bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

De novo draft assembly of the *Botrylloides leachii* genome provides further insight into tunicate evolution.

Simon Blanchoud[1#], Kim Rutherford[1], Lisa Zondag[1], Neil J. Gemmell[1] and Megan J. Wilson[1*]

1 Department of Anatomy, School of Biomedical Sciences, University of Otago, P.O. Box 56, Dunedin 9054, New Zealand

Current address: Department of Zoology, University of Fribourg, Switzerland

* Corresponding author:

Email: meganj.wilson@otago.ac.nz

Ph. +64 3 4704695

Fax: +64 479 7254

Keywords: chordate, regeneration, *Botrylloides leachii*, ascidian, tunicate, genome, evolution

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract (250 words)

Tunicates are marine invertebrates that compose the closest phylogenetic group to the vertebrates. This chordate subphylum contains a particularly diverse range of reproductive methods, regenerative abilities and life-history strategies. Consequently, tunicates provide an extraordinary perspective into the emergence and diversity of chordate traits. To gain further insights into the evolution of the tunicate phylum, we have sequenced the genome of the colonial Stolidobranchian *Botrylloides leachii*.

We have produced a high-quality (90 % BUSCO genes) 159 Mb assembly, containing 82 % of the predicted total 194 Mb genomic content. The *B. leachii* genome is much smaller than that of *Botryllus schlosseri* (725 Mb), but comparable to those of *Ciona robusta* and *Molgula oculata* (both 160 Mb). We performed an orthologous clustering between five tunicate genomes that highlights sets of genes specific to some species, including a large group unique to colonial ascidians with gene ontology terms including cell communication and immune response.

By analysing the structure and composition of the conserved gene clusters, we identified many examples of multiple cluster breaks and gene dispersion, suggesting that several lineage-specific genome rearrangements occurred during tunicate evolution. In addition, we investigate lineage-specific gene gain and loss within the Wnt, Notch and retinoic acid pathways. Such examples of genetic change within these highly evolutionary conserved pathways commonly associated with regeneration and development may underlie some of the diverse regenerative abilities observed in the tunicate subphylum. These results supports the widely held view that tunicate genomes are evolving particularly rapidly.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Tunicates are a group of marine suspension-feeding hermaphrodites found worldwide in the inter- or sub-tidal region of the seas. This subphylum of invertebrates is part of the Chordata phylum, phylogenetically positioned between the more basal Cephalochordata and the higher Vertebrata, of which they are considered the closest relatives (Fig. 1A; (Delsuc et al. 2006)). These organisms include a wide range of reproductive methods, regenerative abilities, developmental strategies and life cycles (Lemaire et al. 2008). Importantly, and despite a drastically different body plan during their adult life cycle, tunicates have a tissue complexity incipient to that of vertebrates (Fig. 1A), including a heart, a notochord, an endostyle and a vascular system (Millar 1971). In addition, this group of animals is undergoing rapid genomic evolution compared to higher vertebrates, with a greater nucleotide substitution rate observed in both their nuclear and mitochondrial genomes (Tsagkogeorga et al. 2010, 2012; Rubinstein et al. 2013; Berna and Alvarez-Valin 2014). Therefore, this chordate subphylum provides an excellent opportunity to study the origin of vertebrates, the emergence of clade specific traits and the function of conserved molecular mechanisms. Biological features that can be investigated in tunicates include, among others, the evolution of colonialism, neoteny, sessilness, and budding. However, there are currently only seven Tunicata genomes publicly available, of which three have been well annotated. There is thus a paucity in the sampling of this very diverse subphylum.

Tunicates are separated into seven orders contained in three classes (Fig. 1A): Appendicularia (order Copelata), Thaliacea (orders Pyrosomida, Salpida and Doliolida) and Ascidiacea (orders Aplousobranchia, Phlebobranchia and Stolidobranchia). Appendicularia is a class of planktonic free-swimming organisms that possess chordate traits common to all tunicate larvae including a notochord, neural tube and pharyngeal slits. These social neotenic animals form dioecious communities where each individual lives inside a special external

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

63 mucous structure, termed house, which concentrates and funnels their food. *Oikopleura dioica*
64 is the sole example of the Appendicularian to have its genome sequenced, showing
65 exceptional compaction (70 Mb; (Seo et al. 2001)). Whether these animals can undergo
66 regeneration has not yet been assessed.
67 Thaliacea is an order of planktonic pelagic animals forming cylindrical free-floating
68 compound colonies (Piette and Lemaire 2015). These organisms can reproduce both sexually,
69 through autogamy to initiate novel colonies, as well as asexually, through stolonial budding to
70 increase the size of the colony. Owing to their peculiar life cycle and habitat, these tunicates
71 have not been extensively studied, no genome has been sequenced and whether they can
72 undergo regeneration remains unknown.
73 Ascidiacea consist of both solitary and colonial sessile benthic organisms. Solitary
74 ascidians (Phlebobranchian and some families among the Stolidobranchian) reproduce
75 sexually, releasing eggs through their atrial siphon for external fertilization, hence producing
76 a motile larva. These larvae will colonize novel environments, attach to a submersed substrate
77 and undergo metamorphosis into a sessile filter-feeding adult. These ascidians are capable of
78 regenerating a limited set of organs, including their oral siphon (Auger et al. 2010) although
79 regeneration capability reduces as they age (Jeffery 2015). These are currently the most
80 sampled of the tunicates with five published genomes (*Ciona robusta* [formerly known as *C.*
81 *intestinalis* type A; (Brunetti et al. 2015; Gissi et al. 2017)], *Ciona savigny, Molgula oculata*,
82 *Molgula occulta*, *Molgula occidentalis;* (Dehal et al. 2002a; Small et al. 2007; Stolfi et al. 2014),
83 two yet unpublished species (*Phallusia mamilata*, *Phallusia fumigata*; (Brozovic et al. 2016))
84 and two currently being assembled (*Halocynthia rorezi*, *Halocynthia aurantium*; (Brozovic et
85 al. 2016)).
86

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

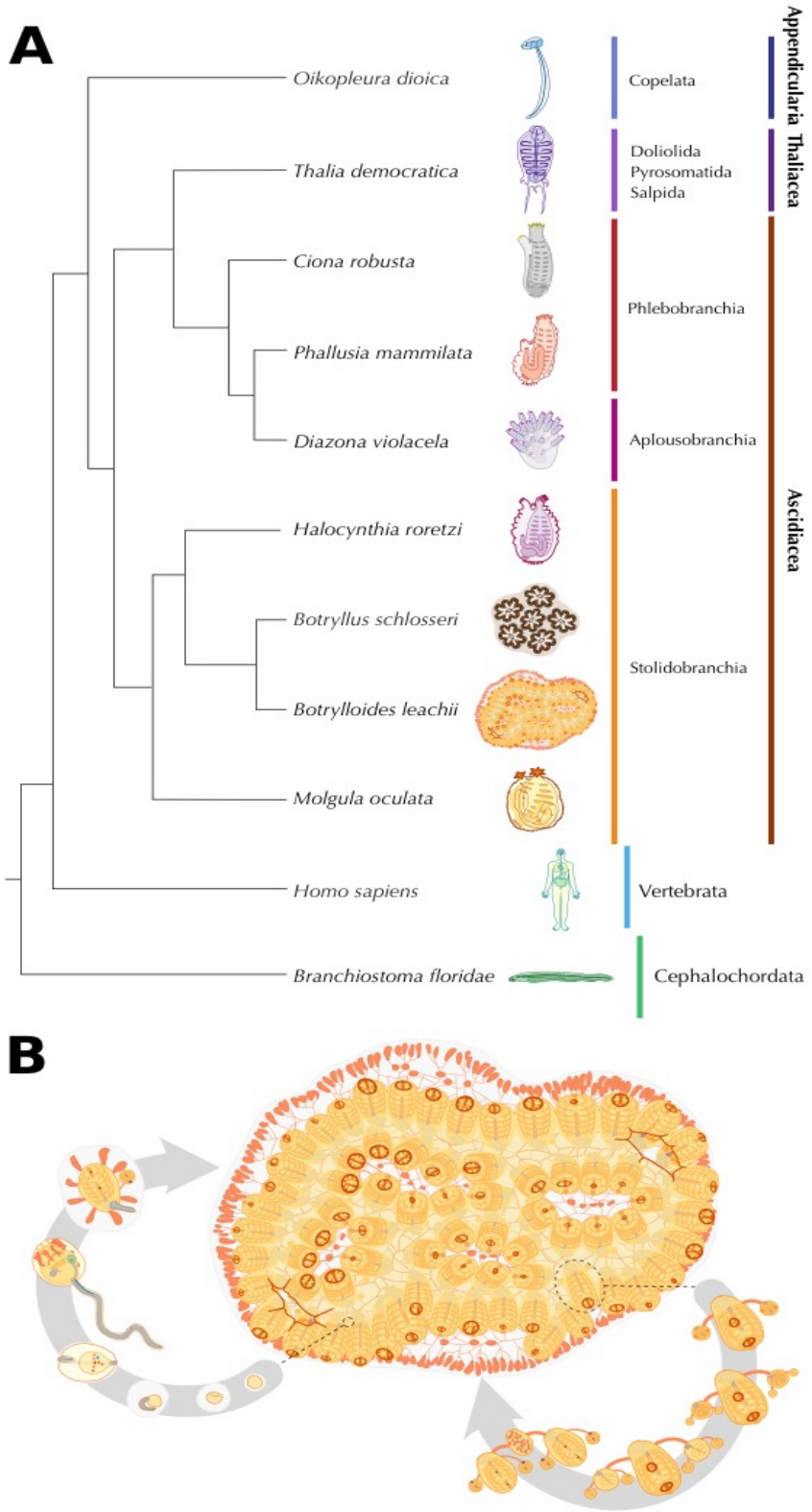


Figure 1 *B. leachii* phylogenetic position and life cycle.
A. Schematic showing phylogeny of tunicates with respect to the chordate clade. **B.** Life cycle of *B. leachii*. The colony expands and grows by asexual reproduction (right loop). During favourable conditions such as warmer water temperatures, members of the colonies start sexual reproduction (left loop). The embryos remain with the colony in brood pouches until release. Hatched larvae attach to nearby substrates and begin metamorphosis into a zooid.

87

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Colonial tunicates (Aplousobranchian and some families among the Stolidobranchian) are capable of both sexual, through autogamy, and asexual reproduction, through a wide range of budding types (palleal, vascular, stolonial, pyloric and strobilation; (Brown and Swalla 2012)). In addition, these compound organisms can undergo whole-body regeneration (WBR; reviewed in (Kürn et al. 2011)). Colonial ascidians are emerging as unique and increasingly popular model organisms for a variety of studies including immunobiology, allorecognition, angiogenesis and WBR (Rinkevich et al. 1995; Ballarin et al. 2001; Rinkevich et al. 2007a; Manni et al. 2007; Gasparini et al. 2008; Franchi et al. 2011; Lauzon et al. 2013; Rinkevich et al. 2013). However, colonial tunicates were only represented by *Botryllus schlosseri*, an ascidian that shows a significant expansion of its genome size when compared to the other available Tunicata genomes (725 Mb). To further investigate this fascinating subphylum and assess whether genome expansion is a prerequisite for coloniality and WBR, we have assembled and analysed the genome sequence of *Botrylloides leachii* (class Ascidiacea, order Stolidobranchia; (Savigny 1816)).

The viviparous colonial ascidian *B. leachii* can reproduce sexually through a tadpole stage that allows the settlement of a novel colony onto a new substrate (Fig. 1B). Each colony is composed of genetically identical adults (termed zooids) organized in ladder-like systems and embedded in gelatinous matrix (tunic). While each adult has its own heart, they all share a common vascular system embedded within the tunic. In the presence of sufficient food supply, the size of the colony doubles approximately every 20 days through synchronized asexual reproduction, known as palleal budding. During this process each adult produces two daughter zooids that ultimately replace the mother, which is then resorbed by the colony. In addition, upon loss of all zooids from the colony, *B. leachii* can undergo whole-body regeneration and restore a single fully-functional adult in as little as 10 days from a small

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

113 piece of its vascular system (Rinkevich et al. 1995). Furthermore, when facing unfavourable
114 environmental conditions, these colonial tunicates can enter into hibernation, whereby all
115 zooids undergo regression and are resorbed by the remaining vascular system. When a
116 favourable environment is restored, mature adults will be restored to re-establish the colony
117 (Burighel et al. 1976).

118 We have assembled and annotated the first *de novo* draft genome of the *B. leachii* by
119 taking advantage of our recently published transcriptomes (Zondag et al. 2016). Using this
120 genome, we have then undertaken a large-scale comparison of the four best-annotated
121 ascidian genomes *(B. schlosseri*, *C. robusta, M. oculat*a and *O. dioica*) to gain insights into some
122 of the diverse biological abilities that have evolved within the Tunicata.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Results

Genome assembly and annotation

To minimize contamination from marine algae and bacteria typically present in the pharyngeal basket of feeding *B. leachii*, we isolated genomic DNA from embryos of a single wild *B. leachii* colony. Genomic DNA was used to produce two libraries: one short-range consisting of 19,090,212 fragments (300 bp) of which 100 bp were paired-end sequenced - important for obtaining high coverage - and a second long-range mate pair with 31,780,788 fragments (1.5 - 15 kb size range, median ~3 kb) of which 250 bp were paired-end sequenced – important for scaffolding the assembly. Following quality checks, low quality reads were removed and sequencing adaptors were trimmed, thus resulting in a high-quality dataset of 86,644,308 paired-end and 12,112,004 single-end sequences (100 % with a mean Phred score >= 30, < 1 % with an adapter sequence, Fig. S1).

We then followed a reference-free genome characterization (Simpson 2014); provided with statistics from the human, fish (*Maylandia zebra*; (Bradnam et al. 2013)), bird (*Melopsittacus undulatus;* (Bradnam et al. 2013)) and oyster (*Crassostrea gigas*, (Zhang et al. 2012)) genomes for comparison; to estimate three properties of the *B. leachii* genome. First, k-mer count statistics were used to estimate the genome size to be 194.2 Mb (194,153,277 bp). This size is similar to that of the solitary *C. robusta, C. savigny and M. oculata* (160 Mb, 190 Mb and 160 Mb, respectively; (Dehal et al. 2002a; Small et al. 2007; Stolfi et al. 2014), larger than the compacted 70 Mb genome of *O. dioica* but appreciably smaller than the predicted 725 Mb genome of the related colonial ascidian *B. schlosseri*, of which 580 Mb have been sequenced (Voskoboynik et al. 2013a). Second, by quantifying the structure of the de Brujin graph obtained using the k-mer counts, the computational complexity of the assembly

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

was estimated (sequencing errors 1/213, allelic differences 1/233, genomic repeats 1/2,439). With a cumulative occurrence of 1/106, the *B. leachii* genome is similar to that of bird, more variable than those of fish and human, but still quite less complex than the notably difficult oyster genome (Fig. S1). Third, sequence coverage was estimated using the distribution of 51-mers counts, showing a well-separated bimodal distribution with a true-genomic k-mers maximum at 31x coverage, similar to the human genome but higher than both the fish and the bird. Overall, these metrics suggest that *B. leachii* has a genome well suited for *de novo* assembly and that our sequencing could result in a high quality assembly.

De novo assembly using Metassembler (Wences and Schatz 2015) produced a genome of 159,132,706 bp (estimated percentage of genome assembled is 82 %), with an average sequencing coverage of 66x (after adaptor trimming). The assembly is composed of 1,778 scaffolds, with a N50 scaffold length of 209,776 and a L50 scaffold count of 223. The 7,783 contigs, with a N50 length of 48,085, and a L50 count of 781, represent a total size of 146,061,259 (92 %, Table 1). To evaluate the completeness of our assembly, we used the Benchmarking Universal Single-Copy Orthologs (BUSCO; (Simão et al. 2015)). This tool provides a quantitative measure of genome completeness by verifying the presence of a set of manually curated and highly conserved genes. Out of the 843 orthologs selected in metazoans, 760 (90 %) were found in our assembly of the *B. leachii* genome (File S1), a relatively high score when compared to the BUSCO score of frequently used genome assemblies such as *Homo sapiens* (89 %, GCA_000001405.15). In addition, we took advantage of our previous assembly of the *B. leachii* transcriptome (Zondag et al. 2016) to further assess the quality of our genome. Using BLAT (Kent 2002), we were able to map 93 % of transcript sequences (48,510/52,004) onto our assembly. Overall, these results indicate that our *de novo* genome is largely complete and suitable for annotation.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1. *B. leachii* genome assembly statistics

Total length of assembly	159,132,706 bp
Predicted genome size	194 Mb
Number of scaffolds	1,778
Median scaffold length	43,485 bp
N50 contig length	209,776 bp
Estimated genome coverage before adaptor trimming	101x
Estimated genome coverage after adaptor trimming	66x
Number of predicted genes	15,839
% of the *B. leachii* reference transcriptome aligning to the genome	80 %
% of *Ciona* proteins that have a significant match to the *B. leachii* genome	71%
BUSCO score	90 % (760/843)

Ab initio genome annotation was performed using MAKER2 (Holt and Yandell 2011) and predicted 15,839 coding genes, of which 13'507 could be classified using InterProScan (Jones et al. 2014). Comparing these predictions with our mapping of the transcriptome, we found out that 83 % of our aligned cDNA (40,188/48,510) mapped to a predicted gene locus thus spanning 78 % of the annotated genes (12,395/15,839). In addition, a total of 4,213 non-coding sequences were predicted using Infernal (Nawrocki and Eddy 2013), Snoscan (Lowe 1999) and tRNAscan-SE (Lowe 1999). Finally, repetitive elements were annotated using RepeatMasker (Smit et al. 2015) and a species-specific library created using the RepeatModeler module (Smit and Hubley 2015). Eighteen percent of the genome was identified as containing repetitive elements (Table 2 and Table S1), a majority (17%) of these being interspersed repeats. This proportion is similar to that found in other tunicates including *C. robusta* (17%), M. oculata (22%) and *O. dioica* (15%), while being lower than that in *B. schlosseri* (60%).

To further characterize the genome of *B. leachii*, we set out to compare it to four available Tunicata genomes. First, we quantified the number of sequences from their proteomes which mapped onto our assembly using tBLASTn (Camacho et al. 2009): *C. robusta* 71% (10,507/14,740), *M. oculata* 77 % (12,788 / 16,616), *B. schlosseri* 71 % (32,944/ 46,519) but only 30% for *O. dioica* (9,009/29,572). Secondly, we performed an all-to-all search for protein orthologs using the OrthoMCL clustering approach (Li et al. 2003) to identify any

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

orthologs between the tunicate genomes (Fig. 2A). Clustering the combined protein set from all five genomes resulted in 17,710 orthologous groups. By classifying each group based on which tunicate has proteins in it, we identified the presence of five larger set of orthologs: those shared by all species (17% of all groups), those shared by all sessile tunicates (11%), those between the two colonial species (11%) and two groups unique to *B. schlosseri* and *O. dioica* (15 % and 12 %, respectively; Fig. 2A). Thirdly, the proteins specific to an organism were removed from the corresponding proteome, and a new mapping to the *B. leachii* genome was performed. Mapping of the proteome using only the conserved sequences is 93 % for *B. schlosseri* and 45 % for *O. dioica*.

Table 2. Comparison of sequenced tunicate genomes and their most prominent biological features.

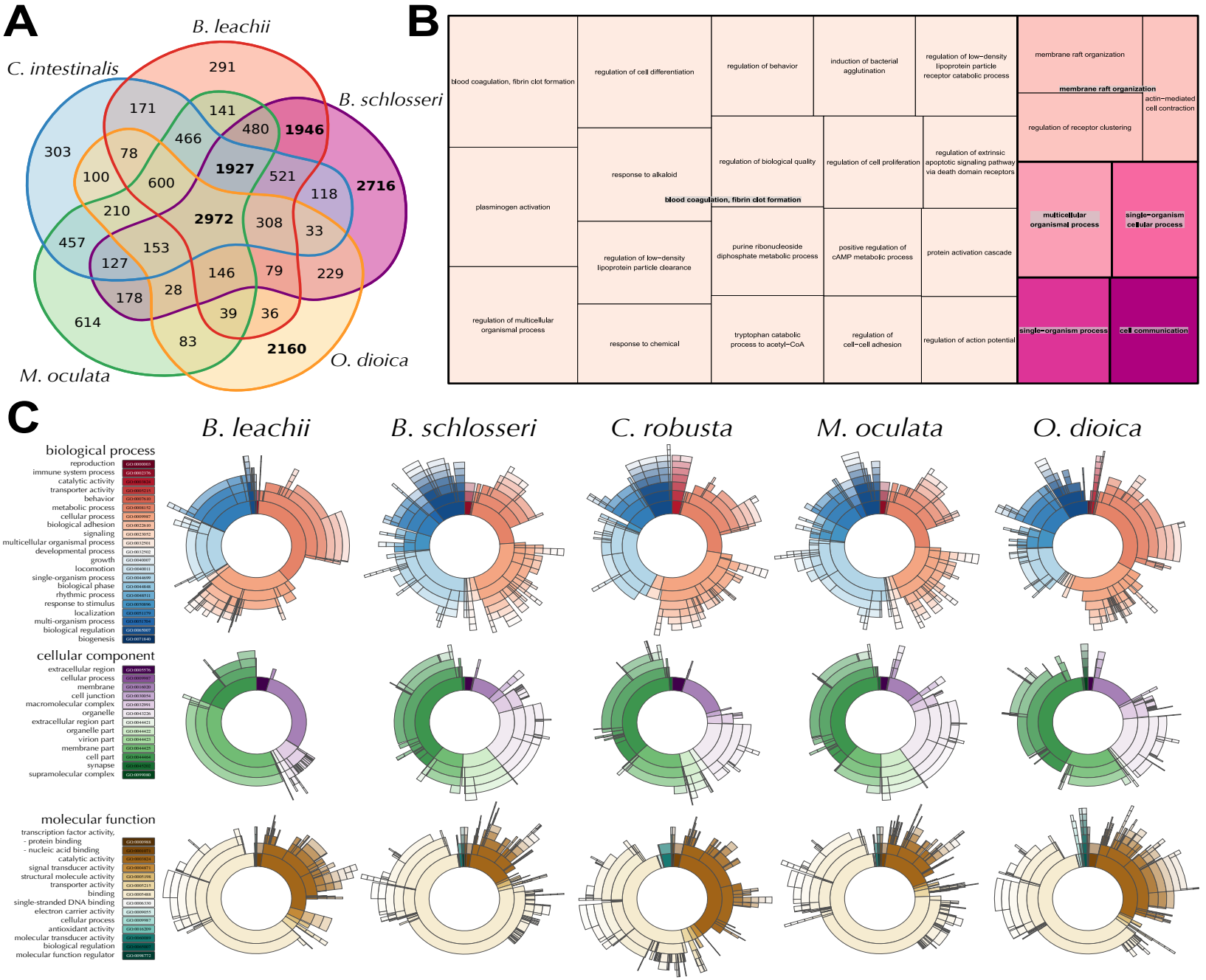
	Ciona robusta	***Botrylloides leachii***	***Botryllus schlosseri***	***Oikopleura dioica***	***Molgula oculata***
Genome size	160 Mb	194 Mb	725 Mb	72 Mb	160 Mb
Number of sequences	4,390	1,778	121,094	1,260	10,554
Fraction of repetitive DNA	17%	18%	60%	15%	22%
Predicted gene number	16,671	15,839	27,463	16,749	15,313
GC content	36%	41%	41%	40%	36%
Body structure	solitary, sessile	colony, sessile	colony, sessile	solitary. motile	solitary, sessile
Reproduction	sexual, hermaphrodite	asexual sexual, hermaphrodite	asexual sexual hermaphrodite	sexual, separate sexes	sexual, hermaphrodite
Regenerative ability	specific organs	WBR	WBR	unknown	unknown

To get insights into the potential biological function underlying these ortholog groups, we analysed the distribution of Gene Ontology (GO) terms for each cluster (Fig. S2). Interestingly, an important fraction of the shared orthologs are related to G-protein signalling, a conserved family of proteins involved in a variety of transduction mechanisms (Iwasa et al. 2003; Philips et al. 2003; Murata et al. 2001). Additionally, given that these proteins are potentially novel to colonial tunicates, a cross-species approach for GO enrichment was

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

performed using the Human GO database as background (Fig. 2B, (Primmer et al. 2013)). Finally, we compared the composition of GO terms at the genomic level (Fig. 2C). Despite *B. schlosseri* having a larger predicted gene number compared to the other analyzed tunicates, the overall proportion of GO groups terms were distributed similar between all genomes (Fig. 2C), indicating no expansion of one particular functional group in *B. schlosseri*. Overall, these analyses showed that our assembly and annotation is consistent with the other tunicate genomes and provides novel insights into their shared mechanisms and potentially for evolutionarily conserved mechanisms as well.

Figure 2 (Following page). Comparison of the tunicate genomes.
A. Clustering of orthologous protein sequences. Indicated are the number of cluster groups, each of which contains at lest two proteins. **B.** Treemap representation of the overrepresented GO Biological Processes terms within the ortholog groups shared between *B. leachii* and *B. schlosseri* genomes but not with *C. robusta, O. dioica* and *M. oculata*. Each rectangle size is proportional to the p-value for the GO term. **C.** Distribution of the three classes of GO terms for each species. The color-codes (left) are common for the entire row.


A
B. leachii
C. intestinalis
B. schlosseri
M. oculata
O. dioica
291
171
141
480
1946
303
78
466
1927
521
2716
100
600
118
210
2972
308
33
457
153
127
146
79
229
178
28
39
36
614
83
2160
B
blood coagulation, fibrin clot formation
regulation of cell differentiation
regulation of behavior
induction of bacterial agglutination
regulation of low-density lipoprotein particle receptor catabolic process
membrane raft organization
membrane raft organization
actin-mediated cell contraction
regulation of receptor clustering
plasminogen activation
response to alkaloid
blood coagulation, fibrin clot formation
regulation of biological quality
regulation of cell proliferation
regulation of extrinsic apoptotic signaling pathway via death domain receptors
multicellular organismal process
single-organism cellular process
regulation of low-density lipoprotein particle clearance
purine ribonucleoside diphosphate metabolic process
positive regulation of cAMP metabolic process
protein activation cascade
regulation of multicellular organismal process
response to chemical
tryptophan catabolic process to acetyl-CoA
regulation of cell-cell adhesion
regulation of action potential
single-organism process
cell communication
C
B. leachii
B. schlosseri
C. robusta
M. oculata
O. dioica
biological process
reproduction
immune system process
catalytic activity
transporter activity
behavior
metabolic process
cellular process
biological adhesion
signaling
multicellular organismal process
developmental process
growth
locomotion
single-organism process
biological phase
rhythmic process
response to stimulus
localization
multi-organism process
biological regulation
biogenesis
cellular component
extracellular region
cellular process
membrane
cell junction
macromolecular complex
organelle
extracellular region part
organelle part
virion part
membrane part
cell part
synapse
supramolecular complex
molecular function
transcription factor activity, - protein binding - nucleic acid binding
catalytic activity
signal transducer activity
structural molecule activity
transporter activity
binding
single-stranded DNA binding
electron carrier activity
cellular process
antioxidant activity
molecular transducer activity
biological regulation
molecular function regulator

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ancient gene linkages are fragmented in tunicate genomes

Ancient gene linkages are highly conserved sets of genes that are spatially restricted, commonly occurring in clusters (Garcia-Fernàndez 2005). These clusters arose in a common ancestor and were preserved because of a common regulatory mechanism such as cis-regulatory elements located within the cluster. The homeobox-containing *Hox* gene family, typically composed of 13 members in vertebrates (Hoegg and Meyer 2005), is among the best-studied examples of such an ancient gene cluster and is critical for the correct embryonic development (Pearson et al. 2005). The linear genomic arrangement of genes within the *Hox* cluster reflects their spatial expression along the anterior-posterior body axis (Pascual-Anaya et al. 2013), which establishes regional identity across this axis. The basal cephalochordate *B. floridae* has all 13 *hox* genes located in a single stereotypical cluster, along with an additional 14th gene (Fig. 3B; (Takatori et al. 2008)), suggesting that the chordate ancestor also had an intact cluster. However in tunicates, this clustering appears to be lost. In *C. robusta*, the nine identified *Hox* genes are distributed across five scaffolds, with linkages preserved only between *Hox2, Hox3* and *Hox4; Hox5* and *Hox6*; *Hox12* and *Hox13* (Fig. 3; (Spagnuolo et al. 2003; Wada et al. 2003)). In *O. dioica,* the total number of *Hox* genes is further reduced to eight, split between 6 scaffolds, including a duplication of *Hox9* (Fig. 3A; (Edvardsen et al. 2005)). In *M. oculata* we could identify only six *Hox* genes, divided between 4 scaffolds, with clustering retained for the *Hox10*, *Hox11* and *Hox12* genes (Fig. 3). In Botryllidae genomes, the same seven *Hox* genes are conserved (Fig 3B), with a preserved linkage between *Hox10*, *Hox12* and *Hox13* in *B. leachii* and three copies of *Hox5* present in *B. schlosseri*. Altogether, the separation of the tunicate *Hox* cluster genes supports the hypothesis that reduction and separation of this ancient gene linkage occurred at the base of tunicate lineage (Edvardsen et al. 2005). In addition, *Hox9* appears to be specifically retained in neotenic Tunicates while there is no pattern of conserved *Hox* cluster genes specific to colonial ascidians.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

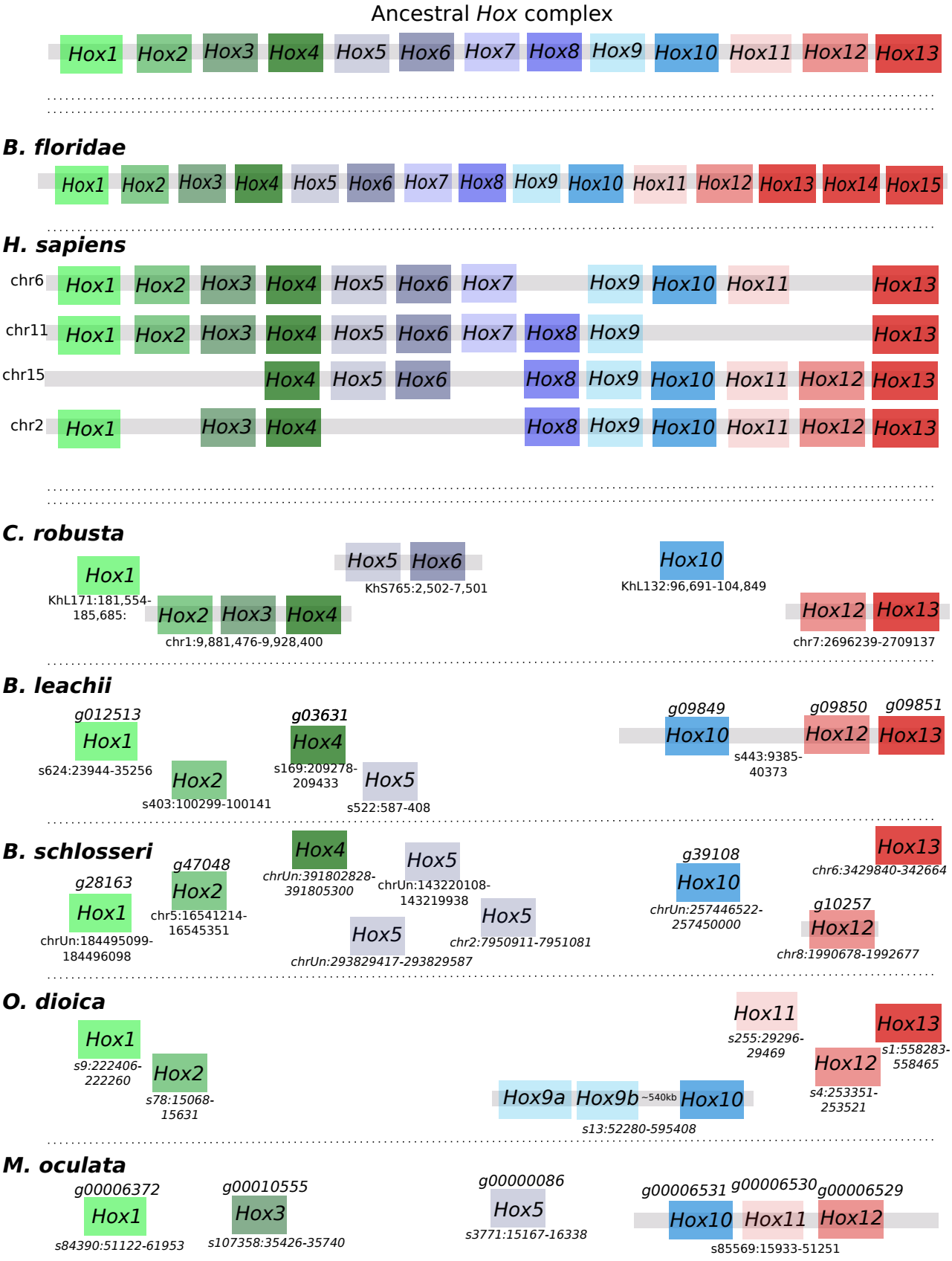


Figure 3. ***Hox*** **genes are dispersed and reduced in number within tunicate genomes.** Schematic depicting *Hox* gene linkages retained in five tunicate genomes in comparison to the ancestral *Hox* complex, which included thirteen genes. Orthologous genes are indicated by common colours. Chromosome or scaffold number is shown, along with gene ID when available for newly annotated genomes.

247

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A second ancient homeobox-containing gene linkage is the *NK* cluster. This cluster, predicted to be present in the last common ancestor of bilaterians (Luke et al. 2003), consists of *Msx, Lbx*, *Tlx*, *NKx1, NKx3, NKx4* and *NKx5* (Fig. 4). In *B. floridae*, linkages between *Msx, NKx4* and *NKx3*; as well as between *Lbx* and *Tlx* provide evidence of retained ancestral clustering while *NKx5* was lost (Fig. 4; (Luke et al. 2003)). However in vertebrates, NKx5 is still present, while only the gene linkages between *Lbx* and *Tlx* as well as between *Nkx4* and *Nkx3* remain (Fig. 4; (Garcia-Fernàndez 2005)). To further clarify the evolution of this ancestral cluster in tunicates, we determined the structure of the NK cluster within five ascidian genomes. In all these species, *NKx1* is absent and no evidence of clustering could be found with all identified orthologs located on different scaffolds (Fig. 4). In *C. robusta*, *M. oculata* and *O. dioica* only five members of this cluster remain, with the loss of either *Lbx* or *Tlx* as well as of *NKx3* and the duplication of the ortholog of *NKx4* (Fig. 4). By contrast, in the colonial tunicates *B. leachii* and *B. schlosseri*, *Tbx*, *Lbx* and *NKx3* are all present. In *B. schlosseri*, *Msx1* is absent and *NKx4* duplicated. In the *B. leachii* genome, *NK1* is the only ancestral cluster member to be missing and *Nk5* has been duplicated (Fig. 4). Altogether, these results suggest that there has been a loss of *NKx5* in Cephalochordate, one of *NKx1* in Tunicate and that the combination of *NKx3*, *Lbx* and *Tbx* may be specific to colonial ascidians.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

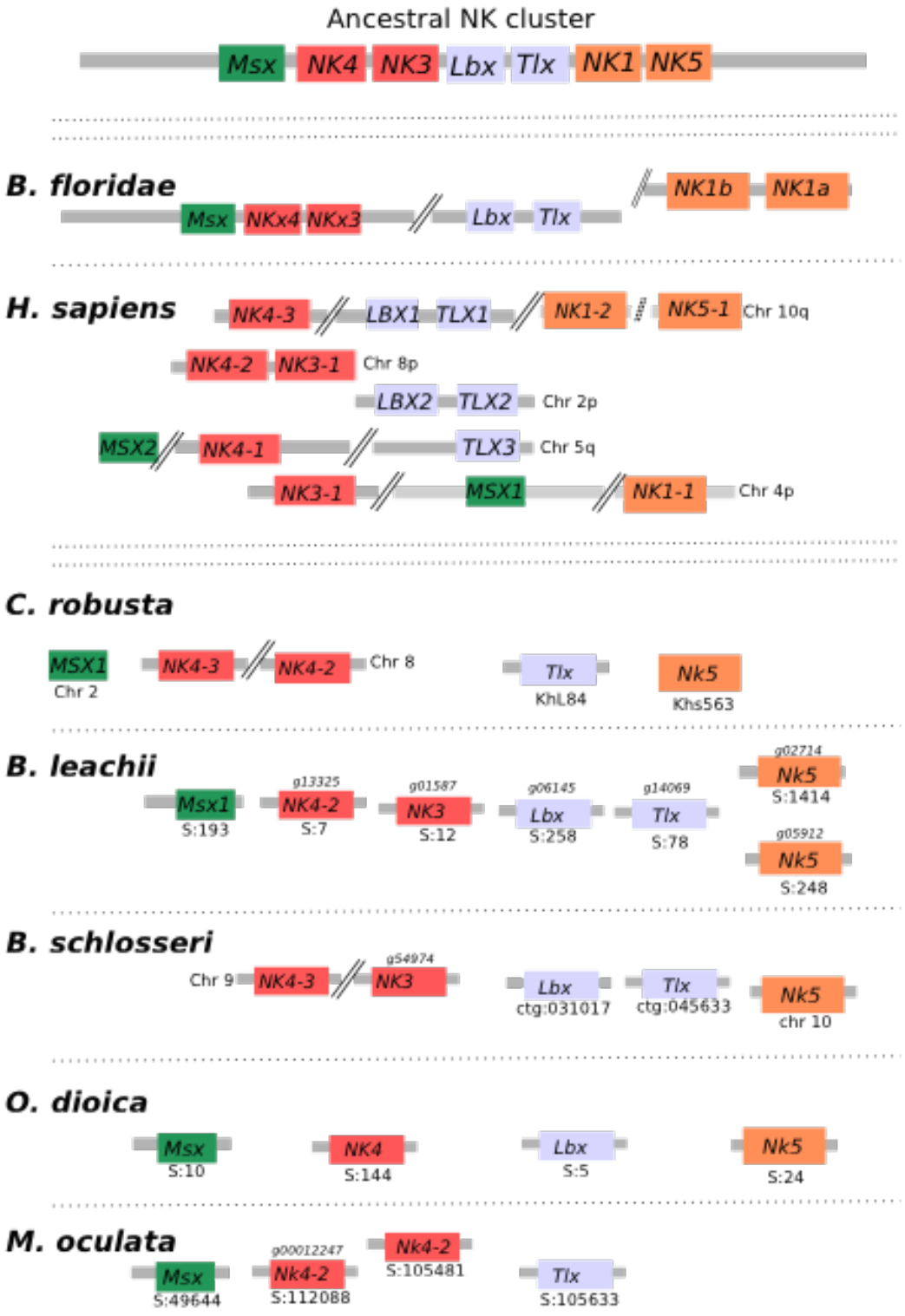


Figure 4. NK homeobox cluster genes are fragmented within tunicate genomes.
Schematic depicting the organization of the *NK homeobox* cluster genes among the studied chordate genomes. Double-parallel lines indicate > 1Mb distance between genes. Chromosome or scaffold number is shown, along with gene ID when available for newly annotated genomes. Orthologous genes are indicated by common colours.

265

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A third ancient linkage that we investigated is the pharyngeal cluster, a gene group present in hemichordates, echinoderm and vertebrates genomes that is considered to be Deuterosome specific (Simakov et al. 2015). The cluster groups *foxhead domain protein* (*FoxA*), *NKx2 (NKx2.2 and Nkx2.1)*, *Pax1/9*, *mitochondrial solute carrier family 25 member 21* (*slc25A21)*, *mirror-image polydactyly 1 protein* (*mipol1*), *egl nine homolog 3* (*egln3*) and *dehydrogenase/reductase member* 7 (*dhrs7*). Among these, s*lc25a21, Pax1/9*, *mipol1* and *FoxA* pairs are also found in protostomes suggesting an even more ancient origin (Simakov et al. 2015). The pharyngeal cluster is thought to have arisen due to the location of the regulatory elements of *Pax1/9* and *FoxA* within the introns of *slc25A21* and *mipol1* (Santagati et al. 2003; Wang et al. 2007), constraining the genes to remain in tight association with each other. In the *B. floridae* genome, the entire cluster is located on the same scaffold, with the exception of the *Nkx2.1* and *Nk2.2* gene pair located on separate scaffold. In *C. robusta*, only orthologs of *FoxA*, slc25a29, *Pax1* and *Pax9* could be identified. Nevertheless, all of them are located on the same chromosome (Fig. 5). In *O. dioica*, the cluster appears even further reduced. While orthologs of *FoxA*, *Pax1/9* and *Nkx2.2* genes were found on different scaffolds, only one rather distant linkage (> 1 Mb) between *a Pax-like* gene and *slc25A21* is retained. For both *B. schlosseri* and *M. oculata*, there was no evidence of clustering between genes (Fig. 5). In the *B. leachii* genome, *mipol1* is the sole missing gene from this cluster. However, only the pairing of a *Pax-like* and *slc25A21* genes remains (Fig. 5). Altogether, these results suggest that most of the Tunicates did not conserve the structure of this ancient linkage, but it is unknown what consequences this would have to their expression and function.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

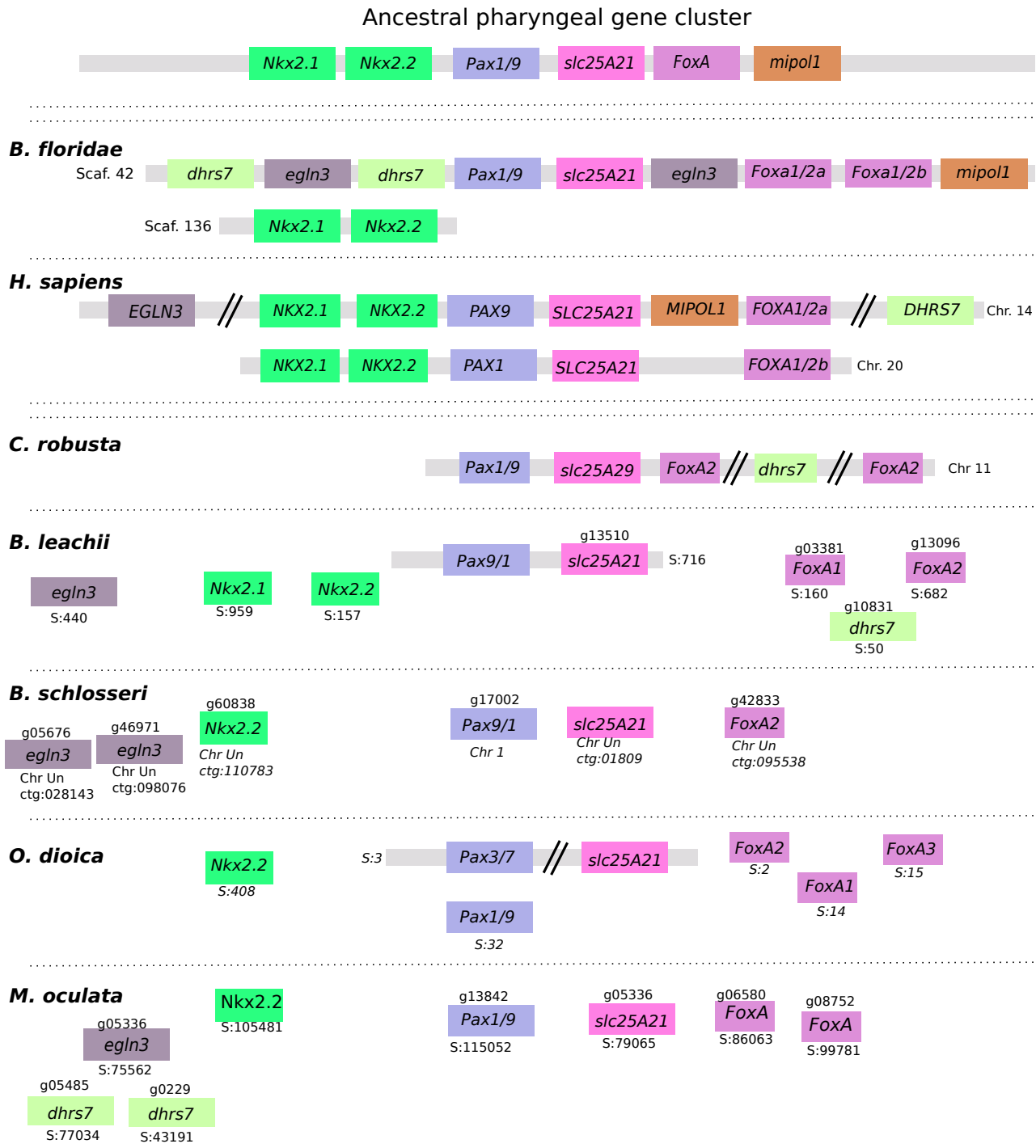


Figure 5. Ancestral gene linkages remain between a few pharyngeal cluster genes in tunicate genomes. Gene order of the six pharyngeal cluster genes, *NK2.1, NK2.2, Pax1/9* and *FoxA* in chordate genomes. Double-parallel lines indicate > 1 Mb distance between genes. Chromosome or scaffold number is shown, along with gene ID when available for newly annotated genomes. Orthologous genes are indicated by common colours.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

289 ***Lineage-specific changes to cell-signalling pathways in Botryllidae genomes.***

290 To dissect more specifically the evolution of colonial ascidians, we examined the
291 genomes of *B. leachii* and *B. schlosseri*, looking for key components of signalling pathways
292 required for metazoan development and regeneration. Of particular interest, we focused on
293 the Wingless-related integration site (Wnt), Notch and Retinoic acid (RA) signalling pathways.
294 All three of these pathways have been implicated in WBR and asexual reproduction in colonial
295 tunicates (Rinkevich et al. 2008, 2007b; Zondag et al. 2016).

296

297 **Wnt pathway**

298 Wnt ligands are secreted glycoproteins that have roles in axis patterning,
299 morphogenesis and cell specification (Loh et al. 2016). The ancestral gene family appears to
300 originate very early on during multi-cellular evolution and to be composed of eleven members
301 (Kusserow et al. 2005; Guder et al. 2006). The *Wnt* gene family expanded to 19 members in
302 the human genome, while independent gene loss has reduced this family to 7 genes in
303 *Drosophila melanogaster* and *Caenorhabditis elegans* (Prud'homme et al. 2002). Consequently,
304 we set out to investigate whether the *Wnt* gene family has either expanded or contracted
305 during Tunicata speciation.

306 We found an increase in the number of *Wnt5a* genes among Styelidae genomes. In *B.*
307 *schlosseri*, we identified 15 *Wnt* members, including seven *Wnt5a-like* genes on multiple
308 scaffolds (Fig. 6, Table S2). In the *B. leachii* genome, fourteen *Wnt* ligand genes were
309 identified, including four *Wnt5a* genes located on the same scaffold near *Wnt4* (Fig. 6). *M.*
310 *oculata* has only 7 Wnt ligand genes, including three *Wnt5a-like* genes (Fig. 6, Table S2). In
311 comparison, *C. robusta* has a total of 11 *Wnt* genes, including a single *Wnt5a* gene (Fig. 6,
312 Table S2; (Wada et al. 2003)). In the compact *O. dioica* genome, this number has reduced to 6
313 (Wnts 3, 4, 7, 11 and 16), none of which are *Wnt5a* orthologs (Table S2). Overall, this suggests

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that an expansion through gene duplication of the Wnt5 family occurred during tunicate evolution, but was lost in some lineages.

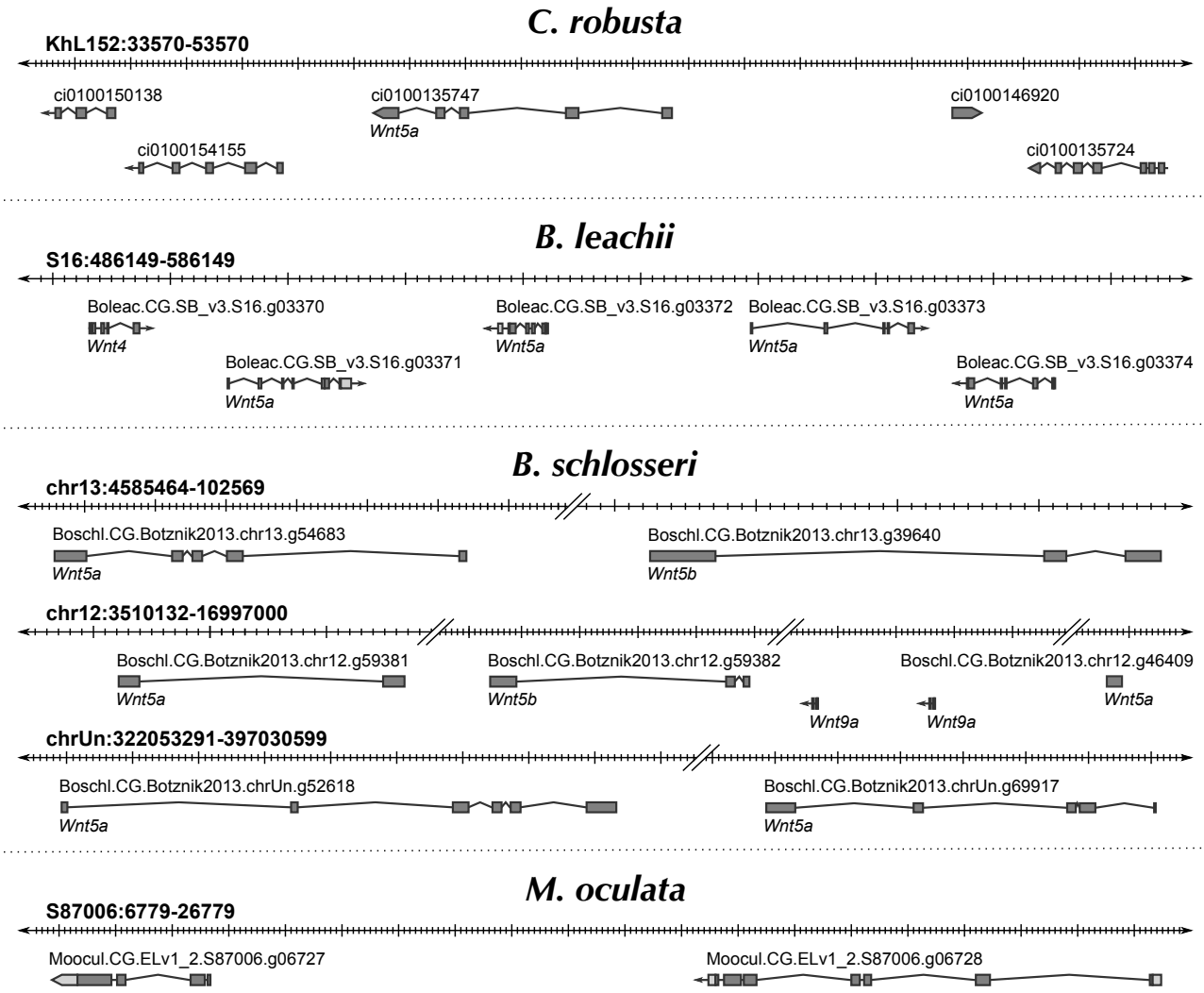


Figure 6. Duplication of *Wnt5a* genes in tunicate genomes.
Schematic showing the genomic location of *Wnt5*-like genes within each indicated genome. Note that no *Wnt5a* ortholog is present in the *O. dioica* genome. Double-parallel lines indicate > 1Mb distance between genes.

To assess the functionality of the Wnt pathway in Tunicates, we set out to assess whether its downstream effectors are themselves present in the available genomic data. The downstream pathways activated by Wnt ligands are divided into canonical, non-canonical calcium and non-canonical planar cell polarity. The Wnt5a ligand is associated with both of

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the non-canonical pathways through binding of membrane receptors that include *frizzled* (*Fzd4*), *receptor tyrosine kinase-like orphan receptor 1/2* (*Ror1/2*) and *atypical tyrosine kinase receptor* (*Ryk*). Further downstream, disheveled (dsh), β-catenin (Cnntb), Axin, low-density lipoprotein receptor-related protein 5/6 (LRP5/6) and nuclear factor of activated T-cells (NFAT) are proteins essential for triggering intracellular responses to Wnt signalling (MacDonald et al. 2009). We identified orthologs for each of these signalling transduction molecules in all Tunicata genomes (Table S2), with no evidence of further gene duplication events. This supports the interpretation that signalling through the Wnt pathway is functional in tunicates.

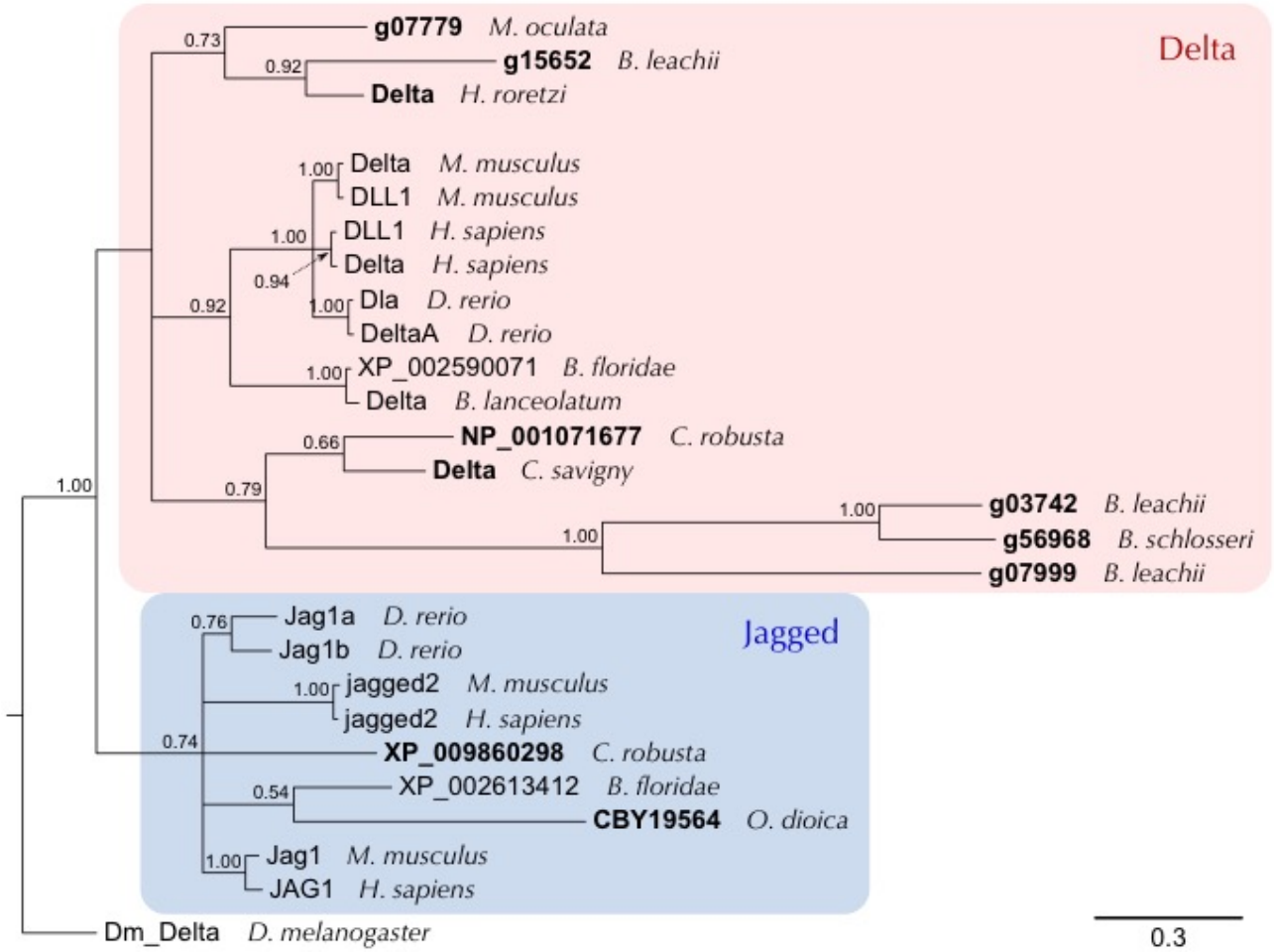
Notch pathway

Notch receptors are transmembrane proteins that are involved in cell-cell signalling during development, morphogenesis and regeneration (Hamada et al. 2015). Following activation through the binding of the delta or jagged/serrate ligands, the intracellular domain of Notch is cleaved and induces the expression of downstream target genes including the *hes (hairy and enhancer of split*) gene family members (Guruharsha et al. 2012). The presence of both Notch and the Delta/Serrate/lag-2 (DSL) proteins in most metazoan genomes suggests that their last common ancestor had a single copy of each gene (Gazave et al. 2009). To establish how this pathway has evolved in tunicates, we screened these genomes for the Notch receptor using the conserved lin-Notch repeat (LNR) domain, and for genes encoding probable Notch ligands such as genes from the DSL family .

In all examined genomes, only a single *Notch* receptor gene was identified while the number of ligand genes varied (Table S3). The *C. robusta* genome contains two *DSL* genes, while *O. dioica, M. oculata* and *B. schlosseri* possess only a single *DSL*. By contrast, we found three DSL genes in *B. leachii* (Table S3). To determine the relationships between these

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

346 identified tunicate DSL-like genes, a phylogeny was constructed along with other chordate
347 DSL proteins. All three *B. leachii* genes are Delta orthologs, two of them related to the *B.*
348 *schlosseri* and *Cionidae* copy; the third one closer to the *M. oculata* and *H. roretzi* variant. The
349 mouse, human and zebrafish delta and delta-like (DLL) proteins form a discrete clade loosely
350 related to the genes found in Cephalochordates and Tunicates (Fig. 7, shaded box). Jagged
351 proteins form a separate clade (Fig. 7). The tunicate DSL-like proteins present long
352 phylogenetic branches, suggestive of greater diversity, which is also observed in the protein
353 alignment (Fig. S3). This suggests that the tunicate DSL proteins are diverging rapidly from
354 each other, indicative of lineage specific evolution of DSL-like genes.



355

Figure 7. ***B. leachii*** **Notch pathway**
Bayesian phylogenetic tree depicting the relationship between tunicate and vertebrate DSL proteins, using *Drosophila* Delta to root the tree. Tunicate proteins are shown in bold and shaded areas correspond to Delta and Jagged groupings. Branch support values (probabilities) are indicated.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

356

357

358 **Retinoic acid signalling**

359 Retinoic acid (RA) is an extracellular metabolite that is essential for chordate
360 embryonic development. RA is synthesized from retinol (vitamin A) by two successive
361 oxidation steps. In the first step, retinol dehydrogenase (RDH) transforms retinol into retinal.
362 Then RA is produced by aldehyde dehydrogenase (ALDH), a superfamily of enzymes with
363 essential roles in detoxification and metabolism (Jackson et al. 2011). RA influences the
364 expression of downstream target genes by binding to the RA receptors, RAR and RXR (Fig. 8A
365 (Cunningham and Duester 2015)). Finally, RA is metabolized by the cytochrome P450 family
366 26 (Cyp26) enzyme, which absence of expression can restrict RA-induced responses to
367 specific tissues or cell types (Ross and Zolfaghari 2011). Components of this pathway have
368 been found in non-chordate animals, suggesting a more ancient origin (Canestro et al. 2006).
369 This pathway has previously been shown to be required for *B. leachii* WBR and *Ciona*
370 development, yet several genes required for RA signalling appear to be missing in *O. dioica*
371 (Martí-Solans et al. 2016).

372 Rdh10 is the major dehydrogenase associated with the first steps of RA production,
373 although the Rdh16 and RdhE2 enzymes can also substitute this function (Belyaeva et al.
374 2009; Lee et al. 2009; Belyaeva et al. 2015). The *O. dioica* genome has no orthologs for either
375 *Rdh10* or *Rdh16* but it does have four genes that encode for RdhE2 proteins (Martí-Solans et
376 al. 2016). *O. dioica* also lacks both an *Aldh1*-type gene as well as a *Cyp26* gene but has a single
377 RXR-ortholog (Table S4, (Martí-Solans et al. 2016)). In contrast, the *C. robusta* genome,
378 contains single copies of *Rdh10*, *Rdh16* and *RdhE2* genes and a total of four *Aldh1* genes,
379 located on two chromosomes (Canestro et al. 2006). Consistent with *C. robusta*, *M. oculata*, *B.*

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

380 *leachii* and *B. schlosseri* genomes all have single copies of *Rdh10, Rdh16* and *RdhE2* genes, as
381 well as three *Aldh1a/b* genes on separate scaffolds (Table S4).

382 Three retinoic acid receptor genes were identified within the *B. leachii* genome, one of
383 which had been cloned previously (*g03013,* (Rinkevich et al. 2007b) All three were also found
384 in *C. robusta*, *M. oculata* and *B. schlosseri* genomes (Table S4). While there is only one potential
385 *Cyp26* gene in *M. oculata*, four paralogs were identified in *B. leachii* and *B. schlosseri*. A
386 phylogenetic analysis showed that these 4 genes group with CYP26 proteins (Fig. 8B, Table
387 S4). Altogether, these results show a loss of key RA-pathway genes in *O. dioica* (*Rdh10, Rdh16,*
388 *Cyp26* and *Aldh1a*), while copy numbers in other tunicate genomes increase.

389

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

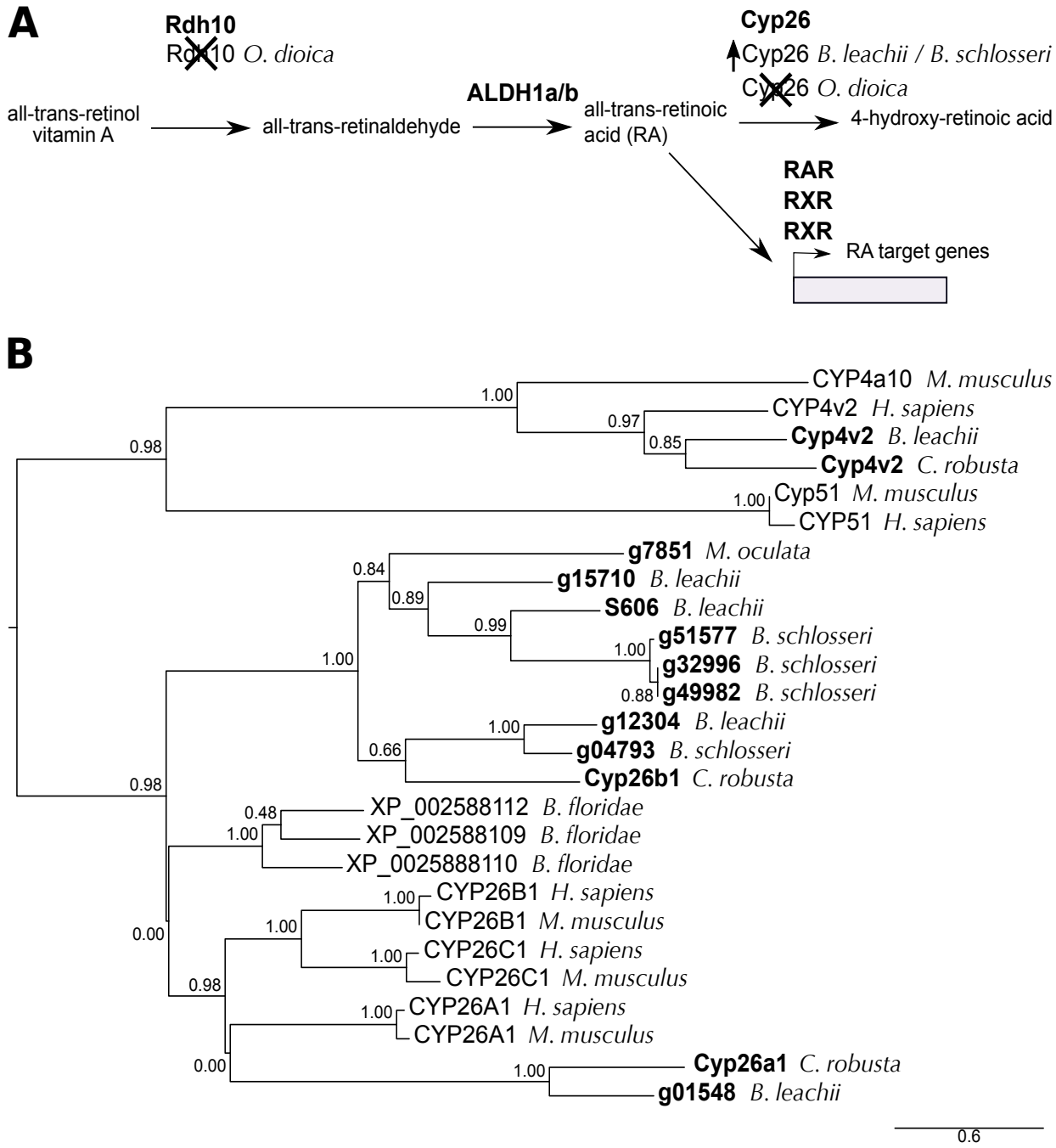


Figure 8. Evolution of the RA pathway in tunicates

(A) Overview of the RA synthesis and degradation pathway. In bold are the major proteins that contribute to RA signalling during animal development. Indicated below these are changes to the number of copies present in examined genomes. **(B)** ML phylogenetic tree depicting the relationship between invertebrate and vertebrate CYP26 proteins using CYP4 and CYP51 proteins as an out-group. Tunicate proteins are shown in bold. No *Cyp26* gene has been identified in the *O. dioica* genome. Values for the approximate likelihood-ratio test (aLRT) are indicated.

390

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

Genomic diversity within the Stolidobranchia

The *B. leachii* genome, along with previous genomic analyses of other ascidian species, support the widely held view that ascidian genomes are diverse and rapidly evolving, which is particularly evident in the Stolidobranchia group (Seo et al. 2001; Dehal et al. 2002b; Voskoboynik et al. 2013a; Stolfi et al. 2014; Tsagkogeorga et al. 2010; Bock et al. 2012; Tsagkogeorga et al. 2012; Rubinstein et al. 2013; Griggio et al. 2014). Nevertheless, botryllids are sufficiently similar in external appearance and morphology for early researchers to have suggested that *Botrylloides* could be a subgenus of *Botryllus* (Saito et al. 2001; Nydam et al. 2017). Strikingly however, the *B. schlosseri* genome differs from that of *B. leachii*, as well as from other sequenced tunicate genomes (Table 2). Particularly striking is the comparison between the *B. leachii* and *B. schlosseri*, where differences in genome sizes (194 Mb vs 725 Mb), the fraction of repetitive sequences (18 % vs 60 %; 65 % in (Voskoboynik et al. 2013a)) and the predicted gene number (15,839 vs 27,463; (Voskoboynik et al. 2013a)) suggest divergent genome architectures. Altogether, these comparisons indicate that the *B. schlosseri* genome has undergone a significant increase in its genomic content, including retrotransposon expansion (Table S1). In particular, there are at least two additional families in the *B. schlosseri* hAT transposon superfamily and counts of common hAT elements, such as hAT-Charlie, can differ dramatically (e.g. hAT-Charlie 366 in *B. leachii* vs 46,661 in *B. schlosseri*). DNA methylation is a key suppressor of transposon activity, changes to the methylation of transposable elements is a known driver of increased transposition (O'Neill et al. 1998; Maumus and Quesneville 2014; Simmen et al. 1999; Suzuki et al. 2007). DNA methylation in tunicate species has only been studied in *C. robusta,* and is described as mosaic, gene body methylation, whereas non-coding regions including transposons remain

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

unmethylated (Suzuki et al. 2007), it is unknown how retrotransposons are suppressed in tunicate genomes. Nevertheless, the observed increase in transpositions could be a consequence of low non-coding DNA methylation, which may contribute to the rapid genome evolution observed in tunicate species, even between closely related species such as *B. schlosseri* and *B. leachii*.

Rapid genome evolution, and active transposable elements in particular, are proposed to aid adaptation to new environments for invasive species (Stapley et al. 2015). Differences in the colonization ability of tunicates has been noted, not only between related species such as *B. leachii* and *B. schlosseri* (Brunetti 1976; Brunetti et al. 1980; Brunetti 1974), but even at the molecular level within *B. schlosseri* populations (Bock et al. 2012; Nydam et al. 2017). It is thus possible that the observed success in tunicate invasion (Zhan et al. 2015) is supported by their plasticity in genome characteristics like transposon diversity and gene number.

Ancient homeobox gene clusters whose structure has been retained over millions of years of evolution in many organisms are fragmented in tunicate genomes. Because, the expression of each *Hox* gene across the anterior-posterior axis relates to their location within the *Hox* gene cluster (Pascual-Anaya et al. 2013), cluster breaks are predicted to have consequences for patterning processes. However, an adult body plan with correct spatial orientation of its body axes during tissue development in ascidians also needs to be established during sexual, asexual and WBR. Early patterning events in tunicate species have only been characterized during sexual reproduction in *Ciona*. Early stages of development (prior to gastrulation) follow a mosaic pattern of developmental axis formation, where inheritance of maternally provided factors establishes the body axes (Nishida 2005). *Hox* gene knockdown experiments in *C. robusta* revealed that they have very limited roles, with defects only observed in larval neuronal and tail development upon loss of *Ci-Hox11* and *Ci-Hox12* function (Ikuta et al. 2010). It thus appears that patterning events in *C. robusta* are less

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

dependent upon anterior-posterior spatial expression of *Hox* genes to establish regional identity. Previously in *B. schlosseri*, the entry point of the connective test vessel into the developing bud determines the posterior end of the new zooid (Sabbadin et al. 1975). Therefore it is possible that ascidians incorporate environmental and physical cues to compensate for the lost gene cluster during polarity establishment. A wider analysis comprising multiple tunicate species will be necessary to investigate the exact consequences of homeobox cluster dispersion and whether the compensatory mechanism observed in *C. robusta* is the norm or an exception.

Gene orthology analysis and coloniality candidate pathways

Among the tunicate orthologous clusters that we obtained, we identified several groups of genes that are not shared by all the tunicate genomes (Fig. 2A). Given the rapid genomic evolution of these organisms, it is more likely that these genes have either been lost or that their sequence has highly diverged, rather than independent gains of novel genes.

Of particular interest are the genes found only in the *B. schlosseri* and *B. leachii* genomes, as these may function in biological processes unique to colonial tunicates. Many of these genes have orthologs not only in vertebrates, but also in more evolutionarily distant animals such as *C. elegans* (File S4). This suggests that these genes have a more ancient origin, which was retained specifically in Botryllidae genomes. The overrepresented genes (File S4) have annotated functions including circulation (GO:0003018, GO:0003013, GO:0050880), wound healing (GO:0072378) and cell communication (GO:0007154); as well as regulation of immune cell differentiation (GO:0033081, GO:0033089), immune system process (GO:0002376) and interferons (GO:0032608). Unlike solitary tunicate species, colonial ascidians possess a complex system of single cell-lined vessels used to transport haemocytes and facilitate communication between zooids within the colony (Mukai et al. 1978). In

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

addition, immune response is known to have roles roles in wound healing, vasculogenesis, allorecognition and regeneration (Voskoboynik et al. 2013b; Taketa et al. 2015; Gutierrez and Brown 2017; Sattler 2017). Therefore, it is possible that these genes, found only in *Botryllus* and *Botrylloides*, contribute to biological pathways and cellular processes that have important roles in colonialism.

Both *O. dioica* and *B. schlosseri* had a high number (2160 and 2716 respectively) of clusters unique to their genomes (Fig. 2A). While the *O. dioica* genome has undergone considerable loss of ancestral genes (Albalat and Cañestro 2016; Seo et al. 2001), the total number of genes in this specie is similar to that of other tunicates (Table 2). Taken together, these observations suggest that there has been a duplication of the retained genes such as *Otx* (3 copies in *O. dioica*, one in *Ciona*(Cañestro et al. 2005)), potentially involving roles in their peculiar neotenic and dioecious life cycle. The *B. schlosseri* genome has an ~10,000 higher predicted gene number compared to other tunicates (Table 2). Such massive increase in numbers suggests partial genome duplication. Further analysis will be required to determine whether these are novel or duplicated genes, hence providing important insights in the evolution of Tunicata genomes.

Lineage-specific changes to evolutionarily conserved cell communication pathways

Cell signalling pathways are critical for morphogenesis, development and adult physiology. In particular, we have focused our analysis on three highly conserved pathways: Wnt, Notch and Retinoic Acid signalling. Representatives of all twelve *Wnt* gene subfamilies are found in metazoans, suggesting that they evolved before evolution of the bilaterians (Janssen et al. 2010). We identified members of each Wnt subfamily in tunicate genomes, along with numerous examples of lineage-specific gene loss and/or duplication. The most striking of these events was an increase in *Wnt5a* gene copy number in *B. leachii*, *B. schlosseri*

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and *M. oculata* genomes. Indeed, most invertebrates genomes, including the basal chordate *B. floridae*, contain a single *Wnt5* gene while most vertebrate genomes have two *Wnt5a* paralogs, believed to be a result of whole genome duplication (Martin et al. 2012). However, in the analysed tunicate genomes, up to 15 copies of this gene were identified, potentially these additional genes may have been co-opted into novel roles and were retained during tunicate evolution. Wnt5a ligands have numerous biological roles, including a suppressive one during zebrafish regeneration (Stoick-Cooper et al. 2007) and a promotive one during amphioxus regeneration (Somorjai et al. 2012). Furthermore, components of both Wnt signalling pathways are differentially expressed during WBR (Zondag et al. 2016). It is thus conceivable that *Wnt5a* gene number has expanded in colonial tunicates to sustain WBR. A functional characterization of the role of these numerous copies of Wnt5a would thus be highly interesting and potentially reveal evolutionary insights into chordate regeneration.

All components of the Notch pathway are present in the genomes we investigated. Of particular interest, the DSL Notch ligand appears to be rapidly evolving in the tunicates. This indicates that tunicate DSL proteins are under less pressure, than vertebrate orthologous proteins, to conserve their sequences. Given that the interaction between the DSL domain and the Notch receptor is core to signaling pathway activation (Chillakuri et al. 2012), it will be interesting to assess whether the functional ligand-receptor interactions between tunicate DSL proteins and tunicate Notch proteins have adapted accordingly.

Components of the RA signalling pathway have also been identified in all the tunicate genomes. However, *Oikopleura* has seemingly lost a functional RA synthesis pathway, while still forming a functional body plan. This suggests that either uniquely RA is not involved in critical developmental events in this species, that the RA signalling function has been replaced or that *O. dioica* utilizes an alternative synthesis approach. Conversely, lineage specific

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

515 increases in RA pathway gene numbers have been observed in *C. robusta* (Aldh1, (Sobreira et
516 al. 2011)) and *Botrylloides* (*CYP26* genes, Fig. 8).

517 RA, Notch and Wnt pathways play roles in regeneration and development in many
518 species, including Stolidobranchian tunicates (Rinkevich et al. 2007b, 2008; Zondag et al.
519 2016) and *Cionidae* (Hamada et al. 2015; Jeffery 2015). The observed loss of RA signalling
520 genes may result in reduced regeneration ability for *O. dioica*, however it's regenerative
521 abilities have not been characterized. Given the unique chordate WBR potential developed by
522 colonial tunicates, it is conceivable that there is selective pressure on their genomes to retain
523 these pathways. We thus predict that these pathways play a similar role in colony reactivation
524 following hibernation.

525 Among tunicates there exist significant differences in life cycle, reproduction and
526 regeneration ability, even between closely related species of the same family, which likely
527 reflect an underlying diversity in genomic content. For instance, differences in both asexual
528 and sexual reproduction have been observed within the Botryllidae family (Oka and
529 Watanabe 1957; Brunetti 1974, 1976, Berrill 1951, 1947, 1941). Furthermore, *B. schlosseri*
530 can only undergo WBR during a short time frame of their asexual reproductive cycle when the
531 adults are reabsorbed by the colony (Voskoboynik et al. 2007; Kürn et al. 2011) while *B.*
532 *leachii* can undergo WBR throughout their adult life (Rinkevich et al. 2007b). Overall, this
533 indicates that despite a generally similar appearance, the rapid evolution of the Tunicata
534 subphylum has provided diversity and innovations within its species. It will be interesting to
535 investigate how such genomic plasticity balances between adaptation to new challenges and
536 constraint, preserving common morphological features, in future studies.

537 In conclusion, our assembly of the *B. leachii* genome provides an essential resource for
538 the study of this colonial ascidian as well as a crucial point of comparison to gain further
539 insights into the remarkable genetic diversity among tunicate species. In addition, the genome

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of *B. leachii* will be most useful for dissecting WBR in chordates; particularly through comparison with *B. schlosseri* for understanding how the initiation of WBR can be blocked during specific periods of their life cycle. Furthermore, given the key phylogenetic position of Tunicates with respect to vertebrates, the analysis of their genomes will provide important insights in the emergence of chordate traits and the origin of vertebrates.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Sampling, library preparation and sequencing

B. leachii colonies were collected from Nelson harbour (latitude 41.26°S, longitude 173.28°E) in New Zealand. To reduce the likelihood of contamination, embryos were dissected out of a colony and pooled before carrying out DNA extraction using E.Z.N.A SP Plant DNA Mini Kit. A total of 2 µg each was sent to New Zealand Genomics Limited (NZGL) for library preparation and sequencing. Short read sequencing of Illumina TruSeq libraries in a HiSeq2500 generated 19,090,212 paired-end reads of 100 bp (average fragment size: 450 bp, adaptor length: 120 bp). A second sequencing (Illumina Nextera MiSeq Mate Pair) not size-selected generated 31,780,788 paired-end sequences of 250 bp (fragment size: 1.5 - 15 kb, median size: ~3 kb, adaptor length: 38 bp).

PreQC report was generated using the String Graph Assembler software package (Simpson 2014) and quality metrics before assembly with both FastQC (Andrews 2010) as well as MultiQC (Ewels et al. 2016) (Fig. S1). These analyses revealed that 91 % of sequences had a mean Phred quality score >= 30, 96 % of bases a mean Phred quality score >= 30, and 39 % of sequences an adapter sequence (either Illumina or Nextera). Adaptor trimming was performed with NxTrim (O'Connell et al. 2015) for the mate pair library, followed by Trimmomatic (Bolger et al. 2014) with the following options: MINLEN:40 ILLUMINACLIP:2:30:12:1:true LEADING:3 TRAILING:3 MAXINFO:40:0.4 MINLEN:40 for both libraries. After trimming, 86,644,308 paired-end (85 %) and 12,112,004 (12 %) single-end sequences remained (100 % with a mean Phred quality score >= 30, < 1 % with an adapter sequence).

Genome assembly

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

570 *De novo* assembly was performed in three consecutive iterations following a Meta-
571 assembly approach (Table S5). First, both libraries were assembled together in parallel, using
572 a k-mer size of 63 following the results from KmerGenie (Chikhi and Medvedev 2014)
573 whenever available, by five assemblers: AbySS (Simpson et al. 2009), Velvet (Zerbino and
574 Birney 2008), SOAPdenovo2 (Luo et al. 2012), ALLPATHS-LG (Gnerre et al. 2011), MaSuRCA
575 (Zimin et al. 2013). The MaSuRCA assembler was run twice, once running the adapter filtering
576 function (here termed "MaSuRCA-filtered"), the other without (termed simply "MaSuRCA").
577 Their respective quality was then estimated using three different metrics: the N50 length, the
578 BUSCO core-genes completion (Simão et al. 2015) and the Glimmer number of predicted
579 genes (Delcher et al. 1999). Second, these drafts were combined by following each ranking
580 using Metassembler (Wences and Schatz 2015), hence producing three new assemblies
581 (limiting the maximum insert size at 15 kb). Third, the *B. leachii* transcriptome (Zondag et al.
582 2016) was aligned to each meta-assembly using STAR (Dobin et al. 2013), which were then
583 combined thrice more using Metassembler following their alignment percentage and limiting
584 the maximum insert size at 3 kb, 8 kb and 15 kb. Finally, the quality of the meta-meta-
585 assemblies was estimated using the BUSCO score and the best one (Table S5) selected as the
586 reference *de novo* assembly.
587
588 **Data access**
589 All data was retrieved from the indicated sources in January 2016. Note that *Ciona*
590 *intestinalis* type A (Dehal et al. 2002b) has recently been recognized as a distinct species
591 (*Ciona robusta,* (Brunetti et al. 2015)) and that this study has been undertaken before it was
592 renamed.
593 *B. leachii*, *B. schlosseri*, *C. robusta*, *M. oculata*: Ascidian Network for *In Situ* Expression and
594 Embryonic Data (ANISEED, https://www.aniseed.cnrs.fr/aniseed/, (Tassy et al. 2010)).

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

595 *O. dioica*: Oikopleura Genome Browser
596 (http://www.genoscope.cns.fr/externe/GenomeBrowser/Oikopleura/, (Seo et al. 2001)).
597 *B. floridae*, *H. sapiens*: Joint Genome Institute (JGI, http://genome.jgi.doe.gov, (Grigoriev et al.
598 2012))
599
600 ***Repeat region analysis***
601 A *de novo* repeat library was build for each tunicate genome using RepeatModeler
602 (Smit and Hubley 2015). This utilizes the RECON tandem repeats finder from the RepeatScout
603 packages to identify species-specific repeats in a genome assembly. RepeatMasker (Smit et al.
604 2015) was then used to mask those repeats.
605
606 **Gene annotation**
607 *Ab initio* genome annotation was performed using MAKER2 (Holt and Yandell 2011)
608 with Augustus (Stanke and Waack 2003) and SNAP (Korf 2004) for gene prediction. In
609 addition, we used our previously published transcriptome (Zondag et al. 2016) and a
610 concatenation of UniProtKB (UniProt Consortium 2015), *C. robusta* and *B. schlosseri* proteins
611 into a custom proteome as evidence of gene product. Using the predicted genes, Augustus and
612 SNAP were then trained to the specificity of *B. leachii* genome. A second round of predictions
613 was then performed, followed by a second round of training. The final annotation of the
614 genome was obtained after running a third round of predictions, and the provided trained
615 Augustus and SNAP configurations after a third round of training. Non-coding RNA sequences
616 were then annotated using Infernal (Nawrocki and Eddy 2013) with Rfam library 12.0
617 (Nawrocki et al. 2015), tRNAscan-SE (Lowe and Eddy 1997) and snoRNA (Lowe 1999).
618 Finally, the identified sequences were characterized by InterProScan (Jones et al. 2014).
619

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

620 **Analysis of Gene Ontology terms**

621 Distribution of Gene Ontology (GO) terms were computed for each species as follows.
622 GO terms were extracted from the genome annotation and the number of occurrence for each
623 term determined using a custom Python script. The resulting list of frequencies was then
624 simplified using REVIGO (similarity factor "small" of 0.5, (Supek et al. 2011)) and the
625 `TreeMap` output retrieved. The hierarchy of every GO term present was reconstructed
626 following the schema defined by the core gene ontology (`go.obo`, (The Gene Ontology
627 Consortium 2015)) using a custom Python script selecting the shortest path to the root of the
628 tree, favouring smaller GO terms identification number in case of multiple paths. Finally,
629 frequencies were displayed using the `sunburstR` function of the Data-Driven Documents
630 library (D3, (Bostock et al. 2011)).

631 Predicted amino-acid sequences for all species were retrieved and clustered into
632 17,710 groups by OrthoMCL (Li et al. 2003). Protein sequences within each group were then
633 aligned into a Multiple Sequence Alignment (MSA) by Clustal-Omega, and the corresponding
634 consensus sequence inferred by EMBOSS cons. Consensus sequences were matched to the
635 Swiss-Prot curated database using BLASTp (e-value cut-off of 10^{-5}), and the GO terms
636 corresponding to the best match retrieved. GO terms frequencies were analysed as described
637 above and displayed using REVIGO's treemap. The overrepresentation analysis was
638 performed using GOrilla (Eden et al. 2009) with *Homo sapiens* as the organism background,
639 using a *p*-value threshold of 10^{-3} and REVIGO treemap (similarity factor "medium" of 0.7) for
640 visualization.

641

642 **Analysis of specific gene families**

643 Genes and transcripts for each examined genome were identified by a tBLASTn search
644 with an e-value cut-off at 10^{-5} using the SequencerServer software (Priyam et al. 2015). This

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

was followed by a reciprocal BLAST using SmartBLAST (NCBI Resource Coordinators 2016), to confirm their identity.

Delta serrate ligand conserved protein domain (PF01414) was used to identify the corresponding proteins in tunicate genomes. To identify *Notch* receptor genes the conserved LNR (lin-notch repeat) domain (PF00066) was used. ALDH-like genes were identified by tBLASTn search (PF00171) and classified using SMART blast.

Phylogenetics

Sequences were aligned with ClustalX (Jeanmougin et al. 1998) before using ProtTest 3 (Abascal et al. 2005) to determine the best-fit model of evolution. The best-fit model for the DSL phylogeny was WAG+I+G and, for CYP26 proteins, was LG+I+G.

Bayesian inference (BI) phylogenies were constructed using MrBayes (Ronquist and Huelsenbeck 2003) with a mixed model for 100,000 generations and summarized using a Sump burn-in of 200. Maximum Likelihood (ML) phylogenies were generated by PhyML (Guindon et al. 2010), using the estimated amino acid frequencies.

Accession numbers are provided in File S3 and sequence alignments are provided in Figure S3. Analyses carried out with BI and ML produced identical tree topologies.

Trees were displayed using FigTree v1.4.2 (Rambaut 2016).

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgements

Funding support was provided to M.J.W. by the Otago BMS Deans Bequest and Department of Anatomy. S.B. was supported by the Swiss National Science Foundation (SNSF) grant number P2ELP3_158873. We would like to thank Peter Maxwell and the New Zealand eScience Infrastructure (NeSI); Christelle Dantec and ANISEED for help and advice during the annotation process, as well as for the accompanying *B. leachii* genome browser.

Supplementary Figures

Fig. S1. SGA analysis. Including genome size estimation and de Brujin graph quantification.

Fig. S2. Gene Ontology terms identified in the larger orthologs clusters between the tunicate genomes.

Fig. S3. Protein sequence alignments used to generate the Notch and RA phylogenies.

Supplementary Files

File S1. BUSCO scores for the *B. leachii* genome assembly

File S2. Results of Repeatmasker analysis using de novo repeat libraries (Repeatmodeler)

File S3. Results from OrthoMCL group including the REVIGO GO term listed each orthologue group, used in the comparison of the GO terms between the tunicate genomes.

File S4. BLASTp, GOrillia and REVIGO results used used for overrepresentation analysis of GO terms present in the *B. leachii* and *B. schlosseri* only orthologue group analysis.

File S5. Corresponding Gene and Transcript IDs for *B. leachii* genes of interest. Accession numbers for protein sequences used in phylogeny construction.

Supplementary Tables

Table S1. Repetitive elements identified in the *B. leachii* and *B. schlosseri* genomes using Repeatmoduler and RepeatMasker.

Table S2. Comparison of the number of Wnt pathway genes.

Table S3. Comparison of the number of Notch pathway genes.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

704 **Table S4.** RA pathway components across annotated tunicate genomes.
705
706 **Table S5.** Iterative results of the meta-assembly approach followed for the *de novo* assembly
707 of the *B. leachii* genome
708

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**: 2104–2105. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti263.

Albalat R, Cañestro C. 2016. Evolution by gene loss. *Nat Rev Genet* **17**: 379–391. http://www.nature.com/doifinder/10.1038/nrg.2016.39.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Auger H, Sasakura Y, Joly J-S, Jeffery WR. 2010. Regeneration of oral siphon pigment organs in the ascidian Ciona intestinalis. *Dev Biol* **339**: 374–389. http://linkinghub.elsevier.com/retrieve/pii/S0012160609014651.

Ballarin L, Franchini A, Ottaviani E, Sabbadin A. 2001. Morula cells as the major immunomodulatory hemocytes in ascidians: Evidences from the colonial species *Botryllus schlosseri*. *Biol Bull* **201**: 59–64.

Belyaeva O V., Chang C, Berlett MC, Kedishvili NY. 2015. Evolutionary origins of retinoid active short-chain dehydrogenases/reductases of SDR16C family. *Chem Biol Interact* **234**: 135–143. http://linkinghub.elsevier.com/retrieve/pii/S000927971400324X.

Belyaeva O V., Lee S-A, Kolupaev O V., Kedishvili NY. 2009. Identification and characterization of retinoid-active short-chain dehydrogenases/reductases in Drosophila melanogaster. *Biochim Biophys Acta - Gen Subj* **1790**: 1266–1273. http://linkinghub.elsevier.com/retrieve/pii/S0304416509001688.

Berna L, Alvarez-Valin F. 2014. Evolutionary genomics of fast evolving tunicates. *Genome Biol Evol* **6**: 1724–1738. https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evu122.

Berrill NJ. 1951. Regeneration and Budding in Tunicates. *Biol Rev* **26**: 456–475. http://doi.wiley.com/10.1111/j.1469-185X.1951.tb01207.x.

Berrill NJ. 1947. The developmental cycle of *Botrylloides*. *Q J Microsc Sci* **88**: 393–407.

Berrill NJ. 1941. The development of the bud in *Botryllus*. *Biol Bull* **80**: 169. http://www.jstor.org/stable/1537595?origin=crossref.

Bock DG, MacIsaac HJ, Cristescu ME. 2012. Multilocus genetic analyses differentiate between widespread and spatially restricted cryptic species in a model ascidian. *Proc R Soc B Biol Sci* **279**: 2377–2385. http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2011.2610.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120. http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btu170.

Bostock M, Ogievetsky V, Heer J. 2011. D3: Data-Driven Documents. *IEEE Trans Vis Comput Graph*. http://vis.stanford.edu/papers/d3.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, et al. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**: 10. http://www.gigasciencejournal.com/content/2/1/10.

Brown FD, Swalla BJ. 2012. Evolution and development of budding by stem cells: Ascidian coloniality as a case study. *Dev Biol* **369**: 151–162. http://dx.doi.org/10.1016/j.ydbio.2012.05.038.

Brozovic M, Martin C, Dantec C, Dauga D, Mendez M, Simion P, Percher M, Laporte B, Scornavacca C, Di Gregorio A, et al. 2016. ANISEED 2015: A digital framework for the comparative developmental biology of ascidians. *Nucleic Acids Res* **44**: D808–D818.

Brunetti R. 1976. Biological cycle of *Botrylloides leachi* (Savigny) (Ascidiacea) in the Venetian lagoon. *Vie Milieu* **XXVI**: 105–122.

Brunetti R. 1974. *Observations on the Life Cycle of Botryllus Schlosseri (Pallas) (Ascidiacea) in*

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the Venetian Lagoon. http://www.tandfonline.com/doi/abs/10.1080/11250007409430119.

Brunetti R, Beghi L, Bressan M, Marin M. 1980. Combined effects of temperature and salinity on colonies of *Botryllus schlosseri* and *Botrylloides leachi* (Ascidiacea) from the Venetian Lagoon. *Mar Ecol Prog Ser* **2**: 303–314. http://www.int-res.com/articles/meps/2/m002p303.pdf.

Brunetti R, Gissi C, Pennati R, Caicci F, Gasparini F, Manni L. 2015. Morphological evidence that the molecularly determined Ciona intestinalis type A and type B are different species: Ciona robusta and Ciona intestinalis. *J Zool Syst Evol Res* **53**: 186–193. http://doi.wiley.com/10.1111/jzs.12101.

Burighel P, Brunetti R, Zaniolo G. 1976. Hibernation of the colonial ascidian *Botrylloides leachi* (Savigny): histological observations. *Ital J Zool* **43**: 293–301.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421. http://www.ncbi.nlm.nih.gov/pubmed/20003500.

Cañestro C, Bassham S, Postlethwait J. 2005. Development of the central nervous system in the larvacean Oikopleura dioica and the evolution of the chordate brain. *Dev Biol* **285**: 298–315. http://linkinghub.elsevier.com/retrieve/pii/S0012160605004410.

Canestro C, Postlethwait JH, Gonzalez-Duarte R, Albalat R. 2006. Is retinoic acid genetic machinery a chordate innovation? *Evol Dev* **8**: 394–406. http://doi.wiley.com/10.1111/j.1525-142X.2006.00113.x.

Chikhi R, Medvedev P. 2014. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**: 31–37. http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btt310.

Chillakuri CR, Sheppard D, Lea SM, Handford PA. 2012. Notch receptor–ligand binding and activation: Insights from molecular studies. *Semin Cell Dev Biol* **23**: 421–428. http://linkinghub.elsevier.com/retrieve/pii/S1084952112000134.

Cunningham TJ, Duester G. 2015. Mechanisms of retinoic acid signalling and its roles in organ and limb development. *Nat Rev Mol Cell Biol* **16**: 110–123. http://www.nature.com/doifinder/10.1038/nrm3932.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, et al. 2002a. The Draft Genome of Ciona intestinalis: Insights into Chordate and Vertebrate Origins. *Science (80-)* **298**: 2157–2167. http://www.sciencemag.org/content/298/5601/2157.abstract.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso AW, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, et al. 2002b. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**: 2157–67. http://www.ncbi.nlm.nih.gov/pubmed/12481130 (Accessed August 25, 2014).

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**: 4636–41. http://www.ncbi.nlm.nih.gov/pubmed/10556321.

Delsuc F, Brinkmann H, Chourrout D, Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**: 965–968. http://www.ncbi.nlm.nih.gov/pubmed/16495997 (Accessed July 10, 2014).

Dobin A, Davis C a, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3530905&tool=pmcentrez&rendertype=abstract.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**: 48. http://www.biomedcentral.com/1471-2105/10/48.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Edvardsen RB, Seo H-C, Jensen MF, Mialon A, Mikhaleva J, Bjordal M, Cartry J, Reinhardt R, Weissenbach J, Wincker P, et al. 2005. Remodelling of the homeobox gene complement in the tunicate Oikopleura dioica. *Curr Biol* **15**: R12–R13. http://linkinghub.elsevier.com/retrieve/pii/S096098220400973X.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* btw354.

Franchi N, Schiavon F, Carletto M, Gasparini F, Bertoloni G, Tosatto SCE, Ballarin L. 2011. Immune roles of a rhamnose-binding lectin in the colonial ascidian Botryllus schlosseri. *Immunobiology* **216**: 725–736. http://linkinghub.elsevier.com/retrieve/pii/S0171298510001993.

Garcia-Fernàndez J. 2005. The genesis and evolution of homeobox gene clusters. *Nat Rev Genet* **6**: 881–92. http://www.ncbi.nlm.nih.gov/pubmed/16341069.

Gasparini F, Burighel P, Manni L, Zaniolo G. 2008. Vascular regeneration and angiogenic-like sprouting mechanism in a compound ascidian is similar to vertebrates. *Evol Dev* **10**: 591–605.

Gazave E, Lapébie P, Richards GS, Brunet F, Ereskovsky A V, Degnan BM, Borchiellini C, Vervoort M, Renard E. 2009. Origin and evolution of the Notch signalling pathway: an overview from eukaryotic genomes. *BMC Evol Biol* **9**: 249. http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-9-249.

Gissi C, Hastings KEM, Gasparini F, Stach T, Pennati R, Manni L. 2017. An unprecedented taxonomic revision of a model organism: the paradigmatic case of Ciona robusta and Ciona intestinalis. *Zool Scr*. http://doi.wiley.com/10.1111/zsc.12233.

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci* **108**: 1513–1518. http://www.pnas.org/cgi/doi/10.1073/pnas.1017351108.

Griggio F, Voskoboynik A, Iannelli F, Justy F, Tilak M-KM-K, Xavier T, Pesole G, Douzery EJP, Mastrototaro F, Gissi C. 2014. Ascidian Mitogenomics: Comparison of Evolutionary Rates in Closely Related Taxa Provides Evidence of Ongoing Speciation Events. *Genome Biol Evol* **6**: 591–605. https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evu041.

Grigoriev I V., Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA, et al. 2012. The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res* **40**: D26–D32. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr947.

Guder C, Philipp I, Lengfeld T, Watanabe H, Hobmayer B, Holstein TW. 2006. The Wnt code: cnidarians signal the way. *Oncogene* **25**: 7450–7460. http://www.nature.com/doifinder/10.1038/sj.onc.1210052.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* **59**: 307–321. https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/syq010.

Guruharsha KG, Kankel MW, Artavanis-Tsakonas S. 2012. The Notch signalling system: recent insights into the complexity of a conserved pathway. *Nat Rev Genet* **13**: 654–666. http://www.nature.com/doifinder/10.1038/nrg3272.

Gutierrez S, Brown FD. 2017. Vascular budding in *Symplegma brakenhielmi* and the evolution of coloniality in styelid ascidians. *Dev Biol* **423**: 152–169. http://dx.doi.org/10.1016/j.ydbio.2017.01.012.

Hamada M, Goricki S, Byerly MS, Satoh N, Jeffery WR. 2015. Evolution of the chordate regeneration blastema: Differential gene expression and conserved role of notch signaling during siphon regeneration in the ascidian Ciona. *Dev Biol* **405**: 304–315. http://linkinghub.elsevier.com/retrieve/pii/S0012160615300701.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Hoegg S, Meyer A. 2005. Hox clusters as models for vertebrate genome evolution. *Trends Genet* **21**: 421–424. http://linkinghub.elsevier.com/retrieve/pii/S0168952505001654.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**: 491.

Ikuta T, Satoh N, Saiga H. 2010. Limited functions of Hox genes in the larval development of the ascidian Ciona intestinalis. *Development* **137**: 1505–1513. http://dev.biologists.org/cgi/doi/10.1242/dev.046938.

Iwasa T, Mishima S, Watari A, Ohkuma M, Azuma T, Kanehara K, Tsuda M. 2003. A novel G protein alpha subunit in embryo of the ascidian, Halocynthia roretzi. *Zoolog Sci* **20**: 141–51. http://www.ncbi.nlm.nih.gov/pubmed/12655177.

Jackson B, Brocker C, Thompson DC, Black W, Vasiliou K, Nebert DW, Vasiliou V. 2011. Update on the aldehyde dehydrogenase gene (ALDH) superfamily. *Hum Genomics* **5**: 283–303. http://www.ncbi.nlm.nih.gov/pubmed/21712190.

Janssen R, Le Gouar M, Pechmann M, Poulin F, Bolognesi R, Schwager EE, Hopfen C, Colbourne JK, Budd GE, Brown SJ, et al. 2010. Conservation, loss, and redeployment of Wnt ligands in protostomes: implications for understanding the evolution of segment formation. *BMC Evol Biol* **10**: 374. http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-10-374.

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. 1998. Multiple sequence alignment with Clustal X. *Trends Biochem Sci* **23**: 403–5. http://www.ncbi.nlm.nih.gov/pubmed/9810230.

Jeffery WR. 2015. Regeneration, Stem Cells, and Aging in the Tunicate Ciona: Insights from the Oral Siphon. *Int Rev Cell Mol Biol* **319**: 255–82. http://linkinghub.elsevier.com/retrieve/pii/S1937644815000581.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**: 1236–1240.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656–64. http://www.ncbi.nlm.nih.gov/pubmed/11932250.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**: 59.

Kürn U, Rendulic S, Tiozzo S, Lauzon RJ. 2011. Asexual propagation and regeneration in colonial ascidians. *Biol Bull* **221**: 43–61. http://www.ncbi.nlm.nih.gov/pubmed/21876110.

Kusserow A, Pang K, Sturm C, Hrouda M, Lentfer J, Schmidt HA, Technau U, von Haeseler A, Hobmayer B, Martindale MQ, et al. 2005. Unexpected complexity of the Wnt gene family in a sea anemone. *Nature* **433**: 156–160. http://www.nature.com/doifinder/10.1038/nature03158.

Lauzon RJ, Brown C, Kerr L, Tiozzo S. 2013. Phagocyte dynamics in a highly regenerative urochordate: insights into development and host defense. *Dev Biol* **374**: 357–73. http://www.ncbi.nlm.nih.gov/pubmed/23174529 (Accessed March 31, 2014).

Lee S-A, Belyaeva O V., Kedishvili NY. 2009. Biochemical characterization of human epidermal retinol dehydrogenase 2. *Chem Biol Interact* **178**: 182–7. http://www.ncbi.nlm.nih.gov/pubmed/18926804.

Lemaire P, Smith WC, Nishida H. 2008. Ascidians and the plasticity of the chordate developmental program. *Curr Biol* **18**: R620-31. http://www.ncbi.nlm.nih.gov/pubmed/18644342.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178–89. http://www.genome.org/cgi/doi/10.1101/gr.1224503.

Loh KM, van Amerongen R, Nusse R. 2016. Generating Cellular Diversity and Spatial Form: Wnt Signaling and the Evolution of Multicellular Animals. *Dev Cell* **38**: 643–655. http://linkinghub.elsevier.com/retrieve/pii/S153458071630586X.

Lowe TM. 1999. A Computational Screen for Methylation Guide snoRNAs in Yeast. *Science (80-)* **283**: 1168–1171. http://www.sciencemag.org/cgi/doi/10.1126/science.283.5405.1168.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: A program for improved detection of transfer RNA

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

genes in genomic sequence. *Nucleic Acids Res* **25**: 955–964.
Luke GN, Castro LFC, McLay K, Bird C, Coulson A, Holland PWH. 2003. Dispersal of NK homeobox gene clusters in amphioxus and humans. *Proc Natl Acad Sci* **100**: 5292–5295. http://www.pnas.org/cgi/doi/10.1073/pnas.0836141100.
Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**: 18. http://www.gigasciencejournal.com/content/1/1/18.
MacDonald BT, Tamai K, He X. 2009. Wnt/β-Catenin Signaling: Components, Mechanisms, and Diseases. *Dev Cell* **17**: 9–26. http://linkinghub.elsevier.com/retrieve/pii/S1534580709002573.
Manni L, Zaniolo G, Cima F, Burighel P, Ballarin L. 2007. *Botryllus schlosseri*: a model ascidian for the study of asexual reproduction. *Dev Dyn* **236**: 335–52. http://www.ncbi.nlm.nih.gov/pubmed/17191252 (Accessed March 31, 2014).
Martí-Solans J, Belyaeva O V., Torres-Aguila NP, Kedishvili NY, Albalat R, Cañestro C. 2016. Coelimination and Survival in Gene Network Evolution: Dismantling the RA-Signaling in a Chordate. *Mol Biol Evol* **33**: 2401–2416. https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw118.
Martin A, Maher S, Summerhurst K, Davidson D, Murphy P. 2012. Differential deployment of paralogous Wnt genes in the mouse and chick embryo during development. *Evol Dev* **14**: 178–195. http://doi.wiley.com/10.1111/j.1525-142X.2012.00534.x.
Maumus F, Quesneville H. 2014. Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. *Nat Commun* **5**. http://www.nature.com/doifinder/10.1038/ncomms5104.
Millar RH. 1971. The biology of ascidians. *Adv Mar Biol* **9**: 1–100.
Mukai H, Sugimoto K, Taneda Y. 1978. Comparative studies on the circulatory system of the compound ascidians, *Botryllus*, *Botrylloides* and *Symplegma*. *J Morphol* **157**: 49–78.
Murata Y, Okado H, Kubo Y. 2001. Characterization of heteromultimeric G protein-coupled inwardly rectifying potassium channels of the tunicate tadpole with a unique pore property. *J Biol Chem* **276**: 18529–39. http://www.ncbi.nlm.nih.gov/pubmed/11278535.
Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, et al. 2015. Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Res* **43**: D130–D137.
Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**: 2933–2935.
NCBI Resource Coordinators. 2016. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **44**: D7-19. http://www.ncbi.nlm.nih.gov/pubmed/26615191.
Nishida H. 2005. Specification of embryonic axis and mosaic development in ascidians. *Dev Dyn* **233**: 1177–1193. http://doi.wiley.com/10.1002/dvdy.20469.
Nydam ML, Giesbrecht KB, Stephenson EE. 2017. Origin and Dispersal History of Two Colonial Ascidian Clades in the Botryllus schlosseri Species Complex ed. T.-Y. Chiang. *PLoS One* **12**: e0169944. http://dx.plos.org/10.1371/journal.pone.0169944.
O'Connell J, Schulz-Trieglaff O, Carlson E, Hims MM, Gormley N a., Cox a. J. 2015. NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics* **31**: btv057. http://bioinformatics.oxfordjournals.org/content/early/2015/02/05/bioinformatics.btv057.short?rss=1.
O'Neill RJ, O'Neill MJ, Graves JA. 1998. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**: 68–72. http://www.nature.com/doifinder/10.1038/29985.
Oka H, Watanabe H. 1957. Vascular budding, a new type of budding in *Botryllus*. *Biol Bull* **112**: 225. http://www.jstor.org/stable/10.2307/1539200?origin=crossref.
Pascual-Anaya J, D'Aniello S, Kuratani S, Garcia-Fernàndez J. 2013. Evolution of Hox gene

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

clusters in deuterostomes. *BMC Dev Biol* **13**: 26. http://bmcdevbiol.biomedcentral.com/articles/10.1186/1471-213X-13-26.

Pearson JC, Lemons D, McGinnis W. 2005. Modulating Hox gene functions during animal body patterning. *Nat Rev Genet* **6**: 893–904. http://www.ncbi.nlm.nih.gov/pubmed/16341070.

Philips A, Blein M, Robert A, Chambon J-P, Baghdiguian S, Weill M, Fort P. 2003. Ascidians as a vertebrate-like model organism for physiological studies of Rho GTPase signaling. *Biol cell* **95**: 295–302. http://www.ncbi.nlm.nih.gov/pubmed/12941527.

Piette J, Lemaire P. 2015. Thaliaceans, The Neglected Pelagic Relatives of Ascidians: A Developmental and Evolutionary Enigma. *Q Rev Biol* **90**: 117–145. http://www.journals.uchicago.edu/doi/10.1086/669266.

Primmer CR, Papakostas S, Leder EH, Davis MJ, Ragan MA. 2013. Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research. *Mol Ecol* **22**: 3216–3241. http://doi.wiley.com/10.1111/mec.12309.

Priyam A, Woodcroft BJ, Rai V, Munagala A, Moghul I, Ter F, Gibbins MA, Moon H, Leonard G, Rumpf W, et al. 2015. Sequenceserver: a modern graphical user interface for custom BLAST databases. *bioRxiv*. http://biorxiv.org/content/early/2015/11/27/033142.abstract.

Prud'homme B, Lartillot N, Balavoine G, Adoutte A, Vervoort M. 2002. Phylogenetic analysis of the Wnt gene family. Insights from lophotrochozoan members. *Curr Biol* **12**: 1395. http://www.sciencedirect.com/science/article/pii/S0960982202010680.

Rambaut A. 2016. FigTree. http://tree.bio.ed.ac.uk/software/figtree/.

Rinkevich B, Shlemberg Z, Fishelson L. 1995. Whole-body protochordate regeneration from totipotent blood cells. *Proc Natl Acad Sci U S A* **92**: 7695–9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=41212&tool=pmcentrez&rendertype=abstract (Accessed March 31, 2014).

Rinkevich Y, Douek J, Haber O, Rinkevich B, Reshef R. 2007a. Urochordate whole body regeneration inaugurates a diverse innate immune signaling profile. *Dev Biol* **312**: 131–46. http://www.ncbi.nlm.nih.gov/pubmed/17964563 (Accessed March 31, 2014).

Rinkevich Y, Paz G, Rinkevich B, Reshef R. 2007b. Systemic bud induction and retinoic acid signaling underlie whole body regeneration in the urochordate *Botrylloides leachi*. *PLoS Biol* **5**: e71. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1808485&tool=pmcentrez&rendertype=abstract (Accessed March 31, 2014).

Rinkevich Y, Rinkevich B, Reshef R. 2008. Cell signaling and transcription factor genes expressed during whole body regeneration in a colonial chordate. *BMC Dev Biol* **8**: 100. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2576188&tool=pmcentrez&rendertype=abstract (Accessed March 20, 2014).

Rinkevich Y, Voskoboynik A, Rosner A, Rabinowitz C, Paz G, Oren M, Douek J, Alfassi G, Moiseeva E, Ishizuka KJ, et al. 2013. Repeated, long-term cycling of putative stem cells between niches in a basal chordate. *Dev Cell* **24**: 76–88. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3810298&tool=pmcentrez&rendertype=abstract (Accessed March 30, 2014).

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–4. http://www.ncbi.nlm.nih.gov/pubmed/12912839.

Ross AC, Zolfaghari R. 2011. Cytochrome P450s in the regulation of cellular retinoic acid metabolism. *Annu Rev Nutr* **31**: 65–87. http://www.ncbi.nlm.nih.gov/pubmed/21529158.

Rubinstein ND, Feldstein T, Shenkar N, Botero-Castro F, Griggio F, Mastrototaro F, Delsuc F, Douzery EJP, Gissi C, Huchon D. 2013. Deep sequencing of mixed total DNA without barcodes allows efficient assembly of highly plastic Ascidian mitochondrial genomes. *Genome Biol Evol* **5**: 1185–1199. http://www.ncbi.nlm.nih.gov/pubmed/23709623.

Sabbadin A, Zaniolo G, Majone F. 1975. Determination of polarity and bilateral asymmetry in

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

palleal and vascular buds of the ascidian Botryllus schlosseri. *Dev Biol* **46**: 79–87.
Saito Y, Shirae M, Okuyama M, Cohen S. 2001. Phylogeny of Botryllid Ascidians. In *The Biology of Ascidians*, pp. 315–320, Springer Japan, Tokyo http://link.springer.com/10.1007/978-4-431-66982-1_50.
Santagati F, Abe K, Schmidt V, Schmitt-John T, Suzuki M, Yamamura K-I, Imai K. 2003. Identification of Cis-regulatory elements in the mouse Pax9/Nkx2-9 genomic region: implication for evolutionary conserved synteny. *Genetics* **165**: 235–42. http://www.ncbi.nlm.nih.gov/pubmed/14504231.
Sattler S. 2017. The Role of the Immune System Beyond the Fight Against Infection. In *The Immunology of Cardiovascular Homeostasis and Pathology* (eds. S. Sattler and T. Kennedy-Lydon), pp. 3–14, Springer International Publishing http://link.springer.com/10.1007/978-3-319-57613-8_1.
Savigny J-C. 1816. *Mémoires sur les animaux sans vertèbres*. Dufour, G., Paris http://www.biodiversitylibrary.org/bibliography/9154.
Seo H-CC, Kube M, Edvardsen RB, Jensen MF, Beck A, Spriet E, Gorsky G, Thompson EM, Lehrach H, Reinhardt R, et al. 2001. Miniature genome in the marine chordate *Oikopleura dioica*. *Science* **294**: 2506. http://www.sciencemag.org/content/294/5551/2506.short.
Simakov O, Kawashima T, Marlétaz F, Jenkins J, Koyanagi R, Mitros T, Hisata K, Bredeson J, Shoguchi E, Gyoja F, et al. 2015. Hemichordate genomes and deuterostome origins. *Nature* **527**: 459–465. http://www.nature.com/doifinder/10.1038/nature16150.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V. 2015. BUSCO : assessing genome assembly and annotation completeness with single-copy orthologs. *Genome Anal* 9–10.
Simmen MW, Leitgeb S, Charlton J, Jones SJ, Harris BR, Clark VH, Bird A. 1999. Nonmethylated transposable elements and methylated genes in a chordate genome. *Science* **283**: 1164–7. http://www.ncbi.nlm.nih.gov/pubmed/10024242.
Simpson JT. 2014. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics* **30**: 1228–1235.
Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Res* **19**: 1117–1123. http://genome.cshlp.org/cgi/doi/10.1101/gr.089532.108.
Small KS, Brudno M, Hill MM, Sidow A. 2007. A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. *Genome Biol* **8**: R41. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-3-r41.
Smit A, Hubley R. 2015. RepeatModeler Open-1.0.
Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0.
Sobreira TJP, Marletaz F, Simoes-Costa M, Schechtman D, Pereira AC, Brunet F, Sweeney S, Pani A, Aronowicz J, Lowe CJ, et al. 2011. Structural shifts of aldehyde dehydrogenase enzymes were instrumental for the early evolution of retinoid-dependent axial patterning in metazoans. *Proc Natl Acad Sci* **108**: 226–231. http://www.pnas.org/cgi/doi/10.1073/pnas.1011223108.
Somorjai IML, Escrivà H, Garcia-Fernàndez J. 2012. Amphioxus makes the cut—Again. *Commun Integr Biol* **5**: 499–502. http://www.tandfonline.com/doi/abs/10.4161/cib.21075.
Spagnuolo A, Ristoratore F, Di Gregorio A, Aniello F, Branno M, Di Lauro R. 2003. Unusual number and genomic organization of Hox genes in the tunicate Ciona intestinalis. *Gene* **309**: 71–9. http://www.ncbi.nlm.nih.gov/pubmed/12758123.
Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**: 215–225.
Stapley J, Santure AW, Dennis SR. 2015. Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Mol Ecol* **24**: 2241–2252. http://doi.wiley.com/10.1111/mec.13089.
Stoick-Cooper CL, Weidinger G, Riehle KJ, Hubbert C, Major MB, Fausto N, Moon RT. 2007.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Distinct Wnt signaling pathways have opposing roles in appendage regeneration. *Development* **134**: 479–89. http://dev.biologists.org/cgi/doi/10.1242/dev.001123.

Stolfi A, Lowe EK, Racioppi C, Ristoratore F, Brown CT, Swalla BJ, Christiaen L. 2014. Divergent mechanisms regulate conserved cardiopharyngeal development and gene expression in distantly related ascidians. *Elife* **3**: e03728.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms ed. C. Gibas. *PLoS One* **6**: e21800. http://dx.plos.org/10.1371/journal.pone.0021800.

Suzuki MM, Kerr ARW, De Sousa D, Bird A. 2007. CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res* **17**: 625–631. http://www.genome.org/cgi/doi/10.1101/gr.6163007.

Takatori N, Butts T, Candiani S, Pestarino M, Ferrier DEK, Saiga H, Holland PWH. 2008. Comprehensive survey and classification of homeobox genes in the genome of amphioxus, Branchiostoma floridae. *Dev Genes Evol* **218**: 579–590. http://link.springer.com/10.1007/s00427-008-0245-9.

Taketa DA, Nydam ML, Langenbacher AD, Rodriguez D, Sanders E, De Tomaso AW. 2015. Molecular evolution and in vitro characterization of Botryllus histocompatibility factor. *Immunogenetics* **67**: 605–623. http://link.springer.com/10.1007/s00251-015-0870-1.

Tassy O, Dauga D, Daian F, Sobral D, Robin F, Khoueiry P, Salgado D, Fox V, Caillol D, Schiappa R, et al. 2010. The ANISEED database: Digital representation, formalization, and elucidation of a chordate developmental program. *Genome Res* **20**: 1459–1468. http://genome.cshlp.org/cgi/doi/10.1101/gr.108175.110.

The Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res* **43**: D1049–D1056. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku1179.

Tsagkogeorga G, Cahais V, Galtier N. 2012. The Population Genomics of a Fast Evolver: High Levels of Diversity, Functional Constraint, and Molecular Adaptation in the Tunicate Ciona intestinalis. *Genome Biol Evol* **4**: 852–861. https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evs054.

Tsagkogeorga G, Turon X, Galtier N, Douzery EJP, Delsuc F. 2010. Accelerated Evolutionary Rate of Housekeeping Genes in Tunicates. *J Mol Evol* **71**: 153–167. http://link.springer.com/10.1007/s00239-010-9372-9.

UniProt Consortium. 2015. UniProt: a hub for protein information. *Nucleic Acids Res* **43**: D204–D212. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku989.

Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B, Fan HC, Mantalas GL, Palmeri KJ, et al. 2013a. The genome sequence of the colonial chordate, *Botryllus schlosseri*. *Elife* **2**: 1–24. http://elifesciences.org/lookup/doi/10.7554/eLife.00569.

Voskoboynik A, Newman AM, Corey DM, Sahoo D, Pushkarev D, Neff NF, Passarelli B, Koh W, Ishizuka KJ, Palmeri KJ, et al. 2013b. Identification of a Colonial Chordate Histocompatibility Gene. *Science (80-)* **341**: 384–387. http://www.sciencemag.org/cgi/doi/10.1126/science.1238036.

Voskoboynik A, Simon-Blecher N, Soen Y, Rinkevich B, De Tomaso AW, Ishizuka KJ, Weissman IL. 2007. Striving for normality: whole body regeneration through a series of abnormal generations. *FASEB J* **21**: 1335–44. http://www.ncbi.nlm.nih.gov/pubmed/17289924 (Accessed March 31, 2014).

Wada S, Tokuoka M, Shoguchi E, Kobayashi K, Di Gregorio A, Spagnuolo A, Branno M, Kohara Y, Rokhsar D, Levine M, et al. 2003. A genomewide survey of developmentally relevant genes in Ciona intestinalis. *Dev Genes Evol* **213**: 222–234. http://link.springer.com/10.1007/s00427-003-0321-0.

Wang X-P, Suomalainen M, Felszeghy S, Zelarayan LC, Alonso MT, Plikus M V, Maas RL, Chuong C-M, Schimmang T, Thesleff I. 2007. An integrated gene regulatory network controls

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

stem cell proliferation in teeth. *PLoS Biol* **5**: e159. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1885832&tool=pmcentrez&rendertype=abstract (Accessed October 28, 2010).

Wences AH, Schatz MC. 2015. Metassembler: merging and optimizing de novo genome assemblies. *Genome Biol* **16**: 207. http://genomebiology.com/2015/16/1/207.

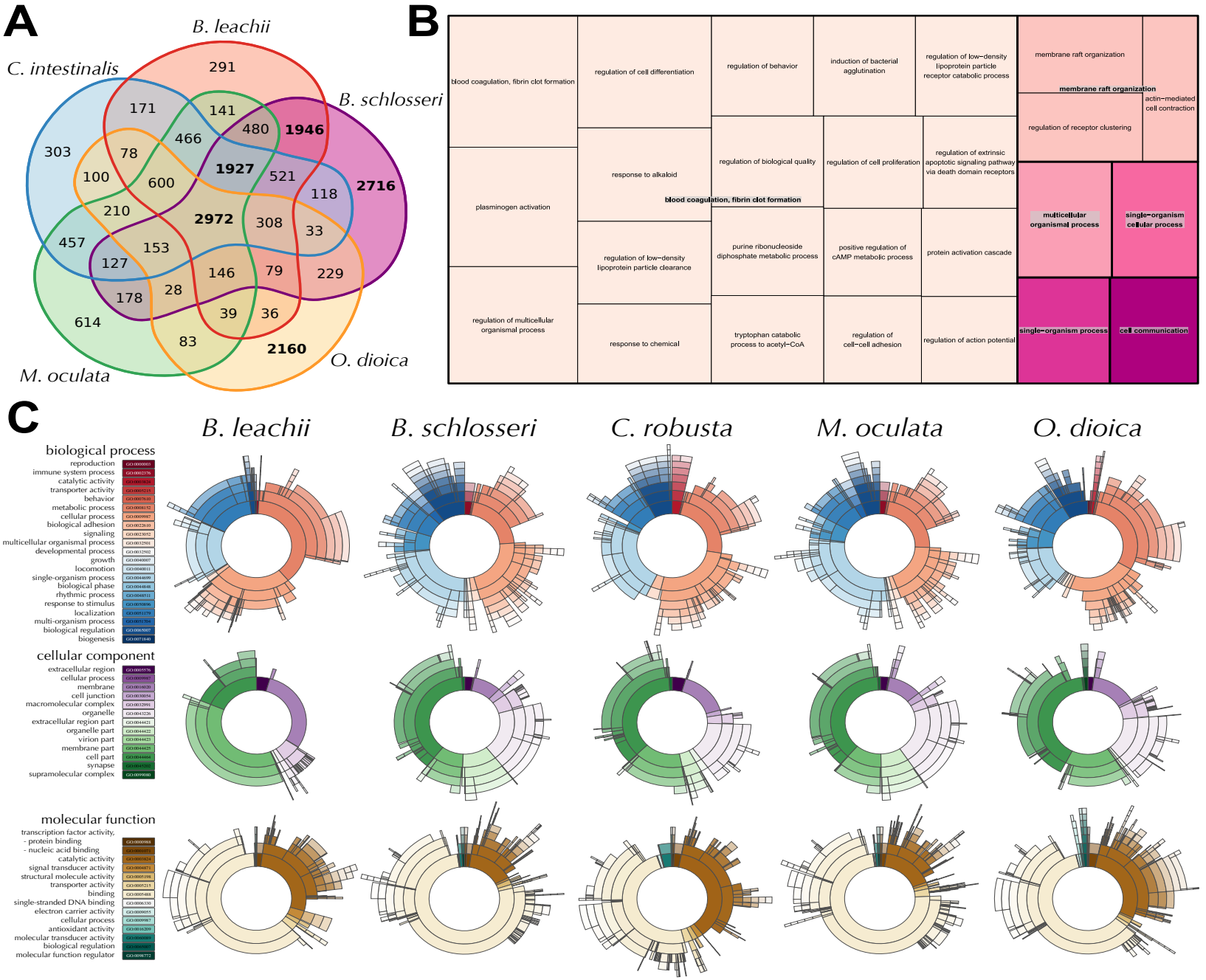
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–9. http://www.ncbi.nlm.nih.gov/pubmed/18349386.

Zhan A, Briski E, Bock DG, Ghabooli S, MacIsaac HJ. 2015. Ascidians as models for studying invasion success. *Mar Biol* **162**. http://link.springer.com/10.1007/s00227-015-2734-5.

Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, Yang P, Zhang L, Wang X, Qi H, et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**: 49–54. http://dx.doi.org/10.1038/nature11413.

Zimin A V., Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**: 2669–2677. http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btt476.

Zondag LE, Rutherford K, Gemmell NJ, Wilson MJ. 2016. Uncovering the pathways underlying whole body regeneration in a chordate model, *Botrylloides leachi* using de novo transcriptome analysis. *BMC Genomics* **17**: 114. http://www.biomedcentral.com/1471-2164/17/114.


A
B. leachii
C. intestinalis
B. schlosseri
M. oculata
O. dioica
291
171
141
480
1946
303
78
466
1927
521
2716
100
600
118
210
2972
308
33
457
153
127
146
79
229
178
28
39
36
614
83
2160
B
blood coagulation, fibrin clot formation
regulation of cell differentiation
regulation of behavior
induction of bacterial agglutination
regulation of low-density lipoprotein particle receptor catabolic process
membrane raft organization
membrane raft organization
actin-mediated cell contraction
regulation of receptor clustering
plasminogen activation
response to alkaloid
blood coagulation, fibrin clot formation
regulation of biological quality
regulation of cell proliferation
regulation of extrinsic apoptotic signaling pathway via death domain receptors
multicellular organismal process
single-organism cellular process
regulation of low-density lipoprotein particle clearance
purine ribonucleoside diphosphate metabolic process
positive regulation of cAMP metabolic process
protein activation cascade
regulation of multicellular organismal process
response to chemical
tryptophan catabolic process to acetyl-CoA
regulation of cell-cell adhesion
regulation of action potential
single-organism process
cell communication
C
B. leachii
B. schlosseri
C. robusta
M. oculata
O. dioica
biological process
reproduction
immune system process
catalytic activity
transporter activity
behavior
metabolic process
cellular process
biological adhesion
signaling
multicellular organismal process
developmental process
growth
locomotion
single-organism process
biological phase
rhythmic process
response to stimulus
localization
multi-organism process
biological regulation
biogenesis
cellular component
extracellular region
cellular process
membrane
cell junction
macromolecular complex
organelle
extracellular region part
organelle part
virion part
membrane part
cell part
synapse
supramolecular complex
molecular function
transcription factor activity,
- protein binding
- nucleic acid binding
catalytic activity
signal transducer activity
structural molecule activity
transporter activity
binding
single-stranded DNA binding
electron carrier activity
cellular process
antioxidant activity
molecular transducer activity
biological regulation
molecular function regulator

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

222 ***Ancient gene linkages are fragmented in tunicate genomes***
223
224 Ancient gene linkages are highly conserved sets of genes that are spatially restricted,
225 commonly occurring in clusters (Garcia-Fernàndez 2005). These clusters arose in a common
226 ancestor and were preserved because of a common regulatory mechanism such as cis-
227 regulatory elements located within the cluster. The homeobox-containing *Hox* gene family,
228 typically composed of 13 members in vertebrates (Hoegg and Meyer 2005), is among the best-
229 studied examples of such an ancient gene cluster and is critical for the correct embryonic
230 development (Pearson et al. 2005). The linear genomic arrangement of genes within the *Hox*
231 cluster reflects their spatial expression along the anterior-posterior body axis (Pascual-Anaya
232 et al. 2013), which establishes regional identity across this axis. The basal cephalochordate *B.*
233 *floridae* has all 13 *hox* genes located in a single stereotypical cluster, along with an additional
234 14th gene (Fig. 3B; (Takatori et al. 2008)), suggesting that the chordate ancestor also had an
235 intact cluster. However in tunicates, this clustering appears to be lost. In *C. robusta*, the nine
236 identified *Hox* genes are distributed across five scaffolds, with linkages preserved only
237 between *Hox2, Hox3* and *Hox4; Hox5* and *Hox6*; *Hox12* and *Hox13* (Fig. 3; (Spagnuolo et al.
238 2003; Wada et al. 2003)). In *O. dioica,* the total number of *Hox* genes is further reduced to
239 eight, split between 6 scaffolds, including a duplication of *Hox9* (Fig. 3A; (Edvardsen et al.
240 2005)). In *M. oculata* we could identify only six *Hox* genes, divided between 4 scaffolds, with
241 clustering retained for the *Hox10*, *Hox11* and *Hox12* genes (Fig. 3). In Botryllidae genomes, the
242 same seven *Hox* genes are conserved (Fig 3B), with a preserved linkage between *Hox10*,
243 *Hox12* and *Hox13* in *B. leachii* and three copies of *Hox5* present in *B. schlosseri*. Altogether, the
244 separation of the tunicate *Hox* cluster genes supports the hypothesis that reduction and
245 separation of this ancient gene linkage occurred at the base of tunicate lineage (Edvardsen et
246 al. 2005). In addition, *Hox9* appears to be specifically retained in neotenic Tunicates while
247 there is no pattern of conserved *Hox* cluster genes specific to colonial ascidians.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

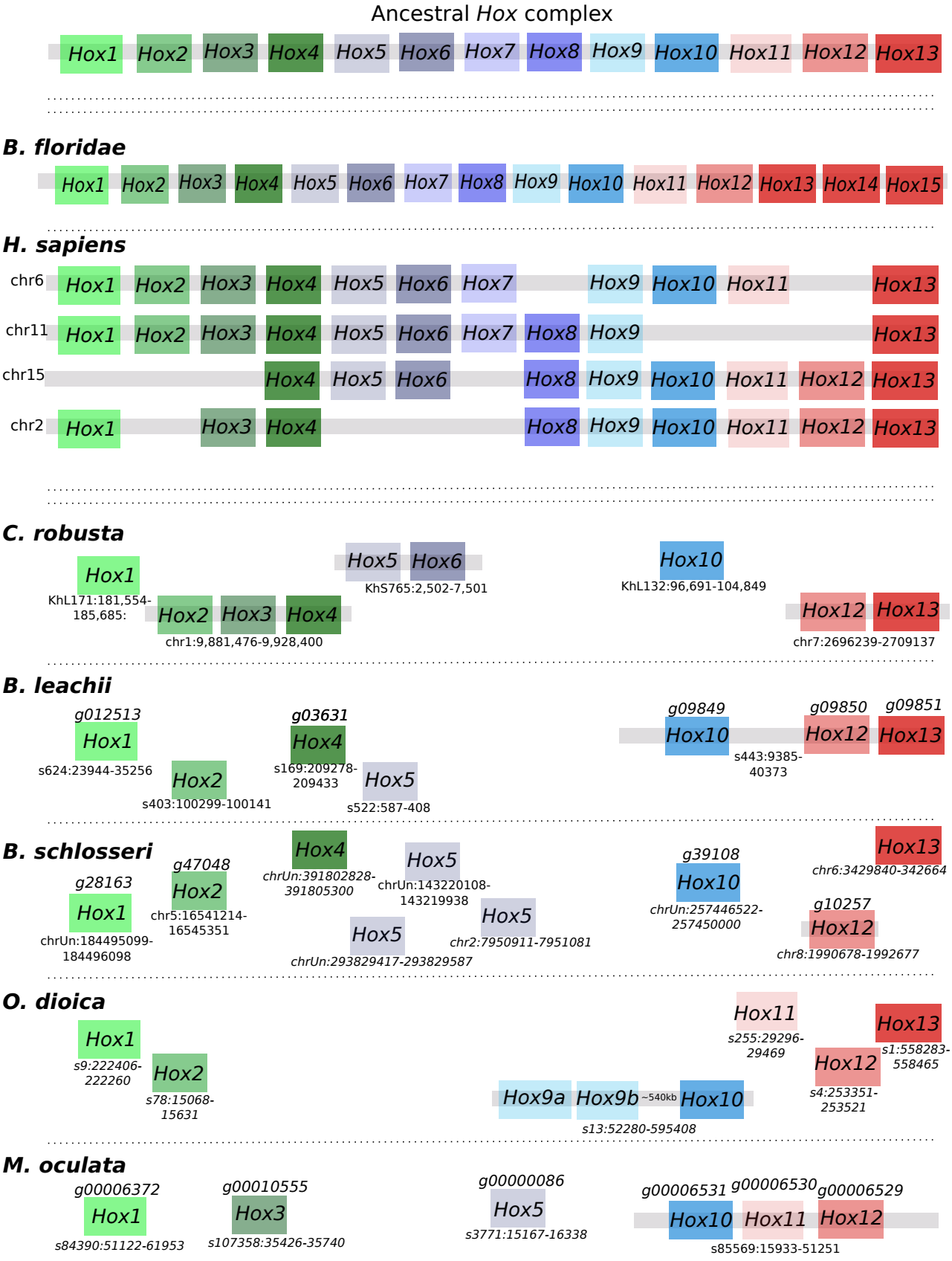


Figure 3. ***Hox*** **genes are dispersed and reduced in number within tunicate genomes.** Schematic depicting *Hox* gene linkages retained in five tunicate genomes in comparison to the ancestral *Hox* complex, which included thirteen genes. Orthologous genes are indicated by common colours. Chromosome or scaffold number is shown, along with gene ID when available for newly annotated genomes.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A second ancient homeobox-containing gene linkage is the *NK* cluster. This cluster, predicted to be present in the last common ancestor of bilaterians (Luke et al. 2003), consists of *Msx, Lbx*, *Tlx*, *NKx1, NKx3, NKx4* and *NKx5* (Fig. 4). In *B. floridae*, linkages between *Msx, NKx4* and *NKx3*; as well as between *Lbx* and *Tlx* provide evidence of retained ancestral clustering while *NKx5* was lost (Fig. 4; (Luke et al. 2003)). However in vertebrates, NKx5 is still present, while only the gene linkages between *Lbx* and *Tlx* as well as between *Nkx4* and *Nkx3* remain (Fig. 4; (Garcia-Fernàndez 2005)). To further clarify the evolution of this ancestral cluster in tunicates, we determined the structure of the NK cluster within five ascidian genomes. In all these species, *NKx1* is absent and no evidence of clustering could be found with all identified orthologs located on different scaffolds (Fig. 4). In *C. robusta*, *M. oculata* and *O. dioica* only five members of this cluster remain, with the loss of either *Lbx* or *Tlx* as well as of *NKx3* and the duplication of the ortholog of *NKx4* (Fig. 4). By contrast, in the colonial tunicates *B. leachii* and *B. schlosseri*, *Tbx*, *Lbx* and *NKx3* are all present. In *B. schlosseri*, *Msx1* is absent and *NKx4* duplicated. In the *B. leachii* genome, *NK1* is the only ancestral cluster member to be missing and *Nk5* has been duplicated (Fig. 4). Altogether, these results suggest that there has been a loss of *NKx5* in Cephalochordate, one of *NKx1* in Tunicate and that the combination of *NKx3*, *Lbx* and *Tbx* may be specific to colonial ascidians.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

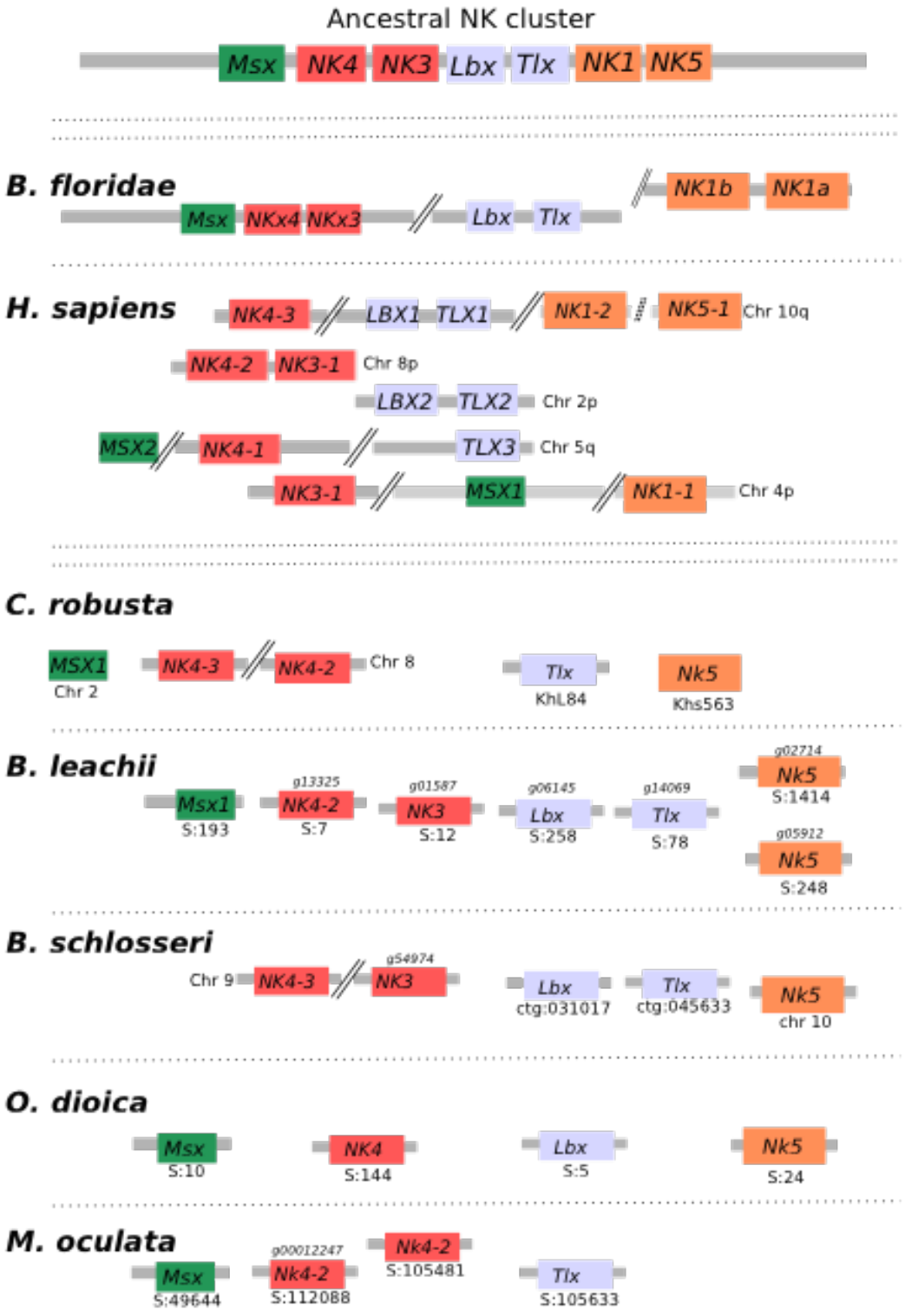


Figure 4. NK homeobox cluster genes are fragmented within tunicate genomes.
Schematic depicting the organization of the *NK homeobox* cluster genes among the studied chordate genomes. Double-parallel lines indicate > 1Mb distance between genes. Chromosome or scaffold number is shown, along with gene ID when available for newly annotated genomes. Orthologous genes are indicated by common colours.

266

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A third ancient linkage that we investigated is the pharyngeal cluster, a gene group present in hemichordates, echinoderm and vertebrates genomes that is considered to be Deuterosome specific (Simakov et al. 2015). The cluster groups *foxhead domain protein* (*FoxA*), *NKx2 (NKx2.2 and Nkx2.1)*, *Pax1/9*, *mitochondrial solute carrier family 25 member 21* (*slc25A21)*, *mirror-image polydactyly 1 protein* (*mipol1*), *egl nine homolog 3* (*egln3*) and *dehydrogenase/reductase member* 7 (*dhrs7*). Among these, s*lc25a21, Pax1/9*, *mipol1* and *FoxA* pairs are also found in protostomes suggesting an even more ancient origin (Simakov et al. 2015). The pharyngeal cluster is thought to have arisen due to the location of the regulatory elements of *Pax1/9* and *FoxA* within the introns of *slc25A21* and *mipol1* (Santagati et al. 2003; Wang et al. 2007), constraining the genes to remain in tight association with each other. In the *B. floridae* genome, the entire cluster is located on the same scaffold, with the exception of the *Nkx2.1* and *Nk2.2* gene pair located on separate scaffold. In *C. robusta*, only orthologs of *FoxA*, slc25a29, *Pax1* and *Pax9* could be identified. Nevertheless, all of them are located on the same chromosome (Fig. 5). In *O. dioica*, the cluster appears even further reduced. While orthologs of *FoxA*, *Pax1/9* and *Nkx2.2* genes were found on different scaffolds, only one rather distant linkage (> 1 Mb) between *a Pax-like* gene and *slc25A21* is retained. For both *B. schlosseri* and *M. oculata*, there was no evidence of clustering between genes (Fig. 5). In the *B. leachii* genome, *mipol1* is the sole missing gene from this cluster. However, only the pairing of a *Pax-like* and *slc25A21* genes remains (Fig. 5). Altogether, these results suggest that most of the Tunicates did not conserve the structure of this ancient linkage, but it is unknown what consequences this would have to their expression and function.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

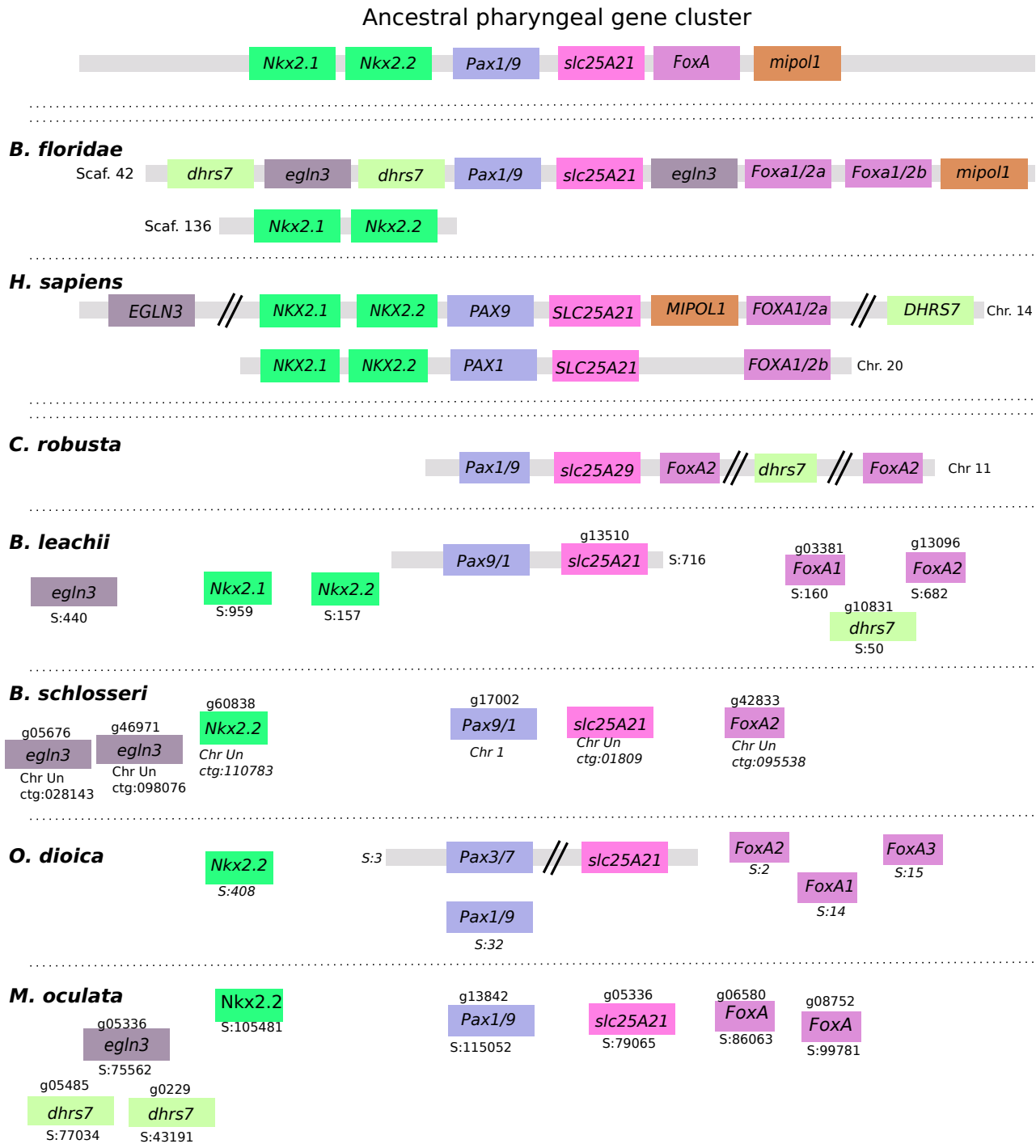


Figure 5. Ancestral gene linkages remain between a few pharyngeal cluster genes in tunicate genomes. Gene order of the six pharyngeal cluster genes, *NK2.1, NK2.2, Pax1/9* and *FoxA* in chordate genomes. Double-parallel lines indicate > 1 Mb distance between genes. Chromosome or scaffold number is shown, along with gene ID when available for newly annotated genomes. Orthologous genes are indicated by common colours.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

290 ***Lineage-specific changes to cell-signalling pathways in Botryllidae genomes.***

291 To dissect more specifically the evolution of colonial ascidians, we examined the
292 genomes of *B. leachii* and *B. schlosseri*, looking for key components of signalling pathways
293 required for metazoan development and regeneration. Of particular interest, we focused on
294 the Wingless-related integration site (Wnt), Notch and Retinoic acid (RA) signalling pathways.
295 All three of these pathways have been implicated in WBR and asexual reproduction in colonial
296 tunicates (Rinkevich et al. 2008, 2007b; Zondag et al. 2016).

297

298 **Wnt pathway**

299 Wnt ligands are secreted glycoproteins that have roles in axis patterning,
300 morphogenesis and cell specification (Loh et al. 2016). The ancestral gene family appears to
301 originate very early on during multi-cellular evolution and to be composed of eleven members
302 (Kusserow et al. 2005; Guder et al. 2006). The *Wnt* gene family expanded to 19 members in
303 the human genome, while independent gene loss has reduced this family to 7 genes in
304 *Drosophila melanogaster* and *Caenorhabditis elegans* (Prud'homme et al. 2002). Consequently,
305 we set out to investigate whether the *Wnt* gene family has either expanded or contracted
306 during Tunicata speciation.

307 We found an increase in the number of *Wnt5a* genes among Styelidae genomes. In *B.*
308 *schlosseri*, we identified 15 *Wnt* members, including seven *Wnt5a-like* genes on multiple
309 scaffolds (Fig. 6, Table S2). In the *B. leachii* genome, fourteen *Wnt* ligand genes were
310 identified, including four *Wnt5a* genes located on the same scaffold near *Wnt4* (Fig. 6). *M.*
311 *oculata* has only 7 Wnt ligand genes, including three *Wnt5a-like* genes (Fig. 6, Table S2). In
312 comparison, *C. robusta* has a total of 11 *Wnt* genes, including a single *Wnt5a* gene (Fig. 6,
313 Table S2; (Wada et al. 2003)). In the compact *O. dioica* genome, this number has reduced to 6
314 (Wnts 3, 4, 7, 11 and 16), none of which are *Wnt5a* orthologs (Table S2). Overall, this suggests

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

315 that an expansion through gene duplication of the Wnt5 family occurred during tunicate
316 evolution, but was lost in some lineages.

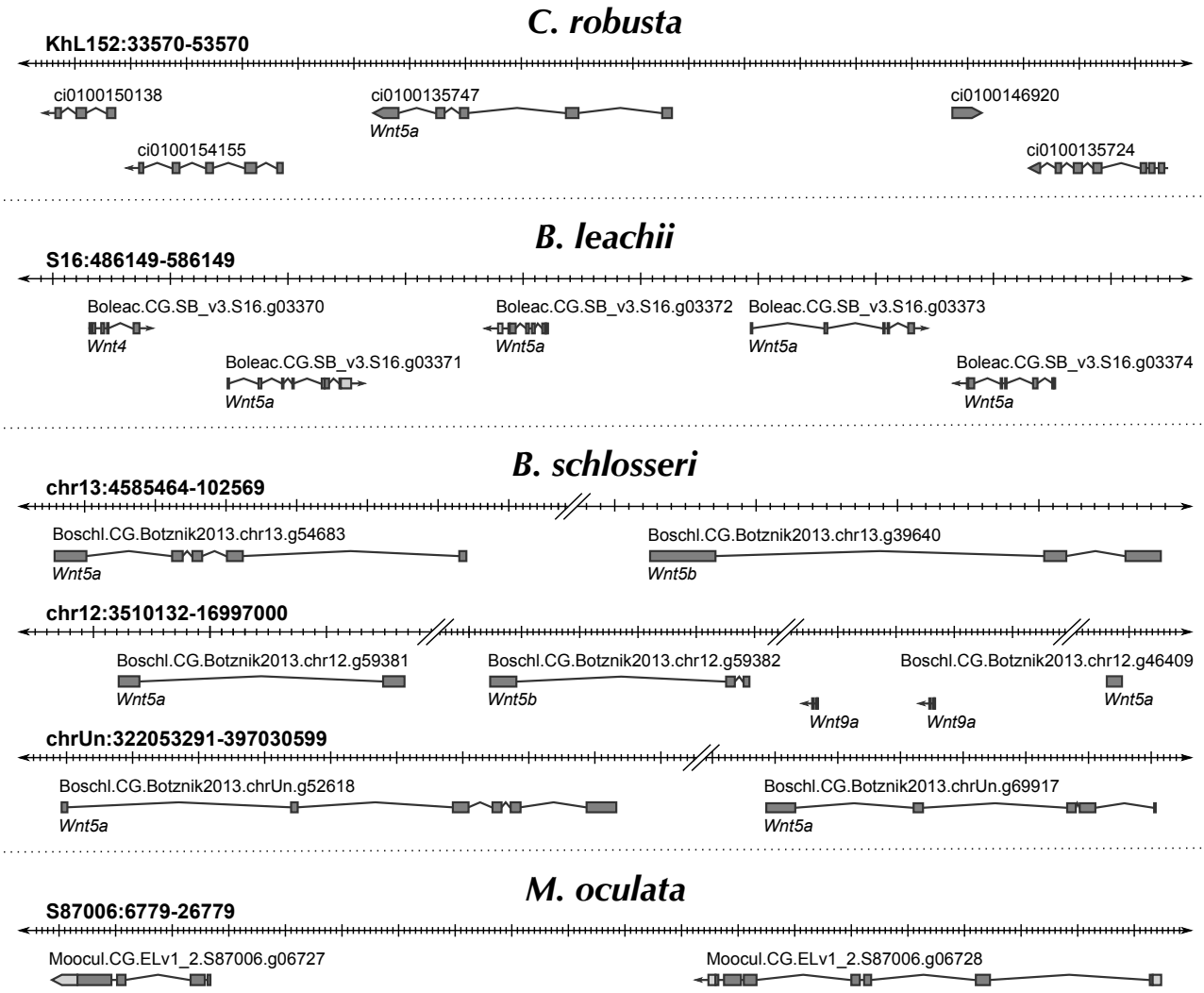


Figure 6. Duplication of *Wnt5a* genes in tunicate genomes.
Schematic showing the genomic location of *Wnt5*-like genes within each indicated genome. Note that no *Wnt5a* ortholog is present in the *O. dioica* genome. Double-parallel lines indicate > 1Mb distance between genes.

317

318 To assess the functionality of the Wnt pathway in Tunicates, we set out to assess
319 whether its downstream effectors are themselves present in the available genomic data. The
320 downstream pathways activated by Wnt ligands are divided into canonical, non-canonical
321 calcium and non-canonical planar cell polarity. The Wnt5a ligand is associated with both of

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the non-canonical pathways through binding of membrane receptors that include *frizzled* (*Fzd4*), *receptor tyrosine kinase-like orphan receptor 1/2* (*Ror1/2*) and *atypical tyrosine kinase receptor* (*Ryk*). Further downstream, disheveled (dsh), β-catenin (Cnntb), Axin, low-density lipoprotein receptor-related protein 5/6 (LRP5/6) and nuclear factor of activated T-cells (NFAT) are proteins essential for triggering intracellular responses to Wnt signalling (MacDonald et al. 2009). We identified orthologs for each of these signalling transduction molecules in all Tunicata genomes (Table S2), with no evidence of further gene duplication events. This supports the interpretation that signalling through the Wnt pathway is functional in tunicates.

Notch pathway

Notch receptors are transmembrane proteins that are involved in cell-cell signalling during development, morphogenesis and regeneration (Hamada et al. 2015). Following activation through the binding of the delta or jagged/serrate ligands, the intracellular domain of Notch is cleaved and induces the expression of downstream target genes including the *hes (hairy and enhancer of split*) gene family members (Guruharsha et al. 2012). The presence of both Notch and the Delta/Serrate/lag-2 (DSL) proteins in most metazoan genomes suggests that their last common ancestor had a single copy of each gene (Gazave et al. 2009). To establish how this pathway has evolved in tunicates, we screened these genomes for the Notch receptor using the conserved lin-Notch repeat (LNR) domain, and for genes encoding probable Notch ligands such as genes from the DSL family .

In all examined genomes, only a single *Notch* receptor gene was identified while the number of ligand genes varied (Table S3). The *C. robusta* genome contains two *DSL* genes, while *O. dioica, M. oculata* and *B. schlosseri* possess only a single *DSL*. By contrast, we found three DSL genes in *B. leachii* (Table S3). To determine the relationships between these

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

identified tunicate DSL-like genes, a phylogeny was constructed along with other chordate DSL proteins. All three *B. leachii* genes are Delta orthologs, two of them related to the *B. schlosseri* and *Cionidae* copy; the third one closer to the *M. oculata* and *H. roretzi* variant. The mouse, human and zebrafish delta and delta-like (DLL) proteins form a discrete clade loosely related to the genes found in Cephalochordates and Tunicates (Fig. 7, shaded box). Jagged proteins form a separate clade (Fig. 7). The tunicate DSL-like proteins present long phylogenetic branches, suggestive of greater diversity, which is also observed in the protein alignment (Fig. S3). This suggests that the tunicate DSL proteins are diverging rapidly from each other, indicative of lineage specific evolution of DSL-like genes.

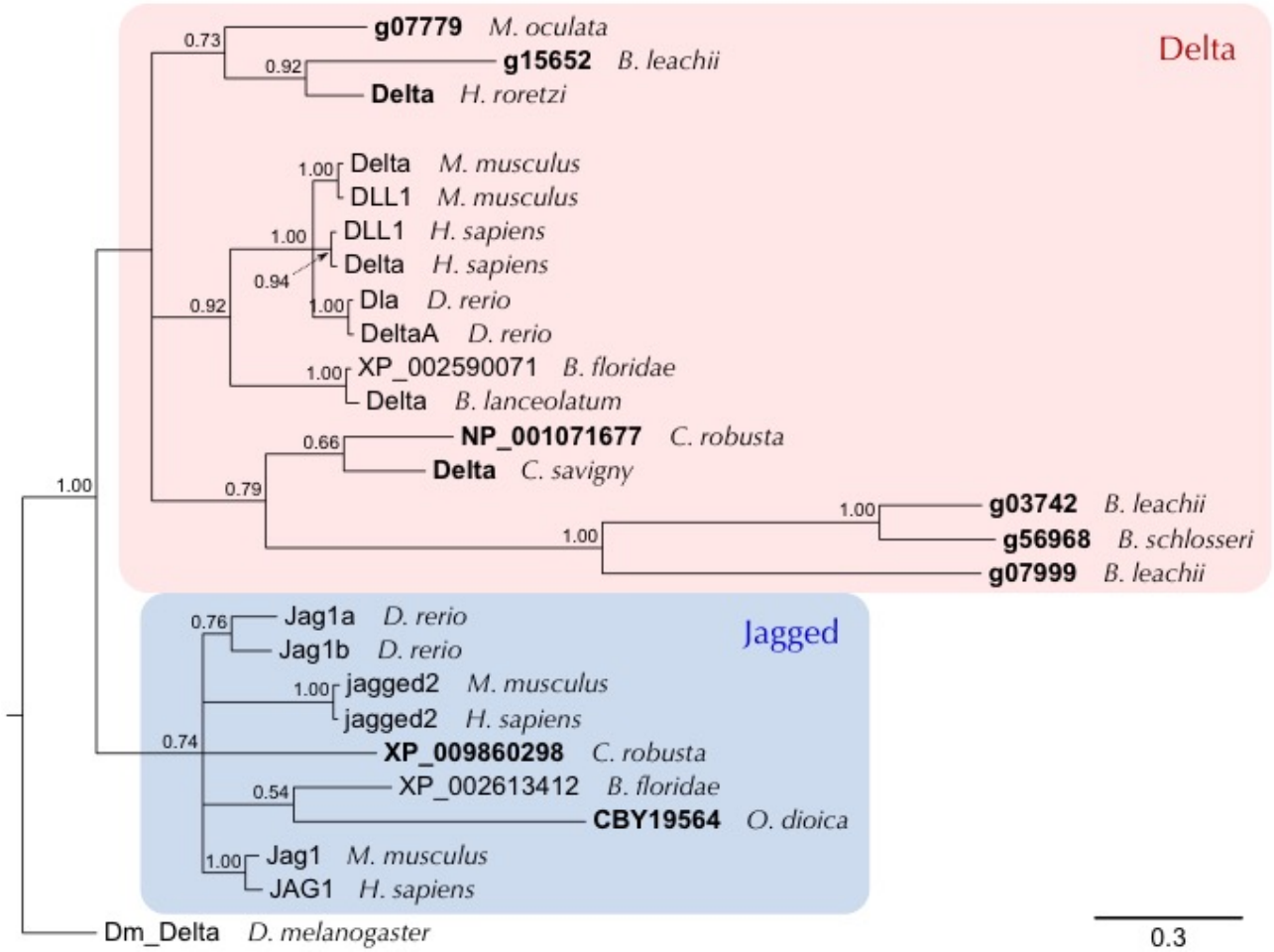


Figure 7. ***B. leachii*** **Notch pathway**
Bayesian phylogenetic tree depicting the relationship between tunicate and vertebrate DSL proteins, using *Drosophila* Delta to root the tree. Tunicate proteins are shown in bold and shaded areas correspond to Delta and Jagged groupings. Branch support values (probabilities) are indicated.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Retinoic acid signalling

Retinoic acid (RA) is an extracellular metabolite that is essential for chordate embryonic development. RA is synthesized from retinol (vitamin A) by two successive oxidation steps. In the first step, retinol dehydrogenase (RDH) transforms retinol into retinal. Then RA is produced by aldehyde dehydrogenase (ALDH), a superfamily of enzymes with essential roles in detoxification and metabolism (Jackson et al. 2011). RA influences the expression of downstream target genes by binding to the RA receptors, RAR and RXR (Fig. 8A (Cunningham and Duester 2015)). Finally, RA is metabolized by the cytochrome P450 family 26 (Cyp26) enzyme, which absence of expression can restrict RA-induced responses to specific tissues or cell types (Ross and Zolfaghari 2011). Components of this pathway have been found in non-chordate animals, suggesting a more ancient origin (Canestro et al. 2006). This pathway has previously been shown to be required for *B. leachii* WBR and *Ciona* development, yet several genes required for RA signalling appear to be missing in *O. dioica* (Martí-Solans et al. 2016).

Rdh10 is the major dehydrogenase associated with the first steps of RA production, although the Rdh16 and RdhE2 enzymes can also substitute this function (Belyaeva et al. 2009; Lee et al. 2009; Belyaeva et al. 2015). The *O. dioica* genome has no orthologs for either *Rdh10* or *Rdh16* but it does have four genes that encode for RdhE2 proteins (Martí-Solans et al. 2016). *O. dioica* also lacks both an *Aldh1*-type gene as well as a *Cyp26* gene but has a single RXR-ortholog (Table S4, (Martí-Solans et al. 2016)). In contrast, the *C. robusta* genome, contains single copies of *Rdh10*, *Rdh16* and *RdhE2* genes and a total of four *Aldh1* genes, located on two chromosomes (Canestro et al. 2006). Consistent with *C. robusta*, *M. oculata*, *B.*

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

381 *leachii* and *B. schlosseri* genomes all have single copies of *Rdh10, Rdh16* and *RdhE2* genes, as
382 well as three *Aldh1a/b* genes on separate scaffolds (Table S4).

383 Three retinoic acid receptor genes were identified within the *B. leachii* genome, one of
384 which had been cloned previously (*g03013,* (Rinkevich et al. 2007b) All three were also found
385 in *C. robusta*, *M. oculata* and *B. schlosseri* genomes (Table S4). While there is only one potential
386 *Cyp26* gene in *M. oculata*, four paralogs were identified in *B. leachii* and *B. schlosseri*. A
387 phylogenetic analysis showed that these 4 genes group with CYP26 proteins (Fig. 8B, Table
388 S4). Altogether, these results show a loss of key RA-pathway genes in *O. dioica* (*Rdh10, Rdh16,*
389 *Cyp26* and *Aldh1a*), while copy numbers in other tunicate genomes increase.
390

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

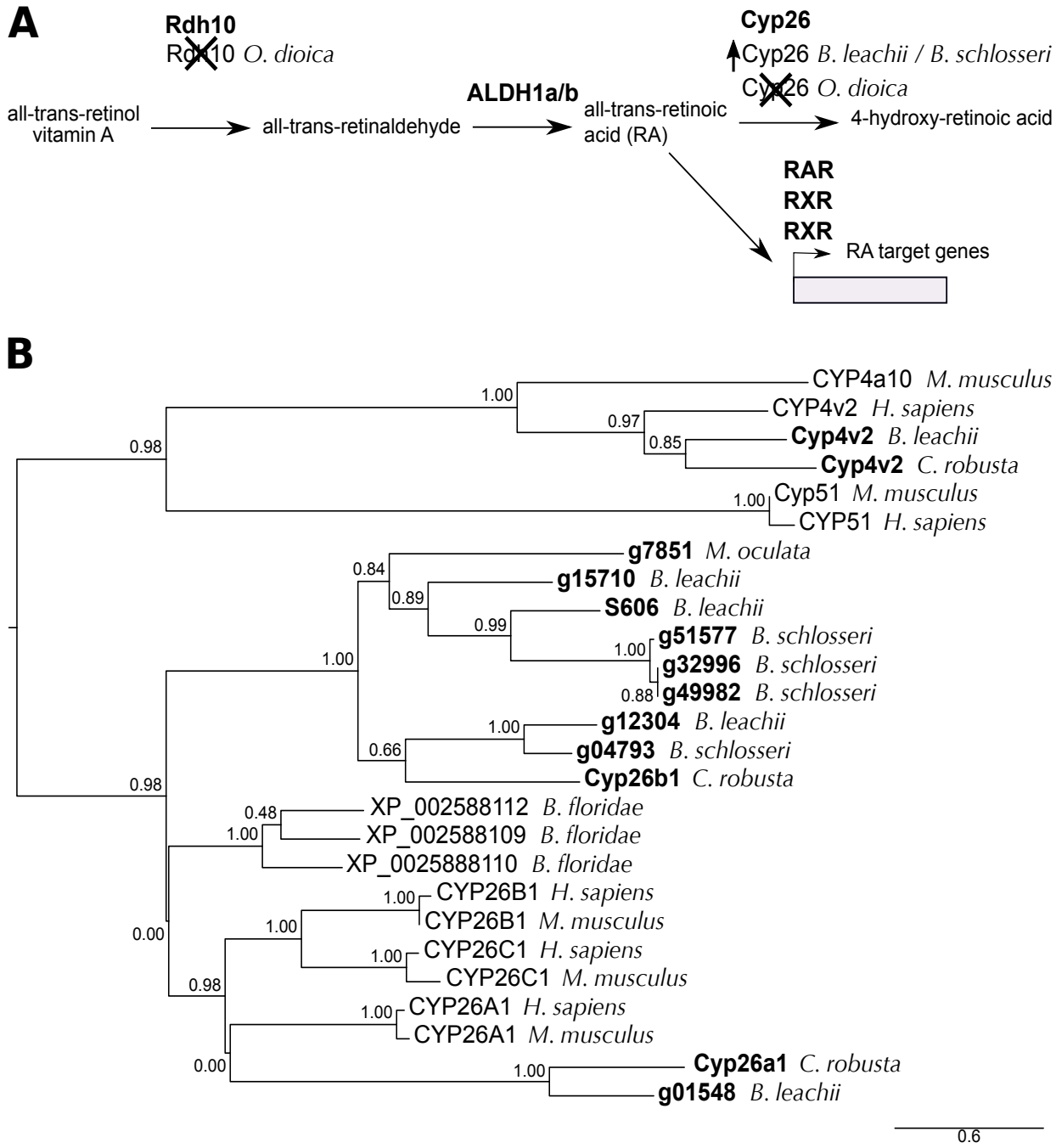


Figure 8. Evolution of the RA pathway in tunicates

(A) Overview of the RA synthesis and degradation pathway. In bold are the major proteins that contribute to RA signalling during animal development. Indicated below these are changes to the number of copies present in examined genomes. **(B)** ML phylogenetic tree depicting the relationship between invertebrate and vertebrate CYP26 proteins using CYP4 and CYP51 proteins as an out-group. Tunicate proteins are shown in bold. No *Cyp26* gene has been identified in the *O. dioica* genome. Values for the approximate likelihood-ratio test (aLRT) are indicated.

391

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

Genomic diversity within the Stolidobranchia

The *B. leachii* genome, along with previous genomic analyses of other ascidian species, support the widely held view that ascidian genomes are diverse and rapidly evolving, which is particularly evident in the Stolidobranchia group (Seo et al. 2001; Dehal et al. 2002b; Voskoboynik et al. 2013a; Stolfi et al. 2014; Tsagkogeorga et al. 2010; Bock et al. 2012; Tsagkogeorga et al. 2012; Rubinstein et al. 2013; Griggio et al. 2014). Nevertheless, botryllids are sufficiently similar in external appearance and morphology for early researchers to have suggested that *Botrylloides* could be a subgenus of *Botryllus* (Saito et al. 2001; Nydam et al. 2017). Strikingly however, the *B. schlosseri* genome differs from that of *B. leachii*, as well as from other sequenced tunicate genomes (Table 2). Particularly striking is the comparison between the *B. leachii* and *B. schlosseri*, where differences in genome sizes (194 Mb vs 725 Mb), the fraction of repetitive sequences (18 % vs 60 %; 65 % in (Voskoboynik et al. 2013a)) and the predicted gene number (15,839 vs 27,463; (Voskoboynik et al. 2013a)) suggest divergent genome architectures. Altogether, these comparisons indicate that the *B. schlosseri* genome has undergone a significant increase in its genomic content, including retrotransposon expansion (Table S1). In particular, there are at least two additional families in the *B. schlosseri* hAT transposon superfamily and counts of common hAT elements, such as hAT-Charlie, can differ dramatically (e.g. hAT-Charlie 366 in *B. leachii* vs 46,661 in *B. schlosseri*). DNA methylation is a key suppressor of transposon activity, changes to the methylation of transposable elements is a known driver of increased transposition (O'Neill et al. 1998; Maumus and Quesneville 2014; Simmen et al. 1999; Suzuki et al. 2007). DNA methylation in tunicate species has only been studied in *C. robusta,* and is described as mosaic, gene body methylation, whereas non-coding regions including transposons remain

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

417 unmethylated (Suzuki et al. 2007), it is unknown how retrotransposons are suppressed in
418 tunicate genomes. Nevertheless, the observed increase in transpositions could be a
419 consequence of low non-coding DNA methylation, which may contribute to the rapid genome
420 evolution observed in tunicate species, even between closely related species such as *B.*
421 *schlosseri* and *B. leachii*.

422 Rapid genome evolution, and active transposable elements in particular, are proposed
423 to aid adaptation to new environments for invasive species (Stapley et al. 2015). Differences
424 in the colonization ability of tunicates has been noted, not only between related species such
425 as *B. leachii* and *B. schlosseri* (Brunetti 1976; Brunetti et al. 1980; Brunetti 1974), but even at
426 the molecular level within *B. schlosseri* populations (Bock et al. 2012; Nydam et al. 2017). It is
427 thus possible that the observed success in tunicate invasion (Zhan et al. 2015) is supported by
428 their plasticity in genome characteristics like transposon diversity and gene number.

429 Ancient homeobox gene clusters whose structure has been retained over millions of
430 years of evolution in many organisms are fragmented in tunicate genomes. Because, the
431 expression of each *Hox* gene across the anterior-posterior axis relates to their location within
432 the *Hox* gene cluster (Pascual-Anaya et al. 2013), cluster breaks are predicted to have
433 consequences for patterning processes. However, an adult body plan with correct spatial
434 orientation of its body axes during tissue development in ascidians also needs to be
435 established during sexual, asexual and WBR. Early patterning events in tunicate species have
436 only been characterized during sexual reproduction in *Ciona*. Early stages of development
437 (prior to gastrulation) follow a mosaic pattern of developmental axis formation, where
438 inheritance of maternally provided factors establishes the body axes (Nishida 2005). *Hox* gene
439 knockdown experiments in *C. robusta* revealed that they have very limited roles, with defects
440 only observed in larval neuronal and tail development upon loss of *Ci-Hox11* and *Ci-Hox12*
441 function (Ikuta et al. 2010). It thus appears that patterning events in *C. robusta* are less

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

dependent upon anterior-posterior spatial expression of *Hox* genes to establish regional identity. Previously in *B. schlosseri*, the entry point of the connective test vessel into the developing bud determines the posterior end of the new zooid (Sabbadin et al. 1975). Therefore it is possible that ascidians incorporate environmental and physical cues to compensate for the lost gene cluster during polarity establishment. A wider analysis comprising multiple tunicate species will be necessary to investigate the exact consequences of homeobox cluster dispersion and whether the compensatory mechanism observed in *C. robusta* is the norm or an exception.

Gene orthology analysis and coloniality candidate pathways

Among the tunicate orthologous clusters that we obtained, we identified several groups of genes that are not shared by all the tunicate genomes (Fig. 2A). Given the rapid genomic evolution of these organisms, it is more likely that these genes have either been lost or that their sequence has highly diverged, rather than independent gains of novel genes.

Of particular interest are the genes found only in the *B. schlosseri* and *B. leachii* genomes, as these may function in biological processes unique to colonial tunicates. Many of these genes have orthologs not only in vertebrates, but also in more evolutionarily distant animals such as *C. elegans* (File S4). This suggests that these genes have a more ancient origin, which was retained specifically in Botryllidae genomes. The overrepresented genes (File S4) have annotated functions including circulation (GO:0003018, GO:0003013, GO:0050880), wound healing (GO:0072378) and cell communication (GO:0007154); as well as regulation of immune cell differentiation (GO:0033081, GO:0033089), immune system process (GO:0002376) and interferons (GO:0032608). Unlike solitary tunicate species, colonial ascidians possess a complex system of single cell-lined vessels used to transport haemocytes and facilitate communication between zooids within the colony (Mukai et al. 1978). In

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

addition, immune response is known to have roles roles in wound healing, vasculogenesis, allorecognition and regeneration (Voskoboynik et al. 2013b; Taketa et al. 2015; Gutierrez and Brown 2017; Sattler 2017). Therefore, it is possible that these genes, found only in *Botryllus* and *Botrylloides*, contribute to biological pathways and cellular processes that have important roles in colonialism.

Both *O. dioica* and *B. schlosseri* had a high number (2160 and 2716 respectively) of clusters unique to their genomes (Fig. 2A). While the *O. dioica* genome has undergone considerable loss of ancestral genes (Albalat and Cañestro 2016; Seo et al. 2001), the total number of genes in this specie is similar to that of other tunicates (Table 2). Taken together, these observations suggest that there has been a duplication of the retained genes such as *Otx* (3 copies in *O. dioica*, one in *Ciona*(Cañestro et al. 2005)), potentially involving roles in their peculiar neotenic and dioecious life cycle. The *B. schlosseri* genome has an ~10,000 higher predicted gene number compared to other tunicates (Table 2). Such massive increase in numbers suggests partial genome duplication. Further analysis will be required to determine whether these are novel or duplicated genes, hence providing important insights in the evolution of Tunicata genomes.

Lineage-specific changes to evolutionarily conserved cell communication pathways

Cell signalling pathways are critical for morphogenesis, development and adult physiology. In particular, we have focused our analysis on three highly conserved pathways: Wnt, Notch and Retinoic Acid signalling. Representatives of all twelve *Wnt* gene subfamilies are found in metazoans, suggesting that they evolved before evolution of the bilaterians (Janssen et al. 2010). We identified members of each Wnt subfamily in tunicate genomes, along with numerous examples of lineage-specific gene loss and/or duplication. The most striking of these events was an increase in *Wnt5a* gene copy number in *B. leachii*, *B. schlosseri*

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and *M. oculata* genomes. Indeed, most invertebrates genomes, including the basal chordate *B. floridae*, contain a single *Wnt5* gene while most vertebrate genomes have two *Wnt5a* paralogs, believed to be a result of whole genome duplication (Martin et al. 2012). However, in the analysed tunicate genomes, up to 15 copies of this gene were identified, potentially these additional genes may have been co-opted into novel roles and were retained during tunicate evolution. Wnt5a ligands have numerous biological roles, including a suppressive one during zebrafish regeneration (Stoick-Cooper et al. 2007) and a promotive one during amphioxus regeneration (Somorjai et al. 2012). Furthermore, components of both Wnt signalling pathways are differentially expressed during WBR (Zondag et al. 2016). It is thus conceivable that *Wnt5a* gene number has expanded in colonial tunicates to sustain WBR. A functional characterization of the role of these numerous copies of Wnt5a would thus be highly interesting and potentially reveal evolutionary insights into chordate regeneration.

All components of the Notch pathway are present in the genomes we investigated. Of particular interest, the DSL Notch ligand appears to be rapidly evolving in the tunicates. This indicates that tunicate DSL proteins are under less pressure, than vertebrate orthologous proteins, to conserve their sequences. Given that the interaction between the DSL domain and the Notch receptor is core to signaling pathway activation (Chillakuri et al. 2012), it will be interesting to assess whether the functional ligand-receptor interactions between tunicate DSL proteins and tunicate Notch proteins have adapted accordingly.

Components of the RA signalling pathway have also been identified in all the tunicate genomes. However, *Oikopleura* has seemingly lost a functional RA synthesis pathway, while still forming a functional body plan. This suggests that either uniquely RA is not involved in critical developmental events in this species, that the RA signalling function has been replaced or that *O. dioica* utilizes an alternative synthesis approach. Conversely, lineage specific

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

516 increases in RA pathway gene numbers have been observed in *C. robusta* (Aldh1, (Sobreira et
517 al. 2011)) and *Botrylloides* (*CYP26* genes, Fig. 8).

518 RA, Notch and Wnt pathways play roles in regeneration and development in many
519 species, including Stolidobranchian tunicates (Rinkevich et al. 2007b, 2008; Zondag et al.
520 2016) and *Cionidae* (Hamada et al. 2015; Jeffery 2015). The observed loss of RA signalling
521 genes may result in reduced regeneration ability for *O. dioica*, however it's regenerative
522 abilities have not been characterized. Given the unique chordate WBR potential developed by
523 colonial tunicates, it is conceivable that there is selective pressure on their genomes to retain
524 these pathways. We thus predict that these pathways play a similar role in colony reactivation
525 following hibernation.

526 Among tunicates there exist significant differences in life cycle, reproduction and
527 regeneration ability, even between closely related species of the same family, which likely
528 reflect an underlying diversity in genomic content. For instance, differences in both asexual
529 and sexual reproduction have been observed within the Botryllidae family (Oka and
530 Watanabe 1957; Brunetti 1974, 1976, Berrill 1951, 1947, 1941). Furthermore, *B. schlosseri*
531 can only undergo WBR during a short time frame of their asexual reproductive cycle when the
532 adults are reabsorbed by the colony (Voskoboynik et al. 2007; Kürn et al. 2011) while *B.*
533 *leachii* can undergo WBR throughout their adult life (Rinkevich et al. 2007b). Overall, this
534 indicates that despite a generally similar appearance, the rapid evolution of the Tunicata
535 subphylum has provided diversity and innovations within its species. It will be interesting to
536 investigate how such genomic plasticity balances between adaptation to new challenges and
537 constraint, preserving common morphological features, in future studies.

538 In conclusion, our assembly of the *B. leachii* genome provides an essential resource for
539 the study of this colonial ascidian as well as a crucial point of comparison to gain further
540 insights into the remarkable genetic diversity among tunicate species. In addition, the genome

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of *B. leachii* will be most useful for dissecting WBR in chordates; particularly through comparison with *B. schlosseri* for understanding how the initiation of WBR can be blocked during specific periods of their life cycle. Furthermore, given the key phylogenetic position of Tunicates with respect to vertebrates, the analysis of their genomes will provide important insights in the emergence of chordate traits and the origin of vertebrates.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Sampling, library preparation and sequencing

B. leachii colonies were collected from Nelson harbour (latitude 41.26°S, longitude 173.28°E) in New Zealand. To reduce the likelihood of contamination, embryos were dissected out of a colony and pooled before carrying out DNA extraction using E.Z.N.A SP Plant DNA Mini Kit. A total of 2 µg each was sent to New Zealand Genomics Limited (NZGL) for library preparation and sequencing. Short read sequencing of Illumina TruSeq libraries in a HiSeq2500 generated 19,090,212 paired-end reads of 100 bp (average fragment size: 450 bp, adaptor length: 120 bp). A second sequencing (Illumina Nextera MiSeq Mate Pair) not size-selected generated 31,780,788 paired-end sequences of 250 bp (fragment size: 1.5 - 15 kb, median size: ~3 kb, adaptor length: 38 bp).

PreQC report was generated using the String Graph Assembler software package (Simpson 2014) and quality metrics before assembly with both FastQC (Andrews 2010) as well as MultiQC (Ewels et al. 2016) (Fig. S1). These analyses revealed that 91 % of sequences had a mean Phred quality score >= 30, 96 % of bases a mean Phred quality score >= 30, and 39 % of sequences an adapter sequence (either Illumina or Nextera). Adaptor trimming was performed with NxTrim (O'Connell et al. 2015) for the mate pair library, followed by Trimmomatic (Bolger et al. 2014) with the following options: MINLEN:40 ILLUMINACLIP:2:30:12:1:true LEADING:3 TRAILING:3 MAXINFO:40:0.4 MINLEN:40 for both libraries. After trimming, 86,644,308 paired-end (85 %) and 12,112,004 (12 %) single-end sequences remained (100 % with a mean Phred quality score >= 30, < 1 % with an adapter sequence).

Genome assembly

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

571 *De novo* assembly was performed in three consecutive iterations following a Meta-
572 assembly approach (Table S5). First, both libraries were assembled together in parallel, using
573 a k-mer size of 63 following the results from KmerGenie (Chikhi and Medvedev 2014)
574 whenever available, by five assemblers: AbySS (Simpson et al. 2009), Velvet (Zerbino and
575 Birney 2008), SOAPdenovo2 (Luo et al. 2012), ALLPATHS-LG (Gnerre et al. 2011), MaSuRCA
576 (Zimin et al. 2013). The MaSuRCA assembler was run twice, once running the adapter filtering
577 function (here termed "MaSuRCA-filtered"), the other without (termed simply "MaSuRCA").
578 Their respective quality was then estimated using three different metrics: the N50 length, the
579 BUSCO core-genes completion (Simão et al. 2015) and the Glimmer number of predicted
580 genes (Delcher et al. 1999). Second, these drafts were combined by following each ranking
581 using Metassembler (Wences and Schatz 2015), hence producing three new assemblies
582 (limiting the maximum insert size at 15 kb). Third, the *B. leachii* transcriptome (Zondag et al.
583 2016) was aligned to each meta-assembly using STAR (Dobin et al. 2013), which were then
584 combined thrice more using Metassembler following their alignment percentage and limiting
585 the maximum insert size at 3 kb, 8 kb and 15 kb. Finally, the quality of the meta-meta-
586 assemblies was estimated using the BUSCO score and the best one (Table S5) selected as the
587 reference *de novo* assembly.
588
589 **Data access**
590 All data was retrieved from the indicated sources in January 2016. Note that *Ciona*
591 *intestinalis* type A (Dehal et al. 2002b) has recently been recognized as a distinct species
592 (*Ciona robusta,* (Brunetti et al. 2015)) and that this study has been undertaken before it was
593 renamed.
594 *B. leachii*, *B. schlosseri*, *C. robusta*, *M. oculata*: Ascidian Network for *In Situ* Expression and
595 Embryonic Data (ANISEED, https://www.aniseed.cnrs.fr/aniseed/, (Tassy et al. 2010)).

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

596 *O. dioica*: Oikopleura Genome Browser
597 (http://www.genoscope.cns.fr/externe/GenomeBrowser/Oikopleura/, (Seo et al. 2001)).
598 *B. floridae*, *H. sapiens*: Joint Genome Institute (JGI, http://genome.jgi.doe.gov, (Grigoriev et al.
599 2012))
600
601 ***Repeat region analysis***
602 A *de novo* repeat library was build for each tunicate genome using RepeatModeler
603 (Smit and Hubley 2015). This utilizes the RECON tandem repeats finder from the RepeatScout
604 packages to identify species-specific repeats in a genome assembly. RepeatMasker (Smit et al.
605 2015) was then used to mask those repeats.
606
607 **Gene annotation**
608 *Ab initio* genome annotation was performed using MAKER2 (Holt and Yandell 2011)
609 with Augustus (Stanke and Waack 2003) and SNAP (Korf 2004) for gene prediction. In
610 addition, we used our previously published transcriptome (Zondag et al. 2016) and a
611 concatenation of UniProtKB (UniProt Consortium 2015), *C. robusta* and *B. schlosseri* proteins
612 into a custom proteome as evidence of gene product. Using the predicted genes, Augustus and
613 SNAP were then trained to the specificity of *B. leachii* genome. A second round of predictions
614 was then performed, followed by a second round of training. The final annotation of the
615 genome was obtained after running a third round of predictions, and the provided trained
616 Augustus and SNAP configurations after a third round of training. Non-coding RNA sequences
617 were then annotated using Infernal (Nawrocki and Eddy 2013) with Rfam library 12.0
618 (Nawrocki et al. 2015), tRNAscan-SE (Lowe and Eddy 1997) and snoRNA (Lowe 1999).
619 Finally, the identified sequences were characterized by InterProScan (Jones et al. 2014).
620

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

621 **Analysis of Gene Ontology terms**

622 Distribution of Gene Ontology (GO) terms were computed for each species as follows.
623 GO terms were extracted from the genome annotation and the number of occurrence for each
624 term determined using a custom Python script. The resulting list of frequencies was then
625 simplified using REVIGO (similarity factor "small" of 0.5, (Supek et al. 2011)) and the
626 `TreeMap` output retrieved. The hierarchy of every GO term present was reconstructed
627 following the schema defined by the core gene ontology (`go.obo`, (The Gene Ontology
628 Consortium 2015)) using a custom Python script selecting the shortest path to the root of the
629 tree, favouring smaller GO terms identification number in case of multiple paths. Finally,
630 frequencies were displayed using the `sunburstR` function of the Data-Driven Documents
631 library (D3, (Bostock et al. 2011)).

632 Predicted amino-acid sequences for all species were retrieved and clustered into
633 17,710 groups by OrthoMCL (Li et al. 2003). Protein sequences within each group were then
634 aligned into a Multiple Sequence Alignment (MSA) by Clustal-Omega, and the corresponding
635 consensus sequence inferred by EMBOSS cons. Consensus sequences were matched to the
636 Swiss-Prot curated database using BLASTp (e-value cut-off of 10^{-5}), and the GO terms
637 corresponding to the best match retrieved. GO terms frequencies were analysed as described
638 above and displayed using REVIGO's treemap. The overrepresentation analysis was
639 performed using GOrilla (Eden et al. 2009) with *Homo sapiens* as the organism background,
640 using a *p*-value threshold of 10^{-3} and REVIGO treemap (similarity factor "medium" of 0.7) for
641 visualization.

642

643 **Analysis of specific gene families**

644 Genes and transcripts for each examined genome were identified by a tBLASTn search
645 with an e-value cut-off at 10^{-5} using the SequencerServer software (Priyam et al. 2015). This

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

was followed by a reciprocal BLAST using SmartBLAST (NCBI Resource Coordinators 2016), to confirm their identity.

Delta serrate ligand conserved protein domain (PF01414) was used to identify the corresponding proteins in tunicate genomes. To identify *Notch* receptor genes the conserved LNR (lin-notch repeat) domain (PF00066) was used. ALDH-like genes were identified by tBLASTn search (PF00171) and classified using SMART blast.

Phylogenetics

Sequences were aligned with ClustalX (Jeanmougin et al. 1998) before using ProtTest 3 (Abascal et al. 2005) to determine the best-fit model of evolution. The best-fit model for the DSL phylogeny was WAG+I+G and, for CYP26 proteins, was LG+I+G.

Bayesian inference (BI) phylogenies were constructed using MrBayes (Ronquist and Huelsenbeck 2003) with a mixed model for 100,000 generations and summarized using a Sump burn-in of 200. Maximum Likelihood (ML) phylogenies were generated by PhyML (Guindon et al. 2010), using the estimated amino acid frequencies.

Accession numbers are provided in File S3 and sequence alignments are provided in Figure S3. Analyses carried out with BI and ML produced identical tree topologies.

Trees were displayed using FigTree v1.4.2 (Rambaut 2016).

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgements

Funding support was provided to M.J.W. by the Otago BMS Deans Bequest and Department of Anatomy. S.B. was supported by the Swiss National Science Foundation (SNSF) grant number P2ELP3_158873. We would like to thank Peter Maxwell and the New Zealand eScience Infrastructure (NeSI); Christelle Dantec and ANISEED for help and advice during the annotation process, as well as for the accompanying *B. leachii* genome browser.

Supplementary Figures

Fig. S1. SGA analysis. Including genome size estimation and de Brujin graph quantification.

Fig. S2. Gene Ontology terms identified in the larger orthologs clusters between the tunicate genomes.

Fig. S3. Protein sequence alignments used to generate the Notch and RA phylogenies.

Supplementary Files

File S1. BUSCO scores for the *B. leachii* genome assembly

File S2. Results of Repeatmasker analysis using de novo repeat libraries (Repeatmodeler)

File S3. Results from OrthoMCL group including the REVIGO GO term listed each orthologue group, used in the comparison of the GO terms between the tunicate genomes.

File S4. BLASTp, GOrillia and REVIGO results used used for overrepresentation analysis of GO terms present in the *B. leachii* and *B. schlosseri* only orthologue group analysis.

File S5. Corresponding Gene and Transcript IDs for *B. leachii* genes of interest. Accession numbers for protein sequences used in phylogeny construction.

Supplementary Tables

Table S1. Repetitive elements identified in the *B. leachii* and *B. schlosseri* genomes using Repeatmoduler and RepeatMasker.

Table S2. Comparison of the number of Wnt pathway genes.

Table S3. Comparison of the number of Notch pathway genes.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

705 **Table S4.** RA pathway components across annotated tunicate genomes.
706
707 **Table S5.** Iterative results of the meta-assembly approach followed for the *de novo* assembly
708 of the *B. leachii* genome
709

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**: 2104–2105. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti263.

Albalat R, Cañestro C. 2016. Evolution by gene loss. *Nat Rev Genet* **17**: 379–391. http://www.nature.com/doifinder/10.1038/nrg.2016.39.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Auger H, Sasakura Y, Joly J-S, Jeffery WR. 2010. Regeneration of oral siphon pigment organs in the ascidian Ciona intestinalis. *Dev Biol* **339**: 374–389. http://linkinghub.elsevier.com/retrieve/pii/S0012160609014651.

Ballarin L, Franchini A, Ottaviani E, Sabbadin A. 2001. Morula cells as the major immunomodulatory hemocytes in ascidians: Evidences from the colonial species *Botryllus schlosseri*. *Biol Bull* **201**: 59–64.

Belyaeva O V., Chang C, Berlett MC, Kedishvili NY. 2015. Evolutionary origins of retinoid active short-chain dehydrogenases/reductases of SDR16C family. *Chem Biol Interact* **234**: 135–143. http://linkinghub.elsevier.com/retrieve/pii/S000927971400324X.

Belyaeva O V., Lee S-A, Kolupaev O V., Kedishvili NY. 2009. Identification and characterization of retinoid-active short-chain dehydrogenases/reductases in Drosophila melanogaster. *Biochim Biophys Acta - Gen Subj* **1790**: 1266–1273. http://linkinghub.elsevier.com/retrieve/pii/S0304416509001688.

Berna L, Alvarez-Valin F. 2014. Evolutionary genomics of fast evolving tunicates. *Genome Biol Evol* **6**: 1724–1738. https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evu122.

Berrill NJ. 1951. Regeneration and Budding in Tunicates. *Biol Rev* **26**: 456–475. http://doi.wiley.com/10.1111/j.1469-185X.1951.tb01207.x.

Berrill NJ. 1947. The developmental cycle of *Botrylloides*. *Q J Microsc Sci* **88**: 393–407.

Berrill NJ. 1941. The development of the bud in *Botryllus*. *Biol Bull* **80**: 169. http://www.jstor.org/stable/1537595?origin=crossref.

Bock DG, MacIsaac HJ, Cristescu ME. 2012. Multilocus genetic analyses differentiate between widespread and spatially restricted cryptic species in a model ascidian. *Proc R Soc B Biol Sci* **279**: 2377–2385. http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2011.2610.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120. http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btu170.

Bostock M, Ogievetsky V, Heer J. 2011. D3: Data-Driven Documents. *IEEE Trans Vis Comput Graph*. http://vis.stanford.edu/papers/d3.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, et al. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**: 10. http://www.gigasciencejournal.com/content/2/1/10.

Brown FD, Swalla BJ. 2012. Evolution and development of budding by stem cells: Ascidian coloniality as a case study. *Dev Biol* **369**: 151–162. http://dx.doi.org/10.1016/j.ydbio.2012.05.038.

Brozovic M, Martin C, Dantec C, Dauga D, Mendez M, Simion P, Percher M, Laporte B, Scornavacca C, Di Gregorio A, et al. 2016. ANISEED 2015: A digital framework for the comparative developmental biology of ascidians. *Nucleic Acids Res* **44**: D808–D818.

Brunetti R. 1976. Biological cycle of *Botrylloides leachi* (Savigny) (Ascidiacea) in the Venetian lagoon. *Vie Milieu* **XXVI**: 105–122.

Brunetti R. 1974. *Observations on the Life Cycle of Botryllus Schlosseri (Pallas) (Ascidiacea) in*

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the Venetian Lagoon. http://www.tandfonline.com/doi/abs/10.1080/11250007409430119.

Brunetti R, Beghi L, Bressan M, Marin M. 1980. Combined effects of temperature and salinity on colonies of *Botryllus schlosseri* and *Botrylloides leachi* (Ascidiacea) from the Venetian Lagoon. *Mar Ecol Prog Ser* **2**: 303–314. http://www.int-res.com/articles/meps/2/m002p303.pdf.

Brunetti R, Gissi C, Pennati R, Caicci F, Gasparini F, Manni L. 2015. Morphological evidence that the molecularly determined Ciona intestinalis type A and type B are different species: Ciona robusta and Ciona intestinalis. *J Zool Syst Evol Res* **53**: 186–193. http://doi.wiley.com/10.1111/jzs.12101.

Burighel P, Brunetti R, Zaniolo G. 1976. Hibernation of the colonial ascidian *Botrylloides leachi* (Savigny): histological observations. *Ital J Zool* **43**: 293–301.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421. http://www.ncbi.nlm.nih.gov/pubmed/20003500.

Cañestro C, Bassham S, Postlethwait J. 2005. Development of the central nervous system in the larvacean Oikopleura dioica and the evolution of the chordate brain. *Dev Biol* **285**: 298–315. http://linkinghub.elsevier.com/retrieve/pii/S0012160605004410.

Canestro C, Postlethwait JH, Gonzalez-Duarte R, Albalat R. 2006. Is retinoic acid genetic machinery a chordate innovation? *Evol Dev* **8**: 394–406. http://doi.wiley.com/10.1111/j.1525-142X.2006.00113.x.

Chikhi R, Medvedev P. 2014. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**: 31–37. http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btt310.

Chillakuri CR, Sheppard D, Lea SM, Handford PA. 2012. Notch receptor–ligand binding and activation: Insights from molecular studies. *Semin Cell Dev Biol* **23**: 421–428. http://linkinghub.elsevier.com/retrieve/pii/S1084952112000134.

Cunningham TJ, Duester G. 2015. Mechanisms of retinoic acid signalling and its roles in organ and limb development. *Nat Rev Mol Cell Biol* **16**: 110–123. http://www.nature.com/doifinder/10.1038/nrm3932.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, et al. 2002a. The Draft Genome of Ciona intestinalis: Insights into Chordate and Vertebrate Origins. *Science (80-)* **298**: 2157–2167. http://www.sciencemag.org/content/298/5601/2157.abstract.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso AW, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, et al. 2002b. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**: 2157–67. http://www.ncbi.nlm.nih.gov/pubmed/12481130 (Accessed August 25, 2014).

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**: 4636–41. http://www.ncbi.nlm.nih.gov/pubmed/10556321.

Delsuc F, Brinkmann H, Chourrout D, Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**: 965–968. http://www.ncbi.nlm.nih.gov/pubmed/16495997 (Accessed July 10, 2014).

Dobin A, Davis C a, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3530905&tool=pmcentrez&rendertype=abstract.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**: 48. http://www.biomedcentral.com/1471-2105/10/48.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Edvardsen RB, Seo H-C, Jensen MF, Mialon A, Mikhaleva J, Bjordal M, Cartry J, Reinhardt R, Weissenbach J, Wincker P, et al. 2005. Remodelling of the homeobox gene complement in the tunicate Oikopleura dioica. *Curr Biol* **15**: R12–R13. http://linkinghub.elsevier.com/retrieve/pii/S096098220400973X.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* btw354.

Franchi N, Schiavon F, Carletto M, Gasparini F, Bertoloni G, Tosatto SCE, Ballarin L. 2011. Immune roles of a rhamnose-binding lectin in the colonial ascidian Botryllus schlosseri. *Immunobiology* **216**: 725–736. http://linkinghub.elsevier.com/retrieve/pii/S0171298510001993.

Garcia-Fernàndez J. 2005. The genesis and evolution of homeobox gene clusters. *Nat Rev Genet* **6**: 881–92. http://www.ncbi.nlm.nih.gov/pubmed/16341069.

Gasparini F, Burighel P, Manni L, Zaniolo G. 2008. Vascular regeneration and angiogenic-like sprouting mechanism in a compound ascidian is similar to vertebrates. *Evol Dev* **10**: 591–605.

Gazave E, Lapébie P, Richards GS, Brunet F, Ereskovsky A V, Degnan BM, Borchiellini C, Vervoort M, Renard E. 2009. Origin and evolution of the Notch signalling pathway: an overview from eukaryotic genomes. *BMC Evol Biol* **9**: 249. http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-9-249.

Gissi C, Hastings KEM, Gasparini F, Stach T, Pennati R, Manni L. 2017. An unprecedented taxonomic revision of a model organism: the paradigmatic case of Ciona robusta and Ciona intestinalis. *Zool Scr*. http://doi.wiley.com/10.1111/zsc.12233.

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci* **108**: 1513–1518. http://www.pnas.org/cgi/doi/10.1073/pnas.1017351108.

Griggio F, Voskoboynik A, Iannelli F, Justy F, Tilak M-KM-K, Xavier T, Pesole G, Douzery EJP, Mastrototaro F, Gissi C. 2014. Ascidian Mitogenomics: Comparison of Evolutionary Rates in Closely Related Taxa Provides Evidence of Ongoing Speciation Events. *Genome Biol Evol* **6**: 591–605. https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evu041.

Grigoriev I V., Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA, et al. 2012. The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res* **40**: D26–D32. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr947.

Guder C, Philipp I, Lengfeld T, Watanabe H, Hobmayer B, Holstein TW. 2006. The Wnt code: cnidarians signal the way. *Oncogene* **25**: 7450–7460. http://www.nature.com/doifinder/10.1038/sj.onc.1210052.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* **59**: 307–321. https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/syq010.

Guruharsha KG, Kankel MW, Artavanis-Tsakonas S. 2012. The Notch signalling system: recent insights into the complexity of a conserved pathway. *Nat Rev Genet* **13**: 654–666. http://www.nature.com/doifinder/10.1038/nrg3272.

Gutierrez S, Brown FD. 2017. Vascular budding in *Symplegma brakenhielmi* and the evolution of coloniality in styelid ascidians. *Dev Biol* **423**: 152–169. http://dx.doi.org/10.1016/j.ydbio.2017.01.012.

Hamada M, Goricki S, Byerly MS, Satoh N, Jeffery WR. 2015. Evolution of the chordate regeneration blastema: Differential gene expression and conserved role of notch signaling during siphon regeneration in the ascidian Ciona. *Dev Biol* **405**: 304–315. http://linkinghub.elsevier.com/retrieve/pii/S0012160615300701.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Hoegg S, Meyer A. 2005. Hox clusters as models for vertebrate genome evolution. *Trends Genet* **21**: 421–424. http://linkinghub.elsevier.com/retrieve/pii/S0168952505001654.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**: 491.

Ikuta T, Satoh N, Saiga H. 2010. Limited functions of Hox genes in the larval development of the ascidian Ciona intestinalis. *Development* **137**: 1505–1513. http://dev.biologists.org/cgi/doi/10.1242/dev.046938.

Iwasa T, Mishima S, Watari A, Ohkuma M, Azuma T, Kanehara K, Tsuda M. 2003. A novel G protein alpha subunit in embryo of the ascidian, Halocynthia roretzi. *Zoolog Sci* **20**: 141–51. http://www.ncbi.nlm.nih.gov/pubmed/12655177.

Jackson B, Brocker C, Thompson DC, Black W, Vasiliou K, Nebert DW, Vasiliou V. 2011. Update on the aldehyde dehydrogenase gene (ALDH) superfamily. *Hum Genomics* **5**: 283–303. http://www.ncbi.nlm.nih.gov/pubmed/21712190.

Janssen R, Le Gouar M, Pechmann M, Poulin F, Bolognesi R, Schwager EE, Hopfen C, Colbourne JK, Budd GE, Brown SJ, et al. 2010. Conservation, loss, and redeployment of Wnt ligands in protostomes: implications for understanding the evolution of segment formation. *BMC Evol Biol* **10**: 374. http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-10-374.

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. 1998. Multiple sequence alignment with Clustal X. *Trends Biochem Sci* **23**: 403–5. http://www.ncbi.nlm.nih.gov/pubmed/9810230.

Jeffery WR. 2015. Regeneration, Stem Cells, and Aging in the Tunicate Ciona: Insights from the Oral Siphon. *Int Rev Cell Mol Biol* **319**: 255–82. http://linkinghub.elsevier.com/retrieve/pii/S1937644815000581.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**: 1236–1240.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656–64. http://www.ncbi.nlm.nih.gov/pubmed/11932250.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**: 59.

Kürn U, Rendulic S, Tiozzo S, Lauzon RJ. 2011. Asexual propagation and regeneration in colonial ascidians. *Biol Bull* **221**: 43–61. http://www.ncbi.nlm.nih.gov/pubmed/21876110.

Kusserow A, Pang K, Sturm C, Hrouda M, Lentfer J, Schmidt HA, Technau U, von Haeseler A, Hobmayer B, Martindale MQ, et al. 2005. Unexpected complexity of the Wnt gene family in a sea anemone. *Nature* **433**: 156–160. http://www.nature.com/doifinder/10.1038/nature03158.

Lauzon RJ, Brown C, Kerr L, Tiozzo S. 2013. Phagocyte dynamics in a highly regenerative urochordate: insights into development and host defense. *Dev Biol* **374**: 357–73. http://www.ncbi.nlm.nih.gov/pubmed/23174529 (Accessed March 31, 2014).

Lee S-A, Belyaeva O V., Kedishvili NY. 2009. Biochemical characterization of human epidermal retinol dehydrogenase 2. *Chem Biol Interact* **178**: 182–7. http://www.ncbi.nlm.nih.gov/pubmed/18926804.

Lemaire P, Smith WC, Nishida H. 2008. Ascidians and the plasticity of the chordate developmental program. *Curr Biol* **18**: R620-31. http://www.ncbi.nlm.nih.gov/pubmed/18644342.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178–89. http://www.genome.org/cgi/doi/10.1101/gr.1224503.

Loh KM, van Amerongen R, Nusse R. 2016. Generating Cellular Diversity and Spatial Form: Wnt Signaling and the Evolution of Multicellular Animals. *Dev Cell* **38**: 643–655. http://linkinghub.elsevier.com/retrieve/pii/S153458071630586X.

Lowe TM. 1999. A Computational Screen for Methylation Guide snoRNAs in Yeast. *Science (80-)* **283**: 1168–1171. http://www.sciencemag.org/cgi/doi/10.1126/science.283.5405.1168.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: A program for improved detection of transfer RNA

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

genes in genomic sequence. *Nucleic Acids Res* **25**: 955–964.
Luke GN, Castro LFC, McLay K, Bird C, Coulson A, Holland PWH. 2003. Dispersal of NK homeobox gene clusters in amphioxus and humans. *Proc Natl Acad Sci* **100**: 5292–5295. http://www.pnas.org/cgi/doi/10.1073/pnas.0836141100.
Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**: 18. http://www.gigasciencejournal.com/content/1/1/18.
MacDonald BT, Tamai K, He X. 2009. Wnt/β-Catenin Signaling: Components, Mechanisms, and Diseases. *Dev Cell* **17**: 9–26. http://linkinghub.elsevier.com/retrieve/pii/S1534580709002573.
Manni L, Zaniolo G, Cima F, Burighel P, Ballarin L. 2007. *Botryllus schlosseri*: a model ascidian for the study of asexual reproduction. *Dev Dyn* **236**: 335–52. http://www.ncbi.nlm.nih.gov/pubmed/17191252 (Accessed March 31, 2014).
Martí-Solans J, Belyaeva O V., Torres-Aguila NP, Kedishvili NY, Albalat R, Cañestro C. 2016. Coelimination and Survival in Gene Network Evolution: Dismantling the RA-Signaling in a Chordate. *Mol Biol Evol* **33**: 2401–2416. https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw118.
Martin A, Maher S, Summerhurst K, Davidson D, Murphy P. 2012. Differential deployment of paralogous Wnt genes in the mouse and chick embryo during development. *Evol Dev* **14**: 178–195. http://doi.wiley.com/10.1111/j.1525-142X.2012.00534.x.
Maumus F, Quesneville H. 2014. Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. *Nat Commun* **5**. http://www.nature.com/doifinder/10.1038/ncomms5104.
Millar RH. 1971. The biology of ascidians. *Adv Mar Biol* **9**: 1–100.
Mukai H, Sugimoto K, Taneda Y. 1978. Comparative studies on the circulatory system of the compound ascidians, *Botryllus*, *Botrylloides* and *Symplegma*. *J Morphol* **157**: 49–78.
Murata Y, Okado H, Kubo Y. 2001. Characterization of heteromultimeric G protein-coupled inwardly rectifying potassium channels of the tunicate tadpole with a unique pore property. *J Biol Chem* **276**: 18529–39. http://www.ncbi.nlm.nih.gov/pubmed/11278535.
Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, et al. 2015. Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Res* **43**: D130–D137.
Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**: 2933–2935.
NCBI Resource Coordinators. 2016. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **44**: D7-19. http://www.ncbi.nlm.nih.gov/pubmed/26615191.
Nishida H. 2005. Specification of embryonic axis and mosaic development in ascidians. *Dev Dyn* **233**: 1177–1193. http://doi.wiley.com/10.1002/dvdy.20469.
Nydam ML, Giesbrecht KB, Stephenson EE. 2017. Origin and Dispersal History of Two Colonial Ascidian Clades in the Botryllus schlosseri Species Complex ed. T.-Y. Chiang. *PLoS One* **12**: e0169944. http://dx.plos.org/10.1371/journal.pone.0169944.
O'Connell J, Schulz-Trieglaff O, Carlson E, Hims MM, Gormley N a., Cox a. J. 2015. NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics* **31**: btv057. http://bioinformatics.oxfordjournals.org/content/early/2015/02/05/bioinformatics.btv057.short?rss=1.
O'Neill RJ, O'Neill MJ, Graves JA. 1998. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**: 68–72. http://www.nature.com/doifinder/10.1038/29985.
Oka H, Watanabe H. 1957. Vascular budding, a new type of budding in *Botryllus*. *Biol Bull* **112**: 225. http://www.jstor.org/stable/10.2307/1539200?origin=crossref.
Pascual-Anaya J, D'Aniello S, Kuratani S, Garcia-Fernàndez J. 2013. Evolution of Hox gene

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

clusters in deuterostomes. *BMC Dev Biol* **13**: 26. http://bmcdevbiol.biomedcentral.com/articles/10.1186/1471-213X-13-26.

Pearson JC, Lemons D, McGinnis W. 2005. Modulating Hox gene functions during animal body patterning. *Nat Rev Genet* **6**: 893–904. http://www.ncbi.nlm.nih.gov/pubmed/16341070.

Philips A, Blein M, Robert A, Chambon J-P, Baghdiguian S, Weill M, Fort P. 2003. Ascidians as a vertebrate-like model organism for physiological studies of Rho GTPase signaling. *Biol cell* **95**: 295–302. http://www.ncbi.nlm.nih.gov/pubmed/12941527.

Piette J, Lemaire P. 2015. Thaliaceans, The Neglected Pelagic Relatives of Ascidians: A Developmental and Evolutionary Enigma. *Q Rev Biol* **90**: 117–145. http://www.journals.uchicago.edu/doi/10.1086/669266.

Primmer CR, Papakostas S, Leder EH, Davis MJ, Ragan MA. 2013. Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research. *Mol Ecol* **22**: 3216–3241. http://doi.wiley.com/10.1111/mec.12309.

Priyam A, Woodcroft BJ, Rai V, Munagala A, Moghul I, Ter F, Gibbins MA, Moon H, Leonard G, Rumpf W, et al. 2015. Sequenceserver: a modern graphical user interface for custom BLAST databases. *bioRxiv*. http://biorxiv.org/content/early/2015/11/27/033142.abstract.

Prud'homme B, Lartillot N, Balavoine G, Adoutte A, Vervoort M. 2002. Phylogenetic analysis of the Wnt gene family. Insights from lophotrochozoan members. *Curr Biol* **12**: 1395. http://www.sciencedirect.com/science/article/pii/S0960982202010680.

Rambaut A. 2016. FigTree. http://tree.bio.ed.ac.uk/software/figtree/.

Rinkevich B, Shlemberg Z, Fishelson L. 1995. Whole-body protochordate regeneration from totipotent blood cells. *Proc Natl Acad Sci U S A* **92**: 7695–9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=41212&tool=pmcentrez&rendertype=abstract (Accessed March 31, 2014).

Rinkevich Y, Douek J, Haber O, Rinkevich B, Reshef R. 2007a. Urochordate whole body regeneration inaugurates a diverse innate immune signaling profile. *Dev Biol* **312**: 131–46. http://www.ncbi.nlm.nih.gov/pubmed/17964563 (Accessed March 31, 2014).

Rinkevich Y, Paz G, Rinkevich B, Reshef R. 2007b. Systemic bud induction and retinoic acid signaling underlie whole body regeneration in the urochordate *Botrylloides leachi*. *PLoS Biol* **5**: e71. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1808485&tool=pmcentrez&rendertype=abstract (Accessed March 31, 2014).

Rinkevich Y, Rinkevich B, Reshef R. 2008. Cell signaling and transcription factor genes expressed during whole body regeneration in a colonial chordate. *BMC Dev Biol* **8**: 100. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2576188&tool=pmcentrez&rendertype=abstract (Accessed March 20, 2014).

Rinkevich Y, Voskoboynik A, Rosner A, Rabinowitz C, Paz G, Oren M, Douek J, Alfassi G, Moiseeva E, Ishizuka KJ, et al. 2013. Repeated, long-term cycling of putative stem cells between niches in a basal chordate. *Dev Cell* **24**: 76–88. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3810298&tool=pmcentrez&rendertype=abstract (Accessed March 30, 2014).

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–4. http://www.ncbi.nlm.nih.gov/pubmed/12912839.

Ross AC, Zolfaghari R. 2011. Cytochrome P450s in the regulation of cellular retinoic acid metabolism. *Annu Rev Nutr* **31**: 65–87. http://www.ncbi.nlm.nih.gov/pubmed/21529158.

Rubinstein ND, Feldstein T, Shenkar N, Botero-Castro F, Griggio F, Mastrototaro F, Delsuc F, Douzery EJP, Gissi C, Huchon D. 2013. Deep sequencing of mixed total DNA without barcodes allows efficient assembly of highly plastic Ascidian mitochondrial genomes. *Genome Biol Evol* **5**: 1185–1199. http://www.ncbi.nlm.nih.gov/pubmed/23709623.

Sabbadin A, Zaniolo G, Majone F. 1975. Determination of polarity and bilateral asymmetry in

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

palleal and vascular buds of the ascidian Botryllus schlosseri. *Dev Biol* **46**: 79–87.
Saito Y, Shirae M, Okuyama M, Cohen S. 2001. Phylogeny of Botryllid Ascidians. In *The Biology of Ascidians*, pp. 315–320, Springer Japan, Tokyo http://link.springer.com/10.1007/978-4-431-66982-1_50.
Santagati F, Abe K, Schmidt V, Schmitt-John T, Suzuki M, Yamamura K-I, Imai K. 2003. Identification of Cis-regulatory elements in the mouse Pax9/Nkx2-9 genomic region: implication for evolutionary conserved synteny. *Genetics* **165**: 235–42. http://www.ncbi.nlm.nih.gov/pubmed/14504231.
Sattler S. 2017. The Role of the Immune System Beyond the Fight Against Infection. In *The Immunology of Cardiovascular Homeostasis and Pathology* (eds. S. Sattler and T. Kennedy-Lydon), pp. 3–14, Springer International Publishing http://link.springer.com/10.1007/978-3-319-57613-8_1.
Savigny J-C. 1816. *Mémoires sur les animaux sans vertèbres*. Dufour, G., Paris http://www.biodiversitylibrary.org/bibliography/9154.
Seo H-CC, Kube M, Edvardsen RB, Jensen MF, Beck A, Spriet E, Gorsky G, Thompson EM, Lehrach H, Reinhardt R, et al. 2001. Miniature genome in the marine chordate *Oikopleura dioica*. *Science* **294**: 2506. http://www.sciencemag.org/content/294/5551/2506.short.
Simakov O, Kawashima T, Marlétaz F, Jenkins J, Koyanagi R, Mitros T, Hisata K, Bredeson J, Shoguchi E, Gyoja F, et al. 2015. Hemichordate genomes and deuterostome origins. *Nature* **527**: 459–465. http://www.nature.com/doifinder/10.1038/nature16150.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V. 2015. BUSCO : assessing genome assembly and annotation completeness with single-copy orthologs. *Genome Anal* 9–10.
Simmen MW, Leitgeb S, Charlton J, Jones SJ, Harris BR, Clark VH, Bird A. 1999. Nonmethylated transposable elements and methylated genes in a chordate genome. *Science* **283**: 1164–7. http://www.ncbi.nlm.nih.gov/pubmed/10024242.
Simpson JT. 2014. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics* **30**: 1228–1235.
Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Res* **19**: 1117–1123. http://genome.cshlp.org/cgi/doi/10.1101/gr.089532.108.
Small KS, Brudno M, Hill MM, Sidow A. 2007. A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. *Genome Biol* **8**: R41. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-3-r41.
Smit A, Hubley R. 2015. RepeatModeler Open-1.0.
Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0.
Sobreira TJP, Marletaz F, Simoes-Costa M, Schechtman D, Pereira AC, Brunet F, Sweeney S, Pani A, Aronowicz J, Lowe CJ, et al. 2011. Structural shifts of aldehyde dehydrogenase enzymes were instrumental for the early evolution of retinoid-dependent axial patterning in metazoans. *Proc Natl Acad Sci* **108**: 226–231. http://www.pnas.org/cgi/doi/10.1073/pnas.1011223108.
Somorjai IML, Escrivà H, Garcia-Fernàndez J. 2012. Amphioxus makes the cut—Again. *Commun Integr Biol* **5**: 499–502. http://www.tandfonline.com/doi/abs/10.4161/cib.21075.
Spagnuolo A, Ristoratore F, Di Gregorio A, Aniello F, Branno M, Di Lauro R. 2003. Unusual number and genomic organization of Hox genes in the tunicate Ciona intestinalis. *Gene* **309**: 71–9. http://www.ncbi.nlm.nih.gov/pubmed/12758123.
Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**: 215–225.
Stapley J, Santure AW, Dennis SR. 2015. Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Mol Ecol* **24**: 2241–2252. http://doi.wiley.com/10.1111/mec.13089.
Stoick-Cooper CL, Weidinger G, Riehle KJ, Hubbert C, Major MB, Fausto N, Moon RT. 2007.

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Distinct Wnt signaling pathways have opposing roles in appendage regeneration. *Development* **134**: 479–89. http://dev.biologists.org/cgi/doi/10.1242/dev.001123.

Stolfi A, Lowe EK, Racioppi C, Ristoratore F, Brown CT, Swalla BJ, Christiaen L. 2014. Divergent mechanisms regulate conserved cardiopharyngeal development and gene expression in distantly related ascidians. *Elife* **3**: e03728.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms ed. C. Gibas. *PLoS One* **6**: e21800. http://dx.plos.org/10.1371/journal.pone.0021800.

Suzuki MM, Kerr ARW, De Sousa D, Bird A. 2007. CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res* **17**: 625–631. http://www.genome.org/cgi/doi/10.1101/gr.6163007.

Takatori N, Butts T, Candiani S, Pestarino M, Ferrier DEK, Saiga H, Holland PWH. 2008. Comprehensive survey and classification of homeobox genes in the genome of amphioxus, Branchiostoma floridae. *Dev Genes Evol* **218**: 579–590. http://link.springer.com/10.1007/s00427-008-0245-9.

Taketa DA, Nydam ML, Langenbacher AD, Rodriguez D, Sanders E, De Tomaso AW. 2015. Molecular evolution and in vitro characterization of Botryllus histocompatibility factor. *Immunogenetics* **67**: 605–623. http://link.springer.com/10.1007/s00251-015-0870-1.

Tassy O, Dauga D, Daian F, Sobral D, Robin F, Khoueiry P, Salgado D, Fox V, Caillol D, Schiappa R, et al. 2010. The ANISEED database: Digital representation, formalization, and elucidation of a chordate developmental program. *Genome Res* **20**: 1459–1468. http://genome.cshlp.org/cgi/doi/10.1101/gr.108175.110.

The Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res* **43**: D1049–D1056. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku1179.

Tsagkogeorga G, Cahais V, Galtier N. 2012. The Population Genomics of a Fast Evolver: High Levels of Diversity, Functional Constraint, and Molecular Adaptation in the Tunicate Ciona intestinalis. *Genome Biol Evol* **4**: 852–861. https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evs054.

Tsagkogeorga G, Turon X, Galtier N, Douzery EJP, Delsuc F. 2010. Accelerated Evolutionary Rate of Housekeeping Genes in Tunicates. *J Mol Evol* **71**: 153–167. http://link.springer.com/10.1007/s00239-010-9372-9.

UniProt Consortium. 2015. UniProt: a hub for protein information. *Nucleic Acids Res* **43**: D204–D212. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku989.

Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B, Fan HC, Mantalas GL, Palmeri KJ, et al. 2013a. The genome sequence of the colonial chordate, *Botryllus schlosseri*. *Elife* **2**: 1–24. http://elifesciences.org/lookup/doi/10.7554/eLife.00569.

Voskoboynik A, Newman AM, Corey DM, Sahoo D, Pushkarev D, Neff NF, Passarelli B, Koh W, Ishizuka KJ, Palmeri KJ, et al. 2013b. Identification of a Colonial Chordate Histocompatibility Gene. *Science (80-)* **341**: 384–387. http://www.sciencemag.org/cgi/doi/10.1126/science.1238036.

Voskoboynik A, Simon-Blecher N, Soen Y, Rinkevich B, De Tomaso AW, Ishizuka KJ, Weissman IL. 2007. Striving for normality: whole body regeneration through a series of abnormal generations. *FASEB J* **21**: 1335–44. http://www.ncbi.nlm.nih.gov/pubmed/17289924 (Accessed March 31, 2014).

Wada S, Tokuoka M, Shoguchi E, Kobayashi K, Di Gregorio A, Spagnuolo A, Branno M, Kohara Y, Rokhsar D, Levine M, et al. 2003. A genomewide survey of developmentally relevant genes in Ciona intestinalis. *Dev Genes Evol* **213**: 222–234. http://link.springer.com/10.1007/s00427-003-0321-0.

Wang X-P, Suomalainen M, Felszeghy S, Zelarayan LC, Alonso MT, Plikus M V, Maas RL, Chuong C-M, Schimmang T, Thesleff I. 2007. An integrated gene regulatory network controls

bioRxiv preprint doi: https://doi.org/10.1101/152983; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

stem cell proliferation in teeth. *PLoS Biol* **5**: e159. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1885832&tool=pmcentrez&rendertype=abstract (Accessed October 28, 2010).

Wences AH, Schatz MC. 2015. Metassembler: merging and optimizing de novo genome assemblies. *Genome Biol* **16**: 207. http://genomebiology.com/2015/16/1/207.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–9. http://www.ncbi.nlm.nih.gov/pubmed/18349386.

Zhan A, Briski E, Bock DG, Ghabooli S, MacIsaac HJ. 2015. Ascidians as models for studying invasion success. *Mar Biol* **162**. http://link.springer.com/10.1007/s00227-015-2734-5.

Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, Yang P, Zhang L, Wang X, Qi H, et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**: 49–54. http://dx.doi.org/10.1038/nature11413.

Zimin A V., Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**: 2669–2677. http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btt476.

Zondag LE, Rutherford K, Gemmell NJ, Wilson MJ. 2016. Uncovering the pathways underlying whole body regeneration in a chordate model, *Botrylloides leachi* using de novo transcriptome analysis. *BMC Genomics* **17**: 114. http://www.biomedcentral.com/1471-2164/17/114.