Please correct your galley proofs and return them within 48h of receipt.
Note: The editors reserve the right to publish your article without your corrections if the proofs do not arrive in time.

Important

Please note that, due to the size constraints of emailed attachments, the attached proofs are low resolution; however high resolution versions will be used for printing.

Check the enclosed galley proofs very carefully, paying particular attention to the formulae, figures, numerical values, and tabulated data. An author query (AQ) in the border indicates unclear or missing information that requires your attention. For the complete query, please consult the list following the proofs in this file. Note that the author is liable for damages arising from incorrect statements, including misprints.

Your corrections should be provided as a list (e.g. in Word format), including the page number, column (left or right) and line at which the correction occurs. In addition, we ask that you provide a scanned pdf of the proof with corrections, to reduce the risk of errors.
Please return your corrections by email to the following address:

E-Mail: wileyvch.j@thomsondigital.com

Please limit corrections to printing errors; costs incurred for any further changes or additions will be charged to the author, unless such changes have been accepted by the editor.

Reprints may be ordered by filling out the accompanying form, which should be returned to the Editorial Office.
The most important correction marks

<table>
<thead>
<tr>
<th>Typesetting error</th>
<th>Correction mark in the text</th>
<th>Correction mark in the right-hand margin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wrong letter, wrong word</td>
<td>Metal type was the key to</td>
<td>p</td>
</tr>
<tr>
<td>Missing letter(s)</td>
<td>the intention of print.</td>
<td>mr</td>
</tr>
<tr>
<td>Missing word</td>
<td>and assembly into printed</td>
<td>tbs</td>
</tr>
<tr>
<td>Wrong word division, missing punctuation mark</td>
<td>was the only step a printer could take until recently</td>
<td>tnt</td>
</tr>
<tr>
<td>Defect letter, soiled text (in case of illegibility to</td>
<td>towards obtaining a surface</td>
<td>18 (2x)</td>
</tr>
<tr>
<td>be treated as a wrong letter or word)</td>
<td>from impressions can</td>
<td></td>
</tr>
<tr>
<td>Letters/words to be deleted</td>
<td>be made. Printing was from</td>
<td>19</td>
</tr>
<tr>
<td>Letters in wrong sequence</td>
<td>movable type stated</td>
<td>nl</td>
</tr>
<tr>
<td>Two words in wrong sequence</td>
<td>1440 around. Since</td>
<td></td>
</tr>
<tr>
<td>Several words in wrong sequence</td>
<td>in this of printing account</td>
<td></td>
</tr>
<tr>
<td>Wrong space, missing space</td>
<td>type founding place has no</td>
<td></td>
</tr>
<tr>
<td>Space between words too large or too small</td>
<td>we may dispense with</td>
<td></td>
</tr>
<tr>
<td>Characters not on a straight line</td>
<td>an account of how type</td>
<td></td>
</tr>
<tr>
<td>Space between lines too narrow</td>
<td>comes into being and starts with the type itself.</td>
<td></td>
</tr>
<tr>
<td>Space between lines too wide</td>
<td>Initially we may concentrate</td>
<td></td>
</tr>
<tr>
<td>Wrong typeface</td>
<td>on a single process -</td>
<td></td>
</tr>
<tr>
<td>Start new paragraph</td>
<td>letterpress - because metal type was once the</td>
<td></td>
</tr>
<tr>
<td>Undesired new paragraph</td>
<td>only printing surface</td>
<td></td>
</tr>
<tr>
<td>Undesired indentation</td>
<td>available to that process.</td>
<td></td>
</tr>
<tr>
<td>Wrong typeface</td>
<td>A historical view of printing would have to take into account a period in which type casting and type-setting remained unchanged. But type has</td>
<td></td>
</tr>
<tr>
<td></td>
<td>continuously</td>
<td></td>
</tr>
<tr>
<td></td>
<td>consolidated and extended</td>
<td></td>
</tr>
<tr>
<td></td>
<td>the domain of the printed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>word. The most important changes came with:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>automatic composing.</td>
<td></td>
</tr>
<tr>
<td>Blockade</td>
<td>(^1)H-NMR (MHz)</td>
<td>300</td>
</tr>
</tbody>
</table>
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 Chavali¹), David A. de Lima Morais²), Julian Gough²) and M. Madan Babu¹)

Recent sequencing of the metazoan Oikopleura dioica 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. Studies of the O. dioica genome suggest 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 miRNAs 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

Abbreviations: lncRNA, long non-coding RNA; LSE, lineage specific expansion; mirRNA, 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 (Fig. 1) displacing cephalochordates [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.
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 invert- 

erates with exceptions such as non-coding regions of devel- 

opmentally 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 (compac- 

tion) or by loss of genes (elimination) [15]. Genome compac-

tion would lead to retention of almost all genes with an 

increase in the gene density and reduction of intergenic dis-


tances (Fig. 2B). As observed in microsporidia and nucleo-

morphs, genome compaction may also lead to overlapping 

transcription, with transcripts either initiating within the 

upstream gene or terminating within or beyond the down-


---

**Figure 2.** Salient features of a eukaryote genome. **A:** Conservation of gene order and 

genomic neighborhood between two species. Changes in genomic neighborhood influ-

ence 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 trans-

posons) typically with a repeat unit length >100 bp.
stream 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>
<thead>
<tr>
<th>Domains</th>
<th>Proteins</th>
<th>Superfamilies</th>
<th>Known functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>812</td>
<td>693</td>
<td>P-loop containing nucleoside triphosphate hydrolases</td>
<td>Found in proteins that often perform functions that assist in the assembly, operation, or disassembly of protein complexes and requires NTPs for its activity</td>
</tr>
<tr>
<td>459</td>
<td>149</td>
<td>EGF/ laminin</td>
<td>Non-collagenous proteins that mediate cell adhesion, growth migration, and differentiation</td>
</tr>
<tr>
<td>453</td>
<td>446</td>
<td>Protein-kinase like (PK-like)</td>
<td>Seen in proteins that catalyze the phosphotransfer reaction fundamental to most signalling and regulatory processes</td>
</tr>
<tr>
<td>363</td>
<td>99</td>
<td>TSP-1 type 1 repeat</td>
<td>Contained in proteins that modulate cell adhesion and mediate cell-cell interaction</td>
</tr>
<tr>
<td>299</td>
<td>73</td>
<td>Immunoglobulin</td>
<td>Proteins containing immunoglobulins have diverse functions including mucosal immunity</td>
</tr>
<tr>
<td>259</td>
<td>153</td>
<td>Beta-beta-alpha zinc fingers</td>
<td>Contained in Zn finger transcription factors that aid in sequence specific DNA binding</td>
</tr>
<tr>
<td>254</td>
<td>92</td>
<td>Growth factor receptor domain</td>
<td>Found in proteins involved in signal transduction by receptor tyrosine kinases</td>
</tr>
<tr>
<td>242</td>
<td>209</td>
<td>ARM (Armadillo) repeat</td>
<td>Seen in proteins involved in intracellular signalling and cytoskeletal regulation</td>
</tr>
<tr>
<td>235</td>
<td>182</td>
<td>RNA-binding domain</td>
<td>Contains proteins that regulate RNA transport and metabolism including splicing</td>
</tr>
<tr>
<td>233</td>
<td>201</td>
<td>Tryptsin-like serine proteases</td>
<td>These proteases cleave peptide bonds following a positively charged amino acid residue such as arginine and lysine</td>
</tr>
</tbody>
</table>

The table presents the superfamilies with the highest frequency in *O. dioica* proteome. Protein superfamilies for *O. dioica* have been obtained from Superfamily database [70].
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 thrombospindin 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 endoglucanases, 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 spliceosome.
Table 2. Unusual superfamily domains in *O. dioica* compared to other eukaryotes

<table>
<thead>
<tr>
<th>Domains in <em>O. dioica</em></th>
<th>Average in other genomes</th>
<th>Superfamily</th>
<th>Known functions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Over-represented domains</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>149</td>
<td>5.1</td>
<td>Phospholipase A2, PLA2</td>
<td>Proteins that aid in adaptation. Cytosolic proteins with PLA2 mediate intracellular signalling in response to external stimulus</td>
</tr>
<tr>
<td>363</td>
<td>71.2</td>
<td>TSP-1 type 1 repeat</td>
<td>Contained in proteins that modulate cell adhesion and mediate cell-cell interaction</td>
</tr>
<tr>
<td>123</td>
<td>11.0</td>
<td>Galactose oxidase, central domain</td>
<td>Seen in extracellular enzyme galactose oxidase that catalyses the oxidation of α-galactose</td>
</tr>
<tr>
<td>459</td>
<td>198.1</td>
<td>EGF/laminin</td>
<td>Non-collagenous proteins that mediate cell adhesion, growth migration, and differentiation</td>
</tr>
<tr>
<td>233</td>
<td>59.8</td>
<td>Trypsin-like serine proteases</td>
<td>These proteases cleave peptide bonds following a positively-charged amino acid residue such as arginine and lysine</td>
</tr>
<tr>
<td>161</td>
<td>40.2</td>
<td>Kelch motif</td>
<td>Proteins with kelch motifs are involved in cytoskeleton function.</td>
</tr>
<tr>
<td>174</td>
<td>59.7</td>
<td>C-type lectin-like</td>
<td>Proteins that contain C-type lectin domains are involved in cell-cell adhesion, immune response to pathogens and apoptosis</td>
</tr>
<tr>
<td>168</td>
<td>55.1</td>
<td>Spermadhesin, CUB (complement C1r/C1s, Uegf, Bmp1) domain</td>
<td>Found almost exclusively in extracellular and plasma membrane-associated proteins, many of which are developmentally regulated.</td>
</tr>
<tr>
<td>24</td>
<td>2.5</td>
<td>Trefoil</td>
<td>Found in proteins (mostly mucins) that provide defence from microbes</td>
</tr>
<tr>
<td>18</td>
<td>1.9</td>
<td>Guanido kinase N-terminal domain</td>
<td>Seen in proteins that play a vital role in energy metabolism</td>
</tr>
</tbody>
</table>

| **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.
ceosome complex (for more details see reference 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, 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

<table>
<thead>
<tr>
<th>Repeat Unit size</th>
<th>No. of repeats (Range)</th>
<th>No. of proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9-165</td>
<td>130</td>
</tr>
<tr>
<td>2</td>
<td>7-15</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>4-22</td>
<td>265</td>
</tr>
<tr>
<td>4</td>
<td>3-57</td>
<td>399</td>
</tr>
<tr>
<td>5</td>
<td>3-18</td>
<td>711</td>
</tr>
<tr>
<td>6</td>
<td>2-14</td>
<td>227</td>
</tr>
<tr>
<td>7</td>
<td>2-16</td>
<td>706</td>
</tr>
<tr>
<td>8</td>
<td>2-20</td>
<td>306</td>
</tr>
<tr>
<td>9</td>
<td>2-13</td>
<td>267</td>
</tr>
<tr>
<td>10</td>
<td>2-7</td>
<td>56</td>
</tr>
<tr>
<td>&gt;10</td>
<td>2-21</td>
<td>203</td>
</tr>
</tbody>
</table>
double strand breaks (for more details see reference 37 and Supplementary Material from reference 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 described. This is a very exciting discovery and opens up several important questions about its prevalence in other genomes.

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.

Figure 4. A schematic representation of how 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 [69]. 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.
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

Prospects & Overviews


