Evolutionary History of the Globin Gene Family in Annelids

Flávia A. Belato[1], Christopher J. Coates[2], Kenneth M. Halanych[3], Roy E. Weber[4], Elisa M. Costa-Paiva[1*]

1. Department of Zoology, Institute of Biosciences, University of Sao Paulo, Sao Paulo, Brazil

2. Department of Biosciences, College of Science, Swansea University, Swansea SA2 8PP, Wales, United Kingdom

3. Department of Biological Sciences, Molette Biology Laboratory for Environmental and Climate Change Studies, Auburn University, Auburn, AL, 36849, USA

4. Zoophysiology, Department of Biology, Aarhus University, DK-8000 Aarhus, Denmark

***Corresponding author e-mail:** elisam.costapaiva@gmail.com

ABSTRACT

Animals depend on the sequential oxidation of organic molecules to survive; thus, oxygen-carrying/transporting proteins play a fundamental role in aerobic metabolism. Globins are the most common and widespread group of respiratory proteins. They can be divided into three types: circulating intracellular, non-circulating intracellular, and extracellular, all of which have been reported in annelids. The diversity of oxygen transport proteins has been underestimated across metazoans. We probed 250 annelid transcriptomes in search of globin diversity in order to elucidate the evolutionary history of this gene family within this phylum. We report two new globin types in annelids, namely androglobins and cytoglobins. Although cytoglobins and myoglobins from vertebrates and from invertebrates are referred to by the same name, our data show they are not genuine orthologs. Our phylogenetic analyses show that extracellular globins from annelids are more closely related to extracellular globins from other metazoans than to the intracellular globins of annelids. Broadly, our findings indicate that multiple gene duplication and neo-functionalization events shaped the evolutionary history of the globin family.

Keywords: androglobin, cytoglobin, extracellular globin, gene tree, transcriptomics, respiratory proteins

List of abbreviations: Adgb, androglobin; Cygb, cytoglobin; Hb, hemoglobin; HBL-Hb, hexagonal bilayer hemoglobin; His, histidine; Mb, myoglobin; MW, molecular weight; nHb, nerve hemoglobin; O_2, oxygen; Phe, phenylalanine; RBC, red blood cell.

Significance Statement: Annelid worms have the greatest diversity of oxygen-carrying proteins, also known as blood pigments, among all animals. However, the real diversity of these proteins has been still underestimated. To access the diversity of globins present in annelids, and to elucidate their evolutionary relationships, we have searched for globin genes in 250 annelid genetic material. We found two new globins in this phylum: androglobins and cytoglobins. Our results indicate that cytoglobins and myoglobins from vertebrates and invertebrates have different evolutionary origins, and that androglobins and extracellular globins originated early in animal's evolution. We show that multiple gene duplication events shaped the complex evolutionary history of the globin famil

INTRODUCTION

Aerobic metabolism relies on the sustained transfer of oxygen (O_2) from environmental sources to the respiring tissues of animals, which is carried out by O_2 transport proteins (also known as respiratory pigments) (Terwilliger, 1998; Burmester and Hankeln, 2004; Coates and Decker, 2017). These globular proteins represent the most widespread respiratory pigments and occur almost ubiquitously amongst organisms, including bacteria, fungi, plants, protists, and animals (Hardison, 1996, 1998; Weber and Vinogradov, 2001; Vazquez-Limon *et al*., 2012 Vinogradov *et al*., 2007, 2013 a, b). Concerning metazoans, intra- and extra-cellular hemoglobins (Hb and HBL-Hb, respectively) and myoglobin (Mb) have been known for over a century (Lankester, 1872). More recent comparative genomic studies revealed the existence of several new globin types in vertebrates, such as cytoglobin (Cygb), androglobin (Adgb), and neuroglobins (Ngb) (Burmester *et al*., 2000; Kawada *et al*., 2001; Burmester *et al*., 2002; Trent and Hargrove, 2002; Burmester and Hankeln, 2004; Hoogewijs *et al*., 2012). Following this trend, studies demonstrated that the known diversity of oxygen-carrying proteins in animals is underestimated, and this also seems to be true in annelids (Bailly *et al*., 2008; Martín-Durán *et al*., 2013; Costa-Paiva *et al.,* 2017, 2018; Belato *et al*., 2019). Despite the conserved tertiary structures of all globins, the recently discovered proteins differ markedly in their amino acid sequences and carry out several alternative cellular functions besides O_2 transport, e.g., oxygen-sensing, enzymic activity, signal transduction, lipid and nitric oxide metabolism, and detoxification of reactive oxygen species, suggesting that the presence of more than one type of oxygen-binding protein in animals is related to those other cellular functions (Weber and Vinogradov, 2001; Burmester and Hankeln, 2014).

The circulating annelid Hbs occur within nucleated red blood cells (RBCs), in contrast to the anucleate RBCs that harbor the intensively studied mammalian Hbs (Storz, 2018). Invertebrate Hbs are found in at least six phyla: Phoronida, Nemertea, Mollusca, Annelida, Arthropoda, and Echinodermata (Terwilliger and Ryan, 2001; Weber and Vinogradov, 2001), where they may occur in closed vascular systems, or dissolved in the coelomic fluid and hemolymph (functional equivalents to blood). Annelid Hbs exhibit the classical "Mb-fold", a structure that comprises five to eight α-helices, named A through H, forming a three-on-three or two-on-two helical sandwich that surrounds the oxygen-binding heme group (Bolognesi *et al*., 1997; Terwilliger, 1998; Weber and Vinogradov, 2001; Vinogradrov and Moens, 2008; Gell, 2018). All invertebrate Hbs contain the characteristic, invariant globin amino acids residues:

His at E7 (seventh amino acid in helix E), His at F8, and Phe at the inter-helical region CD1 (Bolognesi *et al*., 1997; Weber and Vinogradov, 2001). In contrast to the tetrameric vertebrate Hbs, annelid RBC Hbs may be monomeric, dimeric, tetrameric, polymeric or a combination of these states (Weber, 1980; Mangum, 1985). The hexagonal bilayer hemoglobins (HBL-Hbs), also called chlorocruorins and erythrocruorins, are giant (MW ~$3.5x10^6$), extracellular circulating protein complexes that occur freely dissolved in blood equivalents (Weber, 1971; Weber and Vinogradov, 2001). For decades, the HBL-Hbs were considered to be present only in a few annelid species (Vinogradov, 1985; Weber and Vinogradov, 2001). Recently, we demonstrated a much wider phylogenetic distribution of these giant extracellular proteins in invertebrates, including Mollusca, Platyhelminthes, and some deuterostome lineages (Belato *et al*., 2019). The mega-molecular HBL-Hbs are comprised of two types of polypeptides: globin chains, with single oxygen-binding sites that satisfy the criterion of a globin-like fold, and linker chains that lack heme groups and are required for the multimeric (hierarchal) assembly of the vast quaternary structures (Vinogradov, 1985; Lamy *et al*., 1996; Weber and Vinogradov, 2001; Royer *et al*., 2006).

Among the intracellular non-circulating globins, myoglobins (Mbs) are monomers consisting of ~140 amino acids that reside in the cytoplasm of muscle cells of metazoan taxa and function as intracellular O_2 store and in transcellular (facilitated) diffusion of O_2 (Wittenberg, 1970; Suzuki and Imai, 1998). Non-circulating globins also comprise the nerve hemoglobins (nHbs), that occur sporadically in glial cells surrounding the nerve cord and neurons of various invertebrate taxa, including Annelida, Arthropoda, Echiura, Mollusca, Nematoda, and Nemertea (Wittenberg, 1992, Weber and Vinogradov, 2001; Geuens *et al*., 2004; Burmester and Hankeln, 2008). nHbs consist of ~150 amino acid residues and may exhibit the Mb-like structure and exist as homodimers, as seen in the annelid *Aphrodita aculeate* (Wittenberg, 1992; Dewilde *et al*., 1996; Weber and Vinogradov, 2001; Geuens *et al*., 2004). Although all nHbs contain the diagnostic residues (Phe CDl, HisE7, and HisF8), phylogenetic analyses indicate divergent evolutionary origins (Wittenberg, 1992; Dewilde *et al*., 1996; Weber and Vinogradov, 2001; Burmester and Hankeln, 2008). The principal function of invertebrate nHbs is considered to be O_2 storage and supply during hypoxia, sustaining the aerobic metabolism of the nervous system (Kraus *et al*., 1988; Wittenberg, 1992, Weber and Vinogradov, 2001; Geuens *et al*., 2004).

Cytoglobins (Cygbs) are non-circulating globins that are co-located alongside Mbs in the cytoplasm of cells of several different vertebrate tissues. However, Cygbs have longer polypeptide chains, with around 170 amino acid residues, since additional residues flank the *N*- and *C*- terminals and they

thus lack sequence insertions that interrupt the globin fold (Burmester *et al*., 2002). These proteins do not contain signal peptides and are found in the cytoplasm and nucleus of many different cell types (Burmester *et al*., 2002). Vertebrate Cygb shows structural and phylogenetic affinities to vertebrate Mb (Burmester *et al*., 2002; DeSanctis *et al*., 2004); however, these relationships are not resolved for invertebrate Cygbs. Androglobins (Adgbs), the most recently discovered non-circulating globins, are cytoplasmic, large chimeric proteins that exhibit a modular domain structure. They comprise a *N*-terminal calpain-like domain, a rearranged globin domain, where the eight α-helices (A-H) are organized such that helices C-H precede helices A-B, and an IQ calmodulin-binding motif (Hoogewijs *et al*., 2012; Bracke *et al*., 2018). Despite the different globin domain sequence, Adgbs satisfy the globin-fold criterion (Hoogewijs *et al*., 2012; Bracke *et al*., 2018). These chimeric proteins have been recorded in a wide range of metazoan taxa, such as Mollusca, Cnidaria, and Chordata (Hoogewijs *et al*., 2012; Bracke *et al*., 2018).

Annelids thus exhibit the greatest diversity of oxygen-binding proteins among metazoans (Mangum, 1998; Costa-Paiva *et al*., 2017), with three types of globins described so far: 1) non-circulating cytoplasmic globins such as Mbs and nHbs; 2) circulating intracellular red blood cell (RBC) Hbs and 3) extracellular HBL-Hbs dissolved in body fluids (Wittenberg, 1970; Vinogradov *et al*., 1993; Lamy *et al*., 1996; Suzuki and Imai, 1998; Weber and Vinogradov, 2001; Bailly *et al*., 2007). Related to the scarcity of available sequences for these three globin types in annelids, only three families are known to express all simultaneously: Opheliidae, Terebellidae, and Alvinellidae (Weber, 1978; Hourdez *et al*., 2000). To the best of our knowledge, only one study (Bailly *et al*., 2007) has focused exclusively on the evolutionary history of the globin superfamily in annelids. Using 28 annelid globin sequences, Bailly *et al*. (2007) demonstrated that extracellular globin lineages appear to have a separate evolutionary history compared to intracellular circulating and noncirculating annelid globins. Nevertheless, the real diversity of globin genes within annelids is yet to be investigated, and phylogenetic relationships between different globin types in these animals remain uncertain. In order to access the real diversity of globins present in annelids, and to elucidate the evolutionary relationships of these proteins within the phylum, our work represents a systematic analysis of 250 annelid transcriptomes to survey for globin genes. We report the existence of four non-circulating intracellular globin types in annelids: Mbs, nHbs, Cygbs and Adgbs. Our molecular evolutionary analyses indicate a complex evolutionary history for members of the globin superfamily within Annelida, which includes several gene duplication and neo-functionalization events.

METHODS

Transcriptomes of 250 annelid species were used in this work and information about each species is indexed in Supplementary file 1. The transcriptomes were collected as part of the WormNet II project that primarily seeks to resolve annelid phylogeny. Specimens were obtained by several collection techniques, including intertidal sampling, dredging, and box cores. Afterwards, all samples were preserved either in RNALater or frozen at −80 °C. Protocols from Kocot *et al*. (2011) and Whelan *et al*. (2015) were used for RNA extraction, cDNA preparation and high-throughput sequencing. Succinctly, total RNA was extracted using TRIzol (Invitrogen) either from whole small animals, or from the body walls and coelomic regions, in bigger specimens. After extraction, RNAs were purified using the RNeasy kit (Qiagen) with on-column DNase digestion. To reverse transcribe single-stranded RNA template, the SMART cDNA Library Construction Kit (Clontech) was used and double-stranded cDNA synthesis was performed with the Advantage 2 PCR system (Clontech). The Genomic Services Lab at the Hudson Alpha Institute (Huntsville, AL) was responsible for barcoding and sequencing libraries with Illumina technology. Considering that transcriptomic sequencing was conducted from 2012 to 2015, paired-end runs were of 100 or 125 bp in length, utilizing either v3 or v4 chemistry on Illumina HiSeq 2000 or 2500 platforms (San Diego, CA). Finally, in order to facilitate sequence assembly, paired-end transcriptome data were digitally normalized to an average k-mer coverage of 30 using the script normalize-by-median.py (Brown *et al*., 2012) and was assembled using Trinity r2013-02-25 with default settings (Grabherr *et al*., 2011).

Bioinformatic methods employed to search *in silico* for genes of the globin family were similar to those in Belato *et al*. (2019). Transcriptome data was processed through the Trinotate annotation pipeline (http://trinotate.github.io/) (Grabherr *et al*., 2011). The Trinotate pipeline uses a BLAST-based method against two databases, namely EggNOG 4.5.1 (Huerta-Cepas *et al*., 2016) and KEGG (Kanehisa *et al*., 2016), to provide the Gene Ontology (GO) annotation. The GO is a standardized functional classification system for genes that describes the properties of genes and their products using a dynamic-updated controlled vocabulary (Gene Ontology Consortium, 2004). The complete list of software employed by the Trinotate pipeline to provide the annotation of genes is: HMMER 3.2.1, for protein domain identification (Finn *et al*., 2011); tmHMM 2.0, for prediction of transmembrane helices of proteins (Krogh *et al*., 2001); RNAmmer 1.2, for prediction of ribosomal RNA (Lagesen *et al*., 2007);

SignalP 4.1, to predict signal peptide cleavage sites (Petersen *et al*., 2011); GOseq, for prediction of the gene ontology (Young *et al*., 2010); and EggNOG 4.5.1, for searching orthologous groups of genes (Huerta-Cepas *et al*., 2016). As we used transcriptomic data, we can only make inferences about the presence of gene signatures and refrain from drawing conclusions about their absence, since genes may be present in the genome without being expressed in the sampled tissue at time of collection.

Retrieved sequences were manually verified by inspecting each functional annotation made by Trinotate in order to select sequences annotated as hemoglobins (Hbs), myoglobins (Mbs), cytoglobins (Cygbs), androglobins (Adgbs), and nerve hemoglobins (nHbs). In addition, 10 annelid extracellular hemoglobin (HBL-Hbs) sequences from Belato *et al.* (2019) were selected to be used in our analyses because these sequences were obtained employing the same bioinformatic pipeline including the same annotation and validation steps. RNA sequences identified as one of the genes described above were then translated into amino acids using TransDecoder software with default settings (https://transdecoder.github.io/). All translated protein sequences were evaluated using the Pfam domain check (Finn *et al*., 2016) employing the EMBL-EBI protein database with an e-value cutoff of 10^{-5}. This step was necessary because the TransDecoder translation may produce multiple open reading frames (ORFs). Translations returning with a confirmed Pfam domain and that were longer than 130 amino acids residues were retained for further analysis. In order to refine the results, we added a confirmatory step, where we performed a reciprocal BLASTp (Altschul *et al*., 1990) of all sequences annotated as the target genes against the non-redundant protein database (nr) from the National Center for Biotechnology (NCBI). Only sequences that presented a significant top 'hit' with a minimum e-value of 10^{-10} to one of the target genes Hbs, Mbs, Cygbs, Adgbs, nHbs and HBL-Hbs were retained. We have labeled proteins according to their putative functional role, considering local similarity between sequences from our dataset and NCBI database. Sequence similarity and similarity in domain structure are generally indicative of similarity in function (Marcotte *et al*., 1999; Ashburner *et al*., 2000; Gabaldón and Huynen 2004).

Adgbs were manually rearranged in order to remove the IQ motif and concatenate the eight α-helices (A-H) of the globin domain that are inverted in these globins. After all validation steps, the remaining 379 sequences were aligned with MAFFT using the accurate algorithm E-INS-i (Katoh and Standley, 2013), and gap-rich regions in the alignment were removed with trimAl 1.2 (Capella-Gutierrez *et al*., 2009) using a gap threshold of 0.75. The alignment was manually curated using the software

Geneious 11.1.2 (Kearse *et al*., 2012) in order to remove spuriously aligned sequences based on similarity to the protein alignment as a whole. To eliminate dataset redundancy sequences that presented 100% of similarity to each other were also excluded from the alignment. The resulting amino acid alignment of 238 sequences was subsequently used for phylogenetic analyses.

ModelFinder, an ultrafast and automatic model selector implemented in IQ-TREE software (Kalyaanamoorthy *et al*., 2017) was applied to carry out statistical selection of the best-fit model of protein evolution for the dataset using the Akaike and Bayesian Information Criteria (AIC and BIC, respectively) (Darriba *et al*., 2011). Two phylogenetic inference methods were employed: (a) a maximum likelihood inference performed with the IQ-TREE software (Nguyen *et al*., 2015) with branch support obtained with the ultrafast bootstrap approximation (UFBoot) with 1,000 replicates (Minh *et al*., 2013) and (b) a Bayesian inference using MrBayes 3.2.7 (Ronquist and Huelsenbeck, 2003) with two independent runs, each one containing four Metropolis-coupled chains that were run for 10^7 generations and sampled every 500th generation to approximate posterior distributions. To confirm whether chains achieved stationary and to determine an appropriate burn-in, we evaluated trace plots of all MrBayes parameter outputs in Tracer v1.6 (Rambaut *et al*., 2014). The first 25% of samples were discarded as burn-in and a majority rule consensus tree was generated using MrBayes. Bayesian posterior probabilities were used for assessing statistical support of each bipartition. The resultant trees were summarized with FigTree 1.4.3 (Rambaut, 2009) and rooted by midpoint rooting (Farris, 1972; Hess and Russo, 2007). The Phyre2 web portal (Kelley *et al*., 2015) was used to predict the putative tertiary structures of the different globins and models were visualized and inspected using UCSF Chimera (Pettersen *et al*., 2004).

In order to better understand the evolutionary relationship between extracellular globins of annelids and those from other metazoans, another maximum likelihood analysis was performed using the IQ-TREE software (Nguyen *et al*., 2015) with branch support obtained under the ultrafast bootstrap approximation (UFBoot) (Minh *et al*., 2013). This analysis expanded the original 238 sequences alignment with another 15 extracellular globin sequences from five metazoan species obtained from Belato *et al.* (2019): *Astrotoma agassizii* (Echinodermata; MH995909 and MH996362), *Cephalodiscus gracilis* (Hemichordata; MH995925–26), *Hemithiris psittacea* (Brachiopoda; MH996374–75 and MH996036–38), *Phoronis psammophila* (Phoronida; MH996210–11), and *Priapulus* sp. (Priapulida; MH996240–42 and MH996405).

To access the evolutionary relationships between annelid globins and other metazoan globins, we selected a representative panel of 54 globins from deuterostomes (including vertebrates) and other protostomes from NCBI to be used as references of metazoan globins (Supplementary file 2). Together with a subset of 43 annelid globins (Table 1), these 97 sequences of annelid globins + metazoan globins were aligned with MAFFT using the accurate algorithm E-INS-i (Katoh and Standley, 2013), and gap-rich regions in the alignment were removed with trimAl 1.2 (Capella-Gutierrez *et al*., 2009) using a gap threshold of 0.50 (Supplementary file 3). Afterwards, a maximum likelihood analysis was performed using the IQ-TREE software (Nguyen *et al*., 2015) with branch support obtained under the ultrafast bootstrap approximation (UFBoot) (Minh *et al*., 2013).

RESULTS

The initial Trinotate analysis recovered 5,267 nucleotide sequences annotated as Hbs, Mbs, Cygbs, Adgbs, nHbs or HBL-Hbs. After all processing steps, including translation from nucleotides to amino acids, selection by minimum size, reciprocal BLASTp, and Pfam domain evaluation, our *in-silico* analyses recovered 238 unique amino acid sequences. These sequences consisted of 130 sequences of Hbs, 19 sequences of Mbs, 27 sequences of Cygbs, four sequences of Adgbs, 39 sequences of nHbs and 19 sequences of HBL-Hbs genes (Table 1). These genes are actively transcribed in 121 annelid species belonging to 64 different families as detailed in Table 1. Accession numbers of each one of the sequences obtained in this work were deposited in GenBank and are listed in Table 1 and detailed in Supplementary file 4.

The number of expressed Hbs genes in a given species ranged from one in 48 different species to eight in *Terebellides stroemii* (Trichobranchidae). For Mbs, the number of expressed genes in a given species ranged from one in 10 different species to four in *Praxillella pacifica* (Maldanidae). For Cygbs the corresponding numbers were one in 25 different species and two in *Aglaophamus verrilli* (Nephtyidae). One Adgb gene was found in four different species, and nHbs genes ranged from one expressed copy in 24 different species to three in *Alciopa* sp. (Alciopidae). Besides the 10 HBL-Hbs sequences selected from Belato *et al.*'s (2019) previous work that were used as reference (GenBank accession numbers: MH995867–68, MH995870–71, MH995874, MH995876, MH995879–80, MH996356 and MH995881), we found more HBL-Hb genes, ranging from one in two different species to

three in *Heterodrilus* sp. (Naididae). Besides the two well-known intracellular non-circulating globin types found in annelids, Mbs and nHbs, we reveal the presence of cytoglobins (Cygbs) and androglobins (Adgbs) as two new members of this globin type. Additionally, we found three more families that express all globin types (extracellular circulating, intracellular circulating and intracellular noncirculating) simultaneously: 1) Aeolosomatidae (*Aeolosoma* sp.); 2) Arenicolidae (*Abarenicola pacifica*); and 3) Nephtyidae (*Aglaophamus verrilli*).

Following trimming and alignment of translated transcripts the final alignment had the maximum sequence length of 142 residue positions (Supplementary file 5). All sequences in the alignment contained the essential residues of the globin domain (Phe at CD1, His at E7, His at F8), which is an indicator of respiratory function for each one of these proteins (Lecomte *et al*., 2005; Fig. 1). The best-fixed rate model for phylogenetic analyses of the annelid globins dataset and the annelid globins + metazoan extracellular globins dataset was the WAG model. For phylogenetic analyses of the annelid globins + metazoan globins dataset the best-fixed rate model was LG.

Bayesian and maximum likelihood inferences recovered the same topology with several strongly supported clades in the annelid globins tree (Fig. 2), although the topology did not mirror the recent Annelida phylogeny (e.g., Struck *et al*., 2015; Weigert and Bleidorn, 2016). Adgb genes clustered into a highly supported monophyletic group by bootstrap values and posterior probabilities (100%; PP=1; orange clade; Fig. 2), as well as the HBL-Hb genes that also clustered into one monophyletic group with strong statistical support (100%; PP=1; green clade; Fig. 2). When HBL-Hbs sequences from other metazoans were added to the alignment, all extracellular globins from both annelids and other metazoans clustered together in one monophyletic group with high support values (100%; Supplementary file 6).

Mbs clustered in two distinct well-supported clades (100%; PP>0.97; purple clades; Fig. 2), as well as Cygbs which also grouped into two separate highly supported clades (100%; PP>0.99; pink clades; Fig. 2). Concerning nHbs, a majority of sequences (26 sequences from a total of 39) clustered into one large and highly supported clade (100%; PP=1; blue clade; Fig. 2) and four other small clades with very few sequences in each one (blue clades; Fig. 2). The majority of Hb sequences also clustered into one large clade (67%; PP>0.58; yellow clade; Fig. 2), however several smaller clades with few sequences were also formed (yellow clades; Fig. 2). Within the Hb clades some low nodal support values were found. However, such results are common in phylogenetic analyzes within the same protein family (DeSalle, 2015). The tertiary structures the different globin genes inferred using the Phyre2 web portal

resulted in proteins with high similarity among their tertiary structure and putative respiratory function (Supplementary file 7). Model prediction confidence and coverage ranged from 97 – 100%.

In the phylogenetic reconstruction recovered with the annelid globins + metazoan globins dataset Adgbs from both vertebrates and invertebrates clustered in one clade (75%; red clade; Fig. 3), and HBL-Hbs also clustered into one monophyletic group (75%; green clade; Fig. 3). Mb clustered in two separate clades with strong bootstrap support values, one with vertebrate Mbs (85%; purple clade; Fig. 3) and another with invertebrate Mbs (100%; purple clade; Fig. 3). Cygbs formed three different groups with high support values, one with vertebrate Cygbs (85%; pink clade; Fig. 3), and two others with invertebrate Cygbs (86 and 87%; pink clades; Fig. 3). nHbs and vertebrate neuroglobins (Ngb) clustered into two distinct clades (71 and 75%; blue clades; Fig. 3), as well as vertebrate Hbs A and B, which also clustered in two monophyletic groups (95%; gray clades; Fig. 3). Invertebrate Hbs clustered into five different clades (yellow clades; Fig. 3). Some low nodal support values were also found in this gene tree.

DISCUSSION

We recovered two more types of non-circulating globins within annelids: androglobins (Adgbs) and cytoglobins (Cygbs), in addition to the already known nerve Hbs (nHbs) and myoglobins (Mbs), confirming that Annelida has the greatest diversity of oxygen-binding proteins (Mangum, 1998; Weber and Vinogradov, 2001; Costa-Paiva *et al*., 2017, 2018). Three annelid families, Opheliidae, Terebellidae, and Alvinellidae, were previously known to simultaneously express the three globin types – extracellular circulating, intracellular circulating and intracellular noncirculating (Weber, 1978; Hourdez *et al*., 2000). Our data reveal three additional families that express these three globin types: Aeolosomatidae, Arenicolidae, and Nephtyidae. In conjunction with results of Bailly *et al*. (2007), our results demonstrate that co-occurrence of different globin types is much more common than previously documented and that it probably already existed in the annelid ancestor. Our findings corroborate previous studies that suggest that vertebrate Cygbs and Mbs lineages are distinct from invertebrate Cygbs and Mbs lineages (Suzuki and Imai, 1998; Hoffmann *et al*., 2011, 2012; Blank and Burmester, 2012; Storz *et al*., 2013; Pillai *et al*., 2020). Furthermore, our phylogenetic analyses show that Cygbs and Mbs from vertebrates are not true orthologs of invertebrate Cygbs and Mbs (Fig. 3). These proteins received the same name because of their presumed functional role and not their evolutionary relationships.

The phylogenetic hypothesis for invertebrate globin sequences constructed by Goodman *et al*. (1988) divided annelid globins into two monophyletic groups: intracellular and extracellular. Similarly, the phylogenetic analysis carried out by Bailly *et al*. (2007) on the annelid globin genes demonstrated a well-supported division between intracellular and extracellular globins. Interestingly, we did not find one clade of extracellular globins and another of intracellular globins (Fig. 2). In our phylogenetic reconstruction, extracellular globins clustered into one group within a major clade containing Hbs also (Fig. 2). Based on our results, the extracellular globins appear to be phylogenetically closer to the intracellular globins. When additional extracellular globin sequences from other metazoans were added to the analysis, they all clustered together, forming a monophyletic group (Supplementary file 6). These results suggest that the extracellular globins from annelids are more closely related to the extracellular counterparts in other invertebrate phyla (Belato *et al*., 2019) than to the annelid intracellular globins. Studies suggest that extracellular globins arose from an ancient duplication event of an intracellular globin gene (Gotoh *et al*., 1987; Yuasa *et al*., 1996; Bailly *et al*., 2007; Belato *et al*., 2019). Considering that extracellular globins were found in both deuterostome and protostome lineages, such as Echinodermata, Mollusca, Platyhelminthes, and Brachiopoda (Belato *et al*., 2019), this duplication event most likely occurred at least before the protostome-deuterostome split.

Hbs are very ancient proteins and have already undergone several gene duplication events (Goodman *et al*., 1988; Hardison, 1998; Vinogradov *et al*., 2005, 2007; Storz, 2018; Pillai *et al*., 2020). Supporting these observations, we find that Hbs are divided into several subgroups, that may represent paralogs. Annelid Hbs clustered into one large clade that represents the annelid intracellular monomeric RBC Hbs (Goodman *et al*., 1975, 1988; Weber and Vinogradov, 2001; Bailly *et al*., 2007), and some other smaller clades with few sequences (Fig. 2, yellow clades). Moreover, as expected, Hbs from *G. dibranchiata* and *G. americana* clustered into one separate clade compared to other Hbs, which seems to represent the well-known distinct circulating glycerid RBC Hbs (Weber *et al*., 1977, Weber and Vinogradov, 2001).

Herein we present the first record of Adgbs in annelids, with all newly discovered sequences clustered into one well supported group (Fig. 2, orange clade), which is consistent with their conspicuous inversion in the globin domain, with helices C-H preceding helices A-B (Hoogewijs *et al*., 2012; Bracke *et al*., 2018). In the gene tree of annelid globins + metazoan globins, the Adgbs of vertebrates and invertebrates clustered together, corroborating the hypothesis that Adgb genes predate the origin of

metazoans (Hoogewijs *et al*., 2012; Bracke *et al*., 2018). We also present the first record of invertebrate Cygbs within the Annelida. Cygbs grouped into two distinct clades (Fig. 2, pink clades), where one is a sister group to nHbs and the other is sister group to a clade that contains Mbs and Hbs (Fig. 2). These results indicate that Cygbs appear to have a high molecular affinity to other non-circulating intracellular globins, such as invertebrate Mbs and nHbs, similar to vertebrate Cygbs, which show phylogenetic affinities to vertebrate Mbs (Burmester *et al*., 2002). Annelid Mbs clustered into two well-supported clades (Fig. 2, purple clades), and these two distinct groups presumably represent the two major components MbI and MbII that have been reported in the annelid *Arenicola marina* (Weber and Pauptit, 1972; Kleinschmidt and Weber, 1998). Our results are in agreement with those from Suzuki and Imai (1998), which separated Mbs from invertebrates and vertebrates (Fig. 3; purple clades).

Some nerve Hbs clustered into one big clade (Fig. 2, blue clade) with high support values, and the other ones are mixed with other intracellular Hbs sequences (Fig. 2, yellow and blue clades), supporting the hypothesis of divergent evolutionary origins of invertebrate nHbs (Wittenberg, 1992; Weber and Vinogradov, 2001). Dewilde *et al*. (1996) and Burmester *et al*. (2000) have reported that nHbs from the worm *Aphrodita aculeata* reveal higher sequence similarity to the intracellular Hbs from the bloodworm *Glycera dibranchiata* than to other invertebrate nHbs. Our gene genealogy corroborates these results in that the nHbs from *Aphrodita japonica* are more closely related to the Hbs from *G. dibranchiata* than to the other nHbs grouped in the bigger clade (Fig. 2). Vertebrate neuroglobins and nHbs clustered in separate clades (Fig. 3), confirming that although both globins are localized in nerve tissues, they have different evolutionary origins (Weber and Vinogradov, 2001; Blank and Burmester, 2012; Pillai *et al*., 2020). The inferred tertiary structures of the globin genes suggested that they could have a putative respiratory function (Supplementary file 7).

In conclusion, our findings confirm a pattern evident from several recent studies, there is much greater phylogenetic distribution of oxygen-binding proteins than previously established, especially in annelids (Bailly *et al*., 2008; Martín-Durán *et al*., 2013; Costa-Paiva *et al*., 2017, 2018; Belato *et al*., 2019). We found two new intracellular non-circulating globin types within annelids: Adgbs and Cygbs, in addition to the two other documented types. We confirm that Mbs and Cygbs from vertebrates and those from invertebrates have different evolutionary origins. Our analyses demonstrate an intimate relationship between annelid extracellular globins and those from other metazoans, most likely because they were already present in the common ancestor of protostomes and deuterostomes – and reaffirm the crucial

importance of further comprehensive studies on the molecular evolution of the globin super-family across the metazoan evolutionary tree.

DATA DEPOSITION

This project has been deposited at GenBank under accession numbers MT311987 to MT312212.

COMPETING INTERESTS

The authors declare that they have no competing interests.

ACKNOWLEDGMENTS

F.A.B. was supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil). A fellowship to E.M.C.P. was provided by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo, Brazil – 2018/20268–1). The study was supported in part by FAPESP, by thematic project (Proc. 2016/06114-6), coordinated by Prof. Ricardo I. Trindade (Instituto de Astronomia e Ciências Atmosféricas—USP). Use of SkyNet computational resources at Auburn University is acknowledged. This work was funded in part by National Science Foundation (grants DEB–1036537 to K.M.H. and Scott R. Santos and OCE–1155188 to K.M.H.). This is a Molette Biology Laboratory contribution ## and Auburn University Marine Biology Program contribution ###. C.J.C.'s contributions are facilitated by start-up funds from the College of Science, Swansea University. We thank Dr. Diogo Melo and Silvia Mandai for critical reading of the manuscript and helpful suggestions. This work is part of the Ph.D. requirements of Flávia A. Belato at the Graduate Program in Zoology of the Institute of Biosciences of the University of São Paulo.

FIGURE LEGENDS

Figure 1 – Multiple amino acid sequence alignment of annelid Cygb (cytoglobin), Hb (hemoglobin), HBL (hexagonal bilayer hemoglobin), nHb (nerve hemoglobin), Mb (myoglobin), and manually rearranged Adgb (androglobin). Invariant amino acid residues at positions CD1, E7 and F8, which are diagnostic characters of the globin domain, are indicated in bold.

Figure 2 – Maximum likelihood gene genealogy of annelid globin genes rooted by midpoint. Bootstrap support values obtained from the maximum likelihood inference are shown in black, and the posterior

probabilities values obtained from the Bayesian inference are shown in red. To improve clarity, only support values above 80 or 0.8 are shown. Posterior probabilities values Yellow clades represent hemoglobin groups. Blue clades represent nerve hemoglobins. Purple clades represent myoglobins. Green clade represents hexagonal bilayer hemoglobins. Pink clades represent cytoglobins. Orange clade represents androglobins.

Figure 3 – Maximum likelihood gene genealogy of 43 annelid globin genes and 54 metazoan globin genes rooted by midpoint. Bootstrap support values obtained from the maximum likelihood inference are shown above the branches. Dark and light blue clades are nerve hemoglobin and vertebrate neuroglobin, respectively. Green clade is hexagonal bilayer hemoglobin. Yellow clades are invertebrate hemoglobins. Dark and light pink clades are invertebrate and vertebrate cytoglobin, respectively. Dark and light gray clades are vertebrate hemoglobin A and B. Red clade is androglobin. Dark and light purple clades are vertebrate and invertebrate myoglobin, respectively. Vertebrate and invertebrate Cygbs, Hbs, and Mbs do not represent genuine orthologs.

REFERENCES

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

Ashburner M et al. 2000. Gene Ontology: tool for the unification of biology. Nat Genet. 25:25–29.

Bailly X et al. 2007. Globin gene family evolution and functional diversification in annelids. FEBS J. 274(10):2641–2652.

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(1):244.

Belato FA, Schrago CG, Coates CJ, Halanych KM, Costa-Paiva EM. 2019. Newly discovered occurrences and gene tree of the extracellular globins and linker chains from the giant hexagonal bilayer hemoglobin in metazoans. Genome Biol Evol. 11(3):597–612.

Blank M, Burmester T. 2012. Widespread occurrence of N-Terminal acylation in animal globins and possible origin of respiratory globins from a membrane-bound ancestor. Mol Biol Evol. 29(11):3553–3561.

Bolognesi M, Bordo D, Rizzi M, Tarricone C, Ascenzi P. 1997. Nonvertebrate hemoglobins: Structural bases for reactivity. Prog Biophys Mol Biol. 68(1):29–68.

Bracke A, Hoogewijs D, Dewilde S. 2018. Exploring three different expression systems for recombinant expression of globins: *Escherichia coli*, *Pichia pastoris* and *Spodoptera frugiperda*. Anal Biochem. 543:62–70.

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, Ebner B, Weich B, Hankeln T. 2002. Cytoglobin: A novel globin type ubiquitously expressed in vertebrate tissues. Mol Biol Evol. 19(4):416–421.

Burmester T, Hankeln T. 2004. Neuroglobin: A respiratory protein of the nervous system. Physiology. 19(3):110–113.

Burmester T, Hankeln T. 2008. Neuroglobin and Other Nerve Haemoglobins. In: Bolognesi M, di Prisco G, Verde C, editors. *Dioxygen Binding and Sensing Proteins*. Springer Milan: Milano. p. 211–222.

Burmester T, Hankeln T. 2014. Function and evolution of vertebrate globins. Acta Physiol. 211(3):501–514.

Burmester T, Weich B, Reinhardt S, Hankeln T. 2000. A vertebrate globin expressed in the brain. Nature. 407(6803):520–523.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25(15):1972–1973.

Coates CJ, Decker H. 2017. Immunological properties of oxygen-transport proteins: hemoglobin, hemocyanin and hemerythrin. Cell Mol Life Sci. 74(2):293–317.

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

Costa-Paiva EM, Schrago CG, Coates CJ, Halanych KM. 2018. Discovery of novel hemocyanin-like genes in metazoans. Biol Bull. 235(3):134–151.

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.

DeSalle R. 2015. Can single protein and protein family phylogenies be resolved better? J Phylogenetics Evol Biol. 3:116.

DeSanctis D et al. 2004. Crystal Structure of cytoglobin: The fourth globin type discovered in man displays heme hexa-coordination. J Mol Biol. 336(4):917–927.

Dewilde S et al. 1996. Globin and globin gene structure of the nerve myoglobin of *Aphrodite aculeata*. J Biol Chem. 271(33):19865–19870.

Farris JS. 1972. Estimating phylogenetic trees from distance matrices. Am Nat. 106(951):645–668.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39(suppl_2):W29–W37.

Finn RD et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44(D1):D279–D285.

Gabaldón T, Huynen MA. 2004. Prediction of protein function and pathways in the genome era. Cell Mol Life Sci. 61:930–944.

Gell DA. 2018. Structure and function of haemoglobins. Blood Cells, Mol Dis. 70:13–42.

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

Geuens E et al. 2004. Nerve globins in invertebrates. IUBMB Life. 56(11–12):653–656.

Goodman M, Moore GW, Matsuda G. 1975. Darwinian evolution in the genealogy of haemoglobin. Nature. 253(5493):603–608.

Goodman M et al. 1988. An evolutionary tree for invertebrate globin sequences. J Mol Evol. 27(3):236–249.

Gotoh T, et al. 1987. Two globin strains in the giant annelid extracellular haemoglobins. Biochem J. 241(2):441–445.

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

Hardison R. 1996. A brief history of hemoglobins: plant, animal, protist, and bacteria. Proc Natl Acad Sci. 93(12):5675–5679.

Hardison R. 1998. Hemoglobins from bacteria to man: evolution of different patterns of gene expression. J Exp Biol. 201(Pt 8):1099–1117.

Hess PN, Russo CAM. 2007. An empirical test of the midpoint rooting method. Biol J Linnean Soc. 92(4):669–674.

Hoffmann FG, Opazo JC, Storz JF. 2011. Differential loss and retention of cytoglobin, myoglobin, and globin-e during the radiation of vertebrates. Genome Biol Evol. 3:588–600.

Hoffmann FG, Opazo JC, Storz JF. 2012. Whole-genome duplications spurred the functional diversification of the globin gene superfamily in vertebrates. Mol Biol Evol. 29(1):303–312.

Hoogewijs D et al. 2012. Androglobin: A chimeric globin in metazoans that is preferentially expressed in mammalian testes. Mol Biol Evol. 29(4):1105–1114.

Hourdez S et al. 2000. Gas transfer system in *Alvinella pompejana* (Annelida Polychaeta, Terebellida): Functional properties of intracellular and extracellular hemoglobins. Physiol Biochem Zool. 73(3):365–373.

Huerta-Cepas J, et al. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44(D1):D286–D293.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 14(6):587–589.

Kanehisa M, Sato Y, KawashimaM, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44(D1):D457–D462.

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

Kawada N et al. 2001. Characterization of a stellate cell activation-associated protein (STAP) with peroxidase activity found in rat hepatic stellate cells. J Biol Chem. 276(27):25318–25323.

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.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 10(6):845–858.

Kleinschmidt T, Weber RE. 1998. Primary structures of *Arenicola marina* isomyoglobins: Molecular basis for functional heterogeneity. Biochim Biophys Acta - Protein Struct Mol Enzymol. 1383(1):55–62.

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

Kraus DW, Doeller JE. 1988. A physiological comparison of bivalve mollusc cerebro-visceral connectives with and without neurohemoglobin. III. Oxygen Demand. Biol Bull. 174(3):346–354.

Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305(3):567–580.

Lagesen K, et al. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35:3100–3108.

Lamy JN, et al. 1996. Giant hexagonal bilayer hemoglobins. Chem Rev. 96(8):3113–3124.

Lankester ER. 1872. I. A contribution to the knowledge of hæmoglobin. Proc R Soc London. 21(139–147):70–81.

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

Mangum CP. 1985. Oxygen transport in invertebrates. Am J Physiol Integr Comp Physiol. 248(5):R505–R514.

Mangum CP. 1998. Major events in the evolution of the oxygen carriers. Am Zool. 38(1):1–13.

Marcotte EM et al. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science 285(5428): 751–753.

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.

Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 30(5):1188–1195.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32(1):268–274.

Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8(10):785.

Pettersen EF et al. 2004. UCSF Chimera–A visualization system for exploratory research and analysis. J Comput Chem. 25(13):1605–1612.

Pillai AS et al. 2020. Origin of complexity in haemoglobin evolution. Nature. doi: 10.1038/s41586-020-2292-y.

Rambaut A. 2009. FigTree. Tree figure drawing tool [accessed 2019 Oct 19]. Available from: http://tree.bio.ed.ac.uk/software/figtree/.

Rambaut A, Suchard MA, Xie D, Drummond AJ. 2014. Tracer v1:6 [accessed 2019 Oct 30]. Available from: http://beast.bio.ed.ac.uk/Tracer.

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

Royer WE, Sharma H, Strand K, Knapp JE, Bhyravbhatla B. 2006. Lumbricus erythrocruorin at 3.5 Å resolution: architecture of a megadalton respiratory complex. Structure 14(7):1167–1177.

Storz, JF. 2018 *Hemoglobin: Insights into protein structure, function, and evolution*. Oxford University Press.

Storz JF, Opazo JC, Hoffmann FG. 2013. Gene duplication, genome duplication, and the functional diversification of vertebrate globins. Mol Phylogenet Evol. 66(2):469–478.

Struck TH et al. 2015. The evolution of annelids reveals two adaptive routes to the interstitial realm. Curr Biol. 25(15):1993–1999.

Suzuki T, Imai K. 1998. Evolution of myoglobin. Cell Mol Life Sci C. 54(9):979–1004.

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

Terwilliger NB, Ryan M. 2001. Ontogeny of crustacean respiratory proteins. Am Zool. 41(5):1057–1067.

Trent JT, Hargrove MS. 2002. A ubiquitously expressed human hexacoordinate hemoglobin. J Biol Chem. 277(22):19538–19545.

Vázquez-Limón C, Hoogewijs D, Vinogradov SN, Arredondo-Peter R. 2012. The evolution of land plant hemoglobins. Plant Sci. 191–192:71–81.

Vinogradov SN. 1985. The structure of invertebrate extracellular hemoglobins (erythrocruorins and chlorocruorins). Comp Biochem Physiol B. 82(1):1–15.

Vinogradov SN, Moens L. 2008. Diversity of globin function: enzymatic, transport, storage, and sensing. J Biol Chem. 283(14):8773–8777.

Vinogradov, SN et al. 2013a. Microbial eukaryote globins. Adv. Microb. Physiol. 63:391–446.

Vinogradov SN et al. 2005. Three globin lineages belonging to two structural classes in genomes from the three kingdoms of life. Proc Natl Acad Sci. 102(32):11385–11389.

Vinogradov SN, et al. 2007. A model of globin evolution. Gene 398(1–2):132–142.

Vinogradov SN, Tinajero-Trejo M, Poole RK, Hoogewijs D. 2013b. Bacterial and archaeal globins – A revised perspective. Biochim Biophys Acta - Proteins Proteomics. 1834(9):1789–1800.

Vinogradov SN et al. 1993. Adventitious variability? The amino acid sequences of nonvertebrate globins. Comp Biochem Physiol Part B Comp Biochem. 106(1):1–26.

Weber RE. 1971. Oxygenational properties of vascular and coelomic haemoglobins from *Nephtys homerbgii* (Polychaeta) and their functional significance. Netherlands J Sea Res. 5(2):240–251.

Weber RE. 1978. Respiratory pigments. In: Mill PJ, editor. *Physiology of Annelids.* p. 393–446.

Weber RE. 1980. Functions of invertebrate hemoglobins with special reference to adaptations to environmental hypoxia. Am Zool. 20(1):79–101.

Weber RE, Pauptit E. 1972. Molecular and functional heterogeneity in myoglobin from the polychaete *Arenicola marina* L. Arch Biochem Biophys. 148(1):322–324.

Weber RE, Sullivan B, Bonaventura J, Bonaventura C. 1977. The haemoglobin systems of the bloodworms *Glycera dibranchiata* and *G. americana*. Oxygen binding properties of haemolysates and component haemoglobins. Comp Biochem Physiol Part B Comp Biochem. 58(2):183–187.

Weber RE, Vinogradov SN. 2001. Nonvertebrate hemoglobins: functions and molecular adaptations. Physiol Rev. 81(2):569–628.

Weigert A, Bleidorn C. 2016. Status of annelid phylogeny. Org Divers Evol. 16(2):345–362.

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.

Wittenberg JB. 1970. Myoglobin-facilitated oxygen diffusion: role of myoglobin in oxygen entry into muscle. Physiol Rev. 50(4):559–636.

Wittenberg JB. 1992. Functions of Cytoplasmic Hemoglobins and Myohemerythrin. In: Mangum CP, editor. *Blood and Tissue Oxygen Carriers. Advances in Comparative and Environmental Physiology*. Springer, Berlin, Heidelberg. p. 59–85.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 11(2):R14.

Yuasa HJ, et al. 1996. Electrospray ionization mass spectrometric composition of the 400 kDa hemoglobin from the pogonophoran *Oligobrachia mashikoi* and the primary structures of three major globin chains. Biochim Biophys Acta Protein Struct Mol Enzymol. 1296(2):235–244.

Table 1 – List of all taxa analyzed in which globin genes were found and the number of expressed genes in each species. Hb is hemoglobin, Mb is myoglobin, nHb is nerve hemoglobin, Cygb is cytoglobin, Adgb is androglobin, and HBL-Hb is hexagonal bilayer hemoglobin. GenBank accession numbers are also provided here and detailed in Supplementary file 4.

Taxon	Hb genes	Mb genes	nHb genes	Cygb genes	Adgb genes	HBL-Hb genes	Accession number
Acrocirridae							
Macrochaeta sp.	7						MT312144-50
Aeolosomatidae							
Aeolosoma sp.	4		1			2	MT312084-87 MT312044 MH995870-71
Alciopidae							
Alciopa sp.			3				MT312045-47
Alvinellidae							
Paralvinella palmiformis Desbruyères & Laubier, 1986	3				1		MT312166-68 MT311987
Paramphinome jeffreysii (McIntosh, 1868)	1			1			MT312169 MT312037
Ampharetidae							
Amphisamytha galapagensis Zottoli, 1983	1						MT312090
Auchenoplax crinita Ehlers, 1887	1						MT312103
Melinna maculata Webster, 1879	1			1			MT312152 MT312032
Amphinomidae							
Chloeia pinnata Moore, 1911	1						MT312108
Hermodice carunculata (Pallas, 1766)	1						MT312128
Pherecardia striata (Kinberg, 1857)	1						MT312171
Aphroditidae							

Aphrodita japonica Marenzeller, 1879	1	1	2				MT312093 MT311999 MT312048-49
Arenicolidae							
Arenicola loveni Kinberg, 1866		2					MT312000-01
Abarenicola pacifica Healy & Wells, 1959	1	1				2	MT312083 MT311998 MH995867-68
Aspidosiphonidae							
Aspidosiphon laevis Quatrefages, 1865	1		1				MT312102 MT312050
Lithacrosiphon cristatus (Sluiter, 1902)	2						MT312140-41
Branchiobdellidae							
Branchiobdella parasita (Braun, 1805)				1			MT312020
Chaetopteridae							
Chaetopterus variopedatus (Renier, 1804)	1						MT312106
Mesochaetopterus taylori Potts, 1914		1					MT312007
Chrysopetalidae							
Arichlidon gathofi Watson Russell, 2000	4						MT312095-98
Cirratulidae							
Aphelochaeta sp.	1					2	MT312092 MH996356 MH995881
Chaetozone sp.	1						MT312107
Tharyx kirkegaardi Blake, 1991	1						MT312206
Dinophilidae							
Dinophilus gyrociliatus O. Schmidt, 1857			1				MT312054
Dorvilleidae							
Ophryotrocha globopalpata Blake & Hilbig, 1990	1		1				MT312163 MT312073
Eunicidae							
Eunice norvegica (Linnaeus, 1767)	1						MT312113
Eunice pennata (Müller, 1776)	1	1					MT312114 MT312005
Marphysa sanguinea (Montagu, 1813)	1						MT312151
Palola sp.	2			1			MT312164-65 MT312036
Flabelligeridae							
Ilyphagus octobranchus Hartman, 1965			1				MT312066
Poeobius meseres Heath, 1930	1						MT312176
Glossoscolecidae							
Andiorrhinus sp.						2	MH995879-80
Pontoscolex corethrurus (Muller, 1857)				1			MT312039
Urobenus sp.				1			MT312043
Glyceridae							
Glycera americana Leidy, 1855	1						MT312120
Glycera dibranchiata Ehlers, 1868	4						MT312121-24
Hemipodia simplex (Grube, 1857)	3						MT312125-27
Goniadidae							
Goniada brunnea Treadwell, 1906		1					MT312006
Haplotaxidae							
Delaya leruthi (Hrabě, 1958)	1						MT312112
Haplotaxidae gen. sp.			1				MT312059
Hesionidae							

Hesionides sp.	1						MT312129
Microphthalmus listensis Westheide, 1967	1						MT312153
Microphthalmus similis Bobretzky, 1870			1	1			MT312070 MT312033
Histriobdellidae							
Histriobdella homari Beneden, 1858			2				MT312063-64
Hrabeiellidae							
Hrabeiella periglandulata Pizl and Chalupský, 1984	1		1				MT312130 MT312065
Komarekionidae							
Komarekiona eatoni Gates, 1974				1			MT312028
Lessoniaceae							
Eisenia sp.				1			MT312026
Lumbricidae							
Dendrobaena hortensis (Michaelsen, 1890)				1		1	MT312021 MT311991
Lumbrineridae							
Lumbrineris crassicephala Hartman, 1965	1			1			MT312142 MT312031
Ninoe nigripes Verrill, 1873	2			1			MT312159-60 MT312034
Maldanidae							
Axiothella rubrocincta (Johnson, 1901)	1	1					MT312105 MT312002
Clymenella torquata (Leidy, 1855)	1	1	1				MT312109 MT312003 MT312053
Nicomache venticola Blake & Hilbig, 1990	1	3					MT312158 MT312008-10
Praxillella pacifica Berkley, 1929	1	4			1		MT312177 MT312012-15 MT311988
Sabaco elongatus (Verrill, 1873)	2	1					MT312184-85 MT312016
Megascolecidae							
Amynthas sp.	1						MT312091
Pontodrilus litoralis (Grube, 1855)				1			MT312038
Microchaetidae							
Gattyana cirrhosa (Pallas, 1766)			1				MT312057
Kynotus pittarellii Cognetti, 1906				1			MT312029
Moniligastridae							
Drawida sp.				1			MT312025
Naididae							
Aulodrilus japonicus Yamaguchi, 1953	1						MT312104
Bothrioneurum vejdovskyanum Štolc, 1886				1			MT312019
Heterodrilus sp. 1			1			3	MT312062 MT311992-94
Nephtyidae							
Aglaophamus verrilli (McIntosh, 1885)	1			2		2	MT312088 MT312017-18 MH995874 MH995876

Nephtys incisa Malmgren, 1865	1		1				MT312156 MT312071
Nereididae							
Alitta succinea (Leuckart, 1847)	1						MT312089
Octochaetidae							
Dichogaster green tree worm				1			MT312022
Dichogaster guadeloupensis James, 1996				1			MT312023
Oenonidae							
Arabella sp.	1						MT312094
Drilonereis sp.		1					MT312004
Onuphidae							
Diopatra cuprea (Bosc, 1802)				1			
Opheliidae							
Armandia sp.	3						MT312099-101
Ophelina acuminata Örsted, 1843	2						MT312161-62
Orbiniidae							
Leitoscoloplos robustus (Verrill, 1873)	2		2				MT312138-39 MT312067-68
Naineris laevigata (Grube, 1855)	1						MT312155
Proscoloplos cygnochaetus Day, 1954	1						MT312179
Oweniidae							
Galathowenia oculata (Zachs, 1923)	2						MT312118-19
Owenia fusiformis Delle Chiaje, 1844				1			MT312035
Parergodrilidae							
Stygocapitella subterranea 2 Knöllner, 1934	3		1				MT312194-96 MT312081
Parvidrilidae							
Parvidrilus meyssonnieri DesChâtelliers & Martin, 2012	1						MT312170
Pectinariidae							
Pectinaria gouldii (Verrill, 1874)		1					MT312011
Phyllodocidae							
Eulalia myriacyclum (Schmarda, 1861)			1	1			MT312055 MT312027
Nereiphylla sp.	1						MT312157
Pilargidae							
Synelmis sp.	1						MT312197
Polygordiidae							
Polygordius sp.			1				MT312074
Polynoidae							
Halosydna brevisetosa Kinberg, 1855			1				MT312058
Hermenia verruculosa Grube, 1856			2				MT312060-61
Lepidonotus semitectus (Stimpson, 1856)				1			MT312030
Protodriloididae							
Protodriloides chaetifer (Remane, 1926)	4		1				MT312180-83 MT312077
Sabellariidae							
Idanthyrsus sp.	2						MT312131-32
Sabellidae							
Bispira pacifica (Berkeley & Berkeley, 1954)					1		MT311989
Myxicola infundibulum (Montagu, 1808)	1						MT312154
Scalibregmatidae							
Scalibregma inflatum Rathke, 1843	1						MT312186
Serpulidae							

Crucigera zygophora (Johnson, 1901)	2						MT312110-11
Galeolaria caespitosa Lamarck, 1818			1				MT312056
Serpula vermicularis Linnaeus, 1767	1		1				MT312188 MT312078
Spirobranchus kraussii (Baird, 1865)			2				MT312075-76
Siboglinidae							
Lamellibrachia luymesi van der Land & Nørrevang, 1975	1						MT312133
Osedax sp.						2	KT166962-63
Sclerolinum brattstromi Webb, 1964	1						MT312187
Siboglinum ekmani Jägersten, 1956	3						MT312189-91
Sigalionidae							
Sigalion sp.			1				MT312079
Sparganophilidae							
Sparganophilus sp.			1	1			MT312080 MT312040
Spionidae							
Boccardia proboscidea Hartman, 1940			2				MT312051-52
Laonice sp.						2	MT311995-96
Prionospio dubia Day, 1961	1						MT312178
Sternaspidae							
Sternaspis scutata (Ranzani, 1817)					1		MT311990
Sternaspis sp.						1	MT311997
Syllidae							
Odontosyllis gibba Claparède, 1863			1				MT312072
Syllis cf. *hyalina* Grube, 1863				1			MT312041
Terebellidae							
Eupolymnia nebulosa (Montagu, 1819)	3						MT312115-17
Lanicides sp.	4						MT312134-37
Lysilla sp.	1		1				MT312143 MT312069
Pista macrolobata Hessle, 1917	4						MT312172-75
Streblosoma hartmanae Kritzler, 1971	2						MT312192-93
Terebellides stroemii Sars, 1835	8						MT312198-205
Thelepus crispus Johnson, 1901	2						MT312207-08
Themistidae							
Themiste pyroides (Chamberlin, 1919)	1						MT312209
Travisiidae							
Travisia brevis Moore, 1923	3						MT312210-12
Tritogeniidae							
Tritogenia sulcata Kinberg, 1867				1			MT312042
Trochochaetidae							
Trochochaetidae gen. sp.			1				MT312082

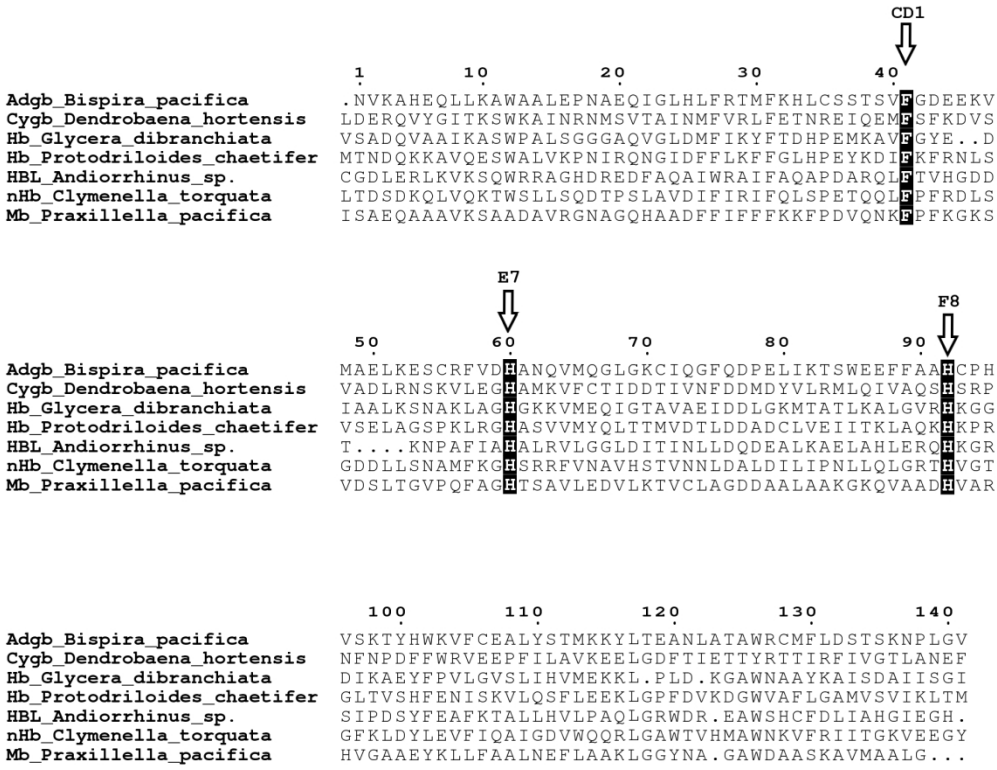

Figure 1 – Multiple amino acid sequence alignment of annelid Cygb (cytoglobin), Hb (hemoglobin), HBL (hexagonal bilayer hemoglobin), nHb (nerve hemoglobin), Mb (myoglobin), and manually rearranged Adgb (androglobin). Invariant amino acid residues at positions CD1, E7 and F8, which are diagnostic characters of the globin domain, are indicated in bold.

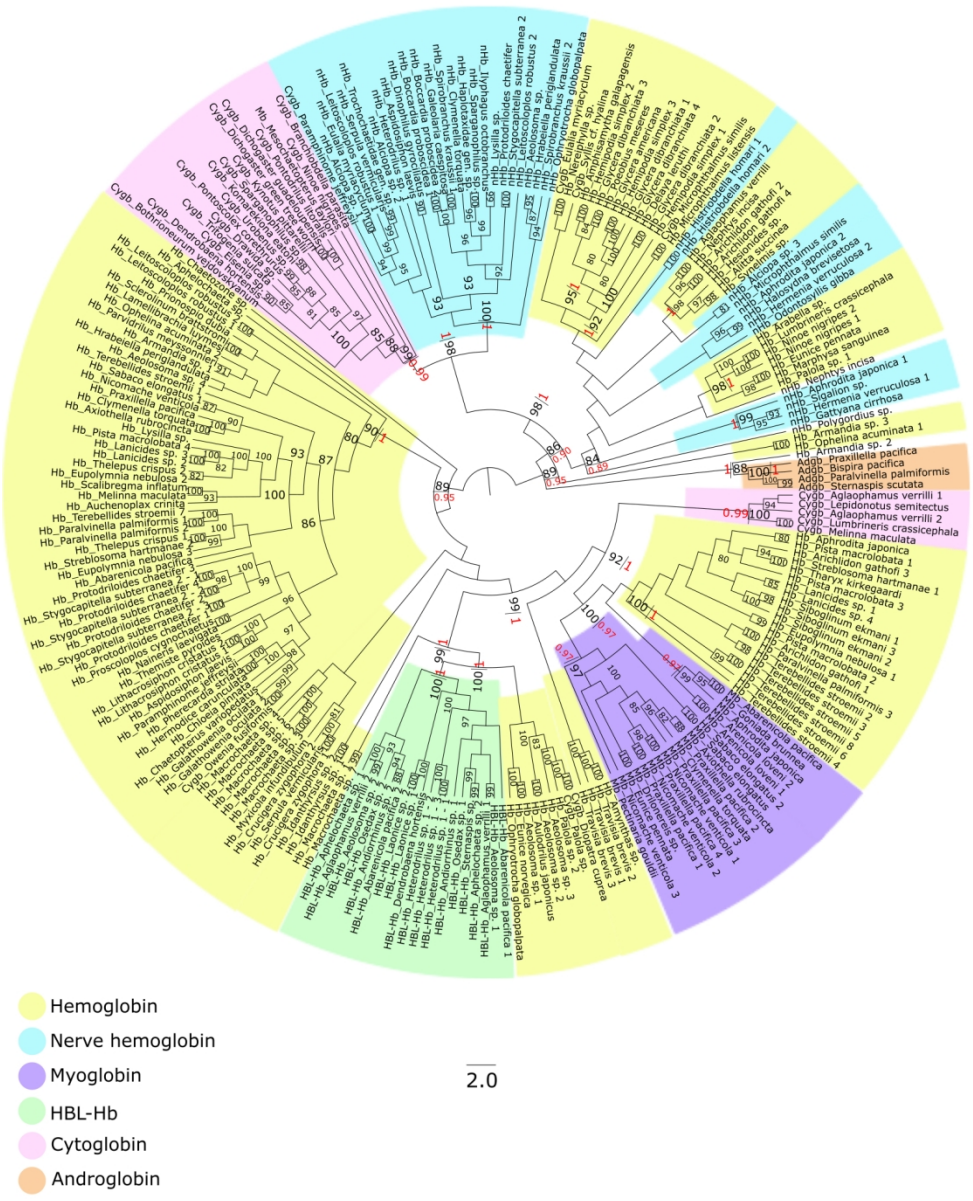

Figure 2 – Maximum likelihood gene genealogy of annelid globin genes rooted by midpoint. Bootstrap support values obtained from the maximum likelihood inference are shown in black, and the posterior probabilities values obtained from the Bayesian inference are shown in red. To improve clarity, only support values above 80 or 0.8 are shown. Posterior probabilities values Yellow clades represent hemoglobin groups. Blue clades represent nerve hemoglobins. Purple clades represent myoglobins. Green clade represents hexagonal bilayer hemoglobins. Pink clades represent cytoglobins. Orange clade represents androglobins.

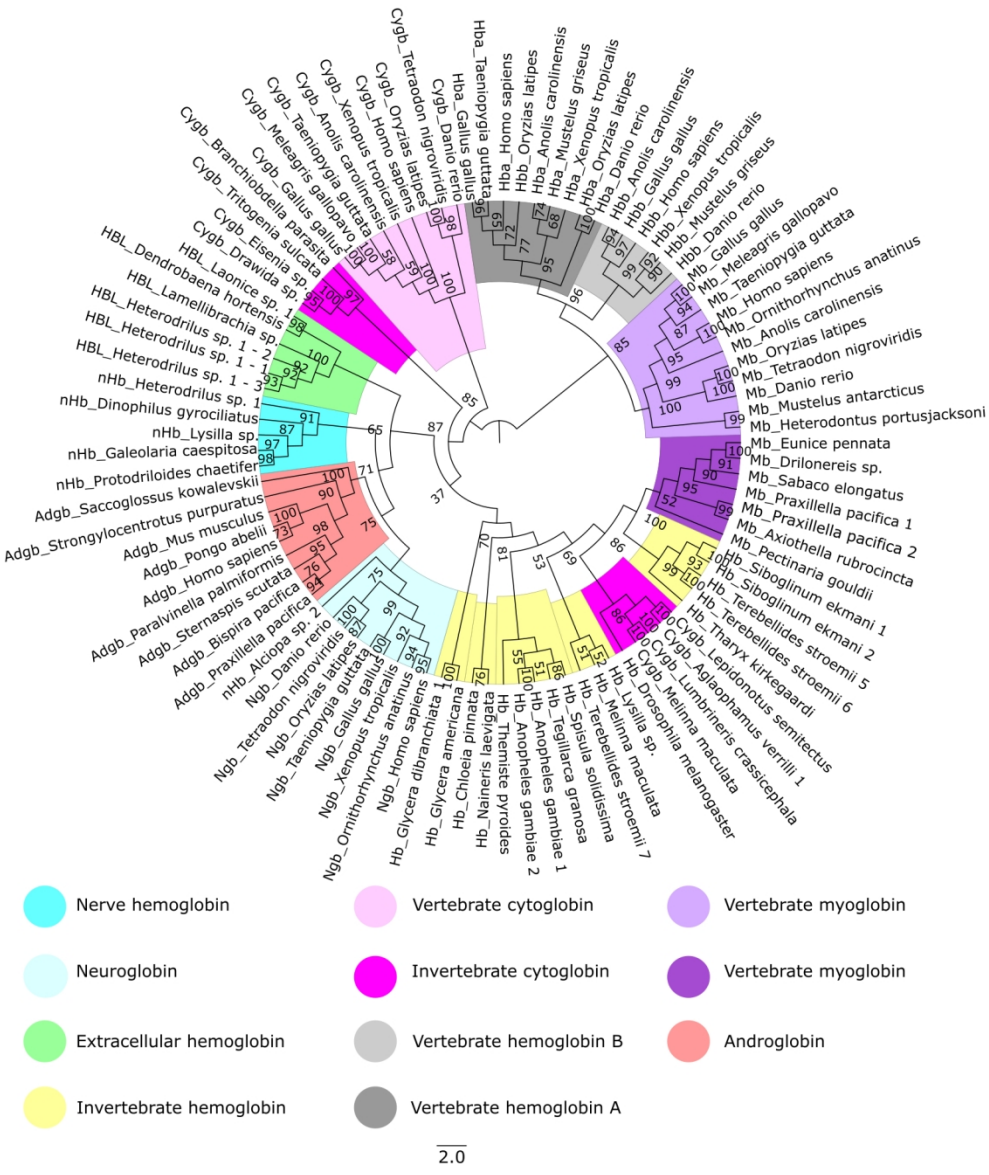

Figure 3 – Maximum likelihood gene genealogy of 43 annelid globin genes and 54 metazoan globin genes rooted by midpoint. Bootstrap support values obtained from the maximum likelihood inference are shown above the branches. Dark and light blue clades are nerve hemoglobin and vertebrate neuroglobin, respectively. Green clade is hexagonal bilayer hemoglobin. Yellow clades are invertebrate hemoglobins. Dark and light pink clades are invertebrate and vertebrate cytoglobin, respectively. Dark and light gray clades are vertebrate hemoglobin A and B. Red clade is androglobin. Dark and light purple clades are vertebrate and invertebrate myoglobin, respectively. Vertebrate and invertebrate Cygbs, Hbs, and Mbs do not represent genuine orthologs.