Supplementary Information for

Neuron-specific coding sequences are highly conserved despite large neuronal cell type and morphological diversity in mammalian evolution

Linhe Xu, Suzana Herculano-Houzel

Suzana Herculano-Houzel
Email: suzana.herculano@vanderbilt.edu

This file includes:

Supplementary text
Figures S1 to S5
Tables S1 to S4
Legends for Datasets S1 to S7
SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S7

Supplementary Information Text

Detailed Methods

Materials. Each analysis was carried out either on the ACCRE cluster at Vanderbilt University or on a 2016 MacBook Pro (8 GB memory).

Selection of Brain Cell Type-specific Genes. The expression level of 22,458 genes in purified astrocytes, neurons, oligodendrocyte precursor cells (OPCs), newly formed oligodendrocytes (NFOs), myelinating oligodendrocytes (MOs), microglia, and endothelial cells from the Barres Lab's 2014 cell type enrichment study(1) was downloaded from https://web.stanford.edu/group/barres_lab/brain_rnaseq.html[1]. Out of the 22,458 genes with expression data, 19,388 are protein-coding genes with one-to-one ortholog in at least one of the 92 mammalian species (*SI Appendix*, Dataset S4) available on Ensembl 98 (now archived at http://sep2019.archive.ensembl.org/index.html). Each gene's expression level in oligodendrocytes is then defined as the mean of its expression level in OPCs, NFOs, and Mos. Each gene's expression level in glial cells is defined as the mean of its expression level in astrocytes, oligodendrocytes, and microglia. Genes with expression level in one cell type greater than the sum of expression level in the other four cell types are defined as cell type-specific genes. For example, if a gene's expression level in neuron is greater than the sum of its expression level in astrocytes, microglia, oligodendrocytes, and endothelial cells, this gene is considered a neuron-specific gene in the mouse brain. Since this threshold is the same as four times the average expression level across four other cell types, a similar definition is carried into defining glial cell-specific genes as genes with expression level greater than four times its average expression level in neurons and endothelial cells. This gives us 1,298 neuron-specific genes, 1,211 glia-specific genes, 1,062 microglia-specific genes, 937 endothelia-specific genes, 824 astrocyte-specific genes, and 521 oligodendrocyte-specific genes.

Selection of Brain Cell Type-expressed Genes. Similar to brain cell type-specific genes, cell type-expressed genes were selected based on the Barres mouse brain cell type enrichment data. Genes with expression level (FPKM) in one cell type greater than 1 are defined as cell type-expressed genes. Only the genes that are protein-coding genes were kept. This gives us 10,028 neuron-specific genes, 10,359 glia-specific genes, 8,458 microglia-specific genes, 8,973 endothelia-specific genes, 9,634 astrocyte-specific genes, and 9,861 oligodendrocyte-specific genes. Unlike cell type-specific gene lists, these cell type-expressed gene lists have great overlaps.

Selection of Benchmark Genes

ATPase. Human ATPases were first retrieved from the HUGO Gene Nomenclature Committee database (2) (https://www.genenames.org/data/genegroup/#!/group/412) on October 10th, 2019. Mouse orthologs of these genes were then found in Ensembl 98 with the BioMart portal.

Housekeeping Genes. Housekeeping genes in this study come from an paper on human housekeeping genes (3) (https://www.tau.ac.il/~elieis/HKG/). Mouse homologs of these genes were then searched on Ensembl 98 with the BioMart portal.

Immune Genes. Immune genes come from the InnateDB database (4) (https://www.innatedb.com/annotatedGenes.do?type=innatedb), accessed on Jan 22, 2020. This list includes genes from multiple species, and a shorter mouse immune gene list is generated from it as benchmark genes.

---

[1] That site is no longer accessible and the interactive portal was transferred to a new site http://www.brainrnaseq.org/. We included the original downloaded Excel file (*SI Appendix*, Dataset S2) for the convenience of replication.

MHC. A list of MHC genes were retrieved from a review article (5) and matched to Ensembl 98 mouse reference genome.

Organ-specific Genes. Organ-specific genes are genes with expression level only in that organ but not anywhere else based on the MGI mouse gene expression database (GXD) (6). Data retrieved on October 3rd, 2019. Liver-specific genes are genes where expression is detected in liver (TS16-28) and not detected anywhere else. Kidney-specific genes are genes detected in metanephros (TS18-28) but not anywhere else. Lung-specific genes are genes detected in lung (TS15-28) but not anywhere else. Skin-specific genes are genes detected in skin (TS20-28) but not anywhere else. Brain-specific genes are genes detected in brain (TS17-28) but not anywhere else. Heart-specific genes are genes detected in heart (TS11-28) but not anywhere else. Pancreas-specific genes are genes detected in pancreas (TS20-28) but not anywhere else. Musculature-specific genes are genes detected in musculature (TS12-28) but not anywhere else. The MGI IDs of these genes can be found at https://github.com/VeritatemAmo/neuron-glia-dNdS/tree/master/data/MGI_organs.

Pairwise dN/dS Ratios. With the mouse reference genome (GRCm38.p6) Ensemnl 98 database has calculated pairwise dN values and dS values as part of the orthologues dataset with 92 mammalian species. These data were retrieved with a bash script (https://github.com/VeritatemAmo/neuron-glia-dNdS/blob/master/bash_scripts/all_mouse_protein_coding_dNdS.sh), including each orthologue's homology type. After orthologues that are not one-to-one paired with mouse are filtered out, dN/dS values were then calculated by simply dividing dN with dS, with dN/dS values defined as zero when both dN and dS equals zero (very high level of negative selection). When only dS is zero but not dN, the dN/dS score is dropped. For each gene, a mean dN/dS score is calculated by averaging all available pairwise dN/dS scores across species. Pairwise dN/dS and averaged dN/dS can be found in *SI Appendix*, Dataset S3 and S4. The dN/dS ratios were then matched to cell type-specific genes by name (https://github.com/VeritatemAmo/neuron-glia-dNdS/blob/master/jupyter_notebooks/mouse.celltype-specific-genes.ipynb).

Gene Ontology. Genes mapped to each GO slim term was retrieved from the Princeton University Lewis-Sigler Institute for Integrative Genomics' Generic GO Term Mapper(7) (https://go.princeton.edu/cgi-bin/GOTermMapper) on October 13th, 2019. The ontology aspects selected was "biological processing" and restricted to species "Mus musculus (MGI)". Additional "regulation of cell size", "regulation of membrane potential", "membrane depolarization", "membrane hyperpolarization", and "membrane repolarization" GO terms that is not included in the slim terms is also added because of our interest in cell size regulation and potential roles of unique properties of neuron. The genes belonged to each GO terms were then transposed and matched onto genes of interest and their dN/dS values with a python script (https://github.com/VeritatemAmo/neuron-glia-dNdS/blob/master/jupyter_notebooks/mouse.GO-celltype-contingency.ipynb).

phastCons of 2,000 bp upstream promoter region. Mouse protein-coding genes' coordinates were downloaded from Ensembl 100 database for the mouse genome GRCm38.p6 (https://github.com/VeritatemAmo/neuron-glia-dNdS/blob/master/bash_scripts/mouse_protein_coding_gene_positions.sh). Genes not on the autosomes, sex chromosomes, or mitochondria genome were filtered out. Then the start and end positions of the 2,000 upstream regions were identified for each cell type-specific gene (https://github.com/VeritatemAmo/neuron-glia-dNdS/blob/master/bash_scripts/parse_mouse_promoter_coord.sh, https://github.com/VeritatemAmo/neuron-glia-dNdS/blob/master/jupyter_notebooks/mouse.gene_position.ipynb).These coordinates were then used to retrieve precalculated phastCons scores (phastCons60way) from the UCSC Genome Browser. For each cell type-specific gene, an average score for all 2,000 base pairs was calculated. Distribution of these phastCons were checked to be not normally distributed, therefore

non-parametric analysis (Mann-Whitney test and Kruskal-Wallis test) were used for further analysis (https://github.com/VeritatemAmo/neuron-glia-dNdS/blob/master/jupyter_notebooks/mouse.phastCons_stats.ipynb).
The 60 vertebrate species included in phastCons60way can be found at https://genome.ucsc.edu/cgi-bin/hgTables .

Statistical Analysis

Confidence Interval. Because dN/dS are log-normally distributed, we chose to use median-based description of dN/dS values. Confidence interval is based on ranked dN/dS values of each gene group. A binomial interval was first fitted to the number of genes with the python scipy.stats.binom.interval() package, with alpha set to 0.95. After ranking the gene in each group on dN/dS value, the genes at the low and high interval threshold of each group is set as the lower and higher bound of the confidence interval (https://github.com/VeritatemAmo/neuron-glia-dNdS/blob/master/jupyter_notebooks/mouse.stats_and_figures.ipynb).

Mann-Whitney U Test. Mann-Whitney U tests were performed with the python package scipy.stats.mannwhitneyu() on pairs of gene group's dN/dS data.

Contingency Analysis. Contingency analysis was performed with a python function we wrote. Essentially a crosstable is generated with the python pandas.crosstab() function between a celltype and a gene ontology, then odds ratios were calculated with the python scipy.stats.fisher_exact() function and the χ^2 values and p values generated by the python - scipy.stats.chi2_contingency() function.

Linear Regression. Distribution normality of log-transformed phastCons and dN/dS were first confirmed with Q-Q plots. Outliers with dN/dS greater than 2 were filtered out. Then a least squares linear regression is performed based on log-transformed phastCons and dN/dS scores of neuron, endothelia, astrocyte, microglia, and oligodendrocyte-specific genes. Python script available at https://github.com/VeritatemAmo/neuron-glia-dNdS/blob/master/jupyter_notebooks/mouse.phastCons_stats.ipynb.

Fig. S1. Neuron-, glia-, and endothelial cell-specific genes' pairwise dN/dS ratios from 92 mammalian species against mouse reference genome, hued with species clades. The red squares highlight the lack of high dN/dS and the abundance of low dN/dS for neuron-specific genes.

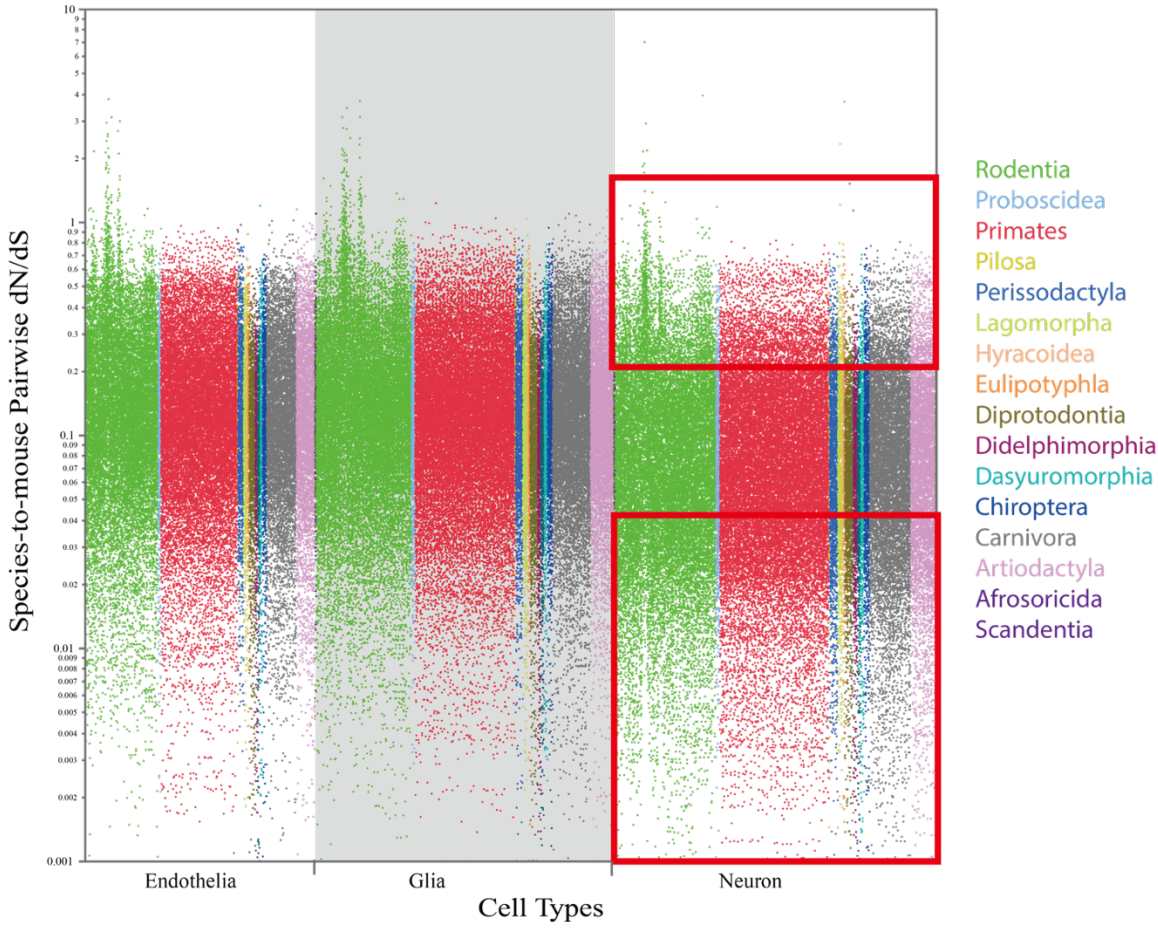

Fig. S2. With mouse reference genome, pairwise dN/dS ratios of cell type-**specific** genes across seven representative mammalian species, as well as averaged dN/dS ratios of all 92 mammalian species (whiskers showing 15th percentile to 85th percentile of each distribution). Except oligodendrocyte-specific genes' dN/dS from rat and megabat, everyone is significantly higher than neuron-specific genes. Kruskal-Wallis H test shows significant differences of the median dN/dS between neuron, glial cells, and endothelial cells (NEG) as well as between neuron, endothelial cells, astrocytes, microglial cells, and oligodendrocytes (NEAMO) for all seven representative species as well as averaged dN/dS. H statistics, U statistics, and p values from Mann-Whitney and Kruskal-Wallis tests can be found at https://github.com/VeritatemAmo/neuron-glia-dNdS/blob/master/results/celltype-specific_inferential_stats.xlsx, and visualization for each species at https://github.com/VeritatemAmo/neuron-glia-dNdS/tree/master/figures/MannWhitney/mouse_reference_genome.

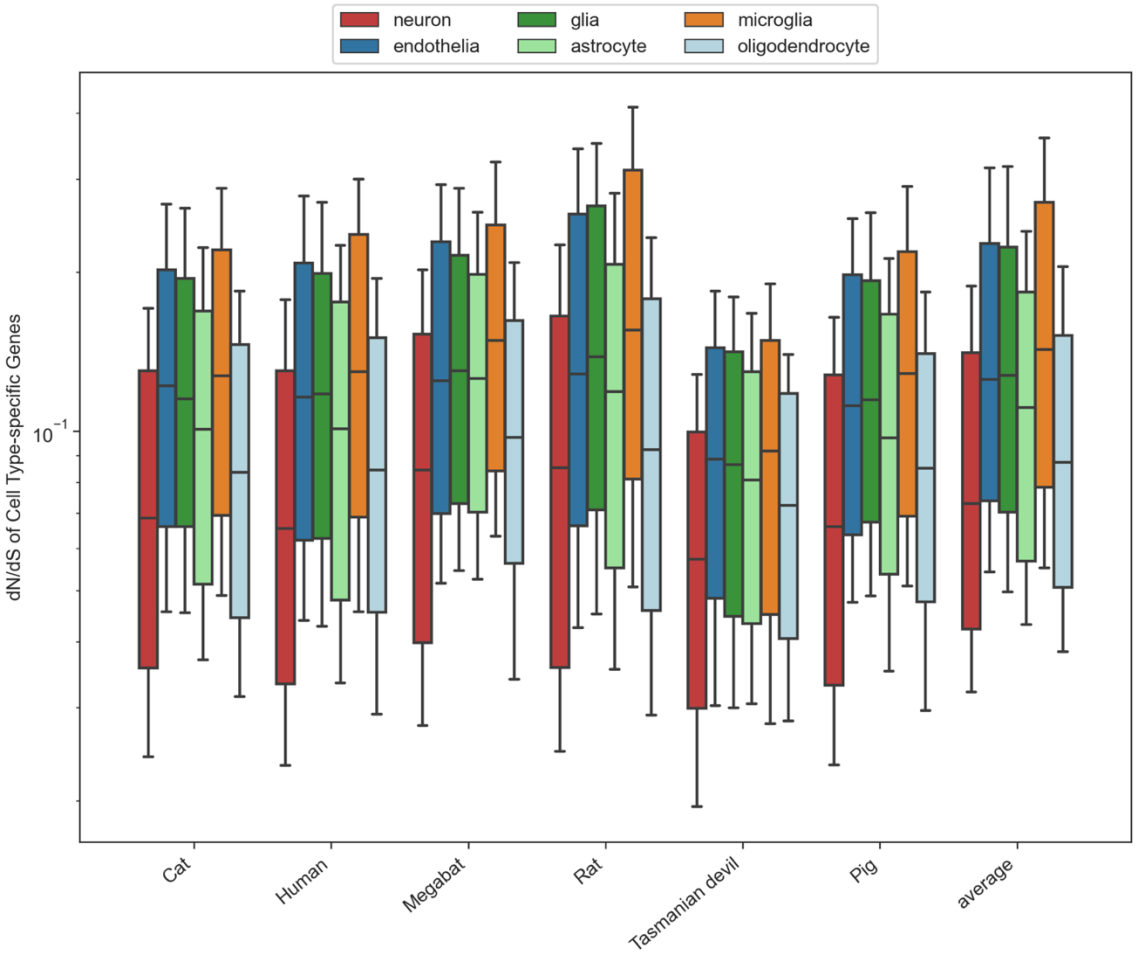

Fig. S3. Neuron, glial cell, glial subtypes, and endothelial cell-**expressed** genes' pairwise dN/dS ratios from seven representative mammalian species against mouse reference genome. In addition, dN/dS averaged from available dN/dS ratios from all 92 mammalian species against mouse are also included. Whiskers show 15th percentile to 85th percentile of distributions.

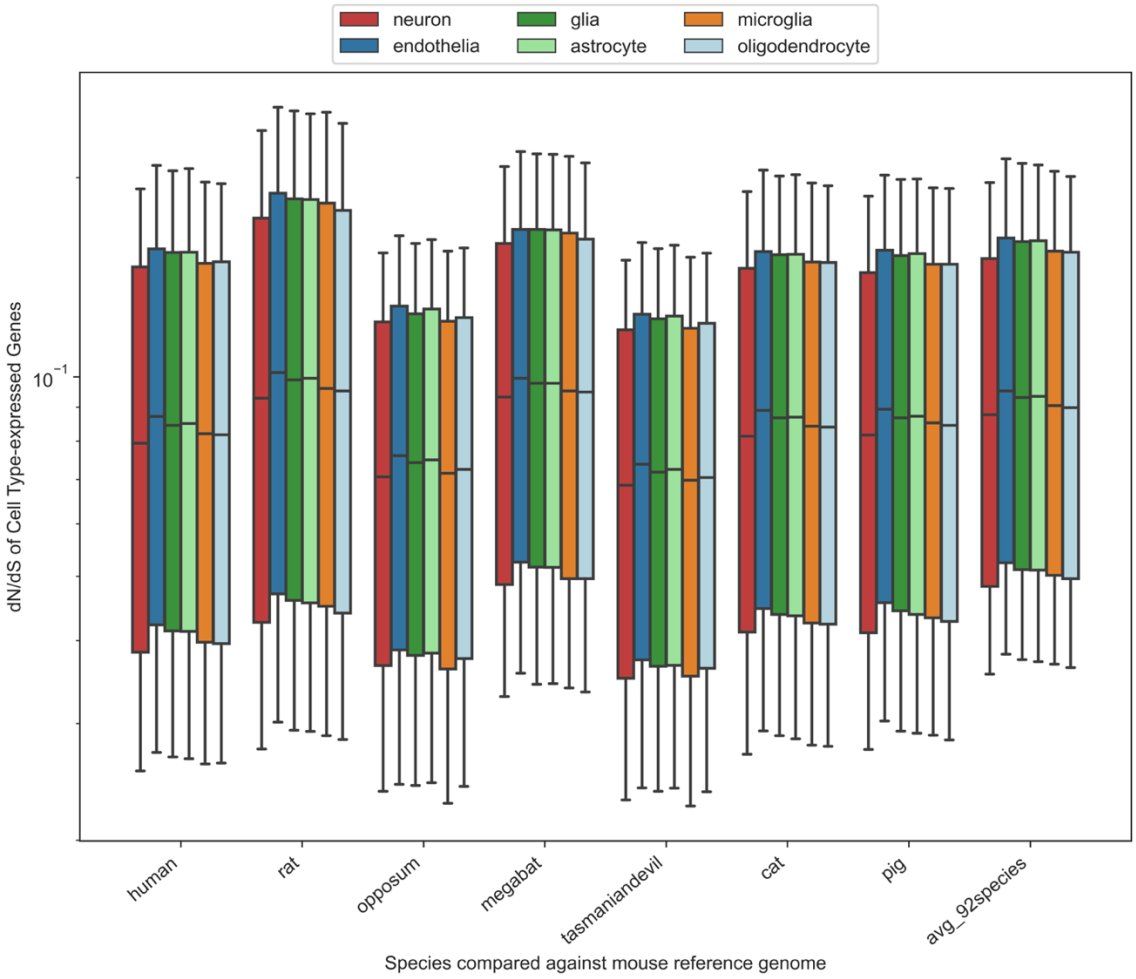

Fig. S4. PhastCons of neuron-specific genes' 2,000 bp upstream promoter region are higher than oligodendrocytes and microglia, but not astrocyte.

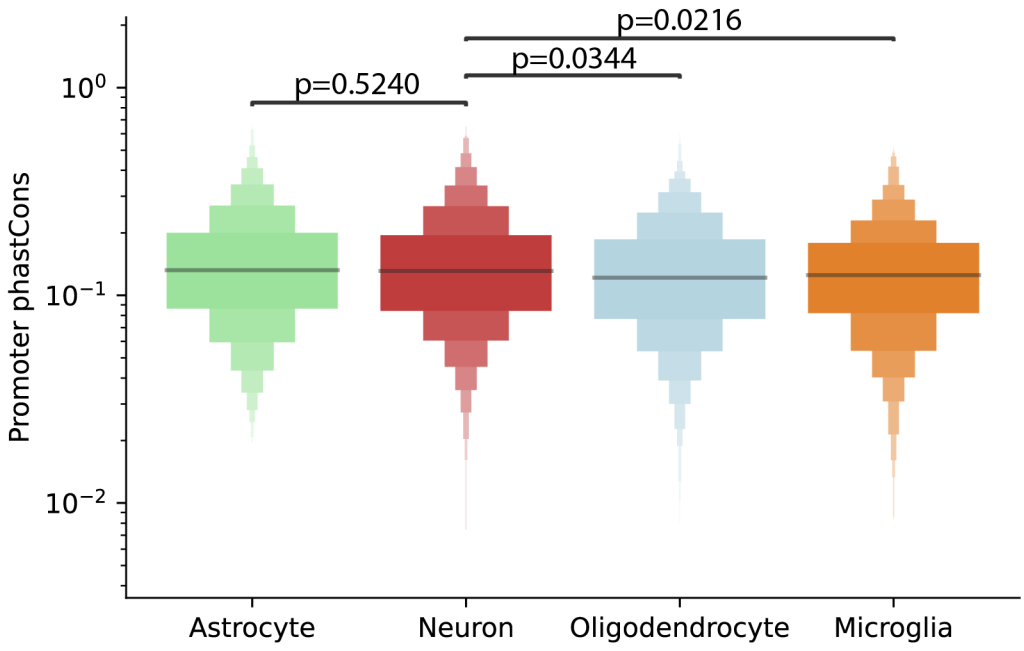

Table S1. Cell type-specific genes dN/dS ratios' inferential statistics with a few representative species. To validate the result of lower neuron-specific genes' dN/dS ratios, neuron-specific genes' dN/dS are compared to glia-specific genes' dN/dS between mouse and one of seven representative mammalian species. Median dN/dS and 95% confidence interval, as well as the Mann-Whitney U test results are reported here.

Species	*Neuron-specific Genes median dN/dS ($CI_{95\%}$)*	*Glia-specific Genes median dN/dS ($CI_{95\%}$)*	*Mann-Whitney U Test p-value*
Human	0.065 (0.061-0.070)	0.118 (0.112-0.124)	2.650×10^{-37}
Rat	0.085 (0.078-0.091)	0.138 (0.130-0.146)	1.711×10^{-29}
Opossum	0.061 (0.058-0.065)	0.085(0.080-0.090)	2.509×10^{-11}
Megabat	0.084 (0.074-0.091)	0.130 (0.118-0.141)	1.021×10^{-15}
Tasmanian Devil	0.057 (0.054-0.061)	0.087 (0.080-0.091)	2.841×10^{-18}
Cat	0.069 (0.063-0.074)	0.115 (0.109-0.124)	6.151×10^{-36}
Pig	0.066 (0.061-0.071)	0.115 (0.108-0.125)	2.286×10^{-40}

Table S2. Cell type-expressed genes dN/dS ratios' inferential statistics. Similar to cell type-specific genes' inferential statistics, median dN/dS of neuron- and glia-specific genes between mouse and seven representative species are reported along with Mann-Whitney U test between glia- and neuron-specific genes. Averaged dN/dS ratios of 92 species are also reported here.

Species	*Neuron-specific Genes median dN/dS ($CI_{95\%}$)*	*Glia-specific Genes median dN/dS ($CI_{95\%}$)*	*Mann-Whitney U Test p-value*
Human	0.079 (0.077-0.082)	0.085 (0.083-0.087)	1.94E-05
Rat	0.093 (0.090-0.095)	0.099 (0.096-0.102)	2.11E-05
Opossum	0.071 (0.069-0.073)	0.074 (0.072-0.076)	0.011
Megabat	0.093 (0.090-0.096)	0.098 (0.095-0.101)	0.007
Tasmanian devil	0.069 (0.067-0.070)	0.072 (0.070-0.074)	0.006
Cat	0.081 (0.080-0.084)	0.087 (0.085-0.089)	3.74E-05
Pig	0.082 (0.080-0.084)	0.087 (0.085-0.089)	1.68E-05
Average of 92 species	0.088 (0.086-0.090)	0.093 (0.091-0.095)	1.29E-06

Table S3. Organ-specific dN/dS compared to neuron. 95% confidence intervals are provided alongside the median dN/dS. Mann-Whitney U tests were performed and p values and common language effect sizes (CLES) are reported here.

Organ	*Median dN/dS ($CI_{95\%}$)*	*Mann-Whitney p-value Against Neuron*	*CLES*
Neuron	0.073 (0.068-0.077)	N/A	N/A
Heart	0.082 (0.063-0.116)	0.232	0.551
Brain	0.108 (0.093-0.117)	3.026×10^{-11}	0.607
Musculature	0.127 (0.060-0.164)	0.089	0.620
Kidney	0.152 (0.093-0.209)	1.643×10^{-5}	0.693
Pancreas	0.154 (0.109-0.184)	6.810×10^{-5}	0.680
Skin	0.161 (0.109-0.255)	2.033×10^{-7}	0.765
Lung	0.169 (0.121-0.231)	2.700×10^{-7}	0.748
Liver	0.219 (0.171-0.274)	2.590×10^{-20}	0.813

Table S4. Neuron-specific genes have significantly lower dN/dS than glia-specific genes even when using difference reference genomes

Reference Genome	*Neuron-specific Genes median dN/dS ($CI_{95\%}$)*	*Glia-specific Genes median dN/dS ($CI_{95\%}$)*	*Mann-Whitney U Test p-value*
Human	0.092 (0.087-0.097)	0.144 (0.136-0.154)	5.479×10^{-29}
Rat	0.070 (0.066-0.076)	0.125 (0.118-0.133)	8.956×10^{-13}
Chicken	0.088 (0.084-0.093)	0.114 (0.105-0.123)	4.308×10^{-10}

Dataset S1 (separate file). Species included in this study. Sheet 1 includes the mammalian species used for calculating pairwise dN/dS values against mouse, rat, and human reference genome. Sheet 2 includes the reptile species used for calculating pairwise dN/dS values against chicken reference genome. Sheet 3 includes the 60 vertebrate species of the 60-way multiple alignment used for phastCons analysis.

Dataset S2 (separate file). Original expression level data from the Barres Lab's 2014 study(1) on mouse brain cell type enrichment. Expression levels were recorded as FPKM.

Dataset S3 (separate file). Pairwise dN/dS ratios, averaged across 92 species, of cell type-specific genes against mouse reference genome. The cell types include neuron, endothelial cells, glial cells, and the glial subtypes (astrocyte, oligodendrocyte, microglia).

Dataset S4 (separate file). Pairwise dN/dS ratios of all protein-coding genes available with mouse, human, rat, or chicken reference genome. While the pairwise dN/dS calculated against mouse, human, and rat genome are the same mammalian species, the pairwise dN/dS calculated against the chicken reference genome are a different set of reptile species.

Dataset S5 (separate file). Genes belong to each GO slim term

Dataset S6 (separate file). Statistics of Contingency Analysis between cell type-specificity and the bottom and top 25% values of dN/dS ratios for genes specific to each GO. The GO terms are ranked by the odds ratio of contingency between neuron-specific genes and genes with the lowest 25% dN/dS ratios within the GO. The lowest two p values between neuron-specific genes and the lowest 25% dN/dS ratios are highlighted in yellow, and they are "Transport" and "Signal Transduction". num_genes: number of brain cell type-specific genes that belong to each GO term; n_med: median dN/dS ratio for neuron-specific genes that belong to that GO term; low_n_chi2: the χ^2 calculated by a χ^2 test of independence between neuron-specific genes and the genes with lowest 25% dN/dS ratios within a given GO term; low_n_p: the p value calculated by a χ^2 test of independence between neuron-specific genes and the genes with lowest 25% dN/dS ratios within a given GO term; low_n_OR: the odds ratio calculated by a Fisher's exact test between neuron-specific genes and the lowest 25% dN/dS ratios within a given GO term; low_n_fisher_p: the p value calculated by a Fisher's exact test between neuron-specific genes and the lowest 25% dN/dS ratios within a given GO term.

Dataset S7 (separate file). Contingency analysis between GO slim terms with cell type-specific genes with high and low dN/dS

SI References

1. Y. Zhang, *et al.*, An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. *J. Neurosci.* **34**, 11929–11947 (2014).
2. B. Yates, *et al.*, Genenames.org: The HGNC and VGNC resources in 2017. *Nucleic Acids Res.* **45**, D619–D625 (2017).
3. E. Eisenberg, E. Y. Levanon, Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
4. K. Breuer, *et al.*, InnateDB: Systems biology of innate immunity and beyond - Recent updates and continuing curation. *Nucleic Acids Res.* **41**, D1228-33 (2013).
5. T. Shiina, A. Blancher, H. Inoko, J. K. Kulski, Comparative genomics of the human, macaque and mouse major histocompatibility complex. *Immunology* **150**, 127–138 (2017).
6. C. M. Smith, *et al.*, The mouse Gene Expression Database (GXD): 2019 update. *Nucleic Acids Res.* **47**, D774–D779 (2019).
7. E. I. Boyle, *et al.*, GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710–3715 (2004).