J Mol Evol (2006) 63:365–374
DOI: 10.1007/s00239-005-0198-9

JOURNAL OF MOLECULAR EVOLUTION

© Springer Science+Business Media, Inc. 2006

Gene Structure and Molecular Phylogeny of the Linker Chains from the Giant Annelid Hexagonal Bilayer Hemoglobins

Christine Chabasse, Xavier Bailly, Sophie Sanchez, Morgane Rousselot, Franck Zal

Equipe Ecophysiologie: Adaptation et Evolution Moléculaires, UPMC—CNRS UMR 7144, Station Biologique, BP 74, 29682 Roscoff cedex, France

Received: 17 August 2005 / Accepted: 31 March 2006

Abstract. Giant extracellular hexagonal bilayer hemoglobin (HBL-Hb), found only in annelids, is an ~3500-kDa heteropolymeric structure involved in oxygen transport. The HBL-Hbs are comprised of globin and linker chains, the latter being required for the assembly of the quaternary structure. The linker chains, varying in size from 225 to 283 amino acids, have a conserved cysteine-rich domain within their N-terminal moiety that is homologous to the cysteine-rich modules constituting the ligand binding domain of the low-density lipoprotein receptor (LDLR) protein family found in many metazoans. We have investigated the gene structure of linkers from *Arenicola marina*, *Alvinella pompejana, Nereis diversicolor, Lumbricus terrestris*, and *Riftia pachyptila*. We found, contrary to the results obtained earlier with linker genes from *N. diversicolor* and *L. terrestris*, that in all of the foregoing cases, the linker LDL-A module is flanked by two phase 1 introns, as in the human LDLR gene, with two more introns in the 3′ side whose positions varied with the species. In addition, we obtained 13 linker cDNAs that have been determined experimentally or found in the EST database LumbriBASE. A molecular phylogenetic analysis of the linker primary sequences demonstrated that they cluster into two distinct families of linker proteins. We propose that the common gene ancestor to annelid linker genes exhibited a four-intron and five-exon structure and gave rise to the two families subsequent to a duplication event.

Key words: Gene structure — Linker — Annelid — Low-density lipoprotein (LDL) receptor — LDL-A

Introduction

The giant extracellular hemoglobins (Hbs) of annelids, called hexagonal bilayer Hbs (HBL-Hbs) because of their characteristic quaternary structure (Vinogradov 1985; Lamy et al. 1996; Weber and Vinogradov 2001), have a molecular mass of ~3.5 MDa and are comprised of globin and linker chains. Based on the crystallographic structure of *Lumbricus terrestris* HBL-Hb, there are 144 globin chains (~17 kDa), which reversibly bind oxygen and others gaseous ligands (Vinogradov et al. 1993; Imai 1999; Weber and Vinogradov 2001), and 36 linker chains (24–32 kDa), which are devoid of heme and are required, together with the presence of Ca^{2+}, to form the HBL complex (Vinogradov et al. 1986; Gotoh et al. 1998; Kuchumov et al. 1999; Lamy et al. 2000). In addition to these structural properties, previous studies have revealed that linker chains may also possess a superoxyde dismutase activity (Liochev et al. 1996).

In contrast to the substantial number of annelid extracellular globin sequences that have become

Correspondence to: Christine Chabasse; Equipe Ecophysiologie: Adaptation et Evolution Moléculaires, Station biologique, BP 74, 29682 Roscoff cedex, France; *email:* chabasse@sb-roscoff.fr

available over the last 20 years (Gotoh et al. 1987; Bailly et al. 2002), only 10 linker amino acid sequences are known from the polychaetes *Neanthes diversicolor* (Suzuki et al. 1994), *Tylorrhynchus heterochaetus* (Suzuki et al. 1990a), and *Sabella spallanzanii* (Pallavicini et al. 2001), the vestimentiferan *Lamellibrachia sp*. (Suzuki et al. 1990b), the oligochaete *Lumbricus terrestris* (Suzuki and Riggs 1993; Fushitani et al. 1996), and the hirudinean *Macrobdella decora* (Suzuki and Vinogradov 2003). The linker chains vary from 185 to 255 amino acids and share a 39-residue cysteine-rich module within their N-terminal moiety that is very similar to low-density lipoprotein receptor class A repeats (LDL-A; also called complement-type repeats). These LDL-A modules are found in the N-terminal region of the low-density lipoprotein receptor (LDLR) in many metazoans, such as *Homo sapiens* (Sudhof et al. 1985a), *Xenopus laevi* (Mehta et al. 1991), and *Caenorhabditis elegans* (Chen et al. 2005). LDL-A modules are also found in other members of the LDLR superfamily, including very low-density lipoprotein receptor (VLDLR) (Takahashi et al. 1992), LDLR-related protein (LRP), and diverse LDLR unrelated functional proteins such as C9 or C8 complementary factor (DiScipio et al. 1984), renal glycoprotein gp330 (Raychowdhury et al. 1989; Saito et al. 1994), the receptor for subgroup A Rous sarcoma virus (Bates et al. 1993), and enterokinase (Kitamoto et al. 1994). The LDL-A modules of these proteins exhibit six invariant cysteine residues and highly conserved acidic residues (aspartatic acid and glutamatic acid). It has been suggested that these acidic residues are involved in ligand binding through ionic interactions with the basic residues of the ligand (Sudhof et al. 1985b; Brown and Goldstein 1986; Mahley 1988; Guo et al. 2004). Moreover, it has been shown that several of these LDL-A modules bind Ca^{2+}, which is essential for the maintenance of their biologically active conformations (van Driel et al. 1987; Dirlam et al. 1996; Guo et al. 2004). Based on structural similarities, it has been inferred that such interactions involving Ca^{2+} and acidic interactions would explain the molecular mechanisms for the subunit assembly in HBL-Hbs (Suzuki and Riggs 1993). In the case of *Lumbricus terrestris* HBL-Hb, it is likely that the binding of Ca^{2+} to the linker module acidic groups makes the linkers conformationally competent to bind the dodecamer globin subunits to form the HBL structure (Kuchumov et al. 1999, 2000).

The evolutionary origin of linkers in annelids remains unclear. Suzuki et al. have proposed that functional linkers could result from globin gene duplication and exon shuffling processes, including the insertion of an LDL-A module (Suzuki et al. 1990b; Suzuki and Riggs 1993). In this paper we report the cDNA sequence and gene structure of a linker chain from the polychaete *Arenicola marina*, two partial cDNA linker sequences and gene structures from the vestimentiferan *Riftia pachyptila*, three cDNA linker sequences and partial gene structures from a deep-sea hydrothermal vent polychaete *Alvinella pompejana*, the revised gene structures of *N. diversicolor* linker L2 (Suzuki et al. 1994) and *L. terrestris* L1 (Suzuki and Riggs 1993), and six linker sequences found in the expressed sequence tag (EST) database LumbriBASE (http://www.earthworms.org/) from two oligochaetes, *Lumbricus rubellus* and *Eiseinia andrei* (Lee et al. 2004). Our results indicate the presence of additional introns in the linker genes from *A. marina*, *R. pachyptila*, *A. pompejana*, *N. diversicolor*, and *L. terrestris* which were not described in an earlier study of *L. terrestris* and *N. diversicolor* linker gene structure (Suzuki and Riggs 1993). Furthermore, a molecular phylogenetic analysis of the 22 annelid linker sequences demonstrates that they cluster into two groups of linkers, which have probably evolved via duplication of a single ancestral linker gene. Since no LDLR- or VLDLR-related genes were found in the databank of annelid ESTs other than the LDL-A module-like in the HBL-Hb linker genes, it appears that the latter represent a rare case of a phylum-specific gene, whose recruitment was a functional and structural innovation restricted to annelids.

Materials and Methods

Collection of Biological Material

Specimens of the polychaetes *Arenicola marina* and *Neanthes diversicolor* were collected at low tide from a sandy shore near Roscoff (Penpoull Beach), Nord Finistère, France, and kept in local running seawater for 24 hr. Juvenile specimens of the hydrothermal vent tube worm *Riftia pachyptila* were collected at the vent site from the ridge segment 9°50′N on the EPR (Riftia field: 9°50.75′N, 104°17.57′W) at a depth of about 2.500 m, during the French *HOT 96* and the American *LARVE99* oceanographic cruises. Additional individuals were collected on the South EPR during the French oceanographic cruise *BIOSPEEDO* (Riftia field: 17°25.47′S, 113°12.281′W). The worms were sampled using the telemanipulated arms of the submersibles *Nautile* and *Alvin*, brought back alive to the surface in a thermally insulated basket, and immediately frozen and stored in liquid nitrogen after their recovery onboard.

Specimens of *Alvinella pompejana* were collected at a depth of 2.500 m on the East Pacific Rise (9°50′) by the manned submersible *Nautile* during the *HOPE'99* cruise. Once onboard, the animals were immediately frozen and stored in liquid nitrogen until used. Specimens of *Lumbricus terrestris* were collected at St Kirio (Ploujean, Nord Finistère, France) and frozen at –80°C until used.

Total RNA Extraction and cDNA Synthesis

Entire specimens of *A. marina*, *A. pompejana*, and *R. pachyptila* were crushed while submerged in liquid nitrogen. Total RNAs were extracted using the RNAble buffer (Eurobio) and poly(A) RNAs were then isolated using an mRNA Purification Kit (Amersham).

Reverse transcpritase PCR was carried out using an anchor 5′-CTC CTC TCC TCT CCT CTT CCT oligo(dT)$_{17}$ primer.

Isolation of Genomic DNA

Specimens of *A. marina*, *A. pompejana*, *N. diversicolor*, *R. pachyptila*, and *L. terrestris* were incubated in 700 µl of PK buffer (50 mM Tris/HCl containing 100 mM NaCl, 25 mM EDTA, and 1% sodium dodecyl sulfate [SDS], pH 8) with 15 µl proteinase K (10 µg/µl) at 65°C for 1 hr. The supernatant was recovered after centrifugation at 12,000*g* for 5 min at 4°C and extracted with 700 µl phenol. The material was extracted once with phenol/chloroform/isoamyl alcohol (25/24/1) and once with chloroform. The resulting DNA was precipitated with isopropanol at –20°C overnight and washed once with 75% ethanol. DNA was finally resuspended in 100 µl TE buffer (10 mM Tris/HCl, 0.1mM EDTA, pH 8) and stored at 4°C.

Amplification of cDNA and Genomic DNA

The PCR reaction was carried out in a 25-µl volume containing 10–50 ng of cDNA/gDNA template, 100 ng of each degenerate primer, 200 µM dNTPs, 2.5 mM $MgCl_2$, and 1 unit of DNA polymerase (Uptima, Interchim). PCR conditions were as follows: an initial denaturation step at 95°C for 5 min, 35 cycles consisting of denaturation at 95°C for 40 sec, annealing at 58°C for 40 sec, extension at 72°C for 50 sec, and a final elongation step at 72°C for 10 min. Primers used are available upon request.

Cloning and Sequencing

The PCR products were the cloned using a TOPO-TA cloning Kit (Invitrogen). The positive recombinant clones were isolated and plasmid DNA was prepared by the FlexiPrep Kit (Amersham). Purified plasmids containing the putative globin insert were used in a dye-primer cycle sequencing reaction, using the Big Dye Terminator V3.1 Cycle Sequencing kit (Applied Biosystems). PCR products were subsequently run on a 3100 Genetic Analyser (Applied Biosystems) at the Roscoff sequencing core facility Ouest genopole platform.

Rapid Amplification of cDNA Ends

cDNAs ends were obtained by PCR using the 5′ and 3′RACE kit from Roche according to the manufacturer's instructions. Buffer, reagents, and other conditions for the nested PCR were as described by the manufacturer. The RACE products were purified, cloned with the TOPO-TA cloning kit (Amersham), and sequenced as described above.

Sequence Analysis

The peptide signal cleavage site was predicted by the SignalP 3.0 Server (Bendtsen et al. 2004) (http://www.cbs.dtu.dk/services/SignalP/).

Linker Multiple Alignment Construction

Analyses were performed on the linker chains of *Lumbricus terrestris* L1 (AAF99389) and L3 (S65723), *Tylorrhynchus heterochaetus* L1 (P18207) and L2 (P18208), *Lamellibrachia* sp. LAV-1 (A35012), *Sabella spallanzanii* L1 (CAB38536) and L3 (CAC37413), *Macrobdella decora* L1 (BAC82449), and *Neanthes diversicolor* L2 (BAA09580). Others sequences have been found in LumbriBASE (http://www.earthworms.org/): *Lumbricus rubellus* L1 (CO046524), L2 (CO046717), L3 (CF839233), and L4 (BF422664 and BG269978),) and *Eisenia andrei* L1 (BP524390), L2 (BP524387), and L3 (BP524672). Amino acid sequences were aligned with the MUSCLE program (Edgar 2004) (http://phylogenomics.berkeley.edu/cgi-bin/muscle/input_muscle.py) and verified manually.

Molecular Phylogeny

An unrooted phylogenetic tree was constructed using the neighbor-joining (NJ) method from the linker amino acid sequence multiple alignment. Gapped domains ambiguously aligned were ignored (Fig. 1). The tree was computed using the PHYLIP program package version 3.63 (J. Felsenstein, Department of Genetics, University of Washington, Seattle) with 1000 bootstrap reiterations and the JTT transition matrix (Jones et al. 1992).

Maximum likelihood (ML) analyses were performed with PHYML software (Guindon and Gascuel 2003), with the JTT model of amino acid substitution (Jones et al. 1992) and 1000 bootstrap reiterations.

Results

Characterization of New Linker Sequences

Arenicola marina. There are two different linker chains, L1 and L2, involved in homo- and heterodimer formation, with a mass of 25,174.1 and 26,829.7 Da, respectively (Zal et al. 1997b). The *A. marina* linker presented here (AM000028) contains an open reading frame of 256 codons including a signal peptide (residues 1 to 16). The cDNA-derived amino acid sequence without the signal peptide has a calculated mass of 26,740.7 Da. We assigned this amino acid sequence to linker L2, even though it exhibits a difference of 89 Da from this polypeptide. This difference may be explained either by the sequencing of an allelic form due to the fact that *A. marina* specimens were harvested in different areas or by posttranslation modifications, such as carboxylation of two glutamic or aspartic acid residues. The cDNA-derived amino acid sequence exhibits 12 cysteine residues, which is in agreement with previous studies on *A. marina* HBL-Hb (Zal et al. 1997b).

Riftia pachyptila. The vestimentiferan *Riftia pachyptila* possesses four types of linker chains (Zal et al. 1996). Two partial linker sequences (AM000032 and AM000033) have been identified and arbitrarily called LA and LB. The signal peptide of the LB sequences includes residues 1 to 19.

Alvinella pompejana. The extracellular HBL-Hb of the polychaete *Alvinella pompejana* involves four types of linker chains (L1 to L4) (Zal et al. 1997a). Three linker sequences have been sequenced here (AM000029, AM000030, and AM000031). The first sequence contains an open reading frame of 225 amino acids, with the first 19 amino acids charac-

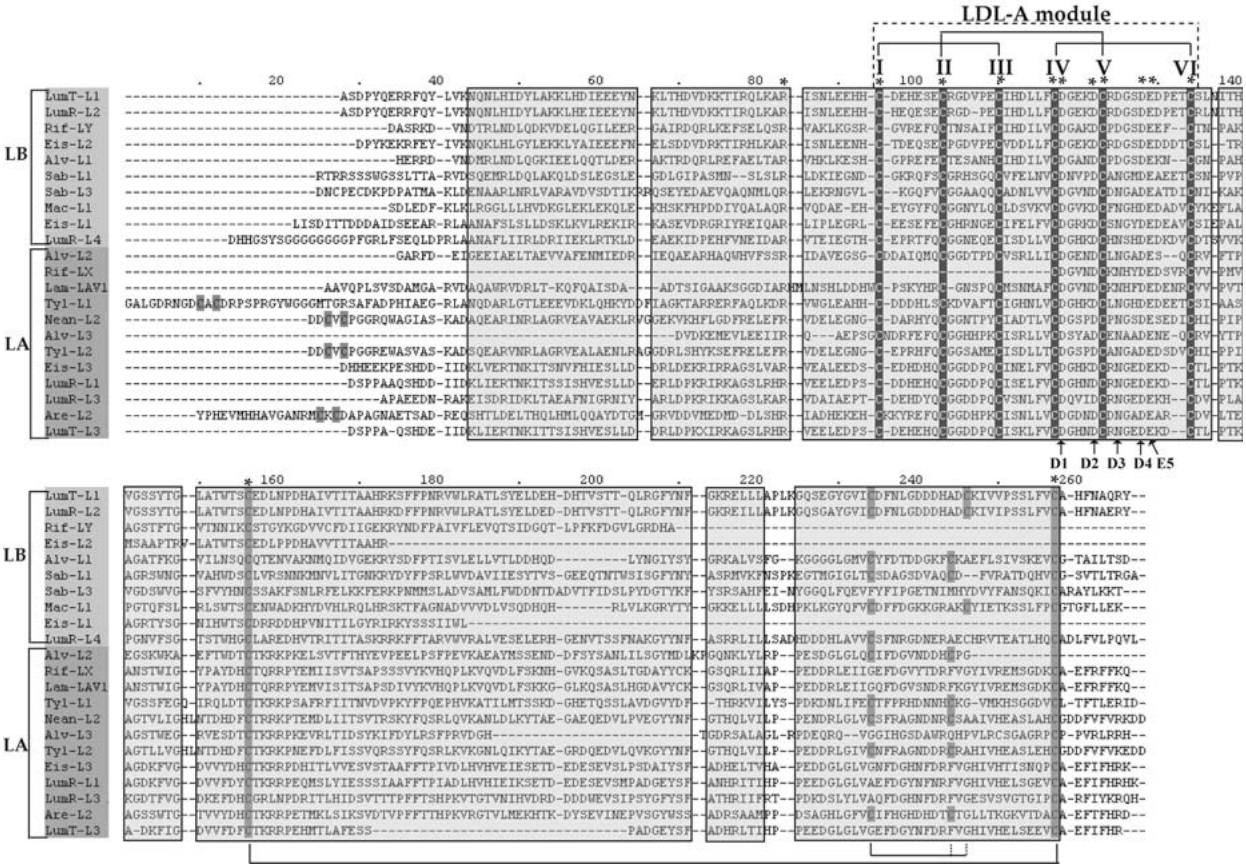

Fig. 1. Multiple alignment of linker amino acid sequences of the oligochaetes *Lumbricus terrestris* (LumT), *Lumbricus rubellus* (LumR), and *Eiseinia andrei* (Eis), the polychaetes *Alvinella pompejana* (Alv), *Sabella spallanzanii* (Sab), *Tylorrhynchus heterochaetus* (Tyl), *Neanthes diversicolor* (Nean), and *Arenicola marina* (Are), the vestimentiferans *Lamellibrachia* sp. (Lam) and *Riftia pachyptila* (Rif), and the hirudinae *Macrobdella decora* (Mac). Signal peptides, when present, have been removed. The conserved residues are marked by an asterisk. The LDL-A module is indicated by dashed lines. The roman numerals refer to the order of appearance of the cysteine residues in the LDL-A module. Solid lines indicate the possible intrachain disulfide bonds. Gray boxes correspond to the restricted alignment used for the phylogeny reconstruction.

terized as a signal peptide. The molecular mass calculated for the polypeptide deduced from the cDNA is 22,895 Da. The amino acid sequence exhibits 10 cysteine residues, which is in agreement with previous studies on *A. pompejana* HBL-Hb, and corresponds to linker L1 (Mr 22,889 Da) (Zal et al. 1997a). Two other partial cDNA sequences have been found and arbitrarily called L2 and L3.

Eisenia andrei. Three linker sequences have been found in the EST database LumbriBASE and called L1 (BP524390), L2 (BP524387), and L3 (BP524672) (Lee et al. 2004). The amino acid sequence derived from the L3 cDNA is composed of 240 amino acids, with the first 19 amino acids constituting the signal peptide. The L1 and L2 sequences are not complete, with the signal peptides comprising the first 20 and 16 residues, respectively.

Lumbricus rubellus. Four complete linker sequences—L1 (CO046524), L2 (CO046717), L3 (CF839233), and L4 (BF422664 and BG269978)—have been found in the EST database LumbriBASE. The sequences contain an open reading frame of 236, 241, 243, and 283 amino acids, respectively. For each sequence, the signal peptide has been identified and corresponds to the residues 1 to15 for L1, 1 to 16 for L2, 1 to 26 for L3, and 1 to 16 for L4. The calculated masses are 24,950.5, 25,0905.8, 24,455.2, and 29,983.4 Da, respectively.

Multiple Alignment

A multiple alignment of the linker amino acid sequences with the signal peptide removed is shown in Fig. 1. Thirteen residues appear to be invariant—Arg-83, Cys-95, Cys-103, Cys-110, Cys-117, Asp-118, Asp-122, Cys-123, Asp-128, Glu-129, Cys-134, Cys-156, and Cys-257—in agreement with previous results (Suzuki et al. 1994; Suzuki and Vinogradov 2003). A remarkable feature of the linker chain is the conservation of a 39-residue cysteine-rich domain (at positions 95–134), comprised of a repeating pattern of cysteinyl residues: C-X_{5-7}-C-X_{5-6}- C-X_6-C-D-X_3-D-C-X_4-D-E-X_{2-4}-C.

Linker Gene Structure

The complete gene structure of *A. marina* linker and partial gene structures of *R. pachyptila* LA and *A. pompejana* L1 have been determined, as well as the revised gene structure from *N. diversicolor* L2 and *L. terrestris* L1 (Figs. 2a and b) The length and position of introns are summarized in Fig. 3.

In the *A. marina* linker sequence, a first intron (434 pb) splits the codon at position 91 in phase 1 (Fig. 2). A second intron position has been identified in *A. marina* linker (348 pb), *N. diversicolor* linker L2 (347 pb), and *L. terrestris* L1 (951 pb) and splits the codon in phase 1 at position 135 in the alignment (Fig. 1). In *R. pachyptila* linker LA, the splice site is at the same position, but the intron sequence is not complete.

A third intron position, variable among the species studied here, has been identified between codon 182 and codon 183 in *A. marina* linker (562 pb), between codon 178 and codon 179 in *N. diversicolor* L2 (391 pb), between codon 193 and codon 194 in *A. pompejana* L1 (382 pb), and between codon 181 and codon 182 in *R. pachyptila* linker LA (673 pb). They all split the codon in phase 0. No intron has been found in this region in *L. terrestris* L1.

A fourth intron position, also variable among the species studied here, has been found at position 266 in phase 1 in *A. marina* linker L2 (757 pb). In *N. diversicolor*, the fourth intron (359 pb) is found four nucleotides downstream from the stop codon. In *R. pachyptila* linker LA, the fourth intron (657 pb) is found just after the stop codon.

The splice junctions have been analyzed and are in agreement with consensus sequences (i.e., the splicing donor GT and acceptor AG are present).

Phylogenetic Analyses of Annelid Linker Chains

Given that linkers genes are specific to the Annelida phylum, it is not possible to provide an outgroup, thus we performed unrooted trees. The neighbor-joining and maximum likelihood trees realized from all the linker amino acid sequences exhibit the same topologies, but with low bootstrap values (50%), which does not allow clear support of the dichotomy between the two clusters LA and LB (Fig. 4).

Discussion

Functional and Structural Inferences from Linker Sequences

Cysteine residue motif in LDL-A modules. Six of the ten to twelve cysteine residues found in linker sequences are part of the cysteine-rich module common to all linkers and the LDL-A modules comprising the ligand binding domains of the LDLR protein superfamily and similar modules in several unrelated proteins. The structure of the mammalian LDL-A module has been determined (Bieri et al. 1995), showing that the connectivity of the three disulfide bridges is Cys(I)-Cys(III), Cys(II)-Cys(V), and Cys(IV)-Cys(VI), with the Roman numerals referring to the order of the cysteine residues in the LDL-A module (Fig. 1). In view of the high similarity between the primary sequences of the linker cysteine-rich module and the LDL-A module, it appears reasonable to expect that the annelid linkers should exhibit the same disulfide connectivity as in the LDL-A module.

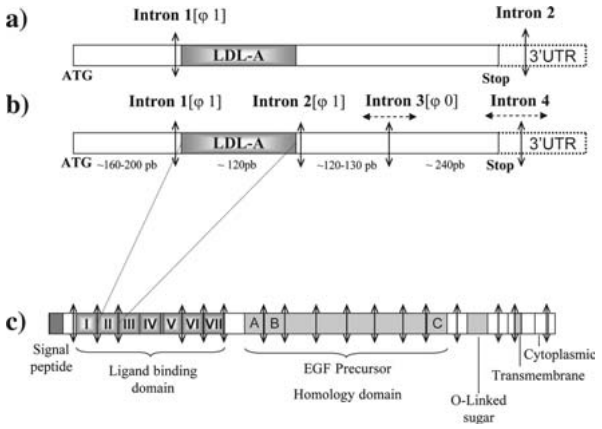

Fig. 2. **a** Linker gene structure as described by Suzuki et al. (Suzuki and Riggs 1993). **b** Present revised linker gene structure. **c** Human LDLR gene structure (Sudhof et al. 1985a). Black arrows indicate intron positions.

Aspartic and glutamic acidic residues in LDL-A modules. Several studies have shown that five conserved aspartatic and glutamic acidic residues interacting with calcium (referred as D^1, D^2, D^3, D^4, and E^5) (Fig. 1) play a role in the conformation and function of the LDL-A module in the LDLR superfamily (Fass et al. 1997; Simonovic et al. 2001; Rudenko et al. 2002). The side chains of four of the conserved acidic residues (D^1, D^2, D^4, and E^5) as well as the carbonyl oxygen group of two nonconserved residues are involved in the calcium coordination of the LDL-A modules of the LDLR (Fass et al. 1997). Indeed, Ca^{2+} is required for correct folding but also for maintaining the structural integrity of the LDL-A module in the LDLR superfamily (van Driel et al. 1987; Atkins et al. 1998; Guo et al. 2004). Residues D^4 and E^5 are present in all of the modules studied (Guo et al. 2004) and are also found in all the linker sequences shown in Fig. 1. Hence, D^4 and E^5 can be considered as invariant residues, like the six conserved cysteine residues. Interestingly, Kuchumov et al. (1999, 2000) have proposed that Ca^{2+} is necessary for the linkers to adopt an assembly-competent conformation in order to allow the reassembly of the HBL structure subsequent to experimental dissociation of *L. terrestris* HBL-Hb in solution.

		Intron	Position	Size (bp)	Splicing donor	BP	Splicing acceptor	
Consensus					HAG **gt** ragt	ctray	yyny **ag** GK	
L. terrestris	L1	Intron 1	91	2011	AAG **gc** aagt	*ttaac*	cttc **ag** AG	(Suzuki and Riggs 1993)
		Intron 2	135	951	GTC **gt** gagt	*ataat*	tcgc **ag** TC	
		Intron 3	NF	NF	NF	NF	NF	
		Intron 4	S+21	717	CAC **gt** acgt	cttaa	ccgc **ag** GT	(Suzuki and Riggs 1993)
N. diversicolor		Intron 1	91	287	AAG **gt** cagt	ctaag	tttc **ag** GA	(Suzuki et al. 1994)
		Intron 2	135	347	GCC **gt** gagt	*ataat*	ttgc **ag** AC	
		Intron 3	178-179	391	CAG **gt** aggt	ctaac	tttc **ag** TC	
		Intron 4	S+4	359	TTG **gt** aatg	ctcat	ttac **ag** GT	
A. marina	L2	Intron 1	91	434	AGA **gt** aagt	ctgaa	ttct **ag** AG	
		Intron 2	135	348	GTG **gt** cagt	*ttgac*	tcac **ag** AT	
		Intron 3	182-183	562	AAG **gt** aaga	*ataac*	cacc **ag** GT	
		Intron 4	266	757	GGG **gt** aggt	*tttat*	ttgc **ag** AT	
R. pachyptila	LA	Intron 1	/	/	/	/	/	
		Intron 2	135	UC	/	cttag	ccgc **ag** TT	
		Intron 3	181-182	673	CAG **gt** aagc	*ttcat*	tttc **ag** CC	
		Intron 4	S+0	657	TAG **gt** gagc	ctaac	ccac **ag** AT	
A. pompejana	L1	Intron 3	193-194	382	CAG **gt** aggc	*ttgat*	ttca **ag** GA	

Fig. 3. Position, length of introns, exon/intron splice junctions, and possible branch points in *A. marina*, *A. pompejana*, *R. pachyptila*, *L. terrestris*, and *N. diversicolor* linker genes. The consensus sequences presented are from Mount et al. (Mount 1982; Keller and Noon 1984). Sequences in italics correspond to branch points where the well-conserved C or T or A is not found. BP, branch point; NF, not found; slash (/), sequence not complete; S, stop codon.

In addition, the acidic residues (except D^2 and D^3) are supposed to be involved also in ligand binding in LDL-A modules of the LDLR (Fass et al. 1997). Surprisingly, in 17 linker sequences among the 22 shown in Fig. 1, the aspartic acid D^3 is substituted by an asparagine residue (Fig. 1), which may be functionally significant for the LDL-A module in annelid linker chains relative to the lipoprotein receptor LDL-A modules in the other organisms. Furthermore, although the LDL-A module has been proposed to be the domain responsible for binding between linker chains and globin dodecamer subunits (Kuchumov et al. 1999), the role of the different acidic residues as well as of the conserved D^3 asparagine residue in linker sequences remains to be elucidated.

Cysteine residue signatures outside the LDL-A module. Four or six cysteine residues are found outside the LDL-A module in linkers in either conserved or variable positions (Fig. 1). From data described for *T. heterochaetus* L1 in Suzuki et al. (1994), we assume that Cys-156 and Cys-257, and Cys-234 and Cys-244/246, would form an intrachain disulfide bridge. As a consequence, the two cysteine residues present in the N-terminal region of the sequence are those probably involved in the interchain disulfide bridge (Suzuki et al. 1990a). Indeed, as suggested by Suzuki et al. (1990a) and confirmed by the new linker sequences, the two cysteine residues (before position 40 in the alignment in Fig. 1) are present in linker chains known to form dimers or trimers (such as *T. heterochaetus*, *S. spallanzanii*, *N. diversicolor*, and *A. marina*) and are absent in linker chains present only as monomers (*L. terrestris*, *A. pompejana*, *M. decora*, and *Lamellibrachia* sp.).

Revised Linker Gene Structure

Intron positions in agreement with previous studies. Earlier studies on the gene structures of linker L1 from *L. terrestris* (Suzuki and Riggs 1993) and linker L2 from *N. diversicolor* (Suzuki et al. 1994)

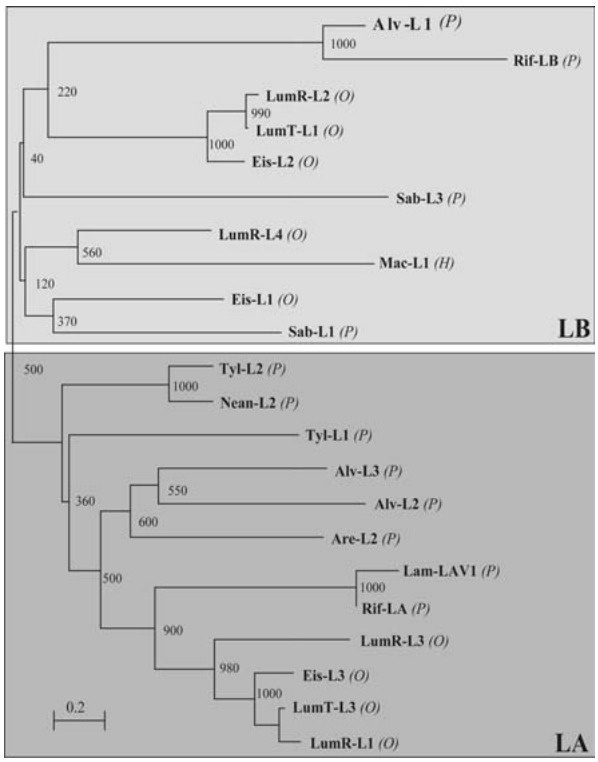

Fig. 4. Neighbor-joining tree of linker amino acid sequences obtained with 1000 bootstrap reiterations. Bootstrap values are indicated at each branch node. (*P*) Polychaete; (*O*) oligochaete; (*H*) hirudinae. See the legend to Fig. 1 for other abbreviations.

have shown a two-intron/three-exon gene pattern (Fig. 2a). The first intron occurs on the 5′ flank of the LDL-A module, i.e., in position 91 (Fig. 1) and is also found at that position in the linker L2 sequence of *A. marina*.

The second intron described in *L. terretris* linker L1 is in the 3'UTR region (Suzuki and Riggs 1993). Even though the position is not strictly conserved in *R. pachyptila* LA and *N. diversicolor* L2, the intron position is also in the 3′UTR region, whereas in *A. marina* L2 the intron position is in the coding sequence (Fig. 2b). Introns in the 3′UTR sequence occur rarely (Nagy and Maquat 1998) and could

play a role in nonsense-mediated reduction of mRNA abundance (Zhang et al. 1998a). Intron position polymorphism between two adjacent codons (phase 0) or inside a codon (phase 1 or 2) can occur, as in the case of the Hb gene (Hardison 1998), but to our knowledge, such an amplitude around the stop codon has not been frequently reported in the literature. Therefore, it is difficult to speculate on a possible intron sliding event or on the biological significance with respect to mRNA transcription or regulation.

Discovery of new intron positions. Surprisingly, we found an additional intron in *A. marina, N. diversicolor, R. pachyptila,* and *L. terrestris* linker sequences flanking 3′ to the LDL-A modules (Fig. 2b). The LDL-A modules of annelids are in fact flanked by two introns, similar to the human LDL-A modules (Fig. 2c). The ligand binding domain of the human LDLR is composed of seven collinear LDL-A modules (Sudhof et al. 1985a). The gene structure shows that the different LDL-A modules are encoded by distinct exons flanked by phase 1 inserted introns, except for domain III-IV-V, encoded by a single exon (Fig. 2c). Interestingly, the LDL-A module of linker sequences is also flanked by two phase 1 introns. Patthy has shown that most of the identified cases of exon shuffling leading to new chimeric protein (for example, interleukin-2 receptor, follistatin, or factor XIIIb subunit) exhibit a typical phase 1 intron pattern (Patthy 1987, 1991, 1995). Suzuki et al. have proposed that the linker sequence would result from the duplication of a globin gene followed by the loss of the first exon and insertion of an LDL-A module by an exon shuffling process (Suzuki and Riggs 1993). The occurrence of these flanking introns supports the assumption of an exon shuffling event to explain the presence of the LDL-A module in linker genes.

A third intron has also been found in *A. marina*, *N. diversicolor*, *R. pachyptila*, and *A. pompejana* linker sequences but not in *L. terrestris* linker L1. Although they all are phase 0 introns, the intron positions are not conserved and are distributed over a region of 16 amino acids. The absence of the third intron in the *L. terrestris* L1 could be due to a loss of this intron in this species. It is, however, important to keep in mind that the gene structure of the *L. terrestris* L1 reported by Suzuki et al. (Suzuki and Riggs 1993) was constructed from three independent genomic fragments, with the putative second intron position in the second fragment. It is possible that the central fragment amplified was a contamination of cDNA, as they did not characterize the second intron. Another possibility is that the gene amplified in this study is a closely related gene (paralogous) different from the one identified by Suzuki et al.

Linker Structure

Suzuki et al. (1994) have proposed that linker sequences can be separated into three domains: N-terminal domain 1 (positions 1 to 94), the cysteine-rich domain (positions 95 to 136), and the C-terminal domain (positions 137 to 268). The new gene structures presented here suggest rather that linker sequences should be separated into four domains, corresponding to the four exons, which suggest a more complicated emergence.

The hypothesis of a common origin for the linker and globin genes has been proposed by Suzuki et al. (Suzuki and Riggs 1993), involving particularly the N-terminal domain of linker sequences. Furthermore, this common origin is supported by the recognition of both globin and linker chains by monoclonal antibodies against *L. terrestris* HBL-Hb (Lightbody et al. 1988), which implies that they share some common epitopes.

On the other hand, this hypothesis of a globin-like N-terminal domain is challenged by the structure of *L. terrestris* HBL-Hb (Royer et al. 2000), which shows the linker chains as long coiled-coil helices attached to a globular domain. Furthermore, this globular domain is composed of both the LDL-A module and the C-terminal domain, and is in contact with the globin subunits, and the N-terminal domain does not appear as a globin-like structure, and is involved only in linker-linker interactions (Royer et al. 2000). A higher-resolution crystal structure will be required to determine the structure of the linker chains in HBL Hbs.

Linker Evolutionary History

Historically, linkers have been named according to their chronological discovery (purification and characterization) such as L1 for the first one, L2 for the second one, etc. This numbering has no sense in a molecular phylogenetic framework taking into account paralogy and orthology relationships between genes or genes products (i.e., proteins).

Despite an obvious divergence between linker sequences which illustrate saturation and almost a limitation for phylogenetic analyses (neighbor-joining analysis in Fig. 4), the linker sequences cluster in two main groups (with quite low bootstrap values) we propose to name arbitrarily LA and LB. (Note that we obtained the same low-bootstrap topology using Bayesian inference with Mr Bayes software [data not shown].) The weak resolution of the tree (low bootstrap values) can also be allocated to the difficulty of getting an unambiguous multiple alignment given sequence divergence, also in terms of insertion and deletion: the possibility of extracting the maximum

number of informative positions from a multiple alignment of linkers will be greatly improved with the three-dimensional structure of a linker. Because different species, such as the polychaete *Alvinella pompejana*, some oligochaetes (*Lumbricus terrestris*, *L. rubellus*, and *Eiseinia andrei*) and the vestimentiferan *Riftia pachyptila*, exhibit distinct copies of linkers in both LA and LB families, we assume that this dichotomy suggests the occurrence of an ancestral duplication and sets of paralogous genes. It is of a prime importance to be aware that, given that there is no possibility of using an outgroup, the split (LA and LB clusters) could be an oversimplification of the real situation, resulting in the lack of additional linker sequences of polychaetes and hirudinae (only one sequence of hirudinae is available to date, from the leech *Macrobdella decora*).

It is noteworthy that the LA family shows an organization consistent with the annelid phylogeny, with a cluster of polychaetes including the vestimentiferans *R. pachyptila* and *Lamellibrachia* sp. clearly separated (i.e., supported by a high bootstrap value) from a cluster made up of oligochetes (Fig. 4). This strongly suggests that the LA group corresponds to a cluster of orthologous genes (with additional duplication events for *A. pompejana* and *T. heterochaetus*). In contrast, the LB cluster does not reflect the annelid phylogeny as the LA cluster does, which suggests that LB corresponds to a cluster of paralogous sequences. Additional sequences from polychaetes, oligochaetes, and hirudinae are needed to clarify the structure of the LB family. We also assume that the common ancestor gene to all extant linker genes had a four-intron/five-exon structure and gave rise to the two families LA and LB after a duplication event.

The LDL-A modules from the LDLR family and the annelid linker sequences exhibit the same conserved amino acids (cysteine and acidic residues) and are similarly flanked by two introns. Based on these observations, all the modules found in the proteins from *Homo sapiens*, *Xenopus laevi*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and annelids could be considered as homologous and resulting from a common ancestor. Moreover, as described above, linker genes are specific to the annelid phylum, and as a consequence, the LDL-A modules in annelids may have been recruited to contribute to the elaboration of linker chains.

So far, there have been no reports in the literature of either LDLR or VLDLR genes in annelids and none were found via BLAST searches in ESTs from *L. rubellus* and *E. andrei*. In contrast, some LDLR-related sequences have been identified from the mollusks *Crassostrea gigas* (CAD82865) and *Lymnaea stagnalis* (CAA80651) (Tensen et al. 1994), the arthropod *Drosophila melanogaster* (AAM50824) (Stapleton et al. 2002), and the nematode *Caenorhabditis elegans* (Q04833) (Yochem and Greenwald 1993). This strongly suggests that an LDLR gene was present in the ancestor common to both protostomes and deuterostomes, and would have been lost in the annelid phylum. On the other hand, it is also possible that the LDLR gene may have existed in the annelid ancestor but was co-opted into a different function, explaining the absence of LDLR-like proteins in annelids.

In this paper, we have presented new linker sequences as well as a revised linker gene structure. These results indicate the presence of two linker families, LA and LB, and provides a new hypothesis regarding the evolutionary history of linkers. However, more data regarding both primary sequences and structural features are necessary to really understand the molecular events that gave rise to the linker genes and, ultimately, to the HBL-Hbs.

Acknowledgments. We are indebted to Serge Vinogradov for his critical reading and comments on the manuscript. We thank the captains and crews of the NO L'Atalante and the pilots, cruises, and team of the submersibles Nautile and Alvin of the *HOT 96*, *HOPE'99*, *LARVE99*, and *BIOSPEEDO*. We thank the referees for helpful discussion. This work was supported by the CNRS, European grant FEDER presage 3814, and the Conseil Régional de Bretagne (contract A2C809).

References

Atkins AR, Brereton IM, Kroon PA, Lee HT, Smith R (1998) Calcium is essential for the structural integrity of the cysteine-rich, ligand-binding repeat of the low-density lipoprotein receptor. Biochemistry 37:1662–1670

Bailly X, Jollivet D, Vanin S, Deutsch J, Zal F, Lallier F, Toulmond A (2002) Evolution of the sulfide-binding function within the globin multigenic family of the deep-sea hydrothermal vent tubeworm Riftia pachyptila. Mol Biol Evol 19: 1421–1433

Bates P, Young JA, Varmus HE (1993) A receptor for subgroup A Rous sarcoma virus is related to the low density lipoprotein receptor. Cell 74:1043–1051

Bendtsen JD, Nielsen H, von Heijne G , Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795

Bieri S, Djordjevic JT, Daly NL, Smith R, Kroon PA (1995) Disulfide bridges of a cysteine-rich repeat of the LDL receptor ligand-binding domain. Biochemistry 34:13059–13065

Brown MS, Goldstein JL (1986) A receptor-mediated pathway for cholesterol homeostasis. Science 232:34–47

Chen N, Harris TW, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Canaran P, Chan J, Chen CK, Chen WJ, Cunningham F, Davis P, Kenny E, Kishore R, Lawson D, Lee R, Muller HM, Nakamura C, Pai S, Ozersky P, Petcherski A, Rogers A, Sabo A, Schwarz EM, Van Auken K, Wang Q, Durbin R, Spieth J, Sternberg PW, Stein LD ((31 co-authors). 2005) WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. Nucleic Acids Res 33:D383–D389

Dirlam KA, Gretch DG, LaCount DJ, Sturley SL, Attie AD (1996) Expression and characterization of a truncated, solu-

ble, low-density lipoprotein receptor. Protein Expr Purif 8: 489–500

DiScipio RG, Gehring MR, Podack ER, Kan CC, Hugli TE, Fey GH (1984) Nucleotide sequence of cDNA and derived amino acid sequence of human complement component C9. Proc Natl Acad Sci USA 81:7298–7302

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Fass D, Blacklow S, Kim PS, Berger JM (1997) Molecular basis of familial hypercholesterolaemia from structure of LDL receptor module. Nature 388:691–693

Fushitani K, Higashiyama K, Asao M, Hosokawa K (1996) Characterization of the constituent polypeptides of the extracellular hemoglobin from *Lumbricus terrestris*: heterogeneity and discovery of a new linker chain L4. Biochim Biophys Acta 1292:273–280

Gotoh T, Shishikura F, Snow JW, Ereifej KI, Vinogradov SN, Walz DA (1987) Two globin strains in the giant annelid extracellular haemoglobins. Biochem J 241:441–445

Gotoh T, Sano T, Shibuya A, Yamaki M, Imai K, Ebina S (1998) Hexagonal bilayer structuring activity of linker chains of an annelid giant hemoglobin from the polychaete *Perinereis aibuhitensis*. Arch Biochem Biophys 360:75–84

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Guo Y, Yu X, Rihani K, Wang QY, Rong L (2004) The role of a conserved acidic residue in calcium-dependent protein folding for a low density lipoprotein (LDL)-A module: implications in structure and function for the LDL receptor superfamily. J Biol Chem 279:16629–16637

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

Hou S, Belisle C, Lam S, Piatibratov M, Sivozhelezov V, Takami H, Alam M (2001) A globin-coupled oxygen sensor from the facultatively alkaliphilic Bacillus halodurans C-125. Extremophiles 5:351–354

Imai K (1999) The haemoglobin enzyme. Nature 401:437–439

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Keller EB, Noon WA (1984) Intron splicing: a conserved internal signal in introns of animal pre-mRNAs. Proc Natl Acad Sci USA 81:7417–7420

Kitamoto Y, Yuan X, Wu Q, McCourt DW, Sadler JE (1994) Enterokinase, the initiator of intestinal digestion, is a mosaic protease composed of a distinctive assortment of domains. Proc Natl Acad Sci USA 91:7588–7592

Kuchumov AR, Taveau JC, Lamy JN, Wall JS, Weber RE, Vinogradov SN (1999) The role of linkers in the reassembly of the 3.6 MDa hexagonal bilayer hemoglobin from *Lumbricus terrestris*. J Mol Biol 289:1361–1374

Kuchumov AR, Loo JA, Vinogradov SN (2000) Subunit distribution of calcium-binding sites in *Lumbricus terrestris* hemoglobin. J Protein Chem 19:139–149

Lamy JN, Green BN, Toulmond A, Wall JS, Weber RE, Vinogradov SN (1996) Giant hexagonal bilayer hemoglobins. Chem Rev 96:3113–3124

Lamy J, Kuchumov A, Taveau JC, Vinogradov SN, Lamy JN (2000) Reassembly of *Lumbricus terrestris* hemoglobin: a study by matrix-assisted laser desorption/ionization mass spectrometry and 3D reconstruction from frozen-hydrated specimens. J Mol Biol 298:633–647

Lee MS, Cho SJ, Lee JA, Park BJ, Cho HJ, Moon JS, Kim SK, Choo JK, Park SC (2004) Transcriptome analysis in the midgut of the earthworm (Eisenia andrei) using expressed sequence tags. Unpublished

Lightbody JJ, Quabar AN, Mainwaring MG, Young JS, Walz DA, Vinogradov SN, Gotoh T (1988) Immunological relatedness of annelid extracellular hemoglobins and chlorocruorins. Comp Biochem Physiol B 90:301–305

Liochev SI, Kuchumov AR, Vinogradov SN, Fridovich I (1996) Superoxide dismutase activity in the giant hemoglobin of the earthworm, *Lumbricus terrestris*. Arch Biochem Biophys 330:281–284

Mahley RW (1988) Apolipoprotein E: cholesterol transport protein with expanding role in cell biology. Science 240:622–630

Mehta KD, Chen WJ, Goldstein JL, Brown MS (1991) The low density lipoprotein receptor in *Xenopus laevis*. I. Five domains that resemble the human receptor. J Biol Chem 266:10406–10414

Mount SM (1982) A catalogue of splice junction sequences. Nucleic Acids Res 10:459–472

Nagy E, Maquat LE (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem Sci 23:198–199

Pallavicini A, Negrisolo E, Barbato R, Dewilde S, Ghiretti-Magaldi A, Moens L, Lanfranchi G (2001) The primary structure of globin and linker chains from the chlorocruorin of the polychaete *Sabella spallanzanii*. J Biol Chem 276:26384–26390

Patthy L (1987) Intron-dependent evolution: preferred types of exons and introns. FEBS Lett 214:1–7

Patthy L (1991) Exons—Original building blocks of proteins? Bioessays 13:187–192

Patthy L (1995) Protein evolution by exon shuffling. In: Molecular Biology Intelligence Unit. Springer, Heidelberg, pp. 136

Raychowdhury R, Niles JL, McCluskey RT, Smith JA (1989) Autoimmune target in Heymann nephritis is a glycoprotein with homology to the LDL receptor. Science 244:1163–1165

Royer WE Jr, Strand K, van Heel M, Hendrickson WA (2000) Structural hierarchy in erythrocruorin, the giant respiratory assemblage of annelids. Proc Natl Acad Sci USA 97:7107–7111

Rudenko G, Henry L, Henderson K, Ichtchenko K, Brown MS, Goldstein JL, Deisenhofer J (2002) Structure of the LDL receptor extracellular domain at endosomal pH. Science 298:2353–2358

Saito A, Pietromonaco S, Loo AK, Farquhar MG (1994) Complete cloning and sequencing of rat gp330/"megalin", a distinctive member of the low density lipoprotein receptor gene family. Proc Natl Acad Sci USA 91:9725–9729

Simonovic M, Dolmer K, Huang W, Strickland DK, Volz K, Gettins PG (2001) Calcium coordination and pH dependence of the calcium affinity of ligand-binding repeat CR7 from the LRP. Comparison with related domains from the LRP and the LDL receptor. Biochemistry 40:15127–15134

Stapleton M, Carlson J, Brokstein P, Yu C, Champe M, George R, Guarin H, Kronmiller B, Pacleb J, Park S, Wan K, Rubin GM, Celniker SE (2002) A Drosophila full-length cDNA resource. Genome Biol 3:RESEARCH0080

Sudhof TC, Goldstein JL, Brown MS, Russell DW (1985a) The LDL receptor gene: a mosaic of exons shared with different proteins. Science 228:815–822

Sudhof TC, Russell DW, Goldstein JL, Brown MS, Sanchez-Pescador R, Bell GI (1985b) Cassette of eight exons shared by genes for LDL receptor and EGF precursor. Science 228:893–895

Suzuki T, Riggs AF (1993) Linker chain L1 of earthworm hemoglobin. Structure of gene and protein: homology with low density lipoprotein receptor. J Biol Chem 268:13548–13555

Suzuki T, Vinogradov SN (2003) Globin and linker sequences of the giant extracellular hemoglobin from the leech *Macrobdella decora*. J Protein Chem 22:231–242

Suzuki T, Takagi T, Gotoh T (1990a) Primary structure of two linker chains of the extracellular hemoglobin from the polychaete *Tylorrhynchus heterochaetus*. J Biol Chem 265:12168–12177

Suzuki T, Takagi T, Ohta S (1990b) Primary structure of a linker subunit of the tube worm 3000-kDa hemoglobin. J Biol Chem 265:1551–1555

Suzuki T, Ohta T, Yuasa HJ, Takagi T (1994) The giant extracellular hemoglobin from the polychaete *Neanthes diversicolor*. The cDNA-derived amino acid sequence of linker chain L2 and the exon/intron boundary conserved in linker genes. Biochim Biophys Acta 1217:291–296

Takahashi S, Kawarabayasi Y, Nakai T, Sakai J, Yamamoto T (1992) Rabbit very low density lipoprotein receptor: a low density lipoprotein receptor-like protein with distinct ligand specificity. Proc Natl Acad Sci USA 89:9252–9256

Tensen CP, Van Kesteren ER, Planta RJ, Cox KJ, Burke JF, van Heerikhuizen H, Vreugdenhil E (1994) A G protein-coupled receptor with low density lipoprotein-binding motifs suggests a role for lipoproteins in G-linked signal transduction. Proc Natl Acad Sci USA 91:4816–4820

van Driel IR, Goldstein JL, Sudhof TC, Brown MS (1987) First cysteine-rich repeat in ligand-binding domain of low density lipoprotein receptor binds Ca^{2+} and monoclonal antibodies, but not lipoproteins. J Biol Chem 262:17443–17449

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

Vinogradov SN, Lugo SD, Mainwaring MG, Kapp OH, Crewe AV (1986) Bracelet protein: a quaternary structure proposed for the giant extracellular hemoglobin of *Lumbricus terrestris*. Proc Natl Acad Sci USA 83:8034–8038

Vinogradov SN, Walz DA, Pohajdak B, Moens L, Kapp OH, Suzuki T, Trotman CN (1993) Adventitious variability? The amino acid sequences of nonvertebrate globins. Comp Biochem Physiol B 106:1–26

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

Yochem J, Greenwald I (1993) A gene for a low density lipoprotein receptor-related protein in the nematode *Caenorhabditis elegans*. Proc Natl Acad Sci USA 90:4572–4576

Zal F, Lallier FH, Wall JS, Vinogradov SN, Toulmond A (1996) The multi-hemoglobin system of the hydrothermal vent tube worm *Riftia pachyptila*. I. Reexamination of the number and masses of its constituents. J Biol Chem 271:8869–8874

Zal F, Green BN, Lallier FH, Toulmond A (1997a) Investigation by electrospray ionization mass spectrometry of the extracellular hemoglobin from the polychaete annelid *Alvinella pompejana*: an unusual hexagonal bilayer hemoglobin. Biochemistry 36:11777–11786

Zal F, Green BN, Lallier FH, Vinogradov SN, Toulmond A (1997b) Quaternary structure of the extracellular haemoglobin of the lugworm *Arenicola marina*: a multi-angle-laser-light-scattering and electrospray-ionisation-mass-spectrometry analysis. Eur J Biochem 243:85–92

Zhang J, Sun X, Qian Y, Maquat LE (1998a) Intron function in the nonsense-mediated decay of beta-globin mRNA: indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm. RNA 4:801–815