Neuron-specific coding and regulatory sequences are the most highly conserved in amniote brains despite neuron-specific cell size diversity

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

[a] Department of Psychology, Vanderbilt University

[b] Vanderbilt Brain Institute, Vanderbilt University

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

Author Contributions: Xu and Herculano-Houzel designed the 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 | molecular evolution | neuronal activity | firing rate

Abstract

Neurons are unique in that they are the only cell type in the body to display massive diversity in cell size, morphology, phenotype, and function both within individuals and across species. Here we use datasets encompassing up to 92 mammalian and 31 sauropsidian species to examine whether neuron-specific diversity occurs with higher evolutionary variation of neuron-specific coding and regulatory sequences compared to non-neuronal cell-specific sequences. We find that the opposite is true: Neuronal diversity in mammalian and sauropsidian evolution arose despite extreme levels of negative selection on neuron-specific protein-coding sequences on par with ATPase coding sequences, the benchmark of evolutionary conservation. We propose that such strong evolutionary conservation is imposed by excitability, which continually exposes cells to the risk of excitotoxic death, and speculate that neuronal cell size diversity is a self-organized consequence of variability in levels of activity, possibly constrained by energy supply to the developing brain.

Introduction

Most single-nucleated cells in the mammalian body, including glial cells in the brain, maintain a constant size across species from mice to elephants[1,2]. The exception is neurons, whose size (the sum of the soma, axon, and dendrite) varies widely both within each brain and across species, often becoming larger in larger brains within each clade, as gauged by neuronal densities spanning at least 1,000-fold variation across mammalian brain structures and species[3]. Moreover, neuronal variation extends beyond simple cell size: Single-cell transcriptomic studies have confirmed that at least several dozen different types of neuronal cells exist within a single cerebral cortex[4,5], onto which different physiological properties can be mapped[6]. In contrast, despite sharing a progenitor cell type with neurons[7], astrocyte and oligodendrocyte cells are much more restricted in subtypes and cell size (all arbors included), varying in density by less than 10-fold within brains and across mammalian species[3,8,9], with consistently only a few cell types defined transcriptomically[4,5].

We have suggested that the high conservation of glial cell types and densities compared to neurons is evidence of strict evolutionary constraints against glial cell variation, whereas the immense phenotypic diversity of neurons suggests evolutionary variation[9]. Brain-specific protein-coding genes have been shown to be under stronger selective pressure than widely expressed genes[10]. However, brains are composed of not only neurons but also glial cells and endothelial cells. A later study that examined all mRNA expressed in neuronal- and non-neuronal cells in mouse reported that neuron-expressed gene sequences tend to show stronger negative correlation between a measure of evolutionary conservation and levels of gene expression than gene sequences expressed in other cell types[11], which the authors attribute to stronger selection pressure against transcriptional and translational errors in more highly expressed sequences[12,13].

48 However, the explanatory power of gene expression level over coding sequence variation reported
49 in the literature is very low, always less than 10% both in whole tissues[14] (e.g., calculated from R
50 values reported from Additional Data File 6 of *Tuller et al., 2008*) and in single cells[11] (calculated
51 as Spearman correlation coefficients in *Hu et al., 2020*). Such a low explanatory power casts doubts
52 over the relevance of the authors' conclusion that there is stronger selective pressure against
53 variation in neuron-expressed genes compared to genes expressed in other cells. More recently,
54 an analysis between human and mouse of the levels of conservation of gene co-expression in
55 different cell types found less divergence in neurons compared to glial cells[15]. Most importantly, all
56 of these studies have been restricted to mouse-human or human-chimpanzee comparisons,
57 leaving unaddressed the wider question of whether the enormous diversity in neuronal cell size
58 compared to all other cell types throughout brain evolution that we have documented across ~100
59 species, including dinosaurs[16], birds[17], human[18], and other mammals[19], can be accounted for by
60 higher variability of neuron-specific genes, or, alternatively, whether that diversity occurs despite
61 strong conservation of these genes, in which case other mechanisms must be sought.

62 Here we compare protein-coding and upstream promoter sequences of neuron-, glial- and
63 endothelial cell-specific genes across 92 mammalian and 31 sauropsidian species (Error!
64 Reference source not found.) to test our initial hypothesis that neuron-specific DNA sequences are
65 more diverse than those specific to other brain cell types on the levels of protein-coding genes,
66 upstream promoter sequences, and cis-regulatory elements. To ensure that our results were not
67 biased or limited to any particular definition of cell type-specific genes, we performed all of our
68 analyses on four datasets, each generated with a different technique, tissue specificity, animal age
69 and sex, and data clustering method. The first is the Barres mouse brain transcriptomic database[20]
70 that lists the expression level of genes in collections of purified neurons, astrocytes, microglia,
71 endothelial cells, and three developmental stages of oligodendrocytes. The second and third are
72 the Linnarsson single cell transcriptomic databases obtained respectively with single cells isolated
73 with microfluidic chip sequenced by Illumina HiSeq 2000[21] or dissociated cell populations
74 sequenced by the droplet-based 10X Genomics Chromium[22], and which also differed in being
75 restricted to regions of cerebral cortical tissue[21] or extending to the entire CNS and PNS[22] of male
76 and female juvenile mice. Finally, the fourth dataset is another single cell transcriptomic database
77 obtained with a different droplet-based method, Drop-seq, in male adult mice[23].

78 As a proxy for evolutionary selective pressure acting on protein-coding sequences, we
79 compared dN/dS ratios[24] of neuron-, glia- and endothelium-specific genes calculated from dN and
80 dS values in the Ensembl 98 database[25], using either mouse, rat, human, or chicken as the
81 reference genome. The dN/dS ratios are calculated pairwise for each gene between two species
82 and represent the ratio between non-synonymous and synonymous substitution rates in protein-
83 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[10]. Additionally, we predicted that neuron-specific genes should have larger values of dN/dS than glial cell-specific genes and endothelium-specific genes given the extreme diversity in neuronal morphologies and cell sizes across mammalian species.

We used two approaches to compare evolutionary conservation of regulatory elements of glial and neuronal cell-specific sequences. First, we examined phastCons scores for the 2,000 bp upstream of coding sequences, as used in comparable studies[26,27], employing the same definition of cell type-specific genes as in the dN/dS analysis. The phastCons scores indicate the probability that an individual base pair in a sequence is evolutionarily conserved, and those scores can be averaged to provide a mean phastCons score for an entire promoter region across all species in the dataset[28–30]. Second, we examined phastCons scores for cis-regulatory elements uncovered by single-cell chromatin accessibility essays in neuronal or non-neuronal cells in the mouse cerebrum[31]. Because cis-regulatory elements are not confined to any particular upstream sequence, this analysis complements any shortcomings of the comparison of regulatory elements in the 2,000 bp upstream region of coding sequences.

Results

Contrary to our expectation that the large evolutionary diversity of neuronal cell size would correspond to less evolutionary conservation of neuron-specific sequences compared to glial cell-specific sequences, we found that in all four datasets, using the mouse genome as reference, the distribution of dN/dS ratios in mammals is sharply shifted towards lower values for neuron-specific genes compared to glial- or endothelial cell-specific genes. This is true whether the dN/dS ratio for *individual* neuronal-, glial- or endothelial-specific genes in each species is considered (**Figure 1A**; Error! Reference source not found. and Error! Reference source not found.); whether the *averaged* dN/dS ratio across *all* genes specific to neuronal, glial or endothelial cells of all subtypes pooled together is plotted in each species (**Figure 1C**; Error! Reference source not found. and Error! Reference source not found.), which circumvents the fact that some genes have missing orthologs in some species; and whether the *averaged* dN/dS across all genes specific to *specific* non-overlapping cell subtypes is plotted (**Figure 1E**; Error! Reference source not found. and Error! Reference source not found.). The same data are depicted as histograms in **Figures 1B and 1D**, which show that the distribution of average dN/dS ratios is significantly shifted towards lower values for neuron-specific genes compared to other cell subtype-specific genes. We find that genes

118 specific and exclusive to each glial subtype all have a distribution of dN/dS ratios that is significantly
119 shifted towards higher values than neuron-specific genes, with the highest median found for
120 microglia-specific genes (**Figure 1D,E**). Similar results were obtained when using the human
121 genome (Error! Reference source not found.) or the rat genome (Error! Reference source not
122 found.) as reference for the calculation of dN/dS, and also across 31 sauropsidian species, using
123 the chicken genome as reference (**Figure 2**), which indicates that our findings in mammalian
124 species generalize to amniotes.

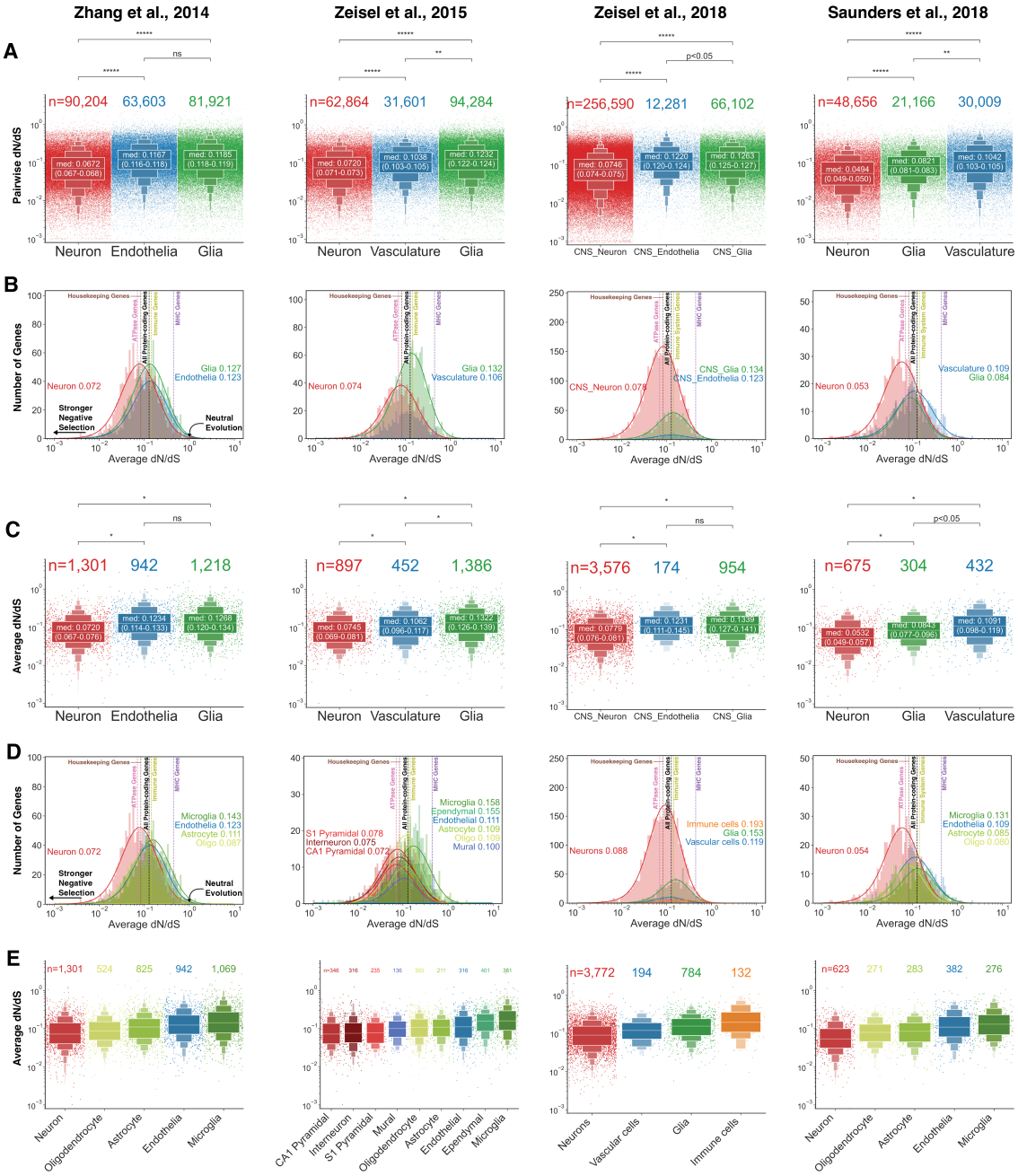

Figure 1: Neuron-specific genes have lower dN/dS ratios than genes specific to glial and endothelial cell types, calculated using the **mouse** genome as reference. **(A)** Enhanced box-strip plots (letter-value plots[32] overlaying on top of strip plots) of **pairwise** dN/dS ratios of neuron-, glial cell-, and endothelial/vascular cell-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. **(B, C)** Distribution plots and enhanced box-strip plots of dN/dS ratios **averaged** across 92 mammalian species against mouse reference genome for neuron-, glial cell-, and

133 endothelial/vascular cell-specific protein-coding genes. Each dot represents averaged dN/dS ratios
134 of one gene. **(D, E)** Distribution plots and enhanced box-strip plots of dN/dS ratios **averaged** across
135 92 mammalian species against mouse reference genome for cell-**subtype**-specific genes. Each
136 dot represents averaged dN/dS ratios of one gene. Each panel shows cell-type-specific genes
137 defined with expression data from four studies, labeled at the top of this figure. For enhanced box
138 plots: the p-values indicate the significance of Mann-Whitney U-tests within each pair of cell types;
139 "ns": $p \geq 0.05$, "$p<0.05$": $p<0.05$, "*": $p<10^{-4}$, "**": $p<10^{-84}$, "***": $p<10^{-164}$, "****": $p<10^{-244}$, "*****":
140 $p<5\times10^{-324}$. Numbers in the box for each cell type are the medians of pairwise dN/dS ratios, with a
141 95% confidence interval for the medians in parentheses. The sample size (number of genes) is
142 labeled for each cell type at the top. For distribution plots: a median dN/dS ratio is labeled for each
143 cell type. ATPases and housekeeping genes are considered benchmarks for very low dN/dS (and
144 strong negative selective pressure), while immune system-related genes, and especially MHCs,
145 represent a benchmark for high dN/dS (and strong positive selective pressure). The median of all
146 protein-coding genes is also included as a benchmark.
147

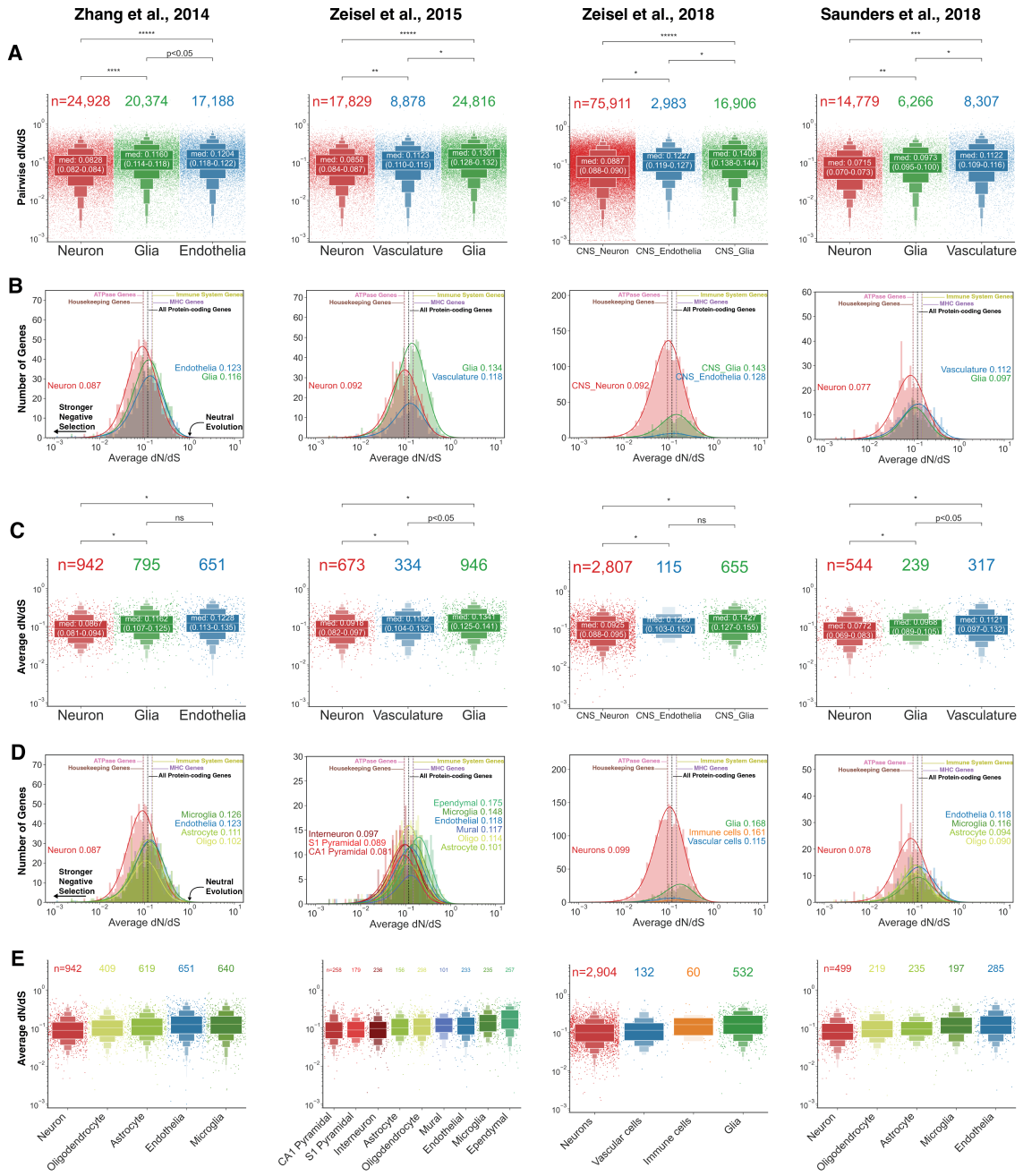

Figure 2: Neuron-specific genes have lower dN/dS ratios than genes specific to glial and endothelial cell types, calculated using the **chicken** genome as reference. **(A)** Enhanced box-strip plots of **pairwise** dN/dS ratios of neuron-, glial cell-, and endothelial/vascular cell-specific protein-coding genes in 31 **sauropsidian** species against the chicken ortholog. Each dot represents one gene's pairwise dN/dS in one species relative to the chicken ortholog. **(B, C)** Distribution plots and enhanced box-strip plots of dN/dS ratios **averaged** across 31 sauropsidian species against chicken reference genome for neuron-, glial cell-, and endothelial/vascular cell-specific protein-coding

genes. Each dot represents averaged dN/dS ratios of one gene. **(D, E)** Distribution plots and enhanced box-strip plots of dN/dS ratios **averaged** across 31 sauropsidian species against chicken reference genome for cell-**subtype**-specific genes. Each dot represents averaged dN/dS ratios of one gene. Each panel shows cell-type-specific genes defined with expression data from four studies, labeled at the top of this figure. For enhanced box plots: the p-values indicate the significance of Mann-Whitney U-tests within each pair of cell types; "ns": $p \geq 0.05$, "$p<0.05$": $p<0.05$, "*": $p<10^{-4}$, "**": $p<10^{-84}$, "***": $p<10^{-164}$, "****": $p<10^{-244}$, "*****": $p<5\times10^{-324}$. Numbers in the box for each cell type are the medians of pairwise dN/dS ratios, with a 95% confidence interval for the medians in parentheses. The sample size (number of genes) is labeled for each cell type at the top. For distribution plots: a median dN/dS ratio is labeled for each cell type.

The 92 mammalian species in our analyses encompass all major mammalian clades (marsupials, afrotherians, xenarthrans, glires, primates, carnivorans, chiropterans, eulipotyphlans, perissodactyls, artiodactyls), with the exception of monotremes. To address whether our mammal-wide results apply to all clades, we compared pairwise dN/dS ratios for neuronal and non-neuronal cell type-specific genes defined with the Barres expression database[20] in individual representative species from different clades: primates (human), glires (rat), marsupials (opossum and Tasmanian devil), chiroptera (large flying fox), carnivorans (cat) and artiodactyls (pig). In all of these species, we found significantly lower values of dN/dS for neuron-specific genes compared to genes specific to glial cells (

176 **Table 1**), glial subtypes (Error! Reference source not found.), and endothelial cells with the
177 exception of oligodendrocyte-specific genes in rat and flying fox (Error! Reference source not
178 found., upper panel). These results hold across the four gene-specificity-defining methods, the four
179 reference genomes, and the individual target species (92 mammalian species and 31 sauropsidian
180 species; Error! Reference source not found. and Error! Reference source not found.).
181 A similar analysis using all genes expressed by each cell type (whether or not exclusively),
182 defined as all protein-coding genes with FPKM higher than 1 for each cell type in the Barres dataset,
183 also confirmed that neuron-expressed genes have a distribution of dN/dS ratios in these select
184 species shifted towards lower values than other cell type-expressed genes (Error! Reference
185 source not found., lower panel; Error! Reference source not found.).

186 **Table 1:** Neuron-specific genes have lower dN/dS ratios than glia-specific genes in representative
187 mammalian species. Inferential statistics for cell type-specific genes dN/dS ratios' calculated
188 against the mouse reference genome for species representative of the different mammalian clades
189 examined, with cell-type-specificity defined according to expression data from the Barres
190 expression database[20]. To validate the result of lower neuron-specific genes' dN/dS ratios, neuron-
191 specific genes' dN/dS are compared to glia-specific genes' dN/dS between mouse and one of
192 seven representative mammalian species. Median dN/dS and 95% confidence interval ($CI_{95\%}$), as
193 well as the Mann-Whitney U test results (p-value and common language effect size), are reported
194 here. Values excerpted from **Supplementary Tables S6** and **S7**.
195

Species	*Neuron-specific Genes median dN/dS ($CI_{95\%}$)*	*Glia-specific Genes median dN/dS ($CI_{95\%}$)*	*Mann-Whitney U Test p-value*	*CLES*
Human	0.063 (0.059-0.068)	0.117 (0.112-0.123)	3.24×10^{-40}	66.27%
Rat	0.075 (0.068-0.083)	0.135 (0.126-0.143)	4.69×10^{-37}	65.30%
Opossum	0.061 (0.058-0.065)	0.085(0.079-0.090)	4.67×10^{-12}	59.23%
Megabat	0.084 (0.074-0.089)	0.130 (0.118-0.141)	4.01×10^{-16}	64.18%
Tasmanian Devil	0.057 (0.054-0.061)	0.087 (0.080-0.091)	7.03×10^{-19}	62.23%
Cat	0.067 (0.062-0.071)	0.114 (0.109-0.123)	3.99×10^{-38}	66.22%
Pig	0.064 (0.061-0.069)	0.114 (0.107-0.124)	1.96×10^{-41}	67.00%

196

197 To provide benchmarks against which to compare the values of dN/dS for neuron-specific
198 genes, we calculated median dN/dS ratios for genes known to be highly or poorly conserved, as
199 well as for all protein-coding genes, across the same 92 mammalian species compared to the
200 mouse (**Figure 1B,D**), human (Error! Reference source not found.), and rat reference genomes
201 (Error! Reference source not found.), and across the same 31 sauropsidian species compared to
202 the chicken reference genome (**Figure 2B,D**). We found that almost half of all neuron-specific genes
203 have values of dN/dS smaller than the median dN/dS of ATPases and housekeeping genes (low
204 dN/dS), which are the gold standards for highly conserved gene sequences[33,34]. Remarkably,
205 neuron-specific genes are as evolutionarily conserved as the benchmarks of conservations, if not
206 more, regardless of the reference genome used, and across four independent gene expression
207 datasets (Error! Reference source not found.; Error! Reference source not found.). Conversely,
208 with human, rat, and mouse reference genomes, microglia-specific genes have a similar distribution
209 of dN/dS ratios to immune system-specific genes, which are known to be highly variable (high
210 dN/dS)[35], although still smaller than the high dN/dS ratio benchmark for major histocompatibility
211 complex (MHC) genes[36].

212 To put into context the low dN/dS ratios of neuron-specific genes (defined according to *Zhang*
213 *et al., 2014*) [20], we compared them to the dN/dS ratios of organ-specific genes in mouse, defined
214 as genes with expression detected only in that organ but not anywhere else in the MGI GXD
215 mouse gene expression database[37]. Interestingly, except for heart- and musculature-specific
216 genes, we found that liver, lung, skin, pancreas, kidney, and brain-specific genes as a whole all
217 have significantly higher average dN/dS ratios than neuron-specific genes (pancreas<0.05, all
218 other p<0.0001; **Figure 3**;

219 **Figure 3.** All organ-specific genes (including the brain) other than heart-specific and musculature-
220 specific genes have significantly higher dN/dS than neuron-specific genes. The box plot shows a
221 comparison of dN/dS between neuron-specific genes, as defined by the Barres dataset, and organ-
222 specific genes in the body, using the mouse genome as reference. The p-values were generated
223 by the Mann-Whitney U test between neuron-specific genes and each gene specificity. Note that
224 brain-specific genes overlap with both glia-specific and neuron-specific genes previously analyzed.
225 Data points that extend the 1.5 interquartile range past the low and high quartiles are identified as
226 outliers and are not shown.

227 **Table 2**).

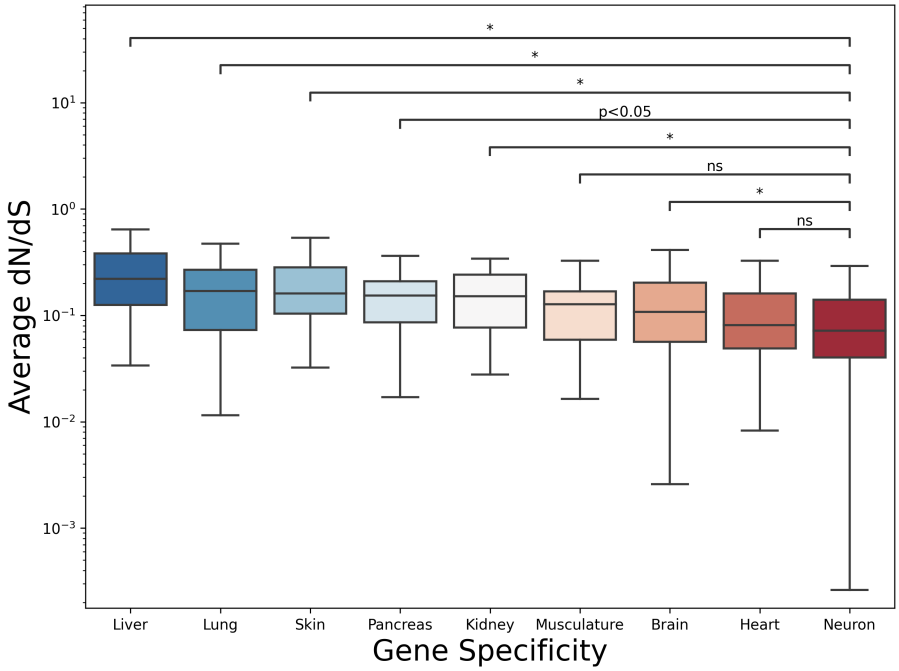

Figure 3. All organ-specific genes (including the brain) other than heart-specific and musculature-specific genes have significantly higher dN/dS than neuron-specific genes. The box plot shows a comparison of dN/dS between neuron-specific genes, as defined by the Barres dataset, and organ-specific genes in the body, using the mouse genome as reference. The p-values were generated by the Mann-Whitney U test between neuron-specific genes and each gene specificity. Note that brain-specific genes overlap with both glia-specific and neuron-specific genes previously analyzed. Data points that extend the 1.5 interquartile range past the low and high quartiles are identified as outliers and are not shown.

Table 2. Organ-specific dN/dS compared to neuron-specific genes defined using the Barres dataset as in **Figure 3**. The dN/dS ratios listed are the average across 92 mammalian target genomes against the mouse reference genome. Organ-specificity is based on the MGI expression database[37]. The 95% confidence intervals are provided alongside the median dN/dS. Mann-Whitney U-test p-values and common language effect sizes (CLES) are reported.

Organ	# Genes	Median dN/dS ($CI_{95\%}$)	M.W.U. p-value Against Neuron	CLES
Neuron	1,301	0.07 (0.07-0.08)	N/A	N/A
Heart	48	0.08 (0.06-0.11)	0.21	55.37%
Brain	412	0.11 (0.09-0.12)	$\mathbf{5.94 \times 10^{-11}}$	60.68%
Musculature	17	0.13 (0.06-0.16)	7.85×10^{-2}	62.40%
Kidney	43	0.15 (0.09-0.21)	$\mathbf{9.83 \times 10^{-6}}$	69.79%
Pancreas	42	0.15 (0.11-0.18)	$\mathbf{2.48 \times 10^{-4}}$	66.59%
Skin	34	0.16 (0.10-0.25)	$\mathbf{1.08 \times 10^{-6}}$	74.46%
Lung	37	0.17 (0.12-0.23)	$\mathbf{1.73 \times 10^{-6}}$	73.02%
Liver	77	0.22 (0.17-0.27)	$\mathbf{1.68 \times 10^{-20}}$	81.43%

Next, we sought to determine, still using the Barres dataset to define cell type-specific genes[20], 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[38] to which the cell type-specific genes belong (Error! Reference source not found.). 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 (concatenated stats in Error! Reference source not found.; visualization and detailed contingency tables in **Supplementary Data S3**). 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. As expected given the mesodermal origin of microglial cells[39], we found highly significant associations between microglial cell-specific genes with the top 25% values of dN/dS and immune system-related GOs (χ^2=36.048, p=1.92×10^{-9}), in agreement with previous reports of enrichment of immune system-related genes in microglial cells[40]. In contrast, neuron-specific genes with low dN/dS ratios were significantly associated with transport-related and signal transduction-related GOs (transport: χ^2=78.41, p=8.39×10^{-19}; signal transduction: χ^2=54.20, p<1.82×10^{-13}). Nonetheless, virtually all GO categories with more than 225 cell type-specific genes represented showed a significant association between cell type being neuronal and dN/dS ratios in the lowest 25% of the distribution.

Finally, we tested whether the large evolutionary variability of neuronal cell size might be attributed to a particularly high evolutionary variability of neuronal cell type-specific **regulatory regions**, including the promoters and the cis-regulatory regions of neuron-specific genes, despite the extreme conservation of the coding regions. We first defined promoter regions as the 2,000 bp upstream of the coding sequences for each cell type-specific gene, according to common usage in the literature[26,27], and defining cell-specificity of genes according to each of the four databases used in the dN/dS analysis[20–23]. We used phastCons scores[41] to quantify the evolutionary conservation level of non-coding regulatory elements per gene. The phastCons score represents the probability of negative selection for each base pair given its flanking base pairs, based on probabilities of state transitions in phylogenetic hidden Markov models, with values 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.

Average phastCons scores for the 2,000 bp upstream **promoter** region for cell type-specific genes were sourced from the UCSC Genome Browser[29] using 1) phastCons60way,which is a multiple alignment of 60 vertebrate species, including 44 mammals, 6 sauropsids (birds and non-avian reptiles), 9 fish, and 1 amphibian, centered around the mouse GRCm38 reference genome, and 2) phastCons77way, which is a multiple alignment of 77 vertebrate species, including 62 sauropsids, 10 fish, 3 amphibians, and 2 mammals. Promoters with less than 80% coverage were

filtered out. At the same $p<0.0001$ level employed for coding sequences, we found no significant differences in promoter region phastCons scores between neuron- and endothelial cell-specific genes; between phastCons of endothelial cell- and glial cell-specific genes; or between neuron- and glia-specific genes. The only exception was found in the Saunders dataset using the chicken-centered phastCons77way score, where promoter regions of neuron-specific genes had significantly higher phastCons scores (therefore more conserved) than those of glial-specific promoter regions (**Figure 4A, C**). Adopting a less stringent p-value threshold of 0.05 would yield promoter phastCons scores that were actually higher for neuron-specific genes compared to glial-cell-specific genes (except in the *Zeisel et al., 2015* dataset with phastCons60way), which would suggest that neuronal promoter sequences are more evolutionarily conserved than glial promoter sequences, not less. Importantly, log-transformed dN/dS ratios of the coding region of cell type-specific genes have virtually no explanatory power over phastCons for the promoter region of the same genes (**Figure 4B,D**). The low correlation that fails to reach statistical significance indicates that the strong conservation of coding sequences unveiled by dN/dS ratios is not compensated by an increased diversity at the 2,000-bp promoter region of neuron-specific genes.

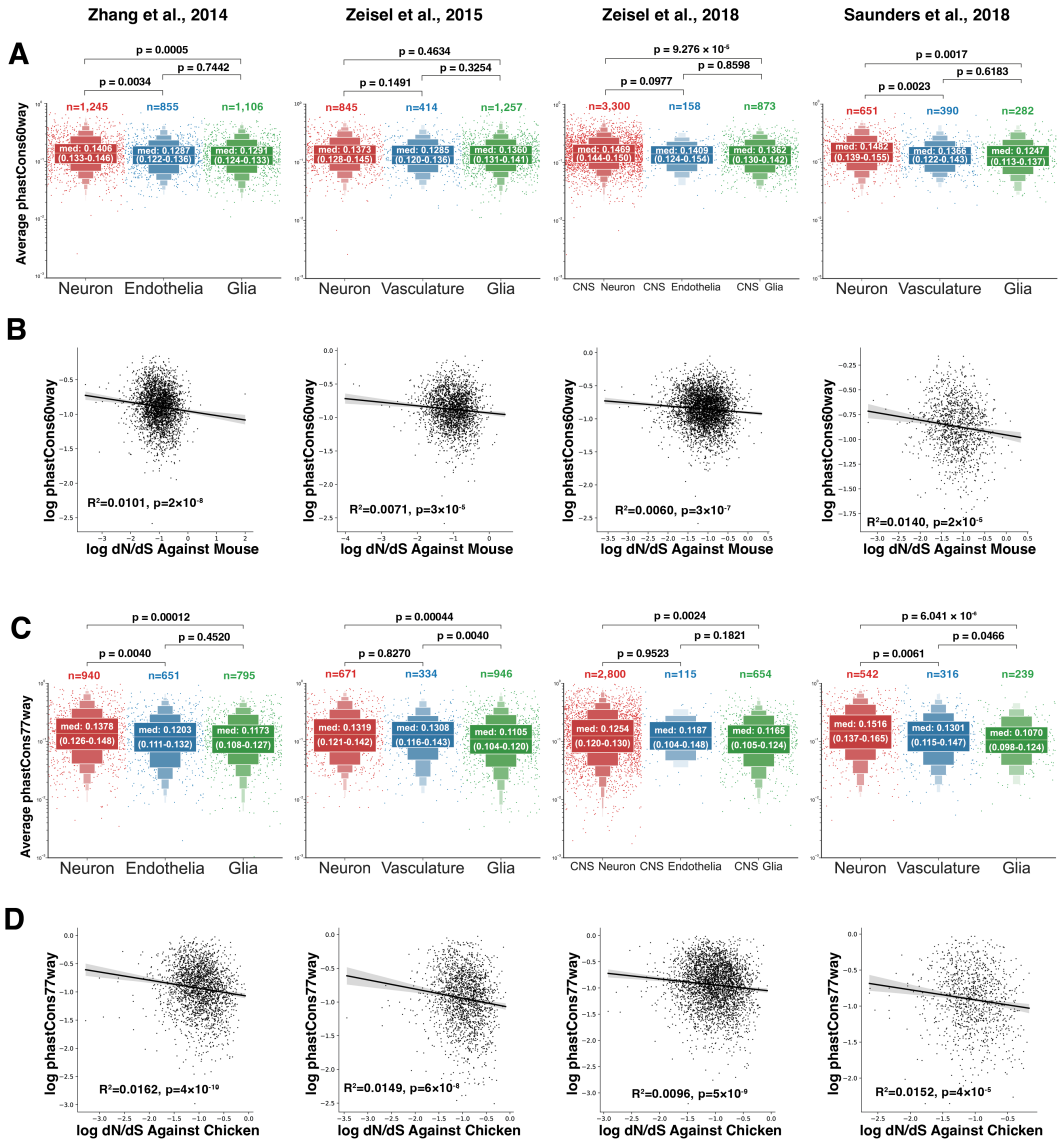

Figure 4: Upstream 2,000 bp regions of neuron-specific genes are not more conserved than those of glial-specific genes. With cell type-specific genes defined with four different datasets, we compared phastCons scores of the 2,000 bp upstream promoter region of cell type-specific calculated with (**A, B**) mouse-referenced multiple alignment of the genomes of 60 vertebrate species and (**C,D**) chicken-referenced multiple alignment of the genomes of 77 vertebrate species. **(A, C)** The phastCons scores show no difference at the p<0.0001 level across neuronal-, glial- and endothelial cell-specific genes, except with phastCons77way score in neuron and glial cell-specific genes defined with the Saunders dataset, where neuron-specific genes' promoters has significantly higher dN/dS (aka, more conserved) than that of glial cell-specific genes. The p-values reported are from Mann-Whitney U tests; sample size n (i.e., number of cell type-specific genes with high-coverage phastCons) are also reported; in the middle boxes of enhanced box plots, median and

308 95% confidence interval (in parentheses) are reported. **(B, D)** Virtually no explanatory power was
309 discovered between log-transformed average dN/dS ratios and phastCons scores of brain cell type-
310 specific genes with linear regression (all R^2 less than 0.02). R^2 and p-value for the linear fits are
311 reported.

312 The 2,000 bp sequence upstream of a coding region functions does however not necessarily
313 function as its promoter. In contrast, chromatin accessibility assays reveal functionally defined
314 cis-regulatory regions, which have been performed in single cell nuclei of neurons and non-
315 neuronal cells in the mouse cerebrum[31]. We calculated the average phastCons scores for these
316 cis-regulatory regions, again sourced from the UCSC Genome Browser[29], using a multiple
317 alignment of 60 vertebrate species centered around the mouse GRCm38 reference genome
318 (phastCons60ways). In line with our dN/dS observations, we found that functionally defined
319 neuronal cis-regulatory DNA elements have significantly higher phastCons than non-neuronal cis-
320 regulatory DNA elements (**Figure 5**), which indicates that both neuronal-specific cis-regulatory
321 DNA elements and coding sequences are more highly conserved than non-neuronal elements
322 and sequences.

323

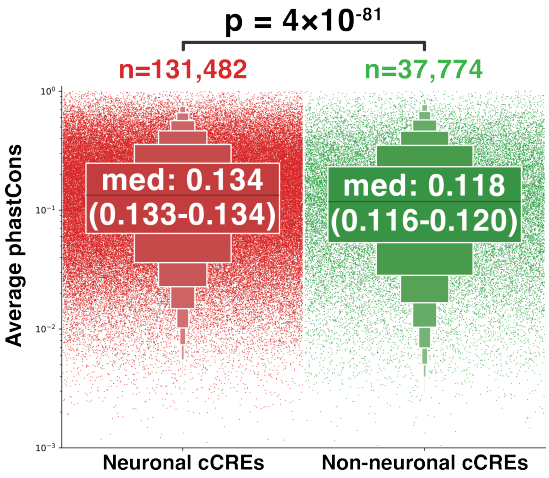

Figure 5. Neuronal cis-regulatory elements (cCREs) have higher phastCons (mouse-centered phastCons60way) than non-neuronal cCREs, indicating they are evolutionarily more conserved. The p-value is reported for Mann-Whitney U test; n: sample size, i.e., number of cCREs; median of phastCons as well as the 95% confidence interval in parentheses reported in the middle box of the enhanced boxplots.

331 We then tested whether the high conservation of neuron-specific sequences is associated with
332 high levels of gene expression in these cells compared to other cell types. Crossing dN/dS scores
333 for each gene with average expression levels per cell defined as fragments per kilobase of
334 transcript sequence per million mapped fragments[20] (FPKM) or as unique molecular identifiers
335 (UMI) counts[22,23], we found that gene expression level has virtually no explanatory power over
336 dN/dS across genes (all three R^2 values smaller than 0.04; **Figure 6**). Additionally, ANCOVA tests
337 for all three datasets reveal statistically significant differences in dN/dS ratios between neuronal,
338 endothelial/vascular cells, and glial cell types even after controlling for expression level (Error!
339 Reference source not found.), with lower dN/dS scores for neuron-specific genes compared to non-
340 neuronal-specific genes with similar expression levels (**Figure 6**).
341

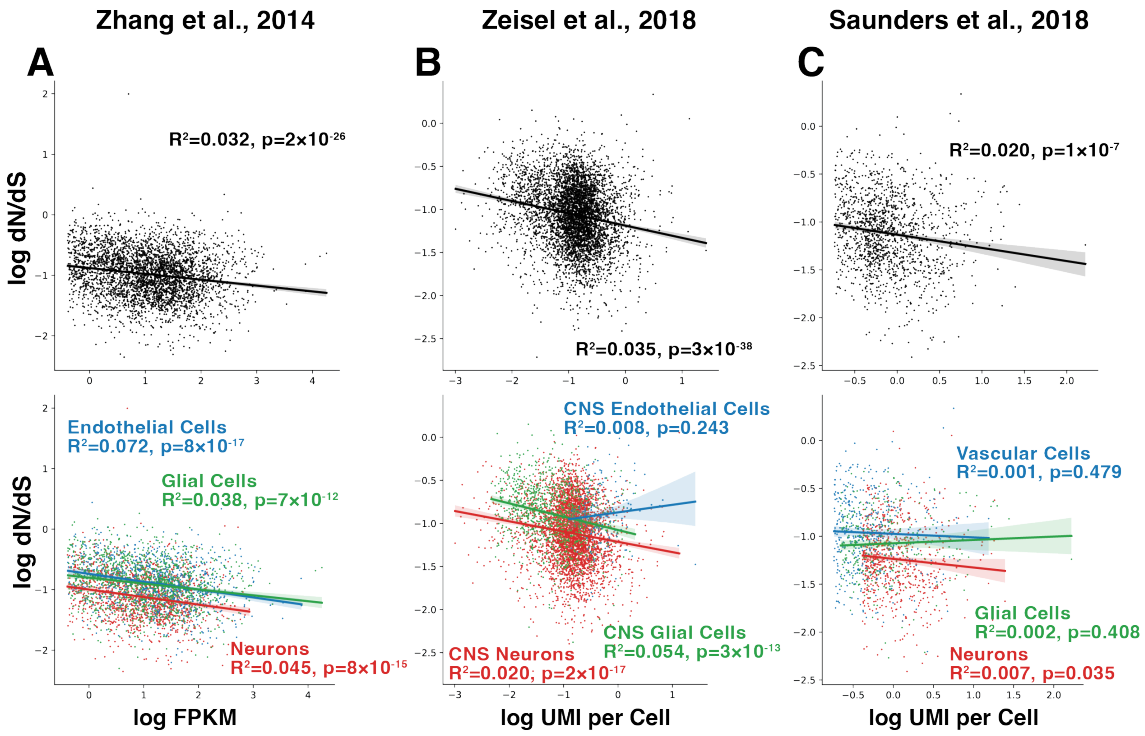

Figure 6. Expression level has virtually no explanatory power over dN/dS of cell type-specific genes, and neuron-specific coding sequences have lower dN/dS values than glial-specific sequences of similar expression level. The upper panel is all cell type-specific genes' dN/dS and their expression level in corresponding cell types. The lower panel is the same data points color coded with cell type. R^2 and p-value for each linear regression between log-transformed dN/dS ratios and expression levels are labeled.

It is paradoxical that neuron-specific coding sequences and cis-regulatory DNA elements are more evolutionarily conserved than their non-neuronal counterparts given the much more evolutionarily diverse morphology and cell sizes of neurons, a paradox that indicates that the source of neuronal cell size variability lies elsewhere than in the genome. How could diversity in mammalian neuronal cell size arise both in development and in evolution without building on genomic diversity? We hypothesize, building on our recent demonstration that energy supply to the adult brain is limiting to brain metabolism[42], that in a scenario where neuronal excitation is energetically costly[43] and the metabolic rate of the developing brain is also supply-limited, there should be a trade-off between neuronal firing rates and growing cell size, such that neurons with lower firing rates would be the ones to have more energy available for cell growth and thus attain larger adult cell sizes.

To test the hypothesis that variations in cell size should be correlated with (gene-based or not) neuron-specific differences in firing activity, such that larger neurons have lower firing rates, we examined the relationship between cell volume and firing rate across human cortical neurons in the Allen Brain Atlas Cell Type Database[44], which covers a range of more than 30-fold variation in reconstructed cell volume, including all arbors in the tissue (but no spines). We restricted our analysis to human cortical neurons, 75% of which in the database are excitatory, because the vast majority of mouse cortical neurons in the database were inhibitory. We analyzed all 131 human cortical neurons with both morphological reconstruction and electrophysiological recordings available and calculated their cell volume in the tissue (including soma and all dendritic and axonal arbors) and representative firing rates (see “Materials and Methods”). We found that firing rate and cell size are indeed negatively correlated amongst excitatory neurons (Spearman $\rho=-0.59$, $p<0.0001$; **Figure 7**), such that larger neurons in the human cerebral cortex have significantly lower firing rates.

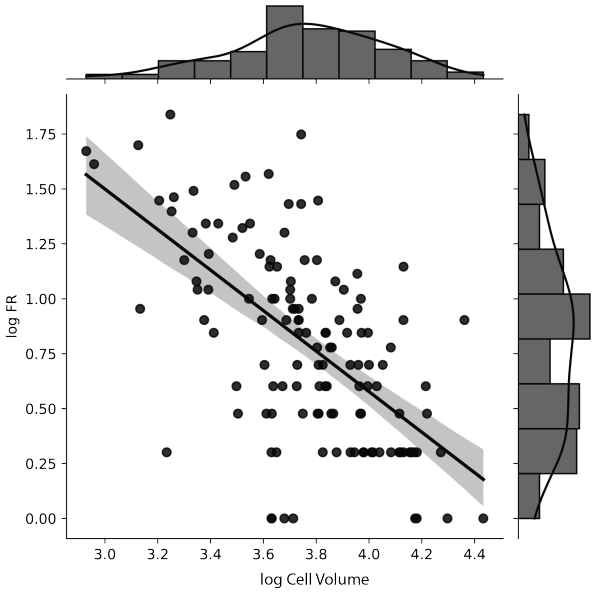

Figure 7. Firing rates and cell volume are negatively correlated across individual neurons in the human cerebral cortex. Linear regression on log-transformed firing rate and cell volume: $y = -0.92x + 4.27$; Pearson's R=-0.61, p=6.814 × 10^{-15}; Spearman's ρ=-0.59, p=1.13 × 10^{-13}.

Discussion

Our initial motivation for this study was to explain the source of the enormous cell size and morphological diversity specific to neurons. Here we refute our initial hypothesis that the enormous evolutionary diversity in neuronal cell size (and therefore density) is due to higher variability in neuron-specific DNA sequences (protein-coding or regulatory) accumulated from mutations in molecular evolution, compared to other cell types. Although it has long been established that differences in protein-coding DNA sequences cannot explain all the morphological differences cross-species[45], changes in gene expression and regulation are believed to be major players in establishing a species' biological uniqueness[46]. However, we find evidence that, across vertebrate species, both neuron-specific coding sequences and cis-regulatory sequences are remarkably more conserved than their endothelial cell- and glial cell-specific counterparts, a higher conservation that cannot be explained simply by higher levels of expression, contrary to what has been suggested for protein-coding genes in yeast[12,13]. Our findings go beyond an earlier discovery that brain-specific genes as a whole are highly conserved[10] by showing that it is specifically the neuron-specific sequences that admit very little evolutionary variation, showing as much conservation as ATPases, the benchmark for evolutionary conservation[33]. Importantly, our findings are reproducible across four methods of defining cell type-specific gene expression in brain cells that range from quantifying gene expression levels in cells isolated by immunopanning to unsupervised clustering of single cell RNA expression, despite the imperfect alignment across

400 methods, as shown in the Venn diagram in **Supplementary Figure S4**. Additionally, we show that
401 the distinctly higher conservation of neuron-specific coding sequences over glial-specific coding
402 sequences occurred in parallel in both reptilian (including birds) and mammalian brain evolution,
403 which implies that fundamental constraints to variation of neuron-specific sequences have been at
404 work across more than 300 million years of evolution[47]. Outside the nervous system, we only find
405 similarly low dN/dS values of genes expressed specifically in other tissues with excitable cells (heart
406 and musculature; **Figure 3** and **Table 2**). Interestingly, we finding that the most conserved neuron-
407 specific genes are enriched in transport-related and signal transduction-related GOs, categories
408 that include genes involved in mediating neuronal excitability, such as ionic channels and
409 metabotropic receptors (**Supplementary Table S9**). We thus hypothesize that molecular pathways
410 associated with electrical excitability events, such as action potentials and membrane
411 repolarization, are particularly sensitive to disruptions that can lead to excitotoxic cell death (for
412 example, calcium overload from prolonged opening of voltage-gated calcium channels leading to
413 mitochondrial release of cytochrome c, triggering the cell death pathway[48]). We postulate that, in
414 the nervous system, this sensitivity strongly restricts the universe of possible functional
415 configurations of those genes expressed specifically in neurons, but not in the other, non-excitable
416 cells that compose brain tissue.

417 Our results have the important implication that the enormous morphological and functional
418 diversity of neurons within and across mammalian brains arises in a context of strong regulatory
419 and coding sequence conservation. Thus, unless post-transcriptional and/or post-translational
420 mechanisms is responsible for neuronal cell volumes that are at least 500 times more variable than
421 non-neuronal cell volume in adults[3], or these are due to evolutionary variation in a very particular
422 subset of genes, both cross-species and within-brain differences in neuronal size must arise
423 through self-organizing mechanisms that build on cell-intrinsic variations in levels of gene
424 expression.

425 However it happens, one expected requirement for the development and evolution of larger
426 neuronal cell sizes is higher energy use, simply because of the larger volume of biosynthesis that
427 builds larger dendritic arbors, axons, and neuronal cell bodies[49]. In neurons, the total energy costs
428 of cell function are compounded by the inflating energetic cost of individual action potentials as
429 neurons become larger, due to the larger membrane capacitance and ATP used to reestablish
430 membrane potential, and further compounded by increased numbers of synapses – unless levels
431 of activity can undergo compensatory regulation[43,50]. Importantly, we have recently demonstrated
432 that brain energetics is constrained by a limited energy supply from blood flow that depends on
433 capillary density in the brain[42,51], such that oxygen delivery rates have very little room to increase
434 even with an increase in blood flow to local tissue due to oxygen diffusivity from the capillary to

435 brain tissues negatively correlating to blood flow[42]. As a result, neurons within a territory supplied
436 by the same capillary and astrocyte share the same limiting energy and oxygen supply.

437 It is in the context of such a limiting energy supply extending to the developing brain that we
438 propose that diversity in neuronal cell size arises from the imposed balance between excitability
439 (firing activity) and cell growth. Specifically, we hypothesize that in a scenario where neuronal
440 excitation is energetically costly[43] and the metabolic rate of the developing brain is also supply-
441 limited, there should be a trade-off between neuronal firing rates and growing cell size, such that
442 neurons with lower firing rates (for instance due to varying firing rate set points, which are
443 characteristic of each neuron through mechanisms yet to be specified[52,53]) would be the ones to
444 have more energy available for cell growth and thus attain larger adult cell sizes. We further
445 propose that differences in combinatorial gene expression across individual neurons in early
446 development[6] directly modulate neuronal excitability and indirectly affect the maximal cell size that
447 a neuron can attain given a limiting blood supply. Such a central role of excitability in self-regulating
448 neuronal cell size would solve the seeming paradox of how come neurons are the only single-
449 nucleus cell type that occurs in a wide range of cell sizes within individuals, when all cells share the
450 same genome (that is, other than variation from aneuploidies[54]).

451 Finally, we speculate that species- and clade-specific supply-limited energetics of the
452 developing nervous system during those stages when neurons are in the process of establishing
453 their adult cell sizes links together cell size variation in development and in evolution. For instance,
454 any factors that modulate the blood supply to the uterus during gestation might impact the rate at
455 which energy is supplied to the embryo, and thus allow for a larger range of adult neuronal cell
456 sizes to be attained through the self-regulation of neuronal cell size proposed above. What we
457 propose is thus a testable mechanism of self-regulation of neuronal cell size that solves the
458 seeming paradox of high variability in neuronal cell size in the presence of extreme conservation
459 of neuron-specific sequences.

460 To summarize, we propose that it is excitability that allows neurons to be uniquely variable in
461 their cell size amongst single-nucleated cells. Neurons are continuously and rapidly excitable
462 thanks to their ability to regenerate the resting membrane potential, without which cells would die[55].
463 The fact that only neurons can dynamically vary their activity and thus the rate of energy use is, we
464 propose, what sets neurons apart from other cells in the body in their enormous elasticity in
465 neuronal cell volume (soma and all neurites included). Regulation of neuronal cell size through
466 electrical excitability would also explain the surprisingly extreme level of conservation of neuron-
467 specific genes, which we show to be on par with the highly conserved ATPase genes. We submit
468 that viability under conditions of fast and relentless excitability is restricted to cells that express only
469 very particular protein sequences, and most evolutionary tampering with them is swiftly met with
470 cell death.

Materials and Methods

Definition of cell-type-specific genes

Cell type-specific genes were selected based on gene expression levels in different cell types in the brain according to each of four different studies, configuring four datasets that we examined to test for congruity in our results. First, we used data from a mouse brain cell type enrichment study from the Ben Barres lab (*Zhang et al., 2014*)[20] with expression levels in FPKM for 22,458 genes in neurons; astrocytes; oligodendrocytes in three states (precursor cell, newly formed, myelinating); microglia; and endothelial cells.[*] For the purpose of this study, we averaged the expression level of cells in all three oligodendrocytic states and used the mean FPKM as the expression level for oligodendrocytes. We arbitrarily defined as cell type-specific genes those genes with expression levels in one brain cell type greater than the sum of expression levels in the other four cell types. For example, if a gene's expression level in neurons is greater than the sum of its expression level in astrocytes, microglia, oligodendrocytes, and endothelial cells, we considered this gene to be a neuron-specific gene in the mouse brain. We repeated the same process to determine astrocyte-, oligodendrocyte-, microglia- and endothelial cell-specific genes. We then averaged the FPKM values in astrocytes, oligodendrocytes, and microglial cells and used that value as the expression level in glial cells as a whole. Effectively, that means that for the five-cell-types (NEAMO) analysis we use a threshold of four times the average expression level across four other cell types. We then identified glial cell-specific genes as genes with expression levels in glial cells greater than four times its average expression level in neurons and endothelial cells, arithmetically similar to our threshold for the five-cell-types analysis. Therefore, cell type-specific genes identified with the Barres dataset are exclusive, meaning that neuronal-, glial-, and endothelial cell-specific genes are non-overlapping. Alternatively, using the same database, we identified cell type-expressed genes, are protein-coding genes with their expression level (FPKM) in one cell type greater than 1, whatever their expression level in other cell types.

Three additional scRNA-seq studies were used to select cell type-specific genes to test the generalizability and robustness of our results. The first, a study from the Linnarsson group on the somatosensory cortex and hippocampal CA1 in P21-P31 mice of both sexes, *Zeisel et al.,* 2015, which has nine major cell types after clustering, provided lists of cell type-specific genes for each cell type in their supplement[21]. This study selected cell type-specific genes based on two criteria: 1) the gene was enriched with 99.9% posterior probability in that cell type and 2) the gene is not enriched in any other cell type. To make this analysis comparable to our analysis with the Barres data, we merged the nine cell types into three cell types: neurons (originally CA1 pyramidal

[*] The original data portal https://web.stanford.edu/group/barres_lab/brain_rnaseq.html is no longer accessible, and the interactive portal was transferred to a new site http://www.brainrnaseq.org/.

504 neurons, S1 pyramidal neurons, and interneurons), glial cells (originally oligodendrocytes,
505 astrocytes, microglial cells, and ependymal cells), and capillary cells (originally endothelial cells
506 and mural cells) and did three-group analyses. Like the neurons, the mural cells are also excitable
507 cells[56,57], including vascular smooth muscle cells and pericytes[21] that regulate the rate of blood
508 flow[58]. As single-cell expression level values for *Zeisel et al., 2015* were not disclosed in the dataset,
509 we could not include a regression analysis between dN/dS and FPKM or UMI count for it as we did
510 for other studies in **Figure 6**.

511 The Linnarsson group later published a more comprehensive scRNA-seq dataset of the entire
512 mouse nervous system, including not only the CNS but also the PNS and the enteric nervous
513 system (*Zeisel et al., 2018*)[22]. After dimensionality reduction through manifold learning and
514 clustering, they generated a curated classification of cells in the mouse nervous system, with 265
515 clusters in total. These clusters were automatically assigned to one of seven major classes, namely
516 neurons, oligodendrocytes, astrocytes, ependymal cells, peripheral glia, immune cells, and
517 vascular cells. We accessed expression level (UMI counts) and additional meta-data of all 265 cell
518 clusters through a multidimensional matrix in loom format[†]. One data layer that *Zeisel et al., 2018*
519 provided is a "trinary score" which indicates whether a gene can be confidently categorized as
520 expressed in a specific cluster. We used a 0.999 trinary score cut-off to define cell type-specific
521 genes. For CNS neuronal, CNS glial, and CNS endothelial cell-specific genes, if a gene has a
522 "trinary score" greater or equal to 0.999 in at least one cluster under the umbrella cell type, but not
523 in any cluster under the other two umbrella cell types, it is defined as cell type-specific. To make
524 this three-cell-type analysis comparable to our analysis using the Barres database, we included
525 microglial clusters as glial clusters. With the same 0.999 trinary score threshold, we also defined
526 cell type-specific genes on umbrella clusters defined for the entire nervous system (i.e., not just
527 CNS) by *Zeisel et al., 2018*, recorded as Rank 1 Taxonomy in their metadata labeling. In this case,
528 we use the Linnarsson criterion and clustered microglial clusters into the immune cell group.

529 The third scRNA-seq dataset (and the fourth RNA-seq study) we used was a study from the
530 McCarroll group (*Saunders et al., 2018*)[23], which developed the Drop-seq sequencing method, that
531 collected ~700,000 single cells in nine regions of the adult mouse brain and clustered them into
532 565 clusters (referred to as "metacells") that are transcriptionally distinct[‡]. The "metacells" can be

---

[†] The aggregated loom file can be downloaded from https://storage.googleapis.com/linnarsson-lab-loom/l5_all.agg.loom. The data portal can be accessed at http://mousebrain.org/. This file type can be processed with the loompy python package developed the Linnarsson group: https://linnarssonlab.org/loompy/.

[‡] Data retrieved from stable data portal http://dropviz.org/. Aggregate UMI counts in https://storage.googleapis.com/dropviz-downloads/static/metacells.BrainCellAtlas_Saunders_version_2018.04.01.csv. Annotations for the 565 clusters in https://storage.googleapis.com/dropviz-downloads/static/annotation.BrainCellAtlas_Saunders_version_2018.04.01.xlsx.

533 categorized into three master classes: neurons, glial cells, or vascular cells. For each "metacell",
534 we calculated an "expression percentage" by dividing each gene's aggregate UMI counts by the
535 total UMI counts in that "metacell". We then define cell type-specific genes as genes with more than
536 0.1% of the total expression of metacells annotated to be in that cell type that at the same time
537 have the expression percentage value greater than four times the average expression percentage
538 in other cell types.

539 **Benchmark Genes**

540 ATPase genes, housekeeping genes, immune genes, and MHC genes were used as
541 benchmarks in this study. Human "ATPase genes" used were retrieved from the HUGO Gene
542 Nomenclature Committee database[59§]; mouse, rat, and chicken ATPase genes are orthologs of
543 the human genes. "Housekeeping genes" used were human housekeeping genes[60**] or mouse,
544 rat, and chicken orthologs of these genes for the other three reference genome. "Immune genes"
545 were mouse or human genes associated with GO:0045087 (innate immune response), selected
546 from the InnateDB database[61††]; rat and chicken immune genes were retrieved from RGD[62] and
547 Ensembl 98[25], respectively. For the mouse and human reference genomes, "MHC genes" were
548 retrieved from the literature[63]; for the rat and chicken reference genomes, "MHC genes" were
549 genes associated with the GO term MHC protein complex from the RGD database or from
550 Ensembl 98, respectively.

551 We defined liver-, kidney- (metanephros-), lung-, skin-, brain-, heart-, pancreas-, and
552 musculature-specific genes as genes with expression recorded only in that organ and no other
553 organ in the MGI mouse gene expression database (GXD)[37].

554 **The dN/dS Ratios**

555 Pairwise dN/dS ratios of the genes of interest for each of 92 mammalian species (excluding
556 the reference species) using as reference the mouse (GRCm38.p6) or human (GRCh38.p13) or
557 rat (Rnor 6.0) genome and were calculated from dN and dS values retrieved from the Ensembl 98
558 database[25‡‡]. A similar procedure was used for the genes of interest expressed in each of 31
559 sauropsidian (including birds) species using the chicken reference genome (GRCg6a) as

---

§ Human ATPase gene names from the HGNC portal at https://www.genenames.org/data/genegroup/#!/group/412. Mouse orthologs matched to human ATPase genes with the online BioMart GUI of Ensembl 98; same for mouse orthologs of other human genes below.
** Retrieved from the stable link: https://www.tau.ac.il/~elieis/HKG/.
†† Retrieved from the stable link: https://www.innatedb.com/annotatedGenes.do?type=innatedb.
‡‡ We access the Ensembl data with a custom bash script that loop through available species with XML BioMart RESTful access. This data can also be accessed through the online BioMart GUI (https://useast.ensembl.org/biomart/martview). Older versions of Ensembl are now archived. To access version 98 used in this study, go to http://sep2019.archive.ensembl.org/index.html.

560 reference. Orthologues that are not one-to-one orthologues across pairs of species were filtered
561 out for that target species.

562 **Gene Ontology**

563 Gene Ontology (GO) slim terms associated with each gene of interest are biological
564 processing slim ontology terms restricted to the mouse species. The genes mapped to each GO
565 slim term were retrieved from the Princeton University Lewis-Sigler Institute for Integrative
566 Genomics' Generic GO Term Mapper[38,§§]. The ontology aspects selected were "biological
567 processing" and restricted to the species *Mus musculus* (MGI). Additional ontology terms
568 "regulation of cell size", "regulation of membrane potential", "membrane depolarization",
569 "membrane hyperpolarization", and "membrane repolarization", which are not included in the slim
570 terms, were also added on top of the slim terms because of our interest in cell size regulation and
571 potential roles of unique properties of neurons.

572 **Regulatory Elements' PhastCons Scores**

573 "Promoter sequences" of cell type-specific genes were defined as the 2,000 base pairs of
574 DNA upstream of each coding sequence in the mouse or chicken genome. Coordinates of these
575 promoters on the autosomes, sex chromosomes, or mitochondria genome were calculated based
576 on the coordinates of the coding sequence on the mouse or chicken genome, and whether the
577 gene is on the positive or negative strand. We downloaded bigWig files for both
578 phastCons60way[***] (from a mammal-rich vertebrate multiple alignment centered around the
579 mouse genome) and phastCons77way[†††] (from a sauropsid-rich multiple alignment centered
580 around the chicken genome), which contain the phastCons scores for all available loci on the
581 mouse and chicken genome. We then fed the start and end position of the promoters and the
582 bigWig files to the Unix executable program bigWigSummary[64] to estimate the average
583 phastCons scores of each promoter, and proceeded to filter out entries with less than 80%
584 phastCons coverage for quality control. These phastCons scores may also be retrieved with the
585 UCSC Genome Browser's Table Browser utility[29]. For the list of species included in the two
586 multiple alignments, see **Error! Reference source not found.**.

587 The coordinates for neuronal and non-neuronal candidate cis-regulatory sequences were
588 defined based on a chromatin accessibility study on the mouse cerebrum using single nucleus
589 ATAC-seq[31]. Neuronal cis-regulatory DNA elements are those associated with both GABAergic
590 and glutamatergic neurons but not non-neuronal cells; non-neuronal cis-regulatory DNA elements

---

[§§] Access at https://go.princeton.edu/cgi-bin/GOTermMapper on October 13th, 2019.
[***] Stable link at: http://hgdownload.cse.ucsc.edu/goldenpath/mm10/phastCons60way/.
[†††] Stable link at: https://hgdownload.soe.ucsc.edu/goldenPath/galGal6/phastCons77way/.

591 are those associated with non-neuronal cell types but not associated with any of the neuronal cell
592 types. The coordinates of these cis-regulatory elements and the mouse-centered
593 phastCons60way bigwig file were similarly fed to bigWigSummary for calculating the average
594 phastCons scores of each element. After filtering out cis-regulatory DNA elements with less than
595 80% coverage of phastCons scores for quality control, we analyzed the phastCons of a resulting
596 131,482 neuronal cis-regulatory DNA elements and 37,774 non-neuronal cis-regulatory DNA
597 elements.

598 **Single Cell Electrophysiology and Morphology Data**

599 Neuronal cell volume and representative firing rate are calculated based on the Allen Brain
600 Atlas Cell Type Database[44]. Data from 156 cerebral cortical neurons from 35 human subjects that
601 had both morphology reconstructions and electrophysiology recordings are publicly available,
602 which we accessed through the Allen SDK[65] with our custom Python scripts. Of the 156 human
603 neurons, 131 has a valid "representative sweep" (see below). The human cortex database
604 consists of both inhibitory (with dendrite type "aspiny") and excitatory neurons (with dendrite type
605 "sparsely spiny" or "spiny"), but predominantly the latter. The mouse counterpart of the dataset
606 consists almost exclusively of inhibitory neurons and for this reason is not included in this study.

607 From the downloaded SWC files denoting the morphology of each cell, we calculated whole
608 cell volumes (soma and all dendritic and axonal arbors in the volume) with the Python package
609 "neuron_morphology", which assumes that each compartment is a conic frustum[‡‡‡]. Dendritic
610 spines are not included in original morphological models and therefore are not included in our
611 volume estimates.

612 For those cells that had morphological reconstructions in the database, we retrieved and
613 studied the quality-controlled current clamp recordings. For each neuron, we first identified the
614 rheobase (the lowest long-square stimulus amplitude at which a cell starts to have spikes) from
615 the cell electrophysiology metadata. We then used the Allen criterion of selecting the current
616 clamp recording with stimulus amplitude 40 to 60 pA above rheobase to select a representative
617 long-square sweep for each neuron. This gave us 131 human cortical neurons with a qualified
618 representative sweep (32 inhibitory neurons and 99 excitatory neurons). We then calculated the

[‡‡‡] The pre-calculated volume included in the cell metadata was estimated with L-Measure, which assumes each component to be a cylinder. Since the SWC files include the radius at each node, we reasoned that the conic frustum-based volume estimation would better capture each neuron's volume in 3D space.

firing rate as the number of spikes during the 1-second stimulation interval for that representative sweep for each cell[§§§].

Statistical Analysis and the Sharing of Reproducible Codes

We conducted all statistical analyses with custom Python scripts, which are publicly available at https://github.com/VeritatemAmo/neuron-glia-dNdS together with all codes and data needed for replicating this study.

Because dN/dS are log-normally distributed, we chose to use a median-based description of dN/dS values. Confidence intervals are 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 package's "stats.binom.interval" package, with alpha set to 0.95. After ranking the genes in each group based on dN/dS value, the genes at the low and high interval threshold of each group are set as the lower and higher bound of the confidence interval.

Mann Whitney U tests were performed with the "mwu" function of the pingouin package or "stats.mannwhitneyu" function of the SciPy package. The p-values that these packages return are in the float64 double-precision floating-point format. This means the smallest positive number that can be represented is $4.9406564584124654 \times 10^{-324}$. Therefore, when we get a p-value of "0.0" from these packages, we report it as $p<5 \times 10^{-324}$ in this paper. We try to include a common language effect size (CLES) to assist our readers in understanding the result whenever possible, but with the huge amount of entries for all 92 species pairwise dN/dS ratios (i.e., when we use raw pairwise dN/dS ratios instead of dN/dS ratios average across the 92 mammalian species), that calculation of CLES exceed our computational capability, and the standard U statistics were reported instead in **Error! Reference source not found.**.

For contingency analyses, we developed a custom statistical and visualization analysis pipeline. Essentially, we generated a crosstable with the Python crosstab function from the pandas library between a cell type and a gene ontology. Then, with the SciPy library, odds ratios were calculated with the "fisher_exact" function, and χ^2 values and p values were generated using the "chi2_contingency" function. The contingency table between each GO term and the genes with the lowest or highest 25% dN/dS ratios, along with distribution plots of cell type-specific genes that belong to each GO term, are included in **Supplementary Data S3** and Error! Reference source not found..

Before running linear regression analyses, we first confirmed the normality of the distribution of log-transformed phastCons and dN/dS values with Q-Q plots. We then ran a least square

[§§§] The pre-calculated firing rate in the Allen electrophysiology metadata is the inverse of inter-spike interval, which yields artificially high firing rates for bursting neurons. Because we are interested in estimating energetic cost of firing over longer time intervals, we calculated firing rates as indicated instead.

651 linear regression of log-transformed phastCons and dN/dS scores of cell-type-specific genes.
652 Since zero values cannot be log-transformed, they were excluded from the linear regression.
653

Acknowledgments

We thank Profs. John A Capra (now at UCSF) and Jon H Kaas for in-person advice and the Vanderbilt University's Advanced Computing Center for Research and Education (ACCRE) for technical support. We thank Drs. Nathan Gouwens and Jeremy Miller at the Allen Brain Institute for assistance in using their dataset. We thank Prof. Arpiar Saunders for assisting our understanding and providing further information about their DropViz data. We thank Prof. Amit Zeisel for helping us interpret the trinary scores reported in their 2018 paper. We thank Prof. Ye Zhang for directing us to the new data portal of the Barres dataset. This work was supported exclusively by Vanderbilt University start-up funds to SHH.

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

670 **References**

671 1. Savage, V. M. *et al.* Scaling of number, size, and metabolic rate of cells with body size in
672 mammals. *Proc. Natl. Acad. Sci.* **104**, 4718–4723 (2007).
673 2. Schmidt-Nielsen, K. *Scaling: why is animal size so important?* (Cambridge University Press,
674 1984).
675 3. Mota, B. & Herculano-Houzel, S. All brains are made of this: a fundamental building block of
676 brain matter with matching neuronal and glial masses. *Front. Neuroanat.* **8**, 127 (2014).
677 4. Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and
678 mouse. *Nature* **598**, 111–119 (2021).
679 5. BRAIN Initiative Cell Census Network (BICCN). A multimodal cell census and atlas of the
680 mammalian primary motor cortex. *Nat. 2021 5987879* **598**, 86–102 (2021).
681 6. Yuste, R. *et al.* A community-based transcriptomics classification and nomenclature of
682 neocortical cell types. *Nat. Neurosci.* **23**, 1456–1468 (2020).
683 7. Kriegstein, A. & Alvarez-Buylla, A. The Glial Nature of Embryonic and Adult Neural Stem
684 Cells. *Annu. Rev. Neurosci.* **32**, 149–184 (2009).
685 8. Herculano-Houzel, S. The glia/neuron ratio: How it varies uniformly across brain structures
686 and species and what that means for brain physiology and evolution. *GLIA* **62**, 1377–1391
687 (2014).
688 9. Herculano-Houzel, S. & Dos Santos, S. You Do Not Mess with the Glia. *Neuroglia* **1**, 193–219
689 (2018).
690 10. Biswas, K., Chakraborty, S., Podder, S. & Ghosh, T. C. Insights into the dN/dS ratio
691 heterogeneity between brain specific genes and widely expressed genes in species of
692 different complexity. *Genomics* **108**, 11–17 (2016).

693 11. Hu, G., Li, J. & Wang, G. Z. Significant Evolutionary Constraints on Neuron Cells Revealed by
694 Single-Cell Transcriptomics. *Genome Biol. Evol.* **12**, 300–308 (2020).
695 12. Drummond, D. A., Raval, A. & Wilke, C. O. A Single Determinant Dominates the Rate of Yeast
696 Protein Evolution. *Mol. Biol. Evol.* **23**, 327–337 (2006).
697 13. Pál, C., Papp, B. & Hurst, L. D. Highly Expressed Genes in Yeast Evolve Slowly. *Genetics* **158**,
698 927–931 (2001).
699 14. Tuller, T., Kupiec, M. & Ruppin, E. Evolutionary rate and gene expression across different
700 brain regions. *Genome Biol.* **9**, 1–11 (2008).
701 15. Pembroke, W. G., Hartl, C. L. & Geschwind, D. H. Evolutionary conservation and divergence
702 of the human brain transcriptome. *Genome Biol.* **22**, 1–33 (2021).
703 16. Herculano-Houzel, S. Theropod dinosaurs had primate-like numbers of telencephalic
704 neurons. *J. Comp. Neurol.* **531**, 962–974 (2023).
705 17. Olkowicz, S. *et al.* Birds have primate-like numbers of neurons in the forebrain. *Proc. Natl.*
706 *Acad. Sci. U. S. A.* **113**, 7255–7260 (2016).
707 18. Azevedo, F. A. C. *et al.* Equal numbers of neuronal and nonneuronal cells make the human
708 brain an isometrically scaled-up primate brain. *J. Comp. Neurol.* **513**, 532–541 (2009).
709 19. Herculano-Houzel, S., Catania, K., Manger, P. R. & Kaas, J. H. Mammalian Brains Are Made of
710 These: A Dataset of the Numbers and Densities of Neuronal and Nonneuronal Cells in the
711 Brain of Glires, Primates, Scandentia, Eulipotyphlans, Afrotherians and Artiodactyls, and
712 Their Relationship with Body Mass. *Brain. Behav. Evol.* **86**, 145–163 (2015).
713 20. Zhang, Y. *et al.* An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons,
714 and Vascular Cells of the Cerebral Cortex. *J. Neurosci.* **34**, 11929–11947 (2014).

715 21. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-
716 seq. *Science* **347**, 1138–1142 (2015).
717 22. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-
718 1014.e22 (2018).
719 23. Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult
720 Mouse Brain. *Cell* **174**, 1015-1030.e16 (2018).
721 24. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood.
722 *Bioinformatics* **13**, 555–556 (1997).
723 25. Cunningham, F. *et al.* Ensembl 2019. *Nucleic Acids Res.* **47**, D745–D751 (2019).
724 26. An, J. Y. *et al.* Genome-wide de novo risk score implicates promoter variation in autism
725 spectrum disorder. *Science* **362**, eaat6576 (2018).
726 27. Shi, W. & Zhou, W. Frequency distribution of TATA Box and extension sequences on human
727 promoters. *BMC Bioinformatics* **7**, S2 (2006).
728 28. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
729 29. Karolchik, D. *et al.* The UCSC table browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–
730 D496 (2004).
731 30. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast
732 genomes. *Genome Res.* **15**, 1034–1050 (2005).
733 31. Li, Y. E. *et al.* An atlas of gene regulatory elements in adult mouse cerebrum. *Nature* **598**,
734 129–136 (2021).
735 32. Hofmann, H., Wickham, H. & Kafadar, K. Letter-Value Plots: Boxplots for Large Data. *J.*
736 *Comput. Graph. Stat.* **26**, 469–477 (2017).

737 33. Gogarten, J. P. & Taiz, L. Evolution of proton pumping ATPases: Rooting the tree of life.
738 *Photosynth. Res.* **33**, 137–146 (1992).
739 34. Zhang, L. & Li, W. H. Mammalian Housekeeping Genes Evolve More Slowly than Tissue-
740 Specific Genes. *Mol. Biol. Evol.* **21**, 236–239 (2004).
741 35. Shultz, A. J. & Sackton, T. B. Immune genes are hotspots of shared positive selection across
742 birds and mammals. *eLife* **8**, (2019).
743 36. Piertney, S. B. & Oliver, M. K. The evolutionary ecology of the major histocompatibility
744 complex. *Heredity* **96**, 7–21 (2006).
745 37. Smith, C. M. *et al.* The mouse Gene Expression Database (GXD): 2019 update. *Nucleic Acids*
746 *Res.* **47**, D774–D779 (2019).
747 38. Boyle, E. I. *et al.* GO::TermFinder--open source software for accessing Gene Ontology
748 information and finding significantly enriched Gene Ontology terms associated with a list of
749 genes. *Bioinformatics* **20**, 3710–3715 (2004).
750 39. Rio-Hortega, P. The Microglia. *The Lancet* **233**, 1023–1026 (1939).
751 40. McKenzie, A. T. *et al.* Brain Cell Type Specific Gene Expression and Co-expression Network
752 Architectures. *Sci. Rep.* **8**, 8868 (2018).
753 41. Kosiol, C. *et al.* Patterns of positive selection in six mammalian genomes. *PLoS Genet.* **4**,
754 (2008).
755 42. Herculano-Houzel, S. & Rothman, D. L. From a Demand-Based to a Supply-Limited
756 Framework of Brain Metabolism. *Front. Integr. Neurosci.* **16**, 7 (2022).
757 43. Attwell, D. & Laughlin, S. B. An energy budget for signaling in the grey matter of the brain. *J.*
758 *Cereb. Blood Flow Metab.* **21**, 1133–1145 (2001).
759 44. Allen Institute for Brain Science. Allen Cell Types Database. celltypes.brain-map.org (2015).

760 45. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**,
761 107–116 (1975).
762 46. Sholtis, S. J. & Noonan, J. P. Gene regulation and the origins of human biological uniqueness.
763 *Trends Genet.* **26**, 110–118 (2010).
764 47. Kemp, T. S. Evolution of the mammal-like reptiles. in *The Origin and Evolution of Mammals*
765 (Oxford University Press, 2005).
766 48. Orrenius, S., Zhivotovsky, B. & Nicotera, P. Regulation of cell death: the calcium–apoptosis
767 link. *Nat. Rev. Mol. Cell Biol.* **4**, 552–565 (2003).
768 49. Lane, N. How energy flow shapes cell evolution. *Curr. Biol.* **30**, R471–R476 (2020).
769 50. Herculano-Houzel, S. Scaling of brain metabolism with a fixed energy budget per neuron:
770 Implications for neuronal activity, plasticity and evolution. *PLoS ONE* **6**, e17514 (2011).
771 51. Ventura-Antunes, aLissa & Herculano-Houzel, S. Energy supply per neuron is constrained by
772 capillary density in the mouse brain. *Front. Integr. Neurosci.* **16**, (2022).
773 52. Turrigiano, G. G. & Nelson, S. B. Homeostatic plasticity in the developing nervous system.
774 *Nat. Rev. Neurosci.* **5**, 97–107 (2004).
775 53. Tyssowski, K. M. *et al.* Firing Rate Homeostasis Can Occur in the Absence of Neuronal
776 Activity-Regulated Transcription. *J. Neurosci.* **39**, 9885–9899 (2019).
777 54. Rehen, S. K. *et al.* Constitutional aneuploidy in the normal human brain. *J. Neurosci.* **25**,
778 2176–2180 (2005).
779 55. Lane, N. & Martin, W. F. The origin of membrane bioenergetics. *Cell* **151**, 1406–1416 (2012).
780 56. Burdyga, T. & Borysova, L. Calcium signalling in pericytes. *J. Vasc. Res.* **51**, 190–199 (2014).
781 57. Kamouchi, M. *et al.* Calcium influx pathways in rat CNS pericytes. *Mol. Brain Res.* **126**, 114–
782 120 (2004).

783 58. Hartmann, D. A. *et al.* Brain capillary pericytes exert a substantial but slow influence on
784 blood flow. *Nat. Neurosci. 2021 245* **24**, 633–645 (2021).
785 59. Yates, B. *et al.* Genenames.org: The HGNC and VGNC resources in 2017. *Nucleic Acids Res.*
786 **45**, D619–D625 (2017).
787 60. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–
788 574 (2013).
789 61. Breuer, K. *et al.* InnateDB: Systems biology of innate immunity and beyond - Recent updates
790 and continuing curation. *Nucleic Acids Res.* **41**, D1228–D1233 (2013).
791 62. Smith, J. R. *et al.* The Year of the Rat: The Rat Genome Database at 20: a multi-species
792 knowledgebase and analysis platform. *Nucleic Acids Res.* **48**, D731–D742 (2020).
793 63. Shiina, T., Blancher, A., Inoko, H. & Kulski, J. K. Comparative genomics of the human,
794 macaque and mouse major histocompatibility complex. *Immunology* **150**, 127–138 (2017).
795 64. Harris, R. S. Improved pairwise alignment of genomic DNA. (PennState, 2007).
796 65. Allen Institute for Brain Science. Allen SDK. allensdk.readthedocs.io (2015).
797
798