Structural and phylogenetic comparison of napsin genes: the duplication, loss of function and human-specific pseudogenization of napsin B

Kazuhisa Nishishita [a,*], Eiko Sakai [a], Kuniaki Okamoto [a], Takayuki Tsukuba [a]

[a] Division of Oral Pathopharmacology, Department of Developmental Reconstructive Medicine, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki 852-8588, Japan

* Corresponding author.

Tel: +81 95 819 7654

Fax: +81 94 819 7655

E-mail:

Kazuhisa Nishishita, kazu@nagasaki-u.ac.jp

Eiko Sakai, eiko-s@nagasaki-u.ac.jp

Kuniaki Okamoto, k-oka@nagasaki-u.ac.jp

Takayuki Tsukuba, tsuta@nagasaki-u.ac.jp

Abstract

Aspartic proteinases form a widely distributed protein superfamily, including cathepsin D, cathepsin E, pepsins, renin, BACE and napsin. Human napsin genes are located on human chromosome 19q13, which comprises napsin A and napsin B. *Napsin B* has been annotated as a pseudogene because it lacks an in-frame stop codon; its nascent chains are cotranslationally degraded. Until recently, there have been no studies concerning the molecular evolution of the napsin protein family in the human genome. In the present study, we investigated the evolution and gene organization of the napsin protein family. *Napsin B* orthologs are primarily distributed in primates, while *napsin A* orthologs are the only napsin genes in other species. The corresponding regions of *napsin B* in the available sequences from primate species contain an in-frame stop codon at a position equivalent to that of human *napsin A*. In addition, a rare single-nucleotide polymorphism (SNP) that creates a proper stop codon in human *napsin B* was identified using HapMap populations. Recombinant protein expression and three-dimensional comparative modeling revealed that napsin B exhibits residual activity toward synthetic aspartic protease substrates compared with napsin A, presumably through a napsin B-specific Arg287 residue. Thus, *napsin B* was duplicated from *napsin A* during the early stages of primate evolution, and the subsequent loss of napsin B function during primate evolution reflected ongoing human-specific *napsin B* pseudogenization.

Keywords: napsin; phylogeny; synteny; pseudogenization; SNP; gene duplication

Abbreviations: NAPSA, napsin A; NAPSB, napsin B; SNP, single-nucleotide polymorphism; cDNA, DNA complementary to RNA; bp, base pair(s); PCR, Polymerase Chain Reaction; kb, kilobase(s) or 1000 bp; NJ, Neighbor-Joining; ML, Maximum-Likelihood; KCNC3, potassium voltage gated channel Shaw-related subfamily member 3; NR1H2, nuclear receptor subfamily 1 group H member 2; atf4, activating transcription factor 4; smcr7l, Smith-Megenis syndrome region candidate 7-like; G418, Geneticin; SDS, sodium dodecyl sulfate; PAGE, polyacrylamide-gel electrophoresis; EDTA, ethylenediaminetetraacetic acid; kDa, kilodalton(s); KLH, Keyhole limpet hemocyanin.

Acknowledgements: This study was supported in part through a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (20592175 and 24592836 to KN).

1. Introduction

Aspartic proteinases are acidic proteolytic enzymes that have a bilobal structure with two domains (Hsu et al., 1977). The active site of the aspartic proteinase contains two aspartate residues positioned in the middle of a cleft between the N- and C-terminal domains of the molecule and is partially covered by a hairpin loop, termed the flap or S1 subsite, protruding from the N-terminal domain (Sepulveda et al., 1975). Aspartic proteinases form a multigenic family that is widely distributed in organisms, and the major members of this family can be arranged into distinct clusters of orthologous groups, including cathepsin D, cathepsin E, nothepsin, renin, BACE, pepsin, and the fetal forms,

pepsin Y and pepsin F (Borrelli et al., 2006; Carginale et al., 2004; Kageyama, 2002; Vassar et al., 1999; Xin et al., 2000).

The human napsin A (*NAPSA*) transcript is primarily expressed in the lung and kidney, but minor expression of *NAPSA* has also been observed in the prostate, connective tissue and the eye. *NAPSA* is expressed in alveolar type II cells and well-differentiated lung adenocarcinomas, whereas *NAPSA* expression is weak in poorly differentiated tumors, making this protein a promising diagnostic marker for primary lung adenocarcinomas (Chuman et al., 1999; Dejmek et al., 2007; Hirano et al., 2003; Ueno et al., 2008). It has also been shown that NAPSA protein is present in human urine. The napsin B (NAPSB) transcript is predominantly expressed in blood and lymphoid tissues, such as tonsil, lymph node, bone marrow, and spleen. In humans, the *NAPSA* and *NAPSB* genes are tandemly located on chromosome 19q13. The napsin genes contain nine exons, and *NAPSA* and *NAPSB* have the same exon organization. Rodents express a single napsin, designated napsin A (Napsa). Thus, Tatnell et al. (1998) proposed that *NAPSA* and *NAPSB* were derived from a relatively recent gene duplication event, in evolutionary terms, after the divergence of mice and humans, though the exact duplication timing and process are unknown until now. *NAPSB* has been annotated as a pseudogene because it lacks an in-frame stop codon at a position equivalent to that of human *NAPSA* (Tatnell et al., 1998). Recently, it has been reported that chimpanzee *NAPSB* contains an in-frame stop codon, suggesting that chimpanzee *NAPSB* encodes a functional aspartic protease (Puente et al., 2005). The reference allele at polymorphic sites has been defined using the most common allele obtained from the alignment of multiple individual genome sequences. However, the

major allele encodes a loss-of-function variant, reflecting an inherent bias toward annotating functional genes in the reference genome; thus, events that might lead to gene inactivation are largely overlooked in automatic annotation processes (Balasubramanian et al., 2011). In this context, *NAPSB* orthologs in other species are typically annotated as *NAPSA*-like genes, and no *NAPSB* orthologs have been described in other species. Here, we investigated the evolution and gene organization of the two napsins, *NAPSA* and *NAPSB*, and traced the evolutionary origin of the subfamily of these genes in animals. Moreover, we compared the enzymatic activity and three-dimensional structure of the NAPSA and NAPSB proteins.

2 Materials and methods

2.1 *In silico* analyses

2.1.1 Identification of NAPSB genes

Napsins and related aspartic protease sequences were identified in the Ensembl and GenBank databases for the following species with available genome sequences: *Homo sapiens* (human), *Pan troglodytes* (common chimpanzee), *Pan paniscus* (bonobo), *Pongo abelii* (orangutan), *Macaca mulatta* (rhesus monkey), *Nomascus leucogenys* (gibbon), *Otolemur garnettii* (galago), *Sus scrofa* (pig), *Bos taurus* (cow), *Equus caballus* (horse), *Ailuropoda melanoleuca* (giant panda), *Canis lupus familiaris* (dog), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Ornithorhynchus anatinus* (platypus), *Gallus gallus* (chicken), *Xenopus laevis* (African clawed frog), *Xenopus tropicalis* (western clawed frog), *Latimeria chalumnae* (coelacanth), *Takifugu rubripes* (pufferfish), *Chionodraco hamatus*

(crocodile icefish), *Clupea harengus* (Atlantic herring), *Danio rerio* (zebrafish), *Sparus aurata* (gilthead seabream), *Oryzias latipes* (medaka), and *Haemonchus contortus* (barber pole worm). The accession numbers for the sequences used in phylogenetic analysis are listed in Table 1. The nucleotide sequences adjacent to poly(A) at the 3'-terminus were aligned manually. Predicted or known nucleotide and protein sequences for the identified loci were aligned using ClustalW.

2.1.2 Sequence analysis

The alignment of the napsin amino acid sequences and related proteases was performed in ClustalW using standard settings (Gonnet weight matrix, gap opening=10 and gap extension = 0.2). All positions containing gaps and missing data were eliminated. The analysis involved 34 amino acid sequences. There were a total of 282 positions in the final dataset. The evolutionary history was inferred using two methods. A Neighbor-Joining (NJ) tree (Saitou and Nei, 1987) was reconstructed in MEGA5 (Tamura, et al., 2011). The bootstrap consensus tree inferred from 1000 replicates (Felsenstein, 1985) represented the evolutionary history of the taxa analyzed (Felsenstein, 1985). The evolutionary distances were computed using the Dayhoff matrix-based method (Schwarz and Dayhoff, 1979) and are presented as the number of amino acid substitutions per site. The rate variation among sites was modeled with a gamma distribution (shape parameter = 2). The same alignment was also used to generate a Maximum Likelihood (ML) tree based on the Jones model (Jones, et al., 1992). Initial tree(s) for the heuristic search were obtained automatically using the following method. When the number of common sites was < 100 or less than one

fourth of the total number of sites, the maximum parsimony method was used; otherwise, the BIONJ method was used, with an MCL distance matrix. A discrete Gamma distribution was used to model evolutionary rate differences among sites (4 categories (+G, parameter = 1.8314)). The trees were drawn to scale, with branch lengths representing the number of substitutions per site. The trees were reconstructed using FigTree version 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/). We estimate synonymous and non-synonymous substitution rates as pairwise comparisons between sequences, using codeml program (run mode = –2, CodonFreq = 2), that is included in PAML 4.6 package (Yang, 2007). Codon-based Z-test to selection ($P < 0.05$) was carried out by using MEGA5 to estimate nucleotide sequence distances from synonymous and non-synonymous sites with the Nei-Gojobori model in standard error determined from 1000 bootstrap replicates (Nei and Gojobori, 1986).

2.1.3 Comparative genomics and neighboring gene families

The synteny of the NAPSA and NAPSB loci and their flanking genes was obtained using the NCBI Map Viewer and Ensembl Genome Browser.

2.1.4 Three-dimensional comparative modeling

The crystal structure of human renin (PDB: 1hrn) was used as a template for 3D modeling. Homology modeling of NAPSA and NAPSB was performed using the SWISS-MODEL Workspace (Arnold et al., 2006). Based on the results obtained, NAPSA and NAPSB were superimposed with human pepsin A (PDB: 1pso) using the PyMOL Version 1.5.0.3 (Schrödinger).

2.2 Recombinant protein expression and analyses

2.2.1 Preparation of the human NAPSA-FLAG and NAPSB-FLAG constructs

The cDNA clone accession numbers BI824756 (IMAGE: 5174913) and BQ073045 (IMAGE: 5757357), encoding *NAPSA* and *NAPSB*, respectively, were purchased from Invitrogen. The putative full-length region encoding NAPSA was PCR-amplified from the BI824756 clone using the Pfu Turbo high-fidelity polymerase (Stratagene) according to the manufacturer's instructions with the primers NapsEcor1atgF (5'-TGCTGGAATTCCCGGGATGTCTCCACCAC-3'; the *EcoRI* site is underlined) and NapsaXhoR (5'-TAGTCCTCGAGGAACTGCGCCTGCGCAG-3'; the *XhoI* site is underlined). The PCR products were digested with *EcoRI* and *XhoI*. A linker containing the DYKDDDDK sequence was prepared from two annealed oligonucleotides: xhoIFLAGxbaIF (5'-TCGAG*GACTACAAGGACGACGATGATAAGTGA*T-3'; the *XhoI* site is underlined, and the FLAG coding sequence with a stop codon is italicized) and xhoIFLAGxbaIR (5'-CTAGA*TCACTTATCATCGTCGTCCTTGTAGTC*C-3'; the *XbaI* site is underlined, and the complementary FLAG coding sequence with a stop codon is italicized). The PCR product and linker were subcloned into the *EcoRI-XbaI* sites of pcDNA3 (Invitrogen). The NAPSB-FLAG construct was constructed from the BQ073045 clone as described above, except NapsbxhoR (5'-TAGTCCTCGAGGTACTGCGCCTGCGCGGTC-3'; the *XhoI* is site underlined) was used as a reverse primer. The *XhoI* site introduced into NAPSA-FLAG was subsequently restored to the wild type sequence using the QuikChange site-directed mutagenesis kit

(Stratagene) and the primer pair 5'-GCAGGCGCAGTTCCCCGGGGACTACAAGGACG-3' (the wild-type sequence is underlined) 5'-CGTCCTTGTAGTCCCCGGGGAACTGCGCCTGC-3'(the wild-type sequence is underlined).

2.2.2 Transfection and purification of recombinant NAPSA and NAPSB

Human HEK293 cells were transfected with the NAPSA-FLAG-pcDNA3 or NAPSB-FLAG-pcDNA3 vector using SuperFect (Qiagen) according to the manufacturer's instructions. The cells were incubated with G418 sulfate, and the G418-resistant colonies were screened through a proteinase assay (see below) or Western blotting using anti-napsin antibodies or the anti-DYKDDDDK (Wako Chemicals) antibody. The lysate from cells transfected with NAPSA-FLAG was centrifuged, adsorbed to anti-FLAG M2 beads and washed with RIPA buffer. The FLAG-containing proteins were eluted using Tris-buffered saline containing 100 μg/ml 3X FLAG peptide (Sigma) or 3 M sodium isothiocyanate.

2.2.3 Protease assay

The aspartic protease activity was determined fluorometrically using intramolecularly quenched peptide substrates. The cell extracts or fractions of the FLAG affinity gel were preincubated with leupeptin for 10 min. The commercially available aspartic proteinase substrates (MOCAc-Ala-Pro-Ala-Lys-Phe-Phe-Arg-Leu-Lys(Dnp)-NH_2, proteinase A and pepsin, 3216-v; MOCAc-Gly-Lys-Pro-Ile-Leu-Phe-Phe-Arg-Leu-Lys(Dnp)-D-Arg-NH_2, cathepsin D/E, 3200-v; MOCAc-Gly-Ser-Pro-Ala-Phe-Leu-Ala-Lys(Dnp)-D-Arg-NH_2, cathepsin E, 3225-v; and MOCAc-Ser-Glu-Val-Asn-Leu-Asp-Ala-Glu-Phe-Arg-Lys(Dnp)-

Arg-Arg-NH_2, BACE, 3212-v; Peptide Institute) were subsequently added and incubated for up to 2 h. At the end of the incubation, the fluorescence was measured at 328 nm and 393 nm as excitation and emission wavelengths, respectively, using a fluorescence spectrophotometer F4000 (Hitachi). The background measurement was conducted using the same manner as described above, except 1 μM of pepstatin A (an aspartic protease inhibitor, Peptide Institute) was added during the preincubation.

3. Results

3.1 A SNP of human napsin B pseudogene that restores the functional gene

The human reference genome is a haploid sequence derived as a composite from multiple individuals. However, the major allele encodes a loss-of-function variant, reflecting an inherent bias toward annotating functional genes in the reference genome; thus, events that might lead to gene inactivation are largely overlooked in automatic annotation processes (Balasubramanian et al., 2011). A search for the SNP in the human dbSNP Build 137 database revealed a SNP, rs11879785, that converts the human *NAPSB* pseudogene to a functional gene. Fig. 1 represents a world map that shows the allele frequency obtained from the HapMap populations (Altshuler et al., 2010). Yoruba in Ibadan, Nigeria (YRI) has the highest frequency of the active NAPSB form at 4.5%. In addition, the active allele was also identified in African Ancestry in SW USA (4.2%), Mexican Ancestry in Los Angeles, CA, USA (3.1%), Chinese in Metropolitan Denver, CO, USA (1.8%), and Gujarati Indians in Houston, Texas, USA (0.6%). The inactive allele was practically fixed in populations with Northern and Western European ancestry (CEU) and Far East Asians (HCB and JPT).

This annotation problem also contributes to erroneous gene annotation in other species. A Blast search using the human NAPSA amino acid sequence revealed annotated or non-annotated genes from tetrapods and fishes but not from birds or invertebrates. To further identify non-annotated genes, blastn searches using the human *NAPSB* mRNA sequence to trace archive Whole Genome Shotgun submission databases were implemented. Several napsin genomic sequences were also available in the gorilla and green anole lizard, but their sequences were incomplete and thereby excluded from the analysis. The *NAPSB* orthologs were distributed only in primates, including the marmoset, while the *NAPSA* orthologs were the only napsin gene in other species, strongly supporting the idea of a recent duplication event. One exception is the prosimian primate greater galago genome, which contains only *NAPSA* (GenBank: NW_003852606). Because the sequence data are represented in an unplaced scaffold reference state, it was unclear whether the *NAPSB* was absent from the grate galago genome. Fig. 2 shows a comparison of the amino acid sequences of the napsins compared with those of the other aspartic proteases. The intramolecular disulfide bonds characteristic of mammalian aspartic proteinases are conserved in all napsins. Several potential N-linked oligosaccharide attachment sites were identified in the same location as the glycosylation motif in nothepsin and cathepsin E (at Asn26 in human NAPSA), human renin and cathepsin D (at Asn67 and Asn183, respectively), but not in other aspartic proteinases (Asn268). The RGD sequence is the most distinctive feature of the NAPSA, NAPSB, and mouse Napsa (Tatnell, et al., 1998). However, the RGD sequence is not conserved in all napsins; the sequence has been replaced with RGN in the dog and giant panda (not shown) and is completely absent in the

platypus, frog, and fish (Fig. 2). Similarly, the C-terminal extension of napsin is absent in the platypus, frog, and fish. The NAPSB contains a unique Arg at position 287, which we will discuss in section 3.5.

We also examined whether the pseudogenization of *NAPSB* is human-specific. Both *NAPSA* and *NAPSB* genes are located reciprocally on human chromosome 19q13 and are organized into nine exons (Fig. 3 and 5). The cDNA sequences of human *NAPSA* and *NAPSB* at their respective 3´ untranslated regions are both short, with 75% identity between the TGA/TGC codon and the AGTAAA putative polyadenylation site motif present in both sequences at approximately 20 bp upstream from the poly(A) tail. The sequence comparisons showed that the stop codon disruption in *NAPSB* is human-specific; the corresponding regions of *NAPSB* in the other available sequences of primate species contain an in-frame stop codon at a position equivalent to that of human *NAPSA*, ruling out an ancestral polymorphism (Fig. 3). Collectively, these data strongly support the conclusion that a duplication event occurred during the early stages of primate evolution to generate *NAPSB*, followed by human-specific *NAPSB* pseudogenization.

3.2 Phylogenetic analyses of napsins

To clarify the duplication timings and the evolutionary relationships between *NAPSA* and *NAPSB*, we constructed phylogenetic trees using NJ and ML. Thus, potential inconsistencies, reflecting the use of a single method, were avoided. Each tree reconstitution showed similar relationships between the retrieved sequences, with minor differences. Both analyses identified two major clades: one clade comprising *pepsins*,

cathepsin E, and *nothepsin* and the other clade included *cathepsin D*, *cathepsin D2*, *napsin*, and *renins*. The *napsin* clades were divided into three species: amniotes, frogs, and fish. A *Sparus aurata* gene annotated as *cathepsin D* (GenBank: AAB88862) was grouped with fish *napsins* and *cathepsin D2* (Fig. 4). Two polytomies remain unsolved: the *napsin* and *cathepsin D* clades represented a triplet branching pattern in the NJ tree and a quadruplet branching pattern in the ML tree. Adding more sequence data for the *napsin* orthologs might resolve these polytomies; however, we currently do not have enough sequence data.

3.3 Arrangement of napsin genes in vertebrate chromosomes

Existing unsolved polytomies prompted us to focus on the gene organization of napsin genes in four eutherian species (human, chimpanzee, marmoset, and mouse), a frog (clawed frog), and two fish (medaka and pufferfish), as the data could provide powerful insights with respect to gene origin. The *napsin* loci were well conserved between the various species, with two distinct settings. In the eutherian species, the napsin genes were located between the *NR1H2* and *KCNC3* genes. A partial cDNA sequence, GenBank accession number XM_002829607, was annotated as orangutan *NAPSA*, however, its genomic topology (*KCNC3- XM_002829607- LOC100455861- NR1H2*) suggested that this sequence represented an *NAPSB* ortholog. In fact, when this orangutan *NAPSA* sequence was used as a query, the top Blast hit was a chimpanzee *NAPSB* (GenBank: XP_530061), and the amino acid identities between orangutan *NAPSA* and chimpanzee *NAPSB* and *NAPSA* (GenBank: XP_524345) were 95% and 85%, respectively. A similar topology was observed in the genomes of the Western clawed frog (Fig. 5B) and the anole lizard

(Ensembl: ENSACAG00000005091), although these primary assembly units have not been identified in any assembled chromosomes or linkage groups. In medaka, an annotated napsin A gene located between *ATF4* and *SMCR7L* was identified as similar to *cathepsin D2* in pufferfish. However, in the corresponding region of the human chromosome (22q13), no napsins were located between these two genes (Fig. 5C). These results indicate that the frog *napsin* gene is orthologous to the eutherian *napsin*, while the fish *napsin* clade, which contains *cathepsin D2*, is not an ortholog. However, it is certain that the napsin gene evolved from an ancestral protease, which could have been present before the divergence of amniotes from amphibians.

3.4 Evidence of purifying selection of the *NAPSB* gene

We next tested for evidence of the purifying selection by estimating the non-synonymous/synonymous substitution rate (*d*N/*d*S) distribution among amino acid sites. Duplication has an important role in the creation of novel genes. Through the redundancy generated by duplication, one of the paralogous copies can escape the pressure of negative selection and accumulate. A codon-based test of purifying selection for analysis between *NAPSA* and *NAPSB* of 4 simian primate species is shown in Table 2. Most of the *d*N/*d*S rates between each pair of napsin genes were significantly lower than 1.0. Interestingly, *d*N/*d*S rates (≈1) observed between macaque *NAPSB* and the other species' *NAPSB* indicate neutral selection. Among the non-synonymous substitutions in the macaque *NAPSB*, a loss-of-function amino acid substitution occurred at one of the active site aspartates (Asp217, Figure 2). These results supports the notion that *NAPSB* is not just a duplicated copy of

NAPSA and that the gene products of human *NAPSB* active allele and chimpanzee *NAPSB* may be functional. Nevertheless, the occurrence of two independent loss-of-function events of *NAPSB* in the primate species strongly suggests that the loss of NAPSB activity may be evolutionary advantageous.

3.5 Recombinant protein expression and analyses

As the pseudogenization event of *NAPSB* were both human-specific and nearly fixed, we explored the functional implications through the enzymatic activities of recombinant human NAPSA and NAPSB. Full-length NAPSA and NAPSB constructs were generated with a C-terminal FLAG tag to facilitate purification and expressed stably in HEK293 cells. HEK293 cells do not express detectable NAPSA (Ueno et al., 2008). Stably transfected clones were established and the protease activity in the cell extract (Fig. 6A) or in partially purified fraction (Fig 6B) was assessed. NAPSA cleaved the synthetic substrates of BACE, proteinase A/pepsin, and cathepsin E; however, we could not detect proteolytic activity toward acid-denatured hemoglobin or ovalbumin, which is widely used as an aspartic proteinase substrate (data not shown). Notably, the cleavage sites KF-FR of 3216-v (proteinase A/pepsin substrate), IL-F-FR of 3200-v (cathepsin D/E substrate), and AF-LA of 3225-v (cathepsin E substrate) match the criteria for NAPSA cleavage sites (Schauer-Vukasinovic et al., 2000), and this is the first report showing that NAPSA cleaves NL-D of 3212-v (BACE substrate, the Swedish mutation of amyloid precursor protein sequence). The proteinase A/pepsin activity in the untransfected cells was attributed to endogenous cathepsin D. In contrast, the NAPSB activity toward aspartic proteinase substrates was

marginal (Fig. 6). However, the active site cleft of NAPSB appears to be accessible, as the NAPSB protein bound to the pepstatin A agarose gel (data not shown). Thus, we cannot rule out the possibility that NAPSB maintains their activity toward biological endogenous substrate(s).

3.6 Three-dimensional homology modeling

We next explored the differences in the 3D structure between NAPSA and NAPSB. NAPSA and NAPSB were superimposed onto human pepsin A (PDB: 1pso) as shown in Fig. 7. The napsin-specific RGD sequence located in the loop at top of the 3D structure was solvent-accessible. The aspartates in the catalytic site (shown in red sticks) were located within the substrate-binding cleft in the enzyme moiety and flanked by the S1 and S1' subsites. These subsites are involved in the binding of the substrate to the enzyme and play an essential role in substrate specificity (Khan, et al., 1997). The S1 subsite (~Tyr75-Gly76-X-Gly78 in pepsin numbering) is conserved and is presented as a flexible loop. The S1'-loops of NAPSA (blue) and NAPSB (orange) are located more centrally than that of pepsin (green loop at left); thus, there is less space in the substrate binding clefts, which might explain why neither NAPSA nor NAPSB exhibit proteolytic activity toward the classical aspartic protease substrate hemoglobin. In addition, all primate NAPSB sequences present a unique Arg at position 287 (in pepsin numbering, shown in orange spheres). Arg is the most basic amino acid, and its side chain is longer than that of Gln, which is present in NAPSA and pepsin (shown in blue and magenta spheres, respectively) at the corresponding position. We also observed that the Arg287 narrows the substrate-binding

cleft of NAPSB compared with that of NAPSA or pepsin (Fig. 7). This distinct residue in the NAPSB sequences might result in the loss of catalytic activity toward the synthetic substrates tested here.

4. Discussion

In this study, we investigated the evolution and gene organization of the napsin protease family. *NAPSB* orthologs are primarily distributed in primates, while *NAPSA* orthologs are the only napsin genes in other species. The corresponding regions of *NAPSB* in the available sequences from primate species contain an in-frame stop codon at a position equivalent to that of human *NAPSA*. In addition, a rare SNP that creates a proper stop codon in human *NAPSB* pseudogene was identified using HapMap database. Thus, the minor allele encoding functional and evolutionarily conserved protein should be annotated as the gene. The human reference genome is a haploid sequence derived as a composite from multiple individuals. However, the major allele encodes a loss-of-function variant, reflecting an inherent bias toward annotating functional genes in the reference genome; thus, events that might lead to gene inactivation are largely overlooked in automatic annotation processes (Balasubramanian et al., 2011). This problem also contributes to erroneous gene annotation in other species. In addition, the tandemly duplicated genes tend to collapse into one gene by low quality sequencing and assembly. In the case of *NAPSB*, we observed no *NAPSB* ortholog was annotated as *NAPSB*, because the reference gene (human *NAPSB)* is annotated as a pseudogene. Thus, annotation based on the current human reference genome

does not provide an accurate and complete set of the genes found across all human populations (Balasubramanian et al., 2011) and the other species.

The phylogenetic analysis involved 35 amino acid sequences, including other members of aspartic proteases, and showed that *napsin* clades were divided into three species: amniotes, frogs, and fish. The fish *napsin* clade includes fish *cathepsin D2*, which has recently been identified in pufferfish as a paralog to *cathepsin D* (Kurokawa et al., 2005). A recent phylogenetic study suggested that the *cathepsin D2* gene was generated from a duplication event in the common ancestors of fish and tetrapods, followed by *cathepsin D2* gene loss in birds and higher vertebrate lineages (Feng et al., 2011). Consistent with this analysis, we observed that the tetrapod and fish *napsins* belong to distinct synteny groups, indicating a potential paralogous relationship between these genes.

Gene inactivation events can have varying effects on human phenotypes. The loss of function due to nonsense mutations has been implicated as disease causing in ~15%-30% of monogenic inherited diseases (Mort et al., 2008). It was previously proposed that, in some cases, pseudogenization could confer a selective advantage. For example, the *CASP12* gene, which encodes a cysteine protease, contains nonsense SNPs leading to premature stop codons that result in the presence of both the active and inactive forms of the genes in the human population. The premature stop variant in *CASP12* is the most common allele in human populations, with a frequency of 100% in many Eurasian populations because it confers increased resistance to severe sepsis (Wang et al., 2005; Xue et al., 2006). The *NAPSB* pseudogenization is also practically fixed in non-African populations. In addition,

the human *NAPSB* transcript is specifically expressed in lymphoid tissues, suggesting the possibility that the pseudogenization of *NAPSB* was advantageous in recent human evolution, presumably against microbes and infections. It was suggested that the translationally active chimpanzee counterpart of this gene might contribute to some of the functional differences between the human and chimpanzee immune systems (Puente et al., 2005). We report here that the enzymatic activity of human NAPSB is marginal, which most likely reflects the Arg287 substitution; however, we cannot rule out the possibility that human NAPSB maintains their activity toward biological substrate(s).Nonstop protein expression is low, and nonstop decay does not fully account for the low level of nonstop protein (Ito-Harashima et al., 2007). Notably, despite the absence of a stop codon in the *NAPSB* gene, the NAPSB protein is expressed in HEK293 cells (patent: US 6225103). Our attempt to express the NAPSB protein encoded by the nonstop-poly(A) *NAPSB* cDNA was unsuccessful (data not shown). Nonstop mutations can lead to the continued and inappropriate translation of mRNA in the 3'-untranslated region. Nonstop mRNA is rapidly degraded, the translation of nonstop mRNA is repressed, and nonstop proteins are cotranslationally degraded (Ito-Harashima et al., 2007). The current model suggests that a polylysine tag at the C-terminus of nonstop proteins, which results from the translation of the poly(A) tail, causes translational repression and the enhanced cotranslational degradation of the nascent peptide (Vasudevan et al., 2002).

Another *NAPSB* SNP, rs634091, which converts the codon GGC to CGC (Gly122Arg), has been proposed; thus, the clone containing the minor allele Arg122 would not generate an active enzyme (Tatnell et al., 1998). The Arg122 minor allele frequency is 0.07 in the

HapMap YRI population. In the present study, the *NAPSB* cDNA clone (GenBank: BQ073045) contained Gly122 and nonstop (rs11879785) major alleles. The distance between rs634091 and rs11879785 is 3 kb. Thus, we attempted to identify haplotypes for the two SNPs in HapMap individuals of YRI (Altshuler et al., 2010), and no individuals possessing both SNPs with minor alleles were found. Thus, it is reasonable to conclude that that Gly122Arg mutation was derived in a single individual with the rs1879785 major allele; that is, this mutation occurred more recently than the human-specific *NAPSB* pseudogenization. In addition, macaque NAPSB protein appears to have lost its protease activity by the active site amino acid substitution. These independent loss-of-function events of *NAPSB* in the primate species strongly suggests that the loss of NAPSB activity may be evolutionary advantageous.

Finally, we have found that the S1'-loops of NAPSA and NAPSB are located more centrally than those of pepsin in the 3D structure homology modeling. Some of the S1' residues are important for the specificity and catalytic efficiency of pepsin A and chymosin (Kageyama, 2004). It has also been proposed that the wider substrate cleft of fish pepsin might accommodate larger substrates more efficiently, thus contributing to the specific activity toward larger substrates, such as hemoglobin and its digestion intermediates (Tanji et al., 2009). On the contrary, our 3D modeling of NAPSA and NAPSB revealed less space in the active site, and it would be interesting to elucidate in detail how the catalytic functions are affected, including the substrate specificity and the specific activity toward large substrates, such as hemoglobin. We also observed that the Arg287 narrows the substrate-binding cleft of NAPSB compared with that of NAPSA or pepsin. This distinct

residue in the NAPSB sequences might result in the loss of catalytic activity toward biological endogenous substrate(s) tested in this study.

In summary, we conclude that the *napsin* family members have been present before the divergence of amniotes from amphibians. The *NAPSB* was duplicated from *NAPSA* during the early stages of primate evolution and the subsequent loss of the *NAPSB* function (i.e. protease activity) during primate evolution. We propose that a minor allele in human *NAPSB* and primate *NAPSB* orthologs, which create a proper stop codon, encode functional, and evolutionarily conserved protein should be annotated as the gene.

Figure legends

Fig. 1. The distribution of *NAPSB* alleles in the HapMap populations. The circles area is proportional to chromosomal sample count size, and the filled area indicates the population of the active NAPSB allele. The abbreviation, allele frequency, and number of samples are ASW, African ancestry in Southwest USA, 0.042, 96; CEU, Utah residents with Northern and Western European ancestry from the CEPH collection, 0.000, 120; HCB, Han Chinese in Beijing, China, 0.000, 90; CHD, Chinese in Metropolitan Denver, Colorado, 0.018, 164; GIH, Gujarati Indians in Houston, Texas, 0.006, 170; JPT, Japanese in Tokyo, Japan, 0.000, 88; LWK, Luhya in Webuye, Kenya, 0.029, 172; MEX, Mexican ancestry in Los Angeles, California, 0.031, 98; MKK, Maasai in Kinyawa, Kenya, 0.022, 274; and YRI, Yoruba in Ibadan, Nigeria, 0.045, 112.

Fig. 2. The alignment of the amino acid sequences for napsin and related proteases. The 35-amino acid sequences, including the NAPSB encoded by the minor active allele, were

aligned in MEGA5 using the ClustalW plugin. The conserved catalytic aspartic acids are shown in bold face and underlined. Green indicates identical residues, and yellow indicates homologous substitutions. The putative N-glycosylation sites are underlined. The positions of the conserved cysteine residues involved in disulfide bond formation and the RGD motif are indicated with C and RGD, respectively. The NAPSB-specific residue Arg287 is indicated with a red face. The species abbreviations are Hsa (*Homo sapiens*), Pat (*Pan troglodytes*), Nol (*Nomascus leucogenys*), Mam (*Macaca mulatta*), Mum (*Mus musculus*), Oan (*Ornithorhynchus anatinus*), Gag (*Gallus gallus*), Xtr (*Xenopus tropicalis*), Xla (*Xenopus laevis*), Lac (*Latimeria chalumnae*), Tru (*Takifugu rubripes*), Chh (*Chionodraco hamatus*), Clh (*Clupea harengus*), Dar (*Danio rerio*), Spa (*Sparus aurata*), and Orl (*Oryzias latipes*).

Fig. 3. The alignment of the 3'-terminus mRNA sequences for mammalian napsins. Both *NAPSA* and *NAPSB* are located reciprocally on human chromosome 19q13, with 9 exons. The coding and untranslated regions are represented by filled and open boxes, respectively. The nucleotide sequence alignment shows the human-specific pseudogenization of *NAPSB*, which lacks an in-frame stop codon. The stop codons are shown in bold and italics. The deduced polyadenylation signals are underlined. Species abbreviations are Hsa (*Homo sapiens*), Pat (*Pan troglodytes*), Pap (*Pan paniscus*), Poa (*Pongo abelii*), Mam (*Macaca mulatta*), Nol (*Nomascus leucogenys*), Caj (*Callithrix jacchus*), Sus (*Sus scrofa*), Bot (*Bos taurus*), Eqc (*Equus caballus*), Caf (*Canis lupus familiaris*), Mum (*Mus musculus*), and Rno (*Rattus norvegicus*).

Fig. 4. Phylogenetic tree of selected aspartic proteases. The evolutionary history was inferred using the Neighbor-Joining method (A) and the Maximum Likelihood method based on the Jones et al. model (B). The robustness of the tree was assessed through 1,000 bootstrap replicates of the data. The numbers at the branches indicate the number of bootstrapping tests that resulted in the marked grouping, and values close to the total used (100) indicate reliable branches. The labels indicate a three-letter abbreviation for the species name with shortened protease names. Species abbreviations are Hsa (*Homo sapiens*), Otg (*Otolemur garnettii*), Mum (*Mus musculus*), Oan (*Ornithorhynchus anatinus*), Gag (*Gallus gallus*), Xla (*Xenopus laevis*), Xtr (*Xenopus tropicalis*), Lac (*Latimeria chalumnae*), Tru (*Takifugu rubripes*), Chh (*Chionodraco hamatus*), Clh (*Clupea harengus*), Dar (*Danio rerio*), Spa (*Sparus aurata*), Orl (*Oryzias latipes*), and Hac (*Haemonchus contortus*).

Fig. 5. The arrangement of napsin genes in vertebrate chromosomes. (A) The genomic organization of the napsin loci in human, chimpanzee, marmoset, and mouse. In humans and chimpanzees, the napsin locus is located on chromosome 19 in the reverse orientation. The arrows on the boxes indicate the direction of transcription. The coding regions are represented by filled boxes. The gray box indicates a pseudogene, and the white box indicates a presumed gene. (B) The genomic organization of *napsin* and its flanking genes in the mouse and the clawed frog. The broken lines drawn between loci indicate an orthologous homology between individual genes. (C) The synteny between fish *napsin/cathepsin D2* loci and human chromosome 22q13. The abbreviations are *KCNC3*, potassium voltage gated channel Shaw-related subfamily member 3; *NAPSA*, napsin A;

NAPSB, napsin B; *NR1H2*, nuclear receptor subfamily 1 group H member 2; *Pold1*, polymerase (DNA directed) delta 1 catalytic subunit; *smcr7l*, Smith-Megenis syndrome region candidate 7-like; *apold1*, apolipoprotein L domain containing 1; *atf4*, activating transcription factor 4; *ctsd2*, cathepsin D2; *acpp*, acid phosphatase, prostate; Akt1S1, Akt1 substrate1; Fuz, fuzzy homolog (Drosophila); and *Med25*, mediator complex subunit 25.

Fig. 6. The protease activityof NAPSA and NAPSB expression in HEK293 cells. (A) Untransfected (293), NAPSA-FLAG, or NAPSB-FLAG-expressing cell proteins were incubated with commercial protease substrates for BACE, proteinase A/pepsin, or cathepsin E for 1 h. Data shown as mean ± SEM, were analyzed by Student's *t* test. Asterisk, $P<0.05$; two asterisks, $P<0.005$ (n = 3-4/group). (B) The lysate of cells transfected with NAPSA-FLAG or NAPSB-FLAG was centrifuged, adsorbed to anti-FLAG M2 beads and washed. The bound proteins were eluted using 3X FLAG peptide. The protease activity was measured using several commercial substrates was determined as described under "Materials and Methods 2.2.3".

Fig. 7. The 3D structures of NAPSA and NAPSB were obtained through homology modeling. The sequence identity of mature region of pepsin A was 45% and 42% with NAPSA and NAPSB, respectively. NAPSA and NAPSB were superimposed onto human pepsin A (PDB: 1pso) and shown as a ribbon model: pepsin A, gray; NAPSA, blue; and NAPSB, orange. Pepsin A is shown as a complex with pepstatin A; the active site Asp (red) and pepstatin A (yellow) residues are shown as a stick model. The S1' loop corresponding to residues 288-298 of human pepsin A is shown as a green ribbon. The residues

corresponding to NAPSB-unique Arg287 are shown as a sphere model. The napsin-specific RGD is shown as a black ribbon.

References

Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P.E., Altshuler, D.M., Gibbs, R.A., de Bakker, P.I., Deloukas, P., Gabriel, S.B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker,P., Yu, F., Chang, K., Hawes, A., Lewis, L.R., Ren, Y., Wheeler, D., Gibbs, R.A., Muzny, D.M., Barnes, C., Darvishi, K., Hurles, M., Korn, J.M., Kristiansson, K., Lee, C., McCarrol, S.A., Nemesh, J., Dermitzakis, E., Keinan, A., Montgomery, S.B., Pollack, S., Price, A.L., Soranzo, N., Bonnen, P.E., Gibbs, R.A., Gonzaga-Jauregui, C., Keinan, A., Price, A.L., Yu, F., Anttila, V., Brodeur, W., Daly, M.J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S.F., Zhang, Q., Ghori, M.J., McGinnis, R., McLaren, W., Pollack, S., Price, A.L., Schaffner, S.F., Takeuchi, F., Grossman, S.R., Shlyakhter, I., Hostetter, E.B., Sabeti, P.C., Adebamowo, C.A., Foster, M.W., Gordon, D.R., Licinio, J., Manca, M.C., Marshall, P.A., Matsuda, I., Ngare, D., Wang, V.O., Reddy, D., Rotimi, C.N., Royal, C.D., Sharp, R.R., Zeng, C., Brooks, L.D., McEwen, J.E., 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467, 52-58.

Arnold, K., Bordoli, L., Kopp, J., Schwede, T., 2006. The SWISS-MODEL Workspace: A web-based environment for protein structure homology modeling. Bioinformatics 22, 195-201.

Balasubramanian, S., Habegger, L., Frankish, A., MacArthur, D.G., Harte, R., Tyler-Smith, C., Harrow, J., Gerstein, M., 2011. Gene inactivation and its implications for annotation in the era of personal genomics. Genes Dev. 25, 1-10.

Borrelli, L., De Stasio, R., Filosa, S., Parisi, E., Riggio, M., Scudiero, R., Trinchella, F., 2006. Evolutionary fate of duplicate genes encoding aspartic proteinases. Nothepsin case study. Gene 368, 101-109.

Carginale, V., Trinchella, F., Capasso, C., Scudiero, R., Riggio, M., Parisi, E., 2004. Adaptive evolution and functional divergence of pepsin gene family. Gene 333, 81-90.

Ito-Harashima, S., Kuroha, K., Tatematsu, T., Inada, T., 2007. Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast. Genes Dev. 21, 519-524.

Chuman, Y., Bergman, A., Ueno, T., Saito, S., Sakaguchi, K., Alaiya, A.A., Franzén, B., Bergman, T., Arnott, D., Auer, G., Appella, E., Jörnvall, H., Linder, S., 1999. Napsin A, a member of the aspartic protease family, is abundantly expressed in normal lung and kidney tissue and is expressed in lung adenocarcinomas. FEBS Lett. 462, 129-134.

Dejmek, A., Naucler, P., Smedjeback, A., Kato, H., Maeda, M., Yashima, K., Maeda, J., Hirano, T., 2007. Napsin A (TA02) is a useful alternative to thyroid transcription factor-1 (TTF-1) for the identification of pulmonary adenocarcinoma cells in pleural effusions. Diagn. Cytopathol. 35, 493-497.

Felsenstein, J., 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39, 783-791.

Feng, T., Zhang, H., Liu, H., Zhou, Z., Niu, D., Wong, L., Kucuktas, H., Liu X., Peatman, E., Liu, Z., 2011. Molecular characterization and expression analysis of the channel catfish cathepsin D genes. Fish Shellfish Immunol. 31, 164-169.

Hirano, T., Gong, Y., Yoshida, K., Kato, Y., Yashima, K., Maeda, M., Nakagawa, A., Fujioka, K., Ohira, T., Ikeda, N., Ebihara, Y., Auer, G., Kato, H., 2003. Usefulness of TA02 (napsin A) to distinguish primary lung adenocarcinoma from metastatic lung adenocarcinoma. Lung Cancer 41, 155-162.

Hsu, I.N., Delbaere, L.T., James, M.N., Hofmann, T., 1977. Penicillopepsin from Penicillium janthinellum crystal structure at 2.8 A and sequence homology with porcine pepsin. Nature 266, 140-145.

Jones, D.T., Taylor, W.R., Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275-282.

Kageyama, T., 2002. Pepsinogens, progastricsins, and prochymosins: structure, function, evolution, and development. Cell. Mol. Life .Sci. 59, 288-306.

Kageyama, T., 2004. Role of S'1 loop residues in the substrate specificities of pepsin A and chymosin. Biochemistry 43, 15122-15130.

Khan, A.R., Cherney, M.M., Tarasova, N.I., James, M.N., 1997. Structural characterization of activation 'intermediate 2' on the pathway to human gastricsin. Nat. Struct. Biol. 4, 1010-1015.

Kurokawa, T., Uji, S., Suzuki, T., 2005. Identification of pepsinogen gene in the genome of stomachless fish, Takifugu rubripes. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 140, 133-40.

Mort, M., Ivanov, D., Cooper, D.N., Chuzhanova, N.A., 2008. A meta-analysis of nonsense mutations causing human genetic disease. Hum. Mutat. 29, 1037-1047.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418-426.

Puente, X.S., Gutiérrez-Fernández, A., Ordóñez, G.R., Hillier, L.W., López-Otín, C., 2005. Comparative genomic analysis of human and chimpanzee proteases. Genomics 86, 638-647.

Saitou, N., Nei M., 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406-425.

Schauer-Vukasinovic, V., Bur, D., Kitas, E., Schlatter, D., Rossé, G., Lahm, H.W., Giller, T., 2000. Purification and characterization of active recombinant human napsin A. Eur. J. Biochem. 267, 2573-2580.

Schwarz, R., Dayhoff, M., 1979. Matrices for detecting distant relationships. In: Dayhoff M., (ed.), Atlas of protein sequences, National Biomedical Research Foundation, pp 353-358.

Sepulveda, P., Marciniszyn, J.Jr., Liu, D., Tang, J., 1975. Primary structure of porcine pepsin. III. Amino acid sequence of a cyanogen bromide fragment, CB2A, and the complete structure of porcine pepsin. J. Biol. Chem. 250, 5082-5088.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol. Biol. Evol. 28, 2731-2739

Tanji, M., Yakabe, E., Kubota, K., Kageyama, T., Ichinose, M., Miki, K., Ito, H., Takahashi, K., 2009. Structural and phylogenetic comparison of three pepsinogens from Pacific bluefin tuna: molecular evolution of fish pepsinogens. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 152, 9-19.

Tatnell, P.J., Powell, D.J., Hill, J., Smith, T.S., Tew, D.G., Kay, J., 1998. Napsins: new human aspartic proteinases. Distinction between two closely related genes. FEBS Lett. 441, 43-48.

Ueno, T., Elmberger, G., Weaver, T.E., Toi, M., Linder, S., 2008. The aspartic protease napsin A suppresses tumor growth independent of its catalytic activity. Lab. Invest. 88, 256-263.

Vassar, R., Bennett, B.D., Babu-Khan, S., Kahn, S., Mendiaz, E.A., Denis, P., Teplow, D.B., Ross, S., Amarante, P., Loeloff, R., Luo, Y., Fisher, S., Fuller, J., Edenson, S., Lile, J., Jarosinski, M.A., Biere, A.L., Curran, E., Burgess, T., Louis, J.C., Collins, F., Treanor, J., Rogers, G., Citron, M., 1999. Beta-secretase cleavage of Alzheimer's amyloid precursor protein by the transmembrane aspartic protease BACE. Science 286, 735-741.

Vasudevan, S., Peltz, S.W., Wilusz, C.J., 2002. Non-stop decay--a new mRNA surveillance pathway. Bioessays 24, 785-788.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Yang, Z., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Li, D., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Zhang, J., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Zheng, H., Ren, Y., Yang, L., Gao,Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G.K., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., Li, S., Yang, H., Wang, J., 2008. The diploid genome sequence of an Asian individual. Nature 456, 60-65.

Xin, H., Stephans, J.C., Duan, X., Harrowe, G., Kim, E., Grieshammer, U., Kingsley, C., Giese, K., 2000. Identification of a novel aspartic-like protease differentially expressed in human breast cancer cell lines. Biochim. Biophys. Acta 1501, 125-137.

Xue, Y., Daly, A., Yngvadottir, B., Liu, M., Coop, G., Kim, Y., Sabeti, P., Chen, Y., Stalker, J., Huckle, E., Burton, J., Leonard, S., Rogers, J., Tyler-Smith, C., 2006. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am. J. Hum. Genet. 78, 659-670.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586-1591.

Fig. 1

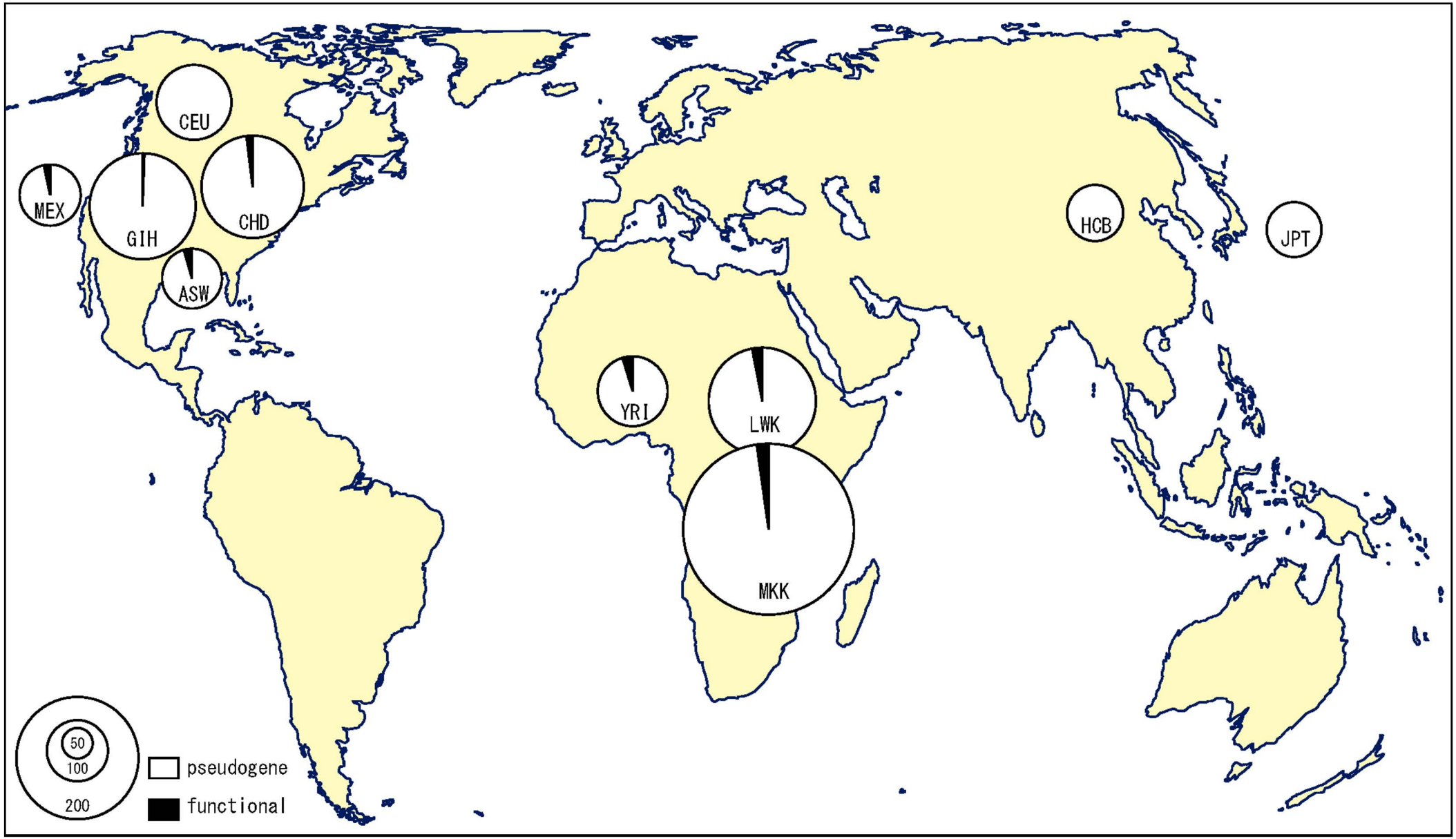
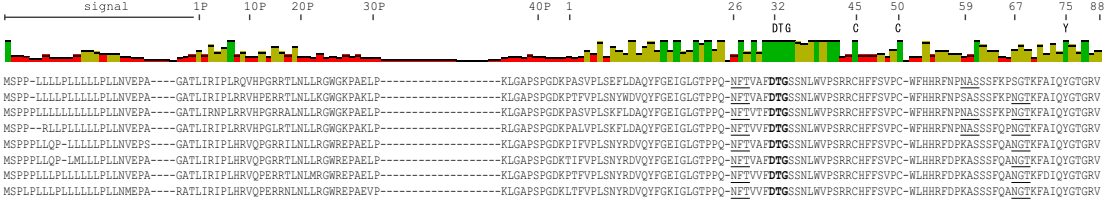
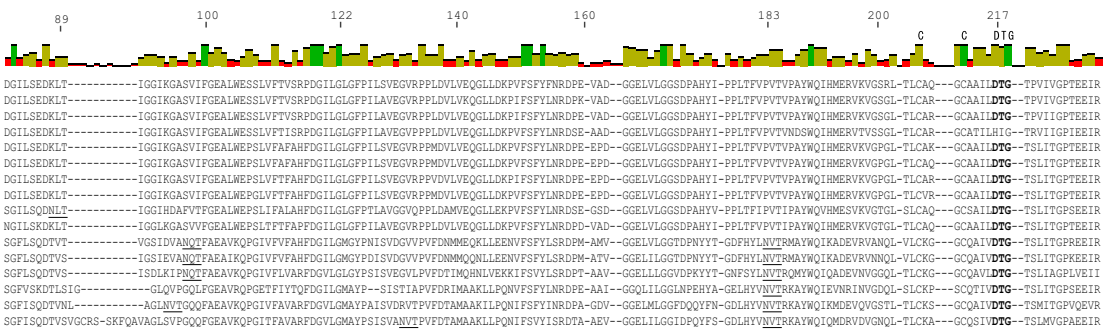

CEU
MEX
GIH
CHD
ASW
YRI
LWK
MKK
HCB
JPT
50
100
200
pseudogene
functional

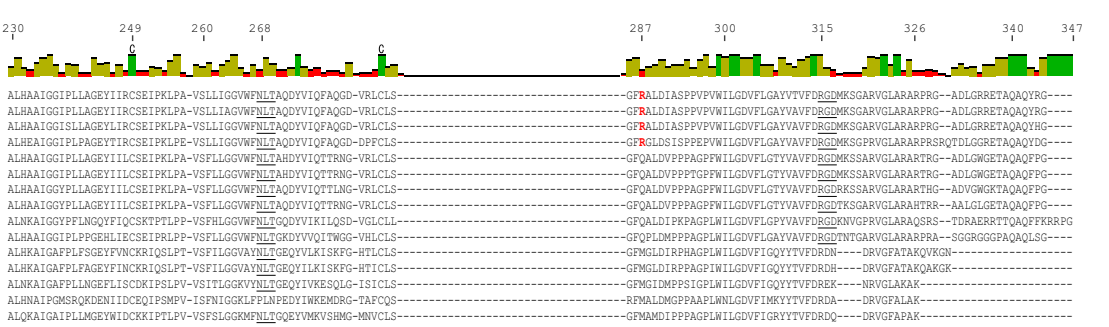
Fig. 2


signal
1P
10P
20P
30P
40P
1
26
32
45
50
59
67
75
DTG
C
C
Y

1. HsaNAPSB
2. PatNAPSB
3. NolNAPSB
4. MamNAPSB
5. HsaNAPSA
6. PatNAPSA
7. NolNAPSA
8. MamNAPSA
9. MumNapsa
10. OanNapsa
11. XtrNapsa
12. XlaNapsa
13. LacNapsa
14. TruCtsd2
15. DarNapsa
16. OrlNapsa
17. SpaCtsd
18. HsaCTSD
19. TruCtsd1
20. ChhCtsd
21. ClhCtsd
22. DarCtsd
23. TruRen
24. DarRen
25. HsaREN
26. TruNts
27. HsaNTS
28. ChhNts
29. DarNts
30. GagCtse
31. HsaCTSE
32. XlaCtse
33. HsaPEPC
34. GagPep
35. HsaPEPA4


89
100
122
140
160
183
200
217
C
C
DTG


1. HsaNAPSB
2. PatNAPSB
3. NolNAPSB
4. MamNAPSB
5. HsaNAPSA
6. PatNAPSA
7. NolNAPSA
8. MamNAPSA
9. MumNapsa
10. OanNapsa
11. XtrNapsa
12. XlaNapsa
13. LacNapsa
14. TruCtsd2
15. DarNapsa
16. OrlNapsa
17. SpaCtsd
18. HsaCTSD
19. TruCtsd1
20. ChhCtsd
21. ClhCtsd
22. DarCtsd
23. TruRen
24. DarRen
25. HsaREN
26. TruNts
27. HsaNTS
28. ChhNts
29. DarNts
30. GagCtse
31. HsaCTSE
32. XlaCtse
33. HsaPEPC
34. GagPep
35. HsaPEPA4


230
249
260
268
287
300
315
326
340
347
C
C

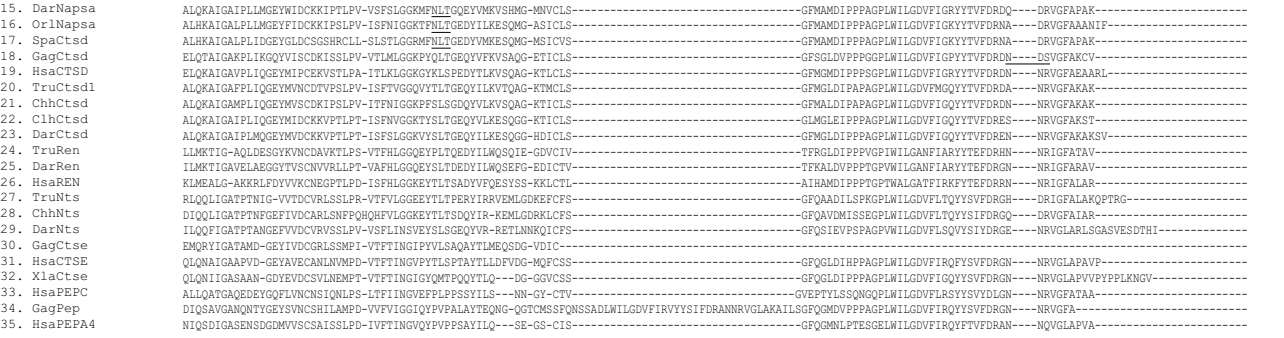

1. HsaNAPSB
2. PatNAPSB
3. NolNAPSB
4. MamNAPSB
5. HsaNAPSA
6. PatNAPSA
7. NolNAPSA
8. MamNAPSA
9. MumNapsa
10. OanNapsa
11. XtrNapsa
12. XlaNapsa
13. LacNapsa
14. TruCtsd2
15. DarNapsa
16. OrlNapsa
17. SpaCtsd
18. HsaCTSD
19. TruCtsd1
20. ChhCtsd
21. ClhCtsd
22. DarCtsd
23. TruRen
24. DarRen
25. HsaREN
26. TruNts
27. HsaNTS
28. ChhNts
29. DarNts
30. GagCtse
31. HsaCTSE
32. XlaCtse
33. HsaPEPC
34. GagPep
35. HsaPEPA4

Fig. 3

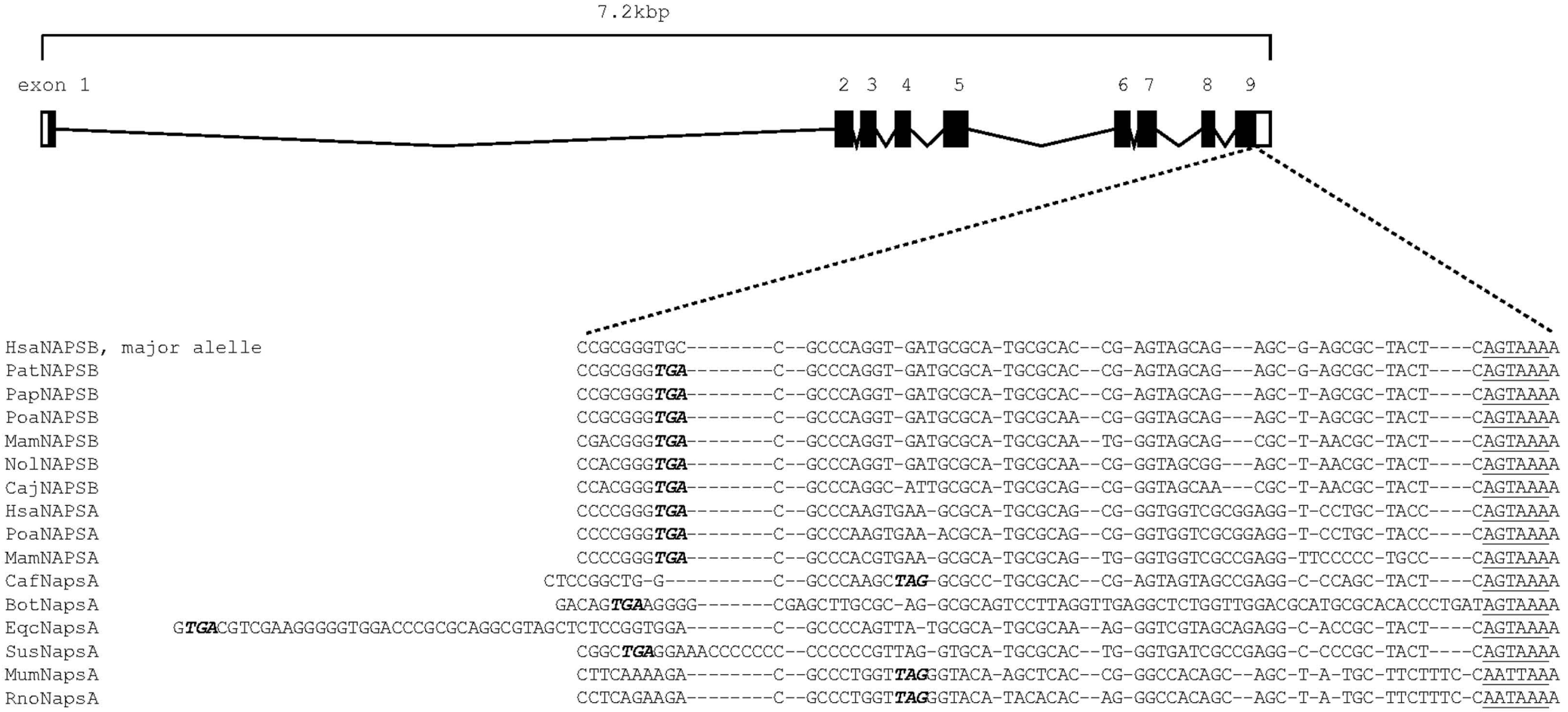

7.2kbp
exon 1
2 3 4 5
6 7 8 9
HsaNAPSB, major alelle CCGCGGGTGC--------C--GCCCAGGT-GATGCGCA-TGCGCAC--CG-AGTAGCAG---AGC-G-AGCGC-TACT----CAGTAAAA
PatNAPSB CCGCGGGTGA--------C--GCCCAGGT-GATGCGCA-TGCGCAC--CG-AGTAGCAG---AGC-G-AGCGC-TACT----CAGTAAAA
PapNAPSB CCGCGGGTGA--------C--GCCCAGGT-GATGCGCA-TGCGCAC--CG-AGTAGCAG---AGC-T-AGCGC-TACT----CAGTAAAA
PoaNAPSB CCGCGGGTGA--------C--GCCCAGGT-GATGCGCA-TGCGCAA--CG-GGTAGCAG---AGC-T-AGCGC-TACT----CAGTAAAA
MamNAPSB CGACGGGTGA--------C--GCCCAGGT-GATGCGCA-TGCGCAA--TG-GGTAGCAG---CGC-T-AACGC-TACT----CAGTAAAA
NolNAPSB CCACGGGTGA--------C--GCCCAGGT-GATGCGCA-TGCGCAA--CG-GGTAGCGG---AGC-T-AACGC-TACT----CAGTAAAA
CajNAPSB CCACGGGTGA--------C--GCCCAGGC-ATTGCGCA-TGCGCAG--CG-GGTAGCAA---CGC-T-AACGC-TACT----CAGTAAAA
HsaNAPSA CCCCGGGTGA--------C--GCCCAAGTGAA-GCGCA-TGCGCAG--CG-GGTGGTCGCGGAGG-T-CCTGC-TACC----CAGTAAAA
PoaNAPSA CCCCGGGTGA--------C--GCCCAAGTGAA-ACGCA-TGCGCAG--CG-GGTGGTCGCGGAGG-T-CCTGC-TACC----CAGTAAAA
MamNAPSA CCCCGGGTGA--------C--GCCCACGTGAA-GCGCA-TGCGCAG--TG-GGTGGTCGCCGAGG-TTCCCCC-TGCC----CAGTAAAA
CafNapsA CTCCGGCTG-G----------C--GCCCAAGCTAG-GCGCC-TGCGCAC--CG-AGTAGTAGCCGAGG-C-CCAGC-TACT----CAGTAAAA
BotNapsA GACAGTGAAGGGG-------CGAGCTTGCGC-AG-GCGCAGTCCTTAGGTTGAGGCTCTGGTTGGACGCATGCGCACACCCTGATAGTAAAA
EqcNapsA GTGACGTCGAAGGGGGTGGACCCGCGCAGGCGTAGCTCTCCGGTGGA--------C--GCCCCAGTTA-TGCGCA-TGCGCAA--AG-GGTCGTAGCAGAGG-C-ACCGC-TACT----CAGTAAAA
SusNapsA CGGCTGAGGAAACCCCCCC--CCCCCCGTTAG-GTGCA-TGCGCAC--TG-GGTGATCGCCGAGG-C-CCCGC-TACT----CAGTAAAA
MumNapsA CTTCAAAAGA--------C--GCCCTGGTTAGGGTACA-AGCTCAC--CG-GGCCACAGC--AGC-T-A-TGC-TTCTTTC-CAATTAAA
RnoNapsA CCTCAGAAGA--------C--GCCCTGGTTAGGGTACA-TACACAC--AG-GGCCACAGC--AGC-T-A-TGC-TTCTTTC-CAATAAAA

Fig. 4

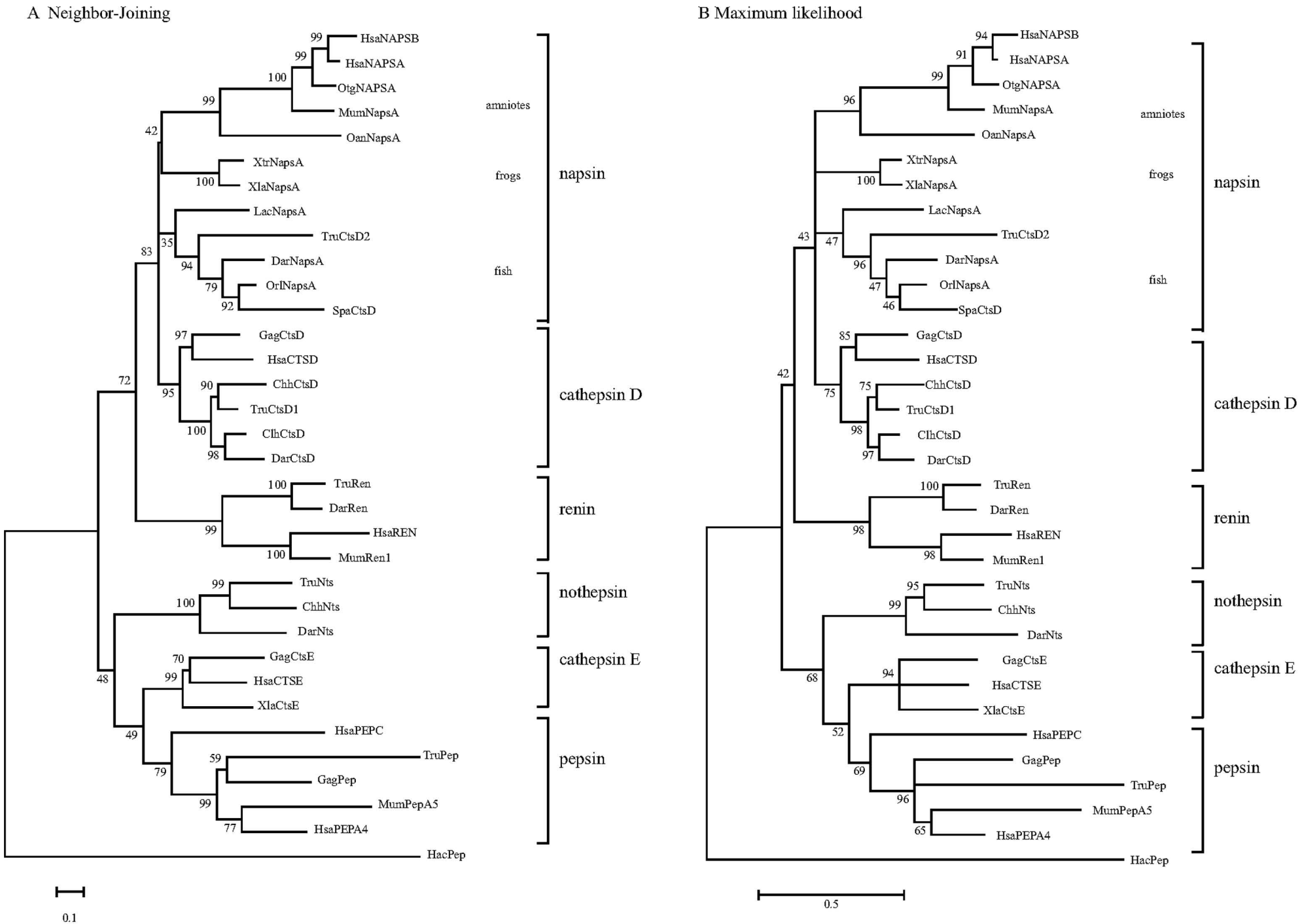

A Neighbor-Joining
HsaNAPSB
HsaNAPSA
OtgNAPSA
MumNapsA
OanNapsA
XtrNapsA
XlaNapsA
LacNapsA
TruCtsD2
DarNapsA
OrlNapsA
SpaCtsD
GagCtsD
HsaCTSD
ChhCtsD
TruCtsD1
ClhCtsD
DarCtsD
TruRen
DarRen
HsaREN
MumRen1
TruNts
ChhNts
DarNts
GagCtsE
HsaCTSE
XlaCtsE
HsaPEPC
TruPep
GagPep
MumPepA5
HsaPEPA4
HacPep
amniotes
frogs
fish
napsin
cathepsin D
renin
nothepsin
cathepsin E
pepsin
0.1
B Maximum likelihood
HsaNAPSB
HsaNAPSA
OtgNAPSA
MumNapsA
OanNapsA
XtrNapsA
XlaNapsA
LacNapsA
TruCtsD2
DarNapsA
OrlNapsA
SpaCtsD
GagCtsD
HsaCTSD
ChhCtsD
TruCtsD1
ClhCtsD
DarCtsD
TruRen
DarRen
HsaREN
MumRen1
TruNts
ChhNts
DarNts
GagCtsE
HsaCTSE
XlaCtsE
HsaPEPC
GagPep
TruPep
MumPepA5
HsaPEPA4
HacPep
amniotes
frogs
fish
napsin
cathepsin D
renin
nothepsin
cathepsin E
pepsin
0.5

Fig. 5

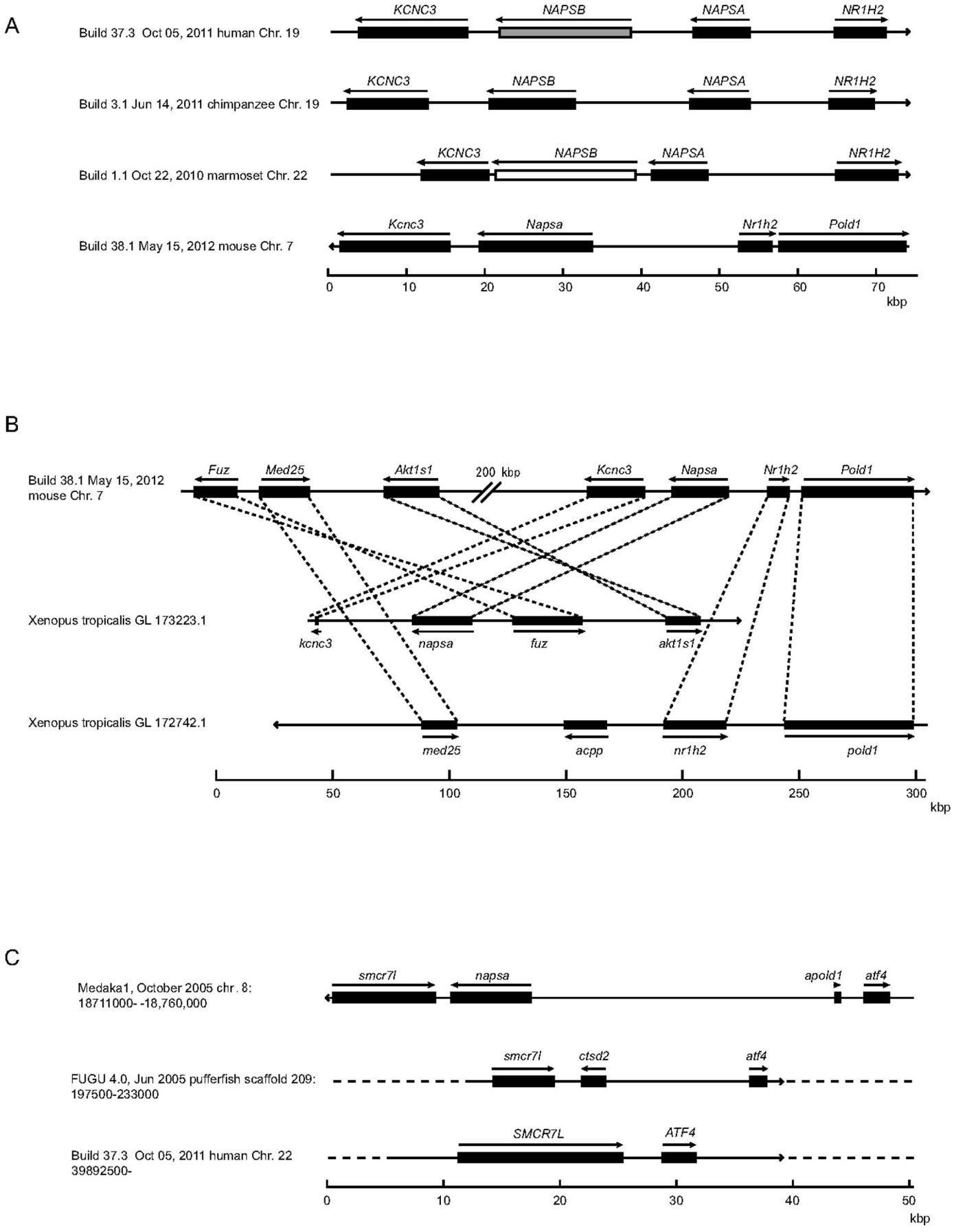
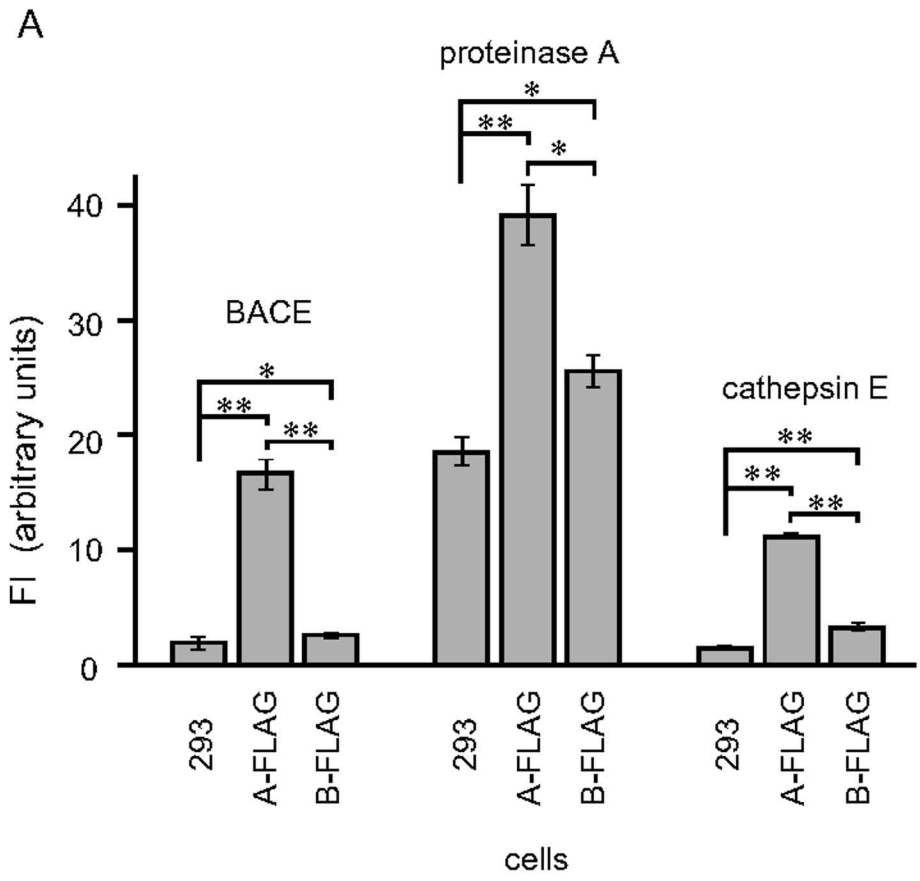

A
Build 37.3 Oct 05, 2011 human Chr. 19
KCNC3
NAPSB
NAPSA
NR1H2
Build 3.1 Jun 14, 2011 chimpanzee Chr. 19
KCNC3
NAPSB
NAPSA
NR1H2
Build 1.1 Oct 22, 2010 marmoset Chr. 22
KCNC3
NAPSB
NAPSA
NR1H2
Build 38.1 May 15, 2012 mouse Chr. 7
Kcnc3
Napsa
Nr1h2
Pold1
0
10
20
30
40
50
60
70
kbp
B
Build 38.1 May 15, 2012 mouse Chr. 7
Fuz
Med25
Akt1s1
200 kbp
Kcnc3
Napsa
Nr1h2
Pold1
Xenopus tropicalis GL 173223.1
kcnc3
napsa
fuz
akt1s1
Xenopus tropicalis GL 172742.1
med25
acpp
nr1h2
pold1
0
50
100
150
200
250
300
kbp
C
Medaka1, October 2005 chr. 8: 18711000- -18,760,000
smcr7l
napsa
apold1
atf4
FUGU 4.0, Jun 2005 pufferfish scaffold 209: 197500-233000
smcr7l
ctsd2
atf4
Build 37.3 Oct 05, 2011 human Chr. 22 39892500-
SMCR7L
ATF4
0
10
20
30
40
50
kbp

Fig. 6


A
proteinase A
BACE
cathepsin E
FI (arbitrary units)
0
10
20
30
40
293
A-FLAG
B-FLAG
cells

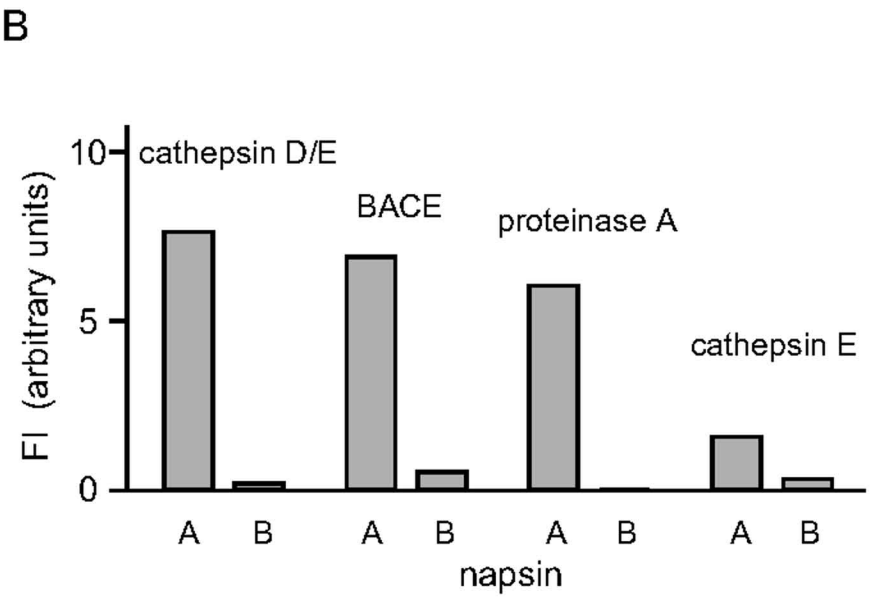

B
cathepsin D/E
BACE
proteinase A
cathepsin E
FI (arbitrary units)
0
5
10
A
B
napsin

Fig. 7

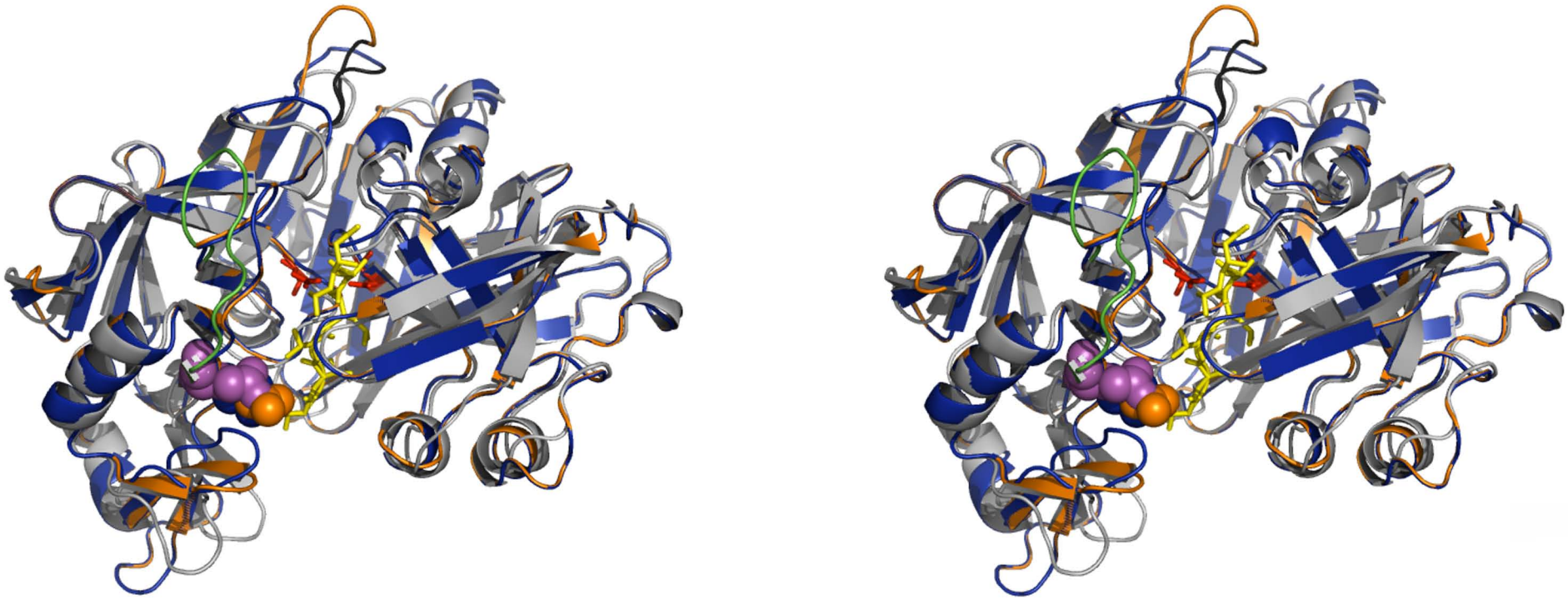


Table 1. List of accession numbers for all aspartic protease sequences used in the phylogenetic analysis.

		Accession number	
Organism	Name	Protein	mRNA
Homo sapiens	HsaNAPSA	NP_004842.1	NM_004851.1
	HsaNAPSBp		NR_002798.1
	HsaCTSE	NP_001901.1	
	HsaCTSD	NP_001900.1	
	HsaREN	NP_000528.1	
	HsaPEPC	NP_002621.1	
	HsaPEPA4	NP_001073276.1	
Pan troglodytes	PatNAPSA	XP_524345.2	XM_524345.3
	PatNAPSB	XP_530061.2	XM_530061.3
Pan paniscus	PapNAPSB		XM_003813624.1
Pongo abelii	PoaNAPSB		XM_002829607.2[a]
Nomascus leucogenys	NolNAPSA		XM_003269801.1
	NolNAPSB	XP_003269848.1	XM_003269800.1
Macaca mulatta	MamNAPSA	XP_001116026.1	ENSMMUT00000018797
	MamNAPSB	ENSMMUP00000031507	ENSMMUT00000038406
Callithrix jacchus	CajNAPSA		XM_003735639.1
	CajNAPSB		XR_144592.1
Otolemur garnettii	OtgNAPSA	XP_003801583.1	
Sus scrofa	SusNapsA		XM_003127363.2
Bos taurus	BotNapsA		XM_002695127.1
Equus caballus	EqcNapsA		XM_001490835.1
Ailuropoda meranoleuca	AimNapsA	XP_002917936.1	
Canis lupus famioliaris	CafNapsA	XP_533610.2	XM_533610.3
Mus musculus	MumNapsA	NP_032463.1	NM_008437.1
	MumPepA5	NP_067428.2	
	MumRen1	NP_112469.1	
Rattus norvegicus	RnoNapsA		NM_031670.2
Ornithorhynchus anatinus	OanNapsA	ENSOANP00000019807	
Gallus gallus	GagPep	ENSGALP00000000593	
	GagCtsE	ENSGALP00000001138	
	GagCtsD	ENSGALP00000010662	
Xenopus laevis	XlaNapsA	NP_001083566.1	
	XlaCtsE	BAC57453.1	
Xenopus tropicalis	XtrNapsA	NP_001005701.1	
Latimeria chalumnae	LacNapsA	ENSLACP00000016743	
Takifugu rubripes	TruPep	NP_001072051.1	
	TruCtsD1	NP_001072052.1	
	TruCtsD2	NP_001072053.1	
	TruRen	NP_001072054.1	
	TruNts	NP_001072055.1	
Chionodraco hamatus	ChhNts	CAA11580.1	
	ChhCtsD	CAA07719.1	
Clupea harengus	ClhCtsD	AAG27733.1	
Danio rerio	DarNapsA	AAH56836.1	
	DarCtsD	NP_571785.1	
	DarRen	AAO31713.1	
	DarNts	NP_571879	
Oryzias latipes	OrlNapsA	ENSORLP00000016894	
Sparus aurata	SpaCtsD	AAB88862	
Haemonchus contortus	HacPep	CAA96571.1	

[a] This sequence is annotated as *NAPSA* in the GenBank.

Table 2. Codon-based Test of Purifying Selection for analysis between sequences.

	HsaNAPSB	PatNAPSB	NolNAPSB	MamNAPSB	HsaNAPSA	PatNAPSA	NolNAPSA	MamNAPSA	MumNapsa
HsaNAPSB		**0.037**	**0.005**	0.497	**0.004**	**0.010**	**0.012**	**0.000**	**0.000**
PatNAPSB	0.246		**0.007**	0.439	**0.001**	**0.004**	**0.004**	**0.000**	**0.000**
NolNAPSB	0.342	0.320		0.095	**0.000**	**0.001**	**0.000**	**0.000**	**0.000**
MamNAPSB	1.039	1.001	0.694		**0.031**	0.064	0.051	**0.006**	**0.000**
HsaNAPSA	0.549	0.508	0.464	0.706		0.333	**0.002**	**0.000**	**0.000**
PatNAPSA	0.585	0.539	0.499	0.756	0.720		**0.007**	**0.000**	**0.000**
NolNAPSA	0.633	0.586	0.477	0.767	0.371	0.415		**0.001**	**0.000**
MamNAPSA	0.524	0.513	0.442	0.651	0.280	0.313	0.352		**0.000**
MumNapsa	0.346	0.339	0.343	0.340	0.323	0.336	0.307	0.301	

The caluculation of d_N / d_S (below diagonal) by PAML codeml program. d_S and d_N are the numbers of synonymous and nonsynonymous substitutions per site, respectively. The probability of rejecting the null hypothesis of strict-neutrality ($d_N = d_S$) in favor of the alternative hypothesis ($d_N < d_S$) (above diagonal) is shown. Values of *P* less than 0.05 are considered significant at the 5% level are shown in bold.