Evolution of single-domain globins in hydrothermal vent scale-worms

Joana Projecto-Garcia, A-S Le Port, T Govindji, D. Jollivet, Steven W Schaeffer, Stéphane Hourdez

▶ **To cite this version:**

Joana Projecto-Garcia, A-S Le Port, T Govindji, D. Jollivet, Steven W Schaeffer, et al.. Evolution of single-domain globins in hydrothermal vent scale-worms. Journal of Molecular Evolution, In press, pp.1-16. 10.1007/s00239-017-9815-7 . hal-01624983

HAL Id: hal-01624983

https://hal.sorbonne-universite.fr/hal-01624983

Submitted on 27 Oct 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Evolution of single-domain globins in hydrothermal vent scale-worms

Projecto-Garcia[1,2*] J, A-S Le Port[1, 2], T. Govindji[3], D. Jollivet[1,2], S. W. Schaeffer[3] and S. Hourdez[1,2]

[1] CNRS UMR 7144, Station Biologique de Roscoff, Place Georges Teissier 29680 Roscoff, France

[2] Sorbonne Universités, UPMC Univ. Paris 06, Laboratoire Adaptation et Diversité en Milieu Marin, Place Georges Teissier, 29680 Roscoff cedex, France

[3] Pennsylvania State University, Department of Biology and Institute of Molecular Evolutionary Genetics, University Park, Pennsylvania 16802, USA

*Corresponding author:

Joana Projecto-Garcia (jucpgarcia@gmail.com)

ORCID: 0000-0002-7802-2029

Current address: Ragsdale Lab, Myers Hall 100, Indiana University, 3rd St Bloomington, IN 47405 USA

Keywords: extracellular globin, single-domain, positive selection, heme, oxygen affinity, Polynoidae

Acknowledgements

The authors would like to thank the crews of the ships and submersibles, as well as the chief scientists, of the cruises ATOS 2001 (project funded by Ifremer and INSU), Lau Basin (projects funded by two NSF grants to C.R. Fisher (NSF OCE 0240985 and NSF OCE 0732333)), and EPR 2001 (project funded by a NSF grant to C.R. Fisher (NSF OCE-0002729)). We would also like to thank Isabelle Boutet-Tanguy and Arnaud Tanguy for technical advice in lab, Matthieu Bruneaux, Anis Bessadok, and Mirjam Czjzek for protein modeling advice. This work is part of the project HYPOXEVO (Région Bretagne), Deep-Sea Annelid Biodiversity and Evolution (Fondation Total), and was supported by the ESTeam research Marie Curie grant under the 6th framework program from the European Commission.

Abstract

Hypoxia at deep-sea hydrothermal vents represents one of the most basic challenges for metazoans, which then requires specific adaptations to acquire oxygen to meet their metabolic needs. Hydrothermal vent scale-worms (Polychaeta; Polynoidae) express large amounts of extracellular single- and multi-domain hemoglobins, in contrast with their shallow-water relatives that only possess intracellular globins in their nervous system (neuroglobins). We sequenced the gene encoding the single-domain (SD) globin from nine species of polynoids found in various vent and deep-sea reduced microhabitats (and associated constraints) to determine if the Polynoidae SD globins have been the targets of diversifying selection.

Although extracellular, all the SD globins (and multi-domain ones) form a monophyletic clade that clusters within the intracellular globin group of other annelids, indicating that these hemoglobins have evolved from an intracellular myoglobin-like form. Positive selection could not be detected at the major ecological changes that the colonization of the deep-sea and hydrothermal vents represents. This suggests that no major structural modification was necessary to allow the globins to function under these conditions. The mere expression of these globins extracellularly may have been sufficiently advantageous for the polynoids living in hypoxic hydrothermal vents. Among hydrothermal vent species, positively selected amino acids were only detected in the phylogenetic lineage leading to the two mussel-commensal species (*Branchipolynoe*). In this lineage, the multiplicity of hemoglobins could have lessened the selective pressure on the SD hemoglobin, allowing the acquisition of novel functions by positive Darwinian selection. Conversely, the colonization of hotter environments (species of *Branchinotogluma*) does not seem to have required additional modifications.

Introduction

Hydrothermal vents are located along oceanic ridges or active convergent margins on the ocean floor. These areas are characterized by harsh and challenging conditions for metazoans because of the presence of heavy metals and sulfide (both toxic compounds), low availability of oxygen (hypoxia), high temperatures, and low pH (Childress and Fisher 1992; Tunnicliffe 1991). Despite such harsh conditions, hydrothermal vent communities are characterized by both a high abundance of specialized fauna (mostly endemic), and a low species richness. This low and specialized biodiversity mainly results from the strong selective constraints that act as a filter to species not adapted to cope with these conditions. The adaptive peculiarities developed by hydrothermal species can be observed at several levels: trophic ability, organ morphology, enzyme activity, respiratory pigment affinity, and ATP synthesis (Childress and Fisher 1992). In particular, response to hypoxia is possibly the most basic challenge that metazoans must overcome to thrive and reap the benefits of the local primary production (Hourdez and Lallier 2007).

As an example, respiratory adaptations found in hydrothermal vent species can affect different organizational levels. They can affect the animal behavior (avoidance of some areas, variations in ventilation), the morphology (increased gills surface areas, reduced diffusion distances), the biochemistry (metabolism, presence of respiratory pigments), and the molecule itself (properties of the respiratory pigments) (for a review, see Hourdez and Lallier 2007). In particular, respiratory pigments usually exhibit high oxygen affinities when compared to littoral species that live in well-oxygenated environments (Hourdez and Weber 2005; Hourdez and Lallier 2007). In some annelids, extracellular hemoglobins that circulate at high concentrations represent a significant form of oxygen storage. In addition, their high oxygen affinity allows oxygen uptake from the environment even when its partial pressure is low. Finally, some hemoglobins have the capacity to reversibly bind both O_2 and sulfide, an ability that is essential for the functioning of the symbiosis in the vestimentiferan tubeworm *Riftia pachyptila* (Arp and Childress 1983; Childress and Fisher 1992; Weber and Vinogradov 2001).

The Polynoidae scale-worms are very diverse in the hydrothermal ecosystem, representing ~10% of all invertebrate species (Tunnicliffe 1991). Different species occupy all the available hydrothermal habitats where metazoa are found, ranging from the coldest areas (~2ºC) to the warmest -and most hypoxic- areas near venting fluids (~40ºC). Before the discovery of hydrothermal vent species, scaleworms (annelids that include Polynoidae)

were thought to only possess intracellular globins, in the muscles (myoglobin) and particularly in the nerve cord (neuroglobin) (Weber 1978; Dewilde et al. 1996). Interestingly, all hydrothermal polynoid species possess red-colored coelomic fluid, due to the presence of extracellular hemoglobins (Hourdez et al. 1999a; Hourdez unpub. data). In the genus *Branchipolynoe* two basic types of extracellular hemoglobins exist, a single-domain and a tetra-domain globin. This latter type was shown to likely be the result of evolutionary tinkering based on the tandem duplication of an ancestral single-domain intracellular globin (Projecto-Garcia et al. 2010). Although tetra-domain hemoglobins are so far only restricted to the genera *Branchipolynoe* (Hourdez et al. 1999a) and *Branchinotogluma* (Hourdez, unpub. data), all the other endemic vent polynoids possess at least single-domain extracellular hemoglobins on which we focused our attention for the present study of their adaptive evolution.

Hypoxic vent environments led to functional innovations in respiratory pigments essential for the survival of species (Bailly et al. 2002, 2003; Projecto-Garcia et al. 2010). Detection of adaptive molecular signatures and of the action of positive selection at the amino acid level can be performed by looking at the variations of the non-synonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$) between either closely-related evolutionary lineages or between codon sites along the coding sequence of a given gene (Yang 1998; Yang and Nielsen 2002). Using this phylogenetic tool, we investigated the possible adaptive role of some amino-acid changes during the evolution of the single-domain extracellular globin in hydrothermal-vent scale worms from a wide range of contrasted conditions and life-styles (and thus different selective constraints), including hydrothermal vents, shallow-water and a non-vent abyssal polynoid species. We were especially interested in testing different lineages, between different ecological groups, for signatures of selection that could be relevant to hemoglobin (Hb) evolution in these contrasted environments: *i*) shallow water *vs* deep-sea; *ii*) deep-sea *vs* hydrothermal vents; *iii*) hydrothermal vents *vs* acquisition of gills and multidomain Hb and finally, within this last group, *iv*) commensal *vs* free-living species.

Materials and methods

Animal collection

The collected species, sampling area, and habitat are detailed in Fig. 1 and Table 1. All the deep-sea specimens were identified on board the research vessel, and immediately frozen and stored at -80ºC until used in the laboratory. The species were chosen to represent various microhabitats at hydrothermal vents, from the coldest with the least hydrothermal influence, to the warmest on the chimney walls (closest to the vent fluid), with temperatures reaching 40˚C near the animals. The pure hydrothermal fluid is anoxic, and its mixing in variable proportions will not only affect temperature but also oxygen contents: the warmer the area, the lower the oxygen concentration. *Branchinotogluma segonzaci* is a representative of the warmest habitat, on the chimney wall (20-40˚C). *B. trifurcus* and *Branchiplicatus cupreus* are usually found in colder areas (10-20˚C for the former, and 2-10˚C for the latter), farther away from the source of the fluid. A still-undescribed species of *Branchinotogluma* sp. inhabits the periphery of the vents, in water at a stable 2-3˚C. *Branchipolynoe seepensis* and *B. symmytilida* live in the mantle cavity of mussels symbiotic with thioautotrophic bacteria (obligatory commensalism: Van Dover et al 1999; Jollivet et al 2000), with temperatures usually ranging between 4 and 10˚C. Besides all these species with gills, *Lepidonotopodium williamsae* represents a free-living, non-branchiate endemic hydrothermal species, collected among mussels, and experiences temperatures in the same range as *Branchipolynoe* spp., and possibly slightly higher. In addition to these vent-endemic species, a deep-sea species of the subfamily Eulagiscinae was captured on bare rocks near hydrothermal vents but was not exposed to any vent influence (stable temperature, around 2-3˚C). *Harmothoe extenuata* is a temperate, shallow-water species, and was collected on the rocky shore in Roscoff, France. *Sthenelais boa* (Sigalionidae), a littoral scale-worm species closely related to polynoids (Norlinder et al 2012) was used as an outgroup. These three latter species do not possess extracellular single- or multi-domain hemoglobins but do have an intracellular globin in their nervous system (neuroglobin) (Weber 1978; Hourdez pers. obs.).

Nucleic acids extraction and cDNA synthesis

A standard phenol/chloroform protocol following proteinase K digestion (Sambrook et al. 1989) was used to extract genomic DNA (gDNA) from *Branchipolynoe symmytilida*, *B. seepensis*, *Branchiplicatus cupreus*, and *Lepidonotopodium williamsae*. For

Branchinotogluma segonzaci, *B. trifurcus*, and the Eulagiscinae, gDNA was isolated following a CTAB + PVPP extraction protocol (Doyle and Doyle 1987). For all species, total RNA was extracted from the anterior part of the worm's body using TRI Reagent® (Sigma) and following the manufacturer's protocol and, cDNA was then synthesized by reverse transcription using MMLV-Reverse Transcriptase with an oligo$(dT)_{18}$ or an anchored oligo(dT) primer (see Table S1 and S2).

cDNA and gene sequencing

Sequences were obtained following two different strategies: amplification by PCR on genomic or cDNA, and search in assembled transcriptomes obtained by assembly of Illumina HiSeq data.

For PCR amplification, degenerate primers were designed based on previous globin sequences from the Polynoidae *Branchipolynoe symmytilida* and *B. seepensis*, as well as neuroglobin from the Aphroditidae *Aphrodita aculeata*. The PCR conditions and the type of template (cDNA or gDNA) differed according to the species used for amplification (see Table S1). The PCR products were visualized on a 1.5% agarose gel containing ethidium bromide under UV light, and cloned with the TOPO TA Cloning kit (Invitrogen). The positive clones were sequenced, and the sequences were used to produce specific primers for all the species (Table S1 and S2). Directional chromosome walking on gDNA (see Projecto-Garcia et al. 2010 for details) was used to sequence the missing parts of the coding sequences, the 5' UTR, and the promoter region of the globin genes for some species. When the sequences were obtained in several fragments, sufficient overlap regions were used to assemble the various fragments into a full-length sequence.

For the two non-vent species (the deep-sea Eulagiscinae and the shallow water species *H. extenuata*), the intracellular globin sequence, was retrieved from RNA-Seq data (unpub. data). Briefly, total RNA was extracted as described above, checked for quality and sent for sequencing. The sequencing was performed at the McGill University platform with the Illumina HiSeq2000 technology. One lane per species was used and provided 80 million, paired-end, 108-base long sequences. For each species, the fragments were assembled with Velvet/Oases, using a Kmer length of 51. The globin sequences were recovered by tblastx on the assembled sequences using a vent species globin sequence as the query.

Protein sequence and phylogenetic analyses

The nucleotide sequences obtained by Sanger sequencing were assembled, checked and, edited based on their chromatograms with CodonCode Aligner 2.0.6 (http://www.codoncode.com/aligner/index.htm). All cDNA sequences were translated into amino acid sequences using the universal genetic code. The obtained sequences have been submitted to GenBank (accession numbers GU121978-GU121983; KJ756506, KJ756507 and KP984527). Multiple nucleotide and amino-acid sequence alignments were performed with multiple sequence alignment algorithm MUSCLE (Edgar 2004, part of software Geneious 7.0.3, created by Biomatters). The optimization was based on minimizing the number of indels, by adjusting the codon alignment to the amino acid sequences alignment using the invariant residue positions associated with the globin fold/heme pocket. This optimization was confirmed by the GUIDANCE filter (Penn et al. 2010), and all regions that were not highly supported (low GUIDANCE scores) were removed before subsequent analyses.

Tree reconstruction

A Bayesian reconstruction of the globin tree (Fig. 2) was performed with the software MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) using all the Polynoidae globin sequences obtained and other extracellular and intracellular annelid amino acid globin sequences (Fig. 3). We used the WAG+I+G+F model of amino acid substitutions (ProtTest 3.0, Darriba et al. 2011) run for 4 000 000 generations, sampling every 10 000 generations and using default priors.

A maximum likelihood (ML) tree (Fig. 4) with the single-domain globin sequences from all the polynoid species was constructed using the PhyML package (Guidon and Gascuel 2003) in Geneious 7.0.3 (Biomatters), using the GTR+I+G model (jModelTest 2.0, Darriba et al. 2012) for nucleotide substitution and NNI for topology search. Prior to this analysis the sequences were analyzed by Gblocks v0.91b (http://molevol.cmima.csic.es/castresana/Gblocks.html) and Gap Strip/Squeeze v2.1.0 (http://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html) to evaluate which gaps to retain/delete for further analyses. The bootstraps from the trees issued from the output alignments of those programs were considerably lower (data not shown) and we chose to proceed using the initial alignment (Fig. S1). This tree was used as the phylogenetic context for the positive selection analyses (Fig. 4).

Positive selection and associated tests (Codeml)

The search for potential positive selection among branches and codon sites was performed by maximum likelihood following the procedure described by Nielsen and Yang (1998), Yang (1998), Yang and Nielsen (2002) and the PAML program instructions (Codeml).

We used the single-domain globin phylogeny for the Polynoidae species as a framework (Fig. 4), using the *Sthenelais boa* (Sboa) sequence as an outgroup. We first tested whether the d_N/d_S (ω) ratios were different among lineages with a likelihood ratio test (LRT = $2\Delta\ell$) between the *one-ratio branch model* (same ω for all branches) and the *free-ratio branch model* (ω free to vary among branches). The LRT results can be compared to a χ^2 distribution, with the number of degrees of freedom equal to the difference in the number of parameters between the two models (Yang 1998). Power and accuracy of the LRT were evaluated by Anisimova et al. (2001), with good results against violation of assumptions. Once the branches with ω values at least twice that of the average value were identified (a possible indication of positive selection), we searched for differences of ω ratio among sites on those specific branches/lineages. Yang and Nielsen (2002) implemented a test that lets the ω ratio vary both among sites and among lineages (branch-site model). We performed a LRT test comparing MA, a combination of the *two-ratio branch model* with the *positive selection site model* (M2a where codons fall in three ω categories ($0<\omega<1$, $\omega=1$, $\omega>1$), Yang and Nielsen 2002), against the nearly neutral site model (M1a where codons fall in 2 ω categories ($0<\omega<1$, $\omega=1$), Yang and Nielsen 2002). A second test, comparing M1a against a MA with fixed $\omega_2 = 1$ ($MA_{\omega=1}$), allowed us to test whether the site variability was actually due to positive selection rather than genetic drift or relaxed selection (Yang & Nielsen 2002; Wong et al. 2004).

Sites under positive selection were identified by a Bayesian analysis, where a posterior probability to belong to a given site class ($0 < \omega < 1$, $\omega = 1$ or $\omega > 1$) is calculated (based on the parameter estimates of the dataset) for each site. By definition sites under positive selection belong to the site class $\omega > 1$. Only sites with posterior probabilities greater than 95% were considered (Yang 2008). We used the Bayes Empirical Bayes (BEB) test performed by the Codeml package. This method accounts for the sampling errors in maximum likelihood estimates of model parameters (compared to the earlier Naive Empirical Bayes analysis), more adapted for small data sets like ours (Yang et al. 2005).

Ancestral sequence reconstruction

Using the same globin phylogeny (Fig. 4) as a reference, the ancestral sequences were reconstructed by Maximum Likelihood based on Bayesian statistics (Koshi and Goldstein 1996; Yang 2008, and the PAML program instructions), through Codeml (model = 0 and NSsites = 0).

Three-dimensional modeling of globins and localization of key amino-acid replacements

To construct a 3D homology protein model of some of the polynoid globin sequences, we used the tools available on the SWISS-MODEL website (https://swissmodel.expasy.org/interactive), using ProMod3 and MODELLER (Arnold et al. 2006; Biasini et al. 2014; Bordoli et al. 2009). Briefly, this modeling tool allowed us to obtain a 3D model from an amino acid sequence of interest based on the available 3D structure of a PDB template sequence that has the best psi-blast score with our sequence. Atomic energy calculations and minimization of the force fields were optimized.

The product of this rough model was visualized using UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (Pettersen et al. 2004). This same software was also used to graphically improve the model, to highlight some important residues, and to insert the heme group into the heme pocket of our model. For the insertion of the heme group, we used the coordinates from the template sequence. The analysis of the structural alignment was done using Pymol Molecular Graphics System v1.8.2.1 (DeLano 2008).

Recombinant globin expression and oxygen binding properties

The full-length coding sequences of *Branchipolynoe symmytilida*, *Branchinotogluma trifurcus*, and the Eulagiscinae globins were cloned into a pET20b vector, preserving the stop codon to prevent fusion with the His-Tag of this vector. Overexpression was performed in BL21 DE3 cells, grown in LB supplemented with ampicillin and in the presence of 1 mM 5-aminolevulinic acid (heme precursor), at 37˚C. After 4 hrs of induction with 1 mM IPTG, the cells were pelleted by centrifugation, resuspended in a lysis buffer (25 mM Tris/400 mM NaCl, pH 7.5) and the cells were lysed with a French press. Cellular debris was eliminated by centrifugation and the globin was purified by size exclusion chromatography from the supernatant onto a Superose 12 column with an elution buffer identical to the lysis buffer.

Oxygen equilibrium curves were obtained with a modified diffusion chamber (Sick and

289 Gersonde 1969) using a step-by-step procedure as previously described (Weber et al.
290 1976). Briefly, small (4 μl) aliquots of purified recombinant globin solution (~0.3 mM
291 heme final concentration) were equilibrated with mixtures of pure N_2 and O_2 prepared by
292 mass-flow meters and the resulting variations of absorption spectra were followed at 430
293 nm with a diode array spectrophotometer (Ocean Optics). The saturation (S) versus PO_2
294 (partial pressure of oxygen) data were linearized according to the Hill equation, log(S/(1-
295 S)) = f(log PO_2), and the values of P_{50} (PO_2 at which the globin is half-saturated with
296 oxygen) and n_{50} (cooperativity at P_{50}) were derived from linear regressions on the data
297 points between 30 and 70% saturation. The sample pH was adjusted by dilution with a
298 buffer solution of greater strength (500 mM Tris/400 mM NaCl).
299

Results

Single-Domain gDNA/cDNA amplification and sequencing

Coding sequences

In this study, we produced globin sequences for *Branchinotogluma segonzaci*, *B. trifurcus*, *Branchiplicatus cupreus*, *Lepidonotopodium williamsae*, a species of Eulagiscinae, and *Harmothoe extenuata*.

For *Branchinotogluma segonzaci*, *B. trifurcus*, *Branchiplicatus cupreus*, and *Lepidonotopodium williamsae*, several slightly different cDNA sequences were obtained, indicating either polymorphism at a single coding locus (*i.e.* alleles) or the presence of different globin loci in these species. For the following analyses, a consensus sequence was produced for all species, considering the most common nucleotides between the sequenced clones and assembling the different parts of the gene where it was possible to align upstream and downstream regions. For *B. cupreus* the sequence differences were such (sequence identity of 90.7% between SD1 and SD2) that we likely have two different loci for each species (transition/transversion rate ratio $\kappa = 2.17$), and only one sequence was considered for the following analyses.

For *Branchipolynoe symmytilida*, *Branchipolynoe seepensis*, *B. segonzaci* and *B. trifurcus,* the complete cDNA sequences from the single-domain globin have a coding sequence of 417 nucleotides including the stop codon. For *Branchinotogluma* sp. nov. and *B. cupreus* we could only amplify 366 bp (122 codons, including the initial methionine) of the coding sequence, and 385 for *L. williamsae*. These partial sequences correspond to the first two exons, and most of the third (and last) exon. Finally, for the Eulagiscinae and *Harmothoe extenuata* the complete coding sequences comprise 423 bp and 417 bp respectively.

Over the shared 354 bp, five indels were found, two common to all Polynoidae species (compared to the Sigalionidae *Sthenelais boa*), the third present in vent species only and the last two solely in *H. extenuata* (Fig. S1). Percentages of nucleotide identity between these single-domain globins is relatively low (37.9%; Fig. S1).

Promoter regions, and UTRs

For *B. symmytilida*, *B. cupreus*, and *L. williamsae*, our sequence covers the full 5'UTR (~68 bp), as well as about 440 bp of the promoter region for *B. symmytilida* and *L. williamsae*. For *B. seepensis*, *B. trifurcus*, and *B. segonzaci*, we successfully sequenced 48

bp of the 5'UTR (Fig. S2).

For *B. symmytilida* and *L. williamsae* the promoter sequences were slightly more conserved than their coding sequences (77.1% and 75.6% of identical sites, respectively). In both sequences the TATA box was located ~30 bp upstream of the beginning of the 5'UTR (Fig. S2). The identity between the amplified common parts of the 5'UTR (48 bp) for all vent polynoid species was ~80%. This value however drops drastically (47.1%) when the 5'UTR of *H. extenuata* is included (data not shown).

Introns

Introns were successfully amplified and sequenced in all species but the Eulagiscinae, *H. extenuata*, and intron 2 in *B. trifurcus*. As reported for *B. seepensis* and *B. symmytilida* (Projecto-Garcia et al. 2010), the single-domain genes all exhibit the typical vertebrate globin gene structure with 3 exons separated by 2 introns. The introns are located in the conserved positions B12.2 and G7.0 in reference to the *Physeter catodon* globin fold.

Intron sequence length differed considerably, especially for intron 1, which length ranged from 306 bp in *B. symmytilida* to 746 bp in *B. seepensis*. Intron 2 sequence length was also variable but with a more limited range, from 180 bp in *L. williamsae* to 295 bp in *B. seepensis*. The alignment between all orthologous intron sequences revealed limited identity (4.9% for intron 1 and 12% for intron 2). Within each genus for which we have two species (i.e. *Branchipolynoe*, and *Branchinotogluma*), however, the identity is higher (16.2% for intron 1 and 47.8% for intron 2).

Amino acid sequences and protein structure

The single-domain (SD) sequences obtained here were aligned with other annelid globins (intra- and extra-cellular), and as a reference we used globin sequences from other representative metazoan groups: invertebrates - two nematode extracellular hemoglobin sequences (*Ascaris suum*, pig intestinal parasite), and a vertebrates myoglobin from sperm whale (*Physeter catodon*) (Fig.2, accession numbers in Fig. 3).

In reference to the *Physeter* myoglobin fold, the alignment exhibits two conserved residues: a phenylalanine in the CD corner (CD1F) and the proximal histidine on the F helix, to which the heme is bound (F8H). The tryptophan in position A14 was conserved in nearly all globin sequences except for the nematode *Ascaris*, *Arenicola*, *Riftia* and *Alvinella*. All sequences also have a conserved tryptophan (H7W) that is not found in the *Physeter* myoglobin. Although extracellular, the Polynoidae globins do not possess the two

well conserved cysteines involved in a disulfide bridge in the typical extracellular globins from annelids (positions A2 and H10). Over the region for which we have a sequence overlap (118 amino acid residues), the Polynoidae sequences exhibit an amino acid identity of 50%. Several important amino acids in the heme pocket exhibit interesting characteristics. Two important residues that have been identified as key to the very high oxygen affinity in *Ascaris* Hb, tyrosine B10 and glutamine E7, are also present in *S. boa* and in all the Polynoidae sequences except the Eulagiscinae, for which the amino acids at both of these positions are replaced by a leucine. The pogonophoran annelid *O. mashikoi* also possesses a glutamine in E7.

Among the polynoid sequences, out of the 30 probable heme contacts (using the sperm whale myoglobin heme contacts as a reference, Fig. 3), only 11 residue positions are affected by changes.

No signal peptide for protein export was found in any of the species for which we obtained sequences upstream of the initial methionine.

Single-domain globin relationship with other globins

In comparison with the *Ascaris* and sperm whale globins , the annelid globins segregate into two initial lineages that separate the globins that form the typical extracellular hexagonal bilayer hemoglobins (HBL-Hb) from all other annelid globins (Bayesian phylogenetic tree, Fig. 2). The topology of the clade that comprises intracellular annelid globins and extracellular polynoid globins reflects the current knowledge of annelid phylogeny (Weigert and Bleidorn 2016). The Phyllodocida include all scaleworms (Aphroditidae, Sigalionidae, and Polynoidae) and Glyceridae in our tree. All the Polynoidae sequences group together, regardless of their extracellular or intracellular state.

Variation of d_N/d_S ratios among branches and tests for positive selection

Variations among lineages (branch model)

Tests for the past action of positive selection were performed using the Maximum Likelihood tree topology based on the 443 bp alignment of the globin gene (Fig. 4). From the two different single-domain globins SD1 and SD2 obtained for *Branchiplicatus cupreus*, only SD1 was used for the following analyses. The same analyses were also performed with SD2 and produced very similar results (data not shown).

The LRT between the *one-ratio branch model* and the *free-ratio branch model* was significantly different from zero, indicating that ω (d_N/d_S) ratios vary among lineages (LRT

401 = 28.98, df = 15, p < 0.025) (Yang 1998). The $\hat{\kappa}$ values (transition/transversion rate ratio)
402 were very similar between the different models, ranging from 1.66 to 1.71. Under the *one*
403 *ratio model* ω_0 is 0.148, indicating an overall moderate purifying selection (Table 2).
404
405 Focus on key evolutionary branches: (branch-site model)
406 We searched for signatures of evolutionary change in branches (Fig. 4) that correspond
407 to ecological transitions (littoral vs deep-sea and deep-sea vs hydrothermal vents),
408 anatomical/physiological transitions (absence of gills and multi domain Hb (hydrothermal
409 vents) vs the presence of gills and multi-domain Hb).
410 For all the ecological transitions, ω did not exceed 0.209, suggesting there was no major
411 non-synonymous substitutions accumulation in this protein to adapt its function between
412 littoral environments and deep-sea environments or the hypoxic habitats such as
413 hydrothermal-vents (Fig. 4). Two branches (***a*** and ***b*** on Fig. 4) exhibit infinite values for ω,
414 as a result of the absence of synonymous substitutions. For both branch ***a***, (genera
415 *Branchipolynoe* and *Branchinotogluma,* a lineage that developed gills and multi-domain
416 Hbs) and branch ***b***, we could not find any signature of positive selection (Fig. 4, Table 2).
417 Branch ***c***, leading to the two species of the genus *Branchipolynoe* (all commensal
418 species), exhibits a LRT significantly different from zero, indicating that there is a
419 signature of positive selection (Table 2) on this branch. The comparison between M1a and
420 MA showed that the latter best fit the data and additional tests corroborated this result (MA
421 *vs* $MA_{\omega=1}$, Table 2). The BEB analysis identified two residues significantly affected by
422 positive selection: 56T (position E11) and 82S (position F6).
423
424 **Ancestral globin reconstruction**
425 These analyses were performed to follow the amino acid substitutions that took place at
426 the nodes of each clade. Overall, the accuracy of the reconstruction had values of posterior
427 probability (PB) for codon change higher than 89%, except for the reconstructed node
428 leading to the outgroup *S. boa* (~66%). This latter node was therefore not taken into
429 consideration. *S. boa*, *H. extenuata* and Eulagiscinae exhibited more amino acid
430 substitutions compared to other sequences (Fig. S3). Interestingly several residues are
431 shared by the littoral *H. extenuata* and the deep-sea Eulagiscinae (node PB ~91%). These
432 residues are located in the B, D and G helices and CD and EF corners (Fig. S3). The
433 identity is greater for the species found at hydrothermal vents but the confidence of the

reconstruction of this node is below 0.95 (PB ~89%). Curiously, the ancestral node corresponding to branch ***b*** (PB ~95%) seems to be the departure point for several new residues specific to this clade (44S, 49I, 79T and 116G), with the exception of *B. trifurcus* (Fig. S3). On the lineage leading to *Branchipolynoe* (node PB ~99%), three residues are uniquely shared (23V, 56T and 82S), two of them are the same that were found to be under positive selection (Table 2).

Single-domain globin 3D modeling approach

Homology models were created only for species for which we had a complete sequence, *Branchipolynoe symmytilida*, *Branchinotogluma trifurcus* and the Eulagiscinae (Fig. 5). For the first species, the automatically chosen PDB template sequence was the monomer chain of the hemoglobin from *Lumbricus terrestris* (PDB: 1ASH, a high-resolution structure) that had 20% of amino acid identity with our sequences. Although this is close to the 'twilight zone' (<20% of amino acid identity), Pascual-García et al. (2010) showed that if two proteins are known to perform the same function, structural prediction is reliable even below this threshold. For *B. trifurcus* and the Eulagiscinae, the automatically chosen template with the highest structural identity was the sequence from the monomeric hemoglobin from *Glycera dibranchiata* (PDB: 1JF4), with 38% and 28% of amino acid identity, respectively.

Positively-selected residues in the *Branchipolynoe* lineage (branch ***c*** in Fig. 4) are highlighted on the *B. trifurcus* and *B. symmytilida* models for comparison (Fig. 5, dotted residues). In *Branchipolynoe* spp. E11T (E11V in *B. trifurcus*) is also located in the distal region of the heme pocket, and points in the same direction as E7Q and B10Y (Fig. 5 *a* and *b*), therefore potentially affecting ligand binding. The last amino acid under positive selection, F6S, in *Branchipolynoe* spp. (F6Q in *B. trifurcus*) is located in a helix region that, in other annelid globins, is important for the formation of oligomers (formation of dimers by interaction of helices E and F; Royer et al. 2001, 2005).

The residues highlighted in branches ***b*** and ***c*** by the ancestral reconstruction analyses are located in the B, F, and H helices, and in the DE corner. The substitutions in the B helix and DE corner were mostly from polar to non-polar residues (Fig. S4). On the other hand, the substitutions in the F and H helices were from non-polar to polar residues.

Oxygen binding properties

The oxygen binding properties of recombinant globins from *Branchipolynoe*

symmytilida, *Branchinotogluma trifurcus* and the Eulagiscinae were measured after their overexpression (Table 3). None of the cooperativity coefficients significantly differs from 1, indicating that, if multimers do form, this association does not allow cooperativity. Elution volumes of the different globins on a size exclusion column do not indicate differences of native mass either, suggesting all globins still remain monomeric (data not shown). As can be expected for globins that lack cooperativity, pH has no significant effect on P_{50} (data not shown). The two globins with B10Y and E7Q (*B. symmytilida* and *B. trifurcus*) both exhibit very similar P_{50} values that are much lower (*i.e.* greater affinities) than the globin from the Eulagiscinae (B10L and E7L). Amongst the two former species, the globin from *B. trifurcus* has a significantly greater affinity (lower P_{50}) than that of *B. symmytilida* (unpaired *t* test p=0.0003).

Discussion

Invertebrate hemoglobins exhibit a great structural and functional diversity (Weber and Vinogradov 2001). This diversity results from an early (i.e. more than 500 Mya) and complex evolutionary history and specific adaptations at the molecular level to contrasted environmental conditions (*e.g*. levels of oxygen, temperature), and physiological needs. Hydrothermal vents can be very challenging for aerobic organisms, especially in regard to hypoxia and the presence of sulfide (a potent inhibitor of aerobic metabolism) (Carrico 1978, Childress and Fisher 1992). The scale-worm species studied here also have adapted to a wide range of marine conditions and represent a very successful lineage that colonized the hydrothermal vent ecosystem (Fig. 4, Table 1), the usual deep-sea and the intertidal zone. Such challenging conditions can lead to functional innovations essential for the survival of the species.

Hemoglobin expression in vent species

Endemic hydrothermal vent polynoids typically possess extracellular hemoglobins in their coelomic fluid that confer them their red color (S. Hourdez, unpub. data). The sheer expression of hemoglobins in deep-sea polynoids can be regarded as an adaptation to hypoxic conditions as these proteins represent a form of oxygen storage that buffers variations of external oxygen concentrations (Hourdez et al. 1999b). It was estimated for *Branchipolynoe seepensis* that the amount of oxygen bound on hemoglobins could provide about 90-minute worth of aerobic metabolic needs if the worm is exposed to complete

anoxia (Hourdez and Lallier 2007). Although extracellular single-domain globins exist in all hydrothermal vent endemic polynoids, tetra-domain globins were only detected in the genera *Branchipolynoe* (Hourdez et al. 1999a; Zhang et al. 2017) and *Branchinotogluma* (S. Hourdez, unpub. data). The phylogenetic relationships indicate that all the studied polynoid extracellular globins (single- and tetra-domain) all derive from a common ancestral gene, which was probably intracellular (Projecto-Garcia et al. 2010, Fig. 2). The extracellular origin of these globins is distinct from the other annelid extracellular globins that diverged from the intracellular ones about 570 million years ago (Goodman et al. 1988).

All the globins sequenced here lack a signal peptide. In *Harmothoe extenuata* and the Eulagiscinae, this is not surprising because the globin is not free in the coelomic fluid but rather contained in cells (mostly in the nervous system, and possibly in muscles). The lack of a signal peptide, although surprising for the vent polynoid species, was already observed in the single- and tetra-domain globin from *Branchipolynoe seepensis* and *B. symmytilida* (Projecto-Garcia et al. 2010). In the vent species *Lepidonotopodium piscesae*, mass spectrometry data indicated a perfect match in molecular mass for both the myoglobin and the hemoglobin found in the coelomic fluid (unpub. data). This observation was used as evidence that the sequenced genes in *Branchipolynoe* spp. likely correspond to the hemoglobin found in the coelomic fluid and that it is released by holocrine secretion (Projecto-Garcia et al., 2010). The detection of a TATA box 30-base pair upstream of the 5'UTR start position in the promoter supports the absence of alternative splicing variants that would have a signal peptide for excretion.

Interestingly, the 5'UTR and the promoter regions are well conserved in most of the vent species. Although this may indicate some structural or regulatory function(s) for these regions, the physiological relevance of the presence of several regulatory motifs (e.g. CAC binding protein and GATA motifs, data not shown) in SD globins is yet to be ascertained.

Amino acid positions under positive selection

The heme pocket of all the polynoid single-domain globin sequences, except the Eulagiscinae, exhibit two conserved amino acid residues that are not under positive selection, B10Y and E7Q. These residues are therefore not recent innovations in the Polynoidae family but could be inherited from ancestral species that evolved under hypoxic conditions. B10Y and E7Q have been shown to be responsible for the very high oxygen affinity of the *Ascaris suum* globins (pig intestinal parasite), mostly through the

low oxygen dissociation rate that they provide (Davenport 1949 in Peterson et al. 1997; De Baere et al. 1994; Peterson et al. 1997). The replacement of the conserved distal histidine (E7H) by a glutamine (E7Q) and the B10L by a tyrosine (B10Y) seems a common convergent feature in many invertebrate globins (Weber and Vinogradov 2001), and could represent an adaptation to hypoxia. Even so, not all invertebrate globins possess the same high oxygen affinity that is observed in *A. suum*. The following invertebrate species, in terms of oxygen affinity, have values that represent at least 10 times higher P_{50} (i.e. lower Hb-O_2 affinity) than *Ascaris* Hb. This property is mostly dependent on the heme pocket conformation (Peterson et al. 1997).

The homology model of the structure of two polynoid globins, *B. symmytilida* and *B. trifurcus*, show that the B10Y and E7Q point towards the heme group. It is tempting to suggest that these residues are likely to participate, like in *A. suum*, on the high oxygen affinity measured in *Branchipolynoe* for both tetra-domain hemoglobins found in its coelomic fluid (Hourdez et al. 1999b). But such a residue configuration would be expected since the template used for this analysis also had the same residues pointing to the heme group.

However, the data obtained by the functional analyses done with recombinant globins of the vent species show a P_{50} 26-32 times lower than in the Eulagiscinae globin that possesses a leucine at both of these positions. Many other substitutions are found in the Eulagiscinae globin that could participate to the observed difference in affinity, but the two positions discussed have been experimentally shown to most profoundly affect oxygen binding in other invertebrates (extensively reviewed in Weber and Vinogradov 2001). The slight difference between *B. symmytilida* and *B. trifurcus* P_{50} values could be due to the sole replacement of a valine by a threonine in the heme pocket (position E11). Although allotropic effects due to amino-acid changes elsewhere in the molecule cannot be discounted, the E11 position is the only one position of the distal heme contacts that is different between the two species.

Despite many substitutions, the branches between the littoral species and the deep-sea species do not exhibit any signature of positive selection, suggesting there is no necessary important change for this protein to function under the high hydrostatic pressure experienced by all the other species in our study. This agrees with the fact that hydrostatic pressure does not induce denaturation or protein structural changes when temperature is constant (Mozhaev et al. 1996), like in deep-sea environments.

In the *Branchipolynoe* lineage some important amino acids, 56T (position E11) and 82S

(position F6) were found to be under positive selection, suggesting that this lineage experienced a more recent adaptive change. The replacement of 56V for a threonine, a residue similar in size but with a hydroxyl group capable of hydrogen bonds, in the E helix and facing the heme group, could influence O_2 binding. The 82S in the F helix, with a smaller side chain than glutamine and a lesser capability of forming bonds, could affect hydrophobicity around it.

Likelihood Ratio Tests can be especially conservative for small-length proteins (~100 codons; Anisimova 2003), close to the *ca*. 135 codons of globins. This could explain why the residue at the position B7 was not identified as under positive selection, even though B7V is shared in the *Branchipolynoe* lineage (and found in the Eulagiscinae globin). The substitution from asparagine (position 23), a polar and hydrophilic residue, for a valine, non-polar and with a short side chain, could reinforce the hydrophobic characteristics of the central part of the B helix.

Residues located in B7 and F6, could affect subunit interactions between single-domain globins in *Branchipolynoe*. The dimer interactions in *Lumbricus terrestris* hemoglobin are established through residues in the E and F helices (Royer et al. 2000), an interaction in which F6S could participate. In *L. terrestris*, dimers form tetramers mainly by the interaction of the loop formed by the AB corner. B7V is close to the AB corner and could be involved in interactions to form a multimer. The formation of multimeric assemblages may be beneficial as these hemoglobins are extracellular and larger molecular weight minimizes excretion (Weber and Vinogradov 2001). The absence of differences in native mass (as estimated by the elution volume by size-exclusion chromatography) between the recombinant *B. symmytilida* globin and that of the two other species argues against a difference in polymerization state. The absence of homotropic (cooperativity) or heterotropic (*e.g.* Bohr effect) characteristics also argues for an absence of polymerization. Even so, other multimeric globins can also exhibit the absence of these same characteristics (Royer et al. 2001), such as *Branchipolynoe* tetra-domain Hbs (Hourdez et al., 1999b) and *Ascaris* Hb (Gibson and Smith 1965; Okazaki and Wittenberg 1965).

Positive selection and molecular innovation

The hydrothermal vent scale-worms studied here are all exposed to generally hypoxic conditions (Hourdez and Lallier 2007). As one gets closer to the source of fluid, its proportion in the mix increases, the temperature rises, and the amount of oxygen decreases. The affinity for oxygen of the globins parallels this oxygen gradient, with the highest P_{50}

603 (i.e. lowest affinity) for the species exposed to the greatest oxygen partial pressure
604 (Eulagiscinae) and the lowest P_{50} (highest affinity) for the species exposed to the lowest
605 average oxygen partial pressure (*B. trifurcus*).
606 Interestingly, the event of positive selection did not take place in any branch
607 representative of major ecological shift. It occurred on the branch that comprises both
608 *Branchipolynoe* species. In this genus, there are two main tetradomain hemoglobins in the
609 coelomic fluid, and these exhibit different sensitivity to CO_2 (Hourdez et al. 1999b). This
610 is reminiscent of 'class II' fish in which hemoglobins found in the erythrocytes have
611 different functional properties and sensitivities to effectors that reflect a division of labor
612 (Weber 2000). In *Branchipolynoe*, this division of labor may be extended to the single-
613 domain globins, also found in the coelomic fluid. In the coelomic fluid of
614 *Branchinotogluma* (sister clade of *Branchipolynoe*), there is only one tetradomain
615 hemoglobin (S. Hourdez, unpub. data). The positively selected position in the
616 *Branchipolynoe* clade could correspond to a consequence of the appearance of the second
617 tetradomain globin. Species of this genus live inside the mantle cavity of Bathymodiolin
618 mussels where hypoxia can be severe. Females indeed stay within the valves of the host
619 and are quite territorial while they only tolerate mobile 'dwarf' males for reproduction
620 (Jollivet et al. 2000). These mussels rely on symbiotic thioautotrophic and/or
621 methanotrophic bacteria for at least part of their nutrition (Childress and Fisher 1992) and
622 flow water laden with sulfide and/or methane to meet their bacteria's metabolic needs.
623 This hypoxic water however also surrounds all other vent species, the level of hypoxia
624 depending on the amount of hydrothermal fluid in the mix. When the mussel closes, the
625 worms could be exposed to more severe hypoxic conditions and the modifications found
626 could be involved in dealing with these conditions.
627 The finding of absence of positive selection in branches representing ecological shifts
628 could be due to limitations of the method used. Indeed, globins tend to accumulate
629 substitutions at greater rate than other proteins. If an episode of positive selection
630 happened in much deeper branches, the accumulation of mutations since that time could
631 make the detection of the event more difficult. As we move deeper into the phylogeny of
632 these fast-evolving molecules, our confidence in the reconstruction of the ancestral state of
633 each position also decreases greatly and limits our ability to detect older events of positive
634 selection. However, in the tetradomain hemoglobins from *Branchipolynoe*, a study showed
635 that the initial domain duplication was accompanied by positive selection on amino acids
636 at the interface between two domains, possibly a response to structural constraints

637 (Projecto-Garcia et al. 2015).

638

639 References

640 Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in
641 detecting adaptive molecular evolution. Mol Biol Evol 18:1585-1592.

642 Anisimova M (2003) Detecting positive selection in the protein coding genes. Dissertation,
643 University College London.

644 Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL Workspace: A web-based
645 environment for protein structure homology modelling. *Bioinformatics*,**22**,195-201.

646 Arp AJ, Childress JJ (1983) Sulfide Binding by the Blood of the Hydrothermal Vent Tube Worm
647 Riftia pachyptila. *Science* 219: 295–297.

648 Bailly X, Jollivet D, Vanin S, Deutsch J, Zal F, Lallier F, Toulmond A (2002) Evolution of
649 the Sulfide-Binding Function Within the Globin Multigenic Family of the Deep-
650 Sea Hydrothermal Vent Tubeworm Riftia pachyptila. Mol Biol Evol 19:1421-1433.

651 Bailly X, Leroy R, Carney S, Collin O, Zal F, Toulmond A, Jollivet D (2003) The loss of
652 the hemoglobin H2S-binding function in annelids from sulfide-free habitats reveals
653 molecular adaptation driven by Darwinian positive selection. PNAS. 100:5885–
654 5890.

655 Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino
656 TG, Bertoni M, Bordoli L, Schwede T (2014) SWISS-MODEL: modelling protein
657 tertiary and quaternary structure using evolutionary information Nucleic Acids Res
658 42 (W1): W252-W258.

659 Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede, T (2009) Protein structure
660 homology modelling using SWISS-MODEL Workspace. Nature Protocols, **4**,1.

661 Carrico RJ, Blumberg WE, Peisach J (1978). The reversible binding of oxygen to
662 sulfhemoglobin. J Biol Chem 253:7212-7215.

663 Childress JJ and Fisher CR (1992) The biology of hydrothermal vent animals: physiology,
664 biochemistry, and autotrophic symbioses. Oceanogr Mar Biol - An Annual Review 30:337-
665 441.

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

668 Darriba D, Taboada GL, Doallo R and Posada D (2012) jModelTest 2: more models, new
669 heuristics and parallel computing. Nature Methods 9: 772.

670 Davenport, HE (1949) *Ascaris* Haemoglobin as an indicator of the oxygen produced by isolated
671 chloroplasts. P R Soc London B 136:281-290.

672 De Baere I, Perutz MF, Kiger L, Marden MC, Poyart C (1994) Formation of two hydrogen bonds
673 from the globin to the heme-linked oxygen molecule in *Ascaris* hemoglobin. P Natl Acad
674 Sci USA. 91:1594-1597.
675 DeLano, WL (2008) The PyMOL Molecular Graphics System. DeLano Scientific LLC, Palo Alto,
676 CA, USA. https://pymol.org
677 Dewilde S, Blaxter M, Hauwaert, M-L, Vanfleteren J, Esmans EL, Marden M, Griffon N, Moens
678 L (1996) Globin and Globin Structure of the Nerve Myoglobin of *Aphrodite aculeata*. J
679 Biol Chem 271:19865-19870.
680 Doyle JJ and Doyle JL (1987) A rapid DNA isolation procedure for small quantities of
681 fresh leaf tissue. Phytochem Bull 19:11-15.
682 Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high
683 throughput. Nucleic Acids Res 32:1792-1797.
684 Gibson QH, and Smith MH (1965) Rates of Reaction of *Ascaris* Haemoglobins with
685 Ligands. P Roy Soc Lond B Bio163: 206–14.
686 Goodman M, Pedwaydon J, Czelusniak J, Suzuki T, Gotoh T, Moens L, Shishikura F,
687 Walz D, Vinogradov SN (1988) An evolutionary tree for invertebrate globin
688 sequences. J Mol Evol 27:236-249.
689 Guidon S and Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large
690 phylogenies by maximum likelihood. Syst Biol 52: 696-704.
691 Hourdez S, Lallier FH, Green BN, Toulmond A (1999a) Hemoglobins from deep-sea
692 hydrothermal vent scale-worms of the genus *Branchipolynoe*: A new type of quaternary
693 structure. Proteins. 34:427-434.
694 Hourdez S, Lallier FH, Martin-Jézéquel V, Weber RE, Toulmond A (1999b)
695 Characterization and functional properties of the extracellular coelomic
696 hemoglobins from the deep-sea, hydrothermal vent scale-worm *Branchipolynoe*
697 *symmytilida*. Proteins. 34:435-442.
698 Hourdez S and Lallier F (2007) Adaptations to hypoxia in hydrothermal-vent and cold-seep
699 invertebrates. Rev Environ Sci Biotechnol 6:143-159.
700 Hourdez S. and Weber RE (2005) Molecular and functional adaptations in deep-sea hemoglobins.
701 J Inorg Biochem 99:130-141.

702 Jollivet D., Empis A, Baker MC, Hourdez S, Comtet T, Jouin-Toulmond C, Desbruyères
703 D, Tyler PA (2000) Reproductive Biology, Sexual Dimorphism, and Population

Structure of the Deep Sea Hydrothermal Vent Scale-Worm, *Branchipolynoe Seepensis* (Polychaeta: Polynoidae). J Mar Biol 80: 55–68.

Huelsenbeck J P and Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755.

Koshi JM, Goldstein R A (1996) Probabilistic Reconstruction of Ancestral Protein Sequences. J Mol Evol 42 :313-320.

Mozhaev VV, Heremans K, Frank J, Masson P, Balny C (1996) High Pressure Effects on Protein Structure and Function. Proteins: Structure, Function and Genetics 24:84-91.

Nielsen R and Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Norlinder E, Nygren A, Wiklund H, Pleijel F (2012). Phylogeny of scale-worms (Aphroditiformia, Annelida), assessed from 18SrRNA, 28SrRNA, 16SrRNA, mitochondrial cytochrome c oxidase subunit I (COI), and morphology. Mol Phylogenet Evol 65(2): 490-500.

Okazaki T, and Wittenberg JB (1965) The Hemoglobin of Ascaris Perienteric Fluid. BBA-Gen Subjects 111: 485–495.

Pascual-García A, Abia D, Méndez R, Nido GS, Bastolla U (2010) Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation. Proteins 78:181–96.

Penn O, Privman E, Ashkenazy H, Landan G, Graur D, and Pupko T (2010) GUIDANCE; a web server for assessing alignment confidence scores. Nucleic Acids Res 38: W23-W28.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera - A visualization system for exploratory research and analysis. J Comput Chem 25:1605-1612.

Peterson ES, Huang S, Wang J, Miller LM, Vidugiris G, Kloek AP, Goldberg DE, Chance MR, Wittenberg JB, Friedman JM (1997) A comparison of functional and structural consequences of the tyrosine B10 and glutamine E7 motifs in two invertebrate hemoglobins (*Ascaris suum* and *Lucina pectinata*). Biochemistry 36:13110-13121.

Projecto-Garcia J, Zorn N, Didier J, Shaeffer SW, Lallier FH and Hourdez S (2010) Origin and evolution of the unique tetra-domain hemoglobin from the hydrothermal vent scale-worm Branchipolynoe. Mol Biol Evol 27:143-152.

Projecto-Garcia J, Jollivet D, Mary J, Lallier FH, Schaeffer SW, Hourdez H (2015) Selective

737 forces acting during multidomain protein evolution: the case of multi-domain globins.
738 SpringerPlus 4:354.
739 Ronquist F, Huelsenbeck JP (2003) Mr Bayes 3: Bayesian phylogenetic inference under mixed
740 models. Bioinformatics 19:1572-1574.
741 Royer Jr WE, Strand K, van Heel M, Hendrickson WA (2000) Structural hierarchy in
742 erythrocruorin, the giant respiratory assemblage of annelids. P Natl Acad Sci USA
743 97:7107-7111.
744 Royer Jr WE, Knapp JE, Strand K, Heaslet HA (2001) Cooperative Hemoglobins: Conserved
745 Fold, Diverse Quaternary Assemblies and Allosteric Mechanisms. Trends Biochem Sci 26:
746 297–304.
747 Royer Jr WE, Zhu H, Gorr TA, Flores JF, Knapp JE (2005) Allosteric hemoglobin assembly:
748 Diversity and similarity. J Biol Chem 280:27477-27480.
749 Sambrook J, Fritsch EF, Maniatis T (1989) Molecular Cloning: A Laboratory Manual, vol.
750 I. 2nd edition. Cold Spring Harbor Laboratory Press.
751 Schlitzer R (2015) Ocean Data View 4, http://odv.awi.de.
752 Sick H and Gersonde K (1969) Method of continuous registration of O2 binding curves of
753 hemoproteins by means of a diffusion chamber. Ana. Biochem. 32 :362-376.
754 Tunnicliffe, V (1991) "The Biology of Hydrothermal Vents: Ecology and Evolution."
755 Oceanogr. Mar. Biol. Ann. Rev. 29: 319-407
756 Van Dover CL, Trask J, Gross J, Knowlton A (1999) Reproductive biology of free-living
757 and commensal polynoid polychaetes at the Lucky Strike hydrothermal vent field
758 (Mid-Atlantic Ridge). Marine Ecology Progress Series 181:201-214.
759 Weber RE (1978) Respiratory Pigments. Physiology of Annelids. Mill, P. J. London, Academic
760 Press Inc.
761 Weber RE (2000) Adaptations for oxygen transport: Lessons from fish hemoglobins. In
762 *Hemoglobin Function in Vertebrates, Molecular Adaptation in Extreme and Temperate*
763 *Environments* (ed. G. Di Prisco, B. Giardina and R. E. Weber), pp. 23-37. Milan: Springer-
764 Verlag Italia.
765 Weber RE, Lykkeboe G, Johansen K (1976) Physiological properties of eel haemoglobin :
766 hypoxic acclimation, phosphate effects and multiplicity. J. Exp. Bio. 64:75-88.
767 Weber RE and Vinogradov SN (2001) Nonvertebrate hemoglobins: functions and molecular
768 adaptations. Physiol Rev 81:569-628.
769 Weigert A, Bleidorn C (2016). Current status of annelid phylogeny. Org Div Evol 16(2):

345-362.
Wong WSW, Yang Z, Goldman N, Nielsen R (2004) Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites. Genetics 168:1041–1051.
Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568-573.
Yang Z (2008) Computational Molecular Evolution. Oxford Uni. New York.
Yang Z and Nielsen R (2002) Codon-substitution models for detecting molecular adaptations at individual sites along specific lineages. Mol Biol Evol 19:908-917.
Yang Z, Wong WSW, Nielsen R (2005) Bayes Empirical Bayes Inference of Amino Acid Sites under Positive Selection. Mol Biol Evol 22: 1107–18.
Zhang Y, Sun J, Chen C, Watanabe HK, Feng D, Zhang Y, Chiu JMY, Qian P-Y, Qiu J-W (2017) Adaptation and evolution of deep-sea scale worms (Annelida: Polynoidae): insights from transcriptome comparison with a shallow-water species. Sci Rep 7:46205.

786 Tables

787 **Table 1** Sampling areas and habitat of the different Polynoidae species (in alphabetical
788 order).

Species	**Sampling area, coordinates, and depth**	**Habitat**
Branchinotogluma segonzaci	Lau Basin 1. ABE (20°46'S, 176°11'W) 2150 m 2. Tow Cam (20°06'S, 176°34'W) 2700 m	Chimney walls, free-living
Branchinotogluma sp. nov.	Lau Basin, Kilo Moana (20°03'S, 176°08'W) 2600 m	Peripheral areas, free-living
Branchinotogluma trifurcus	Lau Basin 1. Kilo Moana (20°03'S, 176°08'W) 2600 m 2. Tu'i Malila (21°59'S, 176°34'W) 1900 m	*Ifremeria nautilei* aggregations, free-living
Branchiplicatus cupreus	East Pacific Rise, 9º50'N area (9º46'N, 104º21'W) 2500 m	Mussel beds, free-living
Branchipolynoe symmytilida	East Pacific Rise, 9º50'N area (9º46'N, 104º21'W) 2500 m	Mussel beds (commensal in mussel mantle cavity)
Branchipolynoe seepensis	Mid-Atlantic Ridge Lucky Strike site (37º18'N, 32º16'W) 1700 m	Mussel beds (commensal in mussel mantle cavity)
Eulagiscinae	Lau Basin Kilo Moana (20°03'S, 176°08'W) 2600 m	Peripheral areas
Harmothoe extenuata	Roscoff, France. 4-6 m	Underneath rocks

Lepidonotopodium williamsae	East Pacific Rise, 11˚N area (11˚25'N, 103˚47'W) 2500 m	Mussel beds and tubeworm aggregations, free-living

789

790

Table 2 Codeml parameters obtained under different codon substitution models. ln*L*= natural log of likelihood value, **κ** = transition/transversion rate ratio, LRT = Likelihood ratio test and degrees of freedom (df), BEB = Bayes Empirical Bayes. NA= Not Applicable. *p = 0.025, **p = 0.001

Model	ln*L*	κ	np	Model estimates	LRT (df)	Sites under positive selection (BEB>0.95)
Branch_model						
M0	-2152.58	1.708	18	$\omega = 0.148$		
M1	-2138.09	1.656	33	$0.001 < \omega < \infty$	28.98^{*} (15)	NA
Site_model						
M1a 'nearly neutral'	-2126.42	1.808	19	$\omega_0 = 0.101$ (83.2%) $\omega_1 = 1.000$ (16.8%)		
M2a 'positive selection'	-2126.42	1.808	21	$\omega_0 = 0.101$ (83.2%) $\omega_1 = 1.000$ (8.8%) $\omega_2 = 1.000$ (8%)	0.00^{NS} (2)	NA
Branch-site_model						
MA_branch ***a*** (BngnovSD + gills and multi-domain Hb)	-2126.25	1.803	21	$\omega_0 = 0.099$ (60.8%) $\omega_1 = 1.000$ (12.5%) $\omega_{2a} = 1.000$ (22.2%) $\omega_{2b} = 1.000$ (4.5%)	0.33^{NS} (2)	None
MA_branch ***b*** (gills and multi-domain Hb)	-2124.70	1.825	21	$\omega_0 = 1.001$ (80.7%) $\omega_1 = 1.000$ (17.2%) $\omega_{2a} = \infty$ (1.7%) $\omega_{2b} = \infty$ (0.4%)	3.44^{NS}(2)	None
MA_branch ***c*** (genus	-2116.77	1.792	21	$\omega_0 = 0.099$ (82%) $\omega_1 = 1.000$ (15.1%)	19.29^{**}	56T (E11T)[1] 82S (F6S)[1]

Model	ln*L*	κ	np	Model estimates	LRT (df)	Sites under positive selection (BEB>0.95)
Branchipolynoe)				$\omega_{2a} = \infty$ (2.5%) $\omega_{2b} = \infty$ (0.4%)		

[1] The position in the protein is given in parentheses as the name of the helix, the amino acid position in that helix, and the identity of the amino acid. This nomenclature is based in the sperm-whale myoglobin structure.

803 **Table 3** Oxygen binding properties of the different recombinant globins at 15˚C and
804 *Ascaris* Hb (at 20ºC), for comparison. P_{50}: partial pressure of oxygen necessary to reach
805 50% saturation of the binding sites. n_{50}: cooperativity coefficient at P_{50}. No significant pH
806 effect was detected, and reported values represent averages and standard deviations for the
807 different pH values tested. The amino acids at the positions responsible for the *A. suum* Hb
808 high affinity for O_2 (B10 and E7, shaded in gray) are indicated, along with the residue
809 positions that were under positive selection in *B. symmytilida*.

	P_{50} (mm Hg)	n_{50}	Amino acid in position			
			B10	E7	E11	F6
Branchipolynoe symmytilida	0.47±0.02 n=7	0.96±0.04 n=7	Y	Q	T	S
Branchinotogluma trifurcus	0.38±0.02 n=6	1.02±0.04 n=6	Y	Q	V	Q
Eulagiscinae	12.3±1.2 n=8	1.01±0.06 n=8	L	L	V	Q
Ascaris	0.001- 0.004[a]	1.0[a]	Y	Q	I	D/E[b]

810 [a]Gibson and Smith 1965, and Okazaki and Wittenberg 1965.
811 [b]De Baere et al 1992
812
813

814 Figures

815

816 **Fig.** 1 World map showing the locations of sampled species. Lau Basin: ABE (20˚46'S,
817 176˚11'W) 2150 m depth, Tow Cam (TC, 20˚06'S, 176˚34'W) 2700 m depth, Kilo Moana
818 (KM, 20˚03'S, 176˚08'W) 2600 m depth, Tu'i Malila (21º59'S, 176º34'W) 1900 m depth;
819 East Pacific Rise: 9º50'N area (9º46'N, 104º21'W) 2500m depth, 11ºN area (11º25'N,
820 103º47'W) 2500 m depth; Mid-Atlantic Rodge: Lucky Strike (LS, 37º18'N, 32º16'W)
821 1700 m depth; Roscoff, France, 4-6 m depth. Map obtained and edited through Ocean
822 View Data 4 (Schlitzer 2015).

823

824 **Fig. 2** Bayesian phylogenetic tree based on annelid globins residues corresponding to the
825 alignment in Fig. 3. The type of each globin sequence is identified in the figure. Zoom area
826 represents the Polynoidae single-domain globins. Posterior probability (PP) values when
827 indicated are near the respective branch or represented as such: ***: ≥0.95, **: ≥0.8, *:
828 ≥0.7. Values below 0.7 were not represented (lowest PP=0.5). The conserved amino acid
829 residues are indicated in each color-coded group; yellow: all sequences, green: all
830 sequences but *Ascaris*, *Arenicola*, *Riftia* and *Alvinella*, salmon: all sequences but sperm-
831 whale (Phyca). See Fig. 3 for abbreviations.

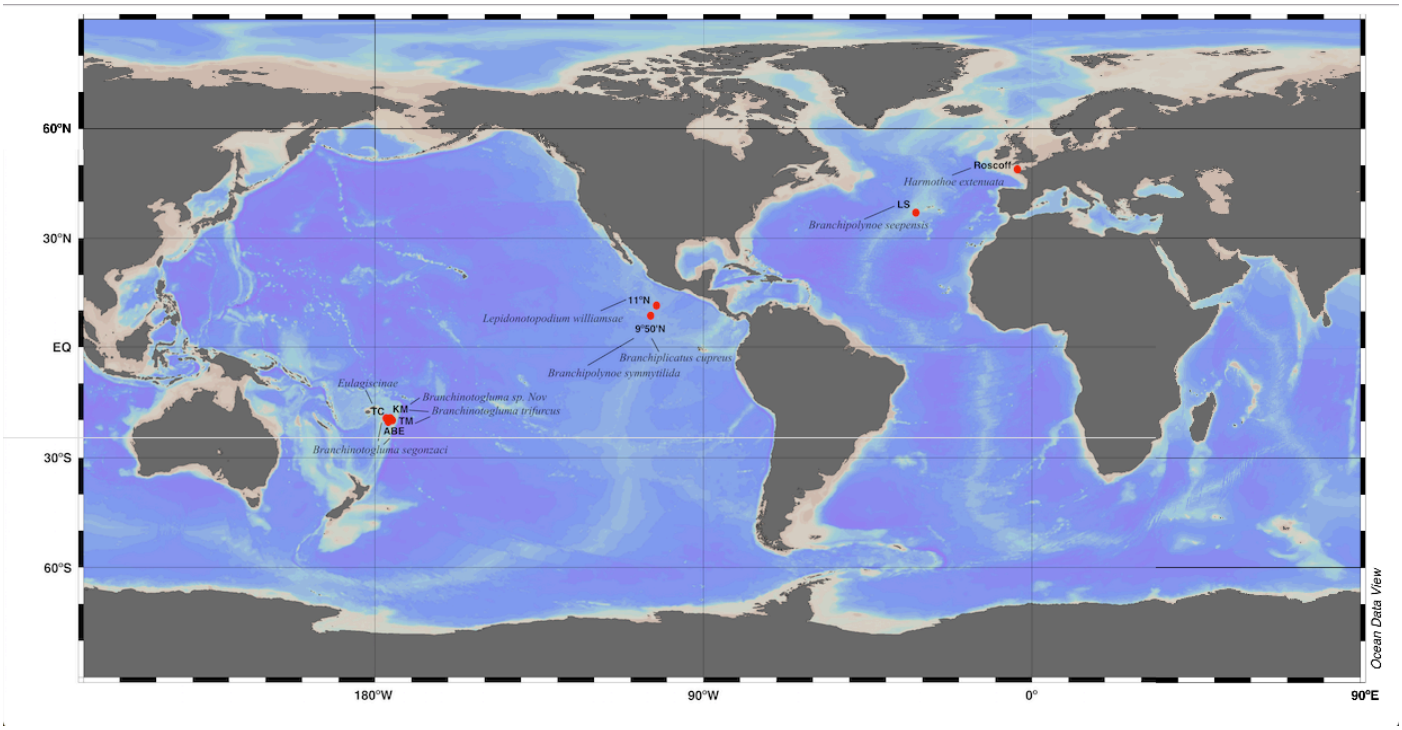
832

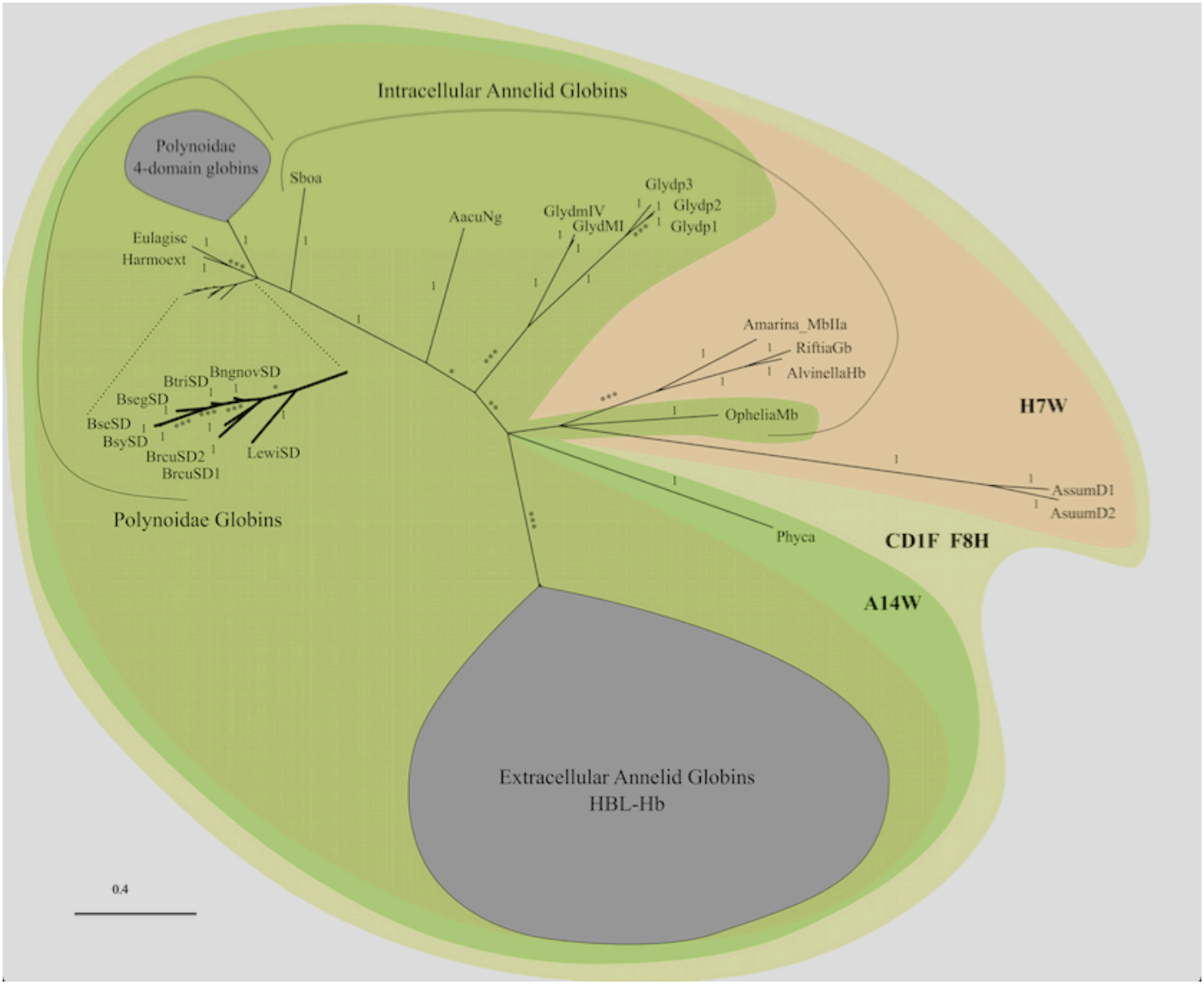
833 **Fig. 3** Alignment of globin sequences from annelids, nematodes and a vertebrate (sperm-
834 whale, in bold). Polynoidae single- and tetra-domain globin sequences are shaded in light
835 gray. Conserved residues are shown in bold (CD1F and F8H), heme pocket residues that
836 explain the high O_2 affinity in *Ascaris* where shaded in dark gray in the Polynoidae, and
837 other species. Cysteines forming an intrachain disulfide bridge in typical extracellular
838 annelid globins (A2C and H10C) are underlined. Arrows indicate the residues under
839 positive selection in *Branchypolynoe*. Intron (I1 and I2) conserved positions shown above
840 the sequences. *d* and *p* represent distal and proximal contacts with the heme group, having
841 the Phyca myoglobin as a reference. Polynoidae sequences: Bsy: *B. symmytilida*; Bse: *B.*
842 *seepensis*; Bseg: *B. segonzaci*; Btri: *B. trifurcus*; Bngnov: *Branchinotogluma sp. nov.*;
843 Brcu: *B. cupreus*; Lewi: *L. williamsae*; Eulagisc: Eulagiscinae; Harmoext: *H. extenuata*;.
844 Other globin sequences: Sboa: *Sthenelais boa* neuroglobin; Aacu: *Aphrodite aculeata*; Gly:
845 *Glycera* sp.; Tylo: *Tylorhynchus heterochaetus*; Lumt: *Lumbricus terrestris*; Tubifex:
846 *Tubifex tubifex*; Phese: *Pheretima seiboldi*; Rifb: Riftia pachyptila HBL-Hb and Riftia: *R.*

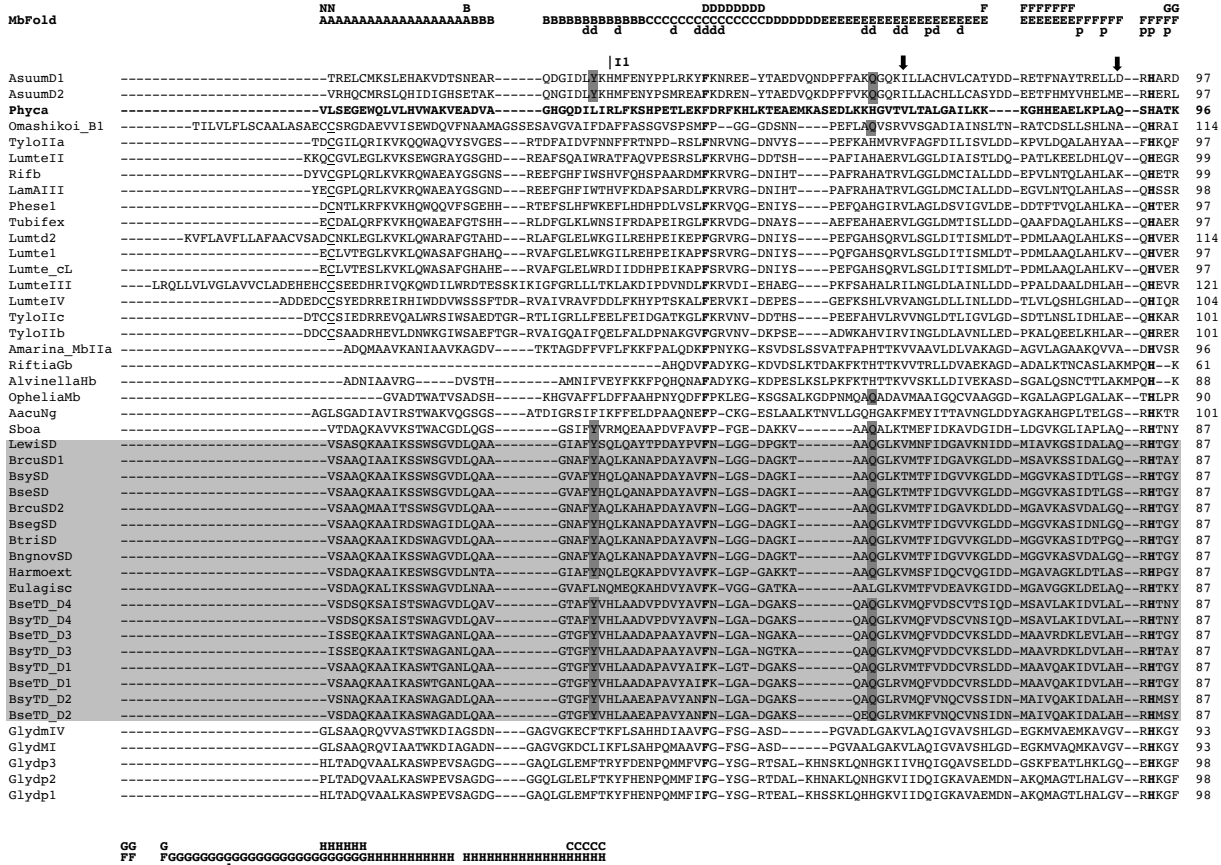
pachyptila intracellular globin; Lam: *Lamellibrachia* sp.; Amarina: *Arenicola marina*; Alvinella: *Alvinella pompejana*; Ophelia: *Ophelia bicornis*; Asuum: *Ascaris suum*; Omashikoi: *Oligobrachia mashikoi*; Phyca: *Physeter catodon*. SD: single-domain; D1-D4: multi-domain globin type; Ng: neuroglobin; Mb: myoglobin; Hb: hemoglobin.

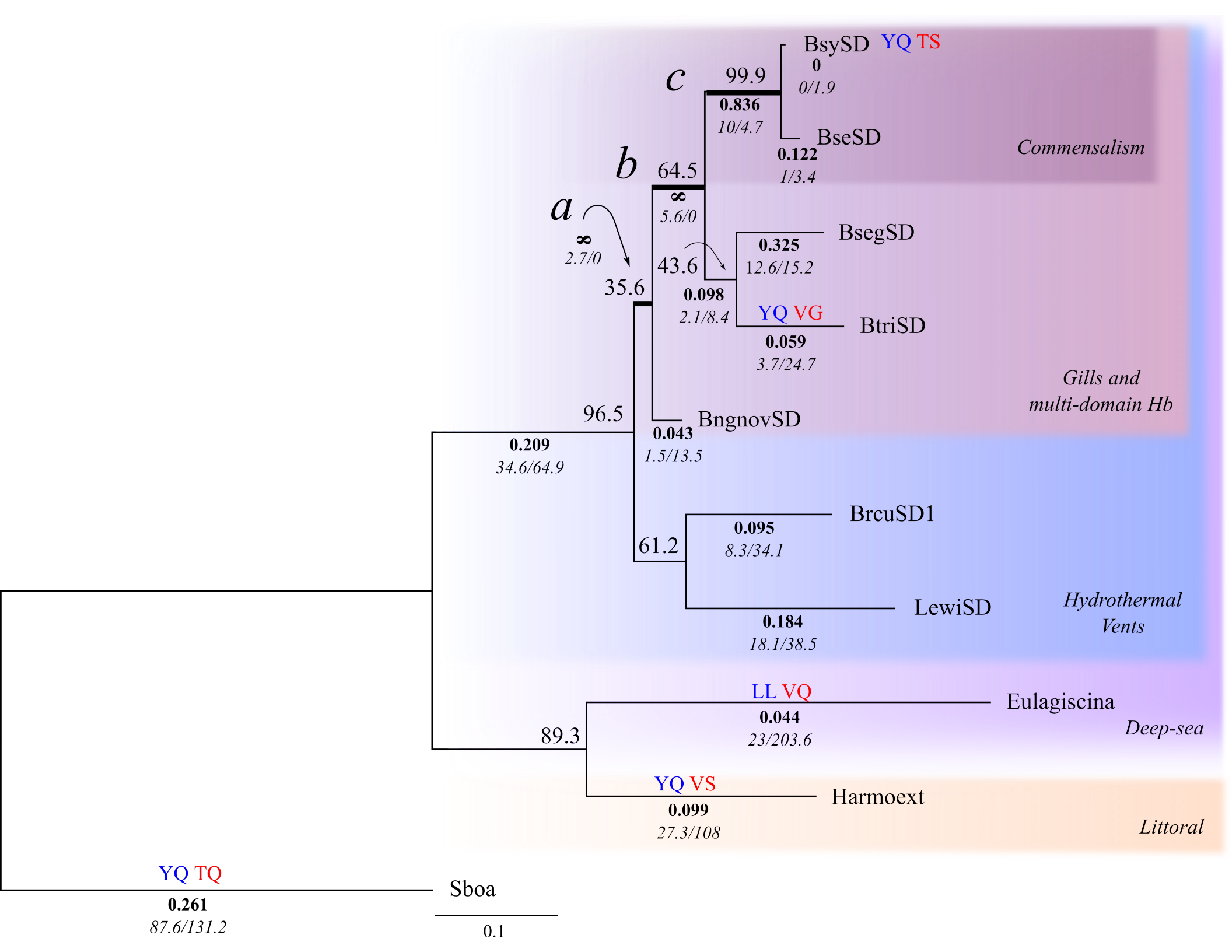
Fig. 4 Maximum likelihood globin tree (443 bp alignment). Bootstrap values are represented on top of each branch; for each lineage ω is represented in bold and ratios indicate the maximum likelihood estimates of the numbers of non-synonymous (d_N) over the synonymous (d_S) substitutions for the entire globin gene ; ***a, b*** and ***c*** represent the chosen lineages for the *branch-site model* test (see results). In relevant clades, amino acids in blue represent the positions correspondent to B10 and E7 (high O_2 affinity in *Ascaris*), and in red to E11 and F6 (positive selection in the *Branchipolynoe* branch). Species distribution and important characteristics are represented on the right of the tree. Sboa: *Sthenelais boa*, Harmoext: *Harmothoe extenuata*, Lewi: *Lepidonotopodium williamsae*, Brcu: *Branchiplicatus cupreus*, Bngnov: *Branchinotogluma* sp, Btri: *Branchinotogluma trifurcus*, Bseg: *B. segonzaci*, Bse: *Branchipolynoe seepensis*, Bsy: *B. symmytilida*. SD: single-domain

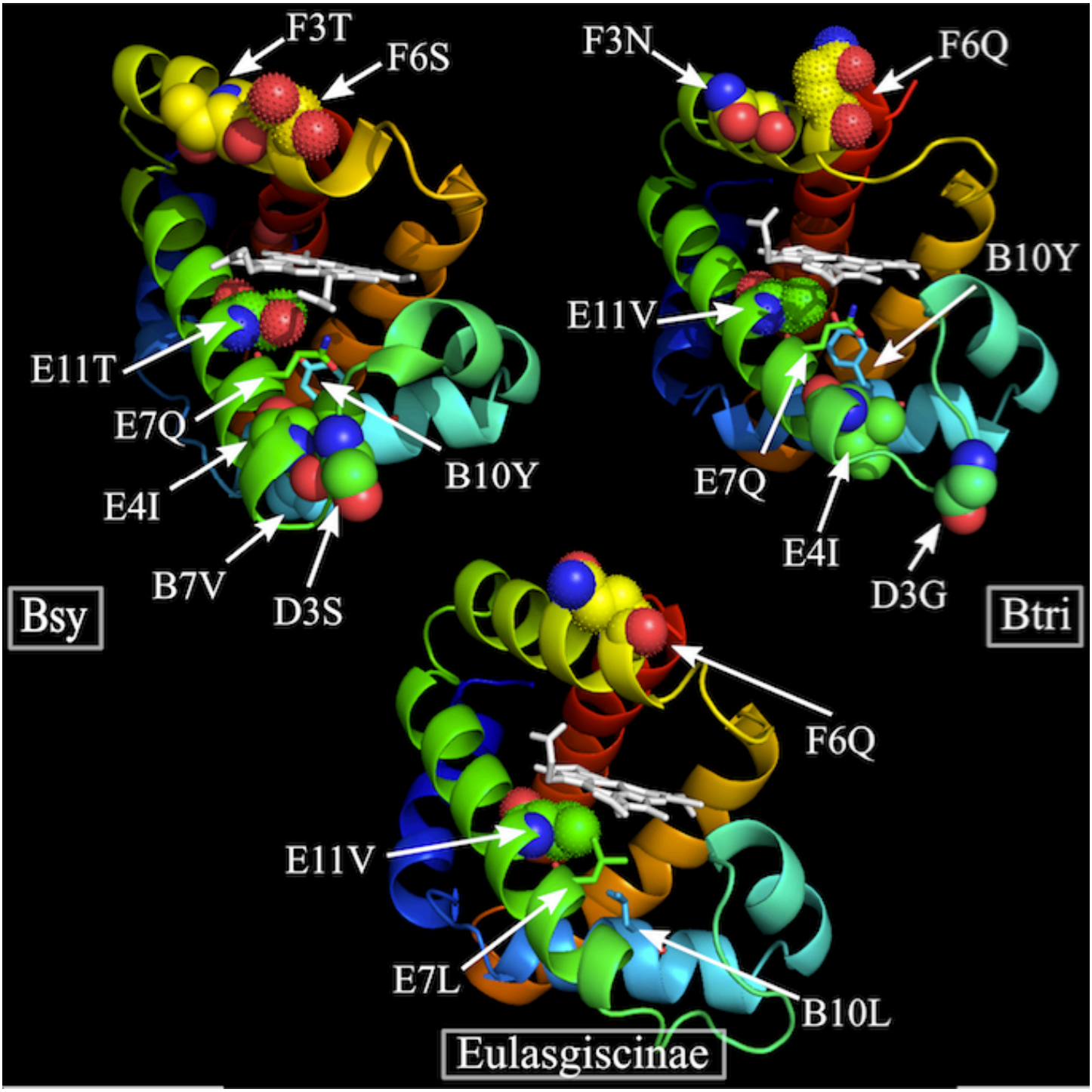
Fig. 5 3D structural model of *B. symmytilida* (Bsy), *B. trifurcus* (Btri) and Eulagiscinae single-domain globin. The amino acid residues that are invariant in Fig. 3 in both vent species (B10Y, and E7Q) are represented as sticks, residues target of positive selection in *Branchipolynoe* (E11T and F6S) are represented as rugged spheres (also depicted in the *B. trifurcus* and Eulagiscinae 3D models), residues highlighted by the ancestral reconstruction analyses (B7V, E11T, F6S in *Branchipolynoe* and D3S/G, E4I and F3T/N in branch ***b***) are represented as spheres.


60°N
30°N
EQ
30°S
60°S
180°W
90°W
0°
90°E
Roscoff
Harmothoe extenuata
LS
Branchipolynoe seepensis
11°N
Lepidonotopodium williamsae
9°50'N
Branchiplicatus cupreus
Branchipolynoe symmytilida
Eulagiscinae
Branchinotogluma sp. Nov
TC
KM
Branchinotogluma trifurcus
TM
ABE
Branchinotogluma segonzaci
Ocean Data View


Intracellular Annelid Globins
Polynoidae
4-domain globins
Sboa
AacuNg
GlydmIV
GlydMI
Glydp3
Glydp2
Glydp1
Eulagisc
Harmoext
Amarina_MbIIa
RiftiaGb
AlvinellaHb
H7W
OpheliaMb
BtriSD
BngnovSD
BsegSD
BseSD
BsySD
BrcuSD2
BrcuSD1
LewiSD
AssumD1
AsuumD2
Polynoidae Globins
Phyca
CD1F F8H
A14W
Extracellular Annelid Globins
HBL-Hb
0.4


MbFold
|I1
|I2
Acession numbers
AsuumD1 AAA29374 (part 1)
AsuumD2 AAA29374 (part 2)
Phyca P02185
Omashikoi_B1 BAD86545
TyloIIa P09966
LumteII P02218
Rifb P80592
LamAIII P15469
Phese1 P11740
Tubifex P18202
Lumtd2 AAC14536
Lumte1 P08924
Lumte_cL LTU055073
LumteIII P11069
LumteIV P13579
TyloIIc P13578
TyloIIb P02220
Amarina_MbIIa AJ880692
RiftiaGb EF648516
AlvinellaHb AJ880693
OpheliaMb AY926578
AacuNg AAC47259
Sboa FR770818
LewiSD GU121982
BrcuSD1 GU121980
BsySD GQ369757
BseSD GQ369758
BrcuSD2 GU121981
BsegSD GU121978
BtriSD GU121979
BngnovSD KJ756506
Harmoext KJ756507
Eulagisc KP984527
BseTD_D4 GQ369756
BsyTD_D4 GQ369555
BseTD_D3 GQ369754
BsyTD_D3 GQ369753
BsyTD_D1 GQ369749
BseTD_D1 GQ369750
BsyTD_D2 GQ369751
BseTD_D2 GQ369752
GlydmIV AAB31684
GlydMI P02216
Glydp3 AAA29161
Glydp2 AAA29160
Glydp1 CAA37995


BsySD YQ TS
0
0/1.9
c
99.9
0.836
10/4.7
BseSD
0.122
1/3.4
Commensalism
b
64.5
∞
5.6/0
a
∞
2.7/0
BsegSD
0.325
12.6/15.2
43.6
35.6
0.098
2.1/8.4
YQ VG
BtriSD
0.059
3.7/24.7
Gills and
multi-domain Hb
96.5
BngnovSD
0.043
1.5/13.5
0.209
34.6/64.9
BrcuSD1
0.095
8.3/34.1
61.2
LewiSD
0.184
18.1/38.5
Hydrothermal
Vents
LL VQ
Eulagiscina
0.044
23/203.6
89.3
Deep-sea
YQ VS
Harmoext
0.099
27.3/108
Littoral
YQ TQ
0.261
87.6/131.2
Sboa
0.1


F3T
F6S
E11T
E7Q
E4I
B7V
D3S
B10Y
Bsy
F3N
F6Q
B10Y
E11V
E7Q
E4I
D3G
Btri
F6Q
E11V
E7L
B10L
Eulasgiscinae