Biochemical and Biophysical Research Communications 446 (2014) 504–518

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

Revising angiotensinogen from phylogenetic and genetic variants perspectives

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

[a] Department of Genetics & Molecular Biology in Botany, Institute of Botany, Christian-Albrechts-University at Kiel, Kiel, Germany
[b] Master Program Agrigenomics, Christian-Albrechts-University at Kiel, Kiel, Germany
[c] Department of Zoophysiology II, Christian-Albrechts-University at Kiel, Kiel, Germany

ARTICLE INFO

Article history:
Received 7 February 2014
Available online 12 March 2014

Keywords:
Angiotensinogen
Serpin A8
Group V2
Synteny
Phylogenetic analysis
AGT variants

ABSTRACT

Angiotensinogen (AGT) belongs to the serpin superfamily. It acts as the unique substrate of all angiotensin peptides, which generates a spectrum of angiotensin peptides in the renin-angiotensin system and regulates hypertension. This serpin belongs to the multiple member group V2 of the intron encoded vertebrate serpin classification. Despite huge advancements in the understanding of angiotensinogen based on biochemical properties and its roles in the RAS, phylogenetic history of AGT remains forgotten. To date, there is no comprehensive study illustrating the phylogenetic history of AGT. Herein, we investigated phylogenetic traits of AGT gene across vertebrates. Gene structures of AGT gene from selected ray-finned fishes varied in exon I and II with insertions of two novel introns in the core domain for ray-finned fishes at the position 77c and 233c. We that found AGT loci is conserved from lampreys to human and estimated to be older than 500 MY. By comparing AGT protein in 57 vertebrate genomes, we illustrated that the reactive center loop (RCL) of AGT protein became from inhibitory (in lampreys, GTEAKAETVVGIMPI†SMPPT) to non-inhibitory (in human, EREPTESTQQLNKPE†VLEVT) during period of 500 MY. We identified 690 AGT variants by analysis of 1092 human genomes with top three variation classes belongs to SNPs (89.7%), somatic SNVs (5.2%) and deletion (2.9%). There are 32 key residues out of 121 missense variants, which are deleterious for AGT protein, computed by combination of SIFT and PolyPhen V2 methods. These results may have clinical implications for understanding hypertension.

1. Introduction

The renin–angiotensin system (RAS) is an enzyme-linked hormonal cascade that plays an important role in body fluid and cardiovascular regulation. The system is initiated by the action of renin on the precursor protein, angiotensinogen (AGT), which yields the active hormone, angiotensin II. AGT belongs to serine protease inhibitor (serpin) is a single chain glycoprotein with a size of 49.761 kDa, which is the only known precursor of all the peptides generated in the renin–angiotensin system. This gene belongs to clade A in the clade based serpin classification and also known as Serpin A8 [1] and the group V2 in the intron encoded group-wise vertebrate serpin classification system [2]. The protein structure of serpin domain consists of three β-sheets, sA-sC and 8–9 α-helices, hA–hI [3]. The hallmark of serpin biology 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′ [3]. Some of the serpins have mutations in the RCL region leading into non-inhibitory serpins with specialized functions other than inhibition and human AGT belongs to this category of serpins. R.F. Doolittle identified the AGT gene for the first time in 1983 [4]. In last three decades, AGT has been extensively characterized by biochemical and biophysical methods to demonstrate its roles in the RAS and ultimately controlling blood pressure. However, it lacks a comprehensive molecular phylogenetic analysis, primarily due to fact that notorious problems are countered during the reconstruction of phylogenetic relationships among animal serpins, as several paralogs are found in various animals, particularly for groups V1 and V2 [5]. This suggests that there is a requirement of an investigation on molecular phylogenetic perspectives. Herein, our data disclosed that the detailed molecular phylogeny of AGT genes by combining sequence, genetic variants, gene structures and genomic organization from 57 vertebrates. Furthermore, we have identified 690 genetic variants of human AGT with 121 missense mutations from which 32 are deleterious for AGT protein, which serves an excellent platform for understanding hypertension regulations.

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

http://dx.doi.org/10.1016/j.bbrc.2014.02.139
0006-291X/

2. Materials and methods

2.1. Sequence collection

We extracted genomic DNA and protein sequences from different vertebrate genomes via Ensembl release 73 (September 2013) [6] using BLAST suite for AGT are provided in Table S1.

2.2. Predicting gene structures and mapping intron positions

We predicted gene structures using AUGUSTUS suite [7] and we combined with gene structure prediction within the Ensembl [6], which ensured accuracy. We used mature human α_1-antitrypsin as the standard sequence for intron position mapping and numbering of intron positions, followed by suffixes a–c for their location as reported previously [5].

2.3. Mining genetic variants of human AGT

We computed genetic variants of human AGT using 1092 human genomes from 14 different populations available in 1000 genomes project [8]. Sorting Intolerant From Tolerant (SIFT) is a software tool, which predicts whether an amino acid substitution affects protein function and it helps in prioritizing substitutions for further study [9]. The SIFT value $\leqslant$0.05 indicates the deleterious effects of missense variants on protein function [9]. 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 [10]. PolyPhen-V2 score (close to 1) indicates the damaging effects of missense variants on protein function. We used these two methods to predict the impact of these variants on function of AGT.

2.4. Synteny analysis

We scanned chromosomal locus for AGT gene for each species using Ensembl genome browser [6] and mapviewer from the NCBI (website: http://www.ncbi.nlm.nih.gov/mapview/). We constructed synteny maps from fishes to mammalian genomes.

2.5. Sequence and structural analysis

We created protein alignment of AGT using the MUSCLE [11] and edited and visualized in GENEDOC [12] as shown in Fig. S1. We constructed sequence logos of conserved motifs in AGT proteins by Weblogo 3.3 [13].

2.6. Phylogenetic analysis

We constructed two phylogenetic trees aided by Bayesian (2 runs, until average standard deviation of split frequencies was lower than 0.0098, 25% burn-in-period) and Neighbor-Joining methods (1000 bootstraps) using Mrbayes 3.2.1 [14] and MEGA 5.2.2 [15] as following vertebrate serpins (259 serpins) and AGT proteins (57 sequences), respectively. These two trees are based on WAG [5 categories (+*G*, parameter = 4.6121)] and JTT [5 categories (+G, parameter = 1.3836)] protein substitution models, respectively, as their best models were computed in MEGA 5.2.2 [15].

2.7. Protein modelling of AGT protein from Petromyzon marinus

We created structural model of AGT protein from *Petromyzon marinus* using the I-TASSER [16] and we visualized the resulting model using YASARA [17].

3. Results

3.1. Variations in gene structures of AGT in fishes

Human AGT possesses four exons I–IV separated by three introns at the canonical positions 192a, 282b, and 331c in the conserved domain (Fig. 1A). These three introns are maintained in AGT gene from all investigated genomes, which assured membership in the group V2. However, there are changes observed in ray-finned fishes. The canonical exon I is divided into two pieces in ray-finned fishes (except in zebrafish, cave fish and spotted gar) as exons Ia and IIb by an intron at the position 77c with its length ranged from 75 bp (in Fugu) to 267 bp (in Atlantic cod).

Interestingly, exon II is also invaded by a novel intron at the position 233c in selected ray-finned fishes, forming two exons IIa and IIb with sizes 125 bp and 149 bp, respectively. This novel intron has length ranged from 80 bp (in Fugu and medaka) to 137 bp (in Atlantic cod). This insertion of intron within exon II is not observed in lamprey, cave fish, spotted gar and zebrafish. This aspect corroborates that it is a specific feature of selected ray-finned fishes evolved after separation from zebrafish. Exon III is conserved in all vertebrates with size 151–157 bp. Exon IV is ranged from 138 bp (in lamprey) to 192 bp (in chicken). The intron at the position 192a is ranged from 83 bp (in stickleback) to 7002 bp (in Chinese softshell turtle), where as the intron at the position 282b is ranged from 79 bp (in zebrafish) to 2387 bp (in *Petromyzon*) and the intron at the position 331c is ranged from 77 bp (in *Fugu*) to 1693 bp (in zebrafish). These two novel introns are localized in the helix C and the sheet s1B (Fig. S1), respectively. During this study, complete gene structures of AGT gene from spotted gar and *Tetraodon* was not possible to construct due to incomplete assembly. This problem also exists for AGT from lamprey in the Ensembl (December 2013). But, when we used previous version, PMAR3.0 assembly (accession id – GENSCAN00000089208), we are able to construct full-length gene. Noteworthy, AGT genes from *Petromyzon* possess all introns, which are > 1kb. In the nutshell, we report change of the gene structure patterns in selected ray-finned fishes in comparison to that of lampreys, cave fish, spotted gar, zebrafish and tetrapods.

3.2. AGT is conserved on the same genomic fragment from ~500 MY during vertebrate evolution

AGT gene is localized in the human chromosome 1, flanking a heptad of genes, ABCB10, TAF5L, URB2, GALNT2, PGBD5 and COG2 (details in Table S2) on the one side, whereas a tetrad of genes (CAPN9, ARV1, FAM89A, and TRIM67) is present (Fig. 1B). This micro-synteny is maintained in several other mammals including mouse (chromosome 8), rat (chromosome 19) horse (chromosome 1), opossum (chromosome 2) and sheep (chromosome 25). This genomic fragment is also conserved in several birds such as chicken (chromosome 3), duck (scaffold KB743153.1), flycatcher (scaffold JH603366.1), turkey (chromosome 2) and zebra finch (chromosome 3). Reptiles also possess AGT gene flanking same marker genes as of mammals and birds as shown for anole lizard (chromosome 1) and Chinese softshell turtle (scaffold JH208515.1). Amphibians also have these loci but only a heptad of genes are conserved on the one side and other side has genes which are not localized flanking AGT in any other vertebrates analyzed and hence considered as variable region, left blank in Fig. 1B. All fishes have single copy of AGT gene on same syntenic segment as of tetrapods but this segment has some variations. AGT gene is only flanked by COG2 gene on the scaffold JH126749.1 in coelacanth, living fossil lobe-finned fish. Ray-finned fishes with compact genomes possess AGT gene flanking a conserved heptad of genes as

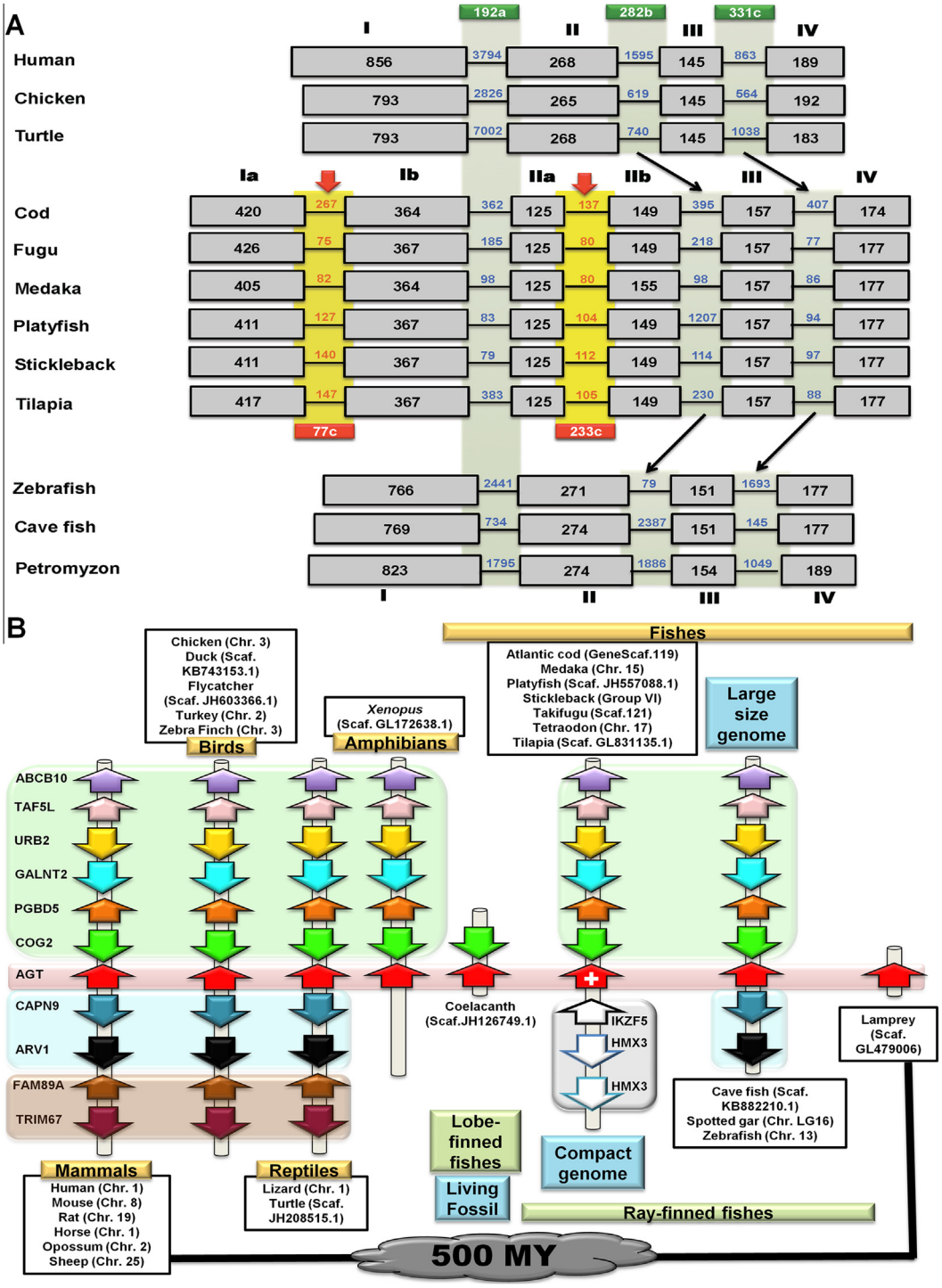
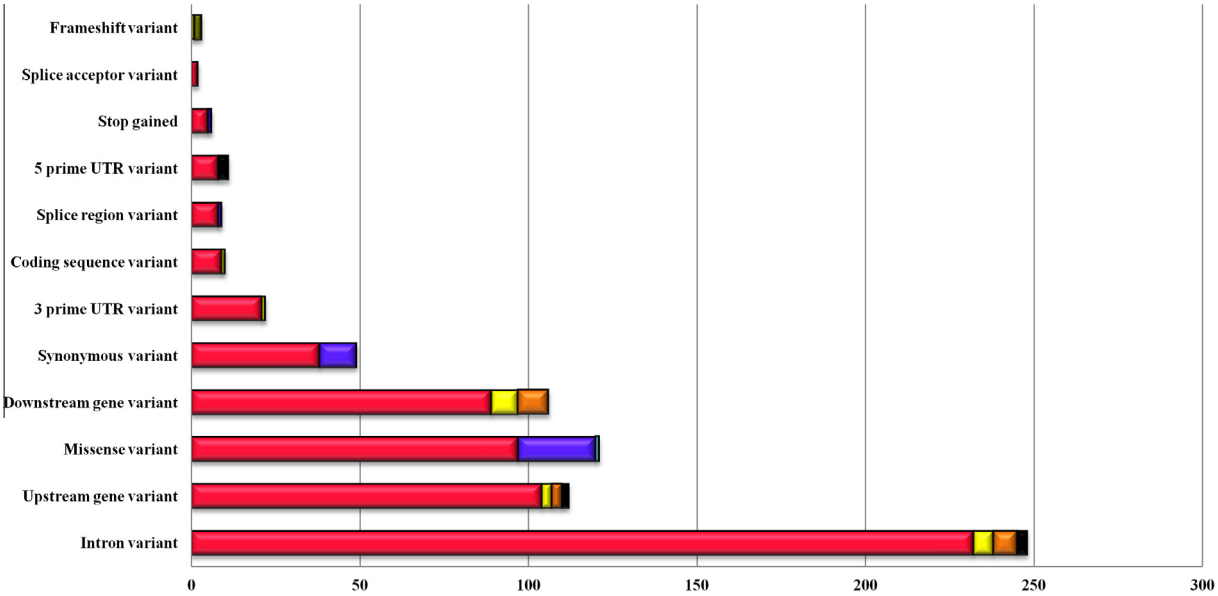


Fig. 1. Gene and Genomic organizations of angiotensinogen. (A) Variation in gene structure of AGT among vertebrates illustrates creation of two novel introns at the positions 77c and 233c (marked in red) in selected ray-finned fishes. (B) Variation in syntenic organization of AGT loci in vertebrates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in tetrapods on the one side. However, there is variation of other other side with a conserved triad of transcription factor genes namely IKZF5 HMX2 and HMX3 (see Table S2) as shown for following fishes – Atlantic cod (Genescaffold 119), medaka (chromosome 15), platyfish (scaffold JH557088.1), stickleback (group VI), *Takifugu* (scaffold 121), *Tetraodon* (chromosome 17) and tilapia (scaffold GL831135.1). AGT gene from these genomic fragments has two novel intron insertion events and is marked by + sign in Fig. 1B. In contrast, ray-finned fishes with large genomes have conserved synteny as of mammals and reptiles. However, instead of a tetrad of genes, only a dyad of genes (CAPN9-ARV1) is conserved as illustrated for cave fish (scaffold KB882210.1), spotted gar (chromosome LG16) and zebrafish (chromosome 13). Details of flanking genes are supplied in Table S2. Ancient fish - lamprey possesses only AGT gene at the beginning of small scaffold GL479006 and hence marker genes are not possible to detect. This suggests that



	Intron variant	Upstream gene variant	Missense variant	Downstream gene variant	Synonymous variant	3 prime UTR variant	Coding sequence variant	Splice region variant	5 prime UTR variant	Stop gained	Splice acceptor variant	Frameshift variant
SNP	232	104	97	89	38	21	9	8	8	5	2	0
Deletion	6	3	0	8	0	1	1	0	0	0	0	1
Insertion	7	3	0	9	0	0	0	0	0	0	0	0
Seq. alteration	3	2	0	0	0	0	0	0	3	0	0	0
Somatic Deletion	0	0	0	0	0	0	0	0	0	0	0	2
Somatic SNV	0	0	23	0	11	0	0	1	0	1	0	0
Substitution	0	0	1	0	0	0	0	0	0	0	0	0

Fig. 2. Summary of genetic variants of AGT reveals majority of these variants are SNPs.

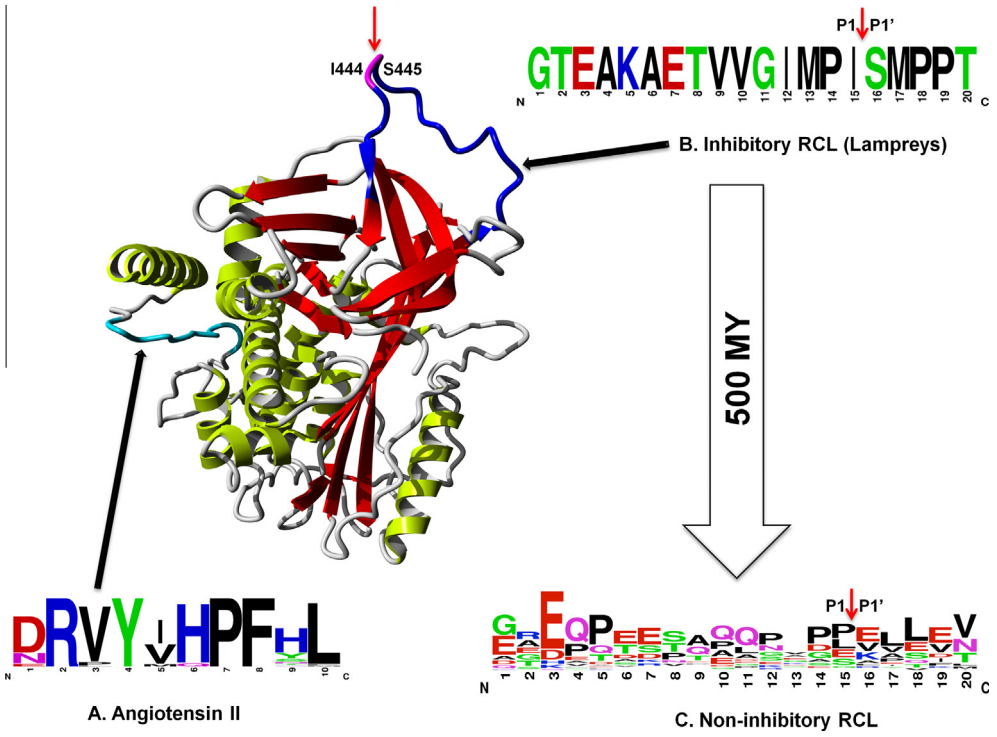


Fig. 3. Homology model of AGT protein from *Petromyzon* reveals angiotensin II (A) and inhibitory RCL (B), which turned into non-inhibitory RCL in tetrapods (C) during 500 MY of vertebrate evolution.

Table 1
Summary of missense AGT variants in 1092 human genomes from 14 ethnicities. There are 32 deleterious mutational sites, marked by bold letters.

Variations	Structural elements	ID	Chromosomal location	Alleles	gmaf	Class	Source	Status	SIFT (Score)	Polyphen V2 (Score)
R2Q	**Signal peptide**	**rs61731498**	**1:2308465921:230846592**	**C/T**	–	**SNP**	**dbSNP**	**Frequency, ESP**	**1 Deleterious (0)**	**991 Probably damaging (0.99)**
R4P	**Signal peptide**	**rs61762541**	**1:2308465861:230846586**	**C/G**	–	**SNP**	**dbSNP**	**Multiple_observations, frequency, ESP**	**1 Deleterious (0)**	**1000 Probably damaging (0.999)**
R4L	**Signal peptide**	**COSM1244653**	**1:2308465861:230846586**	**C/A**	–	**Somatic_SNV**	**COSMIC**	–	**1 Deleterious (0)**	**1001 Probably damaging (1)**
M10	**Signal peptide**	**rs199476082**	**1:2308465691:230846569**	**T/A**	–	**SNP**	**dbSNP**	–	**21 Deleterious (0.02)**	**884 Possibly damaging (0.883)**
G14S	Signal peptide	TMP_ESP_1_230846557	1:2308465571:230846557	C/T	–	SNP	ESP	ESP	641 Tolerated (0.64)	2 Benign (0.001)
T20P	**Signal peptide**	**rs13306155**	**1:2308465391:230846539**	**T/G**	–	**SNP**	**dbSNP**	**Frequency**	**31 Deleterious (0.03)**	**816 Possibly damaging (0.815)**
G29D	Signal peptide	rs143545998	1:2308465111:230846511	C/T	–	SNP	dbSNP	Frequency, ESP	281 Tolerated (0.28)	13 Benign (0.012)
A31G	**Signal peptide**	**rs149973083**	**1:2308465051:230846505**	**G/C**	–	**SNP**	**dbSNP**	**ESP**	**41 Deleterious (0.04)**	**480 Possibly damaging (0.479)**
L43F	**N-terminal end after signal peptide**	**rs41271499**	**1:2308464701:230846470**	**G/A**	–	**SNP**	**dbSNP**	**Multiple_observations, frequency, ESP**	**1 Deleterious (0)**	**987 Probably damaging (0.986)**
V44I	**N-terminal end after signal peptide**	**rs146773738**	**1:2308464671:230846467**	**C/T**	–	**SNP**	**dbSNP**	**Frequency, ESP**	**11 Deleterious (0.01)**	**455 Possibly damaging (0.454)**
S49N	N-terminal end after signal peptide	TMP_ESP_1_230846451	1:2308464511:230846451	C/T	–	SNP	ESP	ESP	81 Tolerated (0.08)	454 Possibly damaging (0.453)
T50S	**N-terminal end after signal peptide**	**rs201777691**	**1:2308464491:230846449**	**T/A**	**0.001 (A)**	**SNP**	**dbSNP**	–	**21 Deleterious (0.02)**	**34 Benign (0.033)**
C51R	N-terminal end after signal peptide	rs61731497	1:2308464461:230846446	A/G	0.001 (G)	SNP	dbSNP	Multiple_observations, frequency, ESP	781 Tolerated (0.78)	3 Benign (0.002)
K61N	N-terminal end after signal peptide	COSM1501176	1:2308464141:230846414	C/A	–	Somatic_SNV	COSMIC	–	291 Tolerated (0.29)	136 Benign (0.135)
P62T	N-terminal end after signal peptide	rs142417928	1:2308464131:230846413	G/T	–	SNP	dbSNP	Frequency	401 Tolerated (0.4)	22 Benign (0.021)
D64Y	**N-terminal end after signal peptide**	**TMP_ESP_1_230846407**	**1:2308464071:230846407**	**C/A**	–	**SNP**	**ESP**	**ESP**	**11 Deleterious (0.01)**	**964 Probably damaging (0.963)**
D64Y	N-terminal end after signal peptide	COSM171702	1:2308464071:230846407	C/A	–	Somatic_SNV	COSMIC	–	751 Tolerated (0.75)	20 Benign (0.019)
F67L	**N-terminal end after signal peptide**	**TMP_ESP_1_230846396**	**1:2308463961:230846396**	**G/C**	–	**SNP**	**ESP**	**ESP**	**1 Deleterious**	**1000 Probably damaging (0.999)**

P71L	**N-terminal end after signal peptide**	**rs112711075**	**1:2308463851:230846385**	**G/A**	**0.001 (A)**	**SNP**	**dbSNP**	**Multiple_observations, frequency, ESP**	**(0) 1 Deleterious (0)**	**960 Probably damaging (0.959)**
P71T	N-terminal end after signal peptide	COSM1236224	1:2308463861:230846386	G/T	–	Somatic_SNV	COSMIC	–	**1 Deleterious (0)**	**975 Probably damaging (0.974)**
A74V	N-terminal end after signal peptide	rs201569036	1:2308463761:230846376	G/A	–	SNP	dbSNP	–	**1 Deleterious (0)**	**997 Probably damaging (0.996)**
A74P	N-terminal end after signal peptide	rs141724549	1:2308463771:230846377	C/G	–	SNP	dbSNP	Frequency, ESP	161 Tolerated (0.16)	11 Benign (0.01)
A83V	**N-terminal end after signal peptide**	**rs140901725**	**1:2308463491:230846349**	**G/A**	–	**SNP**	**dbSNP**	**Multiple_observations, frequency, ESP**	**21 Deleterious (0.02)**	**952 Probably damaging (0.951)**
L84P	N-terminal end after signal peptide	rs61762540	1:2308463461:230846346	A/G	–	SNP	dbSNP	Multiple_observations, frequency, ESP	171 Tolerated (0.17)	21 Benign (0.02)
D86V	N-terminal end after signal peptide	COSM425564	1:2308463401:230846340	T/A	–	Somatic_SNV	COSMIC	–	**1 Deleterious (0)**	**1001 Probably damaging (1)**
E98K	**N-terminal end after signal peptide**	**rs11568032**	**1:2308463051:230846305**	**C/T**	**0.004 (T)**	**SNP**	**dbSNP**	**Multiple_observations, frequency, HapMap, 1000 genomes, ESP**	**11 Deleterious (0.01)**	**943 Probably damaging (0.942)**
K100E	N-terminal end after signal peptide	rs11557882	1:2308462991:230846299	T/C	–	SNP	dbSNP	–	**1 Deleterious (0)**	**990 Probably damaging (0.989)**
R102K	Helix hA	rs200223350	1:2308462921:230846292	C/T	0.001 (T)	SNP	dbSNP	–	221 Tolerated (0.22)	189 Benign (0.188)
A103T	Helix hA	COSM210306	1:2308462901:230846290	C/T	–	Somatic_SNV	COSMIC	–	1001 Tolerated (1)	5 Benign (0.004)
A104T	**Helix hA**	**rs201406560**	**1:2308462871:230846287**	**C/T**	–	**SNP**	**dbSNP**	–	**1 Deleterious (0)**	**910 Probably damaging (0.909)**
A110V	Helix hA	rs11557881	1:2308462681:230846268	G/A	–	SNP	dbSNP	–	101 Tolerated (0.1)	13 Benign (0.012)
G114C	Helix hA	rs2229389	1:2308462571:230846257	C/A	–	SNP	dbSNP	Multiple_observations, frequency	831 Tolerated (0.83)	18 Benign (0.017)
R116C	**Helix hA**	**rs147355405**	**1:2308462511:230846251**	**G/A**	–	**SNP**	**dbSNP**	**Frequency**	**11 Deleterious (0.01)**	**983 Probably damaging (0.982)**
I117R	Helix hA	rs199864970	1:2308462471:230846247	A/C	–	SNP	dbSNP	–	91 Tolerated (0.09)	288 Benign (0.287)
E123K	Helix hA	rs149236456	1:2308462301:230846230	C/T	–	SNP	dbSNP	Frequency	421 Tolerated (0.42)	4 Benign (0.003)
V127M	Loop between helix hA and sheet s6B	TMP_ESP_1_230846218	1:2308462181:230846218	C/T	–	SNP	ESP	ESP	221 Tolerated (0.22)	195 Benign (0.194)
L134F	Sheet s6B	TMP_ESP_1_230846197	1:2308461971:230846197	G/A	–	SNP	ESP	ESP	441 Tolerated (0.44)	58 Benign (0.057)
T137M	Helix hB	rs34829218	1:2308461871:230846187	G/A	0.001 (A)	SNP	dbSNP	Multiple_observations, frequency, 1000 genomes, ESP	441 Tolerated	54 Benign (0.053)

(*continued on next page*)

Table 1 (*continued*)

Variations	Structural elements	ID	Chromosomal location	Alleles	gmaf	Class	Source	Status	SIFT (Score)	Polyphen V2 (Score)
									(0.44)	
T137A	Helix hB	rs138340265	1:2308461881:230846188	T/C	–	SNP	dbSNP	Multiple_observations, frequency, ESP	161 Tolerated (0.16)	433 Benign (0.432)
A138S	Helix hB	rs61762539	1:2308461851:230846185	C/A	0.001 (A)	SNP	dbSNP	Multiple_observations, frequency	51 Deleterious (0.05)	998 Probably damaging (0.997)
A138T	Helix hB	COSM210305	1:2308461851:230846185	C/T	–	Somatic_SNV	COSMIC	–	–	–
V139L	Helix hB	rs144347709	1:2308461821:230846182	C/G	0.001 (G)	SNP	dbSNP	Multiple_observations, frequency, 1000 genomes, ESP	361 Tolerated (0.36)	1 Benign (0)
A144T	Helix hB	rs138660113	1:2308461671:230846167	C/T	–	SNP	dbSNP	Frequency	1001 Tolerated (1)	1 Benign (0)
L151W	Between helices hB and hC	rs150161533	1:2308461451:230846145	A/C	–	SNP	dbSNP	ESP	881 Tolerated (0.88)	3 Benign (0.002)
A155V	Helix hC	rs143437550	1:2308461331:230846133	G/A	–	SNP	dbSNP	Frequency, ESP	781 Tolerated (0.78)	783 Possibly damaging (0.782)
D156G	Helix hC	rs61757173	1:2308461301:230846130	T/C	–	SNP	dbSNP	Frequency, ESP	231 Tolerated (0.23)	421 Benign (0.42)
D156H	Helix hC	COSM1182244	1:2308461311:230846131	C/G	–	Somatic_SNV	COSMIC	–	421 Tolerated (0.42)	304 Benign (0.303)
H178Y	**Between helices HC and hD**	**COSM242875**	**1:2308460651:230846065**	**G/A**	–	**Somatic_SNV**	**COSMIC**	–	**21 Deleterious (0.02)**	**256 Benign (0.255)**
A186V	**Helix hD**	**rs61757175**	**1:2308460401:230846040**	**G/A**	–	**SNP**	**dbSNP**	**Frequency**	**1 Deleterious (0)**	**977 Probably damaging (0.976)**
V187L	**Helix hD**	**COSM425563**	**1:2308460381:230846038**	**C/A**	–	**Somatic_SNV**	**COSMIC**	–	**1 Deleterious (0)**	**1001 Probably damaging (1)**
Q188L	Helix hD	TMP_ESP_1_230846034	1:2308460341:230846034	T/A	–	SNP	ESP	ESP	541 Tolerated (0.54)	751 Possibly damaging (0.75)
A201T	Sheet s2A	rs138015904	1:2308459961:230845996	C/T	–	SNP	dbSNP	Frequency,	201 Tolerated (0.2)	82 Benign (0.081)
S206F	Sheet s2A	rs11557883	1:2308459801:230845980	G/A	–	SNP	dbSNP	-	311 Tolerated (0.31)	204 Benign (0.203)
T207M	**Sheet s2A**	**rs4762**	**1:2308459771:230845977**	**G/A**	**0.104 (A)**	**SNP**	**dbSNP**	**Multiple_observations, frequency, HapMap, 1000 genomes, cited, ESP**	**1 Deleterious (0)**	**1001 Probably damaging (1)**
T213K	**Sheet s2A**	**COSM373955**	**1:2308459591:230845959**	**G/T**	–	**Somatic_SNV**	**COSMIC**	–	**11 Deleterious (0.01)**	**1001 Probably damaging (1)**
L219Q	**Loop between sheet s2A and helix hE**	**rs141302625**	**1:2308459411:230845941**	**A/T**	–	**SNP**	**dbSNP**	**Frequency**	**1 Deleterious (0)**	**1001 Probably damaging (1)**
P222L	Helix hE	rs146566988	1:2308459321:230845932	G/T/A	–	SNP	dbSNP	Multiple_observations, frequency, ESP	721 Tolerated	2 Benign (0.001)

									(0.72)	
P222Q	Helix hE	rs146566988	1:2308459321:230845932	G/T/A	–	SNP	dbSNP	Multiple_observations, frequency, ESP	181 Tolerated (0.18)	824 Possibly damaging (0.823)
F223L	Helix hE	COSM679126	1:2308459301:230845930	A/G	–	Somatic_SNV	COSMIC	–	471 Tolerated (0.47)	441 Possibly damaging (0.44)
A228D	Helix hE	rs143666867	1:2308459141:230845914	G/T	–	SNP	dbSNP	Frequency	481 Tolerated (0.48)	7 Benign (0.006)
R237L	**Sheet s1A**	**rs145882750**	**1:2308458871:230845887**	**C/A**	–	**SNP**	**dbSNP**	**Frequency**	**11 Deleterious (0.01)**	**1000 Probably damaging (0.999)**
R237C	**Sheet s1A**	**rs61762537**	**1:2308458881:230845888**	**G/A**	–	**SNP**	**dbSNP**	**Multiple_observations, frequency, ESP**	**41 Deleterious (0.04)**	**956 Probably damaging (0.955)**
L244R	Helix hF	rs5041	1:2308458661:230845866	A/C	0.001 (C)	SNP	dbSNP	Multiple_observations, frequency, HapMap, ESP	321 Tolerated (0.32)	249 Benign (0.248)
V246A	Helix hF	TMP_ESP_1_230845860	1:2308458601:230845860	A/G	–	SNP	ESP	ESP	1001 Tolerated (1)	2 Benign (0.001)
A248V	Helix hF	rs140964843	1:2308458541:230845854	G/A	–	SNP	dbSNP	Frequency	541 Tolerated (0.54)	61 Benign (0.06)
W261R	Loop between helix hF and sheet s3A	TMP_ESP_1_230845816	1:2308458161:230845816	A/G	–	SNP	ESP	ESP	151 Tolerated (0.15)	79 Benign (0.078)
T263P	Loop between helix hF and sheet s3A	COSM905376	1:2308458101:230845810	T/G	–	Somatic_SNV	COSMIC	–	461 Tolerated (0.46)	11 Benign (0.01)
G264S	Loop between helix hF and sheet s3A	rs147721058	1:2308458071:230845807	C/T	–	SNP	dbSNP	Multiple_observations, frequency, ESP	231 Tolerated (0.23)	722 Possibly damaging (0.721)
M268I	Loop between helix hF and sheet s3A	rs11568053	1:2308457931:230845793	C/T	–	SNP	dbSNP	Multiple_observations, frequency	181 Tolerated (0.18)	14 Benign (0.013)
M268T	Loop between helix hF and sheet s3A	rs699	1:2308457941:230845794	A/G	0.338 (A)	SNP	dbSNP	Multiple_observations, frequency, HapMap, 1000 genomes, cited, ESP	91 Tolerated (0.09)	94 Benign (0.093)
M268T	**Loop between helix hF and sheet s3A**	**COSM425562**	**1:2308457941:230845794**	**A/G**	–	**Somatic_SNV**	**COSMIC**	–	**21 Deleterious (0.02)**	**910 Probably damaging (0.909)**
GA269GS	**Loop between helix hF and sheet s3A**	**rs141900991**	**1:2308457891:230845789-230845790**	**CT/AC**	–	**Substitution**	**dbSNP**	–	**1 Deleterious (0)**	**973 Probably damaging (0.972)**
Y281C	**Sheet s3A**	**rs56073403**	**1:2308457551:230845755**	**T/C**	**0.001 (C)**	**SNP**	**dbSNP**	**Multiple_observations, frequency, 1000 genomes, ESP**	**1 Deleterious (0)**	**1000 Probably damaging (0.999)**
V282I	Sheet s3A	TMP_ESP_1_230845753	1:2308457531:230845753	C/T	–	SNP	ESP	ESP	1001 Tolerated (1)	12 Benign (0.011)
A295S	Helix hF1	TMP_ESP_1_230841920	1:2308419201:230841920	C/A	–	SNP	ESP	ESP	701 Tolerated (0.7)	20 Benign (0.019)
E296K	Loop between sheet s3B and helix hF1	rs139685563	1:2308419171:230841917	C/T	0.001 (T)	SNP	dbSNP	Multiple_observations, frequency, 1000 Genomes, ESP	201 Tolerated (0.2)	315 Benign (0.314)
P297S	Sheet s4C	rs61762530	1:2308419141:230841914	G/A	–	SNP	dbSNP	Multiple_observations, frequency, ESP	181 Tolerated	53 Benign (0.052)

(*continued on next page*)

Table 1 (*continued*)

Variations	Structural elements	ID	Chromosomal location	Alleles	gmaf	Class	Source	Status	SIFT (Score)	Polyphen V2 (Score)
									(0.18)	
F300L	Sheet s4C	COSM1340017	1:2308419031:230841903	G/T	–	Somatic_SNV	COSMIC	–	101 Tolerated (0.1)	123 Benign (0.122)
S305N	Loop between sheets s4C and s3C	rs200008829	1:2308418891:230841889	C/T	0.001 (T)	SNP	dbSNP	–	81 Tolerated (0.08)	279 Benign (0.278)
S307L	Sheet s3C	rs151246535	1:2308418831:230841883	G/A	–	SNP	dbSNP	–	81 Tolerated (0.08)	84 Benign (0.083)
V310L	Sheet s3C	rs201352496	1:2308418751:230841875	C/G	–	SNP	dbSNP	ESP	191 Tolerated (0.19)	7 Benign (0.006)
L313F	**Sheet s3C**	**COSM1340016**	**1:2308418661:230841866**	**G/A**	–	**Somatic_SNV**	**COSMIC**	–	**41 Deleterious (0.04)**	**451 Possibly damaging (0.45)**
W322R	Sheet s1B	TMP_ESP_1_230841839	1:2308418391:230841839	A/G	–	SNP	ESP	ESP	171 Tolerated (0.17)	2 Benign (0.001)
D327H	Loop between sheets s1B and s2B	rs151194891	1:2308418241:230841824	C/G	–	SNP	dbSNP	ESP	81 Tolerated (0.08)	22 Benign (0.021)
P335S	**Sheet s2B**	**rs17856352**	**1:2308418001:230841800**	**G/A**	–	**SNP**	**dbSNP**	–	**21 Deleterious (0.02)**	**476 Possibly damaging (0.475)**
S339N	Loop between sheets s2B and s3B	rs142892394	1:2308417871:230841787	C/T	0.001 (T)	SNP	dbSNP	Multiple_observations, frequency, ESP	801 Tolerated (0.8)	50 Benign (0.049)
A340T	**Sheet s3B**	**rs143215026**	**1:2308417851:230841785**	**C/T**	–	**SNP**	**dbSNP**	**Frequency, ESP**	**11 Deleterious (0.01)**	**978 Probably damaging (0.977)**
I345S	Sheet s3B	rs147736976	1:2308417691:230841769	A/C	–	SNP	dbSNP	Frequency	1001 Tolerated (1)	1 Benign (0)
Q346H	**Sheet s3B**	**TMP_ESP_1_230841765**	**1:2308417651:230841765**	**C/A**	–	**SNP**	**ESP**	**ESP**	**41 Deleterious (0.04)**	**92 Benign (0.091)**
P347L	Sheet s3B	rs201501261	1:2308417631:230841763	G/A	0.001 (A)	SNP	dbSNP	–	131 Tolerated (0.13)	20 Benign (0.019)
D352N	Loop between sheet s3B and helix hG	rs150452789	1:2308417491:230841749	C/T	0.001 (T)	SNP	dbSNP	ESP	81 Tolerated (0.08)	682 Possibly damaging (0.681)
L353V	Helix hG	rs61762529	1:2308417461:230841746	G/C	–	SNP	dbSNP	Frequency, ESP	51 Deleterious (0.05)	459 Possibly damaging (0.458)
L366I	**Helix hH**	**COSM1501178**	**1:2308417071:230841707**	**G/T**	–	**Somatic_SNV**	**COSMIC**	–	**41 Deleterious (0.04)**	**868 Possibly damaging (0.867)**
M369K	Helix hH	rs4932	1:2308416971:230841697	A/T	–	SNP	dbSNP	Multiple_observations	91 Tolerated (0.09)	735 Possibly damaging (0.734)
R375Q	Sheet s2C	rs74315283	1:2308416791:230841679	C/T	–	SNP	dbSNP	Multiple_observations, frequency, ESP	331 Tolerated (0.33)	10 Benign (0.009)
R375W	Sheet s2C	rs201162475	1:2308416801:230841680	G/A	–	SNP	dbSNP	ESP	221 Tolerated (0.22)	14 Benign (0.013)
M381R	Sheet s2C	rs137858911	1:2308400661:230840066	A/C	–	SNP	dbSNP	Frequency, ESP	301 Tolerated	3 Benign (0.002)

									(0.3)	
P382A	Sheet s2C	rs61762527	1:2308400641:230840064	G/C	–	SNP	dbSNP	Frequency	681 Tolerated (0.68)	1 Benign (0)
V385A	Sheet s6A	rs61731499	1:2308400541:230840054	A/G	0.004 (G)	SNP	dbSNP	Multiple_observations, frequency, 1000 genomes, ESP	741 Tolerated (0.74)	10 Benign (0.009)
L392M	Helix hI	rs1805090	1:2308400341:230840034	G/T	0.001 (T)	SNP	dbSNP	Multiple_observations, frequency, 1000 genomes, ESP	1001 Tolerated (1)	8 Benign (0.007)
A397S	Helix hI	rs61751065	1:2308400191:230840019	C/T/A	–	SNP	dbSNP	Multiple_observations, frequency	271 Tolerated (0.27)	23 Benign (0.022)
A397T	**Helix hI**	**rs61751065**	**1:2308400191:230840019**	**C/T/A**	–	**SNP**	**dbSNP**	**Multiple_observations, frequency**	**21 Deleterious (0.02)**	**748 Possibly damaging (0.747)**
A403T	Helix hI1	rs147243938	1:2308400011:230840001	C/T	–	SNP	dbSNP	Frequency, ESP	221 Tolerated (0.22)	2 Benign (0.001)
L409P	Loop between helix hI1 and sheet s5A	rs199817559	1:2308399821:230839982	A/G	0.001 (G)	SNP	dbSNP	–	101 Tolerated (0.1)	10 Benign (0.009)
S415G	Loop between helix hI1 and sheet s5A	rs61751066	1:2308399651:230839965	T/C	–	SNP	dbSNP	Frequency	441 Tolerated (0.44)	565 Possibly damaging (0.564)
N416S	**Loop between helix hI1 and sheet s5A**	**COSM464129**	**1:2308399611:230839961**	**T/C**	–	**Somatic_SNV**	**COSMIC**	–	**11 Deleterious (0.01)**	**936 Probably damaging (0.935)**
R418C	**Loop between helix hI1 and sheet s5A**	**rs200712921**	**1:2308399561:230839956**	**G/A**	**0.001 (A)**	**SNP**	**dbSNP**	–	**1 Deleterious (0)**	**1000 Probably damaging (0.999)**
R418H	Loop between helix hI1 and sheet s5A	COSM1176529	1:2308399551:230839955	C/T	–	Somatic_SNV	COSMIC	–	**1001 Tolerated (1)**	**1 Benign (0)**
V421M	Loop between helix hI1 and sheet s5A	rs61751067	1:2308399471:230839947	C/T	–	SNP	dbSNP	Frequency	151 Tolerated (0.15)	33 Benign (0.032)
E423A	**Sheet s5A**	**COSM1182243**	**1:2308399401:230839940**	**T/G**	–	**Somatic_SNV**	**COSMIC**	–	**1 Deleterious (0)**	**1001 Probably damaging (1)**
E433A	**Loop between sheets s5A and s4A**	**COSM905371**	**1:2308390471:230839047**	**T/G**	–	**Somatic_SNV**	**COSMIC**	–	**11 Deleterious (0.01)**	**931 Probably damaging (0.93)**
A434V	**Loop between sheets s5A and s4A**	**rs61751076**	**1:2308390441:230839044**	**G/A**	–	**SNP**	**dbSNP**	**Multiple_observations, frequency, ESP**	**31 Deleterious (0.03)**	**25 Benign (0.024)**
E441D	Sheet s4A	rs200657291	1:2308390221:230839022	C/G	–	SNP	dbSNP	-	201 Tolerated (0.2)	956 Probably damaging (0.955)
S442F	Sheet s4A	rs61751077	1:2308390201:230839020	G/A	0.001 (A)	SNP	dbSNP	Multiple_observations, frequency,1000 genomes, ESP	571 Tolerated (0.57)	508 Possibly damaging (0.507)
P449S	Sheet s4A	rs61751078	1:2308390001:230839000	G/A	–	SNP	dbSNP	Multiple_observations, frequency, ESP	441 Tolerated (0.44)	49 Benign (0.048)
R458C	Loop between sheets s1C and s4B	TMP_ESP_1_230838973	1:2308389731:230838973	G/A	–	SNP	ESP	ESP	281 Tolerated (0.28)	54 Benign (0.053)
H473Y	Sheet s5B	COSM238730	1:2308389281:230838928	G/A	–	Somatic_SNV	COSMIC	–	131	12 Benign (0.011)

(*continued on next page*)

Table 1 (*continued*)

Variations	Structural elements	ID	Chromosomal location	Alleles	gmaf	Class	Source	Status		SIFT (Score)	Polyphen V2 (Score)
										Tolerated (0.13)	
R477H	Sheet s5B	rs146284519	1:230838915:230838915	C/T	0.001 (T)	SNP	dbSNP	Multiple_observations, frequency		81 Tolerated (0.08)	5 Benign (0.004)
R477C	**Sheet s5B**	**TMP_ESP_1_230838916**	**1:230838916:230838916**	**G/A**	–	**SNP**	**ESP**	**ESP**		**21 Deleterious (0.02)**	**476 Possibly damaging (0.475)**
A479S	C-terminal end	COSM1127109	1:230838910:230838910	C/A	–	Somatic_SNV	COSMIC	–		431 Tolerated (0.43)	35 Benign (0.034)
P481L	C-terminal end	rs143479528	1:230838903:230838903	G/A	0.001 (A)	SNP	dbSNP	Multiple_observations, frequency, ESP		361 Tolerated (0.36)	70 Benign (0.069)

AGT gene originated at the very early stages of vertebrate evolution at about 500 MYA.

3.3. Catalogue of AGT genetic variants using 1092 human genomes

We computed variations in the AGT gene in 1092 human genomes from 14 different populations and details are provided in Fig. 2 with majority of these are SNPs. There are 690 variations in total (Table S3) with major components of 613 SNPs, 36 somatic SNVs, 29 deletions, and 19 insertions, shared in 12 variant types (Fig. 3). Top 5 variant types were intron variants, upstream gene variants, downstream gene variants, missense variants and synonymous variants. Out of 690 AGT variants, 68% were validated variations. From Fig. 2, we further examined 121 missense variations to examine what are the critical changes, which can cause alteration in both AGT amino acid sequence and thus secondary structural elements (compiled in Table 1). Out of all mutations, 67% are validated at least by one source (or study) where as 29% of missense mutants were validated by either multiple sources or studies. Various structural elements of AGT have at least 1 genetic variants also summarized in Table 2. We combined impact of these AGT variants using SIFT [9] and PolyPhen V2 [10] tools in Table 1. SIFT and PolyPhen V2 methods entail a total of 32 key amino acids as deleterious and probably damaging (marked bold in Table 1), respectively. These deleterious mutations are marked by orange stars (where as other mutations are marked by green stars), above protein sequence alignment of AGT in Fig. S1. About half of these deleterious mutations localized in the N-terminal extension of AGT gene with 5 and 12 in the signal peptide and the N-terminal region after signal peptide, respectively. Helix A has 8 missense mutations and half of which are deleterious mutations. Other variants are distributed across different secondary structural elements of AGT protein (Fig. S1).

3.4. Sequence analysis of AGT

AGT is highly conserved in all vertebrates from lamprey to human with variable sequence identities and similarities (Table 2). Human AGT shares 36/55, 27/49, 23/38, 24/45 and 20/38 percentage identities/similarities with turkey, coelacanth, medaka, cave fish and lampreys, respectively. AGT from three lamprey species shares 94–98% identities with each other, while AGT from other vertebrates shares 20–24% identities with AGT from lampreys. There are 20, 64 and 85 amino acids in AGT alignment, which are conserved 100%, 70–99%, and 50–69%, respectively (Table 3). Additionally, there are 51 amino acids conserved in the core serpin domain in more than 70% of serpins [18] and from these 51 residues, 36 were maintained in all vertebrate AGT and 15 only in few species (Table 3) with C-terminal conserved residues only present in AGT protein in lampreys (Fig. S1).

3.4.1. Angiotensin II is maintained from lamprey to human

Angiotensin II is created by removal of two C-terminal residues from angiotensin I by the enzyme angiotensin-converting enzyme (ACE). We found that angiotensin II is conserved in all vertebrates (Fig. 3A) with some variations at the positions 1, 3, 5, 6 and 9.

3.4.2. Serpin motifs are maintained

AGT protein is characterized by three conserved serpin motifs, which are spanned across three major structural regions (Fig. S1). Motif-I is localized in s3A-breach-s4C and it has four human genetic variants. Motif-II is conserved in the s5A–s4A, which starts just before RCL and ends in P8 position of the RCL and it has also four human genetic variants. Motif-III is located immediately after RCL at the C-terminal end, 1C-turn-s4B and this motif has four human genetic variants.

Table 2
Percentage identities and similarities values of AGT for selected vertebrates.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.
1. Human		50	35	36	27	25	25	23	23	24	23	24	24	24	20	20	20
2. Opossum	67		36	36	28	26	27	26	24	26	24	23	24	26	21	21	21
3. Chicken	55	55		92	31	30	30	29	25	28	29	28	28	28	23	23	23
4. Turkey	55	57	95		30	30	30	29	25	29	29	27	28	28	23	22	22
5. Coelacanth	49	52	53	53		28	29	28	28	28	30	27	29	29	23	24	23
6. Cod	44	48	51	52	50		59	61	59	68	67	35	52	54	22	22	21
7. Fugu	47	50	51	52	51	74		60	57	68	66	35	48	51	21	21	21
8. Medaka	44	46	51	52	49	77	77		64	71	69	36	46	50	21	20	20
9. Platyfish	45	46	49	49	49	73	72	79		68	66	34	47	49	20	20	20
10. Tilapia	46	47	51	52	50	80	79	85	80		76	36	50	55	22	21	21
11. Stickleback	44	47	51	51	52	79	79	82	79	86		37	51	57	23	22	22
12. Spottedgar	38	41	46	46	46	52	55	54	53	54	55		37	39	21	21	21
13. Zebrafish	45	48	50	50	49	70	68	67	67	70	71	55		68	24	24	23
14. Cavefish	45	47	51	52	51	72	69	71	69	75	74	56	81		22	22	22
15. Lethenteron	38	41	43	42	44	43	43	44	41	44	45	39	42	42		98	94
16. Lampetra	38	41	43	42	45	42	43	43	41	43	45	39	42	42	98		94
17. Petromyzon	38	41	43	42	45	43	43	43	41	43	45	39	42	42	96	96	

Sequence identities (right of the table)

Sequence similarities (below the table)

Table 3
Summary of amino acid conservations and number of AGT variants in secondary structural elements of AGT protein in vertebrates.

Structural components	Id-100	Id-70–99	Id-50–69	Number of AGT genetic variants	Status of 51 conserved amino acids proposed by Irving et al. (2000) [18] Bold – Missing Italics – Conserved in few species				
Signal peptide	0	0	6	7					
N-terminal end after signal peptide	5	5	13	16					
hA	3	5	6	8	*F33*				
s6B	0	1	0	1	N49	S53			
hB	2	2	2	4	P54	*S56*	L61	G67	
hC	1	2	1	2	T72	L80			
hD	0	4	5	3					
s2A	1	1	3	4					
hE	1	0	3	3	F130				
s1A	1	1	0	1					
hF	0	4	7	3	F147	*I157*	N158	V161	*T165*
Loop between hF/s3A	0	2	4	5	*I169*	T180			
s3A	0	2	0	2	*L184*	*N186*	F190	*K191*	G192
hF1	0	1	2	1					
s4C	1	1	3	2	F198	*T203*	F208		
s3C	0	4	3	3	V218	*M220*	M221		
s1B	0	1	0	1					
s2B	0	2	4	1	*Y244*				
s3B	1	3	1	4	L254	P255			
hG	2	0	1	1					
hH	0	1	1	2					
s2C	1	2	1	3					
s6A	0	2	2	1	P289	K290			
hI	1	1	1	2	L299	L303	*G307*		
hI1	0	1	1	1					
Loop between hI/s5A	0	3	7	5	*F312*	A316	*L327*		
s5A	0	3	3	1	*H334*	E342			
s4A (RCL)	0	3	1	3	G344	*A347*			
s1C	0	0	0	0					
s4B	0	2	2	0	P369	F370			
s5B	0	3	2	2	*L383*	F384	G386		
C terminus	0	2	0	2	P391				

Id = Identity.

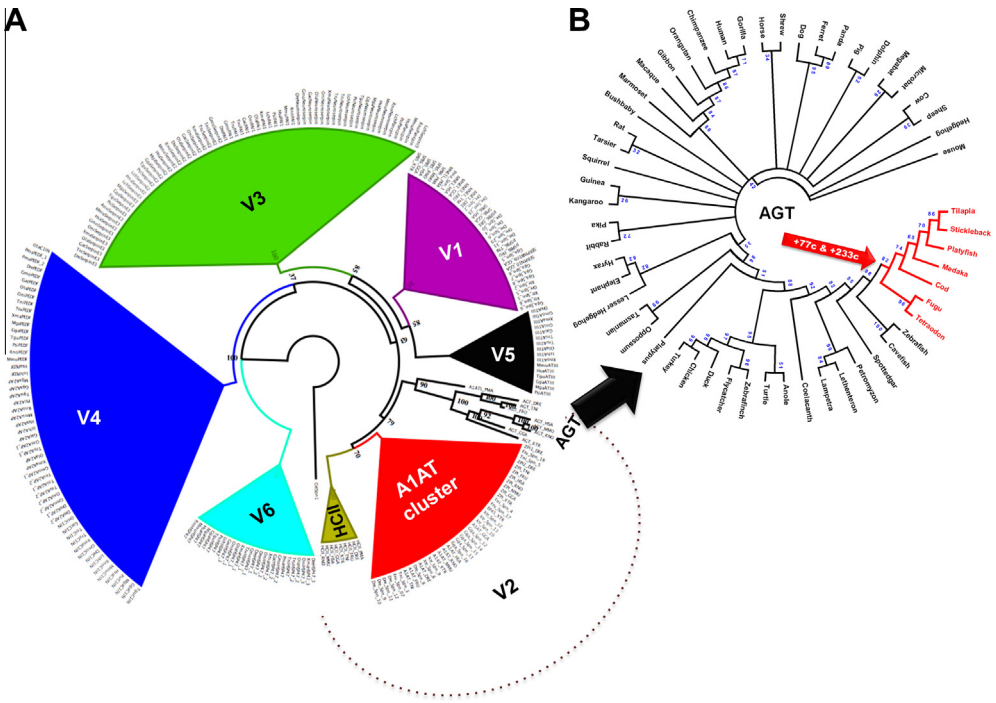


Fig. 4. Phylogenetic history of vertebrate serpins. (A) Bayesian phylogenetic tree of vertebrate group-wise (V1–V6) illustrates AGT is a member of group V2 serpins along with A1AT-cluster (red shade) and HCII as visualized by FigTree V1.4 (website: http://tree.bio.ed.ac.uk/software/figtree/) (B) Neighbor-Joining based phylogenetic tree of AGT demonstrates novel introns (red arrow) in conserved domain are inserted after separation of lampreys, cave fish, spotted gar zebrafish from other ray-finned fishes. Branches corresponding to partitions reproduced in less than 25% bootstrap replicates are collapsed for visualization in MEGA 5.2.2 [15]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.4.3. Reactive center loop changed from inhibitory to non-inhibitory during 500 MY

We analysed AGT from three different lamprey species and found that AGT is inhibitory in nature and also maintained angiotensin II (Fig. 3). Lampreys have an inhibitory RCL as GTEAKAETVVGIMPI†SMPPT with active site P1–P1 residues as I–S (Fig. 3B and Table 4). But, during vertebrate evolution, RCL region has mutations that made this gene non-inhibitory via mutations in 20 RCL residues (Fig. 3). The position P15 is generally a glycine, which is maintained only lampreys and ray-finned fishes (except in platyfish) and five other substitutions has happened at this position (Fig. 3C). The position P14 has normally either predominantly a threonine (T) or sometimes a serine (S) in all inhibitory serpins and this position has a T in lampreys, *Fugu* and birds (Table 4) and four replacements have occurred at this position (Fig. 3C). The position P13 has normally a glutamic acid and it is predominant in AGT as well (Fig. 3C), but it is also replaced by four other amino acids. Positions P12–P8 are normally small amino acids as in lampreys (Fig. 3B), but AGT from other vertebrates have rapid mutations predominantly by a bulky residue such as a proline or a glutamine. Fig. 3C illustrates four to five replacements have been occurring in RCL regions of AGT other than lampreys (Fig. 3B) with P13–P1 region has frequently either a proline or a glutamine. This infers that RCL region became non-inhibitory after separation of lampreys and other vertebrates (Table 4). Lampreys maintained this inhibitory AGT gene, while constant mutations have occurred in RCL region other vertebrates, leading into various types of non-inhibitory RCLs, such as in Flycatcher (GTDQPADPAAQKEDG†VYLDV) and in human (EREPTESTQQLNKPE†VLEVT).

This corroborates lampreys have an inhibitory AGT, which became non-inhibitory during vertebrate evolution in ~500 MY.

3.5. Phylogenetic analysis of AGT

Bayesian phylogenetic analysis classified vertebrate serpins into six groups as intron encoded classification as depicted in different colors (Fig. 4A). Group V2 is composed of majority of genes in A1AT cluster (red shade in Fig. 4A), originated by tandem duplications and two independent genes namely HCII and AGT, localized on separate cluster without tandem duplication. This illustrates that AGT gene is a member of the group V2 (Fig. 4A). To evaluate location of novel intron insertions, species-wide phylogenetic tree was constructed (Fig. 4B), which reveals after separation of zebrafish from other-ray-finned fishes, these novel introns are embedded in AGT gene.

4. Discussion

The unique position of AGT in the renin–angiotensin system and its distinct features make this serpin become an attractive target in developing effective therapeutic strategies in many human diseases such as atherosclerosis and obesity. However, despite great efforts on discovering the molecular mechanisms and clinical relevance of AGT gene and protein functions, questions about genetic variants and detailed phylogenetic origin remain unanswered. This study answers these issues and provides an updated repository of the AGT gene from 57 vertebrate species (Table S1) and summarizes major concepts revolving around sequence, structure and phylogeny of AGT across vertebrate genomes. AGT gene is characterized by four exons I–IV in vertebrates with two novel intron insertions; one each within exons I and II (at positions 77c and 233c, respectively) in ray-finned fishes (Fig. 1). Interestingly, exon III and IV are fully conserved without any insertions. This intron gain event in the AGT gene illustrates an excellent example, which

Table 4
Reactive center loops (RCL) of AGT for selected vertebrates depicts inhibitory to non-inhibitory RCL formation during 500 MY.

Species	P15	P15	P13	P12	P1	P10	P9	P8	P7	P6	P5	P4	P3	P2	P1	P1′	P2′	P3′	P4′	P5′
Human	*E*	*R*	E	*P*	*T*	E	S	*T*	Q	Q	L	N	K	P	E	V	L	E	V	T
Platypus	G	*P*	E	A	*P*	E	E	*P*	T	S	A	S	V	D	S	E	P	L	E	V
Opossum	Q	*R*	E	Q	*D*	S	F	*Q*	Q	Q	N	E	A	V	P	L	E	V	K	M
Shrew	A	S	E	Q	*P*	S	E	*N*	G	Q	S	*P*	H	E	P	E	S	L	E	V
Elephant	*E*	*G*	K	Q	*P*	P	E	*S*	A	Q	Q	*P*	G	P	E	A	L	E	V	T
Gibbon	*E*	*R*	E	*P*	*T*	E	S	*T*	Q	Q	L	N	R	P	E	V	L	E	V	T
Orangutan	*E*	*R*	E	*P*	*T*	E	S	*T*	Q	E	L	N	R	P	E	V	L	E	V	T
Chimpanzee	*E*	*R*	E	*P*	*T*	E	S	*T*	Q	Q	L	N	K	P	E	V	L	E	V	T
Gorilla	*E*	*R*	E	*P*	*T*	E	S	*T*	Q	Q	L	N	K	P	E	V	L	E	V	T
Guinea	D	R	K	Q	*P*	*T*	E	S	T	P	Q	S	S	A	P	E	A	L	E	V
Mouse	E	*E*	E	Q	*P*	*T*	T	S	V	Q	Q	*P*	G	S	P	E	A	L	D	V
Rat	E	*E*	E	Q	*P*	*T*	E	S	A	Q	Q	*P*	G	S	P	E	V	L	D	V
Bushbaby	D	*G*	D	Q	*P*	*T*	E	S	A	Q	Q	*P*	D	G	P	E	V	L	E	L
Squirrel	E	*E*	E	Q	*P*	A	E	S	A	Q	Q	*P*	S	M	S	E	A	L	E	V
Rabbit	D	*R*	E	Q	*P*	V	E	S	A	P	Q	*P*	A	G	P	E	V	L	E	V
Microbat	K	*G*	E	Q	*P*	*T*	E	P	A	P	Q	*P*	T	G	P	E	A	L	E	V
Dolphin	E	*G*	E	Q	*P*	*T*	E	S	A	P	L	*P*	A	G	P	E	V	L	E	V
Sheep	G	*E*	Q	A	*P*	E	S	V	P	Q	P	A	G	P	E	A	L	E	V	T
Dog	Q	*E*	E	Q	*P*	*T*	E	S	A	P	Q	P	D	G	P	E	V	L	E	V
Ferret	E	*G*	E	Q	*P*	*T*	E	S	A	P	Q	*P*	G	E	P	K	A	L	E	V
Panda	E	*G*	E	Q	*P*	*T*	E	S	A	P	Q	*P*	D	G	P	L	E	V	T	L
Anole	G	A	D	E	*A*	E	A	P	S	E	E	N	E	A	T	E	T	L	K	M
Turtle	G	A	E	E	L	L	E	E	N	G	D	S	L	P	L	E	I	Q	L	N
Zebrafinch	G	T	D	Q	*P*	E	D	P	A	A	Q	K	E	D	G	A	Y	L	D	V
Flycatcher	G	T	D	Q	*P*	A	D	P	A	A	Q	K	E	D	G	V	Y	L	D	V
Duck	E	T	D	P	*P*	E	D	P	T	A	Q	K	E	D	S	G	P	L	E	V
Chicken	A	T	H	Q	*P*	E	D	A	T	A	Q	E	E	D	S	V	P	Q	E	V
Turkey	G	T	H	Q	*P*	E	D	T	T	A	Q	E	K	D	S	V	P	Q	E	V
Coelacanth	*Q*	*E*	E	T	D	V	E	N	F	Q	Q	S	N	S	S	D	I	L	E	I
Zebrafish	G	S	E	V	Q	N	R	T	D	D	G	R	A	P	H	K	V	T	F	N
Cave fish	G	S	E	E	Q	S	K	P	Q	D	D	R	A	P	L	K	L	T	V	N
Cod	G	*A*	E	S	E	D	R	N	P	E	G	G	V	P	L	K	L	T	I	N
Platyfish	*A*	*A*	E	P	Q	A	R	T	T	E	A	G	V	A	L	R	L	S	I	N
Fugu	G	T	E	V	Q	E	S	V	E	G	P	S	S	P	L	K	L	S	F	N
Medaka	G	A	E	*P*	Q	D	K	K	E	A	A	G	V	P	L	R	L	S	F	N
Tilapia	G	A	E	*P*	Q	D	P	T	Q	E	E	G	V	P	L	K	L	S	I	N
Stickleback	G	A	E	L	R	D	K	V	Q	E	A	G	V	P	L	K	L	S	I	N
Lethenteron	**G**	**T**	**E**	**A**	**K**	**A**	**E**	**T**	**V**	**V**	**G**	**I**	**M**	**P**	**I**	**S**	**M**	**P**	**P**	**T**
Lampetra	**G**	**T**	**E**	**A**	**K**	**A**	**E**	**T**	**V**	**V**	**G**	**I**	**M**	**P**	**I**	**S**	**M**	**P**	**P**	**T**
Petromyzon	**G**	**T**	**E**	**A**	**K**	**A**	**E**	**T**	**V**	**V**	**G**	**I**	**M**	**P**	**I**	**S**	**M**	**P**	**P**	**T**

shows the typical exon/intron pattern of vertebrate group V2 serpins in lampreys, tetrapods and in three ray-finned fishes (cave fish, spotted gar and zebrafish), Chronologically, the introns at positions 77c and 233c found in orthologs of AGT gene occurred after the split of the zebrafish from other actinopterygians. Likewise, all other non-standard introns found in the vertebrate serpins are also confined to these ray-finned fishes. Eukaryotes have genes with patterns of exons and introns and the hallmark of eukaryotic introns are their splicing mechanisms. However, the mystery about their creation remains puzzling [19], ever since discovery of introns. There are total 24 conserved introns in vertebrate serpins encompassing groups V1–V6 [5] with six additional introns that were gained in selected ray finned fishes among serpin genes [20]. Genome compaction and associate genome repair processes were attributed with several examples of intron creations in selected ray-finned fishes whose genome underwent compaction events in the serpin superfamily [20] and in the GPCR superfamily [21]. To facilitate genome function after genome compactness, DNA breakage and repair processes are essential. These processes are accountable for gains of introns in ray-finned fishes. Intron gain events are believed to be very rare in many metazoan lineages, and the mechanisms underlying creation of spliceosomal introns are largely unknown. This study and recent other studies on GPCRs [21] and serpins [20] confirms that creations of introns in vertebrates are not as exceptional as previously thought.

AGT is localized on the same genomic organization from from lamprey to human maintained for ~500 MY (Fig. 1B). At protein sequence level, following features are maintained: a conserved angiotensin in N-terminal region, which generates spectrum of small peptides, capable of regulating hypertension (Fig. 3A), three serpin motifs and conserved inhibitory RCL (in lampreys), which is mutated to become non-inhibitory with four to five mutations on each site in RCL regins of AGT from 54 vertebrates analyzed (Fig. 3C). AGT proteins from lampreys are thrombin inhibitors [22] and also active angiotensin producers [23]. Antithrombin III (ATIII) gene is missing in lampreys [24]. Haemostasis regulations in vertebrates appear to be so critical that it requires to be controlled by two serpins in all vertebrates including lampreys. AGT gene kept this requirement until ATIII gene is originated around somewhere between 450 to 500 MY lampreys [24]. These corroborates that thrombin inhibition are critical process and AGT gene compensates for ATIII gene in lampreys. After ATIII gene arose in ray-finned fishes [24], AGT lost its inhibitory function by mutations in RCL regions (Fig. 3) and also some of the highly conserved 51 amino acids (Fig. S1). This function is acquired by ATIII gene in other vertebrates.

AGT is a group V2 member (Fig. 4A) . Lampreys have only two group V2 serpins namely AGT and HCII (data not shown), this demonstrates that thrombin inhibition is ancient traits of group V2 serpins and A1AT cluster originated around 450 MY ago in ray-finned fishes. Group V2 serpins are known for considerable clustering of genes by tandem duplication events leading to the expansion of these genes from fishes to mammals [25]. Notably, this gene belongs to only two genes of group V2 without tandem duplications,

and other one is HCII. This also suggested that these two serpins genes are originated by chromosomal duplications events at the beginning of vertebrate evolution. Some of these serpins remained single gene in the chromosomal fragments such as HCII and ATIII gene [24], where as others underwent rapid tandem duplications that originated several serpin paralogs on the same locus such as clade A and B members.

690 AGT variants were identified from 1092 human genomes by combining missense AGT variants (Fig. 2) with top three variation classes being SNPs (89.7%), somatic SNVs (5.2%) and deletion (2.9%). These include 121 missense mutations with 32 are deleterious variants. These findings open up new surveys of these genetic variants in the context of blood pressure regulations.

In conclusion, AGT gene is revisited from sequence-structural, phylogenetic and variants perspective in the post-genomics era. Gene structures of AGT gene are variable in fishes, however synteny of this gene is conserved for ~500 MYA. Several sequence features are maintained in AGT protein from 57 vertebrates analyzed. 690 AGT variants were also identified from 1092 human genomes.

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.2014.02.139.

References

[1] G. Silverman, P. Bird, R. 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.
[2] H. Ragg, T. Lokot, P.B. Kamp, et al., Vertebrate serpins: construction of a conflict-free phylogeny by combining exon-intron and diagnostic site analyses, Mol. Biol. Evol. 18 (2001) 577–584.
[3] 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.
[4] R.F. Doolittle, Angiotensinogen is related to the antitrypsin–antithrombin–ovalbumin family, Science 222 (1983) 417–419.
[5] A. Kumar, H. Ragg, Ancestry and evolution of a secretory pathway serpin, BMC Evol. Biol. 8 (2008) 250.
[6] P. Flicek, I. Ahmed, M.R. Amode, et al., Ensembl 2013, Nucleic Acids Res. 41 (2013) D48–D55.
[7] 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.
[8] 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.
[9] P.C. Ng, S. Henikoff, SIFT: predicting amino acid changes that affect protein function, Nucleic Acids Res. 31 (2003) 3812–3814.
[10] I.A. Adzhubei, S. Schmidt, L. Peshkin, et al., A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248–249.
[11] R.C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics 5 (2004) 113.
[12] K.B. Nicholas, H.B. Nicholas Jr., D.W.I. Deerfield, GeneDoc: analysis and visualization of genetic variation, EMBNEW.NEWS 4 (1997) 14.
[13] G.E. Crooks, G. Hon, J.M. Chandonia, et al., WebLogo: a sequence logo generator, Genome Res. 14 (2004) 1188–1190.
[14] F. Ronquist, M. Teslenko, P. van der Mark, et al., MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space, Syst. Biol. 61 (2012) 539–542.
[15] 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.
[16] A. Roy, A. Kucukural, Y. Zhang, I-TASSER: a unified platform for automated protein structure and function prediction, Nat. Protoc. 5 (2010) 725–738.
[17] E. Krieger, G. Koraimann, G. Vriend, Increasing the precision of comparative models with YASARA NOVA–a self-parameterizing force field, Proteins 47 (2002) 393–402.
[18] 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.
[19] S.W. Roy, W. Gilbert, The evolution of spliceosomal introns: patterns, puzzles and progress, Nat. Rev. Genet. 7 (2006) 211–221.
[20] 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.
[21] 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.
[22] Y. Wang, H. Ragg, An unexpected link between angiotensinogen and thrombin, FEBS Lett. 585 (2011) 2395–2399.
[23] M.K.S. Wong, Y. Takei, Characterization of a native angiotensin from an anciently diverged serine protease inhibitor in lamprey, J. Endocrinol. 209 (2011) 127–137.
[24] A. Kumar, A. Bhandari, S.J. Sarde, et al., Sequence, phylogenetic and variant analyses of antithrombin III, Biochem. Biophys. Res. Commun. (2013).
[25] 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.