McCoy and Fire *BMC Genomics* (2020) 21:360
https://doi.org/10.1186/s12864-020-6760-4

BMC Genomics

RESEARCH ARTICLE Open Access

Intron and gene size expansion during nervous system evolution

Matthew J. McCoy[1,2*] and Andrew Z. Fire[2*]

Abstract

Background: The evolutionary radiation of animals was accompanied by extensive expansion of gene and genome sizes, increased isoform diversity, and complexity of regulation.

Results: Here we show that the longest genes are enriched for expression in neuronal tissues of diverse vertebrates and of invertebrates. Additionally, we show that neuronal gene size expansion occurred predominantly through net gains in intron size, with a positional bias toward the 5′ end of each gene.

Conclusions: We find that intron and gene size expansion is a feature of many genes whose expression is enriched in nervous systems. We speculate that unique attributes of neurons may subject neuronal genes to evolutionary forces favoring net size expansion. This process could be associated with tissue-specific constraints on gene function and/or the evolution of increasingly complex gene regulation in nervous systems.

Keywords: Genome evolution, Gene size, Intron size, Nervous system evolution, Long genes, Long introns

Background

Genome size and complexity greatly increase from prokaryotes to multicellular eukaryotes [1]. The changes include increases in the size and number of introns, in large part through dramatic proliferation and accumulation of mobile genetic elements [2–4] and through shifts in the balance of deletions and insertions [5]. The differential expansion of intronic sequences has led to orders of magnitude variation in overall gene lengths within genomes and among species across diverse phylogenetic lineages (see Additional file 1: Figure S1 for an example in which gene size and density vary dramatically among animal orthologs of *NRXN1*, which encode cell-surface receptors localized at neuronal synapses).

The evolutionary consequences of gene size expansion are still being uncovered. Although the causes and consequences may be intertwined, both are areas of valuable investigation. Longer genes tend to have more splice variants [6], which give rise to alternative isoforms with distinct molecular functions and expression patterns. Additionally, long genes can have greater numbers of regulatory elements, with these arrayed in more diverse patterns [7]. Over 10% of human protein-coding genes are greater than 100 kb in length, with the time required to transcribe these long genes being significant on a cellular, developmental and organismal level [8]. Although exact transcription rates are unknown, assuming the estimated average rates of transcription in eukaryotes of 1–4 kb per minute [9, 10], the 2.3 Mb human gene *CNTNAP2* would require upwards of 10 h to generate one transcript, which is substantially longer than the 10 min required for the average mammalian gene ~ 10 kb in length [11]. This phenomenon, termed intron delay, biases patterns of expression and evolution of gene structure for a substantial portion of the transcriptome [12, 13]. We hypothesize that this potentially limits expression of long genes to long-lived post-mitotic cells, such as neurons. Supporting this, recent transcriptomic studies in rodents and humans have revealed that

* Correspondence: mjmccoy@stanford.edu; afire@stanford.edu
[1] Grass Fellowship Program, Marine Biological Laboratory, Woods Hole, MA 02543, USA
[2] Departments of Pathology and Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

expression of the longest genes is enriched in neuronal tissues [7, 14–17]. Additionally, many of these genes encode ion channels, receptors, and cell adhesion molecules [15], which greatly contribute to the extensive molecular diversity that exists among neuronal subtypes within rodents and humans [7, 17, 69]. Finally, a role for long genes in genetic causation of cancer and neurological disease is becoming apparent [14, 15, 70, 71], potentially from elevated vulnerability to mutation [70, 71]. Using evidence from more extensive phylogenomics and transcriptomics, we show here that long gene expression enrichment in neuronal tissues is a phenomenon extending beyond rodents and humans. Additionally, we show that net expansion of neuronal intron and gene size is a feature of many nervous systems.

Results

Enrichment of long gene expression in neuronal tissues has previously been observed in rodents and humans [7, 14–17]. By analyzing systems with well-defined tissue transcriptomes [18–20], we asked whether neuronal long gene expression is a feature shared across diverse taxonomic classes both within the chordate phylum and among invertebrates. We examined gene expression as a function of gene length (see Gene Expression Analysis in Methods) in multiple tissues for diverse animals and found that as gene length increases, so does expression enrichment in neuronal tissues (Fig. 1a).

Neuronal long gene expression is also apparent in invertebrates, such as the insect arthropod, *Drosophila melanogaster*, and the cephalopod mollusc, *Octopus bimaculoides* (Fig. 1a). As observed in comparing read counts in different tissues with annotated gene lengths, both assembled genomes exhibit enrichment for neuronal transcripts in size ranges at the high end of the gene-length distribution (e.g. the top ~ 14% of genes in *D. melanogaster,* which are > 10 kb, and the top ~ 14% of genes in *O. bimaculoides*, which are > 30 kb). Cephalopods are particularly interesting in part because they have the largest known invertebrate nervous systems [32], along with sophisticated behaviors [33]. Cephalopods are estimated to have shared their most recent common ancestor with insects more than 750 million years ago [22], which is only ~ 50 million years later than the estimated pairwise divergence between vertebrates and invertebrates [22]. Together, these observations support either an ancient origin of neuronal long gene preference dating back to the evolution of the first nervous systems or convergence toward this phenomenon in multiple animal lineages.

Within eukaryotic genes, individual intron lengths negatively correlate with intron ordinal position [34, 35], with 5′ introns being larger and containing more conserved sequences relative to downstream introns [36, 37]. Here we analyzed intron length versus intron ordinal position for genes enriched in different tissues. To calculate a rough metric for tissue-specificity, we stratified genes based on maximal expression in each tissue, then calculated fold-difference in expression over the next-highest expressing tissue. In genes for which expression was substantially enriched in neuronal tissues (> 5-fold, Fig. 1b), we observed striking enrichment for long introns, particularly near the 5′ ends of genes. This enrichment declines toward the 3′ ends of genes, and in genes with lesser neuronal expression enrichment (< 2-fold, Fig. 1b; > 2-fold, Additional file 3: Figure S2).

The rapidly developing invertebrate nematode, *Caenorhabditis elegans* has a very compact genome and compact gene sizes (Fig. 1a). Nonetheless, indication of longer introns in neuronally expressed *C. elegans* genes is evident, as is some indication of neuronal enrichment in the longest genes (Fig. 1a; see Additional file 4: Figures S3, Additional file 5: Figures S4, Additional file 6: Figures S5, Additional file 7: Figures S6 for alternative normalization and plotting; see Gene Expression Analysis in Methods for further discussion). A detailed tissue-specific intron and gene length analysis was also possible for *Zea mays* (Fig. 1), showing compact genes and no evident expression enrichment for the longest genes in any tissues.

Large 5′ introns and the overall length increases of genes enriched in neuronal tissues of both vertebrates and invertebrates may be linked to broader trends in animal genome evolution. For example, although many orthologs of long neuronal genes are relatively long in other animals, gene size and density are strikingly different between vertebrates and invertebrates (Additional file 1: Figure S1). To highlight these differences, we compared protein-coding gene lengths across 325 eukaryotic species currently available on the Ensembl genome database [38]. We chose a measurement of gene length spanning the start of the first exon to the end of the last exon, in part to avoid known differences in annotation completeness of untranslated regions (UTRs) [39]. To compare gene lengths across eukaryotic species, we calculated the median gene lengths within each species (Fig. 2a), including in Protista ($\bar{x}$ = 1.36 ± 0.33 kb [standard deviation]), Plantae ($\bar{x}$ = 2.42 ± 0.71 kb), Fungi ($\bar{x}$ = 1.46 ± 0.26 kb), Metazoa ($\bar{x}$ = 2.79 ± 1.71 kb; excluding chordates), and Chordata ($\bar{x}$ = 12.88 ± 5.43 kb). Gene length variation was most dramatic for the top 10% longest genes, which were an order of magnitude longer in vertebrates ($\bar{x}$ = 129.23 ± 64.12 kb) compared to non-animal eukaryotes ($\bar{x}$ = 7.29 ± 4.26 kb)(Fig. 2b). Consistent with other reports [39, 40], the vast majority of the increases in gene length are intronic (Fig. 2c), with increases in intron length coinciding with increases in the average number of exons per gene (Fig. 2d;

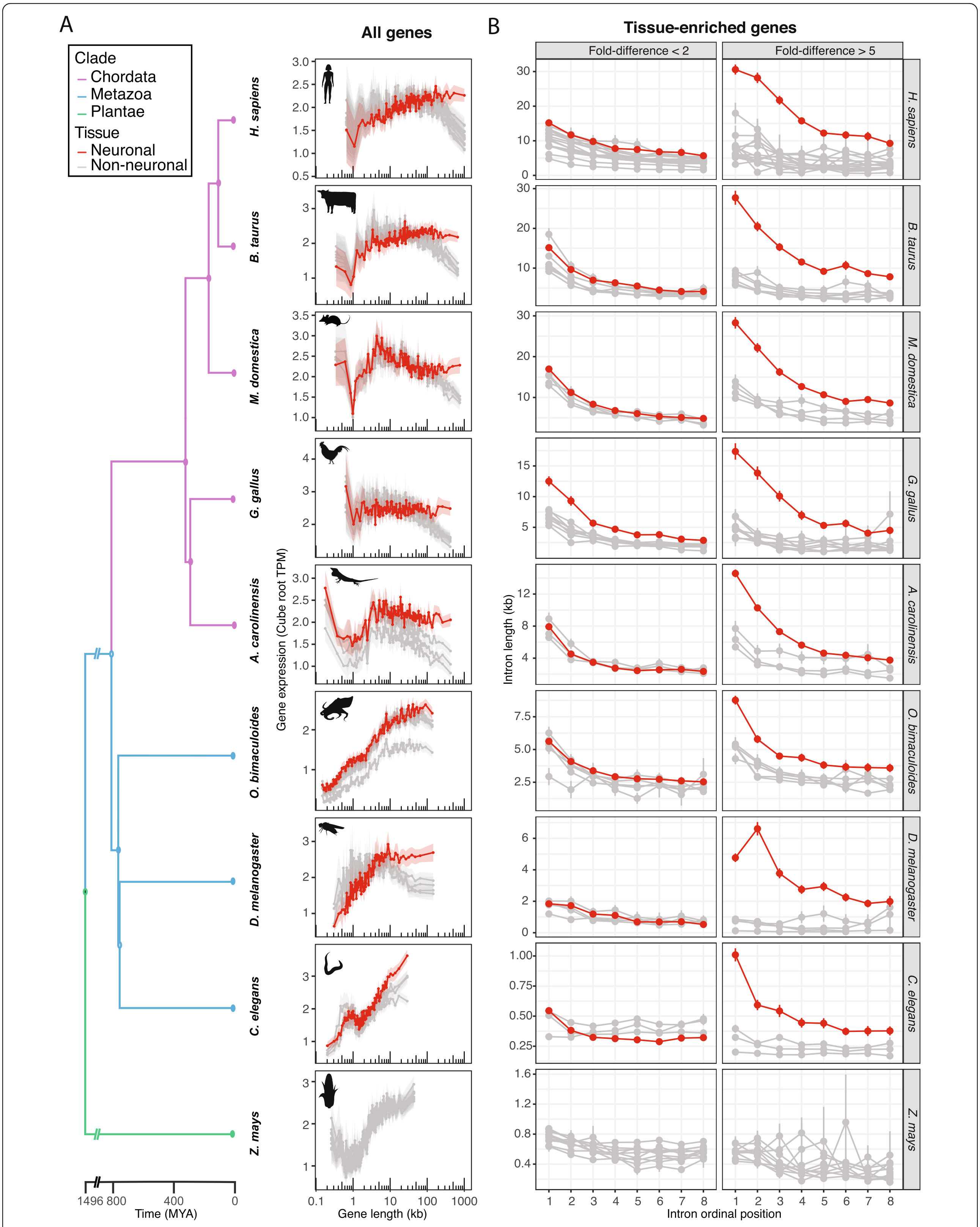

Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Neuronal intron and gene size expansion in diverse animal species. **a** Expression of the longest genes is enriched in neuronal tissues of vertebrates and invertebrates. Each plot shows smooth-quantile-normalized transcriptome data across multiple tissues or cell types from individual species. Each line represents mean gene expression in cube root transcripts per million (TPM) versus gene length (kb). Genes were segregated into 100 bins according to gene length, and points show average gene length of each bin. Transparent ribbons show 95% confidence intervals. Red lines show neuronal tissues and grey lines show non-neuronal tissues (see Additional file 2: Table S1 for a full list of tissues for each species). Tissue transcriptomes were obtained from the EMBL Expression Atlas [21] for all species except *Drosophila melanogaster* [18] and *Octopus bimaculoides* [19]. Branch lengths for the phylogenetic tree were obtained from TimeTree.org [22]. Organism outlines were depicted by author MJM from the following image references in accordance with respective licensing: *Homo sapiens* [23]; *Bos taurus* [24]; *Monodelphis domestica* [25]; *Gallus gallus* [26]; *Anolis carolinensis* [27]; *Octopus bimaculoides* [28]; *Drosophila melanogaster* [29]; *Caenorhabditis elegans* [30]; and *Zea mays* [31]. **b** Mean intron length (kb) versus ordinal position in neuronal (red) and non-neuronal tissues (grey) across the same species presented in Fig. 1a. Genes with only slight expression fold-difference of the top expressing tissue over the next highest-expressing tissue (< 2 fold-difference; left) are contrasted against genes with more than 5 fold-difference in expression (right). Error bars show standard error

Additional file 8: Figure S7). Within Metazoa, multicellular animals lacking nervous systems (e.g. placazoa and sponges) had significantly shorter genes than most animals possessing nervous systems ($p < 2.2e\text{-}16$, wilcoxon signed rank test with continuity correction), even when only considering invertebrates ($p = 4.0e\text{-}10$) (Fig. 2a). Many of the observed differences were consistent within species of the same clade, some of which include more extensively annotated organisms (e.g. *Homo sapiens*, *Drosophila melanogaster*, *Arabidopsis thaliana*, etc.), but since annotation quality and completeness are continually improving [39], the exact values presented here for individual species are expected to change in the future. Although the evolutionary dynamics of gene size expansion alone do not explain how the longest genes became enriched for neuronal expression in particular, the observed median sizes are consistent with metazoan gene size expansion beginning during the evolutionary divergence of animals in the Precambrian eon [41]. Together with evidence of neuronal long gene expression in vertebrates and invertebrates, this suggests that intron and gene size expansion of neuronal genes could have begun during the evolution of the first nervous systems.

Discussion

The evolutionary course of intron and gene size net expansion or net contraction is expected to be a combined function of insertion/deletion frequencies, of selective pressure, and of neutral processes such as genetic drift. Determining the individual and collective contribution of these forces is not trivial, with a primary challenge coming from the currently necessary use of static distributions of gene sizes across Eukarya as presented here. Nonetheless, certain features emerge from the analysis even with the static picture. Our observation that 5′ end introns of neuronal genes are longer than in genes enriched in other tissues is one such feature, potentially indicating differences between tissues that influence tolerance for insertion/deletion events. For eukaryotic genes in general, intron length bias near 5′ ends of genes is generally thought to emerge over evolutionary time from a shift in the insertion/deletion spectrum in favor of insertions, since longer deletions more frequently disrupt functional elements that are enriched near 5′ ends of genes [36, 37, 72].

Regardless of how introns undergo net size expansion, as introns lengthen, they have higher probabilities of gaining novel exons, which contribute to alternative isoform diversity [42, 43]. Furthermore, new exon-containing isoforms are typically rare, which potentiates the generation of novel functions with minimal damage to the existing functional repertoire [44]. These previous observations have contributed to hypotheses regarding the role of alternative splicing in the evolution of overall phenotypic complexity [45, 46]. We speculate that intron and gene size expansion parallels functional element complexity of neuronal genes, and that size expansion may potentiate further increases in gene regulatory complexity and molecular diversity of nervous systems.

There are several outliers for gene and genome size among animals. The microbat (*Myotis lucifugus*) has one of the smallest mammalian genomes at ~ 2 Gb, likely from extensive DNA loss [47] and correspondingly smaller gene sizes (Fig. 2a). There is some evidence that the metabolic cost of powered flight is associated with smaller genomes [48], and that the smaller sizes observed in bats and birds results from DNA loss through segmental deletions rather than decreased repetitive element expansion [47]. In other seeming exceptions, reduced gene sizes could potentially arise from technical limitations as opposed to bona fide biological differences. One possible example of this is the American black bear (*Ursus americanus*) whose genome assembly has low contiguity (contig N50 = 27,247 bp; scaffold N50 = 189,900 bp) [49], which could limit gene annotation completeness. Certain fish also exhibit compact gene and genome sizes, such as the teleosts which underwent an additional round of whole-genome duplication approximately 320–400 million years ago [50]. Genes with larger gene families tend to have shorter genes [6], which has been proposed to result from the partial nature of duplications [51] as longer genes

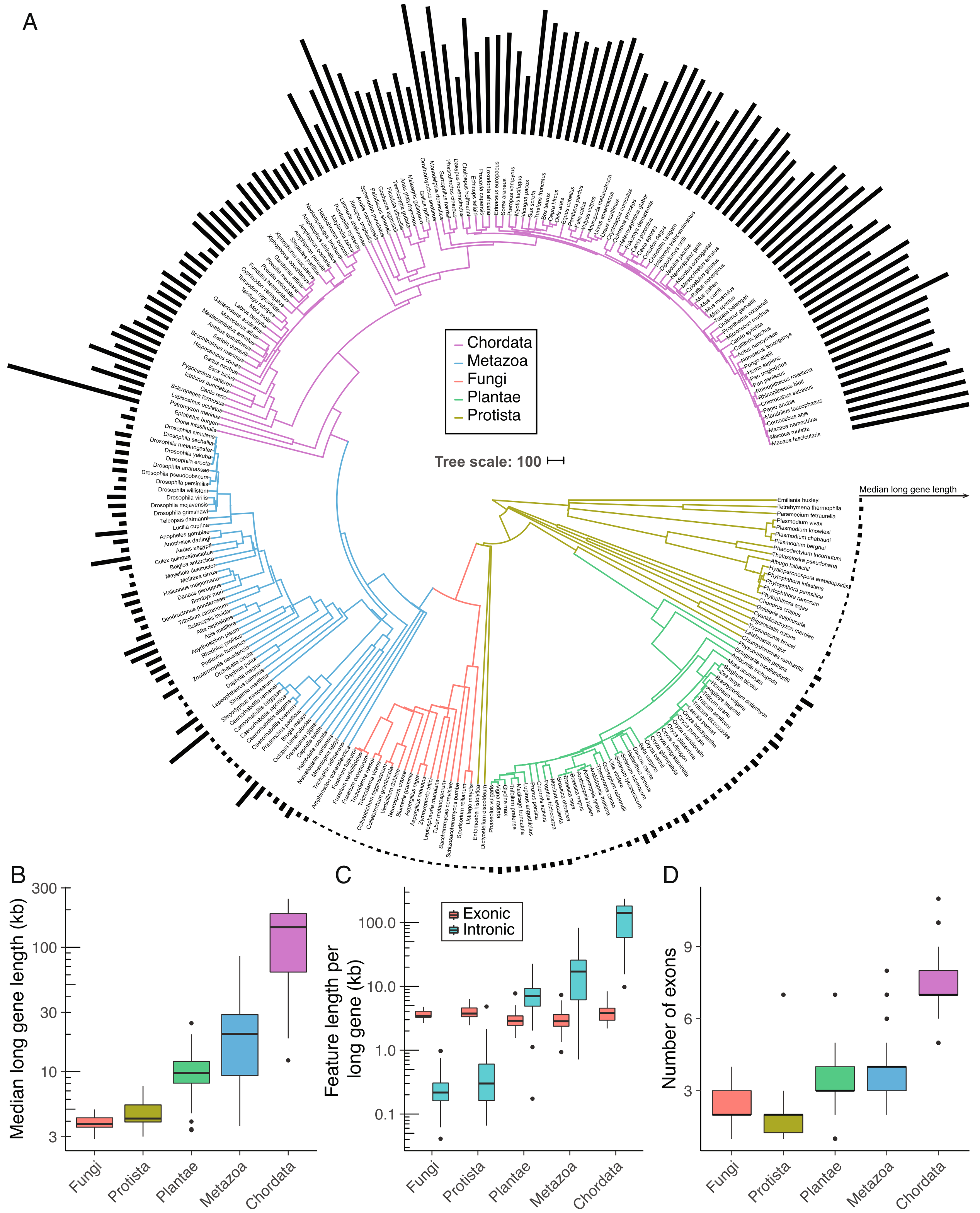

Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 Differential gene size expansion during animal evolution. **a** Phylogenomic tree showing median gene length of the top 10% longest genes in each genome as a bargraph. All eukaryotic kingdoms are represented, with Metazoa further subdivided into animals with notochords (Chordata) and animals without notochords (here labeled as Metazoa). Branch lengths were obtained from TimeTree.org [22], and scale bar shows 100 million years. **b** Boxplots of median gene length of long genes of each genome in kilobases aggregated by clade. **c** Boxplots of long gene intronic sequence versus exonic sequence per clade. **d** Boxplots showing median number of exons of each gene for each species grouped by clade. This extends similar analyses by Lynch et al. [40] and Francis and Wörheide [39] of exon/intron content

undergo incomplete duplication more frequently than shorter genes. However, in the context of duplicated whole-genomes, smaller gene sizes may reflect reduced selective pressure to maintain functional sequences within introns of a duplicated gene. Conversely, hagfish (*Eptatretus burgeri*) have one of the larger median gene lengths (Fig. 2a). Both hagfish and lamprey are jawless vertebrates and the only known chordates to undergo dramatic remodeling of their genome, eliminating hundreds of millions of base pairs from many somatic cells during embryonic development [52]. It remains unknown how these extreme rearrangements contribute to gene and genome size evolution. While these examples highlight several parameters potentially affecting the dynamics of gene and genome size evolution at the organismal level, our study indicates that biological parameters at the tissue and cellular level may also shape aspects of gene and genome architecture.

Conclusions

The molecular components of nervous systems are similar across animals despite vast differences in nervous system complexity, with some of the same sets of proteins and signaling molecules being used repeatedly in different contexts [53]. Intron and gene size expansion provide one mechanism by which existing components can be modified to increase transcriptomic potential from a limited gene complement, contributing to the evolution of novelty [6, 54] and potentially to the molecular complexity that is a hallmark of nervous systems. In this study, we show that many of the longest genes are preferentially expressed in neuronal tissues across diverse vertebrates and invertebrates. Additionally, we find longer introns in genes with neuronal expression enrichment, with introns near the 5′ ends of such genes being particularly lengthy. Determining the degree to which intron and gene size expansion are under the influence of adaptive and/or non-adaptive forces will be critical to understanding their evolutionary trajectory and phenotypic consequences. We hypothesize that intron and gene size expansion is uniquely tolerated, or perhaps uniquely advantageous in neuronal genes, and that expansion may enable the acquisition of new exons, alternative isoforms, and enhancers active within nervous tissue.

Methods

Statistical analysis

All statistical tests were performed in R version 3.5.0 (R Core Team 2018) and RStudio version 1.1.453 (RStudio Team 2015). All analyses are available as R scripts accompanied by data tables.

NRNX1 Ortholog display

Orthologs of the human *NRXN1* gene were obtained by using the Drosophila RNAi Screening Center Integrative Ortholog Prediction Tool (DIOPT, version 8.0) [55], which integrates ortholog predictions from Ensembl Compara, HomoloGene, Inparanoid, Isobase, OMA, orthoMCL, Phylome, RoundUp, and TreeFam. The highest scoring reciprocal hits were selected for display.

Gene expression analysis

RNA-seq data were obtained from NCBI's Sequence Read Archive (SRA) for `O. bimaculoides` (PRJNA285380) [19] and `D. melanogaster` (SRR070409, SRR070410, SRR070408, SRR070405, SRR070406, SRR070392, SRR070393, SRR070407, SRR070425, SRR100268) [18]. `O. bimaculoides` reads was aligned to `Octopus bimaculoides` v2_0 reference genome and `D. melanogaster` reads to `Drosophila melanogaster` Release 6 reference genome using the R package 'Rsubread' [56], and read counts were normalized to TPM. All other transcriptomes were obtained as TPM-normalized counts from the EMBL Expression Atlas [21]: `H. sapiens` (E-MTAB-513) [57], `B. taurus` (E-MTAB-2596) [58], `M. domestica` (E-MTAB-3719) [59], `G. gallus` (E-MTAB-2797) [60], `A. carolinensis` (E-MTAB-3727) [57], `C. elegans` (E-MTAB-2812) [61], and `Z. mays` (E-GEOD-62778) [62].

Smoothed conditional means of gene expression conditional on gene length (or gene length percentiles) were estimated using the 'geom_smooth' function from R package ggplot2 version 2.2.1 [63]. Two different smoothing methods were used: 'gam', which is a generalized additive model with formula = y ~ s(x, bs = "cs") using the restricted maximum likelihood (REML) approach for smoothing parameter estimation; and 'loess', which is a local polynomial regression fitting with formula = y ~ x and span (neighborhood) of proportion 0.1. To visualize the data without the underlying assumptions of smoothing across gene lengths varying by several orders of magnitude, we also (i) analyzed gene expression across gene length percentiles and observed a similar expression enrichment in neuronal tissues for the longest genes (Additional file 5: Figures S4, Additional file 6: Figures S5, Additional file 7: Figures S6), and (ii) analyzed mean gene expression of genes binned into 100 bins (Fig. 1a; Additional file 4: Figures S3, Additional file 5: Figures S4, Additional file 7: Figures S6). In addition to TPM normalization, data in Fig. 1a and Additional file 6: Figures S5, Additional file 7: Figures S6 were subjected to a generalization of quantile normalization called smooth quantile normalization (function 'qsmooth' from R package qsmooth version 1.2.0), which is based on the assumption that the statistical distribution of each sample should be the same within biological groups (e.g. tissue type), but allowing that they may differ between groups [64]. All analytical pipelines gave essentially the same result for gene expression versus gene length, with the possible exception of `C. elegans`.

For tissue-enrichment analyses (Fig. 1b), genes were only considered tissue-enriched if their expression value in their maximal tissue exceeded 50 times the lowest value for that tissue.

Gene architecture quantification

Gene lengths in each species were obtained from Ensembl BioMart version 2.40.4 [38]. Gene start positions were subtracted from gene end positions to obtain a measure of gene length for each gene that excludes 5′ and 3′ UTRs. Protein coding genes were selected using gene biotype information. Human orthologs were also obtained from Ensembl BioMart [38]. Exon numbers were estimated after determining the union of known exon-coding sequences (Fig. 2d), or by calculating the average number of exons across all transcripts of each gene (Additional file 8: Figure S7). Intron lengths were calculated from exon positions of each transcript of each gene. Exonic and intronic content were determined from the union of all annotated exon and intron sequences for each isoform of each gene, respectively (Fig. 2c). Protein-coding gene lengths for each clade were reported in the main text first by obtaining the median gene lengths of either all genes or only the top 10% longest genes in each genome, then obtaining the mean and standard deviation of these values across each clade. Significant differences in gene lengths between a species and all other species of a clade were determined by wilcoxon signed rank test with continuity correction.

Quantification of divergence

The branch lengths obtained from TimeTree.org [22] were combined with gene length annotations using the R packages ape version 5.1 [65] and phylobase version 0.8.4 [66]. Data was visualized using the R packages ggplot2 version 2.2.1 [63] and RColorBrewer version 1.1–2 [67], as well as using the Interactive Tree of Life (itol.embl.de) [68].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-6760-4.

Additional file 1: Figure S1. An exemplary gene size and density comparison at *NRXN1* ortholog loci. For each species, a *NRXN1* ortholog (black; *NRXN1* in *Homo sapiens*, *Nrxn1* in *Mus musculus*, and *nrx-1* in *Drosophila melanogaster* and *Caenorhabditis elegans*) is shown centrally, while other genes (random non-black colors) are jittered vertically while maintaining their linear position in the genome. Each locus is shown in the context of a 10 million bp window, with one additional inset for *C. elegans* showing a 200 kb window surrounding *nrx-1*.

Additional file 2: Table S1. Tissues analyzed in each species of Fig. 1.

Additional file 3: Figure S2. Intron length versus ordinal position for genes with greater than 2-fold tissue enrichment.

Additional file 4: Figure S3. Non-smooth-quantile-normalized gene expression conditional on gene length. Each plot shows transcriptome data across multiple tissues from species presented in Fig. 1. Neuronal tissues or cells are shown in red and all other tissues and cell types are shown in grey, and transparent ribbons show 95% confidence intervals. (**A**) Genes were segregated into 100 bins according to gene length (kb), and points show average gene length of each bin. (**B**) Loess smoothing of gene expression conditional on gene length (kb). (**C**) Generalized additive model of gene expression conditional on gene length (kb).

Additional file 5: Figure S4. Non-smooth-quantile-normalized gene expression conditional on gene length quantiles. Each plot shows transcriptome data across multiple tissues from species presented in Fig. 1. Neuronal tissues or cells are shown in red and all other tissues and cell types are shown in grey, and transparent ribbons show 95% confidence intervals. Gene lengths are shown as quantiles (100 bins). (**A**) Genes were segregated into 100 bins according to gene length, and points show average gene length of each bin. (**B**) Loess smoothing of gene expression conditional on gene length quantiles. (**C**) Generalized additive model of gene expression conditional on gene length quantiles.

Additional file 6: Figure S5. Smooth-quantile-normalized gene expression conditional on gene length. Each plot shows transcriptome data across multiple tissues from species presented in Fig. 1. Neuronal tissues or cells are shown in red and all other tissues and cell types are shown in grey, and transparent ribbons show 95% confidence intervals. (**A**) Loess smoothing of gene expression conditional on gene length (kb). (**B**) Generalized additive model of gene expression conditional on gene length (kb).

Additional file 7: Figure S6. Smooth-quantile-normalized gene expression conditional on gene length quantiles. Each plot shows transcriptome data across multiple tissues from species presented in Fig. 1. Neuronal tissues or cells are shown in red and all other tissues and cell types are shown in grey, and transparent ribbons show 95% confidence intervals. Gene lengths are shown as quantiles (100 bins). (**A**) Genes were segregated into 100 bins according to gene length, and points show average gene length of each bin. (**B**) Loess smoothing of gene expression conditional on gene length quantiles. (**C**) Generalized additive model of gene expression conditional on gene length quantiles.

Additional file 8: Figure S7. Gene features across clades. Joy plots showing distribution of gene features (median gene length, median intron length, median exon length, median number of exons per gene, and number of genes).

Abbreviations

bp: Base pair; DNA: Deoxyribonucleic acid; GAM: Generalized additive model; Gb: Gigabase; kb: Kilobase; LOESS: Locally estimated scatterplot smoothing;

MYA: Million years ago; REML: Restricted maximum likelihood; SRA: Sequence read archive; TPM: Transcripts per million; UTR: Untranslated region

Acknowledgements
Helpful input was provided by A. Herbert, B. Kim, N. Jain, M. Shoura, K. Artiles, L. Wahba, D. Jeong, N. Hall, M. Kasowski, K. Feller, E. Kroppf, T. Montague, A. Schnell, N. Takahashi, C. Vissers, J. Willett, J. Yoshida, K. Khodakhah, H. Rhodes, G. Bejerano, D. Petrov, and H. Fraser.

Authors' contributions
MJM conceived of the study. MJM performed all investigation and analyses with assistance from AZF. The authors read and approved the final manuscript.

Funding
This study was supported by the following programs, grants, and fellowships: 2018 Grass Fellowship in Neuroscience (Grass Foundation), 2019 Whitman Fellowship at the Marine Biological Laboratory, and the Stanford Genomics Training Program (5T32HG000044–22; PI: M. Snyder) to MJM, and R01GM37706/R35GM130366 to AZF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
Transcriptome data for *Octopus bimaculoides* was obtained from SRA BioProject PRJNA285380 [19]. Transcriptome data for *D. melanogaster* was obtained from SRA (SRR070409, SRR070410, SRR070408, SRR070405, SRR070406, SRR070392, SRR070393, SRR070407, SRR070425, SRR100268) [18]. All other transcriptomes were obtained as TPM-normalized counts from the EMBL Expression Atlas: *H. sapiens* (E-MTAB-513) [57], *B. taurus* (E-MTAB-2596) [58], *M. domestica* (E-MTAB-3719) [59], *G. gallus* (E-MTAB-2797) [60], *A. carolinensis* (E-MTAB-3727) [57], *C. elegans* (E-MTAB-2812) [61], and *Z. mays* (E-GEOD-62778) [62]. The datasets generated and analyzed for the current study are available in the GitHub repository, https://github.com/mjmccoy/BMC_Genomics_2020

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 30 January 2020 Accepted: 28 April 2020
Published online: 14 May 2020

References

1. Lynch M, Conery JS. The origins of genome complexity. Science. 2003; 302(5649):1401–4.
2. Huff JT, Zilberman D, Roy SW. Mechanism for DNA transposons to generate introns on genomic scales. Nature. 2016;538(7626):533–6.
3. Rodriguez F, Arkhipova IR. Transposable elements and polyploid evolution in animals. Curr Opin Genet Dev. 2018;49:115–23.
4. Wendel JF, Lisch D, Hu G, Mason AS. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. Curr Opin Genet Dev. 2018;49:1–7.
5. Petrov DA. Evolution of genome size: new approaches to an old problem. Trends Genet. 2001;17(1):23–8.
6. Grishkevich V, Yanai I. Gene length and expression level shape genomic novelties. Genome Res. 2014;24(9):1497–503.
7. Sugino K, Clark E, Schulmann A, Shima Y, Wang L, Hunt DL, et al. Mapping the transcriptional diversity of genetically and anatomically defined cell populations in the mouse brain. eLife. 2019;8 [cited 2019 Sep 4]. Available from: https://elifesciences.org/articles/38619.
8. Gubb D. Intron-delay and the precision of expression of homoeotic gene products inDrosophila. Dev Genet. 1986;7(3):119–31.
9. Singh J, Padgett RA. Rates of in situ transcription and splicing in large human genes. Nat Struct Mol Biol. 2009;16(11):1128–33.
10. Ardehali MB, Lis JT. Tracking rates of transcription and splicing in vivo. Nat Struct Mol Biol. 2009;16(11):1123–4.
11. Shamir M, Bar-On Y, Phillips R, Milo R. SnapShot: Timescales in Cell Biology. Cell. 2016;164(6):1302–1302.e1.
12. Artieri CG, Fraser HB. Transcript length mediates developmental timing of gene expression across Drosophila. Mol Biol Evol. 2014;31(11):2879–89.
13. Graf A, Krebs S, Zakhartchenko V, Schwalb B, Blum H, Wolf E. Fine mapping of genome activation in bovine embryos by RNA sequencing. Proc Natl Acad Sci. 2014;111(11):4139–44.
14. King IF, Yandava CN, Mabb AM, Hsiao JS, Huang H-S, Pearson BL, et al. Topoisomerases facilitate transcription of long genes linked to autism. Nature. 2013;501(7465):58–62.
15. Gabel HW, Kinde B, Stroud H, Gilbert CS, Harmin DA, Kastan NR, et al. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature. 2015;522(7554):89–93.
16. Zylka MJ, Simon JM, Philpot BD. Gene Length Matters in Neurons. Neuron. 2015;86(2):353–5.
17. McCoy MJ, Paul AJ, Victor MB, Richner M, Gabel HW, Gong H, et al. LONGO: an R package for interactive gene length dependent analysis for neuronal identity. Bioinformatics. 2018;34(13):i422–8.
18. Smibert P, Miura P, Westholm JO, Shenker S, May G, Duff MO, et al. Global patterns of tissue-specific alternative Polyadenylation in Drosophila. Cell Rep. 2012;1(3):277–89.
19. Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, et al. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature. 2015;524(7564):220–4.
20. Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, et al. Expression atlas update—an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res. 2016;44(D1): D746–52.
21. Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, et al. Expression atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Res. 2018;46(D1):D246–51.
22. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, Timetrees, and divergence times. Mol Biol Evol. 2017;34(7):1812–9.
23. Häggström M. Female shadow [Internet]. [cited 2020 Apr 23]. Available from: https://commons.wikimedia.org/wiki/File:Female_shadow.png. Depiction by MJM, https://creativecommons.org/publicdomain/zero/1.0/legalcode.
24. Weller K, USDA. Cow female black white [Internet]. [cited 2020 Apr 23]. Available from: https://commons.wikimedia.org/wiki/File:Cow_female_black_white.jpg. Depiction by MJM, https://creativecommons.org/publicdomain/zero/1.0/legalcode.
25. Samollow P, Southwest Foundation for biomedical research, San Antonio. Monodelphis domestica93-300 [Internet]. [cited 2020 Apr 23]. Available from: https://commons.wikimedia.org/wiki/File:Monodelphis_domestica93-300.jpg. Depiction by MJM, https://creativecommons.org/publicdomain/zero/1.0/legalcode.
26. Gratwicke B. Gallus gallus Bandhavgarh [Internet]. [cited 2020 Apr 23]. Available from: https://commons.wikimedia.org/wiki/File:Gallus_gallus_Bandhavgarh.jpg. Depiction by MJM, https://creativecommons.org/licenses/by/2.0/legalcode.
27. Wikimedia User:Pcb21. Green anole [Internet]. [cited 2020 Apr 23]. Available from: https://commons.wikimedia.org/wiki/File:Green_anole.jpg. Depiction by MJM, https://creativecommons.org/licenses/by-sa/3.0/legalcode.
28. Jeremyse at en.wikipedia. CaliforniaTwoSpotOctopus1 [Internet]. [cited 2020 Apr 23]. Available from: https://commons.wikimedia.org/wiki/File:CaliforniaTwoSpotOctopus1.jpg. Depiction by MJM, https://creativecommons.org/publicdomain/zero/1.0/legalcode.
29. Karwath A aka A. Drosophila melanogaster - side (aka) [Internet]. [cited 2020 Apr 23]. Available from: https://commons.wikimedia.org/wiki/File:Drosophila_melanogaster_-_side_(aka).jpg. Depiction by MJM, https://creativecommons.org/licenses/by-sa/2.5/legalcode.
30. Goldstein B. CelegansGoldsteinLabUNC [Internet]. [cited 2020 Apr 23]. Available from: https://commons.wikimedia.org/wiki/File:CelegansGoldsteinLabUNC.jpg. Depiction by MJM, https://creativecommons.org/licenses/by-sa/3.0/legalcode.
31. Zell H. Zea mays 004 [Internet]. [cited 2020 Apr 23]. Available from: https://commons.wikimedia.org/wiki/File:Zea_mays_004.JPG. Depiction by MJM, https://creativecommons.org/licenses/by-sa/3.0/legalcode.
32. Young JZ. The anatomy of the nervous system of *Octopus vulgaris*. Oxford: Clarendon Press; 1971. p. 690.

33. Hanlon RT, Messenger JB. Cephalopod behaviour [Internet]. 2018 [cited 2019 Oct 15]. Available from: https://doi.org/10.1017/9780511843600.
34. Smith MW. Structure of vertebrate genes: a statistical analysis implicating selection. J Mol Evol. 1988;27(1):45–55.
35. Bradnam KR, Korf I. Longer First Introns Are a General Property of Eukaryotic Gene Structure. PLoS One. 2008;3(8):e3093 Christoffels A, editor.
36. Sironi M, Menozzi G, Comi GP, Bresolin N, Cagliani R, Pozzoli U. Fixation of conserved sequences shapes human intron size and influences transposon-insertion dynamics. Trends Genet. 2005;21(9):484–8.
37. Gaffney D, Keightley P. Genomic Selective Constraints in Murids. PLoS Genet. 2005;preprint(2006):e204.
38. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015;43(W1):W589–98.
39. Francis WR, Wörheide G. Similar ratios of introns to Intergenic sequence across animal genomes. Genome Biol Evol. 2017;9(6):1582–98.
40. Lynch M, Bobay L-M, Catania F, Gout J-F, Rho M. The Repatterning of eukaryotic genomes by random genetic drift. Annu Rev Genomics Hum Genet. 2011;12(1):347–66.
41. Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, et al. The Trichoplax genome and the nature of placozoans. Nature. 2008; 454(7207):955–60.
42. Irimia M, Rukov JL, Penny D, Garcia-Fernandez J, Vinther J, Roy SW. Widespread evolutionary conservation of alternatively spliced exons in Caenorhabditis. Mol Biol Evol. 2008;25(2):375–82.
43. Roy M, Kim N, Xing Y, Lee C. The effect of intron length on exon creation ratios during the evolution of mammalian genomes. RNA. 2008;14(11):2261–73.
44. Sorek R. The birth of new exons: mechanisms and evolutionary consequences. RNA. 2007;13(10):1603–8.
45. Bush SJ, Chen L, Tovar-Corona JM, Urrutia AO. Alternative splicing and the evolution of phenotypic novelty. Philos Trans R Soc B Biol Sci. 2017; 372(1713):20150474.
46. Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. Mol Biol Evol. 2014;31(6): 1402–13.
47. Kapusta A, Suh A, Feschotte C. Dynamics of genome size evolution in birds and mammals. Proc Natl Acad Sci U S A. 2017;114(8):E1460–9.
48. Wright NA, Gregory TR, Witt CC. Metabolic 'engines' of flight drive genome size reduction in birds. Proc R Soc B. 2014;281(1779):20132780.
49. Howe KL, Contreras-Moreira B, De Silva N, Maslen G, Akanni W, Allen J, et al. Ensembl genomes 2020—enabling non-vertebrate genomic research. Nucleic Acids Res. 2020;48(D1):D689–95.
50. Meyer A, Van de Peer Y. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays. 2005;27(9):937–45.
51. Katju V. On the formation of novel genes by duplication in the Caenorhabditis elegans genome. Mol Biol Evol. 2006;23(5):1056–67.
52. Smith JJ, Antonacci F, Eichler EE, Amemiya CT. Programmed loss of millions of base pairs from a vertebrate genome. Proc Natl Acad Sci. 2009;106(27): 11212–7.
53. Kristan WB. Early evolution of neurons. Curr Biol. 2016;26(20):R949–54.
54. Niven JE, Chittka L. Evolving understanding of nervous system evolution. Curr Biol. 2016;26(20):R937–41.
55. Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, et al. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics. 2011;12(1):357.
56. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 2019;47(8):e47.
57. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012;338(6114):1587–93.
58. Liao X, Bao H, Meng Y, Plastow G, Moore S, Stothard P. Sequence, structural and expression divergence of duplicate genes in the bovine genome. PLoS One. 2014;9(7):e102868.
59. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;478(7369):343–8.
60. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science. 2012;338(6114):1593–9.
61. Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH. Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. Genome Res. 2009;19(4):657–66.
62. Zhan J, Thakare D, Ma C, Lloyd A, Nixon NM, Arakaki AM, et al. RNA sequencing of laser-capture microdissected compartments of the maize kernel identifies regulatory modules associated with endosperm cell differentiation. Plant Cell. 2015;27(3):513–31.
63. Wickham H. ggplot2: elegant graphics for data analysis. 2nd ed. Cham: Springer; 2016. p. 260. Use R!.
64. Hicks SC, Okrah K, Paulson JN, Quackenbush J, Irizarry RA, Bravo HC. Smooth quantile normalization. Biostatistics. 2018;19(2):185–98.
65. Paradis E, Claude J, Strimmer K. APE: analyses of Phylogenetics and evolution in R language. Bioinformatics. 2004;20(2):289–90.
66. Bolker B, Butler M, Cowan P, de Vienne D, Eddelbuettel D, Holder M, et al. phylobase: an R package that provides a base S4 class for comparative methods, incorporating one or more trees and trait data [Internet]. 2020. Available from: https://github.com/fmichonneau/phylobase.
67. Brewer C, Harrower M, Sheesley B, Woodruff A, Heyman D. ColorBrewer 2.0 Color advice for cartography. 2020. Available from: https://colorbrewer2.org/#.
68. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44(W1):W242–5.
69. Zipursky SL, Sanes JR. Chemoaffinity Revisited: Dscams, Protocadherins, and Neural Circuit Assembly. Cell. 2010;143(3):343–353.
70. Sahakyan AB, Balasubramanian S. Long genes and genes with multiple splice variants are enriched in pathways linked to cancer and other multigenic diseases. BMC Genomics. 2016;17(1).
71. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, et al. Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. Cell. 2012;151 (7):1431–1442.
72. Ptak SE, Petrov DA. How intron splicing affects the deletion and insertion profile in Drosophila melanogaster. Genetics. 2002 Nov;162(3):1233–44.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC