were mapped to

318   human and marmoset transcriptomic clusters (Fig. 1c,d) based on correlated expression of cell type

319   marker genes. SNARE-seq2 transcriptomes were also independently clustered, with both approaches

320   giving consistent results (Extended Data Fig. 5c-f). Consensus clusters were more highly resolved in

321   transcriptomic compared to AC data (Extended Data Fig. 5g), and so an integrative approach was used

322   to achieve best matched AC-level cluster annotations (Extended Data Fig. 5h-k). AC peak calling at

323   multiple levels of cellular identity (for RNA consensus clusters, resolved AC clusters, subclasses and

324   classes) yielded a combined total of 273,103 (human) and 134,769 (marmoset) accessible sites, with

325   an average of 1527 or 1322 unique accessible peak fragment counts per nucleus, respectively. Gene

326   activity estimates based on cis-regulatory interactions predicted from co-accessible promoter and distal

327   peak regions using Cicero [39] were highly correlated with RNA expression values. This highlights the

328   ability of SNARE-Seq2 to meaningfully characterize AC at RNA-defined cellular resolution that cannot

329   be achieved using only AC data (Extended Data Fig. 6a-b). The AC-level clusters (Fig. 3a,b) that

330   showed similar coverage across individual samples (Extended Data Fig. 6c-f) revealed regions of open

331   chromatin that are extremely cell type specific (Fig. 3c). These regulatory regions were relatively more

332   abundant in glutamatergic compared to GABAergic neuron subpopulations (Fig. 3c-d, Supplementary

333   Table 14).

334

335   To better understand the interplay of gene regulation and expression, we compared transcript counts

336   and open chromatin measured in the same nuclei. For example, the GABAergic neuron marker *GAD2*

337   and the L2/3 glutamatergic neuron marker *CUX2* showed cell-type specific chromatin profiles for co-

338   accessible sites that were consistent with their corresponding transcript abundances (Fig. 3e-g).

339   Transcription factor binding site (TFBS) activities were calculated using chromVAR [40], permitting

340   discovery of differentially active TFBSs between cell types. To investigate the regulatory factors that

341   may contribute to marker gene expression, we evaluated active TFBSs for their enrichment within

342   marker gene co-accessible sites. This permitted direct cell type mapping of gene expression and

343   activity levels with the expression and activity of associated regulatory factors (Fig. 3g). Using this

344   strategy, we identified TFBS activities associated with subclass (Fig. 3h-i) and AC-cluster level

345   differentially expressed genes (DEGs) in human and marmoset (Supplementary Table 15). DEG

346   transcript levels and AC-inferred gene activity scores showed high correspondence (Fig. 3h). While

347   most subclasses also showed distinct TFBS activities, correspondence between human and marmoset

348   was higher for glutamatergic rather than GABAergic neurons (Fig. 3h,j). For GABAergic neuron

349   subclasses, gene expression profiles were more conserved than TFBS activities, consistent with fewer

350   differences between GABAergic subpopulations based on AC sites (Fig. 3a,b). This observation is also

351   consistent with fewer distinct TFBS activities among some inhibitory neuron subclasses (*Lamp5*, *Sncg*)

352   in human compared to marmoset (Fig. 3h), despite these cell types having a similar number of AC peak

353   counts (Extended Data Fig. 6d-f). Interestingly, glutamatergic neurons in L5 and L6 showed higher

354   correspondence between primates based on TFBS activities compared to average expression,

355 suggesting that gene regulatory processes are more highly conserved in these subclasses than target

356 gene expression.

357

358 **Methylomic profiling reveals conserved gene regulation**

359 We used snmC-seq2 [41] to profile the DNA methylome from individual cells in M1. Single-nuclei were

360 labeled with an anti-NeuN antibody and isolated by fluorescence-activated cell sorting (FACS), and

361 neurons were enriched (90% NeuN+ nuclei) to increase detection of rare types. Using snmC-seq2, we

362 generated single-nucleus methylcytosine datasets from M1 of human (n = 2 donors, 6,095 nuclei),

363 marmoset (n = 2, 6,090), and mouse (9,876) (Liu et al. companion paper) (Supplementary Table 16).

364 On average, 5.5 ± 2.7% (mean ± s.d.) of human, 5.6 ± 2.9% of marmoset and 6.2 ± 2.6% of mouse

365 genomes were covered by stringently filtered reads per cell, with 3.4×10^4 (56%), 1.8×10^4 (62%) and

366 4.5×10^4 (81%) genes detected per cell in the three species, respectively. Based on the DNA

367 methylome profiles in both CpG sites (CG methylation or mCG) and non-CpG sites (CH methylation or

368 mCH), we clustered nuclei (Methods) to group cell populations into 31 cell types in human, 36 cell types

369 in marmoset, and 42 cell types in mouse (Fig. 4a and Extended Data Fig. 7a,b). For each species, cell

370 type clusters could be robustly discriminated using a supervised classification model and had distinct

371 marker genes based on DNA methylation signatures for neurons (mCH) or non-neuronal cells (mCG)

372 (Methods). Differentially methylated regions (DMR) were determined for each cell type versus all other

373 cell types and yielded 9.8×10^5 DMRs in human, 1.0×10^6 in marmoset, and 1.8×10^6 in mouse.

374

375 We determined a consensus molecular classification of cell types in each species by integrating single-

376 nucleus methylomic data with the Cv3 transcriptomic data described above using measurements of

377 gene body differential methylation (CH-DMG) to approximate expression levels. Nuclei from the two

378 data modalities mixed well as visualized in ensemble UMAPs (Fig. 4b,c). Methylation clusters have

379 one-to-one, one-to-many, or many-to-many mapping relation to transcriptomic clusters (Fig. 1c-e and

380 Extended Data Fig. 7d-f). DMRs were quantified for each subclass versus all other subclasses (Fig.

381   4d), and glutamatergic neurons had more hypo-methylated DMRs compared to GABAergic neurons.

382   Methylome tracks at subclass level can be found at http://neomorph.salk.edu/aj2/pages/cross-species-

383   M1/. To identify enriched transcription factor binding sites (TFBS) in each species and subclass, we

384   performed motif enrichment analysis with hypo-methylated DMRs from one subclass against other

385   DMRs of the same species, and identified 102 ± 57 (mean ± s.d.) TFBS in each subclass (Extended

386   Data Fig. 8 and Supplementary Table 17). We repeated the enrichment analysis using TFBS motif

387   clusters [42] and found similarly distinct subclass signatures (Supplementary Table 18). Although

388   subclasses had unique marker genes (Fig. 2c, genes listed in Supplementary Table 8) and CH-DMG

389   across species, they had remarkably conserved TFBS motif enrichment (Fig. 4e,f and Extended Data

390   Fig. 8). For example, *TCF4* is robustly expressed in L5 IT neurons across species and shows

391   significant TFBS enrichment in hypo-methylated DMRs and AC sites. DMRs and AC sites provide

392   independent epigenomic information (Extended Data Fig. 7f,g) and can identify different TFBS

393   enrichment, such as for *ZNF148* in L5 IT neurons. These results are consistent with previous

394   observations of conserved TF network architectures in neural cell types between human and mouse

395   (Stergachis et al. 2014). Conserved sets of TFs have the potential to determine conserved and

396   divergent expression in consensus types based on shared or altered genomic locations of TFBS motifs

397   across species.

398

399   **Layer 4-like neurons in human M1**

400   M1 lacks a L4 defined by a thin band of densely packed "granular" neurons that is present in other

401   cortical areas, such as MTG (Fig. 5a), although prior studies have identified neurons with L4-like

402   synaptic properties in mice [14] and expression of *RORB*, a L4 marker, in non-human primate M1 [43]. To

403   address the potential existence of L4-like neurons in human M1 from a transcriptomic perspective, we

404   integrated snRNA-seq data from M1 and the granular MTG, where we previously described multiple L4

405   glutamatergic neuron types [2]. This alignment revealed a broadly conserved cellular architecture

406   between M1 and MTG (Fig. 5b,c, Extended Data Fig. 9) including M1 neuron types Exc L3 *RORB*

407   *OTOGL* (here, *OTOGL*) and Exc L3-5 *RORB LINC01202* (here, *LINC01202*) that map closely to MTG

408   neurons in deep L3 to L4 (Fig. 5c, red outlines). Interestingly, four MTG L2/3 IT types (*LTK*, *GLP2R*,

409   *FREM3*, and *CARM1P1*) whose distinct physiology and morphology are reported in a companion paper

410   [44] had less clear homology in M1 than other types (Extended Data Fig. 9a-c), pointing to more

411   variability across cortical areas of superficial as compared to deep glutamatergic neurons. To compare

412   laminar positioning in M1 and MTG, the relative cortical depth from pia for each neuron was estimated

413   based on the layer dissection and average layer thickness [45]. Transcriptomically similar cell types were

414   found at similar cortical depths in M1 and MTG, and the *OTOGL* and *LINC01202* types were located in

415   deep L3 and superficial L5 in M1 (Fig. 5d).

416

417   MTG contains three main transcriptomically-defined L4 glutamatergic neuron types (*FILIP1L*, *TWIST2*

418   and *ESR1*) and a deep L3 type (*COL22A1*) that is found on the border of L3 and L4 (Fig. 5e-g). The M1

419   types *OTOGL* and *LINC01202* matched one-to-one with MTG *COL22A1* and *ESR1*, whereas there

420   were no matches for the other two MTG L4 types (Fig. 5f). Based on snRNA-seq proportions, the L4-

421   like *OTOGL* type was much sparser in M1 than the *ESR1* type in MTG (Fig. 5e). Multiplex fluorescent in

422   situ hybridization (mFISH) with probes to cell type marker genes confirmed these findings. The MTG

423   *ESR1* type was highly enriched in L4, [2], and the homologous M1 *LINC01202* type was sparser and

424   more widely distributed across L3 and L5 (Fig. 5g). The MTG *COL22A1* type was tightly restricted to

425   the L3/4 border [2], and the M1 *OTOGL* type was similarly found at the L3/5 border. Quantification of

426   labeled cells as a fraction of DAPI+ cells in L3-5 showed similar frequencies of M1 *OTOGL* and MTG

427   *COL22A1* types and 4-fold sparser M1 *LINC01202* type versus MTG *ESR1* type (Fig. 5h). These data

428   indicate a conservation of deep L3 glutamatergic types and proportions across human cortical areas,

429   but with reduced diversity and sparsification of L4-like neurons to a single (ESR1) type in M1,

430   distributed more broadly where L4 would be if tightly aggregated.

431

432   **Chandelier cells share a core molecular identity across species**

433   Conserved transcriptomic and epigenomic features of consensus types likely contribute to cell function

434   and generate hypotheses about the gene regulatory mechanisms underlying cell type identity. Focused

435   analysis of *Pvalb*-expressing GABAergic neurons illustrates the power of these data to predict such

436   gene-function relationships. Cortical *Pvalb*-expressing neurons comprise two major types — basket

437   cells (BCs) and ChCs — that have fast-spiking electrical properties and distinctive cellular

438   morphologies. BCs selectively synapse onto the perisomatic region of glutamatergic pyramidal

439   neurons. ChCs, also called axo-axonic cells [46], selectively innervate the axon initial segment (AIS) of

440   pyramidal cells and have unique synaptic specializations called axon cartridges. These cartridges run

441   perpendicular to their post-synaptic target axon, giving a characteristic morphological appearance of

442   candlesticks on a chandelier. This highly conserved feature is shown with biocytin-filled cells from

443   mouse, rhesus macaque, and human (Fig. 6a). To reveal evolutionarily conserved transcriptomic

444   hallmarks of ChCs, we identified DEGs in ChCs versus BCs in each species using an ROC test. 357

445   DEGs were identified in at least one species, and marmoset ChCs shared more DEGs with human (61

446   genes) than mouse (29; Fig. 6b, Supplementary Table 19). Remarkably, only 25 DEGs were conserved

447   across all three species. One conserved gene, *UNC5B* (Fig. 6c), is a netrin receptor involved in axon

448   guidance and may help target ChC to pyramidal neuron AIS. Three transcription factors (*RORA*,

449   *TRPS1*, and *NFIB*) were conserved markers and may contribute to gene regulatory networks that

450   determine the unique attributes of ChCs.

451

452   To determine if ChCs had enriched epigenomic signatures for *RORA* and *NFIB* (*TRPS1* lacked motif

453   data), we compared DMRs between ChCs and BCs. In all species, *RORA* and *NFIB* had significant

454   CH-DMGs in ChCs not BCs (Fig. 6d), consistent with differential expression. To discern if these TFs

455   may preferentially bind to DNA in ChCs, we tested for TF motif enrichment in hypo-methylated (mCG)

456   DMRs and AC sites genome-wide. We found that the RORA motif was significantly enriched in DMRs

457   in primates (Fig. 6d) and in AC sites of ChCs in all species (Fig. 6e, Supplementary Table 14). The

458   NFIB binding motif was only significantly enriched in AC sites of mouse ChCs, possibly because

459   enrichment was transient during development or NFIB specificity is due to expression alone. Three

460   independent genomic assays converge to implicate *RORA* as a ChC-specific TF among *Pvalb*-

461   expression neurons. Notably, 60 of 357 DEGs contained a ROR-motif in DMRs and AC regions in at

462   least one species, further implicating *RORA* in defining ChC identity.

463

464   **Primate Betz cell specialization**

465   In mouse cortex, L5 glutamatergic neurons have distinct long-range projection targets (ET versus IT)

466   and transcriptomes [1]. L5 ET and IT neuron subclasses clearly align between human and mouse using

467   snRNA-seq in M1 (Extended Data Fig. 3) and in temporal [2] and fronto-insular cortex [12]. Betz cells in L5

468   of primate M1 connect to spinal motor-neurons via the pyramidal tracts and are predicted to be L5 ET

469   neurons. The species aligned transcriptomic types allow for the identification of genes whose

470   expression may contribute to conserved ET versus IT features and primate-specific physiology,

471   anatomy, and connectivity. Furthermore, Patch-seq methods that jointly measure the transcriptome,

472   physiological properties and morphology of cells, allow the direct identification and characterization of

473   L5 ET and IT neurons across mouse, non-human primate, and human. As primate physiology

474   experiments are largely restricted to macaque, we also profiled L5 of macaque M1 with snRNA-seq

475   (Cv3) to allow accurate Patch-seq mapping.

476

477   L5 ET neurons had many DEGs compared to L5 IT neurons in all 4 species. Approximately 50 DEGs

478   were conserved across all species and similarity to human varied as a function of evolutionary distance

479   (Fig. 7a, Supplementary Table 20). Several genes encoding ion channel subunits were enriched in ET

480   versus IT neurons in all species, potentially mediating conserved ET physiological properties (Fig. 7b).

481   A number of additional potassium and calcium channels were primate-enriched (Fig. 7c), potentially

482   underlying primate-specific ET or Betz cell physiology. Interestingly, many of these primate-specific ET-

483   enriched genes showed gradually increasing ET specificity in species more closely related to human.

484   To explore this idea of gradual evolutionary change further, we identified genes with increasing L5 ET

485   versus IT specificity as a function of evolutionary distance from human (Fig. 7d, Supplementary Table

486   21). Interestingly, this gene set was highly enriched for genes associated with axon guidance including

487   members of the Robo, Slit and Ephrin gene families. These genes are potential candidates for

488   regulating the cortico-motoneuronal connections associated with increasingly dexterous fine motor

489   control across these species [23].

490

491   To investigate if transcriptomically defined L5 ET types contain anatomically-defined Betz cells, FISH

492   for L5 ET neurons was combined with immunolabeling against SMI-32, a protein enriched in Betz cells

493   and other long-range projecting neurons in macaque [47–49] (Fig. 7e). Cells consistent with the size and

494   shape of Betz cells were identified in two L5 ET clusters (Exc L5 *FEZF2 ASGR2* and Exc L5 *FEZF2*

495   *CSN1S1*). Similar to previous reports on von Economo neurons in the insular cortex [12], ET clusters in

496   M1 also included neurons with non-Betz morphologies.

497

498   To facilitate cross-species comparisons of Betz cells and mouse ET neurons we made patch clamp

499   recordings from L5 neurons in acute and cultured slice preparations of mouse and macaque M1. For a

500   subset of recordings, Patch-seq analysis was applied for transcriptomic cell type identification

501   (Extended Data Fig. 10h). To permit visualization of cells in heavily myelinated macaque M1, we used

502   AAV viruses to drive fluorophore expression in glutamatergic neurons in macaque slice culture

503   (Extended Data Fig. 10g). As shown in Figure 7f, Patch-seq neurons mapping to the macaque Betz/ET

504   cluster (Exc L5 *FEZF2 LOC114676463*) had large somata (diameter > 65 μm) and long "tap root" basal

505   dendrites, canonical hallmarks of Betz cell morphology [17,50]. A unique opportunity to record from

506   neurosurgical tissue excised from human premotor cortex (near the confluence of the precentral and

507   superior frontal gyri) during an epilepsy treatment surgery using the same methods as for macaque

508   yielded multiple neurons that mapped transcriptomically to one of the Betz-containing cell types and

509   had canonical Betz cell morphology (Fig. 7g). Macaque and human ET neurons were grouped for

510   physiological analysis because intrinsic properties were not significantly different, and many

511   corticospinal axons originate from premotor cortex[23].

512

513   Shared transcriptomic profiles of mouse, primate, and human L5 ET neurons predicted conservation of

514   some physiological properties of rodent and primate neurons. Transcriptomically-defined ET neurons

515   across species expressed high levels of genes encoding an HCN channel-subunit and a regulatory

516   protein (*HCN1* and *PEX5L;* Fig. 7b). We hypothesized that HCN-dependent membrane properties,

517   which are used to distinguish rodent ET from IT neurons [51], would similarly separate cell types in

518   primates. Some primate L5 neurons possessed distinctive HCN-related properties such as a lower

519   input resistance (R_N) and a peak resonance (f_R) in voltage response around 3-9 Hz (Fig. 7h,i), similar to

520   rodent ET neurons. To determine whether HCN-related physiology is a conserved feature of L5

521   neurons, we grouped all neurons into physiologically defined ET and non-ET neurons based on their R_N

522   and f_R. We asked whether these physiologically-defined neurons corresponded to genetically-defined

523   ET/Betz or non-ET neurons using Patch-seq and cell-type specific mouse lines. For mouse M1, the ET-

524   specific *Thy1*-YFP[21,52] and IT specific *Etv1*-EGFP[53] mouse lines preferentially labeled physiologically

525   defined ET and non-ET neurons, respectively (Fig. 7j). For primates, transcriptomically-defined Betz

526   cells were physiologically defined ET neurons, whereas transcriptomically defined non-ET neurons

527   were physiologically defined non-ET neurons (Fig. 7k). Thus, there was broad correspondence

528   between physiologically-defined and genetically-defined ET neurons in both mouse and primate M1.

529   There were notable differences in physiology between mouse and primate ET neurons, however. A

530   greater fraction of primate ET neurons exhibited an exceptionally low R_N compared to mouse (Fig. 7l).

531   Additional differences in action potential properties across cell types and species may be explained in

532   part by differences in the expression of ion channel-related genes (Fig. 7c, Extended Data Fig. 10).

533

534   Most strikingly, primate Betz/ET neurons displayed a distinctive biphasic-firing pattern during long spike

535   trains. The firing rate of both primate and mouse non-ET neurons decreased to a steady state within

536   the first second of a 10 second depolarizing current injection, whereas the firing rate of mouse ET

537   neurons increased moderately over the same time period (Fig. 7m,n; Extended Data Fig. 10m,n). The

538   acceleration in rodent ET neurons has been attributed to the expression of Kv1-containing voltage-

539   gated K$^+$ channels that are encoded by genes like the conserved ET gene *KCNA1*. In macaque and

540   human ET/Betz neurons, a distinctive biphasic pattern was characterized by an early cessation of firing

541   followed by a sustained and dramatic increase in firing later in the current injection. Thus, while ET

542   neurons in both primate and rodent M1 displayed spike frequency acceleration, the temporal dynamics

543   and magnitude of this acceleration appears to be a unique feature of primate ET/Betz neurons. These

544   data emphasize how transcriptomic data from this specialized neuron type can be linked to shared and

545   unique physiological properties across species.

546

547   Discussion

548   Comparative analysis is a powerful strategy to understand brain structure and function. Species

549   conservation is strong evidence for functional relevance under evolutionary constraints that can help

550   identify critical molecular and regulatory mechanisms [54,55]. Conversely, divergence indicates adaptive

551   specialization, which may be essential to understand the mechanistic underpinnings of human brain

552   function and susceptibility to human-specific diseases. In the current study, we applied a comparative

553   approach to understand conserved and species-specific features of M1 at the level of cell types using

554   single-nucleus RNA-seq (Cv3 and SSv4), open chromatin (SNARE-seq2 and ATAC-seq) and DNA-

555   methylation (snmC-seq2) technologies. Integrated analysis of over 450,000 nuclei in human, non-

556   human primates (marmoset, a New World monkey, and to a lesser degree macaque, an Old World

557   monkey that is evolutionarily more closely related to humans), and mouse (see also companion paper

558   [6]) yielded a high-resolution, multimodal classification of cell types in each species, and a coarser

559   consensus classification conserved between rodent and primate lineages. Robust species conservation

560   strongly argues for the functional relevance of this consensus cellular architecture. Species

561 specializations are also apparent, both in the additional granularity in cell types within species and

562 differences between conserved cell types. A comparative evolutionary approach provides an anchor

563 point to define the cellular architecture of any tissue and to discover species-specific adaptations.

564

565 A key result of the current study is the identification of a consensus classification of cell types across

566 species that allows the comparison of relative similarities in human compared to common mammalian

567 model organisms in biomedical research. Prior studies have demonstrated that high resolution cellular

568 taxonomies can be generated in mouse, non-human primate and human cortex, and that there is

569 generally good concordance across species [2,11]. However, inconsistencies in the methods and

570 sampling depths used made strong conclusions difficult, compounded by the analysis of different

571 cortical regions in different species. The current study overcame these challenges by focusing on M1, a

572 functionally and anatomically conserved cortical region across mammals, and comparing a variety of

573 methods on similarly isolated tissues (and the same specimens from human and marmoset). Several

574 important points emerged from these integrated analyses. First, with deeper sampling and the same

575 methodology (snRNA-seq with Cv3), a similar cellular complexity on the order of 100 cell types was

576 seen in all three species. The highest resolution molecular classification was seen with RNA-seq

577 compared to epigenomic methods, and among RNA-seq methods with those that allow the most cells

578 to be analyzed. Strikingly, the molecular classifications were well aligned across all methods tested,

579 albeit at different levels of resolution as a function of the information content of the assay and the

580 number of cells profiled. All methods were consistent at the level of subclasses as defined above, both

581 across methods and species; significantly better alignment was achieved among species based on

582 transcriptomics, and with epigenomic methods in some subclasses. Mismatches in cellular sampling

583 affect the ability to compare across species; for example, higher non-neuronal sampling in mouse and

584 marmoset increased detection of rare cell types compared to human. One important comparison was

585 between plate-based (SSv4) and droplet-based (Cv3) RNA-seq of human nuclei, where we compared

586 results between approximately 10,000 SSv4 and 100,000 Cv3 nuclei. On average, SSv4 detected 30%

587 more genes per nucleus and enabled comparisons of isoform usage between cell types, albeit with 20-

588 fold greater sequencing depth. However, SSv4 cost 10 times as much as Cv3 and did not allow

589 detection of additional cell types.

590

591 The snmC-seq2 clustering aligned closely with the transcriptomic classification, although with

592 significantly lower resolution in rarer subclasses. Hypo-methylated sites correlated with gene

593 expression and specific transcription factor binding motifs were enriched in cell type specific sites.

594 Multi-omic SNARE-seq2 measured RNA profiles of nuclei that allowed high confidence assignment to

595 transcriptomic clusters. Examining accessible chromatin (AC) regions within the same nuclei led to

596 strong correlations between cell subclass or type gene expression and active regulatory regions of

597 open chromatin. Using this strategy, gene regulatory activities could be identified within RNA-defined

598 cell populations (including RNA consensus clusters) that could not be resolved from AC data alone

599 (Extended Data Fig. 6a, Supplementary Table 15). By joint consideration of these epigenomic

600 modalities, glutamatergic neurons were found to have more hypo-methylated DMRs and differentially

601 accessible chromatin, consistent with having larger somata and expressing more genes. Within-

602 species, cell types have many more unique AC sites than uniquely expressed marker genes. At the

603 same time, there is striking conservation across species of subclass TFBS motif enrichment within AC

604 and hypo-methylated DMRs. Most subclasses have distinct motifs, although L2/3 and L6 IT and *Lamp5*

605 and *Sncg* subclasses share many motifs and are more clearly distinguished based on gene expression.

606 Taken together, these results show a robust cell type classification that is consistent at the level of

607 subclasses both across transcriptomic and chromatin measures and across species, with additional cell

608 type-level granularity identified with transcriptomics.

609

610 Alignment across species allowed a comparison of relative similarities and differences between

611 species. A common (and expected) theme was that more closely related species are more similar to

612 one another. This was true at the level of gene expression and epigenome patterning across cell types,

613   and in the precision with which transcriptomically-defined cell types could be aligned across species.

614   For example, human and marmoset GABAergic types could be aligned at higher resolution than human

615   and mouse. Human was more similar to macaque than to marmoset. This indicates that cell type

616   similarity increases as a function of evolutionary distance to our closest common ancestors with mouse

617   (~70 mya), marmoset (~40 mya), and macaque (~25 mya). Interestingly, many gene expression

618   differences may change gradually over evolution. This is apparent in the graded changes in expression

619   levels of genes enriched in L5 ET versus L5 IT neurons and in the reduced performance of cell type

620   classification based on marker gene expression that is correlated with evolutionary distance between

621   species.

622

623   Several prominent species differences in cell type proportions were observed. First, the ratio of

624   glutamatergic excitatory projection neurons compared to GABAergic inhibitory interneurons was 2:1 in

625   human compared to 3:1 in marmoset and 5:1 in mouse and leads to a profound shift in the overall

626   excitation-inhibition balance of the cortex. A similar species difference has been described based on

627   histological measures (reviewed in [28]), indicating that snRNA-seq gives a reasonably accurate

628   measurement of cell type proportions. Surprisingly, the relative proportions of GABAergic subclasses

629   and types were similar across species. These results suggest a developmental shift in the size of the

630   GABAergic progenitor pool in the ganglionic eminences or an extended period of neurogenesis and

631   migration. A decreased proportion of the subcortically targeting L5 ET neurons in human was also

632   seen, as previously shown in temporal [2] and frontoinsular [12] cortex. This shift likely reflects the

633   evolutionary increase in cortical neurons relative to their subcortical targets [56] and was less prominent

634   in M1, suggesting regional variation in the proportion of L5 ET neurons. Finally, a large increase in the

635   proportion of L2 and L3 IT neurons was seen in human compared to mouse and marmoset. This

636   increase parallels the disproportionate expansion of human cortical area and supragranular layers that

637   contain neurons projecting to other parts of the cortex, presumably to facilitate greater corticocortical

638   communication. Interestingly, L2 and L3 IT neurons appear to be particularly highly variable across

639   cortical areas and species, and also are more diverse and specialized in human compared to mouse

640   (see companion paper [44]).

641

642   A striking and somewhat paradoxical observation is the high degree of species specialization of

643   consensus types. The majority of DEGs between cell types were consistently species-specific. This

644   result suggests that the conserved cellular features of a cell type are largely due to a minority of DEGs

645   with conserved expression patterns. The current study demonstrates this point for one of the most

646   distinctive brain cell types, the cortical *Pvalb*-expressing GABAergic ChC. ChCs in mouse, non-human

647   primate, and human have 100-150 genes with highly enriched expression compared to other *Pvalb*-

648   expressing interneurons (BCs); however, only 25 of these ChC-enriched genes are shared across

649   species. This small overlapping gene set includes several transcription factors and a member of the

650   netrin family (*UNC5B*) that could be responsible for AIS targeting. Binding sites for these TFs are

651   enriched in ChC cluster regions of open chromatin and in hypo-methylated regions around ChC-

652   enriched genes. While these associations between genes and cellular phenotypes for conserved and

653   divergent features remain to be tested, a comparative strategy can identify these core conserved genes

654   and make strong predictions about the TF code for cell types and the genes responsible for their

655   evolutionarily constrained functions.

656

657   M1 is an agranular cortex lacking a L4, although a recent study demonstrated that there are neurons

658   with L4-like properties in mouse [14]. Here we confirm and extend this finding in human M1. We find a L4-

659   like neuron type in M1 that aligns to a L4 type in human MTG and is scattered between the deep part of

660   L3 and the superficial part of L5 where L4 would be if aggregated into a layer. However, MTG

661   contained several additional L4 types not found in M1, and with a much higher frequency. The human

662   M1 L4-like type is part of the L5 IT_1 consensus cluster that includes several IT types in all species,

663   including two L4-like types in mouse (L4/5 IT_1 and L4/5 IT_2) that also express the canonical L4

664   marker *Rorb* (see companion paper [6]). Therefore, it appears that M1 has L4-like cells from a

665   transcriptomic perspective, but only a subset of the types compared to granular cortical areas, at much

666   lower density, and scattered rather than aggregated into a tight layer.

667

668   The most distinctive cellular hallmark of M1 in primates and cats is the enormous Betz cell, which

669   contributes to direct corticospinal connections to spinal motoneurons in primates that participate in fine

670   motor control [15,16,57–59]. Intracellular recordings from cats have shown highly distinctive characteristics

671   including HCN channel-related membrane properties, spike frequency acceleration, and extremely fast

672   maintained firing rates [19,20]. However, they have never been recorded in primates using patch clamp

673   physiology due to the high degree of myelination in M1 that prevents their visualization, and the inability

674   to obtain motor cortex tissue from neurosurgical procedures which are careful to be function-sparing. A

675   goal of the current project was to identify the transcriptomic cluster corresponding to Betz cells and use

676   this to understand gene expression that may underlie their distinctive properties and species

677   specializations. We have recently taken a similar approach to study von Economo neurons in the

678   fronto-insular cortex, showing they are found within a transcriptomic class consisting of ET neurons [12].

679   Betz cells are classical ET neurons that, together with the axons of smaller corticospinal neurons, make

680   up part of the pyramidal tract from the cortex to the spinal cord [16,60]. We show that neurons with Betz

681   cell morphology label with markers for the M1 ET clusters. Like von Economo neurons, there does not

682   appear to be an exclusively Betz transcriptomic type. Rather, M1 ET clusters are not exclusive for

683   neurons with Betz morphology, and we find more than one ET cluster contains neurons with Betz

684   morphology.

685

686   Although comparative transcriptomic alignments provide strong evidence for functional similarity, the

687   distinctions between corticospinal neurons across species or even between L5 ET and IT neuron types

688   in primates or humans has not been demonstrated physiologically. We recently developed a suite of

689   methodologies for studying specific neuron types in human and non-human tissues, including triple

690   modality Patch-seq to combine physiology, morphology and transcriptome analysis, acute and cultured

691   slice physiology in adult human neurosurgical resections and macaque brain, and AAV-based neuronal

692   labeling to allow targeting of neurons in highly myelinated tissues (companion paper [44; 61] . Specifically,

693   these tools allow the targeting of L5 neurons in mouse and non-human primate and the assignment of

694   neurons to their transcriptomic types using Patch-seq, which we facilitated by generating and aligning a

695   L5 transcriptomic classification in macaque where such analyses could be performed. We show here

696   that several of the characteristic features of L5 ET versus IT neurons are conserved, and can be

697   reliably resolved from one another in mouse and non-human primate. Furthermore, macaque neurons

698   with Betz-like morphologies mapped to the Betz-containing clusters. However, as predicted by

699   differences in ion channel-related gene expression, not all physiological features were conserved

700   between macaque and mouse ET neurons. Betz/ET neurons had the distinctive pauses, bursting and

701   spike-frequency acceleration described previously in cats but not seen in rodents [19,20]. Finally, we had

702   access to an extremely rare human neurosurgical case where a region of premotor cortex was

703   resected. Similar to macaque M1, this premotor region contained large neurons with characteristic

704   Betz-like morphology that mapped transcriptomically to the Betz-containing clusters. Together these

705   results highlight the predictive power of transcriptomic mapping and cross-species inference of cell

706   types for L5 pyramidal neurons including the Betz cells. Furthermore, these data are consistent with

707   observations that Betz cells may not in fact be completely restricted to M1 but distribute across other

708   proximal motor-related areas that contribute to the pyramidal tract [62]. Finally, a number of ion channels

709   that may contribute to conserved ET versus IT features as well as species specializations of Betz cell

710   function were identified that provide candidate genes to explore gene-function relationships. For

711   example, axon guidance-associated genes are enriched in Betz-containing ET neuron types in

712   primates, possibly explaining why Betz cells in primates directly contact spinal motor neurons rather

713   than spinal interneurons as in rodents. Thus, as the comparative approach is helpful in identifying core

714   conserved molecular programs, it may be equally valuable to understand what is different in human or

715   can be well modeled in closer non-human primate relatives. This is particularly relevant in the context of

716   Betz cells and other ET neuron types that are selectively vulnerable in amyotrophic lateral sclerosis,

717   some forms of frontotemporal dementia, and other neurodegenerative conditions.

718

719   References

720   1.   Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**,

721   72–78 (2018).

722   2.   Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex.

723   *Nature* **573**, 61–68 (2019).

724   3.   Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat.*

725   *Neurosci.* **19**, 335–346 (2016).

726   4.   Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of

727   the human brain. *Science* **352**, 1586–1590 (2016).

728   5.   Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in

729   mammalian cortex. *Science* **357**, 600–604 (2017).

730   6.   Yao, Z. *et al.* An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell

731   types. *bioRxiv* 2020.02.29.970558 (2020) doi:10.1101/2020.02.29.970558.

732   7.   Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the

733   human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).

734   8.   Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by

735   single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).

736   9.   Gray, L. T. *et al.* Layer-specific chromatin accessibility landscapes reveal regulatory networks in

737   adult mouse visual cortex. *Elife* **6**, (2017).

738   10.   Lee, D.-S. *et al.* Simultaneous profiling of 3D genome structure and DNA methylation in single

739   human cells. *Nat. Methods* **16**, 999–1006 (2019).

740   11.   Krienen, F. M., Goldman, M., Zhang, Q. & del Rosario, R. Innovations in primate interneuron

741   repertoire. *bioRxiv* (2019).

742 12. Hodge, R. D. *et al.* Transcriptomic evidence that von Economo neurons are regionally specialized

743   extratelencephalic-projecting excitatory neurons. *Nat. Commun.* **11**, 1172 (2020).

744 13. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem

745   Cells. *Cell* **167**, 566–580.e19 (2016).

746 14. Yamawaki, N., Borges, K., Suter, B. A., Harris, K. D. & Shepherd, G. M. G. A genuine layer 4 in

747   motor cortex with prototypical synaptic circuit connectivity. *Elife* **3**, e05422 (2014).

748 15. Betz, W. Anatomischer Nachweis zweier Gehirncentra. *Zentralbl Med Wiss* **12**, (1874).

749 16. Lassek, A. M. The Human Pyramidal Tract II. A Numerical Investigation of the Betz Cells of the

750   Motor Area. *J. Nerv. Ment. Dis.* **94**, 225–226 (1941).

751 17. Jacobs, B. *et al.* Comparative morphology of gigantopyramidal neurons in primary motor cortex

752   across mammals. *J. Comp. Neurol.* **526**, 496–536 (2018).

753 18. Kaiserman-Abramof, I. R. & Peters, A. Some aspects of the morphology of Betz cells in the

754   cerebral cortex of the cat. *Brain Res.* **43**, 527–546 (1972).

755 19. Spain, W. J., Schwindt, P. C. & Crill, W. E. Post-inhibitory excitation and inhibition in layer V

756   pyramidal neurones from cat sensorimotor cortex. *The Journal of Physiology* vol. 434 609–626

757   (1991).

758 20. Chen, W., Zhang, J. J., Hu, G. Y. & Wu, C. P. Electrophysiological and morphological properties of

759   pyramidal and nonpyramidal neurons in the cat motor cortex in vitro. *Neuroscience* **73**, 39–55

760   (1996).

761 21. Miller, M. N., Okaty, B. W. & Nelson, S. B. Region-Specific Spike-Frequency Acceleration in Layer

762   5 Pyramidal Neurons Mediated by Kv1 Subunits. *Journal of Neuroscience* vol. 28 13716–13726

763   (2008).

764 22. Gu, Z. *et al.* Control of species-dependent cortico-motoneuronal connections underlying manual

765   dexterity. *Science* **357**, 400–404 (2017).

766 23. Lemon, R. N. Descending pathways in motor control. *Annu. Rev. Neurosci.* **31**, 195–218 (2008).

767   24. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).

768   25. Oberheim, N. A. *et al.* Uniquely Hominid Features of Adult Human Astrocytes. *Journal of*

769   *Neuroscience* vol. 29 3276–3287 (2009).

770   26. Colombo, J. A. The interlaminar glia: from serendipity to hypothesis. *Brain Struct. Funct.* **222**,

771   1109–1129 (2017).

772   27. Bakken, T. E. *et al.* Single-nucleus and single-cell transcriptomes compared in matched cortical

773   cell types. *PLoS One* **13**, e0209648 (2018).

774   28. Džaja, D., Hladnik, A., Bičanić, I., Baković, M. & Petanjek, Z. Neocortical calretinin neurons in

775   primates: increase in proportion and microcircuitry structure. *Front. Neuroanat.* **8**, 103 (2014).

776   29. DeFelipe, J., Alonso-Nanclares, L. & Arellano, J. I. Microstructure of the neocortex: comparative

777   aspects. *J. Neurocytol.* **31**, 299–316 (2002).

778   30. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types

779   defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9**, 884 (2018).

780   31. Boldog, E. *et al.* Transcriptomic and morphophysiological evidence for a specialized human cortical

781   GABAergic cell type. *Nat. Neurosci.* **21**, 1185–1195 (2018).

782   32. Paul, A. *et al.* Transcriptional Architecture of Synaptic Communication Delineates GABAergic

783   Neuron Identity. *Cell* **171**, 522–539.e20 (2017).

784   33. Verkhratsky, A. & Nedergaard, M. The homeostatic astroglia emerges from evolutionary

785   specialization of neural cells. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **371**, (2016).

786   34. Nortley, R. *et al.* Amyloid β oligomers constrict human capillaries in Alzheimer's disease via

787   signaling to pericytes. *Science* vol. 365 eaav9518 (2019).

788   35. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

789   36. Riccomagno, M. M. *et al.* The RacGAP β2-Chimaerin selectively mediates axonal pruning in the

790   hippocampus. *Cell* **149**, 1594–1606 (2012).

791   37. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin

792   accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).

793   38. Nongluk Plongthongkum, Dinh Diep, Song Chen, Blue B. Lake, Kun Zhang. Scalable Dual-omic

794   Profiling with Single-nucleus Chromatin Accessibility and mRNA Expression Sequencing 2

795   (SNARE-Seq2). (2020).

796   39. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin

797   Accessibility Data. *Mol. Cell* **71**, 858–871.e8 (2018).

798   40. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-

799   factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).

800   41. Luo, C. *et al.* Robust single-cell DNA methylome profiling with snmC-seq2. *Nat. Commun.* **9**, 3824

801   (2018).

802   42. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding

803   profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).

804   43. Bernard, A. *et al.* Transcriptional architecture of the primate neocortex. *Neuron* **73**, 1083–1099

805   (2012).

806   44. Berg, J. *et al.* Human cortical expansion involves diversification and specialization of supragranular

807   intratelencephalic-projecting neurons. *bioRxiv* 2020.03.31.018820 (2020)

808   doi:10.1101/2020.03.31.018820.

809   45. von Economo, C. & Koskinas, G. N. *Die Cytoarchitektonik der Hirnrinde des Erwachsenen*

810   *Menschen*. (J. Springer, 1925).

811   46. Somogyi, P., Freund, T. F. & Cowey, A. The axo-axonic interneuron in the cerebral cortex of the

812   rat, cat and monkey. *Neuroscience* **7**, 2577–2607 (1982).

813   47. Hof, P. R., Nimchinsky, E. A. & Morrison, J. H. Neurochemical phenotype of corticocortical

814   connections in the macaque monkey: quantitative analysis of a subset of neurofilament protein-

815   immunoreactive projection neurons in frontal, parietal, temporal, and cingulate cortices. *J. Comp.*

816   *Neurol.* **362**, 109–133 (1995).

817   48. Tsang, Y. M., Chiong, F., Kuznetsov, D., Kasarskis, E. & Geula, C. Motor neurons are rich in non-

818   phosphorylated neurofilaments: cross-species comparison and alterations in ALS. *Brain Res.* **861**,

45–58 (2000).

49. Preuss, T. M., Stepniewska, I., Jain, N. & Kaas, J. H. Multiple divisions of macaque precentral motor cortex identified with neurofilament antibody SMI-32. *Brain Research* vol. 767 148–153 (1997).

50. Scheibel, M. E., Davies, T. L., Lindsay, R. D. & Scheibel, A. B. Basilar dendrite bundles of giant pyramidal cells. *Exp. Neurol.* **42**, 307–319 (1974).

51. Baker, A. *et al.* Specialized Subpopulations of Deep-Layer Pyramidal Neurons in the Neocortex: Bridging Cellular Properties to Functional Consequences. *J. Neurosci.* **38**, 5441–5455 (2018).

52. Feng, G. *et al.* Imaging neuronal subsets in transgenic mice expressing multiple spectral variants of GFP. *Neuron* **28**, 41–51 (2000).

53. Groh, A. *et al.* Cell-Type Specific Properties of Pyramidal Neurons in Neocortex Underlying a Layout that Is Modifiable Depending on the Cortical Area. *Cerebral Cortex* vol. 20 826–836 (2010).

54. Tosches, M. A. *et al.* Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science* **360**, 881–888 (2018).

55. Arendt, D. *et al.* The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 (2016).

56. Herculano-Houzel, S., Catania, K., Manger, P. R. & Kaas, J. H. Mammalian Brains Are Made of These: A Dataset of the Numbers and Densities of Neuronal and Nonneuronal Cells in the Brain of Glires, Primates, Scandentia, Eulipotyphlans, Afrotherians and Artiodactyls, and Their Relationship with Body Mass. *Brain Behav. Evol.* **86**, 145–163 (2015).

57. Rivara, C.-B., Sherwood, C. C., Bouras, C. & Hof, P. R. Stereologic characterization and spatial distribution patterns of Betz cells in the human primary motor cortex. *Anat. Rec. A Discov. Mol. Cell. Evol. Biol.* **270**, 137–151 (2003).

58. Evarts, E. V. Representation of movements and muscles by pyramidal tract neurons of the precentral motor cortex. in *Neurophysiological basis of normal and abnormal motor activities* 215–253 (Raven Press New York, 1967).

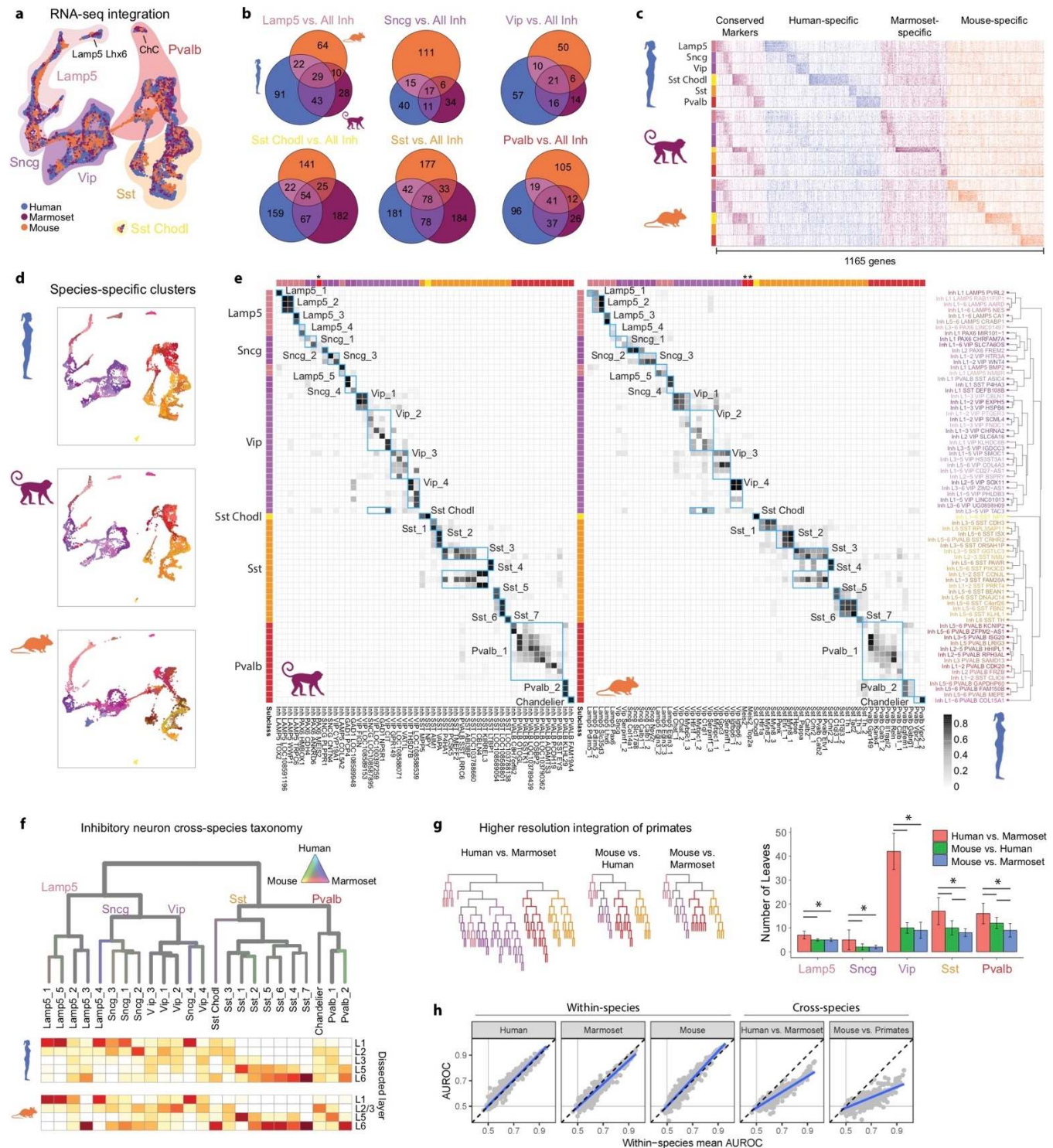59. Evarts, E. V. RELATION OF DISCHARGE FREQUENCY TO CONDUCTION VELOCITY IN

845   PYRAMIDAL TRACT NEURONS. *J. Neurophysiol.* **28**, 216–228 (1965).

846   60.   Lassek, A. M. THE PYRAMIDAL TRACT: A STUDY OF RETROGRADE DEGENERATION IN THE

847   MONKEY. *Arch NeurPsych* **48**, 561–567 (1942).

848   61.   Ting, J. T. *et al.* A robust ex vivo experimental platform for molecular-genetic dissection of adult

849   human neocortical cell types and circuits. *Sci. Rep.* **8**, 8407 (2018).

850   62.   Vigneswaran, G., Kraskov, A. & Lemon, R. N. Large Identified Pyramidal Cells in Macaque Motor

851   and Premotor Cortex Exhibit 'Thin Spikes': Implications for Cell Type Classification. *Journal of*

852   *Neuroscience* vol. 31 14235–14242 (2011).

853   63.   Krimer, L. S. *et al.* Cluster Analysis–Based Physiological Classification and Morphological

854   Properties of Inhibitory Neurons in Layers 2–3 of Monkey Dorsolateral Prefrontal Cortex. *J.*

855   *Neurophysiol.* **94**, 3009–3022 (2005).

856   64.   Rotaru, D. C. *et al.* Functional properties of GABA synaptic inputs onto GABA neurons in monkey

857   prefrontal cortex. *J. Neurophysiol.* **113**, 1850–1861 (2015).

858   65.   Gouwens, N. W. *et al.* Toward an integrated classification of neuronal cell types: morphoelectric

859   and transcriptomic characterization of individual GABAergic cortical neurons. *bioRxiv*

860   2020.02.03.932244 (2020) doi:10.1101/2020.02.03.932244.

861   66.   Fetz, E. E., Cheney, P. D., Mewes, K. & Palmer, S. Control of forelimb muscle activity by

862   populations of corticomotoneuronal and rubromotoneuronal cells. *Prog. Brain Res.* **80**, 437–49;

863   discussion 427–30 (1989).

864   67.   Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more

865   genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic*

866   *Acids Res.* **47**, D419–D426 (2019).

867   68.   Koopmans, F. *et al.* SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the

868   Synapse. *Neuron* **103**, 217–234.e4 (2019).

869   69.   Luo, C. *et al.* Single nucleus multi-omics links human cortical cell regulatory genome diversity to

870   disease risk variants. *bioRxiv* 2019.12.11.873398 (2019) doi:10.1101/2019.12.11.873398.

871   70. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic

872   Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).

873   71. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data

874   analysis. *Genome Biol.* **19**, 15 (2018).

875   72. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected

876   communities. *Sci. Rep.* **9**, 5233 (2019).

877   73. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using

878   Support Vector Machines. *Mach. Learn.* **46**, 389–422 (2002).

879   74. Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The Balanced Accuracy and Its

880   Posterior Distribution. in *2010 20th International Conference on Pattern Recognition* 3121–3124

881   (2010).

882   75. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A Python Toolbox to Tackle the

883   Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **18**, 1–5 (2017).

884   76. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes

885   using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).

886   77. He, Y. *et al.* Spatiotemporal DNA Methylome Dynamics of the Developing Mammalian Fetus.

887   doi:10.1101/166744.

888   78. McLeay, R. C. & Bailey, T. L. Motif Enrichment Analysis: a unified framework and an evaluation on

889   ChIP data. *BMC Bioinformatics* **11**, 165 (2010).

890   79. Palmer, C., Liu, C. & Chun, J. Nuclei Isolation for SNARE-seq2 v1 (protocols.io.8tvhwn6).

891   doi:10.17504/protocols.io.8tvhwn6.

892   80. Lun, A. T. L. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-

893   cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).

894   81. Gayoso, A. & Shor, J. *GitHub: DoubletDetection*. (2019). doi:10.5281/zenodo.2678042.

895   82. Hahne, F. & Ivanek, R. Visualizing Genomic Data Using Gviz and Bioconductor. *Methods Mol. Biol.*

896   **1418**, 335–351 (2016).

897   83. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation.

898   *Nature* **523**, 486–490 (2015).

899   84. Cusanovich, D. A. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial

900   cellular indexing. *Science* vol. 348 910–914 (2015).

901   85. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**,

902   355–364 (2014).

903   86. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that

904   Correlate with Prognosis. *Cell* **162**, 184–197 (2015).

905   87. Graybuck, L. T. *et al.* Prospective, brain-wide labeling of neuronal subclasses with enhancer-driven

906   AAVs. (2019) doi:10.1101/525014.

907   88. Ting, J. T., Daigle, T. L., Chen, Q. & Feng, G. Acute brain slice methods for adult and aging

908   animals: application of targeted patch clamp analysis and optogenetics. *Methods Mol. Biol.* **1183**,

909   221–242 (2014).

910   89. Gouwens, N. W. *et al.* Classification of electrophysiological and morphological neuron types in the

911   mouse visual cortex. *Nat. Neurosci.* **22**, 1182–1195 (2019).

912   90. Chan, K. Y. *et al.* Engineered AAVs for efficient noninvasive gene delivery to the central and

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

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

915   repetitive firing. *Trends Neurosci.* **24**, 517–526 (2001).
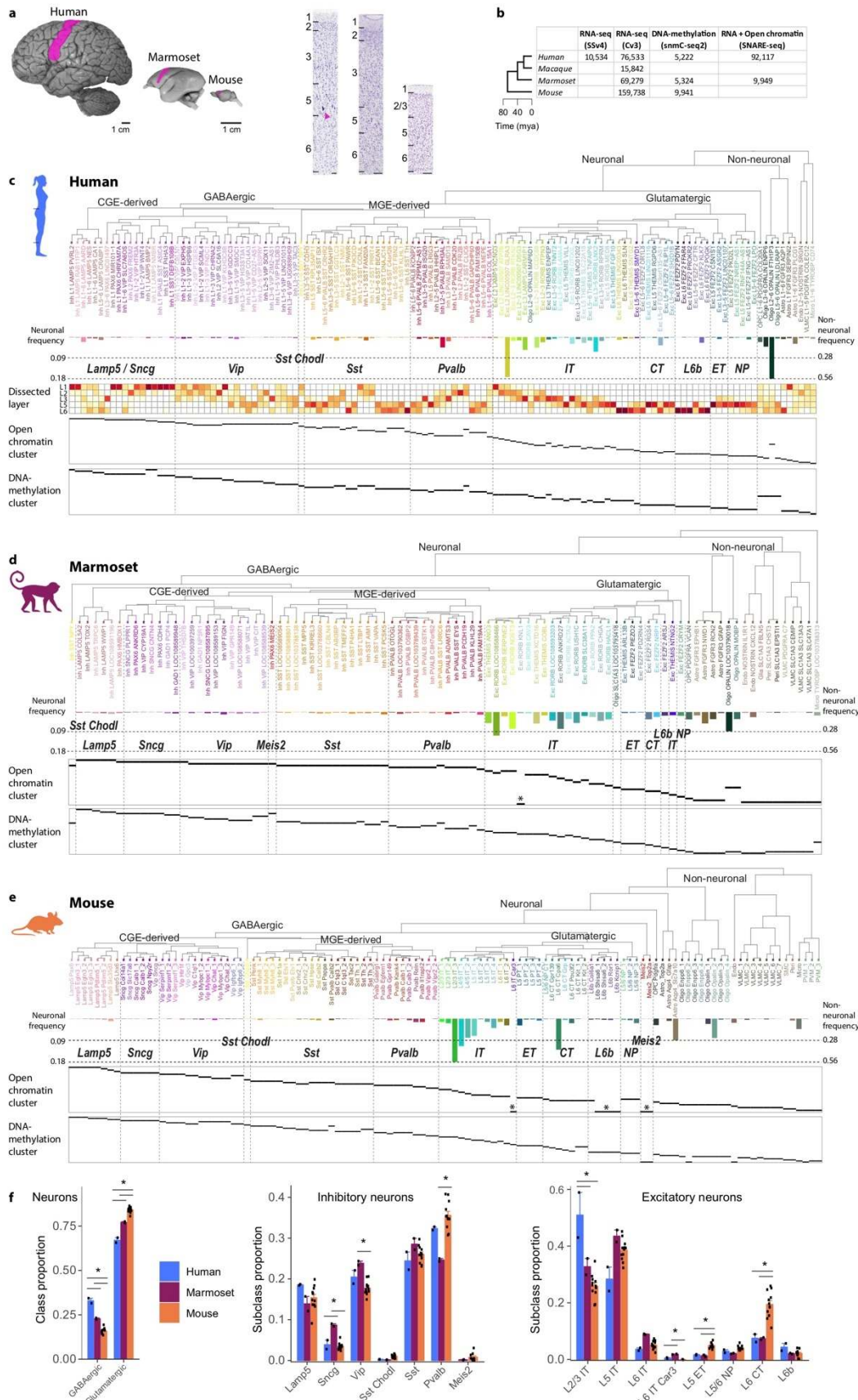
917 **Figure 1. Molecular taxonomy of cell types in M1 of human, marmoset, and mouse. a**, M1

918 highlighted in lateral views of neocortex across species. Nissl-stained sections of M1 annotated with

919 layers and showing the relative expansion of cortical thickness, particularly L2 and L3 in primates, and

920 large pyramidal neurons or 'Betz' cells in human L5 (arrowhead). Scale bars, 100 μm. **b**, Phylogeny of

921 species and number of nuclei included in analysis for each molecular assay. All assays used nuclei

922 isolated from the same donors for human and marmoset. SSv4, SMART-Seq v4; Cv3, Chromium v3;

923 mya, millions of years ago. **c-e**, Dendrograms of cell types defined by RNA-seq (Cv3) for human (**c**),

924 marmoset (**d**), and mouse (**e**) and annotated with cluster frequency and dissected layer (human only).

925 Epigenomic clusters (in rows) aligned to RNA-seq clusters as indicated by horizontal black bars.

926 Asterisks denote RNA clusters that lack corresponding epigenomic clusters. **f**, Relative proportions of

927 cells in several classes and subclasses were significantly different between species based on an

928 ANOVA followed by Tukey's HSD tests (asterisk, adjusted $P < 0.05$).

929

930

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.

934 DEGs were determined by ROC tests of each subclass versus all other GABAergic subclasses within a

935 species. **c,** Heatmap of all DEGs from **b** ordered by subclass and species enrichment. Heatmap shows

936 gene expression scaled by column for up to 50 randomly sampled nuclei from each subclass for each

937 species. **d,** UMAP projection from **a**, separated by species, and colored by within-species clusters. **e,**

938 Cluster overlap heatmap showing the proportion of nuclei in each pair of species clusters that are

939 mixed in the cross-species integrated space. Cross-species consensus clusters are indicated by

940 labeled blue boxes. Human clusters (rows) are ordered by the dendrogram reproduced from **Figure 1c**.

941 Marmoset (left columns) and mouse (right columns) clusters are ordered to align with human clusters.

942 Color bars at top and left indicate subclasses of within-species clusters. Asterisks indicate marmoset

943 and mouse Meis2 subclasses, which were not present in human. **f,** Dendrogram of GABAergic neuron

944 consensus clusters with edges colored by species mixture (grey, well mixed). Below: Estimated spatial

945 distributions of clusters based on layer dissections in human (top) and mouse (bottom). **g,**

946 Dendrograms of pairwise species integrations, colored by subclass. Bar plots quantifying well-mixed

947 leaf nodes. Significant differences (adjusted P < 0.05, Tukey's HSD test) between species are indicated

948 for each subclass. **h,** Scatter plots of MetaNeighbor analysis showing the performance (AUROC) of

949 gene-sets to classify GABAergic neurons within and between species. Blue lines, linear regression fits;

950 black lines, mean within species performance; grey lines, performance equivalent to chance.

951

952
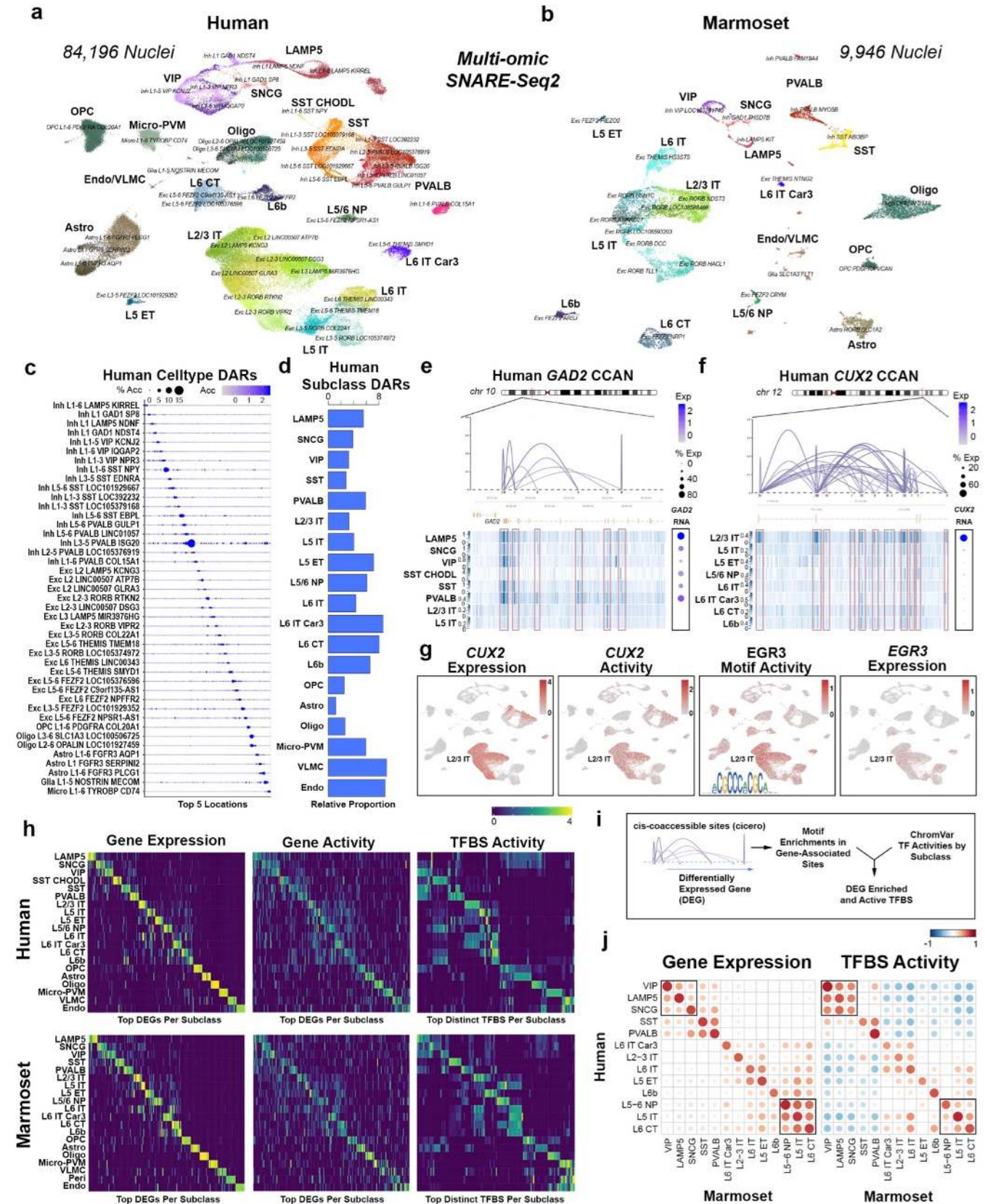
953   **Figure 3. Dual-omic expression and chromatin accessibility reveals regulatory processes**

954   **defining M1 cell types. a-b** UMAP visualizations of human (**a**) and marmoset (**b**) M1 SNARE-Seq2

955   data (2 individuals per species) indicating both subclass and accessibility-level cluster identities. **c**, Dot

956   plot showing proportion and scaled average accessibility of differentially accessible regions (DARs)

957   identified between human AC clusters (adjusted P < 0.001, log-fold change > 1, top 5 distinct sites per

958   cluster). **d**, Proportion of total human or marmoset DARs identified between subclasses (adjusted P <

959   0.001, log-fold change > 1) after normalization to cluster sizes. **e-f**, Connection plots for cis-co-

960   accessible network (CCAN) sites associated with the human *GAD2* (**e**) and *CUX2* (**f**) genes.

961   Corresponding AC read pile-up tracts for GABAergic and select glutamatergic subclasses are shown.

962   Right panels are dot plots showing the percentage of expressing nuclei and average gene expression

963   values (log scale) for *GAD2* or *CUX2* within each of the clusters indicated. **g**, UMAP plots as in Figure

964   5a (human) showing (scaled from low—gray to high—red) *CUX2* gene expression (RNA) and activity

965   level predicted from AC data. UMAP plots for activity level of the EGR3-binding motif, identified using

966   chromVAR and found to be enriched within *CUX2* co-accessible sites, and the corresponding

967   expression (RNA) of the *EGR3* gene are shown. **h**, Heatmaps for human (top) and marmoset (bottom)

968   showing TFBS enrichments, according to the scheme outlined in (**i**), within genes differentially

969   expressed between subclasses and having at least two cis-co-accessible sites. Left panels show

970   scaled average (log scale) gene expression values (RNA) for the top DEGs (adjusted P < 0.05, log-fold

971   change > 1, top 10 distinct sites per cluster visualized), middle panels show the corresponding scaled

972   average cicero gene activity scores and the right panels show scaled values for the corresponding top

973   distinct chromVAR TFBS activities (adjusted P < 0.05, log-fold change > 0.5, top 10 distinct sites per

974   cluster visualized). **j**, Correlation plots comparing scaled average gene expression profiles (left panel)

975   or chromVAR TFBS activity scores (right panel) between human and marmoset matched subclasses.

976

977
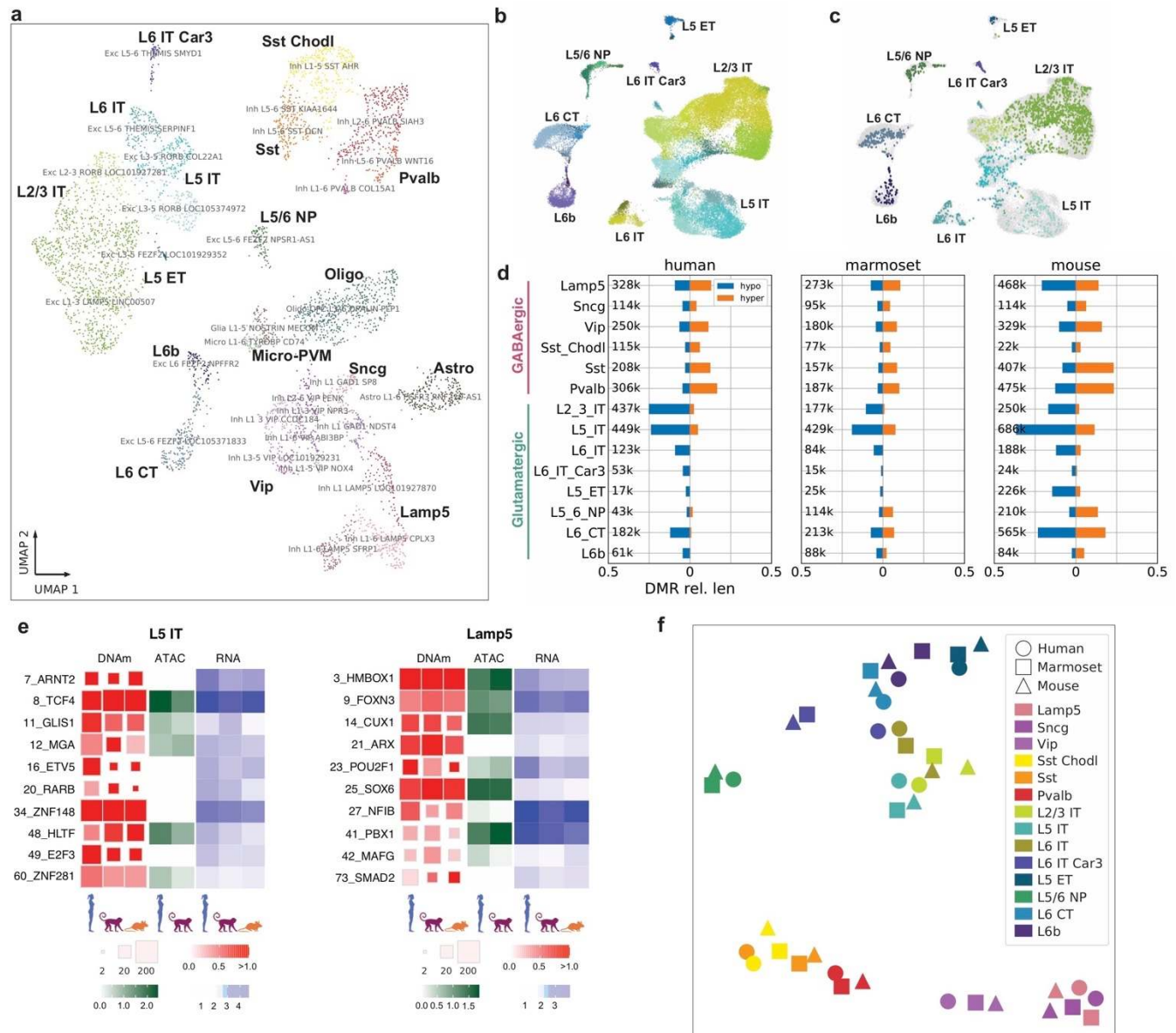
Figure 4. DNA methylation differences across clusters and species. a, UMAP visualization of

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

984   are listed (k, thousands). **e**, Distinct TF motif enrichment for L5 IT and *Lamp5* subclasses across

985   species. **f**, t-SNE visualization of subclass TF motif enrichment that is conserved across species.

986
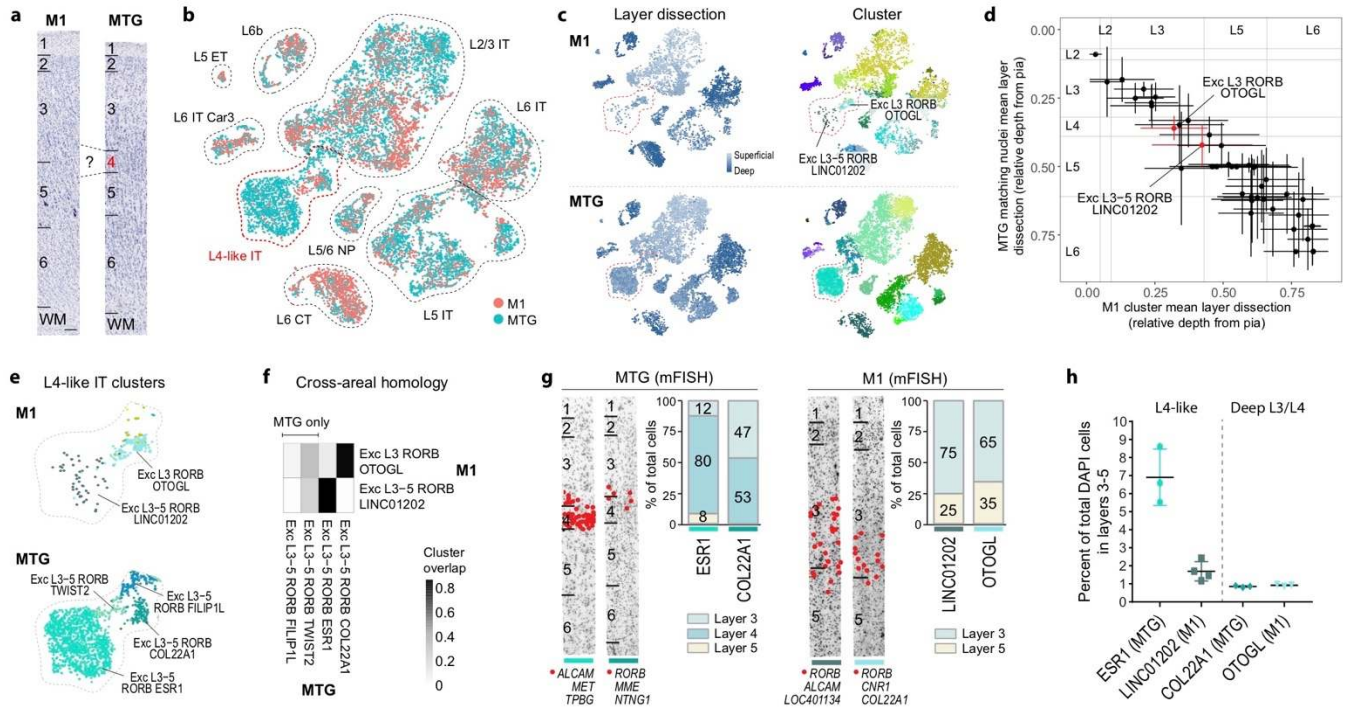
Figure 5. L4-like neurons identified in M1 based on cross-areal cell type homology. a, t-SNE projection of glutamatergic neuronal nuclei from M1 and MTG based on similarity of integrated expression levels. Nuclei are intermixed within all cell subclasses. **b**, Nuclei annotated based on the relative 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

002   inverted image of a DAPI-stained cortical column. **h**, ISH estimated frequencies of homologous clusters

003   shows M1 has a 4-fold sparser L4-like type and similarly rare deep L3 type.

004
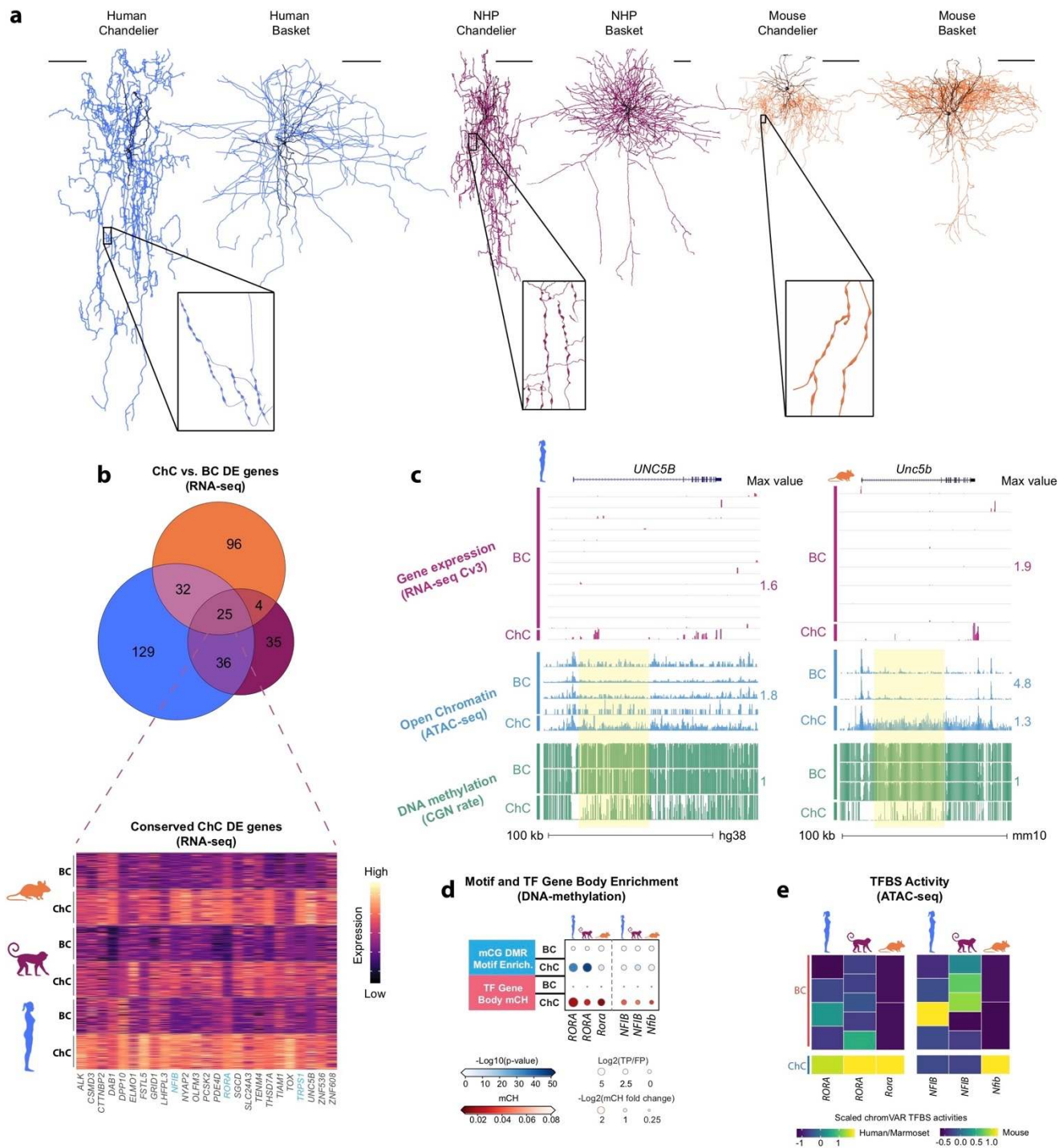
Figure 6. Chandelier neurons have a core set of conserved molecular features. a, Representative ultrastructure reconstructions of a ChC and BC from human (left), macaque (middle), and mouse (right). Insets show higher magnification of ChC axon cartridges. Macaque reconstructions were from source data available in Neuromorpho [63,64]. Mouse cells also appear in [65]. **b,** Venn diagram indicating

010   the number of shared ChC-enriched genes across species (top). DEGs were determined by a ROC test

011   of ChCs against BCs within a species. Heatmap showing scaled expression of the 25 conserved DEGs

012   in 100 randomly selected ChC and BC nuclei for each species (bottom); transcription factors are

013   colored in blue. **c**, Genome browser tracks showing *UNC5B* locus in human (left) and mouse (right)

014   ChCs and BCs. Tracks show aligned transcripts, regions of accessible chromatin, CGN methylation

015   rate, and CHN methylation rate. Yellow highlights mark examples of ChC-enriched regions of

016   accessible chromatin with hypo-methylated CGN. **d,** Heatmaps of TF gene body hypo-methylation

017   (mCH) state (bottom half, red) and genome-wide enrichment of TF motif across mCG DMRs in ChCs

018   and BCs (top half, blue). **e,** Scaled TFBS activities identified from SNARE-seq2 for human and

019   marmoset according to the scheme in Figure 4i and from mouse snATAC-seq data, using genes

020   enriched in ChC versus BC (Supplementary Table 19). Rows correspond to BC and ChC clusters

021   identified in snATAC-seq and SNARE-seq2 datasets.

022
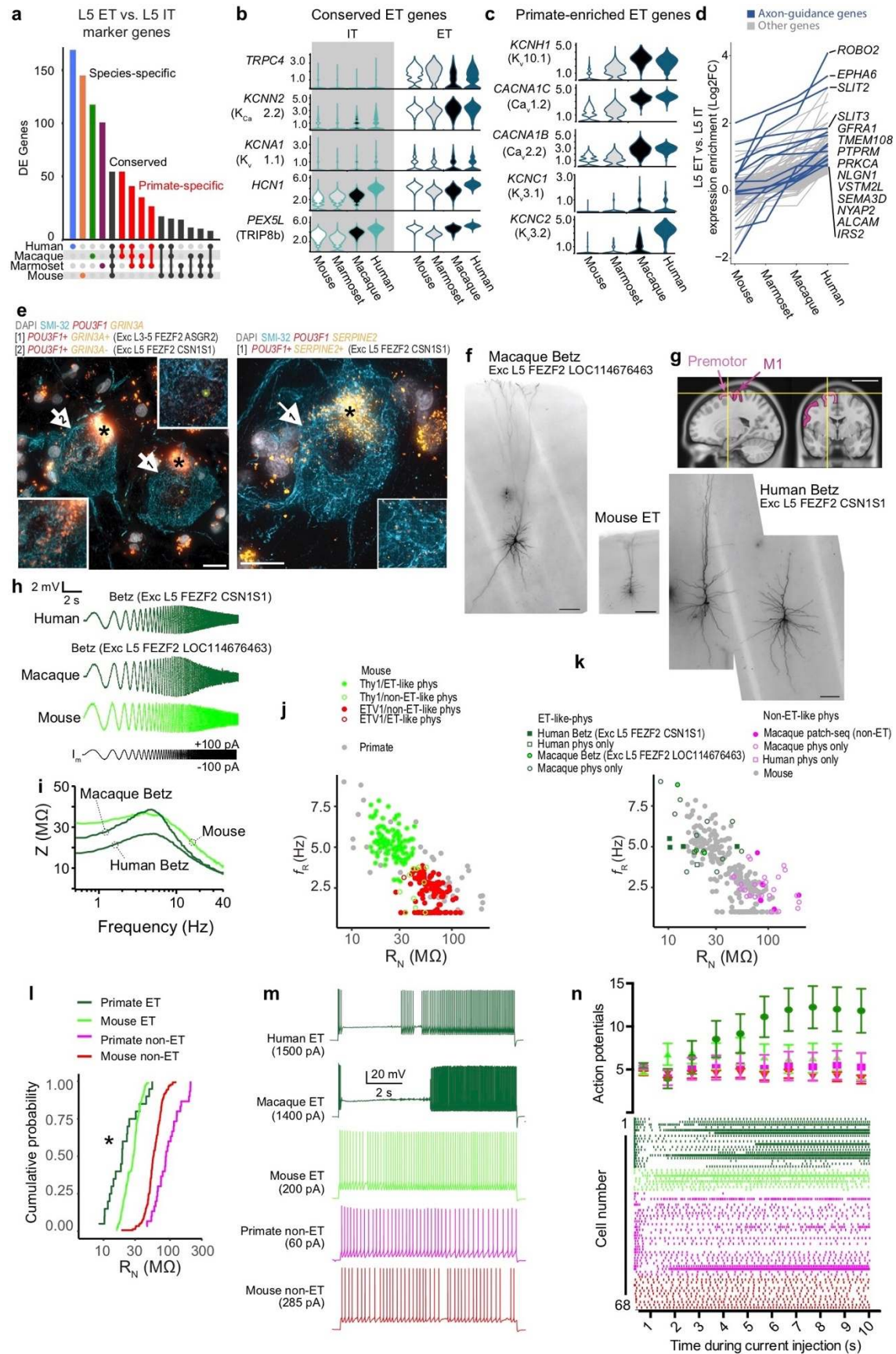
024   **Figure 7. Betz cells have specialized molecular and physiological properties. a,** Upset plot

025   showing conserved and divergent L5 ET glutamatergic neuron marker genes. DEGs were determined

026   by performing a ROC test between L5 ET and L5 IT within each species. **b, c,** Violin plots of ion

027   channel-related gene expression for genes that are enriched in (**b**) ET versus IT neurons and in (**c**)

028   primate versus mouse ET neurons. Protein names are in parentheses. **d,** Line graph of 131 genes with

029   expression enrichment in L5 ET versus IT neurons in human (>0.5 $\log_2$ fold-change) that decreases

030   with evolutionary distance from human. **e,** Two example photomicrographs of ISH labeled, SMI-32 IF

031   stained Betz cells in L5 human M1. Cells corresponding to two L5 ET clusters are labeled based on two

032   sets of marker genes. Insets show higher magnification of ISH in corresponding cells. Asterisks mark

033   lipofuscin; scale bar, 20 μm. **f, g,** Exemplar biocytin fills obtained from patch-seq experiments in

034   human, macaque and mouse brain slices. The example human and macaque neurons mapped to a

035   Betz cell transcriptomic cell type. Scale bars, 200 μm. **g,** MRI image in sagittal and coronal planes and

036   approximate location of excised premotor cortex tissue (yellow lines) and adjacent M1. **h,** Voltage

037   responses to a chirp stimulus for the neurons shown in **f** and **g** (left human neuron). **i,** Corresponding

038   ZAP profiles. All neurons were clustered into putative ET and non-IT neurons based upon their

039   resonant frequency and input resistance. **j,** For mouse L5 neurons (*Thy1*-YFP line H, n=117; *Etv1*-

040   EGFP line, n=123; unlabeled, n=21) 99.2 % of neurons in the *Etv1*-EGFP line possessed non-ET-like

041   physiology, whereas, 91.4% of neurons in the *Thy1*-YFP line H had ET-like physiology. **k,** For primate

042   L5 neurons (human, n=8, macaque n=42), all transcriptomically-defined Betz cells (human, n=4,

043   macaque n=3) had ET-like physiology (human n=6, macaque, n=14), whereas all transcriptomically-

044   defined non-ET neurons (human n=2, macaque n=3) had non-ET like physiology (human n=2,

045   macaque n=28). **l,** Cumulative probability distribution of L5 neuron input resistance for primate versus

046   mouse. * p = 0.0064, Kolmogorov-Smirnov test between mouse and primate ET neurons. **m,** Example

047   voltage responses to 10s step current injections for monkey, mouse and human ET and non-ET

048   neurons. The amplitude of the current injection was adjusted to produce ~5 spikes during the first

049   second. **n,** Raster plot (below) and average firing rate (above) during 1 s epochs during the 10s DC

050   current injection. Primate ET neurons (pooled data from human and macaque) displayed a distinctive

051   decrease followed by a pronounced increase in firing rate over the course of the current injection.

052   Notably, a similar biphasic-firing pattern is observed in macaque corticospinal neurons *in vivo* during

053   prolonged motor movements [66], suggesting that the firing pattern of these neurons during behavior is

054   intimately tied to their intrinsic membrane properties.

055

056 Methods

057 ***Ethical compliance***

058 Postmortem adult human brain tissue was collected after obtaining permission from decedent next-of-

059 kin. Postmortem tissue collection was performed in accordance with the provisions of the United States

060 Uniform Anatomical Gift Act of 2006 described in the California Health and Safety Code section 7150

061 (effective 1/1/2008) and other applicable state and federal laws and regulations. The Western

062 Institutional Review Board reviewed tissue collection processes and determined that they did not

063 constitute human subjects research requiring institutional review board (IRB) review.

064

065 ***Postmortem human tissue specimens***

066 Male and female donors 18–68 years of age with no known history of neuropsychiatric or neurological

067 conditions ('control' cases) were considered for inclusion in the study (Extended Data Table 1). Routine

068 serological screening for infectious disease (HIV, Hepatitis B, and Hepatitis C) was conducted using

069 donor blood samples and only donors negative for all three tests were considered for inclusion in the

070 study. Only specimens with RNA integrity (RIN) values ≥7.0 were considered for inclusion in the study.

071 Postmortem brain specimens were processed as previously described [2]. Briefly, coronal brain slabs

072 were cut at 1cm intervals and frozen for storage at -80°C until the time of further use. Putative hand and

073 trunk-lower limb regions of the primary motor cortex were identified, removed from slabs of interest, and

074 subdivided into smaller blocks. One block from each donor was processed for cryosectioning and

075 fluorescent Nissl staining (Neurotrace 500/525, ThermoFisher Scientific). Stained sections were

076 screened for histological hallmarks of primary motor cortex. After verifying that regions of interest

077 contained M1, blocks were processed for nucleus isolation as described below.

078

079 ***Human RNA-sequencing, QC and clustering***

080   SMART-seq v4 nucleus isolation and sorting. Vibratome sections were stained with fluorescent Nissl

081   permitting microdissection of individual cortical layers (dx.doi.org/10.17504/protocols.io.7aehibe).

082   Nucleus isolation was performed as previously described (dx.doi.org/10.17504/protocols.io.ztqf6mw).

083   NeuN staining was carried out using mouse anti-NeuN conjugated to PE (FCMAB317PE, EMD

084   Millipore) at a dilution of 1:500. Control samples were incubated with mouse IgG1k-PE Isotype control

085   (BD Biosciences, 555749). DAPI (4',6-diamidino-2-phenylindole dihydrochloride, ThermoFisher

086   Scientific, D1306) was applied to nuclei samples at a concentration of 0.1µg/ml. Single-nucleus sorting

087   was carried out on either a BD FACSAria II SORP or BD FACSAria Fusion instrument (BD

088   Biosciences) using a 130 µm nozzle. A standard gating strategy based on DAPI and NeuN staining was

089   applied to all samples as previously described [2]. Doublet discrimination gates were used to exclude

090   nuclei aggregates.

091

092   SMART-seq v4 RNA-sequencing. The SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (Takara

093   #634894) was used per the manufacturer's instructions. Standard controls were processed with each

094   batch of experimental samples as previously described. After reverse transcription, cDNA was amplified

095   with 21 PCR cycles. The NexteraXT DNA Library Preparation (Illumina FC-131-1096) kit with

096   NexteraXT Index Kit V2 Sets A-D (FC-131-2001, 2002, 2003, or 2004) was used for sequencing library

097   preparation. Libraries were sequenced on an Illumina HiSeq 2500 instrument using Illumina High

098   Output V4 chemistry.

099

100   SMART-seq v4 gene expression quantification. Raw read (fastq) files were aligned to the GRCh38

101   human genome sequence (Genome Reference Consortium, 2011) with the RefSeq transcriptome

102   version GRCh38.p2 (current as of 4/13/2015) and updated by removing duplicate Entrez gene entries

103   from the gtf reference file for STAR processing. For alignment, Illumina sequencing adapters were

104   clipped from the reads using the fastqMCF program. After clipping, the paired-end reads were mapped

105   using Spliced Transcripts Alignment to a Reference (STAR) using default settings. Reads that did not

106   map to the genome were then aligned to synthetic construct (i.e. ERCC) sequences and the *E. coli*

107   genome (version ASM584v2). Quantification was performed using summerizeOverlaps from the R

108   package GenomicAlignments. Expression levels were calculated as counts per million (CPM) of exonic

109   plus intronic reads.

110

111   10x Chromium RNA-sequencing. Nucleus isolation for 10x Chromium RNA-sequencing was conducted

112   as described (dx.doi.org/10.17504/protocols.io.y6rfzd6). After sorting, single-nucleus suspensions were

113   frozen in a solution of 1X PBS, 1% BSA, 10% DMSO, and 0.5% RNAsin Plus RNase inhibitor

114   (Promega, N2611) and stored at -80°C. At the time of use, frozen nuclei were thawed at 37°C and

115   processed for loading on the 10x Chromium instrument as described

116   (dx.doi.org/10.17504/protocols.io.nx3dfqn). Samples were processed using the 10x Chromium Single-

117   Cell 3' Reagent Kit v3. 10x chip loading and sample processing was done according to the

118   manufacturer's protocol. Gene expression was quantified using the default 10x Cell Ranger v3 pipeline

119   except substituting the curated genome annotation used for SMART-seq v4 quantification. Introns were

120   annotated as "mRNA", and intronic reads were included in expression quantification.

121

122   Quality control of RNA-seq data. Nuclei were included for analysis if they passed all QC criteria.

123   SMART-seq v4 criteria:

124   > 30% cDNA longer than 400 base pairs

125   > 500,000 reads aligned to exonic or intronic sequence

126   > 40% of total reads aligned

127   > 50% unique reads

128   > 0.7 TA nucleotide ratio

129   Cv3 criteria:

130   > 500 (non-neuronal nuclei) or > 1000 (neuronal nuclei) genes detected

131   < 0.3 doublet score

132

133   Clustering RNA-seq data. Nuclei passing QC criteria were grouped into transcriptomic cell types using

134   a previously reported iterative clustering procedure (Tasic et al. 2018; Hodge, Bakken et al., 2019).

135   Briefly, intronic and exonic read counts were summed, and $\log_2$-transformed expression was centered

136   and scaled across nuclei. X- and Y-chromosomes and mitochondrial genes were excluded to avoid

137   nuclei clustering based on sex or nuclei quality. DEGs were selected, principal components analysis

138   (PCA) reduced dimensionality, and a nearest neighbor graph was built using up to 20 principal

139   components. Clusters were identified with Louvain community detection (or Ward's hierarchical

140   clustering if N < 3000 nuclei), and pairs of clusters were merged if either cluster lacked marker genes.

141   Clustering was applied iteratively to each subcluster until clusters could not be further split.

142

143   Cluster robustness was assessed by repeating iterative clustering 100 times for random subsets of

144   80% of nuclei. A co-clustering matrix was generated that represented the proportion of clustering

145   iterations that each pair of nuclei were assigned to the same cluster. We defined consensus clusters by

146   iteratively splitting the co-clustering matrix as described (Tasic et al. 2018; Hodge, Bakken et al., 2019).

147   The clustering pipeline is implemented in the R package "scrattch.hicat", and the clustering method is

148   provided by the "run_consensus_clust" function (https://github.com/AllenInstitute/scrattch.hicat).

149

150   Clusters were curated based on QC criteria or cell class marker expression (*GAD1*, *SLC17A7*,

151   *SNAP25*). Clusters were identified as donor-specific if they included fewer nuclei sampled from donors

152   than expected by chance. To confirm exclusion, clusters automatically flagged as outliers or donor-

153   specific were manually inspected for expression of broad cell class marker genes, mitochondrial genes

154   related to quality, and known activity-dependent genes.

155

156   **Marmoset sample processing and nuclei isolation**

157   Marmoset experiments were approved by and in accordance with Massachusetts Institute of

158   Technology IACUC protocol number 051705020. Two adult marmosets (2.3 and 3.1 years old; one

159   male, one female; Extended Data Table 2) were deeply sedated by intramuscular injection of ketamine

160   (20-40 mg/kg) or alfaxalone (5-10 mg/kg), followed by intravenous injection of sodium pentobarbital

161   (10–30 mg/kg). When pedal withdrawal reflex was eliminated and/or respiratory rate was diminished,

162   animals were transcardially perfused with ice-cold sucrose-HEPES buffer. Whole brains were rapidly

163   extracted into fresh buffer on ice. Sixteen 2-mm coronal blocking cuts were rapidly made using a

164   custom-designed marmoset brain matrix. Coronal slabs were snap-frozen in liquid nitrogen and stored

165   at -80°C until use.

166

167   As with human samples, M1 was isolated from thawed slabs using fluorescent Nissl staining

168   (Neurotrace 500/525, ThermoFisher Scientific). Stained sections were screened for histological

169   hallmarks of primary motor cortex. Nuclei were isolated from the dissected regions following this

170   protocol: https://www.protocols.io/view/extraction-of-nuclei-from-brain-tissue-2srged6 and processed

171   using the 10x Chromium Single-Cell 3' Reagent Kit v3. 10x chip loading and sample processing was

172   done according to the manufacturer's protocol.

173

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

175   RNA-sequencing. Libraries were sequenced on NovaSeq S2 instruments (Illumina). Raw sequencing

176   reads were aligned to calJac3. Mitochondrial sequence was added into the published reference

177   assembly. Human sequences of RNR1 and RNR2 (mitochondrial) and RNA5S (ribosomal), were

178   aligned using gmap to the marmoset genome and added to the calJac3 annotation. Reads that mapped

179   to exons or introns of each assembly were assigned to annotated genes. Libraries were sequenced to a

180   median read depth of 5.95 reads per unique molecular index (UMI). The alignment pipeline can be

181   found at https://github.com/broadinstitute/Drop-seq.

182

183   Cell filtering. Cell barcodes were filtered to distinguish true nuclei barcodes from empty beads and PCR

184   artifacts by assessing proportions of ribosomal and mitochondrial reads, ratio of intronic/exonic reads (>

185   50% of intronic reads), library size (> 1000 UMIs) and sequencing efficiency (true cell barcodes have

186   higher reads/UMI). The resulting digital gene expression matrix (DGE) from each library was carried

187   forward for clustering.

188

189   Clustering. Clustering analysis proceeded as in Krienen et al (2019, bioRxiv). Briefly, independent

190   component analysis. (ICA, using the fastICA package in R) was performed jointly on all marmoset

191   DGEs after normalization and variable gene selection as in (Saunders et al 2018, *Cell*). The first-round

192   clustering resulted in 15 clusters corresponding to major cell classes (neurons, glia, endothelial). Each

193   cluster was curated as in (Saunders et al 2018, *Cell*) to remove doublets and outliers. Independent

194   components (ICs) were partitioned into those reflecting artifactual signals (e.g. those for which cell

195   loading indicated replicate or batch effects). Remaining ICs were used to determine clustering (Louvain

196   community detection algorithm igraph package in R); for each cluster nearest neighbor and resolution

197   parameters were set to optimize 1:1 mapping between each IC and a cluster.

198

199   ***Mouse snRNA-seq data generation and analysis***

200   Single-nuclei were isolated from mouse primary motor cortex, gene expression was quantified using

201   Cv3 and SSv4 RNA-sequencing, and transcriptomic cell types and dendrogram were defined as

202   described in a companion paper [6].

203

204   ***Integrating and clustering human Cv3 and SSv4 snRNA-seq datasets***

205   To establish a set of human consensus cell types, we performed a separate integration of snRNA-seq

206   technologies on the major cell classes (Glutamatergic, GABAergic, and Non-neuronal). Broadly, this

207   integration is comprised of 6 steps: (1) subsetting the major cell class from each technology (e.g. Cv3

208   GABAergic and SSv4 GABAergic); (2) finding marker genes for all clusters within each technology; (3)

209 integrating both datasets with Seurat's standard workflow using marker genes to guide integration

210 (Seurat 3.1.1); (4) overclustering the data to a greater number of clusters than were originally identified

211 within a given individual dataset; (5) finding marker genes for all integrated clusters; and (6) merging

212 similar integrated clusters together based on marker genes until all merging criteria were sufficed,

213 resulting in the final human consensus taxonomy.

214

215 More specifically, each expression matrix was $\log_2(\text{CPM} + 1)$ transformed then placed into a Seurat

216 object with accompanying metadata. Variable genes were determined by downsampling each

217 expression matrix to a maximum of 300 nuclei per scrattch.hicat-defined cluster (from a previous step;

218 see scrattch.hicat clustering) and running select_markers (scrattch.io 0.1.0) with n set to 20, to

219 generate a list of up to 20 marker genes per cluster. The union of the Cv3 and SSv4 gene lists were

220 then used as input for anchor finding, dimensionality reduction, and Louvain clustering of the full

221 expression matrices. We used 100 dimensions for steps in the workflow, and 100 random starts during

222 clustering. Louvain clustering was performed to overcluster the dataset to identify more integrated

223 clusters than the number of scrattch.hicat-defined clusters. For example, GABAergic neurons had 79

224 and 37 scrattch.hicat-defined clusters, 225 overclustered integrated clusters, and 72 final human

225 consensus clusters after merging for Cv3 and SSv4 datasets, respectively. To merge the overclustered

226 integrated clusters, up to 20 marker genes were found for each cluster to establish the neighborhoods

227 of the integrated dataset. Clusters were then merged with their nearest neighbor if there were not a

228 minimum of ten Cv3 and two SSv4 nuclei in a cluster, and a minimum of 4 DEGs that distinguished the

229 query cluster from the nearest neighbor (note: these were the same parameters used to perform the

230 initial scrattch.hicat clustering of each dataset).

231

232 ***Integrating and clustering MTG and M1 SSv4 snRNA-seq datasets***

233 To compare cell types between our M1 human cell type taxonomy and our previously described human

234 MTG taxonomy [2], we used Seurat's standard integration workflow to perform a supervised integration

235   of the M1 and MTG SSv4 datasets. Intronic and exonic reads were summed into a single expression

236   matrix for each dataset, CPM normalized, and placed into a Seurat object with accompanying

237   metadata. All nuclei from each major cell class were integrated and clustered separately. Up to 100

238   marker genes for each cluster within each dataset were identified, and the union of these two gene lists

239   was used as input to guide alignment of the two datasets during integration, dimensionality reduction,

240   and clustering steps. We used 100 dimensions for all steps in the workflow.

241

242   ***Integrating Cv3 snRNA-seq datasets across species***

243   To identify homologous cell types across species, we used Seurat's SCTransform workflow to perform

244   a separate supervised integration on each cell class across species. Raw expression matrices were

245   reduced to only include genes with one-to-one orthologs defined in the three species (14,870 genes;

246   downloaded from NCBI Homologene in November, 2019) and placed into Seurat objects with

247   accompanying metadata. To avoid having one species dominate the integrated space and to account

248   for potential differences in each species' clustering resolution, we downsampled the number of nuclei to

249   have similar numbers across species at the subclass level (e.g. *Lamp5*, *Sst*, L2/3 IT, L6b, etc.). The

250   species with the largest number of clusters under a given subclass was allowed a maximum of 200

251   nuclei per cluster. The remaining species then split this theoretical maximum (200 nuclei times the max

252   number of clusters under subclass) evenly across their clusters. For example, the L2/3 IT subclass had

253   8, 4, and 3 clusters for human, marmoset, and mouse, respectively. All species were allowed a

254   maximum of 1600 L2/3 IT nuclei total; or a maximum of 200 human, 400 marmoset, and 533 mouse

255   nuclei per cluster. To integrate across species, all Seurat objects were merged and normalized using

256   Seurat's SCTransform function. To better guide the alignment of cell types from each species, we found

257   up to 100 marker genes for each cluster within a given species. We used the union of these gene lists

258   as input for integration and dimensionality reduction steps, with 30 dimensions used for integration and

259   100 for dimensionality reduction and clustering. Clustering the human-marmoset-mouse integrated

260   space provided an additional quality control mechanism, revealing numerous small, species-specific

261   integrated clusters that contained only low-quality nuclei (low UMIs and genes detected). We excluded

262   4836 nuclei from the marmoset dataset that constituted low-quality integrated neuronal clusters.

263

264   To identify which clusters in our three species taxonomy aligned with macaque clusters from our L5

265   dissected Cv3 dataset, we performed an identical integration workflow on Glutamatergic neurons as

266   was used for the three species integration. Macaque clusters were assigned subclass labels based on

267   their corresponding alignment with subclasses from the other species. The annotated L5 dissected

268   macaque Cv3 dataset was then used as a reference for mapping macaque patch-seq nuclei (see

269   section below).

270

271   ***Estimation of cell type homology***

272   To identify homologous groups from different species, we applied a tree-based method

273   (https://github.com/AllenInstitute/BICCN_M1_Evo and package:

274   https://github.com/huqiwen0313/speciesTree). In brief, the approach consists of 4 steps: 1) metacell

275   clustering, 2) hierarchical reconstruction of a metacell tree, 3) measurements of species mixing and

276   stability of splits and 4) dynamic pruning of the hierarchical tree.

277

278   Firstly, to reduce noise in single-cell datasets and to remove species-specific batch effects, we

279   clustered cells into small highly similar groups based on the integrated matrix generated by Seurat, as

280   described in the previous section. These cells were further aggregated into metacells and the

281   expression values of the metacells were calculated by averaging the gene expression of individual cells

282   that belong to each metacell. Correlation was calculated based on the metacell gene expression matrix

283   to infer the similarity of each metacell cluster. Then hierarchical clustering was performed based on the

284   metacell gene expression matrix using Ward's method. For each node or corresponding branch in the

285   hierarchical tree, we calculated 3 measurements, and the hierarchical tree was visualized based on

286   these measurements: 1) Cluster size visualized as the thickness of branches in the tree; 2) Species

287 mixing calculated based on entropy of the normalized cell distribution and visualized as the color of

288 each node and branch; 3) Stability of each node. The entropy of cells was calculated as: $H =$

289 $-\sum_i p_i log p_i$, where p_i is the probability of cells from one species appearing among all the cells in a

290 node. We assessed the node stability by evaluating the agreement between the original hierarchical

291 tree and a result on a subsampled dataset calculated based on the optimal subtree in the subsampled

292 hierarchical trees derived from subsampling 95% of cells in the original dataset. The entire subsampling

293 process was repeated 100 times and the mean stability score for every node in the original tree was

294 calculated. Finally, we recursively searched each node in the tree. If the heuristic criteria (see below)

295 were not met for any node below the upper node, the entire subtree below the upper node was pruned

296 and all the cells belonging to this subtree were merged into one homologous group.

297 To identify robust homologous groups, we applied criteria in two steps to dynamically search the cross-

298 species tree. Firstly, for each node in the tree, we computed the mixing of cells from 3 species based

299 on entropy and set it as a tuning parameter. For each integrated tree, we tuned the entropy parameter

300 to make sure the tree method generated the highest resolution of homologous clusters without losing

301 the ability to identify potential species-specific clusters. Nodes with entropy larger than 2.9 (for inhibitory

302 neurons) or 2.75 (for excitatory neurons) were considered as well-mixed nodes. For example, an

303 entropy of 2.9 corresponded to a mixture of human, marmoset, and mouse equal to (0.43, 0.37, 0.2) or

304 (0.38, 0.30, 0.32). We recursively searched all the nodes in the tree until we found the node nearest the

305 leaves of the tree that was well-mixed among species, and this node was defined as a well-mixed

306 upper node. Secondly, we further checked the within-species cell composition for the subtrees below

307 the well-mixed upper node to determine if further splits were needed. For the cells on the subtrees

308 below the well mixed upper node, we measured the purity of within-species cell composition by

309 calculating the percentage of cells that fall into a specific sub-group in each individual species. If the

310 purity for any species was larger than 0.8, we went one step further below the well mixed upper node

311 so that its children were selected. Any branches below these nodes (or well-mixed upper node if the

312 within-species cell composition criteria was not met) were pruned and cells from these nodes were

313 merged into the same homologous groups, then the final integrated tree was generated.

314 As a final curation step, the homologous groups generated by the tree method were merged to be

315 consistent with within-species clusters. We defined consensus types by comparing the overlap of

316 within-species clusters between human and marmoset and human and mouse, as previously described

317 [2]. For each pair of human and mouse clusters and human and marmoset clusters, the overlap was

318 defined as the sum of the minimum proportion of nuclei in each cluster that overlapped within each leaf

319 of the pruned tree. This approach identified pairs of clusters that consistently co-clustered within one or

320 more leaves. Cluster overlaps varied from 0 to 1 and were visualized as a heatmap with human M1

321 clusters in rows and mouse or marmoset M1 clusters in columns. Cell type homologies were identified

322 as one-to-one, one-to-many, or many-to many so that they were consistent in all three species. For

323 example, the Vip_2 consensus type could be resolved into multiple homologous types between human

324 and marmoset but not human and mouse, and the coarser homology was retained. Consensus type

325 names were assigned based on the annotations of member clusters from human and mouse and

326 avoided specific marker gene names due to the variability of marker expression across species.

327

328 To quantify cell type alignment between pairs of species, we pruned the hierarchical tree described

329 above based on the stability and mixing of two species. We performed this analysis for human-

330 marmoset, human-mouse, and marmoset-mouse and compared the alignment resolution of each

331 subclass. The pruning criteria were tuned to fit the two-species comparison and to remove bias, and we

332 set the same criteria for all comparisons (entropy cutoff 3.0). Specifically, for each subclass and

333 pairwise species comparison, we calculated the number of leaves in the pruned tree. We repeated this

334 analysis on the 100 subsampled datasets and calculated the mean and standard deviation of the

335 number of leaves in the pruned trees. For each subclass, we tested for significant differences in the

336 average number of leaves across pairs of species using an ANOVA test followed by post-hoc Tukey

337 HSD tests.

338

Marker determination for cell type clusters by NS-Forest v2.1

340 NS-Forest v2.1 was used to determine the minimum set of marker genes whose combined expression

341 identified cells of a given type with maximum classification accuracy (T. Bakken et al. 2017; Aevermann

342 et al. 2018). (https://github.com/JCVenterInstitute/NSForest/releases). Briefly, for each cluster NS-

343 Forest produces a Random Forest (RF) model using a one vs all binary classification approach. The

344 top ranked genes from RF are then filtered by expression level to retain genes that are expressed in at

345 least 50% of the cells within the target cluster. The selected genes are then reranked by Binary Score

346 calculated by first finding median cluster expression values for a given gene and dividing by the target

347 median cluster expression value. Next, one minus this scaled value is calculated resulting in 0 for the

348 target cluster and 1 for clusters that have no expression, while negative scaled values are set to 0.

349 These values are then summed and normalized by dividing by the total number of clusters. In the ideal

350 case, where all off-target clusters have no expression, the binary score is 1. Finally, for the top 6 binary

351 genes optimal expression level cutoffs are determined and all permutations of genes are evaluated by

352 f-beta score, where the beta is weighted to favor precision. This f-beta score indicates the power of

353 discrimination for a cluster and a given set of marker genes. The gene combination giving the highest f-

354 beta score is selected as the optimal marker gene combination. Marker gene sets for human, mouse

355 and marmoset primary motor cortex are listed in Supplementary Tables 4, 5, and 6, respectively, and

356 were used to construct the semantic cell type definitions provided in Supplementary Table 1.

357

Calculating differentially expressed genes (DEGs)

359 To identify subclass level DEGs that are conserved and divergent across species, we used the

360 integrated Seurat objects from the species integration step. Seurat objects for each major cell class

361 were downsampled to have up to 200 cells per species cell type. Positive DEGs were then found using

362 Seurat's `FindAllMarkers` function using the ROC test with default parameters. We compared each

363 subclass within species to all remaining nuclei in that class and used the SCT normalized counts to test

364 for differential expression. For example, human *Sst* nuclei were compared to all other GABAergic

365 human neurons using the ROC test. Venn diagrams were generated using the eulerr package (6.0.0) to

366 visualize the relationship of DEGs across species for a given subclass. Heatmaps of DEGs for all

367 subclasses under a given class were generated by downsampling each subclass to 50 random nuclei

368 per species. SCT normalized counts were then scaled and visualized with Seurat's `DoHeatmap`

369 function.

370

371 To identify ChC DEGs that are enriched over BCs, we used the integrated Seurat objects from the

372 species integration step. The *Pvalb* subclass was subset and species cell types were then designated

373 as either ChCs or BCs. Positive DEGs were then found using Seurat's `FindAllMarkers` function

374 using the ROC test to compare ChCs and BCs for each species. Venn diagrams were generated using

375 the eulerr package (6.0.0) to visualize the relationship of ChC-enriched DEGs across species.

376 Heatmaps of conserved DEGs were generated by downsampling the dataset to have 100 randomly

377 selected BCs and ChCs from each species. SCT normalized counts were then scaled and visualized

378 with Seurat's `DoHeatmap` function.

379

380 We used the four species (human, macaque, marmoset, and mouse) integrated Glutamatergic Seurat

381 object from the species integration step for all L5 ET DEG figures. L5 ET and L5 IT subclasses were

382 downsampled to 200 randomly selected nuclei per species. A ROC test was then performed using

383 Seurat's `FindAllMarkers` function between the two subclasses for each species to identify L5 ET-

384 specific marker genes. We then used the UpSetR (1.4.0) package to visualize the intersections of the

385 marker genes across all four species as an upset plot. To determine genes that decrease in expression

386 across evolutionary distance in L5 ET neurons, we found the log-fold change between L5 ET and L5 IT

387 for each species across all genes. We then filtered the gene lists to only include genes that had a trend

388 of decreasing log-fold change (human > macaque > marmoset > mouse). Lastly, we excluded any gene

389   that did not have a log-fold change of 0.5 or greater in the human comparison. These 131 genes were

390   then used as input for GO analysis with the PANTHER Classification System [67] for the biological

391   process category, with organism set to Homo sapiens. All significant GO terms for this gene list were

392   associated with cell-cell adhesion and axon-guidance, and are colored blue in the line graph of their

393   expression enrichment.

394

395   ***Estimating differential isoform usage between human and mouse***

396   To assess changes of isoform usage between mouse and human, we used SSv4 data with full

397   transcript coverage and estimated isoform abundance in each cell subclasses. To mitigate low read

398   depth in each cell, we aggregated reads from all cells in each subclass. We estimated the relative

399   isoform usage in each subclass by calculating its genic proportion (P), defined as the ratio (R) of

400   isoform expression to the gene expression, where $R = (P_{human} - P_{mouse}) / (P_{human} + P_{mouse})$. For a common

401   set of transcripts for mouse and human, we used UCSC browser TransMapV5 set of human transcripts

402   (hg38 assembly, Gencode v31 annotations) mapped to the mouse genome (mm10 assembly)

403   http://hgdownload.soe.ucsc.edu/gbdb/mm10/transMap/V5/mm10.ensembl.transMapV5.bigPsl. We

404   considered only medium to highly expressed isoforms, which have abundance > 10 TPM (Transcripts

405   per Million) and P > 0.2 in either mouse or human and gene expression > 10 TPM in both mouse and

406   human.

407

408   Calculating isoform abundance in each cell subclass:

409   1) Aggregated reads from each subclass

410   2) Mapped reads to the mouse or human reference genome with STAR 2.7.3a using default

411   parameters

412   3) Transformed genomic coordinates into transcriptomic coordinates using STAR parameter: --

413   quantMode TranscriptomeSAM

414   4) Quantified isoform and gene expression using RSEM 1.3.3 parameters: --bam --seed 12345 --

415   paired-end --forward-prob 0.5 --single-cell-prior --calc-ci

416

417   Estimating statistical significance:

418   1) Calculated the standard deviation of isoform genic proportion (P_{human} and P_{mouse}) from the

419   RSEM's 95% confidence intervals of isoform expression

420   2) Calculated the P-value using normal distribution for the (P_{human} - P_{mouse}) and the summed

421   (mouse + human) variance

422   3) Bonferroni-adjusted P-values by multiplying nominal P-values by the number of medium to

423   highly expressed isoforms in each subclass

424

425   ***Species cluster dendrograms***

426   DEGs for a given species were identified using Seurat's `FindAllMarkers` function with a Wilcox test

427   and comparing each cluster to every other cluster under the same subclass, with logfc.threshold set to

428   0.7 and min.pct set to 0.5. The union of up to 100 genes per cluster with the highest avg_logFC were

429   used. The average $\log_2$ expression of the DEGs were then used as input for the `build_dend` function

430   from scratch.hicat to create the dendrograms. This was performed on both human and marmoset

431   datasets. For mouse dendrogram methods, see the companion paper [6].

432

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

434   Fresh-frozen human postmortem brain tissues were sectioned at 14-16 µm onto Superfrost Plus glass

435   slides (Fisher Scientific). Sections were dried for 20 minutes at -20°C and then vacuum sealed and

436   stored at -80°C until use. The RNAscope multiplex fluorescent v1 kit was used per the manufacturer's

437   instructions for fresh-frozen tissue sections (ACD Bio), except that fixation was performed for 60

438   minutes in 4% paraformaldehyde in 1X PBS at 4°C and protease treatment was shortened to 5minutes.

439   Primary antibodies were applied to tissues after completion of mFISH staining. Primary antibodies used

440   were mouse anti-GFAP (EMD Millipore, MAB360, 1:250 dilution) and mouse anti-Neurofilament H

441   (SMI-32, Biolegend, 801701). Secondary antibodies were goat anti-mouse IgG (H+L) Alexa Fluor

442   conjugates (594, 647). Sections were imaged using a 60X oil immersion lens on a Nikon TiE

443   fluorescence microscope equipped with NIS-Elements Advanced Research imaging software (version

444   4.20). For all RNAscope mFISH experiments, positive cells were called by manually counting RNA

445   spots for each gene. Cells were called positive for a gene if they contained ≥ 3 RNA spots for that gene.

446   Lipofuscin autofluorescence was distinguished from RNA spot signal based on the larger size of

447   lipofuscin granules and broad fluorescence spectrum of lipofuscin.

448

449   ***Gene family conservation***

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

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

452   Nomenclature Committee (HGNC) at the European Bioinformatics Institute

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

454   [68] (downloaded February 2020). HGNC annotations were propagated via the provided group hierarchy

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

456   were included in the analysis.

457

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

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

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

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

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

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

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

465

466 **_Replicability of clusters_**

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

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

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

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

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

472 as input for the MetaNeighborUS function, which was run using the fast_version and one_vs_best

473 parameters set to TRUE. Using the one_vs_best parameter means that only the two closest

474 neighboring clusters are tested for their similarity to the training cluster, with results reported as the

475 AUROC for the closest neighbor over the second closest. AUROCs are plotted in heatmaps in

476 Extended Data Figures 2 and 3. Data to reproduce these figures can be found in Supplementary Table

477 9, and scripts are on GitHub (http://github.com/gillislab/MetaNeighbor).

478

479 **_Single-cell methylome data (snmC-seq2): Sequencing and quantification_**

480 Library preparation and Illumina sequencing. Single nuclei were isolated from human and marmoset M1

481 tissue as described above for RNA-seq profiling and for mouse as detailed in [6]. Detailed methods for

482 bisulfite conversion and library preparation were previously described for snmC-seq2[5,41]. The snmC-

483 seq2 libraries generated from mouse brain tissues were sequenced using an Illumina Novaseq 6000

484 instrument with S4 flowcells and 150 bp paired-end mode.

485

486 Mapping and feature count pipeline. We implemented a versatile mapping pipeline (http://cemba-

487 data.rtfd.io) for all the single-cell methylome based technologies developed by our group [5,41,69]. The

488 main steps of this pipeline included: 1) demultiplexing FASTQ files into single-cell; 2) reads level QC; 3)

489 mapping; 4) BAM file processing and QC; and 5) final molecular profile generation. The details of the

490 five steps for snmC-seq2 were described previously [41]. We mapped all the reads from the three

491 corresponding species onto the human hg19 genome, the marmoset ASM275486v1 genome, and the

492   mouse mm10 genome. After mapping, we calculated the methyl-cytosine counts and total cytosine

493   counts for two sets of genome regions in each cell: the non-overlapping chromosome 100-kb bins of

494   each genome, the methylation levels of which were used for clustering analysis, and the gene body

495   regions, the methylation levels of which were used for cluster annotation and integration with RNA

496   expression data.

497

498   ***snmC-seq2: Quality control and preprocessing***

499   <u>Cell filtering.</u> We filtered the cells based on these main mapping metrics: 1) mCCC rate < 0.03. mCCC

500   rate reliably estimates the upper bound of bisulfite non-conversion rate [5]; 2) overall mCG rate > 0.5; 3)

501   overall mCH rate < 0.2; 4) total final reads > 500,000; and 5) bismark mapping rate > 0.5. Other metrics

502   such as genome coverage, PCR duplicates rate, and index ratio were also generated and evaluated

503   during filtering. However, after removing outliers with the main metrics 1-5, few additional outliers can

504   be found.

505

506   <u>Feature filtering.</u> 100kb genomic bin features were filtered by removing bins with mean total cytosine

507   base calls < 250 or > 3000. Regions overlap with the ENCODE blacklist [70] were also excluded from

508   further analysis.

509

510   <u>Computation and normalization of the methylation rate.</u> For CG and CH methylation, the computation of

511   methylation rate from the methyl-cytosine and total cytosine matrices contains two steps: 1) prior

512   estimation for the beta-binomial distribution and 2) posterior rate calculation and normalization per cell.

513   Step 1. For each cell we calculated the sample mean, m, and variance, v, of the raw mc rate *(mc / cov)*

514   for each sequence context (CG, CH). The shape parameters (α, β) of the beta distribution were then

515   estimated using the method of moments:

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

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

518   This approach used different priors for different methylation types for each cell and used weaker prior to

519   cells with more information (higher raw variance).

520

521   Step 2. We then calculated the posterior: $\widehat{mc} = \frac{\alpha+mc}{\alpha+\beta+cov}$., We normalized this rate by the cell's global

522   mean methylation, $m = \alpha/(\alpha + \beta)$. Thus, all the posterior $\widehat{mc}$ with 0 *cov* will be constant 1 after

523   normalization. The resulting normalized *mc* rate matrix contains no NA (not available) value, and

524   features with less *cov* tend to have a mean value close to 1.

525

526   Selection of highly variable features. Highly variable methylation features were selected based on a

527   modified approach using the scanpy package *scanpy.pp.highly_variable_genes* function [71]. In brief, the

528   *scanpy.pp.highly_variable_genes* function normalized the dispersion of a gene by scaling with the

529   mean and standard deviation of the dispersions for genes falling into a given bin for mean expression of

530   genes. In our modified approach, we reasoned that both the mean methylation level and the mean *cov*

531   of a feature (100kb bin or gene) could impact *mc* rate dispersion. We grouped features that fall into a

532   combined bin of mean and *cov*, and then normalized the dispersion within each *mean-cov* group. After

533   dispersion normalization, we selected the top 3000 features based on normalized dispersion for

534   clustering analysis.

535

536   Dimension reduction and combination of different mC types. For each selected feature, *mc* rates were

537   scaled to unit variance, and zero mean. PCA was then performed on the scaled *mc* rate matrix. The

538   number of significant PCs was selected by inspecting the variance ratio of each PC using the elbow

539   method. The CH and CG PCs were then concatenated together for further analysis in clustering and

540   manifold learning.

541

542   ***snmC-seq2: Data analysis***

543   <u>Consensus clustering on concatenated PCs.</u> We used a consensus clustering approach based on

544   multiple Leiden-clustering [72] over K-Nearest Neighbor (KNN) graph to account for the randomness of

545   the Leiden clustering algorithms. After selecting dominant PCs from PCA in both mCH and mCG

546   matrix, we concatenated the PCs together to construct a KNN graph using *scanpy.pp.neighbors* with

547   Euclidean distance. Given fixed resolution parameters, we repeated the Leiden clustering 300 times on

548   the KNN graph with different random starts and combined these cluster assignments as a new feature

549   matrix, where each single Leiden result is a feature. We then used the outlier-aware DBSCAN

550   algorithm from the scikit-learn package to perform consensus clustering over the Leiden feature matrix

551   using the hamming distance. Different epsilon parameters of DBSCAN are traversed to generate

552   consensus cluster versions with the number of clusters that range from minimum to the maximum

553   number of clusters observed in the multiple Leiden runs. Each version contained a few outliers that

554   usually fall into three categories: 1) cells located between two clusters that had gradient differences

555   instead of clear borders; 2) cells with a low number of reads that potentially lack information in essential

556   features to determine the specific cluster; and 3) cells with a high number of reads that were potential

557   doublets. The amount of type 1 and 2 outliers depends on the resolution parameter and is discussed in

558   the choice of the resolution parameter section. The type 3 outliers were very rare after cell filtering. The

559   supervised model evaluation then determined the final consensus cluster version.

560

561   <u>Supervised model evaluation on the clustering assignment.</u> For each consensus clustering version, we

562   performed a Recursive Feature Elimination with Cross-Validation (RFECV) [73] process from the scikit-

563   learn package to evaluate clustering reproducibility. We first removed the outliers from this process,

564   and then we held out 10% of the cells as the final testing dataset. For the remaining 90% of the cells,

565   we used tenfold cross-validation to train a multiclass prediction model using the input PCs as features

566   and *sklearn.metrics.balanced_accuracy_score* [74] as an evaluation score. The multiclass prediction

567   model is based on *BalancedRandomForestClassifier* from the imblearn package that accounts for

568   imbalanced classification problems[75]. After training, we used the 10% testing dataset to test the model

569   performance using the *balanced_accuracy_score* score. We kept the best model and corresponding

570   clustering assignments as the final clustering version. Finally, we used this prediction model to predict

571   outliers' cluster assignments, we rescued the outlier with prediction probability > 0.3, otherwise labeling

572   them as outliers.

573

574   Choice of resolution parameter. Choosing the resolution parameter of the Leiden algorithm is critical for

575   determining the final number of clusters. We selected the resolution parameter by three criteria: 1. The

576   portion of outliers < 0.05 in the final consensus clustering version. 2. The ultimate prediction model

577   accuracy > 0.95. 3. The average cell per cluster ≥ 30, which controls the cluster size to reach the

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

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

580   meeting the criteria using a grid search.

581

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

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

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

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

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

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

588   the same way.

589

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

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

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

593   Cv3 analysis using scanpy [71]. We then used Scanorama [76] to integrate the two modalities.

594

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

596   (https://github.com/yupenghe/methylpy) as previously described [77]. Briefly, we first called CG

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

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

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

600   covered in sequencing.

601

602   TFBS motif enrichment analysis. For each cell subclass (cluster), we performed TFBS motif enrichment

603   analysis for its hypo-methylated DMRs against the hypo-methylated DMRs from other cell subclasses

604   (clusters) using software AME [78]. DMRs and 250bp regions around were used in the analysis.

605

606   **SNARE-Seq2: Sample preparation**

607   Human and marmoset primary motor cortex nuclei were isolated for SNARE-seq2 according to the

608   following protocol: https://www.protocols.io/view/nuclei-isolation-for-snare-seq2-8tvhwn6 [7,79].

609   Fluorescence-activated nuclei sorting (FANS) was then performed on a FACSAria Fusion (BD

610   Biosciences, Franklin Lakes, NJ) gating out debris from FSC and SSC plots and selecting DAPI$^+$

611   singlets (Extended Data Fig. 5a). Samples were kept on ice until sorting was complete and were used

612   immediately for SNARE-seq2.

613

614   **SNARE-Seq2: Library preparation and sequencing**

615   A detailed step-by-step protocol for SNARE-Seq2 has been outlined in a companion paper [38]. The

616   resulting AC libraries were sequenced on MiSeq (Illumina) (R1: 75 cycles for the 1st end of AC DNA

617   read, R2: 94 cycles for cell barcodes and UMI read, R3: 8 cycles for i5 read, R4: 75 cycles for the 2nd

618   end of AC DNA read) for library validation, then on NovaSeq6000 (Illumina) using 300 cycles reagent

619   kit for data generation. RNA libraries were combined at equimolar ratio and sequenced on MiSeq

620   (Illumina) (Read 1: 70 cycles for the cDNA read, Index 1: 6 cycles for i7 read, Read 2: 94 cycles for cell

621 barcodes and UMI read) for library validation, then on NovaSeq6000 (Illumina) using 200 cycles

622 reagent kit for data generation.

623

624 ***SNARE-Seq2: Data processing***

625 A detailed step-by-step SNARE-seq2 data processing pipeline has been provided in a companion

626 paper [38]. For RNA data, this has involved the use of dropEst to extract cell barcodes and STAR

627 (v2.5.2b) to align tagged reads to the genome (GRCh38 version 3.0.0 for human; GCF 000004665.1

628 Callithrix jacchus-3.2, marmoset). For AC data, this involved snaptools for alignment to the genome

629 (cellranger-atac-GRCh38-1.1.0 for human, GCF 000004665.1 Callithrix jacchus-3.2, marmoset) and to

630 generate snap objects for processing using the R package snapATAC.

631

632 ***SNARE-Seq2: Data analysis***

633 RNA quality filtering. For SNARE-Seq2 data, quality filtering of cell barcodes and clustering analysis

634 were first performed on transcriptomic (RNA) counts and used to inform on subsequent accessible

635 chromatin quality filtering and analysis. Each cell barcode was tagged by an associated library batch ID

636 (for example MOP1, MOP2… etc.), RNA read counts associated with dT and n6 adaptor primers were

637 merged, libraries were combined for each sample within each experiment and empty barcodes

638 removed using the emptyDrops() function of DropletUtils [80], mitochondrial transcripts were removed,

639 doublets were identified using the DoubletDetection software [81] and removed. All samples were

640 combined across experiments within species and cell barcodes having greater than 200 and less than

641 7500 genes detected were kept for downstream analyses. To further remove low quality datasets, a

642 gene UMI ratio filter (gene.vs.molecule.cell.filter) was applied using Pagoda2 (https://github.com/hms-

643 dbmi/pagoda2).

644

645 RNA data clustering. For human SNARE-seq2 RNA data, clustering analysis was first performed using

646 Pagoda2 where counts were normalized to the total number per nucleus and batch variations were

647   corrected by scaling expression of each gene to the dataset-wide average. After variance

648   normalization, the top 6000 over-dispersed genes were used for principal component analysis.

649   Clustering was performed using an approximate k-nearest neighbor graph (k values between 50 – 500)

650   based on the top 75 principal components and cluster identities were determined using the infomap

651   community detection algorithm. Major cell types were identified using a common set of broad cell type

652   marker genes: *GAD1/GAD2* (GABAergic neurons), *SLC17A7/SATB2* (glutamatergic neurons),

653   *PDGFRA* (oligodendrocyte progenitor cells), *AQP4* (astrocytes), *PLP1/MOBP* (oligodendrocytes),

654   *MRC1* (perivascular macrophages), *PTPRC* (T cells), *PDGFRB* (vascular smooth muscle cells), *FLT1*

655   (vascular endothelial cells), *DCN* (vascular fibroblasts), *APBB1IP* (microglia) (Extended Data Fig. 5c).

656   Low quality clusters that showed very low gene/UMI detection rates, low marker gene detection and/or

657   mixed cell type marker profiles were removed. Oligodendrocytes were over-represented (54,080 total),

658   possibly reflecting a deeper subcortical sampling then intended, therefore, to ensure a more balanced

659   distribution of cell types, we capped the number of oligodendrocytes at 5000 total and repeated the

660   PAGODA2 clustering as above. To achieve optimal clustering of the different cell types, different k

661   values were used to identify cluster subpopulations for different cell types (L2/3 glutamatergic neurons,

662   k = 500; all other glutamatergic neurons, astrocytes, oligodendrocytes, OPCs, k = 100; GABAergic

663   neurons, vascular cells, microglia/perivascular macrophages, k = 50). To assess the appropriateness of

664   the chosen k values, clusters were compared against SMARTer clustering of data generated on human

665   M1 through correlation of cluster-averaged scaled gene expression values using the corrplot package

666   (https://github.com/taiyun/corrplot) (Extended Data Fig. 5d). For cluster visualization, uniform manifold

667   approximation and projection (UMAP) dimensional reduction was performed in Seurat (version 3.1.0)

668   using the top 75 principal components identified using Pagoda2. For marmoset, clustering was initially

669   performed using Seurat, where the top 2000 variable features were selected from the mean variance

670   plot using the 'vst' method and used for principal component analysis. UMAP embeddings were

671   generated using the top 75 principal components. To harmonize cellular populations across platforms

672   and modalities, snRNA-seq within-species cluster identities were then predicted from both human and

673   marmoset data. We used an iterative nearest centroid classifier algorithm (Methods, 'Mapping of

674   samples to reference taxonomies') to generate probability scores for each SNARE-seq2 nuclei mapping

675   to their respective species' snRNA-seq reference cluster (Cv3 for marmoset and SMART-Seqv4 for

676   human). Comparing the predicted RNA cluster assignment of each nuclei with Pagoda2-identified

677   clusters showed highly consistent cluster membership using Jaccard similarity index (Extended Data

678   Fig. 5e), confirming the robustness of these cell identities discovered using different analysis platforms.

679

680   <u>AC quality filtering and peak calling</u>. Initial analysis of corresponding SNARE-Seq2 chromatin

681   accessibility data was performed using SnapATAC software (version 2)

682   (https://github.com/r3fang/SnapATAC) (https://doi.org/10.1101/615179). Snap objects were generated

683   by combining individual snap files across libraries within each species. Cell barcodes were included for

684   downstream analyses only if cell barcodes passed RNA quality filtering (above) and showed greater

685   than 1000 read fragments and 500 UMI. Read fragments were then binned to 5000 bp windows of the

686   genome and only cell barcodes showing the fraction of binned reads within promoters greater than 10%

687   (15% for marmoset) and less than 80% were kept for downstream analysis. Peak regions were called

688   independently for RNA cluster, subclass and class groupings using MACS2 software

689   (https://github.com/taoliu/MACS) using the following options "--nomodel --shift 100 --ext 200 --qval 5e-2

690   -B --SPMR". Peak regions were combined across peak callings and used to generate a single peak

691   count matrix (cell barcodes by chromosomal peak locations) using the "createPmat" function of

692   SnapATAC.

693

694   <u>AC data clustering</u>. The peak count matrices were filtered to keep only locations from chromosomes 1-

695   22, x or y, and processed using Seurat (version 3.1.0) and Signac (version 0.1.4) software [24]

696   (https://satijalab.org). All peaks having at least 100 counts (20 for marmoset) across cells were used for

697   dimensionality reduction using latent semantic indexing ("RunLSI" function) and visualized by UMAP

698   using the first 50 dimensions (40 for marmoset).

699

700   <u>Calculating gene activity scores</u>. For a gene activity matrix from accessibility data, cis-co-accessible

701   sites and gene activity scores were calculated using Cicero software (v1.2.0) [39] (https://cole-trapnell-

702   lab.github.io/cicero-release/). The binary peak matrix was used as input with expression family variable

703   set to "binomialff" to make the aggregated input Cicero CDS object using the AC peak-derived UMAP

704   coordinates and setting 50 cells to aggregate per bin. Co-accessible sites were then identified using the

705   "run_cicero" function using default settings and modules of cis-co-accessible sites identified using the

706   "generate_ccans" function. Co-accessible sites were annotated to a gene if they fell within a region

707   spanning 10,000 bp upstream and downstream of the gene's transcription start site (TSS). The Cicero

708   gene activity matrix was then calculated using the "build_gene_activity_matrix" function using a co-

709   accessibility cutoff of 0.25 and added to a separate assay of the Seurat object.

710

711   <u>Integrating RNA/AC data modalities</u>. For reconciliation of differing resolutions achievable from RNA and

712   accessible chromatin (Extended Data Fig. 5f-k), integrative analysis was performed using Seurat.

713   Transfer anchors were identified between the activity and RNA matrices using the

714   "FindTransferAnchors" function. For human, transfer anchors were generated using an intersected list

715   of variable genes identified from Pagoda2 analysis of RNA clusters (top 2000 genes) and marker genes

716   for clusters identified from SSv4 data (2492 genes having β-scores > 0.4), and canonical correlation

717   analysis (CCA) for dimension reduction. For marmoset, transfer anchors were generated using an

718   intersected list of variable genes identified using Seurat (top 2000 genes) and DEGs identified between

719   marmoset consensus clusters (Cv3 snRNA-seq data, $P < 0.05$, top 100 markers per cluster). Imputed

720   RNA expression values were then calculated using the "TransferData" function from the Cicero gene

721   activity matrix using normalized RNA expression values for reference and LSI for dimension reduction.

722   RNA and imputed expression data were merged, a UMAP co-embedding and shared nearest neighbor

723   (SNN) graph generated using the top 50 principal components (40 for marmoset) and clusters identified

724   ("FindClusters") using a resolution of 4. Resulting integrated clusters were compared against

725   consensus RNA clusters by calculating jaccard similarity scores using scratch.hicat software. Cell

726   populations identified as T-cells from Pagoda2 analysis (human only) and those representing low

727   quality integrated clusters, showing a mixture of disparate cell types, were removed from these

728   analyses. RNA clusters were assigned to co-embedded clusters based on the highest jaccard similarity

729   score and frequency and then merged to generate the best matched co-embedded clusters, taking in

730   account cell type and subclass to ensure more accurate merging of ambiguous populations. This

731   enabled AC-level clusters that directly matched the RNA-defined populations (Extended Data Fig. 5k).

732   For consensus cluster and subclass level predictions (Extended Data Fig. 5g) the Seurat

733   "TransferData" function was used to transfer RNA consensus cluster or subclass labels to AC data

734   using the pre-computed transfer anchors and LSI dimensionality reduction.

735

736   Final AC peak and gene activity matrices. A final combined list of peak regions was then generated

737   using MACS2 as detailed above for all cell populations corresponding to RNA consensus (> 100

738   nuclei), accessibility-level, subclass (> 50 nuclei) and class level barcode groupings. The corresponding

739   peak by cell barcode matrix generated by SnapATAC was used to establish a Seurat object as outlined

740   above, with peak counts, Cicero gene activity scores and RNA expression values for matched cell

741   barcodes contained within different assay slots.

742

743   Transcription factor motif analyses. Jaspar motifs (JASPAR2020, all vertebrate) were used to generate

744   a motif matrix and motif object that was added to the Seurat object using Signac ("CreateMotifMatrix",

745   "CreateMotifObject", "AddMotifObject") and GC content, region lengths and dinucleotide base

746   frequencies calculated using the "RegionStats" function. Motif enrichments within specific chromosomal

747   sites were calculated using the FindMotifs function. For motif activity scores, chromVAR

748   (https://greenleaflab.github.io/chromVAR) was performed according to default parameters (marmoset)

749   or using Signac "RunChromVAR" function on the peak count matrix (human). The chromVAR deviation

750   score matrix was then added to a separate assay slot of the Seurat object and differential activity of

751   TFBS between different populations were assessed using the "Find[All]Markers" function through

752   logistic regression and using the number of peak counts as a latent variable.

753

754   Differentially accessible regions (DARs) between cell populations (Fig. 4b) were identified using the

755   "find_all_diff" function (https://github.com/yanwu2014/chromfunks) and p-values calculated using a

756   hypergeometric test. For visualization, the top DARs (q value < 0.001 and log-fold change > 1) were

757   selected and the top distinct sites visualized by dot plot in Seurat. For motif enrichment analyses, peak

758   counts associated with the clusters selected for comparison (all subclasses, all AC-level clusters,

759   *PVALB*-positive for ChC analyses) were used to identify cis-co-accessible site networks or CCANs

760   using cicero as indicated above. Peak locations were annotated to the nearest gene (10,000 bases

761   upstream and downstream of the TSS) and only genes identified from SNARE-seq2 RNA data as being

762   differentially expressed (Seurat, Wilcoxon Rank Sum test) within the clusters of interest (adjusted P <

763   0.05, average log-fold change > 0.5) were used. Genes having more than one co-accessible site were

764   assessed for motif enrichments within all overlapping sites using the "FindMotifs" function in Signac

765   (using peaks for all cell barcodes for subclass and AC-level, or only peaks for ChC or L5 ET cells).

766   Motifs were then trimmed to only those showing significant differential activity (chromVAR) between the

767   clusters of interest (P < 0.05) as assessed using the "FindMarkers" function on the chromVAR assay

768   slot using Seurat and using the number of total peaks as a latent variable. The top distinct genes

769   (subclass, AC-level) or all genes (ChC, Betz) used for motif enrichment analysis were visualized for

770   scaled average RNA expression levels and scaled average cicero gene activities using the ggHeat

771   plotting function (SWNE package, https://github.com/yanwu2014/swne). Top chromVAR TFBS activities

772   were also visualized using ggHeat.

773

774   Correlation plots. For correlation of RNA expression and associated AC activities for consensus and

775   AC-level clusters (Extended Data Fig. 6a-b), average scaled expression values were generated and

776   pairwise correlations performed for marker genes identified from an intersected list of variable genes

777   identified from Pagoda2 analysis of RNA clusters (top 2000 genes) and marker genes for clusters

778   identified from SSv4 data (2492 genes having β-scores > 0.4). For correlation across species,

779   expression values for genes used to integrate human and marmoset GABAergic and glutamatergic

780   clusters (Cv3 scRNA-seq data), or chromVAR TFBS activity scores for all Jaspar motifs were averaged

781   by subclass, scaled (trimming values to a minimum of 0 and a maximum of 4) for each species

782   separately, then correlated and visualized using corrplot.

783

784   Plots and figures. All UMAP, feature, dot, and violin plots were generated using Seurat. Connection

785   plots were generated using cicero and peak track gradient heatmaps were generated using Gviz [82] from

786   bedGraph files generated during peak calling using SnapATAC. Correlation plots were generated using

787   the corrplot package.

788

789   **_Mouse chandelier cell ATAC-Seq: Data acquisition and analysis_**

790   Chandelier cells are rare in mouse cortex and were enriched by isolating individual neurons from

791   transgenically-labelled mouse primary visual cortex (VISp). Many of the transgenic mouse lines have

792   previously been characterized by single-cell RNA-seq [1]. Single-cell suspensions of cortical neurons

793   were generated as described previously [1] and subjected to tagmentation (ATAC-seq) [83,84]. Mixed

794   libraries, containing 60 to 96 samples were sequenced on an Illumina MiSeq. In total, 4,275 single-cells

795   were collected from 36 driver-reporter combinations in 67 mice. After sequencing, raw FASTQ files

796   were aligned to the GRCm38 (mm10) mouse genome using Bowtie v1.1.0 as previously described [9].

797   Following alignment, duplicate reads were removed using samtools rmdup, which yielded only single

798   copies of uniquely mapped paired reads in BAM format. Quality control filtering was applied to select

799   samples with >10,000 uniquely mapped paired-end fragments, >10% of which were longer than 250

800   base pairs and with >25% of their fragments overlapping high-depth cortical DNase-seq peaks from

801   ENCODE [85]. The resulting dataset contained a total of 2,799 samples.

802

803   To increase the cell-type resolution of chromatin accessibility profiles beyond that provided by driver

804   lines, a feature-free method for computation of pairwise distances (Jaccard) was used. Using Jaccard

805   distances, principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE)

806   were performed, followed by Phenograph clustering [86]. This clustering method grouped cells from

807   class-specific driver lines together, but also segregated them into multiple clusters. Phenograph-defined

808   neighborhoods were assigned to cell subclasses and clusters by comparison of accessibility near

809   transcription start site (TSS ± 20 kb) to median expression values of scRNA-seq clusters at the cell type

810   and at the subclass level from mouse primary visual cortex [87]. From this analysis, a total of 226

811   samples were assigned to *Pvalb* and 124 samples to *Pvalb Vipr2* (ChC) clusters. The sequence data

812   for these samples were grouped together and further processed through the Snap-ATAC pipeline.

813

814   Mouse scATAC-seq peak counts for *Pvalb* and ChC were used to generate a Seurat object as outlined

815   for human and marmoset SNARE-Seq2 AC data. Cicero cis-co-accessible sites were identified, gene

816   activity scores calculated, and motif enrichment analyses performed as outlined above. Genes used for

817   motif enrichment were ChC markers identified from differential expression analysis between *PVALB-*

818   *positive* clusters in mouse Cv3 scRNA-seq data (adjusted P < 0.05).

819

820   ***Patch-seq neuronal physiology, morphology, and transcriptomics***

821   <u>Subjects</u>. The human neurosurgical specimen was obtained from a 61-year old female patient that

822   underwent deep tumor resection (glioblastoma) from the frontal lobe at a local hospital (Harborview

823   Medical Center). The patient provided informed consent and experimental procedures were approved

824   by the hospital institute review board before commencing the study. Post-hoc analysis revealed that the

825   neocortical tissue obtained from this patient was from a premotor region near the confluence of the

826   superior frontal gyrus and the precentral gyrus (Fig. 7g). All procedures involving macaques and mice

827   were approved by the Institutional Animal Care and Use Committee at either the University of

828   Washington or the Allen Institute for Brain Science. Macaque M1 tissue was obtained from male (n=4)

829   and female (n=5) animals (mean age= 10 ± 2.21 years) designated for euthanasia via the Washington

830   National Primate Research Center's Tissue Distribution Program. Mouse M1 tissue was obtained from

831   4-12 week old male and female mice from the following transgenic lines: *Thy1h*-eyfp (B6.Cg-Tg(*Thy1*-

832   YFP)-HJrs/J: JAX Stock No. 003782), *Etv1*-egfp Tg(*Etv1*-EGFP)BZ192Gsat/Mmucd (etv1) mice

833   maintained with the outbred Charles River Swiss Webster background (Crl:CFW(SW) CR Stock No.

834   024), and C57BL/6-Tg(*Pvalb*-tdTomato)15Gfng/J: JAX stock No. 027395.

835

836   Brain slice preparation. Brain slice preparation was similar for *Pvalb*-TdTomato mice, macaque and

837   human samples. Upon resection, human neurosurgical tissue was immediately placed in a chilled and

838   oxygenated solution formulated to prevent excitotoxicity and preserve neural function [88]. This artificial

839   cerebral spinal fluid (NMDG aCSF) consisted of (in mM): 92 with N-methyl-D-glucamine (NMDG), 2.5

840   KCl, 1.25 NaH_2PO_4, 30 $NaHCO_3$, 20 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 25

841   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

842   the NMDG aCSF was titrated to pH 7.3–7.4 with concentrated hydrochloric acid and the osmolality was

843   300-305 mOsmoles/Kg. The solution was pre-chilled to 2-4°C and thoroughly bubbled with carbogen

844   (95% O_2/5% CO_2) prior to collection. Macaques were anesthetized with sevoflurane gas during which

845   the entire cerebrum was extracted and placed in the same protective solution described above. After

846   extraction, macaques were euthanized with sodium-pentobarbital. We dissected the trunk/limb area of

847   the primary motor cortex for brain slice preparation. *Pvalb*-TdTomato mice were deeply anesthetized by

848   intraperitoneal administration of Advertin (20mg/kg IP) and were perfused through the heart with NMDG

849   aCSF (bubbled with carbogen).

850

851   Brains were sliced at 300-micron thickness on a vibratome using the NMDG protective recovery

852   method and a zirconium ceramic blade [61,88]. Mouse brains were sectioned coronally, and human and

853   macaque brains were sectioned such that the angle of slicing was perpendicular to the pial surface.

854   After sections were obtained, slices were transferred to a warmed (32-34° C) initial recovery chamber

855   filled with NMDG aCSF under constant carbogenation. After 12 minutes, slices were transferred to a

856   chamber containing an aCSF solution consisting of (in mM): 92 NaCl, 2.5 KCl, 1.25 NaH_2PO_4, 30

857    $NaHCO_3$, 20 HEPES, 25 glucose, 2 thiourea, 5 Na-ascorbate, 3 Na-pyruvate, 2 $CaCl_2 \cdot 4H_2O$ and 2

858    $MgSO_4 \cdot 7H_2O$ continuously bubbled with 95% O_2/5% CO_2. Slices were held in this chamber for use in

859   acute recordings or transferred to a 6-well plate for long-term culture and viral transduction. Cultured

860   slices were placed on membrane inserts and wells were filled with culture medium consisting of 8.4 g/L

861   MEM Eagle medium, 20% heat-inactivated horse serum, 30 mM HEPES, 13 mM D-glucose, 15 mM

862    $NaHCO_3$, 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

863   insulin (Ting et al 2018). The slice culture medium was carefully adjusted to pH 7.2-7.3, osmolality of

864   300-310 mOsmoles/Kg by addition of pure H_2O, sterile-filtered and stored at 4°C for up to two weeks.

865   Culture plates were placed in a humidified 5% CO_2 incubator at 35°C and the slice culture medium was

866   replaced every 2-3 days until end point analysis. 1-3 hours after brain slices were plated on cell culture

867   inserts, brain slices were infected by direct application of concentrated AAV viral particles over the slice

868   surface (Ting et al 2018).

869

870   Thy1 and Etv1 mice were deeply anesthetized by IP administration of ketamine (130 mg/kg) and

871   xylazine (8.8 mg/kg) mix and were perfused through the heart with chilled (2-4°C) sodium-free aCSF

872   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

873    $CaCl_2$, 1.3 Na-ascorbate, 3 Na-pyruvate bubbled with carbogen (95% O_2/5% CO_2). Near coronal slices

874   300 microns thick were generated using a Leica vibratome (VT1200) in the same sodium-free aCSF

875   and were transferred to warmed (35°C) holding solution (in mM): 125 NaCl, 2.5 KCl, 1.25 NaH_2PO_4, 26

876    $NaHCO_3$, 2 $CaCl_2$, 2 $MgCl_2$, 17 dextrose, and 1.3 sodium pyruvate bubbled with carbogen (95% O_2/5%

877   CO_2). After 30 minutes of recovery, the chamber holding slices was allowed to cool to room

878   temperature.

879

880 <u>Patch clamp electrophysiology</u>. Macaque, human and *Pvalb*-TdTomato mouse brain slices were placed

881 in a submerged, heated (32-34°C) recording chamber that was continually perfused (3-4 mL/min) with

882 aCSF under constant carbogenation and containing (in mM) 1): 119 NaCl, 2.5 KCl, 1.25 NaH_2PO_4, 24

883 $NaHCO_3$, 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

884 Olympus BX51WI microscope and infrared differential interference contrast (IR-DIC) optics and a 40x

885 water immersion objective. The infragranular layers of macaque primary motor cortex and human

886 premotor cortex are heavily myelinated, which makes visualization of neurons under IR-DIC virtually

887 impossible. To overcome this challenge, we labeled neurons using various viral constructs in

888 organotypic slice cultures (Extended Data Fig. 10g).

889 Patch pipettes (2-6 MΩ) were filled with an internal solution containing (in mM): 110.0 K-gluconate,

890 10.0 HEPES, 0.2 EGTA, 4 KCl, 0.3 Na2-GTP, 10 phosphocreatine disodium salt hydrate, 1 Mg-ATP, 20

891 µg/ml glycogen, 0.5U/µL RNAse inhibitor (Takara, 2313A) and 0.5% biocytin (Sigma B4261), pH 7.3.

892 Fluorescently labeled neurons from *Thy1* or *Etv1* mice were visualized through a 40x objective using

893 either Dodt contrast with a CCD camera (Hamamatsu) and/or a 2-photon imaging/ uncaging system

894 from Prairie (Bruker) Technologies. Recordings were made in aCSF: (in mM): 125 NaCl, 3.0 KCl, 1.25

895 NaH_2PO_4, 26 $NaHCO_3$, 2 $CaCl_2$, 1 MgCl2, 17 dextrose, and 1.3 sodium pyruvate bubbled with

896 carbogen (95% O_2/5% CO_2) at 32-35°, with synaptic inhibition blocked using 100 µM picrotoxin.

897 Sylgard-coated patch pipettes (3-6 MΩ) were filled with an internal solution containing (in mM): 135 K-

898 gluconate, 12 KCl, 11 HEPES, 4 MgATP, 0.3 NaGTP, 7 K_2-phosphocreatine, 4 Na_2-phophocreatine (pH

899 7.42 with KOH) with neurobiotin (0.1-0.2%), Alexa 594 (40 µM) and Oregon Green BAPTA 6F (100

900 µM).

901

902 Whole cell somatic recordings were acquired using either a Multiclamp 700B amplifier, or an AxoClamp

903 2B amplifier (Molecular Devices) and were digitized using an ITC-18 (HEKA). Data acquisition software

904 was either MIES (https://github.com/AllenInstitute/MIES/) or custom software written in Igor Pro.

905 Electrical signals were digitized at 20-50 kHz and filtered at 2-10 kHz. Upon attaining whole-cell current

906   clamp mode, the pipette capacitance was compensated and the bridge was balanced. Access

907   resistance was monitored throughout the recording and was 8-25 MΩ.

908

909   Data analysis. Data were analyzed using custom analysis software written in Igor Pro. All

910   measurements were made at resting membrane potential. Input resistance (R_N) was measured from a

911   series of 1 s hyperpolarizing steps from -150 pA to +50 pA in +20 pA increments. For neurons with low

912   input resistance (e.g. the Betz cells) this current injection series was scaled by upwards of 4x. Input

913   resistance (R_N) was calculated from the linear portion of the current−steady state voltage relationship

914   generated in response to these current injections. Resonance (f_R) was determined from the voltage

915   response to a constant amplitude sinusoidal current injection (Chirp stimulus). The chirp stimulus

916   increased in frequency either linearly from 1-20 Hz over 20 s or logarithmically from 0.2-40 Hz over 20s.

917   The amplitude of the Chirp was adjusted in each cell to produce a peak-to-peak voltage deflection of

918   ~10 mV. The impedance amplitude profile (ZAP) was constructed from the ratio of the fast Fourier

919   transform of the voltage response to the fast Fourier transform of the current injection. ZAPs were

920   produced by averaging at least three presentations of the Chirp and were smoothed using a running

921   median smoothing function. The frequency corresponding to the peak impedance (Z_{max}) was defined as

922   the resonant frequency. Spike input/output curves were constructed in response to 1 s step current

923   injections (50 pA-500 pA in 50 pA steps). For a subset of experiments, this current injection series was

924   extended to 3A in 600 pA steps to probe the full dynamic range of low R_N neurons. Spike frequency

925   acceleration analysis was performed for current injections producing ~10 spikes during the 1 s step.

926   Acceleration ratio was defined as the ratio of the second to the last interspike interval. To examine the

927   dynamics of spike timing over longer periods, we also measured spiking in response to 10 s step

928   current injections in which the amplitude of the current was adjusted to produce ~5 spikes in the first

929   second. Action potential properties were measured for currents near rheobase. Action potential

930   threshold was defined as the voltage at which the first derivative of the voltage response exceeded 20

931   V/s. AP width was measured at half the amplitude between threshold and the peak voltage. Fast AHP

932   was defined relative to threshold. We clustered mouse, macaque and human pyramidal neurons into

933   two broad groups based on their R_N and f_R using Ward's algorithm.

934

935   Biocytin histology. A horseradish peroxidase (HRP) enzyme reaction using diaminobenzidine (DAB) as

936   the chromogen was used to visualize the filled cells after electrophysiological recording, and 4,6-

937   diamidino-2-phenylindole (DAPI) stain was used identify cortical layers as described previously [89].

938

939   Microscopy. Mounted sections were imaged as described previously [89]. Briefly, operators captured

940   images on an upright AxioImager Z2 microscope (Zeiss, Germany) equipped with an Axiocam 506

941   monochrome camera and 0.63x optivar. Two-dimensional tiled overview images were captured with a

942   20X objective lens (Zeiss Plan-NEOFLUAR 20X/0.5) in brightfield transmission and fluorescence

943   channels. Tiled image stacks of individual cells were acquired at higher resolution in the transmission

944   channel only for the purpose of automated and manual reconstruction. Light was transmitted using an

945   oil-immersion condenser (1.4 NA). High-resolution stacks were captured with a 63X objective lens

946   (Zeiss Plan-Apochromat 63x/1.4 Oil or Zeiss LD LCI Plan-Apochromat 63x/1.2 Imm Corr) at an interval

947   of 0.28 µm (1.4 NA objective; mouse specimens) or 0.44 µm (1.2 NA objective; human and non-human

948   primate specimens) along the Z axis. Tiled images were stitched in ZEN software and exported as

949   single-plane TIFF files.

950

951   Morphological reconstruction. Reconstructions of the dendrites and the full axon were generated based

952   on a 3D image stack that was run through a Vaa3D-based image processing and reconstruction

953   pipeline as described previously [89].

954

955   Viral vector production and transduction. Recombinant AAV vectors were produced by triple-

956   transfection of ITR-containing enhancer plasmids along with AAV helper and rep/cap plasmids using

957   the AAV293 cell line, followed by harvest, purification and concentration of the viral particles. The

958   AAV293 packaging cell line and plasmid supplying the helper function are available from a commercial

959   source (Cell Biolabs). The PHP.eB capsid variant was generated by Dr. Viviana Gradinaru at the

960   California Institute of Technology [90] and the DNA plasmid for AAV packaging is available from Addgene

961   (plasmid#103005). Quality control of the packaged AAV was determined by viral titering to determine

962   an adequate concentration was achieved ($>5E^{12}$ viral genomes per mL), and by sequencing the AAV

963   genome to confirm the identity of the viral vector that was packaged. Human and NHP L5 ET neurons

964   including Betz cells were targeted in cultured slices by transducing the slices with viral vectors that

965   either generically label neurons (AAV-hSyn1-tdTomato), or that enrich for L5 ET neurons by expressing

966   reporter transgene under the control of the msCRE4 enhancer [87].

967

968   Processing of Patch-seq samples. For a subset of experiments, the nucleus was extracted at the end of

969   the recording and processed for RNA-sequencing. Prior to data collection for these experiments, all

970   surfaces were thoroughly cleaned with RNAse Zap. The contents of the pipette were expelled into a

971   PCR tube containing lysis buffer (Takara, 634894). cDNA libraries were produced using the SMART-

972   Seq v4 Ultra Low Input RNA Kit for Sequencing according to the manufacturer's instructions. We

973   performed reverse transcription and cDNA amplification for X PCR cycles. Sample proceeded through

974   Nextera NT DNA Library Preparation using Nextera XT Index Kit V2 Set A(FC-131-2001).

975

976   Mapping of samples to reference taxonomies. To identify which cell type a given patch-seq nuclei

977   mapped to, we used our previously described nearest centroid classifier [1]. Briefly, a centroid classifier

978   was constructed for Glutamatergic reference data (human SSv4 or macaque Cv3) using marker genes

979   for each cluster. Patch-seq nuclei were then mapped to the appropriate species reference 100 times,

980   using 80% of randomly sampled marker genes during each iteration. Probabilities for each nuclei

981   mapping to each cluster were computed over the 100 iterations, resulting in a confidence score ranging

982   from 0 to 100. We identified four human patch-seq nuclei that mapped with > 85% confidence and four

983   macaque nuclei that mapped with > 93% confidence to a cluster in the L5 ET subclass.

984

985   ## Data availability

986   Raw sequence data are available for download from the Neuroscience Multi-omics Archive

987   (https://nemoarchive.org/) and the Brain Cell Data Center (https://biccn.org/data). Visualization and

988   analysis tools are available at NeMO Analytics (Individual species:

989   https://nemoanalytics.org//index.html?layout_id=ac9863bf; Integrated species:

990   https://nemoanalytics.org//index.html?layout_id=34603c2b) and Cytosplore Viewer

991   (https://viewer.cytosplore.org/). These tools allow users to compare cross-species datasets and

992   consensus clusters via genome and cell browsers and calculate differential expression within and

993   among species. A semantic representation of the cell types defined through these studies is available in

994   the provisional Cell Ontology (https://bioportal.bioontology.org/ontologies/PCL; Supplementary Table

995   1).

996

997   ## Code availability

998   Code to reproduce figures will be available for download from

999   https://github.com/AllenInstitute/BICCN_M1_Evo.

000

001   ## Acknowledgements

002   We thank the Tissue Procurement, Tissue Processing and Facilities teams at the Allen Institute for

003   Brain Science for assistance with the transport and processing of postmortem and neurosurgical brain

004   specimens; the Technology team at the Allen Institute for assistance with data management; M.

005   Vawter, J. Davis and the San Diego Medical Examiner's Office for assistance with postmortem tissue

006   donations. We thank Ximena Opitz-Araya and Allen Institute for Brain Science Viral Technology team

031 ## Author contributions

032 RNA data generation: A.M.Y., A.Re., A.T., B.B.L., B.T., C.D.K., C.R., C.R.P., C.S.L., D.B., D.D., D.M.,

033 E.S.L., E.Z.M., F.M.K., G.F., H.T., H.Z., J.C., J.Go., J.S., K.C., K.L., K.Si., K.Sm., K.Z., M.G., M.K.,

034 M.T., N.Dee, 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:

035 A.B., A.I.A., A.Ri., C.L., H.L., J.D.L., J.K.O., J.R.E., J.R.N., M.M.B., R.G.C. ATAC data generation:

036 A.E.S., B.B.L., B.R., B.T., C.R.P., C.S.L., D.D., J.C., J.D.L., J.K.O., K.Z., L.T.G., M.M.B., N.P., S.P.,

037 W.J.R., X.H., X.W. Electrophysiology, morphology, and Patch-seq data generation: A.L.K., B.E.K.,

038 D.M., E.S.L., G.D.H., J.Go., J.T.T., K.Sm., M.T., N.Dem., P.R.N., R.D., S.A.S., S.O., T.L.D., T.P.,

039 W.J.S. Data archive and infrastructure: A.E.S., A.M., B.R.H., H.C.B., J.A.M., J.Go., J.K., J.O., M.M.,

040 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:

041 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.Gi., J.Go.,

042 J.R.E., J.T.T., K.Sm., M.C., N.Dem., N.L.J., O.P., P.V.K., Q.H., R.F., R.H.S., R.Z., S.F., S.O., T.E.B.,

043 T.H., W.D., W.T., Y.E.L., Z.Y. Data interpretation: A.D., A.Re., B.B.L., B.E.K., B.T., C.K., C.L., E.S.L.,

044 F.X., H.L., H.Z., J.Gi., J.Go., J.R.E., J.T.T., M.C., M.H., M.M.B., N.Dem., N.L.J., P.R.H., P.V.K., Q.H.,

045 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.,

046 C.K., E.S.L., F.M.K., M.C., N.Dem., N.L.J., P.R.H., Q.H., R.H.S., T.E.B., W.J.S., W.T.

047

048 ## Competing interests

049 A.R. is an equity holder and founder of Celsius Therapeutics, a founder of Immunitas, and an SAB

050 member in Syros Pharmaceuticals, Neogene Therapeutics, Asimov, and Thermo Fisher Scientific. B.R.

051 is a shareholder of Arima Genomics, Inc. K.Z. is a co-founder, equity holder and serves on the

052 Scientific Advisor Board of Singlera Genomics. P.V.K. serves on the Scientific Advisory Board to

053 Celsius Therapeutics Inc.

054

055   ## Materials & Correspondence

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

057
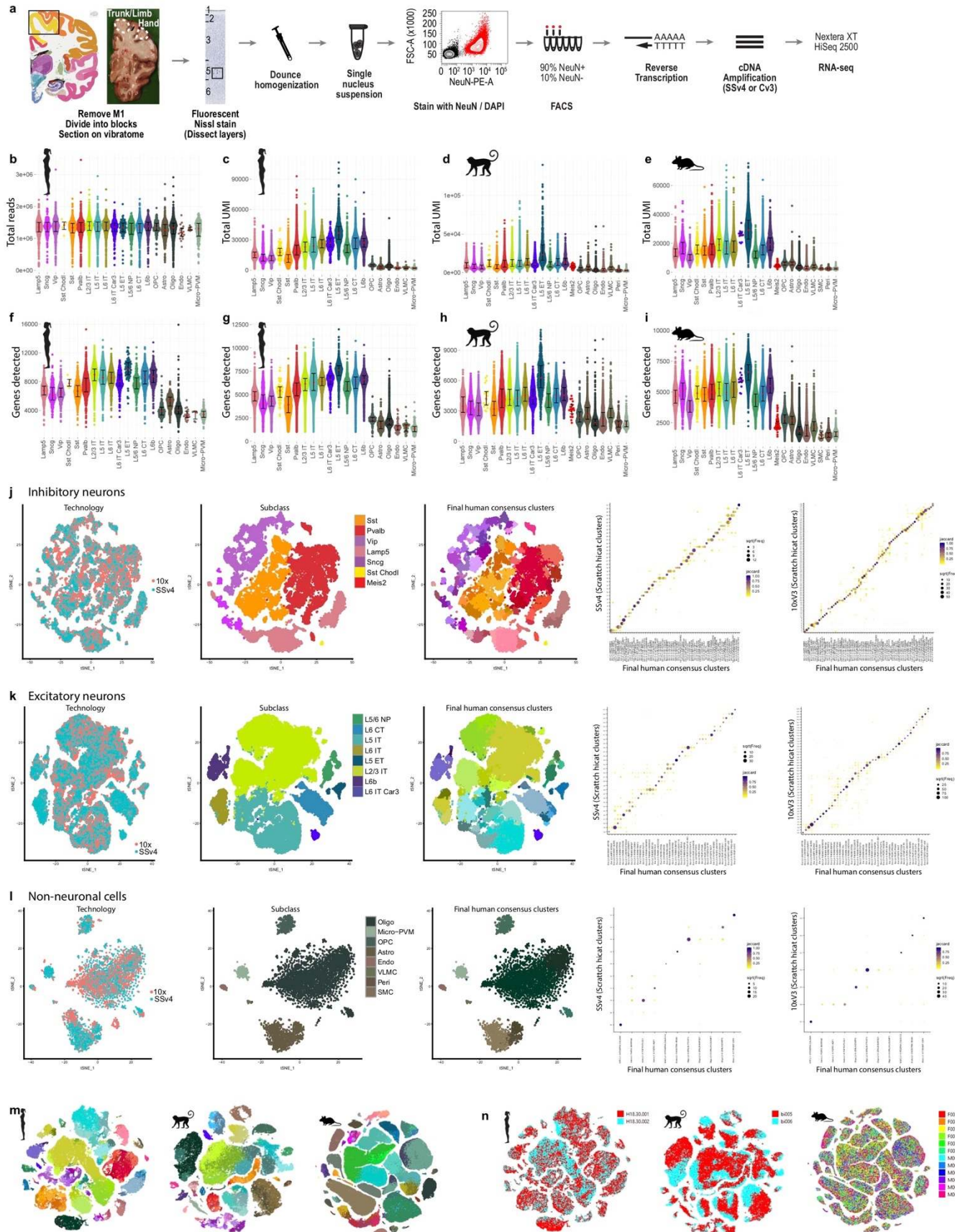
059 **Extended Data Figure 1. RNA-seq quality metrics and integration of human datasets. a**,

060 Schematic of single-nucleus isolation from M1 of post-mortem human brain and profiling with RNA-seq.

061 Box in the Nissl image highlights a cluster of Betz cells in L5. **b**, Using SSv4, > 1 million total reads

062 were sequenced across all subclasses in human. **c-e**, Using Cv3, total unique molecular identifiers

063 (UMI) varies between subclasses, and these differences are shared across species. **f-i**, Gene detection

064 (expression > 0) is highest in human using SSv4 (**e**) and lowest for marmoset using Cv3 (**h**). Note that

065 the average read depth used for SSv4 was approximately 20-fold greater than for Cv3 (target 60,000

066 reads per nucleus). **j-k**, tSNE projections of single nuclei based on expression of several thousand

067 genes with variable gene expression and colored by cluster label (**j**) or donor (**k**). **l-n**, Integration of

068 SSv4 and Cv3 RNA-seq datasets from human single nuclei isolated from GABAergic (**l**) and

069 glutamatergic (**m**) neurons and non-neuronal cells (**n**). Left: UMAP visualizations colored by RNA-seq

070 technology, cell subclass, and unsupervised consensus clusters. Right: Confusion matrices show

071 membership of SSv4 and Cv3 nuclei within integrated consensus clusters.

072
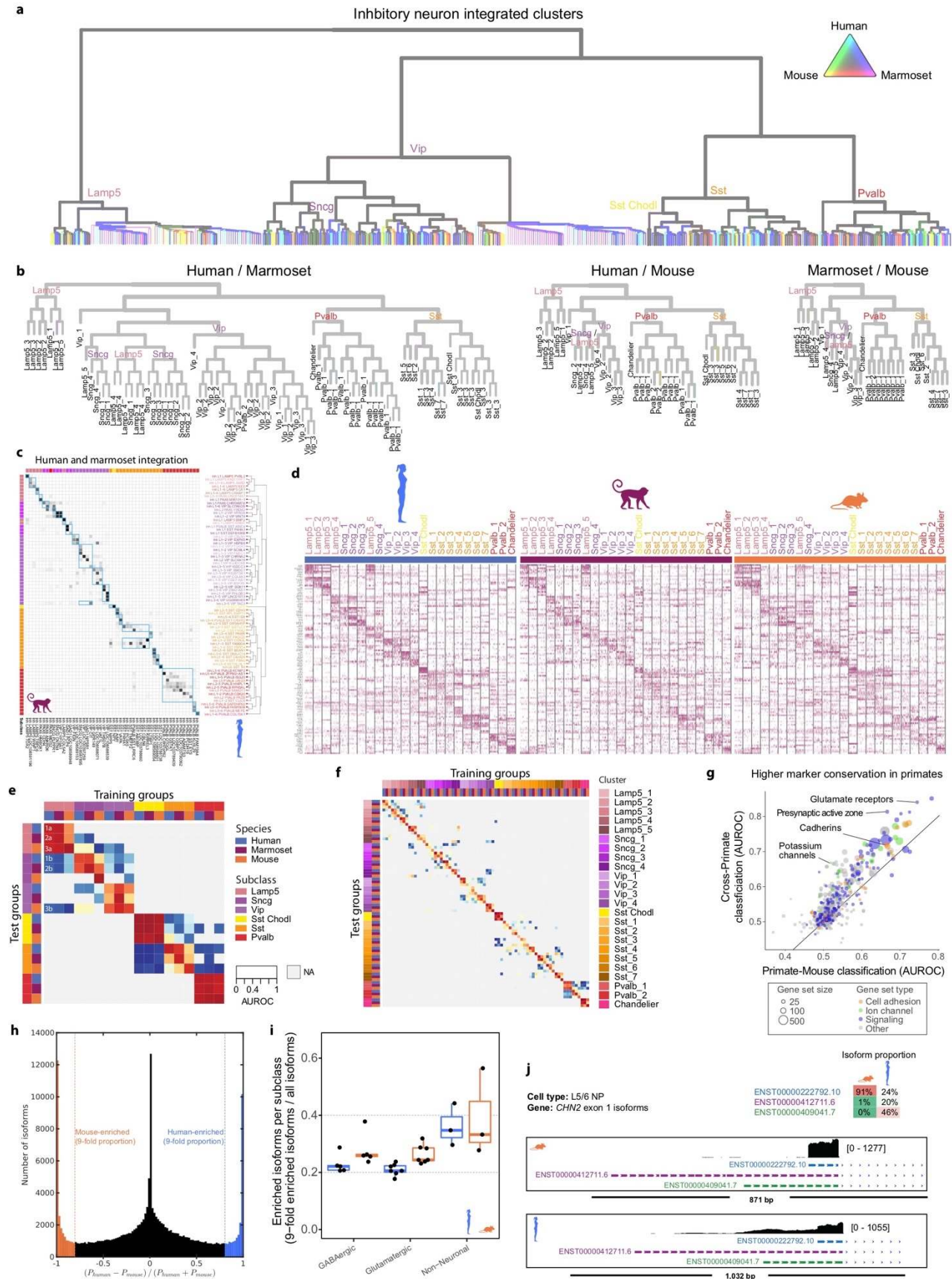
074 **Extended Data Figure 2. RNA-seq integration of GABAergic neurons across species. a**,

075 Dendrogram of GABAergic neuron clusters from unsupervised clustering of integrated RNA-seq data

076 from human, marmoset and mouse. Edge thickness indicates the relative number of nuclei, and edge

077 color indicates species mixing (grey is well mixed). Major branches are labeled by subclass.

078 Dendrogram shown in **Figure 2f** is derived from this tree based on pruning species-specific branches.

079 **b,** Dendrograms of pairwise species integrations from **Figure 2g** with leaves labeled by cross-species

080 clusters and edges colored by species mixing. **c,** Cluster overlap heatmap from human-marmoset

081 pairwise Seurat integration showing the proportion of within-species clusters that coalesce within

082 integrated clusters. Columns and rows are ordered as in **Figure 2e** with cross-species consensus

083 clusters indicated by blue boxes. Top and left color bars indicate subclasses of within-species clusters.

084 **d,** Heatmaps showing scaled expression of the top 5 marker genes for each GABAergic cross-species

085 cluster, and 5 marker genes for *Lamp5* and *Sst*. Initial genes were identified by performing a Wilcox test

086 of every integrated cluster against every other GABAergic nuclei. Additional DEGs were identified for

087 *Lamp5* and *Sst* cross-species clusters, by comparing one of the cross-species clusters to all other

088 related nuclei (e.g. *Sst*_1 against all other *Sst*). **e-f,** Heatmap of 1-vs-best MetaNeighbor scores for

089 GABAergic subclasses (**e**) and clusters (**f**). Each column shows the performance for a single training

090 group across the three test datasets. AUROCs are computed between the two closest neighbors in the

091 test dataset, where the closer neighbor will have the higher score, and all others are shown in gray

092 (NA). For example, in **e** the first column contains results of training on human *Lamp5*, labeled with

093 numbers to indicate test datasets, where 1 is human, 2 is marmoset and 3 is mouse, and letters to

094 indicate closest (**a**) and second-closest (**b**) neighboring groups. Dark red 3x3 blocks along the diagonal

095 indicate high transcriptomic similarity across all three species. **g**, Scatter plot of MetaNeighbor analysis

096 showing the performance (AUROC) of gene sets to classify GABAergic neuron consensus types by

097 training with human or marmoset data and testing with the other species (Cross-Primate, y-axis) or

098 training with primate data and testing with mouse (Primate-Mouse, x-axis). Gene set size and type are

099 indicated by point size and color, respectively. **h**, Histogram of the relative difference in isoform genic

100   proportion (P) between human and mouse for all subclass comparisons. All moderately to highly

101   expressed isoforms were included (gene TPM > 10 in both species; isoform TPM > 10 and proportion >

102   0.2 in either species). Vertical lines indicate >9-fold change in mouse or human. **i**, Proportion of all

103   isoforms in **h** that switch between species (FDR P < 0.05; >9-fold change in P) summarized by

104   subclass and grouped by cell class. **j**, Comparison between species of isoform genic proportions for the

105   top three most common isoforms of Chimerin 2 (*CHN2*) expressed in the L5/6 NP subclass. Genome

106   browser tracks of RNA-seq (SSv4) reads in human and mouse at the *CHN2* locus.

107
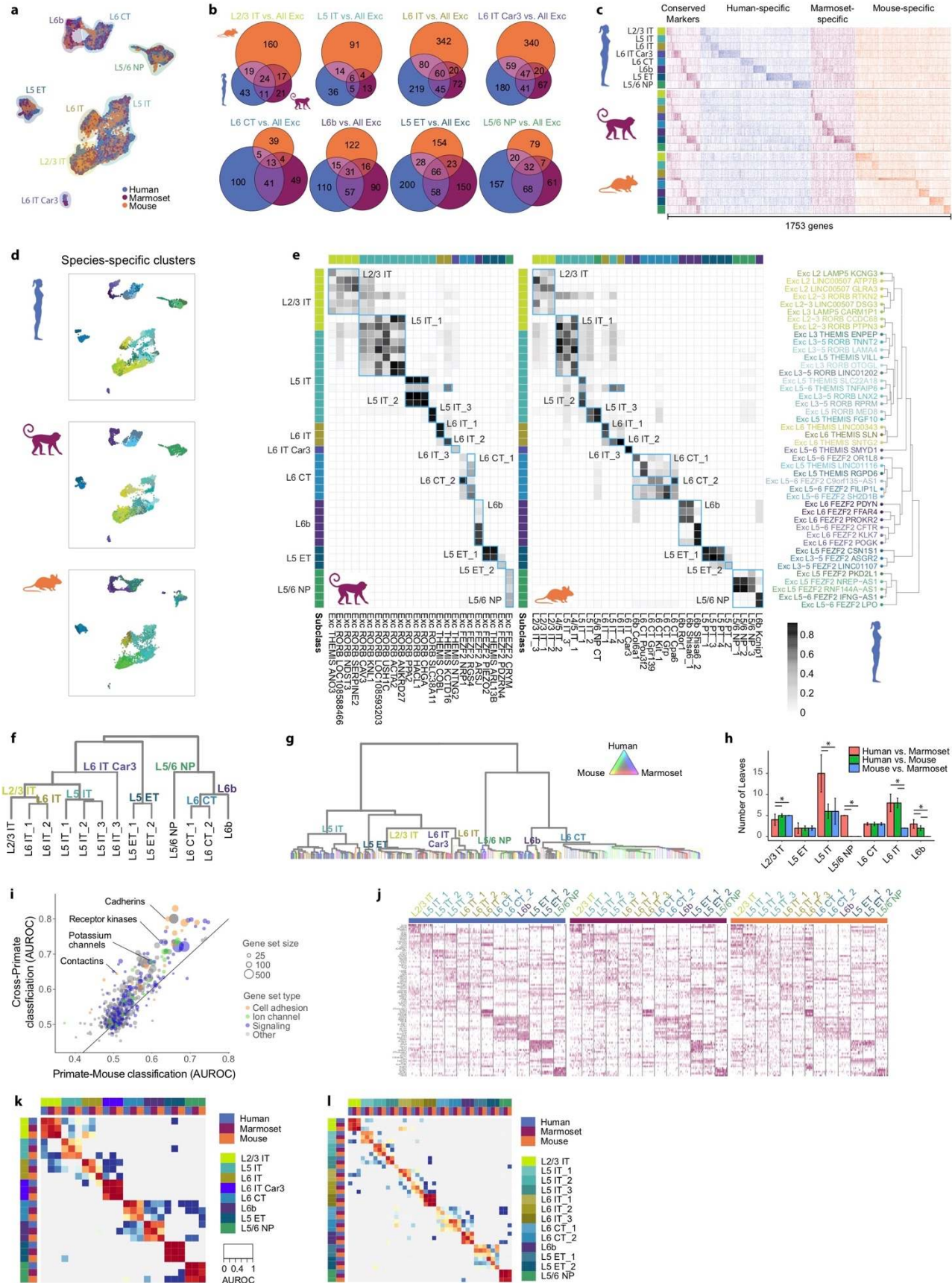
109   **Extended Data Figure 3. Glutamatergic neuron cell type homology across species. a,** UMAP

110   visualization of integrated snRNA-seq data from human, marmoset, and mouse glutamatergic neurons.

111   Highlighted colors indicate subclass. **b,** Venn diagrams indicating number of shared DEGs across

112   species by subclass. DEGs determined by ROC test of subclass against all other glutamatergic

113   subclasses within a species. **c,** Heatmap of all DEGs from **b** ordered by subclass and species

114   enrichment. Heatmap shows expression scaled by column for up to 50 randomly sampled nuclei from

115   each subclass for each species. **d,** UMAP visualization of integrated snRNA-seq data with projected

116   nuclei split by species. Colors indicate different within-species clusters. **e,** Cluster overlap heatmap

117   showing the proportion of within-species clusters that coalesce with a given integrated cross-species

118   cluster. Cross-species clusters are labelled and indicated by blue boxes with human-marmoset overlap

119   shown to the left and human-mouse overlap shown to the right. Top and left axes indicate the subclass

120   of a given within-species cluster by color. Bottom axis indicates marmoset (left) and mouse (right)

121   within species clusters. Right axis shows the glutamatergic branch of the human dendrogram from

122   **Figure 1c**. **f,** Dendrogram of glutamatergic neuron cross-species clusters. **g,** Unpruned dendrogram of

123   glutamatergic neuron clusters from unsupervised clustering of integrated RNA-seq data. Edge

124   thickness indicates the relative number of nuclei, and edge color indicates species mixing. Major

125   branches are labeled by subclass. **h,** Bar plots quantifying the number of well-mixed clusters from

126   unsupervised clustering of pairwise species integrations. Significant differences (adjusted P < 0.05,

127   Tukey's HSD test) between species are indicated for each subclass. **i,** Scatter plot of MetaNeighbor

128   analysis showing the performance (AUROC) of gene sets to classify glutamatergic neuron consensus

129   types by training with human or marmoset data and testing with the other species (Cross-Primate, y-

130   axis) or training with primate data and testing with mouse (Primate-Mouse, x-axis). Gene set size and

131   type are indicated by point size and color, respectively. **j,** Heatmaps showing scaled expression of

132   marker genes for each glutamatergic cross-species cluster. The top 5 marker genes for each cross-

133   species cluster are shown, with an additional 5 genes for L5 ET, L5 IT, and L6 IT. Initial genes were

134   identified by performing a Wilcox test of every integrated cluster against every other glutamatergic

135   nuclei. Additional DEGs were identified for L5 ET, L5 IT, and L6 IT cross-species clusters, by

136   comparing one of the cross-species clusters to all other related nuclei (e.g. L5 IT_1 against all other L5

137   IT). **k**, **l,** Heatmap of 1-vs-best MetaNeighbor scores for glutamatergic subclasses (**k**) and clusters (**l**).

138   Results are displayed as in **Extended Data Fig. 2e,f**.

139

140
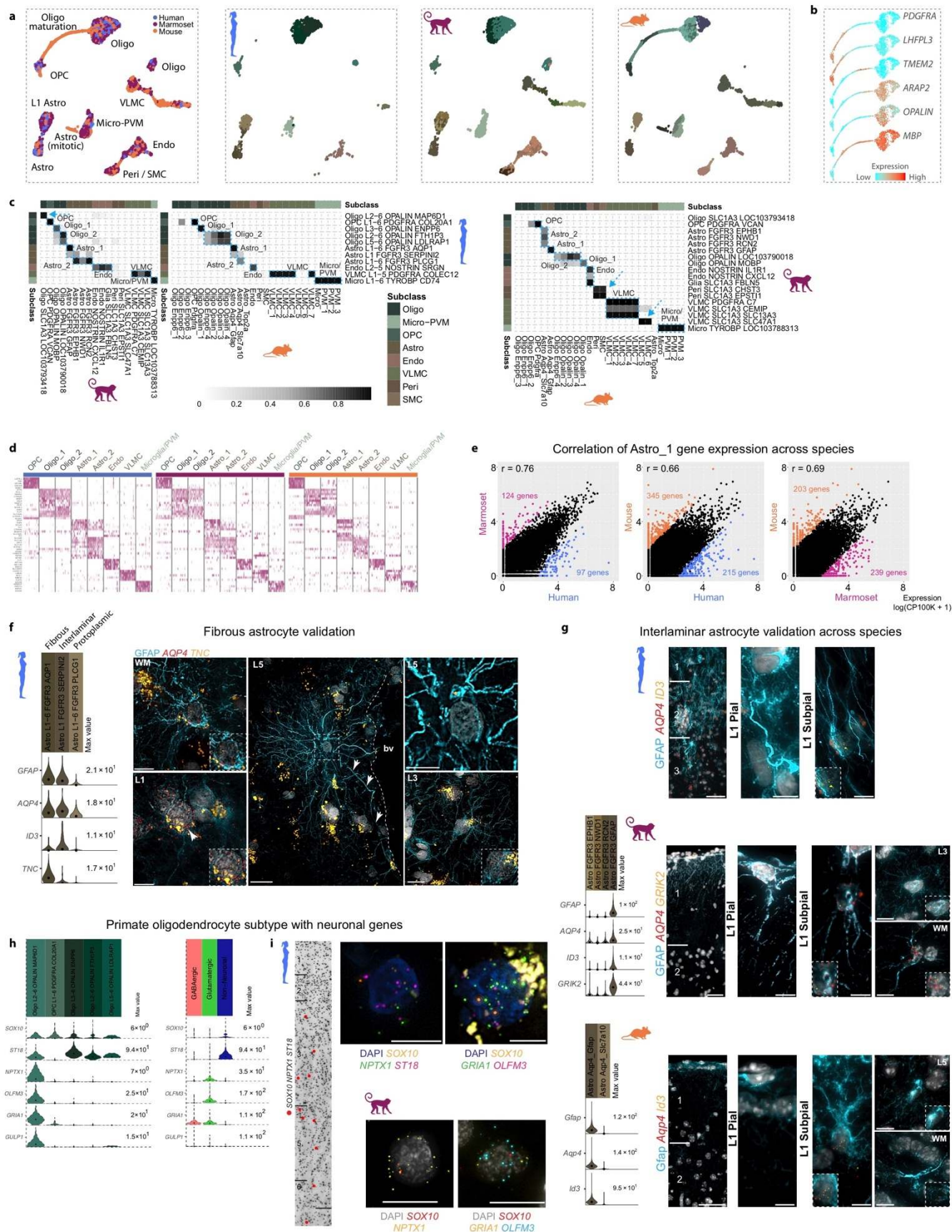
141   **Extended Data Figure 4. Non-neuronal cell type homology across species. a**, UMAP plots of

142   integrated RNA-seq data for non-neuronal nuclei, colored by species and within-species clusters. Note

143   that some cell types are present in only one or two species. **b**, UMAP of mouse oligodendrocyte

144   precursors and mature cells showing expression levels of marker genes for different stages of cell

145   maturation. **c**, Heatmaps of the proportion of nuclei in each species-specific cluster that overlap in the

146   integrated RNA-seq analysis. Blue boxes define homologous cell types that can be resolved across all

147   three species. Arrows highlight clusters that overlap between two species and are not detected in the

148   third species, due to differences in sampling depth of non-neuronal cells, relative abundances of cell

149   types between species, or evolutionary divergence. **d**, Conserved marker genes for homologous cell

150   types across species. **e**, Pairwise comparisons between species of log-transformed gene expression of

151   the Astro_1 type. Colored points correspond to significantly differentially expressed (DE) genes (FDR <

152   0.01, log-fold change > 2). r, Spearman correlation. **f**, Fibrous astrocyte in situ validation. Violin plots of

153   marker genes of human astrocyte clusters that correspond to fibrous, interlaminar, and protoplasmic

154   types based on in situ labeling of types. Left ISH: Fibrous astrocytes located in the white matter (WM,

155   top) and a subset of L1 (bottom) astrocytes express the Astro L1-6 *FGFR3 AQP1* marker gene *TNC*.

156   Middle ISH: Image of putative varicose projection astrocyte located in cortical L5 adjacent to a blood

157   vessel (bv) and extending long GFAP-labeled processes (white arrows) does not express the marker

158   gene *TNC*. The white dashed box indicates the area shown at higher magnification in the top right

159   panel. Likewise, the L3 protoplasmic astrocyte shown in the bottom right panel does not express *TNC*.

160   **g**, Combined GFAP immunohistochemistry and RNAscope FISH for markers of L1 astrocytes in

161   human, mouse, and marmoset. In human (top), pial and subpial interlaminar astrocytes are labeled with

162   *AQP4* and *ID3* and extend long processes from L1 down to L3. In marmoset (middle), both pial and

163   subpial L1 astrocytes express *AQP4* and *GRIK2* and extend GFAP-labeled processes through L1 that

164   terminate before reaching L2. An image of a marmoset protoplasmic astrocyte located in L3 shows that

165   this astrocyte type does not express the marker gene *GRIK2*. A subset of marmoset fibrous astrocytes

166   located in the white matter (WM) express *GRIK2*, suggesting that fibrous and L1 astrocytes have a

167   shared gene expression signature as shown in human [2]. L1 astrocytes in mouse (bottom) consist of pial

168   and subpial types that differ morphologically but are characterized by their expression of the genes

169   *Aqp4* and *Id3*. Pial astrocytes in mouse extend short Gfap-labeled processes that terminate within L1

170   whereas mouse subpial astrocytes appear to extend processes predominantly toward the pial surface.

171   Protoplasmic astrocytes (example shown in L5) do not express *Id3*, whereas fibrous astrocytes in

172   mouse share expression of *Id3* with L1 astrocyte types. Inset images outlined with white dashed boxes

173   illustrate cells in each of the accompanying images at higher magnification to show RNAscope spots for

174   each gene labeled. Scale bars, 20 μm. **h**, Violin plots of marker genes of oligodendrocyte lineage

175   clusters in human. Transcripts detected in the Oligo L2−6 *OPALIN MAP6D1* cluster include genes

176   expressed almost exclusively in neuronal cells. Scale bars, 20 μm. **i**, Left: Inverted DAPI image

177   showing a column of cortex labeled with markers of the human Oligo L2-6 *OPALIN MAP6D1* type. Red

178   dots show cells triple labeled with *SOX10*, *NPTX1*, and *ST18*. Top right: Examples of cells labeled with

179   marker gene combinations specific for the human Oligo L2-6 *OPALIN MAP6D1* type. Bottom right:

180   Example of a marmoset cell labeled with the marker genes *OLIG2* and *NRXN3*. Scale bars, 20 μm.

181

182
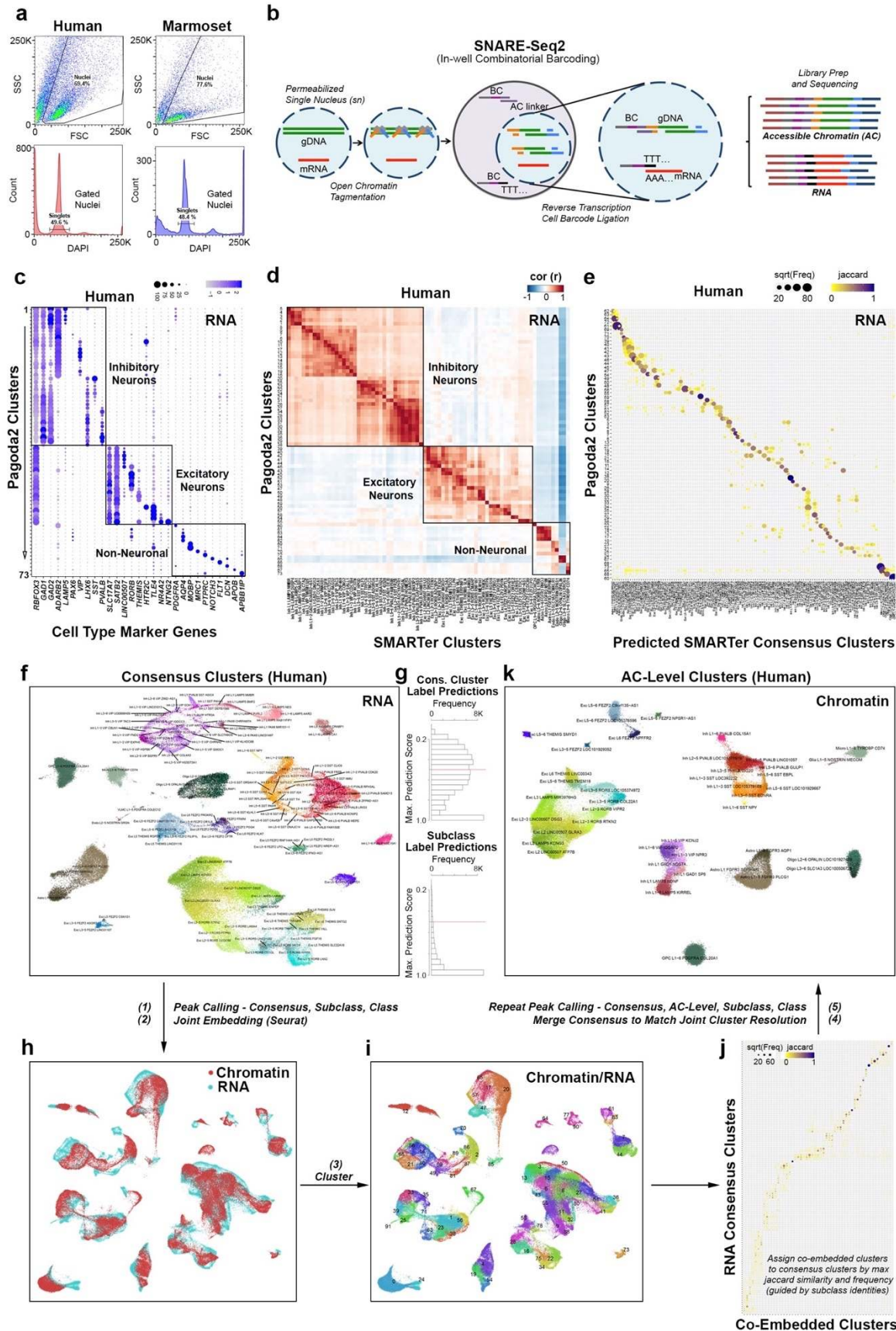
183   **Extended Data Figure 5. SNARE-seq2 transcriptomic profiling resolves M1 cell types**. **a-b**, FACS

184   gating parameters used for sorting human and marmoset single nuclei (**a**) that were used for SNARE-

185   seq2 as outlined in (**b**), to generate both RNA and accessible chromatin (AC) libraries having the same

186   cell barcodes. **c**, Dot plot showing averaged marker gene expression values (log scale) and proportion

187   expressed for clusters identified in a preliminary analysis of SNARE-seq2 RNA using Pagoda2. **d**,

188   Correlation heatmap of averaged scaled gene expression values for Pagoda2 clusters against SSv4

189   clusters from the same M1 region. **e**, Jaccard similarity plot for cell barcodes grouped according to

190   Pagoda2 clustering compared against the predicted SSv4 consensus clustering. **f-k**, Overview of AC-

191   level cluster assignment using RNA-defined clusters indicating the five main steps of the process. **f**,

192   Consensus clusters visualized by UMAP on RNA expression data and that were used to independently

193   call peaks from AC data. **g**, Histograms showing maximum prediction scores for consensus cluster

194   (top) and subclass (bottom) labels from RNA data to corresponding accessibility data (cicero gene

195   activities). **h**, Consensus cluster peaks, as well as those identified from subclass and class level

196   barcode groupings, were combined and the corresponding peak by cell barcode matrix was used to

197   predict gene activity scores using Cicero for integrative RNA/AC analyses. UMAP shows joint

198   embedding of RNA and imputed AC expression values using Seurat/Signac. **i**, UMAP showing clusters

199   identified from the joint embedding (**h**). **j**, Jaccard similarity plot comparing cell barcodes either grouped

200   according to RNA consensus clustering or joint RNA/AC clustering (**i**). RNA consensus clusters were

201   merged to best match the cluster resolution achieved from co-embedded clusters. Chromatin peak

202   counts generated from peak calling independently on consensus, AC-level, subclass, and class

203   barcode groupings were used to generate a final peak by cell barcode matrix. **k**, Final AC-level clusters

204   visualized using UMAP.

205

206
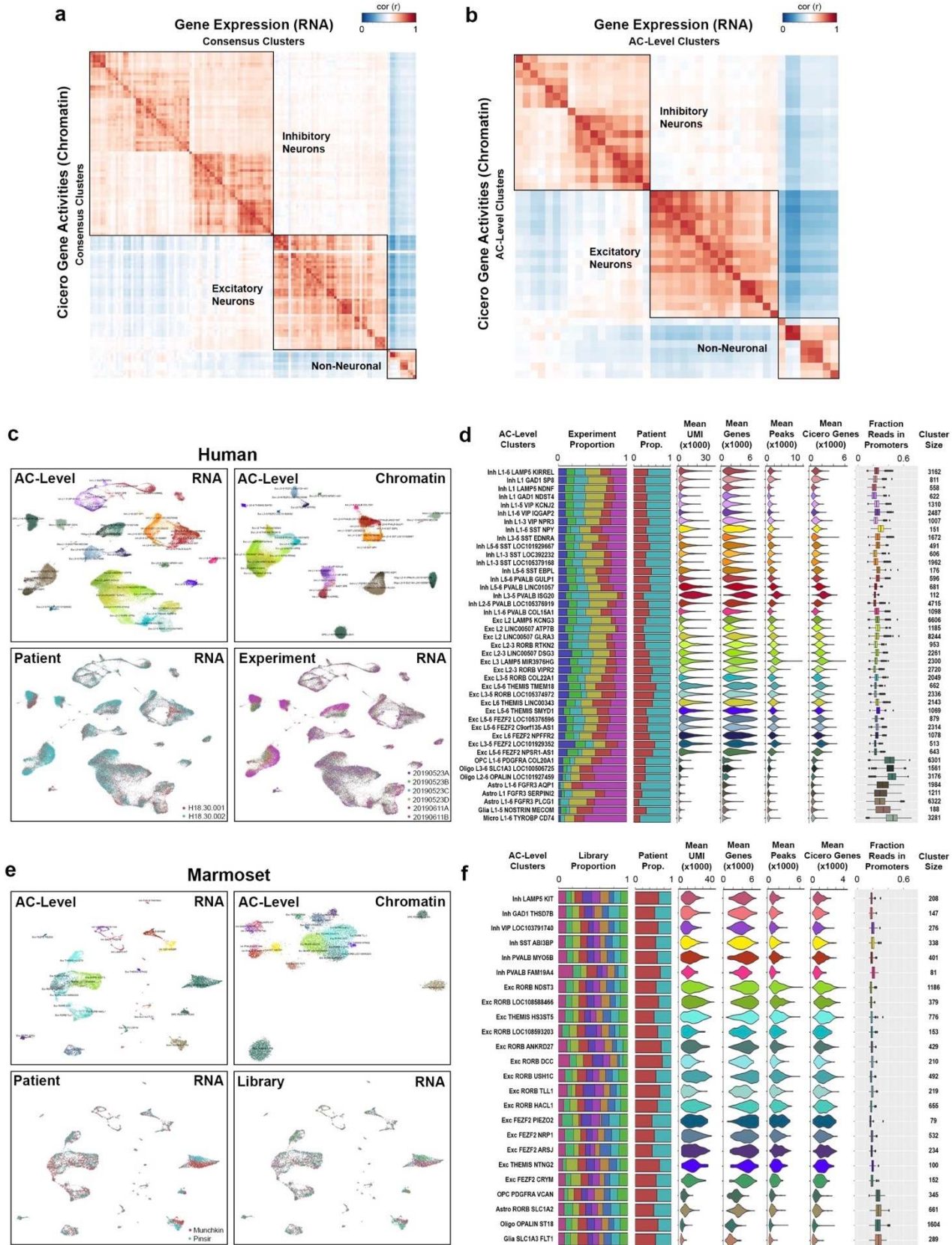
207   **Extended Data Figure 6. SNARE-Seq2 quality statistics. a-b**, Correlation heatmaps of average

208   scaled gene expression values against average scaled Cicero gene activity values for consensus

209   clusters (**a**) and AC-level clusters (**b**). **c**, UMAP plots showing human AC-level clusters for both RNA

210   and chromatin data, as well as the corresponding patient and experiment identities for the RNA

211   embeddings. **d**, Bar, violin and box plots for human AC-level clusters showing proportion contributed by

212   each experiment or patient, mean UMI and genes detected from the RNA data, the mean peaks and

213   cicero active genes detected from AC data, the fraction of reads found in promoters for AC data, and

214   the number of nuclei making up each of the clusters. **e**, UMAP plots showing marmoset AC-level

215   clusters for both RNA and chromatin data, as well as the corresponding patient and library identities for

216   the RNA embeddings. **f**, Bar, violin and box plots for marmoset AC-level clusters showing proportion

217   contributed by each library or patient, mean UMI and genes detected from the RNA data, the mean

218   peaks and cicero active genes detected from AC data, the fraction of reads found in promoters for AC

219   data, and the number of nuclei making up each of the clusters.

220

221

222   **Extended Data Figure 7. DNA-methylation cell type and integration with RNA-seq data. a-b,**

223   UMAP visualization of marmoset M1 and mouse MOp DNA methylation (snmC-seq2) data and cell

224   clusters. **c-e,** Mapping between DNAm-seq and RNA-seq clusters from human (**c**), marmoset (**d**), and

225   mouse (**e**). Number of nuclei in each cluster are listed in parentheses. **f,** Numbers of hypo- and hyper-

226   methylated DMRs and overlap with chromatin accessible peaks in each subclass of human. **g,**

227   Numbers of chromatin accessible peaks and overlap with DMRs in each subclass of human.

228
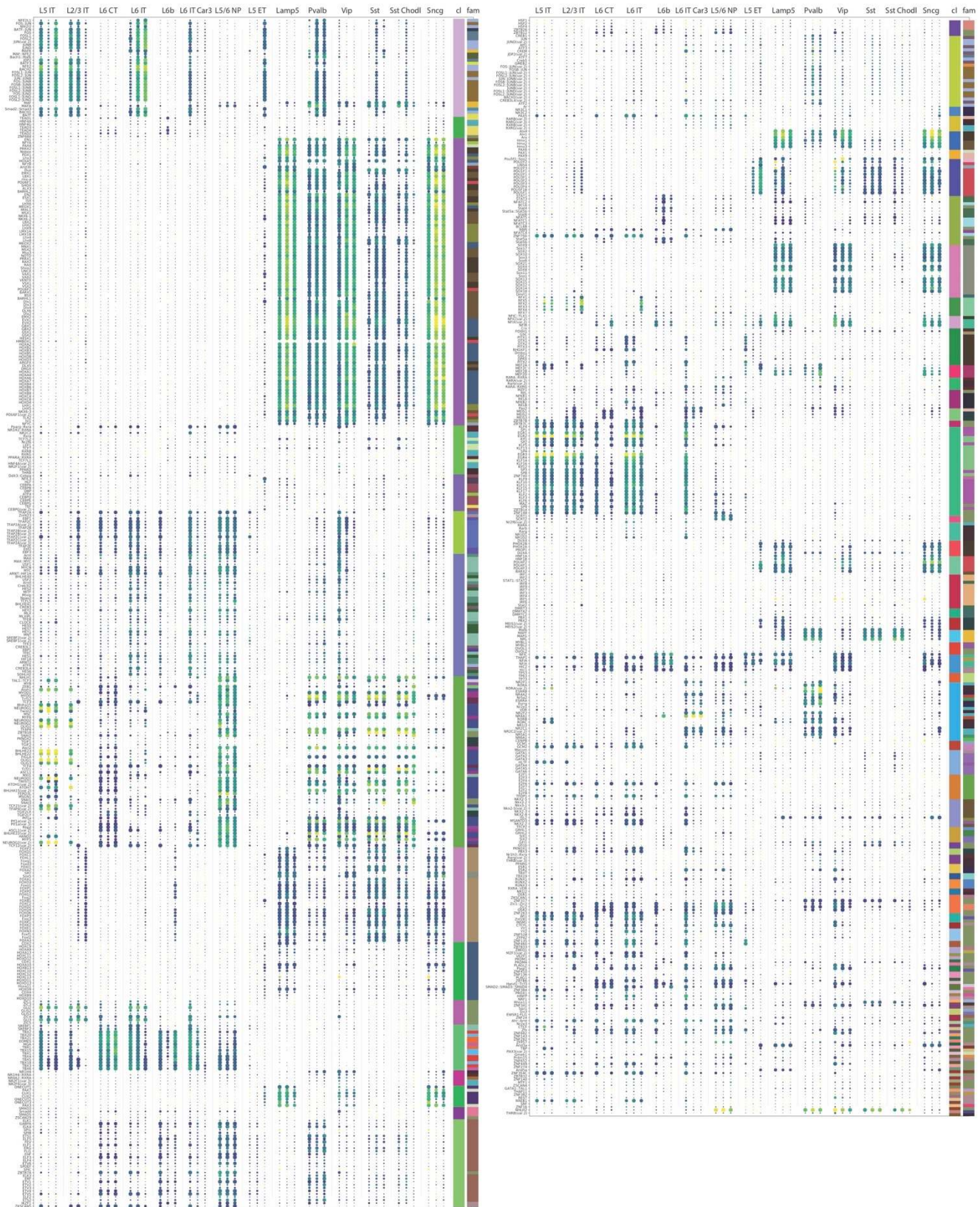
230 **Extended Data Figure 8. TFBS enrichment analysis on hypo-methylated DMRs at subclass level**

231 **show conservativity of gene regulation across species.** Motif enrichment analysis of TFBS were

232 conducted using JASPAR's non-redundant core vertebrata TF motifs for neuronal subclasses in each

233 species. Each subclass tri-column shows the results of human, marmoset and mouse, respectively

234 from left to right. The size of a dot denotes the p-value of the corresponding motif, while the color

235 denotes the fold change. The rightmost two columns show TF clusters (cl) identified from motif profiles

236 and TF family (fam) identified from TF structures.

237

238
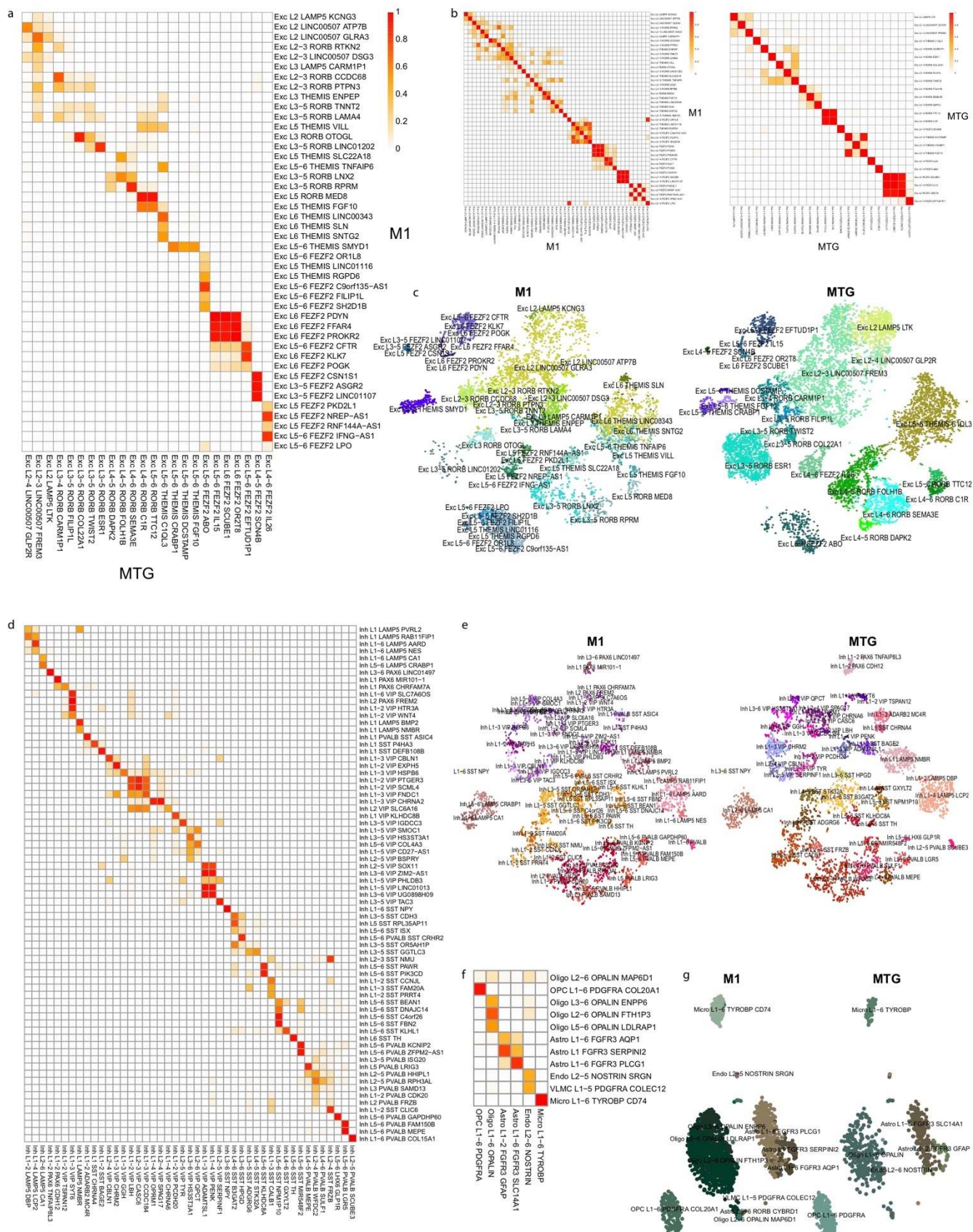
239   **Extended Data Figure 9. Cell type homologies between human cortical areas based on RNA-seq**

240   **integration. a**, Heatmap of glutamatergic neuron cluster overlap between M1 and MTG. **b**, Heatmaps

241   of glutamatergic neuron cluster overlap for M1 and MTG test datasets. Clusters were split in half and

242   two datasets were integrated using the same analysis pipeline as the M1 and MTG integration. Most

243   clusters mapped correctly (along the diagonal) with some loss in resolution between closely related

244   clusters (red blocks). **c**, tSNE plots of integrated glutamatergic neurons labeled with M1 and MTG

245   clusters. **d-g**, Cluster overlap heatmaps and tSNE plots of integrations of GABAergic neurons (**d**, **e**)

246   and non-neuronal cells (**f**, **g**), as described for glutamatergic neurons.

247
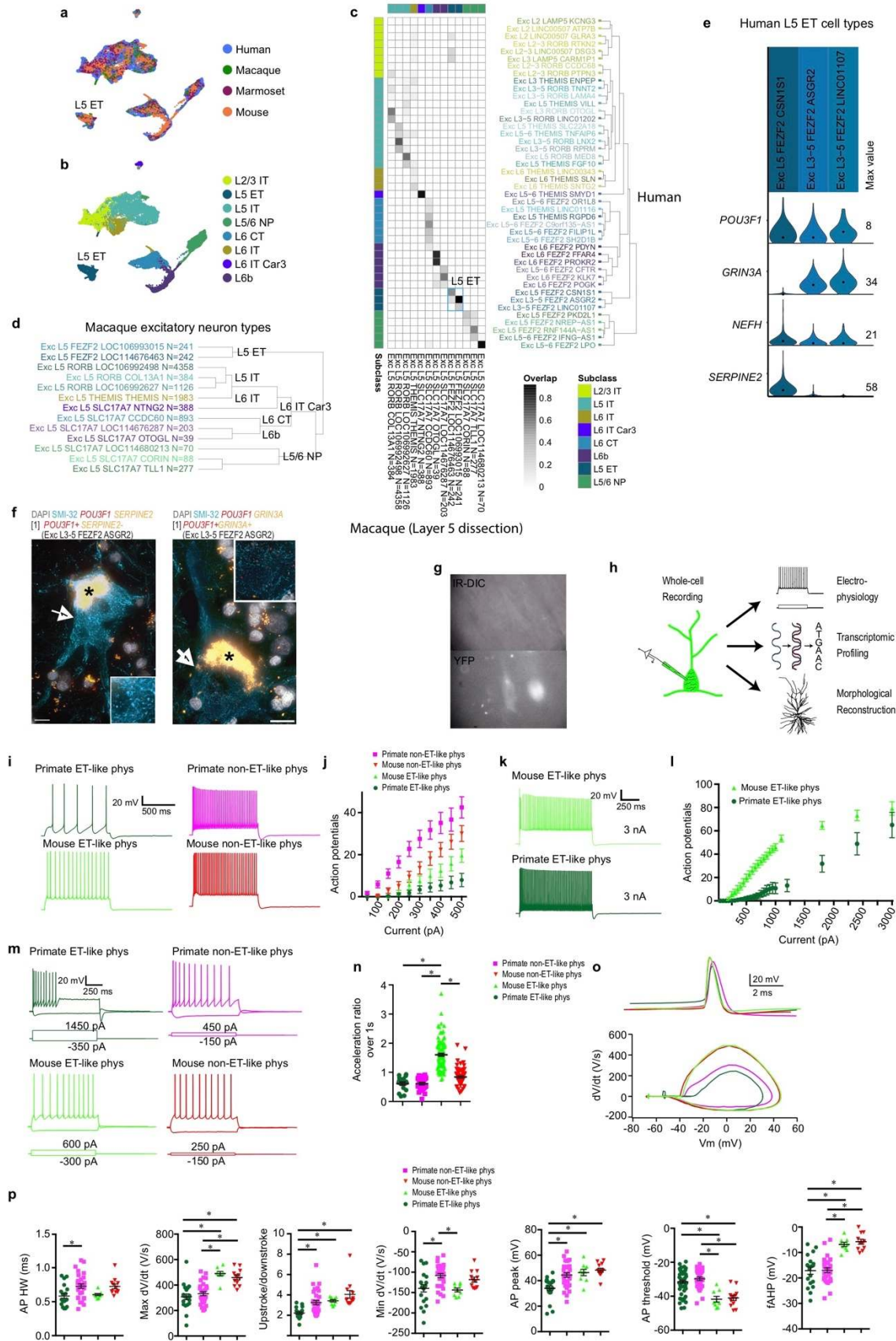
249   **Extended Data Figure 10. Cross-species alignment of glutamatergic neurons and differences in**

250   **L5 neuron spike trains and single spike properties. a, b,** UMAP visualizations of cross-species

251   integration of snRNA-seq data for glutamatergic neurons isolated from human, macaque (L5 dissection

252   only), marmoset, and mouse. Colors indicate species (**a**) or cell subclass (**b**). **c**, Cluster overlap

253   heatmap showing the proportion of nuclei from within-species clusters that are mixed within the same

254   integrated clusters. Human clusters (rows) are ordered by the dendrogram reproduced from **Figure 1c**.

255   Macaque clusters (columns) are ordered to align with human clusters. Color bars at top and left indicate

256   subclasses of within-species clusters. Blue box denotes the L5 ET subclass. **d,** Dendrogram showing

257   all macaque clusters from L5 dissection with subclasses denoted to the right. **e,** Violin plot showing

258   expression of marker genes for human L5 ET neuron subtypes. **f,** Two examples of ISH labeled, SMI-

259   32 IF stained Betz cells in L5 of human M1 that correspond to the L5 ET cluster Exc L3-5 *FEZF2*

260   *ASGR2*. Insets show higher magnification of ISH-labeled transcripts in corresponding cells. Scale bars,

261   20 μm. Asterisks mark lipofuscin. **g,** Example IR-DIC (top) and fluorescent (bottom) images obtained

262   from a macaque organotypic slice culture. Note the inability to visualize the fluorescently labeled

263   neurons in IR-DIC because of dense myelination. **h**, patch-seq involves the collection of morphological,

264   physiological and transcriptomic data from the same neuron. Following electrophysiological recording

265   and cell filling with biocytin via whole cell patch clamp, the contents of the cell are aspirated and

266   processed for RNA-sequencing. This permits a transcriptomic cell type to be pinned on the

267   physiologically-probed neuron. **i**, Example voltage responses to a 1 s, 500 pA step current injection. **j.**

268   Action potentials as a function of current injection amplitude. Primate ET neurons display shallowest

269   action potential-current injection relationship, perhaps partially because of their exceptionally low input

270   resistance. **k**, Voltage responses to a 1 s, 3 nA step current injection. **l**, Action potentials as a function

271   of current injection for a subset of experiments in which current injection amplitude was increased

272   incrementally to 3 nA. While both mouse and primate ET neurons could sustain high firing rates,

273   primate neurons required 3 nA of current over 1s to reach similar average firing rates as mouse ET

274   neurons. **m**, Example voltage responses to 1 s depolarizing step current injections. The amplitude of

275 the current injection was adjusted to produce ~10 spikes. Also shown are voltage responses to a

276 hyperpolarizing current injection. **n**, The firing rate of primate ET and IT neurons decreased during the 1

277 s step current injection, whereas, the firing rate of mouse ET neurons increased. Acceleration

278 ratio=2nd/last interspike interval. **o**, Example single action potentials (above) and phase plane plots

279 (below). **p**, Various action potential features are plotted as a function of cell type. Notably, action

280 potentials in primate ET neurons were reminiscent of fast spiking interneurons in that they were shorter

281 and more symmetrical compared with action potentials in other neuron types/species. Intriguingly, K$^+$

282 channel subunits K$_v$3.1 and K$_v$3.2 that are implicated in fast spiking physiology[91] are encoded by highly

283 expressed genes (*KCNC1* and *KCNC2*) in primate ET neurons (Fig. 7c) * p < 0.05, Bonferroni

284 corrected t-test.

285

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

286 **Extended Data Table 1.** Summary of human tissue donors. RIN, RNA integrity number. Data type:

287 SMART-Seqv4 (SSv4), 10x Genomics Chromium Single Cell 3' Kit v3 (Cv3), Single-Nucleus Chromatin

288 Accessibility and mRNA Expression sequencing (SNARE-seq2), Single-nucleus methylcytosine

289 sequencing (snmC-seq2).

290

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

291 **Extended Data Table 2.** Summary of marmoset specimens. Data type: 10x Genomics Chromium

292 Single Cell 3' Kit v3 (Cv3). ACD Bio multiplex fluorescent in situ hybridization (FISH).

293

294 Supplementary Table legends

295 **Supplementary Table 1.** Provisional cell ontology (pCL) terms for human, mouse, and marmoset

296 primary motor cortex cell types. Column headers are described as follows: pCL_id is a unique

297 alphanumeric identifier assigned to each provisional cell type. CL_id is the Cell Ontology (CL) identifier

298 for those parent cell type classes already represented in CL. pCL_name and Transcriptome data

299 cluster are labels given according to each species naming convention that combines information about

300 cortical layer enrichment and genes expressed in data cluster transcriptomes. TDC_id is a unique

301 identifier assigned to the transcriptome data cluster. The part_of (uberon_id) and part_of

302 (uberon_name) columns contain unique identifiers and names for tissue anatomic regions from which

303 the experiment specimen was derived, in this case primary motor cortex. The is_a (CL or pCL_id) and

304 is_a (CL or pCL_name) columns contain parent cell type or provisional cell type identifiers and names,

305 respectively. Cluster_size indicates the number of single-nucleus or cell transcriptomes that were

306 assigned membership to the transcriptome data cluster. Marker_gene_evidence indicates the number

307 of marker genes that are necessary and sufficient to define the transcriptome cell type data cluster with

308 maximal classification accuracy based on the NS-Forest v2.1 algorithm (see Supplementary Tables 4-

309 6). F-measure_evidence is the f-beta score of classification accuracy from the NS-Forest v2.1 algorithm

310 using the marker genes listed. The selectively_expresses column lists the minimum set of marker

311   genes necessary and sufficient to define the transcriptome cell type data cluster. The definition brings

312   together features to form a data driven ontological representation for each cell type cluster. The pCL

313   annotations are available at https://github.com/mkeshk2018/Provisional_Cell_Ontology and

314   https://bioportal.bioontology.org/ontologies/PCL.

315

316   **Supplementary Table 2.** Cluster annotations for human, marmoset, and mouse in separate

317   worksheets. Cluster_label column identifies the RNA-seq cluster within each species. Cluster_size

318   column denotes the number of nuclei that reside within each cluster (cluster_label). Class column

319   identifies which cell class each cluster belongs to. Subclass column identifies which cell subclass each

320   cluster belongs to. Cross-species cluster column indicates the cross-species consensus cluster

321   taxonomy. DNAm_cluster_label column identifies the transcriptomic cluster (cluster_label) that is

322   aligned to DNAm-determined clusters. ATAC_cluster label column identifies the transcriptomic cluster

323   (cluster_label) that is aligned to ATAC-determined clusters.

324

325   **Supplementary Table 3**. Application of Allen Institute nomenclature schema to mouse, marmoset, and

326   human M1 taxonomies. The "taxonomy_ids" tab lists ids and descriptions for the 11 taxonomies

327   included and which tab those taxonomies are shown on. The "preferred_aliases" tab shows a list of

328   preferred aliases for linking between taxonomies, as well as descriptions for these. The next five tabs

329   show nomenclatures for each of the taxonomies and have the following column headers: "tree_order" is

330   the order shown in the tree (if any); "cell_set_alias", "cell_set_label", and "cell_set_accession" are

331   unique identifiers, as described in the Allen Institute nomenclature page (https://portal.brain-

332   map.org/explore/classes/nomenclature), with "cell_set_alias" including the names used in this

333   manuscript; "cell_set_preferred_alias" indicates which clusters correspond to the "preferred_alias"es

334   from the previous tab, if any; "cell_set_alias_integrated" shows linkages between single species

335   transcriptomics taxonomies and the integrated taxonomy; "cell_set_labels_CS191213#" columns

336   indicate linkages between cell sets in the transcriptomics and other modalities within a single species;

337   "cell_set_descriptor" shows the type of cell set (or level of ontology); and "taxonomy_id" links to the

338   "taxonomy_id" tab. Finally, the "Cell class hierarchy" tab shows the ordered class, level2, and subclass

339   hierarchy and associated colors used as cell sets in previous tabs.

340

341   **Supplementary Table 4.** NS-Forest v2.1 was used to determine cell type cluster marker genes for all

342   annotated levels of the human primary motor cortex cell type taxonomy defined by RNA-seq (Cv3).

343   "clusterName" corresponds to the annotation label, either a cell type cluster name or a parent cell type

344   class in the taxonomy. "markerCount" gives the optimal number of marker genes in the set that best

345   discriminates the label. The "f-measure" column gives the f-beta score for classification using the set of

346   markers. The next four columns "True Negative", "False Positive", "False Negative", "True Positive" give

347   the confusion matrix for the label given the set of markers. Finally, "Marker 1-5" lists the gene symbols

348   corresponding to the optimal set of markers.

349

350   **Supplementary Table 5.** NS-Forest v2.1 was used to determine cell type cluster marker genes for all

351   annotated levels of the mouse primary motor cortex cell type taxonomy defined by RNA-seq (Cv3).

352   "clusterName" corresponds to the annotation label, either a cell type cluster name or a parent cell type

353   class in the taxonomy. "markerCount" gives the optimal number of marker genes in the set that best

354   discriminates the label. The "f-measure" column gives the f-beta score for classification using the set of

355   markers. The next four columns "True Negative", "False Positive", "False Negative", "True Positive" give

356   the confusion matrix for the label given the set of markers. Finally, "Marker 1-5" lists the gene symbols

357   corresponding to the optimal set of markers.

358

359   **Supplementary Table 6.** NS-Forest v2.1 was used to determine cell type cluster marker genes for all

360   annotated levels of the marmoset primary motor cortex cell type taxonomy defined by RNA-seq (Cv3).

361   "clusterName" corresponds to the annotation label, either a cell type cluster name or a parent cell type

362   class in the taxonomy. "markerCount" gives the optimal number of marker genes in the set that best

363   discriminates the label. The "f-measure" column gives the f-beta score for classification using the set of

364   markers. The next four columns "True Negative", "False Positive", "False Negative", "True Positive" give

365   the confusion matrix for the label given the set of markers. Finally, "Marker 1-5" lists the gene symbols

366   corresponding to the optimal set of markers.

367

368   **Supplementary Table 7.** DEGs determined by ROC test between each GABAergic neuron subclass

369   and all other GABAergic nuclei within each species. Columns are labeled myAUC, which contains AUC

370   scores > 0.7; avg_diff, which contains difference in expression between target subclass and all other

371   GABAergic neurons; power; pct.1, which indicates the percent of nuclei that express the gene in the

372   target cluster; pct.2, which indicates the percent of non-target nuclei that express the gene; cluster,

373   which denotes the target cluster; gene, indicating the gene that was identified as DE; and species,

374   which indicates the species the test was performed in.

375

376   **Supplementary Table 8.** List of DEGs (from Supplementary Table 7) that is sorted according to the

377   order the genes appear within the heatmap.

378

379   **Supplementary Table 9.** Supervised MetaNeighbor results, within- and across-species. Each row

380   corresponds to a unique entry for a given gene set and a given cell class, either Glutamatergic or

381   GABAergic. The first five columns provide information about the gene sets, namely their provenance

382   (SynGO or HGNC); numerical IDs; descriptive labels; manual classifications for plotting and

383   interpretation; and finally the number of genes included in the analysis (after subsetting to genes with 1-

384   1 orthologs across all three species). The sixth column indicates cell class. The remaining columns

385   contain MetaNeighbor AUROCs for various analyses: within_species_meanROC (column 7) provides

386   the mean of within-mouse (column 8), within-marmoset (column 9) and within-human (column 10)

387   performance. For each species, tests were run with random 3-fold cross-validation, and the average

388 across folds is reported. Columns 11 and 12 contain results from cross-species analyses, detailed in

389 the methods. Results are sorted by their AUROC across primates (column 12).

390

391 **Supplementary Table 10.** DEGs determined by ROC test between each glutamatergic neuron

392 subclass and all other glutamatergic nuclei within each species. Columns are labeled myAUC, which

393 contains AUC scores > 0.7; avg_diff, which contains difference in expression between target subclass

394 and all other glutamatergic neurons; power; pct.1, which indicates the percent of nuclei that express the

395 gene in the target cluster; pct.2, which indicates the percent of non-target nuclei that express the gene;

396 cluster, which denotes the target cluster; gene, indicating the gene that was identified as DE; and

397 species, which indicates the species the test was performed in.

398

399 **Supplementary Table 11.** List of DEGs (from Supplementary Table 10) that is sorted according to the

400 order the genes appear within the heatmap.

401

402 **Supplementary Table 12**. Average expression of isoforms in human and mouse subclasses and

403 estimates of isoform genic proportions (P) based on the ratio of isoform to gene expression. Isoforms

404 were included if they had at least moderate expression (TPM > 10) and P > 0.2 in either human or

405 mouse and at least moderate gene expression (TPM > 10) in both species.

406

407 **Supplementary Table 13.** SNARE-Seq2 metadata, cluster annotations and quality statistics. Tab 14a

408 indicates SNARE-Seq2 experiment level metadata (experiment name, library, patient, species,

409 purification, age, sex) and mapping statistics for RNA (mean UMI detected, mean genes detected) and

410 AC (mean fraction of reads in promoters or FRiP, mean uniquely mapped fragments grouped by 5000

411 base pair chromosomal bins, mean unique fragment counts per final peak locations, total number of

412 final nuclei). Tab 14b indicates the SNARE-Seq2 local RNA clusters for human M1 generated using

413 Pagoda2 (local cluster, annotated cluster name, broad cell type and abbreviation, k value used for

414   Pagoda2 clustering, broad cell type markers, level 1 and level2 classes and associated markers,

415   unique cluster markers). Tabs 14c-d indicates SNARE-Seq2 consensus or harmonized RNA and AC-

416   Level cluster annotations for human and marmoset M1, respectively, including annotated cluster name,

417   cluster order, associated subclass and class, and the number of datasets making up the clusters. Tabs

418   14e-f lists all metadata outlined in tabs 14a-d for all SNARE-Seq2 cell barcodes from human and

419   marmoset M1 samples, respectively.

420

421   **Supplementary Table 14.** SNARE-Seq2 differentially accessible regions for human and marmoset M1.

422   Tabs 15a and 15b show SNARE-Seq2 differentially accessible regions (DARs, q value < 0.001, log-fold

423   change > 1) identified by AC-Level clusters (15a) or subclass level (15b) for human M1, indicating for

424   each chromosomal location the p value (hypergeometric test), q value (Benjamini-Hochberg adjusted p

425   value), log-fold change and associated cluster or subclass. Tab 15b shows subclass DARs (q value <

426   0.001, log-fold change > 1) for marmoset subclasses as in tab 15b. Tab 15d shows a summarization of

427   human and marmoset DARs detected by matched subclasses, indicating actual number of DARs

428   detected (tabs 15b and 15c) and the values normalized to cluster size and total number of DARs

429   detected per species.

430

431   **Supplementary Table 15.** Cis-co-accessible sites, TF motif enrichments and differential TFBS

432   activities for human and marmoset M1. Tab 16a (human M1) and 16b (marmoset M1) show cis-

433   coaccessible network (CCAN) sites for subclass distinct markers genes (Wilcoxon Rank Sum test,

434   adjusted P < 0.05, average log-fold change > 0.5). pct.1 indicates the percent of nuclei that express the

435   gene in the target cluster, pct.2 indicates the percent of non-target nuclei that express the gene. For

436   each cluster and marker gene, corresponding motif enrichment values (hypergeometric test) for gene-

437   associated CCAN sites are shown ("observed" indicates number of features containing the motif,

438   "background" indicates the total number of features from a random selection of 40000 features that

439   contain the motif), and the motif associated differential chromVAR activity values identified using

440 logistic regression. The full list of chromVAR differentially active TFBS activities are also provided. Tab

441 16c summarizes the number of CCAN-associated marker genes, associated TFBSs enriched and or

442 active by subclass for both human and marmoset M1. Tabs 16d and 16e show cis-co-accessible sites,

443 TFBS enrichments and differential activities by AC-level clusters for human and marmoset M1,

444 respectively, similar to that provided in tabs 16a and 16b. Tab 16f shows chromVAR differentially active

445 TFBS activities by consensus or harmonized cluster using logistic regression. Tabs 16g, 16h, and 16i

446 show cis-co-accessible sites, TF motif enrichments and differential TFBS activities for human,

447 marmoset and mouse M1 ChCs compared against BCs.

448

449 **Supplementary Table 16**. snmC-seq2 metadata. The table shows experiment level metadata,

450 including species, sample name, gender, purification information, experiment nuclei numbers and pass-

451 QC nuclei numbers.

452

453 **Supplementary Table 17**. Subclass TFBS enrichment results. TFBS enrichment analysis was done

454 with AME [78] using JASPAR2020 motifs . Within a species, hypo-methylated DMRs in each subclass

455 were tested against hypo-methylated DMRs of all the other subclasses (background). DMRs and 250bp

456 around regions were used in the analysis. This table includes p-values and effect sizes (log2(TP/FP)) of

457 the analysis results.

458

459 **Supplementary Table 18**. Subclass TFBS enrichment at TF cluster level. TFs in SI Tab 18 were

460 grouped using clusters defined in Ref [42]. The table lists the most significant p-values and the largest

461 effect size of each TF cluster group.

462

463 **Supplementary Table 19.** DEGs determined by ROC test between chandelier cells and basket cells

464 within each species. Columns are labeled as species, with true/false values indicating if a gene was

465 enriched in chandelier cells for that species.

466

467   **Supplementary Table 20.** DEGs determined by ROC test between L5 ET subclass and L5 IT subclass

468   within each species. Columns are labeled as species, with values of 1 indicating a gene was enriched

469   in the L5 ET subclass for that species. A value of 0 indicates that the gene was not enriched in the L5

470   ET subclass for that species.

471

472   **Supplementary Table 21.** Genes with expression enrichment in L5 ET versus L5 IT that decreases

473   with evolutionary distance from human (human > macaque > marmoset > mouse). Columns are labeled

474   by species, and values indicate the log-fold change between L5 ET and L5 IT for that species. Genes

475   were included if they had a minimum log-fold change equal to 0.5 in human.