Biochemical and Biophysical Research Communications 440 (2013) 714–724

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Sequence, phylogenetic and variant analyses of antithrombin III

Abhishek Kumar [a,*], Anita Bhandari [b], Sandeep J. Sarde [c], Chandan Goswami [d]

[a] Dept. of Genetics & Molecular Biology in Botany, Institute of Botany, Christian-Albrechts-University at Kiel, Kiel, Germany
[b] Zoological Institute, Christian-Albrechts-University at Kiel, Kiel, Germany
[c] Master program Agrigenomics, Christian-Albrechts-University at Kiel, Kiel, Germany
[d] National Institute of Science Education and Research, Bhubaneswar, Orissa, India

ARTICLE INFO

Article history:
Received 21 September 2013
Available online 9 October 2013

Keywords:
Antithrombin III
SerpinC1
Group V5
Synteny
Phylogenetic analysis
ATIII variants

ABSTRACT

Antithrombin III (ATIII) performs a critical anticoagulant function by precluding the activation of blood clotting proteinases, aided by its two cofactors, heparin and heparan sulfate. Though several studies have been carried out on physiological, biochemical and structural perspectives on ATIII, so far there are limited studies on the molecular evolution of ATIII. Herein, we carried out molecular phylogenetic analyses of ATIII genes, combining gene structures, synteny and sequence-structural features for ATIII spanning 50 vertebrate genomes. ATIII is maintained over 450 MY on same genomic loci in vertebrates with few changes in ray-finned fishes and lost one intron 262c in tetrapods and coelacanth. In ray-finned fishes, ATIII gene has additional intron at the position 262c, which shared by group V1 members, corroborating that it is lost in other vertebrates and also in lobed-finned fish coelacanth (*Latimeria chalumnae*). We found that heparin binding basic residues, hD helix, three pairs of Cys–Cys salt bridges, N-glycosylation sites, serpin motifs and inhibitory reactive center loop (RCL) of ATIII protein are highly conserved. Using 1092 human genomes available from 1000G project, we also compiled 1997 ATIII variants, which reveals 76.2% single nucleotide polymorphisms (SNPs), 11.8% deletions and 8.1% insertions as three major classes of gene variations. These understandings may have medical importance as well.

1. Introduction

Haemostasis is an important process which regulates bleeding control mechanism in response to vascular injury in vertebrates. There are several proteases and their inhibitors, which are involved in this process. Unlike other factors, ATIII is the major yet rare inhibitor of blood coagulation which requires binding of poly-sulphated glycosaminoglycan heparin for full activation [1]. ATIII belongs to the serine protease inhibitors (serpin) superfamily that covers a highly divergent spectrum of functions. Serpins are primarily inhibitors of serine and/or cysteine proteases, but some family members are known to perform completely other tasks such as collagen-specific chaperone, HSP47 [1]. ATIII is an exclusive member of clade C (serpinC1) within the clade-based classification [2]. Under introns-encoded classification, this gene belongs to group V5 of six vertebrate serpin groups (V1–V6) [2]. The conserved three-dimensional structure of serpins is mainly consists of three β-sheets (sA–sC) and 8–9 α-helices (hA–hI) [1,2]. Like other serpins, ATIII inhibits its target proteases by an unusual branched pathway known as "suicide substrate mechanism" by forming an irreversible kinetic trap [1]. The hallmark of the inhibitory serpin mechanism is the exposed flexible loop (~17–20 residues) known as reactive center loop (RCL), which serves as a bait mimicking a protease substrate that is cleaved between the active sites P1 and P1′ [1,2].

Human ATIII is reported to be involved in other processes such as in anti-angiogenesis, inflammation and it also has antiviral properties [3]. Previous studies of ATIII sequence analysis were carried with 10–13 vertebrate species only [4,5]. In addition, there are limited reports on molecular evolution of ATIII genes, primarily because groups V1 and V2 possess several paralogs in vertebrates. In this study, we investigated the detailed molecular phylogeny of ATIII genes by combining sequence, protein structure, gene structures and synteny analysis for ATIII from 50 vertebrate genomes. We corroborate this molecular evolution studies in human population and we have created catalog of ATIII gene variants from 1092 human genomes.

2. Materials and methods

2.1. Data collection

The genomic DNA/cDNA/protein sequences from different eukaryotes were extracted via different BLAST [6] searches using human ATIII as query sequence from Ensembl Release 73 (September 2013) [7] and are listed in Table S1 from vertebrate genome assemblies (Table S2).

[*] Corresponding author.
E-mail address: akumar@bot.uni-kiel.de (A. Kumar).

0006-291X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2013.09.134

2.2. Gene structure prediction and mapping introns positions

To ensure correct gene structures of ATIII genes, gene structure prediction within the Ensembl [7] was taken and combined with predictions of AUGUSTUS gene prediction tool [8]. Mature human α_1-antitrypsin was used as standard sequence for intron position mapping and numbering of intron positions, followed by suffixes a–c for their location as reported previously [9].

2.3. Chromosomal mapping of ATIII locus

Chromosomal mapping of ATIII locus was performed using ENSEMBL genome browser [7] and NCBI mapviewer.

2.4. Sequence alignment of different serpins

We aligned ATIII protein sequences from different vertebrates under consideration using the MUSCLE alignment tool [10] with default settings. We edited and visualized alignments for different ATIII characteristics using GENEDOC [11] as shown in Fig. 1S. During this study, amino acid numbering of full length ATIII protein was taken into consideration. Sequence logos of conserved motifs in ATIII proteins were constructed by Weblogo 3.3 [12].

2.5. Phylogenetic analyses

Two phylogenetic trees were constructed using Maximum Likelihood method using MEGA 5 [13], namely vertebrate serpins (259 serpins) and ATIII proteins (45 sequences), respectively. Vertebrate genomes used in this study are listed in Table S2. These two trees are based on WAG [5 categories (+*G*, parameter = 46,121)] and JTT [5 categories (+*G*, parameter = 12,246)] protein substitution models, respectively, as their best models computed in MEGA 5 [13] with 1000 bootstraps. Branches corresponding to partitions reproduced in less than 25% bootstrap replicates are collapsed for visualization in MEGA 5 [13].

2.6. Structural analyses

Structural model of coelacanth ATIII protein was created using I-TASSER [14] and visualized using Pymol [15].

2.7. Scanning ATIII variants in 1092 human genomes

ATIII variants were computed from 1092 human genomes from 14 different populations available in 1000 genomes project [16]. Sorting Intolerant From Tolerant (SIFT) is a software tool that predicts whether an amino acid substitution affects protein function and it helps in prioritizing substitutions for further study [17]. Polymorphism Phenotyping v2 (PolyPhen-v2) is a tool that predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations [18]. Evaluation of the ATIII variant impact on human ATIII protein was performed using these two methods.

3. Results

3.1. Ray-finned fishes possess an additional intron at the position 262c in conserved core domain of ATIII

After compiling ATIII genes from 50 different vertebrates (Table S1), we determined gene structures and illustrated selected species as depicted in Fig. 1A. In the conserved core of this gene, the gene structure is group V5-specific with introns at positions 78c, 148c, 191c, 320a and 339c, which are maintained in all vertebrates. In human and coelacanth, the 5′ region of ATIII genes possesses one and two exons of sizes 41, 82 and 34 base pairs, respectively. These non-canonical introns cause extension of the N-terminal region in ATIII protein (Fig. 1S). ATIII gene from ray-finned fishes possesses at an additional intron at the position 262c, splits the largest exon IV (size 391–394) into two exons E4–E5 with sizes 219 and 175 base pairs, respectively (exception in Atlantic cod, exon size is reduced by one codon). The size of intron at the position 262c varied from 80 (in *Takifugu*) to 2128 base pairs (in Medaka). This is a remarkable difference between ATIII orthologs of ray-finned fish and the orthologs of tetrapods. This intron position is normally a characteristic for group V1 serpins and 7-exons/8-introns gene structure is the signature of group V1a. Group V1a serpins are only found in ray-finned fishes and the ancestral locus of group V1 serpins possesses a gene with 7-exons/8-introns gene structure [5,19]. Hence, it is postulated that group V5 is originated by duplication and diversification of group V1a, very early in the vertebrate evolution (ca ∼500 MYA [20]). As of September 2013, genomic assemblies of sea lamprey (Pmarinus_7.0) and elephant shark (Eshark 1.4X assembly) lack ATIII gene, it hinders delineating ATIII gene and origin of 262c intron.

Table 1
Percentage identities and similarities values of ATIII for selected vertebrates.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. Human		70	66	66	66	53	64	61	69	55	55	48	57	54	54	55	55	55
2. Platypus	81		63	64	62	51	64	59	65	55	56	47	56	55	54	54	55	54
3. Chicken	80	76		86	92	66	63	64	72	55	57	50	59	57	58	55	59	60
4. Duck	80	78	91		87	65	62	63	75	57	54	48	57	54	55	53	56	57
5. Turkey	80	77	95	93		64	60	65	72	56	55	47	56	55	55	52	56	58
6. Zebrafinch	67	65	76	74	74		50	53	59	45	46	41	47	47	46	44	48	47
7. Xenopus	76	77	79	78	76	66		60	63	54	57	48	57	57	56	56	57	55
8. Lizard	75	73	79	80	79	65	73		63	53	56	48	56	57	55	54	58	58
9. Turtle	83	79	84	87	86	71	77	80		58	56	48	57	56	56	54	56	57
10. Coelacanth	71	70	71	72	71	60	69	69	73		49	43	51	49	49	49	50	48
11. Fugu	72	71	74	72	72	62	74	73	73	69		72	68	74	70	73	75	65
12. Tetrandon	66	63	67	65	65	57	66	65	66	61	80		60	65	61	63	65	58
13. Cod	72	70	75	71	72	62	72	71	72	69	83	75		70	69	68	72	63
14. Tilapia	72	69	75	72	73	62	73	71	72	69	85	77	84		76	74	78	68
15. Medaka	71	69	74	72	73	63	72	70	71	68	83	73	83	87		72	72	66
16. Platyfish	70	67	74	72	72	61	71	70	71	67	84	75	83	85	84		72	65
17. Stickleback	71	69	75	73	73	64	72	73	72	68	86	78	85	88	85	84		67
18. Zebrafish	70	71	76	74	74	65	72	73	74	69	77	70	78	80	78	78	79	

Sequence identities

Sequence similarities

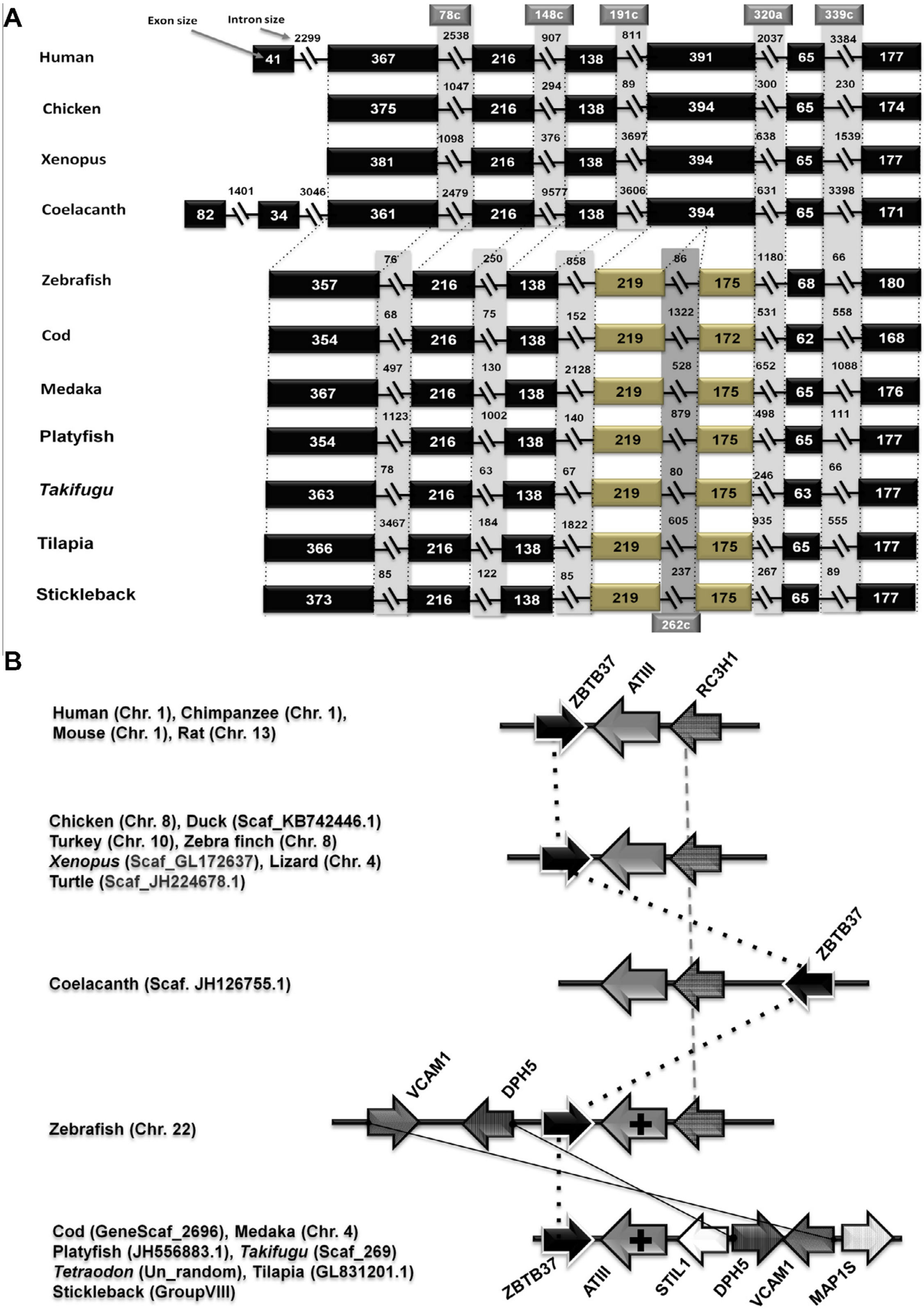

Fig. 1. Comparisons of gene structures and chromosomal localization of ATIII genes in vertebrates. (A) Exon–intron structures of ATIII genes from selected vertebrates illustrates that ray-finned fishes have an intron at the position 262c, but not in other vertebrates including coelacanth. Introns localized into the conserved serpin domain are considered in this comparison. ATIII genes from human and coelacanth have additional introns in non-conserved serpin domain. Standard nomenclature of inton position of serpin is according to amino acid numbering of mature human α_1-antitrypsin coupled by three intron phases (A–C) [9]. (B) Chromosomal mapping of ATIII loci reveals that this loci is conserved with a few variations in fishes. Fishes with introns at the position 262c is marked by + sign.

3.2. Synteny analysis of ATIII genes

To evaluate the syntenic conservation of ATIII locus in different vertebrates, chromosomal localization analysis was carried out (Fig. 1B). The ATIII gene in the human chromosome 1 is flanked by two transcription factor genes namely ring finger and CCCH-type domains 1 (RC3H1) on one side and the zinc finger and BTB domain containing 37 (ZBTB37) gene (opposite orientation) on the other side. This genomic architecture is maintained in chimpanzee (chr. 1), gorilla (chr. 1), mouse (chr. 1), rat (chr. 13) and several other mammals as documented in Table S1. Similarly, this syntenic organization is conserved in non-mammalian tetrapods including chicken (Chr. 8), duck (Scaf_KB742446.1), turkey (chr. 10), zebra finch (chr. 8), *Xenopus* (Scaf_GL172637), anole lizard (chr. 4) and Chinese softshell turtle (Scaf_JH224678.1). In the coelacanth genome, ZBTB37 gene moved to the other side, after the RC3H1 gene and inverted its orientation. Zebrafish possess genomic organization same as tetrapods but a dyad of vascular cell adhesion molecule 1 (VCAM1) and yeast DPH5 homolog (DPH5) are localized before ZBTB37 gene. In other ray-finned fishes, the ATIII-ZBTB37 synteny is conserved, but on the other side, a tetrad of genes namely microtubule-associated protein 1S (MAP1S), VCAM1, DPH5 and SCL/TAL1 interrupting locus (STIL). However, location and direction of VCAM1 and DPH5 differed from that on ATIII in zebrafish loci. In the nutshell, this analysis demonstrates that the ATIII gene synteny is conserved in different vertebrates on a single locus with some variations in fishes.

3.3. ATIII proteins are highly conserved

The ATIII gene is highly conserved in vertebrates (Fig. 1S and Table). Human ATIII shares 70/81, 66/80, 69/83 and 57/72 percentage identities/similarities with platypus, chicken, turtle and Atlantic cod, respectively. There are 47 and 213 amino acids in ATIII alignment, which are conserved 100% and 70–99%, respectively (Table 2). Additionally, there are 51 amino acids conserved in the core serpin domain in more than 70% of serpins [21] and from these 51 residues, 47 were maintained in vertebrate ATIII (Table 2). From this alignment (Fig. 1S), several signature sequences have been deduced and are described further in details.

3.4. Highly conserved eight basic residues are essential for heparin binding in vertebrate ATIII

The heparin binding site is defined by the basic residues present in the N-terminal region of human AIII, namely within the flexible N-terminus, the amino-terminal tip of the helix-A (hA) and within the helix-D (hD) [22]. There are eight basic residues, which are important in heparin binding:- two in the N-terminal segment at positions K43 and K45 (numbering according to the full length human ATIII), which are completely conserved in the all vertebrates (Fig. 2A) with one exception as K43G mutation in coelacanth ATIII (Fig. 1S). The helix hA possesses two basic residues as R78 and R79 (Fig. 2B). There is R78P replacement in 21 ATIII from different species including four mammals and non-mammalian vertebrates. In contrast, the R79 is highly conserved and only two exceptions are R79Y and R79-(gap) mutations in coelacanth and in *Tetraodon* (Fig. 1S), respectively. There are four basic residues (K157, R161, [RK]164, and R165) in the highly conserved hD region (Fig. 2C) in the full-length human ATIII. Residues K157 and R161 are conserved in all species taken into consideration, where as both R and K are possible at the [RK]164, hence named so. The R165 of ATIII is conserved in majority of species with one exception in gorilla due to R165L mutation. The helix hD is marked red in the protein model of coelacanth ATIII (Fig. 3A). No other vertebrate serpins have these specific arrangements of basic residues in the hD region [5].

Table 2
Summary of amino acid conservation in secondary structural elements of ATIII protein in vertebrates.

Structural component	Id-100	Id-70–99	Status of 51 conserved amino acids proposed by Irwing et al. (2000) bold – missing				
Signal peptide	0	7					
N-terminal end after signal peptide	6	24					
hA	6	14	Phe33				
s6B	0	2	Asn49	Ser53			
hB	6	6	**Pro54**	Ser56	**Leu61**	Gly67	
hC	2	4	Thr72	Leu80			
hD	3	14					
s2A	2	7					
hE	0	9	**Phe130**				
s1A	0	2					
hF	1	11	Phe147	Ile157	Asn158	Val16	Thr165
Loop between hF/s3A	1	13	Ile169	Thr180			
s3A	0	8	Leu184	Asn186	Phe190	Lys191	Gly19
hF1	0	1					
s4C	0	1	Phe198	Thr203	Phe208		
s3C	0	8	Val218	Met220	Met221		
s1C	0	1					
s2B	0	6	Tyr244				
s3B	0	8	Leu254	Pro255			
hG	0	4					
hH	0	4					
s2C	0	4					
s6A	0	6	Pro289	**Lys290**			
hI	0	4	Leu299	Leu303	Gly307		
hI1	0	3					
Loop between hI/s5A	1	14	Phe312	Ala316	Leu327		
s5A	2	8	His334	Glu342			
s4A (RCL)	9	9	Gly344	Ala347			
s1C	1	0					
s4B	4	2	Pro369	Phe370			
s5B	2	6	Leu383	Phe384	Gly386		
C terminus	1	3	Pro391				
Total	47	213					

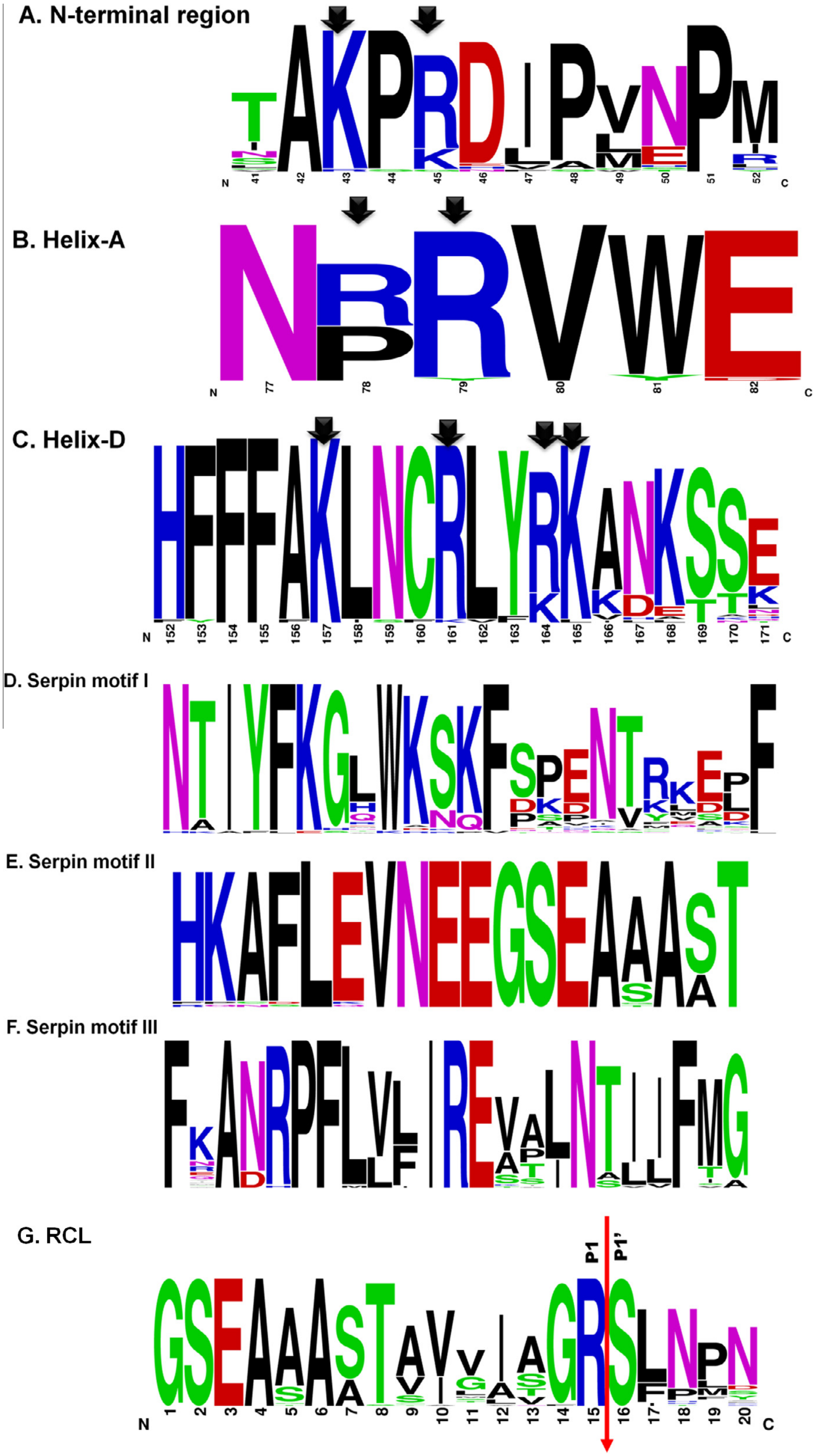
A. N-terminal region
T A K P R D I P V N P M
N 41 42 43 44 45 46 47 48 49 50 51 52 C
B. Helix-A
N R R V W E
N 77 78 79 80 81 82 C
C. Helix-D
H F F F A K L N C R L Y R K A N K S S E
N 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 C
D. Serpin motif I
N T I Y F K G L W K S K F S P E N T R K E P F
E. Serpin motif II
H K A F L E V N E E G S E A A A S T
F. Serpin motif III
F K A N R P F L V L I R E V A L N T I I F M G
G. RCL
G S E A A A S T A V V I A G R S L N P N
P1 P1'
N 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 C

Fig. 2. Conservation of sequence motifs of ATIII proteins. Basic residues (black arrows) are essential for heparin binding located in the N-terminal region (A), at the start of helix-A (B) and within highly conserved helix-D (C). ATIII possess all three segments of serpin motifs as I–III (D–F). Highly conserved RCL region (G) with the cleavage site between P1–P1′ is marked with an arrow.

3.5. Serpin motifs are highly conserved in vertebrate ATIII protein

ATIII are characterized by three conserved serpin motifs, which are spanned across three major structural regions (Fig. 1S). Motif-I is localized in s3A–breach–s4C and is conserved with variable residues at some positions (Fig. 2D), gorilla ATIII causes maximum divergence in this region. Motif-II is highly conserved in the s5A–s4A, which starts just before RCL and ends in P8 position of the RCL (Fig. 2E) without much variation. Motif-III is immediately after RCL at the C-terminal end, 1C-turn-s4B and this position is also highly conserved (Fig. 2F).

3.6. Inhibitory reactive center loop (RCL) is highly conserved

The inhibitory RCL region is another highly conserved region with P1–P1′ position (R–S) maintained in all vertebrates as shown in Fig. 2G (also marked in red shade in Fig. 1S). The hinge region residues P14–P15 as G–S are 100% conserved in all vertebrate ATIII protein, whereas in majority of other serpins possess G–T. Other 100% conserved positions are P13–P12, P10, P8, P2–1 and P1′ (Fig. 1S and Fig. 2G). The positions P3–P7 shows three to four small amino acid replacements where as positions P9 and P11 have two amino acid replacements (either alanine or serine). Conserved RCL is marked blue in the protein model of coelacanth ATIII (Fig. 3A). Overall a highly conserved RCL is maintained in ATIII during vertebrate evolution of over 450 MY.

3.7. Three pairs of Cys–Cys bridges are conserved throughout the 450 MY of vertebrate evolution

It has been reported that three pairs of disulfide bridges are required in human ATIII in order to bind heparin with high affinity and to inhibit proteinases [1,23]. Majority of vertebrate ATIII proteins investigated have maintained these six essential cysteines, which are involved in formation of C40–C160, C53–C133 and C279–C462 (marked C1, C2, and C3 pairs in Fig. 1S and Fig. 3B) disulfide bridges respectively. Among these six residues, it is evident that four cysteines residues namely C40, C53, C160 and C462 are fully conserved in all species analyzed. The only exceptions are in gorilla and zebra finch where C133F and C279Y mutations are present.

3.8. N-glycosylation sites are maintained with a few variations

N-glycosylations are modifications at asparagine specific sites as designed by NX[ST], where X can be any amino acid except P. There are four N-glycosylation sites in the full-length human ATIII protein at positions N128, N167, N187 and N224 [28]. These sites are conserved in ATIII with some variations (Fig. 1S). For example the N-glycosylation site at N128 (in the helix hC) is maintained in several mammals but not in all mammals. Similarly, the same glycosylation position is not retained in birds, reptiles, coelacanth and Atlantic cod, but is present in all other ray-finned fishes. The N-glycosylation site, N167 is located at the end of helix hD in tetrapods and in coelacanth but not in any ray-finned fishes and in gorilla. The N-glycosylation site at N187 is maintained at the start of the helix hE in majority of vertebrates with exception of guinea pig, gorilla, Atlantic cod, *Fugu* and *Tetraodon*. There is another N-glycosylation site at N192, which is maintained within the helix hE in frog, birds, reptiles, coelacanth and ray-finned fishes but not in zebrafish and any mammal. The N-glycosylation site, N224 is maintained in majority of vertebrates with exception of coelacanth, frog and gorilla. This analysis reveals that at least four N-glycosylation sites are maintained in ATIII throughout the vertebrate evolution.

3.9. Phylogenetic analysis of ATIII reveals loss of intron at the position 262c in tetrapod and coelacanth

In order to examine the ancestry of intron at the position 262c, phylogenetic trees were constructed using maximum likelihood method. Fig. 4A illustrates that vertebrate serpins classified into six groups V1–V6. Group V1 and V5 shared common ancestry and serpins of group V1 share 262c intron with ATIII genes from ray-finned fishes, but not in other vertebrates including coelacanth. Hence, this supports that this intron is lost in tetrapods and coelacanth. This suggests that groupV1a is the ancestor of group V5.

3.10. Gorilla possesses highly diverged ATIII

Gorilla ATIII is highly divergent, which shares 50/64, 37/54 and 30/49 percentage identities/similarities with human, chicken and *Fugu*, respectively (Fig. 1S and Table 3). This is also notable as marked by a black arrow in detailed species wise phylogenetic distribution of ATIII proteins (Fig. 4B). Sequence identities and similarities are also low for gorilla ATIII in comparison with apes (marked in bold in Table 3). We confirmed orthology by comparing synteny with other species (Fig. 1B) and searching this sequence using BLASTP and we received hits as ATIII sequences as the closest homolog in different databases (data not shown).

3.11. Catalog of genetic variation in ATIII using 1092 human genomes

We computed variations in the ATIII gene in 1094 human genomes from 14 different populations as summarized (Table 4) and details are provided in Table S3. There are 1997 variations in total with major components of 1522 SNPs, 236 deletions, 162 insertions, 9 indels, 54 sequence alterations, 8 somatic SNV and 21 substitutions, shared in 15 variant types (Table 4). Out of 1997 ATIII variants, 559 were validated variations. From Table 4, we further examined 63 missense variation to examine what changes these

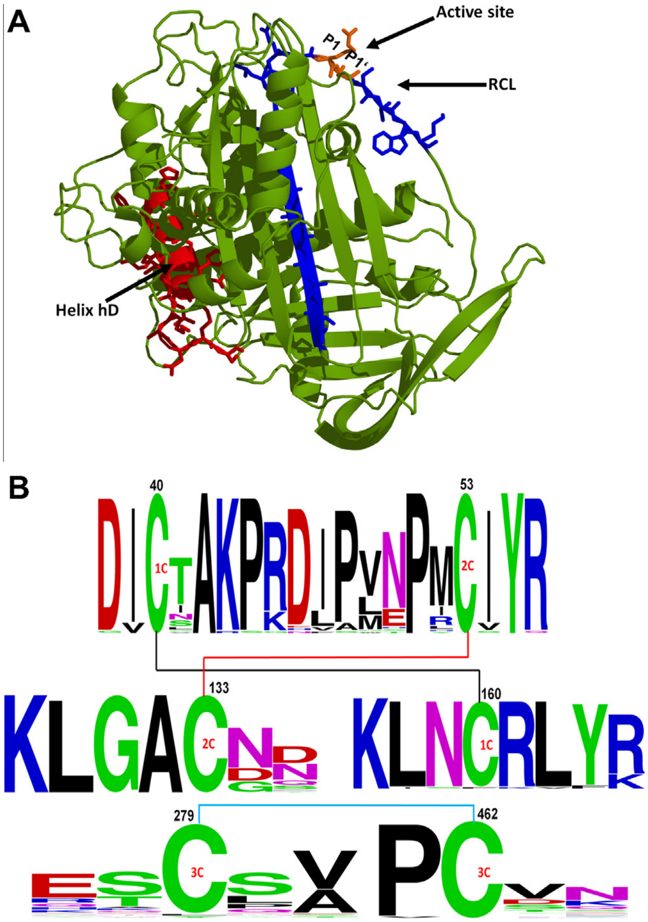

Fig. 3. Protein model and Cys–Cys bridges. (A) Protein model of coelacanth ATIII depicting helix hD (red) and RCL (blue) with active site (orange). (B) Conservation of three Cys–Cys bridges in ATIII proteins. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

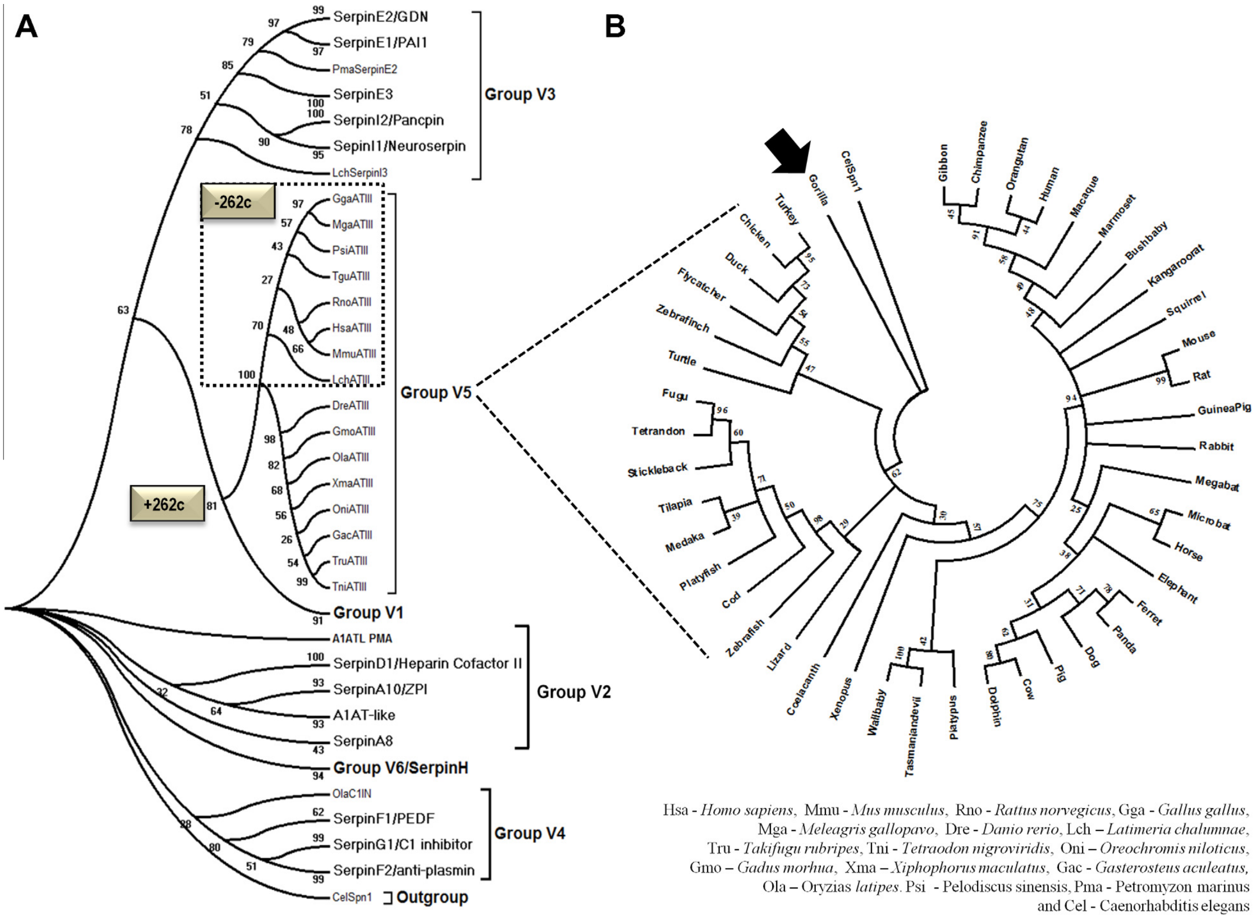

Fig. 4. Phylogenetic history of vertebrate serpins. (A) Evolutionary history of vertebrate group-wise (V1–V6) illustrates loss of 262c intron in ATIII genes of tetrapods and coelacanth where as it is maintained in ATIII genes of ray-finned fishes and in group V1 serpins. (B) Species-wise distribution of ATIII reveals gorilla ATIII is highly diverged. A serpin from *Caenorhabditis elegans*, CelSpn1 (Genbank accession id – NP_503315) was used as outgroup in these two phylogenetic trees. Both trees were prepared by using Maximum Likelihood methods.

causes to ATIII amino acid sequence as well as secondary structural elements (compiled in Table 5). Various structural elements of ATIII have at least one genetic variant, but RCL has maximum number of mutations, as seven distinct mutations are possible. We combined impact of these ATIII variants using SIFT [17] and PolyPhen V2 [18] tools in Table 5. These mutations are marked above protein sequence alignment of ATIII in Fig. 1S.

4. Discussion

This study provides an updated repository of the ATIII gene from 50 vertebrate species (Table 1S) and summarizes major concepts revolving around sequence, structure and phylogeny of ATIII across vertebrate genomes. ATIII gene is highly conserved (Fig. 1S) on the chromosomal locus with a few changes in fishes (Fig. 1B) revealed by sequence and synteny analyses. Though serpins are known for considerable clustering of genes by tandem duplication events leading to the expansion of these genes from fishes to mammals (such as in group V1 serpins [19] and in V2 serpins [24]). Notably, this is not the case for the group V5.

ATIII from ray-finned fishes possesses the intron of 262c (Fig. 1A), which is shared by group V1, and narrowing down to V1a serpins with 7-exons/8-introns gene structure. The intron 262c is shared by group V1 and V5 and is lost after separation of ray-finned fishes from other vertebrates (Fig. 4B).

Although, we could not detect ATIII gene in the lamprey genome, however, we detected three copies of group V1a. Taken together, these data suggests that ATIII gene is present in vertebrates from 450 MY but not in ~500 MY old ancestor lamprey [20]. This also corroborates that ATIII most probably originated from an ancestral serpin with 7-exons/8-introns gene structure in the 50 MY gap period for origin of ray-finned fishes from lampreys. Hence, it is postulated that group V5 is originated by duplication and diversification of group V1a, during very early evolution of the vertebrates, dating 450–500 MYA [20].

It is known that well-defined splicing machinery governs spliceosomal introns in nuclear genomes. Both this splicing machinery and these introns serve major component of eukaryotic genomes. However, the enigma over the origin of these introns remains debatable [25]. There are total 24 conserved introns in vertebrate serpins encompassing group V1–V6 [9] with six additional

Table 3
Comparison of gorilla ATIII with representative species illustrates that gorilla ATIII is highly diverged. Lower percentage identities and similarities values of gorilla ATIII is marked in bold.

Species	1	2	3	4	5	6	7	8	9	10
1. Gorila		**50**	**50**	**50**	**50**	**46**	**45**	**38**	**37**	**30**
2. Human	**64**		99	99	98	87	85	70	66	55
3. Chimpanzee	**64**	99		98	98	86	85	70	66	55
4. Orangutan	**64**	99	99		98	87	85	70	66	55
5. Gibbon	**64**	99	99	99		86	85	70	66	56
6. Mouse	**61**	92	91	92	91		94	70	64	57
7. Rat	**61**	92	92	92	92	97		69	64	57
8. Platypus	**55**	82	82	82	82	83	82		63	57
9. Chicken	**54**	80	80	79	80	78	79	76		57
10. Fugu	**49**	73	73	73	73	73	73	71	71	

Sequence identities

Sequence similarities

Table 4
ATIII variants in 1092 human genomes from 14 ethnicities.

Variant types	Classses							
	SNP	Deletions	Insertions	Indel	Sequence alteration	Somatic SNV	Substitution	All variants
Splice donor variant	2	1	0	0	5	0	0	8
Splice acceptor variant	1	0	0	0	7	0	0	8
Stop gained	4	0	0	0	0	0	0	4
Frameshift variant	0	2	0	0	0	0	1	3
Inframe deletion	0	1	0	0	0	0	0	1
Missense variant	60	0	0	0	0	0	3	63
Splice region variant	12	0	0	0	8	0	2	22
Synonymous variant	22	0	0	0	0	0	1	23
Coding sequence variant	129	42	22	2	4	0	4	200
5 prime UTR variant	2	0	0	0	0	0	0	2
3 prime UTR variant	2	0	0	0	0	0	0	2
Intron variant	307	61	16	0	4	0	2	374
Non coding exon variant	168	33	14	3	2	2	1	224
NC transcript variant	309	32	39	3	10	2	2	398
Upstream gene variant	211	16	24	1	1	0	5	258
Downstream gene variant	293	48	47	0	13	4	0	407
Total	1522	236	162	9	54	8	21	1997

Table 5
List of 63 missense ATIII variants in human with predictions of SIFT [17] and PolyPhen V2 [18] tools.

ATIII mutation	ATIII structural element	Variant ID	Chr:bp	Alleles	Global MAF	Class	Source	Validation	SIFT	PolyPhen
T10N	Signal peptide	rs61736655	1:173886369	G/T	0.001 (T)	SNP	dbSNP	Freq,	11deleterious (0.01)	7benign (0.006)
S20F	Signal peptide	TMP_ESP_1_173884040	1:173884040	G/A	–	SNP	ESP	–	41deleterious (0.04)	4benign (0.003)
C29R	Signal peptide	rs147918976	1:173884014	A/G	–	SNP	dbSNP	–	41deleterious (0.04)	2benign (0.001)
V30E	Signal peptide	rs2227624	1:173884010	A/T	0.002 (T)	SNP	dbSNP	Freq,	1deleterious (0)	234benign (0.233)
H33Q	Signal peptide	rs147676453	1:173884000	G/T/A	–	SNP	dbSNP	Freq,	241tolerated (0.24)	3benign (0.002)
D38N	N-terminal region after signal peptide	rs145771113	1:173883987	C/T	–	SNP	dbSNP	–	1deleterious (0)	574possibly damaging (0.573)
I39N	N-terminal region after signal peptide	rs121909558	1:173883983	A/T	–	SNP	dbSNP	–	1deleterious (0)	971probably damaging (0.97)
R56C	N-terminal region after signal peptide	rs28929469	1:173883933	G/A	–	SNP	dbSNP	Cluster,	1deleterious (0)	1000probably damaging (0.999)
E69V	N-terminal region after signal peptide	TMP_ESP_1_173883893	1:173883893	T/A	–	SNP	ESP	–	11deleterious (0.01)	10benign (0.009)
P73L	N-terminal region after signal peptide	rs121909551	1:173883881	G/A	0.001 (A)	SNP	dbSNP	Freq,	1deleterious (0)	999probably damaging (0.998)
R78W	helix hA, heparin binding basic residue	TMP_ESP_1_173883867	1:173883867	G/A	–	SNP	ESP	–	21deleterious (0.02)	568possibly damaging (0.567)
R79H	helix hA, heparin binding basic residue	rs121909552	1:173883863	C/T	0.001 (T)	SNP	dbSNP	–	1deleterious (0)	279benign (0.278)
R79C	helix hA, heparin binding basic residue	rs121909547	1:173883864	G/A	–	SNP	dbSNP	–	1deleterious (0)	989probably damaging (0.988)
R79S	helix hA, heparin binding basic residue	rs121909553	1:173883864	G/T	–	SNP	dbSNP	–	1deleterious (0)	958probably damaging (0.957)
R89C	helix hA	rs147266200	1:173883834	G/A	0.001 (A)	SNP	dbSNP	–	1deleterious (0)	989probably damaging (0.988)
D100G	helix hA	TMP_ESP_1_173883800	1:173883800	T/C	–	SNP	ESP	–	61tolerated (0.06)	276benign (0.275)
S101C	helix hA	rs199895690	1:173883797	G/C	0.001 (C)	SNP	dbSNP	–	1deleterious (0)	754possibly damaging (0.753)
D104V	hA-turn-s6B	rs200118419	1:173883788	T/A	–	SNP	dbSNP	–	51deleterious (0.05)	95benign (0.094)
T117M	helix hB	rs139392083	1:173883749	G/A	–	SNP	dbSNP	–	1deleterious (0)	999probably damaging

(continued on next page)

Table 5 (*continued*)

ATIII mutation	ATIII structural element	Variant ID	Chr:bp	Alleles	Global MAF	Class	Source	Validation	SIFT	PolyPhen
										(0.998)
M121I	helix hB	TMP_ESP_1_173883736	1:173883736	C/T	–	SNP	ESP	–	1deleterious (0)	861possibly damaging (0.86)
C127R	hB-turn-hC, conserved 2C	rs121909573	1:173883720	A/G	–	SNP	dbSNP	–	851tolerated (0.85)	1000probably damaging (0.999)
L131F	helix hC	rs121909567	1:173883708	G/A	–	SNP	dbSNP	–	41deleterious (0.04)	1000probably damaging (0.999)
T147A	hC-hD	rs2227606	1:173881122	T/C	0.002 (C)	SNP	dbSNP	Freq,	111tolerated (0.11)	207benign (0.206)
S148P	hC-hD	rs121909569	1:173881119	A/G	–	SNP	dbSNP	–	21deleterious (0.02)	998probably damaging (0.997)
R161Q	helix hD, heparin binding basic residue	rs121909563	1:173881079	C/T	–	SNP	dbSNP	–	141tolerated (0.14)	957probably damaging (0.956)
N167T	helix hD, N-glycosilation site N167	rs121909570	1:173881061	T/G	–	SNP	dbSNP	–	481tolerated (0.48)	402benign (0.401)
R177C	s2A	rs143521873	1:173881032	G/A	–	SNP	dbSNP	Freq,	1deleterious (0)	1001probably damaging (1)
N219D	helix hF	rs121909571	1:173879999	T/C	–	SNP	dbSNP	–	1deleterious (0)	995probably damaging (0.994)
S223P	helix hF	rs121909572	1:173879987	A/G	–	SNP	dbSNP	–	181tolerated (0.18)	873possibly damaging (0.872)
N224D	helix hF, N-glycosilation site N224	rs146733468	1:173879984	T/C	–	SNP	dbSNP	–	141tolerated (0.14)	42benign (0.041)
S236L	hF-turn-s3A	TMP_ESP_1_173879947	1:173879947	G/A	–	SNP	ESP	–	201tolerated (0.2)	3benign (0.002)
N240S	hF-turn-s3A	rs200861147	1:173879935	T/C	0.001 (C)	SNP	dbSNP	–	161tolerated (0.16)	11benign (0.01)
T250I	s3A	rs144084678	1:173879905	G/A	–	SNP	dbSNP	–	1deleterious (0)	998probably damaging (0.997)
K254E	s3A	COSM131070	1:173879894	T/C	–	somatic_SNV	COSMIC	–	81tolerated (0.08)	695possibly damaging (0.694)
Q286H	s3C	rs139463995	1:173878985	C/G	–	SNP	dbSNP	–	1deleterious (0)	997probably damaging (0.996)
R291H	s3C-turn-s1C	TMP_ESP_1_173878971	1:173878971	C/T	–	SNP	ESP	–	231tolerated (0.23)	10benign (0.009)
V295M	s1C	rs201381904	1:173878960	C/T	0.001 (T)	SNP	dbSNP	–	11deleterious (0.01)	817possibly damaging (0.816)
A296P	s1C to s2B	TMP_ESP_1_173878957	1:173878957	C/G	–	SNP	ESP	–	351tolerated (0.35)	2benign (0.001)
R356H	s6A	TMP_ESP_1_173878776	1:173878776	C/T	–	SNP	ESP	–	21deleterious (0.02)	49benign (0.048)
R356C	s6A	TMP_ESP_1_173878777	1:173878777	G/A	–	SNP	ESP	–	31deleterious (0.03)	103benign (0.102)
D374N	helix hI1	COSM208077	1:173878723	C/T	–	somatic_SNV	COSMIC	–	51deleterious (0.05)	963probably damaging (0.962)
L375M	helix hI1	rs149006854	1:173878720	G/T	–	SNP	dbSNP	–	11deleterious (0.01)	987probably damaging (0.986)
S381P	hI1-turn-s5A	rs121909565	1:173878702	A/G	–	SNP	dbSNP	–	1deleterious (0)	508possibly damaging (0.507)
R391Q	hI1-turn-s5A	rs201541724	1:173876634	C/T	–	SNP	dbSNP	–	341tolerated (0.34)	3benign (0.002)
D398E	s5A	TMP_ESP_1_173876612	1:173876612	A/T	–	SNP	ESP	–	711tolerated(0.71)	8benign (0.007)
A403S	s5A	rs138743710	1:173876599	C/A	–	SNP	dbSNP	–	491tolerated (0.49)	450possibly damaging (0.449)
A414T	RCL, P12 (s4A)	rs121909557	1:173873182	C/T	–	SNP	dbSNP	–	11deleterious (0.01)	999probably damaging (0.998)
A416P	RCL, P10 (s4A)	rs28930978	1:173873176	C/G	–	SNP	dbSNP	–	1deleterious (0)	1000probably damaging

Table 5 (*continued*)

ATIII mutation	ATIII structural element	Variant ID	Chr:bp	Alleles	Global MAF	Class	Source	Validation	SIFT	PolyPhen
										(0.999)
A416S	RCL, P10 (s4A)	rs121909548	1:173873176	C/G/A	0.001 (A)	SNP	dbSNP	Freq,	11deleterious (0.01)	999probably damaging (0.998)
A416P	RCL, P10 (s4A)	rs121909548	1:173873176	C/G/A	0.001 (A)	SNP	dbSNP	Freq,	1deleterious (0)	1000probably damaging (0.999)
A419V	RCL, P7 (s4A)	rs121909568	1:173873166	G/A	–	SNP	dbSNP	–	111tolerated (0.11)	56benign (0.055)
V420L	RCL, P6 (s4A)	TMP_ESP_1_173873164	1:173873164	C/G	–	SNP	ESP	–	51deleterious (0.05)	82benign (0.081)
G424D	RCL, P2 (s4A)	rs121909566	1:173873151	C/T	–	SNP	dbSNP	–	81tolerated (0.08)	999probably damaging(0.998)
R425P	RCL, P1 (s4A)	rs121909549	1:173873148	C/G	–	SNP	dbSNP	–	121tolerated (0.12)	924probably damaging (0.923)
R425H	RCL, P1 (s4A)	rs121909556	1:173873148	C/T	–	SNP	dbSNP	–	171tolerated (0.17)	996probably damaging (0.995)
R425C	RCL, P1 (s4A)	rs121909554	1:173873149	G/A	0.001 (A)	SNP	dbSNP	–	211tolerated (0.21)	1000probably damaging (0.999)
S426L	RCL, P1′ (s4A)	rs121909550	1:173873145	G/A	–	SNP	dbSNP	–	41deleterious (0.04)	995probably damaging (0.994)
S426L	RCL, P1′ (s4A)	COSM76381	1:173873145	G/A	–	somatic_SNV	COSMIC	–	41deleterious (0.04)	995probably damaging (0.994)
A436T	s1C-turn-s4B	rs121909546	1:173873116	C/T	–	SNP	dbSNP	–	11deleterious (0.01)	994probably damaging (0.993)
P439L	s1C-turn-s4B	rs121909555	1:173873106	G/A	–	SNP	dbSNP	–	1deleterious (0)	1001probably damaging (1)
V447D	s4B-turn-s5B	TMP_ESP_1_173873082	1:173873082	A/T	–	SNP	ESP	–	41deleterious (0.04)	628possibly damaging (0.627)
P448S	s4B-turn-s5B	TMP_ESP_1_173873080	1:173873080	G/A	–	SNP	ESP	–	281tolerated (0.28)	5benign (0.004)
P461L	C-terminal end	rs121909564	1:173873040	G/A	–	SNP	dbSNP	–	1deleterious (0)	1000probably damaging (0.999)

introns (four in group V2 and two in group V6) that were gained in selected ray finned fishes among serpin genes [26]. But, this is the only of intron loss in serpin superfamily in vertebrates. Genome compaction was attributed with several examples of intron creations in selected ray-finned fishes whose genome underwent compaction events in the serpin superfamily [26] and in the GPCR superfamily [27]. In contrast, there are two hypotheses that explains the mechanism of intron losses: (i) by deletion in genomes and (ii) homologous recombination between the genomic copy of a gene and the cDNA produced by the reverse transcription of its mature mRNA or partially spliced pre-mRNA [25]. Such perfect genomic region deletion that leads only loss of 262c intron is highly unlikely. Hence, this precise deletion of intron 262c can be best explained by homologous recombination based loss mechanism.

Additionally, the intron at position 339c of the ATIII gene is found in several serpins from an array of evolutionary distant organisms, such as *Caenorhabditis elegans*, *Brugia malayi*, lancelets, and *Ciona intestinalis* [5]. But current genomic datasets are not sufficient to unravel evolutionary history of the 339c intron. However, it will be possible after millions of high throughput genome sequences of eukaryotes are available in near future.

Vertebrate ATIII possesses several conserved signatures (Fig. 2). In addition to conserved six cysteine residues (Fig. 3A), there are two cysteine residues that are conserved in ATIII proteins at positions C29 and C32 in all mammals and the later is also found in zebrafish. Notably, a non-sense mutation at the C29 position leads into a stop codon (TGC → TGA) causing recurrent venous thrombosis [28]. The ATIII ortholog in gorilla is highly diverged (Fig. 4B and Fig. 1S) but maintained on the same genomic fragment (Fig. 1B). Additionally, 63 mutational hotspots of ATIII were identified from 1092 human genomes by combining missense ATIII variants (Table 5 and Fig. 1S). These key understandings may have medical importance in the context of human pathology related to blood clotting. In conclusion, ATIII gene is revisited from sequence-structural, phylogenetic and variants perspective in the era of desktop genomics.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bbrc.2013.09.134.

References

[1] S.T. Olson, B. Richard, G. Izaguirre, et al., Molecular mechanisms of antithrombin-heparin regulation of blood clotting proteinases. A paradigm for understanding proteinase regulation by serpin family protein proteinase inhibitors, Biochimie 92 (2010) 1587–1596.

[2] G.A. Silverman, P.I. Bird, R.W. Carrell, et al., The serpins are an expanding superfamily of structurally similar but functionally diverse proteins. Evolution, mechanism of inhibition, novel functions, and a revised nomenclature, J. Biol. Chem. 276 (2001) 33293–33296.

[3] G.A. Silverman, D.A. Lomas (Eds.), Molecular and Cellular Aspects of the Serpinopathies and Disorders in Serpin Activity, World Scientific Pub., 2007.

[4] M. Backovic, P.G.W. Gettins, Insight into residues critical for antithrombin function from analysis of an expanded database of sequences that includes frog, turtle, and ostrich antithrombins, J. Proteome Res. 1 (2002) 367–373.

[5] A. Kumar, Phylogenomics of Vertebrate Serpins, University of Bielefeld, Bielefeld, NRW, Germany, 2010.

[6] S.F. Altschul, T.L. Madden, A.A. Schaffer, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.

[7] P. Flicek, I. Ahmed, M.R. Amode, et al., Ensembl 2013, Nucleic Acids Res. 41 (2013) D48–D55.

[8] M. Stanke, B. Morgenstern, Augustus: a web server for gene prediction in eukaryotes that allows user-defined constraints, Nucleic Acids Res. 33 (2005) W465–W467.

[9] A. Kumar, H. Ragg, Ancestry and evolution of a secretory pathway serpin, BMC Evol. Biol. 8 (2008) 250.

[10] R.C. Edgar, Muscle: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics 5 (2004) 113.

[11] K.B. Nicholas, Nicholas H.B. Jr., D.W.I. and Deerfield, GeneDoc: Analysis and Visualization of Genetic Variation, EMBNEW.NEWS 4 (1997) 14.

[12] G.E. Crooks, G. Hon, J.M. Chandonia, et al., WebLogo: a sequence logo generator, Genome Res. 14 (2004) 1188–1190.

[13] K. Tamura, D. Peterson, N. Peterson, et al., MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, Mol. Biol. Evol. 28 (2011) 2731–2739.

[14] A. Roy, A. Kucukural, Y. Zhang, I-TASSER: a unified platform for automated protein structure and function prediction, Nat. Protoc. 5 (2010) 725–738.

[15] Schrodinger LLC., The PyMOL molecular graphics system, version 1.3r1, 2010.

[16] G.R. Abecasis, A. Auton, L.D. Brooks, et al., An integrated map of genetic variation from 1092 human genomes, Nature 491 (2012) 56–65.

[17] P.C. Ng, S. Henikoff, SIFT: predicting amino acid changes that affect protein function, Nucleic Acids Res. 31 (2003) 3812–3814.

[18] I.A. Adzhubei, S. Schmidt, L. Peshkin, et al., A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248–249.

[19] C. Benarafa, E. Remold-O'Donnell, The ovalbumin serpins revisited: perspective from the chicken genome of clade B serpin evolution in vertebrates, Proc. Natl. Acad. Sci. USA 102 (2005) 11367–11372.

[20] J.J. Smith, S. Kuraku, C. Holt, Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution, Nat. Genet. 45 (2013) 415–421. 421e411-412.

[21] J.A. Irving, R.N. Pike, A.M. Lesk, et al., Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function, Genome Res. 10 (2000) 1845–1864.

[22] R. Skinner, J.P. Abrahams, J.C. Whisstock, et al., The 2.6 a structure of antithrombin indicates a conformational change at the heparin binding site, J. Mol. Biol. 266 (1997) 601–609.

[23] Z.R. Zhou, D.L. Smith, Location of disulfide bonds in antithrombin III, Biomed. Environ. Mass Spectrom. 19 (1990) 782–786.

[24] S. Forsyth, A. Horvath, P. Coughlin, A review and comparison of the murine alpha1-antitrypsin and alpha1-antichymotrypsin multigene clusters with the human clade A serpins, Genomics 81 (2003) 336–345.

[25] S.W. Roy, W. Gilbert, The evolution of spliceosomal introns: patterns, puzzles and progress, Nat. Rev. Genet. 7 (2006) 211–221.

[26] H. Ragg, A. Kumar, K. Koster, et al., Multiple gains of spliceosomal introns in a superfamily of vertebrate protease inhibitor genes, BMC Evol. Biol. 9 (2009) 208.

[27] A. Kumar, A. Bhandari, R. Sinha, et al., Spliceosomal intron insertions in genome compacted ray-finned fishes as evident from phylogeny of MC receptors, also supported by a few other GPCRs, PLoS One 6 (2011) e22046.

[28] S. Saleun, P. De Moerloose, A. Bura, et al., A novel nonsense mutation in the antithrombin III gene (Cys-4→stop) causing recurrent venous thrombosis, Blood Coagul. Fibrinolysis 7 (1996) 578–579.