Genetica (2012) 140:249–257
DOI 10.1007/s10709-012-9676-3

Evolutionary history of exon shuffling

Gustavo S. França · Douglas V. Cancherini · Sandro J. de Souza

Received: 2 January 2012 / Accepted: 23 August 2012 / Published online: 5 September 2012

Abstract Exon shuffling has been characterized as one of the major evolutionary forces shaping both the genome and the proteome of eukaryotes. This mechanism was particularly important in the creation of multidomain proteins during animal evolution, bringing a number of functional genetic novelties. Here, genome information from a variety of eukaryotic species was used to address several issues related to the evolutionary history of exon shuffling. By comparing all protein sequences within each species, we were able to characterize exon shuffling signatures throughout metazoans. Intron phase (the position of the intron regarding the codon) and exon symmetry (the pattern of flanking introns for a given exon or block of adjacent exons) were features used to evaluate exon shuffling. We confirmed previous observations that exon shuffling mediated by phase 1 introns (1-1 exon shuffling) is the predominant kind in multicellular animals. Evidence is provided that such pattern was achieved since the early steps of animal evolution, supported by a detectable presence of 1-1 shuffling units in *Trichoplax adhaerens* and a considerable prevalence of them in *Nematostella vectensis*. In contrast, *Monosiga brevicollis,* one of the closest relatives of metazoans, and *Arabidopsis thaliana*, showed no evidence of 1-1 exon or domain shuffling above what it would be expected by chance. Instead, exon shuffling events are less abundant and predominantly mediated by phase 0 introns (0-0 exon shuffling) in those non-metazoan species. Moreover, an intermediate pattern of 1-1 and 0-0 exon shuffling was observed for the placozoan *T. adhaerens*, a primitive animal. Finally, characterization of flanking intron phases around domain borders allowed us to identify a common set of symmetric 1-1 domains that have been shuffled throughout the metazoan lineage.

Keywords Exon shuffling · Metazoan evolution · Protein domains · Introns

Electronic supplementary material The online version of this article (doi:10.1007/s10709-012-9676-3) contains supplementary material, which is available to authorized users.

G. S. França · D. V. Cancherini · S. J. de Souza (✉)
Ludwig Institute for Cancer Research, São Paulo Branch, São Paulo 01323-903, Brazil
e-mail: sandro@compbio.ludwig.org.br; sandro@neuro.ufrn.br

G. S. França
Departamento de Bioquímica, Instituto de Química, Programa de Pós-Graduação, Universidade de São Paulo, São Paulo 05508-900, Brazil

Present Address:
S. J. de Souza
Brain Institute, UFRN, Av. Nascimento de Castro 2155, Natal, RN 59056-450, Brazil

Introduction

One of the most important ways to create new genes is through intron-mediated recombination, a phenomenon proposed by Gilbert more than 30 years ago and named by him "exon shuffling" (Gilbert 1978). It is known today that exon shuffling is one of the most common mechanisms to create new genes and extensive reports have been published describing several mechanistic and evolutionary aspects of exon shuffling (de Souza 2003; Eickbush 1999; Kaessmann et al. 2002; Liu and Grigoriev 2004; Long and Langley 1993; Patthy 1987; Patthy 1999).

A very influential factor in intron and exon evolution is what is called intron phase, which refers to the position of a given intron within the codon. Phase 0 introns lie between two codons, phase 1 and phase 2 introns are located after

the first and second nucleotides of the codon, respectively. Removal/insertion of introns or exons have very different effects on the reading frame of a gene. Intron insertion or removal frequently occurs without any modification in the coding sequence. However, the insertion or removal of an exon or a block of consecutive exons in a gene may disrupt the correct reading frame of all exons downstream from the insertion point if the exon or exonic block did not contain a number of nucleotides that is an exact multiple of three. This latter case will take place when the phase of the two introns bordering the exon or exonic block is the same in both extremities. When such type of exon or exonic block is involved in an exon shuffling event, it is said that the shuffling unit is symmetric. Because they preserve the reading frame of genes created after exon shuffling events, they are believed to be less likely a target for purifying selection.

Indeed, the excess of symmetric exons in most of the genomes analyzed up to now suggests that exon shuffling has been a major player in shaping the genome of eukaryotic species (de Souza et al. 1998; Long et al. 1995). Another feature, mainly observed in metazoans, is the prevalence of shuffling units flanked by introns of phase 1 (1-1 exon shuffling). It has been shown that there is a large excess of 1-1 exons in animal species compared to what would be expected from the total proportions of intron phases in the corresponding genomes (Patthy 1999; Long et al. 1995). The same excess is also observed for protein domains flanked by phase 1 introns (Kaessmann et al. 2002; Liu et al. 2005; Vibranovski et al. 2005).

In spite of all these advances, important information on the mechanism and functional impact of exon shuffling along evolution remains unknown. For example, was exon shuffling a frequent mechanism for the creation of new genes in non-metazoan species? Although there are isolated examples of exon shuffling in non-metazoan species (Elrouby and Bureau 2010; Long et al. 1996; Morgante et al. 2005), little is known about the overall frequency and mode of exon shuffling in these species. When exactly did 1-1 exons (and domains flanked by phase 1 introns) start to expand during evolution? Reports from Patthy's group (Patthy 1999; Patthy 2003) indicate that 1-1 exon shuffling is specific to metazoans, although a more complete comparative analysis is still missing. Was 0-0 and 2-2 exon shuffling relevant in any stage of eukaryotic evolution?

The questions posed above were addressed by generating a catalog of putative exon shuffling events in several eukaryotic species, including non-metazoans and early branching metazoans. We observed that 1-1 exon shuffling became the most frequent type of exon shuffling soon after the emergence of Metazoa. Furthermore, non-metazoan species have a predominance of 0-0 exon shuffling. Intriguingly, *Trichoplax adhaerens*, which is believed to be one of the most basal animals (Dellaporta et al. 2006; Srivastava et al. 2008), presents an intermediary frequency of 1-1 and 0-0 shuffling units. Moreover, 2-2 exon shuffling seems to have had no significant contribution in the building of modular proteins in any analyzed species. The availability of genome sequence for all these species allowed us to draw a map of domain shuffling occurrence along eukaryotic evolution, revealing the expansion of certain domains during specific periods of animal evolution.

Results and discussion

Distribution of intron phase and exon symmetry

As shown by us and others, intron phase distribution in most eukaryotic species is biased toward a predominance of phase 0 introns (de Souza et al. 1998; Fedorov et al. 1992; Long et al. 1995; Nguyen et al. 2006; Qiu et al. 2004). This has been a central issue in the debate about the existence of introns in the ancestor of eukaryotes and prokaryotes. To evaluate if the dataset used by us was in agreement with these previous observations, we calculated the number of phase 0, 1 and 2 introns for all analyzed species. Figure 1a shows that indeed phase 0 is the most prevalent type of intron (42 to 57 %) in all species, followed by phase 1 and 2. Next, we evaluated the relative excess of symmetric exons (flanked by introns of same phases) and asymmetric exons (flanked by introns of distinct phases) from the expected values in a hypothetic scenario where intron phases are randomly distributed in respect to symmetry (Fig. 1b). The main message from Fig. 1b is the significant excess of symmetric exons for almost all analyzed species (p values by Chi-square test are shown in Table S1, and range from 10^{-2} to 10^{-145}). We believe that the best currently available explanation for favored exon symmetry is that exon shuffling has repeatedly occurred during evolution and left remnant signals, even if sequence similarity cannot identify some past shuffling events anymore. We thus speculate that the symmetric exon excesses observed by us reflect a pervasive occurrence of shuffling events in eukaryotic genomes. As seen before, higher excess was observed for 1-1 exons, especially for metazoans (de Souza et al. 1998; Long et al. 1995; Vibranovski et al. 2005). Humans, for example, have an excess of 11, 19 and 8 % for 0-0, 1-1 and 2-2 symmetric exons, respectively. Interestingly, plants have smaller excess for symmetric exons (9, 9 and 13 % for 0-0, 1-1 and 2-2 exons, respectively). The choanoflagellate *Monosiga brevicollis* showed a small excess of symmetric exons that resulted only from an excess of 0-0 exons.

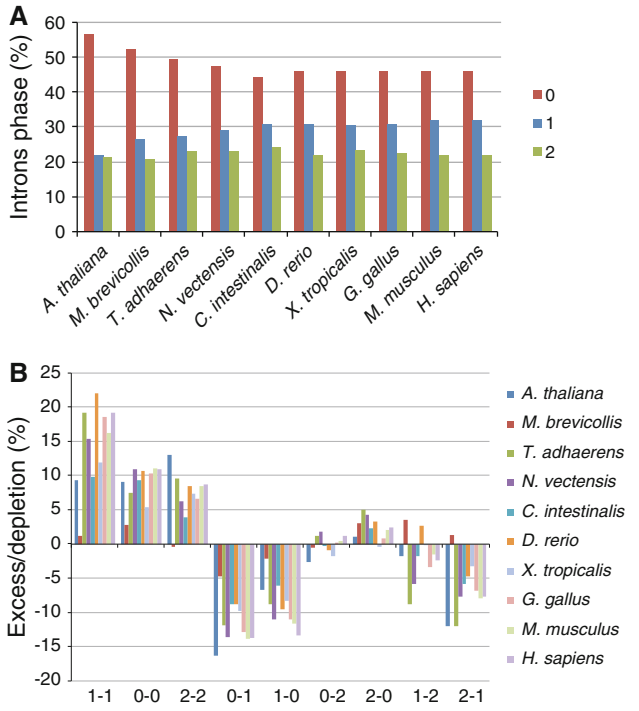

Fig. 1 **a** Proportions of the three intron phases and **b** percentage of excess (positive values)/depletion above/under expectation for the nine possible intron phase combinations around exons (see "Materials and methods") in protein coding genes for 10 eukaryotic species

Generation of a catalog of exon shuffling events

Although the analyses shown in Fig. 1 are relevant, they are indirect evidence of the role of exon shuffling in the creation of new genes. To search for more direct evidence, we identified protein regions that possibly evolved by exon shuffling. Our approach was similar to the one used by us (Vibranovski et al. 2005; Cancherini et al. 2010) and others (Long et al. 1995; Saxonov and Gilbert 2003) and was based in an all-against-all comparison between all proteins of a given species. We excluded alignments between gene duplication products both by discarding alignments between known paralogs and requiring the aligned region not to exceed 40 % of the length of the shorter protein. An important and stringent criterion was the presence of introns flanking both homologous regions in the non-homologous genes (see Fig. S1 for a schematic view of the strategy). Table 1 gives the number of exon shuffling events detected by our strategy for each species. While 6.4 % of all human genes have evidence of at least one exon shuffling event, the same number is 1 % for *Arabidopsis thaliana*. Overall, we still believe that our numbers are underestimated because of the stringent criteria used for the identification of homologous regions. We used the same approach for *Cryptococcus neoformans*, *Rhizopus oryzae* and *Batrachochytrium dendrobatis*, three intron rich or moderately rich fungi species. However, the number of putative exon shuffling events was insufficient to perform further analyses (data not shown). This may be, in part, due to processes of both intron gain and intron loss in extant fungi species (Stajich et al. 2007), obscuring the required signal of flanking introns.

With this dataset, we proceeded to evaluate the distribution of intron phases flanking all shuffling units, i.e., exons or groups of exons flanked by introns. As expected, we observed a high prevalence of symmetric units within the set of exon shuffling examples, which is something around 60 % for all species. We also observed a high prevalence of 1-1 exon shuffling in metazoans (ranging from 25 to 44 %—Fig. 2). Although there was not sufficient genome information at the time for a thorough investigation, Patthy had already suggested that 1-1 exon shuffling would have increased dramatically in early phases of metazoan evolution, likely associated with the emergence of extracellular matrix (ECM), coagulation factors, cell adhesion molecules and several cell surface proteins (Patthy 1999, 2003).

The recent genome sequencing of two early diverging metazoans, *Trichoplax adhaerens* (Srivastava et al. 2008) and *Nematostella vectensis* (Putnam et al. 2007), and the choanoflagellate *Monosiga brevicollis* (King et al. 2008), allowed us to look more closely at the earliest branches of Metazoa and one of its closest relatives, giving for the first time direct support to the concept that 1-1 exon shuffling initiated and expanded concomitant with the appearance of the first animals. Interestingly, the proportions of intron phase combinations in *T. adhaerens* are intermediate between other metazoans, on one hand, and non-metazoans, on the other. In this species, 0-0 and 1-1 exon shuffling represent 23 and 25 % of total exon shuffling cases, respectively. The frequency of 1-1 in this species is almost two fold lower than the average frequency for other metazoans (25 vs. 41 %). For the cnidarian *N. vectensis,* this proportion is ~40 %, quite similar to what is seen in other metazoans analyzed by us. Conversely, 1-1 exon shuffling was highly underrepresented in the choanoflagellate *M. brevicollis* and in the plant *A. thaliana* (Fig. 2). This suggests that shuffling of these 1-1 exonic units might have contributed to the gain of functional complexity in metazoans since the divergence of the most primitive animals. Gene ontology terms that were significantly enriched in the set of exon shuffling genes for *Homo sapiens* included several genes related to critical features that allowed the rise of animal multicellularity, such as cell adhesion, ECM, basement membrane, blood coagulation and immune response (Table S2).

We next evaluated whether a similar picture could be seen for exon shuffling events involving protein domains. We selected from our dataset all shuffling units containing whole protein domains and evaluated the distribution of phase combinations for introns flanking such regions.

Table 1 Number of genes involved in exon shuffling events

Species	Total of genes	ES genes	ES genes (%)	ES events
H. sapiens	23686	1546	6.4	2884
M. musculus	24496	1037	4.2	1610
G. gallus	16736	899	5.4	1144
X. tropicalis	18025	790	4.4	976
D. rerio	21322	911	4.3	1139
C. intestinalis	14180	442	3.2	386
N. vectensis	27273	818	3.0	984
T. adhaerens	11520	483	4.2	349
M. brevicollis	9196	233	2.5	192
A. thaliana	26814	272	1.0	185

Aligned pairs were considered as ES events

We considered a gene involved in exon shuffling (ES) event if its protein has at least one aligned region with a non-homologous protein, sharing identity ≥30 % and length ≤40 % of the shorter protein, flanked by introns in both of its extremities

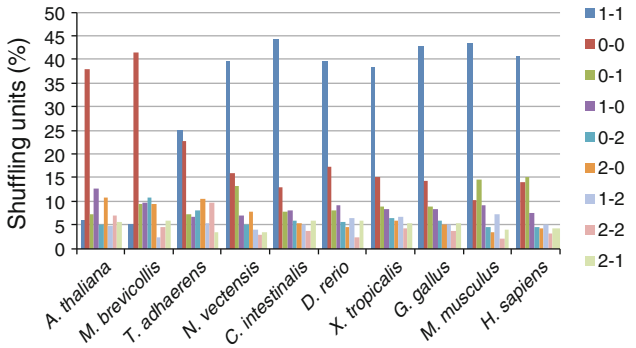

Fig. 2 Patterns of intron phase combinations around shuffling units. *Bars* represent percentages of the total number of shuffling units. A shuffling unit was defined as a protein region sharing sequence similarity with another non-homologous protein and flanked by introns at both boundaries

Figure 3a shows that the prevalence of 1-1 shuffling depends on the presence of one or more protein domains within the shuffling unit. In our datasets for metazoans, around 65 % of shuffling units containing protein domains are bordered by phase 1 introns, while for units not containing protein domains this percentage is only around 15 %. Since protein domains are one of the most important functional units in proteins, this result suggest that the spread of 1-1 exons is linked to its functional impact on proteins and likely a target for positive selection. For shuffling units that contain no protein domains, no clear trend is evident and a higher frequency of non-symmetric exons may suggest that this dataset is enriched with cases whose evolutionary dynamics is neutral (Fig. 3b). Furthermore, this distinction between shuffling units with and without protein domains highlights the importance of domains shuffling in the evolution of genes and proteins in animals.

Signatures of domain shuffling revealed by flanking introns

To complement the analysis discussed above, we evaluated phase combinations for introns around protein domains, based on their occurrence in all proteins of the respective species, as done by us and others (Kaessmann et al. 2002; Liu and Grigoriev 2004; Liu et al. 2005; Vibranovski et al. 2005). As seen in Fig. 4, a high frequency (∼60 %) of domains flanked by phase 1 introns is clearly associated with the emergence of Metazoa. Furthermore, non-metazoan species have predominantly 0-0 domains (∼40 %). Again, *T. adhaerens* appears with intermediate figures between non-metazoans and other metazoan species, with 19 and 48 % of 0-0 and 1-1 domains, respectively Fig. 4 also reveals that shuffling of 2-2 domains is extremely rare (less than 5 % frequency) in all analyzed species.

The identification of all intron-flanked domains in a variety of species allowed us to have a broader view of domain occurrence in eukaryotes. We clustered intron-flanked domains based on relative frequencies of phase combinations of flanking introns (Fig. 5a, b, Fig. S2). For this specific analysis we used another invertebrate deuterostome, the sea urchin *Strongylocentrotus purpuratus*, instead of *Ciona intestinalis*, due to the lack of sensibility of our method when applied to species that, like sea squirt, have undergone profound intron loss (Putnam et al. 2007). Despite the possibility of underestimating the frequency of intron-flanked domains because of the strictness of our criteria, interesting patterns could be observed. First, the great majority of domains identified by us are in agreement with previous works concerning exon or domain shuffling (Liu et al. 2005; Kaessmann et al. 2002; Kawashima et al. 2009; Vibranovski et al. 2005). Second, there is a common set of protein domains predominantly flanked by phase 1

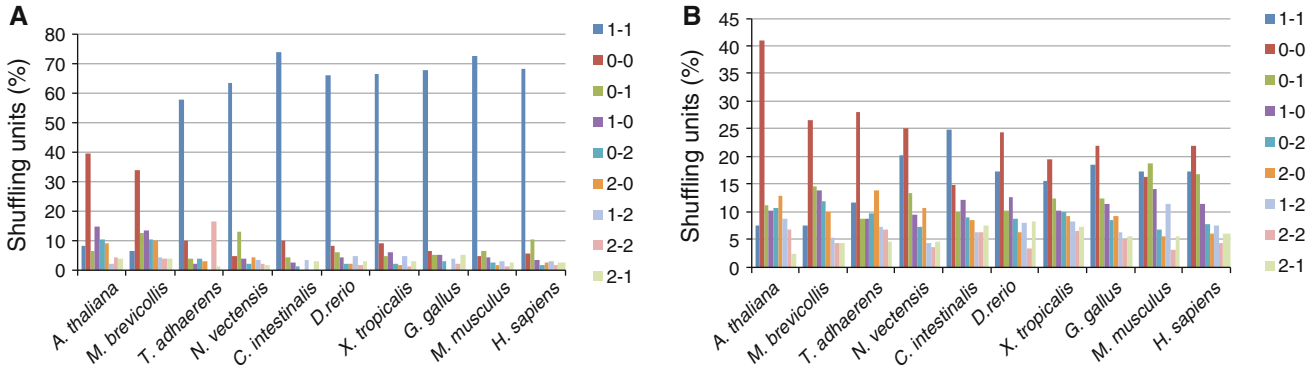

Fig. 3 Patterns of intron phase combinations flanking shuffling units containing (**a**) and not containing (**b**) whole protein domains. Only Pfam-A domains with e-value $\leq 10^{-2}$ were included in the analysis

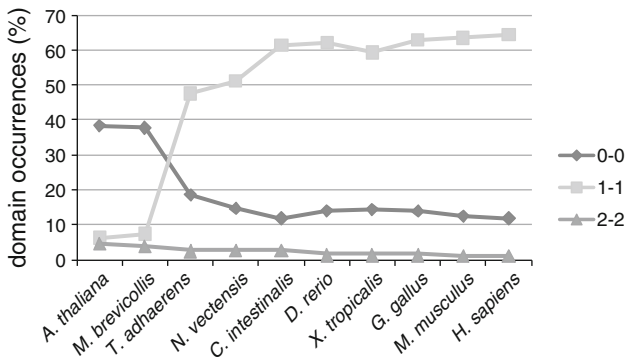

Fig. 4 Proportion of symmetric domains in eukaryotic species. Values represent percentages of domain occurrences for each symmetric class in respect to total domain occurrences flanked by introns

introns along most of the metazoan lineage (Fig. 5a). Several domains associated with ECM, cell adhesion and other animal-specific features (Patthy 1996, 1999, 2003; Kawashima et al. 2009; Hynes 2012), such as EGF-like, TSP_1, fn3, CUB, Sushi, MAM, VWC, LDL-A, V-set, I-set, etc. (Fig. 5a) have a strong signal of flanking introns since the divergence of the early branching animal *T. adhaerens*. Shuffling of this common set of domains would be involved in the establishment of core proteins related with hallmarks of animal multicellularity, such as collagen-based ECM cell–cell adhesion mediated by cell-ECM interactions (Aouacheria et al. 2004; Tyler 2003). Curiously, instances of fn3 and EGF-like domains appear flanked by phase 1 introns in *M. brevicollis*. Although the choanoflagellate have poor symmetric 1-1 signal of flanking introns, its genome encodes for several of those "metazoan-like" domains in a considerable number of copies, while in plants they are virtually absent (Table S3). This reinforces the statement that the common ancestor of metazoans and choanoflagellates already had the central building blocks for the origin of ECM, cell adhesion molecules and other animal-specific characteristics (King et al. 2008). It is then reasonable to assume that soon after the divergence of Metazoa, those functional units would have been recurrently shuffled.

It is worth mentioning that green regions in the heatmap should not be interpreted as the absence of certain domains, but the absence of signal of flanking introns. With the help of Table S3, we can hypothesize that domains like Zona_pellucida, Trefoil, Kringle and PAN, for example, lost most of their signal of flanking introns in ancestor species, since they were found in *N. vectensis* and/or *T. adhaerens* (Fig. 5a; Table S3). Kunitz_BPTI may be an example of domain expansion after the divergence of cnidarians. Based on the number of domain occurrences and 1-1 signal of flanking introns, domains such as fn1, Immunoglobulin C1-set, MHC_I and Xlink, were probably spread through exon shuffling in the vertebrate lineage (Fig. 5a; Table S3). Indeed, we have not found a signal of flanking introns for these domains in any other non-vertebrate species that were examined. Moreover, there are fewer occurrences of those domains in non-vertebrate species (Table S3).

In respect to 0-0 domains, we have found a less marked contrast in terms of abundance and signal of flanking introns between metazoan and non-metazon species in comparison to 1-1 domains (Fig. 5b; Table S3). A very small number of domain types, such as Ankyrin, WD40, RhoGEF, Kinesin, LSM and PX, displayed flanking by phase 0 introns both in multicellular animals and *M. brevicollis* or *A. thaliana*. Other domain types, including lactophilin/CL-1 like GPS, collagen triple helix repeat, KH and 7 transmembrane receptor domains, are abundantly flanked by phase 0 introns specifically in metazoans (Fig. 5b), although they occur in the choanoflagellate (Table S3). The heatmap also revealed groups of asymmetrically flanked 1-0, 2-1 and 0-1 domains (Fig S2). Most commonly, these patterns of intron flanking first appeared in a particular metazoan species, with more recently branching species usually retaining the pattern. There are, however, domains that frequently exhibit more than one combination of flanking intron phase: SAM, for

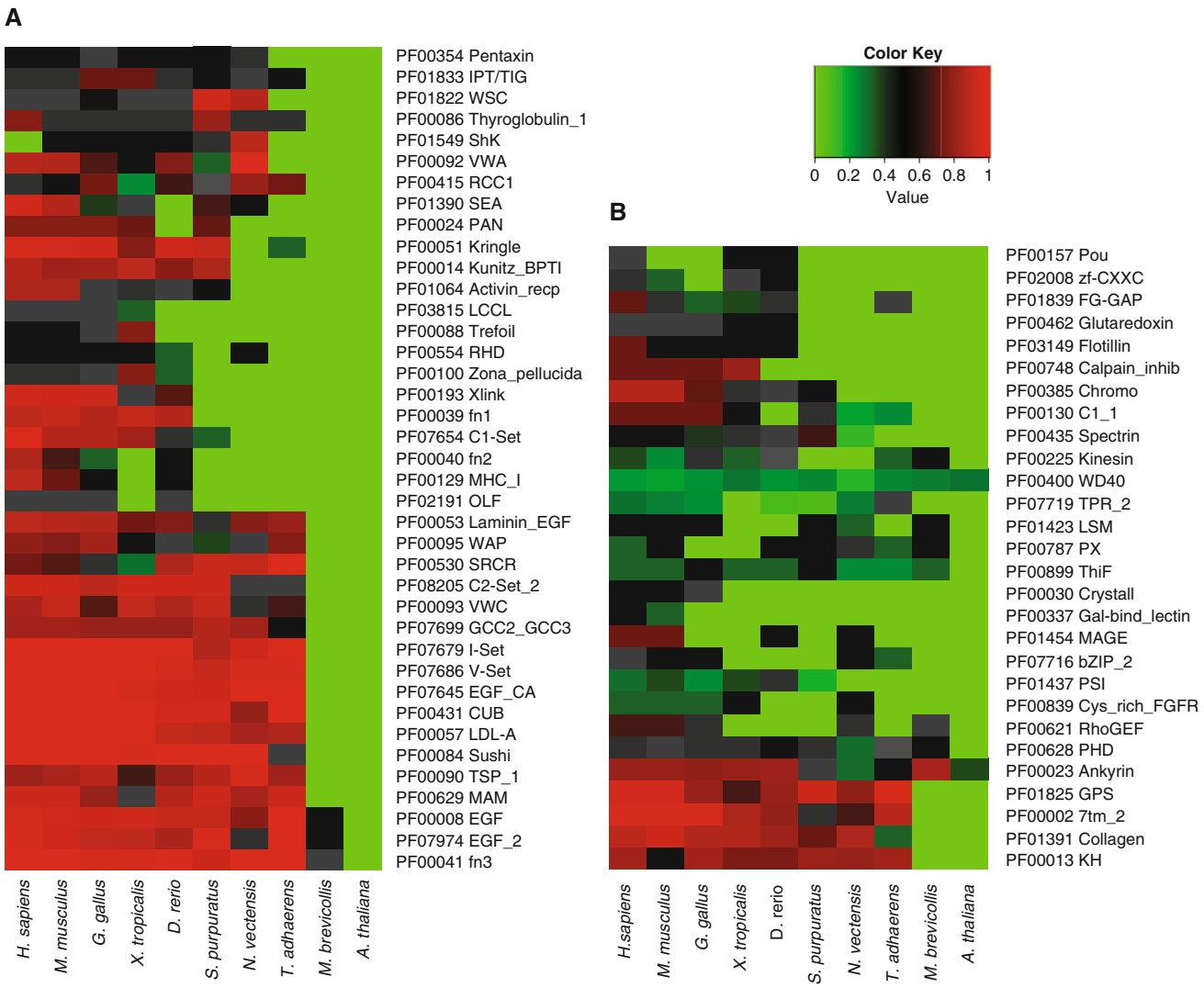

Fig. 5 Heatmap of domains flanked by 1-1 (**a**) and 0-0 (**b**) intron phase combinations. *Red* and *green* regions correspond, respectively, to high and low fractions of the particular intron phase combination. The figure represents a selected portion of the complete heatmap, including only domains which were visibly clustered on the basis of a shared 1-1 or 0-0 signal of flanking introns. The heatmap was generated considering only domains which had more than 2 instances flanked by introns and frequencies of intron flanking higher than 10 % of total occurrences of a particular domain. The complete heatmap showing the pattern of intron flanking for all nine intron phase combinations for all domains fulfilling our above-mentioned criteria can be seen in Fig. S2. The total number of occurrences of these domains is in Table S3. We used Pfam abbreviations for domain names. (Color figure online)

instance, exist as 1-1 and 1-0 domains, and Surp module, as 1-1 and 2-0 domains (Fig. S2). Nevertheless, it remains to be evaluated whether exon shuffling was involved in the spreading of these groups of asymetrically flanked domains.

To precisely trace the evolutionary path of each domain in regard to its expansion through exon or domain shuffling would require a more comprehensive set of species and specific methodology, which is beyond the scope of this study. However, our results clearly demonstrate that some domains have a restricted pattern for phase of flanking introns, and they have been spread in specific lineages. The work of Kawashima et al. (2009) illustrates how shuffling of some of those domains observed in Fig. 5 can give rise to proteins with new functions. The authors recognized more than 1,000 new domain pairs in the vertebrate lineage, and some of them are related to vertebrate-specific structures. One interesting example is aggrecan, an essential component of cartilage, which is composed by a new combination of Xlink, a domain with vertebrate-specific 1-1 intron flanking signal (Fig. 5a) and copy number expansion (Table S3), and Immunoglobulin (V-set) domain, whose 1-1 flanking signal and copy number expansion in metazoans predate the divergence of Placozoa (*T. adhaerens*) (Fig. 5a). Another example might be Tectorin-alpha and Cochlin, proteins

involved in the vertebrate auditory system that appeared by the acquisition of Zona_Pellucida and LCCL domains, respectively. Kawashima et al. (2009) also identified genes created by domain shuffling involved in chordate-specific structures such as endostyle, Reissner's fiber of the neural tube and notochord. In spite of this effort, uncovering relationships between such evolutionary mechanism and phenotypic consequences is a major challenge in biology, opening a vast and exciting field for future investigations.

Concluding remarks

In conclusion, a picture of the evolution of exon shuffling signatures in primitive and derived metazoans and two non-metazoan species is provided in this report. Two major features are evident. First, the frequency of 1-1 shuffling increases dramatically since the emergence of the first animals. Second, before the emergence of metazoans, ancient exon shuffling events are more likely to be associated with units flanked by phase 0 introns. The participation of 2-2 exon shuffling was minimal in all analyzed species. How this scenario emerged? One possibility is that phase 0 introns are more ancient than phase 1 and 2 introns as has been suggested by de Souza (2003) and de Souza et al. (1998). In that case, the frequency of phase 0 introns would be higher in the early stages of eukaryotic evolution, which would explain the prevalence of phase 0 introns in the shuffling events in non-metazoan species. We also identified a core set of protein domains flanked by phase 1 introns across the whole metazoan lineage, and some domains flanked in more restricted phylogenetic groups. The genetic and selection mechanisms that lie behind the preference for shuffling 1-1 exons or domains remain largely unknown. The acquisition of a protein domain must overcome structural limitations. Domains possessing small and flexible interfaces tend to fold independently (Han et al. 2007), so they could confer to shuffling units more acceptance in a multidomain protein context. In this respect, phase one introns more frequently interrupt glycine codons, thus generating shuffling units bordered by glycine residues, which contain the smallest side chain (Fedorov et al. 2001). Obviously, selective forces subsequently acting on functional properties could also take part in spreading of those units. The biological implications of exon and domain shuffling events, how they affect the rise of specific characteristics and what is the impact of such events in terms of protein interaction networks (Cancherini et al. 2010) and genome complexity are pivotal questions quite unexplored. Studies addressing these issues are just beginning and much effort is still necessary for a better understanding of the influence of exon shuffling in shaping the genome architecture and the evolution of new functions.

Materials and methods

Species and public databases

The species used in this study were chosen considering features like intron richness, completely sequenced genomes, availability of public data, and representativeness across main branches of eukaryotes. The collection of protein coding sequences and annotation files were obtained from Ensembl version 47 (http://www.ensembl.org) for *Homo sapiens*, *Mus musculus*, *Gallus gallus*, *Xenopus tropicalis*, *Danio rerio* and *Ciona intestinalis,* and from JGI site (http://www.jgi.doe.gov) for *Nematostella vectensis* (filtered gene models, genome version 1.0), *Trichoplax adhaerens* (filtered gene models, genome version 1.0), and *Monosiga brevicollis* (filtered gene models, genome version 1.0). For *Strongylocentrotus purpuratus* data was obtained from HGSC at Baylor College of Medicine (GLEAN models, March 2007) and for *Arabidopsis thaliana,* from NCBI site (December 2007 version).

Intron positions and intron phase distribution

Intron positions were derived from annotation files regarding coding sequence coordinates. Intron phases were defined according to their position within the codons. Phase 0 introns lie between two codons, phase 1 introns, between the first and second nucleotides, and phase 2 introns, between the second and third nucleotides of the codon. We determined frequencies of intron phases and excesses of intron phase combinations considering the longest protein for each gene. The excess (or depletion) of each of the nine possible intron phase combinations was calculated according to Long et al. (1995), as being $(O - Ne)/(Ne)$, where, N is the total number of exons, e is the expected frequency of exons of a given intron phase combination and O is the number of observed exons with such combination. The expected frequency e is $P_x P_y$, where P_x and P_y are the frequencies of the introns of phase x and y, located at the 5′ and 3′ borders of the exons, respectively. To calculate significance levels of the excesses of symmetric exons, we individually compared symmetric classes (0-0, 1-1, 2-2) with the sum of all other remaining classes by means of Chi-square (X^2) test at significance level of p value <0.01 ($df = 1$). X^2 test was also used to compare the sum of all symmetric classes against the sum of all asymmetric classes (0-1, 0-2, 1-0, 1-2, 2-0, 2-1). Such statistical approach was implemented by Kaessmann et al. (2002).

Identification of protein regions possibly evolved by exon shuffling

Our strategy to identify protein regions which were probably acquired by exon shuffling events was essentially the same

used by Cancherini et al. (2010) and it was based on all-against-all alignments within each proteome by using Blastp 2.2.17 (Altschul et al. 1997) with default parameters. For genes that have more than one product, we just considered the protein of maximal length. In order to obtain alignments between non-homologous proteins sharing similar regions (HSPs—high-scoring segment pairs), we filtered Blast results through Perl scripting, requiring identity greater than 30 % and coverage less than 40 % of the shorter protein. Then, we sought for introns flanking the borders of these conserved regions as an evidence of exon shuffling. Introns were searched within an interval of 3 nucleotides towards the inside portion of the HSP and 32 nucleotides towards the outside portion of the HSP (Fig. S1). Scripts used to perform this analysis and instruction on how to use them will be available under request.

Domains flanked by introns

Protein domain annotations were performed by running HmmPfam with Pfam-A version 22 domain information (Finn et al. 2008) on known or predicted protein sequences of analyzed species, requiring an e-value $\leq 10^{-2}$. Introns flanking the borders of domain instances were sought within the same nucleotide interval used for HSPs. A similar interval to search for introns bordering protein domains was used by Liu and Grigoriev (2004) and Liu et al. (2005). To avoid overestimation of the signal of flanking introns due to gene duplication, we looked for paralog relationships among proteins. For Ensembl species, lists of paralogous proteins were obtained with the help of Biomart (Smedley et al. 2009). For all other species, we locally defined the relationships from the Blastp output. Two given proteins were considered paralogous if their alignment had an e-value $\leq 10^{-6}$, identity ≥ 30 %, and coverage greater than 70 % of the longer protein. Protein clustering into families was performed through the single-linkage clustering method, that is, all connected components of the graph of paralogy relationships were taken as a paralogy family.

For each domain type, frequency of flanking introns for the nine possible intron phase combinations was determined. In order to avoid giving excessive weight to families of repeatedly duplicated proteins sharing the same pattern of intron flanking, all domains in a paralog family with n different genes were counted as $1/n$ of a domain for purposes of both counting the total number of domains and the number of flanked domains. Scripts used to perform this analysis and instruction on how to use them will be available under request.

Hierarchical clustering of domains flanked by introns

In order to build a hierarchical clustering of domain types and signatures of intron flanking, we started by collecting frequencies of flanking introns for all domains. We only considered domains which had more than 2 instances flanked by introns and whose frequencies of flanking were higher than 10 %. For estimating the flanking percentages for domain types, we used a simple Bayesian model where we started with an a priori uniform distribution between 0 and 1 for flanking probability of each domain type and calculated a posterior flanking probability distribution assuming that each domain instance is flanked with a fixed probability characteristic of that domain type. In order to avoid overestimating these flanking percentages for the domains with few occurrences, the final estimates for the flanking percentages were taken as equal to the 0.25 quantile of the posterior probability distribution. A similar Bayesian approach was used in order to estimate, for each domain type, probabilities for the nine possible types of intron flanking, concerning intron phase combinations. All flanking percentages were separately calculated for each analyzed species. The clustering by domain type and by flanking percentages was performed using the R statistical language (http://www.r-project.com), with Euclidean distances. Scripts used to perform this analysis and instruction on how to use them will be available under request.

Acknowledgments Gustavo S. França and Douglas V. Cancherini were supported by FAPESP scholarships.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res 25:3389–3402

Aouacheria A, Cluzel C, Lethias C, Gouy M, Garrone R, Exposito JY (2004) Invertebrate data predict an early emergence of vertebrate fibrillar collagen clades and anti-incest model. J Biol Chem 279:47711–47719

Cancherini DV, França GS, de Souza SJ (2010) The role of exon shuffling in shaping protein-protein interaction networks. BMC Genomics 11(Suppl 5):S11

de Souza SJ (2003) The emergence of a synthetic theory of intron evolution. Genetica 118:117–121

de Souza SJ, Long M, Klein JR, Roy S, Lin S, Gilbert W (1998) Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proc Natl Acad Sci USA 95:5094–5099

Dellaporta SL, Xu A, Sagasser S, Moreno MA, Buss L, Schierwater B (2006) Mitochondrial genome of *Trichoplax adhaerens* supports Placozoa as the basal lower metazoan phylum. Proc Natl Acad Sci USA 103:8751–8756

Eickbush TH (1999) Exon shuffling in retrospect. Science 283: 1465–1467

Elrouby N, Bureau TE (2010) Bs1, a new chimeric gene formed by retrotransposon-mediated exon shuffling in maize. Plant Physiol 153:1413–1424

Fedorov A, Suboch G, Bujakov M, Fedorova L (1992) Analysis of nonuniformity in intron phase distribution. Nucleic Acids Res 20:2553–2557

Fedorov A, Cao X, Saxonov S, de Souza SJ, Roy SW, Gilbert W (2001) Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. Proc Natl Acad Sci USA 98:13177–13182

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A (2008) The Pfam protein families database. Nucleic Acids Res 36:281–288

Gilbert W (1978) Why genes in pieces? Nature 271:501

Han J, Batey S, Nickson AA, Teichmann SA, Clarke J (2007) The folding and evolution of multidomain proteins. Nat Rev Mol Cell Biol 8:319–330

Hynes RO (2012) The evolution of metazoan extracellular matrix. J Cell Biol 196:671–679

Kaessmann H, Zöllner S, Nekrutenko A, Li WH (2002) Signatures of domain shuffling in the human genome. Genome Res 12:1642–1650

Kawashima T, Kawashima S, Tanaka C, Murai M, Yoneda M, Putnam NH, Rokhsar DS, Kanehisa M, Satoh N, Wada H (2009) Domain shuffling and the evolution of vertebrates. Genome Res 19:1393–1403

King N, Westbrook MJ, Young SL et al (2008) The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature 451:783–788

Liu M, Grigoriev A (2004) Protein domains correlate strongly with exons in multiple eukaryote genomes: evidence of exon shuffling? Trends Genet 20:399–403

Liu M, Walch H, Wu S, Grigoriev A (2005) Significant expansion of exon-bordering domains during animal proteome evolution. Nucleic Acids Res 33:95–105

Long M, Langley CH (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. Science 293:91–95

Long M, Rosenberg C, Gilbert W (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. Proc Natl Acad Sci USA 95:219–223

Long M, de Souza SJ, Rosenberg C, Gilbert W (1996) Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. Proc Natl Acad Sci USA 93:7727–7731

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37:997–1002

Nguyen HD, Yoshihama M, Kenmochi N (2006) Phase distribution of spliceosomal introns: implications for intron origin. BMC Evol Biol 6:69

Patthy L (1987) Intron-dependent evolution: preferred types of exons and introns. FEBS Lett 214:1–7

Patthy L (1996) Exon shuffling and other ways of module exchange. Matrix Biol 15:301–310

Patthy L (1999) Genome evolution and the evolution of exon suffling-a review. Gene 238:103–114

Patthy L (2003) Modular assembly of genes and the evolution of new functions. Genetica 118:217–231

Putnam NH, Srivastava M, Hellsten U et al (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94

Qiu W, Schisler N, Stoltzfus A (2004) The evolutionary gain of spliceosomal introns: sequence and phase preferences. Mol Biol Evol 21:1252–1263

Saxonov S, Gilbert W (2003) The universe of exons revisited. Genetica 118:267–278

Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A (2009) Biomart—biological queries made easy. BMC Genomics 10:22

Srivastava M, Begovic E, Chapman J et al (2008) The *Trichoplax adhaerens* genome and the nature of placozoans. Nature 454:955–960

Stajich JE, Dietrich FS, Roy SW (2007) Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. Genome Biol 8:R223

Tyler S (2003) Epithelium—the primary building block for metazoan complexity. Integr Comp Biol 43:55–63

Vibranovski MD, Sakabe NJ, de Oliveira RS, de Souza SJ (2005) Signs of ancient and modern exon shuffling are correlated to the distribution of ancient and modern domains along proteins. J Mol Evol 61:341–350