Transcriptomic Evolution of Neuronal Cell Classes and Cell Types in Human, Chimpanzee, and Rat

Leo Dai
Los Gatos High School, Los Gatos, CA, 95032, USA
*Correspondence should be addressed to (leod91791@gmail.com)

The primary (M1) motor cortex is a highly conserved part of the mammalian brain, responsible for the direct control of voluntary muscles. While it is known that different species have differently structured motor cortices due to their unique needs and evolution, there is little knowledge about the cross-species differences between motor cortex cell types at the molecular level, and no comparison of humans and our closest living relatives, chimpanzees. A cross-species molecular level understanding of the M1 cortex would reveal the development of specific cell types and genes in line with evolutionary progression in motor skills. Using single nucleus RNA sequencing (snRNA-seq) on samples gathered from humans, chimpanzees, and rats, this research identified over 50 excitatory neuron cell types in three species, along with differences in celltype/subclass proportions, marker genes, and some potentially unidentified cell types. These results suggest the existence of Layer 4-like excitatory neurons in primates. Additionaly, differentially overexpressed genes of human specific cell types were enriched in pathways implicated in ADHD, autism, and other human diseases.
Keyword: snRNA-seq, human, motor cortex, neuron, excitatory

Introduction

First mapped in 1874, the motor cortex has since been the subject of continuous study. Today, it is known that the motor cortex can be divided into the primary motor cortex (M1), premotor cortex, and supplementary motor cortex. Responsible for controlling voluntary movement, the human M1 cortex is organized topographically; each body part has a corresponding location on the M1 cortex, while also being composed of six vertical layers of neurons. Recent advances in brain mapping technology have unveiled portions of neuron circuits in the human M1 cortex responsible for coordinating complex muscle movements like speech and running.

However, a new technology called Single Nucleus RNA Sequencing (snRNA-seq), is revolutionizing tissue mapping by giving precise RNA reads of hundreds of thousands of single cells in a tissue sample. This provides data that can be used to identify unique cell types by gene expression. Given spatial data, snRNA-seq can generate a map of cell types over the tissue, providing a higher resolution of organ structure and gene expression than possible with previous techniques.

By applying snRNA-seq to human M1 cortices, and our closest relatives; chimpanzees, this paper addresses a gap in knowledge between the cell types and gene expression of human and chimpanzee M1 cortices. The overall genetic sequence identity between human and chimpanzee is 98.77%[1]. If comparing the coding DNA, the similarity is 99%, meaning the genetic code is nearly identical. Focusing on the pinpoint differences and their implications on cell function will help identify uniquely human genetic features and diseases.

This analysis focused on excitatory neurons and identified over 50 excitatory (Exc), or glutamatergic (glut) cell types across humans, chimpanzees, and rats. There were also significant differences in the proportions of excitatory cells, with the rat sample being approximately 75% excitatory, chimp being 65%, and human being about 65% as well. 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.

Much of the analysis of primate excitatory neurons in this study points to the potential existence of layer 4-like (L4-like) excitatory neurons in primates. L4 neurons are known to be found in rats, but are

predominantly believed not to exist in primates[2]. However, recent studies have identified visual evidence of unique cell type combinations in the region between L3 and L5 in mammals[3]. 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, Cellchat [4] and 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[5]. 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.

Additionally, gene ontology analysis of human-specific cell types revealed the presence of pathways associated with autism and ADHD. Both autism and ADHD appear to be largely human-specific phenomena, although similar behavior patterns manifest in other animals[6].

Results

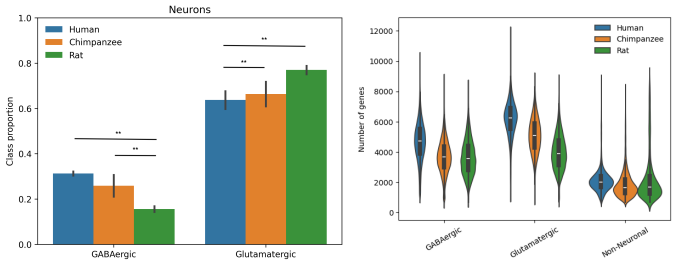
Conserved transcriptomic cells cross specie

From the data collected in BICCN [7] and previously published human M1 data[8], more than 235,000 cells passed quality control with a roughly equal number of cells from each species (Table 1). In the collection process the cells were enriched for neurons, resulting in more than 90% of the population being neurons. Cells were labeled by known markers of three broad classes, GABAergic, 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 chimpanzees (72% vs. 28%) and rats (83% vs. 17%) (Figure 1 (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 1 (b)).

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

Table 1. Number of nuclei included in the analysis.

For each class, unsupervised clustering of the snRNA-seq data identified subclasses (Method section) (Figure 2 a-c). Every subclass contains cells from each donor. Cells of the same classes were grouped together by the transcriptional profile across species (Figure 2 (d)). Non-neuronal markers are not as consistent as neuronal cells between rats and the others (humans and chimpanzees). Consistent with previously identified[9], GABAergic cells can be grouped into two sets: Pvalb, Sst, and Sst Chold express ADARB2, and Vip, Lamp5, and Sncg express LHX6. While the proportion of cells expressing Lhx6 in rats is relatively low, no consensus marker was identified for non-neuronal cells of all three species. NCKAP5 marks OPC, Astrocytes, and Oligo cells for both human and chimpanzees. MBP and SLC1A3 together marks all non-neuronal cells except Endo/peri cells in humans. The proportion of GABAergic subclasses between humans and chimpanzees are similar, except Vip. (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



(a) (b)

Figure 1: Overview of cell population across three species and genes detected. (a) Relative proportions of two neuronal cell classes are significantly different between species (a). The bar plot shows the mean+/- standard deviation cross donor specimens for humans (n=2), chimpanzees (n=10), and rats (n=10). (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 (***). **(b)** Box plot showing the number of genes detected in each cell class across species.

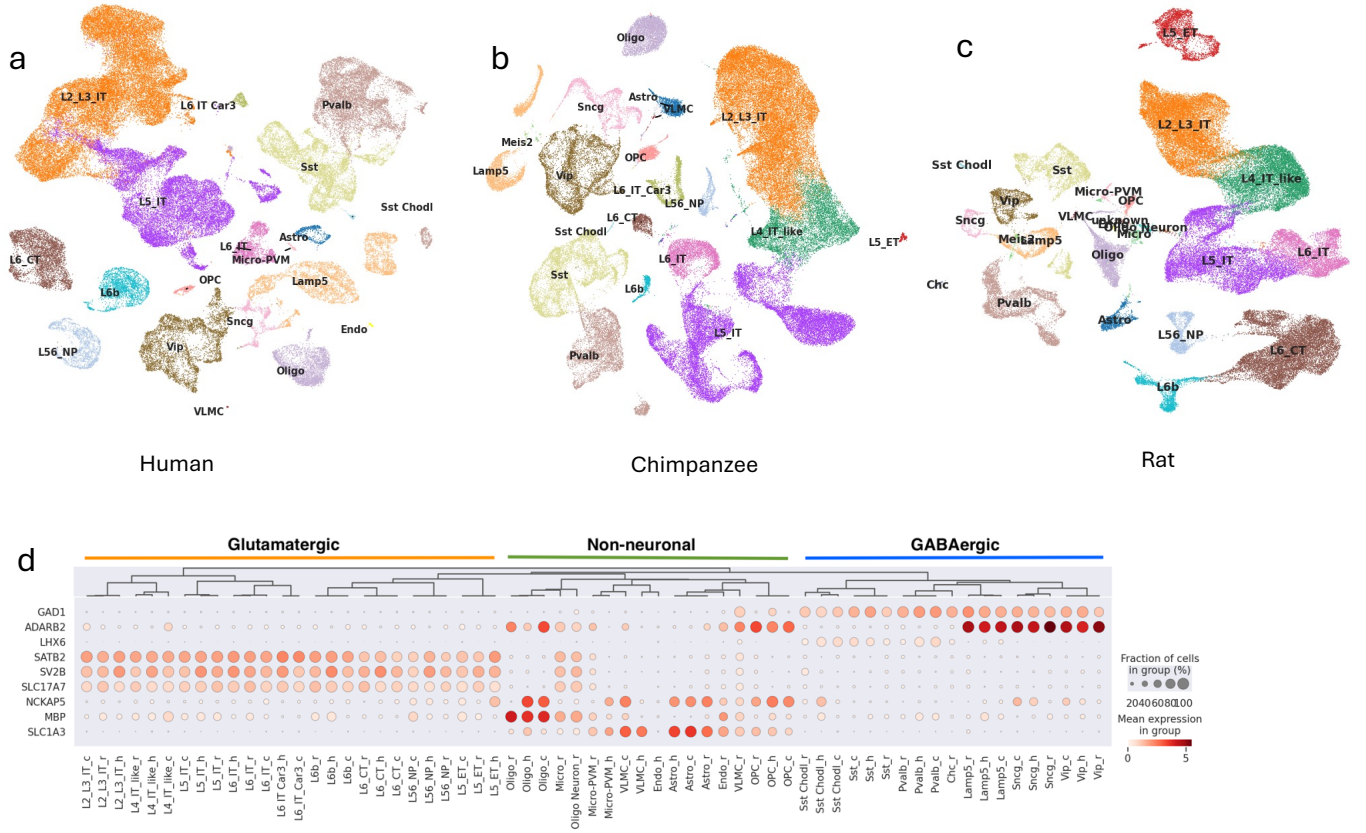


Figure 2: Overview of subclasses and subclass marker genes in each species. (a) UMAP of all human data with subclass annotations. The colors for each subclass are consistent between graphs (a), (b), and (c). **(b)** UMAP of all chimpanzee data with subclass annotations. **(c)** UMAP of all rat data with subclass annotations. **(d)** Dotplot of marker genes for subclass, annotated with species and cell class.

to be more variable between species than GABAergic cells. Superficial layer (layer two and three) excitatory neurons are dominant among the detected glutamatergic neurons, >35% in both human and chimpanzees, but not rats. (Figure 4 (b)). On the other hand, L6 corticothalamic 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 neurons is much greater.

	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

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

Integrated GABAergic and Glutamatergic cells across three species separately reveal the transcriptomic conservedness and diversity of subclasses. Six GABAergic subclasses with well mixed cells across species were detected (Figure 3 (a)). Human and chimpanzee data has far more overlap with each other than with rat data (Figure 3 (b)). Comparing GABAergic and Glutamatergic neurons by clustering integrated cells using subclass labels revealed bio-conservedness of subclasses. Four clustering comparison metrics were calculated, normalized mutual information (NMI), adjusted 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.

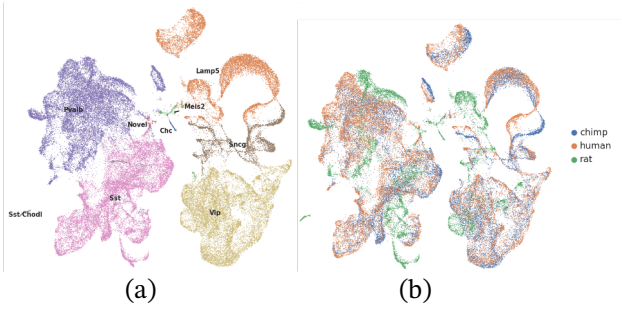
Glutamatergic neuron conserveness and diversity

Markers of glutamatergic cells are well conserved across species (Figure 5 (a)). Nine Glutamatergic

subclasses with well mixed cells across species were detected. Integrated clustering of human and chimpanzee glutamatergic cells show a well-mixed cross species and clear separation between subclasses (Figure 5 (b)). CUX2 is distinguishably highly expressed 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[3]) 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 orange (Figure 5 (b)). It is composed of cells originally grouped into L2/3 IT and L5 IT. High CUX2 expression combined with low expression 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. In categorizing L6 neurons, the presence of high ADAMTS3 and THEMIS are markers of 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 expression 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 expressed in human and chimpanzee L5 IT cells.

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 chimpanzees than between either primate and rat in all seven subclasses (Figure 5 (c)). However, many markers had high expression in only one species. Non-IT cells tend to have more differentially expressed genes than IT cells (Figure 5 (c-d)).

To evaluate the magnitude of difference in gene expression patterns between primates, five distance metrics, spearman distance, mean absolute error, Euclidean distance, mean squared error, and Pearson distance were calculated between the centroids of human subclasses and chimpanzee subclasses. L6

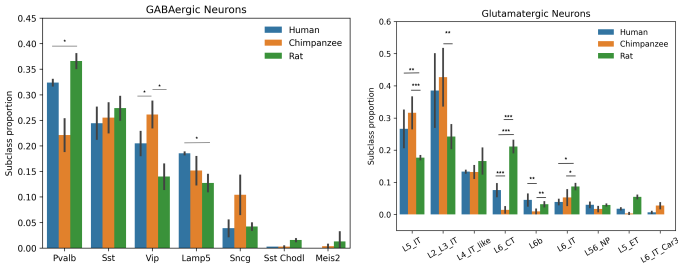


(a) (b)

Figure 3: UMAP of integrated GABAergic cells cross three species. **(a)** Color-coded by subclass, **(b)** color-coded by species.

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, conservedness varies more between cell subclasses than physical locations (layers) (Figure 5 (e)).

35 cell types were detected in chimpanzees with higher resolution clustering, and mostly aligned with previously annotated [8] 44 human cell types (Figure 5 (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 humans and



(a) (b)

Figure 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 cross 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 (***).

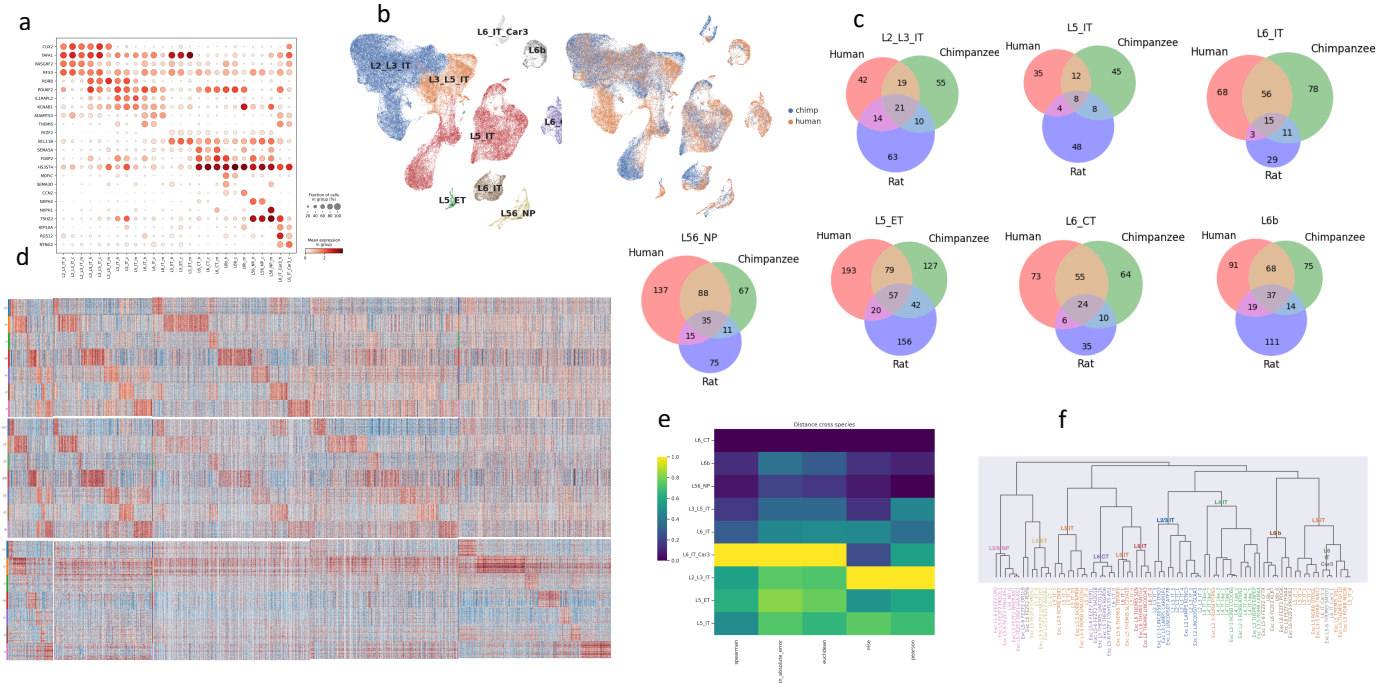


Figure 5: Glutamatergic neuron conservation. (a) Dot plot showing the expression levels and proportions of marker genes in integrated glutamatergic neuron subclasses. (The last character of each cell subclass label denotes the species, e.g., OPC_h is human OPC cells.) Markers identified previously in 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. Human 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 subclasses 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.

chimpanzees. The small clusters under a subclass tree, formed by single species cells indicate species specific cell types.

Layer four neurons

Further high-resolution unsupervised clustering was applied to intratelencephalic cells from the integrated snRNA-seq data to identify potential L4 IT cells. Clusters were labeled with the expression pattern of known markers from rodents. 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 6 (a)). However, there was less clear separation of human and chimpanzee potential 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 6 (c)), and those that expressed RORB and L5 markers were designated as L4/5 IT cells (Figure 6 (a)). It is known that RORB is expressed in the bottom of layer three and upper part of layer five[10]. Any potential L4 IT cells should be contained in the L3/4 IT and L4/5 IT clusters.

The separation of L4 IT like cells (consisting of the primate exclusive L3/4 IT and L4/5 IT cells) from L2/3 IT and L5 IT was evaluated using centroid distance (Figure 6 (b)). The centroid distance in the space of principal components one and two shows a clear separation, where the L4 like centroid is essentially right between the L2/3 and L5 centroids. The blurring of subclasses also indicates that cell types often cross between layer three to layer five, appearing in all those layers.

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 6 (c)). L3/4 IT and L4 IT have a different expression pattern of GRM family genes and DSCAML1 (involved in neuronal differentiation). GRM genes are associated with the glutamatergic receptor complex. L4 IT cells, especially in rat data, have an over-expression of GRM3, but very low expression of GRM1, GRM8, and DSCAML1, whereas L3/4 IT cells have an over-expression of GRM1, GRM3, GRM8, and DSCAML1. L4/5 IT cells are different as they highly express IL1RAPL2, an L5 IT marker.

The number of L4 IT like cells is less than half the number of L2/3 IT or L5 IT cells. In L4 IT like cells, 209 DEGs were detected (Figure 6 (d-e)) by pairwise differential expression analysis. There is no

overlap between these L4 like DEGs and the DEGs identified in L5 IT and L2/3 IT, while about 40% and 24% of L4 IT DEGs overlap with L5 IT and L2/3 IT respectively. Synaptic gene ontology analysis of these DEGs show that L4 IT like cells are more enriched in the presynaptic component and process, while L2/3 IT highly expressed genes contribute more to the postsynaptic component and process (Figure 6 (f-g)). L5 IT cells are equally enriched in both pre and post synaptic processes, with less statistical significance than L2/3 IT and L4 IT like cells. A few biological processes were identified as enriched in L4 IT cells only, like presynaptic processes involved in chemical synaptic transmission.

Human specific glutamatergic cell types

Human specific cell types were defined using the transcriptomic pattern of human cell types to label chimpanzee cells (Figure 7 (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 was determined as human specific candidates. 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 is no chimpanzee cell types that have close transcriptomic pattern as the human cell type, then the human cell type was identified as 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 7 (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 [11]. These genes are mostly enriched in presynaptic and postsynaptic membranes, corresponding to organizational, presynaptic, and postsynaptic process functions (Figure 7 (c)). Furthermore, gene ontology with a focus on genetic disease pathways revealed that DEGs from human specific cell types had significantly higher expression of genes in pathways associated with autism and ADHD (Figure 7 (d)) compared to an average cell. Twenty-eight signaling pathways in the Cellchat database[4] were enriched by the human-specific cell types, either as the signal sender or receiver (Figure 7 (e)). Cell types of different subclasses were

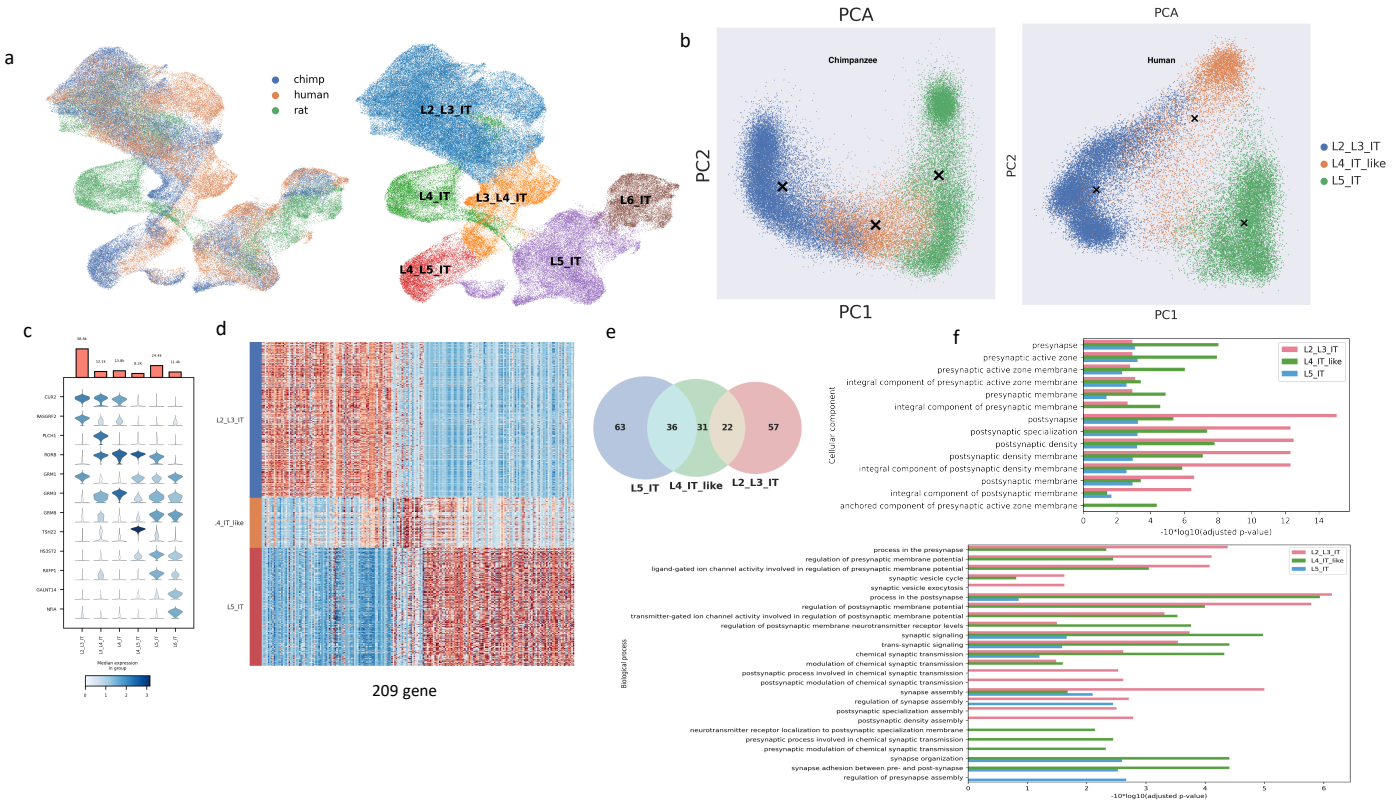


Figure 6: Layer Four IT cells. (a) UMAP of integrated IT cells cross three species, color-coded by species (left) and subclass (right). **(b)** Centroid distance between each pair of L2/3 IT, L4 IT like (contains L3/4 IT, L4/L5 IT cells), and L5 IT in the principal component space for chimpanzees (left) 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.

enriched in distinct 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 morphology 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 proportions of excitatory cells, with the rat sample being approximately 75% excitatory, chimp being 65%, and human being about 65% as well. Furthermore, as excitatory neurons move up the vertical layers of the motor cortex, they become less conserved across species. Much of the analysis of primate excitatory neurons in this study points to the potential existence of layer four-like (L4-like) excitatory neurons in primates. L4 neurons are known to be found in rats, but are predominantly believed not to exist in primates. Additionally, gene ontology analysis of human-specific cell types revealed the presence of pathways associated with autism and ADHD.

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 layer 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 overexpressed within the cells between L3 and L5. This points to a gray zone of unclassified neurons between 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 neurons 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. These results suggest that the L4-like neurons appear to be a subclass distinct from either L3 or 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

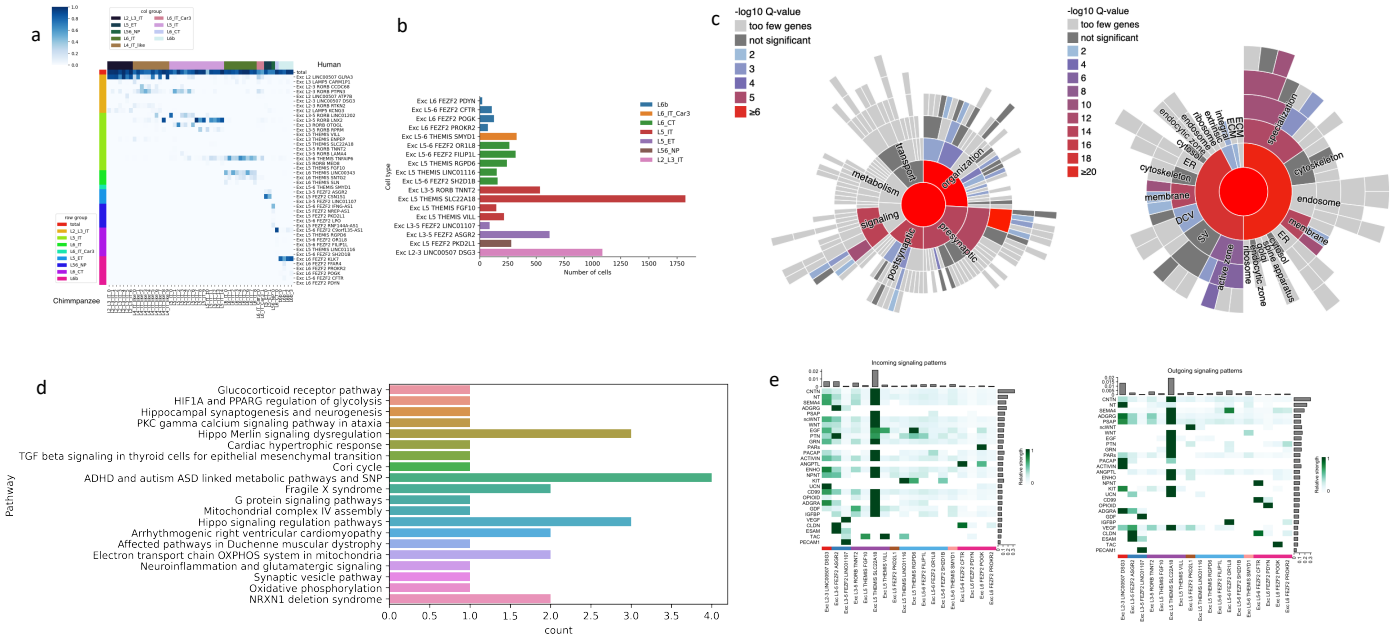


Figure 7: Human-specific cell types. (a) Confusion matrix showing the predicted annotation of chimpanzee Glutamatergic cells using the human cell type label. **(b)** Number of cells in human specific cell types, colored by subclass. **(c)** Synaptic gene ontology analysis of DEGs from human specific 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.

presynaptic structures as well as a couple unique presynaptic processes, whereas L2/3 neurons are predominantly postsynaptic. This pattern fits the known function of L4 cells, that they transmit signals from the thalamus to L2/3 in the motor cortex. In addition, Cellchat analysis of differentially expressed genes of L4 revealed unique cell communication pathways and patterns within this cluster. Based on the communication pathway distinctness and transcriptomic similarity to rodent L4, as well as L3 and L5, these cells appear to point towards further evidence of existence of L4-like cells within primates.

ADHD and Autism associated pathways in human-specific cells

Gene ontology analysis of human-specific cell types compared with shared cell types revealed the overexpression of genes implicated in pathways associated with autism and ADHD. Admittedly these results are preliminary and speculative, especially as the genetic causes for autism and ADHD are not fully understood[12 13]. In addition, gene ontology analysis is typically very rough, and does have an element of randomness present. If these results were to be validated, it would provide more evidence for the idea that autism and ADHD are human specific conditions. While autism and ADHD have not been diagnosed in animals other than humans, there are behavioral patterns with similar symptoms to autism in dogs[14]. 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 conservedness across species

The literature has established general trends in evolutionary conservedness 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[15]. 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[16]. 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 subclasses 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[17], 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 conclusions drawn from the data analysis could not be definitively verified. In addition, the study could not collect more data than what was already available on BICCN.

Regarding the data itself, there is a significant gap between the target species. 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. The data also featured significant batch effects. While tools like Harmony can reduce the impact of batch effects, on finer clustering resolutions like those used for cell classes, batch effects can still seriously impact clustering. Without using batch effect correction, even cell class clustering was seriously impacted, demonstrating the strength of batch effects within the data.

Using Google Colab, there were limitations on the methods available for data processing. Due to RAM limitations, SAMAP[18], which other studies found were ideal for this dataset, could not be used.

Future work

Future studies ideally verify the patterns and hypotheses presented in this study with lab work or expand the scope of analysis presented here. In addition, access to spatial scRNAseq data would allow a contextualization of the patterns found in this study with actual neuron circuits. It could also augment the transcriptomic evidence for L4 neurons in the primate M1 cortex, 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 fairly 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. Sakate, R. *et al.* Analysis of 5'-end sequences of chimpanzee cDNAs. *Genome Res* **13**, 1022–1026 (2003).
2. Shipp, S., Adams, R. A. & Friston, K. J. Reflections on agranular architecture: predictive coding in the motor cortex. *Trends in Neurosciences* **36**, 706–716 (2013).
3. Yamawaki, N., Borges, K., Suter, B. A., Harris, K. D. & Shepherd, G. M. G. A genuine layer 4 in motor cortex with prototypical synaptic circuit connectivity. *Elife* **3**, e05422 (2014).
4. Jin, S., Plikus, M. V. & Nie, Q. CellChat for systematic analysis of cell-cell communication from single-cell and spatially resolved transcriptomics. *bioRxiv* (2023) doi:10.1101/2023.11.05.565674.
5. Bopp, R., Holler-Rickauer, S., Martin, K. A. C. & Schuhknecht, G. F. P. An Ultrastructural Study of the Thalamic Input to Layer 4 of Primary Motor and Primary Somatosensory Cortex in the Mouse. *J Neurosci* **37**, 2435–2448 (2017).
6. Patterson, P. H. Modeling autistic features in animals. *Pediatr Res* **69**, 34R-40R (2011).
7. Hawrylycz, M. *et al.* A guide to the BRAIN Initiative Cell Census Network data ecosystem. *PLoS Biol* **21**, e3002133 (2023).
8. Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111–119 (2021).
9. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
10. Goralski, T. M. *et al.* Spatial transcriptomics reveals molecular dysfunction associated with cortical Lewy pathology. *Nature Communications* **15**, 2642 (2024).
11. Koopmans, F. *et al.* SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron* **103**, 217-234.e4 (2019).
12. Thapar, A. & Stergiakouli, E. An Overview on the Genetics of ADHD. *Xin Li Xue Bao* **40**, 1088–1098 (2008).
13. Genovese, A. & Butler, M. G. The Autism Spectrum: Behavioral, Psychiatric and Genetic Associations. *Genes (Basel)* **14**, (2023).
14. Tiira, K. *et al.* Environmental effects on compulsive tail chasing in dogs. *PLoS One* **7**, e41684 (2012).
15. Vanderhaeghen, P. & Polleux, F. Developmental mechanisms underlying the evolution of human cortical circuits. *Nature Reviews Neuroscience* **24**, 213–232 (2023).
16. Pembroke, W. G., Hartl, C. L. & Geschwind, D. H. Evolutionary conservation and divergence of the human brain transcriptome. *Genome Biol* **22**, 52 (2021).
17. Kharchenko, P. V. Publisher Correction: The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods* **18**, 835–835 (2021).
18. Tarashansky, A. J. *et al.* Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *eLife* **10**, e66747 (2021).
19. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).
20. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281-291.e9 (2019).
21. Harrison, P. W. *et al.* Ensembl 2024. *Nucleic Acids Research* **52**, D891–D899 (2023).
22. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **16**, 1289–1296 (2019).
23. Gayoso, A. *et al.* A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology* **40**, 163–166 (2022).
24. Nguyen, H. C. T., Baik, B., Yoon, S., Park, T. & Nam, D. Benchmarking integration of single-cell differential expression. *Nature Communications* **14**, 1570 (2023).
25. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, New York, NY, USA, 2016). doi:10.1145/2939672.2939785.
26. Galdos, F. X. *et al.* devCellPy is a machine learning-enabled pipeline for automated annotation of complex multilayered single-cell transcriptomic data. *Nature Communications* **13**, 5271 (2022).
27. Saygili, G. & OzgodeYigin, B. Continual learning approaches for single cell RNA sequencing data. *Scientific Reports* **13**, 15286 (2023).

Acknowledgements

Thank you to Salwan Butrus for your advice and guidance throughout this research. Thanks to the BICCN and BICCN production laboratories.

Code Availability

The source code and libraries used in this manuscript are available at: https://github.com/Leo-d-D/primary_mortor_cortex_M1.

Materials and Methods

Data collection

The 10X single nucleus RNA-sequencing data was downloaded from the Brain Initiative Cell Census Network (BICCN) [7], 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 [8] also sourced from BICCN. Gene annotation and genome sequences 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 (Ensembl release 111): 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 scRNA data

Scanpy package was used to preprocess the cell by gene read matrices [19]. 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 which expressed below 500 (non-neuronal cells) or 1,000 (neuronal cells), or greater than 10,000 genes were removed as well. In addition, XY chromosome genes and genes expressed in less than three cells were removed. Suspected doublets were removed with the Scrublet package[20]. Clustering was used to determine doublet cutoff score for ambiguous samples. The maximum sublet score was 0.3.

Initial clustering was conducted on the filtered cell by gene matrices. The read count of each cell was normalized and log transformed. The top 2000 highly variable genes were selected using seruat_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. Each cluster was evaluated based on the quality control criteria. A cluster was filtered if it has a high percentage of mitochondrial or ribosomal gene expression, or high doublet score. No sample specific cluster was detected, which confirmed the batch effect removal was effective.

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[8]. 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. 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.

The same clustering procedure was applied to each cell class separately for both chimpanzees and rats. A higher resolution on leiden yielded finer clusters. Then a hierarchical clustering approach was applied to the leiden clusters on the selected top 50 PCs. The leiden clusters that clustered together were grouped into cell subclasses and annotated according 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 subclasses. 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[8]. Cell types with less than 40 cells were ignored. Only clusters which reappeared during rounds of reclustering 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 represented one cell class. The integration and clustering were conducted on both cell class and subclass levels. Ortholog genes were downloaded from Ensembl 111 release[21]. 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[22]. 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 analysis, but glutamatergic cells had the most variation. The scVI[23] and scANVI 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 ran through scANVI. The cross species subclass annotation of each cluster was derived based on the subclass labels that fit the majority of the cluster. Majority of clusters have well mixed cells from both species, and mostly consist of cells from a single subclass (> 90%) which indicates strong conservation 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 Expression analysis for marker genes

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[8], 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 cell by gene matrices[24]. Trimmed mean of the M-values (TMM) normalization log transformation to counts per million were used to normalize the limmatrend_cov results. The batch information was also added to the design matrix as a confounding factor. From the resulting DEG list, candidates were identified if the log fold change was > 1.4 and the adjusted p-value was < 0.05.

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 (XGboost25) is a well-known supervised machine learning method, and was adapted in cell annotations using the package devCellpy[26]. 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 regularized model and sparsity aware algorithm makes XGboost resistant to overfitting and missing data to a degree[27].

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 were annotated as "unclassified". Unclassified cells were denoted as chimpanzee specific cell types. To get the 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 was used for GO analysis. While the results of GO enrichment are not granular enough to paint a cohesive image of the cell type function, 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.

To understand the intercellular communications of primate glutamatergic cells, CellChat can infer the major molecule interactions and the major signaling roles of each cell type 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 were used. The cell by gene count matrix was normalized 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. The probability was assigned zero if the p-value of a permutation test was greater than 0.05. 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.