Broad Phylogenetic Occurrence of the Oxygen-Binding Hemerythrins in Bilaterians

Elisa M. Costa-Paiva[1,2,*], Carlos G. Schrago[1], and Kenneth M. Halanych[2]

[1]Laboratório de Biologia Evolutiva Teórica e Aplicada, Departamento de Genética, Universidade Federal do Rio de Janeiro, Brazil

[2]Molette Biology Laboratory for Environmental and Climate Change Studies, Department of Biological Sciences, Auburn University

*Corresponding author: E-mail: elisapolychaeta@gmail.com.

Accepted: September 5, 2017

Data deposition: This project has been deposited at GenBank under accession numbers KY929214 to KY929271.

Abstract

Animal tissues need to be properly oxygenated for carrying out catabolic respiration and, as such, natural selection has presumably favored special molecules that can reversibly bind and transport oxygen. Hemoglobins, hemocyanins, and hemerythrins (Hrs) fulfill this role, with Hrs being the least studied. Knowledge of oxygen-binding proteins is crucial for understanding animal physiology. Hr genes are present in the three domains of life, Archaea, Bacteria, and Eukaryota; however, within Animalia, Hrs has been reported only in marine species in six phyla (Annelida, Brachiopoda, Priapulida, Bryozoa, Cnidaria, and Arthropoda). Given this observed Hr distribution, whether all metazoan Hrs share a common origin is circumspect. We investigated Hr diversity and evolution in metazoans, by employing in silico approaches to survey for Hrs from of 120 metazoan transcriptomes and genomes. We found 58 candidate Hr genes actively transcribed in 36 species distributed in 11 animal phyla, with new records in Echinodermata, Hemichordata, Mollusca, Nemertea, Phoronida, and Platyhelminthes. Moreover, we found that “Hrs” reported from Cnidaria and Arthropoda were not consistent with that of other metazoan Hrs. Contrary to previous suggestions that Hr genes were absent in deuterostomes, we find Hr genes present in deuterostomes and were likely present in early bilaterians, but not in nonbilaterian animal lineages. As expected, the Hr gene tree did not mirror metazoan phylogeny, suggesting that Hrs evolutionary history was complex and besides the oxygen carrying capacity, the drivers of Hr evolution may also consist of secondary functional specializations of the proteins, like immunological functions.

Key words: metazoa, transcriptome, evolutionay history, oxygen-binding protein.

Introduction

Oxygen-binding proteins are ancient molecules that probably evolved from enzymes that protected the organism against the toxic oxygen (Terwilliger 1998). Considering that metabolism in metazoans requires oxidation of organic molecules, natural selection has likely favored proteins that can reversibly bind and transport oxygen to body tissues (Schmidt-Rhaesa 2007). In metazoans, four families of oxygen-binding proteins are known, usually divided into two main groups: proteins that are use iron to bind oxygen, including hemoglobins and hemerythrins (Hr), and two nonhomologous families of hemocyanins that are use copper (Terwilliger et al. 1976; Burmester 2002). Although these molecules can reversibly bind oxygen, their binding affinities and evolutionary origins differ and the diversity of blood pigments in animals is clearly underestimated (Martín-Durán et al. 2013; Koch et al. 2016; Costa-Paiva et al. 2017).

The evolution of both hemoglobins and hemocyanins has been extensively studied (Burmester 2002, 2015; Lecomte et al. 2005; Vinogradov et al. 2006; Decker et al. 2007), however knowledge of Hr genes is still limited (Vanin et al. 2006). Hemerythrin is an ancient protein family present in all three domains of life (Fukami-Kobayashi et al. 2007; Bailly et al. 2008; Alvarez-Carreño et al. 2016). However, in animals, Hr records are restricted to marine invertebrates within Annelida (which include sipunculids; Struck et al. 2007; Weigert et al. 2014), Brachiopoda, Priapulida, Bryozoa, and a single species of both Cnidaria (*Nematostella vectensis*) and Arthropoda (*Calanus finmarchicus*) (Klippenstein 1980; Vanin et al. 2006; Bailly et al. 2008; Martín-Durán et al. 2013;

Costa-Paiva et al. 2017). Bailly et al. (2008) suggested that the Hr gene was lost in the ancestor of deuterostomes and conserved only in a few protostomes, leading to questions of Hr homology across metazoans (Bailly et al. 2008; Martín-Durán et al. 2013). A complex evolutionary history of lateral gene transfer, duplications, and gene loses appear to have play an important role in Hr evolution in animals (Alvarez-Carreño et al. 2016).

Hr sequences were originally characterized from sipunculids (Sanders-Loehr and Loehr 1979) for which three Hr sequences were recorded, two from coelomic hemerythrocytes or circulating Hr (cHrs) from *Phascolopsis gouldii* and *Themiste dyscritum* and one myohemerythrin (myoHr) from retractor muscle of *Themiste zostericola*. The difference between cHrs and myoHr sequences is a five-residue insertion in the myoHr sequence between residues 90 and 91 flanked by the C and D helices (Sanders-Loehr and Loehr 1979; Kurtz 1992). Previous workers reported that there were four distinct subtypes of Hrs: polymeric cHrs and monomeric myoHrs, ovohemerythrins (ovoHr), and neurohemerythrins (nHr) (Baert et al. 1992; Coutte et al. 2001; Vergote et al. 2004). However more recent work (Vanin et al. 2006; Costa-Paiva et al. 2017) has confirmed that there are only two types of Hrs (myoHr and cHr). Recent studies show that occurrence, diversity, and expression of Hrs in animals is much greater than currently recorded (Martín-Durán et al. 2013; Costa-Paiva et al. 2017). We employed a stringent approach to scan for Hrs in a diverse array of metazoan transcriptomes and genomes. We examine Hr evolutionary history in the light of animal phylogeny (Whelan et al. 2015; Halanych 2016; Kocot et al. 2017).

Materials and Methods

Sample Collection

Information on species employed herein is provided in table 1. Transcriptomes of these species were collected as part of the WormNet II project to resolve annelid phylogeny with a variety of techniques, including intertidal sampling, dredge and box cores. All samples collected were preserved in RNALater or frozen at −80 °C.

Data Collection and Sequence Assembly

RNA extraction, cDNA preparation and high-throughput sequencing generally followed Kocot et al. (2011) and Whelan et al. (2015). Total RNA was extracted from either whole animals (for small specimens) or the body wall and coelomic region (for larger specimens). RNAs were purified after extraction using TRIzol (Invitrogen) or the RNeasy kit (Qiagen) with on-column DNase digestion, respectively. In order to reverse transcribe single stranded RNA template, we used the SMART cDNA Library Construction Kit (Clonetech) and double stranded cDNA synthesis was completed with The Advantage 2 PCR system (Clontech). Libraries were barcoded and sequenced with Illumina technology by The Genomic Services Lab at the Hudson Alpha Institute (Huntsville, Alabama, USA). Because sequencing was performed from 2012 to 2015, Paired End (PE) runs were of 100 or 125 bp lengths, utilizing either v3 or v4 chemistry on Illumina HiSeq 2000 or 2500 platforms (San Diego, California). To facilitate sequence assembly, paired-end transcriptome data were digitally normalized to an average k-mer coverage of 30 using normalize-by-median.py (Brown et al. 2012) and assembled using Trinity r2013-02-25 with default settings (Grabherr et al. 2011).

Data Mining and Gene Identification

Methods employed were similar to those in Costa-Paiva et al. (2017). Two complementary approaches were utilized to mine transcriptomic data from 100 metazoan species and two choanoflagellate species for putative Hr genes in silico (table 1). Additionally, we surveyed genomes, a transcriptome, and ESTs from Genbank for 20 species (table 1) including chordates, cnidarians, ctenophores, acoels, placozoan, and arthropods in order to search for Hr similarity.

The first approach employed BLASTX (Altschul et al. 1990) with *e*-value cutoff of 10^{-6} in order to compare each assembled transcriptome contig ("queries") to a protein database composed of 19 Hrs sequences from the National Center for Biotechnology (NCBI) database (supplementary file 2, Supplementary Material online) of at least 110 amino acid residues and previously identified as Hrs ($n = 7$), myoHrs ($n = 10$), or "nHr" ($n = 2$). The BLASTX approach assured that any transcriptome contig with a significant "hit" to an Hr would be further evaluated in the pipeline. Initial contigs recovered from BLAST searches were then utilized in BLASTX searches against the NCBI protein database (minimum *e*-value of 10^{-10}) and only top hits longer than 300 nucleotides were retained and considered putative Hr genes.

A second approach processed the transcriptomic data from the same species (table 1) through the Trinotate annotation pipeline (http://trinotate.github.io/) (Grabherr et al. 2011), which utilizes a BLAST-based approach to provide, among others, GO annotation (The Gene Ontology Consortium 2004). Transcripts annotated as Hrs, using the 10^{-6} *e*-value cutoff obtained by using BLASTX, were also considered putative Hr-like gene orthologs.

Contigs putatively identified as Hr genes by both approaches were subsequently translated into amino acids using TransDecoder with default settings (Haas et al. 2013). Since TransDecoder can produce multiple open reading frames (ORFs), all translations were additionally subject to a Pfam domain evaluation using the EMBL-EBI database with an *e*-value cutoff of 10^{-5}. Translations returning an Hr Pfam domain and that were longer than 100 amino acids residues were retained for subsequent analyses. Moreover, we

Table 1
List of All Taxa Analyzed, Including Total Number of Contigs after Assembly, and Number of Putative Hr Genes (for Undelined Taxa)[a]

Taxon	Total Contigs Number	Hr Genes Number	Accession Number
CHOANOFLAGELATA		—	—
Acanthoeca spectabilis W.Ellis, 1930	198,922		
Salpingoeca pyxidium Kent, 1881	202,399		
METAZOA			
Ctenophora			
Beroe abyssicola Mortensen, 1927	83,798	—	—
Coeloplana astericola Mortensen, 1927	222,614	—	—
Dryodora glandiformis (Mertens, 1833)	101,598	—	—
Euplokamis dunlapae Mills, 1987	321,550	—	—
Mnemiopsis leidyi A. Agassiz, 1865	385,798	—	
Pleurobrachia bachei A. Agassiz, 1860	38,856	—	
**Pleurobrachia bachei* A. Agassiz, 1860	–	—	—
Vallicula multiformis Rankin, 1956	339,814	—	—
Porifera			
**Amphimedon queenslandica* Hooper and van Soest, 2006	—	—	—
Hyalonema populiferum Schulze, 1899	58,839	—	—
Kirkpatrickia variolosa (Kirkpatrick, 1907)	100,231		
Latrunculia apicalis Ridley and Dendy, 1886	76,210		
Rossella fibulata Schulze and Kirkpatrick, 1910	40,103	—	—
Sympagella nux Schmidt, 1870	85,237	—	—
Placozoa			
**Trichoplax adhaerens* Schulze, 1883	—	—	—
Cnidaria			
**Acropora digitifera* (Dana, 1846)	—	—	—
Gersemia antarctica (Kukenthal, 1902)	20,023	—	—
**Hydra vulgaris* Pallas, 1766	—	—	—
* *Nematostella vectensis* Stephenson, 1935	—	—	—
Hit with NW_001833871.1, Pfam domain not confirmed			
Hit with NW_001834356.1, Pfam domain not confirmed			
**Orbicella faveolata* (Ellis and Solander, 1786)	—	—	—
Periphylla periphylla (Peron and Lesueur, 1810)	212,658	—	—
**Pseudodiploria strigosa* (Dana, 1846)	—	—	—
Acoela	(reads)	—	—
Childia submaculatum SRX1534054	(29,856,889)	—	—
Convolutriloba macropyga SRX1343815	(210,917,52)	—	—
Diopisthoporus gymnopharyngeus SRX1534055	(33,284,316)	—	—
Diopisthoporus longitubus SRX1534056	(44,491,819)	—	—
Eumecynostomum macrobursalium SRX1534057	(47,195,086)	—	—
Isodiametra pulchra SRX1343817	(268,267,139)	—	—
Echinodermata			
Apostichopus californicus (Stimpson, 1857)	134,640	1	KY929257
Astrotoma agassizii Lyman, 1875	156,062	—	—
Labidiaster annulatus Sladen, 1889	108,871	—	—
Labidiaster sp.	168,720	1	KY929242
Leptosynapta clarki Heding, 1928	242,126	1	KY929245
Hemichordata			
Balanoglossus aurantiaca Girard, 1853	143,815	3	KY929217-9
Cephalodiscus gracilis Harmer, 1905	57,139	4	KY929226-9
Cephalodiscus hodgsoni Ridewood, 1907	200,052	1	KY929230
Cephalodiscus nigrescens Lankester, 1905	11,565	1	KY929231
Harrimaniidae gen sp. (from Iceland)	230,054	—	—
Harrimaniidae gen sp. (from Norway)	274,434	—	—
Ptychodera bahamensis Spengel, 1893	115,310	—	—

(continued)

Table 1 Continued

Taxon	Total Contigs Number	Hr Genes Number	Accession Number
Rhabdopleura sp.	4,790	—	—
Saccoglossus mereschkowskii Wagner, 1885	145,937	—	—
Schizocardium brasiliense Spengel, 1893	101,493	—	—
Stereobalanus canadensis Spengel, 1893	12,741	6	KY929266-71
Torquaratoridae gen. sp.	102,971		
Staurozoa gen. sp.	45,023	—	—
Chordata			
**Oikopleura dioica* Fol, 1872	—	—	—
**Homo sapiens* Linnaeus, 1758	—	—	—
Annelida			
Arenicola loveni Kinberg, 1866	27,028	—	—
Arhynchite pugettensis Fisher, 1949	20,724	1	KY929214
Aulodrilus japonicus Yamaguchi, 1953	109,361	2	KY929215-6
Capilloventer sp.	221,627	5	KY929220-4
Chloeia pinnata Moore, 1911	130,037	1	KY929232
Dichogaster saliens (Beddard 1893)	98,665	1	KY929233
Diopatra cuprea (Bosc, 1802)	138,779	1	KY929234
Dodecaceria pulchra Day, 1955	229,501	1	KY929235
Eunice norvegica (Linnaeus, 1767)	122,784	1	KY929236
Hermodice carunculata (Pallas, 1766)	110,813	—	—
Lumbrineris crassicephala Hartman, 1965	196,426	2	KY929246-7
Marphysa sanguinea (Montagu, 1813)	110,924	2	KY929248-9
Ophiodromus pugettensis (Johnson, 1901)	92,341	—	—
Ophryotrocha globopalpata Blake and Hilbig, 1990	129,450	1	KY929253
Palola sp.	211,279	1	KY929254
Sphaerodorum papillifer Moore, 1909	52,411		
Brachiopoda			
Glottidia pyramidata (Stimpson, 1860)	131,562	1	KY929237
Hemithiris psittacea (Gmelin, 1791)	103,581	2	KY929239-40
Laqueus californicus (Koch, 1848)	133,086	1	KY929243
Macandrevia cranium (O. F. Müller, 1776)	9,695	—	—
Phoronida			
Phoronis psammophila Cori, 1889	193,702	1	KY929259
Phoronopsis harmeri Pixell, 1912	283,821		
Novocrania anomala (O. F. Müller, 1776)	117,369	1	KY929251
Mollusca			
Alexandromenia crassa Odhner, 1920	111,729	—	—
Amphimeniidae gen. sp.	130,196	—	—
Aplacophora gen. sp.	109,736	—	—
Cavibelonia sp.	144,105	—	—
Entonomenia tricarinata (Salvini-Plawen, 1978)	147,128	—	—
Epimenia babai Salvini-Plawen, 1997	71,819	—	—
Falcidens caudatus (Heath, 1918)	132,816	—	—
Graptacme eborea (Conrad, 1846)	144,601	1	KY929238
Helluoherpia aegiri Handl and Buchinger, 1996	95,935	—	—
Hypomenia sp.	93,699	1	KY929241
Kruppomenia borealis Odhner, 1920	142,815	—	—
Leptochiton rugatus (Carpenter in Pilsbry, 1892)	115,512	1	KY929244
Macellomenia sp.	107,525	—	—
Meiomenia swedmarki Morse, 1979	118,867	—	—
Micromenia fodiens (Schwabl, 1955)	230,891	1	KY929250
Neomenia carinata Tullberg, 1875	172,727	—	—
Nuculana pernula (O. F. Müller, 1779)	34,274	1	KY929252
Phyllomenia sp.	170,739	—	—
Prochaetoderma californicum Schwabl, 1963	293,209	—	—

(continued)

Table 1 Continued

Taxon	Total Contigs Number	Hr Genes Number	Accession Number
Proneomeniidae gen. sp.	99,165	2	KY929262-3
Scutopus ventrolineatus Salvini-Plawen, 1968	221,900		
Simrothiella margaritacea (Koren and Danielssen, 1877)	99,722	—	—
Spathoderma clenchi Scheltma, 1985	111,974	—	—
Nemertea			
Malacobdella grossa (Müller, 1779)	79,313		
Paranemertes peregrina Coe, 1901	99,203	2	KY929255-6
Parborlasia corrugatus (McIntosh, 1876)	911,662		
Tubulanus polymorphus Renier, 1804	109,120		
Bryozoa			
Pectinatella magnifica (Leidy, 1851)	191,465	1	KY929258
Cycliophora			
Symbion americanus Obst, Funch and Kristensen, 2006	135,725		
Entoprocta			
Barentsia gracilis M. Sars, 1835	146,310		
Loxosoma pectinaricola Franzen, 1962	144,339		
Platyhelminthes			
Acipensericola petersoni Bullard, Snyder, Jensen and Overstreet, 2008	152,140		
Cardicola currani Bullard and Overstreet, 2004	86,962	1	KY929225
Cardicola palmeri Bullard and Overstreet, 2004	52,837		
Elaphrobates euzeti Bullard and Overstreet, 2003	118,013		
Elopicola sp.	64,384		
Hapalorhynchus sp.	42,863		
Myliobaticola richardheardi Bullard and Jensen, 2008	15,147		
Myliobaticola sp.	73,883		
Psettarium anthicum Bullard and Overstreet, 2006	39,616		
Sanguinicola sp.	145,041		
Selachohemecus olsoni Short, 1954	135,169	2	KY929264-5
Orthonectida			
Orthonectida gen. sp.	231,032		
Arthropoda			
Calanus finmarchicus (Gunnerus, 1770)	–		
Hit with ES3871551, Pfam domain not confirmed			
Colossendeis megalonyx **Hoek, 1881**	114,203		
**Limulus polyphemus* (Linnaeus, 1758)	–		
Priapulida			
Priapulus sp.	50,034	2	KY929260-1

[a]GenBank accession numbers are also provided here and detailed in supplementary file 1, Supplementary Material online. Genomes are marked with asterisks, all others are transcriptomes. Transcriptomes of acoels presented total reads numbers, instead of contigs.

manually evaluated the presence of residues involved in iron binding, which are: histidine residues (His) in positions 26, 56, 75, 79, and 108; glutamic acid residue (Glu) in position 60; and aspartic acid residue (Asp) in position 113, numbered by reference sequence *T. zostericola*. Presence of these signature residues indicates putative respiratory function for Hrs. Transcripts passing the criteria described above were considered Hr genes (table 1).

For additional genomes, transcriptome, and ESTs from Genbank, we employed BLASTP, tBLASTn, or BLASTN (Altschul et al. 1990) depending on the type of data available for each species (table 1), at an *e*-value cutoff of 10^{-6}. We compared the database with the query composed of 19 Hrs sequences from NCBI database (supplementary file 2, Supplementary Material online) as above. Sequences with a significant "hit" to an Hr were additionally subject to a Pfam domain evaluation using the EMBL-EBI database with an *e*-value cutoff of 10^{-5}.

Sequence Alignment

The protein data set consisted of 77 sequences, including 19 Hr sequences previous used as "queries" (supplementary file 2, Supplementary Material online), and a remaining 58 sequences from translated transcripts (supplementary file 1, Supplementary Material online). All sequences were initially aligned with MAFFT using the "accurate E-INS-i" algorithm

(Katoh and Standley 2013), followed by visual inspection and manual curation in order to remove spuriously aligned sequences based on similarity to the protein alignment as a whole. Subsequently, ends of aligned sequences were manually trimmed in Geneious 9.1.3 (Kearse et al. 2012) to exclude 5′ residues leading to the putative start codon and 3′ residues following the first two amino acids subsequent to the end of the D α-helix. The resulting alignment was used for all subsequent analyses (supplementary file 3, Supplementary Material online).

Phylogenetic Analysis

ProtTest3.4 was applied to carry out statistical selection of best-fit models of protein evolution for the data set using the Akaike and Bayesian Information Criteria (AIC and BIC, respectively) methods (Darriba et al. 2011). Bayesian phylogenetic inference was performed with MrBayes 3.2.1 (Ronquist and Huelsenbeck 2003) with two independent runs with four Metropolis-coupled chains were run for 10^7 generations, sampling the posterior distribution every 500 generations. In order to confirm if chains achieved stationary and determine an appropriate burn-in, we evaluated trace plots of all MrBayes parameter output in Tracer v1.6 (Rambaut et al. 2014). The first 25% of samples were discarded as burn-in and a majority rule consensus tree generated using MrBayes. Bayesian posterior probabilities were used for assessing statistical support of each bipartition.

Evolutionary Rate Analyses

The protein alignment (supplementary file 3, Supplementary Material online) was also used in DIVERGE (Gu et al. 2013) to examine site-specific shifted evolutionary rates and assesses whether there has been a significant change in evolutionary rate after duplication or speciation events by calculating the coefficient of divergence (θD) and determining if the null hypothesis of no functional divergence between Hrs with the five residue indel between C and D α-helices could be statistically rejected. We employed a cutoff of 0.8 for detection of site-specific shifted evolutionary rates (supplementary file 4, Supplementary Material online).

Results

Our in silico analyses (fig. 1) recovered 238 unique nucleotide sequences of hemerythrin-like genes from 108 transcriptomes and 11 genome gene models encompassing 20 metazoan phyla and two choanoflagellate species (table 1). Following translation, Pfam domain evaluation (presence of Hr domain), and removal of sequences with <100 amino acid residues, 58 putative novel Hr genes were retained from all taxa examined in this study, representing 36 metazoan species distributed in 11 different phyla (table 1, supplementary file 1, Supplementary Material online). Hrs had been reported previously in four out of these 11 phyla, namely, Annelida, Brachiopoda, Priapulida, and Bryozoa (Bailly et al. 2008; Martín-Durán et al. 2013; Costa-Paiva et al. 2017). However, we report Hrs in Echinodermata, Hemichordata, Mollusca, Nemertea, Phoronida, and Platyhelminthes. Tertiary structure of Hrs was inferred using I-TASSER (Yang et al. 2015) and putative respiratory function and high similarity among their tertiary structure was confirmed for representative Hr genes in each newly recorded phylum (fig. 2). We did not find any Hr genes in either choanoflagellate species, Acoela, Arthropoda, Chordata, Cnidaria, Ctenophora, Cycliophora, Entoprocta, Placozoa, Porifera, and Orthonectida (table 1).

Alignment of translated transcripts included 122 residue positions. All sequences started with a methionine residue and contained signature residues involved in iron binding, indicating putative respiratory function (Thompson et al. 2012). For the 58 putatively novel Hr sequences, 34 were unique and 24 were identical for at least two species at the amino acid level. New sequences were combined with 19 publically available Hrs to produce a final data set of 77 Hr sequences (supplementary files 2 and 3, Supplementary Material online; Figshare file DOI: 10.6084/m9.figshare.4715092).

Bayesian inference analysis (fig. 3) recovered several strongly supported clades, as well as less resolved regions (which are often observed in gene genealogies; DeSalle 2015). We found Hr orthologs in 12 additional annelid species (table 1), augmenting previous counts (Bailly et al. 2008; Costa-Paiva et al. 2017). All 58 novel Hr sequences included the five-residue insertion before the D α-helix, consistent with myoHrs (Costa-Pavia et al. 2017). Those sequences were distributed throughout the gene tree in clades with representatives from other phyla (fig. 3, orange and gray clades). As expected (Costa-Paiva et al. 2017), leech "nHr" was strongly supported ($P = 1$) as sister lineage to a myoHr sequence from the same species. Similarly, the priapulid "nHr" was a strongly supported as sister lineage to a myoHr sequence from the same priapulid species (fig. 3, yellow clades, $P \geq 0.99$).

All new 58 sequences possessed the five-residue insertion before the D α-helix characteristic of myoHrs (Bailly et al. 2008; Costa-Paiva et al. 2017) (fig. 3, except blue clade). We used DIVERGE software (Gu et al. 2013) to look for differences in evolutionary rates between annelid cHrs and other sequences, as well as relative rates of change in different positions were calculated and for helix regions. A and B α-helices each had five sites with an elevated evolutionary rate, whereas C and D α-helices had eight and seven sites, respectively, with elevated rates indicating that later helices are likely evolving faster than the others.

The topology of the Hr gene tree, as expected, did not mirror recent phylogenies of Metazoa based on phylogenomic data sets (Whelan et al. 2015; Halanych 2016; Kocot et al. 2017). We found two clades with exclusively protostome composition. One clade (fig. 3, gray clade, $P < 0.8$)

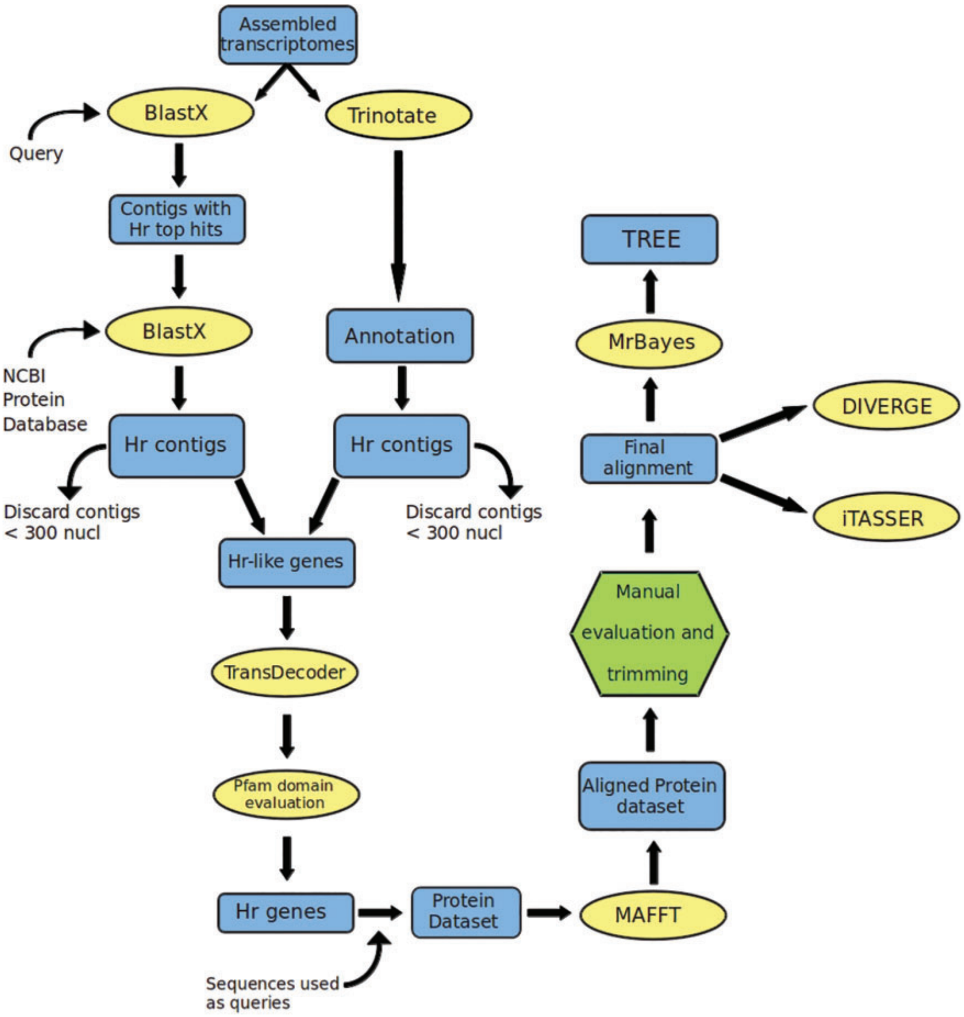

Fig. 1.—Flow chart of bioinformatics pipeline. Rounded blue rectangles represent input/output files, yellow ovals represent software or scripts, and the green hexagon represents a step which involving manual evaluation. Nineteen metazoan Hrs sequences previous used as query sequences from Genbank (supplementary file 2, Supplementary Material online) were also included in the data set.

included representatives of Annelida, Mollusca, Platyhelminthes, Brachiopoda, Nemertea, and Priapulida, and the second clade, the cHr clade (fig. 3, blue clade) included only annelids. Moreover, we found Hrs in eight species of Echinodermata and Hemichordata and observed strongly supported clades (fig. 3, green clades, $P = 1$) exclusively comprised of deuterostome Hr sequences. However, we also found supported clades (fig. 3, orange clades, $P > 0.9$) where deuterostomes Hr sequences were clustered with Hr sequences from protostomes as annelids, brachiopods, and mollusk. Because most of our data are from transcriptomes, we must be cautious about comments on the absence of genes as they may be in the genome but were not expressed in the sampled tissue at time of collection. Nonetheless, we did not find any Hr genes in transcriptomes of choanoflagellates, arthropods, cnidarians, ctenophores, cycliophorans, entoprocts, orthonectids, sponges, and acoels (table 1). However, we also screened the available genomes of arthropods, cnidarians, a ctenophore, chordates, a placozoan, and a sponge for Hr genes and obtained negative results, including previous Hr records from *Nematostella vectensis* and *Calanus finmarchicus* (Martín-Durán et al. 2013). Sequences from *N. vectensis* (XP_001634535.1 and XP_001622541.1) did not match any Pfam domain and when a BLASTp search was performed, neither sequence was similar to Hr sequences. Sequence from *C. finmarchicus* (ES387155), also available in Genbank and assigned as a myoHr, did not match any Pfam domain and overall similarity with other Hr sequences.

Discussion

Distribution of Hr genes spans the breadth of Bilateria, but the absence in nonbilaterians contradicts earlier reports. Previously, the distribution of Hrs was thought to be limited

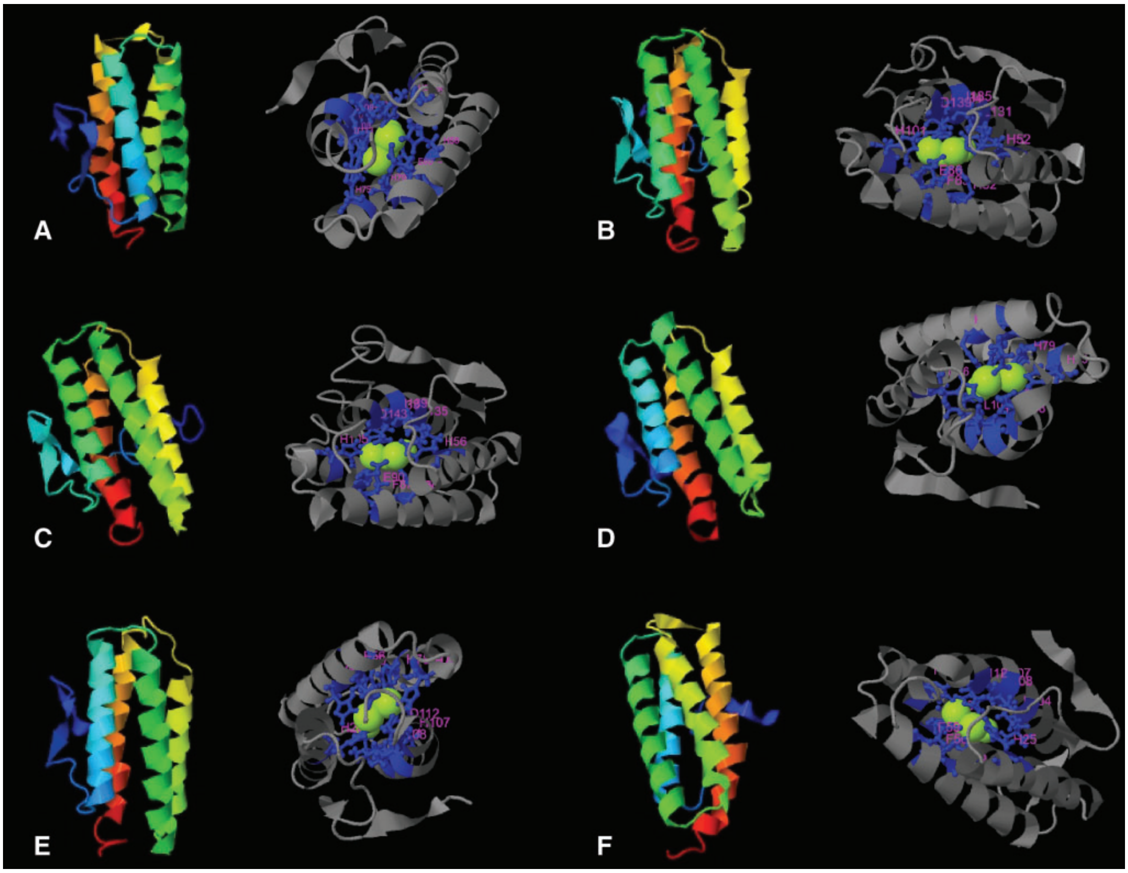

Fig. 2.—Tertiary structure of Hrs from a representative of each newly recorded phylum was inferred using ITASSER (Yang et al. 2015) and confirmed that all sequences have a putative respiratory function and also showed the high similarity among their tertiary structure. Each figure on the right indicated the position of amino acids related to iron binding. (*A*) Echinodermata—*Leptosynapta clarki*; (*B*) Hemichordata—*Balanoglossus aurantiaca*; (*C*) Mollusca—*Nuculana pernula*; (*D*) Nemertea—*Paranemertes peregrina*; (*E*) Phoronida—*Phoronis psammophila*; (*F*) Platyhelminthes—*Selachomecus olsoni*.

to protostomes, in a single ecdysozan phyla (Priapulida) and three lophotrochozoan phyla (Annelida, Brachiopoda, and Bryozoa) (Vanin et al. 2006; Bailly et al. 2008). Here, we discovered actively transcribed Hrs in 36 species from 11 phyla, including the first ever record in deuterostomes. A previous study (Bailly et al. 2008) suggested that deuterostomes lost Hr genes and there was limited conservation of Hrs in protostome lineages after the protostome–deuterostome split. Nonetheless, since we found expressed Hrs in three species of Echinodermata and five species of Hemichordata, we suggest that, at least the ancestor of Deuterostomes and the ancestor of Ambulacraria had at least one copy of an Hr gene (fig. 4). Moreover, we do not observe Hrs in animal lineages that branched off from other metazoans prior to bilaterians. Note, the prior report of Hrs in a cnidarian (Martín-Durán et al. 2013) that is not homologous to the bilaterian Hrs. Both sequences from *N. vectensis* previously attributed to Hrs did not match the Hr Pfam domain structure and, also, did not present a significant similarity with Hr sequences when a BLAST strategy was employed. The homology of the previously reported crustacean *C. finmarchicus* Hr (Martín-Durán et al. 2013) is also not supported based on sequence data, although our analysis of ecdysozoans were limited.

Martín-Durán et al. (2013) suggested that the metazoan ancestor already had a respiratory functional Hr gene followed by frequent gene losses in various lineages. Our findings suggested a scenario with an Hr-bearing nephrozoan ancestor and not the last common bilaterian ancestor, since we did not find any evidence of the presence of Hrs in non-bilaterian metazoans or in the earliest branching bilaterian lineage, Acoela (fig. 4). Bayesian reconstruction of the Hr gene tree was incongruent to current knowledge of metazoan phylogeny (Whelan et al. 2015; Halanych 2016; Kocot et al. 2017), which is not surprising as this gene family is presumably under heavy selection to supply different demands to carry oxygen in various animal lineages. A similar evolutionary pattern is observed across the three domains of life (for Alvarez-Carreño et al. 2016), or even within annelids (Costa-Paiva et al. 2017). The fact that the gene tree includes subclades with disparate taxa (e.g., echinoderms and annelids, orange clades on fig. 3) suggested that Hrs have a

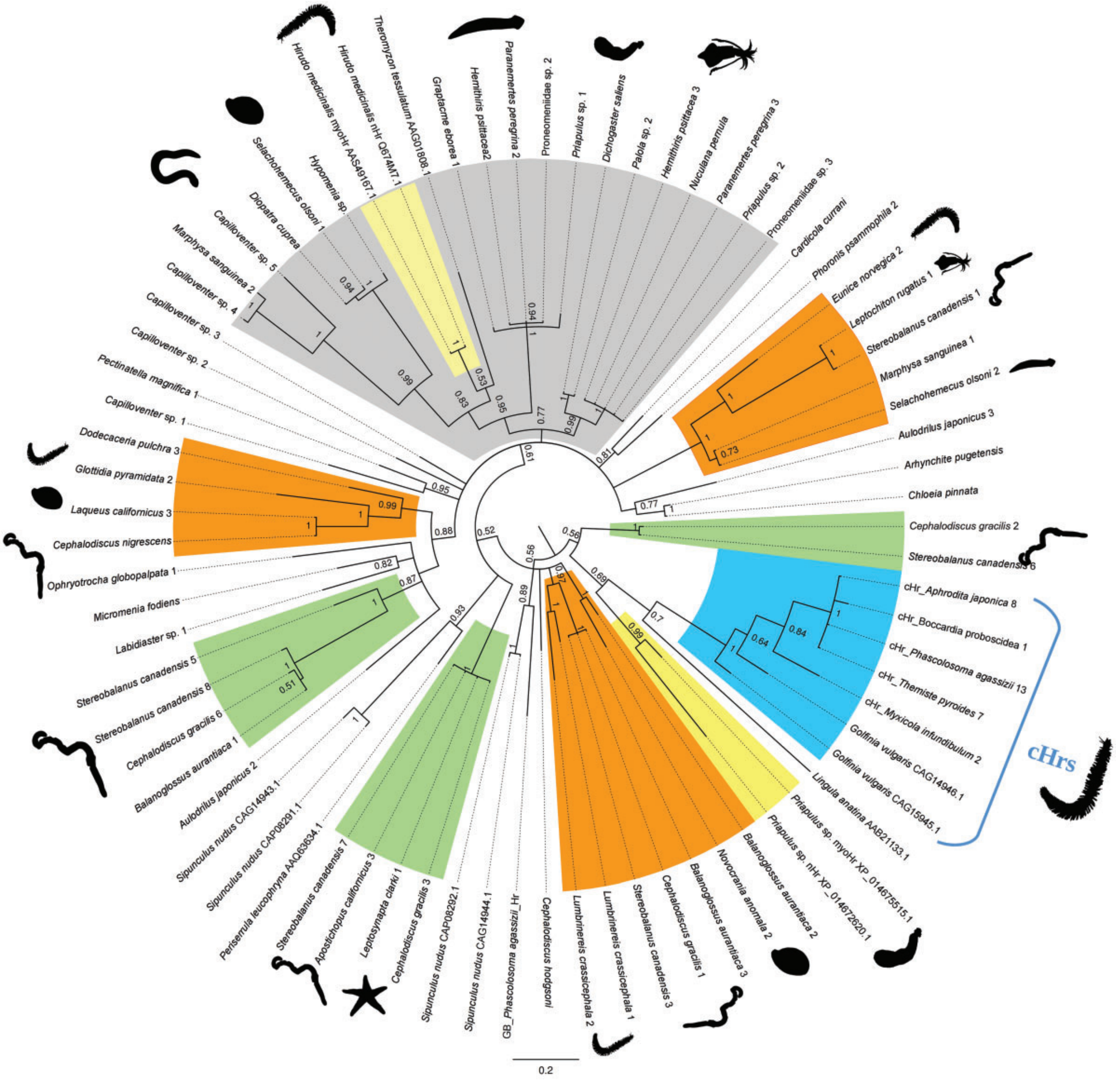

FIG. 3.—Bayesian tree using MrBayes 3.2.1 (Ronquist and Huelsenbeck 2003) midpointed rooted. The blue clade represents cHrs with the five residue deletion between C and D α-helices; gray clades represent clades with protostome myoHr sequences; orange clades represent clades with protostomes and deuterostomes myoHr sequences; green clades represent clades with only deuterostomes myoHr sequences and yellow clades represents sequences of myoHr and "nHr" from a leech and a priapulid. The number after the name of each sequence indicates the GenBank accession numbers for each Hr gene and it is indicated in supplementary file 1, Supplementary Material online.

complex history, supporting notions of gene loss and duplication, and possibly lateral gene transfer (Martín-Durán et al. 2013; Alvarez-Carreño et al. 2016).

Traditional classification of specific Hr subtypes in cHrs, myoHrs, ovoHrs, and nHrs (Baert et al. 1992; Coutte et al. 2001; Vergote et al. 2004) was not validated by the gene genealogy. Although many of our transcriptomes used whole organisms (including reproductive and nerve tissues), our results failed to recover Hr proteins that corresponded to ovoHrs or nHrs, corroborating Costa-Paiva et al.'s (2017) previous findings. Classification of myoHrs and cHrs had traditionally relied on differences regarding the monomeric or polimeric form, respectively, and the presence or absence of a five-amino-acid indel between the C and D α-helices

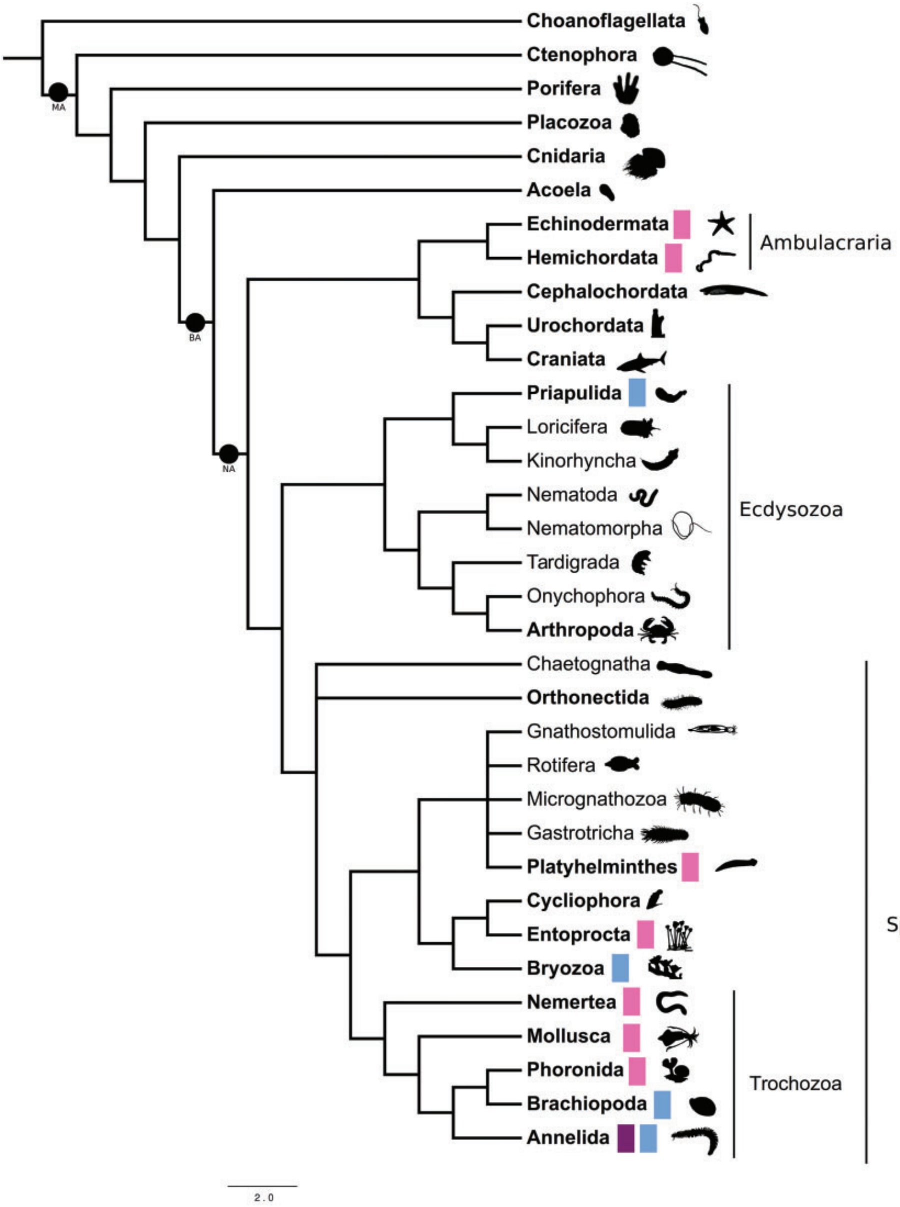

Fig. 4.—Hypothesized relationships among metazoan phyla derived from recent phylogenomic studies (Whelan et al. 2015; Halanych 2016; Cannon et al. 2016; Kocot et al. 2017). Pink rectangles represent new Hr records, blue rectangles represent previous records confirmed by our results, and purple rectangle represents exclusively annelid cHrs. MA is metazoan ancestor, BA is bilaterian ancestor, and NA is nephrozoan ancestor.

(Sanders-Loehr and Loehr 1979; Kurtz 1992; Vanin et al. 2006; Costa-Paiva et al. 2017). The cHr subtype of Hrs, which lacks the five residues between C and D α-helices, is present exclusively in Annelida (fig. 3, blue clade) and represent a novelty in bilaterian Hr evolution. Although rare, there are few records of Hrs present in the vascular systems of priapulids (Weber et al. 1979; Weber and Fange 1980) and brachiopods (Richardson et al. 1987), however they possess five-residues between C and D α-helices. Our findings were able in recognize only two primary types of Hrs, based on molecular differences, myoHrs and exclusive annelid cHrs (Costa-Paiva et al. 2017). Furthermore, our results from DIVERGE, concerning differences in evolutionary rates between annelid cHrs and other sequences, as well as relative rates of change in different positions showed that Hr molecules presented differences in an evolutionary rate, with the C and D α-helices (C and D) evolving faster than the A and B α-helices.

Although the distribution of Hrs in animals is likely tied to the need to deliver oxygen to tissues, as corroborated by our results from modeling the tertiary structure of observed Hr genes (fig. 2), many of metazoan lineages we examined also possess hemoglobins to carry oxygen (Mangum 1992; Coutte et al. 2001). We suggest that although the observed pattern could be explained by the need to carry oxygen, secondary functional specializations could also be important for driving diversification (Coates and Decker 2016). Additional studies of the gene structure of Hr proteins and physiological aspects of organisms are the next important steps toward a better understanding of the evolutionary patterns involved in this family of oxygen carrying proteins.

Supplementary Material

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

Competing Interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Fernando Avila Queiroz for valuable help with the figures. A Fellowship for to E.M.C-P. was provided by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil). Use of SkyNet computational resources at Auburn University is acknowledged. This work was funded by National Science Foundation grant DEB-1036537 to K. M. Halanych and Scott R. Santos and OCE-1155188 to K. M. Halanych. This is Molette Biology Laboratory contribution 69 and Auburn University Marine Biology Program contribution 165.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3): 403–410.

Alvarez-Carreño C, Becerra A, Lazcano A. 2016. Molecular evolution of the oxygen-binding hemerythrin domain. PLoS One 11 (6): e0157904.

Baert JL, Britel M, Sautière P, Malécha J. 1992. Ovohemerythrin, a major 14-kDa yolk protein distinct from vitellogenin in leech. Eur J Biochem. 209(2): 563–569.

Bailly X, Vanin S, Chabasse C, Mizuguchi K, Vinogradov SN. 2008. A phylogenomic profile of hemerythrins, the nonheme diiron binding respiratory proteins. BMC Evol Biol. 8: 244.

Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. 2012. A reference-free algorithm for computational normalization of shotgun sequencing data. arXiv:1203.4802

Burmester T. 2002. Origin and evolution of arthropod hemocyanins and related proteins. J Comp Physiol B. 172 (2): 95–107.

Burmester T. 2015. Evolution of respiratory proteins across the Pancrustacea. Integr Comp Biol. 55 (5): 765–770.

Cannon JT, et al. 2016. Xenacoelomorpha is the sister group to Nephrozoa. Nature 530(7588): 89–93.

Coates CJ, Decker H. 2016. Immunological properties of oxygen-transport proteins: hemoglobin, hemocyanin and hemerythrin. Cell Mol Life Sci. 2016: 1–25.

Costa-Paiva EM, et al. 2017. Discovery and evolution of novel hemerythrin genes in annelid worms. BMC Evol Biol. 17(1): 85–96.

Coutte L, Slomianny MC, Malecha J, Baert JL. 2001. Cloning and expression analysis of a cDNA that encodes a leech hemerythrin. Biochim Biophys Acta 1518 (3): 282–286.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27(8): 1164–1165.

Decker H, et al. 2007. Recent progress in hemocyanin research. Integr Comp Biol. 47 (4): 631–644.

DeSalle R. 2015. Can single protein and protein family phylogenies be resolved better?. J Phylogenetics Evol Biol. 3: e116.

Fukami-Kobayashi K, Minezaki Y, Tateno Y, Nishikawa K. 2007. A tree of life based on protein domain organizations. Mol Biol Evol. 24(5): 1181–1189.

Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32(1): D258–D261.

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7): 644–652.

Gu X, et al. 2013. An update of DIVERGE software for functional divergence analysis of protein family. Mol Biol Evol. 30(7): 1713–1719.

Haas BJ, et al. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 8(8): 1494–1512.

Halanych KM. 2016. How our view of animal phylogeny was reshaped by molecular approaches: lessons learned. Org Divers Evol. 16(2): 319–328.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30 (4): 772–780.

Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28 (12): 1647–1649.

Klippenstein GL. 1980. Structural aspects of hemerythrin and myohemerythrin. Am Zool. 20(1): 39–51.

Koch J, et al. 2016. Unusual diversity of myoglobins genes in the lungfish. Mol Biol Evo. 33(12): 3033–3041.

Kocot KM, et al. 2011. Phylogenomics reveals deep molluscan relationships. Nature 477(7365): 452–456.

Kocot KM, et al. 2017. Phylogenomics of Lophotrochozoa with consideration of systematic error. Syst Biol. 66(2): 256–282.

Kurtz DM Jr. 1992. Molecular structure/function of the hemerythrins. In: Mangum CP, editor. Advances in comparative and environmental physiology, Vol 13: Blood and tissue oxygen carriers. Berlin Heidelberg: Springer-Verlag.

Lecomte JT, Vuletich DA, Lesk AM. 2005. Structural divergence and distant relationships in proteins: evolution of the globins. Curr Opin Struc Biol. 15 (3): 290–301.

Mangum CP. 1992. Physiological function of the hemerythrins. In: Mangum CP, editor. Advances in comparative and environmental physiology, Vol. 13: Blood and tissue oxygen carriers. Berlin Heidelberg: Springer-Verlag.

Martín-Durán JM, De Mendoza A, Sebé-Pedrós A, Ruiz-Trillo I, Hejnol A. 2013. A broad genomic survey reveals multiple origins and frequent losses in the evolution of respiratory hemerythrins and hemocyanins. Genome Biol Evol. 5(7): 1435–1442.

Rambaut A, Suchard MA, Xie D, Drummond AJ. 2014. Tracer v1.6 [Accessed 15 May 2017]. Available from: http://beast.bio.ed.ac.uk/Tracer

Richardson DE, Emad M, Reem RC, Solomon EI. 1987. Allosteric interactions in sipunculid and brachiopod hemerythrins. Biochemistry 26(4): 1003–1013.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19 (12): 1572–1574.

Sanders-Loehr J, Loehr TM. 1979. Hemerythrin: a review of structural and spectroscopic properties. Adv Inorg Biochem. 1: 235–252.

Schmidt-Rhaesa A. 2007. The evolution of organs systems. Oxford: Oxford University Press. p. 385.

Struck TH, et al. 2007. Annelid phylogeny and the status of Sipuncula and Echiura. BMC Evol Biol. 7(1): 57–68.

Terwilliger NB. 1998. Functional adaptations of oxygen-transport proteins. J Exp Biol. 201(Pt 8): 1085–1098.

Terwilliger RC, Terwilliger NB, Schabtach E. 1976. Comparison of chlorocruorin and annelid hemoglobin quaternary structures. Comp Biochem Phys A. 55 (1): 51–55.

Thompson JW, et al. 2012. Structural and molecular characterization of iron-sensing hemerythrin-like domain within F-box and leucine-rich repeat protein 5 (FBXL5). J Biol Chem. 287(10): 7357–7365.

Vanin S, et al. 2006. Molecular evolution and phylogeny of sipunculan hemerythrins. J Mol Evol. 62(1): 32–41.

Vergote D, et al. 2004. Up-regulation of neurohemerythrin expression in the central nervous system of the medicinal leech, *Hirudo medicinalis*, following septic injury. J Biol Chem. 279(42): 43828–43837.

Vinogradov SN, et al. 2006. A phylogenomic profile of globins. BMC Evol Biol. 6: (1): 31.

Weber RE, Fange R. 1980. Oxygen equilibrium of *Priapulus* hemerythrin. Experientia 36(4): 427–428.

Weber RE, Fange R, Rasmussen KK. 1979. Respiratory significance of priapulid hemerythrin. Mar Biol Lett. 1: 87–97.

Weigert A, et al. 2014. Illuminating the base of the annelid tree using transcriptomics. Mol Biol Evol. 31 (6): 1391–1401.

Whelan NV, Kocot KM, Moroz LL, Halanych KM. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Acad Sci U S A. 112(18): 5773–5778.

Yang J, et al. 2015. The I-TASSER suite: protein structure and function prediction. Nat Methods 12(1): 7–8.

Associate editor: Liliana Milani