GBE

Genome Biology and Evolution

SMBE

Comparative Analysis of Teleost Genome Sequences Reveals an Ancient Intron Size Expansion in the Zebrafish Lineage

Stephen P. Moss, Domino A. Joyce, Stuart Humphries, Katherine J. Tindall, and David H. Lunt*

Department of Biological Sciences, The University of Hull, Kingston-Upon-Hull, United Kingdom

*Corresponding author: E-mail: d.h.lunt@hull.ac.uk.

Accepted: 30 August 2011

Abstract

We have developed a bioinformatics pipeline for the comparative evolutionary analysis of Ensembl genomes and have used it to analyze the introns of the five available teleost fish genomes. We show our pipeline to be a powerful tool for revealing variation between genomes that may otherwise be overlooked with simple summary statistics. We identify that the zebrafish, *Danio rerio*, has an unusual distribution of intron sizes, with a greater number of larger introns in general and a notable peak in the frequency of introns of approximately 500 to 2,000 bp compared with the monotonically decreasing frequency distributions of the other fish. We determine that 47% of *D. rerio* introns are composed of repetitive sequences, although the remainder, over 331 Mb, is not. Because repetitive elements may be the origin of the majority of all noncoding DNA, it is likely that the remaining *D. rerio* intronic sequence has an ancient repetitive origin and has since accumulated so many mutations that it can no longer be recognized as such. To study such an ancient expansion of repeats in the *Danio*, lineage will require further comparative analysis of fish genomes incorporating a broader distribution of teleost lineages.

Key words: genome evolution, teleosts, introns, repeat elements, comparative genomics pipeline.

Introduction

Introns are a major component of metazoan genomes, comprising approximately 24% of the human genome compared with only 1.1% for exons (Venter et al. 2001). Even in species with genomes considerably smaller than humans and representing taxonomically diverse lineages, introns can account for a substantial proportion of the genome. The nematode *Caenorhabditis briggsae*, for example, has introns containing 1.3 times as many nucleotides as do exons, which together account for approximately 30% of the entire genome sequence (Stein et al. 2003). Intron sequence in general evolves at a high rate, close to that of 4-fold degenerate sites, pseudogenes, and noncoding regions (Hughes and Yeager 1997; Chamary and Hurst 2004; Gaffney and Keightley 2006). Despite this, introns may also contain gene regulatory elements (Majewski and Ott 2002; Gaffney and Keightley 2006), and their impact on translation, via alternative splicing, can also be substantial (Mironov et al. 1999; Kim et al. 2007). Even without the presence of regulatory elements within introns, they may still contribute strongly to the deleterious mutation rate. Correct splicing requires the maintenance of specific splicing signals at the start and end of each intron, an interior branch point adenine, and a number of other sequences imperfectly conserved across eukaryotes involved in the recruitment of the spliceosome (Schwartz et al. 2008). Together, these sequences increase the mutational load of intron-containing genes because mutations in any of the required splicing signals can lead to nonfunctionalization of the locus. The several hundred thousand introns in a vertebrate genome are therefore a considerable mutational burden, and it has been estimated that perhaps a third of all human genetic disorders involve mutations affecting splice-site recognition (Frischmeyer and Dietz 1999; López-Bigas et al. 2005). The study of introns can therefore greatly aid in our understanding of the genome's mutational dynamics and in the selectively maintained regulation of surrounding coding regions.

There are diverse mechanisms by which introns may be gained including reverse splicing, local duplications, transposable elements, and transfer from paralogs by unequal crossing over (Sharp 1985; Rogers 1989; Hankeln et al. 1997; Iwamoto et al. 1998; Roy and Gilbert 2006). Subsequent to its origin, introns will change in size due to the accumulation of small

insertions and deletions, nonhomologous recombination, and the action of transposable elements. Repetitive sequences such as transposable elements occupy from 33% to 52% of sequenced vertebrate genomes, and it has been shown that 20–60% of vertebrate introns contain transposable elements (Mills et al. 2007; Sela et al. 2010). Intron frequency and mean intron size are known to vary considerably across animal taxa (Deutsch and Long 1999; Lynch and Conery 2003; Roy and Gilbert 2006; Yandell et al. 2006; Gazave et al. 2007; Zhu et al. 2009), though few investigations have been able to compare the intron composition of entire genomes within and between closely related taxa. There are a small number of previous whole genome studies of introns, although these have often been limited to one-to-one comparisons or groups of phylogenetically very divergent organisms (Coghlan and Wolfe 2004; Marais et al. 2005; Yandell et al. 2006; Gazave et al. 2007; Stajich et al. 2007; Sharpton et al. 2008; Li et al. 2009). A full understanding of the processes shaping intron diversity and evolution will require a large-scale comparative genomic approach making full use of the rapidly increasing number and diversity of whole genome sequences. Such evolutionary comparative genomic studies, however, are slowed by substantial analytical technical challenges presented to most biologists in dealing with these huge amounts of data. In this study, we present a bioinformatics pipeline that can be used to compare the size distribution and content of introns in a comparative genomics study. We investigate the potential of such a genomic approach by comparing introns in the genomes of five teleost fish available at Ensembl (Flicek et al. 2010)—the zebrafish (*Danio rerio*), three-spined Stickleback (*Gasterosteus aculeatus*), Medaka (*Oryzias latipes*), Fugu (*Takifugu rubripes*), and Tetraodon (*Tetraodon nigroviridis*). These fish have been used as model organisms in the laboratory for a number of years, and a great deal of research has been undertaken focusing on their anatomical and physiological structure (Haffter et al. 1996; Aparicio et al. 2002; Jaillon et al. 2004; Kimura et al. 2004; Roest Crollius and Weissenbach 2005). We feel that the information that can be elucidated from their genomes, in relation to the biological processes driving or constraining their genomic evolution is therefore of particular interest. Our pipeline (Genome Comparison and Analysis Toolkit) has allowed us to characterize, in detail, the composition and diversity of approximately 1 million introns in these teleost genomes and provides a valuable open source extensible platform for comparative genomics of introns and other genomic components.

Materials and Methods

Sequences Used

The intron data were retrieved from the Ensembl Core online database, release number 61. The individual fish database versions were danio_rerio_core_61_9a, gasterosteus_aculeatus_core_61_1n, oryzias_latipes_core_61_1m, takifugu_rubripes_core_61_4o, and tetraodon_nigroviridis_core_61_8f.

Method of Access

The data were accessed using a novel bioinformatics pipeline, built using the Perl Ensembl Core Software Libraries (Stabenau et al. 2004), along with BioPerl (Stajich et al. 2002) and several open-source Comprehensive Perl Archive Network (CPAN) libraries. Some information was verified manually using Ensembls' BioMart website and the Ensembl MySQL databases. The pipeline code is available at http://github.com/gawbul/gcat. Information on the specific software requirements is available in the supplementary materials (Supplementary Material online), along with an overview of the pipeline workflow (supplementary fig. S2, Supplementary Material online).

Intron Sequence Retrieval

Intron sequences were retrieved using the canonical transcript for each gene, as defined by the Ensembl Core database. The database and application programming interface (API) are designed in such a way that the intron sequences can only be retrieved automatically via their associated transcript, but because there can be multiple transcripts per gene, this can result in redundant intron data. We chose to use the canonical transcript for each gene, as these are explicitly annotated in the Ensembl database. We could also have taken a transcript at random or the largest transcript for each gene, but we believe our method presents the least bias. It is possible that some exons in additional transcripts of the gene could overlap the introns of the canonical transcript but we believe this error to be small and to not significantly affect our results. Introns aren't explicitly defined in the database and are instead implicitly defined from the exon coordinates by the Ensembl Perl API, and our pipeline was used to automate the intron retrieval process. Because we anticipated that annotation of non-protein–coding genes would vary with genome annotation quality, we restricted our analyses to introns in genes matching the biotype "protein coding," which represented greater than 98% of all introns in all fish.

Frequency Distributions

The frequency distributions were built for each of the five fish using our pipeline via the Statistics::Descriptive CPAN package and plotted using custom-made R scripts. The Comprehensive R Archive Network package gdata was used to provide functionality for concatenating multiple columns of csv data, but all other calculations were made using novel R code, built on top of the core R functionality. The calculations for the sliding window means and confidence intervals were calculated from a subset of the intron frequency data, consisting of successive 25-bp windows between 1

and 5,000 bp. This resulted in 200 points being plotted and reduced any noise due to variation in intron size frequency within each window, but did not affect the overall shape of the distributions.

Determining Repeat Element Content and Unique Intron Size

Ensembl explicitly defines repeat elements, as determined by the RepeatMasker, DUST, and TRF software (Benson 1999; Smit et al. 2011; Morgulis et al. 2006), as annotation features in its database, and these were retrieved by our pipeline for the canonical transcript of each gene matching the protein-coding biotype. We also used WindowMasker (Morgulis et al. 2006) to check for repeats, as the quality and coverage of the RepBase repeat libraries (Jurka et al. 2005) used by RepeatMasker has previously been questioned (Bergman and Quesneville 2007). A novel bioinformatics script (see count_wm_repeats.py in the git repository) was developed to parse the WindowMasker results, in order to determine the unique sequence length of each intron by removing its total repeat element length.

Intron Position and Type

Intron frequencies per gene region (5′-UTR, coding sequences [CDS], 3′-UTR) were calculated according to Ensembl annotations. We corrected for size differences between regions by calculating introns per bond, where the number of phosphodiester bonds in each region is equal to the nucleotide count for the UTRs and the nucleotide count minus one for the CDS because UTRs are defined by reference to the CDS coordinates in our pipeline.

Additional to this, we calculated the intron type based on explicit splice-site nucleotides, matching 5′ GU-AG 3′ and 5′ AU-AC 3′ for the U2 and U12 intron categories, respectively. Any introns not matching these definitions were placed in an "other" category.

Results

Intron Retrieval and Characterization

Table 1 presents intron size and frequency data as provided by Ensembl. We retrieved between 185,494 (*O. latipes*) and 221,589 (*D. rerio*) introns per genome totaling 982,544 introns between these five fish. The smallest total intron lengths were those of the pufferfish *T. rubripes* and *T. nigroviridis* at 90,447,562 bp and 108,524,412 bp, respectively. *Danio rerio* has the largest at 622,476,590 bp as well as the lowest intron density of all the teleost genomes with 8.93 introns per gene compared with 9.80 to 10.51 introns per gene for the other four fish. Despite this, at 622 Mb of intronic DNA, *D. rerio* has from 2.8 to 6.9 times more intronic sequence than the other fish.

Frequency Distributions of Teleost Intron Size

Figure 1*a* shows a frequency plot of intron size class in all five fish, with 5% and 95% confidence intervals. The ordinate is a log-scaled count, and the abscissa represents the mean of 25-bp sliding windows of intron size. We observe a change in the shape of the intron distribution in *D. rerio* that is not present in the other fish. The minimum, mode, and maximum intron sizes for each fish are given in table 1. Above the 5,000-bp cutoff in figure 1*a*, the number of instances of each individual size class is very low, causing a great scatter in values, although the trend does not differ.

Repeat Element Content and Unique Intron Size

The length of repeat elements determined by RepeatMasker (see Materials and Methods) ranges from 942,285 (*T. nigroviridis)* to 13,406,652 bp (*D. rerio)* comprising between 0.66% (*O. latipes)* and 2.15% (*D. rerio)* of total intronic sequence (supplementary table S1, Supplementary Material online). A summary of the subsequent WindowMasker analysis is shown in table 2, giving a breakdown of the repeat elements, and the unique intron sizes calculated. WindowMasker calculated between 20,313,082 (*T. nigroviridis*) and 291,676,913 bp (*D. rerio*) with from 2.71 to 20.69 repeats per intron. This accounts for between 22.46% (*O. latipes)* and 46.86% (*D. rerio)* of total intronic sequence. We used the WindowMasker results to replot intron size frequency, as shown in figure 1*a*, using the unique intron sequence frequency distributions of all introns after repeat element trimming (fig. 1*b*).

Large Introns

The maximum intron size found in each genome is presented in table 1. These are not solitary outliers, however, with 1,228 (0.6%) *D. rerio* introns greater than 50,000 bp in size (here referred to as "large introns" after Shepard et al. (2009)). There are between 16 and 221 introns in the other fish (supplementary table S2, Supplementary Material online) accounting for between 0.9% (*T. nigroviridis)* and 17% (*D. rerio)* of total intron length. Our figure for *D. rerio* large introns is different from the 756 reported by Shepard et al. (2009), perhaps because their data were retrieved from a custom database and represents an earlier version of the *D. rerio* genome. However, our teleost large intron values do fall within the range of 7 (mosquito) to 3,473 (human), previously reported for metazoan (Shepard et al. 2009).

Small Introns

We refer to "small introns" as those less than 80 bp, which approximates the mode of the pooled teleost data set. These comprise from 11,473 (*D. rerio*) to 44,755 (*T. nigroviridis*) introns accounting for between 0.12% and 2.97% of the total intronic sequence, respectively (supplementary fig. S1 and table S3, Supplementary Material online).

Table 1
The Summary Statistics for the Five Teleost Fish

	Danio rerio	*Gasterosteus aculeatus*	*Oryzias latipes*	*Takifugu rubripes*	*Tetraodon nigroviridis*
Genome size	1,412,464,843	461,533,448	868,983,502	393,312,790	358,618,246
Number of genes	32,312	22,456	20,422	19,388	20,562
Number of transcripts	51,569	29,245	25,397	48,706	24,078
Protein coding genes	24,803	20,109	18,920	17,876	18,872
Canonical transcripts	24,803	20,109	18,920	17,876	18,872
Introns per gene	8.93	9.93	9.80	10.51	9.96
Number of introns	221,589	199,624	185,494	187,962	187,875
Maximum intron length	378,145	175,269	295,125	93,537	631,227
Total intron length	622,476,590	151,619,269	219,591,667	108,524,412	90,447,562
Mean length	2,809	760	1,184	577	481
Median length	984	219	252	143	118
Mode length	84	85	77	78	76
25th percentile length	138	104	90	84	80
75th percentile length	2,563	615	1,026	450	350
GC content	50.58%	50.48%	47.10%	40.39%	49.21%
Percentage of genome	44.07%	32.85%	25.27%	27.59%	25.22%

Note.—We include total genome size, total number of genes, and total number of transcripts, but our study focuses on the introns found within the genes matching Ensembl's protein_coding biotype.

Intron Location

Within protein-coding transcripts, introns may occur in the coding region (CDS) or either of the terminal untranslated regions (5′-UTR or 3′-UTR). Of those 24,803 *D. rerio* transcripts containing all three regions, 2.08% of introns were in 5′-UTR, 0.57% in 3′-UTR, and 97.35% in the CDS. Similar percentages were found in the other fish (supplementary table S4, Supplementary Material online). Correcting for the sizes for these three regions, we find 3.4×10^{-4} introns per bond in the CDS, 1.5×10^{-4} introns per bond in the 5′-UTR, and 0.6×10^{-4} introns per bond in the 3′-UTR.

Splice Signals

This teleost introns data set contained introns bounded by the typical GU-AG splice signal (U2-type), AU-AC splice signal (U12-type), and those employing other splice signals.

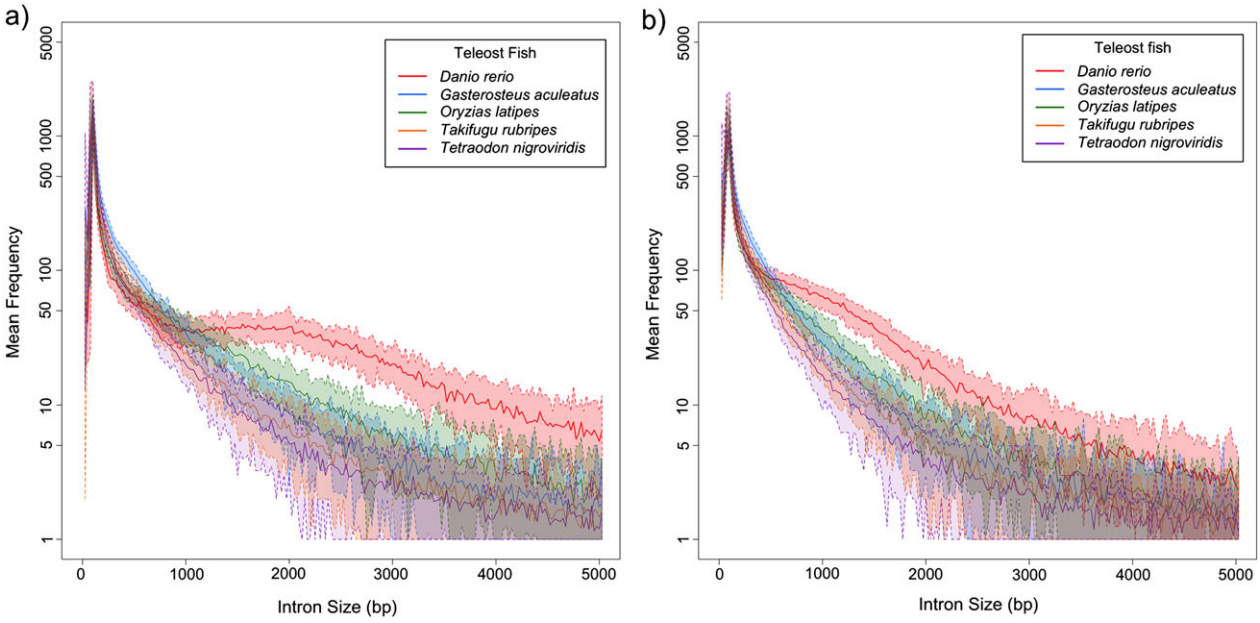

Fig. 1.—(*a*) A frequency distribution plot of intron size in the five teleost fish. Each point represents the mean of intron sizes within a 25-bp sliding window. The lower and upper dashed lines represent the 5% and 95% confidence intervals, respectively. All fish present an initial peak of approximately 80 bp and then decay in a similar pattern, with the exception of *Danio rerio*, which has a second peak between 500 and 2,000 bp and, subsequently, decays parallel to the others. (*b*) A frequency distribution plot of unique intron size in the five teleost fish, representing the intron sizes after removal of repeat sequences.

Table 2
A Summary of Repeat Element Content in the Five Teleost Fish, Determined Using the WindowMasker Software

	Danio rerio	*Gasterosteus aculeatus*	*Oryzias latipes*	*Takifugu rubripes*	*Tetraodon nigroviridis*
Number of repeat elements	4,583,943	891,753	1,498,499	591,789	509,271
Length of repeat elements	291,676,913	31,910,164	74,289,913	20,701,619	20,313,082
Number of repeat elements per intron	20.69	4.47	8.08	3.15	2.71
Percentage of intron length	46.86%	21.05%	33.83%	19.08%	22.46%
Length of unique introns	330,799,677	119,709,105	145,301,754	87,822,793	70,134,480

Tetraodon nigroviridis, at 82.51%, has the lowest percentage of typical GU-AG introns and *D. rerio*, at 93.57%, the highest. All fish have a similar number of U12-type introns, with some variation in other introns (supplementary table S5, Supplementary Material online).

Discussion

We have employed a novel comparative genomic pipeline to perform detailed comparison of the intron characteristics of five teleost fish genomes. This allowed us to identify the diversity of intron content and characteristics across the whole genome and to partition these data into biologically relevant categories. Previous approaches to such characterization have typically either restricted themselves to single comparisons or else incorporated exceptionally divergent organisms (Coghlan and Wolfe 2004; Marais et al. 2005; Yandell et al. 2006; Gazave et al. 2007; Stajich et al. 2007; Sharpton et al. 2008; Li et al. 2009). Because our bioinformatic pipeline has been designed to build on the high-quality genome annotations present at Ensembl and use open-source software libraries such as BioPerl, this approach can be easily integrated into more general studies in comparative genomics. For the analysis of teleost genome data presented here, our pipeline has proved itself to be highly automated, yet flexible, fast, and to lend itself to evolutionary and statistical approaches to comparative genomics.

Intron Size Distributions

Our characterization of teleost introns shows that *D. rerio*, the species with the largest total genome size, has more and larger introns than any of the other fish genomes. Although simple summary statistics such as "average intron length" are commonly applied to the description of a genome's intron content in the literature, these can be significantly influenced by outlier values and miss many of the important differences between taxa. The mean intron length for *D. rerio* is 2,809 bp, yet 50% of all introns are found below 985 bp in length with the modal size only 84 bp. Figure 1*a* also shows the shape of intron frequency for each fish up to intron sizes of 5,000 bp. *Oryzias latipes* has more than twice the mean intron size of *T. nigroviridis* and *T. rubripes*, yet the distribution of intron sizes in figure 1*a* shows them to be remarkably similar. In contrast to the mean, modal intron size is relatively tightly grouped among these five fish, in the range 76 to 85 bp, despite approximately 150 My divergence (Benton and Donoghue 2007) (table 1; fig. 1*a*). For the pooled set of teleost introns, the mean size is 1,214 bp (range 481 to 2,809 bp), yet the mode intron size is a mere 81 bp with up to 37% of introns within 20 bp of this mode value. The zebrafish *D. rerio* has a modal intron size only 1 bp different from the stickleback *G. aculeatus*, yet contains 4.1 times as much intronic DNA, an extra 471 Mb. Most introns across fish are small and similar in length, yet introns much bigger than this mode size vary and contribute extensively to the differences between fish. Although 50% of all introns in *D. rerio* are less than 985 bp, these account for only 4.8% of all intronic nucleotides.

The comparisons of intron size frequency distributions generated here highlight the unique pattern present in the *D. rerio* genome. The multimodal distribution we see with zebrafish contrasts with the monotonically decreasing pattern in the other fish (fig. 1*a*). The shape of this curve represents separate genomic processes generating an intron size distribution with a broad peak of approximately 500 to 2,000 bp in addition to the usual teleost approximately 80-bp mode size.

Our analyses emphasize that overreliance on simple summary statistics, such as mean or mode intron size, can obscure real biological trends and differences that would be revealed with much more detailed investigation of the distribution of the data as a whole.

Repeat Element Content as an Explanation of Intron Size Differences

Zebrafish has both more and larger introns than the other fish (fig. 1*a*, table 1), accounting for between 402 and 532 million extra nucleotides compared with the other fish genomes. Repetitive elements are known to be the major cause of genome size variation (Mills et al. 2007; Sela et al. 2010), and we were interested to see if they also accounted for the difference in intron size between these teleosts, in particular the increased intron content of *D. rerio*. We took two different approaches to determine this. The first relied on the annotations available at Ensembl, which uses the RepeatMasker software and compares data against a curated library of repeats using local alignment methods. The standard repeat libraries however may not

have optimal quality and coverage for some taxa (Morgulis et al. 2006; Bergman and Quesneville 2007). The second approach used the WindowMasker program, which compares the genome against itself to identify repeats and is therefore independent of previous repeat curation in closely related taxa. It implements the DUST and WinMask algorithms to identify low-complexity regions and global repeats, respectively, by identifying and scanning for repetitive regions within the genome sequence.

Using the Ensembl annotations, we detected repeat elements accounting for from 0.66% to 2.15% of the total intronic length. A much larger proportion of intronic sequence was characterized as repetitive using WindowMasker (table 2) with *D. rerio* introns containing 46.86% repeat sequences. This result for *D. rerio* agrees with the values obtained by Sela et al. (2010). WindowMasker doesn't annotate the repeats; however, thus one can't determine the class of repetitive elements they belong to.

The increased percentage of repeat elements within the *D. rerio* intron sequences accounts for some of the difference in its frequency distribution (fig. 1*b*). It is possible that the additional proportion of this sequence was formerly repetitive and has since decayed beyond our ability to recognize it as such. Because repetitive elements are likely to be the origins of the majority of all noncoding DNA (Smit 1999; Lander et al. 2001), we propose that the *Danio* lineage experienced an early burst of repeat element expansion that has been decaying for many millions of years. Figures 2*a* and *b* show the frequency distribution of repeat elements within the major class of introns (500–2,000 bp), which includes the region comprising the second intron size peak in *D. rerio* (fig. 1*a*). If there had been a recent expansion of particular repeat elements figure 2*a* would be expected to show peaks in the frequency of specific size classes. Contrary to this, our analysis reveals a gradual decline in the repeat element size frequency distribution, indicating no recent large-scale repeat expansions. Figures 2*a* and *b* also show that the frequencies of both individual and cumulative repeat element sizes are greater in *D. rerio* within the size range expected to contribute to the second zebrafish peak in figure 1*a*. We consider it likely therefore that repeat elements have contributed importantly to the second *D. rerio* intron size peak but that this striking repeat expansion was an ancient rather than recent genomic change.

The differences in the distributions may also represent a continuum that with increased sampling within the teleostei infraclass, particularly of those species intermediate to those presented here, would fill the gap. *Oryzias latipes* exhibits a very subtle difference in its intron size class distribution and in being more closely related to *D. rerio* than any of the other fish, adds some weight to this argument. As the genomes of more fish become available it will allow us to continue this kind of research and test specific hypotheses on the differences we have observed. This would be best approached within a phylogenetic comparative framework in order to observe how much of the variation we see can be accounted for by phylogeny. By comparing orthologous loci, for example, we can make more accurate inferences on the likely ancient nature of these introns, which could then be applied to more taxa.

Large Introns

Large introns can present several problems for organisms, including the expense of transcription and the difficulty of splicing large introns (Shepard et al. 2009). The 1,228 large introns in *D. rerio* consist of 107,485,505 nucleotides, which is 17.3% of all *D. rerio* intronic nucleotides and 7.6% of the entire genome sequence. Such large introns may be very costly with regard to both the time and energy required for synthesis (Wagner 2005). Intronic nucleotides are removed from the mRNA before its export from the nucleus and the synthesis and subsequent degradation of introns has a cost approximately proportional to the length of those introns multiplied by the frequency of transcription. Large introns constitute 15.8% of the transcribed section of the genome in *D. rerio* and therefore account for approximately 1 h in additional transcription time per large intron, at a cost of at least 175,000 molecules of ATP, a significant extra metabolic cost to the cell (Castillo-Davis et al. 2002).

In addition to metabolic costs, splicing large introns may also introduce conformational problems. A key step of intron splicing is the formation of the loop-like "lariat" structure as the recently cleaved 5′ end of the intron is attached to the branch point sequence close to the 3′ intron junction. Since a 100-Kb intron may extend out over 30 microns, its size may become a problem for the approximately 5 micron cell (Shepard et al. 2009). It has been proposed that especially large introns require different splicing mechanisms than standard introns and that these recursively splice the intron at a series of internal "ratcheting points" rather than in one piece (Hatton et al. 1998; Burnette et al. 2005; Shepard et al. 2009). It is as yet unclear to what extent this large intron ratcheting also occurs in fish.

Wagner (2005) discusses the cost of gene duplication in yeast in terms of extra energy expenditure from increased nucleotides transcribed and finds a significant cost to duplication in terms of extra transcription. We can therefore infer that there must also be a significant cost to large introns. It is possible that these large introns are recent recipients of extensive repetitive sequence expansions and selection has not had time to favor their reduction in size. Our analyses support this, revealing that greater than 70.61% of all large *D. rerio* intron sequence is repeat DNA, also reducing the number of introns greater than 50,000 bp to 426. It is possible that these remaining 426 introns also contain a portion of decayed repeats that cannot be recognized using the novel identification algorithms. Previous work has also focused on the effects of gene expression on intron length, finding that

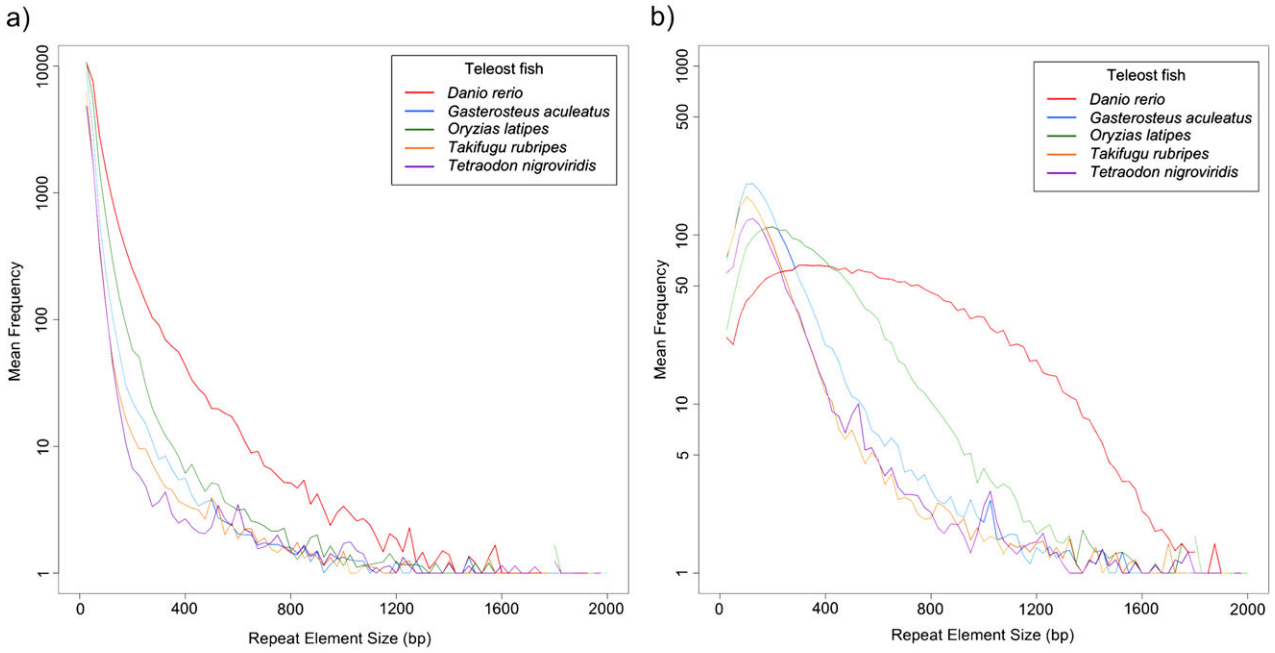

Fig. 2.—(*a*) A frequency distribution of individual repeat element sizes in introns between 500 and 2,000 bp in size. Each point represents the mean of intron sizes within a 25-bp sliding window. (*b*) Frequency distribution of cumulative repeat element size produced by pooling all repeat elements within individual introns.

introns appear to be smaller in more highly expressed genes and therefore must be under the influence of selective pressures to reduce them in size. This certainly seems to be the case in lower eukaryotes, which contain proportionally more intronless genes and shorter introns than in vertebrates, although the latter are less well sampled. Lower eukaryotes have larger effective population sizes and shorter reproductive cycles, as well as having less tolerance to environmental stress, which will likely impact the speed at which natural selection can process any deleterious traits. The role of transposable elements on intron length was also examined, and introns were found to be significantly shorter in more highly expressed genes, although selection against transposition overall remains unclear (Castillo-Davis et al. 2002; Jeffares et al. 2008). In future studies, we could examine the effects of gene expression and the involvement of specific genes and gene families on intron content in orthologous loci across a more widely sampled group of taxa, in order to further clarify these findings.

Small Introns as a Proxy for Annotation Quality

The minimum intron size reported in a previous Ensembl release (version 59) of *D. rerio* was zero nucleotides, with a further 882 introns less than 5 bp. The existence of 0 bp introns is a result of the way the Ensembl API identifies introns based on the exon coordinates. Given that intron splicing requires a "minimum" of five nucleotides (GU-AG plus an A for the branch point), these introns cannot be real and/or functional. In practice, both for steric requirements of intron bending during splicing and due to the need for other signal sequences, minimum intron sizes are likely to be larger (Schwartz et al. 2008). Certainly in yeast (*Saccharomyces cerevisiae*), there is a conserved 8-bp branch site that is typically 18 to 40 bp upstream of the 3′ splice site (Zhuang et al. 1989). This implied 30-bp minimum size in yeast may well be different from vertebrates where branch site sequences are not conserved but given that the branch point must still be displaced from the intron boundaries and a 3′ polypyrimidine tract interacting with the U2 snRNP auxiliary factor of the spliceosome is common (Zhuang et al. 1989; Adams et al. 1996) typical introns will be considerably larger. For all these reasons, we do not consider introns of 1–5 nucleotides to be biologically realistic. In *D. rerio*, the smallest intron for either U2 or U12 is 11 bp, whereas the other splice site category has 412 introns smaller than this. We suggest that since these introns have nonstandard splice signals and a different size range to standard introns they should be treated with caution until they are experimentally validated. Although we included all introns annotated by Ensembl in our analyses, small introns comprise less than 0.19% of all introns and do not influence our conclusions.

Danio rerio is widely considered to be a reasonably high-quality genome annotation, though it undoubtedly contains intron annotation errors, as indeed will all genomes. We note that the extreme intron size outliers in the *D. rerio* genome have changed considerably with releases 59–61 of Ensembl. Not only have the two zero-size introns been removed but also a 2-Mb intron that was previously the

largest. It is likely that automated intron annotation errors can particularly skew the extremes of the intron size distribution since these have relatively few members. As an example of an additional source of error in the annotation of genomic introns, we can envisage that if a gene was annotated by comparison to cDNA from a paralog containing a small coding indel or to a transcript that had spliced out a small exon, the extra sequence present in the genomic copy would likely be identified as intronic. Because these coding regions must necessarily be a multiple of 3 bp they will lead to a 3-bp size periodicity of any coding region misannotated as intronic and we would expect introns present in the CDS but not 5′-UTR or 3′-UTR to show such a periodicity. Supplementary figure S1 (Supplementary Material online) shows exactly this 3-bp pattern of periodicity for small introns between approximately 11 and 60 bp. This pattern was present in CDS introns but could not be detected in 5′-UTR or 3′-UTR introns. This indicates that CDS introns smaller than approximately 60 bp have a significant quantity of misannotated coding region.

U2 and U12 Introns

Given the difficulties of studying the interaction of the spliceosomes with identified introns, we have based our determination of U2 and U12 introns on the splicing signals they contain. Although this may contain errors because the U12 spliceosomes can interact with U2-type splicing signals (Lin et al. 2010), this is not the normal situation, and our error is likely to be very small. The frequencies of intron type are shown in supplementary table S5 (Supplementary Material online) and reveal that, as expected, the vast majority of introns are of the U2 type. For all fish except *D. rerio*, there are 13.9–17.4% of introns that we classify as other because they do not possess the classical splicing signals encountered with either U2- or U12-type introns. *Danio rerio*, the highest quality genome, has considerably fewer of these other introns (6.4%) and a similarly higher percentage of the major U2 type introns, suggesting that the other category is dominated by poorly annotated regions.

Conclusions

Understanding the diversity of genome variation using comparative genomics requires a bioinformatics approach that can be tailored and modified by the end user. We have developed a comparative genomics pipeline based on the well-tested and open-source code of the Perl Ensembl Core Software Libraries and BioPerl APIs (Stajich et al. 2002; Stabenau et al. 2004). Our analysis of the five currently available fish genomes indicates that although the intron content of these genomes is very similar in many respects, different genomic processes appear to be shaping the genomic intron content. The five fish differ not only in scale (number and total amount of intronic sequence) but also the frequency distribution of different intron size classes. The zebrafish *D. rerio* in particular does not have monotonically decreasing intron frequency with size from an approximately 80-bp mode, as the other fish appear to have, but rather has a second peak of introns in the 500- to 2,000-bp range. Repetitive DNA including transposable elements, satellites sequences, and simple repeats are known to be largely responsible for the differences in genome size between species that do not vary in ploidy (Neafsey and Palumbi 2003; Boulesteix et al. 2006; Hawkins et al. 2006; Bosco et al. 2007), and it is likely therefore that much noncoding DNA will have this origin, even if it has accumulated so many mutations that its previous repetitive nature can no longer be recognized. Our diverse approaches to characterizing repetitive elements in *D. rerio* introns revealed that approximately 47% of intronic sequence could be identified as repetitive. Repeating our analyses only with nonrepetitive intron sequences still revealed a unique size distribution for *D. rerio* introns, indicating that this has not been caused by a recent expansion of repetitive sequences, as these would have been readily recognizable as repetitive. Instead, we suggest that a more ancient expansion of repeats has created this intronic pattern and little signal of their repetitive origins still remains. A broader sampling of teleost genome sequences in a robust phylogenetic design would help to locate such an event and better clarify the origins of intron expansion across these lineages.

Supplementary Material

Supplementary figures S1–S2 and tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We are grateful to Dr. Chris D. Venditti, who helped with statistical methods of data analysis and provided input on the draft manuscript; to Dr. Casey Bergman, who advised on repeat sequence analysis; and to Dr. Andrew Davidson for providing input on the draft manuscript. The work was supported partly by Natural Environment Research Council (NER/S/A/2004/12984 to D.H.L. and D.A.J.) and The University of Hull.

Literature Cited

Adams MD, Rudner DZ, Rio DC. 1996. Biochemistry and regulation of pre-mRNA splicing. Curr Opin Cell Biol. 8:331–339.

Aparicio S, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297:1301–1310.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573–580.

Benton MJ, Donoghue PCJ. 2007. Paleontological evidence to date the tree of life. Mol Biol Evol. 24:26–53.

Bergman CM, Quesneville H. 2007. Discovering and detecting transposable elements in genome sequences. Brief Bioinform. 8:382–392.

Bosco G, Campbell P, Leiva-Neto JT, Markow TA. 2007. Analysis of Drosophila species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics 177:1277–1290.
Boulesteix M, Weiss M, Biémont C. 2006. Differences in genome size between closely related species: the Drosophila melanogaster species subgroup. Mol Biol Evol. 23:162–167.
Burnette JM, Miyamoto-Sato E, Schaub MA, Conklin J, Lopez AJ. 2005. Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. Genetics 170:661–674.
Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. Nat Genet. 31:415–418.
Chamary J-V, Hurst LD. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. Mol Biol Evol. 21:1014–1023.
Coghlan A, Wolfe KH. 2004. Origins of recently gained introns in Caenorhabditis. Proc Natl Acad Sci U S A. 101:11362–11367.
Deutsch M, Long M. 1999. Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res. 27:3219–3228.
Flicek P, et al. 2010. Ensembl 2011. Nucleic Acids Res. 39:D800–D806.
Frischmeyer PA, Dietz HC. 1999. Nonsense-mediated mRNA decay in health and disease. Hum Mol Genet. 8:1893–1900.
Gaffney DJ, Keightley PD. 2006. Genomic selective constraints in murid noncoding DNA. PLoS Genet. 2:e204.
Gazave E, Marqués-Bonet T, Fernando O, Charlesworth B, Navarro A. 2007. Patterns and rates of intron divergence between humans and chimpanzees. Genome Biol. 8:R21.
Haffter P, et al. 1996. The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. Development 123:1–36.
Hankeln T, Friedl H, Ebersberger I, Martin J, Schmidt ER. 1997. A variable intron distribution in globin genes of Chironomus: evidence for recent intron gain. Gene 205:151–160.
Hatton AR, Subramaniam V, Lopez AJ. 1998. Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon–exon junctions. Mol Cell. 2:787–796.
Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res. 16:1252–1261.
Hughes AL, Yeager M. 1997. Comparative evolutionary rates of introns and exons in murine rodents. J Mol Evol. 45:125–130.
Iwamoto M, Maekawa M, Saito A, Higo H, Higo K. 1998. Evolutionary relationship of plant catalase genes inferred from exon-intron structures: isozyme divergence after the separation of monocots and dicots. Theor Appl Genet. 97:9–19.
Jaillon O, et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431:946–957.
Jeffares DC, Penkett CJ, Bähler J. 2008. Rapidly regulated genes are intron poor. Trends Genet. 24:375–378.
Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.
Kim E, Magen A, Ast G. 2007. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 35:125–131.
Kimura T, et al. 2004. Large-scale isolation of ESTs from medaka embryos and its application to medaka developmental genetics. Mech Dev. 121:915–932.
Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
Li W, Tucker AE, Sung W, Thomas WK, Lynch M. 2009. Extensive, recent intron gains in Daphnia populations. Science 326:1260–1262.
Lin C-F, Mount SM, Jarmołowski A, Makałowski W. 2010. Evolutionary dynamics of U12-type spliceosomal introns. BMC Evol Biol. 10:47.
López-Bigas N, Audit B, Ouzounis C, Parra G, Guigó R. 2005. Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett. 579:1900–1903.
Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401–1404.
Majewski J, Ott J. 2002. Distribution and characterization of regulatory elements in the human genome. Genome Res. 12:1827–1836.
Marais G, Nouvellet P, Keightley PD, Charlesworth B. 2005. Intron size and exon evolution in Drosophila. Genetics 170:481–485.
Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome? Trends Genet. 23:183–191.
Mironov AA, Fickett JW, Gelfand MS. 1999. Frequent alternative splicing of human genes. Genome Res. 9:1288–1293.
Morgulis A, Gertz EM, Schäffer AA, Agarwala R. 2006. WindowMasker: window-based masker for sequenced genomes. Bioinformatics 22:134–141.
Neafsey DE, Palumbi SR. 2003. Genome size evolution in pufferfish: a comparative analysis of diodontid and tetraodontid pufferfish genomes. Genome Res. 13:821–830.
Roest Crollius H, Weissenbach J. 2005. Fish genomics and biology. Genome Res. 15:1675–1682.
Rogers J. 1989. How were introns inserted into nuclear genes? Trends Genet. 5:213–216.
Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet. 7:211–221.
Schwartz SH, et al. 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. 18:88–103.
Sela N, Kim E, Ast G. 2010. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. Genome Biol. 11:R59.
Sharp PA. 1985. On the origin of RNA splicing and introns. Cell 42:397–400.
Sharpton TJ, Neafsey DE, Galagan JE, Taylor JW. 2008. Mechanisms of intron gain and loss in Cryptococcus. Genome Biol. 9:R24.
Shepard S, McCreary M, Fedorov A. 2009. The peculiarities of large intron splicing in animals. PloS One. 4:e7853.
Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev. 9:657–663.
Smit AFA, Hubley R, Green P. 2011. RepeatMasker Open-3.3.0 [computer program]. Seattle (WA): Institute for Systems Biology [cited 2011 September 16]. Available from: http://www.repeatmasker.org
Stabenau A, et al. 2004. The Ensembl core software libraries. Genome Res. 14:929–933.
Stajich JE, Dietrich FS, Roy SW. 2007. Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. Genome Biol. 8: R223.
Stajich JE, et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12:1611–1618.
Stein LD, et al. 2003. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. PLoS Biol. 1:E45.
Venter JC, et al. 2001. The sequence of the human genome. Science 291:1304–1351.

Wagner A. 2005. Energy constraints on the evolution of gene expression. Mol Biol Evol. 22:1365–1374.

Yandell M, et al. 2006. Large-scale trends in the evolution of gene structures within 11 animal genomes. PLoS Comput Biol. 2:e15.

Zhu L, et al. 2009. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics. 10:47.

Zhuang YA, Goldstein AM, Weiner AM. 1989. UACUAAC is the preferred branch site for mammalian mRNA splicing. Proc Natl Acad Sci U S A. 86:2752–2756.

Associate editor: Greg Elgar