The genome of the "Sea Vomit" *Didemnum vexillum*

Ernesto Parra-Rincón

Universidad Nacional de Colombia

Cristian A Velandia-Huerto

Universitat Leipzig

Jörg Fallmann

Universitat Leipzig

Adriaan Gittenberger

Universiteit Leiden

Federico D Brown

Universidade de Sao Paulo

Peter F Stadler (✉ studla@bioinf.uni-leipzig.de)

Universitat Leipzig <https://orcid.org/0000-0002-5016-5191>

Clara I Bermúdez-Santana (✉ cibermudezs@unal.edu.co)

Universidad Nacional de Colombia

Research article

Keywords: Tunicata, *Didemnum vexillum*, microRNAs, genome annotation

Posted Date: June 2nd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-23360/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Life on December 10th, 2021. See the published version at <https://doi.org/10.3390/life11121377>.

RESEARCH

The genome of the “Sea Vomit” *Didemnum vexillum*

Ernesto Parra-Rincón^{1*}, Cristian A. Velandia-Huerto^{7,1*}, Jörg Fallmann⁷, Adriaan Gittenberger^{2,3,4}, Federico D. Brown^{5,6}, Peter F. Stadler^{7,8,9,1,10†} and Clara I. Bermúdez-Santana^{1†}

Abstract

Background: Tunicates are the sister group of vertebrates and thus occupy a key position for investigations into vertebrate innovations as well as into the consequences of the vertebrate-specific genome duplications. Nevertheless, tunicate genomes have not been studied extensively in the past and comparative studies of tunicate genomes have remained scarce. The carpet sea squirt *Didemnum vexillum* is a colonial tunicate considered an invasive species with substantial ecological and economical risk.

Results: We report a newly re-assembled genome of *Didemnum vexillum*. We used a hybrid approach that combines two genome sequencing technologies and also its first transcriptome. Started from 28.5 Gb Illumina and 12.35 Gb of PacBio data a new hybrid scaffolded assembly was obtained comprised of a total size of 517.55 Mb that increases contig length about 8-fold compared to previous, Illumina-only assembly. While still highly fragmented (L50=25,284, N50=6539), the assembly is sufficient for comprehensive annotations of both protein-coding genes and non-coding RNAs.

Conclusions: The draft assembly of the “sea vomit” genome provides a valuable resource for comparative tunicate genomics and for the study of the specific properties of colonial ascidians.

Availability: Genome and annotation data as well as a link to a UCSC Genome Browser hub are available at <http://tunicatadvexillum.bioinf.uni-leipzig.de/>.

Keywords: Tunicata; *Didemnum vexillum*; microRNAs; genome annotation

Background

The carpet sea squirt *Didemnum vexillum* [1], a.k.a. “sea vomit”, is a colonial tunicate presumably native to Japan that has appeared as an invasive species in Europe, the Americas, and New Zealand [2]. It negatively affects established benthic species and damages ship hulls as well as the infrastructure in marinas, ports, and shellfish farms.

Rapid colony growth or regression in response to the dynamics of the habitat [3], water temperature [4], colony fragmentation as a reproductive and dispersal strategy [5], fast asexual budding that allows attachment to a variety of living and/or non-living substrata, and relatively few predators [3] have facilitated *D. vexillum* to become a well-recognized world-wide invader.

The invasion potential of *D. vexillum* has an important economic impact on the aquaculture industry as it affects the conditions of bivalve and shellfish cultures (see e.g. [6] and the references therein), and increases the cost of maintenance to avoid the fouling process on mussel cages and facilities [7].

Despite the economic impact of tunicates and their pivotal phylogenetic position as sister group of the vertebrates, genomic studies and comparative analyses have remained relatively scarce. So far, the genomes of three solitary tunicates have been assembled and annotated in substantial depth: for the closely related sessile ascidians *Ciona savignyi* and *Ciona robusta* assemblies of their 14 chromosomes are available [8–11]; and for the pelagic larvacean *Oikopleura dioica* only 6 chromosomes have been reported [12–14]. In addition, draft assemblies recently have become available for the pelagic colonial thaliacian *Salpa thompsoni*, which was used to analyse the high mutation rates in the genomes of tunicates [15].

Full list of author information is available at the end of the article

*Equal contributor

†Correspondence: studla@bioinf.uni-leipzig.de (PFS) or cibermedezs@unal.edu.co (CIBS)

Other genomic studies in tunicates include: for solitary ascidians, the genome of *Halocynthia roretzi* was used to predict microRNAs [16], and the genomes of three species of *Molgula* (*M. occidentalis*, *M. oculata* and *M. occulta*) were used [17] to study developmental system drift during cardiopharyngeal development. At the same time, the ANISEED database [18] served as a hub for ongoing sequencing projects for other ascidian species, including: *Phallusia mammillata*, *Phallusia fumigata* and *Halocynthia aurantium*. For colonial ascidians, the genome of *Botryllus schlosseri* has been assembled to 13 incomplete chromosomes (of the 16 chromosomes in total) [19], and a draft assembly comprising 1778 scaffolds has been reported for the related species *Botrylloides leachii* [20]. A very fragmented assembly of the “sea vomit” *Didemnum vexillum* was also recently sequenced by our group to analyze non-coding RNAs (ncRNAs) [21]. Here we report on a substantial improvement of this assembly.

Comparisons between tunicate and other chordate genomes have identified both expansions of gene families and innovations such as the horizontally transferred genes of cellulose synthase from Actinobacteria [22], but also substantial losses, e.g., of parts of the *homeobox* (HOX) gene cluster [20]. The genomic organization of tunicates, as exemplified by *Ciona* and *Oikopleura* shows substantial differences compared to both vertebrates and amphioxus, the common outgroup to the Olfactores [23], and has led to the hypothesis that tunicates have undergone major genomic re-structuring because of an accelerated rate of evolution that was linked to changes in the organization of the entire gene complements [15, 24, 25]. In contrast, other chordate lineages have maintained a fairly constant rate of evolution [15, 24, 25].

In this study we expand the assembly and annotation of tunicate genomic resources, and improve the current genome assembly of the colonial tunicate *D. vexillum* to contribute to the understanding of chordate evolution, as well as to help understand the genomic changes involved in the novel mechanisms of asexual reproduction of colonial animals.

Results

Assembly of the *D. vexillum* Genome

An improved assembly of the *D. vexillum* genome was obtained by integrating PacBio and Illumina sequencing (see Methods for details). This resulted in a new assembly at scaffold level, comprising about 517.55 Mb. This amounts to a reduction in the number of genome fragments by a factor of $\sim 8\times$ and a corresponding increase in the N50 length from 918 bp to ~ 6.5 Kb.

The new assembly also decreases the estimated genome size by about 25 Mb. While only about 15% of the contigs in our previous study [21] were longer than 1 kb, this threshold is now exceeded by almost 96% of the scaffolds in the new assembly and thus allows at least a comprehensive gene-level analysis. The newly analyzed nucleotide composition was consistent with our previous study [21].

D. vexillum has a similar genome size and GC content as other deuterostome genomes, including 10 tunicate genomes. Among tunicates, solitary organisms appear to have smaller genomes (≤ 250 Mb) than colonial ones (with range from 160 to 723 Mb). The *D. vexillum* genome thus appears in the typical size range for colonial tunicates, and in terms of its size, the *D. vexillum* genome is comparable to the amphioxus genome (Figure 1).

Tunicates have a GC content $\leq 43\%$, with lowest values reported for solitary ascidians, particularly in the molgulids. *D. vexillum* is similar in GC content to the salp *S. thompsoni* and solitary ascidian species to the zebrafish *D. rerio*, but also to the two outgroup species, the ambulacrarians: (*S. purpuratus* and *S. kowalevskii*). Although we do not find a clear relationship between GC content and genome size in the deuterostomes, when we compare both factors (i.e. genome size and GC content) together, there is a tendency for tunicates to have both lower GC content and smaller genome sizes when compared to other deuterostomes, and other chordates in particular. Moreover, within tunicates, solitary species show even lower GC content and genome size compared to colonial species (Figure 1). It remains an open question what the biological consequences of this trend are for tunicates in general and colonial tunicates in particular.

Partial degradation of genomic DNA

Despite considerable efforts we could not avoid partial degradation of gDNA isolated from *D. vexillum* even though the same protocols were used by one of us (F. Brown) for other ascidians such as *Styela*, *Perophora* and *Clavelina* spp., where no problems of DNA degradation were encountered. Fragment size was quantified using Agilent 2100 Bioanalyzer at the University of Washington PacBio Sequencing Service Facility showing that the peak size was 2 Kb but some longer material was presented.

We hypothesize that the particularly high gDNA degradation in *D. vexillum* may be linked to the presence of tunic bladder cells, which are restricted to some groups of ascidians, including the Didemnidae. The bulk of their cytoplasm comprises a large vacuole containing sulfuric acid, which accounts for a tunic pH

<3.0 in didemnids [26] that may be involved in chemical defense. In contrast, tunic pH>6.0 was measured for *Perophora* and *Clavelina* species. The acidic pH may account for the observed gDNA degradation, possibly due to increased deamination rates [27, 28]. The partial degradation of gDNA is a confounding factor for genome assembly and accounts for the limited quality of our assembly.

Transcriptome Sequencing and Assembly

In order to support the annotation of the Sea vomit genome, the transcriptome was assembled from RNA-seq reads. **Trinity** assembled a total of 55.1 million paired-end Illumina reads into 90,938 transcripts. After two training rounds of **Maker**, only 64,424 transcripts were annotated, with a median contig length of 375 nt with a positive skewed, long-tailed distribution. There are transcripts with a length > 10 kb, both corresponding to the uncharacterised proteins **Dex_pep14095** and **Dvex_pep554** (as shown in Additional File 1: Tables 1 and 2). Both of which have homologs in *C. robusta*, containing the **TILa** (PF12714) and **von Willebrand factor type C** protein domains (PF00093).

Genome Annotation

Detection and analysis of repetitive regions

To identify regions prone to have repetitive elements, a combined strategy using **RepeatModeler** and **RepeatMasker** was used to generate a *de novo* library. Additionally, including also the reported repeats library from *C. robusta*, the *D. vexillum* genome was soft-masked, as explained in Methods. About 300.66 Mb, i.e. 57.89% of the assembled *D. vexillum* genome, consists of repetitive elements (see Additional File 1). Most of the repetitive elements are interspersed repeats (56.96%) as well as retroelements (12.86%), DNA transposons (7.65%), leaving about 100 Mb of the repeats as unclassified elements (35.96%). This is similar to the repeat content of *B. schlosseri* (~59.85%) [20]. The most abundant family of repeats in *D. vexillum* are 100,404 copies of SINE/tRNA-Lys, a class of repetitive elements that have not been reported for other tunicates species. The other highly abundant families (i.e. LINE/L2 and DNA/hAT-Charlie) are also prevalent in other tunicates, see Additional File 1: Table 6 for details.

Annotation of protein-coding genes

We identified 62,194 coding genes accounting for 64,424 distinct protein products (Table 1). About

97.5% of coding genes, have 0.97 kb in median, generate only one transcript and thus a single predicted protein product. Those genes that reported more than one transcripts have minimum and maximum median sizes of 0.44 to 7.07 kb, respectively (Additional File 1: Table 1). The largest annotated gene, *Divexi.CG.Dive2019.scaffold1-size56789.g1453*, has a size of 33.74 kb and comprises as single transcript product, which accounts for a protein with domains as Laminin N-terminal (PF00055), Laminin EGF (PF00053) and Carbohydrate Binding Module 6 (PF03422). At the same time, it presented high homology to the *C. robusta* laminin alpha 5 subunit protein XP_026696566.1. The gene with largest number of transcripts is a homolog of *Dynein heavy chain* proteins. It covers 16 exons in a region of only 7.07 kb and produces 10 observed isoforms, see Additional File 1: Figure 2 and Additional File 1: Table 2.

In order to assess the quality of both the genome assembly and the predicted gene set, we used BUSCO to compare them to metazoan orthologous genes (Figure 2). For the *D. vexillum* genome, from the 978 orthologs, 50.8% were found complete. Overall, the BUSCO results are comparable to other, published tunicate genomes (Figure 2), indicating that current assembly of *D. vexillum* is comparable to the *S. thompsoni* assembly in terms of completeness and annotation. In general terms, most of the reported tunicate genomes displayed $\geq 75.4\%$ of complete BUSCO orthologs.

In this initial annotation, we specifically searched for *homeobox* transcription factors using a combined **blastp/tblastn** strategy (see Methods) that identified 48 coding sequences with their corresponding same number of genes located in 47 scaffolds. The most frequent found proteins are homologs from the families: *ZEB2*, *LHX2* and *Irx* transcription factors. In an alternative approach we used the genome-wide alignments to compare existing annotations of *homeobox* genes in six tunicate and one cephalochordate genomes to our *D. vexillum* assembly (see Methods). Only one of the 48 *homeobox loci* had annotated homologs in four of the six query species, which corresponds to a *Hox2* gene, located on the *scaffold16549-size8805*. Several other Hox genes, however, were not recognized by the default homology annotation pipeline because of incomplete overlaps, and in some cases, no gene was annotated for *D. vexillum* (Figure 3).

By comparison with *H. roretzi* and *Ciona* spp., we expected to find three anterior, three middle-group, and three posterior Hox genes as in other tunicate genomes [29, 30]. Based on the data outline above and a more detailed manual search with genome alignments as support, we found evidence for two anterior

genes (*Hox2* and *Hox3*), two central genes (*Hox4* and *Hox6/7*-like), and the three expected posterior genes, as referred on Figure 3. What the consequences are for *D. vexillum* for the presumable absence of *Hox1* and *Hox5* remains to be studied. The assembly of the HOX gene region unfortunately is too fragmented to conclusively rule out the presence of *Hox1* and *Hox5* or to provide any linkage information of the reported Hox genes.

Annotation of noncoding RNAs

Noncoding RNAs were annotated using a homology-based strategy combining **blastn** searches, HMM profiles, and covariance models (CMs) as described in [21] with some modifications detailed in Methods. Not counting tRNAs, we identified 2153 ncRNA *loci* corresponding to 271 distinct ncRNA families. A search with **tRNAscan-SE** resulted in 18,343 predicted loci, including pseudogenes and undetermined isotype candidates. In addition we mapped the 206 families of ncRNAs identified in a preliminary draft of the *D. vexillum* genome [21] to the current assembly (see Methods and Additional File 1). As in other genomes, in particular the pol-III transcribed RNAs including 5S rRNA, tRNAs, U6 RNA as well as the snRNAs transcribed by pol-II appear in multiples copies [31]. The data are summarized in Table 2. While most ncRNAs were visible in the automatized annotation pipeline, several additional ncRNAs could be added by manual curation only. **RFAM** IDs for the RNA families mentioned below can be found in Additional File 4.

Transfer RNAs. We found 2724 tRNAs and 15,619 tRNA pseudogenes or with undetermined isotype (23). The most abundant tRNA is *tRNA_{Thr}* with 1395 copies, while only a single copy of *tRNA_{Sec}* was observed. Surprisingly, **tRNAscan-SE** reports a large number of suppressor tRNAs: 153 (*tRNA_{Suppressor-TCA}*:145, *tRNA_{Suppressor-TTA}*:7, and *tRNA_{Suppressor-CTA}*:1). Detailed information on the tRNA annotation is compiled in Additional File 1: Figures 3 and 4.

Ribosomal RNAs. As in most eukaryotes, the small and large subunit (SSU 18S and LSU 28S) rRNAs are organized in repetitive units of the rRNA operon. It also contain the 5.8S rRNAs. In this case, *D. vexillum* reported 6 clusters of rRNAs: two clusters are composed of repetitions of 5S rRNA (*scaffold1545-size16374* and *scaffold22447-size6833*), two clusters contain SSU 18S, 5.8S and LSU 28S rRNA elements within (*scaffold4839-size12187* and *scaffold9164-size12300*), and one cluster contains repetitions of 5.8 rRNAs with a loci of LSU 28S rRNA (*scaffold4349-size12561*). At the same time for the subunit 5S rRNA 71 loci were detected and from them

52 are located on the *scaffoldUncertain*. For the other rRNAs elements, in total were found 6 5.8S, 3 SSU, and 4 LSU rRNAs.

Spliceosomal RNAs. All RNA components of the spliceosome machinery were found in the new genome assembly. As usual, the snRNAs of the major spliceosome appear in multiple copies U6 (46), U5 (9), U1 (21), U2 (27), U4 (3). Among the snRNAs of the minor spliceosome, U12 appear once, while there are 2 loci coding for U4atac, U6atac, and 4 U11 genes.

Other small nuclear RNAs. We identified the expected genes for the RNA component of the signal recognition particle as well as the RNase P RNA, RNase MRP RNA, and 7SK RNA. No homologs were found for the telomerase RNA, U7 snRNA, vault RNA, and Y RNAs, although their presence in the genome is expected. These groups are notoriously difficult to be detected by homology search without the benefit of known homologs in closely related species [32]. A thorough search along reported Tunicata genomes successfully reported vault snRNA *loci*, except for *D. vexillum*, other families were not detected, indicating that specific CMs should be redefined with a broad set of sequences to improve the annotation from those families on *D. vexillum* and another tunicate species (Additional File 1: Table 5).

MicroRNAs. The miRNA annotation pipeline, described in the Methods section, identified 2065 *loci* encoding members of 248 distinct miRNA families. An additional 20 *loci*, which harbor two additional families, correspond to previously reported miRNAs [21] which were successfully mapped into the new assembly. To avoid the annotation of false positives due to the modification of the threshold values (see Additional File 1: Figure 12), the position of the *mature* sequence was evaluated using **MIRFix** [33] which used both, the **RFAM** database for the miRNA families alignments and **miRBase** as source for the annotated *mature* sequences (as explained in more detail in Methods and Additional File 1: Figure 5). As a result, the definition of a true miRNA candidate relies not only on the homology results given by the sequence/secondary structure comparison, but also in the annotation of their *mature* sequence. In addition, we also require miRNA-specific features, such as a conserved position of the mature products within the defined miRNA family. To this end, candidates that reported homologous *mature* regions were compared against their corrected **stockholm** alignments, by the calculation of the *tree edit distance* between generated consensus secondary structures, as described on Additional File 1.

In this way, a number of 1582 *loci* were reported, from which 1394 fulfill all the designed filters and reported a set of *mature* sequences harbored at the predicted hairpin structure, the other 188 have broken the

conservation block in the defined family alignment, despite having shown a high conservation at hairpin level. Taking into account those detected miRNAs with *mature* annotation, the distribution of *loci* shows that 75% of miRNA families have less than 6 *loci*. The corresponding 25% of miRNA families have a higher median of ~ 11.5 *loci*. Within these miRNA families, mir-544 (65), mir-578 (70), and mir-944 (97), had the highest number of *loci*.

We also analyzed the phylogenetic distribution of the miRNAs in the Rfam *seed* alignment, the corresponding species were retrieved along with their annotated *kingdom*, *phylum* and *subphylum*, as described in Methods. The annotated miRNA families and their *loci* in *D. vexillum* were compared as shown in Additional File 1: Figure 7. We found 18 miRNA families that were represented in more than 2 *phyla*: **mir-124**, mir-598, mir-7, **let-7**, **mir-1**, **mir-133**, mir-33, *lin-4*, **mir-137**, **mir-153**, mir-2, **mir-31**, mir-449, **mir-183**, **mir-190**, **mir-210**, **mir-219** and **mir-8**. Families highlighted in bold showed a conserved structure (panel labeled as *VALID_STR*), even when the *D. vexillum* sequences were included into the alignment. In this analysis, we uncovered two additional families: **ciona-mir-92** (RF01117) and **mir-281** (RF00967), to the previously reported **mir-1497** (RF00953) [16], candidate in *D. vexillum*. In contrast, a subset of 13 miRNA family candidates did not fit into the corrected *stockholm* alignment (classified as *NO_VALID_STR*), despite our previous homology validation.

In a previous study of the miRNA complement in the solitary species *H. roretzi* [16] a more extensive list of tunicate-specific miRNAs was reported (21). From these only one (**mir-1497**, (RF00953)), was detected in our study because of the corresponding covariance model used to validate their secondary structure. From the *conserved* families of miRNAs in Metazoa (25) we identified 21 in *D. vexillum*. Other families, including **mir-9**, **mir-182**, **mir-184**, **mir-200**, and **mir-218**, were not found. These families (except **mir-200**) were also found to be absent in other tunicates such as *C. savignyi* and *O. dioica* [16]. Absence of these families were also reported in a preliminary analysis along bilaterian species [34].

From our previously reported set of miRNAs [21], 16 families were detected only in *D. vexillum* and not in other tunicates. From this set, 10 families were annotated in our new assembly and four were discarded because their mature sequences could not be annotated (**mir-130**, **mir-460**, **mir-185**, and **mir-233**), one does not have covariance model (**mir-4068**), and another was not found in the new assembly (**mir-9**). From the set of shared families in colonial tunicates, all were annotated and validated by our strategy, except **mir-340** (RF00761). The latter showed a good

homology but did not pass the conditions of the current structural alignment strategy, which used only vertebrate sequences to assign homology. In this study, we report **mir-31**, as the sole miRNA candidate that passed all our present filtering criteria to be exclusively found in solitary ascidian species. We also excluded 502 candidates based on the lack of conserved *mature* sequences inside the hairpins.

Small Nucleolar RNAs. Conserved snoRNA families were detected by the automatized homology-search strategy. We found 3 U3, 2 copies for SNORD14, SNORD18, snoZ39, and SNORA36, as well as a single copy of SNORD29, SNORD33, SNORD35, SNORD36, SNORD52, SNORD63, and SNORD83.

LncRNAs and other structured RNA elements. Two structured lncRNAs were found, corresponding to the *Rhabdomyosarcoma 2 associated transcript conserved region*: RMST8 (1) and RMST 9 (7), the latter one had already been previously annotated [21]. As a result of the iteration and re-building of the correspondent CM with newly detected tunicate sequences, (see Additional File 1: Figure 8) we now report the occurrence of the complete RMST family in deuterostomes. RMST 8 and 9 were detected in all deuterostomes. We found two additional RMST families (RMST 6 and 7) in the coelacanth suggesting an initial expansion in the ancestor of lobe finned fishes (Sarcopterygii). The complete set RMST 1, 2, 3, 4, 5, 6, 7 and 10 were detected in mammals. Because of their relevance in neural development [35], it would be interesting to study the evolution of RMSTs in the tetrapods, and the ancestral role of RMST 8 and 9 in the deuterostomes, the tetrapods and mammals.

Finally, by using a specific search with HMMs and CMs we identified 326 loci carrying the Histone 3' UTR stem-loop, 6 instances of the Potassium channel RNA editing signal, one for the Iron response element II and 9 loci for the Hammerhead ribozyme (type I).

Functional annotation and comparison of proteins across the tunicates

To obtain functional annotations for the predicted *D. vexillum* proteins we used the pre-clustered orthology groups from the **eggNOG** database [36] together with the protein annotation of eleven chordates (see Methods). We obtained 8349 orthology groups of which 6279 were represented in at least two of the chordates included in our reference set. Figure 5A, shows that 57.1% (4584) of the orthologs were shared with at least one sequence of each of the major branches of the chordates (Cephalochordata, Tunicata, and Vertebrata). Only 3.63% of orthologs were shared exclusively with at least two species of tunicates, while 15.81% of orthologs were shared only with vertebrates.

Along all the detected set of orthologs shared exclusively among two or more tunicates (292), 5 were found present in all Tunicata species. From this subset, the *lytic polysaccharide monooxygenase* (ENOG5028N9R, 20) was involved in cellulose fibrillation and degradation. The other orthologs were unknown proteins with sulfotransferase family domains (ENOG502CNPV, 444; ENOG502CXMB, 23), pleckstrin homology domains (ENOG502EA0P, 24) or transmembrane domains with unknown function (ENOG502EQW0, 32). Another set, composed by 5 orthology groups, was found when *O. dioica* was excluded. From those groups, only one did not have functional annotation and the other 4 revealed functional annotations related to *dopamine monooxygenase activity* (ENOG502C4CE, 33), *regulatory subunits of protein phosphatases* (ENOG502CZS6, 17), the *reducing fluoride concentration levels in the cell* (COG0239, 15) and *AMP binding* (COG0589, 71).

Five orthology groups were found exclusively in all three available colonial tunicate genomes (*B. schlosseri*, *B. leachii* and *D. vexillum*), and 21 orthology groups were shared between *D. vexillum* and at least one other colonial botryllid. The set of orthology groups shared by all colonial ascidian genomes (5) contain proteins that present the following shared domains: Methyltransferase domains (ENOG502AI1Z), GIY-YIG endonucleases (ENOG502D72K), THAP DNA binding domains, DDE endonucleases (ENOG502E46Z), Antistatin domains (ENOG502FAER) and domains that lack annotation as the DUF3605 domain (ENOG502E7AU). To reach a more universal understanding of ortholog proteins during the evolution of coloniality, we need to better characterize and assign cellular functions to the conserved proteins found in colonial tunicates that evolved from independent events of coloniality. The above mentioned groups of proteins present in *D. vexillum* and the two botryllids will provide a starting point to address their biological roles for ascidian colonies.

In contrast, no orthology groups were shared between all solitary tunicates (*C. robusta*, *C. savignyi*, *O. dioica*, *M. oculata* and *M. occidentalis*), but three were shared among the four solitary ascidians, excluding the larvacean. The orthology groups shared by all solitary ascidians included proteins and presented the following shared domains: (i) *Tumor necrosis factor receptor / nerve growth factor receptor repeats* (ENOG502D0E2); (ii) a group of proteins with the *BESS motif* (ENOG502DZH1), usually associated with DNA binding as molecular function (GO:0003677); and (iii) ARM-like, ARM-type fold and Rotatin domains of the Rotatin gene family members that have functions related to cilium organization.

In spite of an absence of 1737 orthology groups of predicted proteins in the *D. vexillum* genome, most of these orthologs became detectable when the Chordata group were analyzed (79.2%). Moreover, in Olfactores and Tunicata (include *D. vexillum*), we found 12.7% and 8.1 of orphan genes, respectively. Although we report the functional profile of orphan genes in *D. vexillum*, which represent the majority of the orthologs recovered, we were not able uncover a clear functional annotation for many of these genes (see Additional File 2: Table 1 for details).

Despite the difficulties in the assignment of ortholog candidates across all genome datasets, comparisons against clustered groups allowed us to detect and annotate orthologs in the *D. vexillum* genome. Because the Didemnidae can mineralize calcium to form spicules in their tunics, we decided to search for key proteins involved in skeletogenesis [37]: *Sox*, *Hedgehog* (Hh), and *RUNX*, which corresponded to the ortholog groups: *KOG0527* (SOX), *KOG3638* (Hh), and *KOG3982* (RUNX) on the eggNOG database. Gene phylogenies for these ortholog groups (including the chordate sequences used as reference and the orthologs annotated in the eggNOG database) are shown in Figure 4. In *D. vexillum*, we found seven members of the SOX family belonging to SoxB1, SoxB2, SoxC, SoxD and SoxE subgroups as defined in [38]. Overall, we found two paralogs for the SOXC (*SOX4/SoxC#32* and *SOX4/SoxC#33*) and SoxB2 (*SOX14/SoxB2#5* and *SOX14/SoxB2#6*) in our annotation of the *D. vexillum* genome, see Figure 4A and Additional File 2: Figure 3 for the complete tree.

All tunicates except *O. dioica* reported members of the Hh families (Figure 4B). The basal Hh family, previously reported in *Ciona* [39] and in amphioxus [40], was detected in all ascidians. In the vertebrates, we confirmed the presence of the three Hh genes: Desert (DhH), Indian (IHh) and Sonic-hedgehog (SHh) [40, 41]. In ascidians, we found several clades of Hh genes. There are at least three Hh families in the ascidians: Hh clade A (with medium bootstrap support of 61), Hh clade B (with full bootstrap support in *Ciona*) and Hh clade C (with full bootstrap support in the botryllids). The *D. vexillum* Hh does not group with any of the other clades. Our analysis supports an independent diversification of the Hh family in ascidians.

We did not find the key regulators of skeletogenesis RUNX-related transcription factor (RUNX) proteins in *D. vexillum*. This does not necessarily indicate a true loss, however, because in a detailed domain-based homology search (data not show), we found parts of the domain *Runt* (PF00853) along 15 proteins from *D. vexillum*, albeit with truncated sequences. The phylogenetic distribution of the orthologs found (Additional

File 2: Figure 2), shows a defined clade of tunicate sequences that belong to the ancestral RUNX family, which has been detected in this study in amphioxus and is known to be expressed in *Ciona* and *Oikopleura* [42]. This suggests that RUNX proteins may not be truly absent in *D. vexillum*. We note in passing that the RUNX family has undergone additional duplications in the lampreys (Additional File 2: Figure 2).

Ortholog groups determined by the eggNOG database were used to transfer the annotation to the corresponding orthologs in *D. vexillum*. General ontology terms (e.g. cellular, metabolic, multi-organismal processes, reproductive processes, regulation and locomotion) were commonly annotated for *D. vexillum* proteins (Additional File 2: Figure 4). Enrichment analysis with REVIGO [43] based on the frequencies of ontology terms (see Methods and Additional File 2 for details) detected seven distinct overrepresented semantic clusters in *D. vexillum*: *Positive regulation of phosphatidylinositol 3-kinase signaling*, *tRNA catabolism*, *secondary metabolism*, *chaperone-mediated protein folding*, *protein folding*, *protein autophosphorylation*, and *phosphorus metabolism*, see Figure 5B.

A detailed annotation of the *D. vexillum* genome comparing the GO assignments from selected chordates genomes is provided in Additional File 2: Figures 6 and 7. We found a total of 237 tunicate-specific enriched GO terms, when compared to the annotations in *B. floridae*, *P. marinus* and *L. chalumnae*. All tunicates, except *O. dioica* shared 8 assignments. Where related terms were found, we indicate these relationships ($\rightarrow$ “is a”, $\mapsto$ “part of”) as follows: Oogenesis (GO:0048477) $\rightarrow$ Germ cell development (GO:0007281) $\mapsto$ Gamete generation (GO:0007276) ($\leftarrow$ Female gamete generation (GO:0007292)), cellular process involved in reproduction in multicellular organism (GO:0022412) $\rightarrow, \mapsto$ Multicellular organismal reproductive process (GO:0048609) $\mapsto$ Multicellular organism reproduction (GO:0032504) $\rightarrow$ Reproduction (GO:0000003).

Based on the previously described semantic clusters, the functional interactions of involved *D. vexillum* proteins were inferred using STRING (v.11) [44], comparing them with their homologous proteins annotated in *C. robusta*. As an example, Figure 5C shows the annotations for *C. robusta* that have been detected as homologs of the proteins in *D. vexillum* involved in tRNA catabolism processes. As a result, it was possible to detect 5 protein clusters, each one with a specific interaction, as follows: cluster 3 was related to the autophagy pathways (KEGG pathways cin04136, cin04140, and cin04137), while clusters 2 and 5, are involved to the ribosome biogenesis and

RNA transport pathways (cin03008, cin03013), respectively. Clusters 1 and 4 contain proteins without any clear association. For the other ontology clusters (Additional File 2: Figure 4) we carried out the same analysis. As expected, for the proteins involved in protein folding, the functional annotation pointed out processes related with the endoplasmic reticulum. Regarding the pathways of secondary metabolism, we detected process involved in starch, sucrose, and porphyrin metabolism, as well in chlorophyll metabolic pathways. In more detail, 3847 proteins are involved in the *Positive regulation of phosphatidylinositol 3-kinase signaling*, and in processes such as: endocytosis (cin04144), autophagy (cin04140), mTOR, FoXO, Wnt, and Inositol signaling pathways (cin04150, cin04068, cin04310 and cin00562) and RNA transport (cin03013). In addition, 1056 and 1053 proteins were found related to phosphorus metabolism and protein autophosphorylation, respectively. This detected proteins reported the same interactions in: metabolic pathways (cin01100), Inositol phosphate metabolism (cin00562), phosphatidylinositol signaling (cin04070), FoxO signaling (cin04068), purine metabolism (cin00230) and autophagy (cin04140). In summary, we were able to infer several candidate functional networks on the basis of the semantic clusters detected in *D. vexillum* with the help of homologous proteins from the solitary tunicate *C. robusta*. However, a much more extensive annotation effort will be necessary not only for *D. vexillum* but also for tunicate genomes in general, in order to produce a more complete picture of the functional landscape.

Genome Browser and analysis of genomic coordinates

We updated the data on the *D. vexillum* genome site <http://tunicatadvexillum.bioinf.uni-leipzig.de/> with the data reported here. In particular, it provides the new assembly and a link to a UCSC genome browser hub [45] as described in Methods. Genome coordinates for ncRNAs and annotated genes were concatenated, sorted and crossed by incrementing the starting position for each scaffold and by reporting the genome coordinates. We labeled the ncRNAs as suggested in the guidelines for tunicate elements [46]. Accordingly, we found that a total number of 2378 genes have a ncRNA nearby or within their gene structure. These corresponded to 1832 ncRNAs, where 53.93% were tRNAs, 36.14% (183) miRNA families, 6.66% (3) cis regulatory RNAs and 0.27% miscellaneous RNAs (1) and ribozymes (2). Other house keeping RNAs summed up 3%, including rRNAs (3 families), snoRNAs (10) and snRNAs (6).

Mitochondrial Genome

The mitochondrial genome of *D. vexillum* maps to a single scaffold *scaffold1656-size16126* and very closely matches the two previously reported mitogenomic sequences [47], known as Clade A and Clade B. The mt-LSU is 99.9% identical to Clade A, and diverges about 3.6% from Clade B, confirming that the collected organisms belongs to clade A, see also [21]. Mapping the currently reported elements from mtDNA, resulted in the gene order depicted on Additional File 1: Figure 10. In this case, intergenic distances were reduced, but the size and the order of the genes in the new assembly were conserved. The 37 expected elements from the mtDNA were mapped into the new assembly. The gene order of the mitogenome matches that of clade A but differs from other tunicate species, as shown in the multiple alignment of the mitogenomes in Additional File 1: Figure 11.

Discussion

The draft genome assembly using a combination of PacBio and Illumina data reported here provides a comprehensive resource to study *D. vexillum* at least at the level of genes. The contiguity of the assembly still falls short of those available for other ascidians. Despite considerable efforts, a partial degradation of the genomic DNA, presumably due to the unusually acidic milieu of the tunic, could not be avoided entirely, limiting the achievable PacBio read lengths. The genome assembly as well as the annotation data are made available for both download and interactive use in the browser <http://tunicatadvexillum.bioinf.uni-leipzig.de/>. The mitochondrial genome is assembled to a single scaffold. Analysis of its mt-LSU RNA showed that the specimen used for sequencing belongs to *D. vexillum* Clade A.

Functional annotation of the predicted proteome of *D. vexillum* by comparison with 11 chordates resulted in 8349 orthology groups. The vast majority is shared among chordates. We identified 292 orthology groups in tunicates only (present in more than one tunicate). Among them five functional groups shared by all tunicates, including *lytic polysaccharide monooxygenase and cellulose-degrading processes* (ENOG5028N9R). Other shared orthology groups did not have a specific annotation, however in some cases protein domains (e.g., *sulfotransferase and pleckstrin families* and some transmembrane domains) were recognizable. Of all of the chordate orthology groups available, 1737 groups were not recovered in our *D. vexillum* assembly. Most notably, we did not find any member of the *RUNX*

family, which correspond to key regulators of skeletogenesis together with *HH* and *SOX* family members. We observed that tunicates, except *O. dioica*, showed a tunicate specific expansion of HH members. We found seven members of the *SOX* family. A phylogenetic analysis revealed duplication events for SoxC and SoxB2 in *D. vexillum*. We also identified seven of nine tunicate *homeobox* transcription factors of HOX family, the contiguity of the assembly is insufficient to conclusively rule out the absence of the remaining two genes (*Hox1* and *Hox5*) or to determine the genomic organization of the HOX gene cluster.

The new assembly increased the number of detected ncRNA families to 4877 genomic loci corresponding to 271 families. From these, most of the detected *loci* were *housekeeping* ncRNAs (rRNAs, tRNAs, snRNAs and snoRNAs) and those *loci* where found in a conserved cluster organization, as seen on tRNAs, rRNAs and snRNAs. At the same time, a new set of regulatory ncRNAs (miRNAs, Cis-regulatory RNAs and lncRNAs) were detected, in case of miRNAs by the modification of the default threshold value on the structural alignment step and on the validation of the position of the mature sequences inside the detected hairpin. As expected, the conserved set of miRNAs were annotated: mir-124, mir-598, mir-7, let-7, mir-1, mir-133, mir-33, lin-4, mir-137, mir-153, mir-2, mir-31, mir-449, mir-183, mir-190, mir-210, mir-219 and mir-8. In comparison to previous miRNA tunicate surveys [16, 34], we validated previous reports of tunicate-specific miRNA mir-1497 (RF00953), by detecting the mature position and evaluating it along a secondary family specific structural multiple alignment, and also reported additional specific families, such as: cionamir-92 (RF01117) and mir-281 (RF00967). Additionally, we discarded previously reported 6 *D. vexillum* miRNA families because of a lacking of the mature annotation or missing of covariance model, and one other colonial specific mir-340 (RF00761), that could not be supported using the current multiple structural alignment. Further studies will allow us to continue to refine the complete miRNA complement in *D. vexillum* and allow us to reconstruct the evolutionary history of miRNAs in the tunicates. We were not able to identify homologs of other expected ncRNA families, as: *vault*, *U7* and *Y* RNA and *Telomerase* RNA.

Concluding Remark

This new assembly of the *D. vexillum* genome provides an integrated effort to contribute for the ongoing Tunicata genome projects and constitutes the first annotation dataset for a species in the Aplousobranchia. We hope that the new *D. vexillum* genome annotation

presented here triggers more biological studies in a representative of a highly invasive species with a colonial life history.

Methods

Re-assembly of *D. vexillum* genome

DNA extraction

On 10 June 2015, during an alien species focused survey of oyster beds in the marine lake Grevelingen, The Netherlands, a colony of *D. vexillum* was collected with a mussel dredge (coordinates: N51° 45.073, E3° 55.664). Directly after collection, the colony was preserved on ethanol 96% and stored in a refrigerator at 4 °C (GiMaRIS collection number AG4844). On 14 December 2015 eight little pieces, with a diameter of about 4 mm each, were cut out from this colony with a sterile scalpel for DNA-extraction. A Kingfisherflex robot was used to extract the DNA from these pieces with the Nucleomag Tissue kit from Macherey Nagel. To lyse the cells, 200µl T1 lysis buffer was added to the wells of a 96-well plate. Eight of these wells were used for the DNA extraction of *D. vexillum*. After adding a small piece of tissue, 25µl of Proteinase K (20mg/ml) was added and incubated at 56°C overnight. After the cells were lysed, 225µl of the sample was added to the MB2 plate containing 360µl MB2 binding buffer (35–55% ethanol, 20–40% sodium perchlorate) and 25µl Magnetic beads. The robot then mixed the mixture and transferred the DNA that was attached to the magnetic beads to a series of wash buffers (20–30% ethanol). The MB3 plate was filled with 600µl MB3 wash buffer, the MB4 plate with 600µl MB4 wash buffer, and the MB5 plate with 600µl MB5 wash buffer. To release the DNA from the magnetic beads, the robot proceeded after the wash buffer to the MB6 plate with 150µl MB6 elution buffer (5 mM Tris/HCl, pH 8.5). The DNA dilution was stored in the fridge at 4°C. With the Nanodrop ND1000 the quality and quantity was tested for each of the 8 DNA extractions. Based on these analyses two samples of 100µl each were selected and sent to University of Washington PacBio Sequencing Services for further analyses. Agilent 2100 Bioanalyzer platform was used for the quantitation and sizing of gDNA. PacBio sequencing started from 386ng/µl and 192ng/µl quantified by a Qubit assay for each sample respectively.

RNA extraction

A $\sim 10\text{cm}^2$ large piece of a *D. vexillum* colony was collected on 14/Dec/2009 from the upside of a settlement plate that was deployed about six months earlier on 25/Mar/2009 at a depth of 1 meter from the

south pier of the islet Hompelvoet (Grevelingen, The Netherlands) in an enclosed marine lake with minimal tidal differences. One piece of this colony was used for the first draft in 2016, [21], while another piece of the same colony was used for transcriptome analyses. This piece was preserved in RNAlater (Ambion) at -20°C prior to RNA extraction and subsequent sequencing in February 2010. Total RNA was extracted using the RNeasy kit according to manufacturers instructions (QIAGEN GmbH, Hilden). A transcriptome library was prepared from 10 mg total RNA, using the Illumina mRNA-Seq Sample Preparation Kit according to the manufacturers instructions (Illumina Inc., San Diego, USA). The mRNA-Seq library with a read length of 2×76 nucleotides was sequenced using the next generation sequencing apparatus Illumina GAIIx according to the manufacturers description at ZF-Screens.

Genome sequencing and data preprocessing

The Illumina data used for the genome assembly are described in more detail in [21]. The comprise a mix of paired end reads of 76 and 151 nt, respectively, with a total coverage of about $30\times$ obtained on a Illumina GAIIx instrument.

PacBio sequencing data was obtained using P6/C4 chemistry in an instrument PacBio RSII at University of Washington PacBio Sequencing Services. SMRT libraries of size 20 kb, 10 kb and 5kb were run on eleven SMRT cells and prepared without previous DNA shearing or size-selection [48] due to low DIN of the samples ($\text{DIN} \leq 3.8$). A total of 12.35 Gb sequence data obtained corresponds to 5 millions of subreads with $\text{N50} = 2.3$ Kbp. Size distribution of subreads sequenced is shown on (Figure 6).

We opted for a hybrid, non-conservative *de novo* assembly approach. Therefore, PacBio subreads and high-confidence Illumina paired end reads used previously to drafting the genome of *D. vexillum* [21] were used to collectively provide an improved new genome assembly. Before performing the assembly, PacBio sequence data were error corrected and pre-processed in three independent steps. First, PacBio reads of size $\geq 250\text{bp}$ and quality ≥ 0.83 were pre-assembled using the protocols RS_PreAssembler and RS_ReadsOfInsert implemented on SMRT pipe v.2.3. A total of 1.4 Gpb comprised of 823,758 pre-assembled reads (error-corrected reads) with $\text{N50} = 1.8\text{Kbp}$ were obtained. Second, a total of 450 Mbp distributed in 220,514 CCSs with quality $\geq 99\%$ and $\text{N50} = 2.1$ Kbp were obtained using RS_PreAssembler by processing PacBio reads of complete sequencing cycles ≥ 2 . Third PacBio subreads of size $> 150\text{bp}$ and quality > 0.87

were retrieved using **dextract**[49] to be error corrected by the alignments of Illumina PE reads using **Proovread-2.13.13**[50]. The module **ccseq** was utilized to improve correction performance. Then, 2.7 Gpb of sequence corrected data comprising of 776,295 of untrimmed error-corrected subreads $N50=3.4$ kbp and 391 Mbp of trimmed error-corrected subreads corresponding to 288,198 $N50=1.7$ kbp were obtained. Finally a total of (4.94 Gb) of error-corrected data were assembled. The size of the data used to be run on the new hybrid assembly is shown in Table 3.

Genome assembly

De novo hybrid assembly was performed using Celera Assembler Approach [51], Version 8.3rc2, without popping bubbles. Command-line parameters used were `utgErrorRate=0.12`, `utgErrorLimit=2.5`, `ovlErrorRate=0.15`, `cgwErrorRate=0.15` and `kmer=17`. A first version produced an assembly of 566.4 Mbp comprising 130,707 contigs with $N50 = 5.97$ kb and $GC = 36\%$. Summary of general steps followed to perform the genome assembly are shown in Figure 7. Duplicated contigs were filtered using **fasta2homozygous** [52]. A total of 16839 contigs of size ≤ 500 bp and similarity $\geq 95\%$ corresponding to 47.4 Mb were removed. Finally only 519 Mbp were subjected to genome scaffolding.

Genome Scaffolding

LoRDEc [53] was run to correct high quality CCS by processing together Illumina short reads and CCS subreads. Bruijn graphs were built with Illumina data using k-mers of size 13, 15, 17, 19, 21, 31, 41 and 51 and guided by 462,447 CCS subreads (985 Mb, $N50 = 2.1$ kb) retrieved by **unanimity** v.3.0 [54]. On a further step, error-corrected CCS aligned by **daligner** [55] were input into **daccord** [56] to get consensus of CCSs. Those error-corrected consensus CCSs and 519 Mbp of genome data assembled by Celera were used as input into **SSPACE-Long** [57] to genome scaffolding. The final assembly resulted in a 517.5 Mb genome sequence (109,769 scaffolds with $N50$ of 6.54 kb).

Assembly polishing

QUIVER v.2.1 [58] from the BAM.Resequencing Beta.1 SMRT pipe v2.3.0 was used to provide SNPs and high quality base calling for each scaffold.

Assessment of genome assembly quality

Genome assembly completeness was evaluated by Benchmarking Universal Single-Copy Orthologs

(BUSCO) [59], using the metazoan lineage data resulting in scores to be comparable with other tunicate species.

RNA data assembly

Illumina sequence data (PE reads of size 76 bp) were trimmed using BBtools [60]. After trimming a total of ~ 2.6 Mbp comprising of 55.1 millions of PE reads of size 50 bp and $\text{Phred} \geq 30$ were input to perform a genome-guided Trinity *de novo* transcriptome assembly using Trinity v2.4.0 [61]. Reads were first aligned to the reassembled genome of *D. vexillum* with Gmap (Version 2019-06-10) [62] to get groups of overlapping reads into clusters used for further steps for the *de novo* transcriptome assembly. Finally 39 Mbp comprising 90,938 transcripts were assembled and processed by TransDecoder [63] to find coding regions within transcripts.

Genome annotation

Gene structure were predicted using **Maker** v.3.01.02 [64] in two rounds. First Maker annotation round run Augustus 3.3 [65] and **RepeatMasker** version open-4.0.5 [66] with *Ciona robusta* models and RepBase (RepBase20.03) [67]. This first draft annotation was further improved on a second round by incorporating transcripts, peptide and filtered RNAseq raw data previously used to assembly *D. vexillum* transcriptome. Genomic variants previously detected by Quiver were used to guide SNAP [68] to align mRNAseq data. Besides, Repeatmodeler [69] was used to construct our *de novo* repeat annotation library which was used in combination with RepBase (RepBase20.03) by RepeatMasker to asses for the total repetitive elements content of *D. vexillum*. Semi-HMM-based Nucleic Acid Parser (version 2006-07-28), GeneMark, GeneMark.hmm eukaryotic, version 3.54 [70], Nucleotide-Nucleotide BLAST 2.4.0+ [71] were used in the steps of Maker annotations. GO annotation was assigned with Uniref90 [72] from Uniprot, PFAM 31.0 or RefSeq accessions. Finally eggNOG v.5 was used to identify clusters of Orthologous groups as described below.

Genome Browser construction

Resulting gene, ncRNAs and mtDNA annotations were generated as GFF3 files and were processed using MakeHub [73] as preprocessing step to generate the input files of the hub. The input files were used to create a genome Hub hosted on the UCSC hub site [45].

Identification of contamination along sequenced *D. vexillum* genome

A modification of the protocol described on [21] to detect possible contamination was performed and it is described in more detail on Additional File 1. A final number of 4 scaffolds were removed from the original genome assembly, it means ~ 18.65 kb of the final assembly. Those sequences were removed from the final available genome, reporting a final genome with: 109,769 scaffolds and 517.55 Mb.

Annotation of non-coding RNAs

Annotated ncRNA candidates from the first assembly of *D. vexillum* were mapped in the new assembly as described on Additional File 1. At the same time, homology **blastn** and HMM strategies with their corresponding metazoan-specific CMs and default CMs evaluation have been applied following the methodology proposed in [21], in order to annotated those candidates that have not been detected on the mapping strategy. The tRNAs genes were found using **tRNAscan-SE** v.2.0.3 using default parameters. Final check of candidates was performed to ensure that the reported families belongs from **Rfam** families that contains at least one sequence from Metazoa clade into their original seed alignment. These last step was performed to report possible false-positive families that could be retrieved applying the default **Rfam** models directly to the genome.

In order to annotate the position of *mature* sequences from miRNA candidates, **MIRfix** [33] was used. The **miRBase** (v.22) mature and hairpin sequences were used as initial sequences resource. Via **RNAcentral** database [74], the cross-link between **miRBase** and **Rfam** (v.14.1) was retrieved, and a list of **Rfam** families were classified as annotated in both databases. In this case, by **MIRfix** the reported *mature* position was corrected along their corresponding *hairpin* sequence. After that, the remaining *seed* sequences that have not been annotated on **miRBase** were included to be evaluated by the same methodology, but with the *mature* family-specific sequences. Final correction and annotation of those families, allowed the re-build of multiple sequence and structural alignments from the **Rfam** defined sequences, as a stockholm alignment. Given those results, the *D. vexillum* miRNA sequences annotated in this study, were processed as subject to annotate their *mature* sequences, based on previously detected matures in the **Rfam** family. At the end, the results were the positions of the most probable *mature* sequences plus the correspondent alignments in stockholm format for each miRNA **Rfam** family. Those genome annotations can be assessed on the described genome browser.

Computational validation of miRNAs

Precursors from predicted miRNAs were obtained in **fasta** format. The position of the mature sequence (s) from **Rfam** seed sequences were evaluated using **MIRfix** [33]. Based on the complete set of corrected seed sequences an evaluation of *D. vexillum* miRNAs were performed, miRNAs hairpins that contains a candidate mature annotation and are supported by the family structural alignment were considered as true candidates (for details see Additional File 1).

In order to identify the phylogenetic distribution of the **Rfam** sequences, the taxonomic distribution, generated by the annotated *kingdom*, *phylum* and *sub-phylum*, was retrieved for each of the species from the **Rfam** stockholm alignments from NCBI Taxonomy Browser^[1]. Once defined these groups, was possible to calculate the number of *phyla* that reported sequences on the annotated miRNAs from **Rfam**.

Mitochondrial genes

Mitochondrial complete genome from isolated clade A (NC_026107) and isolated clade B (KM259617.1) from *D. vexillum* were retrieved from **GenBank** as reported by [47]. Both set of sequences searched with **blastn** against new *D. vexillum* genome. Best candidates were retrieved adjusting identity $\leq 95\%$ and E-value ≤ 0.001 and coverage 100%. Final coordinate files was organized as **GFF3** file. Reduction of the intergenic coordinates was performed by a **Perl** script and this output was depicted with **LuaTeX** package **pgfmlbio**. Annotated Tunicata mitochondrial genomes were collected from NCBI. Multiple mitochondrial genome alignments were calculated using **progressiveMauve** [75] as referenced in Additional File 3.

Functional annotation of annotated proteins

Protein **fasta** files were obtained from **Ensembl** v.81: *C. savignyi*, *P. marinus*, *L. chalumnae*; **Aniseed** [18]: *C. robusta*, *B. schlosseri*, *M. oculata*, *M. occidentalis* and *B. leachii*. Proteins from *B. floridae* were retrieved in **JGI** [76] and for *O. dioica* from **Oikoaarrays** [77].

Functional annotation from all retrieved species were assessed using **eggNOG-Mapper** v.2 [78], based on the database **eggNOG** v.5 [36] applying **DIAMONT** mode as referred on [79].

Protein enrichment analysis

Enrichment analysis was calculated with **goatools** [80] taking as *background* group the

^[1]<https://www.ncbi.nlm.nih.gov/taxonomy>

complete set of proteins reported on studied chordata species and the *comparison* group, the list of proteins for each specie, the *association* file between proteins and GO, was generated based on **eggNOG-Mapper** results, all the command line methods are described on Additional File 3. Final results of enrichment were plotted using **ggplot2**, **tidyverse** [81, 82] and **grid** [83] R packages. TreeMap plots were performed with **REVIGO** [43]. Calculated *p-values* from **goatools** were used as input data to **REVIGO** webserver against whole **UniProt** and **SimRel** as semantic similarity measure.

Interaction analysis of proteins

Proteins that reported the same semantic terms on the **REVIGO** results were clustered and subject to a protein-protein interaction analysis using **STRING** (v.11) [44]. Proteins from *D. vexillum* were compared against the entire Chordata protein set. Since *C. robusta* was the species with the largest number of recognizable homologs, this species was used as reference. Only connected nodes with “high” or “highest” confidence were analyzed and plotted.

Annotation of *homeobox* proteins

A collection of reported *homeobox* proteins from human (from the family *Homeoboxes* (516)^[2] from HGNC database on 10 October, 2019 [84]), *C. robusta*, *C. savignyi* (both species from **Ensembl**), *B. leachii* [20], *H. roretzi* [29, 30] and a variety of species from the **HomeoDB** [85] were retrieved from the corresponding references. All the described subset was searched along the annotated transcriptome and proteins sequences from *D. vexillum* using **tblastn** and **blastp**, respectively. Best candidates were obtained if reported an identity percent ≥ 35 , E-value $\leq 10^{-5}$ and a query coverage of 70%. For command line details refer to Additional File 3.

As a complement, pairwise genome alignments with the new assembly from *D. vexillum* and close species that reported annotations of *homeobox* genes: *B. floridae*, *B. leachii*, *B. schlosseri*, *C. savignyi*, *C. robusta*, *H. roretzi* and *O. dioica*, were performed with **LASTZ** [86]. References from *homeobox* genes were obtained from **Aniseed** [18] using the Gene Builder with the term *hox*, except from *B. floridae* where updated annotations (from v.2) were searched and retrieved from **LanceletDB** [87]. Cross-matching of shared regions and reported genes and homology searches were performed to support the identification of *homeobox* candidates.

Detection of orthologous proteins involved in skeletogenesis

We searched for **RUNX**, **SOX**, and **Hh** homologs in the output of **eggNOG-Mapper** for all studied chordate species. The corresponding orthology groups have the accession numbers: **K0G3982**, **K0G0527** and **K0G3638**, respectively. Due to the lack of true runt orthologous on *D. vexillum*, we performed an additional analysis to confirm the presence of some homology signal. We retrieved the **RUNX** sequences reported on [88], from available 16 chordates from NCBI: AN08565.1, AAN08567.1, AAQ88389.1, AAS02047.1, AAS21356.1, BAA03485.1, BAF36001.1, BAF36011.1, EAX04278.1, EDL03777.1, EDL29993.1, ENSCINT00000004611.3, NP_001001890.1, NP_001092121.1, NP_004341.1 and NP_571678.1. Those sequences were searched with **blastp** into the proteome from *D. vexillum* and the following 10 species: *B. floridae*, *B. leachii*, *B. schlosseri*, *C. robusta*, *C. savignyi*, *M. oculata*, *M. occidentalis*, *O. dioica*, *P. marinus*, and *L. chalumnae*. On the other hand, the **PFAM** domain *Runt* (PF00853) was searched along all the reported proteomes from the described species using **hmmsearch** (**HMMER** v.3.1b1) [89]. Filtering was based on the *gathering score* reported by **PFAM** and a low E-value < 0.001 .

Phylogenetic analysis from mentioned proteins was performed on the set of orthologs from the described target species and their correspondent orthologous sequences that have been obtained after the **eggNOG-Mapper** analysis. As an outgroup, we obtained from NCBI, the following sequences for **RUNX**: NP_999779.1 and XP_781626.2; for **Hh**: FBpp0121221, KDR14772 and XP_008546836.1. For the analysis of **RUNX**, we included the reported sequences from the lamprey (*Lethenteron camtschaticum*) [42], annotated on NCBI with the following accession numbers: AJM44878.1, AJM44883.1 and AJM44886.1. The complete phylogenetic analyses were performed using the **ETE 3 Toolkit** [90], using Maximum Likelihood (ML) tree with the JTT+G+I, substitution model generating 100 bootstraps. Specific command line is described on Additional File 3. Gene IDs were replaced by “human readable” names in Figure 4. A version with the database IDs is provided in Additional File 4.

^[2]<https://www.genenames.org/cgi-bin/genegroup/download?id=516&type=branch>

Availability of data and material

The reported data can be accessed at <http://tunicatadvexillum.bioinf.uni-leipzig.de/Home.html>.

List of Abbreviations

GO gene ontology
miRNA microRNA
ML maximum likelihood
ncRNA non-coding RNA

Declarations

Acknowledgments

CIB acknowledges Universidad Nacional de Colombia for the time granted to do this research.

Funding

This work and the computational analysis were partially supported by the equipment donation from the German Academic Exchange Service-DAAD to the Faculty of Science at the Universidad Nacional de Colombia. The design and the complete project was supported by Colciencias (project no. 110165843196, contract 571-2014). FDB was supported by FAPESP grants 2019/06927-5, 2015/50164-5. PFS acknowledges support by the DAAD/CAPES exchange program, grant number 57390771 and the part funding from the German Federal Ministry for Education and Research (BMBF 031A538A (de.NBI/RBC)). CAVH acknowledges the support by DAAD scholarship: Forschungsstipendien-Promotionen in Deutschland, 2018/19 (Bewerbung 57299294). Publication costs are supported by the German Research Foundation (DFG) and Leipzig University within the program of Open Access Publishing. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Author's contributions

CIB, AAG, FDB, and PFS designed the study. AAG was responsible for gDNA and RNA extraction. EPR and CIB assembled the genome and transcriptome sequences. CAVH implemented the computational workflows and analyzed the data with contributions by JF and PFS. CAVH and PFS drafted the manuscript. All authors contributed to the interpretation of the data and the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

The colonies of *Didemnum vexillum* were collected in 2009 and 2015 during surveys conducted from ships of the responsible national authorities (Staatsbosbeheer & the Dutch ministry). They were conducted as part of the continuous monitoring, i.e. within the SETL-project run by GiMaRIS, and the detection of marine invasive species in the Grevelingen. No further permits or permissions were required to collect specimens according to institutional, national, or international laws or guidelines. The results of the surveys concerned are described in reports on alien fouling organisms issued by the *Office for Risk Assessment and Research of the Netherlands Food and Consumer Product Safety Authority, of the Dutch Ministry of Economic Affairs* [91, 92].

Consent for publication

Not applicable.

Author details

¹Biology Department, Universidad Nacional de Colombia, Carrera 45 # 26-85, Edif. Uriel Gutiérrez, Bogotá D.C., Colombia. ²Institute of Biology, Leiden University, P.O. Box 9505, 2300 RA Leiden, Netherlands. ³GiMaRIS, Rijksweg 75, 2171ak, Sassenheim, Netherlands. ⁴Naturalis Biodiversity Center, Darwinweg 2, 2333 CR, Leiden, The

Netherlands. ⁵Departamento de Zoologia, Instituto Biociências, Universidade de São Paulo, Rua do Matão, Tr. 14 no. 101, São Paulo, SP, Brazil. ⁶Centro de Biologia Marinha, Universidade de São Paulo, Rod. Manuel Hypólito do Rego km. 131.5, Praia do Cabelo Gordo, São Sebastião, Brazil. ⁷Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstraße 16–18, D-04107 Leipzig, Germany. ⁸Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany. ⁹Department of Theoretical Chemistry, University of Vienna Währinger Straße 17, A-1090 Vienna, Austria. ¹⁰Santa Fe Institute, 1399 Hyde Park Rd., NM87501 Santa Fe, USA.

References

- Kott, P.: A complex didemnid ascidian from Whangamata, New Zealand. *J. Marine Biol. Assoc. UK* **82**, 625–628 (2002). doi:10.1017/S0025315402005970
- Lambert, G.: Adventures of a sea squirt sleuth: unraveling the identity of *Didemnum vexillum*, a global ascidian invader. *Aquatic Invasions* **4**, 5–28 (2009). doi:10.3391/ai.2009.4.1.2
- Valentine, P.C., Carman, M.R., Blackwood, D.S., Heffron, E.J.: Ecological observations on the colonial ascidian *Didemnum sp.* in a New England tide pool habitat. *J. Exp. Marine Biol. Ecol.* **342**, 109–121 (2007). doi:10.1016/j.jembe.2006.10.021
- Ordóñez, V., Pascual, M., Fernández-Tejedor, M., Pineda, M.C., Tagliapietra, D., Turon, X.: Ongoing expansion of the worldwide invader *Didemnum vexillum* (Ascidacea) in the Mediterranean Sea: high plasticity of its biological cycle promotes establishment in warm waters. *Biological Invasions* **17**, 2075–2085 (2015). doi:10.1007/s10530-015-0861-z
- Bullard, S.G., Sedlack, B., Reinhardt, J.F., Litty, C., Gareau, K., Whitlatch, R.B.: Fragmentation of colonial ascidians: Differences in reattachment capability among species. *J. Exp. Marine Biol. Ecol.* **342**, 166–168 (2007). doi:10.1016/j.jembe.2006.10.034
- Cottier-Cook, E.J., Minchin, D., Geisler, R., Graham, J., Mogg, A., Sayer, M.D., Matejusova, I.: Biosecurity implications of the highly invasive carpet sea-squirt *Didemnum vexillum* kott, 2002 for a protected area of global significance. *Management Biol. Invasions* **10**, 311–323 (2019). doi:10.3391/mbi.2019.10.2.07
- Bullard, S.G., Lambert, G., Carman, M.R., Byrnes, J., Whitlatch, R.B., Ruiz, G., Miller, R.J., Harris, L., Valentine, P.C., Collie, J.S., Pederson, J., McNaught, D.C., Cohen, A.N., Asch, R.G., Dijkstra, J., Heinonen, K.: The colonial ascidian *Didemnum sp.* A: Current distribution, basic biology and potential threat to marine communities of the northeast and west coasts of North America. *J. Exp. Marine Biol. Ecol.* **342**, 99–108 (2007). doi:10.1016/j.jembe.2006.10.020
- 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.*: The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**, 2157–2167 (2002). doi:10.1126/science.1080049
- Small, K.S., Brudno, M., Hill, M.M., Sidow, A.: A haplome alignment and reference sequence of the highly polymorphic *Ciona savignyi* genome. *Genome biology* **8**(3), 41 (2007). doi:10.1186/gb-2007-8-3-r41
- Hill, M.M., Broman, K.W., Stupka, E., Smith, W.C., Jiang, D., Sidow, A.: The *C. savignyi* genetic map and its integration with the reference sequence facilitates insights into chordate genome evolution. *Genome Res.* **18**, 1369–1379 (2008). doi:10.1101/gr.078576.108
- Satou, Y., Nakamura, R., Yu, D., Yoshida, R., Hamada, M., Fujie, M., Hisata, K., Takeda, H., Satoh, N.: A nearly complete genome of *Ciona intestinalis* type A (*C. robusta*) reveals the contribution of inversion to chromosomal evolution in the genus *Ciona*. *Genome Biol. Evol.* **11**, 3144–3157 (2019). doi:10.1093/gbe/evz228
- Colombero, D., Fenaux, R.: Chromosome form and number in the Larvacea. *Boll. Zoologia* **40**, 347–353 (1973). doi:10.1080/11250007309429248
- 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.*: Miniature genome in the marine chordate *Oikopleura dioica*. *Science* **294**, 2506–2506 (2001). doi:10.1126/science.294.5551.2506
- 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.*: Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* **330**, 1381–1385 (2010). doi:10.1126/science.1194167
15. Jue, N.K., Batta-Lona, P.G., Trusiak, S., Obergfell, C., Bucklin, A., O'Neill, M.J., O'Neill, R.J.: 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 (2016). doi:10.1093/gbe/evw215
 16. Wang, K., Dantec, C., Lemaire, P., Onuma, T.A., Nishida, H.: Genome-wide survey of miRNAs and their evolutionary history in the ascidian, *Halocynthia roretzi*. *BMC Genomics* **18**, 314 (2017). doi:10.1186/s12864-017-3707-5
 17. Stolfi, A., Lowe, E.K., Racioppi, C., Ristoratore, F., Brown, C.T., Swalla, B.J., Christiaen, L.: Divergent mechanisms regulate conserved cardiopharyngeal development and gene expression in distantly related ascidians. *eLife* **3**, 03728 (2014). doi:10.7554/eLife.03728
 18. Brozovic, M., Dantec, C., Dardaillon, J., Dauga, D., Faure, E., Gineste, M., Louis, A., Naville, M., Nitta, K.R., Piette, J., Reeves, W., Scornavacca, C., Simion, P., Vincentelli, R., Bellec, M., Aicha, S.B., Fagotto, M., Guérault-Bellone, M., Haeussler, M., Jacox, E., Lowe, E.K., Mendez, M., Roberge, A., Stolfi, A., Yokomori, R., Brown, C., Cambillau, C., Christiaen, L., Delsuc, F., Douzery, E., Dumollard, R., Kusakabe, T., Nakai, K., Nishida, H., Satou, Y., Swalla, B., Veeman, M., Volff, J.-N., Lemaire, P.: ANISEED 2017: extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets. *Nucleic Acids Res.* **46**, 718–725 (2017). doi:10.1093/nar/gkx1108
 19. Voskoboinik, A., Neff, N.F., Sahoo, D., Newman, A.M., Pushkarev, D., Koh, W., Passarelli, B., Fan, H.C., Mantalas, G.L., Palmeri, K.J., Ishizuka, K.J., Gissi, C., Griggio, F., Ben-Shlomo, R., Corey, D.M., Penland, L., White 3rd, R.A., Weissman, I.L., Quake, S.R.: The genome sequence of the colonial chordate, *Botryllus schlosseri*. *Elife* **2**, 00569 (2013). doi:10.7554/eLife.00569
 20. Blanchoud, S., Rutherford, K., Zondag, L., Gemmell, N.J., Wilson, M.J.: *De novo* draft assembly of the *Botryllodes leachii* genome provides further insight into tunicate evolution. *Sci. Rep.* **8**, 5518 (2018). doi:10.1038/s41598-018-23749-w
 21. Velandia-Huerto, C.A., Gittenberger, A., Brown, F.D., Stadler, P.F., Bermúdez-Santana, C.I.: Automated detection of ncRNAs in the draft genome sequence of a basal chordate: The carpet sea squirt *Didemnum vexillum*. *BMC Genomics* **17**, 591 (2016). doi:10.1186/s12864-016-2934-5
 22. Sasakura, Y., Ogura, Y., Treen, N., Yokomori, R., Park, S.-J., Nakai, K., Saiga, H., Sakuma, T., Yamamoto, T., Fujiwara, S., Yoshida, K.: Transcriptional regulation of a horizontally transferred gene from bacterium to chordate. *Proc. Roy. Soc. B: Biol. Sci.* **283**, 20161712 (2016). doi:10.1098/rspb.2016.1712
 23. Delsuc, F., Brinkmann, H., Chourrout, D., Philippe, H.: Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965–968 (2006). doi:10.1038/nature04336
 24. Putnam, N.H., Butts, T., Ferrier, D.E., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.K., Benito-Gutiérrez, E.L., Dubchak, I., García-Fernández, J., Gibson-Brown, J.J., Grigoriev, I.V., Horton, A.C., de Jong, P.J., Jurka, J., Kapitonov, V.V., Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegawa, K., Pennacchio, L.A., Salamov, A.A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin-I, T., Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L.Z., Holland, P.W., Satoh, N., Rokhsar, D.S.: The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071 (2008). doi:10.1038/nature06967
 25. Berná, L., Alvarez-Valín, F.: Evolutionary genomics of fast evolving tunicates. *Genome Biol. Evol.* **6**, 1724–1738 (2014). doi:10.1093/gbe/evu122
 26. Hirose, E.: Acid containers and cellular networks in the ascidian tunic with special remarks on ascidian phylogeny. *Zoological Science* **18**, 723–731 (2001). doi:10.2108/zsj.18.723
 27. Shapiro, R., Klein, R.S.: The deamination of cytidine and cytosine by acidic buffer solutions. mutagenic implications. *Biochemistry* **5**, 2358–2362 (1966). doi:10.1021/bi00871a026
 28. Lindahl, T., Nyberg, B.: Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13**, 3405–3410 (1974). doi:10.1021/bi00713a035
 29. Sekigami, Y., Kobayashi, T., Omi, A., Nishitsuji, K., Ikuta, T., Fujiyama, A., Satoh, N., Saiga, H.: Hox gene cluster of the ascidian, *Halocynthia roretzi*, reveals multiple ancient steps of cluster disintegration during ascidian evolution. *Zoological Lett.* **3**, 17 (2017). doi:10.1186/s40851-017-0078-3
 30. Sekigami, Y., Kobayashi, T., Omi, A., Nishitsuji, K., Ikuta, T., Fujiyama, A., Satoh, N., Saiga, H.: Note to: Hox gene cluster of the ascidian, *Halocynthia roretzi*, reveals multiple ancient steps of cluster disintegration during ascidian evolution. *Zoological Lett.* **5**, 8 (2019). doi:10.1186/s40851-019-0121-7
 31. Marz, M., Kirsten, T., Stadler, P.F.: Evolution of spliceosomal snRNA genes in metazoan animals. *J. Mol. Evol.* **67**, 594–607 (2008)
 32. Menzel, P., Gorodkin, J., Stadler, P.F.: The tedious task of finding homologous non-coding RNA genes. *RNA* **15**, 2075–2082 (2009)
 33. Yazbeck, A.M., Stadler, P.F., Tout, K., Fallmann, J.: Automatic curation of large comparative animal microRNA data sets. *Bioinformatics* **35**, 4553–4559 (2019). doi:10.1093/bioinformatics/btz271
 34. Velandia-Huerto, C.A., Brown, F.D., Gittenberger, A., Stadler, P.F., Bermúdez-Santana, C.I.: Nonprotein-coding RNAs as regulators of development in tunicates. In: Kloc, M., Kubiak, J. (eds.) *Marine Organisms as Model Systems in Biology and Medicine. Results Probl. Cell Differ.*, vol. 65, pp. 197–225. Springer, Cham (2018). doi:10.1007/978-3-319-92486-1_11
 35. Chodroff, R.A., Goodstadt, L., Sirey, T.M., Oliver, P.L., Davies, K.E., Green, E.D., Molnár, Z., Ponting, C.P.: Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol.* **11**, 72 (2010). doi:10.1186/gb-2010-11-7-r72
 36. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L., vonMering, C., Bork, P.: eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, 309–314 (2018). doi:10.1093/nar/gky1085
 37. Obradovic Wagner, D., Aspenberg, P.: Where did bone come from? *Acta Orthopaedica* **82**, 393–398 (2011). doi:10.3109/17453674.2011.588861
 38. Guth, S.I.E., Wegner, M.: Having it both ways: Sox protein function between conservation and innovation. *Cell. Mol. Life Sci.* **65**, 3000–3018 (2008). doi:10.1007/s00018-008-8138-7
 39. Takatori, N., Satou, Y., Satoh, N.: Expression of hedgehog genes in *Ciona intestinalis* embryos. *Mech. Dev.* **116**, 235–238 (2002). doi:10.1016/S0925-4773(02)00150-8
 40. Shimeld, S.M.: The evolution of the hedgehog gene family in chordates: insights from amphioxus hedgehog. *Dev. Genes Evol.* **209**, 40–47 (1999). doi:10.1007/s004270050225
 41. Ingham, P.W., McMahon, A.P.: Hedgehog signaling in animal development: paradigms and principles. *Genes Dev.* **15**, 3059–3087 (2001). doi:10.1101/gad.938601
 42. Nah, G.S.S., Tay, B.-H., Brenner, S., Osato, M., Venkatesh, B.: Characterization of the Runx gene family in a jawless vertebrate, the Japanese lamprey (*Lethenteron japonicum*). *PLOS ONE* **9**, 113445 (2014). doi:10.1371/journal.pone.0113445
 43. Supek, F., Bošnjak, M., Škunca, N., Šmuc, T.: REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLOS ONE* **6**, 1–9 (2011). doi:10.1371/journal.pone.0021800
 44. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., Jensen, L.J., Mering, C.: STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, 607–613 (2018). doi:10.1093/nar/gky1131
 45. Raney, B.J., Dreszer, T.R., Barber, G.P., Clawson, H., Fujita, P.A., Wang, T., Nguyen, N., Paten, B., Zweig, A.S., Karolchik, D., Kent, W.J.: Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC genome browser. *Bioinformatics* **30**, 1003–1005 (2013). doi:10.1093/bioinformatics/btt637
 46. Stolfi, A., Sasakura, Y., Chalopin, D., Satou, Y., Christiaen, L.,

- Dantec, C., Endo, T., Naville, M., Nishida, H., Swalla, B.J., Volff, J.-N., Voskoboinik, A., Dauga, D., Lemaire, P.: Guidelines for the nomenclature of genetic elements in tunicate genomes. *Genesis* **53**, 1–14 (2015). doi:10.1002/dvg.22822
47. Smith, K.F., Abbott, C.L., Saito, Y., Fidler, A.E.: Comparison of whole mitochondrial genome sequences from two clades of the invasive ascidian, *Didemnum vexillum*. *Marine Genomics* **19**, 75–83 (2015). doi:10.1016/j.margen.2014.11.007
 48. Wang, S., Harting, J., Tseng, E., Baybayan, P.: Getting the most out of your PacBio libraries with size selection. Technical report, PacBio (2019)
 49. Myers, G.: DEXTRACTOR. <https://github.com/thegenemyers/DEXTRACTOR> (2015)
 50. Förster, F., Schultz, J., Hedrich, R., Hackl, T.: proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004–3011 (2014). doi:10.1093/bioinformatics/btu392
 51. Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H.J., Remington, K.A., Anson, E.L., Bolanos, R.A., Chou, H.H., Jordan, C.M., Halpern, A.L., Lonardi, S., Beasley, E.M., Brandon, R.C., Chen, L., Dunn, P.J., Lai, Z., Liang, Y., Nusskern, D.R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G.M., Adams, M.D., Venter, J.C.: A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000). doi:10.1126/science.287.5461.2196
 52. Pryszcz, L.P., Gabaldón, T.: Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* **44**, 113–113 (2016). doi:10.1093/nar/gkw294
 53. Salmela, L., Rivals, E.: LoRDEC: accurate and efficient long read error correction. *Bioinformatics* **30**, 3506–3514 (2014). doi:10.1093/bioinformatics/btu538
 54. PacBio: Unanimity. <https://github.com/PacificBiosciences/ccs> (2015)
 55. Myers, E.: DALIGNER. <https://github.com/thegenemyers/DALIGNER> (2016)
 56. Tischler, G., Myers, E.W.: Non hybrid long read consensus using local de Bruijn graph assembly. *bioRxiv*, 106252 (2017). doi:10.1101/106252
 57. Boetzer, M., Pirovano, W.: SSPACE-longread: scaffolding bacterial draft genomes using long read sequence information. *BMC bioinformatics* **15**, 211 (2014). doi:10.1186/1471-2105-15-211
 58. PacBio: QUIVER. <https://github.com/PacificBiosciences/GenomicConsensus> (2016)
 59. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M.: BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015). doi:10.1093/bioinformatics/btv351
 60. Bushnell, B.: BBmap. <https://sourceforge.net/projects/bbmap> (2016)
 61. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A.: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotech.* **29**, 644 (2011). doi:10.1038/nbt.1883
 62. Wu, T.D., Reeder, J., Lawrence, M., Becker, G., Brauer, M.J.: GMAP and GSNAP for genomic sequence alignment: Enhancements to speed, accuracy, and functionality. In: Mathé, E., Davis, S. (eds.) *Statistical Genomics: Methods and Protocols*, pp. 283–334. Springer, New York, NY (2016). doi:10.1007/978-1-4939-3578-9_15
 63. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A.: *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**, 1494–1512 (2013). doi:10.1038/nprot.2013.084
 64. Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A.S., Yandell, M.: MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008). doi:10.1101/gr.6743907
 65. Stanke, M., Schöffmann, O., Morgenstern, B., Waack, S.: Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006). doi:10.1186/1471-2105-7-62
 66. Smit, A.F.A., Hubley, R., Green, P.: RepeatMasker Open-4.0. 2013–2015. <http://www.repeatmasker.org> (2015)
 67. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J.: Repbase Update, a database of eukaryotic repetitive elements. *Cytogen. Genome Res.* **110**, 462–467 (2005). doi:10.1159/000084979
 68. Korf, I.: Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004). doi:10.1186/1471-2105-5-59
 69. Smit, A.F., Hubley, R.: RepeatModeler Open-1.0. <http://www.repeatmasker.org> (2008)
 70. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., Borodovsky, M.: Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005). doi:10.1093/nar/gki937
 71. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990). doi:10.1016/S0022-2836(05)80360-2
 72. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., UniProt Consortium: UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2014). doi:10.1093/bioinformatics/btu739
 73. Hoff, K.J.: MakeHub: Fully automated generation of UCSC genome browser assembly hubs. *Genomics Proteomics Bioinformatics* **17**, 546–549 (2019). doi:10.1016/j.gpb.2019.05.003
 74. The RNAcentral Consortium: RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res.* **47**, 221–229 (2018). doi:10.1093/nar/gky1034
 75. Darling, A.E., Mau, B., Perna, N.T.: progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLOS ONE* **5**, 11147 (2010). doi:10.1371/journal.pone.0011147
 76. Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I.V., Dubchak, I.: The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.* **42**, 26–31 (2013). doi:10.1093/nar/gkt1069
 77. Danks, G., Campsteijn, C., Parida, M., Butcher, S., Doddapaneni, H., Fu, B., Petrin, R., Metpally, R., Lenhard, B., Wincker, P., Chourrout, D., Thompson, E.M., Manak, J.R.: OikoBase: a genomics and developmental transcriptomics resource for the urochordate *Oikopleura dioica*. *Nucleic Acids Res.* **41**, 845–853 (2012). doi:10.1093/nar/gks1159
 78. Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., Bork, P.: Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017). doi:10.1093/molbev/msx148
 79. Buchfink, B., Xie, C., Huson, D.H.: Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60 (2014). doi:10.1038/nmeth.3176
 80. Klopfenstein, D.V., Zhang, L., Pedersen, B.S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C.J., Yunes, J.M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., Tang, H.: GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 10872 (2018). doi:10.1038/s41598-018-28948-z
 81. Wickham, H.: ggplot2: Elegant Graphics for Data Analysis. (2016). <http://ggplot2.org>
 82. Wickham, H.: tidyverse: Easily Install and Load the 'Tidyverse'. (2017). R package version 1.2.1. <https://CRAN.R-project.org/package=tidyverse>
 83. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2019). <https://www.R-project.org/>
 84. Yates, B., Braschi, B., Gray, K.A., Seal, R.L., Tweedie, S., Bruford, E.A.: Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.* **45**, 619–625 (2016). doi:10.1093/nar/gkw1033
 85. Zhong, Y.-F., Butts, T., Holland, P.W.H.: HomeoDB: a database of homeobox gene diversity. *Evol. Dev.* **10**, 516–518 (2008). doi:10.1111/j.1525-142X.2008.00266.x

86. Harris, R.S.: Improved pairwise alignment of genomic DNA. PhD thesis, Pennsylvania State University, University Park, PA, USA (2007)
87. You, L., Chi, J., Huang, S., Yu, T., Huang, G., Feng, Y., Sang, X., Gao, X., Li, T., Yue, Z., Liu, A., Chen, S., Xu, A.: LanceletDB: an integrated genome database for lancelet, comparing domain types and combination in orthologues among lancelet and other species. *Database* **2019** (2019). doi:10.1093/database/baz056
88. Hecht, J., Stricker, S., Wiecha, U., Stiege, A., Panopoulou, G., Podsiadlowski, L., Poustka, A.J., Dieterich, C., Ehrich, S., Suvorova, J., Mundlos, S., Seitz, V.: Evolution of a core gene network for skeletogenesis in chordates. *PLOS Genetics* **4**, 1000025 (2008). doi:10.1371/journal.pgen.1000025
89. Eddy, S.R.: Accelerated profile HMM searches. *PLOS Comp. Biol.* **7**, 1002195 (2011). doi:10.1371/journal.pcbi.1002195
90. Huerta-Cepas, J., Serra, F., Bork, P.: ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016). doi:10.1093/molbev/msw046
91. Gittenberger, A., Rensing, M., Niemantsverdriet, P., Schrieken, A. N. D'Hont, Stegenga, H.: Soorteninventarisatie oesterputten en oesterpercelen, 2015. I.o.v. bureau risicobeoordeling & onderzoeksprogrammering, nederlandse voedsel- en warenautoriteit, ministerie van economische zaken. Technical Report 2015_19, GiMaRIS (2015). <https://www.nvwa.nl/binaries/nvwa/documenten/dier/dieren-in-de-natuur/exoten/risicobeoordelingen/soorteninventarisatie-oesterpunten-en-oesterpercelen-2015-gimaris/Soorteninventarisatie+oesterpunten+en+oesterpercelen+2015%2C+GiMaRIS.pdf>
92. Gittenberger, A., Wesdorp, K.H., Rensing, M.: Biofouling as a transport vector of non-native marine species in the Dutch Delta, along the North Sea coast and in the Wadden Sea. commissioned by Office for Risk Assessment and Research, the Netherlands Food and Consumer Product Safety Authority. Technical Report 2017_09, GiMaRIS (2017). http://www.gimaris.com/Portals/3/GiMaRIS.2017_09.Hullfouling.pdf

Figures

Figure 6 Raw data heatmap of PacBio sequences showing the density distribution of subreads produced on PacBio sequencing. X-axis: subread length, Y axis: read quality (RQ) score.

Figure 7 General procedure of the hybrid assembly of *D. vexillum* using error-corrected subreads. Numbers 1, 2, 3 and 4 correspond to the data size (shown in Table 3).

Table 1 Comparison of first draft [21] and the new draft assembly of the *D. vexillum* genome.

Assembly	Estimated Size (kb)	Number contigs (c) / Scaffolds (s)	L50	N50	GC content	IUPAC	Gene Number	Protein Number
Draft [21]	542,259	882,106 (c)	152,090	918	0.366 ± 0.063	0.000	N/A	N/A
This work	517,553	109,769 (s)	25,281	6539	0.362 ± 0.024	0.0155	62,194	64,424

Table 2 Annotated ncRNAs families and *loci* (in parentheses) in the *D. vexillum* genome. *Homology* corresponds to previously reported numbers of ncRNAs by homology [21], *Mapped* corresponds to the number of ncRNAs that were mapped in the first genome draft [21]. *Final* corresponds to the current list of candidate ncRNAs. *NA*: Not available.

ncRNA Family	Homology	Mapped	Final
Cis-Reg	3 (333)	0	3 (333)
miRNAs	248 (2065)	17 (20)	235 (1582)
misc RNAs	1 (1)	1 (1)	2 (2)
lncRNAs	2 (8)	0	2 (8)
Ribozyme	3 (11)	0	3 (11)
rRNAs	4 (84)	0	4 (84)
snoRNAs	6 (9)	6 (9)	12 (18)
snRNAs	9 (87)	2 (34)	9 (115)
tRNAs	23 (2724)	NA	23 (2724)
↪ mt-tRNAs	0	21	21
↪ mt-rRNAs	0	2	2
Total	277 (5322)	26 (64)	271 (4877)

Table 3 Data used to be run on the new hybrid assembly

	Correction	Software	Reads	N50	Size
1	Hybrid	Proovread	776,295	3.94kbp	2.7Gp
2	Hybrid	Proovread	288,198	1.7kbp	391Mbp
3	Pre-assembled reads	SMRT pipe	823,758	1.8Kbp	1.4Gpb
4	CCS	SMRT pipe	220,514	2.1Kbp	450 Mbp

Additional Files

- **Additional File 1:** Genome annotation methodology.
- **Additional File 2:** Orthologs and Protein GO enrichment analysis.
- **Additional File 3:** Command line methods.
- **Additional File 4:** List of SOX, RUNX and Hedgehog sequence names and their accession numbers used for phylogenetic analysis.

Figures

Figure 1

Distribution of estimated genome size and GC content of Deuterostome taxa. We included Hemichordata (*S. kowalevskii*, Saco), Echinodermata (*P. miniata* Pami and *S. purpuratus* Stpu) and Chordata species. Filled circles are tunicates and their lifestyle, colonial or solitary, is highlighted in orange or blue accordingly. Species labels: *B. oridae* (Br), *B. belcheri* (Brbe), *O. dioica* (Oidi), *M. occidentalis* (Mlis), *M. oculata* (Mata), *M. occulta* (MLta), *B. schlosseri* (Bosc), *H. roretzi* (Harro), *S. thompsoni* (Sath), *B. leachii* (Bole), *D. vexillum* (Dive), *C. robusta* (Ciro), *C. savignyi* (Cisa), *P. marinus* (Pema), *D. rerio* (Dare) and *L. chalumnae* (Lach).

BUSCO Assessment Results

Figure 2

Completeness of tunicate genomes assessed by BUSCO [59] in comparison to metazoan orthologs.

Figure 3

Detection of Homeobox genes on *D. vexillum*. A. Model of detection, a shared region between genomes A and B is detected and referenced as gray boxes. Correspondence is denoted by dotted lines between genomes. The dark gray box in genome B represents an annotated gene whereas the dark gray box mark represents the putative orthologous region. Figures B-D show examples of putative orthologous Hox gene assignment in *D. vexillum*. Specific details are explained in the main text. Figure E summarize the

complete Homeobox genes annotation in *D. vexillum* (Dive) in comparison to reported genes on *C. robusta* (Ciro) and *H. roretzi* (Haro). Genomic locations were retrieved from ANISEED, Hox cluster of the chordate ancestor is depicted [29, 30]. Uncertain positions of some genes are represented as a dotted box, eg. Hox1 and Hox4 in *H. roretzi*. For specific genome coordinates see Additional File 1: Table 3.

Figure 4

Phylogenetic analysis of skeletogenesis proteins found in *D. vexillum*. A. SoxB1/B2 family, B Hh family. The sea vomit is highlighted in gray. A tree of the complete SOX family can be found in Additional File 2: Figure 3. Trees were built using Maximum Likelihood (ML) with the JTT+G+I substitution model generating 100 bootstrap replicates.

Figure 5

Comparative genome analyses of *D. vexillum* predicted proteins. A. EggNOG functional annotation of orthologous groups across chordate genomes. B. TreeMap representation from REVIGO [43] *D. vexillum* of the enriched GO specific terms. Boxes' area correspond to the adjusted p-value, calculated with goatoools [80]. C. Functional interactions of homologous proteins in *C. robusta* which have shown homology with the functionally annotated proteins on *D. vexillum* with tRNA catabolism processes. Nodes correspond to single, protein-coding locus. Edges does not represent physically binding, but functional association determined by STRING [44]. The legend was obtained and modied from web server (<https://string-db.org/>).

Figure 6

Raw data heatmap of PacBio sequences showing the density distribution of subreads produced on PacBio sequencing. X-axis: subread length, Y axis: read quality (RQ) score.

Figure 7

General procedure of the hybrid assembly of *D. vexillum* using error-corrected subreads. Numbers 1, 2, 3 and 4 correspond to the data size (shown in Table 3).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplemental1.pdf](#)
- [supplemental2.pdf](#)

- [supplemental3.pdf](#)
- [Accessionnumbers.txt](#)