Intron size minimisation in teleosts

Lars Martin Jakt (lmjakt@gmail.com)
Nord Universitet - Bodo Campus https://orcid.org/0000-0002-3787-8138
Arseny Dubin
Nord Universitet - Bodo Campus https://orcid.org/0000-0001-8360-5475
Steinar Daae Johansen
Nord Universitet - Bodo Campus

Research article

Keywords: genome size, intron length, teleost, vertebrate evolution

Posted Date: March 31st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-362877/v1

RESEARCH

Intron size minimisation in teleosts

Lars Martin Jakt[1*], Arseny Dubin[1,2] and Steinar Daae Johansen[1]

[*]Correspondence: lars.m.jakt@nord.no
[1]Faculty for bioscience and aquaculture, Nord University, Universitetsalléen 11, 8026 Bodoe, Norway
Full list of author information is available at the end of the article

Abstract

Background:

Spliceosomal introns are parts of primary transcripts that are removed by RNA splicing. Although introns apparently do not contribute to the function of the mature transcript, in vertebrates they comprise the majority of the transcribed region increasing the metabolic cost of transcription. The persistence of long introns across evolutionary time suggests functional roles that can offset this metabolic cost. The teleosts comprise one of the largest vertebrate clades. They have unusually compact and variable genome sizes and provide a suitable system for analysing intron evolution.

Results:

We have analysed intron lengths in 172 vertebrate genomes and show that teleost intron lengths are relatively short, highly variable and bimodally distributed. Introns that were long in teleosts were also found to be long in mammals and were more likely to be found in regulatory genes and to contain conserved sequences. Our results argue that intron length has decreased in parallel in a non-random manner throughout teleost evolution and represent a deviation from the ancestral state.

Conclusion:

Our observations indicate an accelerated rate of intron size evolution in the teleosts and that teleost introns can be divided into two classes by their length. Teleost intron sizes have evolved primarily as a side-effect of genome size evolution and small genomes are dominated by short introns (<256 bp). However, a non-random subset of introns has resisted this process across the teleosts and these are more likely have functional roles in all vertebrate clades.

Keywords: genome size; intron length; teleost; vertebrate evolution

Introduction

The presence of spliceosomal introns within both coding and non-coding transcripts is a fundamental property of eukaryotes that separates them from eubacteria and archae [1, 2]. It is well-known that spliceosomal introns can be alternatively spliced increasing the functional complexity of the transcriptome [3], but the presence of introns by it self also has important roles in translation efficiency and non-sense mediated decay [4]. Introns can also encode regulatory RNAs such as microRNAs, snoRNAs or lncRNAs [5, 6]. A very interesting example is the processing of functional snoRNAs from introns in pseudogenes [7, 8]. Nevertheless, it is unclear as to what proportion of introns have functional roles.

Conservation of intron positions within orthologous transcripts has been observed across all eukaryotic kingdoms [9, 10]. Interestingly the extent of this conservation is not well correlated with phylogenetic distance with more conservation found between human and *Arabidopsis* than between human and *Drosophila* [9]. Remarkably, around 80% of intron positions in orthologous transcripts are conserved between human and the sea anemone *Nematostella vectensis* [11]. In contrast, 76% of introns in the chordate *Oikopleura* are unique with only 17% found at ancestral positions [12]. Hence, although introns (and their positions) are generally well conserved, it appears that intron loss and gain has been accelerated in specific clades.

The presence of splice sites in a transcript has a range of beneficial effects leading to more effective and better regulated translation (eg. intron mediated enhancement [13] and nonsense-mediated decay [4]), however, the effect of intron length is unclear. In *Drosophila* species it seems that long and very short introns are selected against and thus likely to be detrimental to evolutionary success [14]. Similarly, housekeeping genes and other highly expressed genes tend to have shorter introns, perhaps simply reflecting the metabolic cost of excessive transcription [15].

Nevertheless, long introns are maintained in many species, and intron length is well correlated with genome size (Fig. S1), suggesting that long introns in general may arise as a side-effect rather than through the accumulation of function. However, evolutionary conservation of intron size across diverse clades would argue for function since it is clear that introns can both grow and shrink in size. This is supported by the tendency for developmental regulators to contain long introns [12] and the observation that mammalian introns containing conserved sequences are longer [16] than those that do not.

In mammals, 3 to 6% of transcribed sequence is exonic, while the vast majority is composed of introns. Consistent with their smaller genome sizes, teleosts have much shorter introns with the transcribed regions being around 10-20% exonic (Fig. S1). During the annotation of the genome of *Lophius piscatorius*[17] we noticed that the distribution of log transformed intron sizes was clearly bimodal, with a very large number of short introns (shorter than 256 bp) forming a sharp peak separated from a broad peak of larger introns indicating that introns in *Lophius piscatorius* can be divided into two separate classes (data not shown).

In order to determine whether a bimodal intron size distribution is a general and or specific property of teleosts we extended our analysis to include the majority of vertebrate genomes present in the Ensembl database [18]. Our observations show that a) teleost introns do in general have bimodal size distributions, b) teleost intron length predicts intron length in other vertebrates, c) long introns are associated with specific biological functions and d) are more likely to contain sequence conserved across the vertebrate clade. We also show that intron size variation in teleosts is likely to have arisen as a result of the parallel evolution towards small introns throughout the clade, and that a similar process of intron size reduction is likely to have occurred specifically in the Aves, but not other Sauria clades. Both of these processes are likely to have resulted from an evolutionary pressure for smaller genomes or as a side-effect of directional neutral evolution.

Results

Teleost introns can be divided into long and short introns

Bimodal distributions across the vertebrates

We made use of the Ensembl database to obtain gene coordinates from 172 different vertebrate species and calculated the distributions of intron sizes for each species (Figs. 1 and S2). Teleost distributions were very variable, but the majority of species had bimodal distributions with a separating antimode (trough) at approximately 256 bp (2^8). In smaller genomes the short intron peak (< 256bp) was almost completely dominant obscuring the much smaller peak of long introns (Figs 1A, S2, additional file 1).

We also observed bimodality in the distributions of genomes belonging to both the Mammalia and Sauria clades with a minor peak (at 105 bp) of introns shorter than 150 bp (Figs. 1 B, C, S2 and additional file 1) as previously reported [19]. In addition the intron size distributions in the Sauria could be clearly divided into two groups corresponding to Aves and non-Aves species. Aves genomes contained shorter introns consistent with their generally smaller genome sizes (Fig. 1C).

The bimodal distribution observed in the teleosts was not seen in other aquatic vertebrates. The most closely related non-teleost analysed here, *Lepisosteus oculatus*, had a distribution more similar to mammals and birds then to teleosts. We also observed a very clear bimodal distribution in the jawless vertebrate *Eptatretus burgeri* (hagfish), but this distribution was distinctly different from that observed in teleosts (additional file 1). We thus infer that the bimodality observed across the teleosts is teleost specific.

In general, the vast majority of intron sizes reported were larger than 75 base-pairs (bp) and it is likely that this represents a lower bound for the size of a vertebrate intron. The proportion of very small (<32 bp) introns varied across the vertebrates in a manner unrelated to phylogeny. Genomes with higher proportions of such introns were usually from highly fragmented assemblies with short scaffold lengths (Fig. S3). Hence, intron sizes reported to be smaller than this lower-bound are more likely to represent annotation errors than real introns.

The proportion of long introns reflects genome size

Since the short intron peak seemed to dominate the small teleost genomes, we plotted the proportion of introns shorter than 256 bp against genome size to determine whether this represents a general property (Fig. 1 C). Indeed a strong correlation between genome size and the fraction of short (< 256bp) introns was readily observed across the teleosts. This suggests that the relationship observed between median intron and genome size [20] does not simply reflect differences in the size of all introns, but may also arise from a binary change in intron size status.

First introns are more likely to be long in teleosts

The first intron of a transcript is generally both longer and more likely to contain regulatory elements [21, 22, 23] and we asked whether first introns were more likely to belong to the longer class. Indeed, first introns were markedly more likely to be longer than 256 bp across all teleosts (Fig. 1D and S2). Although first introns were in general longer across the vertebrates we did not observe a major partition around 256 bp for non-teleost species and the differences in distributions were greater for teleost species (Fig. S4). Interestingly, the fractions of long first introns in teleosts were comparable to the fractions of long introns in Aves species with similar genome sizes (Fig. 1E).

Evolution of vertebrate intron sizes

An intron orthology

To investigate how intron size has evolved across the vertebrates we constructed an intron orthology by pairwise alignments of orthologous genes. To simplify the orthology we concentrated on gene families that mostly have only a single orthologue in each species. Since we are primarily interested in intron evolution in teleosts we made use of an orthology based on alignments to *Danio rerio* transcripts as these are derived from the most complete teleost genome annotation. This intron orthology contains a total of 63,068 introns from 5,752 genes. Although we did not find orthologues in all species for all *D. rerio* introns, the majority of introns were represented by at least 156 species and 90% (154/172) of species had orthologues to at least 47,609 *D. rerio* introns (Fig. S5).

Long teleost introns predict long mammalian introns

We first asked to what extent intron size in teleosts predicts intron size in other clades by comparing the median lengths of orthologue sets in mammals and teleosts (Fig. 2 A-C). The median (across species) intron size in teleosts and mammals were highly correlated for introns whose teleost median lengths were above 1000 (2^{10}) bp (Fig. 2 A-C). This represents 20% of the full intron set (Fig2. D); 73% of the mammalian orthologues of these introns had median intron lengths longer than the 50th percentile (i.e. median of median lengths) and 20% were longer than the 95th percentile (Fig. 2E). In general, the longer the teleost median length, the stronger the prediction (observed / expected) of long intron size in mammals (Fig. 2F).

These observations suggest that introns which have remained long across the majority of teleosts do not represent a random sub-set, and that these contain functional sequences which are conserved in mammalian species.

Intron size mutual information

To determine to what extent intron size correlates between species we calculated the mutual information in intron size between pairs of species. For this analysis we excluded genomes where the fraction of predicted introns shorter than 32 bp was larger than 2.5% (Fig. S3) since they obscured otherwise clear general trends. We observed clearly significant levels of mutual information for all pairs of species indicating some conservation of intron sizes across all vertebrates (significance judged by Monte-Carlo sampling of 10,000 random permutations). The amount of mutual information between species was clearly a function of taxonomic grouping with high mutual information observed within both the mammalian and Sauria clades (Fig. 3A).

Interestingly, the extent of mutual information between teleost species was much lower than that observed for mammal and sauria pairs. This is suggestive of an increased rate of intron size evolution in teleosts. To ensure that this was not merely a function of the evolutionary distances of the represented teleost species we also calculated pairwise Kimura-two factor distances for all species based on pairwise alignments of 472 (gene) orthologues. The mutual information for a given Kimura distance was lowest for pairs of teleost species and highest for pairs of mammalian species (Fig. 3B). This argues for an increased rate of intron size evolution in teleosts.

Intron size reduction in Teleostei and Aves

Teleost genomes and intron sizes are in general smaller than those of other vertebrates, suggesting a decrease in genome size that is specific to teleosts (i.e. a synapomorphy). However, it is also possible that the common vertebrate ancestor possessed a small genome with small introns and that this expanded in size in non-teleost clades. To evaluate the likelihood of these alternatives we used Sankoff maximum parsimony on a neighbour-joining tree (Fig. S6) created from pairwise Kimura two factor distances (as used in Fig. 3) to infer the sizes of introns in ancestors. This inference (Fig. 4) is consistent with intron sizes in the common vertebrate ancestor having been more similar to extant mammalian and Sauria species, and that intron sizes in the teleost clade decreased drastically both before and after divergence from *L. oculatus*.

Three teleost species have unusually large genomes and corresponding introns; *Danio rerio*, *Astyanax mexicanus* and *Pygocentrus nattereri*. All three are derived from a branch diverging early in teleost evolution and *A. mexicanus* and *P. nattereri* have an exlusive common ancestor. Nevertheless, all these three species appear to have experienced independent increases in intron sizes since their common ancestors suggesting an ongoing process of genome size expansion presumably through the accumulation of transposable elements (TEs) [24].

Perhaps most striking is the apparent intron size reduction in avian species; avians are members of the Sauria clade, but compared to other Sauria species they have much smaller intron sizes. The majority of this intron size reduction appears to have occurred in the evolution of their common ancestor with smaller changes after divergence which is consistent with an inference of small genomes in the extinct pterosaur lineage [25] and in extant alligators [26]. Other Sauria species appear to have experienced intron size expansion rather than contraction with *Sphenodon punctatus* having one of the largest genomes in the data set.

We also asked to what extent intron size minimisation is an ongoing process by inspecting the difference in intron size between extant species and their most recent inferred ancestor. Most species appear to have experienced recent increases in intron size (Fig. S7). However, the teleost species with the smallest genomes had shorter introns than their most recent ancestors, suggesting that introns in these species are likely to become shorter in the descendants of these species.

Importantly, there is relatively little conservation in intron size within the teleosts suggesting an ongoing process of intron size diversification. Although the majority of decrease in intron size seems to have occurred prior to the most recent common teleost ancestor, most teleost species have introns considerably smaller than this. This suggests a subsequent parallel decrease in intron size across the teleost clade.

Selective retention of long introns

If loss of teleost intron length during teleost evolution is a random process then the set of long introns retained (as long) by descendants of a common ancestor should be random subsets of the ancestral ones. In contrast, if a subset of long introns are selectively retained, then we should see an enrichment of common long introns in pairs of descendants. We tested this by considering how intron size has evolved from the extinct ancestor represented by node 288 (Fig. 4 and S6) in our neighbour-joining tree. The extant descendants of node 288 are all members of the Percomorphaceae clade. Three descendants are Tetraodontiformes (*T. rubripes, T. nigroviridis, Mola mola*), one of which (*M. mola*) has been misplaced (compared to the NCBI taxonomy database) within our tree outside of the Tetraodontiformes branch. Node 288 is the most recent common ancestor of the Tetraodontiformes and *Betta splendens* which has the smallest genome and shortest introns outside of *T. rubripes* and *T. nigroviridis.* The inferred intron sizes of node 288 are much reduced from the teleost common ancestor (Fig. 4), and apart from the Tetraodontiformes and a small number of species (eg. *B. splendens* and *Parambassis ranga*) most of its descendants appear to have gained rather than lost intron length.

We evaluated the number of long introns (> 256 bp, inferred by Sankoff maximum parsimony, Fig. 4) in the ancestral node 288 that had been commonly retained as long within *T. rubripes* and each non-Tetraodontiformes descendant (including *M. mola* due to its placement in our tree). The proportion of long introns retained in the descendants varied from less than 70% in the species with the smallest genomes (*T. rubripes*, *Betta splendens*) to more than 90% in species with larger genomes (Fig. 5A). We observed larger than expected overlaps between long introns retained in both *T. rubripes* and the other descendants (Fig 5 B-F). Although the excess of common retained introns was small (up to 15% for *B. splendens*) it was negatively correlated with the proportion of long introns retained in either species (Fig. 5B) indicating that the smaller the number of introns retained as long, the larger the excess of common long introns.

The observed enrichments were highly significant (down to $p < 1e - 300$, hypergeometric distribution) and were correlated with genome length (Fig. 5C, E). Importantly, there was no correlation between either the enrichment (observed / expected ratios) or the associated p-values and phylogenetic distance indicating that the commonality was not due to inheritance (Fig. 5D, F).

We repeated these analyses for node 266 by comparing selected descendants of node 267 (*D. rerio*, *Electrophorus electricus* and *Denticeps clupeoides*) against all descendants of node 276 (these descend from 266 but not from node 267, Fig 4, S6). Again we saw the strongest effects for pairs of small genomes and no correlation with phylogenetic distance. These effects were also markedly stronger (Fig. S8-10) in line with the larger amount of intron sequence lost since the more ancient ancestor represented by node 266 (Fig. 4, S06).

Functional association

Genes with long introns are enriched for specific biological functions

Retention of intron sequence/length across the vertebrates suggests the presence of some regulatory function. If this is the case, then genes which do not contain any long introns should be less likely to be highly regulated and this ought to be reflected in their biological and molecular function. To test this we selected nested sets of genes containing at least one intron that is long ($> T_i$ where T was a series of thresholds) across the teleosts and performed gene ontology (GO) enrichment analysis using human annotation. Sets of genes containing long introns were strongly enriched for specific biological process (BP), molecular function (MF) and cellular compartment (CC) terms. The enriched BP and MF terms were related to system development, cell signalling and transcriptional regulation (data not shown). This is consistent with genes containing long introns being more regulated than other genes and previous observations of short introns in highly expressed and house-keeping genes [15, 27].

Since this gene selection could have been biased towards genes containing large numbers of introns we also selected genes by randomly selecting introns from the intron orthology and performed the same analyses. Although such genes were strongly enriched for specific GO terms as would be expected (due to preferential selection of genes with many introns), these terms did not overlap with those found by selecting genes with long introns.

To further investigate the functional associations we ordered genes by their longest minimal teleost intron length, determined the number of genes belonging to enriched GO terms for all nested sets (from short to long) and calculated depletion statistics (Fig. 6, Fig. S11-13, additional file 2). For most terms the strongest depletion probability occurred between 2^8-2^9 (Fig. S11-13, additional file 2). This position roughly corresponds with the observed antimode observed in teleosts lengths and is consistent with distinct functional roles for short and long introns in teleosts.

Long introns are enriched for conserved sequences

If conservation of intron length is related to the retention of regulatory sequences then this should be reflected in an increase in sequence conservation. To test this we aligned sets of intron sequences from *D. rerio* chosen by their lengths across the teleosts to their orthologues in other species. Because regulatory function may be present in several regions across the intron we performed local alignments and extracted all non-overlapping alignments to *D. rerio* sequences. High scores were more easily found from alignments of introns that were consistently long across the teleosts (Fig. S14, additional file 3) than for variable length introns (Fig. S15, additional file 4). These included alignments across the full length of *D. rerio* introns, as well as single and multiple windows of conservation. Although high scoring alignments were most common to other teleosts, a number of *D. rerio* regions could be aligned across the vertebrate clade.

To compensate for the fact that alignments of long sequences to each other are intrinsically more likely to yield high scoring alignments, we made use of alignments to non-orthologous introns to create models representing operational null hypotheses (Fig. S16) relating the frequency of maximally scoring alignments to the total search space (see Fig. S18-S27 and additional files 5-14 for examples of alignments). Sequences from introns long in all teleosts were much more likely to diverge from these models (Fig. 7 and Fig. S17). This effect increased with minimum teleost intron length with significant local alignments to *D. rerio* sequences against teleost and mammalian introns found in up to 60% and 20% of introns longer than 2000 bp.

Discussion

Intron sizes are exponentially distributed

It has previously been reported that introns specifically in *Danio rerio*, but not in other teleosts have a bimodal distribution [24]. We show here that bimodality in intron length is a general property of teleost introns. Our analysis differs from Moss *et. al* [24] in that we log transformed intron lengths. We believe that this is correct because the probability of an intron changing in size is likely to be a function of its length [16] and this means that the expected distributions are exponential. Furthermore, the fact that log-transformation gives rise to distributions that approximate normality argues for the suitability of log-transformation.

Moss *et al.* interpreted the appearance of a second peak of larger introns as a result of expansion of intron lengths due to transposable element (TE) expansion specifically in the *Danio* lineage. Our results, based on a large number of recently available vertebrate genomes, suggest instead that bimodality is the norm across the teleosts and may have arisen as a result of loss of ancestral sequences rather than expansion.

Two types of teleost introns

Clear bimodal distributions are rare in genomic data and indicate the division of the underlying entities into distinct classes. Our observations argue that in teleosts introns can be divided into two classes; short (< 256 bp) and long (> 256 bp) introns. The simplest explanation for our observations is that introns shorter than the antimode contain only sequences necessary for efficient spliceosome association [28] whereas longer introns, especially those that are long across the teleosts, also contain regulatory or other functional elements. However, the presence of a separate peak of shorter (peak at $\sim$ 100 bp) introns in the mammalian and Sauria (Fig. S02 and [19]) clades argues for a shorter minimal intron size. This shorter peak can also be observed across the teleosts (Fig. S02) though it is obscured by the overall teleost distribution. A sharp peak of minimal intron size has been observed in both plants and several metazoan clades and it has been argued that the position of the distribution peak represents a clade specific ($\sim$ 50 bp in *Caenorhabditis elegans*) optimal intron size [19]. Hence, the teleost specific major short intron peak is unlikely to merely contain a set of minimal introns and the cause for the position of the antimode ($\sim$256 bp) remains unclear.

An ongoing process of intron size diversification

Our observations suggest that vertebrate ancestral intron sizes were in general longer than in the teleosts indicating an active loss of intron sequence across the teleosts. The alternative hypothesis, that ancestral introns (and presumably genomes) were small would indicate a reduced speed of intron growth in teleosts compared to other vertebrates. This seems less likely given that intron sizes are more variable within the teleosts than other vertebrate clades. An active process of intron and associated genome size reduction within the teleosts also seems more likely since the teleosts descend from an ancestor that underwent an additional round of genome duplication [29] and as such would be expected to initially have a larger than typical genome.

A similar decrease in genome and intron sizes has previously been reported in birds (Aves) and in flying vertebrates generally [30, 25, 31]. In birds it has been suggested that both genome and intron size reduction is related to the increased metabolic demands of flight; however, alligators also appear to have smaller introns and it is unclear as to whether genome shrinkage occurred prior to powered flight

[26]. Similar explanations have been suggested for genome size reduction in teleosts, though the lack of a correlation between genome size and metabolic rate makes this less likely [32].

Introns with function

Why some introns retain length across the vertebrates even in species where there is an apparent evolutionary pressure for genome size minimisation is unclear, but argues for function, either of the length itself, or encoded in sequence. Introns can contain regulatory regions [21] and indeed, in mammals introns containing higher densities of conserved sequences are longer than those lacking such sequences [16]. Such conserved sequences may include not only cis-regulatory elements but can also encode functional molecules such as miRNAs, snoRNAs or lncRNAS, and even exons of overlapping genes. Intron length is more strongly conserved in genes associated with embryonic development [33] and the length of introns from genes that are co-expressed or that belong to the same protein complexes appear to have co-evolved [34] suggesting that the intron length per-se can have functional implications. This role might be partly explained by the fact that intron length can affect the delay between transcription and protein expression which can change the timing of regulatory networks [35, 23].

The observation that introns that are long across the teleost clade are also likely to be long in other vertebrates supports the notion that these introns have specifically retained length for functional reasons. This is consistent with vertebrate introns being divided into two classes; those containing longer functional regions and those that do not. Although both classes would be subject to minimisation by evolutionary processes, the minimal size of the former would be larger than the latter. In teleosts, which have evolved small genomes, loss of sequence from the latter class of introns would have been more prevalent than from the former class. This scenario is supported by the observation of an increased rate of common retention of long introns in species with smaller genomes (Fig. 6, S8-10) and suggests that there exists a subset of introns in the teleosts whose size must be maintained. This is similar to the differing distributions of the lengths of introns with or without multi-species conserved sequences (MCS) observed in mammals [16].

It is also possible that sequence has been lost from a specific set of introns rather than there being a selective retention of sequence in introns with function. This could conceivably occur as a result of increased recombination rates at such introns, or because of an increased evolutionary selection against length at such loci (eg. as a result of high gene expression [16]). Untangling these alternatives is difficult because selective retention and loss both result in similar outcomes. However, a selective loss of intron sequence would be more likely to be associated with genes or genome regions rather than specific introns and this could potentially be used to argue for one scenario.

Intron and genome size

The small sizes of teleost genomes and introns are intrinsically linked, with intron size reflecting genome size. Since a reduction in intron size cannot markedly reduce genome size (Fig. S1) it is likely that the short introns are a side effect of the evolution of small genomes in the teleost lineage.

Why teleosts specifically should have small genome sizes remains unclear, but may be related to the high fecundity and associated developmental strategies of the teleosts. Most teleosts produce large numbers of small eggs giving rise to extreme number of offspring, and within the teleosts, there appears to be a relationship between egg size and genome size [32]. Both a small propagule size (the size of the stage that leaves its parents) and a high fecundity are strongly associated with a large effective population size (N_e) in metazoans [36].

The distribution of intron sizes in teleosts with smaller genomes (*T. rubripes*, *T. nigroviridis*, *B. splendens*) is reminiscent of those in *Drosophila* species [14, 37, 38], where about half of introns are between 45 and 110 bases long. In *Drosophila*, large introns are found preferentially at sites of low recombination leading to the idea that an amelioration of Hill-Robertson interference (the competition between two linked loci for fixation in the absence of recombination) may be a driver for increased intron size and that species with lower N_e will have larger introns [37]. In addition, a large N_e supports more effective purifying selection [39, 40] that might be able to reduce the rate of fixation of mutations leading to larger introns even where these have only very small negative effects. These ideas are consistent with the decreased N_e and increased genome sizes of groundwater living asellid isopods [40]. Hence, the small size of teleost genomes (and introns) may simply be a consequence of the typical teleost breeding strategy of having large numbers of small offspring. This is supported by the observation that marine fish tend to have larger N_e, but smaller genomes than freshwater species [41].

It has also been argued that the process of recombination itself leads to an increased deletion bias and underlies genome contraction in Avian species [42]. Recombination rates may thus affect intron size both during the generation and selection of genetic variance. Our observations show that the lengths of long introns are generally conserved across the vertebrates; if these are to be explained by recombination rate variance, then recombination rates at these introns should be low throughout the vertebrates. This is testable given adequate recombination rate maps.

Conclusion

Teleosts have both genome and intron sizes that are unusually small for vertebrate species. We show here that teleost intron sizes are likely to have decreased from those in the last common vertebrate ancestor. This decrease appears to have started in the ancestor of the teleosts but to have continued in the independent teleost lineages, suggesting either selection by common evolutionary pressure or molecular mechanisms specific to the teleosts.

A subset of introns have escaped from this process and maintained their lengths. Members of this set are also more likely to be long in both mammalian and avian species suggesting a function dependent on the sequence content or the length itself that is conserved across the vertebrates.

It is unclear as to why introns have shrunk in the teleost lineages, but it is likely to be a side-effect of a general evolution towards smaller genomes. This may have been driven by neutral evolution or by selective forces. In either case, the decrease in genome size may have facilitated teleost specific developmental strategies.

Methods

All analyses, unless specifically mentioned were performed using a mixture of SQL, Perl, R and C. All data visualisations were created using core R functions. Detailed descriptions of methods, species and genes used are provided in the supplement.

Intron sizes and orthology

Exon and intron sizes were inferred from genomic coordinates obtained from 172 (54 Teleost, 80 Mammalia, 31 Sauria and 7 other vertebrate species) locally installed Ensembl [18] core 98 databases. The intron orthology was constructed from a set of 6114 protein families which had a single orthologue in at least 130 of the 172 species according to the protein family classification of the Ensembl compara 98 database. Orthologous introns were identified by aligning transcript sequences modified to have 'sticky' meta-characters representing introns at exon-exon boundaries to their *D. rerio* orthologues using a Needleman-Wunsch algorithm [43] implemented as an extension to R in C. Introns aligning to each other were considered orthologous.

Intron size conservation

Mutual information estimates were obtained using the `entropy` package `https://CRAN.R-project.org/package=entropy` [44]. Kimura two factor distances [45] were based on all against all (species) alignments of the members of 472 protein families that had members in at least 170 of the 172 species. These distances were used to construct a neighbour-joining tree using the `nj` function of the `ape` `https://cran.r-project.org/web/packages/ape/index.html` [46]. The ancestral lengths of introns of the internal nodes of this tree were inferred using an implementation of Sankoff maximum parsimony [47] with intron lengths discretized to integral $10 \times log_2$ values and transition costs being the difference in intron length state. The Sankoff maximum parsimony was implemented as an R plug-in in C and Kimura two factor distances calculated in R.

The probability of a given number or larger, of introns remaining longer than 256 bp in two descendant species was calculated using the hypergeometric distribution considering the sets of long introns in the descendants (d1, d2) as samples of the long introns in the most recent common ancestor (MRCA). P-values were calculated using $phyper(q-1, m, n, k, lower.tail = FALSE)$ in `R`, where q, m, n and k were defined by the sizes of the subsets of of introns long in the MRCA and d1 and d2. The parameters were defined as: q, long in both d1 and d2, m long in d1, n short in d1, k long in d2, such that $m+n$ denotes the number of long introns in the MRCA. The expected number of common long introns was calculated as $mk/(m+n)$.

Gene ontology

Initial GO statistics were obtained using human annotation from the org.Hs.eg.db [48] and GOstats [49] packages. Detailed analyses of enriched / depleted GO terms were based on the org.Hs.eg.db annotation using the phyper function to calculate enrichment and depletion probabilities of nested sets of genes ordered by their minimal teleost length [R_gene_ontology/gene_ontology.R].

Intron sequence conservation

We used a recursive Smith-Waterman [50] implementation to identify all local alignments between pairs of intron sequences. The implementation first identifies the optimal local alignment between the pair, and then recurses to find additional alignments that do not overlap within the seed species (*D. rerio* for all analyses shown). To allow comparisons between introns with different types of length distributions across the teleosts we selected five sets of introns, 'long': 1594 introns with a min teleost length of 1024 bp, 'med': 2000 introns with a min teleost length between 256 and 1024 bases, two sets of 2000 'short' introns with min teleost lengths between 90 and 256 bases, and a set of 'ctl' introns with a median intron length less than 256 bp but with a *D. rerio* length of more than 1024 bp. The 'ctl' and one of the 'short' sets were sampled randomly from the allowable subset, the 'long' set included all introns satisfying the criteria. All other sets used the 2000 introns with the smallest variance in teleost intron length. Alignments where the individual search space ($l_1 \times l_2$ where l_1 and l_2 are the lengths of the two sequences) was larger than 1e8 where excluded to reduce memory requirements.

To simplify the statistical assessment we focused on the top scoring alignment for each *D. rerio* intron. To compensate for the effect of the differences in search space on the observed top scores we performed alignments for the 'ctl', 'med' and 'long' intron sets but using a non-orthologous *D. rerio* sequence of similar length. The resulting alignment scores were used to obtain a linear model ($log(S) \propto log(l_{dr} \times \sum l_{\notin dr})$, where S is the maximum score observed and l_{dr} is the *D. rerio*, and $l_{\notin dr}$ the other intron lengths) representing the likelihood of observing a given score for a given alignment search space (Fig. S16). In order to remove the effect of simple repeats and other repetitive sequences from the effects we screened all aligned sequences (for both the orthologous and non-orthologous alignments) by BLASTn of *D. rerio* sequences against the *D. rerio* genome. Any aligned *D. rerio* sequence which had more than 2 sequences aligned to 50% of its length or no sequences at all (indicating simple repeats) were removed from further consideration. Not surprisingly, far more introns were removed from the non-orthologous alignments (Fig. S16, S17). The divergence from this model was calculated for both orthologous and non-orthologous sets and considered as a measure of effective sequence conservation.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
All data used in this study is publicly available from Ensembl http://www.ensembl.org version 98. Code written for the analyses carried out is described in the supplementary section and is available from https://github.com/lmjakt/ in the repositories intron_size_evo_1 and intron_size_evo_2.

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

Author's contributions
The work was initiated based on observations by AD. LMJ designed and performed the analyses. The manuscript was written by LMJ, AD and SDJ.

Funding
All work was funded internally by Nord University.

Acknowledgements

We are grateful to Joost Andre M Raeymaekers (Nord University) for a detailed and critical reading of the manuscript.

Author details

[1]Faculty for bioscience and aquaculture, Nord University, Universitetsalléen 11, 8026 Bodoe, Norway. [2]Currently at: Parental Investment and Immune Dynamics, GEOMAR Helmholtz Centre for Ocean Research, Düsternbrookerweg 20, D-24105 Kiel, Germany.

References

1. Rogozin IB, Carmel L, Csuros M, Koonin EV. Origin and evolution of spliceosomal introns. Biology Direct. 2012 Apr;7(1):11. Available from: https://doi.org/10.1186/1745-6150-7-11.
2. Irimia M, Roy SW. Origin of Spliceosomal Introns and Alternative Splicing. Cold Spring Harbor Perspectives in Biology. 2014 Jan;6(6):a016071. Number: 6. Available from: http://cshperspectives.cshlp.org/content/6/6/a016071.
3. Lee Y, Rio DC. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annual Review of Biochemistry. 2015;84:291–323.
4. Le Hir H, Saulière J, Wang Z. The exon junction complex as a node of post-transcriptional networks. Nature Reviews Molecular Cell Biology. 2016 Jan;17(1):41–54.
5. Sollner-Webb B. Novel intron-encoded small nucleolar RNAs. Cell. 1993 Nov;75(3):403–405. Available from: http://www.sciencedirect.com/science/article/pii/009286749390374Y.
6. Caffarelli E, Fatica A, Prislei S, De Gregorio E, Fragapane P, Bozzoni I. Processing of the intron-encoded U16 and U18 snoRNAs: the conserved C and D boxes control both the processing reaction and the stability of the mature snoRNA. The EMBO Journal. 1996 Mar;15(5):1121–1131. Publisher: John Wiley & Sons, Ltd. Available from: https://www.embopress.org/doi/abs/10.1002/j.1460-2075.1996.tb00450.x.
7. Tycowski KT, Shu MD, Steitz JA. A mammalian gene with introns instead of exons generating stable RNA products. Nature. 1996 Feb;379(6564):464–466. Number: 6564 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/379464a0.
8. Bortolin ML, Kiss T. Human U19 intron-encoded snoRNA is processed from a long primary transcript that possesses little potential for protein coding. RNA. 1998 Jan;4(4):445–454. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. Available from: http://rnajournal.cshlp.org/content/4/4/445.
9. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. Remarkable Interkingdom Conservation of Intron Positions and Massive, Lineage-Specific Intron Loss and Gain in Eukaryotic Evolution. Current Biology. 2003 Sep;13(17):1512–1517. Publisher: Elsevier. Available from: https://www.cell.com/current-biology/abstract/S0960-9822(03)00558-X.
10. Fedorov A, Merican AF, Gilbert W. Large-scale comparison of intron positions among animal, plant, and fungal genes. Proceedings of the National Academy of Sciences. 2002 Dec;99(25):16128–16133. Publisher: National Academy of Sciences Section: Biological Sciences. Available from: https://www.pnas.org/content/99/25/16128.
11. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, et al. Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. Science. 2007 Jul;317(5834):86–94. Publisher: American Association for the Advancement of Science Section: Research Article. Available from: https://science.sciencemag.org/content/317/5834/86.
12. Denoeud F, Henriet S, Mungpakdee S, Aury JM, Silva CD, Brinkmann H, et al. Plasticity of Animal Genome Architecture Unmasked by Rapid Evolution of a Pelagic Tunicate. Science. 2010 Dec;330(6009):1381–1385. Publisher: American Association for the Advancement of Science Section: Report. Available from: https://science.sciencemag.org/content/330/6009/1381.
13. Shaul O. How introns enhance gene expression. The International Journal of Biochemistry & Cell Biology. 2017 Oct;91:145–155. Available from: http://www.sciencedirect.com/science/article/pii/S1357272517301541.
14. Carvalho AB, Clark AG. Intron size and natural selection. Nature. 1999 Sep;401(6751):344–344. Number: 6751 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/43827.
15. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. Selection for short introns in highly expressed genes. Nature Genetics. 2002 Aug;31(4):415–418. Number: 4 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/ng940z.
16. Pozzoli U, Menozzi G, Comi GP, Cagliani R, Bresolin N, Sironi M. Intron size in mammals: complexity comes to terms with economy. Trends in Genetics. 2007 Jan;23(1):20–24. Available from: http://www.sciencedirect.com/science/article/pii/S0168952506003465.
17. Dubin A, Jørgensen TE, Moum TB, Johansen SD, Jakt LM. Complete loss of the MHC II pathway in an anglerfish, Lophius piscatorius. 6. 2019;Accepted: 2020-06-16T11:13:45Z Publisher: The Royal Society. Available from: https://nordopen.nord.no/nord-xmlui/handle/11250/2658241.
18. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic Acids Research. 2019 Jan;47(D1):D745–D751. Publisher: Oxford Academic. Available from: https://academic.oup.com/nar/article/47/D1/D745/5165265.
19. Yu J, Yang Z, Kibukawa M, Paddock M, Passey DA, Wong GKS. Minimal Introns Are Not "Junk". Genome Research. 2002 Jan;12(8):1185–1189. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. Available from: http://genome.cshlp.org/content/12/8/1185.
20. Hara Y, Yamaguchi K, Onimaru K, Kadota M, Koyanagi M, Keeley SD, et al. Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates. Nature Ecology & Evolution. 2018

Nov;2(11):1761–1771. Number: 11 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41559-018-0673-5.

21. Marais G, Nouvellet P, Keightley PD, Charlesworth B. Intron Size and Exon Evolution in Drosophila. Genetics. 2005 May;170(1):481–485. Publisher: Genetics Section: Note. Available from: https://www.genetics.org/content/170/1/481.
22. Chorev M, Carmel L. Computational identification of functional introns: high positional conservation of introns that harbor RNA genes. Nucleic Acids Research. 2013 Jun;41(11):5604–5613. Publisher: Oxford Academic. Available from: https://academic.oup.com/nar/article/41/11/5604/2411186.
23. Pai AA, Henriques T, McCue K, Burkholder A, Adelman K, Burge CB. The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene architecture. eLife. 2017 Dec;6:e32537. Publisher: eLife Sciences Publications, Ltd. Available from: https://doi.org/10.7554/eLife.32537.
24. Moss SP, Joyce DA, Humphries S, Tindall KJ, Lunt DH. Comparative Analysis of Teleost Genome Sequences Reveals an Ancient Intron Size Expansion in the Zebrafish Lineage. Genome Biology and Evolution. 2011 Sep;3:1187–1196. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3205604/.
25. Organ CL, Shedlock AM. Palaeogenomics of pterosaurs and the evolution of small genome size in flying vertebrates. Biology Letters. 2009 Feb;5(1):47–50. Publisher: Royal Society. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rsbl.2008.0491.
26. Waltari E, Edwards SV. Evolutionary dynamics of intron size, genome size, and physiological correlates in archosaurs. The American Naturalist. 2002 Nov;160(5):539–552.
27. Eisenberg E, Levanon EY. Human housekeeping genes are compact. Trends in Genetics. 2003 Jul;19(7):362–365. Available from: http://www.sciencedirect.com/science/article/pii/S0168952503001409.
28. Wieringa B, Hofer E, Weissmann C. A minimal intron length but no specific internal sequence is required for splicing the large rabbit β-globin intron. Cell. 1984 Jul;37(3):915–925. Available from: http://www.sciencedirect.com/science/article/pii/0092867484904264.
29. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature. 2004 Oct;431(7011):946–957.
30. Gregory TR. A Bird's-Eye View of the C-Value Enigma: Genome Size, Cell Size, and Metabolic Rate in the Class Aves. Evolution. 2002;56(1):121–130. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0014-3820.2002.tb00854.x. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0014-3820.2002.tb00854.x.
31. Zhang Q, Edwards SV. The Evolution of Intron Size in Amniotes: A Role for Powered Flight? Genome Biology and Evolution. 2012;4(10):1033–1043. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3490418/.
32. Hardie D, Hebert P. Genome-size evolution in fishes. Canadian Journal of Fisheries and Aquatic Sciences. 2004 Sep;61:1636–1646.
33. Seoighe C, Korir PK. Evidence for intron length conservation in a set of mammalian genes associated with embryonic development. BMC bioinformatics. 2011 Oct;12 Suppl 9:S16.
34. Keane PA, Seoighe C. Intron Length Coevolution across Mammalian Genomes. Molecular Biology and Evolution. 2016;33(10):2682–2691. Number: 10.
35. Swinburne IA, Miguez DG, Landgraf D, Silver PA. Intron length increases oscillatory periods of gene expression in animal cells. Genes & Development. 2008 Sep;22(17):2342–2346. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2532923/.
36. Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, et al. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature. 2014 Nov;515(7526):261–263. Number: 7526 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature13685.
37. Comeron JM, Kreitman M. The Correlation Between Intron Length and Recombination in Drosophila: Dynamic Equilibrium Between Mutational and Selective Forces. Genetics. 2000 Nov;156(3):1175–1190. Publisher: Genetics Section: Investigations. Available from: https://www.genetics.org/content/156/3/1175.
38. Mount SM, Burks C, Herts G, Stormo GD, White O, Fields C. Splicing signals in Drosophila: intron size, information content, and consensus sequences. Nucleic Acids Research. 1992 Aug;20(16):4255–4262. Publisher: Oxford Academic. Available from: https://academic.oup.com/nar/article/20/16/4255/1105205.
39. Lynch M, Conery JS. The Origins of Genome Complexity. Science. 2003 Nov;302(5649):1401–1404. Number: 5649. Available from: https://science.sciencemag.org/content/302/5649/1401.
40. Leféburе T, Morvan C, Malard F, François C, Konecny-Dupré L, Guéguen L, et al. Less effective selection leads to larger genomes. Genome Research. 2017 Jan;27(6):1016–1028. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. Available from: http://genome.cshlp.org/content/27/6/1016.
41. DeWoody JA, Avise JC. Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. Journal of Fish Biology. 2000;56(3):461–473. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1095-8649.2000.tb00748.x. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1095-8649.2000.tb00748.x.
42. Nam K, Ellegren H. Recombination Drives Vertebrate Genome Contraction. PLOS Genetics. 2012 May;8(5):e1002680. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002680.
43. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 1970 Mar;48(3):443–453. Number: 3. Available from: http://linkinghub.elsevier.com/retrieve/pii/0022283670900574.
44. Hausser J, Strimmer K. Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. Journal of Machine Learning Research. 2009;10(50):1469–1484. Available from: http://jmlr.org/papers/v10/hausser09a.html.

45. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. 1980 Jun;16(2):111–120. Available from: http://link.springer.com/10.1007/BF01731581.
46. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019 Feb;35(3):526–528. Publisher: Oxford Academic. Available from: https://academic.oup.com/bioinformatics/article/35/3/526/5055127.
47. Sankoff D. Minimal Mutation Trees of Sequences. SIAM Journal on Applied Mathematics. 1975;28(1):35–42. Publisher: Society for Industrial and Applied Mathematics. Available from: https://www.jstor.org/stable/2100459.
48. org.Hs.eg.db;. Available from: http://bioconductor.org/packages/org.Hs.eg.db/.
49. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007 Jan;23(2):257–258. Publisher: Oxford Academic. Available from: https://academic.oup.com/bioinformatics/article/23/2/257/204776.
50. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology. 1981 Mar;147(1):195–197. Number: 1. Available from: http://linkinghub.elsevier.com/retrieve/pii/0022283681900875.

Figures

Figure 1 Vertebrate intron size. A-C, distributions of $\log_2$intron size in teleosts (A), mammals (B) and Sauria (C). Colour of lines indicate log genome size from blue (small) to purple (large); x-axis, $\log_2$ intron size, y-axis density. Distributions for teleosts are much more varied than for mammals or Sauria, but are generally bimodal with an antimode at 2^8 bases (indicated by vertical line). The distributions of mammals and Sauria also suggest bimodality with a peak of shorter introns at 105 bp. Distributions for Sauria separate naturally into two groups of different genome size ranges; these correspond to Aves (birds) and non-Aves species, with Aves species (blues) having both smaller genomes and introns. D, $\log_2$ genome size (x-axis) plotted against the fraction of introns shorter than 2^8 base pairs (y-axis). Colour indicates taxonomic membership as indicated. E, points plotted as in (D), but divided into first introns (filled circles, rank = 1) and other introns (rank > 1). Pairs of points for each species are joined by vertical lines. First introns are much less likely to be smaller than 2^8 bp in teleosts.

Figure 2 Teleost intron size predicts mammalian intron size. A, median $\log_2$ intron size plotted against each other for teleosts (x-axis) and mammals (y-axis) indicate a strong correlation in size for longer introns. The vast majority of introns in both teleosts and mammals are longer than 75 bases (indicated by the dashed red lines); however we observe indicated sizes shorter than these in both clades that are likely to represent annotation artefacts rather than real short introns. B, and C, two-dimensional histograms for points in A. Colours range from dark blue (low) to bright pink (high) and indicate either the log transformed counts of the number of points within the indicated ranges (B) or, counts scaled by column to have means and standard deviations of 0 and 1 respectively. D, quantile plot of $\log_2$ median teleost and mammal intron sizes. 20% of introns have a median length of 10^2 (1024) bp or more in teleosts. E, proportion of introns having a median length above the indicated quantiles in mammals plotted against $\log_2$ teleost median length. Of the 20% of introns whose median teleost length are longer than 2^{10} bp, 73 and 20% have mammal lengths longer than 50 and 95% of mammalian introns (see right inset). F, excess of long introns in mammals ($\log_2$ observed / expected ratios) for sets indicated in E.

Figure 3 Conservation of intron sizes across the vertebrates. A, Mutual information between intron size for all pairs of species having less than 2.5% of introns shorter than 2^5 bp. Species are arranged by taxonomic order obtained from a neigbhour-joining tree based on exon alignments (Fig. SX). High mutual information is seen within taxonomic groupings, but is much lower for teleosts than mammals and the Sauria. B, Kimura-two factor distances (x-axis, based on exon alignments) plotted against mutual information (y-axis) for all within clade species pairs. Mutual information is generally lower within the teleosts than mammals for equalivent exon derived distances indicating a higher rate of change in intron size within the teleosts.

Figure 4 Evolution of intron size. Neighbour joining tree derived from exon alignments with the y-position of nodes determined by changes in intron size inferred by Sankoff maximum parsimony. Positions along the x-axis indicate cumulative Kimura two-factor distance from ancestors, positions along the y-axis indicate cumulative changes in mean ($\log_2$) intron size from ancestors. Colours of lines indicate taxonomic clade membership of leaf and inferred ancestral nodes. Hollow red lines indicate descendants of node 288 analysed in Fig. 5; other nodes discussed in the main text are also indicated. Intron sizes in teleosts (red) and aves (cyan) appear to have decreased dramatically from their respective ancestral states.

Figure 5 Retention of length in common introns. A) Proportion of introns long (> 256 bp) in node 288 (Fig. S06) that are also long in *T. rubripes* (open circles) and non-Tetraodontiformes descendants of node 288 (filled circles) plotted against the size of the non-Tetraodontiformes species. The proportion of retained introns in *T. rubripes* varies because the proportions were determined pairwise counting only introns identified in both species. B) Enrichment ($observed/expected$ ratios) of introns retained as long both in *T. fugu* and other node 288 descendants plotted against the total number of introns retained in either species. C, D) Probability of observing the given number or more of commonly retained introns (hypergeometric distribution) plotted against genome size (C) or the Kimura two-factor distance between the members of the pair (D). E, F) Observed over expected ratios plotted against genome size and phylogenetic distances as in C,D. Gray points indicate species that are members of Eupercaria; these have a closer taxonomic relationship to Tetraodontiformes and may thus have a more recent common ancestor. Nevertheless they have not in general retained more common long introns with *T. rubripes*.

Figure 6 Depletion of gene ontology (GO) categories in genes without long introns. Depletion probabilities for gene sets with short exons only calculated by the hypergeometric distribution. Genes were ordered by the minimum teleost length of their longest introns and the probability of observing the observed, or fewer, number of members of the indicated GO categories calculated for all nested sub-sets. Analyses were calculated separately for A) biological process (BP), B) molecular function (MF) and C) cellular compartment (CC). The probabilites (y-axis, -$\log_{10}$p) are shown plotted against the size ($\log_2$ transformed) of the largest intron of the nested set (x-axis).

Figure 7 Conservation of intron sequences. Intron orthologue sets were chosen by their minimal length in teleosts and the *D. rerio* intron sequence locally aligned (Smith-Waterman) to the orthologues of the set (B, D). As a control, we also aligned a non-orthologous *D. rerio* sequence to each orthologue set (A, C). The top-scoring alignment for each orthologue set was identified and filtered to remove repeat sequences. The remaining control alignments were used to construct a null model relating the search space ($l_{dr} \times \sum l_{\notin dr}$, where l_{dr} is the *D. rerio*, and $l_{\notin dr}$ the other intron lengths) to the frequency of alignment scores). Plots show the deviation from the null models for teleosts (A, B) and mammals (C, D) plotted against 10th percentile teleost intron length with each point representing the maximally scoring alignment for a given orthologue set. Points in red are outside of the 95th percentile of model residual values. Bars show the proportion of points within the indicated range that are outside of the 95th percentile (right y-axis). Left panels (A, C) show data from the control alignments of non-orthologous sequences and indicate that the model only partially compensates for the increase in search space since we observe higher residuals for longer teleost introns. Nevertheless it is clear that an increased teleost intron length is associated with an increased sequence conservation and that this effect is not limited to teleost sequences.

Additional Files

Additional file 1 — Vertebrate intron size distributions
log_2 transformed intron size distributions for teleosts. Blue and red lines indicate the distributions of all and first introns respectively. Panels are ordered by genome size from small to large. The major vertebrate clade (teleost, sauria, mammals and others) are indicated for each plot. Dashed vertical lines indicates the inferred mammalian small peak (105 bp) and the mammal and teleost antimodes at approximately 150 and 256 bp respectively.

Additional file 2 — Gene ontology depletion statistics
Individual panels show statistics associated with an under-representation of members of gene ontology groups in nested sets of human genes ordered by the minimum size of their largest orthologous teleost intron. The statistics are shown plotted against the size (log_2 transformed) of the largest teleost intron of the nested set (x-axis). The statistics shown are: $log_2(observed/expected)$ (black), $-log_{10}$ hypergeometric probabilities associated with the observations (red), q the number of members of the gene ontology set (blue) and k, the total number of genes in the nested set (purple). The dashed vertical lines show the position of the minimal hypergeometric p-values. The minimal p-values are associated with an inflexion in the observed / expected ratios and occur close to the antimode of the typical teleost intron length distribution. Pages 1-7: Biological Process (BP), 8: Molecular Function (MF), 9-11 Cellular Compartment (CC).

Additional file 3 — Local alignments of *D. rerio* intron sequences (long)
Local alignments of *D. rerio* intron sequences to introns long (< 1024 bp) in all teleosts and vertebrate intron orthologues. Alignments were performed recursively to identify all non-overlapping alignments to the *D. rerio* sequence.

Additional file 4 — Local alignments of *D. rerio* intron sequences (ctl)
Local alignments of *D. rerio* intron sequences to introns with variable teleost length and vertebrate intron orthologues. introns were long (> 1024 bp) in *D. rerio* but had a median teleost length less than 256 bp. Alignments were performed recursively to identify all non-overlapping alignments to the *D. rerio* sequence.

Additional file 5-14 — Transcript and intron alignments for points in Fig. S17.
Each panel shows the maximally scoring alignment between *D. rerio* and teleost intron orthologues (lower) and transcript alignment (upper) used to establish the intron orthology. Grey, white and red parts indicate aligned exonic sequence, gaps and positions of intron meta-characters respectively. Colours in intron alignment represent bases (A blue, C cyan, G green, T brown, N grey, gap white). Curves lying between sequence representations show a normal kernel density smoothed estimate of local similarity (9 bp window, standard deviation two); vertical lines indicate matches. Region between exon and intron alignments indicates the location of the maximally scoring alignment in the introns. Upper sequence *D. rerio*. Files 5-9 and 10-14 contain alignments to teleost and mammalian sequences respectively. Each file corresponds to one panel in Fig. S17 and to specific teleost size class: Files 5,10: long (E,J), 6,11: medium (D,I), 7,12: short.2 (C,H), 8,13 short (B, G) and 9,14 (A,F).

Figures

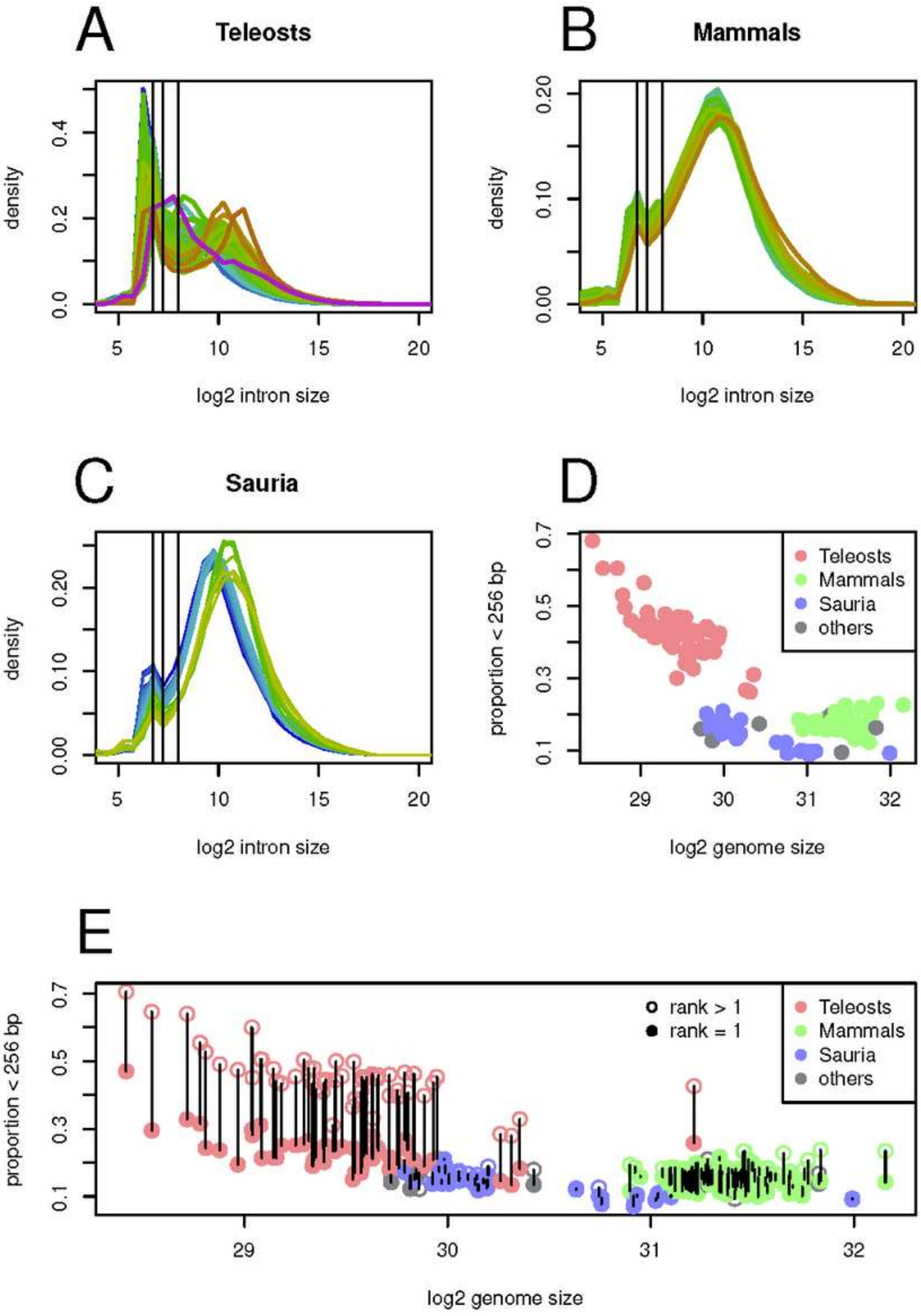

Figure 1

Vertebrate intron size. A-C, distributions of log2intron size in teleosts (A), mammals (B) and Sauria (C). Colour of lines indicate log genome size from blue (small) to purple (large); x-axis, log2 intron size, y-axis density. Distributions for teleosts are much more varied than for mammals or Sauria, but are generally bimodal with an antimode at 28 bases (indicated by vertical line). The distributions of mammals and Sauria also suggest bimodality with a peak of shorter introns at 105 bp. Distributions for Sauria separate naturally into two groups of different genome size ranges; these correspond to Aves (birds) and non-Aves species, with Aves species (blues) having both smaller genomes and introns. D, log2 genome size (x-axis) plotted against the fraction of introns shorter than 28 base pairs (y-axis). Colour indicates taxonomic membership as indicated. E, points plotted as in (D), but divided into first introns (filled circles, rank = 1) and other introns (rank > 1). Pairs of points for each species are joined by vertical lines. First introns are much less likely to be smaller than 28 bp in teleosts.

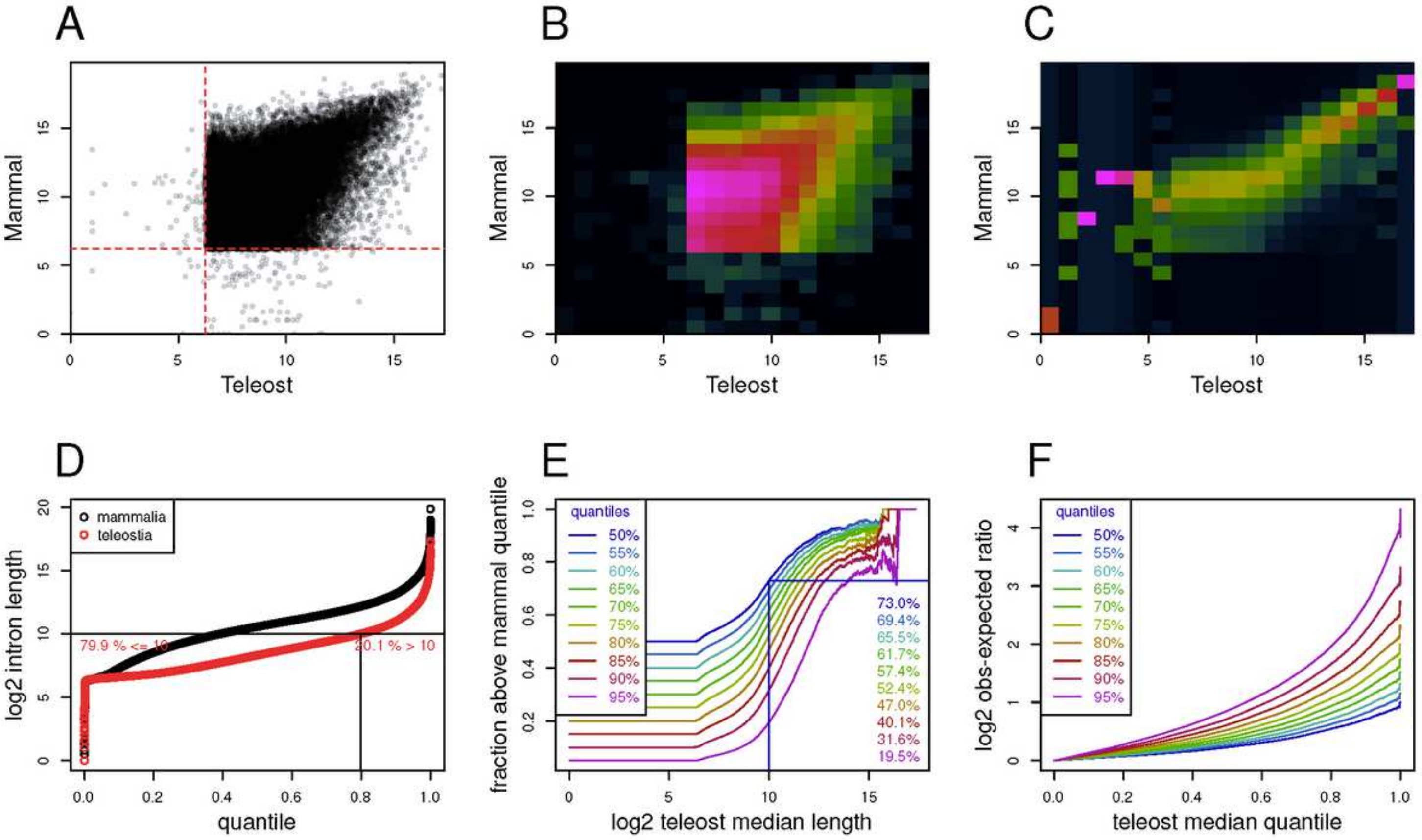

Figure 2

Teleost intron size predicts mammalian intron size. A, median log2 intron size plotted against each other for teleosts (x-axis) and mammals (y-axis) indicate a strong correlation in size for longer introns. The vast majority of introns in both teleosts and mammals are longer than 75 bases (indicated by the dashed red lines); however we observe indicated sizes shorter than these in both clades that are likely to represent annotation artefacts rather than real short introns. B, and C, two-dimensional histograms for points in A. Colours range from dark blue (low) to bright pink (high) and indicate either the log transformed counts of the number of points within the indicated ranges (B) or, counts scaled by column to have means and standard deviations of 0 and 1 respectively. D, quantile plot of log2 median teleost and mammal intron sizes. 20% of introns have a median length of 102 (1024) bp or more in teleosts. E, proportion of introns having a median length above the indicated quantiles in mammals plotted against log2 teleost median length. Of the 20% of introns whose median teleost length are longer than 210 bp, 73 and 20% have mammal lengths longer than 50 and 95% of mammalian introns (see right inset). F, excess of long introns in mammals (log2 observed / expected ratios) for sets indicated in E.

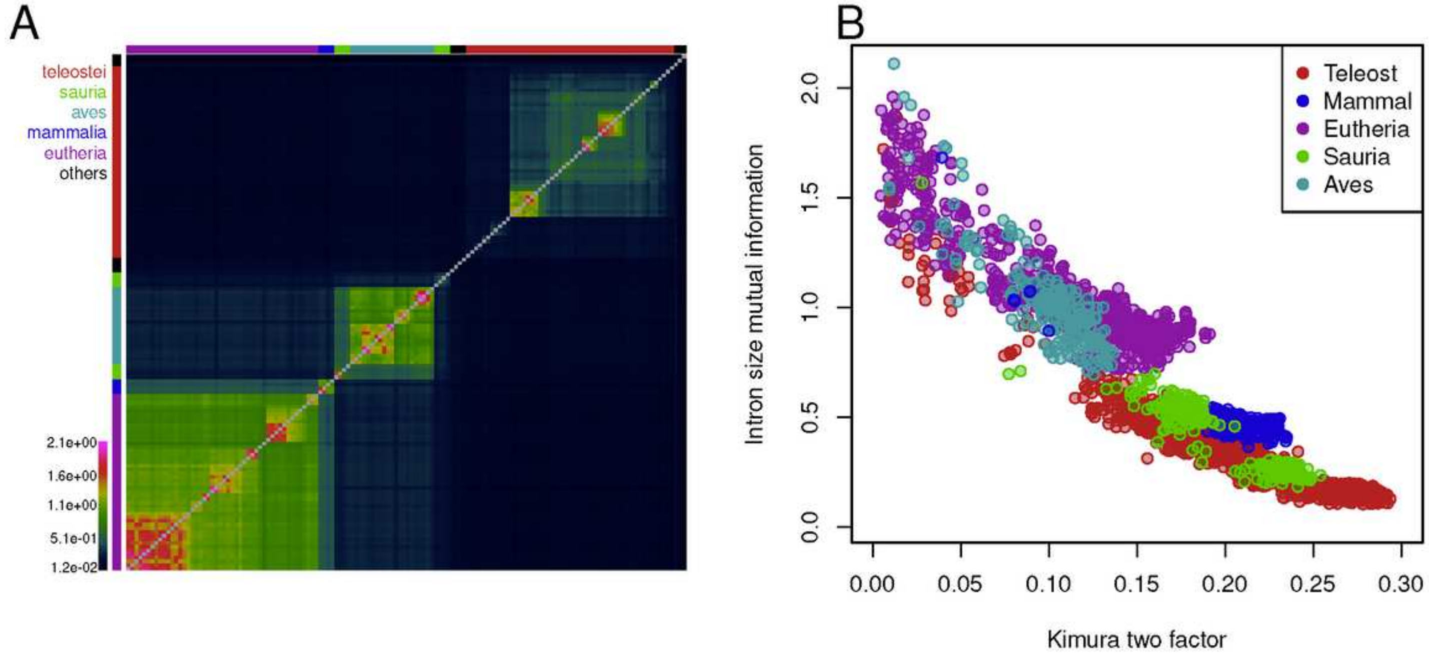

Figure 3

Conservation of intron sizes across the vertebrates. A, Mutual information between intron size for all pairs of species having less than 2.5% of introns shorter than 25 bp. Species are arranged by taxonomic order obtained from a neigbhour-joining tree based on exon alignments (Fig. SX). High mutual information is seen within taxonomic groupings, but is much lower for teleosts than mammals and the Sauria. B, Kimura-two factor distances (x-axis, based on exon alignments) plotted against mutual information (y-axis) for all within clade species pairs. Mutual information is generally lower within the teleosts than mammals for equalivent exon derived distances indicating a higher rate of change in intron size within the teleosts.

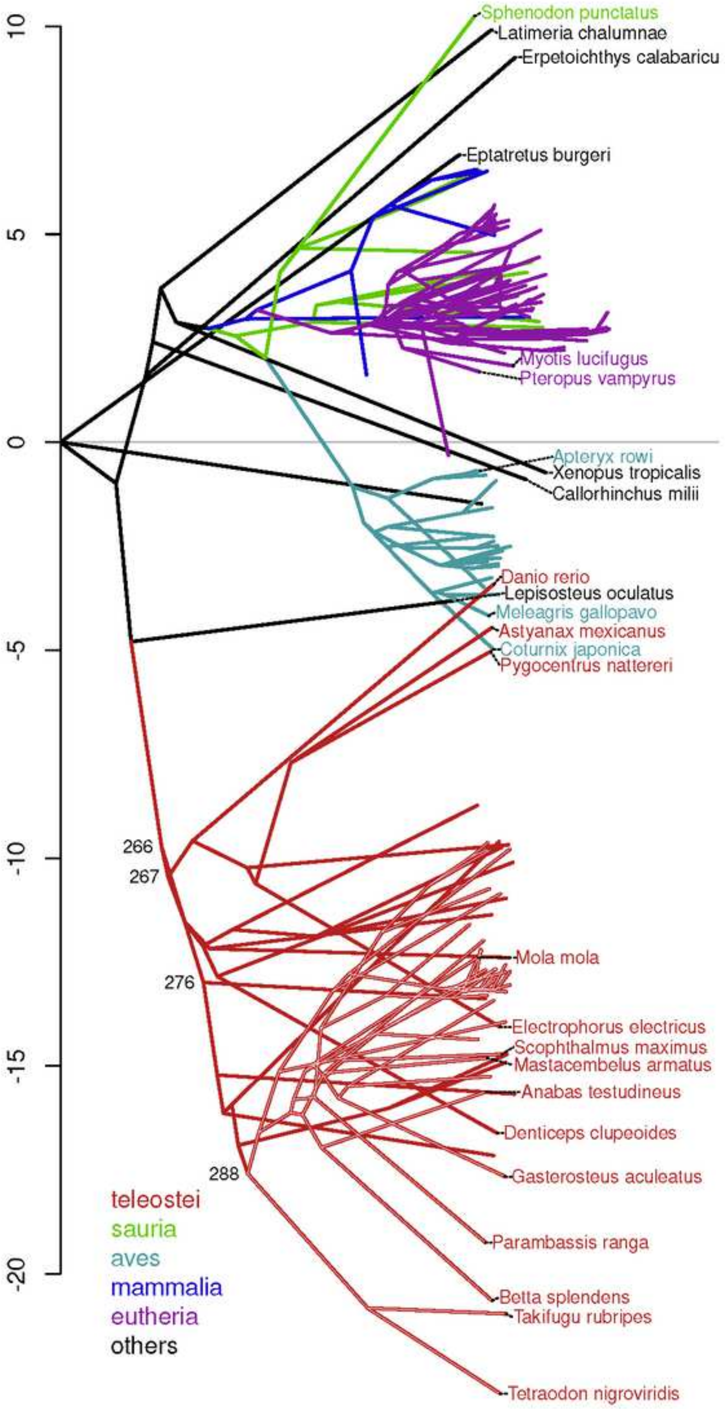

Figure 4

Evolution of intron size. Neighbour joining tree derived from exon alignments with the y-position of nodes determined by changes in intron size inferred by Sankoff maximum parsimony. Positions along the x-axis indicate cumulative Kimura two-factor distance from ancestors, positions along the y-axis indicate cumulative changes in mean (log2) intron size from ancestors. Colours of lines indicate taxonomic clade membership of leaf and inferred ancestral nodes. Hollow red lines indicate descendants of node 288 analysed in Fig. 5; other nodes discussed in the main text are also indicated. Intron sizes in teleosts (red) and aves (cyan) appear to have decreased dramatically from their respective ancestral states.

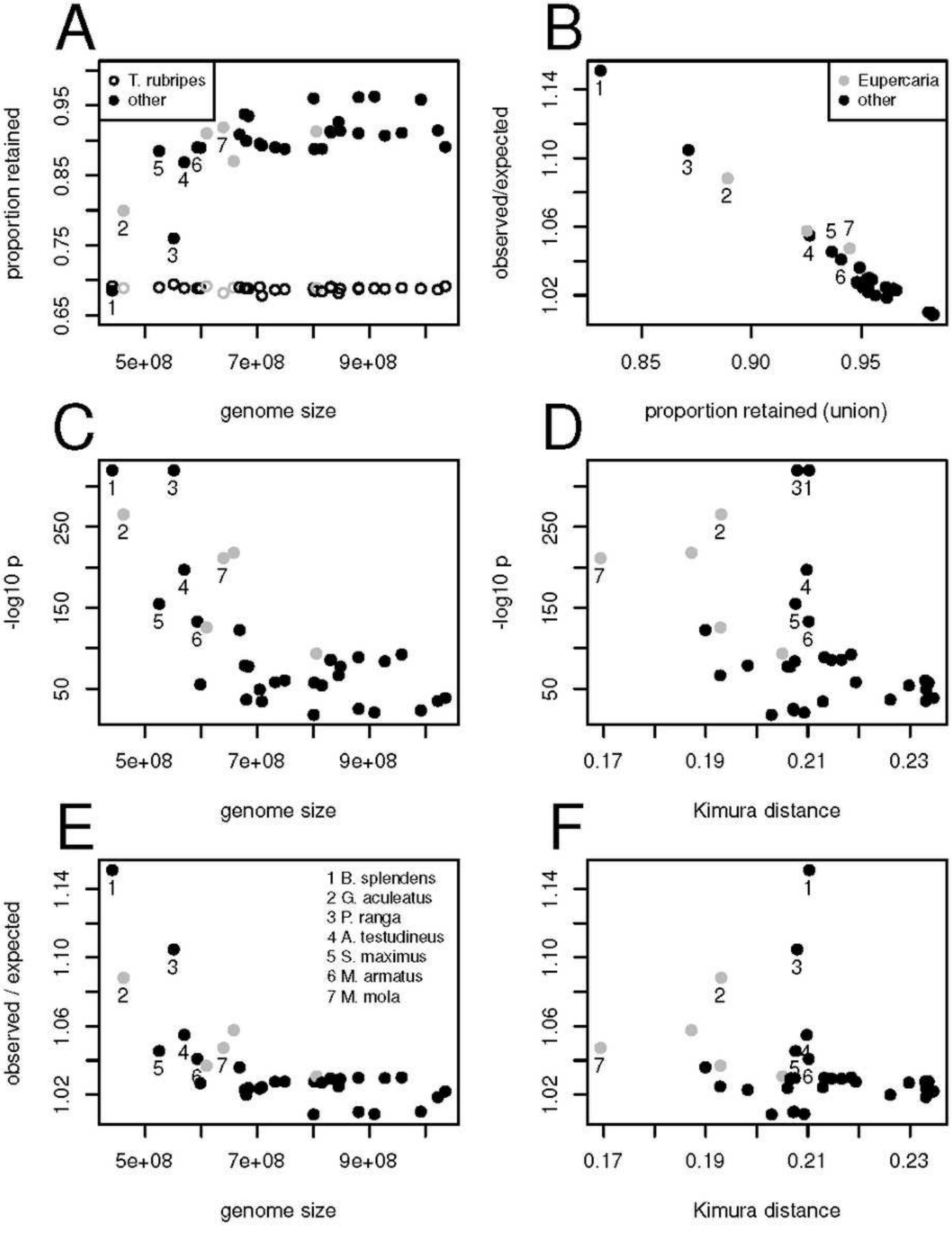

Figure 5

Retention of length in common introns. A) Proportion of introns long (> 256 bp) in node 288 (Fig. S06) that are also long in T. rubripes (open circles) and non-Tetraodontiformes descendants of node 288 (lled circles) plotted against the size of the non-Tetraodontiformes species. The proportion of retained introns in T. rubripes varies because the proportions were determined pairwise counting only introns identied in both species. B) Enrichment (observed=expected ratios) of introns retained as long both in T. fugu and other node 288 descendants plotted against the total number of introns retained in either species. C, D) Probability of observing the given number or more of commonly retained introns (hypergeometric distribution) plotted against genome size (C) or the Kimura two-factor distance between the members of the pair (D). E, F) Observed over expected ratios plotted against genome size and phylogenetic distances as in C,D. Gray points indicate species that are members of Eupercaria; these have a closer taxonomic relationship to Tetraodontiformes and may thus have a more recent common ancestor. Nevertheless they have not in general retained more common long introns with T. rubripes.

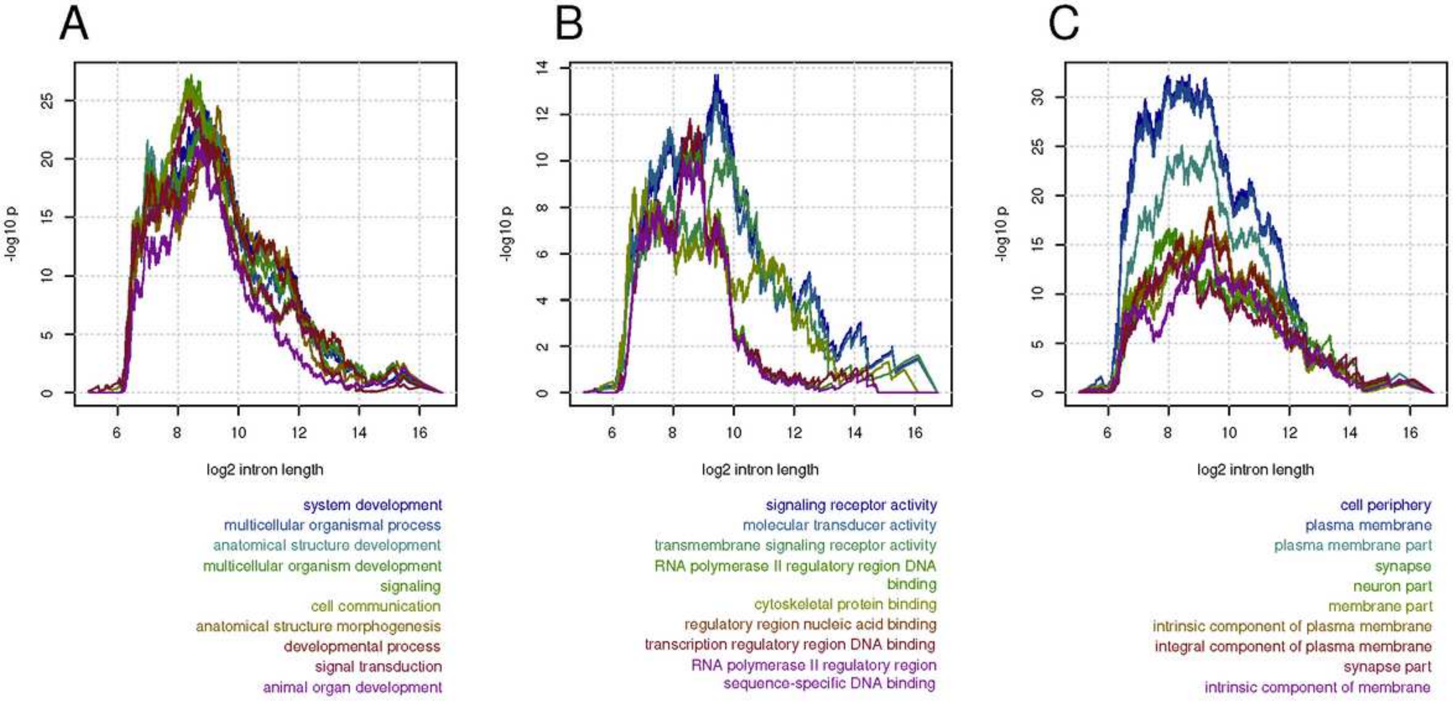

Figure 6

Depletion of gene ontology (GO) categories in genes without long introns. Depletion probabilities for gene sets with short exons only calculated by the hypergeometric distribution. Genes were ordered by the minimum teleost length of their longest introns and the probability of observing the observed, or fewer, number of members of the indicated GO categories calculated for all nested sub-sets. Analyses were calculated separately for A) biological process (BP), B) molecular function (MF) and C) cellular compartment (CC). The probabilites (y-axis, -log10p) are shown plotted against the size (log2 transformed) of the largest intron of the nested set (x-axis).

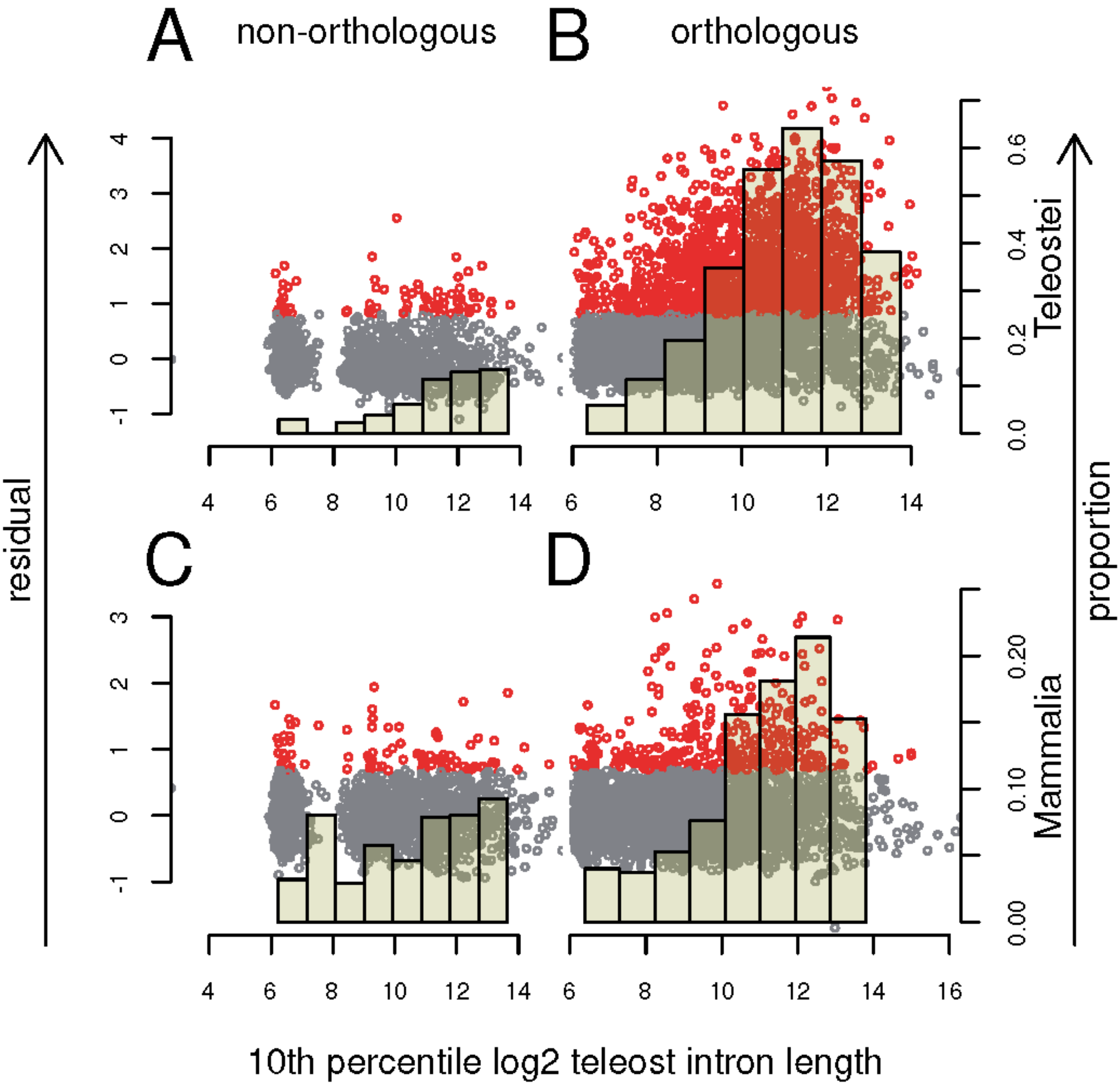

Figure 7

Please see the Manuscript Doc file for the complete figure caption.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- additionalfile1.pdf
- additionalfile10.pdf
- additionalfile11.pdf
- additionalfile12.pdf
- additionalfile13.pdf
- additionalfile14.pdf
- additionalfile15.pdf
- additionalfile16.pdf
- additionalfile2.pdf
- additionalfile3.pdf
- additionalfile4.pdf
- additionalfile5.pdf
- additionalfile6.pdf
- additionalfile7.pdf
- additionalfile8.pdf
- additionalfile9.pdf
- supplementarymethods.pdf