J. Plant Biochem. Biotechnol. (April–June 2016) 25(2):142–146
DOI 10.1007/s13562-015-0319-5

ORIGINAL ARTICLE

Intron gain, a dominant evolutionary process supporting high levels of gene expression in rice

Rupesh K. Deshmukh[1,2] · Humira Sonah[2] · Nagendra K. Singh[1]

Received: 4 October 2014 / Accepted: 25 May 2015 / Published online: 7 July 2015

Abstract Presence of introns in eukaryotic genes, their evolution and biological function has been a subject of considerable debate ever since their discovery in 1977. To understand the effect of number of introns on the structural and functional characteristics of rice genes, we carried out whole genome analysis of the relationship of the number of introns per gene with predicted cDNA sequence (CDS) length, average exon length and gene expression patterns. There was a direct correlation between the number of introns and the average CDS length among the expressed rice genes, as determined by expressed sequence tags (EST) support. The percentage of expressed genes in groups of rice genes representing different intron numbers showed a significant positive correlation with the number of introns providing evidence for higher level of expression for intron-rich genes. This was further supported by higher abundance of ESTs for the intron-rich genes in the rice EST database. Higher number of introns may be providing post-transcriptional stability to the mRNA leading to higher expression levels. Here we first report the detailed genome wide analysis of distribution pattern of introns in rice that provides important insight in to understanding the evolution, structure and expression of genes in plant species. Particularly, the complex gene structure and functional advantage of the intron containing genes supports the gain of intron theory for the evolution of eukaryotic genes.

Keywords Intron gain · Intron number · Gene expression · Exon · CDS · Rice genome

Abbreviations

CDS	Coding DNA sequence
EST	Expressed sequence tags

Introduction

Presence of introns is a characteristic feature of the eukaryotic genes, although genes with introns have also been reported with very low frequency in bacteriophages and bacterial genomes (Edgell et al. 2000). On the other hand, many eukaryotic genes lack introns. The number of introns per gene varies drastically among the eukaryotes. Most vertebrates have several introns per gene whereas only two characterized introns have been found in *Giardia lamblia* a flagellated protozoan parasite (Aparicio et al. 2002; Nixon et al. 2002). There is no simple phylogenetic pattern of intron-rich and intron-poor species across the eukaryotic tree (Jeffares et al. 2006). Such differences largely reflect stronger genomic streamlining in unicellular organisms than in multicellular species (Gilbert 1987). There is differential efficiency of intron selection in species with different population sizes, but none of the models predict high intron densities in early eukaryotes and unicellular species (Lynch and Conery 2003; Roy and Gilbert 2005; Loftus et al. 2005). Thus, intron number is controlled by sensitive natural selection that implies an important role for

Electronic supplementary material The online version of this article (doi:10.1007/s13562-015-0319-5) contains supplementary material, which is available to authorized users.

✉ Rupesh K. Deshmukh
rupesh0deshmukh@gmail.com

✉ Nagendra K. Singh
nksingh@nrcpb.org

[1] National Research Centre on Plant Biotechnology, Indian Agricultural Research Institute, New Delhi, India

[2] Division of Plant Science, University of Missouri, Columbia, 44 Agriculture Building, Columbia, MO, USA

mutational mechanisms of intron gain and loss. In general intron sequences are under low selection pressure than exons, consequently the introns have a higher rate of gain and loss than exons (Lin et al. 2006).

Presence of introns before the divergence of prokaryotes and eukaryotes is yet to be confirmed, but there are definite instances of both loss and gain of introns later in the evolution of species (Jeffares et al. 2006). There is no conclusive theory as yet on the mechanisms and forces that underlie gain and loss of introns, but the evolution of splicesomal introns has broad implications for many fundamental evolutionary questions. Presently there are two opposing views on the origin of introns. The intron-early hypothesis suggests that the introns were present in the genes of common ancestor of all the presently living organisms and splicing mechanism is ancient one (Gilbert et al. 1997). The intron-late theory advocates the view that the introns were inserted into their present location in the genes and the splicing mechanism has evolved late in eukaryotes due to its selective advantage (Rzhetsky and Ayala 1999). Research on mechanisms and causes of splicesomal intron evolution has been very active in the past few years in the post genomic era, resolving some old controversies and sparking some new ones (Roy and Gilbert 2006).

The completion of high quality rice genome sequence has provided an excellent opportunity to study the evolution of introns, possible mechanisms involved in loss and gain of introns, distribution of introns in the genes and their relation with other structural and functional features of the genes. The aim of present study was to particularly analyze the relationship between intron number and gene expression in rice.

Materials and methods

The complete set of 62,820 CDS sequences of predicted rice genes distributed over the twelve rice chromosomes was downloaded in batches from the TIGR built 4.0 database (http://www.tigr.org/tdb/e2k1/osa1/overview.shtml). A subset of 27,330 expressed genes of these were extracted manually. For the evidence of expression, ESTs and full-length cDNA information was used (Kikuchi et al. 2003). The number of introns per gene for all the predicted rice genes, including both expressed and non-expressed categories were predicted individually employing a semi-automated procedure using GffUtils tools (https://pythonhosted.org/gffutils/). The outputs were arranged chromosome wise and grouped in to classes based on the number of introns per gene. The average CDS length, average exon length and percent of expressed genes were calculated for each group of expressed rice genes and plotted against the number of introns contained in these genes.

First 200 genes (started from short-arm terminal of chromosome) of 0 to 30 intron categories of expressed genes were extracted manually and all the ESTs and cDNA, for the individual genes were downloaded in batches from the TIGR site. The average number of ESTs for each of genes possessing the 0 to 20 intron was categorized separately; however, the average number of ESTs for 21 to 30 intron genes was calculated together because there were less number of genes with high intron numbers in these categories. The number of ESTs was plotted against the number of introns in these genes. As some rice genes have multiple splice forms, average of all the alternative isoforms were considered for the analysis. Same set of 200 genes under each category was also used for the functional classification of genes. The predicted function information for each rice genes was extracted from the TIGR website and classified as per the Plant GOSlim ontologies (http://www.geneontology.org/).

Results and discussion

Intron frequency in the rice genes

The 62,820 predicted rice genes contained total 2,29,556 introns of which 1,36,792 introns were present in 27,330 expressed genes with cDNA evidence. In the rice genome there is preponderance of genes with no introns or fewer number of introns as against the popular perception that most eukaryotic genes have introns (Fig. 1). About 60 % of the rice genes have less than 4 introns per gene and only ninety one genes had more than 30 introns per gene (Fig. 1, Table S1). The genes with higher number of introns coded for structural proteins and high molecular weight protein for instance, kinesin motor domain containing protein (316.6kD, 37 introns), HEAT repeat family protein (250kD, 53 introns). Genes with protein binding function were the largest class (8.94 %) of intron-rich gene followed by those with hydrolase activity (7.82 %) and motor activity (6.15 %) (Fig. S1). Response to abiotic stimulus and secondary metabolic process is mainly coded by the intronless genes. Genes for transferase activity was a major class of intronless genes followed by

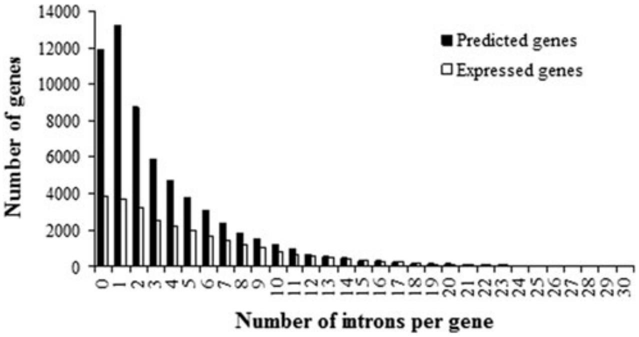

Fig. 1 Frequency of genes with different number of introns per gene in the rice genome. Expressed genes are those supported by cDNA match in dbEST

transcription factor activity (3.93 %) (Fig.S1). Mitochondrion, response to endogenous stimulus, kinase activity and catalytic activity genes showed diverse number of intron (Fig.S1).

Relationship of intron frequency with exon length and CDS length in expressed rice genes

The average exon length showed an overall negative correlation (r=−0.377) with the number of introns in a gene. As the number of intron increased the average exon size fell down initially but there was no further reduction in the average exon size in the genes with more than ten introns (Fig. 2). The correlation coefficient between the number introns and average exon size was much higher (r=−0.79) up to ten intron per gene. Exons of less than 50 bp are too short for the splicesome to operate and exons that are too long (greater than 300 bp) are difficult to locate by the splicesome (14). This may be the reason for genes with larger exon size having very few or no introns to avoid difficulty in splicing. Intron early theory hypothesizes that the very first genes and exons represented small polypeptide chains 15–20 amino acids and then large genes evolved by fusion of these smaller genes (Gilbert et al. 1997). But according to our results, Average exon size in the expressed rice genes is 215 bp. It is assumed by the intron-early theory that the exons of today are the results of, on an average, two to three acts of fusion from the original 15–20 amino acids long exons (Gilbert et al. 1997).

Average CDS length for the expressed rice genes was 1292 bp, with significant variation between genes. There was a strong positive correlation (r=0.89) between intron number and CDS length of the expressed genes (Fig. 3). As there was a lower limit to the average size of exons, increasing number of intron resulted in increase in the length of CDS.

Relationship between intron frequency and gene expression

Present analysis revealed a strong positive correlation between the number of introns per gene and percent of expressed genes in that category. The percentage of expressed genes increased with increasing number of introns in the genes up to about 13 introns after which it stabilized at about 80 % (Fig. 4). This is more than two fold higher than 32.74 % observed for intron less genes and 28.13 % for genes with single intron. This finding supports the hypothesis that more introns per gene leads to higher gene expression.

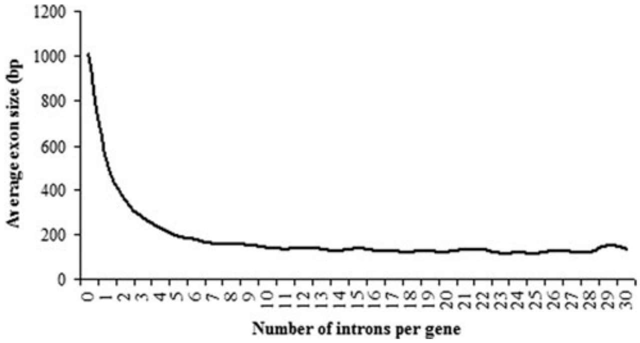

Fig. 2 Relationship between exon size and number of introns per gene based on 27,330 expressed genes in the rice genome

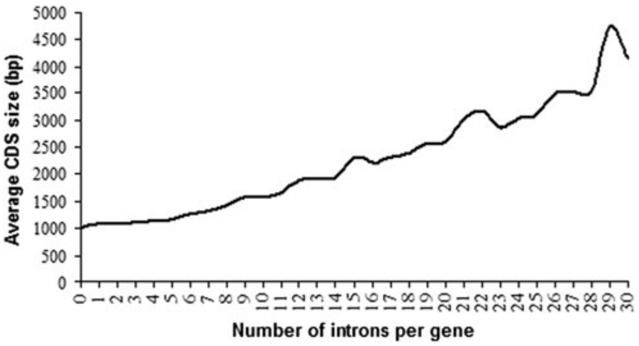

Fig. 3 Relationship between average size of the predicted coding sequence (CDS) and number of introns per gene in 27,330 expressed genes in the rice genome

To make a quantitative estimation of the relationship between level of expression and number of introns per gene, we performed correlation analysis between average number of ESTs representing particular gene locus and the number of introns in that gene. A significant positive correlation (r= 0.45) was observed between the average number of ESTs for particular gene and the average number of introns in the gene. The rice EST database was the second largest behind human with total 1,220,261 entries on April 25th 2008 in NCBI (Release 042508) and an average of 102 ESTs per gene loci analyzed in the present study. The average number of EST matches increased with increasing number of introns per gene but not in a strictly linear pattern (Fig. 5). The category of genes with 15 introns per gene was most highly expressed with an average of 236 ESTs matches per gene in the database. In contrast, intronless genes were very low in expression, with an average of only 32 ESTs per locus. These observations are based on a valid assumption that genes having low expression

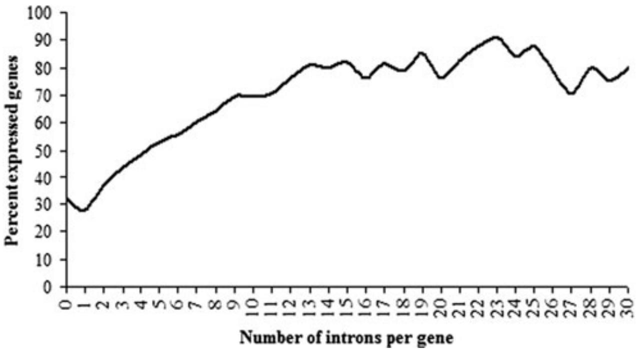

Fig. 4 Relationship between percentage of expressed genes and number of introns per gene in the rice genome. Percentage of expressed genes was calculated by dividing number of genes with EST support by total number of predicted genes in that category

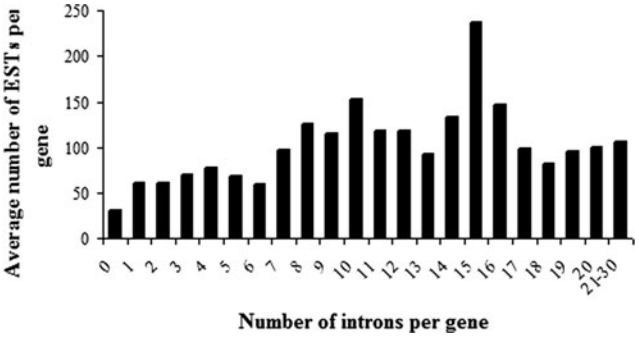

Fig. 5 Average number of EST matches in the TIGR database (give web address) for each category of genes with different number of introns. Average number of ESTs was calculated for the first 200 genes identified in each category of genes with 0 to 30. Category 21 to 30 have less than 200 genes each hence average was calculated on combined frequency

level or very specific expression conditions will accumulate less in the EST database. Recently, the abundance within the EST database method has been proposed for estimation of expression levels (Marais and Piganeau 2002). It has been reported that highly expressed genes have more and longer introns in rice and *Arabidopsis* which is consistent with our results (Ren et al. 2006). Presence of intron is known to enhance the gene expression level in transgenic plants also, and examples of intraspecific intron presence/absence polymorphism also supports role of introns in enhanced expression level (Llopart et al. 2002). The enhancement of gene expression level using intron sequence has been verified in wide range of plant species including monocots and dicots (Morita et al. 2012; Patil et al. 2010). Recently, intron of the *Gmubi* gene found to be contribute to very high levels of expression in soybean transiently and stably transformed tissues (Carola and Finer 2015).

In humans and C*aenorhabditis elegans,* the highly expressed genes have fewer and shorter introns. This compact nature of highly expressed genes is explained by transcriptional efficiency, regional mutation bias or genomic design (Vinogradov 2005; Urrutia and Hurst 2003; Sanderson et al. 2004). This hypothesis however does not fit well with the observations in plants. Whatever selection was responsible for the presence of more introns in the expressed genes as compared to the non-expressed rice genes might be due to divergence of animals and plant about 1600 million year ago (Sanderson et al. 2004).

The theory of intron gain with joining of adjacent exons can be evaluated by analyzing open reading frame across all exons of the gene. Consequently, the phase information of all the rice exons was retrieved in gff3 format from phytozome database (ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v8.0/Osativa/). The phase indicates whether the exon started with reference to the reading frame or not. On the basis of exon start with nucleotide number in a codon, the exons are classified as phase 0, 1, and 2. The phase analysis revealed over dominance of exons started with proper reading frame in rice genome. More particularly, 60.15 % phase 0 exons and only 33.28 % exons either phase 1 or phase 2 were observed. However, 6. 5 % of total exons were totally non-coding and represent only untranslated regions. The results of exon phase analysis strongly support the theory of intron gain with joining of adjacent exons.

Intron gain also occurs by the insertion of transposable elements into the existing coding sequences leading to progressively lower frequencies of genes with more introns (Ren et al. 2006). In contrast, loss of introns during evolution will lead to accumulation of genes with fewer or no introns. There are two models for intron loss in genes, the classical model, in which genomic copy of a gene undergoes gene conversion by double recombination with a reverse-transcribed copy leading to loss of one or more adjacent introns or creation of new genes by reverse transcription of mRNA (intronless) followed by insertion of this cDNA into the genome (retrotransposons like mechanism), and second is genomic deletion model in which introns could be lost by (near) exact genomic deletion (Hu 2006). The two models make several distinct predictions. First, recombination with RT-mRNAs should excise introns exactly, whereas genomic deletion should be less tidy, sometimes deleting adjacent coding sequences and leaving residual intron sequence (Roy and Gilbert 2006). Formation of new gene copy as per classical model was rare in *Drosophila* (Betrán et al. 2002). But we assume this mechanism might be predominant in plants, as retrotransposons are particularly abundant in plants, where they are often a principal component of nuclear DNA. For instance, in wheat about 90 % of the genome is made up of retrotransposons, whereas it is 49–78 % in maize (Li et al. 2004; Sanmiguel and Bennetzen 1998).

There would be a possibility of formation of defective copies of genes along with the functional copy during the intron loss. It has been reported that evolutionarily conserved genes in rice have more number of introns than newly evolved genes (25). Further, the results agreed with our presumption about the newly evolved intronless or genes with few introns being faulty. While loss of introns theory can explain formation of defective genes there is no evidence of any selective advantage for such genes during evolution of eukaryotic genes. Particularly, our results point more towards the selective advantage of intron-rich genes due to their high expression level.

Intron features like length and abundance in gene have been routinely used to correlate with evolution of gene family (Patil and Nicander 2013; Deshmukh et al. 2013). The gene families mostly expands with segmental or whole genome duplications. Many other mode of gene duplication are also known which involves reverse transcription of RNA, horizontal gene transfer and uneven recombination events. All these different mechanisms enforced different level of selection pressure on introns. Xu et al. (2012) have analyzed 612 pairs of sibling paralogs from seven representative gene families and 300 pairs of one-to-one orthologs from different species and their results suggested that the structural divergences have a more important role during the evolution of duplicate than non-duplicate genes.

Despite increasing number of available genome sequences and advance analytical tool very less efforts are employed to understand intron evolution and genomic scale. The present study is focused on rice genome and there is possibility of having diverse pattern of intron distribution in other plant genomes particularly in dicots and primitive plant species. This study will be helpful to verify the facts and enrich understanding of intron distribution in genome. We have shown here a typical genome wide distribution pattern of introns in the rice genes, their correlation with exon length, CDS length and gene expression. The result presented here, would be useful in understanding the structural organization of genes with respect to the presence of introns.

Acknowledgments Authors are thankful to Dr. Swarup K. Parida (NIPGR, New Delhi, India) and Dr. Amit A. Deokar (University of Saskatchewan, Canada) for valuable input for data analysis

References

Aparicio S, Chapman J, Stupka E, Putnam N, Chia J-M, Dehal P, Christoffels A, Rash S, Hoon S, Smit A (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297(5585):1301–1310

Betrán E, Thornton K, Long M (2002) Retroposed new genes out of the X in Drosophila. Genome Res 12(12):1854–1859

Carola M, Finer JJ (2015) The intron and 5′ distal region of the soybean Gmubi promoter contribute to very high levels of gene expression in transiently and stably transformed tissues. Plant Cell Rep 34(1):111–120

Deshmukh RK, Vivancos J, Guérin V, Sonah H, Labbé C, Belzile F, Bélanger RR (2013) Identification and functional characterization of silicon transporters in soybean using comparative genomics of major intrinsic proteins in Arabidopsis and rice. Plant Mol Biol 83(4–5):303–315

Edgell DR, Belfort M, Shub DA (2000) Barriers to intron promiscuity in bacteria. J Bacteriol 182(19):5281–5289

Gilbert W (1987) The exon theory of genes. In: Cold Spring Harbor symposia on quantitative biology. Cold Spring Harbor Laboratory Press, pp 901–905

Gilbert W, De Souza SJ, Long M (1997) Origin of genes. Proc Natl Acad Sci 94(15):7698–7703

Hu K (2006) Intron exclusion and the mystery of intron loss. FEBS Lett 580(27):6361–6365

Jeffares DC, Mourier T, Penny D (2006) The biology of intron gain and loss. Trends Genet 22(1):16–22

Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, Kishimoto N, Yazaki J, Ishikawa M, Yamada H, Ooka H (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. Science 301(5631):376–379

Li W, Zhang P, Fellers JP, Friebe B, Gill BS (2004) Sequence composition, organization, and evolution of the core Triticeae genome. Plant J 40(4):500–511

Lin H, Zhu W, Silva JC, Gu X, Buell CR (2006) Intron gain and loss in segmentally duplicated genes in rice. Genome Biol 7(5):R41

Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M (2002) Intron presence–absence polymorphism in Drosophila driven by positive Darwinian selection. Proc Natl Acad Sci 99(12):8121–8126

Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA (2005) The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. Science 307(5713):1321–1324

Lynch M, Conery JS (2003) The origins of genome complexity. Science 302(5649):1401–1404

Marais G, Piganeau G (2002) Hill-Robertson interference is a minor determinant of variations in codon bias across Drosophila melanogaster and Caenorhabditis elegans genomes. Mol Biol Evol 19(9):1399–1406

Morita S, Tsukamoto S, Sakamoto A, Makino H, Nakauji E, Kaminaka H, Masumura T, Ogihara Y, Satoh S, Tanaka K (2012) Differences in intron-mediated enhancement of gene expression by the first intron of cytosolic superoxide dismutase gene from rice in monocot and dicot plants. Plant Biotechnol 29(1):115–119

Nixon JE, Wang A, Morrison HG, McArthur AG, Sogin ML, Loftus BJ, Samuelson J (2002) A spliceosomal intron in Giardia lamblia. Proc Natl Acad Sci 99(6):3701–3705

Patil G, Nicander B (2013) Identification of two additional members of the tRNA isopentenyltransferase family in Physcomitrella patens. Plant Mol Biol 82(4–5):417–426

Patil G, Kumar V, Sharma P, Deokar A, Kondawar V, Jain PK, Srinivasan R Promoter Element of an ERF Gene of Arabidopsis Drives Trichome-Specific Expression and Retains Its Specificity in Brassica juncea. In: In vitro cellular & developmental biology-animal, 2010. Springer 233 Spring ST, New York, NY 10013 USA, pp S153-S154

Ren X-Y, Vorst O, Fiers MW, Stiekema WJ, Nap J-P (2006) In plants, highly expressed genes are the least compact. Trends Genet 22(10): 528–532

Roy SW, Gilbert W (2005) Rates of intron loss and gain: implications for early eukaryotic evolution. Proc Natl Acad Sci U S A 102(16): 5773–5778

Roy SW, Gilbert W (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet 7(3):211–221

Rzhetsky A, Ayala F (1999) The enigma of intron origins. Cell Mol Life Sci 55(1):3–6

Sanderson MJ, Thorne JL, Wikström N, Bremer K (2004) Molecular evidence on plant divergence times. Am J Bot 91(10):1656–1665

Sanmiguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot 82(suppl 1):37–44

Urrutia AO, Hurst LD (2003) The signature of selection mediated by expression on human genes. Genome Res 13(10):2260–2264

Vinogradov AE (2005) Genome size and chromatin condensation in vertebrates. Chromosoma 113(7):362–369

Xu G, Guo C, Shan H, Kong H (2012) Divergence of duplicate genes in exon–intron structure. Proc Natl Acad Sci 109(4):1187–1192