

ANISEED 2017: extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets

Matija Brozovic, Christelle Dantec, Justine Dardaillon, Delphine Dauga, Emmanuel Faure, Mathieu Gineste, Alexandra Louis, Magali Naville, Kazuhiro R. Nitta, Jacques Piette, et al.

▶ **To cite this version:**

Matija Brozovic, Christelle Dantec, Justine Dardaillon, Delphine Dauga, Emmanuel Faure, et al.. ANISEED 2017: extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets. Nucleic Acids Research, 2018, 46 (D1), pp.D718-D725. 10.1093/nar/gkx1108 . hal-01636650

HAL Id: hal-01636650
https://hal.science/hal-01636650

Submitted on 18 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

D718–D725 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 15 November 2017*
doi: 10.1093/nar/gkx1108

ANISEED 2017: extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets

Matija Brozovic[1,†], Christelle Dantec[1,†], Justine Dardaillon[1,†], Delphine Dauga[2,†], Emmanuel Faure[3,4], Mathieu Gineste[1], Alexandra Louis[5], Magali Naville[6], Kazuhiro R. Nitta[7], Jacques Piette[1], Wendy Reeves[8], Céline Scornavacca[9], Paul Simion[9], Renaud Vincentelli[10], Maelle Bellec[11], Sameh Ben Aicha[12], Marie Fagotto[11], Marion Guéroult-Bellone[2], Maximilian Haeussler[13], Edwin Jacox[1], Elijah K. Lowe[14,15], Mickael Mendez[7], Alexis Roberge[11], Alberto Stolfi[16], Rui Yokomori[17], C. Titus Brown[15,18], Christian Cambillau[10], Lionel Christiaen[19], Frédéric Delsuc[9], Emmanuel Douzery[9], Rémi Dumollard[12], Takehiro Kusakabe[20], Kenta Nakai[17], Hiroki Nishida[21], Yutaka Satou[22], Billie Swalla[15,23], Michael Veeman[8], Jean-Nicolas Volff[6] and Patrick Lemaire[1,3,*]

[1]CRBM, Université de Montpellier, CNRS, Montpellier, France, [2]Bioself Communication; 28 rue de la Bibliothèque, F-13001 Marseille, France, [3]Institut de Biologie Computationnelle, Université de Montpellier, Montpellier, France, [4]Team VORTEX, Institut de Recherche en Informatique de Toulouse, Universities Toulouse I and III, CNRS, INPT, ENSEEIHT; 2 rue Camichel, BP 7122, F-31071 Toulouse Cedex 7, France, [5]DYOGEN, IBENS, Département de Biologie, Ecole Normale Supérieure, CNRS, Inserm, PSL Research University, F-75005, Paris, France, [6]Institut de Génomique Fonctionnelle de Lyon, Université de Lyon, Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, CNRS; 46 allée d'Italie, F-69364 Lyon, France, [7]IBDM, Aix-Marseille Université, CNRS, Campus de Luminy, Case 907; 163 Avenue de Luminy, F-13288 Marseille Cedex 9, France, [8]Division of Biology, Kansas State University, Manhattan, Kansas, [9]ISEM, Université de Montpellier, CNRS, IRD, EPHE, Montpellier, France, [10]AFMB, Aix-Marseille Université, CNRS, Campus de Luminy, Case 932, 163 Avenue de Luminy, F-13288 Marseille Cedex 9, France, [11]Université de Montpellier, Montpellier, France, [12]Laboratoire de Biologie du Développement de Villefranche-sur-mer (LBDV), Sorbonne Universités, Université Pierre-et-Marie-Curie, CNRS; Quai de la Darse, F-06234 Villefranche-sur-Mer Cedex, France, [13]Santa Cruz Genomics Institute, MS CBSE, University of California, 1156 High Street, Santa Cruz, CA 95064, USA, [14]Department of Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Napoli, Italy, [15]BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, MI48824, USA, [16]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA, [17]Human Genome Center, the Institute of Medical Science, the University of Tokyo, 4-6-1 Shirokanedai, Minato, Tokyo 108-8639, Japan, [18]Population Health and Reproduction, UC Davis, Davis, CA 95616, USA, [19]New York University, Center for Developmental Genetics, Department of Biology, 1009 Silver Center, 100 Washington Square East, New York City, NY10003, USA, [20]Department of Biology, Faculty of Science and Engineering, Konan University, Kobe 658-8501, Japan, [21]Department of Biological Sciences, Graduate School of Science, Osaka University, 1-1 Machikaneyama-cho, Toyonaka, Osaka 560-0043, Japan, [22]Department of Zoology, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan and [23]Friday Harbor Laboratories, 620 University Road, Friday Harbor, WA 98250-9299, USA

Received September 18, 2017; Revised October 22, 2017; Editorial Decision October 23, 2017; Accepted November 09, 2017

ABSTRACT

ANISEED (www.aniseed.cnrs.fr) is the main model organism database for tunicates, the sister-group of vertebrates. This release gives access to annotated genomes, gene expression patterns, and anatomical descriptions for nine ascidian species. It pro-

*To whom correspondence should be addressed. Email: patrick.lemaire@crbm.cnrs.fr
†These authors contributed equally to this work as first authors.

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D718/4628129 by BIU Montpellier user on 18 December 2019

vides increased integration with external molecular and taxonomy databases, better support for epigenomics datasets, in particular RNA-seq, ChIP-seq and SELEX-seq, and features novel interactive interfaces for existing and novel datatypes. In particular, the cross-species navigation and comparison is enhanced through a novel taxonomy section describing each represented species and through the implementation of interactive phylogenetic gene trees for 60% of tunicate genes. The gene expression section displays the results of RNA-seq experiments for the three major model species of solitary ascidians. Gene expression is controlled by the binding of transcription factors to *cis*-regulatory sequences. A high-resolution description of the DNA-binding specificity for 131 *Ciona robusta* (formerly *C. intestinalis* type A) transcription factors by SELEX-seq is provided and used to map candidate binding sites across the *Ciona robusta* and *Phallusia mammillata* genomes. Finally, use of a WashU Epigenome browser enhances genome navigation, while a Genomicus server was set up to explore microsynteny relationships within tunicates and with vertebrates, *Amphioxus*, echinoderms and hemichordates.

INTRODUCTION

The tunicate clade of marine invertebrate deuterostomes includes the sessile ascidians and the pelagic appendicularians and thaliaceans (1,2). A major interest of the scientific community for tunicates stems from the phylogenetic position of this chordate group, now considered to be the long-diverged sister group of vertebrates (3). Solitary ascidians and all appendicularians reproduce sexually only. By contrast, colonial ascidians and all thaliaceans reproduce both sexually and asexually by budding (1).

Ascidians, a paraphyletic group (4), are the largest and most studied group of tunicates. Major solitary ascidian model species, studied for their embryogenesis, aging and regeneration properties are *Ciona robusta* (formerly *Ciona intestinalis* type A), *Ciona intestinalis* (formerly *Ciona intestinalis* type B), *Ciona savignyi*, *Phallusia mammillata*, *Halocynthia roretzi* and *Molgula oculata*. The colonial ascidian *Botryllus schlosseri* is well known for studies on whole-body regeneration and stem cells (5), and on self/non-self recognition (6). Comparison of ascidian and vertebrate strategies can provide cues on the origin of vertebrate novelties such as the neural crest (7) or the second heart field (8). Their rapid molecular evolution is also informative on the plasticity of the chordate program of gene regulation (8–11) or of histocompatibility (6). In addition to this basic science focus, many solitary and colonial ascidian species are invasive, major contributors to bio fouling and, as such, pests for the aquaculture industry (12,13).

Over the past 20 years, an extensive functional genomics toolbox (14) has been developed to make best use of the simplicity of the embryo and genome (15) of *Ciona intestinalis* and *robusta*. This tool has now been extended/adapted to other tunicate species (16,17). Interference with gene function, initially relying on morpholino oligonucleotides, now includes the use of TALENs (18) and CRISPR/cas9 strategies (19,20).

ANISEED - Ascidian Network for In Situ Expression and Embryological Data—is the major database system in the Tunicate community, hosting and giving access to genomic, genetic and anatomical data on ascidian development. First published in 2010 (21), the database hosts and gives access to genomic, genetic and anatomical data on tunicates. Its refactoring using the Chado schema philosophy and its extension to 11 ascidian species was completed in 2015 (22). Since then, tunicate laboratories have rapidly embraced genomics, including ChIP-seq (23,24), RNA-seq of whole embryos (24) or FACS-sorted single cells (25), bisulphite sequencing (26), SELEX-seq and TSS-seq (27). The aim of the ANISEED 2017 update is to accompany the entry of tunicate biology into the genome-scale era through three types of improvements: extension/refactoring of mining and display tools, extension of pre-existing datasets and development of modules to host and give access to novel types of genomics data.

NEW DATA MINING/DISPLAY TOOLS

Taxonomy pages

A taxonomy page accessible from the home page was introduced in ANISEED 2017 to place the covered species into their phylogenetic context and facilitate the interpretation of comparative studies. The page shows a cladogram of the twelve species currently described at the anatomical or genetic level in the database. Each node gives access to a page describing the corresponding species. Following the novel taxonomy (28) and a community decision at the 2017 international Tunicate Meeting, *Ciona intestinalis* type A has been renamed *Ciona robusta* throughout the database; while *Ciona intestinalis* type B is now referred to as *Ciona intestinalis*. No molecular data are currently associated to this latter species in ANISEED, as its genome draft is of insufficient quality to build gene models.

Each species page briefly describes the species, its distribution and its ecology, lists the number of ANISEED entries and provides a link to the corresponding ANISEED genome browser. This page also links to other tunicate molecular databases (DBTGR (29), CiPro (30), Ghost (31) and FABA (32) for *Ciona robusta*), to the CRISPOR CRISPR/Cas9 guide RNA design tool for ascidian species (33), as well as to major ecological databases, including the Ascidiacea World Database from the World Registry of Marine species (WoRMS) (34), the Global Biodiversity Information Facility (GBIF) (35), or the Global Invasive Species Database (http://www.invasivespecies.net/). Finally, for practical purposes, the page lists for each species the animal providers used in the scientific community.

Interactive gene phylogenies. The diversity of ascidian species with an assembled genome provides a framework to study the evolution of developmental mechanisms. To identify orthologous genes within tunicates and with chordates, ANISEED 2015 used an orthology pipeline based on the OrthoMCL software (36) and listed for each gene the cluster of homologous genes it was associated to. In many cases,

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D718/4628129 by BIU Montpellier user on 18 December 2019

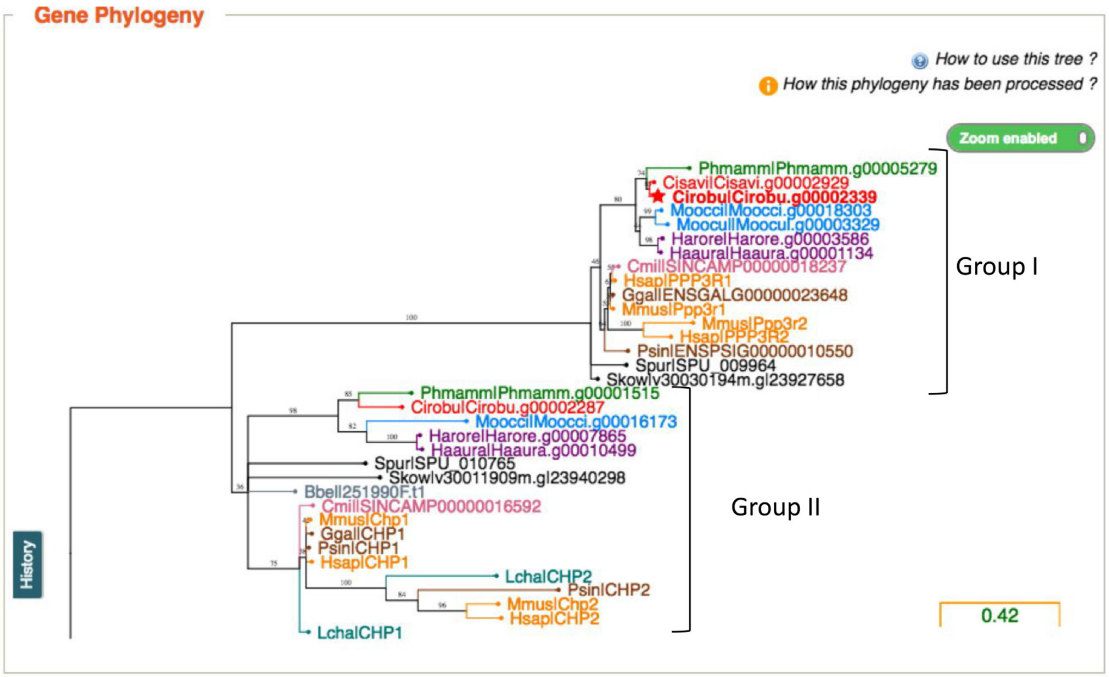


Figure 1. Interactive gene Phylogenies. Screenshot from a part of the phylogenetic gene tree including genes Cirobu.g00002339 and Cirobu.g00002287. Species are color-coded and the tree can be compressed or extended lengthwise and zoomed in and out. Note the difference between the two groups of genes. In group I, the gene phylogeny follows the species phylogeny, and tunicate genes are considered orthologous to the vertebrate genes of the group. In group II by contrast, the gene phylogeny does not follow the species phylogeny, and we consider that the tunicate genes have no unequivocal vertebrate orthologs.

clusters were large, and in the absence of phylogenetic trees, the system did not allow the identification of accurate 1-to-1 orthology relationships within the list.

We thus refactored the ANISEED 2017 orthology pipeline. We first used SiLiX (37) to build clusters of homologous genes from 18 species (6 vertebrates, 9 ascidians, 1 lancelet and 2 non-chordate deuterostome outgroups). A phylogenetic tree was then inferred for each cluster using RaxML (38) under the LG+Γ4+F evolution model (See supplementary methods). Trees for approximately 60% of the coding gene models of each ascidian species are displayed interactively on the corresponding gene card using PhyloCanvas (http://phylocanvas.org) (Figure 1). This procedure identifies unequivocal vertebrate orthologs for ∼30% of ascidian genes (average number of human orthologs/*C. robusta* gene: 2.8). A similar frequency of tunicate genes with clear human orthologs is also observed in ENSEMBL 90 (28% of *C. robusta* genes with high-confidence human orthologues). While this situation may reflect the large genetic distance between tunicates and vertebrates, we noticed that orthology relationships could also be confounded by the position the non-olfactorian outgroups (Figure 1). Approximately 25% of ascidian genes without clear vertebrate orthologs have at least one ortholog in an ascidian of a different order (Supplementary Table S1).

New genome browsers

ANISEED 2015 featured Gbrowse 2.0 (39) genome browsers, which showed limitations as the size of the dataset increased. The WashU Epigenome browser (40) was selected as a replacement and adapted with the help of its developer. It offers increased speed and the ability to represent genome-wide chromatin interaction data (41). Within or between genome comparisons are facilitated by the display in a unique window of concatenations of sets of gene loci within a species, or between species (e.g. orthologous loci to explore synteny relationships).

Tracks presented are grouped into three categories. 'Annotation tracks' provide general background information (gene models, repeated elements, inter-species genome sequence comparison, genome alignment, operons, *cis*-regulatory sequences, Ns and GC content). 'Public track hubs' group tracks presenting the information gathered within a specific project or using a specific genome-scale technique (RNA-seq, TSS-seq, methyl-seq, SELEX-seq, ChIP-seq: see below). For each public hub, a precise description of the content and methods is provided. Finally, 'Custom tracks' can be uploaded as individual tracks or hubs. Each session can be saved and shared with other scientists. A useful wiki handbook is provided by the developer (wiki.wubrowse.org).

The WashU and Gbrowse 2.0 browsers will coexist until summer 2018, after which only the WashU browser will remain. During this 1-year transition period, the addition of new tracks, some of which are presented here, will be restricted to the WashU browser.

Tunicate genomes rapidly diverge at the local sequence level, as well as in their organization. To further explore the conservation of genome organization within deuterostomes, we implemented a Genomicus synteny browser (42),

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D718/4628129 by BIU Montpellier user on 18 December 2019

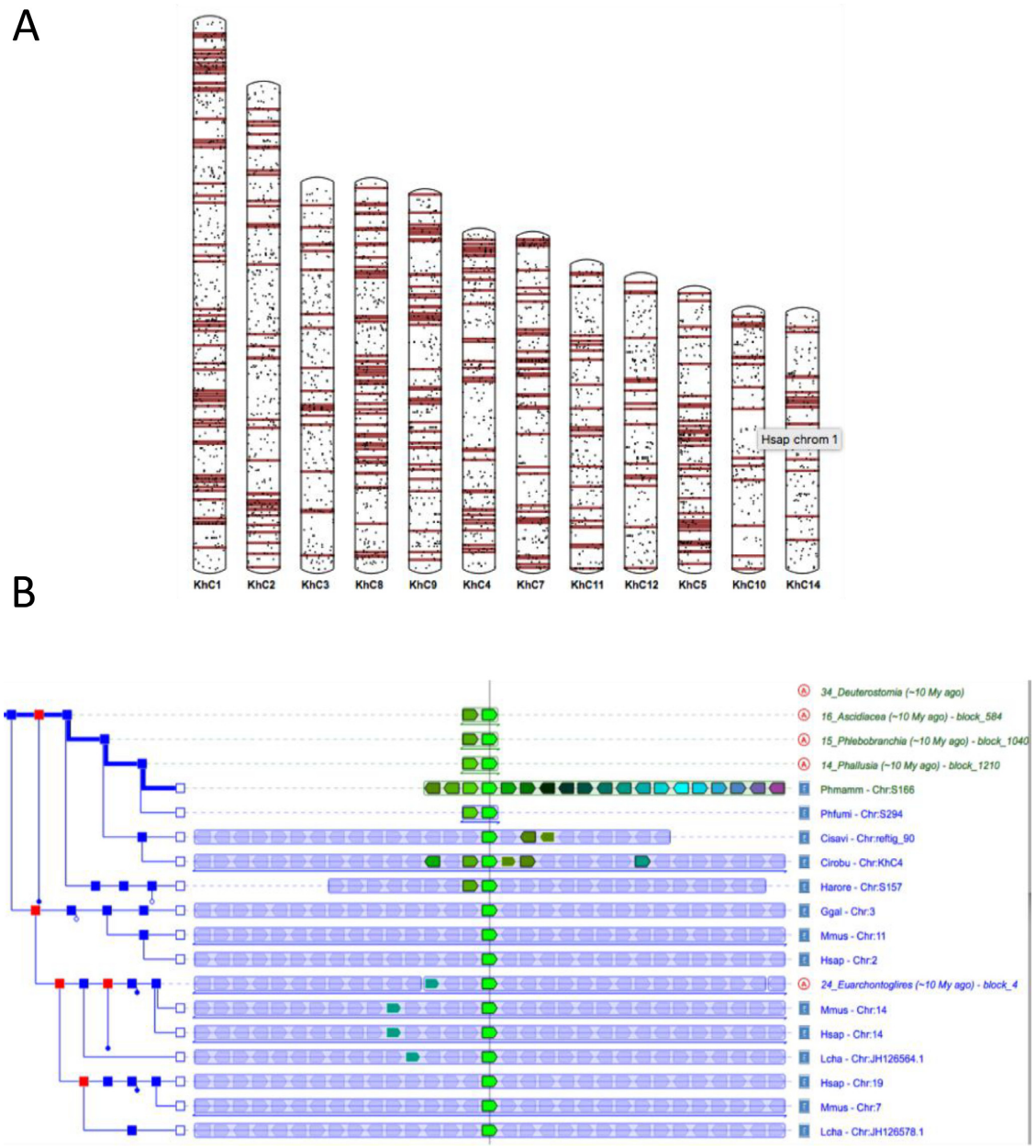


Figure 2. Genomicus and the study of global and local genome rearrangements. (**A**) Karyotype view showing the mapping of segments of Human Chromosome 1 onto *Ciona robusta* chromosomes. (**B**) Phyloview of the *Phallusia mammillata* gene Phmamm.g00004580 showing rapid loss of microsynteny within tunicates and complete loss with vertebrates.

based on the refined orthology relationships described above. This tool provides powerful visualization interfaces, such as karyotype View to explore genome organization (Figure 2A) and PhyloView to display the evolution of microsynteny relationships within ascidians and with vertebrates or non-chordate deuterostomes (Figure 2B).

EXTENSION OF PREEXISTING DATA TYPES

Functional gene annotations

Functional genome annotation (transcript and gene models, orthology to other tunicates, to vertebrates and to other deuterostomes, best blastp hit to human proteins, Interpro and GO annotations) was extended to all species with a genome assembly of sufficient quality to support gene model building, including three species previously only represented on the genomic browser: *Phallusia fumigata*, *Halocynthia aurantium* and *Molgula occidentalis*.

Gene models are annotated on each gene page and also provided as annotation tracks on the browsers. A link to alternative NCBI and ENSEMBL gene models is provided on each gene card. Manually-curated gene names respecting the tunicate nomenclature guidelines (43) are currently being generated by a dedicated committee and will be gradually added until the next official release. Genes can currently be searched for in ANISEED and in the WashU browser by

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D718/4628129 by BIU Montpellier user on 18 December 2019

Table 1. Evolution of the number of entries in ANISEED between October 2010 and October 2017

Class of entries	2010		2015				2017				
Species	*Halocynthia roretzi*	*Ciona robusta*	*Halocynthia roretzi*		*Ciona robusta*		*Halocynthia roretzi*		*Ciona robusta*		*Phallusia mammillata*
Measure;	*Number*	*Number*	*Number*	*% Increase /2010*	*Number*	*% Increase /2010*	*Number*	*% Increase /2015*	*Number*	*% Increase /2015*	*Number*
Articles in Pubmed (query: species + gene + development)	-	214	NA	NA	280	31%	79	NA	339	21%	7
Articles in ANISEED	1	160	1	0	217	36%	4	300%	253	17%	0
Molecular Tools	0	619	0	NA	769	24%	0	NA	874	14%	0
Regulatory Regions Analyzed (literature)	0	528	0	NA	875	66%	0	NA	1102	26%	0
Spatio-temporal patterns of cis-regulatory activity (literature)	0	777	0	NA	944	21%	0	NA	1152	22%	0
Gene Expression Patterns in Mutant Embryo (literature)	0	1152	0	NA	1294	12%	51	NA	1860	44%	0
Gene Expression Patterns (literature)	28	4259	51	82%	6545	54%	154	202%	7717	18%	0
Gene Expression Patterns (large scale)	5825	21017	5825	0%	21017	0%	5825	0%	21538	2%	262
Total Gene Expression Patterns	5853	25276	5876	0%	27562	9%	5979	2%	29255	6%	262
Nb of genes described	933	4000	937	0%	4500	13%	943	1%	4750	6%	31

unique gene or gene model Ids, by name of their most related human gene, or through the ANISEED blast server, which can now blast protein or nucleotide queries against both whole genomes or transcript model sets.

The functional annotation of ascidian gene models is based on Gene Ontology (44) terms. Because of the complexity and depth of the GO ontology, containing more than 40 000 terms, genes with similar function can be annotated with distinct terms, which can confound GO term enrichment analyses (45). To overcome this issue, we used an objective framework Information Content algorithm (46, stringency parameter 0.2) to define a smaller Tunicate GO Slim of 633 terms (MF:129; BP: 374; CC: 130), which provide higher level and more consistent annotations. We mapped global GO annotations onto this GO slim and propose this annotation in the download section of the database. For *Ciona robusta*, this process led to the annotation of 69% of genes with at least one term of each MF, BP and CC category.

Manually curated data

ANISEED hosts both large-scale data, automated annotations, and manually curated data from the literature or communicated by partner laboratories. This latter category, which now forms 30% of total gene expression patterns in *Ciona robusta*, is crucial to contextualize large-scale datasets, and is the primary source of expression profiles in altered developmental conditions (gene knock-down or overexpression; 6% of *Ciona robusta* dataset). Table 1 presents the growth of major sections of ANISEED between 2010 and 2017. The majority of the new data originates from the manual curation of 33 articles describing the development of *Ciona robusta*, published between 2012 and 2016, including 19 describing neural development, a very active field of *Ciona* research. The newly curated dataset set also includes 3 *Halocynthia roretzi* articles.

NEW DATASETS

Annotated Repeated elements

Repeated elements, previously not described, were detected for all nine species, and each element was analysed and named according to the guidelines for the nomenclature of tunicate genetic elements (43). These elements are displayed in the 'Annotation tracks' section of the WashU genomic browser.

RNA-seq

RNA-seq provides a quantitative estimation of the steady-state expression level of genes in wild-type and experimentally perturbed conditions. Gene cards in ANISEED 2017 include in a new section of the expression tab a summary of gene expression dynamics across all wild-type and mutant RNA-seq experiments analysed (Supplementary Figure S1). For each experiment two complementary but comparable normalizations are displayed: FPKM/RPKM and an 'unlogged' version of RLE (47). The former is best to compare the expression of different genes in the same experiment, while the latter is a better choice to compare the expression of a given gene across experiments. Note that 'unlogged' RLE values are not suitable for statistical analysis. The download button of the page thus returns the true logarithmic RLE value.

Two sets of data of immediate interest to the community were quantified. The first dataset is composed of unpublished stranded RNA-seq data for whole embryos of seven developmental stages ranging from egg to hatching larva in three species (*Ciona robusta, Phallusia mammillata and Halocynthia roretzi*). This dataset allows comparing the expression dynamics of orthologous genes in these three divergent ascidian species. Note that the *Ciona* data are from *Ciona intestinalis* embryos, but in the absence of a high quality *C. intestinalis* genome assembly, were mapped onto the *Ciona robusta* genome and considered to reflect the expression of the *Ciona robusta* orthologs. The second dataset (NCBI bioproject PRJDB3843) compares wild-type, Bmp-treated and Bmp-inhibited (Dorsomorphin) *Ciona robusta* late gastrula embryos. Additional datasets will be progressively integrated as they become public. Each of these datasets is also integrated as a public hub in the WashU browser.

RNA-seq usually poorly represents immature unspliced transcripts. This technique is therefore inadequate to identify Transcription Start Sites (TSS), in particular in ascidians, which make frequent use of trans-splicing (10). Re-

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D718/4628129 by BIU Montpellier user on 18 December 2019



Figure 3. Simultaneous visualization of candidate transcription factor binding sites in two known regulatory sequences. The WashU epigenome browser allows splitting the display to show two independent loci from the same genome. Here the display shows regulatory sequences of the Cirobu *Otx* (Cirobu.g00006940; left part of the window) and *Nodal* (Cirobu.g00010576; right part of the window) genes, and predicted local *in silico* binding affinity for transcription factors of the ETS and GATA families, known to regulate these enhancers.

cently, high-throughput RNA-sequencing techniques were developed to produce precise and quantitative genome-wide mapping of TSS, thereby allowing a better analysis of promoters (48). The *Ciona robusta* WashU browser of ANISEED 2017 includes a public hub with TSS-seq mapping for a set of adult tissues, and whole larvae (27).

ChIP-seq

In addition to the *Ciona robusta* ChIP-chip data for 11 transcription factors previously displayed on the *Ciona robusta* genome browser (49), ChIP-seq data for three additional transcription factors at the 32-cell stage (23) are now accessible as public hubs in the *C. robusta* WashU browser. ChIP-seq for the promoter-specific H3K4me3 histone mark in *C. robusta* and *P. mammillata* are also provided to help, combined with TSS-seq data, with the identification of basal promoters.

SELEX-seq

Gene expression is controlled at the transcriptional level by non-coding *cis*-regulatory sequences that act as binding platforms for transcription factors, which recognize specific DNA sequences. The knowledge of this binding specificity is a major help in deciphering the gene regulatory networks that underlie biological function. ANISEED 2017 describes the *in vitro* DNA-binding specificity of 131 of the estimated 500 transcription factors encoded in the *Ciona robusta* genome. These DNA-binding specificities, determined by SELEX-seq (50), are accessible via a specific tab in the gene card, which describes the details of the enrichment procedure and DNA-binding specificity and the relatedness of DNA-binding specificity to mammalian orthologs (50,51). 6-mer enrichments values and Position Weight Matrices can also be downloaded from this page. In addition, public hubs on the *Ciona robusta* and *Phallusia mammillata* WashU browsers group tracks of local predictions of binding affinity for each of these 131 *C. robusta* transcription factors and their 79 *P. mammillata* orthologs (Figure 3).

HOSTING OF PRIVATE NEW TUNICATE GENOMES

ANISEED offers to host on its genome browsers and blast servers, password-protected, private genomes sequenced by members of the community. Two genome projects are currently hosted and kept private until their publication in peer reviewed journals: the colonial stolidobranch ascidian *Botrylloides leachii* and the solitary phlebobranch *Corella inflata*.

WEBSERVICES AND DOWNLOADS

To facilitate the programmatic extraction of datasets and the development of external modules relying on ANISEED data, we developed an Application Programming Interface (API), accessible from the main menu. Currently, the main functions allow retrieving genes according to their expression patterns, articles and authors, *cis*-regulatory regions and *cis*-regulatory constructs and cDNA clones.

The download section was enriched with standardized GAF files with functional gene annotations for all species. Two functional annotations are proposed for each species, one based on the full Gene Ontology, the other on the tunicate GOslim we developed, which can also be downloaded. The gene trees are also available (.tre format). Finally, 6-mer enrichments and Position Weight Matrices for all rounds for

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D718/4628129 by BIU Montpellier user on 18 December 2019

131 SELEXed *Ciona robusta* transcription factors are accessible.

FUTURE PESPECTIVES

ANISEED 2017 marks a substantial improvement in the content and organization of the main tunicate model organism database. The diversity and depth of the added data attests of the dynamism of this small scientific community of less than a hundred labs worlwide.

One major evolution over the past two years was the discovery that the species initially called *Ciona intestinalis* comprises at least two independent species, *Ciona robusta* (formerly called *Ciona intestinalis* type A and found in all major oceans) and *Ciona intestinalis* (formerly called *Ciona intestinalis* type B, and mostly restricted to the North Atlantic) (28). In this version of ANISEED, we therefore reallocated all *Ciona intestinalis* data to *Ciona robusta*, which is now the most studied *Ciona* species. To facilitate interoperability and comparison between databases and to avoid confusion between species, it will be important for the *Ciona* community to also press for a reassignment of data to their correct species in generalists databases including NCBI, ENSEMBL, Uniprot or UCSC, which have so far not adapted to the new taxonomy.

We would also like to extend the database, beyond ascidians, to appendicularians and thaliaceans. While genomic information remains very partial in thaliaceans (52), high-quality genomic (53), transcriptomic (54) and functional (55) data exist for the appendicularian *Oikopleura dioica*, and we have established contact with the scientific community working on these animals. Computing accurate orthology relationships for most genes of these fast evolving taxa may, however, prove particularly challenging, and may require further optimization of our procedures.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the Si2C2 IT service of CRBM/IGMM/IRIM for support throughout the project, in particular Danielle Avinens and Patrice Langlois. We are grateful for the advice provided by the WashU Epigenome Browser team, in particular Daofeng Li, during the adaptation of this browser. This is contribution ISEM 2017-259 of the Institut des Sciences de l'Evolution de Montpellier (ISEM).

FUNDING

Agence Nationale de la Recherche (ANR) [Equipex Morphoscope2 ANR-11-EQPX-0029 to P.L.; Chor-Reg-Net NT05–2_42083 to P.L., C.C.; Chor-Evo-Net ANR-08-BLAN-0067-01 to P.L.; Institut de Biologie Computationnelle ANR-11-BINF-0002 to P.L.; Renabi-IFB, ANR-11-INBS-0013 to P.L.; TED, ANR-13-BSV2–0011-01 to P.L., E.D.; ANR-10-BINF-01-03, ANR-10-LABX–54 MEMO-LIFE and ANR-10-IDEX-0001-02 PSL* Research University to A.L.; ANR-16-CE92-0019 evobooster to J.N.V.]; Dopaminet European project (to P.L.); National Institutes of Health (NIGMS) [R01 GM096032 to L.C.]; Japan Society for the Promotion of Science (JSPS) KAKENHI [16H04724 to K.N.]; ANR and CNRS (to K.R.N.); ANR TED project (to P.S.); Marie Curie IIF *cis*-reg-logic (to E.J.); National Institutes of Health [R00 HD084814 to A.S.]; National Institutes of Health (NHGRI) [5U41HG002371-15 to M.H.]. Funding for open access charge: Institut de Biologie Computationnelle, Montpellier, France.
Conflict of interest statement. None declared.

REFERENCES

1. Lemaire,P. and Piette,J. (2015) Tunicates: exploring the sea shores and roaming the open ocean. A tribute to Thomas Huxley. *Open Biol.*, **5**, 150053.
2. Lemaire,P. (2011) Evolutionary crossroads in developmental biology: the tunicates. *Development*, **138**, 2143–2152.
3. 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.
4. Tsagkogeorga,G., Turon,X., Hopcroft,R.R., Tilak,M.-K., Feldstein,T., Shenkar,N., Loya,Y., Huchon,D., Douzery,E.J.P. and Delsuc,F. (2009) An updated 18S rRNA phylogeny of tunicates based on mixture and secondary structure models. *BMC Evol. Biol.*, **9**, 187.
5. Voskoboynik,A. and Weissman,I.L. (2015) Botryllus schlosseri, an emerging model for the study of aging, stem cells, and mechanisms of regeneration. *Invertebr. Reprod. Dev.*, **59**, 33–38.
6. Litman,G.W. and Dishaw,L.J. (2013) Histocompatibility: clarifying fusion confusion. *Curr. Biol.*, **23**, R934–R935.
7. Stolfi,A., Ryan,K., Meinertzhagen,I.A. and Christiaen,L. (2015) Migratory neuronal progenitors arise from the neural plate borders in tunicates. *Nature*, **527**, 371–374.
8. Stolfi,A., Gainous,T.B., Young,J.J., Mori,A., Levine,M. and Christiaen,L. (2010) Early chordate origins of the vertebrate second heart field. *Science*, **329**, 565–568.
9. Sobral,D., Tassy,O. and Lemaire,P. (2009) Highly divergent gene expression programs can lead to similar chordate larval body plans. *Curr. Biol.*, **19**, 2014–2019.
10. Matsumoto,J., Dewar,K., Wasserscheid,J., Wiley,G.B., Macmil,S.L., Roe,B.A., Zeller,R.W., Satou,Y. and Hastings,K.E.M. (2010) High-throughput sequence analysis of Ciona intestinalis SL trans-spliced mRNAs: alternative expression modes and gene function correlates. *Genome Res.*, **20**, 636–645.
11. Satou,Y., Mineta,K., Ogasawara,M., Sasakura,Y., Shoguchi,E., Ueno,K., Yamada,L., Matsumoto,J., Wasserscheid,J., Dewar,K. *et al.* (2008) Improved genome assembly and evidence-based global gene model set for the chordate Ciona intestinalis: new insight into intron and operon populations. *Genome Biol*, **9**, R152.
12. Aldred,N. and Clare,A.S. (2014) Mini-review: Impact and dynamics of surface fouling by solitary and compound ascidians. *Biofouling*, **30**, 259–270.
13. Fitridge,I., Dempster,T., Guenther,J. and de Nys,R. (2012) The impact and control of biofouling in marine aquaculture: a review. *Biofouling*, **28**, 649–669.
14. Stolfi,A. and Christiaen,L. (2012) Genetic and Genomic Toolbox of the Chordate Ciona intestinalis. *Genetics*, **192**, 55–66.
15. Dehal,P., Satou,Y., Campbell,R.K., Chapman,J., Degnan,B., De Tomaso,A., Davidson,B., Di Gregorio,A., Gelpke,M., Goodstein,D.M. *et al.* (2002) The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science*, **298**, 2157–2167.
16. Roure,A., Lemaire,P. and Darras,S. (2014) An otx/nodal regulatory signature for posterior neural development in ascidians. *PLoS Genet.*, **10**, e1004548.
17. Stolfi,A., Lowe,E.K., Racioppi,C., Ristoratore,F., Brown,C.T., Swalla,B.J. and Christiaen,L. (2014) Divergent mechanisms regulate conserved cardiopharyngeal development and gene expression in distantly related ascidians. *Elife*, **3**, e03728.
18. Treen,N., Yoshida,K., Sakuma,T., Sasaki,H., Kawai,N., Yamamoto,T. and Sasakura,Y. (2014) Tissue-specific and ubiquitous gene knockouts by TALEN electroporation provide new approaches to investigating gene function in Ciona. *Development*, **141**, 481–487.

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D718/4628129 by BIU Montpellier user on 18 December 2019

19. Stolfi,A., Gandhi,S., Salek,F. and Christiaen,L. (2014) Tissue-specific genome editing in Ciona embryos by CRISPR/Cas9. *Development*, **141**, 4115–4120.
20. Sasaki,H., Yoshida,K., Hozumi,A. and Sasakura,Y. (2014) CRISPR/Cas9-mediated gene knockout in the ascidian Ciona intestinalis. *Dev. Growth Differ.*, **56**, 499–510.
21. Tassy,O., Dauga,D., Daian,F., Sobral,D., Robin,F., Khoueiry,P., Salgado,D., Fox,V., Caillol,D., Schiappa,R. *et al.* (2010) The ANISEED database: digital representation, formalization, and elucidation of a chordate developmental program. *Genome Res.*, **20**, 1459–1468.
22. Brozovic,M., Martin,C., Dantec,C., Dauga,D., Mendez,M., Simion,P., Percher,M., Laporte,B., Scornavacca,C., Di Gregorio,A. *et al.* (2016) ANISEED 2015: a digital framework for the comparative developmental biology of ascidians. *Nucleic Acids Res.*, **44**, D808–D818.
23. Oda-Ishii,I., Kubo,A., Kari,W., Suzuki,N., Rothbächer,U. and Satou,Y. (2016) A Maternal System Initiating the Zygotic Developmental Program through Combinatorial Repression in the Ascidian Embryo. *PLoS Genet.*, **12**, e1006045.
24. Tokuhiro,S.-I., Tokuoka,M., Kobayashi,K., Kubo,A., Oda-Ishii,I. and Satou,Y. (2017) Differential gene expression along the animal-vegetal axis in the ascidian embryo is maintained by a dual functional protein Foxd. *PLoS Genet.*, **13**, e1006741.
25. Racioppi,C., Kamal,A.K., Razy-Krajka,F., Gambardella,G., Zanetti,L., di Bernardo,D., Sanges,R., Christiaen,L.A. and Ristoratore,F. (2014) Fibroblast growth factor signalling controls nervous system patterning and pigment cell formation in Ciona intestinalis. *Nat. Commun.*, **5**, 4830.
26. Suzuki,M.M., Mori,T. and Satoh,N. (2016) The Ciona intestinalis cleavage clock is independent of DNA methylation. *Genomics*, **108**, 168–176.
27. Yokomori,R., Shimai,K., Nishitsuji,K., Suzuki,Y., Kusakabe,T.G. and Nakai,K. (2016) Genome-wide identification and characterization of transcription start sites and promoters in the tunicate Ciona intestinalis. *Genome Res.*, **26**, 140–150.
28. Brunetti,R., Gissi,C., Pennati,R., Caicci,F., Gasparini,F. and Manni,L. (2015) Morphological evidence that the molecularly determined Ciona intestinalis type A and type B are different species: Ciona robusta and Ciona intestinalis. *J. Zoolog. Syst. Evol. Res.*, **53**, 186–193.
29. Sierro,N., Kusakabe,T., Park,K.-J., Yamashita,R., Kinoshita,K. and Nakai,K. (2006) DBTGR: a database of tunicate promoters and their regulatory elements. *Nucleic Acids Res.*, **34**, D552–D555.
30. Endo,T., Ueno,K., Yonezawa,K., Mineta,K., Hotta,K., Satou,Y., Yamada,L., Ogasawara,M., Takahashi,H., Nakajima,A. *et al.* (2010) CIPRO 2.5: Ciona intestinalis protein database, a unique integrated repository of large-scale omics data, bioinformatic analyses and curated annotation, with user rating and reviewing functionality. *Nucleic Acids Res.*, **39**, D807–D814.
31. Satou,Y., Kawashima,T., Shoguchi,E., Nakayama,A. and Satoh,N. (2005) An integrated database of the ascidian, Ciona intestinalis: towards functional genomics. *Zool. Sci.*, **22**, 837–843.
32. Hotta,K., Mitsuhara,K., Takahashi,H., Inaba,K., Oka,K., Gojobori,T. and Ikeo,K. (2007) A web-based interactive developmental table for the ascidian Ciona intestinalis, including 3D real-image embryo reconstructions: I. From fertilized egg to hatching larva. *Dev. Dyn.*, **236**, 1790–1805.
33. Haeussler,M., Schönig,K., Eckert,H., Eschstruth,A., Mianné,J., Renaud,J.-B., Schneider-Maunoury,S., Shkumatava,A., Teboul,L., Kent,J. *et al.* (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.*, **17**, 148.
34. Costello,M.J., Bouchet,P., Boxshall,G., Fauchald,K., Gordon,D., Hoeksema,B.W., Poore,G.C.B., Soest,R.W.M. van, Stöhr,S., Walter,T.C. *et al.* (2013) Global Coordination and Standardisation in Marine Biodiversity through the World Register of Marine Species (WoRMS) and Related Databases. *PLOS ONE*, **8**, e51629.
35. Dooley,E.E. (2002) GBIF: The Global Biodiversity Information Facility. *Environ. Health Perspect.*, **110**, A669.
36. Li,L., Stoeckert,C.J. and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
37. Miele,V., Penel,S. and Duret,L. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, **12**, 116.
38. Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
39. Stein,L.D. (2013) Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief. Bioinformatics*, **14**, 162–171.
40. Zhou,X., Maricque,B., Xie,M., Li,D., Sundaram,V., Martin,E.A., Koebbe,B.C., Nielsen,C., Hirst,M., Farnham,P. *et al.* (2011) The human epigenome browser at Washington University. *Nat Meth*, **8**, 989–990.
41. Zhou,X., Lowdon,R.F., Li,D., Lawson,H.A., Madden,P.A.F., Costello,J.F. and Wang,T. (2013) Exploring long-range genome interactions using the WashU Epigenome Browser. *Nat. Methods*, **10**, 375–376.
42. Nguyen,N.T.T., Vincens,P., Muffato,M., Roest Crollius,H. and Louis,A. (2017) Genomicus 2018: karyotype evolutionary trees and on-the-fly synteny computing. *Nucleic Acids Res.*, doi:10.1093/nar/gkx1003.
43. Stolfi,A., Sasakura,Y., Chalopin,D., Satou,Y., Christiaen,L., Dantec,C., Endo,T., Naville,M., Nishida,H., Swalla,B.J. *et al.* (2015) Guidelines for the nomenclature of genetic elements in tunicate genomes. *Genesis*, **53**, 1–14.
44. Gene Ontology Consortium (2015) *Nucleic Acids Res.*, **43**, D1049–D1056.
45. Yon Rhee,S., Wood,V., Dolinski,K. and Draghici,S. (2008) Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.*, **9**, 509–515.
46. Davis,M.J., Sehgal,M.S.B. and Ragan,M.A. (2010) Automatic, context-specific generation of Gene Ontology slims. *BMC Bioinformatics*, **11**, 498.
47. Conesa,A., Madrigal,P., Tarazona,S., Gomez-Cabrero,D., Cervera,A., McPherson,A., Szcześniak,M.W., Gaffney,D.J., Elo,L.L., Zhang,X. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13.
48. Haberle,V. and Lenhard,B. (2016) Promoter architectures and developmental gene regulation. *Semin. Cell Dev. Biol.*, **57**, 11–23.
49. Kubo,A., Suzuki,N., Yuan,X., Nakai,K., Satoh,N., Imai,K.S. and Satou,Y. (2010) Genomic cis-regulatory networks in the early Ciona intestinalis embryo. *Development*, **137**, 1613–1623.
50. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpää,M.J. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.
51. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
52. Jue,N.K., Batta-Lona,P.G., Trusiak,S., Obergfell,C., Bucklin,A., O'Neill,M.J. and O'Neill,R.J. (2016) Rapid evolutionary rates and unique genomic signatures discovered in the first reference genome for the southern ocean Salp, Salpa thompsoni (Urochordata, Thaliacea). *Genome Biol. Evol.*, **8**, 3171–3186.
53. 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.
54. Wang,K., Omotezako,T., Kishi,K., Nishida,H. and Onuma,T.A. (2015) Maternal and zygotic transcriptomes in the appendicularian, Oikopleura dioica: novel protein-encoding genes, intra-species sequence variations, and trans-spliced RNA leader. *Dev. Genes Evol.*, **225**, 149–159.
55. Omotezako,T., Nishino,A., Onuma,T.A. and Nishida,H. (2013) RNA interference in the appendicularian Oikopleura dioica reveals the function of the Brachyury gene. *Dev. Genes Evol.*, **223**, 261–267.

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D718/4628129 by BIU Montpellier user on 18 December 2019