Selective conservation of the RSL-encoding, proteinase inhibitory-type, clade L serpins in *Caenorhabditis* species

Cliff J. Luke [1], Stephen C. Pak [1], David J. Askew [1], Yuko S. Askew [1], Justin E. Smith [2] and Gary A. Silverman [1]

[1] *UPMC Newborn Medicine Program, Department of Pediatrics, University of Pittsburgh School of Medicine, Children's Hospital of Pittsburgh and Magee-Womens Research Institute, 204 Craft Avenue, Pittsburgh, PA 15213, [2]Washington and Jefferson College, 60 South Lincoln, Washington, PA 15301*

TABLE OF CONTENTS

1. ABSTRACT

Serpins are a highly conserved superfamily of serine and papain-like cysteine proteinase inhibitors that are divided phylogenetically into clades. Serpins also can be divided anatomically into those that reside predominately outside or inside cells. While the activities of the extracellular serpins are well understood, the biological functions, as well as the overall distribution of the intracellular (serpin$_{IC}$) serpins is less well defined. Conceivably, the biological function of the serpins$_{IC}$ might be revealed by analysis of species with genomes of lower complexity. To this end, we sought to define the clade L serpin repertoire of *Caenorhabditis elegans* and other nematode species. Analysis of the *C. elegans* genome revealed the presence of 9 serpin genes. Five genes encoded for full-length serpins with functional reactive site loops (RSL). By definition, these genes were designated proteinase inhibitory-type, RSL-encoding serpins. Four of the *C. elegans* genes encoded for proteins without an RSL or transcripts with premature termination codons. The high percentage of non-RSL encoding to RSL-encoding serpin genes suggested that the former served a unique biological function rather than residing in the genome as simple pseudogenes. If this hypothesis was correct, we expected these non-RSL encoding genes to be conserved precisely in other *Caenorhabditis* species. However, in contrast to the RSL-encoding serpins that were well conserved and segregated into 3 sub-clades, we failed to detect non-RSL encoding serpin orthologues in the genomes of *Caenorhabditis briggsae* and *Caenorhabditis remanei*. These data suggested that unlike their RSL-encoding paralogues, the relatively high percentage of non-RSL encoding serpins in *C. elegans* was a vestige of recent duplication events and these latter genes were unlikely to serve essential functions in *Caenorhabditis* species.

2. INTRODUCTION

The high molecular weight serine proteinase inhibitor (serpin) superfamily has evolved over 1000 million years (1) and is present in all domains of life (2, 3). Serpins are a structurally well-conserved superfamily of proteins that regulate proteolytic events, such as blood coagulation, fibrinolysis, apoptosis and inflammation (reviewed in (3, 4)). Serpins inhibit their target proteinases via a suicide substrate-like mechanism (5-7), where the active site of the target proteinase binds the exposed substrate-like serpin reactive site loop (RSL) and initiates peptide bond hydrolysis between the P1 and P1' residues (Schechter and Berger numbering scheme, (8)). Cleavage by the proteinase allows the metastable serpin to undergo a

major conformational rearrangement characterized by complete insertion of the N-terminal region of the RSL into β-sheet A. Loop insertion generates a more stable six-stranded, antiparallel structure (5). The crystal structure of the trypsin and α-1 antitrypsin (SERPINA1) complex shows that proteinase inhibition occurs by structural deformation of the proteinase (9). The proteinase, covalently attached to the N-terminal portion RSL, during loop insertion is translocated 70 Å towards the opposite pole of the serpin molecule (9). This conformational change distorts the active site of the proteinase and traps the serpin and enzyme in a covalent acyl-enzyme complex (9-11).

Phylogenetically serpins fall into 17 clades plus more than 10 orphans (3, 12). Analysis of the serpin genes in vertebrates shows that these serpins fall into clades A-I, with orthologous genes from different species segregating into the same clade (e.g., human SERPINB5 and mouse Serpinb5 in clade B (13, 14)). However, in the lower eukaryotic species many of the serpins fall into a phylum or species-specific clade (e.g., the nematode serpins in clade L).

The clade B serpins differ from the other vertebrate clades, as they residue predominately within cells, have accessory C-D loops and lack classical N-terminal signal peptides (15, 16). Although the clade B intracellular serpins (serpins$_{IC}$) are implicated in tumor progression, angiogenesis and apoptosis, the lack of an association between naturally occurring mutations/variations and disease phenotypes has yet to provide clear insight into their biological function (16, 17). The use of targeted deletions in mouse model systems has so far yielded little information regarding biologic activity, as null mutants have yielded no overt phenotypes (18). The exception is Serpinb5 (maspin), a non-inhibitory type serpin, in which homozygous loss leads to embryonic lethality (19). The expanded repertoire of clade B serpins in the mouse, which may result in some functional overlap, may account in part for the lack of overt phenotypes (13, 14). Conceivably the biological function of the serpins$_{IC}$ might be revealed by analysis of species with genomes of lower complexity. To this end, we sought to define the serpin repertoire in *Caenorhabditis elegans* and other nematodes.

Database analysis of the *C. elegans* genome revealed the presence of 9 serpin genes (20, 21). Five genes encoded for a full-length serpins with functional RSLs. By definition, these genes were designated proteinase inhibitory-type, RSL-encoding serpins. Four of the *C. elegans* genes encoded for proteins without an RSL or transcripts with premature termination codons. Whether these genes encoded for sterile transcripts or truncated proteins with non-proteinase inhibitory functions was unknown, but by definition they were all non-RSL-encoding serpins. The high percentage of non-RSL encoding to RSL-encoding serpin genes suggested that the former served a unique biological function rather than residing in the genome as simple pseudogenes. If this hypothesis was correct, we expected these non-RSL

encoding genes to be conserved precisely in other *Caenorhabditis* species. However, in contrast to the RSL-encoding serpins that were well conserved and segregate into 3 sub-clades, we failed to detect non-RSL encoding serpin orthologues in the genomes of *Caenorhabditis briggsae* and *Caenorhabditis remanei*. These data suggested that unlike their RSL-encoding paralogues, the relatively high percentage of non-RSL encoding serpins in *C. elegans* was a vestige of recent duplication events and these latter genes were unlikely to serve essential functions in *Caenorhabditis* species.

3. MATERIALS AND METHODS

3.1. Database searches

Serpin genes in the *C. briggsae* genome (http://www.ensembl.org, http://www.wormbase.org) and *C. remanei* supercontigs (http://genome.wustl.edu) were determined by the use of the TBLASTN algorithm (22) using the derived amino acid sequences of the previously reported 9 *C. elegans* serpin genes (20). For genes that contained early stop codons (*srp-5*, *-9* and *-10*) the WormBase GENEFINDER predicted amino acid sequences were used (http://www.wormbase.org). The start sites and exon/intron boundaries were confirmed using the motifs described (23). RSL sequences contained within other nematode species were obtained from the NemaBLAST server (http://www.nematode.net/BLAST/) using the RSL sequences from all the *C. elegans* serpins with the TBLASTN algorithm (22) against the Nem-No-Ele EST and genomic database. Species in which serpin RSL motifs were determined were *Ascaris suum* (Asu), *Xiphinema index* (Xin), *Brugia malayi* (Bma), *Pristionchus pacificus* (Ppa), *Meloidogyne arenaria* (Mar), *Strongyloides stercoralis* (Sst), *Wuchereria bancrofti* (Wba), *Ancylostoma ceylanicum* (Ace), *Trichinella spiralis* (Tsp), *Necator americanus* (Nam), *Trichostrongylus vitrinus* (Tvi), *Trichuris muris* (Tmu), *Meloidogyne hapla* (Mha), *Globodera rostochiensis* (Gro) and *Dirofilaria immitis* (Dim).

3.2. Sequencing of the *srp-5* regions from other *C. elegans* strains

Genomic DNA was prepared from the *C. elegans* strains N2 Bristol, CB4855 (Sta-5), CB4856 (HA-8) and CB4852 (N3) by propagating the strains of worms at 20°C on standard nematode growth medium agar plates (24) seeded with the E. coli strain OP50 (Caenorhabditis Genetics Center (CGC), http://biosci.umn.edu/CGC/CGChomepage.htm). Worms were then harvested using M9 media (22 mM KH$_2$PO$_4$, 42 mM Na$_2$HPO$_4$, 86 mM NaCl, 1 mM MgSO$_4$) and lysed in worm lysis buffer (100 mM NaCl, 50 mM Tris, 20 mM EDTA, 500 μg/ml Proteinase K, 1% SDS). The genomic DNA was then extracted using phenol:chloroform:isoamyl alcohol (25:24:1) and ethanol precipitated. The genomic DNA was then resuspended in double distilled water and then treated with DNAse-free RNAse (Roche, Indianapolis, IN) for 30 minutes at 37°C. The *srp-5* region of the genomic DNA was then amplified by polymerase chain reaction using primers SRP-5.St (5'-CAGAATGGCTCTTTTTCTCCAATCA-3') and SRP5 RSL1R (5'- CGGTTCGCTGGACTGATCCG -3'). The amplified product from each species was then sequenced by

the Mental Retardation Resource Core sequencing facility (Children's Hospital Boston) using primers SRP-5.ex2.Start (5'-ATGGCTCTTTTTCTCCAATCAGAG-3') and SRP5.1R (5'-CACTGATACTGTCCGGGTTGATAATTG-3').

3.3. Multiple sequence alignments

Amino acid sequences deduced from the cDNAs derived from C. elegans (srp-1 (accession number AY525078); srp-2 (accession number AY525079); srp-3 (accession number AY525080); srp-5 (accession number AY525081); srp-6 (accession number AY525082); srp-7a-c (accession numbers AY525083-AY525085); srp-8 (accession number AY525086); srp-9 (accession number AY525087); srp-10 (accession number AY525088)) were aligned using ClustalX v1.83 (25). Of the C. elegans serpins with the amino acid sequences derived from genomic serpin gene sequences of C. briggsae (CBG00799; CBG06332; CBG06333; CBG06347; CBG06351; CBG06352; CBG11331) and C. remanei (Cre128.5.1, Cre128.5.2, Cre128.5.3, Cre128.6.1, Cre128.6.2, Cre2.18.1 and Cre2.18.2) were also completed. For genes that contained early stop codons (srp-5, -9 and -10) the WormBase GENEFINDER predicted amino acid sequences were used. Alignments were then exported in multiple sequence alignment format (msf) and visualized using SeqVu 1.1 (Gardner, J., The Garvan Institute of Medical Research, Sydney, Australia).

3.4. Phylogeny

Amino acid sequences from the nematode serpins and the other 276 serpin sequences (available from http://www.ncbi.nlm.gov) were aligned using ClustalX v1.83 (25). For genes that contained early stop codons (srp-5, -9 and -10) the WormBase GENEFINDER predicted amino acid sequences were used. A neighbor-joining phylogenetic tree was created using a random seed generator of 11 and bootstrapped 1000 times. The tree was then visualized using NJplot (Laboratoire de Biométrie, Université Claude Bernard, Lyon, France).

4. RESULTS

4.1. The *C. elegans* serpins
4.1.1. RSL-encoding serpins

Previous analysis of the C. elegans genome and transcriptome revealed the presence of nine serpin genes, all situated on chromosome V (Figure 1). Note, much of the data presented in this figure were presented in tabular form by Pak et. al, but are included here to serve as a reference for the comparative genomic studies (20). Five serpin mRNAs (srp-1, srp-2, srp-3, srp-6 and srp-7) encoded for full-length serpins with functional RSLs. Analysis of their primary deduced amino acid sequences, as compared to the canonical human serpin, SERPINA1 (α1-antitrypsin), revealed several structural variations (Figure 2). First, segments of helices A and C as well as the loop just before the beginning of strand 5A were truncated by two or three amino acids (Figure 2). Second, the distal hinge region of the RSL in the C.

elegans serpins was elongated by 4-6 residues (Figure 2). These regions have significant structure-function relationships. For example, helix A participates in heparin binding and activation of SERPIND1 (26) and the distal hinge region regulates latency and the proteinase specificity of SERPINA5 (27).
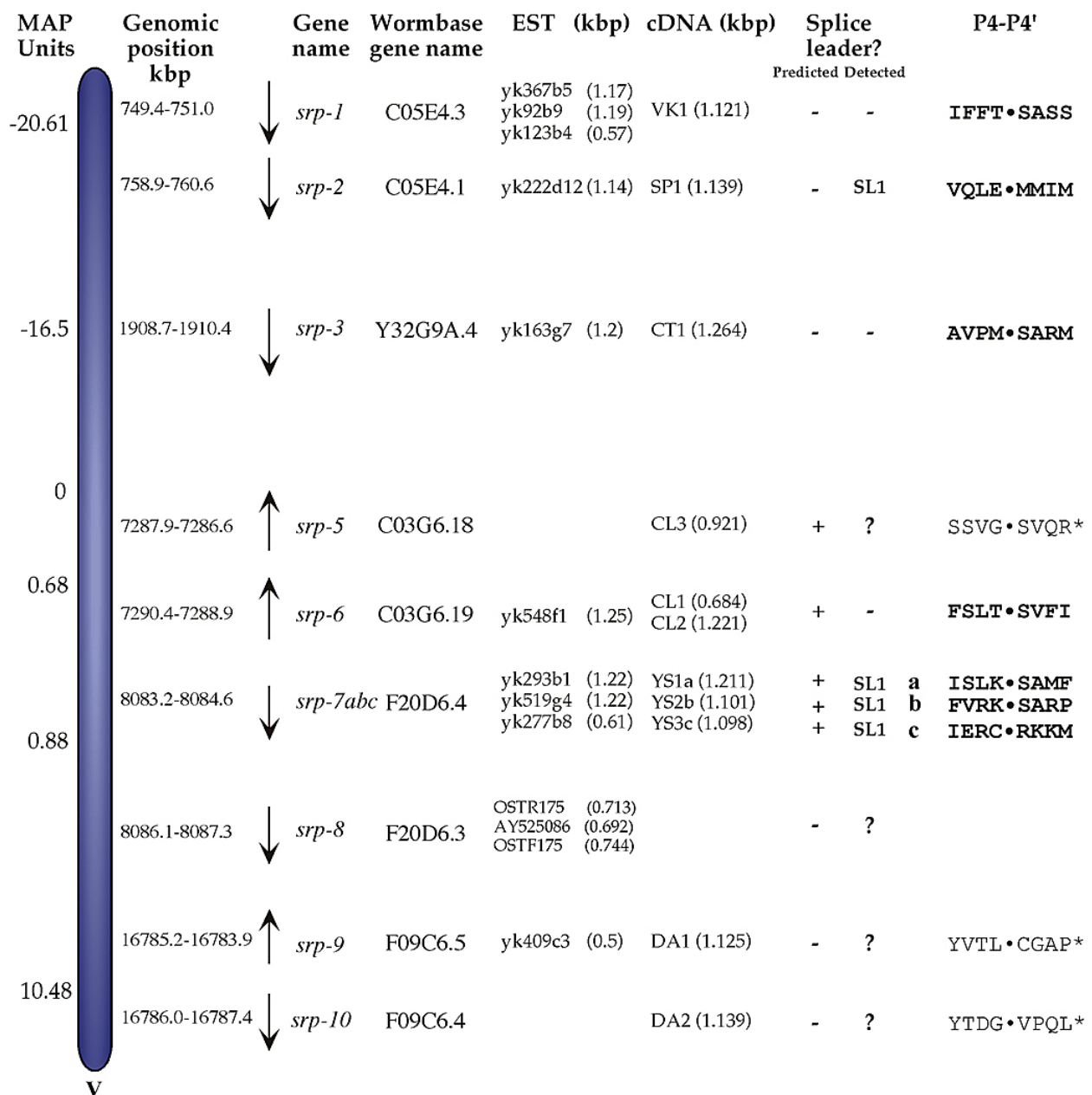
As described computationally and confirmed by our cDNA sequence analysis, *srp-7* encoded for three alternative final exons leading to putative proteins with three different RSLs (20, 28). Alternative splicing between two final exons could occur also with *srp-9*, but premature stop codons within the cDNA predicted that neither protein would be translated (see next section).

4.1.2. Non-RSL-encoding serpins

srp-5, which was nearly identical to *srp-6*, contained a single base pair deletion in the coding region. This deletion should result in a frame-shift and premature termination (Figure 3A). srp-8 encoded for a full-length gene with a highly divergent sequence in the typical location of the RSL (Figure 2). Two serpin mRNAs (srp-9 and srp-10) that appeared to encode for functional RSLs also contained non-sense mutations in exons 1 and 3, respectively (Figure 3A). These mutations should result in premature termination (Figures 2 and 3). Collectively, none of these genes should yield a serpin with a functional RSL. cDNA cloning by our laboratory and others revealed the presence of these variants in the transcripts of SRP-5, -9 and -10 (Figure 3A) (20, 29). By alternative splicing, the GENEFINDER program removed these premature stop codons in all three of these transcripts (Figure 3A). However, if these splicing patterns were to occur in vivo, they would lead to the generation of serpins lacking key structural elements and would most likely lead to protein mis-folding (Figure 3B).

Since three of the four C. elegans non-RSL encoding serpins contained single base-pair mutations, it was conceivable that srp-5, -9 and -10 actually encoded for RSL containing proteins but mutations unique to the N2 strain skewed our assessment. To test this hypothesis, we re-examined the srp-5 gene. When the genomic sequence for srp-5 was compared to that of the nearly identical gene srp-6, srp-5 contained a single base pair deletion at position 368 in exon 2 of the genomic sequence (position 324 in the cDNA) (Figure 3C; black arrowhead). To correct for this single base-pair deletion, the GENEFINDER prediction software spliced out this deletion to obtain a full-length cDNA (Figure 3C). However, the cDNA sequences we obtained did not validate this splicing event and maintained the same splice sites as srp-6. The consequence of this single base-pair deletion was a frame-shift leading to a premature termination codon. To determine if this single base-pair deletion was an N2 strain variant, the srp-5 genomic regions from three other C. elegans strains CB4855 (Sta-5), CB4856 (HA-8) and CB4852 (N3)) were PCR amplified and sequenced. The same single base-pair deletion occurred in all C. elegans strains (Figure 3D).

MAP Units	Genomic position kbp	Gene name	Wormbase gene name	EST (kbp)	cDNA (kbp)	Splice leader? Predicted	Detected		P4-P4'
-20.61	749.4-751.0	srp-1	C05E4.3	yk367b5 (1.17) yk92b9 (1.19) yk123b4 (0.57)	VK1 (1.121)	-	-		**IFFT•SASS**
	758.9-760.6	srp-2	C05E4.1	yk222d12 (1.14)	SP1 (1.139)	-	SL1		**VQLE•MMIM**
-16.5	1908.7-1910.4	srp-3	Y32G9A.4	yk163g7 (1.2)	CT1 (1.264)	-	-		**AVPM•SARM**
0	7287.9-7286.6	srp-5	C03G6.18		CL3 (0.921)	+	?		SSVG•SVQR*
0.68	7290.4-7288.9	srp-6	C03G6.19	yk548f1 (1.25)	CL1 (0.684) CL2 (1.221)	+	-		**FSLT•SVFI**
0.88	8083.2-8084.6	srp-7abc	F20D6.4	yk293b1 (1.22) yk519g4 (1.22) yk277b8 (0.61)	YS1a (1.211) YS2b (1.101) YS3c (1.098)	+ + +	SL1 SL1 SL1	a b c	**ISLK•SAMF** **FVRK•SARP** **IERC•RKKM**
	8086.1-8087.3	srp-8	F20D6.3	OSTR175 (0.713) AY525086 (0.692) OSTF175 (0.744)		-	?		
	16785.2-16783.9	srp-9	F09C6.5	yk409c3 (0.5)	DA1 (1.125)	-	?		YVTL•CGAP*
10.48	16786.0-16787.4	srp-10	F09C6.4		DA2 (1.139)	-	?		YTDG•VPQL*

Figure 1. The *C. elegans* serpins. The genomic positions of the nine *C. elegans* serpins are marked with both MAP units and chromosome V nucleotide sequence positions along an ideogram of chromosome V (*purple shaded oval*). The transcriptional orientations of the genes are indicated (*black arrows*). The approved gene name (http://www.wormbase.org), the original sequence name, the EST sequences determined by Dr. Yuji Kohara (ykohara@lab.nig.ac.jp) and the cDNAs that were amplified and sequenced by our laboratory are also indicated (20). Within WormBase, several of the transcripts were predicted to contain 5' splice leader sequences (+). Transcripts actually containing an SL1 or SL2 splice leader are indicated. Those transcripts with a question mark could not be confirmed due to the low yield of first strand cDNA synthesized. The last column shows the predicted P4-P4' of the RSL sequence deduced from the cDNAs. The dot (•) indicates the position of the putative scissile bond. Unbolded and asterisked (*) motifs indicate the putative RSL residues in transcripts with single premature termination codons (i.e., the RSL sequence was deduced by reading through the stop codon). Note, most of the content of this figure was presented in tabular form in a prior publication (20). The information is presented here to assist in assessing the comparative genomics outlined in this study.
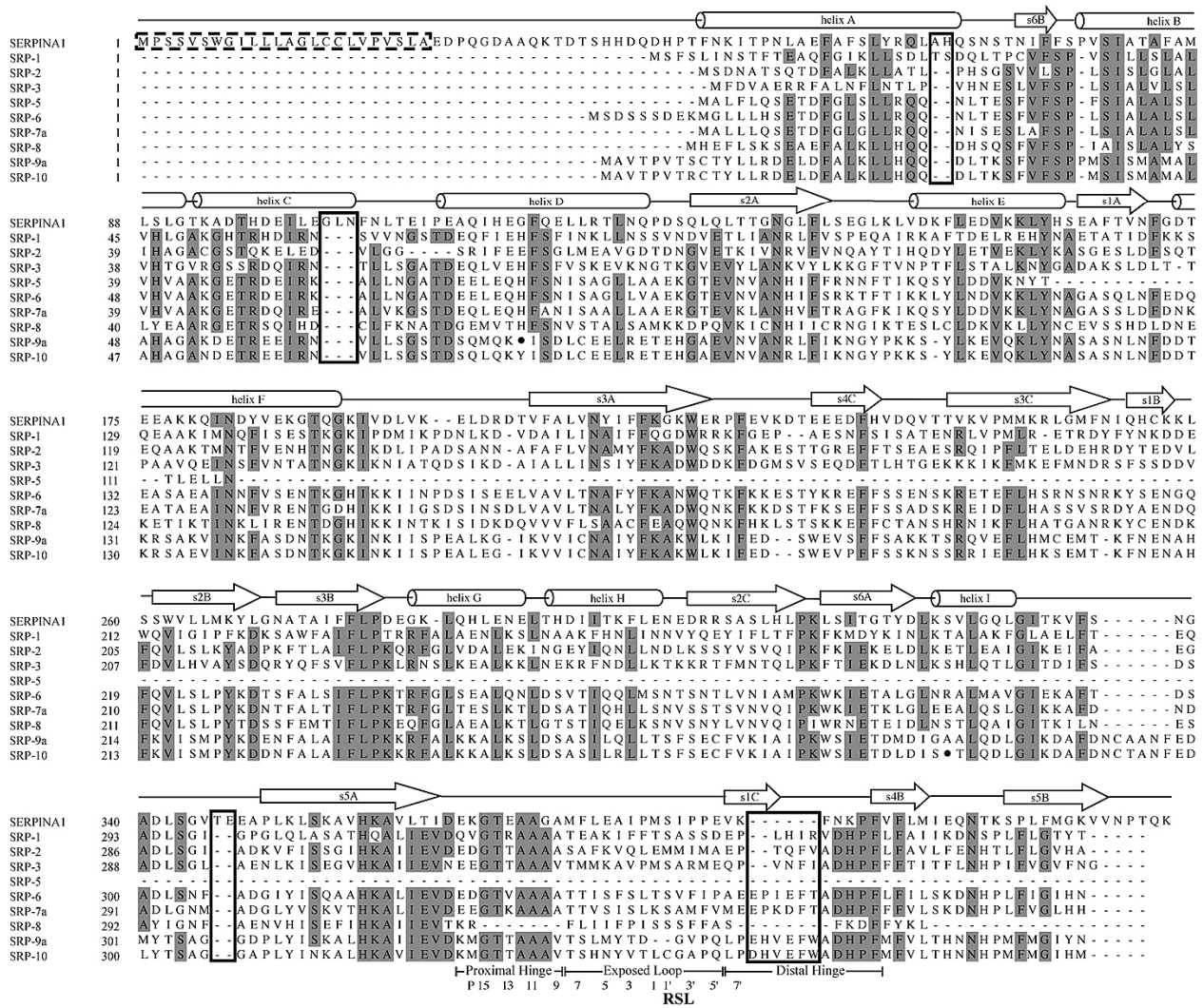
Figure 2. Primary amino acid alignment of the *C. elegans* serpin sequences. The deduced amino acid sequence for the *C elegans* serpins (*srp-1* (accession number AY525078); *srp-2* (accession number AY525079); *srp-3* (accession number AY525080); *srp-5* (accession number AY525081); *srp-6* (accession number AY525082); *srp-7a* (accession number AY525083); *srp-8* (accession number AY525086); *srp-9a* (accession number AY525087) and *srp-10* (accession number AY525088) were aligned with the canonical human serpin, SERPINA1 (α1-antitrypsin, accession number NP_000286), using ClustalX v1.83 (25). The alignment was displayed using SeqVu 1.1 (Gardner, J., The Garvan Institute of Medical Research, Sydney, Australia). The conserved serpin secondary elements (helices [*cylinders*] and sheets [*unfilled arrows*]) are shown above the alignment. The RSL region (*solid line*) is divided into the proximal hinge, exposed loop and distal hinge regions. Translation of the *srp-5* cDNA is a truncated due to a single base-pair deletion resulting in a frame shift and a premature termination codon. The positions of single non-sense mutations in *srp-9* and *srp-10* that generate in-frame stop codons are marked by a dot (•). However, the read-through sequences are provided to indicate that these genes could encode for a functional RSL-containing serpins should the stop codons be suppressed. *srp-8* encodes for a serpin with no recognizable RSL motif. The shading represents regions of identity for over 65% of the sequences. The signal sequence for SERPINA1 is indicated by a dashed box. Note the absence of N-terminal signal peptides in the *C. elegans* serpins. Regions of minor structural differences between the *C. elegans* serpins and SERPINA1 are highlighted (black boxes). Amino acid numbering is relative to the canonical serpin SERPINA1.

These data suggested that for at least *srp-5*, the mutation was not a variant unique to the N2 Bristol strain.

4.2. Comparison of the *C. elegans* serpins with those of *C. briggsae* and *C. remanei*
4.2.1. Genomic comparison

The *C. elegans* genome contained five RSL-encoding and four non-RSL encoding serpins. If the proteinase inhibitory or non-inhibitory functions of these genes provided a selective advantage to the *Caenorhabditis* lineage, we would expect these genes to be conserved among more distantly related *Caenorhabditis* species. As a corollary, strict conservation of the non-RSL encoding genes would suggest that these loci are not simply pseudogenes but rather, encoded for products (RNAs or truncated/alternatively folded proteins) that have important,
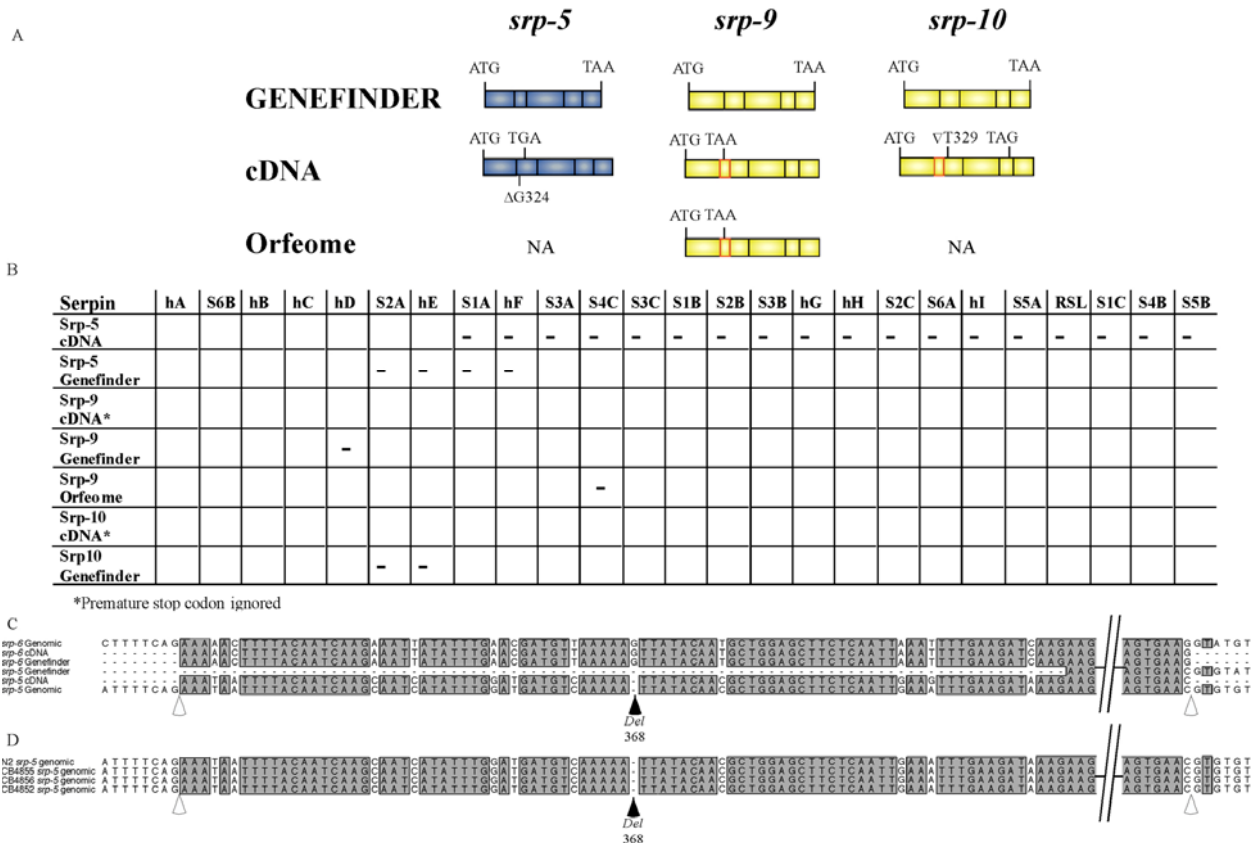
Figure 3. *srp-5, -9 and -10* transcripts contained premature termination codons. A: Schematic representation of the *srp-5*, *-9* and *-10* mRNAs obtained by the GENFINDER prediction program, cDNA cloning and the ORFEOME project v. 1.1 (29). Note the location of the premature termination codons in each of the cDNA sequences for SRP-5, -9 and -10 and the ORFEOME clone for SRP-9. The GENEFINDER prediction program eliminated these stops by alternative splicing. The red boxes indicate the portions of the cDNAs eliminated by GENEFINDER. B: GENEFINDER alternative splicing deletes key structural elements (top of figure ; h = helix, S = beta strand, (-) = loss of structural element). Although there was no experimental evidence suggesting that the alternative splicing of SRP-5, -9 and -10 as predicted by GENEFINDER occurred *in vivo*, these splicing events also would lead to the loss of several key structural elements in the proposed serpin proteins. C: An alignment of the nucleotide sequences from the genomic region of putative exon 2 from *srp-5* and *-6*, the corresponding cDNA sequences and the predicted GENEFINDER spliced sequences. The shading shows the regions of identity. The position of the single base-pair deletion in *srp-5* (*black arrowhead*) and flanking splice sites are marked (*open arrowhead*). D: Nucleotide sequence alignments of the genomic *srp-5* region from four representative *C. elegans* strains; CB4855, CB4856, N2 and CB4852. The position of the single base-pair deletion was present in all four strains (*black arrowhead*). For brevity, 80 base pairs of conserved exon sequence is not depicted (double slash) in both C: and D:.

albeit undefined functions distinct from proteinase inhibition. The nearly completed genomic DNA sequences of *C. briggsae* and *C. remanei* provided an opportunity to make this assessment.

Using the full-length amino acid sequences as well as the available RSL motifs of the *C. elegans* serpins, we used the TBLASTN algorithm to scan the *C. briggsae* supercontigs (release 25; http://www.ensembl.org, http://www.wormbase.org) and the *C. remanei* assemblies (http://genome.wustl.edu/) (22). Initial results from the full-length *C. elegans* serpin BLAST searches, revealed the presence of 7 *C. briggsae* serpin genes (CBG00799, CBG06332, CBG06333, CBG06347, CBG06351, CBG06352 and CBG11331). For simplicity only the RSL amino acid sequences are shown (Figure 4). All seven were

identified previously in the databases cited above. On initial examination, only four of the *C. briggsae* serpin genes (CBG06333, CBG06332, CBG06347 and CBG11331) appeared to contain amino acid sequences reminiscent of a functional RSL, such as the distal hinge signature motif ((A/V)-(D/N/E/Q)-(H/Q)-PF-(L/M)-F), a Thr residue at P14 and Ala residues at P12-P9 (Figure 4) (30).

The RSL of CBG00799 contained a few atypical residues in the hinge region (AlaP12Thr and AlaP9Glu) and a di-prolyl reactive center (P1-P1'). This motif is unusual for the reactive center of an inhibitory-type serpin (12, 16), but could serve as a target for prolyl endopeptidases (31, 32). Thus, we could not predict with confidence whether this gene encoded for a proteinase-
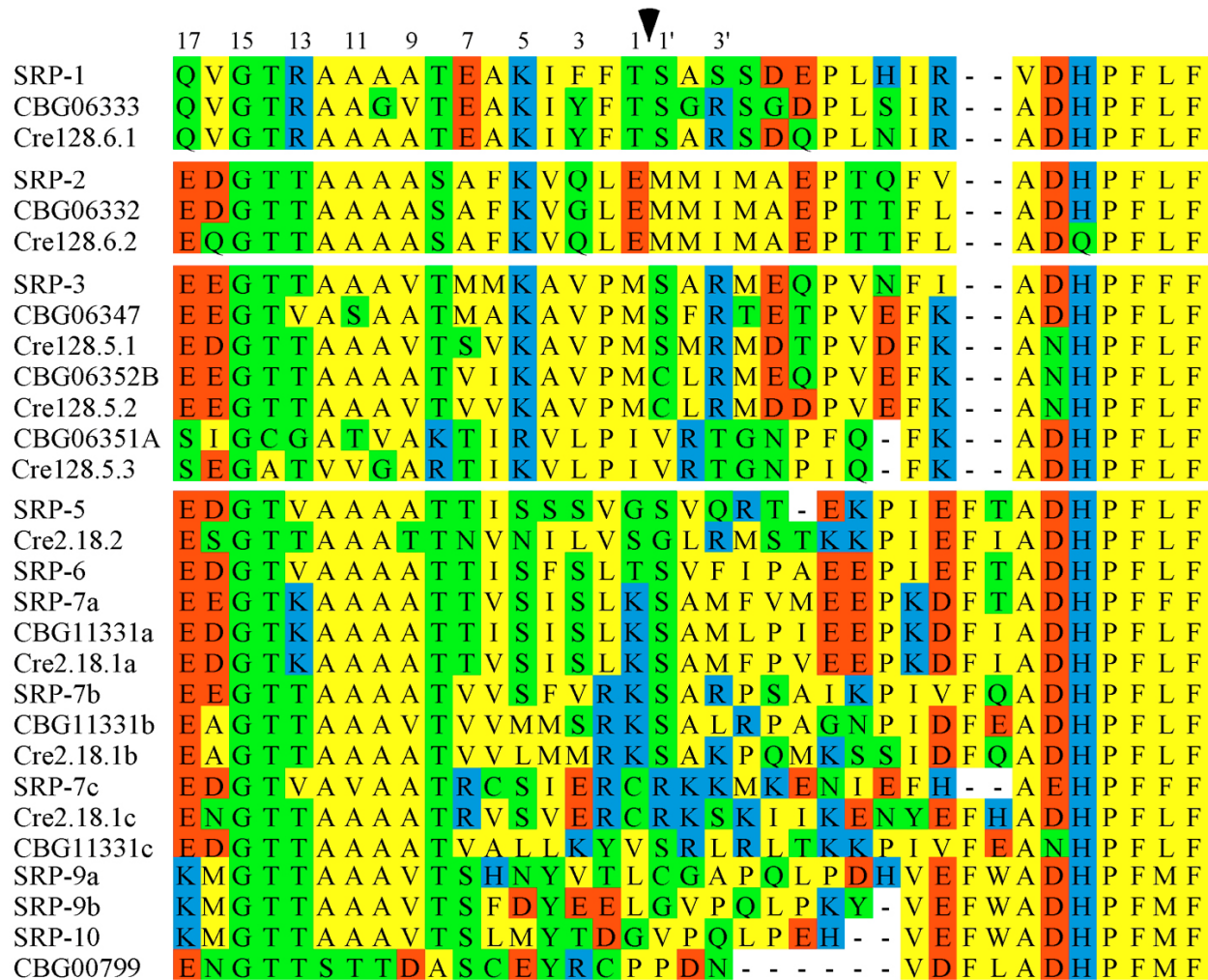
Figure 4. Amino acid alignment of the nematode serpin RSLs. The RSL regions of the nematode serpins, containing the proximal hinge region, the exposed loop and the distal hinge region, were aligned using ClustalX v1.83 and grouped according to their phylogenetic relationships (see Figure 5). The colors indicate polar (*green*), nonpolar/hydrophobic (*yellow*), acidic (*red*), and basic (*blue*) residues. The RSL is numbered from P15 to P3'. The putative scissile bond is marked by an arrowhead.

inhibitory molecule or a non-inhibitory serpin such as SERPINB5 (33).

CBG06352 was relatively large compared to the other serpin genes and was predicted to have seven exons with a large (almost 1 kbp) fourth intron (http://www.wormbase.org). Comparison of this gene with the *C. elegans* serpins showed that the last three exons of CBG06352 contained the entire serpin fold. Moreover, a methionine in exon 5 proved to be a good start site for a *Caenorhabditis* gene (34). Thus, we predicted that CBG06352 was spliced incorrectly by the GENEFINDER prediction software and created an artificial fusion between two genes. We designated the first gene CBG06352A and the last 3 exons containing the actual serpin gene CBG06352B. This serpin gene appeared to contain an inhibitory type RSL (Figure 4).

The original predicted amino acid sequence of CBG06351 also failed to reveal the presence of an RSL

(data not shown). However, examination of the nucleotide sequence revealed an inhibitory-type RSL-containing exon ~300 bp downstream of the second exon. This exon partially overlapped with the predicted exon 3. The RSL-containing exon was flanked 5' by an acceptor splice site and 3' by an appropriately placed stop codon (Figures 4 and 5). Again, we concluded that the original GENEFINDER prediction of the final exon was incorrect and the RSL-containing exon was the actual final exon of CBG06351. The gene was designated CBG06351A (Figure 4).

Within the *C. remanei* assemblies, 7 serpin genes also were found (Cre128.5.1, Cre128.5.2, Cre128.5.3, Cre128.6.1, Cre128.6.2, Cre2.18.1 and Cre2.18.1) (Figure 5). All of these genes encoded for RSL-containing exons. However, one of these genes, Cre128.5.3 had a Val instead of an Ala residue at the P12 position (Figure 4). Thus, we could not predict with confidence whether the RSL-encoding exon of Cre128.5.3 yielded a proteinase-

inhibitory or non-inhibitory serpin. Note this gene was not orthologous to CBG00799 (see below).

Since *C. elegans srp-7* (and conceivably *srp-9*) used alternative 3' exons containing typical RSL sequences, we re-analyzed the databases using only the RSL motifs. This analysis yielded 2 additional RSL-containing exons that were not accounted for previously in both the *C. briggsae* and the *C. remanei* assemblies. Similar to *srp-7* (20), the tandem sets of *C. briggsae* and *C. remanei* RSL-encoding exons were located just 3' to the "last" RSL-encoding exons of CBG11331 and Cre2.18.1, respectively (20). Thus, these genes were designated CBG11331a, -b and -c, and Cre2.18.1a, -b and -c, respectively.

In summary, the *C. elegans* genome encoded for a total of seven different RSL-containing proteins (SRP-1, -2, -3, -6, -7a, -7b, and -7c). All of theses proteins appeared to be proteinase-inhibitory serpins. The other genes encoded for non-RSL encoding mRNAs or non-RSL containing proteins. The *C. briggsae* and *C. remanei* genomes both encoded for 9 RSL-containing proteins, of which at least 8 in each species appeared to be proteinase-inhibitory serpins.

4.2.1.1. Phylogenetic analysis of the nematode serpins

To obtain further insight into the evolutionary relationships between the RSL-containing and non-RSL-containing genes in the *Caenorhabditis* species, we constructed a phylogenetic tree using the neighbor-joining (N-J) method. Serpins from *C. elegans* (n=9), *C. briggsae* (n=7), *C. remanei* (n=7) and the non-redundant NCBI database (n=276; http://www.ncbi.nlm.nih.gov/) were analyzed. As expected, the tree showed that all the *Caenorhabditis* serpin genes fell into clade L (the nematode clade; data not shown).

Within the clade L portion of the tree, three main branches emerged (Figure 5A). One branch contained *srp-1* and orthologues from *C. briggsae* (CBG06333) and *C. remanei* (Cre128.6.1) (Figure 5A, purple). The second branch contained *C. elegans srp-2* and *srp-3* and orthologues for each in *C. briggsae* (CBG06332 and CBG06347) and *C. remanei* (Cre128.6.2 and Cre128.5.1) (Figure 5A, green). This branch also contained two different pairs of orthologous genes unique to *C. briggsae* and *C. remanei* (CBG06351A and Cre128.5.3; CBG06352B and Cre128.5.2) (Figure 5A, green). Although the *C. briggsae* (CBG06351A and CBG06352B) and *C. remanei* (Cre128.5.2 and Cre128.5.3) genes did not have direct orthologues in *C. elegans*, they showed homology to *srp-3*. These data suggested that there was an expanded *srp-3*-like serpin gene family in *C. briggsae* and *C. remanei* or loss of *srp-3*-like paralogues in *C. elegans* (Figure 5A, green). The third branch was the most disparate among the species. *C. elegans srp-7* had orthologues in *C. briggsae* (CBG11331) and *C. remanei* (Cre2.18.1) (Figure 5A, red). However, the *C. elegans srp-5, -6, -8, -9* and *-10* genes had no direct orthologues in either *C. briggsae* or *C. remanei*. However, the tree showed that *C. remanei* (Cre2.18.2) and *C. briggsae* (CBG00799) contained genes with an ancestral

relationship to these *C. elegans* homologues (Figure 5A, red).

4.2.1.2. Genomic organization of the *Caenorhabditis* serpin genes

Another means of examining the evolutionary relationships among orthologous and/or homologous serpin genes was to compare their genomic organization (35). For the *Caenorhabditis* serpin genes, the intron-exon boundaries were determined by aligning the cDNA sequences (20) against their respective genomic regions (Figure 5B). Within the *C. elegans* serpin family the genes contained either 4 (*srp-3*, *srp-5*), 5 (*srp-1*, *srp-7*, *srp-9*, *srp-10*) or 6 (*srp-2*, *srp-6*, *srp-8*) exons. Except for the *srp-7* orthologues, the exon numbers and intron sizes of the *C. briggsae* and *C. remanei* orthologous serpin genes were not well conserved (Figure 5A). Codon phasing at spice junctions also was not well conserved (not shown).

4.2.2 .Analysis of the nematode RSL regions
4.2.2.1. Predicted proteinase inhibitory profiles based on the deduced amino acid sequences of the *Caenorhabditis* RSLs

The genomic and phylogenetic analyses of the *C. elegans* non-RSL-encoding serpins suggested that there was little conservation among the other *Caenorhabditis* species. In contrast, the RSL-encoding, proteinase inhibitory-type serpins were well conserved. To determine if biochemical function of the inhibitory serpins might also be conserved, we examined the amino acid residues surrounding the putative reactive centers of the RSLs. For serine and cysteine proteinases, the substrate amino acid residues opposite the active site S1 and S2 residues, respectively (i.e., the RSL P2 and P1 residues and to a lesser extent the P1' residue) are critical determinants of catalytic and therefore inhibitory specificity. Remarkably, the critical P2-P1' residues were well conserved among the *C. briggsae* and *C. remanei* genes orthologous to *C. elegans srp-1*, *srp-2, srp-3*, *srp-7a* and *srp-7b* (Figure 4). These findings suggested that among these orthologues, proteinase-inhibitory activity also was conserved. Interestingly, the *C. elegans srp-7c* and *C. remanei* Cre2.18.1c RSL sequences differed from the orthologous *C. briggsae* gene (CBG11331c). The P1•P1' amino acids for *srp-7c* and Cre2.18.1c were Cys•Arg, whereas that for CBG11331c was Val•Ser (Figure 4). These data suggested that the proteinase inhibitory profile of CBG11331c would differ from that of *srp-7c* and Cre2.18c.

Five of the seven *C. elegans* serpins (*srp-1*, *srp-2*, *srp-5*, *srp-6* and *srp-7a*) have a Phe, Val or Leu residue at the putative P2 position, as compared to 3/7 and 4/7 of the *C. briggsae* and *C. remanei* serpin genes, respectively (Figure 5). Since papain-like cysteine proteinases prefer the hydrophobic residues Phe, Leu or Val at the P2 position (36-39) and the *C. elegans* genome harbors a large number of papain-like cysteine proteinases (21), most of these serpins may help regulate endogenous papain-like cysteine proteinases.

The *srp-3* RSL contained Met•Ser residues at the P1•P1' positions, respectively, as well as a Pro at the P2
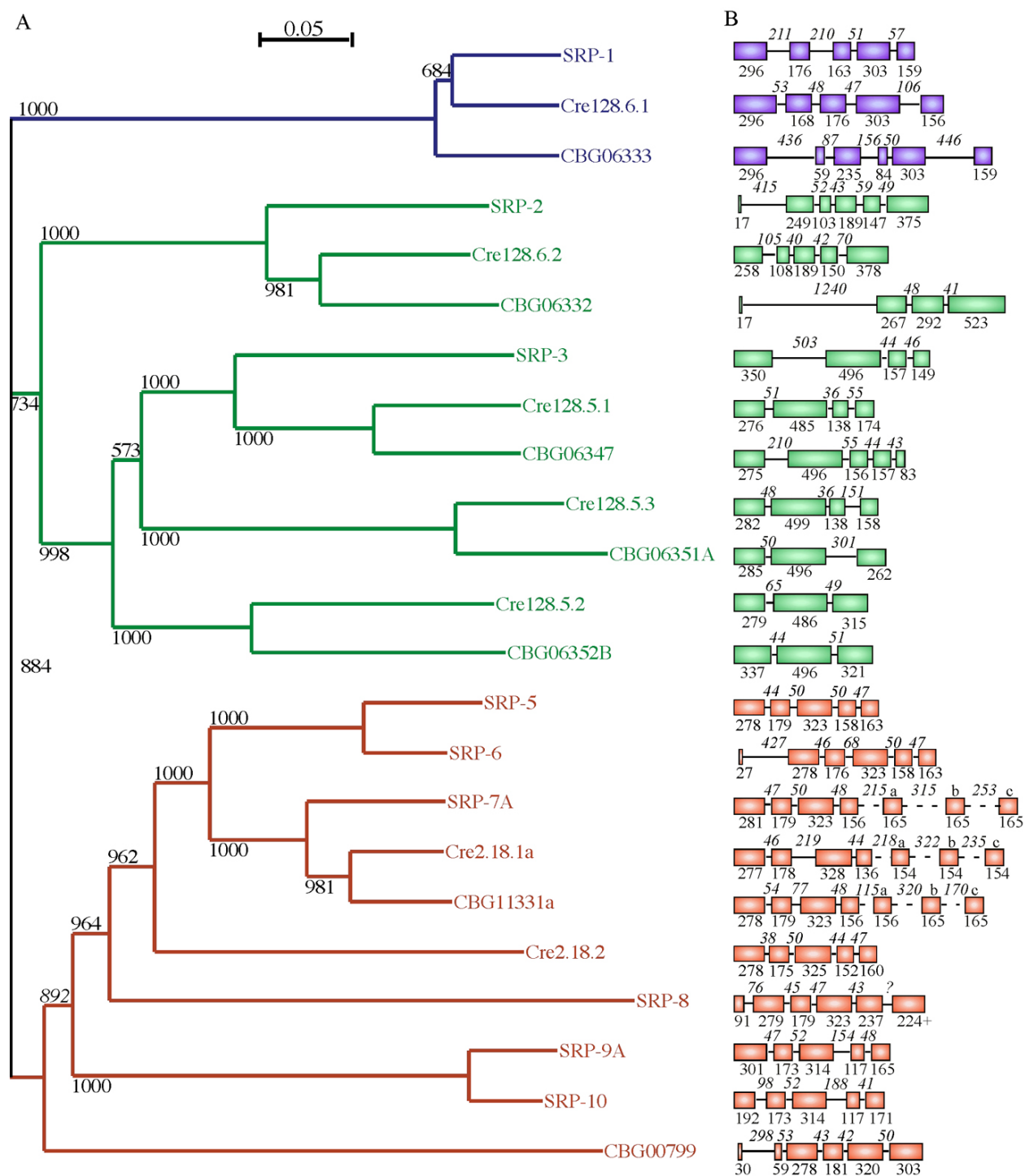
Figure 5. Phylogenetic analysis of the clade L serpins. A: Phylogenetic tree. The nematode serpins were aligned with 276 other serpin sequences and a bootstrapped neighbor-joining tree was created using ClustalX v1.83. For clarity, the tree branch that contained the clade L serpins was isolated using NJplot. The bootstrap values are shown at the branch splits. The length of the branches indicates the number of base-pair substitutions per 100 amino acids (scale bar = 0.5 substituions/100 residues). Three subgroups were identified; *srp-1* family (*purple*); *srp-2* and *srp-3* family (*green*); *srp-7* family (*red*). B: Gene organization within three nematode species. The exon-intron boundaries for the serpin genes found in *C. elegans* (*srp-X,* where X is a gene number), *C. briggsae* (CBGXXXXX) and *C. remanei* (CreXXX) were grouped according to their phylogenetic families (*srp-1* family, *purple*; *srp-2* and *srp-3* family, *green*; *srp-7* family, *red*). The exons (*shaded boxes*) are shown with the number of base-pairs underneath (*bold type*). The introns (*black line*) are shown with the number of base-pairs above (*italic type*).

position. This motif was remarkably similar to human SERPINA1 (α1-antitrypsin) and this serpin may inhibit chymotrypsin and trypsin-like serine proteinases. Indeed, this inhibitory profile has been confirmed *in vitro* using recombinant protein (Pak *et al.*, manuscript in preparation).

srp-7b contained Lys residues at the P1 and P2 position. This motif may be recognized by furin-like, serine proprotein convertases (40, 41).

4.2.2.2. Predicted proteinase inhibitory profiles based on the deduced amino acid sequences of nematode RSLs

Although there were no orthologues of the four *C. elegans* non-RSL encoding serpins within the other two *Caenorhabditis* species, it was conceivable that these genes were the vestiges of important nematode serpins whose functions became non-essential during their evolution. If this were the case, then ancestors of these genes should be present in other nematode genomes. However, many of the other nematode genomes are not fully sequenced, and many of the ESTs contained non-overlapping 5' and 3' reads. Nonetheless, these databases are of great utility since the RSLs are encoded at the 3' end of serpin genes. Since, three of the four non-RSL encoding genes in *C. elegans* (*srp-5*, *-9* and *-10*) could yield proteinase-inhibitory RSLs if it were not for the premature stop codons, identification of any of these hypothetical RSL motifs in the databases might allude to the presence of active proteinase inhibitors in other nematode lineages. The RSL sequences from all the *C. elegans* serpins were analyzed by the NemaBLAST server using the Nem-No-Ele database and the TBLASTN algorithm. This analysis yielded 34 RSL sequences (Figure 6). However, none of deduced RSL sequences from any of the nematode species had a high degree of similarity to any the hypothetical RSLs of the non-RSL encoding *C. elegans* serpin genes.

Based on the P2-P1' RSL residues, there were, however, nematode RSL motifs similar to those of *C. elegans srp-1* (AsuCA303711), *srp-2* (PpaBH827155), *srp-3* (TspBD015390), *srp-7a* (AsuBM278618) and *srp-7c* (TviY12233) (Figure 6). Of the 34 nematode RSL sequences identified in the database, 13, 12, 5 and 2 contained hydrophobic, basic, small polar and acidic residues at the P1 position, respectively (Figure 6). Based on general proteinase active site preferences, these serpins could target chymotrypsin-like serine, trypsin-like serine, elastase-like serine and caspase-like cysteine proteinases, respectively (Figure 6). As with the *Caenorhabditis* serpins, 17 of these deduced amino acid sequences contained a Phe, Leu, or Val residue at (Figure 6, *black arrows*) or near (Figure 6, *gray arrows*) the putative P2 position, suggesting that these nematode serpins could neutralize papain-like cysteine proteinases. Similar to several other serpins$_{IC}$, many of the nematode serpins are likely to serve as dual serine and cysteine proteinase inhibitors (Figure 6, arrowheads) (42, 43).

5. DISCUSSION

The *C. elegans* genome contains nine serpin genes (20). Five of the genes (*srp-1*, *-2*, *-3*, *-6* and *-7*) were RSL-encoding and were likely to inhibit serine and/or

papain-like cysteine proteinases (see (20, 42) and Pak *et al.*, manuscript in preparation). Curiously, the remaining four genes (*srp-5*, *-8*, *-9* and *-10*) appeared to be non-RSL encoding. In the case of both *srp-9* and *-10*, the absence of an RSL was due to a single nucleotide change resulting in a nonsense mutation. A single base-pair deletion in *srp-5* resulted in a frame-shift. Thus, for *srp-5*, *-9* and *-10* premature termination codons would yield a truncated protein void of an RSL, providing that the transcripts were not eliminated by nonsense mediated RNA decay (44, 45). Of note, at least SRP-5, -9 and -10 transcripts have been detected in either EST databases, by hybridization to microarrays, in ORF libraries, by SAGE analysis or by RT-PCR (WormBase and (20, 29, 46, 47). In the case of *srp-8*, the gene encoded for a protein with no recognizable RSL motif at the C-terminus. Since, *srp-5*, *-9* and *-10* would encode for a full-length RSL-containing, proteinase-inhibitory type serpin if it were not for the single base-pair mutations, we sought to determine if at least these genes represented simple sequence variants in the *C. elegans* N2 Bristol strain and would be present in a fully functional form in other *Caenorhabditis* strains. This proved not to be the case as the same single base-pair deletion in *srp-5* was present in other *C. elegans* strains and none of the non-RSL encoding genes was conserved among the more distantly related nematodes (divergence ~80-100 million years ago), *C. briggsae* and *C. remanei*. These data suggested that *srp-5*, *-8*, *-9* and *-10*, arouse from gene duplication events after *C. elegans* diverged from the other nematode species or these genes were descendents from ancestral nematode genes but were lost subsequently from the other *Caenorhabditis* genomes. Regardless of the mechanism of gene ontogeny, the absence of the non-RSL encoding genes from other *Caenorhabditis* genomes argues against these genes playing an essential role in *Caenorhabditis* development or survival. Moreover, we might conclude that these serpins in *C. elegans* most likely represent vestigial (pseudo)genes without any biological significance. However, many RNAs can mediate significant regulatory functions and truncated proteins can have dominate-negative effects depending on the their interactions with other proteins (44, 45, 48, 49). Preliminary analysis showing that homozygous *srp-5* null mutations result in embryonic lethality underscores this possibility (National Bioresource Project, http://shigen.lab.nig.ac.jp/c.elegans/index.jsp). Finally, it is conceivable that cryptic splicing events or the presence of a unique suppressor tRNA, like that for selenocysteine, yield an RSL-encoding SRP-5, SRP-9 or -10 (50, 51). Of note, only SRP-5 contained a premature UGA termination. Conceivably, this codon could lead to insertion of a selenocysteine. However, only one selenoprotein has been identified in *C. elegans* (a thioredoxin reductase) and no other genes encoding for the cis-acting stem-loop structure, the selenocysteine insertion sequence, in the 3' untranslated region of the mRNA have been identified in the database (51). Ultimately, proteomic approaches that help identify low abundance proteins should help determine whether any serpin domains are derived from of these transcripts *in vivo* (52, 53).

In contrast to the non-RSL encoding serpins, the RSL encoding serpins appeared to be well conserved in the
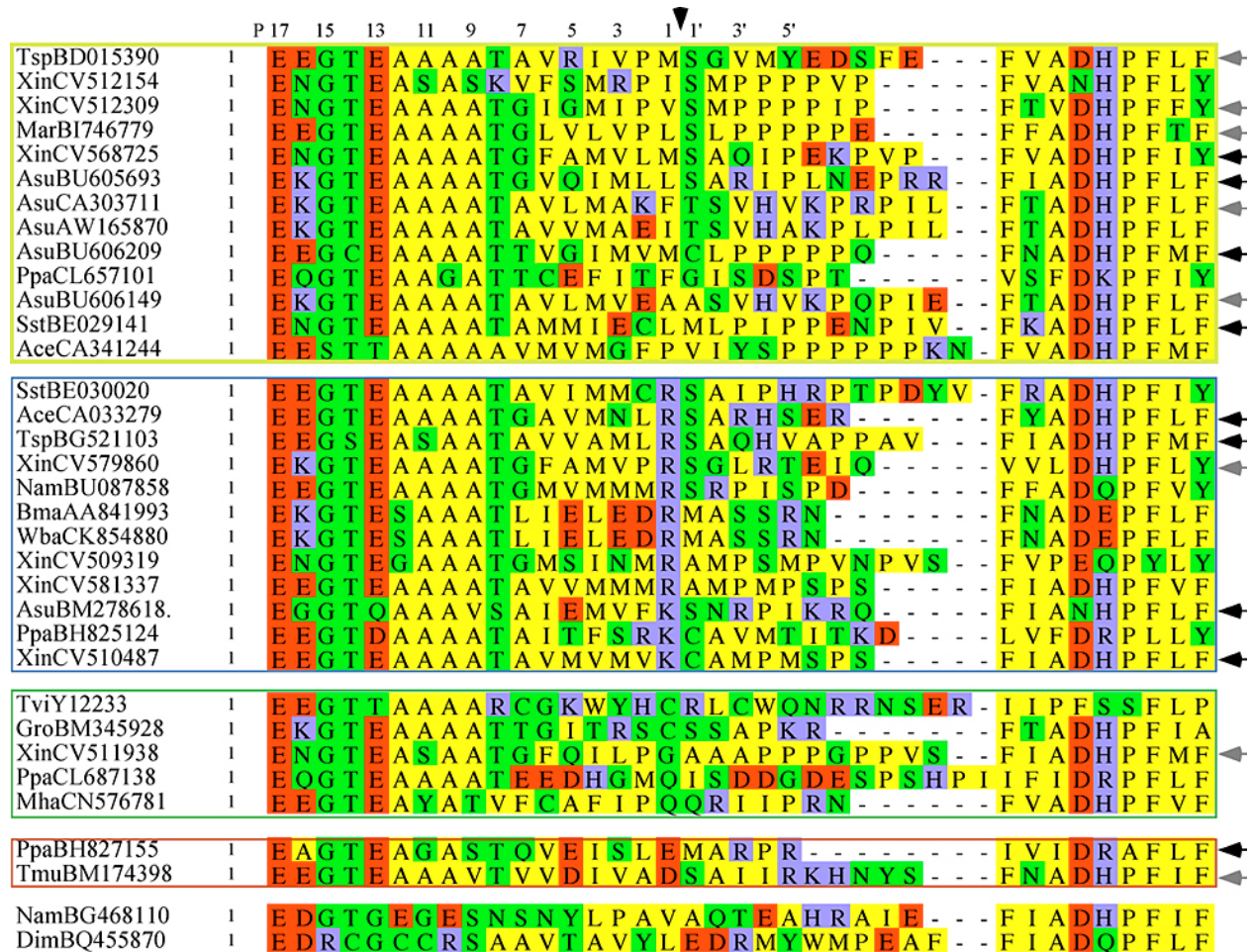
Figure 6. Comparison between non-pathologic and pathologic nematode RSLs. RSL sequences contained within other nematode species were obtained by using the Nem-No-Ele EST and genomic databases and the NemaBLAST server (http://www.nematode.net/BLAST/). RSL peptides from all the *C. elegans* serpins served as the query sequences. Species in which serpin RSL motifs were determined were *Ascaris suum* (Asu), *Xiphinema index* (Xin), *Brugia malayi* (Bma), *Pristionchus pacificus* (Ppa), Meloidogyne *arenaria* (Mar), *Strongyloides stercoralis* (Sst), *Wuchereria bancrofti* (Wba), *Ancylostoma ceylanicum* (Ace), *Trichinella spiralis* (Tsp), *Necator americanus* (Nam), *Trichostrongylus vitrinus* (Tvi), *Trichuris muris* (Tmu), *Meloidogyne hapla* (Mha), *Globodera rostochiensis* (Gro) and *Dirofilaria immitis* (Dim). Genbank accession numbers are indicated next to the organism three-letter name. The colors indicate polar (*green*), nonpolar/hydrophobic (*yellow*), acidic (*red*), and basic (*blue*) residues. The RSL is numbered from P17 to P5'. The RSL sequences are clustered into those that contain a nonpolar/hydrophobic residue (*yellow box*), basic residue (*blue box*), polar uncharged residue (*green box*) or an acidic residue (*red box*) at their putative P1 positions. RSLs that may be cleaved preferentially by papain-like cysteine proteinases due to Phe, Leu or Val residues at (*black arrows*) or near (*gray arrows*) to the P2 position are indicated.

more distantly related *Caenorhabditis* species. The *C. briggsae* and *C. remanei* genomes each contained seven genes, four of which were orthologous to *C. elegans srp-1*, *-2*, *-3* and *-7*. In both *C. briggsae* and *remanei*, the presence of an additional *srp-3*-like gene suggested that either a duplication event occurred after *C. briggsae* and *C. remanei* diverged from *C. elegans* or that a *srp-3*-like gene was lost in the *C. elegans* lineage. Interestingly, in the *C. elegans* genome, all of the *srp* genes were tandemly arrayed along chromosome V, except for *srp-3*. Initially, we suspected that this was an error as we identified two nearly identical cDNA clones that mapped to yeast artificial chromosome clone, Y32G9. One cDNA was identical to

that predicted by *srp-3*. The cDNA sequence of the other clone (which we designated SRP-4) was nearly identical to SRP-3 except for a single base-pair change in the RSL and several nucleotides in the 3' UTR. The nucleotide change in the RSL altered the critical P1 site, converting Val➜Met. Based on the activity of mammalian serpins, the differences in P1 residues should yield proteins with dramatically different inhibitory profiles, and would be analogous to the subtle differences (~95% identical) between the two tandemly arrayed human serpins, *SERPINB3* and *-B4* (54). However, Y32G9 contained only one serpin gene, *srp-3*. Although the absence of "*srp-4*" from Y32G9 could be due to a genomic DNA sequence

assembly error, a Southern blot of *C. elegans* genomic DNA digested with several different restriction enzymes and hybridized with a ^{32}P-labeled, *srp-3/-4* RSL exon probe yielded a single band in all lanes (unpublished data). The failure to find evidence for a local duplication event by either DNA sequencing or Southern blotting suggested that the "SRP-4" cDNA was an artifact of SRP-3 RT-PCR and/or cDNA cloning and that a tandem *srp-3*-like gene (i.e., *srp-4*) was lost from the *C. elegans* genome after its divergence from the other *Caenorhabditis* species.

Amino acid residues around the reactive center (P1-P1') of the RSL are critical determinants of serpin target specificity (3, 4, 55). Of the 23 *Caenorhabditis* serpin RSL sequences identified, 17 had a Phe, Val or Leu at or near the putative P2 position. Since cysteine proteinases prefer these hydrophobic residues at the P2 position (36-39) and a relatively large number of cysteine proteinases reside in the *C. elegans* genome, many of these serpins may help regulate endogenous papain-like cysteine proteinases such as cathepsins B, F, K, L, S and V. Indeed, SRP-2 inhibits papain-like cysteine proteinases (20). Within the parasitic nematode strains, there were 17 serpin RSL sequences that also contained a Phe, Val or Leu residue at or near the putative P2 position. Based on the parasitic lifestyle, these serpins might also serve as virulence factors by neutralizing the potentially lethal effects of host cysteine proteinases.

In conclusion, we have found that only the RSL-encoding, proteinase-inhibitory *C. elegans* serpins were well conserved among other *Caenorhabditis* species. Moreover, based on a comparison between RSL sequences, the serpin inhibitory profiles of the free-living *Caenorhabditis* species appeared to be remarkably similar to those of the pathologic nematodes. These data suggested that the RSL-encoding, proteinase-inhibitory serpins serve similar functions in all nematode species and that perturbations in serpin function that lead to deleterious effects in *Caenorhabditis* species may help identify novel therapeutic targets in parasitic organisms.

6. ACKNOWLEDGEMENT

7. REFERENCES

1. Cole, E. B., D. Miller, D. Rometo, R. M. Greenberg, D. Bromme, S. Cataltepe, S. C. Pak, D. R. Mills, G. A. Silverman & C. J. Luke: Identification and Activity of a Lower Eukaryotic Serine Proteinase Inhibitor (Serpin) from Cyanea capillata: Analysis of a Jellyfish Serpin, Jellypin. *Biochemistry*, 43, 11750-11759 (2004)

2. Irving, J. A., P. J. Steenbakkers, A. M. Lesk, H. J. Op den Camp, R. N. Pike & J. C. Whisstock: Serpins in prokaryotes. *Mol Biol Evol*, 19, 1881-1890 (2002)

3. Silverman, G. A., P. I. Bird, R. W. Carrell, P. B. Coughlin, P. G. Gettins, J. I. Irving, D. A. Lomas, C. J. Luke, R. W. Moyer, P. A. Pemberton, E. Remold-O'Donnell, G. S. Salvesen, J. Travis & J. C. Whisstock: The serpins are an expanding superfamily of structurally similar but funtionally diverse proteins: Evolution, mechanism of inhibition, novel functions, and a revised nomenclature. *J Biol Chem*, 276, 33293-33296 (2001)

4. Gettins, P. G.: Serpin structure, mechanism, and function. *Chem Rev*, 102, 4751-4803 (2002)

5. Carrell, R. W. & D. L. I. Evans: Serpins: mobile conformations in a family of proteinase inhibitors. *Curr. Opin. Struct. Biol.*, 2, 438-446 (1992)

6. Patston, P. A., P. Gettins, J. Beechem & M. Schapira: Mechanism of serpin action: evidence that C1 inhibitor functions as a suicide substrate. *Biochemistry*, 30, 8876-8882 (1991)

7. Travis, J., A. Guzdek, J. Potempa & W. Watorek: Serpins: structure and mechanism of action. *Biol. Chem. Hoppe Seyler*, 71, 3-11 (1990)

8. Schechter, I. & A. Berger: On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Comm.*, 27, 157-162 (1967)

9. Huntington, J. A., R. J. Read & R. W. Carrell: Structure of a serpin-protease complex shows inhibition by deformation. *Nature*, 407, 923-926 (2000)

10. Matheson, N. R., H. van Halbeek & J. Travis: Evidence for a tetrahedral intermediate complex during serpin-proteinase interactions. *J. Biol. Chem.*, 266, 13489-13491 (1991)

11. Plotnick, M. I., L. Mayne, N. M. Schechter & H. Rubin: Distortion of the active site of chymotrypsin complexed with a serpin. *Biochemistry*, 35, 7586-7590 (1996)

12. Irving, J. A., R. N. Pike, A. M. Lesk & J. C. Whisstock: Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function. *Genome Res*, 10, 1845-1864 (2000)

13. Askew, D. J., Y. S. Askew, Y. Kato, C. J. Luke, S. C. Pak, D. Bromme & G. A. Silverman: The amplified mouse squamous cell carcinoma antigen gene locus contains a serpin (serpinb3b) that inhibits both papain-like cysteine and trypsin-like serine proteinases. *Genomics*, 84, 166-175 (2004)

14. Askew, D. J., Y. S. Askew, Y. Kato, R. F. Turner, K. Dewar, J. Lehoczky & G. A. Silverman: Comparative genomic analysis of the clade B serpin cluster at human chromsome 18q21: amplification within the mouse squamous cell carcinoma antigen gene locus. *Genomics*, 84, 176-184 (2004)

15. Remold-O'Donnell, E.: The ovalbumin family of serpin proteins. *FEBS Lett.*, 315, 105-108 (1993)

16. Silverman, G. A., J. C. Whisstock, D. J. Askew, S. C. Pak, C. Luke, S. Cataltepe, J. A. Irving & P. I. Bird: Human clade B serpins (ov-serpins) belong to a cohort of evolutionarily-dispersed intracellular proteinase inhibitor clades that protect cells from promiscuous proteolysis. *Cell. Mol. Life Sci.*, 61, 301-325 (2004)

17. Bird, P. I. & G. A. Silverman: Ovalbumin Serpins. Nature Publishing Group, London (2003)

18. Dougherty, K. M., J. M. Pearson, A. Y. Yang, R. J. Westrick, M. S. Baker & D. Ginsburg: The plasminogen activator inhibitor-2 gene is not required for normal murine development or survival. *Proc Natl Acad Sci U S A*, 96, 686-691 (1999)

19. Gao, F., H. Y. Shi, C. Daughty, N. Cella & M. Zhang: Maspin plays an essential role in early embryonic development. *Development*, 131, 1479-1489 (2004)

20. Pak, S. C., V. Kumar, C. Tsu, C. J. Luke, Y. S. Askew, D. J. Askew, D. R. Mills, D. Bromme & G. A. Silverman: Srp-2 is a cross-class inhibitor that participates in post-embryonic development of the nematode Caenorhabditis elegans: Initial characterization of the Clade L serpins. *J Biol Chem* (2004)

21. Whisstock, J. C., J. A. Irving, S. P. Bottomley, R. N. Pike & A. M. Lesk: Serpins in the Caenorhabditis elegans genome. *Proteins*, 36, 31-41 (1999)

22. Altschul, S. F., W. Gish, W. Miller, E. W. Myers & D. J. Lipman: Basic local alignment search tool. *J. Mol. Biol.*, 215, 403-410 (1990)

23. Emmons, S. W.: The Nematode Caenorhabditis elegans: The Genome. Cold Spring Harbor Laboratory Press, Plainview (1988)

24. Lewis, J. A. & J. T. Fleming: Basic culture methods. *Methods Cell Biol*, 48, 3-29 (1995)

25. Jeanmougin, F., J. D. Thompson, M. Gouy, D. G. Higgins & T. J. Gibson: Multiple sequence alignment with Clustal X. *Trends Biochem Sci*, 23, 403-405. (1998)

26. Hayakawa, Y., Y. Hirashima, M. Kurimoto, N. Hayashi, H. Hamada, N. Kuwayama & S. Endo: Contribution of basic residues of the A helix of heparin cofactor II to heparin- or dermatan sulfate-mediated thrombin inhibition. *FEBS Lett*, 522, 147-150 (2002)

27. Wang, Q. & S. Shaltiel: Distal hinge of plasminogen activator inhibitor-1 involves its latency transition and specificities toward serine proteases. *BMC Biochem*, 4, 5-20 (2003)

28. Kruger, O., J. Ladewig, K. Koster & H. Ragg: Widespread occurrence of serpin genes with multiple reactive centre-containing exon cassettes in insects and nematodes. *Gene*, 293, 97-105 (2002)

29. Reboul, J., P. Vaglio, N. Tzellas, N. Thierry-Mieg, T. Moore, C. Jackson, T. Shin-i, Y. Kohara, D. Thierry-Mieg, J. Thierry-Mieg, H. Lee, J. Hitti, L. Doucette-Stamm, J. L. Hartley, G. F. Temple, M. A. Brasch, J. Vandenhaute, P. E. Lamesch, D. E. Hill & M. Vidal: Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in C. elegans. *Nat Genet*, 27, 332-336 (2001)

30. Huber, R. & R. W. Carrell: Implications of the three-dimensional structure of a 1- antitrypsin for structure and function of serpins. *Biochemistry*, 28, 8951-8966 (1989)

31. Rosenblum, J. S. & J. W. Kozarich: Prolyl peptidases: a serine protease subfamily with high potential for drug discovery. *Curr Opin Chem Biol*, 7, 496-504 (2003)

32. Shan, L., Mathews, II & C. Khosla: Structural and mechanistic analysis of two prolyl endopeptidases: role of interdomain dynamics in catalysis and specificity. *Proc Natl Acad Sci U S A*, 102, 3599-3604 (2005)

33. Pemberton, P. A., D. T. Wong, H. L. Gibson, M. C. Kiefer, P. A. Fitzpatrick, R. Sager & P. J. Barr: The tumor suppressor maspin does not undergo the stressed to relaxed transition or inhibit trypsin-like serine proteases. Evidence that maspin is not a protease inhibitory serpin. *J Biol Chem*, 270, 15832-7 (1995)

34. Riddle, D. L., T. Blumenthal, B. J. Meyer & J. R. Priess: C. elegans II. Cold Spring Harbor Laboratory Press, Plainview, New York (1977)

35. Ragg, H., T. Lokot, P. B. Kamp, W. R. Atchley & A. Dress: Vertebrate serpins: construction of a conflict-free phylogeny by combining exon-intron and diagnostic site analyses. *Mol Biol Evol*, 18, 577-584 (2001)

36. Bromme, D., P. R. Bonneau, P. Lachance & A. C. Storer: Engineering the S_2 subsite specificity of human cathepsin S to a cathepsin L- and cathepsin B-like specificity. *J. Biol. Chem.*, 269, 30238-30242 (1994)

37. Bromme, D., P. R. Bonneau, P. Lachance, B. Wiederanders, H. Kirschke, C. Peters, D. Y. Thomas, A. C. Storer & T. Vernet: Functional expression of human cathepsin S in *Saccharomyces cerevisiae*. *J. Biol. Chem.*, 268, 4832-4838 (1993)

38. Bromme, D., Z. Li, M. Barnes & E. Mehler: Human cathepsin V functional expression, tissue distribution, electrostatic surface potential, enzymatic characterization, and chromosomal localization. *Biochemistry*, 38, 2377-2385. (1999)

39. Bromme, D., K. Okamoto, B. B. Wang & S. Biroc: Human cathepsin O2, a matrix protein-degrading cysteine protease expressed in osteoclasts. Functional expression of human cathepsin O2 in Spodoptera frugiperda and characterization of the enzyme. *J. Biol. Chem.*, 271, 2126-2132 (1996)

40. Henrich, S., A. Cameron, G. P. Bourenkov, R. Kiefersauer, R. Huber, I. Lindberg, W. Bode & M. E. Than: The crystal structure of the proprotein processing proteinase furin explains its stringent specificity. *Nat Struct Biol*, 10, 520-526 (2003)

41. Zhou, A., G. Webb, X. Zhu & D. F. Steiner: Proteolytic processing in the secretory pathway. *J Biol Chem*, 274, 20745-20748 (1999)

42. Al-Khunaizi, M., C. J. Luke, Y. S. Askew, S. C. Pak, D. J. Askew, S. Cataltepe, D. Miller, D. R. Mills, C. Tsu, D. Bromme, J. A. Irving, J. C. Whisstock & G. A. Silverman: The serpin SQN-5 is a dual mechanistic-class inhibitor of serine and cysteine proteinases. *Biochemistry*, 41, 3189-3199 (2002)

43. Luke, C., C. Schick, C. Tsu, J. C. Whisstock, J. A. Irving, D. Bromme, L. Juliano, G. P. Shi, H. A. Chapman & G. A. Silverman: Simple modifications of the serpin reactive site loop convert SCCA2 into a cysteine proteinase inhibitor: a critical role for the P3' proline in facilitating RSL cleavage. *Biochemistry*, 39, 7081-7091. (2000)

44. Maquat, L. E.: Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol*, 5, 89-99 (2004)

45. Mitrovich, Q. M. & P. Anderson: mRNA surveillance of expressed pseudogenes in C. elegans. *Curr Biol*, 15, 963-967 (2005)

46. McKay, S. J., R. Johnsen, J. Khattra, J. Asano, D. L. Baillie, S. Chan, N. Dube, L. Fang, B. Goszczynski, E. Ha, E. Halfnight, R. Hollebakken, P. Huang, K. Hung, V. Jensen, S. J. Jones, H. Kai, D. Li, A. Mah, M. Marra, J. McGhee, R. Newbury, A. Pouzyrev, D. L. Riddle, E. Sonnhammer, H. Tian, D. Tu, J. R. Tyson, G. Vatcher, A. Warner, K. Wong, Z. Zhao & D. G. Moerman: Gene expression profiling of cells, tissues, and developmental stages of the nematode C. elegans. *Cold Spring Harb Symp Quant Biol*, 68, 159-169 (2003)

47. Kim, S. K., J. Lund, M. Kiraly, K. Duke, M. Jiang, J. M. Stuart, A. Eizinger, B. N. Wylie & G. S. Davidson: A gene expression map for Caenorhabditis elegans. *Science*, 293, 2087-2092 (2001)

48. Frischmeyer, P. A. & H. C. Dietz: Nonsense-mediated mRNA decay in health and disease. *Hum Mol Genet*, 8, 1893-1900 (1999)

49. Maquat, L. E.: Nonsense-mediated mRNA decay in mammals. *J Cell Sci*, 118, 1773-1776 (2005)

50. Kryukov, G. V., S. Castellano, S. V. Novoselov, A. V. Lobanov, O. Zehtab, R. Guigo & V. N. Gladyshev: Characterization of mammalian selenoproteomes. *Science*, 300, 1439-1443 (2003)

51. Taskov, K., C. Chapple, G. V. Kryukov, S. Castellano, A. V. Lobanov, K. V. Korotkov, R. Guigo & V. N. Gladyshev: Nematode selenoproteome: the use of the selenocysteine insertion system to decode one codon in an animal genome? *Nucleic Acids Res*, 33, 2227-2238 (2005)

52. Kaji, H., H. Saito, Y. Yamauchi, T. Shinkawa, M. Taoka, J. Hirabayashi, K. Kasai, N. Takahashi & T. Isobe: Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nat Biotechnol*, 21, 667-672 (2003)

53. Mawuenyega, K. G., H. Kaji, Y. Yamuchi, T. Shinkawa, H. Saito, M. Taoka, N. Takahashi & T. Isobe: Large-scale identification of Caenorhabditis elegans proteins by multidimensional liquid chromatography-tandem mass spectrometry. *J Proteome Res*, 2, 23-35 (2003)

54. Schneider, S. S., C. Schick, K. E. Fish, E. Miller, J. C. Pena, S. D. Treter, S. M. Hui & G. A. Silverman: A serine proteinase inhibitor locus at 18q21.3 contains a tandem duplication of the human squamous cell carcinoma antigen gene. *Proc. Natl. Acad. Sci. USA*, 92, 3147-3151 (1995)

55. Potempa, J., E. Korzus & J. Travis: The serpin superfamily of proteinase inhibitors: structure, function, and regulation. *J. Biol. Chem.*, 269, 15957-15960 (1994)

Send all correspondence to: Dr Gary A. Silverman, Magee-Womens Research Institute, 204 Craft Avenue, Pittsburgh, PA 15213, USA, Tel: 412-641-4111, Fax: 412-641-5425, E-mail: gsilverman@magee.edu