Phylogenetics of Lophotrochozoan bHLH Genes and the Evolution of Lineage-Specific Gene Duplicates

Yongbo Bao,[1,2] Fei Xu,[3,4,5,*] and Sebastian M. Shimeld[1,*]

[1]Department of Zoology, University of Oxford, United Kingdom

[2]Zhejiang Key Laboratory of Aquatic Germplasm Resources, College of Biological & Environmental Sciences, Zhejiang Wanli University, Zhejiang, China

[3]Key Laboratory of Experimental Marine Biology, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China

[4]Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China

[5]National & Local Joint Engineering Laboratory of Ecological Mariculture, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China

*Corresponding authors: E-mails: xufei@qdio.ac.cn; sebastian.shimeld@zoo.ox.ac.uk.

Accepted: March 10, 2017

Abstract

The gain and loss of genes encoding transcription factors is of importance to understanding the evolution of gene regulatory complexity. The basic helix–loop–helix (bHLH) genes encode a large superfamily of transcription factors. We systematically classify the bHLH genes from five mollusc, two annelid and one brachiopod genomes, tracing the pattern of bHLH gene evolution across these poorly studied Phyla. In total, 56–88 bHLH genes were identified in each genome, with most identifiable as members of previously described bilaterian families, or of new families we define. Of such families only one, *Mesp*, appears lost by all these species. Additional duplications have also played a role in the evolution of the bHLH gene repertoire, with many new lophotrochozoan-, mollusc-, bivalve-, or gastropod-specific genes defined. Using a combination of transcriptome mining, RT-PCR, and in situ hybridization we compared the expression of several of these novel genes in tissues and embryos of the molluscs *Crassostrea gigas* and *Patella vulgata*, finding both conserved expression and evidence for neofunctionalization. We also map the positions of the genes across these genomes, identifying numerous gene linkages. Some reflect recent paralog divergence by tandem duplication, others are remnants of ancient tandem duplications dating to the lophotrochozoan or bilaterian common ancestors. These data are built into a model of the evolution of bHLH genes in molluscs, showing formidable evolutionary stasis at the family level but considerable within-family diversification by tandem gene duplication.

Key words: bHLH gene, mollusc, annelid, transcription factor, gene duplication.

Introduction

Basic helix–loop–helix (bHLH) proteins form a large and diverse group of transcriptional regulators, characterized by possession of a bHLH domain that is involved in DNA binding and dimerization. One study indicates 125 genes encoding bHLH factors can be identified in the human genome (Ledent et al. 2002), and studies in model organisms have identified numerous roles in development and physiology. For example, some bHLH proteins such as MyoD, Oligo, and Neurogenin act as key regulators of specific cell types during embryo development, whereas others such as Bmal, Clock, and HIF play critical roles in the circadian clock (Bmal, Clock) or the response to hypoxia (HIF). bHLH genes are found in a wide range of eukaryotes and also show a high level of diversity in plant genomes (Carretero-Paulet et al. 2010; Pires and Dolan 2010a, 2010b); however, here we confine analysis to animal bHLH genes.

There have been several attempts to classify bHLH genes on the basis of sequence identity and domain structure, and hence inferred evolutionary relationships (Atchley and Fitch 1997; Ledent and Vervoort 2001; Ledent et al. 2002; Skinner et al. 2010). Most studies find that, as with other transcription factor genes such as the Fox genes (Mazet et al. 2003) and Homeobox genes (Holland et al. 2007), animal bHLH genes generally fall into clear families of orthologs, identifiable across the Bilateria and sometimes in earlier diverging animal phyla such as Cnidaria (Simionato et al. 2007). There is also evidence from molecular phylogenetics and protein domain organization for higher-level groups of such ortholog families; these have been successively named A–D, A–E, and A–F by different authors (Atchley and Fitch 1997; Ledent and Vervoort 2001; Skinner et al. 2010), though it has been noted Group B may not be monophyletic (Atchley and Fitch 1997; Sebe-Pedros et al. 2011).

Several studies have attempted to catalog and classify bHLH genes in specific organisms. These include model organisms (Moore et al. 2000; Ledent and Vervoort 2001; Liu and Zhao 2010), nonmodel vertebrates (Dang, Wang, Zhang, et al. 2011; Liu et al. 2013), *Ciona intestinalis* (Satou et al. 2003), several insects (Wang et al. 2008; Dang, Wang, Chen, et al. 2011; Liu et al. 2012; Zhang et al. 2013), a pearl oyster (Gyoja and Satoh 2013), and nonbilaterian species (Simionato et al. 2007; Gyoja et al. 2012; Gyoja 2014). Such cataloguing exercises reveal patterns of gene loss and gain across animal evolution.

A major gap in our understanding of these patterns of gene loss and gain is a lack of data from lophotrochozoans, to date represented only by a focused study on the pearl oyster *Pinctada fucata* (Gyoja and Satoh 2013) and by broader analyses that included data from some lophotrochozoan species (Simionato et al. 2007). Here, we address this gap by exploiting recent developments in genome sequencing of molluscs to conduct a focused analysis of bHLH gene evolution in this lineage. The molluscs are a diverse Phylum with an estimated 100,000 species, most of which fall into two classes, the Bivalvia (of which *P. fucata* is a member) and the Gastropoda (snails, slugs, and allies). As well as reevaluating the *P. fucata* data, we include another bivalve (the oyster *Crassostrea gigas;* Zhang et al. 2012) and three gastropods (the limpets *Lottia gigantea;* Simakov et al. 2013) and *Patella vulgata* (Kenny et al. 2015) and the fresh water snail *Biomphalaria glabrata*. We also include three other lophotrochozoans with sequenced genomes (the annelids *Helobdella robusta* and *Capitella teleta* and the brachiopod *Lingula anatina* [Simakov et al. 2013; Luo et al. 2015]) to help identify when genes and gene families have been gained or lost. We find evidence for a high level of bHLH family retention in the Lophotrochozoa. We also detect many new genes, most of which have evolved by tandem duplication. Most such duplicates are clearly ascribable to bilaterian bHLH families, but some are not and form new lineage-specific families in the Lophotrochozoa, Mollusca, Gastropoda, or Bivalvia. The evolution of new genes may be linked to new functions, and as a consequence we consider the expression of several of these genes in adult tissues and staged embryos by a combination of transcriptome mining, RT-PCR and in situ hybridization.

Materials and Methods

Data Set Collection and Identification of bHLH Genes

The sequences of *C. gigas* (genome version oyster_v9) bHLHs were retrieved from the OysterBase (http://www.oysterdb.com/; last accessed March 16, 2017), *P. fucata* (genome version 1.0) bHLHs from the OIST Marine Genomics Unit (http://marinegenomics.oist.jp/genomes/gallery; last accessed March 16, 2017) (Takeuchi et al. 2012, 2016) and from Gyoja and colleagues (Gyoja 2014). The genome data of *L. gigantea* (version Lotgi1) were retrieved from The Joint Genome Institute (JGI: http://genome.jgi-psf.org/Lotgi1/Lotgi1.home.html; last accessed March 16, 2017), *Biomphalaria glabrata* (version BglaB1) from VectorBase (https://www.vectorbase.org/organisms/biomphalaria-glabrata; last accessed March 16, 2017), *P. vulgata* from DOI: 10.5287/bodleian:xp68kh25x (Kenny et al. 2015), and *H. robusta* (version 1.0) from the JGI (http://genome.jgi-psf.org/Helro1/Helro1.info.html; last accessed March 16, 2017). Data for the brachiopod *L. anatina* (version 1.0) (Luo et al. 2015) were accessed via the web browser for this organism (http://marinegenomics.oist.jp/lingula/viewer?project_id=47; last accessed March 16, 2017). Lists of previously analyzed genes for three species (*Homo sapiens, Drosophila melanogaster*, and *C. teleta*.) were obtained from the relevant publication (Simionato et al. 2007).

The whole complement of *H. sapiens* and *D. melanogaster* bHLHs were used as query sequences in BLAST searches of mollusc and annelid genome data. Searches were performed at low stringency (e-value ≤ 1) in order to obtain divergent members relative to those of *H. sapiens* and *D. melanogaster*. Hits were then assessed against the PFAM database with an e-value of $1e^{-5}$ to identify the bHLH domain. All identified bHLH proteins were searched against the PFAM (http://pfam.sanger.ac.uk/search; last accessed March 16, 2017) and SMART (http://smart.embl-heidelberg.de; last accessed March 16, 2017) databases to enable trimming of the sequence to the bHLH domain. Preliminary phylogenetic analyses were performed on the whole data set for each species. If we failed to identify members of bHLH families of orthologs that we expected to find based on their distribution in other taxa, additional similarity searches using TBLASTN against the whole-genome assembly for that species were performed specifically using members of the missing families as queries. In a few cases, this identified additional bHLH domains missed in the first BLAST search.

Phylogenetic Analysis and Classification

A multiple alignment was produced using MAFFT 7.221 (Katoh and Standley 2013) with the E-INS-I algorithm for the amino acid sequences of the bHLH domain. The resulting alignment belonging to the bHLH domain region (711 sequences, supplementary files 1 and 2, Supplementary Material online) was used to perform maximum likelihood

(ML) phylogenetic analyses with the program RAxML (Stamatakis 2006) using the evolutionary model LG + Gamma + Invariant; 1,000 replicates were performed to obtain bootstrap support values. Bayesian Inference (BI) phylogenetic analysis was conducted with MrBayes 3.2.2 (Ronquist and Huelsenbeck 2003). Four Markov chains were run for 3×10^6 generations, with sampling performed every 100 generations, to yield a posterior probability distribution of 10^4 trees. The first 25% of the trees were discarded when compiling summary statistics and consensus trees. A second alignment was produced after removal of all sequences from *H. robusta* and *C. teleta*, and most from *D. melanogaster*, though retaining families *Delilah*, and *Clockwork orange* which are absent from *H. sapiens*. The plant *Arabidopsis thaliana* bHLH gene *POPEYE* (NP_190348.1) domain was used as the outgroup in phylogenetic analyses. We also conducted one family-specific phylogenetic analysis, on the new gene family *SOHLH* (see below) to establish which lineages we could detect this gene in. We identified potential *SOHLH* orthologs from GenBank from about 120 species using BLAST searches (supplementary file 3, Supplementary Material online), and analyzed these genes by molecular phylogenetics as above, using human Group A sequences as outgroups.

Gene Expression in *C. gigas* Assessed by Transcriptomics

Transcriptome data from multiple adult organs and developmental stages for *C. gigas* were obtained from the NCBI gene expression omnibus (accession GSE31012) and the supplementary materials of the associated publication (Zhang et al. 2012). Corresponding gene expression levels (measured by fragments per kilobase per million mapped reads: FPKM) were calculated using HISAT2, StringTie, and Ballgown (Pertea et al. 2016). This allowed us to identify gene models and hence expression levels for several bHLH genes not previously annotated (*CgBgasHLH, CgCoe, CgLjuvHLH, CgLobHLH2, CgLcleavHLH_2, CgOrphan1, CgOrphan2*). We divided the 38 sequential developmental stages (names and abbreviations as previously defined; Zhang et al. 2012) into nine main stages: egg (E), cleavage (two cells, TC; four cells, FC; and early morula, EM), postcleavage embryo (morula, M; early gastrula, EG; and gastrula, G), trochophore (trochophore stages 1–5, T1–T5), D-shape larvae (early D-shape larvae, ED1, ED2; D-stage larvae, D1–D7), umbo larvae (early umbo stages, EU1, EU2; umbo stages U1–U6; late umbo stages, LU1, LU2), pediveliger larvae (P1 and P2), spat (S), and juvenile (J). Heat maps of bHLH gene expression were performed using the R packages heatmap 2 and ggplot 2.

Cloning, RT-PCR, and in Situ Hybridization

Primers used for amplification of bHLH and reference (elongation factor 1 alpha, *EF1*α) gene sequences from *P. vulgata* and *C. gigas* are shown in the supplementary table S1, Supplementary Material online. Amplified fragments were cloned into pCRII (Inivitrogen) and verified by sequencing. For in situ hybridization, digoxygenin-labeled probes were synthesized from cloned fragments in both sense and antisense directions. In situ hybridization of *P. vulgata* embryos was carried out as previously described (Shimeld et al. 2010). This method was also adapted for in situ hybridization of *C. gigas* embryos. For all experiments sense and antisense probes were analyzed in parallel, along with a positive control with a gene of known expression pattern.

Criteria for Inference of Evolutionary Relationships

In defining orthology groups using phylogenetic trees we followed the criteria adopted by previous analyses (Ledent and Vervoort 2001; Simionato et al. 2007; Gyoja 2014): If genes from two or more organisms formed a single clade with a bootstrap value > 50 by ML or a posterior probability > 0.50 by BI, they were considered to constitute an orthologous family. For some orthologous families, such as *Bmal* or *HES*, this criterion was relaxed as discussed in previous studies (Simionato et al. 2007; Zheng et al. 2009; Skinner et al. 2010; Liu et al. 2013; Gyoja 2014). Genes that could not be assigned to a family were categorized as "orphan" genes (Simionato et al. 2007).

To further examine the classification of orphan genes and to provide more detailed study of proposed ortholog groups that failed to show strong support in the trees conducted with all sequences, we undertook ingroup analysis, as conducted by Wang et al. (2008). We first selected clades of sequences from the large trees. We added outgroups (deriving from the closest gene families that showed good support in both ML and BI analyses) and realigned the sequences before running ML and BI analyses. This was undertaken for the following groups: The Hey/Hes/HELT/Clockwork Orange group; the Achaete–Scute group (including *ASCa, ASCb, ASCc, MtrochHLH, MgasHLH*, and several orphan genes); the HIF/Trh/Sim group (including *Ltclock*); the Mlx/Gmlx/Tf4 group; the Mnt/Myc/Mad/Max group; and the Oligo/Beta3 group.

Gene linkages can reflect ancestral arrangements or derive from lineage-specific genome reorganization brining previously unlinked genes together. We applied two principles to interpreting these data. First, linked genes from different bilaterian bHLH families were only considered as evidence of ancestral arrangements when they were shared by more than one lineage. Second, the closer together genes were in the genome, the more likely it was considered the pair evolved by tandem duplication.

Inference of Gene Losses and Gains

To evaluate when families of bHLH genes might have been lost or gained, we mapped the distribution of orthologs onto the phylogeny of the organisms we analyzed. The phylogeny used is that supported by recent phylogenomic studies (Kocot et al. 2011; Struck et al. 2011; Weigert et al. 2014). We also included data from studies of bHLH gene diversity in other lineages, including amphioxus and insects, to help identify when in a lineage an inferred evolutionary event occurred. Based on this we inferred the ancestral state (in terms of bHLH gene complement) of each node, assuming no convergence or horizontal transfer had taken place. Gene losses were inferred when a lineage was expected to have a gene based on the inferred ancestral gene family complement, but the gene was not detected. However this can also reflect missing data, and the genomes analyzed here, and in published studies of other species, vary in assembly quality and coverage. Accordingly we only conclude a loss where a gene was absent from two or more sister lineages. Gene gains were inferred when a new gene grouping, showing robust support in phylogenetic analyses but falling outside of other defined families, included genes from more than one lineage.

Results and Discussion

Identification of bHLH Gene Sequences

We extracted the bHLH genes from the *P. vulgata* and *C. gigas* (molluscs), *H. robusta* (annelid) and *L. anatina* genomes, and searched available data for the mollusc *Biomphalaria glabrata*. We also extracted bHLH genes from five other species (*H. sapiens*, *D. melanogaster*, *L. gigantea*, *P. fucata*, and *C. teleta*), which have been studied previously (Moore et al. 2000; Ledent et al. 2002; Simionato et al. 2007; Gyoja and Satoh 2013). We found the same 118 and 59 bHLH genes as in previous studies of *H. sapiens* and *D. melanogaster*, respectively, though added two additional human genes (*SOHLH1* and *SOHLH2*; see below). We identified 20 additional genes in *L. gigantea*, 4 additional genes in *P. fucata*, and 21 additional genes in *C. teleta*. The majority of the additional genes are lineage-specific paralogs, several of which are found in clusters (analyzed further below). However in some cases previously unidentified members of conserved families were also found, for example, *Hand* in *C. teleta* and *Mad* in *L. gigantea*. Thus, a total of 72 bHLH genes from *P. vulgata*, 83 from *L. gigantea*, 67 from *B. glabrata*, 88 from *C. gigas*, 68 from *P. fucata*, 85 from *C. teleta*, 70 from *H. robusta*, 77 from *L. anatina*, 120 from *H. sapiens*, and 59 from *D. melanogaster* were retrieved (table 1, supplementary figs. S3–S18, Supplementary Material online).

Phylogenetic Analyses and Classification of Orthologous Families

We constructed molecular phylogenetic trees to support bHLH gene classification. Recent studies have reported good resolution (with respect to orthologous family classification) in bHLH gene trees derived from genes from the test species together with genes from well-annotated vertebrate (*M. musculus* or *H. sapiens*) and insect (*D. melanogaster*) genomes (Wang et al. 2008; Gyoja et al. 2012). We hence adopted this approach, initially developing alignments of *H. sapiens* and *D. melanogaster* bHLH sequences with one of the mollusc bHLH gene sets (supplementary figs. S3–S12 and tables S2–S6, Supplementary Material online). We also made additional multiple alignments; one with the bHLHs from *H. sapiens* and *D. melanogaster* plus those from the two annelids species, *H. robusta* and *C. teleta* (supplementary figs. S13–S16 and tables S7 and S8, Supplementary Material online), one with sequences from *L. anatina*, *H.sapiens* and *D. melanogaster* (supplementary figs. S17 and S18 and table S9, Supplementary Material online), and one with 711 sequences including all mollusc bHLH genes, for which trees are shown in figure 1 and supplementary figure S1 (BI analyses) and figure S2 (ML analysis), Supplementary Material online. bHLH genes were subdivided into corresponding bHLH orthologous families and named according to the initials of the species name and orthologous family name. For example, the *P. vulgata* and *C. gigas* members of the Coe gene family become *PvCoe* and *CgCoe*, respectively. Where two or more members of a family were identified in one species, they are named "1," "2," "3," and so forth, such as *CgTwist1* and *CgTwist2*. These data are summarized in table 1. In all species, a number of genes did not fall into these previously defined orthology families. We will consider the classified genes, then return to these "orphan" genes and their evolutionary history.

Group A bHLH Genes

Group A (table 1) contains many well-studied genes involved in cell type specification, including *Neurogenin, NeuroD, Hand, MyoD, Oligo*, and *Twist*. Our classification is similar to most previous analyses, with two exceptions. We found evidence for multiple Achaete-Scute (Asc) families (as also described by Gyoja and Satoh [2013]). Within the gene families variously known as *Nato3*, *Fer1–3*, *48-related1–3* and *Ptfa* and *Ptfb(Fer1)*, our analysis also concurs with that of Gyoja and Satoh (2013): We find three families overall, *Ptfa/48-related3/Fer3/Nato3*, *Ptfb/48-related1/Fer1*, and *48-related2/Fer2*. We list these using conjoined names in table 1 and supplementary tables S2–S9, Supplementary Material online, to try and preserve continuity between studies, and use *Ptfa, Ptfb* and *48-related2* as shortened versions in figures.

Other Group A families followed previous classification and hence nomenclature (Ledent and Vervoort 2001; Simionato et al. 2007; Skinner et al. 2010). This group includes 24 families, with mollusc and annelid representatives in all except *Mesp*. As *Mesp* is well defined in *H. sapiens*, and in *D. melanogaster* and other insects (e.g., Dang, Wang, Chen, et al. 2011), we infer it to be lost in the lophotrochozoans we studied, and hence probably early in the lophotrochozoan lineage.

Although all other Group A families were represented in *P. vulgata, L. gigantea, C. gigas*, and *P. fucata*, we failed to identify a small number of families in the other

Table 1
Classification of bHLH Genes from Eight Lophotrochozoan Species, Compared with Human and *Drosophila*

Group	Family (49)	*Homo*	*Drosophila*	*Patella*	*Lottia*	*Biomphalaria*	*Crassostrea*	*Pinctada*	*Capitella*	*Helobdella*	*Lingula*
		sapiens	*melanogaster*	*vulgata*	*gigantea*	*glabrata*	*gigas*	*fucata*	*teleta*	*robusta*	*anatina*
A (24)	ASCa	2	4	3	3	2	3	3	1	1	3
	ASCb	3	0	1	1	1	1	1	1	1	1
	ASCc	0	0	1	1	1	1	1	2	0	1
	Atonal	2	3	1	2	1	2	1	1	0	1
	Beta3	2	1	1	2	1	1	1	1	1	1
	Delilah	0	1	3	3	0	1	1	1	0	2
	E12/E47	4	1	1	1	1	1	1	1	2	2
	Hand	2	1	1	1	1	1	1	1	2	1
	Mesp	3	1	0	0	0	0	0	0	0	0
	Mist	1	1	1	1	1	1	1	1	2	1
	MyoD	4	1	1	1	1	1	1	1	2	1
	MyoRa	2	1	1	1	1	1	1	1	1	2
	MyoRb	2	0	1	1	1	1	1	0	0	1
	Net	1	1	1	1	1	1	1	1	1	2
	NeuroD	4	0	1	1	1	1	1	2	1	2
	Neurogenin	3	1	2	2	1	1	1	1	0	1
	NSCL	2	1	1	1	1	1	1	1	3	1
	Oligo	3	0	2	2	3	2	1	2	5	2
	Paraxis	2	1	1	1	0	1	1	1	1	1
	PTFa/Fer3	1	1	3	3	3	2	3	6	4	4
	PTFb/Fer1	1	1	2	2	2	2	2	1	1	1
	48 related2/Fer2	1	1	1	1	1	1	1	2	1	0
	SCL	3	1	1	1	1	1	1	1	1	1
	Twist	2	1	2	2	2	2	2	2	2	1
B(12)	AP4	1	1	1	1	1	1	1	1	2	1
	Figα	1	0	0	0	0	0	0	0	0	0
	Mad	5	0	1	1	1	1	1	1	1	1
	Max	1	1	1	1	1	1	1	1	1	1
	MITF	5	1	1	1	1	4	0	1	2	1
	Mlx	2	1	2	2	2	2	1	2	2	3
	Mnt	1	1	0	1	1	1	1	1	1	1
	Myc	4	1	1	1	5	1	1	1	2	1
	SRC	3	1	1	1	0	1	1	1	0	1
	SREBP	2	1	1	1	1	1	1	1	2	2
	TF4	1	1	1	1	1	1	1	1	1	2
	USF	3	1	1	1	1	1	0	1	3	1
C (8)	AHR	2	1	1	1	1	1	1	1	3	1
	ARNT	2	1	1	1	1	1	1	4	2	4
	Bmal	2	1	1	1	1	1	1	1	0	0
	Clock	2	1	1	1	1	1	1	1	0	1
	Cranky	1	1	1	0	1	1	1	1	1	1
	Hif	3	1	1	1	1	1	1	2	1	1
	Sim	2	1	1	1	1	1	1	1	2	1
	Trh	2	1	1	1	1	1	1	1	1	1
D (1)	Emc	4	1	1	1	0	1	1	1	1	1
	Pearl	0	0	0	0	1	1	1	4	0	0
E (4)	Hey	4	1	2	1	1	1	1	2	1	1
	Helt	1	0	0	1	0	1	0	0	0	0
	HES	9	11	6	11	5	10	6	5	5	13
	Clockwork Orange	0	1	1	1	1	2	1	1	0	2
F (1)	Coe	4	1	1	1	1	1	1	1	1	1
No classification (15)		3	2	9	13	8	19	12	16	3	12
Total		118	59	72	83	67	88	68	85	70	88

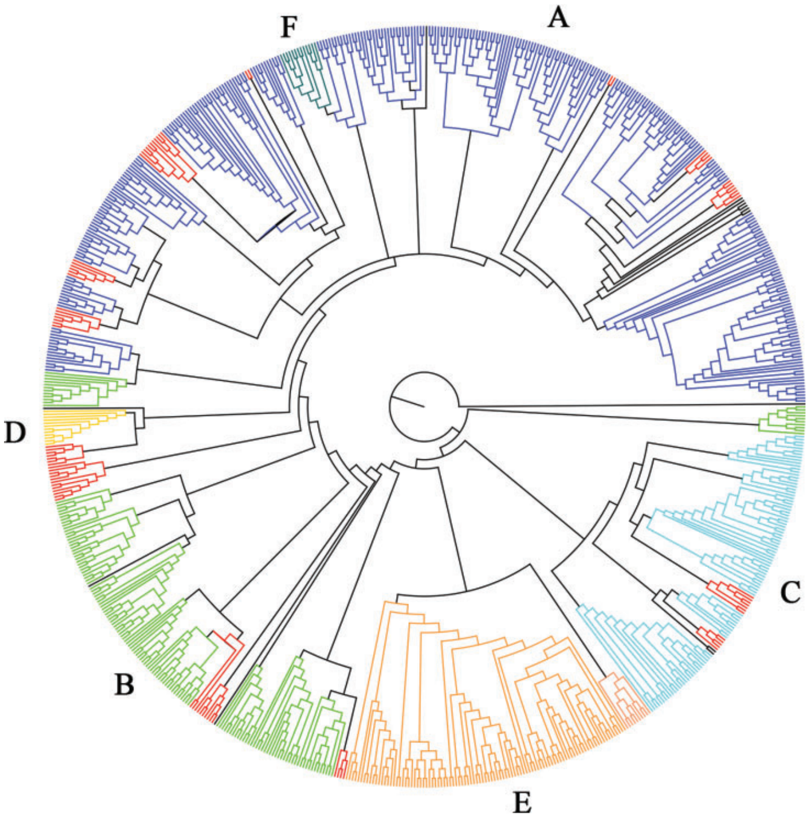

Fig. 1.—Molecular phylogenetic analysis of mollusc bHLH genes. The tree, shows the evolution the bHLH gene complement of ten complete bilaterian genomes (eight lophotrochozoans including five molluscs, two annelids and one brachiopod, plus representatives of Chordata, *Homo sapiens*, and Arthropoda, *Drosophila melanogaster*) and was constructed using BI. The interspersed red branches denote lophotrochozoan- or mollusc-specific genes. Different colors of branches denote Groups A–F. The same tree, with branch lengths, support values and gene names, is shown in the supplementary figure S1, Supplementary Material online, and the same data set analyzed by ML in the supplementary figure S2, Supplementary Material online.

lophotrochozoan species examined (table 1). One, *MyoRb*, could also represent a gene loss as it is absent from both annelid data sets. Some evidence of lineage-specific duplication (considered further below) also emerged: For example, *ASCc* and *Ptfa(Fer3)* show expansion in lophotrochozoans, with six *Ptfa(Fer3)* genes identified in *C. teleta*. The leech *H. robusta* genome also contained more members of some families (*Hand, E12/E47, MyoD, NSCL, Oligo*) than the other lophotrochozoans, suggesting considerable duplication in this lineage.

Group B bHLH Genes

Twelve families are included in Group B as described by previous studies (Simionato et al. 2007; Skinner et al. 2010; Gyoja and Satoh 2013). Lophotrochozoan bHLH genes were distributed in 11 of these families: The missing family is *Figα*, which was previously identified only in *H. sapiens* and other chordates and hence may not deserve bilaterian family status (discussed more below). As with Group A genes, some families (*AP4, MITF, SREBP*, and *USF*) have undergone duplication in the *H. robusta* lineage, whereas the *MITF* family has expanded in *C. gigas* to four genes, and *Myc* has undergone duplication in the *B. glabrata* lineage to give five genes.

Group C bHLH Genes

Group C genes encode a PAS domain as well as the bHLH domain, and include well-characterized genes involved in adult physiology (e.g., *HIF*, *ARNT*) and the circadian clock (*Bmal*, *Clock*). Eight families have been described previously (Ledent and Vervoort 2001; Satou et al. 2003; Gyoja and Satoh 2013), with single lophotrochozoan genes found in most of these (table 1).

Group D, E, and F bHLH Genes

Groups D and F are usually considered to contain a single gene family each, *Emc/Id*, and *Coe*, respectively. Both were in single copy in all the lophotrochozoan genomes examined, with the exception that we failed to identify an *Emc* gene in *B. glabrata*. In our analysis the gene family *Pearl*, defined by Gyoja et al. (2012), also fell into Group D and we include it here in table 1. Group E genes encode a hairy/orange domain as well as the bHLH domain and include the *Hey* and *Hes* families, with well-defined roles in Notch signaling, as well as the more recently described *Clockwork orange* gene family which is absent from *H. sapiens* and other vertebrates (Gyoja and Satoh 2013), and a fourth gene family known as *Heylike* (*HeyL*) or *HELT* (Gazave et al. 2014; Gyoja 2014). We identified putative *HELT* genes from *L. gigantea* and *C. gigas*, but not from the other species.

Both *Hey* and *Hes* families show evidence of duplication, with this most extensive in the *Hes* gene family, with the number of genes in molluscs ranging from 5 to 11 (table 1). Duplications in the *Hes* family have been noted and studied previously in other lineages than those studied here (Jimenez-Delgado et al. 2006; Zhou et al. 2012), and it has proven challenging to resolve orthology and paralogy relationships. We checked the chromosome location of *Hes* genes, and found that many of them located on the same scaffolds (see below) indicating extensive tandem duplication.

A Summary of Classified Mollusc bHLH

In molluscs, most (49/51) of the bHLH gene families that have been described in other bilaterians were found to have at least one member. This suggests good coverage from the draft genomes searched here, and that very few genes have been lost in this lineage. The two missing gene families are *Mesp* and *Figα*. The former we would have expected to find in lophotrochozoans as orthologs are present in both deuterostomes and ecdysozoans. This loss probably happened early in lophotrochozoan evolution as we failed to identify a *Mesp* gene in either annelid species, whereas an analysis of bHLH genes of the platyhelminth *Schmidtea mediterranea* also failed to recover a *Mesp* ortholog (Cowles et al. 2013). *Figα* has only been found in amphioxus, zebrafish, reptile, and mammal genomes to date, and hence may be a chordate-specific gene group and not have the bilaterian-wide orthology family status of other bHLH families (Simionato et al. 2007; Wang et al. 2009; Liu et al. 2013; Wang et al. 2015). We also failed to find members of an additional gene family, *Peridot*, whose existence was suggested by comparison of coral and invertebrate deuterostome genes (Gyoja et al. 2012). This gene is not listed in table 1.

The total number of bHLH genes identified was around 80 or higher in most molluscs (table 1; the figures of 67 in *B. glabrata* and 68 in *P. fucata* are probably underestimates due to poorer sequence data for these species). In *L. gigantea* and *P. fucata* we found 83 and 68 bHLHs, respectively, more than previously reported (Simionato et al. 2007; Gyoja and Satoh 2013). In *L. gigantea* and *C. gigas* 24 and 20 genes, respectively, are inferred to have evolved by duplication within a family We therefore conclude that gene duplications within defined orthology families have had an impact on the mollusc bHLH gene diversity, with such paralogs making up about 25% of the repertoire.

Novel bHLH Genes in Lophotrochozoans

In all our analyses, a number of bHLH genes from each species failed to group with any of the previously named bHLH ortholog families identified through analysis of genes from *H. sapiens, D. melanogaster*, and other model organisms. Two are members of additional bHLH families (*Amber* and *Pearl*) that recent studies have suggested are genuine bilaterian orthology groups, but whose members have been lost by many model species (Gyoja et al. 2012; Gyoja and Satoh 2013). Others have not been described before. All are listed in table 2, together with either the name given to them in previous studies, or the name we now propose for these genes. In proposing family names we have used the phylogenetic restriction of the gene family and/or expression. Families restricted to lophotrochozoans begin with L, molluscs begin with M, bivalves begin with B, and gastropods begin with G. For example, *GmlxlHLH* means Gastropod Mlx-like gene, whereas *LcleavHLH* is a lophotrochozoan gene prominently expressed at cleavage stages.

Of the three families likely to be bilaterian wide, *Pearl* was previously described from *P. fucata* as well as cnidarians, an annelid and a spider, whereas *Amber* was found in mollusc, annelid and cnidarian genomes (Gyoja et al. 2012). Identification of genes from these families in other molluscs and annelids is therefore not surprising. We also found a new family to be bilaterian wide, at least primitively, which we name *SOHLH* after the human genes *SOHLH1* and *SOHLH2*. To examine the distribution of *SOHLH* genes more closely, we extracted candidate genes from a wide diversity of species (supplementary file 3, Supplementary Material online) and evaluated their potential orthology by molecular phylogenetics (supplementary fig. S22, Supplementary Material online). Our analysis shows support for candidate *SOHLH* genes from multiple species including several cnidarians, confirming its status as an ancestral bilaterian gene. Most cnidarians and lophotrochozoans had more than one *SOHLH* gene, which appeared to group into two clades (supplementary fig. S22, Supplementary Material online); however support values within the *SOHLH* clade were generally too low to draw any firm conclusions from this. The distribution of *SOHLH* genes in other lineages was patchy: In vertebrates we identified *SOHLH* genes from numerous mammals, plus the coelacanth, shark and gar (a basal actinopterygian), but not from several model species including zebrafish. In the Ecdysozoa we identified *SOHLH* genes from a spider and a priapulid, but not from any insect. These data suggest multiple losses of *SOHLH* genes in bilaterian evolution. The placement of the *SOHLH* relative to other families is poorly defined in our analyses, though it usually lies close to Group A, B, or E genes. In our summaries we include it within the Group A for now, though this is not definitive.

Eleven other groups of gene that are restricted to lophotrochozoans, gastropods, or bivalves are listed in table 2, and

Table 2
Recently Defined or Novel bHLH Families

Gene Family	Named by or Novel (Reason for Name)	BI Support	Lineage Distribution
Amber	(Gyoja et al. 2012)	1	Prebilaterian (Mollusc, Annelid, Cnidaria)
Pearl	(Gyoja et al. 2012)	1	Prebilaterian (Bivalve, Cnidaria, Annelid, Spider)
SOHLH	(Suzuki et al. 2012)	1	Prebilaterian (Human, Mollusc, Annelid, Ecdysozoan, Cnidaria)
LHLH1	Novel (Lophotrochozoan)	0.85	Lophotrochozoan only
LHLH2	Novel (Lophotrochozoan)	1	Lophotrochozoan only
LtcHLH	Novel (Linked to cranky)	1	Lophotrochozoan only
LoblHLH	Novel (Olig, Beta3-like)	1	Lophotrochozoan only
LtclockHLH	Novel (Linked to clock)	0.9	Lophotrochozoan only
MHLH1	Novel (Mollusc)	0.95	Mollusc only
MHLH2	Novel (Mollusc)	0.96	Mollusc only
MHLH3	Novel (Mollusc)	0.65	Mollusc only
BiHLH	Novel (Bivalve)	1	Bivalve only
GmlxlHLH	Novel (similar to Mlx)	1	Gastropod only
GaHLH	Novel (Gastropod)	1	Gastropod only

will be discussed further below in relation to their expression. In addition some genes remained unclassified; two orphan genes in *C. gigas*, seven in *C. teleta*, and one in *H. robusta*. Attempts to classify them using ingroup analyses (see Materials and Methods) were not successful, as the placement of orphan genes in such trees had similarly low support as in the larger analyses.

Linkage of bHLH Genes in Mollusc Annelid and Brachiopod Genomes

Linkage between homologous genes can reveal their evolutionary history. We examined scaffolds from the genome assemblies to determine which bHLH genes were linked. Numerous linkages were found (supplementary fig. S21, Supplementary Material online), though many did not pass criteria for inference of evolutionary relationships (outlined in the methods above) and are not discussed further. We instead focus on linkages where proximity and/or consistency across lineages allow gene origins to be determined.

Linkage of Bilaterian Orthology Family Genes: Evidence for Ancient Clusters

We found that *Amber* and *Neurogenin* are located close together on the same scaffold in *C. gigas* and *L. gigantea* (fig. 2). *Amber* genes are present in the Mollusca, Annelida and Cnidaria, but lost in *H. sapiens*, *D. melanogaster* and *C. elegans*, and *Neurogenin* is also a bilaterian gene family (Gyoja et al. 2012; Gyoja and Satoh 2013). Gyoja and Satoh's analyses and our study all show that *Amber* is more similar to *NeuroD* than *Neurogenin* based on molecular phylogenetic analysis; however, the tight linkage of *Amber* and *Neurogenin* indicates an origin of these two by duplication of a single ancestral gene. Furthermore *NeuroD* is tightly linked to *Ptfa(Fer3)* genes in both *L. gigantea* and *C. teleta*, suggesting these too were separated by a tandem duplication early in animal evolution and linkage maintained in some lophotrochozoans. We also noted linkage of *ASCc* and *ASCa*. These genes may have separated earlier in evolution (though this is not currently clear as some ASC-related genes in nonbilaterians have proven hard to classify; Gyoja 2014), but are certainly ancestral bilaterian gene families. This is therefore likely to be a very old linkage. All the above are Group A genes; however, in *C. teleta* an *ASCc* gene is linked to one of the *Hes* gene clusters, that is, Group E genes. This hints at ancient clustering of *Hes* and *ASC* genes, though as this is currently supported by only one genome and includes genes widely separated in phylogenetic analyses this cannot be concluded with certainty.

Lineage-Specific Paralog Clusters

A significant phenomenon identified in the linkage analysis was clusters of relatively recent paralogs, that is, clusters of genes within a family. This included linkage of *ARNT*, *Delilah*, *Beta3*, *Oligo*, *Pearl*, *Myc*, *LoblHLH*, *Clockwork orange*, *Twist*, *Ptfb(Fer1), Ptfa(Fer3), Hes*, and *ASC* paralogs in various lineages (fig. 2, supplementary fig. S21, Supplementary Material online). Several of these (*ARNT*, *Delilah, Beta3, Oligo, Pearl, Myc, LoblHLH, Clockwork orange*) derive from gene duplications identifiable in only one of the studied species, and hence are relatively recent. Others are shared by two or more species and are hence likely to be older. *Twist* gene linkage was identified in mollusc and *C. teleta* genomes, suggesting the tandem duplication that formed these occurred early in the lophotrochozoan lineage. *Ptfa(Fer3)* paralogs, as well as being linked to *NeuroD* as discussed above, were linked in mollusc and annelid genomes (fig. 2). This was also observed for *Ptfb(Fer1)* paralog linkage (supplementary fig. S21, Supplementary Material online). Phylogenetic resolution within these families is poor, so we cannot exclude the

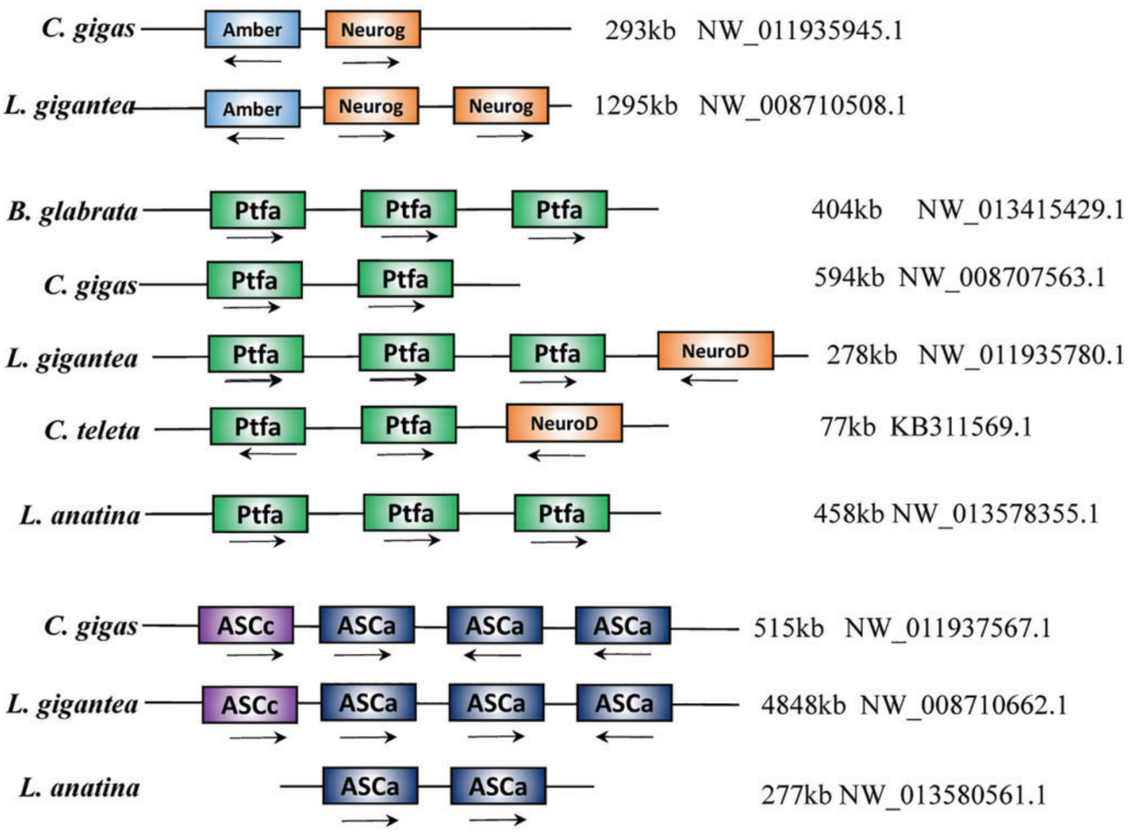

Fig. 2.—A schematic map of selected bHLH gene clusters. Different color rectangles show different genes and arrows below the scaffold indicate the transcriptional orientation of each gene. All genes shown are members of Group A. Additional linkages, plus details on intergenic distances, are described in the supplementary figure S21, Supplementary Material online.

possibility that these clusters arose independently in different lineages; however, we consider it likely they are primitive.

Paralog clusters of *Hes* and *ASC* genes were also identified (fig. 2, supplementary fig. S19, Supplementary Material online). *Hes* genes have been seen to form clusters in several lineages as discussed above. Phylogenetic trees focused on the Group E genes have previously proven insufficiently supported to convincingly determine whether this reflects common ancestry or parallel cluster evolution, though some lineage-specific tandem duplication is supported for some species (Minguillon et al. 2003; Gazave et al. 2014). Our analysis also suffers from this problem, and ingroup analyses (see Materials and Methods) did not provide additional resolution (data not shown). We can clearly conclude though that *Hes* genes are clustered in bivalve, gastropod, brachiopod, and annelid genomes. Linkage of *ASC* paralogs was also observed in gastropod, bivalve, and brachiopod genomes.

Origin of New Gene Families in Lophotrochozoans

Some of the new, lineage-specific, bHLH gene families described above were linked to previously defined orthology families; for example, *LtcHLH* to *Cranky*, *MtrochHLH* to *AHR*, *GMlxlHLH* to *Mlx2*, and *LtclockHLH* to *Clock* (supplementary fig. S21, Supplementary Material online). Although these linkages were only observed in some of the genomes, the genes were generally close together (supplementary fig. S21, Supplementary Material online) and it is unlikely such arrangements evolved by chance. Rather, it suggests these genes evolved by duplication from the linked genes, followed by elevated divergence of one copy such that its paralogy is no longer apparent on the basis of sequence identity.

Gene Duplications in the Leech, H. robusta

Finally, we noted that despite *H. robusta* possessing many lineage-specific paralogs, these genes do not generally seem to be tandemly linked in the *H. robusta* genome, with only clusters of *Oligo* and *Hes* genes showing such an arrangement. This may reflect the differing quality of the genomes analyzed here, but two alternate possibilities present: 1) Gene duplication in *H. robusta* may be qualitatively different to that in the other lineages, occurring more frequently by mechanisms that do not leave paralogs close together; and 2) the rate of break-up of tandem duplicates may be higher in the *H. robusta* lineage. These could both reflect higher rates of genome reorganization in *H. robusta*, something also indicated by the absence of clustering of paralogs detected in other lineages and which we have inferred are ancestral (*Twist, Ptfa(Fer3), Ptfb(Fer1)*).

Gene Expression Indicates Different Evolutionary Trajectories for Different Types of Paralog

Molecular phylogenetic and linkage data indicate a complex evolutionary history for bHLH genes in these taxa, with numerous duplications mapping to different points in their phylogeny and having different outcomes with respect to sequence divergence. To relate this to gene function, we used three types of data: Genome-wide transcriptome data from embryos and adult tissues of *C. gigas*, RT-PCR of selected genes from adult tissues of *P. vulgata*, and in situ hybridization on embryos of both species. Genome-wide transcriptome data cover all bHLH genes present in the *C. gigas* predicted gene set (figs. 3 and 4). We focused RT-PCR analysis (fig. 6) onto the novel genes, that is the new family (*SOHLH*) and lophotrochozoan or mollusc lineage-specific genes, as these are unstudied and we considered analyzing their expression may give clues as to their function and/or origin. However not all such genes were analyzed by RT-PCR in *P. vulgata* as one is bivalve specific (*BgasHLH*), some are missing from the *P. vulgata* assembly (*Pearl, LtcHLH, LmorHLH*), and we did not develop primers for others (*LtclockHLH, GmlxlHLH*). We validated all amplicons by sequencing, and used these for probe generation for in situ hybridization. Some genes (*LjuvHLH, MtrochHLH*) did not show signal in in situ experiments, and the accompanying figure reports only those where signal was observed (fig. 5).

Analysis of Orthology Family bHLH Gene Expression in *C. gigas* by Transcriptomics

Crassostrea gigas benefits from excellent transcriptomic resources, covering 38 developmental stages and nine adult tissues (Zhang et al. 2012). We examined read data for all *C. gigas* bHLH genes for which gene models were developed; 87 in total. An FPKM greater than 1 was considered evidence of expression (supplementary table S10, Supplementary Material online), and heat maps were used to visualize comparative levels of expression across all samples (figs. 3 and 4, supplementary figs. S19 and S20, Supplementary Material online).

Several genes were found to be expressed at all developmental stages, and by all adult organs, including *CgE12/E47, CgSREBP, CgTF4, CgUSF, CgMad, CgMax*, and *CgEMC*. *CgEMC, CgMad*, and *CgMax* were especially highly expressed. Functional analyses of Emc genes have demonstrated that they are required for multiple processes in development, and act by blocking the function of other bHLH proteins by forming heterodimers (Campuzano 2001). Max is also a heterodimeric partner that antagonizes Myc and Mad transcriptional activity (Blackwood and Eisenman 1991; Ayer et al. 1993). Such widespread expression is consistent with such general inhibitory roles.

Most genes showed evidence of differential expression through developmental stages or in different adult tissues. For example, expression of the *CgNet, CgCoe*, and *CgNeuroD* genes were highest from the trochophore stage to D-shape stage, indicating these genes might play role in metamorphosis. Some genes, such as *CgASCc*, *CgAtonal1*, and *CgBeta3*, were either not expressed at all or had relatively low expression (supplementary table S10, Supplementary Material online). These genes might not play a role during embryogenesis and larval development, though as transcriptomic data conflate the level of expression per cell with the number of cells expressing a gene it remains possible they are expressed in a few cells at one or more stages but that this is diluted out in whole-organism transcriptomes.

Examining the transcriptome data at the level of the Groups A–F reveals interesting differences in the general patterns of expression. Group B genes (with the exception of duplicated MITF genes, see below) are generally widely expressed in all developmental stages and adult tissues. The *CgMad* and *CgMax* genes, discussed above, fall into this group, and such wide expression is consistent with general roles in transcriptional regulation in most cell types and tissues. Some Group C genes also show widespread if lower expression, for example, *CgARNT*, *CgBmal*, *CgHif*, and *CgClock*. This is consistent with their predicted roles in physiological processes. Most Group A and E genes show much more specific expression, either through developmental stages, adult tissues, or both. Group E genes are involved in Notch signaling, which is likely to play spatially and temporally specific roles in development and adult tissue homeostasis. Group A genes include many cell type-specific genes which would also be expected to show highly specific expression in both developing stages and adult tissues.

We also examined cases where *C. gigas* showed duplicated genes within an orthology family, including *Ptfa(Fer3), Ptfb(Fer1), Twist*, and *HES*. Interestingly, duplicated paralogs in the *Ptfa(Fer3), Ptfb(Fer1)*, and *Twist* families showed divergent expression in development, but, for *Twist* and *Ptfb(Fer1)*, more similarity in adult tissues (figs. 3 and 4). *HES* genes showed a more complex pattern, with several genes (*CgHES1*, *CgHES2*, *CgHES3*, *CgHES4, CgHES5*) showing high expression in early development, whereas others (*CgHES6*, *CgHES7*) showed expression later in development and in adult organs. Some similarity between linked HES genes was also observed, for example, between *HES1* and *HES2* in developmental stages, but this was not consistent and the complexity of *HES* gene evolutionary history (discussed above) makes this difficult to interpret robustly as mollusc *HES* gene relationships are poorly resolved.

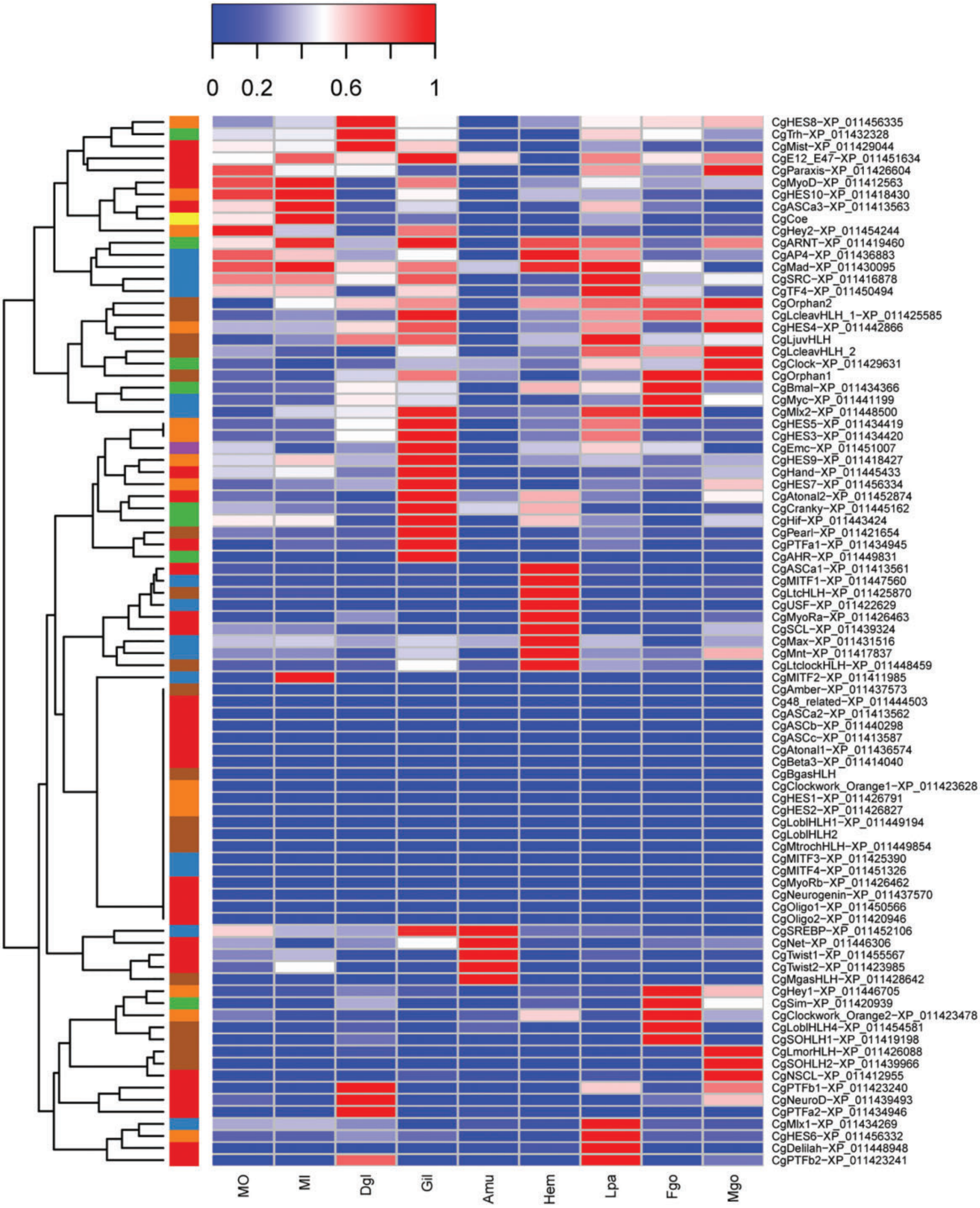

Fig. 3.—The expression of bHLH genes among nine different tissues in *C. gigas*. Hierarchical clustering analysis of *C. gigas* transcriptome data from adult organs for all bHLH genes. Distances were measured for rows (genes) and clustered using pairwise complete-linkage. Normalization and median centering was conducted for each row: red blocks represent a high level of expression whereas blue blocks represent a low level, with white the median. For absolute read counts please refer to the supplementary table S10, Supplementary Material online. Organs are abbreviated as: Mo, the outer edge of mantle; Mi, the inner pallial of mantle; Fgo, female gonad; Mgo, male gonad; Amu, adductor muscle; Hem, hemocyte; Dgl, digestive gland; Gil, gills; and Lpa, labial palp.

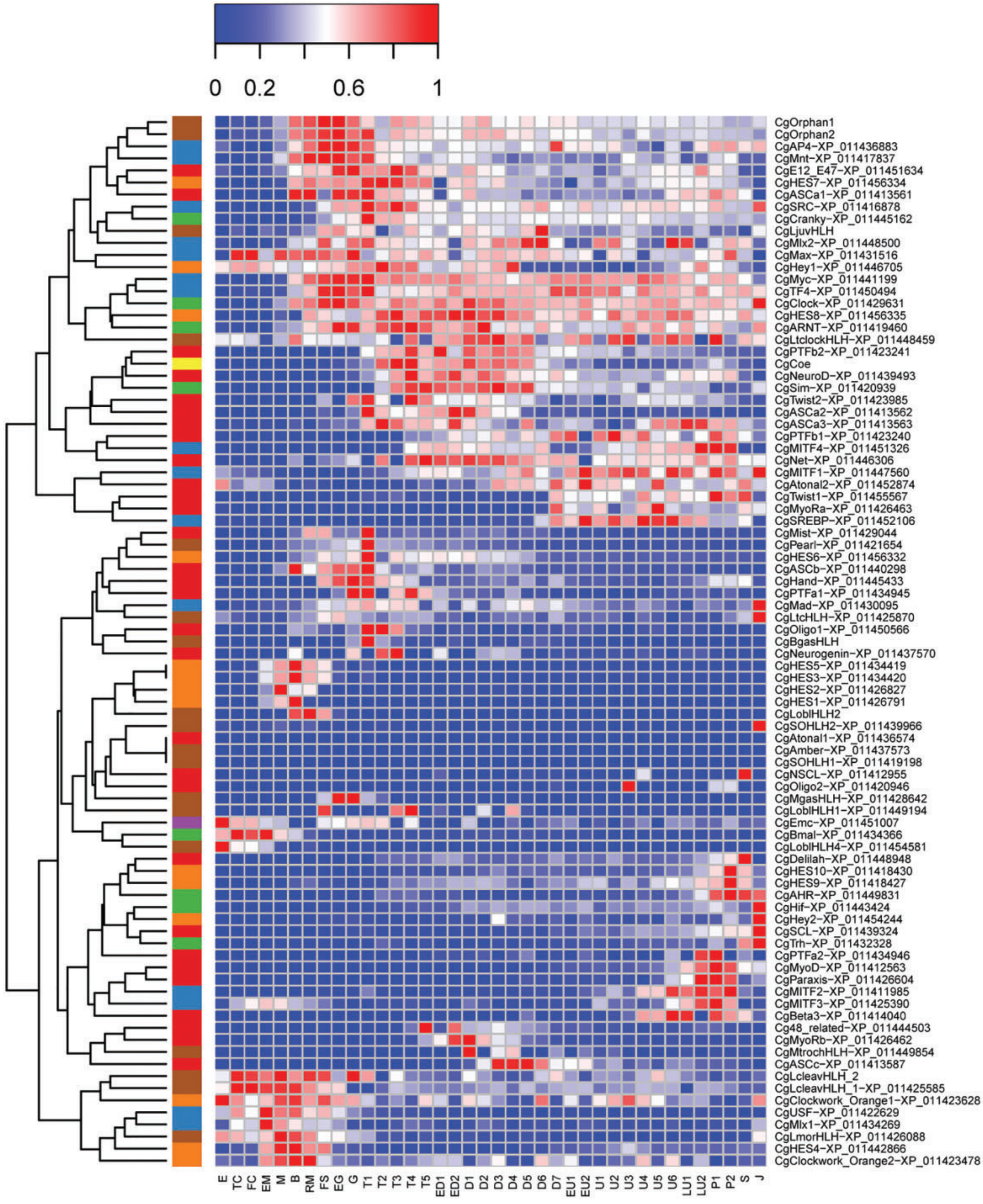

Fig. 4.—Temporal expression pattern for bHLH genes in developing *C. gigas*. Hierarchical clustering analysis of *C. gigas* transcriptome data from developmental stages. Distances were measured for rows (genes) and clustered using pairwise complete-linkage. Normalization and median centering was conducted for each row: red blocks represent a high level of expression whereas blue blocks represent a low level, with white the median. For absolute read counts please refer to the supplementary table S10, Supplementary Material online. The 38 developmental stages are: E, egg; TC, two cells; FC, four cells;

Expression of Novel Bilaterian bHLH Gene Families

The *Amber* and *Pearl* gene families have been previously defined (Gyoja et al. 2012), but their expression has been little studied. *CgAmber* showed little evidence of expression from the transcriptome data at any stage or in any tissue, and we did not further examine its expression in *C. gigas* or *P. vulgata*. *CgPearl*, however, was expressed at all developmental stages and in all tissues with the possible exception of adductor muscle where the FPKM was about 2 (supplementary table S10, Supplementary Material online). This suggests widespread and possibly ubiquitous expression. *Pearl* is missing from the limpets so we could not test expression in *P. vulgata*; however, we were able to examine expression in *C. gigas* embryos by in situ hybridization. Surprisingly, *CgPearl* was expressed very specifically during embryogenesis, by one or two cells in the blastula and gastrula stages (fig. 5*D–F*).

We also defined a new bHLH orthology family, which we named *SOHLH* after the prototypical human and mouse genes. In mammals the *SOHLH* genes function during spermatogenesis and oogenesis (Ballow, Meistrich, et al. 2006; Ballow, Xin, et al. 2006; Pangas et al. 2006). *C. gigas SOHLH1* and *SOHLH2* are not detectable in developmental stages (supplementary table S10, Supplementary Material online) but are strongly expressed in the female and male gonads, respectively (fig. 3), whereas RT-PCR found *P. vulgata SOHLH* to be expressed exclusively in male and female gonad (fig. 6). We hence suggest the *SOHLH* is an ancient gene family with a role in gametogenesis, and it will be interesting to study its function further as it may have a conserved role in the transition from germ cell to gamete.

Expression of Lineage-Specific bHLH Genes

The duplication and divergence of genes encoding transcription factors has been postulated to play a role in the evolution of development. Our analysis identified 11 novel gene families of lineage-specific bHLH genes (table 2), and to gain insight into their possible roles we examined their expression via the *C. gigas* transcriptome data, RT-PCR of adult tissues, and in situ hybridization of staged embryos.

Lophotrochozoan-Specific Genes

Six gene families we consider to be lophotrochozoan specific (on the basis of identification in molluscs and annelids) were defined: *LcleavHLH*, *LjuvHLH, LmorHLH, LtcHLH, LoblHLH*, and *LtclockHLH* (table 2). *LcleavHLH*, *LmorHLH, LtcHLH*, and *LtclockHLH* show expression in a variety of tissues and developmental stages; neither *CgLtcHLH* or *CgLtclockHLH* was grouped next to their linked genes (*CgCranky* and *CgClock*) in transcriptome clustering analysis, suggesting they do not share similarity in expression with their inferred linked paralogs (figs. 3 and 4; supplementary table S10, Supplementary Material online). *LjuvHLH* expression appears to be widespread at a low level in both *C. gigas* and *P. vulgata* (supplementary table S10, Supplementary Material online, fig. 6). We could not detect *LjuvHLH* expression by in situ hybridization in embryos or larvae of either species (not shown) consistent with a low level of expression.

In *P. vulgata, PvLoblHLH* was found to be expressed in male and female gonad and mantle by RT-PCR (fig. 6), and in a very specific pattern in embryos in a pair of meseodermal cells in early trochophores, becoming restricted to a small population of ventral cells posterior to the mouth (fig. 5*G–I*). The LoblHLH gene family has duplicated in *C. gigas*, yielding four genes. Their expression in early development as determined by transcriptome data is dynamic, appearing to switch successively through *CgLoblHLH4* (in oocytes and cleavage stage embryos), *CgLoblHLH2* (gastrula stages) and by the trochophore stage to *CgLoblHLH1*. In adults, *CgLoblHLH* gene expression depends on the paralog considered, but their combined expression includes digestive gland and female gonad, as in *P. vulgata*. These data suggest subfunctionalization of LoblHLH function in *C. gigas*, as the combined expression of the four paralogs is similar to the expression of the single *P. vulgata* gene.

Mollusc-Specific Genes

Two mollusc-specific genes were identified, *MtrochHLH* and *MgasHLH* (table 2). *MgasHLH* is specifically expressed around gastrula stages in *C. gigas* as determined by transcriptomics, but was not detected by in situ hybridization in this species. In *P. vulgata*, however we identified *PvMgasHLH* expression in a specific pair of cells in the posterior mesoderm, lying underneath the shell field (fig. 5*M–O*). Adult expression was identified in adductor muscle in *C. gigas*, and in male gonad in *P. vulgata.* Expression hence appears quite variable between mollusc lineages, with the possible exception of mesodermal expression at the trochophore stage. *CgMtrochHLH* appeared rather broadly expressed at a low level in *C. gigas* transcriptome data, and in situ hybridization supported weak ubiquitous expression (fig. 5*A–C*). Expression of *PvMtrochHLH* was

Fig. 4.—Continued

EM, early morula; M, morula; B, blastula; RM, rotary movement; FS, free swimming; EG, early gastrula stage; G, gastrula; T1, trochophore 1; T2, trochophore 2; T3, trochophore 3; T4, trochophore 4; T5, trochophore 5; ED1, early D-larva 1; ED2, early D-larva 2; D1, D-larva 1; D2, D-larva 2; D3, D-larva 3; D4, D-larva 4; D5, D-larva 5; D6, D-larva 6; D7, D-larva 7; EU1, early umbo larva 1; EU2, early umbo larva 2; U1, umbo larva 1; U2, umbo larva 2; U3, umbo larva 3; U4, umbo larva 4; U5, umbo larva 5; U6, umbo larva 6; LU1, later umbo larva 1; LU2, later umbo larva 2; P1, pediveliger 1; P2, pediveliger 2; S, spat; and J, juvenile.

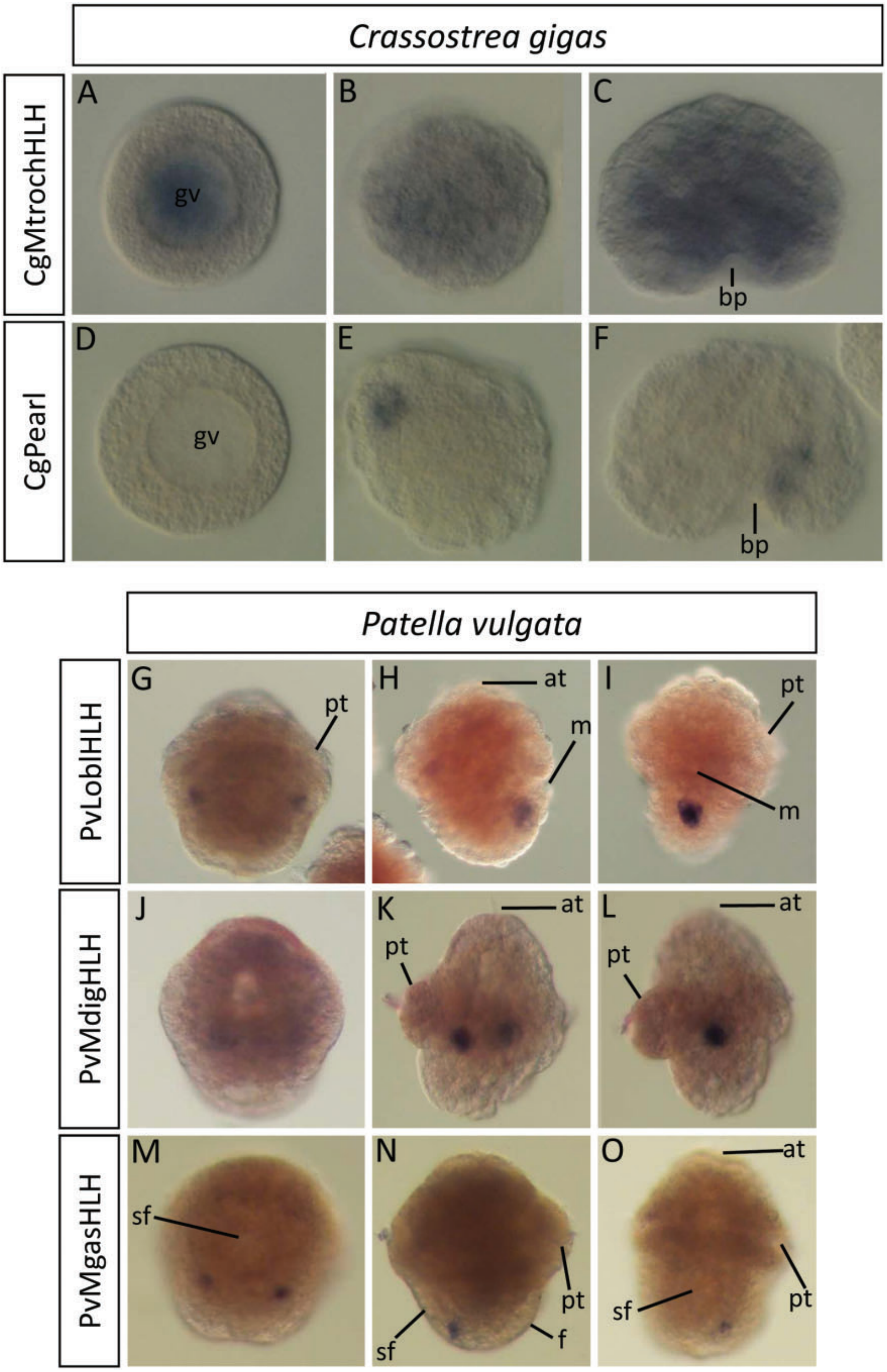

FIG. 5.—In situ hybridization of *C. gigas* and *P. vulgata* bHLH genes during embryogenesis and early larval development. In situ hybridization of *C. gigas* (*A*–*F*) and *P. vulgata* (*G*–*O*). The genes studied are shown to the left of each panel. Developmental stages are; for *C. gigas*, oocytes (*A*, *D*), blastulae (*B*, *E*), and gastrulae (*C*, *F*); for *P. vulgata* early trochophores in ventral (*G*, *J*), dorsal (*M*) and lateral (*N*) view, and late trochophores in lateral (*H*, *L*), ventral (*I*, *O*) and dorsal (*K*) view. at, apical tuft; bp, blastopore; m, mouth; pt, prototroch; sf, shell field. Genes that were studied but for which no expression was detected are not shown.

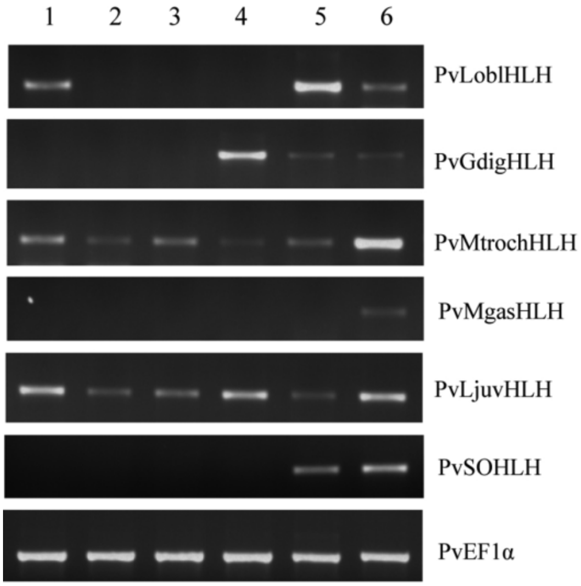

Fig. 6.—RT-PCR analysis of six novel *P. vulgata* bHLH genes expression in adult tissues. Tissues: 1, mantle; 2, foot; 3, head; 4, digestive gland; 5, female gonad; 6, male gonad. *EF1α* is shown as a positive control. Primers and size of amplification products are shown in the supplementary table S1, Supplementary Material online.

not detected by in situ hybridization, but was found in all adult tissues by RT-PCR (fig. 6).

Bivalve- and Gastropod-Specific Genes

A small number of genes were found to be restricted to just one of these mollusc lineages; a single bivalve-specific gene, *BgasHLH*, and two gastropod specific genes, *GmlxlHLH* and *GdigHLH* (table 2). *CgBgasHLH* showed expression at gastrula stages by transcriptome data, but was not detectable by in situ hybridization. *PvGdigHLH* showed expression in adult digestive gland and male and female gonad (fig. 6), and was detected by in situ hybridization, with specific expression in paired mesodermal cells posterior to the prototroch at trochophore stages (fig. 5*J*–*L*).

A Summary of Expression Data

Integrating gene expression, molecular phylogenetic, and linkage data leads to a number of generalizable points regarding bHLH gene evolution:

1. Genes in different superfamily groups differ consistently in their expression. Specifically, Group A, E and F genes tend to show stage and tissue specificity, whereas Group B, C and D genes are generally broadly if not ubiquitously expressed (figs. 3 and 4, supplementary figs. S19 and S20, Supplementary Material online).
2. Old clustered paralogs, that is, those that predate the mollusc-annelid divergence (*Ptfa(Fer3), Ptfb(Fer1), Twist*), show divergent expression in embryos, but there are hints of similar expression in adult tissues (supplementary figs. S19 and S20, Supplementary Material online). We can speculate that coregulation in adult tissues may provide a selective pressure to maintain their linkage.
3. New gene families whose evolutionary origin is inferred from linkage data (e.g., *LtcHLH*) do not show similar expression to their presumed ancestor gene.
4. New clustered paralogs, that is, those that are confined to just one lineage (*LoblHLH, Clockwork Orange*) show very similar expression in embryos, with some evidence of subfunctionalization.

Conclusions

A summary model for the evolution of the mollusc bHLH gene repertoire is shown in figure 7: This is based on our current analysis plus published analyses in other species; however, we consider it may evolve as new data from other bilaterian lineages are added. Molecular phylogenetic analysis shows that the mollusc lineage has retained all except one previously defined orthology families. We also describe one novel family, *SOHLH*, and confirm two others that are absent from many model species, *Pearl* and *Amber*. The missing family is *Mesp*, probably lost early in lophotrochozoan evolution. *Mesp* genes are involved in mesodermal patterning in chordates (Sawada et al. 2000); however, the *Drosophila Mesp* gene *Sage* is involved in salivary gland development (Fox et al. 2013). This divergence makes it hard to establish the ancestral role of *Mesp* genes, and hence why they might have been lost.

Gene duplication has played a role in the evolution of the mollusc bHLH gene repertoire (fig. 7). Our study suggests two different outcomes of such duplications. Duplication within a family can lead to multiple paralogs, which continue to share a high level of sequence identity within the DNA-binding domain; examples include those found in the *Twist, Ptfa(Fer3), Oligo*, and *MITF* families. In many instances such paralogs can be found in clusters, some of which date back to before the separation of annelid and mollusc lineages, over 540 Ma. Expression data indicate relatively recent paralogs retain some similarity in expression, whereas old paralogs diverge in embryonic expression, though not necessarily in adult tissue distribution.

Duplication must also underlie the origin of various lineage-specific genes identified in our study. The timing of these duplications varies, with some predating the annelid–mollusc lineage separation, others confined to subsets of the molluscs only. The outcome of these duplications also differs to that described above. Duplication can only have come from a preexisting gene. Following duplication the sequence has diverged such that the gene from which it duplicated can no longer be determined by molecular phylogenetics, but then stabilized such that orthology relationships are clearly discernable in the descendant species. A possible explanation is neofunctionalization at the level of DNA target sequence it binds

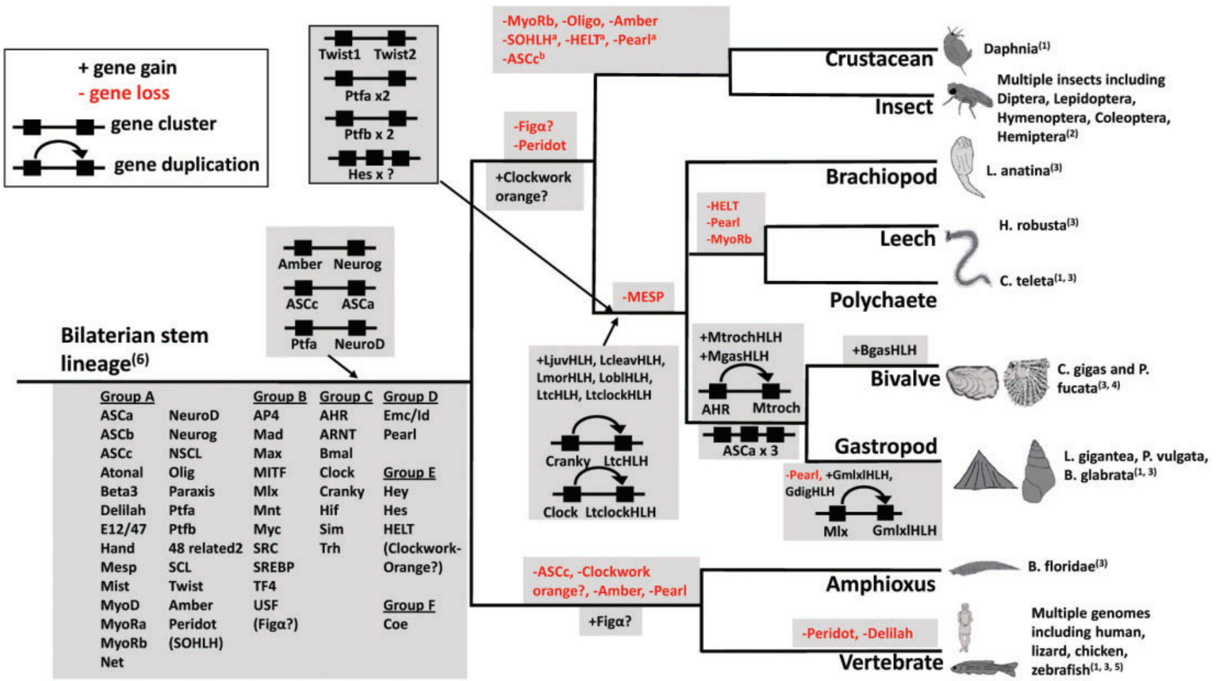

FIG. 7.—A model of bHLH gene family evolution in the Bilateria, focused on the Mollusca. The origin of gene families is indicated by a + sign below the respective branch: most predate the radiation of the Bilateria. Gene loss is mapped on in red with a − sign above the respective branch. Two genes, *Figα* and *Clockwork orange*, appear in multiple places as it is unclear if they are innovations within the Bilateria or have been lost by Protostomia and Deuterostomia, respectively. Selected gene clusters are also shown, adjacent to the branch on which we infer they evolved. For genes that originated in the lophotrochozoan lineage we show, where possible, their linkage relationship with the inferred ancestral gene from which they duplicated. Evolutionary patterns of gain and loss are inferred from bHLH gene analysis in this study combined with previously published work, as indicated by the superscript numbers listed next to each lineage tip and detailed below. In determining losses and gains we have assumed no horizontal transfer or convergence has taken place, and propose what we consider to be the simplest explanation for the distribution of genes given the species phylogeny (please see main text for how this was selected). We have shown the relationship between the major lophotrochozoan lineages as a collapsed node, as the topology here is under debate. *Peridot* is included as a gene family proposed during analysis of coral bHLH genes but absent from most bilaterians (Gyoja et al. 2012). *SOHLH* is shown in brackets at the base of the tree as its group placement is uncertain. [a]*HELT, Pearl* and *SOHLH* genes have been reported from an onychophoran (*HELT*; Gyoja 2014) and spider (*SOHLH, Pearl*; this study, Gyoja et al. 2012). [b]A putative *ASCc* gene has been reported from the insect *Tribolium castaneum*, though not as yet from other insects (Gyoja and Satoh 2013). Thus, these losses have occurred at different points during ecdysozoan evolution and are not shared by all lineages. References cited in figure: [1]Simionato et al. 2007. [2]Wang et al. 2008; Dang, Wang, Chen, et al. 2011; Liu et al. 2012; Zhang et al. 2013; Liu et al. 2014. [3]This study. [4]Gyoja and Satoh 2013. [5]Ledent et al. 2002; Simionato et al. 2007; Wang et al. 2009; Liu and Zhao 2010; Liu et al. 2013. [6]Simionato et al. 2007; Gyoja et al. 2012; Gyoja 2014. Additional data on *Hes/Hey/Helt* genes from Gazave et al. (2014).

to, with relatively rapid change of the DNA-binding domain until a new target sequence becomes established and generates effective purifying selection. In some instances however, tandem gene arrangements provide evidence of the evolutionary origin of these novel genes. With this in mind, it is interesting to note the expression of these lineage-specific genes shows evidence of both similarity (e.g., *LoblHLH*) and difference (e.g., *MgasHLH*) between the two molluscs examined. In the longer term, it may be feasible to further dissect this using RNAi and ChIP approaches.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

Y.B. would like to acknowledge Zhejiang Association for International Exchange of Personnel (ZAIEP) and Ningbo Science and Technology International Cooperation Research Projects (2016D10017) for partial financial support. F.X. would like to acknowledge China Scholarship Council (CSC) and National Natural Science Foundation of China (No. 31402285) for partial financial support. We also thank The High Performance Computing Center (HPCC) at the Institute of Oceanology, Chinese Academy of Sciences.

Literature Cited

Atchley WR, Fitch WM. 1997. A natural classification of the basic helix-loop-helix class of transcription factors. Proc Natl Acad Sci U S A. 94:5172–5176.

Ayer DE, Kretzner L, Eisenman RN. 1993. Mad—a heterodimeric partner for Max that antagonizes Myc transcriptional activity. Cell 72:211–222.

Ballow D, Meistrich ML, Matzuk M, Rajkovic A. 2006. Sohlh1 is essential for spermatogonial differentiation. Dev Biol. 294:161–167.

Ballow DJ, Xin Y, Choi Y, Pangas SA, Rajkovic A. 2006. Sohlh2 is a germ cell-specific bHLH transcription factor. Gene Expr Patterns. 6:1014–1018.

Blackwood EM, Eisenman RN. 1991. Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. Science 251:1211–1217.

Campuzano S. 2001. Emc, a negative HLH regulator with multiple functions in *Drosophila* development. Oncogene 20:8299–8307.

Carretero-Paulet L, et al. 2010. Genome-wide classification and evolutionary analysis of the bHLH family of transcription factors in *Arabidopsis*, poplar, rice, moss, and algae. Plant Physiol. 153:1398–1412.

Cowles MW, et al. 2013. Genome-wide analysis of the bHLH gene family in planarians identifies factors required for adult neurogenesis and neuronal regeneration. Development 140:4691–4702.

Dang C, Wang Y, Zhang D, Yao Q, Chen K. 2011. A genome-wide survey on basic helix-loop-helix transcription factors in giant panda. PLoS One 6:e26878.

Dang CW, Wang Y, Chen KP, et al. 2011. The basic helix-loop-helix transcription factor family in the pea aphid, *Acyrthosiphon pisum*. J Insect Sci. 11:84.

Fox RM, Vaishnavi A, Maruyama R, Andrew DJ. 2013. Organ-specific gene expression: the bHLH protein Sage provides tissue specificity to *Drosophila* FoxA. Development 140:2160–2171.

Gazave E, Guillou A, Balavoine G. 2014. History of a prolific family: the Hes/Hey-related genes of the annelid *Platynereis*. Evodevo 5:29.

Gyoja F. 2014. A genome-wide survey of bHLH transcription factors in the Placozoan *Trichoplax adhaerens* reveals the ancient repertoire of this gene family in metazoan. Gene 542:29–37.

Gyoja F, Kawashima T, Satoh N. 2012. A genomewide survey of bHLH transcription factors in the coral *Acropora digitifera* identifies three novel orthologous families, pearl, amber, and peridot. Dev Genes Evol. 222:63–76.

Gyoja F, Satoh N. 2013. Evolutionary aspects of variability in bHLH orthologous families: insights from the pearl oyster, *Pinctada fucata*. Zool Sci. 30:868–876.

Holland PW, Booth HA, Bruford EA. 2007. Classification and nomenclature of all human homeobox genes. BMC Biol. 5:47.

Jimenez-Delgado S, Crespo M, Permanyer J, Garcia-Fernandez J, Manzanares M. 2006. Evolutionary genomics of the recently duplicated amphioxus Hairy genes. Int J Biol Sci. 2:66–72.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.

Kenny NJ, Namigai EK, Marletaz F, Hui JH, Shimeld SM. 2015. Draft genome assemblies and predicted microRNA complements of the intertidal lophotrochozoans *Patella vulgata* (Mollusca, Patellogastropoda) and *Spirobranchus* (*Pomatoceros*) *lamarcki* (Annelida, Serpulida). Mar Genomics. 24P2:139–146.

Kocot KM, et al. 2011. Phylogenomics reveals deep molluscan relationships. Nature 477:452–456.

Ledent V, Paquet O, Vervoort M. 2002. Phylogenetic analysis of the human basic helix-loop-helix proteins. Genome Biol. 3:RESEARCH0030.

Ledent V, Vervoort M. 2001. The basic helix-loop-helix protein family: comparative genomics and phylogenetic analysis. Genome Res. 11:754–770.

Liu A, et al. 2012. A genome-wide identification and analysis of the basic helix-loop-helix transcription factors in the ponerine ant, *Harpegnathos saltator*. BMC Evol Biol. 12:165.

Liu A, et al. 2013. Classification and evolutionary analysis of the basic helix-loop-helix gene family in the green anole lizard, *Anolis carolinensis*. Mol Genet Genomics. 288:365–380.

Liu WY, Zhao CJ. 2010. Genome-wide identification and analysis of the chicken basic helix-loop-helix factors. Comp Funct Genomics. 210:1–12.

Liu XT, et al. 2014. A genome-wide identification and classification of basic helix-loop-helix genes in the jewel wasp, *Nasonia vitripennis* (Hymenoptera: Pteromalidae). Genome 57:525–536.

Luo YJ, et al. 2015. The *Lingula* genome provides insights into brachiopod evolution and the origin of phosphate biomineralization. Nat Commun. 6:8301.

Mazet F, Yu JK, Liberles DA, Holland LZ, Shimeld SM. 2003. Phylogenetic relationships of the Fox (Forkhead) gene family in the *Bilateria*. Gene 316:79–89.

Minguillon C, Jimenez-Delgado S, Panopoulou G, Garcia-Fernandez J. 2003. The amphioxus Hairy family: differential fate after duplication. Development 130:5903–5914.

Moore AW, Barbel S, Jan LY, Jan YN. 2000. A genomewide survey of basic helix-loop-helix factors in *Drosophila*. Proc Natl Acad Sci U S A. 97:10436–10441.

Pangas SA, et al. 2006. Oogenesis requires germ cell-specific transcriptional regulators Sohlh1 and Lhx8. Proc Natl Acad Sci U S A. 103:8090–8095.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 11:1650–1667.

Pires N, Dolan L. 2010a. Early evolution of bHLH proteins in plants. Plant Signal Behav. 5:911–912.

Pires N, Dolan L. 2010b. Origin and diversification of basic-helix-loop-helix proteins in plants. Mol Biol Evol. 27:862–874.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Satou Y, et al. 2003. A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. I. Genes for bHLH transcription factors. Dev Genes Evol. 213:213–221.

Sawada A, et al. 2000. Zebrafish Mesp family genes, mesp-a and mesp-b are segmentally expressed in the presomitic mesoderm, and Mesp-b confers the anterior identity to the developing somites. Development 127:1691–1702.

Sebe-Pedros A, de Mendoza A, Lang BF, Degnan BM, Ruiz-Trillo I. 2011. Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsaspora owczarzaki*. Mol Biol Evol. 28:1241–1254.

Shimeld SM, Boyle MJ, Brunet T, Luke GN, Seaver EC. 2010. Clustered Fox genes in lophotrochozoans and the evolution of the bilaterian Fox gene cluster. Dev Biol. 340:234–248.

Simakov O, et al. 2013. Insights into bilaterian evolution from three spiralian genomes. Nature 493:526–531.

Simionato E, et al. 2007. Origin and diversification of the basic helix-loop-helix gene family in metazoans: insights from comparative genomics. BMC Evol Biol. 7:33.

Skinner MK, Rawls A, Wilson-Rawls J, Roalson EH. 2010. Basic helix-loop-helix transcription factor gene family phylogenetics and nomenclature. Differentiation 80:1–8.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Struck TH, et al. 2011. Phylogenomic analyses unravel annelid evolution. Nature 471:95–98.

Suzuki H, et al. 2012. SOHLH1 and SOHLH2 coordinate spermatogonial differentiation. Dev Biol. 361:301–312.

Takeuchi T, et al. 2012. Draft genome of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology. DNA Res. 19:117–130.

Takeuchi T, et al. 2016. Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. Zool Lett. 2:3.

Wang XH, et al. 2015. Genome-wide identification and analysis of basic helix-loop-helix domains in dog, *Canis lupus familiaris*. Mol Genet Genomics. 290:633–648.

Wang Y, Chen K, Yao Q, Wang W, Zhu Z. 2008. The basic helix-loop-helix transcription factor family in the honey bee, *Apis mellifera*. J Insect Sci. 8:1–12.

Wang Y, Chen K, Yao Q, Zheng X, Yang Z. 2009. Phylogenetic analysis of zebrafish basic helix-loop-helix transcription factors. J Mol Evol. 68:629–640.

Weigert A, et al. 2014. Illuminating the base of the annelid tree using transcriptomics. Mol Biol Evol. 31:1391–1401.

Zhang DB, et al. 2013. Phylogenetic analyses of vector mosquito basic helix-loop-helix transcription factors. Insect Mol Biol. 22:608–621.

Zhang G, et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490:49–54.

Zheng X, Wang Y, Yao Q, Yang Z, Chen K. 2009. A genome-wide survey on basic helix-loop-helix transcription factors in rat and mouse. Mamm Genome. 20:236–246.

Zhou M, et al. 2012. Comparative and evolutionary analysis of the HES/HEY gene family reveal exon/intron loss and teleost specific duplication events. PLoS One 7:e40649.

Associate editor: Jay Storz