Cross-Species Single-Nucleus RNA Sequencing Analysis of Primary Motor Cortex: Insights from Humans, Chimpanzees, and Rats

Leo Dai

Received November 01, 2024
Accepted February 28, 2025
Electronic access March 31, 2025

The primary motor cortex (M1), a conserved structure in the mammalian brain, governs voluntary muscle movement. However, molecular-level differences between human and chimpanzee cell types remain poorly understood. A cross-species analysis could uncover specific cell types and gene developments aligned with motor skill evolution. This study analyzed single-nucleus RNA sequencing (snRNA-seq) data of more than 235,000 nuclei from human, chimpanzee, and rat samples. It utilized state-of-the-art machine learning techniques along with an innovative uncertainty-aware label transfer approach, facilitating the discovery of potentially human-specific cell types and more refined definitions of cell subclasses. Together with non-neuronal cells, this comparison across species revealed established excitatory and inhibitory neuronal populations, therefore identifying more than 50 different glutamatergic cell types. Excitatory neurons accounted for 60–65% of cells in humans and chimpanzees, compared to 70–75% in rats. This observation underscores the differences in cortical organization across species. This work discovered a small population of L4-like excitatory neurons in primates that express RORB, linking the transcriptomic profiles of layer 3 and layer 5. Should these L4-like cells be confirmed histologically, they may significantly contribute to corticothalamic and thalamocortical communication, providing valuable insights into the evolutionary complexity of the motor cortex. Finally, a group of human specific excitatory cell types exhibited unique gene expression patterns linked to neurodevelopmental disorders, including ADHD and autism. This indicates that these conditions may be unique to humans.

Introduction

Single Nucleus RNA Sequencing (snRNA-seq), is revolutionizing tissue mapping by giving precise RNA reads of hundreds of thousands of single nuclei in a tissue sample. Previous studies leveraging single-nucleus RNA sequencing have demonstrated its utility in uncovering neuronal subtypes and transcriptomic diversity in human cortical regions, paving the way for identifying orthologous neuronal populations and their specific gene expression profiles across species[1].

The primary motor cortex (M1) is one of the major brain areas responsible for controlling and learning skilled voluntary movement[2–4]. Human M1 is organized topographically, each body part has a corresponding representation location on the cortex, and the cortex itself is composed of six layers of neurons. Two main neuronal populations, GABAergic inhibitory neurons and glutamatergic excitatory neurons, were observed in M1. Interestingly, a species-specific adaptation of the proportions of neuron types and gene expressions were detected among humans, mice, and marmosets[1]. To explore the evolutionary conservation and divergence of M1 cell types and their transcriptomes, this study used snRNA-seq to analyze M1 cortices of humans, chimpanzees, and rats. In addition, this analysis utilized a novel uncertainty-aware cell type label transfer method to uncover potential unique cell types by cross species annotation. By comparing human data to chimpanzee data, the most closely genetically related animals to humans[5]. it will help identify uniquely human genetic features including disorders and diseases.

This analysis focused on excitatory neurons and identified over 50 excitatory (Exc), or glutamatergic (Glut) cell types across humans, chimpanzees, and rats. Hierarchical cell labeling was used in the paper, from cell classes, e.g., GABAergic, Glutamatergic, and non-neuronal, and cell subclasses, like "Vip, Sst", to individual cell types, e.g., "L5_IT_0". There were significant differences in the proportions of excitatory cells, with the rat samples exhibiting approximately 70–75% excitatory neurons, with chimpanzees and humans closer to 60–65%. Furthermore, as excitatory neurons move up the vertical layers of the motor cortex, they become less conserved across species. For example, a chimpanzee layer two (L2) neuron shares more marker genes with a human L2 neuron than a chimpanzee layer six neuron (L6) with a human L6. This pattern is not observed in inhibitory neurons, which remain overall much more tightly conserved across all species.

The analysis of primate excitatory neurons in this study suggests the potential existence of layer 4-like (L4-like) excitatory neurons in primates. The motor cortex has traditionally been described as 'agranular' or 'disgranular,' implying a less prominent L4 (Shipp et al., 2013). However, recent functional investiga-

tions suggest that M1 contains a circuit-level equivalent of L4 in the mouse[6,7]. We identified clusters of neurons sharing marker genes with both L3 and L5 cells, and clustering algorithms often struggled to split these clusters. Changing the clustering organization to account for cross layer neurons yielded distinct clusters between L3 and L5. CellChat[8] and synaptic gene ontology analysis of differentially expressed genes within these clusters revealed unique cell communication pathways and patterns within this cluster. In rodents, L4 cells communicate directly with the thalamus; a unique communication pathway compared to other M1 excitatory neuron layers[9]. Based on the communication pathway distinctness and transcriptomic similarity to L3 and L5, these cells appear to point towards further evidence of existence of L4-like cells within primates.

Finally, the gene ontology analysis of human-specific cell types revealed the presence of pathways associated with autism and ADHD. Clinical diagnoses of autism and ADHD are currently restricted to humans, though certain animal models exhibit behaviors that resemble aspects of these conditions (Patterson, 2011). Consequently, it remains unclear to what degree these behaviors in animals share the same neurobiological underpin-nings as human autism or ADHD.

Methods

Data collection

The 10X single nucleus RNA-sequencing data was downloaded from the Brain Initiative Cell Census Network (BICCN)[9], for chimpanzees (https://assets.nemoarchive.org/dat-depxfwd), rats (https://assets.nemoarchive.org/dat-yio1gaw), and the human data is a CV3 snRNA-sequencing data previously published[1] also sourced from BICCN. Gene annotation and genome se-quences were collected from Ensembl release 111 (Rat: https://ftp.ensembl.org/pub/release-111/gtf/rattus_norvegicus/Rattus_norvegicus.mRatBN7.2.111.chr.gtf.gz, Human: https://ftp.ensembl.org/pub/release-111/gtf/homo_sapiens/Homo_sapiens.GRCh38.111.chr.gtf.gz, and Chimpanzee: https://ftp.ensembl.org/pub/release-111/gtf/pan_troglodytes/Pan_troglodytes.Pan_tro_3.0.111.chr.gtf.gz). 10x Genomics Cell Ranger v7.2.0 was used to map the raw sequencing files to the corresponding genome and transcriptome, filter the low-quality reads, and generate the cell x gene read count matrix.

Quality control and preprocessing of snRNA data

Scanpy package was used to preprocess the cell by gene read matrices[10]. Cells which expressed more than 5% ribosomal or mitochondrial genes were removed, as high mitochondrial or ribosomal gene counts are indicative of unhealthy or abnormal cells. Cells ex-pressing fewer than 500 genes (non-neuronal) or 1,000 genes (neuronal) and cells with over 10,000 genes were excluded. Genes with non-zero counts are considered as detected. In addition, XY chro-mosome genes and genes expressed in less than three cells were removed. Suspected doublets were removed with the Scrublet package[11]. Clustering identified doublet cutoff scores for ambiguous samples. The maximum sublet score was 0.3.

Initial clustering was conducted on the fil-tered cell by gene matrices. Cell read counts were normalized and log transformed. The top 2000 highly variable genes were selected using seurat_v3 then scaled to unit variance and centered. Principal component analysis (PCA) was conducted to reduce dimensions. The data was fed into the Harmony package to remove batch effects.

Clustering

The batch corrected cell by gene matrix using the top 30 principal components was used to generate a nearest neighbor graph and the leiden community detection algorithm was applied to find clusters. Clusters were evaluated against quality control criteria. Firstly, the clusters were evaluated with their silhouette score and Calinski-Harabasz Index to measure the effectiveness of clustering parameters. With the clustering parameters set, the final clusters were filtered if it has a high percentage of mitochondrial or ribosomal gene expression, or high doublet score. No sample-specific clusters were de-tected, confirming the effectiveness of batch effect removal.

Clusters were merged and categorized into three classes, GABAergic, Glutamatergic, and non-neuronal, based on the expression level of marker genes associated with these cell classes. There are no defined marker genes for chimpanzees and rats in the literature, so homologous marker genes from humans and mice respectively were applied (Bakken et al., 2021). The assumption is that the three broad cell classes are well conserved between such closely related species so there would be much overlap with marker genes. The human marker genes GAD1, SLC17A7, SV2B, and ST18 were effective for clustering chimpanzee data, and same for the mouse maker genes Gad1, Sv2b, Qki on rats. Despite this, the chimpanzee's data retained clusters that did not show a clear expression of any above human marker genes. A Wilcoxon test of these clusters of interest against the rest of the dataset revealed genes that are specifically expressed in these clusters. The Wilcoxon Rank-Sum Test is a non-parametric test that is used to compare two independent samples to determine whether their population mean ranks differ. This test is robust to outliers and skewed distribution. Based on the known expression specificity of the top genes, these clusters were grouped into one of the three classes. After all clusters were either filtered or grouped, the distribution and the average of the number of genes expressed in each cell class were calculated.

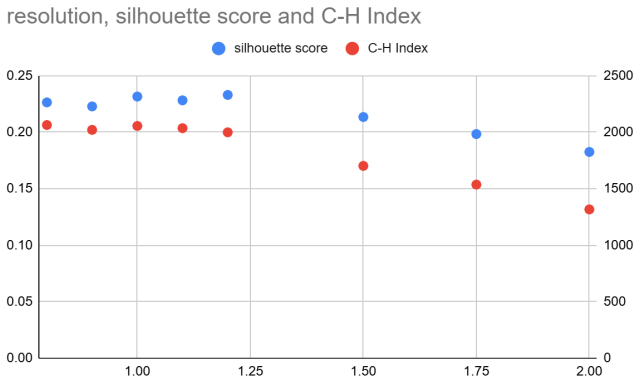

Fig. 1 Clustering Resolution Optimization. The silhouette scores and Carlinski-Harabasz Indices of various clustering resolutions were shown.

The same clustering procedure was applied to each cell class separately for both chimpanzees and rats. Evaluation of clustering metrics using silhouette score and Calinski-Harabasz Index deter-mined that a resolution of 1.0 on leiden yielded bet-ter defined clusters (Figure 1). A hierarchical clustering approach was ap-plied to the leiden clusters on the selected top 50 PCs. The leiden clusters that clustered together were grouped into cell subclasses and annotated accord-ing to the expression of the homolog genes of published human and mouse cell subclass marker genes. However, there were once again clusters that did not show clear expression of known human or mouse markers. The top 20 genes selected by the Wilcoxon test of the leiden clusters were evaluated to refine markers for chimpanzee and rat cell sub-classes. After subclasses were defined, every leiden cluster within a subclass was considered a cell type. Unlike subclasses, cell types are not universally agreed upon, but they represent the finest degree of separation possible with current techniques[1]. Cell types with less than 40 cells were ignored. Only clusters which reappeared during rounds of re-clustering at different resolutions were accepted as cell types.

Cross-species integration

To identify conserved and divergent cell subclasses and cell types across species, the individually clustered datasets for each animal were integrated into a large matrix retaining the cluster labels of each species. To maintain acceptable computing speed, the dataset was split into three, where each dataset rep-resented one cell class. The integration and clustering were conducted on both cell class and subclass levels. Orthologous genes were obtained from Ensembl 111 release[12]. Genes with one-to-one orthologs in each pair of the three species were selected for integration. The number of cells after QC filtering were similar between species (∼50,000 cells each). The raw counts were normalized and log transformed using the same approach as processing each species separately. The package Harmony was used for cross species integration on the cell class level[13]. Harmony can simultaneously account for batch effects of multiple datasets and biological factors. It starts with an initial PCA embedding and clusters the cells into multiple groups, penalizing those with low diversity. For each cluster, both a cluster centroid and dataset-specific centroids are calculated. Correction factors are determined by comparing the dataset centroids to the cluster centroid. Each cell is then corrected using a linear combination of the dataset correction factors, weighted by the clustering. This process repeats until convergence. At convergence, cells are corrected and assigned to clusters that correspond to cell classes or types rather than dataset-specific conditions. Default parameters proposed by the package were applied in this analysis. The "seurat_v3" approach was used to select 1,500 highly variable genes from the integrated data. After harmony correction of the top 50 principal components, the nearest neighbor graph was generated, and cells were clustered using the same leiden clustering approach implemented in the Scanpy package with resolution 0.1. A low resolution was applied to get a coarse clustering at the cell class level. Clusters with low quality, e.g., high mitochondrial or ribosomal expressions were removed. The composition of three species in each leiden cluster were calculated and identified rat and chimpanzee specific clusters and primate specific clusters. Previously identified cell class marker genes of each species were examined and used to annotate clusters into three cell classes. The cell proportion of each cell class in the integration results were calculated.

GABAergic cells are very well conserved across all three species, leaving little room for analy-sis, but glutamatergic cells had the most variation. The scVI[14] and scANVI[15] packages on the glutamatergic cells aided studying the conserveness and divergence of glutamatergic cells across species. scVI provided a probabilistic representation of gene expression in each cell, while scANVI aided cell type annotation and matching cell types across species. scANVI is a semi-supervised approach that builds up on the scVI model. The subclass annotations obtained from clustering were used to evaluate the consistency of subclasses between species to identify conserved and species-specific subclasses. All three species were integrated and examined, but there was a specific focus on primate specific analysis. The label conservation metrics, KMeans (Normalized mutual information) NMI and Adjusted Rand Index (ARI), were used to assess the overall conservedness of clusters overlaps across species. A leiden algorithm was applied on the nearest neighbor graph created on the scANVI derived latent space. The three species clustering was visualized using Uniform manifold approximation and projection (UMAP), and cluster overlaps were measured by NMI and ARI. Many rat specific clusters were observed and clearly separated from primates, and human and chimpanzees were mixed well. (add metric measured values, histogram of cell composition of each cluster).

Human and chimpanzee glutamatergic cells were separated and processed using scANVI. The cross-species subclass annotation of each cluster was derived based on the subclass labels that fit most of the cluster. Most clusters have well mixed cells from both species, and mostly consist of cells from a single subclass ($> 90\%$) which indicates strong conser-vation between species. Notably, a few clusters had high representation of cells from both L2_L3_IT and L5_IT, where each subclass represented greater than 30% of the cluster. These clusters were annotated as the L3_L5_IT subclass, which represented cells between the two layers, which also included L4-like cells. The UMAP of clusters were color coded by species for visualization, and cross species subclass annotations were generated using scANVI generated latent space representation. The composition of species subclass labels in each cross-species subclass were visualized. All cross-species results clustering had $> 98\%$ matching with the individual species clustering, except for the L3_L5_IT subclass. It is composed of 27% of L5_IT cells and 72% of L2_L3_IT cells. This is due to L3_5_IT not existing within individual species clustering, but when integrated, it clearly separated out from other preexisting subclasses.

Differential Gene Expression Analysis

For future categorization of cells, the defining genes for each subclass and cell type were compiled and compared to the previous analysis on human data[1], such that cell types and subclasses could be mapped consistently and compared across species. This step was important for integration, as the human data came from a different source than the rat and chimpanzee data. The paper it came from already clustered the raw data into subclasses and cell types. To identify each subclass and cell type, differentially expressed genes (DEGs) were cataloged using a one-to-all comparison. DEG analysis was applied using the R package limmatrend_cov, which accepted raw cells by gene matrices[16]. Trimmed mean of the M-values (TMM) normalization log transformation to counts per million were used to normalize the limmatrend_cov results. TMM is a normalization meth-od that estimates normalization factors by using a weighted trimmed average of log fold changes. It utilizes precision weights (the inverse of variance) to account for the fact that log fold changes derived from genes with higher read counts exhibit lower variance on the logarithmic scale. The batch information was also added to the design matrix as a confounding factor. A lack of validation techniques meant that it would be impossible to accurately tune the DEG cutoff. Since DEG was utilized to find a smaller number of defining genes for each cell type or subclass, a more stringent cutoff of log fold change > 1.2 and adj p-value < 0.05 were used to identify candidates from the DEG list.

Cell Type Annotation

The visualization of clusters generated from the scANVI model of primate data shows that the conservation in glutamatergic cell types between species is not clear. To compound matters, with the same cutoffs and procedure to find marker genes, some cell types lacked marker genes that matched those cutoffs. Since human cell types were far more published and annotated than the other species, those human annotations were used to classify the chimpanzee data. However, due to the granularity of cell class level clustering, there could be some significant gaps in this approach with species specific clusters. Extreme gradient boosting (XGboost)[17] is a well-known supervised machine learning method, and was adapted in cell annotations using the pack-age devCellpy[18]. XGboost uses a set of gradient boosted decision trees where weights of features/genes contributed to each class allow for automated identification of the marker genes. Furthermore, the reg-ularized model and sparsity aware algorithm makes XGboost resistant to overfitting and missing data to a degree[19]. Hyperparameters for XGboost were tuned by split-ting the human data into training and testing sets and finding a set of parameters that produced the best results. The resulting set of hyperparameters used in the analysis is 'eta': 0.2, 'max_depth': 6, 'subsample': 0.5, 'colsample_bytree': 0.5, 'eval_metric': 'mirror', 'seed': 840.

Applying devCellPy using known cell type annotations on human glutamatergic data with known cell type annotations created a custom classification model. This model was used to classify the chimpanzee's data. 10-fold cross validation was used in model training and cells classified with $<50\%$ certainty was annotated as "unclassified". Unclassified cells were considered either chimpanzee specific cell types or poorly annotated cells and evaluated by gene ontology analysis downstream. To find a list of predictor genes for cell type clustering, devCellpy's incorporated SHAP algorithm ranked the top positive predictor genes.

Gene ontology and cell interaction analysis

The gene expression profiles of each cell subclass and cell types produced by DEG analysis were investigated in gene ontology (GO) enrichment. The online resource gprofiler[20] was used for GO anal-ysis. gprofiler can validate marker genes if the genes map to neuron related pathways. Furthermore, gprofiler can grant a summary of a subclass function, and this was cross-referenced with existing papers to determine the validity of the subclasses.

CellChat was used to infer major molecular interactions and signaling roles of human glutamatergic cells, based on highly expressed and differential genes. Specifically, it uses gene sets to predict likely ligand-receptor pairs and cofactors from a list of roughly 2200 human signaling molecule interactions contained in the CellChatDB. The cell by gene count matrix was normal-

ized to 10000 reads per cell and log transformed. Overexpressed genes of each cell type were identified using the Wilcox test comparing one to all, with p-value < 0.005 and fold change greater than 1.2. The overexpressed signaling pathways were identified by examining the overlap between the overexpressed genes and the ligand-receptor pairs in the database. For each cell pair, 25% truncated average expression of ligands, receptors, and cofactors of the overexpressed signaling pathways were used to derive the communication probability using the hill function. 25% truncated average expression means that the average expression of a gene in a cell group is considered zero if less than 25% of cells in that group express the gene. The Hill function represents the fraction of a receptor bound by a ligand, as a function of the ligand concentration. A Hill coefficient of 1 and the dissociation constant of 0.5 were used. The probability was assigned zero if the p-value of a permutation test was greater than 0.05. The permutation test creates one hundred random permutations of the cell groups. The probability of ligand/receptor interactions between two cell groups are calculated and compared to these 100 permutations. The number of cases where the random permutations showed a higher probability was counted, and the ratio of these occurrences to the total number of permutations (100) are used as the p-value, representing the statistical significance of the communication probability. Within the inferred cell-cell communication network, the dominant interaction sender and receivers, as well as the contribution of signals in terms of outgoing and incoming were identified by calculating the weighted outdegree and weighted indegree of each interaction, producing an interaction summary for each cell type.

Results

Conserved transcriptomic cells cross specie

From the data collected in BICCN[9] and previously published human M1 data[1], more than 235,000 cells passed quality control with a roughly equal number of cells from each species (Table 1).

Table 1 Number of nuclei included in the analysis.

	human	chimpanzee	rat
GABAergic	23992	21987	11512
Glutamatergic	48536	56322	56195
Non-neuronal	4005	7413	5396

In the collection process, the cells were enriched for neurons, resulting in more than 90% of the population being neurons. Known markers were used to label cells into three broad classes, GA-BAergic, Glutamatergic, and Non-neuronal. After finding cell clusters with the markers GAD1 for GABAergic, and SATB2, SV2B, and SLC17A7 for glutamatergic cells, the resulting clusters showed more than twice the number of glutamatergic cells than GABAergic cells across all species, with the specific ratios: humans (67% vs. 33%) to chimpan-zees (72% vs. 28%) and rats (83% vs. 17%) (Figure 2 (a)). Non-neuronal cells have the lowest number of genes detected, on average about 2000 genes per cell for all three species. Glutamatergic cells have the greatest number of genes detected, ranging from 4000 to 6500 genes per cell, with rats having the least genes per cell while humans have the greatest. Overall, humans have more genes detected in all three classes, and chimpanzees and rats have a similar number of average genes detected in GABAergic and non-neuronal classes, but chimpanzees have more genes detected in glutamatergic cells than rats, with about 1000 more genes (Figure 2 (b)).

Subclasses were further identified by unsupervised clustering of cells exclusively assigned to each class. Figure 3(a-c) shows a uniform manifold approximation and projection (UMAP) of subclasses detected in each species separately. Every subclass contains cells from each donor. Cells of the same classes and subclasses were grouped together by the transcriptional profile across species (Figure 3(d)). Consistent with previously identified[21].

GABAergic cells can be grouped into two sets: Pvalb, Sst; Sst and Chold express ADARB2; and Vip, Lamp5, and Sncg express LHX6, while the proportion of cells expressing Lhx6 in rats is relatively low (Figure 3(d)). Non-neuronal markers are not as consistent as neuronal cells between rats and the others (humans and chimpanzees). No consensus marker was identified for non-neuronal cells of all three species. In addition, NCKAP5 marks OPC, Astrocytes, and Oligo cells for both humans and chimpanzees; and MBP and SLC1A3 together marks all nonneuronal cells except Endo/peri cells in humans.

Humans and chimpanzees show similar GABAergic subclass proportions, with the exception of Vip cells. (Figure 4 (a)). Chimpanzees have significantly more Vip cells than both humans and rats. Rats have significantly more Pvalb cells but less Lamp5 cells than humans and chimpanzees. Glutamatergic cell subclass compositions tend to be more variable between species than GABAergic cells. Findings from single-cell transcriptomics suggest conserved gene expression markers among GABAergic neurons across mammalian species, reaffirming their evolutionary stability compared to the more divergent glutamatergic neuron popula-tions (Lake et al., 2016). In humans and chimpanzees, over 35% of detected glutamatergic neurons belong to superficial layers (layers 2 and 3), unlike in rats. (Figure 4 (b)). On the other hand, L6 cortico-thalamic cells and L5 extra-telencephalic cells in primates are much rarer than in rats. Humans have more non-IT (intra-telencephalic) cells than chimpanzees overall, mostly L6 corticothalamic cells and L6b cells. Unlike GABAergic neurons which have similar compositions across species, compositional variation within Glutamatergic

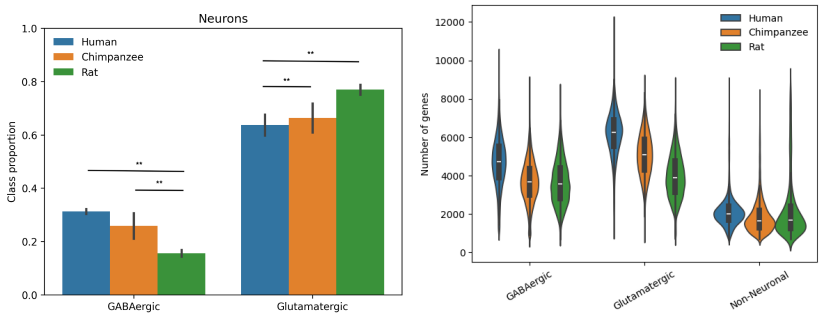

Fig. 2 Overview of cell population across three species and genes detected. (a) Relative proportions of two neuronal cell classes are significantly different between species. (b) Box plot show-ing the number of genes detected in each cell class across species. (p-value < 0.05 ($*$), p-value < 0.01 ($**$), p-value < 0.001 ($***$)).

neurons is much greater.

Integrated GABAergic and Glutamatergic cells across three species separately reveal the tran-scriptomic conservedness and diversity of sub-classes. Six GABAergic subclasses with well mixed cells across species were detected (Figure 5 (a)). Human and chimpanzee data have far more overlap with each other than with rat data (Figure 5 (b)). Comparing GABAergic and Glutamatergic neurons by clustering integrated cells using sub-class labels revealed bio-conservedness of subclasses. Four clustering comparison metrics were calculated to evaluate the bio-conseveness: normalized mutual information (NMI), ad-justed rand index (ARI), silhouette width, and isolated label (Table 2). With all metrics taken into account, GABAergic cells are more conserved than Glutamatergic cells between both mammals and primates.

Table 2 Bio-conserveness of GABAergic and Glutamatergic neu-rons between mammals and primates.

	Mammals		Primates	
	GABAergic	Glutamatergic	GABAergic	Glutamatergic
NMI	0.76	0.68	0.85	0.83
ARI	0.61	0.47	0.62	0.41
Silhouette width	0.64	0.65	0.63	0.67
Isolated label	0.8	0.74	0.71	0.86
Average	0.703	0.635	0.755	0.64

Glutamatergic neuron conserveness and diversity

Markers of glutamatergic cells are well conserved across species (Figure 6 (a)). Nine Glutamatergic subclasses with well mixed cells across species were detected. Integrated clustering of human and chim-panzee glutamatergic cells show a well-mixed cross species and clear separation between subclasses (Figure 6 (b)). CUX2 is distinguishably highly ex-pressed in more than 80% of L2/3 IT cells of all three species. The clusters that over-express RORB (known L4 marker in the mouse M1)[6] and under express L2/3 IT and L5 IT markers are annotated as L3/5 IT, potentially containing L4 IT neurons. The new subclass L3/5 IT was separated from L2/3 IT and L5 IT cells and highlighted in orange (Figure 6 (b)). It is composed of cells originally grouped into L2/3 IT and L5 IT. High CUX2 expression combined with low expres-sion of RORB can distinguish L2/3 IT cells from L3/5 IT cells. Both L3/5 IT cells and L5 IT cells express high levels of RORB in $>$80% of cells, but L3/5 IT are differentiated from L5 IT cells, by L3/5's low expression of IL1RAPL2. High ADAMTS3 and THEMIS expressions mark L6 IT and L6 IT Car3 cells in humans and chimpanzees. L6 IT Car3 cells are differentiated from L6 IT cells by their high expression of HS3ST4. Finally, there is a large difference in marker genes for L6b, as MDFIC and SEMA3D together are markers for primate L6b cells, whereas the rat L6b marker is CCN2. In differentiating layer five glutamatergic cells, L5 ET cells are distinguished by high expres-sion of both TAFA1 and BCL11B. TSHZ2 and NXPH (neurexophilin family) mark L5 NP cells, although there is slight variation in species as rats express NXPH1 while primates express NXPH2. Surprisingly, FEZF2, a previously identified marker for L5 neurons, was not uniquely expressed in any cell subclass. FOXP2 is known to be down-expressed in non IT cells in mice, also ex-pressed in human and chimpanzee L5 IT cells. Table 3 summarized Glutamatergic subclass marker genes conserved between humans and chimpanzees.

Each subclass had a different number of conserved marker genes between species, ranging from eight to 57 markers. Far more conserved markers ($>$27%) were detected between humans and chim-panzees than between either primate and rat in all seven subclasses (Figure 6 (c)). However, many markers had high expression in only one species. Non-IT cells tend to have more differentially ex-pressed genes than IT cells (Figure 6 (c-d)).

To evaluate the magnitude of difference in gene expression patterns between primates, five distance metrics, spearman dis-

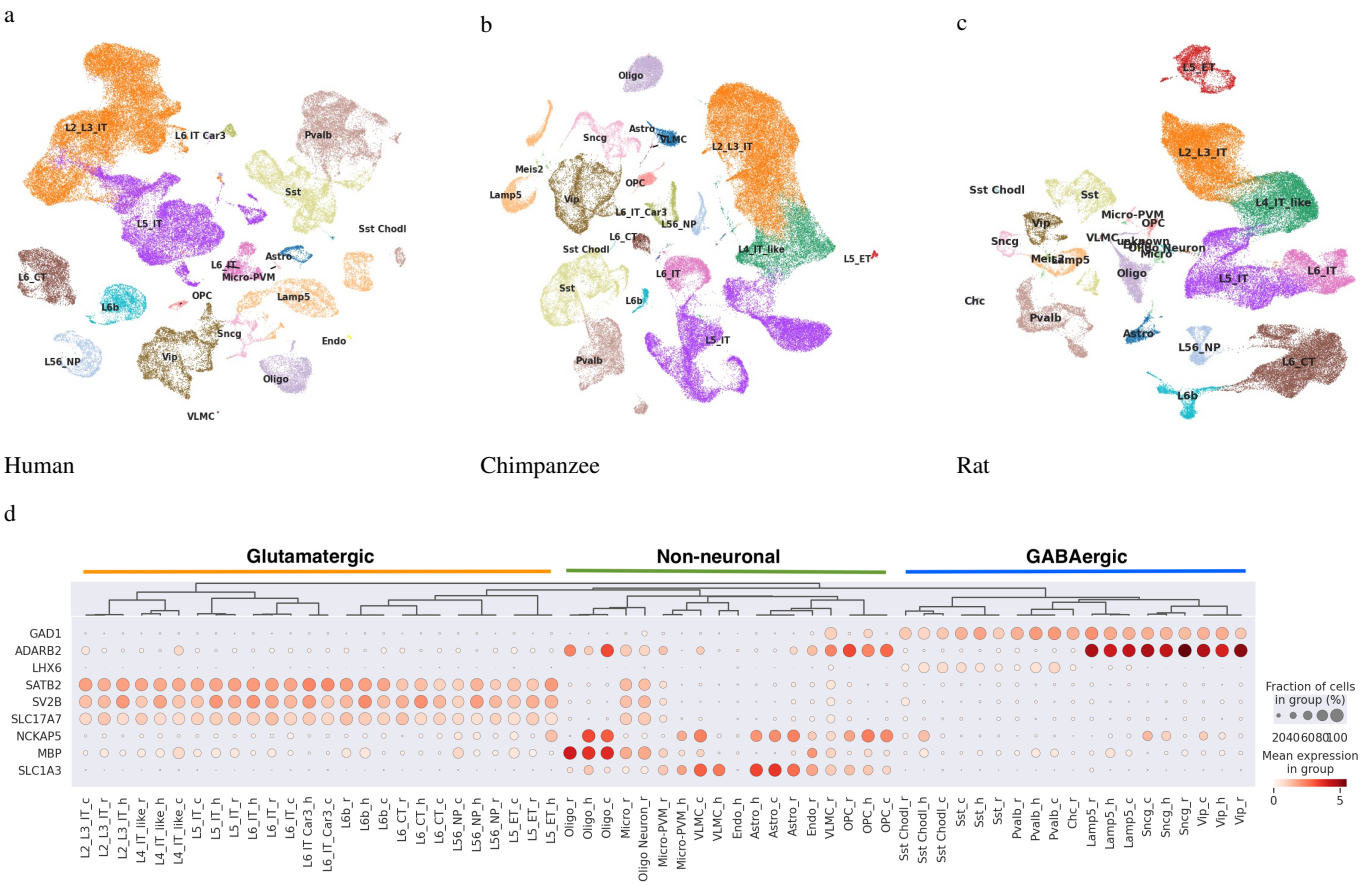

Fig. 3 Overview of subclasses and subclass marker genes in each species. (a) UMAP of all human data with subclass annotations. The colors for each subclass are consistent between graphs (a), (b), and (c). (b) UMAP of all chimpanzee data with subclass annotations. (c) UMAP of all rat data with subclass annotations. (d) Dotplot of marker genes for subclass, annotated with spe-cies and cell class.

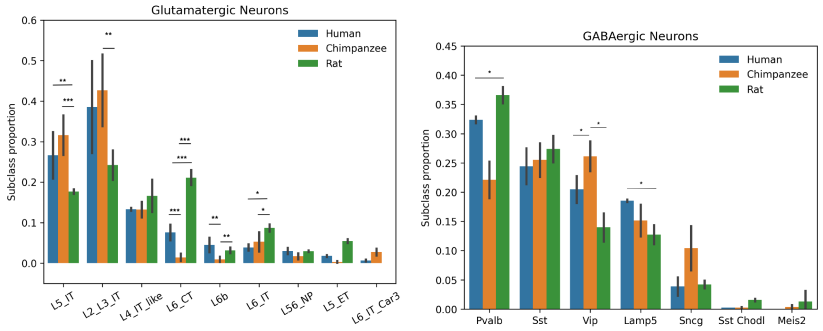

Fig. 4 Overview of cell population of subclasses. The relative proportions of GABAergic (a) and Glutamatergic (b) subclasses across three species. Car3, CAR3 gene; CT, corticothalamic cell; ET, extratelencephalic cell; IT, intratelencephalic cell; NP, near-projecting. The bar plot shows the mean+/- standard deviation across donor specimens for humans (n=2), chimpanzees (n=10), and rats (n=7). (Analysis of variance (ANOVA) followed by Tukey's HSD test shows that the difference between primate and rat is significant, p-value ¡ 0.05 (∗), p-value $<$ 0.01 (∗∗), p-value $<$ 0.001 (∗∗∗).)

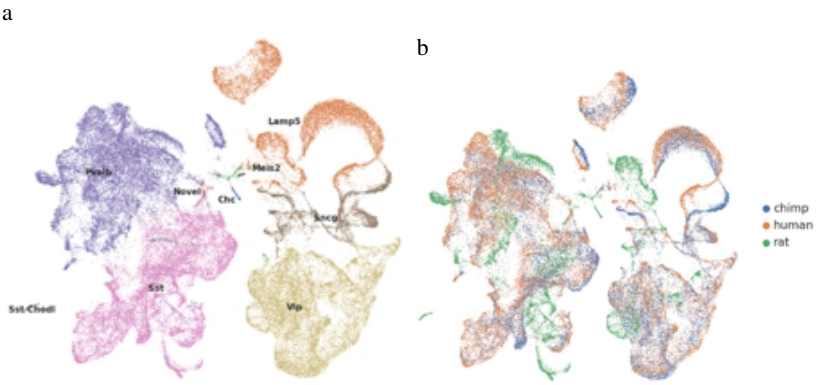

Fig. 5 UMAP of integrated GABAergic cells cross three species. (a) Color-coded by subclass, (b) color-coded by species.

tance, mean absolute error, Euclidean distance, mean squared error, and Pearson distance were calculated between the centroids of human subclasses and chimpanzee sub-classes. L6 corticothalamic cells have the lowest overall distance, followed by L5/6 near projecting cells, and L6b cells. Among the intra telencephalic cells, L3/5 IT and L6 IT have the lowest distance and are followed by L5 IT and L2/3 IT. L2/3 IT and L6 IT Car3 cells have the largest distance overall between species. Besides this broad trend conser-veness varies more between cell subclasses than physical locations (layers) (Figure 6 (e)).

35 cell types were detected in chimpanzees with higher resolution clustering, and mostly aligned with previously annotated [1] 44 human cell types (Figure 6 (f)). Hierarchical clustering of the cell types illustrated this alignment as cell types under the same subclass clustered tightly together regardless of species. However, the number of cell types detected within each subclass varies between hu-mans and chimpanzees. The small clusters under a subclass tree, formed by single species cells indicate species specific cell types.

Table 3 Marker genes for conserved primate subclasses.

Glutamatergic Cell Subclass	Genes conserved between humans and chimpanzees
.5 L3/5 IT	CUX2, RORB
L5 IT	RORB, IL1RAPL2
L6 IT	ADAMTS3, THEMIS
L6 IT Car3	ADAMTS3, THEMIS, HS3ST4
L6b	MDFIC, SEMA3D
L5 ET	TAFA1, BCL11B
L5 NP	TSHZ2, NXPH2

Layer four neurons

Further high-resolution unsupervised clustering was applied to intratelencephalic cells from the integrat-ed snRNA-seq data to identify potential L4 IT cells. Clusters were labeled with the expression pattern of known markers from humans and mice. Rat L4 cells cluster out cleanly, so clusters with an over-expression of layer four marker (RORB) were separated and annotated as L4 IT cells (Figure 7 (a)). However, there was less clear separation of human and chimpanzee L4 IT cells from L2/3 IT and L5 IT cells. Less than 100 human and chimpanzee cells were clustered with the rat L4 IT cluster. Therefore, the clusters that express both RORB and L2/3 markers were designated as L3/4 IT cells, as most of L2/3 IT did not express RORB (Figure 7 (c)), and those that expressed RORB and L5 markers were designated as L4/5 IT cells (Figure 7 (a)). L2/3 IT, L3/4 IT, and L4 IT cells have high expression of CUX2, with strong presence in $> 80\%$ of the population. L4 IT had the strongest expression of RORB and L3/4 IT had the lowest expression level among L3/4 IT, L4 IT, and L4/5 IT (Figure 7 (c)). L3/4 IT and L4 IT have a high ex-pression of GRM3, but L4/5 IT expresses GRM1 instead. Both GRM1 and GRM3 belong to the GRM gene family that are associated with the glu-tamatergic receptor complex. On the other hand, PLCH1 and TSHZ2 are exclusively highly ex-pressed in L3/4 IT and L4/5 IT respectively.

Previous studies show that cell types may cross layers [21], so the accepted division of subclasses into layers may limit the understanding of cell types in finer resolution. It is known that RORB is expressed in the bottom of layer three and upper part of layer five [22]. Any potential L4 IT cells should be contained in the L3/4 IT and L4/5 IT clusters. The separation of L4-like IT cells (consisting of the primate exclusive L3/4 IT and L4/5 IT cells) from L2/3 IT and L5 IT was evaluated using the centroid distance between clusters in the principal component space (Figure 7 (b)). The centroid distances in the space of the first five principal components provides more evidence for L4-like IT cells as the L4-like IT cell centroid is closer to L2/3 IT and L5 IT than they are to each other. The blurring of subclasses also indicates that cell types often cross between layer three and layer five.

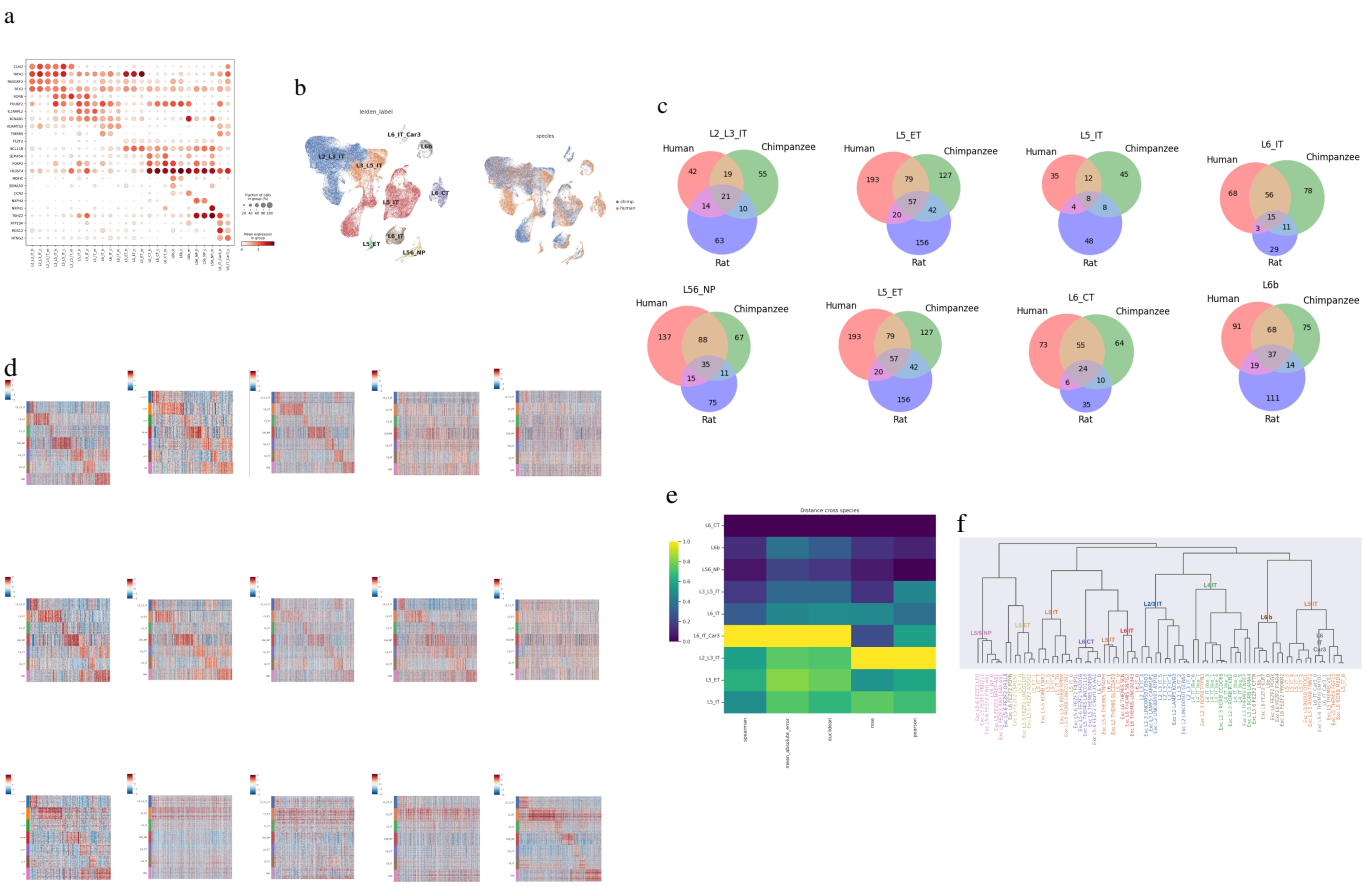

Fig. 6 Glutamatergic neuron conservation. (a) Dot plot showing the expression levels and proportions of marker genes in integrated glutamatergic neuron subclasses. (The last character of each cell subclass label denotes the species, e.g., OPC_h is human OPC cells.) Markers identified previously in the mouse and human M1 study and newly found were listed. (b) UMAP of integrated Glutamatergic neurons in primates, color coded by subclasses (left) and species (right). (c) Overlaps of subclass DEGs across three species. Hu-man and Chimpanzee share more DEGs of all subclasses. (d) Heatmap of DEGs between species, common DEGs followed by species specific DEGs. (e) Relative distance between Glutamatergic cells of humans and chimpanzees. The distance is scaled across sub-classes to make the minimum to 0 and the maximum distance to be 1.0. (f) Dendrogram showing the clustering of primate cell types of humans and chimpanzees. Cell types were color-coded by subclasses.

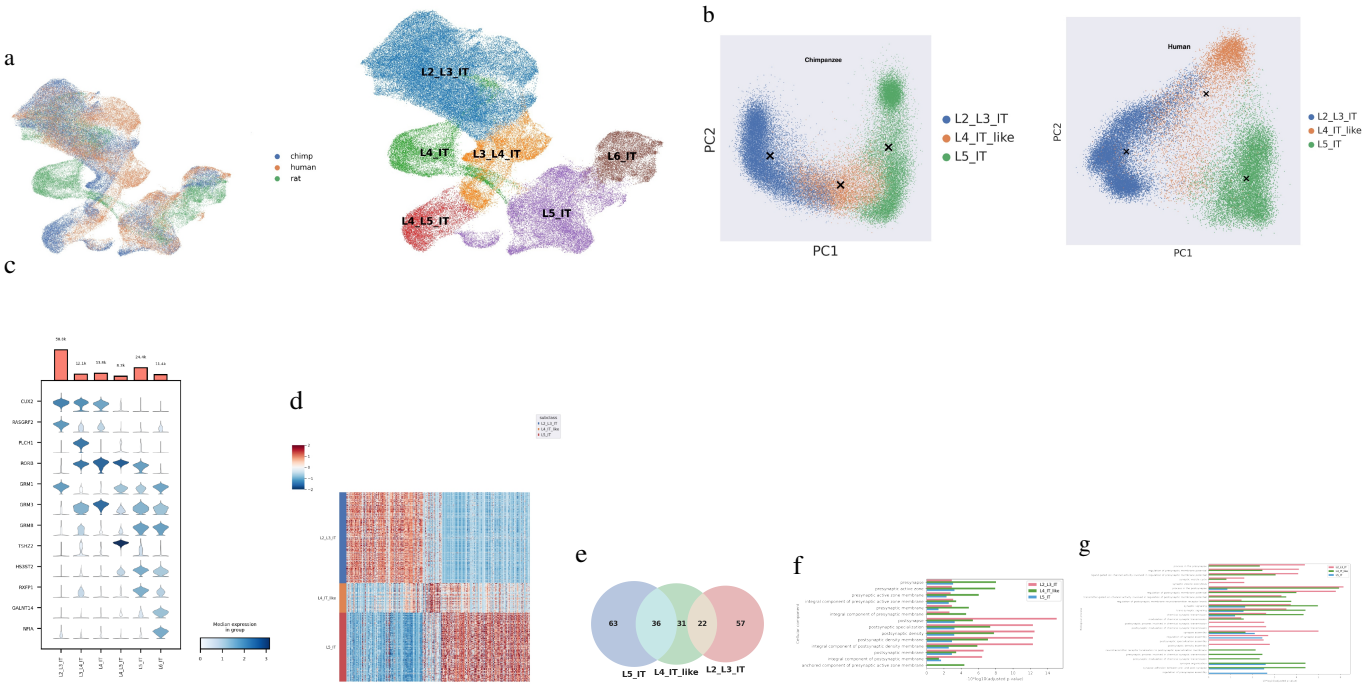

Fig. 7 Layer Four IT cells. (a) UMAP of integrated IT cells cross three species, color-coded by species (left) and subclass (right). (b) Centroid distance between each pair of L2/3 IT, L4 IT like (contains L3/4 IT, L4/L5 IT cells), and L5 IT in the principal component space for chimpanzees (left) and humans (right). (c) Stacked violin plot showing the expression pattern of marker genes. (d) Heatmap showing DEGs between L2/3 IT, L4 IT like, and L5 IT. (e) Overlap of DEGs. Synaptic Gene Ontology analysis of DEGs, (f) cellular component and (g) biological process.

The number of L4-like IT cells is less than half the number of L2/3 IT or L5 IT cells. In L4-like IT cells, 209 DEGs were detected (Figure 7 (d-e)) by pairwise differential expression analysis. About 40% and 24% of L4-like IT DEGs overlap with L5 IT and L2/3 IT DEGs respectively. Synaptic gene ontology[23] analysis of these DEGs show that L4-like IT cells are more enriched in the pre-synaptic component and process, while L2/3 IT cells contribute more to the postsynaptic component and process (Figure 7 (f-g)). It is consistent with the previous knowledge that layer 4 receives signals from the thalamus and passes to superficial layers, like layer two and three[24]. L5 IT cells are equally en-riched in both pre and post synaptic processes, with less statistical significance than L2/3 IT and L4-like IT cells. A few biological processes and cellular components were identified as enriched in L4-like IT cells only, like presynaptic processes involved in chemical synaptic transmission.

Human specific glutamatergic cell types

Hodge et al. demonstrated that while the cellular architecture of human and rodent cortices is broadly conserved, human cortical neurons exhibit unique transcriptional signatures and laminar distributions, highlighting evolutionary adaptations in upper-layer excitatory neurons and species-specific functional specializations[21]. We identified human specific cell types by using the transcriptom-ic pattern of human cell types to label chimpanzee cells (Figure 8 (a)). If a human cell type that labels less than 40 chimpanzee cells and the number of labeled chimpanzee cells are less than 10% of the human cell type size, it is determined as a human specific candidate. And the candidate cell types were further filtered by using chimpanzee cell types as classifiers to label human cells. If the candidate cells were labeled 'undefined', i.e., there are no chimpanzee cell types that have close transcriptomic pattern as the human cell type, then the human cell type was identified as potentially human specific. Four cell types in L6b, four cell types in L5 IT, five cell types in L6 CT, two cell types in L5 ET, one cell type in L6 IT Car3, one cell type in L2/3 IT, and one cell type in L5/6 NP were identified as human specific. Notably, while most human specific cell types were a few hundred cells large, one L5 IT cell type and one L2/3 IT cell type had more than 1000 cells (Figure 8 (b)).

Enriched synaptic cellular components and the biological processes of the overexpressed genes in the human specific cell types were defined using synaptic gene ontology. These genes are mostly en-riched in presynaptic and postsynaptic membranes, corresponding to organizational, presynaptic, and postsynaptic process functions (Figure 8 (c)). Fur-thermore, gene ontology with a focus on genetic disease and disorder pathways revealed that DEGs from human specific cell types had significantly higher expression of genes in pathways associated with

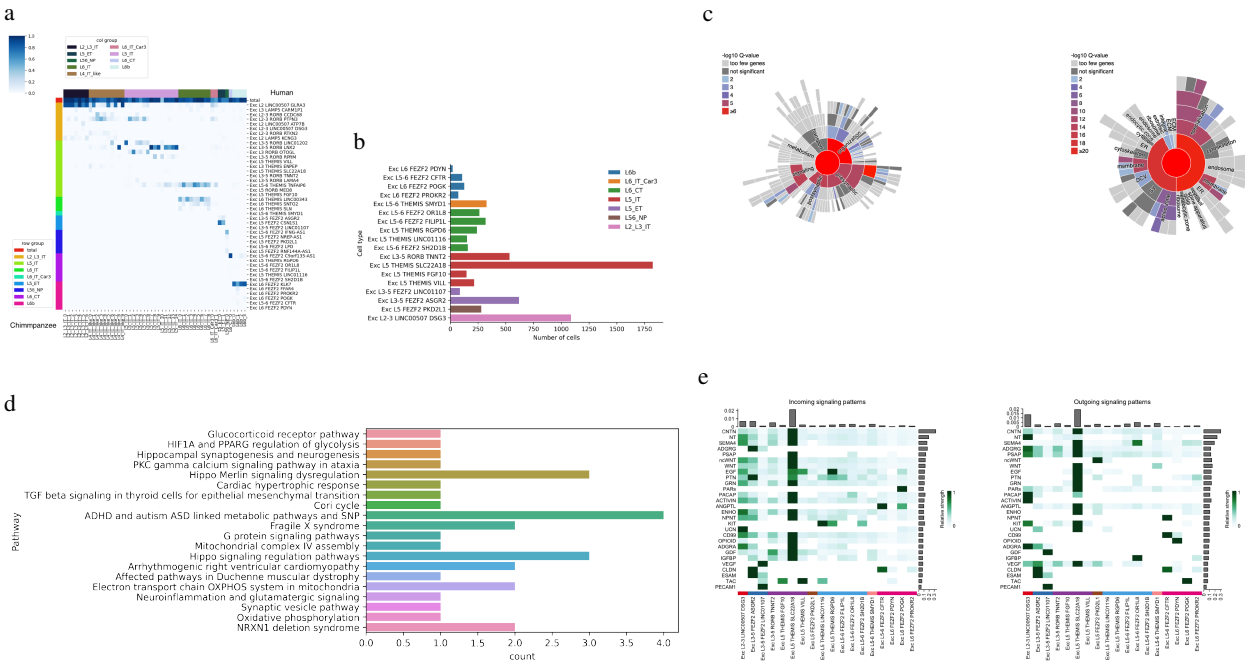

Fig. 8 Human-specific cell types. (a) Confusion matrix showing the predicted annotation of chimpanzee Glutamatergic cells using the human cell type label. (b) Number of cells in human specific cell types, colored by subclass. (c) Synaptic gene ontology analysis of DEGs from human specific cell types. (d) Significantly overexpressed disease pathways within DEGs of human specific cell types. (e) Heatmap of incoming and outgoing signaling pathways enriched by the highly expressed DEGs of human specific cell types.

autism and ADHD (Figure 8 (d)) compared to an average cell. Twenty-eight signaling pathways in the CellChat database[25] were enriched by the human-specific cell types, either as the signal sender or receiver (Figure 8 (e)). Cell types of different subclasses were enriched in dis-tinct pathways. The human-specific L2/3 IT cell type acts as a main signal sender in the PACP-PAC1 receptor pathways. Pituitary adenylate cyclase-activating polypeptide (PACAP) is known to broadly regulate the cellular stress response. A recent study shows that the PACAP-PAC1 receptor pathway plays a role in human psychological stress responses, such as post-traumatic stress disorder (PTSD). The two L5/6 ET cells were enriched in the pathway of cell adhesion molecules. It is known that cell-cell adhesions are important for brain mor-phology and underpin axon-axon contacts, linking neurons with supporting Schwann cells and oligodendrocytes.

Discussion

Summary

This analysis focused on excitatory neurons and identified over 50 excitatory, or glutamatergic(glut) cell types across humans, chimpanzees, and rats. There were also significant differences in the pro-portions of excitatory cells, with the rat sample be-ing approximately 75% excitatory, chimp being 65%, and human being about 65% as well. Further-more, as excitatory neurons move up the vertical layers of the motor cortex, they become less con-served across species. The analysis of primate excitatory neurons in this study points to the potential existence of layer four-like (L4-like) excitatory neu-rons in primates. L4 neurons are known to be found in rats but are predominantly believed not to exist in primates. Gene ontology analysis of human-specific cell types showed enrichment of certain genes previously linked to neurodevelopmental disorders, including autism and ADHD. However, gene ontology annotations are correlative, and follow-up func-tional studies would be needed to establish the roles these genes or pathways might play in such disorders.

Evidence of L4-like excitatory neurons in primates

Much of the analysis of primate excitatory neurons in this study points to the potential existence of lay-er 4-like (L4-like) excitatory neurons in primates. Analysis of primate excitatory neurons in this study found clusters of neurons that shared marker genes with both L3 and L5 cells, and clustering algorithms often struggled to split these clusters. In addition, RORB, a marker for L4 cells in rats, are overex-pressed within the cells between L3 and L5. This points to a gray zone of unclassified neurons be-tween L3 and L5, and the presence of L4 markers suggests some cells with similarity to L4 cells may be present within this

gray zone. Furthermore, a comparison of centroids of L3, L5, and L4-like neu-rons showed that the L4-like neurons appear to be a transcriptomic mix of L3 and L5 neurons. However, the DEG analysis of L4-like neurons shows that despite their relations to L3 and L5 neurons, L4-like cells express a unique set of marker genes. The L4-like neurons likely represent a subclass distinct from L3 and L5, despite retaining some similarities in gene expression. Furthermore, this unique subclass appears to have many transcriptomic similarities with known L4 cells. Looking at synaptic gene ontology, the L4-like cells tend to have more presynaptic structures as well as a couple unique presynaptic processes, whereas L2/3 neurons are predominantly postsynaptic. This pattern aligns with the established function of L4 cells, which transmit sig-nals from the thalamus to L2/3 in the motor cortex.

ADHD and Autism associated pathways in human-specific cells

Gene ontology analysis of human-specific cell types compared with shared cell types revealed the over-expression of genes implicated in pathways associ-ated with autism and ADHD. Notably, the in-volvement of the frontal cortex in neurodevelopmental disorders such as autism has been linked to early postnatal overgrowth, further supporting the idea that human-specific pathways in the motor cor-tex may underlie distinct neurodevelopmental tra-jectories[26]. Genetic studies of ADHD highlight differential heritability estimates and unique ge-nomic associations in adult populations, reflecting the complex interplay between conserved and species-specific pathways observed in primate motor cortex neurons[27]. Admit-tedly these results are preliminary and speculative, especially as the genetic causes for autism and ADHD are not fully understood[28,29]. If these results were to be validated, it would provide more evidence for the idea that autism and ADHD are human specific condi-tions. While autism and ADHD have not been diag-nosed in animals other than humans, there are behavioral patterns with similar symptoms to autism in dogs[30]. Despite observation of autism-like behaviors in these animals, these findings suggest that those behaviors may not be genetically related to autism.

Trends in glutamatergic and GABAergic conser-veness across species

The literature has established general trends in evo-lutionary conserveness between glutamatergic and GABAergic neurons, as well as within the layers of glutamatergic neurons. Firstly, it is known that across the evolution of mammals, L2 neurons grew the most, which means that the L2 layer should be least conserved[31]. This study found the same results, as L2 neurons had the farthest centroid distance between species for any glutamatergic subclass. In addition, comparing marker gene overlap, L5 and most L6 subclasses had greater overlap between species than L2/3 giving further evidence that L2 is less conserved than other subclasses. The second general conclusion is that GABAergic neurons are more conserved than glutamatergic neurons, which is also supported by pre-existing literature[32]. GABAergic neurons had notably higher bio-conservedness metrics than glutamatergic neurons, reinforcing the idea that they are more conserved. Furthermore, each subclass in GABAergic neurons can be clustered using the same marker genes for each species, whereas glutamatergic neurons require different marker genes for some rat subclasses than primate subclasses. This also shows that the sub-classes of GABAergic neurons are more closely related across species than glutamatergic subclasses. These trends are important to the validation of the methods used, as many are very recent[33], so the fact they display trends consistent with previous research is a strong indication of the validity of the techniques.

Limitations

This study was fundamentally limited by the inability to verify results with wet lab work. The data analysis conclusions require experimental verification. In addition, the study could not collect more data than what was already available on BICCN. A significant evolutionary gap exists between the tar-get species, complicating direct comparisons. While chimpanzees and humans are closely related, there exists an evolutionary chasm between those two species and rats. As a result, the results lack nuance; for certain subclasses, there is almost nothing conserved between primates and rats. However, without more intermediate species, it's impossible to tell if that subclass is primate specific or not. This study only focused on protein coding genes, using data enriched for protein coding RNA. Non-coding RNA plays significant roles in epigenetic control and cellular function[34]. In addi-tion, due to limited resources, there is no experimental validation for the putative L4-like neurons and their marker genes.

Future work

Future studies should verify these patterns experimentally and broaden the scope of analysis. In addition, spatial scRNAseq data would contextualize these findings within neuron circuits. It could also augment the transcriptomic evidence for L4 neurons in the primate M1, by locating the cells that express marker genes for L4 neurons. Another direction for future work could be similar analysis presented in this research, on a greater number of species. There is a massive evolutionary void between primates and rats, and similar analysis of a better representation of the class Mammalia could potentially reveal a greater number of genetic pathways as they changed throughout evolutionary

history.

Finally, access to greater computational power could increase the amount of data available. Although the roughly 250,000 cells in this study are standard for the technology, there still were several cell types that were thrown out on account of having too few cells to perform meaningful analysis. With many more cells, some of these smaller cell types could potentially be validated and analyzed.

References

1 T. E. Bakken, N. L. Jorstad, Q. Hu, B. B. Lake, W. Tian, B. E. Kalmbach, M. Crow, R. D. Hodge, F. M. Krienen, S. A. Sorensen, J. Eggermont, Z. Yao, B. D. Aevermann, A. I. Aldridge, A. Bartlett, D. Bertagnolli, T. Casper, R. G. Castanon, K. Crichton and E. S. Lein, *Nature*, 2021, **598**, 111–119.

2 R. G. Burciu and D. E. Vaillancourt, *Movement Disorders*, 2018, **33**, 1688–1699.

3 C. L. Ebbesen and M. Brecht, *Nature Reviews Neuroscience*, 2017, **18**, 694–705.

4 J. Cousineau, V. Plateau, J. Baufreton and M. L. Bon-Jégo, *Neurobiology of Disease*, 2022, **167**, 105674.

5 R. Sakate, N. Osada, M. Hida, S. Sugano, I. Hayasaka, N. Shimohira, S. Yanagi, Y. Suto, K. Hashimoto and M. Hirai, *Genome Research*, 2003, **13**, 1022–1026.

6 N. Yamawaki, K. Borges, B. A. Suter, K. D. Harris and G. M. G. Shepherd, *eLife*, 2014, **3**, e05422.

7 Z. Yao, H. Liu, F. Xie, S. Fischer, R. S. Adkins, A. I. Aldridge, S. A. Ament, A. Bartlett, M. M. Behrens, K. Van den Berge, D. Bertagnolli, H. R. de Bézieux, T. Biancalani, A. S. Booeshaghi, H. C. Bravo, T. Casper, C. Colantuoni, J. Crabtree, H. Creasy and E. A. … Mukamel, *Nature*, 2021, **598**, 103–110.

8 P. H. Patterson, *Pediatric Research*, 2011, **69**, 34R–40R.

9 M. Hawrylycz, M. E. Martone, G. A. Ascoli, J. G. Bjaalie, H.-W. Dong, S. S. Ghosh, J. Gillis, R. Hertzano, D. R. Haynor, P. R. Hof, Y. Kim, E. Lein, Y. Liu, J. A. Miller, P. P. Mitra, E. Mukamel, L. Ng, D. Osumi-Sutherland, H. Peng and B. Zingg, *PLoS Biology*, 2023, **21**, e3002133.

10 F. A. Wolf, P. Angerer and F. J. Theis, *Genome Biology*, 2018, **19**, 15.

11 S. L. Wolock, R. Lopez and A. M. Klein, *Cell Systems*, 2019, **8**, 281–291.e9.

12 P. W. Harrison, M. R. Amode, O. Austine-Orimoloye, A. G. Azov, M. Barba, I. Barnes, A. Becker, R. Bennett, A. Berry, J. Bhai, S. K. Bhurji, S. Boddu, P. R. Branco Lins, L. Brooks, S. B. Ramaraju, L. I. Campbell, M. C. Martinez, M. Charkhchi, K. Chougule and A. D. … Yates, *Nucleic Acids Research*, 2023, **52**, D891–D899.

13 I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. Loh and S. Raychaudhuri, *Nature Methods*, 2019, **16**, 1289–1296.

14 A. Gayoso, R. Lopez, G. Xing, P. Boyeau, V. Valiollah Pour Amiri, J. Hong, K. Wu, M. Jayasuriya, E. Mehlman, M. Langevin, Y. Liu, J. Samaran, G. Misrachi, A. Nazaret, O. Clivio, C. Xu, T. Ashuach, M. Gabitto, M. Lotfollahi and N. Yosef, *Nature Biotechnology*, 2022, **40**, 163–166.

15 C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan and N. Yosef, *Molecular Systems Biology*, 2021, **17**, e9620.

16 H. C. T. Nguyen, B. Baik, S. Yoon, T. Park and D. Nam, *Nature Communications*, 2023, **14**, 1570.

17 T. Chen and C. Guestrin, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

18 F. X. Galdos, S. Xu, W. R. Goodyer, L. Duan, Y. V. Huang, S. Lee, H. Zhu, C. Lee, N. Wei, D. Lee and S. M. Wu, *Nature Communications*, 2022, **13**, 5271.

19 G. Saygili and B. OzgodeYigin, *Scientific Reports*, 2023, **13**, 15286.

20 J. Reimand, M. Kull, H. Peterson, J. Hansen and J. Vilo, *Nucleic Acids Research*, 2007, **35**, W193–W200.

21 R. D. Hodge, T. E. Bakken, J. A. Miller, K. A. Smith, E. R. Barkan, L. T. Graybuck, J. L. Close, B. Long, N. Johansen, O. Penn, Z. Yao, J. Eggermont, T. Höllt, B. P. Levi, S. I. Shehata, B. Aevermann, A. Beller, D. Bertagnolli, K. Brouner and E. S. Lein, *Nature*, 2019, **573**, 61–68.

22 T. M. Goralski, L. Meyerdirk, L. Breton, L. Brasseur, K. Kurgat, D. DeWeerd, L. Turner, K. Becker, M. Adams, D. J. Newhouse and M. X. Henderson, *Nature Communications*, 2024, **15**, 2642.

23 F. Koopmans, P. van Nierop, M. Andres-Alonso, A. Byrnes, T. Cijsouw, M. P. Coba, L. N. Cornelisse, R. J. Farrell, H. L. Goldschmidt, D. P. Howrigan, N. K. Hussain, C. Imig, A. P. H. de Jong, H. Jung, M. Kohansalnodehi, B. Kramarz, N. Lipstein, R. C. Lovering, H. MacGillavry and M. … Verhage, *Neuron*, 2019, **103**, 217–234.e4.

24 M. García-Cabezas and H. Barbas, *The European Journal of Neuroscience*, 2014, **39**, 1824–1834.

25 S. Jin, M. V. Plikus and Q. Nie, *bioRxiv*, 2023.

26 A. P. A. Donovan and M. A. Basson, *Journal of Anatomy*, 2017, **230**, 4–15.

27 B. Franke, S. V. Faraone, P. Asherson, J. Buitelaar, C. H. D. Bau, J. A. Ramos-Quiroga, E. Mick, E. H. Grevet, S. Johansson, J. Haavik, K.-P. Lesch, B. Cormand and A. Reif, *Molecular Psychiatry*, 2012, **17**, 960–987.

28 A. Thapar and E. Stergiakouli, *Xin Li Xue Bao. Acta Psychologica Sinica*, 2008, **40**, 1088–1098.

29 A. Genovese and M. G. Butler, *Genes*, 2023, **14**, year.

30 K. Tiira, O. Hakosalo, L. Kareinen, A. Thomas, A. Hielm-Björkman, C. Escriou, P. Arnold and H. Lohi, *PLOS ONE*, 2012, **7**, e41684.

31 P. Vanderhaeghen and F. Polleux, *Nature Reviews Neuroscience*, 2023, **24**, 213–232.

32 W. G. Pembroke, C. L. Hartl and D. H. Geschwind, *Genome Biology*, 2021, **22**, 52.

33 P. V. Kharchenko, *Nature Methods*, 2021, **18**, 835–835.

34 M. U. Kaikkonen, M. T. Y. Lam and C. K. Glass, *Cardiovascular Research*, 2011, **90**, 430–440.