Neuron-specific coding sequences are the most highly conserved in the mammalian brain

Linhe Xu [a], Suzana Herculano-Houzel *[a,b,c]

[a] Department of Psychological Sciences, Vanderbilt University

[b] Department of Biological Sciences, Vanderbilt University

[c] Vanderbilt Brain Institute, Vanderbilt University

*Corresponding author. Email: suzana.herculano@vanderbilt.edu

Author Contributions: Xu and Herculano-Houzel designed research; Xu conducted the research; Xu and Herculano-Houzel wrote the manuscript together.

Competing Interest Statement: The authors declare that they have no competing interests.

Keywords: dN/dS | phastCons | neuronal diversity | glial cells | negative selection

Abstract

Neurons have become highly diverse in cell size, morphology, phenotype and function in mammalian evolution, whereas glial cells are much less varied in size and types across species. This difference in diversity suggests that neuron-specific protein-coding gene sequences have admitted more variation in evolution, that is, are less evolutionarily conserved than those expressed in glial cells. We calculated values of dN/dS from Ensembl98 for coding sequences expressed specifically in neurons, astrocytes, oligodendrocytes, microglia and endothelial cells in the brain across 92 mammalian species with reference to either mouse or human. Surprisingly, we find that protein-coding sequences that are specifically expressed in neurons are far less variable than those specific to other cell types in the brain. We next analyzed phastCons values for the same genes and found that neuron-specific promoter sequences are at least as conserved as other cell type-specific promoter sequences. Moreover, neuron-specific coding sequences are as highly conserved across mammalian species as ATPase coding sequences, the benchmark of evolutionary conservation, followed by heart and skeletal muscle-specific sequences. Neuronal diversity in mammalian evolution thus arises despite high levels of negative selection on neuron-specific protein-coding sequences. We propose that such strong evolutionary conservation is imposed by excitability, which continually exposes cells to the risk of excitotoxic death, and speculate that variability of neuronal cell sizes arises as a consequence of variability in levels of activity, possibly constrained by energy supply to the developing brain.

Significance Statement

The majority of cells in the mammalian body, including glial cells in the brain, maintain a constant size across species from mice to elephants. The exception is neurons, whose size varies over 100-fold both within each brain and across species, often becoming larger in larger brains. What evolutionary mechanism allows neurons to be so exceptionally diverse in size? We show that neuron-specific genes are some of the most highly conserved in mammalian evolution, which implicates alternate causes of diversity in neuronal cell size beyond heritable genetics. We propose a novel direction of research into how neuronal cell size diversity might be a self-organized response to variations in levels of activity, which are an exclusive property of excitable cells like neurons.

Introduction

Brain-specific genes have been under stronger selective pressure than widely expressed genes (1). However, brains are composed of neurons, glial cells and endothelial cells, and it is not known whether genes expressed in these diverse brain cell types are under equally strong selective pressure. Neurons, in particular, come in multiple shapes, types, and at least 1,000-fold different densities (and thus cell size), while glial cells are much more restricted in cell types and vary in density by less than 10-fold both within brains and across species (2, 3). We have suggested that the high conservation of glial cell types and densities is evidence of strict evolutionary constraints against glial cell variation (4). Additionally, neuron-specific cell diversity could be due to any of a combination of self-regulatory mechanisms like volume- and activity-related energy cost, or genetic constraints, increased alternative splicing in neurons, more variable gene regulatory cascades, earlier specification of neurons in development, before extensive chromatin changes that restrict gene expression (4). Here we focus on testing the possibility that the larger diversity of neuronal versus glial cells is related to glia-specific protein-coding sequences and regulatory sequences being under higher negative selection pressure than neuron-specific sequences (4). We test this hypothesis by comparing protein-coding and regulatory sequences of neuron-, glial- and endothelial cell-specific genes across 92 mammalian species (SI Appendix, Dataset S1), initially using the mouse genome as reference.

To differentiate between genes that are expressed specifically in neuronal, glial or endothelial cell types, we use the Barres mouse brain transcriptomic database (5) that lists the expression level of 22,458 genes in neurons, astrocytes, microglia, endothelial cells, as well as in three developmental stages of oligodendrocytes (*SI Appendix*, Dataset S2). Using a criterion of selective enrichment over other cell types (*SI Appendix*, Detailed Methods), we selected for analysis 1,298 neuron-specific genes, 1,211 glial cell-specific genes, and 937 endothelial cell-specific genes (*SI Appendix*, Dataset S3). As a proxy for evolutionary selective pressure acting on protein-coding sequences, we compared dN/dS ratios (6) of neuron-, glia- and endothelium-specific genes calculated from dN and dS values in the Ensembl 98 database, using the mouse genome as reference. dN/dS ratios are calculated pairwise for each gene between two species, and represent the ratio between non-synonymous and synonymous substitution rates in protein-coding sequences for that gene. A ratio of 1 indicates neutral evolution of the coding sequence. Values of dN/dS greater than 1 indicate positive/directional selection, and values of dN/dS smaller than 1 indicate negative/purifying selection. We expected that if the coding sequences of both neuron- and glia-specific genes are under negative/purifying selection, they should consistently exhibit dN/dS<1, consistent with previous findings in the literature (1). Additionally, we predicted that

neuron-specific genes should have larger values of dN/dS than glial cell-specific genes and endothelium-specific genes.

Results

We pooled pairwise dN/dS ratios from all 92 mammalian species available on Ensembl 98, initially using the mouse genome as reference (*SI Appendix*, Dataset S4). Those values are shown in Figure 1A, where every dot represents the dN/dS ratio for one gene in one species compared to the mouse ortholog. Contrary to our prediction, we found that the distribution of dN/dS ratios is shifted towards lower values for neuron-specific genes compared to glial- or endothelial cell-specific genes (Mann-Whitney U test, both $p<0.0001$), with no difference in dN/dS distribution between the latter ($p=0.337$; Fig. 1A; *SI Appendix*, Fig. S1). Next, because some genes have missing orthologs in some species, we also calculated the average dN/dS for each gene across all species with an ortholog (*SI Appendix*, Dataset S3), and compared these values of dN/dS across cell types. As in Fig. 1A, we find that distribution of average dN/dS ratios is significantly shifted towards lower values for neuron-specific genes compared to glial- and endothelial cell-specific genes (both $p<0.0001$), with no difference in dN/dS distribution between the latter ($p=0.719$).

We then analyzed the distribution of dN/dS ratios averaged across mammalian species for glia-specific genes separately by breaking them down into non-overlapping 824 astrocyte-specific genes, 1,062 microglia-specific genes, 521 oligodendrocyte-specific genes, besides the 1,298 neuron-specific and 937 endothelium-specific genes (*SI Appendix*, Detailed Methods; data in *SI Appendix*, Dataset S3). Interestingly, we find that genes specific and exclusive to each glial subtype all have a distribution of dN/dS ratios that is significantly shifted towards higher values than neuron-specific genes (Fig. 1B; Table 1). Microglia-specific genes, in particular, have a distribution of average dN/dS ratios with an even higher median than endothelial cell-specific genes. Oligodendrocyte-specific genes have a distribution of dN/dS values that is lower than that for other glial subtypes, but still tends towards higher values compared to neuron-specific genes (Table 1). Both oligodendrocyte- and astrocyte-specific genes have dN/dS ratios lower than endothelial cell-specific genes at the significance level of $p<0.0001$.

Pairwise dN/dS ratios for a few representative mammalian species from different clades (human, rat, opossum, the megabat *Pteropus vampyrus*, Tasmanian devil, cat, and pig) are shown in *SI Appendix*, Fig. S2 and Table S1, all sharing the same finding of significantly lower values of dN/dS for neuron-specific genes than genes specific to other cell types. A similar analysis using all genes expressed by each cell type (whether or not exclusively), defined as all protein-coding genes with FPKM higher than 1 for each cell type, also confirmed that neuron-expressed genes have a

distribution of dN/dS ratios in these select species shifted towards lower values than other cell type-expressed genes (*SI Appendix*, Table S2, Fig. S3).

To provide benchmarks against which to compare the values of dN/dS for neuron-specific genes, we calculated median dN/dS ratios for genes known to be highly or poorly conserved, as well as for all protein-coding genes, across the same 92 mammalian species compared to the mouse genome. Surprisingly, we find that almost half of all neuron-specific genes have values of dN/dS smaller than the median dN/dS of ATPases and housekeeping genes, which are the gold standard for highly conserved gene sequences (low dN/dS; Table 1 [7, 8]). The distribution shown in Figure 1B indicates that neuron-specific genes are as conserved as ATPase-coding sequences. Conversely, microglia-specific genes have a similar distribution of dN/dS ratios to immune system-specific genes, which are known to be highly variable (high dN/dS) (9), although still smaller than dN/dS ratios for major histocompatibility complex (MHC) genes (10).

To better appreciate the uniqueness of the low dN/dS ratios of neuron-specific genes, we compared them to dN/dS ratios of organ-specific genes in mouse (Fig. 2). Interestingly, except for heart- ($p=0.232$) and skeletal muscle-specific genes ($p=0.089$), we find that liver, lung, skin, pancreas, kidney, and brain-specific genes as a whole all have significantly higher average dN/dS ratios than neuron-specific genes (all other $p<0.0001$; *SI Appendix*, Table S3).

Next, we sought to determine whether the neuron-specific genes with low dN/dS, and the microglial cell-specific genes with high dN/dS, are disproportionately related to particular biological processes. We first identified the 71 slim gene ontology (GO) categories (11) to which the cell type-specific genes belong (*SI Appendix*, Dataset S5). For each GO category, we then performed a contingency analysis between cell type-specificity and the bottom and top 25% values of the distribution of dN/dS using the χ^2 test of independence (*SI Appendix*, Dataset S6 and S7). This test allows us to find whether any GO category is particularly overrepresented amongst neuron-specific genes with low dN/dS ratios, or amongst microglia-specific genes with high dN/dS ratios. Notably, transport-related and signal transduction-related GOs are significantly associated with neuron-specific genes with low dN/dS ratios (transport: $\chi^2=78.407$, $p<0.0001$; signal transduction: $\chi^2=54.195$, $p<0.0001$). Nonetheless, virtually all GO categories with more than 225 cell type-specific genes represented show a significant association between neuronal cell type and dN/dS ratios in the lowest 25% of the distribution. We also found highly significant associations between immune system-related GOs and the microglial cell-specific genes with the top 25% values of dN/dS ($\chi^2=36.048$, $p<0.0001$), in agreement with previous reports of enrichment of immune system-related genes in microglia and endothelial cells (12).

To ensure that the association between low dN/dS ratios and neuron specificity is not unique to our use of the mouse genome as reference, we replicated the entire study using human, rat, or chicken genomes as reference for the calculation of dN/dS ratios. All four reference genomes yield

similar results, with significantly lower dN/dS ratios for neuron-specific genes than glia-specific genes (*SI Appendix*, Table S4; Dataset S4).

Finally, to test if the strength of negative selection of the promoter region for brain cell type-specific genes agrees with the protein-coding sequences, we used phastCons (13) to compare the 2,000 bp upstream of the coding sequences for each cell type-specific gene. phastCons quantifies the strength of negative selection for each base pair given its flanking base pairs, based on probabilities of state transitions in phylogenetic hidden Markov models. The phastCons score represents the probability of negative selection, ranging from 0 to 1 such that the higher the phastCons value for a nucleotide in the sequence, the more likely that that base pair went through negative selection. For each gene, we calculated the average phastCons score for the 2,000 base pairs upstream of the coding region. phastCons scores were sourced from the UCSC Genome Browser (14) using a multiple alignment of 60 species including the mouse GRCm38 reference genome. At the $p<0.0001$ level employed for coding sequences, we found no significant differences between phastCons scores for the promoter regions of neuron-specific and glia-specific genes ($p=0.003$) or endothelium-specific genes ($p=0.041$), and no significant differences between phastCons of endothelium-specific and glia-specific upstream regulatory sequences ($p=0.466$; *SI Appendix*, Fig. 3A). Further, no significant differences are found when comparing glial subtypes individually (neuron-specific genes vs. oligodendrocyte- microglia- and astroglia-specific genes: $p=0.034$, $p=0.022$, and $p=0.524$, respectively; *SI Appendix*, Fig. S4). Importantly, log-transformed dN/dS ratios of the coding region of cell type-specific genes have very low explanatory power over phastCons for the promoter region of the same genes (linear regression, $r^2=0.008$, $p<0.0001$; Spearman correlation, $\rho=-0.083$, $p<0.0001$; neuron-specific genes only, linear regression, $r^2=0.001$, $p=0.397$; Spearman correlation, $\rho=-0.026$, $p=0.358$; Figure 3B). The low correlation that fails to reach statistical significance for promoter regions of neuron-specific genes indicates that strong conservation of coding sequences is not compensated by increased diversity at the promoter region of neuron-specific genes.

Discussion

Here we find that coding sequences of neuron-specific genes are under remarkably stronger negative selection than coding sequences of endothelial cell- and glial cell-specific genes. Our findings build on the previous discovery that brain-specific genes as a whole are highly conserved (1) by showing that it is specifically the coding sequences of neuron-specific genes that admit very little evolutionary variation.

It has long been established that differences in protein-coding DNA sequences cannot explain most cross-species morphological differences (15); rather, changes in gene expression and regulation are major players in establishing a species' biological uniqueness (16). Such a minor role of coding sequence variations in morphological uniqueness is consistent with coding sequences usually being under strong negative selection, as can be expected from the active structural and functional roles that proteins play in living cells. Still, we find that regulatory sequences for neuron-specific genes appear to be subject to similarly strong negative selection as glial-specific genes. Therefore, our findings imply that the enormous morphological and functional diversity of neurons within and across mammalian brains arises in a context of particularly strong protein sequence conservation of neuron-specific genes that is not compensated by any increased evolutionary variability in the promoter region.

Our previous research on cell densities in different brain structures and species shows that densities vary by about 2-fold in non-neuronal cells but by 1,000-fold across neurons (17). Because cell density serves as an indicator of average cell volume (including dendrites, axons, and cell body; 18) we can infer that neuronal cell volumes are about 500 times more variable than non-neuronal cell volume. Here we find that neurons are much more variable in size and phenotypes than glial cells despite neuron-specific proteins being subjected to much stronger negative selective pressures in mammalian evolution. Given that we find no large differences in selection pressure between the promoter sequences of neuron- and glia-specific genes, remaining possible genetic sources of the enormous neuronal diversity despite highly constrained protein sequence evolution are variable relative levels of expression of conserved sequences that might be due to gene duplications, levels of expression of transcription factors, distant enhancers of neuron-specific genes, or aneuploidy, none of which can be detected by the methods of the present study. Alternatively, there might be a small set of cell size-controlling genes expressed in neurons that are free to vary between species and escaped our analysis. Future studies will address this possibility.

However it happens, one expected consequence of the development and evolution of larger neuronal cell sizes is higher energy cost, simply because of the larger volume of biosynthesis that builds larger dendritic arbors, axons and neuronal cell bodies (19). In neurons, larger cell sizes could become particularly prohibitive from the energetic standpoint due to increased membrane capacitance which requires more energy to recharge the membrane, and would be compounded by increased numbers of synapses – unless levels of activity can undergo compensatory regulation (20, 21). The importance of factoring in energetics when considering the evolution of neuronal diversity is evidenced by the finding that the amount of energy that powers whole mammalian brains is linearly proportional to the numbers of neurons in the brain (21). Similarly, the amount of energy that powers one gram of tissue is proportional to the number of neurons in that gram of tissue, which is almost 3 times larger in mice than in humans (21). Importantly, we have recently

demonstrated that brain energetics is constrained by a limited energy supply that depends on capillary density in the brain (Herculano-Houzel and Rothman, submitted), in which case neurons within a territory supplied by the same capillary and astrocyte compete for the same shared energy supply. In such a scenario, while larger neurons are at a competitive advantage for that supply (because fewer neurons will fit in the same volume), their increased size imposes an unavoidable cost of increased biosynthesis and membrane capacitance (19). At the same time, any increases in activity (for instance through varying firing rate set points, which are characteristic of each neuron; 22) will take away from the energy available for biosynthesis, and thus impose a limit onto the volume that a neuron can attain through growth in brain development.

In line with this energy supply-limited framework, we propose that the marked variation in cell size found in neurons may be a self-regulated consequence of levels of activity which might in turn be determined through cell-specific transcriptomics (23). Such a central role of excitable activity in regulating neuronal cell size would also explain the surprisingly extreme level of conservation of neuron-specific genes, on par with the highly-conserved ATPase genes. We propose that such strong evolutionary conservation of neuron-specific coding sequences is related to the fact that neurons are singular amongst cells in the body in that they are continuously and rapidly excitable, thanks to their ability to regenerate the resting membrane potential, without which cells would die (24). We submit that viability under conditions of fast and relentless excitability is restricted to cells that express only very particular protein sequences, and tampering with them is swiftly met with cell death. We thus speculate that the large variability of neuronal cell sizes might arise as a consequence of a catabolism/anabolism trade-off in the face of cell-specific firing rates, such that energy that supports high levels of activity cannot be diverted into cell growth and extension of neurites. Conversely, in those neurons that have lower rates of activity early in development, there is enough energy to support cell growth. Thus, we propose that what sets neurons apart from other cells in the body in their enormous elasticity in cell volume is the fact that only neurons can dynamically vary their activity and thus rate of energy use. We plan to investigate the role that such a trade-off between firing activity and cell volume might play in both development and evolution of nervous systems.

Materials and Methods

Cell type-specific genes were selected based on gene expression levels in neurons, astrocytes, oligodendrocytes, microglia, and endothelial cells from the Barres Lab's 2014 mouse brain cell type enrichment study (5). Genes with expression level in one brain cell type greater than the sum of expression level in the other four cell types are defined as cell type-specific genes, giving

us 1,298 neuron-specific genes, 1,062 microglia-specific genes, 937 endothelia-specific genes, 824 astrocyte-specific genes, and 521 oligodendrocyte-specific genes. Glial cell-specific genes were defined as genes with expression level greater than four times its average expression level in neurons and endothelial cells, giving us 1,211 glia-specific genes. Based on the same database, protein-coding genes with expression level (FPKM) in one cell type greater than 1 are defined as cell type-expressed genes, giving 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.

ATPase genes, housekeeping genes, immune genes, and MHC genes were used as benchmarks in this study. ATPase genes used were mouse orthologs of human ATPase retrieved from the HUGO Gene Nomenclature Committee database (25). Housekeeping genes used were mouse orthologs of human housekeeping genes (26). Immune genes were mouse genes selected from the InnateDB database (27). The MHC genes were retrieved from the literature (28).

Liver, kidney (metanephros), lung, skin, brain, heart, pancreas, and musculature-specific genes are genes with expression level only in that organ but not anywhere else based on the MGI mouse gene expression database (GXD) (29).

Pairwise dN/dS ratios of the genes of interest between the mouse reference genome (GRCm38.p6) and each of 92 mammalian species were calculated from the dN and dS values from the Ensemnl 98 database, with orthologues that are not one-to-one orthologues filtered out. Pairwise dN/dS and averaged dN/dS can be found in *SI Appendix*, Dataset S3 and S4.

GO slim terms associated with each gene of interest are biological processing slim ontology terms restricted to the mouse species, and were retrieved from the Princeton University Lewis-Sigler Institute for Integrative Genomics' Generic GO Term Mapper (11). Non-slim ontology terms "regulation of cell size", "regulation of membrane potential", "membrane depolarization", "membrane hyperpolarization", and "membrane repolarization" were retrieved on top of the slim terms.

Promoter sequences of cell type-specific genes were defined as the 2,000 base pairs DNA sequence upstream of each gene. Coordinates of these promoters on the autosomes, sex chromosomes, or mitochondria genome were calculated based on the coordinates of the coding sequence and whether the gene is on the 5' or 3' strand. PhastCons scores based on a 60-way multiple alignment (phastCons60way) were then retrieved from the UCSC Genome Browser (14).

Statistical analyses, described in detail in *SI Appendix*, were conducted with python scripts, which we made publicly available at https://github.com/VeritatemAmo/neuron-glia-dNdS together with all codes and data needed for replicating this study.

Acknowledgments

We thank John A Capra and Jon H Kaas for advises and Vanderbilt University's Advanced Computing Center for Research and Education for technical support.

References

1. K. Biswas, S. Chakraborty, S. Podder, T. C. Ghosh, Insights into the dN/dS ratio heterogeneity between brain specific genes and widely expressed genes in species of different complexity. *Genomics* **108**, 11–17 (2016).

2. S. Herculano-Houzel, The glia/neuron ratio: How it varies uniformly across brain structures and species and what that means for brain physiology and evolution. *Glia* **62**, 1377–1391 (2014).

3. B. Mota, S. Herculano-Houzel, All brains are made of this: a fundamental building block of brain matter with matching neuronal and glial masses. *Front. Neuroanat.* **8**, 127 (2014).

4. S. Herculano-Houzel, S. Dos Santos, You Do Not Mess with the Glia. *Neuroglia* **1**, 193–219 (2018).

5. 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).

6. Z. Yang, PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13**, 555–556 (1997).

7. J. P. Gogarten, L. Taiz, Evolution of proton pumping ATPases: Rooting the tree of life. *Photosynth. Res.* **33**, 137–146 (1992).

8. L. Zhang, W. H. Li, Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes. *Mol. Biol. Evol.* **21**, 236–239 (2004).

9. A. J. Shultz, T. B. Sackton, Immune genes are hotspots of shared positive selection across birds and mammals. *Elife* **8** (2019).

10. S. B. Piertney, M. K. Oliver, The evolutionary ecology of the major histocompatibility complex. *Heredity (Edinb).* **96**, 7–21 (2006).

11. 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).

12. A. T. McKenzie, *et al.*, Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Sci. Rep.* **8**, 8868 (2018).

13. A. Siepel, *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

14. D. Karolchik, *et al.*, The UCSC table browser data retrieval tool. *Nucleic Acids Res.* **32**, 493D – 496 (2004).

15. M. C. King, A. C. Wilson, Evolution at two levels in humans and chimpanzees. *Science (80-.).* **188**, 107–116 (1975).

16. S. J. Sholtis, J. P. Noonan, Gene regulation and the origins of human biological uniqueness. *Trends Genet.* **26**, 110–118 (2010).

17. S. Herculano-Houzel, S. E. Dos Santos, You do not mess with the glia. *Neuroglia* **1**, 193-219 (2018).

18. B. Mota, S. Herculano-Houzel. All brains are made of this: A fundamental building block of brain matter with matching neuronal and glial masses. Front. Neuroanat. 8, 127 (2014).

19. N. Lane, How energy flow shapes cell evolution. *Curr. Biol.* **30**, R471–R476 (2020).

20. D. Attwell, S. B. Laughlin, An energy budget for signaling in the grey matter of the brain. *J. Cereb. Blood Flow Metab.* **21**, 1133–1145 (2001).

21. S. Herculano-Houzel, Scaling of brain metabolism with a fixed energy budget per neuron: Implications for neuronal activity, plasticity and evolution. *PLoS One* **6**, e17514 (2011).

22. G. G. Turrigiano, The Self-Tuning Neuron: Synaptic Scaling of Excitatory Synapses. *Cell* **135**, 422–435 (2008).

23. R. Yuste et al. A community-based transcriptomics classification and nomenclature of neocortical cell types. *Nat. Neurosci.* **23**, 1456-1468 (2020).

24. N. Lane, W. F. Martin, The origin of membrane bioenergetics. *Cell* **151**, 1406–1416 (2012).

25. B. Yates, *et al.*, Genenames.org: The HGNC and VGNC resources in 2017. *Nucleic Acids Res.* **45**, D619–D625 (2017).

26. E. Eisenberg, E. Y. Levanon, Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).

27. K. Breuer, *et al.*, InnateDB: Systems biology of innate immunity and beyond - Recent updates and continuing curation. *Nucleic Acids Res.* **41**, D1228-33 (2013).

28. 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).

29. C. M. Smith, *et al.*, The mouse Gene Expression Database (GXD): 2019 update. *Nucleic Acids Res.* **47**, D774–D779 (2019).

Figures and Tables

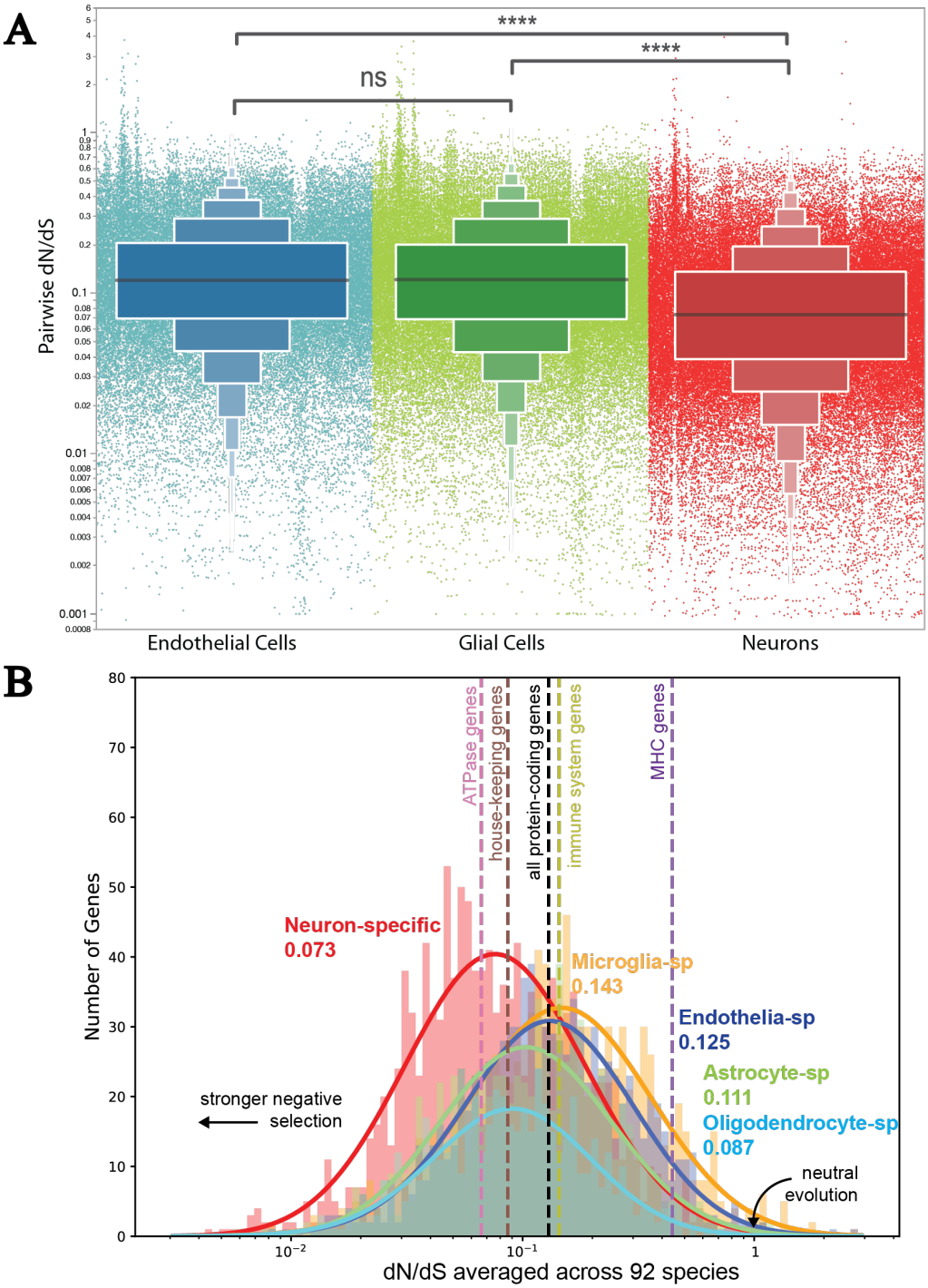

Figure 1. Neuron-specific genes have the lowest dN/dS. (**A**) Enhanced box plot of pairwise dN/dS ratios of neuron- (median: 0.070, $CI_{95\%}$: 0.070-0.071), glia- (median: 0.120, $CI_{95\%}$: 0.119-0.121), and endothelia- (median: 0.119, $CI_{95\%}$: 0.118-0.120) specific protein-coding genes in 92 mammalian species against the mouse ortholog. Each dot represents one gene's pairwise dN/dS in one species relative to the mouse ortholog: There are 86,260 dN/dS ratios for neuron, 61,687 for endothelial cell, and 80,131 for glial cell. Inter-species differences are present, and may be clade-related (*SI Appendix*, Fig. S1). Kruskal-Wallis test shows a significant difference across the medians ($p<0.0001$). Mann-Whitney U test shows differences in the distribution of dN/dS between neuron- and glial-specific genes, as well as between neuron- and endothelial- specific genes (both $p<0.0001$), but not between glial- and endothelial cell-specific genes ($p=0.337$). (**B**) Distribution of dN/dS ratios of neuron-, endothelia-, microglia-, astrocyte-, and oligodendrocyte-specific genes averaged across 92 mammalian species, all calculated against mouse sequences, compared to median dN/dS calculated for benchmark genes. A Kruskal-Wallis H-test shows significant difference across the five brain cell types' medians ($p<0.0001$). Genes specific to each glial subtype have dN/dS ratios that are significantly higher than those for neuron-specific genes (Mann-Whitney U test; all three $p<0.0001$). ATPases and house-keeping genes are considered benchmarks for very low dN/dS (and strong negative selective pressure), while immune system-related genes, and especially MHCs, represent benchmark for high dN/dS (and strong positive selective pressure). Numbers next to each cell type are the median dN/dS for genes specific to that cell type across 92 mammalian species. Neuron-specific genes have dN/dS ratios on par with ATPase genes ($p=0.058$), and lower than the ratios for housekeeping genes ($p=0.0003$) and all protein-coding genes ($p<0.0001$), which indicates a very high level of evolutionary conservation of neuron-specific coding sequences.

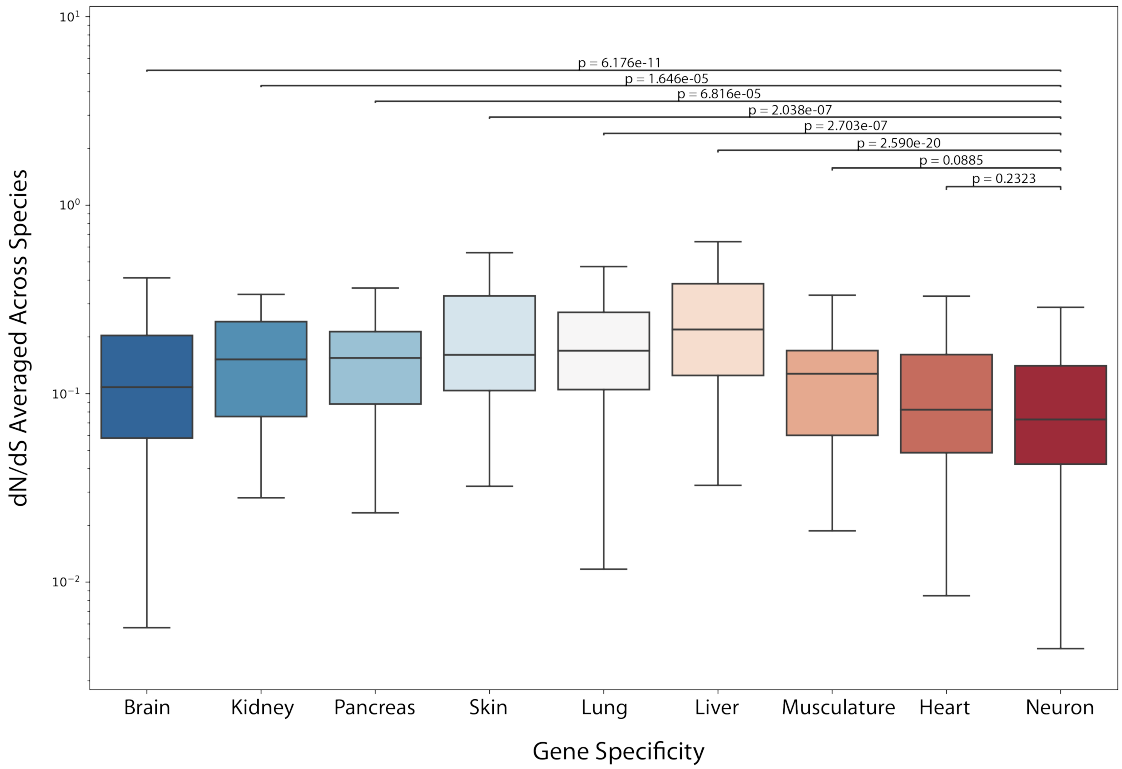

Figure 2. Comparison of dN/dS between neuron-specific genes and organ-specific genes in the body (including brain). Other than heart-specific and musculature-specific genes, all other organ-specific genes (including brain) have significantly higher dN/dS than neuron-specific genes. The p values were generated by Mann-Whitney U test between neuron-specific genes and each gene specificity. Note that brain-specific genes include both glia-specific and neuron-specific genes previously analyzed. Data points that extend 1.5 interquartile range past the low and high quartiles are identified as outliers and are not shown.

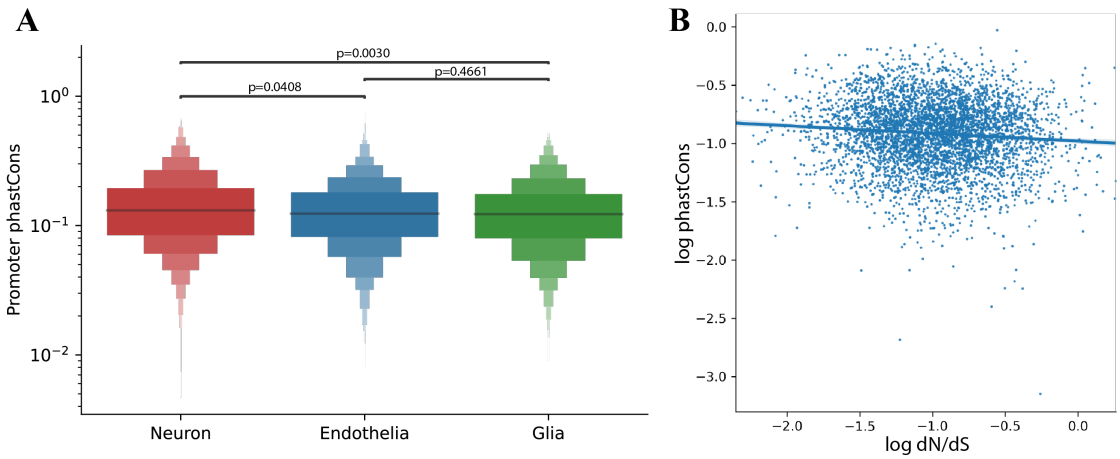

Figure 3. phastCons scores of the 2,000 bp upstream promoter region of brain cell type-specific genes of a mouse-based 60 species multiple alignment. (**A**) phastCons scores show no difference at the p<0.0001 level across neuron-, glial- and endothelia cell-specific genes. (**B**) Virtually no explanatory power was discovered between log transformed average dN/dS ratios and phastCons scores of brain cell type-specific genes with linear regression. Least square fitted line: $y = -0.0661x - 0.98$, F(1,4613)=37.04.

Table 1. Mouse cell type-specific genes' median dN/dS ratios and the p value and common language effect size (CLES) of Mann-Whitney U test against neuron. CLES values indicate what percentage of all possible pairwise comparisons between one group of genes and a reference group with a higher median dN/dS show a higher individual value in the reference group. For example, there is a 60.4% probability that a randomly selected astrocyte-specific gene has a higher dN/dS ratio than a randomly selected neuron-specific gene.

		Mann-Whitney Test Against Neuron		Mann-Whitney Test Against Endothelial Cell	
Cell Type or Benchmark	Median dN/dS ($CI_{95\%}$)	p-value	CLES	p-value	CLES
ATPase	0.066 (0.056-0.081)	0.0577	54.4%	4.7338×10^{-13}	70.3%
Neuron	0.073 (0.068-0.076)	N/A	N/A	2.325×10^{-40}	66.5%
Housekeeping genes	0.086 (0.083-0.090)	3.450×10^{-4}	53.2%	1.998×10^{-39}	64.0
Oligodendrocyte	0.087 (0.081-0.097)	9.071×10^{-4}	55.0%	1.4075×10^{-15}	62.5%
Astrocyte	0.111 (0.101-0.117)	5.749×10^{-16}	60.4%	1.4289×10^{-6}	56.5%
Endothelial Cell	0.125 (0.117-0.136)	2.325×10^{-40}	66.5%	N/A	N/A
Glial Cell	0.128 (0.121-0.134)	6.555×10^{-43}	65.8%	0.7193	49.6%
All Protein-coding genes	0.129 (0.127-0.131)	9.197×10^{-77}	65.3%	0.5652	49.8%
Microglia	0.143 (0.135-0.153)	7.102×10^{-59}	69.3%	1.098×10^{-3}	54.0%
Immune-specific genes	0.143 (0.129-0.160)	2.762×10^{-36}	67.1%	0.0147	53.2%
MHC	0.443 (0.337-0.511)	1.357×10^{-15}	95.2%	7.669×10^{-12}	88.7%