

## ***Caenorhabditis* Globin Genes: Rapid Intronic Divergence Contrasts with Conservation of Silent Exonic Sites**

**Andrew P. Kloek,<sup>1</sup> James P. McCarter,<sup>2</sup> Robert A. Setterquist,<sup>3</sup> Tim Schedl,<sup>2</sup> Daniel E. Goldberg<sup>1</sup>**

<sup>1</sup> Departments of Molecular Microbiology and Medicine, Howard Hughes Medical Institute, Washington University School of Medicine, 600 S. Euclid Ave., St. Louis, MO 63110, USA

<sup>2</sup> Department of Genetics, Washington University School of Medicine, 660 S. Euclid Ave., St. Louis, MO 63110, USA

<sup>3</sup> Department of Biochemical and Biophysical Sciences, University of Houston, Houston, TX 77204-5934, USA

Received: 6 November 1995 / Accepted: 12 February 1996

**Abstract.** Globin genes from the *Caenorhabditis* species *briggsae* and *remanei* were identified and compared with a previously described *C. elegans* globin gene. The encoded globins share between 86% and 93% amino acid identity, with most of the changes in or just before the putative B helix. *C. remanei* was found to have two globin alleles, *Crg1-1* and *Crg1-2*. The coding sequence for each is interrupted by a single intron in the same position. The exons of the two genes are only 1% divergent at the nucleotide level and encode identical polypeptides. In contrast, intron sequence divergence is 16% and numerous insertions and deletions have significantly altered the size and content of both introns. Genetic crosses show that *Crg1-1* and *Crg1-2* segregate as alleles. Homozygous lines for each allele were constructed and northern analysis confirmed the expression of both alleles. These data reveal an unusual situation wherein two alleles encoding identical proteins have diverged much more rapidly in their introns than the silent sites of their coding sequences, suggesting multiple gene conversion events.

**Key words:** Alleles — Evolution — Gene conversion — Introns — Nematode

### **Introduction**

Globins are widespread throughout phylogeny. Comparison of globin amino acid sequences across vertebrates, plants, and nematodes shows sequence identity as low as 10% (Bashford et al. 1987). Yet the tertiary structures of these globins are very much alike (Arutyunyan 1981, Dickerson and Geis 1983). The recently solved crystal structure of *Ascaris* hemoglobin domain 1 appears strikingly similar to that of sperm whale myoglobin despite only 15% amino acid homology (Yang et al. 1995). Vertebrate globins have been studied extensively; consequently their structures and functions are well understood (Dickerson and Geis 1983). The globins of invertebrates are much more diverse, both in structure and affinity for oxygen. Nematode globins, which appear to have arisen from a common metazoan ancestor, are a particularly unusual lot, characterized by tight oxygen binding and diverse quaternary structure (Blaxter 1993). They can be divided into three groups consisting of perenteric hemoglobins, body-wall myoglobins, and cuticular globins (Blaxter 1993). Homology across groups is typically 30–50%; members of the same group from different species are more closely related than members of different groups from the same species. The extracellular hemoglobin of *Ascaris* is perhaps the best known, due to its extremely high oxygen avidity (Davenport 1949; Gibson and Smith 1965; Okazaki and Wittenberg 1965). Mutational analysis (De Baere et al. 1994; Kloek et al. 1994) and crystal structure solution (Yang et al. 1995)

have illuminated the molecular mechanism of this tight oxygen interaction, but a physiological function for this enigmatic molecule remains undetermined, as do the roles of all nematode globins.

Nematode globin gene structures are equally perplexing. The two introns shared by both plant and animal globin genes (Brown et al. 1984; Go 1981; Jensen et al. 1981) are generally accepted as remnants from the primordial globin gene, but the history of the middle intron found in plants and subsequently in other kingdoms is unclear (Dixon and Pohajdak 1992; Dixon et al. 1992; Moens et al. 1992; Sherman et al. 1992; Stoltzfus and Doolittle 1993). Hemoglobin genes from the nematodes *Ascaris suum* (De Baere et al. 1992; Sherman et al. 1992) and *Pseudoterranova decipiens* (Dixon et al. 1992) encode unusual two-domain polypeptides in which the gene sequences have the plant-like three-intron structure. The *C. elegans* globin gene encodes a one-domain globin (Kloek et al. 1993; Mansell et al. 1993; Sulston et al. 1992). Its lack of a secretory leader sequence and homology to the *Ascaris* myoglobin (Blaxter et al. 1994) suggest it is an intracellular body-wall globin. Unexpectedly, the *C. elegans* globin gene contains only a middle intron; neither of the two introns conserved throughout vertebrate and plant globin genes is present. The middle intron interrupts the sequence encoding the E helix in each case, but its precise location is shifted between *Ascaris*, *Pseudoterranova*, *Caenorhabditis*, plants, and certain unicellular organisms (Stoltzfus et al. 1994; Yamauchi et al. 1992). Whether these middle introns are ancient relatives derived from a primordial globin gene or multiple independent insertion events in the same region is the subject of much debate.

We report here a comparison of *Caenorhabditis* globin genes. *C. remanei* has been found to have two alleles of a globin gene that have divergent introns, yet contain highly conserved silent coding positions.

## Materials and Methods

**Strains and Genbank Accession Numbers.** *C. briggsae* strain AF16 and *C. remanei* strain CB5161 were used in this study. It should be noted that the species nomenclature is currently being revised.

The sequences reported in this paper have been deposited in the GenBank data base under the following accession numbers: *Cbg1* cDNA, U48289; *Cbg1* genomic, U48290 and U48291; *Crg1-1* cDNA, U48292; *Crg1-1* genomic, U48294; *Crg1-2* cDNA, U48293; and *Crg1-2* genomic, U48295.

**PCR Amplification of Globin Gene Fragments.** Globin gene fragments from *C. briggsae* (strain AF16) and *C. remanei* (CB5161) were obtained by PCR using oligonucleotides designed from known *C. elegans* globin gene sequences. Forward oligo 5' atgctgatgaaccctcaag 3' (derived from *C. elegans* codons 1–6) and reverse oligo 5' agcattgaattcccttcc 3' (derived from *C. elegans* codons 140–145) were effective in amplifying globin gene fragments from each species. Final PCR conditions were as follows: 1.25 units AmpliTaq (Perkin Elmer Cetus), 0.1  $\mu$ M oligonucleotide primers, 250 ng of genomic DNA, 20 mM

Tris-HCl (pH 8.2), 10 mM KCl, 6 mM  $(\text{NH}_4)_2\text{SO}_4$ , 2 mM  $\text{MgCl}_2$ , 0.1% Triton X-100, 10  $\mu$ g/ml nuclease-free bovine serum albumin, and 0.2 mM deoxynucleotides in a final volume of 0.1 ml. Reactions were cycled 30 times at 95°C for 1 min, 42–48°C for 1 min, and 72°C for 2 min.

**Cloning and Sequencing of PCR Products.** PCR products were cloned directly using the TA Cloning Kit (Invitrogen). Plasmid DNA sequencing was performed with the Sequenase Kit (United States Biochemical Corp.) and the products were resolved on 6% polyacrylamide gels (National Diagnostics). All sequences reported here were confirmed from at least two independent clones except for the 5' untranslated regions and the first three codons for each gene, which were sequenced once. The coding conservation in this region of the four genes (Met, Ser, Met) and successful amplification of gene fragments using oligonucleotides designed from these sequences support the accuracy of the nucleotides reported for this short region.

**Nucleic Acid Isolation and Northern Analysis.** Genomic DNA was isolated by standard protocols (Wood 1988). Total RNA isolation followed the procedure used by Goetinck and Waterston (Goetinck and Waterston 1994). Total *C. remanei* RNA was fractionated on a 2.2 M formaldehyde–1.2% agarose gel (Sambrook et al. 1989). The RNA was transferred to a nylon membrane (Micron Separations Inc.) and probed with radiolabeled *C. remanei* globin cDNA. The blot was washed to a stringency of 60°C in 0.25 $\times$  SSC (SSC is 0.15 M NaCl/0.015 M  $\text{Na}_3\text{ citrate}$  pH 7.6)/0.1% SDS and analyzed by autoradiography. RNA markers (Gibco BRL) were used to calibrate the gel.

**5' and 3' RACE.** The 5' and 3' RACE systems (Gibco BRL) were used according to manufacturer's specifications to amplify overlapping fragments representing full-length cDNA for each gene (Frohman et al. 1988); 5' RACE was performed only on *C. elegans* RNA, and an SL1 primer 5' TTTAATTACCCAAGTTTGAAG 3' was utilized in PCR to obtain the 5' ends of *Crg1-2* and *Cbg1*.

**Genetic Crosses.** Genetic crosses were performed to construct *C. remanei* lines homozygous for *Crg1-1* and *Crg1.2*. Briefly, single pairs of worms consisting of one male and one sexually immature female (larval stage L4) were placed together on NGM plates prepared with *E. coli* strain OP50 and stored at 20°C until the appearance of progeny. The adults were scored for globin genotype by single worm PCR and cataloged. Crosses involving two parent worms homozygous for the same allele were noted and several of their progeny were screened to ensure the exclusion of the other globin allele from the line.

**Single-Worm PCR.** To assess the globin genotypes of *C. remanei* individuals, single-worm PCR was performed (Williams et al. 1992). Briefly, individual worms were placed in 2.5  $\mu$ l of lysis buffer (10 mM Tris-HCl [pH 8], 60  $\mu$ g/ml proteinase K, 50 mM KCl, 2.5 mM  $\text{MgCl}_2$ , 0.45% Tween 20, and 0.05% gelatin) and frozen at –70°C for 15 min. The frozen pellet was overlaid with 60  $\mu$ l of mineral oil and incubated 1 h at 60°C and then 15 min at 95°C. The reaction was then cooled to 4°C and a standard PCR protocol was followed.

## Results

### *Caenorhabditis* Globin Gene Comparison

Oligonucleotides (see Materials and Methods) based on a known sequence from the *C. elegans* globin gene (Kloek et al. 1993; Mansell et al. 1993; Sulston et al. 1992) were used as primers for PCR using *C. remanei* and *C. briggsae*.

```

          AAAAAAAAAAAAAA      BBBBBBBBBBBBBBBB
C. elegans  MSMNRQEIISDLQVKSLEGRMVGTEAQNIEGNFAFYRYFFT
C. briggsae  ...T...Q.....EK...TDKG.A...G...Q....
C. remanei  ...S...Q.....EK.....K.VD...G...Q....

          BCCCCCC      DDDDDDD      EEEEEEEEEEEEEEE
C. elegans  NFPDLRVYFKGAEKYTADDVKKSERFDKQQRILLACHLL
C. briggsae  .....F.E.....I
C. remanei  .....E.....I

          EEE      FFFFFFFF      GGGGGGGGGGGGG
C. elegans  ANVYTNIEEVFKGYVRETIINRHRIVKMDPALWMAFFTIVFTG
C. briggsae  ...F.....A.....
C. remanei  .....A.....V.....

          GGGGGG      HHHHHHHHHHHHHHHHHHH
C. elegans  YLESVGLNDNQKAAWALGKEFNAESQTHLKNLPHV
C. briggsae  ...T.S.....C.V.....Y.H
C. remanei  ..G.T.S.T.....C.E.....Y.H

```

**Fig. 1.** Amino acid sequence comparison of *Caenorhabditis* globins. Dots indicate identity. The eight putative  $\alpha$ -helices (A–H) are indicated by italicized letters above amino acid sequences and are derived from alignment with the *Ascaris* hemoglobin domain 1 structure.

*sae* genomic DNA as template. These reactions produced a single band for *C. briggsae* and two bands differing in size by roughly 200 bases for *C. remanei*. DNA sequencing of the cloned PCR products identified all three as globin gene fragments. Full-length cDNA sequences were generated from overlapping 5' and 3' RACE clones. The two *C. remanei* genes, *Crg1-1* (*Caenorhabditis remanei* globin) and *Crg1-2*, as well as the *C. briggsae* gene, *Cbg1* (*Caenorhabditis briggsae* globin), were found to be homologous to the previously reported *C. elegans* globin gene, in each case consisting of a 480-bp open reading frame interrupted after nucleotide 194 by a single intron. Compared to the *C. elegans* globin gene intron of 298 nucleotides, the *Cbg1* intron is approximately 700 nucleotides, while the *Crg1-1* and *Crg1-2* introns are 747 and 939 nucleotides, respectively, accounting for the PCR doublet. The intron sizes are unusually large for *Caenorhabditis* genes, where the typical intron length is only about 50 nucleotides (Fields 1990). The deduced amino acid sequences of *C. briggsae*, *C. remanei*, and the previously described *C. elegans* globin are aligned in Fig. 1. Remarkably, the two *C. remanei* gene encode identical globins. The *C. briggsae* and *C. remanei* globins are 93% identical to each other and 86% to 87% identical to the *C. elegans* globin. Notably, amino acid changes are clustered in and just before the sequence assigned to the B helix in the globin from the related nematode *Ascaris* (Yang et al. 1995). This is somewhat unexpected, as the B helix is known to line the critical oxygen-binding pocket and play a role in interacting with ligand to enhance oxygen avidity in the *Ascaris* hemoglobin. It is possible that this helix plays a largely structural role, placing the B10 tyrosine in appropriate position to hydrogen-bond to heme-bound oxygen. Substitution of side chains may be well tolerated as long as they do not interfere with this positioning. Moreover, many of the replacements spread throughout these genes are conservative.

Employing the *Ascaris* D1 crystal structure (Yang et al. 1995) and corresponding mutagenesis studies (De

Baere et al. 1994; Kloek et al. 1994) as a comparison model for the *Caenorhabditis* globins, none of the key distal pocket residues or those making critical heme contacts have been altered. Assuming that these very similar globins still function in identical roles, the sites of amino acid changes probably identify positions in these globins which are most flexible to change while still maintaining physiological function.

### Comparison of the Two *C. remanei* Globin Genes

The two *C. remanei* globin genes are aligned in Fig. 2A. The exon nucleotide sequences are 1% divergent, containing six silent third position changes (4% divergent at synonymous sites) and thus encoding identical 160-amino-acid globin polypeptides. Enumerating each insertion and deletion as one event, equivalent to a single nucleotide change, intron divergence is 16%. Further, the size and prevalence of insertions and deletions has significantly altered the sequence content of the two introns. Figure 2B, shows a graphic comparison of *Crg1-1* and *Crg1-2*. Since it is impossible to know whether these changes were insertions in one intron or deletions in the other, all the variations are shown as insertions.

### Comparison of 5' and 3' Untranslated Sequences

The SL1 trans-spliced leader common to many nematode messages was found at the 5' end of the *C. elegans* globin message by 5' RACE (data not shown). The SL1 sequence is separated from the putative initiator methionine codon by only six nucleotides. PCR performed on cDNA from *C. remanei* and *C. briggsae* using an oligonucleotide containing the SL1 sequence was used to amplify the 5' ends of the *Cbg1* and *Crg1-2*. These messages are similarly trans-spliced very close to the initiator codon, with four nucleotides separating SL1 from coding sequence in *Cbg1* and 7 nucleotides in *Crg1-2*.

Past studies have indicated that *C. elegans*, *C. remanei*, and *C. briggsae* have existed as separate species long enough such that unconstrained sequences have diverged considerably, and most highly conserved sequences are actively transcribed or represent functional regulatory elements (Heine and Blumenthal 1986; Heschl and Baillie 1990; Kennedy et al. 1993; Prasad and Baillie 1989; Thomas and Wilson 1991; Zucker-Aprison and Blumenthal 1989). Comparison of the 3' untranslated sequences of these globin genes provides an opportunity to identify potential regulatory elements. The globin gene 3' untranslated regions range in size from 80 to 140 nucleotides and share little sequence homology between species. The sequences were aligned and examined for conserved motifs. The consensus sequence aatgagctttga, containing the four-base palindrome agct, could be identified in each gene 40–45 nucleotides upstream from the putative polyadenylation sequence

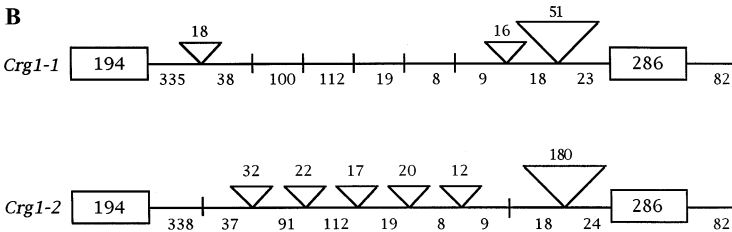
**A**

.....	86
ATGTCGATGAGCCGTCAGAGATCCAAGACTTGTGTCTCAAGTCTTTGGAAAGAGAAAATGGTTGGAACTGAGGCCGAAAAACCTGGA	86
.....T.....C.....	172
TAA CGGCAATGGATTCTATCAATACTTCTTCCACCACTTCCCGATCTTCGCGTCTATTTC AAGGGTGC GAAAAAGTATACCGCCG	172
.....--..a..g..t..t...cc...aa...g...g...t..t..a.....t	255
AGGATGTGAAAGAGAGTGAAGGtgagaagtactattggttttggcttccttagaataactacgaaacagttagcctaacaag	258
..a..g...t.....a..g..c...g.....-.....a...c...t.....	340
gggattttctctcacatggattaggaataaagttagctctccctgtgagtgattctgaaa-tttctagccctcactggagcctga	343
-.....t.....a...-g.....c...t.....	424
gagtcctcgaatcttctctcttgatttccaatgctatgtcc-atatggcgagcttcagatgaagatttccagactcttaggctcc	428
...a.....t..t..c.....t...a.....c.....a..a..cg...c.....a.....t....	510
agtgagggtggaatcttgagaatccac-cacagggacagctaatttcaggccaacaatttgtagaagttcctaaaaatcgtgag	513
...a...-..t.....ttagcगतtagtgataca.....aa.....t...t...g.....	585
aaaagcatttcc-ttgtgag-----ccttctag-gataaagtcagaggttatccggcaaaaaatgagtgcgtt	579
.....t.....a...tc..at.caaa...c...g...-	648
attgcctgctcaagcgctattattgctgctcaagcccaatttggcactcgggcaat-gg-agacct---agt-ggatataga	658
.ga...t.g...c.aaa.--a...g.t...a.....-.....t....a..a..g.....g--....	708
aattttagaattta--tgtgagat-aatccct-tgaaaaagatgaaaaagatccctgactacagggttcgttaacacgaaaaaaa	739
.t...t..g.....a..a.....a..a..c.....t..a..t.....a...t...g.....	794
aagcgcttaagagctaaaaaatccatttttccaaaaatttt-ggtttttgttca-tttttgagtttttctgtaaaatgagcaa	823
.....g.....	831
caaatgatttctagtgtctcacaaggttttcagcagattacaataactcacaacgtgtgcctgtaaaaagagacaaaaaat	909
..aaaaaaaaataaacctc...c.....c...XXX.....t.....-	964
aa-----gcccctgtgactaaaaaXXXtttttgcaaaaactgttttttccagATTGACAAAACAAGGTCAACGTA	1156
.....C.....A.....	1050
TCCCTCTGCTGCCATCTCATCGCTAATGTCTACCCAACGAGGAAGTGTTC AAGGCGTACGTTTCGTGAAACCGTCAACCGTCAC	1242
.....	1136
CGTATTTACAAAATGGACCCAGCCCTCTGGATGGCCTTTTTCACGTGTTCACCGGATATCTCGGATCAACTGGTTCGTTGACTGA	1328
.....C..A.....	1222
TCAACAGAAAAGCTGCATGGATGGCACTCGGAAAAGAAITCAATGCCGAATGTCAGGAGCATCTAAAGAACFCGAATCTTCCTTATG	1414
.....tgct.....g...a.....a.....	1302
TCCACTAaaaaaaatggatatttctgatt--gaaagatatttcaagagctttgaaacatgcatttatgattgtttactctcaa	1497
.....	1313
ttttgtacagaataaaactgtatccga	1524

XXX: *Crg1-1*  
 ttgggtttatgatctgaaattgaaaataaaacaccgatttcttattttt

*Crg1-2*  
 cctttccgaaacgggttacgggtcctgaaaagcgtacaaccctttccatttttagtgagcgt  
 atggtctaaaaatgagattgatcaaatataattcttccagattgaaaaataatttccgctg  
 atcctttctagctttgaaaatgcattaagaagaaaaaaacctgagcagagttcgtg

**B**



**Fig. 2.** Comparison of *Crg1-1* and *Crg1-2*. **A** Nucleotide alignment of *Crg1-1* and *Crg1-2*. *Crg1-1* is represented by the top line and *Crg1-2* the bottom line. Dots indicate identity; dashes indicate gaps in optimal alignment. Coding sequences are shown in uppercase; nontranslated sequences in lowercase. The XXX denotes a gap in the comparison of the two alleles where the sequences do not align. Because of the dra-

matic difference in size, the sequences represented by the gap are shown at the bottom. **B** Schematic representation of *Crg1-1* and *Crg1-2*. Exons are represented as rectangles with size in nucleotides indicated inside. Introns are represented as thin lines; insertions as inverted triangles. Numbers represent sizes of regions in nucleotides.

ataaaa and may represent a site for regulation at the level of polyadenylation or translation. The 3' untranslated sequences of *Crg1-1* and *Crg1-2* are 94% identical, approaching the 96% identity reported for the synonymous positions of the coding sequences.

With intron identity between *Crg1-1* and *Crg1-2* at 84%, the 3' untranslated regions are unexpectedly similar, even considering the presence of regulatory elements. Table 1 displays sequence similarity between regions of *Crg1-1* and *Crg1-2*.

**Table 1.** Comparison of *Crg1-1* and *Crg1-2* homologous regions

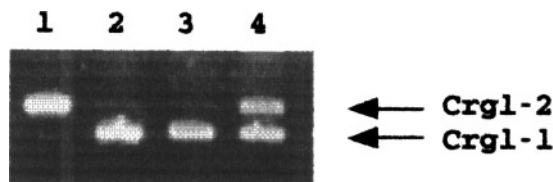
Region	Identify	%
Exons	474/480	98.8
Exons (synonymous sites)	146/152	96.0
Intron	566/676	83.7
3' untranslated	77/82	93.9

### *Crg1-1* and *Crg1-2* Are Alleles

Gene duplication is one possibility for the existence of two closely related genes originating from the same species. If *Crg1-1* and *Crg1-2* had arisen through a duplication event, both genes would be present within any given individual *C. remanei* worm. This possibility was addressed through single-worm PCR (Williams et al. 1992). In this procedure, single *C. remanei* are used as template in separate PCR reactions and the genotype of each individual can be clearly assessed. Care must be exercised in the selection of female individuals as developing embryos within a mated female will also contain paternal genomic DNA and can lead to incorrect genotype scoring. To address this, sexually immature L4 females were selected along with adult males. Due to the 192-nucleotide difference between *Crg1-1* and *Crg1-2*, the presence of one or both genes easily could be scored by PCR amplification using primers flanking the intron and fractionation of the products on agarose gels. The results are shown in Fig. 3. The four individuals depicted provide examples of the three globin genotypes identified. Individual 1 contains only *Crg1-2*, individuals 2 and 3 contain only *Crg1-1*, and the fourth worm contains both globin genes. The two genes were very common in the population and examples of each genotype could be detected by sampling as few as ten worms.

To address the inheritance pattern of the two genes, genetic crosses were carried out and the parents and offspring were then subjected to PCR analysis. Figure 4A shows the results of a single-worm PCR analysis of a genetic cross involving a female carrying only *Crg1-2* and a male carrying only *Crg1-1*. Six F1 offspring of the mating were scored and all but one were shown to contain both genes. The one reaction that failed to generate either globin band may have resulted from inadvertent loss of the worm in the transfer from the plate to the PCR tube. Figure 4B shows the results of a cross between a female carrying both *Crg1-1* and *Crg1-2* with a male containing only *Crg1-2*. Seven offspring were scored and all showed one of the two genotypes of the parents. Additional crosses of parents containing one or both genes produced good Mendelian ratios of offspring. All told, over 40 genetic crosses were performed and the results were always consistent with allelic segregation. An additional culture of *C. remanei* from the Caenorhabditis Genetic Center exhibited the same two-allele structure.

*Crg1-1* and *Crg1-2* homozygote lines were isolated



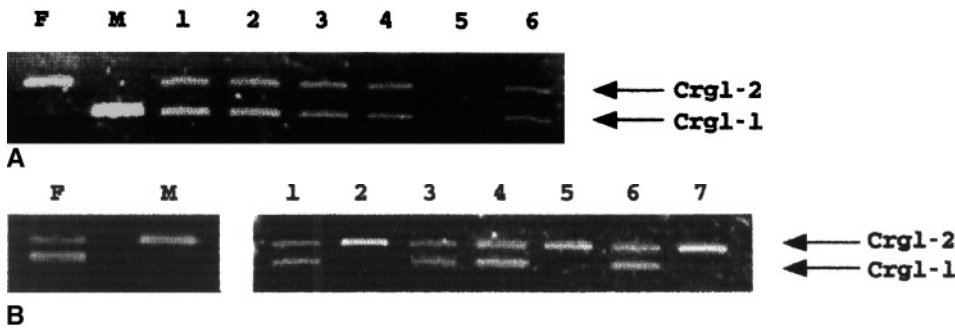
**Fig. 3.** Globin genotype analysis of *C. remanei*. Single-worm PCR was performed using oligonucleotides flanking the intron position in *Crg1-1* and *Crg1-2*. The *Crg1-2* PCR product is 192 base pairs larger than the *Crg1-1* product. Individual 1 contains only *Crg1-2* (one larger band). Individuals 2 and 3 contain only *Crg1-1* (one smaller band). Individual 4 contains both genes (two bands).

and examined by visual inspection as well as northern analysis. No obvious physical differences were apparent between the two lines. Figure 5 shows a northern blot containing RNA from both homozygous lines hybridized with a radiolabeled fragment of *Crg1-2* cDNA. The autoradiograph shows both messages to be present and similar in size (around 600–700 nucleotides) in each line.

### Discussion

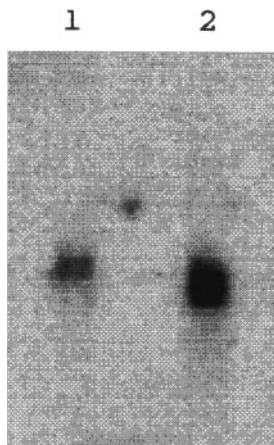
It is unusual that the single introns of *Crg1-1* and *Crg1-2* have undergone such extensive changes with respect to each other, yet the 3' untranslated and coding sequences have been maintained virtually unchanged. Studies of mammalian globin genes show the level of divergence in exonic silent third position sites generally approaches that of the noncoding sequences (Efstratiadis et al. 1980; Koop et al. 1986; Li and Gojobori 1983; Marks et al. 1986). This is not the case for *Crg1-1* and *Crg1-2*. Of 152 allowable silent third position changes, only six are present (4% divergence). This similarity is far higher than the minimum of 16% divergence between the introns. Inspection of codon usage in the two alleles shows little preference, suggesting that codon bias is not the source of the conservation of the silent third positions. However, the possibility that these third positions are under some form of selective pressure cannot be entirely ruled out.

Considering the flexibility of globin sequences in accumulating replacement mutations while retaining the overall globin fold, we expected that some nonsynonymous mutations would have occurred between *Crg1-1* and *Crg1-2*; none were observed. In view of this, the absence of such changes seems more likely a result of genetic mechanisms than natural selection against change. We view the surprisingly high identity of *Crg1-1* and *Crg1-2* in the exons and 3' untranslated region as evidence that multiple gene conversion events have occurred between these alleles. A gene conversion event can involve a single-stranded attack on a locus from any juxtaposed strand of DNA containing similar sequences, leading to the replacement of all or portions of the attacked gene with sequences from the neighboring posi-



**Fig. 4.** Genetic crosses. **A** Single-worm PCR analysis of a female *C. remanei* carrying only *Crg1-2* crossed with a male carrying only *Crg1-1*. Five of six offspring scored received both genes. Reaction five failed. **B** Single-worm PCR analysis of a female *C. remanei* carrying

both *Crg1-1* and *Crg1-2* crossed with a male carrying only *Crg1-2*. Four of seven offspring are *Crg1-1/Crg1-2* heterozygotes and three of seven are *Crg1-2* homozygotes in accordance with the 50/50 ratio expected for the cross.



**Fig. 5.** Northern blot analysis. *Crg1-1* and *Crg1-2* homozygote lines were established and total RNA was isolated from mixed stage populations of each. North analysis was performed on each sample and expression of both alleles was confirmed by hybridization with radiolabeled *Crg1-2* cDNA. Lane 1 contains RNA from *Crg1-2* homozygote lines and lane 2 contains RNA from *Crg1-1* homozygote lines.

tion (Alberts et al. 1994). As alleles, *Crg1-1* and *Crg1-2* probably come into close proximity during meiotic chromosome pairing and could have undergone several conversion events with one another. In fact, Smithies and co-workers have proposed that pairing of homologous sequences may facilitate chromosomal synapsis during meiosis (Powers and Smithies 1986; Smithies and Powers 1986). Throughout the course of their evolution, insertion and deletion events in the introns of *Crg1-1* and *Crg1-2* could have disrupted base pairing and eventually prevented conversion of the intervening sequences and limited the conversion events mainly to the exons. Evidence of gene conversions limited to coding sequences has been reported in other systems (Hill et al. 1984; Slightom et al. 1985). In addition, conversion events have been shown to occur over regions as small as tens of bases (Mellor et al. 1983), lending support to the idea that even short exons can serve as template for conversion.

Gene conversion events are well documented in the

mammalian  $\gamma$ -globin loci (Hill et al. 1984; Powers and Smithies 1986; Shen et al. 1981; Slightom et al. 1980, 1985, 1987; Smithies and Powers 1986). A hallmark of these conversions is the presence of simple dinucleotide repeats consisting of alternating purine-pyrimidine sequences around the apparent boundaries of the conversions (Powers and Smithies 1986; Shen et al. 1981; Slightom et al. 1980, 1985, 1987; Smithies and Powers 1986). It has been suggested that under the appropriate conditions these tracts could form unusual DNA conformations, such as Z-DNA, which could facilitate strand breakage and serve as recognition sites for recombination machinery (Arnott et al. 1980; Shen et al. 1981; Slightom et al. 1980; Wang et al. 1979, 1981). No such repetitive sequences could be identified anywhere within the intron or proximal exon sequences of *Crg1-1* and *Crg1-2*, suggesting a different mechanism of recognition and initiation.

An alternative explanation for the disparity between intron and exon conservation would be that the identity shared between *Crg1-1* and *Crg1-2* synonymous sites of exons actually reflects a short period of evolutionary divergence and that the introns have undergone particularly rapid evolution. As is apparent in Fig. 2, both introns have incurred numerous insertions and deletions with respect to each other, making precise alignment and assessment of evolutionary distance difficult. These introns (and possibly others in *C. remanei*) may contain hot spots for rapid evolution, and assessment of evolutionary divergence based on these sequences could overestimate the time of divergence of *Crg1-1* and *Crg1-2*. Moreover, the conservation of the 3' untranslated sequences may derive more from functional constraints than is apparent. The fact that there are numerous single-base substitutions spread through most of the intron makes this alternative explanation unlikely.

The finding of two divergent alleles in *Caenorhabditis* would be unexpected as most strains used in laboratories are known inbred lines. In addition, a study of calmodulin-like genes from 11 different strains of *C. elegans* collected around the world showed no sequence

differences, even in introns (Thomas and Wilson 1991). Our finding of two divergent alleles in *C. remanei* suggest that the *C. remanei* species may exhibit a more variable population structure than that of *C. elegans*. Currently only common lab strains of *C. remanei* are available. Considering the extensive inbreeding of these strains, it is surprising to find a heterozygous locus.

When comparing sequences suspected to have undergone gene conversions, the most divergent regions in the alignment are expected to best define the evolutionary distance. The divergence of intron sequences in the two alleles suggests that both alleles have existed for millions of years. Since they encode identical proteins, any advantage derived by maintenance of both alleles must come from outside of the coding sequences. One possibility is that selection is not based on the globin genes at all. Another locus, tightly linked to the globin gene position, may be affecting the survival and distribution of *Crg1-1* and *Crg1-2* through its own selective effects. However, both globin alleles are expressed and, in short-term culture, homozygotes survive and reproduce as well as heterozygotes. Thus, we have an unusual situation wherein two ancient alleles encoding identical proteins have been maintained in a nematode population despite unclear heterozygote benefit.

**Acknowledgments.** We thank Jan Geliebter, Martin Kreitman, and W. Kelly Thomas for helpful discussions. We also thank Allan Jones for his generous gift of *Caenorhabditis* RNA. J.P.M. and T.S. were supported by NIH grant HD25614. A.P.K. was supported in part by NIAID grant 5T32AI07172. Some strains used in this study were provided by the *Caenorhabditis* Genetics Center, which is supported by the NIH National Center for Research Resources.

## References

- Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD (1994) Molecular biology of the cell. Garland, New York, NY
- Arnott S, Chandrasekaran R, Birdsall DL, Leslie AGW, Ratliff RL (1980) Left-handed DNA helices. *Nature* 283:743–745
- Arutyunyan EG (1981) The structure of leghemoglobin. *Mol Biol* 15: 27–44
- Bashford D, Chothia C, Lesk AM (1987) Determinants of a protein fold: unique features of the globin amino acid sequences. *J Mol Biol* 196:199–216
- Blaxter ML (1993) Nematoglobins: divergent nematode globins. *Parasitol Today* 9:353–360
- Blaxter ML, Vanfleteren JR, Xia J, Moens L (1994) Structural characterization of an *Ascaris* myoglobin. *J Biol Chem* 269:30181–30186
- Brown GG, Lee JS, Brisson N, Verma DPS (1984) The evolution of a plant globin gene family. *J Mol Evol* 21:19–32
- Davenport H (1949) The haemoglobins of *Ascaris lumbricoides*. *Proc R Soc Lond Biol* 136:255–270
- De Baere I, Liu L, Moens L, Van Beeumen J, Gielens C, Richelle J, Trotman C, Finch J, Gerstein M, Perutz M (1992) Polar zipper sequence in the high-affinity hemoglobin of *Ascaris suum*: amino acid sequence and structural interpretation. *Proc Natl Acad Sci USA* 89:4638–4642
- De Baere I, Perutz MF, Kiger L, Marden MC, Poyart C (1994) Formation of two hydrogen bonds from the globin to the heme-linked oxygen molecule in *Ascaris* hemoglobin. *Proc Natl Acad Sci USA* 91:1594–1597
- Dickerson RE, Geis I (1983) Hemoglobin: structure, function, evolution, and pathology. Benjamin Cummings, Menlo Park, CA
- Dixon B, Pohajdak B (1992) Did the ancestral globin gene of plants and animals contain only two introns? *Trends Biochem Sci* 17:486–488
- Dixon B, Walker B, Kimmins W, Pohajdak B (1992) A nematode hemoglobin gene contains an intron previously thought to be unique to plants. *J Mol Evol* 35:131–136
- Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel JK, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot NJ (1980) The structure and evolution of the human  $\beta$ -globin gene family. *Cell* 21:653–668
- Fields C (1990) Information content of *Caenorhabditis elegans* splice site sequences varies with intron length. *Nucleic Acids Res* 18: 1509–1512
- Frohman MA, Dush MK, Martin GR (1988) Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci USA* 85: 8998–8992
- Gibson Q, Smith M (1965) Rates of reaction of *Ascaris* haemoglobins with ligands. *Proc R Soc Lond [Biol]* 163:206–214
- Go M (1981) Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* 291:90–92
- Goetinck S, Waterston RH (1994) The *Caenorhabditis elegans* muscle-affecting gene *unc-87* encodes a novel thin filament-associated protein. *J Cell Biol* 127:79–93
- Heine U, Blumenthal T (1986) Characterization of regions of the *Caenorhabditis elegans* X chromosome containing vitellogenin genes. *J Mol Biol* 188:301–312
- Heschl MFP, Baillie DL (1990) Functional elements and domains inferred from sequence comparisons of a heat shock gene in two nematodes. *J Mol Evol* 31:3–9
- Hill A, Hardies SC, Phillips SJ, Davies MG, Hutchison CA, Edgell MH (1984) The mouse early embryonic  $\beta$ -globin gene sequences. *J Biol Chem* 259:3739–3747
- Jensen EO, Paludan K, Hyldig-Nielsen JJ, Jorgensen P, Marcker KA (1981) The structure of a chromosomal leghaemoglobin gene from soybean. *Nature* 291:677–679
- Kennedy BP, Aamodt EJ, Allen FL, Chung MA, Heschl MFP, McGhee JD (1993) The gut esterase gene (*gas-1*) from the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *J Mol Biol* 229: 890–908
- Kloek AP, Sherman DR, Goldberg DE (1993) Novel gene structure and evolutionary context of *Caenorhabditis elegans* globin. *Gene* 129: 215–221
- Kloek AP, Yang J, Mathews FS, Frieden C, Goldberg DE (1994) The tyrosine B10 hydroxyl is crucial for oxygen avidity of *Ascaris* hemoglobin. *J Biol Chem* 269:2377–2379
- Koop BF, Miyamoto MM, Embury JE, Goodman M, Czelusniak J, Slightom JL (1986) Nucleotide sequence and evolution of the orangutan  $\epsilon$  globin gene region and surrounding alu repeats. *J Mol Evol* 24:94–102
- Li WH, Gojorbori T (1983) Rapid evolution of goat and sheep globin genes following gene duplication. *Mol Biol Evol* 1:94–108
- Mansell JB, Timms K, Tate WP, Moens L, Trotman CNA (1993) Expression of a globin gene in *Caenorhabditis elegans*. *Biochem Mol Biol Int* 30:643–647
- Marks J, Shaw JP, Shen C-KJ (1986) The orangutan adult  $\alpha$ -globin gene locus: duplicated functional genes and a newly detected mem-

- ber of the primate  $\alpha$ -globin gene family. *Proc Natl Acad Sci USA* 83:1413–1417
- Mellor AL, Weiss EH, Ramachandran K, Flavell RA (1983) A potential donor gene for the *bm1* gene conversion event in the C57BL mouse. *Nature* 306:792–795
- Moens L, Vanfleteren J, De Baere I, Jellie AM, Tate W, Trotman CNA (1992) Unexpected intron location in non-vertebrate globin genes. *FEBS Lett* 312:105–109
- Okazaki T, Wittenberg JB (1965) The hemoglobin of *Ascaris* peritertic fluid: equilibria with oxygen and carbon monoxide. *Biochim Biophys Acta* 3:503–511
- Powers PA, Smithies O (1986) Short gene conversions in the human fetal globin gene region: a by-product of chromosome pairing during meiosis? *Genetics* 112:343–358
- Prasad SS, Baillie DL (1989) Evolutionarily conserved coding sequences in the *dpy-20-unc-22* region of the *Caenorhabditis elegans*. *Genomics* 5:185–198
- Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ch 7
- Shen S, Slightom JL, Smithies O (1981) A history of human fetal globin gene duplication. *Cell* 26:191–203
- Sherman DR, Kloek AP, Krishnan R, Guinn B, Goldberg DE (1992) *Ascaris* hemoglobin gene: plant-like structure reflects the ancestral globin gene. *Proc Natl Acad Sci USA* 89:11696–11700
- Slightom JL, Blechl AE, Smithies O (1980) Human fetal  $\text{G}\gamma$ - and  $\text{A}\gamma$ -globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell* 21:627–638
- Slightom JL, Chang L-YE, Koop BF, Goodman M (1985) Chimpanzee fetal  $\text{G}\gamma$ - and  $\text{A}\gamma$ -globin gene nucleotide sequences provide further evidence of gene conversions in hominine evolution. *Mol Biol Evol* 2:370–389
- Slightom JL, Theisen TW, Koop BF, Goodman M (1987) Orangutan fetal globin genes. *J Biol Chem* 262:7472–7483
- Smithies O, Powers PA (1986) Gene conversions and their relation to homologous chromosome pairing. *Philos Trans R Soc Lond Biol* 312:291–302
- Stoltzfus A, Doolittle WF (1993) Slippery introns and globin gene evolution. *Curr Biol* 3:215–217
- Stoltzfus A, Spencer DF, Zuker M, Logsdon JM, Doolittle WF (1994) Testing the theory of genes: the evidence from protein structure. *Science* 265:202–207
- Sulston J, Du Z, Thomas K, Wilson R, Hillier L, Staden R, Halloran N, Green P, Thierry-Mug J, Qiu L, Dear S, Coulson A, Craxton M, Durbin R, Berks M, Metzstein M, Ainscough R, Waterston R (1992) The *C. elegans* genome sequencing project: a beginning. *Nature* 356:37–41
- Thomas WK, Wilson AC (1991) Mode and tempo of molecular evolution in the nematode *Caenorhabditis*: cytochrome oxidase II and calmodulin sequences. *Genetics* 128:269–279
- Wang AHJ, Quigley GJ, Kolpak FJ, Crawford JL, van Boom JH, van der Marel G, Rich A (1979) Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* 282:680–686
- Wang AHJ, Quigley GJ, Kolpak FJ, van der Marel G, van Boom JH, Rich A (1981) Left-handed double helical DNA: variations in the backbone conformation. *Science* 211:171–176
- Williams BD, Schrank B, Huynh C, Shownkeen R, Waterston RH (1992) A genetic mapping system in *Caenorhabditis elegans* based on polymorphic sequence-tagged sites. *Genetics* 131:609–624
- Yamauchi K, Ochiai T, Usuki I (1992) The unique structure of the *Paramecium caudatum* hemoglobin gene: the presence of one intron in the middle of the coding region. *Biochim Biophys Acta* 1171:81–87
- Yang J, Kloek AP, Goldberg DE, Mathews FS (1995) The structure of *Ascaris* hemoglobin domain I at 2.2Å resolution: molecular features of oxygen avidity. *Proc Natl Acad Sci USA* 92:4224–4228
- Zucker-Aprison E, Blumenthal T (1989) Potential regulatory elements of nematode vitellogenin genes revealed by interspecies sequence comparison. *Mol Evol* 28:487–496