



HAL
open science

Origin and Evolution of the Unique Tetra-Domain Hemoglobin from the Hydrothermal Vent Scale Worm Branchipolynoe

J. Projecto-Garcia, N. Zorn, Didier Jollivet, S. W. Schaeffer, François Lallier, Stéphanne Hourdez

► **To cite this version:**

J. Projecto-Garcia, N. Zorn, Didier Jollivet, S. W. Schaeffer, François Lallier, et al.. Origin and Evolution of the Unique Tetra-Domain Hemoglobin from the Hydrothermal Vent Scale Worm Branchipolynoe. *Molecular Biology and Evolution*, 2010, 27 (1), pp.143-152. 10.1093/molbev/msp218 . hal-01250930

HAL Id: hal-01250930

<https://hal.science/hal-01250930v1>

Submitted on 28 Jan 2016

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.

1
2
3 1 Origin and evolution of the unique tetra-domain hemoglobin from the
4
5
6 2 hydrothermal vent scale-worm *Branchipolynoe*
7
8
9 3

10
11 4 SUBMISSION AS A RESEARCH ARTICLE
12
13 5

14
15
16 6 Projecto-Garcia^{1,2}, J., N. Zorn³, D. Jollivet^{1,2}, S. W. Schaeffer⁴, F. H. Lallier^{2,5} and S.
17
18 7 Hourdez^{1,2}
19
20 8

21
22
23 9 ¹ Équipe GAME, CNRS UMR 7144, Station Biologique de Roscoff, BP 74, Places Georges
24
25 10 Teissier, 29682 Roscoff cedex, France
26

27
28 11 ² Université Pierre et Marie Curie, Laboratoire Adaptation et Diversité en Milieu Marin, Place
29
30 12 Georges Teissier, 29682 Roscoff cedex, France
31

32
33 13 ³LDSM2 Institut de Chimie de Strasbourg, UMR 7177 CNRS-UDS, 1 rue Blaise Pascal - BP
34
35 14 296 R8, 67008 Strasbourg Cedex, France
36

37
38 15 ⁴Department of Biology and Institute of Molecular Evolutionary Genetics, Pennsylvania State
39
40 16 University, University Park, Pennsylvania 16802, USA
41

42
43 17 ⁵ Équipe EIMME, CNRS UMR 7144, Station Biologique de Roscoff, BP 74, Places Georges
44
45 18 Teissier 29682 Roscoff cedex, France
46

47 19

48
49 20 Corresponding author:

50
51 21 Stéphane Hourdez (hourdez@sb-roscoff.fr)
52

53
54 22 Équipe GAME, CNRS UMR 7144, Station Biologique de Roscoff, BP 74, Places Georges
55
56 23 Teissier, 29682 Roscoff cedex, France
57

58
59 24 Tel: +33 (0)2 98 29 23 40
60

25 Fax: +33 (0)2 98 29 23 24

1
2
3 26
4
5
6 27
7
8 28
9
10 29
11
12
13 30
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Keywords: extracellular globins, tetra-domain, tandem duplication, evolution, annelids,

Branchipolynoe

Running head: Scale-worm tetra-domain globin evolution

31 Abstract

32 Hemoglobin is the most common respiratory pigment in annelids. They can be intra- or
33 extra-cellular, and this latter type can form large multimeric complexes. The hydrothermal
34 vent scale-worms *Branchipolynoe symmytilida* and *B. seepensis* express an extracellular tetra-
35 domain hemoglobin that is unique in annelids. We sequenced the gene for the single-domain
36 and tetra-domain globins in these two species. The single-domain gene codes for a mature
37 protein of 137 amino acids, and the tetra-domain gene codes for a mature protein of 552
38 amino acids. The single-domain gene has a typical 3 exons/2 introns structure, with introns
39 located at their typical positions (B12.2 and G7.0). This structure is repeated four times in the
40 tetra-domain gene, with no bridge introns or linker sequences between domains. The
41 phylogenetic position of *Branchipolynoe* globins among known annelid globins revealed that,
42 although extracellular, they cluster within the annelid intracellular globins clade, suggesting
43 that the extracellular state of these hemoglobins is the result of convergent evolution. The
44 tetra-domain structure likely resulted from two tandem duplications, domain 1 giving rise to
45 domain 2 and after this the two-domain gene duplicated to produce domains 3 and 4. The high
46 O₂ affinity of *Branchipolynoe* extracellular globins may be explained by the two key residues
47 (B10Y and E7Q) in the heme pocket in each of the domains of the single and tetra-domain
48 globins, which have been shown to be essential in the oxygen-avid hemoglobin from the
49 nematode *Ascaris suum*. This peculiar globin evolutionary path seems to be very different
50 from other annelid extracellular globins and is most likely the product of evolutionary
51 tinkering associated with the strong selective pressure to adapt to chronic hypoxia that
52 characterizes hydrothermal vents.

53 Introduction

54 Globins produced by invertebrates exhibit considerable heterogeneity in protein sequence
55 and in quaternary structure when compared to vertebrates. Despite this heterogeneity, these
56 proteins fold into the highly conserved globin fold that allows heme binding. The quaternary
57 structure diversity encompasses simple monomers, polymeric subunits made of single-
58 domains, and polymeric multi-domain subunits (see Weber and Vinogradov 2001 for a
59 review). Multi-domain globins can be intracellular, such as in the bivalve *Barbatia* (Grinich
60 and Terwilliger 1980, Suzuki and Arita 1995), or extracellular such as in crustaceans,
61 nematodes, and molluscs (see Weber and Vinogradov 2001 for a review). This type of
62 structure was not known in annelids until its discovery in two closely related hydrothermal
63 vent species of annelids, *Branchipolynoe symmytilida* and *B. seepensis*, which possess tetra-
64 domain globins (Hourdez et al. 1999a).

65 The presence of hemoglobin (Hb) in *Branchipolynoe* spp. may have an adaptive value in
66 their chronically hypoxic environment (Hourdez and Lallier 2007). *Branchipolynoe* belongs
67 to the Polynoidae, a family of scale-worms that is widely distributed in marine ecosystems
68 from the littoral to the deep-sea. They are very diverse at hydrothermal vents (Tunncliffe
69 1991) where they occupy all the microhabitats available to metazoa, from the coldest (and
70 most oxygenated), to the warmest (and usually the most hypoxic) waters. Scale-worms in
71 general (comprising several families, and including the Polynoidae) were known for having
72 only tissue globins (neuroglobin and myoglobin) (Weber 1978; Dewilde et al. 1996). In
73 contrast to their littoral relatives, most hydrothermal vent polynoid species have extracellular
74 Hbs giving them a red-pigmented coelomic fluid (Hourdez et al. 1999a; S.H. pers. obs.).
75 These Hbs can facilitate the diffusion of oxygen from the hypoxic environment, especially
76 given their high affinity for oxygen (Hourdez et al. 1999b), and represent a significant oxygen
77 storage for periods of complete anoxia (Hourdez and Weber 2005).

1
2
3 78 A phylogeny of the annelid globin sequences showed that the extracellular globins all
4
5 79 evolved from a duplicated intracellular ancestral globin, and several duplication events
6
7
8 80 followed to produce the present-day globin diversity found in the complex extracellular
9
10 81 hexagonal bilayer Hb (HBL-Hb; Bailly et al. 2007). The original split between the
11
12 82 intracellular and the extracellular globins occurred about 570 millions years ago in annelids
13
14
15 83 (Goodman et al. 1988).

16
17 84 The hydrothermal vent environment was colonized by scale-worms about 65 million years
18
19 85 ago (Hourdez et al. in prep.). All the vent-endemic subfamilies form a monophyletic group
20
21
22 86 that rapidly radiated after this initial colonization event. All of these species express Hbs
23
24 87 (single- or tetra-domain), suggesting that this trait was already present in the common vent
25
26 88 ancestor (Hourdez et al. in prep.). We sequenced both the single- and tetra-domain (hereafter
27
28
29 89 referred to SD, and TD, respectively) globin genes in *B. symmytilida* and *B. seepensis* to
30
31 90 examine their exon/intron structure and to understand the origin and evolution of the tetra-
32
33
34 91 domain globin.

92 Material and methods

93 **Animal collection**

94 *Branchipolynoe symmytilida* specimens were collected from the 9°50' N locality on the
95 East Pacific Rise in 2001 (9°46' N, 104°21' W, 2515 m depth), and *B. seepensis* were
96 collected from the Lucky Strike site in 2001 (Mid-Atlantic Ridge, 37°18' N, 32°16' W, 1700 m
97 depth). The worms were removed from the pallial cavity of their host mussels. The specimens
98 were identified and immediately frozen in liquid nitrogen on board, transported back to the
99 laboratory and stored at -80°C until analyzed.

100

101 **Nucleic acids extraction, and cDNA synthesis**

102 DNA, RNA extractions, and cDNA synthesis were performed with standard procedures
103 that are described in Supplementary materials.

104

105 **Globin cDNA and gene sequencing for *Branchipolynoe symmytilida***

106 - Protein microsequencing and PCR primer design

107 The two coelomic hemoglobins (HbC1 and HbC2) from *Branchipolynoe symmytilida* were
108 purified as described earlier (Hourdez et al. 1999a). We used Edman degradation for
109 microsequencing and the released amino acids were identified by HPLC (High Performance
110 Liquid Chromatography). The N-terminus in HbC1 was blocked but HbC2 yielded the
111 microsequence N-terminus VSAAQKAAIK. Based on this microsequence, degenerate
112 primers were designed to amplify HbC2 by PCR (Table S1, Supplementary material).

113 - Initial globin amplification and sequencing

114 Primers BSY_E1D1_C2F and an anchored oligo(dT) (Table S1) were used on the
115 synthesized cDNA, to amplify part of the tetra-domain globin gene. Detailed PCR conditions
116 and cloning procedures are given in Supplementary material. This initial PCR reaction

1
2
3 117 produced two insert size classes: 460-600 bp inserts that include one domain (400 bp) with
4
5 118 different sized 3'UTRs, and 900 bp inserts corresponding to two domains (800 bp) with
6
7 119 different sized 3'UTRs (See Results, Fig. S1). Two sequences representing the two size
8
9
10 120 classes of inserts were chosen because they were the most abundant among the clones,
11
12 121 designated BSY3 and BSY7. These sequenced clones were used to design specific primers to
13
14 122 amplify the coding sequence and introns of the corresponding genes (single-domain and tetra-
15
16 123 domain) (Table S1).

17
18
19
20 124 - Bridging oligonucleotide rapid amplification of cDNA ends (BO-RACEing) for the single-
21
22 125 domain globin

23
24 126 The 5'UTR for sequence BSY3 was amplified by BO-RACEing (Shi et al. 2002). Specific
25
26 127 primers and conditions are given in Supplementary material. This approach did not yield any
27
28 128 amplification for sequence BSY7 and chromosome walking was used instead to obtain the
29
30 129 missing portion of the sequence.

31
32
33
34 130 - Chromosome walking on genomic DNA

35
36 131 Sequencing of globin BSY7 from *B. symmytilida* was completed by directional genome
37
38 132 walking using PCR (Mishra et al., 2002). The detailed procedure is given in the
39
40 133 Supplementary material. Directional genome walking was also used to obtain the promoter
41
42 134 regions of both the TD and SD globins, in *B. symmytilida* and in *B. seepensis*.

43
44
45
46 135 - Northern blot

47
48 136 A standard northern blot protocol (Sambrook et al. 1989) was used for *Branchipolynoe*
49
50 137 *symmytilida*. The detailed procedure is given in the Supplementary material.

51
52 138

53
54
55 139 **Globin amplification and sequencing in *Branchipolynoe seepensis***

56
57 140 The amplification of the SD and TD globin genes (cDNA and gDNA) from *B. seepensis*
58
59 141 took advantage of the primers designed for the same genes in *B. symmytilida* and of an EST

1
2
3 142 library from which 2500 clones were sequenced (Hourdez and Tanguy, unpub. data).
4
5 143 Amplification and sequencing followed a standard procedure detailed in Supplementary
6
7 144 materials. Chromosome directional walking (see above) was used to sequence the 5'UTR and
8
9 145 the promoter region of the globin genes.
10
11
12
13
14

15 147 **Phylogenetic Analyses**

16
17 148 The sequences were assembled and nucleotide positions with conflicts were resolved based
18
19
20 149 on their chromatograms with CodonCode Aligner[®] 2.0.6
21
22 150 (<http://www.codoncode.com/aligner/index.htm>). Multiple nucleotide and amino acid sequence
23
24 151 alignments were performed by using ClustalX 2.0.10 (Larkin et al. 2007) and, when
25
26 152 necessary, manually optimized by using the sequence aligner editor Se-Align Carbon
27
28 153 (<http://tree.bio.ed.ac.uk/software/seal/>). During optimization, we minimized the number of
29
30 154 indels and nucleotide alignment was constrained by the amino acid sequences alignment.
31
32
33

34 155 Before all analyses, we verified that there was no saturation of the signal by plotting K2P
35
36 156 pairwise differences using all nucleotide sites against K2P pairwise differences estimated only
37
38 157 on the first two codon positions (data not shown).
39
40

41 158 jModelTest was used to choose a model of nucleotide substitution for use in phylogenetic
42
43 159 analysis that would best fit our data (Guindon and Gascuel 2003, Posada 2008). Neighbor-
44
45 160 joining (Saitou and Nei 1987) trees were constructed using MEGA4 (Tamura et al. 2007),
46
47 161 maximum-likelihood trees were computed using Phylip
48
49 162 (<http://evolution.genetics.washington.edu/phylip.html>) and Phyml Online
50
51 163 (<http://atgc.lirmm.fr/phyml/>, Guindon and Gascuel 2003, Guindon et al. 2005). Finally, a
52
53 164 Bayesian analysis was performed with Mr. Bayes (Ronquist and Huelsenbeck 2003,
54
55 165 Huelsenbeck and Ronquist 2001). PAML 3.14
56
57 166 (<http://abacus.gene.ucl.ac.uk/software/paml.html>, Yang 1997) was later used to test different
58
59
60

1
2
3 167 tree topologies issued from the different phylogenetic analyses, by maximum likelihood
4
5
6 168 relative ratio tests, and the codon substitution model derived from Goldman & Yang (1994).
7
8 169 The tree topologies were rooted only when a molecular clock hypothesis was being tested,
9
10 170 and under every test condition the single-domains were forced to be monophyletic. To test for
11
12 171 molecular clock and possible differences in selective pressures between paralogous domains
13
14
15 172 of the tetra-domain globin, we calculated pairwise distances between species for each domain,
16
17 173 using the K2P model for nucleotides and the Dayhoff matrix for amino acids (MEGA4).
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

174 Results

175 cDNA cloning and sequencing

176 Amplification of *B. symmytilida* cDNA with degenerate primers and oligo(dT) yielded six
177 distinct sequences that can be separated into two types (Fig. S1). The first type has a 400 bp
178 coding region, corresponding to a globin domain, followed by a 3'UTR sequence that differed
179 in length (sequences BSY1 and BSY3). The second type has a 800 bp coding region,
180 corresponding to two globin domains (sequences BSY2, BSY4, BSY5, and BSY7), followed
181 by a 3'UTR sequence also presenting differences in length. These two globin domains will be
182 hereafter referred to as D3 and D4 (from 5' to 3'), in reference to the *Branchipolynoe*
183 hemoglobin model composed of 4 domains (Hourdez et al. 1999a) (the same nomenclature
184 will be used for the upstream domains: D1 and D2). The 3'UTR for sequences BSY2, BSY4,
185 and BSY7 are nearly identical and a few differences appeared in their coding regions. The
186 three remaining 3'UTR sequences (corresponding to sequences BSY1, BSY3, and BSY5) are
187 clearly different. A Northern blot on *B. symmytilida* cDNA samples probed with a portion of
188 sequence BSY3 and BSY7 (the most common sequences among the clones) revealed that
189 these correspond to different transcript sizes. The BSY3 probe revealed a single band of ~610
190 bp in length, the size expected for a full-length mRNA coding for a single-domain globin
191 (Fig. S2A). The BSY7 probe revealed a single band of ~1980 bp in length, consistent with the
192 expected size for full-length mRNA coding for a tetra-domain globin (Fig. S2B).

193 The complete BSY3 cDNA sequence encodes a SD globin (coding sequence of 417 bp,
194 including the stop codon), with a 67-base 5'UTR (Fig. S3), and a 58-base 3'UTR. After
195 removal of the initial methionine, the deduced amino acid sequence (137 codons) would
196 produce a protein with a molecular weight (MW) of 13826.8 Da.

197 The complete BSY7 cDNA sequence has a 79-base 5'UTR, a 1674 bp coding sequence
198 (including the stop codon), and a 102-base 3'UTR (polyA tail not included). The mature

1
2
3 199 protein encoded corresponds to a 4-domain globin (552 codons) (Fig. S4), with a MW of
4
5 200 57691.9 Da, a mass that closely matches the value found for the native HbC2 subunit
6
7
8 201 (Hourdez et al, 1999a).
9

10 202 The cDNA of *B. seepensis* for both the SD and TD globins have the same length and
11
12 203 number of codons as the homologous sequences from *B. symmytilida* (Figs. S3 and S4). The
13
14 204 corresponding proteins have a MW of 13798.7 and 57742.9 Da for the SD and TD globins,
15
16 205 respectively.
17
18

19 206 **Protein primary structure**

20
21
22 207 The protein sequences for both *Branchipolynoe* species were aligned with other
23
24 208 intracellular and extracellular annelid globins, two nematode (*Ascaris suum*) globin
25
26 209 sequences, and the myoglobin from the sperm-whale *Physeter catodon* (Fig. 1). Only two
27
28 210 positions were invariant: the proximal histidine (F8H, i.e. helix F, position 8 with the
29
30 211 *Physeter* myoglobin sequence as a reference), and phenylalanine on the corner between
31
32 212 helices C and D (CD1F). A tryptophan in helix A (A12W) was conserved in all but the
33
34 213 nematode sequences. For all *B. symmytilida* and *B. seepensis* sequences, the distal histidine is
35
36 214 replaced by a glutamine (E7Q), a feature also observed in *Ascaris*. Similarly, the B-helix
37
38 215 tyrosine (B10Y), essential in modulating the oxygen affinity in *Ascaris* (De Baere et al.
39
40 216 1994), is also found in all *Branchipolynoe* spp. sequences. A tryptophan (H8W) is present in
41
42 217 all sequences except in *P. catodon*. The two conserved cysteines in extracellular globins that
43
44 218 form the characteristic intrachain disulfide bridge (A2C and H11C) are not found in any of
45
46 219 the *Branchipolynoe* sequences.
47
48

49
50 220 The TD sequences also have a cysteine (E18C) 11 residues after E7Q, a feature that is found
51
52 221 in some globins of the tubeworms *Riftia pachyptila* and *Lamellibrachia* sp. (from
53
54 222 hydrothermal vents and cold seeps, respectively). For all *Branchipolynoe* sequences, there is
55
56 223 no pre A helix, signal peptide, or linker sequences.
57
58
59
60

224

225 Origin of the globin gene and relationships with other annelid globins

226 For both *Branchipolynoe* species, polymorphism was present in the different amplified
227 sequences, therefore, a consensus nucleotide sequence was produced based on the majority of
228 clones obtained for each species. These consensus sequences were translated and used in all
229 phylogenetic analyses.

230 The unrooted phylogenetic tree clearly separates the globins that are typical extracellular
231 HBL-Hbs from all intracellular globins (Fig. 2). The SD globin and the four domains of the
232 TD globin from both *Branchipolynoe* species clearly group with the intracellular annelid
233 globins, indicating that these extracellular globins have a distinct origin from all other annelid
234 extracellular globins.

235

236 Evolution of the tetra-domain structure

237 A phylogeny of the various *Branchipolynoe* globin sequences was determined using the
238 SD as an outgroup because of its more ancestral status compared to the tetra-domain
239 sequences (Fig. 2). Several nucleotide substitution models were used, and when using Mr.
240 Bayes the codon model was also chosen. These analyses yielded different tree topologies
241 depending on the phylogenetic method used and the selected nucleotide substitution model
242 (Fig. S5).

243 The phylogenetic tree based on amino-acid sequences suggests that D4 is the most
244 ancestral of the domains followed by D3, D2, and D1 in the TD gene. Although this
245 phylogeny is well supported, it may not reflect the history of domain duplications as these can
246 be going through periods of relaxed selective pressure and/or positive selection. Analysis
247 based on nucleotides provided a different result, suggesting that D1 (or D3) is the most
248 ancestral domain. The best resolved trees correspond to 4 equiprobable topologies (Fig. S5).

1
2
3 249 In each tree topology, the globin domains of both species form well-supported pairs,
4
5 250 indicating that duplications giving birth to the TD occurred prior to the radiation between the
6
7
8 251 two *Branchiopolynoe* species. The deeper branches of the nucleotide-based topologies,
9
10 252 however, were not well supported (bootstrap values usually no higher than 65%, and at best
11
12 253 73%), suggesting that duplications occurred very close to each other over time, or that
13
14
15 254 variable selective pressure over the various branches reduced the phylogenetic signal.

16
17 255 The different topologies were tested against each other using maximum likelihood relative
18
19 256 ratio tests in CodeML (Yang 1997). Pairwise comparison of these 4 main topologies revealed
20
21 257 topology (c) was the most probable, regardless of the d_N/d_S ratio model selected. The pairwise
22
23 258 differences between the different TD domains and the SD indicated that D1 possesses the
24
25 259 fewest differences with SD, followed by D3 (Table 1). In addition, D1 also displayed several
26
27 260 motifs of adjacent amino acids that were nearly identical with the SD globin, reinforcing the
28
29 261 hypothesis that D1 is more closely related to the ancestral state. D1 and D3 were also the
30
31 262 domains with the smallest divergence between species for either nucleotide or amino acid
32
33 263 sequences, suggesting that these domains are more constrained by purifying selection. To
34
35 264 verify this hypothesis, we used BaseML (PAML) to test for a molecular clock using the
36
37 265 theoretical topology D1D3 vs D2D4. The results indicate that the molecular clock is rejected
38
39 266 in the evolution of the TD ($L_{\text{clock}=0} = -2171.78932$, $L_{\text{clock}=1} = -2190.984727$; $\text{RRT} = 2 * (L_{\text{clock}=0}$
40
41 267 $- L_{\text{clock}=1}) = 38.390814 \gg 10.83$ ($p = 0.001$). This supports the idea that D1 and D3 may be
42
43 268 under stronger selective constraints and evolving slower than the other domains.
44
45
46
47
48
49
50
51
52

53 270 **Promoters and 5'UTRs**

54
55 271 A portion of the promoter region was sequenced for the SD globin from *B. symmytilida*,
56
57 272 and for the TD globin for both species. For both genes, the TATA box is located 32 bp
58
59 273 upstream of the transcription start in *B. symmytilida*. For the SD sequence, only one base was

1
2
3 274 different for the portion of 5'UTR for which there are data for both species (48 nucleotides,
4
5 275 i.e. 97.9% identity) (Fig. S3). The promoter for the TD globin, as well as the 5'UTR, from
6
7
8 276 both species, also exhibit a high identity level (97.5% for the 5'UTR) (Fig. S4). A search for
9
10 277 transcription regulatory signal binding sites in *B. symmytilida* revealed a site for CP2, a factor
11
12 278 that is known to enhance the transcription of globin genes in erythroid cells (Chae and Kim
13
14
15 279 2003). This element was however not found in *B. seepensis*. Other universal transcription
16
17 280 factors, such as Oct-1, were identified in the promoter of both species (data not shown).
18
19 281 Although these are extracellular Hbs no signal sequence for secretion was found. In *B.*
20
21
22 282 *seepensis*, we sequenced a 700 bp fragment of the promoter. This region contains motifs for
23
24 283 the binding of two transcription factors relevant for this gene: one site for hypoxia inducible
25
26 284 factor 1 (HIF-1); and two sites for GATA-1, which plays an important role in erythroid
27
28
29 285 development (De Maria et al. 1999) and has been reported to enhance erythropoiesis in
30
31 286 response to tissue hypoxia (Krantz 1991, Zon et al. 1991). The sequenced promoter region for
32
33
34 287 *B. symmytilida* does not cover the region where these transcription factors' binding sites are
35
36 288 located in *B. seepensis*.

37
38
39 289

40 41 290 **Introns and gene structure**

42
43 291 In both the SD and TD globin genomic sequences from both species, the typical 3 exons/2
44
45 292 introns globin gene structure was present, and this motif is repeated four times in the TD gene
46
47
48 293 (Fig. 3). For all the genes, the introns are located at positions B12.2 and G7.0, corresponding
49
50 294 to the typical globin introns positions. In the SD globin gene, the second intron exhibits a very
51
52 295 high level of identity between the two species (94.4%). Although this is also true for the first
53
54
55 296 160 bp in intron 1, the identity then drops sharply because *B. seepensis* possesses a large (440
56
57
58 297 bp) insertion (Table S2). In the TD globin genomic sequence, each intron was usually 300 bp
59
60 298 long, with one exception at 700 bp again for the first intron of D1 (Table S2). Interestingly,

1
2
3 299 neither introns nor linker segments separate the four globin domains (Fig. S4). Pairwise intron
4
5 300 sequence comparisons between the two species revealed a very high level of conservation (on
6
7
8 301 average 90% of identity) of all but the second intron of D4. The major differences in the
9
10 302 orthologous introns correspond to indels of varying size, from 3 to 41 bp. Comparisons of
11
12 303 intron sequences among domains failed to reveal significant sequence similarity. In intron 2
13
14
15 304 from D1 and D2 from *B. seepensis*, there is a regulatory signal for GATA-1 and HIF-1,
16
17 305 respectively.
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

306 Discussion

307 The coding sequence for the unique TD and SD globins genes, as well as the position and
308 size of the introns have been determined to understand the origin and evolution of these
309 respiratory pigments. The TD globin with a multi-domain gene structure represents the
310 emergence of a new type of hemoglobin in annelids (Terwilliger 1992).

312 *Clarification of the protein structure*

313 Hourdez et al. (1999a) showed that the *Branchipolynoe* extracellular TD globins form
314 dimers (HbC2) or trimers (HbC1) and our results suggest that we most likely amplified the
315 former because the estimated MW most closely matches the one for HbC2 subunits. Large
316 polypeptidic assemblages are not unusual in invertebrate respiratory pigments (Terwilliger,
317 1998). In all cases, the formation of large polypeptide complexes has been attributed to the
318 necessity to avoid their excretion (Weber and Vinogradov, 2001). In *Daphnia* and *Artemia*,
319 the didomain and nine-domain Hbs, respectively, can form homo- or hetero-dimers (Dewilde
320 et al. 1999, Mansfield et al. 1980). In the nematode *A. suum* the didomain Hbs assemble into
321 octamers of 328 kDa, through a C-terminus extension that forms a β -strand and each strand
322 binds to each other with H-bonds in a zipper-like structure (Darawshe et al. 1987, De Baere et
323 al. 1992). In *Branchipolynoe*, there is no such extension and there are no interchain disulfide
324 bridges to assemble into trimers or dimers, and only inter-subunit non-covalent bonds are
325 involved.

326 Based on subtilisin partial digestion patterns and heme content determination, Hourdez et
327 al. (1999a) suggested that each Hb subunit was composed of 4 globin domains (with one
328 likely truncated) each with one heme group. The sequences we obtained clearly show that
329 there are indeed 4 domains but each of these domains is full-length. This may indicate that the

1
2
3 330 interdomain regions are not well exposed in the native Hb, and susceptible to peptidic
4
5 331 cleavage by a wide-spectrum peptidase.
6
7

8 332 Interestingly, the globins from *Branchipolynoe* do not possess the two cysteine residues
9
10 333 necessary to form the conserved disulfide bridge conserved in all other extracellular annelid
11
12 334 globins (Bailly et al., 2007). This indicates that this character is not necessary for the globins
13
14 335 to function outside of cells, although its conservation in the lineage of the HBL-Hb globins
15
16 336 suggests that purifying selection is constraining these amino acid positions.
17
18
19

20 337

21
22 338 ***Structure/function relationship***
23

24 339 The amino acid alignment revealed the presence of a glutamine instead of a distal histidine
25
26 340 (E7H). Although this amino acid is usually conserved in vertebrates, some invertebrate
27
28 341 species use other amino acids. In the nematode *A. suum* Hb, known to have one of the highest
29
30 342 oxygen affinities among invertebrates (see Weber and Vinogradov 2001 for a review), the
31
32 343 position is also occupied by a glutamine (E7Q). This high affinity can be explained in part by
33
34 344 the interaction of the E7Q with a tyrosine in position B10Y (De Baere et al. 1994, Yang et al.
35
36 345 1995). Interestingly, *Branchipolynoe* also has a tyrosine in position B10. This could explain
37
38 346 the high affinity for oxygen that was found in *B. symmytilida* Hbs (Hourdez et al. 1999b).
39
40
41
42

43 347 In the TD globins, but not in the SD ones, there is a cysteine located 11 residues after the
44
45 348 E7Q (E18C). This cysteine is known to be under strong selection in some annelids living in
46
47 349 reduced habitats (Bailly et al. 2003), suggesting a key function for E18C in these habitats. It
48
49 350 was hypothesized to be a key residue for reversible sulfide binding, a characteristic that is
50
51 351 essential for vestimentiferan tubeworms that live symbiotically with sulfide-oxidizing bacteria
52
53 352 (Zal et al. 1998). It was later shown that the E18 cysteine was not likely the binding site for
54
55 353 transporting sulfide in *Riftia pachyptila* and that, at least for the 400-kDa Hb, zinc atoms were
56
57 354 involved (Flores et al. 2005, Flores and Hourdez 2006). The SD globin from *Branchipolynoe*
58
59
60

1
2
3 355 is devoid of this specific residue, indicating that it most likely represents an adaptive
4
5 356 convergence to life in a sulfide-rich habitat in the TDs. Its presence in the heme pocket may
6
7 357 have a protective role for the heme group that usually reacts with sulfide to irreversibly form
8
9 358 sulfhemoglobin, an altered molecule incapable of binding oxygen.
10
11
12
13 359

14 360 ***Secretion into the coelomic fluid***

15 361 Although *Branchipolynoe* SD and TD Hbs are extracellular (secreted in the coelomic
16 362 cavity), their genes do not possess a signal peptide for secretion whereas all other annelid
17 363 extracellular globins have such a signal (Riggs 1991; Bailly et al. 2007). The absence of a
18 364 signal peptide in all *Branchipolynoe* globins could be due to the fact that we did not obtain the
19 365 whole sequence and that the upstream sequence (promoter) actually corresponds to an intron.
20 366 This however is unlikely as we did locate the TATA box and the sequence corresponding to
21 367 the 5'UTR, and a 700 bp stretch upstream did not reveal any other open reading frame in *B.*
22 368 *seepensis* TD globin gene. This may indicate that the secretion of *Branchipolynoe* Hbs is
23 369 holocrine (by rupture of the cell membrane, and the release of all the components of the
24 370 cytoplasm). This is supported by the fact that, in at least one other vent species
25 371 (*Lepidonotopodium piscesae*), the SD globin expressed in the muscles (myoglobin) has
26 372 exactly the same molecular weight as the one found in the coelomic cavity (Hourdez, unpub.
27 373 data).
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

50 375 ***Origin and evolution of the tetra-domain globin***

51 376 Phylogenetic relationships among annelid globins indicate that extracellular globins have
52 377 evolved independently from the intracellular (circulating or non-circulating) ones, and
53 378 diverged about 570 millions years ago (Goodman et al. 1988), an origin distinct from
54 379 extracellular globins found in other phyla (Bailly et al. 2007). *Branchipolynoe* extracellular

1
2
3 380 globin sequences have a distinct history, not only from other phyla's extracellular globins
4
5 381 (including multidomain globins), but also from the typical extracellular annelid globins. It
6
7
8 382 appears that they arose from an intracellular annelid globin more recently than the original
9
10 383 split between intra and extracellular globins in annelids. Despite this different evolutionary
11
12 384 history, the SD and TD globins conserved the basic globin gene structure of 3 exons and 2
13
14 385 introns, including intron positions. This arrangement differs from that of *Aphrodite aculeata*
15
16 386 myoglobin (Dewilde et al. 1996) in which the first intron is missing. This species belongs to
17
18 387 the scale-worm family Aphroditidae, a close related family to the Polynoidae. This indicates
19
20 388 that the ancestor to all scale-worms most likely did possess this first (B12.2) intron, and that it
21
22 389 was later lost in the lineage leading to *Aphrodite aculeata*. Some of the introns showed a high
23
24 390 level of conservation between the two species, but in both the SD and the TD globin genes
25
26 391 there is a strongly divergent intron (intron 1 in the SD, intron 2 in D4 for the TD). This may
27
28 392 indicate strong selective constraint on the intron sequence for a possible role related to a
29
30 393 regulatory function.

31
32
33
34
35
36 394 The phylogenetic analyses yielded trees with only limited support for the internal branches.
37
38 395 This could be indicative of a rapid succession of the duplication events to lead to this
39
40 396 structure, and/or to different selective pressure affecting the various domains. We found that
41
42 397 D1 and D3 are probably under stronger purifying selection than D2 and D4. This would then
43
44 398 yield to different branch lengths, and the absence of a molecular clock. Although the
45
46 399 phylogenetic trees do not allow us to clearly decide on a likely duplication scenario, other
47
48 400 observations (pairwise distances, conservation of some amino acid motifs), suggest that the
49
50 401 TD structure is likely the result of two tandem duplications, the first one giving rise to D1D2,
51
52 402 and the second one affecting these two domains together to generate the whole TD. Domain
53
54 403 D1 is likely the most ancestral domain, as (1) in both species, intron 1 from D1 has the same
55
56 404 size as in the SD globin, (2) pairwise differences between the SD and D1 are smaller than all

1
2
3 405 the other domains, and (3) there are some conserved amino acids motifs. This duplication
4
5 406 scenario shows that duplication moved in a 5' to 3' direction where downstream duplicates
6
7
8 407 were likely joined by intron losses. The alternative scenario where domain 4 is ancestral, and
9
10 408 duplications proceeded in a 3' to 5' direction seems less likely because with each subsequent
11
12 409 domain duplication the 5' regulatory region may have had to be reconstructed.
13
14
15 410

17 411 ***Duplication events***

19
20 412 The mechanism responsible for these duplication events remains unclear for the moment.
21
22 413 Interestingly, the TD globin gene from these scale-worms does not have any bridge introns.
23
24 414 This contrasts with the didomain globins from the bivalve *Barbatia reeveana* (Naito et al.
25
26 415 1991), the water-flea *Daphnia pulex* (Dewilde et al. 1999), and some nematodes (*Ascaris*
27
28 416 *suum* and *Pseudoterranova decipiens*; Dixon et al. 1992, Sherman et al. 1992). Similarly, in
29
30 417 mollusk hemocyanins, the 7 or 8 domains are separated by bridge introns that are thought to
31
32 418 be the remnants of the duplication of the gene (Lieb and Todt 2008). Amino acid linker
33
34 419 sequences are also missing between the domains of the TD globins from *Branchipolynoe*.
35
36 420 This kind of structure is known in other invertebrate multidomain proteins such as the nine-
37
38 421 domain Hb from *Artemia* (Maning et al. 1990, Trotman et al. 1994), and the didomain Hb of
39
40 422 the bivalve *B. reeveana* (Naito et al. 1991). It is however not found in the nematodes two-
41
42 423 domain Hb (De Baere et al. 1992, Sherman et al. 1992). These linker sequences are thought to
43
44 424 be the remnants of an inter-domain intron that lost its splicing signals and was eventually
45
46 425 integrated into the coding sequence. All *Branchipolynoe* TD domains are full-length,
47
48 426 indicating that there likely was inter-domain introns that were secondarily integrally lost, as
49
50 427 suggested by Naito et al. (1991) for *B. reeveana*, Dewilde et al. (1999) for *D. pulex* Hb, and
51
52 428 Dixon et al. (1992) for *P. decepiens*. The alternative, less likely, hypothesis would require two
53
54 429 end-to-end tandem duplications that preserved (1) the whole length of the domains, and (2)

1
2
3 430 the reading frame. In another water-flea, *Moina macrocopa*, the two-domain globin also lacks
4
5 431 the bridge intron (Kato et al. 2001), and the authors suggest unequal crossing-over as the
6
7
8 432 mechanism for the duplication of the gene, with a subsequent loss of the bridge intron when
9
10 433 the Cladocera (water-fleas) families diverged. Finally, the TD structure could also correspond
11
12 434 to the result of unequal crossing-overs between two copies of an initial didomain structure but
13
14
15 435 this would yield chimeric domains and our search for such possible chimeras proved
16
17 436 unsuccessful (data not shown).

18
19
20 437

21 22 438 **Concluding remarks**

23
24 439 Our work shows that *Branchipolynoe* extracellular Hbs have a different origin and
25
26 440 evolutionary history from other annelid extracellular Hbs. It appears that an ancestral
27
28 441 intracellular myoglobin was duplicated and evolved to form an extracellular globin by
29
30 442 evolutive tinkering. The evolution of these Hbs, constitutively expressed at high levels, is
31
32 443 most likely the result of strong selective pressure due to the chronic hypoxia that characterizes
33
34 444 hydrothermal vents (Hourdez and Lallier, 2007). The multidomain structure allows higher
35
36 445 concentration of oxygen binding sites (and thereby transport/storage capacity), without
37
38 446 increasing the colloid osmotic pressure of the coelomic fluid.

39
40
41
42
43 447

44 448 **Acknowledgments**

45
46 449 The authors would like to thank the crews of the ships and submersibles, as well as the chief
47
48 450 scientists, of the cruises ATOS 2001 (project funded by Ifremer and INSU), and EPR 2001
49
50 451 (project funded by a NSF grant to C.R. Fisher (NSF OCE-0002729)). We would also like to
51
52 452 thank Isabelle Boutet-Tanguy and Arnaud Tanguy for technical advice in lab. This work is
53
54 453 part of the project HYPOXEVO (Région Bretagne), and was supported by the ESTeam
55
56 454 research Marie Curie grant under the 6th framework program from the European Commission.

1
2
3 455 Supplementary material
4
5

6 456 **Materials and methods**
7

8 457 Nucleic acids extraction, and cDNA synthesis
9

10 458 Initial *B. symmytilida* globin amplification and sequencing
11

12 459 Bridging oligonucleotide rapid amplification of cDNA ends (BO-RACEing) for the single-
13

14 domain globin
15

16 460 Northern-blot on Branchipolynoe symmytilida RNA
17

18 461 Directional chromosome walking
19

20 462 *B. seepensis* globin sequencing
21

22 463
23
24 464

25
26
27 465 Table S1, Table S2
28

29 466 Figures S1, S2, S3, S4, and S5
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 467 Literature cited
4

5
6 468 Bailly X, Leroy R, Carney S, Collin O, Zal F, Toulmond A, Jollivet D. 2003. The loss of the
7
8 469 hemoglobin H₂S-binding function in annelids from sulfide-free habitats reveals molecular
9
10 470 adaptation driven by Darwinian positive selection. P Natl Acad Sci USA. 100:5885-5890.
11

12 471
13
14
15 472 Bailly X, Chabasse C, Hourdez S, Dewilde S, Moens L, and Zal F. 2007. Globin gene family
16
17 473 evolution and functional diversification in annelids. FEBS J. 274: 2641-2652.
18

19 474
20
21
22 475 Chae JH, Kim CG. 2003. CP2 Binding to the Promoter is Essential for the Enhanced Transcription of
23
24 476 Globin Genes in Erythroid Cells. Mol Cells. 15: 40-47.
25

26 477
27
28
29 478 Church GM, Gilbert W. 1984. Genomic sequencing. P Natl Acad Sci USA 81:1991-1995.
30

31 479
32
33
34 480 Darawshe S, Tsafadyah Y, Daniel E. 1987. Quaternary structure of erythrocrucorin from the nematode
35
36 481 *Ascaris suum*. Biochem J. 242: 689-694.
37

38 482
39
40
41 483 De Baere I, Liu L, Moens L, Van Beeumen J, Gielens C, Richelle J, Trotman C, Finch J, Gerstein M,
42
43 484 Perutz M. 1992. Polar zipper sequence in the high-affinity hemoglobin of *Ascaris suum*:
44
45 485 Amino acid sequence and structural interpretation. P Natl Acad Sci USA. 89:4638-4642.
46

47 486
48
49
50
51 487 De Baere I, Perutz MF, Kiger L, Marden MC, Poyart C. 1994. Formation of two hydrogen bonds
52
53 488 from the globin to the heme-linked oxygen molecule in *Ascaris* hemoglobin. P Natl Acad Sci
54
55 489 USA. 91:1594-1597.
56

57 490
58
59
60 491 De Maria R, Zeuner A, Eramo A, Domenichelli C, Bonci D, Grignani F, Srinivasula SM, Alnemri

- 1
2
3 492 ES, Testa U, Peschle C. 1999. Negative regulation of erythropoiesis by caspase-mediated
4
5 493 cleavage of GATA-1. *Nature*. 401:489-493.
6
7 494
8
9
10 495 Dewilde S, Blaxter M, Van Hauwaert M-L, Vanfleteren J, Esmans EL, Marden M, Griffon N, Moens
11
12 496 L. 1996. Globin and Globin Structure of the Nerve Myoglobin of *Aphrodite aculeata*. *J Biol*
13
14 497 *Chem*. 271:19865-19870.
15
16
17 498
18
19 499 Dewilde S, Van Hauwaert M-L, Peeters K, Vanfleteren J, Moens L. 1999. *Daphnia pulex* didomain
20
21 500 hemoglobin: structure and evolution of polymeric hemoglobins and their coding genes. *Mol*
22
23 501 *Biol Evol*. 16:1208-1218.
24
25
26 502
27
28 503 Dixon B, Walker B, Kimmins W, Pohajdak B. 1992. A nematode gene contains an intron previously
29
30 504 thought to be unique to plants. *J Mol Evol*. 35:131-136.
31
32
33 505
34
35 506 Flores JF, Fisher CR, Carney SL, Green BN, Freytag JK, Schaeffer SW, Royer Jr WE. 2005. Sulfide
36
37 507 binding is mediated by zinc ions discovered in the crystal structure of a hydrothermal vent
38
39 508 tubeworm hemoglobin. *P Natl Acad Sci USA*. 102:2713-2718.
40
41
42 509
43
44 510 Flores JF, Hourdez S. 2006. The zinc-mediated sulfide-binding mechanism of hydrothermal vent
45
46 511 tubeworm 400-kDa hemoglobin. *Cah Biol Mar*. 47: 371-377.
47
48
49 512
50
51 513 Goldman N, Yang Z. 1994 A codon based model of nucleotide substitution for protein-coding
52
53 514 DNA sequences. *Mol Biol Evol*. 11:725-736
54
55
56 515
57
58 516 Goodman M, Pedwaydon J, Czelusniak J, Suzuki T, Gotoh T, Moens L, Shishikura F, Walz D,
59
60

- 1
2
3 517 Vinogradov SN. 1988. An evolutionary tree for invertebrate globin sequences. *J Mol Evol.*
4
5 518 27:236-249.
6
7
8 519
9
10 520 Grinich N P, Terwilliger RC. 1980. The quaternary structure of an unusual high-molecular-weight
11
12 521 intracellular hemoglobin from the bivalve mollusc *Barbatia reeveana*. *Biochem J.* 189: 1-8.
13
14 522
15
16
17 523 Guindon S, Gascuel O. 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies
18
19 524 by Maximum Likelihood. *Syst Biol* 52: 696-704.
20
21 525
22
23
24 526 Guindon S, Lethiec F, Duroux P, Gascuel O. 2005. PHYML Online-a web server for fast maximum
25
26 527 likelihood-based phylogenetic inference. *Nucl. Acids Res.* 33:W557-W559.
28
29 528
30
31 529 Hourdez S, Lallier FH, Green BN, Toulmond A. 1999a. Hemoglobins from deep-sea hydrothermal
32
33 530 vent scale-worms of the genus *Branchiopolynoe*: A new type of quaternary structure. *Proteins.*
34
35 531 34:427-434.
36
37 532
38
39
40 533 Hourdez S, Lallier FH, Martin-Jézéquel V, Weber RE, Toulmond A. 1999b. Characterization and
41
42 534 functional properties of the extracellular coelomic hemoglobins from the deep-sea,
43
44 535 hydrothermal vent scale-worm *Branchiopolynoe symmytilida*. *Proteins.* 34:435-442.
45
46 536
47
48
49 537 Hourdez S, Lallier FH. 2007. Adaptations to hypoxia in hydrothermal vent and cold-seep
50
51 538 invertebrates. *Rev Env Sci Biotech.* 6: 143-159.
52
53 539
54
55
56 540 Hourdez S, Weber RE. 2005. Molecular and functional adaptations in deep-sea hemoglobins. *J Inorg*
57
58 541 *Biochem.* 99(1): 130-141
59
60

- 1
2
3 542
4
5
6 543 Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees.
7
8 544 Bioinformatics. 17: 754-755.
9
10 545
11
12 546 Kato K, Tokishita S-i, Mandokoro Y, Kimura S, Ohta T, Kobayashi M, Yamagata H. 2001. Two-
13
14 547 domain hemoglobin gene of the water flea *Moina macrocopa*: duplication in the ancestral
15
16 548 Cladocera, diversification, and loss of a bridge intron. *Gene*. 273:41-50.
17
18 549
19
20 550 Krantz SB. 1991. Erythropoietin. *Blood*. 77(3): 419-434.
21
22 551
23
24 552 Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F,
25
26 553 Wallace IM, Wilm A, Lopez R, et al. (3 co-authors). 2007. Clustal W and Clustal X version
27
28 554 2.0. *Bioinformatics* 23:2947-2948.
29
30 555
31
32 556 Lieb B, Todt C. 2008. Hemocyanin in mollusks - A molecular survey and new data on hemocyanin
33
34 557 genes in *Solenogastres* and *Caudofoveata*. *Mol Phylogenet Evol*. 49:382-385.
35
36 558
37
38 559 Manning A, Trotman C, Tate W. 1990. Evolution of a polymeric globin in the brine shrimp *Artemia*.
39
40 560 *Nature*. 348: 653-656.
41
42 561
43
44 562 Mansfield BC, Krissansen GW, Smith MG, Tate WP. 1980. Isolation of intact larval haemoglobin
45
46 563 from the brine shrimp *Artemia salina*. Prevention of degradation in vitro by proteases induced
47
48 564 during larval development. *Biochem Biophys Acta*. 623:163-170.
49
50 565
51
52 566 Mishra RN, Singla-Pareek SL, Nair S, Sopory SK, Reddy MK. 2002. Directional Genome Walking
53
54
55
56
57
58
59
60

- 1
2
3 567 Using PCR. *BioTechniques*. 33:830- 834.
4
5
6 568
7
8 569 Naito Y, Riggs C, Vandergon T, Riggs A. 1991. Origin of a "bridge" intron in the gene for two
9
10 570 domain globin. *P Natl Acad Sci USA*. 88: 6672-6676.
11
12
13 571
14
15 572 Posada D. 2008. *jModelTest: Phylogenetic Averaging*. *Mol Biol Evol*. 25:1253-1256.
16
17 573
18
19
20 574 Riggs AF. 1991. Aspects of the Origin and Evolution of Non-Vertebrate Hemoglobins. *Am Zool*. 31:
21
22 575 535-545.
23
24
25 576
26
27 577 Rodionov MA, Johnson MS. 1994. Residue-residue contact substitution probabilities derived from
28
29 578 aligned three-dimensional structures and the identification of common folds. *Protein Sci*. 3: 2366-
30
31 579 2377.
32
33
34 580
35
36 581 Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed
37
38 582 models. *Bioinformatics*. 19: 1572-1574.
39
40
41 583
42
43 584 Saitou N, Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic
44
45 585 trees. *Mol Biol Evol*. 4:406-425.
46
47
48 586
49
50 587 Sambrook J, Fritsch EF, Maniatis T. 1989. *Molecular cloning: a laboratory manual*. Nolan C. ed.
51
52 588 Cold Spring Harbour :Cold Springs Harbor Press.
53
54
55 589
56
57
58 590 Sherman DR, Kloek AP, Krishnan BR, Guinn B. 1992. *Ascaris* hemoglobin gene: Plant-like structure
59
60 591 reflects the ancestral globin gene. *P Natl Acad Sci USA*. 89:11696-11700.

- 1
2
3 592
4
5
6 593 Shi X, Karkut T, Chahmanhkah M, Alting-Mees M, Hemmingsen SM, Hegedus D. 2002. 5'-
7
8 594 RACEing Across a Bridging Oligonucleotide. *Biotechniques*. 32:480-482.
9
10 595
11
12 596 Suzuki T, Arita T. 1995. Two-Domain Hemoglobin from the Blood Clam, *Barbatia lima*. The cDNA-
13
14 597 Derived Amino Acid Sequence. *J Protein Chem*. 14(7): 499-502.
15
16 598
17
18
19 599 Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis
20
21 600 (MEGA) software version 4.0. *Mol Biol Evol*. 24: 1596-1599.
22
23 601
24
25
26 602 Terwilliger NB. 1992. Molecular Structure of the Extracellular Heme Proteins Vol 13. In: Mangum
27
28 603 CP editor. *Advances in Comparative and Environmental Physiology. Blood and Tissue*
29
30 604 *Oxygen Carriers*, Berlin Heidelberg: Springer-Verlag. 459p.
31
32 605
33
34
35 606 Terwilliger NB. 1998. Functional adaptations of oxygen-transport proteins. *J Exp Biol*. 201:1085-
36
37 607 1098.
38
39 608
40
41
42 609 Trotman C, Manning A, Bray J, Jellie A., Moens L., Tate W. 1994. Interdomain Linkage in the
43
44 610 Polymeric Hemoglobin Molecule of *Artemia*. *J Mol Evol*. 38:628-636.
45
46 611
47
48
49 612 Tunncliffe V. 1991. The Biology of Hydrothermal Vents: Ecology and Evolution. *Oceanogr Mar*
50
51 613 *Biol*. 29:319-407.
52
53 614
54
55
56 615 Weber RE, Vinogradov SN. 2001. Nonvertebrate hemoglobins: functions and molecular adaptations.
57
58 616 *Physiol Rev*. 81:569-628.
59
60

- 1
2
3 617
4
5
6 618 Weber RE. 1978. Respiratory Pigments. Physiology of Annelids. P. J. Mill. London, Academic Press
7
8 619 Inc.
9
10 620
11
12 621 Yang J, Kloek A., Goldberg D, Mathews F. 1995. The structure of *Ascaris* hemoglobin domain I at
13
14 622 2.2 Å resolution: molecular features of oxygen avidity. P Natl Acad Sci USA. 92:4224-4228.
15
16 623
17
18
19 624 Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood.
20
21 625 Comput App BioSci. 13: 555-556.
22
23 626
24
25
26 627 Zal F, Lallier FH, Green BN, Vinogradov SN, Toulmond A. 1996. The Multi-hemoglobin System of
28
29 628 the Hydrothermal Vent Tube Worm *Riftia pachyptila*. II. Complete polypeptide chain
30
31 629 composition investigated by maximum entropy analysis of mass spectra. J Biol Chem.
32
33 630 271:8875-8881.
34
35 631
36
37
38 632 Zal F, Leize E, Lallier FH, Toulmond A, Van Dorsselaer A, Childress J. 1998. S-Sulfohemoglobin
39
40 633 and disulfide exchange: The mechanisms of sulfide binding by *Riftia pachyptila* hemoglobins.
41
42 634 P Natl Acad Sci USA. 95:8997-9002.
43
44 635
45
46
47 636 Zon LI, Youssoufian H, Mather C, Lodish HF, Orkin S. 1991. Activation of the erythropoietin
48
49 637 receptor promoter by transcription factor GATA-1. P Natl Acad Sci USA. 88:10638-10641.
50
51
52
53
54
55
56
57
58
59
60

638 Table

639

640 Table 1. Pairwise differences between orthologous domains and the SD. Nucleotide differences
 641 were calculated by the Kimura-2 parameter model and amino acid differences based on the
 642 Dayhoff index matrix. BSY: *Branchipolynoe symmytilida*, BSE: *B. seepensis*, D: domain, SD:
 643 single-domain. Shaded values represent the lowest values between the SD and all the other
 644 domains.

645

BSY/BSE	Nucleotides (K2P)	Amino Acids (Dayhoff matrix)
D1/D1	0.030	0.022
D2/D2	0.035	0.058
D3/D3	0.025	0.022
D4/D4	0.065	0.058
SD/SD	0.017	0.007
SD/D1	0.458	0.592
SD/D2	0.506	0.725
SD/D3	0.485	0.653
SD/D4	0.564	0.689
D1/SD	0.462	0.549
D2/SD	0.511	0.713
D3/SD	0.480	0.625
D4/SD	0.554	0.657

1
2
3 646 **Figures captions**

4
5 647 Figure S1. Schematic structure of the six initial cDNA sequences obtained for *B. symmytilida*.
6
7 648 The coding sequence is shown in white and the 3'UTR in grey. The length of each part of the
8
9 649 sequences is shown. BSY: *Branchipolynoe symmytilida*

10 650
11
12 651 Figure S2. Northern-Blot results using probes from the sequences BSY3 (A) and BSY7 (B). The size
13
14 652 of the obtained bands is indicated (the molecular weight marker is not visible). Lane 1: 20 µg of total
15
16 653 RNA; Lane 2: 1 µg of poly A + mRNA; Lane 3: 3 µg of poly A + mRNA.

17 654
18
19 655 Figure S3. cDNA sequence of the single-domain globin of *B. symmytilida* (BsySD) and *B.*
20
21 656 *seepensis* (BseSD). The 5'UTR sequence and STOP codon are in lower-case. TATA box
22
23 657 highlighted. Intron positions indicated by vertical lines. 3'UTR not shown.

24 658
25
26 659 Figure S4. cDNA sequence of the tetra-domain globins of *B. symmytilida* (BsyTD) and *B.*
27
28 660 *seepensis* (BseTD). The 5'UTR and STOP codon are in lower-case. The TATA promoter binding
29
30 661 site is highlighted. Domains separated by vertical lines. The 3'UTR not shown.

31 662
32
33 663 Figure S5. Different topologies obtained from the various phylogeny softwares and used in
34
35 664 the PAML analysis. All trees are unrooted. Used softwares are indicated (the method and
36
37 665 nucleotide substitution model are shown between parentheses). The bootstrap or posterior
38
39 666 probability (Mr. Bayes) values are shown above/under the branches and the nucleotide
40
41 667 substitution model to which they correspond is underlined. (a) to (c): topologies based on
42
43 668 nucleotide sequences, (d): topology based on amino acid sequences. Bsy: *B. symmytilida*; Bse:
44
45 669 *B. seepensis*; TD: tetra-domain; SD: single-domain; D1-D4: domains 1 through 4.

46 670
47
48 671 Figure 1. Globin sequences from annelids, nematodes and a vertebrate. *Branchipolynoe*
49
50 672 globin sequences are shaded, TD globins shaded in light grey and SD globins in dark grey.
51
52 673 Conserved residues shown in bold (CD1F and F8H), heme pocket residues that may explain
53
54 674 high O₂ affinity in *Branchipolynoe* are boxed. Cysteines forming an intrachain disulfide
55
56 675 bridge in typical extracellular annelid globins (A2C and H11C) underlined. Cysteine E18
57
58 676 underlined in the TD globins from *Branchipolynoe* spp., and for *Riftia* and *Lamellibrachia*.
59
60 677 Intron (I1 and I2) conserved positions shown above the sequences. Bsy: *B. symmytilida*; Bse:
678 *B. seepensis*; TD: tetra-domain; SD: single-domain; D1-D4: domains 1 through 4; AacuNg:

1
2
3 679 *Aphrodite aculeata* neuroglobin; Gly: *Glycera* sp.; Lumt: *Lumbricus terrestris*; Tubifex:
4
5 680 *Tubifex tubifex*; Phese: *Pheretima seiboldi*; Tylo: *Tylorhynchus heterochaetus*; Rifb: *Riftia*
6
7 681 *pachyptila*; Lam: *Lamellibrachia* sp.; Asuum: *Ascaris suum*; Phyca: *Physeter catodon*.

8
9 682
10
11 683 Figure 2. Bayesian phylogenetic tree based on annelid globins corresponding to the alignment
12
13 684 in Fig. 1. White bubble: extracellular globins; light-grey bubble: *Branchipolynoe* globins
14
15 685 (single-domain: SD and tetra-domain: TD); dark-grey bubble: annelid intracellular globins.
16
17 686 Posterior probability values are indicated above the branches. See Fig. 1 for abbreviations and
18
19 687 accession numbers.

20
21 688
22
23 689 Figure 3. Gene structure from the single- (A) and tetra-domain (B) globins from
24
25 690 *Branchipolynoe* spp. Promoter region: dark grey; UTR: medium grey; exons: light grey;
26
27 691 introns: dark grey with "I" or "Intron". Gene sizes and components are not to scale.

28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

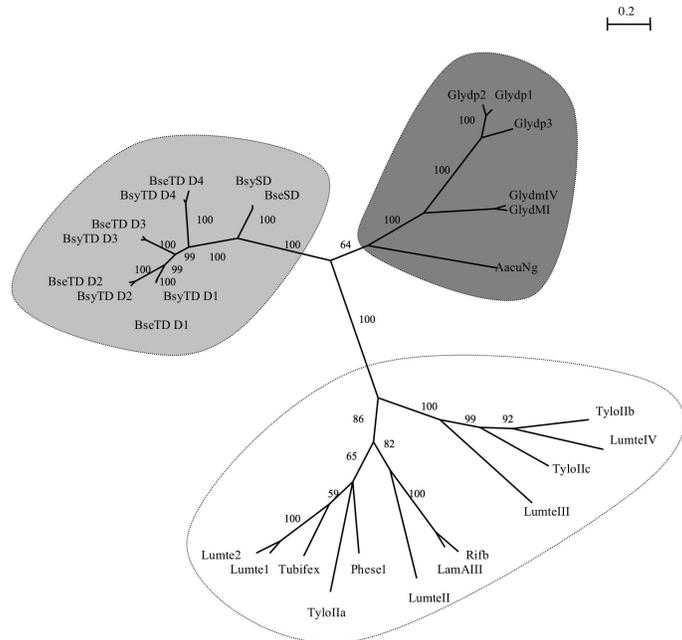


Figure 2. Bayesian phylogenetic tree based on annelid globins corresponding to the alignment in Fig. 1. White bubble: extracellular globins; light-grey bubble: Branchipolynoe globins (single-domain: SD and tetra-domain: TD); dark-grey bubble: annelid intracellular globins. Posterior probability values are indicated above the branches. See Fig. 1 for abbreviations and accession numbers.

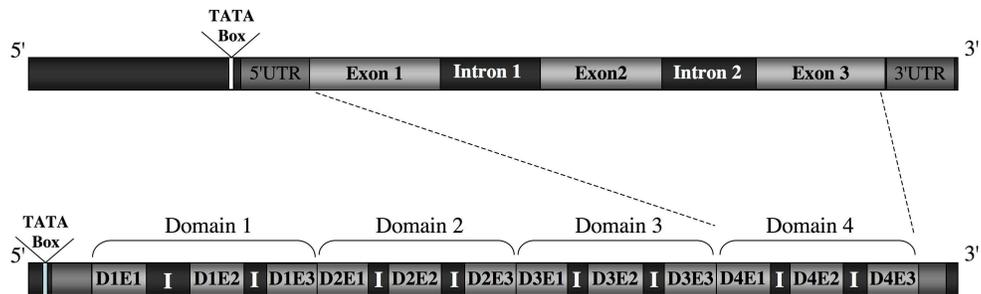


Figure 3. Gene structure from the single- (A) and tetra-domain (B) globins from Branchipolynoe spp. Promoter region: dark grey; UTR: medium grey; exons: light grey; introns: dark grey with "I" or "Intron". Gene sizes and components are not to scale.

Biol. Evol.