A Comprehensive Phylogenetic Analysis of the Serpin Superfamily

Matthew A. Spence ,*[1] Matthew D. Mortimer,[1] Ashley M. Buckle ,[2] Bui Quang Minh ,[3] and Colin J. Jackson*[1,4,5]

[1]Research School of Chemistry, Australian National University, Canberra, ACT, Australia

[2]Department of Biochemistry and Molecular Biology, Biomedicine Discovery Institute, Monash University, Clayton, Melbourne, VIC, Australia

[3]Research School of Computing and Research School of Biology, Australian National University, Canberra, ACT, Australia

[4]Australian Research Council Centre of Excellence for Innovations in Peptide and Protein Science, Research School of Chemistry, Australian National University, Canberra, ACT, Australia

[5]Australian Research Council Centre of Excellence in Synthetic Biology, Research School of Chemistry, Australian National University, Canberra, ACT, Australia

*__Corresponding authors:__ E-mails: colin.jackson@anu.edu.au; matthew.spence@anu.edu.au.

Associate editor: Julian Echave

Abstract

Serine protease inhibitors (serpins) are found in all kingdoms of life and play essential roles in multiple physiological processes. Owing to the diversity of the superfamily, phylogenetic analysis is challenging and prokaryotic serpins have been speculated to have been acquired from Metazoa through horizontal gene transfer due to their unexpectedly high homology. Here, we have leveraged a structural alignment of diverse serpins to generate a comprehensive 6,000-sequence phylogeny that encompasses serpins from all kingdoms of life. We show that in addition to a central "hub" of highly conserved serpins, there has been extensive diversification of the superfamily into many novel functional clades. Our analysis indicates that the hub proteins are ancient and are similar because of convergent evolution, rather than the alternative hypothesis of horizontal gene transfer. This work clarifies longstanding questions in the evolution of serpins and provides new directions for research in the field of serpin biology.

Key words: **serpin, phylogenetics, evolution.**

Introduction

The serine protease inhibitor (serpin) superfamily is the largest group of protease inhibitors in nature (Gettins 2002). Serpins have been identified throughout all kingdoms of life: animals, plants, fungi, protists, archaea, and bacteria (Law et al. 2006). It is notable that eukaryotic serpins are significantly more abundant than their prokaryotic counterparts, which are found in only a few lineages of bacteria and archaea (Irving et al. 2002). Indeed, many eukaryotic serpins have extremely well-understood physiological roles (e.g., alpha-1 antitrypsin, A1AT), whereas the functions of prokaryotic serpins are relatively enigmatic. This has contributed to controversy regarding the progenitor of prokaryotic serpins and the evolutionary origins of the superfamily (Irving et al. 2002; Roberts et al. 2004; Ivanov et al. 2006; Kantyka et al. 2010; Goulas et al. 2017). The ubiquitous presence of serpins in plant and metazoan biology has led to the hypothesis that serpins are a relatively young superfamily that emerged in the last common ancestor of eukaryotes (Irving et al. 2000). Under this model of serpin evolution, prokaryotes acquired serpins via interkingdom horizontal gene transfer from a eukaryote (Irving et al. 2002).

Canonical, inhibitory members of the superfamily have evolved to exploit the energetic difference between two physiologically relevant conformations: in the metastable stressed (native) state, a solvent exposed and flexible reactive center loop (RCL) protrudes from the central β-sheet of the protein (the A-sheet), resembling the typical unstructured peptide substrates of target proteases. Cleavage of the RCL at the scissile bond by an attacking protease induces a conformational change to the relaxed, lower energy cleaved state, in which the cleaved RCL is inserted as an additional strand of the A-sheet. The stressed (S) to relaxed (R) conformational transition kinetically traps the covalent serpin-protease complex, irreversibly inhibiting the attacking protease by distorting the active site residues. Serpins thus function as irreversible, suicide inhibitors. The specificity of serpins toward their cognate protease targets is determined by the amino-acid sequence of the

RCL (Huntington 2011), the complex interplay between dynamics in and around the RCL, and local electrostatics (Marijanovic et al. 2019), as well as auxiliary exosites that stabilize the initial noncovalent serpin-protease association complex prior to RCL cleavage (Gettins and Olson 2009). The functional dependence on the relative energies of the R and S states makes the serpin inhibitory mechanism particularly-well suited to allosteric regulation by cofactors; stabilization or destabilization of one conformation by a ligand can alter this energetic balance and modulate serpin inhibitory function. The archetype of allosteric modulation in the serpins is the regulation of antithrombin by the binding of heparin to an exosite, which accelerates inhibition by acting as a template for both serpin and protease (Johnson et al. 2006; fig. 1).

Although canonical serine protease inhibition is the dominant function within the superfamily and has been characterized in great detail, it is clear that the functional diversity of the serpins extends well beyond serine protease inhibition. For example, cross-class and papain-like cysteine protease inhibitors have been documented (Zhou et al. 1997; Schick et al. 1998; Kantyka et al. 2011; Guo et al. 2015). Many members of the serpin superfamily have also been shown to possess noninhibitory functions. In these instances, the noninhibitory functions typically exploit the conformational transition from S to R states, such as the corticosteroid and thyroxine-binding globulins that transport steroid hormones in higher eukaryotes via differential affinity for steroid ligands in the R and S forms (Zhou et al. 2006, 2008).

Understanding the molecular basis of the functional divergence within protein superfamilies can provide insight into the biophysical properties that permit or constrain functional radiation in protein evolution and reveal the determinants of novel phenotypes, as well as the selective pressures that compel organisms to acquire them. Modeling the evolution of protein superfamilies is often challenging due to poor phylogenetic signal, extensive sequence diversity, and large sequence data sets that have been diverging over geological timescales. This is particularly true of serpins; the superfamily's broad distribution and extensive sequence divergence have limited previous phylogenetic analyses to specific serpin families and taxonomic groups, or obscured the topology of deep branches in the serpin phylogeny (Irving et al. 2000; Heit et al. 2013). Additionally, the number and diversity of serpin sequences belonging to prokaryotes have continued to expand since the advent of the genomic age, further clouding the evolutionary history of serpins. Here, we leverage structural information from the consensus serpin fold to perform comprehensive phylogenetic analysis of the serpin superfamily. This phylogeny provides new insight into the evolution of serpins across the various kingdoms of life and has generated new hypotheses relating to the possible biological function of a number of previously sparsely characterized clades and the origins of the superfamily.

Results and Discussion

A Global Perspective on Serpin Sequence Space from Sequence Similarity Network Analysis

To investigate global serpin sequence diversity, we generated an unbiased and nonredundant data set of 18,233 serpin proteins and putative gene products. This data set was then analyzed as a sequence similarity network (SSN), in which nodes represent sequences (or redundant clusters of sequences) and edges connecting nodes embody similarity scores above an arbitrary similarity threshold (fig. 2). Unlike phylogenetic analysis, an SSN can only convey information relating to the similarity between sequences in a data set and does not explicitly model sequence evolution or phylogenetic relatedness (Atkinson et al. 2009; Copp et al. 2018); however, SSNs have been profoundly insightful and have proven to be useful in revealing sequence–function relationships in a variety of protein superfamilies without the technical and computational limitations and difficulties of full phylogenetic inference (Akiva et al. 2014, 2017; Baier and Tokuriki 2014; Ahmed et al. 2015; Wichelecki et al. 2015).

In the global view of serpin sequence space generated by SSN analysis, 10,123 nodes represent 18,233 amino acid sequences (sequences with >75% similarity are collapsed into a single node) that constitute the serpin protein superfamily (pfam: PF0079) and edges connect nodes that share greater than 40% pairwise sequence identity. We employed the "divide and conquer" workflow of Akiva and Copp (Akiva et al. 2017; Copp et al. 2018) in conjunction with simulated Markov network clustering (MCL) (Morris et al. 2011) to define 481 discrete and highly interconnected clusters of sequences that are expected to be functionally distinct from one another. Indeed, this sequence clustering workflow distinguished annotated subgroups of highly homogenous Group A and Group B serpins from one another exclusively on the basis of SSN topology (supplementary fig. 1, Supplementary Material online), indicating that our classification criteria could effectively differentiate distinct serpin families.

The serpin superfamily SSN shares many characteristics with those of other protein superfamilies and biological systems. The network is approximately scale-free (i.e., the edge distribution follows a power-law), as is often typical of biological and evolutionary networks, and serpins from closely related taxa tend to cluster together. Most sequences belong to a central connected component that exhibits a hub-and-spoke topology, in which divergent groups radially descend from a central cluster of sequences that we hereafter refer to as the "hub." The central component includes chordate Groups A, B, C, E, and I, arthropod Group K and plant Group P, according to the classification scheme defined by (Irving et al. 2000), as well as the majority of unclassified prokaryotic serpins (fig. 2). Many serpins that have diverged from canonical inhibitors, such as uterine associated serpins (UMAP, UFAP) and HSP47 are disconnected from the central network component, reflecting their functional divergence (Ing and Roberts 1989; Widmer et al. 2012).

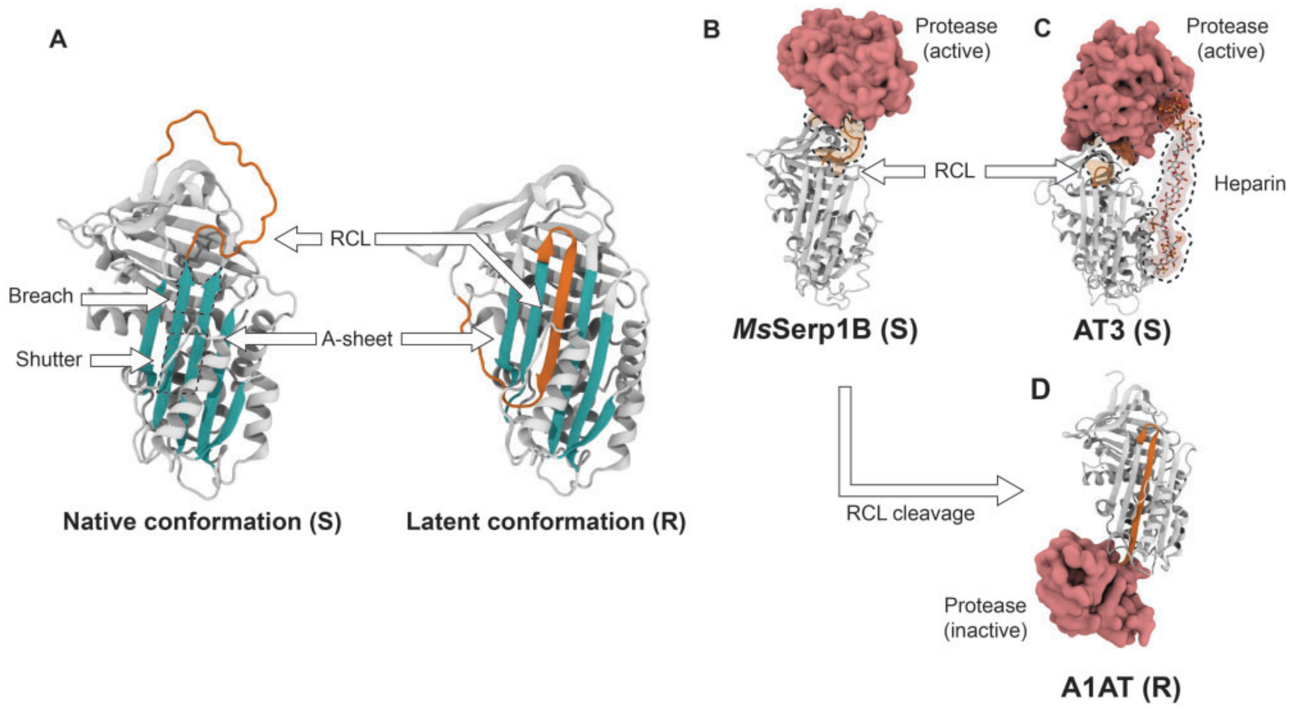

Fig. 1. Serpin fold and mechanistic diversity. (S) and (R) denote stressed and relaxed conformations, respectively. (*A*) Cartoon representation of the S to R state conformational transition, in alpha-1-antitrypsin (A1AT). The RCL in both native (exposed) and latent (inserted) conformations is highlighted, among other key structural components of the serpin fold, such as the breach, shutter, and A-sheet (PDBs: native stressed conformation 3NE4, latent relaxed conformation 1IZ2). (*B*) Noncovalent Michaelis complex between *Manduca sexta* serpin1B (*Ms*Serp1B) and active trypsin protease (PDB: 1K9O). (*C*) Ternary michaelis complex between human antithrombin III (AT3) with heparin templated active thrombin (PDB: 1TB6). (*D*) Covalent complex between cleaved A1AT and inhibited trypsin, post-RCL cleavage (PDB: 1EZX).

The Serpin Network Hub Consists of Similar Chordate and Prokaryotic Serpins

The most prominent topological feature of the serpin SSN is the presence of a heterologous hub group that other clustered sequences radiate from. Hub groups are observed in other protein SSNs and sequences that occupy hub-like positions in SSNs generally share broad similarity across sequence space, because of this, hub-like sequences tend to represent the most consensus- or ancestral-like contemporary sequences (Akiva et al. 2017). The peculiarity of the serpin SSN hub is that it is composed of serpins that are highly similar to each other, yet belong to taxa that are distantly related. This is particularly evidenced by the high similarity shared between chordate Group B serpins with archaeal and bacterial serpins. On the premise of sequence similarity alone, chordate Group B serpins appear to be more closely related to almost all groups of prokaryotic serpins than to other members of the chordate serpin lineage that they are known a priori to belong to (Irving et al. 2000; Heit et al. 2013), such as chordate Groups H, D, and even other members within Group B (UMAP, UFAP).

The incongruence of microbial and chordate serpins sharing significant homology has led to the hypothesis that prokaryotic serpins were acquired via a rare interkingdom horizontal gene transfer event recently in serpin evolution (herein referred to as the HGT hypothesis) (Irving et al. 2002; Roberts et al. 2004; Ivanov et al. 2006; Kantyka et al. 2010; Goulas et al. 2017). One underlying assumption of the HGT hypothesis is the existence of a viable mechanism for genetic transfer between a chordate and recipient prokaryote and a clear selective advantage provided by the transferred genetic material (Ochman et al. 2000; Koonin et al. 2001). This assumption may be valid when considering exclusively commensal and pathogenic prokaryotes that coexist with chordate hosts, such as the prokaryotic serpin miropin from the opportunistic pathogen *Tannerella forsythia*; however, the HGT hypothesis is not supported by all of the available data and is inconsistent with the full breadth of free-living prokaryotes (such as halophilic archaea belonging to *Haloferax*, *Natrialba*, thermophilic archaea belonging to *Thermococcus*, psychrophilic bacteria belonging to *Psychorbacter*, free-living soil bacteria belonging to *Sporangium*, *Chondromyces* and free-living marine sediment bacteria belonging to *Beggiatoa*; supplementary fig. 2, Supplementary Material online) that encode chordate-like serpins. Indeed, most of the prokaryotic hub serpins belong to marine and soil bacteria, sulfur and methane metabolizing prokaryotes and environmental extremophiles (with the exception of few belonging to *Elusimicrobia* spp.) and are hence unlikely to play a role in pathogenicity or host–microbe interaction, as in other clusters of prokaryotic serpins (such as those that include miropin and serpins from known commensals and pathogens). Almost all of the hub-like prokaryotic serpins do share the RCL-hinge sequence motif with known inhibitory serpins (P17–P9: ExGTEAAAA, x: E/K/R)

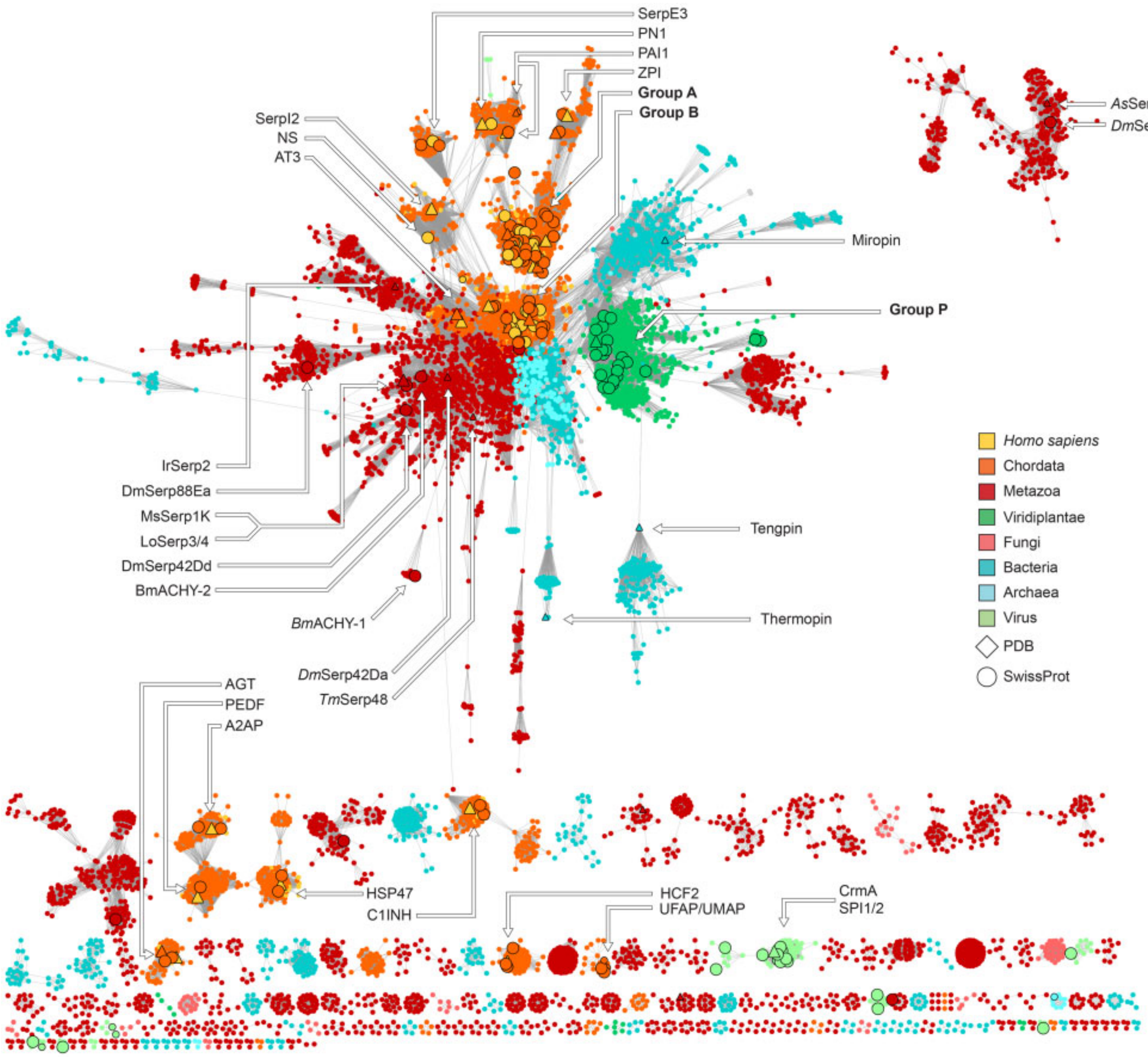

Fig. 2. A nonredundant sequence similarity network of the serpin superfamily. Nodes represent serpin sequences, or clusters of serpin sequences that share 75% pairwise sequence identity. Edges connect sequences that share >40% pairwise sequence identity. Nodes are colored by the taxonomic distribution of organisms that they belong to: nodes that represent chordate and *Homo sapiens* serpins are distinguished at the level of phylum and species for clarity. Sequences with reviewed annotations and solved structures are represented by enlarged circular and triangular nodes with black borders, respectively.

(Hopkins et al. 1993; Irving et al. 2000), indicating that they are competent protease inhibitors that can undergo the stressed-to-relaxed conformational transition.

We hypothesize that the hub-like prokaryotic serpins occupy fundamental, intracellular housekeeping roles, such as controlling cytoplasmic proteolysis and likely resemble ancestral inhibitory serpins. Indeed, the two archaeal serpins within the hub that have experimental annotations (Pnserpin from *Pyrobaculum neutrophilum* and Tkserpin from *Thermococcus kodakaraiensis*) are both potent protease inhibitors that can effectively inhibit endogenous subtilisin and chymotrypsin-like proteases at the extreme temperatures at which the hyperthermophilic organisms reside (Tanaka et al. 2011; Zhang et al. 2017).

Much of Serpin Sequence Space Remains Unexplored

There are many other well-defined sequence clusters beyond the central hub and connected component in the serpin SSN, such as those belonging to angiotensinogen, heat-shock protein 47 (HSP47), UFAP/UMAP, C1 inhibitor (C1INH), and heparin cofactor 2 (HCF2), which have been functionally characterized through comprehensive work on chordate serpins. However, the majority of serpin sequence space lacks annotation. Of the 481 clusters we identified (minimal criteria of at least four unique sequences in a cluster), only 13.5% have representative members with experimental annotation or structural representation. Most of these belong to chordate serpins and despite accounting for the majority of the serpin superfamily's total diversity, only 3.5% of nonchordate metazoan and 5.2% of prokaryotic serpin clusters larger than four unique sequences have at least a single member with experimental annotation. Among the groups that lack any annotation are 40 sequence clusters with comparable size (>20 sequences) to the major Chordate serpin subfamilies that are likely to be functionally distinct. Additionally, the current serpin classification scheme devised from the first published phylogeny of the serpin superfamily includes only a fraction of the sequence diversity that we find here (Irving et al. 2000). There is also no cladistic classification scheme for bacterial and archaeal serpins, emphasizing a broad necessity within the field to experimentally investigate diverse, non-chordate serpins and justifying the need for updated phylogenetic studies on the serpin superfamily.

A Unified Phylogeny of the Serpin Superfamily

To study the evolution of serpins, we performed a comprehensive, superfamily-wide phylogenetic analysis. Whereas SSNs exclusively provide insight on the global similarities shared between protein sequences, full phylogenetic analysis can deconvolute the evolutionary topologies that relate extant sequences to one another at the expense of computational burden. Large-scale phylogenetic inference of full protein superfamilies is technically challenging; homology is often difficult to detect between distantly related members of a superfamily and the difficulty of attaining an accurate sequence alignment is often a barrier to phylogenetic inference.

To overcome this, we devised a workflow to align serpin sequences within our data set based on structural information derived from the conserved serpin fold. Because a protein structure contains more evolutionary information than an amino-acid sequence alone (Shakhnovich et al. 2005), we were able to align a diverse subset of 750 representative serpin sequences (selected as representatives of 750 clusters from the SSN at a 52.5% sequence similarity threshold) to a hidden Markov model (HMM) trained on the distributions of residues observed at each position in 39 structurally aligned and evolutionarily diverse serpin crystal-structures (supplementary fig. 3, Supplementary Material online). Using this structure-guided sequence alignment as a seed for a full sequence data set produced an alignment of 6,000 serpin amino-acid sequences that were aligned at all positions of the consensus serpin fold that are crystallographically resolved. This meant that the RCL (C-terminal of the conserved hinge motif), N- and C-terminal extensions, as well as non-conserved insertions throughout the serpin scaffold that are generally not resolved in or are absent from solved serpin structures were omitted from the alignment. Hypervariability among these sites indicates that they may not be related by homology and should not be included in the alignment regardless, as in previous phylogenetic analyses of the serpins (Irving et al. 2000).

We inferred the serpin superfamily phylogeny by maximum likelihood (ML). Tree inferences were performed with five independent replicates using the free-rate, approximated mixture model LG+R10+F+C10 that explicitly models heterogeneity in both evolutionary rates and amino acid substitution processes across sites using the posterior mean site frequency (PMSF) approximation (Wang et al. 2018). The diversity of sequences in our data set and phylogenetic complexity of modeling evolution in the serpin superfamily warranted the use of such a complex model; indeed using approximated mixture models with PMSF profiles has been shown previously to ameliorate artifactual long-branch attraction in ML phylogenies (Wang et al. 2018). As sequence evolution models should include no more parameters than necessary in phylogenetic reconstruction, the appropriateness of this model was tested against less parameter-rich models (LG, LG+G4, LG+R10, LG+R10+F) by ML and was found to be the model of best fit by both Akaike and Bayesian information criteria (supplementary table 1, Supplementary Material online). We also inferred two replicates of phylogenies under WAG+R10 and LG+R10 to investigate the robustness of our conclusions to model diversity. These additional inferences were uninformative as key bifurcations inferred under these models were reconstructed with very low branch supports (ultrafast bootstrap often <10) that could not confidently place any of the major groups of interest within the superfamily phylogeny. Each of the four topologies inferred under WAG+R10 and LG+R10 were also rejected by the approximately unbiased (AU) test conducted to 10,000 replicates.

Of the five topologies inferred under LG+F+R10+C10, three were rejected by the AU test conducted to 10,000 replicates ($P = 0.0044$, 0.0151, 0.0189) and two were statistically indifferent at explaining the alignment data (Shimodaira 2002). The topology that is presented in figure 3A and used in subsequent analyses failed rejection by the AU test ($P = 0.0703$) and was selected according to a priori expectations from previously published serpin phylogenetic studies (Irving et al. 2000; Krem and Di Cera 2003; Heit et al. 2013) as well as by congruence with the bacterial tree-of-life. The alternate topology that failed rejection by the AU test cannot be excluded as a statistically reasonable evolutionary history contrary to that discussed below; however, its incongruence with prior established understanding of serpin and bacterial evolution led us to favor the topology presented in figure 3. Regardless, this alternate topology (presented and discussed in supplementary fig. 7, Supplementary Material online) supports the new serpin families we define and does not contradict the evolutionary conclusions discussed here as the two topologies that failed rejection by the AU-test differ mainly in the placement of chordate Group C serpins and Actinobacterial serpins. An extensive comparison of each topology is presented in supplementary figure 8, Supplementary Material online.

The phylogeny presented in figure 3A allowed us to define five previously uncategorized families of evolutionarily distinct serpin Groups (Groups Q–U) that belong to clades previously unresolved in phylogenies of the serpin superfamily. Many of the serpin families classified as single groups in previous phylogenetic studies are also much more diverse than originally credited. The Group K arthropod serpins, for example, appear to consist of at least five independent major lineages. However, for simplicity, we maintain the original classification scheme for these groups as the diversity of their

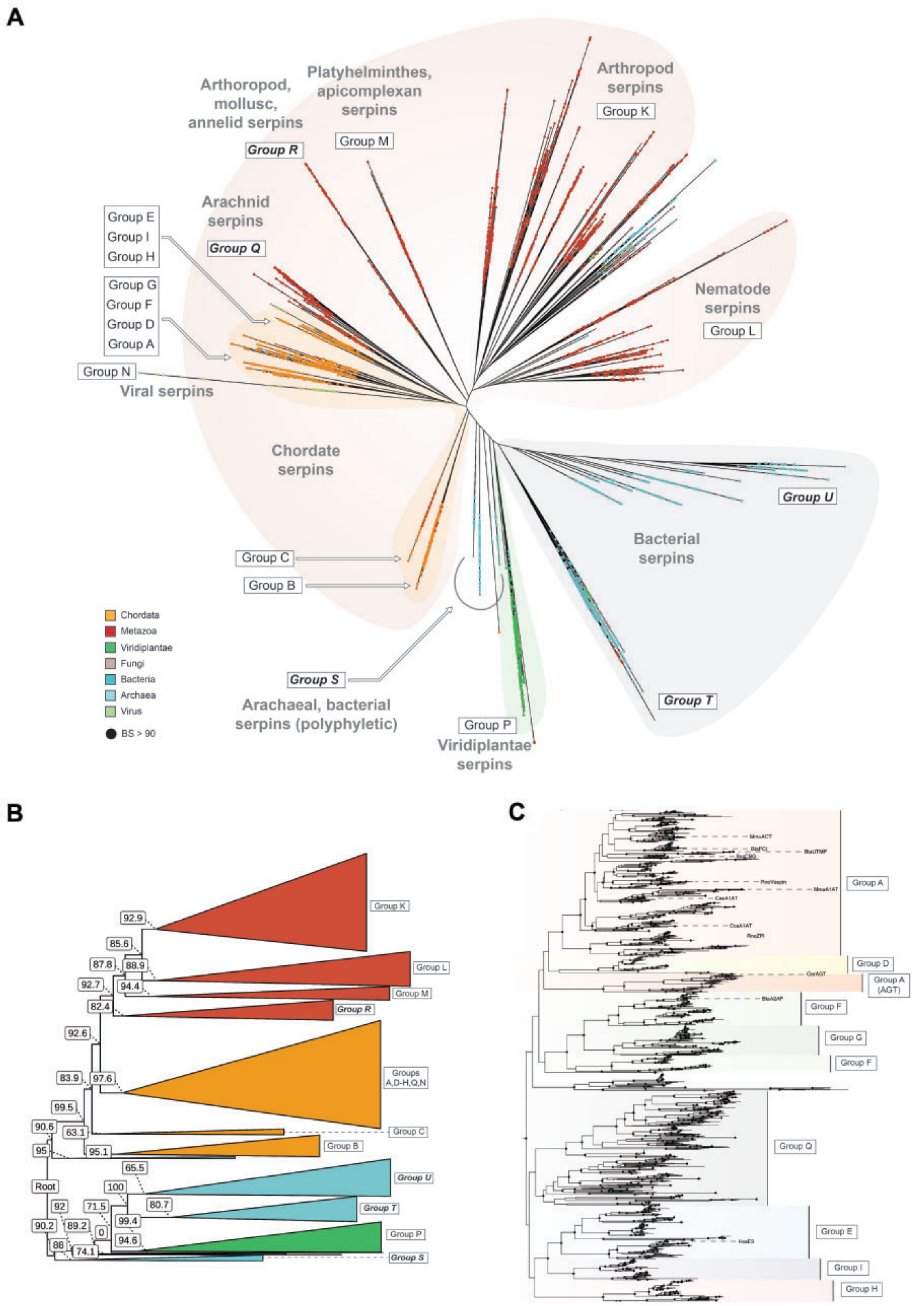

Fig. 3. (*A*) An unrooted maximum likelihood phylogeny of the serpin superfamily. Tips are colored by taxonomic classification and nodes with branch supports (BS) >90 are represented as solid black circles. Well-characterized chordate serpin Groups A, B, C, D, E, F, G, H, and I, known a priori to belong to distinct lineages, are highlighted, as well as Groups K, L, M, and O. New families defined in this study are highlighted in bold. The phylogeny is available in Newick format with full branch support values in S.I. (*B*) View of the phylogeny presented in panel (*A*) with major clades collapsed and BS values overlaid. (*C*) Subtree of chordate serpin Groups A, D–I. Branch color distinguishes each group within the subtree and functionally characterized tips are labeled.

protease targets and cellular functions remain unclear and they are poorly represented by experimentally annotated serpins.

This phylogeny illustrates that the serpin superfamily can be conceptually split into three major groups: one comprising bacterial serpins, one comprising plant serpins, and the other consisting of predominantly metazoan serpins. A smaller polyphyletic group of prokaryotic serpins, including Archaeal serpins (Group S), is also present. The branch that splits the metazoan serpins from the plant and bacterial serpins is the mid-point of the tree and is the most logical position for the root. Serpins belonging to Groups B, P, and S form distinct clades around the root and are the most ancestral members of the superfamily. Metazoan serpins descended from a single common ancestor shared with the Group B serpins and are split into chordate (Groups A and D–I) and nonchordate lineages (Groups K–R) that diverged from one another. Most prokaryotic serpins belong to two lineages (Groups T and U) that shared a single common ancestor early in the superfamily's history. Fungal serpins, together with some prokaryotic and nonchordate metazoan serpins are not resolved as a defined clade, but are instead orphan sequences that were not given a classification.

Serpins in Chordates

The molecular evolution of serpins in chordates has been studied extensively (Irving et al. 2000; Krem and Di Cera 2003; Kumar and Ragg 2008; Heit et al. 2013). The topology of the chordate serpin subtree (fig. 3*B*) is concordant with the established understanding of serpin function and evolution among higher eukaryotes. We find that Group B and C serpins are indeed ancestral to Groups A and D–I, which together form a divergent monophyletic clade. Groups A and D, F and G, and E, I, and H each share a recent common ancestor; however, the cladistic inclusion of Group H with Groups E and I is only marginally supported. Groups B and C belong to sister clades that share an older common ancestor with the other chordate serpin lineages. With the exception of Group H, the topological placement and composition of each of these clades is independently supported by common patterns of intron–exon splice sites and microsynteny (Kumar and Ragg 2008) that were not considered during phylogenetic inference, demonstrating the robustness of the chordate cladistic inclusions.

Serpins in Invertebrates

Nonchordate metazoan serpins are the largest and most diverse group in the superfamily. Despite this, serpins from nonchordate metazoans are paradoxically among the most disproportionately underrepresented in literature and reviewed databases.

Most characterized arthropod serpins belonging to Group K regulate innate immunity. For example, when challenged with infection, *Tenebrio molitor* serpin48 regulates the Toll pathway (Park et al. 2011) and *Drosophila melanogaster* serpin 27A and *Aedes aegypti* serpin2 localize melanin to the site of infection via regulation of the prophenoloxidase proteolytic cascade (An, Budd, et al. 2011; An et al. 2013). Serpin1J regulates both innate immune responses in *Manduca sexta* (Jiang et al. 2003; An, Ragan, et al. 2011), indicating that broad protease specificity can be a feature of arthropod serpins. The Group K serpins also play a role in development. Serpins 16, 18, and 22 from *Bombyx mori* regulate silk-gland development (Guo et al. 2015) and *D. melanogaster* serpins 42 and 27A are involved in developmental protein maturation (Richer et al. 2004) and dorsal–ventral partitioning during embryonic development (Hashimoto et al. 2003; Richer et al. 2004). Many of these paradigms are shared with homologous nematode serpins belonging to Group L (Pak et al. 2004).

Nonchordate metazoan serpins have also evolved ecologically relevant functions in the regulation of exogenous proteases that accommodate parasitic lifestyles or comprise major components in venom. Many of the characterized platyhelminthes serpins belonging to Group M inhibit endogenous and host proteases to dampen inflammation upon infection (Lopez Quezada et al. 2012), alter the proteolytic environment to optimize host penetration (Molehin et al. 2014), or to evade host immune responses (Ghendler et al. 1994; Yang et al. 2014). The inclusion of apicomplexan serpins, which remain largely uncharacterized with some exceptions (Fetterer et al. 2008), within Group M leads us to speculate that this clade comprises serpins predominantly involved in host invasion and accommodating a parasitic lifestyle; however, the functional diversity of this clade remains obscure in our analysis as relatively few Group M serpins have been characterized. Notably, some of the serpins in Group L likely accommodate parasitic life histories in nematodes (Bennuru et al. 2009; Toubarro et al. 2013), and Group K serpins from the parasitoid wasp *Leptopilina boulardi* suppress innate immune responses in *Drosophila* hosts (Colinet et al. 2009), indicating that serpins involved in modulating host immunity and the parasitic microenvironment have emerged independently in multiple lineages over a broad taxonomic range of hosts. Although we maintain the original Group K nomenclature of the arthropod serpins defined by (Irving et al. 2000), here we expand the classification of the family into five major lineages, Groups K1–K5 (supplementary fig. 4, Supplementary Material online).

Among the new serpin families we define in this study are the arachnid Group Q serpins. A peculiarity among the Group Q serpins is their incongruent placement as a sister group to the chordate Groups H, I, and E. A majority of serpins that make up this clade belong to Ixodid ticks, as well as few serpins from the genera *Megacormus* (scorpion), *Tityus* (scorpion), *Parasteatoda* (spider), *Stegodyphus* (spider), and *Sarcoptes* (mite, including *S. scabiei*). The Group Q serpins are also highly similar to chordate serpins and cluster together in SSNs at stringent similarity thresholds, indicating molecular homoplasy between them. It is curious that the majority of taxa that belong to Group Q are chordate parasites (ticks and mites) that have evolved serpins to inhibit host proteases in order to evade adaptive immune responses (Prevot et al. 2009; Chmelar et al. 2011; Xu et al. 2019) or prolong feeding by disrupting hemostatic proteolytic cascades (Chmelar et al. 2011; Mulenga et al. 2013). There is also evidence to suggest that serpins from Group Q form a component of venom in some free-living arachnids (scorpions, spiders) (Gremski et al. 2010; Kazemi and Sabatier 2019), although the protease specificity and precise regulatory function of these proteins remains unknown.

Serpins in Plants

Unlike serpins from other major taxonomic groups, all plant serpins belong to a single, well-defined, and homogenous clade that diverged from a single common ancestor. We maintain the original Group P nomenclature (Irving et al. 2000) that defines the plant serpins as a monophyletic family and is congruent with our cladistic groupings. Like their metazoan counterparts, plant serpins occupy diverse inhibitory and noninhibitory functions (Cohen et al. 2019). Barley protein Z, for example, is a major storage protein during grain filling (Hejgaard et al. 1985). Indeed, grain serpins can comprise as much as 4% of the total storage protein in monocot grains where they are both noninhibitory storage proteins and protease inhibitors that protect storage proteins from proteolytic degradation by endogenous proteases (Evans and Hejgaard 1999). In the cytoplasm, inhibitory plant serpins, such as *Arabidopsis thaliana* Serpin1 (*At*serp1) inhibit ectopic cysteine proteases following vacuolar collapse from abiotic stress (Lampl et al. 2010; Koh et al. 2016) or hypersensitive response when challenged by a pathogen (Lampl et al. 2010; Lema Asqui et al. 2018), thus attenuating programmed cell-death. Plant serpins may also play a role in defence against predation by inhibiting digestive proteases in the midgut of herbivorous arthropods such as aphids (Yoo et al. 2000).

Serpins in Prokaryotes

The physiological role of serpins in many prokaryotes is enigmatic. Prokaryotes lack complex proteolytic signaling pathways and do not utilize the regulatory networks that serpins are often part of in higher eukaryotes. Despite this, serpins are found sparsely distributed throughout some bacterial phyla, with most belonging to *actinobacteria*, *proteobacteria*, *bacteroidetes*, and *firmicutes*. Whereas many of the prokaryotic serpins are indistinguishable from chordate serpins in SSNs, prokaryotic serpins belong to three major families that are phylogenetically distinguished from the chordate serpins. We define these three major prokaryotic groups as S–U. Most bacterial serpins belong to Groups T and U, which shared a common ancestor early in the evolution of the superfamily. Group T is represented by miropin, siropins from *Eubacterium sirium* (Mkaouar et al. 2016), *Bifidobacterium longum* serpin (Ivanov et al. 2006) and tengpin from *Thermoanaerobacter tengcongensis* (Zhang et al. 2007). The only functionally characterized member of Group U is thermopin, belonging to *Thermobifida fusca* (Fulton et al. 2005). Sequence analysis indicates that most of the Group U and T serpins are inhibitory; however, a clade of *Streptomyces* serpins belonging to Group U deviate from a canonical inhibitory serpin in conserved regions of the breach and shutter, suggesting that they may have evolved a noninhibitory function. Despite their highly diverged sequences, these *Streptomyces* serpins share significant homology (E-value <10e-10) with other prokaryotic and eukaryotic serpins in an NCBI BLAST search and many retain the consensus RCL-hinge motif expected in an inhibitory serpin (Hopkins et al. 1993; Irving et al. 2000).

Some bacterial serpins appear to play a role in host–commensal or host–pathogen interaction. Miropin, for example, has recently been demonstrated to efficiently inhibit human plasmin, thereby protecting invading *T. forsythia* cells from plasmin-mediated degradation and attenuating fibrinolysis in the pathogen's local environment (Sochaj-Gregorczyk et al. 2020). Siropins belonging to the commensal bacterium *E. sireum* have been shown to inhibit two human proteases, human neutrophil elastase (HNE) and proteinase3, that are abundant in gastrointestinal (GI) tract where *E. sireum* is prevalent (Mkaouar et al. 2016) and *Bifidobacterium longum* and *B. breve*, both commensal members of the GI microbiome, express serpins capable of inhibiting mammalian proteases present in their microenvironment (Ivanov et al. 2006; Turroni et al. 2010). To investigate the prevalence of prokaryotic serpins in different human microbiota, we performed chemically guided functional profiling (CGFP) (Kaminski et al. 2015) on MCL clusters of prokaryotic serpin sequences. The CGFP workflow matches sequence markers from SSN clusters to markers from metagenomic reads of different human microbiome samples, thus detecting sequence similarity group presence among human microbiota. Many prokaryotic serpins beyond those that have been experimentally characterized have a significant presence in the human oral and GI microbiomes (supplementary fig. 5, Supplementary Material online). The most abundant of these belong to Group T, which includes serpins from many known pathogens and commensal bacteria (such as *Prevotella* spp., *Corynebacterium* spp., and *Lachnospiraceae* spp.) and are closely related to miropin, *Bifidobacterium* spp. serpins and siropins. Many serpins belonging to Group U also have either a presence, or close homologs with a presence, in various human microbiomes. No serpins belonging to Group S, which is composed of both archaeal and bacterial serpins, were detected with significance, indicating that they are likely functionally distinct from the Group S and T serpins in terms of regulatory role and protease specificity.

Serpins Are an Ancient Superfamily

The common ancestry shared between prokaryotic serpins from Groups T and U suggests that bacterial serpins descended from a single common ancestor that was distinct from the last common ancestor of the chordate serpins. We found no phylogenetic evidence supporting the hypothesis that prokaryotes acquired serpins via HGT as no phylogenetic topology (including those that were rejected by the AU-test) placed prokaryotic serpins as descended from chordate serpin lineages, despite their high similarity and coclustering in SSN.

Independent of phylogenetic topology, there are often other molecular markers that may indicate that horizontal gene transfer has occurred. We found few and nonsystematic anomalous $GC\%_{gene}/GC\%_{genome}$ ratios (>1.2 or <0.8) within bacterial serpin genes (supplementary fig. 6, Supplementary Material online). This included many species of *Streptomyces*

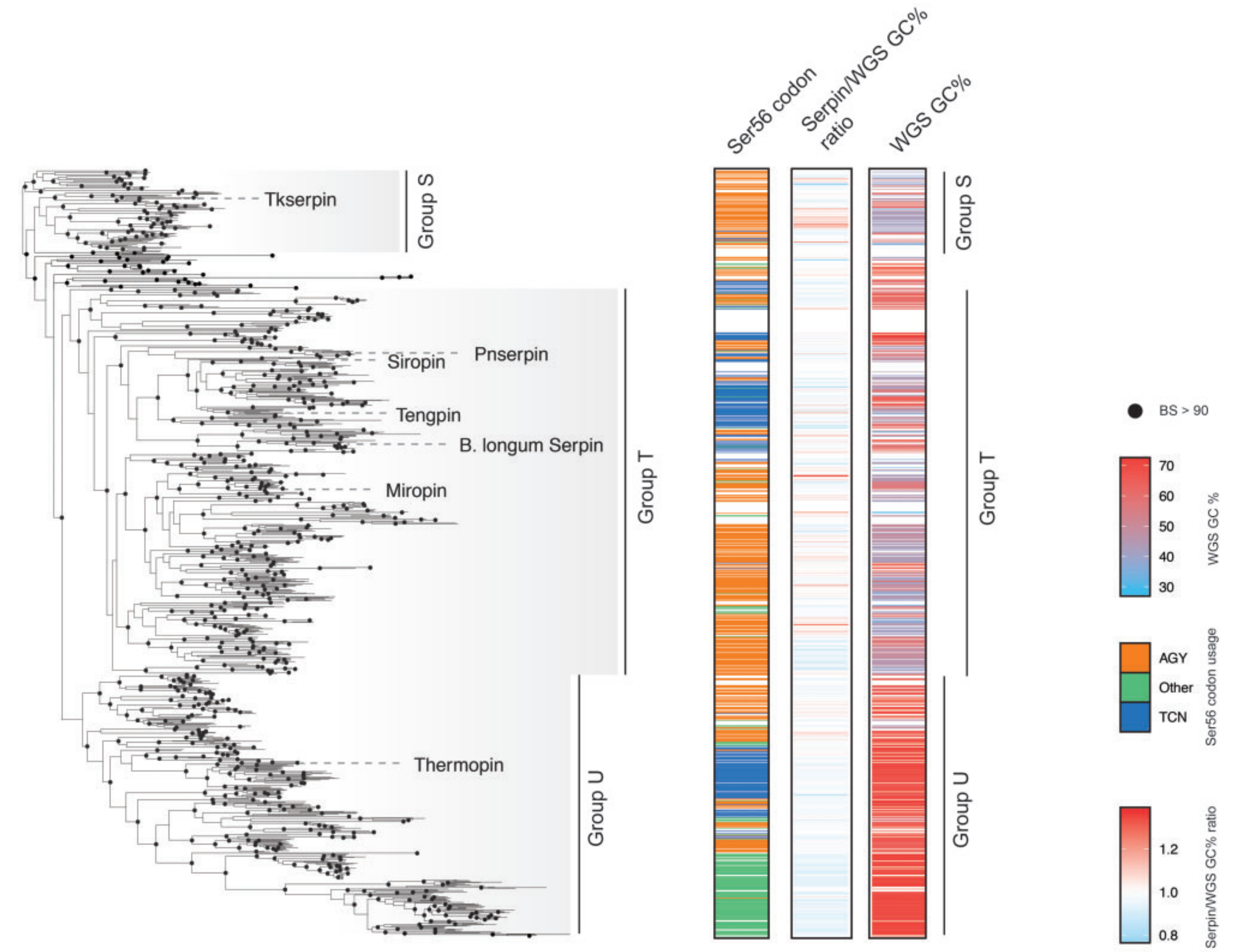

FIG. 4. Genomic and molecular markers of horizontal gene transfer. Where available, codon usage for conserved Ser56, whole-genome sequence (WGS) GC%, and the ratio of serpin gene sequence GC% to WGS GC% are mapped to prokaryotic serpins belonging to Groups S–U. There are no systematic anomalous ratios in serpin gene GC% to WGS GC% over a wide range of whole-genome sequence GC contents (<30% to >70%). Likewise, there is no systematic evidence that Groups S–U were acquired from chordates from the pseudoconserved Ser56 codon usage in prokaryotes.

and *Mycoplasma* that are characterized by high (>70%) and low (<30%) genomic GC content respectively, suggesting that either horizontal gene transfer has not occurred or that genetic transfer events were ancient enough to have ameliorated in extant DNA (fig. 4). The protostome–deuterostome codon usage dichotomy at the conserved position Ser56 (Krem and Di Cera 2003) also provided no evidence of horizontal gene transfer. Ser56 is only marginally conserved in prokaryotic serpins and no particular lineage strictly adheres to either the archaic protostome ACY codon or the TCN codon that is fixed in chordates. Instead, bacterial lineages that use either codon are generally separated by clades where Ser56 has been lost, parsimoniously demonstrating the loss of Ser56 under one codon before refixation under the other (fig. 4). This, corroborated by the absence of a conceivable mechanism of genetic transfer between a chordate and many of the prokaryotes that feature serpins (such as free-living, environmental extremophiles) and the lack of phylogenetic evidence supporting the HGT hypothesis, indicates that the serpin superfamily is ancient and emerged in the last common ancestor of the major bacterial lineages at the latest, independently of eukaryotes. This conclusion is equally supported by the alternative phylogenetic topology that failed rejection by the AU-test (supplementary fig. 7, Supplementary Material online).

Structural Analysis of the Serpin Superfamily Reveals Evidence of Convergent Evolution

A structural comparison between bacterial and chordate serpins reveals several distinct structural adaptations to selective pressures, some of which are shared between families. The majority of these are solutions to the unique structure–function challenge faced by serpins; specifically, how to maintain the relative stabilities of the native and cleaved states in order to function as inhibitors. Serpins from thermophilic bacteria have the added complexity of maintaining this balance at high temperatures. Thermopin, from the moderate thermophilic bacteria *Thermobifida fusca* (optimum growth

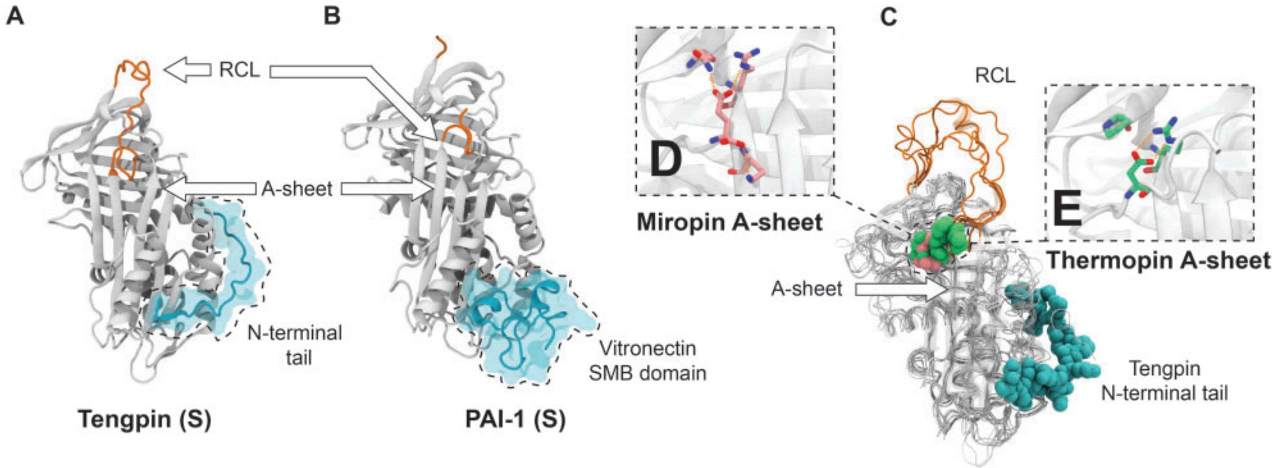

Fig. 5. Structural evidence of convergent evolution in the serpin superfamily. (*A*) Cartoon representation of tengpin in the stressed native state. The N-terminal extension and its contacts in the A-sheet are highlighted, among other structurally important positions, such as the A-sheet and RCL (PDB: 2PEE). (*B*) Cartoon representation of PAI-1 in the stressed native conformation, stabilized by the SMB domain of vitronectin (highlighted in cyan) (PDB: 1OC0). (*C*) A structural alignment of tengpin (PDB: 2PEE), miropin (PDB: 5NCS), thermopin (PDB: 1SNG), SCCA-1 (PDB: 2ZV6), and AT3 (PDB: 3KCG) highlighting the convergent molecular mechanisms exploited to stabilize the native conformation. (*D* and *E*) View of analogous interactions in the C-terminal tails of miropin and thermopin that stabilize the breach, respectively.

temperature 55 °C), contains a unique C-terminal extension that packs against the face of the top of the A-sheet (fig. 5*A*; Fulton et al. 2005). This tail interacts with highly conserved residues in the breach that are known to be critical for controlling serpin stability and function. Tengpin from the extremophilic prokaryote, *Thermoanaerobacter tengcongensis* (optimum growth temperature 75 °C), contains an N-terminal tail that binds to a hydrophobic patch near strand s1A on the body of the protein, stabilizing the native state (Zhang et al. 2007). Indeed, evolution has overcome the thermodynamic challenges imposed by the serpin fold and function many times. Aeropin, from the hyperextremophile archaea *Pyrobaculum aerophilum* (growth temperature 100 °C) (Cabrita et al. 2007), features a C-terminal tail of similar length to that of thermopin and two disulfide bonds that are structurally nonhomologous to disulfides in mesophilic serpins, such as the single disulfide bond in miropin. In both cases, the disulfides contribute significant stabilization energy to the native conformation. Interestingly, structural comparison also shows that the C-terminal region of miropin forms a contact with the breach through a glutamic acid residue, analogous to the aspartate breach interaction in the nonhomologous C-terminal tail of thermopin (fig. 5*C* and *D*). The phylogenetic distance between thermopin (Group U) and the Group T thermophilic serpins from *Pyrobaculum* spp. and tengpin and the analogy between aeropin and miropin disulfide bonds would suggest that thermophilic serpins from different evolutionary backgrounds have independently acquired different and often convergent molecular mechanisms of accommodating the thermodynamic balance required for serpin metastability. Although the structural convergence among bacterial serpins has been acknowledged before (Zhang et al. 2007), their discussion within the broader context of the superfamily phylogeny provides novel insight on convergence among the serpins as a molecular mechanism of overcoming the thermodynamic constraints imposed by the serpin fold.

The biophysical solutions that accommodate metastability have emerged elsewhere in serpin function, as well. The structurally sensitive region of the serpin fold that is stabilized by the N-terminal tail of tengpin is the same site on the mammalian serpin PAI-1 bound by the somatomedin B (SMB) domain of the plasma protein vitronectin (Zhou et al. 2003; Zhang et al. 2007; fig. 5*B*), which functions to stabilize the metastable native state of PAI-1 and thus regulate its function (Declerck et al. 1988). The structural analogy between the N-terminal tail of tengpin and the SMB-binding site of PAI-1 is a stark example of structural convergence where the same biophysical defect in the serpin fold is overcome by analogous mechanisms to achieve different biological results (thermostability in tengpin and regulation by vitronectin in PAI-1). The genetic separation between PAI-1 and the SMB domain of vitronectin is a particularly interesting example of structural convergence that highlights the evolution of a more biologically complex mechanism of regulation in PAI-1 compared with the bacterial tengpin.

Perhaps the most compelling evidence supporting the HGT hypothesis is the apparent structural similarity between the solved crystal structures of miropin and the chordate serpin AT3 (Goulas et al. 2017). Indeed, structural alignment between miropin in a native conformation and tengpin, thermopin, AT3, and SCCA-1 reveals that miropin shares the lowest RMSD with AT3 (1.3 Å) and SCCA-1 (1.6 Å) (compared with 1.8 Å with thermopin and 1.9 Å with tengpin). This is only true when using a single structural model of AT3 in a ternary michaelis complex with thrombin and heparin (PDB: 3KCG). When a more complete conformational ensemble of AT3 is considered in structural analysis with native miropin, including monomeric apo-AT3 structures (PDBs: 2ANT, 1TB6, 1T1F), RMSDs range from 1.6 to 4.1 Å and are on an average 2.7 Å.

The perceived close similarity between miropin and AT3 is likely an artifact stemming from the structural comparison of biologically distinct conformations of miropin (monomeric native) and AT3 (ternary michaelis complex). It is nonetheless an interesting observation (Goulas et al. 2017) that miropin does appear to share more structural features with chordate serpins than what may be expected from proteins that share often <30% sequence identity. In the absence of phylogenetic or molecular evidence supporting the HGT hypothesis, it is most likely that this is yet another example of structural convergent evolution within the serpin superfamily. We hypothesize that structural convergence is such a persistent evolutionary phenomenon among serpins because of the inherent thermodynamic balance that must be maintained by the serpin fold for metastability. It is conceivable that there are only few biophysical solutions to a structure where both stressed and relaxed conformations are accessible. Divergence from that structure would generate noninhibitory serpins, resulting in a fitness landscape with clearly defined fitness peaks that have been converged independently over the evolutionary history of the serpins.

The Evolution of Serpin Function and Location

By identifying N-terminal signal peptides, we were able to classify serpin sequences as likely intracellular, extracellular, periplasmic (bacterial serpins only), or membrane anchored (prokaryotic serpins only) (supplementary fig. 9, Supplementary Material online). The ancestral-like eukaryotic serpins are dominated by sequences that lack N-terminal signal peptides and likely inhibit intracellular proteases, including the Group P plant serpins and the chordate Group B serpins. Extracellular serpin expression emerged within the diverged metazoan lineages, including chordate Groups A, D–H, and Groups K–L of the arthropod and nonchordate metazoan lineages. Notably, many of the Group L nematode serpins lack signal peptides and appear to function in the cytosol. The lower degree of biological organization among prokaryotes makes the functional distinction between intracellular and extracellular prokaryotic serpins less significant than that of eukaryotic serpins. Most prokaryotic serpins feature N-terminal transport peptides, although we were unable to identify systematic trends that relate phylogenetic placement to cellular localization and function among prokaryotic serpins.

We hypothesize that ancestral serpins were intracellularly localized and occupied biologically rudimentary roles, such as protecting cells from promiscuous proteolysis. Intracellular localization is likely a trait that was retained in the ancestral-like extant eukaryotic serpins, which diverged from ancestral functions as biological systems and cellular biology became more complex over evolutionary history. Indeed, there are functional similarities between extant Group B serpins, many of which protect cells from ectopic endolysosomal proteases and a hypothetical primordial serpin that serves solely cytoprotective functions. This paradigm also extends to other ancestral-like Group P plant serpins. The *A. thaliana* *At*serp1, for example, inhibits proteases that have escaped from ruptured vacuoles and ER bodies, thereby protecting cells from programmed death (Lampl et al. 2010). In contrast, the extracellular eukaryotic serpins (as well as few intracellular serpins that do not belong to the ancestral-like lineages, such as HSP47) have diverged into functions that accommodate higher biological organization, such as hormone transport, hemostasis, and immunity, among others. The antithrombin Group C lineages occupy an important position in chordate serpin evolution. Unlike other chordate serpins, those belonging to Group C are not unanimously intracellular or secreted and topologically belong to a monophyletic clade that separates the last common ancestors of Group B and Groups A, D–I, indicating that antithrombin-like serpins were the first to appear extracellularly and likely bridge intracellular and extracellular serpin functions.

Such conclusions are more challenging to draw regarding prokaryotic serpins. Due to their scarce presence in reviewed databases, the extent of functional diversity and protease specificity within Groups S–U is obscure. We can hypothesize, however, that many of the bacterial serpins belonging to Groups T and U have diverged ecological functions involved in host–microbe interaction in vertebrate microbiota, whereas the ancestral-like Group S serpins are undetectable in microbiome metagenomes (supplementary fig. 5, Supplementary Material online). According to our proposed model of serpin evolution, the Group S ancestral-like serpins are likely housekeepers that protect the cell from either endogenous or exogenous environmental proteases, although this conclusion will remain uncertain until our understanding of prokaryotic serpins advances. Additionally, it remains perplexing why most prokaryotic lineages have lost serpins over evolutionary history. Although it is clear from the scarcity of serpins among prokaryotes that the majority of bacteria and archaea have lost their serpin genes over evolutionary history, the lack of commonalities in biochemical niche and life-history traits among the few prokaryotic lineages that have retained serpins makes this difficult to rationalize. For example, serpins are found in both commensal GI bacteria and marine sediment bacteria; understanding the common selective pressures that compel these organisms to retain serpins while the majority of bacteria have lost them remains a major goal in prokaryotic serpin biology. This, with investigating the full diversity of bacteria and their life histories that have serpins, should be the focus of future studies on the evolution of prokaryotic serpins.

Conclusion

In this work, we have leveraged the substantial structural characterization of members of the serpin superfamily to allow us to create a comprehensive phylogeny. This analysis has identified a number of currently uncharacterized, or orphan-function, clades, particularly within nonchordate phlyla. In addition to these diverse and uncharacterized clades, we also observe a large central hub of serpin sequences that includes serpins from bacteria, archaea, metazoa, and plants. We hypothesize that the hub-like prokaryotic serpins occupy primitive, intracellular housekeeping roles, such as controlling rampant cytoplasmic proteolysis and likely resemble ancestral inhibitory serpins. Our analysis indicates that these closely related hub proteins are therefore ancient and are similar because of convergent evolution, rather than the alternative hypothesis of HGT, for which we find little evidence. We hope that this analysis will provide new directions for research in the field of serpin biochemistry, particularly in the characterization of nonchordate metazoa and bacterial serpins.

Materials and Methods

Data Set Collection and SSN

All 18,233 amino acid sequences that comprised the serpin superfamily at the commencement of this study were retrieved from the pfam database (PF00079) using the enzyme function initiative enzyme similarity tool (EFI-EST) server (Gerlt et al. 2015; Zallot et al. 2018). An arbitrary sequence redundancy threshold of 75% sequence identity was imposed on this data set, reducing its size to 10,123 unique serpin sequences before pairwise sequence identities were computed. The resulting SSN was visualized in cytoscape (3.6) using the yFile organic force-directed network layout. Homogenous sequence clusters were defined using the Markov clustering algorithm from the clustermaker cytoscape package (Morris et al. 2011) and validated with experimental evidence from the SwissProt database. Network topology tests, such as centrality metrics, were computed within cytoscape. DNA sequences for serpin encoding genes and prokaryotic genomes were retrieved from NCBI using reference uniprot IDs and organism IDs, respectively. N-terminal signal peptides were identified using the SignalP4.0 algorithm (Petersen et al. 2011).

Sequence Alignment

All sequences that made up the SSN were retrieved from the UniProt database and inspected to confirm their classification as canonical serpins on the basis of the hinge, breach, and shutter motifs. As pairwise identities between sequences in the data set were often <15%, progressive and consistency-based alignment algorithms typically failed to produce robust multiple sequence alignments, which were benchmarked against structural alignments of representative serpin crystal structures. Instead, sequences were aligned according to ensembles of HMMs that were trained iteratively with increasing diversity of sequence data, starting initially from an HMM trained on the distribution of amino acids tolerated at each site of the conserved serpin fold. 39 representative serpin X-ray crystal structures in the stressed conformation (PDBs: 2WXY, 2CEO, 2PEE, 3B9F, 4IF8, 2ZV6, 1JMJ, 2HI9, 1OVA, 1A7C, 3F1S, 4AU2, 5M3Y, 4GA7, 1IMV, 1SEK, 2H4R, 4DTE, 5HGC, 2V95, 1WZ9, 4AJT, 4DY0, 5C98, 1BY7, 2R9Y, 1QMN, 3STO, 3F5N, 1SNG, 1YXA, 5NCS, 3OZQ, 3LE2, 3KCG, 4R9I, 5DU3, 3PZF, 2WXW) belonging to phylogenetically diverse taxa were selected from the serpin SSN and structurally aligned with MUSTANG (Konagurthu et al. 2006, 2010). A global HMM was built from the translation of this structural alignment into a multiple sequence alignment using the default parameters of HMMBuild from the HMMer package (Finn et al. 2011). All independent sequence clusters from the SSN (at a similarity threshold of 52.5% sequence identity) were aligned internally against the structure-guided global HMM and cluster-specific HMMs (single HMM per sequence cluster in the SSN) were built using HMMBuild. The most consensus-like sequence of each cluster was identified by scoring all sequences in that cluster against its cluster-specific HMM. The resulting 750 representative sequences (from 750 sequence clusters) were retrieved and aligned against the structure-guided HMM using HMMalign in HMMer. A guide-tree was inferred from this representative alignment in IQTREE (Nguyen, Schmidt, et al. 2015), using LG+R10 as the ML model (selected by ModelFinder) (Kalyaanamoorthy et al. 2017) and default parameters of the tree-search algorithm, for aligning all serpin sequences in the data set against the representative seed alignment in UPP from the SEPP package (Nguyen, Mirarab, et al. 2015). Columns that contained residues not-resolved in crystal structures (such as the RCL and nonconserved insertions) were not aligned as part of this workflow and hence deleted from the alignment. After manual editing and deletion of poorly aligned sequences, filtering of sequences that were greater than 25% shorter or longer than the median sequence length and manual refinement, the alignment consisted of 6,000 sequences.

Phylogenetic Inference

The serpin superfamily phylogeny was inferred in IQTREE v1.61 (Nguyen, Schmidt, et al. 2015). Phylogenetic inferences were formed with five independent replicates. The number of technical replicates we were capable of performing was limited by computational resources (each independent inference required more than 20,000 CPU hours and 150 GB of system memory). Five initial guide-trees were inferred on the final serpin alignment using default parameters (perturbation strength for nearest neighbor interchange of 0.5, maximum 1,000 iterations of tree-search, tree-search concluded after 100 unsuccessful iterations) with the sequence evolution model LG+R10. This model was chosen by corrected Akaike and Bayesian information criteria using ModelFinder implemented in IQ-TREE on representative guide-trees outlined above. Four additional trees were inferred using the models WAG+R10 and LG+R10 with default tree-search criteria to investigate. Five of the LG+R10 trees (irrespective of topology) was used as a guide-tree to approximate PMSF profiles using the LG+F+R10+C10 complex model (Wang et al. 2018) with extended tree-search parameters (perturbation strength for nearest neighbor interchange of 0.2, maximum 2,000 iterations of tree-search, tree-search concluded after 200 unsuccessful iterations). Branch supports were measured by ultrafast bootstrapping approximated to either 10,000 or 1,000 replicates (Hoang et al. 2018). Topologies that were poorly fitted to the alignment data were rejected using the AU-test (Shimodaira 2002), which was conducted to 10,000 replicates and the single best topology not rejected by the AU-test was selected according to a priori criteria on serpin and prokaryotic evolution (supplementary fig. 7, Supplementary Material online). All tree visualization and analysis were performed in R using the libraries ggtree (Yu et al. 2017), ape (Paradis et al. 2004), tidytree, and phylobase.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

Funding by the Australian Research Council Centre of Excellence for Innovations in Peptide and Protein Science and the Centre of Excellence in Synthetic Biology is gratefully acknowledged. This project was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

Data Availability

The data underlying this article are available in the article and its Supplementary Material online.

References

Ahmed FH, Hafna Ahmed F, Carr PD, Lee BM, Afriat-Jurnou L, Elaaf Mohamed A, Hong N-S, Flanagan J, Taylor MC, Greening C. 2015. Sequence–structure–function classification of a catalytically diverse oxidoreductase superfamily in mycobacteria. *J Mol Biol.* 427(22):3554–3571.

Akiva E, Brown S, Almonacid DE, Barber AE, Custer AF, Hicks MA, Huang CC, Lauck F, Mashiyama ST, Meng EC, et al. 2014. The structure–function linkage database. *Nucleic Acids Res.* 42(Database issue):D521–D530.

Akiva E, Copp JN, Tokuriki N, Babbitt PC. 2017. Evolutionary and molecular foundations of multiple contemporary functions of the nitroreductase superfamily. *Proc Natl Acad Sci U S A.* 114(45):E9549–E9558.

An C, Budd A, Kanost MR, Michel K. 2011. Characterization of a regulatory unit that controls melanization and affects longevity of mosquitoes. *Cell Mol Life Sci.* 68(11):1929–1939.

An C, Ragan EJ, Kanost MR. 2011. Serpin-1 splicing isoform J inhibits the proSpätzle-activating proteinase HP8 to regulate expression of antimicrobial hemolymph proteins in *Manduca sexta*. *Dev Comp Immunol.* 35(1):135–141.

An C, Zhang M, Chu Y, Zhao Z. 2013. Serine protease MP2 activates prophenoloxidase in the melanization immune response of *Drosophila melanogaster*. *PLoS One* 8(11):e79533.

Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. 2009. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One* 4(2):e4345.

Baier F, Tokuriki N. 2014. Connectivity between catalytic landscapes of the metallo-β-lactamase superfamily. *J Mol Biol.* 426(13):2442–2456.

Bennuru S, Semnani R, Meng Z, Ribeiro JMC, Veenstra TD, Nutman TB. 2009. Brugia malayi excreted/secreted proteins at the host/parasite interface: stage- and gender-specific proteomic profiling. *PLoS Negl Trop Dis.* 3(4):e410.

Cabrita LD, Irving JA, Pearce MC, Whisstock JC, Bottomley SP. 2007. Aeropin from the extremophile *Pyrobaculum aerophilum* bypasses the serpin misfolding trap. *J Biol Chem.* 282(37):26802–26809.

Chmelar J, Oliveira CJ, Rezacova P, Francischetti IMB, Kovarova Z, Pejler G, Kopacek P, Ribeiro JMC, Mares M, Kopecky J, et al. 2011. A tick salivary protein targets cathepsin G and chymase and inhibits host inflammation and platelet aggregation. *Blood* 117(2):736–744.

Cohen M, Davydov O, Fluhr R. 2019. Plant serpin protease inhibitors: specificity and duality of function. *J Exp Bot.* 70(7):2077–2085.

Colinet D, Dubuffet A, Cazes D, Moreau S, Drezen J-M, Poirié M. 2009. A serpin from the parasitoid wasp *Leptopilina boulardi* targets the Drosophila phenoloxidase cascade. *Dev Comp Immunol.* 33(5):681–689.

Copp JN, Akiva E, Babbitt PC, Tokuriki N. 2018. Revealing unexplored sequence-function space using sequence similarity networks. *Biochemistry.* 57(31):4651–4662.

Declerck PJ, Alessi MC, Verstreken M, Kruithof EK, Juhan-Vague I, Collen D. 1988. Measurement of plasminogen activator inhibitor 1 in biologic fluids with a murine monoclonal antibody-based enzyme-linked immunosorbent assay. *Blood* 71(1):220–225.

Evans DE, Hejgaard J. 1999. The impact of malt derived proteins on beer foam quality. Part I. The effect of germination and kilning on the level of protein Z4, protein Z7 and LTP1. *J Inst Brew.* 105(3):159–170.

Fetterer RH, Miska KB, Jenkins MC, Barfield RC, Lillehoj H. 2008. Identification and characterization of a serpin from *Eimeria acervulina*. *J Parasitol.* 94(6):1269–1274.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39(Web Server issue):W29–W37.

Fulton KF, Buckle AM, Cabrita LD, Irving JA, Butcher RE, Smith I, Reeve S, Lesk AM, Bottomley SP, Rossjohn J, et al. 2005. The high resolution crystal structure of a native thermostable serpin reveals the complex mechanism underpinning the stressed to relaxed transition. *J Biol Chem.* 280(9):8435–8442.

Gerlt JA, Bouvier JT, Davidson DB, Imker HJ, Sadkhin B, Slater DR, Whalen KL. 2015. Enzyme function initiative-enzyme similarity tool (EFI-EST): a web tool for generating protein sequence similarity networks. *Biochim Biophys Acta.* 1854(8):1019–1037.

Gettins PGW. 2002. Serpin structure, mechanism, and function. *Chem Rev.* 102(12):4751–4804.

Gettins PGW, Olson ST. 2009. Exosite determinants of serpin specificity. *J Biol Chem.* 284(31):20441–20445.

Ghendler Y, Arnon R, Fishelson Z. 1994. *Schistosoma mansoni*: isolation and characterization of Smpi56, a novel serine protease inhibitor. *Exp Parasitol.* 78(2):121–131.

Goulas T, Ksiazek M, Garcia-Ferrer I, Sochaj-Gregorczyk AM, Waligorska I, Wasylewski M, Potempa J, Gomis-Rüth FX. 2017. A structure-derived snap-trap mechanism of a multispecific serpin from the dysbiotic human oral microbiome. *J Biol Chem.* 292(26):10883–10898.

Gremski LH, da Silveira RB, Chaim OM, Probst CM, Ferrer VP, Nowatzki J, Weinschutz HC, Madeira HM, Gremski W, Nader HB, et al. 2010. A novel expression profile of the *Loxosceles intermedia* spider venomous gland revealed by transcriptome analysis. *Mol Biosyst.* 6(12):2403–2416.

Guo P-C, Dong Z, Zhao P, Zhang Y, He H, Tan X, Zhang W, Xia Q. 2015. Structural insights into the unique inhibitory mechanism of the silkworm protease inhibitor serpin18. *Sci Rep.* 5:11863.

Hashimoto C, Kim DR, Weiss LA, Miller JW, Morisato D. 2003. Spatial regulation of developmental signaling by a serpin. *Dev Cell.* 5(6):945–950.

Heit C, Jackson BC, McAndrews M, Wright MW, Thompson DC, Silverman GA, Nebert DW, Vasiliou V. 2013. Update of the human and mouse SERPIN gene superfamily. *Hum Genomics.* 7:22.

Hejgaard J, Rasmussen SK, Brandt A, Svendsen I. 1985. Sequence homology between barley endosperm protein Z and protease inhibitors of the α1-antitrypsin family. *FEBS Lett* 180(1):89–94.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol.* 35(2):518–522.

Hopkins PC, Carrell RW, Stone SR. 1993. Effects of mutations in the hinge region of serpins. *Biochemistry* 32(30):7650–7657.

Huntington JA. 2011. Serpin structure, function and dysfunction. *J Thromb Haemost.* 9:26–34.

Ing NH, Roberts RM. 1989. The major progesterone-modulated proteins secreted into the sheep uterus are members of the serpin superfamily of serine protease inhibitors. *J Biol Chem.* 264(6):3372–3379.

Irving JA, Pike RN, Lesk AM, Whisstock JC. 2000. Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function. *Genome Res.* 10(12):1845–1864.

Irving JA, Steenbakkers PJM, Lesk AM, den Camp HJMO, Pike RN, Whisstock JC. 2002. Serpins in prokaryotes. *Mol Biol Evol*. 19(11):1881–1890.

Ivanov D, Emonet C, Foata F, Affolter M, Delley M, Fisseha M, Blum-Sperisen S, Kochhar S, Arigoni F. 2006. A serpin from the gut bacterium *Bifidobacterium longum* inhibits eukaryotic elastase-like serine proteases. *J Biol Chem*. 281(25):17246–17252.

Jiang H, Wang Y, Yu X-Q, Zhu Y, Kanost M. 2003. Prophenoloxidase-activating proteinase-3 (PAP-3) from *Manduca sexta* hemolymph: a clip-domain serine proteinase regulated by serpin-1J and serine proteinase homologs. *Insect Biochem Mol Biol*. 33(10):1049–1060.

Johnson DJD, Li W, Adams TE, Huntington JA. 2006. Antithrombin–S195A factor Xa-heparin structure reveals the allosteric mechanism of antithrombin activation. *EMBO J*. 25(9):2029–2037.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 14(6):587–589.

Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttenhower C. 2015. High-specificity targeted functional profiling in microbial communities with ShortBRED. *PLoS Comput Biol*. 11(12):e1004557.

Kantyka T, Plaza K, Koziel J, Florczyk D, Stennicke HR, Thogersen IB, Enghild JJ, Silverman GA, Pak SC, Potempa J. 2011. Inhibition of *Staphylococcus aureus* cysteine proteases by human serpin potentially limits staphylococcal virulence. *Biol Chem*. 392(5):483–489.

Kantyka T, Rawlings ND, Potempa J. 2010. Prokaryote-derived protein inhibitors of peptidases: a sketchy occurrence and mostly unknown function. *Biochimie* 92(11):1644–1656.

Kazemi SM, Sabatier J-M. 2019. Venoms of Iranian scorpions (Arachnida, Scorpiones) and their potential for drug discovery. *Molecules* 24(14):2670.

Koh E, Carmieli R, Mor A, Fluhr R. 2016. Singlet oxygen-induced membrane disruption and serpin-protease balance in vacuolar-driven cell death. *Plant Physiol*. 171(3):1616–1625.

Konagurthu AS, Reboul CF, Schmidberger JW, Irving JA, Lesk AM, Stuckey PJ, Whisstock JC, Buckle AM. 2010. MUSTANG-MR structural sieving server: applications in protein structural analysis and crystallography. *PLoS One* 5(4):e10048.

Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. 2006. MUSTANG: a multiple structural alignment algorithm. *Proteins* 64(3):559–574.

Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*. 55:709–742.

Krem MM, Di Cera E. 2003. Conserved Ser residues, the shutter region, and speciation in serpin evolution. *J Biol Chem*. 278(39):37810–37814.

Kumar A, Ragg H. 2008. Ancestry and evolution of a secretory pathway serpin. *BMC Evol Biol*. 8:250.

Lampl N, Budai-Hadrian O, Davydov O, Joss TV, Harrop SJ, Curmi PMG, Roberts TH, Fluhr R. 2010. Arabidopsis AtSerpin1, crystal structure and in vivo interaction with its target protease responsive to desiccation-21 (RD21). *J Biol Chem*. 285(18):13550–13560.

Law RHP, Zhang Q, McGowan S, Buckle AM, Silverman GA, Wong W, Rosado CJ, Langendorf CG, Pike RN, Bird PI, et al. 2006. An overview of the serpin superfamily. *Genome Biol*. 7(5):216–211.

Lema Asqui S, Vercammen D, Serrano I, Valls M, Rivas S, Van Breusegem F, Conlon FL, Dangl JL, Coll NS. 2018. AtSERPIN1 is an inhibitor of the metacaspase AtMC1-mediated cell death and autocatalytic processing in planta. *New Phytol*. 218(3):1156–1166.

Lopez Quezada LA, Sajid M, Lim KC, McKerrow JH. 2012. A blood fluke serine protease inhibitor regulates an endogenous larval elastase. *J Biol Chem*. 287(10):7074–7083.

Marijanovic EM, Fodor J, Riley BT, Porebski BT, Costa MGS, Kass I, Hoke DE, McGowan S, Buckle AM. 2019. Reactive centre loop dynamics and serpin specificity. *Sci Rep*. 9(1):3870.

Mkaouar H, Akermi N, Mariaule V, Boudebbouze S, Gaci N, Szukala F, Pons N, Marquez J, Gargouri A, Maguin E, et al. 2016. Siropins, novel serine protease inhibitors from gut microbiota acting on human proteases involved in inflammatory bowel diseases. *Microb Cell Fact*. 15(1):201.

Molehin AJ, Gobert GN, Driguez P, McManus DP. 2014. Functional characterization of SjB10, an intracellular serpin from *Schistosoma japonicum*. *Parasitology* 141(13):1746–1760.

Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, Bader GD, Ferrin TE. 2011. clusterMaker: a multi-algorithm clustering plugin for cytoscape. *BMC Bioinformatics* 12:436.

Mulenga A, Kim T, Ibelli AMG. 2013. Amblyomma americanumtick saliva serine protease inhibitor 6 is a cross-class inhibitor of serine proteases and papain-like cysteine proteases that delays plasma clotting and inhibits platelet aggregation. *Insect Mol Biol*. 22(3):306–319.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.

Nguyen N-PD, Mirarab S, Kumar K, Warnow T. 2015. Ultra-large alignments using phylogeny-aware profiles. *Genome Biol*. 16:124.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784):299–304.

Pak SC, Kumar V, Tsu C, Luke CJ, Askew YS, Askew DJ, Mills DR, Brömme D, Silverman GA. 2004. SRP-2 is a cross-class inhibitor that participates in postembryonic development of the nematode *Caenorhabditis elegans*: initial characterization of the clade L serpins. *J Biol Chem*. 279(15):15448–15459.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2):289–290.

Park SH, Jiang R, Piao S, Zhang B, Kim E-H, Kwon H-M, Jin XL, Lee BL, Ha N-C. 2011. Structural and functional characterization of a highly specific serpin in the insect innate immunity. *J Biol Chem*. 286(2):1567–1575.

Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*. 8(10):785–786.

Prevot P-P, Beschin A, Lins L, Beaufays J, Grosjean A, Bruys L, Adam B, Brossard M, Brasseur R, Zouaoui Boudjeltia K, et al. 2009. Exosites mediate the anti-inflammatory effects of a multifunctional serpin from the saliva of the tick *Ixodes ricinus*. *FEBS J*. 276(12):3235–3246.

Richer MJ, Keays CA, Waterhouse J, Minhas J, Hashimoto C, Jean F. 2004. The Spn4 gene of Drosophila encodes a potent furin-directed secretory pathway serpin. *Proc Natl Acad Sci U S A*. 101(29):10560–10565.

Roberts TH, Hejgaard J, Saunders NFW, Cavicchioli R, Curmi PMG. 2004. Serpins in unicellular eukarya, archaea, and bacteria: sequence analysis and evolution. *J Mol Evol*. 59(4):437–447.

Schick C, Pemberton PA, Shi G-P, Kamachi Y, Çataltepe S, Bartuski AJ, Gornstein ER, Brömme D, Chapman HA, Silverman GA. 1998. Cross-class inhibition of the cysteine proteinases cathepsins K, L, and S by the serpin squamous cell carcinoma antigen 1: a kinetic analysis. *Biochemistry* 37(15):5258–5266.

Shakhnovich BE, Deeds E, Delisi C, Shakhnovich E. 2005. Protein structure and evolutionary history determine sequence space topology. *Genome Res*. 15(3):385–392.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*. 51(3):492–508.

Sochaj-Gregorczyk A, Ksiazek M, Waligorska I, Straczek A, Benedyk M, Mizgalska D, Thøgersen IB, Enghild JJ, Potempa J. 2020. Plasmin inhibition by bacterial serpin: implications in gum disease. *FASEB J*. 34(1):619–630.

Tanaka S-I, Koga Y, Takano K, Kanaya S. 2011. Inhibition of chymotrypsin- and subtilisin-like serine proteases with Tk-serpin from hyperthermophilic archaeon *Thermococcus kodakaraensis*. *Biochim Biophys Acta*. 1814(2):299–307.

Toubarro D, Avila MM, Hao Y, Balasubramanian N, Jing Y, Montiel R, Faria TQ, Brito RM, Simões N. 2013. A serpin released by an entomopathogen impairs clot formation in insect defense system. *PLoS One* 8(7):e69161.

Turroni F, Foroni E, O'Connell Motherway M, Bottacini F, Giubellini V, Zomer A, Ferrarini A, Delledonne M, Zhang Z, van Sinderen D, et al. 2010. Characterization of the serpin-encoding gene of Bifidobacterium breve 210B. *Appl Environ Microbiol*. 76(10):3206–3219.

Wang H-C, Minh BQ, Susko E, Roger AJ. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst Biol*. 67(2):216–235.

Wichelecki DJ, Vetting MW, Chou L, Al-Obaidi N, Bouvier JT, Almo SC, Gerlt JA. 2015. ATP-binding cassette (ABC) transport system solute-binding protein-guided identification of novel D-altritol and galactitol catabolic pathways in *Agrobacterium tumefaciens* C58. *J Biol Chem*. 290(48):28963–28976.

Widmer C, Gebauer JM, Brunstein E, Rosenbaum S, Zaucke F, Drögemüller C, Leeb T, Baumann U. 2012. Molecular basis for the action of the collagen-specific chaperone Hsp47/SERPINH1 and its structure-specific client recognition. *Proc Natl Acad Sci U S A*. 109(33):13243–13247.

Xu Z, Lin Z, Wei N, Di Q, Cao J, Zhou Y, Gong H, Zhang H, Zhou J. 2019. Immunomodulatory effects of *Rhipicephalus haemaphysaloides* serpin RHS2 on host immune responses. *Parasit Vectors*. 12(1):341.

Yang Y, Hu D, Wang L, Liang C, Hu X, Xu J, Huang Y, Yu X. 2014. Comparison of two serpins of *Clonorchis sinensis* by bioinformatics, expression, and localization in metacercaria. *Pathog Glob Health*. 108(4):179–185.

Yoo BC, Aoki K, Xiang Y, Campbell LR, Hull RJ, Xoconostle-Cázares B, Monzer J, Lee JY, Ullman DE, Lucas WJ. 2000. Characterization of cucurbita maxima phloem serpin-1 (CmPS-1). A developmentally regulated elastase inhibitor. *J Biol Chem*. 275(45):35122–35128.

Yu G, Smith DK, Zhu H, Guan Y, Lam TT. 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol*. 8(1):28–36.

Zallot R, Oberg NO, Gerlt JA. 2018. "Democratized" genomic enzymology web tools for functional assignment. *Curr Opin Chem Biol*. 47:77–85.

Zhang H, Fei R, Xue B, Yu S, Zhang Z, Zhong S, Gao Y, Zhou X. 2017. Pnserpin: a novel serine protease inhibitor from extremophile *Pyrobaculum neutrophilum*. *Int J Mol Sci*. 18(1):113.

Zhang Q, Buckle AM, Law RHP, Pearce MC, Cabrita LD, Lloyd GJ, Irving JA, Ian Smith A, Ruzyla K, Rossjohn J, et al. 2007. The N terminus of the serpin, tengpin, functions to trap the metastable native state. *EMBO Rep*. 8(7):658–663.

Zhou A, Huntington JA, Pannu NS, Carrell RW, Read RJ. 2003. How vitronectin binds PAI-1 to modulate fibrinolysis and cell migration. *Nat Struct Biol*. 10(7):541–544.

Zhou A, Wei Z, Read RJ, Carrell RW. 2006. Structural mechanism for the carriage and release of thyroxine in the blood. *Proc Natl Acad Sci U S A*. 103(36):13321–13326.

Zhou A, Wei Z, Stanley PLD, Read RJ, Stein PE, Carrell RW. 2008. The S-to-R transition of corticosteroid-binding globulin and the mechanism of hormone release. *J Mol Biol*. 380(1):244–251.

Zhou Q, Snipas S, Orth K, Muzio M, Dixit VM, Salvesen GS. 1997. Target protease specificity of the viral serpin CrmA. Analysis of five caspases. *J Biol Chem*. 272(12):7797–7800.