RESEARCH ARTICLE

Open Access

A chromosome-level genome for the nudibranch gastropod *Berghia stephanieae* helps parse clade-specific gene expression in novel and conserved phenotypes

Jessica A. Goodheart[1,2*], Robin A. Rio[3], Neville F. Taraporevala[2,4], Rose A. Fiorenza[2], Seth R. Barnes[2], Kevin Morrill[2], Mark Allan C. Jacob[2], Carl Whitesel[2], Park Masterson[2], Grant O. Batzel[2], Hereroa T. Johnston[2], M. Desmond Ramirez[5,6], Paul S. Katz[5] and Deirdre C. Lyons[2*]

Abstract

Background How novel phenotypes originate from conserved genes, processes, and tissues remains a major question in biology. Research that sets out to answer this question often focuses on the conserved genes and processes involved, an approach that explicitly excludes the impact of genetic elements that may be classified as clade-specific, even though many of these genes are known to be important for many novel, or clade-restricted, phenotypes. This is especially true for understudied phyla such as mollusks, where limited genomic and functional biology resources for members of this phylum have long hindered assessments of genetic homology and function. To address this gap, we constructed a chromosome-level genome for the gastropod *Berghia stephanieae* (Valdés, 2005) to investigate the expression of clade-specific genes across both novel and conserved tissue types in this species.

Results The final assembled and filtered *Berghia* genome is comparable to other high-quality mollusk genomes in terms of size (1.05 Gb) and number of predicted genes (24,960 genes) and is highly contiguous. The proportion of upregulated, clade-specific genes varied across tissues, but with no clear trend between the proportion of clade-specific genes and the novelty of the tissue. However, more complex tissue like the brain had the highest total number of upregulated, clade-specific genes, though the ratio of upregulated clade-specific genes to the total number of upregulated genes was low.

Conclusions Our results, when combined with previous research on the impact of novel genes on phenotypic evolution, highlight the fact that the complexity of the novel tissue or behavior, the type of novelty, and the developmental timing of evolutionary modifications will all influence how novel and conserved genes interact to generate diversity.

Keywords Novelty, Lineage-restricted genes, Nudibranchia, Gastropoda, Differential gene expression

*Correspondence:
Jessica A. Goodheart
jgoodheart@amnh.org
Deirdre C. Lyons
d1lyons@ucsd.edu
Full list of author information is available at the end of the article

Background

One major question in biology is how novel phenotypes originate from conserved genes, processes, and tissues. Research in evolutionary developmental biology often focuses on the conserved modules that have been co-opted for new phenotypes, so-called toolkit genes [1–3]. This approach has also been used to investigate the evolution of homologous adult phenotypes, such as in studies of sensory system evolution (e.g., G protein-coupled receptors [4]). However, a conservation-based approach explicitly excludes genetic elements that may be classified as clade-specific (i.e., taxonomically restricted, lineage-specific, lineage-restricted, or clade-restricted) that contribute to the development or function of a particular phenotype [1, 5–7]. Here we present the chromosome-level genome for the gastropod mollusk, *Berghia stephanieae* (Valdés, 2005) [8], which we used to identify clade-specific genes that may be important for both novel and conserved phenotypes, but have largely remained under investigation.

Many clade-specific genes are known to be involved in novel, or clade-restricted, phenotypes, including a number of cnidarian-specific genes exclusively expressed in specialized cell types called cnidocytes [9–11]; the spiralian-specific gene trochin, expressed in the primary ciliated band [12]; and spidroins in spiders, used for creating spider silk [13, 14], among others (further examples in [5]). A recent review by Wu and Lambert [5] highlighted that clade-specific genes deserve more attention when investigating evolutionary novelties.

In addition to their value in understanding phenotypic novelties, clade-restricted genes also play important roles in what might be considered more conserved phenotypes, and can quickly become essential to the viability of the organism (e.g., *Drosophila*, [15]). These clade-specific genes can become integrated into more conserved systems via some version of system drift (e.g., developmental system drift [16]), which may not result in a drastic change in function that we would classify as an evolutionary novelty. Most research on the impact of clade-specific genes has focused only on the presence or absence of clade-specific gene expression in novelties and how those genes have evolved [17–20], but has not provided comparisons with more conserved phenotypes in the same organism. A few studies have identified more clade-specific gene expression in novel tissues or cell types compared to those that are more conserved [21, 22], moving us closer to appreciating how clade-specific genes—and their expression—impact phenotypic evolution. Based on these previous studies, we can hypothesize that clade-specific genes are likely to be disproportionately upregulated in novel tissues. However, much of this research has centered on well-studied model systems, which limits our ability to generalize across other metazoan lineages.

Investigations into clade-specific genes in understudied groups or phenotypes have a high potential for generating exciting new hypotheses or expanding our technical creativity. Multiple excellent examples of this potential come from the phylum Mollusca, a clade containing taxa such as snails and slugs, cephalopods, bivalves, and chitons. Mollusca is the second most speciose metazoan phylum (after Arthropoda) and contains a great diversity of phenotypes that have already provided many useful insights, including cephalopods as alternative models to vertebrates for the evolution of complex brains and intelligence [23–25], bivalves as a means of understanding the nature of transmissible cancer [26], and gastropods for neuroscience research [27, 28] and as models of parasitism and immunity [29, 30]. However, a lack of genomes and functional biology resources for many members of this phylum has long hindered assessments of genetic homology and function [31]. This lack of resources has limited our ability to even identify clade-restricted genes in mollusk lineages, let alone characterize their expression or test their impact on phenotypes of interest. Luckily, some genomic resources, such as transcriptomes, are being sequenced at a much higher rate than whole genomes [32]. We propose that these resources can also be used to more accurately infer whether genes are clade-specific, so that we might further characterize their impact on the evolution of novel phenotypes.

In this paper, we present a chromosome-level genome for *Berghia* that we use to investigate clade-specific gene expression across multiple tissues. *Berghia stephanieae* (hereafter referred to as *Berghia*) is a species of gastropod in the order Nudibranchia, a clade of marine slugs that lose their shell during metamorphosis [33]. This species has been used as a model for the study of both more conserved systems, such as neurodevelopment [34] and reproductive development [35], as well as clade-restricted phenotypes such as the sequestration of cnidarian nematocysts [36, 37] and endosymbiosis [38, 39]. We combined an inferred proteome from *Berghia* with available genome and transcriptome data from other metazoan species—including mollusks such as cephalopods, bivalves, and other gastropods—to identify clade-specific *Berghia* genes (i.e., restricted to Mollusca, Gastropoda, Heterobranchia, Nudibranchia, Aeolidina, or *Berghia* alone) (Fig. 1). We then describe expression profiles of clade-specific and non-clade-specific genes among adult tissue samples in *Berghia,* including more ancient tissue types (like nervous system tissues; [40]) and clade-restricted ones (such as those associated with nematocyst sequestration; [41]). We find highly upregulated genes that are clade-specific in every tissue

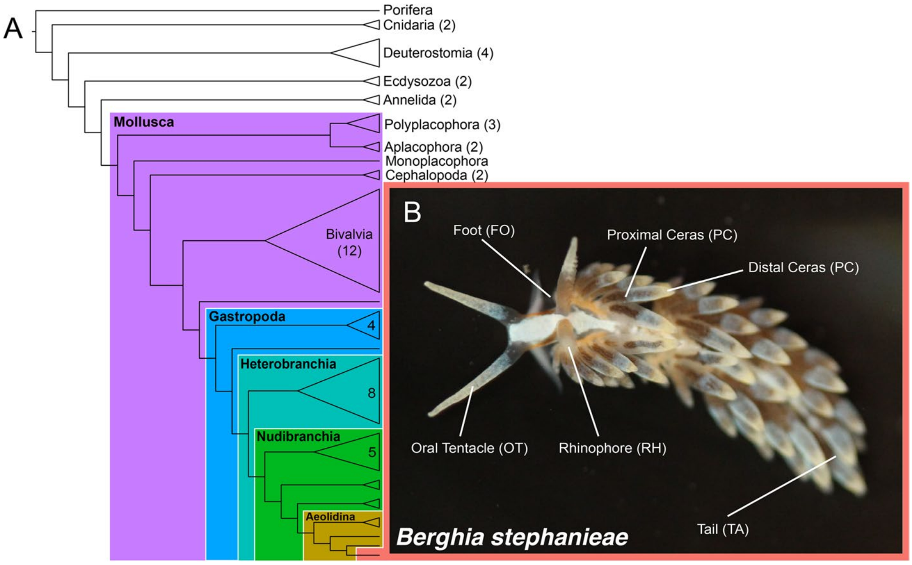

Fig. 1 Cladogram showing broadly where the nudibranch *Berghia stephanieae* falls in the metazoan phylogeny. Colors indicate clades that we investigated for clade-specific genes in the *Berghia* genome: Mollusca (purple); Gastropoda (blue); Heterobranchia (teal); Nudibranchia (green); Aeolidina (gold); and (B) *Berghia stephanieae* (salmon). External tissues used in our analyses are indicated in (B). Major clades for outgroups included in our analyses are also shown, with the numbers in each collapsed clade or next to each name indicating the number of species from that group included in our analysis

investigated, some of which are also highly upregulated in the same tissues during development. The proportion of clade-specific genes upregulated varied across tissues, but with no clear trend between the proportion of clade-specific genes and the novelty of the tissue. For example, the Aeolidina-specific distal ceras did not express a significantly higher proportion of clade-specific genes compared to more "conserved" tissues, such as the foot or tail. Our results support previous assertions that clade-specific genes are important for our understanding of phenotypic evolution, and highlight that future studies in emerging model systems must account for these to truly describe how new phenotypes evolve.

Results

A high-quality *Berghia stephanieae* genome

We constructed a highly contiguous chromosome-level assembly of the 1.05 Gb long *Berghia stephanieae* genome using PacBio long read sequencing (16,476,079 total reads, 166,997 Gigabases, mean read length ~10 kb, N50 = 15,977) that corresponds to ~160x theoretical coverage based on the final assembly size. The assembly was scaffolded with Omni-C Illumina short reads (Additional file 1). Post-assembly but pre-scaffolding, the assembly was 1.1 Gb in length with a contig N50 of 6,921,793 bp and L50 of 46 contigs (Additional file 1). Our scaffolded assembly contained 7945 scaffolds (Table 1) with an N50of 86 Mb (L50 = 5 scaffolds) and N90 of 34 Mb (L90 = 15) and is available through NCBI (Genome Accession JAWQJI000000000). The rest of the scaffolds (7930 scaffolds) encompass less than 8% of the original genome assembly. A GenomeScope 2.0 [42] analysis of our Omni-C data indicated relatively low heterozygosity (0.693%) compared to many other mollusks, including gastropods such as *Elysia chloritica* (3.66% heterozygosity [43]), and many bivalves (1–3% [44–47]). The K-mer spectra and fitted models for *Berghia* are also consistent with a diploid genome (Additional file 2: Fig. S1). The final genome, filtered by GC content, BLASTn hits, and sequence length (see methods for details), contains 18 scaffolds (Table 1) with an N50 length of 85.5 Mb (L50 = 5 scaffolds) and N99 length of 26.7 Mb (L99 = 15 scaffolds). We found 93.3% complete, and 95.9% complete+fragmented, BUSCO core genes represented from the Metazoa (odb10) BUSCO database in the final dataset (Additional file 2: Fig. S2), and 76.6% of PacBio reads map to the final assembly (compared to 77.2% in the unfiltered assembly; Table 1). The 18 scaffolds in the final *Berghia* genome likely represent 15 chromosomes, based on the length distribution of the scaffolds and the linkage map (Fig. 2 A, B) from our Omni-C analyses. This is on the high end of the range for nudibranchs, which are known to have between 12 and 15 chromosomes in their haploid genomes [48]. All further analyses were performed on these 18 scaffolds.

The *Berghia* genome compares favorably to other mollusk genomes in NCBI, with both a very high BUSCO score (when compared to the metazoa_odb10 database)

Table 1 Genome assembly statistics for the *Berghia stephanieae* genome initial assembly (pre-filtering) and final assembly (post-filtering)

	Pre-filtering	Post-filtering
Span (Gb)	1.1	1.05
No. of Scaffolds	7945	18
Scaffold L50	5	5
Scaffold N50 (Mb)	86	86
Scaffold L90	15	12
Scaffold N90 (Mb)	34	44
BUSCO score (metazoa_odb10)	93.6% complete [0.9% duplicated], 2.7% fragmented, 3.7% missing	93.3% complete [0.6% duplicated], 2.6% fragmented, 4.1% missing
PacBio CLR Reads Mapped	12,727,971 (77.2%)	12,622,211 (76.6%)

and scaffold N50 (Fig. 2C; Additional file 3: Table S1). This analysis includes genomes from NCBI classified as either "scaffold," "chromosome," or "complete" from the phylum Mollusca. Of the 190 non-*Berghia* genomes analyzed, only 23 have a higher scaffold N50, the majority of which are larger genomes from cephalopods, bivalves, and scaphopods. Five are other gastropods, including the caenogastropods *Sinotaia purificata*, *Conus ventricosus*, and *Monoplex corrugatus* and the patellogastropods *Patella vulgata* and *P. pellucida*, three of which have higher BUSCO scores. Overall, 61 of the 190 species have higher BUSCO completeness scores, but the *Berghia* genome falls within a cluster of the most contiguous and highest-quality genomes on NCBI (Fig. 2C).

The *Berghia stephanieae* genome is also well-annotated. Our RepeatModeler analysis identified 46.68% of the genome is repetitive elements, with the majority of bases characterized as unclassified repeats (27.45%, further details in Additional file 4: Table S2). BRAKER2 initially predicted 61,662 proteins (covering 59,494 genes; Table 2), which we subsequently filtered to a data set of 26,595 proteins and 24,960 genes for annotation and analysis using a script included with the BRAKER installation (selectSupportedSubsets.py) and the --anySupport flag to only include genes at least partially supported by hints. Prediction filtering resulted in a slightly lower BUSCO score (for both the Metazoa and Mollusca databases; Metazoa—87.2 to 86.0% complete, Mollusca—73.8 to 72.2% complete), though both scores were lower than the original BUSCO result using the whole genome, suggesting that gene predictions from BRAKER2 are incomplete. Prediction filtering also very slightly lowered IsoSeq mapping percentage (95.57 to 95.23% mapped reads), but did not change the percentage of short reads mapped to gene models (74.81% for both). Functional prediction rates, however, were much improved in our filtered data set, for both BLASTP hits (20,820 proteins, 78.3%, in filtered predictions) and InterProScan results (24,469 proteins, 92.0%, in filtered predictions) compared to 58.8% and 78.7%, respectively, in our initial predictions (Table 2). We used our filtered BRAKER2 predictions and functional annotations in subsequent analyses.

Identification of clade-specific genes in *Berghia* genome

Our OrthoFinder analysis compared the predicted *Berghia stephanieae* proteins from BRAKER2 with proteomes from 58 other metazoan species (Additional file 5: Table S3 [24, 49–80]), including 27 gastropods, one scaphopod, 12 bivalves, two cephalopods, one monoplacophoran, 3 polyplacophorans, two aplacophorans, and 11 non-molluscan species. The goal of this analysis was to generate orthologous groups among proteins from all proteomes to assess which *Berghia* genes are restricted to certain clades. It is important to note that genes classified as restricted to narrower taxonomic designations (e.g., Gastropoda) are also by definition restricted to higher taxonomic clades (e.g., Mollusca). We found 25,338 (95.2%) *Berghia stephanieae* proteins clustered into orthogroups, 1027 (3.9%) of which were in *Berghia*-specific clusters of two or more sequences.

Our KinFin analysis, which provides taxon-aware annotation of inferred orthologous groups, identified *Berghia* genes restricted to Mollusca (*n* = 1067, 4.3% of genes), Gastropoda (*n* = 463, 1.9% of genes), Heterobranchia (*n* = 1154, 4.6% of genes), Nudibranchia (*n* = 1030, 4.1% of genes), and Aeolidina (*n* = 108, 0.4% of genes), as well as those genes only found in *Berghia* (*n* = 2188, 8.8% of genes; Fig. 3A). These *Berghia*-specific genes include those clustered in *Berghia*-specific clusters using OrthoFinder in addition to singletons that were not clustered. This level of species-specificity is on the lower end compared to other nudibranchs and fairly average for mollusks more broadly (Additional file 2: Figs. S3-S4) This means that 18,957 *Berghia* genes (75.9% of genes)

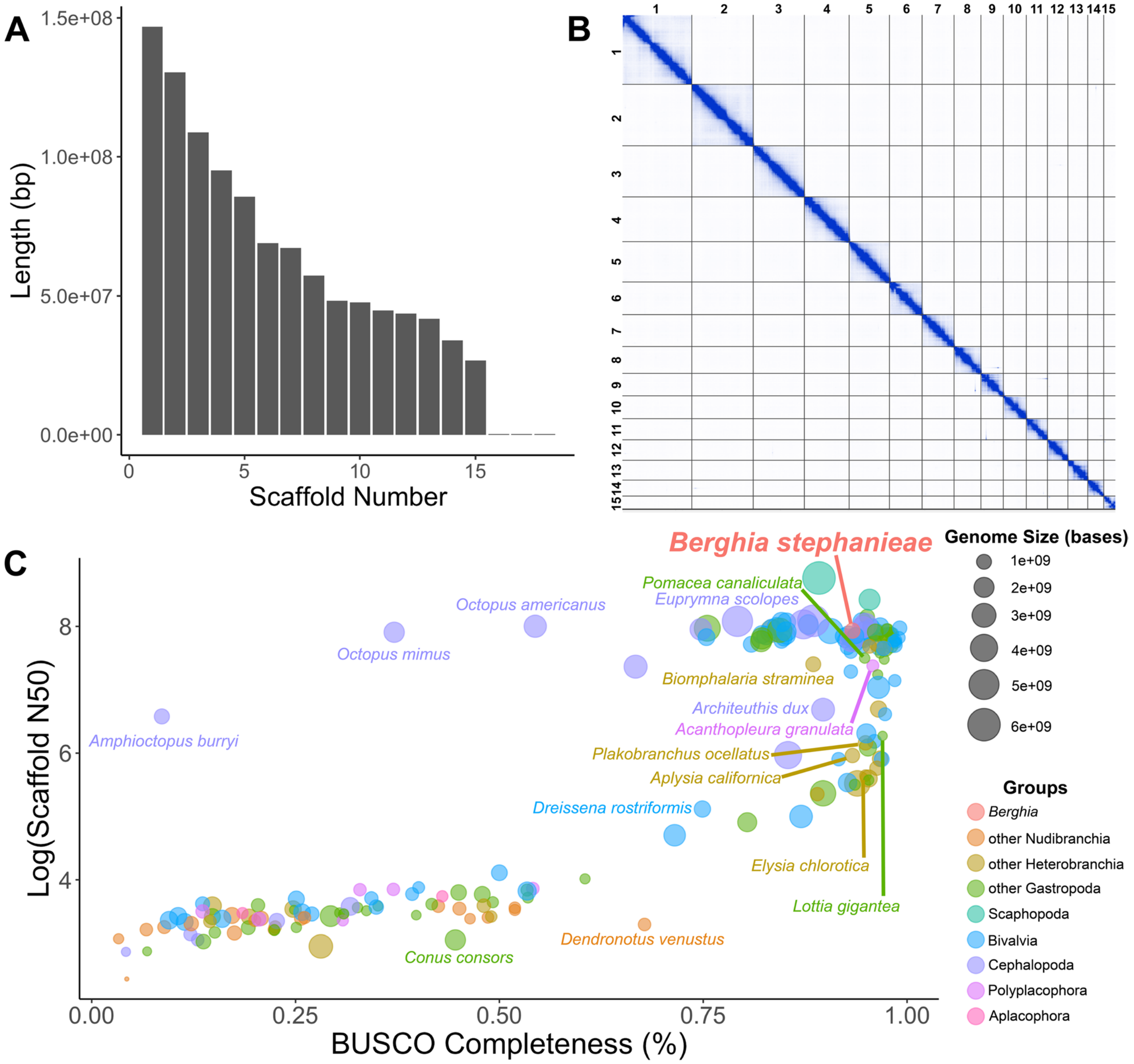

Fig. 2 Summary statistics for the chromosome-level *Berghia stephanieae* genome, with comparisons to other mollusk genomes. **A** Bar chart showing the length (in bp) for each of the retained scaffolds, including 15 putative chromosomes. **B** Hi-C linkage plot showing identified links within and among chromosomes. Darker color of a block indicates a higher frequency of contacts. **C** Plot visualizing the summary statistics for the *Berghia stephanieae* genome compared to assembled genomes of other mollusks by BUSCO completeness score (to the metazoa_odb10), log of the scaffold N50, and total genome assembly length. The *Berghia* genome is nearly average in size but is among the best genomes in terms of contiguity (scaffold N50) and BUSCO completeness score

are not clade-specific at the levels we investigated. Rarefaction curves for each taxonomic level (Additional file 2: Figs. S5-S9) suggest sufficient sampling in all major clades investigated. Both clade-specific and non-clade-specific (Other) genes are well distributed across the genome (Fig. 3B), with higher numbers of both types of genes per Mb in chromosomes 3 and 9 compared to other chromosomes (Additional file 2: Fig. S10). The proportion of clade-specific genes appears to be enriched on smaller chromosomes (Additional file 2: Fig. S11).

As expected, non-clade-specific genes (Other, Fig. 3C) were much more likely to be annotated via UniProt (79.9%) than genes identified as clade-specific. As we get phylogenetically deeper, the percentage of clade-restricted genes that are annotated increases: *Berghia*-specific (7.8% of genes annotated), Aeolidina-specific

Table 2 Gene prediction and annotation statistics for the *Berghia stephanieae* genome, including initial gene models predicted from BRAKER (pre-filtering) and filtered gene models intended to include only those models with external support (post-filtering)

	Initial prediction	Filtered predictions
Gene models—BRAKER2		
No. of genes	59,494	24,960
Average gene length (AA)	327.13	441
No. of predicted proteins	61,662	26,595
% start codon	99.85%	99.73%
% stop codon	99.91%	99.82%
BUSCO results—protein		
Metazoa odb10	*C*:87.2%[*S*:82.9%,*D*:4.3%],*F*:8.6%,*M*:4.2%,*n*:954	*C*:86.0%[*S*:81.8%,*D*:4.2%],*F*:8.7%,*M*:5.3%,*n*:954
Mollusca odb10	*C*:73.8%[*S*:67.3%,*D*:6.5%],*F*:6.5%,*M*:19.7%,*n*:5295	*C*:72.2%[*S*:65.9%,*D*:6.3%],*F*:6.2%,*M*:21.6%,*n*:5295
% reads aligned		
Bulk RNA-seq	74.81%	74.81%
IsoSeq	97.75%	95.23%
Gene model functional annotations		
BLASTP annotations		
Predicted genes annotated	34,594 (58.1%)	19,425 (77.8%)
Predicted proteins annotated	36,259 (58.8%)	20,820 (78.3%)
InterProScan annotations		
Predicted genes annotated	46,535 (78.2%)	22,914 (91.8%)
Predicted proteins annotated	48,501 (78.7%)	24,469 (92.0%)

(17.5% of genes annotated), Nudibranchia-specific (24.2% of genes annotated), Heterobranchia-specific (24.2% of genes annotated), Gastropoda-specific (32.0% of genes annotated), and Mollusca-specific (39.8% of genes annotated). The proportion of matched but uncharacterized genes (including hypothetical proteins) ranged from 1.2% (of Heterobranchia-specific genes) to 18.1% (of Gastropoda-specific genes) across investigated clades.

Clade-specific gene expression across tissues

Bulk tissue RNA-seq data mapped to the genome

We included seven *Berghia* tissues in our differential expression analyses, including (1) the brain, which consists of the paired cerebral-pleural, pedal, buccal, and rhinophore ganglia; (2) rhinophores, which are chemosensory structures restricted to nudibranchs; (3) oral tentacles, which are sensory-motor appendages restricted to the clade Aeolidina; (4) distal cerata, which are also restricted to the nudibranch clade Aeolidina and contain the organ where nematocyst sequestration occurs; (5) proximal cerata, which are common in Aeolidina and a few other nudibranchs and contain branches of the digestive system; and (6) foot; and (7) tail, which are tissues associated with mollusks more broadly. Our RNA-seq samples ranged from 2.9 million (SRR14337001, brain) to 36.5 million (SRR12072210, oral tentacle) read pairs ($\bar{x}$ = 25.3 ± 8.6 million reads per sample; Additional file 6: Table S4). On average, 72.1% (± 7.8%) of read pairs mapped uniquely to the *Berghia stephanieae* genome, and the mapping percentage ranged from 53.4% of read pairs (SRR14337002, brain) to 80.6% of read pairs (SRR12072207, distal ceras). We identified expression (counts >10 across all tissues [81]) in ~97.0% of genes (24,178 out of 24,960 predicted genes).

Differential expression among Berghia tissues

Our differential expression analyses compared the expression of each gene in each tissue to an average of normalized expression of that gene across all other tissues. Genes with a Log2 Fold Change > 2 and adjusted p-value < 0.05 were considered upregulated in a given tissue. We identified 16,691 genes upregulated across all tissues, with the highest number of upregulated genes (Table 3; Additional file 7: Table S5) in brain tissue (15,210 genes), followed by proximal ceras (678 genes), foot (205 genes), distal ceras (188 genes), tail (169 genes), rhinophore (147 genes), and oral tentacle (94 genes), respectively. The proportion of upregulated genes that were clade-specific (Fig. 4A; Additional file 7: Table S6) was variable across tissues, with rhinophore, oral tentacle, distal ceras, and tail having the most similar proportions of clade-specific upregulated genes (28.8–34.4% of upregulated genes),

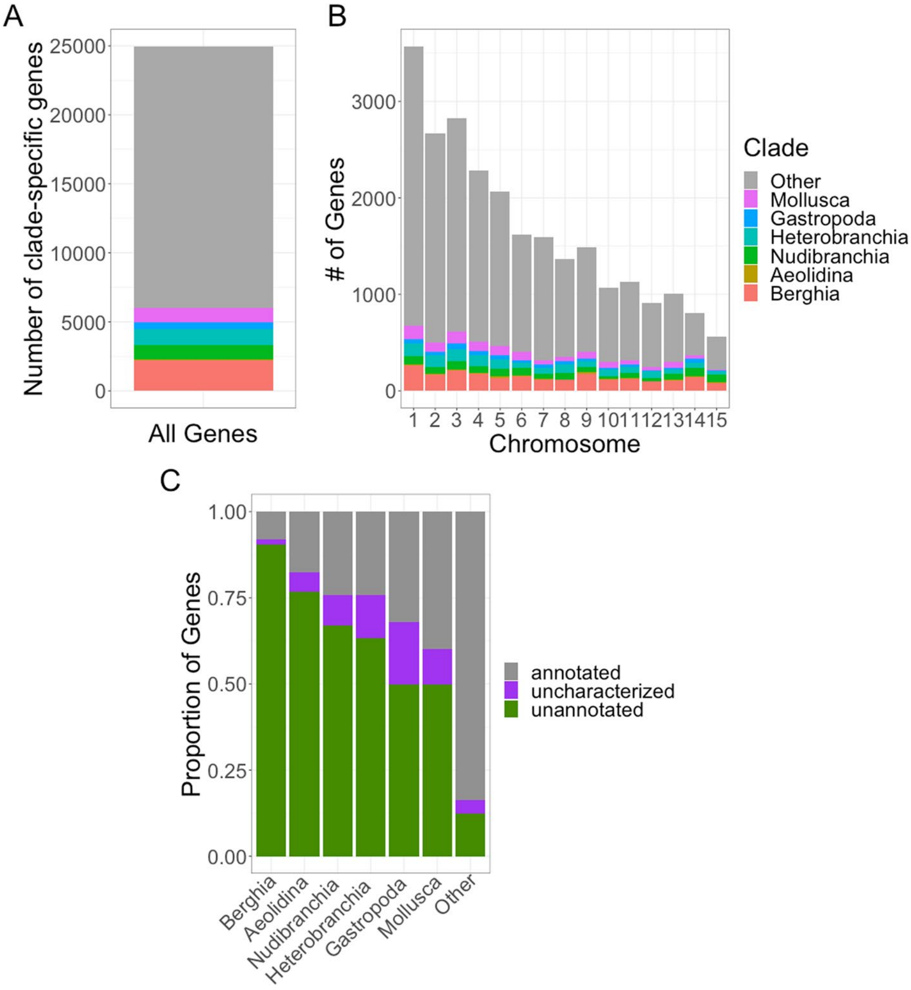

Fig. 3 Results for clade-specific genes found within *Berghia stephanieae*. **A** A bar chart showing the total number of genes in *Berghia stephanieae* parsed by whether they fall into one of the clade-specific groupings or not (Other). **B** A bar chart showing the distribution of genes across the genome, parsed by whether they are clade-specific or not (a size normalized version of this chart is available in Additional file 5: Fig. S5). **C** A bar chart indicating what proportion of genes within each group (clade-specific or Other) were annotated using BLASTP. In some cases, BLASTP found a match to an uncharacterized, hypothetical, or putative protein. These are separated into a different category (uncharacterized)

Table 3 Number upregulated genes for each tissue in *Berghia stephanieae*, by clade-specificity designation. Total numbers of upregulated genes are also reported for each tissue and clade

	Brain	Rhinophore	Oral tentacle	Distal ceras	Proximal ceras	Foot	Tail	Total
Berghia	623	18	8	31	83	15	17	795
Aeolidina	52	2	1	1	3	1	1	61
Nudibranchia	520	18	14	12	93	13	18	688
Heterobranchia	536	14	12	9	35	13	13	632
Gastropoda	238	2	4	3	13	4	6	270
Mollusca	593	2	3	16	62	26	24	726
Other	12648	91	52	116	389	133	90	13519
Total upregulated	15210	147	94	188	678	205	169	16691

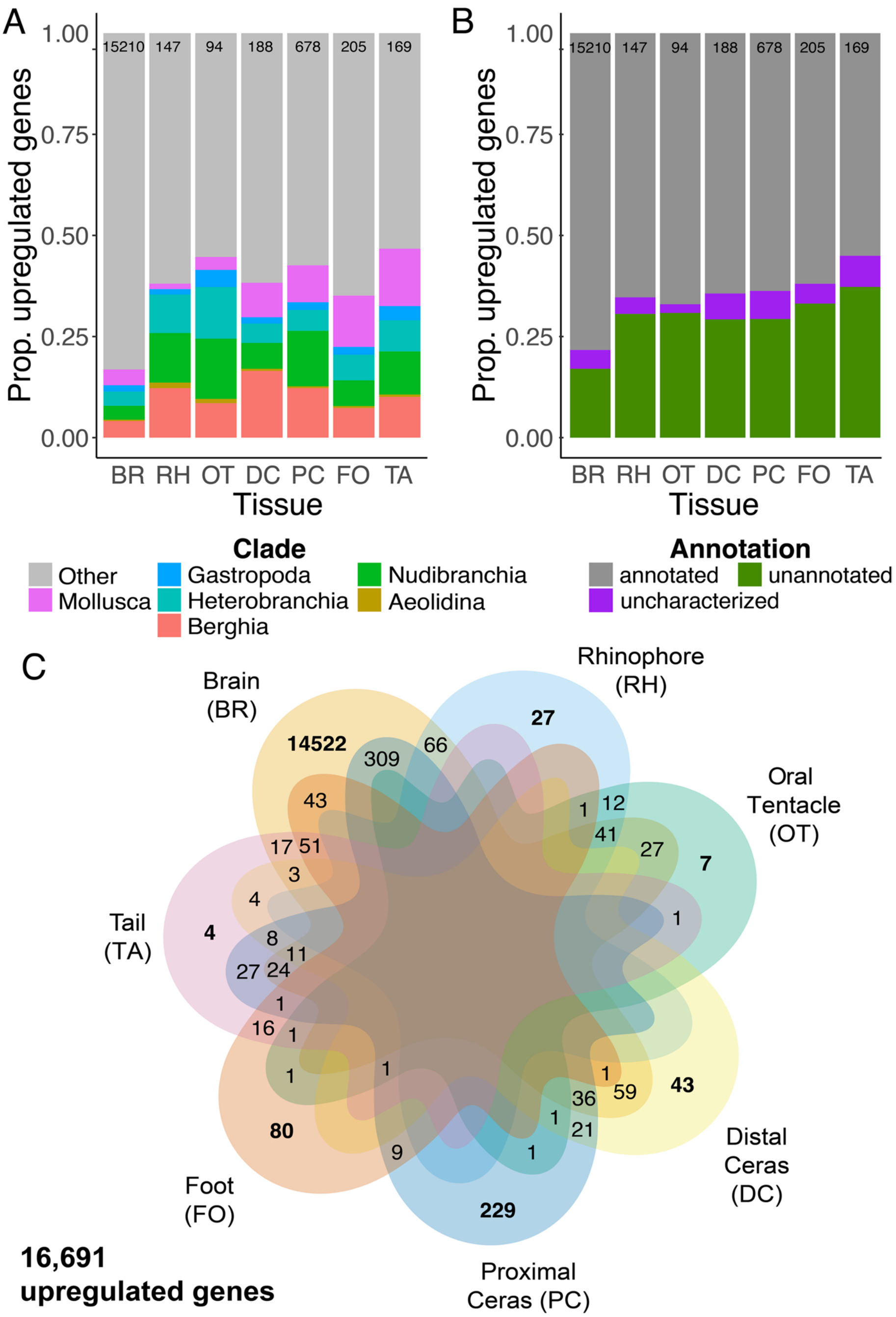

Fig. 4 Results for upregulated genes for each tissue found within *Berghia stephanieae*. **A** A bar chart showing the proportion of upregulated genes that are clade-specific, or not (Other), for each tissue type. **B** A bar chart indicating what proportion of genes upregulated for each tissue type were annotated using BLASTP. In some cases, BLASTP found a match to an uncharacterized, hypothetical, or putative protein. **C** Venn diagram showing the tissues in which all upregulated genes were determined to be upregulated. Some genes are only upregulated in certain tissues, while others are upregulated in multiple tissues (maximum of 4 out of 6, 11 genes upregulated among the brain, distal ceras, proximal ceras, and tail). These are separated into a different category (uncharacterized). Abbreviations: BR, brain; DC, distal ceras; FO, foot; OT, oral tentacle; PC, proximal ceras; RH, rhinophore; TA, tail

followed by foot (21.0% of upregulated genes). The proportion of clade-specific upregulated genes (Fig. 4A) was much lower in brain tissue (12.0% of upregulated genes). Expression data from clade-specific genes with more recent homologs (i.e., Gastropoda-*Berghia*) were more useful for distinguishing all tissues except for the brain

(Additional file 2: Figures S9-S14). Some genes upregulated in one tissue were also upregulated in another tissue (1779 genes, or 10.7% of upregulated genes; Fig. 4C).

We then removed genes that were upregulated in multiple tissues and focused only on those genes upregulated in a single tissue (uniquely upregulated genes; 14,912 genes, or 89.3% of upregulated genes). The highest number of uniquely upregulated genes (Fig. 4C was still in brain tissue (14,522 genes), followed by proximal ceras (229 genes), foot (80 genes), distal ceras (43 genes), rhinophore (27 genes), oral tentacle (7 genes), and tail (4 genes). The clade-specificity and annotation distributions of uniquely upregulated genes differed in some tissues (Fig. 5; Additional file 8: Tables S7-S8), including the rhinophore, oral tentacle, and tail when compared to the distributions of all upregulated genes in those tissues. The proportions of clade-specific and annotated genes changed more significantly in those tissues where far fewer genes were uniquely upregulated (Fig. 5).

With regard to annotation, we noted that in addition to having the highest number of upregulated genes, brain tissue also had the highest proportion of upregulated genes (78.3%) that were annotated via BLASTP (Fig. 5B; Additional file 9: Table S9). The other tissues had slightly lower levels of annotation ranging from 55.0% of upregulated genes (tail) to 67.0% of upregulated genes (oral tentacle) with multiples included, though for some tissues this proportion of upregulated, annotated genes dropped significantly when considering genes upregulated in only one tissue (Fig. 5B). Of the upregulated genes with annotations (Fig. 4B), GO term enrichment analyses were consistent with what might be expected for particular tissues (Additional file 10: Tables S10-S16). For example, we noted enrichment of signal transduction, transmembrane transport, and ion binding GO terms in the brain; G protein-coupled receptor activity and sodium ion transport terms in the sensory oral tentacles and rhinophores; and transmembrane transporter activity, transmembrane transport, and extracellular region terms in the distal ceras, where nematocyst sequestration occurs.

Confirmation of tissue-restricted expression of clade-specific genes

In order to confirm the expression of clade-specific genes inferred to be upregulated in certain tissues, we localized the expression of at least one gene from each of two tissues (rhinophores and distal ceras) using in situ hybridization chain reaction (HCR) techniques [82] in *Berghia* juveniles. We also compared our list of genes upregulated in the *Berghia* brain to HCR gene expression profiling in the brain of adult *Berghia* available from Ramirez et al. [83] Juveniles were selected for our experiments because adult *Berghia* contain pigments that make localizing expression more difficult, and both the

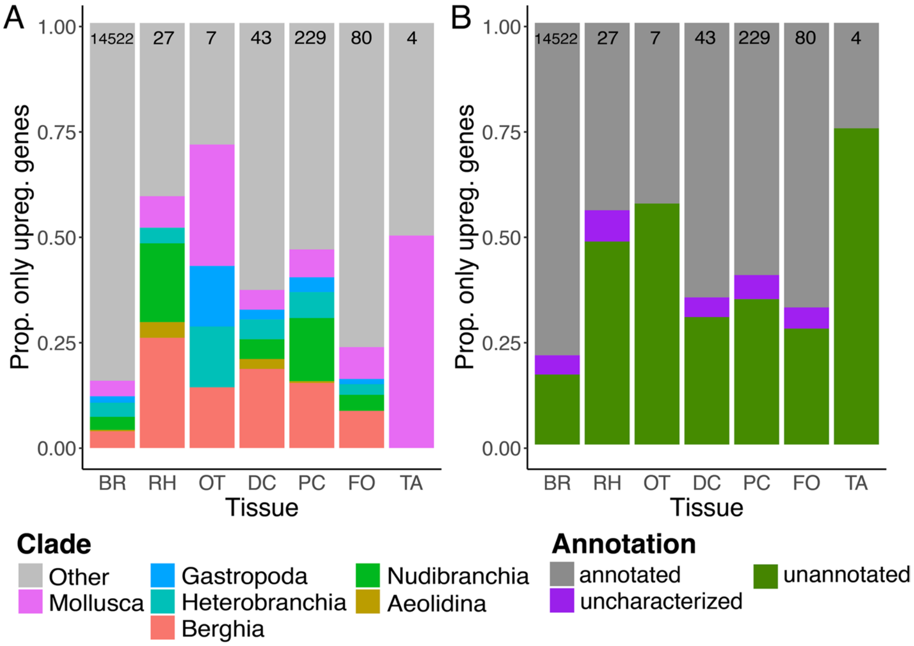

Fig. 5 Results for genes only found upregulated in a single tissue of *Berghia stephanieae* (see bold values in Fig. 4C). **A** A bar chart showing the proportion of upregulated genes that are clade-specific, or not (Other), for each tissue type. **B** A bar chart indicating what proportion of genes upregulated for each tissue type were annotated using BLASTP. In some cases, BLASTP found a match to an uncharacterized, hypothetical, or putative protein. Abbreviations: BR, brain; DC, distal ceras; FO, foot; OT, oral tentacle; PC, proximal ceras; RH, rhinophore; TA, tail

distal ceras and rhinophores are identifiable and functional at an early juvenile stage [34, 36]. We localized a distal ceras upregulated Nudibranchia-specific gene (jg13556; annotated as collagen alpha-1, match to UniProt ID Q7LZR2, *e*-value = 0.000129) in juvenile distal cerata (Fig. 6A–A′′′). This gene appears to be exclusively expressed inside the cnidosac of the juvenile *Berghia*, where nematocyst sequestration is known to occur [36]. We also identified expression of a rhinophore upregulated *Berghia*-specific gene (small domain annotated as a Pancreatic trypsin inhibitor, match to UniProt ID P00974, *e*-value = 1.52E−07) in the juvenile rhinophores where it is expressed in patches on the external epithelium (Fig. 6B–B′′′). For the brain, we found numerous clade-specific genes are expressed in the *Berghia* brain based on single-cell RNA-seq data [83]. These include (1) two genes exclusively upregulated in the brain (jg44129, an unannotated *Berghia*-specific gene expressed in a cluster of glial cells, and jg54950, an unannotated Heterobranchia-specific gene upregulated in nitric oxide synthase (*Nos*) and pigment dispersing factor (*Pdf*)-expressing cells in the rhinophore ganglia (*rhg*)); (2) one gene upregulated in the brain, rhinophore, and oral tentacle in our analysis (jg22847, an unannotated *Berghia*-specific gene in *Nos/Pdf- rhg* cells; and (3) two genes not considered upregulated in the brain in our analysis but appear to be upregulated in certain cell clusters in the brain (jg57406, an unannotated Heterobranchia-specific gene that is upregulated in the distal ceras in our analysis but expressed in mature neurons in all *Berghia* ganglia; jg56194, an unannotated Nudibranchia-specific gene that is not upregulated in any tissue in our analysis but is found in a cluster of glial cells in the brain) [83].

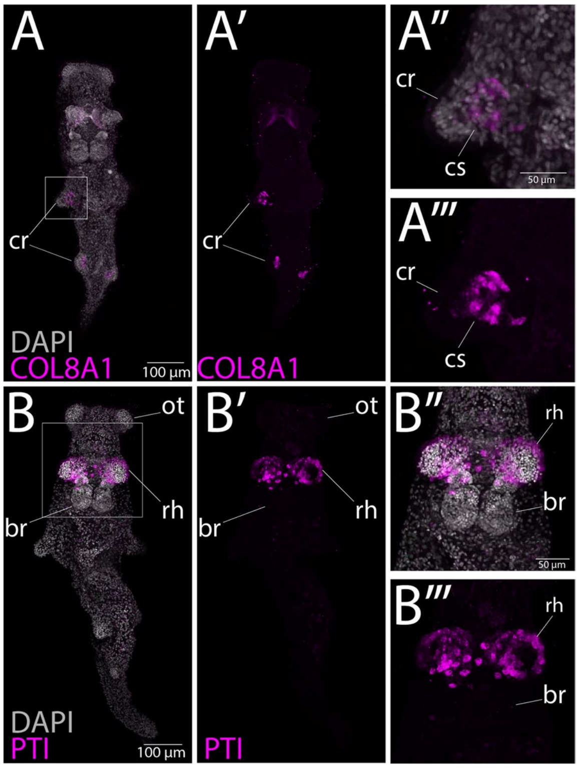

Fig. 6 HCR results in *Berghia stephanieae* juveniles for selected clade-specific genes upregulated in particular adult tissues. **A**–**A**′′′ Juveniles stained for a collagen alpha 1 VIII gene found upregulated in the distal ceras (COL8A1; jg13556) (**A**–**A**′) DAPI and Alexa 647 stained tissues in whole animal, and **A**′′–**A**′′′ close up of cnidosac. **B**–**B**′′′ Juveniles stained for a gene found upregulated in the rhinophores (annotated as a Pancreatic trypsin inhibitor, PTI; jg18351) (**B**–**B**′) DAPI and Alexa 647 stained tissues in whole animal, and **B**′′–**B**′′′ close up of rhinophore. Abbreviated: cr, cerata; cs, cnidosac; br, brain; rh, rhinophore; ot, oral tentacle

Discussion

The *Berghia stephanieae* genome is highly contiguous

The *Berghia stephanieae* genome is among the most contiguous and highest quality gastropod genomes to date (Fig. 2C). The final assembled and filtered *Berghia* genome is comparable to other mollusk genomes [56] in terms of size (1.05 Gb) and number of predicted genes (24,960 genes). The *Berghia* genome also has high Metazoa, and moderate Mollusca, BUSCO scores (Table 2), both comparable to the scores of other high-quality mollusk genomes [84, 85]. Our analysis also identified a high percentage of repetitive elements in the *Berghia* genome (46.68%), similar to rates found in other mollusk species [32]. However, given that only 76.6% of PacBio CLR reads mapped to the final genome assembly (Table 1), it is likely that many repeat regions remain unresolved. The proportion of annotated genes in the *Berghia* genome was also quite high (77.8% with BLASTP hits), consistent with other published gastropod genomes [86–88].

Clade-specific genes are a small percentage of the *Berghia* genome

The vast majority of predicted genes in the *Berghia* genome (18,957 genes, 75.9%; Fig. 3A) are not restricted to the clades of interest in our analysis (Mollusca, Gastropoda, Heterobranchia, Nudibranchia, Aeolidina, and *Berghia*). Of those identified as clade-specific, most (2188 genes, ~8.8%) were classified as *Berghia*-specific genes. This percentage of "orphan" or species-specific genes in *Berghia* is on the lower end of the range compared to other species from across Metazoa (~1 to >30%; [7]), which may simply be a feature of the *Berghia* genome. This may also be because our strategy to perform orthologous gene inference with a clustering-based method using both unannotated transcriptome data and well-annotated, high-quality genome data increased our chances of detecting homologs to *Berghia* genes. Clustering-based algorithmic approaches for inferring gene orthology apply normalization to pairwise similarity scores to account for the sequence length of the query and the length of its hits, ensuring that distantly related sequences receive comparable scores compared to the best-scoring sequences from closely related species [89–91]. This strategy, combined with more data from more closely related taxa, provides greater opportunity for similarity matches and limits the impacts of homology detection failure [92], which occurs when homologs have become undetectable by search algorithms even though they exist. Our analysis suggests that homology detection failure is likely to cause inflated estimates of the proportion of species-specific genes in analyses that (1) include species from less widely distributed taxonomic levels (e.g., only including closely related taxa [19] or having limited outgroup sampling [93]); (2) rely only on pairwise similarity scores for assessing homology [21, 93]. In our analysis, this could mean that a subset of the ~8.8% of *Berghia*-specific genes are likely to be restricted to clades that we have not explicitly investigated, such as those at the genus (*Berghia*) or family (Aeolidiidae) levels. This would mean that the actual percentage of species-specific *Berghia* genes is perhaps even lower than reported here.

As might be expected, we also note that the annotation rate goes down with an increase in clade specificity of the genes, meaning that non-clade-specific genes ("Other") had the highest annotation rate and *Berghia*-specific genes the lowest (Fig. 3C). However, some *Berghia* genes matched to proteins in the NCBI RefSeq database that do not have functional data associated with them (i.e., are predicted, hypothetical, or unidentified proteins; purple color in Fig. 3C). It is possible that some of these putative annotations are due to protein domain-level similarities, which suggests that these clade-specific genes do share some homologous regions with genes in other taxonomic groups.

Novel tissues do not express more novel genes

Clade-specific genes have long been thought to be a driver of morphological novelties [5, 94], and some researchers have hypothesized that novel phenotypes might require a higher frequency of clade-specific genes [22, 94, 95]. This type of increase in the expression of clade-specific genes has been identified in some taxa. These include cnidarians like *Nematostella* in novel cells called nematosomes [22], the mollusk radula [96], and non-morphological novelties like eusocial evolution in honey bees [95]. Our results are inconsistent with this hypothesis. In *Berghia,* we did not see a clear increase in the proportion of upregulated clade-specific genes in morphological novelties (Fig. 4A, B). For example, the distal ceras, where nematocyst sequestration occurs inside a novel organ called the cnidosac [41, 97], appears to have an average level of clade-specific gene upregulation compared to other tissues (Fig. 4A–B). This may be due to the fact that nematocyst sequestration relies on a largely conserved process, namely phagocytosis [36, 41, 98]. Our results also did not show a higher proportion clade-specific genes upregulated in the sensory rhinophores or oral tentacles compared to other tissues, which are functionally similar and homologous to tentacles in other heterobranch gastropods and caenogastropods [34, 99, 100]. However, the clade-specific genes that are upregulated in some of these tissues do appear to be functionally important, as our HCR results indicate that some genes (as shown in Fig. 6) are not only upregulated in the adult tissues but are also highly expressed in those same tissues in early-stage juveniles. The expression of these genes in juveniles suggests that these clade-specific genes may be especially crucial for the core function of these tissues as soon as they form. Alternatively, these genes may be important for upstream developmental processes in addition to downstream functions.

So why are our results inconsistent with the hypothesis that clade-specific genes are likely to be upregulated in novel tissues? For one, we have only collected gene expression data under standard conditions, which means that we are largely capturing constitutive gene expression. It is possible that critical, clade-specific genes for certain tissues may be only expressed in response to certain stimuli. We have likely not captured many of these genes in our analyses. Second, it is known that clade-specific genes are involved in both novel and conserved phenotypes [15, 21]. However, investigating gene expression at the level of tissues likely masks hidden differences in cell type diversity and complexity within tissues. We hypothesize that differences in the number and expression levels of clade-specific genes among novel tissues may be more related to the type of novelty (i.e., functional vs morphological) rather than the level of biological organization. For example, novelties that arise from a loss or modification of function may only require the loss of expression of certain genes. Future investigations focused on the expression of clade-specific genes (facultative and constitutive) at the cellular level—and on the function of these genes—will provide the necessary data to assess the relative impacts of these possibilities.

Tissue complexity—the Berghia brain

Animal nervous systems are perhaps the most complex biological systems, with their diverse components, cell types, and functions (e.g., [101, 102]). Although neurons in general express more genes than many other cell types in a variety of metazoan lineages [103], it has also been noted that genetic novelties appear to underlie cell diversification in multiple lineages (including octopuses

[25] and teleost fish [104]). In these cases, the proportion of clade-specific genes expressed in novel cell types is higher than that in cell types with clear homologs. However, these differences would not be detectable in tissue-level analyses, such as those presented here.

Despite lower read counts and mapping percentages (Additional file 6: Table S4), we identified a large number of genes upregulated in *Berghia* brain tissue (> 15,000; Fig. 4A), which is roughly two-thirds of predicted *Berghia* genes. This is consistent with the hypothesis that neurons express more genes than other cell types [103]. However, our results also show lower proportions of clade-specific genes in the brain (~16.8%; Fig. 4A) compared to other tissues, even though the number of clade-specific genes upregulated in the brain (2,562 genes) is higher overall than the total number of upregulated genes in any other tissue (94–678 genes; Fig. 4B). It is possible then that consideration of clade-specific expression at the tissue level may not provide a comprehensive understanding of novelty.

Although some prior studies identified the expression of more clade-specific genes in novel tissues or structures [21], others have found a higher frequency of clade-specific expression in cell-type novelties [22, 105]. This suggests that the expression of clade-specific genes may have a higher impact on the evolution of novel cell types rather than novel tissues as a whole. This hypothesis is supported by the expression patterns of clade-specific genes in the *Berghia* brain [83]. For example, an unannotated gene that serves as a neuron-specific cell type marker is also a Heterobranchia-specific gene (jg57406). Similarly, serotonergic neurons in the brain express an unannotated gene that is Nudibranchia-specific (jg38442) [83].

Type of novelty—the Berghia cnidosac

The novel-genes-drive-innovation hypothesis might suggest that the novel cnidophage cell type, in the distal ceras [36], may use more (or a higher proportion of) clade-specific genes than other, more conserved cell types in *Berghia*. However, it was previously hypothesized that the cnidophage cell type may simply be a more specialized homolog of a digestive cell type, due to the apparent endodermal nature of cnidophages [36, 41, 97, 106]. A logical inference from this hypothesis is that the novel function of sequestration in cnidophages does not necessarily require novel molecular processes. This is not a new idea. For example, some morphological novelties in plants have been shown to evolve via regulatory evolution [107] and social behavior evolution in ants has been tied to both conserved and novel genetic elements [108]. Under similar conditions, we might not expect an increase in the frequency of clade-specific genes expressed in cnidophages. However, these clade-specific genes may still be functionally important given their high levels of expression in the cnidosac in *Berghia* juveniles (Fig. 6B–B′′).

Conclusions

The *Berghia stephanieae* genome is the first high-quality published genome for the order Nudibranchia and is one of the most contiguous and highest-quality gastropod genomes to date. However, it is likely that many repeat regions remain somewhat unresolved. We used this genome to investigate how clade-specific gene expression is distributed across functionally and evolutionarily diverse tissue types in adult *Berghia* and showed that upregulated genes in novel tissue types are not necessarily more likely to be classified as clade-specific. The proportion of clade-specific genes upregulated varied across tissues, with novel tissues like the distal ceras unexpectedly expressing a fairly average frequency of clade-specific genes compared to other tissues. Our results, when combined with previous research on the impact of novel genes on phenotypic evolution, highlight the value of a more holistic approach to investigating how phenotypes arise and diversify. In particular, the complexity of the novel tissue or behavior, type of novelty [22], and where across development changes may have occurred [109] will all influence how novel and conserved interact to generate new phenotypes.

Methods

Sample preparation and genome sequencing

We isolated one *Berghia* juvenile from the Lyons lab culture prior to mating to minimize genomic contamination. While isolated, we fed the animal ~½ of a medium *Exaiptasia diaphana* (defined by Taraporevala et al. [35]) each day for 34 days. We then starved the animal for 44 days prior to shipping. To minimize residual food in the gut diverticula, cerata were removed with forceps and the remaining body was blotted on a Kimwipe to remove excess water, then the animal was placed in a cryotube and flash frozen in liquid nitrogen and stored at −80 until shipping to Dovetail Genomics (now Cantana Bio, Scotts Valley, CA). Dovetail Genomics used an input of ~101 mg into a slow CTAB protocol to extract high molecular weight DNA. They measured the efficiency of DNA extraction using a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) High Sensitivity Kit. Overall, they obtained 12.1 ug of high molecular weight DNA. They then used a Mini Column for cleanup and resuspended the pellet in 75 µl TE. They then quantified DNA samples using the Qubit. They constructed the PacBio SMRTbell library (~20kb) for PacBio Sequel using SMRTbell Express Template Prep Kit V 2.0 (PacBio, Menlo Park, CA, USA) using the manufacturer-recommended protocol. They then bound the library to polymerase using the Sequel II Binding Kit 2.0 (PacBio) and loaded onto PacBio Sequel II (PacBio) on 8M SMRT cells (SRR25687008).

For scaffolding, Dovetail fixed chromatin in place with formaldehyde in the nucleus for extraction and analysis via Dovetail® Omni-C® proximity ligation. They then digested the fixed chromatin with DNAse I, repaired the chromatin ends, and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter-containing ends. After proximity ligation, they reversed crosslinks and purified the DNA. They treated purified DNA to remove biotin that was not internal to ligated fragments. They generated sequencing libraries using NEBNext Ultra enzymes and Illumina-compatible adapters. They then isolated biotin-containing fragments using streptavidin beads before PCR enrichment of each library. Technicians then sequenced the library on an Illumina HiSeqX platform to produce approximately 30× sequence coverage. They then used HiRise MQ>50 reads for scaffolding (see "read-pair" above for figures).

Short-read RNA sample collection and sequencing

We obtained *Berghia* adult tissue samples, including the (1) brain (2 samples; SRR14337001-SRR14337002); (2) oral tentacles (3 samples; SRR12072210, SRR25598600-SRR25598601); (3) rhinophores (3 samples; SRR12072209, SRR25598592-SRR25598593); (4) foot (2 samples; SRR12072206, SRR25598598); (5) tail (2 samples; SRR12072205, SRR25598599); and (6) proximal (3 samples; SRR12072208, SRR25598594-SRR25598595) and (7) distal ceras (3 samples; SRR12072207, SRR25598596-SRR25598597). We also obtained earlier stage transcriptome data from (8) multiple embryonic stages (bulk sample of 500–600 individual embryos from each time point reared at 27 °C (12, 24, 36, 48, 60 h post oviposition and 4, 7, and 9 d post oviposition; SRR12072213) and (9) juveniles 15 d post oviposition at 27 °C (500 individuals from 3 egg masses laid the same day; SRR12072212). We starved adults for ~1 week prior to the removal of some tissues (all but the brain) to reduce symbiont presence and minimize contamination. We extracted total RNA from most adult tissues (minus the brain) using the RNeasy Kit (QIAGEN, Redwood City, CA) and submitted the extracted total RNA to Novogene Ltd. (Sacramento, CA) for quality assessment, library preparation, and sequencing (Illumina NovaSeq 6000; 150bp paired-end reads). We prepared the adult brain total RNA using the Clontech SmartSeq v4 Ultra-Low Input RNA Kit (Takara). We prepared libraries with the Nextera XT DNA Library Preparation Kit and 96-Sample Index Kit (Illumina, San Diego, CA) and quantified them using Qubit (ThermoFisher Scientific, Waltham, MA) and assessed quality using a Bioanalyzer (Agilent, Santa Clara, CA). We sequenced the brain sample on the Illumina NextSeq 500 (75bp paired-end reads) at the Genomics Resource Laboratory, University of Massachusetts, Amherst. For the first two samples (bulk embryonic stages and juveniles), total RNA was extracted using TRIzol (Ambion) following the standard protocol, quality was assessed using Tapestation (Agilent) and sent to the IGM UCSD Genomic Center for library preparation (TruSeq mRNA stranded library) and sequencing (Illumina NovaSeq 6000; 150bp paired-end reads).

Reference transcriptome construction

Berghia samples used for reference transcriptome construction included a subset of samples to minimize computational cost while maximizing read breadth. These included single samples from multiple adult tissues, selected at random, including the (1) brain (SRR12072211), (2) oral tentacle (SRR12072210), (3) rhinophore (SRR12072209), (4) foot (SRR12072206), (5) tail (SRR12072205), and (6) proximal (SRR12072208) and (7) distal ceras (SRR12072207), as well as samples from (8) embryos (SRR12072213) and (9) juveniles (SRR12072212). We merged all FASTQ output files for the above samples into two files (Read 1 and Read 2) for downstream analysis. We used default parameters for all programs unless otherwise specified. We trimmed and filtered reads using fastp (version 0.20.0; [110]), and assembled transcripts using Trinity (version 2.9.1; [111]). We predicted open reading frames (ORFs) with TransDecoder (version 5.5.0; [112]). Duplication levels were quite high (~56%), so we clustered predicted ORFs using CD-HIT-EST (version 4.8.1; [113, 114]) at 95% identity and word size of 11 (-c 0.95, -n 11). Post-clustering, we filtered transcripts with alien_index (https://github.com/josephryan/alien_index), based on an algorithm described in [115]. We constructed alien index databases using previously constructed metazoan and non-metazoan databases (obtained from http://ryanlab.whitney.ufl.edu/downloads/alien_index) and all "*Symbiodinium*" sequences present on UniProt [116] as of 31 March 2020. We removed all sequences with an alien index greater than 45 from the transcriptome. We then compiled full transcripts for each predicted ORF sequence remaining from the assembled transcriptome using a custom Python script (full_transcripts.py). We then scanned the transcriptome for vectors and possible contaminants via the NCBI VecScreen (https://www.ncbi.nlm.nih.gov/tools/vecscreen/). We removed vectors using a small script (trim_adapters.pl) available through the Trinity Community Codebase (https://github.com/trinityrnaseq/trinity_community_codebase). We removed or trimmed sequences containing contamination using the

Contaminants.txt file provided by NCBI and a custom script (remove_contamination.py). Custom scripts are available at https://github.com/lyons-lab/berghia_reference_transcriptome). We assessed transcriptome quality across all steps using BUSCO v5.1.2 [117–119] scores by comparing assembled transcripts to the metazoa_odb10 (C:98.1%[S:83.1%,D:15.0%],F:0.9%,M:1.0%,n:954) and mollusca_odb10 (C:93.4%[S:76.9%,D:16.5%],F:1.3%,M:5.3%,n:5295) databases. This transcriptome was intentionally generated de novo, in the absence of input from the genome, to provide a completely independent proteome with which to assess potential repeats in the genome.

Long-read RNA sample collection and sequencing

We obtained *Berghia* adult tissue samples from animals starved for at least 4 weeks to minimize gut contaminants, including the (1) head (one animal), (2) oral tentacles (two animals), (3) rhinophores (three animals), (4) cerata (one animal), (5) mantle (one animal), and (6) homogenized mid-body tissue (one animal). We also collected two developmental samples, including (1) embryos from the trochophore (72 h post oviposition; 300 animals) and eyed veliger stages (9–10 days post oviposition; 120 animals), and (2) post-metamorphic and post-feeding juveniles (34 animals). We extracted total RNA using the standard TRIzol Reagent (Life Technologies, Carlsbad, CA, USA) protocol, with some modifications: After the addition of chloroform, we centrifuged samples for 20 min at max speed (16,000 RCF) and precipitated samples in 100% isopropanol for ~1 h at −20 °C. We assessed total RNA sample quality on a 1% agarose gel and quantified the RNA in each sample with a Qubit 2.0 High Sensitivity kit (ThermoFisher Scientific, Waltham, MA). We then pooled developmental stages (embryos and juveniles; DEV), adult rhinophore and oral tentacle samples (RHOT), and adult mantle and cerata samples (MCE) in equivalent amounts. We sent these five total RNA samples (DEV, MCE, RHOT, head, mid-body) to the Roy J. Carver Biotechnology Center at the University of Illinois at Urbana-Champaign for IsoSeq library construction (5 libraries) and sequencing. They performed sequencing (on the five pooled libraries) on a single SMRT 8M cell with the PacBio Sequel II (PacBio, Menlo Park, CA) and a 30-h movie. They then clustered the raw subreads using the IsoSeq v3 clustering workflow (https://github.com/PacificBiosciences/IsoSeq/blob/master/isoseq-clustering.md).

Genome assembly and scaffolding

Dovetail used 167 gigabase pairs of PacBio CLR reads as an input to WTDBG2 v2.5 [120] with genome size 2.0g, minimum read length 20,000, and minimum alignment length 8192. Additionally, they enabled realignment with the -R option and set read type with the option -x sq. They then used BLASTn results of the WTDBG2 output assembly against the nt database as input for blobtools v1.1.1, and scaffolds identified as possible contamination were removed from the assembly. Finally, they used purge_dups v1.2.3 [121] to remove haplotigs and contig overlaps.

Dovetail used input de novo assembly and Dovetail Omni-C library reads (3 samples; SRR25687005-SRR25687007) as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies [122]. They aligned Dovetail Omni-C library sequences to the draft input assembly using BWA with default parameters [123]. They then analyzed separations of Dovetail Omni-C read pairs mapped within draft scaffolds by Hi-Rise to produce a likelihood model for genomic distance between read pairs and used the model to identify and break putative misjoins, to score prospective joins, and make joins above a threshold.

We initially filtered the *Berghia stephanieae* genome with purge_dups v1.2.5 [121] to automatically identify and remove haplotigs and contig/scaffold overlaps from heterozygous sites. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAWQJI000000000. The version described in this paper is version JAWQJI010000000. Following duplicate purging, we assessed completeness with BUSCO v5.1.2 [117–119] by comparing to metazoa_odb10 and mollusca_odb10. BUSCO further used the programs HMMER v3.1 [124] and MetaEuk v4.a0f584d [125] for gene prediction and analysis. We then used Nucleotide-Nucleotide BLAST 2.11.0+ [126, 127] to compare our scaffolds to the nt database (downloaded April 2021) and mapped the original PacBio reads used for assembly via minimap2 v2.18-r1015 [128]. With these results, we used BlobToolKit (Challis et al. 2020) (blobtools2 filter option) to remove additional scaffolds considered contamination. Scaffold selection for removal was based on GC content, PacBio read coverage results, BLASTn hits (we removed no-hit and bacterial contamination), and finally a minimum size threshold (150 kb). This size threshold was selected because it was the point at which the removal of sequences would not change the BUSCO score, as determined by the use of BlobToolKit Viewer v1.1 [129]. Most removed sequences contained differences in GC content and coverage compared to those that were retained in the final annotated genome, in addition to their smaller size. We also performed a Nucleotide-Nucleotide BLAST 2.11.0+ to compare the removed sequences with the final genome to assess duplication rates. Of those sequences removed from the final genome, 92.8% hit to one of the final 18 scaffolds (98.7% of which with an e-value of

0.0), and 1.6% of removed sequences were obvious contaminants. To compare the *Berghia* genome with other available Mollusca genomes, we downloaded assembled genomes from NCBI using the datasets command line function (v.15.24.0)[130] with flags for the taxon Mollusca and genomes at assembly levels "scaffold," "chromosome," or "complete." We then assessed the completeness of each genome with BUSCO v5.1.2 compared to the database metazoa_odb10 and assessed scaffold N50 using the command line tool n50 in SeqFu, a suite of FASTX utilities [131].

Gene prediction and annotation

We analyzed filtered genome scaffolds with RepeatModeler v2.0.1 [132], which used Tandem Repeat Finder (TRF) v4.09 [133], RECON v1.08 [134], RepeatScout v1.0.6 [135], and RepeatMasker v4.1.2 (https://www.repeatmasker.org), to construct a de novo repeat library for *Berghia stephanieae*. This included the ---LTRStruct flag to run an LTR Structural Analysis, which used GenomeTools v1.6.1 (http://genometools.org), LTR_Retriever v2.9.0 [136], Ninja v1.10.2 (https://github.com/ninja-build/ninja), MAFFT v7.480 [137], and CD-HIT v4.8.1 [113, 114]. We then used this species-specific library to detect repeat sequences (via both soft and hardmasking) with RepeatMasker in the *Berghia* genome.

Following repeat masking, we mapped all short RNA-seq reads (unfiltered) to the hardmasked version of the genome using two separate read mapping software programs, HiSat2 v2.2.1 [138] and STAR v2.7.9a [139]. This was intended to account for mapping bias in order to maximize the possibility of support in our gene annotations. For long-read (IsoSeq) mapping, we first obtained Full Length Non-Concatemer (FLNC) reads from step three of the IsoSeq v3 workflow. These FLNC reads were mapped directly to the non-repeatmasked genome using minimap2 v2.22-r1105-dirty [128] with the recommended options according to PacBio (https://github.com/Magdoll/cDNA_Cupcake/wiki/Best-practice-for-aligning-Iso-Seq-to-reference-genome:-minimap2,-deSALT,-GMAP,-STAR,-BLAT). Post-mapping, sam output files were reformatted into bam files and indexed using samtools v1.11 [140]. We used BRAKER v2.1.6 [141–143] for preliminary gene prediction of the *Berghia* genome, which uses GeneMark-EP+ v4 [144, 145], DIAMOND v2.0.8 [146], spaln v2.4.3 [147, 148], and Augustus v3.4.0 [149]. We used long- and short-read RNA-seq mapping results as expression support input. We also used a protein hints file generated by combining the mollusca_odb10 database with *B. stephanieae* sequences identified as BUSCO hits from our initial mollusca_odb10 BUSCO run. We ran BRAKER with the additional flags --etpmode, --gff3, and --softmasking. After initial gene prediction, we generated a filtered predicted gene set using a script included with the BRAKER installation (selectSupportedSubsets.py) and the --anySupport flag to only include genes at least partially supported by hints. IsoSeq and short-read RNA-sequencing data were mapped to both sets of gene models to assess the impact of filtering. Both unfiltered and filtered gene prediction results are provided in Dryad (DOI: 10.6076/D1BS33), but we only used the filtered set (braker_annotations_anysupport.gff3) in subsequent functional annotation and clade-specific gene analyses.

For functional annotation of predicted genes, we used Protein-Protein BLAST 2.11.0+ (BLASTP) and InterProScan v.5.52-86.0. For BLASTP analyses, we used an *e*-value cutoff of 1e-3 with -max_target_seqs of one against three databases: (1) UniProtKB/Swiss-Prot, (2) RefSeq, and (3) Trembl (all downloaded April 2021). We then combined the hits to all three databases in a single blast annotation file. For InterProScan analyses, we used the default parameters with some additional flags, including -goterms to look up gene ontology, -dp to disable pre-calculated match lookup, and -t p to indicate protein sequences.

Assessment of clade-specific genes and their expression

To determine the distribution of clade-specific genes for *Berghia*, we created a proteome dataset containing 47 metazoan species using both published genomes and transcriptomes (Additional file 4: Table S3). Our final dataset included 36 mollusks, including fourteen nudibranchs (five from Anthobranchia and nine from Cladobranchia). We downloaded predicted proteomes from genome datasets from MolluscDB [56]. We downloaded the transcriptomes from the NCBI Sequence Read Archive (SRA), which we then filtered with fastp v0.20.0 [110] and assembled with Trinity v2.9.1 [111]. We predicted ORFs using Transdecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder), with default parameters. Our final proteomes ranged from 17,606 to 72,541 proteins ($\overline{x}$ = 33,691; Additional file 4: Table S3). We identified orthologous gene families among our metazoan proteomes using the OrthoFinder package (Emms and Kelly, 2019) with default parameters and a user-generated species tree as input (Additional file 11). Our user-generated species tree topology was based on recent metazoan phylogenies [150–153], the MolluscDB phylogeny provided on the website [56], and recent Mollusca [154] and nudibranch [67, 71, 155] phylogenetic analyses. We then analyzed orthologous groups using the program KinFin v1.0.3 [156] to determine which predicted genes in *Berghia stephanieae* are clade-specific (meaning that they only cluster with sequences from a particular clade). We used the default parameters, with additional flags

(--infer_singletons --plot_tree -r phylum,class,order,superfamily).

To determine the expression patterns of clade-specific genes, we mapped our short-read RNAseq data (from the brain, oral tentacles, rhinophores, foot, tail, and proximal and distal ceras) to the *Berghia* genome using STAR v2.7.9a [139] with default parameters plus additional flags (--readFilesCommand zcat --outSAMtype BAM SortedByCoordinate --twopassMode Basic --sjdbGTFfeatureExon 'CDS'). We counted reads using the command htseq-count from the HTSeq framework v1.99.2 [157], which is a Python package for analysis of high-throughput sequencing data. We analyzed counts using the DESeq function from DESeq2 v1.26.0 [158] to perform differential analysis and generated results using the results function with contrasts comparing each focal tissue with an average of all other tissues. We considered genes upregulated if the adjusted *p*-value (padj) was greater than 0.05 and log2FoldChange was greater than 2.

In situ hybridization chain reaction (HCR) in *Berghia* juveniles

Probe design

We designed all probe sets using the HCR 3.0 probe maker [159]. The sequences generated by the software were used to order probe sets (50 pmol DNA oPools Oligo Pool) from Integrated DNA Technologies (Coralville, IA USA), which we resuspended to 1 pmol/µl in 50 mL TE buffer (Tris, EDTA).

Hybridization chain reaction

We cultured *Berghia stephanieae* juveniles using the same methods as prior *B. stephanieae* imaging work [35, 36]. We starved juveniles for 5 days prior to fixation to decrease autofluorescence from digestive contents. We relaxed juveniles in 1-part 7.3% MgCl2: 2-part fresh filtered sea water for 30 min prior to fixation. We then washed and incubated samples in 4% PFA (paraformaldehyde diluted in FSW from 16% ampules) overnight at 4 °C. We washed samples three times in 1X phosphate-buffered saline (PBS), followed by a 50% 1X PBS/50% methanol solution wash, followed by three 100% methanol washes. All washes were 10 min. We stored samples in methanol at −20 °C.

We performed in situ HCR using the buffers and protocols detailed in Choi et al. [82] with the following modifications. All steps were performed in 1.5-mL tubes and the volume of washes was decreased to 200-µL to better suit our samples. We rehydrated samples into 5X SSCT from methanol, immediately followed by the detection stage of the protocol. We prepared probe solutions using 100 µL of hybridization buffer and 1.0 pmol/oligo/µL of each probe. Following overnight probe hybridization, we washed samples with 30% probe hybridization wash buffer for 3 × 5 min, followed by 2 × 30 min. Following the 5X SSCT washes during the amplification stage, we placed samples in the hairpin solution (6pmol solution using 2µL of 3µM stock of snap-cooled hairpins in 100 µL of amplification buffer). After the completion of the HCR protocol, we incubated the samples in 1.0 µg/mL DAPI diluted in 5X SSCT for 30 min. Samples were then stored in 5X SSCT at 4 C for up to 5 days until mounting and imaging.

We mounted samples in a 20% 5X SSCT, 80% Glycerol solution, and imaged samples with a Zeiss LSM 710 inverted confocal microscope with an AxioCam HRm camera. We analyzed images using image processing software ImageJ FIJI and Adobe Photoshop [160]. Samples were stitched together using the FIJI Pairwise Stitching Plugin [161]. Figures were created in Adobe Illustrator.

Abbreviations

Berghia	*Berghia stephanieae*
BR	Brain
RH	Rhinophore
OT	Oral Tentacle
DC	Distal ceras
PC	Proximal ceras
FO	Foot
TA	Tail
cr	Cerata
cs	Cnidosac
COL8A1	Gene (jg12556) annotated as collagen alpha 1 VIII
PTI	Gene (jg18351) annotated as a pancreatic trypsin inhibitor
ORF	Open reading frame
PFA	Paraformaldehyde
PBS	Phosphate-buffered saline
HCR	Hybridization chain reaction

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-024-01814-3.

Additional file 1. *Berghia stephanieae* HiRise Scaffolding report from Dovetail.

Additional file 2: Figure S1. K-mer spectra and fitted models for a GenomeScope analysis of our Dovetail Omni-C data for *Berghia stephanieae*. **Figure S2.** Snail plot showing high contiguity of the filtered genome (18 scaffolds), where 99% of the genome is found in the top 15 scaffolds. This plot also indicates the high levels of completeness (93.3%) when compared to metazoan single-copy ortholog datasets (BUSCO), with low levels of duplication (0.6%). **Figure S3.** This chart shows the proteins in each proteome that were classified as species-specific based on the OrthoFinder and KinFin analyses. **Figure S4.** This chart shows the proportion of proteins in each proteome that were classified as species-specific based on the OrthoFinder and KinFin analyses. **Figure S5.** Rarefaction curve for gene sampling at the phylum level. This graph indicates that gene sampling for Mollusca (green) is reaching an asymptote, which suggests that our sampling of these genes is sufficient for identifying *Berghia* genes that would match to other Mollusca proteomes. **Figure S6.** Rarefaction curve for gene sampling at the class level. This graph indicates that gene sampling for Gastropoda (blue) and Bivalvia (green) are reaching an asymptote, which suggests that our sampling of these genes is sufficient for identifying *Berghia* genes that would match to other gastropod proteomes. **Figure S7.** Rarefaction curve for gene sampling at the subclass level. This graph indicates that gene sampling for Heterobranchia (purple) are reaching an asymptote, which suggests that our sampling of these genes is sufficient for identifying *Berghia* genes that would match to other gastropod proteomes. **Figure S8.** Rarefaction curve for gene sampling at the order level. This graph indicates that gene sampling for Nudibranchia (red) is reaching an asymptote, which suggests that our sampling of these genes is sufficient for identifying *Berghia* genes that would match to other nudibranch proteomes. **Figure S9.** Rarefaction curve for gene sampling at the superfamily level. This graph indicates that gene sampling for Aeolidoidea is reaching an asymptote, though this group is not as well sampled as other levels. This still suggests that our sampling of these genes is sufficient for identifying *Berghia* genes that would match to other aeolid proteomes. **Figure S10.** Distribution of both clade-specific and non-clade-specific (Other) genes across the putative chromosomes in our *Berghia stephanieae* genome. To the left is the raw gene counts per chromosome, and to the right is the number of genes per Mb of sequence to account for differences in chromosome lengths. **Figure S11.** This chart shows the proportion of genes on each chromosome that are clade-specific and non-clade-specific (Other). **Figures S12-18.** Gene expression across tissues in genes classified as *Berghia*-specific (**Figure S12**), Aeolidina-specific (**Figure S13**), Nudibranchia-specific (**Figure S14**), Heterobranchia-specific (**Figure S15**),Gastropoda-specific (**Figure S16**), Mollusca-specific (**Figure S17**), and Other (**Figure S18**). Left: Heatmap of expression profiles; Right: PCA-plot showing similarity of expression within and among tissues.

Additional file 3: Table S1. Genome summary statistics for various mollusk genomes available in NCBI and both versions of MolluscDB.

Additional file 4: Table S2. RepeatModeler analysis for the Berghia stephanieae genome.

Additional file 5: Table S3. OrthoFinder cluster analysis statistics for each proteome.

Additional file 6: Table S4. Read statistics for short read transcriptome samples for *Berghia stephanieae*.

Additional file 7: Table S5. Number of genes upregulated in each tissue, categorized by assigned clade-specificity of each gene. **Table S6.** Percentage of genes upregulated in each tissue, categorized by assigned clade-specificity of each gene.

Additional file 8: Table S7. Number of genes uniquely upregulated in each tissue, categorized by assigned clade-specificity of each gene. **Table S8.** Percentage of genes uniquely upregulated in each tissue, categorized by assigned clade-specificity of each gene.

Additional file 9: Table S9. Number and proportion of upregulated genes in each tissue that were classified as annotated, unannotated, or uncharacterized.(XLSX 6 KB)

Additional file 10: Table S10. GO term enrichment results for genes upregulated in the brain. **Table S11**. GO term enrichment results for genes upregulated in the rhinophores. **Table S12.** GO term enrichment results for genes upregulated in the oral tentacles. **Table S13.** GO term enrichment results for genes upregulated in the distal ceras. **Table S14.** GO term enrichment results for genes upregulated in the proximal ceras. **Table S15.** GO term enrichment results for genes upregulated in the foot. **Table S16.** GO term enrichment results for genes upregulated in the tail.

Additional file 11. Newick file of the species tree used in our OrthoFinder analysis.

Additional file 12. List of commands used in the *Berghia* genome annotation.

Acknowledgements

We thank all members of the Lyons and Katz labs for help with animal care and feedback on this project and manuscript. We are also grateful to all members of the *Berghia* Brain Project (https://sites.google.com/umass.edu/berghiabrainproject/) for their feedback on this project. We also thank the Hamdoun Lab for the use of their confocal microscope and Rouse Lab for the use of their Qubit. This work is in memory of our outstanding colleague Dr. Maryna P. Lesoway, who provided significant intellectual and emotional support in executing this work.

Authors' contributions

JAG, MDR, PSK, and DCL conceived of the study; JAG, NFT, RAF, SRB, KM, MACJ, CW, PM, GOB, HTJ, MDR, and DCL collected the data; JAG, RAR, and RAF performed the data analyses; JAG, MDR, RAF, PSK, and DCL participated in the data interpretation. All authors read and approved the final manuscript.

Funding

This work was supported by a Scripps Postdoctoral Fellowship to JAG and NIH BRAIN awards, U01-NS108637 and 1U01-NS123972, to PSK and DCL and an Emerging Research Organisms Grant from the Society for Developmental Biology to DCL This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges-2 system, which is supported by NSF award number ACI-1928147, at the Pittsburgh Supercomputing Center (PSC) and Expanse at the San Diego Supercomputer Center (SDSC) through allocation IDs TG-BIO210019 and TG-BIO210138 [162].

Availability of data and materials

Raw sequencing data are accessible through the NCBI Sequence Read Archive (BioProjects PRJNA1004233 and PRJNA641185). This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAWQJI000000000. The version described in this paper is version JAWQJI010000000. The full commands we used for filtering and annotation steps are provided in Additional file 12. Input and intermediate files for our analyses are available in Dryad (https://doi.org/https://doi.org/10.6076/D1BS33), and custom python, shell, and R scripts used in our analyses are available through GitHub (https://github.com/lyons-lab/berghia_reference_transcriptome and https://github.com/goodheart-lab/berghia_genome).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Division of Invertebrate Zoology, American Museum of Natural History, New York, NY, USA. [2]Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA. [3]Bioengineering Department, Stanford University, Stanford, CA, USA. [4]Department of Wildland Resources, Utah State University, Logan, UT, USA. [5]Department of Biology, University of Massachusetts Amherst, Amherst, MA, USA. [6]Institute of Neuroscience, University of Oregon, Eugene, OR, USA.

Received: 7 August 2023 Accepted: 3 January 2024
Published online: 17 January 2024

References

1. Wilkins AS. "the genetic tool-kit": The life-history of an important metaphor. In: Advances in Evolutionary Developmental Biology. Hoboken: John Wiley & Sons, Inc.; 2013. p. 1–14.
2. Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell. 2008;134:25–36.
3. Newman SA. The developmental genetic toolkit and the molecular homology—analogy paradox. Biol Theory. 2006;1:12–6.
4. de Mendoza A, Sebé-Pedrós A, Ruiz-Trillo I. The evolution of the GPCR signaling system in eukaryotes: modularity, conservation, and the transition to metazoan multicellularity. Genome Biol Evol. 2014;6:606–19.

5. Wu L, Lambert JD. Clade-specific genes and the evolutionary origin of novelty; new tools in the toolkit. Semin Cell Dev Biol. 2022. https://doi.org/10.1016/j.semcdb.2022.05.025.
6. Johnson BR. Taxonomically restricted genes are fundamental to biology and evolution. Front Genet. 2018;9:407.
7. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 2009;25:404–13.
8. Valdés A. A new species of *Aeolidiella* Bergh, 1867 (Mollusca: Nudibranchia: Aeolidiidae) from the Florida keys USA. Veliger. 2005;47:218–23.
9. Kurz EM, Holstein TW, Petri BM, Engel J, David CN. Mini-collagens in *Hydra* nematocytes. J Cell Biol. 1991;115:1159–69.
10. Koch AW, Holstein TW, Mala C, Kurz E, Engel J, David CN. Spinalin, a new glycine- and histidine-rich protein in spines of *Hydra* nematocysts. J Cell Sci. 1998;111(Pt 11):1545–54.
11. Babonis LS, Martindale MQ. Old cell, new trick? Cnidocytes as a model for the evolution of novelty. Integr Comp Biol. 2014;54:714–22.
12. Wu L, Hiebert LS, Klann M, Passamaneck Y, Bastin BR, Schneider SQ, et al. Genes with spiralian-specific protein motifs are expressed in spiralian ciliary bands. Nat Commun. 2020;11:4171.
13. Garb JE, Ayoub NA, Hayashi CY. Untangling spider silk evolution with spidroin terminal domains. BMC Evol Biol. 2010;10:243.
14. Hinman MB, Lewis RV. Isolation of a clone encoding a second dragline silk fibroin. Nephila clavipes dragline silk is a two-protein fiber. J Biol Chem. 1992;267:19320–4.
15. Chen S, Zhang YE, Long M. New genes in *Drosophila* quickly become essential. Science. 2010;330:1682–5.
16. True JR, Haag ES. Developmental system drift and flexibility in evolutionary trajectories. Evol Dev. 2001;3:109–19.
17. Hwang JS, Takaku Y, Momose T, Adamczyk P, Özbek S, Ikeo K, et al. Nematogalectin, a nematocyst protein with GlyXY and galectin domains, demonstrates nematocyte-specific alternative splicing in *Hydra*. Proc Natl Acad Sci U S A. 2010;107:18539–44.
18. Khalturin K, Anton-Erxleben F, Sassmann S, Wittlieb J, Hemmrich G, Bosch TCG. A novel gene family controls species-specific morphological traits in *Hydra*. PLoS Biol. 2008;6:e278.
19. Milde S, Hemmrich G, Anton-Erxleben F, Khalturin K, Wittlieb J, Bosch TCG. Characterization of taxonomically restricted genes in a phylum-restricted cell type. Genome Biol. 2009;10:R8.
20. Santos ME, Le Bouquin A, Crumière AJJ, Khila A. Taxon-restricted genes at the origin of a novel trait allowing access to a new environment. Science. 2017;358:386–90.
21. Jasper WC, Linksvayer TA, Atallah J, Friedman D, Chiu JC, Johnson BR. Large-scale coding sequence change underlies the evolution of post-developmental novelty in honey bees. Mol Biol Evol. 2015;32:334–46.
22. Babonis LS, Martindale MQ, Ryan JF. Do novel genes drive morphological novelty? An investigation of the nematosomes in the sea anemone *Nematostella vectensis*. BMC Evol Biol. 2016;16:114.
23. Saleuddin S, Mukai S. Physiology of Molluscs: A Collection of Selected Reviews, Volume 1. Boca Raton; CRC Press; 2021.
24. Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, et al. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature. 2015;524:220–4.
25. Styfhals R, Zolotarov G, Hulselmans G, Spanier KI, Poovathingal S, Elagoz AM, et al. Cell type diversity in a developing octopus brain. Nat Commun. 2022;13:7392.
26. Metzger MJ, Villalba A, Carballal MJ, Iglesias D, Sherry J, Reinisch C, et al. Widespread transmission of independent cancer lineages within multiple bivalve species. Nature. 2016;534:705–9.
27. Katz PS, Quinlan PD. The importance of identified neurons in gastropod molluscs to neuroscience. Curr Opin Neurobiol. 2019;56:1–7.
28. Byrne JH. Learning and memory in Aplysia and other invertebrates. In: Neurobiology of Comparative Cognition. 1st Ed. East Sussex: Psychology Press; 2014. p. 311–34.
29. Coustau C, Gourbal B, Duval D, Yoshino TP, Adema CM, Mitta G. Advances in gastropod immunity from the study of the interaction between the snail *Biomphalaria glabrata* and its parasites: a review of research progress over the last decade. Fish Shellfish Immunol. 2015;46:5–16.
30. Coustau C, Kurtz J, Moret Y. A novel mechanism of immune memory unveiled at the invertebrate-parasite interface. Trends Parasitol. 2016;32:353–5.
31. Davison A, Neiman M. Mobilizing molluscan models and genomes in biology. Philos Trans R Soc Lond B Biol Sci. 2021;376:20200163.
32. Gomes-dos-Santos A, Lopes-Lima M, Castro LFC, Froufe E. Molluscan genomics: the road so far and the way forward. Hydrobiologia. 2020;847:1705–26.
33. Wägele H, Willan RC. Phylogeny of the Nudibranchia. Zool J Linnean Soc. 2000;130:83–181.
34. Kristof A, Klussmann-Kolb A. Neuromuscular development of *Aeolidiella stephanieae* Valdez, 2005 (Mollusca, Gastropoda, Nudibranchia). Front Zool. 2010;7:5.
35. Taraporevala NF, Lesoway MP, Goodheart JA, Lyons DC. Precocious sperm exchange in the simultaneously hermaphroditic nudibranch *Berghia stephanieae*. Integr Organism Biol. 2022;4:obac030.
36. Goodheart JA, Barone V, Lyons DC. Movement and storage of nematocysts across development in the nudibranch *Berghia stephanieae* (Valdés, 2005). Front Zool. 2022;19:16.
37. Obermann D, Bickmeyer U, Wägele H. Incorporated nematocysts in *Aeolidiella stephanieae* (Gastropoda, Opisthobranchia, Aeolidoidea) mature by acidification shown by the pH sensitive fluorescing alkaloid Ageladine A. Toxicon. 2012;60:1108–16.
38. Silva RXG, Cartaxana P, Calado R. Prevalence and photobiology of photosynthetic dinoflagellate endosymbionts in the nudibranch *Berghia stephanieae*. Animals. 2021;11:2200.
39. Clavijo Melo J, Sickinger C, Bleidißel S, Gasparoni G, Tierling S, Preisfeld A, et al. The nudibranch *Berghia stephanieae* (Valdés, 2005) is not able to initiate a functional symbiosome-like environment to maintain *Breviolum minutum* (J.E.Parkinson & LaJeunesse 2018). Front Marine Sci. 2022;9:934307.
40. Martín-Durán JM, Hejnol A. A developmental perspective on the evolution of the nervous system. Dev Biol. 2021;475:181–92.
41. Goodheart JA, Bely AE. Sequestration of nematocysts by divergent cnidarian predators: mechanism, function, and evolution. Invertebr Biol. 2017;136:75–91.
42. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11:1432.
43. Cai H, Li Q, Fang X, Li J, Curtis NE, Altenburger A, et al. A draft genome assembly of the solar-powered sea slug *Elysia chlorotica*. Sci Data. 2019;6:190022.
44. Sun J, Zhang Y, Xu T, Zhang Y, Mu H, Zhang Y, et al. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. Nat Ecol Evol. 2017;1:121.
45. Uliano-Silva M, Dondero F, Dan Otto T, Costa I, Lima NCB, Americo JA, et al. A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel, *Limnoperna fortunei*. Gigascience. 2018;7:gix128.
46. Du X, Fan G, Jiao Y, Zhang H, Guo X, Huang R, et al. The pearl oyster *Pinctada fucata martensii* genome and multi-omic analyses provide insights into biomineralization. Gigascience. 2017;6:1–12.
47. Jiao W, Fu X, Dou J, Li H, Su H, Mao J, et al. High-resolution linkage and quantitative trait locus mapping aided by genome survey sequencing: building up an integrative genomic framework for a bivalve mollusc. DNA Res. 2014;21:85–101.
48. Thiriot-Quiévreux C. Advances in chromosomal studies of gastropod molluscs. J Molluscan Stud. 2003;69:187–202.
49. Zheng C, Sankoff D. Locating rearrangement events in a phylogeny based on highly fragmented assemblies. BMC Genomics. 2016;17:S1.
50. De Oliveira AL, Wollesen T, Kristof A, Scherholz M, Redl E, Todt C, et al. Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks. BMC Genom. 2016;17:905.
51. Varney RM, Speiser DI, McDougall C, Degnan BM, Kocot KM. The iron-responsive genome of the chiton *Acanthopleura granulata*. Genome Biol Evol. 2021;13:evaa263.
52. Fuller ZL, Mocellin VJL, Morris LA, Cantin N, Shepherd J, Sarre L, et al. Population genetics of the coral: toward genomic prediction of bleaching. Science. 2020;369:eaba4674.

53. Layton KKS, Carvajal JI, Wilson NG. Mimicry and mitonuclear discordance in nudibranchs: new insights from exon capture phylogenomics. Ecol Evol. 2020;10:11966–82.
54. Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, et al. The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature. 2010;466:720–6.
55. Knudsen B, Kohn AB, Nahir B, McFadden CS, Moroz LL. Complete DNA sequence of the mitochondrial genome of the sea-slug, *Aplysia californica*: conservation of the gene order in Euthyneura. Mol Phylogenet Evol. 2006;38:459–69.
56. Liu F, Li Y, Yu H, Zhang L, Hu J, Bao Z, et al. MolluscDB: an integrated functional and evolutionary genomics database for the hyper-diverse animal phylum Mollusca. Nucleic Acids Res. 2021;49:D988-97.
57. Acemel RD, Tena JJ, Irastorza-Azcarate I, Marlétaz F, Gómez-Marín C, de la Calle-Mustienes E, et al. A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. Nat Genet. 2016;48:336–41.
58. Stroehlein AJ, Korhonen PK, Rollinson D, Stothard JR, Hall RS, Gasser RB, et al. *Bulinus truncatus* transcriptome - a resource to enable molecular studies of snail and schistosome biology. Curr Res Parasitol Vector Borne Dis. 2021;1:100015.
59. C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science. 1998;282:2012–8.
60. da Fonseca RR, Couto A, Machado AM, Brejova B, Albertin CB, Silva F, et al. A draft genome sequence of the elusive giant squid, *Architeuthis dux*. Gigascience. 2020;9:giz152.
61. Simakov O, Marletaz F, Cho S-J, Edsinger-Gonzales E, Havlak P, Hellsten U, et al. Insights into bilaterian evolution from three spiralian genomes. Nature. 2013;493:526–31.
62. Sun J, Chen C, Miyamoto N, Li R, Sigwart JD, Xu T, et al. The Scaly-foot Snail genome and implications for the origins of biomineralised armour. Nat Commun. 2020;11:1657.
63. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013;496:498–503.
64. Broughton RE, Milam JE, Roe BA. The complete sequence of the zebrafish (*Danio rerio*) mitochondrial genome and evolutionary patterns in vertebrate mitochondrial DNA. Genome Res. 2001;11:1958–67.
65. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al. The Release 6 reference sequence of the *Drosophila melanogaster* genome. Genome Res. 2015;25:445–58.
66. Maeda T, Takahashi S, Yoshida T, Shimamura S, Takaki Y, Nagai Y, et al. Chloroplast acquisition without the gene transfer in kleptoplastic sea slugs. Elife. 2021;10:e60176.
67. Goodheart JA, Bazinet AL, Collins AG, Cummings MP. Relationships within Cladobranchia (Gastropoda: Nudibranchia) based on RNA-Seq data: an initial investigation. R Soc Open Sci. 2015;2:150196.
68. Caruana NJ, Cooke IR, Faou P, Finn J, Hall NE, Norman M, et al. A combined proteomic and transcriptomic analysis of slime secreted by the southern bottletail squid, *Sepiadarium austrinum* (Cephalopoda). J Proteomics. 2016;148:170–82.
69. Smith SA, Wilson NG, Goetz FE, Feehery C, Andrade SCS, Rouse GW, et al. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature. 2011;480:364–7.
70. Lan Y, Sun J, Chen C, Sun Y, Zhou Y, Yang Y, et al. Hologenome analysis reveals dual symbiosis in the deep-sea hydrothermal vent snail Gigantopelta aegis. Nat Commun. 2021;12:1165.
71. Goodheart JA, Bazinet AL, Valdés Á, Collins AG, Cummings MP. Prey preference follows phylogeny: evolutionary dietary patterns within the marine gastropod group Cladobranchia (Gastropoda: Heterobranchia: Nudibranchia). BMC Evol Biol. 2017;17:221.
72. Zapata F, Wilson NG, Howison M, Andrade SCS, Jörger KM, Schrödl M, et al. Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. Proc Biol Sci. 2014;281:20141739.
73. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 2009;7:e1000112.
74. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. PLoS Biol. 2011;9:e1001091.
75. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science. 2007;317:86–94.
76. Albertin CB, Medina-Ruiz S, Mitros T, Schmidbaur H, Sanchez G, Wang ZY, et al. Genome and transcriptome mechanisms driving cephalopod evolution. Nat Commun. 2022;13:2427.
77. Wang S, Zhang J, Jiao W, Li J, Xun X, Sun Y, et al. Scallop genome provides insights into evolution of bilaterian karyotype and development. Nat Ecol Evol. 2017;1:120.
78. Sato M, Nagashima K. Molecular characterization of a mitochondrial DNA segment from the Japanese scallop (*Patinopecten yessoensis*): demonstration of a region showing sequence polymorphism in the population. Mar Biotechnol. 2001;3:370–9.
79. Nolan JR, Bergthorsson U, Adema CM. atypical mitochondrial gene order among panpulmonates (Gastropoda). J Molluscan Stud. 2014;80:388–99.
80. Sea Urchin Genome Sequencing Consortium, Sodergren E, Weinstock GM, Davidson EH, Cameron RA, Gibbs RA, et al. The genome of the sea urchin *Strongylocentrotus purpuratus*. Science. 2006;314:941–52.
81. Law CW, Alhamdoosh M, Su S, Dong X, Tian L, Smyth GK, et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Res. 2016;5:1408.
82. Choi HMT, Schwarzkopf M, Fornace ME, Acharya A, Artavanis G, Stegmaier J, et al. Third-generation hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. Development. 2018;145:dev165753.
83. Ramirez MD, Bui TN, Katz PS. Mapping neuronal gene expression reveals aspects of ganglionic organization in a gastropod mollusc. 2023. Preprint at https://www.biorxiv.org/content/10.1101/2023.06.22.546160v1.
84. Farhat S, Bonnivard E, Pales Espinosa E, Tanguy A, Boutet I, Guiglielmoni N, et al. Comparative analysis of the *Mercenaria mercenaria* genome provides insights into the diversity of transposable elements and immune molecules in bivalve mollusks. BMC Genom. 2022;23:192.
85. Holmes A, Derbyshire T, Brennan M, McTierney S, Small A, Marine Biological Association Genome Acquisition Lab, et al. The genome sequence of *Gari tellinella* (Lamarck, 1818), a sunset clam. Wellcome Open Research. 2022;7:116.
86. Chen Z, Doğan Ö, Guiglielmoni N, Guichard A, Schrödl M. The *de novo* genome of the "Spanish" slug *Arion vulgaris* Moquin-Tandon, 1855 (Gastropoda: Panpulmonata): massive expansion of transposable elements in a major pest species. 2020. Preprint at https://www.biorxiv.org/content/10.1101/2020.11.30.403303v1.
87. Chueca LJ, Schell T, Pfenninger M. De novo genome assembly of the land snail *Candidula unifasciata* (Mollusca: Gastropoda). G3. 2021;11:jkab180.
88. Linscott TM, González-González A, Hirano T, Parent CE. De novo genome assembly and genome skims reveal LTRs dominate the genome of a limestone endemic Mountainsnail (*Oreohelix idahoensis*). BMC Genom. 2022;23:796.
89. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.
90. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.
91. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238.
92. Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS Biol. 2020;18:e3000862.
93. Aguilera F, McDougall C, Degnan BM. Co-Option and De Novo Gene Evolution Underlie Molluscan Shell Diversity. Mol Biol Evol. 2017;34:779–92.
94. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011;12:692–702.
95. Johnson BR, Tsutsui ND. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. BMC Genom. 2011;12:164.
96. Hilgers L, Hartmann S, Hofreiter M, von Rintelen T. Novel genes, ancient genes, and gene co-option contributed to the genetic basis of the radula, a molluscan innovation. Mol Biol Evol. 2018;35:1638–52.

97. Goodheart JA, Bleidißel S, Schillo D, Strong EE, Ayres DL, Preisfeld A, et al. Comparative morphology and evolution of the cnidosac in Cladobranchia (Gastropoda: Heterobranchia: Nudibranchia). Front Zool. 2018;15:43.
98. Davy SK, Allemand D, Weis VM. Cell biology of cnidarian-dinoflagellate symbiosis. Microbiol Mol Biol Rev. 2012;76:229–61.
99. Brenzinger B, Schrödl M, Kano Y. Origin and significance of two pairs of head tentacles in the radiation of euthyneuran sea slugs and land snails. Sci Rep. 2021;11:21016.
100. Huber G. On the cerebral nervous system of marine Heterobranchia (Gastropoda). J Molluscan Stud. 1993;59:381–420.
101. Striedter GF. Principles of brain evolution. Sunderland; Sinauer Associates Incorporated; 2005.
102. Vickaryous MK, Hall BK. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. Biol Rev Camb Philos Soc. 2006;81:425–55.
103. Moroz LL. On the independent origins of complex brains and neurons. Brain Behav Evol. 2009;74:177–90.
104. Shafer MER, Sawh AN, Schier AF. Gene family evolution underlies cell-type diversification in the hypothalamus of teleosts. Nat Ecol Evol. 2022;6:63–76.
105. Hwang JS, Ohyanagi H, Hayakawa S, Osato N, Nishimiya-Fujisawa C, Ikeo K, et al. The evolutionary emergence of cell type-specific genes inferred from the gene expression analysis of *Hydra*. Proc Natl Acad Sci U S A. 2007;104:14735–40.
106. Greenwood PG, Mariscal RN. The utilization of cnidarian nematocysts by aeolid nudibranchs: nematocyst maintenance and release in *Spurilla*. Tissue Cell. 1984;16:719–30.
107. Rosin FM, Kramer EM. Old dogs, new tricks: regulatory evolution in conserved genetic modules leads to novel morphologies in plants. Dev Biol. 2009;332:25–35.
108. Mikheyev AS, Linksvayer TA. Genes associated with ant social behavior show distinct transcriptional and evolutionary patterns. Elife. 2015;4: e04775.
109. Yang L, Zou M, Fu B, He S. Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish. BMC Genomics. 2013;14:65.
110. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884-90.
111. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
112. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.
113. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.
114. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
115. Gladyshev EA, Meselson M, Arkhipova IR. Massive horizontal gene transfer in bdelloid rotifers. Science. 2008;320:1210–3.
116. UniProt Consortium. The universal protein resource (UniProt). Nucleic Acids Res. 2008;36 Database issue:D190–5.
117. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol Biol. 2019;1962:227–45.
118. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
119. 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;38:4647–54.
120. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17:155–8.
121. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36:2896–8.
122. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 2016;26:342–50.
123. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.
124. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013;41:e121–e121.
125. Levy Karin E, Mirdita M, Söding J. MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome. 2020;8:48.
126. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST : architecture and applications. BMC Bioinformatics. 2009;10:421.
127. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
128. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
129. Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit - Interactive Quality Assessment of Genome Assemblies. G3 . 2020;10:1361–74.
130. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022;50:D20-6.
131. Telatin A, Fariselli P, Birolo G. SeqFu: a suite of utilities for the robust and reproducible manipulation of sequence files. Bioengineering (Basel). 2021;8:59.
132. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117:9451–7.
133. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research. 1999;27:573–80.
134. Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12:1269–76.
135. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21(Suppl 1):i351-8.
136. Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2018;176:1410–22.
137. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
138. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
139. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
140. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
141. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32:767–9.
142. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 2021;3:lqaa108.
143. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. Methods Mol Biol. 2019;1962:65–95.
144. Brůna T, Lomsadze A, Borodovsky M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genom Bioinform. 2020;2:lqaa026.
145. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005;33:6494–506.
146. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
147. Gotoh O. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. Nucleic Acids Res. 2008;36:2630–8.

148. Iwata H, Gotoh O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. Nucleic Acids Res. 2012;40:e161.
149. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44.
150. Laumer CE, Fernández R, Lemer S, Combosch D, Kocot KM, Riesgo A, et al. Revisiting metazoan phylogeny with genomic sampling of all phyla. Proc Biol Sci. 2019;286:20190831.
151. Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, et al. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci. 2009;276:4261–70.
152. Kocot KM, Struck TH, Merkel J, Waits DS, Todt C, Brannock PM, et al. Phylogenomics of Lophotrochozoa with Consideration of Systematic Error. Syst Biol. 2017;66:256–82.
153. Marlétaz F, Peijnenburg KTCA, Goto T, Satoh N, Rokhsar DS. A new spiralian phylogeny places the enigmatic arrow worms among gnathiferans. Curr Biol. 2019;29:312-8.e3.
154. Sigwart JD, Lindberg DR. Consensus and confusion in molluscan trees: evaluating morphological and molecular phylogenies. Syst Biol. 2015;64:384–95.
155. Karmeinski D, Meusemann K, Goodheart JA, Schroedl M, Martynov A, Korshunova T, et al. Transcriptomics provides a robust framework for the relationships of the major clades of cladobranch sea slugs (Mollusca, Gastropoda, Heterobranchia), but fails to resolve the position of the enigmatic genus *Embletonia*. BMC Ecol Evol. 2021;21:226.
156. Laetsch DR, Blaxter ML. KinFin: software for taxon-aware analysis of clustered protein sequences. G3 . 2017;7:3349–57.
157. Putri GH, Anders S, Pyl PT, Pimanda JE, Zanini F. Analysing high-throughput sequencing data in Python with HTSeq 2.0. arXiv [q-bio.GN]. 2021.
158. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
159. Kuehn E, Clausen DS, Null RW, Metzger BM, Willis AD, Özpolat BD. Segment number threshold determines juvenile onset of germline cluster expansion in *Platynereis dumerilii*. J Exp Zool B Mol Dev Evol. 2022;338:225–40.
160. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. Nat Methods. 2012;9:676–82.
161. Preibisch S, Saalfeld S, Tomancak P. Globally optimal stitching of tiled 3D microscopic image acquisitions. Bioinformatics. 2009;25:1463–5.
162. Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, et al. XSEDE: accelerating scientific discovery. Comput Sci Eng. 2014;16:62–74.

Publisher's Note