PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Caurcel C, Laetsch DR, Challis R, Kumar S, Gharbi K, Blaxter M. 2021 MolluscDB: a genome and transcriptome database for molluscs. *Phil. Trans. R. Soc. B* **376**: 20200157.
https://doi.org/10.1098/rstb.2020.0157

Accepted: 5 October 2020

One contribution of 15 to a Theo Murphy meeting issue 'Molluscan genomics: broad insights and future directions for a neglected phylum'.

Subject Areas:
genomics, evolution, bioinformatics

Keywords:
molluscs, database, genome, transcriptome, protein families, shell matrix proteins

Author for correspondence:
Mark Blaxter
e-mail: mb35@sanger.ac.uk

†Present address: The Earlham Institute, Norwich Research Park, Norwich, Norfolk NR4 7UZ, UK.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5324907.

THE ROYAL SOCIETY PUBLISHING

MolluscDB: a genome and transcriptome database for molluscs

Carlos Caurcel[1], Dominik R. Laetsch[1], Richard Challis[1,3], Sujai Kumar[1,3], Karim Gharbi[1,2,†] and Mark Blaxter[1,3]

[1]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK
[2]Edinburgh Genomics, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK
[3]Tree of Life Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

CC, 0000-0003-3836-1402; DRL, 0000-0001-7887-0186; RC, 0000-0002-3502-1122; SK, 0000-0001-5902-6641; KG, 0000-0003-1092-4488; MB, 0000-0003-2861-949X

As sequencing becomes more accessible and affordable, the analysis of genomic and transcriptomic data has become a cornerstone of many research initiatives. Communities with a focus on particular taxa or ecosystems need solutions capable of aggregating genomic resources and serving them in a standardized and analysis-friendly manner. Taxon-focussed resources can be more flexible in addressing the needs of a research community than can universal or general databases. Here, we present MolluscDB, a genome and transcriptome database for molluscs. MolluscDB offers a rich ecosystem of tools, including an Ensembl browser, a BLAST server for homology searches and an HTTP server from which any dataset present in the database can be downloaded. To demonstrate the utility of the database and verify the quality of its data, we imported data from assembled genomes and transcriptomes of 22 species, estimated the phylogeny of Mollusca using single-copy orthologues, explored patterns of gene family size change and interrogated the data for biomineralization-associated enzymes and shell matrix proteins. MolluscDB provides an easy-to-use and openly accessible data resource for the research community.

This article is part of the Theo Murphy meeting issue 'Molluscan genomics: broad insights and future directions for a neglected phylum'.

1. Introduction

Molluscs are important members of many terrestrial and aquatic ecosystems, where they can be dominant herbivores or important predators. They are important indicators of ecosystem health and are highly responsive to environmental change. Molluscs are also major food sources for humans, and understanding their biology is important for improving shellfish productivity. The mollusc shell is a product of active biomineralization, the enzymology of which is of interest to synthetic biology, aquaculture and climate science. Improved understanding of all of these issues and questions can be gained through the application of genomic and post-genomic methodologies to target species of molluscs, and through comparison of mollusc genes and genomes.

Over the past years, sequencing costs have decreased significantly. As a result, sequencing projects are no longer limited to specialized centres or large consortia [1]. This has favoured the study of non-model organisms such as molluscs, for which the number of sequencing-related publications is rapidly growing [2]. During the publication process, raw data and draft genomes are submitted to public databases. However, other analysis outputs such as gene predictions or functional annotations are often not formally submitted, which hinders further analyses by other researchers. Exceptionally, research groups make their results publicly available through their own websites, but such

resources can be hard to maintain, tend to increase format inconsistency between projects and do not offer a stable environment for the exploitation of the data. The situation is worse for species for which only transcriptome data are available, as there is no widely accepted route for submission of transcriptome assembly and annotation data for reuse. Usually, transcriptome data are only available in the form of raw reads, and reuse of these data often requires reassembly and reanalysis, divorcing secondary use from the primary publication [3,4].

While pan-taxonomic databases such as Ensembl [5] host and maintain data in a reliable way and provide a diverse set of tools for data exploration, they cannot accommodate most datasets generated in genomic research owing to strict quality inclusion criteria. For model organisms, data are hosted in taxon-specific databases, such as FlyBase [6] or WormBase [7]. These offer a feature-rich presentation and analysis ecosystem that can focus on meeting the specific needs of their research communities. Owing to their multiple advantages, taxon-specific databases have also been developed for non-model organisms, such as VectorBase [8], Avianbase [9] and Lepbase [10]. This tiered system of databases, from Tier 1 (local, often private, single-species or single-analysis resources) through Tier 2 (multi-species databases) and Tier 3 (pan-species databases such as Ensembl) [11] delivers to the needs of the science community. However, tools for Tier 2 databases that integrate genome and transcriptome data are lacking.

MolluscDB is a Tier 2 genome and transcriptome database for the phylum Mollusca built with GenomeHubs [12]. MolluscDB delivers the core elements of Ensembl, plus additional resources conferred by GenomeHubs. Here, we discuss the features of MolluscDB and highlight its utility by using it to support analyses of mollusc phylogeny, of gene family evolution and of genes associated with biomineralization.

2. Methods

(a) Software tools used

Relevant parameters and version numbers of software tools are given in electronic supplementary material, table S5.

(b) Infrastructure

MolluscDB consists of a set of Docker containers and scripts provided by GenomeHubs. Infrastructural containers (EasyMirror [12], SequenceServer [13], h5ai (https://larsjung.de/h5ai/) and MySQL) were set up on an LXC container running Ubuntu 18.04 on a dedicated server (with access to 4 cores and 16 GB of RAM). Analytical and import-related containers (BLAST [14], BUSCO [15], CEGMA [16,17], InterProScan [18], RepeatMasker [19] and EasyImport [12]) were run when needed in a local compute cluster at the Institute of Evolutionary Biology, Edinburgh (768 cores, 384 GB to 1.5 TB RAM per node). Certain features of the Ensembl instance were modified using a custom plugin (https://github.com/genomehubs/molluscdb-plugin/).

(c) *De novo* transcriptome assembly

Raw reads from different studies were downloaded from the Sequence Read Archive (SRA) [20]. The quality of each dataset was first assessed with FastQC [21]. Raw reads were trimmed to remove adapters and low quality or very short sequences with BBDuk [22]. Trimmed reads were then *de novo* assembled with Trinity [23]. Finally, candidate coding regions within the transcripts were predicted with TransDecoder [24]. Flags to account for strand specificity of the datasets were used for both Trinity and TransDecoder when appropriate. BLAST searches against UniRef90 [25] and hmmscan [26] searches against Pfam [27] were included in the TransDecoder pipeline to maximize sensitivity for capturing open reading frames.

(d) Data import

Transcriptome and genome assemblies from different sources were incorporated into the database (table 1 and electronic supplementary material, table S1). Three species were directly mirrored from Ensembl Genomes [28–31]. Four genome assemblies were imported together with their proteins and gene models from ngenomes.org [32], VectorBase [8,33] and the NCBI Assembly resource [34,35]. Fifteen transcriptome assemblies were added to the database: 4 published assemblies [36–38] and 11 transcriptomes that were *de novo* assembled from publicly available raw read data [3,39–44]. For all the transcriptomes, proteins were predicted and imported together with their assemblies.

(e) GenomeHubs analyses

During importation of genome or transcriptome data, GenomeHubs performs a number of analyses. Genome assemblies were masked with RepeatMasker [19] using the built-in Metazoa repeat database. Sequence similarity of the proteins to SwissProt [25] was determined with BLAST [14]. Domain annotation and GO terms were obtained for each protein via InterProScan [18]. Genome completeness was evaluated with CEGMA [16,17] and BUSCO [15]. The same analyses were performed on transcriptome assemblies except for CEGMA and RepeatMasker.

A full set of orthology predictions and gene tree reconstructions was performed to enable Ensembl Compara [45] data displays within MolluscDB. The full orthology pipeline used Orthofinder 2 [46] for the majority of analysis steps and was implemented in a GenomeHubs Compara container [12]. Protein sequences in MolluscDB were clustered based on pairwise DIAMOND [47] blastp searches using the default inflation parameter of 1.5. For each orthogroup, protein sequences were aligned using MAFFT [48] and the resulting alignments were trimmed to remove poorly aligned regions using trimAl [49]. Approximate maximum-likelihood gene trees were reconstructed using FastTree [50] and reconciled against a species tree generated during the same Orthofinder run from a concatenated alignment of 1163 single-copy or mostly single-copy genes to identify gene duplication events, orthologs and paralogs. Orthofinder results were processed and imported into a Compara database as part of the GenomeHubs import.

(f) Analysis of protein families

Protein, GFF3 and InterProScan files for the 22 species of molluscs detailed in table 1 were downloaded from MolluscDB via the download section. As two gene sets were available for *Lymnaea stagnalis*, only the files corresponding to the AUGUSTUS [51] annotation were used. Additionally, protein and GFF3 files of the annelid *Capitella teleta* were downloaded from Ensembl Genomes [28,31], and InterProScan [18] annotation was generated using the same version and parameters used for the MolluscDB species. Protein files were filtered to remove sequences shorter than 30 residues and predicted peptides with internal stop codons. For transcriptomes or genomes for which isoform information was available (*Biomphalaria glabrata* and *Octopus bimaculoides*), only the longest isoform for each locus was retained. Proteins were clustered with OrthoFinder [52] at an inflation value of 3.0 using BLAST similarity information.

A total of 5 one-to-one single-copy orthologue clusters and 2182 'fuzzy' single-copy orthologue clusters (clusters with

Table 1. Genome and transcriptome assemblies available in MolluscDB[a]. An extended version of this table is available in electronic supplementary material, table S1.

species	class	span (Mb)	scaffold count	scaffold N50 (kb)	contig count	contig N50 (kb)	BUSCO complete (%)[b]
genome data							
Bathymodiolus platifrons	Bivalvia	1658	65 662	343	272 497	13	85.80
Crassostrea gigas	Bivalvia	558	7658	402	30 459	31	94.10
Modiolus philippinarum	Bivalvia	2630	74 573	100	301 873	39	82.60
Octopus bimaculoides	Cephalopoda	2338	151 674	475	700 124	6	86.80
Biomphalaria glabrata	Gastropoda	916	331 400	48	369 696	13	84.50
Lottia gigantea	Gastropoda	360	4469	1870	18 335	96	94.40
Lymnaea stagnalis	Gastropoda	997	148 229	5	328 378	6	88.10
Capitella teleta (outgroup)[a]	Polychaeta	277	20 803	188	49 393	22	98.40
transcriptome data							
Cristaria plicata	Bivalvia	418			523 239	2	100.00
Laternula elliptica	Bivalvia	297			324 119	1	98.60
Mya arenaria	Bivalvia	76			118 239	1	70.30
Mya truncata	Bivalvia	361			684 686	1	96.70
Mytilus edulis	Bivalvia	336			592 134	1	84.80
Mytilus galloprovincialis	Bivalvia	181			227 675	1	92.10
Pecten maximus	Bivalvia	195			298 288	1	85.80
Scutopus ventrolineatus	Caudofoveata	116			246 430	1	72.90
Octopoteuthis deletron	Cephalopoda	114			122 672	2	95.70
Vampyroteuthis infernalis	Cephalopoda	105			149 961	1	88.80
Laevipilina hyalina	Monoplacophora	135			287 179	1	52.50
Acanthochitona crinita	Polyplacophora	208			266 385	1	96.70
Gadila tolmiei	Scaphopoda	181			345 172	1	93.70
Gymnomenia pellucida	Solenogastres	194			266 289	1	86.10
Wirenia argentea	Solenogastres	327			563 852	1	73.00

[a]*Capitella teleta* is not a mollusc so is not present in the database, which is why it is marked as 'outgroup'.

[b]BUSCO loci from the eukaryota_odb9 set identified in the assembly.

maximum 5 proteins per taxon and with members in at least 75% of the taxa) were identified via KinFin [53]. Sequences in these 2187 clusters were aligned with MAFFT [48]. Alignments were trimmed with trimAl [49] and single-gene trees generated via RAxML [54]. Strict orthologues were inferred via PhyloTreePruner [55], yielding 1238 orthologous loci. Alignments of these were concatenated into a supermatrix of 262 970 distinct alignment positions (with 20.31% missing data) with FASconCAT [56] and a phylogenetic tree was inferred using RAxML [54].

An extended KinFin analysis was performed on the orthogroups in order to identify synapomorphic clusters and explore the expansions and contractions of protein families. KinFin defines synapomorphic clusters for each node of the phylogenetic tree using Dollo parsimony, which requires that only proteins of taxa under a given node be members of the cluster, and that proteins of at least one taxon from each child node be present. The topology of the phylogenetic tree of the taxa and the functional annotation of the proteins were supplied as input. For the KinFin analyses, taxa were grouped into the following taxonomic sets: Polyplacophora (*Acanthochitona crinita*), Gastropoda (*Biomphalaria glabrata*, *Lottia gigantea*, *Lymnaea stagnalis*), Bivalvia (*Bathymodiolus platifrons*, *Crassostrea gigas*, *Cristaria plicata*, *Laternula elliptica*, *Mya arenaria*, *Mytilus edulis*, *Mytilus galloprovincialis*, *Modiolus philippinarum*, *Mya truncata*, *Pecten maximus*), Solenogastres (*Gymnomenia pellucida*, *Wirenia argentea*), Scaphopoda (*Gadila tolmiei*), Monoplacophora (*Laevipilina hyalina*), Cephalopoda (*Octopus bimaculoides*, *Octopoteuthis deletron*, *Vampyroteuthis infernalis*), Caudofoveata (*Scutopus ventrolineatus*) and the outgroup (the annelid *Capitella teleta*).

Shell matrix proteins (SMPs) were identified by sequence similarity to a list of experimentally validated SMPs [57]. Proteins shorter than 50 residues were filtered out. Reciprocal best BLAST hits between SMPs and proteins in the clustering were evaluated via rbbh.py (https://github.com/DRL/rbbh).

3. Results and discussion

(a) MolluscDB resources

MolluscDB is available openly at https://molluscdb.org. The database collates the genomic data for 22 species of mollusc including 7 species represented by genome sequences and 15 by transcriptome assemblies. Data are stored in the Ensembl schema and thus could be queried with tools developed by the Ensembl project [5], or custom tools using the Ensembl application programming interface (API). The 22 species in MolluscDB represent eight major classes of molluscs (table 1). For each assembly, a landing page includes a brief description of the species, information on the assembly and

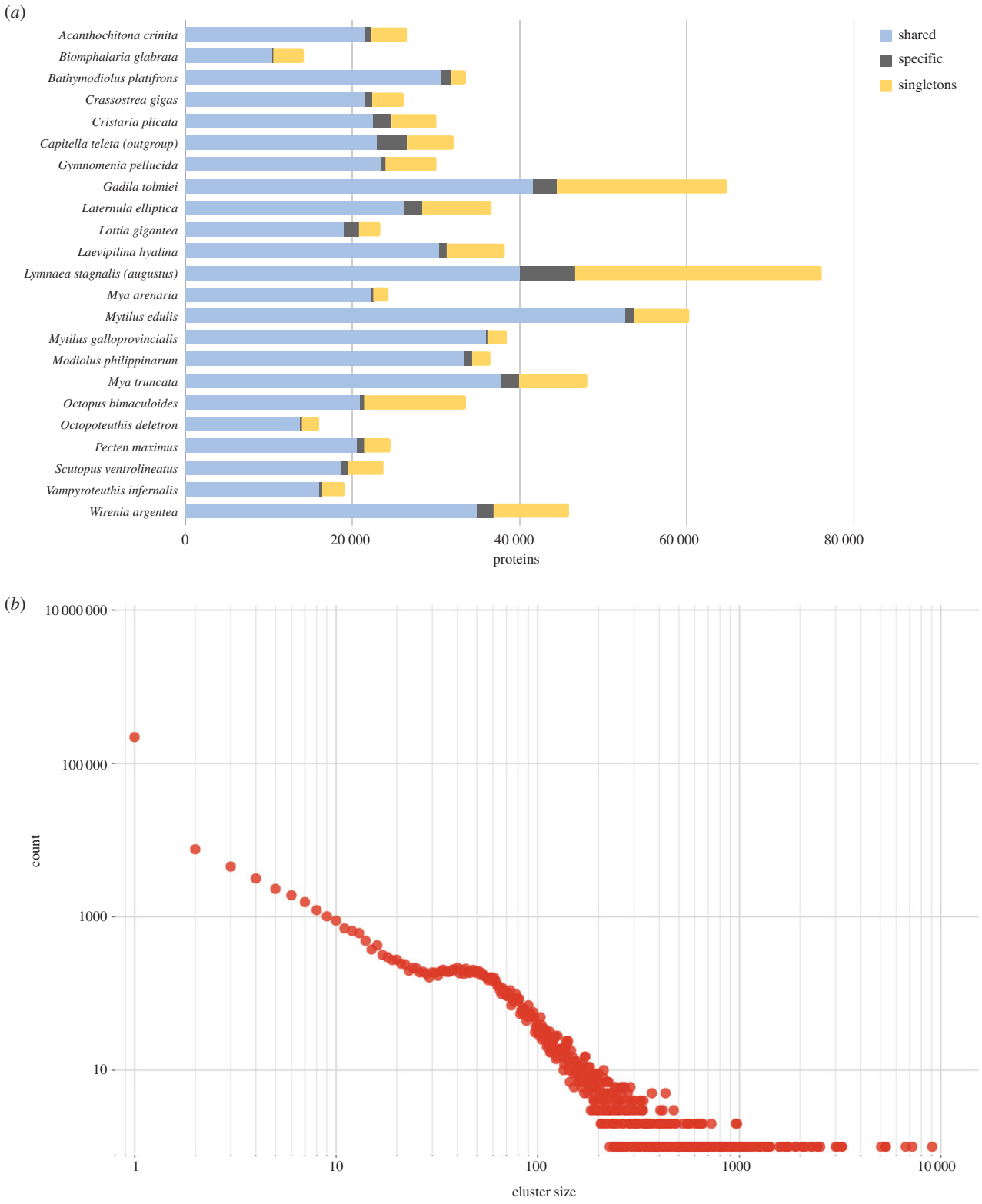

Figure 1. Protein families in species represented in MolluscDB. (*a*) Stacked histogram of proteins in each taxon analysed assigned to: 'shared': proteins in clusters containing proteins from multiple taxa; 'specific': proteins in clusters containing two or more proteins from a single proteome; and 'singleton': proteins in single protein clusters. (*b*) Frequency plot of cluster size in the OrthoFinder clustering of 214 608 orthogroups. (Online version in colour.)

its annotation, and interactive plots displaying assembly metrics [58] and codon usage [59].

The GenomeHubs [12] implementation of the Ensembl infrastructure permits the hosting of multiple assemblies from the same species as well as multiple annotations of one assembly. Thus, for *L. stagnalis*, two annotations of the same assembly can be accessed. A text search function allows the discovery of specific sequences, regions or annotation terms. For each species genome, a gbrowse genome browser representation of the data is available. For the transcriptomes, the data are represented as one contig per assembled transcript, with the open reading frame annotated. Importantly, the annotation of genome-derived and transcriptome-derived proteins is consistent across species, as all the genomes and transcriptomes were decorated with functional inferences derived from searches and comparisons

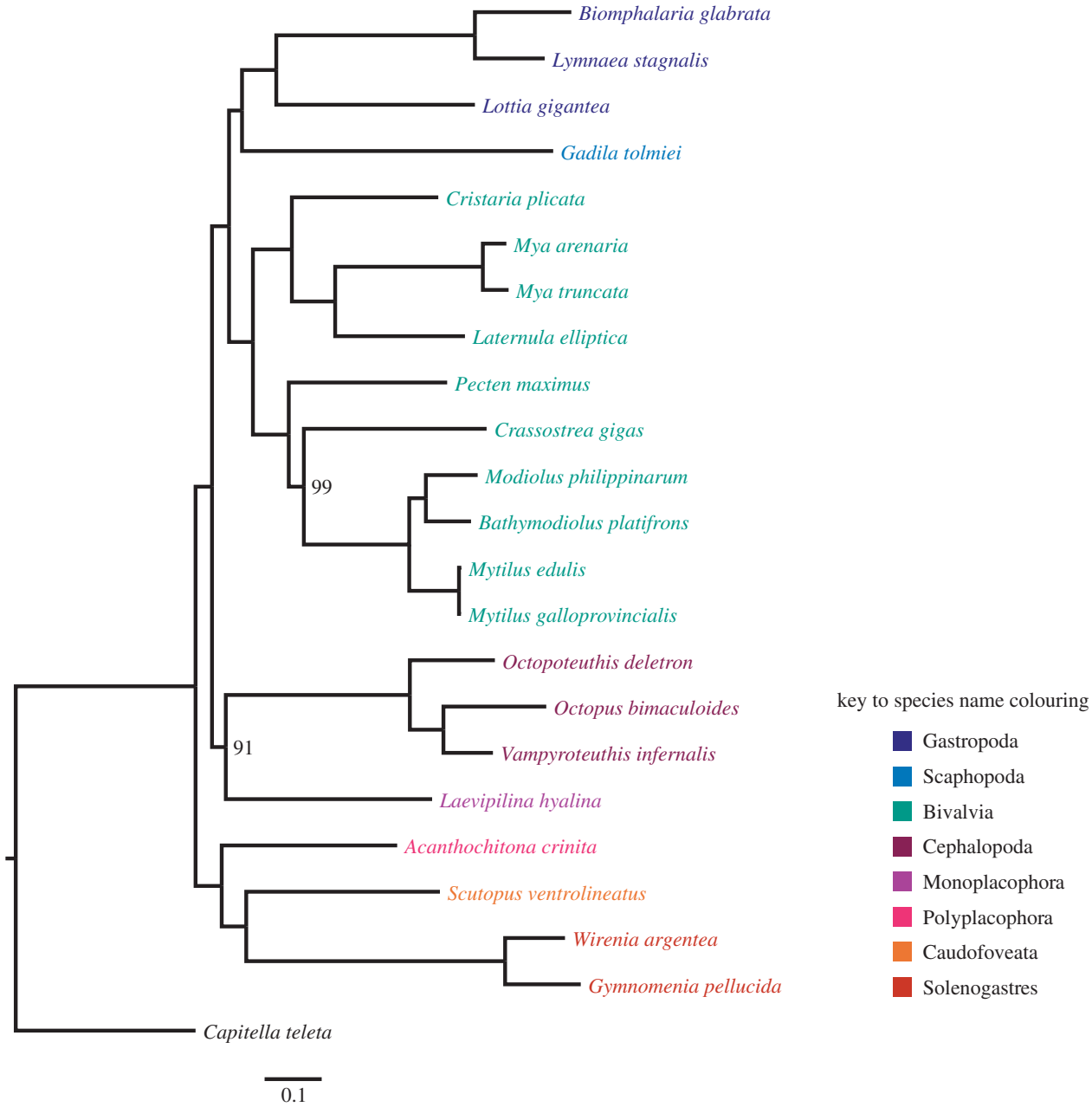

Figure 2. Phylogenetic tree of taxa in MolluscDB. Multilocus phylogeny of the species analysed. Support is 100 at all nodes except indicated.

to the same libraries of reference information. We note that the assembled transcriptomes contain many more contigs (assembled transcripts) and have much longer spans than would be expected of a mollusc genome, and many more than the well-annotated complete genome sequences. This feature of transcriptome assemblies is well known, and results from a preponderance of short contigs supported by very few sequences. We have included the complete assemblies rather than filter by coverage as we would rather not exclude possibly biologically meaningful information.

MolluscDB includes an instance of Sequenceserver [13] so that a user can perform BLAST searches against any sequence hosted in the database at https://blast.molluscdb.org/. Two types of BLAST databases are available: nucleotide databases, which include scaffolds, transcripts and coding sequences (CDS), and protein databases. The BLAST search parameters can be modified from default to facilitate advanced search. A link in the header of every result connects each sequence with its Ensembl entry. The MolluscDB download server at https://download.molluscdb.org/ allows users to download any sequence or analysis hosted in the database. Files are consistently named and formatted.

(b) Using MolluscDB to explore protein family evolution in Mollusca

We used an orthology clustering of the protein-coding genes in MolluscDB to explore the protein family traits of these species. A total of 802 455 proteins were retrieved after the filtering of the 23 proteomes included in the study. OrthoFinder [52] grouped these proteins into 214 608 clusters, 153 141 (71.4%) of which were singletons. *L. stagnalis* and *G. tolmiei* contributed the largest number of proteins to the clustering, and together they accounted for 32.62% of singleton clusters (figure 1*a*). The clusters presented a power-law-like frequency distribution with a deviation at cluster size 23 (figure 1*b*). Similar patterns have been observed before in such analyses [53,60–62]. This peak is largely made up of single-copy orthologues.

Table 2. Protein families that constitute molecular synapomorphies for Bivalvia. An extended version of this table is available in electronic supplementary material, table S2.

cluster ID	taxon count	protein count	MGALL	MEDUL	BPLAT	MPHIL	CGIGA	PMAXI	LELLI	MTRUN	MAREN	CPLIC	domain ID	domain description
OG0013861	10	12	1	1	1	1	1	1	1	2	1	2	IPR001938	thaumatin family
OG0012076	10	16	1	1	1	1	3	2	3	2	1	1	IPR013901[a]	protein of unknown function DUF1772 (PF08592)
OG0015305	9	10	0	1	1	2	1	1	1	1	1	1	IPR025475	protein of unknown function DUF4326
OG0016110	8	9	1	2	1	1	1	1	1	0	0	1	IPR005198	glycoside hydrolase, family 76
OG0016168	8	9	2	1	1	1	1	0	1	1	1	0	IPR005320	peptidase S51
OG0015570	7	10	2	1	0	1	0	2	1	2	1	0	IPR000096	serum amyloid A protein
OG0012573	7	15	1	4	0	2	0	2	0	1	1	4	IPR004305	thiaminase-2/PQQC
OG0007391	7	26	1	9	5	0	0	2	0	3	1	5	IPR005049	STELLO-like
OG0016163	7	9	1	1	2	1	1	1	2	0	0	0	IPR013901[a]	domain of unknown function DUF1772 (PF08592)
OG0015262	7	10	2	2	1	0	1	1	0	0	2	1	IPR028060	big defensin
OG0009745	6	22	2	6	6	1	0	0	0	6	1	0	IPR007743	immunity-related GTPases-like
OG0025374	5	5	1	1	0	1	0	0	0	1	1	0	IPR007855	RNA-dependent RNA polymerase, eukaryotic-type
OG0005710	5	29	0	0	0	3	11	0	3	7	5	0	IPR024518	domain of unknown function DUF3421
OG0016192	5	9	0	1	2	1	0	0	2	3	0	0	IPR029230	Macin

[a]This IPR entry has been retired.

Table 3. Protein families that have significantly different numbers of members in Bivalvia. An extended version of this table is available in electronic supplementary material, table S3.

cluster ID	species with members in cluster	total proteins in cluster	Bivalvia proteins in cluster	status	domain ID	domain description
OG0000078	22	241	189	enriched	IPR002227	tyrosinase copper-binding domain
OG0000280	22	123	97	enriched	IPR008858	TROVE domain
OG0000236	19	134	116	enriched	IPR001073	C1q domain
OG0000462	19	92	73	enriched	IPR006612	zinc finger, C2CH-type
OG0000931	19	63	51	enriched	IPR003886	NIDO domain
OG0000091	18	227	198	enriched	IPR027370	RING-type zinc-finger, LisH dimerization motif
OG0000121	18	198	186	enriched	IPR001611	leucine-rich repeat
					IPR000157	toll/interleukin-1 receptor homology (TIR) domain
OG0000216	18	142	119	enriched	IPR002018	carboxylesterase, type B
OG0000166	17	163	142	enriched	IPR006202	neurotransmitter-gated ion-channel ligand-binding domain
					IPR006029	neurotransmitter-gated ion-channel transmembrane domain
OG0000173	17	159	144	enriched	IPR000210	BTB/POZ domain
					IPR011705	BTB/Kelch-associated
OG0000607	17	79	68	enriched	IPR004245	protein of unknown function DUF229
OG0000275	15	123	107	enriched	IPR003961	fibronectin type III
OG0000451	18	94	8	depleted	IPR007074	LicD family
OG0000752	18	70	7	depleted	IPR001506	peptidase M12A

A well-supported phylogeny was inferred from the set of 1238 orthologous loci (figure 2). All branches display bootstrap support values of 100, with the exception of two: the branch leading to Cephalopoda and Monoplacophora, and a branch leading to a subclade of Bivalves. Our results support a division in two major clades: Aculifera (Caudofoveata, Polyplacophora and Solenogastres) and Conchifera (Bivalvia, Cephalopoda, Gastropoda, Monoplacophora and Scaphopoda). Within Aculifera, Polyplacophora was placed sister to aplacophorans (Caudofoveata and Solenogastres). Within Conchifera, we recovered Bivalvia sister to a clade including Gastropoda and Scaphopoda. These results coincide with previous phylogenetic analyses [41,63]. Monoplacophora was recovered as the sister taxon of Cephalopoda. While this is consistent with Smith *et al.* [41], a recent study including genomic data from a newly sequenced monoplacophoran suggests that Monoplacophora could be a sister taxon to the rest of Conchifera [63]. Clearly, a more exhaustive sampling of loci across and within taxa is needed to create a robust phylogenetic framework for molluscs.

An insight on the distinctive biology of a monophyletic group can be gained through the study of synapomorphies. Just like morphological traits, molecular features such as gene presence and absence or gene family expansion and contraction are informative in a phylogenetic context. Using KinFin [53], we identified protein families that were unique to particular clades. By decorating these with functional attributes based on domain and sequence similarity, we assigned likely biological meaning to these synapomorphies. Here, we present an analysis of species in Bivalvia compared with all other mollusc groups in the database (table 2 and electronic supplementary material, table S2). Analysis of the OrthoFinder [52] clustering identified 14 synapomorphic clusters with the presence of at least five of the ten bivalve species. Annotations associated with these 14 clusters could be grouped into two main classes: immunity and metabolism. The dominant annotations related to immunity to cellular and viral pathogens, including RNA-dependent RNA polymerase (involved in the RNAi response to invading viral nucleic acids), big defensin [64,65], serum amyloid A [66,67], thaumatin [68], macin [69,70] and an immunity-related GTPase. Four Bivalvia-restricted clusters were annotated only as containing matches to domains of unknown function (DUF). One of these domains, DUF3421, has been associated with stress-responsive, sugar-binding natterins in *C. gigas* [71], where they may play roles in immune defence [72,73]. Metabolic annotations included carbohydrate metabolism (glycoside hydrolase family 76, and, possibly, a STELLO-like domain-containing protein), protein metabolism (peptidase S51) and degradation of vitamin B1 (Thiaminase-2/PQQC).

KinFin also facilitates analysis of expansion and contraction of protein families between clades by considering cluster membership count variation in a statistical framework akin to that deployed for gene expression analysis [53]. We identified protein clusters that had significantly different numbers of members in Bivalvia species than in the other taxa analysed (table 3 and electronic supplementary material, table S3). To be selected, the clusters had to have members in at least 7 Bivalvia and 7 other taxa, with Log2 mean ≥ 2 or ≤ -2 and a p-value below 0.05 for the difference between Bivalvia and other taxa.

Table 4. Functionally annotated shell matrix proteins. An extended version of this table is available in electronic supplementary material, table S4.

cluster ID	domain ID	domain description	number of proteins in cluster	number of species represented in cluster
OG0000137	IPR001223	glycoside hydrolase family 18, catalytic domain	182	23
	IPR002557	chitin-binding domain		
OG0000159	IPR001466	beta-lactamase-related	167	23
OG0000231	IPR000668	peptidase C1A, papain C-terminal	136	23
	IPR013201	cathepsin propeptide inhibitor domain (I29)		
OG0000631	IPR019479	peroxiredoxin, C-terminal	77	23
	IPR000866	alkyl hydroperoxide reductase subunit C/ thiolspecific antioxidant		
OG0000962	IPR002130	cyclophilin-type peptidyl-prolyl *cis-trans* isomerase domain	62	23
OG0001572	IPR000741	fructose-bisphosphate aldolase, class-I	49	23
OG0001810	IPR017868	filamin/ABP280 repeat-like	46	23
	IPR001715	calponin homology domain		
OG0000078	IPR002227	tyrosinase copper-binding domain	241	22
OG0000127	IPR019791	haem peroxidase, animal type	190	22
OG0000344	IPR015798	copper amine oxidase, C-terminal	110	22
	IPR015800	copper amine oxidase, N2-terminal		
OG0002836	IPR001660	sterile alpha motif domain	38	22
OG0005847	IPR001715	calponin homology domain	28	22
OG0000686	IPR014044	CAP domain	73	21
OG0009156	IPR001152	beta-thymosin	23	21
OG0000036	IPR001304	C-type lectin-like	369	19
OG0000093	IPR002557	chitin-binding domain	225	19
	IPR002035	von Willebrand factor, type A		
OG0000222	IPR031569	apextrin, C-terminal domain	138	19
OG0001850	IPR002937	amine oxidase	45	12
OG0011795	IPR000867	insulin-like growth factor-binding protein, IGFBP	17	10
OG0013954	IPR001223	glycoside hydrolase family 18, catalytic domain	12	8
OG0015275	IPR003961	fibronectin type III	10	7
OG0017162	IPR002035	von Willebrand factor, type A	8	7
OG0005249	IPR002035	von Willebrand factor, type A	30	6
OG0018497	IPR002223	pancreatic trypsin inhibitor Kunitz domain	7	6
OG0018680	IPR001148	alpha carbonic anhydrase	7	3
OG0020309	IPR015882	beta-hexosaminidase, bacteial type, N-terminal	6	6
	IPR015883	glycoside hydrolase family 20, catalytic domain		
	IPR004866	chitobiase/beta-hexosaminidases, N-terminal domain		
OG0020490	IPR001073	C1q domain	6	3

There were 14 clusters with differential representation in Bivalvia. Twelve had higher family sizes in the bivalves. These included one annotated as tyrosinase, an enzyme implicated in shell formation [74–76], and several families annotated with domains associated with mollusc immunity, including C1q-like proteins [77], Toll/interleukin-1 receptor (TIR) [78] and fibronectin type III [79] domains. These findings mesh with previous descriptions of gene family expansions in bivalves of tyrosinase [80], C1q [77] and TIR [81]. Other enriched clusters have annotations including zinc-finger domains (C2CH-type and RING-type), carboxylesterase type B, neurotransmitter-gated ion-channel ligand-binding and transmembrane domains, TROVE domain, NIDO domain, BTB/POZ domain and a domain of unknown function (DUF229). Two clusters, annotated as containing peptidase M12A and LicD nucleotidyltransferase superfamily members, displayed a significantly lower number of proteins in Bivalvia compared with the other taxa.

(c) Shell matrix proteins

SMPs comprise a heterogeneous set of enzymes and structural proteins implicated in the biomineralization process, either by the demonstration of secretion by the mantle during shell synthesis or repair, or through isolation of peptides from isolated shell material [82]. We used a previously curated list of experimentally validated SMPs [57] to interrogate the gene sets in MolluscDB to identify presence/absence and gene family size change patterns (table 4 and electronic supplementary material, table S4).

Seventy-four clusters presented proteins with significant sequence similarity to SMPs. Of the 48 clusters with three or more species, 27 had annotation matches to InterPro domains. These included highly conserved biomineralization domains such as tyrosinase, carbonic anhydrase, chitin-binding, von Willebrand factor, protease inhibitors and peroxidases [83,84]. We also retrieved proteins and domains associated with immune functions (C1q, fibronectin type III, C-type lectin and apextrin) [77,79,85–87]. Other matching sequences included proteins involved in metabolism (peptidase, fructose aldolase, lactamase, beta-hexosaminidase, oxidases and glycosyl hydrolases), cross-linking (filamin and calponin), protein folding (cyclophilin), actin filament organization (beta-thymosin) and regulation of insulin-like growth factors. These protein families are a strong substrate for future analysis of molecular correlates of mollusc responses to ocean acidification and warming, and for monitoring farmers' shellfish growth, health and disease.

(d) Outlook

By collating genome and transcriptome data in a single database structure, we have been able to explore genomic data for diverse species of molluscs, and identify genes that may have evolved to deliver clade-specific processes. Using GenomeHubs [12] technology, we were able to incorporate genomes from both existing Ensembl instances and genomes that were too fragmented to incorporate in such pan-taxonomic databases. Transcriptome datasets are particularly attractive and economic to generate, as they sample only the expressed genome and allow immediate access to potential genes of interest. We have shown that these data can be rapidly incorporated and coordinated with full genome data in a consistent and accessible way. It is essential to recognize the key differences between transcriptome assembly-derived and genome-derived protein sets, such as the presence of multiple distinct isoforms and gene fragments in transcriptome assemblies. Despite this, the transcriptomes reliably report on the presence of a gene or gene family in a species, and facilitate filtering of lists of target genes to include (or exclude) those with broad phylogenetic representation.

MolluscDB currently presents 22 genomes and transcriptomes from the phylum Mollusca. To date, there are 49 genome assemblies in the International Nucleotide Sequence Database Consortium (INSDC; GenBank, European Nucleotide Archive, DNA Databank of Japan) databases (see https://www.ncbi.nlm.nih.gov/genome/?term=txid6447[Organism:exp], sourced 01 September 2020) with very different completeness and contiguity metrics. Incorporating these assemblies in MolluscDB is a near-future goal for the project. There are nearly 8000 mollusc transcriptome datasets in the short read archive (SRA) from 646 species (see https://www.ncbi.nlm.nih.gov/sra/?term=txid6447[Organism:exp]+and+transcriptomic, sourced 1 September 2020), 601 of which have no genome data. While some of these transcriptome datasets will not be suitable for assembly and presentation owing to low size or complexity of sample (including symbionts or other cobionts), they represent a large, currently untapped resource of information for comparative and functional genomics. Several global and regional projects, such as the Earth BioGenome Project [88] and Darwin Tree of Life project (https://darwintreeoflife.org), intend to sequence and assemble the genomes of large numbers of mollusc and other species, suggesting that the need for analysis hubs will only grow. Current database architectures may struggle to host and display such large amounts of data. For example, the 601 transcriptomes alone may generate 200 million assembled contigs and associated protein predictions and functional annotations. We are, therefore, also developing the GenomeHubs platform to scale to these new demands.

Note added in proof

Since this article was accepted, Liu *et al*. have published their database of molluscan genome data, also called MolluscDB (at http://mgbase.qnlm.ac) (*Nucleic Acids Research* 2021, **49**: D1556. (Liu F, Li Y, Yu H, Zhang L, Hu J, Bao Z, Wang S. 2021 MolluscDB: an integrated functional and evolutionary genomics database for the hyper-diverse animal phylum Mollusca. *Nucleic Acids Research* **49**, D988–D997. (doi:10.1093/nar/gkaa918)). Their presentation includes similar functionality to MolluscDB presented here. We will liaise with the authors to ensure that the community is best served by our complementary efforts.

Data accessibility. All data in MolluscDB are downloadable from the INSDC partner databases and/or from https://download.molluscdb.org/.

Authors' contributions. C.C.: conception, analysis, writing. D.R.L.: conception, software, analysis, writing. S.K.: conception, software, analysis, writing. R.C.: conception, software, analysis, writing. K.G.: conception, funding, writing. M.B.: conception, funding, analysis, writing.

Competing interests. We declare we have no competing interests.

Funding. Specific funding for this project was received from Marie Curie Innovative Training Networks (ITN) grant agreement 605051.

Acknowledgements. We thank members of the Blaxter Laboratory and the ITN CACHE network for their support and suggestions. We appreciate the generous guidance and insights offered by Kevin Kocot. Analyses were performed using the Institute of Evolutionary Biology CLUB compute cluster, which was funded from a mix of UK Research and Innovation (NERC and BBSRC), ERC and charitable sources.

References

1. Muir P *et al*. 2016 The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* **17**, 53. (doi:10.1186/s13059-016-0917-0)
2. Gomes-dos-Santos A, Lopes-Lima M, Castro LFC, Froufe E. 2020 Molluscan genomics: the road so far and the way forward. *Hydrobiologia* **847**, 1705–1726. (doi:10.1007/s10750-019-04111-1)
3. González VL, Andrade SCS, Bieler R, Collins TM, Dunn CW, Mikkelsen PM, Taylor JD, Giribet G. 2015 A phylogenetic backbone for Bivalvia: an RNA-seq approach. *Proc. R.*

Soc. B **282**, 20142332. (doi:10.1098/rspb.2014.2332)

4. Kocot KM *et al*. 2017 Phylogenomics of Lophotrochozoa with consideration of systematic error. *Syst. Biol.* **66**, 256–282. (doi:10.1093/sysbio/syw079)
5. Yates AD *et al*. 2020 Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688. (doi:10.1093/nar/gkz1138)
6. Thurmond J *et al*. 2019 FlyBase 2.0: the next generation. *Nucleic Acids Res.* **47**, D759–D765. (doi:10.1093/nar/gky1003)
7. Harris TW *et al*. 2020 WormBase: a modern Model Organism Information Resource. *Nucleic Acids Res.* **48**, D762–D767. (doi:10.1093/nar/gkz920)
8. Giraldo-Calderón GI *et al*. 2015 VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.* **43**(Database issue), D707–D713. (doi:10.1093/nar/gku1117)
9. Eöry L *et al*. 2015 Avianbase: a community resource for bird genomics. *Genome Biol.* **16**, 21. (doi:10.1186/s13059-015-0588-2)
10. Challis RJ, Kumar S, Dasmahapatra KK, Jiggins CD, Blaxter M. 2016 Lepbase: the Lepidopteran genome database. *bioRxiv*. (doi:10.1101/056994)
11. Parkhill J, Birney E, Kersey P. 2010 Genomic information infrastructure after the deluge. *Genome Biol.* **11**, 402. (doi:10.1186/gb-2010-11-7-402)
12. Challis RJ, Kumar S, Stevens L, Blaxter M. 2017 GenomeHubs: simple containerized setup of a custom Ensembl database and web server for any species. *Database* **2017**, bax039. (doi:10.1093/database/bax039)
13. Priyam A *et al*. 2019 Sequenceserver: a modern graphical user interface for custom BLAST databases. *Mol. Biol. Evol.* **36**, 2922–2924. (doi:10.1093/molbev/msz185)
14. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009 BLAST+: architecture and applications. *BMC Bioinf.* **10**, 421. (doi:10.1186/1471-2105-10-421)
15. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212. (doi:10.1093/bioinformatics/btv351)
16. Parra G, Bradnam K, Korf I. 2007 CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067. (doi:10.1093/bioinformatics/btm071)
17. Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009 Assessing the gene space in draft genomes. *Nucleic Acids Res.* **37**, 289–297. (doi:10.1093/nar/gkn916)
18. Jones P *et al*. 2014 InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240. (doi:10.1093/bioinformatics/btu031)
19. Smith AFA, Hubley R, Green P. 2020 RepeatMasker Open-4.0 2013–2015 [Internet]. See http://www.repeatmasker.org/ (accessed 1 Sep 2020).
20. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. 2011 The sequence read archive. *Nucleic Acids Res.* **39**(Database issue), D19–D21. (doi:10.1093/nar/gkq1019)
21. Andrews S. 2010 FastQC: a quality control tool for high throughput sequence data. See http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed 1 Sep 2020).
22. Bushnell B. 2020 BBMap [Internet]. See https://sourceforge.net/projects/bbmap/ (accessed 1 Sep 2020).
23. Grabherr MG *et al*. 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652. (doi:10.1038/nbt.1883)
24. Haas BJ, Papanicolaou A. 2020 TransDecoder. See https://github.com/TransDecoder/TransDecoder (accessed 1 Sep 2020).
25. UniProt Consortium. 2019 UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**(D1), D506–D515. (doi:10.1093/nar/gky1049)
26. Eddy SR. 2020 HMMER. See http://hmmer.org/ (accessed 1 Sep 2020).
27. El-Gebali S *et al*. 2019 The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**(D1), D427–D432. (doi:10.1093/nar/gky995)
28. Howe KL *et al*. 2020 Ensembl Genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res.* **48**(D1), D689–D695. (doi:10.1093/nar/gkz890)
29. Zhang G *et al*. 2012 The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**, 49–54. (doi:10.1038/nature11413)
30. Albertin CB *et al*. 2015 The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* **524**, 220–224. (doi:10.1038/nature14668)
31. Simakov O *et al*. 2013 Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**, 526–531. (doi:10.1038/nature11696)
32. Davison A *et al*. 2016 Formin is associated with left–right asymmetry in the pond snail and the frog. *Curr. Biol.* **26**, 654–660. (doi:10.1016/j.cub.2015.12.071)
33. Adema CM *et al*. 2017 Whole genome analysis of a schistosomiasis-transmitting freshwater snail. *Nat. Commun.* **8**, 15451. (doi:10.1038/ncomms15451)
34. Kitts PA *et al*. 2016 Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* **44**(D1), D73–D80. (doi:10.1093/nar/gkv1226)
35. Sun J *et al*. 2017 Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nat. Ecol. Evol.* **1**, 121. (doi:10.1038/s41559-017-0121)
36. Sleight VA, Antczak P, Falciani F, Clark MS. 2020 Computationally predicted gene regulatory networks in molluscan biomineralization identify extracellular matrix production and ion transportation pathways. *Bioinformatics* **36**, 1326–1332. (doi:10.1093/bioinformatics/btz754)
37. Sleight VA, Peck LS, Dyrynda EA, Smith VJ, Clark MS. 2018 Cellular stress responses to chronic heat shock and shell damage in temperate *Mya truncata*. *Cell Stress Chaperones* **23**, 1003–1017. (doi:10.1007/s12192-018-0910-5)
38. Francis WR, Christianson LM, Haddock SHD. 2017 Symplectin evolved from multiple duplications in bioluminescent squid. *PeerJ* **5**, e3633. (doi:10.7717/peerj.3633)
39. Yarra T, Gharbi K, Blaxter M, Peck LS, Clark MS. 2016 Characterization of the mantle transcriptome in bivalves: *Pecten maximus*, *Mytilus edulis* and *Crassostrea gigas*. *Mar. Genomics* **27**, 9–15. (doi:10.1016/j.margen.2016.04.003)
40. De Oliveira AL *et al*. 2016 Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks. *BMC Genomics* **17**, 905. (doi:10.1186/s12864-016-3080-9)
41. Smith SA *et al*. 2011 Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature* **480**, 364–367. (doi:10.1038/nature10526)
42. Zapata F *et al*. 2014 Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. *Proc. R. Soc. B* **281**, 20141739. (doi:10.1098/rspb.2014.1739)
43. Bjärnmark NA, Yarra T, Churcher AM, Felix RC, Clark MS, Power DM. 2016 Transcriptomics provides insight into *Mytilus galloprovincialis* (Mollusca: Bivalvia) mantle function and its role in biomineralisation. *Mar. Genomics* **27**, 37–45. (doi:10.1016/j.margen.2016.03.004)
44. Patnaik BB *et al*. 2016 Sequencing, *de novo* assembly, and annotation of the transcriptome of the endangered freshwater pearl bivalve, *Cristaria plicata*, provides novel insights into functional genes and marker discovery. *PLoS ONE* **11**, e0148622. (doi:10.1371/journal.pone.0148622)
45. Herrero J *et al*. 2016 Ensembl comparative genomics resources. *Database* **2016**, bav096. (doi:10.1093/database/bav096)
46. Emms DM, Kelly S. 2019 OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238. (doi:10.1186/s13059-019-1832-y)
47. Buchfink B, Xie C, Huson DH. 2015 Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60. (doi:10.1038/nmeth.3176)
48. Katoh K, Standley DM. 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. (doi:10.1093/molbev/mst010)
49. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009 trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973. (doi:10.1093/bioinformatics/btp348)
50. Price MN, Dehal PS, Arkin AP. 2010 FastTree 2 — approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490. (doi:10.1371/journal.pone.0009490)
51. Keller O, Kollmar M, Stanke M, Waack S. 2011 A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763. (doi:10.1093/bioinformatics/btr010)
52. Emms DM, Kelly S. 2015 OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference

accuracy. *Genome Biol.* **16**, 157. (doi:10.1186/s13059-015-0721-2)

53. Laetsch DR, Blaxter ML. 2017 KinFin: software for taxon-aware analysis of clustered protein sequences. *G3 (Bethesda)* **7**, 3349–3357. (doi:10.1534/g3.117.300233)
54. Stamatakis A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. (doi:10.1093/bioinformatics/btu033)
55. Kocot KM, Citarella MR, Moroz LL, Halanych KM. 2013 PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evol. Bioinform. Online* **9**, 429–435. (doi:10.4137/EBO.S12813)
56. Kück P, Meusemann K. 2010 FASconCAT: convenient handling of data matrices. *Mol. Phylogenet. Evol.* **56**, 1115–1118. (doi:10.1016/j.ympev.2010.04.024)
57. Yarra T. 2019 Molluscan shell matrix proteins. PhD Thesis. UK Polar Data Centre, British Antarctic Survey, Natural Environment Research Council, UK Research & Innovation, Cambridge, UK.
58. Challis R. 2017 Rjchallis/Assembly-Stats 17.02. See https://zenodo.org/record/322347 [zenodo.org].
59. Challis R. 2017 Rjchallis/Codon-Usage 17.02. See https://zenodo.org/record/322348 [zenodo.org].
60. Unger R, Uliel S, Havlin S. 2003 Scaling law in sizes of protein sequence families: from super-families to orphan genes. *Proteins* **51**, 569–576. (doi:10.1002/prot.10347)
61. Enright AJ, Kunin V, Ouzounis CA. 2003 Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.* **31**, 4632–4638. (doi:10.1093/nar/gkg495)
62. Yoshida Y *et al*. 2017 Comparative genomics of the tardigrades *Hypsibius dujardini* and *Ramazzottius varieornatus*. *PLoS Biol.* **15**, e2002266. (doi:10.1371/journal.pbio.2002266)
63. Kocot KM, Poustka AJ, Stöger I, Halanych KM, Schrödl M. 2020 New data from Monoplacophora and a carefully-curated dataset resolve molluscan relationships. *Sci. Rep.* **10**, 101. (doi:10.1038/s41598-019-56728-w)
64. Rosa RD, Santini A, Fievet J, Bulet P, Destoumieux-Garzón D, Bachère E. 2011 Big defensins, a diverse family of antimicrobial peptides that follows different patterns of expression in hemocytes of the oyster *Crassostrea gigas*. *PLoS ONE* **6**, e25594. (doi:10.1371/journal.pone.0025594)
65. Zhao J, Li C, Chen A, Li L, Su X, Li T. 2010 Molecular characterization of a novel big defensin from clam *Venerupis philippinarum*. *PLoS ONE* **5**, e13480. (doi:10.1371/journal.pone.0013480)
66. Rosani U, Domeneghetti S, Gerdol M, Franzoi M, Pallavicini A, Venier P. 2016 Serum amyloid A in marine bivalves: an acute phase and innate immunity protein. *Dev. Comp. Immunol.* **59**, 136–144. (doi:10.1016/j.dci.2016.01.019)
67. Qu F, Xiang Z, Yu Z. 2014 The first molluscan acute phase serum amyloid A (A-SAA) identified from oyster *Crassostrea hongkongensis*: molecular cloning and functional characterization. *Fish Shellfish Immunol.* **39**, 145–151. (doi:10.1016/j.fsi.2014.05.013)
68. Niu X *et al*. 2018 The antifungal activity of a thaumatin-like protein from oyster *Crassostrea gigas*. *Invertebrate Surviv. J.* **15**, 210–222. (doi:10.25431/1824-307X/isj.v15i1.210-222)
69. Gerdol M, De Moro G, Manfrin C, Venier P, Pallavicini A. 2012 Big defensins and mytimacins, new AMP families of the Mediterranean mussel *Mytilus galloprovincialis*. *Dev. Comp. Immunol.* **36**, 390–399. (doi:10.1016/j.dci.2011.08.003)
70. Yang D *et al*. 2019 A macin identified from *Venerupis philippinarum*: investigation on antibacterial activities and action mode. *Fish Shellfish Immunol.* **92**, 897–904. (doi:10.1016/j.fsi.2019.07.031)
71. Zhang Y *et al*. 2015 Proteomic basis of stress responses in the gills of the pacific oyster *Crassostrea gigas*. *J. Proteome Res.* **14**, 304–317. (doi:10.1021/pr500940s)
72. Leprêtre M, Almunia C, Armengaud J, Salvador A, Geffard A, Palos-Ladeiro M. 2019 The immune system of the freshwater zebra mussel, *Dreissena polymorpha*, decrypted by proteogenomics of hemocytes and plasma compartments. *J. Proteomics* **202**, 103366. (doi:10.1016/j.jprot.2019.04.016)
73. Leprêtre M *et al*. 2020 Identification of immune-related proteins of *Dreissena polymorpha* hemocytes and plasma involved in host–microbe interactions by differential proteomics. *Sci. Rep.* **10**, 6226. (doi:10.1038/s41598-020-63321-z)
74. Zhang C, Xie L, Huang J, Chen L, Zhang R. 2006 A novel putative tyrosinase involved in periostracum formation from the pearl oyster (*Pinctada fucata*). *Biochem. Biophys. Res. Commun.* **342**, 632–639. (doi:10.1016/j.bbrc.2006.01.182)
75. Huan P, Liu G, Wang H, Liu B. 2013 Identification of a tyrosinase gene potentially involved in early larval shell biogenesis of the Pacific oyster *Crassostrea gigas*. *Dev. Genes Evol.* **223**, 389–394. (doi:10.1007/s00427-013-0450-z)
76. Miglioli A, Dumollard R, Balbi T, Besnardeau L, Canesi L. 2019 Characterization of the main steps in first shell formation in *Mytilus galloprovincialis*: possible role of tyrosinase. *Proc. R. Soc. B* **286**, 20192043. (doi:10.1098/rspb.2019.2043)
77. Gerdol M, Venier P, Pallavicini A. 2015 The genome of the Pacific oyster *Crassostrea gigas* brings new insights on the massive expansion of the C1q gene family in Bivalvia. *Dev. Comp. Immunol.* **49**, 59–71. (doi:10.1016/j.dci.2014.11.007)
78. Gerdol M *et al.* 2018 Immunity in Molluscs: recognition and effector mechanisms, with a focus on Bivalvia. In *Advances in comparative immunology* (ed. EL Cooper), pp. 225–341. Cham, Switzerland: Springer International Publishing.
79. Huang Q, Yu M, Chen H, Zeng M, Sun Y, Saha TT. 2018 LRFN (leucine-rich repeat and fibronectin type-III domain-containing protein) recognizes bacteria and promotes hemocytic phagocytosis in the Pacific oyster *Crassostrea gigas*. *Fish Shellfish Immunol.* **72**, 622–628. (doi:10.1016/j.fsi.2017.11.049)
80. Aguilera F, McDougall C, Degnan BM. 2014 Evolution of the tyrosinase gene family in bivalve molluscs: independent expansion of the mantle gene repertoire. *Acta Biomater.* **10**, 3855–3865. (doi:10.1016/j.actbio.2014.03.031)
81. Gerdol M, Venier P, Edomi P, Pallavicini A. 2017 Diversity and evolution of TIR-domain-containing proteins in bivalves and Metazoa: new insights from comparative genomics. *Dev. Comp. Immunol.* **70**, 145–164. (doi:10.1016/j.dci.2017.01.014)
82. Clark MS *et al*. 2020 Deciphering mollusc shell production: the roles of genetic mechanisms through to ecology, aquaculture and biomimetics. *Biol. Rev. Camb. Philos. Soc.* **95**, 1812–1837. https://doi.org/10.1111/brv.12640.
83. Arivalagan J *et al*. 2017 Insights from the shell proteome: biomineralization to adaptation. *Mol. Biol. Evol.* **34**, 66–77. (doi:10.1093/molbev/msw219)
84. Kocot KM, Aguilera F, McDougall C, Jackson DJ, Degnan BM. 2016 Sea shell diversity and rapidly evolving secretomes: insights into the evolution of biomineralization. *Front. Zool.* **13**, 23. (doi:10.1186/s12983-016-0155-z)
85. Li D, Nie H, Jahan K, Yan X. 2020 Expression analyses of C-type lectins (CTLs) in Manila clam under cold stress provide insights for its potential function in cold resistance of *Ruditapes philippinarum*. *Comp. Biochem. Physiol. C Toxicol. Pharmacol.* **230**, 108708. (doi:10.1016/j.cbpc.2020.108708)
86. Shi Y *et al*. 2019 Novel Ca^{2+}-independent C-type lectin involved in immune defense of the razor clam *Sinonovacula constricta*. *Fish Shellfish Immunol.* **84**, 502–508. (doi:10.1016/j.fsi.2018.10.036)
87. Estévez-Calvar N, Romero A, Figueras A, Novoa B. 2011 Involvement of pore-forming molecules in immune defense and development of the Mediterranean mussel (*Mytilus galloprovincialis*). *Dev. Comp. Immunol.* **35**, 1017–1031. (doi:10.1016/j.dci.2011.03.023)
88. Lewin HA *et al*. 2018 Earth BioGenome Project: sequencing life for the future of life. *Proc. Natl. Acad. Sci. USA* **115**, 4325–4333. (doi:10.1073/pnas.1720115115)