A peer-reviewed version of this preprint was published in PeerJ on 16 June 2015.

Kumar A. 2015. Bayesian phylogeny analysis of vertebrate serpins illustrates evolutionary conservation of the intron and indels based six groups classification system from lampreys for ~500 MY. PeerJ 3:e1026 https://doi.org/10.7717/peerj.1026

Bayesian phylogeny analysis of vertebrate serpins illustrates evolutionary conservation of the intron and indels based six groups classification system from lampreys for ~500 MY

Abhishek Kumar

The serpin superfamily is characterized by proteins that fold into a conserved tertiary structure and exploits a sophisticated and irreversible suicide-mechanism of inhibition. Vertebrate serpins can be conveniently classified into six groups (V1-V6), based on three independent biological features - genomic organization, diagnostic amino acid sites and rare indels. However, this classification system was based on the limited number of mammalian genomes available. In this study, several non-mammalian genomes are used to validate this classification system, using the powerful Bayesian phylogenetic method. This method supports the intron and indel based vertebrate classification and proves that serpins have been maintained from lampreys to humans for about 500 MY. Lampreys have less than 10 serpins, which expanded into 36 serpins in humans. The two expanding groups V1 and V2 have SERPINB1/SERPINB6 and SERPINA8/SERPIND1 as the ancestral serpins, respectively. Large clusters of serpins are formed by local duplications of these serpins in tetrapod genomes. Interestingly, the ancestral HCII/SERPIND1 locus (nested within PIK4CA) possesses group V4 serpin (A2APL1, homolog of α_2-AP/SERPINF2) of lampreys; hence, pointing to the fact that group V4 might have originated from group V2. Additionally in this study, the phylogenetic history and genomic characteristics of vertebrate serpins were revisited.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1126v1 | CC-BY 4.0 Open Access | rec: 26 May 2015, publ: 26 May 2015

Bayesian phylogeny analysis of vertebrate serpins illustrates evolutionary conservation of the intron and indels based six groups classification system from lampreys for ~500 MY

Abhishek Kumar[1, 2,*]

[1]Department of Genetics & Molecular Biology in Botany, Institute of Botany, Christian-Albrechts-University at Kiel, Kiel, Germany.

[2]Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

*Email:abhishek.abhishekkumar@gmail.com

Abstract

The serpin superfamily is characterized by proteins that fold into a conserved tertiary structure and exploits a sophisticated and irreversible suicide-mechanism of inhibition. Vertebrate serpins can be conveniently classified into six groups (V1-V6), based on three independent biological features - genomic organization, diagnostic amino acid sites and rare indels. However, this classification system was based on the limited number of mammalian genomes available. In this study, several non-mammalian genomes are used to validate this classification system, using the powerful Bayesian phylogenetic method. This method supports the intron and indel based vertebrate classification and proves that serpins have been maintained from lampreys to humans for about 500 MY. Lampreys have less than 10 serpins, which expanded into 36 serpins in humans. The two expanding groups V1 and V2 have SERPINB1/SERPINB6 and SERPINA8/SERPIND1 as the ancestral serpins, respectively. Large clusters of serpins are formed by local duplications of these serpins in tetrapod genomes. Interestingly, the ancestral HCII/SERPIND1 locus (nested within PIK4CA) possesses group V4 serpin (A2APL1, homolog of α_2-AP/SERPINF2) of lampreys; hence, pointing to the fact that group V4 might have originated from group V2. Additionally in this study, the phylogenetic history and genomic characteristics of vertebrate serpins were revisited.

Keywords: Serpins, vertebrates, Bayesian phylogeny, gene structure, intron-exon, gene duplication

1. Introduction

Serine proteinase inhibitors (serpins) are one of the major regulators of cellular proteolysis. The superfamily of serpins is involved in an array of fundamental biological processes such as blood coagulation, cell differentiation, cell migration, complement activation, embryo implantation, fibrinolysis, angiogenesis, and inflammation, and tumor suppression (Silverman et al. 2001). Serpins usually have a single domain (Pfam ID PF00079 or Interpro ID IPR023795) with a conserved core of ~350-400 residues. They often possess N- or C-terminal extensions, and an overall molecular mass of ~40-60 kDa. N- and/or O-glycosylations are frequently observed in extracellular serpins (Gettins 2002; Gettins et al. 1996). The conserved three-dimensional structure of serpins is composed of three β-sheets (βA-βC) and 8-9 α-helices (αA-αI). The hallmark of the serpin inhibitory mechanism is a large-scale conformational change involving the reactive center loop (RCL). The RCL is an exposed flexible loop of about 17-20 residues, which interacts with a target protease (Silverman et al. 2001). RCL acts as a bait imitating the protease substrate, and is cleaved between the positions P1 and P1' (Silverman et al. 2001).

In metazoans, serpins have undergone divergent evolution over a period of about 650-700 million years (Kumar & Ragg 2008). A number of phylogenetic studies have been undertaken using sequence analysis of the serpins. Early investigations suggested the establishment of this multigene family through inter- and intra-chromosomal gene duplications. Several gene clusters have arisen, encoding functionally diverse serpin proteins. In metazoans, serpins display highly variant exon-intron patterns are strongly conserved within some taxa. Gene architecture and other rare genetic characters singularize a robust basis for classifying vertebrate serpins. Based on number, positions, and phases of introns, serpins have been classified into six groups (V1-V6). Vertebrate serpin genes with equivalent gene structures often tend to be organized in clusters

(Benarafa & Remold-O'Donnell 2005). However, close physical linkage cannot always be established. Interestingly, none of the 24-intron positions that have been mapped to the core domain of vertebrate serpins is shared by all of the six gene groups. Nevertheless, characteristic amino acid indels provide further cues to unravel the phylogenetic relationship (Ragg et al. 2001). Previous analyses were performed using limited vertebrate serpins, and mainly focused on human and mouse data. Currently, several non-mammalian vertebrate genomes are known. Hence, combining these genomes for validation of intron-encoded vertebrate serpin classification is possible. In addition, different phylogenetic methods have been applied to vertebrate serpin classification such as maximum-likelihood (ML) and Neighbor-joining (NJ) (Atchley et al. 2001). In the last decade, Bayesian Markov chain Monte Carlo (MCMC) has been enthusiastically corroborated as the state-of-the-art method for phylogenetic reconstruction and was largely driven by the rapid and widespread adoption of MrBayes suite (Ronquist & Huelsenbeck 2003). This method was recently tested for urochordate serpin classification (Kumar & Bhandari 2014). Heretofore, there is no report on the use of this phylogenetic method for large-scale analysis of vertebrate serpins. Herein, Bayesian method was employed along with several non-mammalian genomes for reconstructing vertebrate serpin classification system. This study reveals that Bayesian phylogenic method supports the intron-coded vertebrate serpin classification system.

Moreover, this classification system is conserved from lampreys to human with a few new introns being created in groups V2, V4 and V6 in the serpin core domains of selected ray-finned fishes. Furthermore, different properties of these six serpin groups were also summarized.

2. Materials and methods

2.1. Collection of serpins from selected genomes

Serpin sequences of selected vertebrates **(Table 1)** were obtained from the Ensembl database (release 72, June 2013) using the BLAST suite (Altschul et al. 1997). Details of identified serpins and comprehensive alignments are provided (Kumar 2010).

2.2. Gene structure prediction of serpins

To ensure accuracy, gene structure prediction within the Ensembl (Flicek et al. 2013) was taken and combined with predictions of AUGUSTUS gene prediction tool (Stanke & Morgenstern 2005). Mature human α_1-antitrypsin was used as the standard sequence for intron position mapping and numbering of intron positions, followed by suffixes a–c for their locations as reported previously (Kumar & Ragg 2008).

2.3. Protein sequence alignment

Protein alignments of different vertebrate serpins were created by MUSCLE tool (Edgar 2004) using default parameters.

2.4. Selection of substitution model

Upon evaluation of different amino acid substitution models for this alignment dataset using MEGA5 software suite (Tamura et al. 2011), it turned out that the WAG+G model was the best fit (**Table S1**).

2.5. Bayesian phylogenetic analysis

To infer the evolutionary history of serpins, the Bayesian phylogenetic tree was constructed using the MrBayes 3.2 suite (Ronquist & Huelsenbeck 2003) with the following parameters: 5 generations, until average standard deviation of split frequencies was lower than 0.0098, 25% burn-in-period, WAG+G matrix-based model. Nve_Spn1 from starlet sea anemone (*Nematostella vectensis*) was used as the outgroup.

2.6. Estimation of genome size

Selected vertebrate genome sizes were calculated using the animal genome size database (Gregory 2014).

3. Results and discussion

3.1. Bayesian phylogeny classifies serpins into six groups V1-V6

Bayesian phylogenetic analysis reveals six groups of vertebrate serpins as depicted in different colors (**Figure 1**). Posterior probability values are marked, with the lowest being 47, because several paralogs of group V2 serpins are known in tetrapod genomes (**Figure 1**). Sea anemone serpin (Nve_Spn1) is the out-group for this phylogenetic analysis (marked by a brown arrow). This clustering matches with intron-indel based vertebrate serpins classification system of six groups (V1-V6) as illustrated in **Figure 2**. Lampreys have only eight serpins as evident from BLAST analysis against sea lamprey (*Petromyzon marinus*) genome and cDNA of European river lamprey, *Lampetra fluviatilis* in the Genbank (**Table 2**). These serpins of lampreys are only distributed into four groups (marked by green star in **Figure 2**). We will describe and discuss each of these groups in next sections.

3.2. Group V1 has several clade B members expanded from fishes to human

Group V1 serpins has been defined by a gene structure depicting five introns at positions –78c, 128c, 167a, 212c, and 262c, in their coding region (**Figure 2**). An additional intron at the position 85c is found in some group V1 members and the presence and absence of this intron, defines sub-groups V1a and V1b, respectively. Group V1 is multi-membered, consisting of ovalbumin-like serpins that are involved in different physiological roles and often called as ov-serpins (Benarafa & Remold-O'Donnell 2005). These serpins belong to clade B under clade-based classification system of serpins (Silverman et al. 2001). They are usually inhibitors of serine or cysteine proteases (cross-class inhibition), but some of them are non-inhibitory

members (e.g., maspin/SERPINB5). They are mostly intracellular since they lack N-terminal signal peptide with few exceptions (Benarafa & Remold-O'Donnell 2005; Izuhara et al. 2008; Kaiserman & Bird 2005). They are also deprived of the C-terminal extensions (Benarafa & Remold-O'Donnell 2005; Izuhara et al. 2008; Kaiserman & Bird 2005). These serpins are localized in two clusters in the human genome. Human chromosome 6p25 region harbors three genes - SERPINB1, SERPINB6, and SERPINB9 as the first cluster while, the remaining genes namely, SERPINB2, SERPINB3, SERPINB4, SERPINB5, SERPINB7, SERPINB8, SERPINB10, SERPINB11, SERPINB12, and SERPINB13 are located in the 18q21 region (Benarafa & Remold-O'Donnell 2005; Izuhara et al. 2008; Kaiserman & Bird 2005).

This cluster originated by duplication of SERPINB1/MNEI1-like gene (Benarafa & Remold-O'Donnell 2005). In contrast, the chicken has only one cluster on the chromosome 2q. Therefore, it is corroborated that there was a split after mammal/bird divergence at about 310 MY (Benarafa & Remold-O'Donnell 2005; Izuhara et al. 2008; Kaiserman & Bird 2005). Similar genomic organization of group V1 was also found in frogs and fishes. In addition to this syntenic organization, fishes possess some paralogous clusters of group V1 serpins. While, an additional cluster containing two serpins adjacent to the conserved orthologous cluster is found in frogs. The serpins SERPINB1/SERPINB6 of group V1 are probably the ancestor of all group V1 serpins, as these genes are found in lampreys (**Table 2**) and other fishes and are also conserved across other vertebrate taxa (Kumar 2010). The group V1 serpins may be classified into sub-groups V1a and V1b, since these differ by one intron. It has been argued that a serpin gene of group V1b (7 exons) is the ancestor of group V1a (8 exons) that has emerged in birds after the divergence of frogs (Benarafa & Remold-O'Donnell 2005; Izuhara et al. 2008; Kaiserman & Bird 2005). The first argument coincides with the current data and corroborates that groups V1a

serpins are derived from 7-exon genes (such as MNEI/SPB6). However, the argument that 8-exon genes first arose in chickens does not hold in agreement with the current data, since Xtr-Spn-5 in *X. tropicalis* and pSPB6 in *T. nigroviridis,* are group V1a members have the 8 exons/7 introns architecture (Kumar 2010). Therefore, it is proposed that group V1b is ancestral to all group V1 serpins and group V1a is suggested to have arisen independently several times in different vertebrates from fishes to mammals. The ancestor of group V1 serpins appears to have been generated during the emergence of vertebrates. The oldest group V1 serpins are SPB1/SPB6 orthologs, which are present in lamprey (**Table 2**). A recent study had claimed an ancestor of serpinB6 to be present in urochordates (Benarafa & Remold-O'Donnell 2005). However, another study depicted six different groups of urochordate serpins, based on intron-encoded classification system, which markedly differs from vertebrates six groups (Kumar & Bhandari 2014).

3.3. Group V2 possesses α_1-antitrypsin-like serpins, angiotensinogen (clade A) and heparin cofactor II (clade D)

Group V2 serpins are characterized by three introns at homologous positions - 192a, 282b, and 331c (**Figure 2**) in their coding regions, and most of the members have an intron mapping to the untranslated regions. This group is multi-membered, composed of α_1-antitrypsin like serpins that are involved in different physiological roles, including inhibitors, like α_1-antitrypsin or antichymotrypsin as well as non-inhibitory members, like angiotensinogen [AGT/ SERPINA8 (Kumar et al. 2014d)].

Gene structures of heparin cofactor II (HCII/SERPIND1) are variable in fishes with a novel intron gain at the position 241c, but this gene is nested in the large intron of phosphatidylinositol

194 4-Kinase (PIK4CA) gene for ~500 MYA (Kumar et al. 2014a). Human HCII/SERPIND1 has
195 985 germline variants, identified from 1092 human genomes. This includes 37 statistically
196 deleterious missense variants (Kumar et al. 2014a).
197 Recently, it was reported that the gene structures of AGT from selected ray-finned fishes varied
198 in exons I and II, with insertions of two novel introns in the core domain for ray-finned fishes at
199 positions 77c and 233c, respectively (Kumar et al. 2014d). It was also reported that the AGT loci
200 is conserved from lampreys to human and was estimated to be older than 500 MY (Kumar et al.
201 2014d). Interestingly, the RCL of AGT protein is inhibitory in lampreys and evolved to become
202 non-inhibitory in human over a period of 500 MY (Kumar et al. 2014d). Kumar et al (2014) also
203 detected 690 AGT variants by analyzing 1092 human genomes with the top three variation
204 classes belonging to single nucleotide polymorphisms (SNPs, 89.7%), somatic SNVs (5.2%) and
205 deletion (2.9%) (Kumar et al. 2014d). Furthermore, 121 missense variants of AGT including 32
206 statistically deleterious variants were deciphered (Kumar et al. 2014d).
207 From fishes to humans, group V2 comprises of multiple paralogs of α_1-antitrypsin-like genes.
208 Genuine orthologs of angiotensinogen and HCII were identified from lampreys to humans, using
209 synteny and signature sequences. Concerning the other genes of group V2; since in most
210 tetrapod genomes, α_1-antitrypsin-like gene clusters are derived from recent duplication events,
211 which results in proteins with high sequence similarities, one-to-one orthology allocation proved
212 to be difficult. This poses notorious challenges in detection of orthologs within this cluster and
213 often leads into problems in generating phylogenetic trees (**Figure 1**).
214 In the cluster of α_1-antitrypsin-like genes, the protein Z-dependent protease inhibitor (ZPI/
215 SERPINA10) is localized at the end of the cluster with other conserved marker genes (Kumar
216 2010). This assisted in the detection of ZPI/SERPINA10 orthologs using synteny analysis.

Recently published report on the serpins of channel catfish, *Ictalurus punctatus* (Li et al. 2015) also supported this finding.

Additionally, two group V2 serpins are found only in ray-finned fishes. The first serpin was detected with a novel intron at position 94a and hence it is named as the Spn_94a gene fishes, which have conserved in the same genomic organization in these fishes. This corroborates that fish specific ortholog. Spn_94a shows sequence similarity with ZPI/SERPINA10 gene and hence it is paralog of ZPI/SERPINA10. Similarly, *Fugu* and *T. nigroviridis* possess the second fish-specific group V2 gene with an additional intron at the position 215c (Spn_215c), which indicates that they are orthologs (Kumar 2010). The origin of these genes, however, is unclear.

No orthologs of the hormone binding serpins (corticosteroid-binding globulin [CBG/SERPINA6] and thyroxine-binding globulin [TBG/SERPINA6]) were detected in non-mammalian vertebrates. In short, the conserved set of group V2 comprises only orthologs of AGT/SERPINA8 (Kumar et al. 2014d) and HCII/SERPIND1 (Kumar et al. 2014a). In contrast, some fish-specific group V2 genes and the α_1-antitrypsin-like genes are differentially expanded in vertebrates, particularly in mammalian lineages, such as rodents (Forsyth et al. 2003a) and cattle (Pelissier et al. 2008). The expansion of group V2 members should be further explored by analyzing marsupials and Platypus, which branched out early in mammalian evolution. The presence of group V2 members in the lamprey genome suggests that this group originated during emergence of vertebrates (**Table 2**). Further investigation of group V2 members in the hagfish genome and more lamprey genomes will shed more light on this issue.

3.4. Group V3 is composed of 5 members, which belongs to two clades E and I

Group V3 serpins have seven introns at the positions - 86a/88a/90a, 167a, 230a, 290b, 323a, 352a and 380a in their coding regions (**Figure 2**). The exact location of the first intron is uncertain in different group V3 serpins, due to alignment ambiguities. Group V3 has five inhibitory serpins, which are involved in different physiological processes namely SERPINE1/plasminogen activator inhibitor 1 (PAI1/SERPINE1), glia derived nexin (GDN/SERPINE2/), SERPINE3, neuroserpin/SERPINI1, and pancpin/SERPINI2. PAI1/SERPINE1 is conserved in vertebrates, depicting 38-80% sequence identity and 59-95 % sequence similarity at the amino acid level with human PAI1/SERPINE1. The inhibitory RCL region is conserved and consists of R-M at the P1-P1'. GDN/SERPINE2 is also highly conserved in vertebrates and it shows 51-84% sequence identity and 70-93% sequence similarity with human GDN/SERPINE2. The helix-D region is highly conserved among GDN/SERPINE2 orthologs of different vertebrates and an N-glycosylation site (positions 163-165) is also conserved. The inhibitory RCL region is also strongly conserved. SERPINE3 is maintained in vertebrates, show 27-64% sequence identity, and 37-74 % sequence similarity on the amino acid level with human serpinE3. The inhibitory RCL of SERPINE3 is conserved with a cluster of hydrophobic amino acids preceding the presumptive P1 position.

The neuroserpin/SERPINI1 is highly conserved in vertebrates, and the protein shows 47-81% sequence identity and 65-95% sequence similarity with the human ortholog. The inhibitory RCL region of neuroserpin/SERPINI1 always contains an arginine at the position P1. An N-glycosylation signal at residues 163-165, and a C-terminal extension that has been shown to direct neuroserpin to the regulated secretory pathway (Ishigami et al. 2007) are strongly conserved.

Discriminatory data at the genomic, gene and protein levels offered a comprehensive insight into the phylogenetic history of neuroserpin/SERPINI1 (Kumar & Ragg 2008). Synteny analysis proved to be very instrumental in this respect, demonstrating that rare genomic characters can provide very useful information for decoding of links between protein families with intricate evolutionary history. The strongly conserved syntenic association of PDCD10 and neuroserpin/SERPINI1 orthologs during diversification of deuterostomes is unraveled (Kumar & Ragg 2008). These head-to-head oriented genes (Neuroserpin/SERPINI1 and PDCD10) have common bi-directional and asymmetrical promoter region inserted within the ~0.9 kb intergenic region (Chen et al. 2007). Requirements of common regulatory units could have driven the preservation of this linkage. In the era of next-generation genome sequencing, The rapidly accumulating genome sequences will certainly continue to provide further discriminatory markers, such as codon usage dichotomy (Krem & Di Cera 2003), in order to facilitates robust classification of other metazoan serpins.

Some serpins have signals for sub-cellular localization at the C-terminal end and they have been involved in the secretory pathway. It is not just limited to vertebrate serpins and several examples exists in invertebrates such as in some urochordate serpins (Kumar & Bhandari 2014). Using these signals, ancestral orthologs of neuroserpin/SERPINI1 were deciphered (Kumar & Ragg 2008). A C-terminal KDEL-like motif deters the secretion of soluble endoplasmic reticulum (ER) –resident proteins (Lewis et al. 1990; Raykhel et al. 2007; Semenza et al. 1990). There are 24 possible variants of ER retention signals listed as a PROSITE motif - [KRHQSA]-[DENQ]-E-L in the PROSITE database (Sigrist et al. 2013). In addition, there are some ER retention signals that do not fit into the classical PROSITE motif (Raykhel et al. 2007). These ER

retention signals are present across eukaryotic genomes such as in the BEM46 protein of *Neurospora crassa* (Kumar et al. 2013b). In early diverging deuterostomia, the neuroserpin orthologs in *Strongylocentrotus* (Spu-spn-1), lancelet (Bfl-Spn-1) and sea anemone (Nve-Spn-1) have HEEL, KDEL, and SDEL at their C-terminal ends, respectively. These are variants of the PROSITE motif for ER retention/retrieval signal. In contrast, the C-terminal end of tetrapod neuroserpin is HDFEEL. In HeLa cells that express three different KDEL receptors with overlapping, but differential passenger specificities, the "FEEL" sub-sequence targets attached passenger proteins primarily to the Golgi, though one-fourth of cells depict ER localization (Raykhel et al. 2007). However, in transfected COS cells, intracellular neuroserpin localizes either to the ER or to Golgi (Ishigami et al. 2007). Conversely, in cells with a regulated secretory pathway, neuroserpin/SERPINI1 resides in large dense core vesicles, which is assisted by a C-terminal extension encompassing the last 13 amino acids (ETMNTSGHDFEEL) including the FEEL sequence (Ishigami et al. 2007). Collectively, these data suggest that in the neuroserpin orthologs from deep-branching metazoans, a two amino acid insertion 'FE' in combination with additional residues constitutes a modified sorting signal, which attributes a specialized subcellular localization. Surveillance of the secretory pathway routes by serpins is an ancient and conserved trait in eukaryotes as indicated by the putative neuroserpin ortholog present in the sea anemone genome. It will be interesting to investigate experimentally, whether the C-terminal extensions of neuroserpin orthologs from fishes are functional and mediate differential localization in a fashion similar to mammalian neuroserpin.

Due to variations in their RSL region, ER-localized serpins may work differently in the secretory pathway. Neuroserpin from vertebrates inhibits tissue-type plasminogen activator (tPA) in vitro, using the Arg residue at the P1 position in the RSL region (Osterwalder et al. 1998). The

cleavage site of Bfl-spn-1 is preceded by the dipeptide motif Lys-Arg (KR), a distinct feature for substrates and inhibitors of proprotein convertases (PCs). Similar features were reported for Bla-Spn-1 from *B. lanceolatum* as well (Bentele et al. 2006). Since the serpins Bfl-spn-1 (*B. floridae*), Spu-spn-1 (sea urchin) and Nve-Spn-1 (sea anemone) also possess the Lys-Arg (KR) dipeptide motif. Thus, similar physiological role of these serpins can be expected. Status of the neuroserpin ortholog in the arthropod lineage is recently becoming clear. Examples of classical ER targeting signal (HDEL) possessing serpins were found in *D. melanogaster* as Spn4, which is a furin inhibitor (Oley et al. 2004; Osterwalder et al. 2004; Richer et al. 2004) and its homologous gene in *Anopheles gambiae* as *Spn10* (Danielli et al. 2003). Recently, the crystal structure of fly Spn4 was determined and this serpin exhibits structural properties as of human neuroserpin/SERPINI1 (Ellisdon et al. 2014), which provided first evidences of the orthologous nature of these serpins.

The pancpin/SERPINI2 gene is localized in close proximity to the neuroserpin gene. Pancpin also possesses a C-terminal extension and indels like neuroserpin, suggesting its close relatedness to these proteins. Pancpin/SERPINI2 orthologs are found only in mammals and in *Xenopus*, showing 49-76% sequence identity, and 68-88% sequence similarity at the amino acid level. The C-terminal end is strongly conserved (Kumar 2010). Absence of pancpin/SERPINI2 in fishes hints that the pancpin gene may have originated by tandem duplication of neuroserpin after separation of tetrapods from the fish lineage.

In the human genome, the other group V3 members such as PAI1/SERPINE1 (chromosome 3), GDN/SERPINE2 (chromosome 2) and SERPINE3 (chromosome 13) are present at various genomic locations. This suggests that they originated at independent loci in the vertebrates.

3.5. Group V4 has three serpins - two in the clade F and one in the clade G; surprisingly, fishes have C1IN with two immunoglobulin domains

Group V4 of vertebrate serpins have a gene structure consisting a conserved set of five introns at positions 67a, 123a, 192a, 238c, and 307a in the coding regions (**Figure 2**). In mammals, group V4 serpins consists of three genes - pigment epithelium derived factor (PEDF/SERPINF1), α_2-antiplasmin (α_2-AP/SERPINF2) and C1 inhibitor (C1IN/SERPING1). Group V4 serpins are involved in very different physiological functions. PEDF is a non-inhibitory serpin that possesses neuroprotective and antiangiogenic functions (Sawant et al. 2004; Steele et al. 1993; Tombran-Tink 2005). α_2-antiplasmin is an inhibitor of plasmin and its fibrin bound form is a major regulator of blood clot lysis (Coughlin 2005). C1 inhibitor (C1IN/SERPING1) is a multi-functional serpin, which operates by inactivating various serine proteases in different plasmatic cascades including the complement (classical pathway - C1r and C1s; as well as lectin pathways - MASP1 and MASP2), contact (Factor XII and kallikrein), coagulation (Factor XI and thrombin) and fibrinolytic (tPA and plasmin) systems (Davis et al. 2007; Davis et al. 2010). Fishes have C1IN/SERPING1 with two immunoglobulin-like domains attached at the N-terminal region (**Figure 4**). The RCL regions of C1IN/SERPING1 have variations at the positions P2 and P1' from fishes and tetrapods Gene structures of C1IN/SERPING1 from selected ray-finned fishes varied in the Ig domain region with the insertion of a novel intron splitting exon Im2 into Im2a and Im2b (Kumar et al. 2014b). Kumar et al (2014) depicted that C1IN/SERPING1 gene has remained on the same locus for ~450 MY in 52 vertebrates, but it is missing in frogs and lampreys (Kumar et al. 2014b).

Protein sequence analyses depicts that orthologs of PEDF/SERPINF1 and α_2-AP (SERPINF2)-like genes are conserved throughout vertebrates (Kumar 2010). However, on close scrutiny of

group V4 serpins, orthologs of most human group V4 serpins other than A2AP1_FRU in *Fugu* cannot be found in current genomic sequence versions of fish genomes (Kumar 2010). Fishes have paralogs, probably due to fish-specific genome duplications and diversifications (Kumar 2010). In addition, the sea lamprey genome has two members of group V4 were detected resembling α_2-AP-like genes (A2APL1_PMA and A2APL2_PMA) with orthologs in the European sea lamprey (**Table 1**). This suggests that group V4 serpins existed since the beginning of vertebrates. Recently, it was shown that the A2APL1_PMA gene is present in the nested state in the largest intron of PIK4CA gene along with HCII/SERPIND1 gene in the reverse orientation (Kumar et al. 2014a). However, only HCII/SERPIND1 gene is found as nested gene in PIK4CA in the ray-finned fishes to humans (Kumar et al. 2014a) and hence, it is can postulated that ancestral group V4 gene were originated at adjacent to HCII/SERPIND1, which was mostly likely lost in the other lineages. This also corroborates that the origin of group V4 serpins is associated with group V2, as assumed from the conservation of basal intron at the position 192a (**Figure 2**).

3.6. Group V5 comprises only antithrombin III (ATIII) aka SERPINC1

Group V5 consists of a single member – antithrombin III (ATIII/SERPINC1). This gene encompasses seven exons and six introns with conserved intron positions (**Figure 2**). In the human genome, the ATIII/SERPINC1 gene is located on chromosome 1q23–q25. ATIII/SERPINC1 is the major thrombin inhibitor in the blood coagulation cascade (Jordan 1983), requires heparin for activation and has potent anti-angiogenic activity in certain conformations (Gettins et al. 1996).

The ATIII/SERPINC1 protein is highly conserved, and the sequence identity and similarity from fishes to mammals falls within the range of 50-87% and 67-97% respectively. ATIII/SERPINC1 has been maintained for over 450 MY on the same genomic loci in vertebrates with a few changes in ray-finned fishes. ATIII/SERPINC1 gene has lost an intron (262c) in tetrapods and in the lobed-finned fish coelacanth, *Latimeria chalumnae.* In addition, it has gained an intron at the position 262c in the ray-finned fishes, a characteristic feature, which is shared by group V1 members as well (Kumar et al. 2013a). ATIII/SERPINC1 comprises of several proteins motifs, heparin binding basic residues, the hD helix, 3 pairs of Cys-Cys salt bridges, N-glycosylation sites, serpin motifs and inhibitory RCL (Kumar et al. 2013a). 1997 ATIII/SERPINC1 variants have been identified from 1092 human genomes. These variants have been categorized into 76.2% SNPs, 11.8% deletions and 8.1% insertions (Kumar et al. 2013a).

3.7. Group V6 is composed of HSP47 (SERPINH1) ortholog and fishes have 1-3 paralogs of HSP47

Group V6 is characterized by a gene structure depicting three introns at positions 192a, 225a and 300c in their coding regions (**Figure 2**). This gene encodes for heat shock protein 47 kDa (HSP47/SERPINH1), which possesses a C-terminal endoplasmic reticulum (ER) retention signal (Pelham 1990). HSP47/SERPINH1 is a non-inhibitory serpin that is found in the ER of collagen producing cells where it is involved in the correct folding of procollagen triplet helices. Furthermore, it assists in the transport of procollagen from the ER to the Golgi complex (Hendershot & Bulleid 2000; Lamande & Bateman 1999; Nagata 1996; Sauk et al. 2005). Tetrapods have a single copy of HSP47 genes, while fishes have up to three copies. The first HSP47/SERPINH1 is common to all vertebrates (HSP47_1). The second copy is conserved in

few ray-finned fishes such as *G. aculeatus* and *D. rerio* with conserved syntenic organization (**Figure 4**). The third one is only present in some fishes such as *D. rerio* (**Table 3**). HSP47/SERPINH1-like gene is conserved from lampreys to mammals, and this gene show 22-96% sequence identity and 37-98% sequence similarity with human HSP47/SERPINH1, respectively (**Table 3**). The HSP47_TNI protein is highly diverged from standard HSP47/SERPINH1 protein as well as from all other serpin sequences (**Table 3**).

Orthology of the group V6 gene, lamprey HSP47_PMA (grey) cannot be decided on this basis. Group V6 comprises of the HSP47 gene and its paralogs in different vertebrates. Tetrapods have a single copy of the HSP47/SERPINH1 gene, while there are two or three HSP47-like genes in some fishes (**Figure 1)**.

3.8. Status of thrombin inhibitors

Four human serpins inhibit thrombin, namely ATIII/SERPINC1, HCII/SERPIND1, protein C inhibitor (PCI/SERPINA5) and nexin I//SERPINE2 (Huntington 2013; Huntington 2014). These serpins exhibit higher rates of thrombin inhibition after binding to the glycosaminoglycan (GAG); however they have evolved radically different inhibition mechanisms (Huntington 2013). Apart from these four, recent studies revealed that a fifth serpin, namely AGT/SERPINA8 gene also act as thrombin inhibitor, – at least in lampreys (Kumar et al. 2014o; Wang & Ragg 2011; Wong & Takei 2011). Lamprey AGT/SERPINA8 gene possesses inhibitory RCL and regulates thrombin along with HCII (Kumar et al. 2014d). Lampreys have no ATIII/SERPINC1 gene (Kumar et al. 2013a), but after the emergence of ATIII/SERPINC1 gene in tetrapods, AGT/SERPINA8 gene became non-inhibitory serpin (Kumar et al. 2014d). It is surprising that lampreys have no the major thrombin inhibitor, ATIII/SERPINC1. However, lampreys also lack

other immunologically critical genes such as recombination-activation genes which are essential for V(D)J recombination process that yields and assembles the variable regions of immunoglobulin and T-cell receptor genes in developing B- and T-lymphocytes (Kumar et al. 2015). This suggests that lampreys may not need some of the very essential vertebrate genes and ATIII/SERPINC1 is in this category. This can be explained since AGT/SERPINA8 is the bi-functional serpin in lampreys, which acts as a thrombin inhibitor as well as a blood pressure regulator (Kumar et al. 2014o; Wang & Ragg 2011; Wong & Takei 2011).

Major thrombin regulating thrombin serpins are ATIII/SERPINC1 (Kumar et al. 2013a) and HCII/SERPIND1 (Kumar et al. 2014a), which facilitates thrombin inhibition in two different locations such as in the vascular space and in the extravascular space, respectively.

3.9. Special genomic characteristics of serpins

Various types of gene rearrangements characterize a typical evolution of genome. The evolutionary history of serpins is demarked by several such gene rearrangements. Several duplications were the results for expansions of groups V1 (Benarafa & Remold-O'Donnell 2005) and V2 serpins (Forsyth et al. 2003b), from the basal loci of SERPINB1/SERPINB6 and SERPINA8/SERPIND1 in lampreys, respectively. In addition, several serpins in vertebrates are localized as single serpins by chromosomal duplication events such as AGT/SERPINA8 (Kumar et al. 2014d), ATIII/SERPINC1 (Kumar et al. 2013a) and HCII/SERPIND1 (Kumar et al. 2014a). HCII/SERPIND1 is conserved from lampreys to humans for ~500 MY, as a nested gene in the largest intron of PIK4CA gene. This is the only serpin that is known so far, to be nested or overlapped (Kumar et al. 2014a), and was initially reported in *Takifugu* (Kumar 2009a). Nested

443 gene is a gene that is located within a larger gene (Assis et al. 2008; Kumar 2009b). There are
444 two types of nested genes, either "within intron" genes, which are nested within the intron of the
445 host gene or "non-intronic" genes, nested within the exonic region of the host gene (Kumar
446 2009b).

447 The most interesting part of genomic characters of the serpins is the changes of exon/intron
448 patterns in the vertebrate serpins via either insertion or deletion of spliceosomal introns.
449 Spliceosomal introns and its splicing machinery are the hallmarks of eukaryotes and this adds
450 subtle intricacies to the gene regulation mechanism. However, formation of congruent sequences
451 by mere chance, execution of effective splicing and maintenance of these sequences for several
452 million years due to certain selective forces, remains quite an enigma (Roy & Gilbert 2006).
453 Intron invasion is assumed to have happened early on in evolution. Nevertheless, there are
454 several examples of late insertion of introns.

455 In total, 24 conserved introns are reported in vertebrate serpins encompassing group V1-V6
456 (Kumar & Ragg 2008), with six additional introns that were gained in selected ray finned fishes
457 among serpin genes (Ragg et al. 2009). Notably, the intron gains in the non-serpin domain of
458 C1IN have also been reported in these selected fishes (Kumar et al. 2014b). Selected ray-finned
459 fishes (namely *Fugu*, medaka, platyfish, *Tetraodon*, tilapia and stickleback) have novel introns
460 and these fishes have genomes ranging in size from 350-950 Mb or below 1000 Mb (green box
461 in **Figure 5**). Introns are either gained or lost through out eukaryotic evolution (Fedorov et al.
462 2003; Lee & Chang 2013; Lin et al. 2006; Roy & Irimia 2009; Verhelst et al. 2013; Yenerall et al.
463 2011; Yenerall & Zhou 2012; Zhu & Niu 2013). The only case of intron loss was observed in
464 the ATIII/SERPINC1 gene (Kumar et al. 2013a).

The mechanisms proposed by Yenerall's group for spliceosomal introns insertions (Yenerall et al. 2011; Yenerall & Zhou 2012) are as following: (a) intron transposition with partial recombination, (b) transposon insertion, (c) tandem genomic duplication using duplicated splice sites, (d) double-strand break repair (DSBR), (e) group II intron insertion, (f) intron transfer, and (g) intronization. Double-strand break repair (DSBR) coupled with genome compaction events are the driving forces for several examples of intron insertions in selected ray-finned fishes whose genome underwent compaction events in different group of superfamilies such as in the serpin core domain (Ragg et al. 2009), the non-core domain of serpins (Kumar et al. 2014b) and in the selected G-protein coupled receptors (Kumar et al. 2011).

Intron-exon and higher sequence similarities were maintained in serpins of a particular lineage such as in vertebrates (**Figure 2**), urochordates (Kumar & Bhandari 2014) and within insects (Kumar et al. 2014c). With several 1000 genomes being currently underway, this issue of gene structure pattern will be revisited within the next decade, when several new animal genomes will be available in the databases.

Conclusions

By utilizing Bayesian phylogenetic method, this report corroborated that the six vertebrate serpins groups are conserved from lampreys to humans for circa 500 MY. Moreover, this study provides several vignettes of vertebrate serpins from genomic and phylogenetic perspectives.

Conflict of Interests

The author declares that there are no conflicts of interests regarding the publication of this article.

Acknowledgments

I thank Chitra Rajakuberan for editing the final version of this manuscript.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.

Assis R, Kondrashov AS, Koonin EV, and Kondrashov FA. 2008. Nested genes and increasing organizational complexity of metazoan genomes. *Trends Genet* 24:475-478.

Atchley WR, Lokot T, Wollenberg K, Dress A, and Ragg H. 2001. Phylogenetic analyses of amino acid variation in the serpin proteins. *Mol Biol Evol* 18:1502-1511.

Benarafa C, and Remold-O'Donnell E. 2005. The ovalbumin serpins revisited: Perspective from the chicken genome of clade B serpin evolution in vertebrates. *Proc Natl Acad Sci USA* 102:11367 - 11372.

Bentele C, Kruger O, Todtmann U, Oley M, and Ragg H. 2006. A proprotein convertase-inhibiting serpin with an endoplasmic reticulum targeting signal from Branchiostoma lanceolatum, a close relative of vertebrates. *Biochem J* 395:449-456.

Börner S, and Ragg H. 2008. Functional diversification of a protease inhibitor gene in the genus Drosophila and its molecular basis. *Gene* 415:23 - 31.

Chen PY, Chang WS, Chou RH, Lai YK, Lin SC, Chi CY, and Wu CW. 2007. Two non-homologous brain diseases-related genes, SERPINI1 and PDCD10, are tightly linked by an asymmetric bidirectional promoter in an evolutionarily conserved manner. *BMC Mol Biol* 8:2.

Coughlin PB. 2005. Antiplasmin: the forgotten serpin? *Febs J* 272:4852-4857.

Danielli A, Kafatos F, and Loukeris T. 2003. Cloning and characterization of four Anopheles gambiae serpin isoforms, differentially induced in the midgut by Plasmodium berghei invasion. *J Biol Chem* 278:4184 - 4193.

Davis AE, Cai S, and Liu D. 2007. C1 inhibitor: biologic activities that are independent of protease inhibition. *Immunobiology* 212:313-323.

Davis AE, Lu F, and Mejia P. 2010. C1 inhibitor, a multi-functional serine protease inhibitor. *Thrombosis and haemostasis* 104:886-893.

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

Ellisdon AM, Zhang Q, Henstridge MA, Johnson TK, Warr CG, Law RH, and Whisstock JC. 2014. High resolution structure of cleaved Serpin 42 Da from Drosophila melanogaster. *BMC Struct Biol* 14:14.

Fedorov A, Roy S, Fedorova L, and Gilbert W. 2003. Mystery of intron gain. *Genome Res* 13:2236-2241.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Garcia-Giron C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kahari AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ritchie GRS, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard TJP, Johnson N, Kinsella R, Parker A, Spudich

G, Yates A, Zadissa A, and Searle SMJ. 2013. Ensembl 2013. *Nucleic Acids Research* 41:D48-D55.
Forsyth S, Horvath A, and Coughlin P. 2003a. A review and comparison of the murine alpha1-antitrypsin and alpha1-antichymotrypsin multigene clusters with the human clade A serpins. *Genomics* 81:336-345.
Forsyth S, Horvath A, and Coughlin P. 2003b. A review and comparison of the murine α1-antitrypsin and α1-antichymotrypsin multigene clusters with the human clade A serpins. *Genomics* 81:336-345.
Gettins PG. 2002. Serpin structure, mechanism, and function. *Chem Rev* 102:4751-4804.
Gettins PGW, Patston PA, and Olson ST. 1996. *Serpins: Structure, function and biology*. Austin: R. G. Landes Co.
Gregory TR. 2014. Animal Genome Size Database. http://www.genomesize.com.
Hendershot LM, and Bulleid NJ. 2000. Protein-specific chaperones: the role of hsp47 begins to gel. *Curr Biol* 10:R912-915.
Huntington JA. 2013. Thrombin inhibition by the serpins. *J Thromb Haemost* 11 Suppl 1:254-264.
Huntington JA. 2014. Natural inhibitors of thrombin. *Thromb Haemost* 111:583-589.
Ishigami S, Sandkvist M, Tsui F, Moore E, Coleman TA, and Lawrence DA. 2007. Identification of a novel targeting sequence for regulated secretion in the serine protease inhibitor neuroserpin. *Biochem J* 402:25-34.
Izuhara K, Ohta S, Kanaji S, Shiraishi H, and Arima K. 2008. Recent progress in understanding the diversity of the human ov-serpin/clade B serpin family. *Cell Mol Life Sci*.
Jordan RE. 1983. Antithrombin in vertebrate species: conservation of the heparin-dependent anticoagulant mechanism. *Arch Biochem Biophys* 227:587-595.
Kaiserman D, and Bird P. 2005. Analysis of vertebrate genomes suggests a new model for clade B serpin evolution. *BMC Genomics* 6:167.
Krem MM, and Di Cera E. 2003. Conserved Ser residues, the shutter region, and speciation in serpin evolution. *J Biol Chem* 278:37810-37814.
Kumar A. 2009a. Delving into Vertebrate Serpins for Understanding their Evolution. *Nature Precedings*.
Kumar A. 2009b. An overview of nested genes in eukaryotic genomes. *Eukaryot Cell* 8:1321-1329.
Kumar A. 2010. Phylogenomics of vertebrate serpins. Bielefeld University.
Kumar A, and Bhandari A. 2014. Urochordate serpins are Classified into Six Groups Encoded by Exon-Intron Structures , Microsynteny and Bayesian Phylogenetic Analyses. *Journal of Genomics* 2:131-140.
Kumar A, Bhandari A, Sarde SJ, and Goswami C. 2013a. Sequence, phylogenetic and variant analyses of antithrombin III. *Biochemical and Biophysical Research Communications* 440:714-724.
Kumar A, Bhandari A, Sarde SJ, and Goswami C. 2014a. Genetic variants and evolutionary analyses of heparin cofactor II. *Immunobiology* 219:713-728.
Kumar A, Bhandari A, Sarde SJ, and Goswami C. 2014b. Molecular phylogeny of C1 inhibitor depicts two immunoglobulin-like domains fusion in fishes and ray-finned fishes specific intron insertion after separation from zebrafish. *Biochemical and Biophysical Research Communications* 450:219-226.

585 Kumar A, Bhandari A, Sarde SJ, Muppavarapu S, and Tandon R. 2015. Understanding V(D)J
586 recombination initiator RAG1 gene using molecular phylogenetic and genetic variant
587 analyses and upgrading missense and non-coding variants of clinical importance.
588 *Biochemical and Biophysical Research Communications* (in press),
589 doi:10.1016/j.bbrc.2015.04.125.
590 Kumar A, Bhandari A, Sinha R, Goyal P, and Grapputo A. 2011. Spliceosomal intron insertions
591 in genome compacted ray-finned fishes as evident from phylogeny of MC receptors, also
592 supported by a few other GPCRs. *PLoS One* 6:e22046.
593 Kumar A, Congiu L, Lindstrom L, Piiroinen S, Vidotto M, and Grapputo A. 2014c. Sequencing,
594 De Novo assembly and annotation of the Colorado Potato Beetle, Leptinotarsa
595 decemlineata, Transcriptome. *PloS one* 9:e86012.
596 Kumar A, Kollath-Leiss K, and Kempken F. 2013b. Characterization of bud emergence 46
597 (BEM46) protein: sequence, structural, phylogenetic and subcellular localization analyses.
598 *Biochem Biophys Res Commun* 438:526-532.
599 Kumar A, and Ragg H. 2008. Ancestry and evolution of a secretory pathway serpin. *Bmc*
600 *Evolutionary Biology* 8:250.
601 Kumar A, Sarde SJ, and Bhandari A. 2014d. Revising angiotensinogen from phylogenetic and
602 genetic variants perspectives. *Biochemical and Biophysical Research Communications*
603 446:504-518.
604 Lamande SR, and Bateman JF. 1999. Procollagen folding and assembly: the role of endoplasmic
605 reticulum enzymes and molecular chaperones. *Semin Cell Dev Biol* 10:455-464.
606 Lee YCG, and Chang H-H. 2013. The evolution and functional significance of nested gene
607 structures in Drosophila melanogaster. *Genome Biol Evol* 5:1978-1985.
608 Lewis MJ, Sweet DJ, and Pelham HR. 1990. The ERD2 gene determines the specificity of the
609 luminal ER protein retention system. *Cell* 61:1359-1363.
610 Li Y, Liu S, Qin Z, Yao J, Jiang C, Song L, Dunham R, and Liu Z. 2015. The serpin superfamily
611 in channel catfish: identification, phylogenetic analysis and expression profiling in
612 mucosal tissues after bacterial infections. *Dev Comp Immunol* 49:267-277.
613 Lin H, Zhu W, Silva JC, Gu X, and Buell CR. 2006. Intron gain and loss in segmentally
614 duplicated genes in rice. *Genome Biol* 7:R41.
615 Nagata K. 1996. Hsp47: a collagen-specific molecular chaperone. *Trends Biochem Sci* 21:22-26.
616 Oley M, Letzel M, and Ragg H. 2004. Inhibition of furin by serpin Spn4A from Drosophila
617 melanogaster. *FEBS Lett* 577:165 - 169.
618 Osterwalder T, Cinelli P, Baici A, Pennella A, Krueger S, Schrimpf S, Meins M, and
619 Sonderegger P. 1998. The axonally secreted serine proteinase inhibitor, neuroserpin,
620 inhibits plasminogen activators and plasmin but not thrombin. *J Biol Chem* 273:2312 -
621 2321.
622 Osterwalder T, Kuhnen A, Leiserson W, Kim Y, and Keshishian H. 2004. Drosophila serpin 4
623 functions as a neuroserpin-like inhibitor of subtilisin-like proprotein convertases. *J*
624 *Neurosci* 24:5482 - 5491.
625 Pelham HR. 1990. The retention signal for soluble proteins of the endoplasmic reticulum. *Trends*
626 *Biochem Sci* 15:483-486.
627 Pelissier P, Delourme D, Germot A, Blanchet X, Becila S, Maftah A, Leveziel H, Ouali A, and
628 Bremaud L. 2008. An original SERPINA3 gene cluster: elucidation of genomic
629 organization and gene expression in the Bos taurus 21q24 region. *BMC Genomics* 9:151.

630 Ragg H, Kumar A, Koster K, Bentele C, Wang Y, Frese MA, Prib N, and Kruger O. 2009.
631 Multiple gains of spliceosomal introns in a superfamily of vertebrate protease inhibitor
632 genes. *Bmc Evolutionary Biology* 9:208.
633 Ragg H, Lokot T, Kamp PB, Atchley WR, and Dress A. 2001. Vertebrate serpins: construction
634 of a conflict-free phylogeny by combining exon-intron and diagnostic site analyses. *Mol*
635 *Biol Evol* 18:577-584.
636 Raykhel I, Alanen H, Salo K, Jurvansuu J, Nguyen VD, Latva-Ranta M, and Ruddock L. 2007.
637 A molecular specificity code for the three mammalian KDEL receptors. *J Cell Biol*
638 179:1193-1204.
639 Richer M, Keays C, Waterhouse J, Minhas J, Hashimoto C, and Jean F. 2004. The Spn4 gene of
640 Drosophila encodes a potent furin-directed secretory pathway serpin. *Proc Natl Acad Sci*
641 101:10560 - 10565.
642 Ronquist F, and Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under
643 mixed models. *Bioinformatics* 19:1572-1574.
644 Roy SW, and Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and
645 progress. *Nat Rev Genet* 7:211-221.
646 Roy SW, and Irimia M. 2009. Mystery of intron gain: new data and new models. *Trends Genet*
647 25:67-73.
648 Sauk JJ, Nikitakis N, and Siavash H. 2005. Hsp47 a novel collagen binding serpin chaperone,
649 autoantigen and therapeutic target. *Front Biosci* 10:107-118.
650 Sawant S, Aparicio S, Tink AR, Lara N, Barnstable CJ, and Tombran-Tink J. 2004. Regulation
651 of factors controlling angiogenesis in liver development: a role for PEDF in the formation
652 and maintenance of normal vasculature. *Biochem Biophys Res Commun* 325:408-413.
653 Semenza JC, Hardwick KG, Dean N, and Pelham HR. 1990. ERD2, a yeast gene required for the
654 receptor-mediated retrieval of luminal ER proteins from the secretory pathway. *Cell*
655 61:1349-1357.
656 Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, and Xenarios I.
657 2013. New and continuing developments at PROSITE. *Nucleic Acids Res* 41:D344-347.
658 Silverman GA, Bird PI, Carrell RW, Church FC, Coughlin PB, Gettins PG, Irving JA, Lomas
659 DA, Luke CJ, Moyer RW, Pemberton PA, Remold-O'Donnell E, Salvesen GS, Travis J,
660 and Whisstock JC. 2001. The serpins are an expanding superfamily of structurally similar
661 but functionally diverse proteins. Evolution, mechanism of inhibition, novel functions,
662 and a revised nomenclature. *J Biol Chem* 276:33293-33296.
663 Stanke M, and Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in
664 eukaryotes that allows user-defined constraints. *Nucleic Acids Research* 33:W465-W467.
665 Steele FR, Chader GJ, Johnson LV, and Tombran-Tink J. 1993. Pigment epithelium-derived
666 factor: neurotrophic activity and identification as a member of the serine protease
667 inhibitor gene family. *Proc Natl Acad Sci U S A* 90:1526-1530.
668 Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S. 2011. MEGA5: molecular
669 evolutionary genetics analysis using maximum likelihood, evolutionary distance, and
670 maximum parsimony methods. *Mol Biol Evol* 28:2731-2739.
671 Tombran-Tink J. 2005. The neuroprotective and angiogenesis inhibitory serpin, PEDF: new
672 insights into phylogeny, function, and signaling. *Front Biosci* 10:2131-2149.
673 Verhelst B, Van de Peer Y, and Rouzé P. 2013. The complex intron landscape and massive
674 intron invasion in a picoeukaryote provides insights into intron evolution. *Genome*
675 *biology and evolution* 5:2393-2401.

676 Wang Y, and Ragg H. 2011. An unexpected link between angiotensinogen and thrombin. *FEBS*
677 *Lett* 585:2395-2399.
678 Wong MKS, and Takei Y. 2011. Characterization of a native angiotensin from an anciently
679 diverged serine protease inhibitor in lamprey. *J Endocrinol* 209:127-137.
680 Yenerall P, Krupa B, and Zhou L. 2011. Mechanisms of intron gain and loss in Drosophila. *BMC*
681 *Evol Biol* 11:364.
682 Yenerall P, and Zhou L. 2012. Identifying the mechanisms of intron gain: progress and trends.
683 *Biol Direct* 7:29.
684 Zhu T, and Niu DK. 2013. Mechanisms of intron loss and gain in the fission yeast
685 Schizosaccharomyces. *PLoS One* 8:e61683.
686

Legends of Figures

Figure 1. Bayesian phylogenetic history of vertebrate serpins reveals that exon-intron and rare indel based classification system is retained over a period of 500 MY with conserved patterns from early diverging lampreys. Novel introns are inserted in groups V2 and V6 serpins in core domains are marked by + sign in red color whereas introns inserted in additional Ig domains of fish-specific C1 inhibitor are shown in blue + sign. Sea anemone serpin (Nve_Spn1) is the out-group for this phylogenetic analysis and it is marked by an arrow. Lamprey serpins are marked by green star. HSP47 has two isoforms in lamprey named as HSP47_1_PMA and HSP47_2_PMA.
DRE – Danio rerio, HSA – Homo sapiens, GGA – Gallus gallus, MMU – *Mus musculus*, PMA – *Petromyzon marinus*, RNO – *Rattus norvegicus*, TRU – *Takifugu rubripes*, TNI – *Tetraodon nigroviridis* and XTR – *Xenopus tropicalis*. p – paralog of a gene.

Figure 2. Summary of six groups (V1-V6) classification system of vertebrate serpins, based on introns and rare indels. Conserved intron positions are shown in cyan and yellow boxes for positions 167a and 192a, respectively. Fish-specific inserted introns are illustrated in different colors. Non-inhibitory serpins are shown in square boxes. Presence and absence of sequence indel of two amino acids between positions 173-174 are marked in by red + and – signs, respectively. OVA – Ovalbumin; Gene Y – Chicken gene Y protein; Gene X – Chicken gene X protein; PAI – Plasminogen activator inhibitor; SCCA – Squamous cell carcinoma antigen; α_1-AT – α_1-antitrypsin; α_1-ACT – α_1-antichymotrypsin; CBG – Corticosteroid-binding globulin; TBG – Thyroxine-binding globulin; HCII – Heparin cofactor II; PCI – Protein C Inhibitor; AGT – Angiotensinogen; E3 – SERPINE3; Neuro – Neuroserpin; Panc – Pancpin; A2AP – α_2-Antiplasmin; PEDF – Pigment epithelium derived factor; C1IN – C1-Inhibitor; ATIII – Antithrombin III; HSP47 – Heat shock protein 47kDa.

Figure 3. Serpin motifs of ATIII proteins.

Figure 4. Genomic localization of fish-specific HSP47_2 gene.

Figure 5. Spliceosomal introns are inserted only in selected ray-finned fishes with genome size lower than 1000 Mb.

Legends of tables

Table 1. Vertebrate genomes used during this study.

Table 2. Summary of serpins in two lampreys namely, sea lamprey (*Petromyzon marinus*) and European river lamprey (*Lampetra fluviatilis*).

Table 3. Sequence comparisons of HSP47 homologs in vertebrates. Percentage sequence identity (SI) and percentage sequence similarity (SS) values are shown as compared to HSP47_HSA and A1AT_HSA. Synteny based clustering divides group V6 genes into three sets: set I – true mammalian HSP47 orthologs (bold font), set II - fish specific paralogs (underlined font) and set III (cursive font).

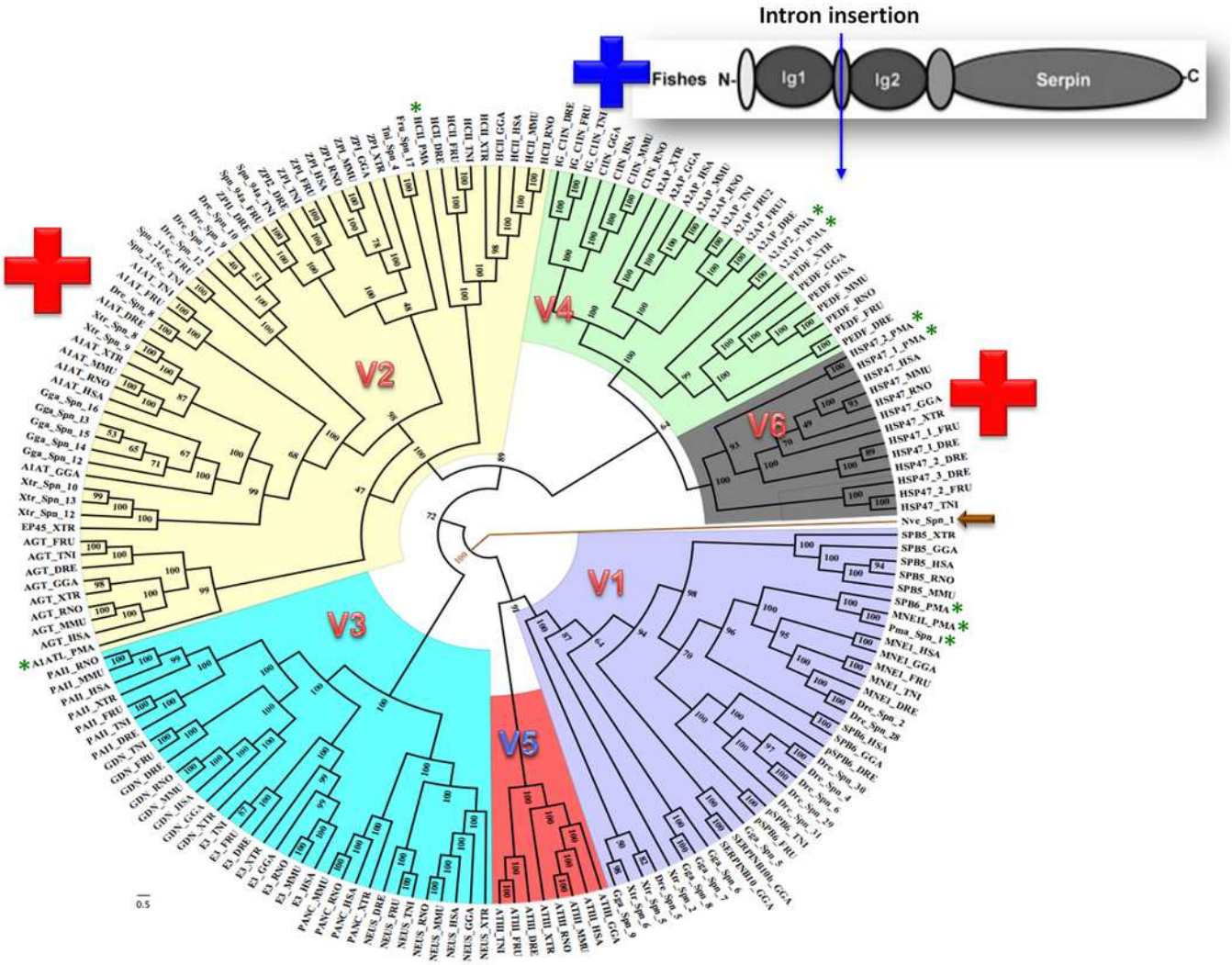
Supplementary files

Table S1. Maximum Likelihood fits of 44 different amino acid substitution models of alignment of serpins using MEGA 5. The lowest BIC scores (Bayesian Information Criterion) are considered for the best fit of the substitution pattern

1

Bayesian phylogenetic history of vertebrate serpins reveals that exon-intron and rare indel based classification system is retained over period of 500 MY with conserved patterns from early diverging lampreys.

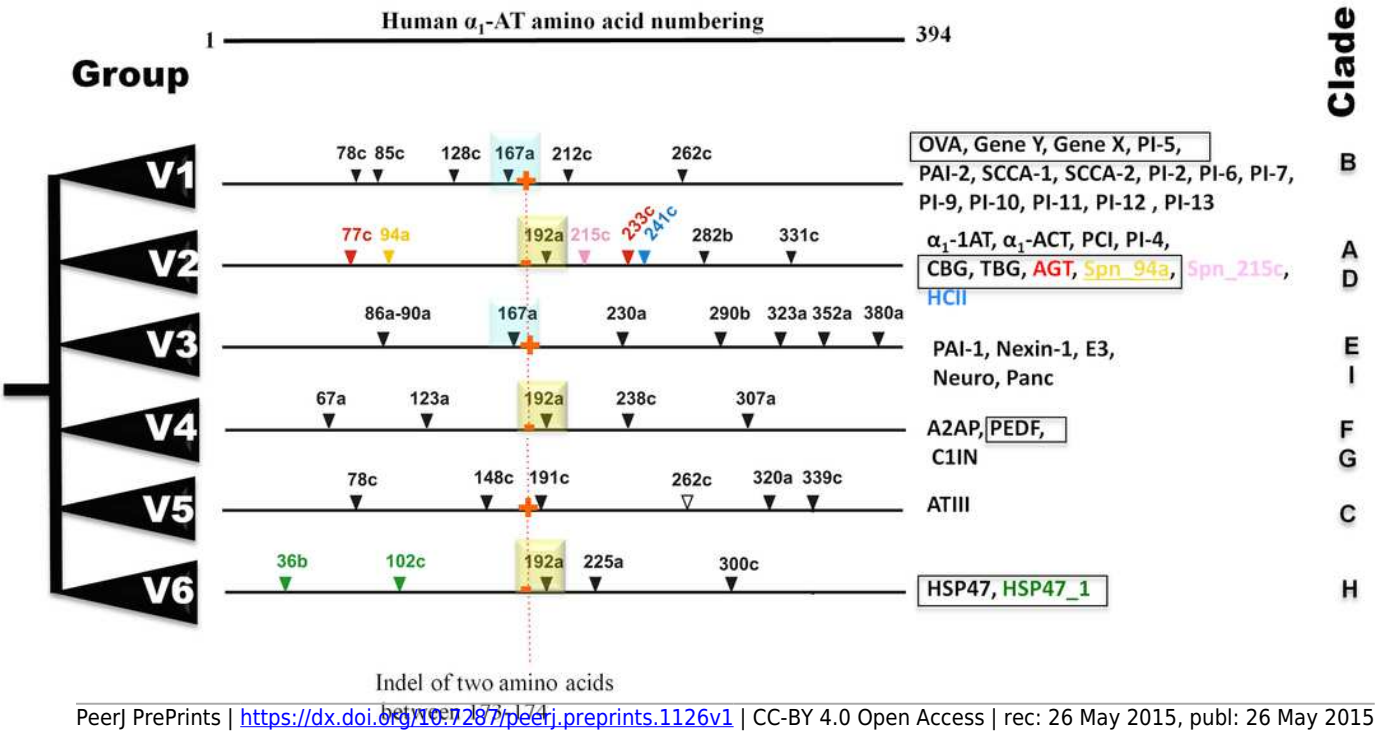
Novel introns are inserted in groups V2 and V6 serpins in core domains are marked by + sign in red color while intron inserted in additional Ig domains of fish-specific C1 inhibitors is shown in blue + sign. A sea anemone serpin (Nve_Spn1) is the out-group for this phylogenetic analysis as marked by an arrow. Lamprey serpins are marked by green star. HSP47 has two isoforms in lamprey named as HSP47_1_PMA and HSP47_2_PMA. DRE – Danio rerio, H SA – Homo sapiens, GGA – Gallus gallus, MMU – Mus musculus, PMA – *Petromyzon marinus*, RNO – *Rattus norvegicus*, TRU – *Takifugu rubripes*, TNI – *Tetraodon nigroviridis* and XTR – *Xenopus tropicalis*. p – paralog of a gene.



2

Summary of six groups (V1-V6) classification system of vertebrate serpins, based on introns and rare indels.

Conserved intron positions are shown in cyan and yellow boxes for positions 167a and 192a, respectively. Fish-specific introns are inserted in selected serpins are illustrated in different colors. Non-inhibitory serpins are shown in square boxes. Presence and absesnce of sequence indel of two amino acid between positions 173-174 are marked in by red + and – signs. OVA – Ovalbumin; Gene Y – Chicken gene Y protein; Gene X – Chicken gene X protein; PAI – Plasminogen activator inhibitor; SCCA – Squamous cell carcinoma antigen; α_1-AT – α_1-antitrypsin; α_1-ACT – α_1-antichymotrypsin; CBG – Corticosteroid-binding globulin; TBG – Thyroxine-binding globulin; HCII – Heparin cofactor II; PCI – ProteinCInhibitor; AGT – Angiotensinogen; E3 – SerpinE3; Neuro – Neuroserpin; Panc – Pancpin; A2AP – α_2-Antiplasmin; PEDF – Pigment epithelium derived factor; C1IN – C1-Inhibitor; ATIII – Antithrombin III; HSP47 – Heat shock protein 47kDa.

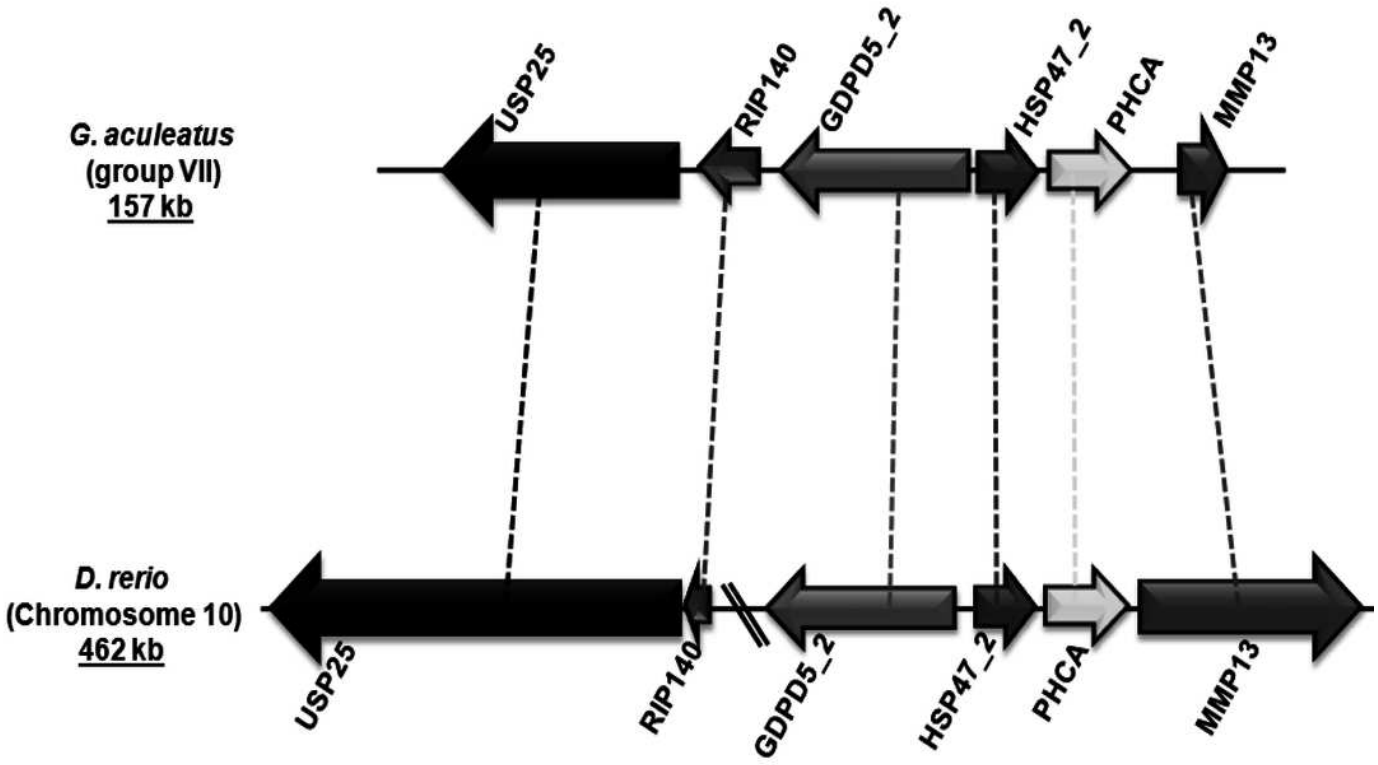


3

Serpin motifs of ATIII proteins.

4

Genomic localization of fish-specific HSP47_2 gene.



5

Spliceosomal introns are inserted only in selected ray-finned fishes with genome size lower than 1000 Mb.

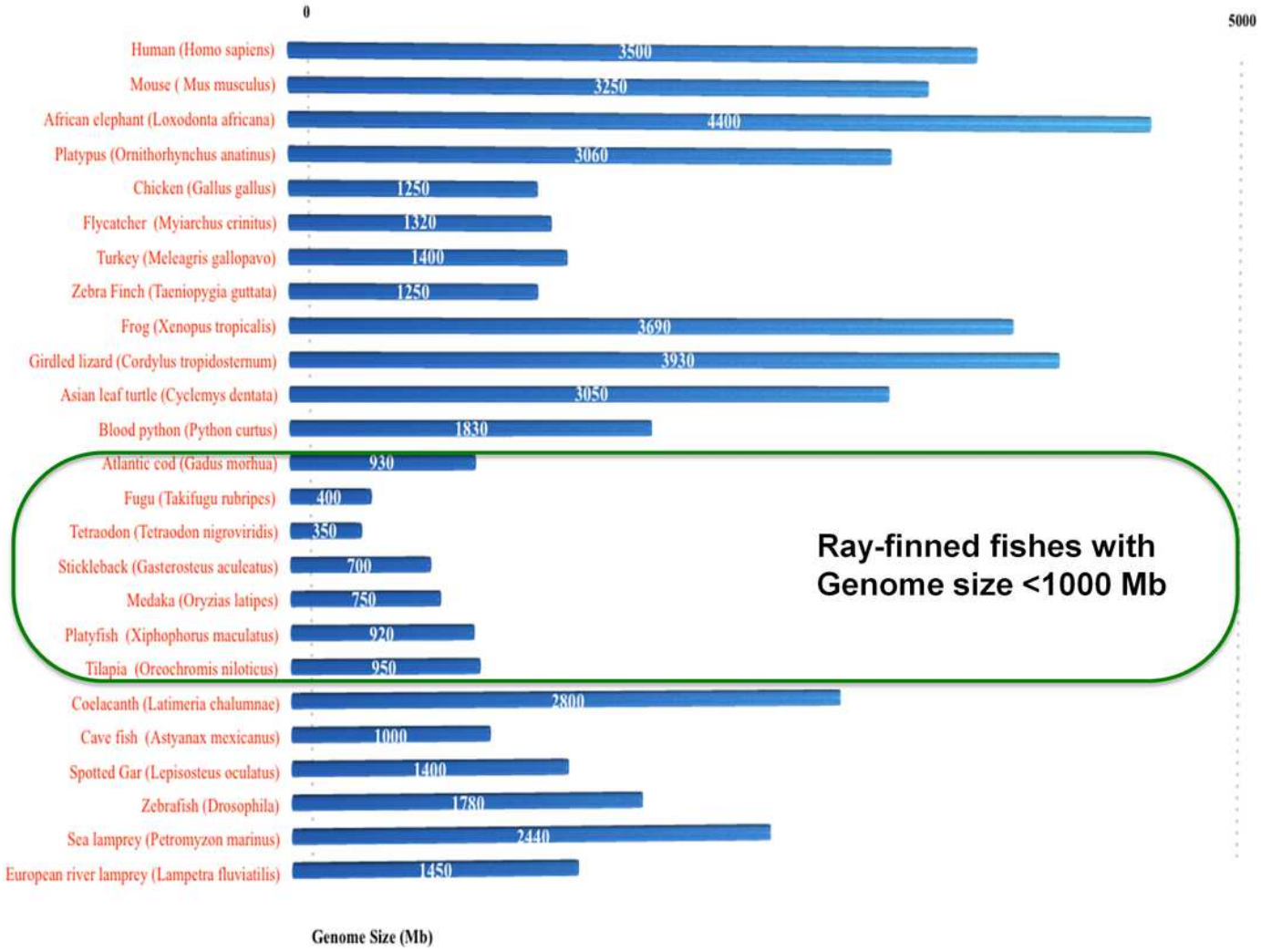


Table 1(on next page)

Vertebrate genomes used during this study

Table 1. Vertebrate genomes used during this study

Genome	Major database used	References
Homo sapiens	http://www.ncbi.nlm.nih.gov/genome/guide/human/	(Venter et al. 2001)
Mus musculus	http://www.ncbi.nlm.nih.gov/genome/guide/mouse/	(Waterston et al. 2002)
Rattus norvegicus	http://www.ncbi.nlm.nih.gov/genome/guide/rat/	(Gibbs et al. 2004)
Gallus gallus	http://www.ncbi.nlm.nih.gov/genome/guide/chicken/	(Hillier et al. 2004)
Xenopus tropicalis	http://genome.jgi-psf.org/Xentr4/Xentr4.home.html	(Hellsten et al. 2010)
Fugu rubripes	http://genome.jgi-psf.org/Takru4/Takru4.home.html	(Aparicio et al. 2002)
Tetraodon nigroviridis	http://www.genoscope.cns.fr/externe/tetranew/	(Jaillon et al. 2004)
Danio rerio	http://www.ensembl.org/Danio_rerio/index.html	(Birney et al. 2006)
Petromyzon marinus	http://www.ensembl.org/Petromyzon_marinus/Info/Index	(Smith et al. 2013)

References:

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A et al. . 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 297:1301-1310.

Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T et al. . 2006. Ensembl 2006. *Nucleic Acids Res* 34:D556-561.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE et al. . 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428:493-521.

Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, Kapitonov V, Ovcharenko I, Putnam NH, Shu S, Taher L et al. . 2010. The genome of the Western clawed frog Xenopus tropicalis. *Science* 328:633-636.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MA, Delany ME et al. . 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695-716.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A et al. . 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* 431:946-957.

Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, Yandell MD, Manousaki T, Meyer A, Bloom OE et al. . 2013. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. *Nat Genet* 45:415-421, 421e411-412.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al. . 2001. The sequence of the human genome. *Science* 291:1304-1351.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P et al. . 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520-562.

Table 2 (on next page)

Summary of serpins in two lampreys namely, sea lamprey (*Petromyzon marinus*) and European river lamprey (*Lampetra fluviatilis*).

Table 2.

Name Given	Ensembl Accession id	Serpin name	Group	Clade	RCL with P1--P1'	Ortholog in *L. fluviatilis* Genbank ID
Pma-Spn-1	ENSPMAG00000001963	SERPINB1-like	V1	B	GTEAAAATAAIVMMR--CARMG	
MNE1L_PMA/ Pma-Spn-2	ENSPMAG00000009027	MNEI1-like/SERPINB1-like	V1	B	GTEAAAATAVTMKLR--CAMPT	
SPB6_PMA (Pma-Spn-3)	ENSPMAG00000009040	SPB6/SERPINB6-like	V1	B	GTEAAAATAISVMLM--CAMPT	
A1ATL_PMA (Pma-Spn-4)	ENSPMAG00000006108	A1AT-like, angiotensinogen, SERPINA8	V2	A	GTEAKAETVVGIMPI--SMPPT	CAV16871.1/CAV29466.1
A1ATL_PMA (Pma-Spn-5)	ENSPMAG00000008131	Heparin cofactor II/SERPIND1	V2	D	GSEAAAVTTVGFTPL--TSHNR	CAX18777.1/AIA57696.1
A2APL1_PMA (Pma-Spn-6)	ENSPMAG00000008124	Alpha-2-antiplasmin-like 1	V4	F	GVKATAATGIMISLM--SVQHS	CAX18777.1/CAX18778.1
A2APL2_PMA (Pma-Spn-7)	ENSPMAG00000002992	Alpha-2-antiplasmin-like 2, A2APL2_PMA	V4	F	GAEAAAVTGVFLSRT--NPIYP	AIE16052.1/AIE16053.1
HSP47_PMA (Pma-Spn-8)	ENSPMAG00000007485	HSP47/SERPINH1	V6	H	GEEYDMSVHGHPDM--RNPHL	

Table 3 (on next page)

Sequence comparisons of HSP47 homologs in vertebrates.

Percentage sequence identity (SI) and percentage sequence similarity (SS) values are shown as compared to HSP47_HSA and A1AT_HSA. Synteny based clustering divides group V6 genes into three sets: set I – true mammalian HSP47 orthologs (bold font), set II - fish specific paralogs (underlined font) and set III (cursive font).

Table 3.

Human Serpins	Values (%)	HSP47_MMU	HSP47_RNO	HSP47_GGA	HSP47_XTR	HSP47_1_FRU	HSP47_2_FRU	HSP47_TNI	HSP47_1_DRE	HSP47_2_DRE	HSP47_3_DRE	HSP47_PMA
HSP47_HSA	SI	96	96	76	70	**63**	*29*	*22*	65	**64**	*29*	46
	SS	98	98	88	83	**82**	*46*	*37*	83	**82**	*52*	65
A1AT_HSA	SI	23	23	25	24	**24**	*18*	*14*	25	**24**	*17*	23
	SS	45	45	45	46	**44**	*35*	*26*	45	**46**	*37*	41