Evolution of eukaryotic genome architecture: Insights from the study of a rapidly evolving metazoan, *Oikopleura dioica*

Non-adaptive forces such as elevated mutation rates may influence the evolution of genome architecture

Sreenivas Chavali[1)], David A. de Lima Morais[2)], Julian Gough[2)] and M. Madan Babu[1)]*

Recent sequencing of the metazoan *Oikopleura dioica* genome has provided important insights, which challenges the current understanding of eukaryotic genome evolution. Many genomic features of *O. dioica* show deviation from the commonly observed trends in other eukaryotic genomes. For instance, *O. dioica* has a rapidly evolving, highly compact genome with a divergent intron-exon organization. Additionally, *O. dioica* lacks the minor spliceosome and key DNA repair pathway genes. Even with a compact genome, *O. dioica* contains tandem repeats, comparable to other eukaryotes, and shows lineage-specific expansion of certain protein domains. Here, we review its genomic features in the context of current knowledge, discuss implications for contemporary biology and identify areas for further research. Analysis of the *O. dioica* genome suggests that non-adaptive forces such as elevated mutation rates might influence the evolution of genome architecture. The knowledge of unique genomic features and splicing mechanisms in *O. dioica* may be exploited for synthetic biology applications, such as generation of orthogonal splicing systems.

Keywords:

- gene duplication; introns; protein domains; tandem repeats; transposable elements

Introduction

Genomic architecture of eukaryotes differs considerably from that of prokaryotes primarily due to the presence of spliceosomal introns, synteny conservation over long evolutionary spans, mostly monocistronic mRNAs and non-random organization of genes preponderantly clustered based on function or expression [1]. Sequencing of the genomes of various eukaryotes has provided a wealth of information on various aspects of the global genome architecture and different genomic features such as intron-exon organization, regulatory regions, splice sites, repeating sequences, transposable elements (TEs) and non-coding RNAs (ncRNAs). This information has constantly increased and has challenged our understanding of genomes and the underlying principles that drive genome

DOI 10.1002/bies.201100034

[1)] MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK
[2)] Department of Computer Science, University of Bristol, The Merchant Venturers Building, Bristol, UK

***Corresponding author:**
Sreenivas Chavali
E-mail: schavali@mrc-lmb.cam.ac.uk

Abbreviations:
lncRNA, long non-coding RNA; **LSE,** lineage specific expansion; **miRNA,** microRNA; **ncRNAs,** non-coding RNA; **TE,** transposable element; **TSP-1,** thrombospondin type 1; **WGD,** whole genome duplication.

evolution. Recent analyses of genome sequences of organisms from different phyla present interesting and contradicting insights. In this review, we discuss recent advances in this field of research and place special emphasis on the recently sequenced *Oikopleura dioica* genome [2].

O. dioica, an appendicularian pelagic tunicate with a conserved fundamental chordate body plan, feeds using a unique gelatinous feature known as 'house'. Unique among tunicates, *O. dioica* has separate sexes, which are genetically determined. It is becoming an important model organism owing to: (i) its relation to other chordates including humans, (ii) its ease of rearing in laboratories and (iii) its short life cycle [3]. Genome analysis of *O. dioica* has provided some fundamental evolutionary insights. For instance, contrary to previous belief, phylogenetic analysis using the then available partial genome sequence of *O. dioica* established tunicates to be the closest relatives of vertebrates displacing cephalochordates (Fig. 1) [4]. This might have resulted from the probable evolution of tunicate genome through simplification from a more complex chordate ancestor. Nevertheless, *O. dioica* is different compared to tunicate ascidians, cephalochordates and vertebrates owing to features such as: (i) a rapidly evolving small genome compared to cephalochordates and vertebrates, (ii) a loss of the retinoic acid signalling mechanism that is important for the development of the chordate body plan and (iii) an apparent loss of notochord genes observed in cephalochordates [5, 6]. Additionally, in contrast to other chordates, but similar to tunicate ascidians [7], *O. dioica* adopts 'determinate cleavage' during embryonic development in which the cell fate is set early, with reduced cell numbers. Here, we consider the implications for deriving principles of evolution from the genome sequence of this interesting organism and discuss how the recent findings might aid synthetic biology applications.

Organization of genes in *O. dioica*

Most eukaryotic genes are monocistronic, often with corresponding regulatory elements organized non-randomly. In the following sections, we discuss different facets of gene organization and how the knowledge obtained from the *O. dioica* genome expands our current understanding.

Gene order is less conserved

Eukaryotic genomes show a considerable conservation of gene order (synteny) in a large number of orthologues over long evolutionary spans (Fig. 2A) [8, 9]. Gene order is conserved among genes that are imprinted, co-expressed or functionally related [10]. Conservation of synteny may be attributed to short intergenic distances that hinder recombination, divergent and co-directional transcription brought about by bidirectional promoters or constitutive expression of genes across many tissues as shown for humans [8, 10, 11]. Changes in the genomic neighbourhood of genes during evolution result in altered expression patterns. The genomic neighbourhood-induced alterations in expression of key genes have been

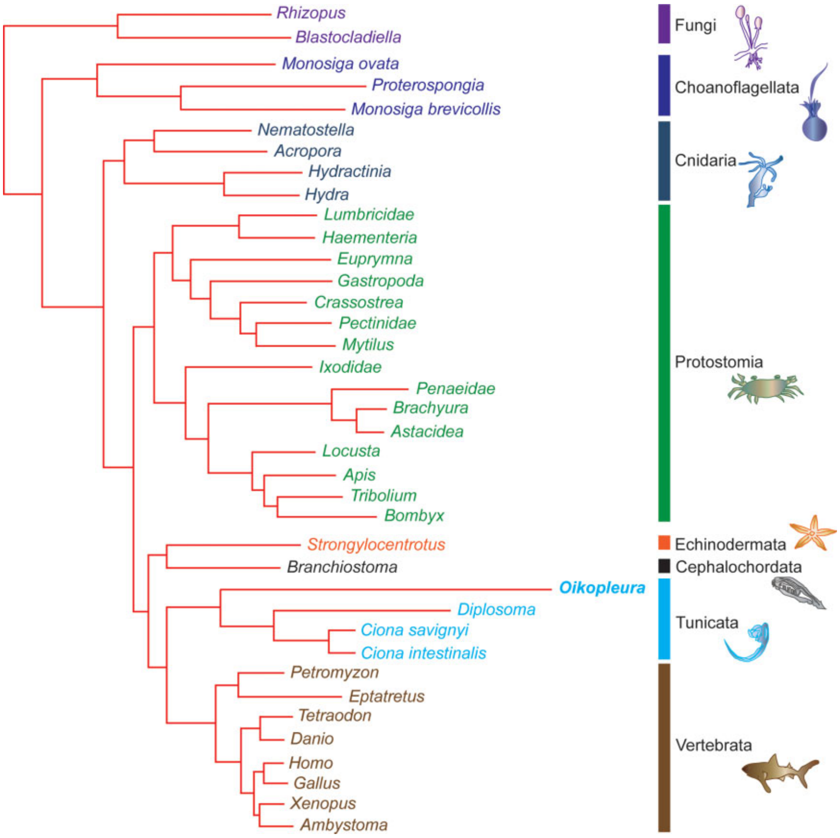


Figure 1. Phylogenetic analysis using genomic data shows *O. dioica* to be the closest living relative of vertebrates. Figure adapted from [4] with permission.

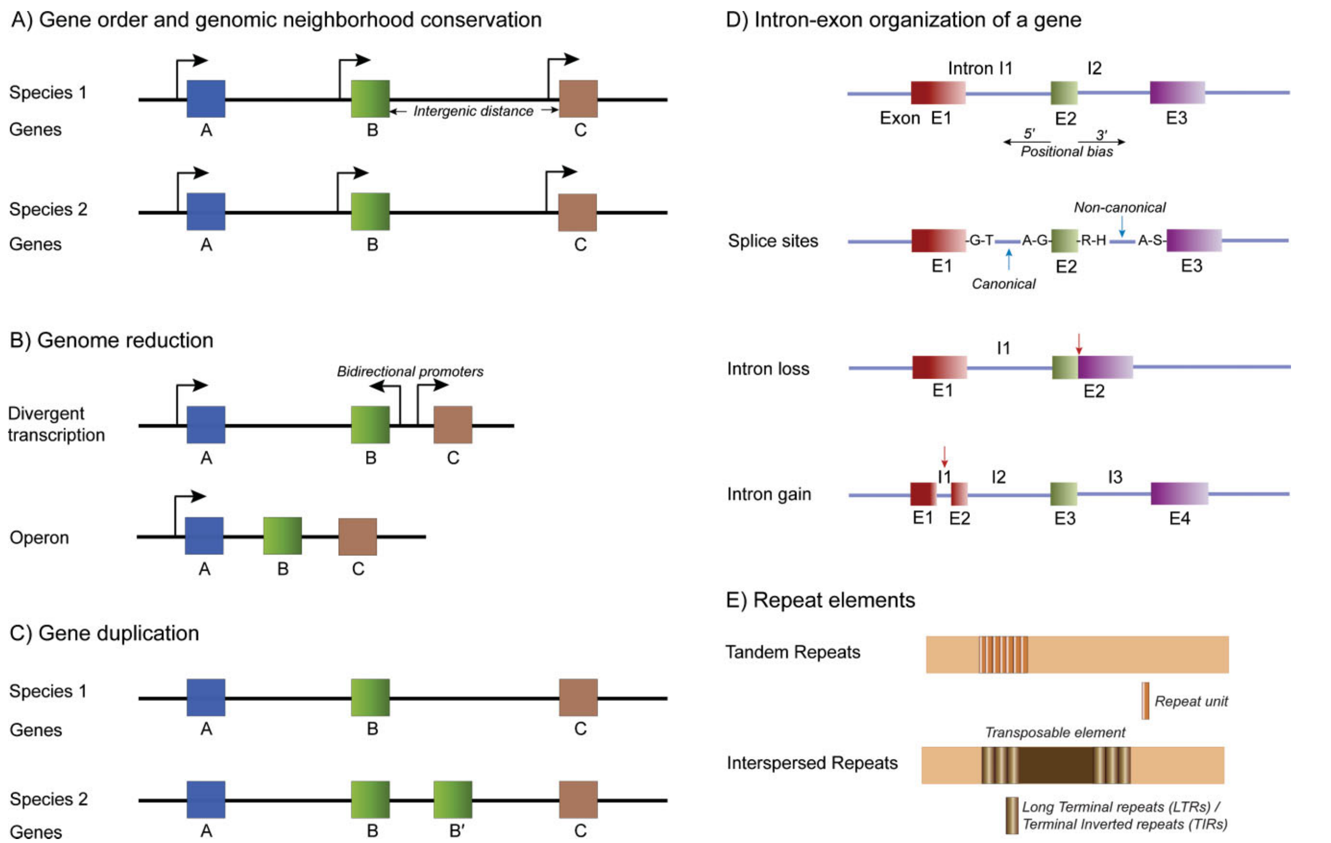


Figure 2. Salient features of a eukaryote genome. **A:** Conservation of gene order and genomic neighbourhood between two species. Changes in genomic neighbourhood influence transcription regulation and may facilitate speciation. **B:** Reduction in genome driven by compaction. All genes are retained with depletion of the intergenic regions. This is brought about by divergent transcription using bidirectional promoters and organization of functionally relevant genes into operons. **C:** Gene duplication resulting from segmental duplication. The gene B is duplicated to B′ in species 2, which might acquire a new function. **D:** Intron-exon organization in a eukaryotic gene. Eukaryotic genomes show a 5′-biased distribution of introns. Old introns in eukaryotic genomes are mostly canonical with 5′ GT-AG 3′ splice sites. A minor fraction of introns (mostly newly acquired) are non-canonical with non-GT 5′ splice sites (RH) including GA, GC or AT and (AS) AG or AC at 3′ end. Loss of introns also leads to reduction in genome size. In the figure, intron 2 (I2) is lost. In eukaryotes there is a general 3′ bias in intron loss. Here, intron 1 is gained due to the insertion within E1, leading to the formation of a new exon (E2). The red arrows indicate intron loss and intron gain, as appropriate. **E:** Repeat elements in eukaryotic genomes. Repeats are characterized by repeating unit (e.g. CAG) and repeat length, which is the number of times the unit is repeated. Tandem repeats are continuous repeats with a repeating unit size ranging from 1 to 100 bp (micro- and minisatellites). Interspersed repeats include transposable elements (TEs). TEs have long terminal repeats (Type I TEs or retrotransposons) or terminal inverted repeats (Type II TEs or DNA transposons) typically with a repeat unit length >100 bp.

implicated in speciation (Fig. 2A) [12, 13]. In this context, *O. dioica* was previously reported to show extensive gene loss, lack of synteny conservation and disintegration of the *Hox* cluster [14]. Consistently, the complete genome analysis showed no conservation of chromosomal synteny with invertebrates with exceptions such as non-coding regions of developmentally regulated genes and demonstrated modest local gene order conservation when compared to humans [2]. These observations suggest that constraints that maintain gene order in metazoans may actually be relaxed in *O. dioica*. Whether this is an abrupt or an incrementally mediated change during the course of evolution is unclear. Studying why such constraints were relaxed might help understand the need for such extensive rearrangement. Alternatively, gene order conservation in other metazoans might be a result of passive and slow evolution for most of the genome, which is in contrast to *O. dioica*. Conservation or changes in the genomic neighbourhood compared to immediate ancestors in *O. dioica* also needs to be investigated.

Genome reduction and short intergenic distances

In general, eukaryotic genomes are large, consisting of gene-dense and gene-sparse regions, which are defined by intergenic distances. Reduction in genome size may be driven by packaging genes into a smaller space (compaction) or by loss of genes (elimination) [15]. Genome compaction would lead to retention of almost all genes with an increase in the gene density and reduction of intergenic distances (Fig. 2B). As observed in microsporidia and nucleomorphs, genome compaction may also lead to overlapping transcription, with transcripts either initiating within the upstream gene or terminating within or beyond the downstream gene or both [16].

On the other hand, elimination of genes reduces the coding capacity of the genome and might compromise some basic processes. Notably, genome reduction in eukaryotes has predominantly been reported for intracellular parasites or endosymbionts. *O. dioica* has a small genome comprising of 70 megabases encoding ~18 000 predicted genes, compared to 117 and 174 Mb of the tunicate ascidians *Ciona intestinalis* and *C. savignyi*, respectively. The reduction in genome size might have resulted from both compaction and elimination. The organization of genes into operons, reduced intergenic distances, small introns and low density of the TEs might have contributed to genome compaction. Apparent loss of genes such as the notochord genes may also contribute to the reduced genome size. Additionally, a minimized immune system, the lack of the minor spliceosome and the absence of non-homologous end joining (NHEJ) pathway genes could have been a direct consequence of genome reduction due to gene elimination [2]. Highly reduced intergenic distances indicate the possibility of divergent transcription using bidirectional promoters [10] in *O. dioica*, which remains to be investigated. However, some developmentally regulated genes have relatively large introns and intergenic distances, which might be due to the need to maintain extensive *cis*-regulatory elements [9, 17, 18].

Operonic organization of genes

Exceptions to the monocistronic eukaryotic genes include the polycistronic genes regulated by a single regulatory element (operon) observed in kinetoplasts of protists, ascidians, platyhelminthes, flies and nematodes (Fig. 2B) [19, 20]. *O. dioica* can be added to this extending list of exceptions with 1293 operons. Although ascidians (close relatives of *O. dioica*) show operon organization, collinearity of genes within the operons is not conserved. In eukaryotes, organization of operons containing functionally related genes was first observed in nematodes. Such an organization, with genes involved in basic processes such as RNA processing, protein modification and transport, is observed in *O. dioica* operons [2]. An extensive analysis of operons from different lineages might provide insights into the constraints that drive the evolution of eukaryotic operons and into the mechanisms by which they are processed into monocistronic transcripts [20].

Lineage-specific gene duplications

Gene duplication has been proposed to be a major driving force in the evolution of eukaryotic genomes. It can result from either segmental duplication or genome duplication, increasing the copy number of the genes. In addition to recombination, gene duplication also drives the evolution of the protein repertoire in several eukaryotes (Fig. 2C) [21]. Two rounds of whole genome duplication (WGD) occurred at the base of the vertebrate evolution, leading to a large scale genome reorganization [22]. This has introduced an overwhelming functional consequence affecting a majority of signalling genes and transcription factors involved in development [23, 24]. Large families of proteins might have been ancestrally derived from, or resulted from, lineage-specific expansions (LSEs) [25, 26]. *O. dioica* shows a massive retention of duplicates for developmental genes, which is exceptional among invertebrates. In addition, large families of domains have been over-represented in the proteome of *O. dioica* (Table 1). The high abundance of certain superfamilies of protein domains suggests that they have resulted from gene duplication events [27].

The unusually high abundance of certain superfamilies in *O. dioica* represents LSE (Table 2). LSE contributes to both the

Table 1. Protein family assignments in *O. dioica*

Domains	Proteins	Superfamilies	Known functions
812	693	P-loop containing nucleoside triphosphate hydrolases	Found in proteins that often perform functions that assist in the assembly, operation, or disassembly of protein complexes and requires NTPs for its activity
459	149	EGF/laminin	Non-collagenous proteins that mediate cell adhesion, growth migration, and differentiation
453	446	Protein-kinase like (PK-like)	Seen in proteins that catalyze the phosphotransfer reaction fundamental to most signalling and regulatory processes
363	99	TSP-1 type 1 repeat	Contained in proteins that modulate cell adhesion and mediate cell-cell interaction
299	73	Immunoglobulin	Proteins containing immunoglobulins have diverse functions including mucosal immunity
259	153	Beta-beta-alpha zinc fingers	Contained in Zn finger transcription factors that aid in sequence specific DNA binding
254	92	Growth factor receptor domain	Found in proteins involved in signal transduction by receptor tyrosine kinases
242	209	ARM (Armadillo) repeat	Seen in proteins involved in intracellular signalling and cytoskeletal regulation
235	182	RNA-binding domain	Contained in proteins that regulate RNA transport and metabolism including splicing
233	201	Trypsin-like serine proteases	These proteases cleave peptide bonds following a positively charged amino acid residue such as arginine and lysine

The table presents the superfamilies with the highest frequency in *O. dioica* proteome. Protein superfamilies for *O. dioica* have been obtained from Superfamily database [68].

adaptation of the organism to the environment during evolution and the unique biology associated with the organism. For example, building the house requires extensive interaction between different cells and cell types and the function of several over-represented families are involved in cell adhesion e.g. EGF/lamin, and the thrombospondin type 1 (TSP-1) repeat in the extracellular matrix milieu (Tables 1 and 2). Oikosin 1 with Cys domain repeats is known to be important in the molecular patterning of the oikoplastic epithelium that generates the 'house' for filter feeding [28]. Comparisons with other chordates showed LSE of invertebrate chitin-binding domains and barwin-like endogluconases, which are involved in chitin binding and glycoside hydrolysis that might help in defence. Since WGD is believed to have occurred at the base of vertebrate lineage, the relationship between the temporal events leading to the massive reorganization of the eukaryotic genome, as seen in *O. dioica*, and the WGD seen in vertebrates needs to be studied thoroughly.

Divergent non-coding RNA pool

Recent genome-wide transcriptome analyses have led to the identification of a huge repertoire of ncRNAs [29]. Of these, the regulatory ncRNAs can be classified based on their size as the small ncRNAs, including microRNAs (miRNAs), endogenous small interfering RNAs (endo-siRNAs), PIWI-interacting RNAs (piRNAs) and large intergenic ncRNAs (lncRNAs) [30]. These regulate gene expression in a spatio-temporal manner by diverse mechanisms ranging from transcriptional or translational repression to chromatin modification. These affect different biological processes such as embryogenesis, cell-fate specificity and repression of retrotransposons [31]. *O. dioica* possesses miRNA biogenesis machinery and produces miRNAs throughout its life cycle with some of them stocked as maternal determinants, whereas the sex-specific miRNAs apparently aid gonadal differentiation [32]. While the majority of mammalian miRNAs are encoded by introns, miRNA loci of *O. dioica* are located in antisense orientations to protein-coding genes. Such an accommodative mechanism of miRNA transcription could have evolved as a result of severe genome compaction. Compared to ascidians, *O. dioica* has lost and acquired many miRNA families, suggesting a large scale reshaping of the miRNA repertoire [32]. With the depletion of intergenic distances, a severe loss of lncRNAs might be anticipated in *O. dioica*. Nevertheless, with the developmentally regulated transcription factors having huge intergenic distances in *O. dioica*, it would be interesting to investigate if some key lncRNAs that are known to play a vital role in embryonic development [33, 34] are conserved in tunicates.

Introns in *O. dioica*

The spread of introns through eukaryotic genome evolution has been proposed to have given rise to the nucleus-cytosol compartmentalization, thus breaking the prokaryotic paradigm of spatially coupled transcription and translation [35]. Evolution of introns has been a topic of intense debate with contradicting views concerning when introns originated [36, 37]. Introns were initially recognized as sequences intervening between the protein-coding regions in DNA. However, with the increasing knowledge of the intronic location of regions encoding ncRNAs, the functional role of introns is beginning to be better appreciated [36, 38, 39]. Here we discuss functional and evolutionary aspects of introns and relate it to what is observed in *O. dioica*.

Highly divergent intron-exon organization

Considerable intron-retention is observed among eukaryotes [40, 41]. However, varying density, position and length of introns in eukaryotes suggest an ample gain or loss of introns or both across different lineages [42]. This loss or gain of introns coupled with their slow turnover determines intron-exon organization (Fig. 2D). Eukaryotic genomes show an excess of introns in the 5' region (i.e. positional bias). This trend is more pronounced in genomes with intron paucity [43, 44]. *O. dioica* has short introns except in a few genes such as developmental genes [2, 45]. However, introns in *O. dioica* show a 3'-biased distribution, observed only in genes contained in operons. This deviation from the observed trend appears to be a result of 5'-biased intron loss, especially in the operons, which are highly expressed. Besides this, *O. dioica* also shows a high divergence and considerable variability in intron-exon organization, a feature attributed to genome compaction and its short life cycle [40, 46].

High intron turnover and acquisition of new introns

Three modes of intron dynamics leading to intron-exon structure evolution have been proposed: (i) balanced mode, suggesting an intron gain-loss balance, (ii) elevated intron loss and (iii) elevated intron gain indicating bursts of intron invasion [47]. *O. dioica* shows a very high intron turnover with most of the introns newly acquired. Moreover, there is a trend that the largest introns are more often old. Old introns are mostly phase 0 introns (i.e. lying before the first base of the codon), while most of the newly acquired introns are phase 1 (i.e. lying between the first and second base of the codon). While most of the old introns host canonical GT-AG splice sites, the majority of the new introns show non-canonical GA-AG splice sites, confirming the hypothesis that these were inserted when the current codon usage already existed (Fig. 2D). However, new introns with both canonical and non-canonical splice site, suggest that the dynamic process of intron creation in *O. dioica* has occurred at different times and not as a result of one burst of intron invasion.

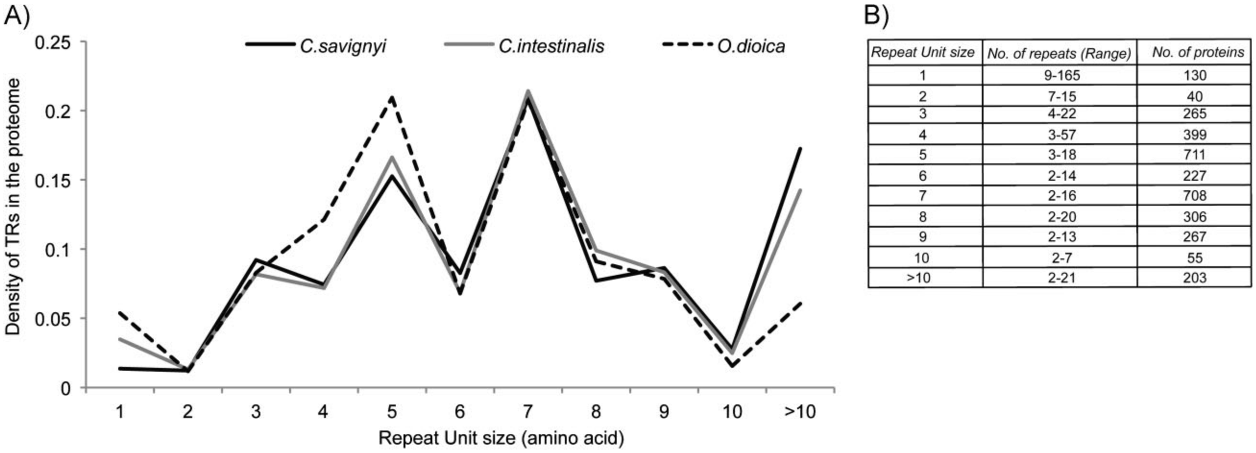
Lack of the minor spliceosome but highly frequent non-canonical introns

Intron splicing is initiated upon splice-site recognition by the spliceosome. Thus, the evolution of splice sites and that of the spliceosomes can be viewed to be coupled. Canonical introns with characteristic GT-AG splice sites are of very early origin in evolution and are present in the most common ancestors of living eukaryotes (Fig. 2D) [48]. These account for the vast majority of introns in eukaryotes and are spliced by the major U2-dependent spliceosomes. A minority of introns with atypical splice sites are excised by the minor U12-dependent spli-

Table 2. Unusual superfamily domains in *O. dioica* compared to other eukaryotes

Domains in *O. dioica*	Average in other genomes	Superfamily	Known functions
Over-represented domains			
149	5.1	Phospholipase A2, PLA2	Proteins that aid in adaptation. Cytosolic proteins with PLA2 mediate intracellular signalling in response to external stimulus
363	71.2	TSP-1 type 1 repeat	Contained in proteins that modulate cell adhesion and mediate cell-cell interaction
123	11.0	Galactose oxidase, central domain	Seen in extracellular enzyme galactose oxidase that catalyses the oxidation of D-galactose
459	198.1	EGF/laminin	Non-collagenous proteins that mediate cell adhesion, growth migration, and differentiation
233	59.8	Trypsin-like serine proteases	These proteases cleave peptide bonds following a positively-charged amino acid residue such as arginine and lysine
161	40.2	Kelch motif	Proteins with kelch motifs are involved in cytoskeleton function
174	59.7	C-type lectin-like	Proteins that contain C-type lectin domains are involved in cell-cell adhesion, immune response to pathogens and apoptosis
168	55.1	Spermadhesin, CUB (complement C1r/C1s, Uegf, Bmp1) domain	Found almost exclusively in extracellular and plasma membrane-associated proteins, many of which are developmentally regulated
24	2.5	Trefoil	Found in proteins (mostly mucins) that provide defence from microbes
18	1.9	Guanido kinase N-terminal domain	Seen in proteins that play a vital role in energy metabolism
Under-represented domains			
0	38.4	Zn2/Cys6 DNA-binding domain	Contained in proteins involved in arginine, proline, pyrimidine, quinate, maltose and galactose metabolism; amide and GABA catabolism; leucine biosynthesis
0	36.0	KRAB (Kruppel-associated box) domain	Members with KRAB domain are involved in transcriptional repression of RNA polymerase I, II, and III promoters, binding and splicing of RNA, and control of nucleolus function
259	754.5	Beta-beta-alpha zinc fingers	Contained in Zn finger proteins aiding in DNA binding
1	18.0	ACP (Acyl Carrier Protein)-like	Contained in proteins involved in fatty acid metabolism, polyketide antibiotics, biotin precursor, membrane-derived oligosaccharides, and activation of toxins
0	14.0	FAS1 (Fasciclin-like) domain	Found in extracellular cell adhesion proteins
0	5.8	Major-surface antigen p30, SAG1	Proteins with SAG1 mediate attachment of parasites to host cells and interface with the host immune response to regulate its virulence
0	19.7	DNA-binding domain	Recognizes double or single-stranded DNA
1	13.6	ACT-like	Found predominantly in basic helix loop helix transcription factors
2	26.6	DEATH domain	DEATH-domain containing proteins are mostly involved in immune response
0	5.2	Probable ACP-binding domain of malonyl-CoA ACP transacylase	Contained in proteins that are involved in fatty acid biosynthesis

Average domain representations in other genomes were derived from the genomes of 274 model eukaryotes including 98 animals, 102 fungi, 32 plants and 41 protists. Only one representative genome was considered for all organisms in this analysis.



Repeat Unit size	No. of repeats (Range)	No. of proteins
1	9-165	130
2	7-15	40
3	4-22	265
4	3-57	399
5	3-18	711
6	2-14	227
7	2-16	708
8	2-20	306
9	2-13	267
10	2-7	55
>10	2-21	203

Figure 3. Tandem repeats in *O. dioica* proteome. **A:** Density of amino acid tandem repeats in the proteome of *O. dioica*, *C. savignyi* and *C. intestinalis*. Non-overlapping tandem repeats in proteins of annotated genes in *O. dioica* were identified using T-REKS program with a threshold of similarity set at 0.7 [69]. **B:** Distribution of tandem repeats in *O. dioica* proteome.

ceosome complex (for more details see ref. [36]). As noted above, most of the introns in *O. dioica* are short and hence do not display long stretches of polypyrimidine tracts and the branch point consensus, which along with AG at the splice boundary comprise the 3′ splice sites in mammals. Non-canonical introns constituted by GA-AG, GC-AG and GG-AG splice sites with specific acceptor site (3′) are unusually frequent in *O. dioica*, comprised mostly by the newly acquired introns. However, *O. dioica* lacks the minor spliceosome and is proposed to have evolved a mechanism consisting of a single and permissive major spliceosome with U1SnRNP and U2AF, which recognizes both canonical and non-canonical donor and acceptor sites.

Alternative splicing generates different transcripts (isoforms) by deriving different combinations of exons from the same gene [49]. Although the number of genes in the genome appears to be fixed, alternative splicing aids in generating a more diverse transcriptome and proteome. Accordingly, alternative splicing is more prevalent in higher eukaryotes than in lower eukaryotes [50]. Given the smaller size of the genome, and the LSE of the RNA-binding domains (Table 2), which are important constituents of the splicing machinery, generation of isoforms might be an active process in *O. dioica*, which needs to be explored.

A 3′-biased intron loss in genes

No clear phylogenetic trend has been established for the intron loss or gain in eukaryotes. Intron loss might be mediated by a double recombination event between the genomic copy of a gene and the reverse-transcribed intronless cDNA of the corresponding spliced mRNA or genomic deletion or both [37]. While the former leads to precise mRNA excision and concerted loss of adjacent introns, genomic deletion results mostly in less accurate single intron loss, either deleting some part of the coding region or leaving residual intron sequences. As reverse transcription (RT) occurs from the 3′ to the 5′ end of mRNA, often resulting in incomplete transcripts, RT-mediated double recombination might lead to a 3′-biased intron loss, which might also cause, and result from, 5′-biased intron accumulation (Fig. 2D). However, lack of a 3′ bias in intron loss could be due to self-priming, which may randomly occur inside the transcripts and not at the polyA tail [51]. *O. dioica* shows a 3′ bias for accumulation of introns and loss of adjacent introns, especially in the highly transcribed operons. Taken together, a double recombination involving random self-priming, along with preferential conservation of 3′ introns due to selection or a higher likelihood of recombination at the 5′ end of genes might lead to 5′-loss bias [2]. The role of genomic deletion leading to intron loss in the *O. dioica* genome has yet to be explored.

A 5′-biased intron gain in genes

Several mechanisms have been proposed to drive the origin of new introns (Fig. 2D). These include intron transposition, transposon insertion, tandem genomic duplication, intron transfer from paralogues, conversion of type II introns to spliceosomal introns, creation of splice sites within introns and intron insertions from transcribed region by repair of double strand breaks (for more details see ref. [37] and Supplementary Material from ref. [2]). Nevertheless, these mechanisms are not mutually exclusive and exactly what and how many of these mechanisms aid intron gain in a genome requires rigorous investigation. Most of the introns in *O. dioica* are new. A 5′ bias in intron gain, especially in operons, has been suggested. Although *O. dioica* genome shows a scarcity of TEs, both in terms of quantity and variety, transposon insertions have acted as a primary source for novel introns [2]. In addition, the first reported candidates of novel introns generated through intron transposition or reverse splicing have been observed. This is a very exciting discovery and opens up several important questions about its prevalence in other genomes.

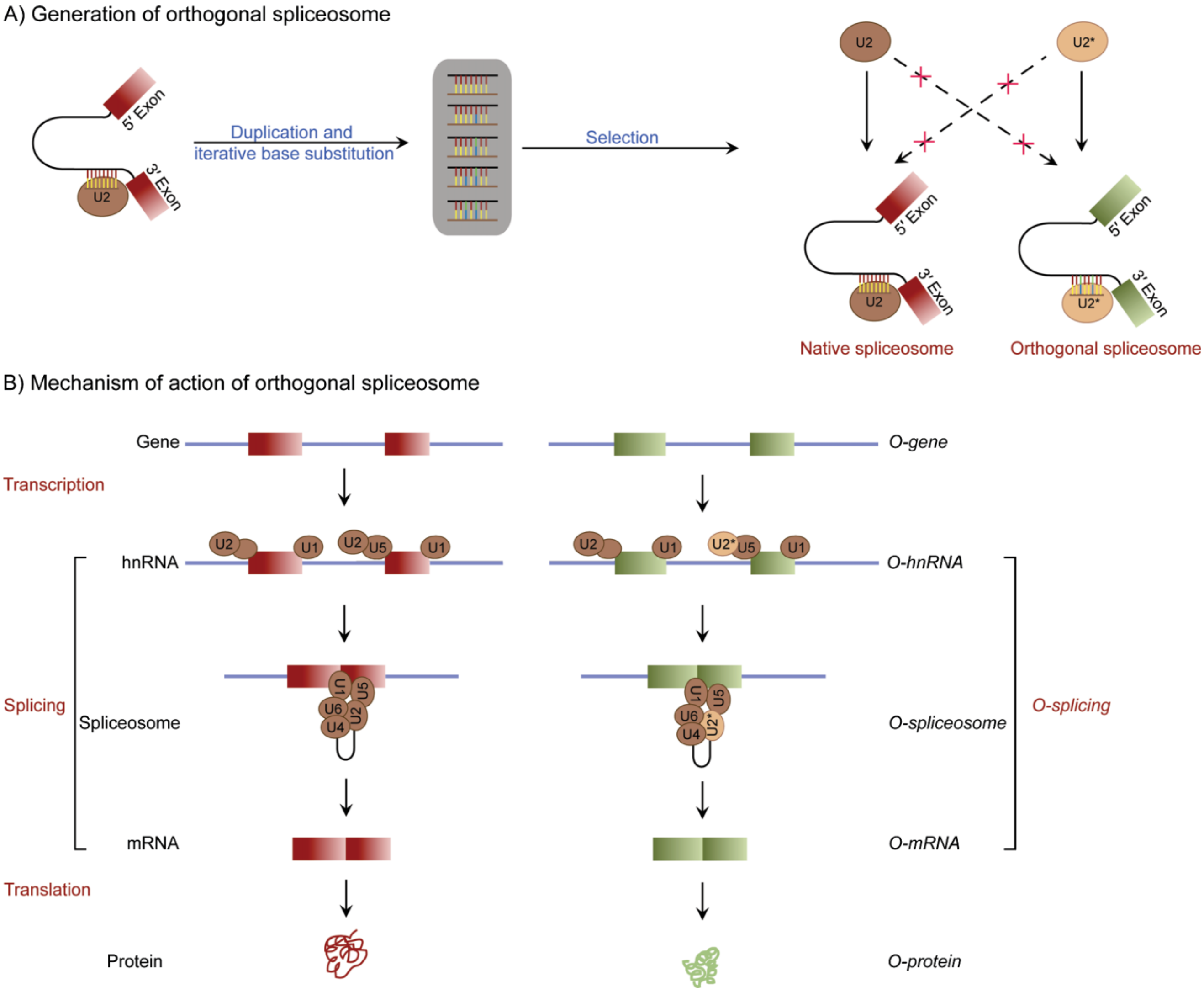


Figure 4. A schematic representation of how the knowledge obtained from *O. dioica* can be exploited for generating orthogonal spliceosome. **A:** Generation of orthogonal spliceosome by duplicating and mutating the binding site on the intron and the U2 recognition site as previously described [70]. The evolution of a single, permissive major spliceosome with U1SnRNP and U2AF, which recognize both canonical and non-canonical donor and acceptor sites in *O. dioica*, and the lack of a consensus branch point provide a great advantage for synthetic biology applications. The orthogonal gene containing the splice site and orthogonal U2 (U2*) can be genetically engineered to be under an inducible promoter, thereby providing the opportunity to regulate splicing.
B: The orthogonal spliceosome executes splicing without interfering with the host splicing mechanism as shown. It should be noted that in an unlikely event of cross-talk between the orthogonal U2 and the native mRNA or vice versa, the mis-spliced transcripts will most likely be degraded in the cell by the non-sense-mediated decay pathway.

Repeat elements in *O. dioica*

Repeating sequences are widely prevalent in all eukaryotic genomes. Earlier, owing to the limited knowledge of the biological importance and functions of repeats, they were believed to be junk or selfish DNA [52]. However, with increasing knowledge, the biological importance of repeat elements is being appreciated [53]. In the following sections, we review the important features of repeat elements and what the *O. dioica* genome adds to our existing understanding.

Retention of tandem repeats

Nearly 10–20% of eukaryotic genes contain continuous repeating sequences, commonly referred to as tandem repeats (TRs; Fig. 2E) [53]. Variable TRs are known to confer phenotypic variability. TRs accelerate evolution by influencing characteristics associated with genetic and epigenetic changes. Variable TRs with a repeat unit size of 6–100 nucleotides (minisatellites) are known to be hotspots for homologous recombination [54]. Given that *O. dioica* has a small but rapidly evolving genome, we checked whether the constraint on the genome size has compromised TR content of the genome or if TRs have been retained to the same extent as in other eukaryotes. We observed that 16.6% of annotated genes in *O. dioica* contain TRs, emphasizing the importance of TRs in eukaryotic genome. Comparable to that of tunicate ascidians, the proteome of *O. dioica* contains 14.8% proteins with TRs of varying repeat unit size and number of repeating units (Figs. 3A and B). As *O. dioica* genome has high mutation rates, the role of TRs in bringing this about, along with functional variability, is an interesting aspect for future research.

Scarcity of interspersed repeats

Eukaryotic genomes are abundant with various types of TEs [55]. They are primarily comprised of interspersed repeat elements (Fig. 2E). Integration of TEs often leads to disruption of protein-coding regions, chromosome breakage, illegitimate recombination and genome rearrangement and hence influences mutation rates. Since the successful integration of TEs often compromises the host fitness, host genomes have evolved various mechanisms to epigenetically suppress TEs [56]. Importantly, pervasive anti-sense transcription from loci encoding TEs is a classic signature of piRNAs and endo-siRNA that regulate TE invasion in several eukaryotes [57, 58]. The *O. dioica* genome shows few TEs, with an absence of most pan-animal TEs. The uneven distribution and low copy numbers suggest a tight regulation on the proliferation of TEs, potentially mediated through small ncRNAs. These mechanisms of regulation need further investigation.

Conclusions and outlook

Genome sequencing of different organisms has posed new challenges for deriving general principles underlying the evolution of eukaryotic genome architecture. *O. dioica* presents one such example, deviating a lot from the current understanding of genomic architecture. Detailed investigation of such violations would provide deeper insights into underlying mechanisms. Notably, *O. dioica* genome highlights the importance of non-adaptive forces, such as elevated mutation rates, in the evolution of genome architecture. It also provides us with an exciting opportunity to uncover the existence of multiple biological solutions to the same problem (e.g. the differences in requirements of splicing certain introns in this genome and the mechanism of reverse splicing). The discovery of such new biology can be exploited in synthetic biology applications such as gene silencing and in artificial control of splicing. The new understanding can also inspire designing orthogonal systems, for instance an orthogonal splicing system, that can control biological process through external means without interfering with the endogenous host processes (Fig. 4) [59–67]. In summary, sequencing the genomes of diverse organisms at the base of vertebrate evolution provides us with a better understanding of genome evolution and represents a great source for uncovering further biological novelties that can be potentially exploited for various benefits.

Acknowledgments
We acknowledge MRC for funding. We thank Kai Kruse, A. J. Venkatakrishnan, Andrew Deonarine and Marija Buljan for helpful comments. M.M.B. acknowledges the EMBO YI Programme and funding from BBSRC, partners of the ERASysBio+ initiative supported under the EU ERA-NET Plus scheme in FP7. S.C. is a career development fellow and acknowledges funding from ERASysBio+. J.G. acknowledges the funding from BBSRC (BB/G022771/1).

References

1. **Koonin EV.** 2009. Evolution of genome architecture. *Int J Biochem Cell Biol* **41**: 298–306.
2. **Denoeud F**, **Henriet S**, **Mungpakdee S**, **Aury JM**, et al. 2010. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* **330**: 1381–5.
3. **Bouquet JM**, **Spriet E**, **Troedsson C**, **Ottera H**, et al. 2009. Culture optimization for the emergent zooplanktonic model organism *Oikopleura dioica*. *J Plankton Res* **31**: 359–70.
4. **Delsuc F**, **Brinkmann H**, **Chourrout D**, **Philippe H.** 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**: 965–8.
5. **Holland LZ.** 2007. Developmental biology: A chordate with a difference. *Nature* **447**: 153–5.
6. **Canestro C**, **Postlethwait JH.** 2007. Development of a chordate anterior-posterior axis without classical retinoic acid signaling. *Dev Biol* **305**: 522–38.
7. **Stach T**, **Winter J**, **Bouquet JM**, **Chourrout D**, et al. 2008. Embryology of a planktonic tunicate reveals traces of sessility. *Proc Natl Acad Sci USA* **105**: 7229–34.
8. **Davila Lopez**, **Martinez M**, **Guerra JJ**, **Samuelsson T.** 2010. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLoS One* **5**: e10654.
9. **Kikuta H**, **Laplante M**, **Navratilova P**, **Komisarczuk AZ**, et al. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res* **17**: 545–55.
10. **Hurst LD**, **Pal C**, **Lercher MJ.** 2004. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* **5**: 299–310.
11. **Lercher MJ**, **Urrutia AO**, **Hurst LD.** 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* **31**: 180–3.
12. **De S**, **Teichmann SA**, **Babu MM.** 2009. The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. *Genome Res* **19**: 785–94.
13. **De S**, **Babu MM.** 2010. Genomic neighbourhood and the regulation of gene expression. *Curr Opin Cell Biol* **22**: 326–33.
14. **Seo HC**, **Edvardsen RB**, **Maeland AD**, **Bjordal M**, et al. 2004. Hox cluster disintegration with persistent anteroposterior order of expression in *Oikopleura dioica*. *Nature* **431**: 67–71.
15. **Keeling PJ**, **Slamovits CH.** 2005. Causes and effects of nuclear genome reduction. *Curr Opin Genet Dev* **15**: 601–8.
16. **Williams BA**, **Slamovits CH**, **Patron NJ**, **Fast NM**, et al. 2005. A high frequency of overlapping gene expression in compacted eukaryotic genomes. *Proc Natl Acad Sci USA* **102**: 10936–41.
17. **Woolfe A**, **Goodson M**, **Goode DK**, **Snell P**, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**: e7.
18. **Sandelin A**, **Bailey P**, **Bruce S**, **Engstrom PG**, et al. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**: 99.
19. **Blumenthal T**, **Gleason KS.** 2003. *Caenorhabditis elegans* operons: Form and function. *Nat Rev Genet* **4**: 112–20.
20. **Blumenthal T.** 2004. Operons in eukaryotes. *Brief Funct Genomic Proteomic* **3**: 199–211.
21. **Chothia C**, **Gough J**, **Vogel C**, **Teichmann SA.** 2003. Evolution of the protein repertoire. *Science* **300**: 1701–3.
22. **Nakatani Y**, **Takeda H**, **Kohara Y**, **Morishita S.** 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**: 1254–65.
23. **Huminiecki L**, **Heldin CH.** 2010. 2R and remodeling of vertebrate signal transduction engine. *BMC Biol* **8**: 146.
24. **Kassahn KS**, **Dang VT**, **Wilkins SJ**, **Perkins AC**, et al. 2009. Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates. *Genome Res* **19**: 1404–18.

Recently in press

25. **Lespinet O**, **Wolf YI**, **Koonin EV**, **Aravind L.** 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* **12**: 1048–59.
26. **Henikoff S**, **Greene EA**, **Pietrokovski S**, **Bork P**, et al. 1997. Gene families: The taxonomy of protein paralogs and chimeras. *Science* **278**: 609–14.
27. **Chothia C**, **Gough J.** 2009. Genomic and structural aspects of protein evolution. *Biochem J* **419**: 15–28.
28. **Spada F**, **Steen H**, **Troedsson C**, **Kallesoe T**, et al. 2001. Molecular patterning of the oikoplastic epithelium of the larvacean tunicate *Oikopleura dioica*. *J Biol Chem* **276**: 20624–32.
29. **Mattick JS.** 2009. The genetic signatures of noncoding RNAs. *PLoS Genet* **5**: e1000459.
30. **Pauli A**, **Rinn JL**, **Schier AF.** 2011. Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet* **12**: 136–49.
31. **Mattick JS**, **Amaral PP**, **Dinger ME**, **Mercer TR**, et al. 2009. RNA regulation of epigenetic processes. *BioEssays* **31**: 51–9.
32. **Fu X**, **Adamski M**, **Thompson EM.** 2008. Altered miRNA repertoire in the simplified chordate, *Oikopleura dioica*. *Mol Biol Evol* **25**: 1067–80.
33. **Ponting CP**, **Oliver PL**, **Reik W.** 2009. Evolution and functions of long noncoding RNAs. *Cell* **136**: 629–41.
34. **Mercer TR**, **Dinger ME**, **Mattick JS.** 2009. Long non-coding RNAs: Insights into functions. *Nat Rev Genet* **10**: 155–9.
35. **Martin W**, **Koonin EV.** 2006. Introns and the origin of nucleus-cytosol compartmentalization. *Nature* **440**: 41–5.
36. **Rodriguez-Trelles F**, **Tarrio R**, **Ayala FJ.** 2006. Origins and evolution of spliceosomal introns. *Annu Rev Genet* **40**: 47–76.
37. **Roy SW**, **Gilbert W.** 2006. The evolution of spliceosomal introns: Patterns, puzzles and progress. *Nat Rev Genet* **7**: 211–21.
38. **Mattick JS**, **Gagen MJ.** 2001. The evolution of controlled multitasked gene networks: The role of introns and other noncoding RNAs in the development of complex organisms. *Mol Biol Evol* **18**: 1611–30.
39. **Le Hir H**, **Nott A**, **Moore MJ.** 2003. How introns influence and enhance eukaryotic gene expression. *Trends Biochem Sci* **28**: 215–20.
40. **Raible F**, **Tessmar-Raible K**, **Osoegawa K**, **Wincker P**, et al. 2005. Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. *Science* **310**: 1325–6.
41. **Carmel L**, **Rogozin IB**, **Wolf YI**, **Koonin EV.** 2007. Patterns of intron gain and conservation in eukaryotic genes. *BMC Evol Biol* **7**: 192.
42. **Rogozin IB**, **Wolf YI**, **Sorokin AV**, **Mirkin BG**, et al. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol* **13**: 1512–7.
43. **Mourier T**, **Jeffares DC.** 2003. Eukaryotic intron loss. *Science* **300**: 1393.
44. **Lin K**, **Zhang DY.** 2005. The excess of 5’ introns in eukaryotic genomes. *Nucleic Acids Res* **33**: 6522–7.
45. **Seo HC**, **Kube M**, **Edvardsen RB**, **Jensen MF**, et al. 2001. Miniature genome in the marine chordate *Oikopleura dioica*. *Science* **294**: 2506.
46. **Edvardsen RB**, **Lerat E**, **Maeland AD**, **Flat M**, et al. 2004. Hypervariable and highly divergent intron-exon organizations in the chordate *Oikopleura dioica*. *J Mol Evol* **59**: 448–57.
47. **Carmel L**, **Wolf YI**, **Rogozin IB**, **Koonin EV.** 2007. Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res* **17**: 1034–44.
48. **Simpson AG**, **MacQuarrie EK**, **Roger AJ.** 2002. Eukaryotic evolution: Early origin of canonical introns. *Nature* **419**: 270.
49. **Keren H**, **Lev-Maor G**, **Ast G.** 2010. Alternative splicing and evolution: Diversification, exon definition and function. *Nat Rev Genet* **11**: 345–55.
50. **Artamonova II**, **Gelfand MS.** 2007. Comparative genomics and evolution of alternative splicing: The pessimists’ science. *Chem Rev* **107**: 3407–30.
51. **Niu DK**, **Hou WR**, **Li SW.** 2005. mRNA-mediated intron losses: Evidence from extraordinarily large exons. *Mol Biol Evol* **22**: 1475–81.
52. **Orgel LE**, **Crick FH.** 1980. Selfish DNA: The ultimate parasite. *Nature* **284**: 604–7.
53. **Gemayel R**, **Vinces MD**, **Legendre M**, **Verstrepen KJ.** 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet* **44**: 445–77.
54. **Vergnaud G**, **Denoeud F.** 2000. Minisatellites: Mutability and genome architecture. *Genome Res* **10**: 899–907.
55. **Wicker T**, **Sabot F**, **Hua-Van A**, **Bennetzen JL**, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**: 973–82.
56. **Slotkin RK**, **Martienssen R.** 2007. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8**: 272–85.
57. **Saito K**, **Siomi MC.** 2010. Small RNA-mediated quiescence of transposable elements in animals. *Dev Cell* **19**: 687–97.
58. **Jacquier A.** 2009. The complex eukaryotic transcriptome: Unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet* **10**: 833–44.
59. **Channon K**, **Bromley EH**, **Woolfson DN.** 2008. Synthetic biology through biomolecular design and engineering. *Curr Opin Struct Biol* **18**: 491–8.
60. **Khalil AS**, **Collins JJ.** 2010. Synthetic biology: Applications come of age. *Nat Rev Genet* **11**: 367–79.
61. **Agapakis CM**, **Silver PA.** 2009. Synthetic biology: Exploring and exploiting genetic modularity through the design of novel biological networks. *Mol Biosyst* **5**: 704–13.
62. **Carr PA**, **Church GM.** 2009. Genome engineering. *Nat Biotechnol* **27**: 1151–62.
63. **Lu TK**, **Khalil AS**, **Collins JJ.** 2009. Next-generation synthetic gene networks. *Nat Biotechnol* **27**: 1139–50.
64. **An W**, **Chin JW.** 2009. Synthesis of orthogonal transcription-translation networks. *Proc Natl Acad Sci USA* **106**: 8477–82.
65. **Leisner M**, **Bleris L**, **Lohmueller J**, **Xie Z**, et al. 2010. Rationally designed logic integration of regulatory signals in mammalian cells. *Nat Nanotechnol* **5**: 666–70.
66. **Chalancon G**, **Babu MM.** 2010. Nanobiotechnology: Scaling up synthetic gene circuits. *Nat Nanotechnol* **5**: 631–3.
67. **Chin JW.** 2006. Modular approaches to expanding the functions of living matter. *Nat Chem Biol* **2**: 304–11.
68. **de Lima Morais DA**, **Fang H**, **Rackham OJ**, **Wilson D**, et al. 2011. SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res* **39**: D427–34.
69. **Jorda J**, **Kajava AV.** 2009. T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics* **25**: 2632–8.
70. **Smith DJ**, **Konarska MM**, **Query CC.** 2009. Insights into branch nucleophile positioning and activation from an orthogonal pre-mRNA splicing system in yeast. *Mol Cell* **34**: 333–43.