Published online 26 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D845–D853***
doi:10.1093/nar/gks1159

OikoBase: a genomics and developmental transcriptomics resource for the urochordate *Oikopleura dioica*

Gemma Danks[1,2], Coen Campsteijn[2,3,4], Mrutyunjaya Parida[5], Stephen Butcher[5], Harsha Doddapaneni[6], Bolei Fu[6], Raul Petrin[7], Raghu Metpally[5], Boris Lenhard[1,2,8], Patrick Wincker[9], Daniel Chourrout[2], Eric M. Thompson[2,10,*] and J. Robert Manak[5,6,7,*]

[1]Computational Biology Unit, [2]Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen, N-5008, [3]Centre for Cancer Biomedicine, Faculty of Medicine, University of Oslo, [4]Department of Biochemistry, Institute for Cancer Research, Norwegian Radium Hospital, Oslo University Hospital, N-0310, Oslo, Norway, [5]Department of Biology, University of Iowa, [6]Department of Biology, Carver Center for Genomics, [7]Department of Pediatrics, Carver College of Medicine, University of Iowa, Iowa City, IA 52242 USA, [8]Department of Molecular Sciences, Imperial College London and MRC Clinical Sciences Centre, London W12 0NN, UK, [9]Commissariat à l'Énergie Atomique (CEA), Genoscope, 91057 Evry, France and [10]Department of Biology, University of Bergen, Bergen, N-5020, Norway

Received July 31, 2012; Revised October 12, 2012; Accepted October 28, 2012

ABSTRACT

We report the development of OikoBase (http://oikoarrays.biology.uiowa.edu/Oiko/), a tiling array-based genome browser resource for *Oikopleura dioica*, a metazoan belonging to the urochordates, the closest extant group to vertebrates. OikoBase facilitates retrieval and mining of a variety of useful genomics information. First, it includes a genome browser which interrogates 1260 genomic sequence scaffolds and features gene, transcript and CDS annotation tracks. Second, we annotated gene models with gene ontology (GO) terms and InterPro domains which are directly accessible in the browser with links to their entries in the GO (http://www.geneontology.org/) and InterPro (http://www.ebi.ac.uk/interpro/) databases, and we provide transcript and peptide links for sequence downloads. Third, we introduce the transcriptomics of a comprehensive set of developmental stages of *O. dioica* at high resolution and provide downloadable gene expression data for all developmental stages. Fourth, we incorporate a BLAST tool to identify homologs of genes and proteins. Finally, we include a tutorial that describes how to use OikoBase as well as a link to detailed methods, explaining the data generation and analysis pipeline. OikoBase will provide a valuable resource for research in chordate development, genome evolution and plasticity and the molecular ecology of this important marine planktonic organism.

INTRODUCTION

The urochordate (or tunicate) appendicularian, *Oikopleura dioica,* is a coastal marine planktonic chordate with a pan-global distribution. As an abundant component of mesozooplankton communities (1), appendicularians are noted for their ability to rapidly expand population size in response to algal blooms. In a defining feature, urochordates are partially or completely enclosed in extracellular, cellulose-based 'tunics' or 'houses' involved in filter-feeding, making them the only animals known to synthesize cellulose (2). Appendicularians frequently resynthesize their house during their short life cycle, and discarded houses (marine snow) play significant, and sometimes dominant, roles in oceanic vertical carbon flux (3,4), thus impacting global carbon cycles. The urochordates are the closest extant relatives to the vertebrates (5) and have a simplified chordate body plan with a notochord, dorsal neural tube, gill slits and endostyle. *Oikopleura dioica* remains transparent throughout its short (less than 1 week) chordate life cycle, exhibits high fecundity [more than 300 eggs per female (6)] and can be cultured in the laboratory for hundreds of generations (7). It is the only known dioecious urochordate species. Compared with ascidians, *O. dioica* undergoes a morphologically simple

*To whom correspondence should be addressed. Tel: +1 319 335 0180; Fax: +1 319 335 1069; Email: john-manak@uiowa.edu
Correspondence may also be addressed to Eric M. Thompson. Tel: +47 55584346; Fax: +47 55584307; Email: eric.thompson@sars.uib.no

metamorphosis, in which the tail is not resorbed but only shifts from a posterior directed orientation in the tadpole to become orthogonal to the trunk at metamorphosis. Development is very rapid, with metamorphosis completed after 12–14 h at 15°C. Throughout the life cycle, the tail retains the notochord as its axial structure in order to function in pumping water through the filters of the house for feeding. The animal is generally found from 0 to 200 m depth and tolerates a wide range of temperatures and salinities (8). The 70-Mb compact, sequenced genome containing more than 18 000 predicted genes ranks among the smallest known metazoan genomes (9,10), with both gene regulatory and intronic regions highly reduced in size. This rapidly evolving lineage exhibits profound alteration of deeply conserved features of metazoan genome architecture, thus offering interesting perspectives in the study of genome plasticity (10) and developmental gene regulation.

Here we present a tiling array-based developmental transcriptome genome browser resource, OikoBase (http://oikoarrays.biology.uiowa.edu/Oiko/; based on the popular GBrowse software), at ultra-high resolution in terms of both developmental chronology and the underlying genome sequence features. The unbiased transcriptome datasets described here comprise 12 key developmental stages encompassing the entire life span of *O. dioica*, in addition to testis, ovary and somatic body-specific sample sets. Finally, three transcriptome datasets generated from animals under growth/developmental arrest are included. The compact genome of *O. dioica*, as well as its short lifespan, permits a level of comprehensiveness and resolution that rivals that of well-established model organisms. A total of 62 Mb of the non-repeat genome (93% of the total genome) was tiled on a single microarray (2.1 million features, with a 2-fold genome coverage) using isothermal oligonucleotide probes with a median size of 53 and 24 bp overlaps between adjacent probes, permitting reliable detection of short introns (47-bp peak size in the *Oikopleura* genome). This platform and its ancillary features, which will be used to explore both the transcriptome as well as global chromatin structure, will serve as a rich resource in understanding the developmental and evolutionary biology of this ecologically important organism in the context of its transcriptional response to changing experimental and environmental inputs.

BIOLOGICAL SOURCE MATERIALS AND DATA GENERATION

Animal culture and collection

Oikopleura dioica were maintained in culture at 15°C (7). Oocytes were collected from mature females, and developmental stages up to and including metamorphosis were generated by *in vitro* fertilization (11), with embryos left to develop in artificial sea water (Red Sea, final salinity 30.4–30.5 g/l) at room temperature to the desired stage. Day1–Day5 animals were placed in artificial seawater, chased from their houses, left for 30 min to empty their gut, anesthetized in cold ethyl 3-aminobenzoate methanesulfonate salt (MS-222, 0.125 mg/ml; Sigma), and then collected. Ovary and testes extracts were collected as previously described (12). To obtain the trunk samples, Day5 animals were washed in cold phosphate-buffered saline and anesthetized in cold MS-222, transferred to cold Buffer A [10 mM Tris (pH 7.5), 360 mM sucrose, 75 mM NaCl, 10 mM ethylenediaminetetraacetic acid, 10 mM ethyleneglycol bis[aminoethylether]-tetraacetic acid, 3 mM dithiothreitol, 1 mM phenylmethylsulfonyl fluoride and 1:500 RNase OUT (Invitrogen)] and gonads were punctured and removed. Remaining trunks were washed once in cold Buffer A, pooled and processed for RNA isolation. For collection of the Day2–Day4 dense samples, the normal dilution of animal spawn at Day1 was omitted (7), yielding a culture with ~5-fold higher animal density. In partial compensation for elevated animal numbers while suppressing algal overgrowth, the food concentration was doubled at each developmental stage. The elevated animal density and feeding regime were maintained throughout the experiment, with animal collection performed at time points as indicated.

Gene annotation

A very detailed reporting of gene annotation protocols is provided in the supplementary information of (10). A summary version is presented here. Semi-HMM-based Nucleic Acid Parser (SNAP) *ab initio* gene prediction (13) was used to create a clean set of *O. dioica* genes. This set was used to train gene prediction algorithms and optimize their parameters. SNAP was launched using the *Caenorhabditis elegans* configuration file, and only models with all introns confirmed by at least one *O. dioica* cDNA were retained. Models that contained at least one exon that overlapped a cDNA intron were rejected. Three hundred models were randomly selected to create the *O. dioica* clean training set.

Exofish (14) comparisons were done with Biofacet (www.gene-it.com). When ecores (Evolutionarily COnserved REgions) were contiguous in the two genomes, they were included in the same ecotig (contig of ecores). Exofish comparisons were performed between *O. dioica* and four organisms: *Tetraodon nigroviridis*, *Strongulocentrus purpuratus*, *Ciona savignyi* and *Ciona intestinalis*. High scoring segment pairs were filtered by length and percent identity.

The Uniprot (15) database was then used to detect genes conserved between *O. dioica* and other species. Since GeneWise (16) is computationally intensive, sequences in the Uniprot database were first aligned with the *O. dioica* genome assembly using BLAST-Like Alignment Tool (BLAT) (17). Each significant match was chosen for a GeneWise alignment. The default GeneWise gene parameter file was modified to account for unusual splice sites of *O. dioica* genes. Geneid (18) and SNAP *ab initio* gene prediction software were then trained on the 300 genes from the training set.

Expressed Sequence Tags (ESTs) generated from three full-length-enriched cDNA libraries were prepared from a cultured outbred population [large pools of unfertilized eggs, embryos at mixed stages from 1 to 3 h post-

fertilization (pf), tadpoles 6–10 h pf and Day 4 animals] and 180 000 cDNA clones were end sequenced. After assembly of 5′- and 3′-sequences, a total of 177 439 sequences were aligned to the *O. dioica* genome using the following pipeline: after masking of polyA tails and spliced leaders, the sequences were aligned with BLAT on the genome assembly and matches with scores within 99% of the best score were extended by 5 kb on each end and realigned with the cDNA clones using Exonerate (19), allowing for non-canonical splice sites, with the following parameters; –model est2genome –minintron 25 –maxintron 15 000 –gapextend -8 –dnahspdropoff 12 –intronpenalty -23.

An National Center for Biotechnology Information (NCBI) (20) collection of ~1 500 000 public tunicate ESTs was then aligned with the *O. dioica* genome assembly using BLAT alignments refined with Est2Genome (21). Significant matches were chosen for alignment with Est2Genome. BLAT alignments used default parameters between translated genomic and translated ESTs.

The above resources were combined to automatically build *O. dioica* gene models using GAZE (22). Individual predictions from each program (Geneid, SNAP, Exofish, GeneWise, Est2genome and Exonerate) were broken into segments (coding, intron and intergenic) and signals (start codon, stop codon, splice acceptor, splice donor, transcript start and 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 were created from the span of each mRNA, with a negative score (forcing GAZE not to split genes). Predicted repeats were used as intron and intergenic segments, to avoid prediction of protein-coding genes in such regions.

The whole genome was scanned to find signals (splice sites, start and stop codons). In order to annotate genes containing non-canonical splice sites, all G* (GT, GA, GC and GG) donor sites were authorized. Segments predicting exon boundaries were used by GAZE only if GAZE chose the same boundaries. Each segment or signal from a given program was accorded a value reflecting our confidence in the data, and these values were used as scores for the arcs of GAZE automaton. All signals were assigned 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 reliability and accuracy in predicting gene models. This weight acted 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. The final proteome of 18 020 gene models was obtained by adding 907 gene models from an allelic assembly that were not present in the reference assembly.

RNA isolation and cDNA synthesis

RNA isolation was performed using RNeasy (Qiagen) and RNA quality was assessed using a Bioanalyzer (Agilent). Purified total RNA was briefly treated (15 min at room temperature) with 0.5 units of DNase I (AMP grade; Invitrogen) per microgram of RNA to remove residual DNA, followed by ethanol precipitation and purification of RNA. Double-stranded (ds) cDNA was synthesized using SuperScript^{TM} Double-Stranded cDNA Synthesis Kit (Invitrogen) according to Roche NimbleGen protocols, followed by phenol:chloroform:isoamylalcohol (25:24:1) as well as chloroform:isoamylalcohol (24:1) back extraction and ds cDNA was recovered by ethanol precipitation. Typically, 4–10 μg of input total RNA yielded 3–9 μg of ds cDNA.

Tiling array hybridization and processing

A tiled genomic microarray was designed to interrogate the entire non-repeat genome of *O. dioica* based on genome sequence [GENBANK/European Molecular Biology Laboratory accession numbers are CABV01000001-CABV01005917, CABW01000001-CABW01006678, FN653015-FN654274, FN654275-FN658470, FP700189-FP710243, FP710258-FP791398 and FP791400-FP884219; (10)] assembled in collaboration between the Sars International Centre for Marine Molecular Biology (www.sars.no) and Genoscope (www.genoscope.cns.fr/secure-nda/Oikopleura). For each developmental stage (or stressor), cDNA was labeled and amplified using the Roche NimbleGen gene expression protocol (http://www.nimblegen.com/products/lit/05434505001_NG_Expression_UGuide_v6p0.pdf) with the following modifications: 15 μg of the labeled cDNA was hybridized to each array, and hybridization was performed on three separate arrays (three technical replicates; four for the oocyte sample). The arrays were washed and processed as per the manufacturer's recommendations, and the arrays were scanned on Molecular Devices GenePix Axon scanners 4000B or 4400 A at 5 or 2.5 μm resolution, respectively. The scanned array data were converted into pair and GFF files. These were processed to generate normalized, background subtracted probe intensities that comprise the transcriptomics tracks in the browser (together with $\log_2$-transformed data tracks). Raw and processed probe intensities for all samples have been deposited in NCBI's Gene Expression Omnibus (23) and are accessible through GEO Series accession number GSE39568.

We define a transcriptionally active region (TAR) as any stretch of consecutive positive probes in a particular sample. To compare TARs between samples, we constructed a set of 'superTARs' that represent maximal continuous regions in which transcription occurs in any one or more samples. Interestingly, we identified 37 941 superTARs (covering 5.4 Mb), not overlapping annotated genes, suggesting that these TARs represent either unannotated exons of known genes or novel transcripts.

Gene ontology and InterPro domain annotation

We used Blast2GO (24) to annotate all predicted gene model protein sequences with gene ontology (GO) terms and protein names using the non-redundant protein sequences (nr) database at an *E*-value cut-off of 1×10^{-3}

and default weighting parameters in the GO term annotation step (see Blast2GO documentation for further details). This gave us a set of 9667 gene models with GO annotations. We also used Blast2GO to annotate each protein with InterPro (25) domains using InterProScan. The resulting GO terms and InterPro domains associated with each gene model provide researchers with valuable information on the putative functions of these *Oikopleura* proteins and allow further analyses of overrepresented GO terms on genes of interest. To facilitate the investigation of individual genes, we provide users of the browser with GO and InterPro domain identifiers and name codes associated with each gene (Figure 2). Furthermore, we provide links for each to the relevant entry in the GO/InterPro databases. The full sets of GO annotations and InterPro domain annotations are also available for download at OikoBase using the Downloads link.

Developmental transcriptome of *O. dioica*

To assemble a complete developmental transcriptome for *O. dioica*, we subdivided organismal development into 12 segments. We began with unfertilized oocytes ('Oocyte') that are arrested in Metaphase I of meiosis. Oocytes were collected from females that were allowed to spawn naturally by rupture of the gonad wall. The next stage consisted of two to eight cell embryos that encompass the maternal to zygotic transition in the transcriptional regulatory control of development. This was followed by a sample at 1 h pf, a stage at which blastomere fates for most tissue types have already been determined (26). Continued rapid cell proliferation within determined blastomere lineages was captured by the 'tailbud' and 'hatched' stages, the latter stage occurring when the animal emerges from the protective chorion. The 'early tadpole' stage is a period of very active organogenesis. The 'tailshift' stage represents the completion of metamorphosis and initiation of filter-feeding. This was then followed by samples isolated from stages in which the bulk of animal somatic growth occurs ('Day1', 'Day2' and 'Day3'). From the fourth day of development onwards, nutritional resources are increasingly allocated to growth and differentiation of the male and female reproductive organs ('Day4' and 'Day5'). To specifically interrogate development of the reproductive organs, we collected samples derived from dissected and isolated testes or ovaries of Day5.5 animals, as well as a complementary animal sample consisting of Day5.5 animals with their gonads removed ('trunk'). Importantly, we observed that culturing *O. dioica* at higher animal density ('restrictive conditions') causes it to cease development and somatic growth just prior to germline differentiation at Day3 (27). Therefore, we included three time points covering the entry into this developmental arrest ('Day2 dense', 'Day3 dense' and 'Day4 dense'). Material derived from these 18 samples was processed and hybridized to tiling arrays.

We used our processed tiling array data to calculate gene expression levels, at each stage, for each *Oikopleura* gene model covered by our array (see the 'Methods' section of OikoBase for details and qPCR validation) and include a matrix of these values in OikoBase. This gives researchers access to the developmental expression profiles of genes of interest as well as their expression during developmental stasis and in the male and female gonads and the trunk of adult animals. These data are provided in a format that facilitates genome-wide analyses.

In total, out of the 16 749 annotated genes that are covered by our tiling array, we find that 13 081 are expressed at some point during *O. dioica* development. The remaining 3668 genes (1161 of which have zero positive probes; the others have <50% positive CDS probes in any stage) could represent silent pseudogenes, genes expressed during very short time points not represented by our time course or genes having environment-specific roles such as during stress responses. Gene expression traffic is very dense in *O. dioica*, and an example of the complex interlaced developmental transcriptome profile of a short 12-kb region of the genome is shown in Figure 1. Note that the tiling array data are of sufficient resolution to identify a distal novel 5′-exon of the right-most gene in this interval (see legend of Figure 1 for details). The array data also accurately portray gene expression profiles associated with developmental life history characteristics of the animal. For example, secretion of the first *O. dioica* pre-house rudiment is detected in the tailshift stage and inflation of this rudiment into a functional house coincides with the beginning of active filter-feeding. Tissue-specifically, the houses are produced from a specialized oikoplastic epithelium (28–30) covering the trunk of the animal. Culturing *O. dioica* at higher animal density and reduced relative food availability causes animal growth and development to cease at the Day3 stage. Despite this, animals continue feeding and replace their house regularly upon house degeneration and filter clogging, demonstrating continued house production even in the absence of apparent animal and cellular growth (Day2 dense to Day4 dense). The full expression profile of a structural component of the house, oikosin 19 (31) faithfully recapitulates all of these features (Supplementary Figure S1).

DATA MINING OF THE OIKOPLEURA GENOME

Navigating OikoBase

The OikoBase tiling array-based genome browser contains various functions to query the *O. dioica* genome and the high-resolution developmental transcriptome. Entering the GBrowse link on the Facepage allows queries based on genomic coordinates or annotated gene model, gene transcript or protein identifiers (GSOIDG, GSOIDT and GSOIDP codes, respectively). For each region of interest, unique gene identifiers, CDS, transcript and GC-content can be visualized. Processed transcriptomics tracks for any, up to all, of the 18 samples can be activated and customized in linear or $\log_2$ scales. Please note that $\log_2$-transformed data appear 'smoothened' and less choppy, which may be preferable for some users who wish to view larger scale dynamic expression changes. Selecting the untransformed data files should be preferable for viewers who wish to identify more subtle changes in gene expression.

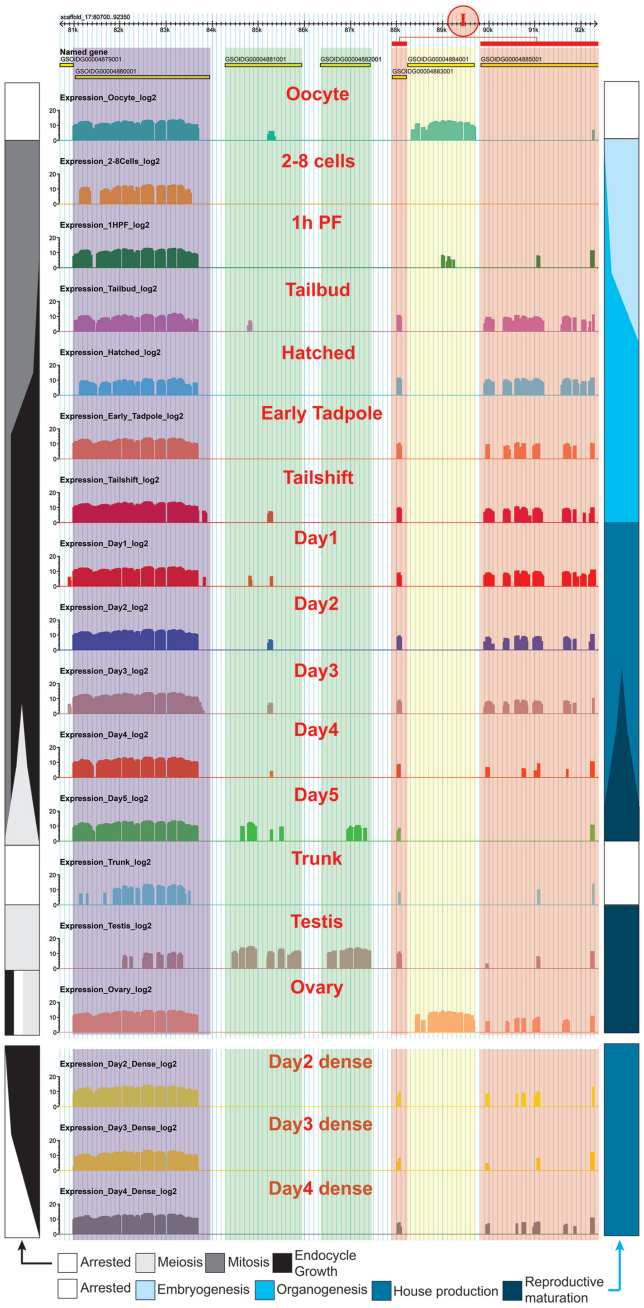

Figure 1. Complex developmental gene expression traffic in the compact *O. dioica* genome as visualized in the OikoBase genome browser. In addition to the normal browser output, the developmental stages are indicated in larger font in red type for clarity. Flanking the browser expression data, vertical timelines are included to illustrate key cell cycle (grey-scale, left-hand side) and organismal (blue-scale, right-hand side) processes and transitions covered by our transcriptomics analysis. The illustrated region includes testes-specific gene models (GSOIDG00004881001 and GSOIDG00004882001, green shading) that are also more weakly expressed in Day5 animals, a sub-portion of which contain testes at an earlier developmental stage. Immediately to the left of these is a ubiquitously expressed gene model (GSOIDG00004880001, purple shading). Notably, gene models GSOIDG00004883001 and GSOIDG00004885001 (pink shading) are expressed predominantly from tailbud to Day3, and RACE analysis indicated that they are part of the same gene (unpublished data; the red 'I' indicates their connectedness), expressed from a promoter left of GSOIDG00004883001. However, the short 5′-exon (GSOIDG00004883001) is not expressed in the ovary at which point GSOIDG00004885001 expression is driven by an internal bidirectional promoter also driving GSOIDG0004884001 (yellow shading), as confirmed by RACE and qRT–PCR analysis (unpublished data). This high diversity of developmentally regulated expression occurs in a region measuring <15 kb and indicates that our transcriptome analysis at high temporal and spatial resolution can improve gene model annotation. The Y-axis scale is exponential (log_2).

In order to create an annotation database that allows for mining of gene and protein information, we have built a pipeline that facilitates identification of both GO terms as well as InterPro terms on a gene-by-gene basis. Figure 2 illustrates these features in a composite genome browser shot of a 50-kb segment of scaffold one, containing several gene models. The operating mode for this pipeline is described in the Figure 2 legend. Options to directly download transcript (GSOIDT code) and protein (GSOIDP code) sequences of interest are also provided, in addition to links that access GO and InterPro domain information. These features should prove very useful in rapidly obtaining additional information for genes expressed in developmental patterns of interest identified in the transcriptomics portion of the OikoBase browser.

BLAST pipeline

We have incorporated a BLAST pipeline that allows researchers to find relevant *Oikopleura* homologs to their genes of interest from other organisms. Entering via the BLAST option on the OikoBase Facepage, the complete *O. dioica* reference genome as well as ESTs and gene model-based transcript and proteome resources can be directly queried. An example of this workflow using the human cytoplasmic actin protein sequence as a query is developed in Figure 3. This function permits rapid assessment of expression profiles of putative homologs and additional annotation as well as corresponding nucleotide and protein sequences can be retrieved (Figure 2) for any desired subsequent analyses. Finally, following the Gene Expression Matrix link on the Facepage, it is possible to access the entire set of processed expression values for each annotated gene model across developmental time in an HTML-sortable table. Selecting a desired gene model identifier (GSOIDG code) allows retrieval of the developmental expression profile of that gene model.

CONCLUSIONS AND PERSPECTIVES

Here we have presented a tiling array-based genome browser resource for the research community which provides unbiased whole-genome transcription data across the key events in development of the urochordate *O. dioica* as well as providing somatic-, testes- and ovary-specific transcriptional outputs. In addition to the comprehensive chronological resolution of the developmental transcriptome profile, the spatial resolution of the data is such that the small (47-bp peak size) introns that predominate in the compact *O. dioica* genome are reliably detected. These data are useful in improving the gene model annotation of the *Oikopleura* genome and in

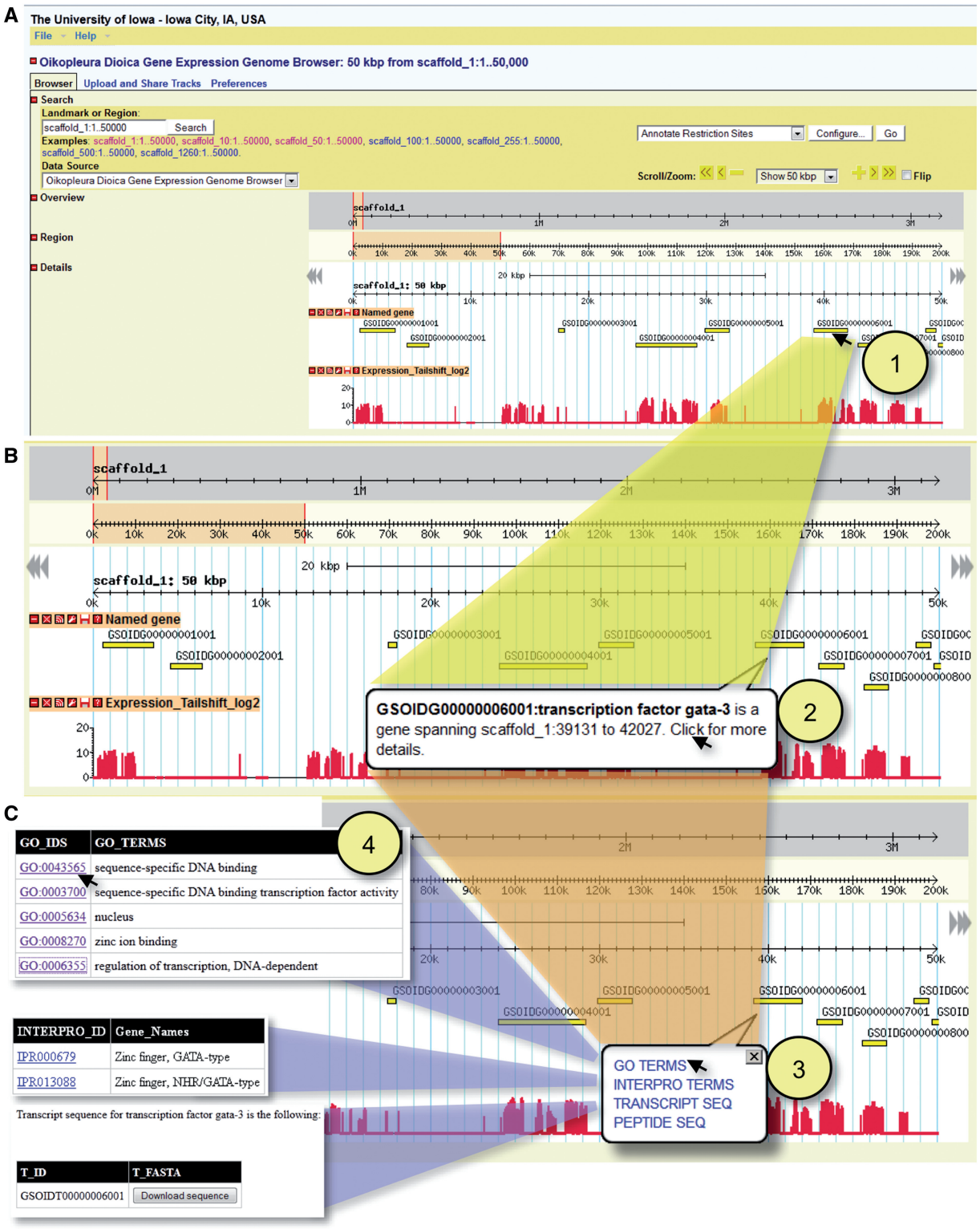

Figure 2. Data mining in OikoBase. (**A**) Browser window of a 50-kb segment of scaffold 1 showing several genes, with the tailshift expression track selected as an example of a time point. (**B**) By placing the pointer over a gene model (1) (here, the sixth gene from the left) a balloon appears that specifies the gene name, gata-3 (assigned by Blast2GO) and its genomic coordinates (2). (**C**) By clicking on the gene model, a different balloon (3) is opened in its place which allows links to both the GO website (http://www.geneontology.org/) (4) and the InterPro website (http://www.ebi.ac.uk/interpro/). By clicking on either link, the browser is redirected to the relevant GO or InterPro terms associated with this gene which catalogue information regarding biological processes, molecular functions and cellular components as well as evolutionarily conserved protein domains. Options to directly download nucleotide (TRANSCRIPT SEQ) and protein (PEPTIDE SEQ) are also provided in this balloon.

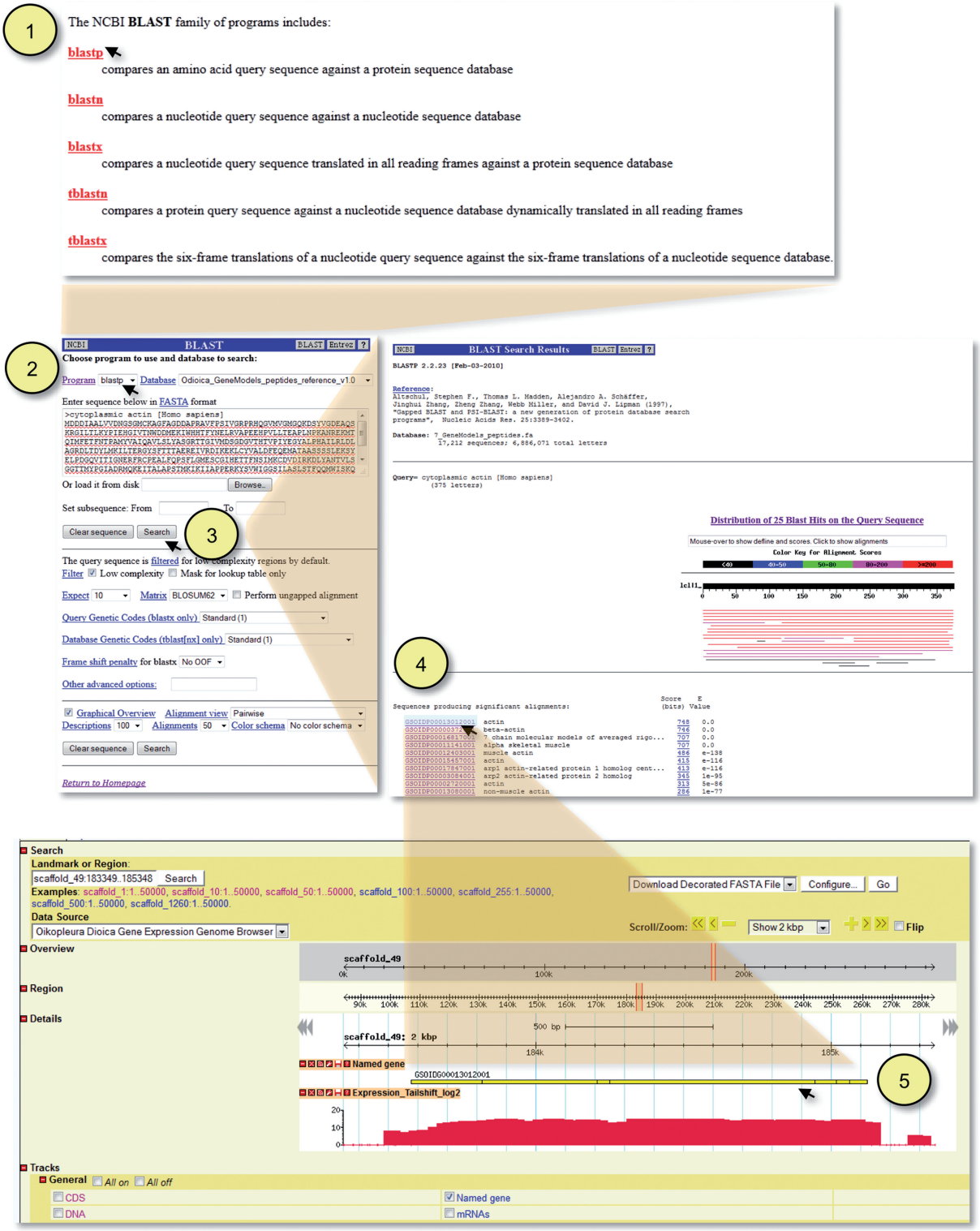

Figure 3. Obtaining developmental expression profiles for *O. dioica* homologs of genes-of-interest from other organisms. To provide an example of how to use this pipeline, the human cytoplasmic actin protein has been selected as the query sequence. First, the 'Programs available for the BLAST search' page (1) is accessed by clicking the Facepage BLAST link. The desired Program should then be matched up with the appropriate Database using the dropdown menus (2). Note that selecting 'blastp' (as in this example) or 'blastx' will automatically pair these programs with the '*O. dioica* peptides reference' database. After executing the search (3), the desired putative *O. dioica* homologs are displayed on the BLAST search results page where unique GSOID identifiers and corresponding gene names are provided (4). By clicking on the desired live link GSOIDP identifier the user is taken to the location of the putative homolog (5). Subsequently, any desired additional annotation or transcript tracks can be activated in the browser. Here, the expression of an *O. dioica* cytoplasmic actin homolog is shown at the tailshift stage with the view zoomed to 2 kb.

identifying transcribed unannotated DNA which likely represent new transcript isoforms of known genes as well as uncharacterized novel genes. The rapid evolution of *O. dioica* has included lineage-specific expansion of gene families including central developmental genes (10). The analysis of retention and diversification of expression patterns among paralogs should therefore provide insights into mechanisms of sub- and neofunctionalization. For example, a preliminary transcriptomics analysis shows that genes containing the most overrepresented domains (32) have highly clustered expression profiles, correlating well with the scaffolding of specific house substructures, a process that requires coordination of cell growth and positioning in addition to tuning of metabolic output within cell fields and between cell fields. The generation of multiple gene paralogs would allow for modulation of cell-specific functional output, as has previously been shown for paralogs of the house structural components, 'oikosins' (28,29,31). Numerous regulators of mitotic cell division have also been amplified in *O. dioica* (12), and our data indicate that expression of these paralogs is frequently anti-correlated, with novel variants expressed in non-mitotic tissues, such as endocycling cells and differentiating testes. Therefore, combining the developmental transcriptomics with DNA sequence divergence of these paralogs could yield novel insights into the evolutionary constraints on mitotic determinants as well as the genetic adaptations required to accommodate variant cell cycles.

Complementing the existing Genoscope *Oikopleura* database, OikoBase now provides transcriptomics data across the animal's entire life cycle. OikoBase also annotates Genoscope gene models with predicted functions including GO terms, InterPro domains and informative gene names, giving researchers readily accessible, biologically relevant data on these genes through a browser interface. We have also implemented a BLAST function that allows researchers to search for homologs of their gene of interest. As a more systemic library of temporal and spatial *O. dioica in situ* images becomes available to significantly complement the temporal transcriptome profiles, such a library will also be incorporated into OikoBase. Genoscope provides sequence resources for a large and expanding number of species, whereas OikoBase is a more dedicated and focused database that will remain an active and improving urochordate resource. Studies are underway to provide more detailed analysis of developmental programs and we plan to layer into the OikoBase resource further transcriptomics interrogations of animal responses to environmental stressors (both tiling array and RNA-seq data), developmental CAGE data for accurate annotation of transcription start sites, and global analysis of histone modifications (regulating chromatin structure and transcriptional states in specific developmental contexts) through ChIP-chip and ChIP-seq.

IMPLEMENTATION

Genetic Model Organism Database (GMOD; http://www.gmod.org) tools MySQL version 5.1.47 (x86_64) and GBrowse 2.07 were implemented to create the OikoBase genome browser. Perl scripts were used to manage and load the appropriate databases. All transcriptome array data were converted to the Genetic Feature Format Version 3 (GFF3). We also installed a BLAST feature accessible from the Facepage to query sequences against the *Oikopleura* genome. The Facepage was created using JavaScript, Jquery, CSS and HTML5 scripts.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figure 1.

ACKNOWLEDGEMENTS

The authors thank Jean-Marie Bouquet and the staff of Appendic Park, Sars Centre, for their efforts in producing a continuous supply of *Oikopleura* and David Osborne for identifying the phenomenon of density induced developmental stasis. The authors also thank Willem Haagmans, Roche NimbleGen, for initial consultations on the design of the custom *Oikopleura* tiling array.

FUNDING

The Norwegian Research Council [183690/S10 NFR-FUGE to D.C. and E.M.T., 204891/F20 NFR-FRIBIO to J.R.M. and E.M.T.]; NFR [133335/V40 to E.M.T.]; Funding for open access charge: J Robert Manak start-up funds.

Conflict of interest statement. None declared.

REFERENCES

1. Gorsky,G. and Fenaux,R. (1998) The role of Appendicularia in marine food webs. In: Bone,Q. (ed.), *The Biology of Pelagic Tunicates*. Oxford University Press, New York, pp. 161–169.
2. Sagane,Y., Zech,K., Bouquet,J.M., Schmid,M., Bal,U. and Thompson,E.M. (2010) Functional specialization of cellulose synthase genes of prokaryotic origin in chordate larvaceans. *Development*, **137**, 1483–1492.
3. Alldredge,A.L. (2005) The contribution of discarded appendicularian houses to the flux of particulate organic carbon from ocean surface waters. In: Gorsky,G., Yongbluth,M.J. and Deibel,D. (eds), *Response of Marine Ecosystems to Global Change*. Contemporary Publishing International, Paris, pp. 309–326.
4. Robison,B.H., Reisenbichler,K.R. and Sherlock,R.E. (2005) Giant larvacean houses: rapid carbon transport to the deep sea floor. *Science*, **308**, 1609–1611.
5. Delsuc,F., Brinkmann,H., Chourrout,D. and Philippe,H. (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, **439**, 965–968.
6. Troedsson,C., Bouquet,J.M., Aksnes,D.L. and Thompson,E.M. (2002) Resource allocation between somatic growth and reproductive output in the pelagic chordate *Oikopleura dioica* allows opportunistic response to nutritional variation. *Mar. Ecol. Prog. Ser.*, **243**, 83–91.
7. Bouquet,J.M., Spriet,E., Troedsson,C., Ottera,H., Chourrout,D. and Thompson,E.M. (2009) Culture optimization for the emergent zooplanktonic model organism *Oikopleura dioica*. *J. Plankton Res.*, **31**, 359–370.
8. Fenaux,R., Bone,Q. and Deibel,D. (1998) Appendicularian distribution and zoogeography. In: Bone,Q. (ed.), *The Biology of Pelagic Tunicates*. Oxford University Press, New York, pp. 251–264.

9. Seo,H.C., Kube,M., Edvardsen,R.B., Jensen,M.F., Beck,A., Spriet,E., Gorsky,G., Thompson,E.M., Lehrach,H., Reinhardt,R. *et al*. (2001) Miniature genome in the marine chordate *Oikopleura dioica*. *Science*, **294**, 2506.
10. Denoeud,F., Henriet,S., Mungpakdee,S., Aury,J.M., Da Silva,C., Brinkmann,H., Mikhaleva,J., Olsen,L.C., Jubin,C., Cañestro,C. *et al*. (2010) Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science*, **330**, 1381–1385.
11. Schulmeister,A., Schmid,M. and Thompson,E.M. (2007) Phosphorylation of the histone H3.3 variant in mitosis and meiosis of the urochordate *Oikopleura dioica*. *Chrom. Res*., **15**, 189–201.
12. Campsteijn,C., Øvrebø,J.I., Karlsen,B.O. and Thompson,E.M. (2012) Expansion of cyclin D and CDK1 paralogs in *Oikopleura dioica*, a chordate employing diverse cell cycle variants. *Mol. Biol. Evol*., **29**, 487–502.
13. Korf,I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.
14. Roest Crollius,H., Jaillon,O., Bernot,A., Dasilva,C., Bouneau,L., Fischer,C., Fizames,C., Wincker,P., Brottier,P., Quétier,F. *et al*. (2000) Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet*., **25**, 235–238.
15. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*., **40**, D71–D75.
16. Birney,E., Clamp,M. and Durbin,R. (2004) GeneWise and Genomewise. *Genome Res*., **14**, 988–995.
17. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
18. Parra,G., Blanco,E. and Guigo,R. (2000) GeneID in *Drosophila*. *Genome Res*., **10**, 511–515.
19. Slater,G.S. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.
20. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al*. (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **40**, D13–D25.
21. Mott,R. (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci*., **13**, 477–478.
22. Howe,K.L., Chothia,T. and Durbin,R. (2002) GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res*., **12**, 1418–1427.
23. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al*. (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res*., **39**, D1005–D1010.
24. Götz,S., García-Gómez,J.M., Terol,J., Williams,T.D., Nagaraj,S.H., Nueda,M.J., Robles,M., Talón,M., Dopazo,J. and Conesa,A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*., **36**, 3420–3435.
25. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al*. (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res*., **40**, D306–D312.
26. Stach,S., Winter,J., Bouquet,J.-M., Chorrout,D. and Schnabel,R. (2008) Embryology of a planktonic tunicate reveals traces of sessility. *Proc. Natl Acad. Sci. USA*, **105**, 7729–7234.
27. Ganot,P., Kallesøe,T. and Thompson,E.M. (2007) The cytoskeleton organizes germ nuclei with divergent fates and asynchronous cycles in a common cytoplasm during oogenesis in the chordate *Oikopleura*. *Dev. Biol*., **302**, 577–590.
28. Spada,F., Steen,H., Troedsson,C., Kallesoe,T., Spriet,E., Mann,M. and Thompson,E.M. (2001) Molecular patterning of the oikoplastic epithelium of the larvacean tunicate *Oikopleura dioica*. *J. Biol. Chem*., **276**, 20624–20632.
29. Thompson,E.M., Kallesøe,T. and Spada,F. (2001) Diverse genes expressed in distinct regions of the trunk epithelium define a monolayer cellular template for construction of the oikopleurid house. *Dev. Biol*., **238**, 260–273.
30. Ganot,P. and Thompson,E.M. (2002) Patterning through differential endoreduplication in epithelial organogenesis of the chordate, *Oikopleura dioica*. *Dev. Biol*., **252**, 59–71.
31. Hosp,J., Sagane,Y., Danks,G. and Thompson,E.M. (2012) The evolving proteome of a complex extracellular matrix, the *Oikopleura* house. *PLoS One*, **7**, e40172.
32. Chavali,S., Morais,D.A., Gough,J. and Babu,M.M. (2011) Evolution of eukaryotic genome architecture: Insights from the study of a rapidly evolving metazoan, *Oikopleura dioica*: non-adaptive forces such as elevated mutation rates may influence the evolution of genome architecture. *Bioessays*, **33**, 592–601.