Genomic resources and annotations for a colonial ascidian, the light-bulb sea squirt Clavelina lepadiformis

Vladimir Daric, Maxence Lanoizelet, Hélène Mayeur, Cécile Leblond, Sébastien Darras

▶ **To cite this version:**

Vladimir Daric, Maxence Lanoizelet, Hélène Mayeur, Cécile Leblond, Sébastien Darras. Genomic resources and annotations for a colonial ascidian, the light-bulb sea squirt Clavelina lepadiformis. Genome Biology and Evolution, 2024, 10.1093/gbe/evae038 . hal-04511455

HAL Id: hal-04511455
https://hal.sorbonne-universite.fr/hal-04511455v1

Submitted on 19 Mar 2024

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.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Genomic resources and annotations for a colonial ascidian, the light-bulb sea squirt *Clavelina lepadiformis*

Vladimir DARIC, Maxence LANOIZELET, Hélène MAYEUR, Cécile LEBLOND and Sébastien DARRAS#

Sorbonne Université, CNRS, Biologie Intégrative des Organismes Marins (BIOM), F-66650, Banyuls/Mer, France

#: author for correspondence (sebastien.darras@obs-banyuls.fr)

Abstract

Ascidian embryos have been studied since the birth of experimental embryology at the end of the 19th century. They represent textbook examples of mosaic development characterized by a fast development with very few cells and invariant cleavage patterns and lineages. Ascidians belong to tunicates, the vertebrate sister group, and their study is essential to shed light on the emergence of vertebrates. Importantly, deciphering developmental gene regulatory networks has been carried out mostly in two of the three ascidian orders, Phlebobranchia and Stolidobranchia. To infer ancestral developmental programs in ascidians, it is thus essential to carry out molecular embryology in the third ascidian order, the Aplousobranchia. Here, we present genomic resources for the colonial aplousobranch *Clavelina lepadiformis*: a transcriptome produced from various embryonic stages, and an annotated genome. The assembly consists of 184 contigs making a total of 233.6 Mb with a N50 of 8.5 Mb and a L50 of 11. The 32,318 predicted genes capture 96.3% of BUSCO orthologs. We further show that these

© The Author(s) 2024. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://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

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

resources are suitable to study developmental gene expression and regulation in a comparative framework within ascidians. Additionally, they will prove valuable for evolutionary and ecological studies.

Key words: *Clavelina lepadiformis*, colonial ascidian, Aplousobranchia, tunicate, genome, transcriptome, evo-devo

Significance

Clavelina lepadiformis belongs to Aplousobranchia, one of the three ascidian orders, that includes only colonial animals and that has been under-explored at the molecular level. This species is a promising model for developmental, evolutionary and ecological studies. We present a transcriptome and an annotated genome, and show how these resources are immediately useful for comparative analysis of embryonic development in ascidians.

Introduction

Ascidians belong to the tunicates, the vertebrate sister group. These marine filter-feeding animals share with vertebrates and cephalochordates (amphioxus) a typical chordate body plan during embryonic life (most prominently visible by the presence of a notochord and a dorsal neural tube). Ascidians have a simple and stereotyped invariant embryonic development with very few cells (100 at gastrulation and 2,500 in the tadpole larva) allowing deciphering developmental mechanisms at cellular resolution. In addition, these externally developing embryos are easily amenable to experimentation, and are particularly well suited for functional

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

genomics (Satoh 2014; Lemaire 2011). These classical features of ascidians actually correspond to a few species that have been used as laboratory animals. Since the advent of molecular approaches, *Ciona* (represented by two closely related species: *C. robusta* (or *C. intestinalis* type A) from the Pacific Ocean and the Mediterranean Sea, and *C. intestinalis* (or *C. intestinalis* type B) from the Atlantic Ocean) has been the best studied and became the reference organism. In recent years, the progress of sequencing technologies and the generally small size of ascidian genomes has led to whole genome sequencing for a number of species (Dardaillon et al. 2020). Our current understanding is that ascidian genomes have been extensively rearranged in this fast-evolving lineage. Consequently, synteny is overall absent and DNA sequence conservation is limited to the coding parts of the genomes. This drastic divergence of the genomes appears to be contradictory to the fact that embryogenesis is remarkably conserved in distantly related species. This raises the question of whether the molecular control of embryonic development is the same in different species. Since ascidians have extensively diversified into around 3,000 species (Shenkar & Swalla 2011), they offer a great opportunity to evaluate the evolution of developmental mechanisms. Current ascidian phylogenetics support a traditional classification into 3 orders: the Phlebobranchia, the Aplousobranchia and the Stolidobranchia (Delsuc et al. 2018; Kocot et al. 2018). Developmental biology research has largely focused on the embryonic and non-embryonic (asexual reproduction and regeneration) development in Phlebobranchia and Stolidobranchia. Aplousobranchia make a group of strictly colonial ascidians that have been so far overlooked. The few studies dedicated to their development are limited to rather ancient and descriptive literature. Some genomic resources are available for three species: a very fragmented but annotated assembly for *Didemnum vexillum* (Parra-Rincón et al. 2021; Velandia-

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

Huerto et al. 2016), and non-annotated chromosome level assemblies for *Aplidium turbinatum* (Bishop et al. 2022) and *Clavelina lepadiformis* (https://www.ebi.ac.uk/ena/browser/view/PRJEB57668). RNA-seq data are also available for a specific period in the life cycle of the latter species, the dormancy (Hiebert et al. 2022). Here, we describe complementary transcriptomic and genomic resources for *C. lepadiformis*, and show that they are immediately valuable for comparative developmental biology of ascidians.

Results and discussion

Genome assembly

The Cvlepa_BANY2021 assembly of the PacBio reads led to an estimated genome size of 233.6 Mb allocated to 184 contigs, the largest one being 15.4 Mb (Table 1). The completeness of the genome seems quite high with a BUSCO score of 96.3%, and a mapping of 95.4% of the RNA-seq reads that we have generated (Table S1). Interestingly, 90% of the genome was found in the 30 longest contigs. This number is higher than the anticipated number of chromosomes taking other ascidian species as references (14 chromosomes in *Ciona robusta* and *Ciona intestinalis* (Satou et al. 2021), and 16 in *Styela clava* and *Aplidium turbinatum* (Zhang et al. 2021; Bishop et al. 2022)), or the karyotype that indicated 9 chromosomes (Colombera 1971). At the time we planned to reach chromosome level assembly using HiC, a 210.1 Mb assembly for *C. lepadiformis* that includes 9 chromosomes was released at ENA (kaClaLepa1.1 assembly. https://www.ebi.ac.uk/ena/browser/view/PRJEB58329). The sequenced animal was collected from the Atlantic Ocean while we sequenced an animal from the Mediterranean Sea. We thus

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

wondered how close these two samples were. We first extracted *Cox1* sequences and compared them to available sequences for *Clavelina* species (Fig S1). As previously demonstrated, there are two *C. lepadiformis* clades suggestive of separate species, Atlantic *vs* Mediterranean (Turon et al. 2003). Our sample that was collected in a harbor falls in the Atlantic clade as other similar Mediterranean samples. This likely corresponds to a recent colonization of Mediterranean marinas by the Atlantic type (Turon et al. 2003). Furthermore, the alignment of the 14,478 bp mitogenomes from the kaClaLepa1.1 and the Cvlepa_BANY2021 assemblies displayed an identity of 99.7%, confirming that the two specimen belong to the same species (Table S2).

The overall comparison of the two assemblies indicated that the 9 chromosomes of the kaClaLepa1.1 assembly match to 30 scaffolds of the Cvlepa_BANY2021 assembly, with 2 to 6 scaffolds for 1 chromosome (Fig S2, Table S2). This fraction of the genomes corresponded to approximately 206 Mb and captured 98.6% of the RNA-seq reads that map to our assembly (Table S2,S3). They could be aligned over 173 Mb with an average identity of 67.3%. The parts that did not align correspond most likely to centromeric regions: they are central in the chromosomes, and have a low gene density and a low RNA-seq coverage (Table S2,S3,S4). The remaining fraction of the assemblies were 3 Mb (kaClaLepa1.1) and 28 Mb (Cvlepa_BANY2021), and did not match to each other. The rather large size of this fraction in our assembly suggests that it might be artifactual, may correspond to repeated regions or unresolved haplotypes. Accordingly, RNA-seq coverage was overall very low with the exception of a few scaffolds. In conclusion, for the Atlantic type of *C. lepadiformis*, genome assemblies from two individuals (one from the Mediterranean sea and one from the Atlantic ocean) are now available and should be useful for studying recent genome evolution and adaptation to different

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

environments. Such a line of research would be beneficially complemented by sequencing the genome of *C. lepadiformis* of endogenous Mediterranean type.

Transcriptome assembly and genome annotation

With a focus on embryonic development, we performed RNA-seq using Illumina on three classes of embryonic stages (egg to neurula, tailbud stages, and larval stages) (Fig 1). To complement this data set, we also sequenced adult pharyngeal tissue. The DRAP-assembled transcriptome contains 31,035 transcripts of which 22,048 are coding (a large fraction of them, 87.4% have a hit against *C. robusta* proteome). The completeness of this dataset is lower (93% BUSCO score) than the genome, probably owing to the non-extensive sampling of tissues/life cycle.

The genome annotation indicates 32,318 genes, 25,067 of which are coding for proteins. Surprisingly, only 62.2% of these coding genes had a hit on Swissprot using blastp (Table 1, Table S5). This percentage increased to 79.4% when considering the proteome of the best annotated ascidian genome *Ciona robusta* (Satou et al. 2022). Yet, almost 20% of the predicted genes had no equivalent, suggesting that the prediction was partly inaccurate, and/or some novel genes have to be found in the aplousobranch ascidian lineage. However, similar numbers were found when the latest version of the *C. robusta* proteome was analyzed in the same manner (Table S5).

Applications of genomic resources for comparative developmental biology

Comparisons of developmental gene expression patterns

Ciona is the reference ascidian species for developmental studies with extensive information on gene expression and gene function during embryogenesis that are accessible in dedicated

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

databases such as Aniseed and Ghost (Dardaillon et al. 2020; Satou et al. 2005). To compare the expression of developmental regulators, we performed *in situ* hybridization for the transcription factor coding genes, *Dmrta*, *Foxn1/4*, *Klf1/2/4/17*, *Isl*, and *Sp6/7/8/9*, and for the neural marker *Celf3/4/5/6* (Fig S3A). For each gene, phylogenetic analysis indicated the presence of a single ortholog in *Clavelina* (File S1). These genes displayed similar patterns as their *Ciona* counterparts: early neural precursors in gastrulae for *Dmrta*, tail epidermis midlines for *Klf1/2/4/17*, adhesive papillae for *Isl* and *Sp6/7/8/9*, and neural cells for *Celf3/4/5/6*. *Foxn1/4* expression has not been described in *Ciona*. It seemed to be expressed in a pattern reminiscent of *Ciona* presumptive germ cells.

Testing cis-regulatory activity

Transcriptional regulation is an essential aspect of developmental gene networks. *In vivo* evaluation of transcriptional activity is readily feasible in different ascidian species through plasmid DNA introduction in the fertilized egg by electroporation (Lemaire 2011; Darras 2021; Coulcher et al. 2020). Since *Clavelina* embryos are not yet amenable to experimentation, we aimed at testing a candidate regulatory region in *Ciona* embryos. We focused on the *CesA* gene that codes for a cellulose synthase, a gene acquired by horizontal transfer that gives tunicates their unique capacity of synthesizing cellulose (Matthysse et al. 2004; Sasakura et al. 2005). In *Ciona*, this gene is expressed in the entire epidermis under the control of the transcription factor Tfap2-r.b, a determinant of epidermal fate (Sasakura et al. 2016). In *Clavelina*, we identified a single *CesA* gene harboring 3 putative Tfap2 binding sites within 1 kb upstream of the start codon (Fig S3B). When this region was placed upstream of a minimal promoter and *LacZ* as a

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

reporter, and tested in *Ciona intestinalis*, it showed activity in the epidermis (32% of the embryos, n=820, results from 3 independent experiments). We obtained similar results when the construct was tested in another phlebobranch ascidian species, *Phallusia mammillata* (29% of the embryos with activity in the epidermis, n=558, results from 3 independent experiments). These results strongly suggests that *CesA* has a conserved expression and regulation between *Clavelina* and Phlebobranchia. Furthermore, it constitutes a proof of concept for the study of *cis*-regulatory elements of *Clavelina*.

Conclusion

The resources that we have presented in this study will be directly beneficial for comparative analyses of embryonic development in ascidians in order to study the evolution of developmental mechanisms. They will complement other resources that are being generated for a number of tunicate species. In addition, they will be very useful in other fields, such as the study of non-embryonic development (Alié et al. 2020), adaptation and evolution.

Material and methods

Sample collection, nucleic acid extraction and sequencing

Colonies of *Clavelina lepadiformis* were collected on ropes in the harbor of Saint-Cyprien, France (42°36'56.2"N 3°02'12.0"E) (Fig 1B). The gonad of a single sexually mature adult was dissected for genomic DNA extraction (Fig 1C,E). The gonad was dissociated in sea water with the help of a disposable plastic pestle, and by using a combination of pipetting and vortexing. The cells were washed and collected by repeated centrifugation (6000 rpm at 4°C for 2 min) and resuspension

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

in sea water. High molecular weight genomic DNA was extracted using the Monarch HMW DNA Extraction Kit for Cells & Blood (T3050, New England Biolabs) following manufacturer's protocol. DNA quantity and quality was evaluated by agarose gel electrophoresis, spectrophotometry (Nanodrop, Thermo Fisher Scientific) and fluorometry (Quantus, Promega). Genome sequencing was performed by the GENTYANE platform (INRAe, Clermont-Ferrand) using PacBio Sequel II. Circular consensus sequencing (CCS) protocol was used in order to obtain highly accurate long read sequences that can be accessed through the BioProject PRJEB64590.

C. lepadiformis is a viviparous ascidian, and embryos at staggered stages can be found in the atrial cavity of mature zooids (Fig 1C,D,F). Embryos from several zooids were released from the adults by dissection, and separated into 4 samples according to stages: early (egg to neurula stages, from 10 zooids), intermediate stages (initial tailbud to late tailbud stages, from 10 zooids), and late stages (late tailbud to larva stages, from 4 zooids). These 3 samples and an adult sample (pharynx from 4 zooids) were immediately flash frozen using liquid nitrogen. Total RNA was extracted using the NucleoSpin Tissue kit (Macherey-Nagel) following the provided classical tissue protocol, except for the early embryo samples where the modifications from the 'difficult-to-lyse tissue' (Rev. 01) were applied (the classical protocol was fully inefficient on such types of samples). RNA quality and concentration was determined using a Bioanalyzer 2100 (Agilent Technologies). The RNAs had a RIN>9.8 and were sequenced using the Illumina technology (paired ends 2x150 bp on NextSeq550) by the BioEnvironnement facility (UPVD, Perpignan). 39 to 41 million reads were produced for each of the 4 samples, and can be accessed through the BioProject PRJEB64590.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

De novo transcriptome and genome assemblies

We used the RNA-seq data to perform a *de novo* transcriptome assembly using the DRAP pipeline (v1.91) (Cabau et al. 2017), with Oases as an assembler with kmers 37, 47, 57 and 63. PacBio sequences were assembled with hifiasm assembler (v0.15.4-r342) with default parameters (Cheng et al. 2021, 2022).

The mitogenome was assembled and annotated with MitoHiFi (v 3.0.0q 1.4.1) (Uliano-Silva et al. 2023; Allio et al. 2020) using *Ciona robusta* mitogenome as reference. This 14,478 bp mitogenome was used in the final assembly to replace a duplicated version (scaffold S176) that was present in the initial hifiasm assembly.

RNA-seq reads were mapped to the genome using STAR (2.7.5a) (Dobin et al. 2013).

Annotation

The genome was annotated with the funannotate pipeline (v1.8.14) (https://github.com/nextgenusfs/funannotate). As a first step, 17 repetitive contigs were removed from the primary hifiasm assembly using the funnannotate *clean* script. The resulting assembly is the current version of the genome that we named Cvlepa_BANY2021 and that we deposited at ENA (BioProject PRJEB64590). The annotation process involved three main stages: (1) funnannotate *train* script (–max_intronlen 8000 –busco_db metazoa) was used to perform a *de novo* genome-guided transcriptome assembly for RNA-seq data with HISAT (v2.2.1) (Kim et al. 2019), Trinity (v2.8.5) (Grabherr et al. 2011), StringTie (v.2.2.1) (Shumate et al. 2022), PASA (v2.4.1) (Haas et al. 2003) and Kallisto (v0.46.1) (Bray et al. 2016) to identify the best probable transcript at each locus. (2) Gene prediction was performed with funnannotate *predict* script (--

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

organism other --repeats2evm --busco_db metazoa --ploidy 2 --optimize_augustus). This script uses Evidence Modeler (v1.1.1) to select consensus gene models from Augustus (v3.3.2), GlimmerHMM (v3.0.4), snap (v2006-07-28) predictions. The prediction yielded a total of 32,318 genes, among which 25,067 are protein-coding genes, exhibiting an average gene length of 3,188 bp. (3) Finally the funnannotate *annotate* script was used to assign functional annotation to the protein coding gene models using evidence from InterProScan5 (v5.52-86) and UniProt DB (v2022_05) databases. This procedure yielded 19,055 InterPro annotations, 16,229 PFAM annotations, 13,862 GO terms, and 1,012 MEROPS annotations (Table S6). The quality of the assembly was assessed with Quast-LG (v5.0.2) (Mikheenko et al. 2018) and BUSCO (v5.1.2) (Manni et al. 2021) with the metazoan lineage orthologs dataset. Genetic elements were named following the nomenclature of the tunicate community (Stolfi et al. 2015).

Repeated regions were identified using RepeatModeler (v2.0.4) (Flynn et al. 2020) and RepeatMasker (v4.1.5) (https://www.repeatmasker.org/RepeatMasker/). They correspond to 43.9% of the assembly and their description is available in File S2.

Sequence analyses and phylogenies

Whole genome assemblies were aligned using minimap2 through the D-genies web interface (https://dgenies.toulouse.inra.fr/) (Cabanettes & Klopp 2018)

Predicted proteins from the genome annotation were compared with proteins from *Ciona robusta* (KY21) (Satou et al. 2022), *Branchiostoma lanceolatum* (BraLan3, NCBI) (Brasó-Vives et al. 2022), *Corella inflata* (DeBiasse et al. 2020), *Styela clava* (ASM1312258v2, NCBI) (Wei et al. 2020), *Homo sapiens* (GRCh38.p13, NCBI) and Swiss-Prot dataset (Release 2023_03) (The

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

UniProt Consortium 2023) using blastp (v2.11.0), or with the whole genome assemblies of *Aplidium turbinatum* (kaAplTurb1.1) (Bishop et al. 2022) using tblastn (hits were considered positive for evalue <5.10^{-5}). Similarly, coding transcripts from the DRAP transcriptome assembly were assessed against *C. robusta* and Swiss-Prot datasets using blastx (hits were considered positive for evalue <5.10^{-5}).

We used blastp against whole proteomes to recover sequences of potential orthologs from *Homo sapiens* (GRCh38.p13, NCBI), *Danio rerio* (GRCz11, NCBI), *Xenopus tropicalis* (UCB_Xtro_10.0, NCBI), *Scyliorhinus canicula* (sScyCan1.1, NCBI), *Callorhinchus milii* (IMCB_Cmil_1.0, NCBI), *Ciona robusta* (KH, NCBI), *Phallusia mammillata* (MTP2014, Aniseed; and nr from NCBI), *Halocynthia roretzi* (MTP2014, Aniseed), *Styela clava* (ASM1312258v2, NCBI), *Molgula occidentalis* (ELv1-2, Aniseed), *Branchiostoma lanceolatum* (BraLan3, NCBI), *Branchiostoma belcheri* (Haploidv18h27, NCBI), *Strongylocentrotus purpuratus* (Spur_5.0, NCBI) and *Saccoglossus kowalevskii* (Skow_1.1, NCBI). All sequences were aligned with the MUSCLE program using the EMBL-EBI website (Edgar 2004; Madeira et al. 2022). Maximum-likelihood phylogenies were inferred using IQ-TREE (http://iqtree.cibiv.univie.ac.at/) (Trifinopoulos et al. 2016).

A similar approach was used for *C. lepadiformis* and *C. oblongata* partial *Cox1* sequences that were retrieved from NCBI, together with the ones extracted from the kaClaLepa1.1 and Cvlepa_BANY2021 genomic assemblies.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

In situ hybridization

We used the same method that prove efficient in several ascidian species (*Ciona intestinalis*, *Phallusia mammillata*, *Molgula appendiculata* and *Halocynthia roretzi*) (Coulcher et al. 2020). The main modification was as follows. Embryos were released from the adult by dissection with scissors and tweezers. The protective chorion that surrounds the embryos was removed by chemical digestion with 0.1% trypsin in sea water (15 to 30 min depending on the stage). Antisense digoxigenin-labeled RNA probes were synthesized from plasmids (RT-PCR-amplification and cloning in pGEM-T) or synthetic double-stranded DNA (eBlock, Integrated DNA Technologies, Leuven, Belgium) (Table S7) as described previously (Chowdhury et al. 2022).

Cvlepa.CesA locus analysis and transcriptional assay

Cvlepa.CesA was identified from our data by blast using various tunicate CesA protein sequences as queries. Actually, 3 neighboring gene models, Cvlepa.CG.BANY2021.S15.g019272, Cvlepa.CG.BANY2021.S15.g019273 and Cvlepa.CG.BANY2021.S15.g019274, represented significant hits. But only Cvlepa.CG.BANY2021.S15.g019272 coded for the 2 domains GT2 and GH6 present in tunicate CesA, and was further considered (File S1). We examined the local synteny of this locus (10 genes on each side of *CesA*). As expected from the study of other tunicate genomes (Dardaillon et al. 2020), it was poorly preserved with a possible better synteny with *Aplidium turbinatum*, another aplousobranch ascidian, than with *Ciona robusta*, a phlebobranch ascidian (Fig S4). Putative Tfap2 binding sites were mapped on the *CesA* locus using FIMO (http://meme-suite.org/tools/fimo) (Grant et al. 2011) with matrices collected from the Jaspar database (Fornes et al. 2020) and the $GCCN_{3/4}GGC$ motif (Eckert et al. 2005). A 993 bp

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

fragment was amplified from genomic DNA using PCR with Fwd ACTTCCCAGCGGTACAGTCA and Rev TGTGACACGGTTCTTTCACCG, placed upstream of the *Ciinte*.*Fog* basal promoter and *LacZ* using the Gateway technology (Invitrogen) (Roure et al. 2007; Coulcher et al. 2020). Transcriptional assay was performed using *Ciona intestinalis* and *Phallusia mammillata* embryos as previously described (Coulcher et al. 2020; Darras 2021).

Conflict of interest.

The authors declare that they have no conflict of interest.

Acknowledgements

We thank the Bio2Mar facility (Sorbonne Université/CNRS, Banyuls-sur-mer) for access to molecular biology equipment (BioAnalyzer, Quantus), the Gentyane facility (INRAe, Clermont-Ferrand) for PacBio HiFi sequencing and the BioEnvironnement facility (UPVD, Perpignan) for Illumina RNA-seq. We are grateful to the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for providing help and/or computing and/or storage resources. We thank X. Turon for a helpful discussion on the situation of *C. lepadiformis* in the Mediterranean Sea. We also thank J. Dainat (@Juke34) for his helpful support for EMBLmyGFF3 (Norling et al. 2018) tool on gitHub.

Funding

VD, HM and SD are CNRS staff. This work was supported by CNRS and Sorbonne Université, and by specific grants from the CNRS (*AscidianDiversity* project funded by the DBM2020 call from

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

the INSB) and from Sorbonne Université (*AnimalCellulose* project within the Emergence 2021 framework).

Authors' contributions

VD and SD designed the project and wrote the manuscript. VD and HM performed the bioinformatic analysis work. ML, CL and SD performed the experimental work. SD obtained funding and supervised the project. All authors edited the manuscript, read and approved the final version.

Data availability

RNA-seq data, PacBio data and the genomic assembly Cvlepa_BANY2021 are available under the BioProject accession PRJEB64590. All other data generated or analyzed during this study are included in the manuscript and supporting files.

References

Alié A, Hiebert LS, Scelzo M, Tiozzo S. 2020. The eventful history of nonembryonic development in tunicates. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution. n/a. doi: 10.1002/jez.b.22940.

Allio R et al. 2020. MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Molecular Ecology Resources. 20:892–905. doi: 10.1111/1755-0998.13160.

Bishop J et al. 2022. The genome sequence of Aplidium turbinatum (Savigny 1816), a colonial sea squirt. Wellcome Open Res. 7:106. doi: 10.12688/wellcomeopenres.17785.1.

Brasó-Vives M et al. 2022. Parallel evolution of amphioxus and vertebrate small-scale gene

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

duplications. Genome Biol. 23:1–24. doi: 10.1186/s13059-022-02808-6.
Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 34:525–527. doi: 10.1038/nbt.3519.
Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ. 6:e4958. doi: 10.7717/peerj.4958.
Cabau C et al. 2017. Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies. PeerJ. 5:e2988. doi: 10.7717/peerj.2988.
Cheng H et al. 2022. Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol. 40:1332–1335. doi: 10.1038/s41587-022-01261-x.
Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 18:170–175. doi: 10.1038/s41592-020-01056-5.
Chowdhury R et al. 2022. Highly distinct genetic programs for peripheral nervous system formation in chordates. BMC Biol. 20:1–25. doi: 10.1186/s12915-022-01355-7.
Colombera D. 1971. The Karyology of Clavelina Lepadiformis Mueller (Ascidiacea). Caryologia. 24:59–64. doi: 10.1080/00087114.1971.10796413.
Coulcher JF et al. 2020. Conservation of peripheral nervous system formation mechanisms in divergent ascidian embryos Kuraku, S, editor. eLife. 9:e59157. doi: 10.7554/eLife.59157.
Dardaillon J et al. 2020. ANISEED 2019: 4D exploration of genetic data for an extended range of tunicates. Nucleic Acids Res. 48:D668–D675. doi: 10.1093/nar/gkz955.
Darras S. 2021. En masse DNA Electroporation for in vivo Transcriptional Assay in Ascidian Embryos. Bio-protocol. 11:e4160.
DeBiasse MB, Colgan WN, Harris L, Davidson B, Ryan JF. 2020. Inferring Tunicate Relationships and the Evolution of the Tunicate Hox Cluster with the Genome of Corella inflata. Genome Biol Evol. 12:948–964. doi: 10.1093/gbe/evaa060.
Delsuc F et al. 2018. A phylogenomic framework and timescale for comparative studies of tunicates. BMC Biology. 16:39. doi: 10.1186/s12915-018-0499-2.
Dobin A et al. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29:15–21. doi: 10.1093/bioinformatics/bts635.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

Eckert D, Buhl S, Weber S, Jäger R, Schorle H. 2005. The AP-2 family of transcription factors. Genome Biol. 6:1–8. doi: 10.1186/gb-2005-6-13-246.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 32:1792–1797. doi: 10.1093/nar/gkh340.

Flynn JM et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences. 117:9451–9457. doi: 10.1073/pnas.1921046117.

Fornes O et al. 2020. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 48:D87–D92. doi: 10.1093/nar/gkz1001.

Grabherr MG et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 29:644–652. doi: 10.1038/nbt.1883.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics. 27:1017–1018. doi: 10.1093/bioinformatics/btr064.

Haas BJ et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research. 31:5654–5666. doi: 10.1093/nar/gkg770.

Hiebert LS et al. 2022. Comparing dormancy in two distantly related tunicates reveals morphological, molecular, and ecological convergences and repeated co-option. Sci Rep. 12:12620. doi: 10.1038/s41598-022-16656-8.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 37:907–915. doi: 10.1038/s41587-019-0201-4.

Kocot KM, Tassia MG, Halanych KM, Swalla BJ. 2018. Phylogenomics offers resolution of major tunicate relationships. Molecular Phylogenetics and Evolution. 121:166–173. doi: 10.1016/j.ympev.2018.01.005.

Lemaire P. 2011. Evolutionary crossroads in developmental biology: the tunicates. Development. 138:2143–52. doi: 10.1242/dev.048975.

Madeira F et al. 2022. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research. 50:W276–W279. doi: 10.1093/nar/gkac240.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution. 38:4647–4654. doi: 10.1093/molbev/msab199.

Matthysse AG et al. 2004. A functional cellulose synthase from ascidian epidermis. Proc Natl Acad Sci U S A. 101:986–91.

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 34:i142–i150. doi: 10.1093/bioinformatics/bty266.

Norling M, Jareborg N, Dainat J. 2018. EMBLmyGFF3: a converter facilitating genome annotation submission to European Nucleotide Archive. BMC Res Notes. 11:1–5. doi: 10.1186/s13104-018-3686-x.

Parra-Rincón E et al. 2021. The Genome of the 'Sea Vomit' Didemnum vexillum. Life. 11:1377. doi: 10.3390/life11121377.

Roure A et al. 2007. A multicassette Gateway vector set for high throughput and comparative analyses in ciona and vertebrate embryos. PLoS ONE. 2:e916.

Sasakura Y et al. 2016. Transcriptional regulation of a horizontally transferred gene from bacterium to chordate. Proceedings of the Royal Society B: Biological Sciences. 283:20161712. doi: 10.1098/rspb.2016.1712.

Sasakura Y et al. 2005. Transposon-mediated insertional mutagenesis revealed the functions of animal cellulose synthase in the ascidian Ciona intestinalis. Proc Natl Acad Sci U S A. 102:15134–9.

Satoh N. 2014. *Developmental genomics of ascidians*. John Wiley & Sons, Inc: Hoboken, New Jersey.

Satou Y et al. 2022. A Manually Curated Gene Model Set for an Ascidian, Ciona robusta (Ciona intestinalis Type A). jzoo. 39. doi: 10.2108/zs210102.

Satou Y et al. 2021. Chromosomal Inversion Polymorphisms in Two Sympatric Ascidian Lineages. Genome Biology and Evolution. 13. doi: 10.1093/gbe/evab068.

Satou Y, Kawashima T, Shoguchi E, Nakayama A, Satoh N. 2005. An integrated database of the ascidian, Ciona intestinalis: towards functional genomics. Zoological science. 22:837–43.

Shenkar N, Swalla BJ. 2011. Global Diversity of Ascidiacea Browman, H, editor. PLoS ONE.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

6:e20657. doi: 10.1371/journal.pone.0020657.
Shumate A, Wong B, Pertea G, Pertea M. 2022. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLOS Computational Biology. 18:e1009730. doi: 10.1371/journal.pcbi.1009730.
Stolfi A et al. 2015. Guidelines for the nomenclature of genetic elements in tunicate genomes. genesis. 53:1–14. doi: 10.1002/dvg.22822.
The UniProt Consortium. 2023. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research. 51:D523–D531. doi: 10.1093/nar/gkac1052.
Trifinopoulos J, Nguyen L-T, von Haeseler A, Minh BQ. 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Research. 44:W232–W235. doi: 10.1093/nar/gkw256.
Turon X, Tarjuelo I, Duran S, Pascual M. 2003. Characterising invasion processes with genetic data: an Atlantic clade of Clavelina lepadiformis (Ascidiacea) introduced into Mediterranean harbours. Hydrobiologia. 503:29–35. doi: 10.1023/B:HYDR.0000008481.10705.c2.
Uliano-Silva M et al. 2023. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. BMC Bioinformatics. 24:1–13. doi: 10.1186/s12859-023-05385-y.
Velandia-Huerto CA, Gittenberger AA, Brown FD, Stadler PF, Bermúdez-Santana CI. 2016. Automated detection of ncRNAs in the draft genome sequence of a colonial tunicate: the carpet sea squirt Didemnum vexillum. BMC Genomics. 17:1–15. doi: 10.1186/s12864-016-2934-5.
Wei J et al. 2020. Genomic basis of environmental adaptation in the leathery sea squirt (Styela clava). Molecular Ecology Resources. 20:1414–1431. doi: 10.1111/1755-0998.13209.
Zhang J, Wei J, Yu H, Dong B. 2021. Genome-Wide Identification, Comparison, and Expression Analysis of Transcription Factors in Ascidian Styela clava. International Journal of Molecular Sciences. 22:4317. doi: 10.3390/ijms22094317.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

Table and figure legend

Figure 1. Overall strategy and results. (A) Simplified phylogenetic tree for ascidian species used in developmental biology with the three orders highlighted in colored boxes (based on (Delsuc et al. 2018)). Pictures of *Clavelina lepadiformis* biological samples: a colony (B), close up of a mature zooid (C) where the embryos developing in the pharynx are circled in red (D) and the gonad circled in black (E). Note that *C. lepadiformis* is hermaphordite, the gonad is thus made of an ovary and a testis. Embryos at the 2-cell, 4-cell, late tailbud and early larva stages (F). RNA from embryos and pharynx was sequenced using Illumina and assembled using *DRAP* (red panel). Genomic DNA was sequenced using PacBio HiFi and assembled with *hifiasm* (black panel). The assembled genome was annotated using RNA-seq data with *Funannotate* (blue panel).

Genome assembly	**Cvlepa_BANY2021**
Genome size	233.6 Mb
Number of contigs	184
Mean length	1.3 Mb
Largest contig	15.4 Mb
N50	8.5 Mb
N90	1.7 Mb
L50	11
L90	30
GC	35.6%
BUSCO (Metazoa n=954	
Complete single cc	91.1%
Complete duplicate	2.3%
Fragmented	2.9%
Missing	3.7%
Transcriptome	
Transcripts	31 035
Protein coding transcrip	22 048

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024

with *Cirobu* hit	19 263 (87.4%)
with SwissProt hit	16 818 (76.3%)
BUSCO (Metazoa n=954	
Complete single cc	59.0%
Complete duplicate	31.2%
Fragmented	2.7%
Missing	7.0%
Genome annotation	
Genes	32 318
Transcripts	35 178
Protein coding genes	25 067
with *Cirobu* hit	19 894 (79.4%)
with SwissProt hit	15 599 (62.2%)
BUSCO (Metazoa n=954	
Complete single cc	86.8%
Complete duplicate	6.4%
Fragmented	1.6%
Missing	5.2%

Table 1. Statistics for the assemblies and annotation.

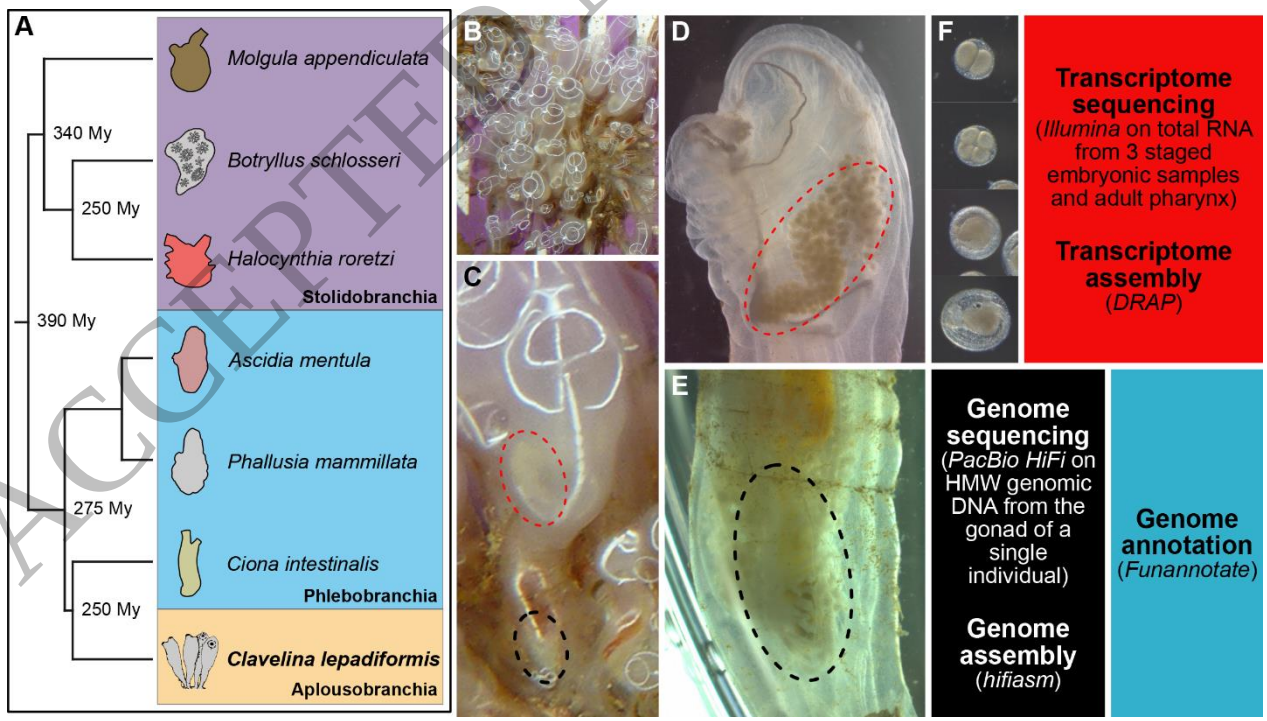

Figure 1
198x111 mm (DPI)

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evae038/7619261 by BIU Scientifique de Jussieu additional IP addresses user on 19 March 2024