***Ascaris* hemoglobin gene: Plant-like structure reflects the ancestral globin gene**

(intron/exon/phylogeny/nematodes)

DAVID R. SHERMAN, ANDREW P. KLOEK, B. RAJENDRA KRISHNAN, BRETT GUINN,
AND DANIEL E. GOLDBERG

Departments of Molecular Microbiology and Medicine, Washington University School of Medicine, and The Jewish Hospital of St. Louis, Box 8230, 660 South Euclid Avenue, St. Louis, MO 63110

Communicated by Helen M. Ranney, September 8, 1992 (received for review July 27, 1992)

ABSTRACT Animal globin genes have two introns at strictly conserved positions, while plant globin genes have both of these as well as an additional, central intron. It has been proposed that a common ancestor gene had three introns, one of which was subsequently lost from animal but not plant globin genes. We have elucidated the cDNA sequence and gene structure of a hemoglobin from the parasitic nematode *Ascaris suum* and found a plant-like central intron, providing strong evidence for a three-intron ancestor of modern globin genes.

The three-dimensional structures of globins are strikingly similar despite extensive variation of amino acid sequence (1, 2). The array of helices that constitutes the globin protein fold has been conserved throughout the evolution of plants and animals (2, 3). Gene structure is similarly conserved. The globin genes of all but one animal studied to date contain three exons and two introns, the positions of which are invariant (4). Only the globin genes of *Chironomus thummi*, which are intron-less and may have derived from an unusual conversion event, are known to violate this pattern (5). The two introns found in animal globin genes are present in identical positions within plant globin genes (6, 7). In addition, plant globin genes have a third, central intron dividing the coding region into the structural modules predicted by Go (8). It has been proposed that the ancestral globin gene contained three introns, one of which was subsequently lost during animal evolution (4, 7, 8). Animal globin genes with only two introns were already established by the divergence of earthworms about 600 million years ago, suggesting that the putative ancestral third intron be sought in the globin genes of more primitive lineages (4).

The parasitic nematode *Ascaris suum* contains a perienteric hemoglobin of remarkably strong oxygen avidity (9–11). The protein exists as an octamer of $M_r \approx 328,000$ (12). Its amino acid sequence was recently determined (13), yet the structural features that strengthen oxygen binding are not evident. The *Ascaris* sequence retains the invariant amino acid residues found throughout globin phylogeny, although overall homology to vertebrate globins is only 10–15%. Each monomer contains two distinct, closely related heme-binding domains (12, 13). Two-domain globins are widespread among the invertebrates and are presumably products of a gene duplication event (14). To address issues of *Ascaris* globin structure, function, and evolution, we have undertaken an analysis of the *Ascaris* hemoglobin gene. We report here that the *Ascaris* perienteric hemoglobin gene possesses three introns per domain, as is characteristic of plant but not animal globin genes.*

MATERIALS AND METHODS

RNA was isolated from adult female *A. suum* (Carolina Biological Supply) by standard procedures (15). Briefly, ≈ 4 g of worm tissue was subjected to homogenization in 4 M guanidinium thiocyanate solution, followed by centrifugation at $2000 \times g$ for 10 min to pellet debris. Supernatants were layered onto cushions of 5.7 M CsCl solution and subjected to overnight ultracentrifugation in an SW50.1 rotor at 36,000 rpm and 15°C. RNA pellets were dissolved in 0.5 ml of sterile, distilled H₂O and stored at -78°C .

Ascaris hemoglobin cDNA was generated by 3' RACE technology (BRL) using tagged oligo(dT) and a gene-specific primer (with *Eco*RI linker) for the reverse transcriptase and PCRs (16). The gene-specific primer (5'-GAGCTATGCAT-GAAATCGCTCGAGCATGCCAAGGTTGA-3') is a best-guess oligonucleotide based on the N-terminal amino acid sequence of purified *Ascaris* hemoglobin, residues 5–17 (D.R.S., B.G., and D.E.G., unpublished data) and the codon preferences of the globin gene from the related organism *Pseudoterranova decipiens* (17). PCR was performed in a TC480 thermal cycler (Perkin-Elmer/Cetus) with Amplitaq DNA polymerase and cDNA made from 0.1 μg of total RNA. Final reaction conditions were 0.1 μM oligonucleotide primers, 10 mM Tris-HCl, 1.5 mM MgCl₂, 50 mM KCl, 0.1 mg of gelatin per ml, 0.2 mM deoxynucleotides (pH 8.3) in a final vol of 0.1 ml. The reactions were cycled 30 times at $94^\circ\text{C} \times 1$ min, $55^\circ\text{C} \times 1$ min, and $72^\circ\text{C} \times 2$ min. After ethanol precipitation, PCR products were digested at 37°C for 16 h with 3 units of primer-specific restriction enzymes *Eco*RI and *Not* I (Boehringer Mannheim). The DNA was then purified by adsorption to a glass slurry using the GeneClean II kit (Bio 101, La Jolla, CA) and subcloned into phagemid pBSSKII+ (Stratagene).

An *Ascaris* genomic library, constructed in λ GEM-11 (Promega), was provided by the Tim Nilsen laboratory (Case Western Reserve University, Cleveland). *Escherichia coli* KW251 (Promega) was transfected with the phage library as described by Sambrook *et al.* (18); 2.5×10^5 plaques were screened with a 39-mer best-guess oligonucleotide based on the amino acid sequence from amino acids 26–38 of the *Ascaris* globin. Five positive clones were isolated. The insert from one λ isolate was subcloned into the phagemid pBSSKII+ (Stratagene) for subsequent sequence analysis.

Intron sizes were determined by PCR using pairs of oligonucleotide primers corresponding to the coding sequence flanking both sides of each intron. The primers were chosen so that each set would amplify targets of approximately the same size from cDNA. PCR amplifications were performed on both purified genomic DNA and cDNA, and products were compared by 1% agarose gel electrophoresis. The validity of each PCR product was confirmed by sequencing

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*The sequence reported in this paper has been deposited in the GenBank data base (accession no. L03351).

(see below). Intron length was calculated by subtracting the size of each cDNA PCR product from that of its corresponding genomic PCR product. The oligonucleotide pairs used were as follows: intron AB, 5'-TTCACGATTCTTGAAGTATTTCCT-3' and 5'-GAGCTATGCATGAAATCGCTCGAGCATGCCAAGGTTGA-3'; intron BC, 5'-TAGTTCCTCTAGTGTAGGCG-3' and 5'-AGGAAATACTTCAAGAATCGTGA-3'; intron CD, 5'-GGGCTCATCGAGTGTCTGT-3' and 5'-CGCCTACACTAGAGAACTA-3'; intron DA', 5'-GCAGGCTTGGCATGAAATCGG-3' and 5'-GCAGTTTCACTATGCCCAATGTCTG-3'; intron A'B', 5'-GTCCTCGGCTGTGTAGTTC-3' and 5'-CGACATTGGGCATAGTGAACTGC-3'; intron B'C', 5'-GAACACACAGCCGAGGAC-3' and 5'-CATGGACGTACATATGAAAC-3'; intron C'D', 5'-GTTTCATATGTACGTCCATG-3' and 5'-GTGCTCACACAAGTGCAC-3'. Intron PA was sized by sequencing it in its entirety.

Plasmid DNA sequencing was performed with the Sequenase kit (United States Biochemical), used according to the manufacturer's specifications, and the products were resolved on 6% polyacrylamide gels (National Diagnostics, Manville

NJ). λ DNA and PCR products were subjected to direct sequencing with a fmol DNA sequencing kit (Promega), with the following modifications: 3 times the suggested amounts of primer and radiolabeled ATP, 7 units of T4 polynucleotide kinase (New England Biolabs), and 4 times the suggested amount of Promega *Taq* DNA polymerase were used. Two individual cDNA clones were sequenced in their entirety to avoid errors and were confirmed with genomic sequence. Sequence determination of genomic versions of both domains was accomplished with amplified genomic DNA as described above. In addition, the sequence of domain one was confirmed with the genomic subclone. Genomic sequence was obtained across each exon-intron boundary; a consensus splice site was obtained in each case. There was a difference between the genomic and cDNA sequences at cDNA position 328. The cDNA sequence read CTG (leucine), while the genomic sequence at this position read TTG (also leucine).

RESULTS AND DISCUSSION

To elucidate *Ascaris* hemoglobin gene structure, cDNA and genomic copies of the gene were isolated and compared. Fig.

FIG. 1. cDNA and deduced amino acid sequence of *Ascaris* hemoglobin gene. Positions of introns in the genomic sequence are marked with arrows.

1 shows the cDNA sequence, which consists of a 1014-nucleotide open reading frame coding for an 18-amino acid hydrophobic prepiece followed by a 320-amino acid mature protein. Sequencing of the genomic DNA revealed a total of eight introns interrupting the coding sequence (arrows in Fig. 1). The size of each intron was determined by PCR using oligonucleotides corresponding to the coding sequences flanking each intron (Fig. 2). The size difference between corresponding products made from genomic and cDNA templates is a measure of intron size (tabulated in Table 1 and schematized in Fig. 3a). When aligned by the rules of Bashford *et al.* (3), each domain has two introns that map precisely to the invariant positions common to all animal and plant globin genes (Fig. 3b). In addition, each domain contains a third, central intron heretofore reported only in plant globin genes (Fig. 3c). The sequence of this portion of the *Ascaris* protein displays little similarity to plant hemoglobins, making precise alignment of these regions difficult. However, the *Ascaris* intron maps to the middle of the sequence encoding the E helix of other globins (asterisk in Fig. 4), as does the analogous plant intron. The ancestral globin gene has been predicted to contain an intron in precisely this position (8).

Two other introns are present in the *Ascaris* hemoglobin gene. The two domains of the gene are separated by a bridge intron; such a structure has previously been reported for the two-domain globin gene of the clam *Barbatia reeveana* (14). Furthermore, an intron separates the DNA encoding the signal sequence and two amino acids of the mature protein from the body of the gene. Many secretory protein genes in eukaryotes have an intron near the junction of the DNA encoding the signal sequence with that of the mature protein (20, 21). However, such an intron is not present in the genes of the other secreted globins analyzed thus far (4, 5).

FIG. 2. PCR analysis of introns in *Ascaris* hemoglobin gene. Purified genomic DNA and cDNA templates were amplified by using oligonucleotide primers designed from coding sequence flanking the intron sites. Shown here is the pattern produced by primers from the second heme-binding domain. A similar pattern was observed with the primers corresponding to the first heme-binding domain. Lanes: 1, PCR products using sequences flanking intron DA'; 2, PCR products using sequences flanking intron A'B'; 3, PCR products using sequences flanking intron B'C'; 4, PCR products using sequences flanking intron C'D'. Intron nomenclature is depicted in Fig. 3. Lanes marked c and g, cDNA and genomic templates, respectively. Markers (100-base-pair and 1-kilobase ladder standards; BRL) are labeled in kilobases.

Table 1. Position and size of *Ascaris* hemoglobin gene introns

	Intron							
	PA	AB	BC	CD	DA'	A'B'	B'C'	C'D'
Position in cDNA	61	149	247	381	508	596	694	828
Size, bp	350	570	470	1500	2300	380	920	620

Introns are named by the exons they divide using the nomenclature of Fig. 3. Positions were determined by analysis of genomic and cDNA sequences as in Fig. 1. Sizes were determined by PCR analysis with oligonucleotides from the cDNA sequences that flank the introns (see Fig. 2) and are accurate to within 50 base pairs (bp).

The *Ascaris* hemoglobin cDNA sequence (Fig. 1) largely confirms the protein sequence of De Baere *et al.* (13). Three discrepancies with the previously reported protein sequence could not be resolved. These are (i) aspartic acid instead of leucine at amino acid 116, (ii) a previously unreported tryptic tetrapeptide His-Ala-Val-Arg between Arg-166 and His-171, and (iii) a previously unreported pentapeptide Lys-Glu-Glu-Gln-His at the C terminus of the protein (amino acids 334–338). These disagreements could reflect strain differences or the presence of multiple hemoglobin genes in *Ascaris*. Of note, our nucleotide sequence predicts one more copy of the His-Lys-Glu-Glu charged repeat than the published amino acid sequence. Perutz and co-workers have postulated that the repeat units, which alternate positive and negative charges, may mediate the octamerization of the *Ascaris* protein by forming a β -barrel of alternating antiparallel strands (13).

The deduced *Ascaris* hemoglobin protein sequence can be compared to that of a related parasitic nematode, the anisakid *P. decipiens* (17). Each has an 18-amino acid prepiece; these putative leader peptides are 50% identical. As shown in Fig. 4, the proteins share extensive amino acid homology in their two heme-binding domains—73% identity in the first domain and 58% identity in the second domain. However, there is less homology in their C-terminal tails. In this region, there are four repeats of the His-Lys-Glu-Glu sequence in *Ascaris* hemoglobin but only one in the *Pseudoterranova* protein. Indeed, these multiple repeats appear unique to *Ascaris* hemoglobin and may form the basis of its exceptional oxygen avidity. If a stable, charge-aligned structure can only form in the oxy- state, this conformation would be greatly favored and oxygen affinity would be enhanced. Site-directed mutagenesis will address this possibility. In this regard, it is noteworthy that the hemoglobin of *Anisakis physeteris*, a parasitic nematode very closely related to *Pseudoterranova*, has a poor oxygen avidity when compared with *Ascaris* hemoglobin (22). It would be valuable to know the oxygen avidity of the *Pseudoterranova* protein and whether it can form multimers despite its deficiency of charged repeats.

Plants and animals apparently diverged 1–1.4 billion years ago (6) and nematodes first appeared at about the same time (23). The presence of a plant-like central intron in each domain of the *Ascaris* hemoglobin gene suggests that the ancestral globin gene structure is at least 1 billion years old. In addition to this similarity, plants and nematodes share other characteristics. For example, the glyoxylate cycle is present only in plants, microorganisms, and the early stages of nematode development; also, only plants, protozoa, and nematodes are known to synthesize polyunsaturated fatty acids *de novo* (23). Might plants and nematodes share a common ancestor distinct from the main body of animal evolution? It will be interesting to examine the globin genes of more divergent eukaryotes, such as the protozoa, for the presence of the central intron. We cannot exclude the possibility that the central introns of plants and *Ascaris* arose from independent insertion events. It seems unlikely, however, that the insertion would occur independently in the

FIG. 3. Schematic representation of *Ascaris* hemoglobin gene. (a) *Ascaris* hemoglobin gene with exons and introns drawn approximately to scale. Exons are represented by solid rectangles; exon sizes are shown above each rectangle. Introns are represented by thin lines; intron sizes are shown below each line. (b) Comparison of amino acid sequences around the proximal (P) and distal (D) intron sites of sperm whale myoglobin (SWM; residues 23–40 and 93–110), human α -globin (HUA; residues 23–40 and 87–104), soybean leghemoglobin (SOY; residues 24–41 and 92–109), and *Ascaris* globin domain 1 (ASC; residues 42–59 and 114–132). Sequences are aligned for maximum homology based on previous alignments (1, 19). Horizontal lines with asterisk represent intron insertion sites. (c) Nonproportional representation of *Ascaris* hemoglobin gene. Structure of plant and vertebrate globin genes is provided for comparison. Sizes of individual introns and exons are not to scale.

same position in both groups of organisms but not in any of the other organisms whose globin gene structures have been elucidated.

It is clear from the common features of amino acid sequence, tertiary protein structure, and gene structure of plant and animal globins that this protein predates the evolution of either. What was the function of this primordial globin? The primitive environment was one of low oxygen tension. Many organisms, such as *E. coli* (24), *Alcaligenes* (25), *Vitreoscilla* (26, 27), and *Candida* (28), that experience at least transient oxygen deficiency today elaborate globins that both bind oxygen and exhibit catalytic reductase activities (29) (the *Vitreoscilla* reductase is an associated but separate polypeptide). The hemoglobin of *Ascaris*, which lives in the oxygen-poor intestinal folds, has a reductase activity as well (30). In the two cases studied to date, *Ascaris* and *Candida*, the hemoglobins have extraordinarily high oxygen affinities (9–11, 31). It thus seems possible that globins evolved as molecules to sequester scarce oxygen directly for use in

biosynthetic reactions requiring molecular oxygen. Perhaps the transport function of modern hemoglobins arose only after oxygen became more available and eukaryotes gained in complexity. Clues to their original function may come from globins, like that of *Ascaris*, whose physiological role has not yet been established.

Note. Dixon *et al.* (32) have recently reported a three-intron structure for the first domain of the globin from the related nematode *P. decipiens*. Interestingly, the second domain of this gene contains only one intron, corresponding to the first intron of other globins.

We are grateful to Drs. Gary Ackers, Alan Templeton, and Igor Brikun for helpful discussions. Drs. Hon Ip and Tim Nilsen generously provided the genomic library and *Ascaris* genomic DNA. Aron Knickerbocker provided invaluable technical assistance. This work was supported by a grant to D.E.G. from the Lucille P. Markey Charitable Trust. B.R.K. was supported by grants from the Public Health Service (HG00563) and the Department of Energy (DEFG0289ER60862) to D. E. Berg.

	A	-----	B	-----	*	C	-----	D	-----	E			
ASCDOM1	NKTREL	CMKSL	EHAKV	DTSE	ARQD	GIDLY	KHMF	ENYPP	LRKY	FKNREE	YTAEDV	QNDP	
ASCDOM2	HAVRHQ	CMRSL	QHID	IGHSE	TAKNG	IDLYK	HMFEN	YPSM	REAFK	DRENY	TAEDV	QKDP	
PDDOM1	SKTREL	CMKSL	EHAKV	GTSE	AKQD	GIDLY	KHMF	EHYP	PAMKK	YFKHRE	NYTPA	DVQKDP	
PDDOM2	HSVRDH	CMNSL	EYIAI	GDK	EHQK	QNGID	LYK	HMF	EHYP	PMRKA	FKGREN	FTKEDV	QKDA

	*	-----	F	-----	G	-----	*							
ASCDOM1	FFAKQ	GQKIL	LACHV	LCATY	DDRET	FNAYT	RELLDR	HARDH	VHMP	PEVW	TD	FWKL	FEFY	
ASCDOM2	FFAKQ	GQKIL	LACHV	LCATY	DDRET	FNAYT	RELLDR	HARDH	VHMP	PEVW	TD	FWKL	FEFY	
PDDOM1	FFIKQ	GQKIL	LACHV	LCATY	DDRET	FNAYT	RELLDR	HARDH	VHMP	PEVW	TD	FWKL	FEFY	
PDDOM2	FFVKQ	GQKIL	LALML	CSSY	DDPT	FDYF	VDAL	MDRHI	KDDI	HL	PQ	QW	HEFWKL	FAFY

	-----	H	-----			
ASCDOM1	LGKKT	T--L	DEPTKQ	AWHEIG	REFAKE	INKHGR
ASCDOM2	LEKKSH	--L	CEHTKH	AWVIG	KEFAYE	ATRHGKEHHEKE--EHKEEHKEEHKEEQH
PDDOM1	LGSKT	T--L	DEPTKQ	AWHEIG	KEFSHE	ISHHGR
PDDOM2	LNEKSH	QHLT	EA	EKHAW	STIGED	FAHEADKHAKAEKDHHGEHKEEH

FIG. 4. Deduced amino acid sequence of *Ascaris* (ASC) hemoglobin gene compared with that of *P. decipiens* (PD) (17). Letters above sequences denote positions of helices in the corresponding regions of human hemoglobin (13). *Ascaris* gene intron positions are marked with asterisks.

1. Dickerson, R. E. & Geis, I. (1983) *Hemoglobin* (Benjamin/Cummings, Menlo Park, CA).
2. Arutyunyan, E. G. (1981) *Mol. Biol.* **15**, 27–44.
3. Bashford, D., Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196**, 199–216.
4. Jhiang, S. M., Garey, J. R. & Riggs, A. F. (1988) *Science* **240**, 334–336.
5. Antoine, M. & Niessing, J. (1984) *Nature (London)* **310**, 795–797.
6. Brown, G. G. (1984) *J. Mol. Evol.* **21**, 19–32.
7. Jensen, E. O., Paludan, K., Hyldig-Nielsen, J. J., Jorgensen, P. & Marcker, K. A. (1981) *Nature (London)* **291**, 677–679.
8. Go, M. (1981) *Nature (London)* **291**, 90–92.
9. Gibson, Q. H. & Smith, M. H. (1965) *Proc. R. Soc. London Ser. B* **163**, 206–214.
10. Okazaki, T. & Wittenberg, J. B. (1965) *Biochim. Biophys. Acta* **111**, 503–511.
11. Davenport, H. E. (1949) *Proc. R. Soc. London Ser. B* **136**, 255–270.
12. Darawshe, S., Tsafadyah, Y. & Daniel, E. (1987) *Biochem. J.* **242**, 689–694.
13. De Baere, I., Liu, L., Moens, L., Van Beeuman, J., Gielens, C., Richelle, J., Trotman, C., Finch, J., Gerstein, M. & Perutz, M. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4638–4642.
14. Naito, Y., Riggs, C. K., Vandergon, T. L. & Riggs, A. F. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 6672–6676.
15. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **18**, 5294–5299.
16. Frohman, M. A., Dush, M. K. & Martin, G. R. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 8998–8992.
17. Dixon, B., Walker, B., Kimmins, W. & Pohajdak, B. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 5655–5659.
18. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY), pp. 2.108–2.125.
19. Appleby, C. A., Bogusz, D., Dennis, E. S. & Peacock, W. J. (1988) *Plant Cell Environ.* **11**, 359–367.
20. Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–373.
21. Gilbert, W. (1985) *Science* **228**, 823–824.
22. Viglierchio, D. R. & Gortz, J. H. (1972) *Exp. Parasitol.* **32**, 211–216.
23. Vanfleteren, J. R., Evers, E. A. I. M., Van de Werken, G. & Van Beeuman, J. J. (1990) *Biochem. J.* **271**, 613–620.
24. Vasudevan, S. G., Armarego, W. L. F., Shaw, D. C., Lilley, P. E., Dixon, N. E. & Poole, R. K. (1991) *Mol. Gen. Genet.* **226**, 49–58.
25. Weihs, V., Schmidt, K., Schneider, B. & Friedrich, B. (1989) *Arch. Microbiol.* **151**, 546–550.
26. Dikshit, K. L., Spaulding, D., Braun, A. & Webster, D. A. (1989) *J. Gen. Microbiol.* **135**, 2601–2609.
27. Wakabayashi, S., Matsubara, H. & Webster, D. A. (1986) *Nature (London)* **322**, 481–483.
28. Oshino, R., Asakura, T., Takio, K., Oshino, N. & Chance, B. (1973) *Eur. J. Biochem.* **39**, 581–590.
29. Zhu, H. & Riggs, A. F. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 5015–5019.
30. Sherman, D. R., Guinn, B., Perdok, M. M. & Goldberg, D. E. (1992) *Science*, in press.
31. Oshino, R., Asakura, T., Tamura, M., Oshino, N. & Chance, B. (1972) *Biochem. Biophys. Res. Commun.* **46**, 1055–1060.
32. Dixon, B., Walker, B., Kimmins, W. & Pohajdak, B. (1992) *J. Mol. Evol.* **35**, 131–136.