SCIENCE CHINA
Life Sciences

SPECIAL ISSUE: Non-coding RNAs
• RESEARCH PAPER •

October 2013 Vol.56 No.10: 968–974
doi: 10.1007/s11427-013-4540-y

Systematic analysis of intron size and abundance parameters in diverse lineages

WU JiaYan[1,2†], XIAO JingFa[1,2†], WANG LingPing[1,2,3], ZHONG Jun[1,2,3], YIN HongYan[1,2,3], WU ShuangXiu[1,2], ZHANG Zhang[1,2] & YU Jun[1,2*]

[1]*Key Lab of Genomics Science and Information, Chinese Academy of Sciences, Beijing 100101, China;*
[2]*Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China;*
[3]*University of Chinese Academy of Sciences, Beijing 100049, China*

Received July 20, 2013; accepted August 10, 2013; published online September 9, 2013

All eukaryotic genomes have genes with introns in variable sizes. As far as spliceosomal introns are concerned, there are at least three basic parameters to stratify introns across diverse eukaryotic taxa: size, number, and sequence context. The number parameter is highly variable in lower eukaryotes, especially among protozoan and fungal species, which ranges from less than 4% to 78% of the genes. Over greater evolutionary time scales, the number parameter undoubtedly increases as observed in higher plants and higher vertebrates, reaching greater than 12.5 exons per gene in average among mammalian genomes. The size parameter is more complex, where multiple modes appear at work. Aside from intronless genes, there are three other types of intron-containing genes: half-sized, minimal, and size-expandable introns. The half-sized introns have only been found in a limited number of genomes among protozoan and fungal lineages and the other two types are prevalent in all animal and plant genomes. Among the size-expandable introns, the sizes of plant introns are expansion-limited in that the large introns exceeding 1000 bp are fewer in numbers and transposon-free as compared to the large introns among animals, where the larger introns are filled with transposable elements and appear expansion-flexible, reaching several kilobasepairs (kbp) and even thousands of kbp in size. Most of the intron parameters can be studied as signatures of the specific splicing machineries of different eukaryotic lineages and are highly relevant to the regulation of gene expression and functionality. In particular, the transcription-splicing-export coupling of eukaryotic intron dispensing leads to a working hypothesis that all intron parameters are evolved to be efficient and function-related in processing and routing the spliced transcripts.

spliceosomal intron, intron length distribution, minimal intron, splice protein, lineage

Citation: Wu J Y, Xiao J F, Wang L P, et al. Systematic analysis of intron size and abundance parameters in diverse lineages. Sci China Life Sci, 2013, 56: 968–974, doi: 10.1007/s11427-013-4540-y

Eukaryotic genes have introns with variable size, number, and sequence context [1–4]. The number and size parameters by and large reflect the nature and efficiency of the intron splicing machinery of particular species or lineages [5–9]. The context parameters are most complicated, concerning the sequence content and context of nucleotide composition (such as GC and purine contents), transposable elements, and functional elements (such as splicing enhancers and the branch point) [10–12]. There are at least three possibilities for the existence and the absence of introns in eukaryotic genes: intronless (no intron), small introns in fixed sizes, and large introns in variable sizes. However, the rules of these intron parameters across diverse taxa have yet to be thoroughly summarized.

The spliceosomal machinery is very complex, containing different molecular complexes of proteins and RNAs, which are partitioned into both the nucleus and the cytosol [13,14].

†Contributed equally to this work
*Corresponding author (email: junyu@big.ac.cn)

Although the exact formations of these spliceosomes for their precise functionality and how these numerous cellular machineries are organized in the cellular compartments all remain to be illustrated, the relationship between introns and their processing machineries can be addressed by defining the variables of the substrates based on the knowledge of intron sequences as footprints or signatures of the machineries. For instance, the human minimal intron-containing genes are known to distribute differently on the chromosomes and are enriched in certain functions [6].

In this study, we investigate intron characteristics through diverse—within and across—lineages and try to understand some of the major rules of intron existence and possible functionality. Based on stratification of different parameters and vast literature on intron splicing mechanisms and processes, we classify all spliceosomal introns into three basic categories: half-sized, minimal, and size-expandable. We also develop a working hypothesis that states: intron size and number of a given species are evolutionarily selected to be optimal for the synthesis, splicing, and export of the intron-containing genes. Detailing the molecular mechanisms of intron processing, we propose that the fraction of minimal introns are optimized for transcript routing, where a maximum of 1/3 intron-containing transcripts are minimal intron-containing and to be diverted into a separate exporting route through the nuclear pores. Other speculations and proposals are also described to provide essential background knowledge for further discussions.

1 Materials and methods

All data were retrieved from Ensemble (https://www.ensemble.org) on April 29th, 2013. Sequence alignment was carried out by using BLAT algorithms. The BLAT results were processed to select for reliable alignments. We calculated the insert length between two adjacent hits of the query and the reference sequence and examined the intron pattern and orientation according to the consensus sequence of the intron. The alignment results were clustered based on splicing-site-sharing for multi-exon transcripts and exon-overlapping for single-exon transcripts [6]. When a locus had multiple alternatively spliced transcripts, the one with the greatest number of exons and/or length was selected as the representative. Introns length data were calculated from the clustered results containing the representative locus. We drew the density distribution plots based on the statistics of the datasets.

Alternative splicing (AS) isoforms of *Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana* and *Oryza sativa* were retrieved from AS-ALPS (alternative splicing-induced alteration of protein structure) database (version 1.8) [15]. Non-redundant datasets in all four species were accessed. All unique and sharing AS isoforms were counted according to the non-redundant dataset described by Venny diagram [16].

2 Results and discussion

2.1 Intron number and size distribution across and within lineages

To understand the intron number and size distribution across and within lineages, we collected genome sequences for all sequenced species and classified them based on various conventional lineages. Here we choose some of the results from a few representative species for selected lineage to display the different characteristics of intron number and size distribution (Figure 1).

Due to the massive scale of sequenced whole protozoan genomes and their highly variable intron number and size, we selected several species that have their intron numbers greater than 1000. The sizes of these introns scattered widely from 10^1 to 3.16×10^5 bp and with a variety of density peaks spreading from 48 to 501 bp. Although a majority of the protozoan species have only one major peak, there are still several genomes in the collection, whose intron sizes are bimodal, i.e., there are two peaks in the curves showing different intron size distributions. For instance, the genes of *Toxoplasma gondii* have intron size partitioned into two peaks and so do *Guillardia theta* and *Plasmodium chabaudi* (Figure 1A). It seems that the two peaks are either major vs. minor or similar in heights; the major peak can be ~10× higher in density than the minor. It is interesting that in some species the minor peaks sometimes are smaller in size than the major peaks but such a reverse order is never seen in some other species. In other words, the small minor peaks may be processed by unique machinery that is rather cryptic so that it cannot process too many introns at a time. Taken together, these results support the idea that there may be more than one splicing mechanism to cut-and-paste spliceosomal introns even in the primitive protozoan species.

The intron size distribution of fungal (Figure 1B) and invertebrate (Figure 1C) species is much narrower as compared with that of protozoan, higher plant and animal species, which ranges from 10^1 to 10^3 bp for fungi and from 10^1 to 3.16×10^3 bp for invertebrates. The intron distributions of most fungal species have two peaks; the smaller intron density peaks are always higher (more in numbers) than the larger peaks in contrast to the situation in protozoan introns where the opposite is true: introns in the smaller peaks are always less in numbers. Even in the case of *Y. lipolytica*, its intron size distribution, two bumps in 58 and 380 bp, is still the fungal type (Figure 1B).

There have been a large number of sequenced invertebrate genomes but their intron patterns are rather limited in diversity. We choose the two best annotated genomes, *Drosophila melanogaster* and *Caenorhabditis elegans* as representatives for an intron size analysis. Both species have two types of introns: minimal and size-expandable

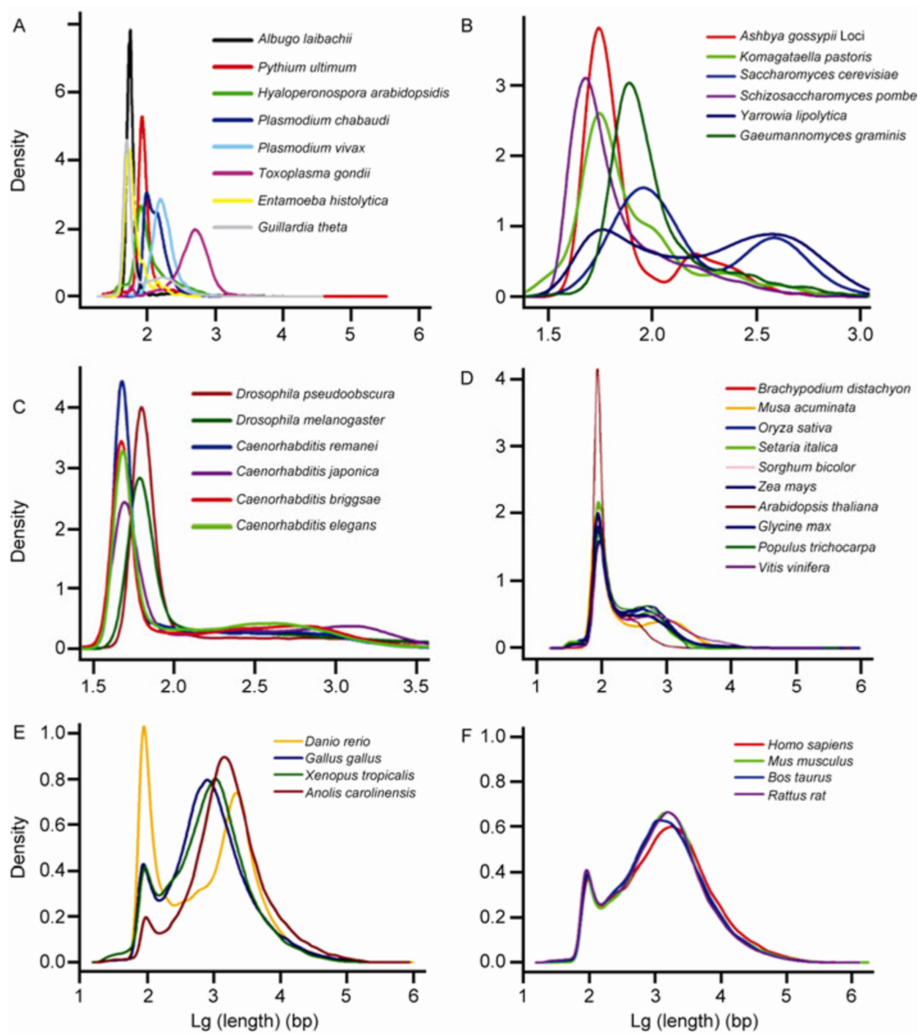

Figure 1 Lineage-specific intron size distribution. *X*- and *Y*-axes show intron length and the density of intron distribution, respectively. Intron sizes are plotted in logarithmic scales for representative species of protozoa (A), fungi (B), metazoa (C), plants (D), vertebrates without mammals (E), and mammals (F).

introns (Figure 1C). Although their minimal introns have slightly different size ranges, 48 bp for *Caenorhabditis* and 63 bp for *Drosophila*, the larger modes of their introns are actually very different (data not shown). If we extend the horizontal scale, the larger introns of the *Caenorhabditis* species always show as small bumps after the minimal intron peaks and such distributions indicate that their intron sizes are rather limited, whereas the larger introns of the *Drosophila* species are always smoothly extended into a larger size range.

The most conserved intron size parameters are observed in two more evolutionarily advanced lineages: plants and vertebrates (Figure 1D and F). Both dicotyledons and monocotyledons show a very similar pattern—a sharp minimal intron peak at ~10^2 bp and a larger peak that is much lower in density; the intron size distribution ranges from 16 to 10^3 bp. Other than those of the miniature model plant *A. thaliana*, most of the larger plant introns have a similar size distribution but the size is rather limited and it seldom goes beyond 10^4 bp. Similar topics have been discussed previously [17–19] and we propose that the size limit of plant introns reflects the nature of the plant splicing machinery that has constraints on how large an intron can be efficiently cut-and-pasted and ready for functioning. In terms of the number parameter, plant genomes have somewhat lower percentage of alternative splicing events when compared to those of human and other mammals [3,20].

The vertebrate introns are also unique in two ways, aside from its wider size distribution similar to those of insects (except the fact that insects have a limited number of larger introns) ranging from 16 to 2×10^6 bp. First, the vertebrate minimal introns are highly conserved in size throughout the lineage from lower vertebrates to mammals. Second, vertebrate introns have been constantly expanding in evolution to the extent that the mammals all have a great number of large introns, which are hundreds of kilobasepairs (kbp) in size (Figure 1F). Looking through evolutionary time scales, we find that the larger introns have been proportionally increasing over time but at the mean time the number of minimal introns per gene has been decreasing (also see Tables 1 and 2). For instance, *Danio rerio* has the highest minimal intron peak and the lowest larger intron peak as compared to four other vertebrates shown in the same figure. The smallest amount of minimal introns is found in *Anolis carolinesis*, while it is not surprising that *A. carolinesis* has the largest proportion of larger introns. The intron distributions among mammals are almost all identical and the densities of the minimal introns are always much lower than that of the larger introns. The norm of such a distribution is that among mammalian genomes ~30% of the total genes have minimal introns and ~10% of the total introns are minimal introns.

2.2 The existence of different splicing machineries for spliceosomal introns

The intron parameters, both number and size, are never stabilized among diverse taxa albeit seemingly stable within discrete lineages that are late in the evolutionary time scale, such as plants and vertebrates. Therefore, the rules of these parameters are inevitably useful for lineage-based analyses and biological interrogations. As far as size parameter is concerned, there are at least three types of transcripts in a typical genome: intronless, with minimal introns, and with large introns. The expandability of the large-intron mode is also part of the size parameter that varies among lineages, i.e., we observe higher degree of expandability in insects and vertebrates and it becomes significantly limited in Nematodes and plants. When number (as densities in the plots) parameter is considered, the most variable taxonomic groups are protozoan and fungal species, where the intron-containing genes can be either the majority of genes or a small fraction of them. The numbers of minimal and size-expandable introns are also variable within lineages, such as lower vertebrates vs. higher vertebrates, where the formal tend to have more minimal introns and the latter have more size-expandable introns than minimal introns.

In general, most of the lower eukaryotic introns are small and even half-sized as compared to the minimal introns of the higher eukaryotes. The number of intron per gene has also been increasing within the vertebrate lineages from fish to mammals. In addition, the ratios between intron of different size classes also change among taxa, as we have shown in Figure 1.

This and our previous studies have classified introns into four basic groups: half-sized, minimal, and size-expandable but expansion-limited, and size-expandable and expansion-flexible introns (Table 1). It is obvious that, aside from prokaryotes, all eukaryotic intron-containing genes fall into four basic categories that correspond to the four intron groups. First, three types of splicing variations—Splice-1, 2, and 3—have been found among protozoan and fungal species, despite that some of them may be incomplete and cryptic, such as in the case of parasitic organisms. Although large introns in a size of 1 kbp or more are present in the primitive eukaryotic genomes, they are not seen in large numbers. The result indicates that the protozoan and fungal splicing machineries may be simpler than what have been described in higher animals and plants. Second, the plant splicing machineries—Splice-2 and 3—are less complex. They are able to process a large number of introns but not lengthy introns as compared to the animal spliceosomal machineries. The plant introns are largely free of transposable elements that are massive, taking account of 50% (such as *Arabidopsis* and rice) to even 95% (such as barley and wheat) of the plant genomes [3,7,17,18]. Third, the animal spliceosomal machineries are diversified among lineages, such as Nematodes, insects (represented by *Drosophila* here), and vertebrates. Nematodes transcribe genes in a different way as compared to the rest of animal taxa, whose transcripts are in large arrays of genes and split subsequently, guided by special guiding RNAs [21,22]. Insects have large introns although not massive in numbers, which distinguish insects from Nematodes. The largest genes characterized thus far are the *Drosophila*'s fertility genes and most of them are Mbp in size [23,24]. These large introns are processed in a peculiar way that the introns are removed in a consecutive piece-meal fashion and aided by a particular sequence motif serving as a splice site. The process is called recursive splicing, starting from the 5′-end of the transcript and the sequence motif is called the "ratcheting point" or RP-sites in a consensus of $(Y)_n$NCAG|GUAAGU, where the splice junction is shown as a vertical bar [25–27]. In addition, since genome duplication has not been seen in both lineages, as well as in other animal lineages other than vertebrates [28], sophisticated splicing machinery provides an alternative way of generating novel genes. Fourth, vertebrate introns can be both large and more massive in numbers [3,9,10]. Therefore, the vertebrate splicing machinery must be different from that of *Drosophila* in its molecular details and evidence has been accumulating, where co-transcriptional efficiency is reported to be different between *Drosophila* and mouse [29].

Table 1 Different types of splicing machineries determine intron parameters as well as structures of genes and genomes

Lineage	Half-sized intron	Minimal intron	Size-expandable intron	
			Limited	Flexible
Protozoan	+[a)]	+[b)]	+[c)]	
Plants		+	+	
Nematodes		+	+	
Drosophila		+	+	+[d)]
Vertebrates		+	+	+

a) Splicing-1, half-sized introns, ~25–50 bp. b) Splicing-2, minimal introns, ~100 bp. c) Splicing-3, size-expandable but expansion-limited introns, >200 but <4000 bp. d) Splicing-4, size-expandable and expansion-flexible >3000 to 30 kbp.

2.3 The unique distributions of minimal introns and their possible functional role

The minimal intron group is one of the two universal intron types other than size-expandable introns. It not only has a size constraint but also often large in numbers; both are highly variable across lineages and conservative within lineages. The size constraint suggests proteinaceous nature of the machinery where the conservation of its physical dimensions varies across lineages. To examine the specific distribution of the minimal intron among genes between vertebrate and plant species, we carry out a broad survey and list the percentages of minimal introns (the size of the minimal intron is defined as <150 bp) out of the total introns and the fraction of minimal-intron-containing genes from eight species, including two mammals, four non-mammal vertebrates, and two plants (Table 2).

The most size-conservative minimal introns in both size and fraction belong to mammals. While 1/4–1/3 of their genes contain minimal introns, the total numbers of minimal introns for all mammalian species are ~10%. In addition, the vertebrate lineage also appears to maintain a significant fraction of minimal-intron-containing genes as seen in *A. carolinensis*, whose genome has 5.57% minimal introns of the total and 30.52% genes containing minimal introns. The case for a lower vertebrate, such as zebrafish shown here, is slightly different due to its large amount of small introns while the minimal introns do not distribute randomly with higher fraction exceeding our expectation. To emphasize the uniqueness of the minimal intron distribution of vertebrates, we also list the minimal intron distributions from the two model plants, *Arabidopsis* and rice, where the fractions of their minimal introns are almost broken even among all genes. We have gone one step further to show that there are large amount of alternatively spliced transcripts in higher plants and animals, based on the AS-ALPS database collection [15] (Figure 2). Aside from the large amount of shared transcripts even between animal and plant genes, there are also over 10000 splicing variants unique to humans. The lower number in mouse is unusual and it may be due to sampling biases and incomplete nature of the database. The numbers of splicing variants between *Arabidopsis* and rice are rather comparable as rice has nearly twice as many

Table 2 Minimal introns and the minimal-intron-containing genes from selected plant and vertebrate species

Species	Minimal intron (%)	Gene (%)
H. sapiens	10.58	32.08
M. musculus	10.35	25.62
G. gallus	12.65	44.83
A. carolinensis	5.57	30.52
X. tropicalis	13.79	51.61
D. rerio	25.74	60.28
A. thaliana	72.29	57.08
O. sativa	47.18	51.82

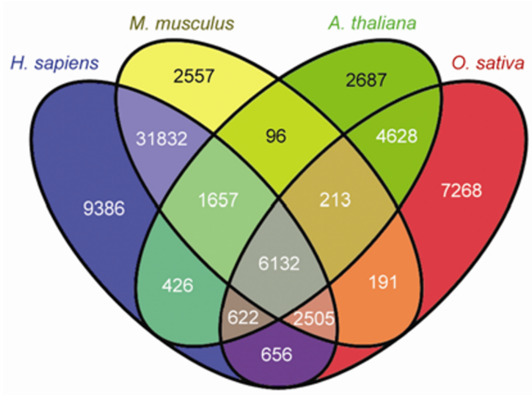

Figure 2 Alternatively-spliced transcript-encoded proteins classified among *H. sapiens*, *M. musculus*, *A. thaliana*, and *O. sativa*.

genes as *Arabidopsis* due to an extra genome-wide duplication [30–32]. The important point here is the fact that the average size of the rice genes has been expending to twice as that of the *Arabidopsis* and at the mean time the percentage of the rice minimal introns is also decreasing to 48.18% as compared to 72.29% [30].

Both size and number parameters of minimal introns appear unique in several ways. First, the size limit (~100 bp) of minimal introns is relatively indispensible in a way that it appears unique to lineages, especially in the cases of vertebrates and plants. It demonstrates that their splicing machineries as complexes of proteins and RNAs are highly conserved to the extent that most the proteinaceous and RNA components and their overall physical dimensions remain nearly unchanged for billions of years. Second, the number limit (>1/3 of the total genes) of minimal introns is evitable. Although in some lineages the smaller introns are more prevalent than the larger introns, minimal introns appear holding their ground tightly, especially when the size increase of the larger introns is intensive (such as the case of mammalian genomes); such an increase reduces both the number of minimal introns in the genomes and the number of minimal-intron-containing genes out of the total number of introns. The trend is clear and firm over the vertebrate lineage. Third, minimal-intron-containing genes are a unique group of genes in both functionality and genomic characteristics [6,8,33,34]. We have previously showed that genes with minimal introns tend to play certain housekeeping roles and are enriched on certain chromosomal territories [35]. They are also abundant (~10% minimal introns in ~30% genes), larger in size, which tend to be universally expressed as compared to genes with only larger introns and intronless genes and preferentially to locate toward the 3′ end of the transcripts. We have further pointed out that genes with minimal introns replicate earlier and preferentially reside in the vicinities of the open chromatins and that they occupy unique nuclear positions relevant to the regulation of transcription regulation and transcript export [6]. Therefore, we proposed a Routing Hypothesis for the functionality of minimal introns in the coupled process of transcription-splicing-export.

2.4 The Routing Hypothesis of intron processing

Pre-mRNA processing happens primarily in the nucleus. The related fields have now agreed upon a unified picture [36] and the spatiotemporally organized movements of cellular components within the nucleus, between the nucleus and the cytosol, and among the subcellular organelles and structures are of essence for the understanding of cellular physiology. In such a context, minimal introns are most relevant to the nuclear architectures and their dynamics [6,37].

Since the other two classes of introns, half-sized and size-expandable, are structurally non-uniformed, the most striking feature of minimal introns is the fact that they are very much fixed in physical dimensions, and export timing should be prioritized since speed is a function of both time and distance (Figure 3). In a spatial sense, the two splicing complexes for minimal introns (Figure 3B), one for each exon-intron junction, can be physically interacting, forming a much larger complex than that for the size-expandable introns. Such larger complex can be easily distinguished from the smaller "monomer" splicing complex (Figure 3A and C). In a temporal sense, the larger complexes can be recognized earlier and more precisely than the size-expandable intron complex that can be extremely large with the intron sequence, and the timing for processing them is unpredictable and lengthy as transcription-couple DNA repair mechanism may be at work that interferes and aborts transcription from time to time. Therefore, a working Routing Hypothesis assumes that the minimal-intron processing

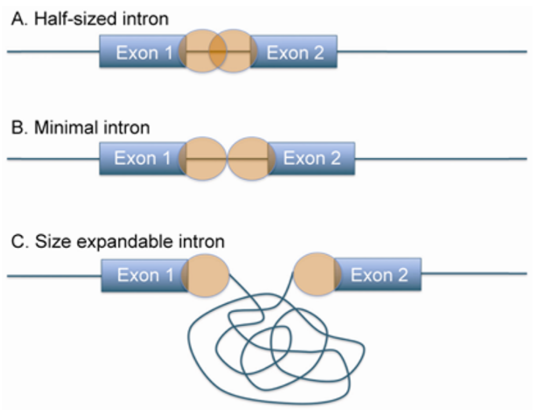

Figure 3 A schematic illustration of the three types of introns and their processing models. A, Half-sized introns (the thin blue lines) are assumed to be processed by either a single protein-RNA complex (splicing machinery, semitransparent brown circle) or a primitive one that is cryptic incapable of processing larger and large number of introns. B, Minimal introns are assumed to be processed by a dimeric complex that has a nominal size specific to species or lineages. C, Size-expandable introns are processed by two distinct and separate complexes that each recognizes one side of the intron flanking by the two exons (blue bars) to be joined.

complex serves as a "landmark" for the processing machinery to find the transcript in time, making sure that these large transcripts can be processed and routed differently from intronless transcripts and transcripts without minimal introns (such as those with half-sized and larger introns). When genes and their introns enlarge over evolutionary time scale, such as in the case of mammalian genes, minimal introns are selected to be more conserved and clustered toward the 3′-end of the minimal-intron-containing transcripts for better discrimination at the late stage of transcription and the early stage of routing and export.

The evidence supports the Routing Hypothesis is ample albeit mostly deemed indirect by molecular biologists. First, comparative studies of splicing machineries have pointed out the tempo-spatial features of the different types of introns that all have unique characteristics as we have discussed above in this article and previously elsewhere [6]. Second, evolutionary studies have proven that minimal introns are selected through negative selections on minimal-intron-containing genes and have special sequence contexts in human populations [8,11]. The original design of these experiments is to investigate whether the intron size constraint of human minimal introns is selected in a human population but it has also revealed sequence context relevance [38,11]. It is not difficult to extend the conclusion in any other population data beyond mammals. Third, a large body of evidence is attributable to the unique features and functional relevance of minimal-intron-containing genes [6]. For instance, the positioning of highly abundant minimal introns in the human genome is uniquely concentrated at the 3′ end of the genes. Minimal-intron-containing genes are larger than the gene size average and distributed non-randomly, enriched on certain chromosomes, and they are replicated earlier and tend to reside on open chromatin structures. The functions of these genes are also unique: they are mostly playing house-keeping roles and involved in phosphorylation in function and cellular structures (such as cytoskeletons) and trafficking in operation [6]. Nevertheless, it is still valuable to have experimental evidence in fields of biochemistry and cell biology since most of the other lines of evidence are either statistical in nature or considered as indirect, even though any direct evidence demonstrated based on a single gene or artificial gene construct is also hard to be generalized.

3 Conclusion

Introns are one of the major structural elements for eukaryotic genes. It must have functional roles, perhaps very limited catalytic but unlimited operational, compartmental, and homeostatic [28]. On the operational framework, for instance, introns can be the overwhelming majority of a genome in space, locate discretely in chromosomal territories, and govern both replication and transcription timing. The larger the introns are, the longer they take to be transcribed. Such a fine-tuning in time and space is perfect for development and organogenesis, especially for those species and lineages that are built in very complex and precise ways, such as humans and mammals. Introns are also a serious burden for cellular trafficking, from chromosomal territories to nuclear pores and from the nucleus to the cytosol. In terms of homeostasis, introns and their degraded fragments and nucleotides all have to flow in and out the nucleus, consuming energy and requiring intimate regulation of membrane potentials. After all, we are just in a process of understanding cellular details at the single cell level with molecular resolutions. The parameters of genes and genomes are much higher in priority for information scrutiny that not only calls for basic information, such as protein-coding sequences and their functionally relevant variations, but also for extended information, such as the conservation of their structural elements, introns and their sequence contexts. We are therefore always at the dawn of knowledge acquisition and eager for novel ideas and hypotheses.

This work was supported by the National Natural Science Foundation of China (31101063, 31271386) and National Basic Research Program of China (2010CB126604, 2011CB944100, 2011CB944101).

1 Berget S M, Moore C, Sharp P A. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci USA, 1977, 74: 3171–3175
2 Chow L T, Gelinas R E, Broker T R, et al. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell, 1977, 12: 1–8
3 Yu J, Wong G K S, Wang J, et al. Shotgun sequencing. In: Encyclopedia of Molecular Cell Biology and Molecular Medicine. 2nd ed. Wiley-VCH, 2005, 13: 71–114
4 Zhang Z, Wong G K S, Yu J. Protein coding. In: eLS. Chichester: John Wiley & Sons Ltd., 2013
5 Hong X, Scofield D G, Lynch M. Intron size, abundance, and distribution within untranslated regions of genes. Mol Biol Evol, 2006, 23: 2392–2404
6 Zhu J, He F, Wang D, et al. A novel role for minimal introns: routing mRNAs to the cytosol. PLoS ONE, 2010, 5: e10144
7 Wang J, Li S T, Zhang Y, et al. Vertebrate gene predictions and the problem of large genes. Nat Rev Genet, 2003, 4: 741–749
8 Yu J, Yang Z, Kibukawa M, et al. Minimal introns are not "junk". Genome Res, 2002, 12: 1185–1189
9 Wong G K S, Passey D, Yu J. Most of the human genome is transcribed. Genome Res, 2001, 11: 1975–1977
10 Wong G K S, Passey D, Huang Y, et al. Is "junk" DNA mostly intron DNA? Genome Res, 2000, 10: 1672–1678
11 Wang D, Yu J. Both size and GC-content of minimal introns are selected in human population. PLoS ONE, 2011, 6: e17945
12 Wong G K S, Wang J, Passey D, et al. Codon-usage gradients in Gramineae genomes. Genome Res, 2002, 12: 851–856
13 Jamison S F, Crow A, Garcia-Blanco M A. The spliceosome assembly pathway in mammalian extracts. Mol Cell Bio, 1992, 12: 4279–4287
14 Pessa H K, Will C L, Meng X, et al. Minor spliceosome components are predominantly localized in the nucleus. Proc Nat Acad Sci USA, 2008, 105: 8655–8660
15 Shionyu M, Yamaguchi A, Shinoda K, et al. AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. Nucleic Acids Res, 2009, 37: D305–D309
16 Oliveros J C. VENNY. An interactive tool for comparing lists with Venn Diagrams. 2007. http://bioinfogp.cnb.csic.es/tools/venny/index.html
17 Wendel J F, Cronn R C, Alvarez I, et al. Intron size and genome size in plants. Mol Biol Evol, 2002, 19: 2346–2352
18 Hawkins J S, Kim H, Nason J D, et al. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res, 2006, 16: 1252–1261
19 Reddy A S. Alternative splicing of pre-messenger RNAs in plants in the genomic era. Ann Rev Plant Biol, 2007, 58: 267–294
20 Naeem H, Kalyna S, Marquez Y, et al. Alternative splicing in plants—coming of age. Trend Plant Sci, 2012, 17: 616–623
21 Nilsen T W. Trans-splicing of nematode premessenger RNA. Ann Rev Micro, 1993, 47: 413–440
22 Zemann A, Bekke A, Kiefmann M, et al. Evolution of small nucleolar RNAs in nematodes. Nucleic Acids Res, 2006, 34: 2676–2685
23 Kurek R, Reugels A M, Lammermann U, et al. Molecular aspects of intron evolution in dynein encoding mega-genes on the heterochromatic Y chromosome of *Drosophila* sp. Genetica, 2000, 109: 113–123
24 Reugels A M, Kurek R, Lammermann U, et al. Mega-introns in the dynein gene DhDhc7(Y) on the heterochromatic Y chromosome give rise to the giant threads loops in primary spermatocytes of *Drosophila* hydei. Genetics, 2000, 154: 759–769
25 Lopez A J. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. Ann Rev Genet, 1998, 32: 279–305
26 Hatton A R, Subramaniam V, Lopez A J. Generation of alternative ultrabithorax isoforms and stepwise removal of large intron by re-splicing at exon exon junctions. Mol Cell, 1998, 2: 787–796
27 Burnette J M, Miyamoto-Sato E, Schaub M A, et al. Subdivision of large introns in *Drosophila* by recursive splicing at nonexonic elements. Genetics, 2005, 170: 661–674
28 Yu J. Life on two tracks. Genomic Prot Bioinf, 2012, 10: 123–126
29 Khodor Y L, Menet J S, Tolan M, et al. Cotranscriptional splicing efficiency differs dramatically between *Drosophila* and mouse. RNA, 2012, 18: 2174–2186
30 Yu J, Hu S, Wang J, et al. A draft sequence assembly of the rice (*Oryza sativa* ssp. indica) genome. Science, 2002, 296: 79–93
31 Yu J, Wang J, Lin W, et al. The genomes of *Oryza sativa*: a history of duplications rice genomes. PLoS Biol, 2005, 3: e38
32 Wang J, Zhang J, Li R, et al. Evolutionary transients in the rice transcriptome. Genomic Prot Bioinf, 2010, 8: 223–228
33 Hong X, Scofield D G, Lynch M. Intron size, abundance, and distribution within untranslated regions of genes. Mol Biol Evol, 2006, 23: 2392–2404
34 Havlioglu N, Wang J, Fushimi K, et al. An intronic signal for alternative splicing in the human genome. PLoS ONE, 2007, 2: e1246
35 Cremer T, Cremer M, Dietzel S, et al. Chromosome territories—a functional nuclear landscape. Curr Opin Cell Biol, 2006, 18: 307–316
36 Darnell J E Jr. Reflections on the history of pre-mRNA processing and highlights of current knowledge: a unified picture. RNA, 2013, 19: 443–460
37 Caudron-Herger M, Rippe K. Nuclear architecture by RNA. Curr Opin Genet Dev, 2012, 22: 179–187
38 Yu J. Challenges to the common dogma. Genomic Prot Bioinf, 2012, 10: 55–57