OXFORD
G3 Genes | Genomes | Genetics

G3, 2023, 13(10), jkad181
https://doi.org/10.1093/g3journal/jkad181
Advance Access Publication Date: 9 August 2023
Genome Report

De novo genome assembly and comparative genomics for the colonial ascidian *Botrylloides violaceus*

Jack T. Sumner,[1,‡] Cassidy L. Andrasz,[1,‡] Christine A. Johnson,[1] Sarah Wax,[1] Paul Anderson,[2] Elena L. Keeling,[1,*] Jean M. Davidson[1,*]

[1]Department of Biological Sciences, California Polytechnic State University, San Luis Obispo, CA 93407, USA
[2]Department of Computer Science and Software Engineering, California Polytechnic State University, San Luis Obispo, CA 93407, USA

*Corresponding author: Department of Biological Sciences, California Polytechnic State University, San Luis Obispo, CA 93407, USA. E-mail: ekeeling@calpoly.edu;
*Corresponding author: Department of Biological Sciences, California Polytechnic State University, San Luis Obispo, CA 93407, USA. E-mail: jdavid06@calpoly.edu
‡These authors contributed equally to this work and share first authorship.

Ascidians have the potential to reveal fundamental biological insights related to coloniality, regeneration, immune function, and the evolution of these traits. This study implements a hybrid assembly technique to produce a genome assembly and annotation for the botryllid ascidian, *Botrylloides violaceus*. A hybrid genome assembly was produced using Illumina, Inc. short and Oxford Nanopore Technologies long-read sequencing technologies. The resulting assembly is comprised of 831 contigs, has a total length of 121 Mbp, N50 of 1 Mbp, and a BUSCO score of 96.1%. Genome annotation identified 13 K protein-coding genes. Comparative genomic analysis with other tunicates reveals patterns of conservation and divergence within orthologous gene families even among closely related species. Characterization of the Wnt gene family, encoding signaling ligands involved in development and regeneration, reveals conserved patterns of subfamily presence and gene copy number among botryllids. This supports the use of genomic data from nonmodel organisms in the investigation of biological phenomena.

Key words: genome assembly; Illumina; Oxford Nanopore; comparative genomics; tunicate; Wnt; *Botrylloides violaceus*

Introduction

High-quality reference genomes of core model organisms have transformed modern biology (Bonini and Berger 2017); however, scarce genomics research on nonmodel species has historically limited insights from across the tree of life. Genomics technologies have become more amenable to discoveries in nonmodel organism biology with the rise of high throughput sequencing (da Fonseca *et al.* 2016, 2020; Whitacre *et al.* 2017; Etherington *et al.* 2020). Genome sequence data from a more diverse range of organisms will enable deeper understanding of both patterns of genome evolution and the genetic underpinnings of diverse biological phenomena.

The hybrid de novo genome assembly method has enabled cost-effective genome projects to improve existing reference genomes and to assemble novel drafts (Mostovoy *et al.* 2016; Tan *et al.* 2018; Jiang *et al.* 2019; Wallberg *et al.* 2019; Xing *et al.* 2019). This method integrates multiple technologies to produce assemblies of greater quality by complementing the strengths and weaknesses of each approach (Utturkar *et al.* 2014). A common hybrid assembly method combines high-coverage, short-read and low-coverage, long-read sequencing data to more accurately resolve complex repetitive regions (Zimin *et al.* 2013; Utturkar *et al.* 2014; Tan *et al.* 2018). This genome assembly technique provides low-cost yet functional sequence data. We define "functional" as sufficient to uncover novel biological insights, especially when compared to closely related species. Here, we use this approach for a colonial ascidian, a member of the tunicate lineage.

Tunicates are a diverse group of marine invertebrate chordates that have been studied in the context of developmental biology, evolutionary genomics, and regenerative medicine. As the closest extant relative of vertebrates, tunicates are key to understanding the rise of vertebrate traits and the history of chordate evolution (Delsuc *et al.* 2006). Tunicate genomes are rapidly evolving and diverse (Denoeud *et al.* 2010; Berna and Alvarez-Valin 2014). Several published genomes from across the three classes of tunicates have facilitated the rise of comparative and functional genomics in this lineage (Dehal *et al.* 2002; Jue *et al.* 2016; Blanchoud *et al.* 2018; Satou *et al.* 2019; DeBiasse *et al.* 2020; Bliznina *et al.* 2021). These have revealed substantial genomic changes such as a high degree of gene loss (Dehal *et al.* 2002; Berna and Alvarez-Valin 2014) and unusually fast sequence evolution (Denoeud *et al.* 2010; Berna and Alvarez-Valin 2014). In addition, massive reorganization events have diverged their global genomic architecture from other metazoans; only faint syntenic relationships are observed even between higher taxa of tunicates (Denoeud *et al.* 2010). These high rates of genomic reorganization and compaction make tunicate genomes uncommonly plastic (Denoeud *et al.* 2010; Berna and Alvarez-Valin 2014), making them a valuable system for further insights about genome evolution.

Tunicate development and life histories are also highly variable; coloniality has evolved multiple times and regenerative ability varies (Lemaire 2011). Colonial ascidians, containing many discrete bodies called zooids that share a peripheral vasculature, have acquired asexual budding that mediates colony growth

Received on 25 January 2023; accepted on 12 July 2023

(Berrill 1941; Watanabe and Newberry 1976; Zeng *et al.* 2006; Alié *et al.* 2018). This innovation may be associated with extensive regenerative capacity; until recently, colonial ascidians in the botryllid clade were the only known chordates capable of whole-body regeneration (WBR) (Rinkevich *et al.* 1995; Blanchoud *et al.* 2018; Kassmer *et al.* 2020). Botryllids, consisting of the *Botrylloides* and *Botryllus* genera, are capable of WBR but show differences in its regulation (Rinkevich *et al.* 2007; Voskoboynik *et al.* 2007; Brown *et al.* 2009). The recent discovery of a solitary ascidian capable of regenerating all body structures from an amputated fragment (Gordon *et al.* 2021) reveals additional complexity in the evolution of regenerative abilities and further emphasizes the potential of ascidians as model organisms for studies of regeneration and stem cell biology.

Regenerative abilities across species are diverse and differ in phylogenetic distribution, with evidence for losses in various lineages (Bely 2010; Lai and Aboobaker 2018). Regeneration in vertebrates is generally more limited, including muscle (Zullo *et al.* 2020), liver (Delgado-Coello 2021), or distal structures such as limbs or tails in various fish and amphibian species (Yokoyama 2008; Darnet *et al.* 2019; Ferrario *et al.* 2020). Historically the best models for whole-body regeneration have been planarians (*Platyhelminthes*) and Cnidaria such as *Hydra* (Reddien and Alvarado 2004; Reddy *et al.* 2019), but there is increasing recognition of the wide range of regenerative abilities across species (Lai and Aboobaker 2018; Mokalled and Poss 2018; Ferrario *et al.* 2020; Edgar *et al.* 2021).

One fundamental question is the extent to which the genetic pathways underlying regeneration are conserved. The Wnt signaling pathway has been implicated in the regulation of stem cells and regeneration across many taxa, including ascidians (Clevers *et al.* 2014; Zondag *et al.* 2016; Nusse and Clevers 2017; Garcia *et al.* 2018; Leucht *et al.* 2019; Kassmer *et al.* 2020; Borisenko *et al.* 2021). The Wnt gene family encodes secreted glycoprotein ligands that initiate signaling pathways leading to changes in transcription associated with cell fate and proliferation, as well as cell and tissue organization (Niehrs 2012; Willert and Nusse 2012; Anthony *et al.* 2020). Evolution of Wnt genes, including changes in gene copy number, may play a role in varying regenerative capabilities as well as other aspects of development (Somorjai *et al.* 2018; Martí-Solans *et al.* 2021).

Although scarce genomic data have historically limited molecular investigations of botryllid biology, published genomes from the two sister genera, *Botrylloides leachii* and *Botryllus schlosseri*, are now available (Voskoboynik *et al.* 2013; Blanchoud *et al.* 2018). Additional ascidian genome sequences will help reveal lineage-specific diversification of genes, such as Wnts, and provide the basis for further investigation of potential roles in regulating regeneration. Undergraduate students carried out the entire genome project, from obtaining local samples, optimizing DNA extraction, constructing libraries and carrying out sequencing, as well as all subsequent analysis.

This study used Illumina, Inc. short reads and Oxford Nanopore Technologies long reads to produce the first draft hybrid assembly genome sequence of the colonial ascidian *Botrylloides violaceus* (Fig. 1). Annotation of the genome allowed initial comparative genomics with other tunicates and a comprehensive survey of Wnt genes. This genome will allow further exploration of genome evolution and mechanisms involved in coloniality and regeneration.

Materials and methods

Sample acquisition

Wild *B. violaceus* colonies were collected from Morro Bay, CA (CDFW permit GM-190280002-19028) and subsequently cleared of visible detritus with forceps and brush. Colonies were cultivated and starved in filtered (0.22 μm) seawater at 10°C for approximately two days to reduce potential microbial contamination from the gut. Approximately, 100 μL of blood was extracted from a colony using a 1 mL needle and syringe. DNA for Illumina sequencing was isolated using a Qiagen DNeasy Blood & Tissue Kit according to manufacturer instructions. Initial ONT sequencing using the same DNA used for Illumina sequencing produced low quality sequencing data (data not shown). From another colony, enriched blood was acquired by macerating 1 g of colony with a syringe plunger against a 40 μm sieve (Falcon™ Cell Strainer—Fisher Science) as previously described (Rosental *et al.* 2018). Over 5 μg of DNA was extracted per 200 μL enriched blood using the Zymo Quick-DNA HMW MagBead Kit. Manufacturer's instructions were followed except for increasing the volume of magnetic beads to 50 μL total. Samples were visually assessed via gel electrophoresis and quantified with Qubit Fluorometric Quantitation (ThermoFisher, Inc.).

Whole-genome sequencing and processing

Illumina sequencing

DNA was processed using the Nextera XT v3 Library Preparation Kit (Illumina, San Diego, CA, USA) to construct two libraries for short-read sequencing. 6.3 Gbp of data in 41.8 million reads (2 × 150 cycles) were produced using the MiniSeq System (Illumina, San Diego, CA, USA). Illumina short reads were quality checked using FastQC v0.11 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Forward and reverse reads from each library were concatenated into one pair of fastq files. Adapter sequences and poor quality data were removed from reads with Trimmomatic v0.36 (ILLUMINACLIP:2:30:10, LEADING:5, TRAILING:5, SLIDINGWINDOW:4:20, MINLEN:50) (Bolger *et al.* 2014). The Nextera XT adapter sequence (5′-CTGTCTCTTATACACAT CT-3′) was supplied for trimming in fasta format. Bacterial, archaeal, and viral environmental contaminants were deconvoluted from presumptive *B. violaceus* sequence data using Kraken v1.1.1 (Wood and Salzberg 2014) with the MiniKraken 8GB database. This processing culminated in approximately 37.4 million "clean" paired-end reads.

ONT sequencing

Libraries for ONT sequencing were constructed using SQK-RAD004 Rapid Sequencing Kit (Oxford Nanopore Technologies, Oxford, UK) following the manufacturer's instructions with the exception that the DNA mass input was increased to approximately 1 μg. ONT libraries were sequenced using the MinIon system (Oxford Nanopore, Oxford, UK) to produce 4.08 Gbp of sequencing data with an N50 of 9,585 bp. Nanopore long reads were basecalled using Guppy v4.0.15 (–config dna_r9.4.1_450bps_fast.cfg) (Oxford Nanopore Technologies). Reads were then quality filtered to include those with quality scores greater than or equal to seven (approximately 85% basecall accuracy) using NanoFilt v2.7.1 (De Coster *et al.* 2018). Filtering resulted in 3.56 Gbp of long-read data, with a read length N50 of 9,799 bp.

Genome assembly and annotation

Genome assembly

Standard short-read, long-read, and hybrid assembly approaches were performed on appropriate data. We employed the MaSuRCA v3.4.2 (Zimin *et al.* 2017) algorithm for hybrid de novo assembly of both reads types with the Flye assembler for final assembly of mega reads, rather than previous CABOG versions, for improved

Fig. 1. *Botrylloides violaceus* imaged in Morro Bay, CA.

efficacy and efficiency. Short- and long-read only assemblies were produced with ABySS v2.0.2 (Jackman *et al.* 2017) and Flye v2.8.1 (Kolmogorov *et al.* 2019), respectively. To estimate completeness, BUSCO analysis (Fig. 2) was performed on each method and key genome assembly statistics were calculated with QUAST (see *Genome Statistics* methods).

Genome annotation

Annotation of the *B. violaceus* hybrid assembly was completed using the Genome Sequence Annotation Server (GenSAS) v6.0 (Humann *et al.* 2019). GenSAS allowed all listed annotation analyses to be streamlined into one pipeline. Prior to annotation, the hybrid assembly was filtered to only contain scaffolds and contigs 1,000 bp or greater, which resulted in removing 101 contigs (filtered assembly length 120,854,445 bp). Removing these contigs did not change the BUSCO score. Since this is a de novo annotation, no prior *B. violaceus* data or results were uploaded to assist in annotation. However, since the sister species *B. leachii* has an annotated genome assembly and transcriptome, the *B. leachii* transcripts and were obtained from ANISEED (Tassy *et al.* 2010; Brozovic *et al.* 2018) and uploaded to GenSAS to use as evidence in downstream programs. RepeatMasker v4.0.7 (http://www.repeatmasker.org) was used with default parameters, NCBI RMBlast v2.2.27 search engine and another well-studied tunicate, *Ciona intestinalis*, as the premasked DNA source to aid in finding repetitive regions in the *B. violaceus* genome. RepeatModeler v1.0.11 (http://www.repeatmasker.org/RepeatModeler/) was used to find repetitive elementsde novo in the hybrid assembly. The results from RepeatMasker and RepeatModeler were combined into a consensus masked genome that was used in downstream annotation (GenSAS built-in masked consensus tool). For structural annotation of the genome, *B. leachii* transcripts were aligned to the *B. violaceus* hybrid assembly using blastn v2.7.1 (Camacho *et al.* 2009) with an e-value cutoff of 1e−30 and all other parameters left default. Ab initio gene predictions were made using GeneMark-ES v4.38 (Lomsadze *et al.* 2005) with the max_contig set to 6,000,000 bp, larger than the hybrid assembly's longest contig as to not split the contigs, and the min_contig set to 1,000 bp to include all contigs in the training of the algorithm; all other parameters were left default. EVidenceModeler v1.1.1 (Haas *et al.* 2008) combined the *B. leachii* transcript alignments from blastn and the ab initio gene predictions from GeneMark-ES into one consensus structural annotation using the default weights of 10 for the transcripts and 1 for the ab initio gene predictions. This consensus of gene predictions was assigned as the Official Gene Set (OGS) of which functional annotation would take place using InterProScan v5.44-79.0 (Jones *et al.* 2014). Additional functional annotation of structurally annotated proteins was conducted outside of GenSAS to create an EggNOG-based annotation using eggNOG-mapper v1.0.3 (Huerta-Cepas *et al.* 2017).

Genome statistics

Genome statistics including number of contigs or scaffolds, N50, and total length were obtained using the Quality Assessment Tool for genome assemblies, QUAST v5.0.2 (Gurevich *et al.* 2013). Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.1.3 assessment (Manni *et al.* 2021) was used to estimate assembly completion using the eukaryota_odb10 (Eukaryota) and metazoa_odb10 (Metazoa) datasets compiled from OrthoDB v10 (Kriventseva *et al.* 2019). BUSCO assessments were completed independently for each genome using the genome mode. Furthermore, analysis of the annotation of the *B. violaceus* hybrid assembly was completed using BUSCO in protein mode using all protein sequences from the annotation.

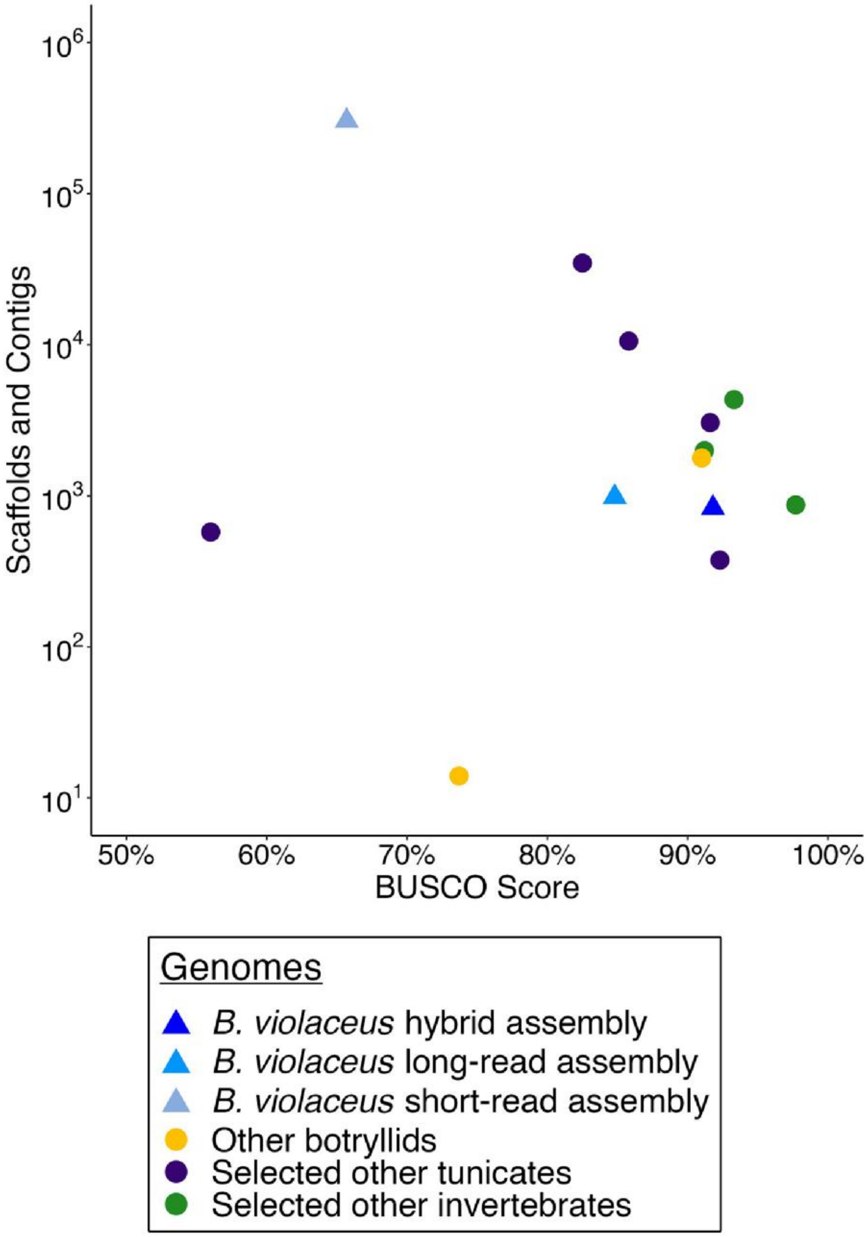


Fig. 2. Using BUSCO score and the total number of scaffolds and contigs in genome assemblies to assess their quality and completeness. *Botrylloides violaceus* assemblies are compared to genomes of other botryllids (*B. leachii*, *B. schlosseri*), other selected tunicates (*O. dioica*, *P. fumigata*, *C. savignyi*, *H. roretzi*, *M. oculata*), and other selected invertebrates (*S. purpuratus*, *A. californica*, *H. robusta*). Triangular points are *B. violaceus* assemblies color coded by assembly type, while circular points represent other genomes color coded by grouping. BUSCO scores are the percentage of complete BUSCOs based on the Metazoa dataset.

Comparative genomics

Acquisition of relevant genomic reference data

Several genomes were acquired for comparative genomics analysis. Tunicate genomes including *B. leachii*, *B. schlosseri*, *Halocynthia aurantium*, *Halocynthia roretzi*, *Mogula occidentalis*, *Mogula oculata*, *Phallusia mammilata*, *Phallusia fumigata*, *Ciona savignyi*, and *Oikopleura dioica* were acquired from ANISEED (Tassy et al. 2010; Brozovic et al. 2018). Selected other invertebrate genomes used to assess the quality of the *B. violaceus* hybrid assembly included *Strongylocentrotus purpuratus*, *Aplysia californica*, and *Helobdella robusta* and have RefSeq accession numbers of GCF_000002235.5, GCF_000002075.1, and GCF_000326865.1, respectively.

OrthoFinder analysis

To better understand the evolution of botryllid-specific traits and to distill relevant information from high-dimensional genomic data, we implemented a gene-centric comparative genomics approach. To that end, pairwise comparisons of protein-coding genes were performed to identify groups of evolutionarily conserved genes across tunicates using OrthoFinder (Emms and Kelly 2019). This high-dimensional dataset was then refined using a stepwise analytical approach that uses publicly available bioinformatics software and in-house scripts to identify copy number variation patterns and putative functions that are unique to botryllids.

The OrthoFinder algorithm was implemented to identify groups of homologous genes (orthogroups) from across the tunicate lineage. Briefly, annotated proteomes from select tunicates (i.e. *B. leachii*, *B. schlosseri*, *H. aurantium*, *H. roretzi*, *M. occidentalis*, *M. oculata*, *P. mammilata*, *P. fumigata*, *C. savignyi*, and *O. dioica*) were downloaded from ANISEED and processed, along with the putative *B. violaceus* proteome from the GenSAS annotation, using OrthoFinder v2.5.2 with default parameters. After initial orthogroups were identified, basic quality control was performed to broadly assess the reliability of these data, as described in the OrthoFinder documentation (Emms and Kelly 2019). The inferred phylogenetic tree was visualized using iTOL (Letunic and Bork 2021) to confirm its consistency with previous reports as reliable phylogenetic reconstruction is essential to downstream analyses.

To dissect botryllid-specific patterns of copy number variations, orthogroups were evaluated using a phylogenetically informed approach. Briefly, gene trees from each node of the inferred phylogeny are clustered using Orthofinder to create lineage-specific hierarchical orthogroups (HOGs) (Altenhoff *et al.* 2013). We then selected HOGs corresponding to the botryllid lineage for our remaining analysis and hereto referred to as HOGBs. Gene counts from HOGBs were used to create a species by HOGB matrix that was organized using unsupervised hierarchical clustering; this cladogram was then cut into eighteen clusters, each containing unique sets of HOGBs. The number of clusters was determined by cutting the cladogram into 1–20 clusters and computing the average silhouette score, a metric used to assess cluster quality of fit, at each point; the local maxima was found at eighteen clusters and was thus used for the remaining analysis (Supplementary Fig. S1).

Gene ontology analysis

KEGG (Kanehisa and Goto 2000) orthology terms were assigned to each *B. violaceus* gene based on our EggNOG-based annotation. Cluster enrichment was performed using the compareCluster and enrichKEGG functions in the R package clusterProfiler v4.2.1 (Wu *et al.* 2021). Additionally, we assessed the presence of relevant developmentally related pathways in each cluster. HOG cluster output provided information on which *B. violaceus* gene was within each cluster, and the corresponding KEGG annotation information for that gene. Genes with KEGG terms corresponding to investigated pathways were recorded within each cluster. The following KEGG pathways were investigated: Wnt (ko04310), TGF-beta (ko04350), Hedgehog (ko04340), Notch (ko04330), ErbB/EGF (ko04012), TNF (ko04668), cytokine–cytokine receptor interaction (ko04060), and retinol metabolism (ko00830).

Wnt family analysis

Annotation data were used to identify putative Wnt pathway gene orthologs in the *B. violaceus* genome. We then focused on characterizing the repertoire of Wnt subfamilies, using BLAST and multiple sequence alignment analyses.

Using the functional annotation data from InterProScan, genes with the IPR term IPR005817 (Wnt description) were aligned against the NCBI nonredundant nucleotide collection (blastn with nr/nt) and nonredundant protein sequences (blastp with nr) databases to confirm Wnt gene orthologs. Indeterminate results showing multiple putative subfamilies for a single Wnt sequence were clarified using the ANISEED built-in BLAST tool with tblastn (Brozovic *et al.* 2018); similarity searches were performed against two Gene Model databases, *B. leachii* (SBv3) and *B. schlosseri* (2013), with default parameters. Additional methods were required to support identification of Wnt genes 1, 11, and A in *B. violaceus* utilizing Integrative Genomics Viewer (IGV) v2.9.2 (Thorvaldsdóttir *et al.* 2013) and additional BLAST and MSA analyses. Upon manual curation and phylogenetic analysis (described below) of Wnt sequences, we identified several *B. leachii* Wnts that had greater identity to alternative subfamilies than previously described (Blanchoud *et al.* 2018) and reclassified them accordingly for our analysis (supplemental data on GitHub).

For phylogenetic analysis of Wnt subfamilies, orthologous Wnt protein sequences from 12 different species (tunicates and human) were obtained from Martí-Solans *et al.* (2021). Wnt protein sequences from *B. leachii*, obtained from ANISEED, and *B. violaceus*, identified in this study, were added to the those 12 species for comprehensive phylogenetic analysis. A custom, automated phylogeny workflow was built in NGPhylogeny (Lemoine *et al.* 2019) and was used to construct the Wnt phylogeny. The Wnt protein sequences were aligned in Clustal Omega (Sievers *et al.* 2011). The maximum likelihood phylogeny constructor PhyML v3.3 (Guindon *et al.* 2010) was used with the default parameters for amino acid data and SH-like aLRT (Shimodaira–Hasegawa-like approximate likelihood ratio test) branch supports. The phylogeny was annotated and edited in iTOL (Letunic and Bork 2021). Determination of Wnt subfamily presence and copy number in other species was based on published literature (Nayak *et al.* 2016; Somorjai *et al.* 2018; Martí-Solans *et al.* 2021). Species relationships used the current accepted phylogeny of tunicates (DeBiasse *et al.* 2020).

Table 1. Genome statistics for *B. violaceus* assemblies.

	ABySS Short-read assembly	Flye Long-read assembly	MaSuRCA Hybrid assembly
Number of contigs	302,459	981	831
Number of contigs ≥1,000 bp	31,964	840	730
Percent contigs ≥1,000 bp	10.6	85.6	87.8
Number of contigs ≥50,000 bp	0	175	245
Largest contig (bp)	44,514	5,304,203	5,758,483
Total length (bp)	160,170,160	122,263,676	120,925,257
N50 (bp)	1,073	1,647,705	1,028,308
L50	29,444	18	28
Number N's per 100 Kbp	3.73	1.31	1.41

Results and discussion

Genome assembly and annotation

The ABySS (short-read), Flye (long-read), and MaSuRCA (hybrid) genome assemblers produced assemblies of lengths 160.1, 122.2, and 120.9 Mbp, respectively (Table 1). Although the hybrid assembly technique did produce the smallest draft genome, it produced the most contiguous genome. The MaSuRCA hybrid assembly contains the smallest number of contigs, the greatest number of contigs over 50 Kbp, and the greatest length of largest contig. The Flye assembly had a slightly larger N50 than the MaSuRCA assembly; however, both substantially surpassed the N50 of the ABySS assembly by 1,000 fold (Table 1). Though the hybrid assembly did not produce the largest N50, it did have the largest L50 indicating a higher percentage of the hybrid assembly consists of long contigs varying in length between its N50 and largest contig (Table 1). The ABySS assembly had few contigs over the length of 1,000 bp, which indicates little realistic viability of the genome for downstream analysis. While the Flye and MaSuRCA assemblies are similar in structural statistics, the MaSuRCA hybrid assembly represents the most complete and functional draft genome for *Botrylloides violaceus* and is used as the reference genome for the remainder of the analysis on this species.

MaSuRCA estimated the *B. violaceus* genome size to be 139 Mb, suggesting that around 86% of the genome has been assembled (Table 2). The hybrid assembly and estimated genome size are the smallest of the botryllids where current draft genome assemblies are sized at 159 Mbp for *B. leachii* and 580 Mbp for *B. schlosseri*, and estimated genome sizes are 194 Mbp for *B. leachii* and 725 Mbp for *B. schlosseri* (Voskoboynik *et al.* 2013; Blanchoud *et al.* 2018).

BUSCO scores for the hybrid assembly show over 90% of highly conserved orthologs present from both the Metazoa and Eukaryota datasets (Table 2). To further assess the quality of the *B. violaceus* assemblies, comparisons were drawn using the other botryllid genomes (*B. leachii* and *B. schlosseri*), other tunicates, and other invertebrates (Fig. 2). BUSCO score (complete single-copy and duplicated BUSCOs) from the Metazoa dataset is used to assess genome completion and the total number of scaffolds and contigs in each genome is used to assess the contiguity and quality. Compared to all other assemblies, the *B. violaceus* hybrid assembly is in the top 25% of BUSCO scores while having an intermediate scaffold and contig number. Both the *B. violaceus* short and long-read assemblies have a lower BUSCO completeness score and greater numbers of scaffolds and contigs compared to the hybrid assembly. The other selected invertebrates tend to have higher BUSCO scores compared to tunicates, which is consistent with previous accounts of genomic loss events in the tunicate ancestor (Seo *et al.* 2001). Our *B. violaceus* hybrid assembly is therefore of high quality and completeness when compared to tunicates, and is comparable to other published invertebrate genomes (Fig. 2).

Annotation of the *B. violaceus* hybrid genome assembly provided sufficient predictions to use in future analyses. The GenSAS annotation of the hybrid genome assembly predicted 13.4 K genes, 12.9 K protein-coding genes, with 78% of protein-coding genes being functionally annotated with InterProScan. In total, 21% of the genome was masked as repetitive regions. Additionally, the protein annotation had a BUSCO score of 86% using the Metazoan database (Table 2).

Initially, annotation used evidence including transcripts from related species and gene models produced from ab initio gene predictors such as SNAP (7/28/2006 version) (Korf 2004) and GeneMark-ES. However, when using SNAP to predict genes based on the *Ciona intestinalis* HMM, the resulting annotation contained significant fragmentation of the genes in the OGS in comparison to orthologs in the other botryllids, *B. schlosseri* and *B. leachii*. Removing SNAP from the OGS and using only the *B. leachii* transcript alignments and GeneMark-ES predictions rectified the fragmentation of genes and significantly decreased the number of predicted genes from 23 to 13 K, which is comparable to the 15 K predicted genes in *B. leachii* (Blanchoud *et al.* 2018). Additionally, when identifying members of the Wnt family based on the final annotation, no Wnt 1 was found in the annotation despite its presence in closely related species. However, Wnt 1 was successfully located due to the GeneMark-ES predictions. Inconsistencies in this gene annotation may be due to heavier weighting of the *B. leachii* transcripts over the ab initio GeneMark-ES predictions when combining evidence to support the final gene models.

Table 2. Additional hybrid assembly statistics, including relevant annotation statistics.

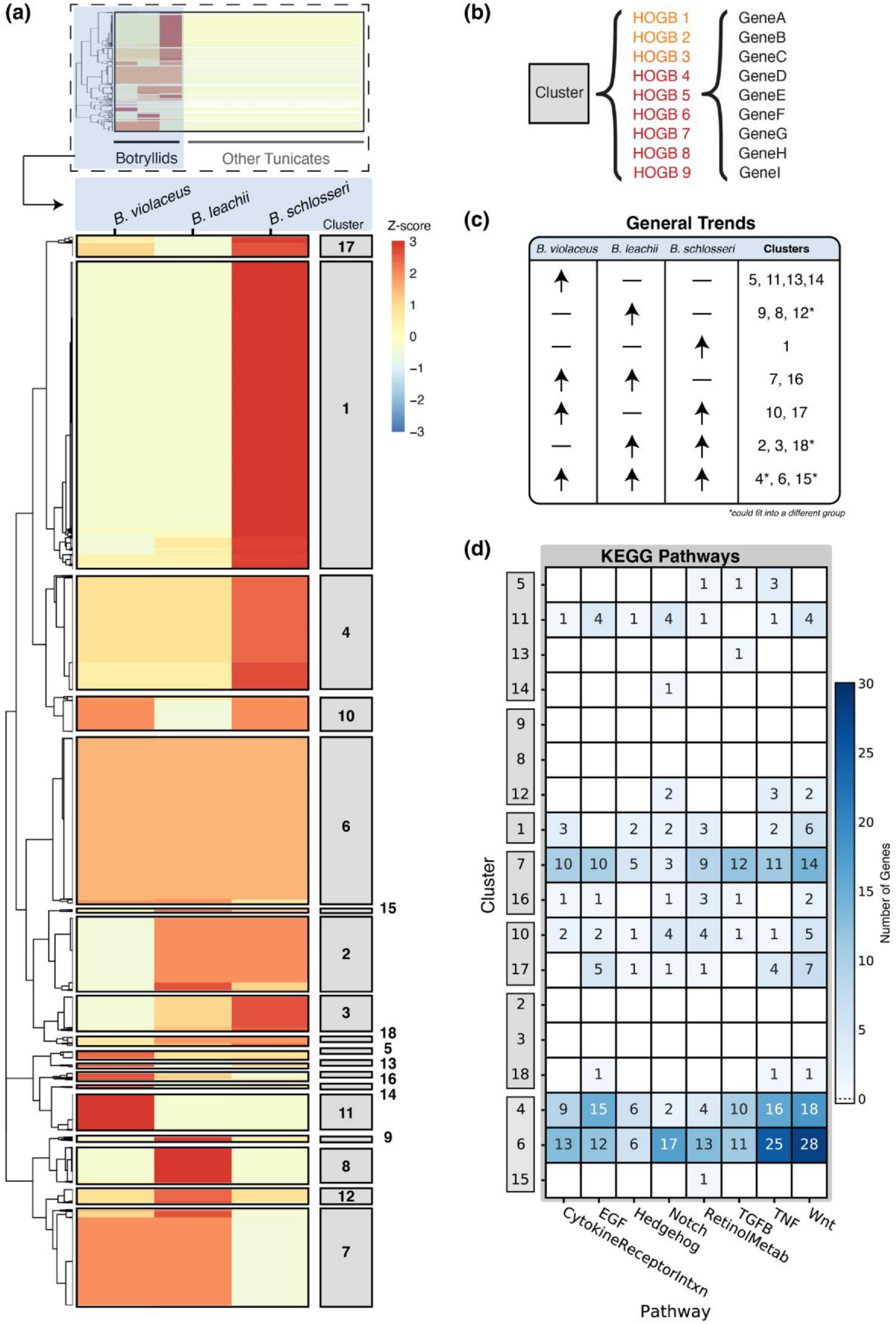
	MaSuRCA Hybrid assembly
Assembly statistics	
Metazoa ODB BUSCO score (*n* = 954)	
Complete BUSCOs	876 (91.9%)
Complete and single-copy BUSCOs	864 (90.6%)
Complete and duplicated BUSCOs	12 (1.3%)
Fragmented BUSCOs	26 (2.7%)
Missing BUSCOs	52 (5.4%)
Eukaryota ODB BUSCO score (*n* = 255)	
Complete BUSCOs	245 (96.1%)
Number of scaffolds	15
Number of contigs	816
Estimated genome size (bp)	139,493,337
Annotation statistics	
Metazoa ODB BUSCO score (*n* = 954)	
Complete BUSCOs	824 (86.4%)
Complete and single-copy BUSCOs	813 (85.2%)
Complete and duplicated BUSCOs	11 (1.2%)
Fragmented BUSCOs	24 (2.5%)
Missing BUSCOs	106 (11.1%)
Eukaryota ODB BUSCO score (*n* = 255)	
Complete BUSCOs	223 (87.5%)
Number of genes	13,430
Number of protein-coding genes	12,933
Number of functionally annotated protein-coding genes (InterPro)	10,104 (78.1%)
RepeatModeler (de novo) bp masked	21.27%
RepeatMasker (*Ciona*) bp masked	0.73%
Consensus Repeat Masking bp masked	21.72%

Comparative genomics

Genomic datasets are reservoirs of rich, high-dimensional data that can be used to infer evolutionary relationships and molecular functions through the lens of comparative genomics. Yet, distilling persuasive evidence for these biological concepts remains challenging. To that end, we implemented a step-wise analytical approach to identify copy number variation patterns that are unique to botryllids (Fig. 3a–d). Pairwise comparison of protein-coding genes has identified patterns of genome evolution unique to the botryllid lineage (see *Materials and methods* for details). Interpreted within an evolutionary framework, we hypothesize that these botryllid-specific patterns are representative of botryllid-specific traits such as high regenerative capacity and coloniality, which thus provide a resource for experimentally disentangling these complex phenomena.

Gene family evolution plays an essential role in trait acquisition (Capra *et al.* 2010). We thus implemented the OrthoFinder algorithm to identify groups of homologous genes (orthogroups) from across the tunicate lineage (Fig. 3a). Although some orthogroups do include genes from all tested species, not all species will be represented in all orthogroups (Emms and Kelly 2019). From the eleven tunicate proteomes used in this analysis, 229,318 genes were assigned into 21,690 orthogroups which includes 88.6% of the original gene set. The relative size of each orthogroup varies considerably and the number of homologous genes in an orthogroup ranges from 2 to 447 (mean = 6.0, median = 9.4). Interestingly, 10.6% of all orthogroups represent single species; this suggests that novel or retained gene lineages have expanded in single species, supporting previous reports of rapid diversification in the tunicate lineage (Denoeud *et al.* 2010; Berna and Alvarez-Valin 2014; Jue *et al.* 2016).

The degree to which species overlap in orthogroups is useful for identifying genome scale duplication and contraction events



General Trends

B. violaceus	B. leachii	B. schlosseri	Clusters
↑	—	—	5, 11,13,14
—	↑	—	9, 8, 12*
—	—	↑	1
↑	↑	—	7, 16
↑	—	↑	10, 17
—	↑	↑	2, 3, 18*
↑	↑	↑	4*, 6, 15*

*could fit into a different group

KEGG Pathways

Cluster	CytokineReceptorIntxn	EGF	Hedgehog	Notch	RetinolMetab	TGFB	TNF	Wnt
5					1	1	3	
11	1	4	1	4	1		1	4
13						1		
14				1				
9								
8								
12				2			3	2
1	3		2	2	3		2	6
7	10	10	5	3	9	12	11	14
16	1	1		1	3	1		2
10	2	2	1	4	4	1	1	5
17		5	1	1	1		4	7
2								
3								
18		1					1	1
4	9	15	6	2	4	10	16	18
6	13	12	6	17	13	11	25	28
15					1			

Fig. 3. Comparative genomic analysis using hierarchical orthogroups reveals patterns of conservation and divergence in the botryllid lineage. a,b) HOGBs clustered using variation in gene copy number per species. A species by HOGB copy number matrix was clustered to identify groups with similar patterns of gene family evolution in botryllids. The appropriate number of clusters (18) was chosen using the silhouette method to account for variation between clusters. c) HOGB clusters binned by duplication pattern similarity. d) Number of *B. violaceus* genes in each cluster that are members of developmentally relevant pathways according to KEGG annotation. Clusters ordered by general trends depicted in (c).

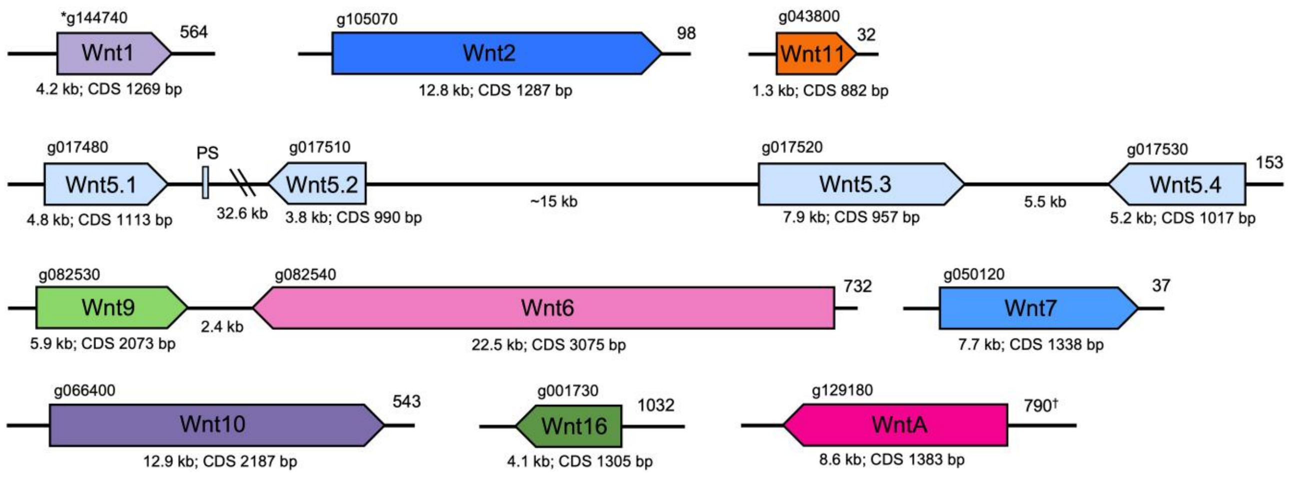


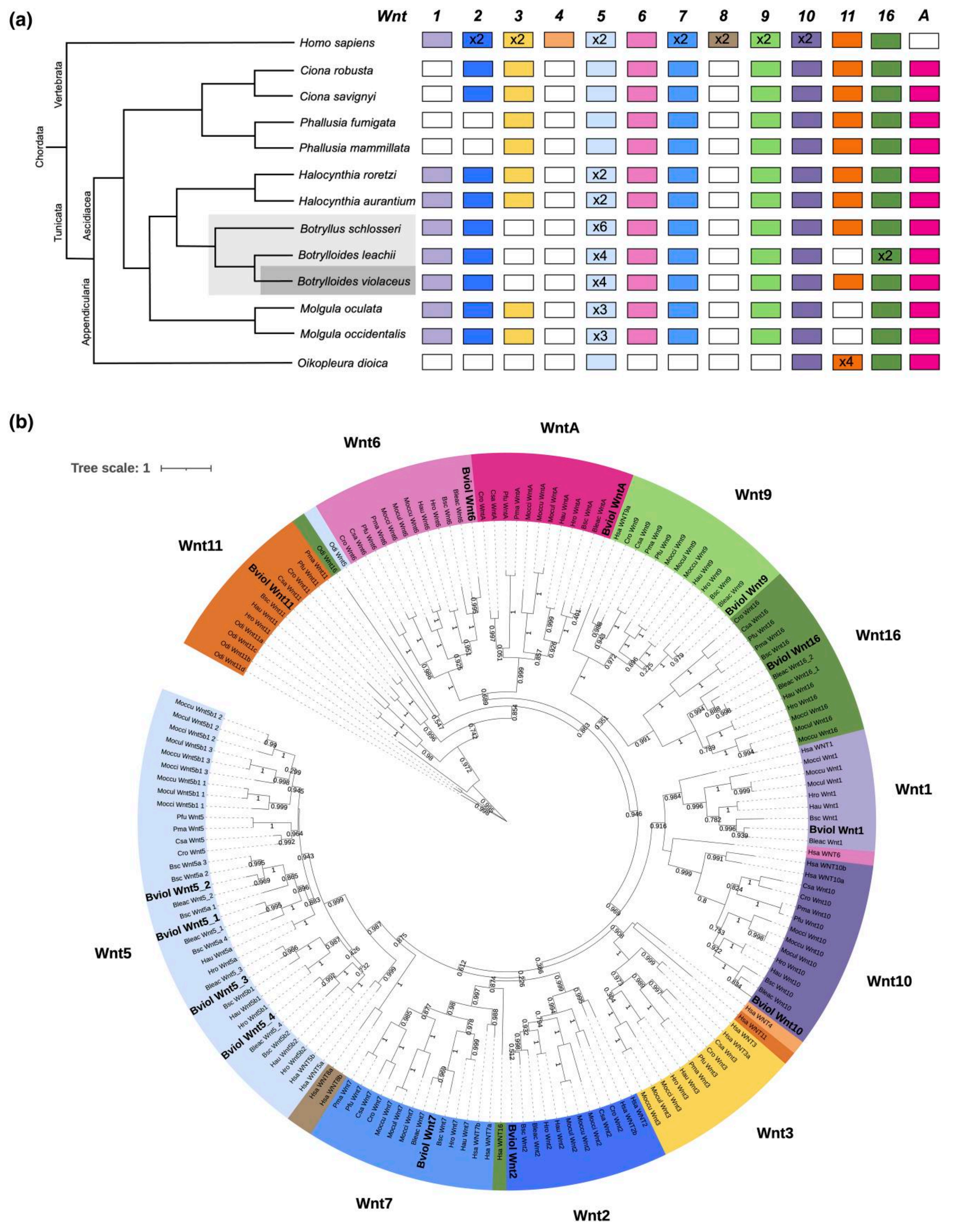
Fig. 4. Relative position and orientation of the Wnt genes in *B. violaceus*. Distances between genes and gene sizes are to scale, with the exception of Wnt11 which was slightly enlarged for clarity. PS indicates a Wnt5 pseudogene. Gene identification number from GenSAS annotation is on top. Numbers at rightmost position of fragments refer to specific contig or scaffold† in the hybrid assembly. Kb refers to genomic size; CDS bp refers to coding DNA sequence length; double-parallel lines indicate the intergenic region is >30 kb. *Wnt1 gene identification number is from GeneMark-ES gene model predictions.

(Emms and Kelly 2019). Pairwise correlation analysis of gene counts per orthogroup reveals the extent to which species show similar duplication and loss patterns (Supplementary Fig. S2). Indeed, we observe that *O. dioica* has substantially reduced orthogroup overlap and gene counts with other tunicate lineages (Supplementary Figs. S2 and S3) which is consistent with its dramatically compacted genome (Seo *et al.* 2001). In contrast, *B. schlosseri* retains greater orthogroup overlap and gene count correlations with other botryllids than with nonbotryllid tunicates (which are strikingly weak). These data support the hypothesis that secondary gene duplication events in *Botryllus* gave rise to an expanded genome after its divergence with the shared botryllid ancestor (Supplementary Figs. S2 and S3); however, they do not refute the alternative, though not mutually exclusive, hypothesis that the *B. schlosseri* genome expanded due to horizontal gene transfer (Voskoboynik *et al.* 2013).

Hierarchical orthogroups are comprised of the orthologs and inparalogs of a specific clade. Lineage-specific hierarchical orthogroups were identified for botryllids (HOGBs) (see *Materials and methods* for details; Fig. 3a). To dissect botryllid-specific patterns of copy number variations, we performed cluster analysis using the gene copy number in a HOGB for each species (Fig. 3a,b). The 18 clusters of HOGBs represent changes in copy number variation among botryllids (Fig. 3a); each cluster contains unique sets of HOGBs, and each cluster represents several HOGBs which in turn include varying numbers of homologous genes (Fig. 3b). Patterns of copy number variation can be classified into seven broad categories that communicate relative changes of variation both within botryllids and between botryllids and nonbotryllid tunicates (Fig. 3c). Generally, we observe that all botryllid genomes encode similar or greater HOGB gene copies than nonbotryllid tunicate genomes, as expected (Fig. 3a). However, the degree to which HOGB gene copies encoded in botryllid genomes differ from nonbotryllid tunicates varies in a species-dependent manner (Fig. 3c). For instance, the HOGB copy number of cluster 6 is mutually expanded in all botryllids relative to nonbotryllid tunicates (Fig. 3a). This pattern suggests that gene duplications in the botryllid ancestor that were then conserved in its descendants. Similarly, cluster 16 contains HOGBs that have expanded in *Botrylloides* spp. but remained largely unchanged in *B. schlosseri*, likely indicating expansion events in the *Botrylloides* ancestor (Fig. 3a). As *Botrylloides* spp. regenerate more readily than *B. schlosseri* (Brown *et al.* 2009; Nourizadeh *et al.* 2021), these patterns may highlight genetic components involved in regulation of whole-body regeneration. Overall, these data suggest that the botryllid lineage has undergone large-scale gene expansion events at multiple points in their evolutionary lineage.

In addition to overall patterns of gene losses and gains, functional annotations associated with binned clusters suggest how patterns may relate to phenotypes. For example, retinoic acid signaling has a well characterized role in regeneration and development (Rinkevich *et al.* 2007). Genes involved in synthesis and breakdown of retinoic acid (retinol metabolism) have been differentially lost or retained in the tunicate lineage (Blanchoud *et al.* 2018). KEGG enrichment analysis reveals that genes involved in retinol metabolism are enriched in cluster 16, representing HOGBs expanded in *Botrylloides* spp. (Supplementary Fig. S4). In addition, four paralogs of CYP26A, a P450-related gene involved in intracellular breakdown of retinoic acid, were found in the *B. violaceus* assembly and previously reported in *B. leachii* (Blanchoud *et al.* 2018). Interesting, one CYP26A ortholog found in cluster 7 was not represented in *B. schlosseri* but was present in both *Botrylloides* species, suggesting either a gene loss or gain event with potential implication for regeneration.

Other developmental signaling pathways are also important for coordination and regulation of regeneration. For example, HDAC2/3, which are effector proteins involved in Notch signaling regulation (KEGG: hsa04330), were identified in our KEGG search (Fig. 3d). HDAC2 is represented in all three botryllids (cluster 6) while *B. violaceus* HDAC3 is represented in a *Botrylloides*-enriched group (cluster 7). Inhibition of HDAC suppresses WBR in *B. leachii* (Zondag *et al.* 2019), and this mechanism is likely conserved in *B. violaceus*, although experimental testing is necessary to confirm. The copy number of *B. violaceus* genes putatively involved in Wnt and TNFs signaling pathways was notably high in clusters generally associated with clusters expanded in botryllids relative to nonbotryllids (clusters 4, 6; Fig. 3a,d). TNFs play a major role in innate immunity, inflammation, and regulation of cell death (Webster and Vucic 2020). Ascidians are useful invertebrate models of allorecognition and the evolution of the immune system in chordates (Voskoboynik *et al.* 2013; Franchi and



Wnt	1	2	3	4	5	6	7	8	9	10	11	16	A
Homo sapiens	■	x2	x2	■	x2	■	x2	x2	x2	x2	■	■	□
Ciona robusta	□	■	■	□	■	■	■	□	■	■	■	■	■
Ciona savignyi	□	■	■	□	■	■	■	□	■	■	■	■	■
Phallusia fumigata	□	□	■	□	■	■	■	□	■	■	■	■	■
Phallusia mammillata	□	□	■	□	■	■	■	□	■	■	■	■	■
Halocynthia roretzi	■	■	■	□	x2	■	■	□	■	■	■	■	■
Halocynthia aurantium	■	■	■	□	x2	■	■	□	■	■	■	■	■
Botryllus schlosseri	■	■	□	□	x6	■	■	□	■	■	■	■	■
Botrylloides leachii	■	■	□	□	x4	■	■	□	■	■	□	x2	■
Botrylloides violaceus	■	■	□	□	x4	■	■	□	■	■	■	■	■
Molgula oculata	■	■	■	□	x3	■	■	□	■	■	□	■	■
Molgula occidentalis	■	■	■	□	x3	■	■	□	■	■	□	■	■
Oikopleura dioica	□	□	□	□	■	□	□	□	□	■	x4	■	■

Fig. 5. Wnt gene family analysis. a) Summary of Wnt repertoire for comparative analysis of selected tunicates and humans. Coloring of the boxes indicates the presence of the indicated Wnt subfamily, while a white box indicates the absence. Total copy number of the gene is provided in the respective box. Botryllids are highlighted in grey on the phylogeny. b) Maximum likelihood phylogeny of tunicate Wnt protein sequences. Branch support values derived from SH-like aLRT method. Species represented: *Botrylloides violaceus* (Bviol), *Botrylloides leachii* (Bleac), *Botryllus schlosseri* (Bsc), *Ciona savignyi* (Csa), *Ciona robusta* (Cro), *Halocynthia roretzi* (Hro), *Halocynthia aurantium* (Hau), *Mogula occulta* (Moccu), *Mogula oculata* (Mocul), *Mogula occidentalis* (Mocci), *Phallusia fumigata* (Pfu), *Phallusia mammillata* (Pma), *Oikopleura dioica* (Odi), and *Homo sapiens* (Hsa). *Oikopleura dioica* Wnt10 was removed from the Wnt10 cluster due to long branch length in order to increase visibility of other leaves. Scale bar represents amino acid substitutions.

Ballarin 2017). For example, over 10 orthologs of the BIRC2 gene, which negatively interacts with TNF signaling cascades, were found in the *B. violaceus* genome. These genes putatively function as inhibitors of apoptosis and several inhibitor of apoptosis genes have been implicated in botryllid regeneration and torpor (Rosner *et al.* 2019).

As HOGB clusters are unique to the botryllid lineage, this analysis is expected to be more sensitive to botryllid-lineage gene duplication events than it is to gene-loss events; gene-loss events that occurred early after ancestral botryllid diversification may require greater genomic resolution to detect (i.e. sampling more botryllid genomes). Furthermore, gene loss inferences may be confounded by possibly incomplete genome coverage. However, gene-loss events can be inferred when expansions are conserved between sister genera but not within the same genus. Cluster 10 is comprised of HOGB clusters with expanded repertoires in both *B. violaceus* and *B. schlosseri*, but not in *B. leachii* (Fig. 3c). Therefore, one may infer that primary duplication events in the botryllid ancestor followed by secondary gene loss events in the *B. leachii* ancestor may have occurred. The biological significance remains elusive, although KEGG enrichment analysis reveals that genes associated with neuronal development or function may facilitate some unknown trait that may have been lost in *B. leachii* (Supplementary Fig. S4). Thus, we provide evidence that comparative genomics analysis of closely related species allows low-resolution inference of their shared ancestor's genomic content and that interpretation within an evolutionary context is useful in generating hypotheses for future experimental testing.

Wnt family analysis

Gene predictions from the hybrid assembly annotation and additional analysis allowed us to characterize the repertoire of Wnt genes found in *B. violaceus*. Characterizing Wnt genes in *B. violaceus* demonstrates the functionality of our genome and provides genetic targets for investigating the unique regenerative abilities of this organism. Comprehensive analysis of Wnt genes in the *B. violaceus* genome indicates the presence of 10 Wnt subfamilies (Fig. 4). Our final Wnt gene repertoire is corroborated by similar patterns in other tunicate genomes (Fig. 5a) (Somorjai *et al.* 2018; Martí-Solans *et al.* 2021). The initial BLAST annotations identified most, but not all, of these Wnt genes, emphasizing the necessity of additional manual annotation methods.

Our workflow for manually confirming the Wnt gene subfamilies included protein multiple-sequence alignments, phylogenetic relationship analyses (Fig. 5b), BLAST results, and occasionally protein modeling. This workflow was performed on the three represented botryllids in order to confirm and compare their Wnt repertoires. Analysis indicates that the botryllids share very similar Wnt repertoires, including extensive duplication of Wnt5. Interestingly, the botryllids lack Wnt3 that other ascidians retain. However, consistent with other ascidians, botryllids lack Wnt4 and Wnt8, but do have a WntA (Fig. 5a).

Particular attention to manual annotation was required for Wnts whose subfamily was ambiguous upon initial analysis, including Wnt1, 11, and A. As discussed with the annotation, Wnt1 was identified only in the GeneMark-ES predicted gene models. Wnt11 annotated as multiple different Wnt subfamilies; it is likely that highly conserved sequences between Wnts may have led to different subfamily categorizations when using different databases. The presence of Wnt11 in *B. violaceus* confirms the variability of Wnt11 in tunicates: absent in *B. leachii*, *M. occulata*, and *M. occidentalis*, while present in *B. schlosseri* and other surveyed tunicates (Somorjai *et al.* 2018; Martí-Solans *et al.* 2021). Wnt11 activates both the traditionally defined canonical and noncanonical Wnt signaling pathways; studies in *Xenopus* have found that the canonical pathway required for axis formation is activated by Wnt11, an interesting implication for further research into ascidian development (Tao *et al.* 2005; Lemaire *et al.* 2008). Initial databases used lacked the classification of WntA, so the sequence first appeared in manual annotation as a Wnt4-like gene. Phylogenetic analysis represented by Fig. 5b contained well-annotated WntA protein sequences obtained from recent literature (Martí-Solans *et al.* 2021), confirming the presence of WntA and lack of Wnt4, consistent with other tunicates.

We identified only one Wnt gene with multiple copies in the *B. violaceus* genome: four distinct Wnt5 genes (Fig. 4). Wnt5 gene duplications have previously been found in the greater order to which botryllids belong (Somorjai *et al.* 2018). The *B. violaceus* Wnt 5 genes are found in tandem, mirroring the Wnt5 gene architecture of *B. leachii* (Blanchoud *et al.* 2018; Somorjai *et al.* 2018). In addition, a potential Wnt 5 pseudogene was found between Wnt5.1 and Wnt5.2, which was similarly identified in *B. leachii*; this classification as a pseudogene is due to a smaller gene and coding sequence length when compared to other Wnt 5 genes. The parallels in the Wnt5 cluster likely indicate that the duplications and pseudogene developed in the last common ancestor of these sister species. Pseudogene conservation across species has been previously observed (Mahmudi *et al.* 2015). Additionally, our phylogeny supports independent Wnt 5 duplications after separation of the *Botrylloides* and *Molgula* genera; this pattern was previously suggested for *Molgula* and *Halocynthia* (Somorjai *et al.* 2018). Some Wnt 5 duplications are shared between the more closely related *Botrylloides* and *Halocynthia* genera, but additional duplications appear to have occurred in the botryllids (Fig. 5a). The Wnt5 gene expansion is absent in other invertebrate chordates and may be related to the unique regenerative capabilities of ascidians (Somorjai *et al.* 2018). Wnt5 activates the noncanonical Wnt signaling pathways, along with Wnt4, Wnt6, and Wnt11 (Schubert and Holland 2013; Komiya and Habas 2008), and noncanonical Wnt signaling has been implicated in tissue regeneration (Hu *et al.* 2008; Zondag *et al.* 2016). In addition, the extensive Wnt5 gene duplications may partially substitute for the functional role of Wnt4 in *B. violaceus* and *B. leachii*, and for the functional role of Wnt11 in *B. leachii* in certain contexts (Croce *et al.* 2006; Somorjai *et al.* 2018).

Conclusions

Nonmodel organisms often display interesting biological processes but lack the availability of high-quality reference genomes. Our study provides a quality hybrid genome assembly and annotation for a nonmodel organism, *B. violaceus*. This genome adds to a growing dataset of ascidian genomes, facilitating comparative genomic analyses that can provide testable hypotheses about gene and genome evolution, as well as about the genetic changes underlying processes such as regeneration.

Data availability

NCBI BioProject and BioSample identifiers are PRJNA875143 and SAMN30603932, respectively. Illumina, Inc. and Oxford Nanopore Technologies sequencing data are available on NCBI SRA, SRX17396406, and SRX17396405, respectively. *B. violaceus* genome assembly is available on NCBI, JASFYC000000000. Supporting scripts and other data generated from analysis, including annotation files, are available on GitHub at https://github.com/calpoly-bioinf/botrylloides.

Supplemental material is available at G3 online.

Acknowledgments

We thank Dr. David Keeling for supplying images of *Botrylloides violaceus* colonies and Dr. Dena Block for her suggestions regarding DNA extraction.

Funding

This research was funded in part by the Frost Undergraduate Research Fund, the Warren J. Baker Endowment, and the Robert D. Koob Endowments at Cal Poly. Additionally, high performance computing resources were funded by the NSF XSEDE startup.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

JTS: conceptualization, experimental optimization, whole-genome sequencing and assembly, comparative genomics analysis. CLA: genome assembly and annotation, comparative genomics analysis, Wnt family analysis. CAJ and SW: Wnt family analysis. PA: computational resources and guidance. ELK conceptualization, sample acquisition, supervision. JMD: conceptualization, supervision. All authors contributed to writing and editing the final manuscript.

Literature cited

Alié A, Hiebert LS, Simion P, Scelzo M, Prünster MM, Lotito S, Delsuc F, Douzery EJ, Dantec C, Lemaire P, *et al.* Convergent acquisition of nonembryonic development in styelid ascidians. Mol Biol Evol. 2018. 35:1728–1743. doi:10.1093/molbev/msy068

Altenhoff AM, Gil M, Gonnet GH, Dessimoz C. Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS One. 2013;8:e53786. doi:10.1371/journal.pone.0053786

Anthony CC, Robbins DJ, Ahmed Y, Lee E. Nuclear regulation of Wnt/β-catenin signaling: it's a complex situation. Genes. 2020;11:886. doi:10.3390/genes11080886

Bely AE. Evolutionary loss of animal regeneration: pattern and process. Integr Comp Biol. 2010;50:515–527. doi:10.1093/icb/icq118

Berna L, Alvarez-Valin F. Evolutionary genomics of fast evolving tunicates. Genome Biol Evol. 2014;6:1724–1738. doi:10.1093/gbe/evu122

Berrill N. The development of the bud in *Botryllus*. Biol Bull. 1941;80: 169–184. doi:10.2307/1537595

Blanchoud S, Rutherford K, Zondag L, Gemmell NJ, Wilson MJ. De novo draft assembly of the *Botrylloides leachii* genome provides further insight into tunicate evolution. Sci Rep. 2018;8:1–18. doi: 10.1038/s41598-018-23749-w

Bliznina A, Masunaga A, Mansfield MJ, Tan Y, Liu AW, West C, Rustagi T, Chien HC, Kumar S, Pichon J, *et al.* Telomere-to-telomere assembly of the genome of an individual *Oikopleura dioica* from okinawa using nanopore-based sequencing. BMC Genomics. 2021;22:1–18. doi:10.1186/s12864-021-07512-6

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

Bonini NM, Berger SL. The sustained impact of model organisms—in genetics and epigenetics. Genetics. 2017;205:1–4. doi:10.1534/genetics.116.187864

Borisenko I, Bolshakov FV, Ereskovsky A, Lavrov AI. Expression of Wnt and TGF-beta pathway components during whole-body regeneration from cell aggregates in demosponge *Halisarca dujardinii*. Genes. 2021;12:944. doi:10.3390/genes12060944

Brown FD, Keeling EL, Le AD, Swalla BJ. Whole body regeneration in a colonial ascidian, *Botrylloides violaceus*. J Exp Zool B: Mol Dev Evol. 2009;312:885–900. doi:10.1002/jez.b.v312b:8

Brozovic M, Dantec C, Dardaillon J, Dauga D, Faure E, Gineste M, Louis A, Naville M, Nitta KR, Piette J, *et al.* Aniseed 2017: extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets. Nucleic Acids Res. 2018;46:D718–D725. doi:10.1093/nar/gkx1108

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. Blast+: architecture and applications. BMC Bioinform. 2009;10:1–9. doi:10.1186/1471-2105-10-421

Capra JA, Pollard KS, Singh M. Novel genes exhibit distinct patterns of function acquisition and network integration. Genome Biol. 2010;11:1–16. doi:10.1186/gb-2010-11-6-306

Clevers H, Loh KM, Nusse R. An integral program for tissue renewal and regeneration: Wnt signaling and stem cell control. Science. 2014;346:1–7. doi:10.1126/science.1248012

Croce JC, Wu SY, Byrum C, Xu R, Duloquin L, Wikramanayake AH, Gache C, McClay DR. A genome-wide survey of the evolutionarily conserved Wnt pathways in the sea urchin *Strongylocentrotus purpuratus*. Dev Biol. 2006;300:121–131. doi:10.1016/j.ydbio.2006.08.045

da Fonseca RR, Albrechtsen A, Themudo GE, Ramos-Madrigal J, Sibbesen JA, Maretty L, Zepeda-Mendoza ML, Campos PF, Heller R, Pereira RJ. Next-generation biology: sequencing and data analysis approaches for non-model organisms. Mar Genom. 2016;30: 3–13. doi:10.1016/j.margen.2016.04.012

da Fonseca RR, Couto A, Machado AM, Brejova B, Albertin CB, Silva F, Gardner P, Baril T, Hayward A, Campos A, *et al.* A draft genome sequence of the elusive giant squid, *Architeuthis dux*. GigaScience. 2020;9:giz152. doi:10.1093/gigascience/giz152

Darnet S, Dragalzew AC, Amaral DB, Sousa JF, Thompson AW, Cass AN, Lorena J, Pires ES, Costa CM, Sousa MP, *et al.* Deep evolutionary origin of limb and fin regeneration. Proc Natl Acad Sci USA. 2019;116:15106–15115. doi:10.1073/pnas.1900475116

DeBiasse MB, Colgan WN, Harris L, Davidson B, Ryan JF. Inferring tunicate relationships and the evolution of the tunicate Hox cluster with the genome of *Corella inflata*. Genome Biol Evol. 2020;12: 948–964. doi:10.1093/gbe/evaa060

De Coster W, D'hert S, Schultz DT, Cruts M, Van Broeckhoven C. Nanopack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34:2666–2669. doi:10.1093/bioinformatics/bty149

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, *et al.* The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science. 2002;298:2157–2167. doi:10.1126/science.1080049

Delgado-Coello B. Liver regeneration observed across the different classes of vertebrates from an evolutionary perspective. Heliyon. 2021;7:e06449. doi:10.1016/j.heliyon.2021.e06449

Delsuc F, Brinkmann H, Chourrout D, Philippe H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature. 2006;439:965–968. doi:10.1038/nature04336

Denoeud F, Henriet S, Mungpakdee S, Aury JM, Da Silva C, Brinkmann H, Mikhaleva J, Olsen LC, Jubin C, Cañestro C, *et al.* Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science. 2010;330:1381–1385. doi:10.1126/science.1194167

Edgar A, Mitchell DG, Martindale MQ. Whole-body regeneration in the lobate ctenophore *mnemiopsis leidyi*. Genes. 2021;12:867. doi: 10.3390/genes12060867

Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238. doi:10.1186/s13059-019-1832-y

Etherington GJ, Heavens D, Baker D, Lister A, McNelly R, Garcia G, Clavijo B, Macaulay I, Haerty W, Di Palma F. Sequencing smart: de novo sequencing and assembly approaches for a non-model mammal. GigaScience. 2020;9:giaa045. doi:10.1093/gigascience/giaa045

Ferrario C, Sugni M, Somorjai IM, Ballarin L. Beyond adult stem cells: dedifferentiation as a unifying mechanism underlying regeneration in invertebrate deuterostomes. Front Cell Dev Biol. 2020;8:587320. doi:10.3389/fcell.2020.587320

Franchi N, Ballarin L. Immunity in protochordates: the tunicate perspective. Front Immunol. 2017;8:674. doi:10.3389/fimmu.2017.00674

Garcia AL, Udeh A, Kalahasty K, Hackam AS. A growing field: the regulation of axonal regeneration by Wnt signaling. Neural Regen Res. 2018;13:43. doi:10.4103/1673-5374.224359

Gordon T, Upadhyay AK, Manni L, Huchon D, Shenkar N. And then there were three…: extreme regeneration ability of the solitary chordate *Polycarpa mytiligera*. Front Cell Dev Biol. 2021;9:793. doi:10.3389/fcell.2021.652466

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. Syst Biol. 2010;59:307–321. doi:10.1093/sysbio/syq010

Gurevich A, Saveliev V, Vyahhi N, Tesler G. Quast: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–1075. doi:10.1093/bioinformatics/btt086

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. Genome Biol. 2008;9:1–22. doi:10.1186/gb-2008-9-1-r7

Hu DJK, Yun J, Elstrott J, Jasper H. Non-canonical Wnt signaling promotes directed migration of intestinal stem cells to sites of injury. Nat Commun. 2008;12:7150. doi:10.1038/s41467-021-27384-4

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, Bork P. Fast genome-wide functional annotation through orthology assignment by eggnog-mapper. Mol Biol Evol. 2017;34:2115–2122. doi:10.1093/molbev/msx148

Humann JL, Lee T, Ficklin S, Main D. Structural and functional annotation of eukaryotic genomes with GenSAS. Methods Mol Biol. 2019;1962:29–51. doi:10.1007/978-1-4939-9173-0_3

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, *et al.* Abyss 2.0: resource-efficient assembly of large genomes using a bloom filter. Genome Res. 2017;27:768–777. doi:10.1101/gr.214346.116

Jiang JB, Quattrini AM, Francis WR, Ryan JF, Rodriguez E, McFadden CS. A hybrid de novo assembly of the sea pansy (*Renilla muelleri*) genome. GigaScience. 2019;8:giz026. doi:10.1093/gigascience/giz026

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, *et al.* Interproscan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–1240. doi:10.1093/bioinformatics/btu031

Jue NK, Batta-Lona PG, Trusiak S, Obergfell C, Bucklin A, O'Neill MJ, O'Neill RJ. Rapid evolutionary rates and unique genomic signatures discovered in the first reference genome for the Southern Ocean Salp, *Salpa thompsoni* (Urochordata, Thaliacea). Genome Biol Evol. 2016;8:3171–3186. doi:10.1093/gbe/evw215

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30. doi:10.1093/nar/28.1.27

Kassmer SH, Langenbacher AD, De Tomaso AW. Integrin-alpha-6+ candidate stem cells are responsible for whole body regeneration in the invertebrate chordate *Botrylloides diegensis*. Nat Commun. 2020;11:1–11. doi:10.1038/s41467-020-18288-w

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–546. doi:10.1038/s41587-019-0072-8

Komiya Y, Habas R. Wnt signal transduction pathways. Organogenesis. 2008;4:68–75. doi:10.4161/org.4.2.5851

Korf I. Gene finding in novel genomes. BMC Bioinform. 2004;5:1–9. doi:10.1186/1471-2105-5-59

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. Orthodb v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2019;47:D807–D811. doi:10.1093/nar/gky1053

Lai AG, Aboobaker AA. Evoregen in animals: time to uncover deep conservation or convergence of adult stem cell evolution and regenerative processes. Dev Biol. 2018;433:118–131. doi:10.1016/j.ydbio.2017.10.010

Lemaire P. Evolutionary crossroads in developmental biology: the tunicates. Development. 2011;138:2143–2152. doi:10.1242/dev.048975

Lemaire P, Smith WC, Nishida H. Ascidians and the plasticity of the chordate developmental program. Curr Biol. 2008;18:620–631. doi:10.1016/j.cub.2008.05.039

Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen-Boulakia S, Gascuel O. Ngphylogeny. fr: new generation phylogenetic services for non-specialists. Nucleic Acids Res. 2019;47:W260–W265. doi:10.1093/nar/gkz303

Letunic I, Bork P. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49:W293–W296. doi:10.1093/nar/gkab301

Leucht P, Lee S, Yim N. Wnt signaling and bone regeneration: can't have one without the other. Biomaterials. 2019;196:46–50. doi:10.1016/j.biomaterials.2018.03.029

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005;33:6494–6506. doi:10.1093/nar/gki937

Mahmudi O, Sennblad B, Arvestad L, Nowick K, Lagergren J. Gene-pseudogene evolution: a probabilistic approach. BMC Genomics. 2015;16:S12. doi:10.1186/1471-2164-16-S10-S12

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. Busco update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38:4647–4654. doi:10.1093/molbev/msab199

Martí-Solans J, Godoy-Marín H, Diaz-Gracia M, Onuma TA, Nishida H, Albalat R, Cañestro C. Massive gene loss and function shuffling in appendicularians stretch the boundaries of chordate Wnt family evolution. Front Cell Dev Biol. 2021;9:700827. doi:10.3389/fcell.2021.700827

Mokalled MH, Poss KD. A regeneration toolkit. Dev Cell. 2018;47:267–280. doi:10.1016/j.devcel.2018.10.015

Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, Lee J, Chu C, Lin C, Džakula Ž, *et al.* A hybrid approach for de novo human genome sequence assembly and phasing. Nat Methods. 2016;13:587–590. doi:10.1038/nmeth.3865

Nayak L, Bhattacharyya NP, De RK. Wnt signal transduction pathways: modules, development and evolution. BMC Syst Biol. 2016;10:44. doi:10.1186/s12918-016-0299-7

Niehrs C. The complex world of Wnt receptor signalling. Nat Rev Mol Cell Biol. 2012;13:767–779. doi:10.1038/nrm3470

Nourizadeh S, Kassmer S, Rodriguez D, Hiebert LS, De Tomaso AW. Whole body regeneration and developmental competition in two botryllid ascidians. EvoDevo. 2021;12:15. doi:10.1186/s13227-021-00185-y

Nusse R, Clevers H. Wnt/β-catenin signaling, disease, and emerging therapeutic modalities. Cell. 2017;169:985–999. doi:10.1016/j.cell.2017.05.016

Reddien PW, Alvarado AS. Fundamentals of planarian regeneration. Annu Rev Cell Dev Biol. 2004;20:725–757. doi:10.1146/cellbio.2004.20.issue-1

Reddy PC, Gungi A, Unni M. Cellular and molecular mechanisms of hydra regeneration. Results Probl Cell Differ. 2019;68:259–290. doi:10.1007/978-3-030-23459-1_12

Rinkevich Y, Paz G, Rinkevich B, Reshef R. Systemic bud induction and retinoic acid signaling underlie whole body regeneration in the urochordate *Botrylloides leachi*. PLoS Biol. 2007;5:e71. doi:10.1371/journal.pbio.0050071

Rinkevich B, Shlemberg Z, Fishelson L. Whole-body protochordate regeneration from totipotent blood cells. Proc Natl Acad Sci USA. 1995;92:7695–7699. doi:10.1073/pnas.92.17.7695

Rosental B, Kowarsky M, Seita J, Corey DM, Ishizuka KJ, Palmeri KJ, Chen SY, Sinha R, Okamoto J, Mantalas G, *et al.* Complex mammalian-like haematopoietic system found in a colonial chordate. Nature. 2018;564:425–429. doi:10.1038/s41586-018-0783-x

Rosner A, Kravchenko O, Rinkevich B. IAP genes partake weighty roles in the astogeny and whole body regeneration in the colonial urochordate *Botryllus schlosseri*. Dev Biol. 2019;448:320–341. doi:10.1016/j.ydbio.2018.10.015

Satou Y, Nakamura R, Yu D, Yoshida R, Hamada M, Fujie M, Hisata K, Takeda H, Satoh N. A nearly complete genome of *Ciona intestinalis* type a (*C. robusta*) reveals the contribution of inversion to chromosomal evolution in the genus *Ciona*. Genome Biol Evol. 2019;11:3144–3157. doi:10.1093/gbe/evz228

Schubert M, Holland LZ. The Wnt gene family and the evolutionary conservation of Wnt expression. Madame Curie Bioscience Database [Internet]. 2013.

Seo HC, Kube M, Edvardsen RB, Jensen MF, Beck A, Spriet E, Gorsky G, Thompson EM, Lehrach H, Reinhardt R, *et al.* Miniature genome in the marine chordate *Oikopleura dioica*. Science. 2001;294:2506–2506. doi:10.1126/science.294.5551.2506

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol. 2011;7:539. doi:10.1038/msb.2011.75

Somorjai IM, Martí-Solans J, Diaz-Gracia M, Nishida H, Imai KS, Escrivà H, Cañestro C, Albalat R. Wnt evolution and function shuffling in liberal and conservative chordate genomes. Genome Biol. 2018;19:1–17. doi:10.1186/s13059-018-1468-3

Tan MH, Austin CM, Hammer MP, Lee YP, Croft LJ, Gan HM. Finding nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (*Amphiprion ocellaris*) genome assembly. GigaScience. 2018;7:137. doi:10.1093/gigascience/gix137

Tao Q, Yokota C, Puck H, Kofron M, Birsoy B, Yan D, Asashima M, Wylie CC, Lin X, Heasman J. Maternal Wnt11 activates the canonical Wnt signaling pathway required for axis formation in *Xenopus* embryos. Cell. 2005;120:857–871. doi:10.1016/j.cell.2005.01.013

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

Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–192. doi:10.1093/bib/bbs017

Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, Brown SD. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics. 2014;30:2709–2716. doi:10.1093/bioinformatics/btu391

Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B, Fan HC, Mantalas GL, Palmeri KJ, *et al.* The genome sequence of the colonial chordate, *Botryllus schlosseri*. elife. 2013;2:e00569. doi:10.7554/eLife.00569

Voskoboynik A, Simon-Blecher N, Soen Y, Rinkevich B, De Tomaso AW, Ishizuka KJ, Weissman IL. Striving for normality: whole body regeneration through a series of abnormal generations. FASEB J. 2007;21:1335–1344. doi:10.1096/fsb2.v21.7

Wallberg A, Bunikis I, Pettersson OV, Mosbech MB, Childers AK, Evans JD, Mikheyev AS, Robertson HM, Robinson GE, Webster MT. A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. BMC Genomics. 2019;20:1–19. doi:10.1186/s12864-019-5642-0

Watanabe H, Newberry AT. Budding by oozooids in the polystyelid ascidian *Metandrocarpa taylori* Huntsman. J Morphol. 1976;148:161–176. doi:10.1002/(ISSN)1097-4687

Webster JD, Vucic D. The balance of TNF mediated pathways regulates inflammatory cell death signaling in healthy and diseased tissues. Front Cell Dev Biol. 2020;8:365. doi:10.3389/fcell.2020.00365

Whitacre LK, Hoff JL, Schnabel RD, Albarella S, Ciotola F, Peretti V, Strozzi F, Ferrandi C, Ramunno L, Sonstegard TS, *et al.* Elucidating the genetic basis of an oligogenic birth defect using whole genome sequence data in a non-model organism, *Bubalus bubalis*. Sci Rep. 2017;7:1–9. doi:10.1038/srep39719

Willert K, Nusse R. Wnt proteins. Cold Spring Harb Perspect Biol. 2012;4:a007864. doi:10.1101/cshperspect.a007864

Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:1–12. doi:10.1186/gb-2014-15-3-r46

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, *et al.* ClusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation. 2021;2:100141. doi:10.1016/j.xinn.2021.100141

Xing Y, Liu Y, Zhang Q, Nie X, Sun Y, Zhang Z, Li H, Fang K, Wang G, Huang H, *et al.* Hybrid de novo genome assembly of Chinese chestnut (*Castanea mollissima*). GigaScience. 2019;8:giz112. doi:10.1093/gigascience/giz112

Yokoyama H. Initiation of limb regeneration: the critical steps for regenerative capacity. Dev Growth Differ. 2008;50:13–22. doi:10.1111/dgd.2008.50.issue-1

Zeng L, Jacobs MW, Swalla BJ. Coloniality has evolved once in Stolidobranch ascidians. Integr Comp Biol. 2006;46:255–268. (doi:10.1093/icb/icj035

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29:2669–2677. doi:10.1093/bioinformatics/btt476

Zimin AV, Puiu D, Luo MC, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. 2017;27:787–792. doi:10.1101/gr.213405.116

Zondag L, Clarke RM, Wilson MJ. Histone deacetylase activity is required for *Botrylloides leachii* whole-body regeneration. J Exp Biol. 2019;222:jeb203620. doi:10.1242/jeb.203620

Zondag LE, Rutherford K, Gemmell NJ, Wilson MJ. Uncovering the pathways underlying whole body regeneration in a chordate model, *Botrylloides leachi* using de novo transcriptome analysis. BMC Genomics. 2016;17:1–11. doi:10.1186/s12864-016-2435-6

Zullo L, Bozzo M, Daya A, Di Clemente A, Mancini FP, Megighian A, Nesher N, Röttinger E, Shomrat T, Tiozzo S, *et al.* The diversity of muscles and their regenerative potential across animals. Cells. 2020;9:1925. doi:10.3390/cells9091925

Editor: N. Whiteman