Current Biology

Parallel gene size and isoform expansion of ancient neuronal genes

Highlights

- Relative gene size is maintained among diverse eukaryotes
- Large neuronal and synaptic genes share ancient origins and sequence constraint
- Many synaptic genes underwent parallel gene size and isoform expansion in animals

Authors

Matthew J. McCoy, Andrew Z. Fire

Correspondence

mjmccoy@stanford.edu (M.J.M.),
afire@stanford.edu (A.Z.F.)

In brief

Many large, multi-isoform genes are expression enriched in the brain and are frequently mutated or misregulated in neurological disorders. By studying gene size evolution, McCoy and Fire show that many neuronal and synaptic genes are ancient and sequence constrained yet have also undergone parallel gene size and isoform expansion in diverse species.

McCoy & Fire, 2024, Current Biology *34*, 1–11
April 22, 2024 © 2024 Elsevier Inc.
https://doi.org/10.1016/j.cub.2024.02.021

CellPress

Current Biology

CellPress

Article

Parallel gene size and isoform expansion of ancient neuronal genes

Matthew J. McCoy[1,3,*] and Andrew Z. Fire[1,2,*]
[1]Department of Pathology, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 94305, USA
[2]Department of Genetics, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 94305, USA
[3]Lead contact
*Correspondence: mjmccoy@stanford.edu (M.J.M.), afire@stanford.edu (A.Z.F.)
https://doi.org/10.1016/j.cub.2024.02.021

SUMMARY

How nervous systems evolved is a central question in biology. A diversity of synaptic proteins is thought to play a central role in the formation of specific synapses leading to nervous system complexity. The largest animal genes, often spanning hundreds of thousands of base pairs, are known to be enriched for expression in neurons at synapses and are frequently mutated or misregulated in neurological disorders and diseases. Although many of these genes have been studied independently in the context of nervous system evolution and disease, general principles underlying their parallel evolution remain unknown. To investigate this, we directly compared orthologous gene sizes across eukaryotes. By comparing relative gene sizes within organisms, we identified a distinct class of large genes with origins predating the diversification of animals and, in many cases, the emergence of neurons as dedicated cell types. We traced this class of ancient large genes through evolution and found orthologs of the large synaptic genes potentially driving the immense complexity of metazoan nervous systems, including in humans and cephalopods. Moreover, we found that while these genes are evolving under strong purifying selection, as demonstrated by low dN/dS ratios, they have simultaneously grown larger and gained the most isoforms in animals. This work provides a new lens through which to view this distinctive class of large and multi-isoform genes and demonstrates how intrinsic genomic properties, such as gene length, can provide flexibility in molecular evolution and allow groups of genes and their host organisms to evolve toward complexity.

INTRODUCTION

Gene size varies among organisms and can change due to the addition of domains to proteins with increasing complexity.[1] Although protein sizes remain consistent among eukaryotes,[2] absolute gene sizes within and among species can vary greatly and have grown particularly large within animals.[3–7] The majority of differences result from expansions of non-coding DNA, specifically within introns.[3,5] Average intron sizes correlate with genome size[8,9] and can impact a range of ecological and cellular processes.[10,11]

The consequences of gene size variation are only beginning to be understood. Many of the largest animal genes are commonly expressed in nervous tissue[7,12–15] and are frequently mutated or misregulated in human conditions such as autism spectrum[12] and Rett syndrome.[13] These genes are particularly enriched for functions at synapses,[12,13] which underlie the precise wiring of nervous systems and are an important character of neurons. Synapses are assembled from multiprotein complexes of diverse protein classes, such as ion channels, receptors, cell adhesion and cytoskeletal proteins, kinases and phosphatases, scaffolding proteins, and signaling molecules.[16] Given that animals have greatly expanded gene sizes relative to other organisms,[7] gene size expansion would have provided numerous opportunities and challenges in the evolution of genes encoding large neuron- and synapse-specific proteins. As examples, larger genes are slower to transcribe[17] and less likely to undergo full duplication while being more likely to exhibit alternative splicing,[6,18] thus providing unusual constraints and flexibilities in the emergence and diversification of neural cell types and synapses. Such properties highlight the particular relevance of elucidating larger-scale genome dynamics and chromosome function in understanding the evolution and function of nervous systems.

Recent tools allow concrete orthology assignments of genes in diverse species.[19,20] This development provides the opportunity to move from averages to individual trajectories of gene and protein size during evolution, including that for large, complex animal genes. Here, we compare the size, age, and architecture of animal genes to provide insight into the origins of molecular diversity and complexity in many animals and their nervous systems.

RESULTS

Relative gene size is preserved among species

To determine the evolutionary origins of large neuronal genes, we set out to define and characterize this set of genes across diverse species. Changes in individual gene size can reflect variance in coding and/or intron content, while overall genome sizes

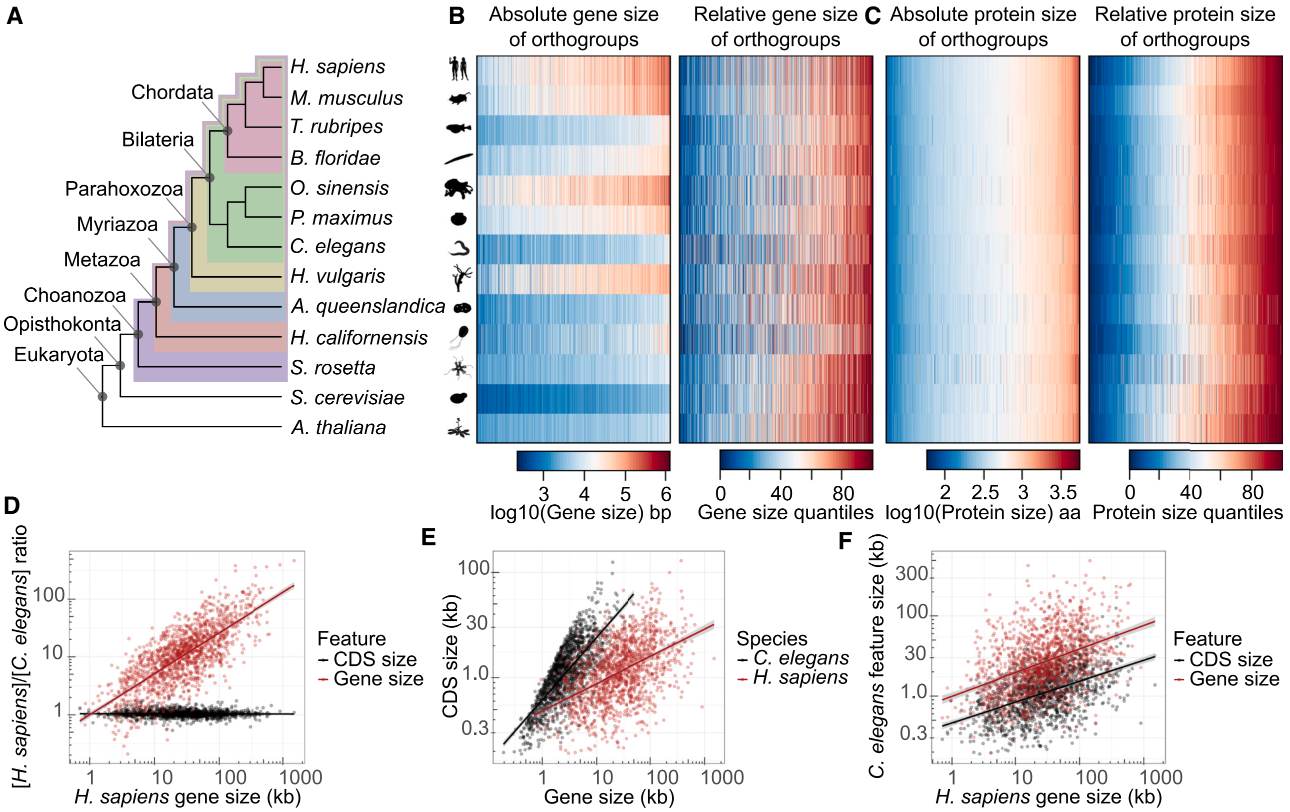

Figure 1. Absolute gene sizes vary by orders of magnitude among diverse species, while relative gene sizes are maintained

(A) Phylogeny of species used in this study, based on chromosome-scale gene linkages.[41]

(B) Comparison of mean gene size per orthogroup across species. (Left) Heatmap of absolute mean gene size ($\log_{10}$ bp) per orthogroup. (Right) Heatmap of relative mean gene size (gene size quantiles) per orthogroup, with each orthogroup binned into 100 quantiles to show the size ranking for the same orthogroups in different species.

(C) (Left) Heatmap of absolute protein size ($\log_{10}$ aa). (Right) Heatmap of relative protein size (protein size quantiles), with each orthogroup binned into 100 quantiles. Orthogroups (columns) in each heatmap are ordered by the average feature (absolute/relative gene/protein size) across all species.

(D–F) Gene and CDS size of Ensembl one-to-one, high-confidence orthologs between *Homo sapiens* and *Caenorhabditis elegans*. Solid lines show linear models with 95% confidence intervals as ribbons. (D) CDS size remains relatively invariant, while gene size varies substantially. Ratios of (*H. sapiens*)/(*C. elegans*) gene and CDS size. (E) *C. elegans* gene and CDS size are both strongly correlated with orthologous gene sizes in *H. sapiens*. (F) Gene size is correlated with CDS size within individual genomes. See also Figure S1, Table S1, and Data S1 and S2.

reflect these as well as intergenic content and gene number. The ratio of introns to intergenic sequences is nearly 1:1 in numerous model animals.[3] Hence, larger animal genomes typically have larger intronic content and thus larger genes.[3,21] Together with previous studies showing that orthologous proteins are encoded by genes with similar-sized coding sequences (CDS),[2] this would suggest that changes in gene sizes are in part a function of changes in genome size. Although previous studies have compared aggregate measures of gene size or coding and non-coding DNA in different species,[3,5,22] gene-by-gene comparisons provide an opportunity to investigate gene size variation during evolution and its impact on gene expression patterns and gene architecture.

We asked whether gene sizes in one species covary with orthologous gene sizes in distantly related species. We addressed this question by comparing rank orders of gene size between species. We focused our analysis on several diverse eukaryotes with chromosome-level genome assemblies, in part because gene annotation quality is related to genome assembly completeness.[3] For this analysis, we identified orthologs and orthogroups (a set of genes from multiple species that descended from a single gene in the last common ancestor) using OrthoFinder,[20] an orthology inference tool that accounts for gene-length bias in detecting orthologs.[22] We then determined absolute and relative gene and protein sizes for each ortholog (see definitions in STAR Methods; Data S1), and compared their averages for each orthogroup among species. Despite orders-of-magnitude variation in absolute gene size, we found that relative gene size is largely maintained across species (Figures 1A, 1B, and S1). This is true not only among vertebrates, which typically have significantly larger genes than invertebrates, but also in comparisons with cephalopods (see *Octopus sinensis* in Figures 1A and 1B), which are of particular interest due in part to their evolution of large and complex nervous systems independent of vertebrates.[23] For the purpose of juxtaposition with gene sizes, Figure 1C displays protein sizes, which are nearly invariant among eukaryotes.[2] These results support the hypothesis that gene sizes are shifting together at the

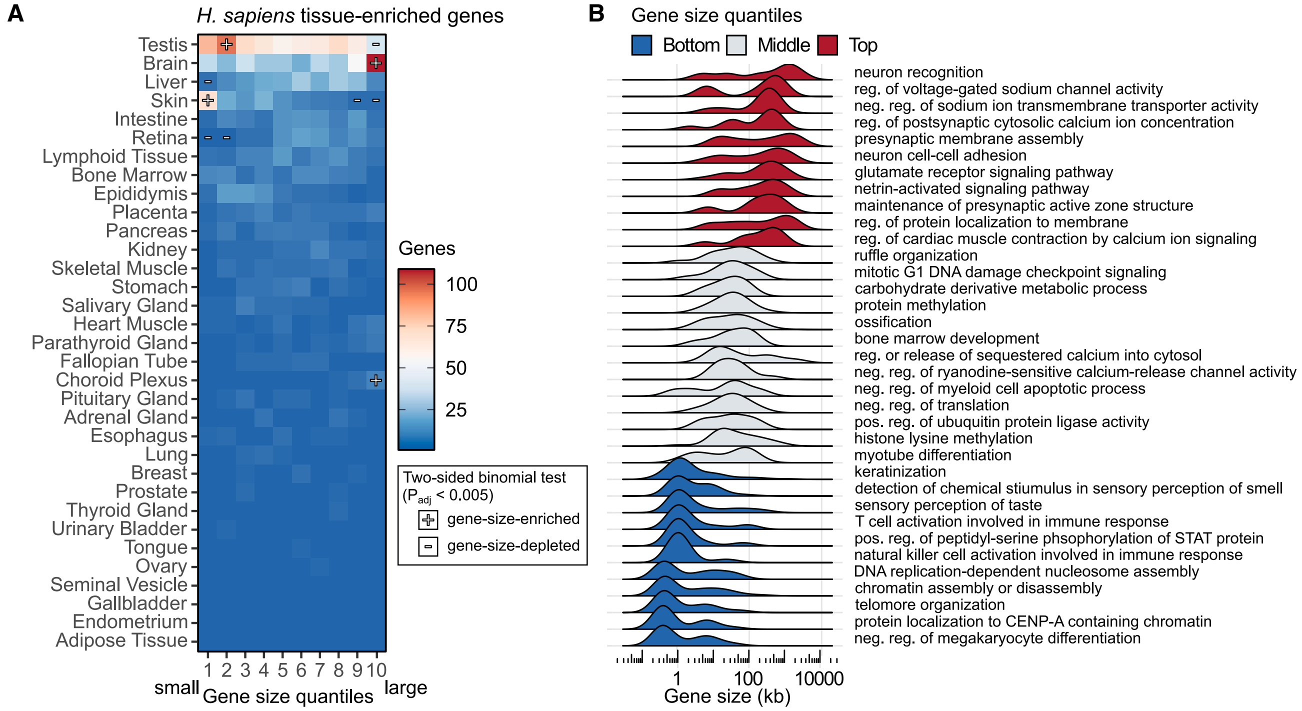

Figure 2. Brain tissue and neural functions are enriched for large genes

(A) Heatmap of Human Protein Atlas tissue-enriched genes binned by gene size quantiles (10 bins) (Data S3A). Heat colors show the number of genes in each bin. Tissues are ordered by the total number of enriched genes across all gene sizes in each tissue. Enrichment (+) or depletion (−) of genes of particular gene sizes per tissue was determined by a two-sided binomial test with Benjamini-Hochberg adjusted p values < 0.005 (Data S3B).

(B) Stacked density plots (joy plots) showing human GO biological terms filtered to display terms with the lowest deviation from median gene sizes (top 10, middle 10, bottom 10) (Data S3E). See also Figure S2.

macroevolutionary scale. This also indicates that the largest genes in one species are among the largest in distantly related species but can vary in absolute size by orders of magnitude.

We sought additional evidence of the relationship between gene and CDS size (and hence protein size) by comparing one-to-one, high-confidence orthologs (obtained from Ensembl[24]; Data S2) in *Homo sapiens* and the invertebrate nematode, *Caenorhabditis elegans*, which have some of the best-characterized animal genomes. Humans shared a common ancestor with nematodes likely more than 550 million years ago,[25] and since then our haploid genome has expanded to more than 3 billion base pairs, roughly 30 times the size of the *C. elegans* haploid genome at around 100 million base pairs.[26] We found that while the CDS size of each orthologous gene is nearly invariant between species, the largest human genes can be more than 100 times the size of their orthologs in *C. elegans* (Figure 1D). We also found that within *H. sapiens* and *C. elegans* genomes, CDS size is strongly correlated with gene size (Figure 1E). When we compared the correlation of gene size in humans with either CDS size or gene size in *C. elegans*, we found these relationships to be similarly strong, suggesting that protein size and gene size are closely related on a macroevolutionary scale (Figure 1F). These results are consistent with the known conservation of orthologous protein sizes among diverse eukaryotes,[2] while highlighting significant differences in absolute gene size that may underlie important aspects of gene function and expression.

Specific neuronal functions enriched for large genes

One unusual feature of nervous tissue is the high number of genes with tissue-specific expression.[27] Previous studies observed that many of the largest genes are enriched for expression in the brain.[7,12,13,15,28–30] Using tissue-enriched genes provided by the Human Protein Atlas[27] (HPA; at least 4-fold higher mRNA level in a particular tissue compared with any other tissue), we quantified the number of tissue-enriched genes across gene sizes (Data S3A and S3B) and found more brain-enriched genes in the top 10% largest genes than in any other size range (Figure 2A; two-sided binomial test, Benjamini-Hochberg adjusted p value = 9.06e−22). We also found that of the 109 human genes that are large and brain-enriched by these criteria, 43 are synaptic or contribute to synaptic function by gene ontology (GO; Figure S2A; Data S3C). This contrasts with the high number of small genes enriched for expression in the testis and skin (Figure 2A).

Although previous studies have noted the enrichment of large gene expression within neurons specifically, it was unclear whether other neural cell types might express large genes. We analyzed gene expression data of single cells provided by the HPA and found widespread expression of large and synaptic genes in several glial cell types, albeit at lower levels than in neurons (Figure S2B; Data S3D). This is intriguing and consistent with observations of functional synapses in glia.[31]

The largest genes are known to be enriched for GO terms associated with synaptic function.[13] We examined gene size distributions for GO terms associated with individual functions, which provided a striking picture in which some functions were associated with a majority of genes in a specific size class (Figure 2B; Data S3E). In particular, many GO terms composed mainly of large genes are involved in neuronal function (e.g., neuron recognition, presynaptic membrane assembly, neuron cell-cell adhesion, etc.) (Figure 2B). These results suggest that there are classes of genes whose functions may benefit from: (1) small, condensed gene sizes, such as highly expressed genes,[32–34] and genes involved in rapid stress response[35]; or (2) expanded gene sizes, such as neuronal genes with numerous isoforms. There may also be a third class of genes (3) that do not benefit from either small or expanded gene sizes, or whose gene sizes are determined by currently unknown forces.

Most large neuronal genes are ancient

Previous studies found that older genes on average are larger, experience stronger purifying selection, and evolve more slowly than younger genes.[36–38] However, these aggregate measures obscure certain features, such as the fact that many short, ancient genes are evolving under strong purifying selection (e.g., histone genes[39]). We therefore sought a more detailed analysis on genes of specific ages and sizes, including the large neural and synaptic genes.

Our analysis in Figure 1 focused on genes with orthologs across diverse eukaryotes and thus was necessarily limited to ancient conserved genes. To address whether most large genes are ancient, we used estimates of gene age based on the phylogenetic distribution of orthologs as described by Tong et al.[40] (Data S4A). We found that most of the larger protein-coding genes are indeed ancient, with the top 10% largest human genes averaging an inferred age (primarily based on molecular clock estimates, STAR Methods) of over 900 million years old, whereas the top 10% shortest genes have an average inferred age of 320 million years old (Figures 3A and 3B; Data S4B). When we specifically analyzed the age and size of all human synaptic genes (1,612 genes in total; Data S4C and S4D), we found that many synaptic genes are large and old (Figure 3C). As an alternative to primarily molecular clock estimates of gene age, we also quantified orthogroups identifiable in different clades (Data S4E and S4F). Starting from a list of large, neural-enriched (brain or retina) genes from the HPA, we found that 81 out of 105 orthogroups (which includes one-to-one, one-to-many, and many-to-many orthologs; STAR Methods) were conserved between humans and invertebrates (Bilateria). More than half (57) of the 105 orthogroups could be identified prior to the lineage divergence separating humans and the sponge *Amphimedon queenslandica* (the recently proposed Myriazoa[41])—which lack obvious neurons and nervous tissue[42]—and a third (35) were identified prior to the divergence of humans and the closest non-animal outgroup, the choanoflagellate *Salpingoeca rosetta* (Choanozoa). We also quantified orthogroups containing tissue-enriched genes of all gene sizes (Figure 3D), and for neural genes we found enrichment for the top 10% largest genes present in Bilateria or older (two-sided binomial test, Benjamini-Hochberg adjusted p value = 6.2e−7), Myriazoa or older (p = 8.81e−7), and in Choanozoa or older (p = 6.1e−8). Of these non-animal orthogroups, we identified 24 as containing synaptic genes by GO terms, which map to 101 human genes. This complements previous studies that identified orthologs of specific synaptic gene families outside of organisms with nervous systems, such as synaptosomal-associated proteins (SNAPs) in sponges,[43] choanoflagellates,[44] and plants.[45] Whether sponges lost neurons and nervous systems or whether ctenophores, as the sister group to all animals,[41,46,47] independently evolved them,[48] there is accumulating evidence that many genes with synaptic expression in extant metazoans originated prior to the evolution of nervous systems.[49] Our observations complement these findings by highlighting that many of these genes share a feature of large gene sizes and raises the possibility that parallel changes in gene size may have played a role in their joint evolution.

Large, ancient genes are sequence constrained yet have gained the most isoforms

A valuable metric for measuring sequence constraint has been the dN/dS ratio (the ratio of non-synonymous to synonymous substitutions; see Jeffares et al. for a review[50]). To estimate the degree of sequence constraint of large, ancient neural and synaptic genes, we next compared the dN/dS ratios of Ensembl one-to-one orthologs between mouse (*M. musculus*) and human (Data S5A and S5B). Although genes with the lowest dN/dS ratios (dN/dS quantile 1/10; median dN/dS ratio = 0.01) were distributed relatively evenly among genes of different sizes, we found that the next group of genes with low dN/dS ratios were enriched for the top 10% largest genes (dN/dS quantile 2/10, median dN/dS ratio = 0.03, p = 1.91e−9; dN/dS quantile 3/10, median dN/dS ratio = 0.05, p = 3.97e−3; two-sided binomial test, Benjamini-Hochberg adjusted p values). By contrast, genes with the highest dN/dS ratios were depleted for the top 10% largest genes (dN/dS quantile 9/10, median dN/dS ratio = 0.26, p = 6.21e−7; dN/dS quantile 10/10, median dN/dS ratio = 0.41, p = 8.53e−9), enriched for the smallest genes (dN/dS quantile 10/10, median dN/dS ratio = 0.41, p = 1.86e−9) (Figures 4A and S3A), and typically younger (Figure S3C). This complements the recent observation of lower dN/dS ratios in genes with larger RNA transcripts (mature RNA transcript length excluding introns).[51] (Because neural genes of all sizes appear to have lower dN/dS ratios in this human-mouse comparison, the enrichment of low dN/dS ratios for larger genes in this case seems likely attributable to the number of neural genes with large gene sizes [Figures S3D–S3F] and not necessarily a relationship between gene size and dN/dS ratios.)

We observed that animals with expanded genomes have ancient, highly conserved genes that are acquiring new isoforms, mainly in larger genes (Figure 4). Isoform numbers were obtained by quantifying annotated peptides for each gene and were averaged for each orthogroup, and we compared between all orthologs as well as focusing only on orthologs of human synaptic genes. When we compared the set of large, ancient genes among animals, we found that while orthologs of these genes are typically among the largest in each genome, they have become absolutely larger and more complex both in vertebrates and (independently) in the case of octopus (Figures 1B, 5C, and S4). This indicates that despite showing signs of strong purifying selection, these large, ancient genes are acquiring many new sequences that may undergo positive selection and drive gene evolution in parallel.

Please cite this article in press as: McCoy and Fire, Parallel gene size and isoform expansion of ancient neuronal genes, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.02.021

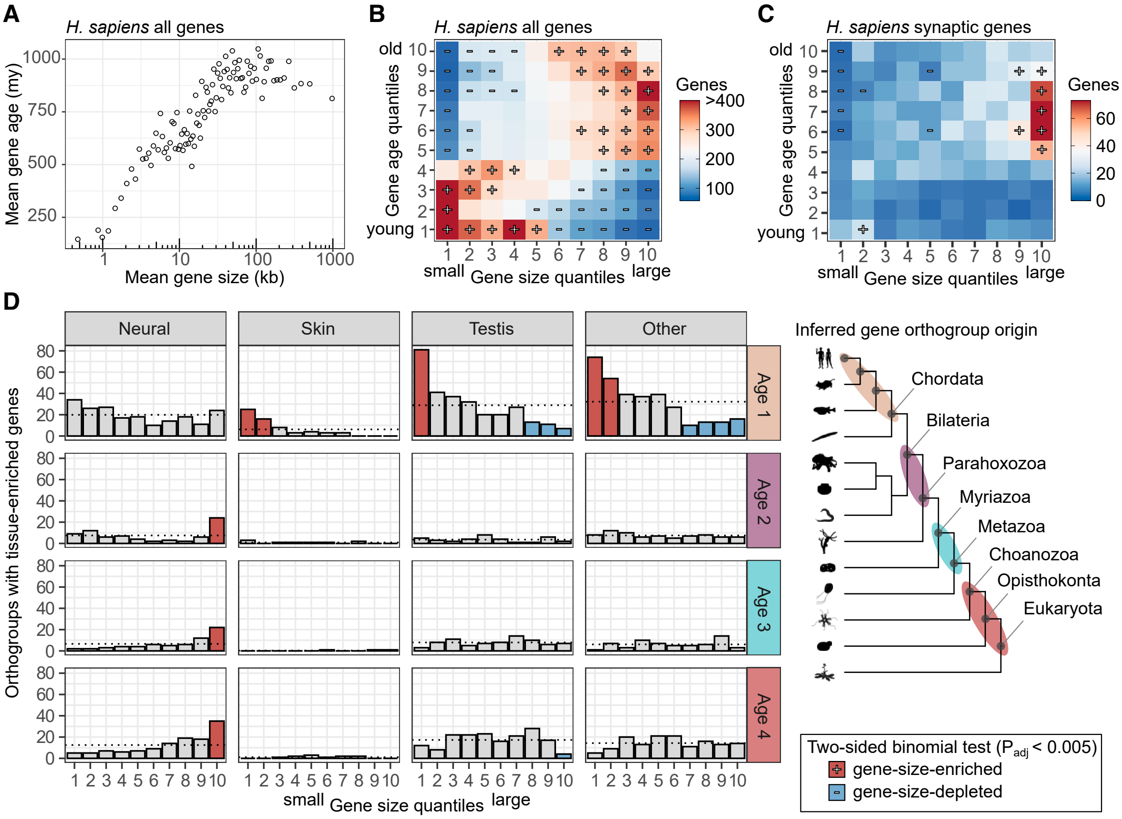

Figure 3. Most large genes are ancient, while most young genes are small

Quantification of gene size versus gene age of human genes.

(A) Scatterplot of mean gene size (kb, kilobases; values derived from Ensembl) versus mean gene age (my, million years; values from GenOrigin[40]) of genes binned by size (100 bins) in *H. sapiens*.

(B) Heatmap of all gene sizes versus gene age in *H. sapiens* (Data S4A and S4B). Genes were binned independently into 10 age bins and 10 gene size bins. Heat colors show the number of genes within each tile and are capped at 400 genes.

(C) Heatmap of synaptic gene size versus gene age in *H. sapiens* (Data S4C and S4D). All genes were binned independently into 10 age bins and 10 gene size bins, then filtered for GO gene terms containing "synaptic" or "synapse." Heat colors show the number of synaptic genes within each tile.

(D) Bar graph quantifying the number of orthogroups containing orthologs of human tissue enriched from the Human Protein Atlas (Data S4E and S4F). All orthogroups from Figure 1A were categorized as age 1 (present within Chordata), age 2 (Bilateria or Parahoxozoa), age 3 (Myriazoa or Metazoa), or age 4 (Choanozoa or older). Mean human gene sizes of all orthologs were estimated per orthogroup, binned into 10 quantiles, then filtered to only contain orthogroups with orthologs of tissue-enriched genes in humans. Enrichment (red) or depletion (blue) of observations per tissue, bin and clade were determined by two-sided binomial test with Benjamini-Hochberg adjusted p values < 0.005. Dotted line shows the mean number of orthogroups per tissue and age. For "Other," tissue-enriched genes from all other tissues (see Figure 2A) were combined.

DISCUSSION

Determinants of optimal gene size

By comparing the genomes and transcriptomes of diverse eukaryotes, we have outlined the contribution of gene size variation to the parallel evolution of large neuronal genes. We propose the adaptive value in gene size expansion does not come from net gains directly but rather from adding sites capable of sustaining beneficial mutations. Any change to individual gene sizes might disrupt coexpression dynamics. However, if these are balanced by net changes in coexpressed gene sizes, coexpression might be maintained while simultaneously generating raw material for selection to act on. This could effectively add new sites capable of sustaining beneficial mutations and potentiate gene architecture complexity in expanded genes. As the largest genes will have the largest absolute expansion of sequence space, these genes have the most potential to gain novel functions and expression patterns.

Gene size and expression timing

Gene size directly affects expression timing and thus may contribute to the precise coordination of gene expression required by many biological processes. The effect of gene size on expression timing was first appreciated in the long, late operons of lambda phage.[52] When the size and abundance of introns in eukaryotic genes were discovered, these were likewise anticipated to have substantial effects on gene expression timing. This idea was articulated in the intron delay hypothesis, which postulates that intron size contributes to a time delay and aids the orchestration of gene expression patterns.[53] Several studies have since provided evidence that intron and gene size play a role in embryonic development by affecting

Please cite this article in press as: McCoy and Fire, Parallel gene size and isoform expansion of ancient neuronal genes, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.02.021

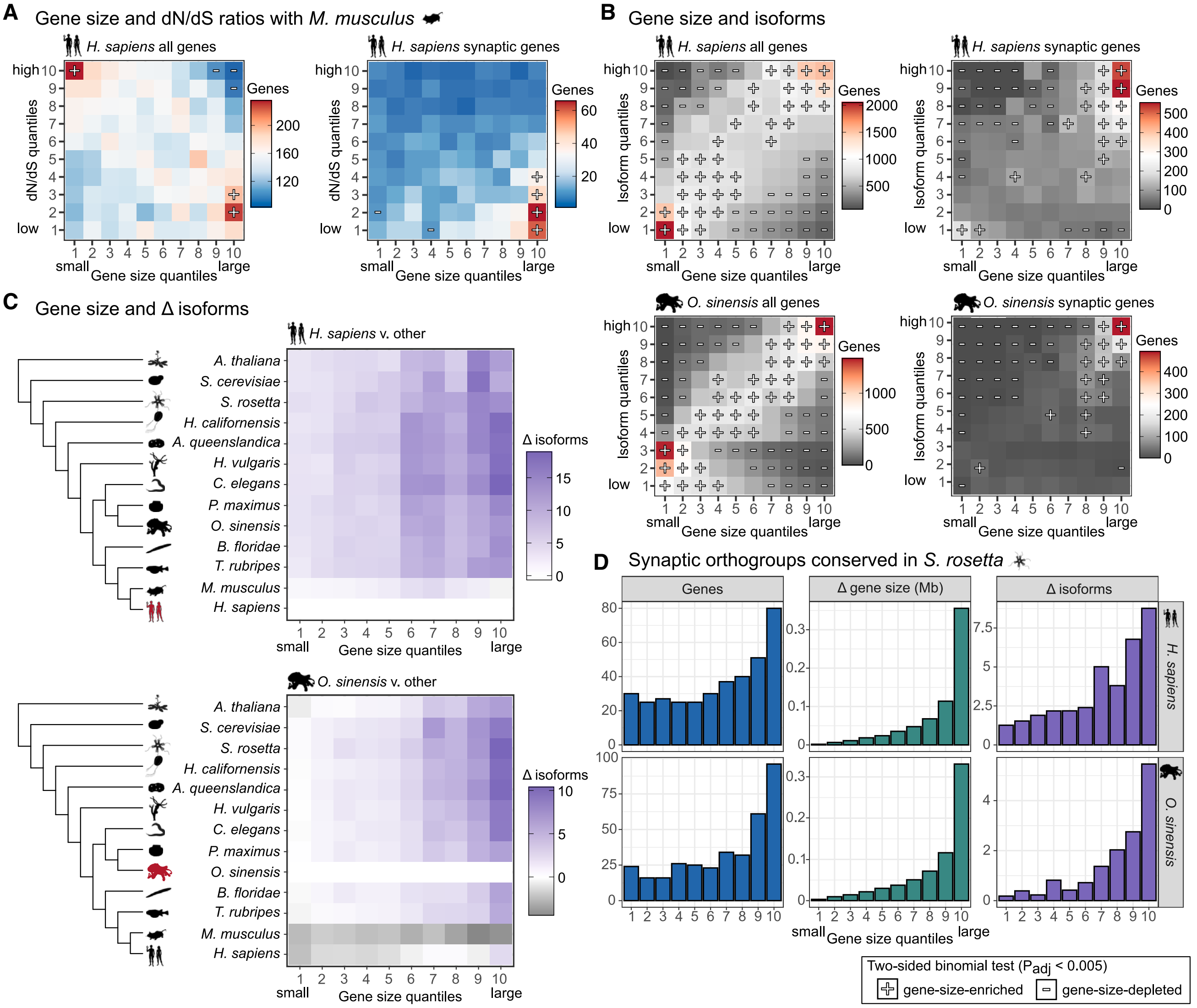

Figure 4. Large ancient genes are sequence constrained yet have gained the most isoforms

(A) Heatmap showing number of Ensembl one-to-one orthologs between *M. musculus* and *H. sapiens* binned independently by human gene size (10 bins) and dN/dS ratios (10 bins) for all one-to-one orthologs (left) or only synaptic genes (right) (Data S5A and S5B).

(B) Heatmaps showing number of genes binned independently by gene sizes (10 bins) and isoform number (10 bins) for *H. sapiens* (top) or *O. sinensis* (bottom), for all genes (left) or only synaptic genes (right) (Data S5C and S5D).

(C) Heatmaps showing the change in isoforms (Δ isoforms) between *H. sapiens* versus all other species for human genes binned by gene size (10 bins) (top), or for *O. sinensis* versus all others species for octopus genes binned by gene size (10 bins) (bottom).

(D) Quantification of the number of synaptic orthogroups (left), change in gene size (middle; Mb, megabases), and change in isoforms (right) between *H. sapiens* and *S. rosetta* (top) or *O. sinensis* and *S. rosetta* (bottom) (Data S5E and S5F). Enrichment (+) or depletion (−) of observations per gene size bin was determined by a two-sided binomial test with Benjamini-Hochberg adjusted p values < 0.005. See also Figure S3.

transcriptional kinetics (see Swinburne and Silver[54] for a review). Additionally, highly expressed genes[32–34] and genes involved in rapid stress response[35] tend to have shorter introns, suggesting that selection for efficiency acts to reduce the time and energy costs of transcription.

Assuming estimated transcription rates of eukaryotes of 1–4 kb per minute, the 2.3 Mb human gene *CNTNAP2* would require upward of 10 h to generate a transcript.[7,55–57] This is dramatically longer, for instance, than the typical intronless histone gene, which, at ∼500 bp, would theoretically take less than a minute to transcribe. Genes encoding subunits of the same protein complex tend to have similar gene sizes, which has been hypothesized to prevent dosage imbalance from uncoordinated gene expression that can be toxic to cells.[58] Additionally, outside of homeostasis, any dynamically expressed set of genes (e.g., synaptic genes upregulated in response to neural activity) could potentially benefit from factors affecting their coordination. It is therefore possible that many biological processes involve genes with similar sizes and that gene sizes may be evolving in part from selective pressure for expression timing.

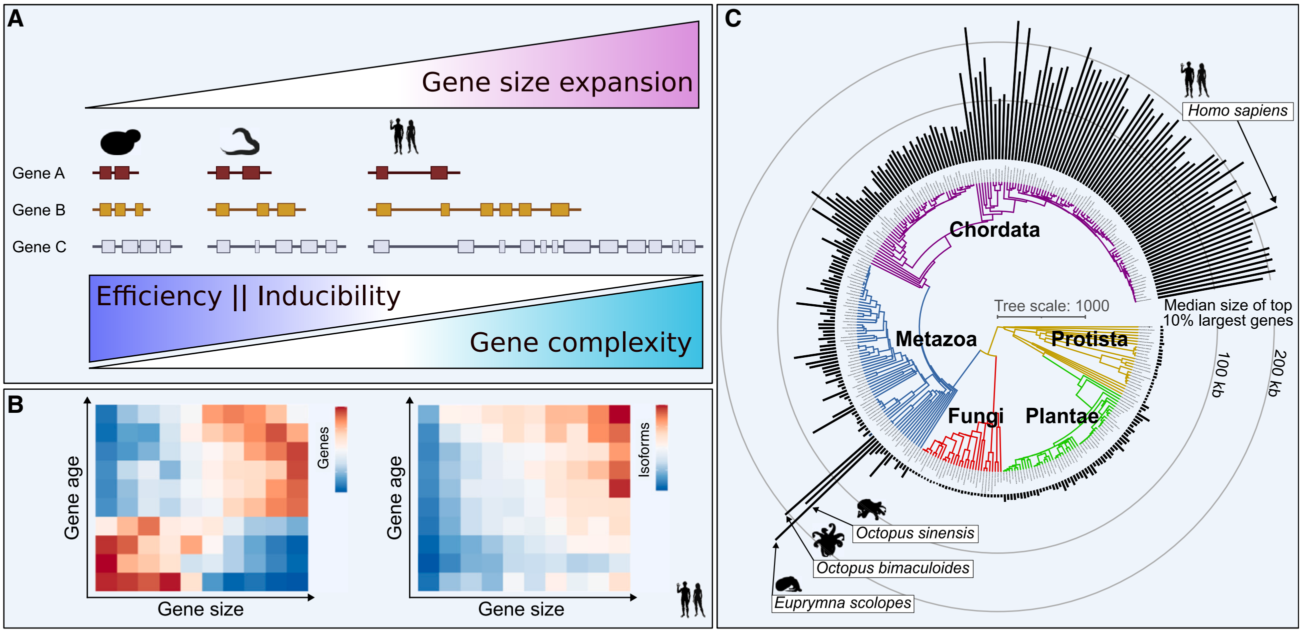

Figure 5. Model of gene size variation

(A) As genomes expand or contract, so does the intronic content of genes and hence gene sizes of eukaryotes. Larger genes are able to become more complex, but at the cost of inducibility and efficiency of expression.

(B) Genes become large by being very old and become complex by being large.

(C) Species-specific differences in gene size variation may contribute to important differences in the potential for complex genes and phenotypes. See also Figure S4 and Data S6.

Large genes in other tissues

There are a handful of very large genes expressed outside the brain (Data S3A). These include skin-expression-enriched genes encoding enzymes involved in melanin biosynthesis (i.e., *TYR*, *DCT*). Several skeletal-muscle-expression-enriched genes are part of a family of giant sarcomeric structural/signaling proteins, including *OBSCN*, *NEB*, and *TTN*. There are around 39 large genes with expression enrichment in the testis, though it is unclear whether these have testis-specific functions or whether their high expression is part of widespread "transcriptional scanning" that appears to occur specifically in the testis.[59] Interestingly, some large genes not in the dedicated neural categories also have evidence of expression within the brain, such as the heart-expression-enriched genes *CORIN* (a serine-type endopeptidase involved in atrial natriuretic peptide [*NPPA*] and brain natriuretic peptide [*NPPB*] processing), *MTUS2* (a microtubule-associated protein), *RYR2* (calcium channel), *TNNI3K* (MAPKKK), and *CCDC141* (predicted to be involved in axon guidance and cell adhesion).

Gene size expansion and the addition of adaptive sites

The rate at which a gene under selection accrues beneficial substitutions is thought to be rapid at first, eventually slowing with the depletion of sites capable of sustaining beneficial mutations (often referred to as "adaptive sites").[60,61] Under the "increasing constraint" model,[62] a newly born gene evolves under weak negative or positive selection, and later evolves primarily under strong negative selection. More recent evidence supports a variation of this idea, which is that young genes experience more variable dN/dS values than old genes.[36]

Our study provides evidence that gene size expansion in genes under high constraint (i.e., large and ancient genes under strong negative selection) can facilitate acquisition of sites capable of sustaining beneficial mutations in the form of new exons and regulatory regions. These new DNA sequences are likely under weaker constraint than the original sequences and can thus contribute to evolution. Many new exons arise from within introns and tend to be cassette exons that are rarely incorporated into final transcripts (i.e., they are spliced out).[63,64] Similar to neo-functionalization of duplicated genes, because the original function is maintained by major isoforms, new isoforms are less constrained by negative selection[64] and can thereby contribute to adaptive evolution.[63] Thus, we speculate that gene size expansion may be one mechanism by which genes under high constraint can gain new raw material under weak constraint and contribute to the evolution of molecular diversity.

Previously, it has been argued that weaker constraint is unlikely to have contributed to the evolution of primate nervous systems because their complexity necessitates a greater precision in gene function.[65] Conversely, based on the results of this study, we speculate that this *weaker* constraint (through gene size expansion) may have set the conditions for the evolution of complex nervous systems by providing substrate for adaptive evolution.

Gene size expansion and nervous system evolution

Gene size expansion has been hypothesized to facilitate the evolution of complex nervous systems.[7,14,66] This is in large part because most of the largest animal genes are multi-isoformic, enriched for expression in nervous tissue, and predominantly encode synaptic proteins underlying the precise wiring of the nervous system.[7,12–15] Additionally, large genes have been shown to contribute to the extensive molecular diversity and complexity of vertebrate brains.[14] However, because most studies of complex nervous systems have focused on vertebrates, it remains unclear whether any such relationship arose from historical contingency. Did any invertebrate animals with complex nervous systems independently undergo gene size expansion?

Although many vertebrates have large brains, as well as some of the largest genomes and gene sizes among animals,[3,5,7] there are outlier species among invertebrates, such as the cephalopods. Cephalopods have the largest invertebrate nervous systems and exhibit complex behaviors rivaling many vertebrates.[23] It has been more than 550 million years since cephalopods and vertebrates shared a last common ancestor,[25,67] which likely had a compact genome and gene sizes as well as a simple nervous system.[68] Several chromosome-level genome assemblies for cephalopods have recently been published,[69–71] and in our analyses we found a striking expansion of gene sizes similar to that seen in the vertebrate lineage (Figures 1B, 5C, and S4). The fact that many large, complex genes are enriched for neuronal expression and function across diverse animals is consistent with the hypothesis that gene size expansion contributed to the tremendous molecular diversity and complexity observed within nervous systems.

Several studies have traced the evolutionary history of individual families of neuronal genes in different animal lineages. For example, divergent lineage-specific events have been characterized for the metabotropic glutamate receptors (mGluRs),[72] some of which we note are among the largest genes in the human genome (*GRM1*: 410 kb; *GRM3*: 221 kb; *GRM4*: 137 kb; *GRM5*: 561 kb; *GRM7*: 971 kb; *GRM8*: 815 kb). Sponges, while lacking nervous systems, have been noted to contain a diversity of mGluRs, as well as metabotropic gamma-aminobutyric acid (GABA) receptors and SNAPs that have been hypothesized to allow for sensation of the environment in the absence of rapid, synaptic-type electro-chemical signaling. Our findings complement the recent phylogenetic studies of the ancient origins of these specific genes, highlighting the potential role of parallel gene size and isoform expansion in the lineage-specific evolution of neuronal genes.

Of considerable interest in the context of models in which gene size expansion accompanies nervous system diversification are a number of counterexamples. For example, there are some animal genomes that underwent significant expansion (salamanders,[73] whale sharks,[74] lungfish,[75] grasshoppers,[76] etc.) without obvious increases in the complexity of their nervous systems relative to other animals. We speculate that gene size expansion is insufficient for gene architectural complexification, but may only set the conditions for further evolution by selection. It is also possible that the mechanisms by which genes and genomes expand impacts the mechanisms that generate novel regulatory elements and exons. For example, the diversity and composition of transposable element pools[77] differs in a species-specific manner; more diverse transposable element pools may limit the acquisition of additional sequences by recombination, while some populations of transposable elements may be more or less likely to introduce regulatory modulation when inserted.

Gene and genome size contraction

The focus of this study was on gene and genome size expansion, but there are numerous examples of gene and genome size contraction as well. One example is the tomato russet mite, *Aculops lycopersici*, one of the smallest animals with the smallest-known arthropod genome at 32.5 Mb.[78] There are few transposable elements (<2% of the genome), small intergenic regions, and more than 80% of coding genes are intronless. Interestingly, 3′ introns were predominantly lost, which complements findings from other studies that 5′ introns are enriched for regulatory elements.[79,80] There are also cases of genome reduction among vertebrates, for example, within the teleost fish, *Takifugu* (*T. rubripes*; 300 Mb).[81]

If gene size expansion sets the conditions for added complexity, does that mean gene size contraction reduces the potential for complexity and adaptation? Future studies are needed to investigate these questions—in particular, whether small genomes are evolutionary dead ends—which have implications for our understanding of how complex systems are generated or degenerated.

Ancient events enabling recent adaptation

The genome design model[82] posits that tissue-specific proteins have more complex architectures that explain the increase in their size. Extending this model, it has been argued that the complexity of large genes was already present at the base of the metazoan common ancestor.[83] Conversely, our results suggest that increases in the size of genes encoding tissue-specific proteins precede and potentiate the evolution of their more complex architecture. Rather than looking for the origins of gene architectural and regulatory complexity in the recent evolutionary history of any one species, our analysis suggests that ancient events established the necessary underlying conditions. The initial size of these genes may predispose them, over time, to becoming extremely large and accumulating sequences that selection can act on to generate complexity.

In conclusion, in this study we found that relative gene size is being maintained for most genes in each genome, despite sometimes orders-of-magnitude changes in absolute gene sizes in orthologs among species. We found that most young genes are small, while virtually all larger genes are ancient. This includes the set of large genes with neuronal expression in extant metazoans, whose origins appear to predate the diversification of animals and, in many cases, the emergence of neurons and nervous systems. An intriguing possibility is that maintaining relative gene size during evolution may facilitate the coordination of gene expression, while increases in absolute gene size may contribute to the evolution of novel gene structures and regulatory elements.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE

- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ Gene and protein sizes
  - ○ Identification of orthologs
  - ○ Gene ontology
  - ○ Species phylogeny
  - ○ Gene ages
  - ○ Species selection
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Ortholog feature comparisons
  - ○ Enrichment of gene size within gene features and gene sets
  - ○ Single-cell gene expression analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2024.02.021.

ACKNOWLEDGMENTS

Helpful input was provided by A. Herbert, D. Schultz, T. Rogers, O. Simakov, D. Lipman, J. Rosenthal, K. Artiles, I. Zheludev, J. Chen, D. Galls, N. Hall, N. Jain, L. Wahba, M. Shoura, D. Jeong, U. Enam, E. Greenwald, L. Del Rosario, and O. Ilbay. This work was funded by Stanford Wu Tsai Neurosciences Institute Interdisciplinary Scholar Fellowship (M.J.M.), Stanford Genomics Training Program (5T32HG000044-22; M.J.M.), Whitman Early Career Award (M.J.M.), and NIH grant R35GM130366 (A.Z.F.).

AUTHOR CONTRIBUTIONS

M.J.M. and A.Z.F. conceived of the study. M.J.M. performed all investigations and analyses with assistance from A.Z.F. The authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 22, 2023
Revised: December 16, 2023
Accepted: February 11, 2024
Published: March 8, 2024

REFERENCES

1. Koonin, E.V., Aravind, L., and Kondrashov, A.S. (2000). The Impact of Comparative Genomics on Our Understanding of Evolution. Cell *101*, 573–576.
2. Wang, D. (2005). A General Tendency for Conservation of Protein Length Across Eukaryotic Kingdoms. Mol. Biol. Evol. *22*, 142–147.
3. Francis, W.R., and Wörheide, G. (2017). Similar Ratios of Introns to Intergenic Sequence across Animal Genomes. Genome Biol. Evol. *9*, 1582–1598.
4. Lynch, M. (2007). The Origins of Genome Architecture (Sinauer Associates).
5. Lynch, M., Bobay, L.M., Catania, F., Gout, J.F., and Rho, M. (2011). The Repatterning of Eukaryotic Genomes by Random Genetic Drift. Annu. Rev. Genomics Hum. Genet. *12*, 347–366.
6. Grishkevich, V., and Yanai, I. (2014). Gene length and expression level shape genomic novelties. Genome Res. *24*, 1497–1503.
7. McCoy, M.J., and Fire, A.Z. (2020). Intron and gene size expansion during nervous system evolution. BMC Genomics *21*, 360.
8. Moriyama, E.N., Petrov, D.A., and Hartl, D.L. (1998). Genome size and intron size in Drosophila. Mol. Biol. Evol. *15*, 770–773.
9. Vinogradov, A.E. (1999). Intron–Genome Size Relationship on a Large Evolutionary Scale. J. Mol. Evol. *49*, 376–384.
10. Gregory, T.R., and Hebert, P.D.N. (1999). The Modulation of DNA Content: Proximate Causes and Ultimate Consequences. Genome Res. *9*, 317–324.
11. Knight, C.A., and Ackerly, D.D. (2002). Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. Ecol. Lett. *5*, 66–76.
12. King, I.F., Yandava, C.N., Mabb, A.M., Hsiao, J.S., Huang, H.S., Pearson, B.L., Calabrese, J.M., Starmer, J., Parker, J.S., Magnuson, T., et al. (2013). Topoisomerases facilitate transcription of long genes linked to autism. Nature *501*, 58–62.
13. Gabel, H.W., Kinde, B., Stroud, H., Gilbert, C.S., Harmin, D.A., Kastan, N.R., Hemberg, M., Ebert, D.H., and Greenberg, M.E. (2015). Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature *522*, 89–93.
14. Sugino, K., Clark, E., Schulmann, A., Shima, Y., Wang, L., Hunt, D.L., Hooks, B.M., Tränkner, D., Chandrashekar, J., Picard, S., et al. (2019). Mapping the transcriptional diversity of genetically and anatomically defined cell populations in the mouse brain. eLife *8*, e38619.
15. McCoy, M.J., Paul, A.J., Victor, M.B., Richner, M., Gabel, H.W., Gong, H., Yoo, A.S., and Ahn, T.H. (2018). LONGO: an R package for interactive gene length dependent analysis for neuronal identity. Bioinformatics *34*, i422–i428.
16. Ryan, T.J., and Grant, S.G.N. (2009). The origin and evolution of synapses. Nat. Rev. Neurosci. *10*, 701–712.
17. Tennyson, C.N., Klamut, H.J., and Worton, R.G. (1995). The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. Nat. Genet. *9*, 184–190.
18. Kopelman, N.M., Lancet, D., and Yanai, I. (2005). Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nat. Genet. *37*, 588–589.
19. Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. *16*, 157.
20. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238.
21. Elliott, T.A., and Gregory, T.R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140331.
22. Lynch, M. (2006). The Origins of Eukaryotic Gene Structure. Mol. Biol. Evol. *23*, 450–468.
23. Young, J.Z. (1971). The Anatomy of the Nervous System of Octopus vulgaris (Clarendon Press).
24. Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M.J., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics resources. Database (Oxford) *2016*, baw053.
25. Budd, G.E., and Mann, R.P. (2020). Survival and selection biases in early animal evolution and a source of systematic overestimation in molecular clocks. Interface Focus *10*, 20190110.
26. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.
27. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419.

Please cite this article in press as: McCoy and Fire, Parallel gene size and isoform expansion of ancient neuronal genes, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.02.021

28. Mabb, A.M., Kullmann, P.H.M., Twomey, M.A., Miriyala, J., Philpot, B.D., and Zylka, M.J. (2014). Topoisomerase 1 inhibition reversibly impairs synaptic function. Proc. Natl. Acad. Sci. USA *111*, 17290–17295.

29. Cates, K., McCoy, M.J., Kwon, J.S., Liu, Y., Abernathy, D.G., Zhang, B., Liu, S., Gontarz, P., Kim, W.K., Chen, S., et al. (2021). Deconstructing Stepwise Fate Conversion of Human Fibroblasts to Neurons by MicroRNAs. Cell Stem Cell *28*, 127–140.e9.

30. Lu, Y.L., Liu, Y., McCoy, M.J., and Yoo, A.S. (2021). MiR-124 synergism with ELAVL3 enhances target gene expression to promote neuronal maturity. Proc. Natl. Acad. Sci. USA *118*, e2015454118.

31. Bergles, D.E., Roberts, J.D.B., Somogyi, P., and Jahr, C.E. (2000). Glutamatergic synapses on oligodendrocyte precursor cells in the hippocampus. Nature *405*, 187–191.

32. Seoighe, C., Gehring, C., and Hurst, L.D. (2005). Gametophytic Selection in Arabidopsis thaliana Supports the Selective Model of Intron Length Reduction. PLoS Genet. *1*, e13.

33. Urrutia, A.O., and Hurst, L.D. (2003). The Signature of Selection Mediated by Expression on Human Genes. Genome Res. *13*, 2260–2264.

34. Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V., and Kondrashov, F.A. (2002). Selection for short introns in highly expressed genes. Nat. Genet. *31*, 415–418.

35. Jeffares, D.C., Penkett, C.J., and Bähler, J. (2008). Rapidly regulated genes are intron poor. Trends Genet. *24*, 375–378.

36. Vishnoi, A., Kryazhimskiy, S., Bazykin, G.A., Hannenhalli, S., and Plotkin, J.B. (2010). Young proteins experience more variable selection pressures than old proteins. Genome Res. *20*, 1574–1581.

37. Cai, J.J., and Petrov, D.A. (2010). Relaxed Purifying Selection and Possibly High Rate of Adaptation in Primate Lineage-Specific Genes. Genome Biol. Evol. *2*, 393–409.

38. Wolf, Y.I., Novichkov, P.S., Karev, G.P., Koonin, E.V., and Lipman, D.J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc. Natl. Acad. Sci. USA *106*, 7273–7280.

39. Piontkivska, H., Rooney, A.P., and Nei, M. (2002). Purifying Selection and Birth-and-death Evolution in the Histone H4 Gene Family. Mol. Biol. Evol. *19*, 689–697.

40. Tong, Y.B., Shi, M.W., Qian, S.H., Chen, Y.J., Luo, Z.H., Tu, Y.X., Xiong, Y.L., Geng, Y.J., Chen, C., and Chen, Z.X. (2021). GenOrigin: A comprehensive protein-coding gene origination database on the evolutionary timescale of life. J. Genet. Genomics *48*, 1122–1129.

41. Schultz, D.T., Haddock, S.H.D., Bredeson, J.V., Green, R.E., Simakov, O., and Rokhsar, D.S. (2023). Ancient gene linkages support ctenophores as sister to other animals. Nature *618*, 110–117.

42. Sakarya, O., Armstrong, K.A., Adamska, M., Adamski, M., Wang, I.F., Tidor, B., Degnan, B.M., Oakley, T.H., and Kosik, K.S. (2007). A Post-Synaptic Scaffold at the Origin of the Animal Kingdom. PLoS One *2*, e506.

43. Kenny, N.J., Francis, W.R., Rivera-Vicéns, R.E., Juravel, K., De Mendoza, A., Díez-Vives, C., Lister, R., Bezares-Calderón, L.A., Grombacher, L., Roller, M., et al. (2020). Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge Ephydatia muelleri. Nat. Commun. *11*, 3676.

44. Kloepper, T.H., Kienle, C.N., and Fasshauer, D. (2008). SNAREing the Basis of Multicellularity: Consequences of Protein Family Expansion during Evolution. Mol. Biol. Evol. *25*, 2055–2068.

45. Sanderfoot, A. (2007). Increases in the Number of SNARE Genes Parallels the Rise of Multicellularity among the Green Plants. Plant Physiol. *144*, 6–17.

46. Moroz, L.L., Kocot, K.M., Citarella, M.R., Dosung, S., Norekian, T.P., Povolotskaya, I.S., Grigorenko, A.P., Dailey, C., Berezikov, E., Buckley, K.M., et al. (2014). The ctenophore genome and the evolutionary origins of neural systems. Nature *510*, 109–114.

47. Moroz, L.L., and Kohn, A.B. (2016). Independent origins of neurons and synapses: insights from ctenophores. Philos. Trans. R. Soc. Lond. B Biol. Sci. *371*, 20150041.

48. Burkhardt, P. (2022). Ctenophores and the evolutionary origin(s) of neurons. Trends Neurosci. *45*, 878–880.

49. Burkhardt, P. (2015). The origin and evolution of synaptic proteins – choanoflagellates lead the way. J. Exp. Biol. *218*, 506–514.

50. Jeffares, D.C., Tomiczek, B., Sojo, V., and Dos Reis, M. (2015). A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome. In Parasite Genomics Protocols Methods in Molecular Biology, C. Peacock, ed. (Springer), pp. 65–90.

51. Lopes, I., Altab, G., Raina, P., and de Magalhães, J.P. (2021). Gene Size Matters: An Analysis of Gene Length in the Human Genome. Front. Genet. *12*, 559998.

52. Watson, J.D. (1970). Molecular Biology of the Gene, Second Edition (Benjamin Cummings).

53. Gubb, D. (1986). Intron-delay and the precision of expression of homoeotic gene products in Drosophila. Dev. Genet. *7*, 119–131.

54. Swinburne, I.A., and Silver, P.A. (2008). Intron delays and transcriptional timing during development. Dev. Cell *14*, 324–330.

55. Singh, J., and Padgett, R.A. (2009). Rates of in situ transcription and splicing in large human genes. Nat. Struct. Mol. Biol. *16*, 1128–1133.

56. Ardehali, M.B., and Lis, J.T. (2009). Tracking rates of transcription and splicing in vivo. Nat. Struct. Mol. Biol. *16*, 1123–1124.

57. Shamir, M., Bar-On, Y., Phillips, R., and Milo, R. (2016). SnapShot: timescales in cell Biology. Cell *164*, 1302–1302.e1.

58. Chen, X., Shi, S., and He, X. (2009). Evidence for Gene Length As a Determinant of Gene Coexpression in Protein Complexes. Genetics *183*, 751–754.

59. Xia, B., Yan, Y., Baron, M., Wagner, F., Barkley, D., Chiodin, M., Kim, S.Y., Keefe, D.L., Alukal, J.P., Boeke, J.D., et al. (2020). Widespread Transcriptional Scanning in the Testis Modulates Gene Evolution Rates. Cell *180*, 248–262.e21.

60. Hartl, D.L., Dykhuizen, D.E., and Dean, A.M. (1985). Limits of adaptation: the evolution of selective neutrality. Genetics *111*, 655–674.

61. Kryazhimskiy, S., Tkačik, G., and Plotkin, J.B. (2009). The dynamics of adaptation on correlated fitness landscapes. Proc. Natl. Acad. Sci. USA *106*, 18638–18643.

62. Albà, M.M., and Castresana, J. (2005). Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes. Mol. Biol. Evol. *22*, 598–606.

63. Zhang, X.H.-F., and Chasin, L.A. (2006). Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. Proc. Natl. Acad. Sci. USA *103*, 13427–13432.

64. Cusack, B.P., and Wolfe, K.H. (2005). Changes in Alternative Splicing of Human and Mouse Genes Are Accompanied by Faster Evolution of Constitutive Exons. Mol. Biol. Evol. *22*, 2198–2208.

65. Dorus, S., Vallender, E.J., Evans, P.D., Anderson, J.R., Gilbert, S.L., Mahowald, M., Wyckoff, G.J., Malcom, C.M., and Lahn, B.T. (2004). Accelerated Evolution of Nervous System Genes in the Origin of Homo sapiens. Cell *119*, 1027–1040.

66. Sahakyan, A.B., and Balasubramanian, S. (2016). Long genes and genes with multiple splice variants are enriched in pathways linked to cancer and other multigenic diseases. BMC Genomics *17*, 225.

67. Kumar, S., Suleski, M., Craig, J.M., Kasprowicz, A.E., Sanderford, M., Li, M., Stecher, G., and Hedges, S.B. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. Mol. Biol. Evol. *39*, msac174.

68. Martín-Durán, J.M., Pang, K., Børve, A., Lê, H.S., Furu, A., Cannon, J.T., Jondelius, U., and Hejnol, A. (2018). Convergent evolution of bilaterian nerve cords. Nature *553*, 45–50.

69. Albertin, C.B., Simakov, O., Mitros, T., Wang, Z.Y., Pungor, J.R., Edsinger-Gonzales, E., Brenner, S., Ragsdale, C.W., and Rokhsar, D.S. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature *524*, 220–224.

Please cite this article in press as: McCoy and Fire, Parallel gene size and isoform expansion of ancient neuronal genes, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.02.021

70. Albertin, C.B., Medina-Ruiz, S., Mitros, T., Schmidbaur, H., Sanchez, G., Wang, Z.Y., Grimwood, J., Rosenthal, J.J.C., Ragsdale, C.W., Simakov, O., et al. (2022). Genome and transcriptome mechanisms driving cephalopod evolution. Nat. Commun. *13*, 2427.

71. Songco-Casey, J.O., Coffing, G.C., Piscopo, D.M., Pungor, J.R., Kern, A.D., Miller, A.C., and Niell, C.M. (2022). Cell types and molecular architecture of the octopus visual system. Curr. Biol. *32*, 5031–5044.e4.

72. Ramos-Vicente, D., Ji, J., Gratacòs-Batlle, E., Gou, G., Reig-Viader, R., Luís, J., Burguera, D., Navas-Perez, E., García-Fernández, J., Fuentes-Prior, P., et al. (2018). Metazoan evolution of glutamate receptors reveals unreported phylogenetic groups and divergent lineage-specific events. eLife *7*, e35774.

73. Nowoshilow, S., Schloissnig, S., Fei, J.F., Dahl, A., Pang, A.W.C., Pippel, M., Winkler, S., Hastie, A.R., Young, G., Roscito, J.G., et al. (2018). The axolotl genome and the evolution of key tissue formation regulators. Nature *554*, 50–55.

74. Weber, J.A., Park, S.G., Luria, V., Jeon, S., Kim, H.M., Jeon, Y., Bhak, Y., Jun, J.H., Kim, S.W., Hong, W.H., et al. (2020). The whale shark genome reveals how genomic and physiological properties scale with body size. Proc. Natl. Acad. Sci. USA *117*, 20662–20671.

75. Wang, K., Wang, J., Zhu, C., Yang, L., Ren, Y., Ruan, J., Fan, G., Hu, J., Xu, W., Bi, X., et al. (2021). African lungfish genome sheds light on the vertebrate water-to-land transition. Cell *184*, 1362–1376.e18.

76. Gosalvez, J., López-Fernandez, C., and Esponda, P. (1980). Variability of the DNA Content in Five Orthopteran Species. Caryologia *33*, 275–281.

77. McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. USA *36*, 344–355.

78. Greenhalgh, R., Dermauw, W., Glas, J.J., Rombauts, S., Wybouw, N., Thomas, J., Alba, J.M., Pritham, E.J., Legarrea, S., Feyereisen, R., et al. (2020). Genome streamlining in a minute herbivore that manipulates its host plant. eLife *9*, e56689.

79. Smith, M.W. (1988). Structure of vertebrate genes: A statistical analysis implicating selection. J. Mol. Evol. *27*, 45–55.

80. Bradnam, K.R., and Korf, I. (2008). Longer First Introns Are a General Property of Eukaryotic Gene Structure. PLoS One *3*, e3093.

81. Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B., and Aparicio, S. (1993). Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. Nature *366*, 265–268.

82. Vinogradov, A.E. (2004). Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet. *20*, 248–253.

83. Vinogradov, A.E., and Anatskaya, O.V. (2021). Growth of Biological Complexity from Prokaryotes to Hominids Reflected in the Human Genome. Int. J. Mol. Sci. *22*, 11640.

84. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. *49*, W293–W296.

85. Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M.H., Baldock, R., Barbiera, G., et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. *43*, W589–W598.

86. Rogers, T.F., Yalçın, G., Briseno, J., Vijayan, N., Nyholm, S.V., and Simakov, O. (2024). Gene modelling and annotation for the Hawaiian bobtail squid, Euprymna scolopes. Sci. Data *11*, 40.

87. Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. *19*, 327–335.

88. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Please cite this article in press as: McCoy and Fire, Parallel gene size and isoform expansion of ancient neuronal genes, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.02.021

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
supplemental information tables and code	Figshare	Figshare: https://doi.org/10.6084/m9.figshare.24003750
Software and algorithms		
R (v4.1.2)	R Core Team 2021[88]	https://www.r-project.org/
RStudio (v2023.06.0)	Rstudio Team	https://www.rstudio.com/
Orthofinder	Emms and Kelly[19,20]	https://github.com/davidemms/OrthoFinder
iTOL (v6)	Letunic and Bork[84]	https://itol.embl.de/
Other		
Gene ontology	Ensembl BioMart[85]	https://www.ensembl.org
Gene age estimates	GenOrigin[40]	http://genorigin.chenzxlab.cn/
Evolutionary timescales	Timetree[67]	https://timetree.org/
Single-cell expression data	Human Protein Atlas[27]	https://www.proteinatlas.org/download/rna_single_cell_type.tsv.zip
Species outlines	Phylopic (v2.0)	https://www.phylopic.org/
dN/dS ratios	Ensembl BioMart[85]	https://www.ensembl.org
Amphimedon queenslandica genome	NCBI	Refseq: GCF_000090795.1
Arabidopsis thaliana genome	NCBI	Refseq: GCF_000001735.4
Branchiostoma floridae genome	NCBI	Refseq: GCF_000003815.2
Caenorhabditis elegans genome	NCBI	Refseq: GCF_000002985.6
Euprymna scolopes genome	NCBI	GenBank: GCA_024364815.1
Euprymna scolopes genome annotations	Rogers et al.[86]	https://github.com/TheaFrances/E.scolopes-V2.2-BRAKER2-gene-annotation
Homo sapiens genome	NCBI	Refseq: GCF_009914755.1
Hormiphora californensis genome	NCBI	GenBank: GCA_020137815.1
Hydra vulgaris genome	NCBI	Refseq: GCF_022113875.1
Mus musculus genome	NCBI	Refseq: GCF_000001635.27
Octopus bimaculoides genome	NCBI	Refseq: GCF_001194135.2
Octopus sinensis genome	NCBI	Refseq: GCF_006345805.1
Pecten maximus genome	NCBI	Refseq: GCF_902652985.1
Saccharomyces cerevisiae genome	NCBI	Refseq: GCF_000146045.2
Salpingoeca rosetta genome	NCBI	Refseq: GCF_000188695.1
Takifugu rubripes genome	NCBI	Refseq: GCF_901000725.2

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Matthew J. McCoy (mjmccoy@stanford.edu).

Materials availability

No new unique reagents were generated in this study.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. The data are publicly available on figshare (Figshare: https://doi.org/10.6084/m9.figshare.24003750).

Please cite this article in press as: McCoy and Fire, Parallel gene size and isoform expansion of ancient neuronal genes, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.02.021

- All code for data analysis and production of figures is publicly available on figshare (Figshare: https://doi.org/10.6084/m9.figshare.24003750).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Gene and protein sizes

In this work we use several terms to describe aspects of size associated with gene expression patterns and function. The term "gene size" refers to the length from the start of the first annotated exon in the genome to the end of the last annotated exon, including introns. This definition excludes 5' and 3' untranslated regions, because these are often under-annotated.[3] We measure and compare size in two ways: the "absolute size" and the "relative size". The term "absolute size" refers to the number of base pairs or amino acids. The term "relative size" refers to the ranked size relative to other genes within the same genome. We use the term "CDS size" to refer to the span of nucleotides within a mature RNA transcript that will eventually be translated into protein, which excludes introns and untranslated regions. "Protein size" is measured by the number of amino acids.

Absolute gene and protein sizes in each species were obtained from reference genome assemblies and annotations from the National Center for Biotechnology and Information (NCBI). Gene start positions from the most 5' exon were subtracted from gene end positions (+1) of the most 3' exon to obtain a measure of absolute gene size for each gene that excludes explicitly annotated 5′ and 3′ UTRs.

Identification of orthologs

OrthoFinder[20] was used to identify orthologs across several representative eukaryotes with chromosomal-level genome assemblies (excepting *S. rosetta* and *A. queenslandica*). OrthoFinder identifies groups of orthologous genes (orthogroups), which may include paralogs. Ensembl was used for other "high-confidence", one-to-one orthologs as indicated in the text.

Gene ontology

H. sapiens gene ontology (GO terms) were obtained from Ensembl (ensembl.org),[85] Ensembl genes 108, GRCh38.p13. Synaptic genes were defined by GO term names containing "synapse" or "synaptic".

Species phylogeny

Species phylogeny for Figure 1A was based on recent chromosome-scale gene linkages.[41] Divergence times were obtained from TimeTree (timetree.org).[67] These incorporate molecular clock assumptions, which may lead to overestimates (discussed in Budd and Mann[25]). Phylogeny for Figure 5C was obtained from TimeTree and initially plotted using the Interactive Tree of Life.[84] Species outlines were obtained from phylopic.org.

Gene ages

Gene ages were obtained from the GenOrigin database (genorigin.chenzxlab.cn).[40] GenOrigin systematically infers gene age using a protein-family based pipeline (FBP) with Wagner parsimony algorithm, phylogeny derived from TimeTree,[67] and orthology information from Ensembl Compara.[24,87]

Species selection

The species analyzed in this study (Table S1) were chosen for the completeness of their genome assemblies, which has a significant impact on the quality and completeness of gene annotations. However, most complete genomes are biased for model organisms chosen for unique biological features with potential impacts on genome organization. As new genomes are sequenced to completion, the generality of these observations can be tested.

QUANTIFICATION AND STATISTICAL ANALYSIS

Ortholog feature comparisons

The mean ortholog size (gene and protein) within each orthogroup was used for comparisons among multiple species. Relative ortholog sizes were estimated using quantiles. In Figures 1 and S1, the relationship between features was visualized by a simple linear regression model using the function geom_smooth(method = "lm", …) from the ggplot2 R package. In the legend of Figure S1, spearman's rank correlation was estimated with the function cor(method = "spearman", …) from base R. To address the issue of multiple comparisons, p values were adjusted using the Benjamini-Hochberg correction method with the function p.adjust(method = "BH", …) from base R. All reported adjusted p values were less than 1.67e-143.

Enrichment of gene size within gene features and gene sets

To estimate enrichment of gene size within gene features (e.g., gene expression) and gene sets (e.g., GO terms, gene age, etc.), gene features or gene sets were first binned by gene size quantiles, then the probability of randomly obtaining the number of features or

Please cite this article in press as: McCoy and Fire, Parallel gene size and isoform expansion of ancient neuronal genes, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.02.021

genes within each gene size quantile was determined by a two-sided binomial test. The binomial test was executed with the function binom.test(x = observations, n = total_trials, p = 0.1, alternative = "two.sided") from base R, where x is the number of observations per bin, n is the total number of observations across all bins, and p is the expected probability of an observation per bin (10 bins) if assignment is random. To address the issue of multiple comparisons, p values were adjusted using the Benjamini-Hochberg correction method with the function p.adjust(method = “BH”, …) from base R, and an adjusted p value of < 0.005 was used as a cutoff for designating enrichment or depletion of observations per bin.

Single-cell gene expression analysis

For Figure S2B, single-cell gene expression data was obtained from the Human Protein Atlas. The top 10% largest genes in the dataset were used (2,009 genes), and z-score normalized gene expression values (normalized transcripts per million; nTPM) were analyzed. Cell types were ranked by a one-sided Mann-Whitney U test of mean gene expression compared to all other cell types using the function wilcox.test(x, y, alternative = "greater") from base R. To address the issue of multiple comparisons, p values were adjusted using the Benjamini-Hochberg correction method with the function p.adjust(method = “BH”, …) from base R, and the top 32 cells were shown with “Other” representing the average of the remaining 49 cell types. Cell types were ordered by euclidean distance of z-scale normalized expression data using the combined function hclust(dist(data, method = "euclidean"), method = "ward.D") in base R.

Current Biology, Volume *34*

Supplemental Information

Parallel gene size and isoform expansion of ancient neuronal genes

Matthew J. McCoy and Andrew Z. Fire

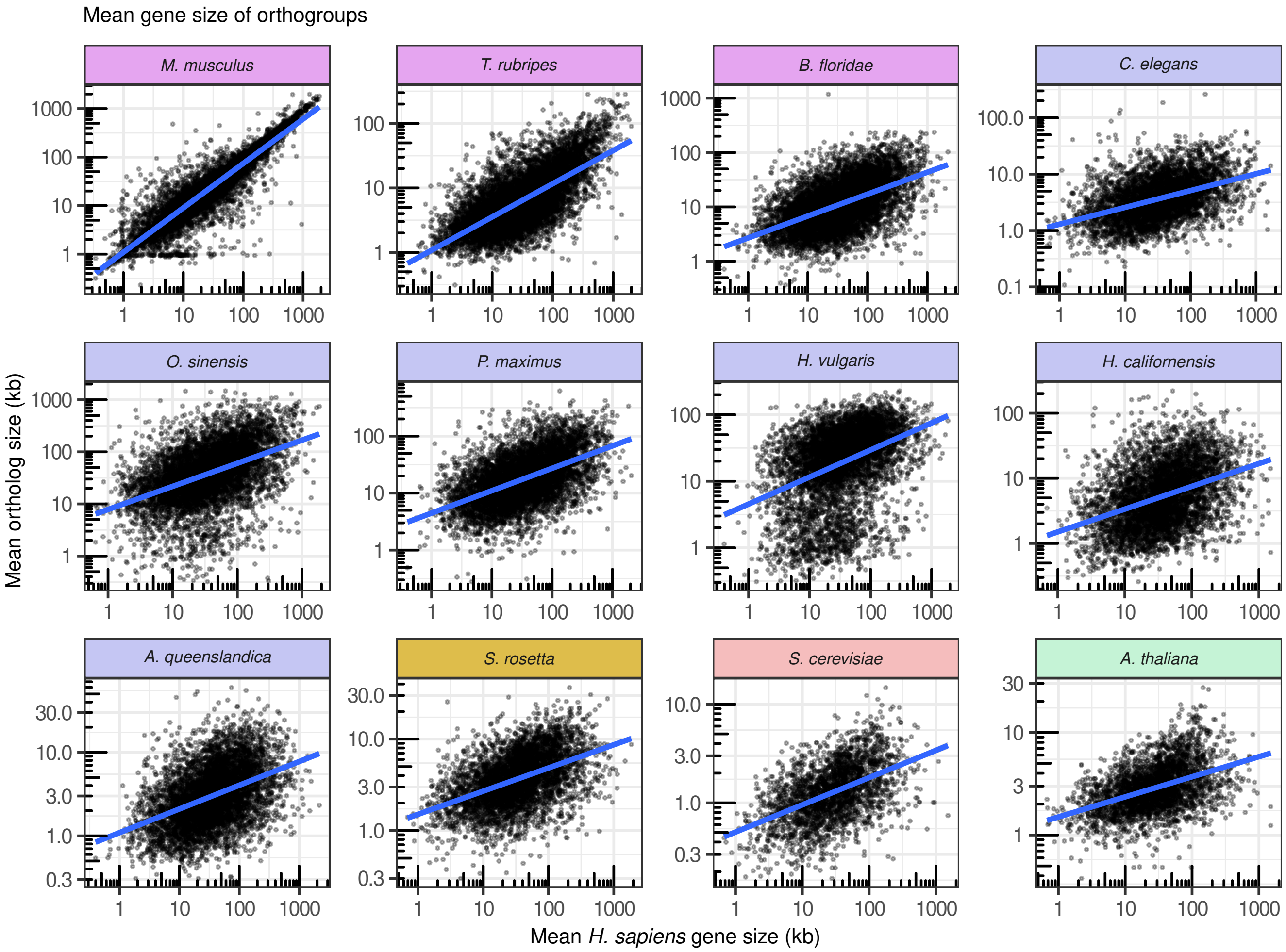
Mean gene size of orthogroups
M. musculus
T. rubripes
B. floridae
C. elegans
O. sinensis
P. maximus
H. vulgaris
H. californensis
A. queenslandica
S. rosetta
S. cerevisiae
A. thaliana
1000
100
10
1
100.0
10.0
1.0
0.1
30.0
3.0
0.3
30
3
Mean ortholog size (kb)
Mean H. sapiens gene size (kb)

Figure S1. Scatter plots of mean gene size of orthogroups, Related to Figure 1. For each group of orthologous genes between any two species (orthogroups), the mean human gene size is shown versus the mean ortholog size in other species. Solid blue lines show linear models with 95% confidence intervals as ribbons. Box colors: purple = vertebrates, blue = invertebrates, yellow = protists, red = fungi, green = plants. Spearman's rank correlation adjusted p < 1.67e-143 (Benjamini-Hochberg corrected) for all comparisons. Spearman's rho: *M. musculus* = 0.937; *T. rubripes* = 0.680; *B. floridae* = 0.528; *C. elegans* = 0.437; *O. sinensis* = 0.483; *P. maximus* = 0.526; *H. vulgaris* = 0.413; *H. californensis* = 0.390; *A. queenslandica* = 0.439; *S. rosetta* = 0.477; *S. cerevisiae* = 0.488; *A. thaliana* = 0.435.

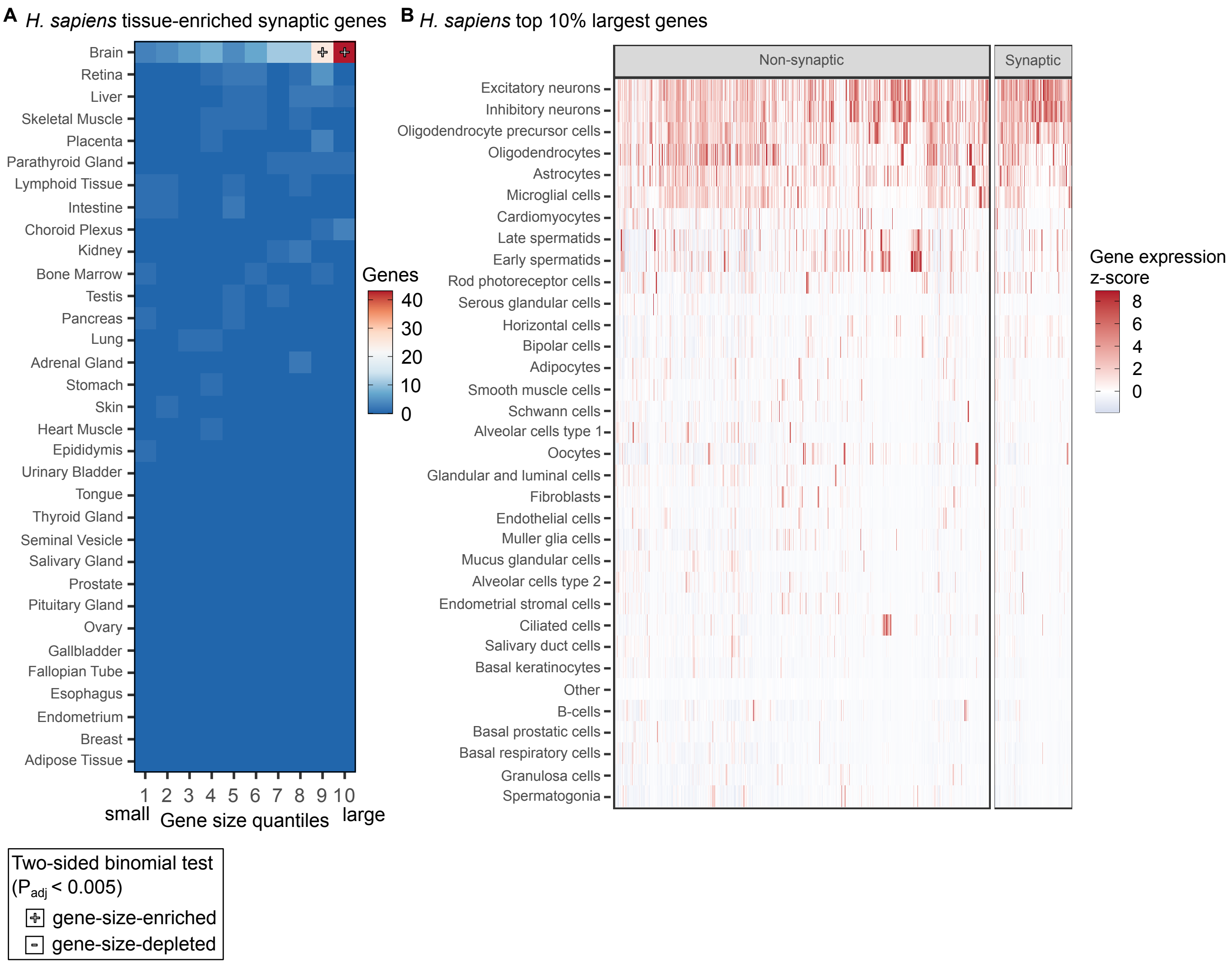
A
H. sapiens tissue-enriched synaptic genes
Brain
Retina
Liver
Skeletal Muscle
Placenta
Parathyroid Gland
Lymphoid Tissue
Intestine
Choroid Plexus
Kidney
Bone Marrow
Testis
Pancreas
Lung
Adrenal Gland
Stomach
Skin
Heart Muscle
Epididymis
Urinary Bladder
Tongue
Thyroid Gland
Seminal Vesicle
Salivary Gland
Prostate
Pituitary Gland
Ovary
Gallbladder
Fallopian Tube
Esophagus
Endometrium
Breast
Adipose Tissue
1 2 3 4 5 6 7 8 9 10
small
Gene size quantiles
large
Genes
40
30
20
10
0
Two-sided binomial test
(P_{adj} < 0.005)
⊞ gene-size-enriched
⊟ gene-size-depleted
B
H. sapiens top 10% largest genes
Non-synaptic
Synaptic
Excitatory neurons
Inhibitory neurons
Oligodendrocyte precursor cells
Oligodendrocytes
Astrocytes
Microglial cells
Cardiomyocytes
Late spermatids
Early spermatids
Rod photoreceptor cells
Serous glandular cells
Horizontal cells
Bipolar cells
Adipocytes
Smooth muscle cells
Schwann cells
Alveolar cells type 1
Oocytes
Glandular and luminal cells
Fibroblasts
Endothelial cells
Muller glia cells
Mucus glandular cells
Alveolar cells type 2
Endometrial stromal cells
Ciliated cells
Salivary duct cells
Basal keratinocytes
Other
B-cells
Basal prostatic cells
Basal respiratory cells
Granulosa cells
Spermatogonia
Gene expression
z-score
8
6
4
2
0

Figure S2. Brain tissue and neural cell types are enriched for large synaptic genes, Related to Figure 2. (**A**) Heatmap of Human Protein Atlas[S1] tissue-enriched synaptic genes (GO term matching "synaptic" or "synapse") binned by gene size quantiles (10 bins)(**Data S3A**). Heat colors show the number of synaptic genes in each bin. Tissues are ordered by the total number of enriched synaptic genes across all gene sizes in each tissue. Gene size enrichment (+) or depletion (-) of observations per tissue was determined by a two-sided binomial test ($p = 0.1$) with Benjamini-Hochberg adjusted p-values < 0.005 based on the value expected for uniform size distribution for each tissue (**Data S3E**). (**B**) Heatmap of single-cell gene expression from the Human Protein Atlas[S1]. Heat colors show z-score normalized gene expression values (nTPM) across the 10% largest human genes (2009 genes). Cell types were ranked by one-sided Mann-Whitney U test adjusted p-values (Benjamini-Hochberg) of mean gene expression compared to all other cell types, and the top 32 cells are shown with "Other" representing the average of the remaining 49 cell types. Genes were binned into "synaptic" (GO term matching "synaptic" or "synapse") or "non-synaptic" (**Data S3C**).

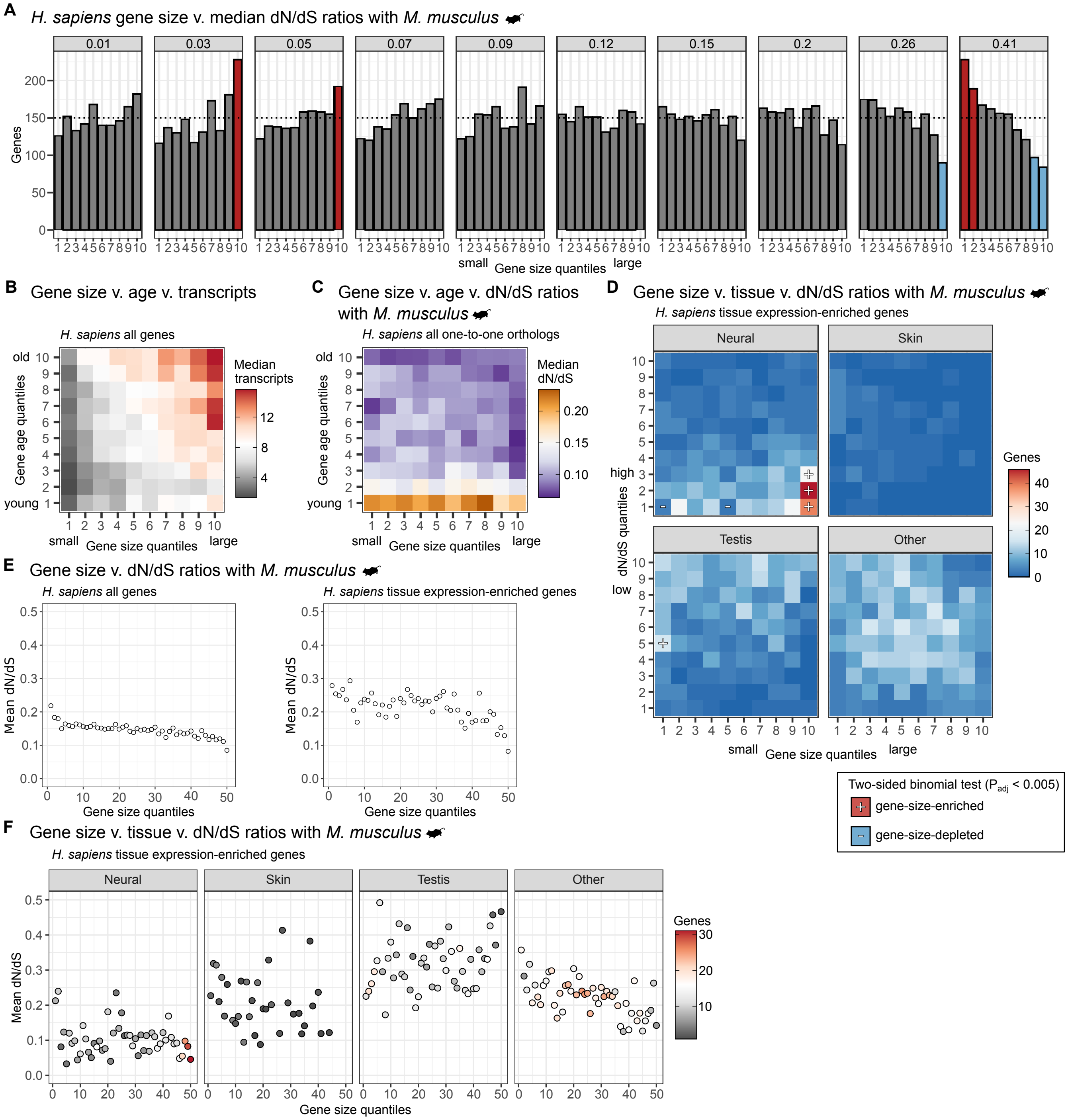
A
H. sapiens gene size v. median dN/dS ratios with M. musculus
0.01
0.03
0.05
0.07
0.09
0.12
0.15
0.2
0.26
0.41
Genes
0
50
100
150
200
1 2 3 4 5 6 7 8 9 10
small
Gene size quantiles
large
B
Gene size v. age v. transcripts
H. sapiens all genes
Gene age quantiles
old 10
9
8
7
6
5
4
3
2
young 1
1 2 3 4 5 6 7 8 9 10
small
Gene size quantiles
large
Median transcripts
12
8
4
C
Gene size v. age v. dN/dS ratios with M. musculus
H. sapiens all one-to-one orthologs
Gene age quantiles
old 10
young 1
small
Gene size quantiles
large
Median dN/dS
0.20
0.15
0.10
D
Gene size v. tissue v. dN/dS ratios with M. musculus
H. sapiens tissue expression-enriched genes
Neural
Skin
Testis
Other
dN/dS quantiles
high
low
1 2 3 4 5 6 7 8 9 10
small
Gene size quantiles
large
Genes
40
30
20
10
0
Two-sided binomial test (P_{adj} < 0.005)
+ gene-size-enriched
− gene-size-depleted
E
Gene size v. dN/dS ratios with M. musculus
H. sapiens all genes
H. sapiens tissue expression-enriched genes
Mean dN/dS
0.0
0.1
0.2
0.3
0.4
0.5
0
10
20
30
40
50
Gene size quantiles
F
Gene size v. tissue v. dN/dS ratios with M. musculus
H. sapiens tissue expression-enriched genes
Neural
Skin
Testis
Other
Mean dN/dS
Gene size quantiles
Genes
30
20
10

Figure S3. Relationship between gene size, transcript number, tissue, and dN/dS ratios, Related to Figure 4. (**A**) Bar graphs of human genes binned independently by gene size (10 bins) and dN/dS ratio (10 bins; *H. sapiens - M. musculus*). Median dN/dS ratios per bin are shown (**Data S5M,N**). (**B**) Heatmap of median transcript counts for human genes binned independently by gene size (10 bins), and gene age (10 bins; values from GenOrigin[S2])(**Data S5I**). (**C**) Heatmap of gene median dN/dS ratios (*H. sapiens - M. musculus*) for human genes binned independently by gene size (10 bins) and gene age (10 bins) for Ensembl[S3] one-to-one orthologs (**Data S5J**). (**D**) Heatmaps of gene number for human genes binned independently by gene size (10 bins) and dN/dS ratio (10 bins) for Human Protein Atlas[S1] tissue-enriched genes. Neural, Skin and Testis are shown with "Other" combining all other tissues (**Data S5K,L**). (**E**) Scatter plots of human genes binned by gene size (50 bins), showing the mean dN/dS ratio between mouse and human per gene size bin. All human genes (left), tissue-enriched genes (right). (**F**) Scatter plots of human tissue-enriched genes binned together by gene size (50 bins), showing the mean dN/dS ratio between mouse and human per gene size bin for each tissue. Colors show the number of genes from each tissue in each gene size bin. Enrichment (+/red) or depletion (-/blue) of observations per dN/dS bin and gene size bin was determined by a two-sided binomial test ($p = 0.1$) with Benjamini-Hochberg adjusted p-values < 0.005.

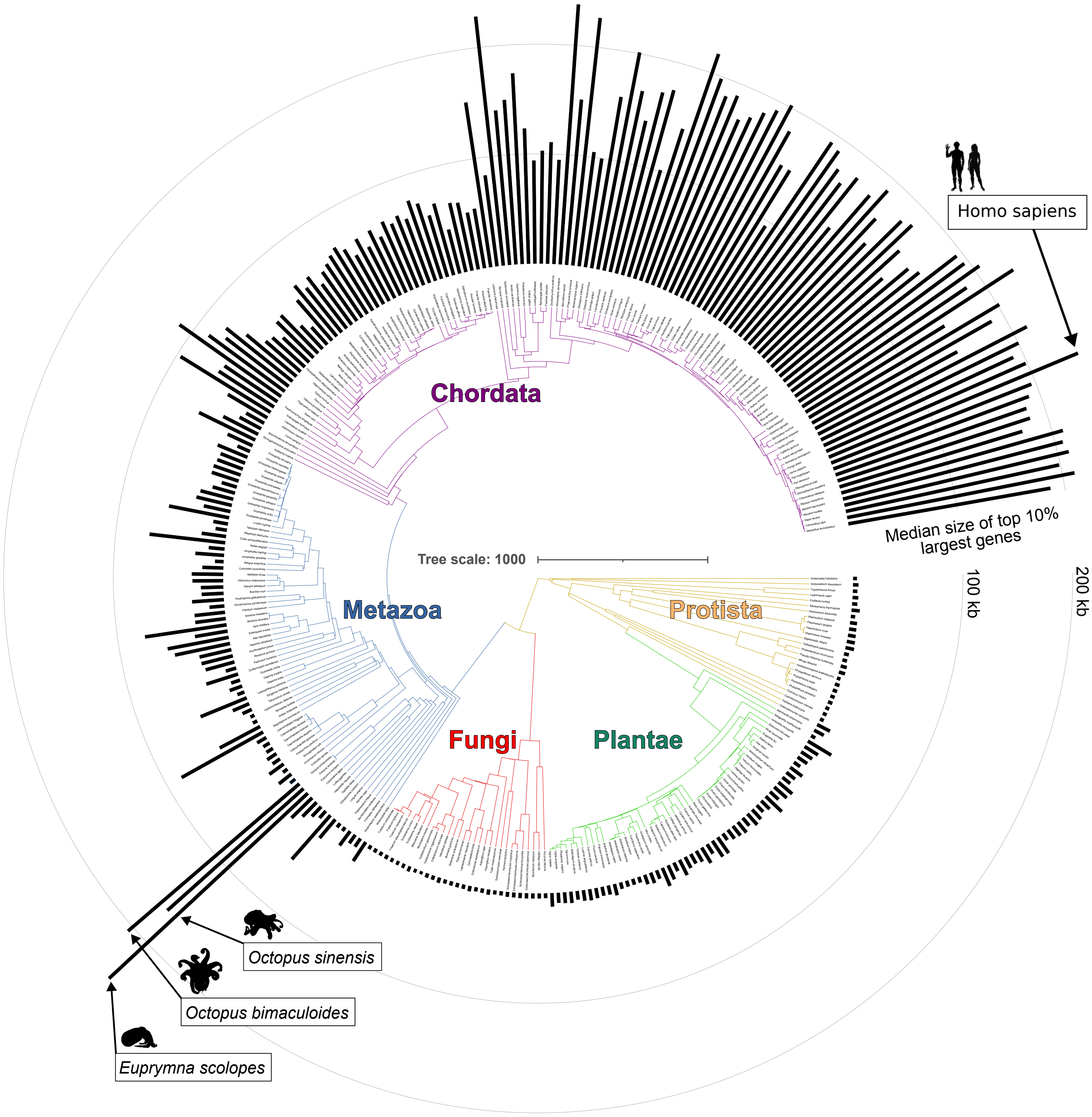
Homo sapiens
Median size of top 10%
largest genes
100 kb
200 kb
Tree scale: 1000
Chordata
Metazoa
Protista
Fungi
Plantae
Octopus sinensis
Octopus bimaculoides
Euprymna scolopes

Figure S4. Independent gene size expansion in diverse species, Related to Figure 5. Phylogenomic tree showing median gene size of the top 10% largest genes in each genome as a bargraph. Phylogeny and branch lengths were obtained from TimeTree.org[S4], with the exception of *Mnemiopsis leidyi* and *Amphimedon queenslandica*, which were placed according to Schultz et al.[S5]. The tree was plotted with iTOL[S6]. Scale bar shows 100 million years. Adapted from McCoy and Fire[S7] to include updated gene size estimates for cephalopods *Octopus bimaculoides, Octopus sinensis,* and *Euprymna scolopes*. Related to **Data S6**.

Species	RefSeq/GenBank	Assembly	Level
Amphimedon queenslandica	GCF_000090795.1	v1.0	Scaffold
Arabidopsis thaliana	GCF_000001735.4	TAIR10.1	Chromosome
Branchiostoma floridae	GCF_000003815.2		Chromosome
Caenorhabditis elegans	GCF_000002985.6	WBcel235	Chromosome
Euprymna scolopes	GCA_024364815.1	ASM2436480v1	Chromosome
Homo sapiens	GCF_009914755.1		Chromosome
Hormiphora californensis	GCA_020137815.1	Hcv1a1d20200309	Chromosome
Hydra vulgaris	GCF_022113875.1		Chromosome
Mus musculus	GCF_000001635.27		Chromosome
Octopus bimaculoides	GCF_001194135.2	ASM119413v2	Chromosome
Octopus sinensis	GCF_006345805.1		Chromosome
Pecten maximus	GCF_902652985.1		Chromosome
Saccharomyces cerevisiae	GCF_000146045.2	R64	Chromosome
Salpingoeca rosetta	GCF_000188695.1	Proterospongia_sp_ATCC50818	Scaffold
Takifugu rubripes	GCF_901000725.2	fTakRub1.2	Chromosome

Table S1. Primary genome assemblies used in this study. Related to Figure 1.

Supplemental References

S1. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. Science *347*, 1260419. 10.1126/science.1260419.

S2. Tong, Y.-B., Shi, M.-W., Qian, S.H., Chen, Y.-J., Luo, Z.-H., Tu, Y.-X., Xiong, Y.-L., Geng, Y.-J., Chen, C., and Chen, Z.-X. (2021). GenOrigin: A comprehensive protein-coding gene origination database on the evolutionary timescale of life. J. Genet. Genomics *48*, 1122–1129. 10.1016/j.jgg.2021.03.018.

S3. Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M.J., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics resources. Database. 10.1093/database/bav096.

S4. Kumar, S., Suleski, M., Craig, J.M., Kasprowicz, A.E., Sanderford, M., Li, M., Stecher, G., and Hedges, S.B. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. Mol. Biol. Evol. 10.1093/molbev/msac174.

S5. Schultz, D.T., Haddock, S.H.D., Bredeson, J.V., Green, R.E., Simakov, O., and Rokhsar, D.S. (2023). Ancient gene linkages support ctenophores as sister to other animals. Nature *618*, 110–117. 10.1038/s41586-023-05936-6.

S6. McCoy, M.J., and Fire, A.Z. (2020). Intron and gene size expansion during nervous system evolution. BMC Genomics *21*, 360. 10.1186/s12864-020-6760-4.