PLASTICITY OF ANIMAL GENOME ARCHITECTURE UNMASKED BY RAPID EVOLUTION OF A PELAGIC TUNICATE

SUPPORTING TEXT

1. Sequencing, assembly and annotation

1.1. Genomic sequencing and assembly

The DNA used for shotgun sequencing was from sperm of approx. 200 males from a partially inbred line produced in culture conditions, and deriving from natural populations caught in the fjords near Bergen, Norway. Inbreeding consisted in 11 successive brother sister mating beginning from a single pair, and followed by three generations of mass spawning within the line to increase the biomass. The genome was sequenced using a Whole Genome Shotgun strategy on three plasmid libraries containing inserts calibrated at approximately 3, 8 and 12 kb generating more than 1,4 million reads . All data were generated by paired-end sequencing of cloned inserts using Sanger technology on ABI3730xl sequencers. All reads were assembled with Arachne (*S1*). We obtained 43,094 contigs that were linked into 34,559 supercontigs. The contig N50 was 10.3 kb, and the supercontig N50 was 37Kb. The total supercontig size was 148 Mb, remarkably distant of the expected size of 75 Mb. The initial assembly was improved and separated into two assemblies (reference and allelic) by applying a two steps protocol, as described in Figure S1. In a first time, we performed an all-against-all comparison (stringent alignment at the nucleotide level) of the supercontigs. From the resulting alignments, we decided to fuse overlapping supercontigs and to move to the allelic assembly supercontigs included in a larger supercontig. In a second step, we produced a draft annotation and we searched for block of syntenic genes. For each block we moved the largest supercontig to the reference assembly and the other one to the allelic assembly. The reference assembly was finally composed of 1,260 supercontigs with a cumulative size of 70.4 Mb (close to the expected size). The N50 was improved of around ten times. See details in Figure S1 and Table S1.

For a high-order assembly associating distinct scaffolds, a BAC library (15-20X coverage) was prepared from sperm of outbred individuals also from the Bergen area. Pairs of end sequences were produced for 7,500 inserts. Shotgun data near the end of contigs and scaffolds as well as in gaps between scaffolds were exploited through local walks using the in-house developed software MàD (Marche à Droite). Also, links between scaffolds of sex chrosomomes were validated through hybridizations of entire BAC clones high density arrays of 60bp tiles with 24 bp overlaps (Nimblegen). The assembly and its scaffolds are shown in Figure S2. Finally, 34 contigs from fosmids clones from *Oikopleura dioica* from the North American West Coast (collected near Bamfield, Canada, cultivated in Eugene, Oregon) and encompassing targeted developmentally regulated genes were cloned and sequenced for alignments with the genome sequence. Cloning the immediate environment of five of these genes after PCR amplification was also performed on DNA of a single individual from a population of *O. dioica* from Japan (near Osaka).

1.2. Genome Annotation

1.2.1. Construction of the training set

SNAP *ab inito* gene prediction software was used to create a clean set of *Oikopleura dioica* genes. This set was used to train gene prediction algorithms and optimize their parameters, and to calibrate whole genome comparisons. SNAP was launched with the *Caenorhabditis elegans* configuration file, and only models with every introns confirmed by at least one *Oikopleura dioica* cDNA were kept. Moreover models that contained at least one exon that overlapped a cDNA intron were rejected. Finally, we obtained an initial set of 1,882 genes and 6,822 exons. Three hundred models were randomly selected to create the clean training set of *Oikopleura dioica*.

1.2.2. Repeat Masking

Most of the genome comparisons were performed with repeat masked sequences. For this purpose, we searched and masked sequentially several kinds of repeats:

- known repeats and transposons available in Repbase with the Repeat masker program (*S2*)
- tandem repeats with the TRF program (*S3*)
- *ab initio* detection : RepeatScout (*S4*)

1.2.3. Exofish comparisons

Exofish (*S5*) comparisons were performed at the Genoscope, with the Biofacet software package from Gene-IT (*S6*). When ecores (Evolutionarily COnserved REgions) were contiguous in the two genomes, they were included in the same ecotig (contig of ecores) (5). Exofish comparisons were performed between *Oikopleura dioica* and four other organisms: *Tetraodon nigroviridis*, *Strongulocentrus purpuratus*, *Ciona savignyi* and *Ciona intestinalis*. HSPs were filtered according to their length and percent identity.

1.2.4. Genewise

The Uniprot (*S7*) database was used to detect conserved genes between *Oikopleura dioica* and other species. As Genewise (*S8*) is time greedy, the Uniprot database was first aligned with the *Oikopleura dioica* genome assembly using Blat (*S9*). Each significant match was chosen for a Genewise alignment. The default genewise gene parameter file was modified to take into account unusual splice sites of the *Oikopleura dioica* genes.

1.2.5. Geneid and SNAP

Geneid (*S10*) and SNAP (*S11*) *ab inito* gene prediction software were trained on 300 *Oikopleura dioica* genes from the training set.

1.2.6. *Oikopleura dioica* cDNAs

Three full-length-enriched cDNA libraries have been prepared from a cultured outbred population (large pools of 1: unfertilized eggs, 2: embryos at mixed stages from 1 to 3 hpf, 3: larvae 6-10hpf with 4 days-old adults) also from the Bergen area. Poly(A+) RNA were purified before cDNA synthesis. Approximately 180.000 cDNA clones were successfully sequenced. After assembly of 5' and 3' sequences when both were available, 177439 sequences could be aligned to the *Oikopleura dioica* genome assembly with the following pipeline, that was run independently on the reference and the allelic assemblies in order to allow each cDNA sequence to be mapped on both alleles. After masking of polyA tails and spliced leaders, the sequences were aligned with BLAT on the assembly and all matches with scores within 99% of the best score were extended by 5 kb on each end, and realigned with the cDNA clones using the Exonerate software (*S12*) to allow for non canonical splice sites, with the following parameters: --model est2genome --minintron 25 --maxintron 15000 --gapextend -8 --dnahspdropoff 12 --intronpenalty -23.

This procedure defined transcript models with a large fraction (more than 10%) of introns displaying non-canonical (non GT-AG) splice sites. The very vast majority of non canonical introns have the consensus GA-AG, GC-AG or GG-AG.

1.2.7. Oikopleura dioica spliced leader detection

To detect genes processed by transsplicing, we searched the *Oikopleura dioica* spliced leader "ACTCATCCCATTTTTGAGTCCGATTTCGATTGTCTAACAG" (*S13*) in the 5' unmapped portion of cDNA sequences using a Smith and Waterman alignment (*S14*). Since the cDNA sequencing strategy did not allow to reach the 5' end of the transcripts, the cDNAs were considered as being transspliced when they contained at least 15 nucleotides from the spliced leader (at the 3' end of the spliced leader). Overall, 25% of the cDNA clusters were detected as being transspliced. The position where the first nucleotide after she spliced leader was mapped on the genome was provided to the automatic annotation software (*S15*) as a transcript start signal.

1.2.8. Tunicate ESTs

A collection of ~1.500.000 public ESTs (from the tunicate clade) was first aligned with the *Oikopleura dioica* genome assembly using Blat (*S9*). This database was composed of public mRNAs downloaded from the NCBI (*S16*). To refine Blat alignment, we used Est2Genome (*S17*). Each significant match was chosen for an alignment with Est2genome. Blat alignments were made using default parameters between translated genomic and translated ESTs.

1.2.9. Integration of resources using GAZE

All the resources described here were used to automatically build *Oikopleura dioica* gene models using GAZE (*S15*). Individual predictions from each of the programs (Geneid, SNAP, Exofish, Genewise, Est2genome and Exonerate) were broken down into segments (coding, intron, intergenic) and signals (start codon, stop codon, splice acceptor, splice donor, transcript start, transcript stop).

Exons predicted by *ab initio* software, Exofish, Genewise, Est2genome and Exonerate were used as coding segments. Introns predicted by Genewise and Exonerate were used as intron segments. Intergenic segments created from the span of each mRNA, with a negative score (coercing GAZE not to split genes). Predicted repeats were used as intron and intergenic segments, to avoid prediction of genes coding proteins in such regions.

The whole genome was scanned to find signals (splice sites, start and stop codons). In order to annotate correctly the genes containing non-canonical splice sites, all G* (GT, GA, GC and GG) donor sites were authorized. Additionnaly, transcript start and stop signals were extracted from the spliced leader positions and ends of mRNAs (polyA tail positions).

Each segment extracted from a software output which predicts exon boundaries (like Genewise, Exonerate or *ab initio* predictors), was used by GAZE only if GAZE chose the same boundaries. Each segment or signal from a given program was given a value reflecting our confidence in the data, and these values were used as scores for the arcs of the GAZE automaton. All signals were given a fixed score, but segment scores were context sensitive: coding segment scores were linked to the percentage identity (%ID) of the alignment; intronic segment scores were linked to the %ID of the flanking exons. A weight was assigned to each resource to further reflect its reliability and accuracy in predicting gene models. This weight acts as a multiplier for the score of each information source, before processing by GAZE. When applied to the entire assembled sequence, GAZE predicted 17,113 gene models on the reference assembly and 13,527 gene models on the allelic assembly. The final proteome composed of 18,020 gene models was obtained by adding 907 gene models of the allelic assembly that were not present in the reference assembly.

1.3. Identification of orthologous genes

We identified orthologous genes with 8 species : *Ciona intestinalis* (*S18*), *Drosophila melanogaster* (*S19*), *Caenorhabditis elegans* (*S20*), *Ciona savignyi* (*S21*), *Homo sapiens* (*S22*), *Strongylocentrotus purpuratus* (*S23*), *Nematostella vectensis* (*S24*) and *Branchiostoma floridae* (*S25*). Each pair of predicted gene sets was aligned with the Smith-Waterman algorithm, and alignments with a score higher than 300 (BLOSUM62, gapo=10, gape=1) were retained. Two genes, A from genome GA and B from genome GB, were considered orthologs if B is the best match for gene A in GB and A is the best match for B in GA.

1.4. Protein domain analysis

InterProScan (*S26*) was run against all *Ciona intestinalis*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Ciona savignyi*, *Homo sapiens*, *Strongylocentrotus purpuratus* and *Oikopleura dioica* proteins. Matches which fulfilled the following criteria were retained:

- match is tagged as “True Positive” by InterProScan (status=T) ;
- match with an e-value less or equal to 10^{-1}.

A total of 2,926 InterPro domains (with IPR number) were found in *Oikopleura dioica*, and correspond to 8,580 *Oikopleura dioica* proteins (Table S2).

1.5. Functional annotation

1.5.1. Enzyme annotation

Enzyme detection in predicted *Oikopleura dioica* proteins was performed with PRIAM (*S27*), using the PRIAM July 2004 ENZYME release. A total of 734 different EC numbers, corresponding to enzyme domains, are associated with 4,700 *Oikopleura dioica* proteins. Therefore, about 26% of *Oikopleura dioica* proteins contain at least one enzymatic domain.

1.5.2. Association of metabolic pathways with enzymes and *Oikopleura dioica* proteins

From EC numbers, potential metabolic pathways were deduced using the KEGG pathway database (*S28*). Links between EC numbers and metabolic pathways were obtained from the KEGG website. Using this file and the PRIAM results, 2,024 (of the 4,700) *Oikopleura dioica* proteins were assigned to 189 pathways. Following the KEGG pathway hierarchy, pathways from the same family were grouped together. For instance, glycolysis and TCA cycle belong to carbohydrate metabolism. Using this method, the different pathways found in *Oikopleura dioica* define 38 pathway families.

2. Allelic variation and evolutionary parameters

2.1. Comparison of reference and allelic scaffolds

We used BLAT to compare allelic and reference scaffolds with length greater than 10 kb, which corresponds to 1184 allelic scaffolds (33Mb) and 652 reference scaffolds (68 Mb). We built aligned regions from the blat output, keeping together adjacent aligned portions that were separated by unaligned portions smaller than 5,000 nt on both sequences or smaller than 10,000 nt on one of the sequences and smaller than 1,000 nt on the other sequence, so that the resulting aligned regions are not splitted by repeats. The aligned regions span 32 Mb on the reference scaffolds (25Mb in aligned portions, 7 Mb in unaligned portions) and 30 Mb on the allelic scaffolds (25Mb in aligned portions and 5 Mb in unaligned portions). The average %id in the 25Mb of aligned portions is 98.5%. Some of the reference regions are aligned with several allelic regions: in total, 27.7 Mb of the reference assembly (41%) is polymorphic, *i.e.* is matching one or more (up to 4) allelic scaffolds. 24 Mb have two alleles, 3.5Mb have 3 alleles, 134 kb have 4 alleles and 9kb have 5 alleles.

The aligned regions were used to detect short (<50 nt, but mainly a few bp), and large (>=50nt) indels between the allelic and reference scaffolds. We identified 33394 short indels (in 30.7 Mb of aligned regions, after removing intercontig gaps: 1088 indels per Mb), and 3059 large indels (100 indels per Mb). As expected, insertions of transposable elements are avoided in CDS regions compared to other compartments and strongly avoided in intergenic regions of operons (Table S3). We quantified the number of insertions that are likely to inactivate genes, *i.e.* insertions in CDS that are not multiple of 3 or contain stops in phase: 533 genes contain polymorphic insertions that are likely to inactivate one of their alleles, unless they can be spliced out from the pre-mRNA.

2.2. Polymorphism in the genome sequence and estimation of population mutation rate

2.2.1. Existence of two major haplotypes

Despite 11 successive sib-matings, an appreciable level of allelic polymorphism is recovered in the genome sequence. Candidate explanations are either contamination of the line during the inbreeding process, or a “resistance to homozygosity” due to a powerful system of balanced lethals. The contamination by males can most likely be ruled out, as we observe virtually no polymorphism for scaffolds of the Y chromosome region. The second explanation appears probable, as we observe within large scaffolds the expected altermance of non-polymorphic and polymorphic regions (SNPs). These regions appear to constitute real blocks shared by distinct individuals of the inbred line as

shown by the linkage analysis of 5,513 polymorphic markers. These are for one part 3,780 exons chosen in the length range 100-300bp, all validated by EST and by at least two shotgun reads, whose sequences differ due to SNPs between the reference and allelic assemblies. The other part are 1,733 sites of polymorphic indels chosen in the length range 30-300bp. After aligning the shotgun dataset on all those markers with a 100% identity level constraint, a total of 4,085 linkages between shotgun reads matching distinct markers are found, with 2,923 of them at an expected distance of 3kb, and 1,162 at a distance of 8 or 12 kb. The numbers of linkages in both assemblies are almost identical, suggesting an equal representation of two major haplotypes in the inbred line. More than 98% of the linkages are within one or the other assembly, suggesting that genome assembly faithfully distinguished and reproduced those haplotypes. Then, 404 pairs of distinct polymorphic sites of the genome were linked at least two times by independent shotgun clones, and over both assemblies. Of those, 402 pairs of sites showed only intra-assembly links, while two showed inter-assembly links in addition to intra-assembly links. One of these two cases is due to a single shotgun read, and therefore not more than one case may result from recombinants between both major haplotypes (we cannot exclude that it results from a third minor haplotype).

2.2.2. Estimation of the population mutation rate

Given that the distinct blocks of high heterozygosity are intertwined with largely homozygous stretches, the main haplotypes in the genome sequence may be considered equivalent to those of a single individual, though resulting from an inbreeding process, and the shotgun data may be used to estimate evolutionary parameters including the population mutation rate $4N_e\mu$ (*S29*). To estimate $4N_e\mu$, we identified longest blocks of heterozygosity, defined as clusters of maximum spacing between SNPs >500bp, and confirmed that the density of SNPs in them is homogeneous. (Changing the maximum spacing parameter from 200 bp to 2kb did not significantly affect the downstream calculations). We remapped all the shotgun reads to the reference genome assembly using SSAHA *(S30)*. From SSAHAs alignments, we prepared and converted its SAM output from the longest 100 blocks of heterozygozity (total length ~3 Mb) to the format required by mlRho software (standalone v. 1.5) (*S29*). We calculated $4N_e\mu$ with mlRho using only the positions with minimum coverage of 16x. (The results do not change significantly if we use 9-18x coverage).

The $4N_e\mu$ value computed for *Oikopleura* is high (0.0220+-0.003) compared to 0.0120 for *Ciona* and 0.0012 for *Daphnia* (*S29*). The estimation based on the amphioxus (*B. floridae*) genome polymorphism is even higher (0.0562). The result suggests that the *Oikopleura* population effective size is large and/or that the mutation rate per generation is high.

2.3. Evolution pressure on proteins:

Three analyses were performed to estimate the relative rates of non-synonymous and synonymous mutation and substitution.

A first analysis used 80 annotated complete or partial coding sequences from fosmid genomic clones of a population of *Oikopleura dioica* from Western Canada (Bamfield). These were aligned and compared with homologous sequences of the reference genome assembly, all validated by cDNA sequences. PAML was used to estimate dS pairwise for each pair of sequences. A weighted average dS value for these genes was 0.47. In this context dS is expected to be equal to theta, the pairwise sequence diversity at synonymous sites and should correspond to $4N_e\mu$. A weighted average of 0.47 is on the order of what is observed for divergences in Eutherian mammals from their last common ancestor, indicating a very long time of divergence for the *Oikopleura* strains, a very large mutation rate, and/or a very large effective population size. The weighted mean of dN is only 0.09, suggesting strong negative selection on these genes.

Two additional analyses were performed on alleles identified within *O. dioica* from Norway. One utilized 384 partially redundant ESTs from outbred individuals matching 47 of the above mentioned coding sequences. These gave a much lower estimate of dS (0.0075), with 0.12 for dN/dS. The other used 7,191 pairs of 100-300 bp long exons from both genome assemblies that were validated by ESTs. Of these, 3,780 showed nucleotide substitutions that were validated by alignment

with at least two shotgun reads. The estimation of dS is higher (0.025) and dN/dS estimate is 0.12 again.

To summarize, the difference between coding sequences was considerably higher in comparisons between strains of different oceans than within the Bergen gene pool. We ignore whether and how *Oikopleura dioica* is subdivided into multiple species, and therefore the population parameters estimated across distinct oceans should be considered cautiously. On the other hand, the inbred background of the Bergen genome sequence may also introduce some bias. Lastly, the sequences are from within a population and will include segregating variation together with fixed changes and this is expected to include more deleterious non-synonymous changes than will be fixed. With these reservations, the protein sequence analysis reveals for the Bergen genome a high value of dS, while all dN/dS estimates concur to indicate strong pressure of negative selection, overall suggestive of large population size.

3. Phylogenetic analyses of metazoan genomes, with focus on node robustness

3.1. Generation of datasets of orthologous genes

The proteomes from 26 organisms (*Acyrthosiphon pisum, Apis mellifera, Branchiostoma floridae, Brugia malayi, Caenorhabditis elegans, Capitella sp., Ciona intestinalis, Danio rerio, Daphnia pulex, Drosophila melanogaster, Gallus gallus, Helobdella robusta, Ixodes scapularis, Lottia gigantea, Monodelphis domestica, Monosiga brevicollis, Mus musculus, Nasonia vitripennis, Nematostella vectensis, Oikopleura dioica, Ornithorhynchus anatinus, Pediculus humanus, Strongylocentrotus purpuratus, Tribolium castaneum, Trichoplax adhaerens* and *Xenopus tropicalis*) were used. Groups of orthologous genes were determined with orthoMCL (*S31*) using default parameter values. Among the 46,954 clusters of genes, we retained the 5158 clusters in which at least 18 different species were present and in which the number of sequences was between equal and up to twice the number of species (i.e. at most an average of 2 in-paralogs per species were allowed). These clusters were aligned with FSA (*S32*), a software performing statistical alignments. Unambiguously aligned positions were extracted from these alignments according to the following criteria: a position must have a specificity ≥0.8, accuracy ≥0.5, certainty ≥0.5, consistency ≥0.5 and sensitivity ≥0.5 and contains less than 50% of gaps; only continuous blocks of more than ten unambiguously aligned positions were conserved. It is possible that clusters generated by orthoMCL do not only contain in-paralogs. We therefore used SCaFoS (*S33*) to retain only undisputable in-paralogs: we discarded all genes in which at least one species possesses paralogs with an evolutionary distance higher than 25% of the average distance between this species and all other species; distances were calculated using TREE-PUZZLE (*S34*) with the WAG+Γ model. When the paralogs met the criterion, the one with the lower distance to all other sequences was selected. Only 1482 of the original 5158 genes were conserved according to this stringent criterion. These genes were divided according to the number of species they contained (from 18 to 26) and SCaFoS was used to construct nine concatenations, one for each number of species (Table S4). The datasets are named S18…S26 according to the number of species present in individual genes. (note that all the concatenated datasets contain the same 26 species). Although the number of genes and of positions was variable (from 105 to 251 genes and from 22,732 to 67,054 positions), the concatenations are all of large size. As expected, the frequency of missing data increases from 3% for the concatenation of the genes present in 26 species to 35% for the concatenation of the genes present in 18 species.

3.2. Methods of phylogenetic analysis

To analyze our phylogenomic datasets, we used the CAT model (*S35*) with PhyloBayes version 2.3, which has been shown in various contexts to be less prone to LBA artefacts than other models (*S36, S37, S38*). Moreover, this site heterogeneous model has a better fit than site homogeneous models (e.g. WAG or GTR) for all phylogenomic datasets tested (*S35, S37, S38, S39*).

The heterogeneity of rate across sites was modelled using a gamma distribution (4 discrete categories). For the plain posterior estimation, two independent chains were run for a total number of 10,000 cycles (corresponding to ~900,000 generations) saving every ten cycles, and discarding the first 5,000 cycles (burn-in). Each cycle consists in a complicated series of updates of all components of the parameter vector, including an average of 20 topological updates. The posterior consensus tree was obtained by pooling the tree lists of four independent runs. For each node, we compared the posterior probabilities inferred from two independent chains. The maximum difference we observed ranged from 0 to 0.2 for the nine datasets; the average difference being 0 and 0.0042. For the most complete dataset (114 genes for which the 26 species were present), robustness was also evaluated using a standard bootstrap procedure (*S40*). To reduce computational load (by a factor of 4), rate heterogeneity was modelled with a Dirichlet Process (*S41*) instead of a Gamma distribution. We verified that the same topology and the same posterior probability were obtained by the two models (CAT+Γ and CAT+DP). Bootstrap percentages were obtained by running 100 independent pseudo-replicates, for 10,000 cycles each. Trees were collected after the initial burn-in period (5,000 cycles) and for each replicate the consensus of these trees was computed by Phylobayes. These 100 trees fed to CONSENSE (*S42*) to compute the bootstrap support values for each node. Phylogenetic trees were also inferred using a LG+F+Γ model with RAxML (*S43*), including 100 fast bootstrap replicates (-m PROTGAMMALGF -f a).

3.3. Results of phylogenetic analyses

The phylogeny obtained with the concatenation of the 114 genes universally present in our 26 selected genomes is in excellent agreement with current knowledge (*S44*), with the monophyly of e.g. Ecdysozoa, Protostomia, Bilateria. The monophyly of Olfactores (Tunicata+Vertebrata) is highly supported: bootstrap support (BS) of 100% for the CAT model and of 97% for the LG model. Several nodes receive limited support despite the use of 32,650 positions: monophyly of Chordata and of Eumetazoa and paraphyly of Deuterostomia. This lack of resolution was previously observed, and it would probably require a denser taxon sampling to solve this problem, as shown for Chordata (*S45*) and Eumetazoa (*S38*).

Although we cannot improve the current taxon sampling, we can take advantage of complete genome analysis to use another powerful approach to evaluate the inferred phylogeny: the corroboration among independent sets of characters. We chose the number of species present for an orthologous gene (from 18 to 26) in order to define nine independent gene partitions (Table S4). The nine phylogenies (Fig. S3) are well-supported (posterior probability [PP] of 1 for almost all the branches) and remarkably similar. For instance, phylogenies based on S24, S25 and S26 are identical. In fact, the trees differ only for the position of three species (see Table S5 for a summary of the PPs):

1) *Trichoplax*, sister-group to all other animals (S18, S22, S24, S25, S26) or to *Nematostella* (S19, S20, S21, S23),
2) *Strongylocentrotus*, sister-group to all other bilaterians (S19, S20, S21, S24, S25, S26), to Protostomia (S22) or to Chordata (monophyletic Deuterostomia, S18, S23),
3) *Acyrthosiphon*, sister-group to all other insects (S23) or to *Pediculus* (S18, S19, S20, S21, S22, S23, S24, S25, S26).

Interestingly, the three nodes that are unstable among the nine independent datasets do also receive moderate bootstrap support in the analysis of the S26 dataset (maximum BS of 87%). In contrast, three nodes (monophyly of Chordata, Theria and Endopterygota) receive bootstrap support below 90%, but are always recovered with a PP of 1 with the 9 independent datasets. This suggests that in the second case moderate bootstrap support is mainly due to an insufficient number of positions, whereas in the first case this is also due to methodological issues. For instance, all mammals evolved at a similar rate, whereas *Acyrthosiphon* evolved rather fast.

The inference with the LG model gave similar results (Table SC3), especially for the three most complete datasets (S24, S25, S26). However, results obtained with the LG model are more sensitive to long-branch attraction artefacts *(S46*) than the CAT model, as previously observed (e.g. (*S37*)). The monophyly of Ecdysozoa, Tunicata and Olfactores does not receive maximal bootstrap

support (Table S6), to the extent that these groups are not monophyletic (in particular for the S19 and S21). However, the very long branches of *Oikopleura* and of nematodes (Fig. S3) and the lack of realism of site-homogeneous models easily explain these problems.

In summary, our phylogenetic analysis of the 26 complete holozoan genomes strongly confirms the current view of animal phylogeny (in particular the monophyly of Chordata and of Olfactores) but leaves unsolved the question of the monophyly of Eumetazoa and of Deuterostomia), which would require additional taxon sampling and/or an improved model of sequence evolution.

4. The *Oikopleura dioica* mitochondrial genome: structure, codon usage and phylogenetic analyses

4.1. Cloning and characterization of the *O.dioica* mitochondrial genome

Surprisingly, no sequence of candidate mitochondrial genes was detected in the shotgun dataset, while they massively appeared in EST sequences. PCR-based cloning from genomic DNA suggested that this was due to unexpected oligo-dT stretches interrupting open reading frames. These are without exception inserted at TTTTTT sites of the coding sequence. Oligo-T insertions, through much shorter, were observed in mitochondrial genes of one nematode (*S47*). We believe they are corrected by RNA editing, rather than by an intron-like splicing, as potential intermediates of a deletion process are detected in ESTs. Due to cloning and sequencing difficulties, the mitochondrial genome was only partially reconstructed (Figure S4). Its mitochondrial origin is of little doubt, based on its complement of genes, which can be translated only with a mitochondrial genetic code (see further). The gene order is considerably changed compared to the inferred mitochondrial genome of the chordate ancestor, as also observed in all other tunicates thus far examined (*S48*).

4.2. Mitochondrial codon usage

cDNA sequences were obtained for 8 of the 13 canonical mitochondrial protein-coding genes. Translation of these sequences was achieved with the previously identified genetic code for the ascidian mitochondrial genomes (translation table 13 in the NCBI collection), suggesting that these sequences derive from an active mitochondrial genome. If these sequences were derived from pseudogenes (*e.g.*, unused copies of the mitochondrial genome residing in the nuclear genome), one might expect that the codon usage would vary randomly and show different profiles between the eight genes. As shown in Figure S5A, this is not the case. Observations are consistent with the idea that the cDNA sequences analysed here derive from the *bona fide Oikopleura* mitochondrial genome. Furthermore, the eight *Oikopleura* sequences show a similar cumulative codon usage profile as *Amphioxus* and *Ciona*, with a markedly uneven usage of the codons (Figure S5B). This is in contrast to human, sea urchin and acorn worm, which exhibit a more even and different usage of the codons (Figures S5B and S5C). It is noteworthy that *Oikopleura* and *Ciona* share five of their most frequently used codons (TTT, AGA, ATT, ATA and TTA), A/T-rich codons that are relatively infrequent in *e.g.* human. Since the ascidian lineage has changed the meaning of codons AGA and AGG from Ser to Gly, one might expect that tunicates have more Gly and less Ser than the other organisms (except the vertebrates, in which has changed these two triplets encode stop codons). For Gly, this is indeed the case, as *Oikopleura* and *Ciona* mitochondrial genes contain 10,0% and 8,7% Gly, respectively, compared to 4,5-7,5% in the other organisms. For Ser, there is, however, no significant difference. Do the supernumerary Gly replace Ser, and if so, does this happen at conserved Ser-positions? We addressed this question in aligning the COX1 sequences of six species and find no significant replacement between Ser and Gly at conserved positions. Hence, the supernumerary Gly are located at variable (less-conserved) positions in the alignment.

Overall, the mitochondrial genetic code adds to the other divergent features of the *Oikopleura* mitochondrial genome (novel gene order, interrupted coding sequences and non-standard location of rRNA and tRNA coding genes).

4.3. Phylogenetic analyses of mitochondrial genes

Phylogenetic analyses of *Oikopleura* mitochondrial genome sequences showed that despite being highly divergent they nevertheless unambiguously clustered with other tunicates (Figure S6). The Bayesian analyses under the CAT model of the concatenation of the four most conserved mitochondrial genes (COX1, COX2, COX3 and CYTB) indeed found almost maximal support for Tunicate monophyly (PP = 0.99). The phylogenetic position of *Oikopleura doica*, which displayed by far the longest branch in the tree, nevertheless appeared unresolved among tunicates. The elevated evolutionary rate of the *Oikopleura dioica* mitochondrial genome is likely responsible for this irresolution which might stand until additional appendicularians are included to break-up this incredibly long branch.

4.4. Methods

Partial but nearly complete cDNA sequences (ESTs) of *Oikopleura dioica* were obtained for 8 of the canonical mitochondrial protein-coding genes from: ATP6, COX1, COX2, COX3, CYTB, ND1, ND4, and ND5. These, and the corresponding genes from *Ciona intestinalis, Branchiostoma lanceolatum, Strongylocentrotus purpuratus, Saccoglossus kowalevskii, and Homo sapiens,* were subjected to codon usage analysis using the EMBOSS cusp program (*S49*) running under eBioX for Mac. A total of 2203 codons were obtained from *Oikopleura*. To quantify the overall similarities between the codon usage profiles among the six species, euclidian distances were computed for all pairs of codons. These distances were used to compute a distance-based tree using the EMBOSS fneighbor program using the UPGMA method (*S50*). This tree is used only to visualise the similarities between codon usage profiles, and does not take into account the fact that codon frequencies are not strictly independent of each other.

Phylogenetic analyses of the *Oikopleura doica* mitochondrial data have been conducted from the amino-acid sequences of the four most conserved genes (COX1, COX2, COX3 and CYTB) by updating a previously assembled metazoan dataset (*S51*) with newly available tunictae sequences. Sequences from individual genes belonging to 60 taxa have been aligned using ProbCons (*S52*). Ambiguously aligned sites were identified and removed by applying the GBlocks program (*S53*) using default parameters. The concatenation of the four gene data sets resulted in a total of 1185 amino-acid sites. This dataset was analysed in a Bayesian context using PhyloBayes (*S*54) under the CAT-GTR mixture model (*S35*) with rates across sites modelled using a Dirichlet process (DP) (*S55*). Two independent MCMC were run and sampled every 10 cycles until 10000 points were collected. The first 1000 points of each chain were excluded as the burnin (10%) and the 50% majority-rule consensus tree was computed from the 2 x 9000 = 18,000 trees pooled from the two independent MCMC runs (Fig S6).

5. *Oikopleura* DNA damage repairome

To identify genes involved in the DNA damage repair (DDR), we first used a set of human-*Oikopleura* orthologous genes, created by reciprocal BLAST search of merged collection of *Oikopleura* annotated protein models and human Ensembl proteins, using a procedure that also enables detection of non-1:1 orthologs by clustering inparalogs in each species into one set of hits. Using this procedure, we found orthologs for 48/81 of human genes known to be involved in DNA repair (Table S7). Due to rapid evolution in the *Oikopleura* lineage and imperfect gene annotations, we expected to miss a certain number of *Oikopleura* orthologs of DNA repair genes this way. We thus also performed TBLASTN using a broad set of sequence queries against both genome assemblies, as well as against the full set of ESTs. The set of queries consisted of components of the human DDR machinery and their candidate orthologs in several metazoans (*Branchiostoma floridae*, *Caenorhabditis elegans, Ciona intestinalis, Ciona savignyi, Danio rerio, Drosophila melanogaster, Nematostella vectensis, Saccharomyces cerevisiae*). We then used one or several top *Oikopleura*

TBLASTN hits for a BLASTX on the SWISSPROT database. Using this procedure we could identify 17 additional *Oikopleura* candidate DDR genes (Table S7).

Overall, our survey suggests that most of the major repair pathways are present in *Oikopleura* with clear orthologs in other genomes, namely double strand break repair (DSBR), base excision repair (BER), nucleotide excision repair (NER), and mismatch repair (MMR). However, we were unable to identify 16/81 genes - notably *all* those involved in the non-homologous end joining pathway (NHEJ) for DSBR. This absence of all members of a pathway strongly suggests that the DSBR repair mechanisms of *Oikopleura* have markedly diverged. NHEJ genes are conserved from yeast to mammals. Their absence in *Oikopleura*, if real and not due to extreme sequence divergence unparalleled by any pathway of similar overall level of phylogenetic conservation, remains to be explained. The consequences of their absence for the mutation rate in the *Oikopleura* lineage may have been considerable.

5.1. Nucleotide Excision Repair

The NER machinery is well conserved in the *Oikopleura* genome. NER is a versatile mechanism, able to detect and remove a variety of different lesions. This includes UV-induced cyclobutane pyrimidine dimers (CPD) and 6-4 pyrimidine-pyrimidone photoproducts (6-4 PPs) as well as chemical adducts. After detection of the lesion, the NER operates by unwinding a stretch of nucleotides comprising the lesion and by removing it by dual excision. Repair occurs by DNA synthesis mediated by DNA pol δ or ε, in complex with the clamp loader PCNA and RFC. The *Oikopleura* genome encodes three proteins involved in the damage recognition (DDB1, RAD23A/B, XPC) and the nucleases involved in the dual excision (ERCC1, ERCC4, ERCC5). Regarding the transcription-coupled NER (TCR), several components of the TFIIH complex were found, including ERCC2 and ERCC3 whose activities are essential for TCR (*S56*).

5.2. Mismatch repair

The MMR is based on MutS and MutL-related proteins. It has been proposed that the MSH2-MSH6 heterodimer was responsible for the reparation of single base mismatch (*S57*), whereas the MSH2-MSH3 heterodimer was primarily responsible for the reparation of large mispairs (*S58*). Binding of MLH1-PMS1 complex is critical for subsequent reparation of the mismatch. We found MSH2, MSH6, MLH1 and PMS1 homologues in the *Oikopleura* genome, but not MSH3. A rapid evolution could be a reason why MSH3 was not detected in the *Oikopleura* genome and in other genomes (fly, worm, ascidians). On the one hand, the absence of MSH3 could provide a straightforward explanation for increased mutation rate in the *Oikopleura* genome. On the other hand, it has been shown that under certain conditions the MSH2-MSH6 complex is also able to recognize large, palindromic, mispairs (*S59*).

5.3. Base Excision Repair

We found several *Oikopleura* genes encoding DNA glycosylases (NTHL1, UNG, TDG, OGG1), potentially able to detect a variety of modified bases in the genome. Base removal is the first step of the BER pathway, creating an abasic site which is processed by an AP endonuclease, then an AP lyase (present in OGG1). We detected APEX1, a major and well-conserved AP endonuclease but we failed to detect APEX2, which might play a role in the reparation of mitochondrial DNA (*S60*). Depending on the size of the abasic site, two subpathways for DNA repair are possible: the short patch BER replaces a single nucleotide, while the long patch BER replaces multiple oligonucleotides (*S61*). The former involves DNA synthesis by DNA pol β, assisted with XRCC1, while the latter involves DNA synthesis with DNA pol β, δ or ε, creating an oligonucleotide flap which is processed by the flap endonuclease FEN1. Ligation, using either DNA ligase LIG1 or LIG3, concludes both pathways. We did not find homologues for DNA pol β, suggesting that another DNA polymerase with equivalent properties could be involved in the short patch BER. It was recently demonstrated

that human DNA pol θ, for which a homologue exists in *Oikopleura*, displays activities comparable to DNA pol β and can function in short patch BER *in vitro* (*S62*). An *Oikopleura* homologue for FEN1 was found in the EST collection. Several genes encoding ATP-dependent DNA ligases, including LIG1, were found in the *Oikopleura* genome, and a possibility remains that one of them provides a functional complementation to LIG3, which is missing.

5.4. Double Strand Break Repair

The cellular response to double strand break (DSB) usually mobilizes two distinct pathways. A first pathway uses homologous recombination (HR) to repair a broken chromatid, by using an intact sister chromatid. A second pathway, NHEJ, directly seals the broken ends. The selection of one pathway versus another is essentially dependent on the phase of the cell cycle (HR is active during S and G2, NHEJ is preferentially used during G1).

The first participant in HR is the MRN complex, comprising MRE11A, RAD50 and NBN, which resects the DNA ends at the DSB. In order to expose homologous sequences, the resected tracts are extended by the exonuclease EXO1 or by a complex including topoisomerase TOP3A and helicases BLM and DNA2. RAD51, assisted by BRCA2 and RAD52, rapidly forms filaments along the exposed single-stranded DNA and promotes its invasion into donor double stranded DNA. RAD54 stimulates the pairing of the broken end to its homologous sequence and plays a role in exposing the 3' end of the broken chromosome. Donor DNA is then used as a template for the reparation of the broken chromosome by the DNA pol δ, resulting in a double Holliday junction (HJ). HR concludes by the resolution or the dissolution of the HJ. We detected all the factors required for HR in the *Oikopleura* genome, with the exception of NBN and RAD52. It has been suggested that RAD52 is not essential for HR in vertebrates and that it might also be absent from fly and worm genomes (*S63*). In contrast, NBN plays a critical role by providing a nuclear localisation signal required for the import of MRN into the nucleus. Compared to its molecular partners in the MRN complex, NBN seems to evolve faster: NBN from zebrafish conserves only 39% identity compared to human NBN, whereas MRE11A and RAD50 are conserved more than 65%. Thus, rapid evolution seems a plausible argument to explain the non detection of NBN.

Strikingly, the *Oikopleura* genome seems to lack the entire machinery required for performing NHEJ repair of DSB. In that pathway, DNA ends resulting from DSB are recognized by a XRCC5-XRCC6 heterodimer, which then recruits several factors including the DNA-PKc-DCLRE1C complex, LIG4, XRCC4 and NHEJ1. The resulting complex processes the DNA ends and mediates their joining. None of the components of NHEJ were found in *Oikopleura*, whereas at least three of them (XRCC5, XRCC6, LIG4) were readily detected in each other genome examined (Figure S7).

5.5. Signalling of DNA damage

We detected several *Oikopleura* genes encoding factors involved in the signalling of DNA damage. Factors like ATR kinase and RAD17 are activated upon DNA damage recognition, and they in turn phosphorylate a number of molecules. Among those targets, we found the BRCA1-BARD1 heterodimer, CHEK1 and components of the 9.1.1 complex. Those factors are collectively involved in delaying the cell cycle in response to genotoxic stress.

5.6. Possible impact of the set of DDR genes on the genome architecture

Taken together, the results of our analysis show that, despite variations from the canonical set of genes, DDR pathways are well represented in *Oikopleura dioica*, suggesting that most insults to the genome would be handled. However, many DDR mechanisms are known to introduce mutations in the genome and we can speculate that the performance of the proposed DDR machinery over evolution could leave visible marks on the genome architecture.

6. Detection of operons and prediction of operon gene function

The operons were predicted as co-oriented genes separated by 60 nucleotides at most : 1761 operons containing 4997 genes were predicted on the reference assembly. The number of genes per operon is displayed in Figure S8. In parallel, spliced leader (SL) sequences were searched in the EST sequences, allowing the detection of ~30% trans-spliced genes. The coincidence between the trans-spliced genes and the genes in operons is high, but not all trans-spliced genes occur in operons, and some genes in operons are lacking the transsplicing signal, most likely due to lack of EST information (see genome browser). Protein models of genes in operons or not in operons were used as queries in BLASTP search (default values) in the mouse proteome (SWISSPROT database). The top BLASTP hit was selected for each query when the alignment e-value was < 10-10. This happened for 62% (3072/4997) of the operon genes and 41% (4962/12116) of genes that are not in operons. Annotation with Gene Ontology (*S64*) was performed using DAVID tools (*S65*) with default settings and the mouse proteome as a background, with an filtered output of 2244 genes of operons (45%) and 2943 genes out of operons (24%). Functional annotation shows the operon gene set is significantly enriched for genes involved in house-keeping functions or general metabolic processes such as RNA, protein, DNA, lipid and carbohydrate processing and transport. Genes involved in developmental processes such as morphogenesis and organogenesis are significantly over-represented in the non-operon gene set and under-represented in the operon gene set (Tables S8A and S8B).

7. Activity of *Tor* elements (LTR retrotransposons) and identification of small TE-like elements

Examination of *Tor* elements in the reference and allelic assemblies shows that the majority of them harbour conserved, intact or poorly corrupted sequences. Retrotransposition requires the expression of both the reverse transcriptase and the integrase. Integration of novel copies of *Tor-3G* and *Tor-4H* elements (two fairly distant subclades) into exons may be the source of novel introns, justifying a special focus on their current expression. Another indication for their recent activity can be searched for in their level of presence/absence polymorphism. This polymorphism is observed in the genome sequence through comparison of reference and allelic assemblies, but it was also evaluated through genotyping single individuals in the culture population. Finally, phylogenetic analysis by Maximum likelihood (*S66*) shows the diversity of *Tor* elements into several large groups in relation to their distribution over distinct chromosomes (Figure S9).

7.1. Transcription of Tor elements

Inspection of the 5' and 3' ends of their transcripts showed that the LTR of both element types contain candidate sites for the initiation and termination of transcription, leading to a repeated sequence R on both ends. By analogy with known LTR retrotransposons, the presence of this sequence may point to a function in reverse transcription of the elements. Two large transcripts could be cloned, which cover the whole element sequences (Figure S10). Sequencing of cDNAs showed that these transcripts contain the R repeat and do not originate through a splicing event, compatible with a copy of the entire element. Such a transcript may serve as template for the reverse transcriptase during the synthesis of novel copies. With RT-PCR and RACE amplification, we showed that the genes of several *Tor* elements are transcribed. Transcription of the *gag* gene is initiated in the LTR (Figure S11). Expression of the *env* gene is also observed (Figure S12) and is initiated from an internal promoter. Tor-3G elements are frequently inserted into exons and can be transcribed together with their host gene (Figure S13), although transcripts initiated in the LTR are also detected (Figure S11).

7.2. Presence/absence polymorphism of Tor-3G elements

LTR retrotransposons account for a significant part of the indel polymorphism in the *Oikopleura* genome. This polymorphism was experimentally characterized beyong the genome sequence through genotyping in outbred populations in culture, with the following observations from RT-PCR and Southern blotting (Figures S14 and S15) : 1) insertions mapped in the genome are usually observed at low frequencies in independently and randomly chosen individuals (<5%), 2) insertions at sites other then those mapped in the genome are identified by inverse PCR, 3) Southern blotting with a Tor-3G derived probe shows a variety of multicopy hybridization patterns among individuals, also randomly selected. The low allelic frequency of Tor-3G insertions is correlated with the almost exclusive occurrence of heterozygous genotypes in the populations. Moreover, experimental crosses between selected heterozygous parents for the same insertion have thus far not resulted in homozygous offsprings. As most Tor-3G insertions were found in exons of protein-coding genes, the absence of homozygous genotypes did not permit to conclude whether the inserted elements are spliced out like introns. However, in several cases of insertion into exons, chimeric transcripts including *Tor* sequences are initiated from the host gene (Figure S13).

7.3. identification of small TE-like elements in indels

The size distribution of polymorphic indels showed an excess of elements in the 550-700bp range (Figure S16). Their collection and alignments revealed that many of them are faithfully repeated in the genome. CAP3 consensus sequences of these elements were obtained, allowing classification into 9 clades, four represented by a single element and five diversified into several subclades (up to 5). Most of these clades are original and show no sequence relationship to the others. All elements display terminal inverted repeats, suggesting novel small transposable elements.

Strikingly, elements which keep identical size among distinct copies belong to the five diversified clades (Table S9). BlastN on the genome reveals the presence of numerous additional copies in sites other than those of the indels, suggesting that they are often settled in loci of both chromosome homolog. In total, 415 insertion sites are detected in both assemblies with up to 179 for one of the clades. Of 268 insertions in the reference genome, 198 could be placed on relatively large scaffolds mapped in the high-order assembly, revealing the following chromosome distribution: 94 on both autosomes, 42 on pseudo-autosomal region, 32 on the X chromosome, and 30 on the Y chromosome. The relatively low concentration on the Y chromosome might be caused by an elimination of truncated copies from the collection of sequences, but verifications showed that it is not the case.

Consequently, and in contrast to all large autonomous TE elements mapped in the genome sequence, these small elements seem rather evenly distributed among all chromosomes, including the X chromosome-specific region, suggesting a different and perhaps more relaxed mode of control. Precise mapping of the insertion sites was carried out for 171 elements that looked intact or almost intact compared to the consensus sequence: 32 are indels proper; 40 may also be at least in part indels as they are in regions showing gapped alignments between both assemblies; 99 appear monomorphic in the genome sequence. We found that 42 elements overlap introns. At least 27 of them were unambiguously inserted into pre-existing introns. In 15 cases they covered all or almost all of the intron length, and 4 were in introns of UTRs. The 11 others which match introns interrupting coding sequences were carefully inspected through alignments, and only one or two of them might have created the whole intron by their insertion.

8. Highly Conserved Elements detected through genomic alignments between Atlantic and Pacific *Oikopleura dioica*

8.1. Sequence conservation in non-coding regions including large introns

We compared genomic sequences around developmental genes from a population of the North American west coast (here referred to as the Oregon strain; sequences are from fosmid clones) with the genome sequence of the Norwegian population. Highly conserved elements (HCEs) lie around these developmental genes (Figure S17; see also main text and Figure 1C). Cloning and sequencing small PCR-amplified fragments in or close to the same genes, from one individual of a Japanese population of *Oikopleura dioica* (near Osaka), confirmed ultraconservation of the same segments, and therefore strengthened their significance. In parallel, matches for vertebrate CNEs were systematically searched with blastN but did not provide convincing matches in the *Oikopleura* genome.

Spots of sequence ultraconservation are almost systematically located in non coding regions (see main text and figure 1C), including introns that are larger than average in such genes than in others. We also showed a very high turn over of introns (main text), with considerable intron loss though a minority of introns have kept ancestral position in the genes: the old:new intron number ratio is 0,22 for all precisely mapped introns. Here, we show on the genome scale that long introns are more likely to be old than are short ones (Figure S18). Moreover we systematically inspected the relatively large introns (>300bp): 42% (118/282) of the new large introns vs 20% (29/145) of the old large introns are invaded by repetitive elements. Note that part of the large introns that are devoid of repetitive elements may not be real introns as they overlap EST sequences. Overall, new introns are more often large due to genome repeats than are old introns. We assume that old introns are large due to high density of regulatory elements. Since *Oikopleura* experienced both dramatic genome compaction and intron turnover, our hypothesis was that both intron size and the incidence of ancestral introns should both be higher for developmental genes than their average values in the genome. Indeed, of the introns reliably classified as old (ancestral) or new (specific for Oikopleura lineage), we show that the old introns show double the proportion of old introns compared to all annotated genes (Figure S19). As predicted by our hypothesis, the developmental genes are also characterized by longer than average introns, since the presence of regulatory elements is expected to limit the intron's decrease in size and probability of their loss simultaneously.

8.2. Methods

8.2.1. Alignments

Fosmid contigs from the American *Oikopleura*, taken from regions surrounding developmental target genes were aligned to the reference genome sequence of the Bergen strain using BLAT with the following parameter values: stepSize=5, minIdentity=0 and minScore=0. Regions surrounding five of these genes were also cloned by PCR amplification from genomic DNA of a single individual of a Japanese population. Sequences were aligned with those of American and Bergen strains.

8.2.2. Extracting highly conserved elements

The longest match from the BLAT output was extracted for each contig. These longest matches were processed in order to extract aligned blocks of all sizes with 100% conservation. This was done by comparing the sequences of the aligned blocks and using mismatches to split blocks into 100% conserved regions (hereafter referred to as highly conserved elements - HCEs) of any size.

8.2.3. Density of highly conserved elements

The densities of CNEs over different lengths were then calculated across the aligned regions. Cutoffs of > 1;> 30;> 50;> 70;> 90;> 100;> 150, and > 200 base pairs (bp) were used with a sliding window of 1000 bp and a step size of 50 bp (see Figure SH1). Here density is the number of bases that lie within blocks over a specified size divided by the window size.

8.2.4. Search for sequences homologous to vertebrate CNEs

CNEs conserved among amniotes were extracted from a multiZ alignement of 46 genomes available from the UCSC site (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/). Non-repetitive and non-coding regions that are aligned and larger than 50 bp between human, mouse,

dog, cow and chicken (38,045 regions) were aligned to the *Oikopleura* reference genome using BLASTN (-W 8 –q 2 –r 2). An E-value threshold of 0.001 was applied to select 335 matching CNEs in the *Oikopleura* genome.

9. Natural tail fluorescence supporting genetic sex determination with male heterogamety in *Oikopleura dioica*

Oikopleura dioica is the unique dioecious tunicate species. Genetic sex determination was evidenced by testing the transmission of various genetic markers, including a visible variant governing fluorescence in the tail of male or female adults (Figure S20). This marker is found at low frequencies in natural populations, and crosses in the lab show X-like inheritance (Table S10). Deviations are observed from expectations for a monolocus determination with full penetrance and expressivity: in crosses between "fluo" mothers and wild-type fathers, the phenotype is indeed observed in less than 50% of the offsprings, though more often in daughters (approx. 9/10 carriers) than in sons (approx. 6/10 carriers). The segregation of four molecular markers from X-linked scaffolds do not show such deviations. The fluorescence is observed only at late stages, shortly before sexual maturation. Its nature (endogenous or not) is still unknown, but it is recovered after sperm cryopreservation of male carriers.

10. Abundant class of non-canonical introns and evidence for intron turnover and its mechanisms

10.1. discovery of a new class of atypical introns

During the annotation process, when mapping the *Oikopleura dioic*a cDNA sequences to the genome allowing any type of intron boundaries, we noticed that a large fraction (more than 10%) of the resulting introns displayed non-canonical (non GT-AG) splice sites, whereas the usual proportion is around 1%-1.5% in other genomes (*S6, S68*). The very vast majority of the introns provided by the cDNA resource have the consensus G*-AG. Interestingly, we did not detect any AT-AC nor GT-AG U12 introns (*S69, S70*), nor any of the snRNA components of the minor spliceosome (*S71*) in the *Oikopleura* genome: as previously suspected (*S72*), *Oikopleura* appears to be lacking the minor spliceosome. In order to annotate correctly the genes containing non-canonical splice sites, all G*-AG (GT-AG, GA-AG, GC-AG and GG-AG) introns were allowed when mapping the resources (ESTs, proteins) to the genome and when predicting gene structures using the automatic annotation pipeline. The final reference proteome contains 73254 introns, 53178 (72.6%) of which are validated by cDNA sequences, *i.e.* correspond exactly to an intron from an O*ikopleura dioica* cDNA sequence that was mapped to a unique location on the *Oikopleura dioica* reference genome. Only 87.6% of the introns are canonical GT-AG. The second most abundant class of introns (9%) are atypical introns of the type GA-AG. GC-AG and GG-AG introns correspond respectively to 2.7% and 0.7% of the introns.

It is notable (Figure S21) that most GT-AG introns are in the size range between 35 and 55 nucleotides, whereas GA-AG introns are in the size range between 55 and 100 nucleotides and GC-AG introns show a bimodal distribution between these two peaks. The less abundant GG-AG introns show a distribution similar to that of GA-AG introns. We computed the logos for the GT-AG, GA-AG and GC-AG introns in all size ranges and in the ranges 35-54 (typical GT-AG), 55-100 (typical GA-AG) and >100 : Figure S22.

GT-AG introns show a similar logo in all the size ranges, with a short poly-T tract preceding the acceptor (3') splice site, and an "AG" signal preceding the donor (5') site, homologous to the U1 snRNA recognizing the 3'exon-5'intron boundary. We did not identify a consensus for the branch-point, although positions -10 to -13 relative to the acceptor site are preferentially adenosines and might correspond to the branch site position. *Oikopleura dioica* GT-AG introns are reminiscent of

Caenorhabditis elegans introns, that are short, display no long polypyrimidine tract but rather a consensus acceptor site of the type "TTTTCAG" (or UUUUCAG on the RNA) which was shown to bind the splicing factor U2AF (*S73, S74*), and where no branch-point consensus is observed (*S75*). The donor (5') splice site consensus of GA-AG and GC-AG is very similar to that of GT-AG introns. As previously observed in GC-AG introns from other species (*S68*), *Oikopleura* GA/GC-AG introns, which lack complementarity to U1 snRNA at the second position of the donor site, show a higher conservation of the preceding "AG" signal than GT-AG introns, possibly as a compensation so that the recognition by U1snRNP remains possible.

Some striking differences between GA/GC-AG and GT-AG introns occur at the acceptor (3') sites, depending on the intron size. In the range between 35 and 54 nt (typical of GT-AG introns) canonical and non-canonical introns show similar logos (with a short poly-T tract preceding the acceptor), whereas in the size ranges above 55 nt, GA-AG and GC-AG acceptors are remarkably different from GT-AG acceptors, with a AAAG consensus. Notably, it was shown in binding competition assays that a *C. elegans* oligonucleotide with a double substitution of U to A (UUAACAG) binds *C. elegans* U2AF better than a sequence with a single substitution (*S73*). Thus, it does not seem unlikely that the AAAG consensus is able to bind to the *O. dioica* U2AF splicing factor in a similar way as in *C. elegans*. Additionally, in those atypical introns, there is a conserved "C" in position -8 relative to the acceptor (and positions -10 to -11 are preferentially adenosines, and could possibly harbor the branch site). It is tempting to speculate that the conserved "C" could correspond to a recognition signal (intronic splicing enhancer) for an additional splicing factor that would help recruiting the U2AF complex at atypical acceptor (3') sites. This could explain why short introns (<55 nt) never display the atypical acceptor consensus: their short length might not provide enough space for the additional factor to bind. Finally, there seems to be a very strong association between the donor (5') splice site and the acceptor (3') splice site, in the sense that GT donors are never associated with atypical C.….AAAG acceptors (even in long introns). This suggests that *Oikopleura dioica* introns are recognized by intron definition, a mechanism by which the splice junctions bridge across the intron, that is considered as the dominant mode of recognition of small introns (*S76, S77*). The incompatibility between GT donors and C.…AAAG acceptors might be related to the higher affinity (or faster binding) of U1snRNP to GT donors compared to G(nonT) donors, which might compete with the association of the additional factor on the acceptor side (around the conserved "C") of the intron. The splicing mechanisms proposed here are hypothetical and remain to be investigated. Nevertheless, we believe that both intron types (typical GT-TTTAG and atypical G(nonT)-C…..AAAG) are spliced by the same machinery, that is the major spliceosome, for the following reasons. First, the donor sites of both intron types show clear complementarity to the U1snRNA, which suggests that both are recognised by U1snRNP. Second, experimental assays performed in *C.elegans*, where introns are very similar to *Oikopleura* introns, suggest that the U2AF splicing factor might be able to recognize both types of acceptors. Finally, when searching for snRNAs and proteins from the major spliceosome, we could identify only one type of each spliceosomal component in the *Oikopleura* genome (for instance, there is no U1snRNA with a mutation making it complementary to AGGA rather than AGGT donors, and there is no other version of the U2AF35 or of the U2AF65 subunits that could have evolved to bind specifically to atypical acceptors). The splicing of both types of introns by the same spliceosome implies that the *Oikopleura dioica* spliceosome is permissive or, as suggested recently, that nonconsensus intron boundaries are sometimes not more costly than consensus boundaries (*S78*). Figure S23 shows the intron retention rate as a function of the number of introns for each intron type in the reference annotation. The most abundant introns (GT-AG) show lower retention rates than the rarer introns (GA-AG, GC-AG, GG-AG, which indicates that atypical GnonT-AG introns are less efficiently spliced than standard GT-AG introns. The simplest explanation is that the *Oikopleura* permissive spliceosome is able to splice out atypical introns, but apparently with a lower efficiency than that achieved for canonical introns. A more speculative and exciting possibility is that much of the observed intron retention serves some regulatory or other function, in which case the association between donor sequence and intron retention could reflect specific usage of non-consensus introns for gene expression regulation.

Since the atypical introns appear to be spliced less efficiently, they are likely to be counter selected in the situations where their missplicing would be more deleterious. Indeed, we noticed that 24.2% of the UTR introns (the splicing of which is not necessary for reconstitution of the coding sequence) are non GT-AG, whereas only 11.2% of the CDS introns are non GT-AG. Additionally, genes that are expressed at higher rates tend to contain less atypical introns than genes that are expressed at lower rates : see Figure S24. We checked whether this bias was due to the intron length rather than the intron type and it appears that only introns in the size range between 55 and 100 nucleotides (preferential range of atypical introns) are located in genes that tend to be less expressed, whereas larger introns do not show such a bias (data not shown). It is thus likely that the difference in expression level for the different intron types is correlated to their splice sites rather than their length.

As atypical introns tend to be located in less expressed genes, we checked whether they were co-localized in some genes. As shown in Table S11, pairs of adjacent GA-AG introns are significantly more abundant than would be expected if the GA-AG introns were distributed randomly in all the genes (see Methods). However, when constraining the proportion of GA-AG introns per gene, the bias disappears: the co-localization of introns of the same type is due to a global effect on genes (some genes seem more prone to tolerate atypical introns than others) than to a local effect between adjacent introns.

10.2. Phase preference of canonical and atypical introns

A long-standing mystery involves intron 'phase bias,' in which introns are not randomly inserted in coding sequences but show preferences towards certain phases (*S79, S80, S81*). In the genomes studied so far, there are more introns inserted in phase 0 (between two codons) than in phase 1 (after the first nucleotide of a codon) or phase 2 (after the second nucleotide of a codon). Two main scenarios have been proposed to explain this observation. The first one relies on the exon theory of genes (*S82*), that holds that early genes were created through the intron-mediated shuffling of exons: the excess of phase 0 introns would correspond to the ancient introns that persisted over time (*S79*). The other scenario relies on the non-randomness of intron insertion sites, so-called "proto-splice sites" (*S83, S84, S85*). Because of the codon usage, non-random local sequence insertion preferences would lead to different rates of insertion into the three phases. In this model, the introns could either be inserted into a consensus sequence or be inserted randomly but with different rates of fixation. The present-day flanking exonic sequences can not explain fully the phase distributions observed (*S86, S87*). However, Nguyen *et al.* (*S88*) proposed that after introns were inserted, the exon junctions surrounding introns were subject to a much lower mutation rate than the average mutation rate. Consequently, the intron phase distributions predicted using current sequences would not match the observed data, especially in fast-evolving species.

We compared the intron phase distributions for *Oikopleura dioica* GT-AG and GA-AG introns: Table S12. As observed in other eukaryotes, *Oikopleura* GT-AG introns occur preferentially in phase 0, but unusually, GA-AG introns occur preferentially in phase 1. Interestingly, when simulating the insertion of introns using the present-day insertion sites and codon usage, as described in (*S88*), phase 1 introns are expected to be the most abundant for both intron types. We hypothesize that in the very fast evolving species *Oikopleura dioica*, the present-day intron phase distribution does not correspond to the present-day codon usage but rather reflects the codon usage that was in action when the introns were inserted. Since the GA-AG intron phase distribution is closer to the distribution predicted from the current codon usage we speculate that, overall, GA-AG introns were inserted more recently than GT-AG introns. If the subset of *Oikopleura* introns that are ancestral are primarily GT-AG (see above), then the GT-AG phase distribution could simply reflect a combination of ancestral phase zero-biased introns and more recently-inserted introns with a phase 1 bias similar in magnitude to that of the GA-AG introns.

10.3. Positional bias of introns in genes

It was shown that introns tend to have a 5' biased distribution in the genes, and that this bias is stronger for intron-poor genomes than for intron-rich genomes, where different detection methods provide different results (*S89, S90, S91*).

We investigated the intron position bias in *Oikopleura*, as described in (*S91*). We used an intragene analysis where the intron positions were mapped into a (0-1) interval from 5' to 3', and counted the number of introns at positions <0.5 ("n5") and at positions >0.5 ("n3"). A gene was qualified as "5'biased" when n5>=n3+2 and as "3'biased" when n3>=n5+2. If the number of 5'biased genes is significantly higher than the number of 3'biased genes in a genome, we conclude that introns tend to have a 5'biased distribution in this genome.

We first compared *Oikopleura* with *Ciona intestinalis*, *Ciona savignyi*, *Branchiostoma floridae*, *Strongylocentrus purpuratus* and *Homo sapiens* in a set of 2524 orthologous proteins : Table S13A. All other species show a 5'biased distribution of introns, whereas *Oikopleura* shows a 3' biased distribution of introns. Interestingly, this bias is seen only for genes that are occuring in operons but not in monocistronic transcripts (Table S13B). These characteristics might be related to the intron gain and/or loss mechanisms in *Oikopleura*, where intron-exon organizations were shown to be very divergent (*S92*).

10.4. Intron evolution : characterization of old and new introns

Whereas the spliceosomal components are conserved throughout the eukaryotic world (*S93*) and the positions of many introns are shared by orthologous genes in distant eukaryotes (*S94, S95, S96, S97, S98*), the number of introns varies greatly between species, which reflects ongoing processes of intron gain and loss. Numerous large-scale analyses of intron evolution performed over recent years yielded somewhat discrepant results about the rate and timing of intron loss and gain (*S84, S90, S97, S99, S100, S101, S102, S103, S104, S105, S106, S107, S108, S109*): those results are discussed in detail in (*S109*). Nevertheless, it is currently admitted that introns invaded eukaryotic genomes at early stages of eukaryogenesis and that subsequent intron losses and gains occurred, with losses being somewhat more common although bursts of intron insertion have certainly occurred, potentially accompanying major evolutionary transitions (*S104, S109, S110*). Carmel *et al.* (*S109*) propose that there are three modalities of intron gain and loss during eukaryotic evolution : the balanced mode, a universal process that is operating in many eukaryotic lineages and is characterized by approximately proportional intron gain and loss rates, and modes of elevated loss or elevated gain rates that are specific to some lineages (*S111, S112, S113, S114, S115, S116*). Episodes of intron proliferation might be favored by severe population bottlenecks resulting in weakened purifying selection (*S117, S118*).

Although the evolutionary history of introns was extensively studied, the mechanisms by which introns are lost or gained remain poorly understood (*S119*). In many lineages, such as vertebrates, the intron turnover is generally slow (*S120*), and the gains are too few and/or too old to reveal their origins. However, some lineages display very high intron turnovers (*26, 31 , 33 , 50*) and are potentially good models for studying intron loss and gain mechanisms. We took benefit of the sequencing of the whole genome of *Oikopleura dioica* to undertake a large-scale survey of intron evolution by comparison of exon-intron structures among chordates. We investigated intron evolution on the whole genome scale, by comparing exon-intron structures of a set of 2524 orthologous proteins (defined as best reciprocal hits: see Methods) found in *Oikopleura dioica*, *Ciona intestinalis* or *Ciona savignyii*, *Branchiostoma floridae*, *Strongylocentrus purpuratus* and *Homo sapiens*. For each gene, only one of the two *Ciona* orthologues -the one providing the best alignment score with *Oikopleura*- was included. The five protein sequences were aligned and the intron positions were mapped in the alignments in order to classify the introns as new and old introns as described in Methods: new introns are defined as introns that are located in conserved regions of the alignment and found only in *Oikopleura dioica*, and old introns as introns that are also found at the same position in at least two other species (see Figure S25). According to this definition, we were able to detect, among 5589 interrogated introns, 4260 new introns (76.2%), among which 3640 are validated by cDNAs and 930 potentially old introns (16.6%), among which 776 are validated by

cDNAs; the other introns were not classified. Most *Oikopleura* introns are new introns, which reflects a high intron gain rate in this lineage. Additionally, when considering introns ancestral to chordates*, i.e.* introns that are present in at least 3 species (Figure S25), only 930/9437 (9.85%) are retained in *Oikopleura dioica*. This lineage appears to be subject to a very high intron turnover : both intron gains and losses have occurred at high rate.

We compared the characteristics of old and new introns in *Oikopleura*: Table S14. There is a significant difference of intron size distributions between old and new introns (P= 1.2e-09): new introns tend to be shorter than old introns (76.5% and 67.9% of introns with length <=80nt respectively). This suggests that introns are inserted as short introns, that are likely to be spliced more efficiently than longer introns, and then increase in length, for instance by insertion of transposable elements. Alternatively, this could reflect preferential retention of longer old introns, for instance due to functional signals present in a subset of longer introns.

New introns display more atypical splice sites (nonGT-AG) than old introns (8.4% and 2.6% respectively, P=1.7e-08). There are more phase 0 introns among old introns than among new introns (P=3.3e-09). Additionally, when comparing only GT-AG old introns with GT-AG new introns, the new introns are less biased towards phase 0 than the old introns (43.3% vs 57.1% of phase 0 introns respectively : Table S15. We performed the simulation previously described (*S88*) to infer the expected phase distribution using the current codon usage in *Oikopleura* and proto-splice sites from old GT-AG introns, new GT-AG introns, and new GA-AG introns (Table S15). The observed distributions differ largely (P<1e-09) from the expected distributions for old GT-AG and new GT-AG introns whereas the observed distribution for new GA-AG introns is in agreement with the expected distribution. When comparing phase distribution of all GT-AG and GA-AG introns, and with the assumption that present-day intron phase distribution reflects the codon usage at the time when the introns were inserted, we had hypothesized that GA-AG introns had been inserted more recently than GT-AG introns. The comparison between old and new introns confirms this hypothesis: GA-AG new introns appear to have been inserted at a time when the codon usage was already the current one. However, GT-AG new introns do not appear to have been inserted so recently: they contain more phase 1 introns than old GT-AG introns but they still show a preference for phase 0 introns. We speculate that those introns contain a mixture of introns that were created at different times (with different codon usages): among them, some are “really new”, some are “not that new” and some are “quite old” (this fraction of old GT-AG introns would be the one retaining a strong phase 0 bias). This suggests that the process of intron creation in *Oikopleura* did not occur in one burst but continuously over time, and that it is still an ongoing process.

The logos for GT-AG old and new introns are very similar (Figure S26), although we could detect slight differences in the information content (IC) of both types of introns: acceptors from new introns have lower information contents than acceptors from old introns (Table S16). We looked for evidence of splice signal migration from exons to introns as described in Sverdlov *et al* (*S121*). We detected a weak signal around donor (GT) sites, where the exonic positions have higher IC for new introns than for old introns, and the intronic positions have higher IC for old introns than for new introns, but no such bias is observed around acceptor sites, where both exonic and intronic positions have higher IC for old introns than for new introns (see Figure S27 for a comparison with Sverdlov *et al*). This observation is in accordance with a model where new introns would originate from non-intronic sequences and would be preferentially inserted -or inserted randomly and preferentially retained- in sites with a strong signal, that is compatible with splicing. Then, over time, the intronic sequences would adapt by gaining signal so that the splicing becomes more efficient, and the constraints on the exonic sequences would be relaxed. On the other hand, it is also possible that old introns that were retained contain more signal in their intronic sequences (regulatory signals…), or have acquired functional roles, which prevented them from being lost. This has been proposed by Carmel *et al.* (*S122*), who observed a negative correlation between the rate of intron gain and the rate of coding sequence evolution, which suggests that at least some introns are functionally relevant.

10.5. Intron loss

We investigated the pattern of intron loss in *Oikopleura dioica*, using 9437 ancestral introns (*i.e.* introns that are present in at least 3 species other than *Oikopleura*) among which 930 (9.85%) are retained in *Oikopleura dioica*. We first quantified the positional bias of lost introns, using an intragene approach similar to that described in (*S101*) : ancestral introns were mapped into a (0-1) interval from 5' to 3' of the genes and were assigned the value "1" when they were retained in Oikopleura and "0" when they were lost. Only the genes containing at least one lost and one retained ancestral intron were included in the analysis. For each gene individually, the Pearson correlation between the position and the retained/lost status was calculated, and the number of genes with significantly ($P<0.05$) positive correlations (losses biased towards 5') was compared to the number of genes with significantly negative correlations (losses biased towards 3'). Among 446 genes, only 18 showed significant correlations, 17 were positive and 1 was negative. Using an intergene approach, *i.e.* calculating the correlation for all genes together, we obtained a significant positive correlation between the position in the gene and the retained status (rho = 0.1381567; P=1.21e-13). We conclude that in *Oikopleura dioica*, unlike in most (but not all) of the species studied so far, where introns tend to be lost in 3' (*S101*), the introns are preferentially lost in 5' of the genes. We compared the intergene correlation for genes occurring in operons (rho= 0.1873622 ; P= 5.47e-12) and genes outside of operons (rho = 0.09680266 ; P= 1.54e-04) : both correlations are significant but the 5' loss bias is stronger in operons than outside of operons. Additionally, the proportion of lost introns is higher in operons (92.4%) than outside of operons (87.6%), which could be related to the higher expression level of genes in operons or to a mechanism of intron loss that would be favored in operons. If the intron loss mechanism was related to the operon resolution mechanism, we would expect a positional bias of intron losses along the operons rather than along the mature mRNAs. Since we did not observe such a bias, we propose that the higher rate of loss in operons is reflecting the higher expression level of operonic genes. In operons, more introns are lost and the 5' bias for losses is stronger than outside of operons, which probably explains why introns distribution is biased towards the 3'end of the genes in operons, but not outside of operons.

Two main mechanisms have been proposed to explain intron loss (reviewed in (*S106*)) : recombination between the genomic copy of a gene and an intronless cDNA produced by reverse transcription of the corresponding mRNA or genomic deletion. For the latter mechanism we would expect an imprecise elimination of introns. In this study, since we only focused on introns in regions correctly aligned, with no gaps surrounding the introns, it is not possible to quantify losses that occured by genomic deletion. Nevertheless, we identified numerous instances of losses that correspond to exact deletion of introns, which is in agreement with the mRNA-mediated loss mechanism. This mechanism also predicts that genes are expressed in germline cells, where recombination usually takes place (we did not have access to this information in order to test this hypothesis), and that we should observe concerted loss of adjacent introns, but too many losses occurred in the *Oikopleura* lineage for us to be able to quantify this phenomenon: we can not decipher between concerted losses and independent losses. Finally, since the reverse transcriptase processes from the 3' to the 5' end of genes, one would expect a 3'biased distribution of intron loss (*S90, S106, S123*). Notably, a clear 3' bias was not observed in many species studied so far including fungi (*S124*), *Caenorhabditis elegans* (*S101*), pufferfishes (*S125*), or *Cryptococcus* (*S126*). In *Oikopleura*, we rather observe a 5' bias for intron loss. The lack of a 3' bias could be explained by a reverse transcription mediated mechanism where the priming does not occur at the polyA tail but randomly inside the transcripts by self-priming (*S127, S128*). Such a mechanism can lead to a 5' loss bias if the minimal size of the loop required for self-annealing prevents 3' introns from being lost. Selective forces could also play a role in the preferential loss, or rather preferential conservation, of introns in certain parts of the genes: in *Oikopleura*, 3' introns might contain important signals -play important roles- that forbid them from being lost. Additionnaly, the bias we observe might not be related to the reverse transcription step but rather to the recombination step: it is possible that in *Oikopleura*, recombination is more likely to take place in 5' than in 3' of the genes. The fact that we can not identify a 3' bias for intron loss does not allow us to rule out the mRNA-mediated mechanism. The most striking feature of intron losses in *Oikopleura* is the fact that losses are more abundant in genes that occur in operons, that is genes that are significantly more expressed. This finding is strongly supporting a model of intron loss where transcription plays a major role. Given all

our observations, whereas *Oikopleura* appears to have a different pattern of intron loss than most species studied so far that will need to be investigated in more detail, we propose that the mechanism of intron loss in this species is the mRNA mediated mechanism already described (*S90, S106*).

10.6. Intron gain

Since most *Oikopleura dioica* introns are new, this species is a good model to study intron gains. When comparing new introns and old introns in *Oikopleura* genes, we observed that new introns tend to be shorter and contain more atypical (non GT-AG) introns than old introns. Additionally, new introns are preferentially inserted in phase 1 whereas old introns are preferentially inserted in phase 0. Notably, introns gained in *Daphnia* are also inserted preferentially in phase 1 (*S129*). In the assumption that the phase preference reflects the codon usage that was in place when the introns were inserted (*S88*) and since the current codon usage is in accordance with the phase distribution of new GA-AG introns, we suspect that the majority of the most recently gained introns are GA-AG introns. From all these observations, we propose that newly-gained introns tend to be gained as short, not necessarily GT-AG introns, preferentially in phase 1, at sites that are compatible with splicing. Over time, the intronic sequences gain signal to improve splicing efficiency, and some of the introns grow in length, for example by insertion of transposable elements.

In order to decipher between different possible gain mechanisms, we looked at the pattern of intron gain in more detail, using 5589 *Oikopleura dioica* introns to which a conservation profile could be assigned (*i.e.* that were located in good quality regions of the multispecies proteic alignment). Among those introns, 4260 (76.2%) are new *i.e.* only present at this position in *Oikopleura* but not in other species; they are likely a mixture of recently gained introns and less recently gained introns. We first quantified the positional bias of new introns, using an intragene approach similar to that described in (*S101*) for intron losses: *Oikopleura dioica* introns were mapped into a (0-1) interval from 5' to 3' of the genes and were assigned the value "1" when they were new and "0" when they were not new. Only the genes containing at least one new and one "not new" intron were included in the analysis. For each gene individually, the Pearson correlation between position and the new/not new status was calculated, and the number of genes with significantly ($P<0.05$) positive correlations (gains biased towards 3') was compared to the number of genes with significantly negative correlations (gains biased towards 5'). In this comparison, we only obtained 11 genes with significant correlations, 7 positive and 4 negative: the difference is not significant. Using an intergene approach, *i.e.* calculating the correlation for all genes together, we obtained a significant negative correlation between the position in the gene and the retained status (rho = -0.08029903; P= 5.735e-05). In *Oikopleura dioica* the introns appear to be preferentially gained in 5'. We compared the intergene correlation for genes occurring in operons (rho = -0.1292584 ; P = 1.167e-05) and genes outside of operons (rho = -0.0396916 ; P = 0.1432) : the 5' gain bias is observed only in operons. Additionally, the proportion of new introns is higher in operons (79.5%) than outside of operons (72.7%), which might be related to the higher expression level of genes in operons.

We also investigated the co-occurrence of new introns in pairs of adjacent introns (Table S17). The pairs of adjacent introns that are both new or both "not new" are significantly more abundant than would be expected if the new introns were distributed randomly in all the genes, even when constraining the proportion of new introns per gene: the co-localization of introns of the same type is not only due to a global effect on genes but also to an intragene effect. The pairs where the 5'intron is new and the 3' intron is not new (553 pairs) are not significantly more abundant than the pairs where the 5'intron is not new and the 3' intron is new (478 pairs), although we had identified a bias for new introns to be located in 5' of the genes in the intergene analysis. Since the intergene analysis may be affected by differences between genes such as gene length, expression level and recombination rate, there is not enough support to conclude that there is 5'bias for intron gains in *Oikopleura dioica*. Several mechanisms have been proposed for the gain of new introns (*S106, S113, S130*):

Insertion of transposon-like elements (*S131, S132, S133*).

Reverse splicing or intron transposition, a reaction where a spliced out intron reverse splices into

a previously intronless position of an mRNA, that is then reverse transcribed and recombines with the genomic locus (*S113, S119, S134, S135, S136*).

Tandem duplication of an exonic sequence that contains splice sites (AGGT sequences) (*S137*).

Creation of new splice sites within exons (so-called intronization) (*S138, S139, S140*).

Internal gene duplications resulting either in the duplication of existing introns or the activation of cryptic splice sites (*S141*).

Recombination between homologous copies of genes (*S120, S142*). This model is not strictly speaking an intron gain model since it does not explain how introns were gained in one copy of the gene in the first place.

Conversion of a type II intron into a spliceosomal intron (*S143*). This model could explain the ancient origin of introns, but less likely the intron gains that have been occurring more recently, since not all organisms still contain type II introns in their mitochondria.

Conversion into an intron of sequence inserted into a transcribed region by repair of double strand breaks (*S129, S144*).

Creation of introns by specific 'Introner' elements in one species of *Micromonas* green algae (*S145*).

Most of those mechanisms can be investigated, provided we have access to gain events that are recent enough for the homology between the gained intron and the sequence it came from to be detected. *Oikopleura dioica* provides such a framework.

The *Oikopleura* genome is relatively poor in transposable elements, in terms of quantity and variety. However, indications exist for a considerable turn over of mobile elements. It is possible that current or past transposable elements are a source of novel introns. We searched for introns that are good candidates for having been created by insertion of transposable elements, *i.e.* introns covered by repeated elements on almost all their nucleotides (see Methods). We focused on introns that are flanked exactly by short direct repeats, which are likely to be remnants of recent insertions involving double-strand break repair, as was observed recently in *Daphnia (S129)*. In some cases, for instance when the direct repeat flanking the intron contains the sequence AG(…)G*; it is possible to find G*-AG intron boundaries that would splice out completely the insertion. In other cases, no canonical intron boundaries can be found, but such insertions might still be spliced out by the permissive *Oikopleura* spliceosome. We identified 32 candidate introns (Table S18). The repeated elements identified in the candidate introns correspond to various families of transposable elements (TE): there is not one specific TE family that is responsible for intron creation in *Oikopleura*. Six of those candidates were genotyped in *Oikopleura* populations, and 5 show a presence/absence polymorphism. Interestingly, some of the introns that appeared monomorphic among the assembled alleles (*i.e.* for which several alleles were assembled, all containing the intron, or for which only one allele was assembled) turned out to be polymorphic by genotyping. The polymorphic candidates are probably the most recent insertions and might either disrupt one of the alleles of the gene (if not spliced), or correspond to recent intron gains that are still polymorphic in the population (if spliced). On the other hand, the monomorphic introns are submitted to a stronger selective pressure to be spliced out, since otherwise the two alleles would be inactivated. We performed RT-PCR experiments on homozygous individuals for several of the candidate introns and we were able to show that one of the polymorphic candidates is genuinely spliced (see Methods section further). Thus, although we can not rule out that some of the candidate introns correspond to presence/absence polymorphism, we showed that at least a portion of the insertions occuring in coding exons can be spliced : transposon insertions are indeed a source of novel introns.

In other species, where gain events are usually rare, all attempts made to identify the origin of introns by homology failed (*S106, S119, S129, S146, S147*). Since intron gains are common in *Oikopleura* and at least some of the gains are likely very recent, we expect that some of the introns still display homologies with the sequences they derived from. We blasted the introns that have cDNA support and are not containing any repeats against the whole reference genome, and kept only the matches with e-value>5e-05, %id >90%, and no significant hits between the exons surrounding the intron, in order to eliminate orthologous genes (as described in Methods). We identified 8 introns : all match other introns from the *Oikopleura* genome, which results in 4 pairs of strongly homologous introns. Interestingly all 4 pairs are part of the same gene or of the same transcriptional

unit. These introns provide the first evidence of intron gain by reverse splicing and suggest that this mechanism is acting in an intra-molecular way in *Oikopleura*, the spliced out intron being reinserted in the same mRNA molecule. Notably, for the four pairs of homologous introns, a few exonic bases (from 1 to 6) are also conserved, but there is no consensus among all pairs for the intron insertion site. It is possible that the insertion of an intron by reverse splicing requires an exonic context similar to that of the original intron, suggesting a cooperation of intronic and exonic elements in the splicing and reverse splicing mechanisms.

Other competing mechanisms needed to be envisaged. First, introns could be reverse spliced directly into the genome as shown for group II introns (*S148*). Second, possibly many new introns could also originate via the repair of double strand breaks (DSB), as proposed for very new introns in *Daphnia* (*S116*). However, this mechanism is unlikely to explain the four intron pairs because: (i) no clear direct repeats suggesting repair are observed near intron boundaries, (ii) exon boundaries of homologous introns show short matches, but these partially fit the consensus exon-intron junctions, (iii) repairs after DSB would not readily explain the co-localization of homologous introns in transcription units."

Assuming the duplication model of intron creation, we would expect to find homologies between the left part of the intron and the sequence downstream of the intron and the right part of the intron and the sequence upstream of the intron. Our homology search was restricted to whole introns and thus did not enable the identification of such cases. Moreover, it is not straightforward to detect homologies on portions of introns as small as *Oikopleura* introns (about 30 nt). We can not rule out the duplication model. Finally, the intronization model would leave virtually no trace without the availability of close homologs, and so was not tested.

10.7. Conclusion

We identified novel (mostly GA-AG) splice sites in *Oikopleura*, which display peculiar characteristics in terms of length, phase preference, and consensus acceptor sites. We propose that those introns are spliced by the major spliceosome, which has evolved to become more permissive. In fact, we observed a variety of non-canonical intron boundaries in the EST data (obtained from a pool of outbred individuals), but we focused on the major class (G*-AG) since we can not rule out that some of those boundaries correspond to insertions polymorphic in the population rather than non-canonical splicing. The fact that we observe a clear signal (logo) for G*-AG introns and that RT-PCR experiments performed on homozygous individuals revealed that most of those introns are genuinely spliced makes us confident about the reality of these introns. Some of the other types of non-canonical introns could correspond to real cases of splicing occurring at inappropriate sites, an expected side effect from a permissive spliceosome, but this would need to be tested using EST data from a homozygous individual, to be clear of artefactual introns emanating from polymorphic insertions. The fact that a lot of introns are being gained in *Oikopleura* may explain why the *Oikopleura* spliceosome is permissive. Irimia *et al.* showed that the total number of introns in a genome is correlated to the information content of 5'splice sites across a broad range of species (S149). With an information content of 1.57 and a log(number of introns) of 4.86, *Oikopleura* is in perfect agreement with their curve. Roy and colleagues hypothesize that ancestral introns had relatively weak splice sites and a relatively permissive spliceosome, and that mutations making the spliceosome more permissive were less disfavored in intron-poor species than in intron-rich species, where mis-splicing would have more deleterious effects (S149). It is also possible that in the *Oikopleura* ancestor the spliceosome was already relatively strict but that the burst of intron gains that occurred in the *Oikopleura* lineage selected strongly for a more permissive spliceosome.

We verified that *Oikopleura* is subject to a very high intron turnover, with both gains and losses occurring very frequently. Edvardsen *et al.* made several hypotheses to explain the variability of intron positions in *Oikopleura dioica* (*S92*) : in particular, the short intron size and the fact that splicing is likely dependant on intron definition (*S150*) can relax constraints for intron positioning. Additionally, we can speculate that the high intron turnover is related to the very short life cycle in this species. Indeed it has been proposed that if the mechanisms of intron loss and gain imply a recombination step (which is the case for RT-mediated intron loss and reverse splicing intron gain

mechanisms), provided that fixed recombination events occur during meiosis, the rate of intron loss and gain should be inversely correlated to the generation time (*S106*). RT-mediated intron loss and intron gain by reverse splicing share common mechanistic components: both imply transcription, reverse transcription, and recombination, which is in agreement with the balanced mode of intron evolution proposed by Carmel *et al* (*S109*). As pointed by Roy *et al.* (*S119*), as reverse splicing requires an additional step compared to RT-mediated intron loss (the reverse splicing of an RNA intron lariat into a novel site), the rate of intron gain should be lower than the rate of intron loss. However, since we found evidence for an additional intron gain mechanism in *Oikopleura*, that is insertion of transposable elements, it is not surprising that the rates of intron gain and loss are comparable in this species. One could argue that this correlation between intron loss and gain rates might reflect intron sliding, *i.e.* relocation of intron/exon boundaries over short distances (*S151*), rather than independent loss and gain of introns (*S147*). Rogozin *et al.* (*S152*) showed that one base-pair intron sliding, although rare, can not be ruled out as one possible mechanism of intron repositionning. As a matter of fact, we found introns that share the same position in the alignment of orthologous proteins but are not inserted in the same phase in *Oikopleura* compared to the other species (data not shown): those introns are good candidates to have arisen from one base pair intron sliding. However, no evidence has ever been found to support intron sliding at longer distances (*S104, S152*). Moreover, as described in (*S119*), intron gains and losses via recombination might be favored within and near long exons, thus promoting an association of intron gains with adjacent intron losses that would resemble the pattern of intron sliding, although the overall high rate of gain in *Oikopleura* suggests that many of these will reflect independent origins. This model also predicts that intron loss should be more common in already intron-poor genes, which is the case in *Oikopleura*, where we observed that intron losses tend to accumulate in certain genes. Finally, we did identify formative introns in the genome of *Oikopleura dioica*, which demonstrates that some, if not most, of the new introns are originating from gain events rather than intron sliding.

We found evidence for two intron gain mechanisms : insertion of transposable elements and reverse splicing. It is possible that additional mechanisms also contribute to the very high intron turnover in *Oikopleura*. The model of intron creation by insertion of transposable elements is in agreement with the observation that new introns are usually short, atypical (non GT-AG), preferentially inserted in phase 1 at sites that are compatible with splicing, and that there is weak evidence of splice signal migration from exons to introns around donor sites. It is tempting to speculate that the spliceosome has evolved to become more permissive and allow the splicing of a variety of introns, in order for the spread of transposable elements not to be too deleterious. On the other hand, we found the first evidence for reverse splicing, *i.e.* transposition of an intron from one location to another. The fact that new introns are more abundant in operons than outside of operons, *i.e.* in genes that are more transcribed is in agreement with the reverse splicing mechanism. The observation that new introns tend to be co-localized might be related to the fact that reverse splicing appears to happen inside a single transcriptional unit. One could hypothesize that the spliced out introns tend to be reinserted close to their previous position in the same mRNA molecule. It is also likely that some introns are more prone to reverse splicing than others, which would result in the propagation of introns from one region of a transcript to another and explain the co-localization of gained introns: several new introns in a gene would be originating from the same initial intron. Although we detected only 4 instances of probable reverse splicing, we believe that the mechanism is not merely anecdotal in *Oikopleura*: the homology is likely to be lost very quickly after intron creation, which explains why all other attempts made at identifying homology between introns in other species failed. The two intron gain mechanisms identified in *Oikopleura* (transposon insertion and reverse splicing) are very distinct and are probably occurring in different genes/contexts. It is not possible to speculate whether one mechanism is predominant over the other. Recent studies in *Daphnia* and *Drosophila* suggest that introns may tend to be created in the process of double strand break repair. Given that these previous studies were among the first known cases of characterization of new introns, their support of the same model suggested that this model might give the long-elusive general explanation for intron gain. The current results indicate that intron gain models are in fact quite diverse. Thus, determination of the dominant model or models of intron gain awaits further work on large numbers of new introns from a wide diversity of species.

10.8. Methods

10.8.1. Identifying intron retention events

Intron retentions were identified as introns that were validated by cDNAs, *i.e.* with both boundaries supported by at least one *Oikopleura* cDNA, and that were completely included in an exonic portion of at least one other cDNA. 3642 intron retentions were identified : average rate of intron retention = 3642/53178= 6.8%.

10.8.2. Logos and information content

The logos were obtained using the weblogo software (*S153*), with 15 nucleotides on each side of the intron boundaries, only on validated introns.
Information contents (IC) were calculated as described in (*S121*) on 3 positions in exonic portions (GT-3, GT-2, GT-1 for donors and AG+1, AG+2, AG+3 for acceptors) and on 5 positions in intronic portions (GT+1, GT+2, GT+3, GT+4, GT+5 for donors and AG-5, AG-4, AG-3, AG-2, AG-1). For each position, the IC was calculated as:
2+ pA log(pA) + pT log(pT) + pC log(pC) + pG log(pG) (where pA, pT, pC and pG are the frequencies of nucleotides A, T, C, G, respectively, at this position among validated introns).

10.8.3. Co-localization of introns

We tested the co-localization of GA-AG introns and new introns by comparing two categories of introns for each test: GA-AG versus other introns and new introns versus other introns. We restricted the analyses to genes containing at least three introns pertaining to one of the two categories that were being compared. For each test, we performed two simulations. In the first simulation, the overall proportions of introns (on all genes) in both categories were used to distribute randomly the introns in the genes. The second simulation was designed to eliminate inter-gene effects and focus on intra-gene associations between adjacent introns: the simulations of intron insertions were made for each gene independently, using its specific proportions of introns in both categories. We then extracted all pairs of adjacent introns (from the real dataset and the simulated datasets) and counted the number of pairs of different types. Those numbers were compared by a χ^2 test, in order to obtain P-values.

10.8.4. Simulation of intron phase distributions

To calculate the expected phase distribution of introns, we used the all-pattern model described in (*S88*). First, we calculated the frequencies of insertion sites (so-called proto-splice sites: we focused on 2 nucleotides before the intron and 3 nucleotides after the intron) observed in various datasets : all GT-AG introns, all GA-AG introns, old GT-AG introns, new GT-AG introns, new GA-AG introns. Then, we counted dicodons in the set of 5932 *Oikopleura* genes that were completely validated by cDNAs, and used the frequencies of insertion sites to simulate intron insertions in those dicodons. We then counted how many simulated introns were inserted in phase 0 (between two codons), phase 1 (after the first position of a codon) and phase 2 (after the second position of a codon).

10.8.5. Quantifying the positional bias of introns

We investigated the intron positional bias in *Oikopleura*, using an approach similar to that described in (*S91*). We used an intragene analysis where the intron positions were mapped into a (0-1) interval from 5' to 3'. We counted the number of introns at positions <0.5 ("n5") and at positions >0.5 ("n3"). A gene was qualified as "5'biased" when n5>=n3+2 and as "3'biased" when n3>=n5+2. We then compared the number of 5'biased genes and 3'biased genes in the whole genome, or only on operonic or non operonic genes, using a χ^2 test.

10.8.6. Quantifying the positional bias of lost and new introns

We quantified the positional bias of lost and new introns using an intragene approach similar to that described in (*S101*): introns were mapped into a (0-1) interval from 5' to 3' of the genes and

were assigned the value "0" or "1" according to their lost/retained or not new/new status. Only the genes containing at least one intron of the two categories compared were included in the analysis. For each gene individually, the Pearson correlation between the position and the intron status was calculated, and the number of genes with significantly ($P<0.05$) positive correlations (losses biased towards 5' or gains biased towards 3') was compared to the number of genes with significantly negative correlations (losses biased towards 3' or gains biased towards 5') using a χ^2 test. An intergene approach was also applied, where the Pearson correlation between the position and the intron status was calculated on the whole gene set, *i.e.* considering all genes together.

10.8.7. Defining intron conservation profiles in orthologous proteins

We investigated intron evolution by comparing exon-intron structures in a set of 2524 orthologous proteins found in *Oikopleura dioica*, *Ciona intestinalis* or *Ciona savignyii*, *Branchiostoma floridae*, *Strongylocentrus purpuratus* and *Homo sapiens* (21). The proteins were aligned using MUSCLE (*S154*) and the highly conserved blocks were identified using Gblocks (*S53*) with the following parameters: -p=t -s=n –b5=a –b2=5 –b1=5 –b3=6. Intron positions were then mapped in the alignments so that a conservation profile was assigned to each intron, listing in which species it can be found. We retained introns that were either in Gblocks blocks (well conserved parts of the alignment) or found in at least 4 species among the 5 species we compared, and filtered out intron positions that were distant of less than 5 aminoacids, since we could not rule out alignment issues such as gaps around the introns. New introns were defined as introns that are located in conserved regions of the alignment and found only in *Oikopleura dioica*, and old introns as introns that are also found at the same position and same phase in at least two other species. Additionaly, the introns that are present in at least 3 species could confidently be considered as ancestral to the chordate lineage.

10.8.8. Identification of candidate introns to originate from TE insertions

We collapsed all nucleotides that had been masked by the different repeat detection methods and identified introns covered on almost all their nucleotides by repeated nucleotides, that is on more than 90% of their nucleotides and so that the boundaries of the intron are within 10 nucleotides of the boundaries of the repeated element. Then, we selected only introns that were flanked exactly by direct repeats, which are likely to be remnants of recent insertions. Finally, we focused only on introns that were in regions where several alleles could be assembled in order to have information regarding their polymorphism. We identified 22 candidate introns, 9 of which are monomorphic and 13 are polymorphic.

10.8.9. Identification of the origin of introns by homology searches against the whole genome

We selected introns that were validated by cDNAs and were not overlapping any repeated element. They were.aligned using BLAST (*S155*) against the whole reference genome, and only the matches with e-value>5e-05, %id >90%, on more than 85% of the intron length were retained. Then, we retained only the introns that showed no significant hit in the adjacent exonic regions, in order to eliminate orthologous genes. To do so, two filters were employed: the first was a BLAT comparison of the 50 nucleotides upstream and downstream of the introns and the regions they matched to. The second filter was a visual inspection of all the remaining hits. Eight introns were identified, all match other introns from the *Oikopleura* genome, which results in 4 pairs of strongly homologous introns.

10.8.10. Experimental validation of the splicing of gained introns

The large number of new introns in *Oikopleura* genes has allowed to discover recently gained introns. A first set of 4 pairs of homologous introns, each restricted to one gene or one operon is compatible with the hypothesis of intron gain through reverse splicing. All other candidate intron sequences are repeated in the genome and the repeats have created insertions in multiple distinct genes; they are compatible with the hypothesis of intron gain through transposon insertion. Testing whether these introns are genuinely spliced out is crucial, as they might only represent mutagenic insertions, compensated by the presence of an intact allele which is expressed. Since ESTs were not produced from the inbred line used for genome sequencing, the expression must be monitored in

single individuals whose genotype for the candidate intron presence/absence is unambiguously known. In this purpose, we have set up a method allowing simultaneous extraction of genomic DNA and cDNA for their examination with PCR amplification.

We first genotyped a collection of males and females from the culture, chosen randomly and therefore unlikely to share common parents. Individuals homozygous for a given candidate intron, or heterozygous with one intronless allele , or devoid of the candidate intron were identified. Individuals that are unambiguously found homozygous for the intron are directly assayed for expression of the gene by RT-PCR and splicing of intron, verified after cDNA cloning and sequencing.

Pairs of homologous introns (reverse splicing candidates): we genotyped 5 females using primers designed in exons flanking the 8 candidate introns (4 pairs) and found that all of them were homozygous for the targeted intron based on the fragment amplified from RNAse treated genomic DNA. After RT-PCR, a single and smaller fragment was detected and cloning/sequencing confirmed its specificity. Splicing of homologous intron pairs was efficient for intron pairs on scaffold_3, _52 et _926, with the spliced expression product detected for four of the five females (the cDNA from female 4 was probably contaminated by genomic DNA) (Figure S28, Table S19). In contrast, splicing of candidate introns for the gene on scaffold 17 was less eficient, as only three individuals displayed the spliced product.

Repeated and monomorphic candidate introns: we genotyped males and females for the presence of six such introns. For all but one insertion (on the Y chromosome part of scaffold_8), a single PCR product was obtained and showed that the introns were not polymorphic (Figure S29, Table S20). This suggests that ESTs have resulted from splicing of genes containing these introns.

Repeated and polymorphic candidate introns: we observed a presence/absence polymorphism for most repeated candidate introns (Figure S30, Table S20). Some of them are also polymorphic in the genome sequence. Four situations were encountered:

- intron present in the genome sequence but not detected by genotyping (scaffold_5). This insertion probably has a very low allelic frequency.
- intron detected by genotyping, however at low frequency (scaffold_50). Carriers proved to be heterozygous for the intron.
- intron detected by genotyping at variable frequencies (scaffolds_1, _42 and _70). Three genotypes were observed: homozygous with the intron (homo+), homozygous without the insertion (homo-) and heterozygous (hetero). As scaffold_42 belongs to the X chromosome, males are for this particular intron hemizygous (hemi+ or hemi-). The intron frequency for introns of scaffold_42 and _70 is variable among populations. The intron on scaffold_1 may be located in a fast-evolving site, as the locus without intron cannot be amplified in some animals (mostly females).
- Introns detected by genotyping, with occasionally sex-linkage (scaffold_267 and _10, are part of the large pseudo-autosomal region). In some populations, the intron when present is heterozygous, and only in males. In other populations, this intron is absent in males and in females.

We have tested the splicing for one X chromosome intron (scaffold_42:311401..312139). A control without RT was included to rule out a contamination by genomic DNA. Amplifications were carried out with three primers specific of the host gene and of the intron (Figure S31). Two additional primers in the gene encoding the ribosomal protein RbL23 provided positive control for the amplification. Under these conditions, we carried out the RT-PCR using the cDNA template (Figure S31, lane +RT) or the control without RT (Figure S31, lane –RT) ; and a PCR using the genomic DNA of the same individuals as template (Figure S31, lane ADNg). This latter PCR includes primers internal and external to the intron.

Among the individuals tested, we could again reveal the three genotypes. As they were collected before sexual maturation, we were unable to distinguish males hemizygous for the introns from homozygous females. In all individuals possessing the intron, two RT-PCR products were detected, one 114bp-long (intronless product), and the other 76bp-long (product retaining the intron). The same result is obtained for individuals devoid of the intronless allele (homozygous or hemizygous for the intron). This strongly supports that the intron can be spliced out, though only partially.

11. A minimal immune system predicted in *Oikopleura*.

We searched the *Oikopleura* genome sequence for genes possibly involved in the immune system (results compiled in Table S21). Such an exercise may appear very abstract in the absence of basic information on pathogens and defense mechanisms of *Oikopleura* and more generally tunicates. However, with a very short life-span, a small body size, no recognized hemocytes or macrophages, one could expect for *Oikopleura* rapid innate defense mechanisms which do not depend upon cell proliferation. Induction/activation, if they occur, might be essentially relayed by transcriptional up-regulation. Also, constitutively expressed effectors, such as secreted soluble molecules i.e. antimicrobial peptides *sensu lato*, could play a major role as well as for antiviral response, immediate intracellular mechanisms following RNA/DNA sensing and possibly related to RNA interference. These predictions sharply contrast with the typical structure of defense systems described in most metazoans in which a number of "core sensors" protein families show variable levels of diversification. The recent analysis of sea urchin and amphioxus genomes revealed large sets of proteins containing dedicated domains such as CARD, DEATH or TIR (*S156, S157*), that are used in a few signaling pathways of defense systems from arthropods to mammals (*S158*). In *Ciona,* most of the canonical receptors and pathways are represented, but often in a rather minimalist manner except for specific pathways such as the complement (*S159*).

We searched for the members of key protein families by screening canonical domains found in pathogen recognition receptors (PRR) and membrane activation proteins, leucine rich repeat, sushis, PGRP, C-type lectin, TIR, CARD, DEATH, NACHT, B30.2, B-box, TNF, Traf. These domains were searched at the Genoscope using the software InterproScan. The Igsf molecules were inspected for their constituting domains V, C2 and I-set, C1 (*S160*).

Screening for pathogen sensors provided striking evidence that the immune defense of *Oikopleura* is not based on a large array of typical TLR, NLR, 185/333 or IgSF-based sets of molecules as in sea urchin, amphioxus or gnathostomes, as illustrated in Table S21. The *Oikopleura* genome contains many LRR proteins (74 models) but most of them are without transmembrane region. Only one TLR-like protein was identified in *Oikopleura* with a predicted cytoplasmic TIR domain (GSOIDP00015273001). The presence of a single TLR –like molecule is in sharp contrast with the multiplicity of such receptors in the metazoa that use them for their immune system. This single one, like Toll in insects, could perhaps be involved in immunity via a cytokine - but no homolog of *spaetzle* was found in *Oikopleura* - or could play a role in development. Remarkably, no adaptors with DEATh and TIR domains nor counterparts of TIRAP/TICAM molecules are present while they are key components of the immune signaling in arthropods and vertebrates and were found in sea urchin and amphioxus genomes. In their absence, the TLR-like receptor may signal through the TOLLIP-like protein (GSOIDP00008444001). This protein contains a typical C2-like domain, and a C-terminus ubiquitin binding CUE domain that are conserved in verterbrate TOLLIP proteins, and may bind the TIR domain of the TLR, mediating recruitment of a serine/threonine kinase. A few other LRR transmembrane proteins were also predicted with EGF or IgSF domains contain ITIM or GRB2-binding site, suggesting that they may be effective sensors. One predicted protein sequence (GSOIDP00009878001) was similar to some extent (26% homology expect=2^{-11}) to the agnathans' Variable Lymphocyte Receptors (VLR) (*S161*). Several other LRR had an architecture compatible with that of membrane VLR but with poor similarity. These sequences may be useful for tracking the origin of the first vertebrate system for somatic diversification of antigen receptors, and their genomic organization should be established and analyzed in details.

Among the predicted IgSf molecules only the V domain and the I /C2 set were represented. No C1 Igsf member was found, which confirmed the general observation that canonical C1 domains are encountered only in vertebrates (*S162*). The number of extracellular Ig domains in IgSF proteins of *Oikopleura* those could vary from 1 to 9 with the V domain usually in the most distal position. Several such molecules had putative transmembrane and cytoplasmic segments,

whose 5 had ITIM-like motifs. One of the V domains (GSOIDP00000389001) showed a feature typical of immune receptors: a diglycine bulge in the G strand corresponding to the J segment of TCR or BCR V regions. In *Ciona intestinalis*, a set of IgSF members encoded on chromosomes 4 and 10 have been identified as the counterparts of genes constituting a tetrad of paralogs in human and chicken, that includes the leukocyte receptor complexe (LRC). These genes encode many receptors involved in DC-APC/T cells cross talk and leukocyte activation, and may have evolved from a unique primordial LRC region through two duplications and functionalization in vertebrates, while *Ciona* contains only one copy of it. Interestingly, several counterparts of these genes have been retrieved in *Oikopleura*, with five of them located on the same chromosome in the first assembly: several 3-Ig domains IgSF members had the architecture of Poliovirus receptors or Nectin family members with a distal V domain (*S163*), and the CTX-JAM family (*S164*) with 2 Ig domain (1V-1C2) was also represented.

No Ig, TCR, MHC-class I or II-like sequences could be retrieved, confirming previous observations from other non-vertebrate genomes.

Oikopleura possesses only one SRCR, and a few PGRP and C-lectins could be identified but in very reduced numbers compared to other species.

Although many DEAD-containing proteins have been predicted in *Oikopleura*, no typical RIG-I–helicase with a CARD-like domain could be found.

On the side of effectors, no good candidate for homologs of C3, BfC2, or even ficolin, MASP or TEP could be identified, suggesting that the complement cascade is absent.

While the high amount of viruses in sea water – and their high mutation rate in UV exposed superficial layer of pelagic environment - suggests that antiviral defenses should be important for *Oikopleura*, no homolog of any interferon type could be found. However, several homologs of interferon-induced proteins are present, as well as a putative interferon receptor-associated protein. Several predicted proteins show significant similarity either to the RNA-binding domain or to the kinase domain of PKR, an interferon-inducible dsRNA-dependent protein kinase that constitutes a major system for dsRNA recognition in mammals (*S165*). Homologs of an activator and a repressor of the PKR are present in *Oikopleura*, suggesting that it would be interesting to analyse the response induced by ds-RNA in this species.

Several components of the RNAi system are present in *Oikopleura*, which may be part of specific antivirus defense mechanisms (*S166*). Concomitant with the absence of IFN, only two Tripartite motif proteins (TRIM) have been found in *Oikopleura*. TRIMs have recently emerged as important factors in immunity with many different functions from virus sensing to IFN signaling and inhibition of virus assembly. In *Oikopleura*, *trim* genes display the typical structure Ring/ B1/B2/CC, and are most similar to Trim56 and Trim33 respectively. However, they do not contain SPRY-based B30.2 domains, a C-terminus domain that plays an important role in the function of many TRIM involved in immunity. No true B30.2 domain could be identified in contrast to *Nematostella*. The presence of two goose-type lysozyme genes is confirmed.

Among over-represented domains in the *Oikopleura* proteome, the only relevant candidates are a very large number of well-diversified *Oikopleura* PLA2-like genes. Interestingly, a number of these sequences are similar to the Group II PLA2 genes (for example GSOIDP00015948001), that encode many toxins from snake venoms and also antibacterial compounds of mammals. Although several important residues in the catalytic domain are not conserved in the *Oikopleura* PLA2-like proteins, the diversification of these proteins suggests they have a role in the defense system. Should these proteins have lost the PLA2 activity, they would constitute an interesting example of a successful exaptation of the PLA2 domain. No typical short antimicrobial peptides could be found in a first approach.

Thus, *Oikopleura* possesses very few genes with domains corresponding to typical immuno-receptors or immuno-effector, compared to vertebrates, sea urchin, amphioxus and even to *Ciona*. Since *Oikopleura* has no proliferating and specialized hemocytes, the few genes encoding these homologous domains may not even have an immune function. In agreement with the cellular simplicity of the putative *Oikopleura* immune system no homologs of the regulatory molecules involved in more complex chordates such as interleukins/cytokines and their receptors were identified so far. Overall, the *Oikopleura* genome potentially uncovers a highly derived and

simplified strategy of defense which appears well correlated to its peculiar life history and may be focused on viruses. Finally, the high fertility of *Oikopleura* and the inter-individual recombination through sexual reproduction implies that polymorphisms might play an important role in the survival of the populations, opening the possibility of a strong selection of defense mechanisms at this higher level.

12. Lineage-specific duplications of developmental genes in *Oikopleura*

In an attempt to annotate a comprehensive set of *O.dioica* developmental genes, all genome resources (gene models, ESTs, both genome assemblies and the shotgun dataset) were explored with BLAST and SMART-EMBL at low stringencies. Queries were chosen in strictly following the ten chapters of Y. Satou, N. Satoh *et al*. (*S171*). Candidate genes were challenged by reciprocal BLASTX on non-redundant NCBI protein databases and SWISSPROT and given a provisional name. Due to rapid evolution and relatively high level of divergence between candidate duplicates, phylogenetic analysis with a broad taxon sampling and state of the art Maximum likelihood analyses was implemented, with bootstrap replicates performed with RAxML under a LG+F+Γ4 model (F = amino acid equilibrium frequencies estimated from the data; Γ4 = gamma estimate with four discrete categories). Even though long branches were often observed for *Oikopleura,* most genes could thus be given a more robust name. A minority of genes were discovered through later genome annotations, and were named after less elaborate phylogenetic analysis by maximum likelihood (*S66*) (using www.phylogeny.fr/version2_cgi/simple_phylogeny.cgi). This study allowed to frequently pinpoint the absence of key-gene groups, either due to gene loss or to unsuccessful detection (data not shown). Phylogenetic trees also unambiguously revealed a number of lineage-specific gene duplications higher than usual and clearly higher than in *Ciona intestinalis* (see main text). Candidate gene duplicates are listed below. Their degree of robustness is variable. Other potential duplicates were omitted here due to ambiguous tree topology.

Most developmental gene duplicates detected in *Oikopleura* have members that have substantially diverged from each other, possibly because the duplication is ancient or because the evolution after duplication has been very rapid. In the table S22, note that in the column "GENE GROUP", a few groups gather or may gather several ancestral individual gene groups. In those cases however, the number of detected genes is high enough to imply one or several lineage-specific duplications.

Finally, the list below does not include the *Oikopleura* duplicates corresponding to the last two chapters of the above mentioned survey (*S171*) in *Ciona intestinalis* (IX. Muscle structural proteins; X. Genes for cell junctions and Extracellular Matrix). For several gene families addressed in these two chapters, phylogenetic analysis suggests markedly independent amplifications in distinct lineages. However, these chapters should not be considered as exception to the rule, as particularly large numbers of genes were found to encode *Oikopleura* muscle structural proteins (7 tropomyosins or tropomyosin-like proteins, up to 17 troponins, up to 15 Myosin heavy chains, 8 Myosin alkali light chains and 8 Myosin regulatory light chains), as well as large numbers of genes encoding claudin-like or connexin-like proteins (15-30 for each).

13. Dynamics of early gene duplicates in the *Oikopleura* genome

Here we present a characterization of recent duplicates in the *Oikopleura* genome. Parameterizations for models for changes in selection over time are presented in Table S23. The fit of a mixture model describing the presence of duplicate genes of various ages in a genome is shown in Figure S32, while the in-growth of selection is shown in Figure S33 and Figure S34. We first provide a description of methods applied to generate figures and the table in the main paper and this section.

The *Oikopleura* protein set was subjected to a BLAST all vs. all to identify recent duplicates. In two independent workups of the data, the top BLAST hits (with e-values <1e-10) and all BLAST hits with e-values <1e-10 were collected. A control was run to eliminate evident tandem duplicates that appear at scaffold boundaries. For each identified duplicate pair, a pairwise alignment was generated using Muscle (*S154*). Alignments where the % of gapped positions was greater than 20% of the total alignment length were excluded. Alignments were back-translated to the nucleotide level with three gaps for each amino acid alignment gap. Codeml from the PAML package was run to estimate dS and dN/dS under a model where dN/dS was fixed to 1 and under a nested model where it was estimated. The two models were compared with a likelihood ratio test. Large gene families in the all BLAST hit data were detected using single linkage clustering of pairs with dS<0.3. Only one gene family with more than 10 members was identified. This gene family of 103 members was excluded from the analysis. A BLAST search of members of this family against non-redundant Genbank did not turn up a known function for this family. In evaluating the retention of duplicated genes as measured in dS and in evaluating the relationship of dN and dN/dS to dS, models presented in Hughes and Liberles (*S172*) were applied and a new mixture model was developed that better explained the data, but at this stage, does not allow for mechanistic inference. This model was defined at 0.0, unlike the exponential and Weibull distributions which provided ad hoc binning to account for these data points, but this mixture model did not account for decay of these duplicates into slightly older duplicates with nonzero dS. The mixture model was fitted in R (*S173*), while the substitution model was implemented in WinBugs (*S174*), using a Bayesian framework.

To summarize the models, one class of models was applied to duplicate gene retention data. The initial models based upon survival analysis used a classic model based upon an exponential distribution consistent with a constant birth and constant loss rate over time (dS) (*S175, S176*), compared with a Weibull distribution with a declining loss rate over time assuming constant birth (*S172*) using a likelihood ratio test. The parameterization of the Weibull model gives the instantaneous loss rate and the rate of decay of the loss rate.

Here, a mixture of a discrete distribution at 0.0 and a mixture of Weibull distributions truncated at 0.3 that better explained the data was used. The simpler model with a single truncated Weibull distribution was not an adequate explanation of the data, as the posterior densities of the parameters of the component Weibull distributions of the mixture model were well separated. The mixture model was consistent with both a heterogeneous birth process that could encompass larger scale events and gene family expansions as well as with a heterogeneous death process between different duplicates. The parameterization of the model is therefore not directly comparable to the parameterization of Hughes and Liberles (*S172*). Future work will extend upon this model to enable mechanistic inference about the retention of duplicate genes and direct comparison of such parameters between recent duplicates in different genomes.

A second class of models describes the in-growth of negative selection as duplicate genes decay from an instantaneous dN/dS ratio consistent with relaxed selection due to redundancy to a ratio consistent with diverged functions and the evolutionary behavior of orthologous genes. The parameters of the model describe the instantaneous rate ratio, the asymptotic rate ratio, and the rate of decay between the two rate ratios.

14. Expression patterns of homeobox genes, duplicated or not

To apprehend the function of the numerous duplicates of *Oikopleura* developmental genes, in situ RNA-RNA hybridizations (*S177*) were performed for a sample of homeobox genes, belonging to amplified or non-amplified genes groups (Figure S35). Notably, amplified groups showed highly frequent and variable expression in the main middle region of the trunk epithelium in larvae, generally in addition to expression at other sites, including anterior and posterior region of the trunk epithelium, inner trunk and tail (Table S24). The main middle region of the trunk epithelium becomes dedicated to the synthesis and secretion of house components, and therefore expression

signals are interpreted as part of a lineage-specific gene expression programme. Conversely, conserved ancestral functions of homeobox genes are assumed to occur in other territories.

15. Gene collinearity and Chromosome synteny

15.1. Synteny conservation between *Oikopleura dioica* and *Ciona intestinalis*

Strikingly, no signal of synteny conservation is detected between *Oikopleura* and *Ciona* intestinalis, despite more recent split than with the other species utilised in the present study (Figures S36 and S37). This is somewhat expected due to the very long branch of *Oikopleura* and the relatively long branch of *Ciona* observed in phylogenetic trees.

15.2. Conservation of local gene order

We also performed quantitative measures of synteny conservation between the five species and the human genome. We counted the number of cases where two genes - separated by up to 2 genes - in each species had orthologs separated by a maximal distance of 10-100 genes in the human genome (Figure S38). Amphioxus, *Ciona* and *C. elegans* - sea anemone to a much lesser degree - exceeded random expectations by displaying several fold better conserved neighbourhoods than expected by chance. *Oikopleura* showed a local gene order indistinguishable from random for distances smaller than 30 genes, and a modest level of conserved synteny at a wider distance span, for genes that are not biased for any particular functional category .

15.3. Genes of operons

Since both *Ciona* and *Oikopleura* possess operons, it could be postulated that their respective operons derive from ancestral operons in a common urochordate ancestor, and thus would be subject to functional constraints dating from that time. Alternatively, an independent establishment of operons in the respective lineages would still impose some constraints on gene rearrangements. Out of 5,237 *Oikopleura* genes found to possess an orthologous copy in the *Ciona* genome, 2,338 belong to an operon and 2,899 are outside of operons. We measured the rate of local conservation of gene neighbourhood for both gene categories allowing up to four intervening genes between two *Oikopleura* genes either in or out of operons, and measuring the number of cases where the two *Ciona* orthologs are located within a certain distance of each other. Since 75% of the 1,293 *Oikopleura* operons are composed of 8 genes or less, we took this range as threshold to measure the extent of conserved neighbourhood. However, results show that only 6 pairs of *Oikopleura* genes inside operons possess orthologs within this range in *Ciona,* while this is the case for 7 pairs of *Oikopleura* genes outside of operons. It is thus clear that while neither gene categories are well conserved as previously shown, genes within operons do not show a preferential rate of collinearity retention.

15.4. X chromosome genes

We also examined a second situation, which may have imposed constraints on synteny conservation at a more global scale. Since *Oikopleura* possesses a pair of sex chromosomes, we hypothesised that the *Oikopleura* X chromosome may have been subject to increased constraints compared to autosomes, which would result in higher conservation of synteny, as has been shown between fly and nematode. However, comparing the *Oikopleura* X chromosome gene content to *Drosophila*, nematode, human, chicken and medaka sex chromosomes did not reveal significant enrichments of ortholog retention (Chi square test, $P > 0.01$).

15.5. Methods

15.5.1 Orthology between *Oikopleura* and other genomes

To define the set of *Oikopleura* genes that are orthologs to other sequenced metazoan genomes, we incorporated *Oikopleura* genes into Ensembl phylogenetic trees in a two-stage process. First, we performed reciprocal BLAST comparison between *Oikopleura* predicted protein sequences and the complete set of predicted protein sequences from human, mouse, opossum, chicken, zebrafish and medaka. We next incorporated *Oikopleura* protein sequences into Ensembl (version 57) phylogenetic trees if the 6 reciprocal best hits matched proteins that belonged to the same tree. This selective method identifies 6,714 *Oikopleura* genes orthologous to the six vertebrates, which belong to 6,215 Ensembl phylogenetic trees.

15.5.2 Synteny conservation

Dot matrices were constructed by plotting the positions of orthologs (in their order along chromosomes or large scaffolds) between a given metazoan genome and ancestral chordate linkage groups (CGLs). The latter were reconstituted based from reference *(S25)* by rearranging 125 segments of the human genome (table S14 in reference (*S25*) mapped to HG19, out of 136 segments originally described in HG18) into 17 CLGs (table S1 in reference (*S25*)).To quantitatively measure the conservation of synteny with specified maximal distances, we compared the relative positions of orthologs in a given metazoan genome (amphioxus, release 2 of the Joint Genome Institute annotation (*S25*) ; sea anemone (*S24*) ; *Ciona intestinalis* and *Caenorhabditis elegans* where from Ensembl version 57) to those of their orthologs in the human genome. Amphioxus and sea anemone orthologs were incorporated into Ensembl trees as described for *Oikopleura*. All the genes that did not possess an ortholog in the human genome were ignored in this analysis (Table S25). For each pair of genes separated by at most 2 genes in a metazoan genome and that possess an ortholog in the human genome, we measured the distance between their human orthologs. We computed empirical P-values by randomising the order of genes in the metazoan genome 100 times and by calculating the number of pairs of genes obtained at each iteration in the human genome that belong to one of the four distance intervals. The P-value thus represents the probability of obtaining at least the number of observed collinear pairs if the genes in metazoan genome were in a random order. P-value tests were performed independently for each species to account for varying gene and ortholog contents in different genomes. When computing collinearity, gene pairs of the same species that belong to the same phylogenetic tree (paralogs) are excluded to avoid counting as collinear instances of independent tandem duplications. Enrichment for Gene Ontology categories was measured using the FATIGO suite of tools (*S178*). A schematic tree with the six metazoan species examined for synteny conservation is provided in Figure S37, with positioning of the chordate node.

REFERENCES

S1. D. B. Jaffe *et al*., *Genome Res* **13**, 91-6 (2003).

S2. N. Chen, *Curr Protoc Bioinformatics* **Chapter 4**, Unit 4 10 (2004).

S3. G. Benson, *Nucleic Acids Res* **27**, 573-580 (1999).

S4. A. L. Price, N. C. Jones, P. A. Pevzner, *Bioinformatics* **21 Suppl 1**, i351-358 (2005).

S5. H. Roest Crollius *et al., Nat Genet* **25**, 235-238 (2000).

S6. Gene-It. www.gene-it.com.

S7. A. Bairoch *et al., Nucleic Acids Res* **33**, D154-159 (2005).

S8. E. Birney, R. Durbin, *Genome Res* **10**, 547-548 (2000).

S9. W. J. Kent, *Genome Res* **12**, 656-664 (2002).

S10. G. Parra, E. Blanco, R. Guigo, *Genome Res* **10**, 511-515 (2000).

S11. I. Korf, *BMC Bioinformatics* **5**, 59 (2004).

S12. G. S. Slater, E. Birney, *BMC Bioinformatics* **6**, 31 (2005).

S13. P. Ganot, T. Kallesoe, R. Reinhardt, D. Chourrout, E. M. Thompson, *Mol Cell Biol* **24**, 7795-7805 (2004).

S14. T. F. Smith, M. S. Waterman, *J Mol Biol* **147**, 195-197 (1981).

S15. K. L. Howe, T. Chothia, R. Durbin, *Genome Res* **12**, 1418-1427 (2002).

S16. NCBI. http://www.ncbi.nlm.nih.gov/.

S17. R. Mott, *Comput Appl Biosci* **13**, 477-478 (1997).

S18. J. H. Kim, M. S. Waterman, L. M. Li, *Genome Res* **17**, 1101-1110 (2007).

S19. M. D. Adams *et al., Science* **287**, 2185-2195 (2000).

S20. Consortium, C. e. S, *Science* **282**, 2012-2018 (1998).

S21. J. P. Vinson *et al., Genome Res* **15**, 1127-1135 (2005).

S22. E. S. Lander *et al., Nature* **409**, 860-921 (2001).

S23. E. Sodergren *et al., Science* **314**, 941-52 (2006).

S24. N. H. Putnam *et al., Science* **317**, 86-94 (2007).

S25. N. H. Putnam *et al., Nature* **453**, 1064-1071 (2008).

S26. E. M. Zdobnov, R. Apweiler, *Bioinformatics* **17**, 847-848 (2001).

S27. C. Claudel-Renard, C. Chevalet, T. Faraut, D. Kahn, *Nucleic Acids Res* **31**, 6633-6639 (2003).

S28. M. Kanehisa, S. Goto, S. Kawashima, A. Nakaya, *Nucleic Acids Res* **30**, 42-46 (2002).

S29. B. Haubold, P. Pfaffelhuber, M. Lynch, *Mol Ecol* **19,** 277-284 (2010).

S30. Z. Ning, A.J. Cox, J.C. Mullikin, *Genome Res* **11,** 1725-1729 (2001).

S31. L. Li, C. J. Stoeckert, Jr., D. S. Roos, *Genome Res* **13**, 2178-2189 (2003).

S32. R. K. Bradley, A. Roberts, M. Smoot *et al.*, *PLoS Comput Biol* **5**, e1000392 (2009).

S33. B. Roure, N. Rodriguez-Ezpeleta, H. Philippe, *BMC Evol Biol* **7 Suppl 1**, S2 (2007).

S34. H. A. Schmidt *et al.*, *Bioinformatics* **18**, 502-504 (2002).

S35. N. Lartillot, H. Philippe, *Mol Biol Evol* **21**, 1095-1109 (2004).

S36. D. Baurain, H. Brinkmann, H. Philippe, *Mol Biol Evol* **24**, 6-9 (2007).

S37. N. Lartillot, H. Brinkmann, H. Philippe, *BMC Evol Biol* **7 Suppl 1**, S4 (2007).

S38. H. Philippe *et al.*, *Curr Biol* **19**, 706-712 (2009).

S39. N. Lartillot, H. Philippe, *Philos Trans R Soc Lond B Biol Sci* **363**, 1463-1472 (2008).

S40. J. Felsenstein, *Evolution* **39**, 783-791 (1985).

S41. J. P. Huelsenbeck, M. A. Suchard, *Syst Biol* **56**, 975-987 (2007).

S42. J. Felsenstein, PHYLIP (Phylogene Inference Package) (Distributed by the author, Department of Genetics, University of Washington, Seattle, 2001).

S43. A. Stamatakis, T. Ludwig, H. Meier, *Bioinformatics* **21**, 456-463 (2005).

S44. K. M. Halanych, *Annu Rev Ecol Evol S* **35**, 229 (2004).

S45. F. Delsuc *et al.*, *Genesis* **46**, 592-604 (2008); S. J. Bourlat *et al.*, *Nature* **444**, 85-88 (2006).

S46. J. Felsenstein, *Syst Zool* **27**, 401-410 (1978).

S47. A. H. Riepsamen *et al., J Mol Evol* **66,** 197-209 (2008).

S48. F. Iannelli, F. Griggio, G. Pesole, G., C. Gissi, C., *BMC Evol Biol* **7,** 155 (2007).

S49. http://emboss.sourceforge.net/apps/release/6.2/emboss/apps/cusp.html.

S50. http://emboss.sourceforge.net/apps/release/6.2/emboss/embassy/phylipnew/fneighbor.html.

S51. T. R. Singh *et al., BMC Genomics* **10,** 534 (2009).

S52. C. Do, M. Mahabhashyam, M. Brudno, S. Batzoglou, *Genome Res* **15,** 330-340 (2005).

S53. J. Castresana, *Mol Biol Evol* **17,** 540-552 (2000).

S54. N. Lartillot, T. Lepage, S. Blanquart, *Bioinformatics* **25,** 2286-2288 (2009).

S55. J. P. Huelsenbeck, M. A. Suchard, *Syst Biol* **56,** 975-987 (2007).

S56. V. Oksenych, F. Coin, F., *Cell Cycle* **9,** 90-96 (2010).

S57. G.T. Marsischky *et al., Genes Dev* **10,** 407-420 (1996).

S58. E.A. Sia *et al.*, *Mol Cell Biol* **17,** 2851-2858 (1997).

S59. E. Alani, *Mol Cell Biol* **16,** 5604-5615 (1996).

S60. D. Tsuchimoto *et al.*, *Nucleic Acids Res* **29,** 2349-2360 (2001).

S61. A. Memisoglu, L. Samson, L. *Mutat Res* **451,** 39-51 (2000).

S62. R. Prasad *et al.*, *Nucleic Acids Res* **37,** 1868-1877 (2009).

S63. J.E. Haber, *Curr Opin Cell Biol* **12,** 286-292 (2000).

S64. M. Ashburner *et al., Nature Genet* **25**, 25-29 (2000).

S65. G. Dennis *et al., Genome Biol* **4**, P3 (2003)

S66. M. Anisimova, O. Gascuel, *Syst Biol* **55,** 539-552 (2006).

S67. M. Burset, I. A. Seledtsov, V. V. Solovyev, *Nucleic Acids Res* **28**, 4364-4375 (2000).

S68. S. Kitamura-Abe, H. Itoh, T. Washio, A. Tsutsumi, M. Tomita, *J Bioinform Comput Biol* **2**, 309-331 (2004).

S69. A. A. Patel, J. A. Steitz, *Nat Rev Mol Cell Biol* **4**, 960-970 (2003).

S70. C. L. Will, R. Luhrmann, *Biol Chem* **386**, 713-724 (2005).

S71. W. Y. Tarn, J. A. Steitz, *Trends Biochem Sci* **22**, 132-137 (1997).

S72. M. Marz, T. Kirsten, P. F. Stadler, *J Mol Evol* **67,** 594-607. (2008).

S73. C. Hollins, D. A. Zorio, M. MacMorris, T. Blumenthal, *Rna* **11**, 248-253 (2005).

S74. D. A. Zorio, T. Blumenthal, *Nature* **402**, 835-838 (1999).

S75. H. Zhang, T. Blumenthal, *Rna* **2**, 380-388 (1996).

S76. S. M. Berget, *J Biol Chem* **270**, 2411-2414 (1995).

S77. A. J. McCullough, S. M. Berget, *Mol Cell Biol* **17**, 4562-4571 (1997).

S78. M. Irimia *et al.*, *Genome Res* **19**, 2021-2027 (2009).

S79. A. Fedorov, G. Suboch, M. Bujakov, L. Fedorova, *Nucleic Acids Res* **20**, 2553-2557 (1992).

S80. M. Tomita, N. Shimizu, D. L. Brutlag, *Mol Biol Evol* **13**, 1219-1223 (1996).

S81. M. Long, C. Rosenberg, W. Gilbert, *Proc Natl Acad Sci U S A* **92**, 12495-12499 (1995).

S82. W. Gilbert, *Cold Spring Harb Symp Quant Biol* **52**, 901-905 (1987).

S83. N. J. Dibb, A. J. Newman, *Embo J* **8**, 2015-2021 (1989).

S84. W. G. Qiu, N. Schisler, A. Stoltzfus, *Mol Biol Evol* **21**, 1252-1263 (2004).

S85. A. Stoltzfus, *Curr Biol* **14**, R351-R352 (2004).

S86. M. Long, S. J. de Souza, C. Rosenberg, W. Gilbert, *Proc Natl Acad Sci U S A* **95**, 219-223 (1998).

S87. A. Ruvinsky, S. T. Eskesen, F. N. Eskesen, L. D. Hurst, *J Mol Evol* **60**, 99-104 (2005).

S88. H. D. Nguyen, M. Yoshihama, N. Kenmochi, *BMC Evol Biol* **6**, 69 (2006).

S89. A. Sakurai *et al.*, *Gene* **300**, 89-95 (2002).

S90. T. Mourier, D. C. Jeffares, *Science* **300**, 1393 (2003).

S91. K. Lin, D. Y. Zhang, *Nucleic Acids Res* **33**, 6522-6527 (2005).

S92. R. B. Edvardsen *et al.*, *J Mol Evol* **59**, 448-457 (2004).

S93. L. Collins, D. Penny, *Mol Biol Evol* **22**, 1053-1066 (2005).

S94. D. Penny, A. Poole, *Curr Opin Genet Dev* **9**, 672-677 (1999).

S95. A. Fedorov, A. F. Merican, W. Gilbert, *Proc Natl Acad Sci U S A* **99**, 16128-16133 (2002).

S96. J. M. Archibald, C. J. O'Kelly, W. F. Doolittle, *Mol Biol Evol* **19**, 422-431 (2002).

S97. I. B. Rogozin, Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, E. V. Koonin, *Curr Biol* **13**, 1512-1517 (2003).

S98. F. Raible *et al.*, *Science* **310**, 1325-1326 (2005).

S99. J. M. Logsdon, Jr., *Curr Opin Genet Dev* **8**, 637-648 (1998).

S100. S. W. Roy, A. Fedorov, W. Gilbert, *Proc Natl Acad Sci U S A* **100**, 7158-7162 (2003).

S101. S. W. Roy, W. Gilbert, *Proc Natl Acad Sci U S A* **102**, 713-718 (2005).

S102. S. W. Roy, W. Gilbert, *Proc Natl Acad Sci U S A* **102**, 1986-1991 (2005).

S103. H. D. Nguyen, M. Yoshihama, N. Kenmochi, *PLoS Comput Biol* **1**, e79 (2005).

S104. I. B. Rogozin, A. V. Sverdlov, V. N. Babenko, E. V. Koonin, *Brief Bioinform* **6**, 118-134 (2005).

S105. F. Rodriguez-Trelles, R. Tarrio, F. J. Ayala, *Annu Rev Genet* **40**, 47-76 (2006).

S106. S. W. Roy, W. Gilbert, *Nat Rev Genet* **7**, 211-221 (2006).

S107. D. C. Jeffares, T. Mourier, D. Penny, *Trends Genet* **22**, 16-22 (2006).

S108. L. Carmel, I. B. Rogozin, Y. I. Wolf, E. V. Koonin, *BMC Evol Biol* **7**, 192 (2007).

S109. L. Carmel, Y. I. Wolf, I. B. Rogozin, E. V. Koonin, *Genome Res* **17**, 1034-1044 (2007).

S110. W. Martin, E. V. Koonin, *Nature* **440**, 41-45 (2006).

S111. H. Wada *et al.*, *J Mol Evol* **54**, 118-128 (2002).

S112. S. Cho, S. W. Jin, A. Cohen, R. E. Ellis, *Genome Res* **14**, 1207-1220 (2004).

S113. A. Coghlan, K. H. Wolfe, *Proc Natl Acad Sci U S A* **101**, 11362-11367 (2004).

S114. S. W. Roy, D. L. Hartl, *Genome Res* **16**, 750-756 (2006).

S115. M. K. Basu *et al.*, *Mol Biol Evol* **25**, 111-119 (2008).

S116. A. R. Omilian, D. G. Scofield, M. Lynch, *Mol Biol Evol* **25**, 2129-2139 (2008).

S117. M. Lynch, J. S. Conery, *Science* **302**, 1401-1404 (2003).

S118. M. Lynch, *Mol Biol Evol* **23**, 450-468 (2006).

S119. S. W. Roy, M. Irimia, *Trends Genet* **25**, 67-73 (2009).

S120. B. Venkatesh, Y. Ning, S. Brenner, *Proc Natl Acad Sci U S A* **96**, 10267-10271 (1999).

S121. A. V. Sverdlov, I. B. Rogozin, V. N. Babenko, E. V. Koonin, *Curr Biol* **13**, 2170-2174 (2003).

S122. L. Carmel, I. B. Rogozin, Y. I. Wolf, E. V. Koonin, *Genome Res* **17**, 1045-1050 (2007).

S123. A. V. Sverdlov, V. N. Babenko, I. B. Rogozin, E. V. Koonin, *Gene* **338**, 85-91 (2004).

S124. L. Fedorova, A. Fedorov, *Genetica* **118**, 123-131 (2003).

S125. Y. H. Loh, S. Brenner, B. Venkatesh, *Mol Biol Evol* **25**, 526-535 (2008).

S126. T. J. Sharpton, D. E. Neafsey, J. E. Galagan, J. W. Taylor, *Genome Biol* **9**, R24 (2008).

S127. D. K. Niu, W. R. Hou, S. W. Li, *Mol Biol Evol* **22**, 1475-1481 (2005).

S128. A. L. Feiber, J. Rangarajan, J. C. Vaughn, *J Mol Evol* **55**, 401-413 (2002).

S129. W. Li, A. E. Tucker, W. Sung, W. K. Thomas, M. Lynch, *Science* **326**, 1260-1262 (2009).

S130. J. Coulombe-Huntington, J. Majewski, *Mol Biol Evol* **24**, 2842-2850 (2007).

S131. M. J. Giroux *et al.*, *Proc Natl Acad Sci U S A* **91**, 12150-12154 (1994).

S132. M. Iwamoto, H. Nagashima, T. Nagamine, H. Higo, K. Higo, *Mol Gen Genet* **262**, 493-500 (1999).

S133. S. W. Roy, *Genome Biol* **5**, 251 (2004).

S134. P. A. Sharp, *Cell* **42**, 397-400 (1985).

S135. T. Cavalier-Smith, *Nature* **315**, 283-284 (1985).

S136. C. K. Tseng, S. C. Cheng, *Science* **320**, 1782-1784 (2008).

S137. D. Zhuo, R. Madden, S. A. Elela, B. Chabot, *Proc Natl Acad Sci U S A* **104**, 882-886 (2007).

S138. W. Wang, H. Yu, M. Long, *Nat Genet* **36**, 523-527 (2004).

S139. M. Irimia *et al.*, *Trends Genet* **24**, 378-381 (2008).

S140. F. Catania, M. Lynch, *PLoS Biol* **6**, e283 (2008).

S141. X. Gao, M. Lynch, *Proc Natl Acad Sci U S A* **106,** 20818-20823 (2009).

S142. T. Hankeln, H. Friedl, I. Ebersberger, J. Martin, E. R. Schmidt, *Gene* **205**, 151-160 (1997).

S143. T. Cavalier-Smith, *Trends Genet* **7**, 145-148 (1991).

S144. A. Farlow, E. Meduri, M. Dolezal, L. Hua, C. Schlotterer, *PLoS Genet* **6**, e1000819 (2010).

S145. A. Z. Worden *et al.*, *Science* **324**, 268-272 (2009).

S146. A. Fedorov, S. Roy, L. Fedorova, W. Gilbert, *Genome Res* **13**, 2236-2241 (2003).

S147. R. Tarrio, F. J. Ayala, F. Rodriguez-Trelles, *Proc Natl Acad Sci U S A* **105**, 7223-7228 (2008).

S148. B. Cousineau *et al.*, *Cell* **94**, 451-462 (1998).

S149. M. Irimia, D. Penny, S. W. Roy, *Trends Genet* **23**, 321-325 (2007).

S150. L. P. Lim, C. B. Burge, *Proc Natl Acad Sci U S A* **98**, 11193-11198 (2001).

S151. A. Stoltzfus, J. M. Logsdon, Jr., J. D. Palmer, W. F. Doolittle, *Proc Natl Acad Sci U S A* **94**, 10739-10744 (1997).

S152. I. B. Rogozin, J. Lyons-Weiler, E. V. Koonin, *Trends Genet* **16**, 430-432 (2000).

S153. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, *Genome Res* **14**, 1188-1190 (2004).

S154. R. C. Edgar, *BMC Bioinformatics* **5**, 113 (2004).

S155. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389-3402 (1997).

S156. J. P. Rast *et al., Science* **314,** 952-956 (2006).

S157. S. Huang *et al., Genome Res*. **18**,1112-1126 (2008).

S158. M. S. Lee, Y. J. Kim, *Ann Rev Biochem* **76,** 447-480 (2007).

S159. K. Azumi *et al., Immunogenetics* **55,** 570-581 (2003).

S160. A. F. Williams, A. N. Barclay, *Ann Rev Immunol* **6**, 381-405 (1988).

S161. Z. Pancer, *Nature* **430**, 174-180 (2004).

S162. L. Du Pasquier, I. Chrétien, *Res Immunol*. **147**, 218-222 (1996).

S163. Y. Takai, J. Miyoshi, W. Ikeda, H. Ogita, *Nat Rev Mol Cell Biol* **9**, 603-615 (2008).

S164. I. Chrétien *et al.,* CTX, *Eur J Immunol* **28**, 4094-4104 (1998).

S165. I. M. Kerr, R. E. Brown, A. G. Hovanessian, *Nature* **268**, 540–542 (1977).

S166. D. J. Obbard, K. H. Gordon, A. H. Buck, F. M. Jiggins, *Phil Trans R Soc B* **364**, 99–115 (2009).

S167. R.B. Mehta, M.I. Nonaka, M. Nonaka, *Immunogenetics* **61**, 463-481 (2009).

S168. M. Daëron, S. Jaeger, L. Du Pasquier, E. Vivier. *Immunol Rev* **224**, 11-43 (2008).

S169. D. Brites *et al., Mol Biol Evol* **25**, 1429-1439 (2008).

S170. C. Haruta, T. Suzuki, M. Kasahara, *Immunogenetics* **58**, 216-225 (2006).

S171. Y. Satou, N. Satoh *et al*., *Dev Genes Evol* **213,** 211-318 (2003).

S172. T. Hughes, D. A. Liberles, *J Mol Evol* **65,** 574-588 (2007).

S173. R Development Core Team, A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org (2008).

S174. D. J. Lunn, A. Thomas, N. Best, D. Spiegelhalter, *Stat Comput* **10,** 325-337 (2000).

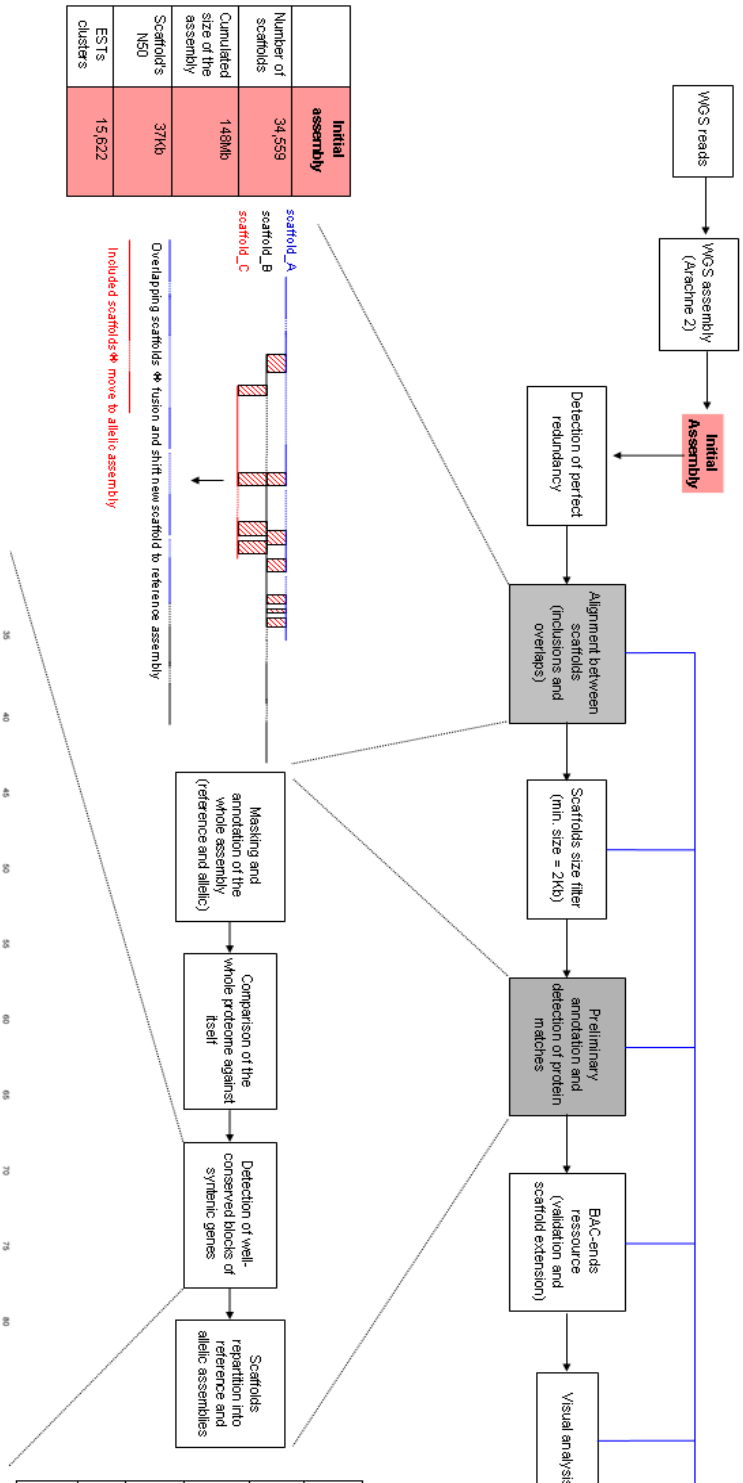
S175. M. Lynch, J. S. Conery, *Science* **290,** 1151-1155 (2000).

S176. Z. Yang, *Mol Biol Evol* **24,** 1586-1591 (2007).

S177. H.C. Seo *et al., Nature* **431,** 67-71 (2004).

S178. F. Al-Shahrour *et al., Nucleic Acids Res* **35**(Web Server issue):W91-96 (2007).

Figure S1: Detection of polymorphism and redundancy in the *O. dioica* genome assembly.



	Initial assembly
Number of scaffolds	34,559
Cumulated size of the assembly	148Mb
Scaffold's N50	37Kb
ESTs clusters	15,622

	Reference Assembly	Allelic assembly
Number of scaffolds	1,260	4,196
Cumulated size of the assembly	70.4Mb	45Mb
Scaffold's N50	395Kb	21.8Kb
ESTs clusters	15,152	11,402

=> Only 470 ESTs clusters are not present in the reference assembly, around 3.1%

Figure S2: Draft chromosome scale assembly based on scaffolds of the reference genome sequence. Ultrascaffolds were obtained using links between BAC end sequences and between shotgun reads located near the ends of scaffolds. Scaffolds are represented by rectangles under which the scaffold number is indicated. The orientation of scaffolds is represented by shadows on top of the rectangles (forward) or on bottom (reverse). In some cases, physical links between contigs of scaffolds appeared questionable, and scaffolds were dissociated into pieces (coordinates in kilobases are indicated with the scaffold number: -80 means from the beginning of the scaffold to the coordinate 80kb, 80+ means from the coordinate 80kb to the end of the scaffold, 80-120 means interval between coordinates 80kb and 120kb of the scaffold). Most scaffolds smaller than 25 kilobases are not represented. Other missing scaffolds are assumed to be part of gaps between scaffolds or ultrascaffolds. Ultrascaffolds have been grouped into chromosomes based on segregation studies of indel variants in full sib families. Ultrascaffolds of X and Y specific regions have received validation using hybridization of entire BAC clones on whole genome tiling arrays.

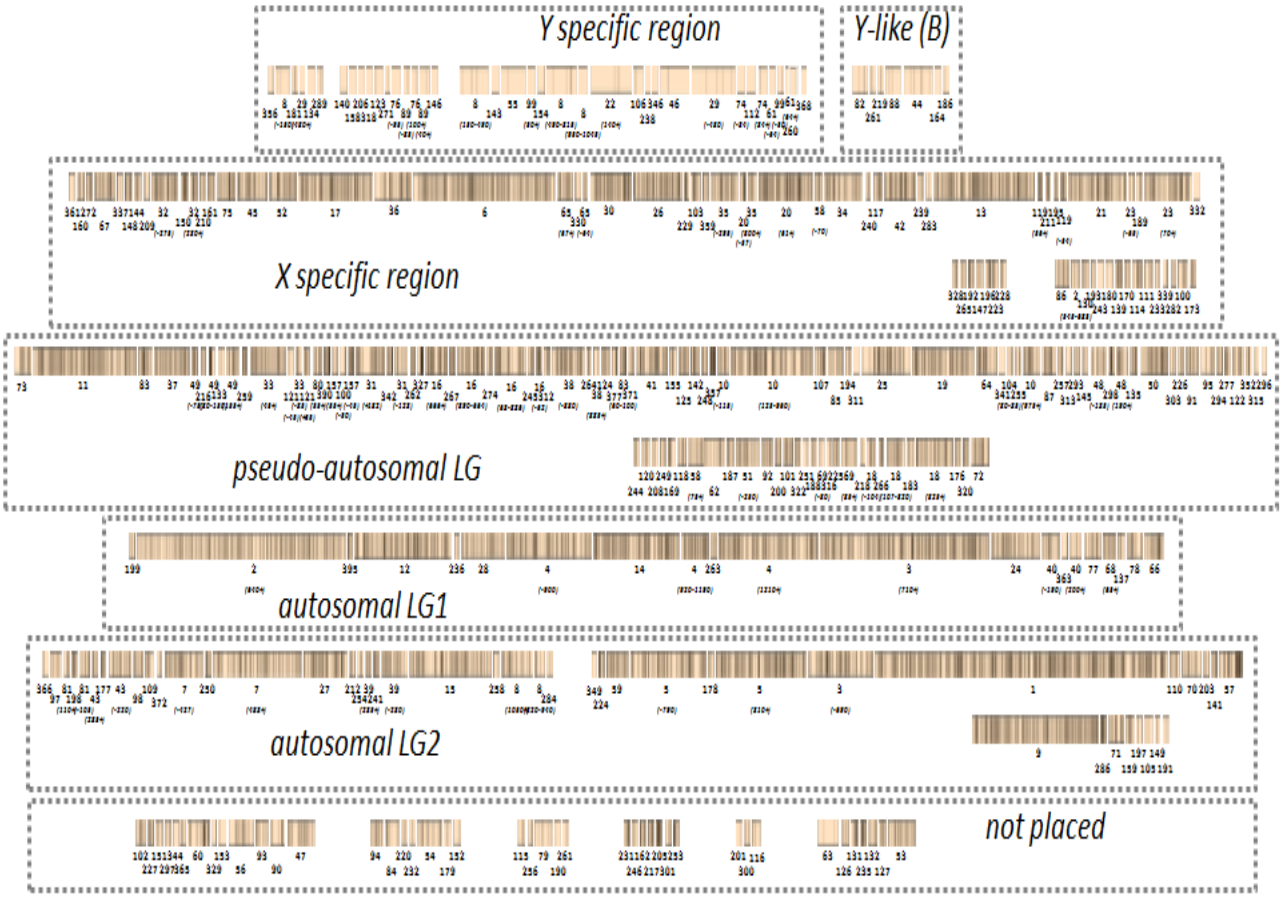


Figure S3A: Trees inferred by PhyloBayes under a CAT+Γ4 model by phylobayes from the nine phylogenomic datasets created from the original pool of orthologous genes in a non-overlapping way by sorting the alignments in function of the number of sequences they contain (from 18 to 26). Only posterior probabilities (PP) different from 1 are presented. All trees are rooted on the choanoflagellate *Monosiga*. The scale bar corresponds to 0.1 substitutions per site; all trees are drawn to this scale.

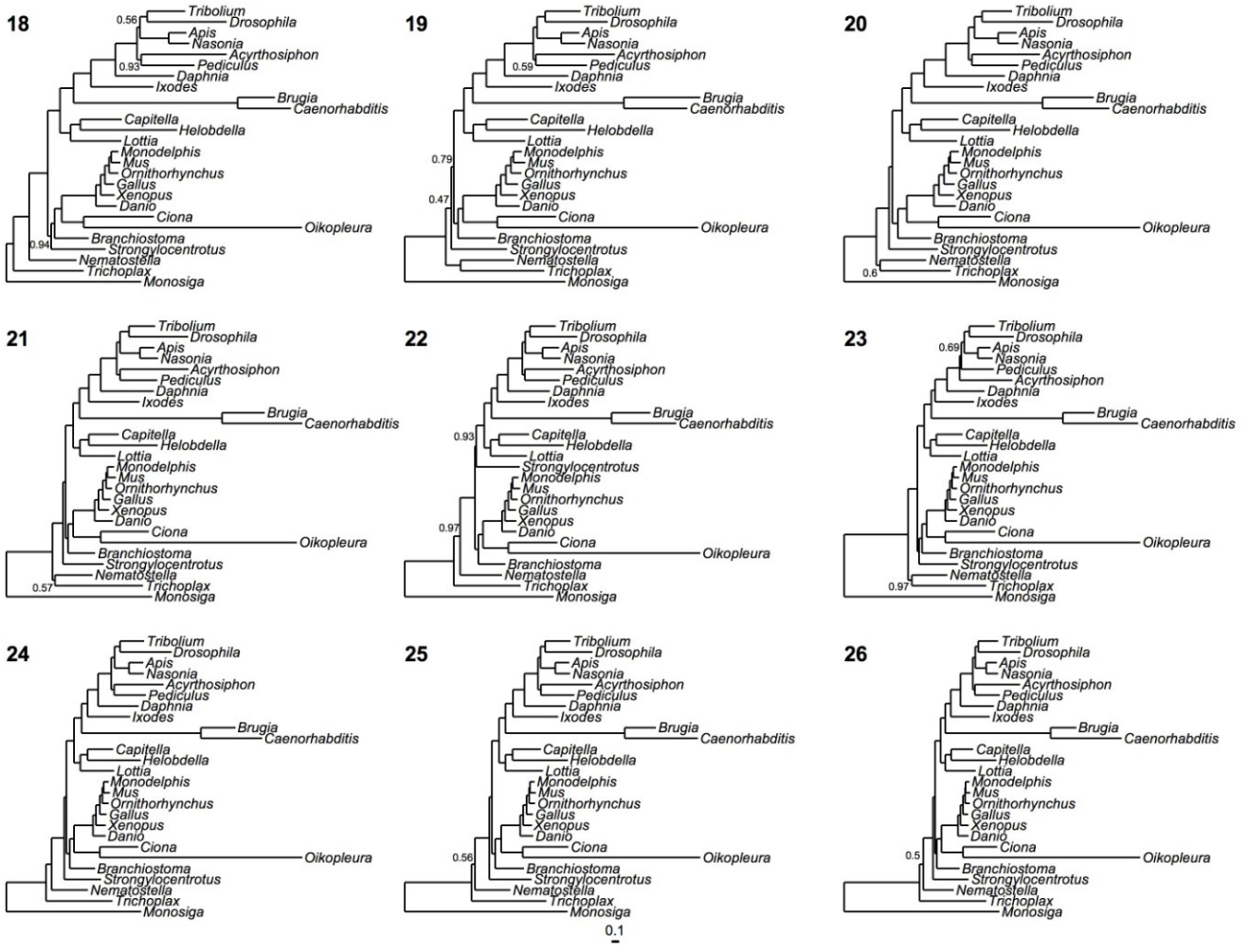


Figure S3B: Phylogenetic trees inferred by RaXML under a LG+F+Γ4 model from the same datasets as in Figure S3A. Bootstrap values (100 pseudo-replicates) using the same program and model of sequence evolution are only presented if they are lower than 100%, but instead of the real numbers they are coded by coloured dots. All other features are identical to Figure S3A.

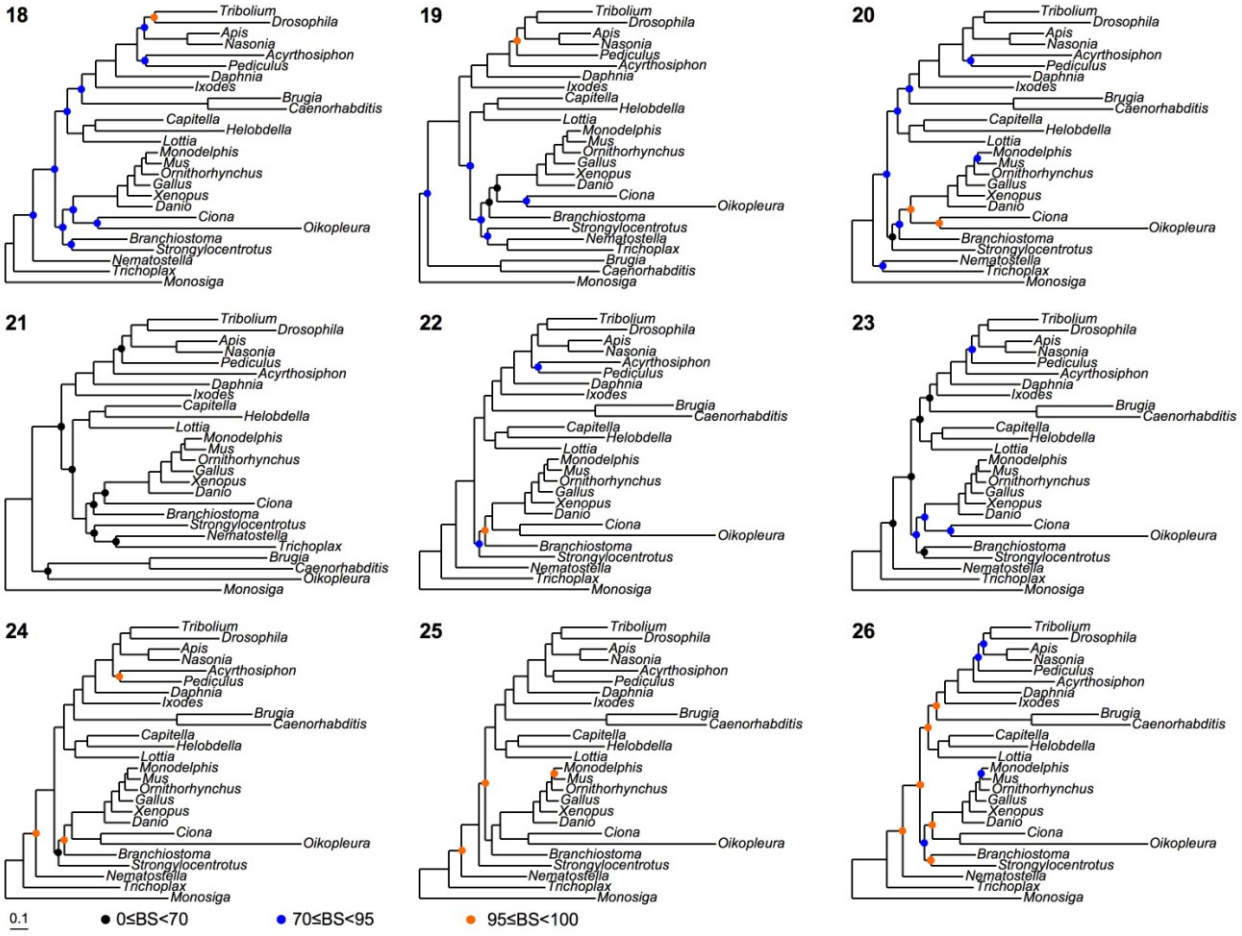


Figure S4: Protein-coding gene order on the mitochondrial genome, mapped through systematic PCR cloning from genomic DNA using EST sequences, which strongly differs from those of amphioxus/human (identical) and other tunicates, here illustrated by the mt-genome of *Ciona intestinalis*. Each TTTTTT site of the *Oikopleura* open reading frames is represented by a red vertical line when occupied by an oligo-dT identified in genomic DNA and a blue line when not interrupted.

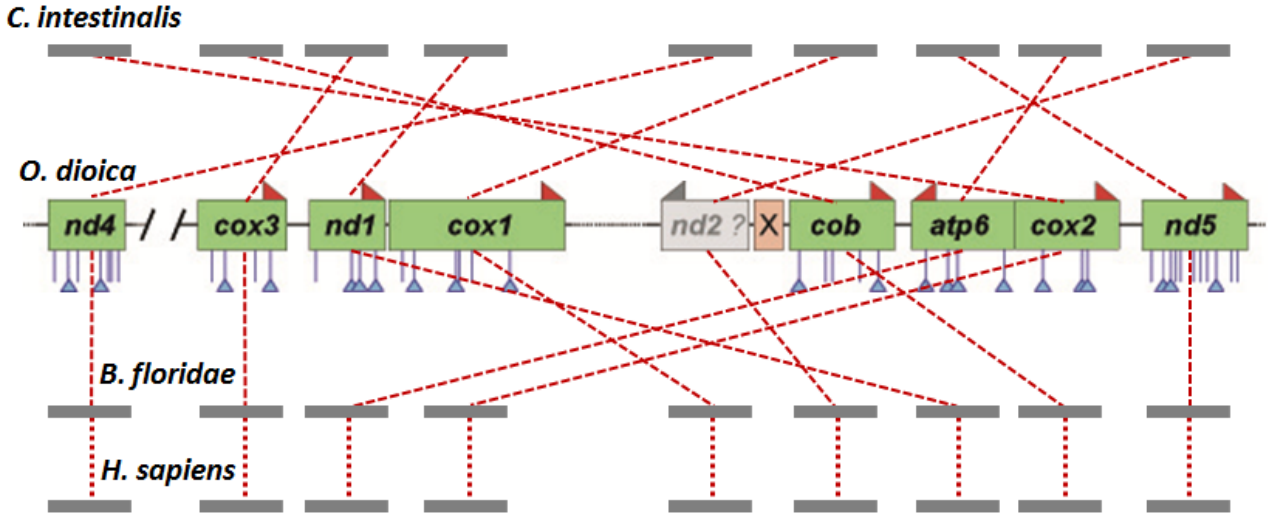


Figure S5: (A) Relative codon frequency profiles for 8 *Oikopleura* mitochondrial genes. (B) Cumulative codon usage profiles for the same 8 genes in *Oikopleura* and seven other species: *Ciona intestinalis, Branchiostoma lanceolatum, Strongylocentrotus purpuratus, Saccoglossus kowalevskii, and Homo sapiens*. (C) Comparison of the codon usage profiles shown in (B) using euclidian distances for all pairs of codons in the six species. See Methods for details.

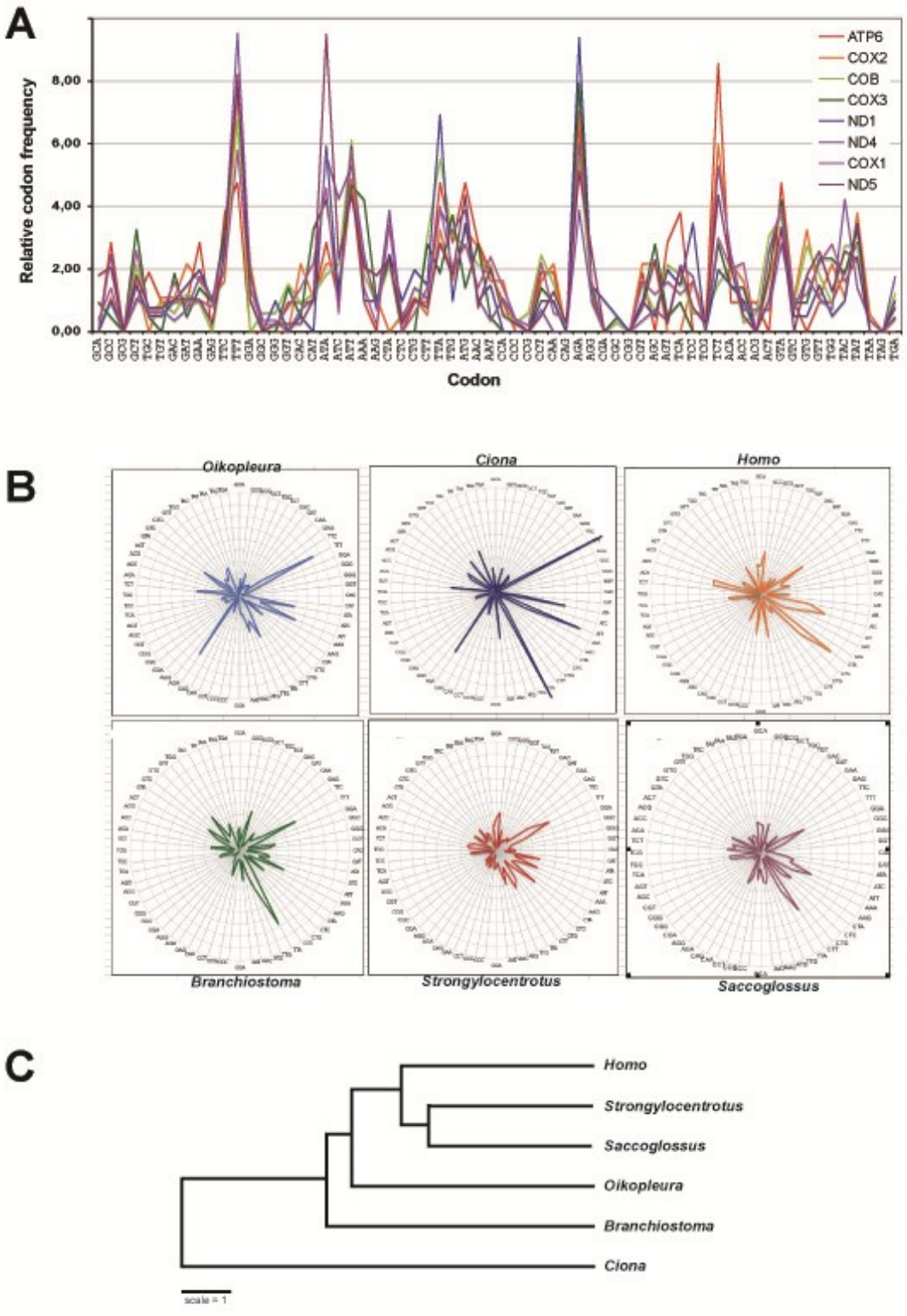


Figure x

Figure S6: Bayesian consensus tree obtained from the phylogenetic analysis of the four most conserved genes of the *Oikopleura dioica* mitochondrial genome (COX1, COX2, COX3 and CYTB). See Methods for details.

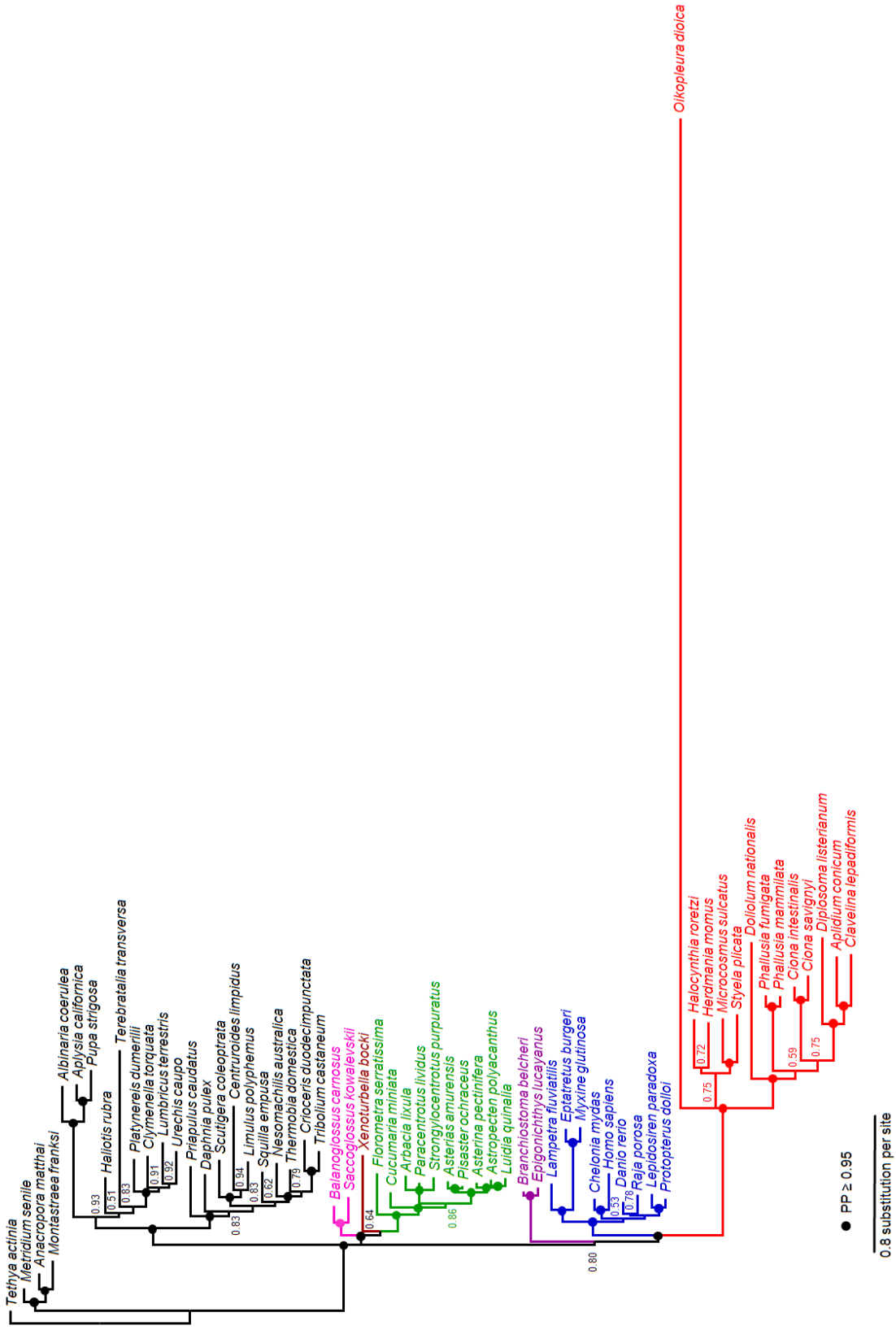


Figure S7: Conservation of amino acids sequences from factors involved in the HR (RAD50, MRE11A, NBN) or the NHEJ (XRCC5, XRCC6, LIG4). The plot shows the percentage of identity after a pairwise alignment between the human homologue and the homolog from a different species. No homolog for NBN was found in the *C. elegans* and *S. cerevisiae* genomes.

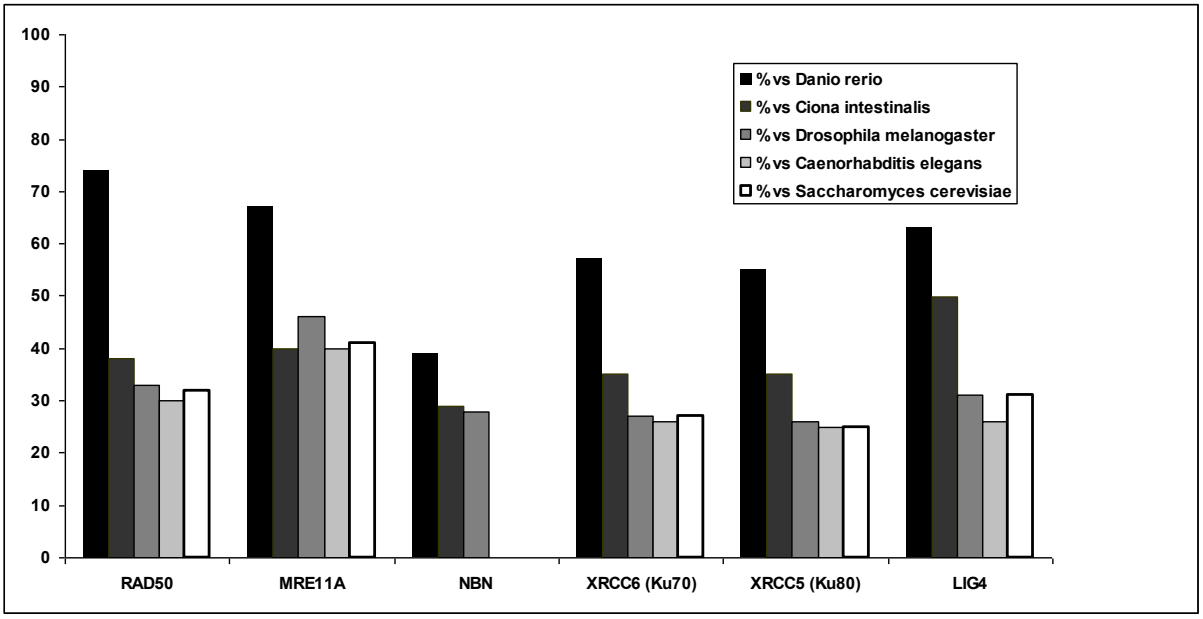


Figure S8: Distribution of operon sizes (in terms of number of genes) for 1761 predicted operons

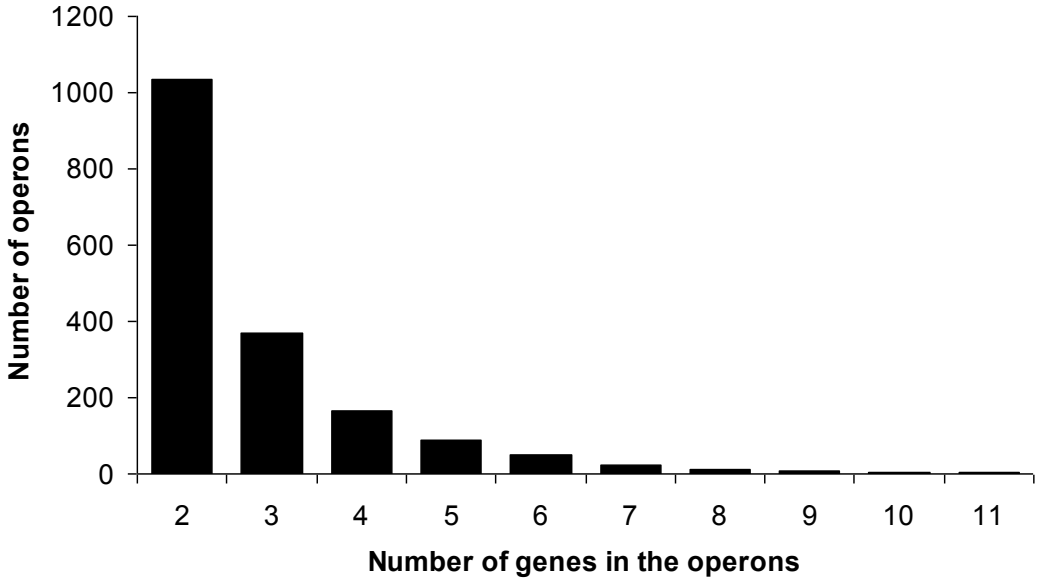


Figure S9: Diversity of *Tor* elements in the *Oikopleura* genome, and their distribution among the chromosomes. Pfam domain RVT_1 (PF00078) were clipped from the *pol* gene of most Tor copies and aligned. The tree was obtained using the "one-click mode" web set of tools with default parameters (MUSCLE for multiple alignment, optionally Gblocks for alignment curation, PhyML for phylogeny and finally TreeDyn for tree drawing). Each copy is referenced by its chromosome (Y, X, B, PAR for pseudo-autosomal region, A for autosomes) and its scaffold in the reference genome. www.phylogeny.fr/version2_cgi/simple_phylogeny.cgi

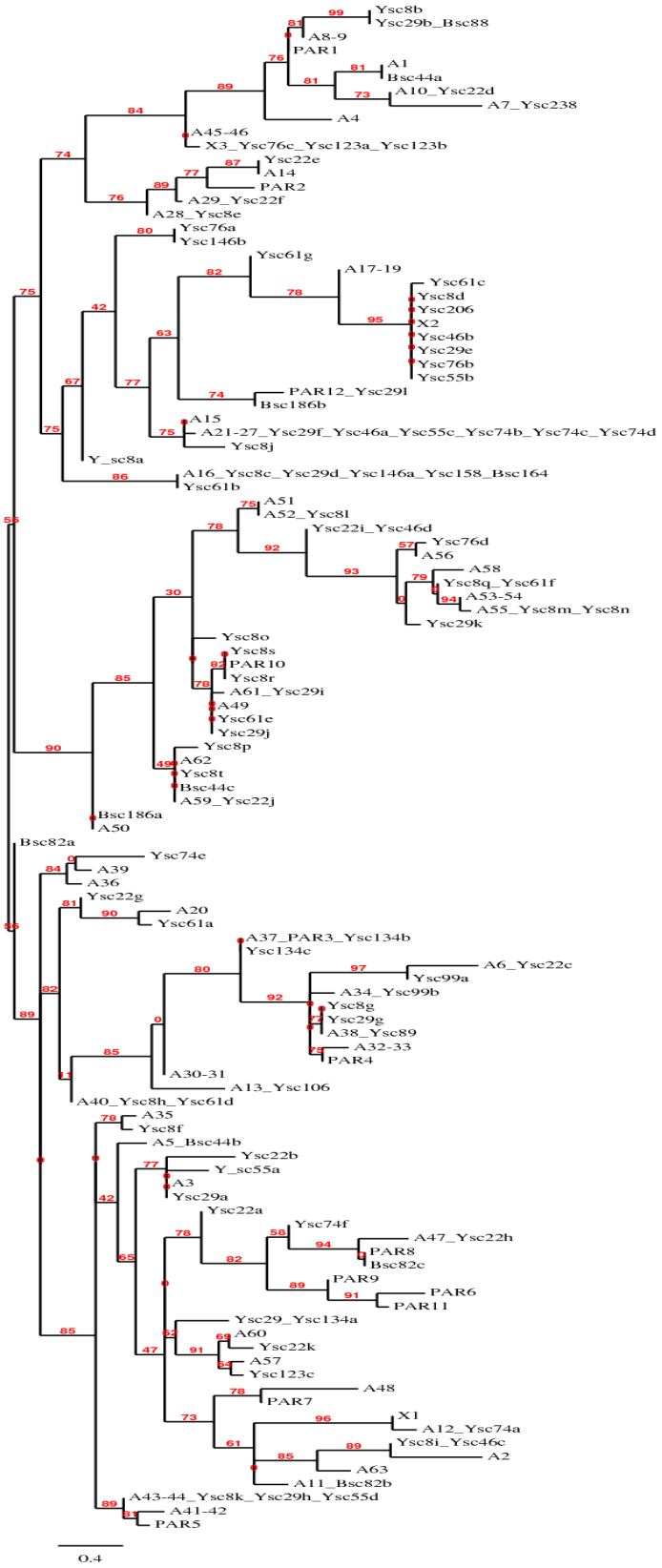


Figure S10: RT-PCR based expression studies of *Tor-3G* and *Tor-4H*, showing the presence of RNA copies of the whole element.

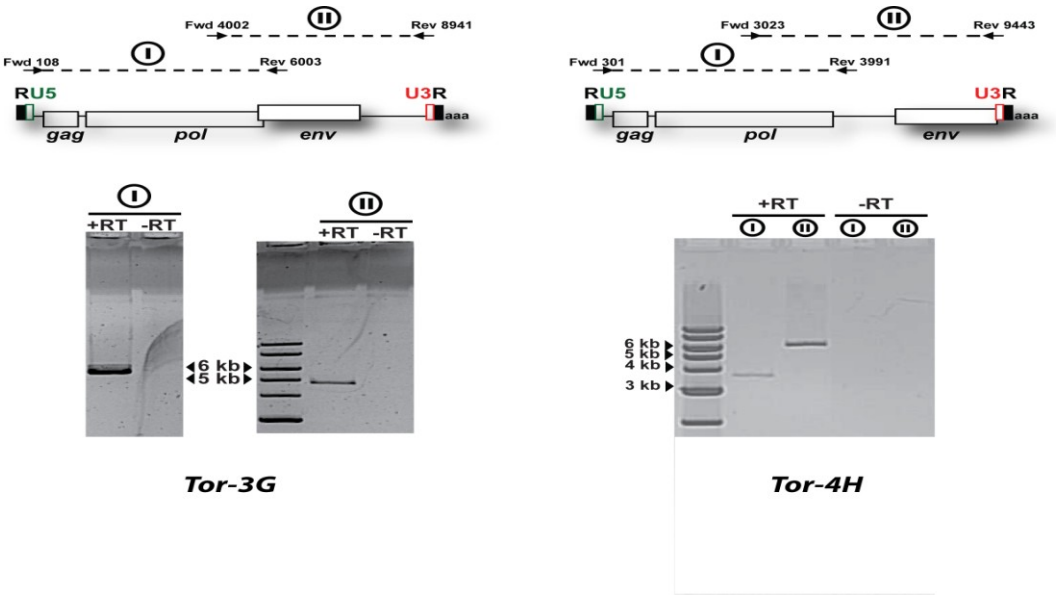


Figure S11: 5' RACE amplification of a transcript including the gag gene for five distinct Tor elements. The RACE products of expected size begin in the LTR, except for Tor-3G which is often inserted into exons and can be transcribed from the host gene.

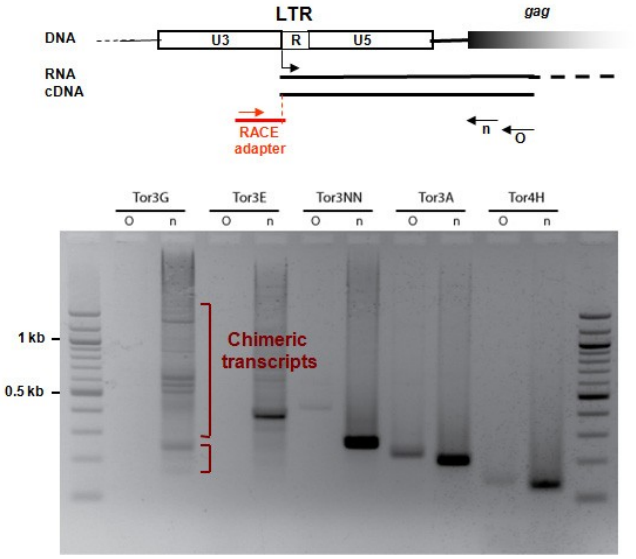


Figure S12: RT-PCR based expression studies of *Tor-3A* and *Tor-4H* downstream regions. For each element, two pairs of primers were used to amplify a transcript which covers the 3'end of the putative *env* gene.

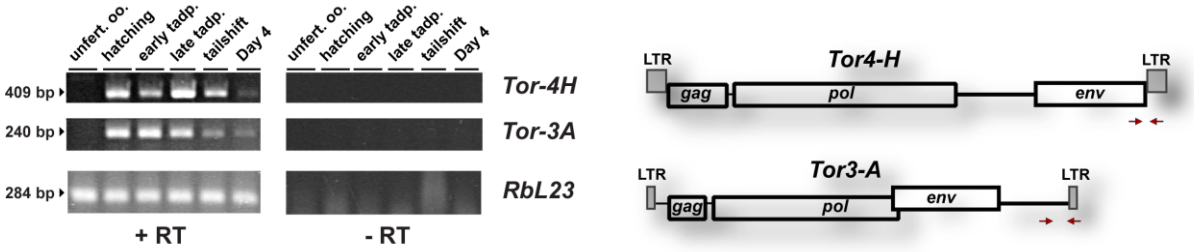


Figure S13: Exonic insertion of element *Tor-3G13* and the structure of a cloned chimeric cDNA overlapping the host exon and the element. The right panel shows the detection of 2 hemizygous individuals in five individuals of the population.

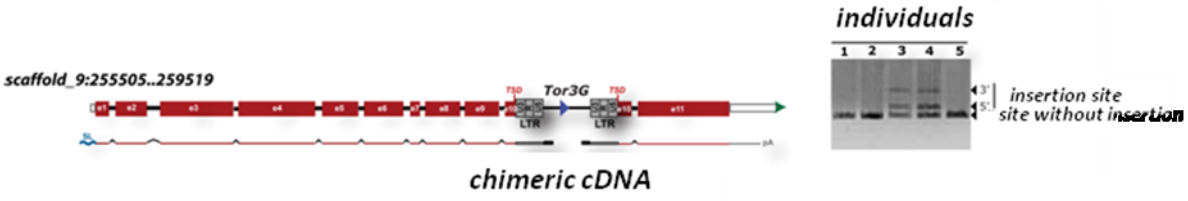


Figure S14: Presence/absence polymorphism of insertion for 5 copies of the Tor-3G element revealed by RT-PCR among 5 individuals.

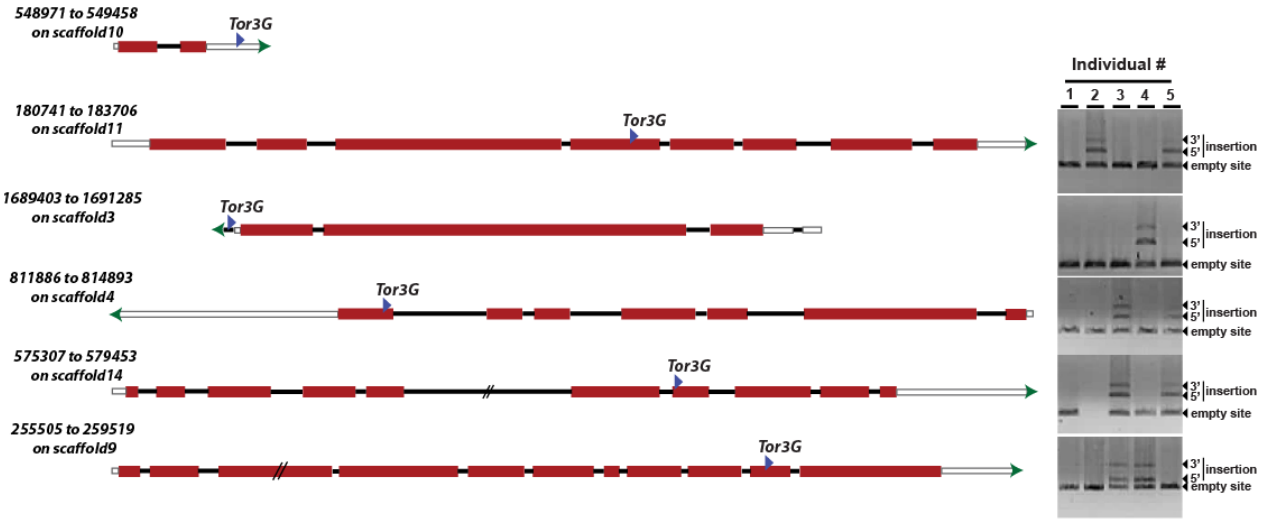


Figure S15: Polymorphism of copy number and insertion sites revealed by Southern blotting among 10 individuals. Probes are derived from a single Tor-3G element.

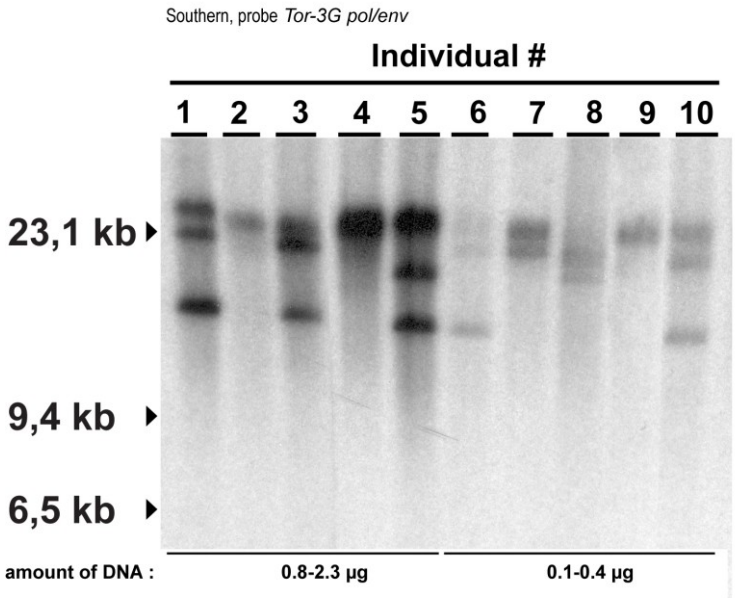


Figure S16 : Size distribution of indels >50bp showing an excess of indels for the size range 550-700bp, in which sequences were analysed for the presence of repeated TE-like elements

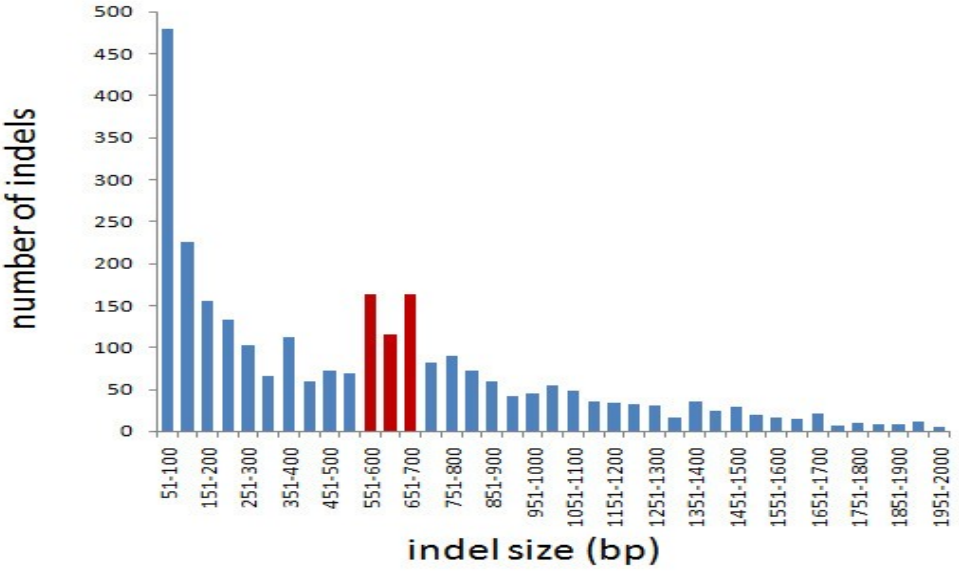


Figure S17: Density of HCEs for scaffolds of the Bergen *Oikopleura* genome and homologous contigs from the Oregon strain of *Oikopleura*. Colours from light to dark signify the density of 100% conserved blocks greater than specified lengths (from >1bp to >200bp - see key to the right). Bars above show the best matching alignment of contigs from the Oregon strain - black represents the matching regions and grey represents gaps in the alignment. Black bars below show the positions of the best scoring genewise uniprot gene annotations of the Bergen strain, with arrows indicating their annotated start sites. Developmental genes in the region are indicated by the labels above and the positions of HCEs of 100bp or greater lengths are shown below (blue).

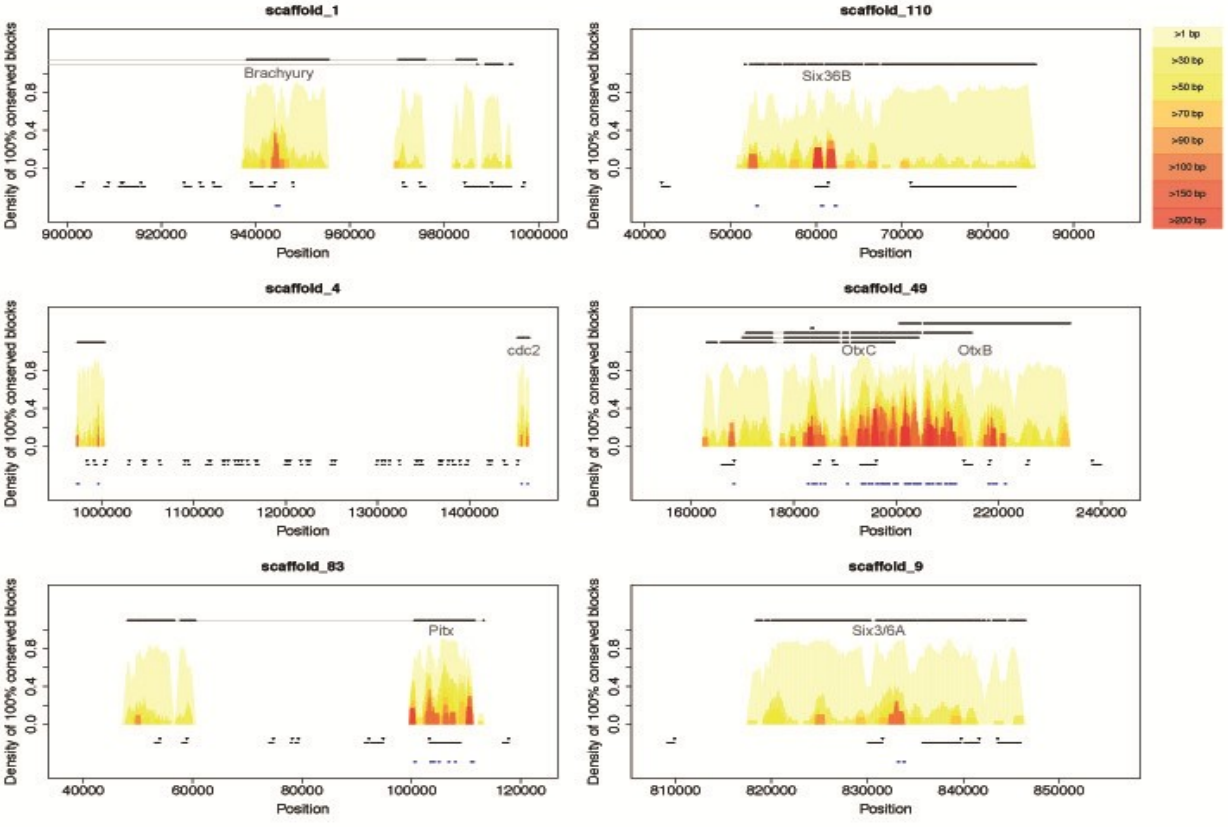


Figure S18: Long introns are more likely to be old. The introns were classified into old (sharing position with those orthologous genes in vertebrates), new (at positions specific for *Oikopleura*) or uncertain. Top chart: Relative proportion of introns (%) reliably classified as old vs. new are shown for introns distributed into bins of different sizes. Bottom chart: Proportions of introns (%) reliably classified as old and new relative to the total number of introns.

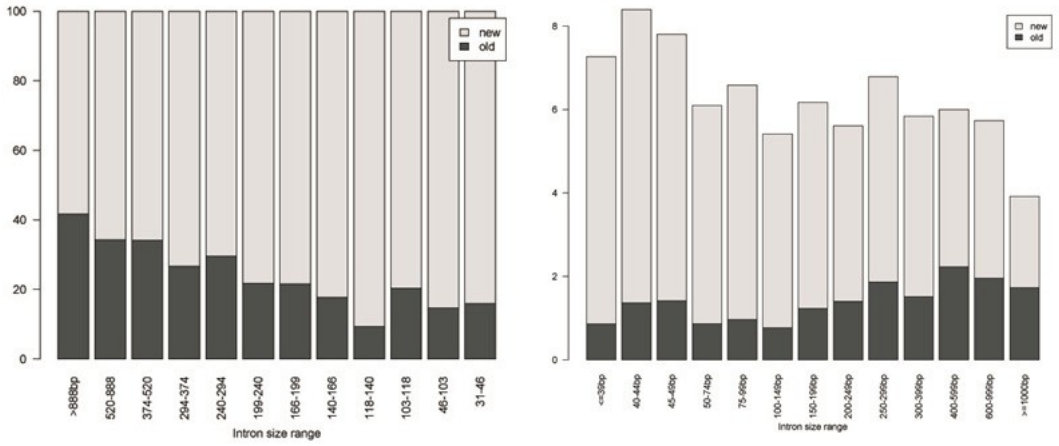


Figure S19: Relative proportion of introns reliably classified as old vs. new are shown for introns distributed into bins of different sizes, for all genes (top) and a subset of genes known to be developmental regulators (bottom). We explain the trend for longer introns to be retained more often by their increased content of regulatory elements (see text). This is corroborated by the observation that developmental genes have a significantly higher than average proportion of long and old introns.

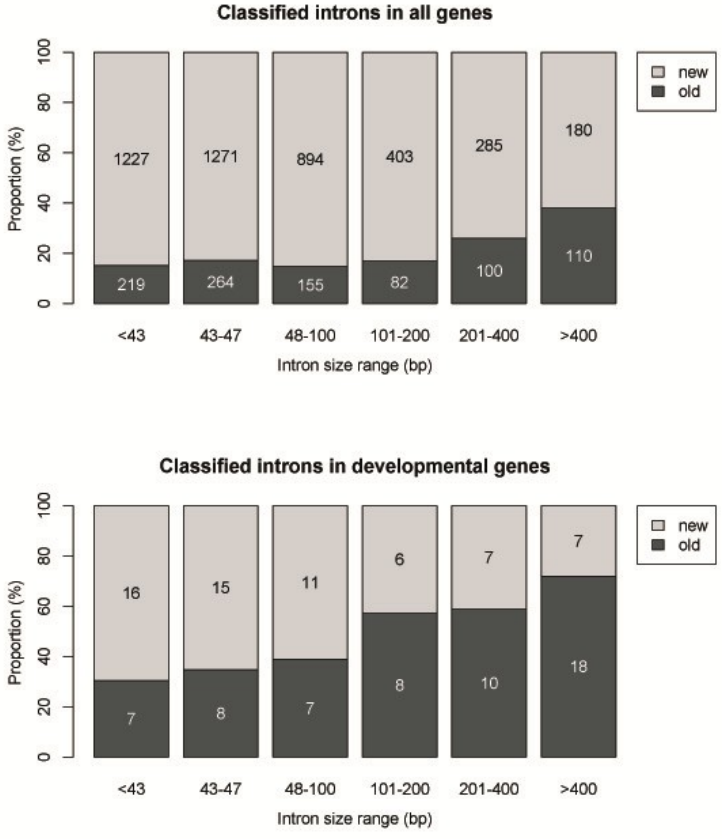


Figure S20: Full-sib family in which the fluorescence-encoding genetic variant is segregating (fluorescent mother and wild type father). The variations in patterns are essentially due to the orientation of specimens.

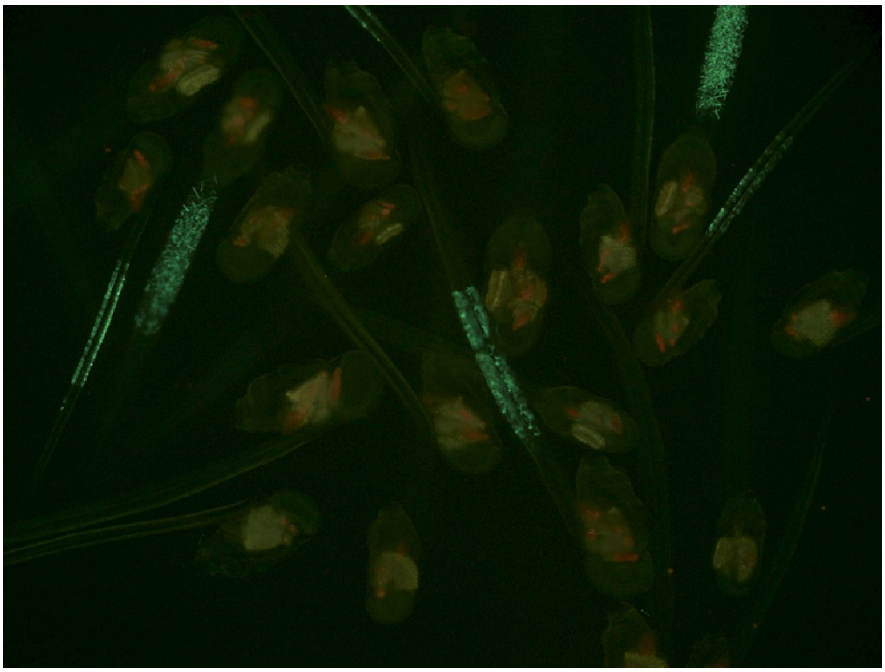


Figure S21 : Intron size distribution. The distribution was plotted for introns that were validated by cDNAs in the size range between 1 and 200 nucleotides, *i.e.* the vast majority of *Oikopleura* introns. The total number of introns in each category (GT-AG, GA-AG, GC-AG, GG-AG) is indicated between brackets.

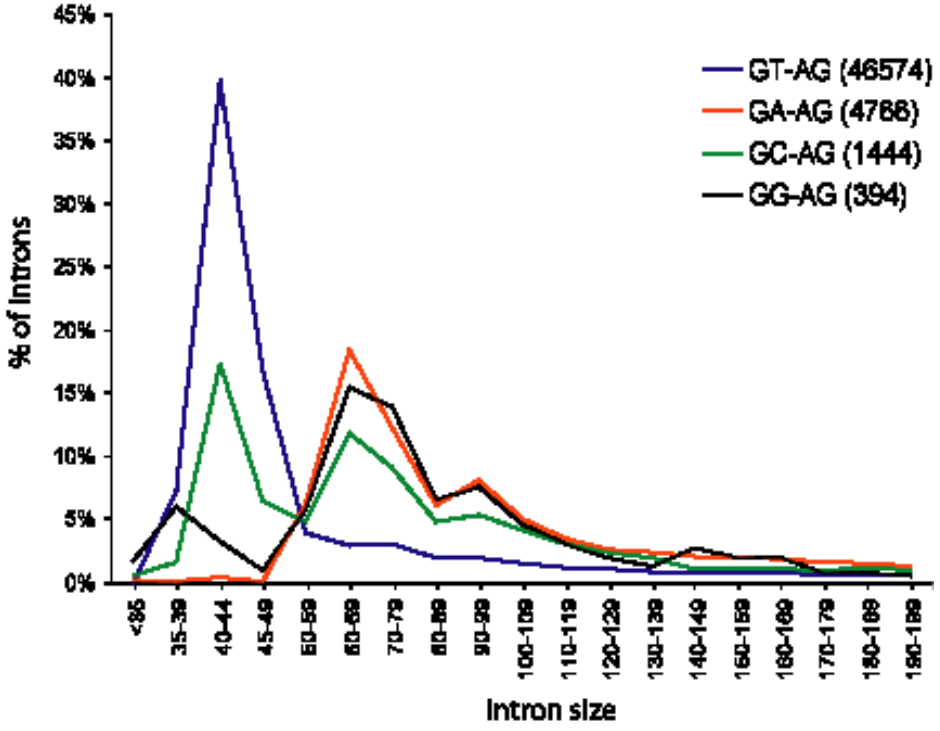


Figure S22: Intron logos. The logos were obtained using the weblogo software (*S153*), for GT-AG and GA-AG introns that were validated by cDNAs in four size ranges : all sizes, < 55 nt, 55 to 100 nt and >100 nt.

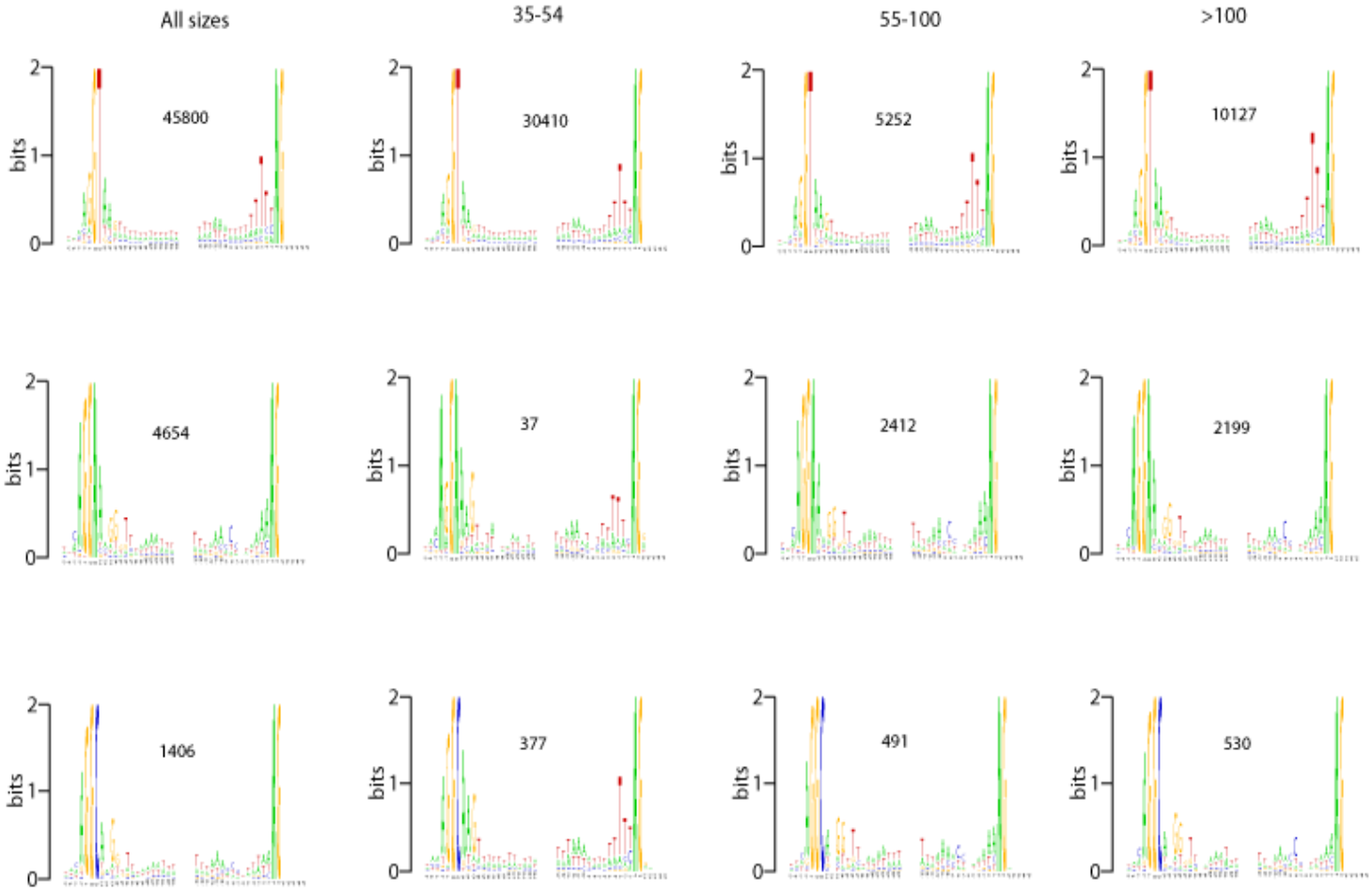


Figure S23: Intron retention rate as a function of the number of introns in each category. Intron retentions events were identified using the cDNA data. They correspond to introns that were spliced out in some cDNA sequences and retained in other cDNA sequences.

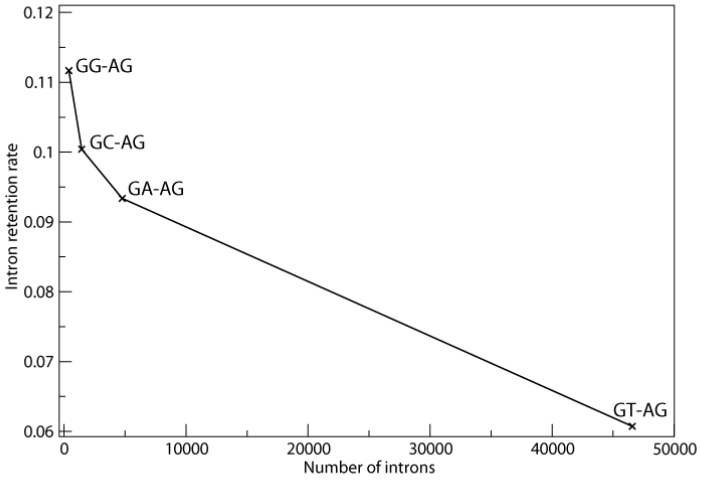


Figure S24: Distribution of expression level for different intron categories. The expression level of each gene was calculated as the number of ESTs that could be mapped uniquely to the reference genome and were overlapping each gene.

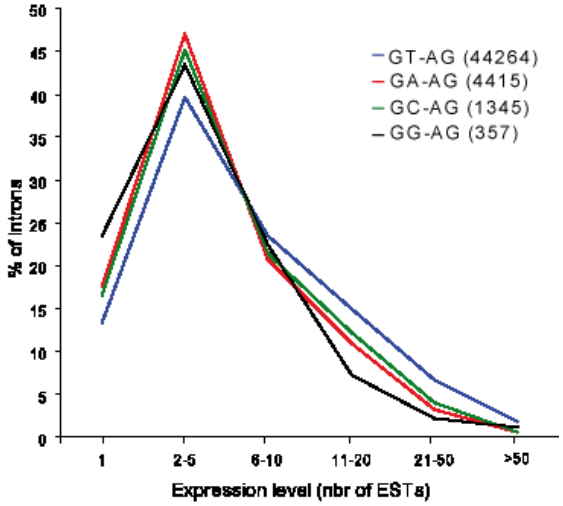
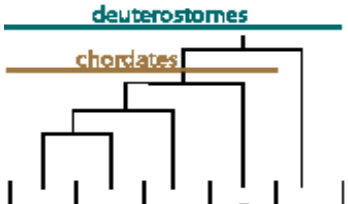


Figure S25: Intron profiles. Intron conservation profiles in the 5 species are encoded as a succesion of "X" (intron present in the species) and "0" (intron absent in the species). The number of introns displaying each profile is indicated as well as the corresponding category (new, old, ancestral to chordates).



Oikopleura	Ciona	Human	Amphioxus	Sea urchin	Number of introns	categories of Oikopleura introns	categories of introns ancestral to chordates
X	0	0	0	0	4260	new	not ancestral
X	0	0	X	0	28	unclassified	not ancestral
X	X	0	0	0	127	unclassified	not ancestral
X	0	X	0	0	29	unclassified	not ancestral
X	0	0	0	X	14	unclassified	not ancestral
X	X	X	X	X	499	Old	ancestral, retained in Od
X	0	X	X	X	144	Old	ancestral, retained in Od
X	X	X	X	0	130	Old	ancestral, retained in Od
X	X	X	0	X	65	Old	ancestral, retained in Od
X	X	0	X	X	23	Old	ancestral, retained in Od
X	0	X	X	0	26	Old	ancestral, retained in Od
X	X	X	0	0	16	Old	ancestral, retained in Od
X	0	X	0	X	12	Old	ancestral, retained in Od
X	0	0	X	X	8	Old	ancestral, retained in Od
X	X	0	X	0	6	Old	ancestral, retained in Od
X	X	0	0	X	7	Old	ancestral, retained in Od
0	X	X	X	X	3917	-	ancestral, lost in Od
0	0	X	X	X	3235	-	ancestral, lost in Od
0	X	X	X	0	605	-	ancestral, lost in Od
0	X	X	0	X	411	-	ancestral, lost in Od
0	X	0	X	X	339	-	ancestral, lost in Od
0	0	X	X	0	746	-	not ancestral
0	0	0	X	X	723	-	not ancestral
0	0	X	0	X	433	-	not ancestral
0	X	X	0	0	312	-	not ancestral
0	X	0	X	0	115	-	not ancestral
0	X	0	0	X	74	-	not ancestral
0	X	0	0	0	2415	-	not ancestral
0	0	0	X	0	902	-	not ancestral
0	0	X	0	0	603	-	not ancestral
0	0	0	0	X	775	-	not ancestral

Figure S26: Logos of new and old GT-AG introns.

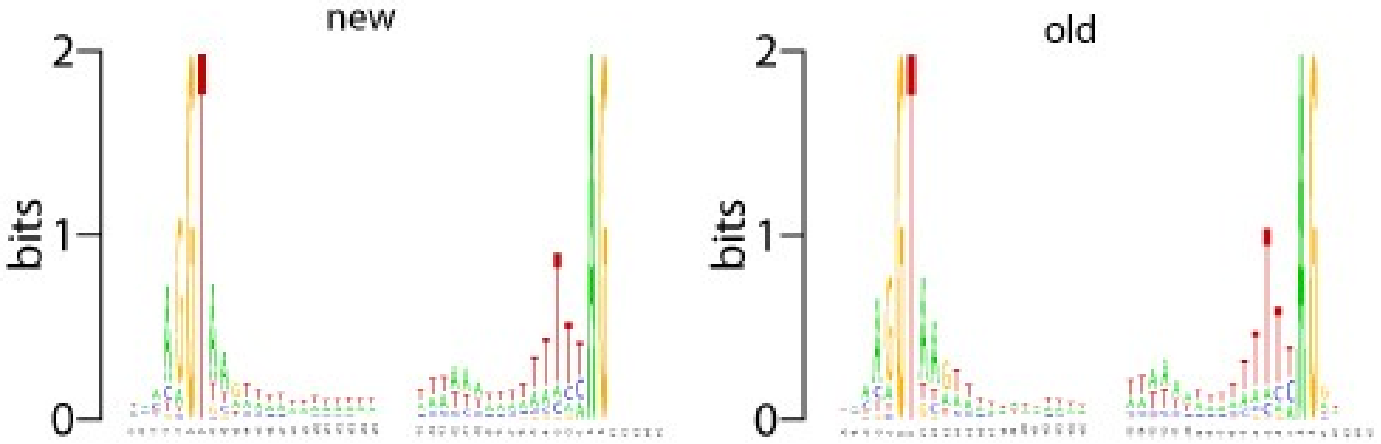


Figure S27. Information contents calculated for new and old introns in Sverdlov *et al*. (*S121*) and in *Oikopleura*.

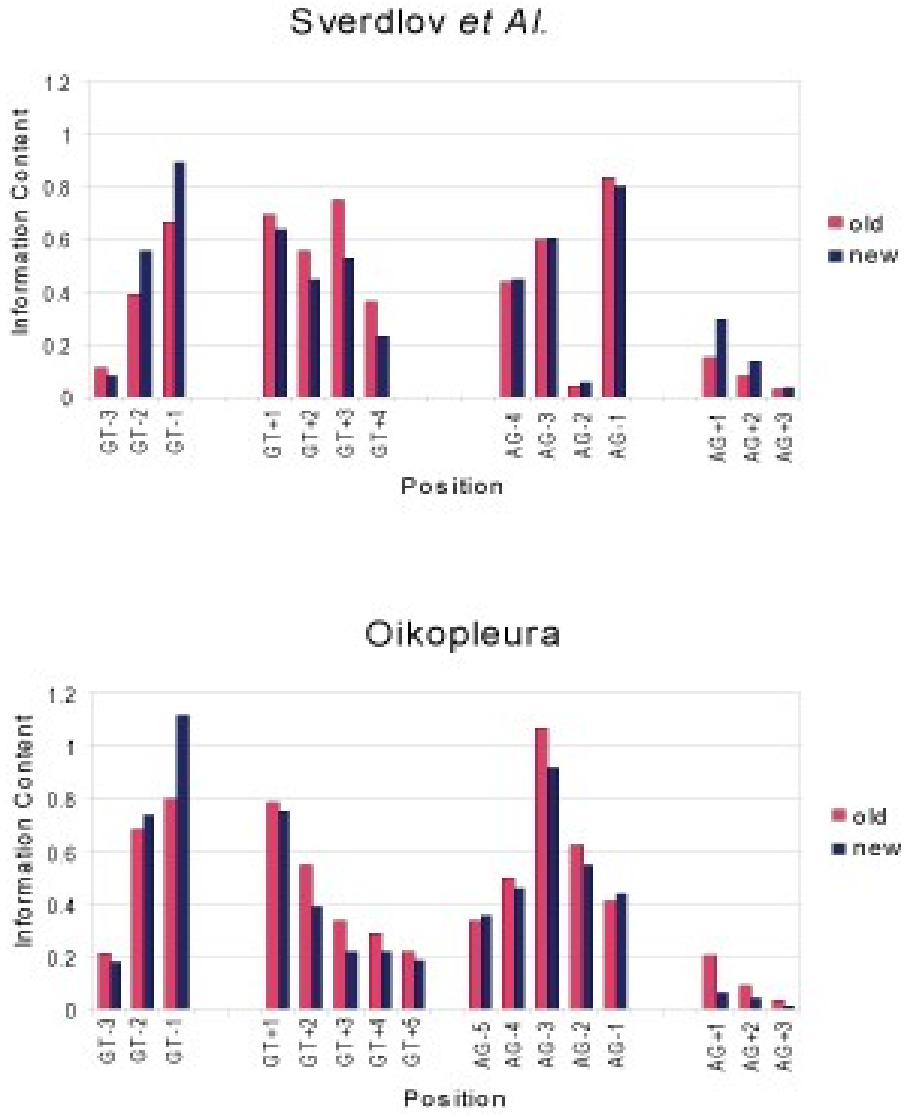


Figure S28: Splicing of homologous intron pair 5 and 6 in gene GSOIDI00008841001 on scaffold_3. A) genomic DNA extracted from 5 females was genotyped and splicing was confirmed by RT-PCR from cDNA extracted from each of the same females. B) Sequence comparison of cloned PCR and RT-PCR products and the genome sequence.

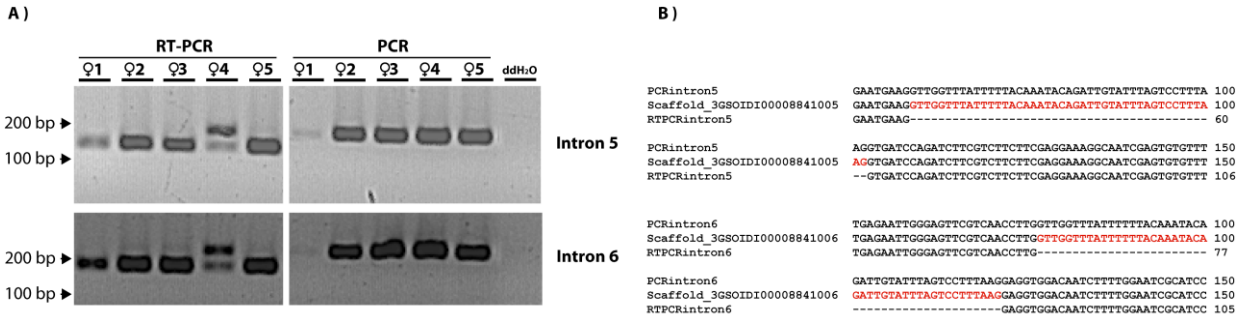


Figure S29: Genotyping of 23 individuals of both sexes for the insertion in scaffold_431:15391..15936.

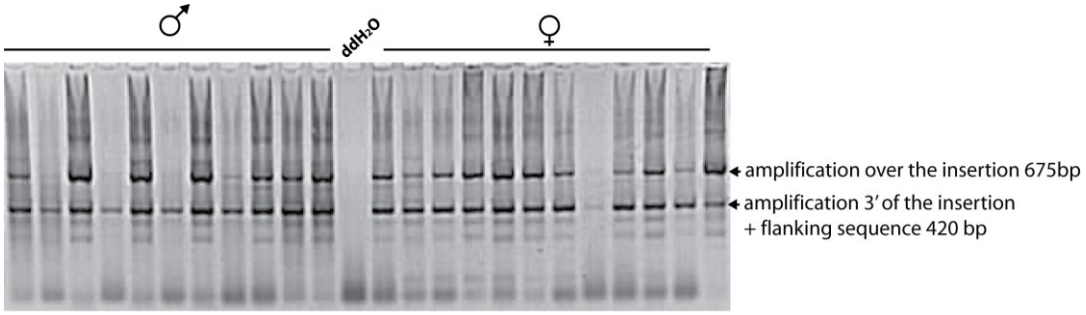


Figure S30 : Genotyping of 10 individuals of both sexes for the insertion in scaffold_70:129142..129424.

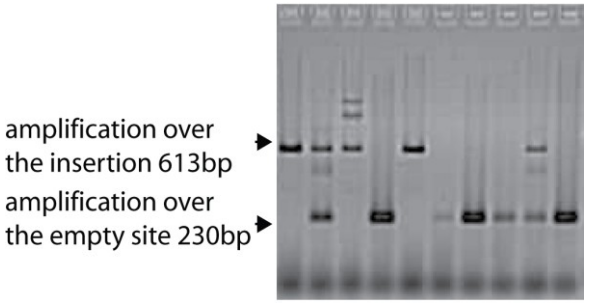


Figure S31: Splicing of an intron located in scaffold_42:311401..312139. A) gene model the intron in the 3'UTR and the expected amplification products B) Presence of the intron was tested by PCR on genomic DNA of 9 individuals, and the presence of the intron in RT-PCR products was tested in parallel.

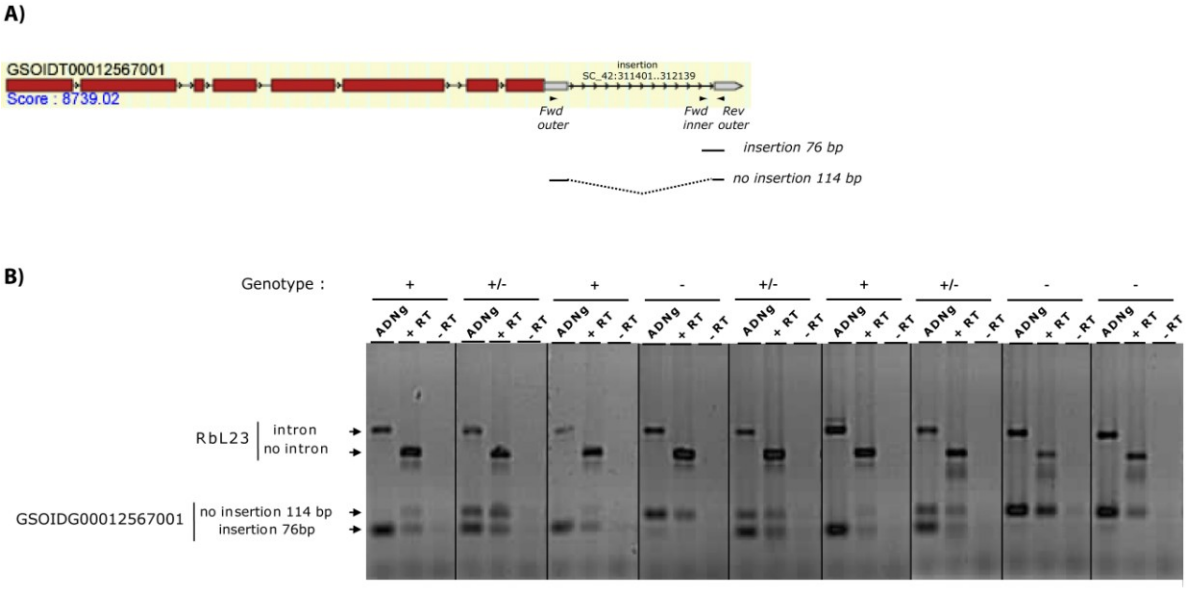


Figure S32: Figure 3A of the main paper presents a histogram and model fit for duplicate gene pairs derived from the top BLAST hit of each gene and fit according to the mixture model described in Supplementary text 13. This is expected to provide an underestimate where slightly older duplicate pairs that each have a more recent retained duplicate will not be counted. An overestimate of these will be provided by counting all BLAST hits, where these older duplication events will then be oversampled by inclusion when identified by BLAST in all underlying duplicate pairs. A single large gene family with 103 members was excluded from this analysis.

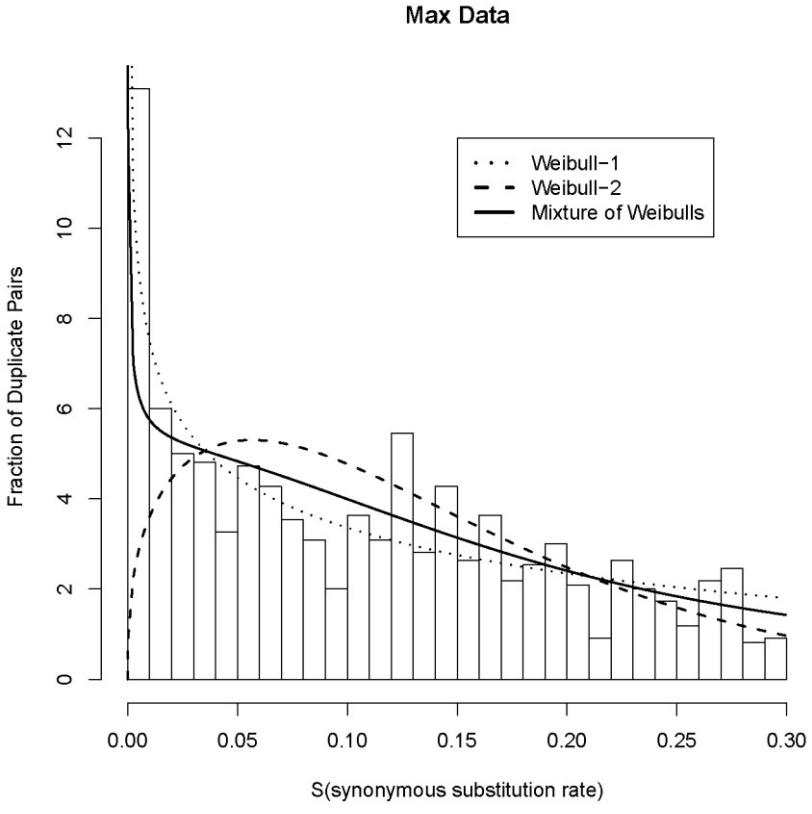


Figure S33: Using PAML, a likelihood ratio was performed between a model where dN/dS was fixed to 1 and where dN/dS was estimated for each of the recent duplicate pairs. The fraction of recent duplicates where dN/dS<>1 was supported is shown for each bin of dS values.

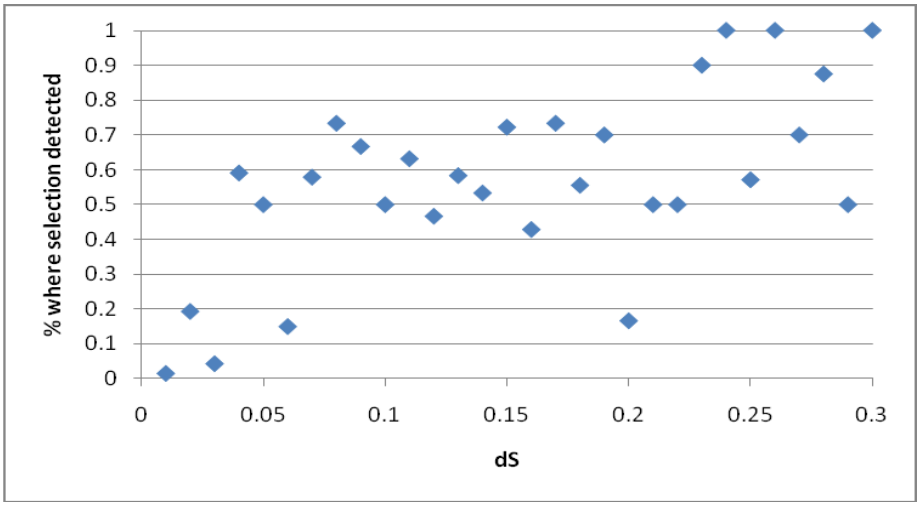


Figure S34: Replotting of Figure 3A (inset), where the accumulation of nonsynonymous substitution and the fit of the model are shown for recent duplicates (up to dS=0.3). This shows the accumulation of nonsynonymous substitutions as synonymous substituions accumulate.

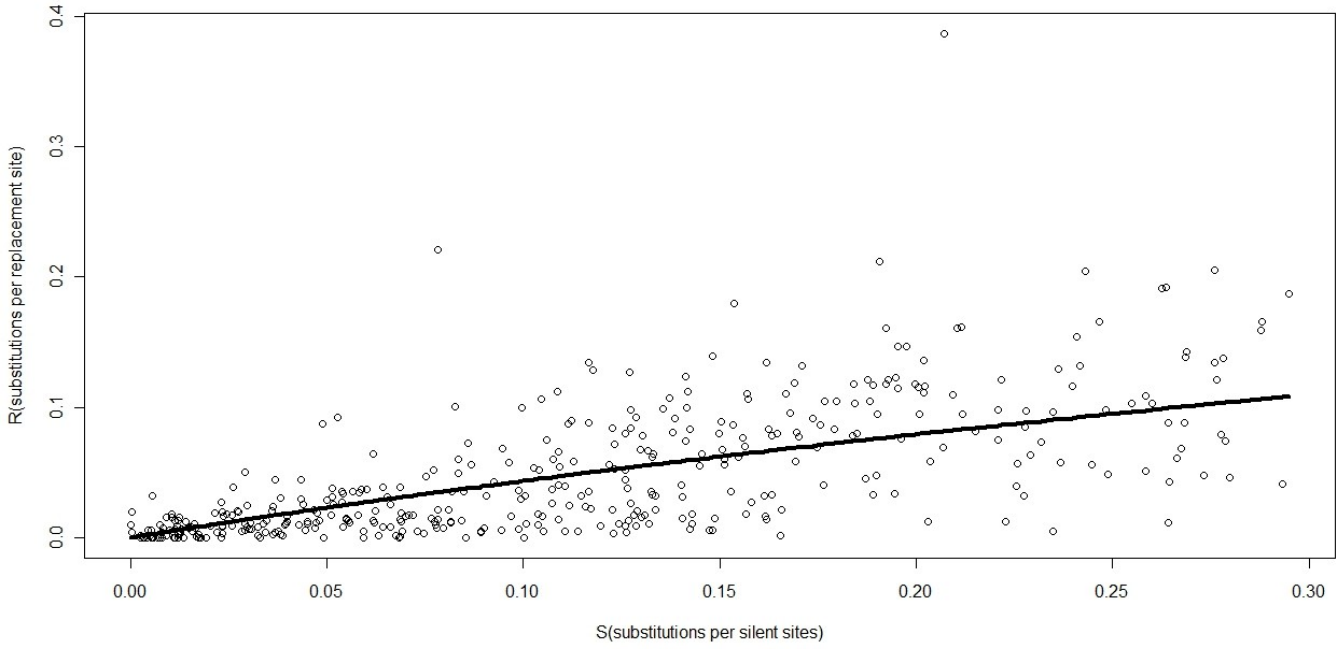
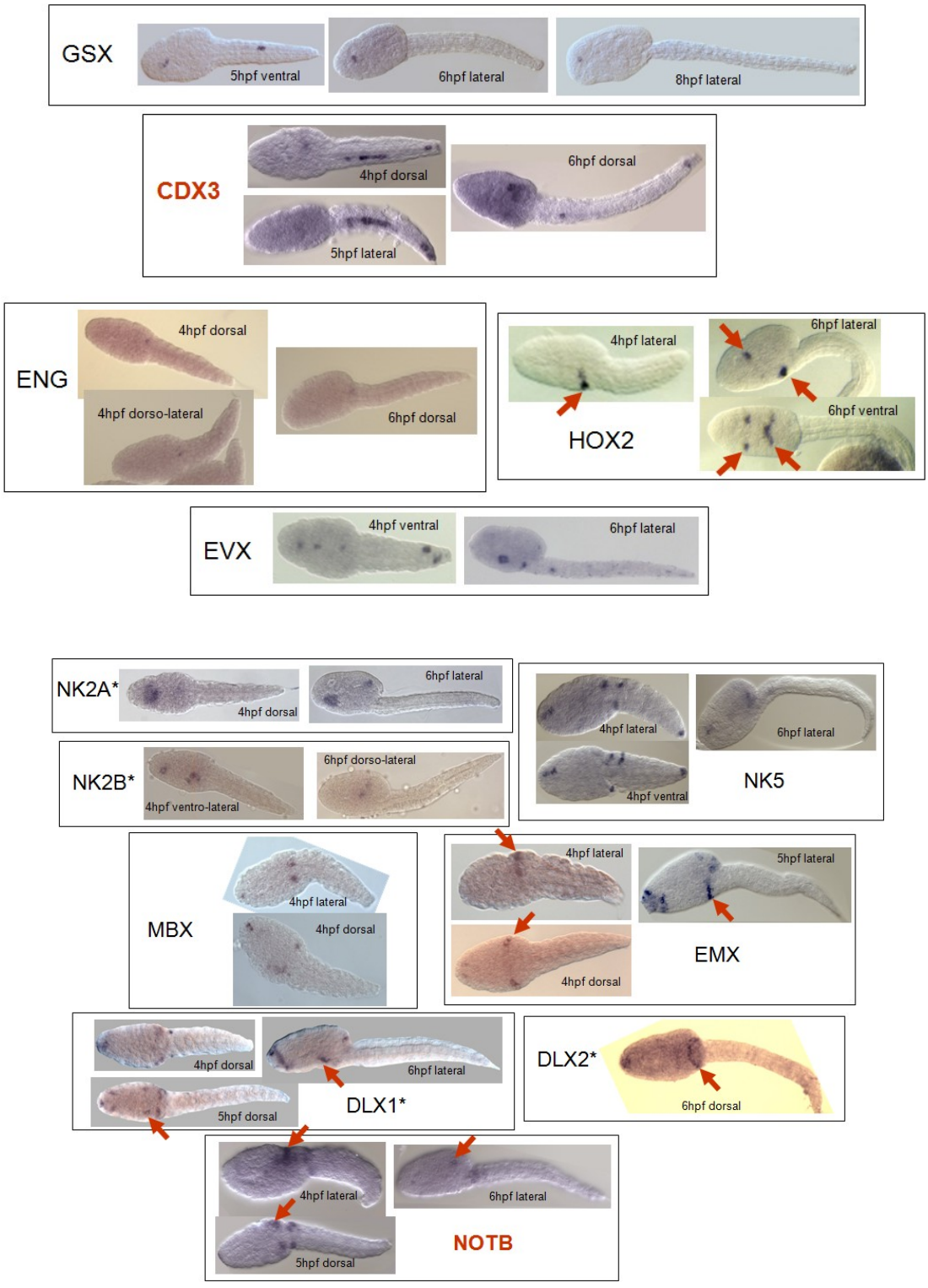
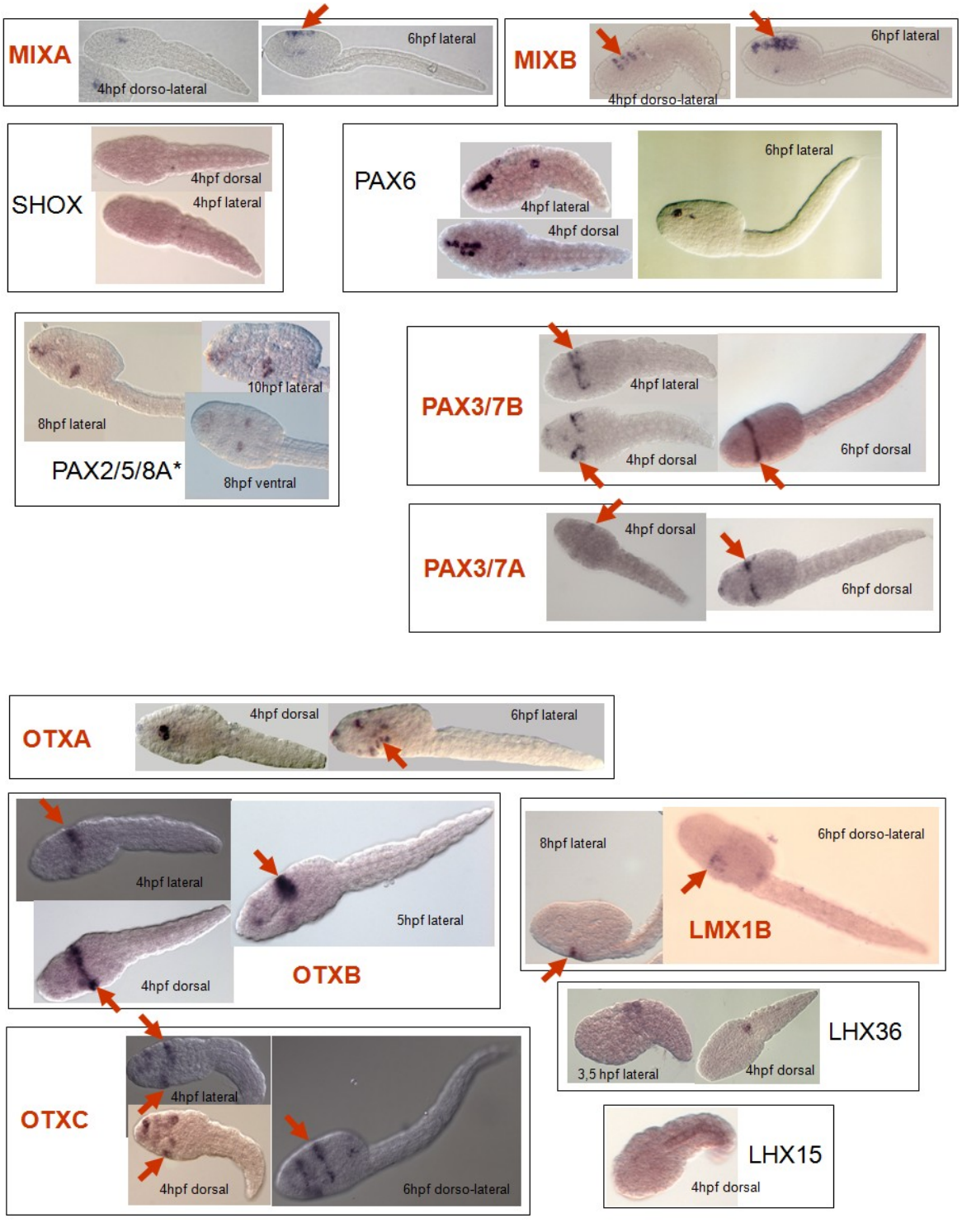
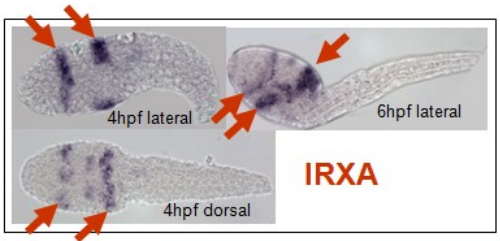
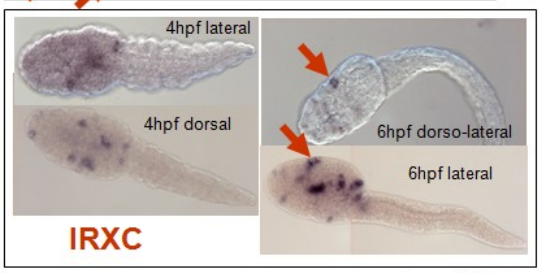


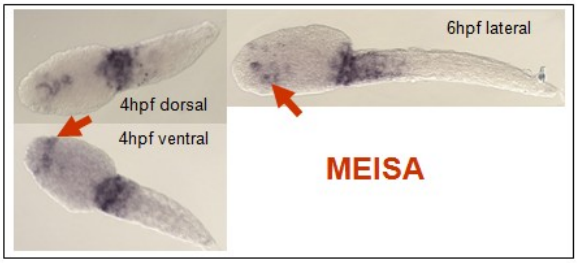
Figure S35: Collection of expression patterns of homeobox genes partially shown in Figure 3B of the main article. Methods for RNA-RNA *in situ* hybridization have been described (S*177*). In all following illlustrations, the name of the gene is in capital letters, either black for genes of non-amplified groups or red for genes of amplified groups. Note that the phylogenetic analysis has confirmed or concluded that NK2A/NK2B, PAX258A/PAX258B, DLX1/DLX2/DLX3 (signalled by an asterisk) are most likely ancient duplicates preceding the *Ciona/Oikopleura* split and therefore considered as members of non-amplified groups.

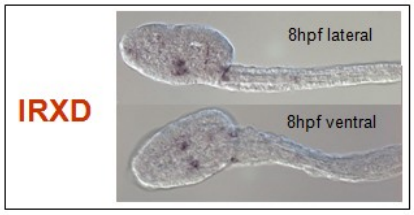


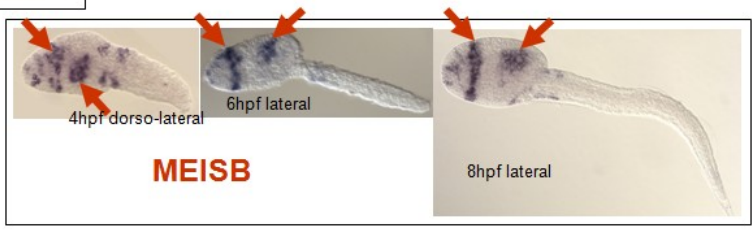

MIXA
4hpf dorso-lateral
6hpf lateral
MIXB
4hpf dorso-lateral
6hpf lateral
SHOX
4hpf dorsal
4hpf lateral
PAX6
4hpf lateral
4hpf dorsal
6hpf lateral
8hpf lateral
10hpf lateral
PAX2/5/8A*
8hpf ventral
PAX3/7B
4hpf lateral
4hpf dorsal
6hpf dorsal
PAX3/7A
4hpf dorsal
6hpf dorsal
OTXA
4hpf dorsal
6hpf lateral
4hpf lateral
5hpf lateral
4hpf dorsal
OTXB
8hpf lateral
6hpf dorso-lateral
LMX1B
3,5 hpf lateral
4hpf dorsal
LHX36
OTXC
4hpf lateral
4hpf dorsal
6hpf dorso-lateral
4hpf dorsal
LHX15

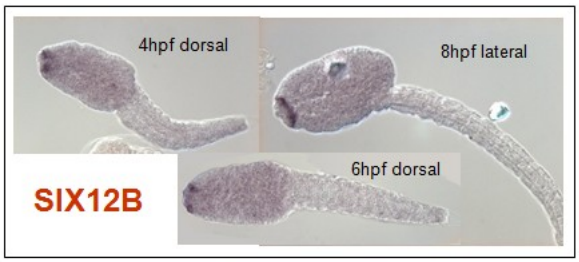

4hpf lateral
6hpf lateral
4hpf dorsal
IRXA

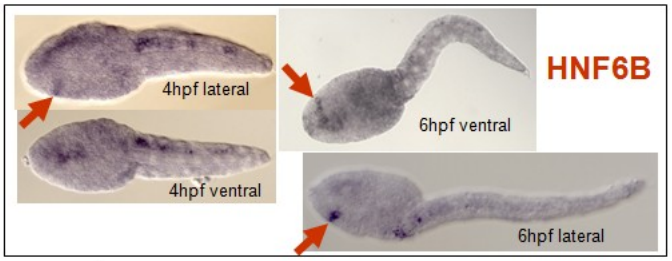

4hpf lateral
4hpf dorsal
6hpf dorso-lateral
6hpf lateral
IRXC


6hpf lateral
4hpf dorsal
4hpf ventral
MEISA

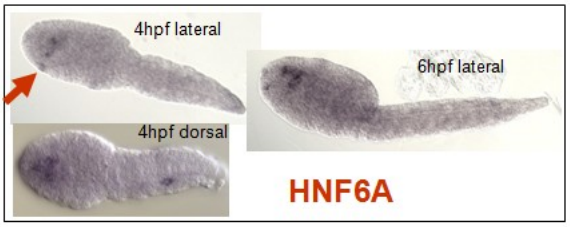

8hpf lateral
IRXD
8hpf ventral


4hpf dorso-lateral
6hpf lateral
MEISB
8hpf lateral


4hpf dorsal
8hpf lateral
6hpf dorsal
SIX12B


4hpf lateral
6hpf ventral
HNF6B
4hpf ventral
6hpf lateral


4hpf lateral
6hpf lateral
SIX36B
4hpf ventral


4hpf lateral
6hpf lateral
4hpf dorsal
HNF6A

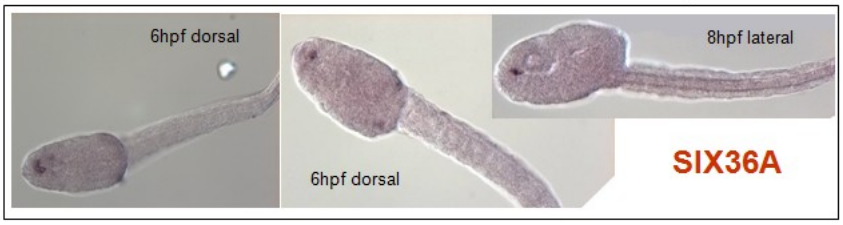

6hpf dorsal
8hpf lateral
6hpf dorsal
SIX36A

Figure S36: Dot matrix representing the distribution of *Oikopleura dioica* - *Ciona intestinalis* orthologous genes along their respective chromosome. Each black dot is an orthologous gene, and the X and Y coordinates are given by the rank of the corresponding extant gene on its chromosome.

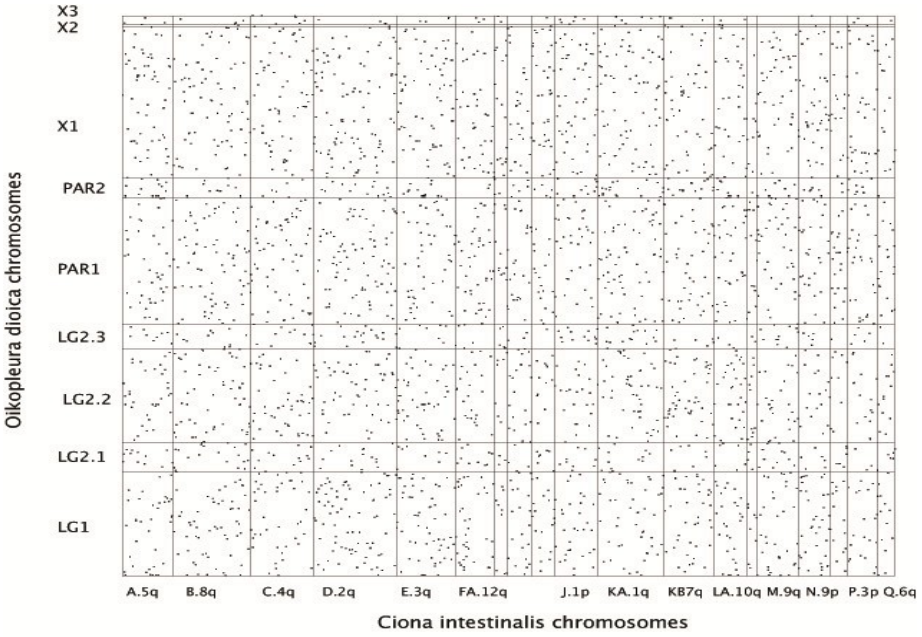


Figure S37. Species tree representing the six species used for the analysis of synteny conservation. The names of their respectivy last common ancestors are indicated. Inferred protochromosomes of the ancestor of chordates (black circle) were used as reference to compare with the five non-human species due to its central location in the tree.

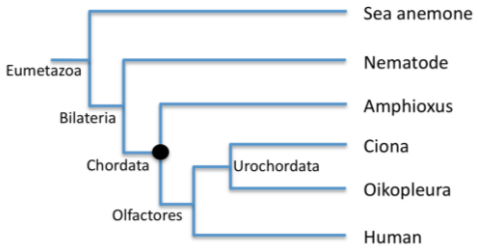


Figure S38. Local rearrangements. Distances between genes in the human genome (number of intervening genes; "d" on the X-axis) were grouped into four classes, if the two corresponding orthologs in a given metazoan genome are separated by less than 3 genes. One hundred randomizations of the position of all the genes in the five metazoan genomes were performed to compute an average distance in the human genome (red circle; error bar = twice the standard deviation).

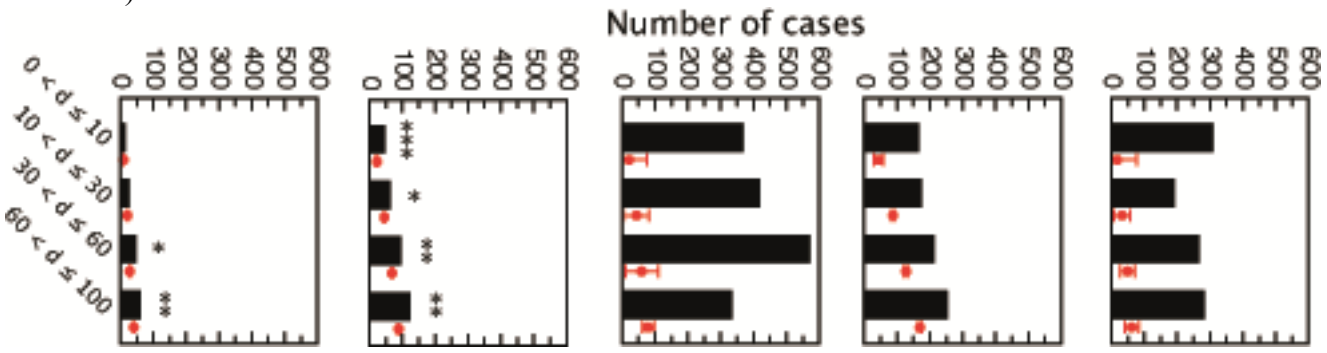


Table S1: Assembly overview (scaffolds). Only 470 ESTs clusters (present in the initial assembly) are not present in the reference assembly, around 3.1%.

	Initial assembly	Reference assembly	Allelic assembly
Number of scaffolds	34,559	1,260	4,196
Cumulated size of the assembly	148Mb	70,4Mb	45Mb
Scaffold's N50	37Kb	395Kb	21,8Kb
ESTs clusters	15,622	15,152	11,402

Table S2: Top 50 Interpro domains in the *O. dioica* genome.

IPR000719	404	Protein kinase
IPR011009	310	Protein kinase-like
IPR002290	235	Serine/threonine protein kinase
IPR001245	204	Tyrosine protein kinase
IPR008271	185	Serine/threonine protein kinase, active site
IPR001254	163	Peptidase S1 and S6, chymotrypsin/Hap
IPR013032	162	EGF-like region
IPR011043	156	Galactose oxidase, central
IPR012677	150	Nucleotide-binding, alpha-beta plait
IPR007087	147	Zinc finger, C2H2-type
IPR011992	142	EF-Hand type
IPR011046	139	WD40-like
IPR001680	136	WD-40 repeat
IPR002110	132	Ankyrin
IPR002048	130	Calcium-binding EF-hand
IPR001211	128	Phospholipase A2
IPR001881	114	EGF-like calcium-binding
IPR009003	114	Peptidase, trypsin-like serine and cysteine
IPR001841	104	Zinc finger, RING-type
IPR001314	103	Peptidase S1A, chymotrypsin
IPR000152	97	Aspartic acid and asparagine hydroxylation site
IPR012287	97	Homeodomain-related
IPR009053	96	Prefoldin
IPR000504	93	RNA-binding region RNP-1 (RNA recognition motif)
IPR011991	91	Winged helix repressor DNA-binding
IPR012335	89	Thioredoxin fold
IPR001356	88	Homeobox
IPR011990	79	Tetratricopeptide-like helical
IPR006210	77	EGF
IPR009057	77	Homeodomain-like
IPR012336	75	Thioredoxin-like fold
IPR003593	73	AAA ATPase
IPR000859	70	CUB
IPR001611	69	Leucine-rich repeat
IPR001452	67	Src homology-3
IPR013783	66	Immunoglobulin-like fold
IPR012337	62	Polynucleotidyl transferase, Ribonuclease H fold
IPR010989	60	t-snare

IPR011989	60	Armadillo-like helical
IPR006209	59	EGF-like
IPR008957	59	Fibronectin, type III-like fold
IPR011993	59	Pleckstrin homology-type
IPR002347	58	Glucose/ribitol dehydrogenase
IPR013091	58	EGF calcium-binding
IPR014001	58	DEAD-like helicases, N-terminal
IPR000884	57	Thrombospondin, type I
IPR009030	57	Growth factor, receptor
IPR001623	56	Heat shock protein DnaJ, N-terminal
IPR009072	56	Histone-fold
IPR001507	55	Endoglin/CD105 antigen

Table S3: « localisation insertions » displays the insertion rate in different genomic compartments.

	All genome					Operons				
	All	CDS	UTR	intronic	Intergenic	All	CDS	UTR	intronic	Intergenic
Size of the compartment	30.7 Mb	7.4 Mb	1.5 Mb	6.4 Mb	15.4 Mb	4 Mb	2.4 Mb	0.4 Mb	0.9 Mb	0.3 Mb
Number of indels (>50nt)	3059	477	175	797	1610	365	126	46	191	2
Indel rate	100/Mb	64/Mb	117/Mb	125/Mb	105/Mb	91/Mb	52.5/Mb	115/Mb	212/Mb	0.5/Mb

Table S4: Statistics of the nine concatenated datasets.

#missing species	#genes	#positions	% missing positions
0 (S26)	114	32,650	3
1 (S25)	220	63,858	7
2 (S24)	251	67,054	11
3 (S23)	211	57,848	16
4 (S22)	181	47,463	20
5 (S21)	140	38,483	23
6 (S20)	126	29,245	27
7 (S19)	134	30,169	31
8 (S18)	105	22,732	35

Table S5: Posterior probability for several groups using the CAT+Γ model.

	18	19	20	21	22	23	24	25	26
Bilateria	1.00	0.47	1	1.00	1.00	1.00	1.00	0.98	1.00
Cnidaria+Placozoa	0.01	1.00	0.61	0.57	0.02	0.98	0.00	0.24	0.44
Eumetazoa	0.99	0.00	0.12	0.34	0.98	0.03	1.00	0.56	0.5
Chordata+Protostomia	0.05	0.79	1.00	1.00	0.06	0.00	1.00	0.98	1.00
Echinodermata+Protostomia	0.00	0.03	0.00	0.00	0.94	0.00	0.00	0.00	0.00
Deuterostomia	0.95	0.00	0.00	0.00	0.01	1.00	0.00	0.00	0.00
Chordata	1.00	0.97	1.00	1.00	1.00	1.00	1.00	0.98	1.00
Olfactores	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	**1.00**
Tunicata	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Protostomia	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00
Ecdysozoa	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00
Arthropoda	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00
Acyrthosiphon+Pediculus	0.94	0.60	1.00	1.00	1.00	0.31	1.00	0.98	1.00

Table S6: Bootstrap support for several groups using the LG+Γ model.

	18	19	20	21	22	23	24	25	26
Bilateria	91	1	88	45	100	52	100	100	97
Cnidaria+Placozoa	10	100	92	61	0	51	4	1	3
Eumetazoa	88	0	1	39	100	49	96	99	97
Chordata+Protostomia	3	1	29	24	26	3	35	96	4
Echinodermata+Protostomia	0	0	3	0	0	0	0	0	0
Deuterostomia	71	14	57	8	74	73	65	4	90
Chordata	9	61	92	16	96	42	96	100	4
Olfactores	**74**	**51**	**99**	**31**	**100**	**86**	**100**	**100**	**99**
Tunicata	74	74	99	49	100	87	100	100	100
Protostomia	76	4	89	45	100	66	100	100	97
Ecdysozoa	77	4	89	45	100	66	100	100	97
Arthropoda	100	100	100	100	100	100	100	100	100
Acyrthosiphon+Pediculus	94	4	90	51	94	20	99	100	30

Table S7: List of proteins involved in mammalian DNA repair and their candidate homologs in *Oikopleura dioica* (continues in next page).

		Protein	Homolog in *Oikopleura dioica* (gene model or EST)
DNA synthesis	DNA polymerases	PCNA	00003167, 00004244, 00012375, 00003105
		RFC1 to 5	00013837, 00000864, 00009093, 00002451, 00006331
		POLA1	00017937
		POLE	00011373
		POLD1	00008839
		POLK	00016471
		POLG	00004251
		POLQ	00009658
		POLB	not found
	dN transferases	REV1	00013895, 00001239
		REV3L	00000937
BER		PARP1	00016878, 00002160
		PARP2	00007472
	DNA glycosylases	NTHL1	00006445
		UNG	00003242
		TDG	00011485
		OGG1	00015021
	AP endonucleases	APEX1	00007717
		APEX2	not found
	Long patch	FEN1	KT0AAA21YC14 (ESTs different stages)
		LIG1	00015612
	Short patch	XRCC1	00010876
		LIG3	not found
		ERCC5	00000288
		ERCC1	00004492
NER		ERCC4	00010617
		RPA1	00012410
		XPA	00005728
	TC-NER	CCNH	00003571
		CDK7	00007022
		MNAT1	00001638
		ERCC2	00014013
		ERCC3	00003289
		GTF2H1 to 5	00011858, 00007056, 00006986, 00000277, 00000417
		ERCC6	00016263
		ERCC8	00000047
		MMS19	00005077
		XAB2	00005623
		DDB1	00006768
	GG-NER	RAD23A/RAD23B	00000766
		XPC	00008773
MMR		MLH1	00014710
		MSH2	00005427
		MSH3	not found
		MSH6	00015076
		PMS1/PMS2	00012470
		EXO1	00001367
Signaling of DNA damage	9-1-1 complex	RAD9	00003312
		RAD1	00007233
		HUS1	00000546
		RAD17	00001465
	Signaling of DSB	ATR	00014911
		CHEK1	00008508
		ATM	not found, ATM subfamily candidates: 00027946, 00031536
		CHEK2	not found
		RAD17	00001465
		BRCA1	00000914
		BARD1	00014237

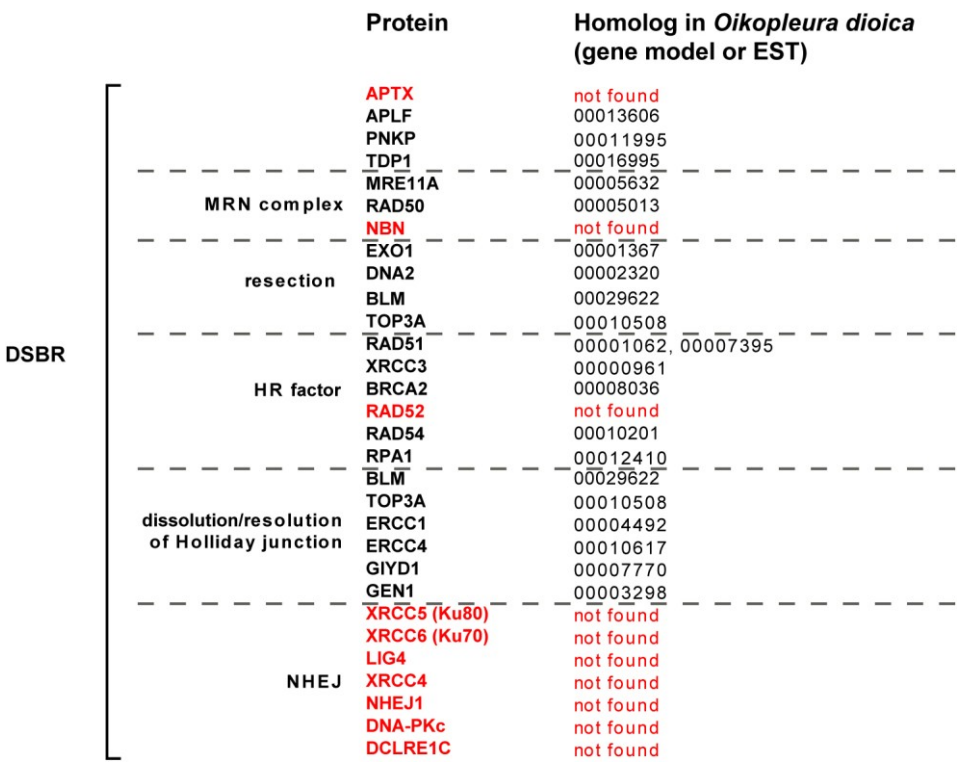


Table S8A: Genes of operons, top 20 terms in Gene Ontology Biological Process, level 5.

Term	Count	%	PValue
GO:0006396~RNA processing	136	4.31%	1,30E-23
GO:0006464~protein modification process	361	11.43%	2,30E-17
GO:0015031~protein transport	186	5.89%	2,56E-16
GO:0016071~mRNA metabolic process	94	2.98%	3,34E-16
GO:0046907~intracellular transport	170	5.38%	1,64E-10
GO:0006497~protein amino acid lipidation	28	0.89%	1,84E-10
GO:0006457~protein folding	68	2.15%	2,35E-10
GO:0046467~membrane lipid biosynthetic process	36	1.14%	3,94E-09
GO:0008654~phospholipid biosynthetic process	31	0.98%	1,21E-08
GO:0000279~M phase	69	2.18%	4,84E-07
GO:0048193~Golgi vesicle transport	33	1.04%	7,25E-07
GO:0042157~lipoprotein metabolic process	29	0.92%	8,36E-07
GO:0006412~translation	116	3.67%	1,04E-06
GO:0006281~DNA repair	58	1.84%	2,26E-06
GO:0009060~aerobic respiration	16	0.51%	3,17E-06
GO:0006399~tRNA metabolic process	38	1.20%	3,42E-06
GO:0006084~acetyl-CoA metabolic process	18	0.57%	5,66E-06
GO:0051028~mRNA transport	23	0.73%	6,05E-06
GO:0000087~M phase of mitotic cell cycle	53	1.68%	6,74E-06
GO:0006886~intracellular protein transport	103	3.26%	8,27E-06

Table S8B: Genes out of operons, top 20 terms in Gene Ontology Biological Process, level 5.

Term	Count	%	PValue
GO:0016310~phosphorylation	349	6.16%	1,49E-25
GO:0006464~protein modification process	640	11.30%	3,52E-23
GO:0009887~organ morphogenesis	215	3.80%	7,30E-10
GO:0006812~cation transport	213	3.76%	1,69E-08
GO:0000279~M phase	114	2.01%	4,06E-08
GO:0006461~protein complex assembly	84	1.48%	5,86E-08
GO:0007417~central nervous system development	113	2.00%	1,06E-07
GO:0000902~cell morphogenesis	205	3.62%	2,15E-07
GO:0030001~metal ion transport	174	3.07%	2,23E-07
GO:0007420~brain development	91	1.61%	3,38E-07
GO:0043009~chordate embryonic development	98	1.73%	6,45E-07
GO:0006281~DNA repair	94	1.66%	6,83E-07
GO:0007067~mitosis	87	1.54%	7,79E-07
GO:0000087~M phase of mitotic cell cycle	87	1.54%	1,02E-06
GO:0048858~cell projection morphogenesis	123	2.17%	1,37E-06
GO:0030030~cell projection organization and biogenesis	123	2.17%	1,37E-06
GO:0032990~cell part morphogenesis	123	2.17%	1,37E-06
GO:0006643~membrane lipid metabolic process	81	1.43%	3,52E-06
GO:0006520~amino acid metabolic process	108	1.91%	3,57E-06
GO:0006897~endocytosis	81	1.43%	4,56E-06

Table S9: CAP3 consensus sequences of the most diversified clade of small TE-like elements (191 intact or nearly intact matches in both genome assemblies). Inverted terminal repeats are underlined.

>subcladeD1
CTTAAGGGCTCCCCCCCCCGTTCTGACTTTTGTCTAGAATTTTATATTTTTTGTCCGGATTCTGATTCTCCTGGTTCTAGTGACTATTTTTGGTCCTTACATCATGGTCCTAGAGACACGGGATCATAGAGAAATC
GAGACCTGAAATTGCACTAGAGTCGACACTTAGAAAAAATTTCGTGAAATATTGATTTTTAAAAGTTGTTGCATGGTGTCAAATAATAGCTTAAAATGATTGTCTCGTGATGCTGATTACGAAAAAATATATTAA
AATAATTTTTGATAGCCATTTAAGGGCGCCAGAGAGCAGCAAACACCCATAAATTAGGTACTTTTTTTCTAAATAAAGAATTTTTTTTTTCTGAAAGTCGGATCGACAAAATCTCTCTAGAGTTTTTAGATAAGG
GACAAAGGTAGACGTTGCACTAGTCAACACTAGCACTGTACAGGTGGTTACAGAGTTATTTCGGCCTGAATATGTGCAAAATTAGCAAACTAGCAATCTTTGGAGAGCTTTTTCGAGCATTTTTTTATACTTAAT
CGACAAAATCTCTCTAGAGTTTGTAGATAAAGGTCCAAACTTTAATATAAGCCATTGAGAGTATAAATCTGTAAGGCCAATCCTGAGTTATATCGCTTCAAAGTTCGAAAAAACGGGGGGGGGGATCCCTTAA
>subcladeD2
CTTAAGGGATCCCCCCCCCGTTCTGACTTTTGTCTAGAATTTTATATTTTTTGCCCGGATTCTGATTCTCCTGGTTCTAGTGACTATTTTTGGTCCTTACATCATGGTCCTAGAGACACGGGATCATAGAGAAATC
GAGACCTGAAATTGCACTAGAGTCGACACTTAGAAAAAATTCCGAGAAATATTGATTTTTAAAAGTTGTTGCATGGTGTCAAATAATAGCTTAAAATGATCGTCTCGTGATGCTGATTACGAAAAAACATAGTAA
AATAATTTTTGATAGCCATTTAAGGGCGCCAGAGAGCAGCAAACACCCATAAATTAGGTACTTTTTTTCTAAATAAAGAATTTTTTTTTTCTGAAAGTCGGATCGCCATAAAGCTCCTTCAGTTTTTAGATGAGGG
ACCAAGGAAGACGATGCAGTAGTCAACTGTAGCACTGTAAAGGTGGTTGCAAAGTTATTAAGGCCTGAATATAGCCTAAATTAGCAAACTGACAATCTTTGGAGAGCTTTTTCGAACATTTTTTTAAACTCGATC
GACAAAAAGCTCCTTCAGTTTTTAGATAAAGGTCTGAACTTTAATTTAAGACCTTAAGAATAAAAATCTGTTAGGCCAAAGCAAAGTTATTAAGCCCTAAAGTTCGAAAAAACGGGGGGGGGATCCCTTAAG
>subcladeD3
CTTAAGGGATCCCCCCCCCCGTTCTGATTTTTGTCATAAATTTAATATTTTTTGCCCGGATTCTGATTCTCCTTGTTCTACTGACCATTTTTGGTCCCTACATCATGGTCCTAAAGTCACGTGATTATTGAGAAAT
CGAGACCTGAACTTGCACTAGAGTCGACACTTAGAAAAAATTTCGAGAAATATTGATTTTTAAAAGTCGTTGCATGGGGTCAATTTACAGCTTAAAATAATCGTCTTGTGATGCTGATTCCGAAAAAACATGAGAA
AATAATTTTTGATTGCCATATAAGGGCGCCAGAGAGCAGCAAACACCCATAAATTAGGTACTTTTTTTCTAAATAAAGAAAATTTTTTTCTGAAAGTCGGACCGACTCAAGGCTAGTCCAATTTGTAGATAAGGG
ACTAAGGAAGACGTTTCACACGTCAGTTCTAGCACTGTAAAGGCGGTTGCAAAGTTATAAGGGCCTAAATATTGGCAAAATTGGCGAACATACAATCTTTGAAGAGCTTTTTCGAGCATTTTTTTATACTCGATC
GACATAAGGCTAGTCCAGTTTTGAGAAAAAGGTCTGAACTTCAAAAAAATCTTTTAAGAATAAAAATCCGAAAGGCCAATCAAAAGTTATAAGGCCCTAAAGTTCGAAAAAACGGGGGGGGGGATCCCTTAAG
>subcladeD4
CTTAAGGGATCCCCCCCCCGTTCTGATTTTTGTCATAAATTTTCTATTTTTTGCCCGGATTCTGATTCTCCTTGTTCTACTGACCATTTTTGGTCCCTACATCATGGTCCTAAAGTCACGTGATCATGGAGAAATC
GAGACCTGAACTTGCACTAGAGTCGACACTTAGAAAAAATTTCGAGAAATATTGATTTTTAAAAGTCGTTGCATGGGGTCAAATAATAGCTTAAAATGATCGTCTCGTGATGCTGATTCCGAAAAAACATAAGAG
AATAATTTTTGATAGCCATATAAGGGCGCCAGAGAGCAGCAAACACTCATAAATTAGGTACTTTTTTTCTAAATAAAGAAAAAATTTTTCTGAAAGTCGGATCGACAAAAAGCTACTTGAATTTTGAGATAAGGG
GCCAAGGAAGACGATGCACTAGTCAGTTCTAGCACTGTAAAGGTGGATTCAGAGTTATATCGGCCTGAATATGTGCTAAATTAGCAAACTTACAATCTTTGGGAAGCTTTTTCGAGCACTTTTATATACTCGATC
GCCAAAAAGCTACTTGAATTTGCAGATAAAGGTCCAAACTTTAATATAAGCCATTGAGAGTAAAAATCTGTAAGGCCGATCCTGAGTTATATCGCTTCAAAGTTCGAAAAAACGGGGGGGGGATCCCTTAAG
>subcladeD5
TTTAAGGGATCCCCCCCCCGTTCTGATCTTTGTCATAAATTTTATATTTTTTGCCCGGATTCTGATTCTCCTTGTTCTACTGACCATTTTTGGTCCCTACATCATGGTCCTAAAGTCACGTGATCATGGAGAAATC
GAGACCTGAACTTGCACTAGAGTCGACACTTAGAAAAAATTTCGAGAAATATTGATTTTTAAAAGTCGTTGCATGGGGTCAAATAATAGCTTAAAATGATCGTCCCGTGATGCTGATTCCGATAAAACATAAGCG
AATAATTTTTGATAGCCATATAAGGGCGCCAGAGAGCAGCAAACACTCATAAATTAGGTACTTTTTTTCTAAATAAAGAAAAATTTTTTCTGAAAGTCGGATCTACAAAAAGCTACTAGAGTTTTTAGATAAGGG
ACCAACAAAGACGATGAAGTAGTCAGCTCTAGCACTGTACAGGTGGATGCAAAGTTATTCAAGCCTGAATATGTGCTAAATTAGCAAATTGACAATCTTTGGAGAGCTTTTTCGAGCACTTTTTTATACTCGATC
GACAAAAAGCTACTAGAGTTTGTAGATAAAGGTCTAAACTTTAATATAAGCCATTAAGAATAAAAATCTGTAAAGCCAATGCAAAGTTATTCAGCCCTAAAGTTCGAAAAAACGGGGGGGGGATCCCTTAAG

Table S10: transmission rates of the fluorescence encoding marker in single pair lab crosses.

			daughters		sons	
mother	father	# offsprings	# wt	# fluo	# wt	# fluo
wt	fluo	106	2	59	45	0
wt	fluo	100	5	49	46	0
wt	fluo	98	2	45	51	0
wt	fluo	115	1	54	60	0
wt	fluo	71	0	33	38	0
wt	fluo	111	0	53	57	1
wt	fluo	141	0	76	65	0
wt	fluo	99	0	46	53	0
wt	fluo	120	0	53	67	0
wt	fluo	141	0	73	67	1
wt	fluo	143	0	81	62	0
wt	fluo	64	4	37	23	0
wt	fluo	127	1	51	75	0
wt	fluo	87	1	26	60	0
fluo	wt	71	21	26	15	9
fluo	wt	60	13	15	20	12
fluo	wt	74	31	12	21	10
fluo	wt	127	37	27	52	11
fluo	wt	46	12	7	24	3
fluo	wt	69	22	17	24	6
fluo	wt	47	15	8	13	11
fluo	wt	53	10	6	29	8
fluo	wt	55	15	7	18	15
fluo	wt	39	5	21	9	4
fluo	wt	53	20	11	13	9

Table S11: Colocalization of introns in genes. Two types of simulations were performed to quantify the inter-gene and intra-gene effects, as described in Methods.

		GA introns distributed randomly across the genes (inter-gene effect)		Constrained number of GA introns per gene (intra-gene effect)	
Pair of adjacent introns	Obs	Exp	Pval	Exp	Pval
nonGA nonGA	11550	11480.39	0.5169	11518.14	0.7641
GA nonGA/nonGA GA	1876	2013.34	2.2e-03	1920.31	0.3125
GA GA	156	88.27	<1e-09	143.56	0.2986

Table S12: Distribution of intron phases for GT-AG and GA-AG introns. The expected intron phase distributions were calculated as described in Methods.

	GT-AG			GA-AG		
Phase	nbr Obs	% Obs	% Exp	nbr Obs	% Obs	% Exp
0	19156	41.89%	35.16%	1591	34.62%	31.52%
1	15690	34.31%	41.05%	1878	40.87%	45.28%
2	10885	23.80%	23.79%	1126	24.50%	23.20%

Table S13: A : Comparison of positional biases in a set of 2524 orthologous genes between *Oikopleura*, *Ciona intestinalis*, *Ciona savignyi*, *Branchiostoma floridae*, *Strongylocentrus purpuratus* and *Homo sapiens*. B: Positional bias in the *Oikopleura* reference genes, and comparison between genes in operons and out of operons.

A	Od	Ci	Cs	Bf	Sp	Hs
Nbr of 5'biased genes	370	280	166	671	642	599
Nbr of 3'biased genes	646	162	94	343	402	473
Pval	<1e-09	1.5e-5	1.5e-3	<1e-09	1.4983e-7	6.5e-3

B	Od		
	all	in operon	not in operon
Nbr of genes 5'biased	3841	897	2944
Nbr of genes 3'biased	4533	1401	3132
Pval	8.9291e-8	<1e-09	0.0880

Table S14: Characteristics of new and old introns.

		New introns 3640	Old introns 776
splice sites	GT-AG	3334 (91.6%)	756 (97.4%)
	GA-AG	209 (5.7%)	11 (1.4%)
	GC-AG	91 (2.5%)	9 (1.2%)
	GG-AG	6 (0.2%)	0 (0%)
phases	0	1657 (45.5%)	447 (57.6%)
	1	1196 (32.9%)	185 (23.8%)
	2	787 (21.6%)	144 (18.6%)
intron sizes	<55	2431 (66.8%)	474 (61.1%)
	55 to 80	354 (9.7%)	53 (6.8%)
	81 to 500	755 (20.8%)	197 (25.4%)
	>500	100 (2.7%)	52 (6.7%)

Table S15: Distribution of intron phases for new and old introns. The expected intron phase distributions were calculated as described in Methods.

	new GT-AG : 3334 introns		new GA-AG : 209 introns		Old GT-AG : 756 introns	
phase	obs	exp	obs	exp	obs	exp
0	46.34%	35.50%	34.93%	32.17%	57.14%	39.15%
1	32.33%	41.87%	39.23%	44.24%	23.94%	39.34%
2	21.33%	22.63%	25.84%	23.59%	18.92%	21.51%

Table S16: Information content around splice sites for new and old introns. The information contents were calculated as described in Methods.

GT-AG introns		new	old
IC donors	all [-3 +5]	3.81	3.88
	Exonic [-3 –1]	2.04	1.70
	Intronic [+1 +5]	1.77	2.18
IC acceptors	all [-5 +3]	2.84	3.28
	Exonic [+1 +3]	2.73	2.95
	Intronic [-5 –1]	0.11	0.33

Table S17: Colocalisation of intron gains in genes. Two types of simulations were performed to quantify the inter-gene and intra-gene effects, as described in Methods.

		new introns distributed randomly across the genes (inter-gene effect)		Constrained number of new introns per gene (intra-gene effect)	
Pair of adjacent introns	Obs	Exp	Pval	Exp	Pval
new - new	2061	1965	3.03 e-02	2639	<10e-09
new - not new	553	632	1.6 e-02	243	<10e-09
not new - new	478	632	<10e-09	243	<10e-09
not new - notnew	341	204	<10e-09	308	5.99e-02

Table S18: Description of introns that are candidates for having originated from transposon insertion (continues in next page).

intron id	intron length	splice sites	Annotated transposable element	Direct repeat length	Direct repeat sequence	Nbr of alleles in the assembly	Polymorphism from genotyping
GSOIDI00000579003	1626	GT-AG	MITE	8	GTGATGAG	other allele identical (with the intron)	polymorphic
GSOIDI00001051001	573	GG-AG	DIRS	8	AGAATCAG	other allele identical (with the intron)	polymorphic
GSOIDI00004756001	97	GA-AG	-	7	TTAGGAT	other allele identical (with the intron)	-
GSOIDI00010634002	185	GA-AG	-	9	GCAAAGGAG	other allele identical (with the intron)	-
GSOIDI00011755001	144	GG-AG	-	9	GGGTAGTAC	other allele identical (with the intron)	-
GSOIDI00013578002	923	GT-AG	MITE	9	GTTGGAATC	other allele identical (with the intron)	-
GSOIDI00015757006	383	GA-AG	-	10	TCAAAAAGGA	other allele identical (with the intron)	-
GSOIDI00016636001	632	GT-AG	MITE	5	CTAAG	other allele identical (with the intron)	-
GSOIDI00012101003	546	GT-AG	MITE	5	CTCAG	other allele identical (with the intron)	monomorphic
GSOIDI00000193001	714	GG-AG	MITE	7	ACCCTAG	no other allele	-

GSOIDI00001055003	730	GC-AG	-	9	AGGCTCGTC	no other allele	-
GSOIDI00003825003	270	GC-AG	mariner	8	CTTTAGGC	no other allele	-
GSOIDI00005013007	786	GT-AG	-	11	GGTTACAAGCT	no other allele	-
GSOIDI00006682004	239	GC-AG	-	8	GATACGAG	no other allele	-
GSOIDI00012567008	739	GT-AG	-	9	GTCAAGGCC	no other allele	polymorphic
GSOIDI00013643007	147	GA-AG	PLE	7	AAGGACC	no other allele	-
GSOIDI00013880002	288	GC-AG	-	8	AATCAGGC	no other allele	-
GSOIDI00015471005	1046	GC-AG	-	7	AGGCATC	no other allele	-
GSOIDI00017365004	248	GC-AG	-	9	GATACGAGG	no other allele	-
GSOIDI00008507002	731	GC-AG	-	7	GCAGCAC	other allele without the intron	-
GSOIDI00007397001	690	GG-AG	LTR	7	TTAAGGG	other allele without the intron	polymorphic
GSOIDI00008884002	268	GT-AG	LINE	7	TTAGGTG	other allele without the intron	-
GSOIDI00002583001	3143	GG-AG	MITE + mav2 + DIRS	10	AGGGGCCAAG	other allele without the intron	-
GSOIDI00013129003	774	GG-AG	LTR	5	TTAAG	other allele without the intron	polymorphic
GSOIDI00000560002	751	GA-AG	-	8	GATCCATG	other allele without the intron	-
GSOIDI00003720006	580	GG-AG	-	8	TTTACGAG	other allele without the intron	-
GSOIDI00015827001	554	GC-AG	mav2	8	GCAGTACT	other allele without the intron	-
GSOIDI00008225007	665	GG-AG	MITE	5	TTAAG	other allele without the intron	-
GSOIDI00003681001	1383	GT-AG	-	7	GGTATGA	other allele without the intron	-
GSOIDI00017920002	2822	GC-AG	mariner + MITE	8	CCAAGGCC	other allele without the intron	-
GSOIDI00011525005	2610	GT-AG	MITE	5	CTGAG	other allele without the intron	-
GSOIDI00018267002	588	GT-AG	-	8	GTCGGAAG	other allele without the intron	-

Table S19: Summary of genotyping and assay for splicing, for the four pairs of homologous introns (reverse splicing hypothesis)

Intron pair	Intron location	Frequency	Splicing (5 individuals tested)
Pair1	sc_3: 299199.. 299244 GSOIDI00008841005	1	5/5
	sc_3: 298806.. 298852 GSOIDI00008841006	1	5/5
Pair 2	sc_17:608316..608360 GSOIDI00005057008	1	3/5
	sc_17:607949..607994 GSOIDI000050570010	1	3/5
Pair 3	sc_52:66075..66126 GSOIDI00014299002	1	5/5
	sc_52:66603..66654 GSOIDI00014298001	1	5/5
Pair 4	sc_926:3765..3813 GSOIDI00016500002	1	5/5
	sc_926:4412..4460 GSOIDI00016500006	1	5/5

Table S20: Summary of genotyping for candidate introns. $F_{(+/+)}$: frequency of homozygotes with the intron, $F_{(-/-)}$: frequency of homozygotes without the intron, $F_{(+)}$: frequency of hemizygotes with the intron, $F_{(-)}$: frequency of hemzygotes without the intron, $F_{(+/-)}$: frequency of heterozygotes.

	Polymorphic ?	Frequency	n samples with this genotype		Remarks
			♀	♂	
sc_1:1900390..1904365	No	$F_{(+/-)} = 0,25$	0	9	
		$F_{(+/+)} = 0,08$	3	0	
		$F_{(-/-)} = 0,67$	0	24	
sc_50:194113..197306	Yes	$F_{(+/-)} = 0,06$	0	1	
		$F_{(-/-)} = 0,94$	3	15	
sc_5:290172..290336	No	$F_{(+/+)} = 1$	12	11	
sc_42:311401..312139	No	$F_{(+)} = 0,21$	0	5	Population A
		$F_{(-)} = 0,28$	0	6	
		$F_{(+/+)} = 0,04$	1	0	
		$F_{(+/-)} = 0,52$	11	0	
		$F_{(+/+)} = 0,37$	9		Population B (collected before maturation)
		$F_{(+/-)} = 0,33$	8		
		$F_{(-/-)} = 0,29$	7		
		$F_{(+/+)} = 1$	5	5	Population F143
		$F_{(-/-)} = 1$	8	2	Population F11
sc_70:129142..129424	No	$F_{(+/+)} = 0,30$	3	0	Population F0
		$F_{(+/-)} = 0,20$	2	0	
		$F_{(-/-)} = 0,50$	5	0	
		$F_{(-/-)} = 0,40$	4	0	Population F11
		$F_{(+/-)} = 0,60$	4	2	
		$F_{(-/-)} = 1$	5	5	Population F143
sc_10:85290..87289	No	$F_{(-/-)} = 0,42$	6	0	
		$F_{(+/-)} = 0,58$	0	8	
sc_267:32792..35973	Yes	$F_{(-/-)} = 0,42$	6	0	
		$F_{(+/-)} = 0,58$	0	8	
sc_8:1322173..1322415	No	$F_{(+/+)} = 0,5$	0	10	
sc_43:198665..198939	No	$F_{(+/+)} = 1$	12	11	
sc_48:98203..98485	No	$F_{(+/+)} = 1$	12	11	
sc_45:185367..185626	No	$F_{(+/+)} = 1$	12	11	
sc_25:72650..72914	No	$F_{(+/+)} = 1$	12	11	
sc_431:15391..15936	No	$F_{(+/+)} = 1$	23	14	

Table S21: Lack of putative sensors and adaptors in the *Oikopleura* genome, and multiplicity of PLA2 (which may include candidate antimicrobial members). A comparison of gene families encoding potential pathogen receptors, signaling adaptors and PLA2 in representative species with sequenced genomes to *O dioica* is provided. *O. dioica* lacks completely several key classes of sensors or adaptors. A "–" was indicated when no occurrence was found, while "?" means that no comprehensive survey was found. The data are from references (*S156, S157, S159, S167, S168, S169, S170).*

Protein models similar to :	*Drosophila melanogaster*[1]	*Strongylocentrotus purpuratus*	*Oikopleura dioica*	*Ciona Intestinalis*	*Branchiostoma floridae*	*Lampetra fluviatilis*	*Homo sapiens*
sensors							
TLR	9	222	**1**	3	48	≈21	10 +1ps
NLR	0	203	**0**	20	92	≈140-220	20
SRCR	14	218	**1**	81	270	≈287 domains	81
PGRP	15	5	**4**	6	>20	?	6
RIG-I-like helicases	0	12	**0**	-	7	?	3
C-type lectins,CTLs	32	104	**31**	120	1215	?	81
IgSF-ITIM	>3	?	**5**	>6	>5	>3	>50
adaptors							
MyD88-like (DEATH-TIR)	1	4	**0**	1	4	-	1
SARM1-like, TIRAP-like, TICAM2-like	1	15	**0**	>2	12	-	3
potential effector							
PLA2	8	65	**128**	7	>7	?	11

Table S22: Protein models (or scaffolds segments) matching lineage-specific duplicates of *Oikopleura* developemental genes (continues in next pages).

CHAPTER	GENE FAMILY	**GENE GROUP**	SOURCE
I	bHLHA	**Ash-a**	GSOIDP00007675001
			GSOIDP00002088001
			GSOIDP00029330001
			GSOIDP00000706001
			GSOIDP00000678001
			GSOIDP00015801001
I	bHLHA	**MyoD**	sca41-192kb
			GSOIDP00004266001
			sca9-341kb
I	bHLHB	**Figα**	sca18-shotgun
			GSOIDP0000941200
			GSOIDP00009414001
I	bHLHE	**Hes**	GSOIDP00014217001
			GSOIDP00008858001
			GSOIDP00002936001
II	CUT	**Onecut**	GSOIDP00007033001
			GSOIDP00012637001
			GSOIDP00016055001
II	SIX	**Six12**	GSOIDP00010834001
			GSOIDP00009116001
II	SIX	**Six36**	GSOIDP00001881001
			GSOIDP00017724001
II	TALE	**Irx (all)**	GSOIDP00001122001
			GSOIDP00010984001
			GSOIDP00017177001
			GSOIDP00013453001
			GSOIDP00012126001
II	TALE	**Meis**	GSOIDP00014391001
			GSOIDP00004388001
II	LIM	**Lmx**	GSOIDP00009252001
			GSOIDP00006999001
II	POU	**Pou3**	GSOIDP00005125001
			GSOIDP00005137001
II	PAX	**Pax37**	GSOIDP00011822001
			GSOIDP00011199001
			GSOIDP00008979001
			GSOIDP00011887001
II	NK	**CG13424**	GSOIDP00010929001
			GSOIDP00010928001
			GSOIDP00018058001
II	NK	**Not**	sca95-41kb
			GSOIDP00016364001
			GSOIDP00003477001
			sca5-shotgun
II	PRD-like	**Otx**	GSOIDP00016424001

			GSOIDP00013015001
			GSOIDP00013014001
II	PRD-like	**Mix**	GSOIDP00012480001
			GSOIDP00012479001
II	HOX-like	**Cdx**	GSOIDP00017556001
			GSOIDP00017339001
			GSOIDP00004905001
III	FOX	**FoxA**	GSOIDP00004124001
			GSOIDP00003756001
			GSOIDP00005965001
III	FOX	**FoxI**	GSOIDP00010324001
			GSOIDP00010950001
			GSOIDP00013037001
			GSOIDP00001629001
			GSOIDP00005691001
III	FOX	**FoxH**	GSOIDP00000155001
			GSOIDP00013454001
			GSOIDP00009515001
			GSOIDP00015435001
III	FOX	**FoxO**	GSOIDP00015791001
			sca21-259kb
			GSOIDP00000952001
			GSOIDP00003334001
			GSOIDP00003475001
III	ETS	**Pointed**	GSOIDP00006792001
			GSOIDP00003683001
III	ETS	**Elk**	GSOIDP00002658001
			GSOIDP00010228001
			GSOIDP00013017001
III	ETS	**Erm/Er81**	GSOIDP00010682001
			GSOIDP00015310001
III	NR	**VDR**	GSOIDP00005963001
			GSOIDP00005043001
			GSOIDP00001685001
			GSOIDP00000593001
			GSOIDP00003479001
			GSOIDP00016406001
III	NR	**ECR/LXR/FXR (all)**	GSOIDP00005829001
			GSOIDP00010629001
			GSOIDP00003168001
			GSOIDP00017885001
			GSOIDP00002087001
			GSOIDP00002349001
			GSOIDP00017450001
			GSOIDP00005828001
			GSOIDP00007515001
			GSOIDP00015935001
III	NR	**ROR**	GSOIDP00004778001
			GSOIDP00005201001

III	NR	**RXR**	GSOIDP00003392001
			GSOIDP00013746001
III	NR	**FTZF**	GSOIDP00018059001
			GSOIDP00006200001
III	NR	**GCNF**	GSOIDP00007953001
			GSOIDP00014229001
			GSOIDP00005196001
III	NR	**NR4A**	GSOIDP00000542001
			GSOIDP00005949001
III	NFAT	**NFAT**	GSOIDP00017918001
			GSOIDP00012710001
IV	SOX	**SoxD**	GSOIDP00015160001
			GSOIDP00014095001
IV	bZIP	**Maf**	GSOIDP00013208001
			sca36-14kb
			GSOIDP00008843001
IV	bZIP	**CREB/ATFII**	GSOIDP00012988001
			GSOIDP00001894001
IV	bZIP	**CREB/ATFIII**	GSOIDP00007699001
			GSOIDP00004607001
			GSOIDP00012510001
			GSOIDP00012506001
IV	bZIP	**CREB/ATFIV**	GSOIDP00005649001
			GSOIDP00004169001
IV	bZIP	**XBP**	GSOIDP00000346001
			GSOIDP00002235001
			GSOIDP00001521001
IV	bZIP	**Jun**	GSOIDP00008627001
			GSOIDP00008219001
IV	bZIP	**Tel/HLF**	GSOIDP00009784001
			GSOIDP00022689001
			sca8-1200kb
			GSOIDP00008325001
			GSOIDP00001562001
IV	bZIP	**C/EBPII**	GSOIDP00014222001
			sca466-7kb
			GSOIDP00009345001
IV	bZIP	**ATF2/7**	GSOIDP00014568001
			sca10-386kb
IV	bZIP	**Fos-like**	GSOIDP00015484001
			GSOIDP00001415001
			GSOIDP00004355001
IV	ZnFGATA	**GATA1/2/3**	GSOIDP00008593001
			GSOIDP00000004001
			GSOIDP00011092001
			GSOIDP00007921001
V	TYRK	**FgfR**	GSOIDP00008404001
			GSOIDP00009373001

			GSOIDP00009004001
			GSOIDP00010261001
V	TYRK	**EphR**	GSOIDP00011750001
			GSOIDP00017622001
			GSOIDP00001449001
			GSOIDP00012999001
			GSOIDP00000495001
			sca4-38kb
V	EPHRIN	**Ephrin (all)**	GSOIDP00010865001
			GSOIDP00006028001
			GSOIDP00010675001
			GSOIDP00001862001
			GSOIDP00015146001
V	RAS	**Rap1 & Rap2**	GSOIDP00004405001
			GSOIDP00001067001
			GSOIDP00010476001
			GSOIDP00028414001
V	RAS	**orphan pair**	GSOIDP00015691001
			GSOIDP00010659001
V	MAPK	**p38**	GSOIDP00011471001/11472001
			GSOIDP00011470001
V	NUMB	**Numb**	GSOIDP00015583001
			GSOIDP00008916001
VI	WNT	**Wnt11-related**	GSOIDP00013757001
			GSOIDP00008816001
			GSOIDP00013856001
			GSOIDP00009921001
VI	FRIZZLED	**Frz3/6**	GSOIDP00008718001
			GSOIDP00006033001
			GSOIDP00000671001
			GSOIDP00007729001
VI	GSK3	**Gsk3**	GSOIDP00003070001
			GSOIDP00014639001
VI	AXIN	**Axin**	GSOIDP00011595001
			GSOIDP00000684001
VI	β-CATENIN	**β-CATENIN**	GSOIDP00004053001
			GSOIDP00002611001
			GSOIDP00011813001
			(GSOIDP00000374001)
			(GSOIDP00000799001)
VI	TGFβR	**ALK1/2**	GSOIDP00010407001
			GSOIDP00005813001
VI	TGFβR	**ALK3/6**	GSOIDP00008331001
			GSOIDP00006382001
			GSOIDP00010442001
			GSOIDP00009375001
VI	TGFβR	**TGFβR II**	GSOIDP00010252001
			GSOIDP00007397001
VI	SMAD	**Smad1/5/8/9**	GSOIDP00008165001

			GSOIDP00008257001
VI	SMAD	**Smad2/3**	GSOIDP00009561001
			GSOIDP00000187001
			GSOIDP00007016001
			GSOIDP00007017001
			GSOIDP00012596001
VI	SMAD	**orphan group**	GSOIDP00014385001
			GSOIDP00006515001
			GSOIDP00012526001
			GSOIDP00008229001
VI	STAT	**STAT5/6**	GSOIDP00004837001/32525001
			GSOIDP00017671001
			GSOIDP00000174001
VII	Par5	**14-3-3ε**	GSOIDP00019558001
			GSOIDP00004848001
VII	Rho/Rac/cdc42	**RhoABC**	GSOIDP00003459001
			GSOIDP00009377001
			GSOIDP00010718001
			GSOIDP00003787001
VII	CAMK	**CAMKII**	GSOIDP00004057001
			GSOIDP00001768001
VII	CAMK	**CAMK-like**	GSOIDP00024548001
			GSOIDP00015133001
VII	PLC	**PLC orphan**	GSOIDP00032006001
			GSOIDP00011888001
VII	WASP/WAVE/SCAR	**WASP**	GSOIDP00011846001
			GSOIDP00010515001
VII	WASP/WAVE/SCAR	**VASP/ENA**	GSOIDP00015088001
			GSOIDP00002881001
VII	ACTIN-like	**ARP1**	GSOIDP00002720001
			GSOIDP00017847001
VII	ACTIN-like	**orphan**	GSOIDP00003432001
			GSOIDP00014724001
VII	ADF-COFILIN	**ADF-COFILIN**	GSOIDP00009430001
			GSOIDP00003179001
			GSOIDP00013139001
VII	PAK	**PAK1/2/3**	GSOIDP00003204001
			GSOIDP00017705001
VIII	RPS6 KINASE	**p70/RPS6KB**	GSOIDP00000569001
			GSOIDP00000567001
VIII	RPS6 KINASE	**p90/RPS6KA123**	GSOIDP00015627001
			GSOIDP00014860001
VIII	PTEN	**PTEN**	GSOIDP00012578001
			GSOIDP00011683001
VIII	4EBP	**4EBP**	GSOIDP00004927001
			GSOIDP00004101001
			GSOIDP00003281001
VIII	CYCLINS	**CYCLIN B**	GSOIDP00012241001

			GSOIDP00003037001
			GSOIDP00003907001
			GSOIDP00004728001
			sca267-shotgun
			GSOIDP00008944001
			GSOIDP00015736001
VIII	CYCLINS	**CYCLIN E**	GSOIDP00011643001
			GSOIDP00007812001
VIII	CDK	**KPT/PCTK**	GSOIDP00006215001
			GSOIDP00000028001
VIII	CDK	**CDK1/CDC2**	GSOIDP00015437001
			GSOIDP00008033001
			GSOIDP00003049001
			GSOIDP00011734001
			GSOIDP00016654001
			GSOIDP00017779001
VIII	CDK	**CRK7/CDL5**	GSOIDP00004437001
			GSOIDP00012412001
VIII	CDKI	**CDKI**	GSOIDP00010338001
			GSOIDP00002106001
VIII	CDC25	**CDC25**	GSOIDP00010933001
			GSOIDP00015421001
			GSOIDP00017717001
			GSOIDP00015083001
			GSOIDP00007837001
VIII	WEE1/MYT1	**WEE1**	GSOIDP00004122001
			GSOIDP00012772001

Table S23: The decay of the dN/dS rate ratio from relaxed selection on recent duplicates towards rates consistent with orthologs is modeled as described in Hughes and Liberles (*S172*). The parameterizations of the instantaneous and asymptotic dN/dS ratios as well as the decay parameter between the two are presented together with their variance. The mammalian values are taken from Hughes and Liberles (*S172*) and the *Oikopleura* values were calculated using a Bayesian implementation with a flat prior to enable direct comparison with the quasi-likelihood estimates given in Hughes and Liberles (*S172*).

genome analysis	θ1	θ2	θ3
Oikopleura	0.18 (0.01)	0.30 (0.02)	3.51 (1.0)
Canis	0.04 (0.01)	0.54 (0.01)	2.26 (0.25)
Homo	0.03 (0.01)	0.70 (0.01)	2.40 (0.17)
Mus	0.07 (0.01)	0.56 (0.01)	2.51 (0.21)
Rattus	0.06 (0.01)	0.57 (0.01)	2.24 (0.18)

Table S24: Expression of lineage-specific amplified homeobox gene groups. Signals in the middle region of trunk epithelium are highly frequent for amplified homeobox gene groups and rare in non-amplified groups.

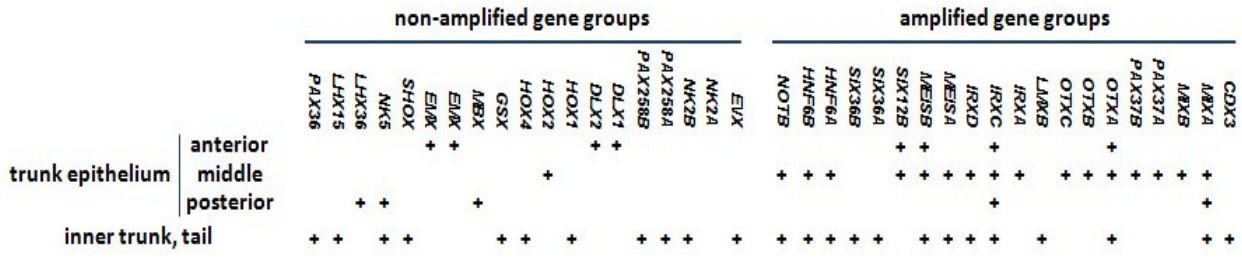


Table S25: Number of orthologous genes included in Figure 3C (main text)

		Dot matrices		Distance measures	
	Total genes	Metazoan Species	Human (on CLGs)	Metazoans	Human
Amphioxus	28666	3646 (25 scaffolds)	2544	7507	10485
Ciona	14180	5354 (chromosomes)	6796	8252	9610
Nematode	20158	6267 (chromosomes)	7141	6022	7700
Sea anemone	27273	1786 (25 scaffolds)	3109	7357	11885
Oikopleura	18020	4491 (chromosomes)	6597	5182	7904